A Review on Designing of Movie Recommendation System

Nikhil Khonde
Department of IT
G H Raisoni College of Engineering
Nagpur, Maharashtra
nikhil.khonde.it@ghrce.raisoni.net

Nityam Lanjewar
Department of IT
G H Raisoni College of Engineering
Nagpur, Maharashtra
nityam.lanjewar.it@ghrce.raisoni.net

Prerit Shende
Department of IT
G H Raisoni College of Engineering
Nagpur, Maharashtra
prerit.shende.it@ghrce.raisoni.net

Ritesh Amilkanthwar
Department of IT
G H Raisoni College of Engineering
Nagpur, Maharashtra
ritesh.amilkanthwar.it@ghrce.raisoni.net

Prof. Shilpa Ghode
Department of IT
G H Raisoni College of Engineering
Nagpur, Maharashtra
shilpa.ghode@raisoni.net

***Abstract*— In today's digital age, Recommendation systems are crucial in directing users to pertinent content, employing diverse methods to comprehend user preferences and provide tailored recommendations**. With **the ever-increasing data complexity and the evolving nature of user behaviors, recommendation systems face ongoing challenges. The objective of this research is to present the outcomes of employing well-known algorithms like Linear Regression, Linear Discriminant Analysis, K-Nearest Neighbors, Decision Trees, Naive Bayes, and Support Vectors for determining the outcome in the recommendation system. Various challenges faced by recommendation system are also studied. This paper is intended for curious learners wanting to know the relation, effect and outcome of commonly used algorithms with recommendation system.**

Keywords— ***algorithms, recommendation system, movie dataset***

I. Introduction

A recommendation system is a technology that offers personalized suggestions based on your preferences and past behaviors, enhancing your online experience by helping you discover content, products, or services that match your interests and needs. These systems are used in platforms like Netflix, Amazon, and social media to make tailored recommendations, enriching user engagement and satisfaction.

Almost all digital media now uses recommendations in one form or another. They are highly effective in relieving search stress i.e., processing. Recommendations help in providing quick results to majority of users hence preserving the need of hassle to search results for every user. The core idea of recommendations was to relieve processing stress. But it also has an engaging effect of quickly providing results to user without search and efforts.

There are various different and advanced methods of recommendation which achieve their core objective and also provide filtered and personalized recommendations to users. But it either requires processing for every user or a huge data of user interaction, behavior and footprint. A startup may not have access to such.

There are common approaches for recommendation systems but we wanted to know about the effect of Basic Algorithms being used for Recommendation system and observe the results.

II. Literature Survey

Miao Xie, Shijun Liu, and Li Pan [1] addresses the challenge of sparsity in recommender systems, particularly in the context of movie recommendations. The paper introduces the concept of implicit relationships between users and movie crew members. For example, users may have preferences for movies based on their favorite actors or directors, even if other aspects of the movie are less appealing. The proposed algorithm integrates movie crew information into the recommendation model. It represents crew members using latent factor vectors and incorporates them into a matrix factorization model, expanding the scope of the recommendation system. This approach allows the model to better understand users' preferences based on their connections with movie crew members.

Pinhao Wang, Wenzhong Li, Zepeng Yu, Baoguo Lu, and Sanglu Lu [2] tackle the challenge of website recommendation by aiming to anticipate users' preferences based on their browsing history and additional details such as visit frequency and duration. Their proposed method, SI-VAE (Side Information Aided Variational Autoencoder), integrates variational autoencoders (VAEs) to facilitate the recommendation process. The SI-VAE model effectively merges user-website interaction data and side information through the framework of a variational autoencoder. VAEs, known as generative models, grasp the reconstruction of input data and the generation of new samples from a latent space. The implementation of SI-VAE involves a neural network that is trained using a reparameterization trick. Additionally, the introduction of a multinomial likelihood objective function aids in ranking the probabilities of user-website interactions, which are then employed to generate top-k website recommendations.

Wei Li, Jian Xu, Qing Bao, Rujia Shen, Hao Yuan and Ming Xu [3] discusses the challenges in determining user weights for group recommendation, including the need for additional information to determine user decision weights and the necessity of adaptive weight modelling to account for different group contexts and item preferences. To tackle the difficulties, the paper suggests an adaptable aggregation technique grounded in item genre or subject information. This method considers item genre information and historical user-item interactions to calculate reasonable user weights for aggregation functions. It aims to make user weights adaptive to different groups and item preferences.

