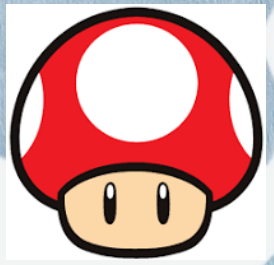


MUSHROOM EDIBILITY CLASSIFICATION

Yanfeiyun Wu (Kathy)

Data Science Institutes at Brown University

Dec. 6th 2023

HTTPS://GITHUB.COM/KATHYWU1201/MUSHROOM_EDIBILITY_CLASSIFICATION_ML_PROJECT

RECAP

- **Question:** If there exists a mushroom that has never seen before, will it be edible or it is poisonous?
- **Why important:** Help them to identify if the mushroom is safe to consume; Reduce the risk of death and intoxication due to poisonous mushrooms.
- **Problem type:** Classification
- **Source:** Kaggle, collected from Patrick Hardin's Mushrooms & Toadstools, and inspired by Jeff Schlimmer's Mushroom Data Set. kaggle
- **Preprocessing:** Most features are categorical data, number of features changes from 20 to 124
- **EDA:** 100% of urban grown mushroom is edible and 100% of path grown mushroom is poisonous (fig1)

Urban habitat

Path habitat

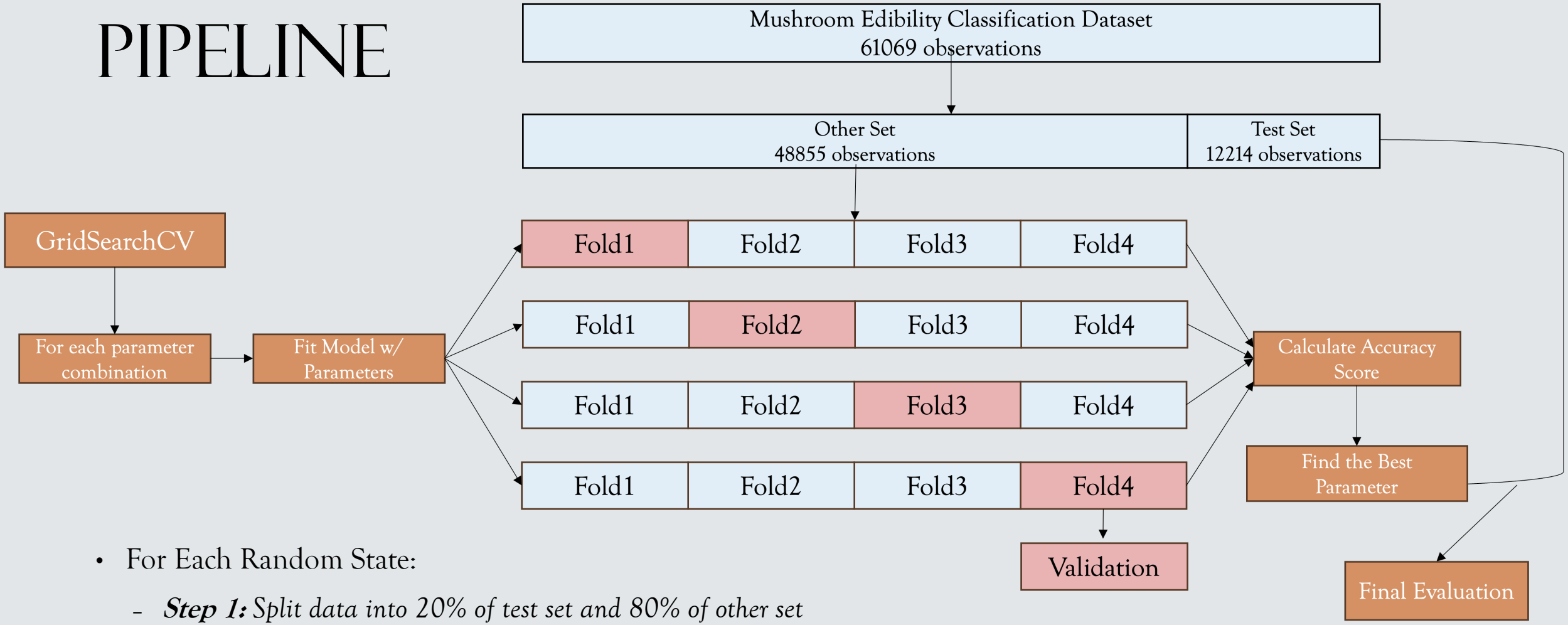
SPLITTING

- **Data Splitting:** use general train_test_split to set 20% of the dataset as test set and the rest 80% as other set, then use Kfold with 4 splits for Cross Validation.
- **Preprocessor:** Standard Scalar & onehot encoder
- **Chosen Models:**
  1. Logistic Regression
  2. Random Forest
  3. KNN
  4. XGBoost
  5. Support Vector Classification

MODEL & PARAMETERS

Model	Parameter(s)
Logistic Regression	C = 1/alpha: [1/0.001, 1/0.01, 1/0.1, 1/1.0]
Random Forest	max_depth: [5, 10, 20, 30] max_features: [0.25, 0.5, 0.75, 1.0] n_estimators: [20, 50, 100]
KNN	n_neighbors: [3,9,12,15,30,50,100]
XGBoost	max_depth: [3, 5, 7, 10] min_child_weight: [1, 3, 5] learning_rate: [0.1] lambda: [0.01, 0.1, 1] # reduce overfitting alpha: [0.01, 0.1, 1] # Used for high dimensionality
SVC	gamma: [1e-2, 1e-1, 1e1, 1e3] C: [1e-1, 1e0, 1e1, 1e2]

