AUGNITO

Enhancing Healthcare Intelligence

Assignment Report : Sound Classification

Submitted by:

Name : Pranshu Nema

Mail ID : pranshu.nema@students.iiit.ac.in

Roll No. : 2022202029

Repository Link

Problem Statement :-

Develop a high-performing sound classification system utilizing the provided dataset, crucial for applications like speech recognition and music genre classification. The objective is to achieve accuracy and versatility in categorizing audio signals based on their distinct acoustic characteristics.

EDA on Dataset :-

Exploratory Data Analysis (EDA) is a data analysis approach that involves visually and statistically exploring datasets to reveal patterns, relationships, and insights, aiding in subsequent in-depth analysis**.**

Here's a brief overview of each of the techniques applied in the proposed model:

1. **Type imbalance** - It refers to analyzing the uneven distribution of categories within a dataset, specifically observing the prevalence of various classes in a classification scenario. Identifying type imbalance is essential, as it can significantly affect the effectiveness and accuracy of machine learning algorithms. To check this, I have plotted different graphs like pie and bar. A pie chart shows the percentage distribution of each category, while a bar graph displays the count of samples for each class.

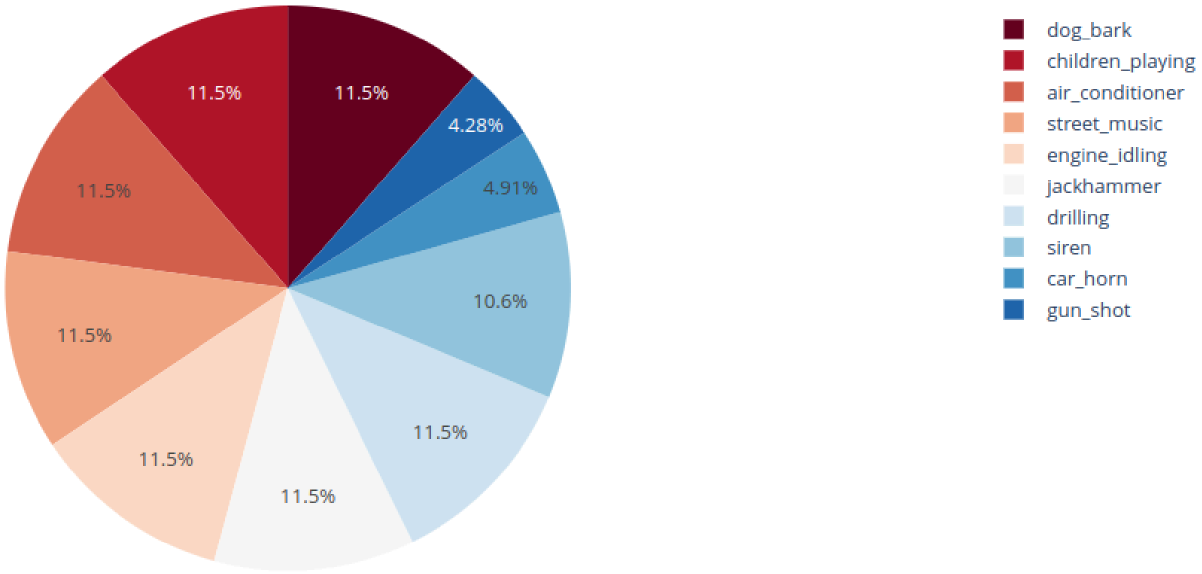


Fig : Pie graph

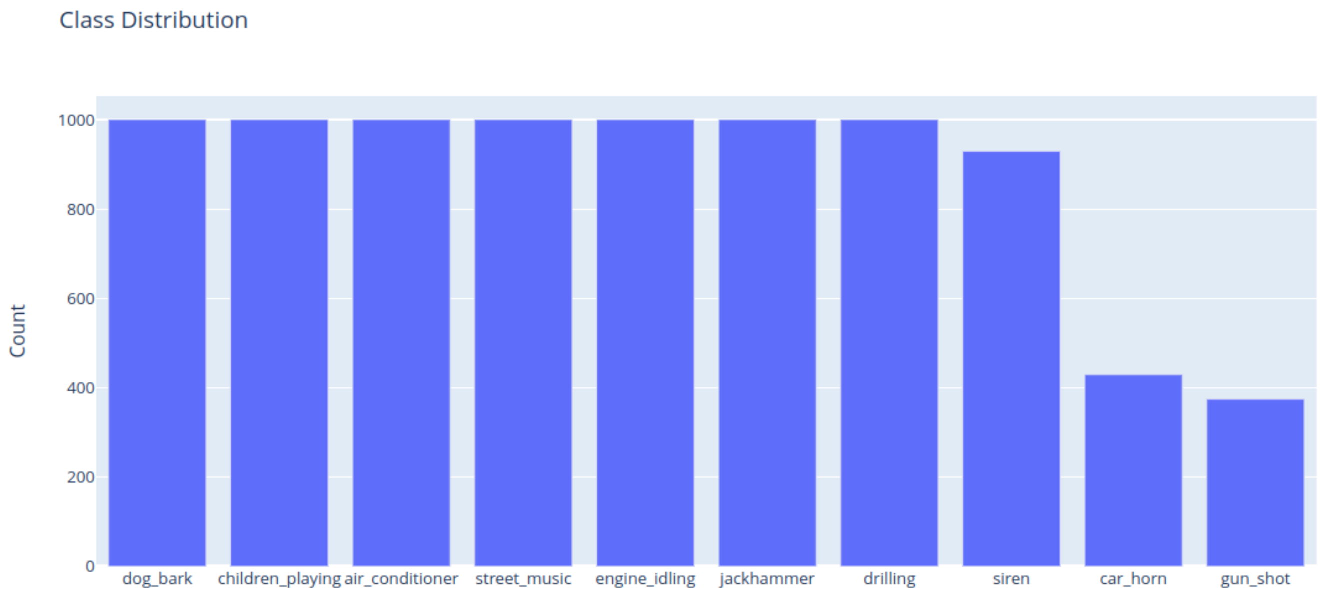


Fig : Bar graph

2. <u>**Plotting waveform**</u> **-** It provides valuable insights into the characteristics of audio signals. Here's why plotting waveforms is beneficial.
   - **Visual Representation of Sound**: Waveforms offer a visual representation of how sound signals vary over time. This visual insight can be intuitive and help to understand the structure of different audio samples.
   - **Identification of Patterns**: By examining waveforms, we can identify patterns, trends, or distinctive features within the audio data. This can be crucial for understanding the characteristics of different sound classes.
   - **Feature Extraction**: EDA on waveforms can guide the selection of relevant features for sound classification. Understanding the distribution of amplitudes and frequencies helps in choosing appropriate features for training machine learning models.

