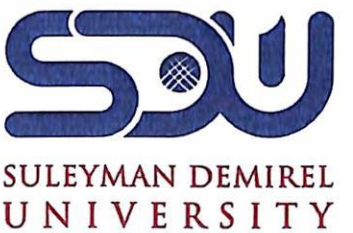
Ministry of Science and Higher Education of the Republic of Kazakhstan

Suleyman Demirel University



Aisaule Bazarkulova

Automatization of diploma processes based on analysis of supervisors' datasets

THESIS

Presented in Partial Fulfilment for the

Master of Technical Sciences Degree in Computer Science

(degree code: 7M06102)

Department of Computer Science

Faculty of Engineering and Natural Sciences

Supervisor: **PhD Assist Prof. Meirambek Zhaparov**

Kaskelen 2023

Suleyman Demirel University
Faculty of Engineering and Natural Sciences
Department of Computer Science

Dean of Faculty

Associate Professor

PhD Zhamanov A.

«06» 06 2023

Topic of the thesis:

Automatization of diploma processes based on analysis of supervisors' datasets

Thesis submitted as part of the requirements for the award of the MSc in "7M06102 - Computer Science" SDU, 2021-2023

Head of Department Assistant Professor, PhD Mukash Zh.

Academic Supervisor Assistant Professor,PhD Zhaparov M.

Master student Aisaule Bazarkulova

Kaskelen 2023

Declaration

I confirm that this is my own work and the use of all material from other sources has been properly and fully acknowledged.

Aisaule Bazarkulova

2023

Acknowledgements

I want to thank my supervisor PhD Assist Prof. Meirambek Zhaparov for valuable and invaluable help during my master's thesis work. Your professional guidance, expertise and mentorship have been invaluable to my education and success in the process. Thanks to your guidance and support, I have been able to expand my knowledge and skills in computer education and to formulate and develop my research . Your mentoring has helped me understand complex issues, find the right methods, and achieve meaningful results. I am immensely grateful for your trust, understanding and encouragement that you have given me throughout the process of working on my dissertation.

Abstract

The primary objective of the dissertation work was to extend and combine the hypothetical and useful information picked up in the learning procedure, and use them to research recommendation systems, in order to simplify the search for supervisors and students in areas of interest . To accomplish this objective, data was collected from different universities supervisors of Kazakhstan in the field of Information Technology. The object of study is to develop a system of recommendations based on the data of supervisors. As a result of the dissertation work, a research was conducted among clustering algorithms. The peculiarity and reliability of the obtained results is justified by the correlation between the rules and generally accepted standards of programming, consistency with the claims of the development environment and the technologies that were used.

Аңдатпа

Диссертациялық жұмыстың негізгі мақсаты оқу процесінде алған теориялық және практикалық білімдерін тереңдету және нығайту және оларды ұсынымдар жүйесін зерттеу үшін пайдалану, қызығушылық танытатын салаларда жетекшілер мен студенттерді іздеуді жеңілдету болды. Осы мақсатқа жету үшін Қазақстанның әртүрлі университеттерінің ақпараттық технологиялар саласындағы жетекшілерінің деректерін жинау орындалды. Зерттеу объектісі – жетекшілердің деректері негізінде ұсыныстар жүйесін әзірлеу. Диссертациялық жұмыстың нәтижесінде кластерлік алгоритмдер арасында зерттеу жүргізілді. Алынған нәтижелердің ерекшелігі мен сенімділігі ережелер мен бағдарламалаудың жалпы қабылданған стандарттары арасындағы корреляциямен, өңдеу ортасының талаптарымен және қолданылған технологиялармен сәйкестігімен негізделеді.

Аннотация

Ключевой целью диссертационной было углубление и укрепление теоретических и практических познаний, приобретенных в процессе изучения, и использование их для исследования рекомендательных систем , чтобы упростить поиск руководителей и студентов в интересующих областях . Для достижения поставленной цели был совершен сбор данных супервайзеров с разных университетов Казахстана в сфере Информационных Технолгии . Объект изучения – разработать систему рекомендаций, основанную на данных супервайзеров. В итоге выполнения диссертационной работы было проведено исследование среди алгоритмов кластеризации . Особенность и достоверность приобретенных результатов обоснована соотношением правилам и общепризнанным меркам программирования, согласованностью с притязаниями среды разработки и технологиями, которые были использованы .

Abbreviations

KNN K-Neighbors Classifier

DBSCAN Density-based spatial clustering of applications with noise

ML Machine Learning

TF Term Frequency

TDF Inverse Document Frequency

CGPA Cumulative grade point average

REP Reduced Error Pruning

WEKA Waikato Environment for Knowledge Analysis

ID3 In decision tree

Table of Contents

Chapter 1

Background and motivations

1.1 Introduction

The relationship between the supervisor and the student plays a critical role in achieving academic degrees [1]. A key factor in the effectiveness of scientific leadership is the ability of the supervisor to meet the requirements of the student and meet the expectations that students place on their supervisor. Mutual understanding and harmonious interaction become integral components of a successful process.

Academic supervisors perform various roles, including adviser, teacher, mentor, guide and critic[2]. They should be ready to help and support their students at any time, including the creation of a methodology, discussion of results, presentation and possible publication of dissertations.

One of the most important tasks is to find suitable students and research supervisors to ensure effective communication, create projects and apply research methods. However, finding students and supervisors who have common interests and goals is often a difficult task. We face difficulties in identifying suitable candidates in various places and areas where they show interest in research [3].

Based on this problem, I decided to choose an algorithm of clustering for a recommendation system for finding students and research supervisors, which will be based on the analysis of similar interests. Our goal is to find an most efficient

algorithm for recommendation system that will improve the search process and ensure successful cooperation between students and academic supervisors.

As part of my dissertation work, I plan to collect data from academic supervisors, including information about the problems they face when working with students, as well as the difficulties encountered in tracking progress.

1.2 Motivation

From the moment I planned to enter the master 's program , the problem of choosing a supervisor was one of my main tasks and . Increasing the effectiveness of scientific leadership is of great importance in the process of obtaining scientific degrees. The relationship between the supervisor and the student should be based on mutual understanding and satisfaction of the requirements of both parties. As a student, I understand that the supervisor performs not only the role of adviser and teacher, but also is a mentor, guide and critic. The whole process, starting with the development of methodology and ending with the presentation and publication of dissertations, requires the active participation of the supervisor and his willingness to help at every stage.

For this reason, I decided to choose this topic for my dissertation work in order to conduct a study of the motivation of scientific supervisors and create an innovative platform based on the analysis of similar interests. In general, I believe that improving the process of scientific leadership, solving difficulties in finding students and scientific supervisors are important motivations in my dissertation work.

1.3 Aims and objectives

The main purpose of the research is to choose an algorithm of clustering for the recommendation system that allows you to find students and research supervisors specializing in the field of information technology. To achieve this goal, various sources of information will be investigated. To find a simplified algorithm of clustering for a recommendation system for what helps to find a supervisor

and a student with similar interests, a study of clustering algorithms will be conducted. The goal is to select the most appropriate algorithm that will effectively group students and supervisors based on their interests. To fulfill the purpose of the work, data collection of scientific supervisors from various universities of Kazakhstan will be carried out. The collected data will be analyzed to identify similarities and differences between the interests and achievements of academic supervisors and students.

My goal is to create a system that will do things like:

- Study of different clustering algorithms .
- Collection of data from supervisors .
- Analyzing data .
- Choosing the most appropriate clustering algorithm for recommendation system .

