HYBRID APPROACH TO COMPLEX NETWORK-BASED LINK PREDICTION FOR RECOMMENDATION SYSTEMS IN TURKISH PUBLICATIONS

2024
MASTER THESIS
COMPUTER ENGINEERING

Ali Asghar Fahad FAHAD

Thesis Advisor
Assist. Prof. Dr. Emrah ÖZKAYNAK

HYBRID APPROACH TO COMPLEX NETWORK-BASED LINK PREDICTION FOR RECOMMENDATION SYSTEMS IN TURKISH PUBLICATIONS

Ali Asghar Fahad FAHAD

Thesis Advisor
Assist. Prof. Dr. Emrah ÖZKAYNAK

T.C.
Karabuk University
Institute of Graduate Programs
Department of Computer Engineering
Prepared as
Master Thesis

KARABUK
January 2024

I certify that, in my opinion, the thesis submitted by Ali Asghar Fahad FAHAD titled "HYBRID APPROACH TO LINK PREDICTION IN COOPERATION NETWORKS CREATED BETWEEN AUTHOR AND PUBLISHER IN TURKISH PUBLICATIONS" is fully adequate in scope and quality as a thesis for the degree of Master of Science.

Assist. Prof. Dr. Emrah ÖZKAYNAK ……...

Thesis Advisor, Department of Computer Engineering

This thesis is accepted by the examining committee with a unanimous vote in the Department of Computer Engineering as a Master of Science thesis. 16/01/2024

Examining Committee Members (Institutions)	Signature
Chairman: Assoc. Prof. Dr. Caner ÖZCAN (KBU)	…….
Member: Assist. Prof. Dr. Emrah ÖZKAYNAK (KBU)	…….
Member: Assist. Prof. Dr. Muhammet ÇAKMAK (SNU)	Online

The degree of Master of Computer Engineering by the thesis submitted is approved by the Administrative Board of the Institute of Graduate Programs, Karabük University.

Assoc. Prof. Dr. Zeynep ÖZCAN .....

Director of the Institute of Graduate Programs

"The information included in this thesis was fully gathered and presented in compliance with academic rules and ethical guidelines, I therefore declare. Furthermore, I have assiduously followed the guidelines and standards set forth by these rules, properly citing any sources that are not the author's own."

Ali Asghar Fahad FAHAD

ABSTRACT

M. Sc. Thesis

HYBRID APPROACH TO COMPLEX NETWORK-BASED LINK PREDICTION FOR RECOMMENDATION SYSTEMS IN TURKISH PUBLICATIONS

Ali Asghar Fahad FAHAD

Karabük University
Institute of Graduate Programs
Department of Computer Engineering

Thesis Advisor:
Assist. Prof. Dr. Emrah ÖZKAYNAK
January 2024, 67 pages

Link prediction is used to uncover missing links in complex networks or to predict new links that may form in the future. Traditional methods proposed for link prediction are based on measures of the similarity between two nodes given the general structure of dynamic networks. Changing the nodes' interaction over time is insufficient for the requirements of link prediction processes. This situation has encouraged researchers to find new prediction methods that can make decisions in complex networks according to the dynamic structure of the network. In this thesis, local similarity index and machine learning techniques have been utilized for author-publisher, and author-author networks created from publications in Turkish literature, and possible collaborations based on predicted. In the data set, publications are divided into time periods according to their publication years, and it has been observed how the created networks will develop in the coming years. The success of the link prediction process

has been measured by the AUC metric. The results of experimental studies have shown that the hybrid approach consisting of traditional neighborhood-based similarity index methods and machine learning methods is more successful than similarity index methods.

Keywords: Complex Networks, Link Prediction, Local Similarity Index, Machine Learning.

Science Code: 92429

ÖZET

Yüksek Lisans Tezi

TÜRKÇE YAYINLARDA ÖNERİ SİSTEMLERİ İÇİN KARMAŞIK AĞ TABANLI BAĞLANTI TAHMİNİNE HİBRİT YAKLAŞIM

Ali Asghar Fahad FAHAD

Karabük Üniversitesi
Lisansüstü Eğitim Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı

Tez Danışmanı:
Dr. Öğr. Üyesi Emrah ÖZKAYNAK
Ocak 2024, 67 sayfa

Bağlantı tahmini, karmaşık ağlardaki eksik bağlantıları ortaya çıkarmak veya gelecekte oluşabilecek yeni bağlantıları tahmin etmek için kullanılır. Bağlantı tahmini için önerilen geleneksel yöntemler, dinamik ağların genel yapısı göz önüne alındığında iki düğüm arasındaki benzerliğin ölçümlerine dayanmaktadır. Düğümlerin etkileşiminin zaman içinde değişmesi, bağlantı tahmin süreçlerinin gereksinimleri açısından yeterli değildir. Bu durum araştırmacıları karmaşık ağlarda ağın dinamik yapısına göre karar verebilecek yeni tahmin yöntemleri bulmaya teşvik etmiştir. Bu tezde, yerel benzerlik indeksi ve makine öğrenmesi teknikleri, Türk edebiyatındaki yayınlardan oluşturulan yazar-yayıncı ve yazar-yazar ağlarına uygulanmış ve olası işbirliklerine dayalı olarak tahminlerde bulunulmuştur. Veri setinde yayınlar yayınlanma yıllarına göre zaman periyotlarına ayrılarak oluşturulan ağların gelecek yıllarda nasıl gelişeceği gözlemlenmiştir. Bağlantı tahmin sürecinin başarısı AUC metriği ile ölçülmüştür. Deneysel çalışmaların sonuçları, geleneksel komşuluk tabanlı

benzerlik indeksi yöntemleri ve makine öğrenmesi yöntemlerinden oluşan hibrit yaklaşımın benzerlik indeksi yöntemlerine göre daha başarılı olduğunu göstermiştir.

Anahtar Kelimeler : Karmaşık Ağ, Bağlantı Tahmini, Yerel Benzerlik İndeksi, Makine Öğrenmesi.

Bilim Kodu **:** 92429

ACKNOWLEDGMENT

I am incredibly appreciative to Allah for giving me the chance to pursue my goal of earning a master's degree at one of the most prestigious colleges in Turkey. Over the past two years, the voyage has been an incredible experience filled with priceless memories and enduring feelings.

I extend my deepest gratitude to my advisor, Dr. Emrah ÖZKAYNAK, whose exceptional support and significant contributions were instrumental in the completion of this thesis.

To my dear Mother and Father, they have been my unwavering pillars of support. Thank them for believing in me and encouraging me to persevere through all the challenges I faced.

I dedicate this achievement to my parents, my homeland, my advisor, and my friends, whose unwavering presence provided the courage and support necessary to see this endeavor through to fruition.

CONTENTS

Page

LIST OF FIGURES

Page

Page

LIST OF TABLES

Page

INDEX OF ABBREVIATIONS

AA : Adamic-Adar
ADS : Adamic-Adar Astrophysics Data System
CN : Common Neighbor
CIEP : Combined Influence and Effective Path
DT : Decision Tree
GB : Gradient Boosting
JC : Jaccard Coefficient
HINs : Heterogeneous Information Networks
KNN : K- Nearest Neighbor
LP : Link Prediction
LR : Logistic Regression
LSI : Local Similarity Index
LGBM : Light Gradient Boosting Machine
LDA : Linear Discriminant Analysis
IDF : Inverse Document Frequency
IS : Information Systems
ML : Machine Learning
IDN : Information-Defined Network
NLP : Natural Language Processing
PA : Preferential Attachment
RA : Resource Allocation
RF : Random Forest
SI : Sorenson Index
SNA : Social Network Analysis
SVM : Support Vector Machine
TF-IDF : Term Frequency-Inverse Document Frequency
TF : Term Frequency
TP : True Positive

TN : True Negative

FN : False Negative

FP : False Positive

ROC : Receiver Operating Characteristic

TPR : True Positive Rate

FPR : False Positive Rate

AUC : Area Under the Curve

PART 1

INTRODUCTION

Link Prediction (LP) is utilized in the analysis of complex networks receiving an increased interest in the collaboration scientific literature field because link prediction's goal is to discover underlying links, it serves as a compass assisting researchers through the wide web of literature relationships. Analysis of these links especially in an environment of scientific collaboration reveals a view of complex relationships [1]. However, Communications not only define the past and present but also provide a roadmap for the future. This takes significance across domains such as social media, global trade, air transport, social interactions, and scientific collaborations [2].

Moreover, the world of social networking of scientific collaborations is becoming more complex as literature referencing networks, which track how scholarly books cite each other. This challenges researchers to find real solutions to issues in this field, such as linking authors and publishers based on years and titles and forming a so-called literature network [3]. Notably, driven by multifaceted research interests, the authors Search for collaboration with diverse groups of co-authors in specific fields, highlighting the need for cooperation to ensure the longevity of their scientific work [4].

Furthermore, predicting connections in this complex network can be achieved by examining similarities between authors and their papers, titles, publishers, years, and publications [5]. Among these metrics, this thesis relied on methods of local similarity index a type of similarity-based metric that is the most straightforward and widely adopted. They allow the researchers who are interested in the LP field to detect common features between nodes which in turn creates a pattern. The latter aids the

researchers in having a deeper understanding of the interactions at hand which makes the prediction process reliable and based on data [6].

However, they should be noted that these algorithms may fail to achieve high accuracy when applied to real-world networks [7]. The goal of network science has embraced the power of Machine Learning (ML) algorithms to significantly enhance predictive power within literary and educational networks, in response to these technical challenges [8]. These algorithms make it possible to leverage data features to make accurate predictions about the relationships between authors and publishers in scholarly and literary publishing [9].The hybrid model aims to improve and raise predictive capabilities, ultimately allowing more accurate predictions within the network of the scientific literature [10].

This thesis focuses on a hybrid approach to increasing the success of LP within a complex network structure. The hybrid approach provides the mere implementation of LP metrics and leverages the power of classification algorithms fixed in supervised ML. The proposed approach uses traditional LSI methods and ML methods in link prediction. Data regarding the networks created, publishers, authors and literary works published in Turkey has been used to apply the model in order to increase cooperation, advisory LP operations.

1.1. MOTIVATION

The abundance of author, publication and publisher data in networks increases the size of the network and makes LP difficult. This creates an important motivation for researchers to find appropriate algorithmic solutions to promote possible collaborations in a complex network. As a result, it encourages researchers to develop more powerful LP methods to reveal strong and weak connections in complex networks consisting of excessive amounts of data. Therefore, the success of LP depends on the combined use of efficient algorithms. In order to increase the collaboration of authors, it is necessary to reveal hidden connections in the networks that form them. The main motivation behind the study is to increase the success of

traditional LP methods together with ML methods for LP that can be used in the recommendation system in collaboration networks.

1.2. PROBLEM STATEMENT

The scientific collaboration network has an exponential growth of published articles and books along with the complexity of the collaboration. The implementation of complex network-based LP in Turkish literary publications brings with it many notable challenges. However, a strong challenge in this thesis is the use of a LP network for the first time in Turkish literature it can rise to a complex network of literature interrelations. The main challenge in this context is to effectively uncover and exploit hidden connections within scientific collaboration networks. The literary scene is characterized by a complex network of contacts that includes authors, publications, and research topics. However, the interaction between these elements creates a network with complex relations that need to be decoded. The challenge is to connect authors and publishers based on factors such as years of publication and titles. Recent research endeavors have made significant strides in this direction, but they have not yet achieved satisfied levels. Moreover, there is need to enhance the accuracy of predictive models in scientific collaboration network publishing by harnessing the power of integrating various algorithms and methodologies to construct the most effective and reliable models.

The main problem is the design and development of an innovative and sophisticated model that goes beyond the limits of traditional approaches, ultimately providing authors and publishers with a dependable model for predicting and understanding relations, collaborations, and trends within the academic publishing.

1.3. RESEARCH OBJECTIVE

The aim of this thesis is to present a model that can be used in recommendation systems by developing and applying a versatile hybrid model consisting of LSI and supervised ML to solve the difficulty of the LP problem in complex networks. The main purpose can be summarized as follows:

- The model offers an approach that can make recommending predictions in networks created from many years of author, publication and publisher data.
- In creating the model, a hybrid approach was adopted by strengthening of the LSI methods with ML.

1.4. CONTRIBUTION

Contribution of this thesis study makes a significant impact on NLP to extract subjects' insights from literature book details, paving the way for a deeper understanding of underlying patterns within the data. Besides that, the development of a hybrid model combining LSI and ML, represents an approach to constructing and analyzing complex collaboration networks. This hybrid model not only advances the field of LP but also provides a significant and versatile tool for researchers to navigate scientific literature network landscapes enhance collaboration and make informed decisions.

1.5. THESIS SCOPE

The thesis scope investigates the LP algorithms in social networks including the LSI, which encompasses local, global, and quasi-local index. The study includes into specific local index such as Common Neighbor (CN), Jaccard Attachment (JA), Preferential Attachment (PA), Adamic-Adar (AA), Resource Allocation (RA), and Sorenson Index (SI) as well. To bolster the effectiveness of LP, the thesis extends its reach to the realm of supervised ML, incorporating a diverse set of classification algorithms including Random Forest (RF), Decision Tree (DT), Support Vector Machine (SVM), Gradient Boosting (GB), K-nearest neighbors (KNN), Logistic Regression (LR) and Light Gradient-Boosting Machine (LGBM). Furthermore, it has discussed its range of vision with the integration of an enhanced model between algorithms LSI and algorithm ML to create networks in the field LP.

1.6. STRUCTURE OF THESIS

In first section functions as an introduction to the thesis, providing contextual information, outlining the inspiration behind the study, and clearly defining the identified problem. Additionally, it delineates the objectives and goals of the research. In second section entails a comprehensive review of the latest research, highlighting the integration of the current study within the broader academic landscape. It emphasizes the originality, significance, and potential impact of the research within this specific field.

In third section provides an extensive exploration of graph complex system networks, delving into the fundamental concepts that underpin the study of the structure and dynamics of interconnected systems.

In fourth section offers an extensive overview of the stages involved in constructing a LP network model within the Turkish literary domain. It begins with a discussion on involving the utilization of a hybrid model integrating the LSI and ML. Various studies conducted throughout the implementation phase are also outlined in this section.

In fifth section, the thesis includes into the complexity of the dataset, explain its structure, processing, analysis, and preparation for the creation of the model.

In sixth section presents the comprehensive results obtained from the complete methodology employed for constructing the hybrid model for LP within the Turkish literary field. It involves a detailed discussion and comparison of the results derived from the proposed approach with those obtained from hybrid classification techniques.

In concluding section provides a succinct summary of the results attained from the research study, emphasizing key findings and contributions. It also offers insights into potential avenues for future research endeavors within the domain of LP.

PART 2

LITERATURE REVIEW

This section of the thesis study is dedicated to crafting an approach for constructing LP networks in the realm of Turkish literature, leveraging the integration of LSI associations and ML techniques. Furthermore, the thesis work provides an extensive literature review that encapsulates studies and initiatives pertinent to the hybrid approach of LP networks.

In this study Muhammad et al. explored similarity-based approaches, providing understanding of the intricate mechanisms involved in discerning shared characteristics among pairs of nodes, as illustrated in Figure 2.1. The research further described these similarity-based approaches into distinct categories, including local, global, and hybrid methods, with a specific focus on local similarity approaches that are complex subcategorized into normalized and unnormalized methodologies. This exploration promises valuable insights into applying similarity-based methods and their classifications [11].

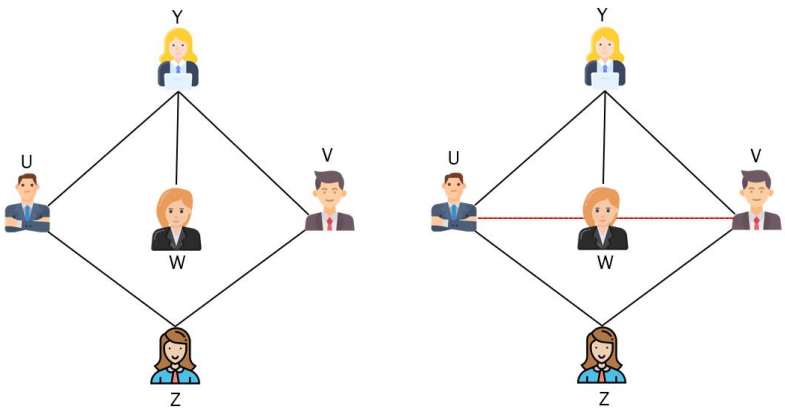


Figure 2.1. Graph at time t graph of time t+n.

In 2022, Shivaansh et al. introduced an innovative perspective on LP in social networks, utilizing ego networks to assess edge strength. Their approach incorporated diverse topological features from ego network layers, enhancing local similarity-based LP algorithms like CN and PA, ultimately facilitating a comprehensive evaluation of ego strength across all edges [12]. Meanwhile, in 2021, Chen et al. conducted a study on partner selection within an interorganizational patent cooperation network, employing LP approaches. Their research focused on collecting cooperative patents filed by organizations in China from 2007 to 2015. The study presented eight commonly utilized LP approaches in social networks. Notably, they emphasized the effective application of the CN index in selecting unfamiliar partners among organizations involved in patent cooperation [13]. In 2022, Aparna et al. conducted a study employing LP techniques in social networks such as Facebook, as well as in E-business platforms like Zomato and Amazon. Their research focused on predict potential associations between users that may not currently exist but have a likelihood of forming in the future as shown in Figure 2.2, the study successfully employed Proximity-Based algorithms, specifically the JC and RA demonstrated effective LP capabilities in the context of social networks and E-business platforms [14].

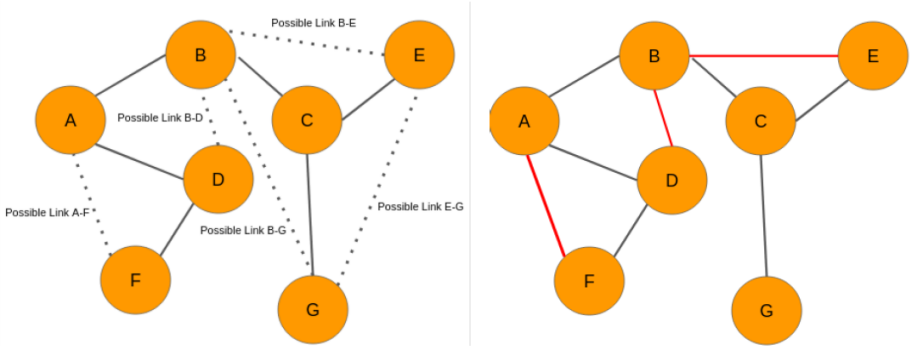


Figure 2.2. Link prediction graph [14].

