APPLYING MACHINE LEARNING TECHNIQUES FOR FUTURE RISING STARS PREDICTION IN ACADEMIA

BY

OMAR JAFER BIN OBAIDELLAH

A Thesis Presented to the
DEANSHIP OF GRADUATE STUDIES

KING FAHD UNIVERSITY OF PETROLEUM & MINERALS

DHAHRAN, SAUDI ARABIA

In Partial Fulfillment of the
Requirements for the Degree of

MASTER OF SCIENCE

In

COMPUTER SCIENCE

MAY 2017

KING FAHD UNIVERSITY OF PETROLEUM & MINERALS
DHAHRAN 31261, SAUDI ARABIA

DEANSHIP OF GRADUATE STUDIES

This thesis, written by **OMAR JAFER BIN OBAIDELLAH** under the direction of his thesis adviser and approved by his thesis committee, has been presented to and accepted by the Dean of Graduate Studies, in partial fulfillment of the requirements for the degree of **MASTER OF SCIENCE IN COMPUTER SCIENCE**.

Thesis Committee

Dr. Ashraf E. Fagih (Adviser)

Dr. Moataz Ahmed (Member)

Dr. Hosam K. Rowaihy (Member)

Dr. Khalid Abdullah Aljasser
Department Chairman

Dr. Salam A. Zummo
Dean of Graduate Studies

2/1/18
Date

To

My Mother
The warm cuddles and the legacy of [Compassion]

My Father
The first and wonderful teacher in my life

My Wife
The reason of my strength and the source of my [Faith]

My Othman and My Hadi
The spirit that I live by

My Family,,

ACKNOWLEDGMENTS

All thanks to ***ALLAH*** *Almighty, for giving me the strength, knowledge, ability and opportunity to undertake this study as well as to complete it satisfactorily.*

It is a genuine pleasure to acknowledge and express my deepest gratitude as well as immeasurable appreciation for the help ,support, ideas and guidance all along to My Advisor Dr. Ashraf E. Fagih.

I would also like to express my deeply indebted and grateful to the Committee Members, Dr. Moataz Ahmed and Dr. Hosam Rowaihy, who have been so helpful and cooperative in giving their support at all times to help me to achieve my goal.

I would also like to extend my thanks to all the professors and faculty members in Computer Science Dept. as well as Deanship of Scientific Research for their collaboration in accessing the data source.

My acknowledgement would be incomplete without thanking the biggest source of my strength, My Family, for their support and prayers during my academic journey.

I would also like to extend my thanks to all My Friends and Colleagues, who were nothing short of supportive, helpful and motivating.

Omar,,

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

THESIS ABSTRACT

NAME:	OMAR JAFER BIN OBAIDELLAH
TITLE OF STUDY:	APPLYING MACHINE LEARNING TECHNIQUES FOR FUTURE RISING STARS PREDICTION IN ACADEMIA
MAJOR FIELD:	COMPUTER SCIENCE
DATE OF DEGREE:	MAY 2017

Newly recruited researchers who are expected to outstandingly surpass their peers in the quality of their work, are often considered as substantial assets in universities and research and development sectors. Foreseeably identifying such rising stars (RSs) is vital for highly competitive and profitable institutes and organizations. In this work, we apply machine learning techniques to propose models based on a set of Scientometric Indicators to predict RSs in academia. In addition, we define the RSs problem in a comprehensive and methodological manner. Data subsets were collected from the Web of Science (WoS) data source and used to train, test and validate the proposed models. Our experimental results show that the proposed models and indicators can be used effectively in predicting future RSs.

ملخص الرسالة

الاسم الكامل: عمر جعفر ربيع بن عبيداالله

عنوان الرسالة: تطبيق تقنيات "التعلم الآلي" في التنبؤ بمستقبل "النجوم الصاعدة" من الباحثين في الجانب الأكاديمي

التخصص: علوم الحاسوب

تاريخ الدرجة العلمية: مايو 2017

الباحثون المبتدئون لمشوارهم البحثي، والذين يُتوقّع بمرور الزمن أن يتفوّقوا على نظرائهم في جودة أعمالهم البحثية، يشكلون حجر زاوية وعنصراً أساسياً في الجامعات وقطاعات البحث والتطوير العلمي، وتُطلق على أولئك الباحثين تسمية "النجوم الصاعدة". وفي الوقت الحاضر، فإنه من الأهمية بمكان ـ خصوصا في المؤسسات العلمية والأكاديمية ذات التنافسية العالية ـ أن يتم تطوير تقنيات لاستشراف ومعرفة احتمالية أي من الباحثين المُعيّنين حديثا مؤهل لأن يُصنّف ضمن تلك "النجوم الصاعدة" في المستقبل المنظور. لذلك، قمنا في هذا العمل بتطبيق تقنيات "التعلّم الآلي" لإنتاج "نماذج" قادرة على حل مشكلة التنبؤ بصعود أولئك النجوم خصوصاً في البيئة الأكاديمية. تلك النماذج تم بنائها بالاعتماد على مؤشرات متعلقة بـ السمات الكمية وخصائص العلوم والبحث العلمي والتي تمت صياغتها، هي الأخرى، واستخدامها للمرة الأولى في هذا العمل. وإضافة إلى إنتاجنا لتلك النماذج وصياغتنا لتلك المؤشرات، فإننا في هذا العمل قد اجتهدنا في تعريف المشكلة ودراستها بأسلوب منهجي وأسس رصينة ومنها اعتمادنا ـ في بناء تلك النماذج ـ على قواعد بيانات جُمعت من واحدة من أشهر وأشمل مصادر البيانات المحكّمة والمتعلقة بالنشاط البحثي الأكاديمي. النتائج التجريبية المتحصلة أثبتت فعالية تلك النماذج وقدرتها الفائقة على التنبؤ بالنجوم الصاعدة مستقبلا.

CHAPTER 1

INTRODUCTION

Scientific research, as an outcome of universities and research institutes, plays a vital role in communities growth. Hence, it is important to study the aspects related to the scientific research which aim at helping in the development of those communities. Authors of scientific researches (the terms ”authors” and ”researchers” will be used interchangeably throughout this text) are the cornerstones of that role with their cooperation with other scholars as well as their published academic articles. With the passage of time, and based on with their behavior, authors in the research community may be placed into four basic groups as depicted in Figure 1.1. The four author groups are: (1) well established authors including those who constantly produce a large amount of work over a time frame, (2) authors with a stable publication rate who produce fewer but regular publications, (3) declining authors whom declined in publications over the same time frame and (4) rising stars who show increased amount of publication as well as high impact of collaborations with other researchers [2].

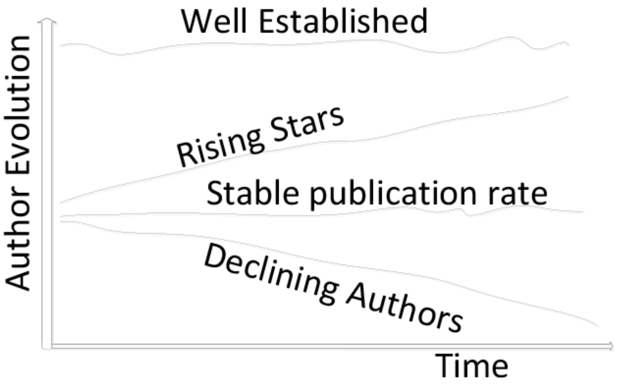

Figure 1.1: The four author's behavior over time (the figure is a verbatim copy of [2])

Although there is no uniform definition for the term "rising star authors", one can define them as "***those scientific research authors who, at present, may not surpass their peers, but have the potential to become prominent and influential contributors in the foreseen future***".

Identifying and evaluating tasks regarding rising star authors in universities and institutes considered a crucial and essential demand. Those tasks will help universities in hiring junior faculty members with the potential to become rising stars so to increase research productivity especially among young faculties, making them members of academic committees to drive benefit from their energetic inputs and outputs, selecting promising supervisors for graduate students, providing funding to qualified groups, grooming and recruiting them to be review conferences and journals, etc. Moreover, markets can also get benefit from identifying rising stars' process where it may be applied in hiring the best talents, appointing skillful directors and supervisors to the boards of companies and organizations, etc.

In this work, we shed light on rising star authors by defining the problem of rising stars in a comprehensive and methodological way as well as proposing models to predict who will be a rising star in the future for a specific region of a certain research community (Computer Science). Machine learning techniques, were adopted to build those models. An actual, trusted and comprehensive data datasets were collected form the Web of Science (WoS) data source to be used in training, testing and validating our models. A novel set of Scientometric indicators were formulated and calculated to be used as feature space in building those models. The experimental results were evaluated using variety of performance evaluation metrics. They showed that the proposed models can be used effectively in predicting future rising stars in academia.

1.1 Rising Star Authors and the Term Scientometrics

Pursuing efforts, evaluating the effectiveness and ranking the productivity of the researchers have been exposed more into the spotlights recently in researches related to what is referred by ***Scientometrics*** [4]. The term Scientometrics, its Russian equivalent "*naukometriya*", was coined in the 1960s by Vassily V.

Nalimov [5]. The term is mainly used for the study of the quantitative aspects of the literature of science and technology and used, for practical purposes, for citation analysis in academic literature. The main themes of Scientometrics are concerned with the ways of measuring the quality and impact of research, understanding citations processes, mapping scientific fields as well as the use of indicators in research policy and management [6].

The term Scientometrics has been, typically, defined as the "***quantitative study of science and technology***", nevertheless, there are other definitions give further insight into the use and clarification of the term. Tague-Sutcliffe [7] declared that the Scientometrics is "***the study of the quantitative aspects of science as a discipline or economic activity. It is part of the sociology of science and has application to science policy-making. It involves quantitative studies of scientific activities, including, among others, publication, and so overlaps bibliometrics to some extent***".

According to Chua et al. [8], within the discipline of Information Systems research, Scientometrics has focused primarily on two goals: (1) the assessment of journal prestige and (2) the assessment of researcher-production. However, researcher-production is also assessed in two ways, either by counting the publications of an author in a given basket of journals (number of articles published within a specific sample of journals) or through the citation analysis.

In this work, we are embracing the term Scientometric by focusing on evaluating the effectiveness of researchers in universities and institutes, together with, examining the Scientometric indicators related to how and where the researches have been published. The aim of that examination is to help in finding the future rising stars authors in Computer Science research community.

1.2 Machine Learning - The Techniques

In machine learning realm techniques are typically grouped into two broad categories, supervised and unsupervised. Supervised techniques known as those techniques of regression and classification while unsupervised techniques known as clustering techniques. Regarding the supervised techniques, there are two main reasons that urging to use them with data observations which include ***inference*** and ***prediction***. Regression is a statistical data analysis that attempts to estimate the strength of the relationship among variables (indicators) of data- *to achieve inference.* Classification is a form of data analysis identifying to which of a set of categories (classes) a new observation belongs- *to achieve prediction.*

While there are many possible regression and classification techniques which one might select to apply with data observations, the selection process is depending on the nature of the targeted problems as well as whether the intended purposes is for inference or prediction. In this work, Multiple Linear Regression

(MLR) [9] in addition to four of the most widely-used classifiers (Naive Bayes (NB)[10], k Nearest Neighbors (kNN) [11], Neural Networks (NN) [12] and Support Vector Machine (SVM)) [13] were applied for proposing the models. The results showed that all those models were performed well in inferencing and predicting authors who will be rising stars in the foreseen future.

1.3 Web of Science - The Trusted Dataset

A comprehensive and accurate source of dataset is required for the successfulness of any work related to the analysis of citations and research, thus in this work we used the WoS as data source for the collected datasets. The WoS [14] is an online subscription-based scientific citation dataset maintained by the Thomson Reuters. The WoS provides the characteristics that required in such a data resource to be used for citation and research related applications with a coverage of around 12,000 of the world's most important and influential journals.

1.4 Scientometric Indicators

To accomplish the goal of our work in finding future rising stars, we are in need to acquire an appropriate feature space to be used along with the already nominated machine learning techniques. Eight indicators were formulated and computed from the collected datasets to be used as the mean feature space for each author in addition to other six supplementary indicators.

The Eight indicators are: (1) *Author Productivity Level*, (2) *Scientific Impact Level*, (3) *Value of Productivity*, (4) *Citations Per Publication*, (5) *Contribution Impact*,(6) *International Collaboration*, (7) *Research Area Relevancy* and (8) *Venue Reputation*. The aforementioned indicators were used as predictor variables against *Rising Star Status* (the response variable).

1.5 Problem Statement and Research Questions

Universities and institutes are in need to know the status and the standing of their researchers in future with the aim of giving concentration and close attention to those may have potentials to become prominent and influential contributors in future- future rising stars.

That motivates us to propose machine learning-based models to predict who will be in the foreseen future, next year, a rising star from those who are presently may has low research profile and not outstanding among peers. That can be achieved by investigating his historical data of last five years through benefiting from the related Scientometric indicators.

The following research questions have been articulated in order to help in coming up with the proposed models:

- **Q1:** Are the future rising stars predictable?

- **Q2:** How accurately can we predict future rising stars?

- **Q3:** Is there any significant difference in obtained results between the applied machine learning techniques (classifiers) in predicting rising stars?

- **Q4:** Is there any significant difference in obtained results between the datasets used in predicting rising stars?

- **Q5:** For the obtained results, is there any significant difference in the interaction between the applied machine learning techniques (classifiers) and the used datasets in predicting rising stars?

- **Q6:** Up to which level are we confident that the proposed models are able to predict the future rising stars when they are used with alternative datasets?

- **Q7:** What are the most important Scientometric indicators in predicting future rising stars?

- **Q8:** How are our proposed models comparable to other models?

Where the key objective of this work is to find answers for the over mentioned questions.

1.6 Research Contributions

While that this work is being one of the few works tackling the problem of finding rising stars in academia, it is further engaging in adding the following contributions:

1. Coming up with simple and reliable models by applying four form the most famous and efficient in performance machine learning techniques (NB, kNN, NN and SVM classifiers) not applied in previous works related to the rising stars problem. That models can be used easily by universities and institutes to predict their rising stars in specific research area (in this work Computer Science).

2. Defining the rising stars problem in a relatively comprehensive and methodological way by insuring and affirming that the researchers are likely to become rising stars in the future not because they were already identified to be so in previous years (excluding the already rising stars in the last year). Moreover, and for experimental purposes, selecting the first 30% of researchers to be labeled as rising stars in contrast with 50% in previous works.

3. Formulating and computing a novel set of eight Scientometric indicators to observe the evolution of each researcher from multiple perspectives. That set of indicators' combination was not addressed in previous works for future rising stars' prediction.

4. Building benchmark datasets for four academic organization (KFUPM, IIT, MIT and CAS) based on a trusted, comprehensive and accurate data source (the WoS) which can be publicly available to the research community.

5. Identifying and evaluating rising stars with a relatively robust and reliable

ground truth criterion by adopting **InCites** author's rank indicator as a benchmark.

6. Using wide variety of popular metrics and statistic methods to analyze and evaluate the efficiency and performance of the proposed models (Confusion Matrix, Prediction Accuracy Level, Sensitivity, Specificity, F1 Score, ROC, AUC, F-statistic and ANOVA).

1.7 Thesis Organizations

The thesis is organized as follows. In Chapter 2 we summarize the literature review of rising stars problem. Machine learning techniques and the Web of Science data source were briefly addressed in Chapter 3. Chapter 4 presents the process of data collection, indicators formulation, and research methodology. Chapter 5 is dedicated to the experiments and results discussion. Finally, the thesis is concluded with future work in Chapter 6.

CHAPTER 2

LITERATURE REVIEW

There is a little work done in the research community related to the rising stars problem. Most of that work was about exploring the factors and features (Scientometric indicators) which finally resulted in finding rising stars. Additionally, that work was focusing on proposing methodologies and approaches based on those explored features to find rising stars.

One of the most referred works in that field is that of Li et al. [15]. They proposed an algorithm named PubRank, influenced by Google's PageRank [16], to find the rising stars through the mutual influence and the static ranking of publication venues. They built their algorithm based on three indicators (1) Researchers mutual influence in terms of co-authorships in the bibliography network. They used the link weighting strategy to model the degree of the mutual influence where a junior researcher who influences the work of his seniors and leverages on his expertise is more likely to have the chance to succeed in future and be an influential researcher- become a star, (2) Venue's weight by placing different weights (ranks) on different venues and keeping a track record of each researcher in terms of averaging the quality of his current publications in different prestige of publication venue and (3) Researchers collaboration by tracing the chronological changes in the networks using a way similar to that in Google's PageRank. Where a researcher who can build up a strong collaborative with different groups of people at different points in time is more likely to become a rising star. Li et al. used the linear regression technique in proposing a model to compute the PubRank score every year for each researcher during a certain time period. According to that work, a researcher is considered rising star if his/ her z-score is greater than zero (larger than the average score). They conducted their experiments to train and evaluate the performance of the proposed algorithm by using publication data gathered from the Digital Bibliography and Library Project (DBLP) dataset for the periods 1950-1955, 1990-1995 and 2002-2007.

