**Guide to the Onboarding task**

**Motive**

Project Echo represents a commitment to leveraging AI/ML tools for bioacoustics analysis, with the ultimate goal of aiding conservationists in monitoring and protecting biodiversity. By focusing on the development of advanced audio classification models, efficient data pipelines, and innovative recording devices, the project addresses real-world ecological challenges through cutting-edge technology.

The onboarding tasks for Project Echo are carefully designed to support the project’s objectives while fostering a collaborative and inclusive team environment. These tasks begin with dataset insertion, requiring new members to manually insert the dataset into their local machines. This step encourages exploration of the Project Echo repository, which, in its current state, can be somewhat complex. Through this hands-on interaction, participants familiarize themselves with the repository structure, documentation, and available resources, including the dataset itself.

Subsequent tasks focus on dataset exploration, preprocessing, and feature extraction. These activities are essential for understanding the raw audio data that powers Project Echo’s bioacoustics tools. By exploring and processing the dataset, participants develop a deeper understanding of the data's structure and its preparation for machine learning models. Feature extraction further introduces them to techniques for deriving meaningful patterns from audio, directly supporting the project’s goal of achieving high model accuracy. Additionally, a team collaboration exercise fosters open communication and cooperation among members, many of whom may not know each other well. This activity sets the tone for a supportive and interactive team environment.

### **Ideation**

The onboarding tasks were developed with the aim of helping new team members integrate into Project Echo while making the process engaging and easy to follow. The tasks were designed to flow naturally, starting with simple activities such as adding and exploring the dataset and gradually progressing to more advanced steps like preprocessing and feature extraction. This incremental approach helps participants build their understanding step by step, ensuring they are not overwhelmed.

To make the experience hands-on and interactive, the onboarding includes small coding exercises that encourage learning by doing. Instant feedback mechanisms are incorporated, providing participants with confidence as they complete each step. All tasks are directly relevant to the project’s objectives, such as working with audio data and preparing it for AI models. By tying the tasks to the broader goals of Project Echo, participants gain a clear understanding of how their efforts contribute to conservation efforts.

Recognizing the importance of teamwork, the onboarding also incorporates a team-building activity to create a welcoming environment. This exercise emphasizes the value of open communication and shows participants that asking for help is encouraged. The goal is not only to make the onboarding process informative but also enjoyable and inclusive for everyone involved.

**Development Environment Setup**

The first phase of onboarding focuses on setting up the development environment, a critical step in ensuring participants can contribute effectively to Project Echo. Configuring the environment for a project of this scale can be daunting, with challenges like dependency conflicts and version mismatches. To address these, a detailed, step-by-step setup guide was integrated into the onboarding process. This guide provides comprehensive instructions for installing essential tools like Anaconda, creating virtual environments, and configuring Jupyter Notebooks.

A detailed setup guide significantly reduces the frustrations new members might encounter during the initial setup phase. Common issues such as missing dependencies, version conflicts, and improper configurations can delay progress and demotivate contributors. Providing clear and comprehensive instructions enables new team members to quickly install the required tools and begin contributing without being hindered by technical roadblocks. This proactive approach ensures a smooth and efficient transition into the project.

A standardized development environment is critical for maintaining consistency across team workflows and ensuring reproducibility in scientific tasks. The guide specifies precise versions of tools and libraries, such as Python 3.9, TensorFlow 2.10, and audio processing utilities, eliminating discrepancies that may arise due to variations in individual setups. This uniformity is essential for achieving reliable and reproducible results, particularly in tasks like model training and testing where accuracy is of utmost importance.



**Familiarizing Members with Key Tools and Practices**

In addition to facilitating technical setup, the guide familiarizes team members with essential tools and best practices used in Project Echo. It introduces libraries such as librosa for audio processing, TensorFlow for machine learning, and Git for version control. Through hands-on exposure, new contributors gain an understanding of how these tools integrate into the project’s broader goals. The guide also emphasizes the importance of practices such as creating virtual environments to isolate dependencies, managing pinned library versions for compatibility, and organizing directories for datasets and scripts. These elements not only streamline the project’s workflows but also equip team members with valuable skills applicable to future endeavours.

Structured to align with the philosophy of gradual learning, the guide breaks down the setup process into manageable steps. New members progress from basic installations to more advanced tasks, such as configuring virtual environments, installing dependencies, and running test scripts. This approach ensures accessibility for all skill levels, fostering confidence and capability in using the tools required for the project.

