ISSN: 0974-5823 Vol. 7 (Special Issue, Jan.-Mar. 2022)

International Journal of Mechanical Engineering

Comparison of Hybrid Novel Pearson Correlation Coefficient with K-Means Clustering Model to Improve Accuracy for Movie Recommendation System

Syed Mohammed Shoaib[1], Jaisharma K[2*]

[1]Research Scholar, Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical And Technical Sciences, Saveetha University, Chennai, Tamilnadu, India, Pincode: 602105.

[2]Project Guide, Corresponding author, Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical And Technical Sciences, Saveetha University, Chennai, Tamilnadu, India, Pincode: 602105.

*** Corresponding author:** Jaisharma K

ABSTRACT

Aim: Comparison of hybrid recommendation model using Hybrid Novel Pearson Correlation Coefficient (HNPCC) with K-Means Clustering (KMC) model to improve accuracy for movie recommendation. **Materials and Methods:** Hybrid movie recommendation system using the HNPCC and the existing system is KMC. The sample size of 30 per group and the pretest power is 0.8. The data from the movielens dataset has 23 attributes and 2725 instances. **Results:** In this study, it is observed that HNPCC has a slight increase in accuracy of 94.3% and KMC is 89.7%. The statistical significance was identified with an independent sample T-test, it shows there exists a significant difference 0.001 between the proposed and existing groups with $p<0.05$ with Confidence Interval (CI) 95%. **Conclusion:** The comparison results show that the HNPCC has better performance than KMC in the aspects of improved accuracy.

Keywords: Recommendation Systems, Hybrid Novel Pearson Correlation Coefficient, K-Means Clustering, Machine Learning, MovieLens, Movie Recommendation System.

INTRODUCTION

The analysis goal of this study is to develop an associate degree correct recommendation system that helps users get acceptable recommendations. K-Means cluster (KMC) techniques became one of the foremost common techniques for providing customized services to users in recent years. KMC techniques collect previous info from users regarding things resembling books, music, moving-picture shows, ideas, etc. The movie recommendation system offers a mechanism to portion the user to realize the renowned film by obtaining an opinion from an identical user (Musa and Zhihong 2020). A recommendation engine leverages this large quantity of knowledge by finding patterns of user behavior (Chen et al. 2021). Recommender systems solve this drawback by looking through large quantities of dynamically generated info to supply users with customized content and services (Chen et al. 2021; Daniel 2020). The web contains an outsized amount of information that should then be filtered to see quality sure users. Recommender systems are a really appropriate tool for this purpose (Chen et al. 2021; Daniel 2020; Walek and Fojtik 2020). In a chilly beginning situation, users might realize the foremost similar neighbors by looking forward to a poor variety of ratings, leading to low-quality recommendations using Machine Learning (Sadowski et al. 2021).

There are approximately 12,318 articles in Google Scholar and 95 IEEE Xplore articles on movie recommendation systems. In the study (Priscilla and Naveena 2020) a system based on probabilistic matrix factorization was implemented, in which the similarity of the cosine is used to assess the similarity between users. They also implemented the NB tree which is used to achieve 85.8% accuracy in creating user links. Movie recommendation systems are prevalent in today's marketplace as people tend to spend a lot of money going to the cinema or renting a movie so they need to make an informed decision about it (Singh et al. 2021). The research article (Singh et al. 2021; Malik et al. 2019) implemented a hybrid filtering approach based on Machine Learning in their article. Various techniques such as grouping, similarity and classification are implemented in their system to achieve a better recommendation accuracy of 89.8%.The movie recommendation system provides a mechanism to assign the user to get the famous movie by getting a review from similar users or a previous rating from the user (Singh et al. 2021; Malik et al. 2019; Satapathy et al. 2020). The most cited article was (Satapathy et al. 2020; Kumar et al. 2019) focused on predicting the accuracy of movie recommendation systems using the HNPCC hybrid recommendation focus algorithm with an accuracy of 89.8%.Previously our team has a rich experience in working on various research projects across multiple disciplines(Ezhilarasan et al. 2021; Balachandar et al. 2020; Muthukrishnan et al. 2020; Kavarthapu and Gurumoorthy 2021; Sarode et al. 2021; Hannah R et al. 2021; Sekar, Nallaswamy, and Lakshmanan 2020; Appavu et al. 2021; Menon et al. 2020; Gopalakrishnan et al. 2020; Arun Prakash et al. 2020)

The existing methods used were outdated, lower accuracy rate, less reliable, and are inefficient in recommending films. By combining the knowledge and experience with various recommender algorithms to develop novel solutions to the problem at hand with the help of Machine Learning. The main objective of this research is to make more efficient film recommendations by implementing and comparing the HNPCC and KMC recommendation system models.

