e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:04/April-2022 Impact Factor- 6.752 www.irjmets.com

MOVIE RECOMMENDATION SYSTEM USING MACHINE LEARNING

Ojas Jawale[*1], Ganesh Senaiyer[*2], Anirban Bhattacharya[*3], Aditi Jha[*4]

[*1,2,3,4]Ramrao Adik Institute Of Technology, Navi Mumbai, Maharashtra, India.

ABSTRACT

A recommendation engine filters the information mistreatment totally different algorithms and recommends the foremost relevant things to users. It 1st captures the past behaviour of a client and supported that, recommends product that the users can be seemingly to shop for. If a totally new user visits an e-commerce website, that website won't have any past history of that user. therefore however will the positioning approach advocating product to the user in such a scenario? One attainable answer might be to recommend the popular product, i.e. the product that arr high in demand.

Another attainable answer might be to advocate the product which might bring the most profit to the business. 3 main approaches are used for our recommender systems. One is Demographic Filtering i.e they provide generalized recommendations to each user, supported picture show quality and/or genre. The System recommends identical movies to users with similar demographic options. Since every user is totally different, this approach is taken into account to be too straightforward.

The basic plan behind this technique is that movies that ar a lot of common and critically acclaimed can have the next likelihood of being likeable by the common audience. Second is content-based filtering, wherever we have a tendency to try and profile the user's interests mistreatment data collected, and advocate things supported that profile. the opposite is cooperative filtering, wherever we have a tendency to try and cluster similar users along and use data regarding the cluster to create recommendations to the user.

I. INTRODUCTION

Recommender systems used in a various form of areas together with movies, music, news, books, analysis articles, search queries, social tags, and merchandise normally. Recommendation System is a filtration program whose prime goal is to predict the movie to a user towards a domain-specific item.

In our case, this domain-specific item is a movie, so the most focus of our recommendation system is to filter and predict solely those movies that a user would favor given some information concerning the user him or herself. There are many alternative ways that to create movie recommendation system however we've selected the content base recommender system in order that user will simply get the foremost similar movies on the user's interest. As our recommender system recommends the top high five movies as like movie that user is selected.

II. LITERATURE REVIEW

There were some studies that were dispensed on the subject associated with our project.

MOVREC may be a moving picture recommendation system conferred by D.K. Yadav et al. supported cooperative filtering approach. cooperative filtering makes use of knowledge provided by user. That info is analysed and a moving picture is usually recommended to the users that ar organized with the moving picture with highest rating initial [1]

Luis M Capos et al has analysed 2 ancient recommender systems i.e. content based mostly filtering and cooperative filtering. As each of them have their own drawbacks he planned a replacement system that may be a combination of Bayesian network and cooperative filtering.[2]

This paper so presents a replacement Bayesian network model to subsume the matter of hybrid recommendation by combining content-based and cooperative options. it's been tailored to the matter in hand and is supplied with a versatile topology. [3]

The user specific info or item specific info is clubbed to make a cluster by Utkarsh Gupta et al. victimisation chameleon. this can be associate degree economical technique supported class-conscious cluster for recommender system. To predict the rating of associate degree item legal system is employed. The planned system has lower error and has higher cluster of comparable things. [4]

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:04/April-2022 Impact Factor- 6.752 www.irjmets.com

III. DATA MINING PROCESS

The basic architecture of the machine learning is as follows:

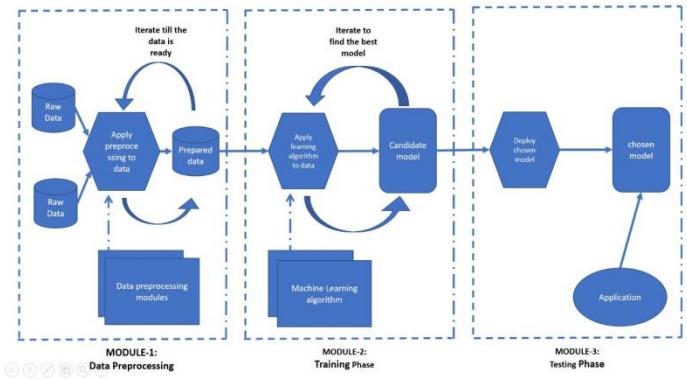


Figure 1: Machine Learning Architecture

The methodology of the proposed system comprises of the following steps:

- Data-set Collection.
- Pre-Processing.
- Training and Validating.
- Prediction.
- Result.

Let's see some brief description on the above-mentioned topics:

A. Data-set collection

The most root part of machine learning process is dataset. So, here we have a valid data-set of 5000 Hollywood movies with different information of the movies.