In [2], Tsatsaronis et al. attempted to learn basic researchers' profiles and properties to know who will become a leader for a group of researchers. They used the bisecting K-Means clustering technique to cluster authors, based on their evolution over time, into four different groups: (1) Well established, (2) Rising stars, (3) Authors with a stable publication rate, and (4) Declining authors. Tsatsaronis et al. applied the Power Graphs Analysis onto co-authorship graphs extracted from bibliographical databases in order to define authors' features witch are (1) Number of the most frequent co-authors an author has at time point, (2) Number of the co-authors of her co-authors, (3) Number of author's publications and (4) The total impact of an author's work (venue impact). The data that used in their experimental work was selected from the DBLP Computer Science Bibliography database within the time frame of 2000-2010 for researchers who have in total a minimum of 5 publications by 2010.

Daud et al. in [17] criticized the work in [15] and declared that the PubRank algorithm suffers from two main blemishes (1) The PubRank in its mutual influence did not take into account a weighting for author contributions- the order in which the authors appear in the paper. (2) PubRank uses a static ranking for publication venues where their ranks may change by time. Consequently, to overcome those limitations they proposed the StarRank algorithm. They overcome the first limitation of the PubRank by adopting the Sekercioglu approach [18] which gives a weight of co-author contribution based on the order that authors appear in the papers where the first author will get the maximum weighting. Additionally, to overcome the second limitation, they used the concept of entropy to dynamically calculate the rank of publication venues. Thus, a venue is considered a high level if it has less entropy, hence, it will be less entropy if it only accepts papers on specific research areas- has a limited title of topics for papers to be accepted. The linear regression and the psychometric test method were used to predict and find the rising stars based on the aforementioned indicators (Author Contribution and Dynamic Publication Venue Rank) by using a dataset collected from the DBLP for the period from 1996 to 2000. The performance of the experiment results was then verified by looking to the authors standing in 2012.

The problem of rising stars was addressed by using classification techniques in Daud et al. in [1]. Four classification techniques were used to predict the future rising stars where two of them are generative techniques (Bayes Network (BN) and Naive Bayes (NB)) and the other two are discriminating (Maximum Entropy Markov Model (MEMM) and Classification and Regression Tree (CART)). The outcome of the classifiers was as a binary classification result which answers the question "*Does a scholar have a potential to become a future rising star (positive value) or not (negative value)?*". Eleven indicators were extensively discussed in that work. From those eleven, four are based on the author: (1) Author Influence, (2) Author Contribution, (3) Author Contribution Based Mutual Influence and (4) Temporal Dimension; four are based on venue specificity: (5) Venue Citations, (6) Venue Specificity Score and (7) Venue Count; and the reaming are based on co-authors relationship: (8) Co-author Citations, (9) Co-author of papers, (10) Co-author Count and finally (11) Co-author venue score. They trained the four techniques on two datasets collected from the DBLP and the ArnetMiner database for a wide discipline during a time window 1995-2000. The author citation information and the Average Relative Increase in Citations (ARIC) were used as ground truth in order to assign the label of classes- the top 50% highly cited authors were labeled as rising stars (positive 1) and other 50% authors with lowest citations were labeled as not rising stars (negative 1). The efficiency of the four classification models was compared and analyzed in terms of F1 scores.

Zhang et al. [19] proposed a method named CocaRank. The CocaRank name was integrated from their proposed indicator named "*COllaboration CAliber*". That indicator was considered as a typical citation indicator to capture the researchers' abilities to collaborate with others based on the concept of entropy- the larger value of entropy the richer ability to collaborate. They proposed that method to overcome the problem they claimed about the traditional citation-based methods. Four indicators were used in that work: (1) Collaboration Caliber- their proposed indicator, (2) Paper Citation Network- which measured by PageRank algorithm [4], (3) Author-Paper Network- which measured by HITS algorithm [20], and (4) Paper-Journal Network- which also measured by HITS algorithm. They adopted "Citation Count" as a measure and ground truth to validate whether a researcher has achieved the expectations to be a rising star or not, then they compared the values with the calculated Spearman Correlation Coefficients. The used dataset was acquired from the American Physical Society (APS) which contains detailed information of each article from 12 physical journals for researchers who published their first articles in 1993, and their academic career is not due to end until 2013.

By using the Multi-Objective Optimization Approach [21] and the Rank Aggregation Approach [22], Wijegunawardana et al. [23] investigated the problem of finding rising stars in heterogeneous social networks. They modeled the problem as a decision level of data fusion problem and addressed it based on three indicators: (1) Co-authorship, (2) Author Citation and (3) Publication Venue. A dataset collected from the ArnetMiner for the time-period of 1990-1995 was used to generate the values of those three indicators. The top 3% of the ranked authors in the citation and co-author networks in 1990 were removed to ensure that the authors are not already stars in the network, but may have the potential to become in the near future. The obtained results were evaluated against authors' ranks in 2006 and 2014 based on three factors include the h-index value, the total number of papers and the citation count.

Although the aforementioned works have spent lots of remarkable efforts in tackling the problem of finding rising stars, they have still suffered from some flaws and concerns. By contrast, our work scores over the previous works in terms of:

- Defining the rising stars problem in a relatively comprehensive and methodological way:
    - Insuring and affirming that the researchers are likely to become rising stars in the future not because they were already identified to be so in

  previous years (excluding the already rising stars in the last year).
  - Selecting the first 30% of the researchers to be labeled as rising stars in contrast with 50% in previous works.

- Formulating and computing a novel set of eight Scientometric indicators to observe the evolution of each researcher from multiple perspectives. That set of indicators' combination was not addressed in previous works for future rising stars' prediction:
  - Considering the importance of the author contribution-oriented and the order in which co-authors appear in a publication, meanwhile, dealing with the situation where only one author presents on a paper.
  - Considering the value of the author international collaboration and the research area relevancy in assessing the researcher activity and productivity.

- Identifying and evaluating rising stars with a relatively robust and reliable ground truth criterion by adopting **InCites** author's rank indicator as a benchmark.

Table 2.1 summarizes and compares the previous works on the aspects of the contributions to the problem of rising stars, the applied techniques, the proposed indicators, the source and time-period of the used dataset, the benchmark used in evaluating the proposed solutions and finally the main concerns regarding the approaches used in those works.

Table 2.1: A Summary and Comparison of the Related Works

Related Work	**Main Contribution**	**Applied Technique**	**Used Dataset**	**Indicators**	**Evaluation Benchmark**	**Main Concerns**
Li et al. [15]	Proposing the PubRank algorithm	Linear Regression	DBLP (Time periods: 1990-1995, 1950-1955 and 2002-2007)	(1)Researchers mutual influence, (2) Venues weight and (3) Researchers collaboration	Author's citation count	(1) Missing weighting of authors contributions and (2) Using a static ranking for publication venues
Tsatsaronis et al. [2]	Clustering authors into four different groups	Bisecting K-Means Clustering	DBLP Computer Science (Time period: 2000-2010)	(1) The most frequent co-authors, (2) The co-authors of authors co-authors, (3) Author's publications and (4) Venue impact	Power Graphs Analysis	Lacking of a robust ground truth for approach evaluation
Daud et al. [17]	Proposing the StarRank algorithm to overcome the limitations of the PubRank	Linear Regression	DBLP (Time period: 1996-2000)	(1) Author Contribution and (2) Dynamic Publication Venue Rank using entropy	Author's citation count	(1) Few indicators are not enough to explain author activity (multiple perspectives required) and (2) Using entropy is not workable to evaluate venue impact
Daud et al. [1]	Predicting future rising stars using classification techniques	Classifications (BN, NB, MEMM and CART)	DBLP and ArnetMiner (Time period: 1995-2000)	(1) Author Influence, (2) Author Contribution, (3) Author Contribution Based Mutual Influence, (4) Temporal Dimension, (5) Venue Citations, (6) Venue Specificity Score, (7) Venue Count, (8) Co-author Citations, (9) Co-author of papers, (10) Co-author Count and (11) Co-author venue score	(1) Author's citation count and (2) Average relative increase in citations (ARIC)	(1) Applying the approaches on widely discipline of research community and (2) No affirmation that a predicted rising star was not a rising star in past
Zhang et al. [19]	Proposing the CocaRank algorithm	Spearman correlation	APS (For researchers with first articles in 1993, and not cut off until 2013)	(1) Collaboration Caliber, (2) Paper Citation-Network, (3) Author-PaperNetwork and (4) Paper-Journal Network	Author's citation count	Using entropy, by calculating the number of words in all the institutions' information, is not a workable measure for the ability to collaborate
Wijegunawardana et al. [23]	Investigating the problem of finding rising stars when heterogeneous data sources are available	Multi-Objective Optimization and Rank Aggregation	ArnetMiner for (Time period: 1990-1995)	(1) Co-authorship, (2) Author Citation and (3) Publication Venue	Authors citation count, (2) h-index value and (3) Total number of papers	Few indicators are not enough to explain author activity (multiple perspectives required)

CHAPTER 3

PRELIMINARIES IN BRIEF

In this chapter, a brief background is presented to familiarize the reader with aspects related to machine learning techniques and the WoS dataset that will be frequently referred to in the forthcoming chapters.

3.1 Machine Learning Techniques

Machine learning techniques have been used frequently in Computer Science and Engineering fields to explore the study and the construction of algorithms that can learn from selected data to make predictions on it [24]. They are about learning some properties of a portion of data which is at hand (training data) and using the other portion of that data (testing data) to test the learned properties and propose a model based on some selected algorithms, where the main idea behind machine learning techniques is that of the selected algorithms.

Machine learning techniques can be classified as supervised or unsupervised techniques. The supervised techniques are basically a synonym for regression and classification where those techniques can be used to predict or infer solutions (outputs) based on a set of inputs (feature space for observations). The unsupervised techniques are essentially a synonym for clustering where those techniques can be used to find and cluster the structure or the relationships between different inputs while there is no supervising output. In other words, clustering is the assignment of a set of observations into subsets (called clusters) so that observations within the same cluster are similar according to some predesignated criterion or criteria, while observations drawn from different clusters are dissimilar [25].

The most mature and widely used techniques is the supervised ones, precisely classification. Classification defined as the problem of identifying to which of a set of categories a new observation belongs based on a training set of data with observations whose category membership is known. In other words, classification is the form of data analysis that extracts models describing important data classes. Such models, called classifiers, *predict* categorical class labels. Hence, a classifier can be defined as a system that inputs a vector of categorical and/or numerical feature values (for observations) and outputs a single discrete value (solution).

A plethora of classification algorithms has been proposed by researchers where, literally, hundreds are available, and tens more are published each year with an extensive use in wide variety of applications. Consequently, the main challenge that someone may face is that of the bewildering variety of algorithms available and which one to be selected? Hence, the objective is to select the classier ones that will more probably achieve the best performance with any dataset [26].

Besides classification, regression techniques as supervised techniques are helpful when the *inference* is the goal by which the relationship among variables (indicators) of data can be estimated. Multiple Linear Regression (MLR) considered a time-honored technique proposed by Karl Pearson, an English mathematician and biostatistician, in 1908. It is about learning more about the relationship between a dependent variable Y (the predicted value) and two or more independent variables $X_1, X_2, ..., X_p$ (the predictors, where p is the number of predictors- in our case the Scientometric indicators). [9]

In the recent years, many applications and problems have adopted the supervised techniques for their proposed solutions. For example, in problems regarding disease diagnosis and cancer outcome prediction [27], in the future in electricity and energy consumption problems [28], in petroleum and oil spills [29], wireless sensor networks applications [30], in applications of operating systems process management [31], etc.

In this work, we applied MLR in addition to four machine learning classifiers picked up from the most widely-used and best performance bracket. The applied classifiers are Naive Bayes (NB), k Nearest Neighbors (kNN), Neural Networks (NN) and Support Vector Machine (SVM). In the following lines we will introduce those applied classifiers, in brief.

3.1.1 Naive Bayes (NB)

Naive Bayes (NB) is a classification algorithms depending on applying Bayes' theorem [32] with the "naive" assumption of the independency between every pair of indicators (features). The NB assumes that the presence of a particular indicator in a class is unrelated to the presence of any other . For example, given our problem of finding future rising stars as a two-class problem where there are only two possible response values for variable y, say class 1 (is a rising star) or class 0 (is not a rising star) based on, for simplicity, a dependent feature vector X consists of indicators X_1 and X_2, the NB classifier can be formulated as following $P_r = (y = 1| X)$. The formula reflects the probability that a researcher will be a rising star (class 1) given that we know the values of his indicators X_1 and X_2.

Figure 3.1, depicts a simulated data set consisting of 100 researchers' observations, with indicators X_1 and X_2 (feature space), in each may belong to one of the two classes to be a rising star or not to be, indicated in blue and in orange regions respectively. The purple dashed line represents the *Bayes Decision Boundary*- the region of the problem space in which the output class is ambiguous. The orange background grid indicates the region in which a test observation will be assigned to the rising star class, and the blue background grid indicates the region in which a test observation will be assigned to not rising star class.

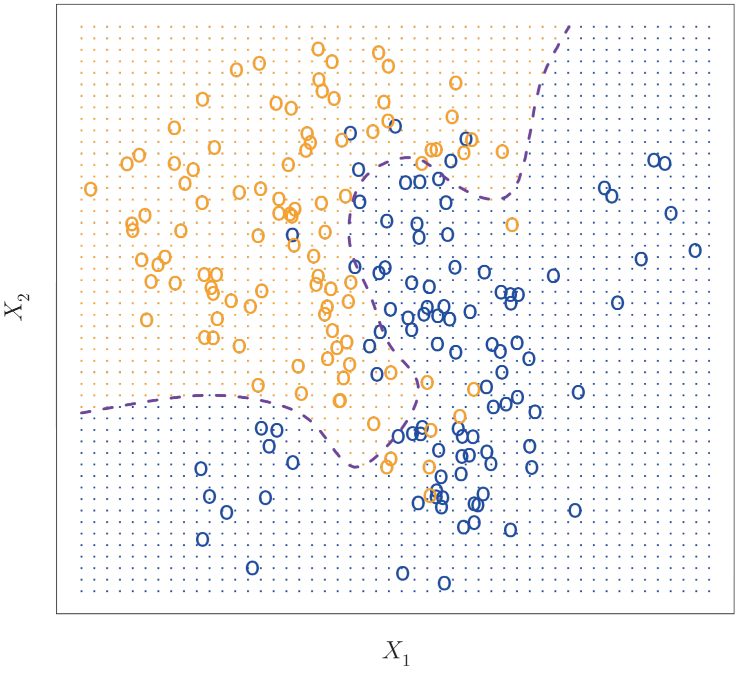

Figure 3.1: A simulated data set consisting of 100 researchers' observations, with indicators X_1 and X_2, in each may belong to one of two classes to be a rising star or not to be, indicated in blue and in orange regions respectively (the figure is a verbatim copy of [3])

NB is easy to build and can be extremely fast compared to the most sophisticated classification algorithms. Along with that, NB is particularly useful and have exhibited high accuracy even when applied to very large datasets [33].

3.1.2 k-Nearest Neighbors (kNN)

k-Nearest Neighbor (kNN) is a non-parametric learning algorithm classifies a data sample (inputs/ observations) based on the labels (output values/ solutions) of the near data samples [34]. Non-parametric means that kNN does not make any assumptions on the underlying data distribution where in the real world it is rare for the practical data to obey the typical theoretical assumptions for data distribution- normal (Gaussian) distribution. Non-parametric algorithms like kNN come to the rescue here.

For example, a researcher (input data point) can be predicted as to be a rising star in the future or not using the average *Measurement of neighboring*, the already trained (data points) researchers within specific diameter limits. Measurement of neighboring implies a distance metric which by default is the Euclidean distance-distance between different data points. kNN does not need to high computational power, as the function is computed relative to local points (i.e., k-nearest points, where k is a small positive integer indicated the k observations that are nearest to a given input test data point). For clarification, in our problem, assume that we have a training set consists of two classes: it is a rising star (RS) and it is not a rising star (Not RS) with five instances apiece, as indicated by blue and orange dots. Only two indicators (X_1 and X_2, 2-dimensional feature space) are used to discriminate between the classes. Now, we are presented with an unlabeled observation (gray input data point) and would like to classify it as either RS or Not RS. To accomplish that, we have to (1) Calculate the distance between the gray point and all other points, (2) Find the nearest neighbors by ranking points by increasing distance. The nearest neighbors (NNs) of the gray point are the ones close in the data space, finally, (3) Vote on the predicted class labels based on the classes of the k nearest neighbors. For the case $k = 3$ (small circle), one neighbor is of Class RS and two are of Not RS, so we classify the gray data point as a member of Not RS; for $k = 6$ (large circle), however, four neighbors are of Class RS and only two are of Class Not RS, so the unlabeled the gray data point is instead classified as a member of RS. Figure 3.2 depicts the example.

