1. **Two paragraph summary of the project as implemented, including the main customer need and how the application meets it, including who the stakeholders are. This will contrast to what you wrote in Iteration 0.**

Our customer is a research group in the A&M College of Education and Human Development. Our primary point of contact is Dr. Beverly Irby and we have also worked closely with other members of the Center for Research & Development in Dual Language & Literacy Acquisition. This research group is working to improve teaching in ESL (English as a Second Language) classrooms. Towards this goal, they have collected videos of ESL classes and manually coded each video with information such as the language of instruction, mode of communication, and curriculum area. The customer's need was for us to automate some part of the classification process using machine learning.

We worked on building a machine learning classifier model for the communication mode attribute, using the audio from the videos. We limited our analysis to 2 classes: aural versus aural/verbal and verbal/aural (those 2 communications modes are combined into 1 class for classification). These 3 communication modes encompass most of the data. We built a data loading pipeline to get the data ready for machine learning, and we tried using several different ML pipelines for binary classification. The best-performing method we tried was a random forest classifier using domain-specific features. This method gave us 71.7% accuracy on test data.

2. **Description of all user stories (including revised/refactored stories in the case of legacy projects). For each story, explain how many points you gave it, explain the implementation status, including those that did not get implemented. Discuss changes to each story as they went. Show lo-fi UI mockups/storyboards you created and then the corresponding screen shots, as needed to explain stories.**

We had no UI-related mockups, since the customer asked us to focus on the machine learning aspects rather than spending time on building a UI. No screenshots are included since our project is primarily Python scripts and screenshots would not be useful.

As a researcher, I want to extract the audio from video files and split the audio into 20 second clips, so that it can be used in a ML model (1 point). Completed in Iteration 1.

As a researcher, I want to combine all the csv files and load them into a dataframe, so that they are ready for a ML model (1 point). Completed in Iteration 1.

As a researcher, I want to add the audio to a dataframe, so that it is ready for the ML model (1 point). Completed in Iteration 2.

As a researcher, I want to load all the spreadsheets and audio clips into a single data structure, so that I can easily work with the dataset (3 points). Completed in Iteration 2.

As a researcher, I want to run scripts on each video to split it, so that we have pre-processed data (1 point). Completed in Iteration 2.

As a researcher, I want to discuss our current ideas with the customer, to ensure they get what they want (1 point). Completed in Iteration 1.

As a researcher, I want to filter out rows with irrelevant communication mode values, so that I can work with only the relevant data (1 point). Completed in Iteration 2.

As a data scientist, I want to be able to load only part of the audio data, so that I can run the data analysis on a machine without much memory (1 point). Completed in Iteration 3.

As a data scientist, I want to try using a PCA model with logistic regression, so that I can get good results on unseen test data (1 point). Completed in Iteration 3.

As a researcher, I want to extract features from audio files like mfcc and energy, so that we can use them to train the model (1 point). Completed in Iteration 3.

As a researcher, I want to try using a Random Forest model, so that I can get good results on unseen test data (2 points). Completed in Release.

As a researcher, I want to try using the Scikitt neural network, so that I can get good results on unseen test data (2 points). Completed in Release.

As a researcher, I want to investigate audio wave patterns, so that I can use trends for the ML model (1 point). Completed in Iteration 3.

As a data scientist, I want to build an accurate model to predict a category, so that I can automate description coding (3 points). Completed in Release.

As a data scientist, I want to try the Scikitt KNN model, so that I can get good results on unseen data (2 points). Completed in Release.

As a data scientist, I want to try the Scikitt SVM model, so that I can get good results on unseen data (2 points). Completed in Release.

As a data scientist, I want to combine the arrays and labels for all videos, so that I can get the overall accuracy on the entire test dataset (3 points). Completed in Release.

3. **For legacy projects, include a discussion of the process for understanding the existing code, and what refactoring/modification was performed on the code, in addition to the user stories listed above.**
   This is not a legacy project.