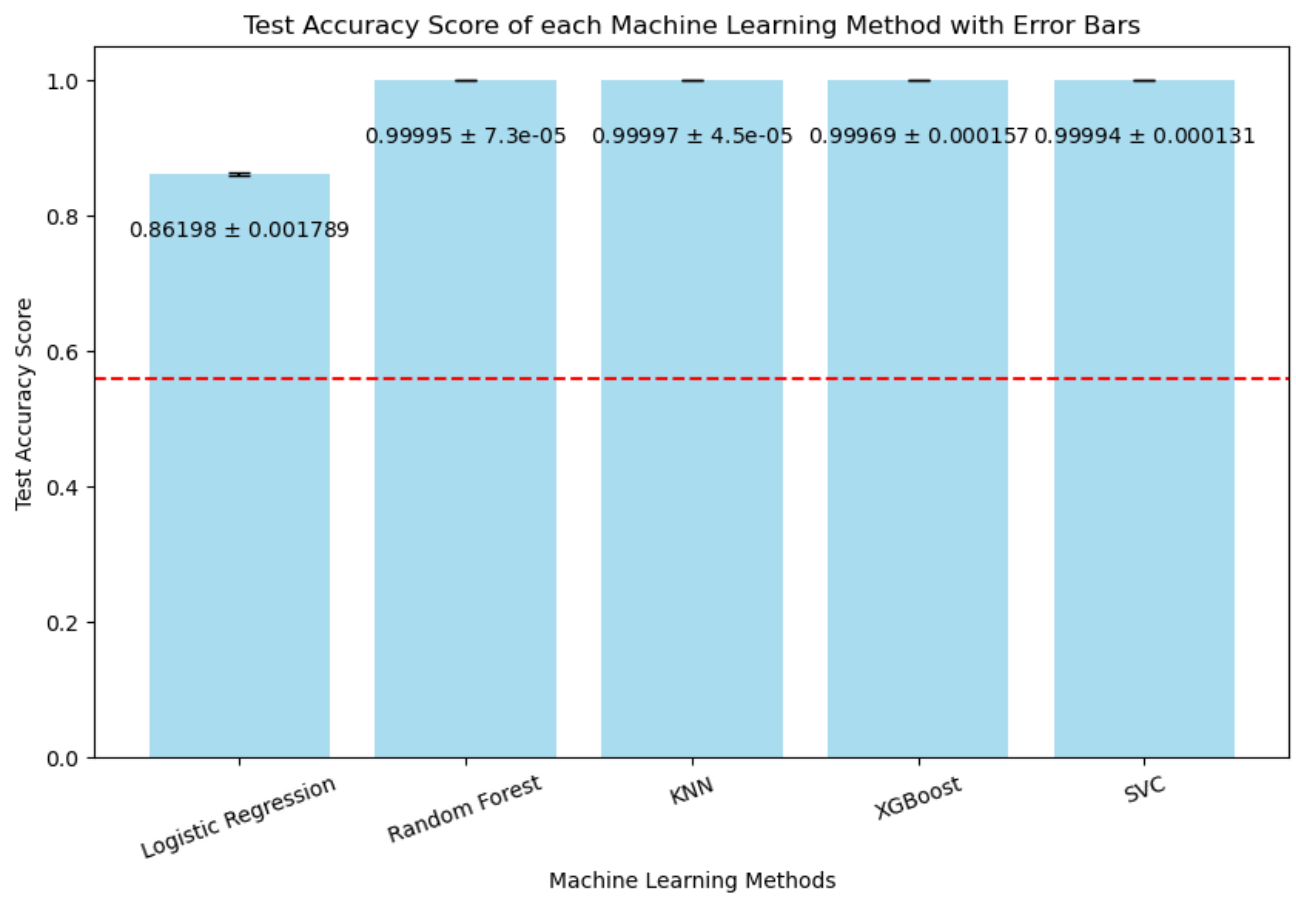
PIPELINE

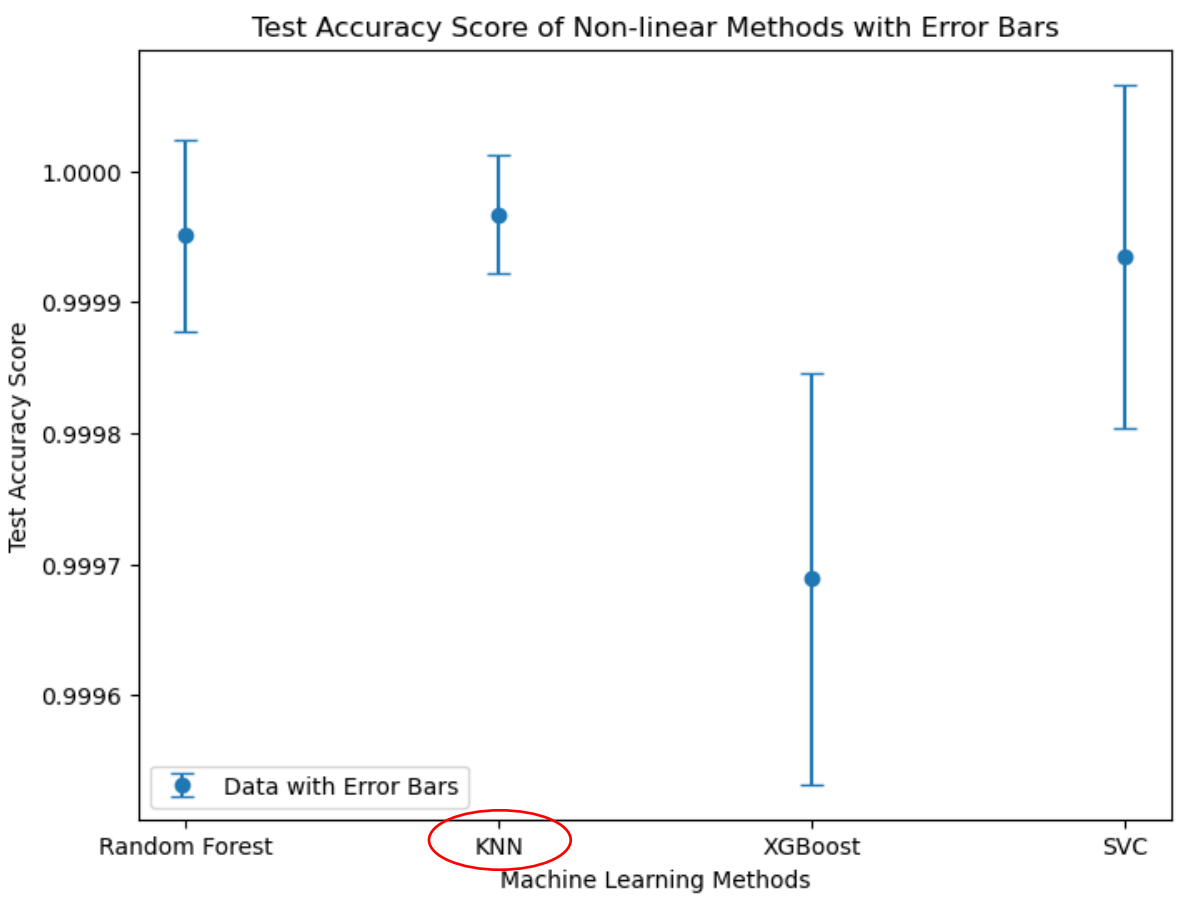


- For Each Random State:
  - ***Step 1:*** *Split data into 20% of test set and 80% of other set*
  - ***Step 2:*** *Use kfold with 4 splits to do cross validation on the other set*
  - ***Step 3:*** *Loop through each possible combination of the parameters, and do CV with 4 splits on each combination*
  - ***Step 4:*** *Find the Best Parameter(s) by evaluating the accuracy score*
  - ***Step 5:*** *Save the Best Parameter(s) and the related test score for each random state*

RESULTS

- Baseline Accuracy = 0.56001 (predicting all points as 1: poisonous)



