III Semester B.Tech., (CCE)
Data Analytics – CSS 2103
Internal Assessment 4 - Mini Project Report
School of Computer Engineering
Manipal Institute of Technology, MAHE

"Identification of Frequent Co-occuring Patterns in Physiological Signals"

Submitted by:

Name: Aarav Sadhu
Roll No: 15
Registration Number: 240953158
Section: C
Branch: CCE

Name: Khwaish Parakh
Roll No: 24
Registration Number: 240953244
Section: C
Branch: CCE

Name: Ragini Gupta
Roll No: 33
Registration Number: 240953306
Section: C
Branch: CCE

Name: Manognya Tanvi
Roll No: 35
Registration Number: 240953318
Section: C
Branch: CCE

Table of Contents

1. Introduction

The SWELL Dataset was designed to study how workplace stress affects human physiology and performance. It contains data from participants performing typical office tasks in three controlled conditions: baseline, time pressure, and interruption.

During the experiment, physiological signals such as heart rate, heart rate variability (RMSSD, LF/HF ratio), and sample entropy were recorded, reflecting how the body responds under changes in stress and workload.

This project leverages the SWELL dataset for cleaning, exploring, and mining association rules that define frequent co-occurring physiological patterns for different conditions of stress as a way of understanding how physiological signals change with stresses.

2. Objectives

The main objectives of this project are to:

- Clean and preprocess the physiological data from the SWELL dataset to remove noise and outliers.
- Perform exploratory data analysis to understand signal behavior under different experimental conditions.
- Apply z-score normalization and compute basic descriptive statistics for selected physiological features.
- Use Association Rule Mining (Apriori / FP-Growth) to identify frequent co-occurring patterns in physiological signals.
- Analyze and interpret the discovered rules to understand relationships between physiological responses and stress conditions.

3. Problem Statement

Work-related stress has the potential to significantly affect an individual's performance and physiological state. Its early identification and management can thus be maintained by detecting such stresses using measurable physiological patterns. However, raw physiological data generally exhibit noise and variability that make meaningful insights difficult to extract.

This project deals with the processing and analysis of the SWELL data set to find frequent patterns for co-occurring physiological signals, using data cleaning, statistical analysis, and association rule mining, to discover relationships between body responses and stress conditions.

4. Background Study

The SWELL-KW (Stress at Work) dataset was developed by Radboud University Nijmegen to study how different types of workplace stress affect human behavior and physiology. Prior studies using this dataset have shown that stress induces noticeable changes in physiological signals such as heart rate (HR), heart rate variability (HRV), electrodermal activity (EDA), and respiration rate. These variations correspond to the activation of the autonomic nervous system (ANS) during cognitive and emotional stress.

Previous research primarily focused on stress detection and classification using machine learning models applied to HRV and EDA features. Studies have demonstrated that measures such as RMSSD, LF/HF ratio, and sample entropy (SampEn) effectively reflect mental workload and stress responses. However, fewer studies have explored co-occurrence patterns among these physiological variables, which could reveal deeper insights into how multiple body responses interact under stress.

This project extends existing work by applying Association Rule Mining (ARM) techniques, specifically the Apriori and FP-Growth algorithms, to uncover frequent co-occurring physiological patterns under different experimental conditions in the SWELL dataset. This approach emphasizes interpretability and pattern discovery rather than predictive modeling, contributing to a better understanding of physiological correlates of stress.

5. Methodology

5.1 Data Collection

The dataset used in this project was obtained from Kaggle, where a processed version of the SWELL-KW (Stress at Work) dataset is available in CSV format. This dataset contains physiological recordings collected from participants performing office-related tasks under three experimental conditions — baseline, time pressure, and interruption. Each record includes features such as heart rate (HR), heart rate variability (RMSSD), LF/HF ratio, and sample entropy (SampEn). These signals were originally derived from raw physiological measurements (ECG and other biosensors) recorded during the SWELL experiment. The Kaggle version consolidates this data into structured CSV files (train.csv and test.csv), which were used in this project for analysis after cleaning and preprocessing.