Betül Bulut, Buket Kaya, Reda Alhajj, and Mehmet Kaya [4] introduce a recommendation system designed to offer tailored academic paper suggestions based on a user's research interests. The research employs various algorithms and techniques to analyze both academic papers and user profiles. The system utilizes methods like TF-IDF, Cosine Similarity, user profiles, and a well-designed recommendation engine to create a personalized paper recommendation system. These techniques collaborate to provide users with relevant academic paper suggestions customized to their preferences. The authors assess the system's performance using evaluation metrics such as F-Measure, which combines precision and recall to evaluate the accuracy of the recommended papers. The authors also compare their system's performance with baseline methods to demonstrate its effectiveness.

Srikar Amara and R. Raja Subramanian [5] propose a user-profile model that leverages tagging mechanisms to improve recommendation accuracy, distinguishing it from current recommender methods. Their approach emphasizes the real-time adaptation of user profiles in response to user actions. Central to their proposal is the integration of POS-taggers via the Natural Language Toolkit (NLTK) framework, aiming to organize and classify dataset information effectively. The paper introduces the T-UP-Tree (Tagger-User Profile-Tree) algorithm, designed to dynamically update user profiles according to user engagement, such as article clicks.

Yongjun Zhang and Hongshuai Wen [6] discusses an improvement to collaborative filtering recommendation algorithms by incorporating semantic relationships between recommended items. The paper proposes enhancing collaborative filtering by incorporating semantic relationships between items. It suggests that items can have relationships based on attributes such as content or category, which can be represented in a knowledge graph. The proposed algorithm combines collaborative filtering with a knowledge graph representation learning approach. It replaces the adjacent set obtained from collaborative filtering with a semantic neighbor set from knowledge graph representation learning.

Ahmed bahaaulddin A. alwahhab [7] addresses the "cold start" problem in recommender systems, particularly for new users, where providing recommendations can be challenging due to a lack of historical data. To tackle this problem, the paper proposes a reinforcement learning-based recommender system that leverages multi-armed bandits to learn and improve its efficiency over time without the need for explicit programming. The paper explores whether Normalized Discounted Cumulative Gain (NDCG) is suitable for measuring the quality of recommender systems that rely on reinforcement learning. It compares NDCG with other evaluation metrics. The results indicate that the proposed system, which operates with a clustered dataset, improves accumulative gain, and decreases RMSE compared to a traditional multi-armed bandit system without clustering. However, the paper suggests that NDCG may not be suitable for measuring the quality of recommender systems based on reinforcement learning due to the random nature of agent choices.

Nan Zhi-hong and Zhao Fei [9] focuses on the challenges of data sparsity and scalability in recommender systems. The paper proposes the use of the semi-supervised co-training method to train recommendation algorithm models from dual perspectives. It divides recommendation algorithms into different types based on their mathematical foundations, such as linear models (e.g., Baseline Only) and matrix decomposition models (e.g., Funk SVD). It reduces the root-mean-square error by 1.4% compared to single-model algorithms trained with only labeled data.

Yisheng Yu, Rui Wei, Kan Hu, Yaru Bu, and Xiaotong Zhang [8] focus on introducing bias into the object-based collaborative filtering (ICF) algorithm, emphasizing the user experience perspective. Their research aims to enhance the comprehensibility and usefulness of recommendations for users, especially in unfamiliar domains. The proposed collaborative filtering algorithm integrates bias to provide more comprehensive reasoning for recommendations, incorporating both the biased component and the contribution of each similar movie to the prediction score. This approach improves the interpretability of the recommendation process, ultimately enhancing the overall user experience. The authors recognize that although the base ICF algorithm may not match the accuracy of deep learning-based methods, it effectively balances real-time performance, interpretability, and accuracy.

III. Background

In today's data-driven age, navigating the overwhelming abundance of choices and information has become an intricate challenge. Recommendation systems, advanced algorithms designed to offer personalized suggestions, have emerged as the essential tools to simplify this complexity. By analyzing user behavior and preferences, these systems help people find content or products tailored to their tastes. Their impact is unmistakable, influencing our online experiences, from e-commerce to streaming services.

