**Introduction**

**Title:** Can machine learning capture snoring moments from audio recordings of sleeping activity?

**Primary Objectives:**

The primary objectives of this project are as follows:

Detect and comprehend the occurrences of snoring sounds within sleeping data, which consists of audio recordings.

Investigate the effectiveness of deep neural networks, including CNN, CNN+LSTM, and Transformer, in accurately identifying snoring events when compared to traditional machine learning algorithms such as k-nearest neighbor (KNN), support vector machine (SVM), and logistic regression.

Based on the above primary objectives, our Exploratory Data Analysis (EDA) process intends to achieve following objectives.

* **Data Understanding:** Gain a comprehensive understanding of the provided dataset, including its structure, size, and the nature of the audio recordings.
* **Data Quality Assessment:** Assess the quality of the data, identifying and addressing any missing values, outliers, or inconsistencies.
* **Descriptive Statistics:** Calculate and present summary statistics for key variables, such as subject numbers and snoring levels, to provide initial insights into the data.
* **Visualization:** Create visualizations to explore the distribution of snoring events over time, across participants, and by snoring level. Visualize any trends or patterns in the data.
* **Feature Engineering:** If necessary, perform initial feature engineering to extract relevant audio features or characteristics that can potentially aid in snoring detection.
* **Data Preprocessing:** Prepare the data for model development by applying any necessary preprocessing steps, such as normalizing audio data or encoding categorical variables.
* **Exploratory Data Insights:** Provide insights and observations from the EDA process, such as the prevalence of snoring events, variations in snoring levels, and any potential correlations with other variables.
* **EDA Findings for Model Selection:** Offer recommendations for the selection of appropriate machine learning models based on the characteristics and distribution of the data uncovered during the EDA.

The data contains 21 participant’s sleeping data (average 50 minutes audio recordings (.wav)). The data will contain the following information:

● Subject number;

● Start timestamp of snoring;

● End timestamp of snoring;

● Snoring level (S1, S2, S3)

**Data Cleaning and Preprocessing**

*Analysis Environment*

The analysis for this project was conducted using the Python programming language within Jupyter Notebook, a popular integrated development environment for data analysis and machine learning tasks. Several essential Python libraries were imported to facilitate data manipulation, visualization, and analysis. The following libraries were utilized:

* pandas (as pd): The pandas library was employed to manage and manipulate structured data efficiently. It provides powerful data structures, such as dataframes, and tools for data cleaning and transformation.
* os: The os library was used for interacting with the operating system, allowing us to handle file operations and directory management as needed during the analysis.
* matplotlib.pyplot (as plt): Matplotlib is a versatile library used for creating a wide range of static, animated, or interactive visualizations in Python. We specifically used the pyplot module for generating plots and charts to visualize our findings.
* seaborn (as sns): Seaborn is a data visualization library built on top of matplotlib. It provides a high-level interface for creating informative and attractive statistical graphics. We utilized seaborn for creating various data visualizations.
* numpy (as np): NumPy is a fundamental library for numerical computing in Python. It provides support for arrays and matrices, along with mathematical functions, making it invaluable for data analysis, especially when working with numerical data.

*Data Preprocessing*

In the initial phase of our analysis, we conducted essential data preprocessing steps on the snoring event data obtained from the CSV file. The primary objectives of these preprocessing steps were to enhance data clarity and facilitate subsequent analysis.

To begin with, we read the CSV file using the pandas library, creating a DataFrame to store the data. This DataFrame, denoted as 'df0,' contains a total of 626 rows and 4 columns, with the columns labeled as 'starting time,' 'ending time,' 'level of snoring,' and 'duration.' Renaming the columns was a crucial step in achieving clarity and ensuring that the data's meaning was easily discernible.

Subsequently, we computed the duration of each snoring event by subtracting the 'start\_time' from the 'end\_time.' This calculated duration was then added as a new column named 'duration' within the DataFrame. The inclusion of this 'duration' feature not only enhanced the interpretability of the data but also provided valuable temporal information that could prove useful in subsequent stages of our analysis.

Next, we combined data from all labels. To consolidate this data from multiple sources, we initialized an empty list to hold individual dataframes. We then iterated through each CSV file, reading its content into separate dataframes, assuming that the files did not include headers. These dataframes were subsequently appended to the list.

Once all relevant data had been collected, we proceeded to combine and merge the individual dataframes into a single, unified dataframe named 'combined\_df.' To enhance the clarity of the data, we renamed the columns within 'combined\_df' as 'start\_time,' 'end\_time,' and 'level.' This consolidation and column renaming process facilitated a comprehensive and standardized representation of the snoring event data from multiple files. The resulting 'combined\_df' would serve as the basis for our subsequent exploratory data analysis and modeling tasks, enabling us to gain meaningful insights into snoring patterns and levels across multiple participants.

