MACHINE LEARNING PROJECTS

Projects:

- **Artifical Neural Networks**
- **Natural Language Processing**

Instructions for the user:

1. Download the data file from the link given in each project
2. Open the .ipynb script from the link given
3. Upload the data to the 'google colab' using the browse option on the left side of the screen
4. Run the cells one after another to execute the script

Project 1:
ARTIFICIAL NEURAL NETWORKS

This project has the data derived from the 'UCI Machine Learning Repository'.

https://archive.ics.uci.edu/ml/datasets/Combined+Cycle+Power+Plant

Data: https://github.com/JavedNihaz/Machine-Learning-Models/blob/main/Folds5x2_pp.xlsx)

.ipynb script : https://colab.research.google.com/drive/1EtIX45a1v04tt5C-cjIQPnhHbKNCGA6N#scrollTo=B9CV13Co_HHM (This opens the code in 'google colab'. Its an online opensource platform that allows any user with a google account to run their python scripts. (Please execute the code by each cell)

Problem type: Regression

Objective:
To train the Artificial Neural Network model on the training data of 8000 rows and compare the results with the test set of 2000 rows. After training with the dataset, the model will be able to predict the Energy output of the power plant with high accuracy.

Data:
The dataset contains 9568 data points collected from a Combined Cycle Power Plant over 6 years (2006-2011), when the power plant was set to work with full load. Features consist of hourly average ambient variables Temperature (T), Ambient Pressure (AP), Relative Humidity (RH) and Exhaust Vacuum (V) to predict the net hourly electrical energy output (EP) of the plant.
A combined cycle power plant (CCPP) is composed of gas turbines (GT), steam turbines (ST) and heat recovery steam generators. In a CCPP, the electricity is generated by gas and steam

turbines, which are combined in one cycle, and is transferred from one turbine to another. While the Vacuum is collected from and has effect on the Steam Turbine, he other three of the ambient variables effect the GT performance.
For comparability with our baseline studies, and to allow 5x2 fold statistical tests be carried out, we provide the data shuffled five times. For each shuffling 2-fold CV is carried out and the resulting 10 measurements are used for statistical testing.

Attribute Information:
Features consist of hourly average ambient variables
- Temperature (T) in the range 1.81°C and 37.11°C,
- Ambient Pressure (AP) in the range 992.89-1033.30 milibar,
- Relative Humidity (RH) in the range 25.56% to 100.16%
- Exhaust Vacuum (V) in teh range 25.36-81.56 cm Hg
- Net hourly electrical energy output (EP) 420.26-495.76 MW
The averages are taken from various sensors located around the plant that record the ambient variables every second. The variables are given without normalization.

Script Explanation:
The script consists 3 parts:

- Part 0: Importing Libraries:
- Part 1: Data Preprocessing:
  This part involves importing the dataset, then splitting the dataset into training set and test set
- Part 2: Building the Artificial Neural Network (ANN)
  This part involves initializing the ANN, then adding the input layer and the first hidden layer, then the second hidden layer and finally the output layer
- Part 3: Training the Artificial Neural Network (ANN)
  This is the part where we train the Model on the given dataset. In this part we compile the ANN, then the train the ANN model on the training set. This runs the model on the entire 8000 rows of training data. Then we predict the results for the Test set that contains 2000 rows of test data.

  We can observe from the output that the loss gets stagnated after 50 epochs and the loss drops down to (26.7392) on the 100^{th} epoch

Result:
Then we print the results of the model (y_pred) and the test data (y_test) vertically to compare the difference between them. In order to compare the predicted results with the real data, we need to compare the rows one by one and we can see that the predictions each time are very close to the real results. We can see that the model predicts the energy output with great accuracy

Project 2:

NATURAL LANGUAGE PROCESSING

NAÏVE BAYES MODEL

Data: (download the data from the link)
https://drive.google.com/file/d/1VaOeqwNjASYlOgpkg2_1ON9S6608xOwg/view?usp=sharing

ipynb script:
https://colab.research.google.com/drive/1okbHTb3T92eeGrRsd_16bgMIC778b4k6#scrollTo=X1kiO9kACE6s

Objective:
The objective of this project is to use 'Natural Language Processing' to train the model to predict if the outcome of a review for a restaurant is positive or negative. This used sentiment analysis concepts to analyze the review and provides the results.

Data:
Contains reviews given by set of customers for a restaurant (1000 rows). The first column consists of the reviews and the second column contains info if the review is positive or negative (1 = positive, 0 = negative)

Script Explanation:
The script consists 3 parts:

- Part 0: Importing Libraries:
- Part 1: Data Preprocessing:
  This part involves importing the dataset and cleaning the Dataset
- Part 2: Creating Bag of Words Model:
  This part involves creating the bag of words model, then splitting the dataset into Training set(X) and Test set(y)
- Part 3: Training the Naïve Bayes Model:
  This part involves training the model on the Training set. Once the model is trained, the test set results are predicted and the result is evaluate using 'confusion matrix'

Result:
We can see that the model predicts whether these English words and reviews are positive or negative with 73% accuracy. The confusion matrix shows that the model has 55 correct predictions of negative reviews and 91 correct predictions of positive reviews, 42 incorrect predictions of positive reviews and 12 incorrect predictions of negative reviews