The Islamic University–Gaza
Research and Postgraduate Affairs
Faculty of Information Technology
Master of Information Technology

الجامعـــــــة الإســـلاميـة – غـزة
شئون البحث العلمي والدراسات العليا
كـليــــة تكنولوجيا المعلومات
ماجستير تكنولوجيا المعلومات

Predicting Lecturer's Performance Using Datamining Techniques Based on Lecturer's Characteristics and Historical Student Evaluation of Lecturer

(Islamic University-Gaza case study)

توقع أداء المحاضر باستخدام تقنيات تنقيب البيانات استناداً إلى خصائص المحاضر وتقييمات الطلبة السابقة للمحاضر

(الجامعة الإسلامية بغزة كحالة دراسية)

Student's Name

Mohammed J. Tahrawi

Supervised by

Dr. Ashraf Alattar

A thesis submitted in partial fulfilment of the requirements for the degree of Master of information technology

August/2016

إقـــــرار

أنا الموقع أدناه مقدم الرسالة التي تحمل العنوان:

Predicting Lecturer's Performance Using Datamining Techniques Based on Lecturer's Characteristics and Historical Student Evaluation of Lecturer

(Islamic University-Gaza case study)

توقع أداء المحاضر باستخدام تقنيات تنقيب البيانات استناداً إلى خصائص المحاضر وتقييمات الطلبة السابقة للمحاضر

(الجامعة الإسلامية بغزة كحالة دراسية)

أقر بأن ما اشتملت عليه هذه الرسالة إنما هو نتاج جهدي الخاص، باستثناء ما تمت الإشارة إليه حيثما ورد، وأن هذه الرسالة ككل أو أي جزء منها لم يقدم من قبل الاخرين لنيل درجة أو لقب علمي أو بحثي لدى أي مؤسسة تعليمية أو بحثية أخرى.

Declaration

I understand the nature of plagiarism, and I am aware of the University's policy on this.

The work provided in this thesis, unless otherwise referenced, is the researcher's own work, and has not been submitted by others elsewhere for any other degree or qualification.

Student's name:	محمد جميل حسن الطهراوي	اسم الطالب:
Signature:	محمد الطهراوي	التوقيع:
Date:	2016. 8. 15	التاريخ:

بسم الله الرحمن الرحيم

الجامعة الإسلامية-غزة
The Islamic University of Gaza

مكتب نائب الرئيس للبحث العلمي والدراسات العليا
هاتف داخلي: 1150

الرقم: ج س غ/35/ Ref:

التاريخ: 2016/08/15 Date:

نتيجة الحكم على أطروحة ماجستير

بناءً على موافقة شئون البحث العلمي والدراسات العليا بالجامعة الإسلامية بغزة على تشكيل لجنة الحكم على أطروحة الباحث/ **محمد جميل حسن الطهراوي** لنيل درجة الماجستير في **كلية *تكنولوجيا المعلومات*** **برنامج تكنولوجيا المعلومات** وموضوعها:

توقع أداء المحاضر باستخدام تقنيات تنقيب البيانات استنادا إلى خصائص المحاضر وتقييمات الطلبة السابقة للمحاضر - الجامعة الإسلامية بغزة كحالة دراسية

Predicting Lecturer's Performance Using Datamining Techniques Based on Lecturer's Characteristics and Historical Student Evaluation of Lecturer Islamic University-Gaza case study

وبعد المناقشة العلنية التي تمت اليوم الاثنين 12 ذو القعدة 1437هـ، الموافق 2016/08/15م الساعة الواحدة ظهراً في قاعة المؤتمرات بمبنى القدس، اجتمعت لجنة الحكم على الأطروحة والمكونة من:

د. أشرف محمد العطار	**مشرفاً و رئيساً**	....
د. راوية فوزي عوض الله	**مناقشاً داخلياً**	Rawia Awad Allah ....
د. محمد عبد اللطيف راضي	**مناقشاً خارجياً**	....

وبعد المداولة أوصت اللجنة بمنح الباحث درجة الماجستير في **كلية *تكنولوجيا المعلومات*/ برنامج تكنولوجيا المعلومات.**

واللجنة إذ تمنحه هذه الدرجة فإنها توصيه بتقوى الله ولزوم طاعته وأن يسخر علمه في خدمة دينه ووطنه.

والله ولي التوفيق ،،،

نائب الرئيس لشئون البحث العلمي والدراسات العليا

أ.د. عبدالرؤوف علي المناعمة

Abstract

One of the most expensive resources in the higher educational process is lecturers, so most of the educational institutes spend a lot of effort and consume the HR resources to allocate the best lecturer to their students. Which will maximize the learning potential of the students according to the lecturer's qualifications, skills and abilities. Therefore, the challenge is how to predict lecturer's performance based on lecturer characterstics ak2nd historical student assessments of previous lecturers.

Our approach uses data mining techniques to analyze existing data composed of lecturer academic and non-academic characteristics and predict prospective lecturer's performance in order to support the decision-making in lecturer selection. We use four datamining techniques: Decision tree, K-Nearest Neighbor, Multinomial Logistic Regression and Naïve Bayesian. The models are trained and evaluated on a subset of the data. The model with the highest prediction outcome is selected.

We used data belonging to the academic staff of Islamic University – Gaza (IUG) that taught in 11 semesters (from second semester 2011/2012 to summer semester 2014/2015). The dataset contains an attribute for the overall evaluation result from the end-of-semester questionnaires routinely filled by students, and 28 attributes of lecturer characteristics. The overall student questionnaire result is aggregated over all sections of a course that is taught by the lecturer in one semester.

Based on training and evaluation of the four techniques mentioned above, and if we suppose that the closest prediction of the true evaluation is true, we can say that the models is predicting the evaluation truly or far from true in one step as next: Multinomial Logistic Regression gave the highest accuracy of 86.5%. Decision tree, K-Nearest Neighbor, and Naïve Bayesian gave accuracies of 85.0%, 86.4%, and 80.9%, respectively.

Key words: Prediction, Lecturer Performance, Datamining, Lecturer's Characteristics, Lecturers Evaluation

Abstract in Arabic

الملخص

يعتبر المدرسون أحد أغلى المصادر في عملية التعليم العالي. تنفق معظم المؤسسات التعليمية الكثير من الجهد وتستهلك الموارد البشرية لتوفير أفضل المحاضرين لطلابهم مما يزيد من قدرة الطلاب على التعلم من الطلاب وفقا لمؤهلات المحاضر، والمهارات والقدرات التي يتمتع بها. و يكمن التحدي هنا في كيفية توقع أداء المدرسين بناء على خصائص المدرس والتقييم السابق للطلاب لمدرسيهم السابقين.

تقوم طريقتنا على استخدام تقنيات تنقيب البيانات من أجل تحليل البيانات المتوفرة حول الخصائص الأكاديمية وغير الأكاديمية للمحاضر الجامعي، والتنبؤ بأداء المحاضرين المحتمل عند توظيفهم و هذا بهدف تعزيز عملية اتخاذ القرار حين اختيار المحاضرين.

و استخدمنا لهذا الغرض أربع تقنيات لتنقيب البيانات و هي: شبكة القرارات، تقنية الجار الأقرب K، الانحدار متعدد الحدود اللوجستية, ,تصنيف نايف بايز. ويتم تدريب النماذج وتقييمها على مجموعة فرعية من البيانات و يتم اختيار النموذج وفقا لأعلى نتائج التنبؤ.

و قد استخدمنا البيانات الخاصة بالمدرسين الاكاديميين في الجامعة الإسلامية بغزة الذين درسوًا لمدة 11 فصلا دراسيا (من الفصل الثاني للعام الجامعي 2012/2011 و حتى الفصل الصيفي للعام الجامعي 2015/2014). و اشتملت مجموعة البيانات على نتائج التقييم الكلي التي تم الحصول عليها من استبيانات يقوم الطلاب بتعبئتها و اشتملت كذلك على 28 خاصية للمحاضر. يتم تجميع النتيجة الإجمالية لاستبيان الطالب على جميع أجزاء المساق الذي يدرسه المحاضر خلال فصل دراسي واحد.

وبناء على تقييم التقنيات الأربعة المذكورة أعلاه، وعلى اعتبار أن القيمة المتنبئة المجاورة للقيمة الصحيحة صحيحة أعطى الانحدار متعدد الحدود اللوجستية أعلى دقة و مقدارها 86.5٪. و أعطت شبكة القرارات، و الجار الأقرب K، تصنيف نايف بايز دقة مقدارها 85.0٪، 86.4٪، 80.9٪ على التوالي.

كلمات مفتاحية: التنبؤ، اداء المدرسين، تنقيب البيانات، خصائص المدرسين، تقييم المدرسين.

بسم الله الرحمن الرحيم

﴿فَصَبْرٌ جَمِيلٌ﴾

[يوسف: 18]

﴿فَغَفَرْنَا لَهُ ذَلِكَ وَإِنَّ لَهُ عِنْدَنَا لَزُلْفَى وَحُسْنَ مَآبٍ﴾

[ص:25]

Dedication

To my beloved mother ***Samia***

To my beloved father ***Dr. Jamel***

To my dear and supporter wife ***Zolfa***

To my eye on the future, my lovely son ***Jamel***

To my ***sisters*** and ***brothers***

To my ***in-Laws***

To my ***relatives***

Acknowledgment

All thanks and praises to ***Allah*** who granted me the strength, support, guidance and eased the difficulties, which I faced during the accomplishment of this thesis .

I would like to thank my supervisor ***Dr. Ashraf Alattar*** for his strong support and guidance throughout the duration of this research. I am very grateful to him for working with me. It has been an honor to work with him .

I would like to express my appreciation to the ***academic staff of Information Technology College*** at the Islamic University-Gaza.

Special thanks for the director of academic affairs in Islamic University-Gaza ***Mr. Khaled Shwedeh*** and the director of Employees affairs ***Mr. Khaled Madi*** for their facilities to get the research data set.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

DM	Data mining
DT	Decision Tree
EDT	Eduactional Data Mining
IUG	Islamic University of Gaza
KNN	K-Neareast Nigbour
MLR	Multinomial Logestic Regression
NB	Naïve Bayeasian
HR	Human Resources

Chapter 1

Introduction

Chapter 1
Introduction

Through the year, the educational institutes undertake the process of allocating its resources to their right position or field, and some of these resources are allocated in a manual way, which adds more effort and consume the HR resources in this process because of the huge number of students and teachers of these institutes. The educational institutes consider students and lecturers their main assets and they look forward to improve their key process indicators by effective and efficient use of their assets (Naik, Sarma, Manjula, & Ramesh, 2011). But unfortunately, little attention has been paid to rethinking the use of existing institutional resources -such as teachers- the most important and expensive resource in the academic institutes (Miles, Darling-Hammond, & Linda, 1998).

One of the main goals of lecturer allocating process is to maximize the learning potential of the students by allocating the best teachers to the right classes according to the teacher's qualifications, skills and abilities. To achieve the previous goal, the academic institutes should be able to manage the resources such as teachers, lectures times tables, students, classes and subjects. This ability depends –in some way- on how to use the existing data about these resources in an efficient way. The existing data in the academic and educational domain offers a fertile ground for many interesting and challenging data mining applications. These applications can help administrators and academics in the academic universities and schools to improve the quality of the education and learning process (Ma, Liu, Wong, Yu, & Lee, 2000).

Data mining is the application of specific algorithms for extracting patterns from data (Fayyad, Piatetsky-Shapiro, & Smyth, 1996), another definition of data mining is the process that involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data set (Phyu, 2009). With the continues growing of educational data, and with the so much need to support taking right decision, the data mining embedding in educational systems will be in great demand on the next years.

The data mining offers a lot of methods and techniques to be benefited from the huge amount of existing data in the academic institutes and schools. Using data mining on analyzing academic related data and extract knowledge from it to support the decision making is called Educational Data mining (EDM). The Educational Data Mining community website, www.educationaldata mining.org, defines educational data mining as follows: “Educational Data Mining is an emerging discipline, concerned with developing methods for exploring the unique types of data that come from educational settings, and using those methods to better students understanding, and the settings which they learn in.”

EDM can play an important role in solving the lecturer recommendation in the academic institutes and schools. This can be done using the different data mining methods and techniques such as association role, classification and prediction depending on the history data available at the Management Information Systems (MIS) as a source, the output of applying data mining on this data can be used as a guide or heuristic base on the academic decision making process.

In this thesis, we get a data set from Islamic university of Gaza that contains some attributes about lecturers characteristics and students assessment of lecturer performance in some courses, we will analyze the patterns and roles that affect students assessment of lecturer, and using these roles to predict the performance of new teacher when he begins to teach in a new semester. This chapter contains six sections. We will display problem statement, objective, scope and limitation, Significance of the project, Research Methodology and research format respectively.

1.1 Statement of the problem

How to predict students ratings of lecturers based on characteristics of the lecturers.

1.2 Objectives

This research has a main objective that can be achieved by achieving numbers of specific, or sub, objectives.

1.2.1 Main objective

The main objective of the research is developing a data mining method to predict student’s ratings of lecturers based on the characterestics of the lecturers.

1.2.2 Specific objectives

1- Obtaining a rich dataset of historical data (student evaluation of previous lecturers).

2- Obtaining data sets of related to lecturer characteristics.

3- Choosing a suitable data mining technique (through training and testing on historical data)

4- Evaluating the performance of the selected model (through testing on actual present cases).

1.3 Scope and limitations

1- The research will focus on higher education and use Islamic University as cast study.

2- The data set, which will be used, corresponds to a specific semesters.

3- The proposed solution not is related to lecturer scheduling, only used to recommend teacher by predicting student's ratings of lecturers.

Note: This research will take into consideration the ethical and privacy issues, and the privacy related attributes in the data set will be extracted and excluding from the research implementation.

1.4 Significance

1- Enhancing lecturers' recommmendation by predicting student's ratings of lecturers.

2- The research focus on solving teacher/lecturer related problem which did not receive adequate attention in the research.

3- The used method can be generalized and applied to other domains.

4- The proposed solution could be integrated easily with the used academic information systems.

1.5 Research Methodology

Our methodology in this research consists of a number of steps each leading into one or more of the objectives presented in section 1.2 above. Table 1.1 describes the steps of our methodology and brief description of each step.

Table 1.1 Research methodology

Step	Description	Objective
1.	**Literature review**: In this step, we will review and study the previous researches conducted in fields of using data mining in educational purposes and using data mining in prediction. Also a deep reading, understanding and analysis will be conducted in the data mining topics in general.	Expanding our knowledge in the problem domain (related works and related techniques used in data mining)
2.	**Collecting and preprocessing data**: in this step we will work to find a suitable data set that contains information about university teachers and their evaluations and qualifications. Then we will apply the data mining preprocessing techniques such as cleaning, discretization, sampling, feature selection and normalization.	Increase data quality and perform data preprocessing to increase the accuracy of the mining. To do serious, effective, real-world data mining.
3.	**Apply and evaluate data mining models**: after we had processed the data set, we will apply the data mining classification techniques such as Decision Tree, Naïve Baysan, KNN and multinomial logistic regression.	Using the most usefulness and most techniques task to be considered the main used techniques.
4.	**Select and evaluate performance of the best model** through testing on actual present cases.	Evaluating the performance of the selected model.

1.6 Research Format

The research is organized in six chapters. Chapter one is an introduction, research problem, objectives, scope and significant. Chapter two is a background and state of the art. Chapter three is a related work. Chapter four is a proposed method and implementation. Chapter five is results and discussion. Chapter six presents conclusions and future work.

Chapter 2

Background and state of the art

Chapter 2
Background and state of the art

Based on the field of this research, this chapter is divided into two main sections: 1) performance evaluation, and 2) data mining. The first section explains the concept of performance evaluation, and focuses on lecturer's performance evaluation. The second section defines data mining, and explains how it is used in performance evaluation. The section also covers data preprocessing techniques, tasks of data mining, classification methods, and performance metrics for predictive modeling.

2.1 Performance evaluation

By investigated many papers we found that there were many terms used interchangeably with the term performance evaluation (Sok-Foon, Sze-Yin, & Yin-Fah, 2012) (Ahmadi & Abadi, 2013) (Wang, Dziuban, Cook, & Moskal, 2009) like performance assessment (Gupta & Suma, 2014) (Mardikyan & Badure, 2011), performance appraisal (Pal & Pal, 2013) (Peleyeju & Ojebiyi, 2013) or performance measurement (Molefe, 2010). We adopted the term performance evaluation in our thesis.

2.1.1 Definition

(Millmore, Lewis, Saunders, Thornhill, & Morrow, 2007) defined performance evaluation as a process of assessing the performance against pre-determined measures of performance, that is based on key success factors (KSF) which may include measures of deviation from the norm, tracking past achievements and measures of output and input.

(Gül, 2010) Add that performance evaluation is a process of management. It has become compulsory in human relations information system in the management of organizations.

It is clear that performance measurement monitors and reports how well someone or something is doing. In theory, it is a broad term applicable to people, things, situations, activities and organizations whilst performance management is a process that helps organizations to formulate, implement and change their strategy in order to satisfy their shareholders' needs (Millmore et al., 2007) .

2.1.2 Lecturer's performance

When we talk about evaluation performance in higher education, the first thing comes to our minds is employee performance, especially the lecturer's performance, because the lecturer is the most important and expensive resource in the academic institutes (Miles et al., 1998). This, translates into pressure on managers in the higher education environment to ensure that their staff (particularly academic staff) is working more productively and their institutions are responsive to the changing demands placed upon them by stakeholders.

To better understand the lecturer's performance evaluation; in this section, we discussed its dimensions and importance. We also presented how lecturer's performance has changed in light of the Web 2.0 technologies.

2.1.2.1 Dimensions

Performance dimensions for a university lecturer's job an approach adopted internationally in line with competency based thinking suggests that the following are some of the competencies that may be associated with lecturers' positions (Molefe, 2010):

Table 2.1 Some of the competencies that may be associated with lecturers' positions (Molefe, 2010)

communication	student and stakeholder orientation	learner assessment
interpersonal skills	innovation and creativity	organizational skills
leadership	decision-making	listening skills
self-development	judgement	project management
development of others	research	change management
change management	subject mastery	originality
commitment to quality	professional relations	critical analytical skills
the ability to challenge conventional views.		

Besides, performance evaluation is supposed to be subjective as it is measured indirectly. Performance evaluation is on the effectiveness of what personnel carry out and the understanding of their performance levels. It can be understood by performance evaluation to what extend an employee carries out the work. At the end of the evaluation, it can be understood not only the evaluation of performance but also the failure (Gül, 2010).

2.1.2.2 Importance

(Samian & Noor, 2012) emphasized the purpose of Lecturer's Performance Assessment by students is to provide teaching staff with information to make informed decisions about improving lecturers teaching. It is a positive process and should be used for the enhancement of staff development and student learning.

(Gül, 2010) show the importance of lecturer's performance in next points:

- Higher education institutions are willing to learn more about the performances of lecturers.
- What is important in evaluation of lecturers are students' perceptions.
- It is essential for higher education institutions to know students' opinions about the lecturers.
- Provides an opportunity to define students' needs and their plans.
- Improves duty and responsibility feelings of students and prevents the probable problems in the communication between students and lecturers.
- Helps the management to be ready for the unknown and unpredictable external factors in order to take necessary precautions in advance.
- Gives clues to the lecturers in terms of their development potential.

(Mardikyan & Badure, 2011) point out that student evaluations of teaching (SET) results are utilized for other purposes by providing information to different parties in educational institutions.

- Administrators use ratings in:
  - Hiring new instructors.
  - Promotion and tenure decisions.
  - Selecting faculty and graduate students for teaching awards.
  - Assigning teachers to courses.
- Instructors use SET results to:
  - Improve their teaching effectiveness.
  - Monitoring the performance of their graduate student assistants.
- Students use the ratings in selecting courses and selecting teachers.

2.1.3 Lecturer’s Performance in the World of Web 2.0

(Wang et al., 2009) point out to important issue that makes using of data mining techniques very important to develop rule based models that best predict what students can evaluate Lecturer.

They said that during the recent decade, the emerging Internet and in particular the concept of Web 2.0 impacted students’ evaluations of their instructors. This phenomenon is interacting with a generation of young people on campus who have been alternatively termed millennials, the net generation, the digital generation, and generation Y, among others. Their learning and technology characteristics are described as operating at twitch speed (miniscule response time), using parallel processing for information intake, preferring information in graphic rather than textual form, using their digital, personal, and mobile technologies to stay continually connected, preferring active rather than passive learning scenarios, incorporating play into their working lifestyles, embracing learning through virtual environments, and seeing technology as fun rather than a challenge. For them, the Web 2.0 with its sharing, communicating, blogging, text messaging, social networking, group writing though wikis, and interactive social opportunities is a seamless and continuous communication medium. These developments present a learning model far different from one-directional, teacher-to-student techniques that served as the prototype for most student evaluation of instruction research of the past decades. Today’s students experience education though online and blended courses (partly face-to-face and partly online) and extending devices, such as podcasts, chat rooms, and worldwide virtual collaborative groups. These trends have implications for students and their instructors. One example of emerging issues is the Web site http://www.ratemyprofessors.com where students formed a worldwide community to share their perceptions about their instructors’ teaching abilities. Further, they share their impressions on social networking tools, such as Facebook and MySpace, or post videos of their instructors in the act of teaching on YouTube. On many campuses students rate their professors online rather than using the paper and pencil scansheets of old. Students respond, not only to their face-to-face courses, but evaluate any number of technology mediated classes in which they might be involved. These emerging trends make it even more important to explore elements that underpin effective teaching in the eyes of students.