MATERIALS AND METHODS

The study environment was carried out in the Data Analysis Laboratory of the Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Chennai. In this study, 2 groups of samples were identified, Group 1 was HNPCC, and Group 2 was KMC. The number of samples per group was identified via the clinical website, the pretest power of 0.8 was taken into account, and a measured sample size of N = 30 iterations was performed for each group (Thulaseedaran et al. 2018).

The data set was taken from the MovieLens 100K dataset, consisting of 100,000 ratings (15) from 943 users for 1,682 films. Each user has rated at least 20 films. Simple demographic information for users (age, gender, occupation, zip code). The complete dataset of MovieLens, 100000 reviews from 943 users on 1682 articles. Each user has rated at least 20 films, users and articles are listed one after the other from 1 randomly arranged data. This is a tab-separated list of (user ID, date, and time of item review) with timestamps are unix seconds from 1.1.1970 UTC. Item Movie Information is a tab-delimited list of Movie ID, Movie Title, Release Date, Video Release Date, IMDb URL, Unknown, Action, Adventure, Animation, Kids, Comedy, Crime, Documentary, Drama, Fantasy, Film Noir, Horror, Musical, mystery, romance, sci-fi, thriller, war, western (Song et al. 2020).

K-Means Clustering (KMC)

The final 19 fields are the genres that suggest that the film belongs to that genre, which suggests that it is now no longer using MovieLens dataset. Movies may be of a couple of genres at an equal time. Movie IDs are used withinside the information set using Machine Learning. Content-primarily based totally advice structures do now no longer incorporate information acquired through customers aside. It handiest facilitates perceived merchandise which are just like the product liked by the movie viewers. KMC is restricted as it does not contain any other user data, and it does not help the user to discover their potential taste (Liu et al. 2020).

Pseudocode: K-Means Clustering (KMC)

Input: Training Dataset

Output: Recommender accuracy

1. Import the python libraries: Numpy, Pandas, Matplotlib, sklearn
2. Read the CSV information as data frames in the user and item variable from MovieLens dataset.
3. Split the data into the training set and test set as a data frame into the variables rating and rating test.
4. Create a utility matrix name utility that tells which user rated which movie.
5. Using the WCSS method, choose the right number of clusters so that the K-means Clustering technique can be applied to classify the movies according to the number of clusters.
6. Define the utility clustered matrix after applying the K-means clustering algorithm.
7. Apply Pearson Correlation metric on utility clustered matrix to calculate the similarity matrix between the users.
8. Normalize the values stored in the utility matrix.
9. Guess() function takes two parameters as input userID and top users which is used by KNN to predict the movie ratings for top similar users.
10. RatingTest data frame ratings are used for comparison while using the guess function for predicting the ratings of test users.
11. RMSE is calculated to evaluate the accuracy of the model.

Hybrid Novel Pearson Correlation Coefficient (HNPCC)

The Pearson correlation coefficient is one of the most widely used similarity measures in collaborative filtering recommender systems for determining how much two users are correlated. This system finds the recommended movie from all movies using the similarity method based on Machine Learning algorithms (Li et al. 2020).

Formula For Pearson correlation coefficient,

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Taking the Kaggle dataset and uploading it into the Jupyter notebook. Data preprocessing might be finished to get rid of noise from the statistics. The statistics are then divided into education and check sets. To educate and compare the algorithm, in addition to examining the expected accuracy using a Machine Learning approach (Daniel 2020).

Pseudocode: Hybrid Novel Pearson Correlation Coefficient (HNPCC)

Input: Training Dataset

Output: Recommender accuracy

1. Read the training dataset as input from MovieLens dataset.
2. Preprocess the dataset and split it to train and test.

3. Removing Noise from the data.

4. Visualizing the data.

5. Removing Sparsity.

6. Making the movie recommendation system model.

7. Making the recommendation function.

8. Getting the recommendations.

The Jupyter lab platform was used to assess the device studying algorithm. The experiments had been completed on an Intel Core i5 pc with 8 GB of RAM. The device turned into a 64-bit Windows OS, X64-primarily based total processor, and a 512-GB SSD. The running device turned into Windows 10, and the device turned into Jupyter labs, which used the Python programming language.

