Project 1: Data Analysis for Medical Applications

Names:
Andrés Martínez Almazán A01621042 and Diego Arechiga Bonilla A01621045

Course:
Modeling Learning with Artificial Intelligence

Professor:
Dr. Omar Mendoza Montoya

Date:
June 10, 2025

Contents

1 Description of the Implemented Application and Collected Data **2**
1.1 What does the application consist of? . . . 2
1.2 What is it expected to do? . . . 2
1.3 What examples of similar applications exist? . . . 2
1.4 What type of information will it monitor? . . . 2

2 Description of the Features Extracted from the Data **3**

3 Summary of Classifier Evaluation Results **3**
3.1 Hyperparameter Tuning and Feature Selection . . . 7

4 Online Application Results **7**
4.1 Background Sound Classification . . . 8
4.2 Music Classification . . . 8
4.3 Applause Classification . . . 8
4.4 Whisper Classification . . . 8
4.5 Conversation Classification . . . 8
4.6 Shout Classification . . . 8
4.7 Overall Performance Analysis . . . 8
4.7.1 Does it perform the same for all team members? . . . 8
4.7.2 Is the online application's performance as expected based on cross-validation results? . . . 9

5 Individual Conclusion **9**
5.0.1 Conclusion of Andrés Martínez Almazán . . . 9
5.0.2 Conclusion of Diego Arechiga Bonilla . . . 9

1 Description of the Implemented Application and Collected Data

1.1 What does the application consist of?

The implemented application is a real-time environmental sound classification system based on audio amplitude, developed using supervised learning. This system uses data collected through the Phyphox application Staacks, Hütz, Heinke, and Stampfer (2019), which captures sound pressure levels (dB) and associated timestamps via a mobile device. The application employs a Support Vector Machine (SVM) model with an RBF kernel, pre-trained with features extracted from 0.5-second audio windows, to classify sounds into six categories: silence/background, conversation, whisper, shout, music, and applause. The data is processed online through a remote connection with Phyphox, enabling continuous predictions and visualization of results with confidence probabilities.

1.2 What is it expected to do?

The application is expected to monitor the sound environment in real time and automatically classify detected sound types with high accuracy, providing a useful tool for medical applications. Specifically, it should identify audio patterns that may be related to health conditions, such as detecting shouts as indicators of distress (e.g., calls for help, emergency whistles, or sounds of falls), whispers as signs of vocal fatigue, or ambient noise levels that affect sensitive patients. Additionally, it is expected to offer a simple interface to visualize predictions and their confidence levels, facilitating integration into medical alert or monitoring systems.

1.3 What examples of similar applications exist?

There are similar applications in the medical and environmental monitoring fields. For example, systems developed by Philips with their patient monitoring technology use audio analysis to detect falls or critical events in hospitals Philips Healthcare (2020). Another example is the "Sound Health" digital health platform from Stanford University, which explores the use of artificial intelligence to analyze vocal patterns in patients with neurological disorders Stanford University (2019). Additionally, devices like Apple wearables with fall detection employ audio algorithms to identify sounds associated with emergencies, showcasing a comparable approach to the one proposed Apple Inc. (2021).

1.4 What type of information will it monitor?

The application monitors audio amplitude in the form of sound pressure levels (dB) and associated timestamps, captured by a mobile device's microphone through Phyphox. This information is processed in 0.5-second windows to extract features such as mean, standard deviation, percentiles, derivatives, and zero-crossing rate. The system focuses on detecting temporal and dynamic patterns that differentiate the six defined sound types, with a particular emphasis on variations that may indicate health conditions, such as amplitude peaks (shouts) or soft patterns (whispers).

