Дослідження методів мультикатегоріальної класифікації даних

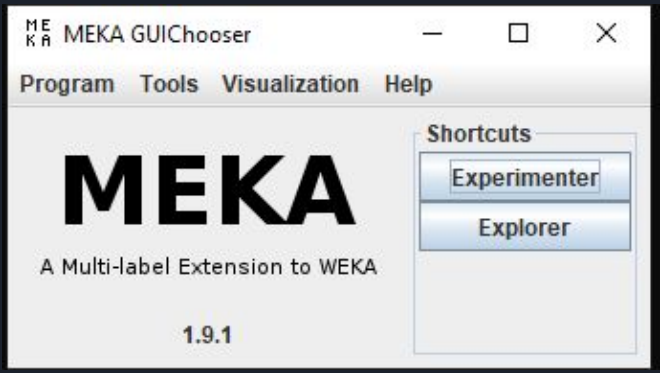
Виконав:
студент Кісіль Іван, Да-61
Науковий керівник:
Сергеєв-Горчинський О. О.

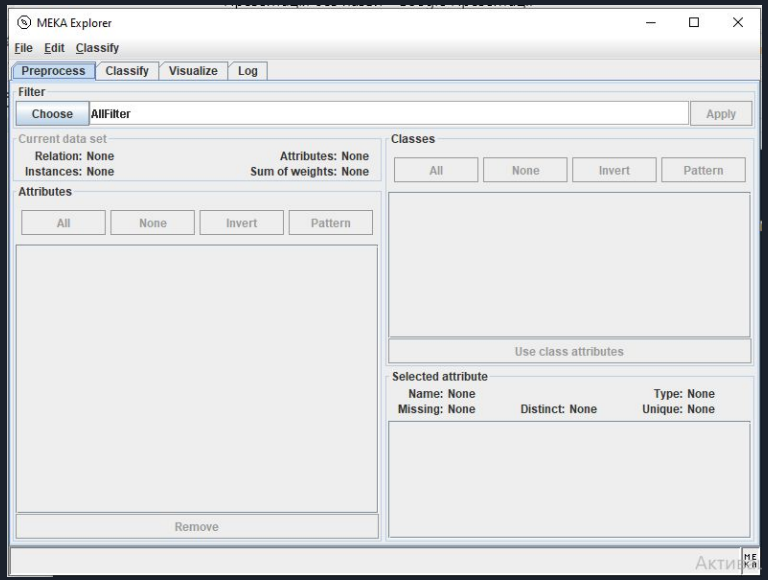
Мета

Метою роботи було вивчення особливостей мультикатегоріальної класифікації даних; експериментальне дослідження наявних методів; проведення експериментів для порівняння методів, враховуючи характеристики вихідних даних.

MEKA framework

Коли мова заходить про роботу з мультикатегоріальними даними, є кілька альтернатив на вибір. Проте в даній роботі я запиню свій вибір на MEKA framework.



