ABSTRACT

A movie recommendation is important in our social life because it has the ability to increase the entertainment for the people. every year there are many movies released and the movie lovers show different choices to particular movies. From the pool of movies user finds it difficult to select and watch the movies of his or her interest.so there comes the need of movie recommendation engine. Recommendation systems made finding things easy that the users need. Most of the existing movie recommendation systems generally provide the overall rating of the movie without any personalization of users. we have implemented and evaluated content based and collaborative based filtering on movie lens dataset by analysing both the models we tried to build a hybrid model to increase the accuracy of the movies recommended to the user. This engine aims for personalization of a user and ease the process of searching of movies based on users interests there by reducing human efforts.

Nowadays the recommendation system has made finding the things easy that we need. Most existing movie services like IMDB do not personalize their recommendations but simply provide an overall rating for a movie. This significantly decreases the value of each recommendation as it does not cater to the individual movie preferences of the user. Unlike these systems, our proposed Recommendation Engine will continually analyse individual user's movie preferences and recommend custom movie recommendations. This aims for personalization of a user, and the overall goal is to ease the movie discovery process. It reduces human effort by suggesting movies based on the user's interests.

CONTENTS

LIST OF FIGURES

I.INTRODUCTION

Basically, there exists some dependency between user and movies they like to watch. An efficient recommendation engine should explore these dependencies and recommend related movies to the users. Many companies are spending billions of dollars on implementing an accurate and personalized recommendation algorithm. Best example for this is Netflix prize competition. This was an open competition for the best collaborative filtering algorithm to find the ratings of unrated movies of the user basing on his previously rated movies without any other data. The competition was held by Netflix in 2006.mostly recommendation systems are of two types which are based on content based filtering and collaborative filtering.it is like if a user likes a movie A, he will also like similar movies like A.

Collaborative filtering filters information by using the interactions and data collected by the system from other users. It's based on the idea that people who agreed in their evaluation of certain items are likely to agree again in the future. The evolution of technology brings us many advanced platforms such as Machine Learning, Deep Learning, Internet of Things, etc. We are using these technologies to satisfy our needs. Recommendation system is the best example for this. Recommendation system is using Machine Learning to recommend Movies, Songs, Products in E-Commerce websites, etc. Methods which are widely used in Recommendation System are Content based filtering, Collaborative filtering. Collaborative filtering is one of the most popular methods to implement a recommender system. Collaborative filtering is of two types Memory based approach and Model based approach.

Dataset:

We have used movie lens 100k dataset. The data set consists of

- 100,000 ratings (1-5) from 943 users on 1682 movies.
- Each user has rated at least 20 movies.

The data was collected through the movie lens website.

1.1 Prerequisites:

1.1.1 Software Requirements:

1.Python version 3.0

2.Python IDE-Jupiter

3.Data science libraries-Pandas, NumPy, Sklearn, Matplotlib

1.1.2 Hardware requirements:

1. CPU - 8 to 16 with each octa core processor in a distributed network.
2. RAM - 128 to 256 GB
3. Storage – 30 to 50 GB

1.2 Python:

Python is a popular programming language. It was created by Guido van Rossum, and released in 1991.

Its often used as a “Scripting language” for web applications. This means that it can automate a specific

It's often used as a “scripting language” for web applications. This means that it can automate a specific series of tasks, making it more efficient. Consequently, Python (and languages like it) is often used in software applications, pages within a web browser, the shells of operating systems and some games.

It is used for:

- Web development (server-side),
- Software development,
- Mathematics,
- System scripting.

Python does the things as follow:

- Python can be used on a server to create web applications.
- Python can be used alongside software to create workflows.
- Python can connect to database systems. It can also read and modify files
- Python can be used to handle big data and perform complex mathematics.
- Python can be used for rapid prototyping, or for production-ready software development.

Advantages of python are mentioned below:

- Python works on different platforms (Windows, Mac, Linux, Raspberry Pi, etc).
- Python has a simple syntax similar to the English language.
- Python has syntax that allows developers to write programs with fewer lines than some other programming languages.
- Python runs on an interpreter system, meaning that code can be executed as soon as it is written. This means that prototyping can be very quick
- Python can be treated in a procedural way, an object-orientated way or a functional way.