4. **List who held each team role, e.g. Scrum Master, Product Owner. Describe any changes in roles during the project.**

Scrum Master: Mehul Varma
Product Owner: Evelyn Crowe

5. **For each scrum iteration, summarize what was accomplished and points completed.**
   Iteration 0: In iteration 0, we created our user stories and worked on completing the training.
   Iteration 1:
   a. As a researcher, I want to extract the audio from video files and split the audio into 20 second clips, so that it can be used in a ML model (1 point).
   b. As a researcher, I want to combine all the csv files and load them into a dataframe, so that they are ready for a ML model (1 point).
   c. As a researcher, I want to discuss our current ideas with the customer, to ensure they get what they want (1 point).
   d. As a researcher, I want to combine all the csv files and load them into a dataframe, so that they are ready for a ML model (1 point).
   e. As a researcher, I want to ruAs a researcher, I want to add the audio to a dataframe, so that it is ready for the ML model (1 point).
   f. As a researcher, I want to run scripts on each video to split it, so that we have pre-processed data(1 point).
   Iteration 2:
   a. As a researcher, I want to add the audio to a dataframe, so that it is ready for the ML model (1 point).
   b. As a researcher, I want to load all the spreadsheets and audio clips into a single data structure, so that I can easily work with the dataset (3 points).
   c. As a researcher, I want to filter out rows with irrelevant communication mode values, so that I can work with only the relevant data (1 point).
   d. As a researcher, I want to run scripts on each video to split it, so that we have pre-processed data (1 point).
   Iteration 3:
   a. As a researcher, I want to investigate audio wave patterns, so that I can use trends for the ML model (1 point).
   b. As a data scientist, I want to be able to load only part of the audio data, so that I can run the data analysis on a machine without much memory (1 point).
   c. As a researcher, I want to filter out rows with irrelevant communication mode values, so that I can work with only the relevant data (1 point).
   d. As a data scientist, I want to try using a PCA model with logistic regression, so that I can get good results on unseen test data (1 point).
   e. As a researcher, I want to extract features from audio files like mfcc and energy, so that we can use them to train a model (1 point).

Release:

a. As a researcher, I want to try using a Random Forest model, so that I can get good results on unseen test data (2 points).
b. As a researcher, I want to try using the Scikitt neural network, so that I can get good results on unseen test data (2 points).
c. As a researcher, I want to investigate audio wave patterns, so that I can use trends for the ML model (1 point).
d. As a data scientist, I want to build an accurate model to predict a category, so that I can automate description coding (3 points).
e. As a data scientist, I want to try the Scikitt KNN model, so that I can get good results on unseen data (2 points).
f. As a data scientist, I want to try the Scikitt SVM model, so that I can get good results on unseen data (2 points).
g. As a data scientist, I want to combine the arrays and labels for all videos, so that I can get the overall accuracy on the entire test dataset (3 points).

**6. List of customer meeting dates, and description of what happened at the meetings, e.g. what software/stories did you demo.**

a) Customer Meeting 1 - 10/25/2021:
This first meeting was before we constructed any user stories. The customers told us what they expected of us, where they wanted us to create a machine learning model to help them fill out a TBOP(Transitional Bilingual Observation Protocol) worksheet based on a video. They asked us to pick one column to focus on.

b) Customer Meeting 2 - 11/03/2021:
At this time we had not received the data we wanted to see before committing to a category. We spoke about doing a data pipeline instead of finding a category because we were concerned about the time restriction of finishing this. We decided to do the data pipeline while also building a model to learn when teacher talking versus student talking.

c) Customer Meeting 3 - 11/08/2021:
We were working on our data pipeline and we decided to make our project where the student is talking or the teacher is talking. We said we would try to get our model's accuracy at 80%, but this is a hard amount to reach. Our achievable goal is an accuracy greater than 50% and our reach goal is an accuracy greater than 80%.

d) Customer Meeting 4 - 11/15/2021:
We spoke about how we wrote code to merge files into dataframe, found issues in currently-cached audio split files, and found some source code for audio data analysis. We told them our next step was to create the model.

e) Customer Meeting 5 - 11/22/2021:
We decided to use three of the different codings in the communication column because these took up a majority of the data in the TBOP worksheets. We let Dr. Tang knew we wanted to use binary codings for the worksheet because it makes it easier to code.

f) Customer Meeting 6 - 12/06/2021:
We updated the customer that our model's score was around 60% and our next task was to increase the accuracy of these models.

