Assignment 2 Implementation

VU Machine Learning 2022 WS

Group 37:
Branimir
Stefan
Rastko

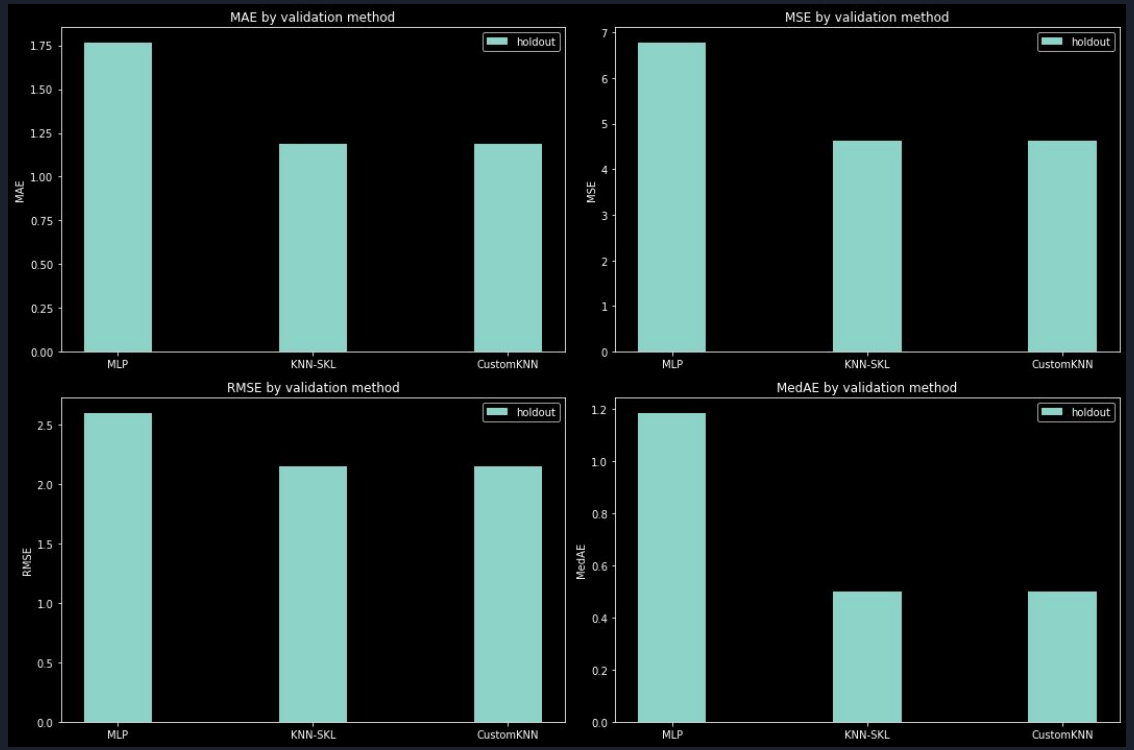
Regression Algorithm Implementation

- Algorithm Implemented: **KNN**
- Two distance metrics available : **‘euclidean’** (default) and **‘manhattan’**
- **Aggregation function** also customizable(i.e. how are the values of the k neighbors aggregated) : **mean** is the default
- Since the Model Implements sklearn’s *BaseEstimator* all hyperparameter optimization algorithms are compatible with the model object (in our code **GridSearchCV** was used)