The facts set used from the Movielens and performed the facts preprocessing. Methods are used for fact cleansing, consisting of putting off superfluous attributes and filling in lacking values. Data exploration gives context for the facts set. The facts are then damaged down into education and check sets, and the advice engine is used to construct and educate a version on facts (Daniel 2020). After education, to make predictions and examine version overall performance primarily based totally on to be had metrics. The advice structures of K-Means Clustering and HNPCC are compared.

Statistical Analysis

The SPSS tool Version 22 (64-bit arch) was used to carry out statistical evaluation of the results for this study. Using SPSS, an independent sample T-Test analysis done among the 2 groups, with the CI adjusted to 95%, $p > 0.05$, and the error bars activated within the comparative evaluation. The independent variables had been User ID, Movie ID, Article ID, Movie Name, and Release Date. The dependent variables had been Genre, Rating, and Timestamp. An unbiased pattern taken a look at is used to examine the overall performance of the algorithms.

RESULTS

With a sample size of 30, the Hybrid Novel Pearson Correlation Coefficient and KMeans clustering algorithms were observed to be run at different times, and accuracy was calculated. In terms of accuracy, the HNPCC algorithm outperforms the KMC algorithm. The Independent Sample T-Test in Table 1 was used to compare the accuracy of HNPCC and KMC, and a statistically significant difference of P 0.05 with a 95 percent confidence level was observed, indicating that hypothesis is correct. The mean accuracy difference was calculated to be 5.4653.

The statistical analysis of 30 samples is shown in Table 2. The HNPCC algorithm produced 0.36 standard deviations with a standard error of 0.04, whereas the KMC algorithm produced 0.16 standard deviations with a standard error of 0.03. The hybrid-based HNPCC has an accuracy rate of 94.3%, while KMC has an accuracy rate of 89.7%. The significance level is set to $p<0.05$ with a 95% confidence interval, and the interference shows that HNPCC is more accurate than KMC with a significance of 0.001.

Figure 1 represents the simple mean bar graph that shows that the Standard deviation of hybrid-based HNPCC is lower than that of the KMeans Clustering algorithm.

DISCUSSION

This study looked at the Hybrid Novel Pearson Correlation Coefficient and KMeans clustering algorithms to predict the accuracy percentage of Movie recommendation systems. In terms of movie recommendation system accuracy (94.3%), Hybrid Novel Pearson Correlation Coefficient outperforms KMeans clustering (89.7%) with a significance of 0.001. The Novel Pearson correlation coefficient characteristic contributes to a better similarity index, which contributes to a growth in accuracy percentage. The consequences display a statistically giant distinction in clustering algorithms among Hybrid and KMeans.

The content that is used to collect prior information from users' accuracy of 87.8% was obtained using Machine Learning (Merki-Feld et al. 2021). The authors (Reddy et al. 2019), the system was implemented using clustering over a clustering approach. The researchers (Sandeep Kumar and Prabhu 2019) implemented a hybrid filtering approach in which the accuracy of 85.4% was obtained. The authors (Musa and Zhihong 2020) developed a system by using a recommendation filtering approach. In their system clustering is achieved among users and their preferences with the help of the Maximization Algorithm an accuracy of 80.7% was obtained. In the research article (Sadowski et al. 2021), implemented using a hybrid filtering approach in their paper Various techniques like indexing, similarity index based on Machine Learning, and classification are implemented in their system to result in better recommendation systems accuracy of 87.8% was obtained.

It has been established that the proposed HNPCC is greater correct than preceding studies articles discussed. The proposed model has an issue in that a real-time dataset with greater parameters might also additionally offer greater correct consequences in phrases of predicting accuracy. In the future, hoping to allow the advice gadget version into any internet or cell application. It will assist the consumer get hold of film guidelines in a greater consumer-pleasant manner.

CONCLUSION

In this work, a movie recommendation system was implemented by utilising two methods HNPCC and KMC. The KMC (89.7%) appears to have lesser accuracy than HNPCC (94.3%). Based on the statistical analysis, the proposed HNPCC performs significantly better than KMC with $p<0.001$.

DECLARATIONS

Conflict of interests

No conflict of interest in this manuscript

Authors Contributions

Author SM was involved in data collection, data analysis, and manuscript writing. Author KJ was involved in conceptualization, data validation, and critical review of the manuscript.

Acknowledgments

The authors would like to express their gratitude towards Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences (Formerly known as Saveetha University) for providing the necessary infrastructure to carry out this work successfully.