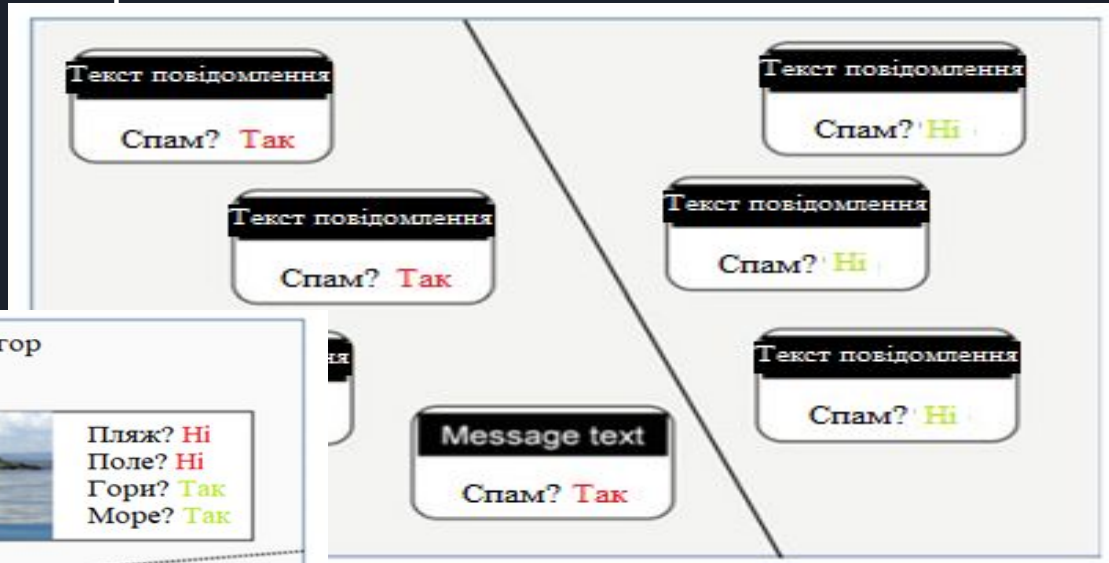


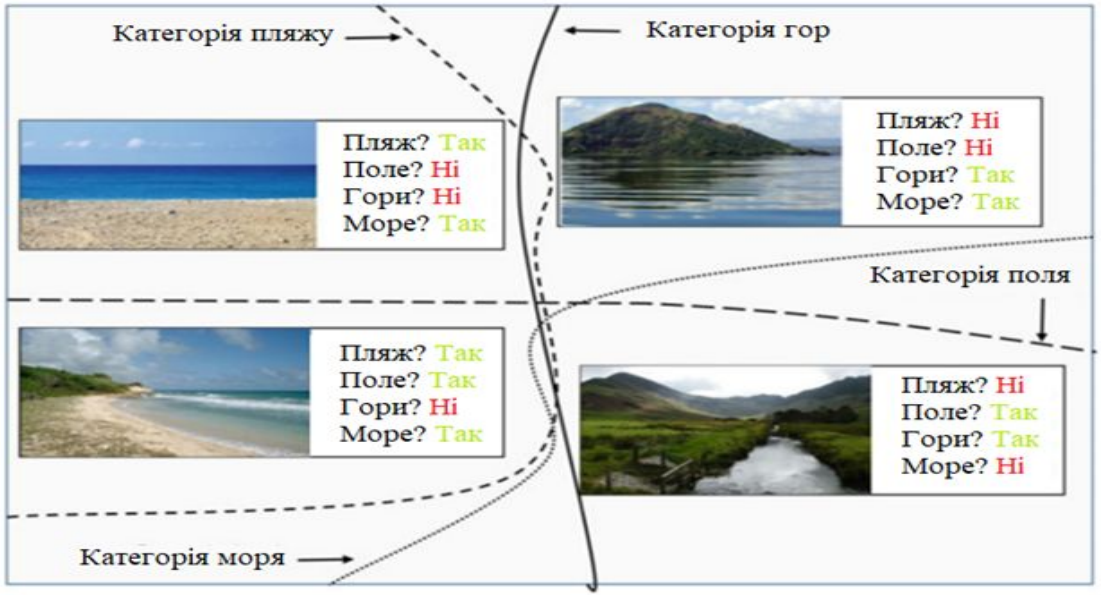
MEKA – це java-фреймворк з відкритим вихідним кодом, заснований на добре відомій бібліотеці WEKA.

Різниця між однозначною та мультикатегріальною класифікаціями

Однозначна класифікація: Чи є текстове повідомлення спамом?

∈ {так, ні}





Мультикатегоріальна класифікація: Які мітки відповідають даному мулюнку?

∈ {Пляж, Поле, Гори, Море}

Методи класифікації для багатьох міток

Методи

У роботі реалізовані наступні методи:

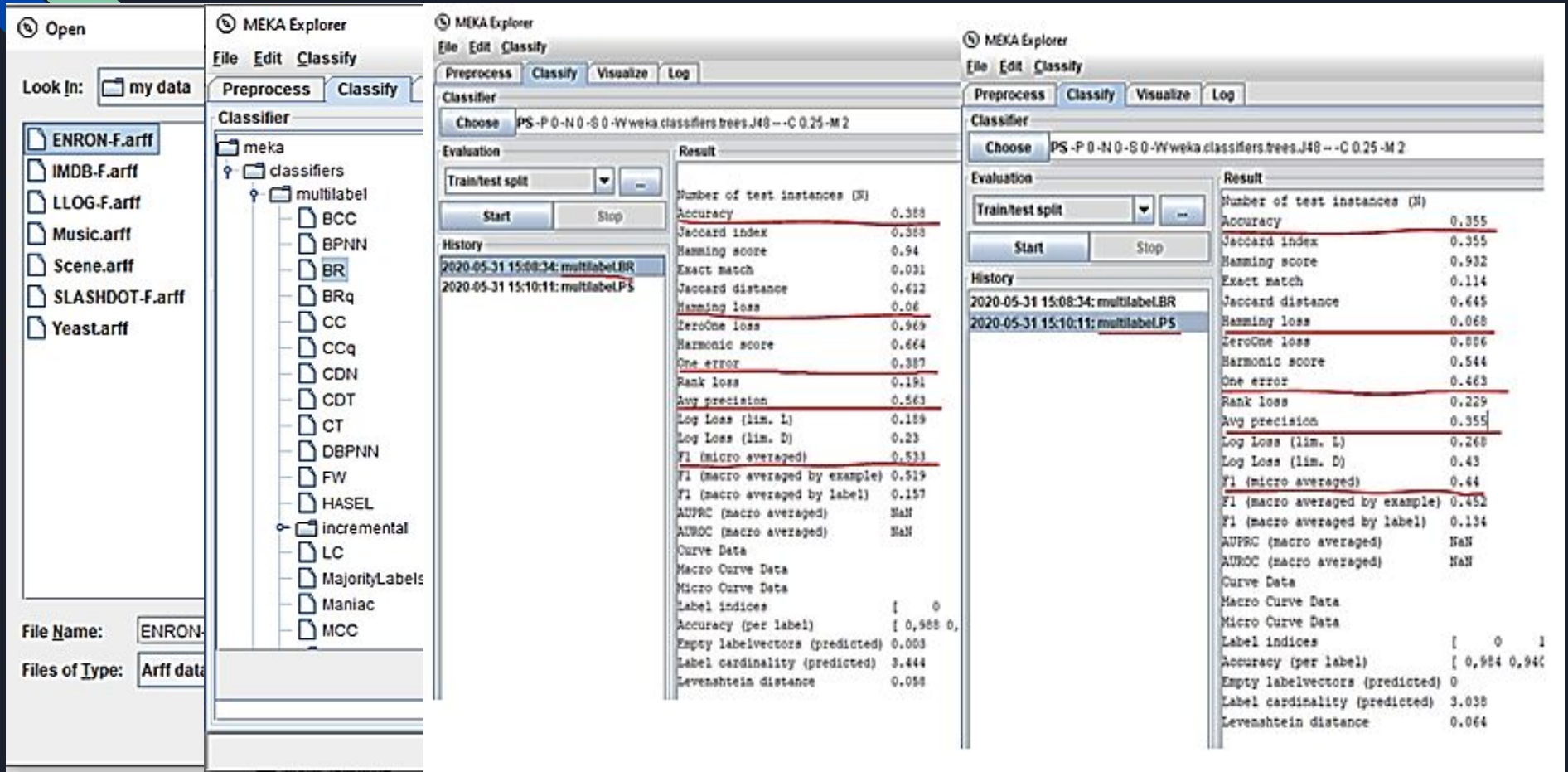
- **Класифікатори на основі трансформації**
  - **BR**
  - **CC**
  - **CLR**
  - **LP**
  - **PS**
- **Класифікатори на основі адаптації**
  - **ML-Tree**
  - **BP-MLL**
  - **Rank-SVM**
  - **BRkNN**
- **Класифікатори на основі ансамблю**
  - **ECC**
  - **EPS**
  - **RAkEL**

