KIET Group of Institutions, Ghaziabad

Information Technology

Internship Report

On

Movie Recommendation System

Using

Machine learning

2022

Summer internship at KIET Group of Institutions

Minor Specialization

Submitted by:

Abhishree Bisht (2100290130008)

Saumya Srivastava (2100290130156)

Geetika Verma (2100290100067)

INDEX

ACKNOWLEDGEMENT

We have got this golden opportunity to express our kind gratitude and sincere thanks to our Head of Institution , KIET Group of Institutions, Ghaziabad for their kind support and necessary counselling in the preparation of the project report .We are also indebted to each and every person responsible for the making up of this project directly or indirectly.

We must also acknowledge our deep debt of gratitude towards each one of our colleague who led this project come out in the way it is. It's our hard work and untiring sincere efforts and cooperation to bring out the project work. Last but not the least , we would like to thank our parents for their support .They have always inspired us and kept our spirit up.

OUR TEAM:

Abhishree Bisht IT (2100290130008)

Saumya Srivastava IT (2100290130156)

Geetika Verma CSE (2100290100067)

SEMESTER: Third

CERTIFICATE

This to certify that the internship project (minor specialization) report entitled Movie recommendation system using ML submitted by Abhishree Bisht, Saumya Srivastava & Geetika Verma of 2nd year in the Department of KIET Group of Institutions , Ghaziabad , affiliated to Dr. A. P. J Abdul Kalam Technical University , Lucknow , Uttar Pradesh ,India, is a record of candidate summer internship under my supervision and guidance and are worthy of consideration for the same.

Signature of Supervisor:

Supervisor's Name: Mr. Amar Singh

Date:

ABSTRACT

Due to the complex nature of Movie recommendation system, it has been a trending area of interest. It has never been easy to predict Song with higher accuracy. Machine learning, which consists of making computers perform tasks that normally requiring human intelligence is currently the dominant trend in scientific research. This paper presents application of machine learning regression based algorithms, K nearest neighbour (KNN) in the prediction of movie based on different-different users by impending different classifiers. Furthermore, the results of the ML system application are given on an accuracy basis. Currently most of these types of recommendation system are made using existing systems like KNN and support vector regression. Machine leaning system provides a user friendly web application based interface for the user. Each machine learning algorithm is first trained with proper historic dataset regarding the artist, song and genera, user interest etc. given as input and is tested against new data. Python programming language is used to make suitable algorithms for the system on platforms like Google colab and Anaconda jupyter notebook.

INTRODUCTION

KNN(K-NEAREST NEIGHBOUR)

K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique. KNN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using KNN algorithm. KNN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems. KNN is a non-parametric algorithm, which means it does not make any assumption on underlying data.

It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.

KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.

Example: Suppose, we have an image of a creature that looks similar to cat and dog, but we want to know either it is a cat or dog. So for this identification, we can use the KNN algorithm, as it works on a similarity measure. Our KNN model will find the similar features of the new data set to the cats and dogs images and based on the most similar features it will put it in either cat or dog category.

FLASK

Flask is a web application framework written in python, in simple terms it helps end users interact with your python code (in this case our ML models) directly from their web browser without needing any libraries, code files, etc. Flask is a Python-based micro framework used for developing small-scale websites.

Flask is very easy to make Restful APIs using python. As of now, we have developed a model i.e. model.pkl , which can predict a class of the data based on various attributes of the data.

Details of task

Our project aimed to develop a model based on machine learning by implementing KNN. It requires us to collect data for training the algorithm and then testing it. Real time data was required because new movies are being released everyday and one wrong entry would have had the assumption go completely wrong and off the grid, which may in turn have caused huge loses for investors. This is the reason why predefined data is avoided in case of subject which tend to change with every second. For better visualization it is first checked that the data is cleaned and every value is clearly located and has a acceptable value. It is done by using data frame set and other tools in pandas libraries. The data gathered is then trained to cleaned of all the NaN and Null values. The data is then visualized using a matplot lib or the seaborn libraries in python libraries. The visualization is usually done in form of graphs and histograms for better understanding by everybody who uses the model.

Details of technical learning

In this project many technical aspects of machine learning and python programming were understood by us and it is tried by us to put them to a good and efficient use. Various libraries in python like numpy, pandas, matplot lib ,seaborn, etc. were understood and developed by us for a better understanding of movie recommendation data collected from Kaggle.

KNN is used for completion of this project. Movie is predicated by the algorithm and is visualized by plotting a graph between valid limits and the historic daily returns of the previous days.

Conclusion

The efficiency of the KNN model of Movie prediction using machine learning is found to be much more efficient, accurate and precise. It can also be made in form of a web application for better user interaction and is also able to provide results over multivariate time depending data.

