Final_Notebook_updated

November 26, 2024

[42]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split, cross_val_score, KFold,␣
↪GridSearchCV
from sklearn.preprocessing import LabelEncoder, StandardScaler
from sklearn.impute import KNNImputer
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import confusion_matrix, ConfusionMatrixDisplay,␣
↪classification_report
import tensorflow
from tensorflow import keras as keras
from keras.models import Sequential
from keras.layers import Dense, Dropout, LSTM, Flatten
from keras.utils import to_categorical
from keras.callbacks import EarlyStopping
from keras.optimizers import Adam
from sklearn.ensemble import RandomForestClassifier
from sklearn.ensemble import GradientBoostingClassifier
from sklearn.svm import SVC
from sklearn.metrics import accuracy_score
from keras.models import Model
from keras.layers import Input, GRU
```

0.1 Load Data from CSV

Data from NOAA, National Oceanic and Atmospheric Administration, daily SST (Sea Surface Temperature) and storms.csv file from Kaggle.com.

Python program fill_storms_sst.ipynb file is used to combine data.

[2]:
```
# GitHub base URL for raw CSV files
base_url = "https://raw.githubusercontent.com/soojirhodes/DATA780_Final_Project/
↪main/"

# List of CSV files to load from the GitHub repository
csv_files = [
    "storms_with_sst_1981-1990.csv",
```

```
    "storms_with_sst_1991-2000.csv",
    "storms_with_sst_2001-2010.csv",
    "storms_with_sst_2011-2021.csv"
]

# Construct full URLs and read each CSV file into a DataFrame
csv_urls = [base_url + file for file in csv_files]

# Read each CSV file into a DataFrame and store them in a list
dataframes = [pd.read_csv(url) for url in csv_urls]

# Concatenate all DataFrames into a single DataFrame
df_all = pd.concat(dataframes, ignore_index=True)
```

[3]:
```
df_all.head()
```

[3]:
```
   Unnamed: 0    name  year  month  day  hour   lat  long  \
0        1268  Arlene  1981      5    6    18  18.4 -83.6
1        1269  Arlene  1981      5    7     0  18.4 -82.7
2        1270  Arlene  1981      5    7     6  18.6 -81.7
3        1271  Arlene  1981      5    7    12  19.0 -80.6
4        1272  Arlene  1981      5    7    18  19.6 -79.7

                status  category  wind  pressure  \
0  tropical depression       NaN    30      1006
1  tropical depression       NaN    30      1006
2  tropical depression       NaN    30      1005
3       tropical storm       NaN    35      1003
4       tropical storm       NaN    40      1000

   tropicalstorm_force_diameter  hurricane_force_diameter  sst
0                           NaN                       NaN  NaN
1                           NaN                       NaN  NaN
2                           NaN                       NaN  NaN
3                           NaN                       NaN  NaN
4                           NaN                       NaN  NaN
```

[4]:
```
df_all.shape
```

[4]:
```
(17799, 15)
```

0.2 Data ETL (Extract, Transform, Load)

[5]:
```
# Select relevant columns and handle missing values

# Select relevant columns
```

```
df = df_all[['wind', 'pressure', 'sst', 'category']].copy()  # Make a copy to
↪avoid warnings
```

Initialize the KNN imputer with a chosen number of neighbors (e.g., 5) for missing data.

Filled missing category data with 0's (representing "not a hurricane").

[6]:
```
# KNN Imputer is used to fill in missing values based on the nearest neighbors,
↪providing more accurate estimates compared to mean/mode imputation.
imputer = KNNImputer(n_neighbors=5)
```

[7]:
```
# Apply KNN imputation only to the numerical columns
df.loc[:, ['wind', 'pressure', 'sst']] = imputer.fit_transform(df[['wind',
↪'pressure', 'sst']])

# Replace missing values in the 'category' column with 0.0 (representing "not a
↪hurricane")
df['category'] = df['category'].fillna(0.0)
```

Adding rolling averages to the DataFrame

[8]:
```
# Rolling averages help capture temporal trends in the data, which may provide
↪useful insights for prediction.
df['sst_rolling_mean'] = df['sst'].rolling(window=3, min_periods=1).mean()
df['pressure_rolling_mean'] = df['pressure'].rolling(window=3, min_periods=1).
↪mean()

# Drop any rows with remaining missing values (optional)
df.dropna(inplace=True)

df.head()
```

[8]:

	wind	pressure	sst	category	sst_rolling_mean	pressure_rolling_mean
0	30	1006	22.288	0.0	22.288000	1006.000000
1	30	1006	22.288	0.0	22.288000	1006.000000
2	30	1005	23.950	0.0	22.842000	1005.666667
3	35	1003	24.032	0.0	23.423333	1004.666667
4	40	1000	24.202	0.0	24.061333	1002.666667

Get distinct values of the 'category' column

[9]:
```
# Checking distinct categories to ensure that the data has been processed
↪correctly.
distinct_categories = df['category'].unique()

# Print the distinct values
print(distinct_categories)
print(df.shape)
```

```
[0. 1. 2. 3. 4. 5.]
(17799, 6)
```

Encode the categorical labels (hurricane categories)

[10]:
```
# Label encoding is used to convert categorical labels into numeric values for machine learning models.
label_encoder = LabelEncoder()
df['category'] = label_encoder.fit_transform(df['category'])
num_classes = df['category'].nunique()  # Number of unique categories
```

[11]:
```
# Split the data into features (X) and labels (y)
X = df[['pressure', 'sst', 'sst_rolling_mean', 'pressure_rolling_mean']].values
y = df['category'].values
```

Scale the features

[12]:
```
# Feature scaling ensures that all features contribute equally to the model, improving convergence during training.
scaler = StandardScaler()
X = scaler.fit_transform(X)
```

1 Training

[13]:
```
# Split the data into training and testing sets
```

[14]:
```
# Splitting the data into training and testing sets to evaluate model performance on unseen data.
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