Content based and Collaborative recommendation systems are primary types of recommendation system and form basis for it. These systems can be further improvised and personalized by using deep learning or neural network approaches.

We want learn where Common Machine Learning Algorithms stand in the recommendation race. Algorithms of study include Logistic Regression, K-Nearest Neighbors, Support Vector Classifier, Decision Trees, and Naive Bayes.

Through this research, we seek to know the best algorithm for the recommendation system from the commonly known machine learning algorithms. We plan to compare them through their accuracy, recall, precision and f1-score. This research will help others in knowing the results of using other algorithms than content-based and collaborative systems on recommendation system.

IV. Proposed Methodology

At first, data collection occurred from MovieLens, including approximately 33 million ratings and 2 million tag applications across 86,000 movies from 330,975 users. Additionally, the dataset contains tag genome data, comprising 14 million relevance scores across 1,100 tags. Nonetheless, we worked with a limited portion of this data for processing, testing, and analysis.

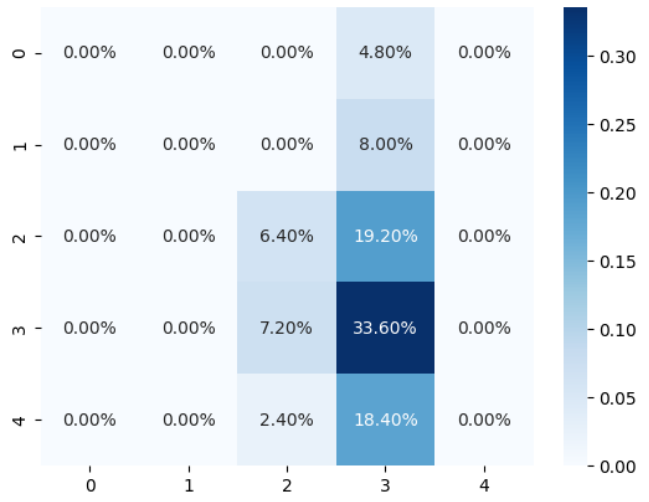
Data preprocessing is the cleaning, transforming, and organizing raw data to make it suitable for analysis or machine learning. This process involves tasks like handling missing values, standardizing formats, and scaling features, ensuring that the data is accurate and consistent, ultimately improving the quality of results. We used to scikit-learn, the Label Encoder is utilized to transform categorical data, as it enables the model to work with these categorical features effectively. The Label Encoder assigns a unique numerical code to each category, allowing the model to interpret and process the data more efficiently during training and prediction.

Feature selection is a crucial technique for enhancing model performance by identifying and retaining the most relevant attributes from a dataset while eliminating fewer valuable ones. This process streamlines the data, reducing dimensionality and potentially enhancing predictive accuracy.

V. Algorithm Tested

Linear Regression -

Linear regression is a statistical method that models the relationship between a dependent variable and one or more independent variables by fitting a straight line to the data, enabling predictions and understanding the connection between variables.

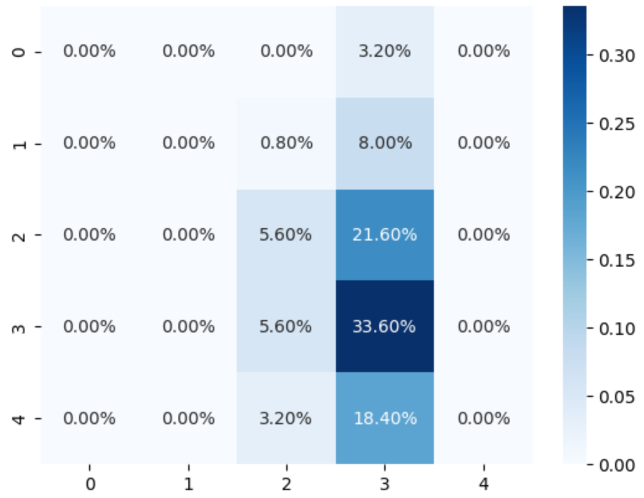


Utilizing linear regression in a recommendation system involves modelling user preferences for items by analyzing historical data, enabling predictions for new recommendations based on a linear relationship between user features and item ratings.

The limitation of using linear regression in a recommendation system is its oversimplified assumption of a linear relationship, often unable to capture the complex patterns and nuances in user preferences for more accurate recommendations.

