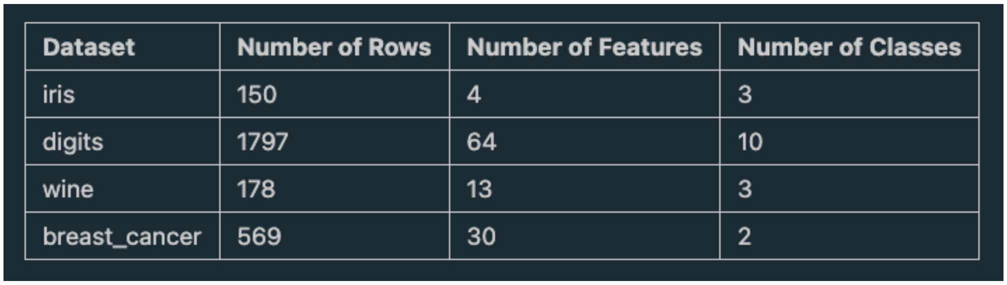
AutoML Project 1

Łukasz Zalewski 329523

Datasets

iris: Iris dataset containing measurements of iris flowers (multiclass classification)
digits: Digits dataset containing 8x8 images of handwritten digits (multiclass classification)
wine: Wine dataset containing attributes of different types of wine (multiclass classification)
breast_cancer: Breast cancer dataset containing features of tumor cells (binary classification)



Dataset	Number of Rows	Number of Features	Number of Classes
iris	150	4	3
digits	1797	64	10
wine	178	13	3
breast_cancer	569	30	2

Models

KNN: The prediction is based on the majority of the k-nearest neighbors.
RandomForest: Builds multiple decision trees and merges their predictions to improve accuracy and control overfitting.
XGBoost: An efficient and scalable implementation of gradient boosting framework.

Sampling Algorithms

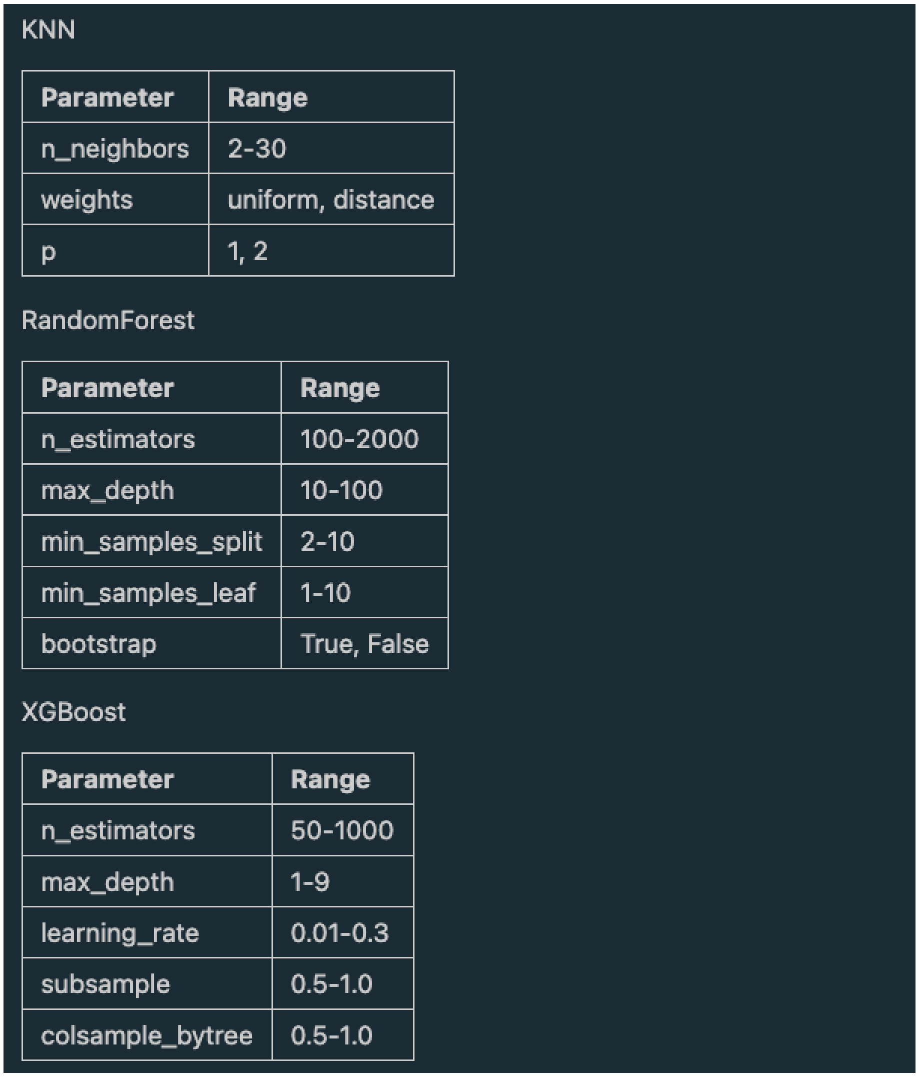
Random Search: Randomly sampling from predefined hyperparameter grid. The library used was RandomizedSearchCV from sklearn.
Bayesian Search: Tree-structured Parzen Estimator (TPE) method which is a type of bayesian optimisation. The library used was Hyperopt.

