

**T.C.
İSTANBUL KÜLTÜR UNIVERSITY
INSTITUTE OF GRADUATE STUDIES**

**ANALYSIS OF ACADEMIC STAFF PERFORMANCE WITH
DATA MINING**

Masters of Applied Science Thesis

Yasemin KALENDER

160004051

Department: Industrial Engineering

Program: Engineering Management

Supervisor: Assist. Prof. Dr. Zeynep GERGİN

JULY 2024

T.C.
İSTANBUL KÜLTÜR UNIVERSITY
INSTITUTE OF GRADUATE STUDIES

ANALYSIS OF ACADEMIC STAFF PERFORMANCE WITH
DATA MINING

Masters of Applied Science Thesis

Yasemin KALENDER

1600004051

Department: Industrial Engineering

Program: Engineering Management

Supervisor and Chairperson: Assist. Prof. Dr. Zeynep GERGİN

Members of Examining Committee: Assist. Prof. Dr. Zeynep GERGİN

Prof. Dr. Fadime YÜKSEKTEPE

Assist. Prof. Dr. Sinem BÜYÜKSAATÇI KİRİŞ

JULY 2024

ACKNOWLEDGEMENT

I would like to express my special thanks to my advisor Assist. Prof. Zeynep GERGİN, who guided me with her valuable knowledge in researching, determining, and detailing my master's thesis, as well as inspiring me to achieve optimal outcomes.

I am also thankful to Prof. Dr. Fadime YÜKSEKTEPE for their valuable contributions.

I am thankful to all my professors at Istanbul Kültür University, Department of Industrial Engineering, for providing me with knowledge that will be instrumental in shaping my personal and academic journey throughout my undergraduate and graduate education.

I am eternally grateful to my biggest chance in life, my mother and father, Nevin KALENDER and Cevdet KALENDER, and my brother Burak KALENDER who have brought me to these days and have never withheld their support from me.

Moreover, I am immensely thankful to Hasan Mürsel TÜRKİŞ, one of the most treasured individuals in my life, for his encouragement and support to me throughout this demanding course.

04.07.2024

Yasemin KALENDER

TABLE OF CONTENTS

ACKNOWLEDGEMENT	iii
LIST OF TABLES	vi
LIST OF FIGURES	vii
LIST OF SYMBOLS	ix
ÖZET	x
ABSTRACT	xi
1. INTRODUCTION	1
2. LITERATURE REVIEW	3
2.1. Data Mining	3
2.2. Educational Data Mining	5
2.3. Clustering Technique in Educational Data Mining	12
3. METHODOLOGY	15
3.1. Data Collection	15
3.2. Data Statistics and Preprocessing	16
3.3. Clustering	17
3.3.1. The Elbow Method	17
3.3.2. K-Means Algorithm	18
3.3.3. Euclidean Distance	19
3.3.4. K-Means++ Initialization Method	20
4. IMPLEMENTATION AND RESULTS	21
4.1. The Dataset	21
4.2. Preprocessing and Input Data	27
4.2.1. Book Score.....	28
4.2.2. Article Score	30
4.2.3. Proceeding Score	33
4.2.4. Course Score	34
4.2.5. Supervised Theses Score	36
4.2.6. Administrative Responsibility Score	37
4.2.7. Data Statistics	39
4.3. K-Means Clustering	41
4.3.1. Identifying the Cluster Size	41

4.3.2. Output Data.....	43
4.4. Interpretation and Discussion of the Results	45
4.4.1. Identifying Performance Segments.....	45
4.4.2. Discussion of the Results.....	50
5. CONCLUSION	59
BIBLIOGRAPHY	61



LIST OF TABLES

Table 4.1: Research domain and faculty associations	23
Table 4.2: Initial and merged data points	25
Table 4.3: Book types and points.....	28
Table 4.4: Article types and points	30
Table 4.5: Proceeding types and points	33
Table 4.6: Course types and points	35
Table 4.7: Administrative types and points	38
Table 4.8: Attributes and their characteristics	40
Table 4.9: Sum of Squared Errors of different k and seed values	42
Table 4.10: The distribution of academic staff across clusters.....	44
Table 4.11: The segments for book scores	45
Table 4.12: The segments for article scores	46
Table 4.13: The segments for proceeding scores.....	46
Table 4.14: The segments for course scores	47
Table 4.15: The segments for supervised theses score scores	47
Table 4.16: The segments for administrative responsibility scores.....	48
Table 4.17: Clustering results showing the attribute averages	49
Table 4.18: Clustering results categorization under performance segments	49

LIST OF FIGURES

Figure 2.1: Data mining techniques	3
Figure 2.2: Affecting factors of teacher’s performance.....	9
Figure 3.1: Flowchart of the methodology	15
Figure 3.2: Screenshot of preprocessing in WEKA.....	16
Figure 3.3: Elbow point	18
Figure 3.4: WEKA K-Means Algorithm	19
Figure 3.5: Euclidean distance.....	20
Figure 4.1: Gender distribution of academic staff	22
Figure 4.2: Academic staff and their faculties distribution	22
Figure 4.3: Input data screenshot.....	27
Figure 4.4: Screenshot of data for published book types and corresponding scores	29
Figure 4.5: Screenshot of data for published articles and corresponding scores.....	32
Figure 4.6: Screenshot of data for published proceedings and corresponding scores	34
Figure 4.7: Screenshot of data for given courses and corresponding scores	34
Figure 4.8: Screenshot of data for supervised graduate thesis types and corresponding scores	36
Figure 4.9: Screenshot of data for administrative responsibilities and corresponding scores	39
Figure 4.10: Histogram diagram of attributes.....	41
Figure 4.11: Clustering SSE Graph for K Means Initialization Method with Seed:10 Value	43
Figure 4.12: Cluster distribution of female academic staff	52
Figure 4.13: Cluster distribution of male academic staff.....	52
Figure 4.14: Proportion of academic staff in Cluster 1 based on research domain	54
Figure 4.15: Proportion of academic staff in Cluster 2 based on research domain	55

Figure 4.16: Proportion of academic staff in Cluster 3 based on research domain56
Figure 4.17: Proportion of academic staff in Cluster 4 based on research domain57
Figure 4.18: Proportion of academic staff in Cluster 5 based on research domain58



LIST OF SYMBOLS

x_i	Data points
c_i	The closest point to the centroid
S	Sets of observations
k	number of sets of observations
x	Observation data point
μ_i	Mean of points in S_i
d	Euclidean distance
x_1	The first point's coordination in the x-axis
y_1	The first point's coordination in the y-axis
x_2	The second point's coordination in the x-axis
y_2	The second point's coordination in the y-axis

Üniversite : İstanbul Kültür Üniversitesi
Enstitü : Lisansüstü Eğitim Enstitüsü
Anabilim Dalı : Endüstri Mühendisliği
Programı : Mühendislik Yönetimi
Tez Danışmanı : Dr. Öğretim Üyesi Zeynep GERGİN
Tez Türü ve Tarihi : Yüksek Lisans – Temmuz 2024

ÖZET

VERİ MADENCİLİĞİ İLE AKADEMİK PERSONEL PERFORMANSININ ANALİZİ

Yasemin KALENDER

Literatürde akademik personelin performansının değerlendirilmesinde birkaç ana kategori ve ilgili alt faktörler değerlendirilmektedir. Bu çalışma, İstanbul'da faaliyet gösteren bir üniversitedeki akademisyenlerin yayınlar (makaleler, kitaplar ve bildiriler), ders yükleri, tez danışmanlıkları ve idari sorumluluklar gibi üniversitenin performans değerlendirme yönergesine bağlı kalınarak seçilmiş altı ana nitelik aracılığıyla performanslarını analiz etmeye odaklanmıştır. Kapsamlı bir veri seti oluşturmak için belirtilen niteliklere ilişkin veriler Yüksek Öğretim Kurumu (YÖK) ve üniversite kayıtlarından elde edilmiştir. Başlangıçta dokuz farklı araştırma alanından 199 akademisyeni temsil eden 12664 veri noktasından oluşan veri seti, 1194 örnek içeren girdi verisine dönüştürülmüştür. Daha sonra veri seti WEKA yazılımı kullanılarak K-means algoritması aracılığıyla beş gruba kümelenecek, oluşan kümelerin farklı performans özellikleri tanımlanarak çıktılar tartışılmıştır. Nihai tartışmalarda akademik personelin cinsiyeti ve araştırma alanları da dikkate alınmıştır. Stratejik kararlar konusunda fikir vermek amacıyla bulgular Üniversite Yönetimi'ne sunulmuştur.

Anahtar Kelimeler: Veri Madenciliği, Eğitsel Veri Madenciliği, K-Means Kümeleme Algoritması, Akademik Personel Performansı, Performans Analizi

University : İstanbul Kültür University
Institute : Institute of Graduate Studies
Department : Industrial Engineering
Program : Engineering Management
Supervisor : Assist. Prof. Zeynep GERGİN
Degree Awarded and Date : MS – July 2024

ABSTRACT

ANALYSIS OF ACADEMIC STAFF PERFORMANCE WITH DATA MINING

Yasemin KALENDER

This study focuses on analyzing the performance of academics at a university in Istanbul based on six main criteria aligned with the university's performance evaluation guidelines. These criteria include publications (articles, books, and conference papers), teaching load, supervision of theses, and administrative responsibilities. A comprehensive dataset was compiled using data sourced from the Higher Education Council (YÖK) and university records pertaining to these criteria. The dataset, which initially consisted of 12664 entries representing 199 academics from nine different research domains, is transformed into input data containing 1194 entries. Subsequently, the dataset is clustered into five groups via K-means algorithm utilizing WEKA software. The distinct performance characteristics of the clusters are identified and discussed. The genders and research domains of the academic staff are also considered in the final discussions. The findings were presented to the University Administration in order to provide insight into strategic decisions.

Keywords: Data Mining, Educational Data Mining, K-means Clustering Algorithm, Academic Staff Performance, Performance Analysis

1. INTRODUCTION

The evaluation of academic staff performance according to the academic performance criteria of universities provides insights for strategic decisions for the overall status of faculties and individual development. The assessment of performance for academic staff involves several primary categories and their respective subfactors. The objective of this thesis study is to determine the scores of academic staff across six main academic performance criteria such as book score, article score, proceeding score, course score, supervised theses score, and administrative score. Subsequently, data mining methodology is used to classify academic staff at the university based on their distinct characteristics.

K-means clustering algorithm is adopted to classify the performance of academic staff across various research domains, and the methodology is implemented for a university in Istanbul (Turkey) that provides education for 27 years, with a total of 705 academic staff, including 62 full-time Professors, 32 Associate Professors, 140 Assistant Professors, 296 part-time faculty members, 81 lecturers, and 94 research assistants in 8 faculties, two vocational schools for undergraduate degrees together with a Graduate Education Institute for the master's and doctorate degrees.

The contribution of this research to academic literature lies in its comprehensive evaluation of academic staff performance using a multidimensional approach. By focusing on six main attributes —book score, article score, proceeding score, course score, supervised theses score, and administrative score— this study offers a nuanced understanding of the factors influencing academic performance and the relation between them. Furthermore, the methodology employed, including dataset construction from authoritative sources like the Higher Education Council and university records, as well as the utilization of data mining techniques, adds depth and rigor to the research findings. The identification of distinct clusters based on performance attributes, coupled with considerations of gender and the research

domains of the staff, provides valuable insights for practical solutions for the strategic decisions of the university management.

Overall, this research contributes to academic literature by offering a systematic and data-driven approach to understanding and evaluating academic staff performance, thereby informing future studies and practices in the field of higher education management and administration.

The thesis report is divided into five main chapters as follows;

The introduction provides an overview of the topic, outlines the main objectives and purposes of the research, and discusses the contribution of the study to the academic literature.

Literature review consists of three main sections namely data mining, educational data mining, and clustering techniques in educational data mining, providing a transition from general to specific approaches. In the data mining section, various data mining techniques are explained, while the educational data mining section focuses on studies conducted in the field of education. The clustering technique in the educational data mining section examines studies conducted with clustering algorithms in the field of education, and this section generally summarizes completed studies related to the research topic in the past.

The methodology delivers the procedure of the study while the implementation and results section encompass the formation of the dataset from various data sources, preprocessing, and statistical analyses. Additionally, the clustering process using the K-means algorithm and findings are discussed in conjunction with the gender and research domain sections.

In the conclusion section, the overall scope of the study is summarized, and an overview of the applied algorithms and methods is provided. The outputs of the study are examined, and the potential changes in the characteristics of attributes in clusters are discussed based on the strategies that the university can adopt. Finally, recommendations are made for expanding the scope of future studies.

2. LITERATURE REVIEW

2.1. Data Mining

Data mining involves the systematic examination of extensive data sets in order to uncover valuable patterns and connections that can be utilized to address business challenges through thorough data analysis. By employing various techniques and tools, enterprises can leverage data mining to anticipate future trends and enhance their decision-making processes (Gillis et al., 2024).

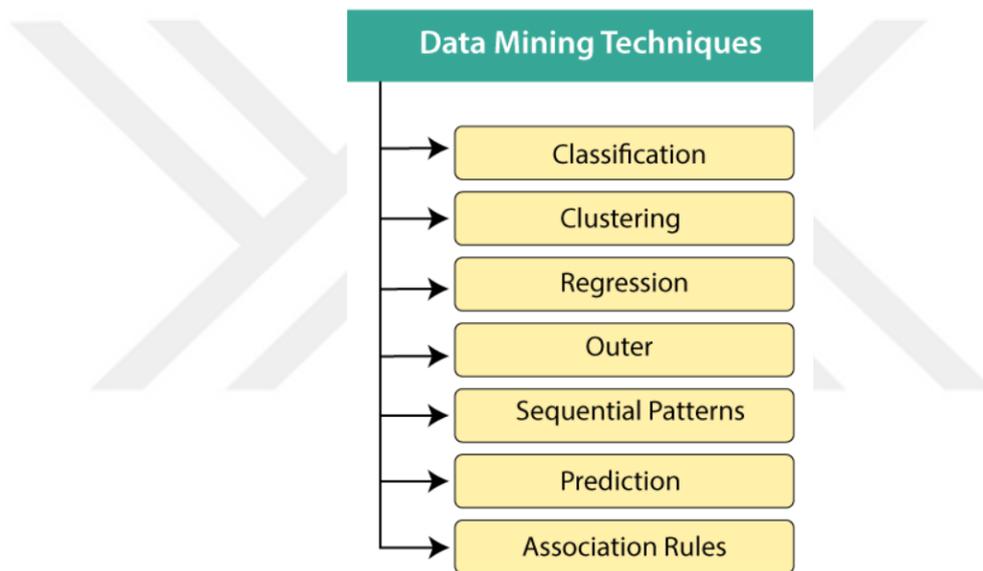


Figure 2.1: Data mining techniques

Recent research in data mining has shown that the selection of data mining techniques is based on the specific characteristics and objectives of the study. These methodologies are illustrated in Figure 2.1.

Classification is an analytical procedure that entails examining data in order to detect patterns and connections. The primary aim of classification is to construct a model that can forecast the class or category of fresh data instances by considering their attributes or characteristics. Classification is a form of supervised learning, which implies that it employs a dataset with labels to develop a predictive model. The classification techniques are Decision Trees, Naïve Bayes, K-Nearest Neighbors (KNN), Support Vector Machines (SVM), and Random Forest (Kumar, 2023).

Clustering involves the process of categorizing data points into distinct groups based on specific criteria, such as shared characteristics or features. These groups, known as clusters, consist of data points that exhibit similarities within the same cluster and differences from data points in other clusters. Through clustering, one can determine the data distribution and characteristics of each cluster and examine unique clusters (Tang & Engelbrecht, 2022).

Regression analysis is a statistical method used to estimate the connection between variables that have a cause and effect relationship. The primary objective of univariate regression is to examine the relationship between a dependent variable and a single independent variable, and establish a linear equation that represents this relationship (Güler & Uyanık, 2013)

Outliers are data points that significantly deviate from the rest of the dataset. The process of identifying outliers is also referred to as anomaly detection. Outlier detection complements the clustering process, which focuses on identifying anomalies and minority groups, whereas clustering primarily groups larger datasets based on their similarity to one another (Al-jabery et al., 2020).

Sequential Pattern Mining highlights the necessity to detect and analyze frequent subsequences of events, items, or tokens occurring within an ordered metric space, which are commonly encountered in datasets.

Predictive analytics is a statistical discipline that revolves around extracting information to forecast trends and behavioral patterns. The development of predictive web analytics entails the calculation of statistical probabilities for future online events. Techniques such as data modeling, machine learning, AI, deep learning algorithms, and data mining are encompassed within the statistical methodologies of predictive analytics. This approach is applicable to various types of unknown data, whether it relates to the past, present, or future, with a primary emphasis on events of interest that are anticipated to happen in the future (Ganesh & Reddy, 2022).

Association Rule mining has emerged as a prominent approach in the realm of data mining. Its primary objective is to detect patterns, correlations, and causal structures among items stored in transaction and relational databases. Through the

analysis of these databases, association rule mining uncovers noteworthy relationships between variables, thereby offering valuable insights into extensive datasets (Kulkarni & Mundhe, 2017).