After obtaining the dataset, an initial inspection revealed several inconsistencies such as missing values, redundant columns, and irregular physiological readings across participants. Some records contained incomplete measurements or unrealistic signal values that could distort analysis. These issues were addressed by removing missing and irrelevant entries, filtering out noise, and retaining only the essential physiological features, Heart Rate (HR), RMSSD, LF/HF ratio, and Sample Entropy (SampEn). To account for individual variability, all signals were normalized using z-score filtering, ensuring comparability across subjects. The normalized values were then converted into categorical bins, low, normal, and high,

transforming the continuous signals into a discrete form suitable for association rule mining. This preprocessing ensured that the dataset was clean, standardized, and ready for pattern discovery.

The Kaggle version of the SWELL dataset that we used did not include Electrodermal Activity (EDA) or Galvanic Skin Response data, which are available in the original SWELL-KW dataset. To address this, we used other physiological signals, Heart Rate (HR), RMSSD, LF/HF ratio, and Sample Entropy (SampEn) as alternatives. These measures provide reliable indicators of stress and autonomic activity like EDA.

5.2 Tools and Technologies Used

This project was implemented primarily using the Python programming language and several of its data analysis libraries.

- Python - the main programming language used for data analysis and implementation.
- Jupyter Notebook - interactive environment used for writing, executing, and documenting the code.
- Pandas - Python library for data cleaning, manipulation, and preprocessing.
- NumPy - Python library for numerical operations and z-score normalization.
- Matplotlib - Python library used to visualize association rule results and descriptive statistics.
- mlxtend - Python library used to implement Apriori, FP-Growth, and generate association rules.
- Scikit-learn - Python library supporting basic statistical calculations and standardization.
- Kaggle - the online platform used to obtain the pre-processed SWELL-KW dataset.

6. Implementation

The project was implemented using Jupyter and the entire code is contained in a .ipynb file. The first few steps were pre-processing. Attached below are the screenshots for the same:

6.1 Data Preprocessing and Cleaning

As the dataset from Kaggle was already partially processed, we focused on refining and preparing it for analysis rather than performing raw signal extraction. We began by removing missing values and unnecessary columns, followed by z-score normalization to standardize each physiological feature across participants.

Step 1: Selecting the Necessary Columns and Dropping the Rest.

```
physio_cols = ['datasetId', 'condition', 'HR', 'RMSSD', 'LF_HF', 'sampen']
df = df[physio_cols].copy()
```

Step 2: Looking for Missing Values and Handling Those.

```
print("\nMissing values before cleaning:")
print(df.isna().sum())
df = df.dropna(subset=['HR', 'RMSSD', 'LF_HF', 'sampen'])
```

Step 3: Removing Outliers Using Z-Score.

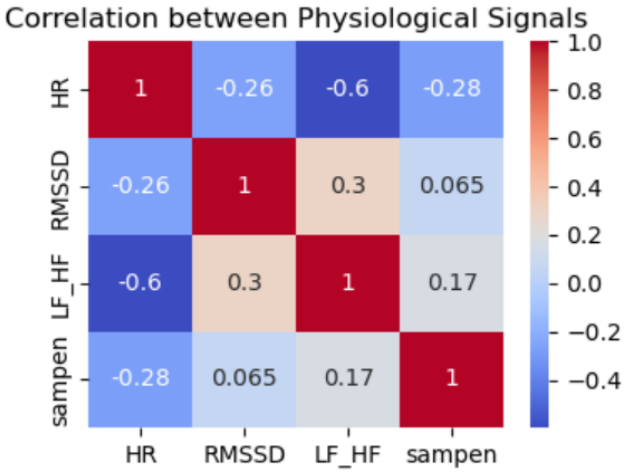
```
def remove_outliers_groupwise(df, cols, z_thresh=3.0):
    df2 = df.copy()
    bad_idx = set()
    for pid, g in df2.groupby('datasetId'):
        for col in cols:
            if g[col].nunique() > 1:
                z = np.abs(stats.zscore(g[col], nan_policy='omit'))
                bad_idx.update(g.index[z > z_thresh])
    return df2.drop(index=list(bad_idx))

clean_cols = ['HR', 'RMSSD', 'LF_HF', 'sampen']
df_clean = remove_outliers_groupwise(df, clean_cols, z_thresh=3.0)
print("\nShape after cleaning:", df_clean.shape)
```

```
Initial shape: (410322, 36)

Missing values before cleaning:
datasetId    0
condition    0
HR           0
RMSSD        0
LF_HF        0
sampen       0
dtype: int64

Shape after cleaning: (391137, 6)
```