Linear Discriminant Analysis -

Linear Discriminant Analysis (LDA) is a statistical method that finds the linear combinations of features that best separate multiple classes or categories in a dataset, aiding in dimensionality reduction and pattern recognition.

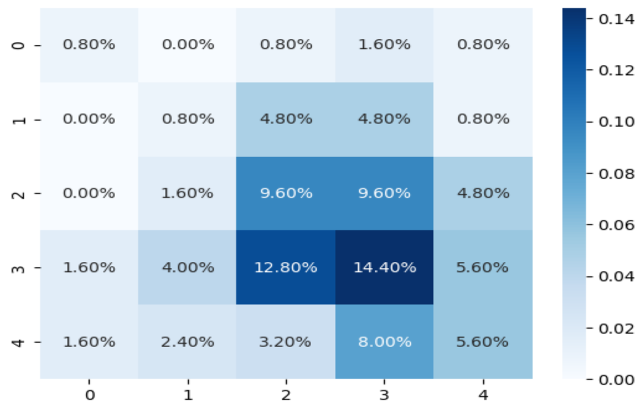


In a recommendation system, Linear Discriminant Analysis (LDA) helps identify linear combinations of user preferences to optimize item recommendations by maximizing the separation between user preferences, enhancing the quality of personalized suggestions.

One limitation of employing Linear Discriminant Analysis in a recommendation system is that it assumes linearity and might not capture complex, nonlinear user-item relationships, potentially limiting the accuracy of recommendations.

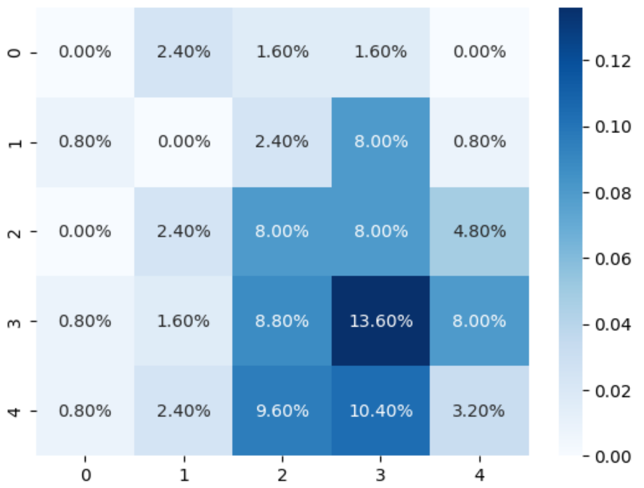
K-Nearest Neighbors -

K-Nearest Neighbors (KNN) is a method utilized in machine learning and data analysis, which discerns the resemblance between data points by examining their nearness in a feature space.



Using K-Nearest Neighbors in recommendation systems provides the advantage of simplicity, enabling effective user-item matching based on similar preferences, resulting in personalized and relevant suggestions. The limitation of K-Nearest neighbors in recommendation systems lies in its computational intensity, making it less efficient for large datasets, and it can also struggle with the "cold start" problem when there is limited user or item data available.

Decision Trees -
Decision trees are a powerful machine learning technique for classification and regression tasks, offering interpretability and adaptability, but they can overfit noisy data and may not generalize well.

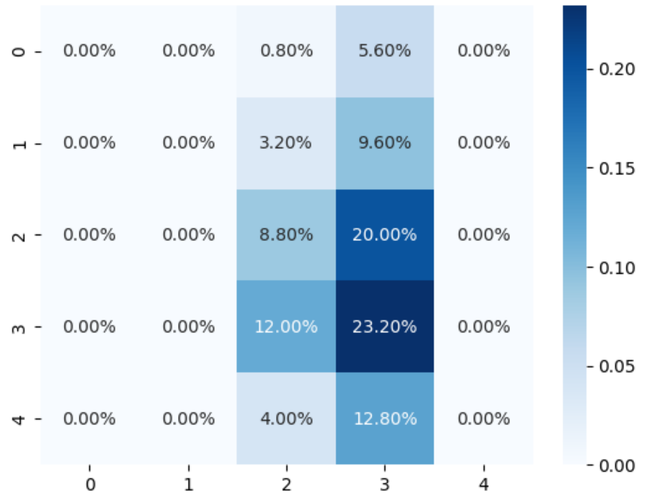


Employing Decision Trees in recommendation systems offers the advantage of interpretability and simplicity, making them ideal for explaining recommendations to users and handling both categorical and numerical data effectively. The challenges in employing Decision Trees for recommendation systems include their susceptibility to overfitting, difficulties in capturing complex patterns in data, and a limited ability to handle large and sparse datasets.