B. Pre-processing

In machine learning, we use "pandas" and "numpy" libraries for pre-processing purpose.

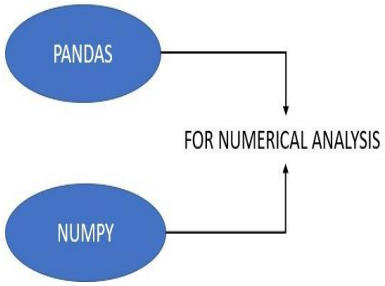


Figure 2: Machine Learning Architecture

NUMPY: NumPy stands for 'Numerical Python' or 'Numeric Python'. it's Associate in Nursing ASCII text file module of Python that provides quick mathematical computation on arrays and matrices. Numpy will be foreign into the notebook using:

>>>import numpy as np

PANDAS: Pandas is one in every of the foremost wide used python libraries in information science. It provides superior, straightforward to use structures and information analysis tools. Hence, with 2nd tables, pandas square measure capable of providing several further functionalities like making pivot tables, computing columns supported different columns and plotting graphs. Pandas will be foreign into Python using:

>>>import pandas as pd

Choosing a model and strategy is extremely vital method wherever in we've victimisation 2 libraries and machine learning techniques specifically Scikit Learn, NLTK (Natural Language Toolkit), and victimisation formula referred to as circular function similarity.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:04/April-2022 Impact Factor- 6.752 www.irjmets.com

SCI-KIT LEARN: Scikit-learn (Sklearn) is that the most helpful and sturdy library for machine learning in Python. It provides a variety of economical tools for machine learning and applied math modeling as well as classification, regression, clump and spatial property reduction via a consistence interface in Python. This library, that is basically written in Python, is made upon NumPy, SciPy and Matplotlib.Rather than that specialize in loading, manipulating and summarising information, Scikit-learn library is targeted on modeling the info. Stop words area unit simply a listing of words you don't wish to use as options. you'll be able to set the parameter stop words='english' to use a integral list. Alternatively, you'll be able to set stop words adequate to some custom list. This parameter defaults to none.

NLTK: NLTK (Natural Language Toolkit) Library could be a suite that contains libraries and programs for applied math language process. it's one in all the foremost powerful NLP libraries, that contains packages to form machines perceive human language associated reply to that with an acceptable response. Stemming and Lemmatization in Python NLTK square measure text standardization techniques for language process. These techniques square measure wide used for text preprocessing. The distinction between stemming and lemmatization is that stemming is quicker because it cuts words while not knowing the context, whereas lemmatization is slower because it is aware of the context of words before process. Stemming could be a methodology of standardization of words in language process. It is a way during which a collection of words in a very sentence square measure born-again into a sequence to shorten its search. during this methodology, the words having identical which means however have some variations consistent with the context or sentence square measure normalized.

```
import nltk

from nltk.stem.porter import PorterStemmer
ps = PorterStemmer()

def stem(text):
    y = []

    for i in text.split():
        y.append(ps.stem(i))

    string = " ".join(y)
```

Stemming and Lemmatization in Python NLTK area unit text normalisation techniques for language process. These techniques area unit wide used for text preprocessing. The distinction between stemming and lemmatization is that stemming is quicker because it cuts words while not knowing the context, whereas lemmatization is slower because it is aware of the context of words before process. Stemming could be a methodology of normalisation of words in language process. it's a method within which a collection of words in an exceedingly sentence area unit reborn into a sequence to shorten its operation. during this methodology, the words having an equivalent which means however have some variations in step with the context or sentence area unit normalized.

C. Training and Validation

Now, the processed data are stored in ".csv " file for further use. The processed data-set is divided into two parts :

- Training.(70 % of the data-set is used)
- Testing.(30 % of the data-set is used)

Now, comes the training part of the models. So, classification models are trained and tested to get the accuracy of the models. Once done with the accuracy part, we need to perform validation for further efficiency of the project.

D. Prediction

The presentation of algorithm associated based on accuracy and performance analysis and will provide a suggestion for the movies to the user whether movies are suggested or not upon user's interest.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:04/April-2022 Impact Factor- 6.752 www.irjmets.com

```
def recommend(movie):
    movie_index = new_df[new_df['title'] == movie].index[0]
    distances = similarity[movie_index]
    movies_list = sorted(list(enumerate(distances)),reverse=True,key=lambda x:x[1])[1:6]

    for i in movies_list:
        print(new_df.iloc[i[0]].title)
```

```
recommend('Avatar')
```

```
Aliens vs Predator: Requiem
Aliens
Falcon Rising
Independence Day
Titan A.E.
```

E. Result

The final result gives the recommendation of the movies.