1.2.1 Machine learning:

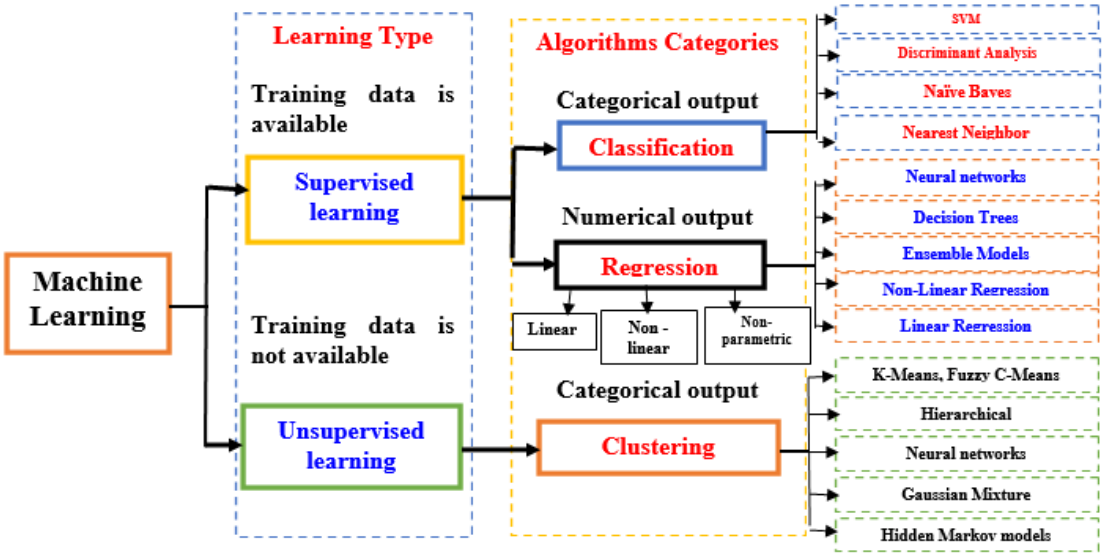

Fig 1.1 Machine learning overview

What is Machine learning:

Machine learning is a branch of artificial intelligence (AI) focused on building applications that learn from data and improve their accuracy over time without being programmed to do so.

In data science, an algorithm is a sequence of statistical processing steps. In machine learning, algorithms are 'trained' to find patterns and features in massive amounts of data in order to make decisions and predictions based on new data. The better the algorithm, the more accurate the decisions and predictions.

Machine learning is learning based on experience. As an example, it is like a person who learns to play chess through observation as others play. In this way, computers can be programmed through the provision of information in which they are trained, acquiring the ability to identify elements or their characteristics with high probability.

There are various stages of machine learning:

- Data collection
- Data storage
- Data analysis
- Algorithm development
- Checking algorithm generated
- The use of an algorithm to further conclusions

Machine learning algorithms are divided into two groups:

- Unsupervised learning
- Supervised learning

With Unsupervised learning, your machine receives only a set of input data. Thereafter, the machine is up to determine the relationship between the entered data and any other hypothetical data. Unlike supervised learning, where the machine is provided with some verification data for learning, independent Unsupervised learning implies that the computer itself will find patterns and relationships between different data sets. Unsupervised learning can be further divided into clustering and association. Supervised learning implies the computer ability to recognize elements based on the provided samples. The computer studies it and develops the ability to recognize new data based on this data. For example, you can train your computer to filter spam messages based on previously received information.