Chapter 2

2.1 Literature review

Researches suggest that the recommendation system can be operated based on an assessment of students' interests and the scientific achievements of the supervisor [4]. This approach is focused on choosing an academic supervisor using Scopus quality metrics and matrix normalization rather than presenting suggested combinations and without creating co-authorship networks. At the same time, other researchers were also taking into consideration the preferences of the supervisors [5]. Some supervisors do not pay attention to students' background knowledge, but others do. Consequently, the authors divided supervisors into three groups: "exploitation" group represents those who prefer to use their existing knowledge, while the "exploration" group represents those who want to work with someone whose knowledge is somewhat different from their own, and a hybrid "moderate" group. For each group there are content-based, collaborative filtering, and hybrid recommendation approach, respectively. In earlier research, there was a similar approach that also contains three filtering techniques for recommender system [6]. They provided extensive categories for collaborative filtering that were divided into two: memory-based and model-based. Each one represents situational usage of various formulae and learning algorithms. Collaborative filtering offers mostly successful recommendations because it bases on object's similarity rather than item's similarity [7]. The authors propose a hybrid recommender system that integrates collaborative filtering with content-based filtering. The use of techniques like clustering, similarity, and classification increases the precision and accuracy of recommendations. There is a proposal based on friendship-based recommenda-

tion mechanism power in social media platforms [8]. Authors proposed the system that gathers data about the preferences of its users.. An algorithm for this system is mainly collaborative filtering using closeness measure as a friendship strength. The tool can utilize the information of closely connected users. Authors provided and verified that the performance of their approach was giving the best results in the term of evaluation metrics. Another study proposed a system where student thesis proposals are accepted by the suggested system, which then provides a list of potential supervisors in descending order of relevance [9]. The proposed recommendation system has two phases: indexing where each supervisor must upload their academic publications and recommending where the system accepts submitted thesis proposal of the student. It uses cosine similarity to compute the relevancy in a vector space model. The same methodology in other studies were used but before using cosine similarity they applied the inverse phrase to document frequency (TF-IDF) word weighting to find the most influential words in title and abstract text of the student research [10]. Both of these approaches were giving good results comparing the recommendations with the actual data. Some experiments demonstrated an utilization of three decision tree algorithms in the higher education system, including J48, random tree, and REP tree [11]. As part of the data mining process, data preparation and pre-processing were done before using WEKA, machine learning software in Java. In the first phase of the proposed model, the three algorithms were applied to the target dataset, and the generated decision trees are set to be compared to choose the best classifier. The chosen classifier is used to manage the student in selecting the major that would result in the highest cumulative grade point average (CGPA) in the subsequent phase. Since using decision trees as a prediction model in recommender systems requires building a large number of trees, researchers proposed a modification to the model in order to handle larger scales [12]. Their proposed algorithm is related to the ID3 algorithm [13] and uses the hybrid approach because the model requires only one tree. As the authors introduce, the considerable difference from the mentioned algorithm is in the leaf nodes of the tree. Proposed work unveiled satisfactory utility of the decision tree after experimental setups. A systematic review from other researchers says that interest in development of e-learning recommender systems is growing in recent years [14]. Comparing different types of

approaches, hybrid recommender systems and approaches based on supervised and unsupervised learning have the highest percentage that shows how effective hybrid and machine learning-based approaches are. Authors call attention to the significance of joining different recommendation methods that tend to raise the quality of recommendations. Another way of recommending is to use the cosine method[15]. Various aspects are considered, such as research interests, expertise, academic qualifications and accessibility. By incorporating these factors into the cosine similarity method, research demonstrates an integrated approach to providing accurate and appropriate advice.

Chapter 3

Method and methodology

3.1 Dataset collection

As part of the thesis work in the field of IT recommendatory systems, data collection plays a key role in understanding and analyzing user preferences and behavior. One effective data collection tool that can be used is Google Forms. Consider the importance and benefits of using Google Forms in the context of my research work.

The first step in data collection is to form questions that will help to obtain the necessary information. Questions can include various aspects. Utilizing Google Forms for collecting information offers many advantages. Google Forms has a simple and intuitive interface that makes creating and configuring surveys simple and accessible for users without programming expertise [16].

Google Forms' adaptability and customizability enable you to create various types of questions, add conditions, and modify access settings. This allows you to tailor the survey to your particular requirements and research objectives.

Google Forms also facilitates the distribution of surveys. You can easily distribute a survey link via email, social media, and other online platforms to your respondents. This enables you to collect information from more participants in less time.

Automatic data processing is an additional benefit of Google Forms. It automatically accumulates and organizes data and makes the results accessible in a convenient format. This simplifies the analysis and interpretation of data, as well as the export of data for further processing.

In addition, Google Forms provide online data access and storage. Your data will be stored in a Google cloud service, which ensures the security and availability of information from any place with the Internet.

In order to obtain a fresh and relevant set of data on the parameters of supervisors, such as age and general work experience, a survey was conducted among supervisors of different universities in the direction of IT of Kazakhstan .

№	Questions
1	Name Surname
2	Age
3	Gender
4	What is your level of study?
5	How long have you been a supervisor?
6	How many PhD and Master students have you supervised in year (average)?
7	Research area
8	What problems did you have during the work with students ?
9	What type of communication do you prefer ?

Table 3.1: Survey questions

Based on the table 3.1 Name, Surname, Age, Gender: These questions helped establish the context and identify the supervisor .

What is your level of study? This question helped to establish the qualifications of the supervisor, as it is an important factor in determining his competence in the field of IT recommendation systems.

Experience as a supervisor: this question helped you to assess his experience, stability and level of knowledge. How long have you been a supervisor? This gives

Table 3.2: Result of the survey

Participants	Quantity
Supervisors:	100
Students:	300

an idea of his ability to provide individual attention to each student.

Research area: Study of the field in which the supervisor conducts research in the field of IT recommendation systems.

What problems did you have during the work with students ? This question will help to understand the various challenges that a supervisor may face, such as limited resources, technical complexities or communication problems.

What type of communication do you prefer with students (online, offline or hybrid)? The question will let you know what approach to communication the supervisor considers most effective in the context of IT recommendation systems.

3.2 Using Google Sheets to analyze data

In the context of research, research analysis, sometimes referred to as data analysis, is the procedure of reviewing and interpreting gathered data in order to draw meaningful findings and make decision-based on evidence. In order to find patterns, relationships, trends, and insights within the data, it includes employing a variety of analytical approaches and statistical procedures. Understanding study findings, confirming hypotheses, and providing accurate interpretations all depend on research analysis.

The effectiveness of using Google Sheets for research purposes

The goal of research analysis is to turn unstructured data into knowledge that may answer research questions, confirm or deny hypotheses, and further our understanding of a given subject.

Important in modern scientific research is the use of efficient data analysis instruments. As an online spreadsheet provided by Google, Google Sheets provides researchers with a practical and affordable data analysis instrument. This article

examines the utility of Google Sheets for research purposes.

The user-friendliness of Google Sheets is one of its primary advantages. Even without extensive programming or statistical knowledge, researchers can rapidly master the fundamental functions and tools of tables. This enables you to spend less time on training and begin data analysis sooner.

Google Sheets provides a vast array of data analysis features and capabilities. Calculations and data aggregation are simplified by functions such as SUM, AVERAGE, COUNTIF, etc. There are also numerous charts and graphs available to visualize the study's findings.

Google Sheets provides the capacity for data collaboration. Researchers can invite coworkers to share tables, which facilitates collaboration and the exchange of information. The possibility of simultaneous data modification allows you to reduce approval time and increase the research team's output.