Funding

We thank the following organizations for providing financial support that enabled us to complete the study.

1. Soft Square Pvt. Ltd.
2. Saveetha University.
3. Saveetha Institute of Medical And Technical Sciences.
4. Saveetha School of Engineering.

REFERENCES

1. Appavu, Prabhu, Venkata Ramanan M, Jayaprabakar Jayaraman, and Harish Venu. 2021. "NOx Emission Reduction Techniques in Biodiesel-Fuelled CI Engine: A Review." *Australian Journal of Mechanical Engineering* 19 (2): 210–20.
2. Arun Prakash, V. R., J. Francis Xavier, G. Ramesh, T. Maridurai, K. Siva Kumar, and R. Blessing Sam Raj. 2020. "Mechanical, Thermal and Fatigue Behaviour of Surface-Treated Novel Caryota Urens Fibre–reinforced Epoxy Composite." *Biomass Conversion and Biorefinery*, August. https://doi.org/10.1007/s13399-020-00938-0.
3. Balachandar, Ramalingam, Logalakshmanan Baskaran, Ananthanarayanan Yuvaraj, Ramasundaram Thangaraj, Ramasamy Subbaiya, Balasubramani Ravindran, Soon Woong Chang, and Natchimuthu Karmegam. 2020. "Enriched Pressmud Vermicompost Production with Green Manure Plants Using Eudrilus Eugeniae." *Bioresource Technology* 299 (March): 122578.
4. Chen, Rong, Tianyi Yan, Yiannis Ventikos, and Jinghao Zhou. 2021. *Computational Methods for Translational Brain-Behavior Analysis*. Frontiers Media SA.
5. Daniel, Aschalew. 2020. "A Hybrid Movie Recommendation System Using Particle Swarm Optimization and K-Means Clustering Algorithm." ASTU. http://213.55.101.20:8080/xmlui/handle/123456789/1566.
6. Ezhilarasan, Devaraj, Thangavelu Lakshmi, Manoharan Subha, Veeraiyan Deepak Nallasamy, and Subramanian Raghunandhakumar. 2021. "The Ambiguous Role of Sirtuins in Head and Neck Squamous Cell Carcinoma." *Oral Diseases*, February. https://doi.org/10.1111/odi.13798.
7. Gopalakrishnan, R., V. M. Sounthararajan, A. Mohan, and M. Tholkapiyan. 2020. "The Strength and Durability of Fly Ash and Quarry Dust Light Weight Foam Concrete." *Materials Today: Proceedings* 22 (January): 1117–24.
8. Hannah R, Pratibha Ramani, WM Tilakaratne, Gheena Sukumaran, Abilasha Ramasubramanian, and Reshma Poothakulath Krishnan. 2021. "Author Response for 'Critical Appraisal of Different Triggering Pathways for the Pathobiology of Pemphigus vulgaris—A Review.'" Wiley. https://doi.org/10.1111/odi.13937/v2/response1.
9. Kavarthapu, Avinash, and Kaarthikeyan Gurumoorthy. 2021. "Linking Chronic Periodontitis and Oral Cancer: A Review." *Oral Oncology*, June, 105375.
10. Kumar, J. Sandeep, J. Sandeep Kumar, B. Rex, S. Irulandi, and S. Prabhu. 2019. "A Review on Diversity, Bio-Ecology, Floral Resources and Behavior of Blue Banded Bees." *International Journal of Current Microbiology and Applied Sciences*. https://doi.org/10.20546/ijcmas.2019.807.072.
11. Li, Runde, Jinshan Pan, Min He, Zechao Li, and Jinhui Tang. 2020. "Task-Oriented Network for Image Dehazing." *IEEE Transactions on Image Processing: A Publication of the IEEE Signal Processing Society*, May. https://doi.org/10.1109/TIP.2020.2991509.
12. Liu, Zenghui, Mengchao Xiao, Zhaofeng Du, Mengwan Li, Huiming Guo, Ming Yao, Xiaochun Wan, and Zhongwen Xie. 2020. "Dietary Supplementation of Huangshan Maofeng Green Tea

Preventing Hypertension of Older C57BL/6 Mice Induced by Desoxycorticosterone Acetate and Salt: Green Tea Preventing Senior Hypertension." *The Journal of Nutritional Biochemistry*, October, 108530.