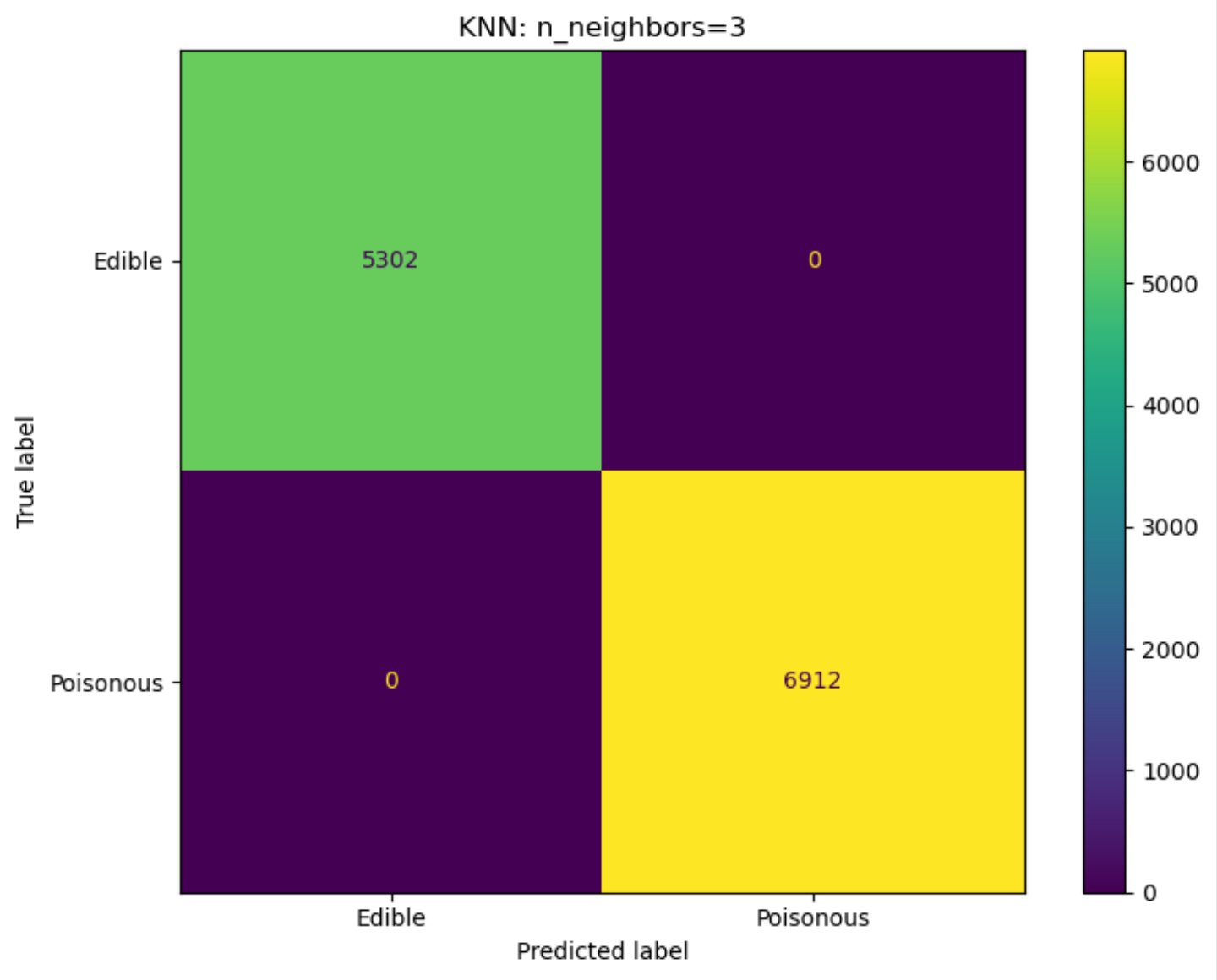


BEST PARAMETER(S) AND TEST SCORES

Model	Best Parameter(s)	Mean Test Scores	Std of Test Scores
Logistic Regression	{'logisticregression__C': [100.0,1000.0]}	0.861978	0.001789
Random Forest	{'randomforestclassifier__max_depth': 20, 'randomforestclassifier__max_features': 0.25, 'randomforestclassifier__n_estimators': [50,20]}	0.999951	0.000073
KNN	{'kneighborsclassifier__n_neighbors': 3}	0.999967	0.000045