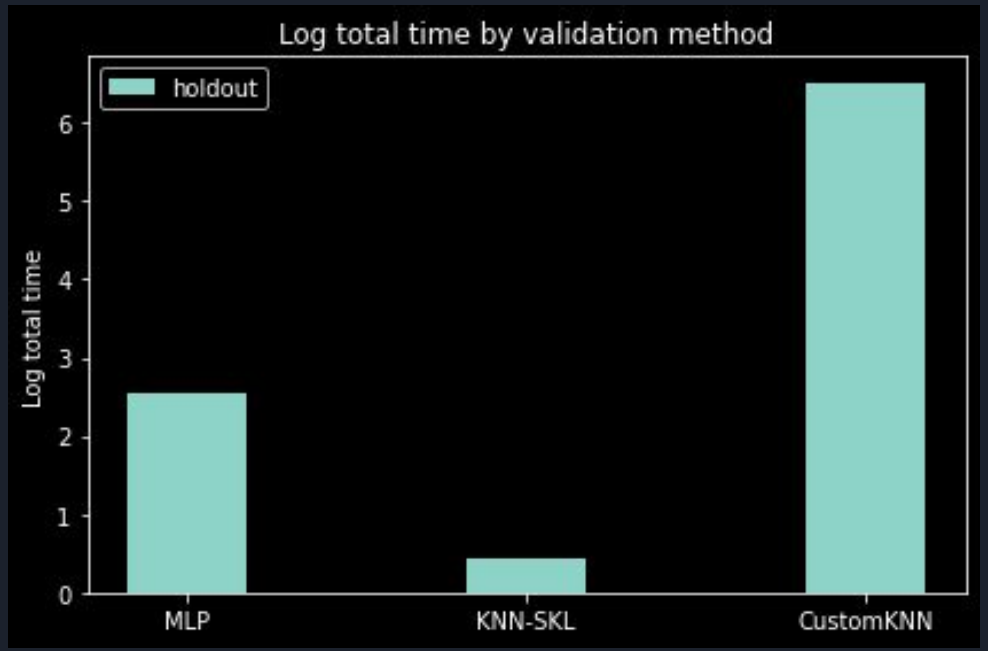
Regression Datasets

- Life Expectancy (WHO) (https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who)
  - Large(r) : ~2.9k observations, 22 variables (2 Categorical)
  - ~56% of observations contain at least one missing value
  - Preprocessing: Missing values imputed, numerical cols scaled and categorical one-hot encoded (resulted in 209 columns after preprocessing)
- Computer Hardware Dataset (https://archive.ics.uci.edu/ml/datasets/Computer+Hardware)
  - Small(er) : ~200 observations, 9 variables (1 Categorical)
  - Preprocessing: Categorical variable one-hot encoded and numerical ones were scaled

Life expectancy Dataset Performance Evaluation



Life expectancy dataset efficiency comparison



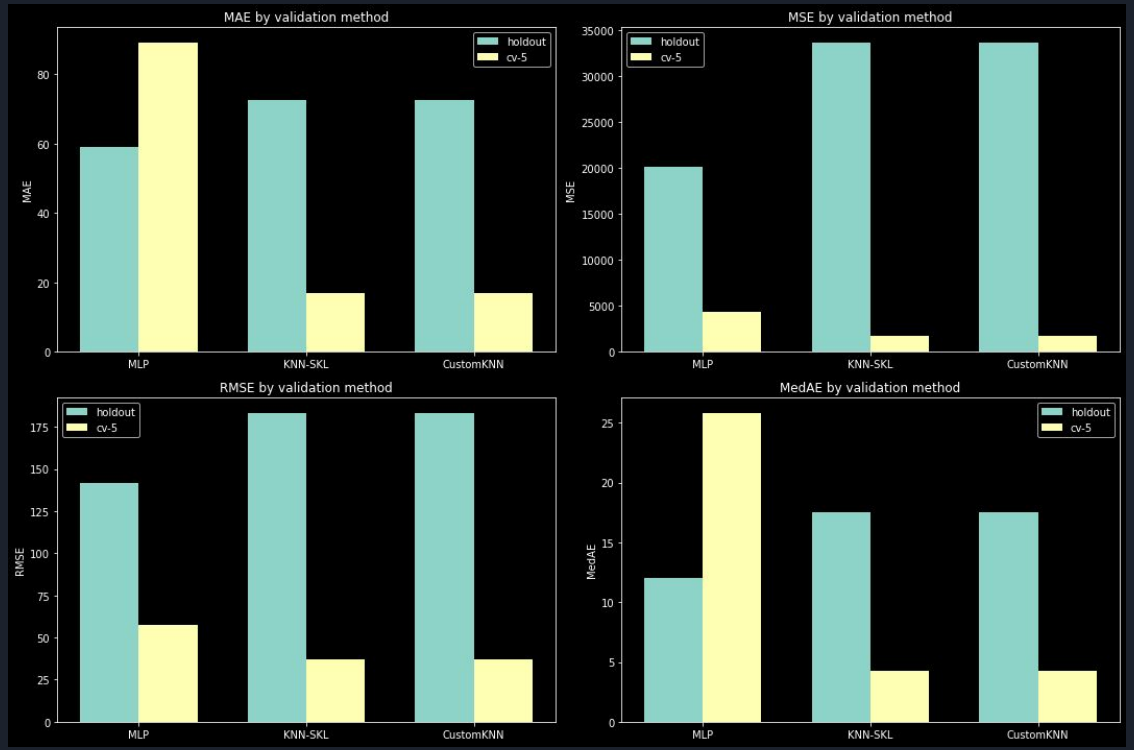
Total time predicting with the CustomKNN:
~11 minutes