1.1 Logistic Regression

[51]:
```
# Convert one-hot encoded `y` to a 1D array of labels
y_labels = np.array(y)

# Splitting the data into training and testing sets
X_train_lr, X_test_lr, y_train_lr, y_test_lr = train_test_split(X, y_labels, test_size=0.2, random_state=42)

# Initialize Logistic Regression model
logistic_model = LogisticRegression(multi_class='multinomial', solver='lbfgs',max_iter=1000, random_state=42)

# Train the model
logistic_model.fit(X_train_lr, y_train_lr)
```

```
/usr/local/lib/python3.10/dist-packages/sklearn/linear_model/_logistic.py:1247:
FutureWarning: 'multi_class' was deprecated in version 1.5 and will be removed
in 1.7. From then on, it will always use 'multinomial'. Leave it to its default
value to avoid this warning.
  warnings.warn(
```

[51]:
```
LogisticRegression(max_iter=1000, multi_class='multinomial', random_state=42)
```

1.2 RandomForest Classifier

[15]:
```python
# Performing grid search to optimize hyperparameters for RandomForest.
rf_param_grid = {
    'n_estimators': [50, 100, 200],
    'max_depth': [None, 10, 20],
    'min_samples_split': [2, 5, 10]
}
rf_grid_search = GridSearchCV(RandomForestClassifier(random_state=42),␣
 ↪rf_param_grid, cv=5, n_jobs=-1, verbose=2)
rf_grid_search.fit(X_train, y_train)
rf_best_model = rf_grid_search.best_estimator_

# Train the best RandomForest model
rf_best_model.fit(X_train, y_train)
```

```
Fitting 5 folds for each of 27 candidates, totalling 135 fits
```

[15]:
```
RandomForestClassifier(min_samples_split=5, n_estimators=200, random_state=42)
```

1.3 SVM Classifier

[16]:
```python
# Performing grid search to optimize hyperparameters for SVM.
svm_param_grid = {
    'C': [0.1, 1, 10],
    'kernel': ['linear', 'rbf'],
    'gamma': ['scale', 'auto']
}
svm_grid_search = GridSearchCV(SVC(probability=True, random_state=42),␣
 ↪svm_param_grid, cv=5, n_jobs=-1, verbose=2)
svm_grid_search.fit(X_train, y_train)
svm_best_model = svm_grid_search.best_estimator_

# Train the best SVM model
svm_best_model.fit(X_train, y_train)
```

```
Fitting 5 folds for each of 12 candidates, totalling 60 fits
```

[16]:
```
SVC(C=10, probability=True, random_state=42)
```

1.4 Gradient Boosting Classifier

[17]:
```
# Gradient Boosting is used for its ability to improve accuracy by combining␣
↪weak learners in a sequential manner.
gb_model = GradientBoostingClassifier(n_estimators=100, learning_rate=0.1,␣
↪max_depth=3, random_state=42)
gb_model.fit(X_train, y_train)
```

[17]:
```
GradientBoostingClassifier(random_state=42)
```

1.5

1.6 ANN Model

[18]:
```
# ANN model with Dropout for regularization
# The ANN model is chosen for its ability to capture complex, non-linear␣
↪relationships in the data.
model = Sequential()
model.add(Dense(128, input_shape=(X.shape[1],), activation='relu'))
model.add(Dropout(0.4))  # Dropout layer to prevent overfitting
model.add(Dense(64, activation='relu'))
model.add(Dropout(0.3))
model.add(Dense(32, activation='relu'))
model.add(Dense(num_classes, activation='softmax'))

model.summary()
```

```
/usr/local/lib/python3.10/dist-packages/keras/src/layers/core/dense.py:87:
UserWarning: Do not pass an `input_shape`/`input_dim` argument to a layer. When
using Sequential models, prefer using an `Input(shape)` object as the first
layer in the model instead.
  super().__init__(activity_regularizer=activity_regularizer, **kwargs)
```

```
Model: "sequential"


 Layer (type)                         Output Shape                          ␣
↪Param #

 dense (Dense)                        (None, 128)                               ␣
↪640

 dropout (Dropout)                    (None, 128)                               ␣
↪  0

 dense_1 (Dense)                      (None, 64)                              ␣
↪8,256
```

```
 dropout_1 (Dropout)                    (None, 64)                          ␣
↪  0

 dense_2 (Dense)                        (None, 32)                         ␣
↪2,080

 dense_3 (Dense)                        (None, 6)                           ␣
↪198
```

```
 Total params: 11,174 (43.65 KB)

 Trainable params: 11,174 (43.65 KB)

 Non-trainable params: 0 (0.00 B)
```

[19]:

```
# Compile the ANN model with a smaller learning rate
# Compiling the model with Adam optimizer and a lower learning rate to achieve␣
 ↪better convergence.
model.compile(optimizer=Adam(learning_rate=0.0005),␣
 ↪loss='sparse_categorical_crossentropy', metrics=['accuracy'])
```