2.2. Educational Data Mining

The incorporation of a performance evaluation system within educational institutions has the potential to enhance the professional growth and advancement of faculty members, thereby benefiting both the institution and the broader education sector. Evaluation data is gathered through a variety of methods, including surveys, ongoing assessments, and end-of-term evaluations, in addition to managerial assessments. Feedback obtained is scrutinized to identify strengths and areas for improvement in professional, scholarly, and other faculty competencies, with appropriate weighting applied during the analysis process. Utilization of Data Mining Techniques is encouraged for effective performance appraisal procedures (Agrawal et al., 2020).

Klašnja-Milićević et al. (2017) explored the relevance of data science and summaries in the context of Big Data and Learning Analytics within the educational domain. They underscored the potential benefits of leveraging big data methodologies in education, emphasizing the substantial advantages that could be derived. Furthermore, the researchers introduced a fundamental framework designed to bolster educational research initiatives.

In their research, Kaur et al. (2015) focused on the identification of slow learners among students. They developed a predictive data mining model using classification-based algorithms to accomplish this. The researchers utilized a dataset of student academic records and applied it to various classification algorithms, such as Multilayer Perception, Naïve Bayes, SMO, J48, and REPTree, using the open-source tool WEKA. The study generated statistics based on all the classification algorithms and conducted a comparison of the five classifiers to predict accuracy and determine the best performing algorithm. Furthermore, the paper presented a knowledge flow model that encompassed all five classifiers.

Bunkar et al. (2012) aimed to apply data mining processes, specifically classification, to enhance the quality of the higher educational system. By evaluating

student data, the main factors that potentially impact student performance in courses can be investigated. The classification rule generation process relies on the utilization of decision trees as a classification method, and the resulting rules were carefully examined and assessed. Furthermore, a system was developed to enable students to predict their final grades in a particular course. This research endeavored to improve the educational system by leveraging data mining techniques and predictive modeling.

Data mining serves as a potent instrument for data analysis and knowledge management, particularly within the educational sector. This research delves into the prediction of instructor performance and the identification of factors influencing student achievements to elevate the quality of the education system. Through the analysis of the Turkey Student Evaluation records dataset using diverse data classifiers like J48 Decision Tree, Multilayer Perception, Naïve Bayes, and Sequential Minimal Optimization, the study was able to make significant discoveries. The results indicate that employing attribute evaluation on the dataset enhances the accuracy of performance prediction, highlighting the crucial role of data mining tools in education (Ahmed et al., 2016).

The study explores the impact of Student Evaluations of Teaching (SET) on teaching quality in Saudi universities. It uses a narrative review of peer-reviewed journal articles to investigate its effectiveness. The study emphasizes the importance of including student feedback in enhancing teacher quality. The findings of Aljendan (2024), suggested that designing, administering, and following SET procedures are crucial for improving teaching quality.

The integration of data mining applications in higher education is increasingly common for analyzing and addressing educational and administrative issues. Nonetheless, the current research tends to prioritize the modeling of student performance over that of instructors. This study undertakes a comparison of four classification techniques, decision tree algorithms, support vector machines, artificial neural networks, and discriminant analysis, to establish classifier models for evaluating instructors' success. Additionally, the study highlights the crucial influence of students' interest in the course on instructors' overall success (Agaoglu, 2016).

Educational Data Mining serves as a valuable approach within educational environments, enabling the extraction of knowledge from educational databases. This

research paper investigates the application of data mining in Teacher Evaluation systems and presents the findings obtained through the utilization of the WEKA tool. The study critically analyzes the performance of a semester's final Teacher Evaluation using WEKA, aiming to gain practical insights into the implementation of data mining techniques in both E-learning systems and traditional education. By uncovering hidden patterns in student evaluations, the research aims to predict which teachers will be invited to faculty classes and which will be refused, thereby facilitating the identification of potential areas for cost reduction in education contract (Ahmedi & Shiri Ahmad Abadi, 2013).

Asanbe et al. (2016) unveiled a two-tiered classifier system aimed at evaluating and predicting the effectiveness of instructors in higher education by leveraging data mining technologies. This system, which integrates Artificial Neural Networks (ANN) and Decision Trees, underwent validation using case study data obtained from a university in Nigeria. The mining process was conducted using the WEKA machine learning tool, highlighting Working Experience and Rank as the key attributes influencing teachers' performance.

Haddawy et al. (2007) conducted a study in which they evaluated the performance of Decision Tree and Bayesian Network algorithms in predicting academic success at Can Tho University and the Asian Institute of Technology. Despite variations in the student demographics at the two institutions, both algorithms demonstrated comparable levels of accuracy. Specifically, the Decision Tree outperformed the Bayesian Network by 3-12% in terms of accuracy. The research contributes valuable insights into the prediction of student performance, offers a comparative analysis of data mining algorithms, and highlights the maturity of open source tools in this field.

The primary objective of Alexandro et al. (2010) is to enhance the effectiveness of higher education systems by comprehending the various factors that influence student loyalty. To achieve this, they employ data mining methods, specifically decision tree classification and VB.net programming. By utilizing these techniques, they are able to generate rules and make predictions regarding student loyalty, thereby enabling more efficient resource management and support within the higher education domain.

Pal and Pal (2013) suggest the utilization of data mining methods for assessing teacher effectiveness through a range of criteria. The objective is to forecast the quality, efficiency, and future potential of educators in different fields, empowering decision-makers at higher levels with valuable insights. Key factors taken into account encompass student evaluations, feedback from the institution, and support provided by the organization. Through data mining techniques, hidden patterns in teacher performance can be unearthed across diverse metrics.

Hemaid and El-Halees (2015) examined the various factors that influence the assessment of teacher performance. In their research, they put forth a model that utilizes data mining techniques such as association, rule induction, K-NN, and Naïve Bayesian (Kernel) to enhance the performance of teachers. The dataset used in this study comprises academic qualifications, experience, and courses obtained from the Ministry of Education and Higher Education in Gaza City. The primary objective of the proposed model is to identify strategies that can improve teachers' performance and provide insights into their classroom effectiveness.

Gangarde and Geremew (2017) employed data mining techniques in the realm of higher education to conduct an in-depth analysis, prediction, and evaluation of both student and teacher performance. Their study introduced a novel system that utilizes a classification model to forecast teacher performance, while also identifying the key factors that influence it. They used decision tree algorithms such as J48 and random tree, along with the Naive Bayes algorithm, all of which were implemented using the WEKA tool.

Anuradha and Velmurugan (2015) scrutinized the final year results of undergraduate students in three private colleges situated in Tamil Nadu, India. They utilized data mining techniques including the decision tree algorithm C4.5 (J48), Bayesian classifiers, k Nearest Neighbor algorithm, and rule learner's algorithms OneR and JRip to predict the academic performance of the students.

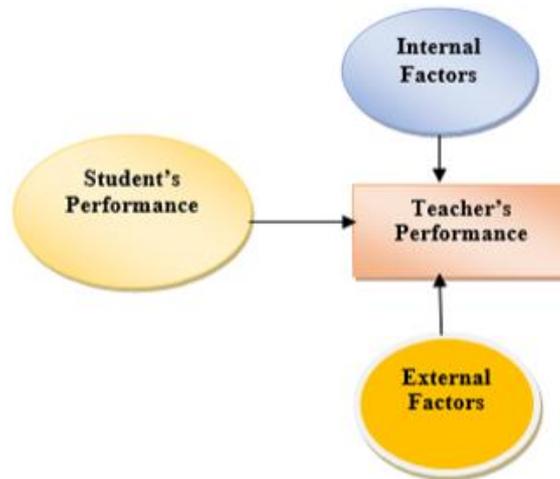


Figure 2.2: Affecting factors of teacher's performance

The significance of competent, adequately supported, and driven educators in enhancing the effectiveness and fairness of education is highlighted in the paper. It proposes that enhancing teaching performance is a strategic approach that can result in substantial advancements in student learning. According to Figure 2.2, teachers' performance is assessed by considering various factors, including student feedback, engagement in professional development activities such as research endeavors, participation in national and international conferences, organization of workshops, publication of research papers, and evaluations. The study indicates that educational data mining has the potential to enhance the education sector by analyzing evaluations from both students and teachers (Glorinthal & Shanmugarajeshwari, 2019).

Bhardwaj and Bhusry (2014) examined two different methodologies in educational data mining to assess the performance of faculty, using data mining techniques: a multiple classifier approach and a single classifier approach. The multiple classifier approach employs K-nearest neighbor (KNN) for classification and rule-based classification for the initial step, whereas the single classifier approach utilizes KNN for both steps. The outcomes are then compared to identify the most effective approach for decision-making in the field of educational data mining.

Badur and Mardikyan (2011) analyzed the various factors influencing the teaching effectiveness of instructors through the application of stepwise regression analysis and decision trees. They gathered data from student evaluations within the

Management Information Systems department at Bogazici University. The findings revealed that questions related to instructors, employment status, course workload, student attendance, and student filling percentage were all found to have a substantial impact on instructor performance.

Chi et al. (2010) introduced a decision tree model designed for evaluating teaching in adult higher education with the goal of enhancing teaching standards and faculty oversight. The model critically assesses the limitations of the ID3 algorithm, outlines its methodology, and provides a practical example through a case study to illustrate its applicability. The findings indicate that the model is successful in fostering informed decision-making and enhancing teaching standards.

The assessment of teaching effectiveness is a multifaceted area that encounters challenges in dealing with incomplete data. The proposed method, the MVCM by Fu and Ping (2009) addressed this issue by employing fuzzy set theory to maximize the likelihood of missing data observations. Empirical investigations were carried out using academic performance data of students to showcase the effectiveness of the proposed approach.

Kavitha (2021) intended to construct a reliable classifier for datasets by utilizing the K-means algorithm, a cost-efficient approach for analyzing vast databases on cloud computing platforms. The application of K-nearest neighbor classification to multi-dimensional data was expected to improve data organization in cloud-based educational environments. The suggested method offered secure information on cloud computing pertaining to teacher recruitment, safeguarding user input queries' privacy, and concealing access patterns.

In a study conducted by Madhumathi et al (2017), the primary focus was on forecasting instructor effectiveness and enhancing the overall educational delivery. Various classification methods were employed for this purpose, including the Decision tree algorithm, Support Vector Machine, Artificial Neural Networks, and Discriminant Analysis. Among these, the Naive Bayesian classifier emerged as the most effective in terms of accuracy, precision, recall, and specificity. Moreover, factors such as feedback from alumni, faculty, and students were found to significantly impact instructor performance.

Dudhe and Sakhare (2018) introduced a system designed to allocate teaching responsibilities and subject ranking to instructors utilizing a range of criteria including time spent teaching, professional background, previous achievements, academic interests, and feedback from students. This comprehensive system facilitates the effective management of teaching duties, providing advantages at both the departmental and institutional levels by taking into account faculty capabilities and equitable workload distribution.

Kamath and Warke (2016) suggested a Data Mining Approach to assess the Performance Based Appraisal System of a group of teachers in Kolhapur City. This methodology facilitated the extraction of valuable insights from large datasets, which could be utilized for curriculum development, teacher placement, and recognition.

Garg and Sharma (2013) conducted research on the application of appropriate data mining algorithms to educational datasets. The primary objective of the study is to perform a comparative analysis of different data mining techniques and algorithms.

Maghari and Zoroub (2017) employed data mining techniques to develop a model for predicting the performance of prospective teachers in high schools within the Gaza Strip. The researchers utilized Decision Tree, Naive Bayes, and KNN algorithms to accomplish this task. Notably, the Decision Tree algorithm exhibited an impressive accuracy rate of nearly 90%.

Tripathi (2015) presented a model for enhancing the educational process by implementing Data Mining methodologies to reveal underlying patterns and trends, and to generate precise predictions through advanced analytical techniques during student counseling sessions.

In their study, Adekitan and Salau (2019) utilized predictive analysis to predict the fifth year and final Cumulative Grade Point Average (CGPA) of engineering students enrolled in a Nigerian University with Konstanz Information Miner (KNIME) based data mining model. They examined six data mining algorithms, achieving a peak accuracy level of 89.15%. The outcomes were authenticated through linear and pure quadratic regression models, facilitating early identification of students in need of academic assistance or at risk of graduation issues.

Rojanavas (2019) investigated the application of two data mining methods in educational data analysis: association rule for admission strategy and decision tree for job forecasting. The findings offer significant implications for enhancing admission planning and student support services.

2.3. Clustering Technique in Educational Data Mining

Joshi and Panchal (2014) conducted an analysis aimed at facilitating the recruitment process for teaching staff. The analysis involved the utilization of the WEKA tool to generate results based on various parameters of the candidates. The researchers employed different data mining techniques to extract the necessary information and uncover hidden relationships within the extensive database. The paper delves into the application, analysis, and explanation of data mining tasks, such as clustering, association, and classification, to determine the suitability of candidates for the positions of assistant professor or associate professor. The parameters considered in this evaluation include the candidate's degree, qualifications, experience, and API score.

Kumari et al. (2014) examined the role of Data mining in an Educational Field. The data collected in this study was utilized to classify and predict both students' behavior and performance, as well as teachers' performance. The use of SDAR enabled the determination of possible grade values. To identify the optimal cluster center for attributes such as attendance, session notes, and assignment notes, the K-means clustering algorithm was employed. Furthermore, the Rule-Based Classification (RBC) method was utilized to extract a set of rules that demonstrate the relationships between the attributes of the dataset and the class label. Additionally, the study delved into the evaluation of academic performance using data mining techniques at the University and College level.

Dol and Jawandhiya (2023) examined 142 research articles published between 2010 and 2020, with a specific focus on the application of classification techniques in Educational Data Mining (EDM). The objective was to classify, analyze, and predict various aspects of students' academic performance, such as dropout rates, and instructors' performance, in order to enhance the teaching and learning process. The study highlighted the utilization of classification techniques including Naïve Bayes,

Random Forest, Support Vector Machine, and J48, while identifying K-means clustering algorithm as the most suitable choice. Furthermore, the article addresses existing research gaps and proposes future directions for EDM to further enhance teaching and learning processes.

Miah and Sabharwal (2024) conducted a study that centered on creating a solution by utilizing machine learning algorithms to evaluate the effectiveness of teachers in the realm of education. The research methodology involved the use of a case dataset obtained from the UCI machine learning repository. The proposed approach customizes learning opportunities and oversees the learning process. The K-means clustering technique was identified as the key component for predicting teachers' effectiveness, demonstrating a 99% accuracy rate with the random forest classifier.

Institutions of higher education place great importance on analyzing and assessing student feedback in order to enhance the educational environment and improve the overall learning experience. Abaidullah et al. (2015) conducted a study in which they employed the K-means clustering algorithm to examine student feedback data. This approach allowed educational stakeholders to make well-informed decisions regarding the effectiveness of educational programs and to enhance the quality of teaching and learning for students.

Erdoğan and Timor (2005) focused on examining the utilization of data mining within an educational setting. They conducted a study using cluster analysis and K-means algorithm techniques with MATLAB to analyze the correlation between students' university entrance exam scores and their academic success.

The academic performance of first-semester undergraduate engineering students at the Modern Academy for Engineering (MAE) was the subject of investigation in a study conducted by Osman and Zeweidy (2013). The researchers compared different data mining techniques to predict this performance, using high school grades as input. The findings of the study reveal that decision trees, clustering, and naive bayes exhibit higher levels of accuracy, while association rules demonstrate the least accuracy. These results contribute to the understanding of student performance prediction and facilitate a comparative analysis of the accuracy of different data mining algorithms.

By employing clustering and association rule mining techniques, Gopal et al. (2011) investigated the performance of faculty members in management disciplines based on student feedback. The effectiveness of faculty was influenced by several parameters, including qualifications, experience, commitment, research activities, institutional support, financial feasibility, and top management support. The primary focus of this paper was to highlight the importance of data mining techniques in clustering faculty performances and identifying meaningful associations that can aid in decision-making processes. The analysis was confined to students enrolled in the Computer Applications Department within accredited institutes.

Križanić (2020) provided an overview of how data mining techniques were utilized to analyze educational data from a higher education institution in Croatia. The data used for this analysis were event logs obtained from an e-learning platform used for a real e-course. The research employed cluster analysis and decision tree techniques to gain insights from the data. Cluster analysis was employed to group patterns of student behavior based on their similarity in using course materials. On the other hand, decision tree analysis was used to generate a representation of decision-making, enabling the identification of different classes of objects for a more comprehensive analysis of student learning patterns.

The main purpose of Anandavally et al. (2015) were illustrating how clustering techniques can be employed to categorize students into clearly defined cluster. Classification enabled educators to make predictions about students' final results based on their academic performance in the early stages.

The research presents a comprehensive review of various data mining techniques, including algorithms such as classification and clustering. It evaluates the efficacy of these techniques using an educational dataset specifically designed for institutions of higher learning (Sharma & Sharma, 2018).

Sharma and Sharma (2018) The research presented a comprehensive review of various data mining techniques, including algorithms such as classification and clustering. The study evaluated the efficacy of these techniques using an educational dataset specifically designed for institutions of higher learning.

