

random_forest_to_debug

December 11, 2025

```
[33]: import pickle
import numpy as np
import pandas as pd
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import classification_report, confusion_matrix
from sklearn.model_selection import train_test_split
import seaborn as sns
import matplotlib as plt
```

```
[34]: train_path = "ift-3395-6390-kaggle-2-competition-fall-2025/train_data.pkl"
test_path  = "ift-3395-6390-kaggle-2-competition-fall-2025/test_data.pkl"
```

```
[35]: with open(train_path, "rb") as f:
    train_data = pickle.load(f)
with open(test_path, "rb") as f:
    test_data = pickle.load(f)

X = train_data["images"]
y = train_data["labels"].reshape(-1)
X_test = test_data["images"]
```

```
[36]: X_tr, X_val, y_tr, y_val = train_test_split(
    X, y, test_size=0.2, random_state=42, stratify=y
)
```

```
[37]: classes_to_augment = [1, 2, 3, 4]
X_extra, y_extra = [], []
for cls in classes_to_augment:
    idx = np.where(y_tr == cls)[0]
    X_extra.append(X_tr[idx])
    y_extra.append(y_tr[idx])

X_tr = np.concatenate([X_tr] + X_extra)
y_tr = np.concatenate([y_tr] + y_extra)
```

```
[38]: def extract_features(img_array, n_bins=8, n_circles=3):
    h, w, _ = img_array.shape
    cy, cx = h//2, w//2
```

```

Y, X_coord = np.ogrid[:h, :w]
radius = np.sqrt((X_coord - cx)**2 + (Y - cy)**2)
max_radius = radius.max()

features = []
r, g, b = img_array[:, :, 0], img_array[:, :, 1], img_array[:, :, 2]
features.extend([r.mean(), g.mean(), b.mean()])
lum = (0.299*r + 0.587*g + 0.114*b).mean()
features.append(lum)

for i in range(n_circles):
    mask = (radius >= i*max_radius/n_circles) & (radius < (i+1)*max_radius/
↪n_circles)
    for ch in [r, g, b]:
        vals = ch[mask]
        if len(vals) == 0:
            vals = np.array([0])
        features.extend([vals.mean(), vals.var(), vals.min(), vals.max()])
        hist, _ = np.histogram(vals, bins=n_bins, range=(0,255))
        features.extend(hist / (vals.size if vals.size>0 else 1))
return np.array(features)

```

```

[39]: X_tr_feat = np.array([extract_features(img) for img in X_tr])
X_val_feat = np.array([extract_features(img) for img in X_val])
X_test_feat = np.array([extract_features(img) for img in X_test])

```

```

[45]: from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import GridSearchCV
from sklearn.metrics import classification_report

# Modèle de base
clf = RandomForestClassifier(
    random_state=42,
    class_weight='balanced',
    n_jobs=-1
)

# Grille d'hyperparamètres
param_grid = {
    'n_estimators': [100, 200, 500],
    'max_depth': [None, 10, 20, 30],
    'min_samples_split': [2, 5, 10],
    'min_samples_leaf': [1, 2, 4],
    'max_features': ['sqrt', 'log2']
}

# Configuration de GridSearchCV

```

```

grid_search = GridSearchCV(
    estimator=clf,
    param_grid=param_grid,
    cv=5,
    scoring='recall_macro',
    n_jobs=-1,
    verbose=2
)

# Entraînement
grid_search.fit(X_tr_feat, y_tr)

# Résultats
print("Best parameters found:", grid_search.best_params_)
print("Best CV recall_macro:", grid_search.best_score_)

# Prédiction sur validation
pred_val = grid_search.predict(X_val)

# Rapport de performance
print(classification_report(y_val, pred_val))

```

```

Fitting 5 folds for each of 216 candidates, totalling 1080 fits
[CV] END max_depth=None, max_features=sqrt, min_samples_leaf=1,
min_samples_split=2, n_estimators=100; total time= 0.2s
[CV] END max_depth=None, max_features=sqrt, min_samples_leaf=1,
min_samples_split=2, n_estimators=100; total time= 0.3s
[CV] END max_depth=None, max_features=sqrt, min_samples_leaf=1,
min_samples_split=2, n_estimators=100; total time= 0.3s
[CV] END max_depth=None, max_features=sqrt, min_samples_leaf=1,
min_samples_split=2, n_estimators=100; total time= 0.3s
[CV] END max_depth=None, max_features=sqrt, min_samples_leaf=1,
min_samples_split=2, n_estimators=100; total time= 0.4s
[CV] END max_depth=None, max_features=sqrt, min_samples_leaf=1,
min_samples_split=2, n_estimators=200; total time= 0.6s
[CV] END max_depth=None, max_features=sqrt, min_samples_leaf=1,
min_samples_split=2, n_estimators=200; total time= 0.6s
[CV] END max_depth=None, max_features=sqrt, min_samples_leaf=1,
min_samples_split=2, n_estimators=200; total time= 0.6s
[CV] END max_depth=None, max_features=sqrt, min_samples_leaf=1,
min_samples_split=2, n_estimators=200; total time= 0.6s
[CV] END max_depth=None, max_features=sqrt, min_samples_leaf=1,
min_samples_split=2, n_estimators=200; total time= 0.6s
[CV] END max_depth=None, max_features=sqrt, min_samples_leaf=1,
min_samples_split=5, n_estimators=100; total time= 0.3s
[CV] END max_depth=None, max_features=sqrt, min_samples_leaf=1,
min_samples_split=5, n_estimators=100; total time= 0.3s
[CV] END max_depth=None, max_features=sqrt, min_samples_leaf=1,