2.2 Data Mining

In this section, we will present some explanations about some related data mining topics as data mining defintions, data mining process, data mining classification techniques and performance metrices for predictive modeling.

2.2.1 Data mining definitions

(Han & Kamber, 2011) simply stated, data mining refers to extracting or "mining" knowledge from large amounts of data. They added that mining is a vivid term characterizing the process that finds a small set of precious nuggets from a great deal of raw material. Thus, such a misnomer that carries both "data" and "mining" became a popular choice. Many other terms carry a similar or slightly different meaning to data mining, such as knowledge mining from data, knowledge extraction, data/pattern analysis, data archaeology, and data dredging. Many people treat data mining as a synonym for another popularly used term, Knowledge Discovery from Data, or KDD. Alternatively, others view data mining as simply an essential step in the process of knowledge discovery. Knowledge discovery as a process is depicted in Figure 2.1 and consists of an iterative sequence steps.

(Larose, 2005) defined data mining as "the process of discovering meaningful new correlations, patterns and trends by sifting through large amounts of data stored in repositories, using pattern recognition technologies as well as statistical and mathematical techniques."

(Hand, Mannila, & Smyth, 2001) Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner. The relationships and summaries derived through a data mining exercise are often referred to as models or patterns.

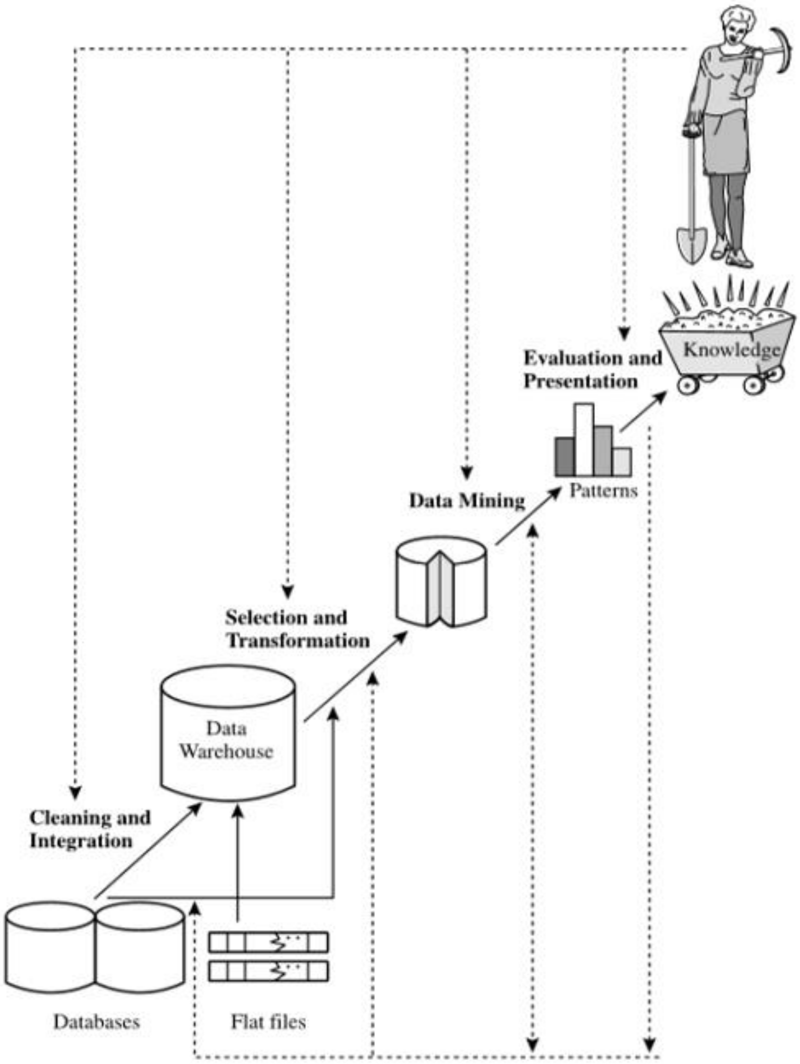

Figure 2.1 Data mining as a step in the process of knowledge discovery (Han & Kamber, 2011)

2.2.2 Data mining process

(Giudici, 2003) divided the data mining process into a series of activities from defining objectives to evaluating results. The seven activities or phases are:

1- Definition of the objectives for analysis

2- Selection, organization and pretreatment of the data

3- Exploratory analysis of the data and subsequent transformation

4- Specification of the statistical methods to be used in the analysis phase

5- Analysis of the data based on the chosen methods

6- Evaluation and comparison of the methods used and the choice of the final model for analysis

7- Interpretation of the chosen model and its subsequent use in decision processes.

We will describe the seven steps briefly in next subsections.

2.2.2.1 Definition of the objectives

Definition of the objectives involves defining the aims of the analysis. It is not always easy to define the phenomenon we want to analyze. In fact, the organization's objectives that we are aiming for, are usually clear, but the underlying problems can be difficult to translate into detailed objectives that need to be analyzed. A clear statement of the problem and the objectives to be achieved are the prerequisites for setting up the analysis correctly. This is certainly one of the most difficult parts of the process since what is established at this stage determines how the subsequent method is organized. Therefore, the objectives must be cleared and there must be no room for doubts or uncertainties.

2.2.2.2 Organization of the data

Once, the objectives of the analysis have been identified, it is necessary to select the data for the analysis. First, it is necessary to identify the data sources. Usually data is taken from internal sources that are cheaper and more reliable. This data also has the advantage of being the result of experiences and procedures of the organization itself. The ideal data source is the company data warehouse, a storeroom of historical data that is no longer subject to changes and from which it is easy to extract topic databases, or data marts, of interest. Once, a data matrix is available it is often necessary to carry out a preliminary cleaning of the data. In other words, a quality control is carried out on the available data, known as data cleaning. We will explain it in details in subsection 2.2.3 data preprocessing.

2.2.2.3 Exploratory analysis of the data

Exploratory analysis of the data involves a preliminary exploratory analysis of the data. An initial evaluation of the data's importance can lead to a transformation of the original variables to better understand the phenomenon or it can lead to statistical methods based on satisfying specific initial hypotheses. Exploratory analysis can highlight any anomalous data-items that are different from the rest. These items data will not necessarily be eliminated because they might contain information that is important to achieve the objectives of the analysis. The author think that an exploratory analysis of the data is essential because it allows the analyst to predict which data mining methods might be most appropriate in the next phase of the analysis. The exploratory analysis might also suggest the need for new extraction of data because the data collection is considered insufficient to achieve the set aims.

2.2.2.4 Specification of DM methods

There are various data mining methods that can be used and there are also many algorithms, so it is important to have a classification of the existing methods. The choice of method depends on the problem being studied or the type of data available. The data mining process is guided by the applications. For this reason, the methods used can be classified according to the aim of the analysis. Then we can distinguish three main classes:

- Descriptive methods.
- Predictive methods.
- Local methods.

We will explain each of them in next subsections.

2.2.2.4.1 Descriptive methods

Methods that are aiming to describe groups of data more briefly; they are also called symmetrical, unsupervised or indirect methods. Observations may be classified into groups not known beforehand (cluster analysis, Kohonen maps); variables may be connected among themselves according to links unknown beforehand (association methods, log-linear models, graphical models). In this way all the variables available are treated at the same level and there are no hypotheses of causality.

2.2.2.4.2 Predictive methods

Methods that are aiming to describe one or more of the variables in relation to all the others; they are also called asymmetrical, supervised or direct methods. This is done by looking for rules of classification or prediction based on the data. These rules help us to predict or classify the results of future of one or more response or target variables in relation to what happens to the explanatory or input variables. The main methods of this type are those developed in the field of machine learning such as the neural and decision trees but also classic statistical models such as linear and logistic regression models. We will explain some of these techniques in subsection 2.2.4 data mining classification techniques.

2.2.2.4.3 Local methods

Methods that are aiming to identify particular characteristics related to subset interests of the database; descriptive methods and predictive methods are global rather than local. Examples of local methods are association rules for analyzing transactional data.

2.2.2.5 Data analysis

Once, the data mining methods have been specified, they must be translated into appropriate algorithms for computing calculations that help us synthesize the results we need from the available database. The wide range of specialized and non-specialized software for data mining means that for most standard applications it is not necessary to develop adhoc algorithms; the algorithms that come with the software should be sufficient. Nevertheless, those managing the data mining process should have a sound knowledge of the different methods as well as the software solutions, so they can adapt the process to the specific needs of the company and interpret the results correctly when taking decisions.

2.2.2.6 Evaluation of statistical methods

To produce a final decision it is necessary to choose the best model of data analysis from the statistical methods available. Therefore, the choice of the model and the final decision rule are based on a comparison of the results obtained with the different methods. This is an important diagnostic check on the validity of the specific statistical methods that are then applied to the available data. In data mining it is rarely a good idea to use just one data mining method to analyze the data. Different methods have the potential to highlight different aspects, aspects which might otherwise have been ignored. To choose the best final model it is necessary to apply and compare various techniques quickly and simply, to compare the results that is produced and then give a business an evaluation of the different rules that is created.

In section 2.2.5 we will explain some methods to evaluate the techniques specially the prediction techniques.

2.2.2.7 Implementation of the methods

Data mining is not just an analysis of the data, it is also the integration of the results into the decision process of the company. Business knowledge, the extraction of rules and their participation in the decision process allow us to move from the analytical phase to the production of a decision engine. Once, the model has been chosen and tested with a data set, the classification rule can be applied to the whole reference population.

2.2.3 Data preprocessing

Real world data tend to be dirty, incomplete, and inconsistent. Data preprocessing techniques can improve the quality of the data, thereby helping to improve the accuracy and efficiency of the subsequent mining process. Data preprocessing is an important step in the knowledge discovery process, because the quality of decisions must be based on quality data. Detecting data anomalies, rectifying them early, and reducing the data to be analyzed can lead to huge payoffs for decision making (Han & Kamber, 2011).

In the following subsections, we introduce the most important steps in data preprocessing as seen in Figure 2.2:

2.2.3.1 Data understanding

This step can include initial data collection, data description, data exploration, and the verification of data quality. Data exploration such as viewing summary statistics (which includes the visual display of categorical variables) can occur at the end of this phase. Models such as cluster analysis can also be applied during this phase, with the intent of identifying patterns in the data (Olson & Delen, 2008)

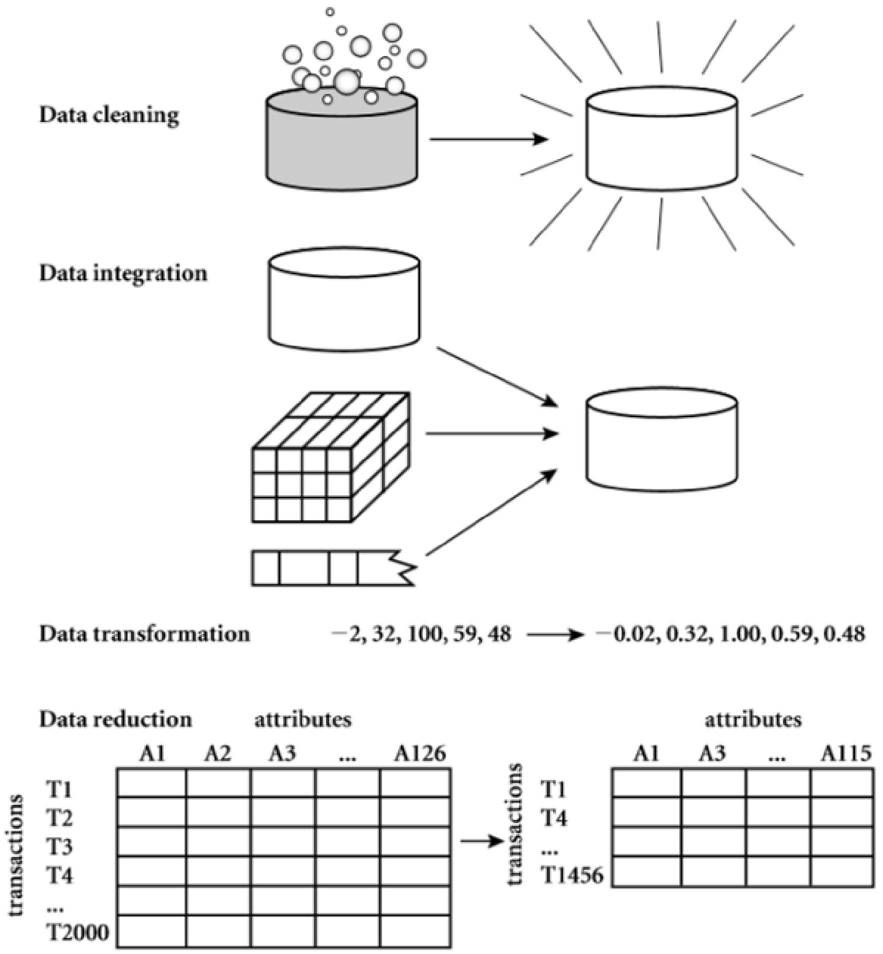

Figure 2.2 Forms of data preprocessing (Han & Kamber, 2011)

2.2.3.2 Data cleaning

(Han & Kamber, 2011) determined some basic methods for data cleaning. They are in the following subsections, missing values, noisy data and inconsistent data

2.2.3.2.1 Missing values

You can go about filling in the missing values for attribute by one of the following methods:

1. Ignore the record: This is usually done when the class label is missing.
2. Fill in the missing value manually: In general, this approach is time-consuming.
3. Use a global constant to fill in the missing value: Replace all missing attribute values by the same constant, such as a label like “Unknown” or $-\infty$.

4. Use the attribute mean to fill in the missing value: For example, suppose that the average of one attribute is X. Use this value to replace the missing value in this attribute.

5. Use the attribute mean for all samples belonging to the same classes that given tuple.

6. Use the most probable value to fill in the missing value: This may be determined with regression, inference-based tools using a Bayesian formalism, or decision tree induction.

2.2.3.2.2 Noisy data

Noise is a random error or variance in a measured variable. We can smooth out the data to remove the noise by the following data smoothing techniques:

1. Binning: Binning methods smooth a sorted data value by consulting its "neighborhood," that is the values around it. The sorted values are distributed into a number of "buckets," or bins. Because binning methods consult the neighborhood of values, they perform local smoothing, you can use the following binning methods:

    a. Partition into (equal-frequency) bins.

    b. Smoothing by bin means.

    c. Smoothing by bin boundaries.

2. Regression: Data can be smoothed by fitting the data to a function, such as with regression. Linear regression involves finding the "best" line to fit two attributes (or variables), so that one attribute can be used to predict the other.

3. Clustering: Outliers may be detected by clustering, where similar values are organized into groups, or "clusters." Intuitively, values that fall outside of these clusters may be considered outliers.

2.2.3.2.3 Inconsistent data

The data analyst should be looked out for the inconsistent data representations such as: "2004/12/25" and "25/12/2004" for date. Field overloading is another source of errors that typically results when developers squeeze new attribute definitions into unused (bit) portions of already defined attributes (e.g., using an unused bit of an attribute whose value range uses only, say, 31 out of 32 bits).

2.2.3.3 Data integration

It is the process of taking operational data from one or more sources and mapping it, field by field, onto a new data structure in the data warehouse. The common identifier problem is one of the most difficult integration issues in building a data warehouse. Essentially, this situation occurs when there are multiple system sources for the same entities and there is no clear way to identify those entities as the same. This is a challenging problem, and in many cases, it cannot be solved in an automated fashion. It frequently requires sophisticated algorithms to pair up probable matches. Another complex data-integration scenario occurs when there are multiple sources for the same data element. In reality, it is a common that some of these values are contradictory, and resolving a conflict is not a straightforward process. Just as difficult as having conflicting values is having no value for a data element in a warehouse. All these problems and corresponding automatic or semiautomatic solutions are always domain-dependent (Kantardzic, 2003).

2.2.3.4 Data transformation

In data transformation, the data are transformed or consolidated into forms appropriate for mining. Data transformation can involve the following (Han & Kamber, 2011):

- Smoothing, which works to remove noise from the data. Such techniques include binning, regression, and clustering.
- Aggregation, where summary or aggregation operations are applied to the data. For example, the daily sales data may be aggregated so as to compute monthly and annual total amounts.
- Generalization of the data, where low-level or “primitive” (raw) data are replaced by higher-level concepts through the use of concept hierarchies. For example, categorical attributes, like street, can be generalized to higher-level concepts, like city or country. Similarly, values for numerical attributes, like age, may be mapped to higher-level concepts, like youth, middle-aged, and senior.

- Normalization, where the attribute data are scaled so as to fall in a small specified range, such as−1.0 to 1.0, or 0.0 to 1.0.
- Attribute construction (or feature construction), where the new attributes are constructed and added from the given set of attributes to help the mining process.

2.2.3.5 Data reduction

Data reduction techniques can be applied to obtain a reduced representation of the data set that is much smaller in volume, yet closely maintains the integrity of the original data. That is, mining on the reduced data set should be more efficient yet produce the same (or almost the same) analytical results (Han & Kamber, 2011).

(Kantardzic, 2003) describe the three basic operations in a data reduction process, delete a column, delete a row, and reduce the number of values in a column. These operations attempt to preserve the character of the original data by deleting data that are nonessential.

2.2.3.5.1 Feature selection (Delete a column)

The operations that may mentioned here are highly application-dependent, so we can mention briefly on one approach. Like the replacement of a set of initial features with a new composite feature. For example, if samples in a data set have two features, person-height and person-weight, it is possible for some applications in the medical domain to replace these two features with only one, body-mass-index, which is proportional to the quotient of the initial two features. Final reduction of data does not reduce the quality of results; in some applications, the results of data mining are even improved. Two standard tasks are associated with producing a reduced set of features, and they are classified as:

- Feature selection: Based on the knowledge of the application domain and the goals of the mining effort, the human analyst may select a subset of the features found in the initial data set. The process of feature selection can be used as a manual or supported by some automated procedures.
- Feature composition: There are transformations of data that can have a surprisingly strong impact on the results of data-mining methods. In this sense,

the composition of features is a greater determining factor in the quality of data-mining results than the specific mining technique. In most instances, feature composition is depending on knowledge of the application, and an interdisciplinary approach to feature the composition tasks to produces significant improvements in the preparation of data.

2.2.3.5.2 Sampling (delete a row)

The largest and the most critical dimension in the initial data set is the number of cases or samples or, in other words, the number of rows in the tabular representation of data. Case reduction is the most complex task in data reduction. Already, in the preprocessing phase, we have elements of case reduction through the elimination of outliers and, sometimes, samples with missing values. But the main reduction process is still ahead. If the number of samples in the prepared data set can be managed by the selected data mining techniques, then there is no technical or theoretical reason for case reduction. In real-world data-mining applications, however, with millions of samples available, that is not the case. Practical sampling possesses one or more of the following advantages: reduced cost, greater speed, greater scope, and sometimes even higher accuracy.

2.2.3.5.3 Discretization (reduce the number of values in a column)

The task of feature-discretization techniques is to discretize the values of continuous features into a small number of intervals, where each interval is mapped to a discrete symbol. The benefits of these techniques are simplified data description and easy-to-understand data and final data-mining results. Also, more data mining techniques are applicable with discrete feature values. An "old fashioned" discretization is made manually, based on our a priori knowledge about the feature. For example, using common sense or consensus, a person's age is given at the beginning of a data-mining process as a continuous value (between 0 and 150 years) may be classified into categorical segments: child, adolescent, adult, middle age, and elderly.

2.2.4 Data mining classification techniques

(Larose, 2005) investigating the classification as a task of data mining techniques, he explains that the classification is a target categorical variable, such as income bracket. For example, could be partitioned into three classes or categories: high income, middle income, and low income. The data mining model examines a large set of records, each record is containing information on the target variable as well as a set of input or predictor variables. For example, consider the excerpt from a data set is shown in Table 2.2. Suppose that the researcher would like to be able to classify the income brackets of persons not currently in the database, based on other characteristics associated with that person, such as age, gender, and occupation. This task is a classification task, very nicely suited to data mining methods and techniques. The algorithm would proceed roughly as follows. First, examine the data set containing both the predictor variables and the (already classified) target variable, income bracket. In this way, the algorithm (software) “learns about” which combinations of variables are associated with income brackets. For example, older females may be associated with the high-income bracket. This data set is called the training set. Then the algorithm would look at new records, for no information about income bracket is available. Based on the classifications in the training set, the algorithm would assign classifications to the new records. For example, a 63-year-old female professor might be classified in the high-income bracket.

Table 2.2 Excerpt from Data Set for Classifying Income (Larose, 2005)

Subject	Age	Gender	Occupation	Income Bracket
1	47	F	Software engineer	High
2	28	M	Marketing consultant	Middle
3	35	M	Unemployed	Low

In the following subsections, we will investigate the most common data mining classification techniques that we used in this thesis.