Характеристики наборів даних

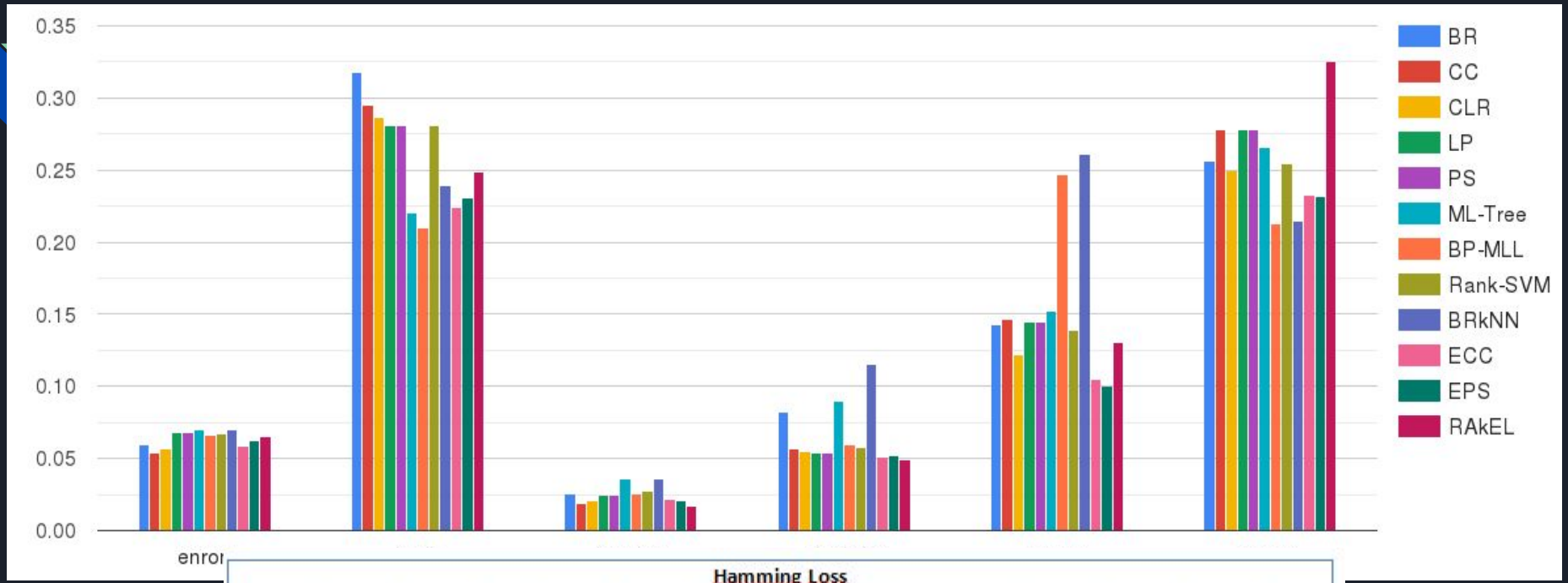
Було обрано шість різнорідних MLD. А саме:

- **enron**: це набір електронних повідомлень.
- **imdb**: набір коротких змістів сюжету фільмів.
- **langlog**: набір публікацій мовних журналів.
- **slashdot**: набір із заголовків новин та резюме.
- **scene**: набір знімків.
- **yeast**: набір генів.