In 2015, a study utilized Wikipedia's editing records to construct a collaboration graph among editors, focusing on predicting social link formations. The AA predictor emerged as the most accurate in predicting link strength between editors' common neighbors [15]. However, Shiping and Yong et al. introduced a hybrid approach in their work, incorporating time-varying link weight information in social networks for link prediction. Their findings showcased the superior performance of the tw-AA, tw-CN,

and tw-JC algorithms, emphasizing the adaptability of time-varied weight in link prediction [16]. In their 2022 study, SONG et al. applied co-authorship data extracted from papers published in 43 statistical journals spanning the years 2001 to 2018, skillfully constructing collaboration networks. They created training and testing networks based on paper timestamps and developed a classification dataset for LP. Utilizing 20 similarity indexes calculated from the training network, their research emerged recommendations for researchers based on collaborative dynamics [17]. In 2019, Tian et al. introduced the Combined Influence and Effective Path (CIEP) index, superior traditional indices in accuracy without added computational complexity across 12 benchmark datasets, marking CIEP as a standout performer in LP [18]. Meanwhile, Wang et al. in 2023 explored the fusion of three fundamental citation relationships (direct citation, co-citation, and coupling) as a predictive framework for potential collaborations using citation information. The study strategically utilized diverse LP indicators, such as CN, AA, PA, and RA, to forecast author collaborations and research topics. This approach effectively enabled the identification of promising scholarly collaborations and research synergies within the academic landscape [19].

In 2017, Li et al. handled the disappearing LP challenge in scientific collaboration networks by employing network embedding, demonstrating that TDL2vec surpasses established methods like CN, JC, PA, AA, RA, and SI, enhancing predictive capabilities [20]. In 2020, Michał et al. presented a comprehensive examination of network complexes, utilizing a blend of global, local, and semi-local algorithms to forecast forthcoming co-authorship relations among researchers at the University of Warsaw. This thorough analysis significantly contributes to an enriched comprehension of the complex network dynamics within the academic environment, shedding light on the evolving landscape of collaborative relationships among researchers [21]. The research conducted by Zhou and Lu et al. explores a basic framework for LP based on node similarity by comparing the performance of nine established local similarity measures across six real networks. The results emphasized the superior performance of the CN measure, closely followed by the AA. Additionally, the study proposed a novel similarity measure inspired by the resource allocation process observed within networks, demonstrating higher prediction accuracy compared to the common neighbor's method [22]. In 2018, Javari et al.

introduced a novel probabilistic approach designed for LP to adapt to the sparsity of data in social networks. The utilization of algorithms for link label modeling based on the local and global. This contribution signifies a notable stride in the ongoing development of LP methodologies, offering valuable insights and tools for navigating the intricacies of sparse network environments [23].

Gimenes et al. introduced a methodology harnessing various topological metrics to offer prompt and prospective link recommendations. Utilizing a blend of these metrics as input for binary classification algorithms, their study singled out the RF as exceptionally effective in identifying potential collaborators, similar interest groups, competitive research efforts, and related works. This emphasized the adaptability and efficiency of RF across diverse networks, affirming its main role in furnishing understandable link suggestions within constrained time frames [24]. In 2021, Mihaljević et al. discussed traditional methods depend on publication metadata, incorporating affiliations, email data, co-authors, and scholarly themes. Their introduction of a semi-supervised algorithm for authorship disambiguation within the Astrophysics Data System (ADS), trained on authorship pairs and subsequent graph clustering, demonstrated the RF algorithm's superior performance in solve authorships. Their evaluation on 39 manually labeled author blocks, covering 9545 authorships from 562 author profiles, strengthened the effectiveness of RF in this nuanced task [25]. In 2020, Ching Chen et al. determined three widely-used datasets with diverse variables, namely the Bank Marketing, Car Evaluation Database, and Human Activity Recognition Using Smartphones. The study systematically evaluated the performance of several classification models, including RF, SVM, KNN, and Linear Discriminant Analysis (LDA). Notably, the results consistently highlighted RF as the most effective model across all experimental groups, emphasizing its robust performance in diverse and complex network data [26]. In 2013, Hurtado, a novel approach is proposed to tackle the data imbalance challenge in social link prediction. The method combines k-nearest neighbor sampling in 2013, Hurtado, a novel approach is proposed to tackle the data instability challenge in social link prediction. The method combines k-nearest neighbor sampling with random sampling, aiming to address the imbalance issue effectively. Using ML algorithms, this innovative approach enhances the accuracy of the LP model, particularly when applied to networks characterized by imbalanced

social connections [27]. In 2020, Kumari et al. investigated into the application of Social Network Analysis (SNA) in various domains, emphasizing its two-fold approach, namely content-based and structure-based analysis. The study employed ML techniques, with the SVM algorithm emerging as the top performer in predicting future link establishment in social networks, emphasizing its significance in understanding and analyzing complex social structures [28]. In 2020, Kumari et al. discussed the emerging domain of link prediction, integrating path-based similarity measures into supervised ML models like KNN, RF, and DT. The SVM emerged as the top performer, solidifying its efficacy in predictive modeling [29]. In 2019, Koker realized a study used on LP algorithms within the context of social networks, employing diverse ML techniques. The investigation utilized LR, SVM, and RF models for link prediction, with the findings revealing that the SVM demonstrated the most effective performance in forecasting future collaborations within the Information Systems (IS) education community. Additionally, the incorporation of the GB technique further augmented the predictive capabilities of the models, emphasizing the effectiveness of these advanced methodologies in advancing social network analysis [30]. In 2022, Kumar et al. introduced a novel approach to LP by merging centralities to create a comprehensive feature set. The study also integrated various similarity measures including local, quasi-local, and global structures alongside ML models, achieving optimal results with the LGBM for accurate and efficient LP [31].

The literature also includes research on the development of hybrid models for link prediction. In 2023, Bath and Amounts presented a hybrid approach between supervised ML and similarity embedding methods, incorporating techniques such as KNN-Deep Walk, Node2Vec-SVM, PA-GB, RA-XGBoost, CN-AdaBoost and JC-RF, among others, to predict future links in a social network [32]. Similarly, Jeong et al.'s 2019 study introduced a LP framework in information networks, employing novel features and a prediction model that leveraged structural characteristics, meta-paths, and path counts to improve LP accuracy. The study implemented a hybrid model integrating similarity indices and classification ML, including combinations such as AA-RF, CN-RF, and Random Walk - RF, thus highlighting the significance of incorporating both structural and predictive elements in LP tasks [33]. Additionally, Aghabozorgi and Khayyambashi's research delved into imbalanced social network

data, proposing a hybrid strategy that combined local similarity indices, supervised learning, and Linear Discriminant Analysis within LGBM, illustrating the potential of hybrid models in navigating complex data structures and significantly boosting predictive accuracy within social networks [34].

In 2022, Krenn et al. conducted a comprehensive study on the construction of dynamic semantic networks using papers from arXiv in computer science categories related to artificial intelligence, spanning the years 1992 to 2020. The study employed NLP methods, particularly the RAKE algorithm, to extract candidate concepts from the titles and abstracts of research papers within the literature network thereby contributing to a deeper understanding of the semantic relationships within the field [35]. In contrast, Hariri's 2023 study succinctly defined NLP as the AI domain focusing on computer-human language interaction. The study highlighted ChatGPT's success due to its language understanding and generation capabilities, emphasizing NLP's role in improving language-based interactions [36]. This section provided a comprehensive exploration through literature reviews, covering diverse aspects like local similarity indices, ML in hybrid models, and NLP's role in topic extraction. Together, these reviews offer insights into methodologies shaping LP networks, illuminating their evolving landscape. Table 2.1 supplements with brief descriptions of the discussed studies.

Table 2.1. The latest research on link prediction.

Author	Year	Method	Objective	Category	Accuracy
Bath and Amounts [32]	2023	SVM+PA, RA+KN, CN+RF, JC+G, more	The study aims to discuss LP social network analysis. Was proposed as a hybrid approach between supervised ML classifier LP and similarity embedding methods.	Hybrid Model	0.92% 0.88% 0.92% 0.91%
Hariri's [36]	2023	NLP	The primary objective is NLP as an AI field focusing on human-computer language interaction. This study specifically aims to underscore ChatGPT's critical position and achievements in NLP, driven by its advanced capabilities in understanding and generating human language.	NLP	-
Shivaansh et al. [12]	2022	CN, PA, RA, JC, ELP	The study aims to propose and evaluate new topological features based on ego network layers to enhance the performance of LSI-based LP algorithms	LSI	0.94% 0.96% 0.95% 0.84% 0.92%
Aparna et al. [14]	2022	RA, JC, RA+JC	the objective LP in social networks like Facebook, and E-business organizations Zomato and Amazon.	LSI	0.95% 0.91% 0.98%
SONG et al. [17]	2022	CN, JC, PA, RWR	The study goal to construct collaboration networks based on the co-authorship information of the papers published in 43 journals from 2001 to 2018.	LSI	0.64% 0.64% 0.61% 0.82%
Kumari et al. [29]	2022	SVM, KNN, DT, LGBM	The objective of the study is to introduce a new approach to making LPs using a well-curated combination of centralities to develop a set of features that can be utilized to make predictions.	ML	0.90% 0.90% 0.65% 0.97%
Krenn et al. [36]	2022	NLP	The objective of the study is to construct a dynamic semantic network using research papers from arXiv within computer science categories related to artificial intelligence, spanning the years 1992 to 2020. This study utilizes NLP methods, titles, and abstracts, enabling the creation of a comprehensive literature network.	NLP	-
Mihaljević et al. [25]	2021	DT, RF, Rule-based, Hist-GBDT	the objective discussed usual approaches relying on publications' metadata such as affiliations, email addresses, co-authors, or scholarly topics.	ML	0.96% 0.97% 0.97% 0.95%
Jeong et al. [33]	2019	AA+RF, CN+RF, Random walk +RF	The primary objective is to enhance the accuracy of LP in Heterogeneous Information networks (HINs) through the introduction of novel features and the development of a hybrid prediction model.	Hybrid Model	0.98% 0.99% 0.99%

PART 3

GRAPH

This section provides a discussion on network science, an interdisciplinary field that investigates into the structure and dynamics of interconnected systems. Within this expansive domain, various subfields intersect to offer profound insights into the complex relationships and patterns of these systems.

3.1. GRAPH THEORY

Graph theory is a wealthy system encompassing of powerful theorems with broad applications [37]. Configurations of nodes and connections find applications across diverse fields such as Computer Science, Networking, Social Media networks, ML, Collaborative Filtering, and Transportation Networks [38]. Furthermore, they use as a representation of real interactions found in ecosystems, sociological relations, databases, or the flow of control within a computer program. These configurations are modeled using connection structures known as graphs, comprised of two sets of nodes and edges with an relation between them [39]. In graph theory, vertices typically denote objects or entities, while edges represent the connections between them [40]. Nodes and edges may possess additional attributes such as color, weight, title, or any other relevant information, rendering graph theory a versatile and powerful tool in various domains [41].

3.1.1. Undirected Graph

An undirected graph, within the context of graph theory, is a mathematical representation of a set of nodes and their binary connections, wherein the edges have no specific direction as shown in Figure 3.1. If there exists an edge between nodes A and B, it implies a mutual connection, with A linked to B and conversely [42].

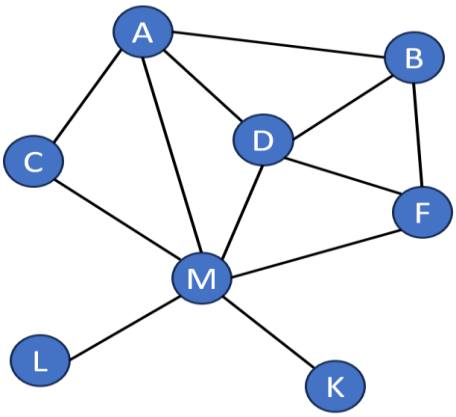


Figure 3.1. Undirected graph.

3.1.2. Directed Graph

A directed graph, also known as a digraph, is a fundamental data structure in graph theory where edges possess direction, indicating one-way relationships from one node to another in the showed graph in Figure 3.2. Thus, if there exists an edge from node A to node B, it does not necessarily imply the existence of an edge from node B to node A. In other words, the relationship is not bidirectional [43].

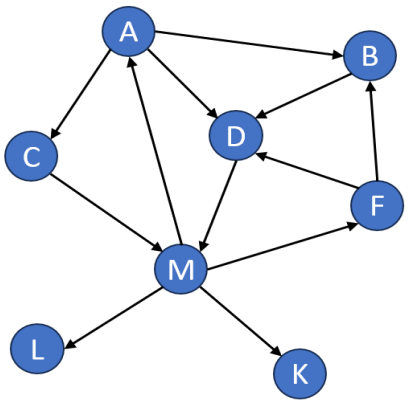


Figure 3.2. Directed graph.

3.1.3. Weighted Graph

A weighted graph is a type of graph where every edge is set as a numerical value or weight in Figure 3.3. These weights represent various attributes associated with the

connections between vertices. Weighted graphs find common usage in various real-world applications [44].

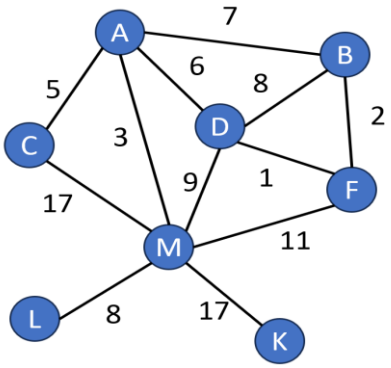


Figure 3.3. Weighted graph.

3.1.4. Bipartite Graph

A bipartite graph is a type of graph that can be divided into two distinct sets of vertices or nodes such that no edges exist between vertices within the same set as in Figure 3.4. However, edges can connect vertices from one set to vertices in the other set. This characteristic makes bipartite graphs useful for modeling and solving problems related to relationships between two different types of entities [45].

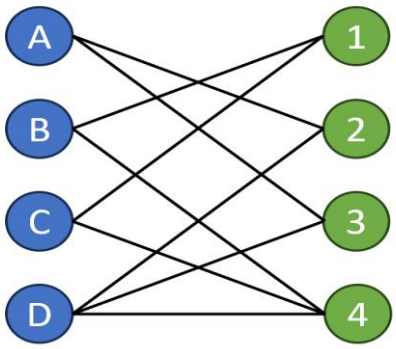


Figure 3.4. Bipartite graph.

3.1.5. Complete Graph

A complete graph is a specific type of simple graph in which every pair of distinct vertices is connected by an edge. Moreover, this characteristic makes a complete graph

fully connected, with an edge present between every pair of vertices as in Figure 3.5. Complete graphs are often showed by the symbol "K" followed by the number of vertices, such as "K3" for a complete graph with 3 vertices or "K6" for a complete graph with 6 vertices [46].

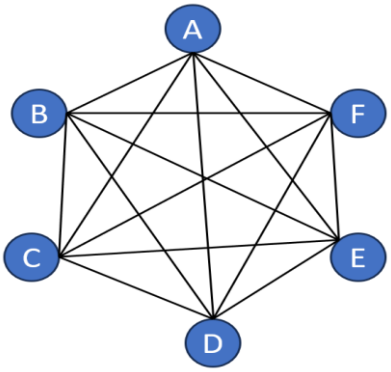


Figure 3.5. Complete graph.

Furthermore, graph theory encompasses various other types of graphs and concepts. The choice of which type of graph to utilize depends on the specific problem in hand, and graph theory offers a rich array of techniques and theories to address a wide range of applications across diverse fields.

3.2. COMPLEX NETWORK

Complex networks are systems that represent structures that are directly or indirectly connected to each other. In contrast to earlier models that treated such systems as random graphs, contemporary understanding recognizes that real networks follow robust organizing principles [47]. Complex networks have gained substantial importance in recent years owing to their diverse practical applications in various domains, encompassing social networks, transportation networks, biological networks, citation networks, the semantic web, and communication networks [48]. What sets these networks apart are their distinctive properties, main among them being a substantial number of linked nodes [49]. These nodes can represent a wide array of entities, involving from individuals and websites to molecules and beyond, contingent on the specific context of the network, as illustrated in Figure 3.6. This diversity of

applications and the underlying principles governing complex networks make them a fascinating and crucial area of study in modern science and technology.

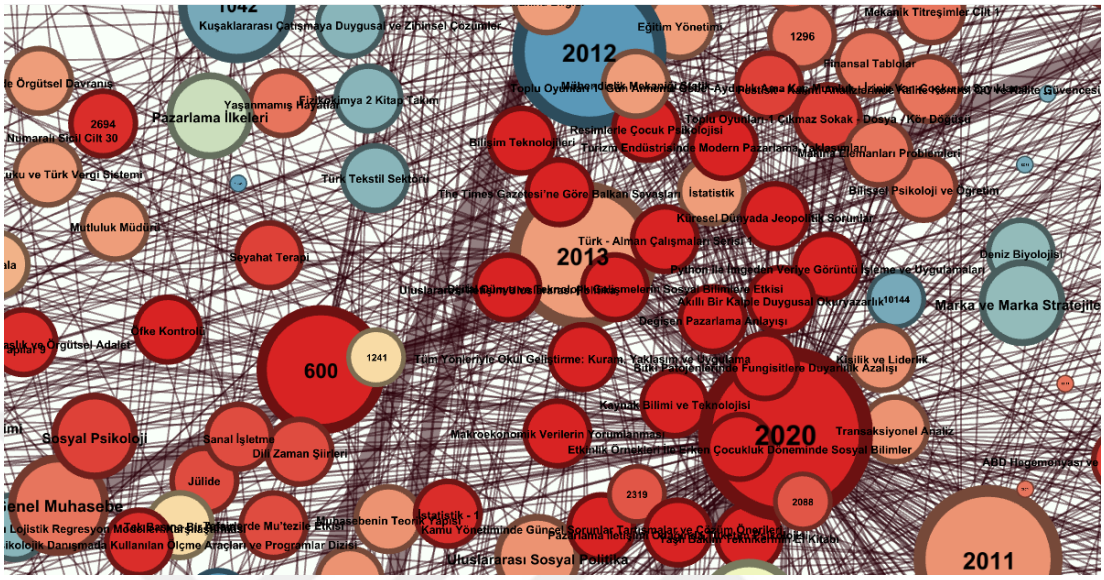


Figure 3.6. Complex network represents author, publisher and year.

In this regard, the complex systems often exhibit various dynamic behaviors depending on multiple factors such as interaction structures, parameter configurations. By analyzing their characteristics, researchers can gain valuable insights into the working of complex systems and devise strategies to optimize their performance. As our world becomes increasingly interconnected, understanding complex networks will continue to grow in importance, opening up new way for innovation and discovery.