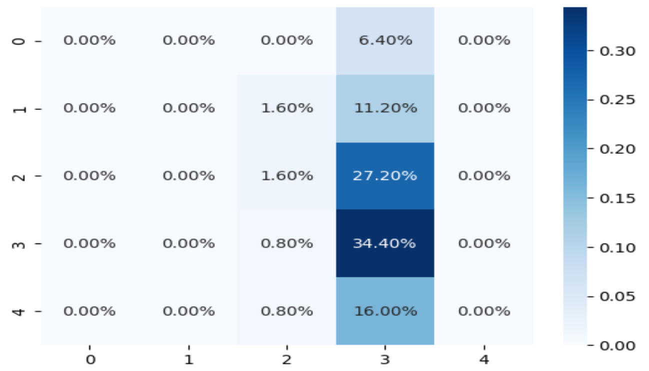
Naive Bayes -
Naive Bayes is a probabilistic classification algorithm that's simple and computationally efficient but assumes feature independence, which may not hold in real-world data.

The advantages of employing Naive Bayes in a recommendation system include its efficiency, simplicity, and ability to handle high-dimensional data effectively, making it well-suited for text-based or categorical recommendations. A limitation of using Naive Bayes for recommendation systems is its strong independence assumption, which can lead to inaccurate predictions when features are not truly independent in the data.

Support Vector Classifier -
The Support Vector Classifier (SVC), while robust and versatile in many classification tasks, can face computational challenges when processing extensive datasets due to its



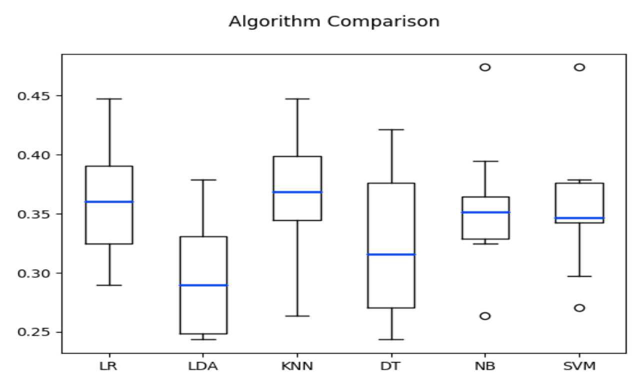
complex optimization process, potentially requiring substantial computational resources.



VI. Result

	Accuracy	Precision	Recall	F1-score
Linear Regression	40.00%	0.27	0.4	0.3
Linear Discriminant Analysis	39.00%	0.26	0.39	0.28
K-Nearest Neighbors	41.00%	0.31	0.31	0.31
Decision Trees	24.00%	0.22	0.25	0.23
Naive Bayes	32.00%	0.2	0.32	0.24
Support Vector	36.00%	0.22	0.36	0.21

From the table, it can be clearly observed that within the tested algorithms K-Nearest Neighbors provides the best overall results. However, the overall accuracy is quite low and we come to realize that using commonly known algorithms for recommendation system leads to lower performance and poor accuracy.



VII. Problem Faced By Recommendation System

Cold Start: The difficulty lies in giving precise suggestions for new users or items with limited or no past data, as the system lacks the necessary information to offer tailored recommendations.

Accuracy: Recommendation systems face accuracy issues, as human behavior is unpredictable and is also affected by many factors. Hence, relevant suggestions are not always accurate.

Diversity: Recommendation systems often face diversity issues where they tend to suggest similar items, limiting the variety of content users encounter. This lack of diversity can lead to users missing out on a broader range of choices and can be a challenge for recommendation algorithms to address.

Scalability: As the user base and the amount of data grow, these systems can struggle to efficiently process and provide accurate recommendations, leading to slower response times and decreased overall performance.

Sparsity: It relates to the challenge of having limited data available for making accurate recommendations. This issue occurs when there are numerous items and users, but very few interactions or ratings between them.