Експерименти у MEKA



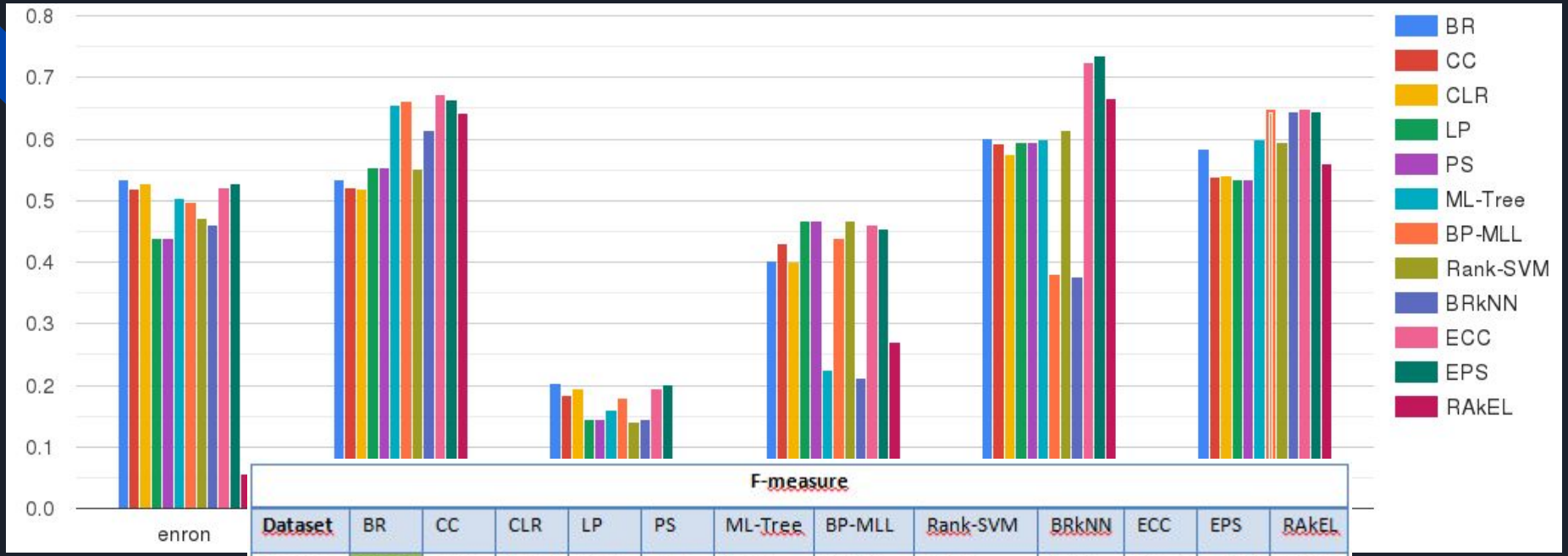
HLoss



Hamming Loss

Dataset	BR	CC	CLR	LP	PS	ML-Tree	BP-MLL	Rank-SVM	BRkNN	ECC	EPS	RAkEL
enron	0.06	**0.054**	0.057	0.068	0.068	0.07	0.066	0.067	0.07	0.059	0.062	0.065
imdb	0.318	0.295	0.287	0.281	0.281	0.22	**0.21**	0.281	0.239	0.224	0.231	0.249
langlog	0.026	0.019	0.021	0.025	0.025	0.036	0.026	0.027	0.036	0.022	0.021	**0.017**
slashdot	0.082	0.057	0.055	0.054	0.054	0.09	0.06	0.058	0.115	0.051	0.052	**0.049**
scene	0.143	0.147	0.122	0.145	0.145	0.152	0.247	0.139	0.261	0.105	**0.1**	0.131
yeast	0.256	0.278	0.25	0.278	0.278	0.266	**0.213**	0.254	0.215	0.233	0.232	0.325