Google Sheets is compatible with Google Docs, Google Slides, and Google Drive. This enables researchers to exchange data and research results using electronic documents and presentations, as well as store data in the cloud for easier accessibility.

The analysis of the effectiveness of using Google Sheets is based on the experience of researchers who regularly use this tool in their work. Performance and user satisfaction data are also taken into account. The results of the analysis allow us to draw conclusions about the advantages and effectiveness of using Google Sheets in research work. It is noted that Google Sheets is a convenient, functional and affordable tool for data analysis. Its advantages include ease of use, collaboration, flexibility of functionality and integration with other tools. The results of the analysis confirm the effectiveness and practicality of using Google Sheets in research work.

To apply the descriptive analysis approach to the data analysis. Researchers may readily study bigger groups of people using this strategy [17]. When a dataset's major qualities need to be summed up and described, descriptive analysis is utilized in the study. It is frequently used in the early phases of data analysis

to improve comprehension of the data and to offer a succinct and unambiguous overview of the data's main characteristics. You may use descriptive analysis on a variety of data kinds, such as textual, numerical, or categorical data.

It also offers graphical displays like histograms, bar charts, and pie charts, as well as statistics like mean, median, mode, and standard deviation. Following are a few typical applications for descriptive analysis:

- Data Exploration: Descriptive analysis is often the first step in data exploration. Researchers use descriptive statistics to calculate measures of central tendency (e.g., mean, median) and measures of dispersion (e.g., standard deviation, range) to understand the distribution and variability of the data. They may also create visual representations like histograms, bar charts, and pie charts to summarize and visualize the data.
- Data Summarization: Descriptive analysis is employed to compress big datasets into digestible summaries. To give a clear picture of the data, researchers may compute summary statistics for various variables, such as mean scores, proportions, or counts. This aids in finding the dataset's most important trends, patterns, and outliers.
- Data Comparison: Descriptive analysis empowers analysts to compare and differentiate groups or categories inside the dataset. For case, they may calculate and compare implies or extents over distinctive statistic bunches or test conditions. This permits for bits of knowledge into contrasts or similarities among different groups or categories of intrigued.
- Preliminary Analysis: Descriptive analysis is very important in the early stages of research to gain insight into the data before proceeding to more complex analyses. This helps researchers understand the characteristics of their data, identify potential problems (such as missing data or outliers), and make informed decisions about subsequent analysis methods and research questions.
- Descriptive analytics provide researchers with key information to effectively report and communicate research results. Researchers can use descriptive

statistics and visualizations to present clear and concise summaries of data in research papers, reports, or presentations. This makes the results easier to understand and interpret for both experts and laypersons.

Generally, descriptive analysis serves as an establishment for assisting information exploration and examination in research. It gives an introductory understanding of the information, distinguishes key designs, and encourages successful communication of research discoveries.

3.3 Data and discussion

Using the results 3.2 of the survey, we conducted a descriptive analysis that resulted in the following:

In table 3.3 shows the average values of the collected data related to the age of the respondents, their work experience as supervisors and the number of students they supervised. These data are important information for understanding the characteristics and experience of supervisors in working with students.

Analysis of average values:

Age of respondents: The average age of respondents is 25-30 years. This indicates that supervisors from different age groups participated in the survey, which may mean a variety of experiences and approaches to working with students. The number of years spent working as a supervisor in total: The typical number of years spent working in this capacity is between one and five. This suggests that among those who responded, there are both relatively new supervisors who are just beginning their careers and more seasoned specialists who have gained a large amount of experience working with students.

Number of students under supervision: The number of students under supervision that has been reported by respondents so far ranges from one to five people on average. This implies that supervisors have some experience dealing with a particular number of students, which may have an effect on their ability to effectively lead and support students as they go through the process of their work.

Table 3.3: Mean Value

Question	Mean Value
Age	25-30
How many PhD and Master students have you supervised in year (average)?	1-5

Through conducting a study of the average values of the data provided by the respondents, we are able to arrive at judgments regarding the qualities and experiences of supervisors. The table demonstrates that the respondents are of varying ages, have varied levels of experience working in supervisory roles, and have previously collaborated with a certain number of student participants. This information is necessary for additional analysis and evaluation of the work done by supervisors in the context of their interactions with students.

Based on the histogram 3.1 following three primary problems have been indicated by the results of the survey of supervisors regarding their interactions with students:

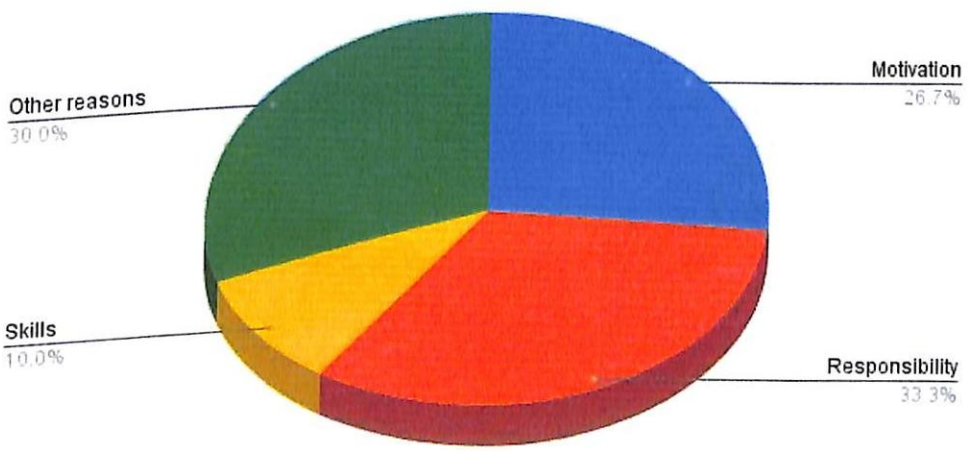


Figure 3.1: Histogram of main problems with working students

- Lack of responsibility on the part of students is one of the most significant factors contributing to the problem. The majority of students begin

projects, but they do not finish them; they miss meetings or arrive late to them; and some students do not even show up for meetings that have been planned. This may result in the completion of projects being delayed and may also make the task of supervisors more difficult.

- A further issue is the students' lack of motivation, which is a problem in and of itself. They are unable to comprehend the project's end outcome and do not acknowledge the existence of well-defined objectives. This might result in students not being as invested in the project as they should be, a lack of drive to reach their goals, and a lack of self-discipline when it comes to completing their assignments.

- Competencies: The third issue is that not all students are equipped with the competencies required to collaborate with an instructor. Some students may not have sufficiently developed skills such as the capacity to successfully interact with others, organize their work, plan their time and duties, evaluate and comprehend data. This may make it more difficult to cooperate with scientific supervisors and have productive interactions with them, which in turn may have an impact on the quality and efficiency of the work.