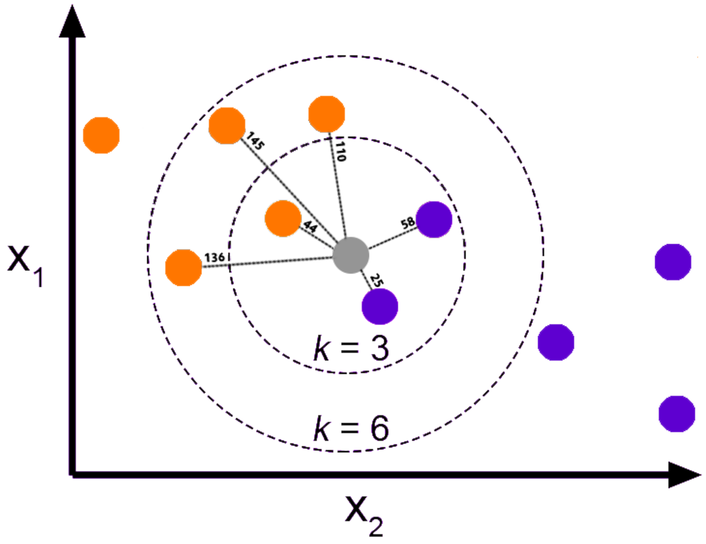

Figure 3.2: A simulated trained data points consisting of 10 researchers' observations, with indicators X_1 and X_2 and two class labels to be a rising star or not to be, indicated in blue and in orange regions respectively. Using kNN to predict to which class a gray data point may belong by considering the value of k and the distance between the gray point and all the other points

kNN is one of the effortless algorithms to understand and easy to implement.

Despite its simplicity, it performs well in many situations and considered one of the top 10 algorithms in data mining [35].

3.1.3 Neural Networks (NN)

Neural Network (NN) is a powerful classification algorithm used in the modern intelligent systems [36]. NN learns a function $f(.) : R^i \rightarrow R^o$ by training on a dataset, where i is the number of dimensions for input (features/ indicators $X = x_1, x_2, ..., x_i$) and o is the number of dimensions for output (solutions/ classes). Artificial Neural Networks (ANN) or Multi-Layer Perceptrons (MLP) are two terms used interchangeably with Neural Network [37].

The NN consists of interconnected units (called neurons) in an acyclic graph and arranged in layers which work to convert an input vector (feature space) into some output. In the learning phase, each unit takes an input, applies an *Activation Function* to it and then passes the output on to one or more *hidden layers* where the number of hidden layers is arbitrary, although in practice, usually only one is used. Finally, the last hidden layer then links to an output layer and produce the solution as output which emits the network's prediction for given observations. An activation function has two parts: (1) the combination function that merges all the inputs into a single value- weighted sum; and (2) the transfer function, which transfers the value of the combination function to the output value of the unit- non-linear function.

For example, suppose the problem of finding rising stars consists of m observations, 2-dimensional feature space (X_1 and $X_2, i = 2$) and 2-dimensional output classes (RS and Not RS, $o = 2$) in Figure 3.3. The input layer accepts the input feature space (X_1 and X_2) through two neurons which then applies the activation function to it and then passes the output to the next layer (hidden layer). Each neuron in the hidden layer transforms the values from the previous layer, again with the activation function which consists from weighted linear summation and a non-linear activation function, to the output layer. The output layer, finally, receives the values from the (last) hidden layer and transforms them into one of the output class values (either a RS or Not RS).

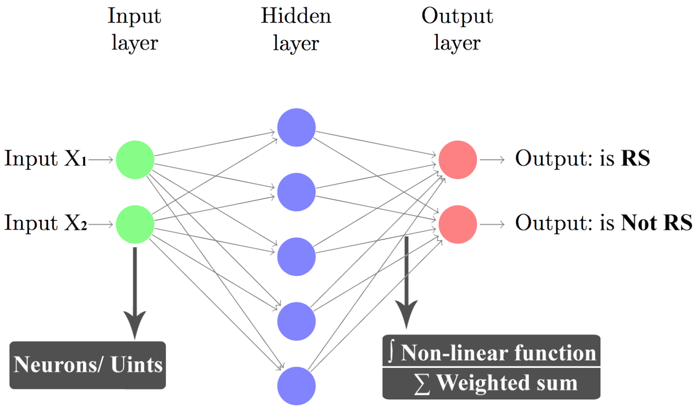

Figure 3.3: A two-layer neural network consists of 2-dimensional feature space (X_1 and X_2) and 2-dimensional output classes (RS and Not RS) with one hidden layer (when it is said N-layer neural network, the input layer is not counted)

3.1.4 Support Vector Machine (SVM)

Support Vector Machine (SVM) is one of the newest classification algorithms that was developed in the 1990s where the first paper on it was presented in 1992 by Vladimir [38] and it has grown in popularity since then. The SVM offers one of the most effective, robust, versatile and accurate methods among all well-known other classification algorithms. Additionally, it is much less prone to *overfitting* problem than other techniques- where the number of feature space dimensions is greater than the number of observations.

The SVM is based on the concept of the *decision planes* that define the decision boundaries. For observations of a training data, it searches for the linear separating hyperplane or decision boundary that separates the observations of one class membership from another. Because there are an infinite number of such separating linear hyperplanes, so the aim is to find the optimal one, that is, one that will have the minimum classification error on testing observations. Technically speaking, the optimal separating hyperplane (also known as the maximal margin hyperplane) is the farthest from the training observations. Geometrically, the margin corresponds to the shortest distance between the closest data points to a point on the hyperplane. Having this geometric definition allows us to explore how to maximize the margin, so that even though there are an uncountable number of hyperplanes, only a few will qualify as the solution to the SVM.

Figure 3.4 depicts an example of our problem to find rising stars by applying SVM algorithm. In the left, the 2-dimensional classes, shown in blue and in orange (to be RS or Not RS, respectively), each of which has measurements on 2-dimensional indicators, X_1 and X_2. Three separating hyperplanes, out of many possible, are shown in black where the aim is to find the optimal one. In the right, the optimal with maximal margin hyperplane is shown as a solid line. The margin is the distance from the solid line to either of the dashed lines. The two blue points and the orange point that lie on the dashed lines are the support vectors, and the distance from those points to the hyperplane is indicated by arrows. The blue and orange grids indicate the decision rule made by the classifier based on this separating hyperplane. In other words, a test observation that falls in the blue portion of the grid will be assigned to the blue class (is RS), and a test observation that falls into the orange portion of the grid will be assigned to the orange class (is Not RS).

The good news about the aforementioned algorithms, and the others, is that all are considered as "Black Boxes" so there is no need to know about their details. All what we have to do is to come with the data observations, accordingly, choose and apply the proper one based on some factors as discussed in [26], [39], [37], [28] and [27].

It is worth noting that there are several popular available toolkits such as Weka

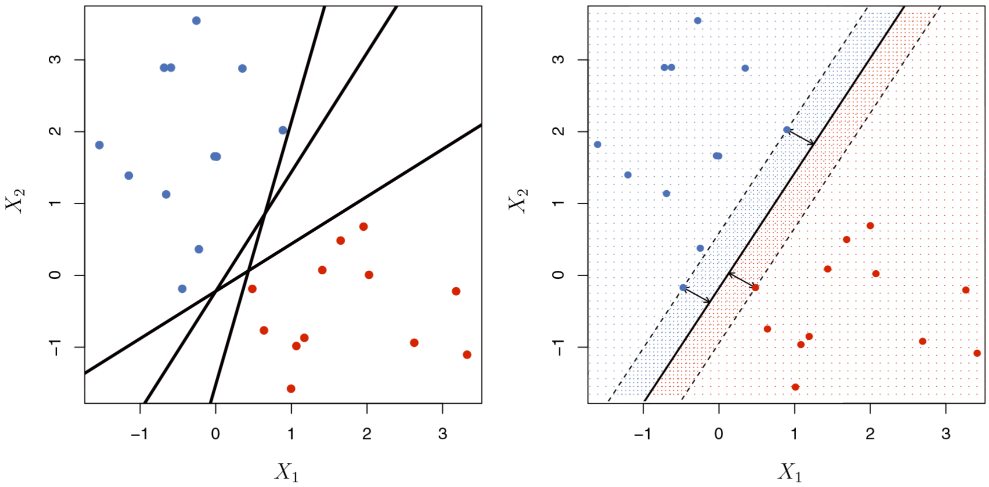

Figure 3.4: Left: A 2-dimensional classes, shown in blue and in orange, each of which has measurements on 2-dimensional indicators, X_1 and X_2. In addition to three separating hyperplanes, out of many possible, are shown in black. Right: The optimal hyperplane is shown as a solid line. The blue and orange grid indicates the decision rule made by a classifier based on this separating hyperplane

[40], Matlab [41] and R [42] where someone can benefit from them in implementing the required techniques and algorithms. In this work, R was chosen for its astonishing power with machine learning, running statistical tests, summarizing obtained results and high capabilities in visualizing data.

3.2 The Web of Science and the Top 3 of Citation Databases

In such works that focus on measuring the scholarly and researchers output, as well as the impact of the social sciences, assessing and selecting the suitable data is an issue of growing importance. While there are different data sources available in the community, the appropriate one can be selected by evaluating the coverage of the social sciences and by assessing the quality of the content.

In the 1950s, Eugene Garfield [43] established the company the ***Institute for Scientific Information*** (ISI) with the aim of establishing the Science Citation Index (SCI) to capture publication citations. The main purpose of this idea is to help researchers to search the literature more effectively [6]. In 1992, the Thomson Corporation acquired the ISI and converted the SCI database into the Web of Science (WoS) [14] as part of their Web of Knowledge platform. Until the fall of 2004, the Thomson Reuters ISI WoS was the major source of citations. It was a specialized data source which provides comprehensive coverage of around 12,000 of the world's most important and influential journals. Such comprehensive and accurate sources of citation are required for the successfulness of the quantitative analysis of citations.

In the fall of 2004, the citation data sources at once was changed considerably by the appearance of Elsevier's Scopus on November 3, 2004 [44]. Scopus database provides consistent full citation coverage from 1996 and onwards. Furthermore, it claims to be "*the largest abstract and citation database of research literature and select results from the web*". Scopus covers 27 million abstracts, 230 million references and 200 million web pages [45]. It is important to notice that Scopus provides citation data only for the items indexed by it.

In the same year as the appearance of Scopus, on November 18, 2004, Google Scholar was launched as the most interesting challenges for the WoS and Scopus [46]. Google Scholar as citation database, works in an entirely different way unlike the others, it searches the web for documents that have references to papers and books rather than collecting data inputted directly from journals.

Several studies have been conducted which have compared and evaluated these three citation data sources remembering the fact that the WoS and Scopus are subscription-based resources while Google Scholar is a free alternative [47]. Assuming that the WoS is accepted as the top one, Norris and Oppenheim [48] presented a study on several alternatives to the WoS for coverage of the social sciences. The study compared databases included Scopus and Google Scholar. They found that Scopus was comparable to the WoS in terms of coverage from 1996 and onwards, but Google Scholar cannot be seriously thought of as a database from which metrics could be used to measure scholarly activity. In contrast, Bauer et al. [49] concluded in their study for analyzing the citation counts of JASIST articles published in 1985 and 2000: "*Based on our preliminary examination and discovery of higher citation counts, we recommend that researchers should consult Google Scholar in addition to Web of Science or Scopus, especially for a relatively recent article, author or subject area*". Accordingly, the WoS excels other data sources in regarding of the coverage and accuracy

In this work, because of a comprehensive and accurate source of a dataset is crucial for the success of work related to the analysis of citations and research, the WoS was nominated as the best choice. Furthermore, its customized and citation-based research evaluation tool, InCite [50], was used as a benchmark and ground truth criterion.

CHAPTER 4

METHODOLOGY

The proposed methodology sheds light on the problem of predicting future rising star authors in a specific region of a certain research community which in this work Computer Science. MLR and four classifiers were adopted to build the inference and perdition models on datasets collected from the WoS. Eight Scientometric indicators were formulated and calculated to be used as a feature space in the process of building models.

The problem definition, the task of collecting and preprocessing the datasets, the formulation and computation of the Scientometric indicators as well as the articulation of the research questions will be detailed and discussed in this chapter.

4.1 Problem Definition

The objective of this work is to find whether a researcher will be a rising star in the future (after one year) by looking to his historical data for research productivity within a specific time-period (5 years before). The problem definition can be stated as following:

> Given n training observations of historical data in specific time period (5 years) $(X_1, y_1), (X_2, y_2)...(X_n, y_n)$, where n is the total number of researchers, X_i is a feature vector (Scientometric indicators) of a researcher r_i, where $X_i \in R^m, m$ is total number of features (indicators) and $y_i \in \{0, 1\}$.
>
> To classify whether a researcher r_i is a rising star or not, the prediction function is defined as:
>
> $$y = IS_{RS}\left(r \mid X\right) \tag{4.1}$$

which will be learned from a training dataset. Where:

$$IS_{RS}\left(r \mid X\right) = \begin{cases} 0, & \text{if } y = 0 \quad \text{not RS} \\ \\ 1, & \text{if } y = 1 \quad \text{not RS} \end{cases} \tag{4.2}$$

The learning task was summarized in $\widehat{IS}_{RS}\left(.\right)$, the predictive function, to predict whether a researcher r is a rising star or not after a given time-period $\triangle t$

(on year). Formally we have:

$$\widehat{y} = \widehat{IS}_{RS}\left(r \mid X, \triangle t\right) \tag{4.3}$$

4.2 Data Collection

A well-known indexing service managed by Thomson Reuters, the WoS, was used as a source of our datasets. The WoS is considered one of the best indexing services as well as trusted and comprehensive data source. Our study was specified on publications conducted in the Computer Science research area as defined in WoS schema (Software Engineering, Theory & Methods, Cybernetics, Information Systems, Artificial Intelligence, Hardware & Architecture and Interdisciplinary Applications). To fulfill the requirement of data generalization and balancing, mainly, four datasets were collected and built in this work for four universities and institutes: (1) King Fahd University of Petroleum & Minerals- KFUPM, (2) Indian Institute of Technology- IIT, (3) Massachusetts Institute of Technology- MIT and (4) Chinese Academy of Sciences- CAS. These organizations have been chosen as proof-of-concept for our methodology and for their notable research outcomes locally and regionally, respectively. The collected data were for the time-period 2011-2015 with the WoS last update in 2016-12-07.

The data collection process lies in the following steps:

Step 1: For each of the aforementioned universities and institutes, the process of collecting data is started by querying and downloading data of authors for the time period 2011-2016. The collected data was stored with the name (**AuthDS2016**).Then, the authors were ordered by their rank as provided by InCites. Preprocessing routines were performed include removing NA values and marking the first 30% of authors as rising stars (with value of 1) and the remaining as not rising stars (with value of 0).

Step 2: The same procedure in **Step 1** was repeated but now for the authors in the time-period 2011-2015. The collected data was stored with the name (**AuthDS2015**).

Step 3: Aiming to ensure that the authors are not rising stars in 2016 because they were already identified as rising stars in 2015- the last year, thus we searched for those authors labeled as rising stars in **AuthDS2015** and excluded them from **AuthDS2016**. The resulted dataset named (**AuthDS**).

Step 4: For each university and institute, the data of publications for the time-period 2011-2015 was queried, downloaded, preprocessed and stored under the name (**DocsDS**).

Step 5: Each author in **AuthDS** was matched with his publications in

DocsDS. The matching task includes, for each author, searching for his publications, collecting the important data and, finally, computing the values of the eight Scientometric indicators (***Author Productivity Level, Scientific Impact Level, Value of Productivity, Citations Per Publication, Contribution Impact, International Collaboration, Research Area Relevancy and Venue Reputation***). The resulted dataset named (**AuthDocsDS**).

For the experimental purposes the **AuthDocsDS** dataset, for each one of the four organizations, was divided randomly into two datasets training and testing datasets (80% **TrainingDS** and 20% **TestingDS**). It is important to mention that, for **AuthDocsDS** with more than 1000 records, a sample of 1000 records was chosen randomly were 30% of them are rising stars and the rest are not. For the chosen samples, the formula proposed by Krejcie and Morgan [51], regarding determining the minimum required sample size, was used to ensure the validity of the used samples in the study where indeed the 1000 samples are more than enough. Figure 4.1 depicts the five steps of data collection process while Table 4.1 shows the results of each dataset for the four selected universities and institutes.

