CAPITAL UNIVERSITY OF SCIENCE AND TECHNOLOGY, ISLAMABAD

# Feature Selection For Author Assessment

by

Muhammad Abuzar

A thesis submitted in partial fulfillment for the degree of Master of Science

in the

Faculty of Computing

Department of Computer Science

2024

*My dissertation work is devoted to My Family, My Teachers, and My Friends. I have a special feeling of gratitude for My beloved parent my brother and my sisters. Special thanks to my supervisor whose uncountable confidence enabled me to reach this milestone.*

# CERTIFICATE OF APPROVAL

## Feature Selection For Author Assessment

by

Muhammad Abuzar

(MCS193024)

### THESIS EXAMINING COMMITTEE

| S. No. | Examiner | Name | Organization |
|---|---|---|---|
| (a) | External Examiner | Dr. Ayyaz Hussain | QAU, Islamabad |
| (b) | Internal Examiner | Dr. Mohammad Masroor Ahmad | CUST, Islamabad |
| (c) | Supervisor | Dr. Abdul Basit Siddiqui | CUST, Islamabad |

Dr. Abdul Basit Siddiqui
Thesis Supervisor
January, 2024

Dr. Abdul Basit Siddiqui
Head
Dept. of Computer Science
January, 2024

Dr. M. Abdul Qadir
Dean
Faculty of Computing
January, 2024

# *Author's Declaration*

I, **Muhammad Abuzar** hereby state that my MS titled "**Feature Selection for Author Assessment**" is my own work and has not been submitted previously by me for taking any degree from Capital University of Science and Technology, Islamabad or anywhere else in the country/abroad.

At any time if my statement is found to be incorrect even after my graduation, the University has the right to withdraw my MS Degree.

**(Muhammad Abuzar)**

Registration No: MCS193024

# *Plagiarism Undertaking*

I solemnly declare that research work presented in this thesis titled "**Feature Selection For Author Assessment**" is solely my research work with no significant contribution from any other person. Small contribution/help wherever taken has been duly acknowledged and that complete thesis has been written by me.

I understand the zero tolerance policy of the HEC and Capital University of Science and Technology towards plagiarism. Therefore, I as an author of the above titled thesis declare that no portion of my thesis has been plagiarized and any material used as reference is properly referred/cited.

I undertake that if I am found guilty of any formal plagiarism in the above titled thesis even after award of MS Degree, the University reserves the right to withdraw/revoke my MS degree and that HEC and the University have the right to publish my name on the HEC/University website on which names of students are placed who submitted plagiarized work.

**(Muhammad Abuzar)**
Registration No: MCS193024

# *Acknowledgement*

Say, ”He is Allah, [who is] One. Allah, the Eternal Refuge. He neither begets nor is born. Nor is there to Him any equivalent.” Al-Quran [112:1-4]. I would like to say Alhamdulillah, for blessing me, with the strength to finish this work. My heartfelt thanks to my mother for her love, prayers, and everything that I need. A special thanks to my father. Last but not least are my friends, with whom this journey becomes a garden full of flowers for me. Special thanks to my esteemed supervisor Dr. Abdul Basit for the support, for his assistance,inspiration, and guidance in the field of research. I have learned not only research under his supervision but also his attitude towards others. Sir ”You are one of those persons in my life, whom I love to follow”. May Allah shower his countless blessing upon all of you, JazakAllah.

**(Muhammad Abuzar)**

# *Abstract*

Reputable international scientific societies recommended candidates for awards each year basedon qualitative evaluations. Qualitative judgment makes subjective judgments based on non quantifiable information. It has not been established or made public how to evaluate the quality of the work, and we do not have any technique for the evaluation of qualitative research. The scientific communities have offered more than 50 quantitative research evaluation parameters so far, including the g-index, h-index, h-index variants, citation count, and publication count. The parameter that can accurately maps experts/researchers'/authors' qualitative evaluation is not determined by the most recent state-of-the-art in the author's assessment. It might be quite difficult to determine the importance of each metric relative to the others in such situations. The dataset in which award winners are 250 and non-award winners are 250 from scientific societies of civil engineering. This study aims to pinpoint the parameter patterns for the discipline of civil engineering that have been primarily used by those who have won such awards. The lists of the top researchers against the top-selected parameters have been used by the Logistic Regression, Vector Machine, Naïve Bayes Support, and Decision Tree to assess the pattern of winners, and got the results of 97% awardees through Logistic Regression.We also applied the correlation method and select parameters that are weakly and moderately correlated with each other. It is observed that 10 features are weakly and moderately correlated and after the evaluation of these parameters, It is found that the occurrence of awardees increased from 67% to 97% percent. Moreover, the experiments are done on the parameters by using wrapper method and find the impact of inclusion of parameter who have strong and very strong correlation. After some experiments it is found that by the inclusion of variables which have correlation count is 1. Here are three parameters which have correlation count is 1 h_coverage, year_first and AR_index. By including h_coverage which is highly positive correlated to g_coverage with 0.93 then accuracy dropped from 97% to 90%. Additionally examination and assessment is done that by including those parameters how have high correlation with other 4 to 5 parameter then accuracy also reduced.On the basis of these experiments, a conclusion is

drown that if parameter or parameters have the correlation greater with the other parameter then accuracy for awardees and non-awardees reduces.

# Contents

# List of Figures

# List of Tables

# Abbreviations

| | |
|---|---|
| **ACI** | American Concrete Institute |
| **ASCE** | American Society of Civil Engineering |
| **CSCE** | Canadian Society for Civil Engineering |
| **ICE** | Institute of Civil Engineering |

# Chapter 1

# Introduction

To find the quality of a researcher's scientific research publication or impact is an utmost task. Its importance can be estimated as it is an utmost demanding task. As we know the quality of the researcher's impact cannot be described very easily, that's why its complexity is very high. The judgment of the quality of research's impact is based on subjective evaluations of non-quantifiable information, one of the examples is research strength. The assessment of authors associated with different fields has become the latest research area.

The author's assessment has various advantages, it can be helpful to find the researcher who can be nominated for awards. Moreover, it can be helpful for journals/organizations to find suitable reviewers for the evaluation of publications. In the past few years, researchers have been nominated for the award based on qualitative judgment [24], which is not quantifiable. We do not have any strong assessment criteria to identify the quality of the researcher's work. There are many assessment techniques such as page Rank on the author citation graph and page Rank on the paper citation graph [12] to nominate the researchers for awards.

The qualitative judgment allows us to define the queries and answer i) who is eligible for a scholarship award? ii) Who should be the editor and article reviewer for scientific journals and conferences? iii) Who should award membership and fellowship to a scientific research society? iv) Who should help the education institutes to hire the staff? v) Who is more competent to become a fellow or a member of the scientific society? Every year the scientific society nominated the award winner based on qualitative assessment.As it is discussed and found the pros and cons of the parameters in the literature review that any parameter can not be used to estimate the quality of the researcher work. It is observed that there is no criteria or any author assessment parameters by which quality of work can be analyse. However, to identify the researcher's research work, many quantitative measures or parameters are proposed.The quantitative parameters are measured based on numeric values. To date, several researchers/authors assessment techniques are proposed in the literature. In this way, all the assessment techniques are evaluated on different assessment parameters, and the researchers/authors are assessed qualitatively through quantitatively judgment or by using author/researchers assessment quantitative features[6]. Scientists have been working on approaches to evaluate the author's work impact in the scientific community. In each approach, the authors/researchers are grouped based on the strength and ability of the new research approach. So, in the literature, several quantitative assessment measures are proposed, and the list of these measures is growing over time very rapidly. All the techniques which have been proposed are based on their publication and citations.

## 1.1 Background

The traditional method of assessment of an author's research work is publication count [36], according to this technique; those researchers nominated for the award have a greater number of publications. But this technique does not work properly and is not universally right, because an author who has few publications may be impactful as compared to an author who has many publications. This can be explained by Cameron with the help of an example [9]. Cameron selected two authors from google scholar the field of these two researchers is a database, in these two authors one has more research publications while the other has a small list of publications. Of these two researchers, one is 'E. F. codd', E.F Codd invents the relational database. The other is 'Hector Garcia Molina'.

The first one has only 49 publications, while the latter has 248 research publications. 'E. F. codd' is regarded as more productive compared to 'Molina', because he won the two Turing award (1981, 1994). But assessment of the researcher's work based on their research publications, the above example shows that 'Hector Garcia Molina' is more productive because he has more research publications. The scientific community could be severely harmed in the scenario of publication count. Therefore, to overcome this problem, another technique, citation count was proposed [26]. According to this technique, an author should be awarded based on citations of papers. Total number of citations a researcher gains is regarded as the researcher is highly profiled. Again, also this technique does not work properly or is inefficient as i) sometimes newly published papers require a lot of time to receive citations. ii) The authors can also increase their citations illegally. iii) The authors may cite the paper for the sake of criticism. iv) Typically, the survey paper has more citations. It is observed that essential issues of the academic community can be address with the qualitative appraisal of researchers' work, as mentioned in [6]. In 2005, Jorge Hirsh proposed a technique in which the number of citations and publications were included and named this technique h-index [18]. This method of assessing the researcher's research work became more popular because it can allow the research community to explore new fields of 3 research. In this technique, there is a point where the number of citations and publications is equal, after that point the very next value is called their h-index. it is found researcher's work can be assess through the h-index. The h-index is the parameter that can measure the current performance of the researcher as well as it can also give the prediction for future research work.

The h-index is a quantitative parameter for the researcher's impact. The h-index resolved several problems regarding the researcher's work evaluation. The world is using H-index for the evaluation and it is the most widely accepted scientific norm. It is the most popular parameter that makes computing reasonably simple. The original h-index publication had 9314 citations till December 19, 2020. According to Hirsh, the h-index will perfectly indicate that a researcher has received the prominent award like the presidency award of National Academy of Sciences or the Nobel Prize. Hirsh calculated h-index for the top ten scientists from the bioscience field, and they found a higher h-index for authors cited [20]. However, there are several limitations of h-index. One of them is, h-index never entertained the increasing number of citations for index publications and does not boost the impact of researcher's work [5, 11].

The limitations of h-index motivated the researchers, in that case many author assessment parameters were developed by the researchers that can meetup the issues and weaknesses of h-index, for assessing the researcher's work. In this way, many researchers have developed many indices some are listed R-index [20], Ar-index [20], g-index [14] , hg-index [2] and A-index [19] etc. These assessment features have dependency over the research publications, their citations, and academic research age. These parameters were categorized into three groups, Primitive , Citation intensity-base and age-base features [38]. There is no effective standard for evaluating a researcher's influence or impact. By deeply analyzing these parameters it is came to know that it is need of the time to develop a technique that defines when the research work of the researcher is determined, the generated data is more accurate. Previous research [36] has used prize winners from scientific societies as the benchmark for measuring how well these parameters perform. They rated the researchers based on the results of these assessment parameters and manually calculated award winner's number in the ranked list [1, 29]. However, it is still unclear which best effective assessment parameter is based on expert qualitative judgment.

To figure out which parameter is the most effective for assessment of authors, these parameters should be weighed using advanced machine learning techniques on datasets from many fields of study to produce the best results. Furthermore, the worth of such parameters must be identified, which can be regarded as the minimum eligibility criteria for award 4 nomination. Thus, the current study aims to select the best feature from primitive parameters, citations-intensity-based, and academic research age-based parameters and apply classifier to predict the award winners from the four civil engineering societies. In these three types of parameters the preemptive parameters include publications count, years count in the field of research, number of citations, Citation/Paper, Author/Paper, h-index and h-core citation. Citation-intensity based features includes hg-index, g-index, R-index, p-index, A-index, E-index, K-index, q2-index, F-index and hm-index. Age-based features includes hl-norm, hc-index, AR-index, M-quotient, AW-index and hi-index [28].

The dataset consists of 250 members of the award winner and 250 members are the non-award winner from the civil engineering field. The civil engineering field has different societies that are used for assessment of the current technique, Institute of Civil Engineering (ICE), American Society for Civil Engineering (ASCE), Canadian Society for Civil Engineering) and American Concrete Institute (ACI). The goal of this study is to discover the key factors that influence worldwide scientific societies for award nominations; for this purpose, the researcher's profiles are derived from the Google Scholar citations repository [3]. Based on different criteria we are going to propose and validate a framework for author assessment parameters. First, the data is collected from 2011 to 2019 to meet the goal of this research, the data is collected from the websites of all above mentioned civil engineering societies. After that, the citation-intensity-based, primitive, and age-based parameters were calculated. When all is done, different techniques will be applied to find out the best feature which have less dependency and will apply classifier to predict the awardees. In the end, proposed a solution on the author's assessment parameters to define the eligibility criteria of the award winner in the field of civil engineering society.

According to Hirsh, the h-index will perfectly indicate that the author has received a prominent award such as the presidency award of National Academy of Sciences or the Nobel Prize. Hirsh computed h-index for the top ten scientists from the bio-science field, and they found a higher h-index for authors cited. To figure out which parameter or combination of parameter is the most effective for the evaluation of authors, these parameters should be used in the advanced machine learning techniques on datasets from many fields of study to produce the best results.