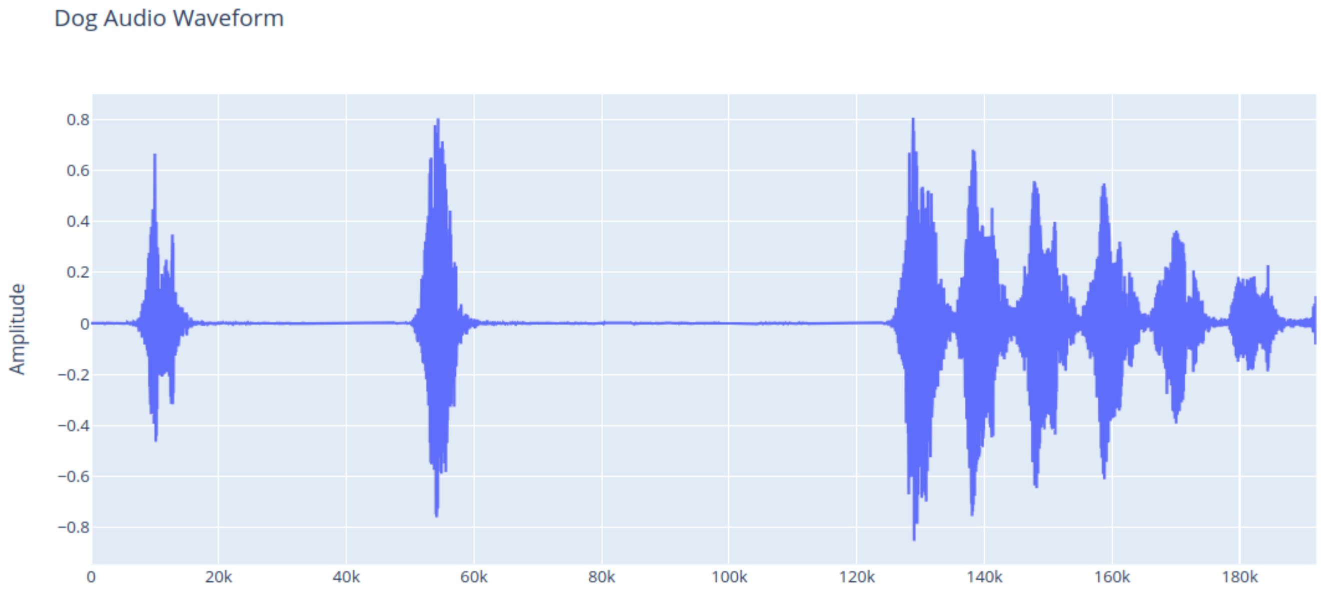


Fig : Waveform of Dogbarking audio

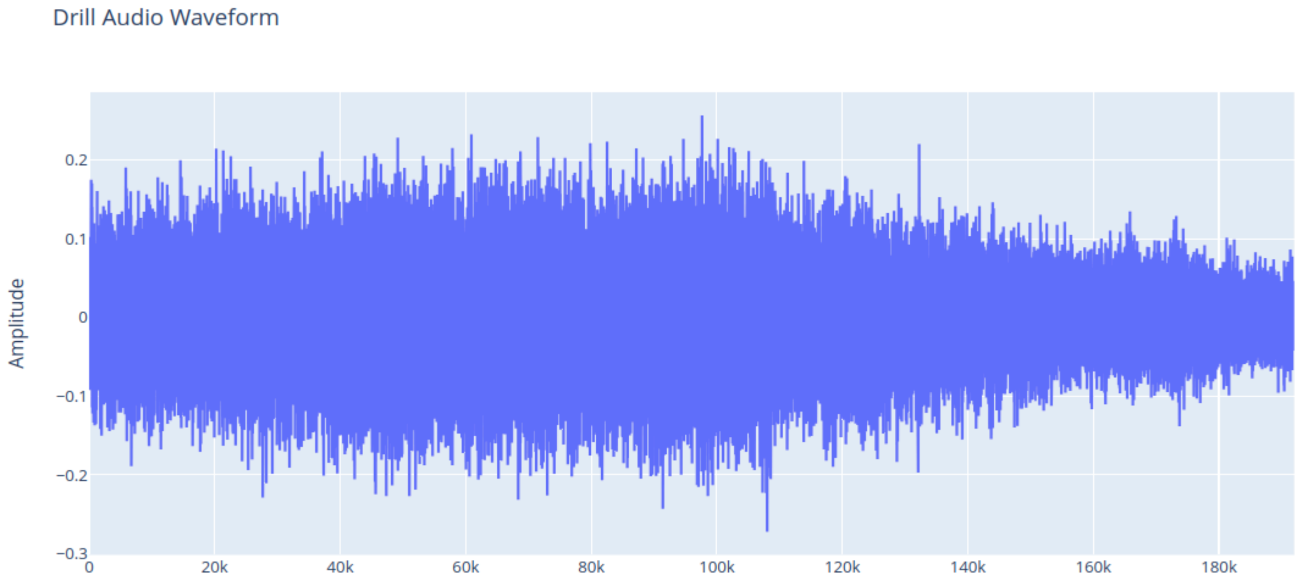