VIII. Conclusion

In this paper, we studied about various common algorithms and their effect on recommendation system. We tested each algorithm on the basis of accuracy, precision, recall and F1-score. K-Nearest Neighbor provide the best overall result in all aspects which can be further enhanced using deep learning and Blockchain technology.

Studying recommendation systems provided knowledge that these systems have evolved to adapt to the increasing complexity of user data and changing contexts. They now incorporate advanced strategies such as context-based filtering, which tailors' recommendations to specific situations, and they grapple with emerging technologies like blockchain, which can challenge traditional collaborative filtering methods by enhancing user privacy. As these systems continue to develop, striking a balance between privacy and accurate recommendations remains a key challenge.

References

[1] M. Xie, S. Liu and L. Pan, "A recommendation algorithm for sparse data via implicit relationship," 2020 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom), Exeter, United Kingdom, 2020, pp. 586-592, doi: 10.1109/ISPA-BDCloud-SocialCom-SustainCom51426.2020.00099

[2] P. Wang, W. Li, Z. Yu, B. Lu, and S. Lu published a paper titled "Website Recommendation with Side Information Aided Variational Autoencoder" in the 2020 IEEE 39th International Performance Computing and Communications Conference (IPCCC) in Austin, TX, USA. The paper spans pages 1-6 and can be accessed via doi: 10.1109/IPCCC50635.2020.9391524.

[3] W. Li, J. Xu, Q. Bao, R. Shen, H. Yuan, and M. Xu authored a paper titled "Group Recommendation: An Adaptive Aggregation Method Based on Movie Genre" presented at the 2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI) in Baltimore, MD, USA. This work can be found on pages 73-78 and is accessible via the DOI: 10.1109/ICTAI50040.2020.00022.

[4] B. Bulut, B. Kaya, R. Alhajj and M. Kaya, "A Paper Recommendation System Based on User's Research Interests," 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), Barcelona, Spain, 2018, pp. 911-915, doi: 10.1109/ASONAM.2018.8508313.

[5] S. Amara and R. R. Subramanian, "Collaborating personalized recommender system and content-based recommender system using TextCorpus," 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 2020, pp. 105-109, doi: 10.1109/ICACCS48705.2020.9074360.

[6] Y. Zhang and H. Wen, "Research on recommendation system based on integrating semantic similarity," 2020 2nd International Conference on Artificial Intelligence and Advanced Manufacture (AIAM), Manchester, United Kingdom, 2020, pp. 18-21, doi: 10.1109/AIAM50918.2020.00010.

[7] Y. Yu, R. Wei, K. Hu, Y. Bu and X. Zhang, "Research on an Interpretable Real-Time Information Recommendation Model based on BAS-ICF algrithm," 2020 Management Science Informatization and Economic Innovation Development Conference (MSIEID), Guangzhou, China, 2020, pp. 304-308, doi: 10.1109/MSIEID52046.2020.00063.

[8] N. Zhi-hong and Z. Fei, "Research on Semi-supervised Recommendation Algorithm Based on Hybrid Model," 2020 2nd International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI), Taiyuan, China, 2020, pp. 344-348, doi: 10.1109/MLBDBI51377.2020.00073.

[9] Alwahhab presented a novel recommender system addressing the cold start problem by employing k-means clustering and a reinforcement learning agent. This work was showcased at the 2020 2nd Annual International Conference on Information and Sciences (AiCIS) in Fallujah, Iraq, with the paper spanning pages 13-21 and bearing the DOI: 10.1109/AiCIS51645.2020.00013.

[10] N. Ifada, T. F. Rahman and M. K. Sophan, "Comparing Collaborative Filtering and Hybrid based Approaches for Movie Recommendation," 2020 6th Information Technology International Seminar (ITIS), Surabaya, Indonesia, 2020, pp. 219-223, doi: 10.1109/ITIS50118.2020.9321014.

[11] C. Cai and L. Wang, "Application of improved k-means k-nearest neighbor algorithm in the movie recommendation system," 2020 13th International Symposium on Computational Intelligence and Design (ISCID), Hangzhou, China, 2020, pp. 314-317, doi: 10.1109/ISCID51228.2020.00076.

[12] Mohamed, Marwa & Khafagy, Mohamed & Ibrahim, Mohamed. (2019). Recommender Systems Challenges and Solutions Survey. 10.1109/ITCE.2019.8646645.Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].