

SELECTED TOPIC IN -AI ASSIGNMENT #1

Team Members

Moustafa Omar Mohammed

20200542

Abdelrahmman Ramadan Aboelela

20200284

Ahmed Hany Ibrahim

20200054

Ezz El-Din Ahmed

20200325

TOPICS FOR DISCUSSION

01 Introduction

02 Libraries

03 Datasets

04 Reports

01 Introduction

Imagine you're training a machine learning model, but labeling your data requires significant time and resources. Active learning offers a solution to this challenge by intelligently selecting the most informative data points for labeling, thus maximizing the efficiency of the learning process.

02 Libraries

scikit-activeml



Implementing an uncertainty sampling strategy for selecting samples with high uncertainty.

Implementing a random sampling strategy for randomly selecting samples.

Representing missing labels in active learning scenarios.

scikit-learn



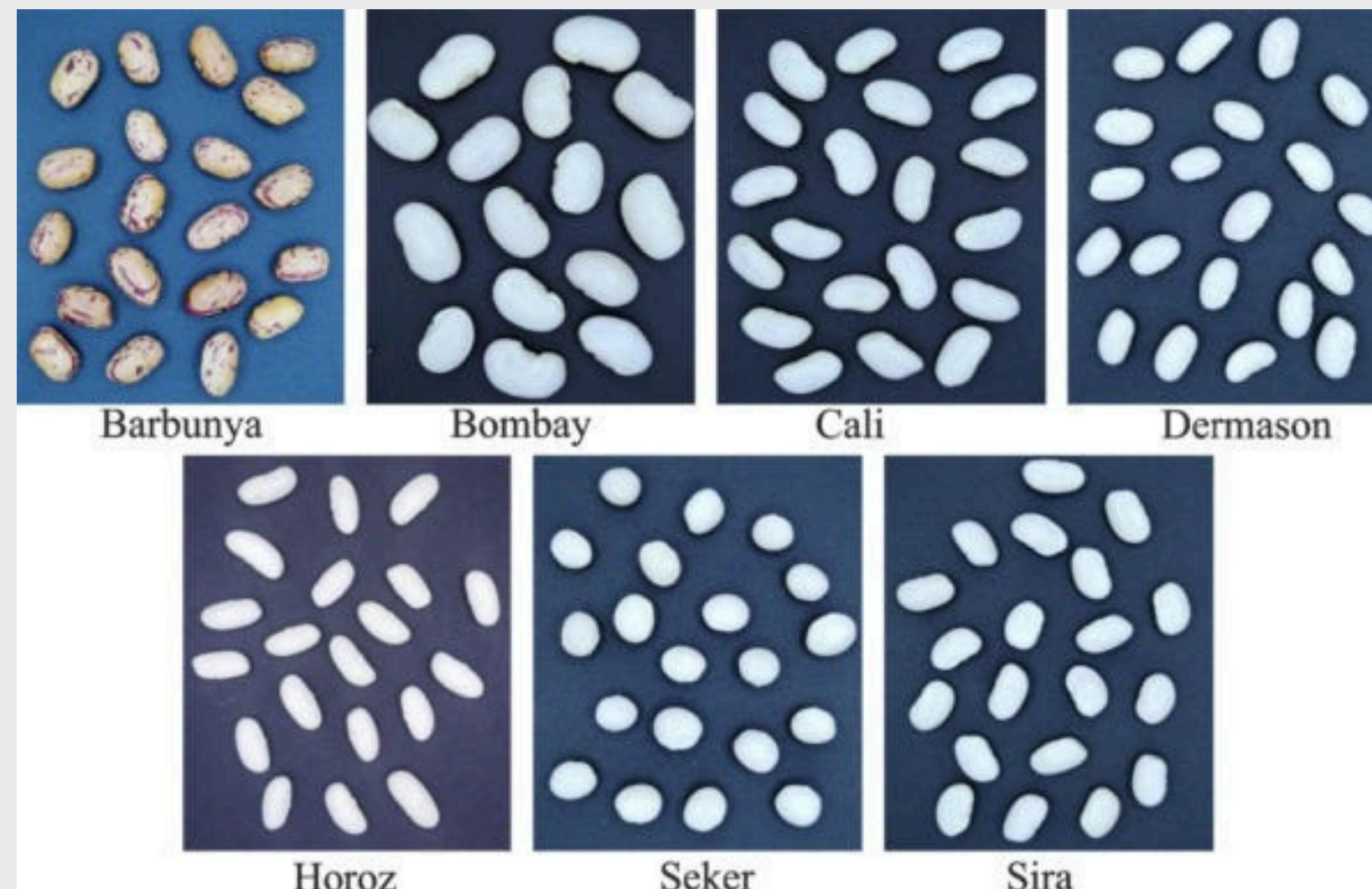
- Evaluation metric
- Data splitting
- Label encoding
- Feature scaling



03 Datasets (Balanced)

Dry Bean Dataset

Images of 13,611 grains of 7 different registered dry beans were taken with a high-resolution camera. A total of 16 features; 12 dimensions and 4 shape forms, were obtained from the grains



03 Datasets cont.

Features Information:

- 1.) Area (A):
- 2.) Perimeter (P):
- 3.) Major axis length (L)
- 4.) Minor axis length (l)
- 5.) Aspect ratio (K)

⋮

- 9.) Extent (Ex)
- 10.) Solidity (S)
- 11.) Roundness (R)

Classes:

(Seker, Barbunya, Bombay, Cali, Dermosan, Horoz and Sira)

Length	MinorAxisLength	AspectRatio	Eccentricity	ConvexArea	EquivDiameter	Extent
117	173.888747	1.197191	0.549812	28715	190.141097	0.763923
796	182.734419	1.097356	0.411785	29172	191.272750	0.783968
130	175.931143	1.209713	0.562727	29690	193.410904	0.778113
999	182.516516	1.153638	0.498616	30724	195.467062	0.782681
882	190.279279	1.060798	0.333680	30417	195.896503	0.773098

.....