2.2.4.1 Decision tree

Decision tree induction is the learning of decision trees from class –labeled training tuples. A decision tree is a flow chart- like tree structure, where each internal node (non leaf node) denotes test on an attribute, each branch represents an outcome of the test, and each leaf node (or terminal node) holds a class label. The topmost node in a tree is the root node. (Han & Kamber, 2011)

Figure 2.3:A decision tree for the concept buys computer, indicating whether a customer is likely to purchase a computer. Each internal (non leaf) node represents a

test on an attribute. Each leaf node represents a class (either buys computer = yes or buys comput (Han & Kamber, 2011)

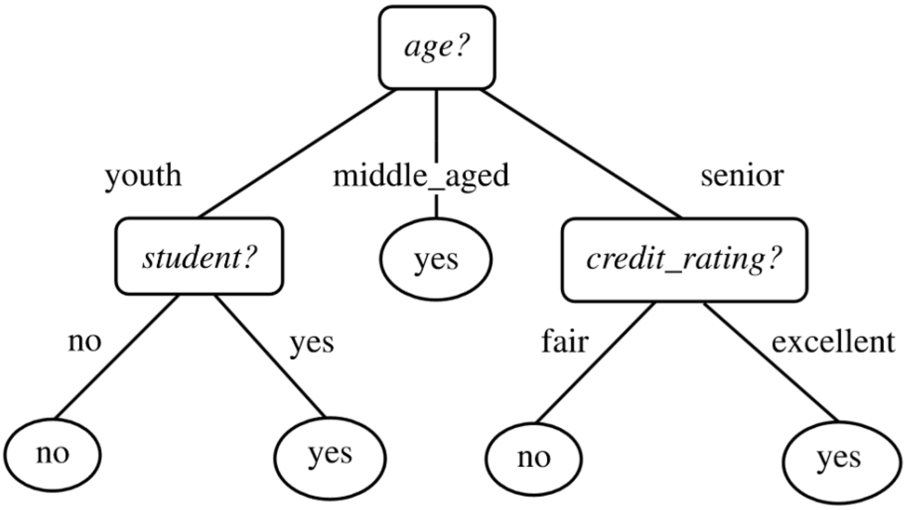

Figure 2.3 an example of decision tree for the concept buys computer

A typical decision tree is shown in Figure 2.3. It represents the concept buys computer, that it predicts whether a customer at some companies are likely to purchase a computer. Internal nodes are denoted by rectangles, and leaf nodes are denoted by ovals. Some decisions tree algorithms produce only binary trees (where each internal node branches to exactly two other nodes), whereas others can produce non binary trees. (Han & Kamber, 2011)

Decision tree classifiers are considered on of the most popular classifiers, because the construction of decision tree classifiers does not require any domain knowledge or parameter setting, and therefore is appropriate for exploratory knowledge discovery. Decision trees can handle high dimensional data. Their representation of acquired knowledge in tree form is intuitive and generally easy to assimilate by humans. The learning and classification steps of decision tree induction are simple and fast. In general, decision tree classifiers have good accuracy. Figure 2.4 summary of the basic decision tree algorithm. (Han & Kamber, 2011)

```
(1)  create a node N;
(2)  if tuples in D are all of the same class, C then
(3)      return N as a leaf node labeled with the class C;
(4)  if attribute_list is empty then
(5)      return N as a leaf node labeled with the majority class in D; // majority voting
(6)  apply Attribute_selection_method(D, attribute_list) to find the "best" splitting_criterion;
(7)  label node N with splitting_criterion;
(8)  if splitting_attribute is discrete-valued and
         multiway splits allowed then // not restricted to binary trees
(9)      attribute_list ← attribute_list − splitting_attribute; // remove splitting_attribute
(10) for each outcome j of splitting_criterion
     // partition the tuples and grow subtrees for each partition
(11)     let D_j be the set of data tuples in D satisfying outcome j; // a partition
(12)     if D_j is empty then
(13)         attach a leaf labeled with the majority class in D to node N;
(14)     else attach the node returned by Generate_decision_tree(D_j, attribute_list) to node N;
     endfor
(15) return N;
```

Figure 2.4 Basic algorithm for inducing a decision tree from training tuples (Han & Kamber, 2011)

2.2.4.2 K-Nearest Neighbor KNN

(Han & Kamber, 2011) The k nearest neighbor method was first described in the early 1950s. The method is intensive when is given large training sets, and did not gain popularity until the1960s when increased computing power became available. It has since been widely used in the area of pattern recognition. Nearest-neighbor classifiers are based on learning by analogy, that is by comparing a given test tuple with training tuples that are similar to it. The training tuples are described by n attributes. Each tuple represents a point in an n-dimensional space. In this way, all of the training tuples are stored in an n-dimensional pattern space. When it gives an unknown tuple, a k-nearest-neighbor classifier searches the pattern space for the k training tuples that are closest to the unknown tuple. These k training tuples are the k "nearest neighbors" of the unknown tuple. "Closeness" is defined in terms of a distance metric, such as Euclidean distance. The Euclidean distance between two points or tuples, say, X1 = (x11, x12, ... , x1n) and X2 =(x21, x22 , ... , x2n), is

$$dist(X_1, X_2) = \sqrt{\sum_{i=1}^{n} (x1i - x2i)^2}$$

Equation 1: The Euclidean distance

In other words, for each numeric attribute, we take the difference between the corresponding values of that attribute in tuple X1 and in tuple X2, square this difference, and accumulate it. The square root is taken of the total accumulated distance count. Typically, we normalize the values of each attribute before using Equation 1 .This helps to prevent attributes with initially large ranges from outweighing attributes with initially smaller ranges. Min-max normalization, for example, can be used to transform a value v of a numeric attribute A to v' in the range [0,1] by computing

$$v' = \frac{v - min_A}{max_A - min_A}$$

Equation 2 Min-max normalization

Where min_A and max_A are the minimum and maximum values of attribute A. For k-nearest-neighbor classification, the unknown tuple is assigned the most common class among its k nearest neighbors. When k = 1, the unknown tuple is assigned the class of the training tuple that is closest to it in pattern space. Nearest neighbor classifiers can also be used for prediction, that is to return a real-valued prediction for the given unknown tuple. In this case, the classifier returns the average value of the real-valued labels associated with the k nearest neighbors of the unknown tuple.

2.2.4.3 Naïve Bayesian

Bayesian classifiers are statistical classifiers. They can predict class membership probabilities, such as the probability that the given tuple belongs to a particular class. It has also exhibited high accuracy and speed when it is applied to large databases.

Naïve Bayesian classifiers assume that the effect of an attribute value on a given class is independent of the values of the other attributes. This assumption is called class conditional independence. It is made to simplify the computations involved and, in this sense, is considered “naïve”. Bayesian belief networks are graphical models, which unlike naïve Bayesian classifiers, allow the representation of dependencies among sub sets of attributes. Bayesian belief networks can also be used for classification.

The naïve Bayesian classifier, or simple Bayesian classifier, works as follows (Han & Kamber, 2011):

> Let D be a training set of tuples and their associated class labels. As usual, each tuple is represented by an n-dimensional attribute vector, X = (x1, x2,..., xn), depicting n measurements made of the tuple from n attributes, respectively, A1,A2,...,An.
>
> Suppose that there are m classes, C1,C2,...,Cm. Given a tuple, X, the classifier will predict that X belongs to the class having the highest posterior probability, conditioned on X. That is, the naïve Bayesian classifier predicts that tuple X belongs to the class Ci if and only if P(Ci|X)> P(Cj|X) for $1\leq j\leq m, j\neq i$. Thus we maximize P(Ci|X). The class Ci for which P(Ci|X) is maximized is called the maximum posteriori hypothesis. By Bayes' theorem
>
> $$\mathrm{P(Ci|X)} = \frac{\mathrm{P(X|Ci)P(Ci)}}{\mathrm{P(X)}}$$
>
> As P(X) is constant for all classes, only P(X|Ci)P(Ci) needed to be maximized. If the class prior probabilities are not known, then it is commonly assumed that the classes are equally likely, that is, P(C1) = P(C2) =···= P(Cm), and we would therefore maximize P(X|Ci). Otherwise, we maximize P(X|Ci) P(Ci). Note that the class prior probabilities may be estimated by P(Ci)=|Ci,D|/|D|, where |Ci,D| is the number of training tuples of class Ci in D.
>
> Given data sets with many attributes, it would be extremely computationally expensive to compute P(X|Ci). In order to reduce computation in evaluating P(X|Ci), the naive assumption of class conditional independence is made. This presumes that the values of the attributes are conditionally independent of one another, given the class label of the tuple (i.e., that there are no dependence relationships among the attributes). Thus,
>
> $$\mathrm{P(X|Ci)} = \prod_{\mathrm{k=1}}^{n} \mathrm{P(xk|Ci)}$$

$$= P(x1|Ci) \times P(x2|Ci) \times \cdots \times P(xn|Ci).$$

We can easily estimate the probabilities P(x1|Ci),P(x2|Ci),...,P(xn|Ci) from the training tuples. Recall that here xk refers to the value of attribute Ak for tuple X.

In order to predict the class label of X, P(X|Ci)P(Ci) is evaluated for each class Ci. The classifier predicts that the class label of tuple X is the classCi if and only if

$$P(X|Ci)P(Ci) > P(X|Cj)P(Cj) \text{ for } 1 \leq j \leq m, j \neq i.$$

In other words, the predicted class label is the class Ci for which P(X|Ci)P(Ci) is the maximum.

2.2.4.4 Multinomial Logistic regression

(Han & Kamber, 2011) Logistic regression models the probability of some events are occurring as a linear function of a set of predictor variables.

(Starkweather & Moske, 2011) Multinomial logistic regression is used to predict categorical placement in or the probability of category membership on a dependent variable based on multiple independent variables. The independent variables can be either dichotomous (i.e., binary) or continuous (i.e., interval or ratio in scale). Multinomial logistic regression is a simple extension of binary logistic regression that allows to more than two categories of the dependent or outcome variable. Like binary logistic regression, multinomial logistic regression uses maximum likelihood estimation to evaluate the probability of categorical membership. Multinomial logistic regression does necessitate careful consideration of the sample size and examination for outlying cases. Like other data analysis procedures, initial data analysis should be thorough and include careful univariate, bivariate, and multivariate assessment. Specifically, multicollinearity should be evaluated with simple correlations among the independent variables. Also, multivariate diagnostics (i.e. standard multiple regression) can be used to assess for multivariate outliers and for the exclusion of outliers or influential cases. Sample size guidelines for multinomial logistic regression indicate a minimum of 10 cases per independent variable (Schwab, 2002). Multinomial logistic regression is often considered an attractive analysis because; it does not assume normality, linearity, or homoscedasticity. A more powerful alternative to multinomial logistic regression is discriminant function analysis which requires these assumptions are met. Indeed, multinomial logistic regression is used more frequently than discriminant function analysis because the analysis does not have such assumptions. Multinomial logistic regression have assumptions, such as the assumption of independence among the dependent variable choices. This assumption states that the choice of or membership in one category is not related to the choice or membership of another category (i.e., the dependent variable). The assumption of independence can be tested with the Hausman-McFadden test. Furthermore, multinomial logistic regression also assumes non-perfect separation. If the groups of the outcome variable are perfectly separated by the predictor(s), then unrealistic coefficients will be estimated and effect sizes will be greatly exaggerated. There are different parameter estimation techniques based on the inferential goals of multinomial logistic regression analysis. One might think of these as ways of applying multinomial logistic regression when strata or clusters are apparent in the data. Unconditional logistic regression (Breslow & Day, 1980) refers to the modeling of strata with the use of dummy variables (to express the strata) in a traditional logistic model. Here, one model is applied to all the cases and the stata are included in the model in the form of separate dummy variables, each reflecting the membership of cases to a particular stata. Conditional logistic regression (Breslow & Day, 1980; Vittinghoff, Shiboski, Glidden, & McCulloch, 2005) refers to applying the logistic model to each of the stata individually. The coefficients of the predictors (of the logistic model) are conditionally modeled based on the membership of cases to a particular stata. Marginal logistic modeling (Vittinghoff, Shiboski, Glidden, & McCulloch, 2005) refers to an aggregation of the stata so that the coefficients reflect the population values averaged across the stata. As a rudimentary example, consider averaging each of the conditional logistic coefficients, from the previous paragraph, to arrive at set marginal coefficients for all members of the population – regardless of stata membership. Variable selection or model specification methods for multinomial logistic regression are similar to those used with standard multiple regression; for example, sequential or nested logistic regression analysis. These methods are used when one dependent variable is used as criteria for placement or choice on subsequent dependent variables (i.e., a decision or flow-chart). For example, many studies indicate the decision to use drugs follows a sequential pattern, with alcohol at an initial stage followed by the use of marijuana, cocaine, and other illicit drugs.

2.2.5 Performance Metrics for Predictive Modeling

(Olson & Delen, 2008) explains that the primary source of performance measurements in classification problems is a coincidence matrix. Figure 2.5 shows a coincidence matrix for a two-class classification problem. The equations of most commonly used metrics that can be calculated from the coincidence matrix is also given below.

		True Class	
		Positive	Negative
Predicted Class	Positive	True Positive Count (TP)	False Positive Count (FP)
	Negative	False Negative Count (FN)	True Negative Count (TN)

Figure 2.5 a simple coincidence matrix (Olson & Delen, 2008).

The numbers along the diagonal from upper-left to lower-right represent the correct decisions made, and the numbers outside this diagonal represent the errors. The true positive rate (also called hit rate or recall) of a classifier is estimated by dividing the correctly classified positives (the true positive count) by the total positive count. The false positive rate (also called false alarm rate) of the classifier is estimated by dividing the incorrectly classified negatives (the false negative count) by the total negatives. The overall accuracy of a classifier is estimated by dividing the total correctly classified positives and negatives by the total number of samples. Other performance measures, such as recall (sensitivity), specificity and F-measure are also used for calculating other aggregated performance measures.

$$True\ Positive\ Rate = \frac{TP}{TP + FN}$$

$$True\ Negative\ Rate = \frac{TN}{TN + FP}$$

$$Accurcy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F - measure = \frac{2}{\frac{1}{Presition} + \frac{1}{Recall}}$$

When the classification problem is not binary, the coincidence matrix gets a little more complicated (see Figure 2.6). In this case the terminology of the performance metrics becomes limited to "per class accuracy rates" and the "overall classifier accuracy". These formulas are given below.

		Actual Classification of Classes in the Dataset			
		Class 1	*Class 2*	*Class 3*	
Model Classification	*Class 1*	22	7	2	
	Class 2	5	18	7	
	Class 3	3	5	21	
	Sum	30	30	30	
	Probability	0.33	0.33	0.33	
	Accuracy	0.73	0.60	0.70	0.68

Figure 2.6 A sample coincidence matrix for a three class classifier (Olson & Delen, 2008).

$$(True\ Classification\ Rate)i = \frac{(True\ Classification)i}{\sum_{i=1}^{n}(False\ Classification)i}$$

$$(Overall\ Classificer\ Accuracy)i = \frac{\sum_{i=1}^{n}(True\ Classification)i}{Total\ number\ of\ Cases}$$

Where *i* is the class number, n is the total number of classes.

Estimating the accuracy of a classifier induced by some supervised learning algorithms is important for the following reasons:

1- It can be used to estimate its future prediction accuracy which could imply the level of confidence one should have in the classifier's output in the prediction system.

2- It can be used for choosing a classifier from a given set (selecting the "best" model from two or more qualification models).

3- It can be used to assign confidence levels to multiple classifiers so that the outcome of a combining classifier can be optimized. Combined classifiers are increasingly becoming more popular due to the empirical results that suggest them producing more robust and more accurate predictions as they are compared to the individual predictors.

For estimating the final accuracy of a classifier one would like an estimation method with low bias and low variance. In some application domains, to choose a classifier or to combine classifiers the absolute accuracies may be less important and one might be willing to trade off bias for low variance.

Chapter 3

Related Works

Chapter 3
Related Works

In this chapter, different related works are studied and investigated. The chapter is divided into three sections. In section 3.1, we will give some related works about using data mining techniques in human performance prediction *in general*. In section 3.2, we will present some related works about *lecturer's* performance using data mining techniques. Finally, in section 3.3 we will give some conclusions about this chapter.

3.1 Performance prediction using DM techniques

In this section, we present some related works that use data mining techniques in predicting performance of people in general. We try to focus on performance in the field of education to be more close to our thesis problem.

(Hajizadeh & Ahmadzadeh, 2014) have used two applicable data mining techniques that known as “Classification” and “Association rule discovery” to explore effective factors on non re taking a course by students. They use evaluation tools as a step toward identifying and extracting factors affecting educational failure in students. They use data set contains a total 5820 evaluation scores and 33 attributes that provided by students from Gazi University in Ankara (Turkey). The attributes are nb.repeat, attendance, difficulty and 28 attributes that taken from the questionnaire: 12 attributes are related to course that taken by students and 16 attributes are related to features of the course instructor. According to their purpose, it was necessary nb.repeat field had to be chosen as the “class” attribute and the numerical values had to be replaced. By using data mining tool WEKA (Waikato Environment for Knowledge Analysis), three types of evaluation have been performed on the data. The first analysis the goal is to investigate the role of providing a specific course by an instructor on not taking that course again by students. The second analysis the goal is to investigate the features of provided course on not taking that course again by students. To evaluate this analysis attributes attendance, difficulty and Q1 to Q12 have been selected. Among classification algorithms that is applied to the selected data, the best result provided by REPTree algorithm with the following features and decision tree: Correctly Classified Instances 84.24% Avg. F-Measure 0.774. The third analysis The goal is to investigate the features of an instructor on not taking more than once the course provided by that instructor. To evaluate this analysis attributes Q1 to Q12 have been selected. Among classification algorithms that is applied to the selected data, the best result provided by REPTree algorithm with the following features and decision tree: Correctly Classified Instances 84.26% Avg. F-Measure 0.772.

The work is related to our thesis because the problem is in the field of educational data mining, and it is a classification problem. Some differentiations between this study and our thesis, 1- our thesis goal is predicting the performance of new instructor not to determine exist instructor to be tech the course again or not. 2- this study was focused on student side by predicting student performance in a course if he will fail or pass, but our thesis focused on the lecturer side.

(Gupta & Suma, 2014) predicted the performance of software project member on the basis of previous database or training set. They have focused on the human aspect of software engineering to achieve good quality of software by building a classification model for predicting employees' performance based on certain attributes. Data mining techniques could identify those attributes required in a project member which will contribute to good performance and thereby enhance software quality and success. Classification techniques like ID3, CART and C4.5 showed similar results. Performance was earlier assumed to be best for candidates having a good college aggregate. The study showed that other talent attributes like programming skills and reasoning skills have proved to be more important, although software companies emphasize upon aggregate percentile and followed the same trend for many years. Due to lack of analytical method in human aspects, software companies were not selecting the right people who could perform well in the software process and thereby failed to achieve the desired quality in the time and cost constraints. They use data mining techniques that give a very interesting patterns and helped in earlier identification of project members who will perform well. The study enables the managers to refocus on human capability criteria and thereby enhance the development process of software project. Just like all process within generic framework of software development is given importance, human aspect also needs a deeper investigation for effective software development. Without the right people even the best process analysis and development are bound to fail. Further scope of this study is to investigate on several projects of different domains and take into account more attributes of project personnel and correlate it with software quality and success.

This paper describes the importance of using data mining in analysis and predicting performance of human resources in organizations, they help software companies to select the right people that will perform well in the software process. This study have the same goals of our thesis with different field, our thesis goal is to help universities to select the best lecturer to teach courses by predicting his performance when he requests to teach based on his characteristics using data mining techniques.

(Bhardwaj & Pal, 2011) gathered data of 300 students from different degree colleges and institutions in Awadh University, Faizabad, India. These data are analyzed using classification method to predict the student's performance. The researchers used the attributes: Students Sex, Students category, Medium of Teaching, Students food habit, Students other habit, Living Location, Student live in hostel or not, student's family size, Students family status, Family annual income status, Students grade in Senior, Secondary education, Students College Type, Fathers qualification, Mother's Qualification, Father's Occupation, Mother's Occupation, Grade obtained in BCA. They use Bayesian classification method on student database to predict the students division on the basis of previous year database. The study help the students and the teachers to improve the division of the student. and work to identify those students which needed special attention to reduce failing ration and taking appropriate action at right time. The study shows that academic performances of the students are not always depending on their own effort. It shows that other factors have got significant influence over students' performance. The variables whose probability values were greater than 0.50 were given due considerations and the highly influencing variables with high probability values have been shown in Table 3.1. These features were used for prediction model construction. For both variable selection and prediction model construction, they have used MatLab.

Table 3.1 High Potential Variables (Bhardwaj & Pal, 2012)

Variable	Description	Probability
GSS	Students grade in Senior Secondary education	0.8642
LLoc	Living Location	0.7862
Med	Medium of Teaching	0.7225
MQual	Mother's Qualification	0.6788
SOH	Students other habit	0.6653
FAIn	Family annual income status	0.5672
FStat	Students family status	0.5225

The researchers focused on student side in educational data mining. Our opinion of this work that he uses a good series of procedures to predict the student performance, that we will use in our thesis, but with differences because we will focus on lecturer performance side, and the amount of data that we will use. Here the researchers used a sample consist of 300 students only, it makes the results less reliable. We will use about 5000 records.

