

Name: - Muhammad Huzaifa Waseem (2303-KHI-DEG-021)

Pair Partner 1: - Muhammad Faizan Rafique (2303.005.KHI.DEG)

Pair Partner 2: - Syed Muhammad Hammad Irshad(2303.KHI.DEG.032)

UNIT 3.4:

Assignment

Random Forest is a popular machine learning algorithm used for analyzing data. To optimize our Random Forest model, we use several tools and processes to streamline our workflow and ensure consistent results.

MLflow_env_vars.sh is one such tool, which allows us to provide environment variables for our project. This ensures that we are using the same environment across all stages of our workflow, from data preprocessing to model training and evaluation.

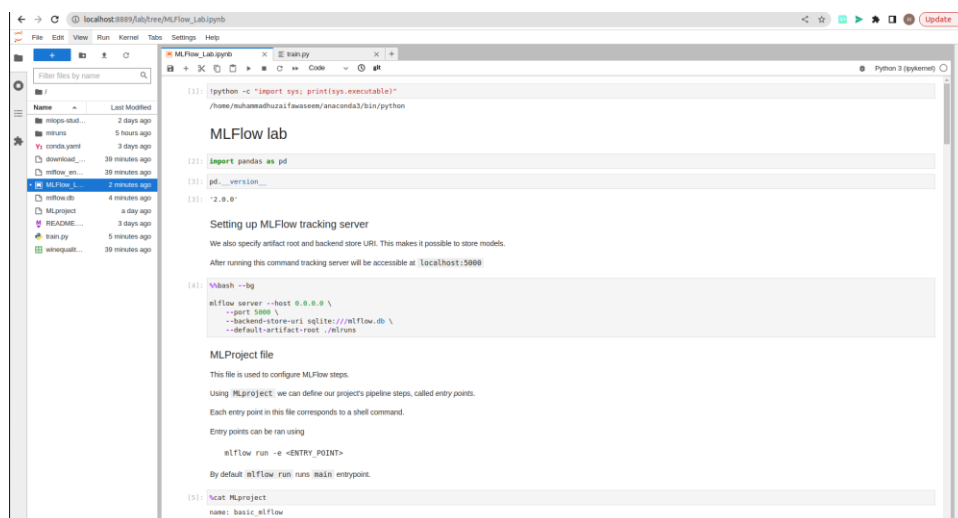
We also utilize a **conda.yaml** file to define dependencies for our project, which makes it easier to manage the libraries and packages required for our analysis.

To train our model, we use **Train.py**, define the necessary functions and parameters for our Random Forest model, including the parameters **max_n** and **max_d**. These parameters are used to define the maximum number of trees in the forest and the maximum depth of the trees. This ensures that our model is optimized for our specific data and use case.

Additionally, **Download_data.sh** is called in our code to import or download data in csv format. This script ensures that we have access to the necessary data required for our analysis.

Finally, we log all our work to ensure transparency and reproducibility. By utilizing these tools and processes, we are able to efficiently and effectively analyze our data using Random Forest.

MLFlow_lab.ipynb(code):



```
[1]: |python -c "import sys; print(sys.executable)"
/home/muhamadhuzaiwaseem/anaconda3/bin/python

MLFlow lab

[2]: |import pandas as pd

[3]: |pd.__version__

[4]: |Setting up MLFlow tracking server
We also specify artifact root and backend store URI. This makes it possible to store models.
After running this command tracking server will be accessible at localhost:5000

$bash --bg
mlflow server --host 0.0.0.0 \
--port 5000 \
--backend-store-uri sqlite://mlflow.db \
--default-artifact-root ./mlruns

MLProject file
This file is used to configure MLFlow steps.
Using MLPROJECT we can define our projects pipeline steps, called entry points.
Each entry point in this file corresponds to a shell command.
Entry points can be ran using
mlflow run -e <ENTRY_POINT>
By default mlflow run runs main entrypoint.

[5]: |cat MLproject
name: basic_mlflow
```

localhost:5589/lab/tree/MLFlow_Lab.ipynb

File Edit View Run Kernel Tabs Settings Help

MLFlow_Lab.ipynb x E train.py x + Python 3 (ipykernel)