Network analysis is a technique employed to search the structural relations between people or organizations [50]. That focuses on the analysis of complex systems represented as networks or graphs. A complex network consists of a structure that encompasses individuals or organizations and the connections between them, with the nature of these relationships varying based on the network type, spanning dimensions such as friendship, collaboration, and scientific associations [51]. In this context, a network is created of nodes connected by edges and these nodes and edges can represent various entities and their relationships [52].

In 2023, a recent study by Setiadi et al. the objective to identify communities or groups of users and borrowed books based on book borrowing records employed complex network analysis. The investigation provided valuable insights into the structure of the book borrowing network, leading to the identification of borrowing communities and book communities. This information can inform a more effective book development policy strategy and targeted book recommendations for specific users [53]. In 2022, Chira et al. gave a talk introducing complex network science, which is about studying how things are connected. Researchers talked about tasks like finding important points in the connections, discovering groups of connected points, and understanding patterns in the connections. In their paper, they briefly explained these tasks and mentioned that there's a lack of research on how things change over time in the financial sector. They suggest that using network science tools can help us understand these changes and patterns in a simple and effective way [54]. Zang et al. (2022) presented a network reliability analysis method based on complex network theory, demonstrating its feasibility in a train control system. The study analyzed the system's topological properties and identified weak nodes and edges, facilitating efforts to improve the system's reliability [55]. In 2021, Jia et al. focused on the significant role of complex networks in advancing various fields like biology, chemistry, social science, computer, and communication engineering. While the use of complex networks for studying communication networks, particularly in designing efficient routing strategies and robust communication networks, has gained popularity, there's still unused potential in applying these principles to explore networks in diverse disciplines beyond telecommunications. To address this problem, their paper introduces an Information-Defined Network (IDN) framework, offering a conceptualization where a complex network can be represented as a communication network intricately linked with multiple intelligent agents [56].

3.3. LINK PREDICTION

LP striving to uncover missing connections or forecast potential links in a network [57]. Particularly fascinating in social complex network analysis it the focus of LP revolves around anticipating the probability or likelihood of a connection between two nodes in a network [58]. The structure drawn of all algorithms used in different types

of link prediction complex networks is depicted in Figure 3.7 [59]. This predictive process is find missing links in complex networks and anticipating the creation or dissolution of links in future networks, this predictive process plays a pivotal role in advancing our understanding of the intricate dynamics and evolution of social complex networks, then making a substantial contribution to their exploration and analysis [60].

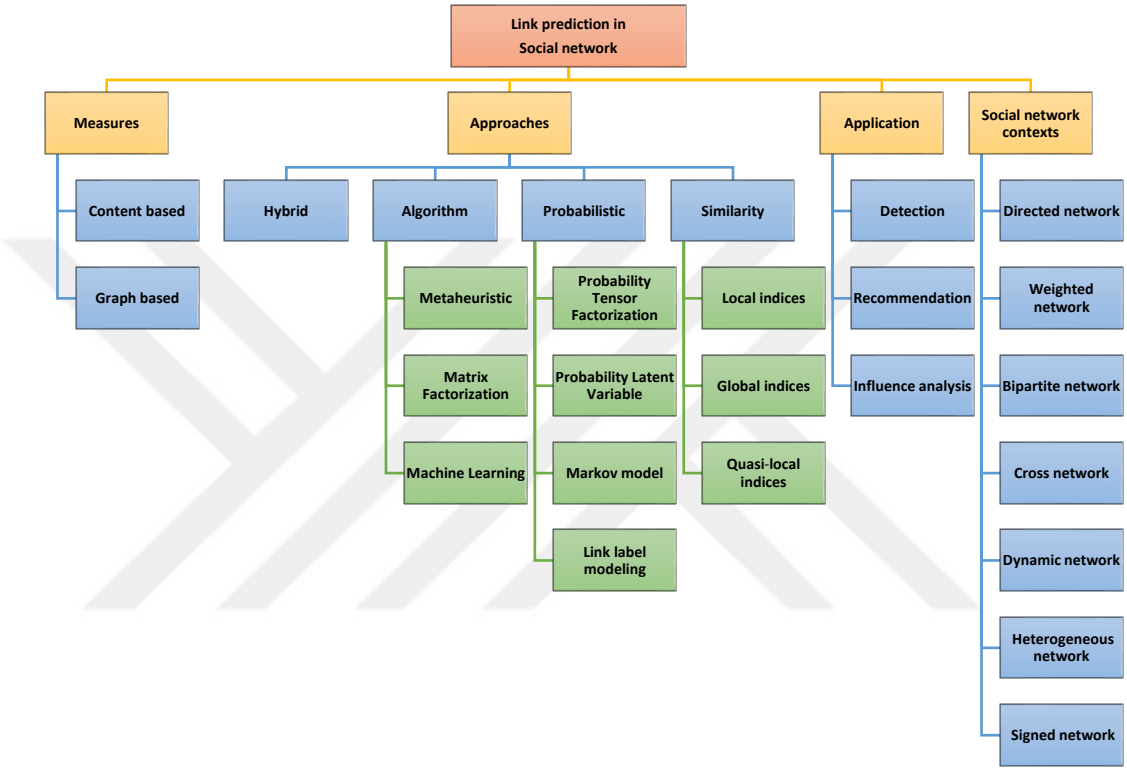


Figure 3.7. Structure link prediction network [59].

3.3.1. Similarity Approaches

Similarity approaches form the foundation of LP algorithms, taking advantage of notion of node similarity to predict links in a network. These approaches involve evaluating the resemblance between nodes based on essential attributes, utilizing similarity local index, global index, and quasi-local index as shown in Figure 3.8. By measuring the similarity between nodes, these methods effectively forecast potential connections within the network, facilitating a comprehensive understanding of its underlying structure and dynamics [61].

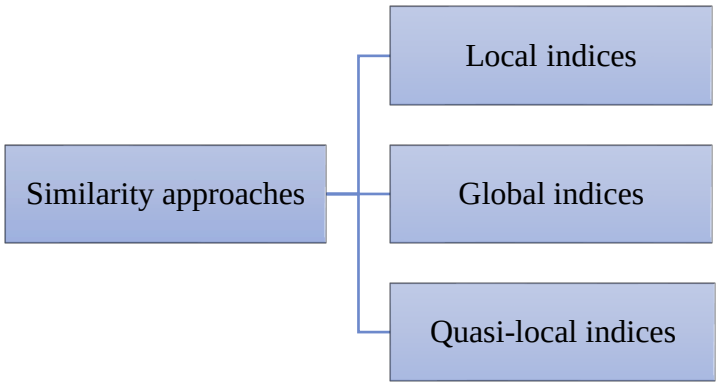


Figure 3.8. Taxonomy similarity approaches.

3.3.2. Probabilistic Approaches

Probabilistic approaches offer a solution to the LP challenge by constructing a statistical probability model aligned with the network's structure [62]. This algorithm defined by parameters calculates a mathematical statistic to generate probability values for each pair of nodes. These probability values are then categorized based on the hypothesis that higher values indicate a greater likelihood of link formation between the node pairs. Notably, the algorithm efficiency stems from extracting simple statistics from the network. Moreover, its adaptability is evident as it can easily extend to operate on networks featuring more than two types of relations [63]. In this thesis, these approaches are categorized into four groups: the Probability Tensor Factorization model, Probability Latent Variable model, Markov model, and Link Label Modeling, as illustrated in Figure 3.9.

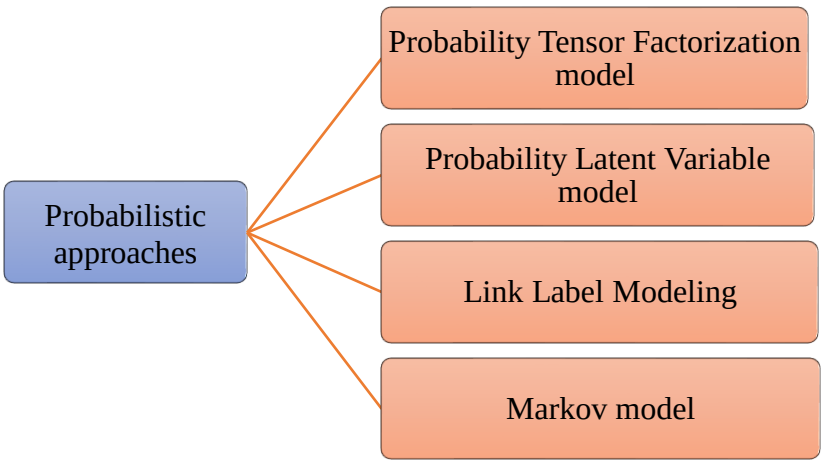


Figure 3.9. Taxonomy probabilistic approaches.

3.3.3. Algorithmic Approaches

Algorithmic approaches represent a high-performance strategy extensively explored by researchers within the realm of LP literature [64]. Researchers and data scientists have employed a multitude of approaches to tackle this challenge, and among them, algorithmic methods have garnered significant attention and recognition. Algorithmic approaches in LP offer a systematic and rule-based framework for uncovering latent relationships within networks, capitalizing on principles of computation, data processing, and automated prediction tasks [65]. In contrast to similarity-based and probabilistic approaches, which primarily depend on observed patterns and statistical inference, algorithmic methods delve deeply into the intricacies of network structures [66]. This category of algorithmic approaches encompasses a variety of methods, as illustrated in Figure 3.10.

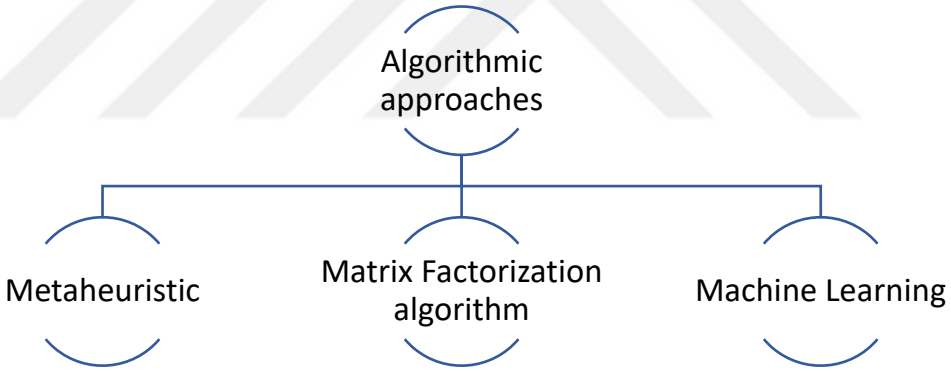


Figure 3.10. Represent algorithmic approaches in link prediction.

PART 4

METHODOLOGY

In the methodology section, the methods employed in the thesis are explained in term of Neighborhood-based LSI methods and ML methods used in experimental studies are explained in this section.

4.1. SIMILARITY LOCAL INDEX (LSI)

LSI method serves in the analysis of correlation prediction, playing a crucial role in assessing the potential likelihood of future interactions within the network. This method has gained widespread acceptance for its efficacy in predicting correlations and encouraging a deeper understanding of complex network dynamics[67]. The research examines six distinct types of algorithms within the LSI domain as shown in Figure 4.1.

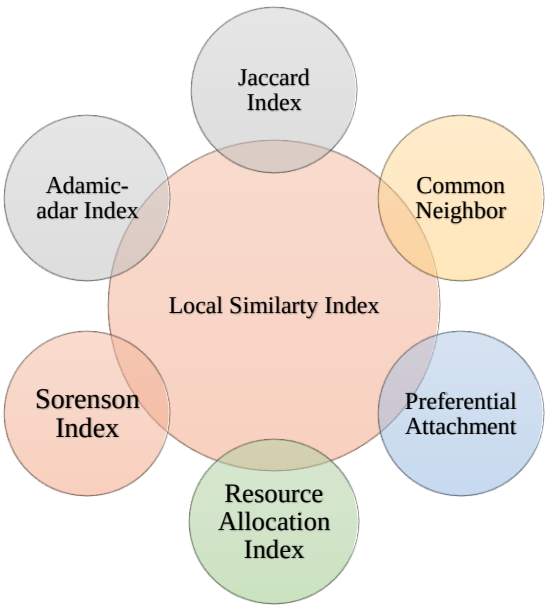


Figure 4.1. Local similarity index methods.

4.1.1. Common Neighbor (CN)

CN is a concept used in various fields, including graph theory, network analysis, and social network analysis. It refers to a measure that quantifies the similarity or connection strength between two nodes in a network based on their shared neighbors or connections [68]. The equation for CN is defined as:

$$CN(x, y) = | A(x) \cap B(y) | \quad (4.1)$$

In the context of network analysis and the CN measure, A(x) and B(y) refer to the neighborhood sets of nodes u and v, respectively. These neighborhood sets represent all the nodes that are directly connected to nodes u and v in a network contributing to the computation of their common neighbor score [69].

4.1.2. Jaccard Coefficient (JC)

The Jaccard Coefficient Index, often denoted as the JC similarity coefficient is a measure of similarity between two sets. It's commonly used in various fields, including data analysis, information retrieval, and network analysis, to compare the similarity between two sets by considering their intersection and union [70]. The equation for JC is defined as:

$$S_{xy} = \frac{|N(x) \cup N(y)|}{|N(x) \cap N(y)|} \quad (4.2)$$

When calculating the similarity between a pair of nodes, JC performs the normalization process by looking at the ratio of common neighbors of two nodes to the total neighbors of the nodes in the network [71].

4.1.3. Adamic-Adar (AA)

AA index is a similarity measure commonly used in network analysis, particularly for assessing the similarity between two nodes in a network based on their common neighbors. This index serves as an extension of the Common Neighbor measure,

emphasizing the varying degrees of informativeness exhibited by shared neighbors [72]. The equation for AA is defined as:

$$AA(x,y) = \sum_{z \in N(x) \cap N(y)} \frac{1}{log|N(z)|} \quad (4.3)$$

The AA index quantifies the similarity between nodes A and B by considering the inverse of the logarithm of the degree of each common neighbor.

4.1.4. Preferential Attachment (PA)

PA index is used in large non-scalable networks. The likelihood of establishing a new link to node X relates directly to A(x). The link's chance between node x and y scales with A(x)×B(y). The link feasibility between two nodes resonates with the count of neighboring nodes in the system [73]. The equation for PA is defined as:

$$S_{xy} = A(x) \times B(y) \quad (4.4)$$

4.1.5. Resource Allocation (RA)

RA index is a similarity measure used in complex network analysis to assess the similarity or potential interactions between nodes in a network based on the allocation of a limited resource, often modeled as a flow of resources or information. The RA index measures how well two nodes can allocate a resource to each other through their common neighbors in the network [74]. The equation for RA is defined as:

$$S_{xy} = \sum_{z \in (\Gamma(x) \cap \Gamma(y))} \frac{1}{k_z} \quad (4.5)$$

RA index calculates the similarities between nodes X and Y based on their common neighbors, even though there is no connection between them. The resources that the X and Y node pair receive from each other determine their similarity ratio [75].

4.1.6. Sorenson Index (SI)

SI is a similarity measure commonly used in various fields, including data analysis, information retrieval, and network analysis. SI measures the similarity between two sets or graphs by comparing their intersection to the size of their union. In network analysis, the SI can be used to quantify the similarity between two sets of nodes, often representing the neighborhoods of two nodes in a network [76]. The equation for SI is defined as:

$$S_{xy} = \frac{2|\Gamma(x) \cap \Gamma(y)|}{k_a + k_b} \tag{4.6}$$

The calculate of SI calculates the size of the intersection of the two sets, denoted as |A ∩ B|, which represents the number of elements that are common to both sets. SI (A, B) indicates the degree of similarity between the sets Higher values indicate greater similarity [77].

4.2. MACHINE LEARNING (ML)

ML represents a decisive field dedicated to focusing on the creation of algorithms and statistical models that capable of enabling computers to learn knowledge from data, make predictions, or take actions based on the learned patterns. It has gained widespread attention due to its potential to revolutionize various domains and improve decision-making processes[78]. In the context of the thesis, ML techniques play a task role in predicting relationships between authors and publishers, contributing to a deeper understanding of the data and facilitating more precise forecasts across diverse networks. The types of structure of ML teaching are composed as in Figure 4.2.

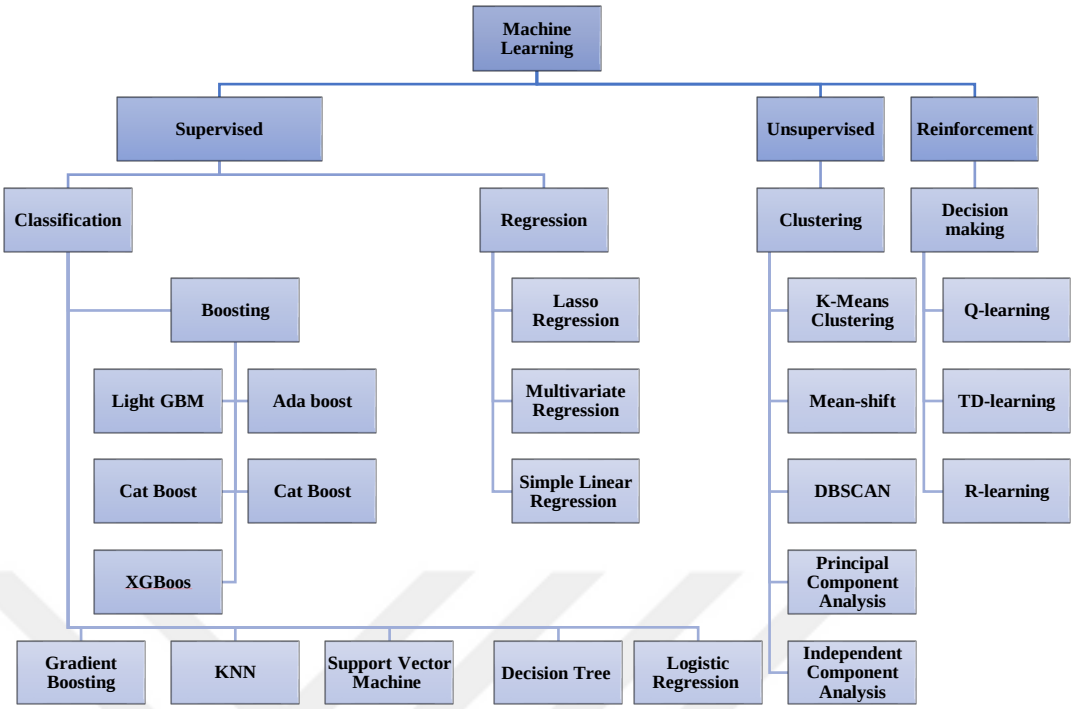


Figure 4.2. Machine learning methods.