## 1.2 Research Gap

Today scholarly data is increasing rapidly. With the increase of scholarly data, it is difficult to assess the researcher's most relevant research work for the decision-making of awards on the bases of given parameters. Every year the scientific society nominated the award winner based on qualitative assessment. The qualitative assessment uses subjective judgment which is not quantifiable. There are many author assessment parameters which have been introduced and calculated till now. Some of the parameters are discussed in the literature review. These parameters are currently focused to find the quality of Researcher research work.In Previous Study [36], it is observed that Prediction is done on the all author assessment parameters by using Logistic regression and found the feature importance by using wrapper method. This techniques figured out the importance of only one parameter not the combination of parameters.It is required to find out the most important author assessment parameter or parameters to evaluate or assess the quality of author's evaluation parameters.

## 1.3 Problem Statement

The subjective evaluation that supports qualitative judgement is based on non-quantifiable information. [24]. In the literature review, it is observed that it is difficult to assess the author's/research's work on the bases of quantitative information or author assessment parameters. In the meanwhile, mapping the quantitative features onto the qualitative parameters is also challenging which is fundamental problem that how to find the quality of the work through quantitative parameters.

## 1.4 Scope

The scope of this thesis is the evaluation and selection of the author's assessment parameters in the broader field of civil engineering. The age-based, preemptive, and citation-intensity-based parameters include, i) citations ii) number of publications iii) years in research iv) citation/year v) citations/paper vi) author/paper vii) h-core citations viii) h-index xi) hg-index x) g-index xi) eindex xii) a-index xiii) r-index xiv) f-index xv) q2-index xvi) k-index xvii) hc-index xviii) ARindex. this study will make use of the dataset from the field of civil engineering to determine the correlation between the aforementioned factors in this domain. This research examines which parameter or parameters contributes most to the assessment of national and international award winners. As a result of this technique, it will give the award wining researchers from a specific field. To figure out which parameter is the most effective for assessment of authors, these parameters should be weighed using advanced machine learning techniques on datasets from many fields of study to produce the best results.

## 1.5 Research Question

- Which of the primitive, age-based, and citation intensity base author assessment parameters contributes/contribute most effectively to the award-winning of the scientific society?

- Which of the classifier perform better based on the selected features which are used for author ranking assessment?

## 1.6 Research Objective

This study aims to map the quantitative (preemptive, citation-intensity based, and research age-based) authors assessment parameters to qualitative perception in the civil engineering field. It is also a type of finding the correlation between the quantitative parameters and qualitative parameters. In this way, a comprehensive dataset is considered. In literature, most of the studies considered the awardees as a benchmark for the evaluation of the author's assessment parameters [22]. The dataset which is under consideration contains 500 researchers belonging to the scientific societies of civil engineering i.e., Institute of Civil Engineering, American Concrete, American Society of Civil Engineering, and Institute Canadian Society of Civil Engineering. Of these 500 researchers,' there are 250 researchers are award winners, and 250 are non-awardees. These assessment parameters could be implemented by different techniques. But although the methodology which is going to adopt is quite different, in this way selection of the author assessment parameters will done first and then will apply classifier.

# Chapter 2

# Literature Review

## 2.1 Introduction

Introduction every year the scientific societies conduct a ceremony to honor the award recipients based on their contribution to the scientific society. Different societies honor their researchers in different ways. For example, the award can be given to a researcher in the form of selecting them as research article reviewers or as an editor of the journal [22]. Determining the influence of researchers is one of the major issues in their research work in the scientific community. Another justification for evaluating the researcher's research impact is to provide him with a postdoctoral post or they should hand over the supervision of the funded project. Various approaches have been introduced in the literature to influence academics, research groups, journals, or universities based on their research impact on the scientific community. Moreover, it is quite difficult to discriminate among researchers, i) one who makes frequent contributions to the scientific community and who publishes many research articles every year. ii) The researcher who does not actively publish the research article but published very few articles with new research ideas [4]. Based on citations, gained by the research articles the two authors Lercher and Smolinsky addressed the influence of several researchers in the domain of mathematics. They claimed that the inequality in citations mostly depends on the published articles and how these articles are cited internally [34]. The h index allows to assess a single researcher's work output. Although h-Index is essentially finite, its basic and uncomplicated application is common in the scientific world [18]. There are many limitations of h-index, one of the limitations of h-index is as it does not accomplish the research career of the researchers. Costas and Bordons define a researcher with a very short research career, because of their short career he has a small number of research publications and these publications gained very fewer citations. In this way, the researcher needs more time to influence the research community that causes the score of the h-index is low. As a result, a parameter that incorporates a better researcher evaluation is required [10]. To address these deficiencies, researchers have proposed many additional h-index variants. The remaining part of this section explained the preemptive parameters like gained citations, the total number of research publications, Citation/Paper, research career of a researcher, Citations/Year, Authors/Paper, Citations in h-index and h-core. Citations intensity-based features like, hg-index, g-index, p-index, E-index, R-Index, q2-index, A-Index, K-index, hm-index and F-index. Age-based parameters like AR-index, hc-index, M-quotient, hg-index and AW-index [28].

## 2.2 Primitive Parameters

The author/paper, cites/paper, cites/year, and h-index, etc. are types of primitive parameters in which researchers can be assessed by the values of these parameters. Cites/paper means the total number of citations divided by total number of papers of that researcher. The h-index is explained below [26].

### h-index

In example 2.1, a researcher having 6 publications along with some citations count is given in decreasing way gained by these publications. As earlier, it is found that the number at which the publication index and total number of citations at that publication are equal is called the h-index. Here the researcher's h-index is 5 in table given table. The h-index can be computed as

| **Publications** | **Citations** |
|---|---|
| 1 | 5 |
| 2 | 4 |
| 3 | 8 |
| 4 | 3 |
| 5 $h - index(5)$ | 6 |
| 6 | 3 |

Table 2.1: h-index

## 2.3 Citations Intensity Base Parameters

The publications and citations of those publications that aren't part of h-core are passed over by the h-index. To overcome the limitation of h-index citiation intensity-based parameters were developed like g-index [14]. The citation intensity-based parameters are those parameters that can minimize the loss of information in the h-index.

### g-index

To overcome the limitation of the h-index the g-index is proposed by Egghe. According to g-index the publication count is either equal to the citations count or the sum of citations is greater than the square of the index number of publication count [2].

| **Publications** | **Citations** | **g2** | $\sum$ |
|---|---|---|---|
| 1 | 18 | 1 | 18 |
| 2 | 9 | 4 | 27 |
| 3 $h - index(3)$ | 4 | 9 | 31 |
| 4 | 0 | 16 | 31 |
| 5 $h - index(5)$ | 1 | 25 | 32 |
| 6 | 0 | 36 | 32 |

Table 2.2: g-index

Table 2.2 has a practical example of the g-index. It has 5 research publications along with their citations in descending order. The data shows that g-index is 5 because the square of index number of publications is lesser than the sum of citations is greater. The formula of g-index is given in equation 2.1.

$$g - index = Sum\ of\ citations\ count >= square\ of\ publications \tag{2.1}$$

### hg-index

The hg-index of the researcher is find by calculated the square root of the (h*g) [2]. We can see that g-index is 5 and h-index is 3 so the hg-index = 3.87. The hg-index can be computed by the formula given in equation 2.2.In equation 2.2, g represent the g-index and h represent the h-index.

$$hg - index = \sqrt{h * g} \tag{2.2}$$

### A-index

To compute A index, first calculate the g-index of a researcher and then compute their cumulative sum of citations against their publications. The A-index is computed by dividing the cumulative sum of citations from their h-index as shown in table 2.3.

| **Publications** | **Citations** | **Citation in h-core** |
|---|---|---|
| 1 | 10 | 10 |
| 2 | 9 | 19 |
| 3 | 8 | 27 |
| 4 | 7 | 34 |
| 5 | 7 | 41 |
| 6 $h - index(6)$ | 6 | 47 $A - index = 47/6 = 7.83$ |

Table 2.3: A-index

In table we have a researcher X who has seven publications and gained their citations. H-index is 6 to find their A-index compute their citations in h-core and then divide h-core citations to the h-index. Here it can be seen that the A-index of the researcher X is 7.83 as shown in table 2.3. The A-index can be calculated as shown in the equation 2.3

$$A - index = \frac{\sum_{j=1}^{h} cit_{\text{j}}}{h} \tag{2.3}$$

In equation 2.3, $\sum$ cit represents the sum of citations, and h represents the h-index.

**R-index**

The R-index employs the h-index as a divisor, as the A-index gives high h-index researchers the lowest rank. The researcher's A-index decreases as his h-index increases. Therefore, it has been suggested by researchers that calculating the square root of the total citation count rather than dividing it by h-index will lead to greater performance. R-index similar to a-index which only considers the number of h-core citations. Which will lead it extremely vulnerable to even a few highly cited papers. A-index separates the h-index in the way that A-index considers the paper that has a higher h-index discussed by Jin et al. Since the estimation function of R-index is identical to the A-index. If a paper has numerous citations the A-index has the higher influence on it. Table 2.4 determines R-index of a researcher who has 7 publications and multiple citations.

| **Publications** | **Citations** | **Citation in h-core** |
|---|---|---|
| 1 | 10 | 10 |
| 2 | 9 | 19 |
| 3 | 8 | 27 |
| 4 | 7 | 34 |
| 5 | 6 | 40 |
| 6 $h - index(6)$ | 6 | 46 $R - index = \sqrt{46} = 6.78$ |
| 7 | 2 | 48 |

Table 2.4: R-index

The R-index is represented by the equation 2.4.

$$R - index = \sqrt{\sum_{j=1}^{h} cit_{\mathrm{j}}} \tag{2.4}$$

In the given equation 2.4, $\sum$ cit represents the sum of citations and h represents the h-index.

### E-index

The e-index is another assessment parameter, the e-index also indicates the citations count in h-core citations, and these h-core citations are discarded by h-index [40]. E-index indicates that h-core has information loss by the h-index. This information loss will be 0, if the h-index handle and include the total number of citations in the h-core, that's why e-index value will be zero which is given in the table 2.5.

The E-index is represented by the equation 2.5.

$$E - index = \sqrt{d^2 - h^2} \tag{2.5}$$

In the given equation 2.5, "h" is the h-index and "d" is the number of citations in the h-core.

| Publications | Citations | Citation in h-core |
|---|---|---|
| 1 | 8 | 10 |
| 2 | 5 | 19 |
| 3 | 4 | 27 |
| 4 | 3 | 34 |
| 5 $h - index(5)$ | 5 | 40 $e - index = \sqrt{25 - 25} = 0$ |

Table 2.5: E-index

### F-index

F-index shows the fold of h2 citations for publications in h-core. It uses the fractional system of counting citations. To calculate the f-index we have already calculated the value of e-index and h-index. Divide the e-index by the h-index and find the whole square of the value. For calculating the f-index we need to compute the e-index and the h-index earlier. Mathematically the f-index can be represented as in equation 2.6.

$$F - index = (\frac{e}{h})^2 \tag{2.6}$$

In equation 2.6, "e" represents the e-index, and "h" represents the h-index.

### P-index

H-index has successfully captured the attention of scientometrics and bibliometrics. It was designed to be simple to measure composite predictors that tried to include just one number. But the h-index makes a correlation between citation and publication without saying how. However, the p-index makes an optimal balance among the mean citation rate and overall citations of concerning publications. P-index offers the best balance between quality and quantity, making it more flexible than h-index [27]. The P-index can be calculated by equation 2.7.

$$P - index = (\frac{c^2}{h})^{1/3} \tag{2.7}$$

In equation 2.7, "c" represents the citation, and "h" represents the h-index.

### q2 -index

The q2 -index is formed from the combination of the m-index and h-index. The m-index can be calculated by taking the median of the citations in the h-core. When compared to the h-index and the m-index separately, the q2 index gives interesting results. The q2-index considers both qualitative and quantitative aspects. The q2-index uses geometrical methods to streamline its derivation and is not easily influenced by other high values [8]. On the other hand, the p-index provides the optimal balance between the mean citation rate and overall citations [27].

The equation 2.8 represents the q2 -index.

$$q^2 - index = \sqrt{h * m} \tag{2.8}$$

In equation 2.8, "h" represents the h-index, and "m" represents the m-index.

**K-index**

There are two parameters, publication and citations that make up the h-index. A collection of references drawn from the publications of both h-core and h-tail. H-index only considered citations in the h-core and disregarded citations in the h-tail. Even in the case of h-core citations, the scientific scholar's impact does not boost by increase in h-core citations. In that way, Maabreh et al have developed and introduced a new index called the k-index that additionally considers the citations in the h-tail [23]. They discovered that when k-index typically falls in real world scenarios then the tail-core ratio often increases. It has been demonstrated that a power-law idea fits with these functional results. Mathematically it is described in equation 2.9.

$$K - index = \frac{(\frac{c}{p})}{\frac{c(h\_tail)}{c(h\_core)}} \tag{2.9}$$

It is found from the equation 2.9 that the k-index considers the citations in h-tail and h-core. Here "c" represents the citation, "p" represents the publication."c(h_tail)" represents citations in h_tail, and c(h_core) represents citations in h_core.

## 2.4 Academic Age-Base Parameters

The academic age-based assessment parameters determine the career age of the researcher's publications. The inactive researchers are ignored by the h-index, the age-based parameters can also enclose the impact of those researchers who are not active in the field of research. Jin 2007 believes that the age of the publications is an essential prerequisite to analyzing the performance changes over time [20].