7. **Explain your BDD/TDD process, and any benefits/problems from it.**
Since our project was machine-learning based instead of web development, we did not use a BDD/TDD approach. Instead, we performed data cleaning and formatting and then tested different models, but our testing approach consisted mostly of trying a model, tweaking a few parameters or training sets, and accepting the results and moving on to different models. The benefits were that we were able to move fairly quickly and sample a wide range of models without sinking too much time into any that weren't working out. The main problem was that as a team, we lacked experience in machine learning, and were sometimes unable to troubleshoot issues or unexpectedly low accuracies.

8. **Discuss your configuration management approach. Did you need to do any spikes? How many branches and releases did you have?**
For configuration management, you only need Python 3 installed to run the Pip packages: Pandas, Numpy, sklearn, pydub. We did not have a legacy project, so we only have a master branch, where we put all the information needed.

9. **Discuss any issues you had in the production release process to Heroku.**
As our project was not web-based, we did not use Heroku.

10. **Describe any issues you had using AWS Cloud9 and GitHub and other tools.**
We did not use Cloud9 since our codebase was mostly Python scripts that were easier to run locally. All team members had experience with GitHub and we did not experience any difficulties. We found PivotalTracker unintuitive and more difficult to use than Jira, Trello, or other boards meant for Scrum purposes, but we experienced no significant issues aside from a slight learning curve.

11. **Describe the other tools/GEMs you used, such as CodeClimate, or SimpleCov, and their benefits.**
MoviePy: We used this library to split the videos into 20 second increments.

Numpy: We used Numpy to hold the audio arrays.
PyDub: We used the AudioSegment from Pydub to write the audio files into .wav files. We also used it to convert the audio files into 2D Numpy arrays.
Sklearn: We used the ML models from SKlearn, such as the logistic regression, RandomForest, neural networks, and SVM in order to build different models.
Pandas: The Pandas library was used to create a dataframe.

12. **Make sure all code (including Cucumber and RSpec!) is pushed to your public GitHub repo.**
    All code is pushed to our public GitHub repo. Since our project was ML-based, we do not have Cucumber/RSpec or a similar testing framework.

13. **Make a separate section discussing your repo contents and the process, scripts, etc., you use to deploy your code. Make very sure that everything you need to deploy your code is in the repo. We have had problems with legacy projects missing libraries. We will verify that everything is in the repo.**

Our code does not have a traditional deployment process since it is not web-based. Nearly everything of technical interest is in the folder data-cleaning-python. The folder data-cleaning is filled with older scripts, mostly from iteration 2. The documentation folder contains reports and usage information.

The rough process to run our models begins with running splitfiles.py, which takes the path to the samples as an argument, cleans the data, and splits it into twenty-second audio segments.

Then you need to run the analysis.py, which combines the code and videos from different files. In order to run this you need to type python.py analysis.py <directory>, where directory is the location of the folder that contains the videos and codes you want to combine.

14. ***Links to your Pivotal Tracker, public GitHub repo, and Heroku deployment, as appropriate. Make sure these are up-to-date.***
    Heroku: No Heroku needed
    Pivotal Tracker: https://www.pivotaltracker.com/n/projects/2536108
    Github Repo: https://github.com/rrshah099/Octonet

15. **Links to your poster video and demo video.**
    Video:
    https://drive.google.com/file/d/1MY3v9SVR_YF9rk-9aBQUJLgHpzSa8ZCN/view?usp=sharing