**Breast Cancer Wisconsin (Diagnostic) Data Set (**[**Link**](http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29)**):**

Data Set Info: This data set is from UCI Machine Learning Repository. Features are computed from a digitized image of a fine needle aspirate(FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image.

Attribute Information:

1. ID Number
2. Diagnosis (M = malignant, B = benign)
3. Radius (mean of distances from center to points on the perimeter)
4. Texture (standard deviation of gray-scale values)
5. Perimeter
6. Area
7. Smoothness (local variation in radius lengths)
8. Compactness (perimeter^2/ ares - 1.0)
9. Concavity (severity of concave portions of the contour)
10. Concave points (number of concave portions of the contour)
11. Symmetry
12. Fractal dimension

(Feature 3-12 are computed for each cell nucleus)

Associated Tasks: Classification. The main purpose of this task is to predict breast cancer of the observation base on the given features.

Possible Algorithm: Random Forest, SVM, Logistic Regression, Neural Networks

**In-vehicle Coupon Recommendation Data Set (**[**link**](http://archive.ics.uci.edu/ml/datasets/in-vehicle+coupon+recommendation)**):**

Data Set Info: This data set is collected via a survey on Amazon Mechanical Turk. The survey included different driving scenarios: destination, current time, weather, passenger, etc Then ask people whether they could accept the coupon if they were the driver.

Attribute Information:

1. Destination: No Urgent Place, Home, Work
2. Weather: Sunny, Rainy, Snowy
3. Temperature:55, 80, 30
4. Time: 2PM, 10AM, 6PM, 7AM, 10PM
5. Coupon: Restaurant(<$20), Coffee House, Carry out & Take away, Bar, Restaurant($20-$50)
6. Expiration: 1d, 2h (the coupon expires in 1 day or in 2 hours)
7. Aender: Female, Male
8. Age: 21, 46, 26, 31, 41, 50plus, 36, below21
9. MaritalStatus: Unmarried partner, Single, Married partner, Divorced, Widowed
10. Has\_Children:1, 0
11. Education: Some college

……

There are more than 30 features, most of them are in numeric. Labels of the observations are 0 or 1. 0 is not accepting and 1 is accepting.

Associate Tasks: Clustering, Classification. The main purpose of this task is to study whether a person will accept the coupon recommended to him in different driving scenarios.

Possible Algorithm: K-means clustering, PCA (unsupervised Learning)

**Coleridge Initiative - Show US the Data(**[**link**](https://www.kaggle.com/c/coleridgeinitiative-show-us-the-data)**):**

Data Set Info: This data set are scientific publications from numerous research areas gathered from [CHORUS](https://www.chorusaccess.org/) publisher members and other sources. These publications include pandemics, climate change, Alzheimer’s disease, child hunger...

Attribute Information:

1. id: publication id
2. pub\_title: title of the dataset that is mentioned within the publication
3. dataset\_title: the title of the dataset that is mentioned within the publication
4. dataset\_label: a portion of the text that indicates the dataset
5. cleaned\_label

Associated Tasks: Classification. The main task is to use natural language processing to automate the discovery of how scientific data are referenced in publications. If we can find how datasets are used in publications quickly and at scale, the government can catalog answers to most pressing questions and apply them to the world’s critical issues.

Possible Algorithm: Navia Bayes, KNN, RNN model(BERT)...