# VARIOUS PROJECT DEVELOPMENT STAGES AND MOVIE RECOMMENDATION USING KNN*

BY

[1]TEJAL PATIL♣, [2]PROF. CHANDRASHEKHAR KUMBHAR

*[1]Student, School of Computer Science and Engineering, Ajeenkya D Y Patil University, Pune*

*tejal.patil@adypu.edu.in*

*[2]Assistant Professor, School of Computer Science and Engineering, Ajeenkya DY Patil University, Pune, India*

*chadrashekhar.k@ainurture.co.in*

**ABSTRACT**

*Developing a project is not that simple as it involves several steps for preparation with correct knowledge and right implementation. Here we will be dealing with some steps of project development as well as a real time project based on recommending movies using the algorithm KNN on the basis of their ratings. KNN generally retrieves the K data points which are nearest in distance. We have also used various steps which include Data mining, cleaning with different data sets and implemented everything in python that is Jupyter Notebook as it was compatible to use. Movie recommendation is a trending topic as there are many real time working scenarios and is a beneficial system for people who are using it as they get their related recommendations for any movies, books, online sales etc. It is not a new concept to be defined because even the retail shops use to recommend customers on the basis of the products that they buy.*

**KEYWORDS**

Stages of project development, KNN, Movie recommendation, Jupyter Notebook.

## 1Introduction

Project Development specifically in data science involves different steps like data set collection, data preparation, data cleaning, data mining and so on. It is important for any working project to achieve these steps for better results. Getting to learn some hands-on experience with these steps for real time project. We tried to implement a real time project which is Movie recommendation system and figures out to use KNN as an algorithm to deploy as it is based on ratings. The basic purpose of any recommendation system is to search interesting and some sort of relevant content to an individual with no time consumed. Recommendation is not only

* Received 06 October 2021, Accepted 25 October 2021, Published 11 November 2021
♣ Corresponding Author

about movies it have wide exposure with books, Music tracks, messages, articles, restaurants etc. There are many projects developed on movie recommendation systems but here in this paper we have used KNN i.e., K nearest neighbors. There are numerous real times working applications of this algorithm like in banking systems it can be used to check weather to give loan to this specific person or not by checking the defaulter's characteristics and comparing it for the purpose of analysis. Medical, text mining and Politics where you need to consider for a vote which will include either a yes or no for a vote. Here the Movie recommendation will generally work on the ratings given by people. The nearest best rated movie will be suggested. The working of KNN algorithm is that it will always suggest its nearest neighbors because it works in regards with the distance. It can be used for both Classification as well as regression problems. The recommender engines are high in demand in today's world of emerging technology as many businesses have their e commerce platform and would wish to have recommendation system so that it benefits for higher sales and customer satisfaction which is needed for any business to grow.

**1.1Approach**

There are various approaches used for the implementation of projects including steps to achieve them. Below are the following steps which are used for project development.

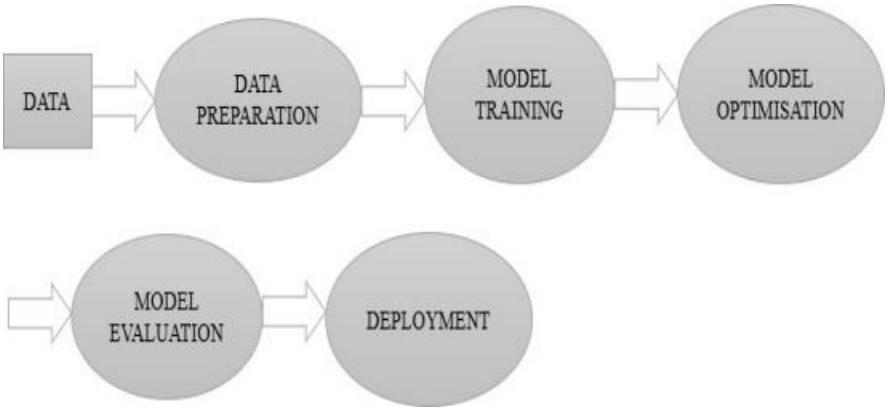


Fig. 1 Development stages of Project

1.Data preparation: In the preparation stage all the raw data is checked that whether they consist any errors or not. Here it will remove all the incomplete, redundant data.

2.Model training: It will formulate the data in a structured way as here the model will be trained as per the requirements.