3. METHODOLOGY

The methodology is designed for applying data mining for the performance evaluation of academicians in order to provide strategic insights for the university administration. In this section, the methodology of the study is presented. The methodology starts by introducing the collection of the dataset, then discusses preprocessing and explains the K-means algorithm for clustering as provided in Figure 3.1.

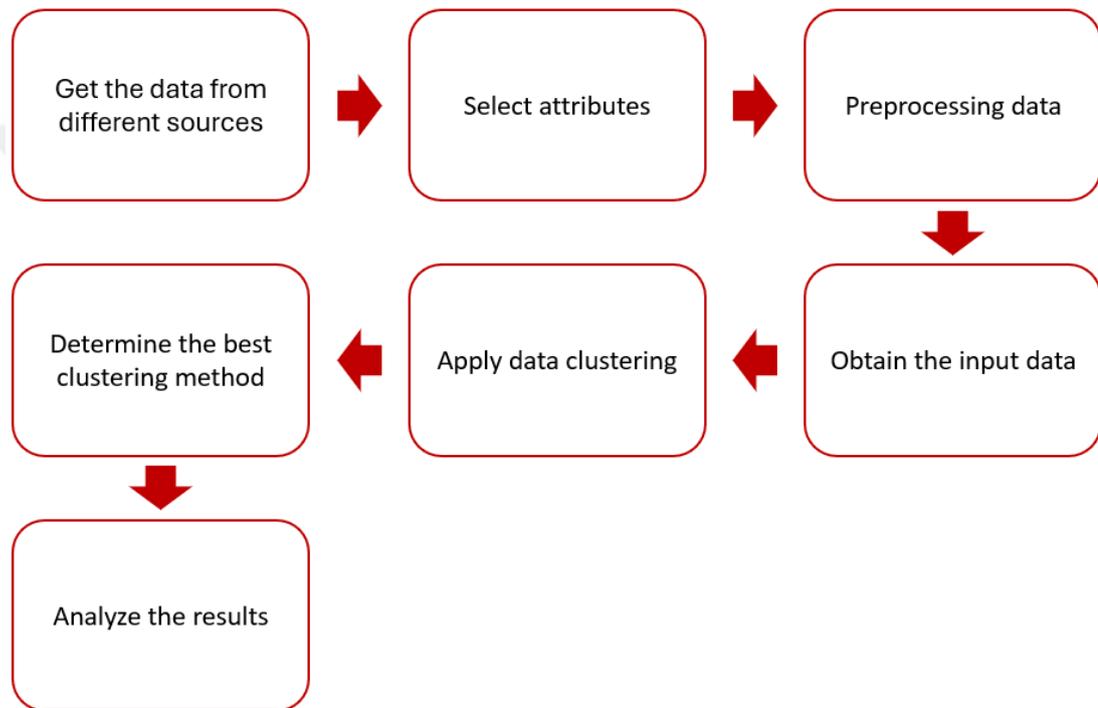


Figure 3.1: Flowchart of the methodology

3.1. Data Collection

The dataset of this study is collected from YÖK profiles of the academic staff and university records on the academic and administrative responsibilities and merged into a single Excel file. The dataset is grouped under six attributes, which are book score, article score, course score, supervised theses score and administrative responsibilities score, consisting of multiple subfactors that is determined to be used in the study. Each attribute's original value is transformed into a score through

specified calculations, and the average yearly score is determined based on the academic performance criteria set by the university, as explained in the Preprocessing section. Consequently, each academician receives a score for each of the defined attributes.

3.2. Data Statistics and Preprocessing

At this stage, data statistics are applied to the data set for defining attribute distributions and exhibiting the statistical parameters of the data via WEKA and MINITAB softwares. Furthermore, the large-scale data consisting of multiple subfactors undergo specific calculation processes in order to be reduced to the attribute level. The results are examined, and the apparent outliers are removed from the dataset. The dataset distribution, mean, and median values are examined using the WEKA software and MINITAB. The screenshot of the preprocessing stage utilized by WEKA is displayed in Figure 3.2.

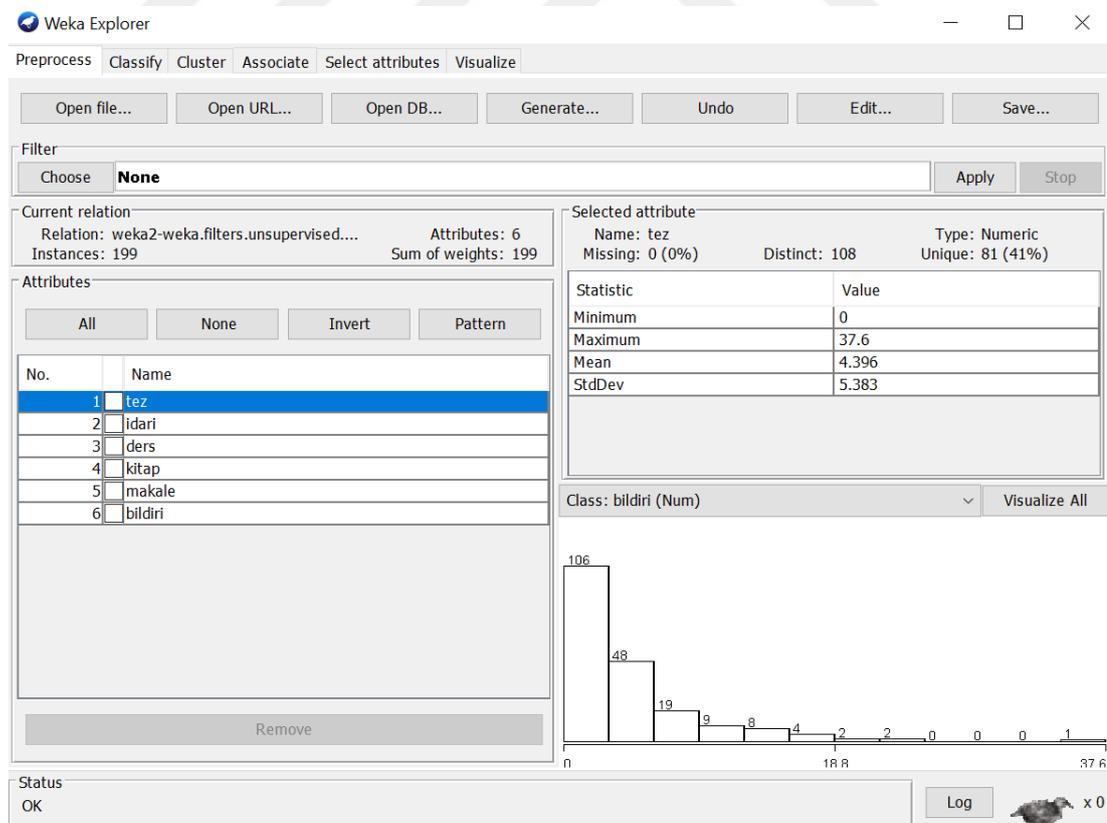


Figure 3.2: Screenshot of preprocessing in WEKA

3.3. Clustering

The main method of this study is the clustering approach applied for the attributes defined for the academicians on their academic and administrative responsibilities.

The objective of clustering is to divide a collection of data points into separate clusters, where each cluster comprises data points that exhibit greater similarity than those in different clusters. Clustering is utilized with the K-means algorithm where clustering entails the aggregation of akin data points according to the defined attributes or traits in this algorithm. In this regard, the K-means algorithm as the clustering algorithm, the K-means++ method as the initialization method, and the Euclidean distance as the distance method implemented with WEKA software. The following sections provide the details to guide the implementation of these methods.

3.3.1. The Elbow Method

The selection of the correct cluster number is crucial for obtaining accurate results and conducting analysis in clustering studies. For this purpose, the Elbow Method is applied to find the optimal cluster number in this stage.

The Elbow Method is the most commonly used technique to determine the optimal number of clusters in the K-means algorithm. The Elbow Method operates by computing the Within-Cluster Sum of Squares (WCSS), the first equation represents the aggregate of squared distances between data points and their respective cluster centers, where x_i is the data points and c_i is the closest point to the centroid as shown in Equation 3.1. However, there comes a juncture where augmenting K ceases to result in a substantial reduction in WCSS, and the pace of decrease decelerates. This juncture is frequently denoted as the elbow.

$$WSS = \sum_{i=1}^m (x_i - c_i)^2 \quad (3.1)$$

Where,

x_i = data points

c_i = the closest point to the centroid

WSS is measured for each k value, and the WSS values are shown in Figure 3.3 where WSS on the y-axis and the number of clusters on the x-axis.

In this example, as the k value increases from 2, there is a gradual change in the WSS value. Therefore, the elbow point value is determined as the optimal k , which is 2. Beyond $k=2$, increasing the number of clusters does not result in a dramatic change in the WSS value.

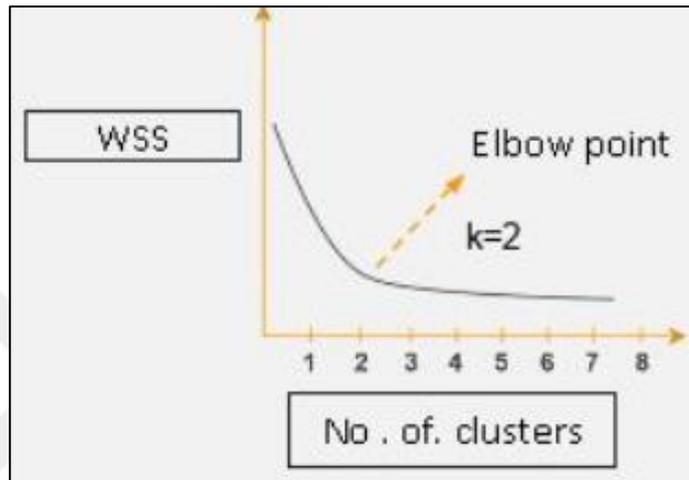


Figure 3.3: Elbow point

3.3.2. K-Means Algorithm

K-means is an unsupervised clustering method that divides data points into k clusters, where each data point is assigned to the cluster with the closest centroid. As indicated Equation 3.2, K-means algorithm partitions data into k clusters by segregating samples, aiming to minimize a measure called inertia or the sum of squares of distances within clusters.

$$arg_s \min = \sum_{i=1}^k \sum_{x \in S_i} \|x - \mu_i\|^2 \quad (3.2)$$

Where,

S = sets of observations

k = number of sets of observations

x = observation data point

μ_i = mean of points in S_i

This model is solved by selecting the K-means algorithm in the WEKA software, which is shown in Figure 3.4.

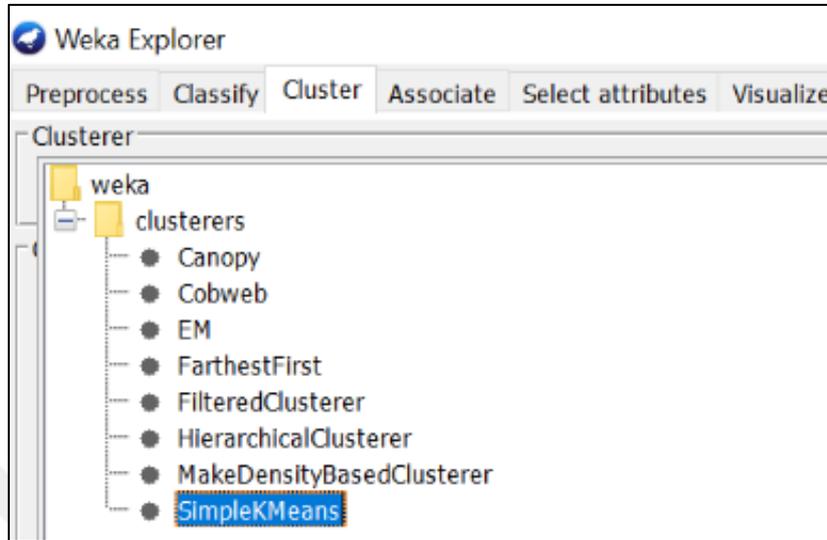


Figure 3.4: WEKA K-Means Algorithm

3.3.3. Euclidean Distance

The Euclidean Distance, which quantifies the separation between two points within Euclidean space, is established by computing the length of the line segment connecting the said points. In Equation 3.3, the calculation of Euclidean distance is demonstrated. It is assumed that when there are two points in a 2-dimensional space, d is Euclidean distance, (x_1, y_1) is the first data point's coordinate, and (x_2, y_2) is the second data point's coordinate. The computation illustrated in Equation 3.3 is visualized for 2 points in Figure 3.5.

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \quad (3.3)$$

Where,

d = Euclidean distance

(x_1, y_1) = the first point's coordination

(x_2, y_2) = the second point's coordination

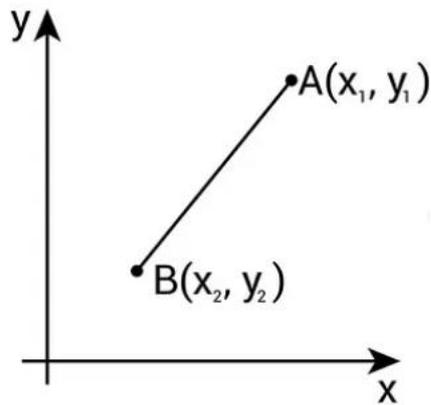


Figure 3.5: Euclidean distance

3.3.4. K-Means++ Initialization Method

K-means++ initialization method is a technique used to determine initial cluster centers. This method aims to obtain better results when choosing the starting points of the k-means algorithm. The K-means++ method aims to create a more balanced beginning by taking into account the distances between data points, instead of choosing the starting centers randomly. In this way, the algorithm is enabled to cluster more consistently.

The algorithm is described in 5 steps:

- 1- Select one center randomly from the data points with equal probability.
- 2- Calculate $D(x)$ for each data point x that has not been selected yet, which represents the distance from x to the closest center already chosen.
- 3- Select a new data point randomly to serve as a new center, utilizing a weighted probability distribution where a point x is chosen with a probability directly proportional to $D(x)^2$.
- 4- Iterate Steps 2 and 3 until k centers have been selected.
- 5- Once the initial centers have been chosen, continue with the standard k -means clustering approach.

4. IMPLEMENTATION AND RESULTS

This study is implemented in a university in Istanbul (Turkey) that provides education for 27 years, with a total of 705 academic staff, including 62 full-time Professors, 32 Associate Professors, 140 Assistant Professors, 296 part-time faculty members, 81 lecturers, and 94 research assistants in 8 faculties, two vocational schools for undergraduate degrees together with a Graduate Education Institute for the master's and doctorate degrees.

4.1. The Dataset

The dataset is prepared using sources from the Higher Education Council (YÖK) and the university records on personal information together with academic and administrative responsibility. The YÖK profiles of 207 academic staff with different research domains at the university are examined, and the data of 8 academic staff with outdated profiles in the YÖK profiles is removed from the dataset for data cleaning. Hence, this study focused on 199 academic staff excluding the lecturers in vocational schools, research assistants, and part-time faculty members. The statistics of the dataset is provided in the following tables and figures.

Figure 4.1 displays the gender information of the dataset. Academics' gender information is obtained by examining YÖK Academic profiles alongside university records. In the study, it is not used as an attribute; rather, it is utilized for analysis during the interpretation of results. Figure 4.1 shows that out of the 199 individuals (instances) in the dataset, 127 are female, and 72 are male.

The academic personnel contributing to the dataset come from 8 faculties namely Faculty of Education, Faculty of Arts and Sciences, Faculty of Law, Faculty of Economics and Administrative Sciences, Faculty of Architecture, Faculty of Engineering, Faculty of Art and Design, and Faculty of Health Sciences. The distribution of the people and their faculties are given in Figure 4.2. The academic staff under different faculties are merged under common research domains such as 1: Education Sciences, 2: Natural Sciences and Mathematics, 3: Philology, 4: Fine Arts, 5: Law, 6: Architecture, Planning, and Design, 7: Engineering, 8: Health Sciences, 9: Social, Humanities, and Administrative Sciences. The corresponding faculty-research

domain associations are displayed in Table 4.1. In this study, research domains are not used as attributes however for interpreting and contextualizing the analysis findings.