151. `!cat Mlproject`

```
name: basic_mlflow

# this file is used to configure Python package dependencies.
# it uses Anaconda, but it can be also alternatively configured to use pip.
conda env: conda.yaml

# entry points can be ran using 'mlflow run -p <project_name> -e <entry_point_name>'
entry_points:
  download_data:
    # you can run any command using MLFlow
    command: "bash download_data.sh"
  # Mlproject file has to have main entry_point. It can be toggled without using -e option.
  main:
    # parameters is a key-value collection.
    parameters:
      file_name:
        type: str
        default: "winequality-red.csv"
      max_n:
        type: int
        default: 2
      max_d:
        type: int
        default: 2
    command: "python train.py (file_name) (max_n) (max_d)"

First we need to download data. We will use weather data from previous machine learning tutorial.
```

161. `!Vbash`

```
source mlflow_env_vars.sh
mlflow run --e download_data
```

2023/05/05 15:26:23 INFO mlflow.utils.config: Conda environment mlflow-d80f6d40409708131458f294963940da3af33 already exists.
2023/05/05 15:26:23 INFO mlflow.projects.utils: Created directory /tmp/tmp5k2zc9r for downloading remote URIs passed to arguments of type 'path' ===
2023/05/05 15:26:23 INFO mlflow.projects.backend.local: Running command 'source /home/muhamadhuzai/anaconda3/bin/.setprofile.d/conda.sh && conda activate mlflow-d80f6d40409708131458f294963940da3af33 && bash download_data.sh' in run with ID '12a1af12614c9095226226244f4' ===
File 'winequality-red.csv' already there, not retrieving.
2023/05/05 15:26:24 INFO mlflow.projects: Run ID '12a1af12614c9095226226244f4' succeeded ==

Training

Now we can train models. See 'train.py'. It contains code from supervised machine learning tutorial, we added tracking metrics and model.

We will train KNN models for $k \in \{1, 2, \dots, 10\}$ using temperature and casual features.

localhost:5589/lab/tree/MLFlow_Lab.ipynb

File Edit View Run Kernel Tabs Settings Help

MLFlow_Lab.ipynb x E train.py x + Python 3 (ipykernel)

Training

Now we can train models. See 'train.py'. It contains code from supervised machine learning tutorial, we added tracking metrics and model.

We will train KNN models for $k \in \{1, 2, \dots, 10\}$ using temperature and casual features.

After running this command you can go to 'localhost:5000' and see the trained models.

171. `!import sklearn`

181. `!sklearn.__version__`

191. `!1.2.2'`

201. `!python --version`

Python 3.10.9

1101. `!Vbash`

```
source mlflow_env_vars.sh
mlflow run --
```

2023/05/05 15:26:28 INFO mlflow.utils.config: Conda environment mlflow-d80f6d40409708131458f294963940da3af33 already exists.
2023/05/05 15:26:28 INFO mlflow.projects.utils: Created directory /tmp/tmp4p2t04f for downloading remote URIs passed to arguments of type 'path' ===
2023/05/05 15:26:28 INFO mlflow.projects.backend.local: Running command 'source /home/muhamadhuzai/anaconda3/bin/.setprofile.d/conda.sh && conda activate mlflow-d80f6d40409708131458f294963940da3af33 && python train.py winequality-red.csv 2' in run with ID '5ff2598105403485050808f4e84b07e' ===
/home/muhamadhuzai/anaconda3/envs/mlflow-d80f6d40409708131458f294963940da3af33/lib/python3.10/site-packages/distutils/backward_..._py33: UserWarning: Setuptools is replacing distutils.
WARNING:root:Setuptools is replacing distutils.
Registered model 'sklearn.RandomForest' already exists. Creating a new version of this model...
2023/05/05 15:26:32 INFO mlflow.tracking.model_registry.client: Waiting up to 300 seconds for model version to finish creation. Model name: sklearn.RandomForest, version 18
Created version '18' of model 'sklearn.RandomForest'.
2023/05/05 15:26:32 INFO mlflow.projects: Run ID '5ff2598105403485050808f4e84b07e' succeeded ==

Inspecting stored models

The trained models are stored in 'mlruns/0'.