Note: Log scale

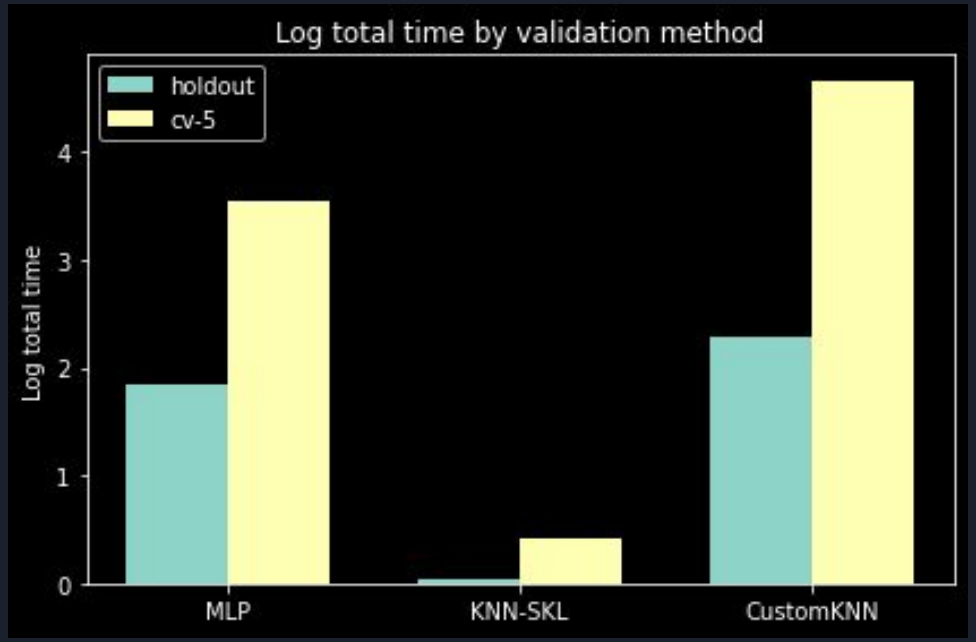
Life expectancy results in tabular form

algorithm	val	MAE	MSE	RMSE	MedAE	fit_time(sec)	predict_time(sec)	total_time(sec)
MLP	holdout	1.765	6.770	2.602	1.186	11.745	0.013	11.758
KNN-SKL	holdout	1.186	4.614	2.148	0.500	0.003	0.537	0.540
Custom KNN	holdout	1.186	4.614	2.148	0.500	0.000	676.840	676.840

Computer Hardware Dataset Performance Evaluation



Computer Hardware Dataset Efficiency Comparison



Note: Log scale

Computer Hardware results in tabular form

algorithm	val	MAE	MSE	RMSE	MedAE	fit_time(sec)	predict_time(sec)	total_time(sec)
MLP	holdout	59.038	20165.931	142.007	12.002	5.259977	0.013009	5.273
MLP	cv	89.060	4302.409	57.592	25.833	None	None	33.717
KNN-SKL	holdout	72.508	33619.992	183.358	17.500	0.011998	0.016	0.028
KNN-SKL	cv	17.021	1733.870	37.317	4.250	None	None	0.522
CustomKNN	holdout	72.508	33619.992	183.358	17.500	0.0	8.884806	8.885
CustomKNN	cv	17.021	1733.870	37.317	4.250	None	None	105.164