[20]:

```
# Train the ANN model with early stopping to prevent overfitting
# Early stopping is used to stop training when the validation loss stops␣
 ↪improving, helping to avoid overfitting.
early_stopping = EarlyStopping(monitor='val_loss', patience=5,␣
 ↪restore_best_weights=True)
history = model.fit(X_train, y_train, validation_split=0.2, epochs=100,␣
 ↪batch_size=16, callbacks=[early_stopping],
                    verbose=1)
```

```
Epoch 1/100
712/712             3s 2ms/step -
accuracy: 0.7403 - loss: 0.7756 - val_accuracy: 0.8732 - val_loss: 0.3136
Epoch 2/100
712/712             2s 2ms/step -
accuracy: 0.8598 - loss: 0.3410 - val_accuracy: 0.8869 - val_loss: 0.2837
Epoch 3/100
712/712             3s 4ms/step -
accuracy: 0.8683 - loss: 0.3196 - val_accuracy: 0.8926 - val_loss: 0.2758
Epoch 4/100
712/712             2s 3ms/step -
accuracy: 0.8751 - loss: 0.3050 - val_accuracy: 0.8824 - val_loss: 0.2657
Epoch 5/100
712/712             2s 2ms/step -
```

```
accuracy: 0.8820 - loss: 0.2897 - val_accuracy: 0.8897 - val_loss: 0.2638
Epoch 6/100
712/712            2s 2ms/step -
accuracy: 0.8808 - loss: 0.2859 - val_accuracy: 0.8901 - val_loss: 0.2583
Epoch 7/100
712/712            2s 2ms/step -
accuracy: 0.8868 - loss: 0.2684 - val_accuracy: 0.8866 - val_loss: 0.2601
Epoch 8/100
712/712            3s 2ms/step -
accuracy: 0.8852 - loss: 0.2749 - val_accuracy: 0.8890 - val_loss: 0.2556
Epoch 9/100
712/712            2s 2ms/step -
accuracy: 0.8795 - loss: 0.2799 - val_accuracy: 0.8887 - val_loss: 0.2629
Epoch 10/100
712/712            3s 3ms/step -
accuracy: 0.8813 - loss: 0.2833 - val_accuracy: 0.8901 - val_loss: 0.2544
Epoch 11/100
712/712            2s 3ms/step -
accuracy: 0.8916 - loss: 0.2640 - val_accuracy: 0.8950 - val_loss: 0.2534
Epoch 12/100
712/712            2s 2ms/step -
accuracy: 0.8849 - loss: 0.2705 - val_accuracy: 0.8957 - val_loss: 0.2494
Epoch 13/100
712/712            2s 2ms/step -
accuracy: 0.8863 - loss: 0.2756 - val_accuracy: 0.8894 - val_loss: 0.2535
Epoch 14/100
712/712            3s 2ms/step -
accuracy: 0.8852 - loss: 0.2685 - val_accuracy: 0.8971 - val_loss: 0.2439
Epoch 15/100
712/712            2s 2ms/step -
accuracy: 0.8826 - loss: 0.2724 - val_accuracy: 0.8957 - val_loss: 0.2532
Epoch 16/100
712/712            2s 2ms/step -
accuracy: 0.8898 - loss: 0.2670 - val_accuracy: 0.9010 - val_loss: 0.2430
Epoch 17/100
712/712            5s 5ms/step -
accuracy: 0.8846 - loss: 0.2703 - val_accuracy: 0.8968 - val_loss: 0.2433
Epoch 18/100
712/712            3s 2ms/step -
accuracy: 0.8874 - loss: 0.2571 - val_accuracy: 0.8961 - val_loss: 0.2504
Epoch 19/100
712/712            3s 2ms/step -
accuracy: 0.8886 - loss: 0.2725 - val_accuracy: 0.9003 - val_loss: 0.2456
Epoch 20/100
712/712            3s 5ms/step -
accuracy: 0.8913 - loss: 0.2707 - val_accuracy: 0.8954 - val_loss: 0.2486
Epoch 21/100
712/712            2s 3ms/step -
```

```
accuracy: 0.8885 - loss: 0.2650 - val_accuracy: 0.9034 - val_loss: 0.2383
Epoch 22/100
712/712              3s 3ms/step -
accuracy: 0.8966 - loss: 0.2480 - val_accuracy: 0.8999 - val_loss: 0.2413
Epoch 23/100
712/712              2s 2ms/step -
accuracy: 0.8909 - loss: 0.2525 - val_accuracy: 0.9006 - val_loss: 0.2394
Epoch 24/100
712/712              3s 2ms/step -
accuracy: 0.8916 - loss: 0.2598 - val_accuracy: 0.9045 - val_loss: 0.2342
Epoch 25/100
712/712              2s 2ms/step -
accuracy: 0.8932 - loss: 0.2478 - val_accuracy: 0.9038 - val_loss: 0.2348
Epoch 26/100
712/712              2s 2ms/step -
accuracy: 0.8991 - loss: 0.2452 - val_accuracy: 0.9070 - val_loss: 0.2360
Epoch 27/100
712/712              5s 7ms/step -
accuracy: 0.8924 - loss: 0.2554 - val_accuracy: 0.8999 - val_loss: 0.2405
Epoch 28/100
712/712              5s 7ms/step -
accuracy: 0.8947 - loss: 0.2603 - val_accuracy: 0.8996 - val_loss: 0.2424
Epoch 29/100
712/712              3s 4ms/step -
accuracy: 0.8909 - loss: 0.2499 - val_accuracy: 0.9017 - val_loss: 0.2356
```

1.7 RNN Model LTSM-based

[21]:
```
# LSTM-based RNN model for sequential learning
# LSTM is used for its ability to capture temporal dependencies in sequential␣
↪data, which is important for time-series analysis.
X_reshaped = X.reshape((X.shape[0], 1, X.shape[1]))  # Reshape data to be 3D␣
↪for LSTM

lstm_model = Sequential()
lstm_model.add(LSTM(50, input_shape=(X_reshaped.shape[1], X_reshaped.shape[2]),␣
↪activation='relu'))
lstm_model.add(Dense(32, activation='relu'))
lstm_model.add(Dense(num_classes, activation='softmax'))
```

```
/usr/local/lib/python3.10/dist-packages/keras/src/layers/rnn/rnn.py:204:
UserWarning: Do not pass an `input_shape`/`input_dim` argument to a layer. When
using Sequential models, prefer using an `Input(shape)` object as the first
layer in the model instead.
  super().__init__(**kwargs)
```

[22]:
```
# Compile the LSTM model
lstm_model.compile(optimizer='adam', loss='sparse_categorical_crossentropy',␣
↪metrics=['accuracy'])
```

[23]:
```
# Train the LSTM model
lstm_history = lstm_model.fit(X_reshaped, y, epochs=50, batch_size=16,␣
↪validation_split=0.2, callbacks=[early_stopping],
                              verbose=1)
```

```
Epoch 1/50
890/890              5s 3ms/step -
accuracy: 0.7905 - loss: 0.7729 - val_accuracy: 0.8612 - val_loss: 0.3449
Epoch 2/50
890/890              6s 6ms/step -
accuracy: 0.8804 - loss: 0.2840 - val_accuracy: 0.8542 - val_loss: 0.3164
Epoch 3/50
890/890              4s 4ms/step -
accuracy: 0.8922 - loss: 0.2552 - val_accuracy: 0.8697 - val_loss: 0.2915
Epoch 4/50
890/890              3s 4ms/step -
accuracy: 0.8942 - loss: 0.2535 - val_accuracy: 0.8705 - val_loss: 0.2868
Epoch 5/50
890/890              8s 6ms/step -
accuracy: 0.8976 - loss: 0.2420 - val_accuracy: 0.8795 - val_loss: 0.2878
```

