International Journal of Information Science and Management
Vol. 22, No. 4, 2024, 155-183
DOI: https://doi.org/10.22034/ijism.2024.2015308.1297 / DOR: https://dorl.net/dor/20.1001.1.20088302.2024.22.4.9.2

Original Research

How Do We Select a Combined Algorithm to Determine High-Quality Aerospace Researchers by Utilizing Data Mining Techniques? *

Somayeh GhaviDel
Ph.D., Department of Knowledge and Information Science, Informatics Services Corporation (ISC), Tehran, Iran.
S_Ghavidel@ISC.CO.IR
ORCID iD: https://orcid.org/0000-0002-3852-3398

Nosrat Riahinia
Professor, Department of Knowledge and Information Science, School of Psychology and Educational Sciences, Kharazmi University, Tehran, Iran.
riahinia@khu.ac.ir
ORCID iD: https://orcid.org/0000-0003-2609-6330

Farshid Danesh
Associate Prof., Information Management Research Unit, Islamic World Science & Technology Monitoring and Citation Institute (ISC), Shiraz, Iran.
Corresponding Author: farshiddanesh@isc.ac
ORCID iD: https://orcid.org/0000-0001-5481-3988

Abdolreza Noroozi Chakoli
Professor, Department of Information Science and Knowledge Studies, Shahed University, Tehran, Iran.
noroozi@shahed.ac.ir
ORCID iD: https://orcid.org/0000-0002-0088-9159

Received: 13 November 2023
Accepted: 12 March 2024

Abstract

The aerospace industry and technology are always considered one of the country's most important and valuable industries. The research area of "Aerospace" is among the priorities of the grand science and technology development strategies, and addressing it is strategically vital. The present research aims to estimate and predict the appropriate algorithm for identifying high-quality aerospace researchers based on Advanced Ensemble Classifier Techniques (AECT) in data mining on the outputs of scientometric analyses and predicting the most essential scientometric-related metrics to identify high-quality researchers. The present study was performed using the protocols of applied research and multiple methods. The studied population includes all aerospace researchers (1945 and 2021) indexed in "The Web of Science Core Collection (WOSCC)". DataLab software and multiple programming languages have been applied in this research. All three algorithms have an accuracy of 0.96 and an F1-score of 0.97, which indicates that the models have high accuracy, validity, sensitivity, and predictive power. The "Blending" algorithm is considered a suitable and predictive model. The output of the LightGBM algorithm is that the most important and robust metric in the evaluation of prominent researchers is a metric from the researchers' effectiveness dimension, the Q parameter. According to the knowledge obtained from the ability to predict AECT in the prediction of high-quality researchers, it is possible to use the metrics mentioned in the evaluation of researchers in the field of scientometrics for more accurate and comprehensive prediction. An algorithm that can lead to the optimal and efficient classification of researchers provides the possibility of in-depth analysis of the available data about researchers and smooths the predictive power of the most high-quality researcher. The use of the proposed algorithms in this research, while suggesting the appropriate algorithm, led to reliable and valuable knowledge in classifying high-quality aerospace researchers.

Keywords Aerospace, Scientometrics, Data Mining (DM), Advanced Ensemble Classifier Techniques (AECT), Light Gradient Boosting Machine (LightGBM), Confusion Matrix.

Introduction

The aerospace industry and technology are considered one of the most essential and valuable industries due to their special and unique features and applications, and it is taken into account as one of the priorities of the countries' science and technology development strategies (Britannica, 2020; Naghshineh, 2007). "Aerospace Engineering" is one of the subject lines of the Web of Science Core Collection (Web of Science Core Collection (WOSCC)[1], which has indexed the scientific publications of "Engineering, Aerospace" (Web of Science Core Collection, 2021). Due to the large data volume and the need to extract helpful knowledge, techniques that can facilitate data analysis and extract valuable knowledge from reliable scientific databases have been discussed. With the development of Data Mining (DM) (textual) techniques, the possibility of extracting useful and high-quality information and knowledge has been provided to discover new knowledge from the text. Moreover, summarization using clustering and classification, finding relationships between concepts, and detecting emerging scientific trends have been provided (Sapkota, Alsadoon, Prasad, Elchouemi & Singh, 2019). In the meantime, classification is one of the most well-known methods and has previously been employed in scientometrics (Web of Science) (Khorrami, 2018). Classification is one of the data mining techniques that provides a model for prediction and new data (Maksood & Achuthan, 2016). In the present investigation, classification refers to the effective techniques of exploring and searching to extract valuable and practical knowledge using algorithms and methods and techniques implementable by programming language and software.The ensemble classifier or group learning combines models to improve the model's accuracy. According to Rokach (2010), ensemble methods (hybrid algorithms) aim to build a prediction model by combining several models. Ensemble methods are divided into two categories: basic ensemble techniques and advanced ensemble techniques (AET). In basic ensemble techniques, there are two methods: "Voting" and "Averaging." While "Boosting," and "Bagging" (James, Witten & Hastie, 2013; Zhou, 2012; Breiman, 1996; Syarif, Zaluska, Prugel-Bennett & Wills, 2012; Pino-Mejías, Cubiles-de-la-Vega, Anaya-Romero, Pascual-Acosta, Jordán-López & Bellinfante-Crocci, 2010), "Stacking" (Rokach, 2010) and "Blending" are also among advanced ensemble techniques. There are standard analysis methods for the classification technique (Appendix 1). In AET, the updated boosting algorithms have been embedded, which are "AdaBoost," "GBM," "XGBM," "Light GBM," and "Catboost."

A part of the valuable data in scientific fields is also the result of the analysis of metrics and scientific indices, which made it possible to discover authentic, modern, and reliable knowledge; in this regard, many indices have constantly been introduced in various fields for the quantitative and qualitative monitoring of researchers' publications, based on which various metrics of scientometrics have been introduced (Hirsch, 2019; Bornmann, Mutz & Daniel, 2008; Van Eck & Waltman, 2008; Rousseau & Ye, 2008; Egghe & Rousseau, 2008; Guns & Rousseau, 2009; Yaminfrooz & Gholinia, 2015; Perry & Reny, 2016; Mazurek, 2018). It is essential to evaluate the research performance of aerospace researchers at the national and international levels; monitoring is performed using various metrics of scientometrics (Biglu, 2008; Ivancheva, 2008; Waltman & Noyons, 2018; Hicks & Melkers, 2012; Hicks, Wouters, Waltman, De Rijcke & Rafols, 2015; Carlsson, Kettis & Söderholm, 2014). This research aims

to provide a predictive model for identifying high-quality aerospace researchers using a combination of classification algorithms in data mining. Also, which scientometrics metrics used to prepare the dataset have been the most essential and decisive criteria in evaluating prominent aerospace researchers? Therefore, in this study, by relying on various metrics and indices, the evaluation of aerospace researchers has been performed along with the data extracted from WOSCC regarding the mentioned researchers to form reliable data for analyzing and discovering helpful knowledge. This research aimed to provide a predictive model to identify high-quality researchers based on Bazeley's (2010) analytical model to evaluate and analyze the quality of researchers' works in four fields (productivity, impact, collaboration, and prestige). For this purpose, the analysis of the data obtained from various metrics and the data related to the researchers of the important and strategic area of aerospace available in the reliable scientific database was used. Moreover, it leads to the presentation of DM's predictive AET (a conceptual model) to identify high-quality aerospace researchers and the most essential and practical criteria for determining high-quality researchers.

In expressing the importance of research, it should be emphasized that the aerospace industry and technology are valuable and vital for countries economically, militarily, and politically (Amir & Weiss, 2023). On the other hand, compiling indices for measuring the quality and productivity of the science and technology system and their monitoring is essential and will lead to a more effective realization of science and technology priorities, especially in the vital field of the aerospace industry. In addition to the existing indices, nowadays, researchers' scientific activity is reflected in various scientific databases for visibility, which leads to the emergence of significant and essential data. Such information is always crucial for the scientific evaluation of researchers, and if they are employed, it will be effective in scientific policy-making and managerial decisions. The development of more opportunities for the benefit of researchers, professionals, and experts in appropriate fields and positions and optimal management of scientific programs are among these effects. Data mining is a suitable tool that can act through the correct and proper analysis of the data available in the studied areas of researchers of research groups (soft power) about pathology and future research in that area and can be effective in identifying weaknesses and turning threats into strengths and opportunities by determining the main clusters and nodes in each location (Mohamadesmaeil, 2020). In the meantime, providing an efficient model for monitoring, evaluating, ranking, and guaranteeing the quality of researchers in priority fields (including aerospace) will significantly help scientific institutions. This subject can also open new horizons for policymakers and decision-makers in the aerospace field. They can increase their ability to attract experts in suitable research centers. They can also develop the research skills of faculty members and researchers and access information resources. They can create transnational research networks to disseminate and exchange knowledge and technology under national priorities and take advantage of global opportunities. They can improve scientific productivity, technology, and innovation indices in the studied field. They can enhance the level of researchers in developmental and applied research centers. And they can strengthen the flow of valuable data in the aerospace technology field. Therefore, the results of the present investigation, clarifying the quality of researchers and presenting a model for the evaluation of the assumption of their outputs outstanding and high-quality, lead to more targeted research credits and financial facilitation to develop more effective research provide more beneficial support for strategic and fundamental works with an emphasis on exploitation from them, as well as the establishment

of international research institutes with outstanding experts for the development of international collaboration. Following the development of a model to identify the level of high-quality researchers, scientists, and specialists, we will also witness the advancement of science and technology in the defense and security realms of aerospace-related fields. Considering the strategic and decisive importance of the "Aerospace Technology" subject area as well as the significance of evaluating and measuring the research performance and productive works of authors and identifying the influential scientific leaders of each field who can accelerate the growth and prosperity of research activities in this field, scientific communications and make the future road map of that field clear and also be effective in the macro management strategy, the above issue was the focus of the present research. Finally, reviewing and evaluating the available evidence and preliminary studies make conducting such research inevitable.

In stating the research's problem, aerospace research activities are one of the areas that have been regarded as a concern by the policymakers and country planners, and the scientific researchers also have a thorough understanding of the current situation as well as the role and contribution in their research which they have presented a comprehensive perspective of the current situation and the role and contribution of aerospace researchers in international journals of priority fields. As a result, we are dealing with a data volume increasingly characterized by exponential expansion. The main objective of the broad activity known as DM is to use this stored data to retrieve useful and practical information. The task of the predictive DM is to make inferences from the current data to make predictions. The forward-looking approach in the predictive model of DM will be the basis for creating a more desirable future. In addition, DM methods such as data classification have been employed in scientometrics studies (Khorrami, 2018).

On the other hand, it is inevitable to address the use of accurate quantitative and qualitative measuring techniques in science and technology. Therefore, the present research focuses on the question of how the analysis and evaluation of the quantity and quality of research in the field of aerospace contribute to the presentation of an appropriate picture of the scientific structure, its improvement, and how it will affect the field of science policy or research management strategy. Providing a proper and accurate model for assessing high-quality aerospace researchers is, in general, the present research's most crucial challenge and is of exceptional relevance. A predictive model from DM studies can be applied to decide on educational policies and strategies, direct resources and actions toward actual needs, and develop short-term, long-term, and forward-looking plans. The DM studies also offer the possibility of extracting new and valuable knowledge, recognizing and discovering meaningful patterns and laws, and providing a predictive model. Another challenge with this research is the ability of DM approaches to predict an unknown value of a particular variable, which frequently occurs in the future, the way to utilize DM algorithms to identify patterns from data, assess the importance and influence of each variable, and then give planners of aerospace subject area new knowledge that is relevant, accurate, intelligible, and referable information. The main objective of the present study is to assess the accuracy of AECT-based models for evaluating and predicting high-quality aerospace researchers using the framework provided. Moreover, the present study aims to explain the strongest and weakest criteria for assessing and predicting high-quality aerospace researchers based on DM analysis and present a suitable and accurate model for identifying them. The following research questions can help determine whether or not the study's objectives have been met:

RQ1. What criteria are used for evaluating and predicting prominent aerospace researchers based on DM analysis?