Fig : Waveform of Drill audio

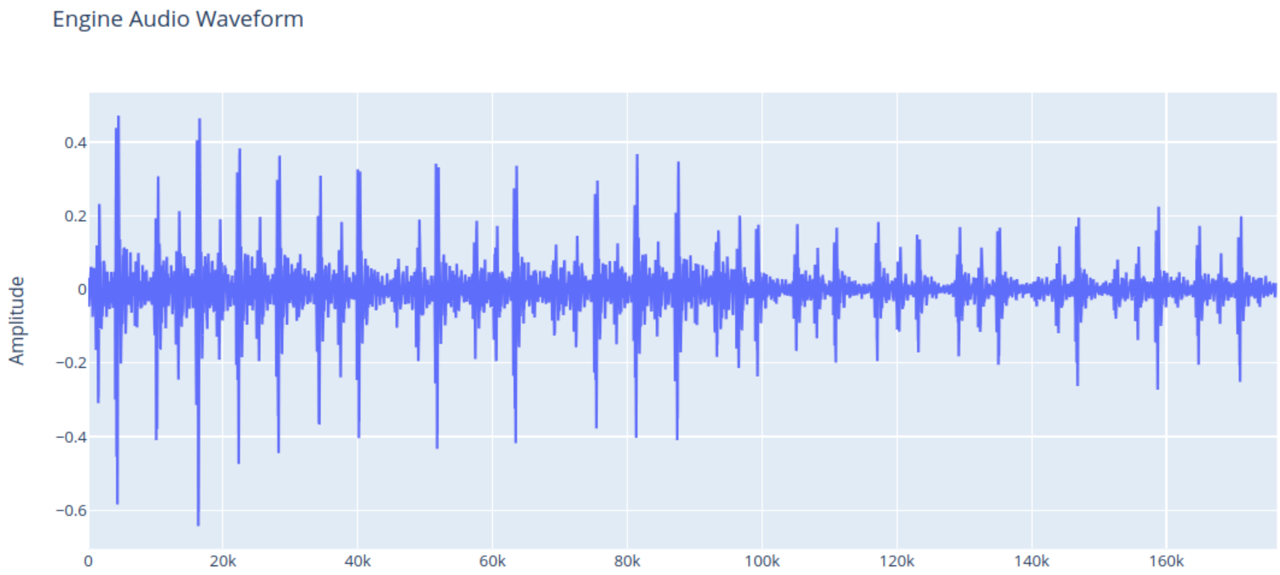