In section 3.1 we try to found some related work that investigated the performance predicting in general, the term performance predicting is a wide term that use in most things: variant organizations, factories, human, machines, etc. we can benefited from the methodologies that used to predict performance, so we try to explore works that closed to our work field, Educational data mining.

3.2 Lecturer's performance using DM techniques

This section presents some works that studied the evaluation of performance of university lecturers using DM techniques.

(Pal & Pal, 2013) proposed to evaluate teacher's performance on the basis of different factors, using data mining techniques. They collected data from post graduate studies at department in the college of engineering over a three-year period of the same students. Their proposed model consider the various aspects of performance measures of teachers, like Students' Feedback (voice modulation, speed of delivery, content arrangement, presentation, communication, overall impression, content delivery, explanation power, overall teaching and regularity, Results, Students attendance, that have deep influence on the teachers' performance in university. All the predictor and response variables, which were derived from the database, are given in Table 3.2.

Table 3.2 Student related variables (Pal & Pal, 2013)

Variables	Description	Possible Values
Name	Teacher Name	Text
SD	Speed of delivery	(1,2,3,4,5)
CA	Content arrangement	(1,2,3,4,5)
PR	Presentation	(1,2,3,4,5)
CO	Communication	(1,2,3,4,5)
OI	Knowledge	(1,2,3,4,5)
CD	Content delivery	(1,2,3,4,5)
EP	Explanation power	(1,2,3,4,5)
DC	Doubts Clearing	(1,2,3,4,5)
DP	Discussion of Problems	(1,2,3,4,5)
OCR	Overall completion of course and regularity	(1,2,3,4,5)
SA	Students attendenc	1- Below, 2 – average, 3- high
RE	Result	1- Pass, 2- fail, 3- promoted
PT	Performance of Teacher	(1,2,3,4,5)

They use Weka, which is an open source software that implements a large collection of machine leaning algorithms and is widely used in data mining applications. Tests were conducted using three tests for the assessment of input variables: Chi-square test, Info Gain test and Gain Ratio test. Different algorithms were provide different results, i.e. each of them accounts the relevance of variables in a different way, therefore the average value of all the algorithms is taken in the final result of variables ranking, instead of selecting one algorithm and evaluating it.

Then they have carried out some experiments in order to evaluate the performance and usefulness of different classification algorithms for predicting students' placement. The algorithm used for classification is Naive Bayes, ID3, CART and LAD. The results of the experiments are shown in Table 3.3

Table 3.3 Performance of the classifiers in (Pal & Pal, 2013)

Evaluation Criteria	Classifiers			
	LAD	CART	ID3	NB
Timing to build model (in Sec)	0.08	0.11	0	0
Correctly classified instances	84	81	73	90
Incorrectly classified instances	28	31	34	22
Accuracy (%)	75.00%	72.32%	65.17%	80.35%

It is clear that the highest accuracy is 80.35% and the lowest is 65.17%. In fact, the highest accuracy belongs to the Naïve Bayes Classifier followed by LAD tree with a percentage of 75.00% and subsequently CART. After tested many attributes, some of them are found effective on the performance prediction. The content arrangement was the strongest attribute, and then the result plays an important role in the performance of teachers. The speed of delivery attribute did not show any clear effect while the overall completion of course and regularity attribute has shown some effect in some of the experiments for predicting the performance. Other attributes had a degree of effect on predicting the performance.

It is clear that the attributes that use by the research belong to students like student attendance and student result, in our thesis, all the attributed belong to lecturer expect the class label (evaluation) that is related to lecturers but it is determined by students.

(Mardikyan & Badure, 2011) conduct a research that aimed at identifying the factors associated with the teaching performance of instructors of the Management Information Systems (MIS) Department of Bogazici University in Turkey. they include in their study variables related to instructor and course characteristics. Student course evaluations are conducted during the last two weeks of each semester by the registration office in Bogazici University. The data set used in this study includes all reported MIS graduate and undergraduate course sections offered from Fall 2004 to Summer Term 2009 except project, seminar, and graduate foundation courses. After this exclusion, the final data set contains totally 259 course sections offered by both full time and part-time instructors of MIS department.

The data include two basic categories of variables. The first group consists of the data obtained from evaluation questionnaires where students response anonymously 13 questions about course (Q1–Q6) and instructor (Q7–Q13) characteristics. In addition to some of the data used concerns us tow attributes: whether or not the instructor is a part time and giving the course for the first time. The average of Q12 and Q13 (named as AVERQ1213) is used as the dependent variable, since among all the variables; these are the only two variables that measure the overall teaching performance of the instructor. To reduce the number of independent variables and to understand the patterns of relationship among them the factor analysis was applied. In this study, two well-known data mining techniques; stepwise regression and decision tree techniques are applied to the data. The results show that, a factor summarizing the instructor related questions in the evaluation form, the employment status of the instructor, the work load of the course, the attendance of the students, and the percentage of the students filling the form are significant dimensions of instructor's teaching performance. The evaluation form in use distinguishes two groups of questions: related to the course and related to the instructor. According to these results, the most important factor to explain the instructors' teaching performance is the instructor attitudes that are primarily measured by the evaluation process. This result is not peculiar to the MIS education since the same evaluation form is used in all departments of Bogazici University. The courses offered by part time instructors tend to receive higher ratings in MIS education. According to the results of stepwise regression and decision tree analyses the instructors' teaching performance is explained with mostly instructor related questions summarized by the variable COMP1. Instructors, who have well prepared course outlines, use satisfactory materials, help the students outside the lectures, grade exams fairly and on time receive higher evaluations. An additional instructor characteristics; the employment status of the instructor that is not included in the questionnaire is found to be significant. Giving the course for the first time is not found to be a significant factor in explaining the teaching performance.

The study present prediction of lecturer performance based on course characteristics and lecturer characteristics from students view, it use only two attributes about personal characteristics: whether or not the instructor is a part time and giving the course for the first time. The first attribute will be one of our data set; we will compare its significant in this study to its significant in our thesis. The second attribute is not available in our data set because it is not in IUG database. In our thesis all the attributes we use is related only to lecturer characteristics.

(Wang et al., 2009) sought to determine which of a number of independent variables (demographic and rating response) would predict student response to an overall rating item for their instructor. The University of Central Florida (UCF) administers an end-of-course student evaluation instrument. The Student Perception of Instruction (SPI) form is a 16-item, Likert-type device that students use to rate their instructors (e.g., excellent, very good, good, fair, poor). Respondents have the opportunity to provide written comments about the instructor, and considerable demographic information (course level, college, department, and instructor) can be obtained from the instrument because the class and date are recorded on the form. After classes end, instructors receive the original forms with student comments and a summary of course-rating responses. Presently, many students have an online response option as well.

The investigators assembled a dataset containing all student ratings of instructors for the 5 academic years beginning 1996/97 through 2000/01. The file contained 588,575 student records with responses to the 16 items (Table 3.4) and corresponding demographic information. The investigators reformatted the file so that it comprised only the responses to the 16 items (five levels) and indicators of course level (lower undergraduate, upper undergraduate, graduate), college (Arts and Sciences, Business Administration, Education, Engineering and Computer Science, Health and Public Affairs), and the academic year. No further identifying information was available in the analysis file. Throughout the study, the investigators preserved department and instructor anonymity. Therefore, this study investigated the independent measures, college, course levels, academic year, and items 1 through 15 on the Student Perception of Instruction instrument for their ability to predict overall rating of the instructor (item 16).

Table 3.4 Student perception of instruction items for the University of Central Florida (Wang et al., 2009)

Source	Questions
Administration	1. Feedback concerning your performance in this course was:
	2. The instructor's interest in your learning was:
	3. Use of class time was:
	4. The instructor's overall organization of the course was:
	5. Continuity from one class meeting to the next was:
	6. The pace of the course was:
	7. The instructor's assessment of your progress in the course was:
	8. The texts and supplemental learning materials used in the course were:
Board of regents	9. Description of course objectives and assignments:
	10. Communication of ideas and information:
	11. Expression of expectations for performance:
	12. Availability to assist students in or outside of class:
	13. Respect and concern for students:
	14. Stimulation of interest in the course:
	15. Facilitation of learning:
	16. Overall assessment of instructor:

In order to explore these data, the authors incorporated decision trees that developed three rules that predicted a high probability that an instructor would receive an overall rating of excellent (Table 3.5) while three other rules led to a poor rating (Table 3.6)

Table 3.5 Decision rules that lead to an overall instructor rating of "excellent"

Question	Rating					Excellent (p)
	E	VG	G	F	P	
Rule 1 (n = 46,805)						
- Facilitation of learning	●					0.96
- Communication of ideas and information	●					
Rule 2 (n = 3,462)						
- Facilitation of learning	●					
- Communication of ideas and information		●				0.85
- Organization of the course	●					
- Assessment of student progress	●	●				
Rule 3 (n = 6,215)						
- Facilitation of learning				●		0.78
- Communication of ideas and learning	●	●				
- Organization of the course	●					
-Instructor interest in your learning	●					

Table 3.6 Decision rules that lead to an overall instructor rating of “poor”

Question	Rating					Excellent (p)
	E	VG	G	F	P	
Rule 4 (n = 1,821)						
- Facilitation of learning				●	●	
- Communication of ideas and information					●	0.83
- Instructor interest in your learning					●	
Rule 5 (n = 1,135)						
- Facilitation of learning				●	●	
- Communication of ideas and information					●	0.58
- Organization of the course					●	
Rule 6 (n = 532)						
- Facilitation of learning				●	●	
- Communication of ideas and learning			●	●		0.54
- Assessment of student progress					●	
- Assessment of student progress					●	

Table 3.7 presents the results of the analyses for the items contributing to an overall excellent rating of the instructor. All items selected by the decision tree contributed to the equation with the Wald chi-square probabilities rounded to 0.00. The model predicted with 97.6% accuracy producing a Somer’s D of 0.945.

Table 3.7 Logistic regression for “excellent” rule items (Wang et al., 2009)

Items	df	Coefficient	Wald x^2	p
Intercept	1	0.895	23,385.10	0.0001
Interest	1	0.796	23,157.20	0.0001
Organization	1	0.798	23,903.10	0.0001
Assessment	1	0.847	12,454.40	0.0001
Communication	1	0.847	25,162.30	0.0001
Facilitation	1	1.092	33,328.60	0.0001
A Percent correctly predicted = 97.6, Somer’s D = .963.				

Table 3.8 Logistic regression for "poor" rule items (Wang et al., 2009)

Items	df	Coefficient	Wald x^2	p
Intercept	1	0.171	349.9	0.0001
Interest	1	-0.691	6,472.9	0.0001
Organization	1	-0.919	12,650.7	0.0001
Assessment	1	-0.667	7,062.0	0.0001
Communication	1	0.924	13,598.8	0.0001
Facilitation	1	-0.961	14,867.4	0.0001
A Percent correctly predicted = 97.6, Somer's D = .963.				

This research predicting the overall evaluation based on the internal answers of end semester evaluation questionnaire. Our opinion that this is can called analysis not prediction, because it depend on the answers of evolution questions, so the overall prediction is exist by simple calculation for the answers, and there is no need for prediction. Our theses is depend on attributes of lecturer characteristics, the lecturer may come to teach for the first time, so all data needed is available when the lecturer request to teach.

(Ahmadi & Abadi, 2013) are collected information and results of a survey about 104 teachers in Sanandaj Daughter Vocational Faculty on teacher's behaviors in classroom then with data mining algorithms such Association Rule and decision trees (j48), it is proceeded to analyze and predict acceptation of a teacher for continuing the teaching in faculty. There are new rules and relations between selected parameters such as Evaluation's score, Teacher's degree, Degree's Type, Teaching experience, Acceptation to next semesters on teacher's evaluation that is interested for education managers. They use the acceptation attributes as class label that take tow values Yes or No. It is done a web base survey from 830 students then it is prepared results of this survey for 104 teachers.

They indicated that evaluation's score of students is very important factor that many universities gather this information on performance of teachers. By using data mining and J48 tree as a decision tree new rules are results that education managers could use these rules in future decisions to submit new teachers and continue with elected old teachers. They generate rules for example in Figure 3.1.

1- IF(Evaluation_score=GOOD)THEN (Acceptation is Yes, means next semester will continue his/her teaching)
2- IF(Evaluation_score=Excellent)AND (Teaching_experience=FALSE means is low)THEN(Acceptation is Yes means next semester will continue his/her teaching)

Figure 3.1 Discovered rules (Ahmadi & Abadi, 2013)

The study limited on four variables related to lecturer, without determining the accuracy of this model, and not exploited the historical data for mining. However, in our study we use many attributes related to lecturer characteristics, with high accuracy model relatively.

(Sok-Foon et al., 2012) try to identify the factors and predictors of lecturer performance among undergraduates in a private university in Malaysia using the existing questionnaire. They use a total of 223 respondents were recruited using multistage sampling. In the first stage, a faculty was randomly chosen. In the second stage, the School of Management was randomly chosen out of two schools. In the third stage, two subjects were randomly chosen from subjects in year 1 of the undergraduate program. A similar selection was undertaken for two subjects from years 2 and 3. A total of six subjects from years 1 to 3 were chosen. A total of 32 items was used to gather the responses to measure: lecturer and tutor characteristics (13 items), subject characteristics (6 items), the studentship (7 items) and learning resources and facilities (4 items). They also added two questions to the existing questionnaire on overall performance (2 items). The authors use Statistical Package for Social Science for Windows (SPSS for Windows Version 13.0) to analyze the collected data. They use Multiple linear regressions and the stepwise method to identify the significant determinants of overall lecturer performance.

The results of the final model showed that lecturer and tutor characteristics, subject characteristics, and learning resources and facilities explained 61.9% of the variance in overall lecturer performance among students. They point out that even though lecturer or tutor characteristics are the main predictors of overall performance and lead to student satisfaction. The results of this study showed that: lecturer and tutor characteristics ($r = 0.722$, $p < 0.01$), subject characteristics ($r = 0.699$, $p < 0.01$), the studentship ($r = 0.472$, $p < 0.01$) and learning resources and facilities ($r = 0.650$, $p < 0.01$) were positively correlated with overall lecturer performance. The side that related to our study that lecturer and tutor characteristics were not statistically different in terms of gender.

This study depends on statistical methods not data mining methods, it deal with samples not with historical data, it can benefited for us that the study results show that the lecturers characteristics are the main predictors of overall performance.

(Agaoglu, 2016) was collected data from one of the randomly selected departments of Marmara University, Istanbul, Turkey. A total of 2850 evaluation scores are obtained. He use 70% of the data for training the classifier models and the remaining 30% for testing. Student evaluation data has 26 variables all except one, which is class label, are responses, measured on an interval scale, to questions in course and instructor performance evaluation forms. Response values of these questions are of the form (1,2,3,4,5) where 1,2,3,4,5 represents the answers ‘‘Never’’, ‘‘Rarely’’, ‘‘Sometimes’’, ‘‘Often’’, ‘‘Always’’ respectively for Q1 to Q4; and 1,2,3,4,5 represents ‘‘Strongly disagree’’, ‘‘Disagree’’, ‘‘Neutral’’, ‘‘Agree’’, and ‘‘Strongly agree’’ respectively for Q5 to Q25. The last variable is dichotomous variable measured on a nominal scale in the form of (1,2) where 1 stands for ‘‘Not satisfactory’’ and 2 for ‘‘Satisfactory’’.

The researcher use seven classification models: two using decision tree algorithms (C5.0, and CART), one using SVM, three using ANNs, and one using DA. The performances of these models are evaluated on the test data in terms of accuracy, precision, recall, and specificity.

He compare all the applied classifiers using evaluation measures. As seen in Figure 3.2, classifiers give similar results on the test dataset.

Model	Accuracy	Precision	Recall	Specificity
C5.0	**92.3 %**	**94.4 %**	92.1 %	**92.5 %**
CART	89.9 %	89.9 %	93.1 %	85.5 %
SVM	91.3%	92.2%	92.9%	89.1%
ANN-Q2H	91.2 %	94.1 %	90.5 %	92.2 %
ANN-Q3H	90.8 %	89.7 %	**95.0 %**	85.0 %
ANN-M	90.5 %	92.1 %	91.5 %	89.1 %
DA	90.5 %	94.1 %	90.6 %	90.4 %

Figure 3.2 Performances of classifiers (Agaoglu, 2016)

Accuracy values, which assess the effectiveness of the models, are all at least approximately 90%. C5.0 classifier is the best in performance according to accuracy followed by SVM, and CART is the worst

This research is similar to (Wang et al., 2009), both of them try to identify the factors that affect the lecturer performance by studying the relation between student satesfaction and the internal points of questionnaire that applied on student at the last two weeks of semester. In our research we deal with the average of the questionair and try to predect the student ratings of lecturer based on lecturers characterstics.

In this section, we reviewed some previous studies that related to our thesis, these studies about prediction of lecturer performance using data mining techniques. Some of these studies shared us in proposed logical methodology like (Hajizadeh & Ahmadzadeh, 2014) (Gupta & Suma, 2014) and (Bhardwaj & Pal, 2011). Most of these studies focused on course characteristics and student characteristics to predict lecturer performance, a few studies involved lecturers characteristics poorly like (Pal & Pal, 2013), (Mardikyan & Badure, 2011) and (Ahmadi & Abadi, 2013).

3.3 Conclusion

In chapter 3 we talk about studies that related to our thesis, we divided these studies into two categories. Section 3.1 contain studies that investigated in the field of performance prediction using data mining techniques, we try to be closed to educational data mining field. In section 3.2, we present studies about prediction of lecturer performance using data mining techniques.

We notice that most studies built its prediction on course characteristics, student characteristics, or lecturer's characteristics poorly. The studies that involved lecturer's characteristics are involved as secondary predictors. In our thesis, we will predict the lecturer's performance depending mainly on lecturers characteristics. We get dataset contain 30 attribute about lecturers characteristics in Islamic University of Gaza. We will describe these attributes in section 4.1.1.

Chapter 4

Proposed Method and Implementation

Chapter 4

Proposed method and implementation

In this chapter, we are going to explain our proposed method of predicting lecturer's performance based on lecturer's characteristics using data mining techniques. This chapter is divided into three sections: Section one describes our methodology to achieve the research problem. Section two explains the steps that is followed to apply the selected techniques, and the implementation environment that has been done. Section three summarizes the proposed method and the implementation steps.

4.1 Methodology

To achieve the research goals, we are going to follow some steps, we are going to describe these steps briefly and we will explain it in details in next subsections.

- Data collection: collecting data sets from various IUG departments.
- Data integration: integrate the data sets to be one data set.
- Data preprocessing: dealing with missing values and noisy data and increasing data quality to give good results.
- Apply data mining techniques: we have chosen four data mining techniques:
    - Decision tree
    - KNN
    - Multinomial Logistic Regression
    - Naïve Bayesian

  We will apply these techniques on our data set.
- Compare results: by comparing the accuracy of each model.
- Choose the proved model: that gives a higher accuracy and depends on it.

The next Figure 4.1 explains the steps of our methodology.

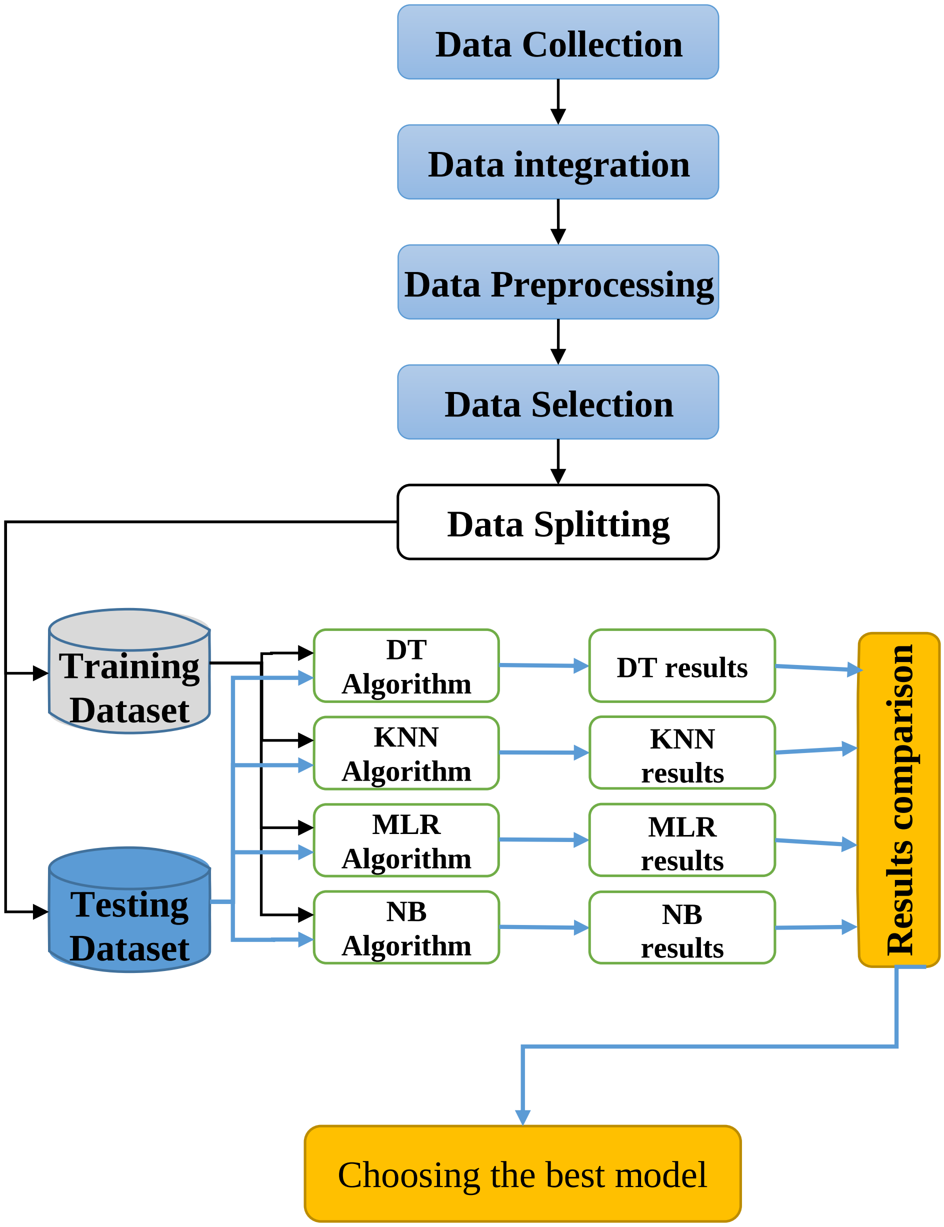

Figure 4.1 Thesis methodology

4.1.1 Data collection

The used data is belonging to the academic staff of Islamic university – Gaza (IUG) that was taught in last 11 semesters (from second semester 2011/2012 to summer semester 2014/2015), we chose these semesters because the evaluation model was stable in this period. Previously IUG used many models of evaluation and there were many editions, so the specialist in IT department cannot extract the overall evaluation for the lecturer per course in this period.