Literature Review Report

1.Machine Learning Algorithms - A Review

Machine learning (ML) is the scientific study of algorithms and statistical models that computer systems use to perform a specific task without being explicitly programmed. These algorithms are used for various purposes like data mining, image processing, predictive analytics, etc. to name a few. The main advantage of using machine learning is that, once an algorithm learns what to do with data, it can do its work automatically. Machine Learning can be a Supervised or Unsupervised. If you have lesser amount of data and clearly labelled data for training, opt for Supervised Learning. Unsupervised Learning would generally give better performance and results for large data sets. If you have a huge data set easily available, go for deep learning techniques. This paper also helps you understand Reinforcement Learning and Deep Reinforcement Learning. It tells about Neural Networks, their applications and limitations. This paper surveys various machine learning algorithms. Machine learning is used from getting a recommended product in online shopping to updating photos in social networking sites. This paper introduces most of the popular machine learning algorithms.

2.Machine Learning in Agriculture: A Review

In this paper, a comprehensive review of research dedicated to applications of machine learning in agricultural production systems is done. The works analysed were categorized in (a) crop management, including applications on yield prediction, disease detection, weed detection crop quality, and species

recognition; (b) livestock management, including applications on animal welfare and livestock production; (c) water management; and (d) soil management.

From the analysis of these articles, it was found that eight ML models have been implemented in total where the most popular models were ANNs (with most. In livestock management category, four ML models were implemented, with most popular models being SVMs. For water management in particular evapotranspiration estimation, two ML models were implemented and the most frequently implemented were ANNs. Finally, in the soil management category, four ML models were implemented, with the most popular once again being the ANN model. ANNs were used mostly for implementations in crop, water, and soil management, while SVMs were used mostly for livestock management.

By applying machine learning to sensor data, farm management systems are evolving into real artificial intelligence systems, providing richer recommendations and insights for the subsequent decisions and actions with the ultimate scope of production improvement.

3.Machine Learning Approaches to Manufacturing

The paper surveys machine learning techniques that seem to be applicable in realizing systems with intelligent behaviour. Symbolic, sub symbolic approaches and their applications in manufacturing are equally treated, together with hybrid solutions which try to integrate the benefits of the individual techniques. It is addressed by several new paradigms, for example just in time manufacturing. Sustainable production

and sustainable competitiveness or learn faster than your competitors are notations that describe current trends. The mass production paradigm is being succeeded by knowledge intensive engineering, etc. Learning organizations continuously update their goals, proclaim and act upon all employee's participation and understanding in how to achieve and define subgoals and systematize their educational efforts. Learning factories, enterprises in manufacturing must be conscious of their duality, i.e., the interdependence of their technical and human constituents. Though having concentrated to machine learning techniques to manufacturing, the authors of this survey are convinced that human and machine learning are equally essential for these learning enterprises.

4.COVID-19 Fake News Classification with Deep Learning

Human monitoring and identification of each false story is a difficult task. With advancement in processing technology, machine learning models, and deep learning techniques, user intervention can be replaced by assigning pattern identification task to computers. But it requires a large dataset of both real and fake news. The size and features of data in dataset, is not enough for building models for accurate classification. Accuracy achieved is not so high. This paper discusses such classification of fake news articles related to COVID-19 using deep learning.

The experiments shows that deep learning is efficient in terms of text classification than base line models. But the choice of accurate models is also necessary. Also, inconsistency in

dataset might lead to deviation in the model. Some news articles contain completely false information while some contains partially false statements. So, trying to learn such inconsistent data could be the reason why the accuracy bar achieved is not so high. One way to tackle this could be sub classification among the type of news articles and building hybrid models for the final classification. This would allow the initial models to only learn important parameters from sub classes. These parameters can be combined to build the final model which would contain necessary parameters belonging to each and every subclass thus contributing in improving the accuracy of the overall classifier.

5.A survey of machine learning for big data processing

In this paper, a literature survey of the latest advances in research on machine learning for big data processing. First, machine learning techniques are elaborated and highlights some promising learning methods in recent studies, such as representation learning, deep learning, distributed and parallel learning, transfer learning, active learning, and kernel-based learning. However, most traditional machine learning techniques are not inherently efficient or scalable enough to handle the data with the characteristics of large volume, different types, high speed, uncertainty and incompleteness, and low value density. In response, machine learning needs to reinvent itself for big data processing. This paper began with a brief review of conventional machine learning algorithms, followed by several current advanced learning methods. Then, a discussion about the challenges of learning with big data and

the corresponding possible solutions in recent research was given. In addition, the connection of machine learning with modern signal processing technologies was analysed through studying several latest representative research papers. To stimulate more interests for the audience of the paper, at last, open issues and research trends were presented.

Future scope

Machine learning is a field of ongoing research and development . Everyday machine learning experts are trying to develop new algorithm which are more efficient ,accurate and precise in every work which need to be performed by them no matter what the field of application is. The scope of machine learning is not limited to the investment sector. Rather ,it is expanding across all fields such as netflix ,information technology ,media & entertainment ,gaming and the automotive industry. As the machine learning scope is very high ,there are some area where researches are working toward revolutionizing the world for the future. Let us discuss them in details .

