Cyber Labs

MACHINE LEARNING

# ML Essentials: Report WOC

## Week-1 (Learning period)

- Completed five weeks of Andrew Ng’s Machine Learning course on Coursera Platform.
- Gone through the Python tutorials and object oriented programming with Python.
- Gone through the essential libraries of machine learning in python.
  Numpy: primarily used for computation
  Pandas: for data manipulation and analysis
  Matplotlib: for visualisation of data by different plots
  Scikit Learn: for predictive data analysis.

## Week-2

### EDA ( Exploratory Data Analysis)

- Read articles on the topic EDA  and explored different steps involved in EDA.
- Basic data exploration: missing values, shape, info.
- Analysed data through different plots (scatter plots, histograms ).
- Handling Outliers: Analysed data through Boxplot if any outliers may be present in values of features.
- Finding correlation between features.
- Univariate analysis:  Label column is analysed by taking individual variable features into consideration.
- Bivariate analysis : Label column is analysed by taking both variable features into consideration.
- Normalization and Scaling : Scaling the values present in feature columns to the same range.It is very useful in case of algorithms based on principle of euclidean distance.

### K-Means Clustering from Scratch

- Read articles on K-Means from Cyberlabs resources.
- Splitted the data for the training and testing part.
- Randomly initialized four points as centroids of clusters.
- For the training part: clusters are assigned to the training dataset on the basis of Euclidean distance from randomly assigned centeroids.
- To reach the correct position of centroids, the mean value of all the data points present in each cluster is assigned  to the centroids. Performed this step for multiple iterations until the position of centroids became almost constant.
- Testing part: data points splitted for testing part is fitted in the model and predictions are made by model.

## MID EVALUATIONS

## Week-3

### Decision Tree from scratch

- Read the Decision tree algorithm from different resources about its different aspects : Nodes, Splitting, Branch, Entropy, Gini impurity, Information gain Types: Regressor and Classifier , Avoiding Overfitting.
- Basic EDA is performed on given data to explore its different aspects.
- As features contain categorical string type values which will affect our algorithm during calculation so Label Encoder is used to avoid such issues.
- Preparing the data: As data may contain some information which is irrelevant to our analysis . Such features are dropped.
- Splitting the dataset for training and testing part (4:1).
- Methods for entropy and information gain are defined in a class named Decision_tree.

## Week-4

### Implementing algorithm from Scikit Learn

- K-Means implementation: K-Means model is imported from "sklearn.cluster". Dataset is splitted into training and testing parts . The Elbow plot is plotted for getting appropriate K-value. Training or Fitting of the model by training data.In the testing part cluster is predicted by model for testing data.

- Decision Tree implementation: Basic EDA is performed .Label encoding. Preparing data by dropping features irrelevant to our analysis. Splitting dataset for training and testing . Importing Decision Tree Classifier from "sklearn.tree". Training or fitting the model by training data. Predicting label for testing dataset . Evaluating the prediction of model by Classification report and Confusion Matrix.

## Personal Details:

Name : Brijesh Kumar
Registration no. - 20JE0279