IUG applies the evaluation model at the last two weeks of the semester, the model consists of 20 Likert-type questions. Students use this model to rate their lecturers by choice (كبيرة جدا، كبيرة، متوسطة، قليلة، قليلة جدا). Students can also provide written comments about their lecturers. The result of student's answers, calculated to be one percentage value per course. This value considers the overall lecturer's evaluation of the students in all classes in one course at the semester.

The used questionnaire in IUG is attached in Appendix A.

In this research, we get the evaluation (one percentage value) of all lecturers that taught in IUG in the last 11 semesters, this data is available in Academic affairs at the university and it is contains 4 attributes and 5170 records. It is for 664 courses that is taught by 551 lecturer. Every lecturer tought one or more courses and the lecturer has one evaluation value per course, without any consideration to the count of the students that ratings and their sex. And we may found repeation to the lecturer and course with different semesters.Table 4.1 describes the attributes and possible values.

Table 4.1 Academic Affairs data

attribute name	attribute description	values
EMP_NO	Employee ID	551 IDs of lecturers
SEMESTER_NO	Semester number	11 semesters
SUBJECT_NO	Course Code	664 course
EVALUATION	Overall evaluation of the lecturer in this course on this semester	numeric value (29.2 - 100)

After we had got these data, we contact with department of employee affairs at IUG to provide us the data that is available at the department about the 551 academic stuff, they provide us the data which contains 30 attribute and 544 records, the attributes are explained in Table 4.2.

Table 4.2 Employee affairs data

no.	attribute name	attribute description	values
1	EMP_NO	Employee ID	serial number
2	JOB_NAME	Job Title	أستاذ مساعد، أستاذ، أستاذ مشارك، مدرس، معيد، محاضر
3	DEPARTMENT_A_NAME	the college of lecturer	12 colleges in IUG
4	Section_A_Name	the department of lecturer	41 academic department in IUG
5	Gender	Lecturer Gender	ذكر، أنثى
6	Birth_Date	date of birth	numeric value (1945 - 1992)
7	MARITAL_STATUS	material status	متزوج، أعزب، مطلق
8	Governorate_Name	Place of residence	شمال غزة، غزة، الوسطى، خانيونس، رفح
9	Education_Name	Qualification	دكتوراه، ماجستير، بكالوريوس، دبلوم عالي
10	Contract Name	contract Type	دائم، إشراف عملي، خاص، العاملون بالساعة، وكالة بالساعة، راتب مقطوع، تجريبي
11	Status_Name	managerial status	منتظم، متجدد، إنتهاء عقد، إجازة علمية، ابتعاث، متقاعد، استقالة، إجازة بدون مرتب، متوفي على رأس عمله، انهاء خدمة
12	Professional_Experience	number of experience years	numeric value (0 - 22)
13	Specialist_A_Name_Bc	specialist in BSc.	38 defferint specialist
14	Specialist_Detail_A_Name_Bc	accurate specialist in Bachelor	177 defferint accurate specialist
15	University_Name_Bc	university that graduate in Bachelor	81 different university
16	Country_Bc	Country that graduate in Bachelor	29 different country
17	End_Date_Bc	date of graduate in Bachelor	numeric value (1966 - 2014)
18	Emp_Level_Bc	grade of Bachelor	مقبول، جيد جدا، ممتاز، جيد
19	Specialist_A_Name_M	specialist in master	47 different specialist
20	Specialist_Detail_A_Name_M	accurate specialist in master	224 defferint accurate specialist
21	University_Name_M	university that graduate in master	122 different university
22	Country_M	Country that graduate in master	30 different country
23	End_Date_M	date of graduate in master	numeric value (1977 - 2014)
24	Emp_Level_M	grade of master	جيد جدا، ممتاز، جيد
25	Specialist_A_Name_Dr	specialist in Ph.D.	45 defferint specialist
26	Specialist_Detail_A_Name_Dr	accurate specialist in Ph.D.	226 defferint accurate specialist
27	University_Name_Dr	university that graduate in Ph.D.	138 different university
28	Country_Dr	Country that graduate in Ph.D.	35 different country
29	End_Date_Dr	date of graduate in Ph.D.	numeric value (1983 - 2014)
30	Emp_Level_Dr	grade of Ph.D.	جيد جدا، ممتاز، جيد

4.1.2 Data preprocessing

After data collection, we proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data and to increase the accuracy of the mining. We use Microsoft excel 2013 to perform the following data preprocessing.

4.1.2.1 Data integration

We merge the two datasets through the common attribute EMP_NO to result 34 attributes and 5122 records.

4.1.2.2 Missing values:

- Delete the attribute Professional_Experience because the missing value is 91%.
- Delete the records (7 records) of the lecturer who is emp_no is 99568 because most of the values are missing.
- Fill some missing value manually based on other attributes.
    - End_Date_Bc based on Birth_Date by adding 22 years.
    - End_Date_M based on End_Date_Bc and End_Date_Dr by average.
    - Emp_Level_Bc based on the mean for all samples belonging to the same classe.
    - Specialist_A_Name_Bc, Specialist_A_Name_M and Specialist_A_Name_Dr based on each other.
- Fill the 12 columns that related to master degree and PhD degree to who is education level is "بكالوريوس" with the value "has_B_Only".
- Fill the 6 columns that related to PhD degree to who is education level is "ماجستير" with the value "has_M_Only"
- Fill the remains of missing values in the columns that are related to qualifications data with the values "un_Known_columnName", for example the missing values in the column Country_Dr we fill it with the value "un_Known_Country_Dr"

4.1.2.3 Inconsistent data:

We edit some values in some attributes manually because there is inconsistency between values, for example the attribute Specialist_Detail_A_Name_Dr, it contains three values with printing errors for the same Specialist ("تفسير وعلوم القرآن "," التفسير وعلوم القرآن"," تفسيروعلوم قرآن "), we consolidate these values in one value " التفسير وعلوم القرآن "

In attribute Specialist_Detail_A_Name_Dr we edit 44 values as illustrated in Table 4.3.

Table 4.3 Inconsistent values in Specialist_Detail_A_Name_Dr attribute

The original values	Modified values
الاختصاص الجنائي العالميس	الاختصاص الجنائي العالمي
إحصاء	الإحصاء
الأسلوبية (المسرحية الإنجليزية	الأسلوبية (المسرحية الإنجليزية)
البورد الفلسطيني+البورد الأردن	البورد الفلسطيني+البورد الأردني
تاريخ	التاريخ
تاريخ إسلامي	التاريخ الإسلامي
تحاليل طبية	التحاليل الطبية
تربية انجليزي	التربية/اللغة الإنجليزية
التعليم الإلكتروني باستخدام ال	التعليم الإلكتروني
تفسير	التفسير
تفسير وعلوم القرآن	التفسير وعلوم القرآن
تفسيروعلوم قرآن	التفسير وعلوم القرآن
الحديث الشريف وعلومه	الحديث وعلومه
حقوق	الحقوق
خدمةإجتماعية	الخدمة الإجتماعية
صحة نفسية	الصحة النفسية
عقيدة	العقيدة
العقيدة	العقيدة الإسلامية
فقة مقارن	الفقه المقارن
فقه مقارن	الفقه المقارن
فقه وأصول فقه	الفقه وأصوله
فقه وأصوله	الفقه وأصوله
فلسفة-مناهج وطرق تدريس	الفلسفة في التربية(مناهج وطرق.
فيزياء	الفيزياء
كيمياء	الكيمياء
لغة عربية	اللغة العربية
اللغة العربية- النحو والصرف	اللغة العربية - نحو وصرف
المحاسبة (كلية الاقتصاد)	المحاسبة
محاسبة	المحاسبة
مناهج وطرق التدريس	المناهج وطرق التدريس

The original values	Modified values
مناهج وطرق تدريس	المناهج وطرق التدريس
نحو وصرف	النحو والصرف
الهندسة الكهربائية والإلكتروني	الهندسة الكهربائية والإلكترونية
هندسة كهربائيةوالكترونية	الهندسة الكهربائية والإلكترونية
هندسة مدنية	الهندسة المدنية
هندسة ميكانيكية	الهندسة الميكانيكية
إدارة أعمال	إدارة أعمال
أحياد دقيقة	أحياء دقيقة
دكتوراة الفلسفة في اللغة العرب	دكتوراة الفلسفة في اللغة العربية
دكتوراه الفلسفة في اللغة العرب	دكتوراة الفلسفة في اللغة العربية
علوم حاسوب	علم الحاسوب
علم نفس	علم النفس
علوم التمريض/ تمريض باطني جراح	علوم التمريض
كهرمغناطيسية	كهرومغناطيسية

This procedure is applied on most attributes, Table 4.4 illustrate these attributes, numbers of values that is edited and numbers of records that is affected by the editing.

Table 4.4 Values and records that are influenced by inconsistent data editing

Attribute name	Altered values	Altered records
Specialist_Detail_A_Name_Bc	49	1946
Specialist_Detail_A_Name_M	37	1441
Specialist_Detail_A_Name_Dr	44	907
University_Name_Bc	2	188
University_Name_M	1	6

4.1.2.4 Data reduction:

a. Delete redundant and irrelevant attributes:

- Delete the attributes that is called emp_no that duplicated after merging the two datasets because they are irrelevant attributes.
- Delete the attributes Course_NO and Semester_NO that is not related to lecturer characterstics.

b. Binning by frequency for the attributes Evaluation, we divided the values into 4 categories based on Quartiles as seen in Table 4.5.

Table 4.5 Evaluation column binning

Class	from	to	Number of records
1	>=29.2	<=75.5	1297
2	>=75.6	<=80.0	1261
3	>=80.1	<=85.6	1282
4	>=85.7	<=100	1275

c. Generalization:

We found that the attribute Education_Name contain only one value "دبلوم عالي", for one employee, we generalize it to be "بكالوريوس".

The final data consists of 29 attributes and 5115 records, Table 4.6 explains these attributes and the final values for each attribute.

Table 4.6 Final data after preprocessing

No.	attribute name	attribute description	values
1	JOB_NAME	Job Title	أستاذ مساعد، أستاذ، أستاذ مشارك، مدرس، معيد، محاضر
2	DEPARTMENT_A_NAME	the college of lecturer	12 colleges in IUG
3	Section_A_Name	the department of lecturer	41 academic department in IUG
4	Gender	Lecturer Gender	ذكر، أنثى
5	Birth_Date	date of birth	47 values (from 1945 to 1992)
6	MARITAL_STATUS	material status	متزوج، أعزب، مطلق
7	Governorate_Name	Place of residence	شمال غزة، غزة، الوسطى، خانيونس، رفح
8	Education_Name	Qualification	دكتوراه، ماجستير، بكالوريوس
9	Contract Name	contract Type	دائم، إشراف عملي، خاص، العاملون بالساعة، وكالة بالساعة، راتب مقطوع، تجريبي
10	Status_Name	managerial status	منتظم، متجدد، إنتهاء عقد، إجازة علمية، ابتعاث، متقاعد، استقالة، إجازة بدون مرتب، متوفي على رأس عمله، انهاء خدمة
11	Specialist_A_Name_Bc	specialist in Bc.	38 different specialist
12	Specialist_Detail_A_Name_Bc	accurate specialist in Bachelor	177 different accurate specialist
13	University_Name_Bc	university that graduate in Bachelor	81 different university
14	Country_Bc	Country that graduate in Bachelor	29 different country
15	End_Date_Bc	date of graduate in Bachelor	45 values (from 1966 to 2014)
16	Emp_Level_Bc	grade of Bachelor	مقبول، جيد جدا، ممتاز، جيد
17	Specialist_A_Name_M	specialist in master	47 different specialist
18	Specialist_Detail_A_Name_M	accurate specialist in master	224 defferint accurate specialist
19	University_Name_M	university that graduate in master	122 different university
20	Country_M	Country that graduate in master	30 different country
21	End_Date_M	date of graduate in master	37 values (from 1977 to 2014)
22	Emp_Level_M	grade of master	جيد جدا، ممتاز، جيد
23	Specialist_A_Name_Dr	specialist in Ph.D.	45 different specialist
24	Specialist_Detail_A_Name_Dr	accurate specialist in Ph.D.	226 different accurate specialist
25	University_Name_Dr	university that graduate in Ph.D.	138 different university
26	Country_Dr	Country that graduate in Ph.D.	35 different country
27	End_Date_Dr	date of graduate in Ph.D.	33 values (from 1983 to 2015)
28	Emp_Level_Dr	grade of Ph.D.	جيد جدا، ممتاز، جيد
29	EVALUATION	The overall evaluation	[29.2 – 75.5] [75.6 – 80.0] [80.1 – 85.6] [85.7 – 100]

4.1.3 Selected models

We used in our research four models: 1- Decision tree, 2- K nearest neighbor, 3- Multinomial Logistic Regression and 4- Naïve Bayesian. The models were described in details in section 2.2.4.

As we found on a lot of researchs and books that stated on chapter two like (Giudici, 2003), (Han & Kamber, 2011), (Witten & Frank), that the predicion models we used consider the most common predictive models in prediction problems, so we chose these models for their common and no specific advantage of anyone of them on the others with respect to our applicati

4.2 Implementation

4.2.1 Implementation environment

By implementation environment, we mean that the hardware and software tools used to run the training and testing experiments. We performed our prediction and data preprocessing on PC machine with the following specification:

1- Operating System: Windows 8.1 Pro 64-Bit
2- Processor: Intel® Core™ i53230M CPU @2.60GHz (4 CPUs)
3- RAM: 6144 MB RAM

We used two software tools:

1- Microsoft Excel:

A spreadsheet was developed by Microsoft for Windows, Mac OS X, and iOS. And its features calculation, graphing tools, pivot tables, and a macro programming language that is called Visual Basic for Applications. Microsoft Excel has the basic features of all spreadsheets, using a grid of cells arranged in numbered rows and letter named columns to organize data manipulations like arithmetic operations. It has a battery of supplied functions to answer statistical, engineering and financial needs. In addition, it can display data as line graphs, histograms and charts, and with a very limited three-dimensional graphical display. It allows sectioning of data to view its dependencies on various factors for different perspectives (using pivot tables and the scenario manager). (Wikipedia, 2016a)

We use version Microsoft excel 2013 in data preprocessing and summarizing our dataset. For its common, many researchers using excel in preparation of data (Bratt & Moodley, 2015) (Natek & Zwilling, 2013) ,some of them beyond that to use excel in more advanced way and perform some data mining models like (Tang, 2008).

2- IBM® SPSS® Modeler:

IBM SPSS Modeler is a data mining and text analytics software application has been built by IBM. It is used to build predictive models and conduct other analytic tasks. It has a visual interface which allows users to leverage statistical and data mining algorithms without programming.

IBM SPSS Modeler was originally named Clementine by its creators, Integral Solutions Limited. This name continued for a while after SPSS's acquisition of the product. SPSS later changed the name to SPSS Clementine, and then later to PASW Modeler. Following IBM's 2009 acquisition of SPSS, the product was renamed IBM SPSS Modeler, its current name.(Wikipedia, 2016b)

We use SPSS Modeler to accomplish our training and testing models of prediction, the software provide wide range of data mining models.

We use the virgin IBM® SPSS® Modeler 14.2.

4.2.2 Applying the models

In this section we describe how each of the classification models (Decision tree, KNN, Multinomial Logistic regression, Naïve Bayesian) where applied on our data, and the specific settings required for each model.

In the four models, we divided our data set into two partitions: 1- Training data set and 2- Testing data set, with ratio 80% training and 20% testing. The following Table 4.7 shows the number of records for each partition.

Table 4.7 Data set partitions details

partition title	No. of records	Ratio
Training	4089	80%
Testing	1026	20%
sum	5115	100%

The experiment workflow illustrated in Figure 4.2, in applying decision tree classifier. The other three models are replacing the decision tree classifier and follows the exactly same workflow.

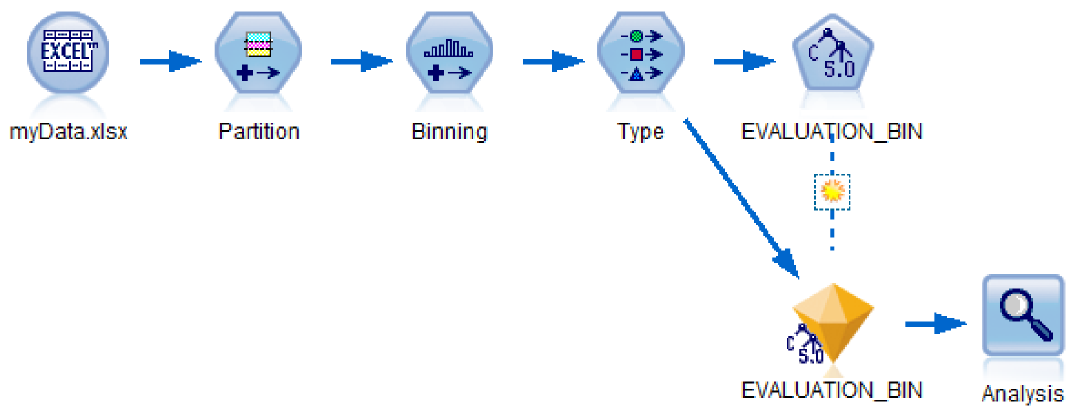

Figure 4.2 Experiment Workflow

4.2.3 Decision tree:

We used the Decision tree model with the following settings that was described in next table:

Table 4.8 Decision tree model settings

Use partitioned data: true	Alpha for Merging: 0.05
Partition: Partition	Epsilon For Convergence: 0.0
Calculate predictor importance: true	Maximum iterations for convergence: 100
Calculate raw propensity scores: false	Use Bonferroni adjustment: true
Calculate adjusted propensity scores: false	Allow splitting of merged categories: false
Continue training existing model: false	Chi-Square method: Likelihood ratio
Use frequency: false	Stopping criteria: Use absolute value
Use weight: false	Minimum records in parent branch: 30
Levels below root: 5	Minimum records in child branch: 5
Alpha for Splitting: 0.05	Use misclassification costs: false

The next figures: Figure 4.3, Figure 4.4, Figure 4.5 and Figure 4.6 show some settings based on the interface of SPSS Modeler:

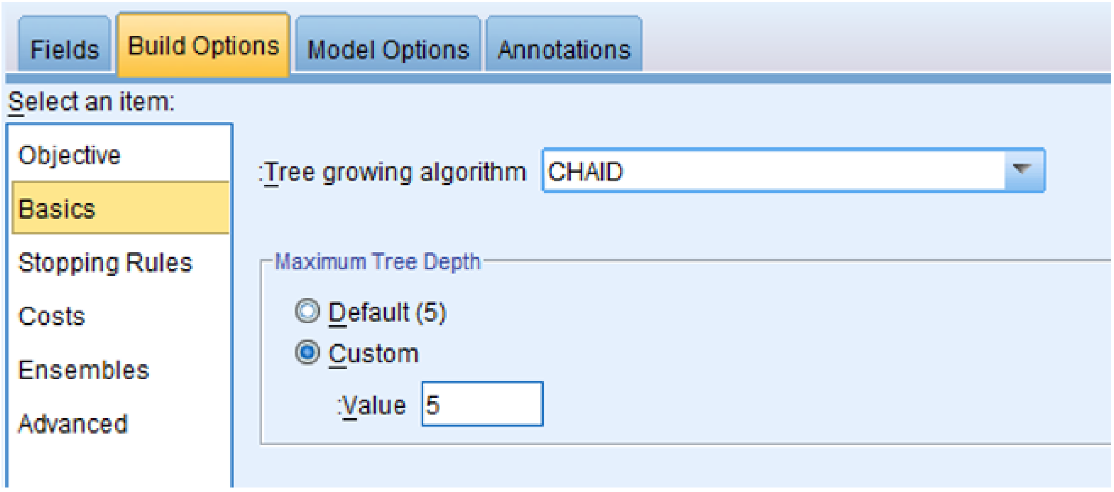

Figure 4.3 Basic Settings for Decision tree

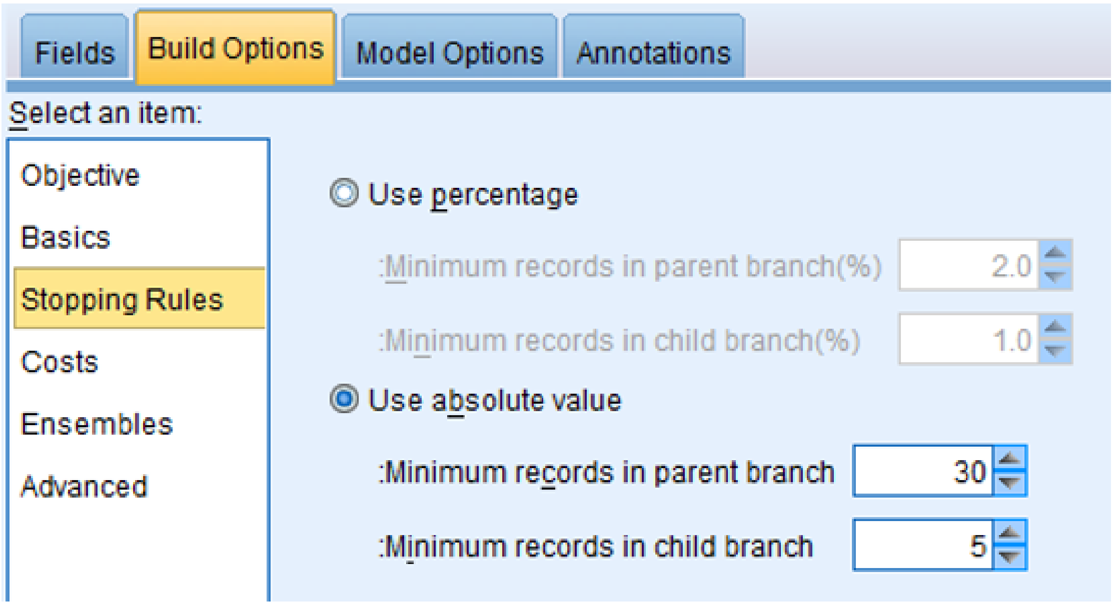

Figure 4.4 Stopping Rules Settings for Decision tree

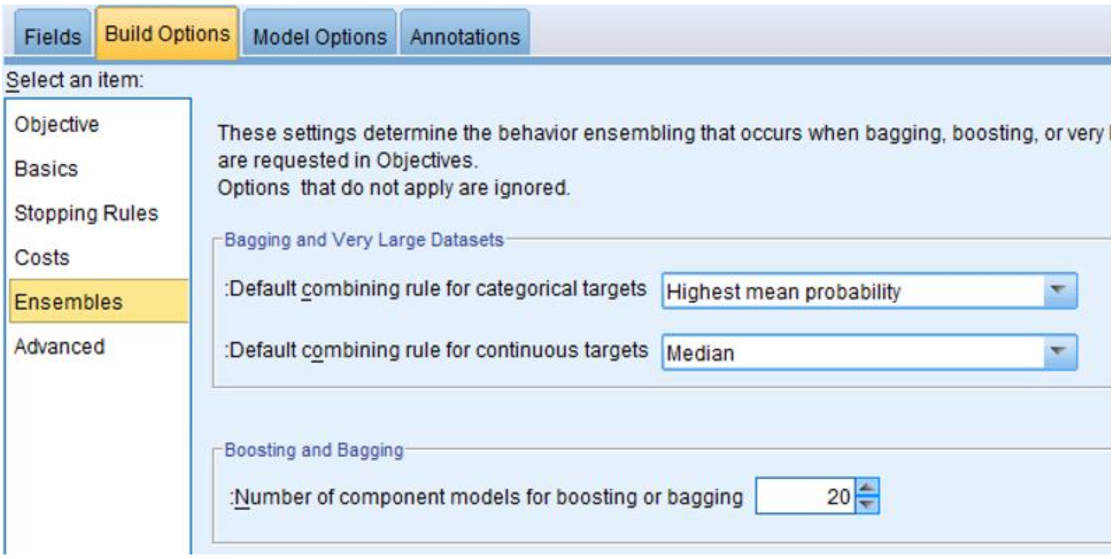

Figure 4.5 Ensembles Settings for Decision tree

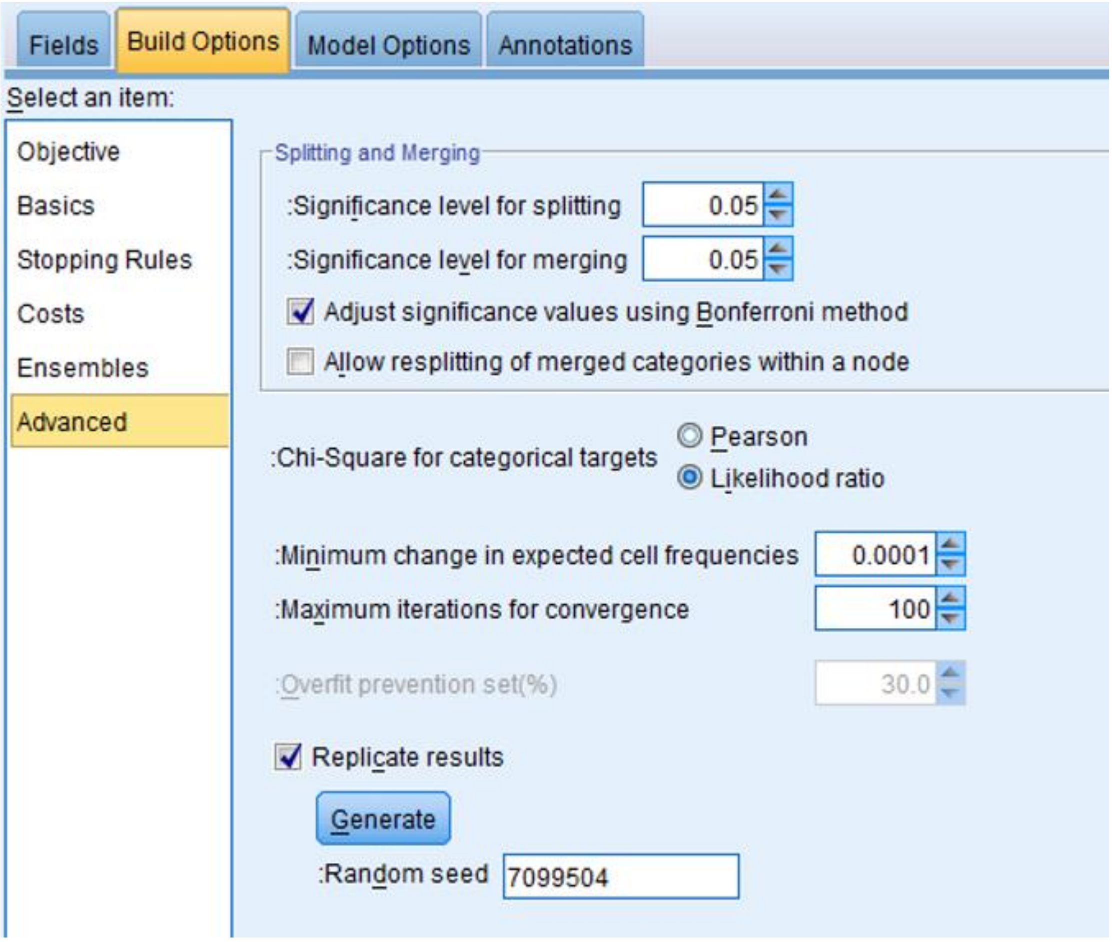

Figure 4.6 Advanced Settings for Decision tree

4.2.4 K-Nearest Neighbour:

KNN model was applied with the following settings that was illustrated in Table 4.9, it is also appeared in Figure 4.7 and Figure 4.8 based on the interface of SPSS Modeler.

Table 4.9 KNN model settings

Use partitioned data: true	K: 3
Partition: Partition	Minimum: 3
Calculate raw propensity scores: false	Maximum: 7
Calculate adjusted propensity scores: false	Perform feature selection: false
What type kof analysis do you want to perform?: Predict a target field	Forced entry: [ ]
What is your objective?: Custom analysis	Stop when the change in the absolute error ratio is less than or equal to the minimum: false
Use case labels: true	Number to select: 10
Case Label variable: Governorate_Name	Minimum change: 0.01
Identify focal record: false	Use field to assign cases: false
Focal Case Value: 0	Folds variable: Partition
Normalize range inputs: true	Number of folds: 10
Distance Computation: Euclidean metric	Set random seed: false
Predictions for Range Target: Mean of nearest neighbor values	Seed: 12,345
Weight features by importance when computing distances: true	Append all probabilities: false
Automatically select k: false	Save distances between cases and k nearest neighbors: false

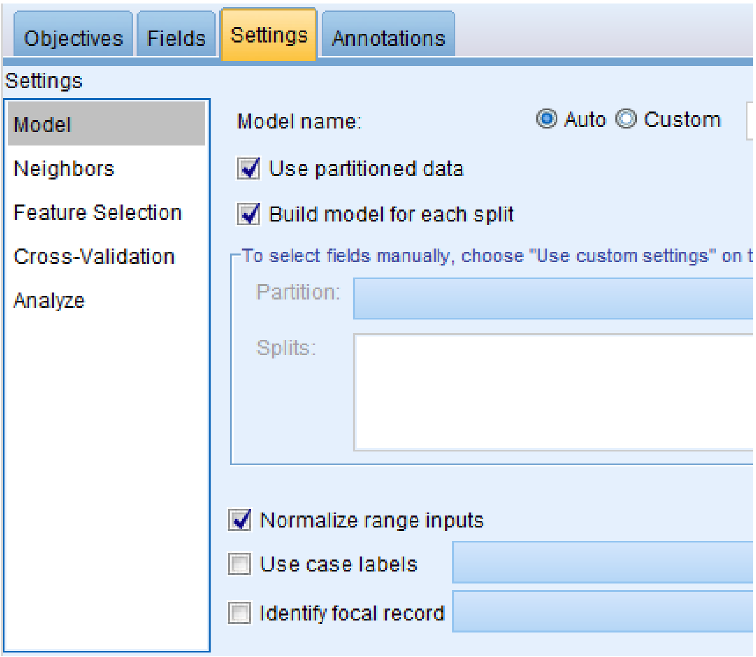

Figure 4.7 Model settings for KNN model

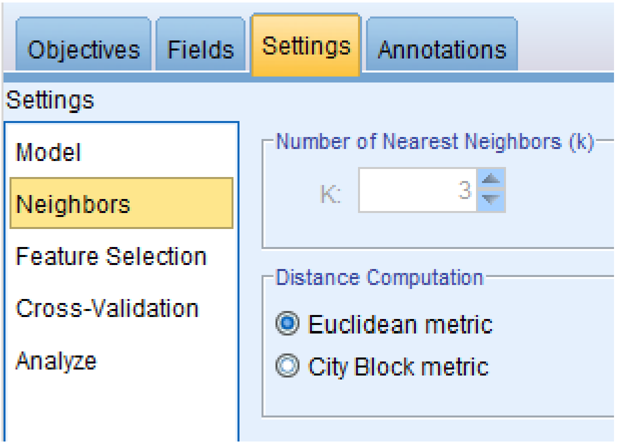

Figure 4.8 Neighbors settings for KNN model

4.2.5 Multinomial Logistic regression:

Figure 4.9 shows the settings that are applied on Multinomial Logistic regression model.

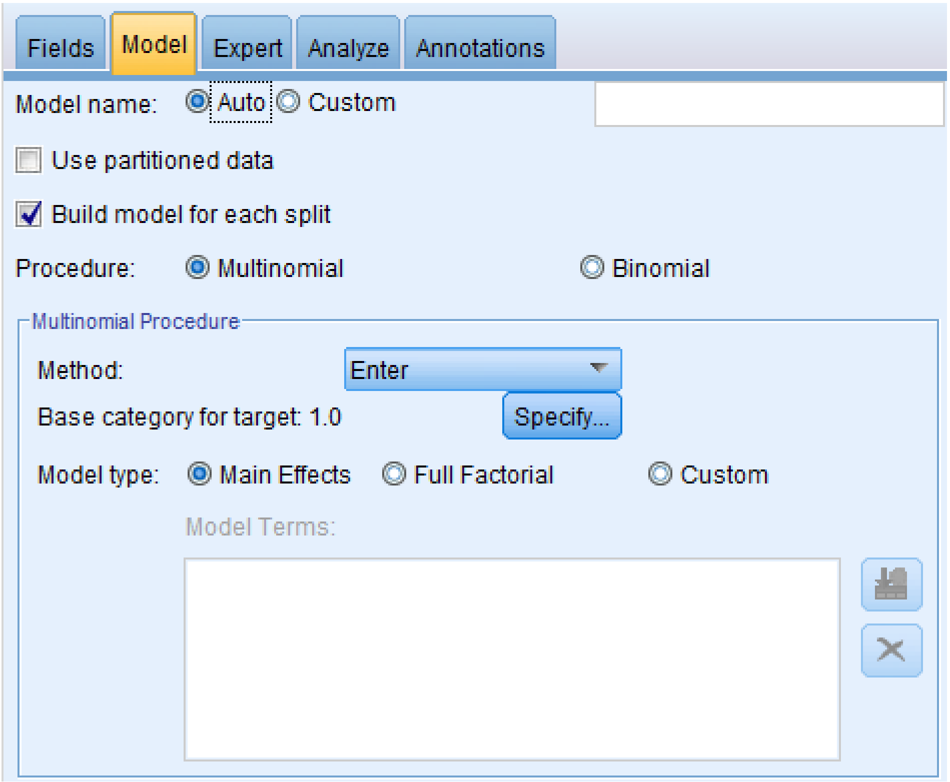

Figure 4.9 Multinomial Logestic Regressition model settings

4.2.6 Naïve Bayesian:

Figure 4.10 show the settings that applied on Naïve Bayesian model.

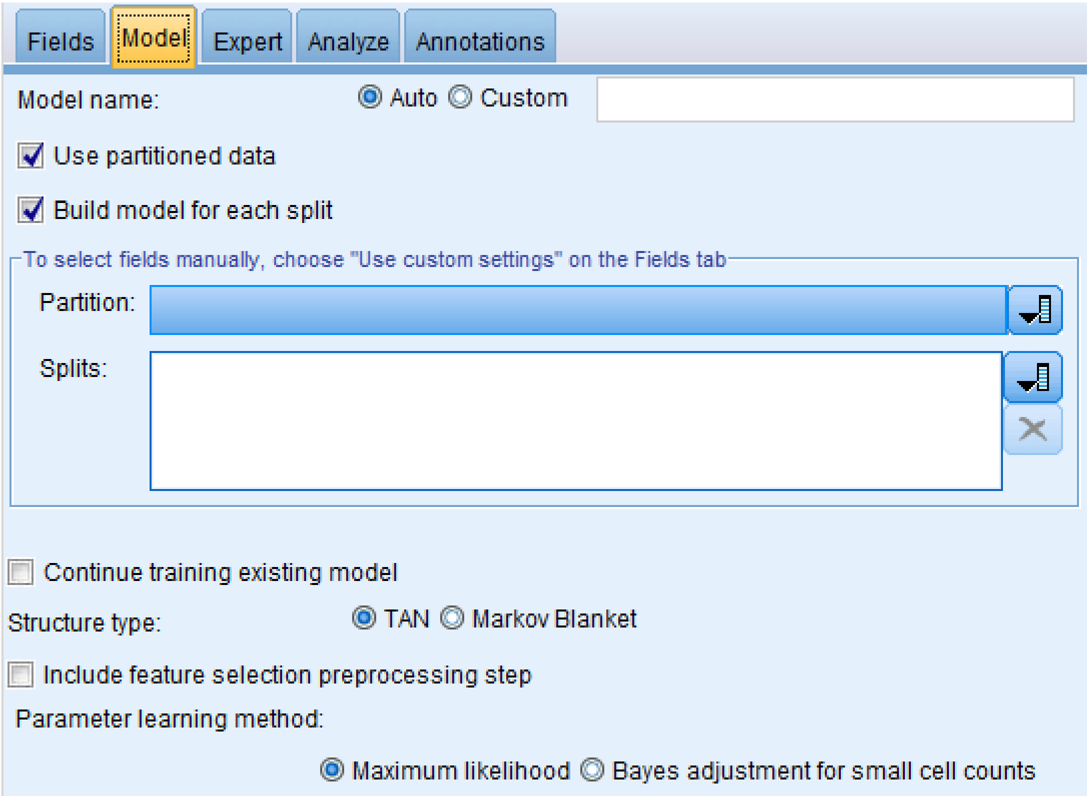

Figure 4.10 Naïve Bayesian model settings

Table 4.10 NB model settings

Use partitioned data: true	Parameter learning method: Maximum likelihood
Partition: Partition	Mode: Simple
Calculate predictor importance: true	Use only complete records: true
Calculate raw propensity scores: false	Append all probabilities: false
Calculate adjusted propensity scores: false	Independence test: Likelihood ratio
Use frequency field: false	Significance level: 0.01
Continue training existing model: false	Maximal conditioning set size: 5
Structure type: TAN	Inputs always selected: []
Include feature selection preprocessing step: false	Maximum number of inputs: 10

4.3 Summary

In this chapter, we explained our proposed method, and showed the training and testing steps that were performed on the data. Section 4.1, Proposed Method, talked about data collection, description, integration and preprocessing in details. Section 4.2, Implementation, explained the implementation environment, settings, and steps to apply the proposed method. The next chapter presents and discusses the results of our research.

Chapter 5

Results and discussion

Chapter 5
Results and discussion

The testing and training experiments were described in the previous chapter were repeated in a sequence of experimental attempts. In each attempt, some aspects were altered from the previous attempts with the aim of achieving a better accuracy. In each attempt, some of the four models processed the data. Otherwise we set the accuracy by the value "n/a" that means that the model is not applicable here. The following table lists and describes each of the experimental attempts we had performed.

Table 5.1 Not acceptable experimental attempts results

No.	Title and Description	Accuracy			
		DT	KNN	MLR	NB
1	Group data set by EMO_NO column: By calculate the average of evaluation column for each lecturer.	60%	55.6%	n/a	n/a
2	Generalize the next attributes to point to the continent instead of the country: Contry_Bc Country_M Country_Dr Delete the attributes : Specialist_Detail_A_Name_BC Specialist_Detail_A_Name_M Specialist_Detail_A_Name_Dr group the next attributes to reduce the values by devided the period to 4 sub periods: Birth_Date End_Date_Bc End_Date_M End_Date_Dr	23.5%	20.1%	14.2%	n/a
3	Delete the 18 attributes that describe bachelore, master and doctora characterestics and change them by 6 attributes describe the last qualification. To avoid messing values existance.	67.8%	65.1%	60.9%	58.7%
4	Dispensing of qualifications columns and put it in rows, for example, Lecturer X has 10 column for personal data and 18 column for qualifications and 1 column for evaluation, we devided the 18 column for 3 records (6 attribute for each record) and repeated the other 11 column with each record of 3 records. Then we delete the column that is has empty data.	42.0%	n/a	n/a	n/a
5	Split the evaluation attribute to 8 periods as next >=90 >=80 and <90	31%	27.3%	n/a	n/a

No.	Title and Description	Accuracy			
		DT	KNN	MLR	NB
	>=70 and <80				
	>=60 and <70				
	>=50 and <60				
	>=40 and <50				
	>=30 and <40				
	<30				

In this chapter, we present results of applying data mining models that was presented in previous chapter. We depend on accuracy to evaluate the use models; we did not found any special cases that need to apply the other evaluation methods like precision, recall and F-measure. This chapter is divided into 5 sections: section 1 Decision tree results, section 2 KNN results, section 3 MLR results, section 4 Naïve Bayesian results, section 5 we summarize the previous sections and identify the best model, then compare and discuss the result of best model with other researches.

5.1 Decision tree

An important topic in classification is how to focus your modeling efforts on the predictor fields that matter most and consider dropping or ignoring those that matter least. The predictor importance chart helps to do this by indicating the relative importance of each predictor in estimating the model. Predictor importance does not relate to model accuracy. It just relates to the importance of each predictor in making a prediction, not whether or not the prediction is accurate (IBM_Knowledge_Center, 2016).

We found that the highest importance on our data set is the attribute Specialist_Detail_A_Name_Dr, so the tree is splitted firstly depending on this column, Figure 5.1 illustrate the highest predictor importance in our research. In addition, Table 5.2 illustrate the values of predictor importance.