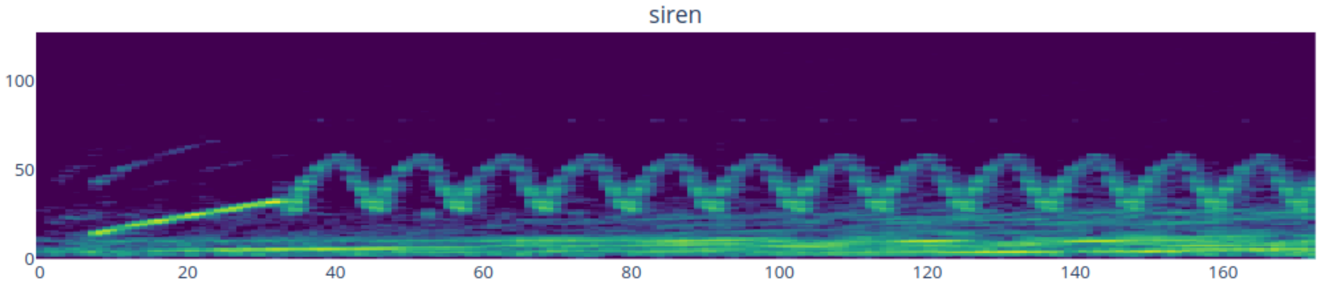
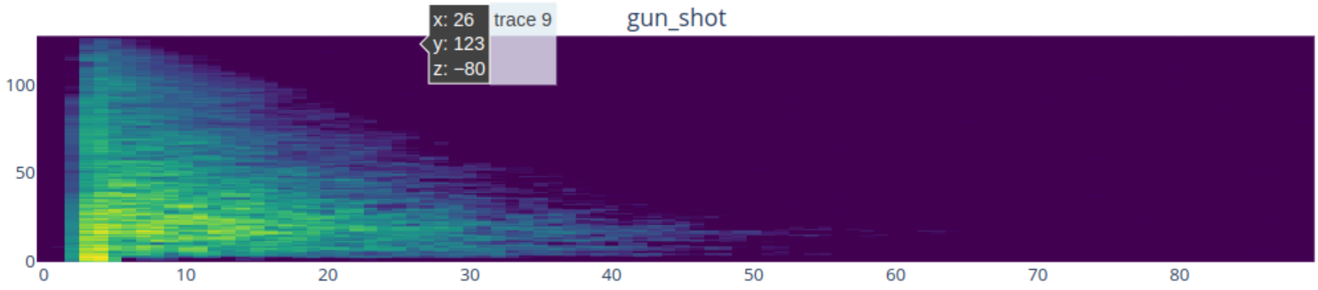
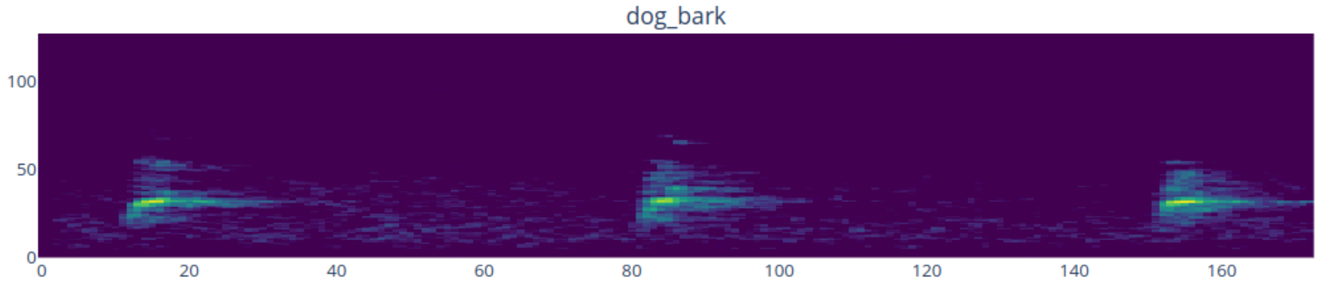