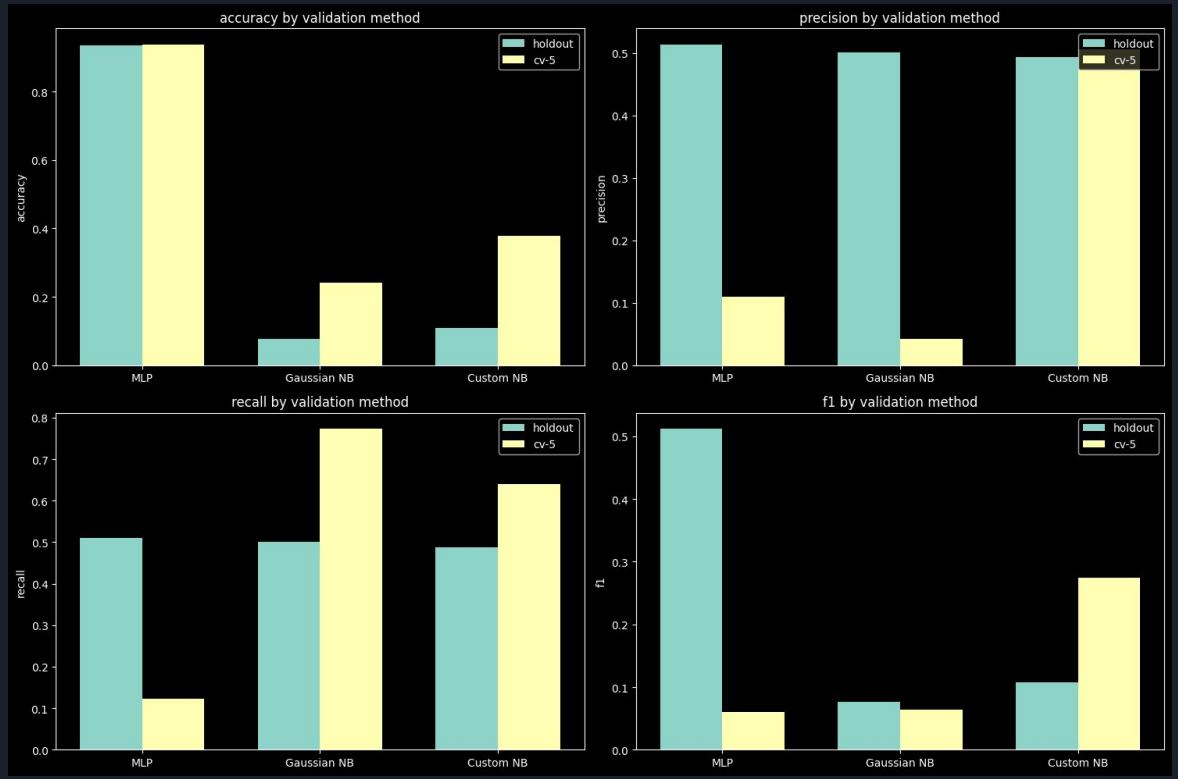
Naive Bayes Implementation

- Created as a class in python
- Has two functions fit and predict
- Calculates and stores the probabilities in a likelihood table
  - Nominal features - calculate the priori probability (frequencies of unique variable values)
  - Numeric features - calculate the mean and standard deviation
- Posterior probability in predict method for numeric features calculated using the normal distribution function

Classification Datasets

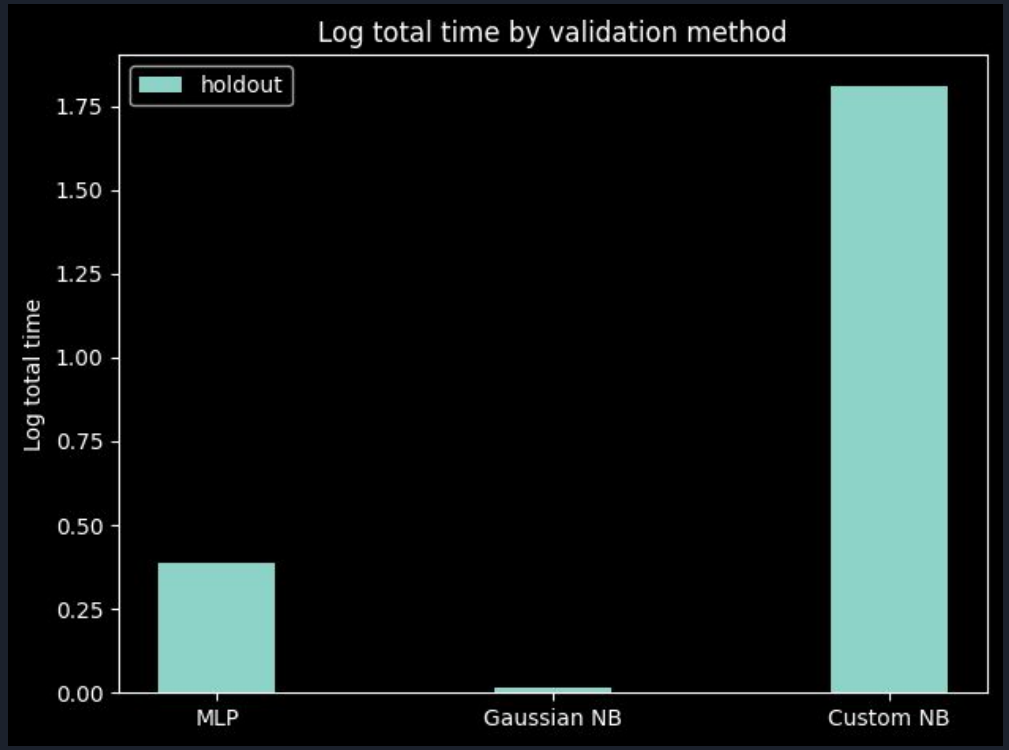
- Taiwanese Bankruptcy Prediction Dataset (https://archive.ics.uci.edu/ml/datasets/Taiwanese+Bankruptcy+Prediction)
  - Large(r) : ~6.8k observations, 95 variables (all numeric)
  - ~56% of observations contain at least one missing value
  - Preprocessing:
    - Drop instances with missing values,
    - Scaling on MLP, but no scaling on NB as algorithm invariant to feature scaling
- Eucalyptus (https://www.openml.org/search?type=data&status=active&id=188)
  - Small(er) : ~734 observations, 20 variables (6 categorical)
  - Dataset contains instances with missing feature values (641 after dropping)
  - Preprocessing: Drop instances, Scale for MLP, no scaling for NB

