

Building the text classifier

In this phase, we focus on NLP processing and vectorization to prepare the cleaned data for classification model development. We will start by transforming the text into a format suitable for machine learning algorithms. The key steps involve tokenization, stemming, and feature extraction, optimizing the data for model training. Moving forward, we will employ classification models (conventional ensemble to learn patterns and relationships within the text, creating a robust and accurate predictive system. As the final touch, the trained model is serialized using pickle, ensuring its preservation and easy deployment for future use. This comprehensive approach ensures the creation of a high-performing and deployable text classifier.

In [2]:

```
#importing basic package
import pandas as pd
```

In [3]:

```
# loading the dataset
data = pd.read_csv('filtered_data.csv')
data.head()
```

Out[3]:

	Text	Classification	Filtered_Text
0		i didnt feel humiliated	sadness
1	i can go from feeling so hopeless to so dammed...	sadness	'i can go from feeling so hopeless to so dammed..
2	im grabbing a minute to post i feel greedy wrong	anger	'im grabbing a minute to post i feel greedy wr...
3	i am ever feeling nostalgic about the firepla...	love	'i am ever feeling nostalgic about the firepla...
4	iam feeling grouchy	anger	'iam feeling grouchy'

In [4]:

```
#seeing the type of labels and their count
data['Classification'].value_counts()
```

Out[4]:

joy	5041
sadness	5243
anger	2429
fear	2157
love	1456
surprise	632

Name: Classification, dtype: int64

In [5]:

```
#function to remove commas
def remove_single_quotes(dataset, column_name):
    dataset[column_name] = dataset[column_name].str.replace("'", "")
    return dataset
```

In [6]:

```
#removing unnecessary column
data.drop(['Text'],axis=1)
data.head()
```

Out[6]:

	Classification	Filtered_Text
0	sadness	'i didnt feel humiliated'
1	sadness	'i can go from feeling so hopeless to so dammed...
2	anger	'im grabbing a minute to post i feel greedy wrong
3	love	'i am ever feeling nostalgic about the firepla...
4	anger	'iam feeling grouchy'

In [7]:

```
#cleaning the text
data = remove_single_quotes(data, 'Filtered_Text')
data.head()
```

Out[7]:

	Classification	Filtered_Text
0	sadness	i didnt feel humiliated
1	sadness	i can go from feeling so hopeless to so dammed...
2	anger	im grabbing a minute to post i feel greedy wrong
3	love	i am ever feeling nostalgic about the firepla...
4	anger	iam feeling grouchy

In [8]:

```
# Rearranging the columns so that text --> Labels
data = data[['Filtered_Text', 'Classification']]
data.head()
```

Out[8]:

	Filtered_Text	Classification
0	i didnt feel humiliated	sadness
1	i can go from feeling so hopeless to so dammed...	sadness
2	im grabbing a minute to post i feel greedy wrong	anger
3	i am ever feeling nostalgic about the firepla...	love
4	iam feeling grouchy	anger

Bonus task: Marking incomplete/grammatically incorrect sentences

We will use spaCy package to analyze the syntax of each sentence and check for syntax errors.

Note: This approach relies on pre-trained language models and may not catch all nuances of grammatical correctness.

In [9]:

```
#getting spacy
import spacy
from spacy.tokens import Doc
```

In [10]:

```
# Loading spaCy English language model
nlp = spacy.load("en_core_web_sm")
```

In [11]:

```
# Function to grammatical correctness
def is_grammatically_correct(text):
    doc = nlp(text)

    return all(sent.has_vector for sent in doc.sents)
```

In [12]:

```
# Sample usage to mark records

def segregate_grammatically_correct(data, text_column):
    grammatically_correct = data[data[text_column].apply(is_grammatically_correct)]
    grammatically_incorrect = data[~data[text_column].apply(is_grammatically_correct)]

    return grammatically_correct, grammatically_incorrect
```