These directories contain artifacts and config that is needed to serve them.

1121. `!Vbash`

```
last_model_paths=$(tr -s '\n' / | tail -1)
cat $last_model_paths | tr -s '\n' / | tail -1
```

artifact_path: RandomForest
flavors: {}
python_function: {}
conda: conda.yaml
virtualenv: python_env.yaml
loader_module: mlflow.sklearn
model_path: model.pkl
predict_fn: predict
python_version: 3.11.3
sklearn: {}
code: null
pickled_model: model.pkl
serialization_format: cloudpickle
sklearn_version: 1.2.2
mlflow_version: 2.9.1
model_uuid: ba16c70bac947f8a0c7c408597351
run_id: a2f6d45174709270570f5e205a
utc_time_created: '2023-05-05 05:05:28.533741'

localhost:5589/lab/tree/MLFlow_Lab.ipynb

File Edit View Run Kernel Tabs Settings Help

MLFlow_Lab.ipynb x E train.py x + Python 3 (ipykernel)

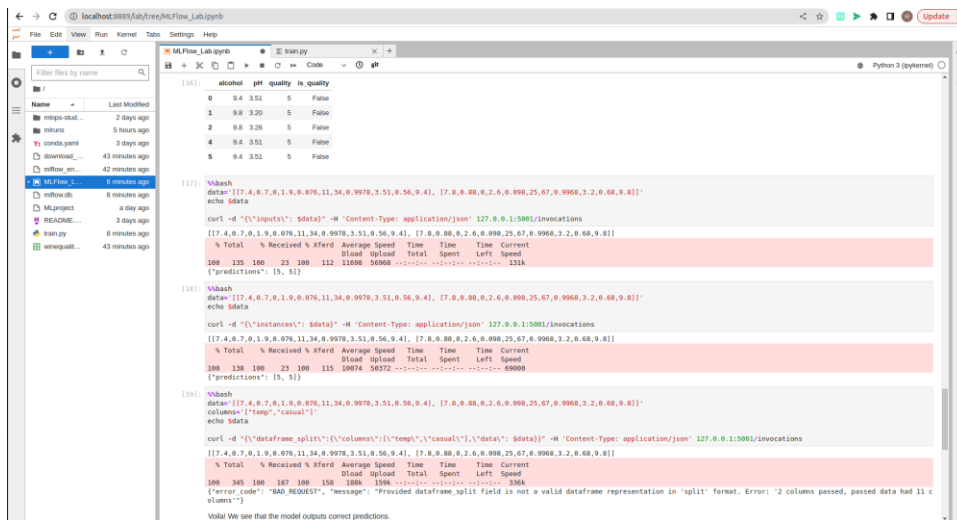
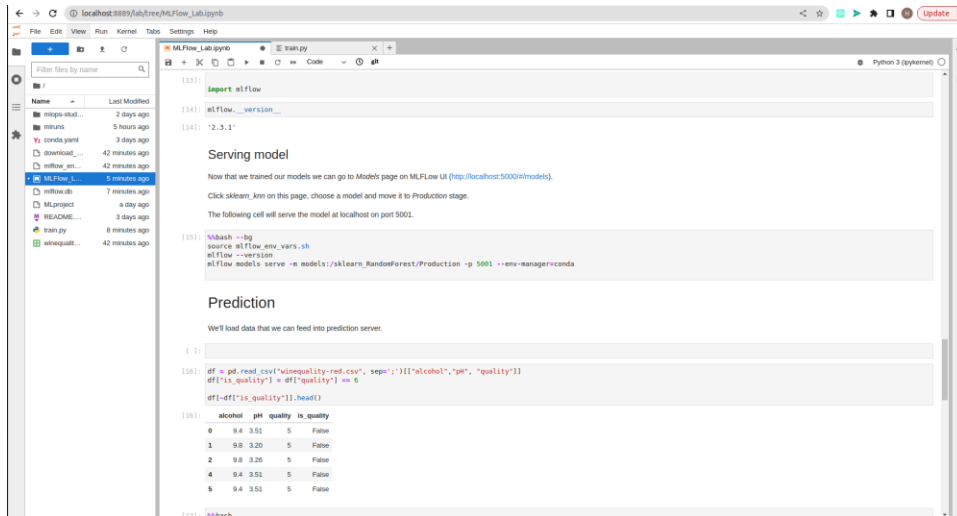
1121. `!Vbash`

```
last_model_paths=$(tr -s '\n' / | tail -1)
cat $last_model_paths | tr -s '\n' / | tail -1
```