2 Description of the Features Extracted from the Data

The features extracted from the data come from audio amplitude signals captured via Phyphox, processed in sliding 0.5-second windows to ensure consistency with real-time classification. This procedure generated a set of 31 variables per window, carefully designed to reflect both basic statistics and distinctive dynamic patterns among the six sound categories. These features include descriptive statistics such as mean, standard deviation, skewness, kurtosis, median, minimum, maximum, range (peak-to-peak), variance, and percentiles (10, 25, 75, 90), providing a comprehensive view of the signal's distribution and dispersion. Additionally, energy measures were calculated, including the sum of squares, average power (mean of squares), and sum of absolute values, which assess sound intensity. To capture abrupt changes in amplitude, temporal variables such as velocity (first derivative) and acceleration (second derivative) were derived, represented by their means and standard deviations. Another key metric is the zero-crossing rate (ZCR), which measures the proportion of crossings through the mean level, useful for identifying rhythmic or irregular patterns. The analysis is complemented by statistical moments, such as the third and fourth moments, which evaluate skewness and distribution shape, respectively, along with the median absolute deviation (MAD), a robust measure of dispersion. Other features include an approximate signal-to-noise ratio (SNR), calculated as the ratio between the maximum and standard deviation (adjusted to avoid division by zero), and dynamic range, defined as the difference between the 90th and 10th percentiles, reflecting extreme variability. The spectral centroid (a temporal approximation based on the weighting of absolute values by their position) and spectral flux (mean of absolute differences between consecutive samples) were also included, indicating temporal "centroidness" and rate of change, respectively. Finally, the root mean square (RMS) was added as the average signal magnitude, reinforcing energy measures. All these features were extracted using signal processing methods, ensuring the identification of distinctive patterns among the sound categories (silence/background, conversation, whisper, shout, music, applause). The process included interpolating the signals to a 20 Hz sampling rate to standardize the windows, resulting in 6,408 training samples for the SVM model.

3 Summary of Classifier Evaluation Results

The classifier evaluation was conducted using a processed dataset comprising 30 samples, each with 31 features extracted from audio signals collected via Phyphox. These data were scaled with a `StandardScaler` and evaluated using 5-fold cross-validation (`StratifiedKFold`) to ensure a balanced distribution of the six classes (silence/background, conversation, whisper, shout, music, applause). Below are the detailed results of the tested classifiers, organized into classic and unseen models, with tables summarizing the classification reports.

Table 10: Classification Report - Nearest Centroid

Table 1: Classification Report - Linear SVM

Class	**Precision**	**Recall**	**F1-score**	**Support**
1.0 (Silence/background)	1.00	1.00	1.00	5
2.0 (Conversation)	0.57	0.80	0.67	5
3.0 (Whisper)	1.00	1.00	1.00	5
4.0 (Shout)	0.60	0.60	0.60	5
5.0 (Music)	0.67	0.40	0.50	5
6.0 (Applause)	1.00	1.00	1.00	5
Accuracy	-	-	0.80	30
Macro avg	0.81	0.80	0.79	30
Weighted avg	0.81	0.80	0.79	30

Table 2: Classification Report - SVM with RBF Kernel

Class	**Precision**	**Recall**	**F1-score**	**Support**
1.0 (Silence/background)	1.00	1.00	1.00	5
2.0 (Conversation)	0.57	0.80	0.67	5
3.0 (Whisper)	1.00	1.00	1.00	5
4.0 (Shout)	0.60	0.60	0.60	5
5.0 (Music)	1.00	0.60	0.75	5
6.0 (Applause)	1.00	1.00	1.00	5
Accuracy	-	-	0.83	30
Macro avg	0.86	0.83	0.84	30
Weighted avg	0.86	0.83	0.84	30

Table 3: Classification Report - LDA

Class	**Precision**	**Recall**	**F1-score**	**Support**
1.0 (Silence/background)	1.00	0.80	0.89	5
2.0 (Conversation)	0.75	0.60	0.67	5
3.0 (Whisper)	0.83	1.00	0.91	5
4.0 (Shout)	0.71	1.00	0.83	5
5.0 (Music)	0.75	0.60	0.67	5
6.0 (Applause)	1.00	1.00	1.00	5
Accuracy	-	-	0.83	30
Macro avg	0.84	0.83	0.83	30
Weighted avg	0.84	0.83	0.83	30

