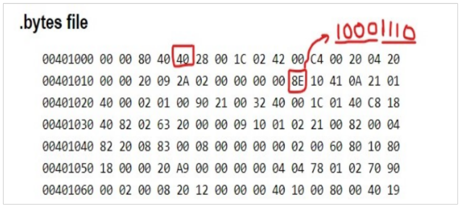
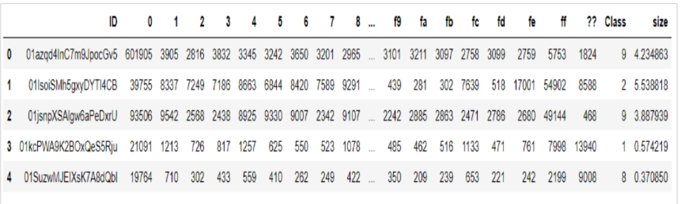
Byte file structure



.bytes file

```
00401000 00 00 80 40 40 28 00 1C 02 42 00 C4 00 20 04 20
00401010 00 00 20 09 2A 02 00 00 00 00 8E 10 41 0A 21 01
00401020 40 00 02 01 00 90 21 00 32 40 00 1C 01 40 C8 18
00401030 40 82 02 63 20 00 00 09 10 01 02 21 00 82 00 04
00401040 82 20 08 83 00 08 00 00 00 00 02 00 60 80 10 80
00401050 18 00 00 20 A9 00 00 00 00 04 04 78 01 02 70 90
00401060 00 02 00 08 20 12 00 00 00 40 10 00 80 00 40 19
```

Uni-gram BOW is created using the frequency of each Keyword (256 hexadecimal keywords + 1 feature as file size) from each byte file



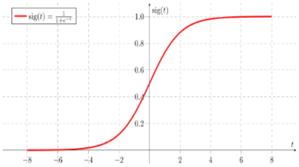
	ID	0	1	2	3	4	5	6	7	8	...	f9	fa	fb	fc	fd	fe	ff	??	Class	size
0	01azqd4InC7m9JpocGv5	601905	3905	2816	3832	3345	3242	3650	3201	2965	...	3101	3211	3097	2758	3099	2759	5753	1824	9	4.234863
1	01IsoiSMh5gxyDYTl4CB	39755	8337	7249	7186	8663	6844	8420	7589	9291	...	439	281	302	7639	518	17001	54902	8588	2	5.538818
2	01jsnpXSAlgw6aPeDxrU	93506	9542	2568	2438	8925	9330	9007	2342	9107	...	2242	2885	2863	2471	2786	2680	49144	468	9	3.887939
3	01kcPWA9K2BOxQeS5Rju	21091	1213	726	817	1257	625	550	523	1078	...	485	462	516	1133	471	761	7998	13940	1	0.574219
4	01SuzwMJEIXsK7A8dQbl	19764	710	302	433	559	410	262	249	422	...	350	209	239	653	221	242	2199	9008	8	0.370850

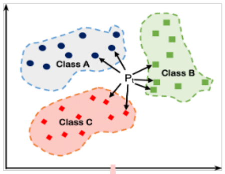
Logistic Regression

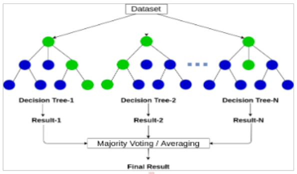
K-Nearest Nieghbor

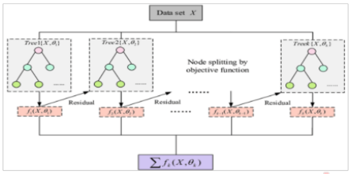
Random Forest

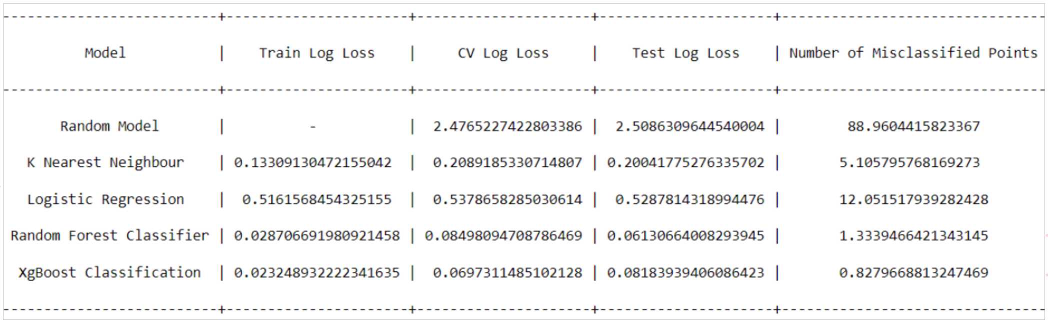
XgBoost (GBDT)











```
-----------------------------+-------------------------+------------------------+-------------------------+-----------------------------------

           Model             |     Train Log Loss      |      CV Log Loss       |      Test Log Loss      | Number of Misclassified Points

-----------------------------+-------------------------+------------------------+-------------------------+-----------------------------------

        Random Model         |            -            |   2.4765227422803386   |   2.5086309644540004    |        88.9604415823367
     K Nearest Neighbour     |   0.13309130472155042   |   0.2089185330714807   |   0.20041775276335702   |        5.105795768169273
     Logistic Regression     |    0.5161568454325155   |   0.5378658285030614   |   0.5287814318994476    |        12.051517939282428
  Random Forest Classifier   |   0.028706691980921458  |   0.08498094708786469  |   0.06130664008293945   |        1.3339466421343145
   XgBoost Classification    |   0.023248932222341635  |   0.0697311485102128   |   0.08183939406086423   |        0.8279668813247469

-----------------------------+-------------------------+------------------------+-------------------------+-----------------------------------
```