1. Software Engineering team Assessment
    a. Name: SETAP (Software Engineering Team Assessment Process)
    b. Description: Focused on assessing software engineering teams during a certain time interval (Time Interval 1). 11 time intervals for this data
    c. Number of features: 84
    d. Number of rows (teams): 64
2. Period Changer
    a. Description: The dataset includes 90 non-toxic molecules designed for functional domain of a core clock protein, CRY1, of which 27 molecules significantly lengthen the period of circadian rhythm and the rest, 63 molecules, are no changers.
    b. Number of features: 1177
    c. Number of molecules : 90
3. Darwin data
    a. Description: The DARWIN dataset includes handwriting data from 174 participants. The classification task consists in distinguishing Alzheimer's disease patients from healthy people.
    b. Number of features: 451
    c. Number of people: 174
4. Toxicity in molecules
    a. The dataset includes 171 molecules designed for functional domains of a core clock protein, CRY1, responsible for generating circadian rhythm. 56 of the molecules are toxic and the rest are non-toxic.
    b. Number of features: 1203
    c. Number of rows: 171
5. Voice Rehabilitation dataset
    a. Description: 126 samples from 14 participants, 309 features. Aim: assess whether voice rehabilitation treatment lead to phonations considered 'acceptable' or 'unacceptable'
    b. Number of features: 309
    c. Number of rows: 126
6. Malware Static and Dynamic Features(Not done)
    a. Description: 3 datasets: staDynBenignLab.csv, features extracted from 595 files (Win 7 and 8); staDynVxHeaven2698Lab.csv, from 2698 files of VxHeaven and staDynVt2955Lab.csv,from 2955 files of Virus Total.
        i. - staDynBenignLab.csv: 1086 features extracted from 595 files on MS Windows 7 and 8, obtained Program Files directory.
        ii. - staDynVxHeaven2698Lab.csv: 1087 features extracted from 2698 files of VxHeaven dataset.
        iii. - staDynVt2955Lab.csv: 1087 features extracted from 2955 provided by Virus Total in 2018.

7. Gastrointestinal Lesions in Regular Colonoscopy(Cannot be used)

    a. Description: dataset that contains features extracted from colonoscopy videos, and the purpose of this dataset is to detect gastrointestinal lesions. The dataset includes a total of 76 lesions, which are categorized into three types: 15 serrated adenomas, 21 hyperplastic lesions and 40 adenoma.
    b. Number of features: 698
    c. Number of instances: 76
8. Taiwanese bankruptcy data(Not done): The data were collected from the Taiwan Economic Journal for the years 1999 to 2009. Company bankruptcy was defined based on the business regulations of the Taiwan Stock Exchange.
    a. Features: 96
    b. Instances: 6819
9. Sales Transactions Dataset Weekly (Cannot be used)
    a. Contains weekly purchased quantities of 800 over products over 52 weeks. Normalized values are provided too.
    b. Features: 107
    c. Instances: 811
10. Parkinson's Disease Classification(Not done) : The data used in this study were gathered from 188 patients with PD (107 men and 81 women) with ages ranging from 33 to 87 (65.1Â±10.9) at the Department of Neurology in CerrahpaÅŸa Faculty of Medicine, Istanbul University. The control group consists of 64 healthy individuals (23 men and 41 women) with ages varying between 41 and 82 (61.1Â±8.9). During the data collection process, the microphone is set to 44.1 KHz and following the physicianâ€™s examination, the sustained phonation of the vowel /a/ was collected from each subject with three repetitions.
    a. Features: 755
    b. Instances: 755
11. Superconductivity Data: (Cannot be used) There are two files: (1) train.csv contains 81 features extracted from 21263 superconductors along with the critical temperature in the 82nd column, (2) unique_m.csv contains the chemical formula broken up for all the 21263 superconductors from the train.csv file. The last two columns have the critical temperature and chemical formula. The original data comes from http://supercon.nims.go.jp/index_en.html which is public. The goal here is to predict the critical temperature based on the features extracted.
    a. Features: 81
    b. Instances: 21263
12. Residential Dataset: (Cannot be used) Data set includes construction cost, sale prices, project variables, and economic variables corresponding to real estate single-family residential apartments in Tehran, Iran.
    a. Features: 109
    b. Instances: 372

13. Amazon Access Samples (Cannot be used) : Amazon's InfoSec is getting smarter about the way Access data is leveraged. This is an anonymized sample of access provisioned within the company.
    a. Features: 4
    b. Instances: 716063
14. APS Failure at Scania Trucks (Not done): The datasets' positive class consists of component failures for a specific component of the APS system. The negative class consists of trucks with failures for components not related to the APS.
    a. Features: 171
    b. Instances: 60000
15. Cargo 2000 Freight Tracking and Tracing: (Cannot be used) Sanitized and anonymized Cargo 2000 (C2K) airfreight tracking and tracing events, covering five months of business execution (3,942 process instances, 7,932 transport legs, 56,082 activities).
    a. Features: 98
    b. Instances: 3943
16. Swarm Behavior(Not done): This dataset achieved from an online survey, which is run by UNSW, Australia. It contains three data of ' Flocking - Not Flocking', 'Aligned - Not Aligned', and 'Grouped - Not Grouped'.
    a. Features: 2400
    b. Instances: 24017
17. Victorian Era Authorship Attribution: (Cannot be used) To create the largest authorship attribution dataset, we extracted works of 50 well-known authors. To have a non-exhaustive learning, in training there are 45 authors whereas, in the testing, it's 50
    a. Features: 1000
    b. Instances: 93600
18. Blog Feedback:(Cannot be used) Instances in this dataset contain features extracted from blog posts. The task associated with the data is to predict how many comments the post will receive.
    a. Features:
    b. Instances:
19. TUANDROMD (Tezpur University Android Malware Dataset)(Not Done): TUANDROMD (Tezpur University Android Malware Dataset)
    a. Features: 241
    b. Instances: 4464
20. Urban Land Cover: Classification of urban land cover using high resolution aerial imagery. Intended to assist sustainable urban planning efforts.
    a. Features: 148
    b. Instances: 168
21. Colon Data
    a. Features: 2000
    b. Instances: 62
22. DLBCL Data

    a. Features: 7070
    b. Instances: 77
23. Gastric Cancer Data
    a. Features: 4522
    b. Instances: 30
24. Gastroenterology data
    a. Features: 698
    b. Instances: 152
25. Leukemia Data
    a. Features: 5147
    b. Instances: 72
26. QSAR androgen receptor: 1024 binary attributes (molecular fingerprints) used to classify 1687 chemicals into 2 classes (binder to androgen receptor/positive, non-binder to androgen receptor /negative)
    a. Features: 1024
    b. Instances: 1687
27. QSAR oral toxicity: Data set containing values for 1024 binary attributes (molecular fingerprints) used to classify 8992 chemicals into 2 classes (very toxic/positive, not very toxic/negative)
    a. Features: 1024
    b. Instances: 8992