2 Validation and Evaluation

2.1 Accuracy

Logistic Regression Accuracy

[55]:
```
# Evaluate Logistic Regression Classifier
lr_predicted_classes = logistic_model.predict(X_test_lr)
lr_accuracy = accuracy_score(y_test_lr, lr_predicted_classes)
print(f"Logistic Regression Test Accuracy: {lr_accuracy * 100:.2f}%")
```

```
Logistic Regression Test Accuracy: 88.88%
```

RandomForest Accuracy

[24]:
```
# Evaluate RandomForest Classifier
rf_predicted_classes = rf_best_model.predict(X_test)
rf_accuracy = accuracy_score(y_test, rf_predicted_classes)
print(f"Random Forest Test Accuracy: {rf_accuracy * 100:.2f}%")
```

```
Random Forest Test Accuracy: 96.12%
```

Gradient Boosting Accuracy

[25]:
```
# Evaluate Gradient Boosting Classifier
gb_predicted_classes = gb_model.predict(X_test)
```

```
gb_accuracy = accuracy_score(y_test, gb_predicted_classes)
print(f"Gradient Boosting Test Accuracy: {gb_accuracy * 100:.2f}%")
```

```
Gradient Boosting Test Accuracy: 94.89%
```

SVM Accuracy

[26]:
```
# Evaluate SVM Classifier
svm_predicted_classes = svm_best_model.predict(X_test)
svm_accuracy = accuracy_score(y_test, svm_predicted_classes)
print(f"SVM Test Accuracy: {svm_accuracy * 100:.2f}%")
```

```
SVM Test Accuracy: 90.73%
```

ANN Accuracy

[27]:
```
# Evaluate the ANN model on the test data
loss, accuracy = model.evaluate(X_test, y_test)
print(f"ANN Test Accuracy: {accuracy * 100:.2f}%")
```

```
112/112              0s 1ms/step -
accuracy: 0.9045 - loss: 0.2401
ANN Test Accuracy: 90.65%
```

[28]:
```
# Make predictions with the ANN model
ann_predictions = model.predict(X_test)
ann_predicted_classes = ann_predictions.argmax(axis=1)
```

```
112/112              0s 2ms/step
```

LSTM Accuracy

[29]:
```
# Evaluate the LSTM model
loss, accuracy = lstm_model.evaluate(X_reshaped, y)
print(f"LSTM Model Accuracy: {accuracy * 100:.2f}%")
```

```
557/557              1s 1ms/step -
accuracy: 0.8686 - loss: 0.3104
LSTM Model Accuracy: 87.34%
```

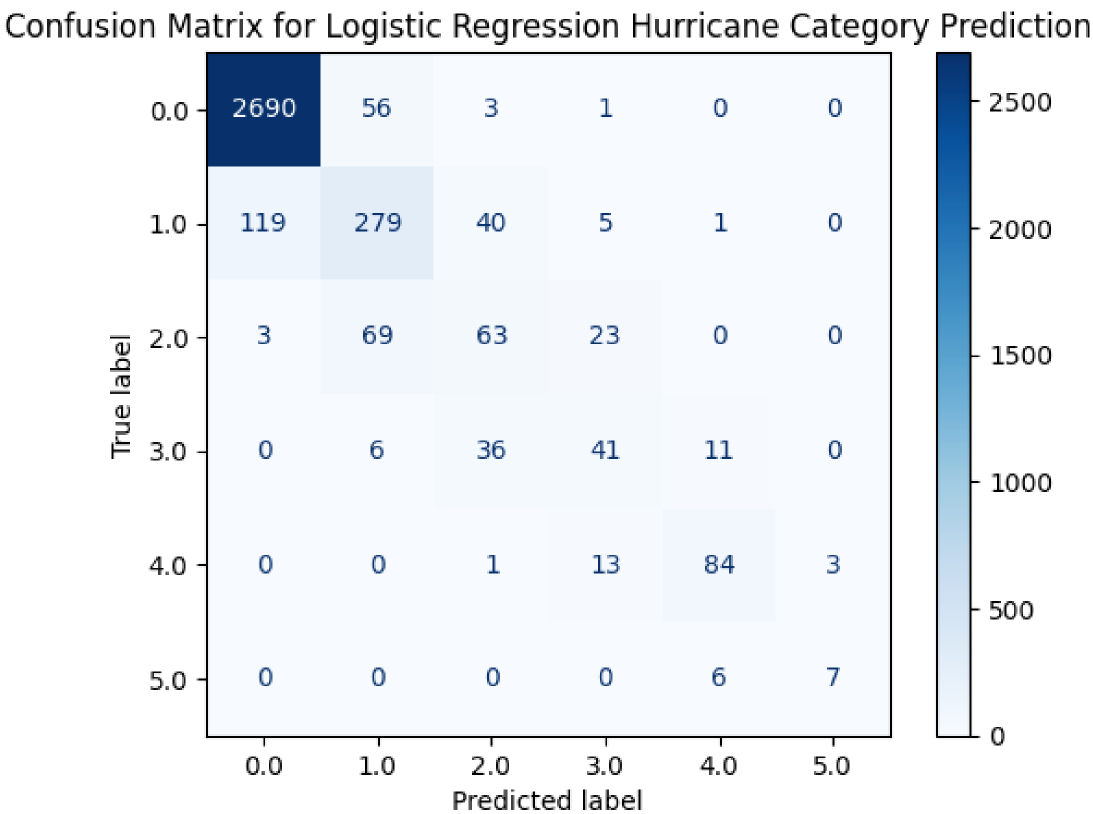
[30]:
```
# Make predictions with the LSTM model
lstm_predictions = lstm_model.predict(X_reshaped)
lstm_predicted_classes = lstm_predictions.argmax(axis=1)
```

```
557/557              1s 2ms/step
```