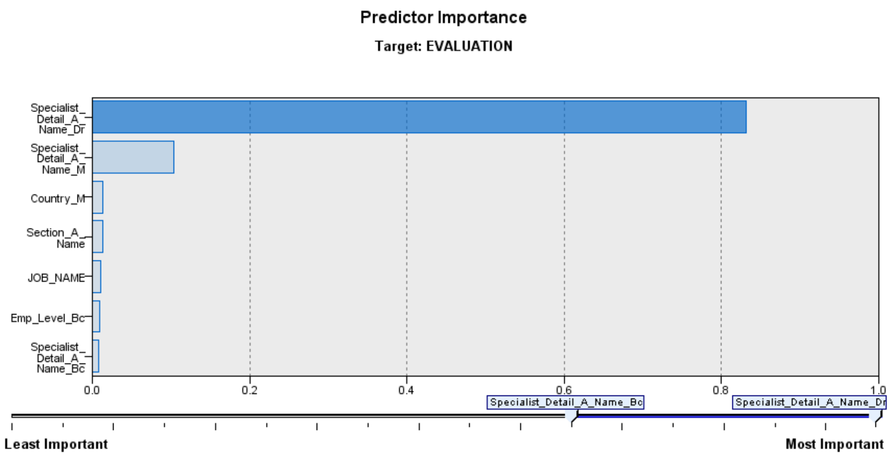

Figure 5.1 Decision Tree Predictor Importance

Table 5.2 Impotrance values for decision tree classification

Nodes	Importance
Specialist_Detail_A_Name_Dr	0.8316
Specialist_Detail_A_Name_M	0.1033
Country_M	0.0134
Section_A_Name	0.0131
JOB_NAME	0.0099
Emp_Level_Bc	0.0089
Specialist_Detail_A_Name_Bc	0.0073
End_Date_Bc	0.0065
Country_Dr	0.0024
Birth_Date	0.0019
Gender	0.0018

After applying the decision tree depending on splitting with the highest importance, the IBM SPSS Modeler generate the tree, Figure 5.2 shows a part of decision tree.

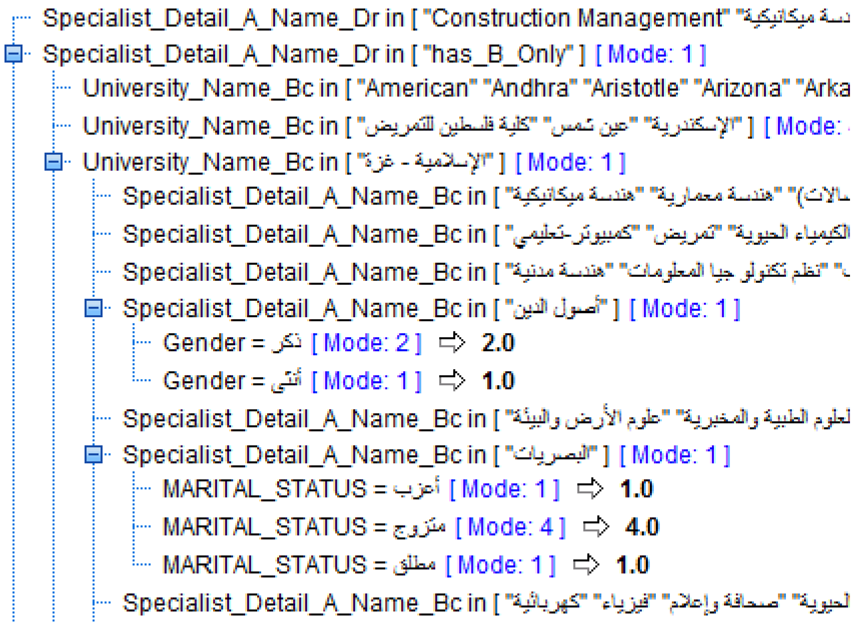

Figure 5.2 Part of Decision tree

One of the rules Rule1 says that ***IF*** Specialist_Detail_A_Name_Dr in Set1 ***Then*** Evaluation = 4.0, that means that lecturer's evaluation is predicted to be greater than or equal 85.7%. Where Set1 is the set of the items that listed in Table 5.3.

More detailed roles of the decition tree is illustrated in Appendix B

Table 5.3 Set1 in Rule1

Construction Management	التنمية الزراعية	تخطيط معماري	ميكانيكا
اتصالات والكترونيات	الصحة الإنجابية	تخطيط وتصميم المواقع	نظم المعلومات الجغرافية
الاختصاص الجنائي العالمي	القانون الدولي	جيوفيزياء	هندسة الاتصالات الكهربائية
الاستزراع المائي	الهندسة الكهربائية والإلكترونية	دراسات إعلامية	هندسة الشواطئ
الإسكان	الهندسة المدنية/ البيئية	رياضيات بحتة	هندسة صناعية
الأدب والنقد	إدارة مشاريع هندسية	صحة عامة	هندسة كهربائية
الآثار/ قديم	أدب ونقد	طب وجراحة العيون/زراعة عدسات	هندسة معمارية
البلاغة العربية	بنوك ومالية	فيزياء جوامد	هندسة موائع
البورد الفلسطيني+البورد الأردني	بيئة	كيمياء المبيدات	
التحاليل الطبية	تاريخ حديث	لغة إنجليزية	
التعليم الإلكتروني	تاريخ عرب حديث	مصادر مياه وبيئة	

After we had applied our testing data set that contains 1026 records, we got the next results that illustrated in Table 5.4.

Table 5.4 Detailed coincidence matrix for Decision tree

'Partition' = Testing		Predicted values				Sum
		1	**2**	**3**	**4**	
Actual values	1	192	25	31	19	267
	2	36	163	24	27	250
	3	32	55	160	20	267
	4	19	25	37	161	242

The header of column is for predicted values and the header of rows for actual values, which means that 267 records of 1026 records of testing data set have the target 1, and the Decision tree model is predicting these values as 192 for target 1, 25 for target 2, 31 for target 3 and 19 for target 4. That is mean the model predicts the target class truly 192 times. And so with 2, 3 and 4 target class labels.

This leads that the overall truly predicted values is (192+163+160+161) = 676 times with 65.89% this called the accuracy of the model, Table 5.5 illustrates the accuracy of the model.

Table 5.5 General coincidence matrix for Decision tree

'Partition'	Testing	Percentage
Correct	676	65.89%
Wrong	350	34.11%
Total	1,026	100.00%

If we suppose that the closest prediction of the true evaluation is true we can say that the model is predicting the evaluation truly or far from true in one step for 873 times, that means that the model accuracy is 85%

5.2 K-Nearest Neighbor

In this research we apply the KNN model with 6 experiments by editing the value K in each time, because the difficulty of determining the k value, Table 5.6 illustrates the accuracy of the model and value of K for each time.

Table 5.6 KNN accuracy with defferent K

K	Accuracy
2	62.77%
3	65.20%
4	62.48%
5	61.60%
6	60.04%
7	59.94%

We found that the model gives the highest accuracy when K = 3, so we depend the model with K=3, Table 5.7 shows the coincidence matrix for KNN model when k=3.

Table 5.7 Detailed coincidence matrix for KNN

'Partition' = Testing		Predicted values				sum
		1	2	3	4	
Actual values	1	194	28	26	13	261
	2	35	149	57	33	274
	3	21	24	170	46	261
	4	25	21	28	156	230

Table 5.8 General coincidence matrix for KNN

'Partition'	Testing	Percentage
Correct	669	65.20%
Wrong	357	34.80%
Total	1,026	100.00%

If we suppose that the closest prediction of the true evaluation is true we can say that the KNN model is predicting the evaluation truly or far from true in one step for 887 times, that means that the model accuracy is 86.4%

5.3 Multinomial Logistic Regression

When implemented Multinomial Logistic Regression, it presented the accuracy 67.64%. Table 5.9 presents the predicted and actual values, and Table 5.10 illustrates the overall correct prediction values.

Table 5.9 Detailed coincidence matrix for Multinomial Logistic Regression

'Partition' = Testing		Predicted values				sum
		1	**2**	**3**	**4**	
Actual values	1	180	31	34	16	261
	2	28	164	54	28	274
	3	18	23	185	35	261
	4	16	26	23	165	230

Table 5.10 General coincidence matrix for Multinomial Logistic Regression

'Partition'	Testing	Percentage
Correct	694	67.64%
Wrong	332	32.36%
Total	1,026	100.00%

If we suppose that the closest prediction of the true evaluation is true we can say that the model is predicting the evaluation truly or far from true in one step for 888 times, that means that the model accuracy is 86.5%

5.4 Naïve Bayesian

Table 5.11 Detailed coincidence matrix for Naïve Bayesian

'Partition' = Testing		Predicted values				Null	sum
		1	2	3	4		
Actual values	1	166	37	35	21	8	267
	2	31	146	34	29	10	250
	3	26	43	154	32	12	267
	4	14	25	30	158	15	242

Table 5.12 General coincidence matrix for Naive Bayesian

'Partition'	Testing	Percentage
Correct	624	60.82%
Wrong	402	39.18%
Total	1,026	100.00%

If we suppose that the closest prediction of the true evaluation is true we can say that the model is predicting the evaluation truly or far from true in one step for 831 times, that means that the model accuracy is 80.9%

5.5 Summary

This chapter described experiments results of predicting lecturer performance based on lecturer characteristics and historical student evaluation of lecturers. We used four data mining techniques: Decision tree, K Nearest Neighbor, Multinomial Logistic Regression and Naïve Bayesian.

The four techniques gives different accuracy values based on data nature, many attempts had done on the data set to give accuracies better than accuracies that were illustrated in previous sections, the final accuracies are listed in Table 5.13 bellow based on considering the predicted values that neigbour the true value is true.

Table 5.13 Used models accuracies based on considering the predicted values that neigbour the true value is true

Model	Accuracy
Decision Tree	85.0%
K Nearest Neighbor	86.4%
Multinomial Logistic Regression	86.5%
Naïve Bayesian	80.9%

As clear in Table 5.13, three models are relatively closed to one another, Decision Tree, K Nearest Neighbor and Multinomial Logistic Regression, the fourth model Naïve Bayesian gives the least accuracies.

Clearly, Multinomial Logistic Regression achieves highest accuracy, and therefore, it should be adopted as the model for predicting lecturer performance.

The researcher attributes the low accuracies in general to the nature of the data set that was entered in very long period of time, and no clear criteria or template for entering data, that makes the quality of the data low, which reflects on the mining of this data and requires a lot of effort to extract knowledge.

We found that our reseach results are consisted with results of (Sok-Foon et al., 2012) were they showed that lecturer and tutor characteristics, subject characteristics, and learning resources and facilities explained 61.9% of the variance in overall lecturer performance among students.

The study of (Mardikyan & Badure, 2011) presents prediction of lecturer performance based on course characteristics and lecturer characteristics from students view, it uses only two attributes about personal characteristics: whether or not the instructor is a part time and giving the course for the first time. The study found that the courses offered by part time instructors tend to receive higher ratings. Hence, Administration should offer a variety of elective courses given by part time instructors. In our research, we found that this attribute has no significant and does not contribute in decision tree as predictor.

(Ahmadi & Abadi, 2013) were reaching to relations between parameters such as Evaluation‘s score, Teacher’s degree, Degree’s Type, Teaching experience, Acceptation to next semesters on teacher’s evaluation . They used the acceptation attributes as class label that take two values Yes or No. The study limited on four variables related to lecturer, and did not exploit the historical data for mining. But in our study we used many attributes related to lecturer characteristics. We found that there is consistency with our research in the common lecturer characteristic that is named Teacher's degree; it has a relation with the evaluation attribute and contribute in prediction in both two researches.

Chapter 6
Conclusion and future work

Chapter 6

Conclusion and future work

6.1 Conclusion

As a research of investigating data mining techniques for predicting Lecturer's performance based on lecturer characteristics and historical evaluation of lecturer, we collect the related data from two department at Islamic University of Gaza, academic affairs and employees affairs departments, the data belonging to the academic stuff that was taught at the university in 11 semesters start from second semester 2011/2012 and at the end of summer semester 2014/2015. The data set contain the evaluation of the lecturers in the courses that was taught by these lecturers. we preprocessed the data and format it to be in predictable format, then we apply four famous classification techniques, Decision tree, K-Nearest Neighbors, Multinomial Logistic Regression and Naïve Bayesian. If we suppose that the closest prediction of the true evaluation is true the four techniques give closed results as follow, Decision tree 85.0%, K-Nearest Neighbors 86.4%, Multinomial Logistic Regression 86.5%, and Naïve Bayesian 80.9%, the Multinomial Logistic Regression model gives the highest result so we adopted it.

6.2 Future work

In future works, the following issues will be considered, we can improve predicting lecturer's performance by using more attributes that may affect the evaluation of lecturer performance, for example an important attribute was ignored from this research is as an experience years, it is ignored for the large missing values. We can try to apply new data mining techniques that may give results that are more accurate. We may use data from another destination like lecturer's characteristics that is found in the civil registry of the ministry of the interior.

Referencies List

Referencies List

Agaoglu, M. (2016). Predicting Instructor Performance Using Data Mining Techniques in Higher Education. *4*.

Ahmadi, F., & Abadi, S. (2013). Data Mining in Teacher Evaluation System using WEKA. *International Journal of Computer Applications, 63*(10), 12-18.

Bhardwaj, B. K., & Pal, S. (2011). Data Mining: A prediction for performance improvement using classification. *(IJCSIS) International Journal of Computer Science and Information Security, 9*.

Bratt, S., & Moodley, K. (2015). *Promoting Public Library Sustainability through Data Mining: R and Excel*. Retrieved March 5,2016, from:http://library.ifla.org/1257/1/180-bratt-en.pdf

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, *17*(3), 37.

Giudici, P. (2003). *Applied Data Mining, Statistical Methods for Business and Industry* (2nd Ed.). USA: Wiley.

Gül, H. (2010). Evaluation of Lecturer Performance Depending on Student Perception in Higher Education. *Egitim ve Bilim, 35*(158).

Gupta, S., & Suma, V. (2014). *Prediction of Human Performance Capability during Software Development Using Classification*. Paper presented at the ICT and Critical Infrastructure: Proceedings of the 48th Annual Convention of Computer Society of India, India.

Hajizadeh, N., & Ahmadzadeh, M. (2014). Analysis of factors that affect students' academic performance-Data Mining Approach. *International Journal of advanced studies in Computer Science and Engineering IJASCSE, 3*(8).

Han, J., & Kamber, M. (2011). *Data Mining: Concepts and Techniques* (2nd ed.). USA: Diane Cerra

Hand, D., Mannila, H., & Smyth, P. (2001). *Principles of Data Mining* Cambridge, Massachusetts London England The MIT Press

IBM_Knowledge_Center. (2016). *Predictor Importance*. Retrieved May 20,2016, from:https://www.ibm.com/support/knowledgecenter/SS3RA7_15.0.0/com.ibm.spss.modeler.help/model_nugget_variableimportance.htm

Kantardzic, M. (2003). *Data Mining: Concepts, Models, Methods, and Algorithms*. USA: John Wiley & Sons.

Larose, D. T. (2005). *Discovering Knowledge in Data An Introduction to Data Mining*: John Wiley & sons.

Ma, Y., Liu, B., Wong, C. K., Yu, P. S., & Lee, S. M. (2000). *Targeting the right students using data mining.* Paper presented at the Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, National University of Singapore,Singapore.

Mardikyan, S., & Badure, B. (2011). *Analyzing Teaching Performance of Instructors Using Data Mining Techniques*. Paper presented at the Informatics in Education, Vilnius University,Lithuania.

Miles, K. H., Darling-Hammond, & Linda. (1998). Rethinking the allocation of teaching resources: Some lessons from high-performing schools. *Educational Evaluation and Policy Analysis, 20*(1), 9-29.

Millmore, M., Lewis, P., Saunders, M., Thornhill, A., & Morrow, T. (2007). *Strategic Human Resources Management Contemporary Issues*. London: Prentice Hall.

Molefe, G. N. (2010). Performance measurement dimensions for lecturers at selected universities: An international perspective. *SA Journal of Human Resource Management, 8*(1), 13 pages.

Naik, R. L., Sarma, S., Manjula, B., & Ramesh, D. (2011). Study of Trends in Higher Education. *International Journal of Computer Trends and Technology*, 6-10.

Natek, S., & Zwilling, M. (2013). *Data mining for small student data set: Knowledge management system for higher education teachers*. Paper presented at the management, knowledge and learning international conference, Zadar.

Olson, D. L., & Delen, D. (2008). *Advanced Data Mining Techniques*. USA: Springer-Verlag Berlin Heidelberg

Pal, A. K., & Pal, S. (2013). Evaluation of Teacher's Performance: A Data Mining Approach. *International Journal of Computer Science and Mobile Computing, 2*(12), 359-369.

Peleyeju, J. O., & Ojebiyi, O. A. (2013). lecturers' performance appraisal and total quality management of public universities in south-western nigeria. *British Journal of Education, 1*(2), 41-47.

Phyu, T. N. (2009). *Survey of classification techniques in data mining.* Paper presented at the Proceedings of the International MultiConference of Engineers and Computer Scientists, Hong Kong.

Samian, Y., & Noor, N. M. (2012). Student's Perception on Good Lecturer based on Lecturer Performance Assessment. *Procedia-Social and Behavioral Sciences, 56*, 783-790.

Sok-Foon, Y., Sze-Yin, J. H., & Yin-Fah, B. C. (2012). Student evaluation of lecturer performance among private university students. *Canadian Social Science, 8*(4), 238-243.

Starkweather, J., & Moske, A. K. (2011). *Multinomial logistic regression*. Retrieved March 13,2016, from:http://researchsupport.unt.edu/class/Jon/Benchmarks/MLR_JDS_Aug2011.pdf

Tang, H. (2008, 12-14 Oct). *A Simple Approach of Data Mining in Excel.* Paper presented at the Wireless Communications, Networking and Mobile Computing, 4th International Conference, China.

Wang, M. C., Dziuban, C. D., Cook, I. J., & Moskal, P. D. (2009). Using data-mining techniques to examine student ratings of instruction *Quality research in literacy and science education* (pp. 383-398). University of Central Florida: Springer.

Wikipedia. (2016a). *Microsoft_Excel*. Retrieved Mar.19.2016, from:https://en.wikipedia.org/wiki/Microsoft_Excel

Wikipedia. (2016b). *SPSS_Modeler*. Retrieved Mar.19.2016, from:https://en.wikipedia.org/wiki/SPSS_Modeler

Witten, I. H., & Frank, E. *Data Mining, Practical Machine Learning Tools and Techniques* (Second ed.): Diane Cerra

Appendices

Appendix A: IUG questionnaire for lecturer's performance

بسم الله الرحمن الرحيم

الشئون الأكاديمية	**عمادة الجودة والتطوير**

استطلاع رأي الطلبة في أداء الأستاذ الجامعي

الفصل الدراسي الأول- العام الجامعي 2016/2015

أخي الطالب ، أختي الطالبة: تهدف هذه الاستمارة إلى استطلاع رأيك فيما يتعلق بتعلمك في هذا المساق من خلال تقييم الأداء التدريسي ، وذلك سعياً للارتقاء بالمستوى الأكاديمي في البرنامج ، وتحقيقاً لأهداف الجامعة في التطوير المستمر ، آملين مراعاة الدقة والموضوعية ، علماً أنه لا توجد إجابات صحيحة وأخرى خاطئة.

الكليـــة:		القسـم:	
المساق:		الشعبة:	
المدرس:		المعدل التراكمي للطالب/ة:	% (اختياري)

نرجو إبداء الرأي في الجوانب المختلفة بوضع علامة (✓) في الخانة التي تعبر عن رأيك باستخدام المقياس التالي:

المفتاح: (5)= بدرجة كبيرة جداً ، (4)= بدرجة كبيرة ، (3)= بدرجة متوسطة ، (2)= بدرجة قليلة ، (1)= بدرجة قليلة جداً

م	البـنـــد	5	4	3	2	1
1	يظهر المدرس تمكناً جيداً لموضوعات المساق.					
2	يقدم محتوى المساق وفصوله بطريقة واضحة ومنظمة.					
3	يشرح مفردات المساق بطريقة معمقة ومتنوعة.					
4	يوصل المدرس أفكار المساق ومفاهيمه بأسلوب واضح وسلس.					
5	يتصف المدرس بالحيوية والنشاط، ويشرح بصوت واضح ومسموع.					
6	يستخدم المدرس أساليب تعلم متنوعة من طرق التدريس (المحاضرة، المناقشة .....).					
7	يستخدم المدرس وسائل تكنولوجيا التعلم الحديثة (البوربوينت، برنامج مودل .....).					
8	يوضح المدرس المفاهيم والجوانب النظرية للمساق بأمثلة مرتبطة بالواقع إن أمكن.					
9	يشجع المدرس الطلبة على المشاركة من خلال طرح أسئلة وتوفير فرصاً للمناقشة.					
10	يحسن المدرس الاستفادة من وقت المحاضرة بالتدريس الفعلي.					
11	يستطيع المدرس ضبط المحاضرة وإدارتها بشكل فاعل.					
12	يلتزم المدرس بمواعيد بدء وانتهاء المحاضرة.					
13	صفحة المدرس الشخصية على الموقع الإلكتروني للجامعة محدّثة وتضم مواد وأبحاث علمية ومواقع داعمة لتعلم الطلبة.					
14	يعامل المدرس الطلبة بود واحترام، ويقدم المساعدة عند الحاجة.					
15	يتقبل المدرس اقتراحات الطلبة بسعة صدر.					
16	يستخدم المدرس وسائل تقييم متنوعة (مثل: أبحاث، اختبارات قصيرة، نصفية ونهائية، تعيينات وواجبات .....).					
17	يعتبر المدرس موضوعياً ومنصفاً في تقييمه للطلبة.					
18	يقدم المدرس تغذية راجعة للأنشطة التقييمية (يحل أسئلة الامتحانات، يوزع أوراق الإجابة، يجيب على أسئلة الطلبة أثناء المحاضرة ......).					
19	يضع المدرس خطة تدريس المساق على صفحته الشخصية الالكترونية أو المودل.					
20	يلتزم المدرس بالساعات المكتبية المحددة للمساق.					