Table 4.1: The Collected Datasets for the Four Selected Universities and Institutions

	KFUPM	**IIT**	**MIT**	**CAS**
AuthDS2016	566	747	2928	9276
AuthDS2015	507	667	2671	8656
AuthDS	388	544	2056	5681
DocsDS	682	5659	3507	11045
AuthDocsDS	334	499	1690	5093
TrainingDS	267	400	800	800
TestingDS	67	99	200	200

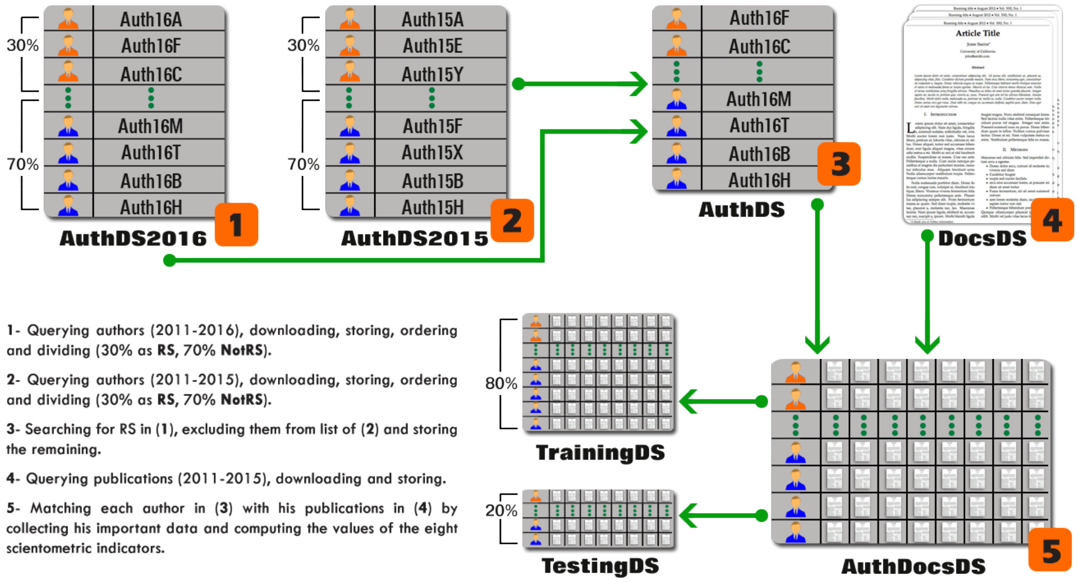
30%
70%
Auth16A
Auth16F
Auth16C
Auth16M
Auth16T
Auth16B
Auth16H
1
AuthDS2016
Auth15A
Auth15E
Auth15Y
Auth15F
Auth15X
Auth15B
Auth15H
2
AuthDS2015
Auth16F
Auth16C
Auth16M
Auth16T
Auth16B
Auth16H
3
AuthDS
Article Title
4
DocsDS
1- Querying authors (2011-2016), downloading, storing, ordering and dividing (30% as RS, 70% NotRS).
2- Querying authors (2011-2015), downloading, storing, ordering and dividing (30% as RS, 70% NotRS).
3- Searching for RS in (1), excluding them from list of (2) and storing the remaining.
4- Querying publications (2011-2015), downloading and storing.
5- Matching each author in (3) with his publications in (4) by collecting his important data and computing the values of the eight scientometric indicators.
80%
TrainingDS
20%
TestingDS
5
AuthDocsDS

Figure 4.1: The five steps of data collection process for the Four Selected Universities and Institutions

By looking back to Table 4.1, it is clear that number of records in **AuthDS** and **DocsDS** are uneven among the four datasets. That is due to the disparity in numbers of faculty members and research activities in that organizations where CAS has the highest number of records among the others.

4.3 Scientometric Indicators: Formulation and Computation

To accomplish the goal of this study we needed to use an appropriate feature space (indicators) with appropriate machine learning techniques. For that purpose, we formulated eight Scientometric indicators and used them as the mean feature space against the response variable. The response variable in our study is "**Rising Star Status**" which indicates whether a researcher is considered a rising star (RS) or not. It has two values, either "TRUE- value of 1" if the researcher was listed in the first 30% of ranks in InCites, otherwise "FALSE- value of 0". The mean feature space is as follows:

Main Feature Space:

1. **Author Productivity Level**: The author productivity is always measured by the number of his publications during specific time period [52]. The level of author productivity can be determined by either an absolute threshold (a fixed threshold is used as a definition) or relative threshold. In this study, we considered three levels of author productivity (Min, Med and Max) where the relative threshold approach was adopted in defining each level. A level is assigned to an author based on the mean value (as a threshold) of all his publications during the time-period. While each author has at least one publication in a time-period, so an author may get one of these levels as follows:

    - Min: Author Publications == One Publication.
    - Med: One Publication $<$ Author Publications $\geq \lceil \text{MEAN} \rceil$.
    - Max: Author Publications $> \lceil \text{MEAN} \rceil$.

    The rationale behind: If a junior author has a high level of productivity since his initial starts, he will have bright chances to become a prominent author in the foreseen future.

2. **Scientific Impact Level**: The author scientific impact is evaluated by counting the citation times of his publications [53]. In this study, there are two levels of scientific impact were considered (Min and Max) while the approach of relative threshold was adopted as in [54]. A level is assigned to an author based on the mean value of the total times that the set of his published docs has been cited by other publications during a time-period. While an author may have zero citations for all his publications in that time-period, so he gets a citation level as follows:

   - Min: Zero Citation $\leq$ Author Publications Times Cited $\leq \lceil$MEAN$\rceil$
   - Max: Author Publications Times Cited $> \lceil$MEAN$\rceil$

   The rationale behind: If a junior author has a high level of a scientific impact since his initial starts, he will have bright chances to become a prominent author in the foreseen future.

3. **Value of Productivity**: According to [55], there are many academic articles were published long time ago and never cited that leads to make them valueless. Herein, the value of the author productivity was proposed to be evaluated as the rate of publications that have been cited one or more times- the inverse of the rate of publications that didn't get cited at all during a

specific time period.

$$\text{Value of Productivity} = 1 - \frac{\text{Publications Not Cited}}{\text{Total of Author's Publications}}$$

The value of this indicator is between 0 and 1, the greater the better.

The rationale behind: If a junior author has a high value of productivity since his initial starts, he will have bright chances to become a prominent author in the foreseen future.

4. **Citations Per Publication**: Contains the rate of citations per paper-mean number of citations that a document has received. It is an indicator to evaluate author research performance (Citation Impact) and is calculated by dividing the total number of author citations by the total number of his publications.

$$\text{Citations Per Publication} \; = \frac{\text{Total of Author Publications' Citation}}{\text{Total of Author's Publications}}$$

The value of this indicator is a numeric, the greater the better.

The rationale behind: If a junior author has high performance since his initial starts, he will have bright chances to become a prominent author in the foreseen future.

5. **Contribution Impact**: For publications with multiple authors (co-authors), it is inherently unfair to treat all co-authors equally in counting publication credit either by issuing full credit to each one (inflationary bias)

or by distributing one credit equally among all authors (equalizing bias) [56]. So, the practices of assigning publication credit in multiple authors publications remains a problematic task and common objection among different research approaches [57].

In this work, we proposed the contribution impact to be the rate of author publication credits in all his publications. Although there are several approaches has been proposed to find author publications credits, the *Credit of Harmonic Authorship Allocation* [58] approach was adopted in our work. Hence, the contribution impact for an author is calculated by dividing the harmonic credit of all his co-author's contributions by the total number of his publications.

The harmonic credit for the i^{th} author ($i = AuthorPosition$) of a publication with k co-authors is calculated as follows:

$$\text{Harmonic Credit } \left(i^{th} AuthorCredit\right) = \frac{\left(\frac{1}{i}\right)}{\left(1 + \frac{1}{2} + \frac{1}{3} + ... + \frac{1}{n-1} + \frac{1}{k}\right)}$$

This formula of Harmonic Authorship Allocation ensures that:

- The total publication credit is shared among all coauthors: removes the inflationary bias (repeatedly issued full publication credit to all co-authors).
- The first author gets the most credit and in general the i^{th} author receives more credit than the $(i+1)^{th}$ author: removes the equaliz-

ing bias (dividing credit uniformly among all co-authors irrespective of their actual contribution).

- The greater the number of authors, the less credit per author.

Figure 4.2, depicts harmonic authorship credit distribution among co-authors.

So, the author contribution impact can be calculated as:

$$\text{Contribution Impact} = \frac{\text{Total of Author's Harmonic Credits}}{\text{Total of Author's Publications}}$$

The value of this indicator is between 0 and 1, the greater the better.

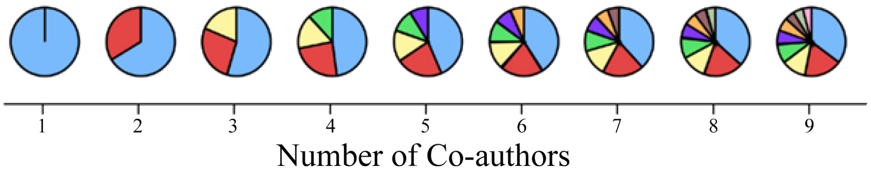

Figure 4.2: The harmonic authorship credit distribution among co-authors

The rationale behind: If a junior author, since his initial starts, can collaborate with senior authors and can influence their work as a main contributor (by appearing at first positions in papers), he will have bright chances to become a prominent author in the foreseen future.

6. **International Collaboration**: Contains the rate of publications that have one or more international co-authors- a publication with at least two authors from different countries. The international collaboration is an indication of

an author's ability to attract the world, In the meanwhile, the researches have shown that citations to articles resulting from international collaborations grew faster than those referring to domestic collaborations [59], [60]. Author international collaboration is calculated as following:

$$\text{International Collaborations} = \frac{\text{Total of Papers with International Co-authors}}{\text{Total of Author's Publications}}$$

The value of this indicator is between 0 and 1, the greater the better.

The rationale behind: If a junior author, since his initial starts, has established relationships with various authors in different scholars, that will help him to broaden his horizons and improve his academic capacities, as well as, give him the chance to become prominent and expert in the foreseen future.

7. **Research Area Relevancy**: Contains the rate of publications' percentiles that are relevant to author subject or research area. For an author, focusing on a specific research area is an excellent evidence that he is versed and well-skilled in that area. Research area relevancy for an author can be calculated by dividing the summation of all the area percentiles (captured from InCite) of all the papers by the total number of his papers (documents).

$$\text{Research Area Relevancy} = \frac{\frac{\text{Total of All papers' Percentiles}}{100}}{\text{Total of Author's Publications}}$$

The value of this indicator is between 0 and 1, the greater the better.

The rationale behind: If a junior author, since his initial starts, has focused on specific research area, he will have the ability to make contributions to the knowledge in that area and being distinguished in the foreseen future.

8. **Venue Reputation**: Authors are evaluated based on their ability to publish in refereed venues. Where publishing venue maybe a journal, conference, symposium, workshop, etc., however, the quality of the venue is measured by it is standards of acceptance. That means if the venue has high acceptance standards for publications then that venue is presumed to be high in quality and reputation [61]. Journals are considered the highest in quality and reputation and they are compared by their *Impact Factors*. To evaluate the venue quality and reputation, we used the ISI Journal Impact Factor which can be found in the Journal Citation Reports. Venue reputation indicator can be calculated by dividing the impact factors of all publishing venues (journals) by the total number of publications- the unlisted venue impact factors in that report is considered to be zero.

$$\text{Venue Reputation} = \frac{\text{Total of Author's Publications Impact Factors}}{\text{Total of Author's Publications}}$$

The value of this indicator is numeric, the greater the better.

The rationale behind: if a junior author, since his initial starts, has published in high-level venues, he will have the ability to be a distinguished and prominent in the foreseen future.

In addition to those main indicators mentioned above, there are other six supplementary indicators for which their values were gathered and used.

Supplementary Indicators:

1. **Author Name**: contains the name of author (surname and initials) as captured from each publication.

2. **Affiliation**: contains the name of institutions which authors' affiliated with.

3. **Rank**: contains a number indicated the author rank among others. The smaller the number the higher the rank. The rank number was taken from InCites.

4. **Time-Period**: contains the period of years where Rank indicators were chosen based on author activity within.

5. **Region**: contains the region which authors' affiliation belongs to.

6. **Rising Star Status:** contains values of "TRUE- value of 1" for the rising star authors (those who listed in the first 30% of the ranks) otherwise "FALSE- value of 0".

Tables from Table 4.2 to Table 4.5 show summary of the four **AuthDocsDS** datasets (KFUPM, IIT, MIT and CAS) with the computed values of their Scientometric indicators starting by response variable "**Rising Star Status**".

Table 4.2: A Summary of AuthDocsDS for KFUPM Dataset (334 Records)

Categorical Indicators			
Rising Star Status	**FALSE**: 257	**TRUE**: 77	
Author Productivity Level	**Max**: 54	**Med**: 78	**Min**: 202
Scientific Impact Level	**Max**: 63	**Min**: 271	

Numerical Indicators	**Min.**	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max.**
Value of Productivity	0	0	0	0.2892	0.5	1
Citations Per Publication	0	0	0	1.146	1	17
Contribution Impact	0.09	0.24	0.33	0.3898	0.55	1
International Collaborations	0	0	0	0.2087	0.33	1
Research Area Relevancy	0	0	0	0.1997	0.39	0.99
Venue Reputation	0	0	0	0.4499	0.655	3.4

Table 4.3: A Summary of AuthDocsDS for IIT Dataset (499 Records)

Categorical Indicators			
Rising Star Status	**FALSE**: 390	**TRUE**: 109	
Author Productivity Level	**Max**: 100	**Med**: 151	**Min**: 248
Scientific Impact Level	**Max**: 79	**Min**: 420	

Numerical Indicators	**Min.**	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max.**
Value of Productivity	0	0	0	0.2447	0.485	1
Citations Per Publication	0	0	0	1.06	1	22
Contribution Impact	0.09	0.24	0.35	0.3941	0.55	1
International Collaborations	0	0	0	0.05475	0	1
Research Area Relevancy	0	0	0	0.1757	0.29	1
Venue Reputation	0	0	0	0.306	0.42	3.72

Table 4.4: A Summary of AuthDocsDS for MIT Dataset (1690 Records)

Categorical Indicators			
Rising Star Status	**FALSE**: 1299	**TRUE**: 391	
Author Productivity Level	**Max**: 256	**Med**: 446	**Min**: 988
Scientific Impact Level	**Max**: 330	**Min**: 1360	

Numerical Indicators	**Min.**	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max.**
Value of Productivity	0	0	0.33	0.4405	1	1
Citations Per Publication	0	0	0.5	4.168	3.33	399.5
Contribution Impact	0.09	0.19	0.33	0.3582	0.49	1
International Collaborations	0	0	0	0.2294	0.33	1
Research Area Relevancy	0	0	0.23	0.3238	0.63	1
Venue Reputation	2	143.6	194	162.6	194	194

Table 4.5: A Summary of AuthDocsDS for CAS Dataset (5093 Records)

Categorical Indicators			
Rising Star Status	**FALSE**: 3906	**TRUE**: 1187	
Author Productivity Level	**Max**: 751	**Med**: 1718	**Min**: 2624
Scientific Impact Level	**Max**: 686	**Min**: 4407	

Numerical Indicators	**Min.**	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max.**
Value of Productivity	0	0	0	0.3201	0.6	1
Citations Per Publication	0	0	0	1.73	2	171
Contribution Impact	0.09	0.16	0.25	0.2889	0.41	1
International Collaborations	0	0	0	0.1147	0	1
Research Area Relevancy	0	0	0	0.2165	0.39	1
Venue Reputation	0	0	0	0.6096	0.91	9.22

4.4 Research Questions

In this work, we are interested in answering the following questions aimed at helping to come up with applicable and reliable models:

- **Q1: Are the future rising stars predictable**? (Is there a relationship between the response (**Rising Star Status**) and each predictor (the eight Scientometric indicators)? Is at least one of the indicator $X_1, X_2, ..., X_8$ useful in predicting the response Y?)
  - This question can be answered by fitting a multiple regression model of response to all the indicators and testing the ***null hypothesis***. The ***F-statistic*** can be used to determine whether or not the null hypothesis should be rejected. In case that the ***p-value*** corresponding to the ***F-statistic*** is very low while the ***F-statistic*** is far larger than 1, so that reveals a clear evidence of a relationship between the response and at least one of the indicators.
- **Q2: How accurately can we predict future rising stars**?
  - The answer of this question can be achieved by applying different classifiers (NB, kNN, NN and SVM) and looking to the resulted confusion matrices in addition to other evaluation metrics such as ROC and AUC to evaluate and compare the classifiers accuracy values.
- **Q3: Is there any significant difference in obtained results between the applied machine learning techniques (classifiers) in predicting**

rising stars?

- **Q4: Is there any significant difference in obtained results between the datasets used in predicting rising stars**?
- **Q5: For the obtained results, is there any significant difference in the interaction between the applied machine learning techniques (classifiers) and the used datasets in predicting rising stars**?
  - Questions **Q3**, **Q4** and **Q5** can be answered by applying ANOVA test and analyzing the values of the ***F-value***, the ***f critical-value*** and the ***p-value***.
- **Q6: Up to which level are we confident that the proposed models are able to predict the future rising stars when they are used with alternative datasets**?
  - To answer this question an untouched dataset can be used to verify the applicability of the proposed models in predicting the future rising stars. The validation process can be evaluated using different performance evaluation metrics.
- **Q7: What are the most important Scientometric indicators in predicting future rising stars**?
  - This question can be answered by selecting the best model/ models and assessing the ***Relative Importance*** of each indicator- assessing percent contribution to each correlated indicator.

- **Q8: How are our proposed models comparable to other models**?
    - To answer this question a comparison between the best proposed models in this work and those proposed in one of the previous works will be conducted and reported.

CHAPTER 5

EXPERIMENTS AND RESULTS

5.1 Experimental Work

In order to predict the future rising stars in universities and research institutes, we experiment using our methodology with the four datasets that collected from the WoS for authors historical research activities during the time-period 2011-2015. We focused on the researches outcome of KFUPM, IIT, MIT and CAS related to the Computer Science domain. The goal of the conducted experiments is to find answers to the research questions aforementioned in the previous chapter. R; the flexible, pretty nifty and powerful language and software environment; was used in all the experiments to implement the opted machine learning techniques.

We started by answering the first research question (**Q1**): **Are the future**

rising stars predictable? *(Is there a relationship between the response "**Rising Star Status**" and the proposed Scientometric indicators).* To answer this question we have to investigate whether is at least one of the indicators $X_1, X_2, ..., X_8$ is useful in predicting the response Y or not. Therefore, to investigate that in a scientific way, the following null hypothesis was stated:

- $H1_{1-0}$: Predicting rising stars ***is not*** achievable by applying the proposed Scientometric indicators (no relationship between **Rising Star Status** and other Scientometric indicators).