IV. ALGORITHMS

Some of the algorithms used in movie recommendation are COUNT VECTORIZER AND COSINE SIMILARITY.

A. COUNT VECTORIZER

In order to use matter information for prophetic modelling, the text should be parsed to get rid of sure words – this method is termed tokenization. These words got to then be encoded as integers, or floating-point values, to be used as inputs in machine learning algorithms. This method is termed feature extraction (or vectorization).

Scikit-learn's CountVectorizer is employed to convert a set of text documents to a vector of term/token counts. It conjointly permits the pre-processing of text information before generating the vector illustration. This practicality makes it a extremely versatile feature illustration module for text.

e.g-

text = ['Hello my name is james, this is my jupyter notebook']

The text is transformed to a sparse matrix as shown below.

	hello	is	james	my	name	notebook	python	this
0	1	2	1	2	1	1	1	1

Count vectorizer makes it easy for text data to be used directly in machine learning and deep learning models such as text classification.

```
1 X.toarray()
```

```
array([[0, 0, 0, 1, 1, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0],
       [1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 1, 0, 0],
       [0, 1, 1, 0, 1, 1, 1, 0, 0, 0, 0, 1, 1, 1, 0, 1, 1]])
```

Text Vectorization is that the method of changing text into numerical illustration. Vectorization is jargon for a classic approach of changing computer file from its raw format (i.e. text) into vectors of real numbers that is that the format that millilitre models support. Here we have a tendency to area unit mistreatment Bag of Words technique to convert text to vectors.

B. COSINE SIMILARITY

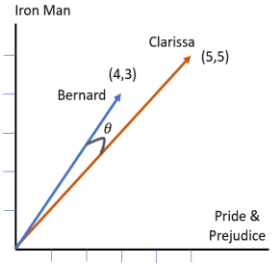
The circular function Similarity mensuration begins by finding the circular function of the 2 non-zero vectors. The output can manufacture a worth starting from -1 to one, indicating similarity wherever -1 is non-similar, zero is orthogonal (perpendicular), and one represents total similarity If 2 vectors area unit diametrically opposed, that means they're familiarised in mere opposite directions, then the similarity mensuration is -1. circular function Similarity is employed in positive area, between the bounds zero and one. circular function Similarity isn't involved, and doesn't live, variations is magnitude (length), and is simply a illustration of similarities in orientation. The library contains each procedures and functions to calculate similarity between sets of knowledge. The operate is best used once calculative the similarity between little numbers of sets. The procedures lay the computation and area unit thus additional acceptable for computing similarities onlargerdatasets.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:04/April-2022 Impact Factor- 6.752 www.irjmets.com

$$similarity = \cos\theta = \frac{b.c}{\|b\|\|c\|}$$

$b.c \Rightarrow Is\ the\ Dot\ product\ of\ the\ two\ vectors$

$\|b\|\|c\| \Rightarrow Is\ the\ product\ of\ each\ vector's\ magnitude$



Calculating:

$b.c = \sum_{i=1}^{n} b_i c_i = (4 \times 5) + (3 \times 5) = 35$

$\|b\| = \sqrt{4^2 + 3^2} = 5$

$\|c\| = \sqrt{5^2 + 5^2} = 5\sqrt{2}$

$similarity = \frac{35}{5 \times 5\sqrt{2}} \sim 0.989$

Theoretically, the perform circular function similarity is any range between -1 and +1 as a result of the image of the cos function, however during this case, there'll not be any negative picture show rating therefore the therefore the are going to be between zero^{o} and 90^{o} bounding the cos similarity between 0 and one. If the angle θ = 0^{o} =>cosine similarity = one, if θ = 90^{o} => cos similarity =0.

C. Cross Validation

In machine learning, we tend to couldn't work the model on the coaching information and can't say that the model can work accurately for the important information. For this, we tend to should assure that our model got the right patterns from the info, and it's not obtaining up an excessive amount of noise. For this purpose, we tend to use the cross-validation technique.

Cross-validation may be a technique within which we tend to train our model exploitation the set of the data-set and so value exploitation the complementary set of the data-set.

The 3 steps concerned in cross-validation square measure as follows :

Reserve some portion of sample data-set.

- Using the rest data-set train the model.
- Test the model using the reserve portion of the data-set.