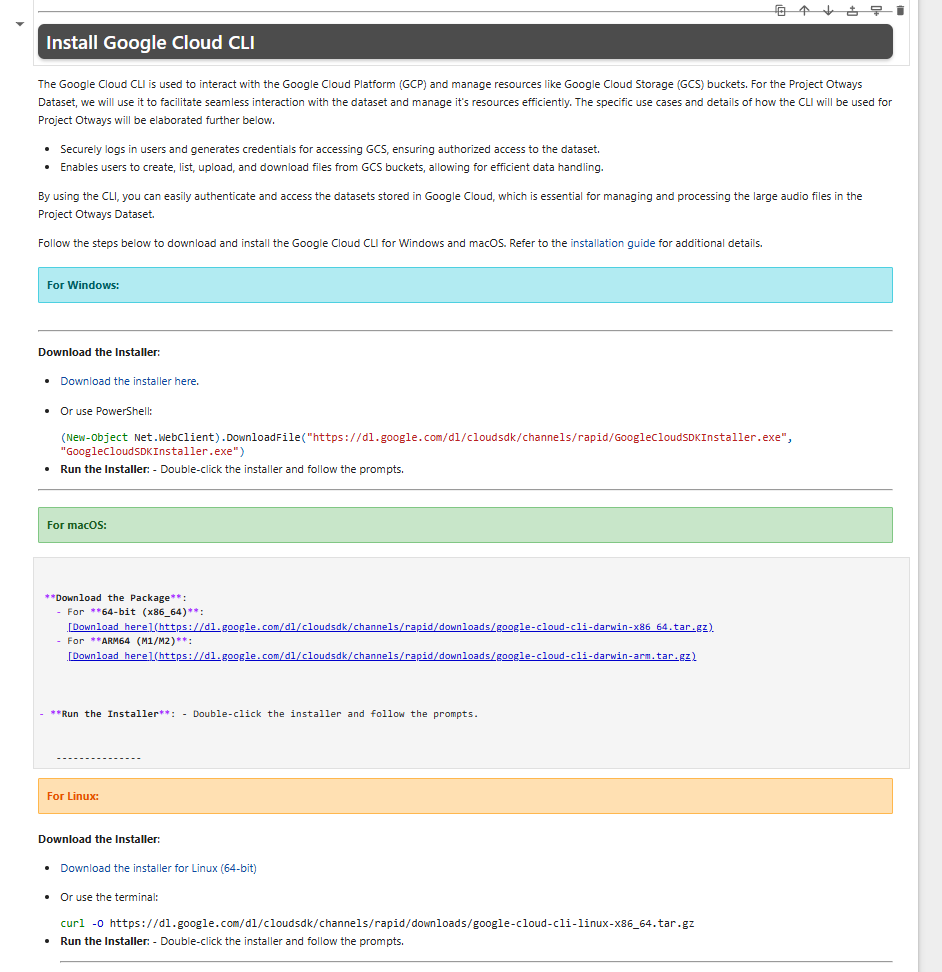
**Dataset Access and Repository Navigation in Project Echo**

The next phase of onboarding focuses on dataset access and repository navigation. The GitHub repository included the following resources:

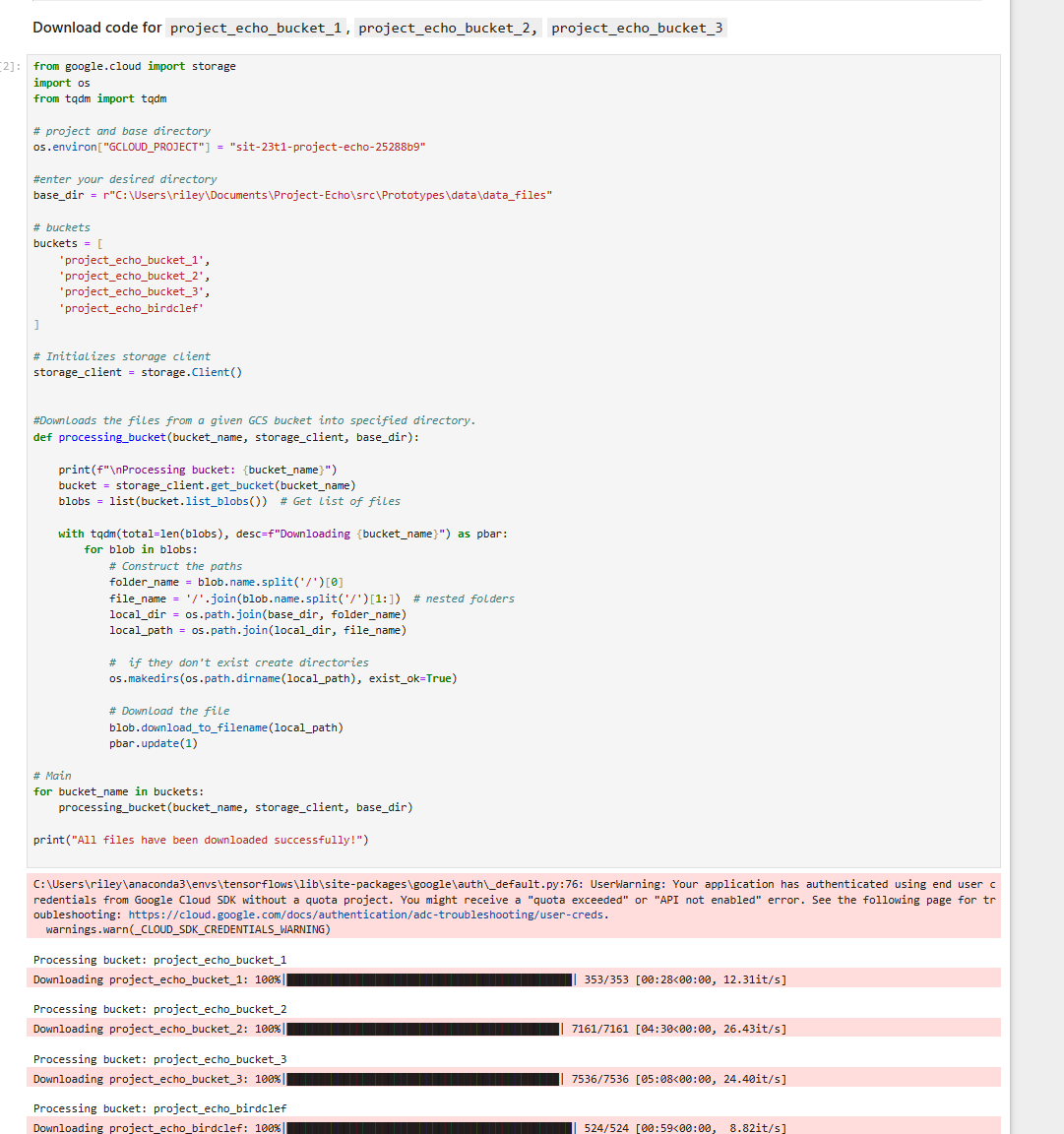
* **GoogleCloud\_download Notebook**: A Jupyter notebook for accessing and downloading datasets from Google Cloud Storage. Provides details on storage
* buckets, authentication, local directory setup, and data transfers.
* **README.md**: Documentation for the Project Otways Dataset, setup instructions for tools and authentication, dataset download steps.

The original onboarding process lacked both thes GoogleCloud\_download notebook and the README.md file, which are critical components of the repository. Without these resources, new users were left without any clear guidance on accessing datasets or understanding the project's structure and purpose. This makes it challenging to set up the environment, navigate the data, or effectively use the available tools. This was not user intuitive at all due to the lack of how to work with the data or understand the purpose of the files in Project Echo. A main goal for our updated onboarding task was to consolidate and improve this information.