The **F-statistic** can be used to infer whether or not we should reject this null hypothesis. In case that the **p-value** corresponding to the **F-statistic** is very low (< 0.05), while the **F-statistic** is far larger than 1, so that provides a compelling evidence against the null hypothesis and the alternative hypothesis will be accepted:

- $H1_{1a}$: Predicting rising stars ***is*** achievable by applying the proposed Scientometric indicators (there is a relationship between **Rising Star Status** and other Scientometric indicators).

MLR has distinct advantages in terms of interpretation and inference, notwithstanding, the general assumption about the MLR is the linear relationship between the response and the indicators which, in reality, that linearity assumption has to be relaxed. Therefore, we applied the MLR to the main four training datasets (**TrainingDS**) for testing the null hypothesis and finding the

Table 5.1: The F-statistic and the p-value for Fitting the Four Training Datasets with Multiple Linear Regression

	KFUPM	**IIT**	**MIT**	**CAS**
F-statistic	60.9	56.36	136.8	92.27
p-value	2.2e-16	2.2e-16	2.2e-16	2.2e-16

question's answer. The experimental results obtained for the **F-statistic** and the **p-value** are shown in Table 5.1.

From Table 5.1, it's obvious that the Scientometric indicators are playing a role in predicting future rising stars in all the datasets where the corresponding **F-statistics** are greater than 1 and the **p-values** are extremely low ($2.2e-16$ *is the smallest possible number the system can show*). That forms a strong evidence of the relationship between the response, rejecting the null hypotheses and accepting the alternative.

The second research question (**Q2**): **How accurately can we predict future rising stars?** To answer this question, the four classifiers NB, kNN, NN and SVM were applied. The package "***caret***", short for *classification and regression training* [62], was used to implement the classifiers with R tool. The intent of adopting "***caret***" package is that because it has several functions that attempt to streamline the model building, using resampling and evaluation process as well as feature selection and other sophisticated techniques. The 10-fold cross-validation resampling method was opted of in the tuning function to ensure the accuracy of the proposed models and to avoid the over-fitting problem

[63]. We investigated the answer of this question in two rounds. In each round we used different metrics for evaluating performance and accuracy of classifiers.

First round: the classifiers were trained on the training datasets "**TrainingDS**" of the four organizations (KFUPM, IIT, MIT and CAS). The categorical variables (indicators), such as **Author Productivity Level** and **Scientific Impact Level**, were converted to ***dummy*** variables, then all the numerical indicators were scaled to zero mean and unit variance. After that preprocessing steps, the models (classifiers) were trained on the training datasets under the following parameters which were chosen based on the obtained *superior performance* measurements with all the four datasets:

- **Naive Bayes**:
  - Laplace correction: fL=0, Distribution type: kernel and Bandwidth adjustment: = 1.
- **k Nearest Neighbors**:
  - Number of considered neighbors: k= 11.
- **Neural Networks**:
  - Number of hidden layerss:1, Number of hidden neurons: 2 and Algorithm: backpropagation.
- **Support Vector Machine**:

  - SVM-Kernel: Radial and Gamma: 0.5.

The predictivity accuracy of the proposed models was evaluated using two popular metrics the *Receiver Operating Characteristic (ROC)* curve [64] and the *Area Under the Curve (AUC)* [65]. ROC is a graphical plot that illustrates the performance of the classifiers by plotting the true positive rate (TPR/ Sensitivity) as a function of the false positive rate (FPR/ 1-Specificity) for different cut-off points. A test with perfect discrimination has a ROC curve that passes through the upper left corner (100% sensitivity, 100% specificity). Yet, the closer the graph is to the top and the left-hand corner, the more accurate the test [66]. AUC measures the area under the ROC curve where an area close to 1 represents a perfect test. Figure 5.1 to Figure 5.4 illustrate the ROC and the AUC of NB, kNN, NN and SVM classifiers with KFUPM, IIT, MIT and CAS training datasets while Table 5.2 shows the AUC values.

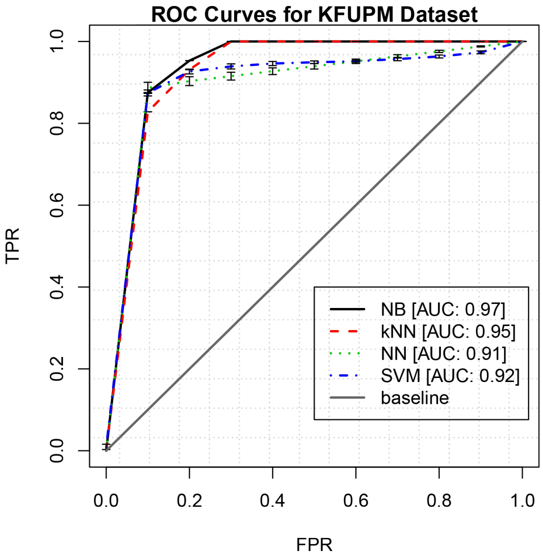

Figure 5.1: The ROC curves and the AUC for KFUPM training dataset with NB, kNN, NN and SVM models

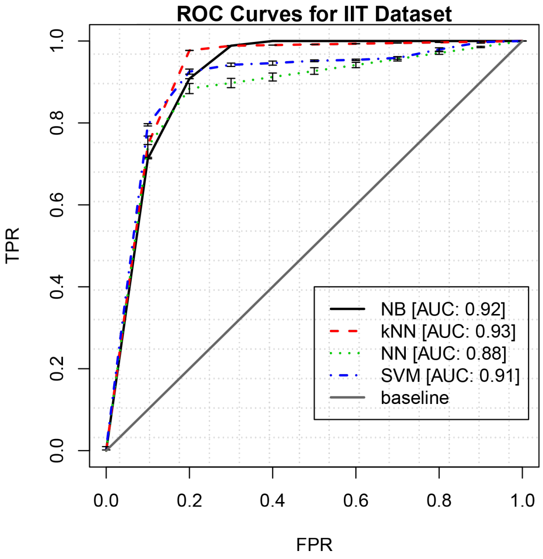

Figure 5.2: The ROC curves and the AUC for IIT training dataset with NB, kNN, NN and SVM models

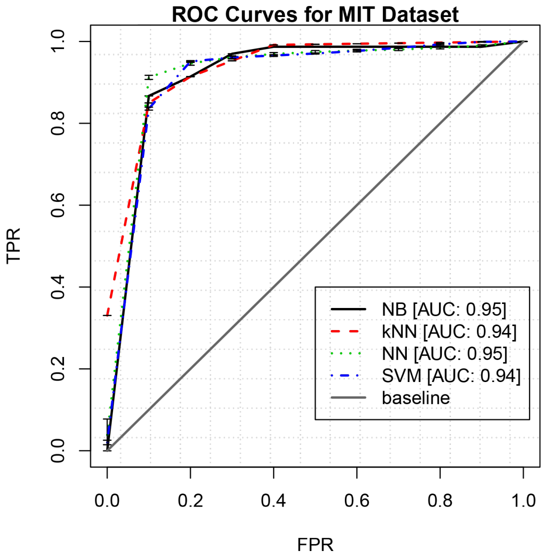

Figure 5.3: The ROC curves and the AUC for MIT training dataset with NB, kNN, NN and SVM models

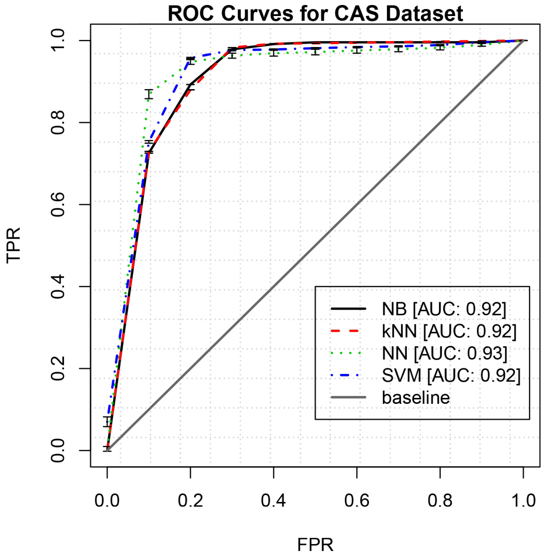

Figure 5.4: The ROC curves and the AUC for CAS training dataset with NB, kNN, NN and SVM models

Looking at the above table and figures, it is obvious that the obtained values for both the ROC and the AUC are promising where the lowest value of the AUC

Table 5.2: The AUC for Fitting NB, kNN, NN and SVM with Training Datasets of KFUPM, IIT, MIT and CAS

	KFUPM	**IIT**	**MIT**	**CAS**
NB	0.97	0.92	0.95	0.92
kNN	0.95	0.93	0.94	0.92
NN	0.91	0.88	0.95	0.93
SVM	0.92	0.91	0.94	0.92

is 0.88 for NN with IIT and the highest is 0.97 for NB with KFUPM. With these obtained values we can give a positive answer to the second question where the future rising stars can be predicted with high accuracy. Anyhow, in reality one may interested in providing an unbiased predictivity evaluation of the proposed models. This can be achieved by using untouched dataset different from that used in fitting models- testing dataset.

Second round: to provide an impartial predictivity evaluation, the proposed models resulted from the **first round** were tested with the testing datasets "**TestingDS**" of the four organizations (KFUPM, IIT, MIT and CAS). The *Confusion Matrix*, a metric showing the predicted and actual classifications, was used here in order to evaluate and determine how many observations were correctly or incorrectly predicted by the proposed models [67]. Table 5.3 to Table 5.6 show the confusion matrix for each one of the four proposed models with the testing datasets of the four organizations. Figure 5.5 shows the accuracy summary of the proposed models.

Table 5.3: A Confusion Matrix Compares Models Predictions to The true Default Statuses for the 67 Observations in KFUPM Testing Dataset

Naive Bayes	Predicted FALSE	Predicted TRUE
Actual FALSE	49	5
Actual TRUE	0	13
Classifier Accuracy	93%	

k Nearest Neighbors	Predicted FALSE	Predicted TRUE
Actual FALSE	50	4
Actual TRUE	3	10
Classifier Accuracy	90%	

Neural Networks	Predicted FALSE	Predicted TRUE
Actual FALSE	50	4
Actual TRUE	2	11
Classifier Accuracy	91%	

Support Vector Machine	Predicted FALSE	Predicted TRUE
Actual FALSE	50	2
Actual TRUE	4	11
Classifier Accuracy	91%	

Table 5.4: A Confusion Matrix Compares Models Predictions to the True Default Statuses for the 99 Observations in IIT Testing Dataset

Naive Bayes	Predicted FALSE	Predicted TRUE
Actual FALSE	71	6
Actual TRUE	7	15
Classifier Accuracy	87%	

k Nearest Neighbors	Predicted FALSE	Predicted TRUE
Actual FALSE	74	3
Actual TRUE	7	15
Classifier Accuracy	89%	

Neural Networks	Predicted FALSE	Predicted TRUE
Actual FALSE	75	2
Actual TRUE	8	14
Classifier Accuracy	90%	

Support Vector Machine	Predicted FALSE	Predicted TRUE
Actual FALSE	74	5
Actual TRUE	3	17
Classifier Accuracy	92%	

Table 5.5: A Confusion Matrix Compares Models Predictions to the True Default Statuses for the 200 Observations in MIT Testing Dataset

Naive Bayes	Predicted FALSE	Predicted TRUE
Actual FALSE	127	6
Actual TRUE	19	48
Classifier Accuracy	89%	

k Nearest Neighbors	Predicted FALSE	Predicted TRUE
Actual FALSE	127	6
Actual TRUE	12	55
Classifier Accuracy	91%	

Neural Networks	Predicted FALSE	Predicted TRUE
Actual FALSE	125	8
Actual TRUE	12	55
Classifier Accuracy	90%	

Support Vector Machine	Predicted FALSE	Predicted TRUE
Actual FALSE	126	12
Actual TRUE	7	55
Classifier Accuracy	91%	

Table 5.6: A Confusion Matrix Compares Models Predictions to the True Default Statuses for the 200 Observations in CAS Testing Dataset

Naive Bayes	Predicted FALSE	Predicted TRUE
Actual FALSE	122	11
Actual TRUE	15	52
Classifier Accuracy	87%	

k Nearest Neighbors	Predicted FALSE	Predicted TRUE
Actual FALSE	124	9
Actual TRUE	17	50
Classifier Accuracy	87%	

Neural Networks	Predicted FALSE	Predicted TRUE
Actual FALSE	124	9
Actual TRUE	16	51
Classifier Accuracy	88%	

Support Vector Machine	Predicted FALSE	Predicted TRUE
Actual FALSE	120	13
Actual TRUE	13	54
Classifier Accuracy	87%	

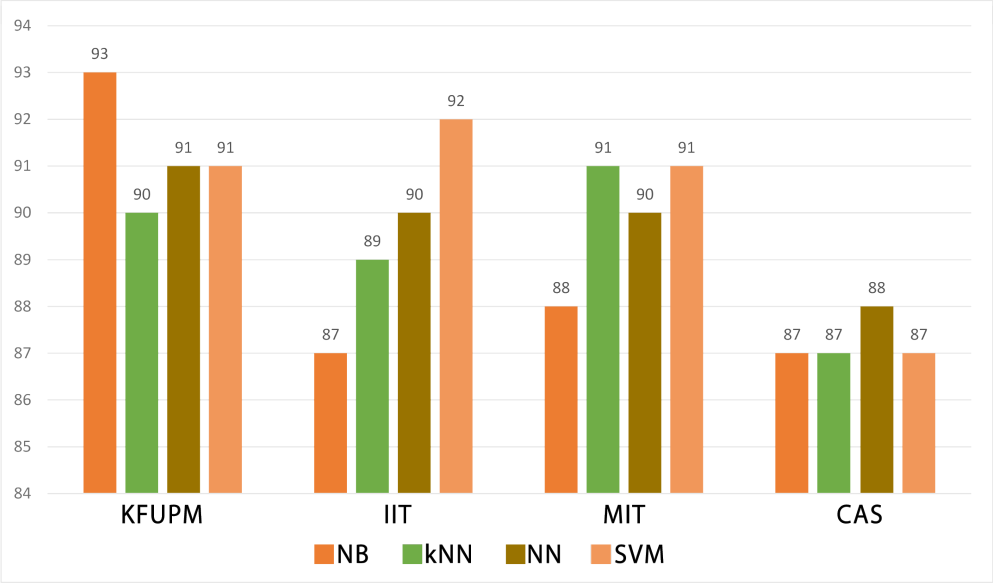

Figure 5.5: The accuracy summary of proposed models with testing datasets

At models level, it is clear that all the four classifiers have outstanding performance with prediction accuracy ranges from 87% (NB with IIT, and NB, kNN and SVM with CAS) to 93% (SVM with KFUPM). On the whole, SVM outperforms the other models on KFUPM, IIT, MIT and CAS testing datasets with 91%, 92%, 91% and 87% respectively. At organizations level, KFUPM got the highest accuracy levels (90% - 93%) while CAS got the lowest accuracy levels (87% - 88%). The reasons why CAS got that values may be due to the observations' selection of the data samples or may be due to the research activities at CAS itself (in the forthcoming lines we will investigate more to reveal the reasons behind). In general, all the obtained results are emboldening and giving answer to **Q2** where we can predict future rising stars with high accuracy using our proposed models, and in the meanwhile, that leads us to achieve our essential goal of this work.

Before progressing to find answers for the questions **Q3**, **Q4** and **Q5**, extra two samples with 1000 observation for each one of MIT and CAS organizing were randomly selected. The purpose of this step is to ensure that the obtained results for answering **Q**2 were not happened by chance due to the observations in the created samples for MIT and CAS. So, with these samples we end up with eight datasets, KFUPM- 334 observations, IIT- 499 observations, MIT1- 1000 (the already used dataset in previous) observations, MIT2- 1000 observations, MIT3- 1000 observations, CAS1- 1000 observations (the already used dataset in previous), CAS2- 1000 observations and CAS3- 1000 observations. For the experiments, each dataset was divided into 80% of observations as training dataset and 20% as testing dataset to be used to train and test the four models, respectively. For each model and dataset, the experiments were run 10 times with the *optimal* values of parameters tuned by the function "*train*" from package "caret" (in contrast with values of parameters in the **first round** which were selected based on the superior performance measurements) with the 10-fold cross-validation resampling method. For investigating the performance of the resulted models, we did not assess them on the testing datasets by just looking to the *prediction accuracy level*, but further, we used additional performance evaluation metrics such as *sensitivity*, *specificity* and *F1 score*. Tables from Table 5.7 to Table 5.10 show the performance evaluation metrics for the models NB, kNN, NN and SVM, respectively, with the eight testing datasets. The same case for figures from Figure 5.6 to Figure 5.9.