Taiwanese Bankrupcy Dataset Performance Evaluation



Note: No scaling applied

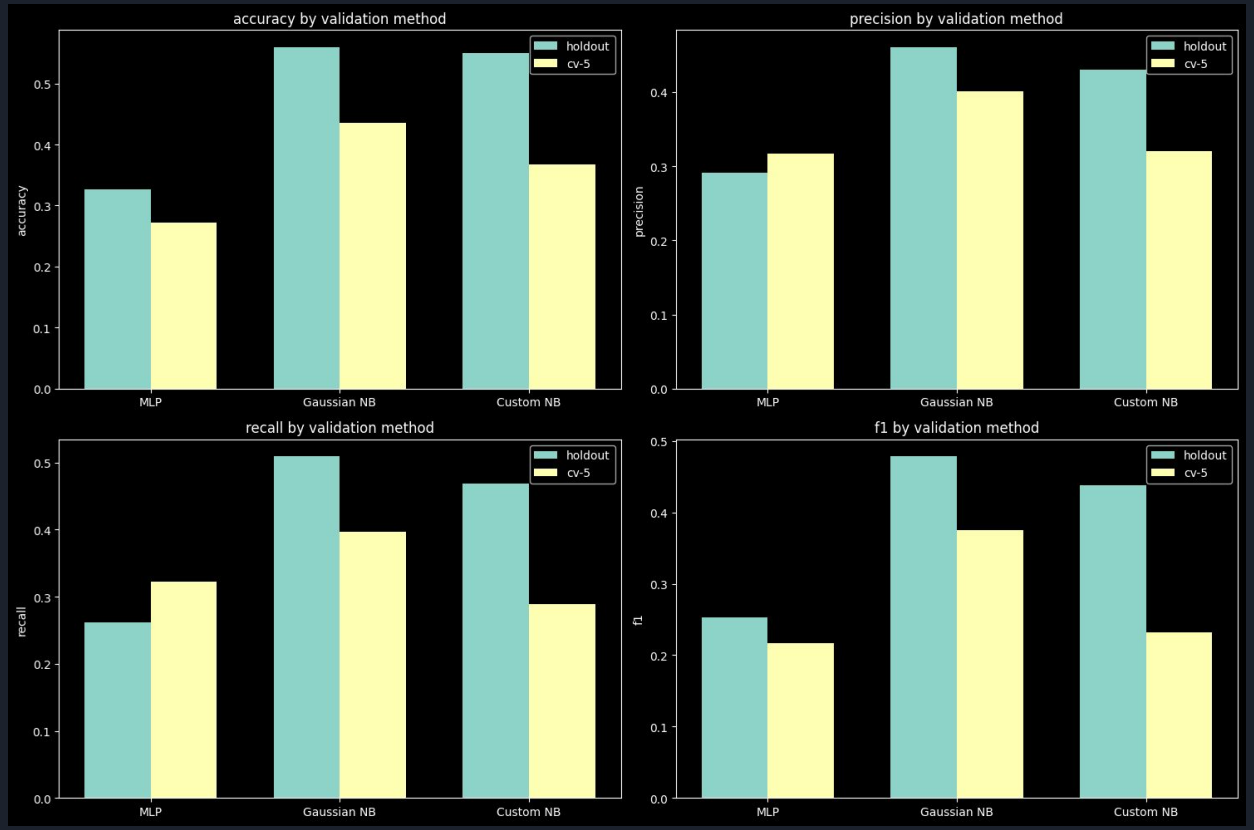
Taiwanese Bankrupcy dataset efficiency comparison