Methodology

There are: **4 datasets x 3 models x 50 trials**
The optimized metric is **accuracy**.
Each trial is evaluated on a **5-fold cross validation** split.

Hyperparameter Ranges

The following ranges were chosen based on the values applied in paper “Tunability: Importance of Hyperparameters of Machine Learning Algorithms”, as well as intuitions of the author.



KNN

Parameter	Range
n_neighbors	2-30
weights	uniform, distance
p	1, 2

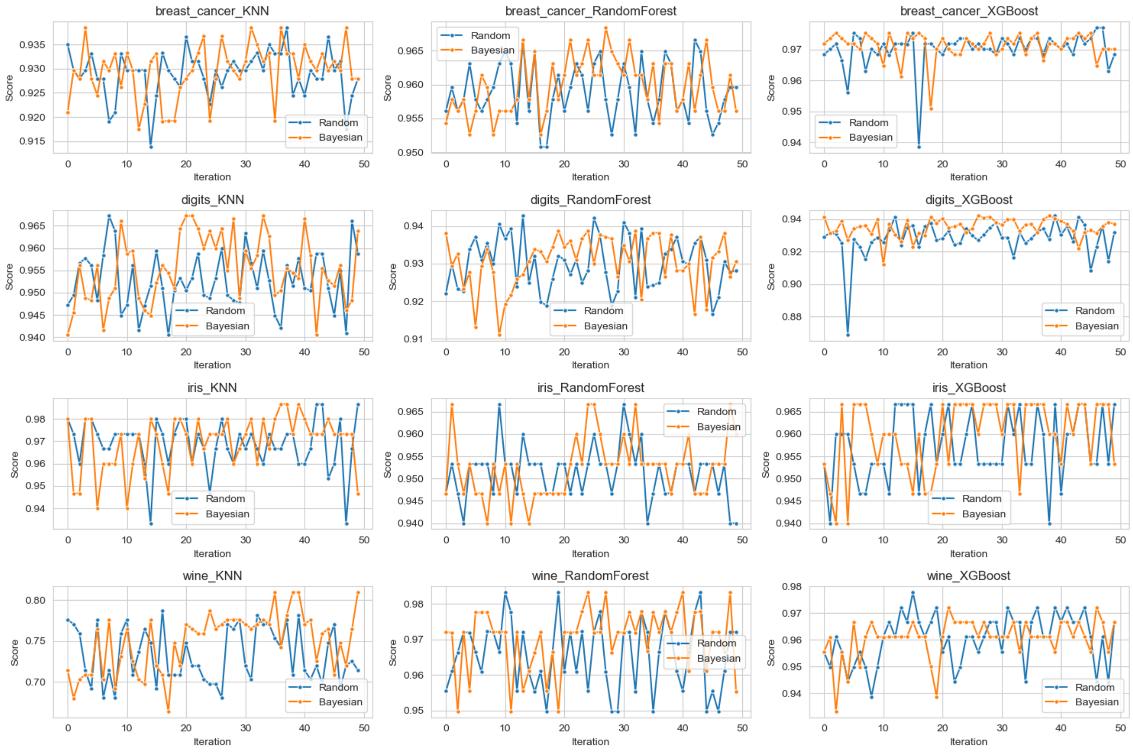
RandomForest

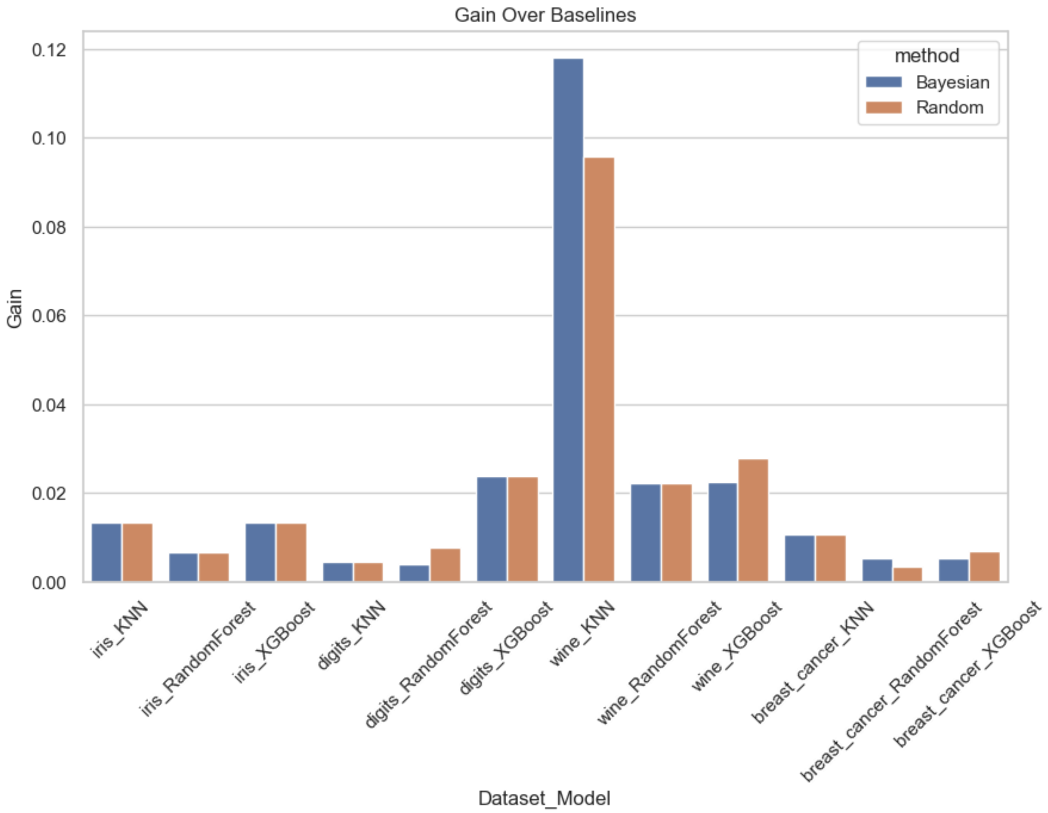
Parameter	Range
n_estimators	100-2000
max_depth	10-100
min_samples_split	2-10
min_samples_leaf	1-10
bootstrap	True, False