RQ2. Estimation power of which model is the most suitable to evaluate high-quality aerospace researchers based on AECT?

RQ3. What is the appropriate and predictive model to identify high-quality aerospace researchers?

Literature Review

According to their analysis of the astronomy collaboration network, Osareh (2006) concluded that the quantity of multi-authored papers was significant. This Science Direct database has 419 indexed astronomy publications from 2761 authors between 2000 and 2006. Moreover, Ganguli (2008) revealed that Singapore is ranked first when comparing the number of aerospace publications to the global population. The top three universities in the world are the "Beijing University of Aeronautics and Astronautics," "NASA," and "Nanjing University of Aeronautics and Astronautics," respectively. In addition, Alonso-Valdivielso and Antonio (2010) used the bibliometric method to analyze the operations of astronomy-specific libraries. These two researchers' works have resulted in the realization that bibliometric techniques are beneficial in various situations, including alternatives for preservation, extraction, publication of results, and facilitation of internal information management. A study on aerospace engineering by Vaziri and Rajabali Baglo (2010) was based on scientific productions and activities. Their findings indicated that authors from the United States published more than 41% of the scientific works produced by aerospace experts. The highest level of collaboration in the studied countries was with the authors from the United States. Most of the authors' scientific works have been produced and published as articles in the journal "Aircraft Engineering and Aerospace Technology." In addition, NASA, NOAA, and USFA are the world's top three aerospace engineering organizations.

Moreover, Vaziri (2010) conducted a scientometrics study using statistics from the WOS to examine a decade of activity in the scientific productions of aerospace engineering. According to the findings of the study, this database includes 152 scientific production indexed items, the majority of which are journals written in English. To discover and propose an initial framework that fulfills certain criteria based on the Domain Driven Data Mining approach, Strohmeier and Piazza (2013) have investigated DM approaches in human resource management. In addition, Galyani-Moghaddam and Taheri (2015) examined 2501 aerospace-related papers from Iran from 01 January 2014 to the end of the year that were indexed in Web of Science (Science Citation Index Expanded [SCI-EXPANDED]). According to their research, the main authors, such as Kumar, Jain, and Lee, have appeared in numerous international journals. In terms of partner countries, 36 countries participated in the production of scientific articles in the field of aerospace with Iran. The United States was ranked first among them, but according to the authors' Collaborative Coefficient (CC), England, Italy, Germany, and France were ranked higher than the United States. Kirimi and Moturi (2016) employed the DM classification to draw important knowledge from the analysis of their prior works and ultimately predict the performance of their employees. In their study, a classification model with many rules was developed using decision trees as the primary data mining method. A predictive model for predicting employee performance has been presented in this study, allowing HR professionals to refocus on human ability measurements and improve their process for

evaluating the performance of their human resources. Furthermore, Asanbe, Osofisan, and William (2016) used the WEKA tool and the DM approach (classification, neural network, and decision tree) to evaluate the effectiveness of teachers in a higher education institution. The findings indicated that Decision Tree C4.5 has an accuracy level of 83.5% compared to the other two algorithms (ID3 and MLP), both of which have high accuracy levels of performance and time required to develop the models. Compared to the other two algorithms in this study, the decision above tree is the most appropriate algorithm for predicting teachers' performance. Dehghani Ashkezari (2016) used data mining methods to analyze professors' performance at Yazd University of Science and Arts and then provided a suitable model. Their findings demonstrated that the neural network algorithm and the support vector machine method had the highest levels of accuracy. In another study, Qian and Ohwada (2018) applied the clustering method to assess DM for occupational change. Their study is considerably useful to occupational change websites since it enables them to select the top candidates from hundreds of applicants and attract more users. Saad (2018) analyzed the data related to employees in the Libyan Textile industry production sector using the "Bagging" algorithm and DM classification to increase the accuracy of predicting workers' performance. Then, four decision tree algorithms were employed to predict the result and identify the significant connection between the input and the result. The "Bagging" group algorithm utilized four algorithms using decision trees to achieve the maximum prediction accuracy of 16.99%. The four algorithms' standard errors were tiny, indicating that the inputs (seven variables) and the output (evaluation) correlated significantly. Gain charts were extremely close to one another, and the algorithms' curves (receiver operating characteristics) had good specificity and sensitivity. In addition, Nachev and Teodosiev (2018) used support vector machines (SVMs) to evaluate employment data as part of their research to offer tools for detecting, quantifying, and evaluating job-related aspects. The outcomes demonstrated that SVMs outperformed other models used on the same data. The employment data processing method presented in this study and the conclusions drawn is a way to extract usable empirical knowledge. In this regard, Pelicioni, Ribeiro, Devezas, Belderrain & Melo (2018) used bibliometric approaches to examine the trend of space technology. Space technology, satellites (artificial moons), space launch vehicles, spacecraft, rockets, and space stations were essential terms for in-depth analysis. The findings demonstrated that small satellites are where new space technology is being developed as the field of space technologies continues to focus on studies related to satellites, particularly satellite launches. Moreover, Amani and Salama (2019) analyzed the performance of faculty members' characteristics using DM techniques. According to their findings, DM technology can be useful for analyzing staff traits and ranking them to help hiring and decision-making. The right selection is made for each cluster based on the selected features, which are then applied to group faculty members in clusters with similar attributes using the K-means method. In future studies, it is suggested that an algorithm be developed based on the outcomes for each cluster for categorizing decisions. In addition to these studies, Pessach, Singer, Avrahami, Ben-Gal, Shmueli & Ben-Gal (2020) presented a comprehensive framework for decision-making in human resources employment in realistic settings based on a distinctive large dataset, which included the employment records of hundreds of thousands of employees over a decade and a diverse range of heterogeneous research communities. The findings revealed that predicting a candidate's success in landing a particular job in the pre-employment stage is feasible, and one may use the predictions to build a global optimization model. Moreover, the findings demonstrated that the developed

framework could offer a balanced recruitment program while enhancing diversity and success in hiring compared to actual hiring decisions. Moreover, Vavilova, Zievako, Pakuliak & Potapovych (2020) used scientometrics and statistical analysis to assess the NAS Ukraine Journal of Space Science and Technology. Their study examined the relationship between the number of articles published on various journal themes and the growth of thematic areas related to space research in Ukraine. In addition, Yoosin, Yeonjin, SeongGwan and Seung (2017) employed text mining to visualize ontology in aerospace technology. The study mentioned above successfully created an ontology of aerospace technology, and some calculations for subject extraction, network connection, and time series analysis were made. According to the findings, the unique ontology of aerospace technology contains 4,000 terms. Ma (2021) presented results through DM and big data analysis that could identify the problems in the salary and benefits of the company's employees. Therefore, the company's human resources department could more easily discover talents. Liu, Qingqing and Liu (2021) researched the organizational human resource management platform with the proposed methods in DM, and their research improved the optimal allocation management of corporate human resources. Wu, Wang and Wang (2021) also facilitated the task of human resource selection by improving DM algorithms and implementing them on human resource allocation data. Arfaee, Bahari & Khalilzadeh (2021) provided a predictive model for planning the training of human resources of tax administration to improve the current program based on neural network model analysis. Nijs, Dries, Van Vlasselaer and Sels (2021) also modeled team attributes and behavior configurations using the decision tree algorithm. In recent work conducted by GhaviDel, Riahinia, Danesh and Noroozi Chakoli (2022), aerospace researchers have been evaluated using scientometric methods (social network analysis of co-authorship, centrality measures, and network analysis at the level of micro and macro network indices) from 1945 to 2021. Their findings demonstrated the co-authorship network's poor density and weak coherence among aerospace researchers.

The literature review reveals that studies have been conducted with the separate application of "Scientometrics" or "DM" in a variety of subject areas that have been examined; the analyses of the participation network in the aerospace domain at the level of micro and macro indices of the network have provided comprehensive and accurate information for the formulation of science policy and the advancement of the desired strategic plans and programs for aerospace research. However, there was no research in the field of "Aerospace" using a hybrid of the techniques, strategies, and tools of "Scientometrics" and "DM." In some studies, such as the Ph.D. theses of Mryglod, Holovatch and Kenna (2018), Ameli (2020), Li (2020), and Shaheen, Ahsan and Iqbal (2021), the metrics accessible in scientometrics and DM approaches were combined in a hybrid form.

Materials and Methods

The present study is an example of applied research with an analytical approach that uses multiple methods. In the current investigation, scientometric data led to a large dataset that requires the application of the DM technique to guide the discovery of knowledge from the provided databases. Therefore, during the research process, the CRISP process—one of the most well-known, widespread, and at the same time most effective methods and analytical models for conducting DM projects—was used to describe the working process of data DM (Chapman et al., 2003; Pérez, Iturbide, Olivares, Hidalgo, Almanza & Martínez, 2015; Wirth

& Hipp, 2000; Jaggia, Kelly, Lertwachara & Chen, 2020). The six stages of this methodology were 1) project comprehension, 2) data comprehension, 3) data preparation, 4) modeling, 5) model evaluation, and 6) model development (Wirth & Hipp, 2000). The scientometric data was collected and processed by direct extraction from the database, and analyses were performed based on the implementation of metrics on the data extracted from the database. Based on these metrics, the "Classification data mining technique" was used to form predictive models for two research target classes, which were "Google Ranking" and "Affiliation Ranking" of aerospace researchers from each of the metrics in the list, as well as a set of metrics. In the present study, the "Naive Bayes Algorithm," "k-nearest Neighbor Algorithm (KNN)," "SVM," and "Classification and Regression Trees (CART)" were also applied as typical analysis approaches for data mining classification techniques.

The present research population included the information of all researchers in the field of aerospace, who were considered the scientific reference of the world's top scientists. Their scientific outputs were indexed as articles in "WOSCC," among the world's oldest, most reliable, and most crucial citation indices (Web of Science Core Collection, 2021; Birkle, Pendlebury, Schnell & Adams, 2020; Codina, Morales-Vargas, Rodríguez-Martínez & Pérez-Montoro, 2020). In this study, no sampling was utilized; the initial search in the "WOSCC" database led to consideration of the search technique based on the subject areas "Engineering, Aerospace," and no time limit was considered. The authors retrieved 153,994 records from 154,450 authors (without sampling). It seems that thoroughness is considered when extracting materials from researchers and their works, and this can effectively demonstrate the intellectual framework of the researchers in the "aerospace" research field. Data extraction using software that combines multiple technologies is very effective and increases extraction speed and accuracy. The data for each aerospace researcher's investigation was extracted using a different method.