3.Model optimization: It is the characteristic key in this process and will optimize the model.

4.Model Evaluation: It will help you to find the best suited model for our project.

5.Deployment: After the finalization we need to check the compatibility of the training model with the environment for the test model.

In Movie recommendation system the approach which we used to build is Machine learning which basically means it trains your machine/computer to learn things. It is a branch of Artificial Intelligence (AI). KNN algorithm formerly known as K nearest Neighbor is what we will be using in movie recommendation system based on ratings as it will find the nearest rating and suggest movies in reference with that It is a simple algorithm to deploy as it uses the entire data set for its training phase. KNN is also called as a non-parametric algorithm which is considered as one of its properties because it won't assume anything about the underlying data. Following the approach, we took machine learning algorithm which is a type of Supervised Learning which can handle both classification as well as regression problems but it is more compatible and used with classification problems for predictive analysis both Classification and regression are categorized under supervised learning. Classification in simpler words classifies the things like predicting a label or category whereas regression is to find the optimal solution for several continuous real values and make predictions. There are further many types of classification and regression naming some of them like Linear regression, Polynomial regression, Random forest algorithm, decision tree algorithm.

2Literature Review

Here, we will be reviewing some papers based on the topic that is the KNN algorithm and Movie recommendation system.

**2.1Literature Review table**

| Sr. No. | Title of Research Paper | Author | Methodology | Strengths | Drawback | Future Scope |
|---|---|---|---|---|---|---|
| 1 | Introduction to Machine Learning – KNN (2016) | Zhongheng Zhang | Explanation and working of KNN usingreal time example with R. | Explanation About KNN. Highlighting important features of algorithm | Nothing mentioned about the applications Of algorithm | Useful to understand basic KNN |
| 2 | Using KNN Algorithm for Classification of Textual Documents (2017) | Aiman Moldagulova<br><br>Rosnafisah Bte. Sulaiman | Used R as they have already implemented and will integrate classifier with Hadoop | KNN algorithm for text classificationas it is one ofthe best classifiers to use. | The technique used in not explained in detail. | Can classify the text using KNN as well as any other algorithm |

| | | | | | | |
|---|---|---|---|---|---|---|
| 3 | Movie recommend ation system through group-level sentiment analysis in microblogs (2016) | Hui Li, Jiangtao Cui | They have used collaborative filtering & built their own system | A proper research work and overviewof system is displayed | Description about the data mining techniques and algorithm isnot specifies | Can create more such recommenda tion systems |
| 4 | Movie Recommend er System Using K-Means Clustering AND K-Nearest Neighbor (2019) | Rishabh Ahuja, Arun Solanki, Anand Nayyar | They have used K means clustering and KNN algorithm to get the result | Detailed Information about the system including project development steps. | The proposed system could have been better. | Can add sentiment analysis as well. |
| 5 | A Survey of Recommend ation System: Research Challenges (2013) | Lalita Sharma, Anju Gera | It was a basic survey with brief introduction for several approaches for recommendati on systems | Briefing about various systems of recommendati on with important concepts. | No live working project/ example | Exploration of more such techniques. |

| Sr. No. | Title of ResearchPaper | Author | Methodology | Strengths | Drawback | Future Scope |
|---|---|---|---|---|---|---|
| 6 | A Movie Recommender System: MOVREC (2015) | Manoj Kumar, D.K. Yadav, Ankur Singh | Used basic K means algorithm | Real time project named Movrec. | Working of project is not shared. | Increase in recommendation systems |
| 7 | A Method to Improve the Accuracy of K-Nearest Neighbor Algorithm (2017) | Maryam Kuhkan | Used various methods to improve Accuracy ofKNN | The various classification methods are used to compare with KNN | Working of project is not shared. | More research can take place |