### M-quotient

There is another assessment parameter, the m quotient can also determine the time issue [35]. One of the major flaws of h-index is that researchers continue to support it even when there is no current study in the field because its impact is still being felt. Therefore, it is impossible to compare researchers of different ages if the h-index is biased towards older researchers. Hirsch developed the m-quotient, often known as Hirsch's m-quotient, in his initial study while taking the career duration issue into account. To compare researchers with various career lengths, he divided the h-index by the number of years since the first publication. Therefore, the m-quotient is helpful when comparing researchers with different career lengths. It can be calculated as.

$$M - quotient = \frac{h - index}{y} \tag{2.10}$$

In equation 2.10, y represents the number of years since the first publication.

### hI-norm

The h-norm assessment parameter considers the normalized citations. The h-norm works in two steps: first, it normalizes the citations of each research publication, and then for normalized citations, it calculates the h-Index. This is a significantly more advanced method. In the initial h-index, citations were not taken [39]. This method is much more advanced. As the normalized citations were considered, which were not considered in the first h-index calculation.

### hc-index

The age of the article does not taken into account of h-index because authors who got a higher h-index from a noteworthy publication and later on for retiring or engaging in inactive research activities, he have not receive a penalty did not do so. Sidiropoulos et al. have developed the modern h-index, a broader version of the h-index [33]. It is biased and gives newly released studies more attention. The hc-index assesses the impact of the article that is used. This generalized index accepts citations but gives newly published articles higher priority. Mathematically it can be represented as

$$hc - index = [\frac{c_{\text{i}}}{(y_{\text{now}}) - (y_{\text{i}} + 1)}] \tag{2.11}$$

from the equation, y(now) represent the current year while yi+1 shows the publication year, and c(i) shows the paper's citations. When the citation count divided by a variety of factors then the value of hc-index is very low. The y factor enters to address this deficiency. According to the equation below, this can frequently lessen the impact of an article with the passing of four years. The equation of the hc-index is given below.

$$hc - index = [\frac{c_1}{1}, \frac{c_2}{2}, \frac{c_3}{3}, ..., \frac{c_{\text{i}}}{n}] \tag{2.12}$$

### Aw-index

It is observed in the literature that the AR-index considers only those papers that are cited more and ignored the less one, but it cannot satisfy the requirements, and this can be cured by another assessment parameter known as the AW-index. The AW-index defines the research career age of each research publication. Instead of merely considering articles with many citations, it considered all the researcher's publications.

The average annual number of citations for all articles is added up to create the aw-index. A successful measure must be impervious to alterations throughout time. The square root of the total yearly average citations from all publications is equal to Aw-index. The equation 2.13 represents Aw-index.

$$Aw - index = \sqrt{\sum_{j=1}^{n} \frac{cit_{\text{j}}}{a_{\text{j}}}} \tag{2.13}$$

In equation 2.13, "cit" represents citation, and "a" represents the total number of publications in a year.

### hI-annual

The proposed hl-annual has addressed a comparison of researchers with varying career lengths. It mentions the yearly average change in the unique h-index [13]. Finally, an intuitive interpretation is also feasible because the metric represents the typical number of unilaterally impacted full publications that a scholar publishes each year.

### Ar-index

The Ar-index took the publication age in addition to the citations count in h-core and increase in h-core citations. The AR-index indicates that when the years are passed the citations gained from the research publications remain the same, this behavior of the Ar-index decreases the impact of the researchers. The researchers who want to keep their rank high, it is necessary for them to be active in the research field. The average citation count made from the articles included in the h-core is added together to create the AR-index [20]. An effective and good metric should be capable of resisting to depict the changes over time, according to Jin et al.
This measure can show how performance changes over time, both upward and downward. The AR-index is equal to the square root of the sum of the yearly average citations for all publications in the h-core. Three elements make up the Ar-index publications, citations, and years since the first publication. Because if a researcher is inactive over time then h-index does not decrease, the AR-index aims to remove bias towards authors who have not published any research in a while. The formula for AR-index is shown below.

$$Ar - index = \sqrt{\sum_{j=1}^{h} \frac{cit_{\mathrm{j}}}{a_{\mathrm{j}}}} \tag{2.14}$$

In equation 2.14, "cit" represents citation, and "a" represents the total number of publications in a year.

**h-variants evaluation**

It is an interesting subject to assess the researcher's work, journals, different groups of researchers, and conferences conducted by the research community. Scientometrics is also for analyzing the impact of a research scholar. Scientometrics research comprises quantitative research features and the impact of scientific research on the research community. The h-index undervalues the significance of the work because it concentrates on the quantity as well as the impact of publications while ignoring the impact of highly cited works [21].

Fukuzawa et al studied the propagation of copyright of research publications and citations. An assessment of their relationship was also included in the investigation. The study covered nearly four thousand publications and top authors were considered from Japan among the correspondent's sample. This investigation revealed a U-shaped relationship between them [17]. Schreiber et al, studied and examined the h-index including its all variants. They conducted a meta-analysis and concluded that there is a strong relationship between the h-index and its variants [31]. Van Rann examines the association and relation between h-index variants using dataset from the field of chemistry research groups in the Netherlands for assessing this technique [37].

## 2.5 International Award

In the field of research still, it is difficult for assessment of researchers. Dunaiski et al, for the first time, use the award winner of the specific domain as a benchmark to evaluate these indexes [7]. The technique which we use in this study is based on a diversified civil engineering field. 250 researchers are under examine who have won the award in the field of civil engineering. These researchers are associated with different award-winning societies from ACI, ASCE, CSCE, and ICE. To figure out which parameter is the most effective for ranking authors, these parameters should be weighed using advanced machine learning techniques on datasets from many fields of study to produce the best results.

## 2.6 Societies of Civil Engineering

### 2.6.1 American Concrete Institute (ACI)

(ACI) is the world's top authority and resource for developing, disseminating, and adopting unity based standards. ACI was founded in January 1905. Its headquarters are currently located in Farmington Hills, Michigan, United States.

### 2.6.2 American Society of Civil Engineering (ASCE)

The American Civil Engineering Society is a tax-exempt organization that was created in 1852 to support the civil engineering industry throughout the world 2. This organization is Reston, Virginia-based and an old organization of civil engineering in the US. It was established in 1848 and is based on one of the old societies, the Boston civil engineering society.

The ASCE society takes an initiative to promote civil engineering in the field of research and also promote human rights by supporting social leaders. There are about 177 countries in which this society works and about 152,000 members are registered in this society all over the world. The main goal of this society is to promote civil engineers and also encourage development in technology in the field of civil engineering.

### 2.6.3 Canadian Society for Civil Engineering (CSCE)

In 1887, the Civil Engineering Society of Canada was established, after that, in 1918 the name of this society was changed from CSCE to the engineering institute of civil engineering, and a few years later in 1972, it was registered in EIC3 . This promotes advancement in the field of civil engineering like geological, structural engineering, geomagnetic, and a few more. Osama Moselhi received the CSCE's best paper award in 2019. Every two years later this society gave him an award in the field of structural engineering.

### 2.6.4 Institute of Civil Engineering (ICE)

The institute of civil engineering came into existence in 1818, about two hundred years ago. It is a nonprofit organization, and also it has professional civil engineers in the UK. This society has about 92,000 workers in just the UK, one-third of whom live in the UK. There is a huge amount from all over the world. ICE headquarters is in London. The ICE's mission is to advance the field of civil engineering by providing ethical development, quality-oriented qualification, and also agreement with the government and business. Its business branch offers services such as recruitment and training of employees, publishing, and contracting. ICE is dedicated to supporting and stimulating education, the administration of professional ethics, and the defense of engineering. As an institution, ICE encourages training in professional ethics administration and engineering status defense. It establishes membership criteria for the organization to work for the betterment of civil engineering professionals.

| **Ref No** | **Targeted Problem** | **Proposed work** | **Techniques** | **Limitations** |
|---|---|---|---|---|
| [36] | Assessing qualitative judgment Through Quantitave parameter | Ranking of author Assessment parameter Using logistic regression | Forward Selection Using Wrapper Method | Dependency Between Assessment Parameter / Multicolinearity |
| [1] | optimum parameter to find the most influential author of a specific domain | evaluation of indices Used in the dataset, Mathematics | Spearman's rank correlation coefficient | Not utilize the all indices for this target problem |
| [29] | Auther ranking | evaluation of indices Used in the dataset, Civil Engineering | Spearman's rank correlation coefficient | Highly correlated Feature are used For the existing problem |
| [6] | measure the performance of an author | measure the performance of an author publication count | Publication Count | Does not Depict The true performance of an author. Like one author has contribute the most |
| [9] | Discovery of Expert | Expertise profiles For ranking expert | Publication Impact | Does not Depict The true performance of an author. Like one author has contribute the most |
| [26] | Issue of publication count | Authors awarded based on the number of citations of papers | Citation Count | Illegally increased citation / paper cited for criticism / survey paper |
| [18] | Publication count and Citations Count | New Index Named h-index | h-index = no. of publication with the citation equal or greater | Handle both publication and citation in different dimension |

Table 2.6: Literature Riview

| | | | | |
|---|---|---|---|---|
| [14] | Limitation of h-index | g-index | publication count is either equal to the citations count or the sum of citations is greater than the square of the index number of publication count | cannot give the Absolute value of Total publication in Given period |
| [2] | Limitation of h-index and g-index | Hg-index | 2.2 | Hg-index is derived from h and g index both are on ordinal scale |
| [27] | h-index makes a correlation between citation and publication | balance between quantity and quality: p-index | 2.7 | Provide best balancing for only non-linear process |
| [8] | Does not considers both qualitative and quantitative aspects | derivation and influencing by the other high values q2 -index | 2.8 | The base of $q^2$ is weak correlation between h and m which does not true represents in the other situations |
| [40] | h-core citations are discarded by the h-index | e-index indicates the information loss in the h-core e-index | 2.5 | Does not uses fractional system of counting citations |
| [16] | uses fractional system of counting citations | F-index uses a fractional system of counting citations | 2.6 | It only calculate the fractional value e and h index. |
| [19] | Limitations of Hg-index | A-index | 2.3 | A-index decreases as his h-index increases |
| [20, 32] | Limitations of A-index | R-Index | 2.4 | R-index $\alpha$ statistical power R-index $\alpha$ 1/publication Biasness |

Table 2.7: Literature Riview

# Chapter 3

# Methodology

## 3.1 Introduction

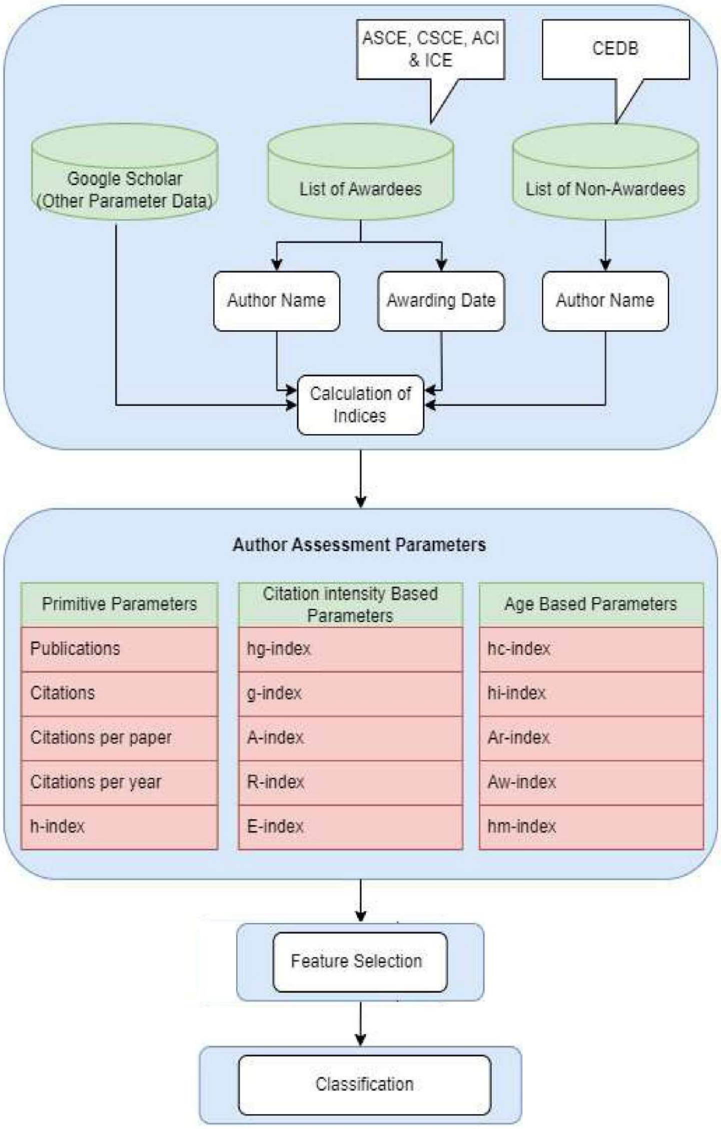
The scientific community is constantly coming up with new methods for assessing scholars/researchers/authors according to their fields of study. According to the findings in chapter 2, researchers are typically grouped based on their total number of publications, their citations, g-index, h-index, or any combination of the two or more features. Still, the efficiency of these author rating criteria has not been thoroughly studied and classified. The evaluation of our implemented technique is based on national and worldwide award winners and non-awardees in the field of civil engineering. In this part, the proposed technique is explained, and the overall schematic method is presented in fig.