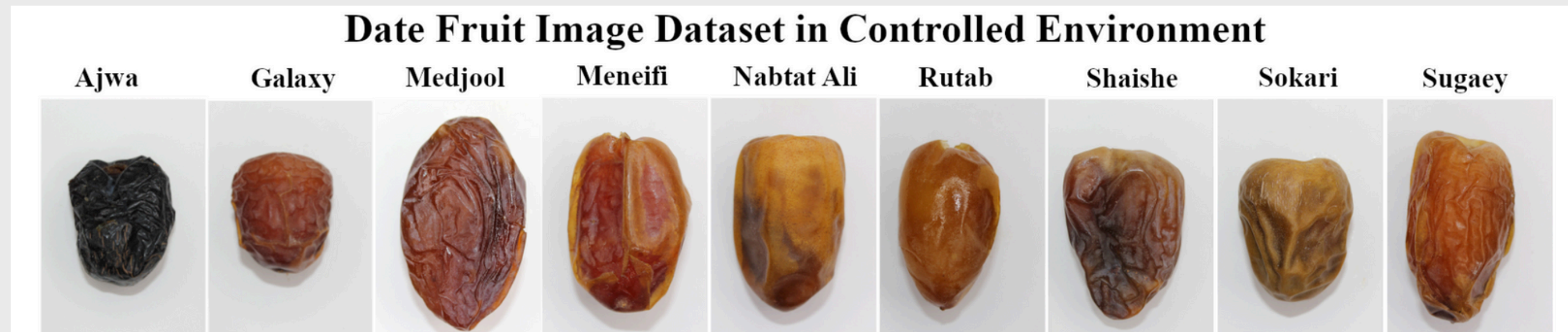
roundness	Compactness	ShapeFactor1	ShapeFactor2	ShapeFactor3	ShapeFactor4	Class
0.958027	0.913358	0.007332	0.003147	0.834222	0.998724	SEKER
0.887034	0.953861	0.006979	0.003564	0.909851	0.998430	SEKER
0.947849	0.908774	0.007244	0.003048	0.825871	0.999066	SEKER
0.903936	0.928329	0.007017	0.003215	0.861794	0.994199	SEKER
0.984877	0.970516	0.006697	0.003665	0.941900	0.999166	SEKER

03

Datasets cont.

Date Fruit Datasets

A great number of fruits are grown around the world, each of which has various types. The factors that determine the type of fruit are the external appearance features such as color, length, diameter, and shape. The external appearance of the fruits is a major determinant of the fruit type. Determining the variety of fruits by looking at their external appearance may necessitate expertise, which is time-consuming and requires great effort. The aim of this study is to classify the types of date fruit



03

Datasets cont.

	AREA	PERIMETER	MAJOR_AXIS	MINOR_AXIS	ECCENTRICITY	EQDIASQ	SOLIDITY	CONVEX_AREA	EXTENT	ASPECT_RATIO	...	KurtosisRR	KurtosisRG	KurtosisRB
0	422163	2378.908	837.8484	645.6693	0.6373	733.1539	0.9947	424428	0.7831	1.2976	...	3.2370	2.9574	4.2287
1	338136	2085.144	723.8198	595.2073	0.5690	656.1464	0.9974	339014	0.7795	1.2161	...	2.6228	2.6350	3.1704
2	526843	2647.394	940.7379	715.3638	0.6494	819.0222	0.9962	528876	0.7657	1.3150	...	3.7516	3.8611	4.7192
3	416063	2351.210	827.9804	645.2988	0.6266	727.8378	0.9948	418255	0.7759	1.2831	...	5.0401	8.6136	8.2618
4	347562	2160.354	763.9877	582.8359	0.6465	665.2291	0.9908	350797	0.7569	1.3108	...	2.7016	2.9761	4.4146