In [13]:

```
#Applying the function and segregating
grammatically_correct, grammatically_incorrect = segregate_grammatically_correct(data, 'Filtered_Text')

print("Grammatically Correct:")
print(grammatically_correct)

print("\nGrammatically Incorrect:")
print(grammatically_incorrect)

Grammatically Correct:
0          i didnt feel humiliated          sadness
1  i can go from feeling so hopeless to so dammed...  sadness
2  im grabbing a minute to post i feel greedy wrong    anger
3  i am ever feeling nostalgic about the firepla...    love
4      iam feeling grouchy                          anger
...
17953 i just keep feeling like someone is being unki...    fear
17954 i feel like a little cranky negative after this...    anger
17955 i feel that i am useful to my people and that ...    joy
17956 im feeling more comfortable with derby i feel ...    joy
17957 i feel all weird when i have to meet w people ...    fear

[17958 rows x 2 columns]

Grammatically Incorrect:
Empty DataFrame
Columns: [Filtered_Text, Classification]
Index: []

Remarks: We can say that we have no grammatically incorrect text in the dataset
```

In [14]:

```
#importing data pre-processing packages
import
from sklearn.feature_extraction.text import TfidfVectorizer
from nltk.corpus import stopwords
from nltk.stem import WordNetLemmatizer
from nltk.tokenize import word_tokenize
import re
import nltk
from sklearn.feature_extraction.text import TfidfVectorizer
```

In [15]:

```
# Loading spaCy English language model
nlp = spacy.load("en_core_web_sm")
```

In [16]:

```
# Function for text preprocessing
def preprocess_dataset(dataset, text_column):

    def preprocess_text(text):
        doc = nlp(text)
        tokens = [token.text for token in doc]

        stop_words = set(stopwords.words('english'))
        tokens = [word for word in tokens if word.lower() not in stop_words]

        lemmatizer = WordNetLemmatizer()
        tokens = [lemmatizer.lemmatize(word) for word in tokens]

        return ' '.join(tokens)

    dataset[text_column] = dataset[text_column].apply(preprocess_text)

    return dataset
```

In [17]:

```
# Applying the function
preprocessed_dataset = preprocess_dataset(data, 'Filtered_Text')
preprocessed_dataset.head()
```

Out[17]:

	Filtered_Text	Classification
0	nt feel humiliated	sadness
1	go feeling hopeless dammed hopeful around some...	sadness
2	grabbing minute post feel greedy wrong	anger
3	ever feeling nostalgic fireplace know still pr...	love
4	feeling grouchy	anger

In [18]:

```
# getting the preprocessed text
preprocessed_text = preprocessed_dataset['Filtered_Text']
preprocessed_text.head()
```

Out[18]:

	nt feel humiliated
1	go feeling hopeless dammed hopeful around some...
2	grabbing minute post feel greedy wrong
3	ever feeling nostalgic fireplace know still pr...
4	feeling grouchy

Name: Filtered_Text, dtype: object

In [19]:

```
# Initializing the TF-IDF vectorizer
tfidf_vectorizer = TfidfVectorizer(max_features=5000)
```

In [20]:

```
# Fitting and transforming the preprocessed text
tfidf_matrix = tfidf_vectorizer.fit_transform(preprocessed_text)
```

In [21]:

```
# Converting the TF-IDF matrix to a DataFrame
data_df = pd.DataFrame(tfidf_matrix.toarray(), columns=tfidf_vectorizer.get_feature_names_out())

print(tfidf_df.head())

aa  abandon  abandoned  abandoning  abandonment  abc  abdomen  abide  \
0  0.0      0.0         0.0         0.0         0.0  0.0  0.0  0.0
1  0.0      0.0         0.0         0.0         0.0  0.0  0.0  0.0
2  0.0      0.0         0.0         0.0         0.0  0.0  0.0  0.0
3  0.0      0.0         0.0         0.0         0.0  0.0  0.0  0.0
4  0.0      0.0         0.0         0.0         0.0  0.0  0.0  0.0

ability  abit  ...  youthful  youtube  yuki  zach  Zealand  zero  zombie  \
0  0.0  0.0  0.0  ...      0.0      0.0  0.0  0.0  0.0  0.0  0.0
1  0.0  0.0  ...      0.0      0.0  0.0  0.0  0.0  0.0  0.0
2  0.0  0.0  ...      0.0      0.0  0.0  0.0  0.0  0.0  0.0
3  0.0  0.0  ...      0.0      0.0  0.0  0.0  0.0  0.0  0.0
4  0.0  0.0  ...      0.0      0.0  0.0  0.0  0.0  0.0  0.0

zone  zoom  zumba
0  0.0  0.0  0.0
1  0.0  0.0  0.0
2  0.0  0.0  0.0
3  0.0  0.0  0.0
4  0.0  0.0  0.0

[5 rows x 5000 columns]
```