13. Malik, V. P., Deepak Kapoor, V. K. Ahluwalia, P. M. Hariz, Anuraag Singh Rawat, P. K. Chakravorty, Anil Chopra, et al. 2019. *CLAWS Journal: Vol. 12 No. 1 (2019): Summer 2019*. IndraStra Global e-Journal Hosting Services.
14. Menon, Soumya, Happy Agarwal, S. Rajeshkumar, P. Jacquline Rosy, and Venkat Kumar Shanmugam. 2020. "Investigating the Antimicrobial Activities of the Biosynthesized Selenium Nanoparticles and Its Statistical Analysis." *BioNanoScience* 10 (1): 122–35.
15. Merki-Feld, Gabriele S., Peter S. Sandor, Rossella E. Nappi, Heiko Pohl, and Christoph Schankin. 2021. "Clinical Features of Migraine with Onset prior to or during Start of Combined Hormonal Contraception: A Prospective Cohort Study." *Acta Neurologica Belgica*, April. https://doi.org/10.1007/s13760-021-01677-3.
16. Musa, Jamilu Maaruf, and Xu Zhihong. 2020. "Item Based Collaborative Filtering Approach in Movie Recommendation System Using Different Similarity Measures." *Proceedings of the 2020 6th International Conference on Computer and Technology Applications*. https://doi.org/10.1145/3397125.3397148.
17. Muthukrishnan, Sivaprakash, Haribabu Krishnaswamy, Sathish Thanikodi, Dinesh Sundaresan, and Vijayan Venkatraman. 2020. "Support Vector Machine for Modelling and Simulation of Heat Exchangers." *Thermal Science* 24 (1 Part B): 499–503.
18. Priscilla, S., and C. Naveena. 2020. "Social Balance Theory Based Hybrid Movie Recommendation System." *Journal of Computational and Theoretical Nanoscience*. https://doi.org/10.1166/jctn.2020.9012.
19. Reddy, S. R. S., Sravani Nalluri, Subramanyam Kunisetti, S. Ashok, and B. Venkatesh. 2019. "Content-Based Movie Recommendation System Using Genre Correlation." In *Smart Intelligent Computing and Applications*, 391–97. Springer, Singapore.
20. Sadowski, Elizabeth A., Ali Pirasteh, Alan B. McMillan, Kathryn J. Fowler, and Joanna E. Kusmirek. 2021. "PET/MR Imaging in Gynecologic Cancer: Tips for Differentiating Normal Gynecologic Anatomy and Benign Pathology versus Cancer." *Abdominal Radiology (New York)*, October. https://doi.org/10.1007/s00261-021-03264-9.
21. Sandeep Kumar, M., and J. Prabhu. 2019. "Hybrid Model for Movie Recommendation System Using Fireflies and Fuzzy C-Means." *International Journal of Web Portals (IJWP)* 11 (2): 1–13.
22. Sarode, Sachin C., Shailesh Gondivkar, Gargi S. Sarode, Amol Gadbail, and Monal Yuwanati. 2021. "Hybrid Oral Potentially Malignant Disorder: A Neglected Fact in Oral Submucous Fibrosis." *Oral Oncology*, June, 105390.
23. Satapathy, Suresh Chandra, Yu-Dong Zhang, Vikrant Bhateja, and Ritanjali Majhi. 2020. *Intelligent Data Engineering and Analytics: Frontiers in Intelligent Computing: Theory and Applications (FICTA 2020), Volume 2*. Springer Nature.
24. Sekar, Durairaj, Deepak Nallaswamy, and Ganesh Lakshmanan. 2020. "Decoding the Functional Role of Long Noncoding RNAs (lncRNAs) in Hypertension Progression." *Hypertension Research: Official Journal of the Japanese Society of Hypertension*.
25. Singh, Sukhmin, Aman Verma, Aakriti Jain, Tarun Goyal, Pankaj Kandwal, and Shobha S. Arora. 2021. "Infection and Utilization Rates of Bone Allografts in a Hospital-Based Musculoskeletal Tissue Bank in North India." *Journal of Clinical Orthopaedics and Trauma* 23 (December): 101635.
26. Song, Han Soo, Dong Hwi Kim, Gwang Chul Lee, Kweon Young Kim, So Yeon Ryu, and Chul Gab Lee. 2020. "Work-Related Factors of Knee Osteoarthritis in Korean Farmers: A Cross-Sectional Study." *Annals of Occupational and Environmental Medicine* 32 (November): e37.

27. Thulaseedaran, N. K., K. G. Sajeeth Kumar, Jayesk Kumar, P. Geetha, N. V. Jayachandran, C. G. Kamalasanan, Sheela Mathew, and Shiji Pv. 2018. "A Case Series on the Recent Nipah Epidemic in Kerala." *The Journal of the Association of Physicians of India* 66 (10): 63–67.