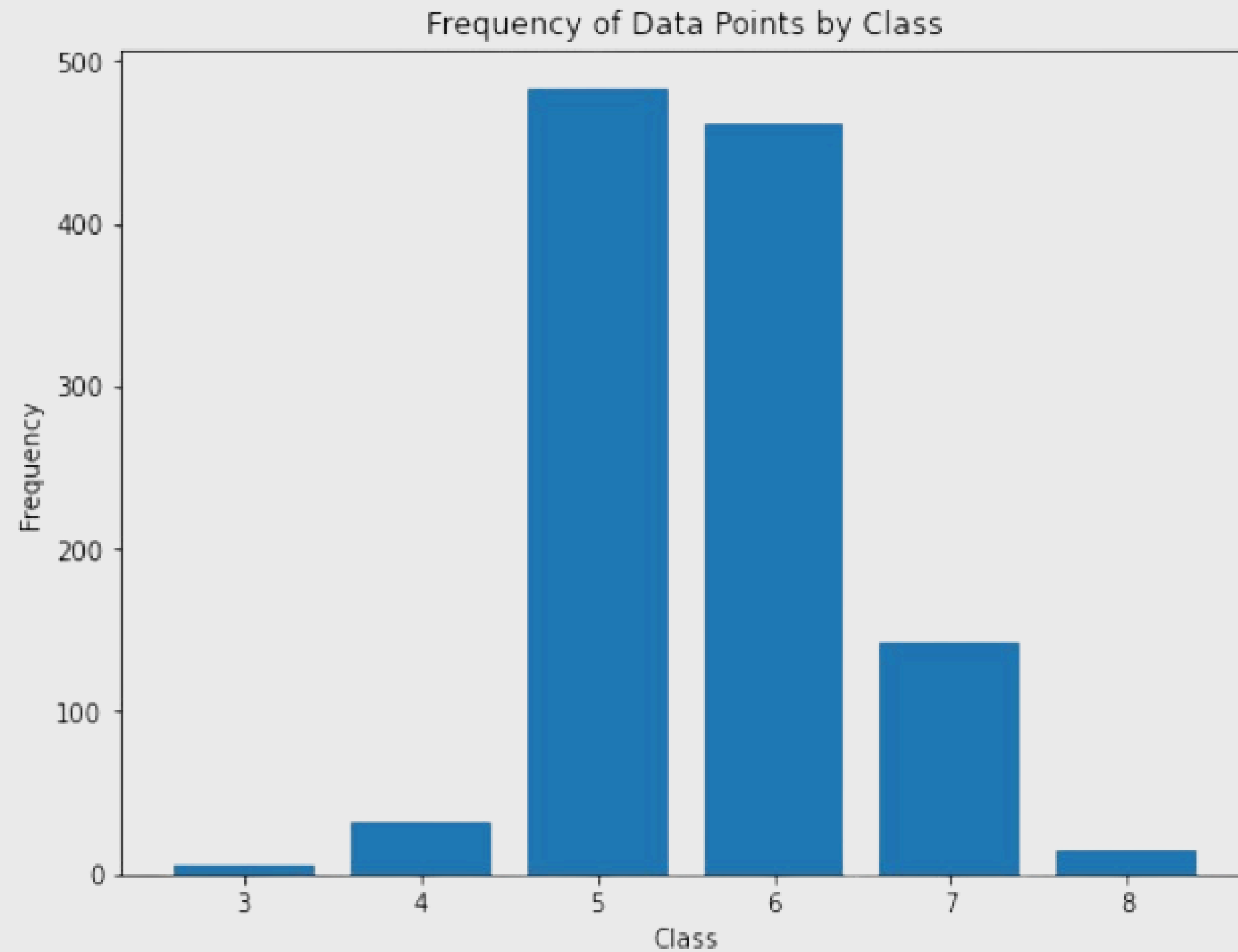
.....

EntropyRR	EntropyRG	EntropyRB	ALLdaub4RR	ALLdaub4RG	ALLdaub4RB	Class
-59191263232	-50714214400	-39922372608	58.7255	54.9554	47.8400	BERHI
-34233065472	-37462601728	-31477794816	50.0259	52.8168	47.8315	BERHI
-93948354560	-74738221056	-60311207936	65.4772	59.2860	51.9378	BERHI
-32074307584	-32060925952	-29575010304	43.3900	44.1259	41.1882	BERHI
-39980974080	-35980042240	-25593278464	52.7743	50.9080	42.6666	BERHI

03

Datasets cont. (Imbalanced)

Wine Quality



03

Datasets cont. (Imbalanced)

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality	Id
0	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5	0
1	7.8	0.88	0.00	2.6	0.098	25.0	67.0	0.9968	3.20	0.68	9.8	5	1
2	7.8	0.76	0.04	2.3	0.092	15.0	54.0	0.9970	3.26	0.65	9.8	5	2
3	11.2	0.28	0.56	1.9	0.075	17.0	60.0	0.9980	3.16	0.58	9.8	6	3
4	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5	4

1 - fixed acidity | 2 - volatile acidity | 3 - citric acid | 4 - residual sugar | 5 - chlorides | 6 - free sulfur dioxide |
7 - total sulfur dioxide | 8 - density | 9 - pH | 10 - sulphates | 11 - alcohol

Output:

Quality (score between 3 to 8)

04 Reports

Blanced Dataset



Passive Learning

Dataset 1 (Dry Bean Dataset):

	precision	recall	f1-score	support
BARBUNYA	0.93	0.90	0.91	261
BOMBAY	1.00	1.00	1.00	117
CALI	0.93	0.95	0.94	317
DERMASON	0.90	0.92	0.91	671
HOROZ	0.97	0.96	0.97	408
SEKER	0.96	0.94	0.95	413
SIRA	0.88	0.87	0.87	536
accuracy			0.92	2723
macro avg	0.94	0.93	0.94	2723
weighted avg	0.92	0.92	0.92	2723

04 Reports

Blanced Dataset

 non-Active
learning.
 Active
learning

Dataset 1 (Dry Bean Dataset):

Round 1:

- Random Sampling Training Accuracy: 71.27 %, Testing Accuracy: 70.33 %
- Least Confident Training Accuracy: 75.57 %, Testing Accuracy: 75.43 %
- Margin Sampling Training Accuracy: 84.30 %, Testing Accuracy: 84.69 %
- Entropy Sampling Training Accuracy: 85.88 %, Testing Accuracy: 85.79 %



Round 2:

- Random Sampling Training Accuracy: 85.19 %, Testing Accuracy: 85.16 %
- Least Confident Training Accuracy: 86.62 %, Testing Accuracy: 86.82 %
- Margin Sampling Training Accuracy: 87.12 %, Testing Accuracy: 87.07 %
- Entropy Sampling Training Accuracy: 87.64 %, Testing Accuracy: 87.04 %

⋮

04 Reports cont.

Blanced Dataset cont.

 non-Active
learning.
 Active
learning

Dataset 1 (Dry Bean Dataset) cont. :

Round 3:

- Random Sampling Training Accuracy: 89.47 %, Testing Accuracy: 88.95 %
- Least Confident Training Accuracy: 90.46 %, Testing Accuracy: 89.97 %
- Margin Sampling Training Accuracy: 90.96 %, Testing Accuracy: 90.41 %
- Entropy Sampling Training Accuracy: 91.08 %, Testing Accuracy: 91.19 %