The fundamental and extensively utilized technique in supervised ML is categorized into two primary types. The first, known as binary classification, focuses on sorting data into two distinct groups, simplifying predictions to a straightforward “yes” or “no” when, for instance, establishing the relationship between authors and publishers. The second type, multi-category classification, extends to predicting more than two outcomes, offering nuanced forecasts such as strong, medium, or weak relationships in the context of authors and publishers. Leveraging classification methods not only enhances the precision of data comprehension but also enables more detailed and nuanced forecasts across various networks [79]. The types of classification structures utilized in the thesis are depicted as shown in Figure 4.3.

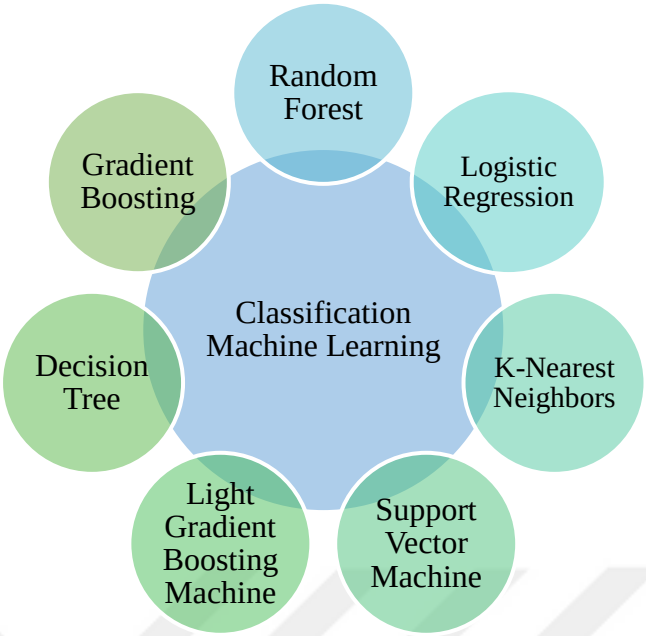


Figure 4.3. Machine learning classification.

4.2.1. Decision Tree (DT)

DT techniques have found widespread use in constructing classification and regression models, as these models closely resemble human reasoning and are easily understandable As illustrated in Figure 4.4 [80][81]. A decision tree assumes a hierarchical structure resembling an upside-down tree, comprising nodes and branches. The top node, known as the root, represents the initial decision or the starting point for the tree. The leaves of the tree symbolize the final outcomes or decisions [82]. Beginning with a root node containing all objects of the training dataset, the DT splits the root node into two disjoint subsets, namely the left child and the right child, and repeats this process recursively over the child's nodes until a specific stopping criterion is met. To split each node, a subset of attributes is chosen, generating multiple splitting criteria based on the type of attribute. Subsequently, a split evaluation measure assesses each splitting criterion at each level of the DT, aiming to select the best one at that level. When the recursive procedure halts, certain sub-trees within the DT are pruned to reduce overall complexity and enhance generalization capability[83]. Various DT approaches, including ID3, C4.5, C5, and CART, employ distinct mathematical techniques to partition the training data, catering to diverse networks and offering flexibility in solving various real-world problems [84],[85]. In the proposed thesis, a

DT algorithm, such as CART or C4.5, is utilized to develop a model for LP. The objective is to train the DT to classify node pairs as either having a link or not having a link, thereby addressing LP prevalent in numerous fields and applications.

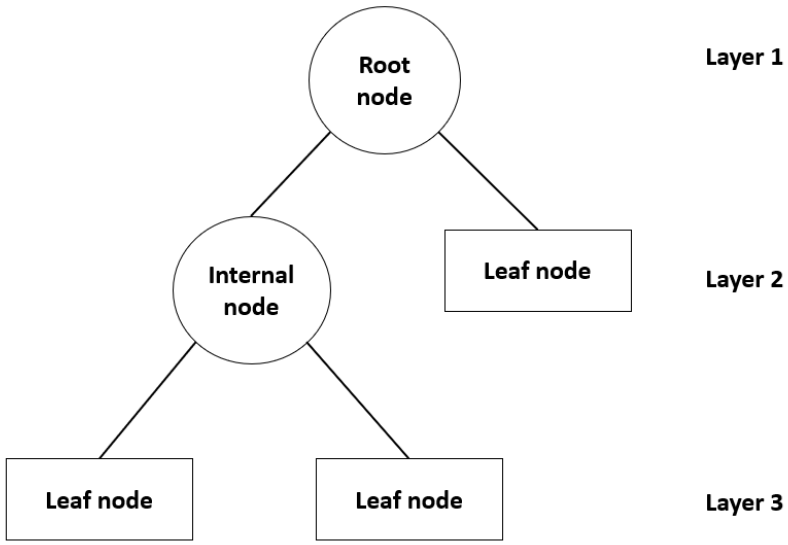


Figure 4.4. Decision tree technique [81].

4.2.2. Support Vector Machine (SVM)

SVM is a versatile and widely used ML algorithm that originated in statistical learning theory and has found extensive application in various fields, including classification and prediction tasks. Its primary objective is to create a decision boundary that effectively separates classes and facilitates the prediction of labels based on one or more feature vectors. Central to the SVM approach are several key concepts, including the notion of linear separability, which allows SVM to handle data that can be distinctly separated by a linear boundary. The kernel trick is another essential aspect of SVM, enabling the algorithm to handle non-linear data by transforming the feature space into a higher-dimensional space, thereby facilitating the better separation of classes, Support vectors, representing the data points closest to the decision boundary, play a crucial role in defining the decision boundary and aiding in the prediction of labels for

unseen data points as shown in Figure 4.5. SVM is a powerful tool in ML for both binary and multi-class classification problems [86-89].

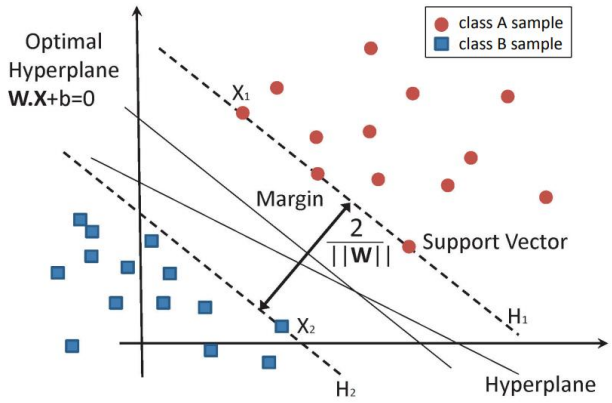


Figure 4.5. Support vector machine technique [89].

4.2.3. K-Nearest Neighbors (KNN)

KNN is a straightforward and intuitive supervised ML algorithm employed for both classification and regression task. KNN regression finds application in various fields, including predicting stock prices, estimating housing values, and forecasting weather pattern. The fundamental idea behind KNN is to identify the class or category that appears most frequently among the KNN of a given data point, as illustrated in Figure 4.6. This approach makes KNN relatively easy to implement as it doesn't require many tuning parameters. Moreover, KNN is a non-parametric and instance-based learning algorithm, which means it doesn't make strong assumptions about the underlying data distribution and bases its predictions on the similarity between data point [90-93].

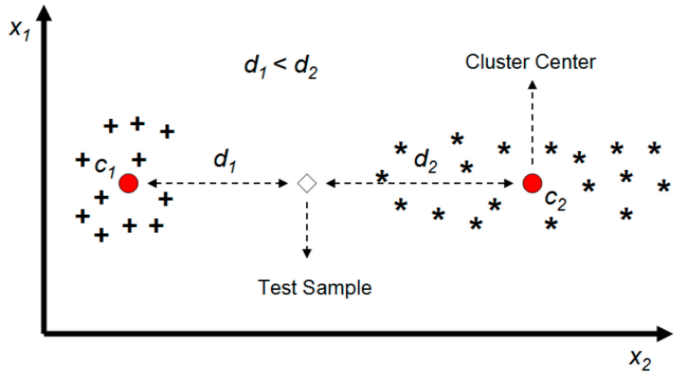


Figure 4.6. KNN technique [92].

4.2.4. Logistic Regression (LR)

LR is a statistical and ML method primarily used for binary classification tasks, although it can also be extended to handle multi-class classification problems. Despite its name, LR is designed to predict the probability of an observation belonging to a particular class, making it a powerful tool for classification tasks It is depicted in Figure 4.7. LR models the relationship between a dependent binary variable (0 or 1) and one or more independent predictor variables by estimating the probability of the dependent variable belonging to the positive class. The predicted probabilities are then transformed using a threshold, typically 0.5, to make final binary predictions. If the probability is greater than or equal to the threshold, the data point is classified as the positive class otherwise, it's classified as the negative class. LR is a straightforward and interpretable method for LP, and it can work well when there are clear patterns or features that indicate the presence or absence of links in a network [94-97].

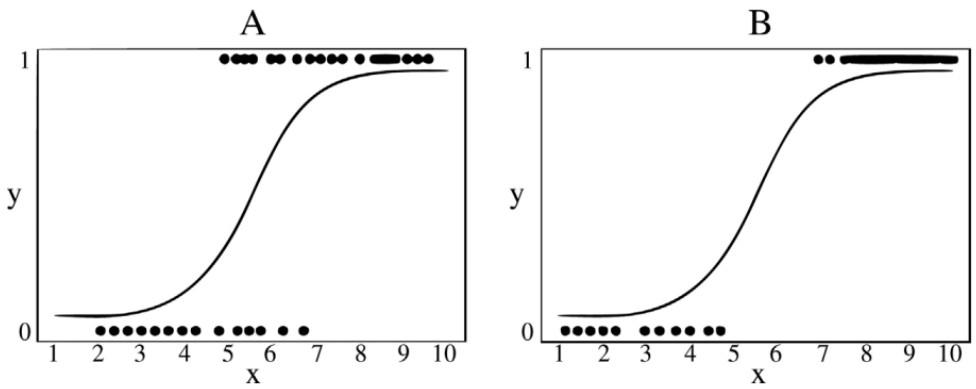


Figure 4.7. Logistic Regression [95].

4.2.5. Random Forest (RF)

RF is a technique common ML algorithm capable of handling both classification and regression tasks with proficiency. This algorithm operates by combining multiple tree predictors, each relying on values from a randomly sampled vector, maintaining the same distribution across all trees within the forest. Bagging randomly selects samples from the dataset to generate a decision tree, with each decision tree relying on a random, independent dataset to make predictions. By combining the predictions of individual trees, RF delivers robust and accurate results, making it a valuable tool in various ML tasks As shown in Figure 4.8. RF is used as a LP algorithm for complex networks to predict the connection probability existing between any two nodes. The use of the RF algorithm in the field of LP its ability to combine multiple DT and aggregate their predictions makes it robust and accurate. Its capabilities can used to solve LP problems in various domains [98-100].

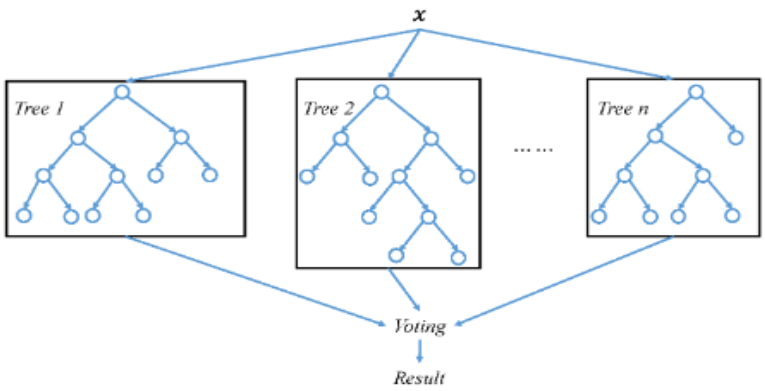


Figure 4.8. Random forest technique [100].

4.2.6. Gradient Boosting (GB)

GB machines are a family of powerful ML techniques that have demonstrated significant success in various practical applications. The GB algorithm operates by sequentially adding weak models, typically decision trees, to the ensemble. Each weak model is trained to minimize the loss function by focusing on the network data that previous models struggled to predict accurately. This process explains how strong predictors can be constructed iteratively by combining weaker modelsAmong the most

popular implementations of GB, decision trees are commonly used as base predictors. The key strength of GB for LP lies in its ability to handle large and complex networks effectively. It can capture non-linear relationships and interactions between entities in the network, leading to more accurate predictions. By leveraging the advantages of ensemble learning and iterative model improvement, as depicted in Figure 4.9. GB can effectively forecast future links in networks. Its capacity to manage complex network structures and noisy data renders it a valuable tool in network analysis and data mining [101-104].

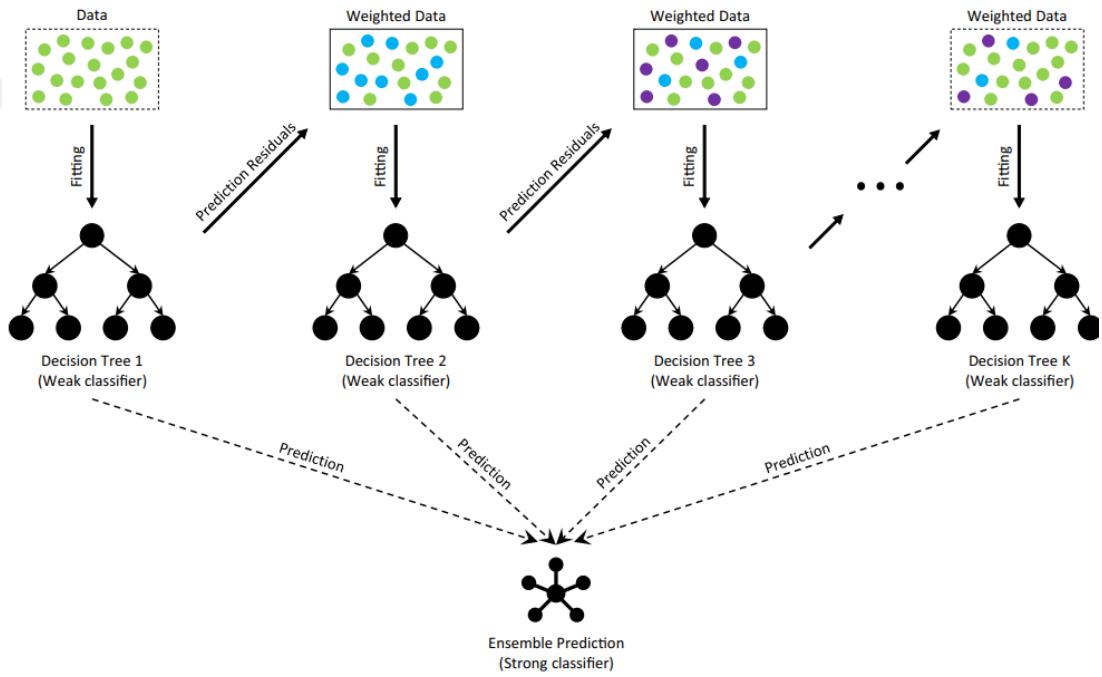


Figure 4.9. Gradient boosting technique [104].

4.2.7. Light Gradient Boosting Machine (LGBM)

The LGBM is a widely used ML algorithm, due to its efficiency, accuracy, and interpretability in ML tasks, such as multi-class classification. LGBM is a gradient-boosting framework that employs a DT based learning algorithm. Its key distinguishing feature lies in the way it grows the trees during the boosting process. Unlike other gradient boosting algorithms that grow trees horizontally, LGBM grows trees vertically, progressing leaf-wise instead of level-wise [105]. This vertical growth strategy allows for a more effective and efficient way of building DT, enabling the

algorithm to handle large datasets and complex tasks with greater speed and less memory consumption. This characteristic is illustrated in Figure 4.10.

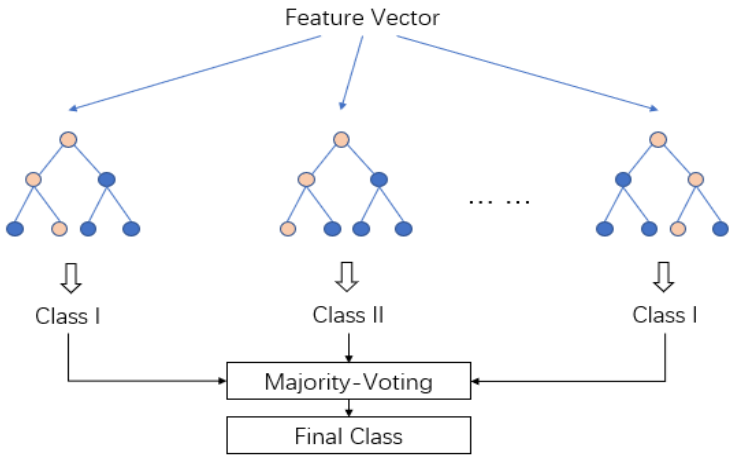


Figure 4.10. LightGBM framework technique [106].

One of the primary reasons for opting for LGBM over other contemporary algorithms is its speed and efficient memory usage, making it a powerful tool for LP tasks, especially when handling large datasets. LGBM's ability to process data swiftly and effectively with minimal memory consumption makes it an ideal choice for various complex ML tasks, including LP and analysis in complex networks [105-106].

4.3. NATURAL LANGUAGE PROCESSING (NLP)

NLP stands for Natural Language Processing, which is a subfield of AI, is a range of computational techniques whose main goal is to provide machines with the capability to understand natural language by understanding its semantics now playing a crucial role in a vast number of knowledge discovery tasks. It includes sentiment analysis, language translation, social networks, academic research [107]. NLP can be applied in the context of complex networks to extract insights and patterns from textual data used to analyze the textual help identify communities or clusters of nodes with similar characteristics. For example, in a social network, you can analyze users' profiles and posts to group them based on common interests [108].

In this thesis, NLP techniques have used to transform complex details in Turkish literature into different topics using TF-IDF. This textual information serves as the basic structure for a complex network connecting authors and allows authors to be thematically identified and clustered. This approach enables the uncovering of hidden patterns and connections in Turkish literature publications, as seen in Figure 4.11. Additionally, TF-IDF, which stands for "Term Frequency-Inverse Document Frequency", is used as a numerical statistic widely used in NLP and information retrieval to evaluate the importance of a term within a document relative to a collection of documents [109].

The TF measures the frequency of a terms appearance in a document, indicating its relative importance within that document [110]. It is calculated as follows:

$$TF_{(t,d)} = \frac{(Number\ of\ times\ term\ t\ appears\ in\ document\ d\ )}{(Total\ number\ of\ terms\ in\ document\ d\ )} \tag{4.7}$$

The IDF evaluates the importance of a term across an entire collection of documents, giving higher weight to rare terms and reducing the weight of common terms [111]. It is calculated as follows:

$$IDF_{(t,d)} = log_{10} \frac{(Total\ number\ of\ documents\ in\ the\ collection\ D\ )}{(Number\ of\ documents\ containing\ term\ t\ )} \tag{4.8}$$

The TF-IDF score for a term in a document is the product of its TF and its IDF, quantifying its importance in a specific document while considering its relevance across the entire document collection [112]. The TF-IDF score is calculated as:

$$TF - IDF_{(t,d,D)} = TF_{(t,d)} * IDF_{(t,D)} \tag{4.9}$$

The TF-IDF process results in a numerical score for each term in each document, providing a numerical representation of the document's content. This representation is valuable for document ranking, text classification, and information retrieval, enabling the comparison of documents based on the similarity of their TF-IDF representations [113].