| | | | | | | |
|---|---|---|---|---|---|---|
| 8 | Personalized Research Paper Recommendation using Deep Learning (2017) | Hebatallah A. Mohamed Hassan | Data collection, Language modelling, Profile creationusing ANN | Research Goals are quite clear | Live working example is not there. | More such need of search engines for research papers. |
| 9 | Collaborative Filtering Techniques in Recommendation Systems (2019) | Sandeep K. Raghuwanshiand R. K.Pateriya | Created user based collaborative filtering | Description about collaborative filtering & challenges faced | Applications are not mentioned. | Can create on other techniques as well. |
| 10 | Storytime: Children for Book Recommendations (2019) | Ashlee Milton, Michael Green, AdamKeener, | They are in a phase to generate recommendation System for children. | Gathering and displaying recommend Tion. | No live working of project | Can help children to find books based on their interests. |
| 11 | An Improved Approach forMovie Recommendation System (2017) | Shreya Agrawal, Pooja Jain | To build recommendation using Hybrid approach | Used various algorithm to get desired solution | The result received was satisfactory | They can consider age ofthe user as well. |
| 12 | Trust Based Recommendation Systems (2013) | Makbule Gulcin Ozsoy and Faruk Polat | Content based approach is used | Included challenges and improvement. | No live working example | More reliable with recommenders as its trust based. |
| 13 | Speech emotion recognition basedon SVM and KNN | Mohammed Jawad, Ahmed Fatlawi | Detect the features and classify them using various algorithms. | Used PCA technique for reducing data. | No live working example | Much more needed technology with more advancements. |

| Sr. No. | Title of ResearchPaper | Author | Methodology | Strengths | Drawback | Future Scope |
|---|---|---|---|---|---|---|
| 14 | COVID-19 Patients Detection Strategy based on hybrid feature selection (2020) | Warda M. Shaban, Asmaa H. Rabie | Used Hybrid feature technique andKNN | EKNN Training | Challenges are not included | Can be used in a more advanced way in future. |

| | | | | | | |
|---|---|---|---|---|---|---|
| 15 | Sentiment Analysis on Movie ReviewData (2019) | Atiqur Rahman, Md. Sharif Hossen | Pre-processing of data and applied classification techniques. | Evaluates the opinion on movie reviewdata | No live working Example | Sentiment analysis will have more demand in future |

Literature Review Description

I.The R package class contains very useful function for the purpose of kNN machine learning algorithm [1]

The author discussed about K nearest neighbor algorithm with its working example of fruits and grains as well as live code demonstration in R programming with a dataset which actually had no meaning in it and used Kappa statistics to check the performance of KNN.

II.Another modification of KNN algorithm is a combination of eager learning with KNN classification [2]

In this paper a text classification was done with document using KNN as it is one of the best classifiers to use rather than naïve byes or any other having many other approaches to solve machine learning is always preferable as KNN is very easy to use.

III.The neighborhood-based approaches are the most popular prediction methods which are widely adopted in commercial collaborative filtering systems [3]

The authors in this paper have implemented a recommendation system, Kbridge which includes multiple information theories and works as online tv/ movie recommendation as well as ads and service recommendation. They even used data mining techniques for social network

IV.A recommendation system is a type of information filtering system which is used to predict the" rating" or" preference" a user would give to an item [4] In this paper the author has given brief explanation about various approaches of recommendation system like hybrid, collaborative and content based. They have implemented this using Python and various data implementation steps.

V.CF systems recommend an item to a user based on opinions of other users [5]

Here it is discussed about the various approaches like collaborative filtering, hybrid process etc. which is being used for recommendation system as well as problems faced like sparsity problem, cold problem and scalability which indirectly challenges the researcher about this.

VI.The original K-means algorithm was proposed by MacQueen [6]

The author Manoj Kumar have used K means algorithm to develop Movrec as well as had a research on Netflix where 2/3rd of movies watched are recommended whereas google

generates more than 38% click troughs. The working is explained in simple mathematical way with its data set description.

VII.Data mining includes the detection of valid, new, and understandable patterns in data sets; in other words, it is a process that extracts knowledge from data sets by using smart techniques [7]

In this paper there are various methods used like the naïve byes, J48 algorithm, LWL algorithm which are implemented on MATLAB and WEKA. In this a new algorithm was introduced based on their weighting to characteristic in order to improve accuracy.

VIII.Hybrid RSs use a combination of content based and collaborative filtering techniques [8]

The author Mohamed Hassan has defined the use of recommendation systems in research paper and is currently in their first stage of development and used deep learning as their methodology to use and have also given description about collaborative and hybrid filtering.

IX.The system should not repeatedly show popular items as this may also leads to reduction in user interest [9]

In this paper they have certainly explained briefly about collaborative filtering and its further types with all the Goals achieved and challenges faced and made two basic recommender systems (memory-based and model- based).