Round 4:

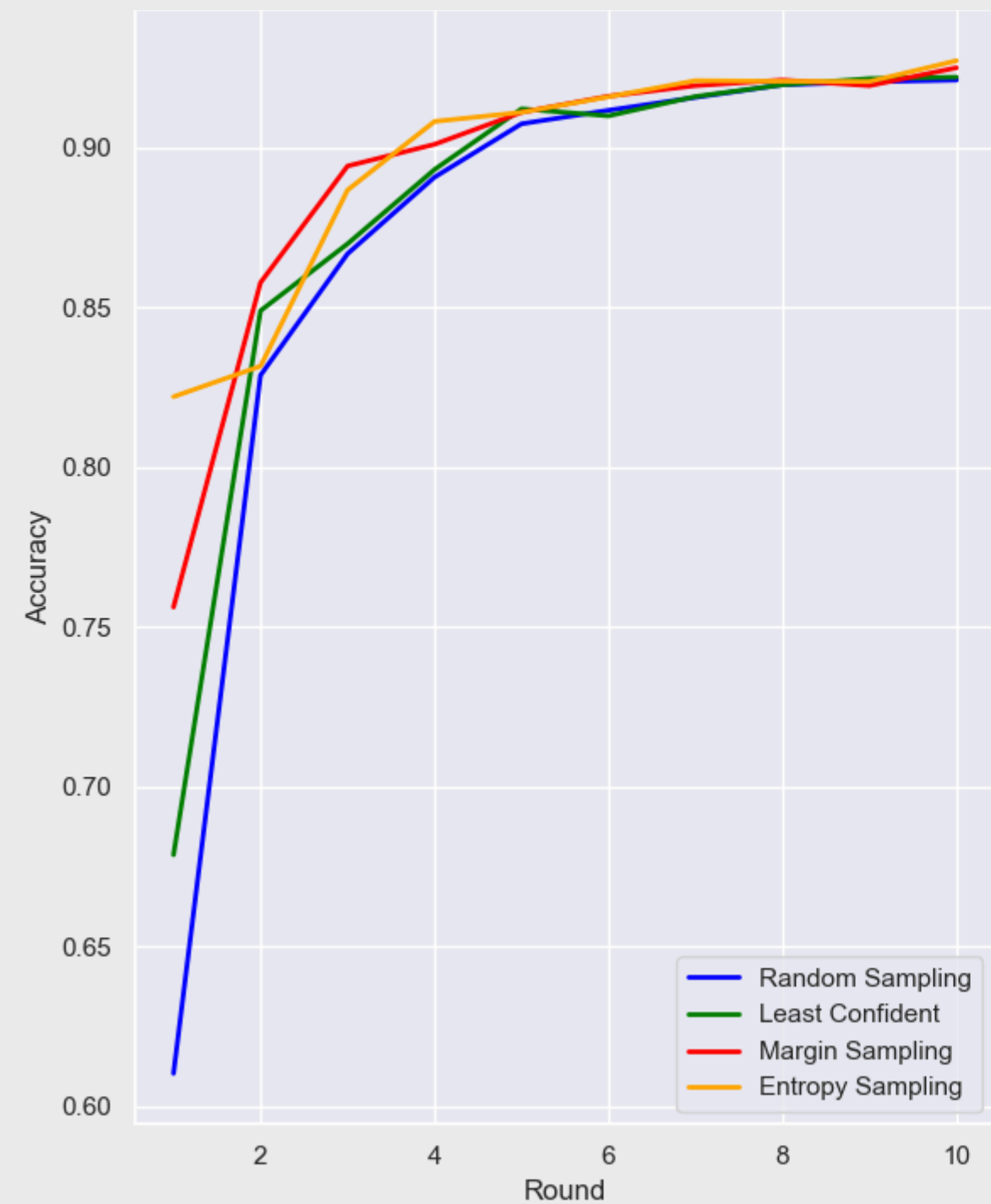
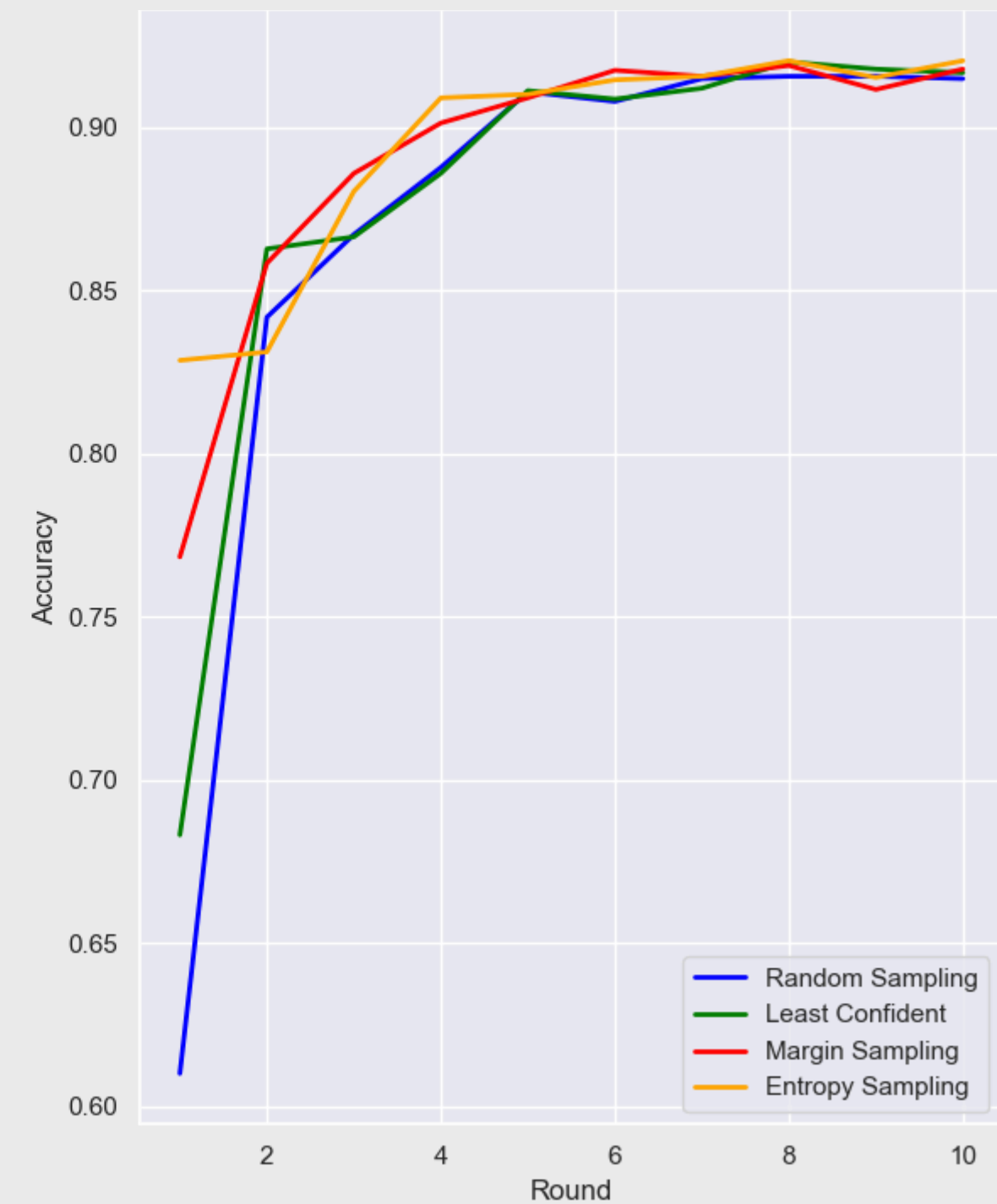
- Random Sampling Training Accuracy: 90.92 %, Testing Accuracy: 90.75 %
- Least Confident Training Accuracy: 91.36 %, Testing Accuracy: 91.33 %
- Margin Sampling Training Accuracy: 91.52 %, Testing Accuracy: 91.30 %
- Entropy Sampling Training Accuracy: 91.34 %, Testing Accuracy: 91.30 %