## 3.2 Domain Selection

Before going toward the experiments, comprehensive data is required. To assess h-variants and basic features according to their usefulness for prize winners in each respective field, field of civil engineering is selected for the assessment of authors/researchers after careful examination of multiple aspects. These aspects are explained below with their section of the respective field.

## 3.3 Civil Engineering

The oldest profession in human history is civil engineering, and significant research is done in this area. Hence civil engineering field chooses the purpose technique. In addition, almost all scientific organizations evaluate the top experts in each field based on the significance of their research. For the evaluation and ranking of h-indices, the researchers in this field have not been fully exploited. The scientific community in this subject may be able to recognize the deserving person and encourage the development of this field by assessing the research assessment criteria. Due to its significance, this area should also be further explored.



| Primitive Parameters | Citation intensity Based Parameters | Age Based Parameters |
|---|---|---|
| Publications | hg-index | hc-index |
| Citations | g-index | hi-index |
| Citations per paper | A-index | Ar-index |
| Citations per year | R-index | Aw-index |
| h-index | E-index | hm-index |

FIGURE 3.1: Methodology Diagram

## 3.4 Taxonomy Building

The development, construction, and maintenance of the natural places and physical environments, such as airports, dams, building structural components and sewerage networks etc., belongs to one of the oldest scientific fields called civil engineering. Ayaz and Afzal's dataset from the field of civil engineering has been taken into consideration to compile a list of non-awardees [5]. It is observed that Ayaz and Afzal discussed the data of award winners and non-award winners from the field of civil engineering. They have also made use of the ASCE's approved civil engineering database (CEDB), a professional project. Another well-known scientific association for civil engineering is the American Society of Civil Engineering (ACSE). Additionally, we evaluated 250 recipients of major scientific discipline for civil engineering from 2011 to 2019 these societies are (ACI, ASCE, CSCE, and ICE). A list of non-award winners in the neuroscience domain was compiled which is used by Ameer and Afzal dataset. They discuss the data of non-award winners and award winners from the field of neuroscience. The recognized classification for the field of neuroscience is a result of the NIH Blueprint for neuroscience research. NIH is a neuroscience information framework. Reputable neuroscience scientific societies have been considered for the list of 250 recipients of neuroscience awards. The neuroscience scientific societies are ANS, SFN, FENS, and CNS [22]. In the field of mathematics, the list of award winners and non-award winners is organized by Ain et al [28]. They made use of a crawler that is developed by Ayaz and Afzal [5]. Ain et al used a crawler to crawl the data from Google Scholar [28]. Here are some scientific societies in mathematics like IMU, LMS, and AMS. These societies have been considered for the list of 250 recipients of mathematics awards.

## 3.5 Search Engine

Google Scholar has been utilized in this study to compile data on researchers, including publications and citations, and compare it to a well-curated list of recipients and non-awardees . Web of Science, Scopus, and Google Scholar all are sources of getting information for researchers.Google Scholar is one of the most influential sources of getting information for researchers because all other resources have limited access [5]. Because Google Scholar data is openly available and offers thorough coverage of every subject, all the researcher's 25 profiles on google scholar are freely available, and everyone can see the profile of the researcher of their field of study, that's why Google Scholar is selected for the collection of data instead of Scopus and Web of Science. There is enough research done to compare Google Scholar's coverage to that of the Web of Science. With the increase of scholarly data Google scholar platform give the facility, that the profile data of all researchers are updated regularly as compared to all other resources. Harzing also stated that Google Scholar is superior to WoS because new data is consistently updated there.

## 3.6 Data Description

Dataset from the diversified field of engineering is under expirement in this research, which has been obtained by the former research by Raheel et al [29]. The dataset has 500 researchers' information. Of these 500 researchers, 250 researchers have won the award of a scientific society, and the remaining 250 researchers have not won the scientific society award. The researchers who win the award belong to civil engineering societies. There are 4 types of civil engineering societies, ACI, ASCE, CSCE and ICE. The award winners belong to these societies. The description of the dataset in detail is discussed in table 9. Table 9 also has information on civil engineering societies where these societies are found, how many members are there, and how many researchers win the award from each society.

| **Name** | **Values** |
|---|---|
| Authors | 500 |
| Publications | 84, 195 |
| Citations | 40, 76, 722 |

TABLE 3.1: Dataset Description

| Name | Values |
|---|---|
| Authors | 250 |
| Publications | $46,465$ |
| Citations | $17,70,447$ |

Table 3.2: Awardees Description

| Name | Values |
|---|---|
| Authors | 250 |
| Publications | $37,730$ |
| Citations | $23,06,275$ |

Table 3.3: Non-Awardees Description

| Name | Found in | Members | Award Winner |
|---|---|---|---|
| ACI | 1904 | $30,000$ | 14 |
| ICE | 1818 | $90,000$ | 14 |
| CSCE | 1887 | $75,000$ | 64 |
| ASCE | 1852 | $152,000$ | 158 |

Table 3.4: Awarding Societies Description

## 3.7 Data Pre-processing

Data pre-processing is the most important step in machine learning model training. Data preprocessing is applied to get meaningful data from the whole dataset. Data pre-processing uses some important libraries to clean the data, these libraries Numpy, Pandas, Scipy, and matplotlib are the major libraries for data pre-processing. Data pre-processing contains data cleaning removing outliers, removing NaN values, and also removing duplicate values. In this dataset, there are some NaN values in Awarding date column, which is removed. In another column, the f-index contains few NaN values, so NaN values replaced with the mean value of that column. In this way, all the steps of data pre-processing followed. Removal NaN values, drop columns, labeled encoding, and replace them with mean values is done. Some of the steps of data preprocessing are data cleaning, data normalization, removal of NaN values, feature selection, etc.

## 3.8 Feature Selection

One of the fundamental ideas in machine learning, feature selection has a significant impact on our proposed solution's performance. The performance can be attained is greatly influenced by the data features which are utilize to train the machine learning classifiers.

### 3.8.1 Filter Method

Filter method is commonly used to improve the performance, efficiency and accuracy of ML classifier [30]. This method gives the subset of features from the data set by using correlation matrix and it is most frequently done by Pearson correlation. In filter method, it is observed that first, find the strong and week correlation between target variable (dependent variable) and independent variables by defining the threshold. For that Correlation analyses is done for this research.

### 3.8.2 Correlation

Correlation is a technique in which we identify the relationship between two or more variables. Basically it's a statistical relationship. It aids in figuring out whether there is a statistical relationship between variables as well as its strength and direction. In order to comprehend the links between multiple elements or variables, correlation analysis is frequently utilized in a variety of disciplines, including economics, finance, social sciences, and natural sciences.Noting that correlation does not indicate causation is crucial. Two variables are not necessarily caused by each other just because they are connected. Simply put, correlation evaluates how closely two variables tend to move in tandem.It describe the strength and direction of the relationship between two variables. It is a numerical measure which lies between -1 to 1. For example if there is correlation between two variables then the value will be measured, lies from -1 to 1. There are three types of correlation.

| **Coefficient Value** | **Type** | **Meaning** |
|---|---|---|
| 0 to 1 | Perfectly Positive correlation | Change in one variable due to increase of another variable in same direction |
| 0 | Zero Correlation | No relationship |
| 0 to -1 | Perfectly Negative Correction | Change in one variable due to the increase of another variable in opposite direction |

Table 3.5: Correlation Analysis

### 3.8.3 Positive Correlation

A statistical relationship in which two variables tend to move in the same direction is called a positive correlation. In other words, when one variable rises, the other tends to rise as well, and vice versa as one variable falls, the other tends to fall. In other words, the two variables consistently and favorably correlate with one another. While a positive correlation may signal that two variables move together in the same direction, it does not necessarily imply causality, it is crucial to keep in mind. To put it another way, simply because two variables have a positive correlation does not necessarily imply that they are connected. There could be additional variables or underlying mechanisms at work that have independent effects on both variables. It is a correlation in which if the value of one variable increases the value of another variable also increases and if the value of one variable decreases the value of the other variable is also decreased. In other words, rate of change in one variable, other variable also change in the same direction. We can say that both variable are directly proportional. As shown in the figure 3.2.

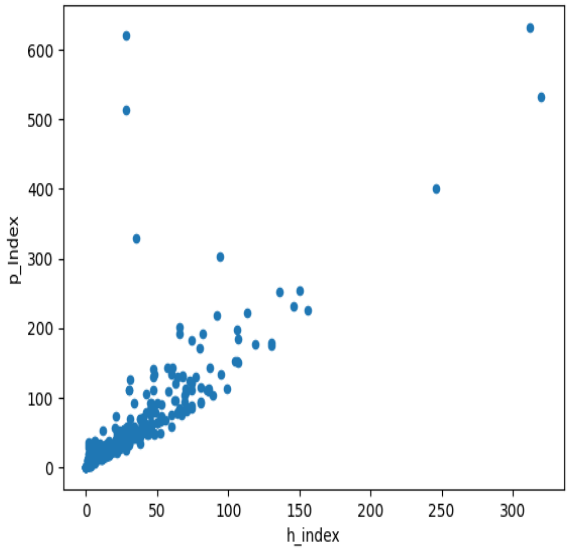


Figure 3.2: Positive Correlation

### 3.8.4 Negative Correlation

A negative correlation is a correlation in which if the values of one variable increase the values of the other variable are decreasing and if the value of one variable is decreased the value of the other variable is also increased. In other words, rate of change in one variable, other variable also change in the opposite direction. We can say that both variable are directly proportional. It's necessary to point out that a negative correlation does not imply connection, just like with positive correlations. Negative correlation between two variables does not imply that one influences the other. Other underlying variables or mechanisms may be in operation and have independent effects on both variables.A negative correlation might have different strengths. Usually, a correlation coefficient, such Pearson's correlation coefficient (r), is used to measure it. A high negative connection is indicated by a correlation coefficient that is close to -1, whereas a weaker negative correlation is indicated by a correlation value that is closer to 0. As was already noted, correlation analysis is a statistical method for determining the nature, magnitude, and direction of links between different variables, including those with a negative correlation. It can be useful for making predictions and decisions in a variety of fields by assisting researchers and analysts in better understanding how changes in one variable connect to changes in another.As shown in the figure 3.3.

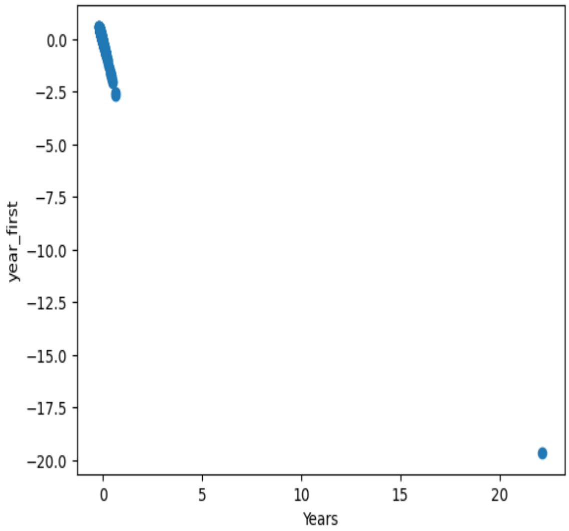


Figure 3.3: Negative Correlation

### 3.8.5 Zero Correlation OR No Correlation

Zero OR No correlation is the correlation in which if the values of one variable are increasing or decreasing there is no change in the values of the other variable. It does not states that there does not exit any relationship at all, it employees that the relationship is not linear.There is no statistical relationship or association between two variables if there is a zero correlation, also known as no correlation. To put it another way, changes in one variable do not translate into changes in the other. It is a strong indication that there is little to no linear relationship between two variables when the correlation coefficient—typically Pearson's correlation coefficient, abbreviated ”r”—between them is close to 0. Like as shown in the figure 3.4. Key points regarding zero correlation.

- **No Predictive Relationship**: When two variables have zero correlation, you cannot predict the value of one variable based on the value of the other variable. They are essentially independent of each other in terms of a linear relationship.
- **Scatterplot Pattern**: In a scatterplot of the data points for two variables with zero correlation, you would typically see a random scattering of points with no discernible pattern. This lack of pattern reinforces the absence of a linear relationship.
- **Independence**:Zero correlation does not necessarily mean that the two variables are completely unconnected. It expressly implies that there isn't a straight path of causality connecting them. They might still be dependent on other people or systems.
- **Causation Consideration**:It's critical to keep in mind that even if there is no correlation between two variables, there may yet be a causal connection between them that isn't captured by their linear association. Other variables or nonlinear interactions may be at work in these situations.

Finding no association between two variables in research and data analysis can be just as instructive as discovering positive or negative correlations.

