KTDK: Kuch Toh Data Kahenge

This event was intended to get students interested in Machine Learning and introduce ML Technique of Feature Selection to them.

Feature Selection:

Feature Selection is the process where you automatically or manually select those features which contribute most to your prediction variable or output in which you are interested in.

Feature Selection is used because:

- ➔ It reduces Overfitting.
- ➔ Improves Accuracy
- ➔ Reduces Training Time for the Model.

Methods for performing Feature Selection

1. **Manual/Intuitional:**

   In this we can select the best applicable features, by visualizing the data in the form of graphs and charts. One can do that with the help of tools like MS-Excel/Google Sheets. But the more used way is using various **visualisation libraries of python like matplotlib, Seaborn Library** etc.

2. **Using Statistical Test:**

   One of the ways for performing feature selection, includes getting the relation between the each of the feature and the dependent variable(i.e. Target Variable). The same can be found by using various **statistical techniques like Chi Square Test, ANOVA Test, Correlation Matrix** etc.

3. **Using ML Techniques:**
   There are various ML Techniques which can be used for feature selection like **Recursive Feature Elimination with Cross-validation (RFECV), Tree Based Classifier Techniques** etc.

All these three above mentioned techniques for feature selection can be easily implemented in python using the **scikit-learn library**.

One can learn more about Feature Selection on this link:
https://machinelearningmastery.com/feature-selection-with-real-and-categorical-data/

For Learning/Doing Feature Selection using **scikit-learn:**
https://scikit-learn.org/stable/modules/feature_selection.html

Convention followed in this Repo:

Test-1/2.py: This file contains the code used to check the accuracy for the various weighted-dataset.

Calculation-1/2.py: This file contains the code used for generating the ideal weight using the above mentioned techniques (namely Intuitional (only for DS1), Mutual Information (Chi Square Test), ANOVA Test, Correlation Matrix, Tree based Classifier).

Weight-1/2.csv: This file contains the weights calculated for the training of the model. (The weights are relative to the each other, i.e.. the most connected feature to the target variable has been assigned weight as 1, and for categorical features threshold value is decided for selection)

Result-1/2.csv: This file contains the accuracy of the respective trained models over test data.

Link to The Leader-Board:
https://docs.google.com/spreadsheets/d/1UKZC7_XVo4OhTlbMc9JXXgU8hDD_NzN4Cis5mv5ptW0

For Getting more details on the dataset:

Source for Dataset-1:
https://www.kaggle.com/radmirzosimov/telecom-users-dataset

Source for Dataset-2:
https://www.kaggle.com/fedesoriano/company-bankruptcy-prediction