XGBoost	{'xgbclassifier__max_depth': 10, 'xgbclassifier__min_child_weight': 1 'xgbclassifier__learning_rate': 0.1, 'xgbclassifier__lambda': 1, 'xgbclassifier__alpha': 1}	0.999689	0.000157
SVC	{'svc__C': 10.0, 'svc__gamma': 0.1}	0.999935	0.000131

CONFUSION MATRIX – (BEST MODEL)

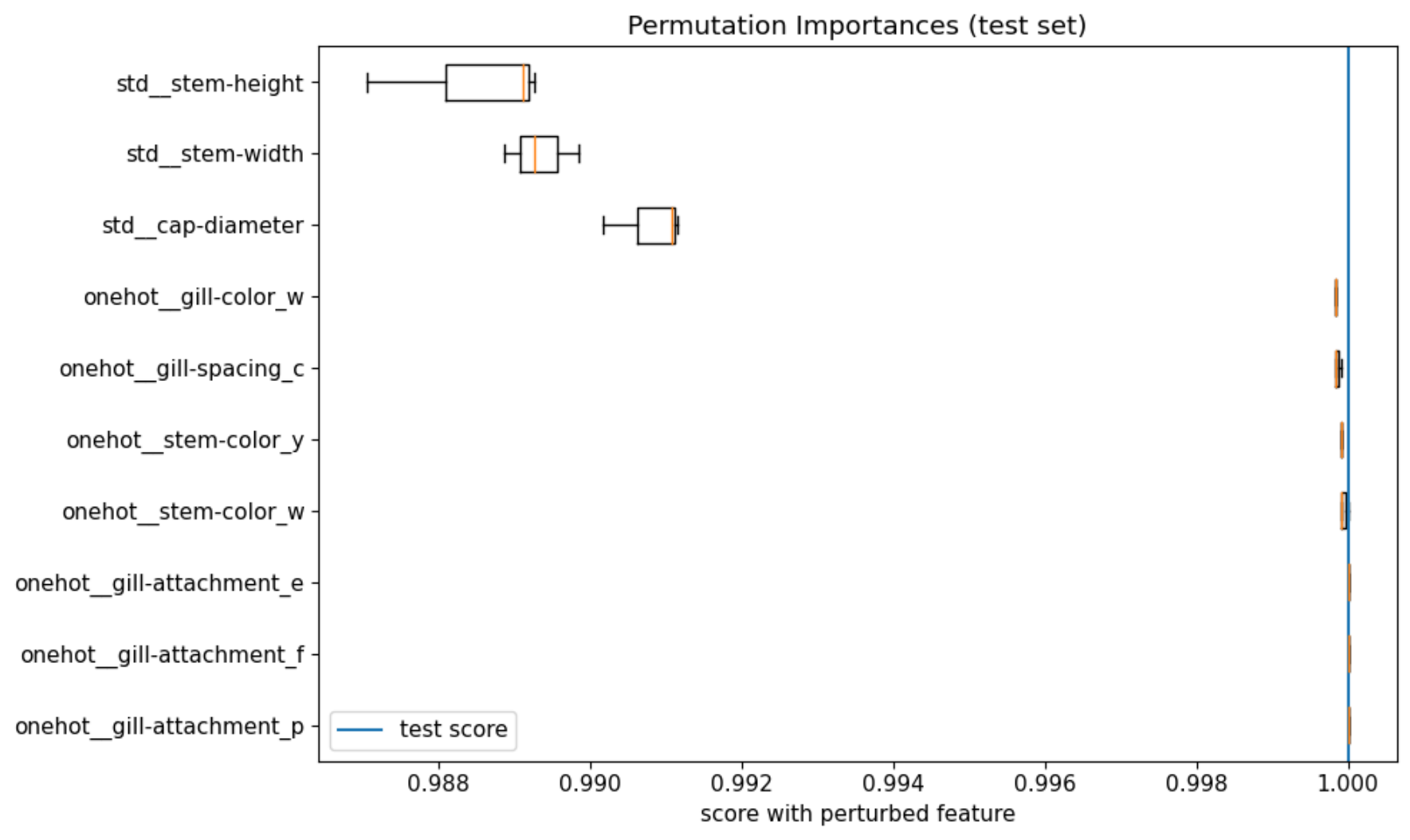
- Baseline Accuracy: 0.56001
- KNN: n_neighbors=3
- Test_accuracy = 1.0
  F1_score = 1.0