### 3.8.6 Pearson Correlation

In order to calculate a final result, similarity scores are calculated by comparing two data objects side by side, attribute by attribute, and typically adding the squares of the magnitude differences for each attribute. This is called correlation as explained in the above Correlation section. The most famous way of correlation

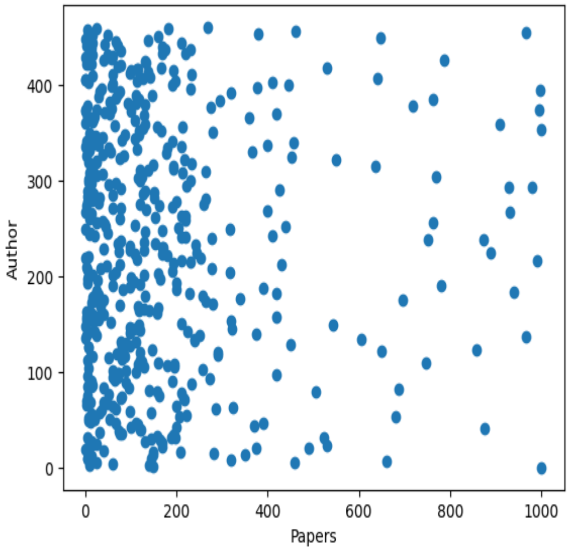


Figure 3.4: Zero Correlation

is Pearson's correlation. Pearson's correlation gives a score that lies from -1 to +1. Objects that have high scores (near + 1) are highly similar and correlated. Objects that are uncorrelated to each other have a Pearson value of zero or near zero. Objects that have a high score or value (near - 1) are highly inversely correlated. Pearson's correlation equation is given below. Pearson correlation can be calculated as the covariance of two variables divided by the multiplication of the standard deviation of X and the standard deviation of Y.

$$Pearson's\ Correlation = Covariance\frac{x, y}{std(x) * std(y)} \tag{3.1}$$

Further simplification,

$$r = \sum(xi - x)(yi - y)/\sqrt{\sum(xi - x)2(yi - y)2}$$

$$r = \frac{\sum(x_{\mathrm{i}} - \overline{x})(y_{\mathrm{i}} - \overline{y})}{\sqrt{\sum(x_{\mathrm{i}} - \overline{x})^2 \sum(y_{\mathrm{i}} - \overline{y})^2}} \tag{3.2}$$

- r = correlation coefficient
- $\mathrm{x}_{\mathrm{i}}$ = values of the x-variable in a sample
- $\overline{x}$ = mean of the values of the x-variable

- $\mathrm{y_i}$ = values of the y-variable in a sample
- $\overline{y}$ = mean of the values of the y-variable

In this case, the dataset which is taken for the this research contains 500 tuples and 33 features. The dataset contains 250 award winners of scientific society and 250 authors who are non-awardees. Pearson's correlation coefficient is used to find correlation between independent variables and Class variable.

Filtered method technique is considered for feature selection. The first step of filter method to find the weakly and strongly correlated feature with target class. After some experiments, it is observed that correlations coefficients of independent variables with class variable (dependent variable) is like if we apply any threshold then many of independent variable which have no correlation with other independent variables (feasible in case of multicolinearity ) extracted from the dataset. For example, by considering the threshold from -0.03 to +0.03 in that way that we select features which are greater than this threshold. In this case Author_Paper, h_coverage and g_coverage are less correlated to the other independent variables which will not create multicolinearity issue as shown in the figure **??**. In over case we have considered all the features for future steps.As so we have calculated the pearson correlation with the each other.
After the experiments on the first step of the filter method to find the weakly and strongly correlated feature with the target class. it is found with some observations that correlation coefficients of independent variables with class variable (dependent variable) are like if we apply any threshold then many of independent variables which have no correlation with other independent variables (feasible in case of multicollinearity) extracted from the data. Consider the threshold from -0.03 to +0.03, for instance in that way we select features that have a greater correlation from the given threshold. In this case, Author_Paper, h_coverage, and g_coverage are least correlated to the other independent variables which will not create a multicollinearity issue. For the aforementioned scenario, every feature has been taken into account for future steps 3.5.
After the computation of the correlation between all the independent variables

as shown in the above figure. The threshold was then set to between -0.5 and +0.5. The independent variables that exhibit correlations between -0.5 and +0.5 have been chosen because between these thresholds correlation strength is weak and moderate as shown in the table3.6. Strengths and weakness of correlation are followed from [15, 29].

| Sr. No. | Coefficient | Strength |
|---|---|---|
| 1 | -0.7 to -1 | Very strong |
| 2 | -0.5 to -0.7 | Strong |
| 3 | -0.3 to -0.5 | Moderate |
| 4 | 0 to -0.3 | Weak |
| 5 | 0 | None |
| 6 | 0 to 0.3 | Weak |
| 7 | 0.3 to 0.5 | Moderate |
| 8 | 0.5 to 0.7 | Strong |
| 9 | 0.7 to 1 | Very strong |

Table 3.6: Correlation Coefficient Analysis

```
Algorithm 1: Feature Selection
Data: Input data
for $i \leftarrow 0$ to $(len(corr\_matrix.columns)$ do
    count =0
    for $j \leftarrow i+1,$ to $(len(corr\_matrix.columns)$ do
        count =0
        if $(corr\_matrix.iloc[i,j]) > 0.5$ or $(corr\_matrix.iloc[i,j]) < -0.5$ then
        colname = corr_matrix.columns $i$;
        col_corr.add(colname)
        count+=1
    if $count \neq 0$ then
    Result: Parameter with Correlation Count
```

The problem statement states that there is a multicollinearity problem with the data, for that we have considered those variables that have weak correlations between each other so we considered those variables that have correlations between

-0.5 and 0.5. A piece of code is written that takes the independent variable that has a Pearson correlation greater than the threshold with other independent variables along with the count. Correlation count means if the variable correlation count is 5 then the parameter is correlated to the 5 other variables under the threshold. Algorithm For that Peace of code is written below.

Through the algorithm 1, parameters that have Pearson correlation with another parameter higher than the threshold with count are obtained as shown in the table 3.7 .

The parameters with a larger count have been eliminated one by one while keeping the other parameters. While doing this if the dropped parameter has a correlation with other parameters or parameters then their count will be dropped by 1. By dropping the citations and then the h-index parameter because the correlation count is greater than the other parameters correlation count the correlation count of Paper decreases by 2. Likewise by dropping the g-index citations count of Paper decreased by 1 again. As shown in the tables 3.7,3.8,3.9.

As so all listed parameters have been dropped one by one. The Paper parameter was automatically removed by the list because all the parameters in which it has correlation were dropped one by one due to a higher citation count. After that Author, Award, g-coverage, Cites-Paper, f-Index, k-Index, year-last, A-Index, and Q-Square-Index parameters which have a correlation with each other between -0.5 to +0.5 have been filtered out.

To address the first and second questions, these filtered features have been used in order to verify the model's accuracy. Feature selection is the most important part of this research. While selecting the features we came to know that the features we have selected under the threshold help to increase the accuracy up to 98%. Even if including these parameters which have a correlation count equal to 1 like h_coverage, year_first and AR_Index accuracy decreases to 74%. So we can conclude that if we consider the parameters that have a correlation is between -0.5 to +0.5 will contribute more in terms of nomination for an award based on some features.

| Sr.no. | Iteration 1 | Iteration 2 | Iteration 3 | Iteration 4 | Iteration 5 | Iteration 6 | Iteration 7 | Iteration 8 |
|---|---|---|---|---|---|---|---|---|
| 1 | Papers 9 | Papers 9 | Papers 8 | Papers 7 | Papers 6 | Papers 5 | Papers 4 | Papers 3 |
| 2 | Citations 17 | | | | | | | |
| 3 | Years 5 | Years 5 | Years 5 | Years 5 | Years 5 | Years 5 | Years 5 | Years 5 |
| 4 | Author_Paper 4 | Author_Paper 4 | Author_Paper 4 | Author_Paper 4 | Author_Paper 4 | Author_Paper 4 | Author_Paper 4 | Author_Paper 4 |
| 5 | h_index 15 | h_index 15 | | | | | | |
| 6 | g_index 14 | g_index 14 | g_index 14 | | | | | |
| 7 | hc_index 13 | hc_index 13 | hc_index 13 | hc_index 13 | | | | |
| 8 | hI_index 12 | hI_index 12 | hI_index 12 | hI_index 12 | hI_index 12 | | | |
| 9 | hI_norm 11 | hI_norm 11 | hI_norm 11 | hI_norm 11 | hI_norm 11 | hI_norm 11 | hI_norm 10 | |
| 10 | AW_index 4 | AW_index 4 | AW_index 4 | AW_index 4 | AW_index 4 | AW_index 4 | AW_index 4 | AW_index 3 |
| 11 | e_index 11 | e_index 11 | e_index 11 | e_index 11 | e_index 11 | e_index 11 | e_index 11 | |
| 12 | hm_index 9 | hm_index 9 | hm_index 9 | hm_index 9 | hm_index 9 | hm_index 9 | hm_index 9 | hm_index 9 |
| 13 | hI_annual 2 | hI_annual 2 | hI_annual 2 | hI_annual 2 | hI_annual 2 | hI_annual 2 | hI_annual 2 | hI_annual 2 |
| 14 | h_coverage 1 | h_coverage 1 | h_coverage 1 | h_coverage 1 | h_coverage 1 | h_coverage 1 | h_coverage 1 | h_coverage 1 |
| 15 | star_count 8 | star_count 8 | star_count 8 | star_count 8 | star_count 8 | star_count 8 | star_count 8 | star_count 8 |
| 16 | year_first 1 | year_first 1 | year_first 1 | year_first 1 | year_first 1 | year_first 1 | year_first 1 | year_first 1 |
| 17 | Cites_Year 7 | Cites_Year 7 | Cites_Year 7 | Cites_Year 7 | Cites_Year 7 | Cites_Year 7 | Cites_Year 7 | Cites_Year 7 |
| 18 | Cites_Paper 1 | Cites_Paper 1 | Cites_Paper 1 | Cites_Paper 1 | Cites_Paper 1 | Cites_Paper 1 | Cites_Paper 1 | Cites_Paper 1 |
| 19 | hg_Index 6 | hg_Index 6 | hg_Index 6 | hg_Index 6 | hg_Index 6 | hg_Index 6 | hg_Index 6 | hg_Index 6 |
| 20 | H_Core 5 | H_Core 5 | H_Core 5 | H_Core 5 | H_Core 5 | H_Core 5 | H_Core 5 | H_Core 5 |
| 21 | R_Index 4 | R_Index 4 | R_Index 4 | R_Index 4 | R_Index 4 | R_Index 4 | R_Index 4 | R_Index 4 |
| 22 | p_Index 3 | p_Index 3 | p_Index 3 | p_Index 3 | p_Index 3 | p_Index 3 | p_Index 3 | p_Index 3 |
| 23 | M_quotient 2 | M_quotient 2 | M_quotient 2 | M_quotient 2 | M_quotient 2 | M_quotient 2 | M_quotient 2 | M_quotient 2 |
| 24 | AR_Index 1 | AR_Index 1 | AR_Index 1 | AR_Index 1 | AR_Index 1 | AR_Index 1 | AR_Index 1 | AR_Index 1 |

Table 3.7: Substitution of Parameters

| Sr.no. | Iteration 9 | Iteration 10 | Iteration 11 | Iteration 12 | Iteration 13 | Iteration 14 | Iteration 15 | Iteration 16 |
|---|---|---|---|---|---|---|---|---|
| 1 | Papers 2 | Papers 2 | Papers 2 | Papers 1 | Papers 1 | Papers 1 | Papers 1 | |
| 2 | Citations 17 | | | | | | | |
| 3 | Years 5 | Years 5 | Years 5 | Years 5 | Years 5 | | | |
| 4 | Authors_Paper 4 | Authors_Paper 4 | Authors_Paper 4 | Authors_Paper 4 | Authors_Paper 4 | Authors_Paper 4 | | |
| 5 | | | | | | | | |
| 6 | | | | | | | | |
| 7 | | | | | | | | |
| 8 | | | | | | | | |
| 9 | | | | | | | | |
| 10 | AW_index 3 | AW_index 3 | AW_index 3 | AW_index 3 | AW_index 3 | AW_index 3 | AW_index 3 | AW_index 3 |
| 11 | | | | | | | | |
| 12 | | | | | | | | |
| 13 | hI_annual 2 | hI_annual 2 | hI_annual 2 | hI_annual 2 | hI_annual 2 | hI_annual 2 | hI_annual 2 | hI_annual 2 |
| 14 | h_coverage 1 | h_coverage 1 | h_coverage 1 | h_coverage 1 | h_coverage 1 | h_coverage 1 | h_coverage 1 | h_coverage 1 |
| 15 | star_count 8 | | | | | | | |
| 16 | year_first 1 | year_first 1 | year_first 1 | year_first 1 | year_first 1 | year_first 1 | year_first 1 | year_first 1 |
| 17 | Cites_Year 7 | Cites_Year 7 | | | | | | |
| 18 | Cites_Paper 1 | Cites_Paper 1 | Cites_Paper 1 | Cites_Paper 1 | Cites_Paper 1 | Cites_Paper 1 | Cites_Paper 1 | Cites_Paper 1 |
| 19 | hg_Index 6 | hg_Index 6 | hg_Index 6 | | | | | |
| 20 | H_Core 5 | H_Core 5 | H_Core 5 | H_Core 5 | | | | |
| 21 | R_Index 4 | R_Index 4 | R_Index 4 | R_Index 4 | R_Index 4 | R_Index 4 | | |
| 22 | p_Index 3 | p_Index 3 | p_Index 3 | p_Index 3 | p_Index 3 | p_Index 3 | p_Index 3 | p_Index 3 |
| 23 | M_quotient 2 | M_quotient 2 | M_quotient 2 | M_quotient 2 | M_quotient 2 | M_quotient 2 | M_quotient 2 | M_quotient 2 |
| 24 | AR_Index 1 | AR_Index 1 | AR_Index 1 | AR_Index 1 | AR_Index 1 | AR_Index 1 | AR_Index 1 | AR_Index 1 |