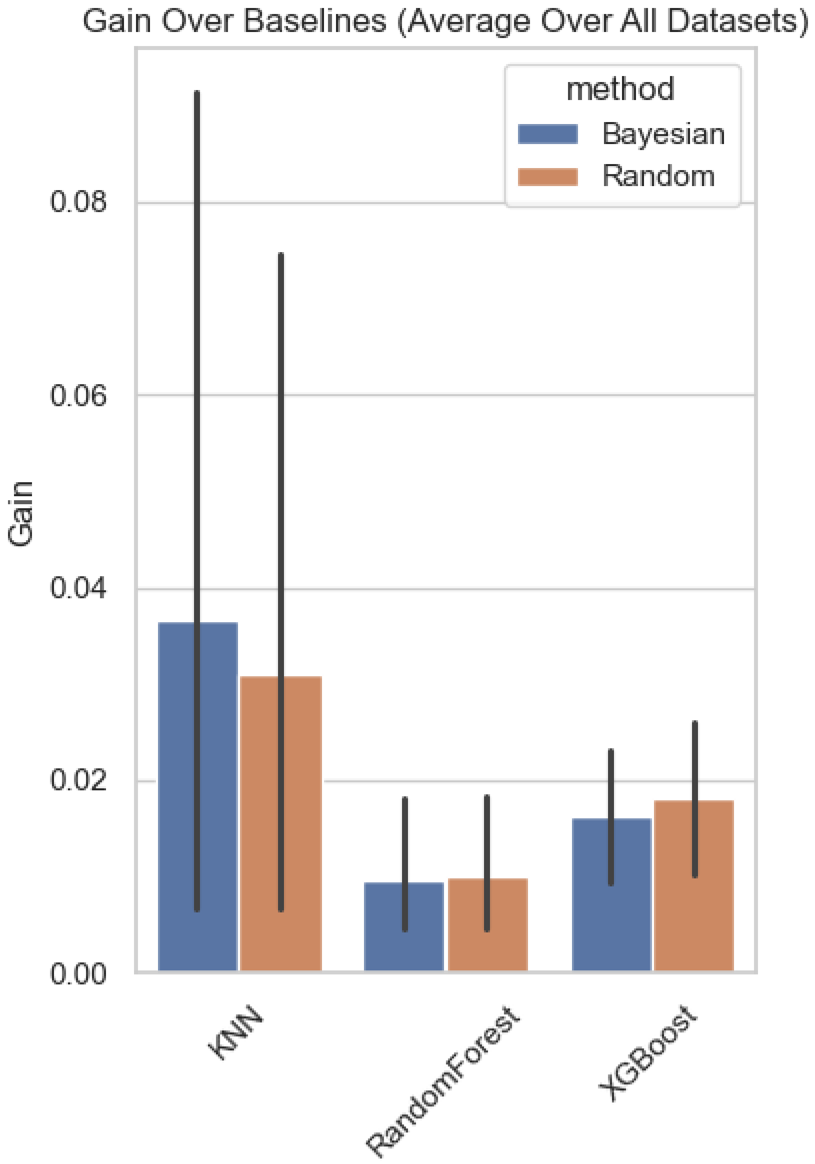
XGBoost

Parameter	Range
n_estimators	50-1000
max_depth	1-9
learning_rate	0.01-0.3
subsample	0.5-1.0
colsample_bytree	0.5-1.0

Results


breast_cancer_KNN
breast_cancer_RandomForest
breast_cancer_XGBoost
digits_KNN
digits_RandomForest
digits_XGBoost
iris_KNN
iris_RandomForest
iris_XGBoost
wine_KNN
wine_RandomForest
wine_XGBoost
Score
Iteration
0
10
20
30
40
50
Random
Bayesian


Gain Over Baselines
method
Bayesian
Random
Gain
0.12
0.10
0.08
0.06
0.04
0.02
0.00
iris_KNN
iris_RandomForest
iris_XGBoost
digits_KNN
digits_RandomForest
digits_XGBoost
wine_KNN
wine_RandomForest
wine_XGBoost
breast_cancer_KNN
breast_cancer_RandomForest
breast_cancer_XGBoost
Dataset_Model


Gain Over Baselines (Average Over All Datasets)
method
Bayesian
Random
Gain
0.00
0.02
0.04
0.06
0.08
KNN
RandomForest
XGBoost