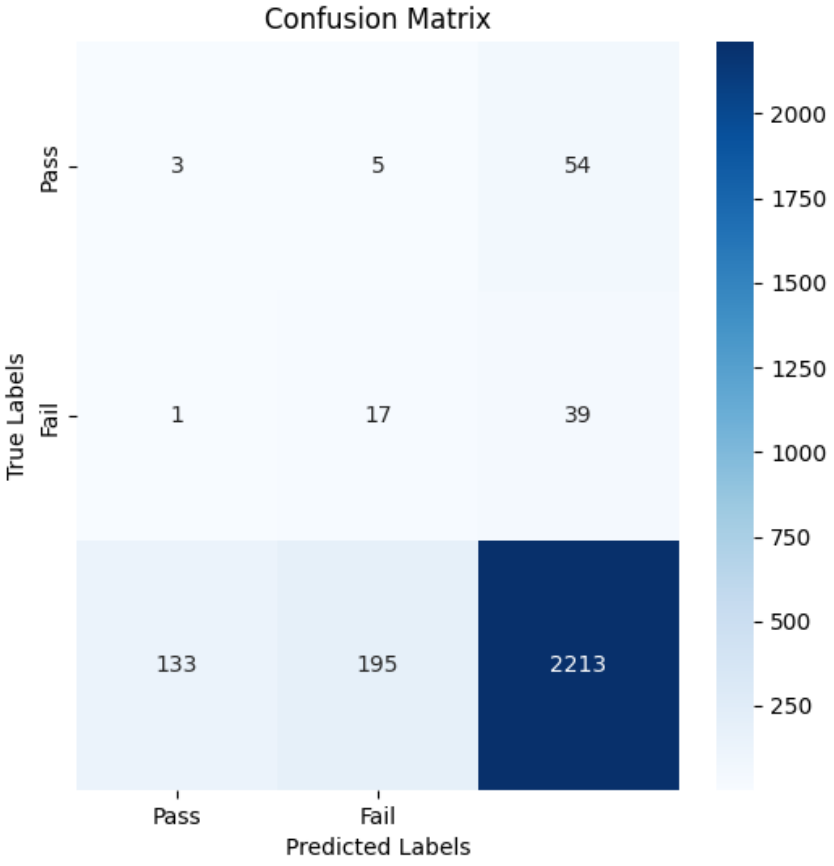
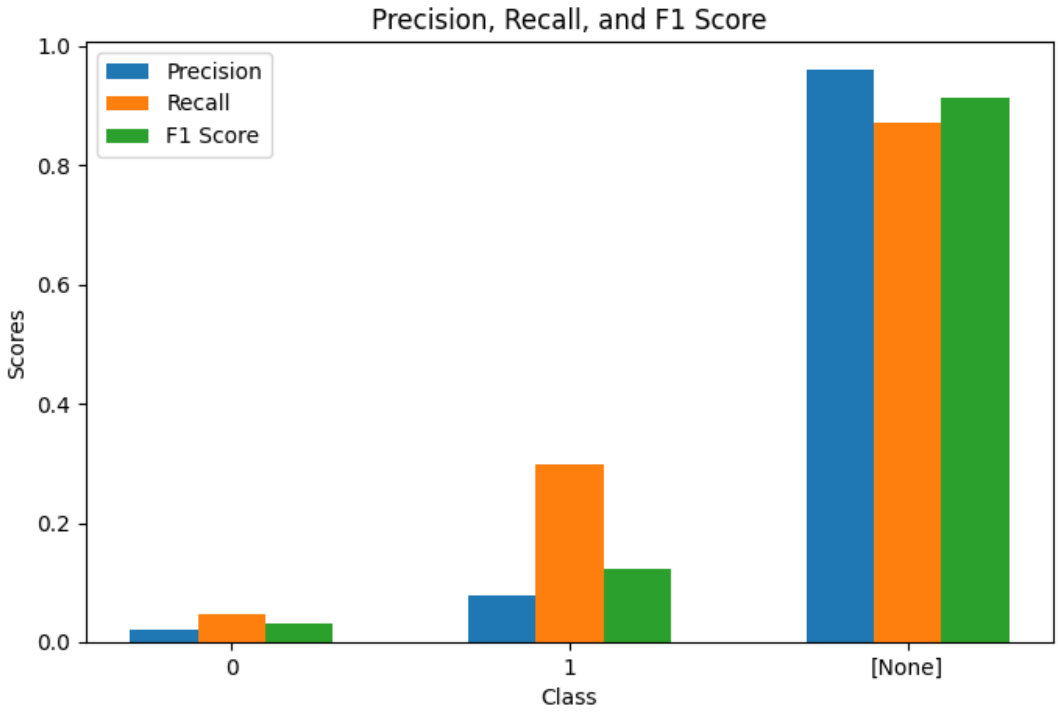
Task 6

I used BeautifulSoup to clean the text data from html entities and to reduce noise and text and remove stopwords used stop_words='english' option in the Tfidf Vectorizer. All text was converted to lowercase to ensure uniformity and avoid duplicating words due to case differences. To transform text into numerical features used TF-IDF and limited maximum number of features to 5000 to focus on the most significant terms. Incorporated Numerical features such as the number of reviews and average rating into the feature set, aiming to provide context beyond just the text. And used Logistic Regression algorithm due to its simplicity, effectiveness in binary classification, and ability to handle multi-class problems and added class_weight='balanced' parameter to adjust for class imbalance, assigning higher weights to underrepresented classes in the training data.
This model achieved the highest Macro F1 Score of 0.3558.




Precision, Recall, and F1 Score
Precision
Recall
F1 Score
Scores
1.0
0.8
0.6
0.4
0.2
0.0
0
1
[None]
Class