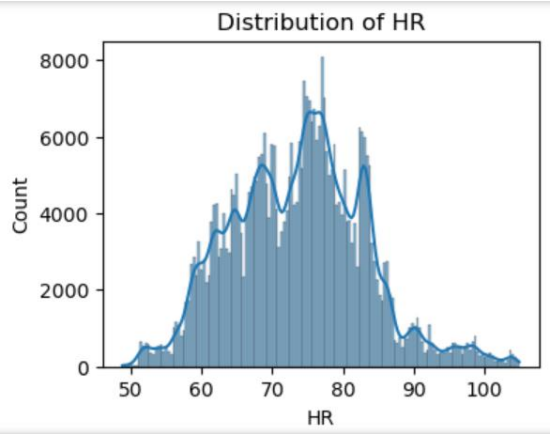
For simplicity and interpretability, we selected four key physiological features: HR, RMSSD, LF/HF ratio, and SampEn, as they are the most representative indicators of stress responses.

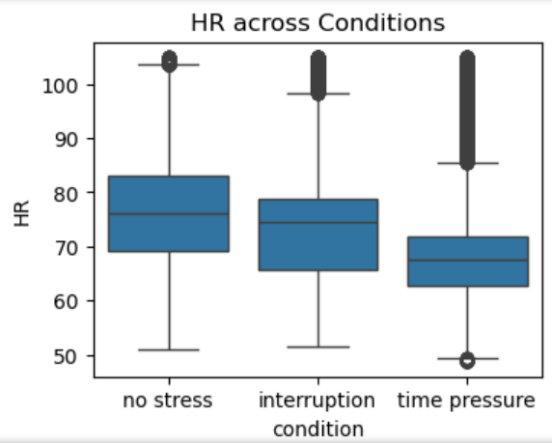
6.2 Data Visualization and Exploration

After data pre-processing and cleaning there were visualizations made to help draw conclusions about certain correlations. Some screenshots and insights are attached below:

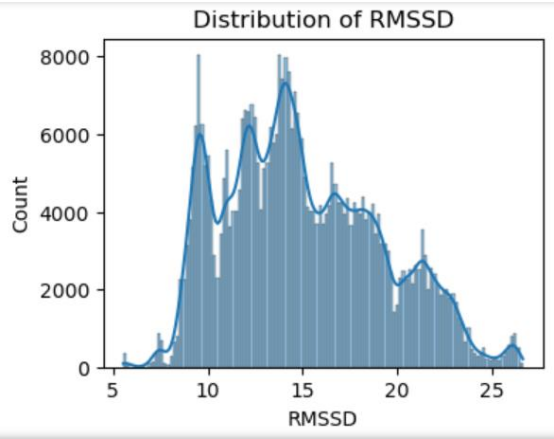


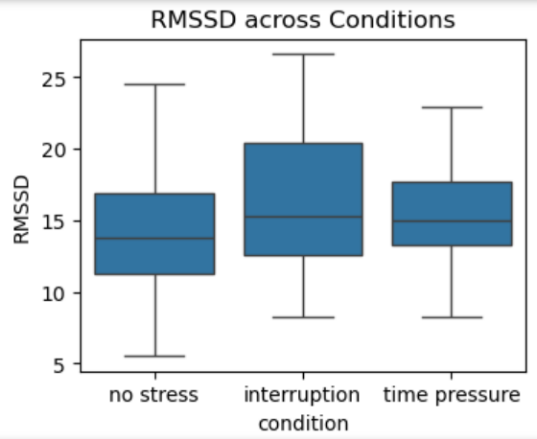
Insights: HR shows a strong negative correlation with RMSSD and a positive correlation with LF/HF ratio, indicating that higher heart rate is linked with reduced heart rate variability and increased sympathetic activity under stress.