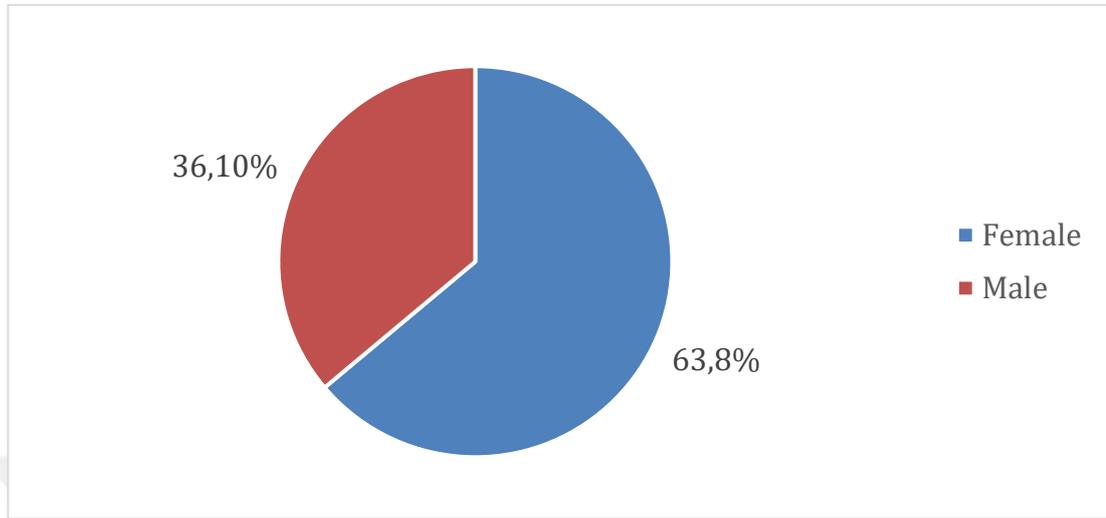


Figure 4.1: Gender distribution of academic staff

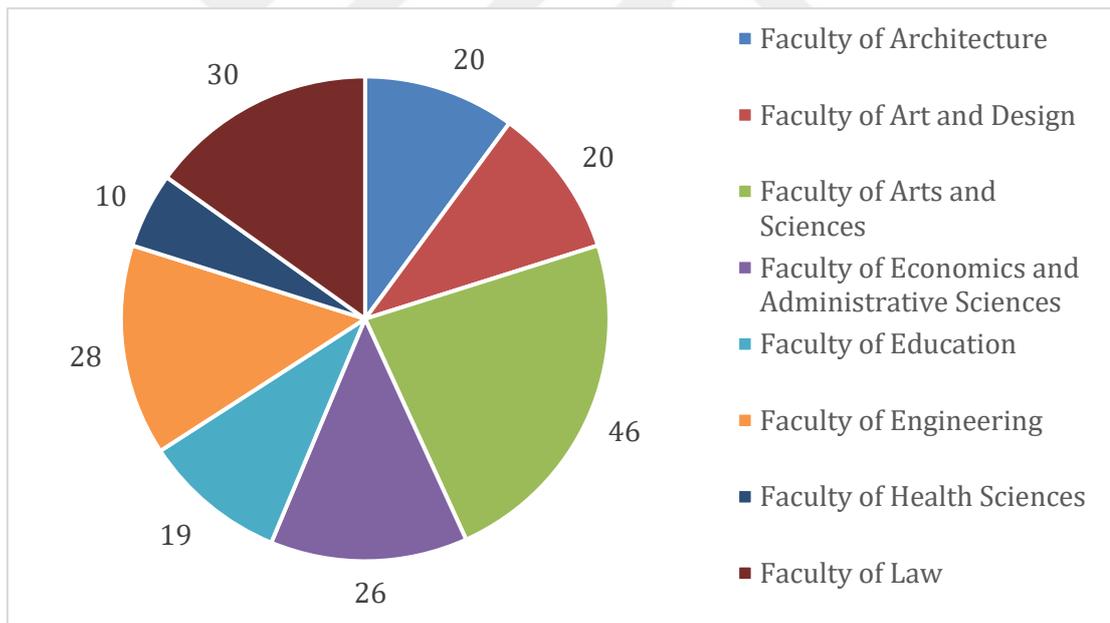


Figure 4.2: Academic staff and their faculties distribution

Table 4.1: Research domain and faculty associations

Research Domain	Faculty
Education Sciences	Education Sciences
Education Sciences	Special Education
Education Sciences	Elementary Education
Education Sciences	Foreign Language Education
Natural Sciences and Mathematics	Physics
Natural Sciences and Mathematics	Mathematics and Computer Sciences
Natural Sciences and Mathematics	Molecular Biology and Genetics
Philology	English Language and Literature
Philology	Turkish Language and Literature
Fine Arts	Animation and Cartoon
Fine Arts	Communication Arts
Fine Arts	Communication and Design
Fine Arts	Arts and Cultural Management
Fine Arts	Radio TV and Cinema
Fine Arts	New Media and Communication
Law	Law
Architecture, Planning, and Design	Architecture
Architecture, Planning, and Design	Interior Architecture and Environmental Design
Engineering	Computer Engineering
Engineering	Electrical and Electronics Engineering
Engineering	Industrial Engineering
Engineering	Civil Engineering
Health Sciences	Nutrition and Dietetics
Health Sciences	Physiotherapy and Rehabilitation

Health Sciences	Nursing
Social, Humanities, and Administrative Sciences	Psychology
Social, Humanities, and Administrative Sciences	Entrepreneurship
Social, Humanities, and Administrative Sciences	Economics
Social, Humanities, and Administrative Sciences	Business Administration
Social, Humanities, and Administrative Sciences	International Relations
Social, Humanities, and Administrative Sciences	International Trade and Finance

Initially, the dataset shown in Table 4.2 had 12664 entries for 199 academic staff (instances) and 6 defined attributes such as supervised theses, administrative responsibilities, courses, publications (i.e. books, articles, and proceedings), and 58 subfactors (i.e. scientific book, SCI, international/full text). This dataset which is consisting of 12664 data points composed of various subfactors is then transformed into a dataset of 1194 data points by merging under 6 main attributes. The merging process consists of various calculations provided under the succeeding sections to convert different units of measures under yearly average scores using the academic performance regulation of the University. The merging process and the scoring are elucidated in the following Preprocessing and Input Data sections.

Table 4.2: Initial and merged data points

Attribute	Subfactor		Number of Entry for Each Subfactor	Total Number of Data for Attributes
1 Book	1.1	Scientific Book	199	1791
	1.2	Scientific Book Chapter	199	
	1.3	Course Book	199	
	1.4	Course Book Chapter	199	
	1.5	Book Translation	199	
	1.6	Book Translation Chapter	199	
	1.7	Encyclopedia Content	199	
	1.8	Research Book (Except Thesis)	199	
	1.9	Research Book (Except Thesis) Chapter	199	
2 Article	2.1	SCI	199	4776
	2.2	SSCI	199	
	2.3	SCI-Expanded	199	
	2.4	AHCI	199	
	2.5	ESCI	199	
	2.6	TR Dizin	199	
	2.7	Field Index	199	
	2.8	EBSCO	199	
	2.9	Not Scanned in Indexes (International)	199	
	2.10	Other Index (International)	199	
	2.11	Art Index (International)	199	
	2.12	Ergonomics Abstract (International)	199	
	2.13	IM: Index Medicus (International)	199	
	2.14	DAAI (Design and Applied Arts Index) (International)	199	
	2.15	The Avery Index to Architectural Periodicals (International)	199	
	2.16	IE: Engineering Index (International)	199	
	2.17	Educational Full Text (International)	199	

	2.18	Journals Indexed in Eric (International)	199	
	2.19	Not Scanned in Indexes (National)	199	
	2.20	Other Index (National)	199	
	2.21	Art Index (National)	199	
	2.22	DAAI (Design and Applied Arts Index) (National)	199	
	2.23	The Avery Index to Architectural Periodicals (National)	199	
	2.24	Non-Refereed	199	
3 Proceeding	3.1	International / Full Text	199	1194
	3.2	International / Abstract	199	
	3.3	International / Poster & Oral	199	
	3.4	National / Full Text	199	
	3.5	National / Abstract	199	
	3.6	National / Poster & Oral	199	
4. Course	4.1.	19-20 Spring Semester	199	2587
	4.2.	19-20 Fall Semester	199	
	4.3	19-20 Summer Semester	199	
	4.4	20-21 Spring Semester	199	
	4.5	20-21 Fall Semester	199	
	4.6	20-21 Summer Semester	199	
	4.7	21-22 Spring Semester	199	
	4.8	21-22 Fall Semester	199	
	4.9	21-22 Summer Semester	199	
	4.10	22-23 Spring Semester	199	
	4.11	22-23 Fall Semester	199	
	4.12	22-23 Summer Semester	199	
	4.13	23-24 Spring Semester	199	
	4.13	23-24 Fall Semester	199	
5 Supervised Theses	5.1	Medical Specialization	199	597
	5.2	Doctorate	199	
	5.3	Master	199	
6 Administrative Responsibilities	6.1	Unit Name	573	1719
	6.2	Administrative Title	573	
	6.3	Administrative Point	573	
Total				12664

4.2. Preprocessing and Input Data

The data for the attributes are measured with different units. For instance, course hours are used for courses while the number of supervised theses is used for the supervised thesis. In the preprocessing stage, the final values of the determined attributes are achieved through a series of calculations. The academicians' YÖK profiles and the university's performance evaluation scores for each criterion are used for this stage, and due to the space limits to provide the whole dataset, the screenshot for the average yearly score based dataset with 1194 data to be used as Input data is given in Figure 4.3. The preceding chapters provide the calculation of the scores for each attribute namely supervised thesis, administrative responsibilities, courses, books, articles, and proceedings.

	A	B	C	D	E	F
1	supervisedtheses	administrative	course	book	article	proceeding
2	3,733333333	0	9,9	2	70	5,266666667
3	0,181818182	0	15,62641026	0	0	0
4	11	0	56,5997619	1,5	113,4	16,2
5	2,666666667	3	30,75333333	5	28	33
6	0	0	21,17935897	0	50	0
7	0	0	23,18611111	6	86,4	18
8	1,066666667	21	23,18923077	1	34,93333333	10,93333333
9	0	18	28,38095238	0	17,33333333	21,66666667
10	2,341463415	6	28,16363636	0	12,24390244	6,43902439
11	3,333333333	0	16,92888889	3	106,6666667	7,2
12	1,142857143	2	34,97183333	0	25,52380952	6,80952381
13	0	0	22,19166667	0	3,333333333	3,933333333
14	8,888888889	26	42,4705	23,33333333	10,11111111	45,55555556
15	8	10	37,96551282	18	25,6	45,8
16	12,8	8	34,37857143	6,4	33,4	49
17	0,75	48	24,7175	1,40625	15,25	4,21875
18	0,571428571	45	30,078	15,07142857	26,07142857	36,14285714
19	3,578947368	0	30,30294872	1,578947368	32,42105263	15,15789474
20	3,428571429	0	41,07638889	2,142857143	5,571428571	6
21	1,6	0	35,11880952	0	2,4	2,8
22	1	26	26,76878788	2,307692308	19,42307692	4,653846154
23	0	8	43,325	6	25	19,1
24	14,25	18	71,55952381	1,875	48,75	17,5
25	0,216216216	9	39,9075641	4,054054054	33,91891892	8,378378378
26	1,333333333	8	18,70333333	3,479166667	8,854166667	16,33333333
27	1,333333333	0	33,26033333	5	75	56,33333333
28	0	0	40,44952381	0	62,5	42
29	0	0	30,04	0	62	90
30	6	21	32,70533333	23,18181818	23,36363636	0
31	23,33333333	0	63,88038462	6,25	18,16666667	17,16666667
32	2,888888889	33	19,02583333	5,555555556	20,2962963	2,222222222
33	0,888888889	11	32,0425641	15	22,88888889	20,88888889
34	37,6	0	57,04527778	7,5	28	23,4
35	2,666666667	13	36,39027778	12	8,222222222	12,11111111
36	2,285714286	0	24,54285714	4,285714286	5,571428571	2,571428571

Figure 4.3: Input data screenshot

4.2.1. Book Score

The book scores of individuals are evaluated from YÖK Academic, by examining their profiles. Table 4.3 displays the points assigned by the University's Academic Performance Regulation, based on the performance grading for nine different types of books.

Table 4.3: Book types and points

Book Types	Points
Scientific Book	30
Scientific Book Chapter	15
Course Book	30
Course Book Chapter	15
Book Translation	15
Book Translation Chapter	3
Encyclopedia Content	1
Research Book (Except Thesis)	2
Research Book (Except Thesis) Chapter	1

The total score is calculated by multiplying the quantity of each pertinent book type published by an academician by its respective score and then summing all the points. Subsequently, the total points are divided by the years elapsed since individuals obtained their doctorate to find average yearly scores as the results. Figure 4.4 shows the screenshot of raw data from which academics' scores are calculated for nine book types.

An example of calculating the average yearly score of a book score of a staff for 9 subfactors is as follows:

$$2.00 = \frac{(1 * 30) + (0 * 15) + (0 * 30) + (0 * 15) + (0 * 15) + (0 * 3) + (0 * 1) + (0 * 2) + (0 * 1)}{2023 - 2008}$$

YÖK Profile Update Date	Research Domain	Gender	PhD Graduation	Book Types and Points										Book Scores
				30	15	15	30	15	15	3	1	2	1	
				Scientific Book Chapter	Scientific Book Chapter	Course Book Chapter	Course Book Chapter	Book Translation Chapter	Book Translation Chapter	Encyclopedia Content	Research Book (Except Thesis) Chapter	Research Book (Except Thesis) Chapter	Research Book (Except Thesis) Chapter	
2023	Engineering	M	2008	1	0	0	0	0	0	0	0	0	0	2,00
2021	Engineering	M	1977	0	0	0	0	0	0	0	0	0	0	0,00
2023	Engineering	M	2013	0	1	0	0	0	0	0	0	0	0	1,50
2023	Engineering	F	2020	0	1	0	0	0	0	0	0	0	0	5,00
2021	Engineering	M	2020	0	0	0	0	0	0	0	0	0	0	0,00
2024	Engineering	M	2019	0	2	0	0	0	0	0	0	0	0	6,00
2024	Engineering	F	2009	0	1	0	0	0	0	0	0	0	0	1,00
2023	Engineering	F	2020	0	0	0	0	0	0	0	0	0	0	0,00
2023	Engineering	M	1982	0	0	0	0	0	0	0	0	0	0	0,00
2021	Engineering	M	2006	0	3	0	0	0	0	0	0	0	0	3,00
2024	Engineering	M	2003	0	0	0	0	0	0	0	0	0	0	0,00
2024	Engineering	M	2009	0	0	0	0	0	0	0	0	0	0	0,00
2024	Engineering	F	2015	0	14	0	0	0	0	0	0	0	0	23,33
2024	Engineering	F	2019	0	6	0	0	0	0	0	0	0	0	18,00
2024	Engineering	M	2019	0	2	0	0	0	0	0	0	0	2	6,40
2024	Engineering	F	1992	0	3	0	0	0	0	0	0	0	0	1,41
2023	Engineering	F	2009	0	14	0	0	0	0	1	0	0	0	15,07
2024	Engineering	M	2005	0	2	0	0	0	0	0	0	0	0	1,58
2024	Engineering	M	2010	0	2	0	0	0	0	0	0	0	0	2,14
2024	Engineering	M	2009	0	0	0	0	0	0	0	0	0	0	0,00

Figure 4.4: Screenshot of data for published book types and corresponding scores

4.2.2. Article Score

The examination of the YÖK Academic profiles of academic staff reveals that published articles are grouped under twenty-four different types. The scores allocated to each article type, as per the university's academic performance regulations, are outlined in Table 4.4. The relevant scores of each academic's article are multiplied by the corresponding types of articles, and the resulting values are summed to calculate a final score. The final score is divided by the number of years since completing the doctoral degree until the last update of the YÖK Academic profile to find average yearly score of the data.

Table 4.4: Article types and points

Article Types	Points
SCI	50
SSCI	50
SCI-Expanded	50
AHCI	50
ESCI	15
TR Dizin	12
Field Index	10
EBSCO	10
Not Scanned in Indexes (International)	5
Other Index (International)	5
Art Index (International)	5
Ergonomics Abstract (International)	5
IM: Index Medicus (International)	5
DAAI (Design and Applied Arts Index) (International)	5
The Avery Index to Architectural Periodicals (International)	5

IE: Engineering Index (International)	5
Educational Full Text (International)	5
Journals Indexed in Eric (International)	5
Not Scanned in Indexes (National)	2
Other Index (National)	2
Art Index (National)	2
DAAI (Design and Applied Arts Index) (National)	2
The Avery Index to Architectural Periodicals (National)	2
Non-Refereed	2

Figure 4.5 presents the raw data utilized in computing the scores of academics across 24 subfactors (article types), each subfactor's points, and the scores of the article.

An example of calculating the average yearly score of an article score of a staff for 24 subfactors is as follows:

$$\begin{aligned}
& (0 * 50) + (5 * 50) + (0 * 50) + (0 * 50) + (0 * 50) + (4 * 50) \\
& + (1 * 10) + (0 * 10) + (0 * 15) + (1 * 15) + (0 * 10) + (0 * 10) \\
& + (0 * 12) + (0 * 12) + (2 * 2) + (1 * 5) + (1 * 2) + (0 * 5) \\
& + (0 * 2) + (0 * 5) + (0 * 5) + (0 * 5) + (0 * 2) + (0 * 5) \\
& + (0 * 2) + (0 * 5) + (0 * 2) + (0 * 5) + (1 * 2) + (0 * 5) \\
& + (0 * 5) + (0 * 2) + (0 * 5) + (0 * 5) + (0 * 50) + (0 * 50)
\end{aligned}$$

$$15.25 = \frac{\hspace{15em}}{2024 - 1992}$$

4.2.3. Proceeding Score

The proceeding information is also obtained from the YÖK Academic profiles of academic staff under six categories, and their scores are determined based on the university's performance criteria. Table 4.5 displays the types of proceedings and corresponding scores from the regulation. For each proceeding category, the quantity of proceedings held by academic staff is multiplied and aggregated, resulting in a total score. To find the average yearly score, it is divided by the number of years elapsed since the completion of the doctoral degree until the last update of the YÖK Academic profile.