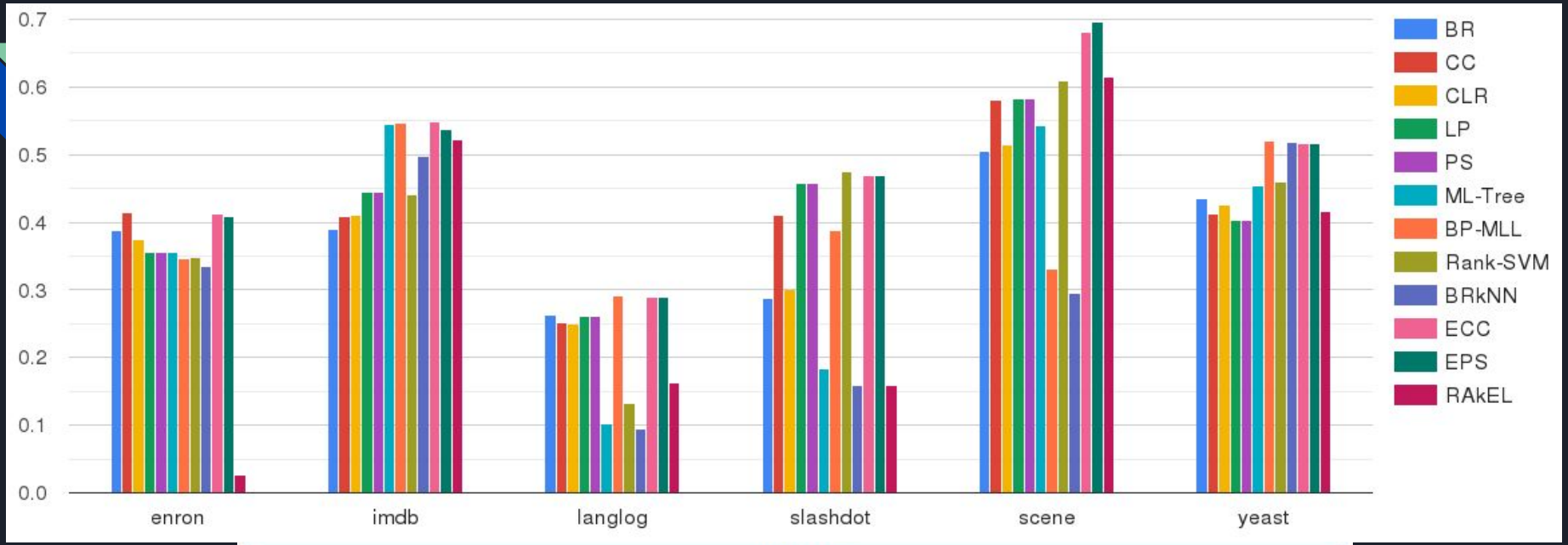
F-measure



F-measure

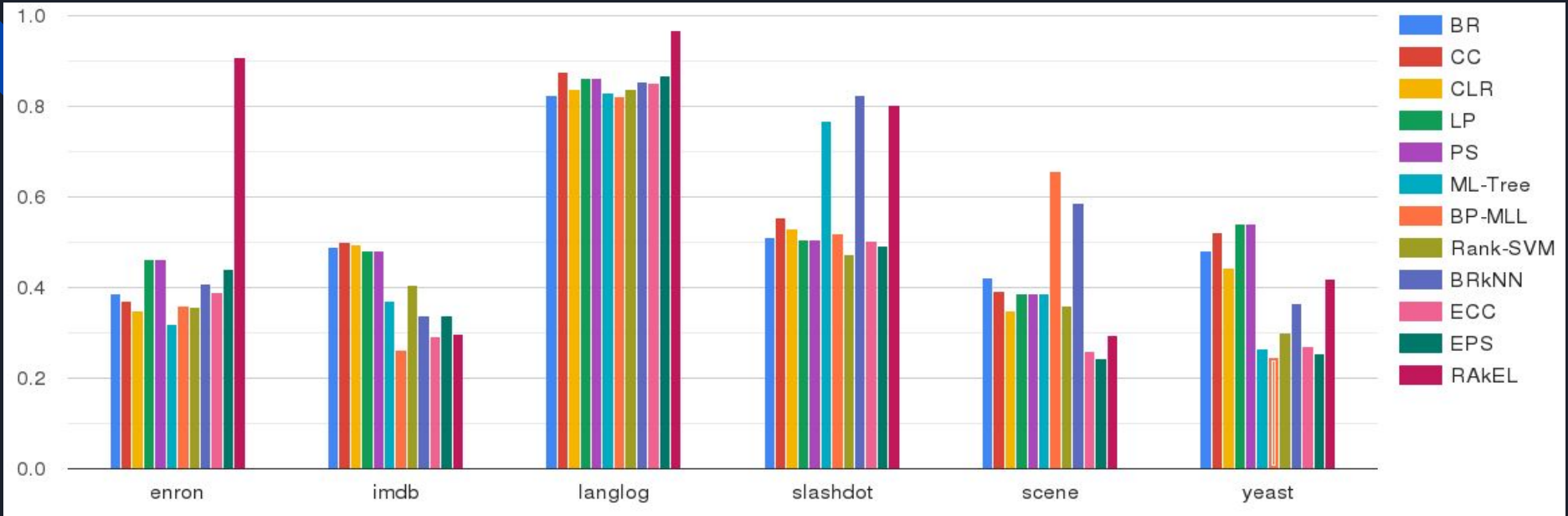
Dataset	BR	CC	CLR	LP	PS	ML-Tree	BP-MLL	Rank-SVM	BRkNN	ECC	EPS	RAkEL
enron	**0.533**	0.52	0.527	0.44	0.44	0.504	0.498	0.472	0.461	0.521	0.527	0.057
imdb	0.533	0.521	0.520	0.554	0.554	0.655	0.661	0.551	0.613	**0.672**	0.664	0.642
langlog	**0.203**	0.184	0.195	0.144	0.144	0.16	0.179	0.14	0.144	0.195	0.201	0.039
slashdot	0.402	0.431	0.4	0.466	0.466	0.224	0.438	**0.467**	0.211	0.46	0.454	0.27
scene	0.602	0.593	0.575	0.594	0.594	0.599	0.38	0.613	0.376	0.725	**0.736**	0.667
yeast	0.583	0.539	0.54	0.535	0.535	0.6	**0.648**	0.595	0.645	**0.648**	0.644	0.561

Accuracy



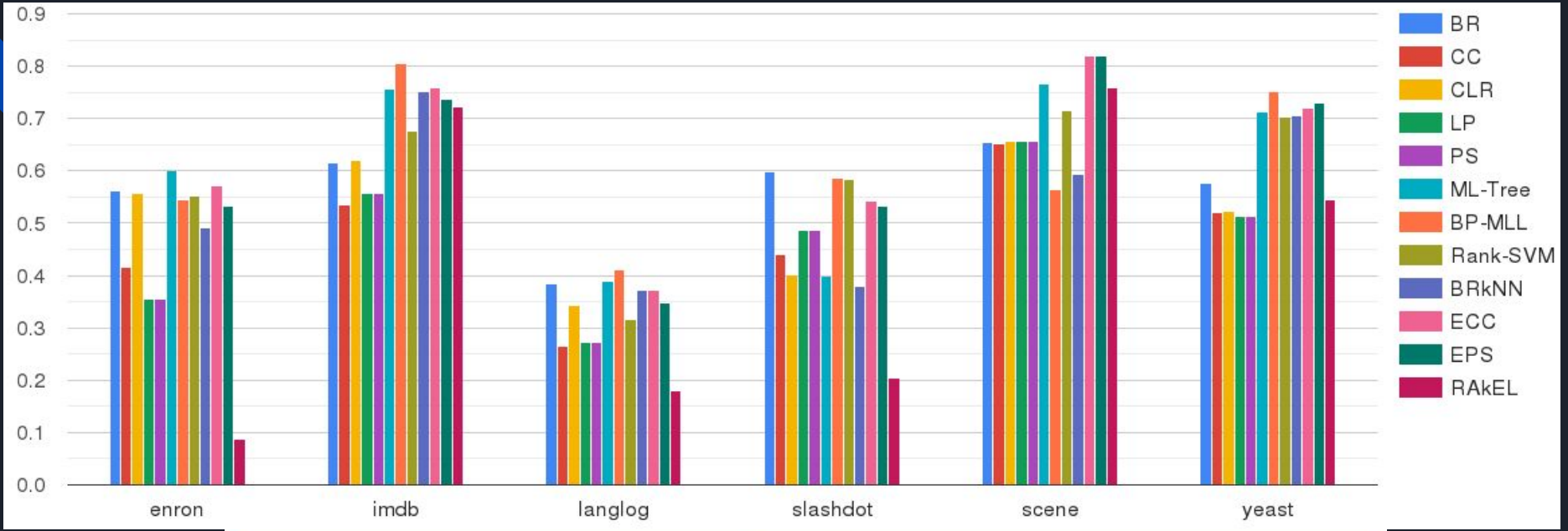
Accuracy												
Dataset	BR	CC	CLR	LP	PS	ML-Tree	BP-MLL	Rank-SVM	BRkNN	ECC	EPS	RAkEL
enron	0.388	**0.414**	0.375	0.355	0.355	0.355	0.347	0.348	0.334	0.412	0.409	0.027
imdb	0.39	0.408	0.410	0.445	0.445	0.545	0.546	0.441	0.497	**0.549**	0.538	0.523
langlog	0.263	0.251	0.25	0.261	0.261	0.103	**0.292**	0.132	0.095	0.289	0.29	0.163
slashdot	0.287	0.41	0.3	0.458	0.458	0.183	0.388	**0.474**	0.158	0.469	0.469	0.158
scene	0.506	0.581	0.515	0.583	0.583	0.543	0.332	0.609	0.296	0.682	**0.696**	0.614
yeast	0.435	0.413	0.425	0.403	0.403	0.455	0.521	0.459	**0.519**	0.517	0.517	0.416