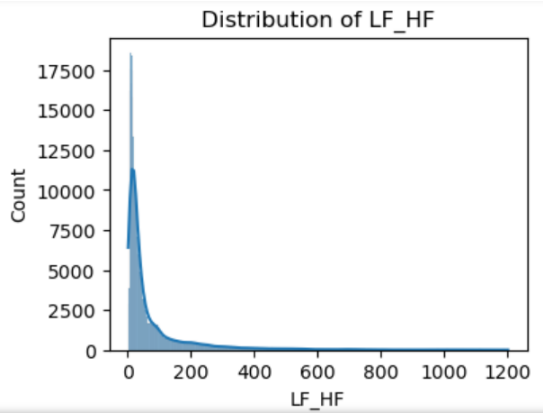


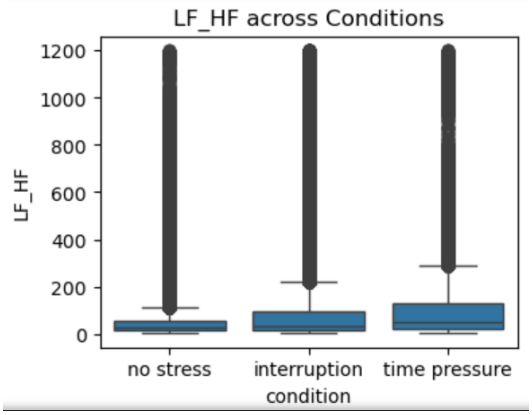
Insights: The distribution is right-skewed with more high HR readings during time pressure. The box plot confirms that HR is highest under time pressure, moderate during interruption, and lowest at baseline.



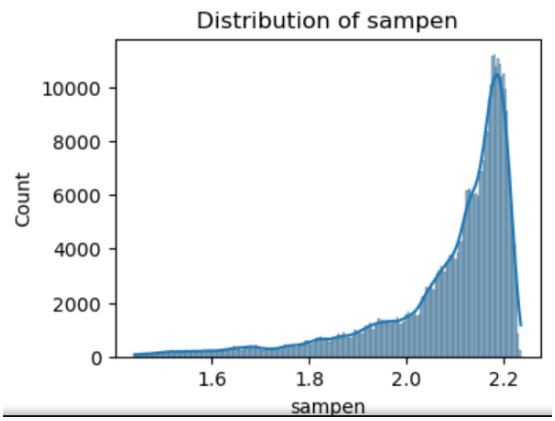


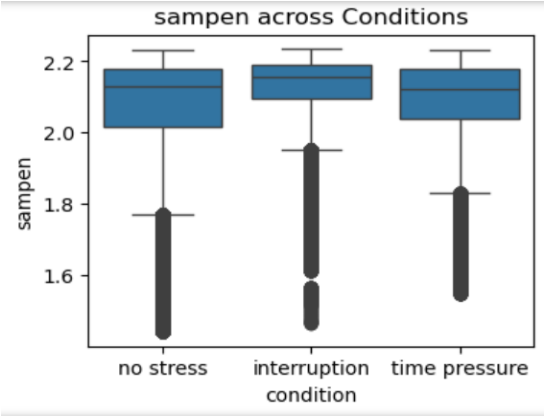
Insights: The distribution shows lower RMSSD values, and the box plot highlights a clear decrease in RMSSD during time pressure, indicating reduced variability and elevated stress.





Insights: The distribution widens under stress conditions, and the box plot shows increased LF/HF ratios during time pressure and interruption, reflecting sympathetic dominance compared to baseline.





Insights: The distribution shifts slightly lower under stress, and the box plot reveals reduced entropy values during time pressure, suggesting more regular and less complex heart rate dynamics when stressed.

6.3 Association Rule Mining

After visualizing and exploring the dataset, association rule mining is applied to identify frequent co-occurring patterns among the physiological signals using the Apriori and FP-Growth algorithms.

Z-score normalization is done per participant and those are then converted into bins, and then into categorial items to describe the state of a participant's physiological signals at a given time. Each record was converted into a transaction representing a participant's physiological

state. The Apriori and FP-Growth algorithms then scan these transactions to find frequent combinations of items.

After running the Apriori and FP-Growth algorithms, the most significant patterns are extracted and displayed as the top 5 association rules. These tables show combinations of physiological states (antecedents) that frequently occur together with certain conditions (consequents), along with their support, confidence, and lift values. Higher lift and confidence indicate stronger, more meaningful relationships between physiological responses and stress conditions.