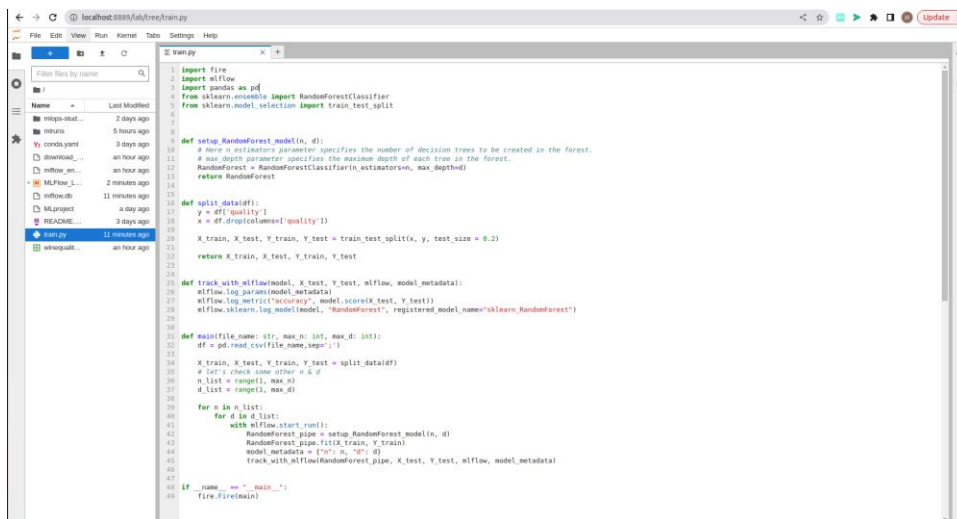
artifact_path: RandomForest
flavors: {}
python_function: {}
conda: conda.yaml
virtualenv: python_env.yaml
loader_module: mlflow.sklearn
model_path: model.pkl
predict_fn: predict
python_version: 3.11.3
sklearn: {}
code: null
pickled_model: model.pkl
serialization_format: cloudpickle
sklearn_version: 1.2.2
mlflow_version: 2.9.1
model_uuid: ba16c70bac947f8a0c7c408597351
run_id: a2f6d45174709270570f5e205a
utc_time_created: '2023-05-05 05:05:28.533741'