21	1- عدد المحاضرات التي تغيب عنها المدرس:	.... محاضرات
	2- عدد المحاضرات التي قام المدرس بتعويضها (المحاضرات البديلة):	.... محاضرات
22	أكثر ثلاث أشياء ساعدت في تعلمي المساق: 1- .... 2- 3-	
23	أكثر ثلاث أشياء لم تعجبني في المساق: 1- .... 2- 3-	
24	في رأيك كيف يمكن تحسين أو تطوير هذا المساق؟ 1- .... 2- 3-	

مع الشكر والتقدير،،،

Appendix B: Rule2 of decition tree result

Rule2 says that ***IF*** Specialist_Detail_A_Name_Dr in ["has_B_Only"], This means that the lecturer's qualification is Bachelor (Education_Name column = "بكالوريوس"), ***Then:***

IF University_Name_Bc in Set1, ***Then*** Evaluation = 1.0 , that means that lecturer evaluation is predicted to be less than or equal 75.5%. Where Set1 is the set of the items that listed in Table 8.1.

- ***IF*** University_Name_Bc in ["كلية فلسطين للتمريض" "عين شمس" "الإسكندرية"] ***Then*** Evaluation = 4.0, that means that lecturer's evaluation is predicted to be greater than or equal 85.7%.
- ***IF*** University_Name_Bc in ["الإسلامية - غزة"] ***Then***:
    - ***IF*** Specialist_Detail_A_Name_Bc in Set2 ***Then*** Evaluation = 1.0, that means that lecturer's evaluation is predicted to be less than or equal 75.5%. Where set2 is the set of items that listed in Table 8.2.
    - ***IF*** Specialist_Detail_A_Name_Bc in [" "آداب إنجليزي" "أصول الدين" "الكيمياء الحيوية" "الكيمياء" "الكمبيوتر التعليمي" "التمريض""] ***Then*** Evaluation = 1.0, that means that lecturer's evaluation is predicted to be less than or equal 75.5%.
    - ***IF*** Specialist_Detail_A_Name_Bc in ["الأحياء"] ***Then*** :
        - ***IF*** Governorate_Name in ["رفح" "الوسطى"] ***Then*** Evaluation = 4.0 , that means that lecturer's evaluation is predicted to be greater than or equal 85.7%.
        - ***IF*** Governorate_Name in ["غزة" "خانيونس"] ***Then*** Evaluation = 4.0 , that means that lecturer's evaluation is predicted to be greater than or equal 85.7%.
        - ***IF*** Governorate_Name in ["شمال غزة"] ***Then*** Evaluation = 1.0, that means that lecturer's evaluation is predicted to be less than or equal 75.5%.
    - ***IF*** Specialist_Detail_A_Name_Bc in ["البصريات الطبية"] ***Then*** :
        - ***IF*** MARITAL_STATUS = أعزب ***Then*** Evaluation = 1.0, that means that lecturer's evaluation is predicted to be less than or equal 75.5%.

- *IF* MARITAL_STATUS = متزوج ***Then*** Evaluation = 4.0 , that means that lecturer's evaluation is predicted to be greater than or equal 85.7%.
- *IF* MARITAL_STATUS = مطلق ***Then*** Evaluation = 4.0 , that means that lecturer's evaluation is predicted to be greater than or equal 85.7%.

- ***IF*** Specialist_Detail_A_Name_Bc in [" الحاسوب" "البيئة وعلوم الأرض" "واساليب تدريسه" "حاسوب" "هندسة مدنية"] ***Then*** Evaluation = 2.0, that means that lecturer's evaluation is predicted to be between 75.6% and 80%.
- ***IF*** Specialist_Detail_A_Name_Bc in [" العلوم" "الرياضيات" "التحاليل الطبية" "الطبية والمخبرية" "علم الحاسوب" "علوم الأرض والبيئة"] ***Then*** Evaluation = 4.0 , that means that lecturer's evaluation is predicted to be greater than or equal 85.7%.
- ***IF*** Specialist_Detail_A_Name_Bc in ["التكنولوجيا الحيوية"] ***Then*** :
  - ***IF*** Governorate_Name in ["رفح" "خانيونس" "الوسطى"] ***Then*** Evaluation = 3.0, that means that lecturer's evaluation is predicted to be between 80.1% and 85.6%.
  - ***IF*** Governorate_Name in ["شمال غزة"] ***Then*** Evaluation = 3.0, that means that lecturer's evaluation is predicted to be between 80.1% and 85.6%.
  - ***IF*** Governorate_Name in ["غزة"] ***Then*** :
    - ***IF*** Emp_Level_Bc = جيد ***Then*** Evaluation = 3.0, that means that lecturer's evaluation is predicted to be between 80.1% and 85.6%.
    - ***IF*** Emp_Level_Bc = جيد جدا ***Then*** Evaluation = 4.0 , that means that lecturer's evaluation is predicted to be greater than or equal 85.7%.
    - ***IF*** Emp_Level_Bc = مقبول ***Then*** Evaluation = 2.0, that means that lecturer's evaluation is predicted to be between 75.6% and 80%.

            - ***IF*** Emp_Level_Bc = ممتاز ***Then*** Evaluation = 2.0, that means that lecturer's evaluation is predicted to be between 75.6% and 80%.

- ***IF*** Specialist_Detail_A_Name_Bc in ["الجغرافيا"] ***Then*** :
    - ***IF*** MARITAL_STATUS = أعزب ***Then*** Evaluation = 3.0, that means that lecturer's evaluation is predicted to be between 80.1% and 85.6%.
    - ***IF*** MARITAL_STATUS = متزوج ***Then*** Evaluation = 4.0 , that means that lecturer's evaluation is predicted to be greater than or equal 85.7%.
    - ***IF*** MARITAL_STATUS = مطلق ***Then*** Evaluation = 3.0, that means that lecturer's evaluation is predicted to be between 80.1% and 85.6%.
- ***IF*** Specialist_Detail_A_Name_Bc in ["الطب البشري"] ***Then*** :
    - ***IF*** MARITAL_STATUS = أعزب ***Then*** Evaluation = 2.0, that means that lecturer's evaluation is predicted to be between 75.6% and 80%.
    - ***IF*** MARITAL_STATUS = متزوج ***Then*** Evaluation = 1.0, that means that lecturer's evaluation is predicted to be less than or equal 75.5%.
    - ***IF*** MARITAL_STATUS = مطلق ***Then*** Evaluation = 1.0, that means that lecturer's evaluation is predicted to be less than or equal 75.5%.
- ***IF*** Specialist_Detail_A_Name_Bc in ["الفيزياء"] ***Then*** :
    - ***IF*** MARITAL_STATUS = أعزب ***Then*** :
        - ***IF*** Governorate_Name in ["الوسطى"] ***Then*** Evaluation = 2.0, that means that lecturer's evaluation is predicted to be between 75.6% and 80%.
        - ***IF*** Governorate_Name in ["خانيونس"] ***Then*** :

- *IF* Specialist_A_Name_Bc in Set3 ***Then*** Evaluation = 1.0, that means that lecturer's evaluation is predicted to be less than or equal 75.5%. Where Set3 is the set of items that listed in Table 8.3.
- *IF* Specialist_A_Name_Bc in ["علـوم"] ***Then*** Evaluation = 3.0, that means that lecturer's evaluation is predicted to be between 80.1% and 85.6%.
- *IF* Specialist_A_Name_Bc in ["فيزيـاء"] ***Then*** Evaluation = 1.0, that means that lecturer's evaluation is predicted to be less than or equal 75.5%.

- *IF* Governorate_Name in ["شـمـال غزة" "رفح"] ***Then*** Evaluation = 2.0, that means that lecturer's evaluation is predicted to be between 75.6% and 80%.
- *IF* Governorate_Name in ["غزة"] ***Then*** :

IF End_Date_Bc in Set4 ***Then*** Evaluation = 3.0, that means that lecturer's evaluation is predicted to be between 80.1% and 85.6%. Where Set4 is the set of items that listed in

- Table 8.4.
- *IF* End_Date_Bc in ["2011"] ***Then*** Evaluation = 2.0, that means that lecturer's evaluation is predicted to be between 75.6% and 80%.
- *IF* End_Date_Bc in ["2014"] ***Then*** Evaluation = 3.0, that means that lecturer's evaluation is predicted to be between 80.1% and 85.6%.

- *IF* MARITAL_STATUS = متزوج ***Then*** Evaluation = 1.0, that means that lecturer's evaluation is predicted to be less than or equal 75.5%.

    - ***IF*** MARITAL_STATUS = مطلق ***Then*** Evaluation = 2.0, that means that lecturer's evaluation is predicted to be between 75.6% and 80%.

- ***IF*** Specialist_Detail_A_Name_Bc in ["اللغة الإنجليزية"] ***Then*** :
    - ***IF*** End_Date_Bc in Set5 ***Then*** Evaluation = 2.0, that means that lecturer's evaluation is predicted to be between 75.6% and 80%. Where Set5 is the set of items that listed in Table 8.5.
    - ***IF*** End_Date_Bc in ["2013"] ***Then*** :
        - ***IF*** Emp_Level_Bc in ["جيد"] ***Then*** Evaluation = 1.0, that means that lecturer's evaluation is predicted to be less than or equal 75.5%.
        - ***IF*** Emp_Level_Bc in ["جيد جدا"] ***Then*** :
            - ***IF*** MARITAL_STATUS = أعزب ***Then*** Evaluation = 3.0, that means that lecturer's evaluation is predicted to be between 80.1% and 85.6%.
            - ***IF*** MARITAL_STATUS = متزوج ***Then*** Evaluation = 2.0, that means that lecturer's evaluation is predicted to be between 75.6% and 80%.
            - ***IF*** MARITAL_STATUS = مطلق ***Then*** Evaluation = 2.0, that means that lecturer's evaluation is predicted to be between 75.6% and 80%.
        - ***IF*** Emp_Level_Bc in ["ممتاز" "مقبول"] ***Then*** Evaluation = 2.0, that means that lecturer's evaluation is predicted to be between 75.6% and 80%.
    - ***IF*** End_Date_Bc in ["2014"] ***Then*** Evaluation = 4.0 , that means that lecturer's evaluation is predicted to be greater than or equal 85.7%.

- ***IF*** Specialist_Detail_A_Name_Bc in ["تطوير البرمجيات"] ***Then*** :

- ▪ ***IF*** Emp_Level_Bc in ["جيد"] ***Then*** Evaluation = 1.0, that means that lecturer's evaluation is predicted to be less than or equal 75.5%.
- ▪ ***IF*** Emp_Level_Bc in ["جيد جدا"] ***Then*** Evaluation = 2.0, that means that lecturer's evaluation is predicted to be between 75.6% and 80%.
- ▪ ***IF*** Emp_Level_Bc in ["ممتاز" "مقبول"] ***Then*** Evaluation = 2.0, that means that lecturer's evaluation is predicted to be between 75.6% and 80%.

o ***IF*** Specialist_Detail_A_Name_Bc in ["تكنولوجيا المعلومات"] ***Then*** :

- ▪ ***IF*** Gender = ذكر ***Then*** :
  - ***IF*** MARITAL_STATUS = أعزب ***Then*** Evaluation = 4.0 , that means that lecturer's evaluation is predicted to be greater than or equal 85.7%.
  - ***IF*** MARITAL_STATUS = متزوج ***Then*** Evaluation = 2.0, that means that lecturer's evaluation is predicted to be between 75.6% and 80%.
  - ***IF*** MARITAL_STATUS = مطلق ***Then*** Evaluation = 4.0 , that means that lecturer's evaluation is predicted to be greater than or equal 85.7%.
- ▪ ***IF*** Gender = أنثى ***Then*** Evaluation = 3.0, that means that lecturer's evaluation is predicted to be between 80.1% and 85.6%.

o ***IF*** Specialist_Detail_A_Name_Bc in ["كهربائية" "صحافة وإعلام"] ***Then*** Evaluation = 3.0, that means that lecturer's evaluation is predicted to be between 80.1% and 85.6%.

o ***IF*** Specialist_Detail_A_Name_Bc in ["هندسة بيئية"] ***Then*** :

- ▪ ***IF*** Gender = ذكر ***Then*** Evaluation = 4.0 , that means that lecturer's evaluation is predicted to be greater than or equal 85.7%.
- ▪ ***IF*** Gender = أنثى ***Then*** Evaluation = 1.0, that means that lecturer's evaluation is predicted to be less than or equal 75.5%.

o ***IF*** Specialist_Detail_A_Name_Bc in ["هندسة حاسوب"] ***Then*** :

- ***IF*** Gender = ذكر ***Then*** Evaluation = 1.0, that means that lecturer's evaluation is predicted to be less than or equal 75.5%.
- ***IF*** Gender = أنثى ***Then*** Evaluation = 4.0 , that means that lecturer's evaluation is predicted to be greater than or equal 85.7%.

- ***IF*** Specialist_Detail_A_Name_Bc in ["هندسة صناعية"] ***Then*** :
  - ***IF*** Gender = ذكر ***Then*** Evaluation = 1.0, that means that lecturer's evaluation is predicted to be less than or equal 75.5%.
  - ***IF*** Gender = أنثى ***Then*** Evaluation = 3.0, that means that lecturer's evaluation is predicted to be between 80.1% and 85.6%.
- ***IF*** Specialist_Detail_A_Name_Bc in ["هندسة كهربائية"] ***Then*** :
  - ***IF*** Governorate_Name in ["رفح" "خـانيونس" "الوســـطى"] ***Then*** Evaluation = 1.0, that means that lecturer's evaluation is predicted to be less than or equal 75.5%.
  - ***IF*** Governorate_Name in ["شـمال غزة"] ***Then*** Evaluation = 2.0, that means that lecturer's evaluation is predicted to be between 75.6% and 80%.
  - ***IF*** Governorate_Name in ["غزة"] ***Then*** :
    - ***IF*** Birth_Date in Set6 ***Then*** Evaluation = 1.0, that means that lecturer's evaluation is predicted to be less than or equal 75.5%. Where Set6 is the set of items that listed in Table 8.6 .
    - ***IF*** Birth_Date in ["1990"] ***Then*** Evaluation = 4.0 , that means that lecturer's evaluation is predicted to be greater than or equal 85.7%.
    - ***IF*** Birth_Date in ["1991"] ***Then*** :
      - ***IF*** University_Name_Bc in ["المنصــورة"] ***Then*** Evaluation = 2.0, that means that lecturer's evaluation is predicted to be between 75.6% and 80%.

- ***IF*** University_Name_Bc in [" المنوفية" "كلية دبي الطبيــة"] ***Then*** Evaluation = 1.0, that means that lecturer's evaluation is predicted to be less than or equal 75.5%.

Table 8.1 Set1 in Rule 2

American	Near East	التحدي	أسيوط
Andhra	New Mexico	التكنولوجيا	أم القري
Aristotle	Nottingham	الجزائر	أم درمان
Arizona	Philipps	الخرطوم	بغداد
Arkansas	Poona	الزقازيق	بيت لحم
Bangalore	Saint-Petersburg state	الزيتونة	بير زيت
Berlin technical	Santo Tomas	السابع من إبريل	بيروت العربية
Colorado	Southwestern	السند	حلوان
Darisruhe	The East	الشلف	سبأ
Darmstadt	Toledo	العلوم التطبيقية	صنعاء
Delhi	Venice	الفاتح	طنطا
Donetsk Poly	Virginia Polytect	القاهرة	عبد المالك السعدي
Engineering & Technology	الإسراء	القدس المفتوحة	عمر المختار البيضاء
Exeter	الإسلامية العالمية	القرآن الكريم	قاريونس
Far Eastern	الإمارات العربية	الكليات العربية	قناة السويس
Guangxi	الإمام محمد	الكويت	كلية التربية(الأقصى)
Heritage College	الأردنية	الملك عبد العزيز	مالطة
Kazakh	الأزهر	النجاح الوطنية	محمد الأول
Middle East	الأمريكية	اليرموك	ناصر

Table 8.2 Set2 in Rule 2

اتصالات وتحكم	اللغة العربية والدراسات الإسلامية	تربية ابتدائية	علوم / تحاليل
الاتصال والعلاقات العامة	اللغة العربية وآدابها	تربية إنجليزي	علوم تربوية
الإدارة المالية والمصرفية	المحاسبة	تربية تاريخ	علوم زراعية
الإرشاد النفسي والتربوي	المناهج وطرق التدريس	تربية عربي	علوم سياسية
الإقتصاد والعلوم السياسية	النحو والصرف	تربية علوم	علوم سياسية ودبلوماسية
الأدب والنقد	النقد والبلاغة	تصميم معماري	علوم فيزياء
البلاغة العربية	الهندسة المعمارية	تعليم التكنولوجيا	علوم قانونية
التاريخ	الهندسة الميكانيكية	تعليم العلوم والتكنولوجيا	علوم كمبيوتر
التمريض العام	إحصاء	تكنولوجيا التعليم	فسيولوجي
الحقوق	إدارة أعمال	تكنولوجيا وعلوم تطبيقية	كمبيوتر
الحيوان والكيمياء	إقتصاد	جيولوجيا	كهرباء آلات

الخدمة الاجتماعية	إقتصادإسلامي	جيولوجيا المياه	كيمياء المبيدات
الشريعة	أحياء وكيمياء	رياضيات /تربية	ليسانس آداب وتربية إنجليزي
الشريعة الإسلامية	أدب إنجليزي	رياضيات/فيزياء	محاسبة وتدقيق
الشريعة والقانون	أنظمة المعلومات الحاسوبية	رياضيات/كمبيوتر	هندسة اتصالات
الطب العام	آداب - جغرافيا	صحافة	هندسة الطرق
الطب والجراحة	آداب لغة عربية	صيدلة	هندسة العوامل البشرية
الطبيعة	آداب وتربية	طب	هندسة إلكترونية
العلاج الطبيعي	بيولوجي	طب مخبري	هندسة تخطيط مدن
العلوم الطبيعية، الصيدلة	تاريخ فرعي وأثار	علاج طبيعي	هندسة تكرير بترول
العلوم والتربية – رياضيات	تخطيط حضري	علم الحيوان	هندسة كهربائية (اتصالات)
القانون	تخطيط عمراني	علم النفس	هندسة معمارية
القانون الخاص	تربية	علم إجتماع	
الكترونيات	تربية / تاريخ	علوم	
اللغة العربية	تربية /كيمياء	علوم-طبيعة	

Table 8.3 Set3 in Rule 2

إدارة أعمال	تربية	طب	هندسة بيئية
إقتصاد	تكنولوجيا معلومات	علم نفس	هندسة حاسوب
أحياء	تمريض	قانون	هندسة صناعية
أصول تربية	جغرافيا	كمبيوتر	هندسة كهربائية
أصول دين	حقوق	كيمياء	هندسة مدنية
بيئة وعلوم أرض	خدمة إجتماعية	لغة إنجليزية	هندسة معمارية
تاريخ	رياضيات	لغة عربية	هندسة ميكانيكية
تجارة	شريعة	محاسبة	
تحاليل طبية	صحافة	مناهج وطرق تدريس	

Table 8.4 Set4 in Rule 2

1966	1981	1992	2003
1968	1982	1993	2004
1971	1983	1994	2005
1972	1984	1995	2006
1973	1985	1996	2007
1975	1986	1997	2008
1976	1987	1998	2009
1977	1988	1999	2010
1978	1989	2000	2012
1979	1990	2001	2013
1980	1991	2002	

Table 8.5 Set5 in Rule 2

1966	1981	1992	2003
1968	1982	1993	2004
1971	1983	1994	2005
1972	1984	1995	2006
1973	1985	1996	2007
1975	1986	1997	2008
1976	1987	1998	2009
1977	1988	1999	2010
1978	1989	2000	2011
1979	1990	2001	2012
1980	1991	2002	

Table 8.6 Set6 in Rule 2

1945	1958	1970	1982
1946	1959	1971	1983
1948	1960	1972	1984
1949	1961	1973	1985
1950	1962	1974	1986
1951	1963	1975	1987
1952	1964	1976	1988
1953	1965	1977	1989
1954	1966	1978	1992
1955	1967	1979	
1956	1968	1980	
1957	1969	1981	