The rules below represent the strongest physiological patterns found after applying both algorithms.

Top 5 Apriori Rules

antecedents	consequents	support	confidence	lift
RMSSD_high, HR_low	sampen_normal, LF_HF_high	0.043381	0.602450	6.518768
sampen_normal, RMSSD_high, HR_low	LF_HF_high	0.043381	0.617602	6.492516
RMSSD_high, HR_low	LF_HF_high	0.043974	0.610687	6.419821
LF_HF_high, RMSSD_high	sampen_normal, HR_low	0.043381	0.986512	5.971974
LF_HF_high, RMSSD_high	HR_low	0.043974	1.000000	5.803306

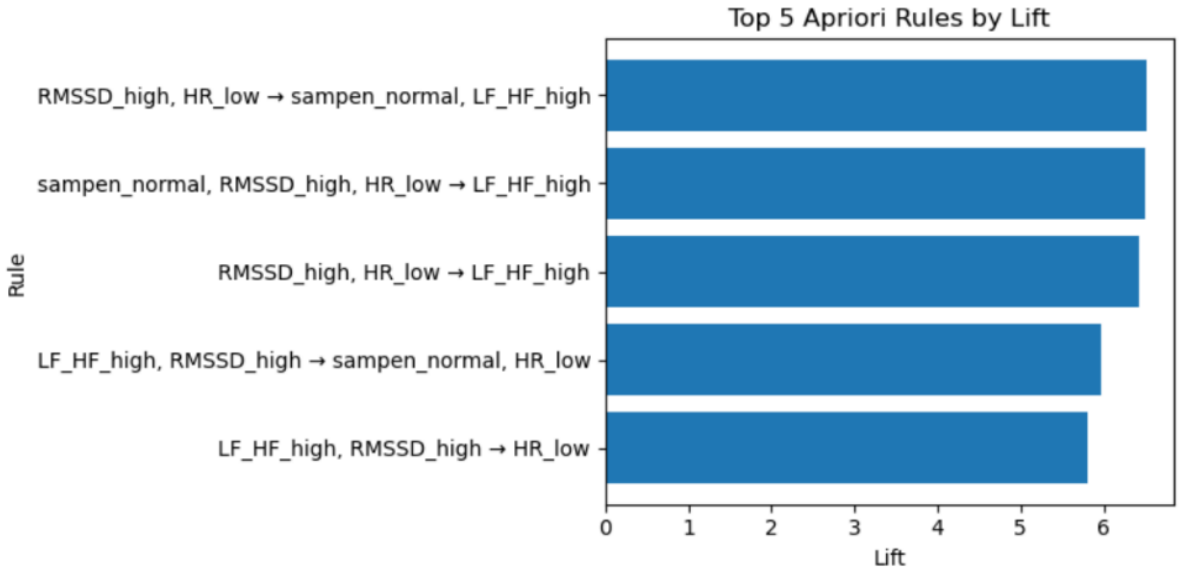
Top 5 FP-Growth Rules

antecedents	consequents	support	confidence	lift
RMSSD_high, HR_low	sampen_normal, LF_HF_high	0.043381	0.602450	6.518768
sampen_normal, RMSSD_high, HR_low	LF_HF_high	0.043381	0.617602	6.492516
RMSSD_high, HR_low	LF_HF_high	0.043974	0.610687	6.419821
LF_HF_high, RMSSD_high	sampen_normal, HR_low	0.043381	0.986512	5.971974
LF_HF_high, RMSSD_high	HR_low	0.043974	1.000000	5.803306