In [22]:

```
#importing model building/testing tools
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import LabelEncoder
from sklearn.naive_bayes import MultinomialNB
from sklearn.linear_model import LogisticRegression
from sklearn.svm import SVC
from sklearn.ensemble import RandomForestClassifier, GradientBoostingClassifier
from xgboost import XGBClassifier
from sklearn.metrics import accuracy_score, classification_report
```

In [23]:

```
# Append the TF-IDF features to the main dataset
data = pd.concat((data, tfidf_df), axis=1)
data.head()
```

Out[23]:

	Filtered_Text	Classification	aa	abandon	abandoned	abandoning	abandonment	abc	abdomen	abide	...	youthful	youtube	yuki	za
0	nt feel humiliated	sadness	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0
1	go feeling hopeless dammed hopeful around some...	sadness	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0
2	grabbing minute post feel greedy wrong	anger	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0
3	ever feeling nostalgic fireplace know still pr...	love	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0
4	feeling grouchy	anger	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0

5 rows x 5002 columns

In [24]:

```
# Encoding the target variable
label_encoder = LabelEncoder()
data['target_class_encoded'] = label_encoder.fit_transform(data['Classification'])
```

In [25]:

```
# Defining features (X) and target variable (y)
X = data.drop(['target_class_encoded', 'Filtered_Text', 'Classification'], axis=1)
y = data['target_class_encoded']
```

In [26]:

```
# Splitting the dataset into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

In [26]:

```
# Training and evaluating conventional classifiers
classifiers = {
    'Multinomial Naive Bayes': MultinomialNB(),
    'Logistic Regression': LogisticRegression(),
    'Support Vector Classifier': SVC()
}

for name, classifier in classifiers.items():
    classifier.fit(X_train, y_train)

    y_pred = classifier.predict(X_test)

    accuracy = accuracy_score(y_test, y_pred)
    classification_rep = classification_report(y_test, y_pred)

    print(f"{name}:")
    print(f"Accuracy: {accuracy:.4f}")
    print(f"Classification Report:\n", classification_rep)
```

Classifier: Multinomial Naive Bayes
Accuracy: 0.7422
Classification Report:

	precision	recall	f1-score	support
0	0.93	0.56	0.70	474
1	0.87	0.44	0.58	1157
2	0.67	0.97	0.79	1157
3	0.95	0.13	0.23	300
4	0.76	0.84	0.84	1118
5	1.00	0.02	0.04	133

accuracy 0.86
macro avg 0.86 0.51 0.53 3592
weighted avg 0.79 0.74 0.70 3592

C:\Users\ajugy\Documents\anaconda3\lib\site-packages\sklearn\linear_model_logistic.py:444: ConvergenceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Increase the number of iterations (max_iter) or scale the data as shown in:
https://scikit-learn.org/stable/modules/preprocessing.html
Please also refer to the documentation for alternative solver options:
https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression
0.06s UserWarning: ConvergenceWarning: lbfgs failed to converge (status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.

Classifier: Logistic Regression
Accuracy: 0.8641
Classification Report:

	precision	recall	f1-score	support
0	0.91	0.83	0.86	474
1	0.83	0.76	0.79	410
2	0.82	0.95	0.88	1157
3	0.86	0.61	0.71	300
4	0.90	0.95	0.92	1118
5	0.91	0.46	0.61	133

accuracy 0.86
macro avg 0.87 0.76 0.80 3592
weighted avg 0.87 0.86 0.86 3592

Classifier: Support Vector Classifier
Accuracy: 0.8383
Classification Report:

	precision	recall	f1-score	support
0	0.90	0.81	0.85	474
1	0.82	0.76	0.79	410
2	0.81	0.96	0.88	1157
3	0.85	0.58	0.69	300
4	0.91	0.93	0.92	1118
5	0.97	0.50	0.66	133

accuracy 0.86
macro avg 0.88 0.75 0.80 3592
weighted avg 0.86 0.86 0.85 3592

Remark: Of all the three conventional models we tried, the highest accuracy was around 86.4% and of the Logistic Regression Model. We will go ahead and try ensemble learning methods now and will see if the accuracy can be improved or not.

In [27]:

```
# Training and evaluating ensemble classifiers
ensemble_classifiers = {
    'Random Forest': RandomForestClassifier(),
    'Gradient Boosting': GradientBoostingClassifier(),
    'XGBoost': XGBClassifier()
}

for name, classifier in ensemble_classifiers.items():
    classifier.fit(X_train, y_train)

    y_pred = classifier.predict(X_test)

    accuracy = accuracy_score(y_test, y_pred)
    classification_rep = classification_report(y_test, y_pred)

    print(f"{name} Ensemble Classifier: (name)")
    print(f"Accuracy: {accuracy:.4f}")
    print(f"Classification Report:\n", classification_rep)
```

Ensemble Classifier: Random Forest
Accuracy: 0.8767
Classification Report:

	precision	recall	f1-score	support
0	0.88	0.87	0.87	474
1	0.79	0.88	0.83	410
2	0.88	0.91	0.90	1157
3	0.80	0.74	0.77	300
4	0.94	0.90	0.92	1118
5	0.79	0.67	0.73	133

accuracy 0.88
macro avg 0.85 0.83 0.84 3592
weighted avg 0.88 0.88 0.88 3592

Ensemble Classifier: Gradient Boosting
Accuracy: 0.8394
Classification Report:

	precision	recall	f1-score	support
0	0.93	0.78	0.85	474
1	0.87	0.73	0.79	410
2	0.75	0.94	0.83	1157
3	0.81	0.70	0.75	300
4	0.95	0.84	0.89	1118
5	0.75	0.80	0.77	133

accuracy 0.84
macro avg 0.84 0.80 0.82 3592
weighted avg 0.85 0.84 0.84 3592

Ensemble Classifier: XGBoost
Accuracy: 0.8842
Classification Report:

	precision	recall	f1-score	support
0	0.89	0.88	0.89	474
1	0.85	0.86	0.86	410
2	0.87	0.91	0.89	1157
3	0.78	0.83	0.80	300
4	0.96	0.89	0.92	1118
5	0.75	0.79	0.77	133

accuracy 0.88
macro avg 0.85 0.86 0.86 3592
weighted avg 0.89 0.88 0.88 3592

Remarks: The XGBoost classifier gave the highest accuracy till now which stands at 88.4% which is a bit of an improvement from the Logistic classifier. However we must try to perform better and we will move on to hyperparameter tuning to see if we can improve.

In [28]:

```
#getting the packages to build an optimised ensemble model
from sklearn.experimental import enable_halving_search_cv
from sklearn.feature_selection import SelectKBest, chi2
from sklearn.model_selection import GridSearchCV
from tensorflow.keras.optimizers import Adam
from tensorflow.keras.callbacks import EarlyStopping
import tensorflow.keras as callbacks
from sklearn.model_selection import HalvingGridSearchCV
```

In [29]:

```
# Using chi-squared statistic for important feature selection
k_best = 100
selector = SelectKBest(chi2, k=k_best)
X_train_selected = selector.fit_transform(X_train, y_train)
X_test_selected = selector.transform(X_test)
```

In [30]:

```
# Defining the XGBClassifier
xgb_classifier = XGBClassifier()
```

In [37]:

```
# Defining the hyperparameter grid
param_grid = {
    "learning_rate": [0.05, 0.1, 0.2],
    "n_estimators": [100, 150, 200, 250, 300],
    "max_depth": [3, 4, 5],
    "min_child_weight": [0.1, 2, 3],
    "subsample": [0.7, 0.8, 0.9],
    "colsample_bytree": [0.7, 0.8, 0.9],
    "gamma": [0, 0.1, 0.2],
}

