penguins-post-2

February 4, 2024

1 Classification

The Palmer Penguins dataset is a common resource for data exploration and demonstration of data analysis techniques. It was brought into the limelight by Dr. Kristen Gorman and the Palmer Station, Antarctica LTER, which is a member of the Long Term Ecological Research Network.

The dataset includes data for 344 penguins from three different species found on three islands in the Palmer Archipelago, Antarctica. The measured attributes in the dataset include:

1. **Species**: The species of the penguin, which can be Adelie, Gentoo, or Chinstrap.
2. **Island**: The island in the Palmer Archipelago, Antarctica, where the penguin observation was made. The options are Torgersen, Biscoe, or Dream.
3. **Culmen Length (mm)**: The length of the penguin's culmen (bill).
4. **Culmen Depth (mm)**: The depth of the penguin's culmen (bill).
5. **Flipper Length (mm)**: The length of the penguin's flipper.
6. **Body Mass (g)**: The body mass of the penguin.
7. **Sex**: The sex of the penguin.

The Palmer Penguins dataset is excellent for practicing data cleaning, exploration, and visualization.

You can find more information about the dataset, including a more detailed explanation of the variables, in this repository: allisonhorst/palmerpenguins.

For more in-depth studies or referencing, you might also consider checking out the publications from Palmer Station LTER: pal.lternet.edu/bibliography.

[1]:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.pipeline import Pipeline
from sklearn.impute import SimpleImputer
from sklearn.preprocessing import StandardScaler, OneHotEncoder
from sklearn.compose import ColumnTransformer
from sklearn.linear_model import SGDClassifier
from sklearn.model_selection import cross_val_score
from sklearn.model_selection import cross_val_predict
from sklearn.metrics import confusion_matrix
from sklearn.metrics import precision_score, recall_score
from sklearn.metrics import f1_score
from sklearn.metrics import precision_recall_curve
from sklearn.metrics import roc_curve
from sklearn.metrics import roc_auc_score
from sklearn.metrics import ConfusionMatrixDisplay
```

[2]:

```
# read penquins dataset from github
penguins = pd.read_csv('https://raw.githubusercontent.com/allisonhorst/
 ↪palmerpenguins/master/inst/extdata/penguins.csv')
penguins.head()
```

[2]:

```
  species     island  bill_length_mm  bill_depth_mm  flipper_length_mm  \
0  Adelie  Torgersen            39.1           18.7              181.0
1  Adelie  Torgersen            39.5           17.4              186.0
2  Adelie  Torgersen            40.3           18.0              195.0
3  Adelie  Torgersen             NaN            NaN                NaN
4  Adelie  Torgersen            36.7           19.3              193.0

   body_mass_g     sex  year
0       3750.0    male  2007
1       3800.0  female  2007
2       3250.0  female  2007
3          NaN     NaN  2007
4       3450.0  female  2007
```

[5]:

```
# drop the year column, it is not useful for our analysis,
# and it has no adequate explanation in the dataset documentation
penguins = penguins.drop('year', axis=1)
penguins.head()
```

[5]:

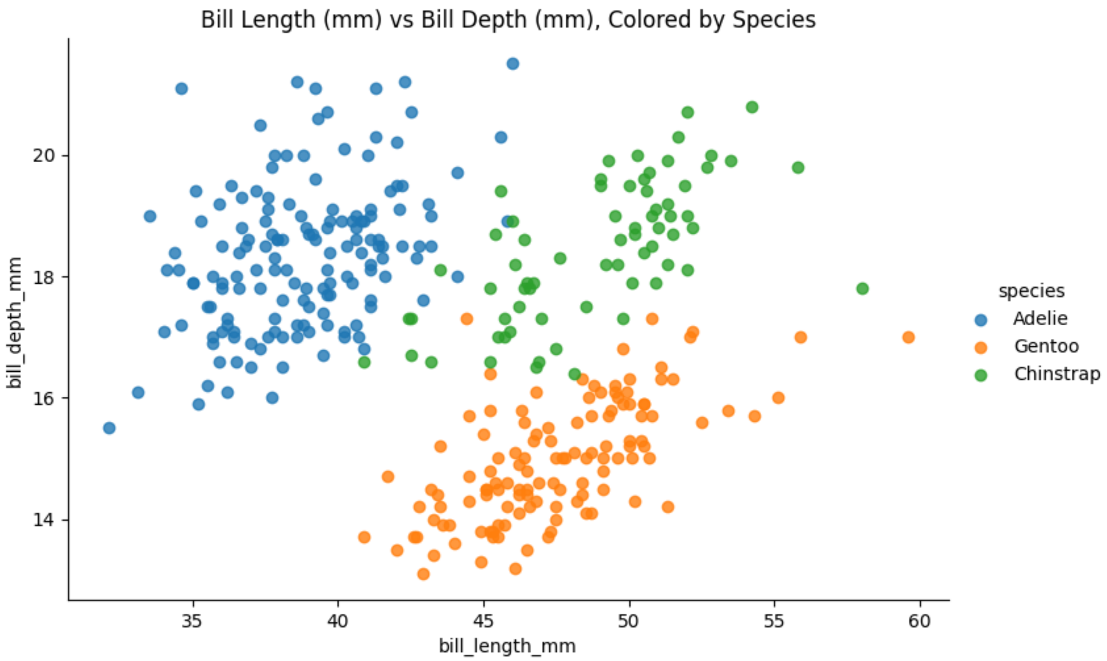
```
  species     island  bill_length_mm  bill_depth_mm  flipper_length_mm  \
0  Adelie  Torgersen            39.1           18.7              181.0
1  Adelie  Torgersen            39.5           17.4              186.0
2  Adelie  Torgersen            40.3           18.0              195.0
3  Adelie  Torgersen             NaN            NaN                NaN
4  Adelie  Torgersen            36.7           19.3              193.0