Moreover, there were 2,617,712 citations in total. A couple of tools, the "MiMFa Scraper," were employed to obtain comprehensive data from each author. This tool is regarded as a component of DataLab[2] software. By scraping the data extracted from the website and fitting the received file data using an Application Programming Interface (API), these two software tools enabled accurate and comprehensive data extraction from each researcher's profile. The above software uses multiple programming languages. The most popular, effective, and superior language used for DM operations is "Python programming language," which offers an extensive range of features (Stancin & Jovic, 2019; Hagberg, Schult & Swart, 2021). In the present research, various languages were applied, such as C, C++, and Java. PHP. Machine learning (ML) enables the system to learn automatically without explicit programming. In other words, machine learning aims to create intelligent computers that acquire knowledge based on data collection and experiences (Cabena, Hadjinian, Stadler, Verhees & Zanasi, 1998; PHP). Further testing was performed on "machine learning" models built on the fundamental "bagging," "boosting," "stacked," and "blending" algorithms. To estimate "Feature significance," the Light Gradient Boosting Machine (LightGBM) algorithm **(**Appendix 2**)** was also employed. The "AdaBoost," "GBM," "GBM and XGBM," "LightGBM," and "CatBoost" algorithms are regarded as meta-techniques in AET-boosting algorithms. A gradient boosting system called LightGBM employs tree-based learning algorithms. This method is developed for distributed and efficient applications with advantages such as quick learning and higher efficiency, reduced memory use, improved accuracy, support for parallel computations, and the

capacity to handle massive amounts of data. Notably, this method's processing speed is remarkable (Ke et al., 2017; Chen et al., 2019). For this purpose, two categories of features were selected. One group of features provided information regarding "Publications," "Citations," and "H-index" of aerospace researchers that was obtained from direct data fetching. The other group of features including "Q-Parameter," "ψ –index," "Co-Author Weighted H-index (CAWH)," "K-index," "PPI & PPB," "MNCS," and finally "h^c-Index" and were evaluated and calculated to rank high-quality aerospace researchers, was introduced to the "LightGBM" algorithm.

After designing and developing a model or an algorithm, one of the most important steps is evaluating its performance, efficacy, accuracy, and precision. For the simplicity of the evaluation criteria of classification algorithms, the authors will present them for a problem with two categories. If something can be evaluated, it can be improved as well. Sensitivity and specificity are two essential indices in the statistical evaluation of classification results. The Specificity (Accuracy [ACC]) parameter is known to be against false positives, and the sensitivity parameter is known to be against false negatives. After analysis, the results are divided into two groups of positive and negative data: true positive, false positive, true negative, and false negative. The result of evaluating the performance of the algorithms is displayed by a matrix called the "confusion matrix" (Chawla, 2009). This matrix shows how the classification algorithm works according to the input dataset according to the types of categories of the classification problem (Elkan, 2001). The evaluation results of the above classification are presented in the form of a confusion matrix. These results are employed to determine which model best predicts the target class. The main criteria in the confusion matrix are accuracy, precision, retrieval, specificity, and F1 score; in the present study, accuracy and F1 score are the consensus average of accuracy and sensitivity (criteria for measuring the accuracy of the test and the complexity of the implementation time in the form of the confusion matrix), was drawn. The F1 score is the harmonic mean of precision and retrieval. This metric can have a maximum score of 1 (complete precision and recall) and a minimum score of 0. In general, the F1 score is the criterion for measuring the accuracy and precision of the model (Chawla, 2009). The process of this study is presented in Table 1.

Table 1
Steps of the Present Research

<table>
<tr><td colspan="4">Step 1. Searching the strategy details</td></tr>
<tr><td colspan="2">Selection of database for data extraction</td><td colspan="2">Web of Science Core Collection (WOSCC)</td></tr>
<tr><td colspan="4">Selection of "Web of Science Categories" from the Combo box of the "WOSCC" database search page</td></tr>
<tr><td colspan="4">Entering the term "Engineering, Aerospace" in the search field</td></tr>
<tr><td colspan="2">Search strategy</td><td colspan="2">WC = (Engineering, Aerospace) (Exclude – Publication Years) and English (Languages) and Articles (Document Types)</td></tr>
<tr><td colspan="2">Statistical population (by records): Results: 153 944</td><td>Source language: English</td><td>Document Type: Article</td></tr>
<tr><td>Database indexes</td><td colspan="3">SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI, CCR-EXPANDED, IC</td></tr>
<tr><td>Period: 1945-2021</td><td colspan="3">The date of data extraction from the database: 2022 04 February, Friday</td></tr>
<tr><td colspan="4">Step 2. Data extraction</td></tr>
<tr><td colspan="4">A) Direct extraction of data from the Web of Science Core Collection (WOSCC)</td></tr>
</table>

To extract data, after applying the search strategy in advanced search, the restrictions considered, including the English language of the articles, limiting from 1945 to 2021, and the subject area of "Engineering, Aerospace," the list of results was examined;

●In the list of results, it is clear that to extract "full record and cited references" in plaintext format, only 500 records can be extracted. On the other hand, the website does not allow the extraction of more than 100,000 records; therefore, you should first sort the results list from the newest to the oldest or vice versa. Selecting the "data newest first" option from the sort by combo box is necessary. In this case, after extracting 100,000 new records, the order can be reversed, and the rest of the data can be extracted from the end of the result list;

●The data storage was performed in 500 plaintext formats, and integration of files; Excel (version 2016) was also applied.

●Among 153,944 search results page records, 6,706 were anonymous; therefore, 147,238 unique (non-repeating) articles were identified, and a .csv file was saved for further processing. In the next step, the information of each record containing the article should be given to the software so that the authors' information can be accurately extracted from the same articles in the search results list.

B) Extracting details of the data of articles and authors

●Preparation of scrape pattern of extracting details of the data of articles and authors indexed in "WOSCC" (Appendix 3);

●Due to the large volume of data and the need for high speed, two Virtual Private Servers (VPS) were prepared. In the VPS environment, software and exes were downloaded. To speed up the data extraction process, 13 WOSCC accounts were prepared so that they could be connected simultaneously and two programs could be run simultaneously.

●First, the data related to the articles of the desired subject area that had resulted from the search strategy was searched in the database and stored in *tsv.* The software received format, and "data fetching" from the website was performed. Therefore, all the articles in the search result list (resulting from the search strategy) were fetched. The output is saved in a result art file, which includes 28 records.

●In the next step and through programming, the distinct operation (to remove duplicate values) was applied with the following formula to the entire data of the articles (Appendix 4).

●Then, to get detailed information about the authors of the first stage articles, the saved files were given to the second software to obtain information such as the full name and surname of the authors, the published name of the authors, the code of the authors in the database, the author's paths, the information related to their publications and citations in the period which is indexed in the above database and this information were saved in *.csv* files. When the researcher's name was not included in his profile, searching for the person's articles in Google was also utilized. In this way, it became possible to extract the data of each researcher to control the interference or movement of people with each other.

●The output of the second scrape was performed with a distinct operation (based on the page address of each author) with the following code (Appendix 4).

●The researchers' information was saved in .csv format. In the process of each of the two Scrape software, three files, "success," "error," and "logs," containing extracted correct records, errors, and logs, respectively, were stored in the given path. Finally, the complete information of the researchers in the aerospace field, equal to 154,450 researchers, was extracted and stored, which can be accurately identified with WoS tracking code and profile.

C) Extracting the details of the data of each of the authors' articles

●Extracting citations, references, citing articles, the number of authors of each article, the position of each author in their articles, and preparing the output in JSON format in PHP.

D) Extracting the details of the authors from the Google knowledge graph to retrieve the Google rank

Google Knowledge API service was provided by creating a personalized API key, searching for the authors' names and badges, and receiving related outputs and a score for links connected to the searched name. This key can be obtained by creating an account in the Google Developers section. Each API key

can only send a limited number of requests to the API. The algorithm applied to find researchers with rank one is designed according to Chia's research (2020). The output algorithm was passed through three filters.
E) Extracting the "affiliation" details of each author from the original fetched data of authors
Step 3. Data refinement, homogenization, and normalization
●Authors' names were also refined in the stage of direct data extraction in terms of non-duplication and non-similarity. ●To normalize the textual and numerical data of the final file related to authors (names, numbers related to years, citations, etc.), the DataLab software was also normalized and homogenized. The codes of each of them can be observed in Appendix 5 in their respective order.
Step 4. Calculation of formula and code related to metrics and indices
●With the consultation of a mathematical expert and formula scientist, the studied metrics from Latin resources were carefully and theoretically examined. The components, details, and parameters of the formula of each metric were specified. ●In consultation with the programmer with data scraping skills, the required data of each formula was determined to extract the data from the Web of Science according to each metric. ●In consultation with the formula programmer and data analyst, the best script for the formula of each metric was coded with Python programming language. ●The indices and metrics that were both operational and impacted the determination of the model's accuracy were calculated. ●Google rank and affiliation ranking of aerospace researchers were identified in the data.
Step 5. Data mining process and machine learning
Calling and specifying data ●Python, C#, and PHP programming languages were specified. ●The primary and essential required libraries, including "Math service" with more features than the C# library and pandas (to work with data, analysis, "preprocessing" and "visualization" of data), Numpy (to use arrays and matrices and mathematical functions), matplotlib.pyplot (for drawing graphs), Scikit-Learn (with a range of "machine learning," "data preprocessing," "cross-validation," and "visualization" algorithms), and time (related to time) were determined and imported. ●Libraries related to a random number (to receive random numbers), such as randint, importing libraries related to the preprocessing library in ML models for preprocessing "machine learning" models, such as preprocessing, train_test_split (for dividing data, controlling model accuracy, assigning data to train and test to evaluate the model) and metrics (model evaluation) inputted into Python software. ●The primary data file in .csv format related to 126,874 researchers and the calculated indices were called to the Python file.
Data preprocessing, preparation, cleaning, refining, or purification ●"Data Type" was specified in the file. The columns that cannot be numbered were determined and separated from the data. ●Unnecessary columns "CITATION REPORT PATH," "PROFILE PATH," "OTHER ORGANIZATIONS," "CURRENT ORGANIZATION," "ADDRESS, "PUBLISHED NAMES, "FILED," AND "P1900" were identified and eliminated from the data. ●Columns with "Integers (int)" values were specified. ●Irrelevant data and noise removal, outliers, "missing values" and "duplicate data," unrealistic words or numbers, and dates were removed- and edited from fields such as the "H-Index" field.
Preparation and preprocessing of features (primary features extracted from data and features calculated with metrics and indices) ●Insertion of "Missing Values," depending on the data type, the fields in the rows and columns were filled with zero or Null values. ●"Selection of features" was coded (Appendix 6);