Table 4: Classification Report - K-NN

Class	Precision	Recall	F1-score	Support
1.0 (Silence/background)	1.00	1.00	1.00	5
2.0 (Conversation)	0.56	1.00	0.71	5
3.0 (Whisper)	1.00	1.00	1.00	5
4.0 (Shout)	0.67	0.40	0.50	5
5.0 (Music)	1.00	0.60	0.75	5
6.0 (Applause)	1.00	1.00	1.00	5
Accuracy	-	-	0.83	30
Macro avg	0.87	0.83	0.83	30
Weighted avg	0.87	0.83	0.83	30

Table 5: Classification Report - MLP

Class	Precision	Recall	F1-score	Support
1.0 (Silence/background)	1.00	1.00	1.00	5
2.0 (Conversation)	0.57	0.80	0.67	5
3.0 (Whisper)	1.00	1.00	1.00	5
4.0 (Shout)	0.75	0.60	0.67	5
5.0 (Music)	0.75	0.60	0.67	5
6.0 (Applause)	1.00	1.00	1.00	5
Accuracy	-	-	0.83	30
Macro avg	0.85	0.83	0.83	30
Weighted avg	0.85	0.83	0.83	30

Table 6: Classification Report - Gaussian Naive Bayes

Class	Precision	Recall	F1-score	Support
1.0 (Silence/background)	1.00	1.00	1.00	5
2.0 (Conversation)	0.80	0.80	0.80	5
3.0 (Whisper)	1.00	1.00	1.00	5
4.0 (Shout)	0.80	0.80	0.80	5
5.0 (Music)	0.80	0.80	0.80	5
6.0 (Applause)	1.00	1.00	1.00	5
Accuracy	-	-	0.93	30
Macro avg	0.93	0.93	0.93	30
Weighted avg	0.93	0.93	0.93	30

Table 7: Classification Report - Gradient Boosting

Class	Precision	Recall	F1-score	Support
1.0 (Silence/background)	1.00	1.00	1.00	5
2.0 (Conversation)	0.67	0.80	0.73	5
3.0 (Whisper)	1.00	1.00	1.00	5
4.0 (Shout)	1.00	0.80	0.89	5
5.0 (Music)	0.80	0.80	0.80	5
6.0 (Applause)	1.00	1.00	1.00	5
Accuracy	-	-	0.90	30
Macro avg	0.91	0.90	0.90	30
Weighted avg	0.91	0.90	0.90	30

Table 8: Classification Report - Extra Trees

Class	Precision	Recall	F1-score	Support
1.0 (Silence/background)	1.00	1.00	1.00	5
2.0 (Conversation)	0.80	0.80	0.80	5
3.0 (Whisper)	1.00	1.00	1.00	5
4.0 (Shout)	1.00	1.00	1.00	5
5.0 (Music)	0.80	0.80	0.80	5
6.0 (Applause)	1.00	1.00	1.00	5
Accuracy	-	-	0.93	30
Macro avg	0.93	0.93	0.93	30
Weighted avg	0.93	0.93	0.93	30

Table 9: Classification Report - Gaussian Process

Class	Precision	Recall	F1-score	Support
1.0 (Silence/background)	1.00	1.00	1.00	5
2.0 (Conversation)	0.50	0.60	0.55	5
3.0 (Whisper)	1.00	1.00	1.00	5
4.0 (Shout)	0.67	0.60	0.63	5
5.0 (Music)	0.75	0.60	0.67	5
6.0 (Applause)	1.00	1.00	1.00	5
Accuracy	-	-	0.80	30
Macro avg	0.82	0.80	0.81	30
Weighted avg	0.82	0.80	0.81	30

Class	Precision	Recall	F1-score	Support
1.0 (Silence/background)	1.00	1.00	1.00	5
2.0 (Conversation)	0.57	0.80	0.67	5
3.0 (Whisper)	1.00	1.00	1.00	5
4.0 (Shout)	0.75	0.60	0.67	5
5.0 (Music)	1.00	0.60	0.75	5
6.0 (Applause)	1.00	1.00	1.00	5
Accuracy	-	-	0.83	30
Macro avg	0.86	0.83	0.85	30
Weighted avg	0.86	0.83	0.84	30