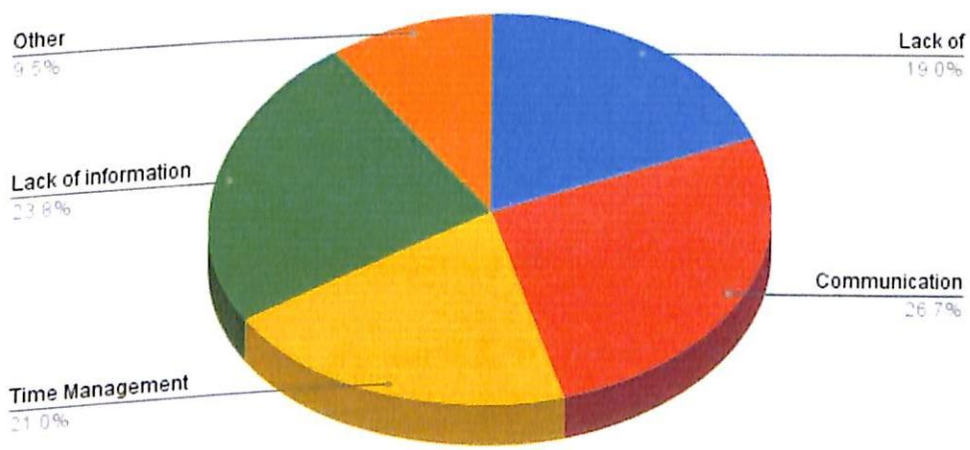


Figure 3.2: Histogram of main problems with working supervisors

Based on the figure 3.2 following is a list of the main problems that students experience when interacting with academic supervisors:

- Insufficient information and responsibility: It is problematic for students to obtain insufficient information from the academic supervisors who are responsible for their coursework. Some supervisors may not provide complete information about the project, goals, expectations, and roles of students. This leads to a vague understanding by students of their responsibility and how exactly they should contribute to the project. Lack of information can also lead to difficulties in completing tasks and achieving goals.
- Communication time management: The problem arises in the field of communication between students and academic supervisors, especially when it comes to planning and time management. Students may experience difficulties in establishing regular and effective communication with supervisors, in clarifying meetings, receiving feedback and support. This can lead to delays in work, misunderstanding of instructions and loss of direction in the research process.

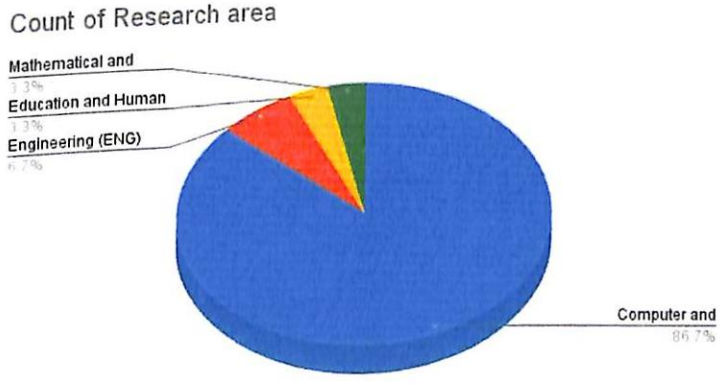


Figure 3.3: Histogram of the Research area

Based on the histogram Figure 3.3, it can be concluded that the majority of respondents conduct research in the field of information technology (IT). This means that the bulk of the research work that respondents do is related to IT.

Information technology is a broad field that covers various aspects, such as software development, network technologies, databases, artificial intelligence, cybersecurity, data analysis and others. It plays an important role in the modern world and has a significant impact on many industries, including business, medicine,

education, finance and others.

The fact that the majority of respondents conduct research in the field of IT may indicate the importance of this field and the interest of researchers in studying its various aspects. Perhaps this is due to the rapid development of technology, constant innovation and the need for relevant research that can help in the progress and improvement of the IT industry.

Chapter 4

Methods of Machine Learning

4.1 Overview

In our research, we have used various modern tools and techniques.

1. Back End - Python, Django, SQLite
2. Front End - Vue.js
3. Machine learning - Agglomerative Clustering,DBSCAN,Spectral Clustering,K-Neighbors Classifier

Python is a modern programming language, that is used in various fields of computer science, such as back end, data science, statistics, analytics, etc. We've decided to use Python due to its simplicity and ability to implement a machine-learning algorithm [18].

Django is a framework for the back end on Python language. It contains many libraries and tools for development under the hood. SQLite is a database system that can be hosted on a computer without any additional programs and drivers. It comes by default with Django.

Vue.js is a popular front-end framework, that is fast, capable, and useful.

We used pandas, numpy and scikit-learn libraries for our recommendation system machine-learning algorithm. These libraries are beneficial for working with

data.

Machine learning algorithms

For our recommendation system, we have tried to use the following algorithms:

- Agglomerative Clustering
- DBSCAN
- Spectral Clustering
- K-Neighbors Classifier

We have decided to use these algorithms of machine learning, because these algorithms are used for solving recommendation problems by top-tech companies like Facebook, Amazon, Google, and Netflix. These algorithms fit the requirements of our problem the most.

Score metrics

For our machine-learning algorithms, we use the following scoring metrics:

- F1-score
- Rand score
- Runtime

F1-score is a metric used to assess the quality of classification models, especially in the case of unbalanced classes. It combines two metrics - precision and recall to get one overall score [19].

Precision is a measure of how accurately the model classifies positive examples. It is defined as the ratio of the number of correctly classified positive examples to the total number of examples that the model assigned to a positive class. Formally, the accuracy is calculated as follows:

$$Precision = \frac{Correctly\ classified\ positive\ examples}{(Correctly\ classified\ positive\ examples + Falsely\ classified\ positive\ examples)} \tag{4.1.1}$$

Recall, also known as sensitivity or completeness, measures what proportion of positive examples a model is able to correctly detect. It is defined as the ratio of the number of correctly classified positive examples to the total number of positive examples in the data. Formally, the recall is calculated as follows:

$$Recall = \frac{Correctly\ classified\ positive\ examples}{(Correctly\ classified\ positive\ examples + False\ negative\ examples)} \tag{4.1.2}$$

The F1 measure is the harmonic mean between accuracy and recall. It represents a balance between accuracy and recall, taking into account both false positive and false negative errors. Formally, the F1-measure is calculated by the following formula:

The calculation of the F1- score allows you to take into account and evaluate both the accuracy and the recall of the model at the same time. This is especially useful when classes are unbalanced and it is important to achieve both high accuracy and high recall for both classes.

Using the F1-score allows you to get one number that characterizes the quality of the model as a whole. The closer the value of the F1 score is to 1, the better the model copes with the classification.

F1-score calculates the precision and recall of the data and then uses the following formula to calculate a score of the model:

$$F1 - score = \frac{2 * (Precision * Recall)}{(Precision + Recall)} \tag{4.1.3}$$

Rand score is a metric that describes a similarity between train and test data sets. It is used to evaluate the performance of the clustering algorithm [20]. Rand score calculates true positive and true negative values for data sets.

The Rand score is based on comparing pairs of objects and determining whether they belong to the same cluster or to different clusters in different partitions. The evaluation is based on the concept of four possible combinations of pairs of objects:

True Positives (TP) - pairs of objects that belong to the same cluster in both the first partition and the second partition.

True Negatives (TN) - pairs of objects that do not belong to the same cluster in both the first partition and the second partition.

False Positives (FP) - pairs of objects that do not belong to the same cluster in the first partition, but belong to the same cluster in the second partition.

False Negatives (FN) - pairs of objects that belong to the same cluster in the first partition, but do not belong to the same cluster in the second partition.

The Rand score calculates the number of true positive (TP) and true negative (TN) pairs of objects in the data partitions and uses them to calculate the consistency coefficient. Formally, the Rand score is calculated as follows:

The value of the range estimate can vary from 0 to 1. A value of 1 indicates a complete coincidence of the data splits, while a value of 0 indicates a complete lack of similarity between the splits. Values close to 0.5 indicate a random split of the data.