For safety, please do not pick and eat any mushroom you see in the nature ! !

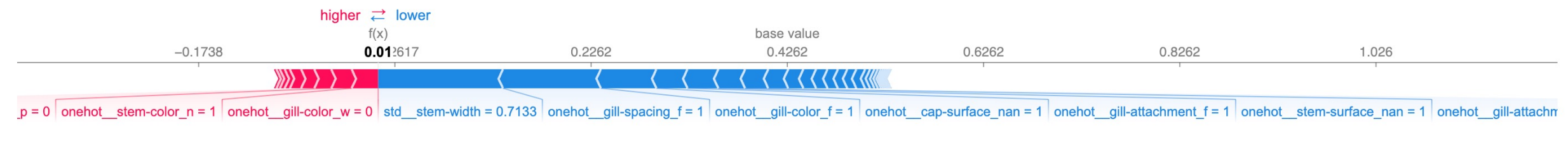
IMPORTANCE FEATURES (GLOBAL)

- Permutation Importance
  - *Stem-height*
  - *Stem-width*
  - *Cap-diameter*
  - *Gill-color: white*
  - *Gill-spacing: close*
  - *Stem-color: white, yellow*

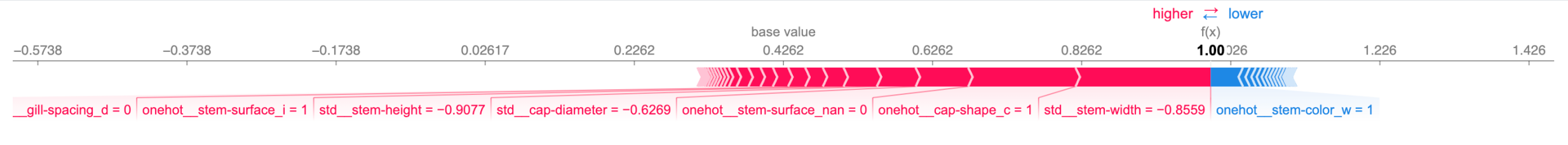


LOCAL FEATURE IMPORTANCE

Index=0



Index=99



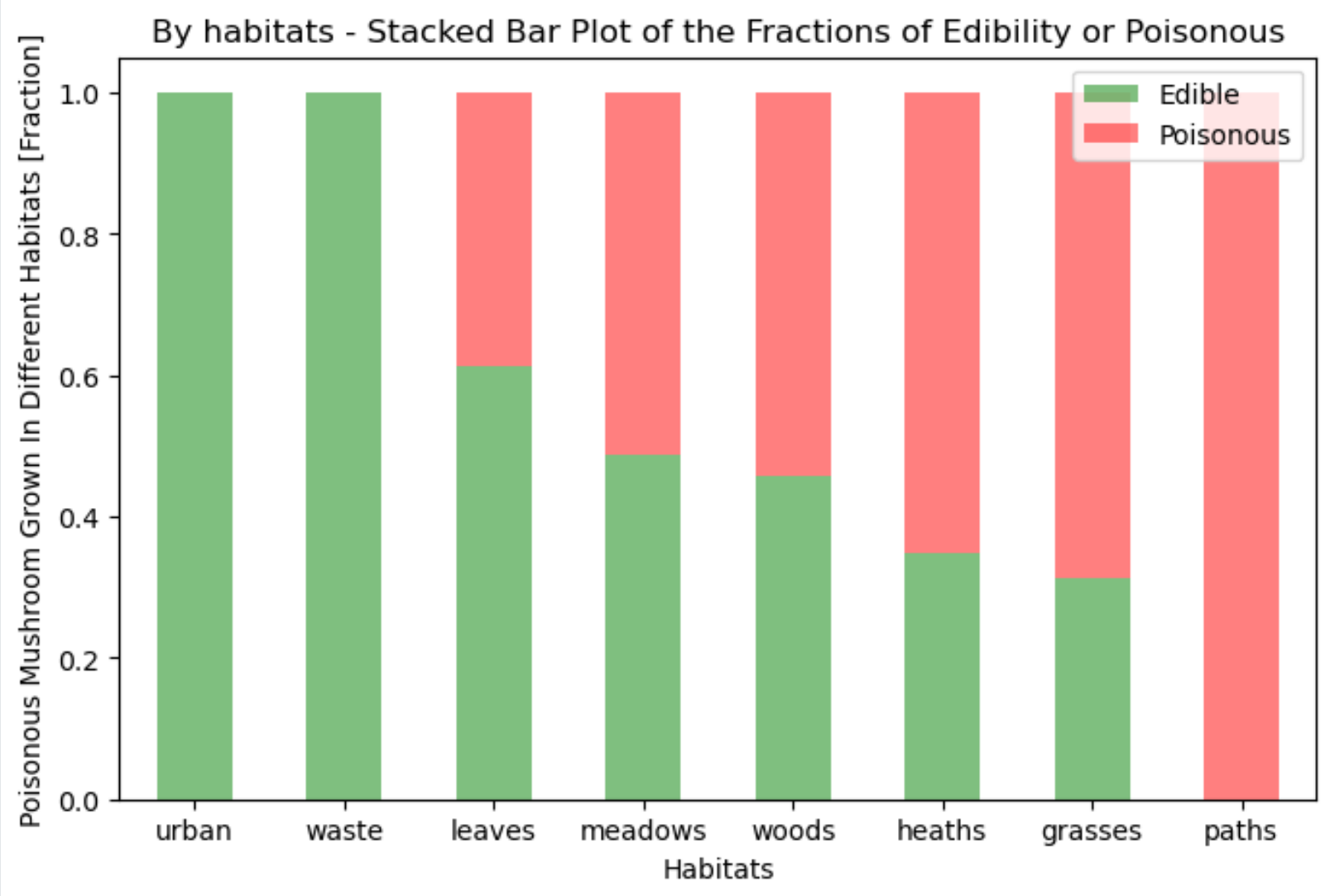
OUTLOOKS

- Feature Selection
  - *Reduce the number of features (124 features)*
  - *Less features, easier to interpret*