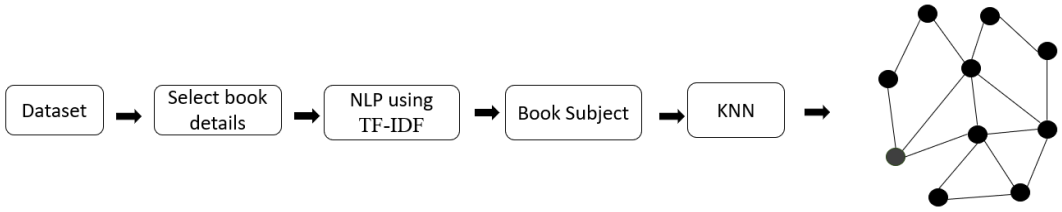


Figure 4.11. NLP technique.

4.4. MEASUREMENT AND EVALUATION

A multitude of measures exist to evaluate a model's performance within ML and hybrid model classes. Parameters like Accuracy, Precision, Recall, and ROC (AUC) constitute some of these essential metrics. Precise estimation of these parameters is crucial in comprehending the constructed model and identifying potential flaws Leveraging these metrics leads to a deeper understanding of ML and hybrid model performance, enabling informed decisions about potential enhancements. Researchers can draw precise conclusions about model accuracy and precision while pinpointing and rectifying potential errors. Ultimately, utilizing these metrics is pivotal for obtaining thorough and accurate evaluations of model performance.

Accuracy is a simple and widely used metric that measures the portion of cases properly categorized relative to the total number of occurrences in the dataset [114]. The equation for Accuracy is defined as:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{4.10}$$

Precision is a valuable metric to evaluate a model's performance, especially in the context of classification problems. TP refer to predicting correctly positives and FP refer to the model incorrectly predicting a positive outcome [115], the equation of precision is presented as following:

$$Precision = \frac{TP}{TP+FP} \tag{4.11}$$

The recall is a crucial metric in the classification mission and refers to the proportion of TP cases in a classification that is accurately identified by models [116]. The Recall is calculated as following.

$$Recall = \frac{TP}{TP+FN} \tag{4.12}$$

One well-known method for evaluating a classification model's effectiveness is the ROC curve, which is particularly useful for differentiating between positive and negative categories. The trade-off between TPR and FPR across different categorization thresholds is visually represented by the ROC curve. The AUC metric, which ranges from 0.5 to 1.0, encapsulates the overall performance of the ROC curve.

A greater AUC denotes a higher TPR and a lower FPR for each decision threshold, indicating the superiority of the classifier. Because it takes into account all possible cutoff values, the AUC measure provides a comprehensive evaluation of classifier performance [117].

PART 5

EXPERIMENTAL STUDY

In this section, the thesis focuses on key aspects such as dataset preprocessing and the experimental study design. It will provide a detailed explanation of the data cleaning process, data partitioning, transformation techniques, and the dataset preparation for model training. Additionally, the section will be focused on the network architecture.

5.1. DATASET

To begin with, the information dataset on Turkish literature is acquired from Kaggle [118]. This dataset encompasses a diverse range of literary works, research papers, and publications across various fields, including literature, psychology, business, science, and etc. moreover, the dataset provides an extensive detail, such as the names of authors, publishers, book titles, book category names, book product codes, publication years, and also comprehensive book details, as illustrated in Table 5.1.

Table 5.1. The structure of the data set used in the experimental study.

N	Book Title	Book Publisher	Book Author	Book Category Name	Book Product Code	Book Page Count	Book Year	Book detail
0	Klinik Uygulamada Bilişsel Davranışçı Terapi - 2	Nobel Akademik Yayıncılık	Mehmet Ak	academic	9.786250e+12	260.0	2022.0	Bilişsel davranışçı terapi (BDT) alanında çalı...
1	Taşra Üniversiteleri	İletişim Yayınevi	Tuğba Tekerek	academic	9.789750e+12	400.0	2023.0	Tugˆba Tekerek, tas¸ra u¨niversitelerini n akad...
2	Çocuklar ve Ergenler İçin Bilişsel Davranışçı ...	Nobel Akademik Yayıncılık	Arzu Aydın	academic	9.786050e+12	304.0	2021.0	Çocuklar ve Ergenler İçin Bilişsel-davranışçı ...
3	İki Dirhem Bir Çekirdek	Kapı Yayınları	İskender Pala	academic	9.789760e+12	212.0	2022.0	Anlatımı güzelleştirmek, savunulan fikir ve dü...

4	Reader at Work 2	ODTÜ Geliştirme Vakfı Yayıncılık	Kolektif	academic	9.789750e+12	550.0	2020.0	This book is the second volume of a collection...
5	Çocuklar İçin İyileştirici Öyküler	Nobel Yaşam	Nancy Davis	academic	9.786060e+12	460.0	2022.0	Hepimizin yaşamı bir öykü… Doğumumuzdan önce ...
6	Adım Adım İyiye	Cenevre Fikir Sanat	V. Özlem Bozkaya	academic	9.786260e+12	273.0	2022.0	Uzayan insan ömrünün, kronik hastalıkların dev...
7	Algoritma Geliştirme ve Programlamaya Giriş	Seçkin Yayıncılık	Doç. Dr. Fahri Vatansever	academic	9.789750e+12	616.0	2020.0	Yenilenmiş 14. baskısını yapan kitaba, okuyucu...
8	Bilişsel Davranışçı Terapi Temelleri ve Ötesi	Nobel Akademik Yayıncılık	Judith S. Beck	academic	9.786050e+12	406.0	2020.0	*Bilişsel Davranışçı Terapi* kitabını, alanda ...
9	Coğrafi Keşifler	Alfa Yayınları	Erol Mütercimler	academic	9.786250e+12	456.0	2022.0	Uygarlık tarihi boyunca insanlık keşifler, ic

5.1.1. Data Pre-Processing

Data pre-preprocessing stands as an important stage within the data mining process. This stage requires cleaning, transforming, and integrating data to prepare it for analysis. The primary objective of data preprocessing is to enhance data quality, and tailor it to suit the specific requirements of the data mining task at hand. Addressing concerns like missing values, outliers, and discrepancies, Moreover, this process ensures that data is appropriately structured and standardized for subsequent analysis and modeling.

In this context, the Data cleaning is identifying and handling missing or erroneous data. This can involve removing or imputing missing values, correcting inconsistent data, or removing outliers. Thus, the possible of issues in analysis prevented by ensuring the quality as well as the integrity of data.

Transforming data text into numerical values is very important, particularly when working with textual data in various data analysis tasks. In additional, this process, often termed text encoding or text vectorization, involves converting text variables into

numerical representations. Utilizing tools like the Label Encoder from the scikit-learn (sklearn) library assigns a unique integer to each distinct category value, as illustrated in Table 5.2. This transformation enables compatibility with ML models while retaining the inherent information within the data. However, it is essential to recognize that label encoding doesn't introduce any new information to the dataset and might not be suitable for all types of categorical data.

Table 5.2. Dataset after labeling and transformation processes.

N	Book Title	Book Publisher	Book Author	Book Category Name	Book Product Code	Book Page Count	Book Released Year	Subject
0	50328	194	1107	2	838	96	1922	5
1	39863	1834	14783	0	872	340	1961	9
2	36369	2538	6538	0	836	172	2007	8
3	41518	2215	17258	0	836	619	1969	7
4	47794	2538	1896	0	836	758	2011	5
5	28222	2171	9798	2	872	151	2008	7
6	52591	2538	23776	0	836	86	1977	5
7	41528	2215	17258	0	836	268	2016	7
8	31640	1897	26081	4	872	95	2013	2
9	43283	2538	23776	0	836	288	2023	5

In this direction, the NLP and ML techniques are utilized to categorize Turkish books using their details. It is worthy to mention that the script involves several key steps, including the importation of essential libraries, preprocessing of data by cleaning and normalizing text, and text vectorization using the TF-IDF technique facilitates the transformation of textual information into a format that is conducive to ML algorithms as Table 5.3, thereby enabling the identification of significant patterns and clusters within the dataset. In this regard, the subsequent application of KNN clustering aids in the automatic categorization of books, simplifying the process of organizing and understanding the diverse collection of Turkish literature based on shared features and attributes. By documenting the steps involved in the process, and the subsequent storage of results in a CSV file.

Table 5.3. Clustering of publication details by subject with NLP.

Book detail		Subject
Bilişsel davranışçı terapi (BDT) alanında çalı...	NLP →	7
Tugˆba Tekerek, tas¸ra u¨niversitelerinin akad...		5
Çocuklar ve Ergenler İçin Bilişsel-davranışçı ...		7
Anlatımı güzelleştirmek, savunulan fikir ve dü...		2
This book is the second volume of a collection...		5
Hepimizin yaşamı bir öykü… Doğumumuzdan önce .		7

5.1.2. Creating Network

In this scenario, the prepare of the structure for two distinct two networks by specifying the nodes and links for the network link prediction. Researcher identify the nodes within the network that hold significance in the LP task. However, these nodes can represent diverse entities, including authors, publishers, and publication years, all of them play a significant role in determining the connections between different network parts. By specifying the nodes, we establish the foundation for analyzing the network's structure and predicting future links. Once the nodes are defined, researcher move on to specifying the links in the network. Subsequently, researcher meticulously delineate the links in the network, representing the various connections or relationships between the specified nodes. Gephi was chosen as the preferred tool for this study [119]. The selection of Gephi was based on its remarkable capability to effectively handle and process substantial volumes of data. Effectively handle and process substantial volumes of data. Additionally, Gephi's capacity to translate complex textual data into clear and intelligible visual representations further solidified our decision. A comprehensive understanding of these links and their underlying characteristics is crucial in predicting the formation of new links within the network. In the thesis study, networks with two different structures have implemented.

Author-Publisher Network: This network entails the specification of nodes, including authors, publishers, and publication years, with book titles serving as the edges as shown in Figure 5.1. Structuring the data in this way facilitates the implementation of

LP, which takes advantage of LSI and ML techniques. This approach enables the construction of a predictive model capable of identifying and foreseeing connections or relationships between authors, publishers, publication years, and book titles in the network. Within this network, our primary objective is to offer recommendations to authors and publishers, guiding them in making informed decisions about their publication strategies for the upcoming years. These recommendations may encompass optimal release timing, target audience selection, and even potential collaborative opportunities, all aimed at enhancing the success and impact.

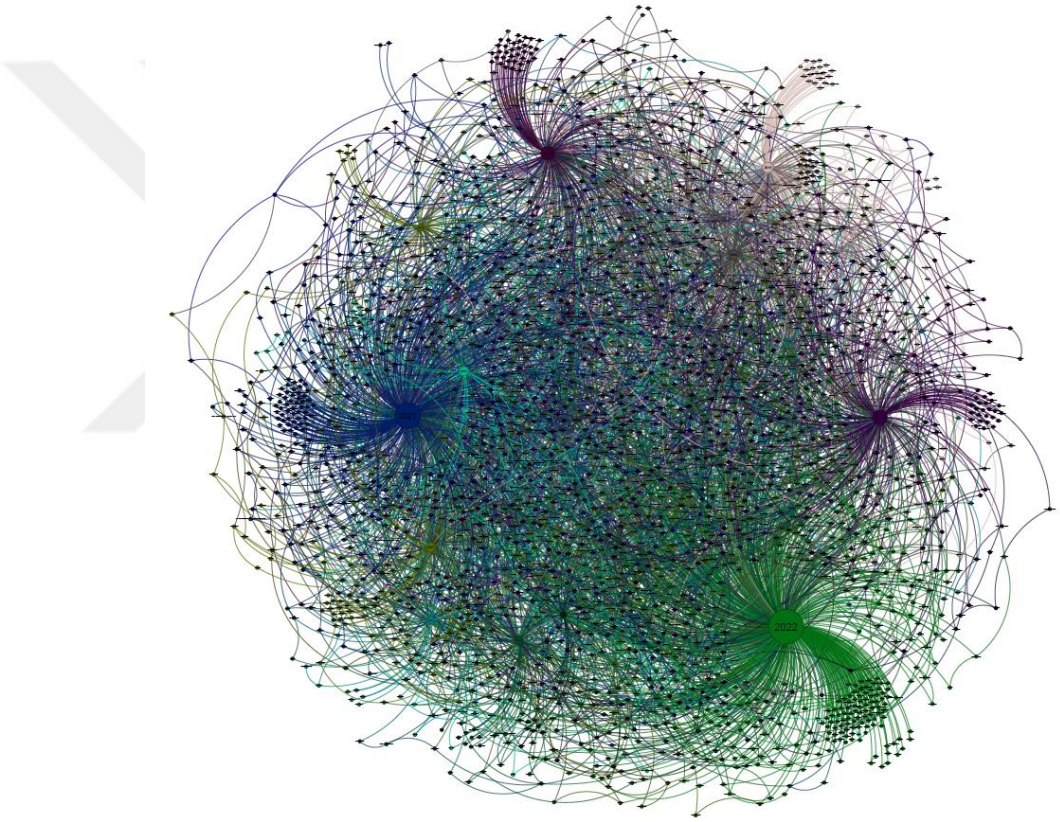


Figure 5.1. Author-Publisher collaboration network.

Author-Author Network: In this network, both the nodes and the links are precisely defined within the network. The nodes represent two distinct authors, while the links signify the shared book titles as illustrated in Figure 5.2. Using LP algorithms and advanced ML methodologies, a predictive model was built to predict possible future collaborations between the two authors. This approach anticipates and comprehends the potential subject areas in which the authors might collaborate based on their shared Turkish works.

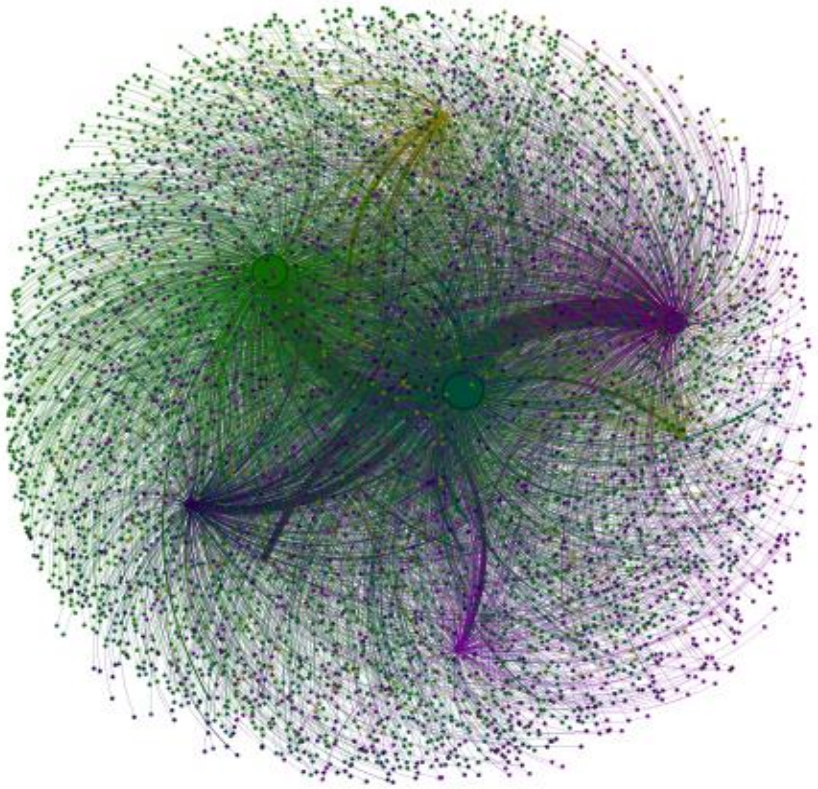


Figure 5.2. Author – Author collaboration network.

Structuring a network is the preliminary work for a comprehensive exploration of network dynamics and fosters a deeper understanding of the underlying relationships and potential collaborations within the dataset. As matter of fact, this strategic approach to data preparation forms the basis for advanced network analysis and fosters the generation of valuable insights and predictions within the network data.

5.1.3. Data Segmentation

Data segmentation involves the process of partitioning a dataset into distinct subsets or segments based on specific criteria or attributes. In the context, the dataset encompasses publications from the years 1920 to 2023, which have been segmented into five distinct periods based on their year of publication. To emphasize, these segmented subsets are subsequently subjected to the application LSI and ML, enabling

the construction of network structures that depict the evolution of relationships over time.

In the first network, the dataset is partitioned into five different time periods based on the publication years as Table 5.4. Within each time a designated portion is allocated for training, encompassing all years within that specific segment. Additionally, another portion is reserved for testing covering the most recent three years within the segment. This strategic division of the dataset ensures that comprehensive training and testing data are available for analyzing the temporal evolution of the network over distinct timeframes.

Table 5.4. Dataset segmentation on years.

Steps	Training	Testing	Nodes	Edges
1	1922-2003	2001, 2002, 2003	2228	4687
2	1922-2008	2006, 2007, 2008	5088	10289
3	1922-2013	2011, 2012, 2013	9730	17196
4	1922-2018	2016, 2017, 2018	14469	23791
5	1922-2023	2021, 2022, 2023	18022	36619

In the author-author networks, the focus is available datasets, utilizing an NLP server to transform the intricate details of publications into distinct subjects. This transformation process enables the conversion of textual information into a structured format that can be readily analyzed and processed. Subsequently its LSI algorithms and advanced ML techniques are applied to the dataset, facilitating a comprehensive analysis of the network's underlying relationships and dynamics.

5.2. PROPOSED HYBRID APPROACH

The experimental study constitutes an extensive exploration of various LP methods applied to complex social networks and interaction networks. However, there remains a paucity of research that specifically assesses the performance of these methods in dynamic networks, particularly concerning the intricate relationships among book publishers within the Turkish publishing dataset. As demonstrated in Table 5.5, the

thesis outlines the construction methods employed in the two networks, encompassing a diverse array of algorithms ranging from LSI to sophisticated ML models. To enhance the overall model performance, a novel hybrid model was introduced, designed to achieve superior accuracy and foster the development of an improved predictive model within the domain of Turkish literature.

Table 5.5. Representation structure of two different networks.

Model type	Networks	nodes	Edges	Objective
1- Model LSI	Author-Publisher Network	Author, publisher, year	Book title	LP model to predict future collaboration between authors and publishers to be based next year
2- Hybrid Model Between LSI and ML	Author-Author Network	Author	subject	Create a LP model to predict subject based connections between authors.