```

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

```

[CV] END max_depth=30, max_features=log2, min_samples_leaf=4,
min_samples_split=10, n_estimators=200; total time= 0.7s
[CV] END max_depth=30, max_features=log2, min_samples_leaf=4,
min_samples_split=5, n_estimators=500; total time= 0.9s
[CV] END max_depth=30, max_features=log2, min_samples_leaf=4,
min_samples_split=10, n_estimators=200; total time= 0.6s
[CV] END max_depth=30, max_features=log2, min_samples_leaf=4,
min_samples_split=5, n_estimators=500; total time= 1.0s
[CV] END max_depth=30, max_features=log2, min_samples_leaf=4,
min_samples_split=5, n_estimators=500; total time= 1.0s
[CV] END max_depth=30, max_features=log2, min_samples_leaf=4,
min_samples_split=5, n_estimators=500; total time= 1.1s
[CV] END max_depth=30, max_features=log2, min_samples_leaf=4,
min_samples_split=5, n_estimators=500; total time= 1.1s
[CV] END max_depth=30, max_features=log2, min_samples_leaf=4,
min_samples_split=10, n_estimators=500; total time= 0.9s
[CV] END max_depth=30, max_features=log2, min_samples_leaf=4,
min_samples_split=10, n_estimators=500; total time= 0.8s
[CV] END max_depth=30, max_features=log2, min_samples_leaf=4,
min_samples_split=10, n_estimators=500; total time= 0.7s
[CV] END max_depth=30, max_features=log2, min_samples_leaf=4,
min_samples_split=10, n_estimators=500; total time= 0.8s
[CV] END max_depth=30, max_features=log2, min_samples_leaf=4,
min_samples_split=10, n_estimators=500; total time= 0.7s
Best parameters found: {'max_depth': 20, 'max_features': 'log2',
'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 100}
Best CV recall_macro: 0.9547585747585747

```

```

-----
ValueError                                Traceback (most recent call last)
Cell In[45], line 39
    36 print("Best CV recall_macro:", grid_search.best_score_)
    38 # Prédiction sur validation
--> 39 pred_val = grid_search.predict(X_val)
    41 # Rapport de performance
    42 print(classification_report(y_val, pred_val))

File ~/Documents/UDEM/A25/IFT3395/kaggle2/kaggle2/lib/python3.13/site-packages/
sklearn/model_selection/_search.py:598, in BaseSearchCV.predict(self, X)
    580 """Call predict on the estimator with the best found parameters.
    581
    582 Only available if ``refit=True`` and the underlying estimator supports
    (...) 595 the best found parameters.
    596 """
    597 check_is_fitted(self)
--> 598 return self.best_estimator_.predict(X)

```



```

File ~/Documents/UDEM/A25/IFT3395/kaggle2/kaggle2/lib/python3.13/site-packages/
↳sklearn/ensemble/_forest.py:903, in ForestClassifier.predict(self, X)
    882 def predict(self, X):
    883     """
    884     Predict class for X.
    885
    (...) 901     The predicted classes.
    902     """
--> 903     proba = self.predict_proba(X)
    905     if self.n_outputs_ == 1:
    906         return self.classes_.take(np.argmax(proba, axis=1), axis=0)

```

```

File ~/Documents/UDEM/A25/IFT3395/kaggle2/kaggle2/lib/python3.13/site-packages/
↳sklearn/ensemble/_forest.py:945, in ForestClassifier.predict_proba(self, X)
    943 check_is_fitted(self)
    944 # Check data
--> 945 X = self._validate_X_predict(X)
    947 # Assign chunk of trees to jobs
    948 n_jobs, _, _ = _partition_estimators(self.n_estimators, self.n_jobs)

```

```

File ~/Documents/UDEM/A25/IFT3395/kaggle2/kaggle2/lib/python3.13/site-packages/
↳sklearn/ensemble/_forest.py:637, in BaseForest._validate_X_predict(self, X)
    634 else:
    635     ensure_all_finite = True
--> 637 X = validate_data(
    638     self,
    639     X,
    640     dtype=DTYPE,
    641     accept_sparse=,
    642     reset=False,
    643     ensure_all_finite=ensure_all_finite,
    644 )
    645 if issparse(X) and (X.indices.dtype != np.intc or X.indptr.dtype != np.
↳intc):
    646     raise ValueError("No support for np.int64 index based sparse_
↳matrices")