Table 3.8: Substitution of Parameters

| Sr. No. | Iteration 17 | Iteration 18 | Iteration 19 | Iteration 20 | Iteration 21 | Iteration 22 |
|---|---|---|---|---|---|---|
| 1 | | | | | | |
| 2 | | | | | | |
| 3 | | | | | | |
| 4 | | | | | | |
| 5 | | | | | | |
| 6 | | | | | | |
| 7 | | | | | | |
| 8 | | | | | | |
| 9 | | | | | | |
| 10 | | | | | | |
| 11 | | | | | | |
| 12 | | | | | | |
| 13 | | hI_annual 2 | hI_annual 2 | | | |
| 14 | h_coverage 1 | h_coverage 1 | h_coverage 1 | h_coverage 1 | h_coverage 1 | |
| 15 | | | | | | |
| 16 | year_first 1 | year_first 1 | year_first 1 | year_first 1 | year_first 1 | year_first 1 |
| 17 | | | | | | |
| 18 | Cites_Paper 1 | | | | | |
| 19 | | | | | | |
| 20 | | | | | | |
| 21 | | | | | | |
| 22 | p_Index 3 | | | | | |
| 23 | M_quotient 2 | M_quotient 2 | M_quotient 2 | | | |
| 24 | AR_Index 1 | AR_Index 1 | AR_Index 1 | AR_Index 1 | | |

Table 3.9: Substitution of Parameters

## 3.9 Principle Component Analysis

A popular dimensionality reduction technique used in signal processing, statistics, and machine learning is principal component analysis, or PCA. Its main objective is to transfer high-dimensional data with as much of the original information as feasible into a lower-dimensional representation. The main elements of the data are found and preserved in order to accomplish this. PCA works as, If the features in data have different units or scales, it's important to standardize them to have a mean of 0 and a standard deviation of 1. This ensures that all features contribute equally to the analysis. PCA analyzes the relationships between different features. The covariance matrix is computed to represent the relationships between pairs of features. It shows how much two variables change together. The next step is to find the eigenvalues and eigenvectors of the covariance matrix. Eigenvectors represent the directions or components of the data, while eigenvalues indicate the magnitude of variance in each direction. The eigenvectors are ranked by their corresponding eigenvalues in descending order. The eigenvectors with the highest eigenvalues (largest variance) are considered the principal components. These components form a new basis for the data. The original data is projected onto the new basis formed by the selected principal components. This results in a reduced-dimensional representation of the data. By applying PCA on the given data set, it reduces the parameters to 9 parameters. Different classifiers were applied on the PCA's reduced parameters, details are discussed in the section 3.10.

## 3.10 Machine Learning Classifiers

### 3.10.1 Logistic Regression

In the removal of attributes or parameters from a feature vector, there are a lot of techniques that have been recommended in the literature using a variety of classifications and feature filtering techniques. To answer the first and second research question the statistical approach called logistic Regression for the proposed research is used.

Logistic Regression works.

$$Odds\ of\ success = \frac{probability\ of\ happening}{probability\ of\ not\ happening} \tag{3.3}$$

can be written as

$$Odds(\theta) = \frac{p}{1-p} \tag{3.4}$$

In the Research Gap It was clearly discussed the Problems and Issues in the proposed research of Usman and Tanveer Afzal [36]. Main issue that Logistic Regression faces ,it assumes that independent variables are not correlated to each other's. If correlations exit then results are not reliable. Logistic Regression uses logit function as shown below.

$$y = \beta + \beta_1 x_0 + \beta_2 x_1 + ... + \beta_{\text{n}} x_{\text{n-1}}$$

To predict odds of success take the log of odd of success

$$log\frac{(p(x)}{(1-p(x))}$$

This equation is "y" over here. So in that case it will be written as

$$log\frac{(p(x)}{(1-p(x))} = \beta + \beta_1 x_0 + \beta_2 x_1 + ... + \beta_{\text{n}} x_{\text{n-1}}$$

After some simplification it will simplify into

$$P(x) = \frac{1}{1 + e^{-(\beta + \beta_1 x_0\ +\ \beta_2 x_1\ +\ ...\ +\ \beta_{\text{n}} x_{\text{n-1}})}} \tag{3.5}$$

The above equation is the equation of sigmoid function. Through this logistic regression will predict. As Logistic Regression uses logit function that indicates that when independent variables correlate to each other than change in Dependent variable does not truly represent by Logistic Regression. In order to do that, the features reduced by PCA and filtered features are used in the preceding section. For this problem, by default parameters were used which are given the table 3.10.

| **Parameter** | **Value** |
|---|---|
| Penalty | l2 |
| Tolerance | 0.0001 |
| Random State | 0 |
| Solver | lbfgs |

Table 3.10: Logistic Regression's Parameters

By applying logistic regression on the features reduced by PCA, 96% accuracy got on the testing data and 99% on the training data. By applying logistic regression on the filtered features , 97% accuracy achieved on the testing data and 96.65% on the training data. Accuracy achieved on the filtered features that were selected through correlation is higher than the accuracy on the features that were reduced by PCA. The increase in accuracy is due to selection of variables which have weak and moderate correlation with each other. In which classifier truly represent. To verify the over fitting of the classifier, wrapper method is used in which parameters which have correlation count is 1 have been included. Here three parameters which have correlation count is 1 h_coverage, year_first and AR_index. By including h_coverage which is highly positive correlated to g_coverage with 0.93 then accuracy dropped from 97% to 90%. By including year_first which is highly negative correlated to f_index with -0.88 then accuracy dropped from 97% to 88%. By including AR_Index which is highly positive correlated to Q_Square_Index with 0.95 then accuracy dropped from 97% to 91%. By including Authors_Paper which has positive correlation with f-index then accuracy reduced from 97% to 96%. R_Index is correlated with Q_square_Index with 0.93, by include this parameter, accuracy reduced from 97% to 90%. p_index correlation with Cites_Paper is 0.77 and Q_square_Index is 0.77.Accuracy reduced form 97% to 95% by including p_index. It can be concluded on the bases of these experiments that if parameter or parameters have the correlation greater than the threshold with the other parameter then accuracy for awardees and non-awardees reduces.

### 3.10.2 Naïve Bayes

An effective probabilistic machine learning technique for classification applications including text categorization, spam detection, and sentiment analysis is called a Naive Bayes classifier. It focuses on Bayes' Theorem, an essential element of probability theory. The assumption that the features used for classification are conditionally independent of one another is where the term "naive" in Naive Bayes originates from. Although this assumption is sometimes oversimplified, it can still function effectively in practice. To answer the second research question, statistical approach taken into account, Gaussian Naïve Bayes for the ranking of author assessment parameters. The Gaussian Naïve Bayes model is based on the Bayes theorem and makes strong assumptions about the independence of the attributes. A machine learning method called Naive Bayes is utilized for many classification and ranking problems. Like the other machine learning model (logistic regression) Naïve Bayes is used for binary class classification. Gaussian Naive Bayes can be used to solve three types of (multinomial, ordinary, or binary) problems.

The binary Naive Bayes model is used when there are only two possible outcomes for a dependent variable that can be observed, either 0 or 1, in our case awardees and non-awardees. The multiclass Naïve Bayes focuses on more than two dependent variables. For example, hepatitis has an unorganized variety, it can be A, B, or C. The Naïve Bayes algorithm solves the issues when there are more than two ordered outcomes from the observable output. IT can be explained with an example the rating of a product is multiple it can be one star, two, three, four, and five stars. For this problem, by default parameters were used which are given the table 3.11. The Naïve base has achieved the accuracy of 91% on features reduced by PCA for testing and 87.25% on the training of awardees and non-awardees. And Naive base achieved 100% on both testing and training on the features filtered by

| Parameter | Value |
|---|---|
| Priors | None |
| var_smoothing | 1e-9 |

Table 3.11: Naïve Bayes Parameters

correlation process.The reasons data is low or very high influencing the predictions and the probabilities differ from what it is observed outside the training data. As a result model is accurate at training time but not prediction time. Hence it can be stated that it is over fitting.

### 3.10.3 Support Vector Machine

A supervised machine learning approach called a Support Vector Machine (SVM) is employed for classification and regression problems. SVMs are effective tools for pattern identification and data analysis. They function by determining the optimum hyperplane (decision boundary) that classifies data points into distinct groups or forecasts a continuous value. Finding the hyperplane that maximizes the margin between two classes while minimizing classification error is the basic goal of an SVM. This margin shows the separation between the hyperplane and the nearest data points (support vectors) from various classes. SVMs work best with data that can be separated into two classes using a straight line in two dimensions, a plane in three dimensions, or a hyperplane in higher dimensions. However, they can also deal with non-linear data by applying a kernel function to map it into a higher-dimensional space called the feature space. The data points closest to the hyperplane are known as support vectors, and they are very important in establishing the margin. These data points are the hardest to categorize and are utilized to improve the SVM. Because research which is under observation is based on classification, in this way another classification algorithm support vector machine for the evaluation of assessment of author evaluating parameters is used. The support vector machine draws a separation line between the data points and divides them into two classes. It selects the data point which is nearest to the separation line and computes the distance, based on the computed distance it includes that data point into their relevant group. For this problem, by default parameters were used which are given the table 3.12.

| **Parameter** | **Value** |
|---|---|
| Kernel | rbf |
| decision_function_shape | ovr |
| shrinking | True |
| gamma | scale |
| random state | 0 |

Table 3.12: Support Vector Machine Parameters

Accuracy achieved on the features reduced by PCA on the testing data is 96.6% and 98% on the training data. In the proposed case, it is analyzed that the support vector machine achieved the results of 54% on the testing data and 53% on the training data, which is also less than 97% of awardees from Logistic Regression.

### 3.10.4 Decision Tree

A supervised machine learning approach used for both classification and regression tasks is the decision tree model. Internal nodes stand in for feature testing, whereas branches represent potential results of those tests, and leaf nodes represent the ultimate determination or prediction. Decision trees are helpful for both data analysis and model explanation since they are simple to comprehend and interpret. Components are given below

- The tree's highest node, representing the initial feature test that offers the best data separation. It serves as the basis for decision-making.
- Based on particular characteristics or traits in the dataset, these nodes reflect intermediate judgments. The movement of the decision-making process is directed by internal nodes.

- Branches represent potential results of the feature tests by linking internal nodes to child nodes. Each branch is associated with a certain tested feature value.
- Leaf nodes are the tree's endpoints and stand for the ultimate result, whether it be a predicted value in a regression or a class label in a classification.

Decision trees are employed in classification tasks to predict the class or category of a data. When splitting a dataset, they often measure its impurity using metrics like Gini impurity or entropy and choose attributes that have the greatest potential to do so. ID3, C4.5, and CART (Classification and Regression Trees) are common classification decision tree techniques. It is a tree-based algorithm in which decisions are made in the form of Yes or No and 0 or 1. Tree based algorithms are the best fit for deep data. Here we will discuss deep and shallow data. Deep data is those data that have many records. Shallow data is those data that have relatively small numbers of records. When there are training data, the deep architecture model For the evaluation of h-indices faces an over fitting issue [25]. For this problem, by default parameters were used which are given the table 3.13.

| **Parameter** | **Value** |
|---|---|
| criterion | gini |
| splitter | best |
| min_sample_split | 2 |
| min_weight_leaf | 1 |
| random state | 0 |

Table 3.13: Decision Tree Classifier Parameters

Decision tree achieved the accuracy of 95% on testing and 100% on the training data by using the features reduced by PCA. Accuracy achieved on the features selected by correlation process is 95% on the testing and 100% on the traing dataset. Because data used is shallow data that's why decision tree classifier overfits and shows the results of 100%.

### 3.10.5 Cross Validation

In above section, it is observed that our technique or proposed solution have increased the accuracy. Prior to applying the classifier, features selection is done first. Then the verification of our techniques is done by including the parameters using wrapper method. Through this, accuracy of the model decreases according to the inclusion of parameter. We cross-validated our suggested work to ensure that our proposed solution is valid on the data which is used by Aoun, S.G., and Afzal, M.T [4]. The proposed solution is applied on the neurosciences data. First Pearson correlation have been calculated for the given parameters. Then applied the threshold which is -0.5 to +0.5. Some features have been filtered out, the correlation among features is weak and moderate. The filtered parameters after the selection procedure are given in figure 3.6 with their Pearson correlation. On these features when we have applied the classifier Logistic Regression then it gives the accuracy rate 97.3%. Which means that our proposed solution out perform on the both civil and neuroscience data.