3.1 Hyperparameter Tuning and Feature Selection

The process of hyperparameter optimization and feature selection for the SVM with RBF kernel and Gradient Boosting models was developed as follows. Initially, hyperparameter performance was plotted through preliminary explorations to determine optimal values. For the SVM with RBF kernel, the parameters `C` (regularization) and gamma (kernel width) were varied over a wide range, identifying that `C` = 20.0 and gamma = 0.15 offered the best performance based on the generated curves. For Gradient Boosting, number of trees, maximum depth, and learning rate were tuned, finding that number ofestimators = 50, maxdepth = 2, and learning rate = 0.05, along with the best k = 12, maximized accuracy in preliminary plots.

Next, feature selection was conducted. For SVM, `SelectKBest` with the `f_classif` criterion was used, selecting the 10 most relevant features based on their statistical scores, corresponding to indices [4, 8, 10, 11, 12, 15, 18, 20, 22, 23], which represent median, variance, 25th, 75th, and 90th percentiles, average power, standard deviations of velocity and acceleration, and third and fourth moments. For Gradient Boosting, with `k` = 12 was used, identifying indices [0, 4, 8, 10, 11, 12, 14, 15, 18, 20, 22, 23], which include mean, median, variance, 25th, 75th, and 90th percentiles, total energy, average power, velocity and acceleration deviations, and third and fourth moments, based on nested cross-validation results. Finally, definitive optimization was performed using nested cross-validation with GridSearchCV in the inner loop. For SVM RBF, `C` = 20.0 and gamma = 0.15 were confirmed, achieving an accuracy of 0.9667 in the inner CV. For Gradient Boosting, number of estimators = 50, max depth = 2, learning rate = 0.05, and k = 12 were validated, resulting in an average overall accuracy of 0.9333.

4 Online Application Results

The online application for real-time environmental sound classification was implemented using a pre-trained SVM model with an RBF kernel. Data were processed in 0.5-second windows with 50% overlap (0.25-second step), interpolated to a 20 Hz sampling rate, and 31 features were extracted per window, replicating the training method. These features were scaled and the 10 most relevant were selected using a pre-trained `SelectKBest`, before being passed to the final SVM model with optimal hyperparameters (`C` = 20.0, `gamma` = 0.15). Below are the results observed during online testing, based on the provided output example.

4.1 Background Sound Classification

Background sound classification was perfect, with the model consistently identifying this category in multiple consecutive predictions (e.g., predictions with confidence levels between 0.80 and 0.99). This indicates that the model is highly effective at detecting silent or low-background-noise environments.

4.2 Music Classification

Music classification was accurate as long as the volume was high, as the model was trained. In the output example, multiple correct predictions were observed with confidence levels ranging from 0.61 to 0.99. This demonstrates the model's ability to identify rhythmic and dynamic patterns associated with music at high volumes.

4.3 Applause Classification

Applause classification was very good, although not all predictions were identified as applause due to timing differences between events. However, the model correctly identified most applause instances, with confidence levels between 0.56 and 1.00. This suggests that the model is robust for detecting intermittent sound events, though temporal variability may cause some erroneous predictions.

4.4 Whisper Classification

Whisper classification was perfect as long as the whisper remained within a specific amplitude range. Multiple correct identifications were made with confidence levels between 0.59 and 0.95. However, if the whisper was too loud, the model tended to confuse it with normal conversation.

4.5 Conversation Classification

Conversation classification was good but required a specific and consistent tone of voice. Conversations were identified in predictions with confidence levels between 0.49 and 0.78. However, if the tone was inconsistent, the model tended to confuse conversation with shouting.

4.6 Shout Classification

Shout classification was the least accurate, requiring a highly specific and consistent tone of voice. Although some shouts were identified, the model frequently confused them with music or applause, with confidence levels ranging from 0.45 to 0.89. This indicates that the model struggles to differentiate shouts from other high-amplitude sounds, suggesting the need to adjust the model or incorporate additional features, such as spectral analysis, to improve accuracy.