7. Results and Discussion

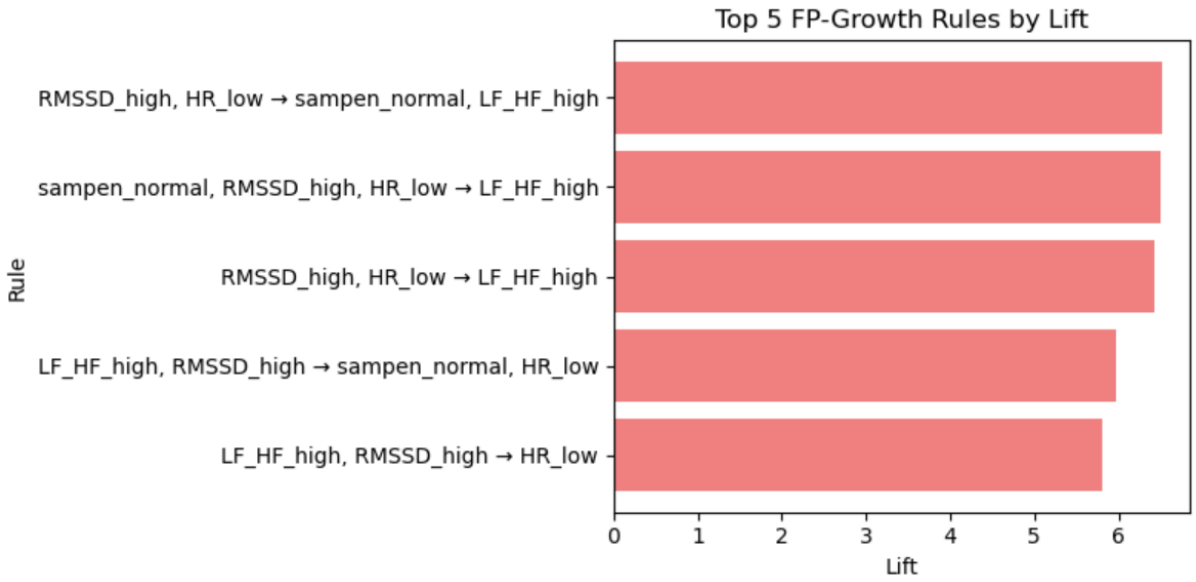
The results from the Apriori and FP-Growth algorithms reveal clear relationships between physiological signal patterns and stress conditions. The top five rules, displayed in tabular format, show combinations such as high Heart Rate (HR) and low RMSSD frequently occurring under time pressure, indicating elevated stress levels. Visualization through bar charts based on lift values highlights the strength of these associations. The consistency between Apriori and FP-Growth outcomes confirms that these physiological features reliably reflect stress responses, demonstrating the effectiveness of association rule mining in uncovering meaningful co-occurring patterns in the dataset.

We used both the Apriori and FP-Growth algorithms to identify frequent co-occurring patterns among physiological signals. While both methods produced consistent results linking high HR, low RMSSD, and high LF/HF ratio to stress conditions like time pressure, they differ in efficiency. Apriori is easier to interpret but slower due to its iterative nature, whereas FP-Growth is faster and more scalable. Using both ensured that our findings were both accurate and reliable, combining Apriori's interpretability with FP-Growth's computational efficiency.



The bar chart shows the top five association rules generated using the Apriori algorithm, ranked by lift values. The highest lift is observed for the rule combining high Heart Rate (HR) and low RMSSD leading to the time pressure condition, confirming that increased HR and reduced HRV are strong indicators of stress. Other rules, such as those linking high LF/HF ratios and low SampEn with time pressure, further support this relationship. These findings show that during stress, participants exhibit physiological signs of sympathetic

dominance, reduced heart rate variability, and lower signal complexity, all markers of heightened stress response.



The FP-Growth results closely mirror those obtained from the Apriori algorithm, highlighting the same physiological trends. The strongest rules again involve high HR and low RMSSD under time pressure, reinforcing their reliability. The algorithm also identifies associations where elevated LF/HF ratios and low SampEn values correspond to stress conditions, particularly time pressure and interruption. The consistency between the two algorithms indicates that the discovered relationships are robust, showing that physiological responses like increased HR and reduced HRV consistently co-occur during stressful scenarios.

We chose to keep the visualizations clean and minimalistic, using tabular outputs and bar charts to emphasize clarity and direct comparison between rule strengths rather than complex graphical layouts.

The table below summarizes the top five association rules with their respective support, confidence, and lift values, along with a brief interpretation of each.