## 3.11 Tools and Technology

- Python
- Microsoft Visual Studio Code
- Jupyter Notebook
- Microsoft Office
- Microsoft Excel
- Draw.io

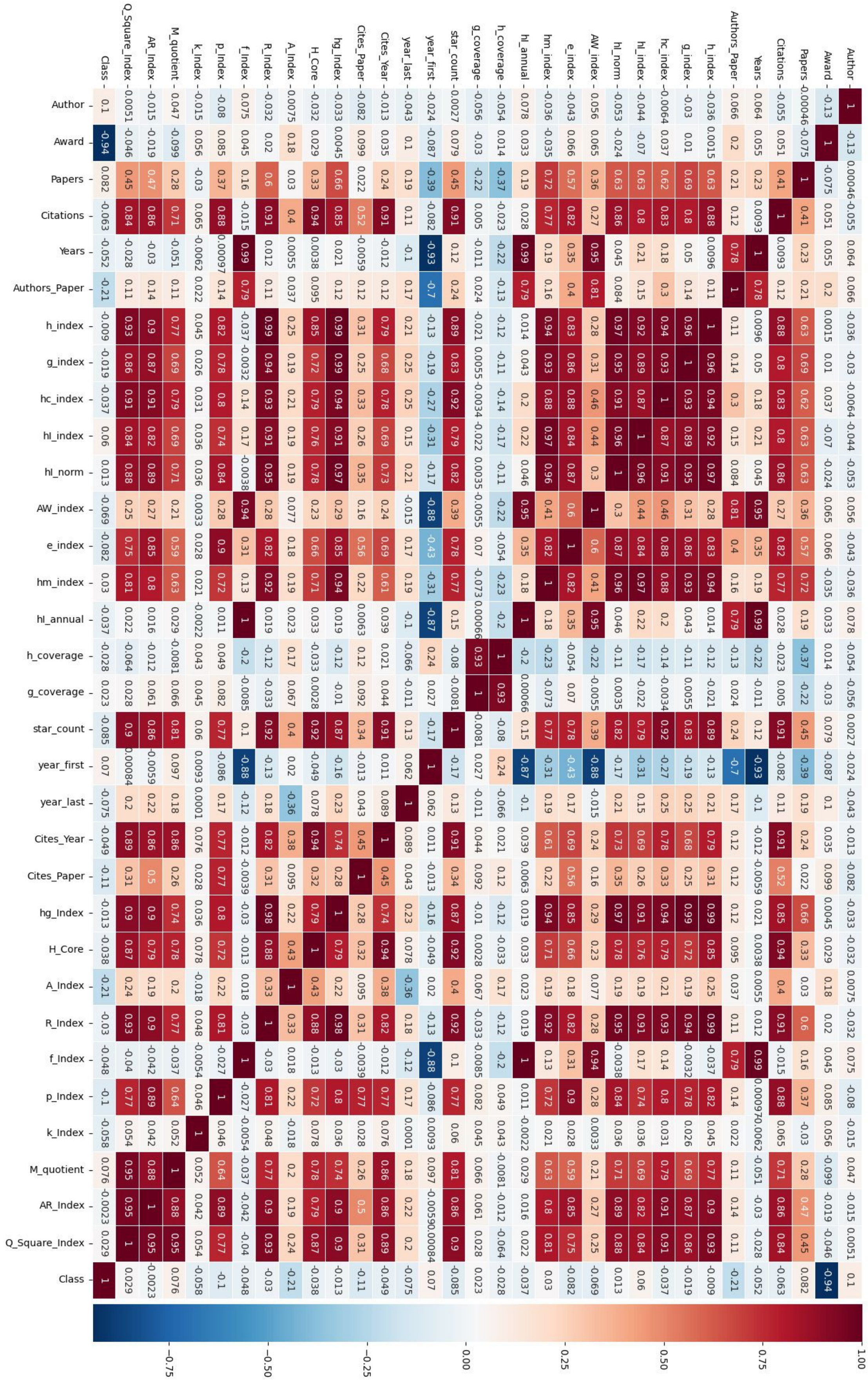


Figure 3.5: Correlation Matrix

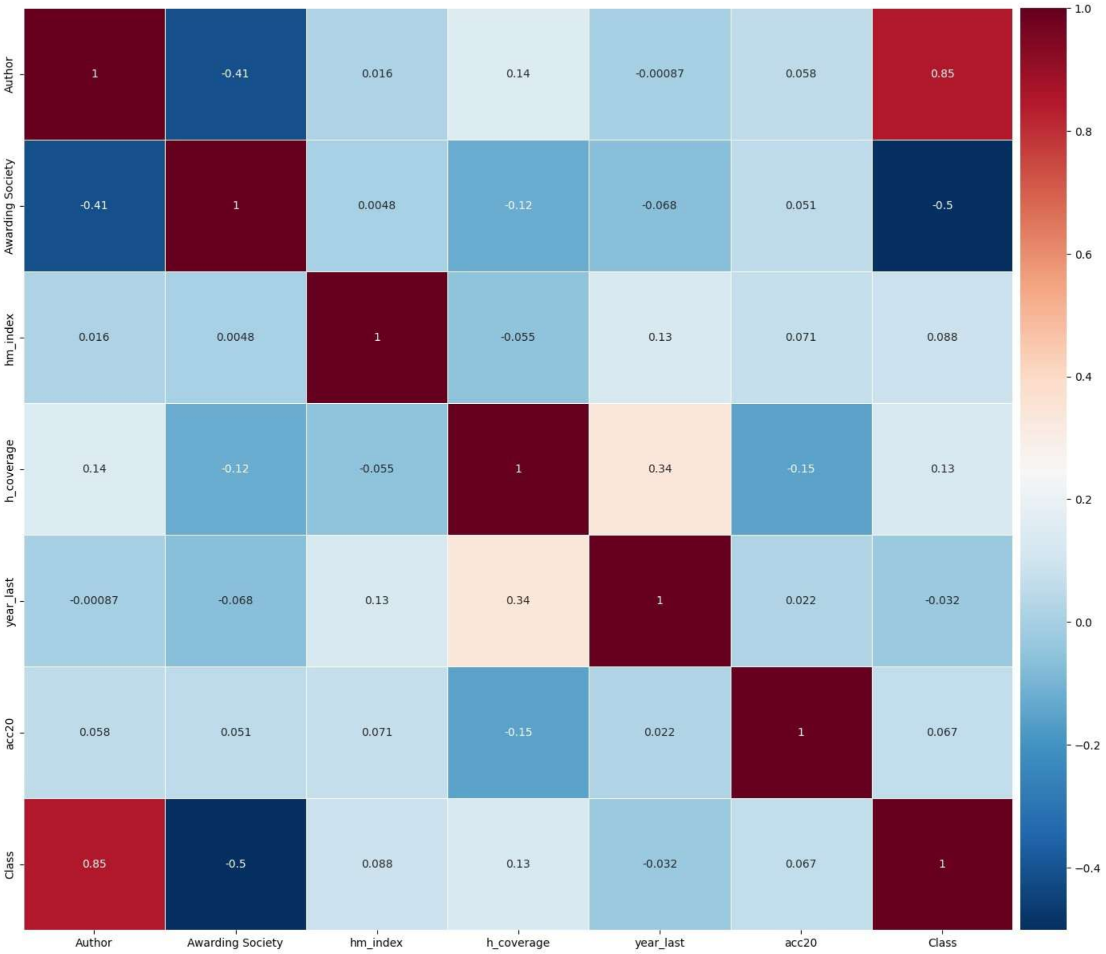


FIGURE 3.6: Pearson Correlation

# Chapter 4

# Result and Discussion

## 4.1 Introduction

In this study, A comprehensive feature selection analysis to identify the most relevant and discriminative features for author assessment has been conducted. The dataset consisted of a collection of civil discipline documents, and our goal was to determine which author assessment parameter or parameters contribute the most to accurately distinguishing between different authors. The results gained are explained in the methodology part of Chapter 3. Over here, a brief comparison or result with previous study[36] will be described. The feature selection process involved several steps. subsectionPearson Correlation Computation As is described and observed in the literature review there are the dependencies between the data that cause the multicollinearity. For that, we employed the Pearson correlation to find the correlation between the variables. Correlation is the best way to compute the dependency between the parameters.

## 4.2 Feature Selection

After computing the correlation, the threshold value has been used to eliminate strongly and extremely strongly associated features. First, the data from the civil engineering field has been considered. Through this Author, Award, Papers, g_coverage, Cites_Paper, f_Index, k_Index, year_last, A_Index, and Q_Square_Index features have been selected. These parameters have weak or moderate correlation with each other. After that, the data from the neurosciences field for the validation of our proposed solution has been considered which is used by[3]. From that Author, Awarding Society, hm_index, h_coverage, year_last, and acc20 have been selected. In both cases, some of the features from the whole dataset are selected. It is found and observed to assume that the dependency between the variables is less than the previous. This technique leads to solving or answering half of the 1st question.

## 4.3 Applying ML Classifier

This section leads us to solve both of the research questions. After feature selection, we applied different classifiers for the prediction of authors like he or she is awardee or non-awardee. In the methodology, it is briefly described the how classifier has been applied and the result of the classifier. Over here some more descriptions will be added to the results. First Logistic regression has been applied because the previous M.Usman has also given importance and applied logistic regression. We give Logistic regression more importance over others because in research the multicollinearity and Logistic regression issues have been highlighted. Logistic regression uses the Logit function which assumes that variables are not correlated to each other. For multicollinearity, the Pearson correlation has been computed and then selected features that have a weak and moderate correlation with each other. Logistic regression has been trained on the selected author assessment parameters with a ratio of 0.80 and performed testing on the data with a ratio of 0.20. It is observed that our model outperforms and gives an accuracy of 97% on the data from the civil engineering discipline. To verify the overfitting of the classifier we use the wrapper method by including variables that have a correlation count is 1. Here we have three parameters that have correlation count 1 h_coverage, year_first, and AR_index. By including h_coverage which is highly positively correlated to g_coverage with 0.93 accuracy dropped from 97% to 90%. By including year_first which is highly negatively correlated to f_index with -0.88 then accuracy dropped from 97% to 88%. By including AR_Index which is highly positively correlated to Q_Square_Index with 0.95 accuracy dropped from 97% to 91%. When we included Authors_Paper which has a positive correlation with the f-index the accuracy reduced from 97% to 96%. R_Index is correlated with Q_square_Index with 0.93 when we include it then accuracy is reduced from 97% to 90%. p_index correlation with Cites_Paper is 0.77 and Q_square_Index is 0.77. When we include p_index then accuracy is reduced from 97% to 95%. We can conclude on the basis of these experiments that if a parameter or parameters have a correlation greater than the threshold with the other parameter then accuracy for awardees and non-awardees reduces. For the cross-validation, when we trained our model on the data from the Department of Neuroscience with the ratio 0.20 then we got an accuracy of 97.33% on the testing data for the predictions of awardees and non-awardees.

## 4.4 Results and Comparisons

Previous research[36] has used the Logistic Regression and found the importance of wrapper method results are given below.

| **Author Assessment Parameters** | **Accuracy** |
|---|---|
| Author paper | 67% |
| hg-index | 19% |
| Rest of Parameters | less then 15% |

Table 4.1: Accuracy Table

Through the proposed solution on the data from the field of civil engineering by applying Logistic Regression, the results are given in the below table.

The same data set was also used by raheel [29]. It is observed that Raheel did find the occurrence of awardees from the ranked list of awardees in 1% to 10%, 11% to 20%,....,91% to 100%.In the ranked list of 1% to 10% f-index, hc-index, $h_T$index, t-index, and Wu-index perform better than the other parameters like Raw h-rate and a-index who have performed low in a ranked list in 10%.

From the ranked list of 10%, Wu-index got 47%, F-index got 46.5%, T-index got 46.5% and the Contemporary h-index got 46.5% which were the highest from the other author assessment parameters. The occurrence of awardees decreased gradually as they moved in the ranked list from 21% to 100%. The percentage occurrence of awardees from 11% to 20% was less then 15% and from 21% to 100% was less than 10%. The occurrence of awardees from the ranked list in 1% to 10% are shown in the table.

| **Author Assessment Parameters** | **Percentage Occurrence of Awardees** |
|---|---|
| Wu-index | 47% |
| H-index | 42.5% |
| Tapered h-index | 45.5% |
| F-index | 46.5% |
| T-index | 46.5% |
| Contemporary H-index | 45.5% |

Table 4.2: Occurrence of Awardees

It is observed from the given table 4.2 that no one from the given table was able to get 50% of the awardees from the ranked list of awardees in 1% to 10%. It is observed that both study[29, 36] have only found out the importance of features one by one not go for the combination of parameters. On the other hand in our proposed solution, all parameters are considered equally before taking correlation. After correlation, only weakly and moderately correlated features were selected for the further procedures. Our proposed solution results are extremely good as show in the table 4.3.

| Features | Accuracy |
|---|---|
| Filtered Features | 97% |
| Filtered Features + h_coverage | 90% |
| Filtered Features + year_first | 88% |
| Filtered Features + AR_index | 91% |
| Filtered Features + R_index | 90% |
| Filtered Features + p_index | 95% |
| Filtered Features + Author_Paper | 96% |

Table 4.3: Accuracy Comparison

## 4.5 Evaluation

Three evaluation metrics have been considered to assess the suggested methodology 1) Accuracy, extracted from the machine learning model,s and which one of the classifiers outperforms on the selected features? 2) The importance of the features selection which increases the possibility of a researcher's name appearing on the list of prize winners.3) The importance of the features selection which increases the possibility of a researcher's name appearing on the list of non-awardees. It is observed that the trends of winning awards among the filtered parameters in the field of Civil Engineering.