Taiwanese Bankrupcy results in tabular form

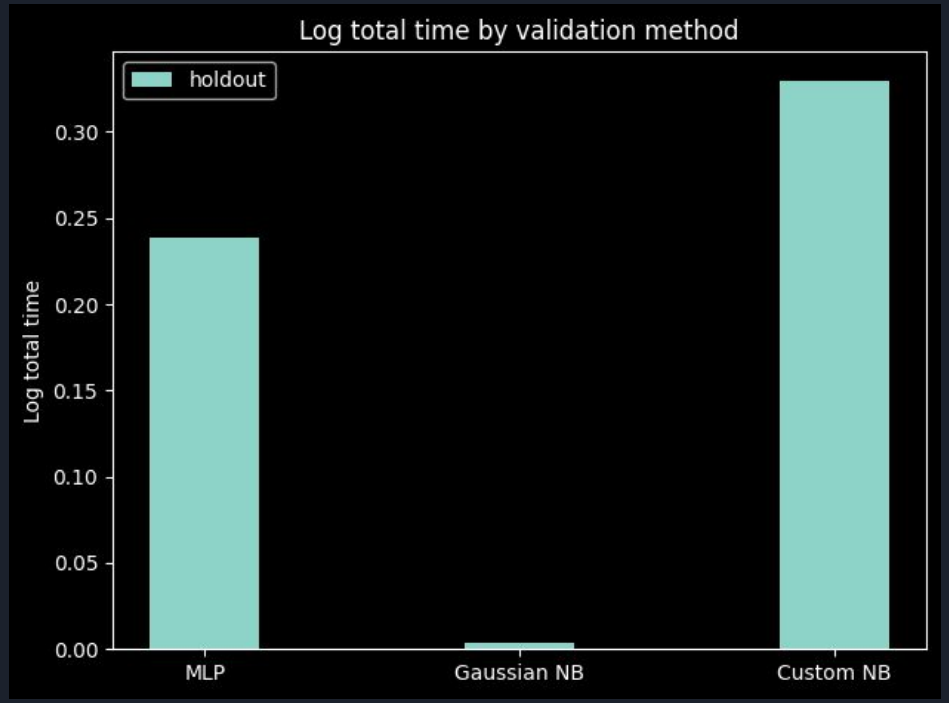
	algorithm	scaler	val	accuracy	precision	recall	f1	total_time (sec)
2	MLP	StandardScaler	Holdout	0.957478	0.689560	0.620837	0.646509	7.405393
0	MLP	StandardScaler	CV	0.949263	0.364046	0.227273	0.297071	45.930193
1	MLP	None	CV	0.937677	0.109656	0.122727	0.059660	8.129031
3	MLP	None	Holdout	0.935484	0.513609	0.510905	0.511871	0.471622
6	Custom NB	None	CV	0.379369	0.505512	0.638980	0.274208	18.394205
4	Gaussian NB	None	CV	0.241064	0.041623	0.772727	0.063983	0.095604
7	Custom NB	None	Holdout	0.108504	0.493558	0.487336	0.107070	5.118539
5	Gaussian NB	None	Holdout	0.076735	0.501476	0.501603	0.076724	0.015644

Eucalyptus Dataset Performance Evaluation



Note: No scaling applied

Eucalyptus dataset efficiency comparison



Eucalyptus results in tabular form

	algorithm	scaler	val	accuracy	precision	recall	f1	total_time (sec)
2	MLP	StandardScaler	Holdout	0.616580	0.611484	0.581157	0.585074	0.572839
5	Gaussian NB	None	Holdout	0.559585	0.460638	0.508969	0.478254	0.004998
7	Custom NB	None	Holdout	0.549223	0.430680	0.468894	0.438098	0.513519
0	MLP	StandardScaler	CV	0.454118	0.544141	0.444773	0.448370	3.159436
4	Gaussian NB	None	CV	0.435308	0.401011	0.396140	0.374909	0.025194
6	Custom NB	None	CV	0.366776	0.320501	0.289003	0.232282	2.060718
3	MLP	None	Holdout	0.326425	0.291379	0.262172	0.253141	0.254051
1	MLP	None	CV	0.271536	0.316765	0.322117	0.216686	1.328717