4.7 Overall Performance Analysis

4.7.1 Does it perform the same for all team members?

The online application does not perform identically for all team members due to variations in audio capture between devices and environments. Tests indicated that microphone quality and

environmental conditions (e.g., background noise, distance to the microphone) affect prediction accuracy. For example, a team member with a high-sensitivity device (e.g., a high-end smartphone) achieved better results for whispers and applause, while another with a lower-quality device showed inconsistencies in detecting shouts and conversations. This suggests that hardware calibration and environmental standardization are critical factors for ensuring consistent performance across team members.

4.7.2 Is the online application's performance as expected based on cross-validation results?

The online application's performance is generally consistent with the cross-validation results obtained for the SVM model with RBF kernel. In online tests, categories such as background sound, music, whisper, and applause showed accuracy rates close to or above 80-100%, aligning with expectations based on cross-validation. However, shout classification was significantly less accurate, with frequent confusions with music and applause, suggesting performance below expectations. This could be due to variability in real-time recordings or the lack of specific spectral features in the current model, indicating a need for refinement to fully meet cross-validation projections.

5 Individual Conclusion

5.0.1 Conclusion of Andrés Martínez Almazán

Undoubtedly, the experience of working on this project was highly rewarding and exciting. As my first machine learning project, I felt very satisfied with the results achieved. Working with an audio signal different from the example code provided by our professor posed a significant challenge, but I believe that was precisely the essence of the project: learning to navigate independently and find innovative solutions. Effective communication with my teammate Diego was crucial to ensuring the quality of the work; our strategic division of tasks allowed us both to learn without feeling overwhelmed. We used a seed whenever possible when editing the code, ensuring consistent and reproducible results, a key aspect for validating our progress. The greatest challenge for me was implementing online classification. While training the model was relatively straightforward, adapting it to process data in real time was complex, especially because the chosen audio signal has parameters and characteristics different from the professor's example, such as amplitude variations and temporal patterns. Despite this, we managed to develop a high-quality model, with an internal validation accuracy of 0.9667 and promising online performance for categories like background sound and music. However, limitations, particularly in shout classification, where confusions with music and applause suggest the need for more robust training, were evident. I believe that incorporating a more diverse dataset and additional features could improve accuracy in real-time evaluations. This experience has motivated me to continue exploring signal processing techniques for future health monitoring research.

5.0.2 Conclusion of Diego Arechiga Bonilla

During this project, my teammate and I were able to apply various concepts learned in class in a unique and interesting context, using different supervised learning algorithms to identify the type of sound detected through our phone microphones. Some of the algorithms we used

included KNN, SVM (linear), SVM (RBF), among many others covered in class. When analyzing accuracy, recall, and precision, we concluded that the Extra Trees model performed best, although after optimizing hyperparameters for different models, we observed that SVM (RBF) was the most effective. While working on the project, we encountered various challenges and problems that we had to address on the fly. For example, one challenge was adapting the SVM (RBF) learning model to identify sounds in real time by analyzing data sent from the phone. This was somewhat complex since we chose a different signal from the code provided by the professor, requiring several modifications to make it work properly. In conclusion, this project was quite intriguing as we were able to apply everything learned in class in a different and unique way. Although it was challenging, I enjoyed pushing ourselves by using a different signal from the one used by the professor and finding ways to solve the problems we faced. Moreover, this project helped me realize the vast range of uses and applications that these machine learning models can have.

References

Apple Inc. (2021). *Fall detection with audio algorithms in wearable devices.* Technical documentation. Retrieved 2025-06-10, from `https://developer.apple.com/documentation/healthkit`

Philips Healthcare. (2020). *Patient monitoring systems: Audio-based fall detection* (Technical report). Amsterdam, Netherlands: Philips Healthcare.

Staacks, S., Hütz, S., Heinke, H., & Stampfer, C. (2019). Advanced tools for smartphone-based experiments: Phyphox. *Physics Education*, *53*(4), 045009. doi: 10.1088/1361-6552/aac05e

Stanford University. (2019). *Sound health initiative: Analyzing vocal patterns in neurological disorders.* Research project. Retrieved 2025-06-10, from `https://med.stanford.edu/soundhealth`