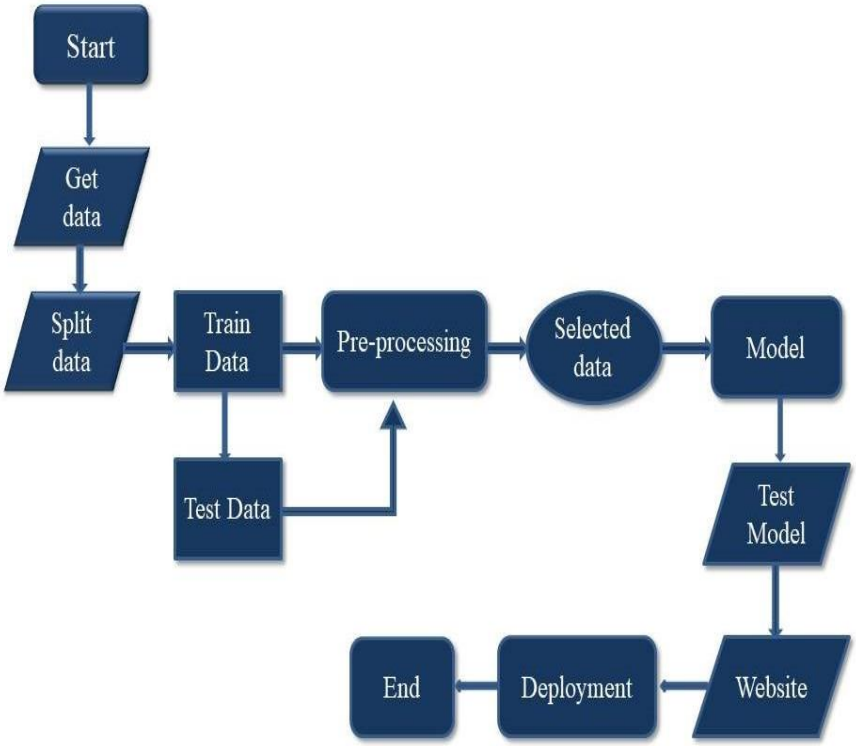
V. PROPOSED METHODOLOGY

The methodology of the project is meant in six steps:

- Installing the Python and SciPy platform. we want to mount our ".ipynb" file on our google drive for more access.
- Loading the dataset. The dataset of picture show recommendation is required to be foreign in ".csv" format.
- Summarizing the dataset. Sorting and improvement of knowledge is that the necessary method to extend the potency of the
- project. we are able to fill the missing information victimisation "imputer" perform.
- Visualizing the dataset. we are able to visualize our "tmdb_5000_movies.csv" and "tmdb_5000_credits dataset through the Kaggle.com and so pre process method thereon.
- Evaluating some algorithms. when visualising the dataset, currently comes coaching and testing part!!! Let's divide {the information|the info|the information} into 7:3 magnitude relation wherever seventieth data are trained and half-hour are tested. Now, let's choose the suitable models and so train them to urge the accuracy of the prediction. we've got used two models: COUNT VECTORIZER AND cos SIMILARITY. when obtaining the accuracy of every model and scrutiny them, lets cross Making some predictions. Now , comes the last stage of the project, i.e., to form predictions. Here, user will manually provide the input and acquire the advice of flick as per his/her interest.
- For content-based recommender system specifically, we have a tendency to conceive to notice a brand new thanks to improve the accuracy of the representative of the flick and suggest high 5 similar flicks to the user as per the interest of movie. Now, to form the project additional easy, we've got designed a frontend as well!!

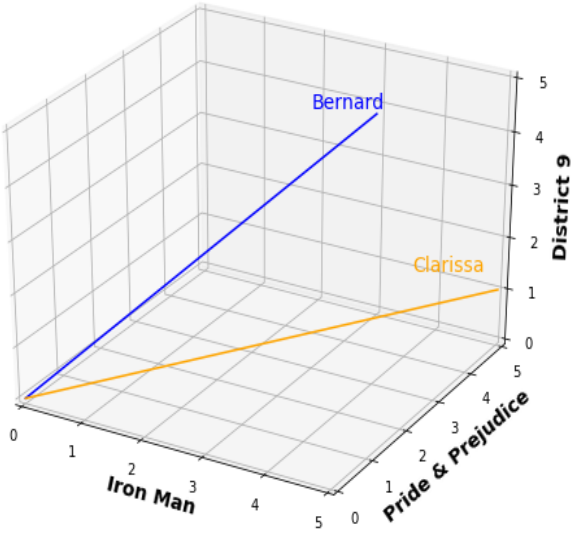
e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:04/April-2022 Impact Factor- 6.752 www.irjmets.com

The face consists of a web site with functions particularly recommendation and show. The face will be created victimisation flask module in python and preparation victimisation Heroku to link.