Fig : Waveform of Engine audio

3. **Plotting Spectrogram** : Spectrograms are crucial in Exploratory Data Analysis (EDA) for sound classification due to their ability to provide a detailed and visual representation of the frequency content of audio signals over time.
   - **Data Cleaning and Preprocessing**: Spectrograms can reveal noise, artifacts, or inconsistencies in the data that may require cleaning or preprocessing before model training.
   - **Anomaly Detection** : Unusual patterns or unexpected frequency components in the spectrogram may indicate anomalies or outliers in the audio data, helping in quality control and anomaly detection during EDA.







Approaches :-

Brainstorming on various possible approaches :-

1. **Use of visual approaches** : One of the most common approaches for solving the problem of sound classification problem is to train/fine-tune CNNs or CNN-based models(ResNet) or visual transformers on the spectrogram of the audio. We can have a dense layer with softmax for multiclass classification.
2. Another very simple approach was to use **vision transformers** to generate embeddings of the spectrogram of the audio and then use a simple 3-4 layered Vanilla Neural Network to classify the embedding. In this approach there is a need to train the classifier thus we'll have to split our data into train/test sets.

- **Final Implementation**

  In this approach, I have fine-tuned OpenAI Whisper Model although it's more like a zero-shot approach. Here in this approach, I did the following.