Table 5.7: The Performance Evaluation Metrics for Naive Bayes with the Eight Testing Datasets

	Accuracy	**Sensitivity**	**Specificity**	**F1 score**
KFUPM	93	0.91	1	0.95
IIT	87	0.92	0.68	0.92
MIT-1	88	0.95	0.72	0.91
MIT-2	89	0.91	0.84	0.92
MIT-3	88	0.9	0.84	0.91
CAS-1	87	0.92	0.78	0.9
CAS-2	83	0.84	0.8	0.88
CAS-3	84	0.84	0.85	0.88

Table 5.8: The Performance Evaluation Metrics for k Nearest Neighbors with the Eight Testing Datasets

	Accuracy	**Sensitivity**	**Specificity**	**F1 score**
KFUPM	89	0.91	0.81	0.93
IIT	89.8	0.96	0.66	0.94
MIT-1	90.6	0.95	0.81	0.93
MIT-2	88.3	0.91	0.82	0.92
MIT-3	86	0.911	0.74	0.9
CAS-1	87	0.92	0.78	0.9
CAS-2	83.7	0.87	0.76	0.89
CAS-3	82	0.88	0.68	0.87

Table 5.9: The Performance Evaluation Metrics for Neural Networks with the Eight Testing Datasets

	Accuracy	**Sensitivity**	**Specificity**	**F1 score**
KFUPM	91.3	0.91	0.91	0.94
IIT	88.5	0.94	0.69	0.93
MIT-1	90.2	0.941	0.829	0.93
MIT-2	90.8	0.93	0.869	0.93
MIT-3	90.3	0.93	0.849	0.93
CAS-1	87.7	0.93	0.76	0.91
CAS-2	88	0.88	0.88	0.92
CAS-3	87.9	0.94	0.75	0.92

Table 5.10: The Performance Evaluation Metrics for Support Vector Machine with the Eight Testing Datasets

	Accuracy	**Sensitivity**	**Specificity**	**F1 score**
KFUPM	90.7	0.92	0.85	0.94
IIT	92	0.96	0.77	0.95
MIT-1	90.2	0.95	0.82	0.93
MIT-2	89.6	0.95	0.75	0.93
MIT-3	86.2	0.91	0.77	0.9
CAS-1	87.2	0.9	0.827	0.9
CAS-2	85.8	0.88	0.8	0.9
CAS-3	86.1	0.89	0.82	0.91

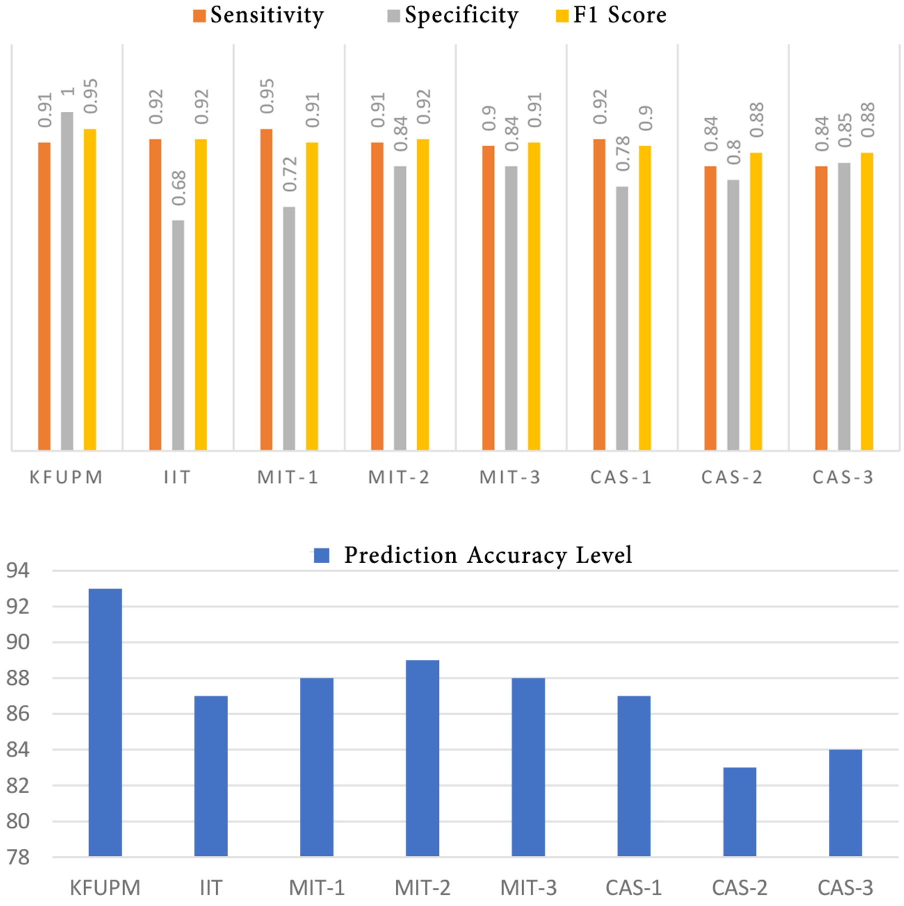

Figure 5.6: The performance evaluation metrics for NB with all datasets

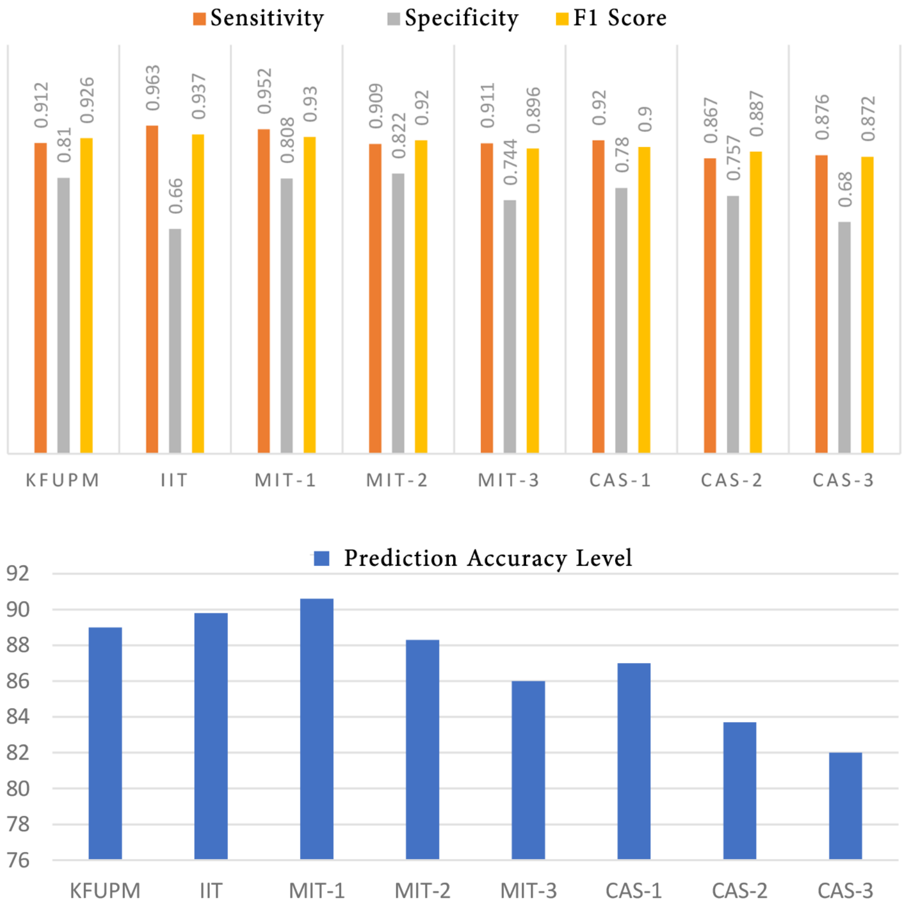

Figure 5.7: The performance evaluation metrics for kNN with all datasets

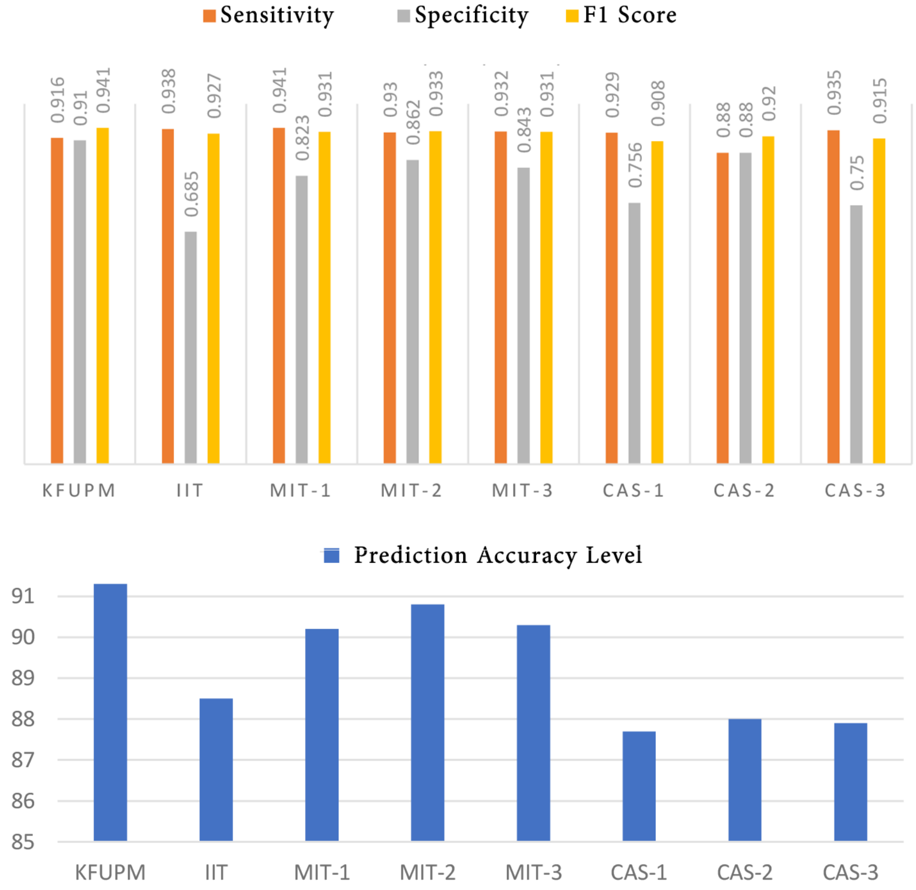

Figure 5.8: The performance evaluation metrics for NN with all datasets

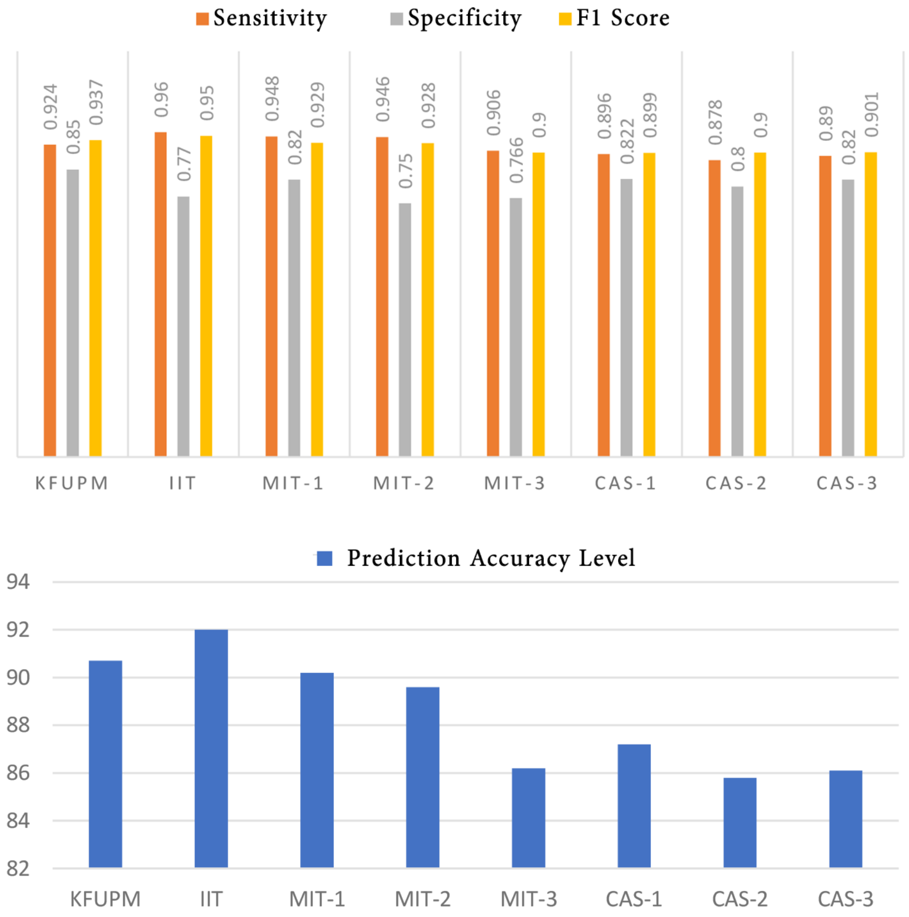

Figure 5.9: The performance evaluation metrics for SVM with all datasets

Based on what is done in above, we took the step forward to answer the questions **Q3**, **Q4** and **Q5**. For that, Two-way ANOVA analysis, short of Analysis Of Variance, was conducted. The ANOVA analysis is a collection of statistical models used for comparing and testing means of groups in our work the group of classifiers and the group of datasets [68]. The primary purpose of the two-way ANOVA is to understand the importance of each group (classifiers and datasets) and to find if there is an interaction between them.

Conventionally, there are three sets of hypothesis come along with the two-way ANOVA. Regarding our work, the null hypotheses for each of those sets are given below:

- $H2_{1-0}$: There ***is no significant*** difference between the applied classifiers in predicting rising stars.
- $H2_{2-0}$: There ***is no significant*** difference between the datasets used in predicting rising stars.
- $H2_{3-0}$: There ***is no significant*** difference in the interaction between the applied classifiers and the used datasets in predicting rising stars.

Consequently, the alternative hypotheses are:

- $H2_{1a}$: There ***is a significant*** difference between the applied classifiers in predicting rising stars.

- $H2_{2a}$: There ***is a significant*** difference between the datasets used in predicting rising stars.
- $H2_{3a}$: There ***is a significant*** difference in the interaction between the applied classifiers and the used datasets in predicting rising stars.

The acceptance and rejection of the null hypotheses will form answers to the research questions **Q3**, **Q4** and **Q5**. Thus, by choosing alpha level equal to (0.05), the null hypothesis is rejected when the obtained ***p-value is smaller than that alpha level and the value of F-ratio is larger than the F-critical (F-statistic)***.

For the purposes of conducting ANOVA analysis, the F1 score results from testing models on the eight testing datasets were obtained for the 10-runs. The reason behind choosing the F1 score is two-fold, the first is to be used as a proof-of-concept for the applicability of our models in predicting the future rising stars as well as to insure the answer of **Q2**. The second reason is to compare the final results with the work of [1] (in answering **Q8**). Table 5.11 shows the F1 scores for the 10-runs experiments.

Table 5.12 presents the results of ANOVA analysis on data values from Table 5.11. The table has columns labeled Sum of Squares (SS), Degrees of freedom (DF), Mean Square (MS), F-ratio, p-value, and F-critical. The only columns those are critical for the intended interpretation are the last three

Table 5.11: The F1 Scores for the 10-Runs Experiments (Testing Models on the Eight Testing Datasets)

	KFUPM	**IIT**	**MIT-1**	**MIT-2**	**MIT-3**	**CAS-1**	**CAS-2**	**CAS-3**
NB	0.95	0.92	0.91	0.92	0.91	0.9	0.88	0.88
	0.95	0.92	0.91	0.92	0.91	0.9	0.88	0.88
	0.95	0.92	0.91	0.92	0.91	0.9	0.88	0.88
	0.95	0.92	0.91	0.92	0.91	0.9	0.88	0.88
	0.95	0.92	0.91	0.92	0.91	0.9	0.88	0.88
	0.95	0.92	0.91	0.92	0.91	0.9	0.88	0.88
	0.95	0.92	0.91	0.92	0.91	0.9	0.88	0.88
	0.95	0.92	0.91	0.92	0.91	0.9	0.88	0.88
	0.95	0.92	0.91	0.92	0.91	0.9	0.88	0.88
	0.95	0.92	0.91	0.92	0.91	0.9	0.88	0.88
kNN	0.92	0.93	0.93	0.92	0.89	0.9	0.88	0.87
	0.93	0.94	0.93	0.92	0.89	0.9	0.89	0.87
	0.94	0.94	0.93	0.92	0.9	0.9	0.89	0.87
	0.92	0.94	0.93	0.92	0.9	0.9	0.89	0.87
	0.92	0.94	0.93	0.92	0.89	0.9	0.88	0.88
	0.92	0.93	0.93	0.92	0.89	0.9	0.89	0.87
	0.94	0.94	0.93	0.92	0.9	0.9	0.89	0.87
	0.92	0.94	0.93	0.92	0.9	0.9	0.88	0.88
	0.92	0.93	0.93	0.92	0.9	0.9	0.89	0.87
	0.93	0.94	0.93	0.92	0.9	0.9	0.89	0.87
NN	0.94	0.88	0.93	0.94	0.93	0.91	0.92	0.92
	0.94	0.94	0.94	0.93	0.92	0.91	0.92	0.91
	0.95	0.94	0.93	0.93	0.93	0.91	0.92	0.91
	0.95	0.94	0.93	0.94	0.93	0.91	0.92	0.92
	0.94	0.92	0.93	0.94	0.93	0.91	0.92	0.91
	0.94	0.89	0.93	0.93	0.93	0.91	0.92	0.92
	0.93	0.94	0.93	0.93	0.93	0.91	0.92	0.92
	0.94	0.94	0.93	0.93	0.94	0.91	0.92	0.91
	0.94	0.94	0.93	0.93	0.94	0.89	0.92	0.92
	0.94	0.94	0.93	0.93	0.93	0.91	0.92	0.91
SVM	0.94	0.95	0.93	0.93	0.9	0.9	0.9	0.9
	0.93	0.95	0.92	0.93	0.9	0.9	0.9	0.9
	0.94	0.95	0.93	0.93	0.9	0.9	0.9	0.9
	0.93	0.95	0.93	0.93	0.9	0.9	0.9	0.9
	0.94	0.95	0.93	0.93	0.9	0.89	0.9	0.9
	0.93	0.95	0.93	0.93	0.9	0.9	0.9	0.9
	0.94	0.95	0.93	0.93	0.9	0.9	0.9	0.91
	0.94	0.95	0.92	0.93	0.9	0.9	0.9	0.9
	0.94	0.95	0.93	0.91	0.9	0.9	0.9	0.9
	0.94	0.95	0.94	0.93	0.9	0.9	0.9	0.9

(F-ratio, p-value, and F-critical). The remaining columns are used mainly for intermediate computational purposes.