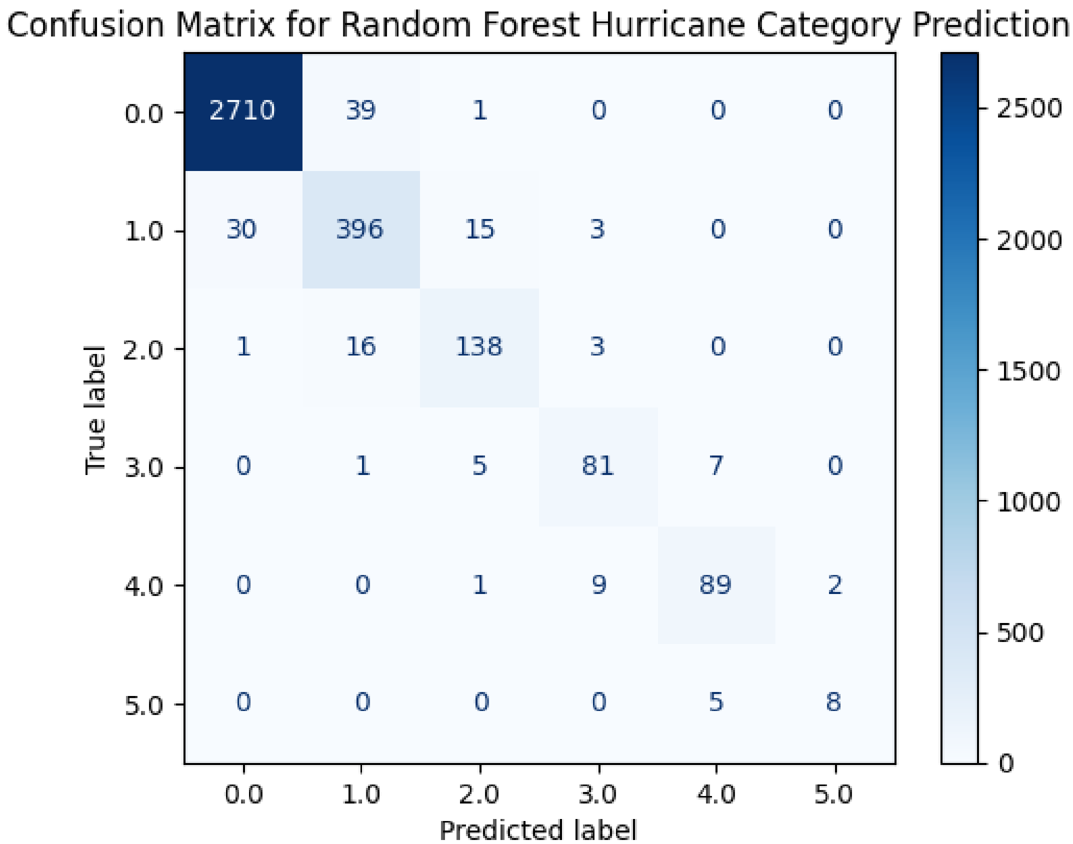
2.2 Confusion Matrix

[57]:
```
# Confusion Matrix Visualizations for All Models
# Logistic Regression Confusion Matrix
lr_cm = confusion_matrix(y_test_lr, lr_predicted_classes, labels=label_encoder.
 ↪classes_)
```

```
lr_disp = ConfusionMatrixDisplay(confusion_matrix=lr_cm,␣
 ↪display_labels=label_encoder.classes_)
lr_disp.plot(cmap=plt.cm.Blues)
plt.title('Confusion Matrix for Logistic Regression Hurricane Category␣
 ↪Prediction')
plt.show()
```

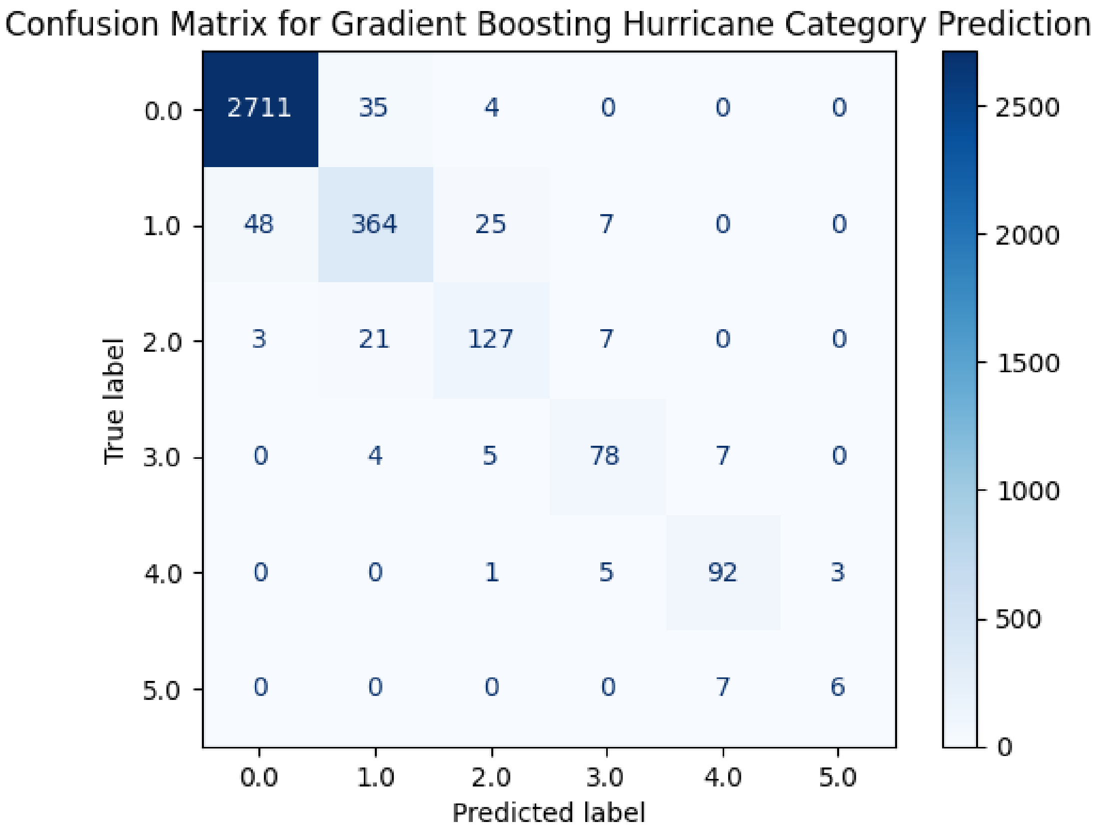


[31]:
```
# RandomForest Confusion Matrix
rf_cm = confusion_matrix(y_test, rf_predicted_classes, labels=label_encoder.
 ↪classes_)
rf_disp = ConfusionMatrixDisplay(confusion_matrix=rf_cm,␣
 ↪display_labels=label_encoder.classes_)
rf_disp.plot(cmap=plt.cm.Blues)
plt.title('Confusion Matrix for Random Forest Hurricane Category Prediction')
plt.show()
```