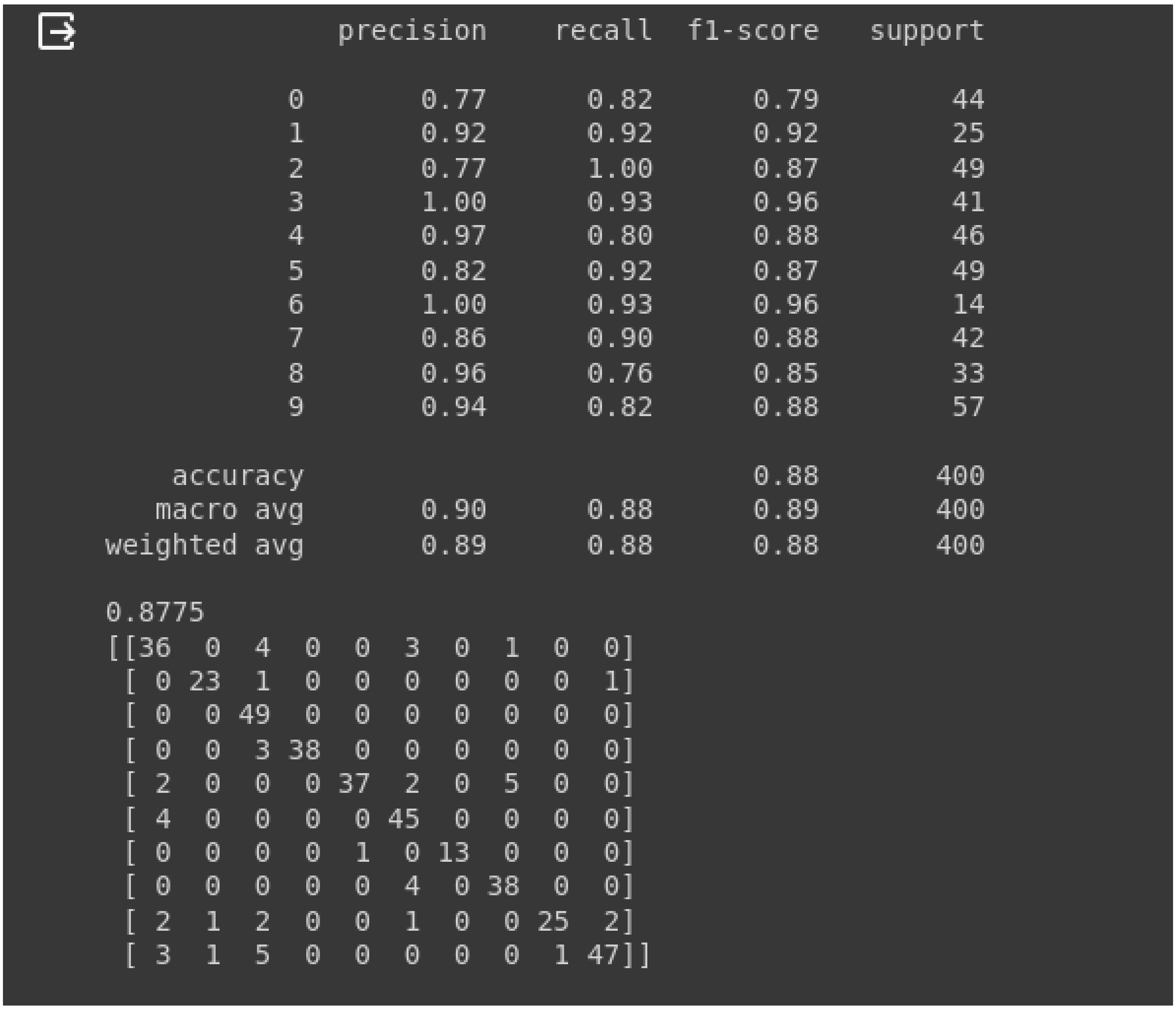
  1. I have used the OpenAI whisper model to extract features from the audio.
  2. Once the audio features are extracted, we have a vanilla neural network with Relu activation to classify the audio in one of the 10 categories.
  3. On a smaller sample of data we have achieved the following results.