Table 4.5: Proceeding types and points

Proceeding	Points
International / Full Text	15
International / Abstract	10
International / Poster & Oral	10
National / Full Text	4
National / Abstract	3
National / Poster & Oral	3

Figure 4.6 displays the raw data employed in calculating the scores of academic staff for six different proceeding types.

An example of calculating the average yearly score of a proceeding score of a staff for 6 subfactors is as follows:

$$5.27 = \frac{(0 * 3) + (1 * 4) + (0 * 3) + (0 * 10) + (5 * 15) + (0 * 10)}{2023 - 2008}$$

				Proceeding Types and Points						
				3	4	3	10	15	10	
YÖK Profile Update Date	Research Domain	Gender	PhD Graduation	National / Abstract	National / Full Text	National / Poster & Oral	International / Abstract	International / Full Text	International / Poster & Oral	Proceeding Score
2023	Engineering	E	2008	0	1	0	0	5	0	5,27
2021	Engineering	E	1977	0	0	0	0	0	0	0,00
2023	Engineering	E	2013	0	18	0	0	6	0	16,20
2023	Engineering	K	2020	0	1	0	2	5	0	33,00
2021	Engineering	E	2020	0	0	0	0	0	0	0,00
2024	Engineering	E	2019	0	0	0	0	6	0	18,00
2024	Engineering	K	2009	0	6	0	5	6	0	10,93
2023	Engineering	K	2020	0	5	0	0	3	0	21,67
2023	Engineering	E	1982	0	6	0	0	16	0	6,44
2021	Engineering	E	2006	1	0	0	9	1	0	7,20
2024	Engineering	E	2003	0	2	0	0	9	0	6,81
2024	Engineering	E	2009	2	2	0	3	1	0	3,93
2024	Engineering	K	2015	14	7	0	7	18	0	45,56
2024	Engineering	K	2019	8	0	0	4	11	0	45,80
2024	Engineering	E	2019	15	0	0	5	10	0	49,00
2024	Engineering	K	1992	15	0	0	3	4	0	4,22
2023	Engineering	K	2009	17	0	0	17	19	0	36,14
2024	Engineering	E	2005	3	1	0	2	17	0	15,16
2024	Engineering	E	2010	2	2	0	1	4	0	6,00
2024	Engineering	E	2009	1	1	0	2	1	0	2,80
2024	Engineering	E	1998	2	0	0	1	7	0	4,65
2023	Engineering	E	2013	1	2	0	0	12	0	10,10

Figure 4.6: Screenshot of data for published proceedings and corresponding scores

4.2.4. Course Score

The course information provided by academic staff over the past five years is obtained from the University's database, and scores are assigned for each semester based on the performance points determined by the university. The formulas determined by the university for performance scoring and used in the study's calculations are also provided in Table 4.6.

Since there are academic staff who have taught for less than five years within the dataset, the total score is divided by the number of semesters to find average semester based of the data. In Figure 4.7, individuals' course grades and average scores are visible for each semester.

19-20 Spring Semester	19-20 Fall Semester	19-20 Summer Semester	20-21 Spring Semester	20-21 Fall Semester	20-21 Summer Semester	21-22 Spring Semester	21-22 Fall Semester	21-22 Summer Semester	22-23 Spring Semester	22-23 Fall Semester	22-23 Summer Semester	23-24 Spring Semester	23-24 Fall Semester	Average Scores
-	-	-	-	-	-	-	-	-	-	-	-	8,06667	-	8,07
-	-	-	16,65	20,88333	-	22,95	18,15	-	13,05	5,53	-	7,35	15,75	15,04
8,04	-	-	-	-	-	-	-	-	-	-	-	-	-	8,04
19,97333	19,28	-	18,46	15,06	-	29,8	-	-	28,25333	32,6	-	-	-	23,35
-	-	-	-	-	7,8	17,3367	20,6933	-	16,88333	13,85	3,873333	-	-	13,41
-	-	-	-	-	-	-	-	-	-	-	-	10,46667	16,89	13,68
-	-	-	-	-	-	4,32667	-	-	-	-	-	7,41	4,006667	7,87
-	-	-	-	-	-	-	-	-	-	8,013333	-	-	4,753333	6,38
-	-	-	-	-	-	-	-	-	-	-	-	7,693333	-	7,69
21,55333	21,76667	14,56	20,25167	17,02667	-	40,25	40,6267	-	46,285	36,86	-	50,35333	58,13333	33,42
-	24,43667	-	-	-	-	-	-	-	-	-	-	-	-	24,44
32,19667	27,51667	-	-	-	-	-	-	-	-	-	-	-	-	29,86
-	-	-	-	-	-	-	-	-	-	-	-	8,55	-	8,55
-	-	-	-	-	-	-	-	-	8,12	11,46	-	4,54	6,29	7,60
4,326667	-	-	4,7	-	-	2,13667	-	-	-	-	-	-	-	0,49
-	-	-	-	-	-	5,65	-	-	5,93	-	-	5,69	-	5,76
-	-	-	-	-	-	-	-	-	-	-	-	14,8	-	14,80
-	-	-	54,39333	49,13667	51,5733	53,8467	61,07	36,6	58,04333	63,46667	22,96333	-	8,55	45,96
62,52667	46,3	37,55	64,57667	79,08667	36,2267	75,8867	69,7933	25,1533	52,85667	47,21333	22,36167	43,17	35,91	49,90
12,94	4,806667	-	-	-	-	-	-	-	7,77	-	-	-	-	8,51
-	-	-	-	-	-	-	-	-	7,426667	20,98	-	-	-	14,20
-	-	-	-	-	-	-	-	-	8,706667	-	-	-	-	8,71
4,14	-	-	4,166667	-	-	4,43333	6,17	-	-	-	-	6,21	4,406667	4,92

Figure 4.7: Screenshot of data for given courses and corresponding scores

Table 4.6: Course types and points

Course	Points
Ongoing Doctoral, Specialization, and Proficiency Thesis / Advisor	5
Ongoing Master's Thesis / Advisor	4
Associate/Bachelor's Course	For each course: Weekly teaching hours $\times [2 + (\text{Number of students} / 100)]$
Associate/Bachelor's Laboratory/Practical Course	For each course: Weekly teaching hours $\times [1 + (\text{Number of students} / 100)]$
Undergraduate Final Project Course	For each course: $2 \times [1 + (\text{Number of students} / 5)]$
Graduate Course	For each course: Weekly teaching hours $\times [2 + (\text{Number of students} / 20)]$
Graduate Seminar Course	For each course: $+ (\text{Number of students} / 10)$

An example of calculating the average semesterly score of a proceeding score of a staff for 14 subfactors is as follows:

$$15.04 = \frac{0 + 0 + 0 + 16.65 + 20.88 + 0 + 22.95 + 18.15 + 0 + 13.05 + 5.53 + 0 + 7.35 + 15.75}{14 - 6}$$

4.2.5. Supervised Theses Score

The number of master's, doctorate, and medical specialties theses that the academic personnel supervised until April 2024 are obtained from their profiles in YÖK Academic. In the performance criteria of the university, supervision scores are determined as 8 for a master's supervising, 10 for a doctorate supervising, and 10 for a medicine specialization thesis supervising. Lastly, the score is divided by the years since doctoral graduation of the academic to find the average yearly score. The Course Score attribute scores calculated in the preceding section includes ongoing thesis supervisions, while the scores for supervision of completed theses are considered in this section. Figure 4.8 illustrates the raw data used to compute the scores of academic staff across three distinct types of supervised theses.

				Supervised Theses Types and Points			Supervised Theses Score
				10	8	10	
YÖK Profile Update Date	Research Domain	Gender	PhD Graduation	Medical Specialization	Master	Doctorate	
2023	Engineering	E	2008	0	7	0	3,73
2021	Engineering	E	1977	0	1	0	0,18
2023	Engineering	E	2013	0	10	3	11,00
2023	Engineering	K	2020	0	1	0	2,67
2021	Engineering	E	2020	0	0	0	0,00
2024	Engineering	E	2019	0	0	0	0,00
2024	Engineering	K	2009	0	2	0	1,07
2023	Engineering	K	2020	0	0	0	0,00
2023	Engineering	E	1982	0	7	4	2,34
2021	Engineering	E	2006	0	5	1	3,33
2024	Engineering	E	2003	0	3	0	1,14
2024	Engineering	E	2009	0	0	0	0,00
2024	Engineering	K	2015	0	10	0	8,89
2024	Engineering	K	2019	0	5	0	8,00
2024	Engineering	E	2019	0	8	0	12,80
2024	Engineering	K	1992	0	3	0	0,75
2023	Engineering	K	2009	0	1	0	0,57
2024	Engineering	E	2005	0	6	2	3,58
2024	Engineering	E	2010	0	6	0	3,43
2024	Engineering	E	2009	0	3	0	1,60
2024	Engineering	E	1998	0	2	1	1,00
2023	Engineering	E	2013	0	0	0	0,00

Figure 4.8: Screenshot of data for supervised graduate thesis types and corresponding scores

An example of calculating the average semesterly score of a supervised theses score of a staff for subfactors is as follows:

$$3.73 = \frac{(0 * 10) + (7 * 8) + (0 * 10)}{2023 - 2008}$$

4.2.6. Administrative Responsibility Score

Many academic staff have administrative duties in addition to their academic responsibilities and this is also accepted as a performance criterion by the Academic Performance Regulation of the University. To assess the workload of these duties, individuals' administrative roles are matched with the scores specified for administrative titles via the university's performance criteria. A person may have more than one administrative responsibility, resulting in an administrative duty score as the sum of all scores coming from corresponding points. The academic titles of each academic are examined and matched with predetermined scores as depicted in Figure 4.9, thereby revealing the administrative responsibilities score based on the titles held by the academic staff. The points corresponding to each administrative duty are shown in Table 4.7.

An example of calculating the average yearly score of an administrative responsibility score of a staff for 3 subfactors is as follows:

$$15 = (\text{Deputy Dean: } 10) + (\text{Faculty Board Member: } 3) \\ + (\text{Assistant Director of Research and Application Center: } 2)$$

Table 4.7: Administrative types and points

Administrative Title	Point
<ul style="list-style-type: none"> • Rector 	40
<ul style="list-style-type: none"> • Vice Rector 	25
<ul style="list-style-type: none"> • Dean • Institute Director • Vocational School Director 	20
<ul style="list-style-type: none"> • Department Head (Graduate/Undergraduate Programs) • Chair of the Discipline 	15
<ul style="list-style-type: none"> • Vice Dean • Assistant Director of the Institute • Vice Head of the Department • Commission Chair 	10
<ul style="list-style-type: none"> • Deputy Department Chair 	8
<ul style="list-style-type: none"> • Assistant Director of the Vocational School 	7
<ul style="list-style-type: none"> • Vocational School Program Chair 	6
<ul style="list-style-type: none"> • Unit Head • Unit Director (Administrative) <ul style="list-style-type: none"> • Senate Member • University Board Member • Director of Research and Application Center 	5
<ul style="list-style-type: none"> • Faculty Board Member • Faculty Council Member • Institute Board Member • Commission Member • Student Club Advisor • Vocational School Board Member 	3
<ul style="list-style-type: none"> • Assistant Director of Research and Application Center <ul style="list-style-type: none"> • Erasmus Coordinator • Board Member • Quality Coordination Unit Member • Commission Chair of Quality Coordination Unit 	2

Academic Staff	Unit Name	Administrative Title	Administrative Point
1	SAĞLIK BİLİMLERİ FAKÜLTESİ	Dekan Yardımcısı	10
1	SAĞLIK BİLİMLERİ FAKÜLTESİ	Fakülte Yönetim Kurulu Üyeliği	3
1	HAREKET ANALİZİ UYG. DEĞ. Y.	Araştırma Uygulama Merkez Müdür Yardımcısı	2
1	HAREKET ANALİZİ UYG. DEĞ. Y.	Komisyon Üyesi	0
1	ÜNİVERSİTE	Komisyon Üyesi	0
2	ULUSLARARASI İLİŞKİLER BÖL.	Erasmus Koordinatörü	2
2	ULUSLARARASI İLİŞKİLER BÖL.	Bütçe Sorumlusu	0
3	İNŞAAT MÜHENDİSLİĞİ BÖLÜMÜ	Bölüm Başkanı	15
3	ÜNİVERSİTE	Senato Üyesi (Fakülte)	5
3	MÜHENDİSLİK FAKÜLTESİ	Fakülte Yönetim Kurulu Üyeliği	3
3	MÜHENDİSLİK FAKÜLTESİ	Fakülte Kurulu Üyeliği	3
3	GEOMATİK UYGULAMA VE AR.	Arş. Ve Uyg. Mrk. Yönetim Kurulu Üyesi	0
3	İNŞAAT MÜHENDİSLİĞİ ANABİL.	Anabilim Dalı Başkanı	0
3	LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ	Enstitü Kurulu Üyeliği	0
3	TEKNOLOJİ TRANSFER OFİSİ	Kurul Üyesi	0
3	ÜNİVERSİTE	Komisyon Üyesi	0
4	HAREKET ANALİZİ UYG. DEĞ. Y.	Arş. Ve Uyg. Mrk. Yönetim Kurulu Üyesi	0
4	HAREKET ANALİZİ UYG. DEĞ. Y.	Komisyon Üyesi	0
5	EĞİTİM FAKÜLTESİ	Dekan	20
5	ÜNİVERSİTE	Senato Üyesi (Dekan)	5
5	ÜNİVERSİTE	Üniversite Yönetim Kurulu Üyesi (Dekan)	5
5	LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ	Enstitü Yönetim Kurulu Üyeliği	3
5	EĞİTİM FAKÜLTESİ	Fakülte Kurulu Başkanı (Dekan)	0
5	REHBERLİK VE PSİK.DAN.UYG.	Arş. Ve Uyg. Mrk. Yönetim Kurulu Başkanı	0
5	EĞİTİM FAKÜLTESİ	Komisyon Başkanı	0
5	EĞİTİM FAKÜLTESİ	Fakülte Yönetim Kurulu Başkanı (Dekan)	0
5	EĞİTİM FAKÜLTESİ	Komisyon Başkanı	0
6	MİMARLIK FAKÜLTESİ	Dekan Yardımcısı	10
7	MOLEKÜLER BİYOLOJİ VE GEN.	Komisyon Üyesi	0
8	LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ	Enstitü Müdür Yardımcısı	10
8	LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ	Komisyon Üyesi	3
8	LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ	Enstitü Kurulu Üyeliği	0
9	GİRİŞİMCİLİK BÖLÜMÜ	Bölüm Başkanı	15

Figure 4.9: Screenshot of data for administrative responsibilities and corresponding scores

4.2.7. Data Statistics

Before the implementation of clustering, input data is analyzed subject to attribute distributions, and the statistical parameters are identified. These statistical information provides the base for interpreting the clusters.

Parameters for the distributions are provided by WEKA software while the distribution analysis is applied in MINITAB. Table 4.8 contains the minimum, maximum, mean, median, and standard deviation values for six attributes—Book Score, Article Score, Proceeding Score, Course Score, Supervised Theses Score, and Administrative Score—all of which are numerically provided.

Table 4.8: Attributes and their characteristics

Attribute No	Attribute Name	Characteristics	
		Type	Properties
1	Book Score	Numerical	Min: 0 Max: 216.9 Mean: 10.687 Median: 6 S. Deviation: 18.28
2	Article Score	Numerical	Min: 0 Max: 212.039 Mean: 23.181 Median: 12.57 S. Deviation: 29.708
3	Proceeding Score	Numerical	Min: 0 Max: 90 Mean: 15.348 Median: 9 S. Deviation: 16.859
4	Course Score	Numerical	Min: 9.9 Max: 109.674 Mean: 40.042 Median: 36.92 S. Deviation: 16.451
5	Supervised Theses Score	Numerical	Min: 0 Max: 37.6 Mean: 4.396 Median: 2.67 S. Deviation: 5.383
6	Administrative Score	Numerical	Min: 0 Max: 50 Mean: 9.075 Median: 5 S. Deviation: 11.357

During the data exploration process in MINITAB software, it has been discerned that the data exhibits an exponential distribution, as illustrated in the histogram graphs depicted in Figure 4.10. Based on the exponential distributions, use of median is decided to be used in defining the cluster segments.