04

Reports cont.

Blanced Dataset cont.

Dataset 1 (Dry Bean Dataset) cont. :

TrainTest

04 Reports cont.

Blanced Dataset cont.



Passive Learning

Dataset 2 (Date Fruit Datasets) :

	precision	recall	f1-score	support
BERHI	0.86	0.67	0.75	9
DEGLET	0.88	0.52	0.65	29
DOKOL	0.83	0.94	0.88	36
IRAQI	0.80	0.92	0.86	13
ROTANA	0.95	0.97	0.96	40
SAFAVI	0.98	0.98	0.98	43
SOGAY	0.50	0.80	0.62	10
accuracy			0.87	180
macro avg	0.83	0.83	0.81	180
weighted avg	0.88	0.87	0.86	180

04 Reports cont.

Blanced Dataset cont.

 non-Active
learning.
 Active
learning

Dataset 2 (Date Fruit Datasets) :

Round 1:

◦ Random Sampling Training Accuracy: 72.01 % Testing Accuracy: 73.89 %

◦ Least Confident Training Accuracy: 74.09 %, Testing Accuracy: 75.56 %

◦ Margin Sampling Training Accuracy: 80.50 %, Testing Accuracy: 82.78 %

◦ Entropy Sampling Training Accuracy: 82.87 %, Testing Accuracy: 83.33 %

Round 2:

◦ Random Sampling Training Accuracy: 83.70 % Testing Accuracy: 82.22 %

◦ Least Confident Training Accuracy: 86.91 %, Testing Accuracy: 83.89 %



◦ Margin Sampling Training Accuracy: 88.30 %, Testing Accuracy: 85.56 %

◦ Entropy Sampling Training Accuracy: 89.97 %, Testing Accuracy: 86.67 %

⋮

04 Reports cont.

Blanced Dataset cont.

 non-Active
learning.
 Active
learning

Dataset 2 (Date Fruit Datasets) cont. :

Round 3:

- Random Sampling Training Accuracy: 89.14 % , Testing Accuracy: 86.11 %
- Least Confident Training Accuracy: 93.04 % , Testing Accuracy: 90.56 %
- Margin Sampling Training Accuracy: 93.45 % , Testing Accuracy: 90.56 %
- Entropy Sampling Training Accuracy: 94.29 % , Testing Accuracy: 88.89 %

Round 4:

- Random Sampling Training Accuracy: 94.99 % , Testing Accuracy: 89.44 %
- Least Confident Training Accuracy: 96.38 % , Testing Accuracy: 90.00 %
- Margin Sampling Training Accuracy: 96.66 % , Testing Accuracy: 90.00 %
- Entropy Sampling Training Accuracy: 96.80 % , Testing Accuracy: 92.22 %