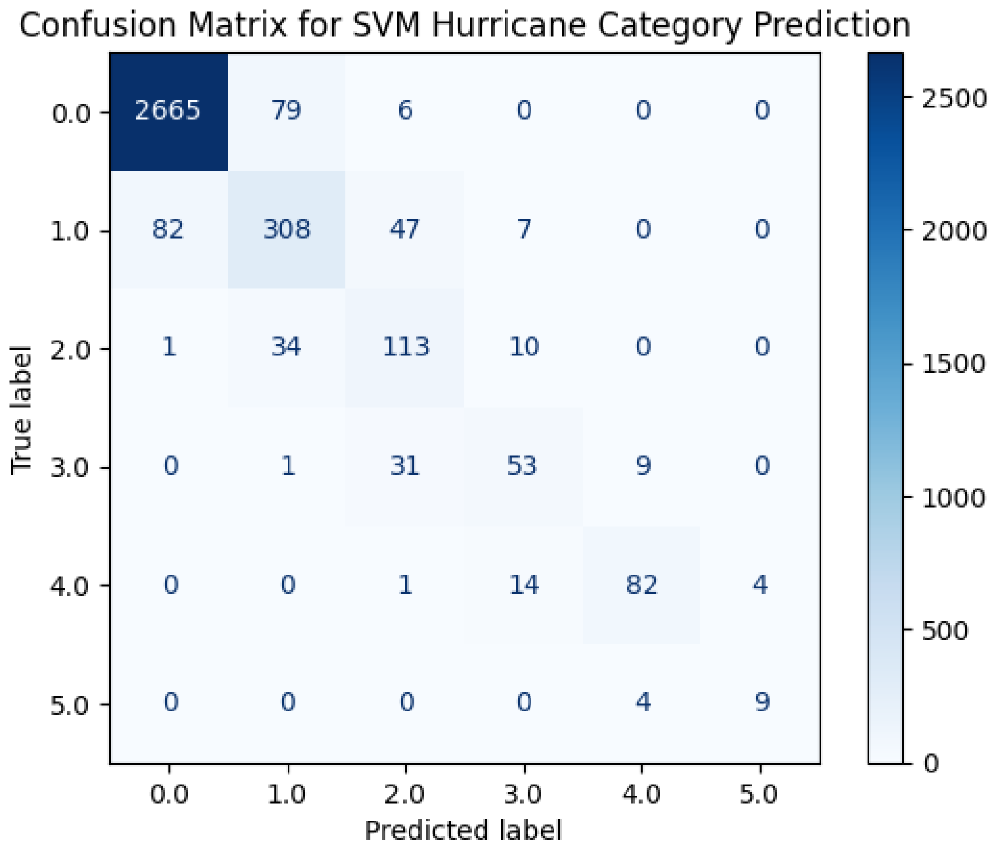


[32]:

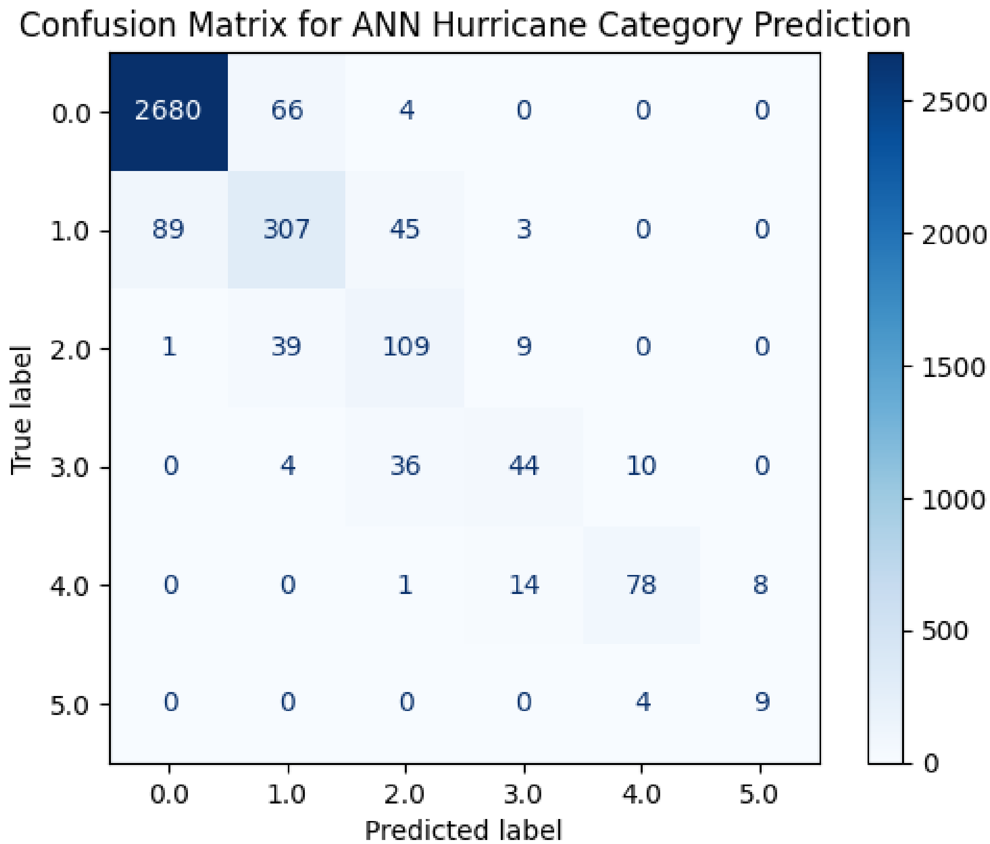
```
# Gradient Boosting Confusion Matrix
gb_cm = confusion_matrix(y_test, gb_predicted_classes, labels=label_encoder.
 ↪classes_)
gb_disp = ConfusionMatrixDisplay(confusion_matrix=gb_cm,␣
 ↪display_labels=label_encoder.classes_)
gb_disp.plot(cmap=plt.cm.Blues)
plt.title('Confusion Matrix for Gradient Boosting Hurricane Category␣
 ↪Prediction')
plt.show()
```