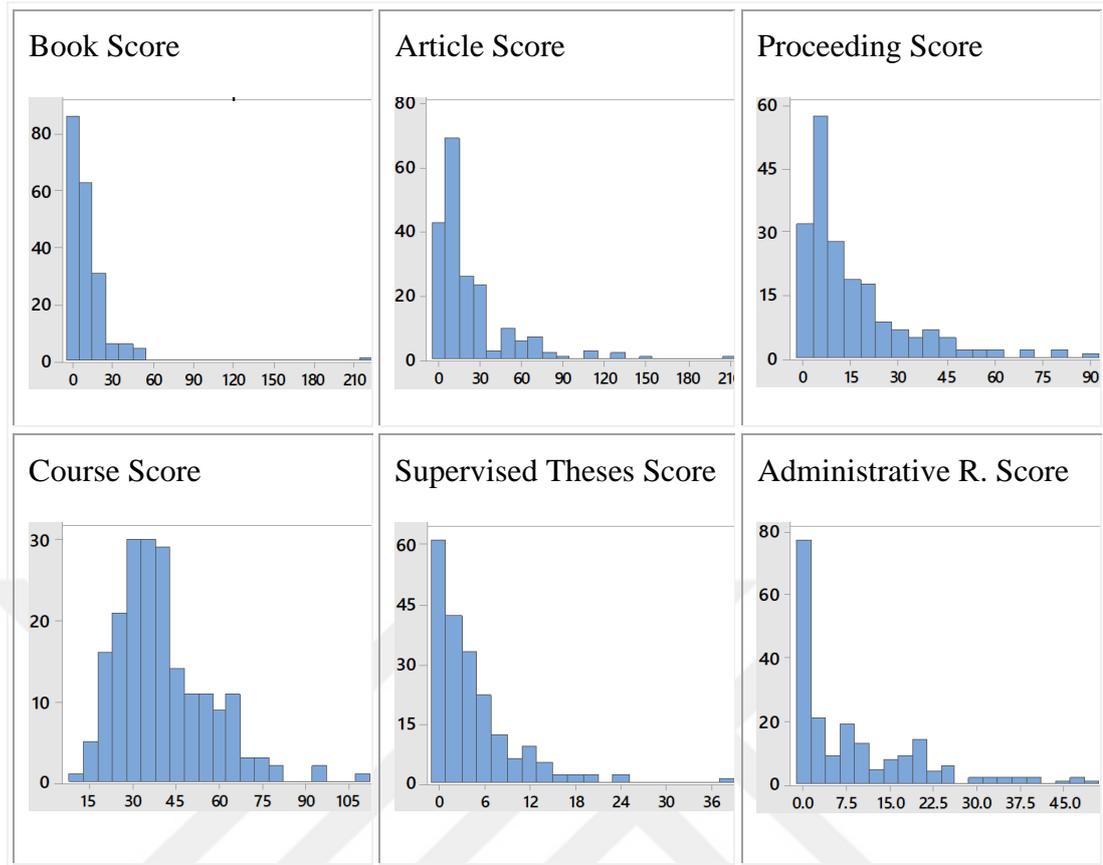


Figure 4.10: Histogram diagram of attributes

4.3.K-Means Clustering

K-means algorithm is adopted in this study which requires a defined cluster size and a selected distance measurement method. For this study, the K-means++ method is selected as the initialization method, and the Euclidean distance as the distance method. The clustering algorithm is implemented with WEKA software and the following sections provide the implementation results.

4.3.1. Identifying the Cluster Size

Initially, the dataset is tested using the K-means algorithm in the WEKA software, incorporating various cluster size k and seed values, along with different initialization methods such as Random, K-means, Canopy, and Farthest First, each applying Euclidean distance. The SSE (Sum of Squared Error) values tested in 72 different scenarios are displayed in Table 4.9.

Table 4.9: Sum of Squared Errors of different k and seed values

Number of Cluster	Seed	Sum of Squared Errors (SSE) for Different Initialization Methods			
		Random	K-means++	Canopy	Farthest First
3	10	19.72	22.71	20.64	26.09
4	10	16.11	18.02	16.18	20.92
5	10	14.99	14.74	14.58	17.70
6	10	13.60	13.62	13.16	16.47
7	10	12.26	12.32	12.58	12.53
8	10	11.48	11.75	11.94	11.15
3	15	20.28	19.80	19.73	26.09
4	15	16.12	16.13	17.64	20.92
5	15	14.42	14.47	14.81	17.70
6	15	13.65	13.69	13.03	16.47
7	15	12.79	12.30	12.56	12.53
8	15	11.80	11.68	11.90	11.15
3	20	20.30	19.80	21.90	26.09
4	20	16.12	16.12	20.31	20.92
5	20	14.75	15.03	14.57	17.70
6	20	13.64	13.75	13.48	16.47
7	20	12.17	12.58	12.09	12.53
8	20	11.67	11.56	11.31	11.15

Values for different cluster sizes and seed numbers are examined for each initialization method. As a result, clustering using the K-means++ initialization method with 10 seed values in WEKA Software. In Figure 4.11, the variation of SSE values with different numbers of clusters is depicted, highlighting a discernible breakpoint at the value of five clusters.

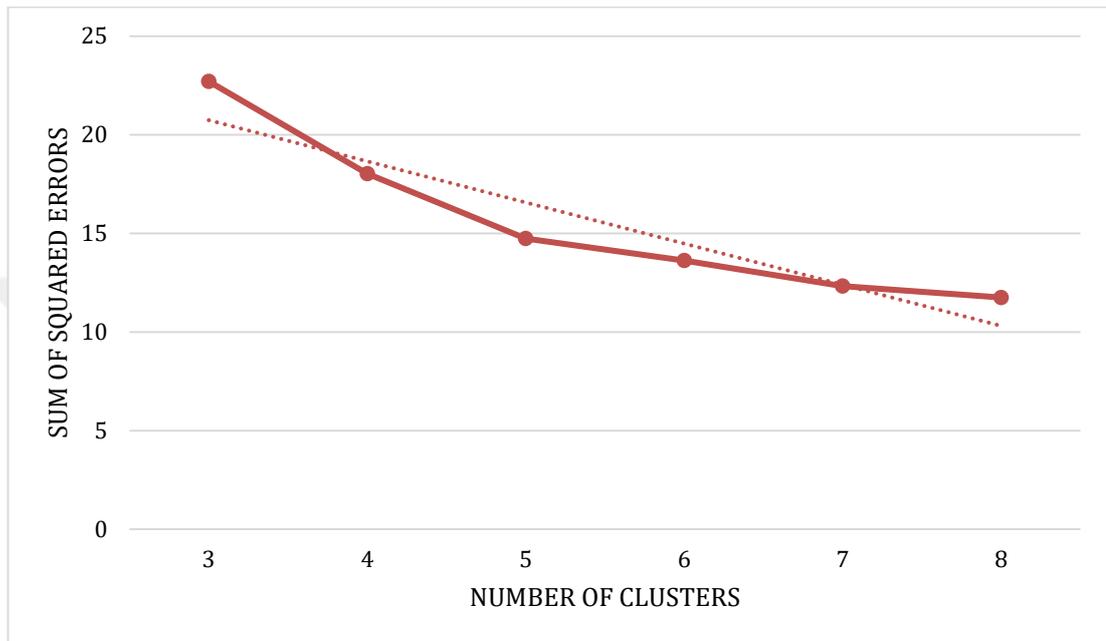


Figure 4.11: Clustering SSE Graph for K Means Initialization Method with Seed:10 Value

4.3.2. Output Data

Table 4.10 comprises the counts of items in each of the five distinct clusters, totaling 199 instances, along with the corresponding percentages representing the proportion of these items, and the respective codes assigned to scholars for privacy concerns.

The output data clustered under 5 clusters reveal that Cluster 3 includes 45 % of the whole data while Cluster 2 is the cluster with the minimum number of members. It's observed that Cluster 3 and Cluster 0 compromise 87 % of the whole data clustered.

Table 4.10: The distribution of academic staff across clusters

Cluster Number	Number of Academic Staff	Distribution	Codes
Cluster 0	44	22%	6, 7, 12, 20, 28, 30, 33, 38, 40, 41, 44, 47, 51, 58, 68, 75, 79, 81, 93, 94, 95, 113, 119, 123, 124, 133, 139, 144, 145, 149, 150, 155, 161, 164, 169, 175, 179, 181, 183, 186, 188, 189, 194, 195
Cluster 1	31	16%	2, 22, 29, 32, 42, 72, 78, 80, 85, 86, 87, 89, 114, 116, 121, 125, 126, 129, 130, 132, 134, 135, 136, 137, 142, 153, 158, 162, 167, 168, 182
Cluster 2	7	4%	15, 16, 48, 60, 84, 171, 172
Cluster 3	90	45%	0, 1, 4, 5, 8, 9, 10, 11, 17, 18, 19, 21, 23, 24, 31, 34, 35, 36, 45, 46, 50, 52, 53, 54, 55, 56, 57, 59, 61, 62, 63, 64, 65, 66, 67, 70, 76, 77, 82, 88, 90, 92, 96, 97, 98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 115, 117, 118, 122, 128, 131, 140, 141, 143, 146, 147, 151, 152, 154, 156, 157, 159, 160, 166, 174, 176, 178, 180, 184, 187, 190, 191, 192, 193, 196, 197
Cluster 4	27	14%	3, 13, 14, 25, 26, 27, 37, 39, 43, 49, 69, 71, 73, 74, 83, 91, 120, 127, 138, 148, 163, 165, 170, 173, 177, 185, 198

4.4. Interpretation and Discussion of the Results

4.4.1. Identifying Performance Segments

In this stage of the implementation, clusters are interpreted based on their distinct characteristics. For this purpose, our focus lies in determining the position of cluster members on their performance in the respective attribute. Based on the data statistics analysis that revealed exponential distributions, use of median is preferred as the parameter to define the performance segments for all 6 attributes. As the cluster size is 5, the data range for each attribute is partitioned into five segments using 20% as percentiles. Correspondingly, attribute values ranging from 0% to 20% are labeled as very low, those between 20% and 40% as low, 40% to 60% as average, 60% to 80% as high, and 80% to 100% as very high.

As depicted in Table 4.11, the distribution percentiles of book scores with corresponding book score values assigned to each percentile range is given. Based on these segments, if the cluster average for the book scores is more than 1.25 and less than 4.1053, that cluster members meant to be showing low performance for this attribute (book score). Similarly, Tables 4.12 to 4.16 display the segment ranges for the other 5 attributes.

Table 4.11: The segments for book scores

Book Score			
Percentile	Definition	Min	Max
20%	Very Low	0.0000	1.2500
40%	Low	1.2500	4.1053
60%	Average	4.1053	8.7500
80%	High	8.7500	16.8750
100%	Very High	16.8750	216.9000

In Table 4.12, it can be interpreted that if the cluster average for the article scores is more than 16.6667 and less than 33.1579, that cluster members meant to be showing high performance for this attribute (article score). Similarly, In Table 4.13, it can be interpreted that if the cluster average for the proceeding scores is less than 3.00, that cluster members meant to be showing very low performance for this attribute (article score).

Table 4.12: The segments for article scores

Article Score			
Percentile	Definition	Min	Max
20%	Very Low	0.0000	4.6000
40%	Low	4.6000	9.0526
60%	Average	9.0526	16.6667
80%	High	16.6667	33.1579
100%	Very High	33.1579	212.0392

Table 4.13: The segments for proceeding scores

Proceeding Score			
Percentile	Definition	Min	Max
20%	Very Low	0.0000	3.0000
40%	Low	3.0000	6.5000
60%	Average	6.5000	12.5714
80%	High	12.5714	25.1154
100%	Very High	25.1154	90.0000

Table 4.14 displays the segments of Course scores that has a range with a maximum value of 109.6742 and a minimum value of 9.90. Similarly, Table 4.15 provides supervised thesis scores with a range between 0.00 to 37.60. Finally, Table 4.16 presents the segments for administrative scores.

Table 4.14: The segments for course scores

Course Score			
Percentile	Definition	Min	Max
20%	Very Low	9.9000	26.7688
40%	Low	26.7688	33.3142
60%	Average	33.3142	39.5883
80%	High	39.5883	53.3233
100%	Very High	53.3233	109.6742

Table 4.15: The segments for supervised theses score scores

Supervised Theses Score			
Percentile	Definition	Min	Max
20%	Very Low	0.0000	0.0000
40%	Low	0.0000	1.6842
60%	Average	1.6842	3.7333
80%	High	3.7333	7.3600
100%	Very High	7.3600	37.6000

Table 4.16: The segments for administrative responsibility scores

Administrative Responsibility Score			
Percentile	Definition	Min	Max
20%	Very Low	0.0000	0,0000
40%	Low	0.0000	2.0000
60%	Average	2.0000	8.0000
80%	High	8.0000	18.0000
100%	Very High	18.0000	50.0000

In this final stage, the cluster results obtained from WEKA are interpreted based on the segments defined in the previous section. The attribute averages of the six attributes in each cluster is discussed based on the defined segments.

As indicated in Table 4.17, values are color-coded according to their respective intervals such as, very high is represented in dark green, high in light green, average in yellow, low in orange, and very low in red. These values are represented in Table 4.18 with their corresponding performance segments.

Final results show that only Cluster 3 exhibits a very low value for course score. Apart from this, no other cluster possesses a very low value for any attribute. Also, none of the clusters contain a low value. The most prevalent performance observed across clusters is initially the average performance, followed by the very high, and subsequently the high, very low, and low performances, respectively.

Table 4.17: Clustering results showing the attribute averages

Attributes	Clusters				
	1	2	3	4	5
Book Score	10.4705	13.1279	13.3218	6.8702	20.2806
Article Score	18.0377	23.0229	56.5337	15.8747	47.4535
Proceeding Score	10.8855	13.8209	26.8639	7.5505	47.3785
Course Score	39.2922	63.735	24.9666	34.6118	36.0664
Supervised Theses Score	4.2679	12.3783	3.3619	2.274	2.7790
Administrative Score	22.1136	5.3548	43.1429	2.600	4.8519

Table 4.18: Clustering results categorization under performance segments

Attributes	Clusters				
	1	2	3	4	5
Book Score	High	High	High	Average	Very High
Article Score	High	High	Very High	Average	Very High
Proceeding Score	Average	High	Very High	Average	Very High
Course Score	Average	Very High	Very Low	Average	Average
Supervised Theses Score	High	Very High	Average	Average	Average
Administrative Score	Very High	Average	Very High	Average	Average

4.4.2. Discussion of the Results

In this study, a dataset comprising 199 instances is clustered into 5 clusters using the clustering method in WEKA. Cluster 1, which consists of 44 individuals, represents 22% of the dataset. According to the WEKA outputs, academic staff in Cluster 1 exhibit very high administrative score. Additionally, they achieve high scores in article and supervised theses scores. Alongside these, academic staff in Cluster 1 attain average levels of proceeding score and course score, with no attributes categorized as low or very low. Despite having a very high administrative score, academic staff in this group might lean towards producing more publications due to their average course loads. The high score in supervised theses score could be a factor positively influencing publication, thus their book score and article score might also be at high levels.

Another group with a very high administrative score is Cluster 3. This cluster exhibits very high values for article score, proceeding score, and administrative score while having an average value for supervised theses score. Cluster 3 is the only cluster in the outputs with a very low value for course score, and it does not have any attributes categorized under low performance. While Cluster 1 has an average course score, Cluster 3's course score is at a very low level. With a very low course score, Cluster 3 has more article score and proceeding score compared to Cluster 1. Consequently, while a high administrative score may serve as a barrier to publication, the addition of course workload may exacerbate the negative impact on publication. Due to the lower course and supervised theses loads in this group, administrative score, proceeding score, article score, and book score might have been higher. Cluster 3, consisting of 7 individuals, constitutes 4% of the dataset and is perceived as the least populous cluster, which may imply that these academic staff with such characteristics are rare.

Cluster 2 exhibits extremely high scores in both course score and supervised theses score, while its book score, article score, and proceeding score are all categorized as high. The only attribute with an average value in Cluster 2 is administrative score, and it does not contain any attributes with low or very low values. Academic staff in this cluster have a higher course workload compared to academic staff in other clusters, and their supervised theses scores are also higher. As a result, with more theses they supervise translating into more publications, their book score,

article score, and proceeding score may be high. A notable feature of this cluster is that, overall, attribute values consist of very high and high values, while administrative score is seen as average. Among the 31 individuals in this cluster (16% of the dataset), they have balanced their administrative workload at an average level while focusing on teaching, thesis supervision, and publication.

Cluster 4, comprising 90 individuals, and it is the largest cluster covering 45% of the dataset consisting of 199 individuals. Academic staff in this cluster have an average score for each attribute and show no characteristics of very low, low, high, or very high in any aspect. While other clusters exhibit tendencies towards specific areas, no such inclination is observed in this cluster. It can be interpreted that the majority of the population performs for every performance indicator at an average performance level.

Cluster 5 consists of 27 individuals and covers 14% of the dataset. The cluster has a very high average for book score, article score, and proceeding score, while having average values for course score, supervised theses score, and administrative score. It does not have any attributes with high, low, or very low values. It is the only cluster among the clusters with a very high book score. Academic staff in this cluster show a tendency towards the publication domain while balancing their course workload, supervised theses score, and administrative score.