   body_mass_g     sex
0       3750.0    male
1       3800.0  female
2       3250.0  female
3          NaN     NaN
4       3450.0  female
```

[6]:

```
# Create a scatterplot of bill length vs bill depth using seaborn, hue by␣
 ↪species.
# Add a title.
```

```
sns.lmplot(x='bill_length_mm', y='bill_depth_mm', data=penguins, hue='species',␣
 ↪aspect=1.5, fit_reg=False)
plt.title('Bill Length (mm) vs Bill Depth (mm), Colored by Species')
```

[6]:
```
Text(0.5, 1.0, 'Bill Length (mm) vs Bill Depth (mm), Colored by Species')
```



[7]:
```
numeric_features = ['bill_length_mm', 'bill_depth_mm',␣
 ↪'flipper_length_mm',         'body_mass_g']
categorical_features = ['island', 'sex']
```

[11]:
```
# create a pipeline to impute missing values with the mean and scale numeric␣
 ↪features
from sklearn.pipeline import make_pipeline
num_pipeline = make_pipeline(SimpleImputer(strategy='mean'), StandardScaler())

# create a pipeline to impute missing values with the most frequent value and␣
 ↪one-hot encode categorical features
cat_pipeline = make_pipeline(SimpleImputer(strategy='most_frequent'),␣
 ↪OneHotEncoder(handle_unknown='ignore'))

# create a column transformer to apply the numeric and categorical pipelines to␣
 ↪the correct features
# use remainder='passthrough' to keep the remaining features in the dataframe
preprocessor = ColumnTransformer([
    ('num', num_pipeline, numeric_features),
     ('cat', cat_pipeline, categorical_features)
     ],
      remainder='passthrough')

# fit_transform the preprocessor on the penguins dataset
# convert the result to a dataframe
# use the preprocessor's get_feature_names_out() method to get the column names
transformed_data = preprocessor.fit_transform(penguins)
transformed_df = pd.DataFrame(transformed_data, columns=preprocessor.
 ↪get_feature_names_out())

# display the first 5 rows of the preprocessed dataframe/Users/juice/uvic/CSC␣
 ↪361/test.txt
print(transformed_df.head())
```

```
  num__bill_length_mm num__bill_depth_mm num__flipper_length_mm  \
0           -0.887081           0.787743              -1.422488
1           -0.813494           0.126556              -1.065352
2            -0.66632           0.431719              -0.422507
3                -0.0                0.0                    0.0
4           -1.328605           1.092905              -0.565361

  num__body_mass_g cat__island_Biscoe cat__island_Dream cat__island_Torgersen  \
0        -0.565789                0.0               0.0                   1.0
1        -0.503168                0.0               0.0                   1.0
2        -1.192003                0.0               0.0                   1.0
3              0.0                0.0               0.0                   1.0
4        -0.941517                0.0               0.0                   1.0

  cat__sex_female cat__sex_male remainder__species
0             0.0           1.0             Adelie
1             1.0           0.0             Adelie
2             1.0           0.0             Adelie
3             0.0           1.0             Adelie
4             1.0           0.0             Adelie
```

[47]:
```
# separate the features from the target
# call the features X and the target y
X = transformed_df.drop('remainder__species', axis=1)
y = transformed_df['remainder__species']
```

[43]:
```
# setup binary classification for Adelie vs. rest of species
# use the Adelie species as the positive class
# create a new target called y_adelie
y_adelie = (transformed_df['remainder__species'] == 'Adelie').astype(int)
```

[48]:

```
# build an SGDClassifier model using X and y
# use random_state=42 for reproducibility
from sklearn.linear_model import SGDClassifier
sgd_clf = SGDClassifier(random_state=42)
sgd_clf.fit(X, y_adelie)
```

[48]:

```
SGDClassifier(random_state=42)
```

[49]:

```
# compute the accuracy using cross_val_score with cv=10
accuracy_scores = cross_val_score(sgd_clf, X, y_adelie, cv=10,␣
 ↪scoring='accuracy')
accuracy_scores
```

[49]:

```
array([1.        , 1.        , 0.97142857, 1.        , 1.        ,
       1.        , 1.        , 1.        , 1.        , 0.97058824])
```

[50]:

```
# compute the mean accuracy
mean_accuracy = accuracy_scores.mean()
mean_accuracy
```

[50]:

```
0.9942016806722689
```

[53]:

```
# predict the target using cross_val_predict with cv=10
# call the result y_train_pred
from sklearn.model_selection import cross_val_predict
y_train_pred = cross_val_predict(sgd_clf, X, y_adelie, cv=10)
```

[57]:

```
# compute the confusion matrix
from sklearn.metrics import confusion_matrix
cm = confusion_matrix(y_adelie, y_train_pred)
cm
```

[57]:

```
array([[150,   2],
       [  0, 192]])
```

[31]:

```
# compute the precision score using precision_score()
precision_score(y_adelie, y_train_pred)
```

[31]:

```
1.0
```

[68]:

```
# compute the recall score using recall_score()
recall_score(y_adelie, y_train_pred)
```

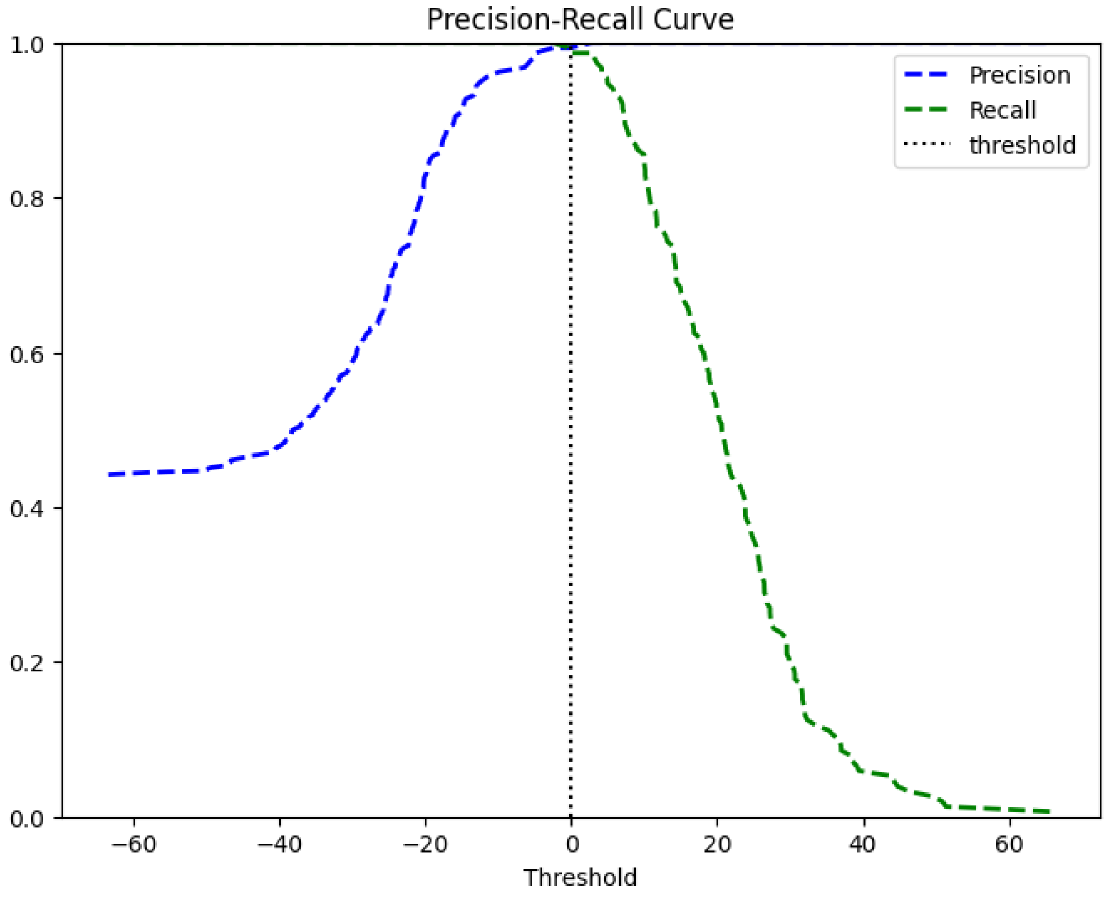
[68]:

```
0.9868421052631579
```

[70]:

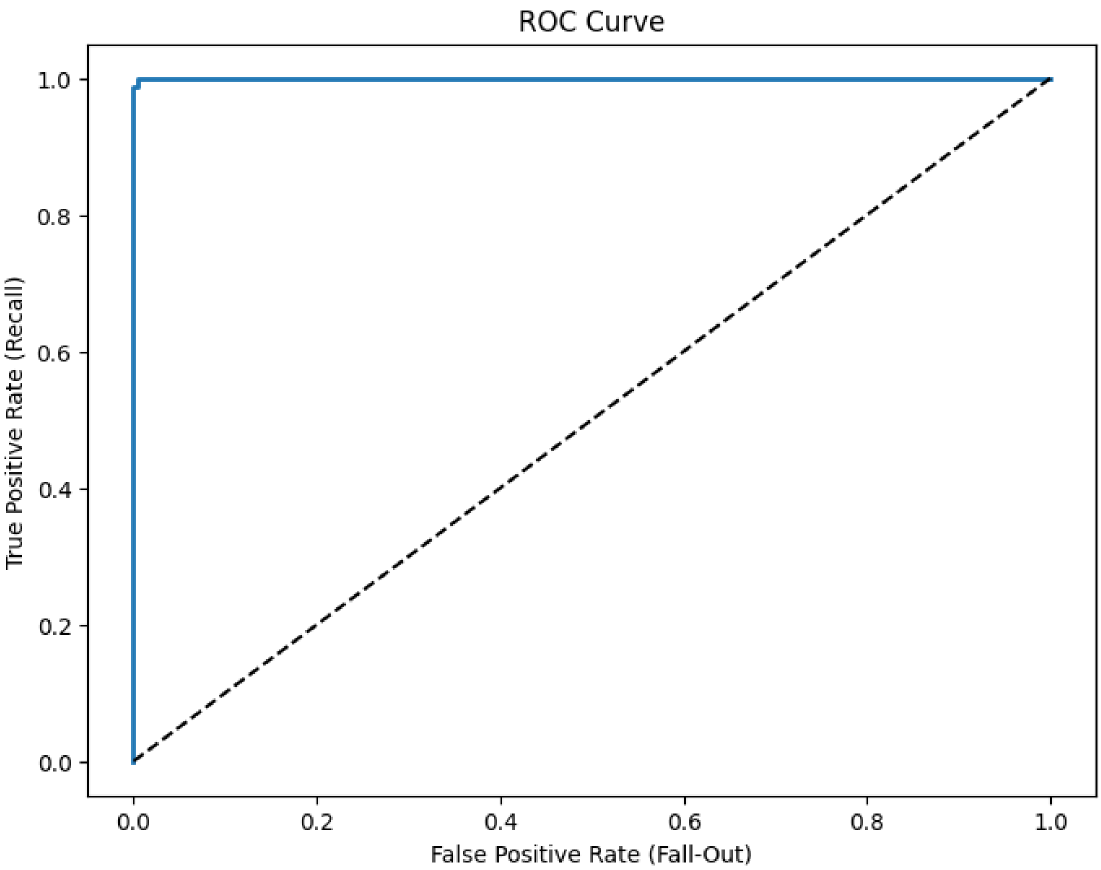
```
# draw the precision-recall curve
# call the result precisions, recalls, thresholds
```

```
from sklearn.metrics import precision_recall_curve
y_scores = cross_val_predict(sgd_clf, X, y_adelie, cv=10,␣
  ↪method='decision_function')
y_scores
precisions, recalls, thresholds = precision_recall_curve(y_adelie, y_scores)
thresholds.shape
plt.figure(figsize=(8, 6))
plt.plot(thresholds, precisions[:-1], 'b--', label='Precision', linewidth=2)
plt.plot(thresholds, recalls[:-1], 'g--', label='Recall', linewidth=2)
threshold = 0
plt.vlines(threshold, 0, 1.0, "k", "dotted", label="threshold")
plt.xlabel('Threshold')
plt.legend(loc='best')
plt.ylim([0, 1])
plt.title('Precision-Recall Curve')
plt.show()
```



[73]:

```
# call the result fpr, tpr, thresholds
from sklearn.metrics import roc_curve
fpr, tpr, thresholds = roc_curve(y_adelie, y_scores)
# plot the roc curve
plt.figure(figsize=(8, 6))
plt.plot(fpr, tpr, linewidth=2)
plt.plot([0, 1], [0, 1], 'k--')
plt.xlabel('False Positive Rate (Fall-Out)')
plt.ylabel('True Positive Rate (Recall)')
plt.title('ROC Curve')
plt.show()
```



[74]:

```
# now let's do multiclass classification
# build an SGDClassifier model using X and y
# use random_state=42 for reproducibility
sgd_clf_multi = SGDClassifier(random_state=42)
sgd_clf_multi.fit(X, y)
```

[74]:

```
SGDClassifier(random_state=42)
```

[77]:

```
# show the mean accuracy using cross_val_score with cv=10
accuracy_scores = cross_val_score(sgd_clf_multi, X, y, cv=10,␣
 ↪scoring='accuracy')
mean_accuracy = accuracy_scores.mean()
mean_accuracy
```

[77]:

```
0.9883193277310924
```

[78]:

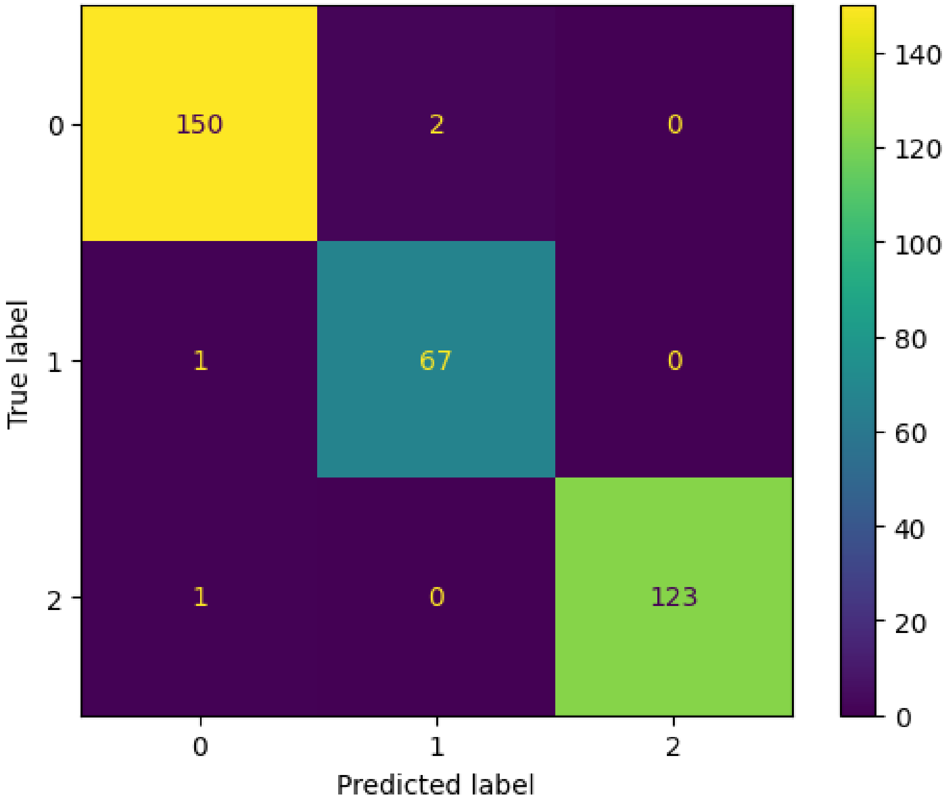
```
# predict the target using cross_val_predict with cv=10
# call the result y_train_pred
# show the confusion matrix
y_train_pred = cross_val_predict(sgd_clf_multi, X, y, cv=10)
cm = confusion_matrix(y, y_train_pred)
cm
```

[78]:

```
array([[150,   2,   0],
       [  1,  67,   0],
       [  1,   0, 123]])
```

[79]:

```
# use ConfusionMatrixDisplay to display the confusion matrix
ConfusionMatrixDisplay(cm).plot()
plt.show()
```


True label
Predicted label