Range estimation is a useful metric for comparing different clustering algorithms or evaluating the quality of clustering. A high value of the Rand score indicates strong consistency between partitions, which indicates good performance of the clustering algorithm.

$$R1 = \frac{(a+b)}{\binom{n}{2}} = \frac{correct\ similar\ pairs + correct\ dissimilar\ pairs}{total\ possible\ pairs} \tag{4.1.4}$$

Runtime is a metric that shows the time of algorithm evaluation. Runtime is an important metric because it will decrease the cost of server maintenance and users will get their response faster.

Runtime is an important indicator because it directly affects the performance and efficiency of the system. A lower runtime means that an operation or algorithm is executed faster, which can be critical for users, especially in the case of

online services or applications where a fast response is needed to meet user needs.

Reducing the runtime has a number of advantages. Firstly, it allows you to reduce the load on the server and reduce maintenance costs, since fast execution of operations requires less server resources. This can be especially important when working with a large amount of data or highly loaded systems.

In addition, the low runtime increases user satisfaction, as they receive the results of an operation or query faster. This can enhance the overall user experience and improve the interaction with the system.

Various methods can be used to estimate the runtime, including measuring the real execution time of an operation or algorithm on specific data, or using asymptotic complexity to evaluate the performance of the algorithm depending on the size of the input data.

Runtime estimation and performance optimization are important tasks in various fields such as software development, computational science, databases, machine learning, and others. Efficient use of resources and reduced execution time are key factors for improving system performance and meeting user needs.

Data

As a data source, we used data of SDU students. This data contains:

- hashed students names
- hashed teachers names
- courses
- grades

We have collected about 100 supervisors and 300 students' data. We knew this was not enough, so we generated augmented data based on the data we had. After the generation of augmented data, we had almost 2000 data fields with different projects for different users. We used 70% percent of the data for training and 30% percent for testing to validate and fine-tune the model. For the machine learning algorithm, used in our diploma project, we have found the best

suitable features using an automatic feature selection technique called Recursive Feature Elimination (RFE). RFE works by recursively removing features from the dataset until it finds the best suitable feature set. There are specific steps for RFE execution:

1. Train the model: We chose our ML models, and the best suitable algorithms for our problem were clustering algorithms (AgglomerativeClustering, DBSCAN, SpectralClustering, KNeighborsClassifier).
2. Feature ranking: The trained model, now assigns a rank to each to each feature, based on how important was this feature for the prediction.
3. Features elimination: We've eliminated features, that didn't have any impact on the target variable.
4. Model retraining: Retrained the model with new features.
5. Comparing results: We've compared results with different features and algorithms.

Agglomerative Clustering

One of the algorithms used in our recommendation system is Agglomerative Clustering. It is a machine-learning approach that groups data points based on how far apart they are from one another[21].

Based on the figure 4.1 process of agglomerative clustering begins with the fact that each piece of data is considered as a separate cluster at the first iteration of the algorithm. After that, a proximity matrix is computed in order to estimate the proximity between clusters. This matrix includes information on the distances or similarities between each pair of clusters and is used to determine the proximity between clusters. The distance measure that is calculated by the algorithm is determined, in part, by the kind of data that is being processed by the system.

During each iteration of agglomerative clustering, the proximity matrix is used to choose the two groups that are geographically closest to the input data. After that, these clusters are brought together to form a single new cluster, and the proximity matrix is modified to reflect these changes. The procedure of merg-

ing existing proximity matrices and updating them takes place in an iterative manner until a predetermined completion condition is satisfied. The completion requirement could be hitting a certain distance threshold, reaching a certain number of clusters, or meeting some other criteria that determines the quality of the clustering.

Agglomerative clustering offers a variety of benefits to its users. To begin, it does not require a specific number of clusters, which enables you to easily modify the method in accordance with the specifics of the data you are working with. Second, it is capable of processing data of many different types, including textual, categorical, and quantitative data. Third, agglomerative clustering makes it possible to investigate the hierarchical structure of the data by revealing information on the nesting of clusters and the connections between them.

Agglomerative clustering does, however, have a few drawbacks to consider. To begin, it may be computationally difficult and call for a significant amount of processing resources, particularly when working with a large amount of data. Second, the quality of the findings and their ability to be interpreted can be considerably impacted by the selection of an appropriate distance metric and cluster aggregation approach. When utilizing agglomerative clustering, therefore, it is essential to ensure that the appropriate parameters are selected and that the algorithm is configured correctly.

When applied to the setting of a recommendation system, agglomerative clustering can be used to group persons or items on the basis of their similarities in order to make recommendations based on the qualities or preferences of the group as a whole. This can be helpful for developing personalized suggestions that take into consideration the similarity between users or items, as well as increasing the overall quality of recommendations that are generated by the system.

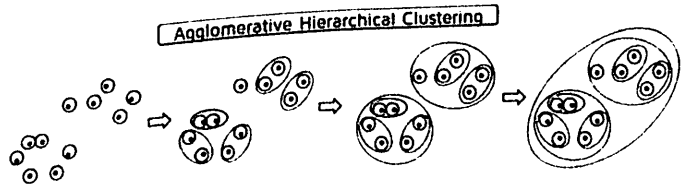


Figure 4.1: Agglomerative Clustering

Agglomerative clustering is a technique that is frequently utilized in recommendation systems like those found on Amazon, Netflix, Spotify, and other similar platforms.

Function *agglomerative_clustering()*

```
start = timer();
agg_clustering = AgglomerativeClustering(n_clusters=3);
agg_clustering.fit(X_train);
y_pred = agg_clustering.fit_predict(X_test);
agg_f1_score = f1_score(y_test, y_pred, average='micro');
agg_rand_score = rand_score(y_test, y_pred);
end = timer();
print("AgglomerativeClustering F1 score: " + agg_f1_score);
print("AgglomerativeClustering Rand score: " + agg_rand_score);
print("AgglomerativeClustering runtime: " + (end - start) + " s.");
```

Algorithm 1: Agglomerative Clustering evaluation

Density-Based Spatial Clustering of Applications with Noise

A complicated method known as quad DBSCAN is used to cluster data points by grouping them together into clusters while taking into account the distance between them and the number of neighbors who are located within their radius. The advantage of this algorithm is that it deals with noise/outliers by removing them [21].

There are the following parameters for the DBSCAN algorithm:

- Epsilon - defines the radius of clusters
- Minimum points - min. number of data points to create a cluster

DBSCAN steps:

1. Distance calculation: For each point in the dataset, the distance to all other points is calculated. This can be the Euclidean distance or another distance metric, depending on the characteristics of the data.
2. Definition of the main points: If the number of points inside the epsilon radius is greater than or equal to the minimum value (minPts), then the

point is considered the main point. The main points are central to the formation of clusters.

3. Cluster formation: For each main point, a cluster is formed by adding all points located within the epsilon radius from this main point, as well as their neighboring points satisfying the minPts condition. If a point is not the main one and does not belong to any of the clusters, it is considered an outlier.

4. Elimination of noise points: At the end of the clustering process, outliers and noise points that do not belong to any cluster are excluded from the resulting dataset.