The results are also analyzed and discussed considering the gender and research domains of the academic staff in the data set. Figure 4.12 depicts the distribution of female academic staff across five clusters. The cluster with the highest proportion of female academic staff is Cluster 4, comprising 45% of the total, followed by Cluster 1 with 23%. The cluster with the lowest representation of female academic staff is Cluster 3, accounting for only 3%. The proportions of female academic staff in other clusters are Cluster 2 (15%) and Cluster 5 (14%). This distribution suggests that female academic staff are more concentrated in certain clusters.

The distribution of male academic staff in Figure 4.13 reveals that the highest proportion of male academic staff is found in Cluster 4 (46%). This indicates a substantial concentration of male representation within the academic domain. The next prominent cluster is Cluster 1, with male academic staff accounting for 21% of this

cluster. Lower proportions of male academic staff are observed in other Clusters c2 (17%), Cluster 5 (13%), and Cluster 3 (4%).

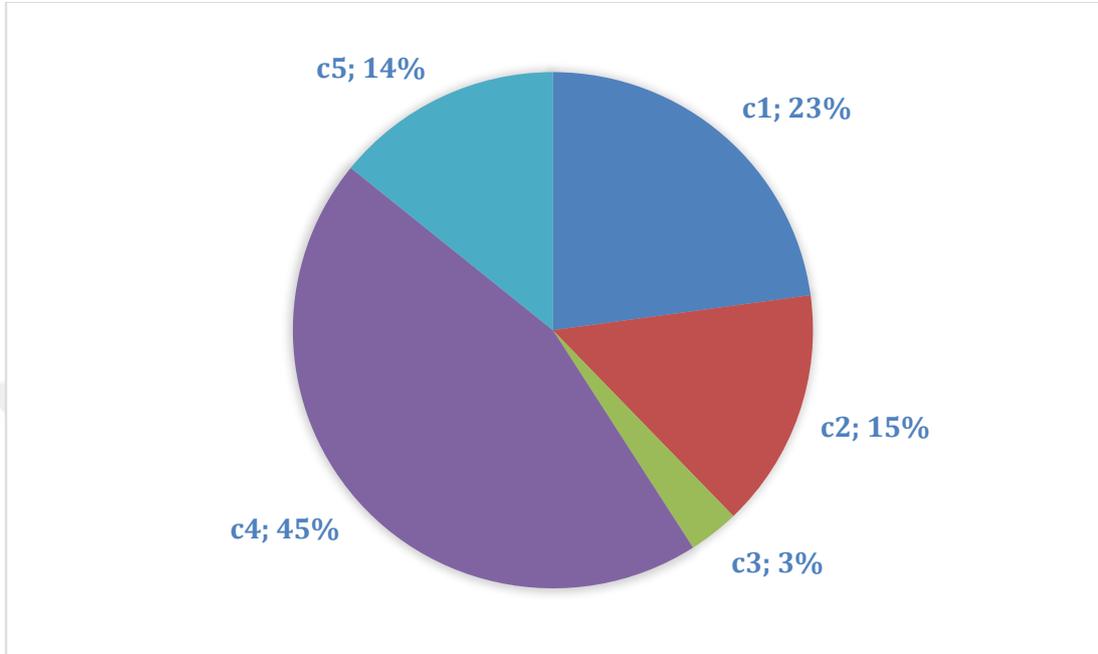


Figure 4.12: Cluster distribution of female academic staff

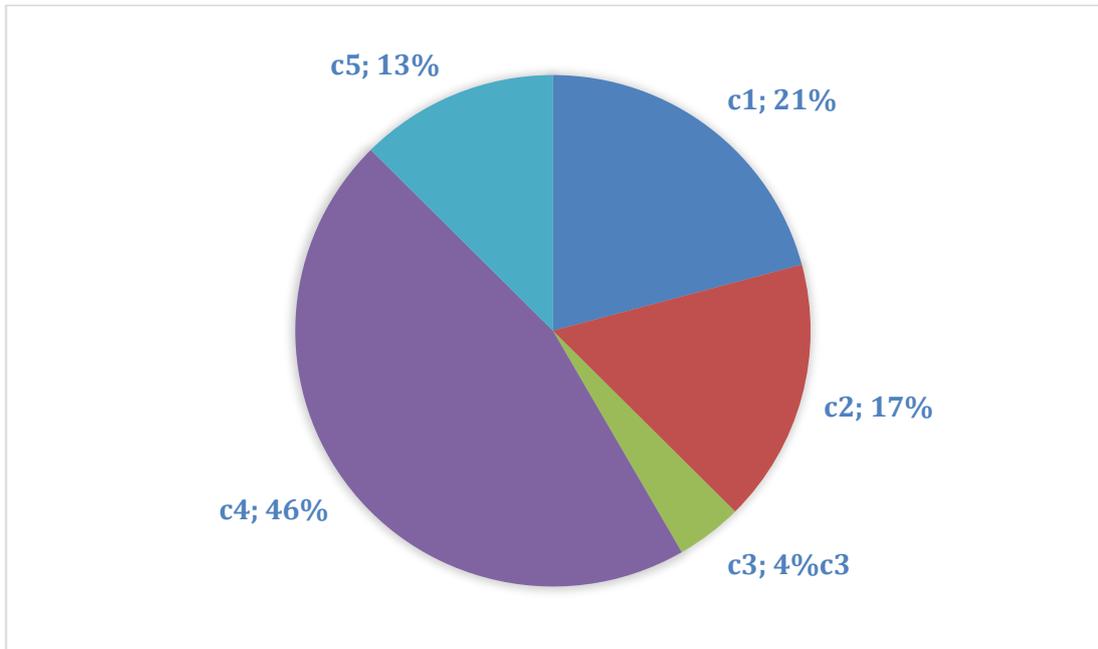


Figure 4.13: Cluster distribution of male academic staff

Both genders have the highest representation in Cluster 4. This suggests a tendency for both female and male academic staff to concentrate on a particular cluster. The second highest representation for both female and male academic staff is observed in Cluster 1. This indicates that female and male academic staff exhibit similar profiles and trends when considering the analyzed attributes and clustering results. Both genders have lower representation in clusters Cluster 2, Cluster 5, and Cluster 3.

Due to the varying numbers of academic staff in different research domains at the university, the clusters are also interpreted for the percentage distribution of research domains with the following formula and the results are displayed in Figure 4.14, Figure 4.15, Figure 4.16, Figure 4.17, and Figure 4.18.

$$\% \text{ research domain} = \frac{\text{\# of individuals belonging to the specific research domain}}{\text{total \# of individuals in that research domain}}$$

Figure 4.14 depicts the distribution of academic staff within Cluster 1 across different research domains. The graph illustrates that the highest proportion of academic staff within Cluster 1 is in Fine Arts, constituting 40% of the total. Following Fine Arts, the research domains with significant representation in Cluster 1 include Education Sciences (37%), Architecture, Planning, and Design (25%), and Social, Humanities, and Administrative Sciences (25%). Other research domains such as Law (17%), Health Sciences (20%), and Natural Sciences and Mathematics (14%) also contribute to Cluster 1 but to a lesser extent. This distribution suggests that Cluster 1 encompasses a diverse range of research domains, with a notable concentration in Fine Arts, followed by Education Sciences and Social Sciences.

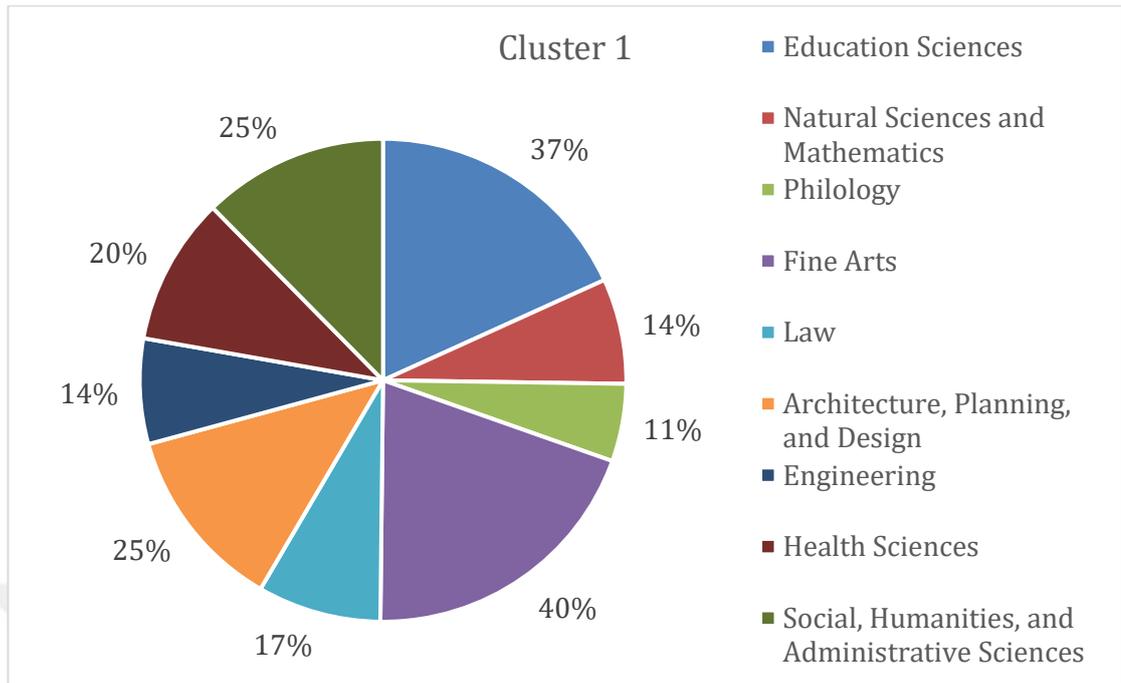


Figure 4.14: Proportion of academic staff in Cluster 1 based on research domain

Figure 4.15 illustrates the distribution of academic staff within Cluster 2 across different research domains. The graph indicates that the highest proportion of academic staff within Cluster 2 is in Social, Humanities, and Administrative Sciences, constituting 38% of the total. Following this, Philology contributes 21%, Architecture, Planning, and Design account for 25%, and Education Sciences represent 16% of the cluster. Other research domains such as Law (10%), Engineering (7%), and Natural Sciences and Mathematics (5%) also contribute to Cluster 2, albeit to a lesser extent. Notably, there is no representation of Health Sciences in this cluster. This distribution suggests that Cluster 2 encompasses diverse research domains, with a significant concentration in Social, Humanities, and Administrative Sciences.

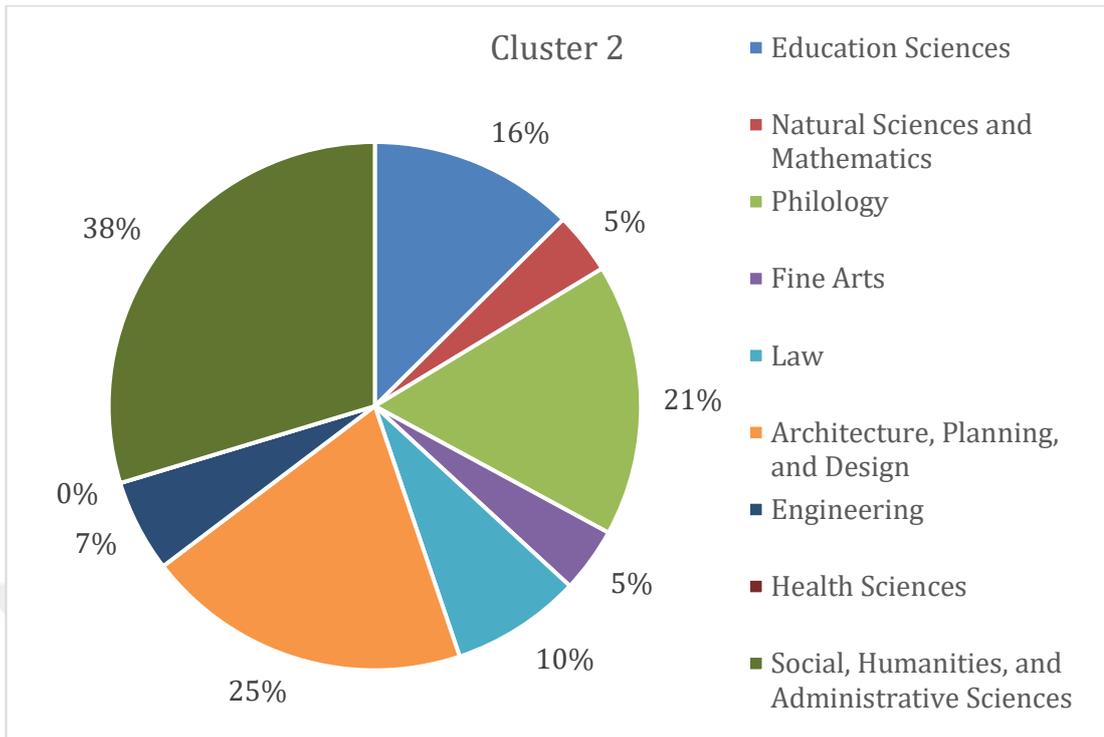


Figure 4.15: Proportion of academic staff in Cluster 2 based on research domain

Figure 4.16 depicts the distribution of academic staff within Cluster 3 across different research domains. The graph reveals that Health Sciences has the highest representation within Cluster 3, constituting 20% of the total. Following this, Natural Sciences and Mathematics contribute 10%, while Engineering accounts for 7% of the cluster. Philology also contributes to Cluster 3 with 5%. Notably, this cluster does not represent academic staff from Education Sciences, Fine Arts, Law, Architecture, Planning, and Design, or Social, Humanities, and Administrative Sciences. This distribution highlights a concentration of academic staff from Health Sciences and Natural Sciences and Mathematics within Cluster 3, suggesting a focus on these research domains within this particular cluster.

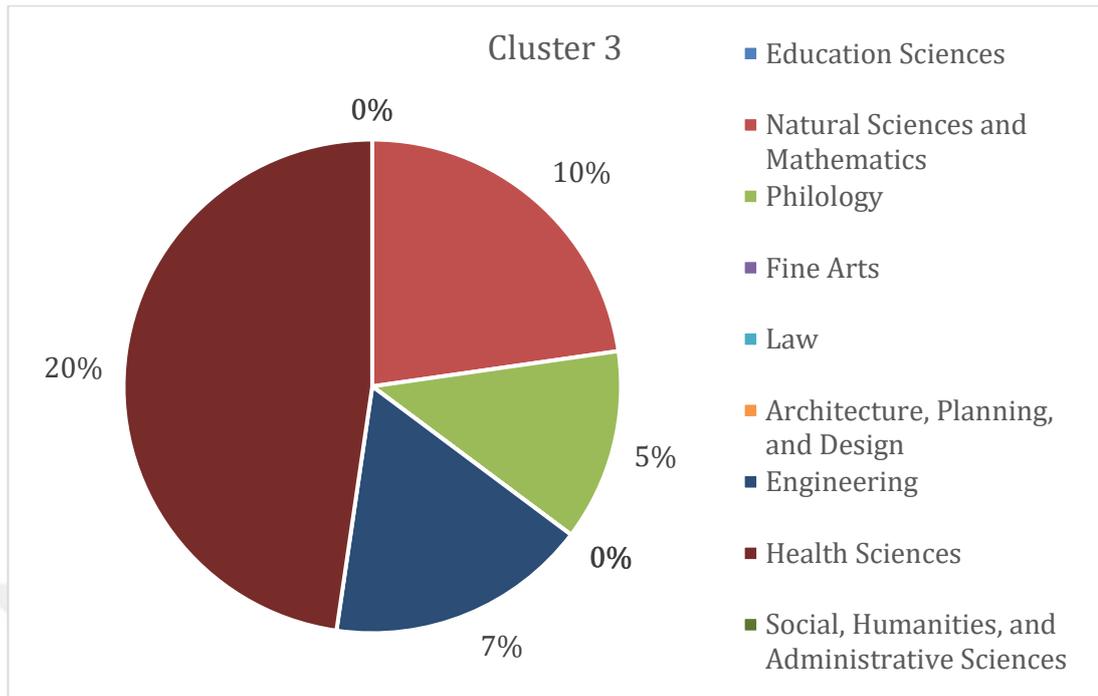


Figure 4.16: Proportion of academic staff in Cluster 3 based on research domain

Figure 4.17 illustrates the distribution of academic staff within Cluster 4 across different research domains. The graph indicates that Law has the highest representation within Cluster 4, constituting 70% of the total. Following this, Natural Sciences and Mathematics contribute 48%, while Engineering accounts for 50% of the cluster. Philology also contributes significantly to Cluster 0 with 53%. Additionally, Fine Arts (45%), Architecture, Planning, and Design (40%), and Education Sciences (32%) have notable representations within this cluster. However, Health Sciences and Social, Humanities, and Administrative Sciences have comparatively lower representations with 30% and 28%, respectively. This distribution suggests a diverse range of research domains within Cluster 4, with a particular emphasis on Law, Natural Sciences and Mathematics, Engineering, and Philology.