In our data analysis, we conducted an essential check for null values within the 'combined\_df' dataframe. During this examination, we identified specific rows where null values were present, particularly in the 'level' column. These instances of missing values could potentially affect the quality and completeness of our analysis and therefore, we decided to address these null values before further analysis of the data.

*Data Cleaning*

In a critical data quality enhancement step, we addressed the issue of null values in our 'combined\_df' dataframe. Specifically, we executed the operation to remove rows with null values by utilizing the 'dropna' function. We decided to remove the null values as null values appeared in the level of snoring and therefore, we are not able to impute the values to complete the data set. This process resulted in the elimination of any rows that had missing data, ensuring that our dataset is free from incomplete or inconsistent entries. By taking this action, we have enhanced the robustness and reliability of our data, which is essential for our subsequent exploratory data analysis and modeling efforts.

In our data preparation and feature engineering process, we calculated the duration of snoring events within the 'combined\_df' dataframe, which encapsulates information related to these events. To accomplish this, we performed a straightforward mathematical operation, subtracting the 'start\_time' from the 'end\_time' for each event, and created a new column labeled 'duration' to store these calculated durations. This 'duration' feature is invaluable as it quantifies the time span of each snoring event, providing an additional dimension to our dataset.

As the next step, we examined the distinct levels of snoring events within the 'combined\_df' dataframe, which holds information about these events. Upon performing this analysis, we identified a variety of unique snoring levels that characterize the dataset. These levels include 'Start to sleep,' 'S1,' 'S2,' 'S3,' 'S1`,' '1,' 'S1 ,' 'S1,' '1S,' 'W1,' 'SS1,' and 'D1.' By observing the output we realized that there were some typos in the level of snoring.

Before making any modifications to our dataset, we create a new dataframe named 'combined\_df\_new' as an exact copy of the original 'combined\_df.' As mentioned before, several snoring level values, such as '1,' 'S1`,' 'S1 ,' 'S!,' '1S,' 'W1,' 'SS1,' and 'D1,' were identified as typos or misclassifications. To ensure data accuracy and uniformity, we systematically replaced these irregular values with the standardized snoring level 'S1.' This correction aimed to harmonize and simplify the snoring level classification within our dataset. By executing this replacement operation, we improved the overall data quality, making it more consistent and suitable for our exploratory data analysis and modeling endeavors. After the correction of typos, we successfully harmonized and simplified the snoring level classification. The unique snoring levels within the dataset now consist of 'S1,' 'S2,' and 'S3.' This consolidation of snoring levels enhances the clarity and consistency of the dataset, making it more suitable for analysis and modeling.

In our data refinement and transformation process, the 'combined\_df\_new' dataframe now exhibits a well-structured and standardized format, making it conducive for in-depth analysis and modeling. The dataframe comprises a total of 19,967 rows and four columns, with each row representing a specific snoring event. The columns include 'start\_time,' 'end\_time,' 'level,' and 'duration.' 'Start\_time' and 'end\_time' denote the temporal boundaries of each snoring event, while 'level' signifies the categorized intensity of the snoring event, with values primarily consisting of 'S1,' 'S2,' and 'S3.' Additionally, 'duration' quantifies the time span of each snoring event.

In our data analysis, we performed a count of snoring levels within the 'combined\_df\_new' dataframe, revealing the distribution of snoring intensities among the study participants. The analysis unveiled that the majority of snoring events were categorized as 'S1,' accounting for 18,994 occurrences. Additionally, 'S2' was observed in 865 events, while 'S3' was identified in 81 instances. This distribution indicates that a significant portion of the dataset is skewed toward the 'S1' level, signifying a notable imbalance in snoring event intensities.

Subsequently, we conducted a count of snoring levels to analyze the distribution and compared it with the provided client overview data. An intriguing observation emerged - a substantial disparity between the total count of 'S1' levels in our combined dataframe and the client's overview data. According to the client's information, the total count for 'S1' was reported as 11,585, whereas our combined dataframe, even after correcting typos, indicated a count of 18,994. This difference was further accentuated when we excluded the typo levels, with 'S1' still accounting for 18,984 occurrences. This discrepancy has raised questions about the specific data that was included or excluded in the client's count. Despite this uncertainty, we decided to proceed with the dataframe in which the typo levels were standardized to 'S1.'

After identifying the situation, we systematically handled our multiple CSV files. Firstly, we initiated the process by listing all CSV files in the folder and subsequently created an empty list to hold dataframes, providing a structured approach to managing the data. For each CSV file, we conducted a series of critical data cleaning and preprocessing steps. These included the removal of rows with null values and the correction of typo levels. Specifically, we identified and replaced irregular level values, such as '1,' 'S1`,' 'S1 ,' 'S!,' '1S,' 'W1,' 'SS1,' and 'D1,' standardizing them to 'S1.' This correction ensured data uniformity and accuracy. Secondly, we calculated the duration of each snoring event by subtracting the 'start\_time' from the 'end\_time,' providing valuable temporal information. To enhance data clarity, we assigned headers to the data, making it more interpretable. Finally, we saved the cleaned and processed data as new output files, preserving the original file names within the 'output\_folder.'

