**NITTE MEENAKSHI INSTITUTE OF TECHNOLOGY**

(AN AUTONOMOUS INSTITUTION, AFFILIATED TO VISVESVARAYA TECHNOLOGICAL UNIVERSITY, BELGAUM, APPROVED BY AICTE & GOVT.OF KARNATAKA

****

**LEARNING ACTIVITY PROPOSAL**

**EXPLORATORY DATA ANALYSIS ON MOVIES**

*Submitted in partial fulfilment of the requirement for the award of Degree of*

*Bachelor of Engineering*

*in*

*Computer Science and Engineering*

**Introduction to Machine Learning (18CSE751)**

*Submitted by:*

|  |  |
| --- | --- |
| MOHAMED RAYAN KHAZI  GURUKIRAN K MOGER  AKSHAY SIMHA S P | 1NT18CS098  1NT18CS047  1NT18CS006 |

Under the Guidance of

Dr. Vani V

Professor, Dept. of CS&E, NMIT



Department of Computer Science and Engineering

**(Accredited by NBA Tier-1)**

2021-22

**ABSTRACT**

Machine learning is a subfield of artificial intelligence, which is broadly defined as the capability of a machine to imitate intelligent human behavior. It encompasses everything from teaching machines to do calculations of simple kind to predicting variables on a complex scale. This project deals with Exploratory Data Analysis (EDA) on a dataset of movies to predict the popularity. This is unavoidable and one of the major step to fine-tune the given data set(s) in a different form of analysis to understand the insights of the key characteristics of various entities of the data set like column(s), row(s) by applying Pandas, NumPy, Statistical Methods, and Data visualization packages.

In this project, we experiment with a real world dataset, and explore how machine learning algorithms can be used to find the patterns in data. We gain experience using a common data-mining and common machine learning libraries and algorithms.

**TABLE OF CONTENTS**

1. Introduction
2. Dataset
3. Machine Learning Methods
4. Presentation and Visualization
5. Roles
6. Schedule
7. Bibliography

**INTRODUCTION**

Machine learning is the science of getting computers to act without being explicitly programmed. In the past decade, machine learning has given us self-driving cars, practical speech recognition, effective web search, and a vastly improved understanding of the human genome. Machine learning is so pervasive today that you probably use it dozens of times a day without knowing it. Many researchers also think it is the best way to make progress towards human-level AI.

In this project, Exploratory Data Analysis has been conducted on the dataset “tmdb-movies” in order to predict the popularity of the movie. This includes a combination of algorithms and hence accuracies have been computed based on the outcomes. Comparison of these accuracies gives us the efficiency of each algorithm and its usability.

The procedure includes step by step processes in Exploratory Data Analysis. The dataset has been worked on by:

1. Reading and exploring dataset.
2. Visualizing the dataset.
3. Cleaning the dataset.

Next procedure includes Feature Selection and training data using the following algorithms:

1. **Random Forest Algorithm**: It is an ensemble learning algorithm. It is a Meta estimator that uses the process of fitting a number of decision tree classifiers on sub-samples of the used dataset and uses averaging to improve efficiency and predictive accuracy and hence control over-fitting. Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.
2. **Support Vector Machine**: The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space (N — the number of features) that distinctly classifies the data points. Support vector machine is another simple algorithm that is highly preferred by many as it produces significant accuracy with less computation power.
3. **K- Nearest Neighbors**: K-Nearest Neighbor is one of the simplest Machine Learning algorithms based on Supervised Learning technique. K-NN 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 K- NN algorithm. 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.

Respective accuracies have been computed and compared for reference.

**DATASET**

The dataset used in this project is called “tmdb-movies” that includes details about movies and related details in order to predict the popularity of a given movie.

The dataset has been obtained from Kaggle.com which has number of datasets to work on, on different aspects and topics.

1. The dataset is a fairly small that includes details of about 10,000 movies (recent and old).
2. The dataset has 10856 valid unique entries and over 20 columns. The columns present in the dataset hat contribute as attributes are :

# Column Non-Null Count Dtype

--- ------ -------------- -----

**0 id 10866 non-null int64**

1 imdb\_id 10856 non-null object

2 popularity 10866 non-null float64

3 budget 10866 non-null int64

4 revenue 10866 non-null int64

5 original\_title 10866 non-null object

6 cast 10790 non-null object

7 homepage 2936 non-null object

8 director 10822 non-null object

9 tagline 8042 non-null object

10 keywords 9373 non-null object

11 overview 10862 non-null object

12 runtime 10866 non-null int64

13 genres 10843 non-null object

14 production\_companies 9836 non-null object

15 release\_date 10866 non-null object

16 vote\_count 10866 non-null int64

17 vote\_average 10866 non-null float64

18 release\_year 10866 non-null int64

19 budget\_adj 10866 non-null float64

20 revenue\_adj 10866 non-null float64

Advantages of the used Dataset:

* Since the dataset is smaller, learning models like KNN and SVM can be used.
* The dataset gives a good picture on which movies have the highest and lowest budget and revenue.
* It also explains which most profitable movies are.
* Null values and outliers can be ignored as a result of the dataset being smaller.

