**NITTE MEENAKSHI INSTITUTE OF TECHNOLOGY**

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

****

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

**LEARNING ACTIVITY PROPOSAL**

**TUMOR DETECTION**

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

*Bachelor of Engineering*

*in*

*Computer Science and Engineering*

*Submitted by:*

|  |  |
| --- | --- |
| H PUNEETH SHETTY  KUMARA H | 1NT18CS049  1NT18CS081 |

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 “Tumor detection” 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.

**DATASET**

1. The dataset has valid entries and over 33 columns. The columns present in the dataset hat contribute as attributes are :

# Column Non-Null Count Dtype

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

0 id 569 non-null int64

1 diagnosis 569 non-null object

2 radius\_mean 569 non-null float64

3 texture\_mean 569 non-null float64

4 perimeter\_mean 569 non-null float64

5 area\_mean 569 non-null float64

6 smoothness\_mean 569 non-null float64

7 compactness\_mean 569 non-null float64

8 concavity\_mean 569 non-null float64

9 concave points\_mean 569 non-null float64

10 symmetry\_mean 569 non-null float64

11 fractal\_dimension\_mean 569 non-null float64

12 radius\_se 569 non-null float64

13 texture\_se 569 non-null float64

14 perimeter\_se 569 non-null float64

15 area\_se 569 non-null float64

16 smoothness\_se 569 non-null float64

17 compactness\_se 569 non-null float64

18 concavity\_se 569 non-null float64

19 concave points\_se 569 non-null float64

20 symmetry\_se 569 non-null float64

21 fractal\_dimension\_se 569 non-null float64

22 radius\_worst 569 non-null float64

23 texture\_worst 569 non-null float64

24 perimeter\_worst 569 non-null float64

25 area\_worst 569 non-null float64

26 smoothness\_worst 569 non-null float64

27 compactness\_worst 569 non-null float64

28 concavity\_worst 569 non-null float64

29 concave points\_worst 569 non-null float64

30 symmetry\_worst 569 non-null float64

31 fractal\_dimension\_worst 569 non-null float64

32 Unnamed: 32 0 non-null float64

dtypes: float64(31), int64(1), object(1)

Challenges of the used Dataset:

* The dataset does not cover many unique values and hence fitting of data may not be optimal.
* 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. 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. To check what value does the Diagnosis field have.

Which will give the what type of diagnosis malignant or others.

1. Give the mean of the tumor type. 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. Mean: Mean is given by the total of the values of the samples divided by the number of samples.
3. Standard deviation: Standard deviation is a number that describes how spread out the values are. A low standard deviation means that most of the numbers are close to the mean (average) value. A high standard deviation means that the values are spread out over a wider range.
4. Recall: Recall literally is how many of the true positives were recalled (found), i.e. how many of the correct hits were also found.
5. Support: The support is the number of samples of the true response that lie in that class.
6. 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.
7. Mean absolute error: mean absolute error (MAE) is a measure of errors between paired observations expressing the same phenomenon.

.

**ROLES**

The roles of the teammates have been stated as follows:

1. H Puneeth Shetty: Conducted the following data preprocessing techniques: Data reading, Data Cleaning, Data visualization.
2. Kumara H: 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. A Distincitve Model to Classify Tumor Using Random Forest Classifier

<https://ieeexplore.ieee.org/abstract/document/9036473>

# Deep random forest-based learning transfer to SVM for brain tumor segmentation

# <https://ieeexplore.ieee.org/abstract/document/7523095>

# 3. A Machine Learning Approach to Brain Tumors Segmentation Using Adaptive Random Forest Algorithm

# <https://ieeexplore.ieee.org/abstract/document/8735072>

# Multi stage classification and segmentation of brain tumor

# <https://ieeexplore.ieee.org/abstract/document/7724543>

# 