Binary 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
```

Assembly file structure

segment	address	bytes	opcode	operands
.text:	10001007	BF 8C D5 EA DB	mov	edi, 0DBEAD58Ch
.text:	1000100C	78 51	js	short loc_1000105F
.text:	1000100E	B6 B7	mov	dh, 0B7h
.text:	10001010	24 8D	and	al, 8Dh
.text:	10001012	42	inc	edx
.text:	10001013	53	push	ebx
.text:	10001014	90	nop	
.text:	10001015	89 8E AF BC 45 9A	mov	[esi-65BA4351h], ecx

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

Uni-gram BOW is created using the frequency of each Keyword (52 opcodes + 1 feature as file size) from each asm file

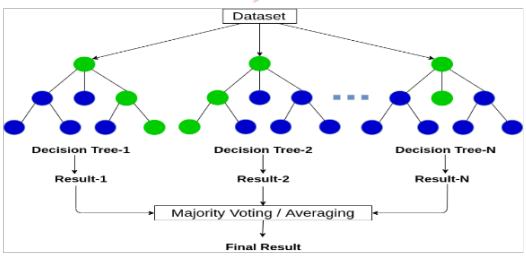
	ID	0	1	2	3	4	5	6	7	8	...	f9	fa	fb	fc	fd	fe	ff	??	Class	size
0	01azqd4InC7m9JpocGv5	601905	3905	2816	3832	3345	3242	3650	3201	2965	...	3101	3211	3097	2758	3099	2759	5753	1824	9	4.234863
1	01IsoiSMh5gxyDYTI4CB	39755	8337	7249	7186	8663	6844	8420	7589	9291	...	439	281	302	7639	518	17001	54902	8598	2	5.538818
2	01jsnpXSAlgw6aPeDxrU	93506	9542	2568	2438	8925	9330	9007	2342	9107	...	2242	2885	2863	2471	2786	2680	49144	468	9	3.687939
3	01kcPWA9K2BOxQeS5Rju	21091	1213	726	817	1257	625	550	523	1078	...	485	462	516	1133	471	761	7998	13940	1	0.574219
4	01SuzwIMJEIXsK7A8dQbI	19764	710	302	433	559	410	262	249	422	...	350	209	239	653	221	242	2199	9008	8	0.370850

	ID	HEADER:	.text:	.Pav:	.idata:	.data:	.bss:	.rdata:	.edata:	.rsrc:	...	esi	eax	ebx	ecx	edi	ebp	esp	eip	Class	size
0	01kcPWA9K2BOxQeS5Rju	19	744	0	127	57	0	323	0	3	...	66	15	43	83	0	17	48	29	1	0.078190
1	1E93CpP60RHFNiT5Qfvn	17	838	0	103	49	0	0	0	3	...	29	48	82	12	0	14	0	20	1	0.063400
2	3ekVow2ajZHbTnBcsDfX	17	427	0	50	43	0	145	0	3	...	42	10	67	14	0	11	0	9	1	0.041695
3	3X2nY7iQaPBIWDrAZqJe	17	227	0	43	19	0	0	0	3	...	8	14	7	2	0	8	0	6	1	0.018757
4	46OZzdsSKDCFV8h7XWxf	17	402	0	59	170	0	0	0	3	...	9	18	29	5	0	11	0	11	1	0.037567

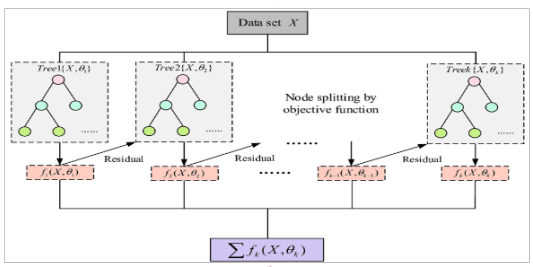
Both asm & byte files are merge / concatenated , both after combining gives total of 257+53 = 310 features to train our model

	0	1	2	3	4	5	6	7	8	9	...	edx	esi	eax	ebx	ecx
0	0.002525	0.000082	0.000013	0.000017	0.000016	0.000012	0.000004	0.000003	0.000099	0.000004	...	0.001622	0.000452	0.002331	0.000678	0.000013
1	0.010740	0.001771	0.000416	0.000489	0.000886	0.000320	0.000332	0.000662	0.000994	0.001810	...	0.015494	0.012723	0.026849	0.016721	0.003503
2	0.005374	0.000624	0.000130	0.000249	0.000129	0.000100	0.000105	0.000168	0.000174	0.000170	...	0.001546	0.000960	0.001085	0.003154	0.000471
3	0.008818	0.000957	0.000176	0.000247	0.000174	0.000207	0.000123	0.000221	0.000222	0.000236	...	0.000324	0.010678	0.000382	0.012012	0.001338
4	0.037465	0.000991	0.000251	0.000246	0.000315	0.000366	0.000252	0.000369	0.000447	0.000460	...	0.008968	0.004260	0.014892	0.007386	0.002357

We are using Random Forest as our first Algorithm to Train Model



We are using XgBoost as our Second Algorithm to Train Model



Model	Train Log Loss	CV Log Loss	Test Log Loss	Number of Misclassified Points
Random Forest Classifier	0.014762155486539353	0.042874337981406434	0.04192004241532452	0.9199632014719411
XgBoost Classification	0.010227838888593169	0.046105797578046656	0.03857589259645944	0.78196872125115