### 4.5.1 Accuracy

Accuracy, which is the total number of accurate predictions divided by the total number of predictions made for a dataset, is one of the often-used assessment metrics in classification tasks. When the target class is balanced, accuracy is useful. But in the case of imbalanced classes do not make for a suitable decision. If our training data consisted of 99 photographs of cars and just 1 image of a bike, our model would just be a line that consistently predicted cars, giving us 99% accuracy. In our case data is balanced as 250 for awardees and 250 for non-awardees. To verify the proposed solution findings, accuracy has been calculated for an evaluation metric. The Pearson correlation method has been used to find the correlated features of given parameters. The proposed technique is evaluated using the accepted formula for accuracy which is shown below.

$$Accuracy = \frac{(TruePositive + TrueNegative)}{(TruePositive + FalseNegative + TrueNegative + FalsePositive)} \tag{4.1}$$

in that way, the accuracy calculated is 97%.

### 4.5.2 Precision

Precision, an essential concept in statistics and machine learning, evaluates how accurately a classification model predicts the future. It is one of the most important measures for assessing how well binary and multiclass classification algorithms perform. When the cost of false positives is significant or want to be sure that positive forecasts are very dependable, precision is especially crucial. A high precision number means that the majority of the model's optimistic predictions came true. This indicates that the model is trustworthy in recognizing true positives because it is typically correct when it predicts a positive class. A low precision number denotes that a sizable fraction of the model's optimistic forecasts were produced in error. In this situation, the model can be prone to false alarms, and its optimistic predictions might not be very trustworthy. It is the ratio between actual true positives and predicted true positives by the model. Precision is inversely proportional to predicted false positives, which means if the model predicts more false positives then precision is less. Precision tells us the ratio of how much our proposed solution predicts the true positive and false positive. This means it tells how many actual awardees our model predicts from all predicted awardees. Which is the main part of the proposed research that to predict awardees which helps society to resolve the issues described in the introduction and literature review. The formula is given below.

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive}$$

$$Precision = 0.959$$

### 4.5.3 Recall

Another significant parameter in the assessment of classification models is recall, sometimes referred to as true positive rate. Out of the total number of real positive examples in a dataset, it assesses a model's capacity to accurately identify all relevant cases (positives). Recall measures the model's capacity to prevent missing any positive cases. When the model successfully captures the majority of the positive cases in the dataset, it has a high recall value. It indicates that there aren't many missing positive examples in the model. A low recall value indicates that a sizable part of the positive cases in the dataset are missing from the model. Because it misses more true positive situations, it has a higher rate of false negatives. The recall is calculated as the ratio between the numbers of Positive samples correctly classified as Positive to the total number of Positive samples. The recall is also a performance measure of the machine learning model. The formula of recall is,

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative}$$

$$Recall = 0.979$$

# Chapter 5

# Conclusion and Future Work

## 5.1 Conclusion

Evaluation of scientists' accomplishments has subsequently grown to be of utmost importance to the scientific community due to its many benefits. Numerous unique criteria, such as an author's number of published articles and their number of citations, have been suggested to identify any scientific academics' involvement in the scientific community. The scientific community evaluates the significance of scholars' research work based on these criteria and assesses the researchers to determine who the most profiled researcher is.In educational institutes, it can helps in the process of hiring the faculty on the bases these criteria. In the student's view, they can help the student choose a researcher to serve as their supervisor to accomplish research objectives. From an academic point of view, they can assist institutes in hiring candidates with distinguished research backgrounds to serve as a faculty member. Hiring the member as an editor or reviewer in the top-rated journals and conferences might also aid them. Through this, the throughput of the system will increases.

To answer the first research question, it is found the most influential parameter or parameters that can increase the possibility of the researcher in the nomination of award recipients. In this technique, we find the most important parameters that is most affected in the assessment of researchers. In the field of civil engineering, assessment of the researchers through Author, Award Papers, g_coverage, Cites_Paper, f_Index, k_Index, year_last, A_Index, and Q_Square_Index is very well than the inclusion of other parameters.With these inclusion of parameters, it achieved the accuracy of 97%. So, It stats that these are the most important features, and should not drop these features in the model training. Naïve Bayes, Support Vector Machine, and Decision Tree techniques have also been applied and found the Logistic Regression outer performed among all these techniques with 97% of accuracy.

To answer the second research question, Logistic Regression machine learning model has been applied to find the accuracy of awardees or non-awardees occurrence on these parameters. It is found from the previsou study[36] the accuracy of the model which is 67%. Accuracy is increased by applying the correlation technique. By using the Pearson Correlation, it is observed that some parameters have positively correlated, and some parameters have negatively correlated. By analyzing the correlation, it is found that nine parameters have weakly and moderately correlated. With the selection of weakly and moderately correlated parameters, the accuracy of Logistic Regression elevated from 67% to 97%. From the foregoing discussion, it can be stated that the suggested technique assisted the scientific community in discovering the parameters on which the author assessed that has a strong correlation with the recipients of awards from famous scientific societies. In addition, the identified rules can direct non-recipients to appear on the list of award recipients, and it aids scientific societies and academic institutions in decisions about membership assignments, the selection of editors and faculty members, and the employment of article reviewers and increase the output of these systems.

There is a significant group of co-author-based author evaluation parameters also have been assessed on the diverse dataset and results are very good on the dataset form the different field but must be assess on a large dataset. To determine whether the current author assessment criteria are sufficient to identify the actual significance of researchers in the scientific community, it is planed to assess co-author-based characteristics in the future, or a more accurate assessment parameter is required.

## 5.2 Future Work

In addition to the age, primitive, and citation intensity-based author assessment parameters, there is a significant group of co-author-based author evaluation parameters also have been assessed on the diverse dataset and results are very good on the dataset form the different field but must be assess on a large dataset. To determine whether the current author assessment criteria are sufficient to identify the actual significance of researchers in the scientific community, it is planed to assess co-author-based characteristics in the future, or a more accurate assessment parameter is required.

# Bibliography

[1] Qurat-ul Ain, Hira Riaz, and Muhammad Tanvir Afzal. Evaluation of h-index and its citation intensity based variants in the field of mathematics. *Scientometrics*, 119:187–211, 2019.

[2] Sergio Alonso, Francisco Cabrerizo, Enrique Herrera-Viedma, and Francisco Herrera. hg-index: A new index to characterize the scientific output of researchers based on the h-and g-indices. *Scientometrics*, 82(2):391–400, 2010.

[3] Madiha Ameer and Muhammad Tanvir Afzal. Evaluation of h-index and its qualitative and quantitative variants in neuroscience. *Scientometrics*, 121(2):653–673, 2019.

[4] Salah G Aoun, Bernard R Bendok, Rudy J Rahme, Ralph G Dacey Jr, and H Hunt Batjer. Standardizing the evaluation of scientific and academic performance in neurosurgery—critical review of the "h" index and its variants. *World neurosurgery*, 80(5):e85–e90, 2013.

[5] Samreen Ayaz and Muhammad Tanvir Afzal. Identification of conversion factor for completing-h index for the field of mathematics. *Scientometrics*, 109(3):1511–1524, 2016.

[6] Krisztian Balog, Leif Azzopardi, and Maarten De Rijke. Formal models for expert finding in enterprise corpora. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 43–50, 2006.

[7] Aparna Basu, Sumit Kumar Banshal, Khushboo Singhal, and Vivek Kumar Singh. Designing a composite index for research performance evaluation at the

national or regional level: ranking central universities in india. *Scientometrics*, 107:1171–1193, 2016.

[8] Francisco Javier Cabrerizo, Sergio Alonso, Enrique Herrera-Viedma, and Francisco Herrera. q2-index: Quantitative and qualitative evaluation based on the number and impact of papers in the hirsch core. *Journal of informetrics*, 4(1):23–28, 2010.

[9] Delroy Huborn L Cameron, Boanerges Aleman-Meza, S Decker, and I Budak Arpinar. *SEMEF: A taxonomy-based discovery of experts, expertise and collaboration networks*. PhD thesis, University of Georgia, 2007.

[10] Rodrigo Costas and María Bordons. The h-index: Advantages, limitations and its relation with other bibliometric indicators at the micro level. *Journal of informetrics*, 1(3):193–203, 2007.

[11] Keith R Dienes. Completing h. *Journal of Informetrics*, 9(2):385–397, 2015.

[12] Marcel Dunaiski, Jaco Geldenhuys, and Willem Visser. Author ranking evaluation at scale. *Journal of Informetrics*, 12(3):679–702, 2018.

[13] Marcel Dunaiski, Willem Visser, and Jaco Geldenhuys. Evaluating paper and author ranking algorithms using impact and contribution awards. *Journal of Informetrics*, 10(2):392–407, 2016.

[14] Leo Egghe. Theory and practise of the g-index. *Scientometrics*, 69(1):131–152, 2006.

[15] James D Evans. *Straightforward statistics for the behavioral sciences.* Thomson Brooks/Cole Publishing Co, 1996.

[16] Youhua Fu. F-index: A new index in citation analysis. *International Information Institute (Tokyo). Information*, 16(2):987, 2013.

[17] Naomi Fukuzawa. An empirical analysis of the relationship between individual characteristics and research productivity. *Scientometrics*, 99(3):785–809, 2014.

[18] Jorge E Hirsch. An index to quantify an individual's scientific research output. *Proceedings of the National academy of Sciences*, 102(46):16569–16572, 2005.

[19] Bihui Jin. H-index: an evaluation indicator proposed by scientist. *Science Focus*, 1(1):8–9, 2006.

[20] Bihui Jin, LiMing Liang, Ronald Rousseau, and Leo Egghe. The r-and ar-indices: Complementing the h-index. *Chinese science bulletin*, 52(6):855–863, 2007.

[21] Parul Khurana and Kiran Sharma. Impact of h-index on author's rankings: an improvement to the h-index for lower-ranked authors. *Scientometrics*, 127(8):4483–4498, 2022.

[22] Liming Liang. h-index sequence and h-index matrix: Constructions and applications. *Scientometrics*, 69(1):153–159, 2006.

[23] Majdi Maabreh and Izzat M Alsmadi. A survey of impact and citation indices: Limitations and issues. *International Journal of Advanced Science and Technology*, 40(4):35–54, 2012.

[24] Fionn Murtagh, Michael Orlov, and Boris Mirkin. Qualitative judgement of research impact: Domain taxonomy as a fundamental framework for judgement of the quality of research. *Journal of Classification*, 35:5–28, 2018.

[25] Kitsuchart Pasupa and Wisuwat Sunhem. A comparison between shallow and deep architecture classifiers on small dataset. In *2016 8th International Conference on Information Technology and Electrical Engineering (ICITEE)*, pages 1–6. ieee, 2016.

[26] Richard Pates, Paul Candon, Kerstin Stenius, Michal Miovsky, Jean O'Reilly, and Thomas Babor. *Publishing addiction science: a guide for the perplexed.* Ubiquity Press, 2017.

[27] Gangan Prathap. The 100 most prolific economists using the p-index. *Scientometrics*, 84(1):167–172, 2010.

[28] Muhammad Sajid Qureshi and Ali Daud. Fine-grained academic rankings: mapping affiliation of the influential researchers with the top ranked heis. *Scientometrics*, 126(10):8331–8361, 2021.

[29] Muhammad Raheel, Samreen Ayaz, and Muhammad Tanvir Afzal. Evaluation of h-index, its variants and extensions based on publication age & citation intensity in civil engineering. *Scientometrics*, 114:1107–1127, 2018.

[30] Noelia Sánchez-Maroño, Amparo Alonso-Betanzos, and María Tombilla-Sanromán. Filter methods for feature selection–a comparative study. In *International Conference on Intelligent Data Engineering and Automated Learning*, pages 178–187. Springer, 2007.

[31] Michael Schreiber. Twenty hirsch index variants and other indicators giving more or less preference to highly cited papers. *Annalen der Physik*, 522(8):536–554, 2010.

[32] A Shakil and U Schimmack. Using the r-index to detect questionable research practices in ssri studies [blog post], 2015.

[33] Antonis Sidiropoulos, Dimitrios Katsaros, and Yannis Manolopoulos. Generalized hirsch h-index for disclosing latent facts in citation networks. *Scientometrics*, 72:253–280, 2007.

[34] Lawrence Smolinsky and Aaron Lercher. Citation rates in mathematics: A study of variation by subdiscipline. *Scientometrics*, 91(3):911–924, 2012.

[35] Dennis F Thompson, Erin C Callen, and Milap C Nahata. New indices in scholarship assessment. *American Journal of Pharmaceutical Education*, 73(6), 2009.

[36] Muhammad Usman, Ghulam Mustafa, and Muhammad Tanvir Afzal. Ranking of author assessment parameters using logistic regression. *Scientometrics*, 126(1):335–353, 2021.

[37] Anthony FJ Van Raan. Comparison of the hirsch-index with standard bibliometric indicators and with peer judgment for 147 chemistry research groups. *scientometrics*, 67:491–502, 2006.

[38] Lorna Wildgaard, Jesper W Schneider, and Birger Larsen. A review of the characteristics of 108 author-level bibliometric indicators. *Scientometrics*, 101:125–158, 2014.

[39] Claes Wohlin. A new index for the citation curve of researchers. *Scientometrics*, 81(2):521–533, 2009.

[40] Chun-Ting Zhang. The e-index, complementing the h-index for excess citations. *PLoS One*, 4(5):e5429, 2009.