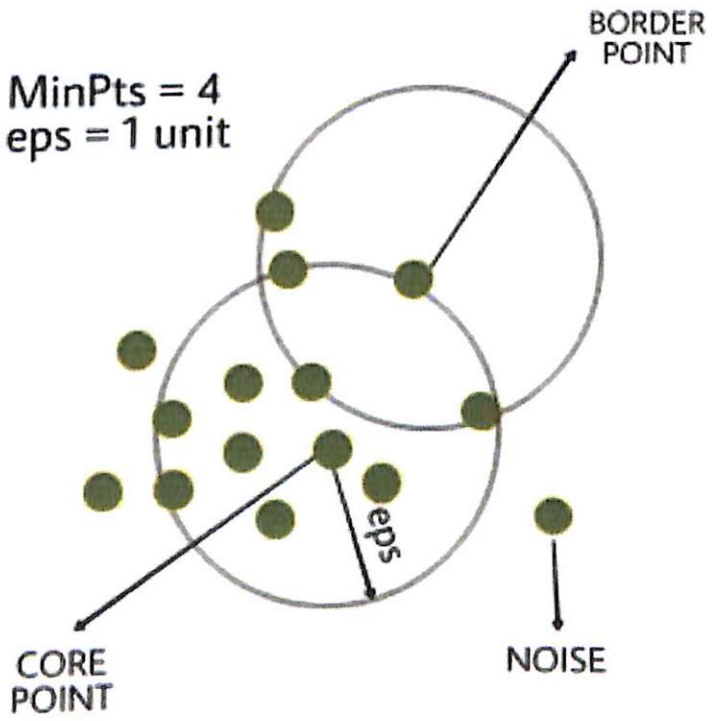


Figure 4.2: Density-Based Spatial Clustering of Applications with Noise

The DBSCAN algorithm allows efficient processing of noise points and differs from other clustering methods such as agglomerative clustering and spectral clustering. It is based on the concept of density, which allows it to identify clusters of arbitrary shape and detect boundaries between clusters. See figure 4.2.

DBSCAN algorithm is used in GPS services (Google Maps, Yandex.Maps), social media and online advertisement, fraud detection, etc.

Function *dbscan()*

```
start = timer();
dbscan = DBSCAN(eps=0.5, min_samples=5);
dbscan.fit(X_train);
y_pred = dbscan.fit_predict(X_test);
dbscan_f1_score = f1_score(y_test, y_pred, average='micro');
dbscan_rand_score = rand_score(y_test, y_pred);
end = timer();
print("DBSCAN F1 score: " + dbscan_f1_score);
print("DBSCAN Rand score: " + dbscan_rand_score);
print("DBSCAN runtime: " + (end - start) + " s.");
```

Algorithm 2: Density-Based Spatial Clustering evaluation

Spectral Clustering

This algorithm has a different approach, in comparison to others used. It uses eigenvectors and eigenvalues of a similarity matrix. This algorithm is very useful for finding complex relationships in the data.

Steps for spectral clustering algorithm:

- Calculation of the similarity matrix: First, a similarity matrix is calculated, which measures the degree of similarity or distance between each pair of data points in the dataset. It can be a distance matrix or a similarity matrix, depending on the chosen metric.
- Laplace graph construction: Based on the similarity matrix, a Laplace graph is constructed. The Laplace graph represents data as a graph where data points are vertices and edges represent connections between points based on their similarity. The graph can be constructed as a complete graph or using stricter criteria such as k-nearest neighbors.
- Calculation of eigenvectors and eigenvalues: The eigenvectors and eigenvalues of the Laplace graph matrix are calculated. The eigenvectors represent a

set of new features for the data, and the eigenvalues show their importance. Usually, the first few eigenvectors corresponding to the smallest eigenvalues are selected.

- The generated eigenvectors are utilized in the clustering process, which was named after its end result. The traditional method involves making use of clustering techniques such as k-means in order to classify new features that have been derived from eigenvectors. You can also make use of spectral clustering, which is a technique that determines the best number of clusters and their boundaries by making use of information about eigenvalues.

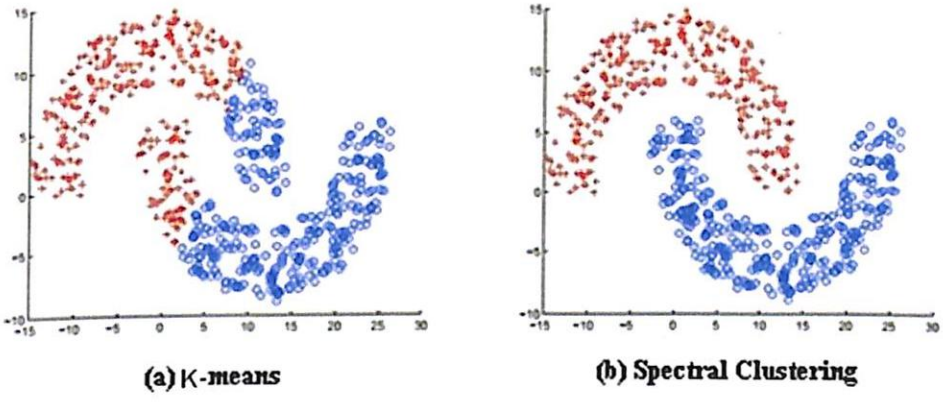


Figure 4.3: Spectral clustering

The spectral clustering method is very helpful when discovering complicated linkages in data, such as nonlinear dependencies or unequal clusters. This is because of the algorithm's ability to recognize these types of relationships. It enables you to take into account the global data structure and has the potential to produce positive results in a variety of contexts [22].

The image segmentation, document clustering, and recommendation system are just few of the applications that are suitable for this clustering approach.See figure 4.3.

Function *spectral()*

```
start = timer();
spectral_clustering = SpectralClustering(n_clusters=3);
spectral_clustering.fit(X_train);
y_pred = spectral_clustering.fit_predict(X_test);
spectral_f1_score = f1_score(y_test, y_pred, average='micro');
spectral_rand_score = rand_score(y_test, y_pred);
end = timer();
print("SpectralClustering F1 score: " + spectral_f1_score);
print("SpectralClustering Rand score: " + spectral_rand_score);
print("SpectralClustering runtime: " + (end - start) + " s.");
```

Algorithm 3: Spectral clustering evaluation

K-Neighbors Classifier

KNN is a straightforward yet highly effective method for classification. In our use case, the performance of our method has been superior to that of other algorithms. This technique makes use of an approach known as instance-based learning, in which the classification of new data is determined by the labels of the class that is geographically closest to it[23].

The KNN algorithm consists of the following steps:

1. Training: In this step, the model receives a training sample consisting of vectors of objects and their corresponding class labels. Each object is represented by a set of attributes, and each class label determines whether an object belongs to a certain class. The training sample is used to build the KNN model.

2. Distance calculation: When new data is received for classification, the KNN algorithm calculates the distance between each data point in the training sample and the new data point. Most often, Euclidean distance or another distance metric is used to determine the proximity between points.

3. Neighbor Selection: KNN then selects the K nearest neighbors of the new data point. K is a user-defined number, usually a hyperparameter of the model.

4. Weighted voting: When the nearest neighbors are selected, each of them gives a vote in determining the class of the new data point. The weight of the voice can be determined by distance, for example, the nearest neighbors may have more weight than the more distant ones. The class to which most of the neighbors belong (by votes) is selected as the predicted class for the new data point.

5. Notation: In the last step, KNN assigns the predicted class to a new data point. This can be classification based on class labels or, in some cases, regression if class labels are numeric values.

KNN algorithm is often used for image classification, handwriting recognition, medical diagnosis, recommendation systems, etc . The algorithm is simple but very useful and handy. In our case, KNN showed the best accuracy of the F1-score and random score. See figure 4.4.