●If the affiliations have a value, it is considered 1, and otherwise, it is regarded as 0 (Appendix 7); ●All the citations and references related to each of the author's articles, which were presented as an array in the citations and references field, were put as a sum of numbers. ●The columns "authors count," authors positions," and "years, which were not needed, were removed from the data. ●Some letters in the metrics, including the PPB measure related to the position of the authors, were numerically equated (Appendix 8) ('A' value = 0, 'B' value = 1, 'C' value = 2, 'D' value = 3, 'E' value = 4, 'F' value = 5, G, value = 6, 'H' value = 7).
●Training sets (80%) and test set data (20%) were determined. ●After data control, 500 out of 126,874 data related to researchers were tested at this stage to accelerate the learning and evaluation of the model; after getting the perfect result from the algorithm, all the data were given to the model. ●Since the classification of people is considered in this research, the target is necessary. Selecting the target increases the accuracy of the model. In the present study, affiliation rank and Google rank were considered targets. However, according to Chia (2020), because affiliation rank has less credibility than Google rank, the focus was placed on Google rank. In the current research data, 80% have at least one affiliation, and 18% have zero affiliation; therefore, two methods were applied to generate Google rank. While using the Google Knowledge API service, the logic considered for affiliation was also utilized for Google production (Appendix 7). For this work and the production of Google rank, the possibility for P0 and P1 were considered. It means that take whatever field has zero in the data and divide it by the total number of data (probability of zero affiliation) and then take the number of affiliations of one from the total data and divide by the total number of data (probability of affiliation of one). Utilizing the mentioned logic and weighting was more appropriate than the random generation of Google Rank. ●In this case, Google's ranking is dependent on affiliation. To prevent the data from being dependent on the affiliation and avoid model error, the affiliation was eliminated.
Feature selection ● The data count, mean, std, min, and max of each column were generated and displayed because it is essential in machine learning regarding model selection and accuracy.
Missing value processing ● Relevant features of the dataset were selected. "Irrelevant" or non-important features (which do not contribute much to the correct prediction of the class label of new examples and, consequently, the performance of the learning models) were removed or ignored. ● According to the high number of columns (in the current research, we have more than 270 columns) and extensive data, feature selection was performed to prevent the model from going astray and remove unnecessary information input to the model, which led to higher performance, accuracy, and desired efficacy of the model. ● "Variance[3], one of the unsupervised filter methods, was utilized for feature selection. This method is one of the best and most effective methods for selecting related features that usually have a higher variance score. The VarianceThreshold function was called from the Feature Selection library. A threshold limit was considered based on the variance drawn for each column. If the variance of each column was higher than the researcher's desired threshold limit, the column was removed. In this step, except for the Google rank column, the rest were checked (Appendix 9). With a threshold limit of 20, it was determined that 140 columns with high variances should not be included in the model algorithm; based on the normal distribution function[4] (min and max), the columns that affect the model's accuracy can be identified (Appendix 9).
Data normalization ●To not diminish the effect of some columns, the min and max of all columns were determined (Appendix 5). ● Then, the distance-dependent data was measured with the K-nearest neighbor algorithm, the simplest

algorithm regarding the classification.
Modeling and evaluation of classification models ● The data (based on 0.2) were split into 80% training and 20% testing. ● In the model, learning is performed based on features. The test data was called, and the model was trained to perform the prediction. We compared the predicted value and the actual value (which we already knew) as well as evaluated the model; ● The data were evaluated based on basic evaluation algorithms and were presented in the form of a confusion matrix; ● "Advanced combined classification" algorithms, including "bagging," "boosting," and "stacking," were called and evaluated based on algorithms related to basic evaluation models (Appendix 1); ● The "blending" algorithm was called and evaluated based on the algorithms related to the basic evaluation models. ● The criteria for evaluating the performance of predicting the methods and models were drawn from the point of view of a "confusion matrix" with "accuracy" indices, "F1 score," or consensus average of accuracy and sensitivity; ● Each of the "bagging," "boosting," and "stacking" models and the combined method was analyzed based on the criterion of time complexity or the execution time of the algorithm and presented in the form of a graph.
Implementation, employment, and dissemination of the model ● Each model was analyzed. ● The appropriate and predictive model with maximum accuracy and efficacy was provided.

Result

Estimation power of which model is the most suitable to evaluate high-quality aerospace researchers based on AECT?

Classification in machine learning has two main stages: 1) learning and 2) prediction. In the first stage, the model is expanded based on training sets, and in the second stage, the model is tested for test data. To answer this question, the primary classification algorithms (Figures 1, 2, 3, and 4) and then the algorithms related to AECT, namely "boosting," "bagging," and "stacking," were implemented (Appendix 10) (Figures 5, 6, 7). The above techniques are visualized in the form of "confusion matrices."

Model Prediction Stage Based on Basic Algorithms

To evaluate the model's performance and the estimation power of the machine learning classification models developed here, the confusion matrix of the models was first assessed (Appendix 10). If any model is placed at a high level regarding "accuracy" parameters and its power in estimating data output, it will be among high-quality models. In this question, the accuracy and precision of a class (category) identification are more important than the accuracy and precision of the overall identification; therefore, the "confusion matrix" concept is helpful here. Approximately 16 metrics-related datasets were modeled with two target variables: google rank and affiliation ranking. Training and test sets were introduced to the models. The models were evaluated in the confusion matrix using Bayes classification algorithms (Figure 1), (Figure 2), "KNN" (Figure 3), "SVM," and "decision tree" (Figure 4), as well as "predicted label" indicates the predicted values and "true label" shows the actual values in the test data.

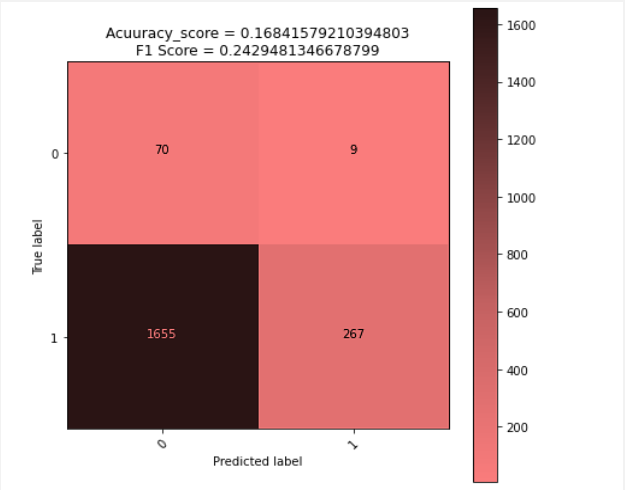

Figure 1: Naive Bayes

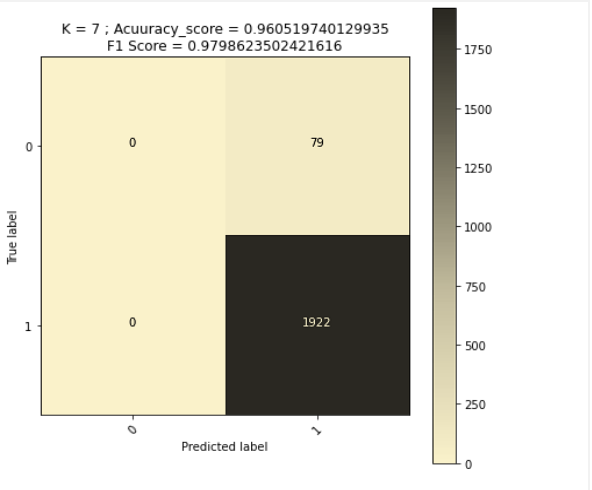

Figure 2: KNN

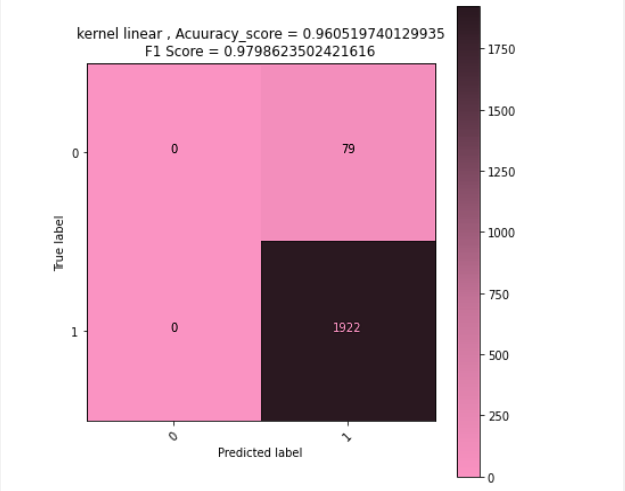

Figure 3: SVM

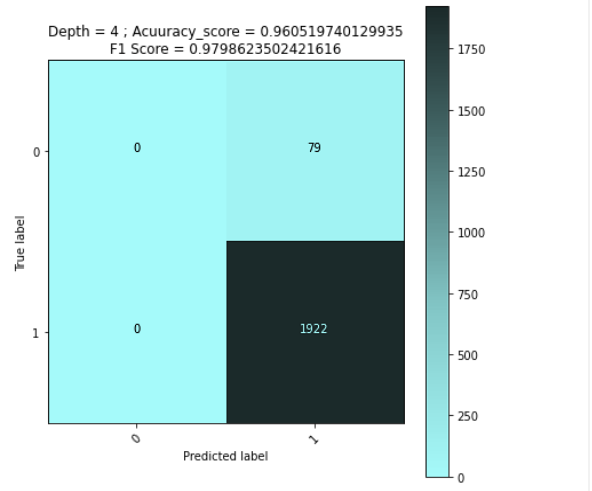

Figure 4: DTree

Figures 1-4: The Models in the Confusion Matrix Using Bayes Classification Algorithms

Figure 1 shows that Naivy Bayas is classified based on probabilities. Training and testing sets were adapted with the GaussianNB function. In this basic algorithm, the "Naive Bayas" algorithm has an accuracy score = 0.16, and the F1 score equals 0.37%. Moreover, in the confusion matrix of the "Naive Bayas" algorithm, 267 cases are predicted to be equal to 1, which is correct, but 1655 cases are predicted to be 1, while they are equal to 0. It also predicted 70 cases to be 0, and they are 0.

Moreover, it predicted 9 cases to be 1, but they are 0. The columns corresponding to 70 and 267 are correct; therefore, the precision is extremely low. According to Figure 2, "KNN" in the classification mode, using the closest points and according to the specified value for K, has calculated the distance of the point we want to label. According to the maximum number of votes of these neighboring points, the decision regarding the label of the desired point has been made. The Euclidean distance method was employed to calculate this distance. In the present study, the "KNN" model was optimized. According to this optimization, the best K has been determined, and the highest accuracy based on trial and error (based on neighbors 5 to 15) has been considered (Appendix 10). The results demonstrated that the best neighbor was 7. The "KNN" algorithm had accuracy score=0.96 and F1=0.97. The KNN confusion matrix predicted 1,922 cases to be 1, which was correct, and 79 cases were predicted to be 1, which was incorrect.

According to Figure 3, SVM was tested and performed with four kernels: "linear," "poly," "RBF," and "sigmoid." The best kernel was the basis for drawing the best model. The innovation of this algorithm was fulfilled by selecting the best kernel and developing a model based on this kernel. The "SVM" algorithm has accuracy score=0.96 and F1=0.97. The "SVM" confusion matrix predicted 1922 cases to be 1, correct, and 79 cases were predicted to be 1, incorrect. According to Figure 4, "entropy" and "information gain" were employed to develop the DTree algorithm. This metric depends on a concept called information entropy. By optimizing the depth of the tree (based on the depth of 4 to 140) in the present research, "max depth" was introduced to the algorithm. The models were developed, and the accuracy of each one was saved. According to this innovation, the best accuracy will be based on any depth. In the present study, the best depth is equal to 4. The "DTree" algorithm has accuracy_score=0.96 and F1=0.97. The "DTree" confusion matrix predicted 1922 cases to be 1, which was correct, and 79 cases were expected to be 1, which was incorrect.

Learning and prediction stages of AECT algorithms

According to Figure 5, first, the Bagging Classification function was written with inputs including training data (X_train, Y_train), test data (X_test, Y_test), the desired model, the number of times estimators and random Number (even if one data is selected). The results were stored in the clf function and were adapted to Gradient Bagging Classifier training and test sets. In the next step, this algorithm was evaluated using test data. Finally, using "confusion matrices," the three evaluation criteria, namely confusion matrix, accuracy, and F1 score, were visualized (Figure 5).