One Error



One Error												
Dataset	BR	CC	CLR	LP	PS	ML-Tree	BP-MLL	Rank-SVM	BRkNN	ECC	EPS	RAkEL
enron	0.387	0.37	0.35	0.463	0.463	**0.32**	0.359	0.358	0.408	0.39	0.44	0.907
imdb	0.49	0.5	0.495	0.48	0.48	0.371	**0.262**	0.406	0.337	0.292	0.337	0.297
langlog	0.825	0.877	0.837	0.863	0.863	0.831	**0.821**	0.837	0.855	0.851	0.867	0.968
slashdot	0.512	0.554	0.529	0.505	0.505	0.768	0.519	**0.472**	0.823	0.503	0.491	0.802
scene	0.422	0.391	0.349	0.387	0.387	0.386	0.657	0.359	0.586	0.26	**0.242**	0.295
yeast	0.481	0.522	0.444	0.541	0.541	0.264	**0.245**	0.299	0.365	0.269	0.254	0.42

Average Precision



Avg precision												
Dataset	BR	CC	CLR	LP	PS	ML-Tree	BP-MLL	Rank-SVM	BRkNN	ECC	EPS	RAkEL
enron	0.563	0.417	0.558	0.355	0.355	**0.601**	0.545	0.553	0.492	0.572	0.532	0.087
imdb	0.615	0.536	0.62	0.557	0.557	0.757	**0.805**	0.677	0.752	0.76	0.738	0.723
langlog	0.384	0.265	0.342	0.273	0.273	0.388	**0.41**	0.316	0.372	0.373	0.349	0.18
slashdot	**0.599**	0.441	0.401	0.487	0.487	0.4	0.587	0.583	0.379	0.542	0.533	0.205
scene	0.654	0.653	0.656	0.657	0.657	0.767	0.564	0.715	0.594	0.819	**0.82**	0.759
yeast	0.577	0.521	0.522	0.514	0.514	0.712	**0.752**	0.702	0.706	0.721	0.729	0.546

Test i Total time

Test time												
Dataset	BR	CC	CLR	LP	PS	ML-Tree	BP-MLL	Rank-SVM	BRkNN	ECC	EPS	RAkEL
enron	0.848	0.428	1.102	1.246	1.061	0.272	0.044	**0.163**	3.148	2.508	1.902	0.15
imdb	**0.002**	**0.002**	0.005	0.014	0.005	0.007	**0.002**	0.007	0.065	0.061	0.038	0.015
langlog	2.585	1.451	2.597	0.342	0.382	0.431	**0.045**	0.167	2.55	3.137	2.746	0.393
slashdot	0.069	0.053	1.015	1.771	1.818	0.166	0.07	0.219	**0.047**	2.395	2.001	0.097
scene	0.029	0.027	0.047	0.042	0.026	0.04	0.03	**0.012**	0.375	0.312	0.431	0.092
yeast	0.059	0.063	0.066	0.083	0.066	0.028	0.017	**0.016**	0.379	0.75	0.743	0.101

Total time												
Dataset	BR	CC	CLR	LP	PS	MLTree	BP-MLL	RankSVM	BRkNN	ECC	EPS	RAkEL
enron	201	275	206	37	**31**	240	61	269	125	560	501	107
imdb	0.328	0.312	0.452	0.198	**0.18**	14	2	0.287	0.419	17	22	1
langlog	149	227	155	**16**	**16**	216	57	28	188	333	300	75
slashdot	313	342	320	528	543	648	153	256	**105**	610	593	458
scene	6	6	7	5	**3**	201	28	**3**	4	35	34	25
yeast	6	5	6	6	5	65	14	8	**3**	40	31	18

Висновки

Провівши аналіз бачимо що з поміж усіх методів найбільше виділяються за заданими показниками BP-MLL, EPS, Rank-SVM. Трохи уступають їм але також повинні бути згадані BR, CC, ECC, ML-Tree.

Підсумки виконаної роботи

У даній роботі було:

- проаналізовано проблему мультикатегоріальної класифікації мультимедійних даних;
- надано структуровану презентацію методів, представлених у літературі;
- виконано серію експериментів порівнявши загалом 12 методів у 6 наборах даних із різних областей та з різними характеристиками за допомогою MEKA framework.