Table 5.12: The ANOVA Analysis Output

Source of Variation	**SS**	**DF**	**MS**	**F-ratio**	**p-value**	**F-critical**
Classifiers	0.01645	3	0.005483333	179.4545455	1.31942E-65	2.635951054
Datasets	0.08679	7	0.012398571	405.7714286	4.0686E-145	2.041441862
Interaction	0.02454	21	0.001168571	38.24415584	1.1004E-70	1.592707936
Within	0.0088	288	3.05556E-05			
Total	0.13658	319				

By looking at Table 5.12, it is obvious that:

- For Classifiers: since the p-value is smaller than the alpha level and the F-ratio is larger than the F-critical, then there is a significant difference between the applied classifiers in predicting rising stars. Yet, the null hypothesis is rejected (the group of classifiers plays an important role in predicting rising stars).

- For Datasets: since the p-value is smaller than the alpha level and the F-ratio is larger than the F-critical, then there is a significant difference between the datasets used in predicting rising stars. Yet, the null hypothesis is rejected (the group of datasets plays an important role in predicting rising stars).

- For Interaction: since the p-value is smaller than the alpha level and the F-ratio is larger than the F-critical, then there is a significant difference in the interaction between the applied classifiers and the used datasets in predicting rising stars. Yet, the null hypothesis is rejected (the interaction between the group of classifiers and the group of datasets plays an important role in predicting rising stars).

Regarding the sixth question (**Q6**): **Up to which level are we confident that the proposed models are able to predict the future rising stars when they are used with alternative datasets?**, we validated our proposed models on a dataset generated from the eight testing datasets (untouched datasets) of the four organizations (called **ValDS**). The observations of the generated **ValDS** were selected randomly with 9% from KFUPM, 17% from IIT and 37% from each one of MIT and CAS. The total number of the selected observations was 150. The models resulted from answering the questions **Q3**, **Q4** and **Q5** were validated on **ValDS**. The performance of the validation process was extensively evaluated using different metrics include of F1 Scores, Accuracy Levels and 95% Confidence Intervals and presented in Table 5.13.

Table 5.13: The F1 Scores, Accuracy Levels and 95% Confidence Intervals Values Resulted form Validating the Proposed Models

Classifier	Metric	KFUPM	IIT	MIT_1	MIT_2	MIT_3	CAS_1	CAS_2	CAS_3
NB	**F1 Score**	0.91	0.84	0.95	0.95	0.95	0.92	0.94	0.93
	Accuracy Level	87	78	91	93	92	88	91	89
	95% CI [LL, UL]	[81, 92]	[70, 84]	[86, 95]	[87, 96]	[86, 96]	[82, 93]	[85, 95]	[83, 94]
kNN	**F1 Score**	0.77	0.75	0.92	0.92	0.92	0.76	0.74	0.75
	Accuracy Level	69	68	87	87	87	69	67	69
	95% CI [LL, UL]	[61, 77]	[60, 75]	[81, 92]	[80, 92]	[80, 92]	[61, 76]	[59, 75]	[61, 76]
NN	**F1 Score**	0.77	0.86	0.95	0.95	0.95	0.77	0.9	0.9
	Accuracy Level	71	78	92	93	92	71	84	83
	95% CI [LL, UL]	[63, 78]	[70, 84]	[86, 96]	[87, 96]	[86, 96]	[63, 78]	[77, 90]	[76, 89]
SVM	**F1 Score**	0.86	0.85	0.94	0.95	0.93	0.8	0.76	0.77
	Accuracy Level	77	76	91	92	89	74	69	71
	95% CI [LL, UL]	[69, 83]	[68, 83]	[86, 95]	[86, 96]	[82, 93]	[66, 81]	[61, 77]	[63, 78]

By looking to Table 5.13 it is clear that, on the level of classifiers, the models based on NB give the highest F1 scores that range from 0.84 to 0.95 and outperforming the others. Models based on NN come next with F1 scores range from 0.77 to 0.95. Those models which were based on SVM got F1 scores range from 0.76 to 0.95 and came in the third place while those models which were based on kNN come in the last place with F1 scores range from 0.74 to 0.92. On the level of organizations (datasets), the models which were built based on MIT-2 gave the highest F1 scores with overwhelmingly 0.95 and outperforming the others while those models built which were based on IIT came in the last place with F1 scores range from 0.75 to 0.86. The same for accuracy levels and confidence intervals either with models-level or with datasets-level. In summary, models based on NB and NN with respect to MIT-2 dataset gave the highest performance regarding the F1 score, accuracy level and confidence interval with 0.95, 93 and [LL: 86, UL: 96], respectively. Hence, we can rely on those models in predicting future rising stars with high level of confidence regarding less the used dataset.

To answer **Q7** regarding **What are the most important Scientometric indicators in predicting future rising stars?** (*Which indicator contribute more to the response?*). The best models resulted in previous question, NB and NN with respect to MIT-2 dataset, were used to analyze and find the most important Scientometric indicators in predicting future rising stars. The models were applied on all the eight testing datasets in addition to**ValDS**- validation dataset. The percent of contribution for each correlated indicator was assessed and assigned. Table 5.14 and Table 5.15 summarize the **Relative Importance Analysis (RIA)** of indicators which were given by the best models (NB and NN with respect to MIT-2) for each dataset, respectively. While Figure 5.10 and Figure 5.11 illustrate the values in Table 5.14 and Table 5.15.

Table 5.14: The Relative Importance Analysis for Indicators with NB

Scientometric Indicators	**KFUPM**	**IIT**	**MIT-1**	**MIT-2**	**MIT-3**	**CAS-1**	**CAS-2**	**CAS-3**	**ValDS**	**Average**
Author Productivity Level	0.035	0.017	0.042	0.025	0.032	0.023	0.018	0.027	0.023	0.027
Scientific Impact Level	0.518	0.354	0.481	0.390	0.368	0.367	0.377	0.454	0.452	0.418
Value of Productivity	0.037	0.020	0.045	0.023	0.035	0.030	0.024	0.034	0.027	0.031
Citations Per Publication	0.218	0.428	0.312	0.494	0.458	0.454	0.468	0.342	0.394	0.396
Contribution Impact	0.005	0.002	0.008	0.004	0.006	0.004	0.003	0.004	0.004	0.004
International Collaboration	0.004	0.064	0.006	0.002	0.005	0.005	0.005	0.005	0.004	0.011
Research Area Relevancy	0.178	0.112	0.102	0.060	0.092	0.114	0.102	0.129	0.093	0.109
Venue Reputation	0.005	0.003	0.004	0.002	0.004	0.003	0.003	0.005	0.003	0.004

Table 5.15: The Relative Importance Analysis for Indicators with NN

Scientometric Indicators	**KFUPM**	**IIT**	**MIT-1**	**MIT-2**	**MIT-3**	**CAS-1**	**CAS-2**	**CAS-3**	**ValDS**	**Average**
Author Productivity Level	0.066	0.017	0.426	0.004	0.221	0.106	0.166	0.003	0.127	0.126
Scientific Impact Level	0.168	0.524	0.211	0.337	0.108	0.309	0.354	0.519	0.338	0.319
Value of Productivity	0.116	0.037	0.000	0.004	0.058	0.052	0.088	0.101	0.001	0.051
Citations Per Publication	0.145	0.278	0.259	0.217	0.083	0.196	0.131	0.282	0.163	0.195
Contribution Impact	0.116	0.009	0.000	0.005	0.123	0.038	0.133	0.021	0.192	0.071
International Collaboration	0.057	0.048	0.104	0.000	0.186	0.051	0.000	0.016	0.093	0.062
Research Area Relevancy	0.142	0.075	0.000	0.002	0.221	0.049	0.084	0.015	0.065	0.073
Venue Reputation	0.190	0.012	0.000	0.431	0.000	0.199	0.044	0.043	0.021	0.104

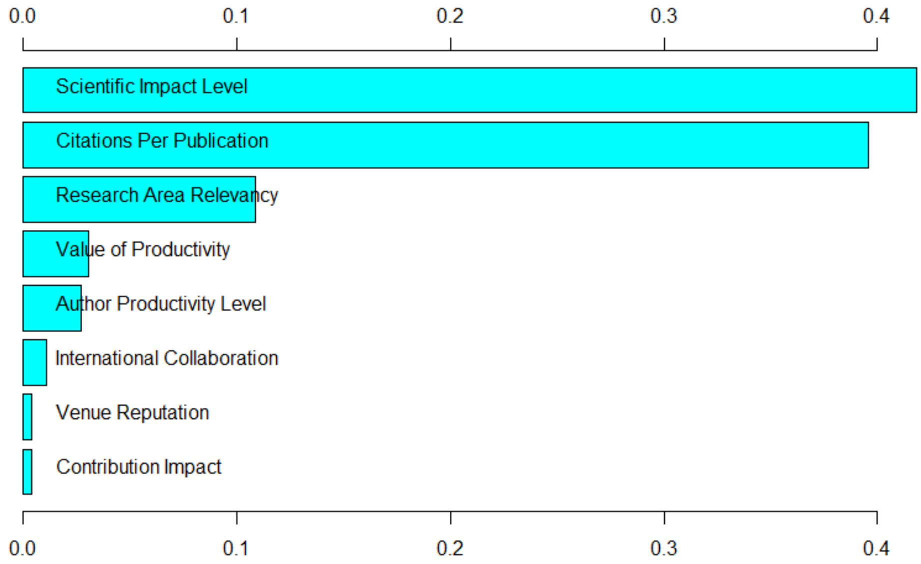

Figure 5.10: The averaged Relative Importance Analysis values for indicators with NB

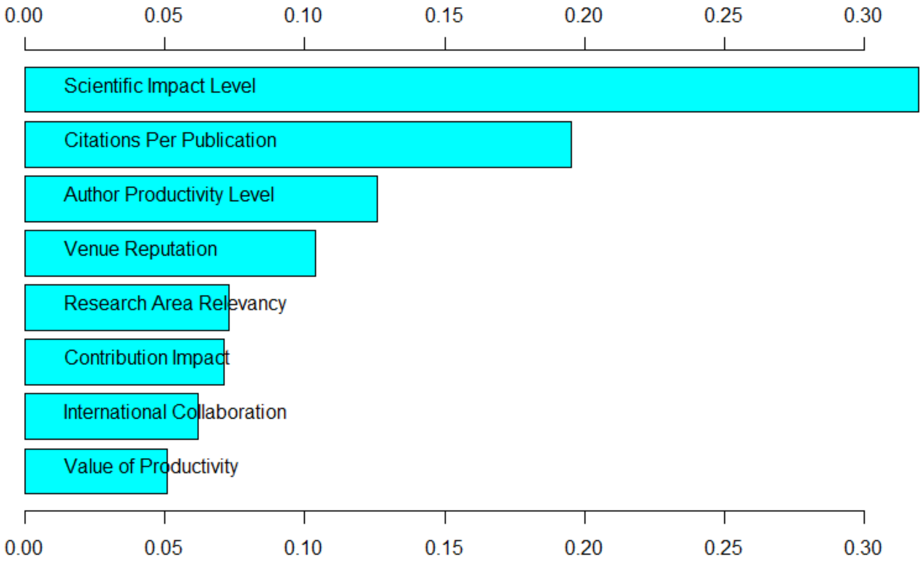

Figure 5.11: The averaged Relative Importance Analysis values for indicators with NN

The overall averaged values of indicators contributions (**Relative Importance Analysis**) is shown in Table 5.16 and depicted in Figure 5.12.

Table 5.16: The Averaged Relative Importance Analysis Values for Indicators with NB and NN

Scientometric Indicators	**RIA Average with NB**	**RIA Average with NN**	**RIA Overall Average**
Author Productivity Level	0.027	0.126	0.076
Scientific Impact Level	0.418	0.319	0.368
Value of Productivity	0.031	0.051	0.041
Citations Per Publication	0.396	0.195	0.296
Contribution Impact	0.004	0.071	0.038
International Collaboration	0.011	0.062	0.037
Research Area Relevancy	0.109	0.073	0.091
Venue Reputation	0.004	0.104	0.054

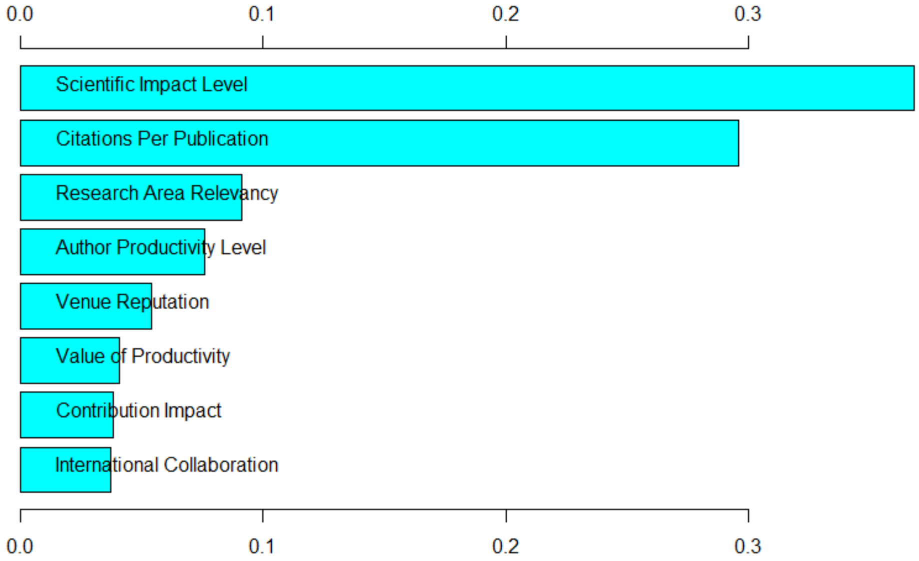

Figure 5.12: The overall averaged Relative Importance Analysis values

From Table 5.16 and Figure 5.12 we can notice that the first three important indicators in predicting future rising stars are *Scientific Impact Level*, *Citations Per Publication* and *Research Area Relevancy* with 0.368, 0.296 and 0.091 all are out of 1, respectively. It is also important to point out that even though the other indicators got less percentages, they are still not neglectable in the prediction process, where this analysis is considered data-model driven. In other words, they are still plays an important role and contributed in the prediction accuracy of the proposed models.

The answer of the last question (**Q8**): **How are our proposed models comparable to other models?**, hard difficult and cannot be achieved straightforward. The reasons behind the difficulty in answering this question can be summarized as follows:

- It is difficult to get the exact data which have been used in building the models proposed by others.
- It remains complicated to reproduce the same feature space (indicators) that was used in other's work.
- It matters to consider the differences from on study to another regarding the definition of rising stars, the adopted ground truth, the applied classifiers, etc.

However, we did compare our models with those of Daud et al. [1]. Daud et al., as mentioned in Chapter 2 of Literature Review, explored eleven features (indicators) with two types of data sets. MEMM, CART, NB and BN were chosen for experiment and results analysis. In Table 5.17 we averaged the F1 scores obtained by applying our proposed models (as detailed in Table 5.13) and the F1 scores obtained by Daud et al. models.

Table 5.17: The F1 Scores of Our Models and Those of Daud et al. [1]

Proposed Models	**F1 Scores**			
Daud et al. [1] Models	**NB**	**MEMM**	**CART**	**BN**
	0.88	0.94	0.91	0.88
Our Models	**NB**	**kNN**	**NN**	**SVM**
	0.92	0.82	0.88	0.86

From Table 5.17, although it seems that the models proposed by Daud et al. [1] overtopped those proposed in our work, however, that can not be relied on by just a quick look at those values. In contrast, if we just consider the F1 score for models based on NB, our proposed model with 0.92 F1 score surpasses that of Daud et al.. However it is important to mentioned that we have to consider the points summarized above in addition to those articulated at the end of Chapter 2 about the advantages of our approach as well as the resulted scores.

5.2 Validity Threats

The approach and methodology followed in this research represent an empirical effort. Thus, it is a quite common practice to face some threats to validity in similar research endeavors. This section discusses the main threats to the validity of the experiment results while the key goal is to minimize those threats.

A common threat to the validity of any machine learning prediction process is the used feature space- indicators, since they may not represent the targeted object very well. In this respect, we therefore carefully formulated a set of Scientometric indicators and used it to observe the evolution of each researcher from multiple perspectives.