> Advanced Ensemble models:

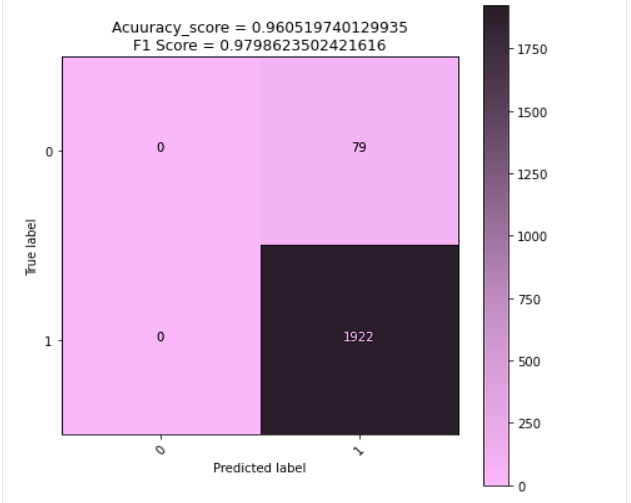

Figure 5: Advanced Bagging Ensemble Classification Technique

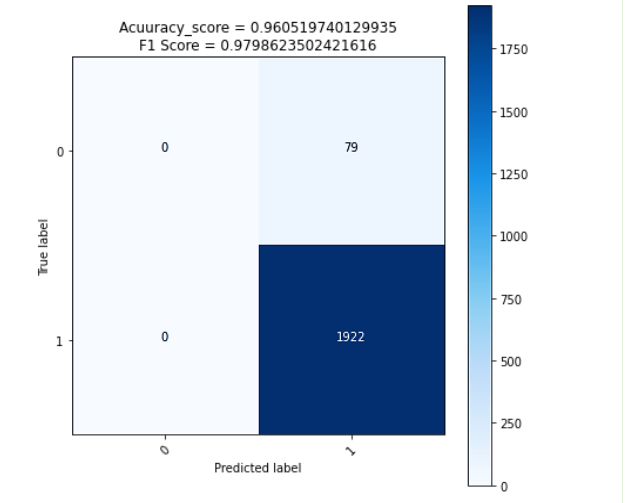

Figure 6: Advanced Boosting Ensemble Classification Technique

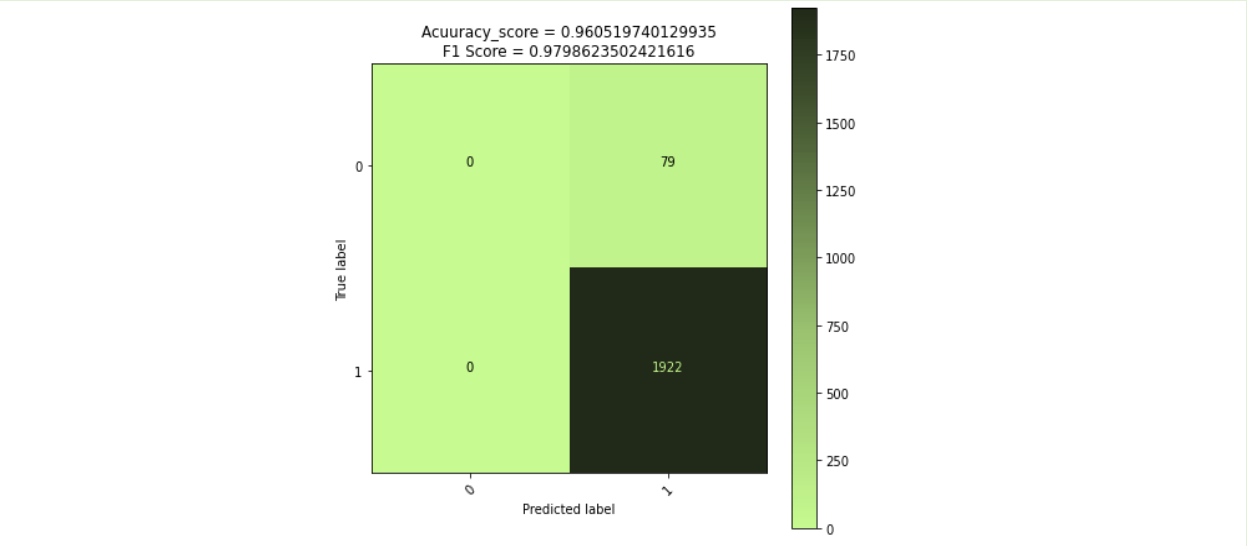

Figure 7: Advanced Stacking Ensemble Classification Technique

In the Boosting Classification function, in addition to the inputs introduced to the Bagging Classification function, a learning rate (learning rate=lr) that represents the speed (step) of updating the weights and is equal to the value of 0.01 was given to the algorithm. The training and test data were also adapted with the Gradient Boosting Classifier. In the next stage, evaluations were applied (Figure 6). In the Stacking Classification function, in addition to training data inputs (X_train, Y_train) and test data (X_test, Y_test), several considered models, including "Naivy Bayas," "KNN," and "SVM" are also entered. Then, the "final model" that provides the final decision was determined by "regression logistics[5]" (Voting). Finally, the evaluation steps of the model have been completed, and the output of the above algorithm is drawn in Figure 7. The innovation that was applied in the "Stacking" algorithm is that it was fed from the optimal basic models, i.e., the best neighbor of 7, the maximum depth of 4, and the best kernel, "linear" of the algorithm, and then the optimal operation of the algorithm was performed for the second time. The results indicated that all three algorithms, namely "bagging," "boosting," and "stacking," have accuracy_score = 0.96 and F1 = 0.97. The "DTree" confusion matrix predicted 1922 cases to be 1, which was correct, and 79 cases were expected to be 1, which was incorrect.

Evaluation of the model based on the criterion of time complexity or algorithm execution time

In machine learning, in addition to the accuracy, the execution time of an algorithm is also necessary. The time complexity or execution time of an algorithm describes the time it takes for the algorithm to execute and stop. Therefore, a time plot was taken from the output of the algorithm (Figure 8).

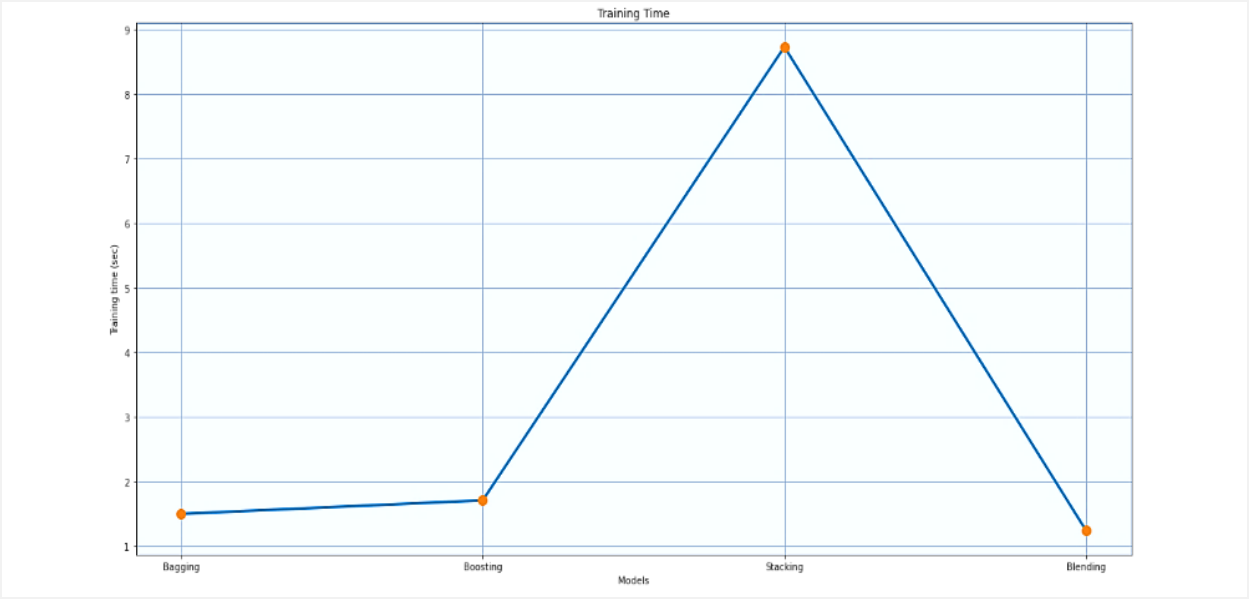

Figure 8: Plot Training Time

According to Figure 8, the fastest algorithm is the Blending algorithm. After the Blending hybrid model, the "bagging" and "boosting" algorithms and then, by a considerable margin, the "stacking" algorithm have fast execution times.

A suitable and predictive model for identifying high-quality aerospace researchers

To identify a suitable and predictive model for identifying high-quality aerospace research hers, in addition to the presented models "boosting," "bagging," and "stacking," the blending method was also investigated. The hybrid method was programmed based on the Blending Classification function by inputting the total data (X, Y), a list of considered models, and a random number (Appendix 10). The training and test data were evaluated and adapted by applying the blending ensemble function's fit. Moreover, the imported models were returned to the algorithm using the return blender function. Finally, the blender was selected using the predict_ensemble function. The final blender was considered "regression logistic" (Voting). The final decision for the final model will be made by "regression logistics"; it should be noted that the "SVM" algorithm was not included in the Blending algorithm due to its unique characteristics (Joseph, Hlomani & Letsholo, 2016).

Figure 9 indicates that the test and prediction data are increased because the data in this model is divided into two test and training parts within the model. Although fewer data were given as learning to the Blending algorithm, it has an accuracy score = 0.96 and F1 = 0.99. A total of 4814 cases were predicted to be one and found to be correct.

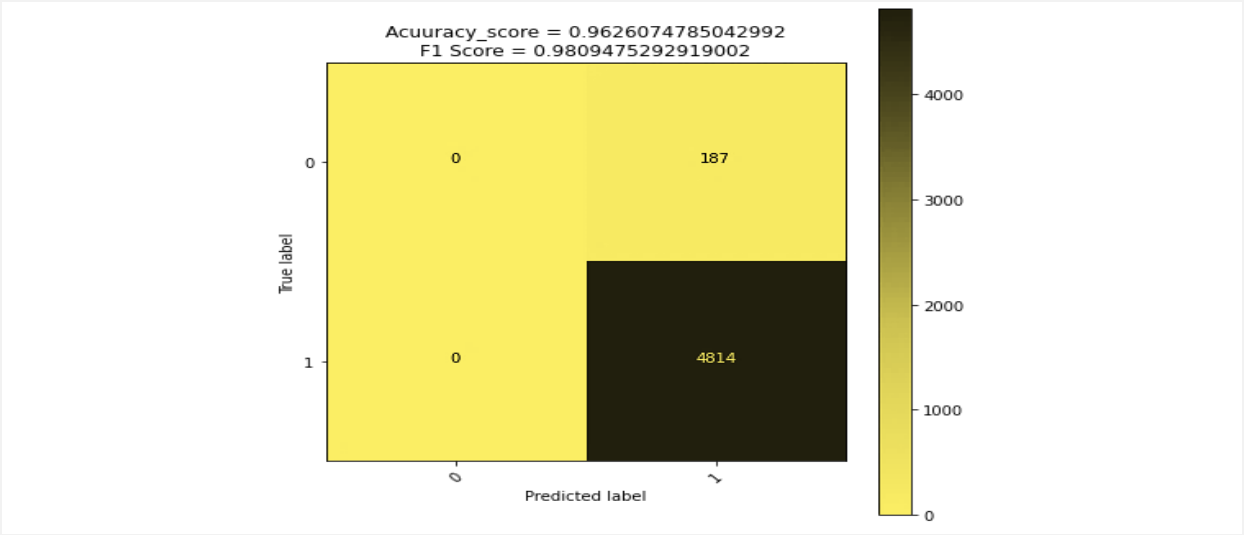

Figure 9: Blending Method

Strongest and weakest metrics for evaluating and predicting prominent aerospace researchers based on dm analyses

LightGBM algorithm has been employed to investigate and evaluate the importance and severity of the impact of the calculated metrics of the four dimensions in evaluating prominent aerospace researchers (Appendix 2). In the "feature selection" section, various metrics were removed from the data; therefore, they were not included in the algorithm in this section. Metrics were entered into the LightGBM algorithm that was calculated on the data, had an operational feature, and could be sufficient in evaluating and presenting the model's accuracy.