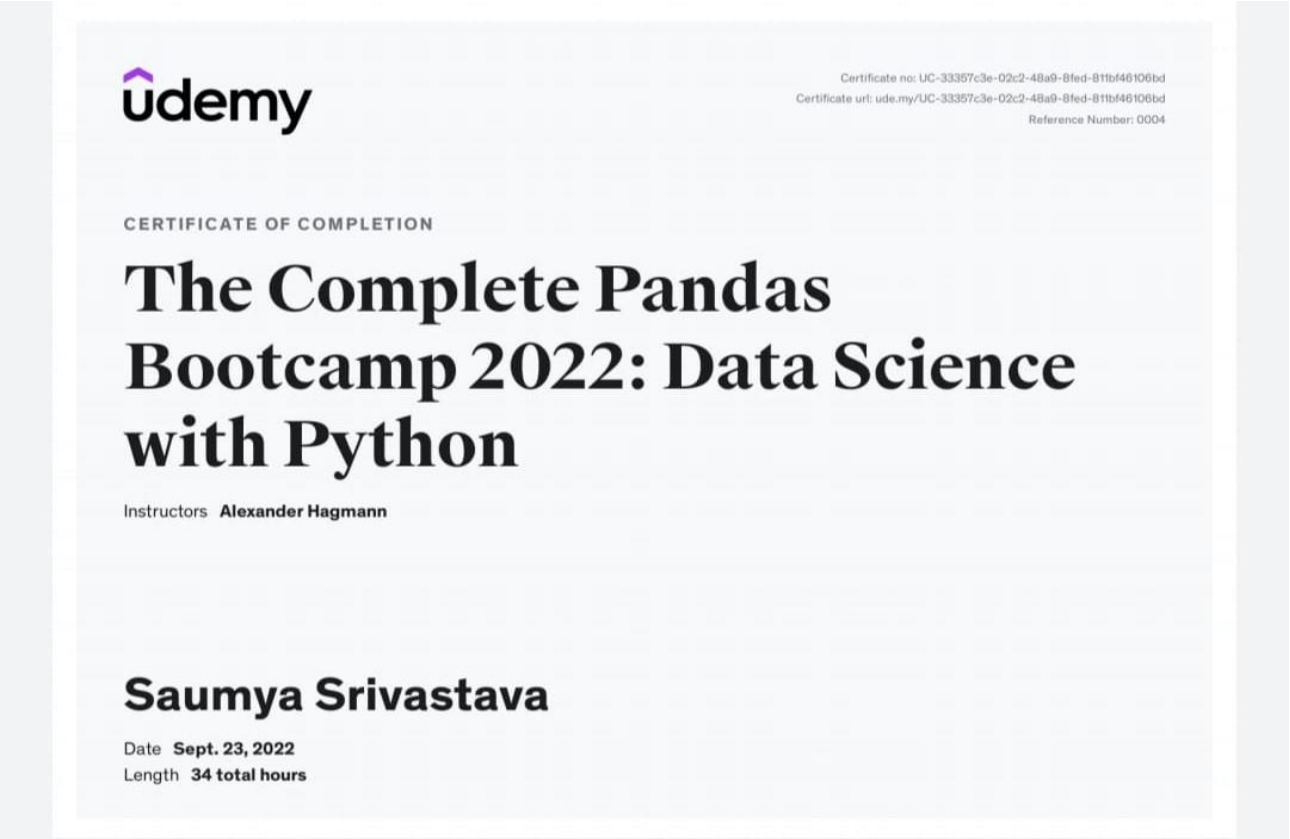
MOOC Certificate

Below are the links for MOOC certificate of all the team members along with the link:

1.Abhishree Bisht :

Link: The Complete Pandas Bootcamp 2022: Data Science with Python | Udemy

2.Saumya Srivastava:



udemy

Certificate no: UC-33357c3e-02c2-48a9-8fed-811bf46106bd
Certificate url: ude.my/UC-33357c3e-02c2-48a9-8fed-811bf46106bd
Reference Number: 0004

CERTIFICATE OF COMPLETION

The Complete Pandas Bootcamp 2022: Data Science with Python

Instructors **Alexander Hagmann**

Saumya Srivastava

Date **Sept. 23, 2022**
Length **34 total hours**

Link: The Complete Pandas Bootcamp 2022: Data Science with Python | Udemy

3.Geetika Verma:

udemy

Certificate no: UC-a93abea3-0ae3-4bef-ab59-b95ab0c26627
Certificate url: ude.my/UC-a93abea3-0ae3-4bef-ab59-b95ab0c26627
Reference Number: 0004

CERTIFICATE OF COMPLETION

The Complete Pandas Bootcamp 2022: Data Science with Python

Instructors **Alexander Hagmann**

Geetika Verma

Date **Sept. 5, 2022**
Length **34 total hours**

Link: The Complete Pandas Bootcamp 2022: Data Science with Python | Udemy