# Building the custom search

def custom_grid_search(model, param_grid, X_train, y_train, X_test, y_test, max_iterations=20, improvement_threshold=0.01,
                        best_model=None, best_accuracy=None, no_improvement_count=0):
    xgb_classifier = XGBClassifier()

    grid_search = HalvingGridSearchCV(estimator=xgb_classifier, param_grid=param_grid, scoring='accuracy', cv=2)
    grid_search.fit(X_train, y_train)

    best_params = grid_search.best_params_

    best_model = xgb_classifier.set_params(**best_params)
    best_model.fit(X_train, y_train)

    best_accuracy = accuracy_score(y_test, best_model.predict(X_test))

    print(f"Initial Accuracy: (best_accuracy:.4f)")
    print(f"Best Parameters: (best_params)")

    if best_accuracy >= baseline_accuracy:
        print("Baseline accuracy achieved.")
        return best_model, best_accuracy

    early_stopping = EarlyStopping(monitor='val_accuracy', patience=5, restore_best_weights=True)

    for iteration in range(1, max_iterations):
        current_best_model = xgb_classifier.set_params(**best_params)
        current_best_model.fit(X_train, y_train, early_stopping_rounds=5, eval_set=((X_train, y_train)))

        y_pred = current_best_model.predict(X_test)

        current_accuracy = accuracy_score(y_test, y_pred)

        print(f"Interaction {iteration}/(max_iterations)")
        print(f"Current Accuracy: (current_accuracy:.4f)")
        print(f"Best Parameters: (best_params)")

        if current_accuracy >= baseline_accuracy:
            print("Baseline accuracy achieved.")
            return current_best_model, current_accuracy

        if current_accuracy > best_accuracy + improvement_threshold:
            best_model = current_best_model
            best_accuracy = current_accuracy
            no_improvement_count = 0
        else:
            no_improvement_count += 1

        if no_improvement_count >= 10:
            print(f"Stopping early as no improvement observed for the last {no_improvement_count} iterations.")
            break

    return best_model, best_accuracy
```

In [39]:

```
# Performing the grid search

best_model, best_accuracy = custom_grid_search(xgb_classifier, param_grid, X_train_selected, y_train, X_test_selected, y_test)
print(f"Final Best Model Accuracy: (best_accuracy:.4f)")
```


[illegible]


```

(181) validation_0-mlogloss=0.78181
(182) validation_0-mlogloss=0.78180
(183) validation_0-mlogloss=0.77845
(184) validation_0-mlogloss=0.77682
(185) validation_0-mlogloss=0.77681
(186) validation_0-mlogloss=0.77332
(187) validation_0-mlogloss=0.77158
(188) validation_0-mlogloss=0.77157
(189) validation_0-mlogloss=0.76820
(190) validation_0-mlogloss=0.76649
(191) validation_0-mlogloss=0.76648
(192) validation_0-mlogloss=0.76332
(193) validation_0-mlogloss=0.76153
(194) validation_0-mlogloss=0.76152
(195) validation_0-mlogloss=0.75825
(196) validation_0-mlogloss=0.75660
(197) validation_0-mlogloss=0.75659
(198) validation_0-mlogloss=0.75330
(199) validation_0-mlogloss=0.75166
(200) validation_0-mlogloss=0.75007
(201) validation_0-mlogloss=0.74845
(202) validation_0-mlogloss=0.74684
(203) validation_0-mlogloss=0.74683
(204) validation_0-mlogloss=0.74342
(205) validation_0-mlogloss=0.74217
(206) validation_0-mlogloss=0.74216
(207) validation_0-mlogloss=0.73906
(208) validation_0-mlogloss=0.73749
(209) validation_0-mlogloss=0.73748
(210) validation_0-mlogloss=0.73441
(211) validation_0-mlogloss=0.73292
(212) validation_0-mlogloss=0.73291
(213) validation_0-mlogloss=0.72930
(214) validation_0-mlogloss=0.72844
(215) validation_0-mlogloss=0.72843
(216) validation_0-mlogloss=0.72547
(217) validation_0-mlogloss=0.72398
(218) validation_0-mlogloss=0.72397
(219) validation_0-mlogloss=0.72100
(220) validation_0-mlogloss=0.71951
(221) validation_0-mlogloss=0.71898
(222) validation_0-mlogloss=0.71666
(223) validation_0-mlogloss=0.71525
(224) validation_0-mlogloss=0.71524
(225) validation_0-mlogloss=0.71239
(226) validation_0-mlogloss=0.71090
(227) validation_0-mlogloss=0.71089
(228) validation_0-mlogloss=0.70784
(229) validation_0-mlogloss=0.70670
(230) validation_0-mlogloss=0.70532
(231) validation_0-mlogloss=0.70393
(232) validation_0-mlogloss=0.70266
(233) validation_0-mlogloss=0.70265
(234) validation_0-mlogloss=0.69984
(235) validation_0-mlogloss=0.69846
(236) validation_0-mlogloss=0.69845
(237) validation_0-mlogloss=0.69583
(238) validation_0-mlogloss=0.69446
(239) validation_0-mlogloss=0.69445
(240) validation_0-mlogloss=0.69182
(241) validation_0-mlogloss=0.69047
(242) validation_0-mlogloss=0.69046
(243) validation_0-mlogloss=0.68783
(244) validation_0-mlogloss=0.68650
(245) validation_0-mlogloss=0.68649
(246) validation_0-mlogloss=0.68389
(247) validation_0-mlogloss=0.68263
(248) validation_0-mlogloss=0.68389
(249) validation_0-mlogloss=0.68062