Майбутні напрямки роботи та досліджень

В майбутньому я маю намір:

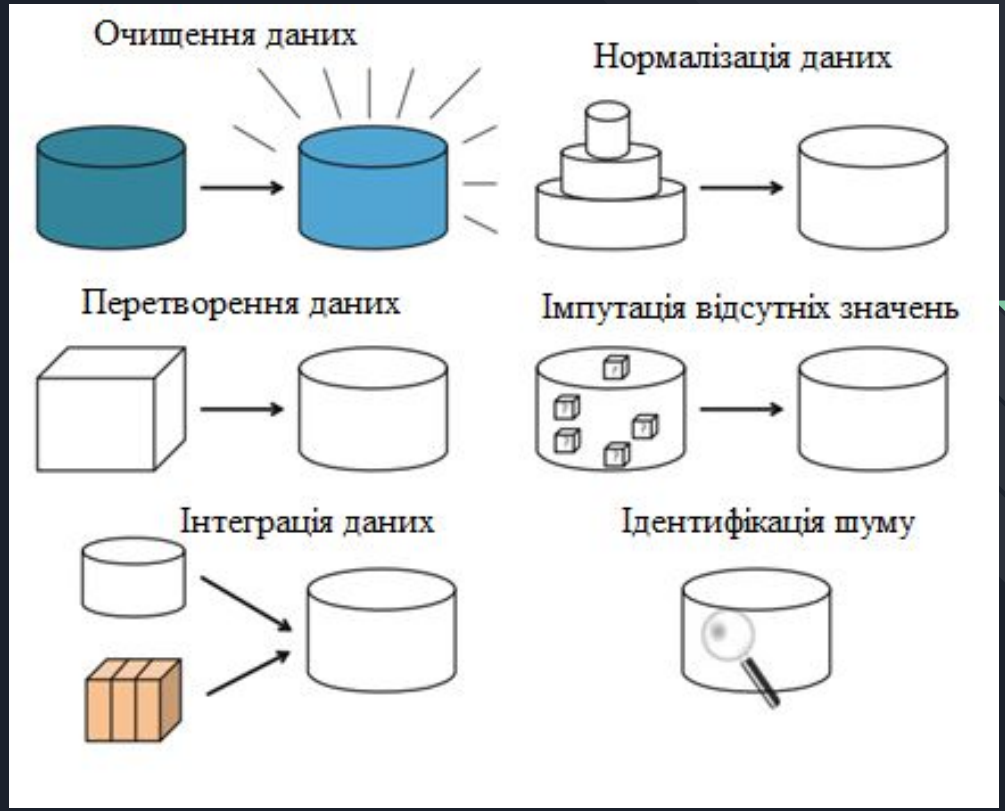
- здійснити детальну структуризацію різноманітних методів мультикатегоріальної класифікації;
- здійснити більш детальні експерименти з більшою кількістю наборів даних та методів для кращого розуміння факторів, обумовлюючих ефективність багатозначної класифікації в конкретних ситуаціях;
- програмно реалізувати Веб-сервіс анотування мультимедійних даних.

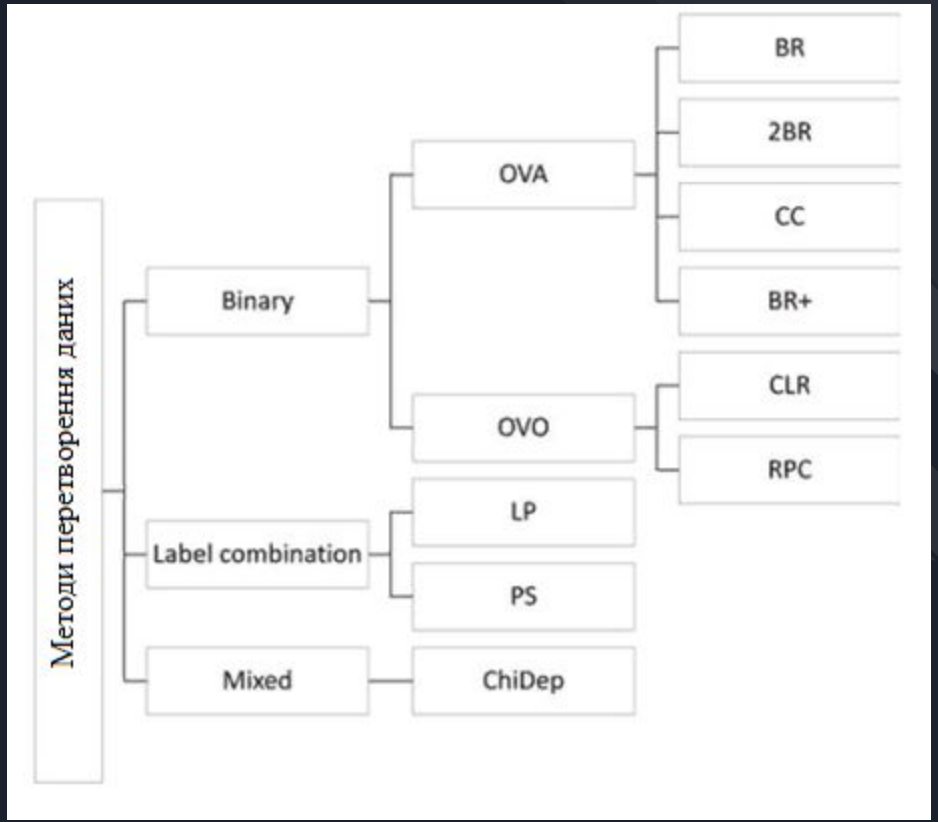
Дякую за увагу!

