Read me

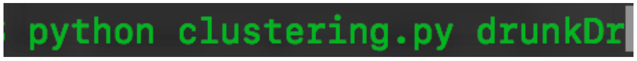
Name: Dwipam Katariya(ddkatari@iu.edu,)
Neelam Tikone(ntikone@iu.edu)

Following code files are included in the zip file:

1. clustering.py, clustering_kmeans.py
2. cluster_analysis.py
3. analysis_R.rmd
4. analysis_R.html

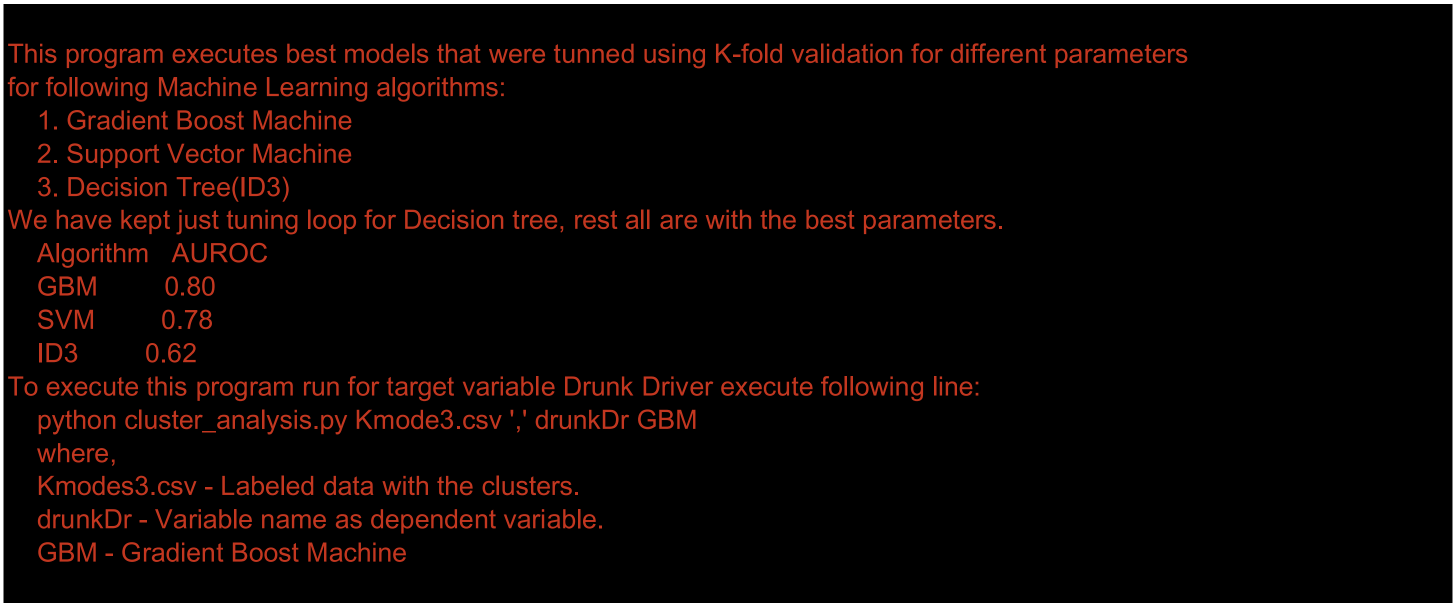
1. clustering.py: (Clustering program) clustering_kmeans.py
   Pre-reqs:
   K-modes library: Unsupervised learning algorithm for categorical data
   Sklearn: Machine Learning library and Metric evaluation
   Pandas: Soring data in tabular format
   Numpy: Scientific computations.

   To run the program:

   ```
   python clustering.py drunkDr
   ```

   Where,
   drunkDr is the variable name to cluster against, because if we want to see effects of clustering for Drunk Driving accidents, we can remove the variable.
2. cluster_analysis.py:
   Pre-reqs:
   Graphlab library : Machine Learning api (Best for ID3)
   Sklearn: Machine Learning api (Best for Linear SVM & Multi SVM) & Metric evaluations
   Imblearn : Balancing data library
   Pandas: Soring data in tabular format
   Numpy: Scientific computations.
   H2O.ai: Machine Learning api (Best for deep learning and gradient boosting)



```
This program executes best models that were tunned using K-fold validation for different parameters
for following Machine Learning algorithms:
   1. Gradient Boost Machine
   2. Support Vector Machine
   3. Decision Tree(ID3)
We have kept just tuning loop for Decision tree, rest all are with the best parameters.
   Algorithm   AUROC
   GBM         0.80
   SVM        0.78
   ID3        0.62
To execute this program run for target variable Drunk Driver execute following line:
   python cluster_analysis.py Kmode3.csv ',' drunkDr GBM
   where,
   Kmodes3.csv - Labeled data with the clusters.
   drunkDr - Variable name as dependent variable.
   GBM - Gradient Boost Machine
```

3. analysis_R.rmd (R Markdown File), analysis_R.html(HTML file for EDA No need to execute RMD)

Following are dataset files included:

1. final_data.csv : Data with selected important variables using domain knowledge
2. Population.csv: Population file in csv state wise
3. Us_county.csv: Geocoding for accidents

First execute .RMD which will output filtered fatalities, then use clustering.py, then use cluster_analysis.py

Following files are exported as the result of clustering.py and clustering_kmeans.py, we have kept both the files as a result of best cluster optimization, so that you can directly run cluster_analysis.py

- Filtered_fatalities
- Kmeans3.csv
- Kmode3.csv

Please execute directly Supervised learning as follows:

```
python cluster_analysis.py Kmode3.csv ',' drunkDr GBM
```