**Enhanced Documentation for Google Cloud CLI Installation and Authentication**

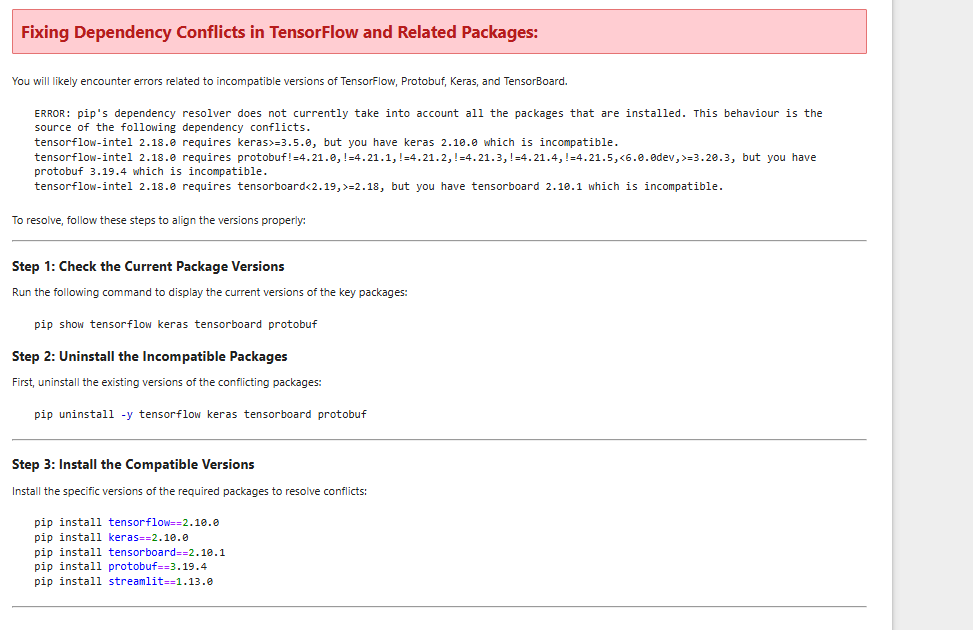
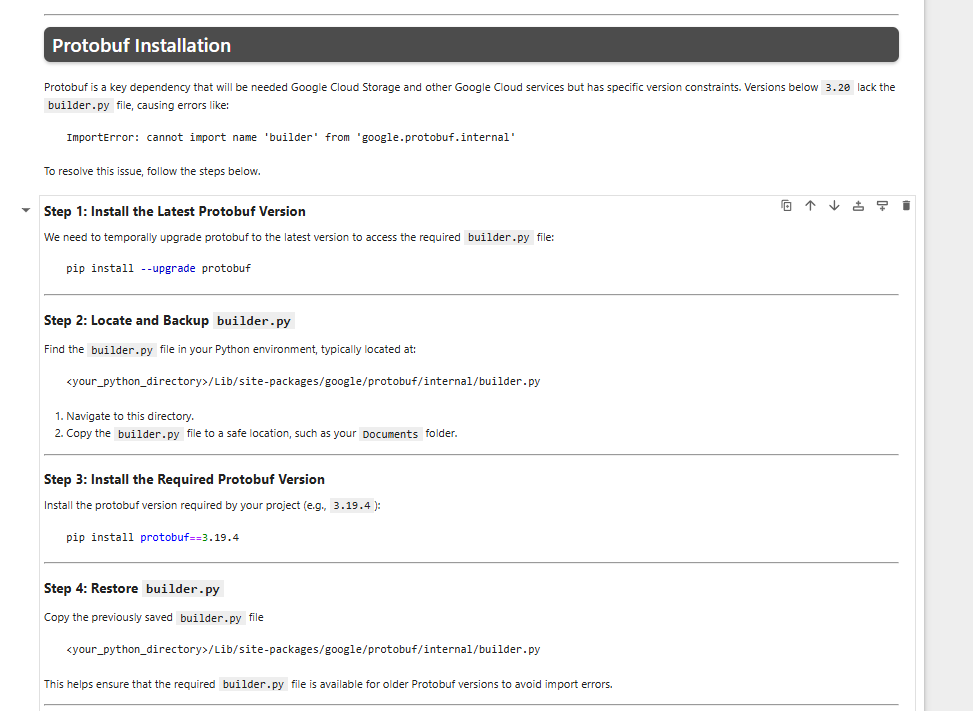
Recognizing that new students may come from diverse technical backgrounds, the guide includes platform-specific instructions for installing and authenticating the Google Cloud CLI on Linux, macOS, and Windows, significantly improving the setup process by making it more accessible and user-friendly.

**Streamlined GCS File Download with Progress Tracking**

A streamlined GCS file download script was also implemented to organize files into a unified local directory structure. The code downloads files from multiple Google Cloud Storage (GCS) buckets into a unified local directory structure (for easy access and efficiency), organizing nested files and showing a progress bar for each bucket, making it easier to be able to manage and track the download process efficiently.

**Troubleshooting Guide for Protobuf and TensorFlow Dependency Issues**

Troubleshooting common issues, such as TensorFlow-Protobuf dependency conflicts, is also addressed through a dedicated section in the guide, providing step-by-step solutions to significantly improving troubleshooting efficiency for users. This reduces frustrations during setup and ensures a smoother onboarding experience.