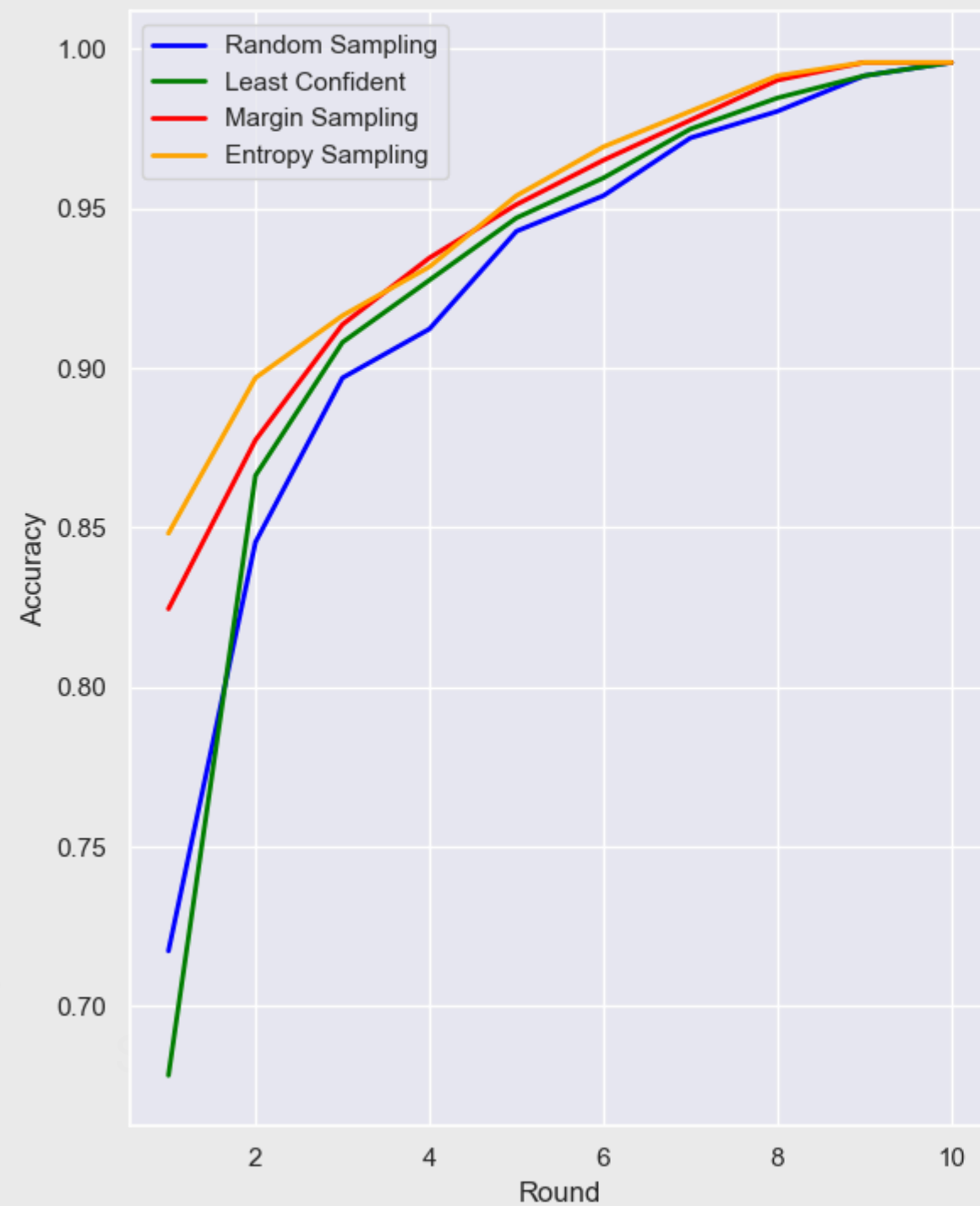
04

Reports cont.

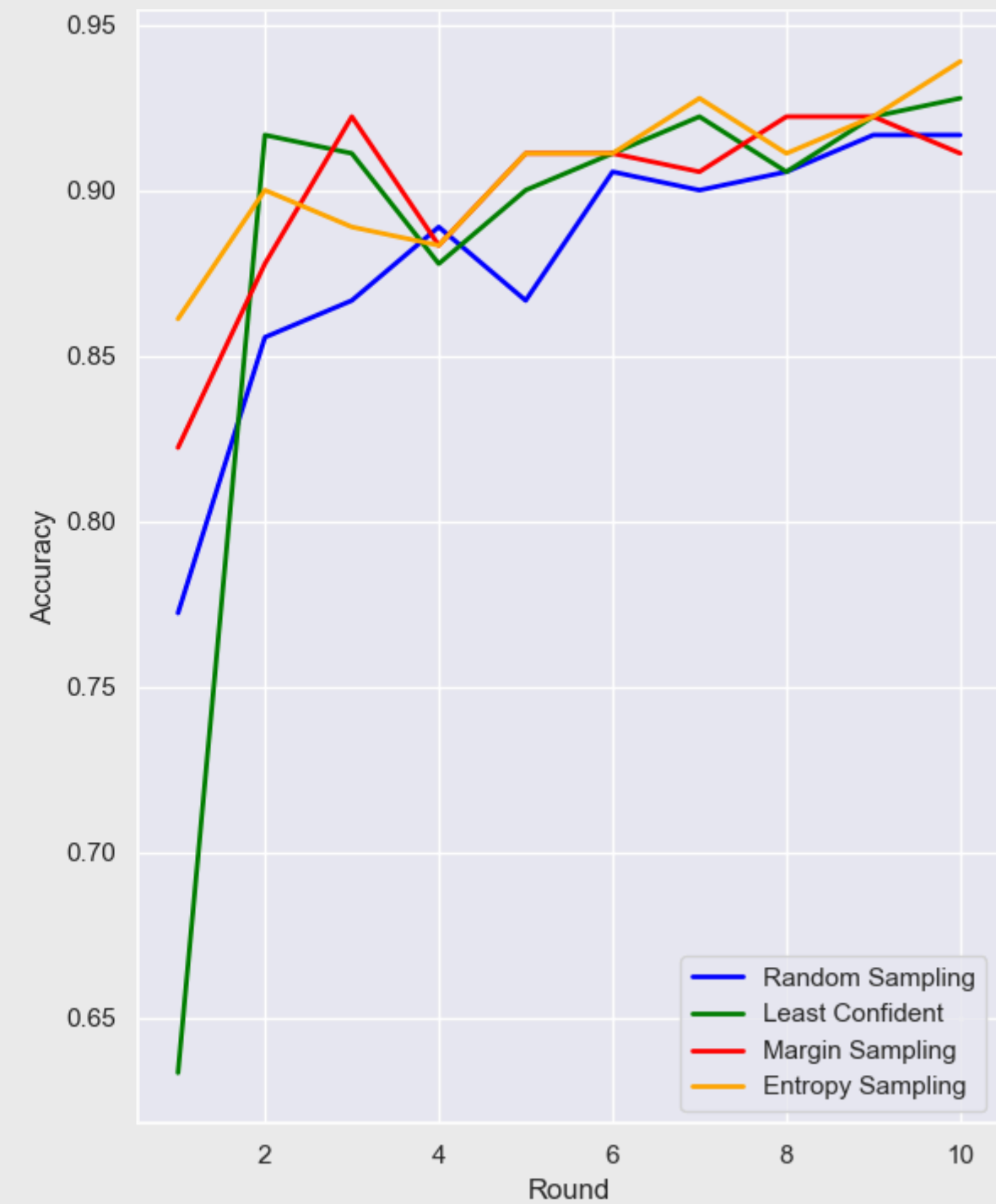
Blanced Dataset cont.