28. Walek, Bogdan, and Vladimir Fojtik. 2020. "A Hybrid Recommender System for Recommending Relevant Movies Using an Expert System." *Expert Systems with Applications*. https://doi.org/10.1016/j.eswa.2020.113452.

TABLES AND FIGURES

Table 1. Independent Sample T-test Results with a confidence interval of 95% and level of significance of 0.05, HNPCC performs significantly better than KMC.

		Levene's Test for Equality of Variances		t-test for Equality of Means					
		F	Sig.	t	df	Sig. (2-tailed)	Mean Diff.	Std. Error Diff.	95% Confidence Interval
									Lower
Accuracy	Equal variances assumed	0.64	0.20	9.334	58	.001	5.46533	0.05	4.29
	Equal variances not assumed			9.334	57.563	.001	5.46533	0.05	4.29

Table 2. Statistical analysis of HNPCC and KMC. Mean accuracy value, Standard deviation and Standard Error Mean for HNPCC and KMC algorithms are obtained for 30 iterations. It is observed that the hybrid-based HNPCC algorithm performed better than the KMC algorithm.

Algorithm	N	Mean	Std. Deviation	Std. Error Mean
Accuracy Hybrid	30	94.3327	.36450	.43170
KMeansClustering(KMC)	30	89.7313	.16680	.39560

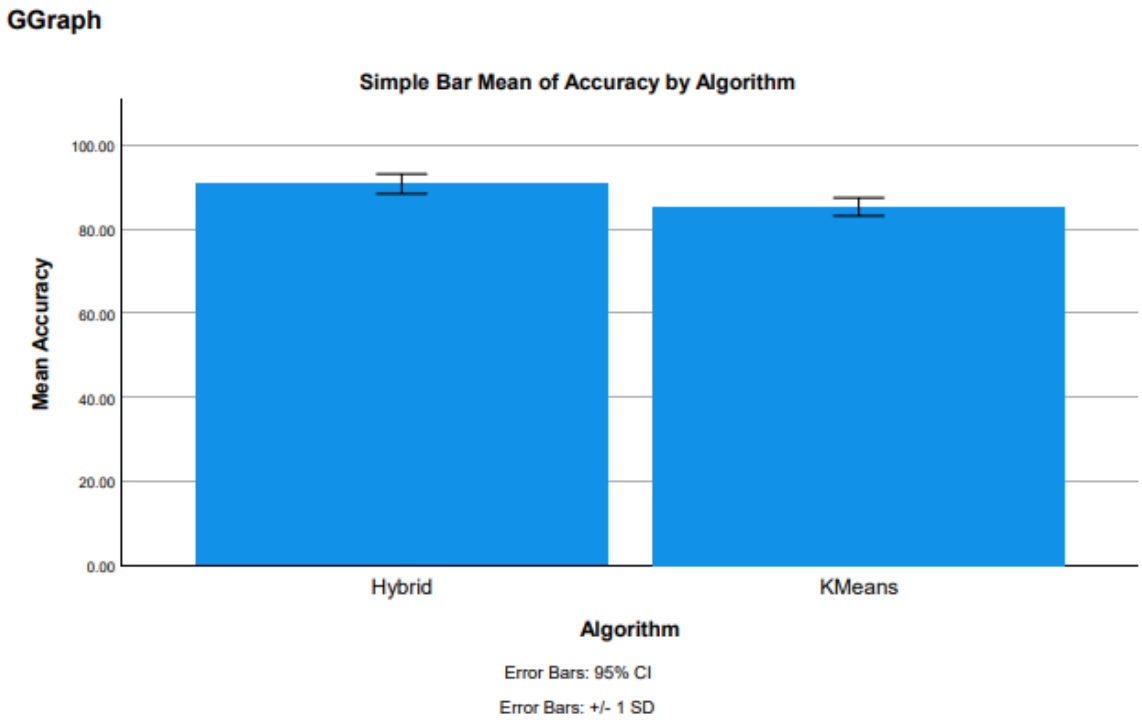

Fig. 1. Comparison of HNPCC algorithm and KMC in terms of mean accuracy. The mean accuracy of HNPCC was better than KMC and the standard deviation of Hybrid is slightly better than KMeans clustering. The X-axis represents HNPCC and KMC algorithm, Y-axis represents the mean accuracy of algorithms with +/- 1 SD.