According to the output of the LightGBM algorithm presented in Figure 10, the metrics of evaluation and prediction of prominent aerospace researchers based on the calculation were "Q-Parameter," "CAWH," "h-index," "Publications," "PPI & PPB," "K-index," "MNCS," "ψ-index," "Citations," and "hc-Index." "Q-Parameter" is the most significant and influential metric in evaluating prominent aerospace researchers, and the "h^c-Index" is less effective in this evaluation.

Training until validation scores do not improve for 50 rounds.
Early stopping, best iteration is:
[12]valid_0's binary_logloss: 0.221407

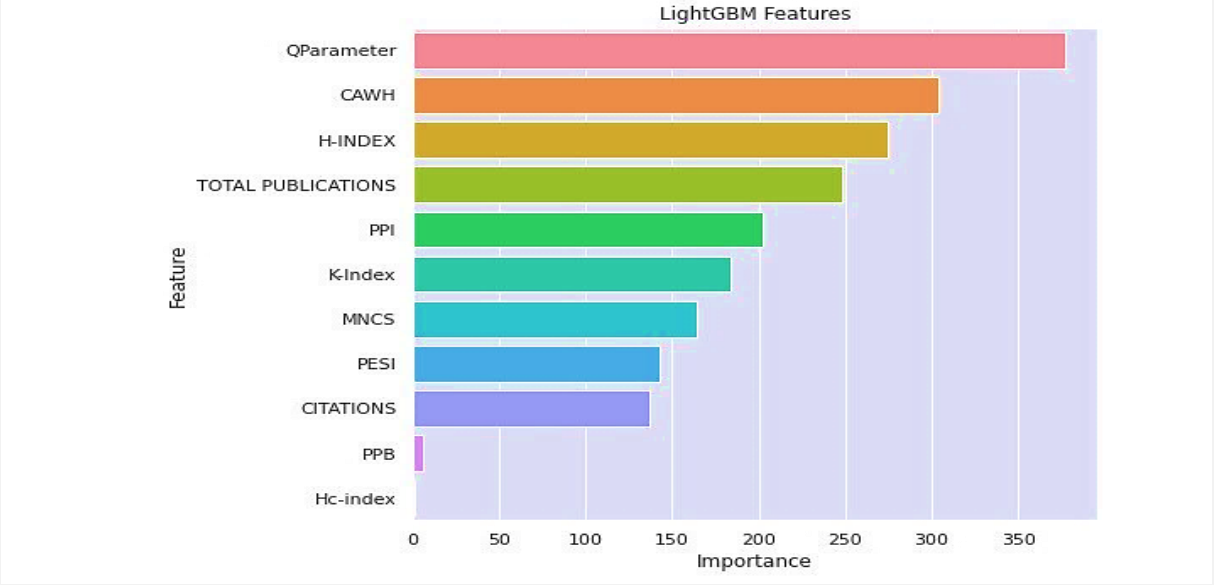

Figure 10: Significance and Severity of the Impact of the Calculated Measure of the Four Dimensions in Evaluating Prominent Aerospace Researchers.

Discussion

In the present study, the prediction of the most significant scientometric parameters to identify high-quality researchers based on AECT on the outputs of scientometric analyses and the prediction of high-quality aerospace researchers while using the LightGBM were taken into account. The data obtained from examining and measuring the quality of researchers' research with the four dimensions and multiple measures raised in the dimensions that result from scientific measurement techniques and measures provided the researcher with a vast and rich data set about aerospace researchers, which were structured and tagged. By using and combining different classification models, we have improved the quality of classification. These algorithms can result in better outcomes using a suitable combination of models. Therefore, a more accurate classification of aerospace researchers has been made possible, and more precise predictions using the Blending algorithm have shown their ability to identify.

The data obtained from examining and measuring the quality of researchers' research with the four dimensions (Productivity, Impact, Collaboration, and Prestige) and multiple measures raised in the dimensions that result from scientific measurement techniques and measures provided the researcher with a vast and rich data set about aerospace researchers, which were structured and tagged. This led to the use of classification instead of clustering people. For this reason, we are not facing one of the problems in the field of DM of texts: the discovery of useful knowledge from semi-structured or unstructured texts. New information was discovered from among the available data. New information was instead found among the already-existing data. Since Kirimi and Moturi (2016) attempted to predict employee performance using DM methods, such as classification, the present study aligns with their findings. The "Naivy Bayas" algorithm has an accuracy score of 0.16 in the present study, while accuracy scores of 0.96 are shared by the "KNN," "SVM," and "DTree" algorithms. The "Naivy Bayas" method has an F1

score of 0.37%, whereas the F1 scores for the "KNN," "SVM," and "DTree" algorithms are all equal to 0.97. Accordingly, the "Naive Bayas" method was found to have the lowest accuracy of the model, which was 0.16, which is correct even though Bayasian and other machine learning algorithms generated an accurate and easy-to-understand model for employing human resources, as claimed in Pessach et al. (2020). Even predictions with this algorithm were only 50% accurate. The outcome mentioned above is consistent with the findings of Bavisi, Mehta and Lopes (2014), Tesema (2019), Guruvayur and Suchithra (2018) regarding the algorithm's poor accuracy. The "Naive Bayas" algorithm, with an F1 score (average consensus of accuracy and sensitivity) of 0.37, again performs poorly in accuracy and sensitivity among the four fundamental indices. All four main algorithms used to predict models have been improved, which has increased the expected accuracy of the model at this level. In the "KNN" model, optimization and innovation have been performed. Since this algorithm's number of neighbors is one of its inputs, this study optimized it from best K to select the best neighbor. Therefore, the highest accuracy is considered when choosing the neighbor, and the process has been moved forward with a better neighbor.

Moreover, in this research, an option named "Maximum Depth" was created in "DTree." The "SVM" algorithm innovation led to selecting the best corner. In this way, the models were requested to be identified based on the best possible accuracy. This innovation and optimization have made "KNN," "SVM," and "DTree" algorithms demonstrate the highest accuracy in terms of accuracy and F1 score. Moreover, Asanbe et al. (2016) employed machine learning algorithms to select the most suitable algorithm for predicting teachers' performance, and DTree was chosen as the best algorithm. Also, Ma (2021) used the DTree algorithm to manage organizational human resources, which could overcome the problems related to employee salaries and benefits and discover their talents quickly. In addition, this approach was discovered by Arfaee et al. (2021) to help predict work performance characteristics and organizational advancement status.

Moreover, Nijs *et al.* (2021) applied the algorithm to assess the team members' talents. Kirimi and Moturi (2016), Arfaee et al. (2021), and Guruvayur and Suchithra (2018) have also considered the use of DM classification in predicting employee performance. The study of Dehghani Ashkezari (2015), in predicting the efficacy level of faculty members, also indicated that the "SVM" algorithm has the highest level of accuracy in estimating the predictive model using the studied data. Moreover, Nachev and Teodosiev (2018) proposed the "SVM" algorithm as a high-precision algorithm for assessing, which is in line with the present study.

According to the literature, various techniques can be employed for classification, including deep learning, traditional machine learning algorithms, and complicated ensemble machine learning techniques such as hybrid methods. Compared to deep learning, hybrid classification algorithms have the advantage of more efficient parallel processing and data division. The accuracy score and F1 values for each of the three hybrid methods, "Bagging," "Boosting," and "Stacking," were 0.96 and 0.97, respectively, demonstrating the high accuracy, precision, sensitivity, and predictive capability of the models. The innovation applied in the "Stacking" algorithm is that from the basic optimal models, i.e., the best neighbor of 7, the maximum depth of 4, and the best corner, "linear," was entered into the algorithm and re-optimized. The current research is consistent with the results of Saad (2018), who studied the application of the "Bagging" algorithm to improve the accuracy of prediction in evaluating the performance of employees and, as a result, presented a predictive model. If the machine learning algorithms

did not demonstrate the desired accuracy and did not solve the need to answer the research question, deep learning was used. In the present study, the algorithms in the advanced ensemble method, while removing the defects of the conventional and basic algorithms such as "KNN," showed high accuracy. In the present research, comparing the ensemble methods with the conventional techniques improved the traditional algorithms, and excellent results were obtained.

Even though the "Blending" ensemble method received fewer input data, the model's high accuracy and predictive capacity are higher than others. Higher performance, better accuracy, and a higher F1 score than the three models mentioned-KNN, SVM, and DTree-offer further justification. The Blending model effectively evaluated the test data as indicated by the 4814 examples predicted to be one and later proved to be correct. The ensemble "Blending" technique can be implemented more rapidly than other models, despite the two "Stacking" and "Blending" models having nearly similar accuracy. It can be concluded from the results that the model based on the "Blending" algorithm is suitable and predictive for identifying top aerospace researchers considering all aspects (estimation power, accuracy, sensitivity, F1 score, or execution time).

What is inferred from the output of the LightGBM algorithm is that the most essential and most effective metric in the evaluation of high-quality aerospace researchers is a metric of the effectiveness of researchers, namely the "Q-Parameter." The nature and characteristics of the Q-Parameter metric alone indicate the significance of this metric in the evaluation of researchers, and the results of the analysis of the LightGBM algorithm also confirmed this. In this metric, the influential articles of individual authors who have contributed to the writing of a work are measured and demonstrate the sustainable ability of a researcher to publish high-impact articles at each stage of the researcher's scientific professional activity (Sinatra, Wang, Deville, Song & Barabási, 2016). Additionally, it was discovered in the present study that the "hc-index" or "modern h-index" metrics have less impact on this assessment. This index considers and weights the citations to the author's recent articles. Therefore, the effects caused by time and receiving citations of people disappear over time (Sidiropoulos, Katsaros & Manolopoulos, 2007). In Chen et al. (2019), the LightGBM algorithm performed well, accurately, and predictably when it identified 36 key aspects of the "mechanical ventilation" subject area concerning 3636 adult patient cases in the Intensive Care Unit (ICU).

It is always necessary to have a complete and correct knowledge of the prospective capabilities of researchers to build scientific policy and progress strategic plans and programs for aerospace research. In addition, this research can:

- It provides a new pattern for detecting high-quality aerospace researchers in the future that can help select suitable colleagues, advisors, reviewers, and scientific collaborations.
- It provides an optimal and efficient method for classifying researchers based on various scientometric indicators that can help evaluate performance, encourage and strengthen researchers, allocate resources, and make managerial decisions.
- It provides a method for in-depth analysis of the data available about aerospace researchers that can help identify patterns, trends, relationships, and factors affecting the quality of research.
- It provides a predictive method for identifying high-quality aerospace researchers that can help detect talents, develop skills, enhance the scientific level, and create motivation in researchers.