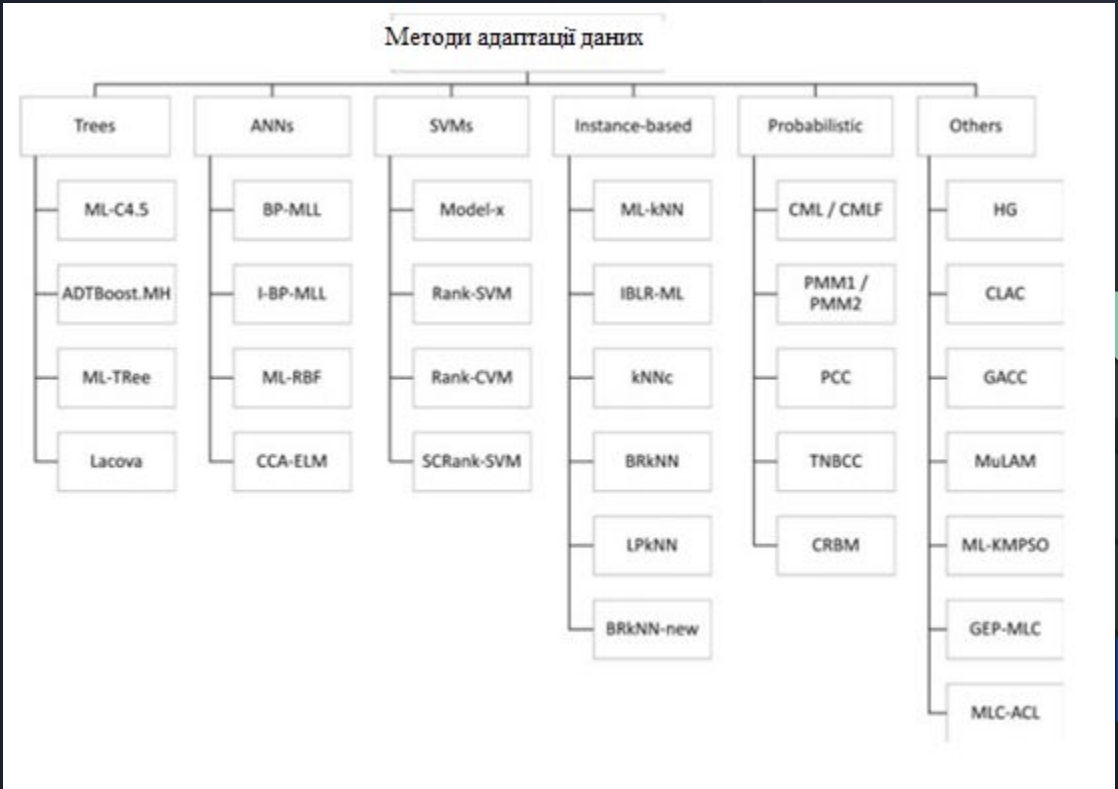

собака
біг
трава
дитина
сидить
чорна собака
дві собаки
собака біжить
біжать по траві
дитина біжить
дитина сидить
дві чорні собаки
дві собаки біжать
собака біжить по траві
дитина біжить по траві
дві собаки біжать по траві

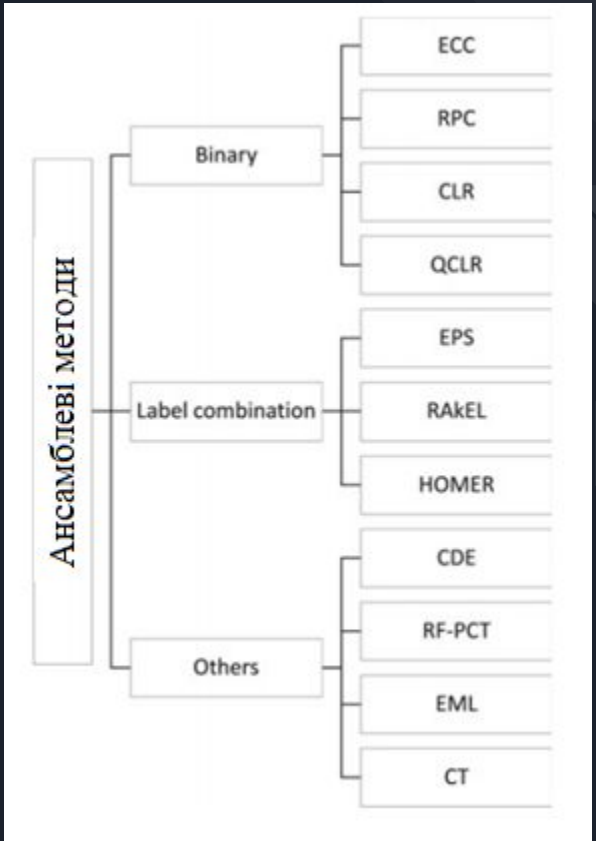

дві собаки біжать по траві


Бібліотека
Колекція
Справа
Анотація
Анотація
Малюнок
Анотація
Анотація
Малюнок
Справа
Анотація
Анотація
Колекція


Очищення даних
Нормалізація даних
Перетворення даних
Імпутація відсутніх значень
Інтеграція даних
Ідентифікація шуму


Методи перетворення даних
Binary
OVA
BR
2BR
CC
BR+
OVO
CLR
RPC
Label combination
LP
PS
Mixed
ChiDep


Методи адаптації даних
Trees
ML-C4.5
ADTBoost.MH
ML-TRee
Lacova
ANNs
BP-MLL
I-BP-MLL
ML-RBF
CCA-ELM
SVMs
Model-x
Rank-SVM
Rank-CVM
SCRank-SVM
Instance-based
ML-kNN
IBLR-ML
kNNc
BRkNN
LPkNN
BRkNN-new
Probabilistic
CML / CMLF
PMM1 / PMM2
PCC
TNBCC
CRBM
Others
HG
CLAC
GACC
MuLAM
ML-KMPSO
GEP-MLC
MLC-ACL


Ансамблеві методи
Binary
ECC
RPC
CLR
QCLR
Label combination
EPS
RAkEL
HOMER
Others
CDE
RF-PCT
EML
CT