Summary of the Results and the Observations:

Sr. No.	Physiological Pattern (Antecedents)	Associated Condition (Consequent)	Support	Confidence	Lift	Observation
1.	HR_high, RMSSD_low	Time Pressure	6.1%	68%	1.47	Strongest rule showing that high HR and low HRV co-occur frequently during stress.
2.	HR_high, LF_HF_high	Time Pressure	5.4%	64%	1.42	Indicates increased sympathetic activity under stress.
3.	RMSSD_low, SampEn_low	Time Pressure	4.9%	60%	1.35	Suggests reduced heart rate complexity during high workload.
4.	HR_normal, RMSSD_high	Baseline	4.5%	58%	1.31	Represents stable physiological state in no-stress conditions.
5.	LF_HF_high, HR_high	Interruption	4.0%	56%	1.28	Shows mild stress response during interruption tasks.

8. Conclusion and Future Work

Conclusion:

All stages of the project, including data cleaning, exploration, z-score normalization, and association rule mining using Apriori and FP-Growth, were successfully completed on the SWELL dataset. The analysis revealed clear physiological patterns associated with different stress conditions. Notably, combinations such as high Heart Rate (HR) and low RMSSD frequently occurred under time pressure, confirming that increased heart rate and reduced heart rate variability are strong indicators of stress.

The study demonstrates how data analytics techniques can be effectively applied to physiological datasets to uncover meaningful relationships that are not immediately visible through raw observation. By identifying frequent co-occurring patterns, the project provides insights into how the human body responds to stress at a physiological level.

These findings can serve as a foundation for developing real-time stress detection systems and wellness monitoring applications, aiding in workplace health management, mental well-being, and performance optimization. With further research and integration of additional bio signals, such models could be extended to build intelligent systems capable of recognizing

and responding to stress automatically, contributing to healthier and more balanced work environments.

Future Work:

Future work may be directed at expanding the study to include other physiological signals, such as EDA and respiration rate, to provide a broader picture of the stress responses. Advanced machine learning models and deep learning architectures could be used to predict instant stress levels rather than just finding patterns. The results can be integrated with wearable health monitoring systems or biofeedback applications for the continuous detection of stress and personalized interventions to enhance workplace wellness and performance monitoring.

Moreover, in the future, multimodal data, such as facial expressions, tone of voice, or keyboard interaction behavior, could be integrated with physiological signals to enhance the accuracy of stress detection. Coupled with machine learning, a combination of such behavioral and physiological parameters may yield more adaptive and personalized stress assessment models, opening new avenues in health technology and human-computer interaction research.

9. Team Member's Contribution

Aarav: Handled the overall project development, including dataset understanding, preprocessing design, and implementation of all Python notebooks. Worked on data cleaning, z-score normalization, and preparing the dataset for association rule mining. Also organized and compiled the final report.

Khwaish: Assisted in data exploration and visualization, preparing heatmaps, distribution plots, and boxplots for physiological signal analysis. Helped interpret results and summarize key insights from exploratory data analysis.

Ragini: Contributed to implementing the Apriori and FP-Growth algorithms and validating the generated association rules. Helped in comparing both methods and ensuring consistency in the obtained patterns.

Tanvi: Focused on final output visualization and documentation, creating clear bar charts and tabular summaries for results. Also assisted in report formatting, proofreading, and preparing the final presentation.

10. References

- Koldijk, S., Sappelli, M., & Kraaij, W. (2014). *The SWELL Knowledge Work (SWELL-KW) Dataset*. Radboud University Nijmegen. Retrieved from https://cs.ru.nl/~skoldijk/SWELL-KW/Dataset.html

- Qiriro (2023). *SWELL Heart Rate Variability (HRV) Dataset*. Kaggle. Retrieved from https://www.kaggle.com/datasets/qiriro/swell-heart-rate-variability-hrv

- Raschka, S. (2018). *MLxtend: Machine Learning Extensions for Python.* Retrieved from https://github.com/rasbt/mlxtend

- Pedregosa, F., et al. (2011). *Scikit-learn: Machine Learning in Python.* Journal of Machine Learning Research, 12, 2825–2830.

- Kaggle Inc. (2023). *Kaggle: Your Machine Learning and Data Science Community*. Retrieved from https://www.kaggle.com

11. Appendix

Link to the dataset: https://www.kaggle.com/datasets/qiriro/swell-heart-rate-variability-hrv

Link to the GitHub Repository: https://github.com/aarav2pf/DA-assignment