Iteration 0/250: Accuracy: 0.8516
Best Parameters: {'colsample_bytree': 0.8, 'gamma': 0.1, 'learning_rate': 0.1, 'max_depth': 3, 'min_child_weight': 1, 'n_estimators': 250, 'subsample': 0.7}
Current Accuracy: 0.8516
fit method is deprecated for better compatibility with scikit-learn, use 'early_stopping_rounds' in construct or 'set_params' instead.
(0) validation_0-mlogloss=1.75207
(1) validation_0-mlogloss=1.71673
(2) validation_0-mlogloss=1.65883
(3) validation_0-mlogloss=1.65568
(4) validation_0-mlogloss=1.62860
(5) validation_0-mlogloss=1.63517
(6) validation_0-mlogloss=1.61334
(7) validation_0-mlogloss=1.60446
(8) validation_0-mlogloss=1.61093
(9) validation_0-mlogloss=1.62249
(10) validation_0-mlogloss=1.60588
(11) validation_0-mlogloss=1.61699
(12) validation_0-mlogloss=1.61429
(13) validation_0-mlogloss=1.65984
(14) validation_0-mlogloss=1.61239
(15) validation_0-mlogloss=1.63315
(16) validation_0-mlogloss=1.62071
(17) validation_0-mlogloss=1.60889
(18) validation_0-mlogloss=1.63649
(19) validation_0-mlogloss=1.63648
(20) validation_0-mlogloss=1.63576
(21) validation_0-mlogloss=1.63676
(22) validation_0-mlogloss=1.65608
(23) validation_0-mlogloss=1.65606
(24) validation_0-mlogloss=1.69864
(25) validation_0-mlogloss=1.63728
(26) validation_0-mlogloss=1.62856
(27) validation_0-mlogloss=1.61552
(28) validation_0-mlogloss=1.60334
(29) validation_0-mlogloss=1.60333
(30) validation_0-mlogloss=1.61870
(31) validation_0-mlogloss=1.61867
(32) validation_0-mlogloss=1.61873
(33) validation_0-mlogloss=1.62579
(34) validation_0-mlogloss=1.62582
(35) validation_0-mlogloss=1.62581
(36) validation_0-mlogloss=1.64255
(37) validation_0-mlogloss=1.63865
(38) validation_0-mlogloss=1.64032
(39) validation_0-mlogloss=1.62609
(40) validation_0-mlogloss=1.61961
(41) validation_0-mlogloss=1.60898
(42) validation_0-mlogloss=1.60771
(43) validation_0-mlogloss=1.60179
(44) validation_0-mlogloss=1.60620
(45) validation_0-mlogloss=1.61940
(46) validation_0-mlogloss=1.64891
(47) validation_0-mlogloss=1.67946
(48) validation_0-mlogloss=1.71425
(49) validation_0-mlogloss=1.68844
(50) validation_0-mlogloss=1.67683
(51) validation_0-mlogloss=1.67682
(52) validation_0-mlogloss=1.65300
(53) validation_0-mlogloss=1.65301
(54) validation_0-mlogloss=1.64311
(55) validation_0-mlogloss=1.63819
(56) validation_0-mlogloss=1.63516
(57) validation_0-mlogloss=1.62853
(58) validation_0-mlogloss=1.62392
(59) validation_0-mlogloss=1.61921
(60) validation_0-mlogloss=1.61464
(61) validation_0-mlogloss=1.61010
(62) validation_0-mlogloss=1.61537
(63) validation_0-mlogloss=1.61050
(64) validation_0-mlogloss=1.60914
(65) validation_0-mlogloss=1.60289
(66) validation_0-mlogloss=1.60855
(67) validation_0-mlogloss=1.60444
(68) validation_0-mlogloss=1.60741
(69) validation_0-mlogloss=1.60736
(70) validation_0-mlogloss=1.60498
(71) validation_0-mlogloss=1.60497