X.Recommendation strategies that most align with Storytime's include Rabbit (Readers' advisory-based book recommendation tool) [10]

The author Ashlee have presented story Time, a book recommender for children. Where they have given a system overview with its existing approach with preferred solutions regarding their recommender system which is developed in Python.

XI.For watching movies online, there are a number of movies to search in our most liked movies [11]

In this paper the authors have achieved their results by testing with other algorithms like Genetic algorithm, SVM and K means clustering and used various methods using Hybrid approach and have compared the results of proposed system and existing system.

XII.The goal of recommendation systems is suggesting a user the items that might be of interest for him/her [12]

Here there is a creation of trust-based recommender systems using reputation networks. There is even a trust-based agent recommender which gives recommendation of movies based on its trust relationships with those users as it uses only trust scores.

XIII.The creation of emotional models requires their careful consideration to compensate the effects of these variables [13]

The author has used KNN and SVM for speech recognition which is based on Feature extraction where the performed this using a database and extracted the features as well as classification algorithms like KNN, SVM and got the desired results.

XIV.KNN is a useful and rapid technique [14]

In this paper the authors have detected the Covid 19 patient's detection using the algorithm KNN because it is compatible and easy to use as well as they have used Hybrid feature selection with proper techniques using deep learning and neural network.

XV.Authors deal the view-level SA on ecommerce data [15]

The author Atiqur Rahman have used Sentiment analysis on movie recommendation as it classifies the people opinion as a positive or negative view. They used data cleaning by removing stop words, punctuation etc and used SVM for fixed dimensions of features.

**3Technology on which we are working**

The Technology we choose to work on is Python as it involves less time to code and very flexible to use and is one of the most popular and compatible language. Jupyter notebook is the application which we are currently working with. It is basically an app which provide client-server application which provide environment for python language. In this paper we have implemented a movie recommendation system using the algorithm KNN. This system is basically recommending the movies as per their popularity and ratings to a particular person. So firstly, we started to code in Jupyter notebook by importing the dataset of movies and ratings after that we moved ahead to code for reading the csv file being done with all this import stuff, we went with merging the two data sets and performing data exploration techniques and classifies the movies based on their popularity like the top 50 or top 100. We will even use pivot table in data exploration technique as it will show the total amount of ratings given per person. Then we will finally implement our algorithm KNN with brute force using the library SciPy sparse so that it will train the dataset and give the desired output by recommending the movies with nearest ratings between that parameter. For even achieving this project we used the project development stages to get our data trained and testified for this real time authentic working project. There are even various types used for recommending any stuff like the collaborative filtering, content-based recommendation etc.

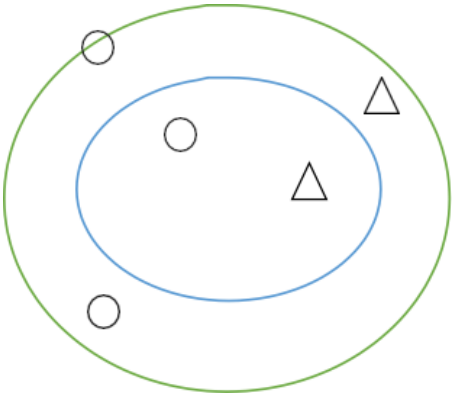

Fig 2. Working of KNN

The working of KNN is very simple and easy to use. Here it will first select the number K of its neighbor then moving ahead we will have to calculate the Euclidean distance of its K number of neighbors then we will take the nearest K neighbor as per the Euclidean distance calculated. We need to count the number of data points in each category so that we can give them new data points in that specific category for which their neighbor number would be maximum and finally the model is ready to use. KNN don't use any specialized training phase as it will use all the sample data set for classification/ regression and simply stores desired results in the memory. We need to do standardization and normalization before applying KNN algorithm and it is very sensitive to missing values and noisy data.