Challenges of the used Dataset:

* The dataset does not cover many unique values and hence fitting of data may not be optimal.
* There is a risk of under-fitting in the dataset.
* Lack of generalization maybe observed in the dataset.
* Data may not be balanced completely.
* There can be a difficulty in optimization.

**MACHINE LEARNING METHODS**

Procedures run on the dataset include;

1. Reading and exploring dataset : First we will import the dataset which is required for our analysis. Exploring the data is one of the initial step in which we can use some statistical techniques to describe dataset characterizations, such as size, quantity, and accuracy, in order to better understand the nature of the data.
2. Visualizing the dataset : Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.
3. Cleaning the dataset : Data cleaning refers to identifying and correcting errors in the dataset that may negatively impact a predictive model. In the present dataset, the number of outliers and null values are small to ignore(since the dataset is small enough).
4. Feature Selection : Feature selection is the process of reducing the number of input variables when developing a predictive model. It is desirable to reduce the number of input variables to both reduce the computational cost of modeling and, in some cases, to improve the performance of the model.
5. Splitting test train data: The train-test split procedure is used to estimate the performance of machine learning algorithms when they are used to make predictions on data not used to train the model. It is a fast and easy procedure to perform, the results of which allow you to compare the performance of machine learning algorithms for your predictive modeling problem.
6. Fitting algorithms and learning models : Model fitting is a measure of how well a machine learning model generalizes to similar data to that on which it was trained. A model that is well-fitted produces more accurate outcomes. A model that is overfitted matches the data too closely. A model that is underfitted doesn't match closely enough.
7. Computing accuracies of algorithms : Computing accuracies and other related statistical details from the classification report generated from the learning models.

Specific Questions investigated on the dataset:

1. Give the names and id of movies having the highest and the lowest popularity.

This question helps to analyze the indexing of the dataset using the minimax algorithm.

1. Give the mean profit of the movies released between 2003 and 2007.

This question helps to find the mean and the variance in the dataset.

**PRESENTATION AND VISUALIZATION**

The results of each learning model have been assessed by generating the classification report.

The classification report consists of the following details:

1. Precision: Precision is the closeness of the measurements to each other.
2. Recall: Recall literally is how many of the true positives were recalled (found), i.e. how many of the correct hits were also found.
3. F1-score: F1 Score is the weighted average of Precision and Recall.
4. Support: The support is the number of samples of the true response that lie in that class.
5. Accuracy: Machine learning model accuracy is the measurement used to determine which model is best at identifying relationships and patterns between variables in a dataset based on the input, or training, data.
6. Macro average: Macro averaging reduces your multiclass predictions down to multiple sets of binary predictions, calculates the corresponding metric for each of the binary cases, and then averages the results together.
7. Weighted average: Weighted average is a calculation that takes into account the varying degrees of importance of the numbers in a data set.
8. Mean absolute error: mean absolute error (MAE) is a measure of errors between paired observations expressing the same phenomenon.
9. Residual sum of scores: The residual sum of squares (RSS) is a statistical technique used to measure the amount of variance in a data set that is not explained by a regression model itself. Instead, it estimates the variance in the residuals, or error term.
10. R2-score: the proportion of the variance in the dependent variable that is predictable from the independent variable(s).

Corresponding histograms and heatmaps have been plotted using the required libraries and modules.

**ROLES**

The roles of the teammates have been stated as follows:

1. Gurukiran K Moger: Conducted the following data preprocessing techniques: Data reading, Data Cleaning, Data visualization.
2. Akshay Simha S P: Investigating relevant questions on the dataset and also conducted feature selection.
3. Mohamed Rayan Khazi: Split the test train data, fit the learning models and algorithms, computed data accuracies and other parameters.

**SCHEDULE**

|  |  |
| --- | --- |
| DATE | TASKS TO BE COMPLETED |
| 20/12/2021 | Proposal report submission |
| 17/01/2022 | Final report submission |

**BIBLIOGRAPHY**

[1] Role of Exploratory Data Analysis in Data Science by A Suresh Rao, B. Vishnu Vardhan, Hafeezuddin Shaik (2021)

[2] Visualization and Prediction of Film Award Nominations by Using of Visual Data Mining (VDM) and Exploratory Data Analysis (EDA) Method by Rayhanali Heiko Amier, Johan Setiawan (2019)

[3] Exploratory Data Analysis and Machine Learning on Titanic Disaster Dataset by Karman Singh, Renuka Nagpal, Rajni Sehgal (2020)

[4] An Exploratory Analysis on Data Features and Analysis Techniques in Social Networks by U. TanujaH.L. ,Gururaj, V. Janhavi (2019)

[5] Prediction of Heart Disease using Different KNN Classifier by Chethana C (2020)

[6] Predicting movie box office success using multiple regression and SVM by V. Subramaniyaswamy, M. Viginesh Vaibhav, R. Vishnu Prasad, R. Logesh (2017)

[7] Decision Tree Algorithms for Accurate Prediction of Movie Rating by Kavya Pradeep, C R TintuRosmin, Sherly Susana Durom, G S Anisha (2020)

[8] <https://www.kaggle.com/> (dataset)