Conclusion

The various scientometric metrics provide indices that show the quality of researchers. This study examined effective and robust methods for predicting high-quality researchers following this policy. A thorough analysis of the available data regarding aerospace researchers is possible using predictive models based on modern algorithms, which can result in the optimal and effective classification of researchers. It smoothes the power of predicting the best researchers. Based on the knowledge obtained from the ability to estimate AECT in the prediction of high-quality researchers, it is possible to use the metrics and indices mentioned in the evaluation of researchers in scientometrics in a more accurate and comprehensive prediction. Using the proposed algorithms in this research while suggesting the appropriate algorithm led to reliable and useful knowledge in classifying high-quality aerospace researchers. While useful for policies relating to science, technology, and innovation in the aerospace sector, the findings are also helpful for identifying and categorizing capable researchers who are qualified for important roles or novel job positions. In addition, based on the value, information, characteristics, and competencies of the researchers, which is the result of the most appropriate algorithms, the planning and direction of applied research are smoother. The most precise and preferred algorithms have also made it possible to divide duties and responsibilities among the elites of society, who conduct higher-quality research in aerospace scientific associations, universities, research centers, and industrial organizations, which will reduce prediction error. The results of my research contribute to solving the problem of identifying and ranking high-quality aerospace researchers, which is important for the development and innovation of the aerospace industry and the advancement of scientific knowledge in this field. The results of my research provide a new understanding of the role and importance of various scientometric indicators in evaluating the performance and impact of aerospace researchers, which measures the researchers' effectiveness in producing high-quality publications. The results of my research have practical and commercial implications for the aerospace sector, as they can be used to design and implement effective policies and strategies for supporting and rewarding outstanding researchers and attracting and retaining talented researchers in this domain. Also suggests some directions for future research, such as exploring the applicability and generalizability of the proposed model and methods to other scientific domains, or investigating the temporal and spatial dynamics of the scientometric indicators and their relation to the quality and impact of the researchers. One of the limitations of the present research was that it focused on metrics and indices that evaluate high-quality researchers, and numerous metrics related to journal evaluation could not be analyzed. On the other hand, in the present study, data related to aerospace researchers were extracted from WOSCC. In contrast, other databases, such as Scopus or SciVal, also contain information about aerospace researchers.

Further study

It is recommended that the data related to aerospace researchers be extracted from Scopus or SciVal in independent research and subjected to DM. Moreover, it is possible to identify and classify some criteria mentioned in the evaluation of journals, some of which are discussed below:

"Weighted impact factor," " Journal Citation Report (JCR)," "Category Normalized

Citation Impact (CNCI)," "Journal Citation Indicator (JCI)," "Citation Score Normalized by Cited References (CSNCR)," "Eigenfactor Score (EF)," "SCImago," "SJR" index, and "Immediacy Index." This classification can provide valuable evaluation data for aerospace researchers and identify the strongest and weakest metrics based on the LightGBM algorithm.

Acknowledgments

We appreciate the constructive collaboration with the anonymous reviewers and the editors.

Funding

This work does not have any financial support.

Conflict of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

***** The present article was extracted from a Ph.D. thesis entitled "Evaluation of High-quality Researchers in Aerospace Field Using Taxonomy of Metrics Approach and Hybrid Techniques of Scientometrics and Data Mining," which is being implemented presently at Khrazmi University, Iran.

Endnotes

1. https://images.webofknowledge.com/images/help/WOS/hp_subject_category_terms_tasca.html
2. http://datalab.mimfa.net
3. In the theory of probability and statistics, variance is a kind of dispersion measurement and is a number that shows how a series of data is spread around the average value.
4. The normal distribution (Gaussian distribution) is one of the most important continuous probability distributions in probability theory. This distribution shows that data close to the average occur more often than data far from the average.
5. Regression logistic is one of the classification methods in the topic of supervised machine learning.

References

Alonso-Valdivielso, M. Á. & Antonio, E. G. (2010, October). Why include bibliometric analysis in the activities of a library specialized in astronomy? -Notes From the libraries of INTA. In *Library and Information Services in Astronomy VI: 21st Century Astronomy Librarianship, From New Ideas to Action* (Vol. 433, p. 95).

Amani, M. & Salama, S. (2019). Discovering performance evaluation features of faculty members using data mining techniques to support decision making. *International Journal of Computer Applications*, 178(49), 25-29. https://doi.org/10.5120/ijca2019919417

Ameli, O. (2020). *A study assessing the characteristics of big data environments that predict high research impact: application of qualitative and quantitative methods*. Doctoral dissertation. Boston University

Amir, A. R. & Weiss, S. I. (2023, July 1). Aerospace industry. Encyclopedia Britannica. Retrieved from https://www.britannica.com/technology/aerospace-industry

Arfaee, M., Bahari, A. & Khalilzadeh, M. (2021). A novel prediction model for educational planning of human resources with data mining approach: A national tax administration case study. *Education and Information Technologies*, 27(2), 2209–2239. https://doi.org/10.1007/s10639-021-10699-6

Asanbe, M. O., Osofisan, A. & William, W. F. (2016). Teachers' performance evaluation in higher educational institution using data mining technique. *International Journal of Applied Information Systems,* 10(7), 10-15. https://doi.org/10.5120/ijais2016451524

Bavisi, S., Mehta, J. & Lopes, L. (2014). A comparative study of different data mining algorithms. *International Journal of Current Engineering and Technology*, 4(5), 3248–3252. Retrieved from https://inpressco.com/wp-content/uploads/2014/09/Paper293248-3252.pdf

Bazeley, P. (2010). Conceptualizing research performance. *Studies in Higher Education*, 35(8), 889-903. https://doi.org/10.1080/03075070903348404

Biglu, M. H. (2008). Editor Scientometric study of patent literature in medicine. In H. Kretschmer & F. Havemann (Eds.): In *Proceedings of WIS 2008, Berlin. Fourth International Conference on Webometrics, Informetrics and Scientometrics & Ninth COLLNET Meeting Humboldt-Universität zu Berlin, Institute for Library and Information Science (IBI*)

Bornmann, L., Mutz, R. & Daniel, H. D. (2008). Are there better indices for evaluation purposes than the h index? A comparison of nine different variants of the h index using data from biomedicine. *Journal of the American Society for Information Science and Technology*, 59(5), 830–837. https://doi.org/10.1002/asi.20806

Breiman, L. (1996). *Bias, variance, and arcing classifiers*. University of California, Berkeley, CA. Retrieved from https://www.stat.berkeley.edu/users/breiman/arcall96.pdf

Birkle, C., Pendlebury, D. A., Schnell, J. & Adams, J. (2020). Web of Science as a data source for research on scientific and scholarly activity. *Quantitative Science Studies*, 1(1): 363–376. https://doi.org/10.1162/qss_a_00018

Britannica, T. Editors of Encyclopaedia (2023, August 19). National Aeronautics and Space Administration. Encyclopedia Britannica. Retrieved from https://www.britannica.com/topic/NASA

Cabena, P., Hadjinian, P., Stadler, R., Verhees, J. & Zanasi, A. (1998). *Discovering data mining: from concept to implementation*. Prentice-Hall, Inc.

Carlsson, H., Kettis, Å. & Söderholm, A. (2014). Research quality and the role of the university leadership. Sveriges universitets- och högskoleförbund (SUHF). Retrieved from https://suhf.se/app/uploads/2019/07/Expertgruppen-kvalitet-2010-2011-Bilaga-4-Research-Quality-and-the-Role-of-the-University-Leadership.pdf

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R. (2000). CRISP-DM 1.0 Step-by-step data mining guide. SPSS Inc. Retrieved from https://www.kde.cs.uni-kassel.de/wp-content/uploads/lehre/ws2012-13/kdd/files/CRISPWP-0800.pdf

Chawla, N.V. (2009). Data mining for imbalanced datasets: An overview. In Maimon, O., Rokach, L. (eds) *Data Mining and knowledge discovery handbook*. Springer, Boston, MA. https://doi.org/10.1007/978-0-387-09823-4_45

Chen, T., Xu, J., Ying, H., Chen, X., Feng, R., Fang, X., Gao, H. & Wu, J. (2019). Prediction of extubation failure for intensive care unit patients using light gradient boosting machine. *IEEE Access*, 7, 150960-150968. Retrieved from https://www.zyicu.cn/wordpress/wp-content/uploads/2020/09/Prediction-of-Extubation-Failure-for-Intensive-Care-Unit-Patients-Using-Light-Gradient-Boosting-Machine.pdf

Chia, V. S. (2020). *New metrics for assessing high-quality researchers*. Professional Doctorate thesis, Queensland University of Technology.

Codina, L., Morales-Vargas, A., Rodríguez-Martínez, R. & Pérez-Montoro, M. (2020). Uso de Scopus y Web of Science para investigar y evaluar en comunicación social: análisis comparativo y caracterización. *Index.comunicación*, 10(3), 235-261. https://doi.org/10.33732/ixc/10/03Usodes [in Spanish]

Dehghani Ashkezari, A. (2016). Predicting the performance of faculty members using data mining (Case study: University of Science and Art: Yazd). The Thesis of M.Sc. in Science & Art Department: Management, Yazd University, Yazd. [in Persian]

Egghe, L. & Rousseau, R. (2008). An h-index weighted by citation impact. *Information Processing and Management*, 44(2), 770–780. https://doi.org/10.1016/j.ipm.2007.05.003

Elkan, C. (2001, August). The foundations of cost-sensitive learning. In *International Joint Conference on Artificial Intelligence (*Vol. 17, No. 1, pp. 973-978). Lawrence Erlbaum Associates Ltd.

Galyani-Moghaddam, G. & Taheri, P. (2015). Mapping co-authorship network and scientific collaborative coefficient of Iranian researchers in the field of aerospace in the Science Citation Index to 2014. *Knowledge Retrieval and Semantic Systems*, 2(3), 23-42. [In Persian]

Ganguli, R. (2008). A scientometric analysis of recent aerospace research. *Current Science*, 95(12), 1670-1672. Retrieved from https://www.currentscience.ac.in/Volumes/95/12/1670.pdf

Ghavidel, S., Riahinia, N., Danesh, F. & Noroozi Chakoli, A. (2022). Aerospace: The study of scientometrics and an analysis of centrality indicators of the co-authorship network of researchers. *Scientometrics Research Journal*, 9(2), 165-204. https://doi.org/10.22070/rsci.2022.15902.1568 [in Persian]

Guruvayur, S. R. & Suchithra, R. (2017, May). A detailed study on machine learning techniques for data mining. In *2017 International Conference on Trends in Electronics and Informatics (ICEI)* (pp. 1187-1192). IEEE.

Guns, R. & Rousseau, R. (2009). Real and rational variants of the h-index and the g-index. *Journal of Informetrics*, 3(1), 64–71. https://doi.org/10.1016/j.joi.2008.11.004

Hagberg, A., Schult, D. & Swart, P. (2021). NetworkX Reference, Release 2.6. 2. Technical Report.

Hicks, D., Wouters, P., Waltman, L., De Rijcke, S. & Rafols, I. (2015). Bibliometrics: the Leiden manifesto for research metrics. *Nature*, 520(2), 429–431. https://doi.org/10.1038/520429a

Hicks, D. & Melkers, J. (2013). Bibliometrics as a tool for research evaluation. In *Handbook on the theory and practice of program evaluation* (pp. 323-349). Edward Elgar Publishing.