Another threat to validity related to those set of Scientometric indicators is the way that values were computed by in addition to generalization insurance. That threat may arise especially with indicators such as Author Productivity and Scientific Impact where they may differ from one organization to another because of the conducted research activities there. Hence, we minimized this threat by considering levels instead of numeric values whilst calculated them based on a relative threshold.

The selection of a certain classification algorithm is likely to have much influence on the resulted model where relying on just one algorithm in building that model is a source of threat. Hence, we tackled this threat by selecting not only just one algorithm, but rather four algorithms were picked-up from the most widely-used and best performance bracket while the superior one in regarding predictability was chosen, accordingly.

It might also be noted that the use of data source may be another threat to validity, since it may not be an accurate nor a reliable source. This issue is discussed extensively in Section 3.2- The Web of Science and the Top 3 of Citation Databases. Therefore, according to what was discussed there the WoS, the comprehensive and trusted source, was the most competent among its competitors.

Dividing the collected datasets into two portions, training and testing datasets, may also be a threat to the validity where building models based on training dataset will required it to be a quite qualified representative. To address that threat, we randomly selected the observations from the original dataset and assigned them to each portion. Another threat related to building models based on training dataset is the over-fitting problem. Therefore, we used 10-fold cross-validation to mitigate that threat.

Another potential threat to validity is the choice of the performance evaluation metrics whereas analysts should not rule out considering only one metric in evaluating the obtained results. Thus, to tackle this threat we considered a wide variety of performance evaluation metrics include confusion matrix, prediction accuracy level, sensitivity, specificity, F1 score, ROC, and AUC.

Finally, there is a threat that the chosen organizations and the opted domains, to conduct the study on their research outcomes, were not qualified representatives for the whole realm of academic research. We minimized this threat in two ways (with the aim to fulfill the requirement of data generalization and balancing). First, we collected data about four organizations that are unlike each other in regards the size and structure due to the differences in research activities. Second, for the collected data, we generated different samples with different sizes. The fact that all datasets are regarding one domain, which is the Computer Science, may introduce another threat to validity. Therefore, we can justify that where it is convenient to compare researchers within the same domain. However, due to the non-similarities in their activates, it is inapplicable to compare together different domains neither there is a guarantee regarding the obtained results- a likely threat in any empirical research.

CHAPTER 6

CONCLUSION AND FUTURE WORK

Conclusion

In this work, machine learning techniques were applied to proposing models aiming to predict future rising star authors in academia, universities and research institutes. At first, four samples of datasets (for four organization KFUPM, IIT, MIT and CAS) were collected form the comprehensive and accurate data source for scientific publications (the WoS) which were used for training and testing tasks. The datasets are like each other regarding the research area and time-period, from 2011 to 2015. Nevertheless, they are different in size and structure due to the difference in research activities in each organization. A novel set of eight Scientometric indicators were formulated and computed for each author to be used as feature space. Throughout this work, a systematic and scientific approach was adopted and followed in deeply investigating the applicability of those proposing models. That was achieved by designing experimental work with aiming at finding answers for eight research questions. The applicability of those indicators was investigated using MLR technique and evaluated in regarding of p-value and F-statistic; the results were promising. NN, kNN, NN and SVM machine learning classifiers were applied for prediction. The predicability were evaluated and compared using a wide variety of performance evaluation metrics include Confusion Matrix, Prediction Accuracy Level, Sensitivity, Specificity, F1 Score, ROC and AUC. Deeply investigation was conducted and extra four datasets were generated, hence we end-up with eight datasets (KFUPM, IIT, MIT-1, MIT-2, MIT-3, CAS-1, CAS-2 and CAS-3). All the models were validated on untouched validation dataset. The validation task results the following: models based on NB and NN with respect to MIT-2 were outperforming others with 0.95 and 93 F1 for score and accuracy level, respectively. Furthermore, indicators' contribution in predicting rising stars was assessed using RIA. Scientific Impact Level, Citations Per Publication and Research Area Relevancy are contributing the most, respectively. Finally, the models in this work were compared against the others from pervious work with taking into account the inapplicability of reusing the exact datasets nor reproducing the same set of indicators in addition to the differences of the used approaches. During answering all the research questions, the obtained results were very auspicious and have shown high predictability whereas all the experiments were carried-out using R.

Future Work

In light of what were discussed in Section 5.2 about threats to validity and aiming for filling the gaps mentioned there, we recommend to extend this work in future with following:

- Ranking rising starts by computing the values of the Scientometric indicators.
- Reformulating some of the Scientometric indicators such as the *International Collaboration*, the *Value of Productivity* and the *Contribution Impact* in addition to use different approaches.
- Collecting datasets about other organizations and other research areas.
- Training with extra classification algorithms.
- Predicting rising starts in academia for long-term future (five years and 10 years in addition to one year).
- Comparing the results of the future work with the results of this work and investigate the differences.

REFERENCES

[1] A. Daud, M. Ahmad, M. Malik, and D. Che, “Using machine learning techniques for rising star prediction in co-author network,” *Scientometrics*, vol. 102, no. 2, pp. 1687–1711, 2015.

[2] G. Tsatsaronis, I. Varlamis, S. Torge, M. Reimann, K. Nørvåg, M. Schroeder, and M. Zschunke, “How to become a group leader? or modeling author types based on graph mining,” *Research and Advanced Technology for Digital Libraries*, pp. 15–26, 2011.

[3] G. James, D. Witten, and T. Hastie, “An introduction to statistical learning: With applications in r.” 2014.

[4] P. B. Lowry, D. Romans, and A. M. Curtis, “Global journal prestige and supporting disciplines: A scientometric study of information systems journals,” 2004.

[5] V. Nalimov and Z. Mulchenko, “Naukometriya. izuchenie razvitiya nauki kak informat-sionnogo protsessa [scientometrics. the study of science development as an information process],” 1969.

[6] J. Mingers and L. Leydesdorff, “A review of theory and practice in scientometrics,” *European Journal of Operational Research*, vol. 246, no. 1, pp. 1–19, 2015.

[7] J. Tague-Sutcliffe, “An introduction to informetrics,” *Information processing & management*, vol. 28, no. 1, pp. 1–3, 1992.

[8] C. Chua, L. Cao, K. Cousins, and D. W. Straub, “Measuring researcher-production in information systems,” *Journal of the Association for information Systems*, vol. 3, no. 1, p. 6, 2002.

[9] G. Grégoire, “Multiple linear regression,” *European Astronomical Society Publications Series*, vol. 66, pp. 45–72, 2014.

[10] D. D. Lewis, “Naive (bayes) at forty: The independence assumption in information retrieval,” in *European conference on machine learning.* Springer, 1998, pp. 4–15.

[11] K. Fukunaga and P. M. Narendra, “A branch and bound algorithm for computing k-nearest neighbors,” *IEEE transactions on computers*, vol. 100, no. 7, pp. 750–753, 1975.

[12] K.-I. Funahashi, “On the approximate realization of continuous mappings by neural networks,” *Neural networks*, vol. 2, no. 3, pp. 183–192, 1989.

[13] V. N. Vapnik and V. Vapnik, *Statistical learning theory.* Wiley New York, 1998, vol. 1.

[14] T. Reuters, “Web of Science.” 2012.

[15] X.-L. Li, C. S. Foo, K. L. Tew, and S.-K. Ng, “Searching for rising stars in bibliography networks,” in *International Conference on Database Systems for Advanced Applications.* Springer, 2009, pp. 288–292.

[16] L. Page, S. Brin, R. Motwani, and T. Winograd, “The pagerank citation ranking: Bringing order to the web.” Stanford InfoLab, Tech. Rep., 1999.

[17] A. Daud, R. Abbasi, and F. Muhammad, “Finding rising stars in social networks,” in *International Conference on Database Systems for Advanced Applications.* Springer, 2013, pp. 13–24.

[18] C. H. Sekercioglu, “Quantifying coauthor contributions,” *Science*, vol. 322, no. 5900, p. 371, 2008.

[19] J. Zhang, F. Xia, W. Wang, X. Bai, S. Yu, T. M. Bekele, and Z. Peng, “Cocarank: a collaboration caliber-based method for finding academic rising stars,” in *Proceedings of the 25th International Conference Companion on World Wide Web.* International World Wide Web Conferences Steering Committee, 2016, pp. 395–400.

[20] J. M. Kleinberg, “Authoritative sources in a hyperlinked environment,” *Journal of the ACM (JACM)*, vol. 46, no. 5, pp. 604–632, 1999.

[21] K. Miettinen, “Introduction to multiobjective optimization: Noninteractive approaches.” *Multiobjective Optimization*, vol. 5252, pp. 1–26, 2008.

[22] S. Lin, “Rank aggregation methods,” *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 5, pp. 555–570, 2010.

[23] P. Wijegunawardana, K. Mehrotra, and C. Mohan, “Finding rising stars in heterogeneous social networks,” in *2016 IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI)*, Nov 2016, pp. 614–618.

[24] P. Domingos, “A few useful things to know about machine learning,” *Communications of the ACM*, vol. 55, no. 10, pp. 78–87, 2012.

[25] R. Xu and D. Wunsch, “Survey of clustering algorithms,” *IEEE Transactions on neural networks*, vol. 16, no. 3, pp. 645–678, 2005.

[26] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim, “Do we need hundreds of classifiers to solve real world classification problems,” *J. Mach. Learn. Res*, vol. 15, no. 1, pp. 3133–3181, 2014.

[27] S. Michiels, S. Koscielny, and C. Hill, “Prediction of cancer outcome with microarrays: a multiple random validation strategy,” *The Lancet*, vol. 365, no. 9458, pp. 488–492, 2005.

[28] G. K. Tso and K. K. Yau, “Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks,” *Energy*, vol. 32, no. 9, pp. 1761–1768, 2007.

[29] M. Kubat, R. C. Holte, and S. Matwin, “Machine learning for the detection of oil spills in satellite radar images,” *Machine learning*, vol. 30, no. 2-3, pp. 195–215, 1998.

[30] M. A. Alsheikh, S. Lin, D. Niyato, and H.-P. Tan, “Machine learning in wireless sensor networks: Algorithms, strategies, and applications,” *IEEE Communications Surveys & Tutorials*, vol. 16, no. 4, pp. 1996–2018, 2014.

[31] T. Helmy, S. Al-Azani, and O. Bin-Obaidellah, “A machine learning-based approach to estimate the cpu-burst time for processes in the computational grids,” in *Artificial Intelligence, Modelling and Simulation (AIMS), 2015 3rd International Conference on.* IEEE, 2015, pp. 3–8.

[32] I. Rish, “An empirical study of the naive bayes classifier,” in *IJCAI 2001 workshop on empirical methods in artificial intelligence*, vol. 3, no. 22. IBM, 2001, pp. 41–46.

[33] H. Zhang, “The optimality of naive bayes,” *AA*, vol. 1, no. 2, p. 3, 2004.

[34] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft, “When is ”nearest neighbor” meaningful?” in *International conference on database theory.* Springer, 1999, pp. 217–235.

[35] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, S. Y. Philip *et al.*, “Top 10 algorithms in data mining,” *Knowledge and information systems*, vol. 14, no. 1, pp. 1–37, 2008.

[36] F. Günther and S. Fritsch, “neuralnet: Training of neural networks,” *The R journal*, vol. 2, no. 1, pp. 30–38, 2010.

[37] G. P. Zhang, "Neural networks for classification: a survey," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 30, no. 4, pp. 451–462, 2000.

[38] V. Vapnik, *The nature of statistical learning theory.* Springer science & business media, 2013.

[39] S. B. Kotsiantis, I. Zaharakis, and P. Pintelas, "Supervised machine learning: A review of classification techniques," 2007.

[40] U. of Waikato, "Weka 3: Data mining software in java," http://www.cs.waikato.ac.nz/ml/weka/, retrieved: 2016-4-1.

[41] MathWorks, "The language of technical computing," http://www.mathworks.com/products/matlab/, retrieved: 2016-4-1.

[42] R. C. Team, "The R project for statistical computing," https://www.r-project.org/, retrieved: 2016-4-1.

[43] E. Garfield, "Citation indexes for science. a new dimension in documentation through association of ideas," *International journal of epidemiology*, vol. 35, no. 5, pp. 1123–1127, 2006.

[44] ELSEVIER, "Scopus comes of age," https://www.elsevier.com/about/press-releases/science-and-technology/scopus-comes-of-age, 2004, retrieved: 2016-4-1.

[45] J. Bar-Ilan, “Which h-index?A comparison of WoS, Scopus and Google Scholar,” *Scientometrics*, vol. 74, no. 2, pp. 257–271, 2008.

[46] D. Payne, “Google scholar welcomed,” http://www.the-scientist.com/?articles.view/articleNo/23178/title/Google-Scholar-welcomed, 2004, retrieved: 2016-4-1.

[47] J. Bar-Ilan, “Citations to the introduction to informetrics indexed by wos, scopus and google scholar,” *Scientometrics*, vol. 82, no. 3, pp. 495–506, 2010.

[48] M. Norris and C. Oppenheim, “Comparing alternatives to the web of science for coverage of the social sciences literature,” *Journal of informetrics*, vol. 1, no. 2, pp. 161–169, 2007.

[49] K. Bauer and N. Bakkalbasi, “An examination of citation counts in a new scholarly communication environment,” *D-Lib magazine*, 2005.

[50] T. Reuters, “Incites indicators handbook,” http://researchanalytics.thomsonreuters.com/m/pdfs/indicators-handbook.pdf, 2014, retrieved: 2016-4-1.

[51] R. V. Krejcie and D. W. Morgan, “Determining sample size for research activities,” *Educational and psychological measurement*, vol. 30, no. 3, pp. 607–610, 1970.

[52] H. Kretschmer, “Author productivity and geodesic distance in bibliographic co-authorship networks, and visibility on the web,” *Scientometrics*, vol. 60, no. 3, pp. 409–420, 2004.

[53] L. Bornmann and H.-D. Daniel, “What do citation counts measure? a review of studies on citing behavior,” *Journal of documentation*, vol. 64, no. 1, pp. 45–80, 2008.

[54] D. W. Aksnes, “Characteristics of highly cited papers,” *Research evaluation*, vol. 12, no. 3, pp. 159–170, 2003.

[55] D. Remler, “Are 90% of academic papers really never cited? reviewing the literature on academic citations..” *Impact of Social Sciences Blog*, 2014.

[56] L. Waltman, “A review of the literature on citation impact indicators,” *Journal of Informetrics*, vol. 10, no. 2, pp. 365–391, 2016.

[57] J. Xu, Y. Ding, M. Song, and T. Chambers, “Author credit-assignment schemas: a comparison and analysis,” *Journal of the Association for Information Science and Technology*, vol. 67, no. 8, pp. 1973–1989, 2016.

[58] N. T. Hagen, “Harmonic allocation of authorship credit: Source-level correction of bibliometric bias assures accurate publication and citation analysis,” *PLoS One*, vol. 3, no. 12, p. e4021, 2008.

[59] C. S. Wagner and L. Leydesdorff, “Network structure, self-organization, and the growth of international collaboration in science,” *Research policy*, vol. 34, no. 10, pp. 1608–1618, 2005.

[60] O. Persson, W. Glänzel, and R. Danell, “Inflationary bibliometric values: The role of scientific collaboration and the need for relative indicators in evaluative studies,” *Scientometrics*, vol. 60, no. 3, pp. 421–432, 2004.

[61] A. Page, "Best practices memo evaluating computer scientists and engineers for promotion and tenure," *Computing*, 1999.

[62] K. Max, "Building predictive models in r using the caret package," vol. 28, 11 2008.

[63] G. C. Cawley and N. L. Talbot, "On over-fitting in model selection and subsequent selection bias in performance evaluation," *Journal of Machine Learning Research*, vol. 11, no. Jul, pp. 2079–2107, 2010.

[64] T. Fawcett, "An introduction to roc analysis," *Pattern recognition letters*, vol. 27, no. 8, pp. 861–874, 2006.

[65] A. P. Bradley, "The use of the area under the roc curve in the evaluation of machine learning algorithms," *Pattern recognition*, vol. 30, no. 7, pp. 1145–1159, 1997.

[66] M. H. Zweig and G. Campbell, "Receiver-operating characteristic (roc) plots: a fundamental evaluation tool in clinical medicine." *Clinical chemistry*, vol. 39, no. 4, pp. 561–577, 1993.

[67] R. Kohavi and F. Provost, "Glossary of terms," *Machine Learning*, vol. 30, no. 2-3, pp. 271–274, 1998.

[68] R. A. Fisher, "Xv.the correlation between relatives on the supposition of mendelian inheritance." *Earth and Environmental Science Transactions of the Royal Society of Edinburgh*, vol. 52, no. 2, pp. 399–433, 1919.

Vitae

- Name: Omar Jafer Bin Obaidellah
- Nationality: Yemen
- Date of Birth: September 3, 1983
- Email: *almdodi@gmail.com*
- Permenant Address: Dhahran, Eastern Region, Saudi Arabia

- Academic Background: Earned the degree of Bachelor of Science (B.Sc.) in Computer Science form Al-Ahgaff University. Currently, pursuing the degree of Master of Science (M.Sc.) in Computer Science at King Fahd University of Petroleum & Minerals. Bin Obaidellah academic interests include Machine Learning, Data Mining, Data Analysis and Big Data.