*Analyse the data set*

*Descriptive Statistics*

In our data analysis, we conducted a comprehensive examination of snoring events, grouping them by their categorized intensity levels within the 'combined\_df\_new' dataframe. This analysis aimed to provide insights into the average duration of snoring events across different intensity levels. Our findings indicate that 'S1' levels, the most prevalent category, had an average duration of approximately 3.08 minutes, with a standard deviation of 0.16 minutes. Snoring events classified as 'S2' exhibited a slightly higher average duration of around 3.09 minutes, with a standard deviation of 0.28 minutes. 'S3' level snoring events had an average duration of approximately 2.98 minutes. These insights offer valuable information about the typical durations of snoring events at different intensity levels, with 'S1' being the most frequent and longest-lasting, followed by 'S2,' and 'S3'. The detailed output is as follows:

A screenshot of a graph

Description automatically generated

In our data visualization process, we created a count plot to illustrate the distribution of snoring levels within the 'combined\_df\_new1' dataframe. The plot showcases the frequency of different snoring levels, providing a visual representation of the data. Each level is represented by a distinct bar, and the height of each bar corresponds to the count of snoring events for that specific level. The plot reveals that 'S1' snoring events are the most prevalent, followed by 'S2' and 'S3,' showcasing the distribution of snoring intensity levels among the study participants. The count labels above each bar further enhance the interpretability of the plot. This visualization aids in better understanding the relative occurrences of snoring levels, offering valuable insights into the distribution of snoring intensity levels.

A graph with a bar and numbers

Description automatically generated with medium confidence

Next, we generated a boxplot to examine the distribution of snoring durations across different snoring levels within the 'combined\_df\_new1' dataframe. The boxplot provides a visual representation of the central tendency and variability in snoring event durations for each intensity level. The x-axis categorizes snoring levels, with 'S1,' 'S2,' and 'S3' presented in descending order of frequency. The y-axis represents the duration of snoring events in seconds. The boxplot showcases the central tendency (median) and spread (interquartile range) of snoring durations for each level. It offers insights into the variations in snoring event durations among different intensity levels. We have identified many outliers in all the three levels and the highest outlier is identified under S1 level.

A diagram of different colored lines

Description automatically generated

*Analysis of mfcc extraction*

In this analysis, we try to analyse our mfcc data which extracted from the sound files (.wav files). Firstly, we create a new DataFrame called filtered\_df, which contains only the data related to the 'wav\_id' 'Video\_00\_001.wav' from the original DataFrame df\_mfcc. It effectively filters the data to focus on a specific audio sample identified by its 'wav\_id'. Then we convert the strings representing MFCC features into actual lists of numerical values for columns such as 'mfcc\_1' through 'mfcc\_13.' This conversion allowed us to work with the MFCC features as numerical data. The resulting 'filtered\_df' DataFrame encapsulates essential information about this specific audio sample, including start and end times, duration, label, and the extracted MFCC feature values. Then we calculated the mean value of the DataFrame filtered\_df and it is - 597.431

After organizing the mfcc data frame, we generate a series of subplots, each depicting a histogram for one of the 13 Mel-frequency cepstral coefficient (MFCC) columns. For MFCC, we identified two main levels as “non-snoring” and “snoring” and then snoring will further classify in to three levels, “S1”, “S2” and S3”. It operates on a DataFrame called filtered\_df, focusing on specific labels, including 'N', 'S1', 'S2', and 'S3'. For each MFCC column, it flattens the corresponding MFCC values across these selected labels and creates histograms to visualize the distribution of these values. Each subplot represents one MFCC column, and the histograms for each label are overlaid for comparison, with distinctive alpha values to distinguish them. The subplots are arranged vertically, providing a clear insight into the frequency distribution of MFCC values for each of the selected labels. This analysis can help in understanding how MFCC features vary among different snoring intensity levels, aiding in the interpretation of audio data.

Need to include few plots

Among all the histograms, we identified few of them are more skewed compared to others. For instance, MFCC\_1, MFCC\_3, MFCC\_8, MFCC\_9, MFCC\_11

Finally, we utilize the Librosa library to generate a spectrogram from an audio file named 'Video\_00\_001.wav.' It begins by loading the audio data and sample rate using Librosa. Subsequently, it calculates the spectrogram by employing the Mel frequency spectrogram (mel-spectrogram) as a visual representation of the audio's spectral content. Then we generate a plot to display the spectrogram, with mel frequency on the vertical axis and time on the horizontal axis. To enhance visualization, the spectrogram's power values are converted to decibels, making it easier to perceive relative amplitudes of different frequency components.