The structure used dataset in the thesis consists of a Turkish literary dataset total comprising approximately 14,537 dataset samples, spanning its inception from 1922 to 2023. In the author-publisher network, its model comprises 18,528 nodes interconnected by 37,742 links, demonstrating a complex and expansive network structure. In contrast, the author-author network presents a model with 11,595 nodes connected by 13,246 links, showcasing a more concise yet interconnected network configuration, thereby encapsulating the evolving dynamics of the literary landscape within the network sphere. The model-building process is visually illustrated in Figure 5.3. The model construction framework is orchestrated through the implementation of two primary methodologies. Initially, LSI algorithms are deployed to discern underlying patterns and relationships within the dataset. Subsequently, ML are integrated into the model, further enhancing its predictive capabilities and performance

metrics performance metrics. are integrated into the model, further enhancing its predictive capabilities and performance metrics.

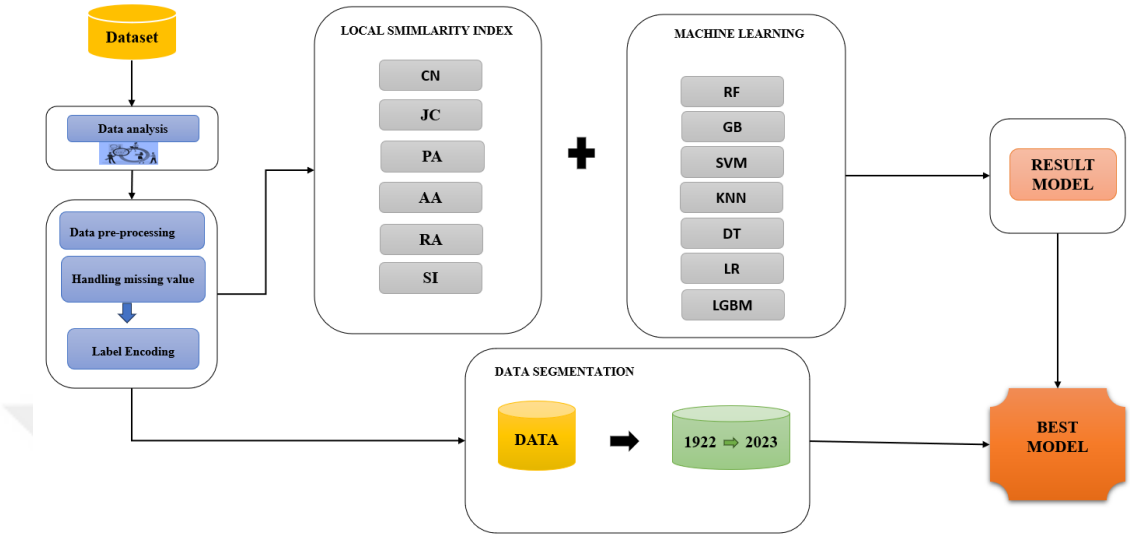


Figure 5.3. Structure link prediction model

PART 6

RESULTS AND DISCUSSION

This section discussed the outcomes utilized two networks and conduct a comparative analysis practically graphical representations and tables data. Then, the identification of the most promising performance metrics will be involved a comparison of methodologies and findings from various research studies.

6.1. LINK PREDICTION RESULTS FOR AUTHOR-PUBLISHER NETWORKS

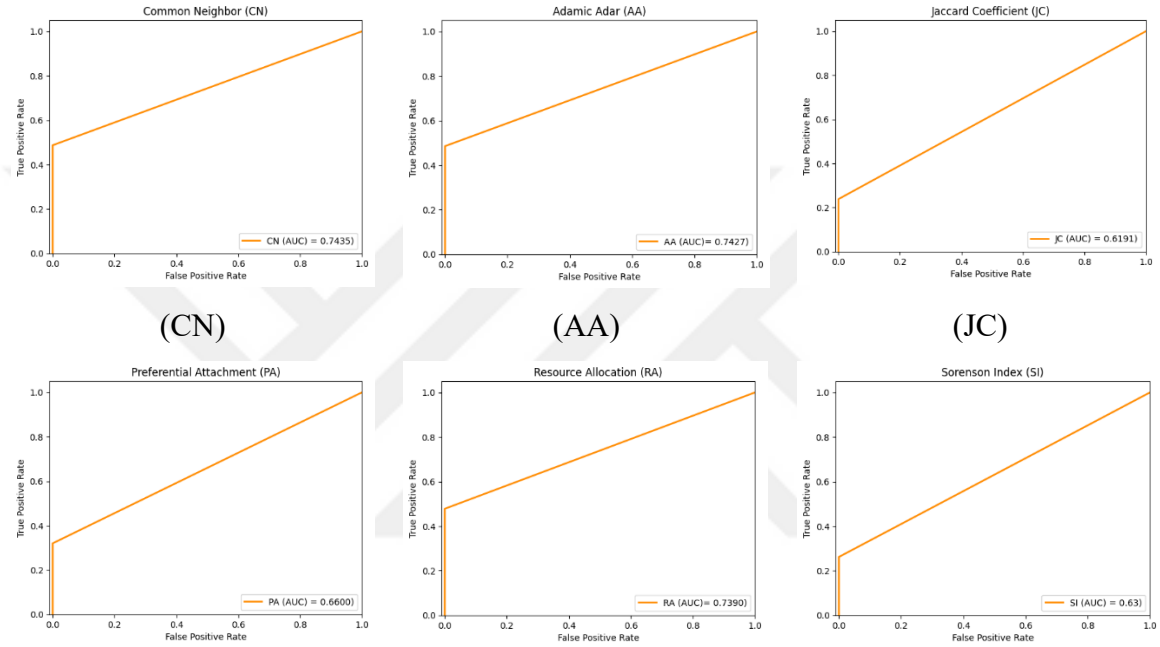
In the author-publisher network, two distinct models were employed, each based on different algorithms, with the results presented for detailed analysis and discussion. This comprehensive approach facilitates evaluation of the outcomes generated by these diverse methodologies, thereby offering valuable insights into their strengths and limitations within the context of the study.

6.1.1. Results of LSI

The LSI measures serve as essential indicators that assess the degree of similarity or relatedness between two nodes within a network, primarily based on their immediate neighbors or local graph structure. Various LSI measures were utilized in this thesis, including CN, JC, AA, PA, RA, and SI. Table 6.1 and Figure 6.1 explain the AUC score obtained to apply the author-publisher network.

Table 6.1. Author-publisher network AUC score obtained by LSI methods.

Similarity local Method	**AUC**
Common Neighbor Index	**0.7435**
Jaccard Coefficient Index	0.6190
Preferential Attachment Index	0.6600
Adamic-Adar Index	0.7426
Resource Allocation Index	0.7390
Sorenson Index	0.6309



(CN) (AA) (JC)

(PA) (RA) (SI)

Figure 6.1. Representation plot of LSI methods applying to author-publisher networks.

The outcomes of LSI manifest discernibly, with the CN, AA metrics attaining the highest accuracies of 0.7435 and 0.7426, respectively. Meanwhile, the RA metric achieved an accuracy of 0.7390, PA demonstrated an accuracy of 0.6600, and SI yielded an accuracy of 0.6309. It is clear that, the JC recorded the lowest accuracy of 0.6190 in assessing the similarity between two nodes.

These findings underscore the superior efficacy of the CN and AA coefficients in gauging the likeness between nodes within the network. Their discernment proves instrumental in identifying pivotal relationships and patterns embedded in the network structure. Conversely, while the PA, RA, and SI metrics rendered commendable

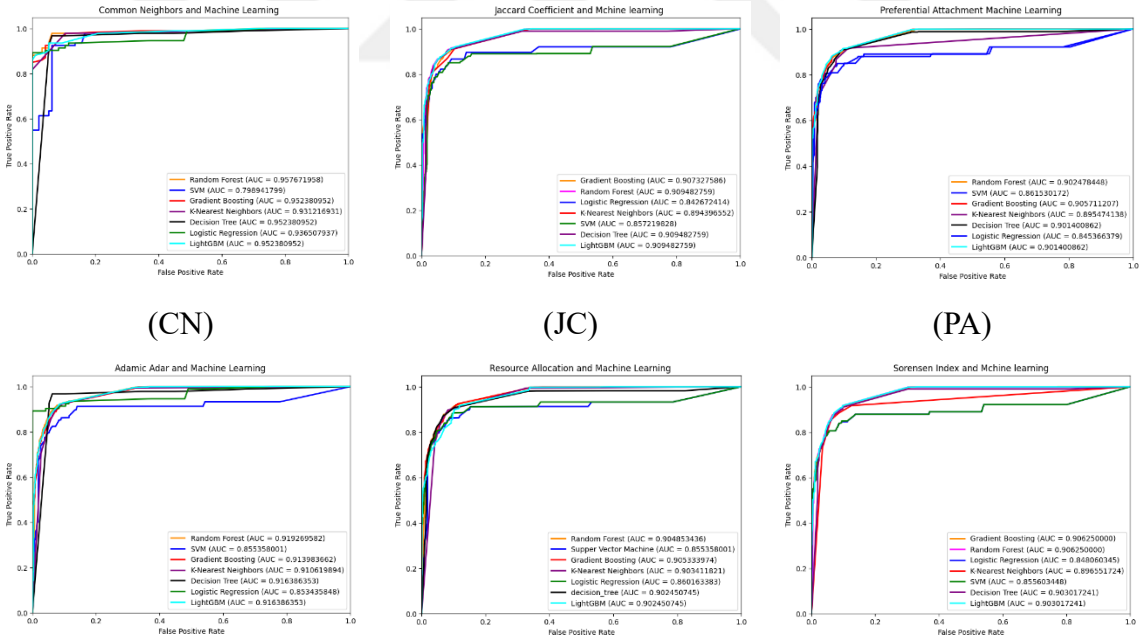
results, the JC metric exhibited a comparatively lower accuracy, indicative of its limitations in capturing nuanced similarities among nodes.

6.1.2. Results of Hybrid Model (LSI and ML)

The AUC results obtained by using LSI methods together with ML in the author-publisher network are shown in Figure 6.2 and Table 6.2.

Table 6.2. AUC score hybrid model the author-publisher network.

LSI	ML						
	DT	**SVM**	**KNN**	**LR**	**RF**	**GB**	**LGBM**
CN	0.9523	0.7989	0.9312	0.9365	**0.9576**	0.9523	0.9523
JC	0.9094	0.8572	0.8943	0.8426	0.9094	0.9073	0.9094
PA	0.9014	0.8615	0.8954	0.8453	0.9024	0.9057	0.9014
AA	0.9163	0.8553	0.9106	0.8534	0.9192	0.9139	0.9163
RA	0.9024	0.8553	0.9034	0.8601	0.9048	0.9053	0.9024
SI	0.9030	0.8556	0.8965	0.8488	0.9062	0.9062	0.9030



(CN) (JC) (PA)

(AA) (RA) (SI)

Figure 6.2. Representation plot of hybrid approach applying to author-publisher networks.

To improve performance of a hybrid model between LSI and ML in the author-publisher network by applying ML techniques to LSI algorithms. In additional, this approach demonstrated that classification significantly improved the accuracy of model results. In this context, the RF algorithm has got the highest accuracy, achieving a good 0.9576 success rate. Thu, the accuracy rate of the same algorithm when applied to the rest of the LSI algorithms ranged from 0.9019 to 0.9576, DT achieved an accuracy range of 0.9014 to 0.9523, GB ranged from 0.9013 to 0.9523, LGBM gained accuracy between 0.9014 and 0.9523, KNN maintained an accuracy range from 0.8943 to 0.9312. LR exhibited an accuracy ranging from 0.88426 to 0.9206, while SVM showed an accuracy ranging from 0.7989 to 0.8615. These results underscore the performance and power of these algorithms illuminating the intricate relationships within the Turkish literary dataset. To conclude, the RF algorithm demonstrated exceptional performance, outperforming other techniques and providing valuable insights into the network dynamics and connections among authors, publishers, publication years, and book titles in Turkish literature.

6.1.3. Results Segmentation Dataset

Following the partitioning of the dataset, the best model that obtained high accuracy is applied to years of publication, divided into five sections as shown in Table 6.3.

Table 6.3. AUC scores best hybrid model to the segmentation dataset based on years.

Steps	Years	Hybrid Model (CN-RF)
1	1922-2003	0.8379
2	1922-2008	0.8792
3	1922-2013	0.9153
4	1922-2018	0.9408
5	1922-2023	0.9583

It is crucial to understand how the network evolves over time after applying the most effective LP model to the Turkish literary dataset.

6.2. LINK PREDICTION RESULTS FOR AUTHOR-AUTHOR NETWORKS

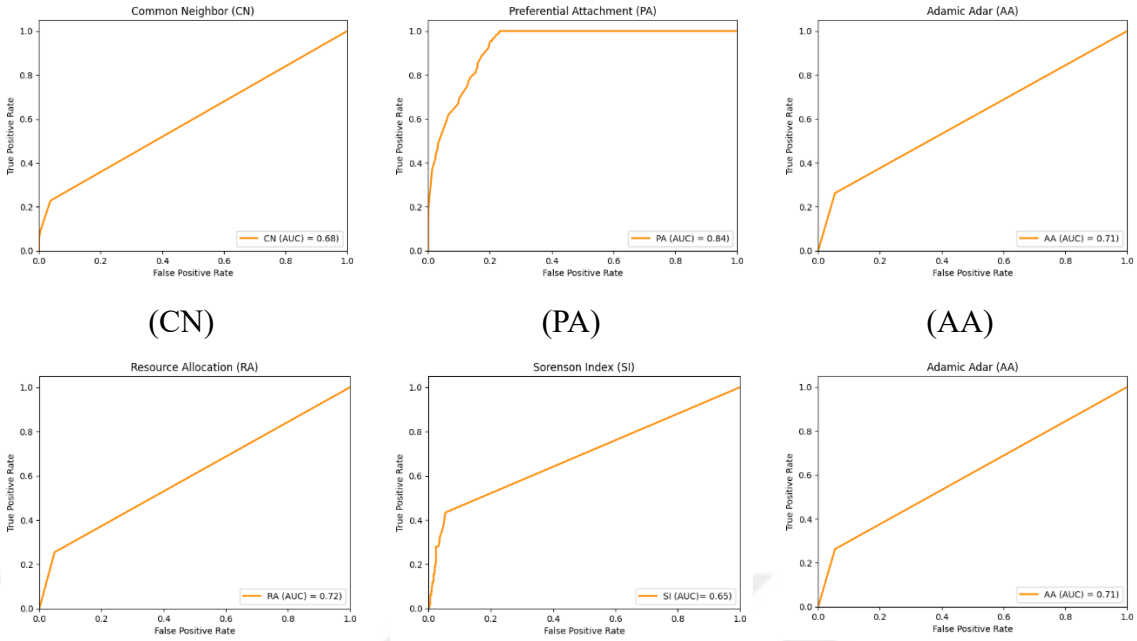
In the author-author network, the implementation of NLP facilitated the conversion of intricate book details into distinct subjects. This transformation enabled the application of two distinct methods based on different algorithms, including LSI and ML. This approach allows for a comprehensive evaluation of the results generated by these methods.

6.2.1. Result of LSI

The LSI is used to measure the similarity between two nodes. the algorithms CN, JC, AA, PA, RA, and SI, has used to apply to its author-author network to find out the similarity between two entities. The results of these analysis are presented comprehensively in and Table 6.4 and Figure 6.3 Representation plot graph has explained the AUC score for each technique.

Table 6.4. AUC score performed using measurement methods to the author-author network.

Similarity local Method	**AUC**
Common Neighbo Index	0.7091
Jaccard Coefficient Index	0.7211
Preferential Attachment Index	**0.8444**
Adamic-Adar Index	0.7052
Resource Allocation Index	0.7211
Sorenson Index	0.7655



(CN) (PA) (AA)

(RA) (SI) (JC)

Figure 6.3. Representation plot of LSI applying to author-author networks.

The outcomes of the LSI revealed distinct accuracy levels, with the PA leading with the highest accuracy at 0.8444, followed by the SI accuracy at 0.7655, JC obtaining accuracy at 0.7211, RA at 0.7211, and CN getting accuracy at 0.7091. Conversely, the AA exhibited the lowest accuracy at 0.7052 in capturing the similarity between the two nodes.

These findings highlight the effectiveness of the PA coefficient in discerning meaningful relationships and patterns within the network structure, underscoring its ability to accurately assess the local similarity between nodes. While the RA and JC measures also delivered commendable results. The comparatively lower accuracy of the CN and AA emphasizes its limitations in capturing the nuanced similarities between nodes within the network.

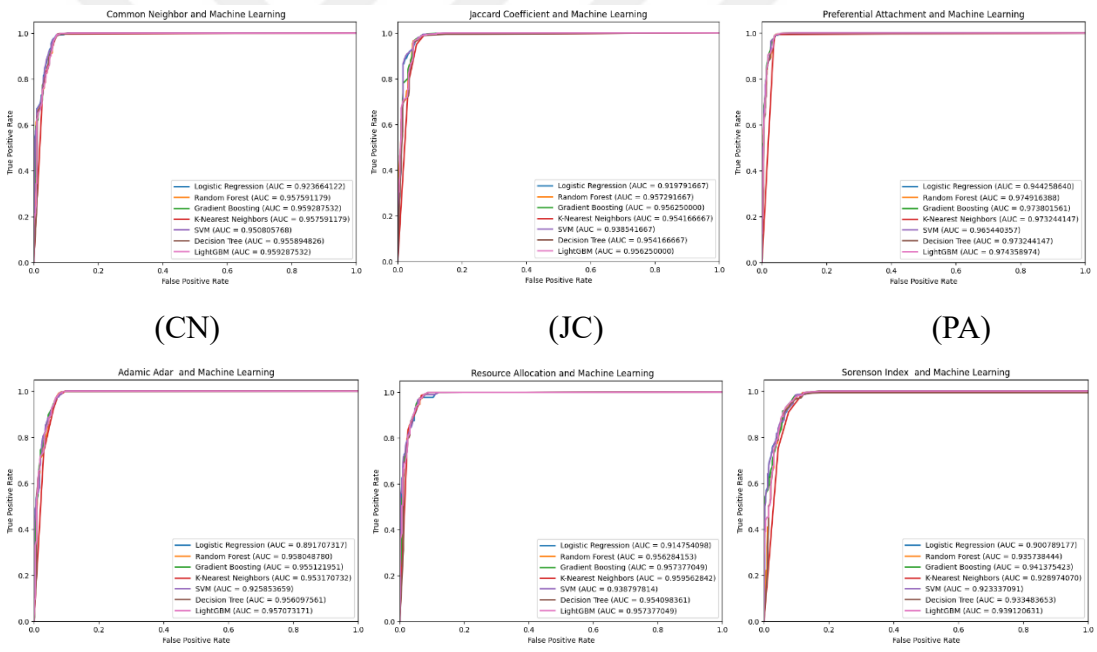
6.2.2. Result of Hybrid Model (LSI and ML)

The author-author network applied a hybrid model between LSI and ML to obtain better results that represent relationships between authors based on their subject. The

results of these analysis are presented comprehensively in and Table 6.5 and Figure 6.4 Representation plot graph has explained the AUC score for each technique.