Hirsch, J. E. (2019). h_α: An index to quantify an individual's scientific leadership. *Scientometrics*, 118(2), 673–686. https://doi.org/10.1007/s11192-018-2994-1

Ivancheva, L. (2008). Scientometrics today: A methodological overview. *Collnet Journal of Scientometrics and Information Management*, 2(3), 47-56. https://doi.org/10.1080/09737766.2008.10700853

Jaggia, S., Kelly, A., Lertwachara, K. & Chen, L. (2020), Applying the crisp-dm framework for teaching business analytics. *Decision Sciences Journal of Innovative Education*, 18(4), 612-634. https://doi.org/10.1111/dsji.12222

James, G., Witten, D. & Hastie, T. (2013). *An introduction to statistical learning with applications in R.* New York, Springer-Verlag.

Joseph, S. R., Hlomani, H. & Letsholo, K. (2016). Data mining algorithms: An overview. *International Journal of Computers & amp; Technology*, 15(6), 6806–6813. https://doi.org/10.24297/ijct.v15i6.1615

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. & Liu, T.-Y. (2017) LightGBM: A highly efficient gradient boosting decision tree. *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Long Beach, CA, December 2017, (pp. 3149-3157).

Kirimi, J.M. & Moturi, C. (2016). Application of data mining classification in employee performance prediction. *International Journal of Computer Applications*, 146(2), 28-35. https://doi.org/10.5120/ijca2016910883

Khorrami, M. (2018). *Scimago analysis to extract geographical and topical similarity relations*. Master's Thesis, Science in Information Technology Engineering-Electronic Business, Faculty of Engineering and Technology, University of Qom. [in Persian]

Li, K. (2020). Research materiality in citation contexts: A quantitative examination based on full-text psychological papers. Doctoral Dissertation. Drexel University. https://doi.org/10.17918/00000265

Liu, P., Qingqing, W. & Liu, W. (2021). Enterprise human resource management platform based on FPGA and data mining. *Microprocessors and Microsystems*, 80(3), 103330. https://doi.org/10.1016/j.micpro.2020.103330

Ma, H. (2021). Enterprise human resource management based on big data mining technology of internet of things. *Journal of Intelligent & Fuzzy Systems*, 3(4),1-7. https://doi.org/10.3233/jifs-219096

Maksood, F. Z. & Achuthan, G. (2016). Analysis of data mining techniques and its applications. *International Journal of Computer Applications*, 140 (3), 6-14.

Mazurek, J. (2018). A modification to Hirsch index allowing comparisons across different scientific fields. *Current Science*, 114(11), 2238–2239. https://doi.org/10.18520/cs%2Fv114%2Fi11%2F2238-2239

Mohamadesmaeil, S. (2020). Data analysis of information behaviors of soft power, society and security research groups of the institute of humanities, cultural and strategic studies of the Ministry of Science, Research and Technology. *Sciences and Techniques of Information Management*, 6(3), 59-80. https://doi.org/10.22091/stim.2020.4014.1290

Mryglod, O., Holovatch, Y. & Kenna, R. (2018, August). Data mining in scientometrics: Usage analysis for academic publications. In *2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP)* (pp. 241-246). IEEE.

Nachev, A. & Teodosiev, T. (2018). Analysis of employment data using support vector machines. *International Journal of Applied Engineering Research*, 13(18), 13525-13535. Retrieved from https://www.ripublication.com/ijaer18/ijaerv13n18_17.pdf

Naghshineh, V. (2007, 25 June). Applications and characteristics of aerospace technology: A searching look at the earth. Iran News. Retrieved from https://www.magiran.com/article/1429509 [In Persian]

Nijs, S., Dries, N., Van Vlasselaer, V. & Sels, L. (2021). Reframing talent identification as a status-organizing process: Examining talent hierarchies through data mining. *Human Resource Management Journal*, 32(1), 169-193. https://doi.org/10.1111/1748-8583.12401

Osareh, F. (2006). Collaboration in astronomy knowledge production: A case study in ScienceDirect from 2000-2004. In P. Ingwersen, B. Larsen (Eds), *Proceedings of ISSI 2005 – the 10th International Conference of the International Society for Scientometrics and Informetrics*, Vol. 2, Stockholm, Sweden, 24-28 July 2005, Karolinska University Press, 2036, 660-661. Retrieved from https://www.issi-society.org/proceedings/issi_2005/Osareh_ISSI2005.pdf

Pelicioni, L. C., Ribeiro, J. R., Devezas, T., Belderrain, M. C. N. & de Melo, F. C. L. (2018). Application of a Bibliometric Tool for Studying Space Technology Trends. *Journal of Aerospace Technology and Management*, 10, e0318. https://doi.org/10.5028/jatm.v10.830

Pérez, J., Iturbide, E., Olivares, V., Hidalgo, M., Almanza, N. & Martínez, A. (2015) A data preparation methodology in data mining applied to mortality population databases. In: Rocha A., Correia A., Costanzo S., Reis L. (Eds) *New Contributions in Information Systems and Technologies. Advances in Intelligent Systems and Computing*, Vol. 353. Springer, Cham.

Perry, M. & Reny, P. J. (2016). How to count citations if you must. *American Economic Review*, 106 (4), 2722–2741. https://goi.org/10.1257/aer.20140850

Pessach, D., Singer, G., Avrahami, D., Ben-Gal, H. C., Shmueli, E. & Ben-Gal, I. (2020). Employee recruitment: A prescriptive analytics approach via machine learning and mathematical programming. *Decision Support Systems*, 134, 113290. https://doi.org/10.1016%2Fj.dss.2020.113290

Pino-Mejías, R., M. D. Cubiles-de-la-Vega, M. Anaya-Romero, A. Pascual-Acosta, A. Jordán-López & N. Bellinfante-Crocci (2010). Predicting the potential habitat of oaks with data mining models and the R system. *Environmental Modelling & Software*, 25(7), 826-836. https://doi.org/10.1016/j.envsoft.2010.01.004

Qian, X. & Ohwada, H. (2018, February). Application of data mining classification in job-changing. In *Proceedings of the 2018 10th International Conference on Machine Learning and Computing* (pp. 107-110).

Rokach, L. (2010). Ensemble-based classifiers. *Artificial Intelligence Review*, 33, 1-39. https://doi.org/10.1007/s10462-009-9124-7

Rousseau, R. & Ye, F. (2008). A proposal for a dynamic h-type index. *Journal of the American Society for Information Science and Technology*, 59(11), 1853–1855. Retrieved from https://sci2s.ugr.es/sites/default/files/files/TematicWebSites/hindex/Rousseauetal2008.pdf

Saad, H. R. (2018). Use the bagging algorithm to improve prediction accuracy for the evaluation of worker performances at a production company. *Industrial Engineering and Management*, 7(2), 1-7. https://doi.org/10.4172/2169-0316.1000257

Sapkota, N., Alsadoon, A., Prasad, P. W. C., Elchouemi, A. & Singh, A. K. (2019, February). Data summarization using clustering and classification: Spectral clustering combined with k-means using NFPH. In *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)* (pp. 146-151). IEEE. https://doi.org/10.1109/COMITCon.2019.8862218

Shaheen, M., Ahsan, A. & Iqbal, S. (2021). Data mining of scientometrics for classifying science journals. *Intelligent Automation & Soft Computing*, 28(3), 873–885. https://doi.org/10.32604/iasc.2021.016622

Stančin, I. & Jović, A. (2019, May). An overview and comparison of free Python libraries for data mining and big data analysis. In *2019 42nd International convention on information and communication technology, electronics and microelectronics (MIPRO)* (pp. 977-982). IEEE. https://doi.org/10.23919/MIPRO.2019.8757088

Sidiropoulos, A., Katsaros, D. & Manolopoulos, Y. (2007). Generalized Hirsch h-index for disclosing latent facts in citation networks. *Scientometrics*, 72(2), 253–280. https://doi.org/10.1007/s11192-007-1722-z

Sinatra, R., Wang, D., Deville, P., Song, C. & Barabási, A.L. (2016). Quantifying the evolution of individual scientific impact. *Science*, 354(6312), aaf5239. https://doi.org/10.1126/science.aaf5239

Strohmeier, S. & Piazza, F. (2013). Domain driven data mining in human resource management: A review of current research. *Expert Systems with Applications*, 40(7), 2410-2420. https://doi.org/10.1016/j.eswa.2012.10.059

Syarif, I., Zaluska, E., Prugel-Bennett, A. & Wills, G. (2012). Application of bagging, boosting and stacking to intrusion detection. In *Machine Learning and Data Mining in Pattern Recognition: 8th International Conference, MLDM 2012, Berlin, Germany, July 13-20, 2012. Proceedings 8* (pp. 593-602). Springer Berlin Heidelberg.

Tesema, W. (2019). Inefficiency of data mining algorithms and its architecture: With emphasis to the shortcoming of data mining algorithms on the output of the researches. *Applied Computer Science*, 15(3), 73-86. https://doi.org/10.23743/acs-2019-23

Van Eck, N. J. & Waltman, L. (2008). Generalizing the h- and g-indices. *Journal of Informetrics*, 2(4), 263–271. https://doi.org/10.1016/j.joi.2008.09.004

Vaziri, I. & Rajabali Baglo, R. (2010). Aerospace engineering of Iran and the world in the mirror of scientology: A study in citation databases. Paper presented at *the 10th conference of Iran Aerospace Society*. March 10-12, 2010, Tehran. [In Persian]

Vaziri, I. (2010). Iranian science in the subject group of aerospace engineering at the international level: A scientometric study based on the statistics of the Institute of Scientific Information (ISI). In *Special issue of the conference on the employment status of aerospace graduates*, October 26, 2010, Amirkabir University, Tehran. [In Persian]

Vavilova I. B., Zievako V. S., Pakuliak, L. K. & Potapovych, L. P. (2020). "Space Science and Technology" journal: Statistics and Scientometrics for 1995–2020. *Space Science and Technology*, 26(6), 94-103. https://doi.org/10.15407/knit2020.06.094

Waltman, L. & Noyons, E. (2018). Bibliometrics for research management and research evaluation: A brief introduction. Leiden, Netherlands: Centre for Science and Technology Studies. Retrieved from https://www.cwts.nl/pdf/CWTS_bibliometrics.pdf

Web of Science Core Collection (2021). *Categories & Collections (Scope Notes).* Retrieved from https://mjl.clarivate.com/help-center

Wirth, R. & Hipp, J. (2000, April). CRISP-DM: Towards a standard process model for data mining. In *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining* (Vol. 1, pp. 29-39).

Wu, Y., Wang, Z. & Wang, S. (2021). Human resource allocation based on fuzzy data mining algorithm. *Complexity*, 9489114. https://doi.org/10.1155/2021/9489114

Yaminfrooz, M. & Gholinia, H. (2015). Multiple h-index: A new scientometric indicator. *The Electronic Library*, 33(3), 547-556. https://doi.org/10.1108/EL-07-2013-0137

Yoosin, K., Yeonjin, SeongGwan & Seung, R. J. (2017). Practical text mining for trend analysis: Ontology to visualization in aerospace technology. *KSII Transactions on Internet and Information Systems*, 11(8), 4133-4145. https://doi.org/10.3837/tiis.2017.08.022

Zhou, Z. H. (2012). *Ensemble methods foundations and algorithms*. Chapman and Hall/CRC.