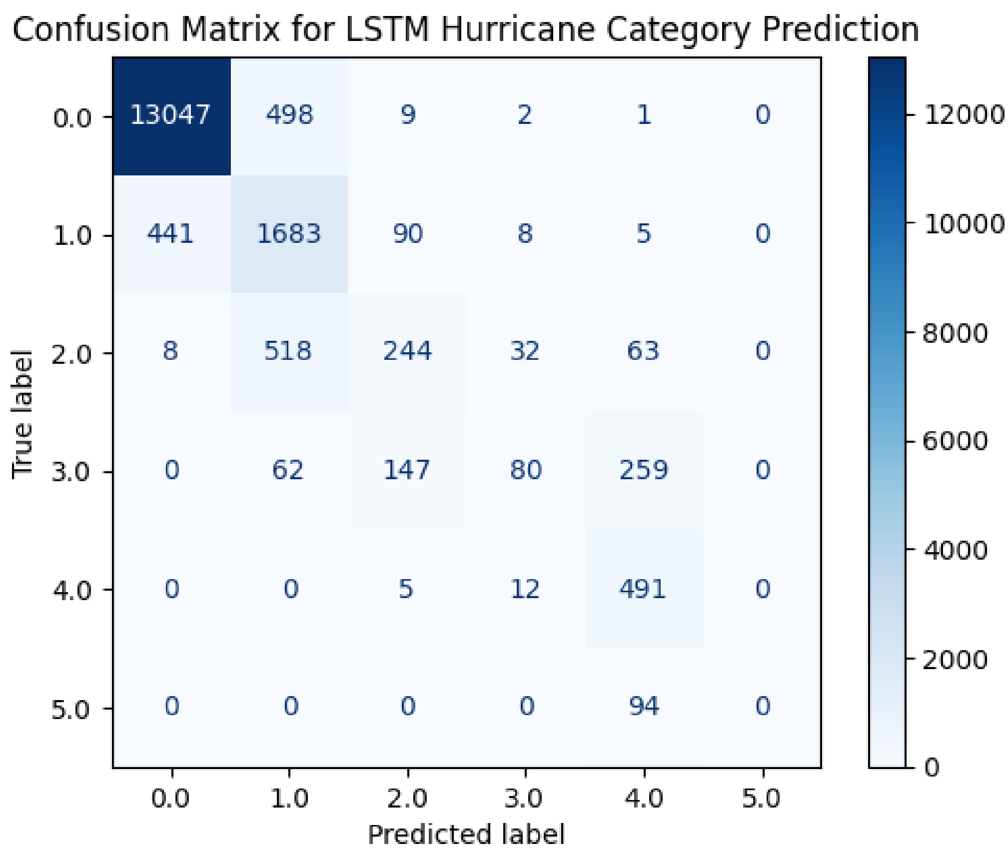
[33]:
```
# SVM Confusion Matrix
svm_cm = confusion_matrix(y_test, svm_predicted_classes, labels=label_encoder.
    ↪classes_)
svm_disp = ConfusionMatrixDisplay(confusion_matrix=svm_cm,␣
    ↪display_labels=label_encoder.classes_)
svm_disp.plot(cmap=plt.cm.Blues)
plt.title('Confusion Matrix for SVM Hurricane Category Prediction')
plt.show()
```



[34]:
```
# ANN Confusion Matrix
ann_cm = confusion_matrix(y_test, ann_predicted_classes, labels=label_encoder.
  ↪classes_)
ann_disp = ConfusionMatrixDisplay(confusion_matrix=ann_cm,␣
  ↪display_labels=label_encoder.classes_)
ann_disp.plot(cmap=plt.cm.Blues)
plt.title('Confusion Matrix for ANN Hurricane Category Prediction')
plt.show()
```



[35]:
```
# LSTM Confusion Matrix
lstm_cm = confusion_matrix(y, lstm_predicted_classes, labels=label_encoder.
 ↪classes_)
lstm_disp = ConfusionMatrixDisplay(confusion_matrix=lstm_cm,␣
 ↪display_labels=label_encoder.classes_)
lstm_disp.plot(cmap=plt.cm.Blues)
plt.title('Confusion Matrix for LSTM Hurricane Category Prediction')
plt.show()
```



2.3 Classification Reports

[58]:
```
# Classification Report for All Models
print("Classification Report for Logistic Regression:")
print(classification_report(y_test_lr, lr_predicted_classes,␣
↪target_names=label_encoder.classes_.astype(str)))
```

```
Classification Report for Logistic Regression:
              precision    recall  f1-score   support

         0.0       0.96      0.98      0.97      2750
         1.0       0.68      0.63      0.65       444
         2.0       0.44      0.40      0.42       158
         3.0       0.49      0.44      0.46        94
         4.0       0.82      0.83      0.83       101
         5.0       0.70      0.54      0.61        13

    accuracy                           0.89      3560
   macro avg       0.68      0.64      0.66      3560
weighted avg       0.88      0.89      0.89      3560
```

[36]:
```
print("Classification Report for Random Forest:")
print(classification_report(y_test, rf_predicted_classes,␣
↪target_names=label_encoder.classes_.astype(str)))
```

```
Classification Report for Random Forest:
              precision    recall  f1-score   support

         0.0       0.99      0.99      0.99      2750
         1.0       0.88      0.89      0.88       444
         2.0       0.86      0.87      0.87       158
         3.0       0.84      0.86      0.85        94
         4.0       0.88      0.88      0.88       101
         5.0       0.80      0.62      0.70        13

    accuracy                           0.96      3560
   macro avg       0.88      0.85      0.86      3560
weighted avg       0.96      0.96      0.96      3560
```