Table 6.5. AUC score hybrid model the author-author network.

LSI	ML						
	DT	**SVM**	**KNN**	**LR**	**RF**	**GB**	**LGBM**
CN	0.9558	0.9508	0.9574	0.9236	0.9575	0.9592	0.9592
JC	0.9541	0.9385	0.9541	0.9197	0.9572	0.9562	0.9562
PA	0.9732	0.9654	0.9732	0.9442	**0.9749**	0.9738	0.9743
AA	0.9560	0.9258	0.9531	0.8917	0.9580	0.9551	0.9570
RA	0.9540	0.9387	0.9595	0.9147	0.9562	0.9573	0.9573
SI	0.9334	0.9233	0.9289	0.9007	0.9357	0.9413	0.9391



(CN) (JC) (PA)

(AA) (RA) (SI)

Figure 6.4. Representation plot of hybrid approach applying to author-author networks.

In the author-author network, the implementation hybrid model of ML to the local similarity index analysis on the Turkish literary data revealed varying accuracy results. Particularly notable has been the RF algorithm, which achieved the highest accuracy at an 0.9749. Furthermore, when applied to the remaining LSI algorithms, the RF algorithm consistently demonstrated accuracy levels, ranging from 0.9357 to 0.9749,

GB ranged from 0.9413 to 0.9738, LGBM got an accuracy between 0.9391 and 0.9743, KNN achieved an accuracy range between 0.9289 and 0.9732, DT demonstrated an accuracy range from 0.9334 to 0.9732, LR exhibited accuracy ranging from 0.8917 to 0.9442, and SVM showed accuracy ranging from 0.9233 to 0.9654. These findings underscore the adaptability and robustness of the ML algorithms across different similarity indexes, providing valuable insights into the relationships between authors and their subject selections within the context of Turkish literary data.

6.3. DISCUSSION

Evaluation and results obtained from the application of various algorithms, methods and strategies on Turkish literary datasets show that LSI and ML methods are successful for future interactions on Turkish literature dataset. These results show that the hybrid model can be used in recommendation systems for collaboration networks.

In the initial phase, the utilization of local similarity index algorithms for LP in both two networks revealed key insights into the nature of relationships between different entities or nodes within the network. In the author-publisher network, the CN demonstrated notable accuracy of approximately 0.7435. The CN Index's ability to identify the number of shared neighbors between nodes serves as a powerful indicator of potential collaborations and fruitful interactions, offering valuable insights for future engagements and recommendations within the literary landscape. In the author-author network, PA Index was the most successful method with an accuracy rate of 0.8444 among the LSI methods. Its ability to prioritize nodes with higher degrees, this index contributes to the emergence of hubs and the overall structural and dynamic evolution of complex networks.

In the second stage, a hybrid approach has created by applying ML algorithms to LSI techniques in the author-publisher network. In particular, the success of CF-R in capturing and predicting relationships between various literary entities has been demonstrated, exhibiting good accuracy levels of 0.9576. Similarly, in the second author-author network, PA-RF performed well with accuracy levels of 0.9749, demonstrating its capacity to capture new connections and predict relationships

between authors and other literary entities. As seen in Table 6.6, hybrid approaches that give the most successful results in both networks can be used for recommendation systems in association networks.

Table 6.6. The most successful models for two network.

Type networks	**Similarty index**	**AUC score**	**Hybrid model**	**AUC score**
Author-Publisher Network	CN	0.74	CN-RF	0.95
Author-Author Network	PA	0.84	PA-RF	0.97

PART 7

CONCLUSION

This thesis concludes two training methodologies proposed for addressing the challenge of LP in complex networks. By comparing the accuracy of two prevalent methods for relationship prediction, including LSI and ML, it demonstrates their efficacy in the context of real-world Turkish literary datasets. The Python-coded experiments evaluate the efficiency and effectiveness of prediction algorithms, with the support of external tools such as the Networkx library, expediting development and ensuring high quality. Moreover, the thesis delves into the application of NLP and ML techniques for the classification of Turkish books based on their details. It involves the utilization of fundamental libraries, text vectorization using TF-IDF, book categorization using KNN clustering, and the recording of outcomes in a CSV file. This thesis includes a multi-step approach that combines different algorithms and strategies, a promising path toward building future predictive models.

The application used LSI algorithms, which is a technique used in network analysis to predict between two nodes within a network based on their local structural similarities. It focuses on identifying potential connections between nodes by examining their common neighbors or shared connections in the network, In the author-publisher network, CN and AA obtained the highest accuracy and the author-author network PA contained better accuracy, where predicting all possible links in the two networks, Represents ML is one of the effective tools for obtaining results that predict more possible connections. This thesis also has appeared on the effect of ML in generating robust connection predictions between two entities.

In two distinct networks, the hybrid model delivered noteworthy results by implementing algorithms ML to techniques LSI achieving a prediction score high by CN-RF in the author-publisher network and obtaining the best score for PA-RF

algorithms in the author-author network of the literature dataset. To further develop the performance of the model. It is considered that using a hybrid model gives more strength to the model to obtain more future predictions and effective recommendations for authors and publishers to choose their topics more accurately in the coming years.

The proposal for future work in leveraging the insights gained from the current thesis and expanding the study to encompass larger datasets, incorporating additional classes and algorithms within ML and hybrid models demonstrates a strategic and forward-thinking approach. The idea of combining LP algorithms with ML and deep learning as the dataset grows reflects an awareness of the evolving landscape of data analysis. Overall, the results from the study have the potential to improve the scope and applicability of the study.

REFERENCES

[1] J. S. Rubin, "What Is the History of the History of Books?," *J. Am. Hist.*, vol. 90, no. 2, p. 555, 2003, doi: 10.2307/3659444.

[2] D. Lande, M. Fu, W. Guo, I. Balagura, I. Gorbov, and H. Yang, "Link prediction of scientific collaboration networks based on information retrieval," *World Wide Web*, vol. 23, no. 4, pp. 2239–2257, 2020, doi: 10.1007/s11280-019-00768-9.

[3] X. Liu, "Full-Text Citation Analysis : A New Method to Enhance," *J. Am. Soc. Inf. Sci. Technol.*, vol. 64, no. July, pp. 1852–1863, 2013, doi: 10.1002/asi.

[4] M. Wei and A. Noroozi Chakoli, "Evaluating the relationship between the academic and social impact of open access books based on citation behaviors and social media attention," *Scientometrics*, vol. 125, no. 3, pp. 2401–2420, 2020, doi: 10.1007/s11192-020-03678-0.

[5] I. I. Abbasi, "A Hybrid Approach for the Recommendation of Scholarly Journals," pp. 1–6, 2020.

[6] L. L. Linyuan and T. Zhou, "Link prediction in complex networks: A survey," *Phys. A Stat. Mech. its Appl.*, vol. 390, no. 6, pp. 1150–1170, 2011, doi: 10.1016/j.physa.2010.11.027.

[7] B. Xu, L. Li, J. Liu, L. Wan, X. Kong, and F. Xia, "Disappearing Link Prediction in Scientific Collaboration Networks," *IEEE Access*, vol. 6, pp. 69702–69712, 2018, doi: 10.1109/ACCESS.2018.2880233.

[8] G. Resce, A. Zinilli, and G. Cerulli, "Machine learning prediction of academic collaboration networks," *Sci. Rep.*, vol. 12, no. 1, pp. 1–16, 2022, doi: 10.1038/s41598-022-26531-1.

[9] Q. Yu, C. Long, Y. Lv, H. Shao, P. He, and Z. Duan, "Predicting Co-Author Relationship in Medical Co-Authorship Networks," *PLoS One*, vol. 9, no. 7, pp. 1–7, 2014, doi: 10.1371/journal.pone.0101214.

[10] X. Bai, "Predicting the Number of Publications for Scholarly Networks," *IEEE Access*, vol. 6, pp. 11842–11848, 2018, doi: 10.1109/ACCESS.2018.2812804.

[11] M. Azam, M. Nouman, A. H. Al-Faouri, A. M. Saleh, and H. Y. Abuaddous, "Evaluations of Similarity Base Link Prediction Techniques in Social Network," *J. Eng. Sci. Technol.*, vol. 18, no. 2, pp. 1055–1082, 2023.

[12] S. Mishra, S. S. Singh, A. Kumar, and B. Biswas, "ELP: Link prediction in social networks based on ego network perspective," *Phys. A Stat. Mech. its*

Appl., vol. 605, p. 128008, 2022, doi: 10.1016/j.physa.2022.128008.

[13] W. Chen, H. Qu, and K. Chi, "Partner selection in China interorganizational patent cooperation network based on link prediction approaches," *Sustain.*, vol. 13, no. 2, pp. 1–16, 2021, doi: 10.3390/su13021003.

[14] P. M. Aparna, G. N. Jayalaxmi, and V. P. Baligar, "Link Prediction in Social Networks Using Proximity-Based Algorithms," *4th Int. Conf. Emerg. Res. Electron. Comput. Sci. Technol. ICERECT 2022*, pp. 1–6, 2022, doi: 10.1109/ICERECT56837.2022.10060562.

[15] F. Molnar, "Link Prediction Analysis in the Wikipedia Collaboration Graph," no. May, 2011, [Online]. Available: http://assassin.cs.rpi.edu/~magdon/courses/casp/projects/Molnar.pdf

[16] S. Huang, Y. Tang, F. Tang, and J. Li, "Link prediction based on time-varied weight in co-authorship network," *Proc. 2014 IEEE 18th Int. Conf. Comput. Support. Coop. Work Des. CSCWD 2014*, pp. 706–709, 2014, doi: 10.1109/CSCWD.2014.6846931.

[17] X. Song, Y. Zhang, R. Pan, and H. Wang, "Link Prediction for Statistical Collaboration Networks Incorporating Institutes and Research Interests," *IEEE Access*, vol. 10, no. October, pp. 104954–104965, 2022, doi: 10.1109/ACCESS.2022.3210129.

[18] Y. Tian, H. Li, X. Zhu, and H. Tian, "Link prediction based on combined influence and effective path," *Int. J. Mod. Phys. B*, vol. 33, no. 22, 2019, doi: 10.1142/S0217979219502497.

[19] S. X. Feifei Wang, Jiaxin Dong, Wanzhao Lu, "Collaboration prediction based on multilayer all-author tripartite citation networks: A case study of gene editing." 2023.

[20] L. Li, W. Wang, S. Yu, L. Wan, Z. Xu, and X. Kong, "A Modified Node2vec Method for Disappearing Link Prediction," *Proc. - 2017 IEEE 15th Int. Conf. Dependable, Auton. Secur. Comput. 2017 IEEE 15th Int. Conf. Pervasive Intell. Comput. 2017 IEEE 3rd Int. Conf. Big Data Intell. Compu*, vol. 2018-Janua, pp. 1232–1235, 2018, doi: 10.1109/DASC-PICom-DataCom-CyberSciTec.2017.197.

[21] M. Bojanowski and B. Chroł, "Proximity-based Methods for Link Prediction in Graphs with R package 'linkprediction,'" *Ask Res. Methods*, vol. 29, no. 1, pp. 5–28, 2020, doi: 10.18061/ask.v29i1.0002.

[22] T. Zhou, L. Lü, and Y. C. Zhang, "Predicting missing links via local information," *Eur. Phys. J. B*, vol. 71, no. 4, pp. 623–630, 2009, doi: 10.1140/epjb/e2009-00335-8.

[23] A. Javari, H. Qiu, E. Barzegaran, M. Jalili, and K. C. C. Chang, "Statistical link label modeling for sign prediction: Smoothing sparsity by joining local and global information," *Proc. - IEEE Int. Conf. Data Mining, ICDM*, vol. 2017-

Novem, pp. 1039–1044, 2017, doi: 10.1109/ICDM.2017.135.

[24] G. P. Gimenes, H. Gualdron, T. R. Raddo, and J. F. Rodrigues, "Supervised-learning link recommendation in the DBLP co-Authoring network," *2014 IEEE Int. Conf. Pervasive Comput. Commun. Work. PERCOM Work. 2014*, pp. 563–568, 2014, doi: 10.1109/PerComW.2014.6815268.

[25] H. Mihaljević and L. Santamaría, "Disambiguation of author entities in ADS using supervised learning and graph theory methods," *Scientometrics*, vol. 126, no. 5, pp. 3893–3917, 2021, doi: 10.1007/s11192-021-03951-w.

[26] R. C. Chen, C. Dewi, S. W. Huang, and R. E. Caraka, "Selecting critical features for data classification based on machine learning methods," *J. Big Data*, vol. 7, no. 1, 2020, doi: 10.1186/s40537-020-00327-4.

[27] J. Hurtado, N. Taweewitchakreeya, X. Kong, and X. Zhu, "A classifier ensembling approach for imbalanced social link prediction," *Proc. - 2013 12th Int. Conf. Mach. Learn. Appl. ICMLA 2013*, vol. 1, pp. 436–439, 2013, doi: 10.1109/ICMLA.2013.88.

[28] A. Kumari, R. K. Behera, K. S. Sahoo, A. Nayyar, A. Kumar Luhach, and S. Prakash Sahoo, "Supervised link prediction using structured-based feature extraction in social network," *Concurr. Comput. Pract. Exp.*, vol. 34, no. 13, 2022, doi: 10.1002/cpe.5839.

[29] A. Kumari, S. P. Sahoo, R. K. Behera, and B. Sahoo, "Supervised Machine Learning for Link Prediction Using Path-Based Similarity Features," *2020 IEEE 17th India Counc. Int. Conf. INDICON 2020*, 2020, doi: 10.1109/INDICON49873.2020.9342531.

[30] Á. Máté, "Supervised machine learning for text classification," no. January, 2019, [Online]. Available: https://aakosm.github.io/QTA_SZISZ_2019/week06_supervised_ml/06_supervised_ml.pdf

[31] S. Kumar, A. Mallik, and B. S. Panda, "Link prediction in complex networks using node centrality and light gradient boosting machine," *World Wide Web*, vol. 25, no. 6, pp. 2487–2513, 2022, doi: 10.1007/s11280-021-01000-3.

[32] M. Badiy and F. Amounas, "Embedding-based Method for the Supervised Link Prediction in Social Networks," *Int. J. Recent Innov. Trends Comput. Commun.*, vol. 11, no. 3, pp. 105–116, 2023, doi: 10.17762/ijritcc.v11i3.6327.

[33] H. J. Jeong and M. H. Kim, "Utilizing adjacency of colleagues and type correlations for enhanced link prediction," *Data Knowl. Eng.*, vol. 125, no. May, p. 101785, 2020, doi: 10.1016/j.datak.2019.101785.

[34] F. Aghabozorgi and M. R. Khayyambashi, "A new similarity measure for link prediction based on local structures in social networks," *Phys. A Stat. Mech. its Appl.*, vol. 501, pp. 12–23, 2018, doi: 10.1016/j.physa.2018.02.010.

[35] M. Krenn *et al.*, “Predicting the Future of AI with AI: High-quality link prediction in an exponentially growing knowledge network,” no. November, pp. 1–13, 2022, [Online]. Available: http://arxiv.org/abs/2210.00881

[36] W. Hariri, “Unlocking the Potential of ChatGPT: A Comprehensive Exploration of its Applications, Advantages, Limitations, and Future Directions in Natural Language Processing,” 2023, [Online]. Available: http://arxiv.org/abs/2304.02017

[37] E. Kay, J. A. Bondy, and U. S. R. Murty, “Graph Theory with Applications,” *Oper. Res. Q.*, vol. 28, no. 1, p. 237, 1977, doi: 10.2307/3008805.

[38] A. Majeed and I. Rauf, “Graph theory: A comprehensive survey about graph theory applications in computer science and social networks,” *Inventions*, vol. 5, no. 1, 2020, doi: 10.3390/inventions5010010.

[39] Ф. Котлер, *No TitleМаркетинг по Котлеру*. 2008.

[40] C. J. Garroway, J. Bowman, D. Carr, and P. J. Wilson, “Applications of graph theory to landscape genetics,” *Evol. Appl.*, vol. 1, no. 4, pp. 620–630, 2008, doi: 10.1111/j.1752-4571.2008.00047.x.

[41] G. Rücker, “Network meta-analysis, electrical networks and graph theory,” *Res. Synth. Methods*, vol. 3, no. 4, pp. 312–324, 2012, doi: 10.1002/jrsm.1058.

[42] V. Fionda and L. Palopoli, “Biological network querying techniques: Analysis and comparison,” *J. Comput. Biol.*, vol. 18, no. 4, pp. 595–625, 2011, doi: 10.1089/cmb.2009.0144.

[43] Z. Tong, Y. Liang, C. Sun, D. S. Rosenblum, and A. Lim, “Directed Graph Convolutional Network,” 2020, [Online]. Available: http://arxiv.org/abs/2004.13970

[44] Z. Alsheekhhussain, T. Réti, and A. Ali, “Weighted Graph Irregularity Indices Defined on the Vertex Set of a Graph,” *J. Math.*, vol. 2022, 2022, doi: 10.1155/2022/7834080.

[45] J. L. Guillaume and M. Latapy, “Bipartite graphs as models of complex networks,” *Phys. A Stat. Mech. its Appl.*, vol. 371, no. 2, pp. 795–813, 2006, doi: 10.1016/j.physa.2006.04.047.

[46] M. Dehmer, F. Emmert-Streib, and M. Grabner, “A computational approach to construct a multivariate complete graph invariant,” *Inf. Sci. (Ny).*, vol. 260, pp. 200–208, 2014, doi: 10.1016/j.ins.2013.11.008.

[47] R. Albert and A. L. Barabási, “Statistical mechanics of complex networks,” *Rev. Mod. Phys.*, vol. 74, no. 1, pp. 47–97, 2002, doi: 10.1103/RevModPhys.74.47.

[48] I. Türker and E. E. Sulak, “A multilayer network analysis of hashtags in twitter via co-occurrence and semantic links,” *Int. J. Mod. Phys. B*, vol. 32, no. 4, 2018, doi: 10.1142/S0217979218500297.

[49] R. van der Hofstad, “Phase Transition for the Erdős-Rényi Random Graph,” *Random Graphs and Complex Networks*, vol. I, pp. 117–149, 2017, doi: 10.1017/9781316779422.006.