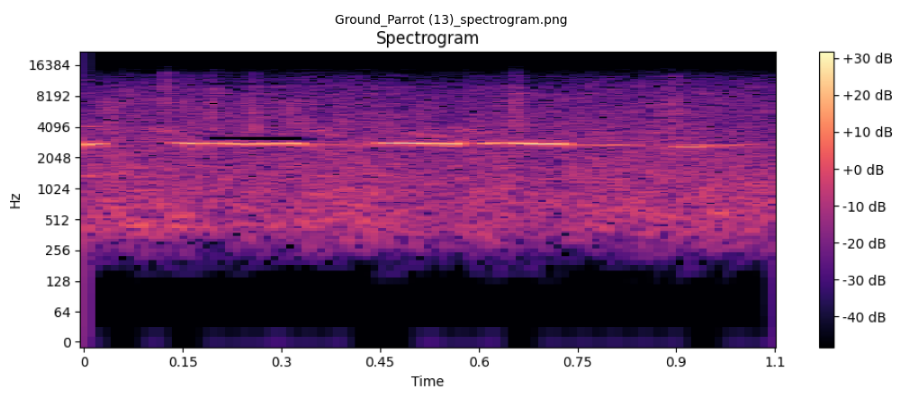
**Defining the Project Otways Dataset Structure and Purpose**

To help students understand the dataset’s importance, a new section clearly defines the Project Otways Dataset and its relevance to Project Echo. This section outlines the contents of each storage bucket, explaining how the data supports the project’s goals. By providing this context, students can better appreciate the significance of the dataset and how it informs their work.



**Implementation**  
  
**Dataset implementation**

The third phase involves exploring and preprocessing the dataset, providing students with practical experience in handling raw data. This section of the onboarding task begins with students downloading the dataset and organizing it on their local machines. This step not only familiarizes them with the Project Echo repository but also encourages them to explore its structure and available resources.

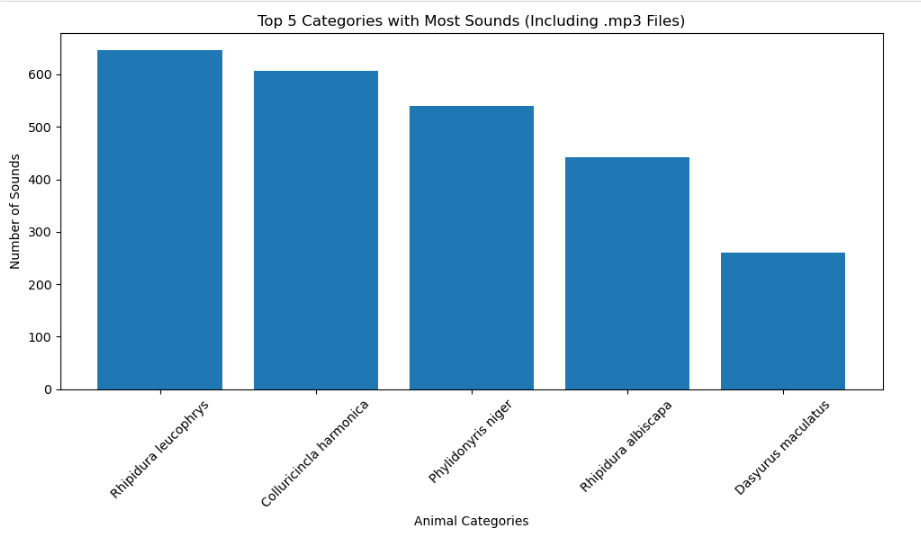
To accommodate low-powered machines and ensure a seamless experience, only 50 random audio files from the dataset are processed. These files are then converted into spectrograms and saved as PNG images. Upon completion, students are prompted to view 10 of the generated spectrograms. This visual feedback enables them to observe the outcomes of their work, reinforcing their understanding of the conversion process and building confidence for future tasks.

*(e.g. of the spectrogram created)*

**Dataset Exploration**

The next step introduces students to dataset exploration; They are tasked with identifying the top five categories with the most sound files. A small code snippet is provided, requiring students to specify the file types (e.g., .wav and .mp3). This approach encourages active engagement with the notebook and ensures they thoroughly interact with the content rather than rushing through it.

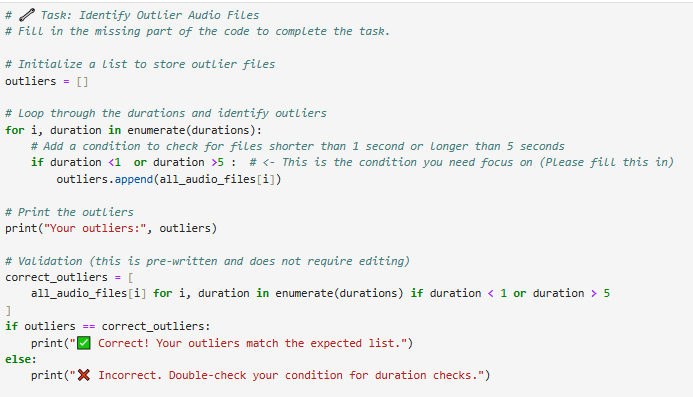
Upon completing the task, the results are displayed in a bar chart. This visualization helps students gain a clearer understanding of the dataset's sound distribution and highlights the importance of data exploration in shaping the project's direction.