Dataset 2 (Date Fruit Datasets) cont. :

Train



Test



04 Reports cont.

Imbalanced Dataset

Dataset (Wine Quality) :

Passive Learning

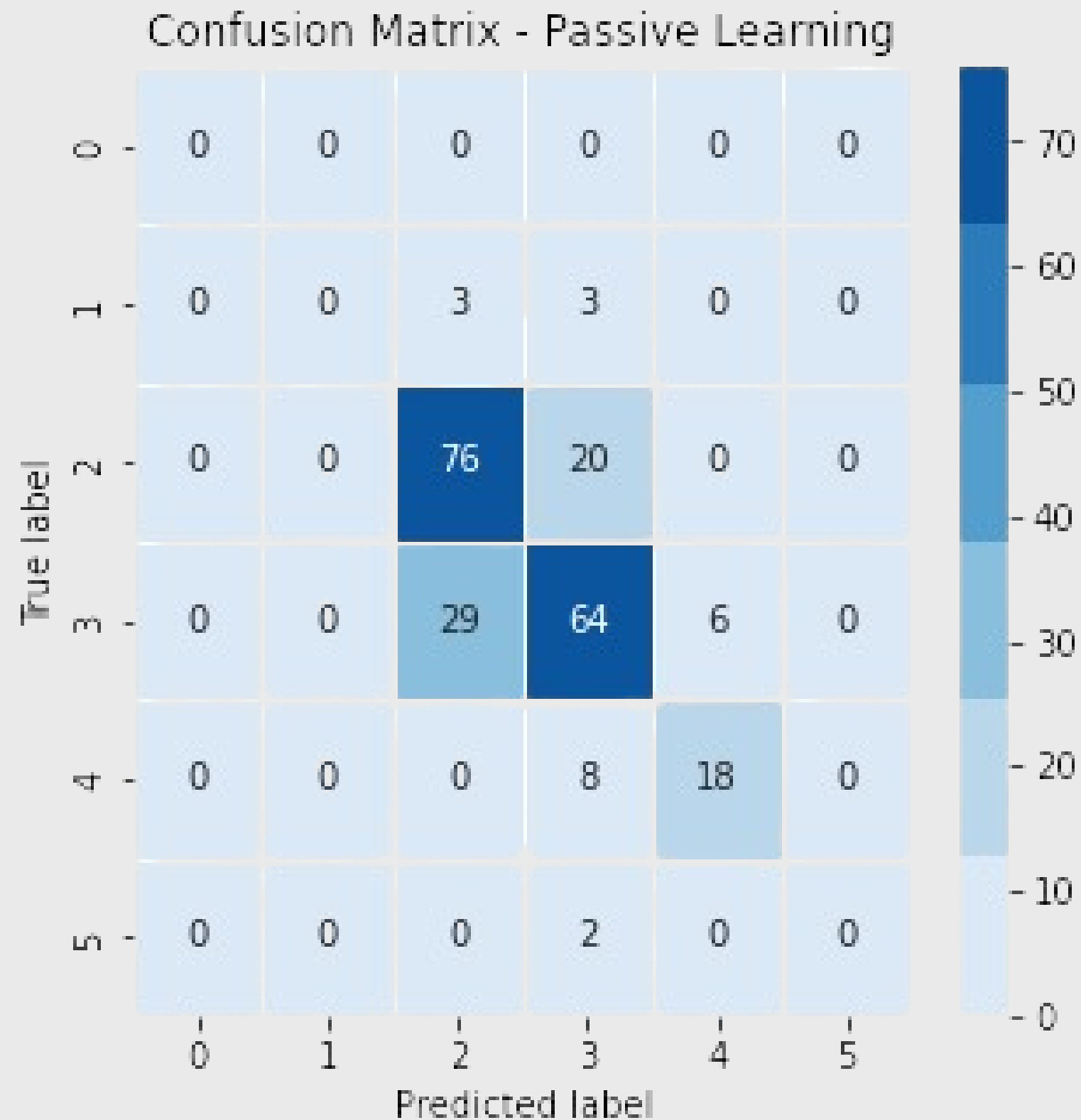
	precision	recall	f1-score	support
1	0.00	0.00	0.00	6
2	0.70	0.79	0.75	96
3	0.66	0.65	0.65	99
4	0.75	0.69	0.72	26
5	0.00	0.00	0.00	2
accuracy			0.69	229
macro avg	0.42	0.43	0.42	229
weighted avg	0.67	0.69	0.68	229