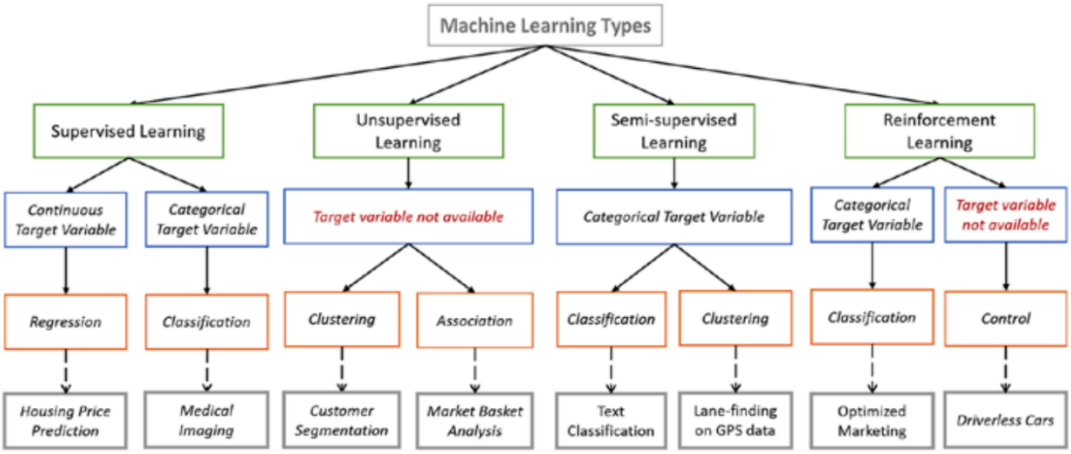

Fig 1.2. Types Of Machine Learning

Some Supervised learning algorithms include:

- Decision trees
- Support-vector machine
- Naive Bayes classifier
- k-nearest neighbours
- linear regression

Some Unsupervised learning algorithms include:

- K-means clustering
- Hierarchical clustering
- Anomaly detection
- Neural Networks
- Principal Component Analysis
- Independent Component Analysis
- Apriori algorithm

1.2.1.1 NumPy

NumPy is the fundamental package needed for scientific computing with Python.

NumPy, which stands for Numerical Python, is a library consisting of multidimensional array objects and a collection of routines for processing those arrays. NumPy is an open-source numerical Python library.

NumPy array is a powerful N-dimensional array object which is in the form of rows and columns. We can initialize NumPy arrays from nested Python lists and access it elements.

This package contains:

- A powerful N-dimensional array object
- Sophisticated (broadcasting) functions
- Basic linear algebra functions
- Basic Fourier transforms
- Sophisticated random number capabilities
- Tools for integrating Fortran code
- Tools for integrating C/C++ cod

Besides its obvious scientific uses, NumPy can also be used as an efficient multi-dimensional container of generic data. Arbitrary data types can be defined. This allows NumPy to seamlessly and speedily integrate with a wide variety of databases.

Firstly, the Data Frame can contain data that is:

- A Pandas Data Frame
- A Pandas Series: a one-dimensional labelled array capable of holding any data type with axis labels or index. An example of a Series object is one column from a Data Frame
- A NumPy ND array, which can be a record or structured.
- A two-dimensional ND array.
- Dictionaries of one-dimensional ND array, lists, dictionaries or Series

The difference between np. ND array and np. array (). The former is an actual data type, while the latter is a function to make arrays from other data structures.

Structured arrays allow users to manipulate the data by named fields: in the example below, a structured array of three tuples is created. The first element of each tuple will be called foo and will be of type int, while the second element will be named bar and will be a float.

Record arrays, on the other hand, expand the properties of structured arrays. They allow users to access fields of structured arrays by attribute rather than by index. You see below that the foo values are accessed in the r2 record array.

1.2.1.3 SK-Learn

Scikit-learn provides a range of supervised and unsupervised learning algorithms via a consistent interface in Python.

Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modelling including classification, regression, clustering and dimensionality reduction via a consistent interface in Python.

Scikit-learn is a library in Python that provides many unsupervised and supervised learning algorithms. It's built upon some of the technology you might already be familiar with, like NumPy, pandas, and Matplotlib!

The functionality that scikit-learn provides include:

- Regression, including Linear and Logistic Regression
- Clustering, including K-Means and K-Means Regression
- Model selection
- Pre-processing, including Min-Max Normalization