(72) validation_0-mlogloss=1.60444
(73) validation_0-mlogloss=1.60498
(74) validation_0-mlogloss=1.60593
(75) validation_0-mlogloss=1.60293
(76) validation_0-mlogloss=1.60418
(77) validation_0-mlogloss=1.60417
(78) validation_0-mlogloss=1.64158
(79) validation_0-mlogloss=1.63799
(80) validation_0-mlogloss=1.63698
(81) validation_0-mlogloss=1.63070
(82) validation_0-mlogloss=1.62710
(83) validation_0-mlogloss=1.62552
(84) validation_0-mlogloss=1.62066
(85) validation_0-mlogloss=1.61656
(86) validation_0-mlogloss=1.61311
(87) validation_0-mlogloss=1.60978
(88) validation_0-mlogloss=1.60644
(89) validation_0-mlogloss=1.60366
(90) validation_0-mlogloss=0.99977
(91) validation_0-mlogloss=0.99644
(92) validation_0-mlogloss=0.99261
(93) validation_0-mlogloss=0.98687
(94) validation_0-mlogloss=0.98686
(95) validation_0-mlogloss=0.98984
(96) validation_0-mlogloss=0.98405
(97) validation_0-mlogloss=0.97738
(98) validation_0-mlogloss=0.97405
(99) validation_0-mlogloss=0.97109
(100) validation_0-mlogloss=0.96817
(101) validation_0-mlogloss=0.96503
(102) validation_0-mlogloss=0.96213
(103) validation_0-mlogloss=0.95920
(104) validation_0-mlogloss=0.95609
(105) validation_0-mlogloss=0.95320
(106) validation_0-mlogloss=0.95026
(107) validation_0-mlogloss=0.94732
(108) validation_0-mlogloss=0.94456
(109) validation_0-mlogloss=0.94172
(110) validation_0-mlogloss=0.93895
(111) validation_0-mlogloss=0.93603
(112) validation_0-mlogloss=0.93326
(113) validation_0-mlogloss=0.93046
(114) validation_0-mlogloss=0.92783
(115) validation_0-mlogloss=0.92516
(116) validation_0-mlogloss=0.92255
(117) validation_0-mlogloss=0.91990
(118) validation_0-mlogloss=0.91730
(119) validation_0-mlogloss=0.91470
(120) validation_0-mlogloss=0.91202
(121) validation_0-mlogloss=0.90944
(122) validation_0-mlogloss=0.90681
(123) validation_0-mlogloss=0.90426
(124) validation_0-mlogloss=0.90179
(125) validation_0-mlogloss=0.89925
(126) validation_0-mlogloss=0.89665
(127) validation_0-mlogloss=0.89417
(128) validation_0-mlogloss=0.89171
(129) validation_0-mlogloss=0.88929
(130) validation_0-mlogloss=0.88692
(131) validation_0-mlogloss=0.88450
(132) validation_0-mlogloss=0.88210
(133) validation_0-mlogloss=
```


Note: Despite an exhaustive process of hyperparameter tuning, the marginal improvement in accuracy falls short of the desired threshold. Given the constraints of time and computing resources, we opt to proceed with this model, acknowledging that even incremental enhancements contribute to overall progress. Regrettably, further pursuit of accuracy is deemed impractical within the current limitations.

In [71]: `#getting the pickle package`

`import pickle`

In [72]: `# Saving the model to a file`

`with open('best_xgb_classifier.pkl', 'wb') as file:`
`pickle.dump(best_xgb_classifier, file)`

`print("Model has been pickled.")`

Model has been pickled.