**4Architecture/ Code snapshots**

Firstly, we have downloaded the two different datasets from Kaggle. Com which was ratings and movies.

```
In [2]: import pandas as pd
        import numpy as np

In [3]: movies_df = pd.read_csv('movies.csv',usecols=['movieId','title'],
                                dtype={'moviesId':'int32','title':'str'})
        movies_df
```

Out[3]:

| | movieId | title |
|---|---|---|
| 0 | 1 | Toy Story (1995) |
| 1 | 2 | Jumanji (1995) |
| 2 | 3 | Grumpier Old Men (1995) |
| 3 | 4 | Waiting to Exhale (1995) |
| 4 | 5 | Father of the Bride Part II (1995) |
| ... | ... | ... |
| 9737 | 193581 | Black Butler: Book of the Atlantic (2017) |
| 9738 | 193583 | No Game No Life: Zero (2017) |
| 9739 | 193585 | Flint (2017) |
| 9740 | 193587 | Bungo Stray Dogs: Dead Apple (2018) |
| 9741 | 193609 | Andrew Dice Clay: Dice Rules (1991) |

9742 rows × 2 columns

Sample Code 1

In the above code we have imported the data set from the csv file.

```
model_knn = NearestNeighbors(metric = 'cosine', algorithm = 'brute')
model_knn.fit(movie_features_df_matrix)

Out[19]: NearestNeighbors(algorithm='brute', metric='cosine')

In [20]: query_index = np.random.choice(movie_features_df.shape[0])
print(query_index)
distances, indices = model_knn.kneighbors(movie_features_df.iloc[query_index,:].values.reshape(1, -1), n_neighbors = 10)

239

In [21]: for i in range(0, len(distances.flatten())):
    if i == 0:
        print('Recommendations for {0}:\n'.format(movie_features_df.index[query_index]))
    else:
        print('{0}: {1}, with distance of {2}:'.format(i, movie_features_df.index[indices.flatten()[i]], distances.flatten()[i]))

Recommendations for Last Samurai, The (2003):

1: Star Wars: Episode III - Revenge of the Sith (2005), with distance of 0.4267086386680603:
2: War of the Worlds (2005), with distance of 0.42777353525161743:
3: X-Men: The Last Stand (2006), with distance of 0.4416844844818115:
4: Minority Report (2002), with distance of 0.44474202394485474:
5: Lord of the Rings: The Return of the King, The (2003), with distance of 0.4491521120071411:
6: Lord of the Rings: The Two Towers, The (2002), with distance of 0.4513309001922607:
7: Spider-Man 2 (2004), with distance of 0.4561489224433899:
8: Star Wars: Episode II - Attack of the Clones (2002), with distance of 0.4622541069984436:
9: Sin City (2005), with distance of 0.4691777229309082:
```

Output 1

Here in the above output code, we have applied the KNN algorithm and we got the desired output as it's showing the recommendation for the movie last samurai.

**5Survey Based on Technology with Results**

Here, in this paper we took a survey based on our technology which we used that is movie recommendation system which works on the basis of ratings received. The survey we took was through google forms which consisted the set of questionnaires like their personal details which included Name, Email id, age, profession and Phone no regardless we also asked do they prefer watching movies, if yes then what kind of movies they would like to watch, do they prefer watching movies based on their ratings and popularity, which recommendation system/engine they like/ prefer the most, does a movie rating matter to them and any suggestions they want to give in regards with changes they expect ? This survey was simply taken to check the response from the people of different age groups to check what are their expectations towards this technology. This form was circulated and got amazing responses from the public.

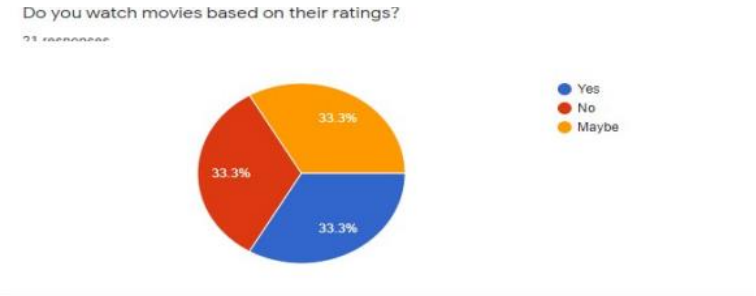


Fig 3. Survey outcome

Getting a good number of responses, the majority age group we got was from 20 to 25 years as well as 15 to 20 years with 38% of it where 90% of them were students. As we can see the figure 3, we got 33% of people interested in watching the movies based on ratings whereas the rest 33% is not while the others have no idea about it. That's why we considered to build a movie recommendation model based on ratings system. People preferred YouTube as their

most used/liked recommender system. We also received the highest percentage of people with 66% watching movies based on their popularity which indirectly requires a good rating. Therefore, our overall survey results helped us a lot to achieve our project in a real time scenario.