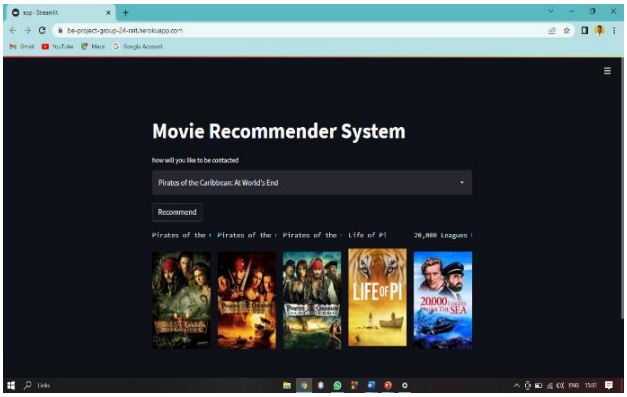


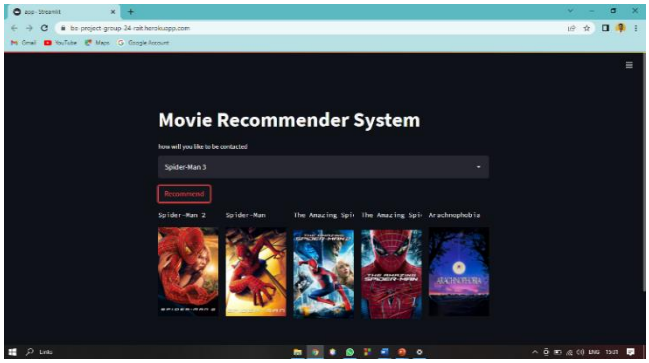
VI. RESULT ANALYSIS

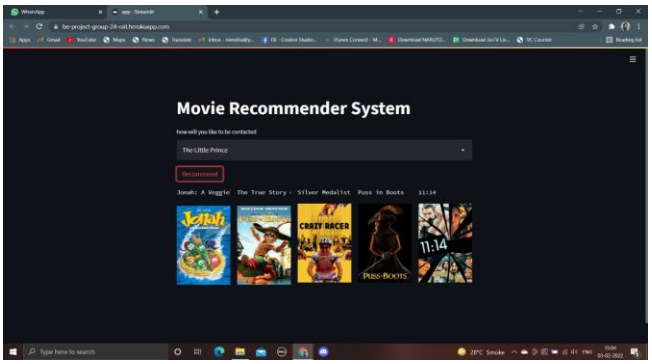
With the model trained, it must be tested to check if it'd operate well in planet things. that's why the a part of the info set created for analysis is employed to examine the model's proficiency. This puts the model during a state of affairs wherever it encounters things that weren't a district of its coaching. In short, for analysis purpose we tend to are mistreatment out tested knowledge and model to verify whether or not model is functioning fine or not. Machine learning is mistreatment knowledge to answer queries. therefore recommendation, or reasoning, is that the step wherever we tend to get to answer some queries. this is often the purpose of all this work, wherever the worth of machine learning is complete. we are able to finally use our model to predict whether or not the similar motion picture is usually recommended to the user or not as per his/her interest , supported the similarity of the films.



e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:04/April-2022 Impact Factor- 6.752 www.irjmets.com







VII. CONCLUSION

The main motivation of creating this project is to spice up every ., in order that we are able to perform our day-after-day of the movie, which are diversity and unique. We have successfully got the output of top high five recommended movies as the user in selected by it's choice. We develop the movie recommendation model using the machine learning and algorithms.

Hence, our project "Movie recommendation system" is justified.

VIII. REFERENCES

[1] D.K.Yadav. A movie recommender system. 2000(1):012101, 2017.

[2] Hongli Lin, Xuedong Yang, and Weisheng Wang. A content-boosted collaborative filtering algorithm for personalized training in interpretation of radiological imaging. Journal of digital imaging, 27(4):449–456, 2014.

[3] Harpreet Kaur Virk, Er Maninder Singh, and A Singh. Analysis and design of hybrid online movie recommender system. International Journal of Innovations in Engineering and Technology (IJIET) Volume, 5, 2015.

[4] Urszula Ku zelewska. Recommendation system engines. Iranian Journal of Energy and Environment, 2019.

[5] Hongli Lin, Xuedong Yang, and Weisheng Wang. A content-boosted collaborative filtering algorithm for personalized training in interpretation of radiological imaging. Journal of digital imaging, 27(4):449–456, 2014.