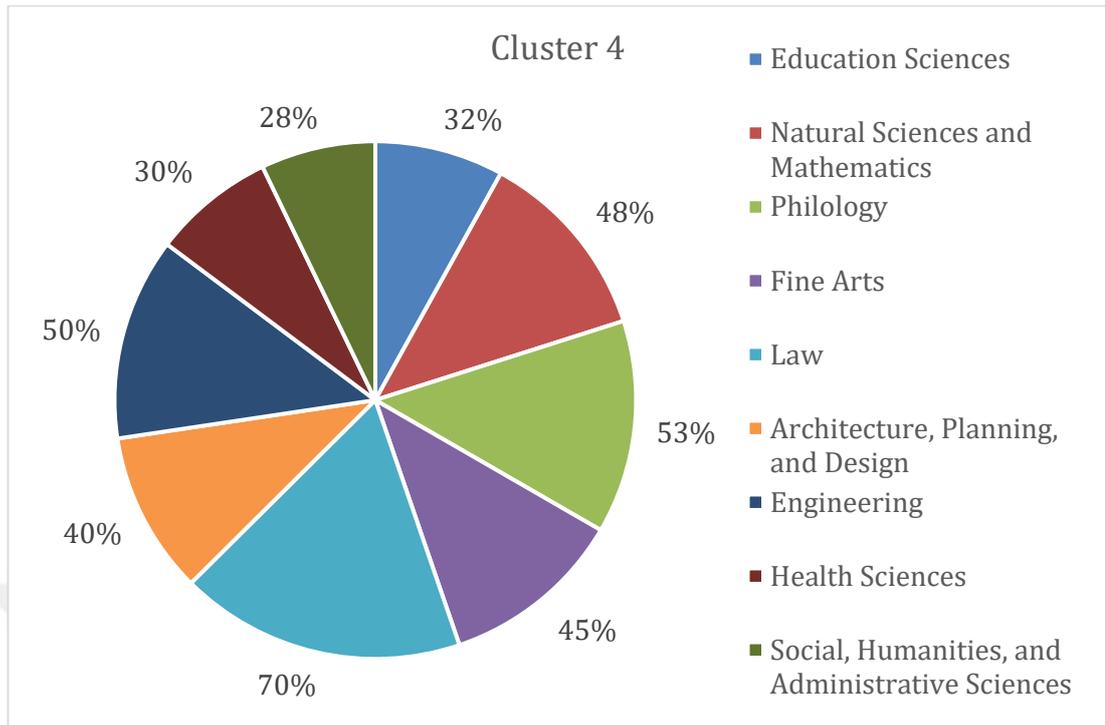


Figure 4.17: Proportion of academic staff in Cluster 4 based on research domain

Figure 4.18 illustrates the distribution of academic staff within Cluster 5 across different research domains. The graph indicates that Health Sciences has the highest representation within Cluster 5, constituting 30% of the total. Following this, Engineering contributes 21%, while Natural Sciences and Mathematics account for 24% of the cluster. Additionally, Education Sciences, Fine Arts, Architecture, Planning, and Design, and Philology have notable representations within this cluster, ranging from 10% to 16%. However, Law and Social, Humanities, and Administrative Sciences have lower representations with 3% and 9%, respectively. This distribution suggests a diverse range of research domains within Cluster 5, with a particular emphasis on Health Sciences, Engineering, and Natural Sciences and Mathematics.

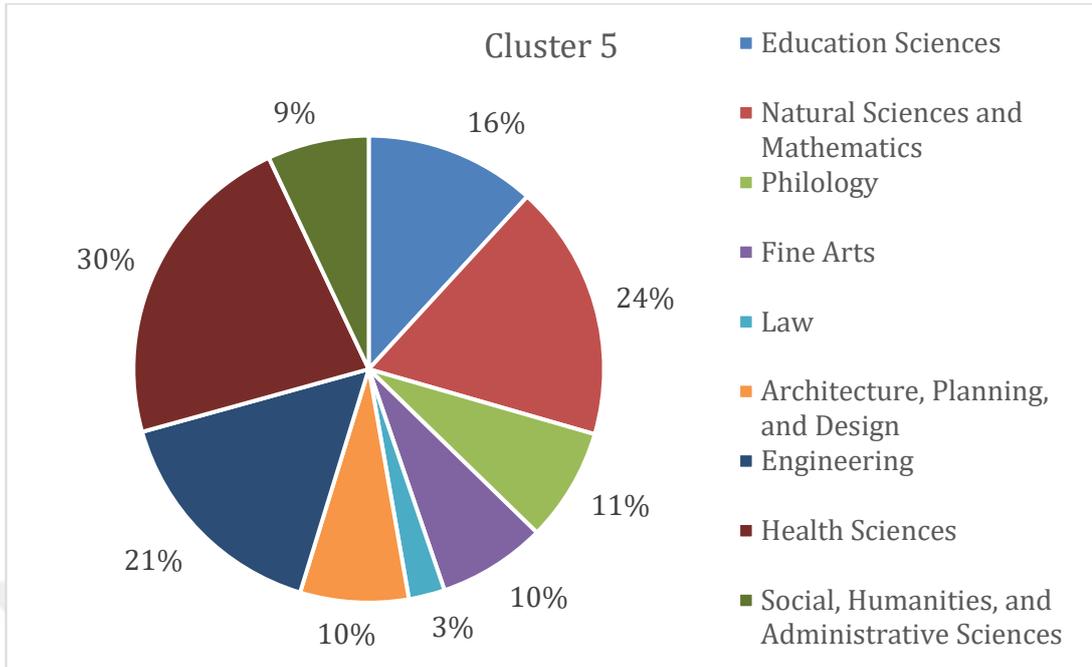


Figure 4.18: Proportion of academic staff in Cluster 5 based on research domain

5. CONCLUSION

The scope of responsibilities and performance criteria for academic staff in universities encompass various subjects and duties, each comprising numerous subfactors. This study is initiated to conduct clustering analysis using the K-means algorithm utilizing WEKA software. The focus is on academic staff employed at a university, and six attributes are considered such as, book score, article score, proceeding score, course score, supervised theses score, and administrative responsibilities score. The objective is to group academics with similar performance characteristics into the same clusters and provide a comprehensive analysis of the results. A dataset comprising information extracted from YÖK profiles and university resources is prepared for 199 academics employed at a university. Each attribute is calculated in line with the university performance evaluation criteria.

The Elbow method is employed to determine the optimal number of clusters, resulting in the identification of 5 clusters. Within the K-means algorithm framework, the Euclidean distance metric and the K-means++ initialization method are utilized. Subsequently, the findings are discussed on the common and distinct performance indicators of the clusters and further analyzed by taking into account gender and research domain effects.

Upon delving into the details of the five clusters, it becomes apparent that individuals with high administrative scores tend to have lower course loads. Among these individuals, some have sought to increase their publication output alongside their administrative duties, while others have turned their attention toward supervising theses. Individuals with a very high course load all exhibit an average administrative score, suggesting a potential negative impact of administrative responsibilities on course-related tasks. Academic staff with very high publication burdens have demonstrated average efforts across course, thesis supervision, and administrative domains. To enhance the university's publication output, efforts can be directed towards balancing the workload across course delivery, thesis supervision, and administrative duties, guiding individuals towards directing their efforts towards publication. Academic staff in Cluster 2, apart from administrative score, have yielded high outputs across the other five attributes. They can serve as examples for other

cluster members with average administrative burdens, and an examination of how these individuals have achieved their outputs can be utilized in balancing the workloads of others. The situations of individuals in Cluster 4, which encompasses 45% of the 199 participants, yielding average outputs across all areas, should be thoroughly examined. Guidance can be provided on how academic staff can improve individual and departmental performance across the six attributes in alignment with the requirements and objectives of their research domains.

In order to enrich and further explore the study in a comprehensive manner, additional attributes can be included for future investigation. These may encompass holding patents, student achievement and feedback, development and updating of course materials, mentoring and guidance of students, leadership and service in scientific communities, citation count, and participation in activities other than academic. Integrating these supplementary attributes would provide a more nuanced understanding of the multifaceted roles and contributions of academic staff within the academic environment, allowing for a more exhaustive examination of their professional endeavors and impact.

BIBLIOGRAPHY

- Abaidullah, A. M, Ahmed, N., & Ali A. (2015). Identifying Hidden Patterns in Students' Feedback through Cluster Analysis. *International Journal of Computer Theory and Engineering*, 7(1). 16-20. doi: 10.7763/IJCTE.2015.V7.923
- Adekitan, A. I., & Salau, O. (2019). The impact of engineering students' performance in the first three years on their graduation result using educational data mining. *Heliyon* 5. e01250. doi: 10.1016/j.heliyon.2019. e01250
- Agaoglu, M. (2016). Predicting instructor performance using data mining techniques in higher education. *IEEE Access*, 4, 2379–2387. doi: 10.1109/ACCESS.2016.2568756
- Agrawal, R., Singh, J. & Ghosh, S.M. (2020). Performance Appraisal of an Educational Institute Using Data Mining Techniques. *Computing in Engineering and Technology: Proceedings of ICCET 2019*, 733-745. Singapore. https://doi.org/10.1007/978-981-32-9515-5_69
- Ahmed, A. M., Rizaner, A., & Ulusoy, A. H. (2016). Using data Mining to Predict Instructor Performance. *Procedia Computer Science*, 102, 137-142. <https://doi.org/10.1016/j.procs.2016.09.380>
- Ahmedi, F. & Shiri Ahmad Abadi, M. E. (2013). Data Mining in Teacher Evaluation System using WEKA. *International Journal of Computer Applications*, 63(10), 0975 – 8887
- Alexandro, V., Rafael, Y., & Shannaq, B., (2010). Student Relationship in Higher Education Using Data Mining Techniques, *Global Journal of Computer Science and Technology*, 10(11), 54-59.
- Aljendan, A. Y. (2024). Improving Teaching Quality in Higher Education through Student Evaluation Feedback in Saudi Universities. *Educational Administration: Theory and Practice*, 30(2), 83-100. doi: 10.52152/kuey.v30i2.845

- Anandavally, R., Anuradha, C. & Velmurugan, T. (2015). Clustering Algorithms in Educational Data Mining: A Review. *International Journal of Power Control and Computation(IJPCSC)*. 7(1), 47-52
- Anuradha, C. & Velmurugan, T. (2015). A Comparative Analysis on the Evaluation of Classification Algorithms in the Prediction of Students Performance. *Indian Journal of Science and Technology*, 8(15), 1-12.
- Asanbe M.O., Osofisan A.O. & William W.F. (2016). Teachers' Performance Evaluation in Higher Educational Institution using Data Mining Technique. *International Journal of Applied Information Systems*. 10, 10-15. DOI=10.5120/ijais2016451524
- Badur, B. & Mardikyan, S. (2011). Analyzing Teaching Performance of Instructors Using Data Mining Techniques. *Informatics Educ.*, 10, 245-257.
- Bhardwaj, A. & Bhusry, M. (2014). Two approach comparisons for relative evaluation of faculty performance using data mining techniques, *2014 Conference on IT in Business, Industry and Government (CSIBIG)*, Indore, India, pp. 1-9, doi: 10.1109/CSIBIG.2014.7056945.
- Bunkar K., Bunkar R., Singh, U. K. & Pandya B. (2012). Data mining: Prediction for performance improvement of graduate students using classification, *Ninth International Conference on Wireless and Optical Communications Networks (WOCN)*, (pp. 1-5), Indore, India, doi: 10.1109/WOCN.2012.6335530.
- Chi, H., Hu, J., Deng, J. & Wu, J. (2010). A Study of Teaching Evaluation in Adult Higher Education Based on Decision Tree. *2010 Second International Conference on Information Technology and Computer Science*, Kiev, Ukraine, pp. 381-385, doi: 10.1109/ITCS.2010.99.
- Dol, S. M., & Jawandhiya, P. M. (2023). Classification Technique and its Combination with Clustering and Association Rule Mining in Educational Data Mining — A survey. *Engineering Applications of Artificial Intelligence*, 122, 106071. <https://doi.org/10.1016/j.engappai.2023.106071>
- Dudhe, A.A., & Sakhare, S.R. (2018). Teacher Ranking System to Rank of Teacher as Per Specific Domain. *ICTACT Journal on Soft Computing*, 8(02).

- Erdoğan, Ş. Z. & Timor, M. (2005). *A Data Mining Application in a Student Database. Journal of Aeronautics and Space Technology*, 2(2), 53-57.
- Fu, D. & Ping, L. (2009). The Disposal of Incomplete Classification Data in Teaching Evaluation System. *2009 Third International Symposium on Intelligent Information Technology Application*, Nanchang, China, pp. 132-135, doi: 10.1109/IITA.2009.234.
- Ganesh, C & Kesavulu Reddy, E. (2022). Overview of the Predictive Data Mining Techniques. *International Journal of Computer Sciences and Engineering*, 10(1). 30. doi: 10.26438/ijcse/v10i1.2836.
- Gangarde, R. & Geremew, D. (2017). Data Mining Techniques to Evaluate and Predict Teachers' Performance in Higher Education. *International Journal of Innovations & Advancement in Computer Science*, 6(10), 441-445.
- Garg, S.K., & Sharma, A.K. (2013). Comparative Analysis of Data Mining Techniques on Educational Dataset. *International Journal of Computer Applications*, 74(5), 1-5.
- Gillis, A. S., Stedman, C., & Hughes, A. (2024, February 13). Data Mining. TechTarget. Retrieved April 16, 2024, from <https://www.techtarget.com/searchbusinessanalytics/definition/data-mining>
- Glorinthal, P. & Shanmugarajeshwari, V. (2019). Survey: An Analysis of Student's and Teachers' Performance on Educational Data Mining. *International Journal of Research in Advent Technology*, 7(5).
- Gopal, A., Mishra, S. & Singh, C. (2011). Extraction and analysis of faculty performance of management discipline from student feedback using clustering and association rule mining techniques. *3rd International Conference on Electronics Computer Technology*, Kanyakumari, India, pp. 94-96, doi: 10.1109/ICECTECH.2011.5941864.
- Güler, N. & Uyanık, G. K. (2013). A Study on Multiple Linear Regression Analysis. *Procedia - Social and Behavioral Sciences*, 106, 234-240. <https://doi.org/10.1016/j.sbspro.2013.12.027>

- Haddawy, P., Nghe, N. T., & Janecek, P. (2007). A comparative analysis of techniques for predicting academic performance, *37th Annual Frontiers In Education Conference - Global Engineering: Knowledge Without Borders, Opportunities Without Passports*, Milwaukee, WI, USA, pp. T2G-7-T2G-12, doi: 10.1109/FIE.2007.4417993.
- Hemaid, R. K. & El-Halees, A. M. (2015). Improving Teacher Performance using Data Mining. *International Journal of Advanced Research in Computer and Communication Engineering*, 4(2), 407-412.
- Joshi, D. A. P., & Panchal, R. K. (2014). *Data Mining for Staff Recruitment in Education System using WEKA (1-4)*. Retrieved from https://www.researchgate.net/publication/345441279_Data_Mining_for_Staff_Recruitment_in_Education_System_using_WEKA1-4
- Kamath, R.S. & Warke, M.V. (2016). Data Mining Approach for the Analysis of Performance Based Appraisal System of Selected Teachers in Kolhapur City. *International Journal of Multifaceted and Multilingual Studies*, 3.
- Kaur, P., Singh, M., & Josan, G. S. (2015). Classification and Prediction Based Data Mining Algorithms to Predict Slow Learners in Education Sector. *Procedia Computer Science*, 57, 500-508. <https://doi.org/10.1016/j.procs.2015.07.372>
- Kavitha, K. (2021). Assessing Teacher's Performance Evaluation and Prediction Model Using Cloud Computing Over Multi-dimensional Dataset. *Wireless Personal Communications*, 119, 3207 - 3221.
- Klašnja-Milićević, A., Ivanović, M., & Budimac, Z. (2017). Data science in education: Big data and learning analytics. *Computer Applications in Engineering Education*, 25(6), 1066-1078. <https://doi.org/10.1002/cae.21844>
- Kulkarni, A. R. & Mundhe, S. D. (2017). Data Mining Technique: An Implementation of Association Rule Mining in Healthcare. *International Advanced Research Journal in Science, Engineering and Technology*, 4(7). doi: 10.17148/IARJSET.2017.4710
- Kumar, G. (2023, March 18). Introduction to Data Mining Classification - Gaurav Kumar - Medium. Medium. Retrieved April 16, 2024. <https://mobigaurav.medium.com/introduction-to-data-mining-classification->

NCON), Nan, Thailand, pp. 142-145, doi: 10.1109/ECTI-NCON.2019.8692274.

Sharma, P., & Sharma, S. (2018). Data Mining Techniques for Educational Data: A Review. *International Journal of Engineering Technologies and Management Research*, 5(2). <https://doi.org/10.29121/ijetmr.v5.i2.2018.641>

Tang, N., & Engelbrecht, A. (Ed.) (2022). Data Clustering. IntechOpen Series. Data Artificial Intelligence, Volume 10. doi: 10.5772/intechopen.95124

Tripathi, P. (2015). Competence Based Management in Academics through Data Mining Approach. *Computer Engineering and Intelligent Systems*. 6(2).