<u>Results</u> :-

The model was tried and tested on various hyper params for this model which includes trying different numbers of layers in the vanilla neural network, and layers of different sizes also trying up with various learning rates. The following configuration worked the best are following given below.

Here I have used a random sample of 1000 audio.

A detailed analysis of precision, recall and f1 score for every class is given in the picture below.



```
              precision    recall  f1-score   support

           0       0.77      0.82      0.79        44
           1       0.92      0.92      0.92        25
           2       0.77      1.00      0.87        49
           3       1.00      0.93      0.96        41
           4       0.97      0.80      0.88        46
           5       0.82      0.92      0.87        49
           6       1.00      0.93      0.96        14
           7       0.86      0.90      0.88        42
           8       0.96      0.76      0.85        33
           9       0.94      0.82      0.88        57

    accuracy                           0.88       400
   macro avg       0.90      0.88      0.89       400
weighted avg       0.89      0.88      0.88       400

0.8775
[[36  0  4  0  0  3  0  1  0  0]
 [ 0 23  1  0  0  0  0  0  0  1]
 [ 0  0 49  0  0  0  0  0  0  0]
 [ 0  0  3 38  0  0  0  0  0  0]
 [ 2  0  0  0 37  2  0  5  0  0]
 [ 4  0  0  0  0 45  0  0  0  0]
 [ 0  0  0  0  1  0 13  0  0  0]
 [ 0  0  0  0  0  4  0 38  0  0]
 [ 2  1  2  0  0  1  0  0 25  2]
 [ 3  1  5  0  0  0  0  0  1 47]]
```

Overview of Metrics:

Precision: Indicates the ratio of true positive predictions to the total predicted positives. It answers the question: "Of all the samples we predicted as a certain class, how many were actually that class?"

Recall: Measures the ratio of true positives to the total actual positives. It reflects the model's ability to detect all relevant instances of a given class.

F1-Score: The harmonic mean of precision and recall, providing a balance between them. It's particularly useful when the class distribution is imbalanced.

Support: The number of actual occurrences of the class in the dataset.
Class-wise Analysis:

Class 0: Moderate precision and recall, indicating a balanced detection and prediction capability for this class.
Class 1: High precision but slightly lower recall, suggesting the model predicts this class very accurately when it does, but may miss some instances.
Class 2: Perfect recall but lower precision, the model captures all instances of this class but also has false positives.
Class 3: Perfect precision but lower recall, indicating no false positives but missed detections.
Class 4: High precision and moderate recall, showing good prediction and detection ability.
Class 5: Balanced precision and high recall, indicating good detection with some false positives.
Class 6: Perfect precision and high recall, indicating excellent performance for this class.
Class 7: High precision and recall, demonstrating strong performance.
Class 8: Very high precision with lower recall, suggesting some instances of this class are being missed.
Class 9: High precision with moderate recall, indicating more reliable predictions than detections.

Overall Performance:

- The model has an accuracy of 87.75%, which means that it correctly classified that percentage of the total samples.
- The macro-average precision, recall, and F1-score are all roughly 0.89-0.90, indicating that on average, the model performs consistently across all classes without significant bias toward any particular class.
- The weighted average takes the support into account, providing a performance metric weighted by the number of samples in each class. These are also around 0.88-0.89, suggesting a consistent performance across classes, weighted by their occurrence in the dataset.

Potential Areas of Improvement:

- Classes 0, 2, and 8 have lower precision or recall compared to others, indicating potential areas for improvement. For instance, more training data, feature engineering, or model tuning could be explored to improve these metrics.
- Since Class 6 has a smaller support (only 14 instances), its perfect scores should be interpreted with caution. The model might be overfitting to the few samples it has seen.

Note: Above analysis is on model trained on a smaller sample of the original set i.e 1000 out of 8000.