04 Reports cont.

Imbalanced Dataset



Dataset (Wine Quality) :

Passive Learning



04 Reports cont.

Imbalanced Dataset

 non-Active
learning.
 Active
learning

Dataset (Wine Quality) :

Round 1:

◦ Random Sampling - Training Accuracy: 47.59 % , Testing Accuracy: 50.22 %

◦ Least Confident Training Accuracy: 49.23 % , Testing Accuracy: 49.78 %

◦ Margin Sampling Training Accuracy: 57.66 % , Testing Accuracy: 57.21 %

◦ Entropy Sampling Training Accuracy: 55.47 % , Testing Accuracy: 56.77 %

⋮

Round 9:

◦ Random Sampling Training Accuracy: 67.40 % , Testing Accuracy: 56.77 %

◦ Least Confident Training Accuracy: 70.35 % , Testing Accuracy: 62.45 %

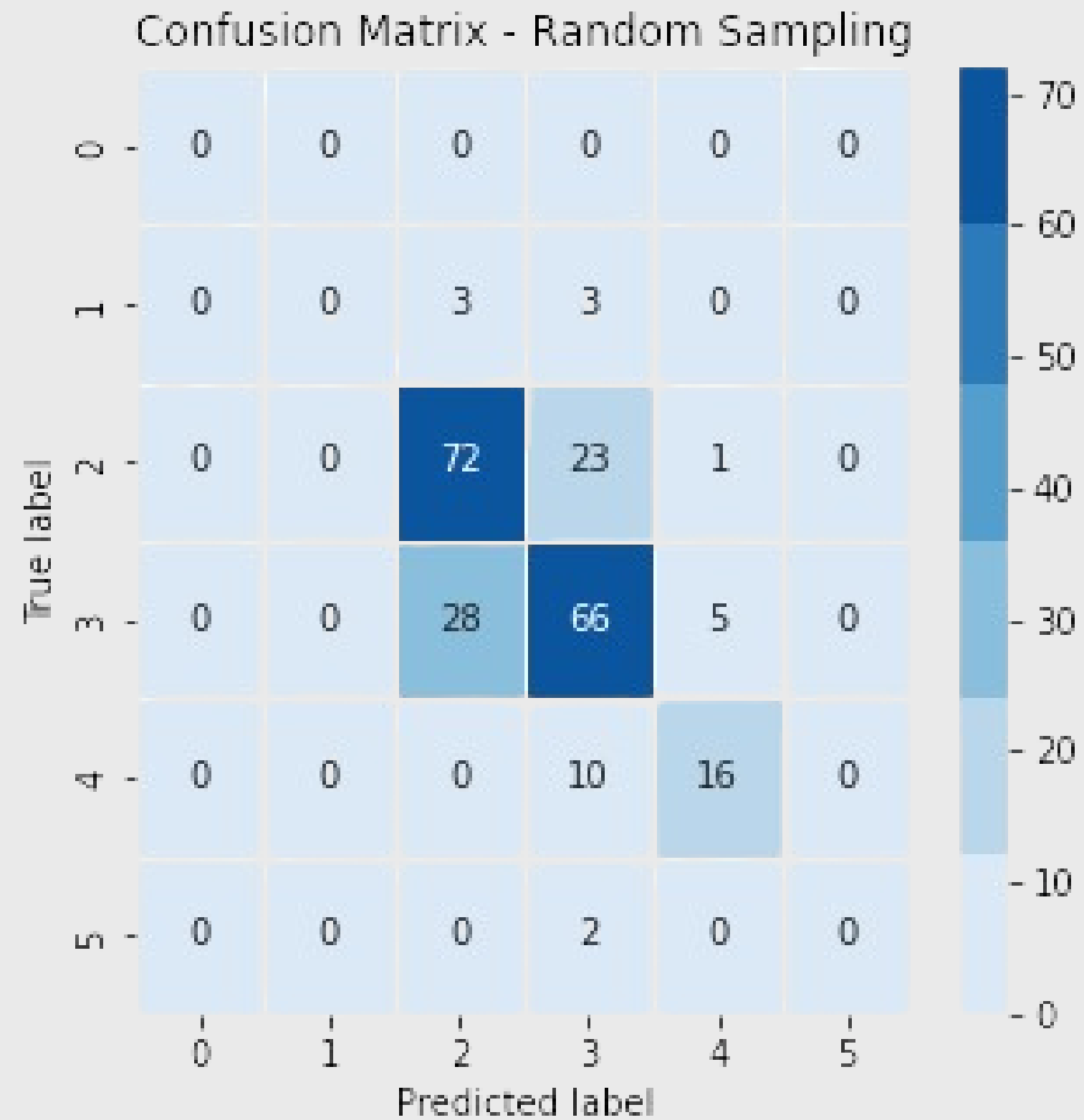
◦ Margin Sampling Training Accuracy: 70.79 % , Testing Accuracy: 61.14 %

◦ Entropy Sampling Training Accuracy: 70.35 % , Testing Accuracy: 60.70 %

04 Reports cont.

Imbalanced Dataset

Dataset (Wine Quality) :



04 Reports cont.

Imbalanced Dataset

Dataset (Wine Quality) :