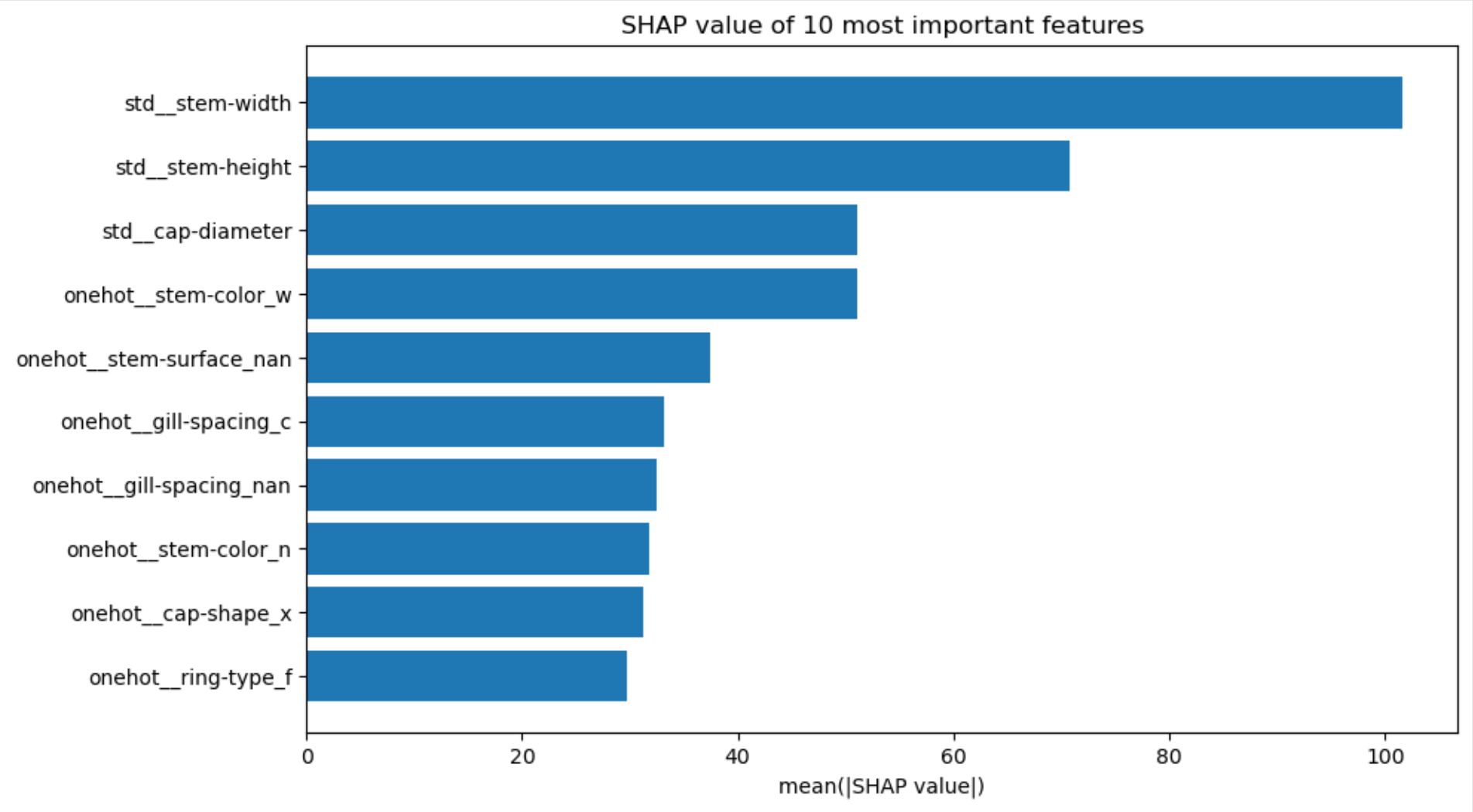
THANKS FOR LISTENING!

APPENDIX.

- fig1



SHAP


SHAP value of 10 most important features
std__stem-width
std__stem-height
std__cap-diameter
onehot__stem-color_w
onehot__stem-surface_nan
onehot__gill-spacing_c
onehot__gill-spacing_nan
onehot__stem-color_n
onehot__cap-shape_x
onehot__ring-type_f
0
20
40
60
80
100
mean(|SHAP value|)

RANDOM FOREST: FEATURE IMPORTANCE