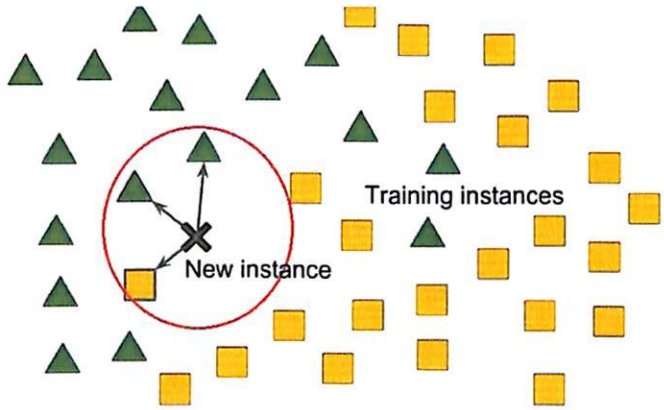


Figure 4.4: KNN

Function *knn_classifier()*

```
start = timer();
knn = KNeighborsClassifier(n_neighbors=3);
knn.fit(X, y);
y_pred = knn.predict(X_test);
knn_f1_score = f1_score(y_test, y_pred, average='micro');
knn_rand_score = rand_score(y_test, y_pred);
end = timer();
print("KNN F1 score: " + knn_f1_score);
print("KNN Rand score: " + knn_rand_score);
print("KNN runtime: " + (end - start) + " s.");
```

Algorithm 4: KNN algorithm evaluation

4.2 Result

To assess the effectiveness of the various clustering and classification algorithms, F1 score, random score and runtime were used. The F1 score is a measure of model accuracy and completeness, and its value closer to 1 indicates more accurate and complete results. A random score is used to compare results with a random model, and a value closer to 1 indicates a higher accuracy. Running time is important, especially when dealing with large amounts of data, as fast performance can be an important factor when choosing an algorithm.

The results of evaluation of algorithms of agglomerative clustering, DBSCAN, spectral clustering and K-neighbor classifier showed the following values:

Table 4.1: Aglomerative Clutering

Aglomerative Clutering	Value
F1 score	0.07
Rand score	0.579108138238573
runtime	0.03682747099998096

Table 4.2: DBSCAN

DBSCAN	Value
F1 score	0.01
Rand score	0.3837235228539576
runtime	0.018435167000006913

Table 4.3: Spectral Clustering

Spectral Clustering	Value
F1 score	0.15333333333333332
Rand score	0.5872017837235228
runtime	0.721807449000039

Table 4.4: K-Neighbors Classifier

K-Neighbors Classifier	Value
F1 score	0.37666666666666665
Rand score	0.6722408026755853
runtime	0.04209975199995597

Table 4.5: K-Neighbors Classifier

Based on the results 4.1,4.2,4.5,4.3 it is possible to conclude that the K-neighbor classifier shows the best efficiency among the considered algorithms. The F1 scores and random scores for KNN are higher, indicating more accurate results than other algorithms. It is worth noting, however, that by adding more data to the learning part of the algorithm, the evaluation indicators can improve, indicating the potential for improving model accuracy with more data.

Thus, on the basis of the analysis, it is possible to recommend the use of the K-neighbor classifier as the most efficient algorithm for a given dataset.

Chapter 5

Conclusions and future work

5.1 Discussion

The F1 score is a measure of model accuracy and completeness, with a value closer to 1 indicating more accurate and complete results. The random score is used as a comparison with a random model, and a value closer to 1 indicates higher accuracy. Runtime is also considered, as faster performance is important when dealing with large amounts of data.

Based on the results in the tables 4.1, 4.2, 4.3, 4.5. It can be concluded that the K-Neighbors Classifier demonstrates the best efficiency among the considered algorithms. It achieves higher F1 scores and random scores, indicating more accurate results compared to the other algorithms. However, it's important to note that the evaluation indicators can further improve with the addition of more data to the learning process, suggesting the potential for enhanced model accuracy with increased data.

In summary, based on the analysis, the research recommends utilizing the K-Neighbors Classifier as the most efficient algorithm for the given dataset.

The provided research text describes the accuracy result of a system that recommends prospective supervisors to students based on their research topics. The accuracy testing was conducted by comparing the system's recommended supervisors with the actual data. The results of the accuracy testing are presented

in Table 13, where each row represents a different query and the recommended supervisors are compared with the actual supervisors[18].

For example, in the first query (Q1), the system recommended supervisors S43, S21, and S39. The actual supervisors were S39 and S7, and the system's recommendation was true. The accuracy testing was performed for 20 experiments, and the overall accuracy rate is calculated as 75% .

The conclusion of the research suggests that the system, which utilizes cosine similarity as a method for determining supervisors, can effectively provide recommendations aligned with the research topics of the students. The system aims to facilitate the process of finding supervisors who match the students' research interests.

In terms of future work, the research suggests automating the classification of research list topics for prospective research supervisors. This can be achieved by incorporating machine learning approaches such as Naïve Bayes classification or Neural Network classification algorithms. Automating the classification process can streamline the addition of new research supervisors to the system's database.

Overall, the accuracy result indicates that the recommendation system performs reasonably well, with a 75% accuracy rate based on the comparison between the system's recommendations and the actual supervisors.

Comparing the two accuracy results:

Clustering and Classification Algorithms: The evaluation focused on the performance of clustering and classification algorithms. Metrics used: F1 score, random score, and runtime. The K-Neighbors Classifier was identified as the most efficient algorithm based on higher F1 scores and random scores, indicating more accurate results. The analysis highlighted the importance of runtime, especially when dealing with large datasets. The potential for improved accuracy with more data was mentioned. Overall, the research recommended the use of the K-Neighbors Classifier for the given dataset. Supervisor Recommendation System: The accuracy testing was conducted for a system that recommends prospective supervisors based on research topics. The evaluation compared the system's recommended

supervisors with the actual data. Metrics used: Comparison of recommended and actual supervisors for 20 queries. The overall accuracy rate was calculated as 75%. The system utilized cosine similarity for determining supervisors. Future work suggested automating the classification of research list topics using machine learning approaches. The system aimed to facilitate the process of finding supervisors aligned with students' research interests. Comparison:

Both research texts focus on assessing the accuracy of different systems/algorithms but in different domains. The Clustering and Classification Algorithms research evaluated the algorithms' performance using metrics like F1 score and random score, considering runtime as well. The Supervisor Recommendation System research tested the accuracy of a recommendation system by comparing recommended and actual supervisors. The Clustering and Classification Algorithms research identified the K-Neighbors Classifier as the most efficient algorithm, while the Supervisor Recommendation System achieved a 75% accuracy rate. Both research texts mentioned the potential for improved accuracy with more data or the use of machine learning approaches. In summary, while both research texts discuss accuracy results, they are based on different methodologies and domains. The Clustering and Classification Algorithms research evaluated algorithms' performance, whereas the Supervisor Recommendation System research focused on a recommendation system for supervisors.

5.2 Conclusion

Students' overall intellectual development is significantly impacted by the contributions of prominent scientific supervisor. The ability of a scientific leader to successfully satisfy the needs, expectations, and facilitate harmonic interaction of their students is an essential part of a successful scientific leadership process.

Scientific supervisors are responsible for a wide variety of tasks, including serving as an advisor, educator, mentor, guide, and critic. They need to be ready to help and support their students throughout the entirety of the research process, beginning with the formulation of the technique and continuing all the way to the publishing of the findings.

The task of finding qualified students and scientific leaders who share similar interests and objectives is a hard one. It is necessary to have a system that is able to effectively connect students and those who make decisions, taking into consideration the preferences and requirements of the scientists involved.

The thesis required the collecting of data from researchers, which included information on the challenges faced by researchers while working with students and the obstacles experienced when measuring progress. The method that would be employed by the recommendation system was chosen with the use of this data.