1122. `!pip install mlflow`

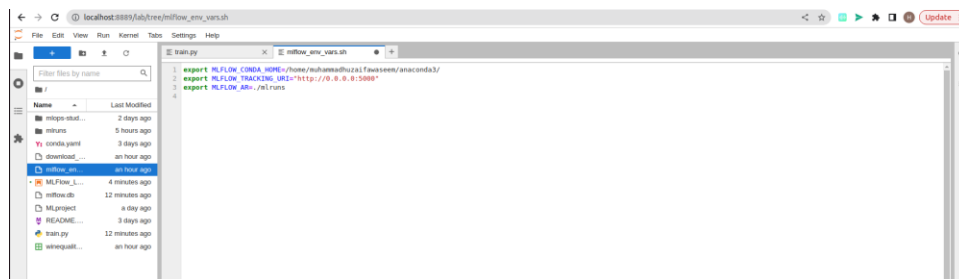
```
Requirement already satisfied: mlflow in /home/muhamadhuzai/anaconda3/lib/python3.10/site-packages (2.9.1)
Requirement already satisfied: click>=7.0 in /home/muhamadhuzai/anaconda3/lib/python3.10/site-packages (from mlflow) (8.0.4)
Requirement already satisfied: cloudpickle in /home/muhamadhuzai/anaconda3/lib/python3.10/site-packages (from mlflow) (2.2.1)
Requirement already satisfied: databricks-cli in /home/muhamadhuzai/anaconda3/lib/python3.10/site-packages (from mlflow) (0.17.4)
Requirement already satisfied: entrypoint in /home/muhamadhuzai/anaconda3/lib/python3.10/site-packages (from mlflow) (0.4)
Requirement already satisfied: gitpython>=2.1.0 in /home/muhamadhuzai/anaconda3/lib/python3.10/site-packages (from mlflow) (3.1.3)
Requirement already satisfied: gpytorch>=1.0.1 in /home/muhamadhuzai/anaconda3/lib/python3.10/site-packages (from mlflow) (1.8.4)
Requirement already satisfied: protobuf>=3.12.0 in /home/muhamadhuzai/anaconda3/lib/python3.10/site-packages (from mlflow) (4.22.3)
Requirement already satisfied: pytz>=2014 in /home/muhamadhuzai/anaconda3/lib/python3.10/site-packages (from mlflow) (2022.7)
Requirement already satisfied: requests>=2.17.3 in /home/muhamadhuzai/anaconda3/lib/python3.10/site-packages (from mlflow) (2.28.2)
Requirement already satisfied: packaging>=24 in /home/muhamadhuzai/anaconda3/lib/python3.10/site-packages (from mlflow) (23.0)
Requirement already satisfied: importlib-metadata>=4.7.0, <=4.7.0 in /home/muhamadhuzai/anaconda3/lib/python3.10/site-packages (from mlflow) (4.11.3)
Requirement already satisfied: sqlalchemy>=1.4.0 in /home/muhamadhuzai/anaconda3/lib/python3.10/site-packages (from mlflow) (1.4.4)
Requirement already satisfied: alembic>=1.10.4, <=1.10.4 in /home/muhamadhuzai/anaconda3/lib/python3.10/site-packages (from mlflow) (1.10.4)
Requirement already satisfied: docker>=4.0.0 in /home/muhamadhuzai/anaconda3/lib/python3.10/site-packages (from mlflow) (6.0.1)
Requirement already satisfied: flask in /home/muhamadhuzai/anaconda3/lib/python3.10/site-packages (from mlflow) (2.2.3)
Requirement already satisfied: numpy>=1.16.0 in /home/muhamadhuzai/anaconda3/lib/python3.10/site-packages (from mlflow) (1.22.5)
Requirement already satisfied: scipy in /home/muhamadhuzai/anaconda3/lib/python3.10/site-packages (from mlflow) (1.10.3)
Requirement already satisfied: pandas>=3 in /home/muhamadhuzai/anaconda3/lib/python3.10/site-packages (from mlflow) (2.0.0)
Requirement already satisfied: qystrys-parser2 in /home/muhamadhuzai/anaconda3/lib/python3.10/site-packages (from mlflow) (1.2.4)
Requirement already satisfied: sqlalchemy>=1.4.0 in /home/muhamadhuzai/anaconda3/lib/python3.10/site-packages (from mlflow) (1.4.4)
Requirement already satisfied: pyarrow>=0.17.0 in /home/muhamadhuzai/anaconda3/lib/python3.10/site-packages (from mlflow) (11.0.0)
Requirement already satisfied: markdown>=3.3 in /home/muhamadhuzai/anaconda3/lib/python3.10/site-packages (from mlflow) (3.4.3)
Requirement already satisfied: matplotlib in /home/muhamadhuzai/anaconda3/lib/python3.10/site-packages (from mlflow) (3.7.1)
```



Train.py:



MLflow_env_vars.sh:



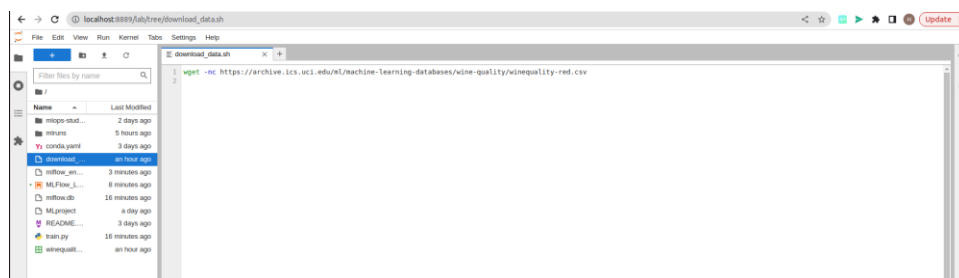
MLproject:



Conda.yaml:



Download_data.sh:



Winequality-red.csv:

		fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	ph	alcohol	quality
1	7.4	0.7	0	1.9	0.076	11	34	0.9978	3.51	0.56	9.4	5
2	7.8	0.88	0	2.6	0.090	25	67	0.9968	3.22	0.68	9.8	5
3	7.9	0.76	0.56	2.3	0.082	15	54	0.997	3.26	0.65	9.8	5
4	11.2	0.28	0.56	1.9	0.075	17	60	0.998	3.15	0.58	9.8	6
5	7.4	0.7	0	1.9	0.076	11	34	0.9978	3.51	0.56	9.4	5
6	7.4	0.66	0	1.8	0.075	13	40	0.9978	3.51	0.56	9.4	5
7	7.9	0.6	0.56	2.5	0.089	25	59	0.9954	3.3	0.62	9.4	5
8	7.3	0.65	0	1.2	0.065	15	21	0.9948	3.39	0.47	10	7
9	7.8	0.58	0.02	2	0.075	9	18	0.9968	3.36	0.57	9.5	7
10	7.5	0.5	0.36	6.1	0.071	17	102	0.9978	3.35	0.8	10.5	5
11	6.7	0.56	0.58	1.8	0.067	16	65	0.9959	3.28	0.54	9.2	5
12	7.5	0.5	0.36	6.1	0.071	17	102	0.9978	3.35	0.8	10.5	5
13	5.6	0.025	0	1.8	0.068	16	59	0.9943	3.58	0.52	9.9	5
14	7.8	0.61	0.29	1.6	0.114	9	29	0.9974	3.26	0.58	9.1	5
15	8.9	0.62	0.19	3.9	0.176	52	149	0.9986	3.18	0.98	9.2	5
16	8.9	0.62	0.19	3.9	0.17	51	148	0.9986	3.17	0.89	9.2	5
17	8.5	0.28	0.58	1.8	0.062	26	103	0.9969	3.3	0.75	10.5	7
18	8.1	0.56	0.28	1.7	0.388	16	58	0.9968	3.11	1.29	9.3	5
19	7.4	0.59	0.58	4.4	0.066	6	28	0.9974	3.38	0.5	9	4
20	7.9	0.32	0.51	1.8	0.342	17	56	0.9969	3.04	1.38	9.2	6
21	8.9	0.22	0.48	1.8	0.077	29	60	0.9968	3.39	0.53	9.4	6
22	7.6	0.39	0.31	2.3	0.082	23	71	0.9962	3.52	0.85	9.7	5
23	7.6	0.43	0.32	1.8	0.136	10	37	0.9968	3.17	0.82	9.5	5
24	8.5	0.49	0.11	2.3	0.064	9	67	0.9968	3.17	0.53	9.4	5
25	6.9	0.4	0.14	2.4	0.085	21	40	0.9968	3.43	0.63	9.7	6
26	6.3	0.39	0.18	1.4	0.08	11	23	0.9955	3.34	0.56	9.3	5
27	7.8	0.41	0.36	1.8	0.198	4	11	0.9962	3.28	0.59	9.5	5
28	7.9	0.43	0.21	1.8	0.139	10	37	0.9968	3.17	0.82	9.5	5
29	7.1	0.71	0	1.9	0.08	14	35	0.9972	3.47	0.55	9.4	5
30	7.8	0.60	0	2	0.062	8	18	0.9964	3.38	0.59	9.8	6
31	6.7	0.675	0.67	2.4	0.099	17	82	0.9958	3.35	0.94	10.1	5
32	6.9	0.885	0	2.3	0.135	12	37	0.9968	3.46	0.57	10.6	6
33	8.3	0.855	0.12	2.3	0.083	15	113	0.9968	3.17	0.86	9.8	5
34	6.9	0.805	0.12	10.7	0.073	40	85	0.9983	3.45	0.52	9.4	5