```

```

File ~/Documents/UDEM/A25/IFT3395/kaggle2/kaggle2/lib/python3.13/site-packages/
↳sklearn/utils/validation.py:2954, in validate_data(estimator, X, y, reset,
↳validate_separately, skip_check_array, **check_params)
    2952     out = X, y
    2953 elif not no_val_X and no_val_y:
-> 2954     out = check_array(X, input_name=, **check_params)
    2955 elif no_val_X and not no_val_y:
    2956     out = _check_y(y, **check_params)

```

```

File ~/Documents/UDEM/A25/IFT3395/kaggle2/kaggle2/lib/python3.13/site-packages/
↳sklearn/utils/validation.py:1099, in check_array(array, accept_sparse,
↳accept_large_sparse, dtype, order, copy, force_writeable, force_all_finite,
↳ensure_all_finite, ensure_non_negative, ensure_2d, allow_nd,
↳ensure_min_samples, ensure_min_features, estimator, input_name)
    1094     raise ValueError(
    1095         "dtype='numeric' is not compatible with arrays of bytes/strings "
    1096         "Convert your data to numeric values explicitly instead."
    1097     )
    1098 if not allow_nd and array.ndim >= 3:
-> 1099     raise ValueError(
    1100         f"Found array with dim {array.ndim},"
    1101         f" while dim <= 2 is required{context}."
    1102     )
    1104 if ensure_all_finite:
    1105     _assert_all_finite(
    1106         array,
    1107         input_name=input_name,
    1108         estimator_name=estimator_name,
    1109         allow_nan=ensure_all_finite == "allow-nan",
    1110     )

ValueError: Found array with dim 4, while dim <= 2 is required by
↳RandomForestClassifier.

```

```

[ ]: print("=== Classification report sur la validation ===")
print(classification_report(y_val, pred_val, digits=3))

```

```

cm = confusion_matrix(y_val, pred_val)
plt.figure(figsize=(8, 6))
sns.heatmap(cm, annot=True, fmt='d', cmap='Blues')
plt.xlabel('Predicted')
plt.ylabel('True')
plt.title('Confusion Matrix')
plt.show()

```

```

-----
NotFittedError                                Traceback (most recent call last)

```

```

Cell In[41], line 1
----> 1 pred_val = clf.predict(X_val)
      2 print("=== Classification report sur la validation ===")
      3 print(classification_report(y_val, pred_val, digits=3))

```

```

File ~/Documents/UDEM/A25/IFT3395/kaggle2/kaggle2/lib/python3.13/site-packages/
↳sklearn/ensemble/_forest.py:903, in ForestClassifier.predict(self, X)

```

```

882 def predict(self, X):
883     """
884     Predict class for X.
885
886     (...) 901         The predicted classes.
887     902     """
--> 903     proba = self.predict_proba(X)
905     if self.n_outputs_ == 1:
906         return self.classes_.take(np.argmax(proba, axis=1), axis=0)

```

File ~/Documents/UDEM/A25/IFT3395/kaggle2/kaggle2/lib/python3.13/site-packages/sklearn/ensemble/_forest.py:943, in ForestClassifier.predict_proba(self, X)

```

921 def predict_proba(self, X):
922     """
923     Predict class probabilities for X.
924
925     (...) 941         classes corresponds to that in the attribute :term:
926     <- `classes_`.
927     942     """
--> 943     check_is_fitted(self)
944     # Check data
945     X = self._validate_X_predict(X)

```

File ~/Documents/UDEM/A25/IFT3395/kaggle2/kaggle2/lib/python3.13/site-packages/sklearn/utils/validation.py:1754, in check_is_fitted(estimator, attributes, msg, all_or_any)

```

1751     return
1753 if not _is_fitted(estimator, attributes, all_or_any):
-> 1754     raise NotFittedError(msg % {"name": type(estimator).__name__})

```

NotFittedError: This RandomForestClassifier instance is not fitted yet. Call
<- 'fit' with appropriate arguments before using this estimator.

```

[ ]: unique, counts = np.unique(pred_val, return_counts=True)
print("\nNombre de prédictions par classe sur validation :")
for u, c in zip(unique, counts):
    print(f"Classe {u}: {c} images")

```

```

[ ]: pred_test = clf.predict(X_test_feat)

```

```

[ ]: unique_test, counts_test = np.unique(pred_test, return_counts=True)
print("\nNombre de prédictions par classe sur test :")
for u, c in zip(unique_test, counts_test):
    print(f"Classe {u}: {c} images")

```

```

[ ]: results_test = [{"ID": idx, "Label": int(label)} for idx, label in
<- enumerate(pred_test, start=1)]

```

```
df_test = pd.DataFrame(results_test)
df_test.to_csv("manual_classification_no_leak_test.csv", index=False)
print("\nCSV test créé avec succès !")
```