In general, the establishment of a reliable system for searching students and supervisors is an important step that needs to be taken to improve academic counselling and improve the overall academic performance of students. The thesis is an analysis of classification algorithms for the recommendation system based on the analysis of similar interests. This system can significantly simplify and improve the process of selecting a research supervisor for students in the future. In addition, the collected data and research results could be used as a basis for further research and development of a recommendation system platform.

5.3 Future work

Integration of numerous data sources may be the subject of additional research in the field of analyzing the data set of academic supervisors and constructing a recommendation system for students. This may be done in order to generate suggestions that are both more accurate and complete. Data about student demographics, academic records, and participation in extracurricular activities can be included in the report, in addition to information regarding academic supervisors. This can include details such as the academic supervisor's professional experience, area of expertise, and connections with students.

Integrating additional data sources, such as demographic data, can be an effective way to gain a more in-depth understanding of the preferences and requirements of students. When proposing a supervisor, for instance, it is helpful to take into account gender, age, nationality, and other demographic aspects. This is because doing so allows one to take into account the potential preferences of

students as well as the socio-cultural milieu in which they find themselves.

The academic histories of students, including details such as their prior achievements, grades, and academic preferences, are also eligible for inclusion in the collection. The examination of these data can lend a hand in determining the nature of the connection that exists between the academic qualities of a student and the academic advisors who are most suited to guide them. Students who have demonstrated prior academic performance or who have particular academic interests may, for instance, obtain suggestions regarding academic supervisors who specialize in the relevant disciplines.

The participation of students in extracurricular activities and the gathering of useful information from those activities is also a possibility. For instance, students who take an active role in scientific conferences, projects, or other extracurricular activities may give preference to research supervisors who have prior experience in these areas and are able to assist them with their research work.

However, for the successful deployment of recommendation systems, it is also vital to pay attention to the development of a user-friendly and intuitive interface for students. This is because students are the ones who will be using the system. When it comes to the suggestions' acceptance and application in a real-world setting, the ease of use and accessibility of the recommendations play a significant influence. When it comes to choosing a academic supervisor , students shouldn't have any trouble gaining access to the recommendations, and the materials that are offered to them should be arranged in such a way that makes it simpler for them to comprehend and use the advice.

In general, prospective future research in the field of analysis of the data set of supervisors and the creation of recommendation systems for students may focus on combining diverse data sources in order to provide guidance that is both more accurate and more detailed. At the same time, it is essential to provide students with an interface that is user-friendly and straightforward, as well as to take into consideration demographic data, academic data, and extracurricular activity data. Additionally, it is essential to keep in mind the fact that students. Students can obtain a greater level of success in their research projects with the assistance

of this type of innovation, which has the potential to considerably improve the process of management selection.

Bibliography

[1] Adrian Eley and Roy Jennings. Effective postgraduate supervision: improving the student/supervisor relationship. Open University Press McGraw-Hill Education, 2005. ISBN 9780335217083.

[2] B. Hon Kam. Style and quality in research supervision: the supervisor dependency factor. pages 81–103, 1997. URL https://doi.org/10.1023/A:1002946922952.

[3] Serek Azamat, Bazarkulova Aisaule, Chazhabayev Abylaikhan, and Akhmetov Adil. Analysis of supervisors and students in the context of diploma defense. 2021. doi: 10.1109/ICECCO53203.2021.9663776.

[4] Vladimir Kazakovtsev, Svyatoslav Oreshin, Alexey Serdyukov, Egor Krasheninnikov, Sergey Muravyov, Albert Bezvinnyi, Alexander Panfilov, Igor Glukhov, Yulia Kaliberda, Daniil Masalskiy, Timofey Podolenchuk, and Maksim Khlopotov. Recommender system for an academic supervisor with a matrix normalization approach. pages 84–87, 10 2020. doi: 10.1145/3437802.3437817.

[5] Ming-Yu ZHANG and Jian-Shan SUN. An intelligent recommendation service for student-selection on research social network: Bridging the gap between students and supervisors—research-in-progress. DEStech Transactions on Social Science, Education and Human Science, 02 2019. doi: 10.12783/dtssehs/icesd2019/28161.

[6] Folasade Isinkaye, Yetunde Folajimi, and Bolanle Ojokoh. Recommendation systems: Principles, methods and evaluation. Egyptian Informatics Journal, 16, 08 2015. doi: 10.1016/j.eij.2015.06.005.

[7] Geetha Gunasekar, Safa Iqubal, Fancy Chelladurai, and D Saranya. A hybrid approach using collaborative filtering and content based filtering for recommender system. Journal of Physics: Conference Series, 1000:012101, 04 2018. doi: 10.1088/1742-6596/1000/1/012101.

[8] Seo Young Duk, Young-Gab Kim, Euijong Lee, and Doo-Kwon Baik. Personalized recommender system based on friendship strength in social network services. Expert Systems with Applications, 69, 10 2016. doi: 10.1016/j.eswa.2016.10.024.

[9] Maresha Wijanto, Rachmi Rahmadiany, and Oscar Karnalim. Thesis supervisor recommendation with representative content and information retrieval. Journal of Information Systems Engineering and Business Intelligence, 6: 143–150, 10 2020. doi: 10.20473/jisebi.6.2.143-150.

[10] Ridwan Rismanto, Arie Rachmad Syulistyo, and Bebby Pramudya Citra Agusta. Research supervisor recommendation system based on topic conformity. International Journal of Modern Education and Computer Science, 12:26–34, 2020.

[11] Mohamed Hegazy and Hoda Waguih. A proposed academic advisor model based on data mining classification techniques. International Journal of Advanced Computer Research, 8:129–136, 05 2018. doi: 10.19101/IJACR.2018.836003.

[12] Amir Gershman, Amnon Meisels, Karl-Heinz Lüke, Lior Rokach, Alon Schclar, and Arnon Sturm. A decision tree based recommender system. pages 170–179, 01 2010.

[13] J. R. Quinlan. Induction of decision trees. Machine Learning, 1(1):81–106, Mar 1986. ISSN 1573-0565. doi: 10.1007/BF00116251. URL https://doi.org/10.1007/BF00116251.

[14] Huiyi Tan, Junfei Guo, and Yong Li. E-learning recommendation system. pages 430–433, 01 2008. doi: 10.1109/CSSE.2008.305.

[15] Hairani. Recommendations of thesis supervisor using the cosine similarity method. september 2022. doi: 10.32520/stmsi.v11i3.2003.

[16] HUI-YIN, SHIANG-KWEI HSU, and WANG. Using google forms to collect and analyze data. 40, april|may 2017.

[17] M.J. Goertzen. Introduction to quantitative research and data. 53, 2017.

[18] Simon, Harry Yuill, and Halpin. Python. 2006.

[19] Joos Korstanje. The f1 score. August 2021. URL https://towardsdatascience.com/the-f1-score-bec2bbc38aa6.

[20] Kay Jan Wong. 7 evaluation metrics for clustering algorithms. December 2022. URL https://towardsdatascience.com/7-evaluation-metrics-for-clustering-algorithms-bdc537ff54d2#7226.

[21] Alboukadel Kassambara. Practical Guide To Cluster Analysis in R. STHDA, 2017. URL http://www.sthda.com.

[22] William Fleshman. Spectral clustering. February 2019. URL https://towardsdatascience.com/spectral-clustering-aba2640c0d5b.

[23] K-nearest neighbors algorithm. URL https://www.ibm.com/topics/knn.