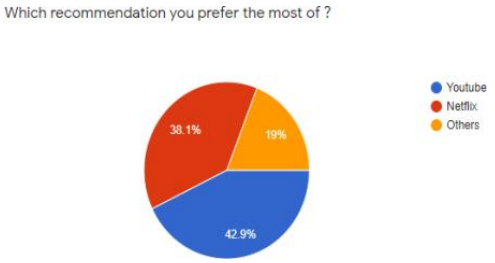


Fig 4. Survey outcome

**6Future Trends**

The forthcoming of Movie recommendation is extending day by day as its increasing demand in today's world because people definitely prefer their own specific choice rather than anything. It's not just only in movie but it has also spread more widely in the fields of E-commerce industries. There are many such platforms which uses movie recommendation system/ engine like the Netflix, Amazon prime, Hotstar etc. There are numerous use cases for social media platforms like Instagram and You tube after all this it has even shifted to online stores like Amazon, Myntra, Flipkart where people get the suggestion/ recommendation based on the product they buy or they frequently search for.

**7Conclusion**

Using the KNN algorithm we built a recommendation system which will recommend the movies based on their ratings and get the desired results. Recommendation has been a very important part in a human life because every person has their different choices and perspective. Even the small retail stores or any shopkeeper will recommend their costumers on the basis of what they are buying. With the growing need of recommendations there will be need for such kind off projects. Data plays the key role here because all these predictions, these systems are based upon the data provided. This is the reason why data has been given so much of importance in the last few years.

**References**

1.Zhang, Z., 2016. Introduction to machine learning: k-nearest neighbors. Annals of translational medicine, 4(11).

2.Moldagulova, A. and Sulaiman, R.B., 2017, May. Using KNN algorithm for classification of textual documents. In 2017 8th Conference on Information Technology (pp. 665-671). IEEE.

3.Li, H., Cui, J., Shen, B. and Ma, J., 2016. An intelligent movie recommendation system through group-level sentiment analysis in microblogs. Neurocomputing, 210, pp.164-173.

4.Ahuja, R., Solanki, A. and Nayyar, A., 2019, January. Movie recommender system using K-Means clustering and K-Nearest Neighbor. In 2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence) (pp. 263-268). IEEE.

5.Sharma, L. and Gera, A., 2013. A survey of recommendation system: Research challenges. International Journal of Engineering Trends and Technology (IJETT), 4(5), pp.1989-1992.

6. Kumar, M., Yadav, D.K., Singh, A. and Gupta, V.K., 2015. A movie recommender system: Movrec. International Journal of Computer Applications, 124(3).

7.Kuhkan, M., 2016. A method to improve the accuracy of k-nearest neighbor algorithm. International Journal of Computer Engineering and Information Technology, 8(6), p.90.

8.Hassan, H.A.M., 2017, July. Personalized research paper recommendation using deep learning. In Proceedings of the 25th conference on user modeling, adaptation and personalization (pp. 327-330).

9.Raghuwanshi, S.K. and Pateriya, R.K., 2019. Collaborative filtering techniques in recommendation systems. Engineering and Applications (pp. 11-21). Springer, Singapore.

10.Milton, A., Green, M., Keener, A., Ames, J., Ekstrand, M.D. and Pera, M.S., 2019, September. StoryTime: Eliciting preferences from children for book recommendations. In Proceedings of the 13th ACM Conference on Recommender Systems (pp. 544-545).

11.Agrawal, S. and Jain, P., 2017, February. An improved approach for movie recommendation system. In 2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC) (pp. 336-342). IEEE.

12.Ozsoy, M.G. and Polat, F., 2013, August. Trust based recommendation systems. In Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (pp. 1267-1274).

13.Al Dujaili, M.J., Ebrahimi-Moghadam, A. and Fatlawi, A., 2021. Speech emotion recognition based on SVM and KNN classifications fusion. International Journal of Electrical and Computer Engineering, 11(2), p.1259.

14.Shaban, W.M., Rabie, A.H., Saleh, A.I. and Abo-Elsoud, M.A., 2020. A new COVID-19 Patients Detection Strategy (CPDS) based on hybrid feature selection and enhanced KNN classifier. Knowledge-Based Systems, 205, p.106270.

15.Rahman, A. and Hossen, M.S., 2019, September. Sentiment analysis on movie review data using machine learning approach. In 2019 International Conference on Bangla Speech and Language Processing (ICBSLP) (pp. 1-4). IEEE.