[37]:
```
print("Classification Report for Gradient Boosting:")
print(classification_report(y_test, gb_predicted_classes,␣
↪target_names=label_encoder.classes_.astype(str)))
```

```
Classification Report for Gradient Boosting:
              precision    recall  f1-score   support

         0.0       0.98      0.99      0.98      2750
         1.0       0.86      0.82      0.84       444
         2.0       0.78      0.80      0.79       158
         3.0       0.80      0.83      0.82        94
         4.0       0.87      0.91      0.89       101
         5.0       0.67      0.46      0.55        13

    accuracy                           0.95      3560
   macro avg       0.83      0.80      0.81      3560
weighted avg       0.95      0.95      0.95      3560
```

[38]:
```
print("Classification Report for SVM:")
print(classification_report(y_test, svm_predicted_classes,␣
↪target_names=label_encoder.classes_.astype(str)))
```

```
Classification Report for SVM:
              precision    recall  f1-score   support

         0.0       0.97      0.97      0.97      2750
         1.0       0.73      0.69      0.71       444
         2.0       0.57      0.72      0.63       158
         3.0       0.63      0.56      0.60        94
         4.0       0.86      0.81      0.84       101
         5.0       0.69      0.69      0.69        13
```

```
    accuracy                           0.91      3560
   macro avg       0.74      0.74      0.74      3560
weighted avg       0.91      0.91      0.91      3560
```

[39]:

```
print("Classification Report for ANN:")
print(classification_report(y_test, ann_predicted_classes,␣
 ↪target_names=label_encoder.classes_.astype(str)))
```

```
Classification Report for ANN:
              precision    recall  f1-score   support

         0.0       0.97      0.97      0.97      2750
         1.0       0.74      0.69      0.71       444
         2.0       0.56      0.69      0.62       158
         3.0       0.63      0.47      0.54        94
         4.0       0.85      0.77      0.81       101
         5.0       0.53      0.69      0.60        13

    accuracy                           0.91      3560
   macro avg       0.71      0.71      0.71      3560
weighted avg       0.91      0.91      0.91      3560
```

[40]:

```
print("Classification Report for LSTM:")
print(classification_report(y, lstm_predicted_classes,␣
 ↪target_names=label_encoder.classes_.astype(str)))
```

```
Classification Report for LSTM:
              precision    recall  f1-score   support

         0.0       0.97      0.96      0.96     13557
         1.0       0.61      0.76      0.67      2227
         2.0       0.49      0.28      0.36       865
         3.0       0.60      0.15      0.23       548
         4.0       0.54      0.97      0.69       508
         5.0       0.00      0.00      0.00        94

    accuracy                           0.87     17799
   macro avg       0.53      0.52      0.49     17799
weighted avg       0.87      0.87      0.86     17799

/usr/local/lib/python3.10/dist-packages/sklearn/metrics/_classification.py:1531:
UndefinedMetricWarning: Precision is ill-defined and being set to 0.0 in labels
with no predicted samples. Use `zero_division` parameter to control this
behavior.
  _warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))
/usr/local/lib/python3.10/dist-packages/sklearn/metrics/_classification.py:1531:
UndefinedMetricWarning: Precision is ill-defined and being set to 0.0 in labels
with no predicted samples. Use `zero_division` parameter to control this
```

```
behavior.
  _warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))
/usr/local/lib/python3.10/dist-packages/sklearn/metrics/_classification.py:1531:
UndefinedMetricWarning: Precision is ill-defined and being set to 0.0 in labels
with no predicted samples. Use `zero_division` parameter to control this
behavior.
  _warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))
```

3 Conclusion

3.0.1 Bottom Line: Which Model Performed Best?

Based on the accuracy results, the **Random Forest** and **Gradient Boosting** models performed well and provided stable accuracy. Here is a summary of their performance: - **Logistic Regression**: Logistic Regression showed stable performance across most classes, excelling in interpretability and serving well as a baseline model. However, it struggled slightly with imbalanced classes and complex, non-linear patterns. - **Random Forest**: Achieved competitive accuracy with strong cross-validation scores. It handled noisy data effectively. - **Gradient Boosting**: Slightly outperformed Random Forest due to its sequential learning, focusing on the more difficult samples. - **SVM**: Performed well but showed more variability in accuracy, indicating that it might not be as robust as the ensemble models. - **ANN**: Showed good results, effectively capturing non-linear relationships. However, it had a tendency to struggle with underrepresented classes. - **LSTM**: Demonstrated the ability to model temporal dependencies, but it was prone to overfitting with the available data.

Best Performing Model: Based on the accuracy and cross-validation scores, **Gradient Boosting** and **Random Forest** emerged as the best performers. Gradient Boosting had a slight edge in focusing on difficult-to-classify instances.

3.0.2 Most Important Feature

From the feature importance analysis of Random Forest and Gradient Boosting, the **most important features** influencing the model were:

1. **Pressure**: The atmospheric pressure had the highest impact on the model's decision-making process.
2. **SST (Sea Surface Temperature)**: SST was the second most influential, especially in distinguishing between hurricanes of different intensities.

The rolling averages of SST and pressure also contributed to the model's performance, although not as strongly as the direct pressure and SST values.

3.0.3 Way Forward

1. **Data Augmentation**: Collect additional data to address underrepresented hurricane categories. This could include incorporating more detailed historical weather data or using synthetic data generation.
2. **Feature Engineering**: Explore more advanced feature engineering, such as deriving features that represent storm evolution over time.
3. **Hyperparameter Tuning**: Further optimize the hyperparameters of Gradient Boosting and Random Forest to potentially improve their performance.

4. **Ensemble Methods**: Consider an ensemble approach that combines Random Forest, Gradient Boosting, and ANN to leverage their strengths for improved overall prediction accuracy.
5. **Handling Imbalanced Data**: Use techniques like SMOTE (Synthetic Minority Oversampling Technique) to address class imbalance, especially for higher hurricane categories.
6. **Sequential Modeling**: For LSTM, consider training on a larger dataset with more sequential information to reduce overfitting and better capture temporal dependencies.

The next steps should focus on addressing data imbalances and incorporating new features or data sources that can enhance the prediction capabilities of these models.

[ ]:
```
!jupyter nbconvert --to PDF "Final_Notebook_updated.ipynb"
```

[ ]:
```

```