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 (52 opcodes + 1 feature as file size) from each asm file

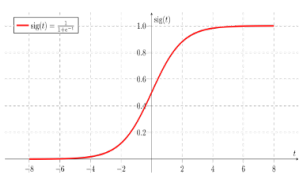
	ID	HEADER:	.text:	.Pav:	.idata:	.data:	.bss:	.rdata:	.edata:	.rsrc:	...	esi	eax	ebx	ecx	edi	ebp	esp	eip	Class	size
0	01kcPWA9K2BOxQeS5Rju	19	744	0	127	57	0	323	0	3	...	66	15	43	83	0	17	48	29	1	0.078190
1	1E93CpP60RHFNiT5Qfvn	17	838	0	103	49	0	0	0	3	...	29	48	82	12	0	14	0	20	1	0.063400
2	3ekVow2ajZHbTnBcsDfX	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

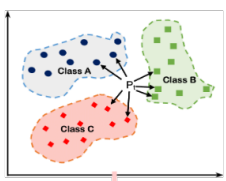
Logistic Regression

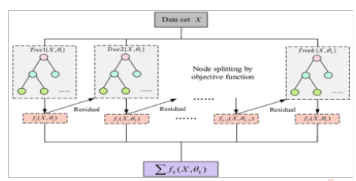
K-Nearest Nieghbor

Random Forest

XgBoost (GBDT)







Model	Train Log Loss	CV Log Loss	Test Log Loss	Number of Misclassified Points
K Nearest Neighbour	0.037475912101338785	0.09551157959211777	0.07985178727072437	1.1039558417663293
Logistic Regression	0.3951231421379966	0.40949221859900775	0.39929985766626985	8.693652253909843
Random Forest Classifier	0.021489116411912302	0.036614779171787376	0.03239270498531275	0.5059797608095675
XgBoost Classification	0.019288923540468395	0.03159265585460439	0.02576461657919649	0.45998160073597055