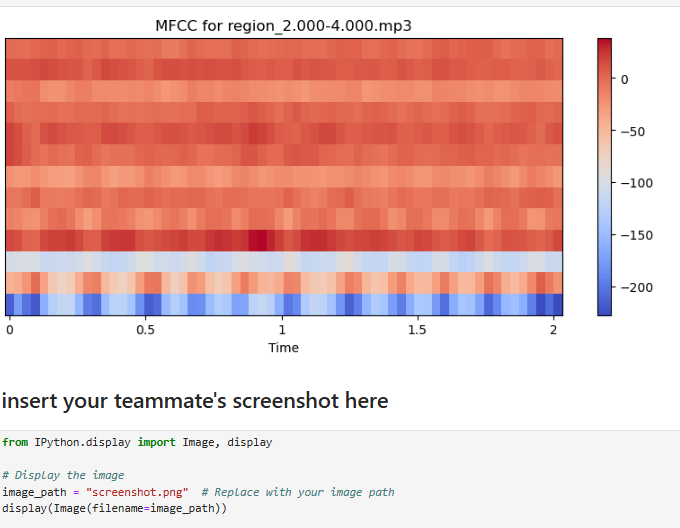
  
*(e.g. of the bar graph)*

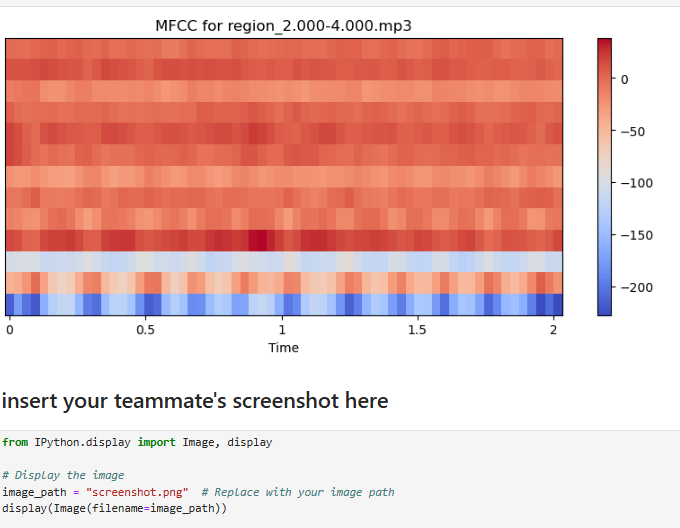
**Data Preprocessing**

In this step, students are tasked with identifying audio files with unusual durations. While real-world preprocessing often involves multiple criteria for detecting outliers, this task focuses solely on duration for simplicity. A straightforward condition is provided to flag files shorter than one second or longer than five seconds as outliers.

This task offers a clear introduction to the importance of data cleaning in machine learning workflows. By looping through file durations and identifying anomalies, students gain practical experience in handling irregularities within a dataset, building their understanding of essential preprocessing techniques.

  
  
**Team Collaboration**

The final step of the onboarding task emphasizes teamwork and collaboration. Students are required to run a program to generate and visualize Mel Frequency Cepstral Coefficients (MFCCs) for a random audio file. To promote collaboration, students are encouraged to share a screenshot of their teammate’s MFCC heatmap. This simple yet effective activity helps break the ice and fosters an environment where team members feel comfortable communicating and supporting one another.



The onboarding process for Project Echo is thoughtfully designed to provide new team members with the resources, knowledge, and confidence they need to contribute meaningfully. It guides them through setting up the development environment, downloading and exploring datasets, and engaging in preprocessing and implementation tasks, with each phase carefully aligned to the project’s overarching goals.

By integrating clear troubleshooting guides, accessible documentation, and collaborative activities, the onboarding experience fosters a welcoming and supportive team culture. Each step builds progressively on the previous one, creating a cohesive and engaging pathway that encourages teamwork and open communication. This structured and inclusive approach ensures that students feel empowered to make impactful contributions to Project Echo.