Logs:

mlflow 2.3.1 Experiments Models											
Experiments											
Experiment ID: 0 Artifact Location: home/muhammadrza/awaseem/Documents/data_engineering_bootcamp_2303/tasks/3_machine_learning_essentials/day_4_mlruns/mlruns/0											
Description Edit											
Table view Chart view metrics mean < 1 and params model = "tree" Sort: max_n Columns Refresh											
Time created: All time State: Active											
	Run Name	Created	Duration	Version	Models	Metrics	Parameters				
						accuracy	d	max_d	max_n	≡	n
<input type="checkbox"/>	tasteful-trout-831	33 minutes ago	5.7s	-	sklearn_Ra_16	0.494	1	2	2	1	
<input type="checkbox"/>	overjoyed-fox-297	36 minutes ago	5.7s	-	sklearn_Ra_17	0.528	1	2	2	1	
<input type="checkbox"/>	carefree-sponge-271	39 minutes ago	4.0s	-	sklearn_Ra_16	0.528	1	2	2	1	
<input type="checkbox"/>	shivering-bug-374	39 minutes ago	-	-	-	-	-	2	2	-	
<input type="checkbox"/>	brawny-bug-28	41 minutes ago	5.3s	-	sklearn_Ra_15	0.434	1	2	2	1	
<input type="checkbox"/>	invincible-smelt-530	1 hour ago	4.1s	-	sklearn_Ra_14	0.434	1	2	2	1	
<input type="checkbox"/>	angry-jay-7	5 hours ago	4.2s	-	sklearn_Ra_13	0.463	1	2	2	1	
<input type="checkbox"/>	industrious-mare-671	5 hours ago	4.0s	-	sklearn_Ra_12	0.469	1	2	2	1	
<input type="checkbox"/>	placid-swan-96	5 hours ago	4.3s	-	sklearn_Ra_11	0.488	1	2	2	1	
<input type="checkbox"/>	monumental-foal-134	5 hours ago	4.0s	-	sklearn_Ra_10	0.544	1	2	2	1	
<input type="checkbox"/>	unique-pug-448	5 hours ago	8.5s	-	sklearn_Ra_9	0.475	1	2	2	1	
<input type="checkbox"/>	rebellious-suk-345	5 hours ago	4.2s	-	sklearn_Ra_8	0.494	1	2	2	1	
<input type="checkbox"/>	languid-toad-764	5 hours ago	5.8s	-	sklearn_Ra_7	0.469	1	2	2	1	
<input type="checkbox"/>	redolent-squid-325	1 day ago	3.7s	-	sklearn_Ra_6	0.5	1	2	2	1	
<input type="checkbox"/>	sincere-bear-946	1 day ago	3.5s	-	sklearn_Ra_5	0.45	1	2	2	1	
<input type="checkbox"/>	wise-fowl-543	1 day ago	3.5s	-	sklearn_Ra_4	0.484	1	2	2	1	
<input type="checkbox"/>	salty-duck-666	1 day ago	5.0s	-	sklearn_Ra_3	0.481	1	2	2	1	
<input type="checkbox"/>	wistful-shrew-103	1 day ago	5.2s	-	sklearn_Ra_2	0.581	1	2	2	1	