[50] Z. K. Gao, M. Small, and J. Kurths, “Complex network analysis of time series,” *Epl*, vol. 116, no. 5, 2016, doi: 10.1209/0295-5075/116/50001.

[51] J. Jabari Lotf, M. Abdollahi Azgomi, and M. R. Ebrahimi Dishabi, “An improved influence maximization method for social networks based on genetic algorithm,” *Phys. A Stat. Mech. its Appl.*, vol. 586, p. 126480, 2022, doi: 10.1016/j.physa.2021.126480.

[52] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D. U. Hwang, “Complex networks: Structure and dynamics,” *Phys. Rep.*, vol. 424, no. 4–5, pp. 175–308, 2006, doi: 10.1016/j.physrep.2005.10.009.

[53] T. Setiadi, R. Ratih, N. Arkiyah, and G. Prestisia, “Detection of Book and Borrower Communities Based on Book Borrowing Records in the Library Using Complex Network Analysis,” *Int. Conf. Inf. Sci. Technol. Innov.*, vol. 2, no. 1, pp. 171–176, 2023, doi: 10.35842/icostec.v2i1.57.

[54] C. Chira, “Complex Network Analysis using Artificial Intelligence Algorithms,” *Proc. - 2022 24th Int. Symp. Symb. Numer. Algorithms Sci. Comput. SYNASC 2022*, pp. 9–11, 2022, doi: 10.1109/SYNASC57785.2022.00010.

[55] Y. Zang and L. Fiondella, “A Network Reliability Analysis Method for Complex Systems based on Complex Network Theory,” *Proc. - Annu. Reliab. Maintainab. Symp.*, vol. 2022-Janua, pp. 1–6, 2022, doi: 10.1109/RAMS51457.2022.9893999.

[56] W. Jia and T. Jiang, “Information-defined networks: A communication network approach for network studies,” *China Commun.*, vol. 18, no. 7, pp. 197–210, 2021, doi: 10.23919/JCC.2021.07.016.

[57] M. Zhou, Q. Han, M. Li, K. Li, and Z. Qian, “Nearest neighbor walk network embedding for link prediction in complex networks,” *Phys. A Stat. Mech. its Appl.*, vol. 620, p. 128757, 2023, doi: 10.1016/j.physa.2023.128757.

[58] E. Gündoan, B. Kaya, and M. Kaya, “Prediction of symptom-Disease links in online health forums,” *Proc. 2017 IEEE/ACM Int. Conf. Adv. Soc. Networks Anal. Mining, ASONAM 2017*, pp. 876–880, 2017, doi: 10.1145/3110025.3119399.

[59] N. N. Daud, S. H. Ab Hamid, M. Saadoon, F. Sahran, and N. B. Anuar, “Applications of link prediction in social networks: A review,” *J. Netw. Comput. Appl.*, vol. 166, p. 102716, 2020, doi: 10.1016/j.jnca.2020.102716.

[60] J. Chen, X. Wang, and X. Xu, “GC-LSTM: graph convolution embedded LSTM for dynamic network link prediction,” *Appl. Intell.*, vol. 52, no. 7, pp. 7513–7528, 2022, doi: 10.1007/s10489-021-02518-9.

[61] W. Liu and L. Lü, “Link prediction based on local random walk,” *Epl*, vol. 89, no. 5, 2010, doi: 10.1209/0295-5075/89/58007.

[62] L. Ballice, “Co-pyrolysis of Turkish Soma lignite and Şirnak asphaltite. Analysis of co-pyrolysis products by capillary gas chromatography total stream sampling technique,” *Oil Shale*, vol. 19, no. 2, pp. 127–141, 2002, doi: 10.3176/oil.2002.2.04.

[63] V. Martínez, F. Berzal, and J. C. Cubero, “A survey of link prediction in complex networks,” *ACM Comput. Surv.*, vol. 49, no. 4, 2016, doi: 10.1145/3012704.

[64] Z. Huang and D. D. Zeng, “A link prediction approach to anomalous email detection,” *Conf. Proc. - IEEE Int. Conf. Syst. Man Cybern.*, vol. 2, no. November, pp. 1131–1136, 2006, doi: 10.1109/ICSMC.2006.384552.

[65] W. Cukierski, B. Hamner, and B. Yang, “Graph-based features for supervised link prediction,” *Proc. Int. Jt. Conf. Neural Networks*, pp. 1237–1244, 2011, doi: 10.1109/IJCNN.2011.6033365.

[66] C. Yu, X. Zhao, L. An, and X. Lin, “Similarity-based link prediction in social networks: A path and node combined approach,” *J. Inf. Sci.*, vol. 43, no. 5, pp. 683–695, 2017, doi: 10.1177/0165551516664039.

[67] D. Sharma, B. Kumar, S. Chand, and R. R. Shah, “Uncovering research trends and topics of communities in machine learning,” *Multimed. Tools Appl.*, vol. 80, no. 6, pp. 9281–9314, 2021, doi: 10.1007/s11042-020-10072-8.

[68] R. Mahmoodi, S. A. Seyedi, F. A. Tab, and A. Abdollahpouri, “Link prediction by adversarial Nonnegative Matrix Factorization,” *Knowledge-Based Syst.*, vol. 280, no. June, p. 110998, 2023, doi: 10.1016/j.knosys.2023.110998.

[69] S. Soundarajan and John E. Hopcroft, “Using Community Information to Improve the Precision of Link Prediction,” *Www 2012*, 2013.

[70] O. Findik and E. Özkaynak, “Link prediction on networks created from UEFA European competitions,” *Lect. Notes Inst. Comput. Sci. Soc. Telecommun. Eng. LNICST*, vol. 325 LNICST, pp. 207–217, 2020, doi: 10.1007/978-3-030-52856-0_16.

[71] J. Ayoub, D. Lotfi, and A. Hammouch, “Mean Received Resources Meet Machine Learning Algorithms to Improve Link Prediction Methods,” *Inf.*, vol. 13, no. 1, 2022, doi: 10.3390/info13010035.

[72] B. Pandey and A. Khamparia, *Hidden Link Prediction in Stochastic Social Networks*, vol. i. 2019. [Online]. Available: www.igi-global.com.

[73] S. Zeng, “Link prediction based on local information considering preferential attachment,” *Phys. A Stat. Mech. its Appl.*, vol. 443, pp. 537–542, 2016, doi: 10.1016/j.physa.2015.10.016.

[74] P. Szyman and D. Barbucha, “Link prediction in organizational social network based on e-mail communication,” *Procedia Comput. Sci.*, vol. 207, no. Kes, pp. 4008–4016, 2022, doi: 10.1016/j.procs.2022.09.463.

[75] Z. Wu and Y. Li, “Link prediction based on multi-steps resource allocation,” *Proc. - 2014 IEEE/WIC/ACM Int. Jt. Conf. Web Intell. Intell. Agent Technol. - Work. WI-IAT 2014*, vol. 1, no. 11172209, pp. 355–360, 2014, doi: 10.1109/WI-IAT.2014.56.

[76] “Prediction technique for Vortex.pdf,” no. Icosec, pp. 782–786, 2020.

[77] M. Coşkun and M. Koyutürk, “Node similarity-based graph convolution for link prediction in biological networks,” *Bioinformatics*, vol. 37, no. 23, pp. 4501–4508, 2021, doi: 10.1093/bioinformatics/btab464.

[78] R. Kumari and S. Kr., “Machine Learning: A Review on Binary Classification,” *Int. J. Comput. Appl.*, vol. 160, no. 7, pp. 11–15, 2017, doi: 10.5120/ijca2017913083.

[79] S. Anand, Rahul, A. Mallik, and S. Kumar, “Integrating node centralities, similarity measures, and machine learning classifiers for link prediction,” *Multimed. Tools Appl.*, vol. 81, no. 27, pp. 38593–38621, 2022, doi: 10.1007/s11042-022-12854-8.

[80] S. B. Kotsiantis, “Decision trees: A recent overview,” *Artif. Intell. Rev.*, vol. 39, no. 4, pp. 261–283, 2013, doi: 10.1007/s10462-011-9272-4.

[81] M.-H. Chiu, L. C. Hao, and Y.-R. Yu, “THE USE OF FACIAL MICRO-EXPRESSION STATE AND TREE-FOREST MODEL FOR PREDICTING CONCEPTUAL-CONFLICT BASED CONCEPTUAL CHANGE Models and modeling in science learning View project Facial micro-expressions in science education View project,” no. January, 2016, [Online]. Available: https://www.researchgate.net/publication/295860754

[82] V. G. Costa and C. E. Pedreira, *Recent advances in decision trees: an updated survey*, vol. 56, no. 5. Springer Netherlands, 2023. doi: 10.1007/s10462-022-10275-5.

[83] S. Biruntha, B. S. Sowmiya, R. Subashri, and M. Vasanth, “Rainfall Prediction using kNN and Decision Tree,” *Proc. Int. Conf. Electron. Renew. Syst. ICEARS 2022*, no. Icears, pp. 1757–1763, 2022, doi: 10.1109/ICEARS53579.2022.9752220.

[84] A. D. Mankar, “A Comparative Study of Recursive Partitioning Algorithms (ID3 , CART , C5 . 0) for Classification Abstract :,” *Int. Res. J. Humanit. Interdisplinary Stud.*, vol. 1, no. 12582–8568, pp. 462–467, 2023.

[85] C. Deng, F. Yi, X. Li, J. Tang, and G. Sun, “Performance Analysis of CHAID Algorithm for Accuracy,” *Proc. - 2023 Int. Conf. Pattern Recognition, Mach. Vis. Intell. Algorithms, PRMVIA 2023*, pp. 182–186, 2023, doi: 10.1109/PRMVIA58252.2023.00036.

[86] Y. Ma and G. Guo, *Support vector machines applications*, vol. 9783319023. 2014. doi: 10.1007/978-3-319-02300-7.

[87] S. C. Atin Roy, "Support vector machine in structural reliability analysis: A review," *ScienceDriect*, vol. Volume 233, no. May 2023, 2023, [Online]. Available: https://doi.org/10.1016/j.ress.2023.109126

[88] R. Rodríguez-Pérez and J. Bajorath, "Evolution of Support Vector Machine and Regression Modeling in Chemoinformatics and Drug Discovery," *J. Comput. Aided. Mol. Des.*, vol. 36, no. 5, pp. 355–362, 2022, doi: 10.1007/s10822-022-00442-9.

[89] E. García-Gonzalo, Z. Fernández-Muñiz, P. J. G. Nieto, A. B. Sánchez, and M. M. Fernández, "Hard-rock stability analysis for span design in entry-type excavations with learning classifiers," *Materials (Basel).*, vol. 9, no. 7, pp. 1–19, 2016, doi: 10.3390/ma9070531.

[90] S. B. Imandoust and M. Bolandraftar, "Application of K-Nearest Neighbor (KNN) Approach for Predicting Economic Events : Theoretical Background," *Int. J. Eng. Res. Appl.*, vol. 3, no. 5, pp. 605–610, 2013.

[91] A. Sumayli, "Development of advanced machine learning models for optimization of methyl ester biofuel production from papaya oil: Gaussian process regression (GPR), multilayer perceptron (MLP), and K-nearest neighbor (KNN) regression models," *Arab. J. Chem.*, vol. 16, no. 7, p. 104833, 2023, doi: 10.1016/j.arabjc.2023.104833.

[92] ARDAAKDERE, "A new k-nearest neighbors classifier for big data based on efficient data pruning," *Mathematics*, 2023. https://www.kaggle.com/datasets/ardaakdere16/turkish-book-dataset?group=bookmarked+%5B20%5DM.+Bastian%2C+S.+Heymann%2C+and+M.+Ja

[93] A. Zamsuri, S. Defit, and G. W. Nurcahyo, "Classification Of Multiple Emotions In Indonesian Text Using The K-Nearest Neighbor Method," *J. Appl. Eng. Technol. Sci.*, vol. 4, no. 2, pp. 1012–1021, 2023, doi: 10.37385/jaets.v4i2.1964.

[94] M. P. LaValley, "Logistic regression," *Circulation*, vol. 117, no. 18, pp. 2395–2399, 2008, doi: 10.1161/CIRCULATIONAHA.106.682658.

[95] A. A. T. Fernandes, D. B. F. Filho, E. C. da Rocha, and W. da Silva Nascimento, "Read this paper if you want to learn logistic regression," *Rev. Sociol. e Polit.*, vol. 28, no. 74, pp. 1/1-19/19, 2020, doi: 10.1590/1678-987320287406EN.

[96] W. Y. Loh, "Logistic Regression Tree Analysis," *Springer Handbooks*, pp. 537–549, 2006, doi: 10.1007/978-1-84628-288-1_29.

[97] H. Ming and H. Yang, "L0 regularized logistic regression for large-scale data," *Pattern Recognit.*, vol. 146, no. August 2023, p. 110024, 2024, doi: 10.1016/j.patcog.2023.110024.

[98] A. K. Balyan *et al.*, “A Hybrid Intrusion Detection Model Using EGA-PSO and Improved Random Forest Method,” *Sensors*, vol. 22, no. 16, pp. 1–20, 2022, doi: 10.3390/s22165986.

[99] Y. L. Chen, C. H. Hsiao, and C. C. Wu, “An ensemble model for link prediction based on graph embedding,” *Decis. Support Syst.*, vol. 157, no. August 2021, p. 113753, 2022, doi: 10.1016/j.dss.2022.113753.

[100] Y. Wang, Z. Pan, J. Zheng, L. Qian, and M. Li, “A hybrid ensemble method for pulsar candidate classification,” *Astrophys. Space Sci.*, vol. 364, no. 8, pp. 1–15, 2019, doi: 10.1007/s10509-019-3602-4.

[101] A. Natekin and A. Knoll, “Gradient boosting machines, a tutorial,” *Front. Neurorobot.*, vol. 7, no. DEC, 2013, doi: 10.3389/fnbot.2013.00021.

[102] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz, *A comparative analysis of gradient boosting algorithms*, vol. 54, no. 3. Springer Netherlands, 2021. doi: 10.1007/s10462-020-09896-5.

[103] A. V. Dorogush, V. Ershov, and A. Gulin, “CatBoost: gradient boosting with categorical features support,” pp. 1–7, 2018, [Online]. Available: http://arxiv.org/abs/1810.11363

[104] S. Barman and S. Srivastava, “Link Prediction in Social Network using Gradient Boosting,” *Proc. 8th Int. Conf. Commun. Electron. Syst. ICCES 2023*, no. Icces, pp. 626–631, 2023, doi: 10.1109/ICCES57224.2023.10192645.

[105] G. Ke *et al.*, “LightGBM: A highly efficient gradient boosting decision tree,” *Adv. Neural Inf. Process. Syst.*, vol. 2017-Decem, no. Nips, pp. 3147–3155, 2017.

[106] İlyurek Kılıç, “No Title.” https://medium.com/@ilyurek/light-gbm-a-powerful-gradient-boosting-algorithm-fe145a1cd8a6

[107] S. Gil-Clavel and T. Filatova, “Using Natural Language Processing and Networks to Automate Structured Literature Reviews: An Application to Farmers Climate Change Adaptation,” 2023, [Online]. Available: http://arxiv.org/abs/2306.09737

[108] J. Sawicki, M. Ganzha, and M. Paprzycki, *The State of the Art of Natural Language Processing—A Systematic Automated Review of NLP Literature Using NLP Techniques*, vol. 5, no. 3. 2023. doi: 10.1162/dint_a_00213.

[109] B. Kabra and C. Nagar, “Convolutional Neural Network based sentiment analysis with TF-IDF based vectorization,” *J. Integr. Sci. Technol.*, vol. 11, no. 3, pp. 1–7, 2023.

[110] C. Dev and A. Ganguly, “Sentiment Analysis of Assamese Text Reviews : Supervised Machine Learning Approach with Combined n-gram and TF-IDF Feature,” vol. 5, no. 2, pp. 18–30, 2023.

[111] W. Du, C. Ge, S. Yao, N. Chen, and L. Xu, “Applicability Analysis and Ensemble Application of BERT with TF-IDF, TextRank, MMR, and LDA for Topic Classification Based on Flood-Related VGI,” *ISPRS Int. J. Geo-Information*, vol. 12, no. 6, 2023, doi: 10.3390/ijgi12060240.

[112] C. Tao *et al.*, “On Defect Grading for the Relay Protection Devices Based on TF-IDF Assignment and Simple Classifiers,” *J. Phys. Conf. Ser.*, vol. 2433, no. 1, 2023, doi: 10.1088/1742-6596/2433/1/012023.

[113] S. Sitara, N. Mohamed, and K. Srinivasan, “SSN MLRG at MEDIQA-SUM 2023 : Automatic Text Summarization using Support Vector Machine and RoBERTa,” vol. 2107, pp. 0–3, 2023.

[114] R. G. Poola, L. Pl, and S. S. Y, “COVID-19 diagnosis: A comprehensive review of pre-trained deep learning models based on feature extraction algorithm,” *Results Eng.*, vol. 18, no. January, p. 101020, 2023, doi: 10.1016/j.rineng.2023.101020.

[115] M. Perumal and M. Srinivas, “DenSplitnet: Classifier-invariant neural network method to detect COVID-19 in chest CT data,” *J. Vis. Commun. Image Represent.*, vol. 97, no. August, p. 103949, 2023, doi: 10.1016/j.jvcir.2023.103949.

[116] M. Hammad, “Enhancing Smart Home Security : Anomaly Detection and Face Recognition in Smart Home IoT Devices Using Logit-Boosted CNN Models,” 2023.

[117] ARDAAKDERE, “The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification,” *BioData Mining*, 2023. https://www.kaggle.com/datasets/ardaakdere16/turkish-book-dataset?group=bookmarked+%5B20%5DM.+Bastian%2C+S.+Heymann%2C+and+M.+Ja

[118] ARDAAKDERE, “Predicting missing links via significant paths,” *Epl*, 2023. https://www.kaggle.com/datasets/ardaakdere16/turkish-book-dataset?group=bookmarked+%5B20%5DM.+Bastian%2C+S.+Heymann%2C+and+M.+Ja

[119] M. Bastian, S. Heymann, and M. Jacomy, “Gephi: An Open Source Software for Exploring and Manipulating Networks Visualization and Exploration of Large Graphs,” *Proc., Int. AAAI Conf. Web Soc. Media*, pp. 361–362, 2009, [Online]. Available: www.aaai.org

RESUME

Ali Asghar Fahad FAHAD Completed High School At (Al Tuz) High School in Tuz Khurmatu District. Then, He obtained a Bachelor's Degree from College Computer Science from Al-Tikrit University in 2019.

In her quest for further academic advancement, Ali Asghar FAHAD, relocated to Karabük, Turkey, in 2021 to pursue her M.Sc. degree. He embarked on his Master's Program at the Department of Computer Engineering at Karabük University, Turkey.