UNIVERSITAT POLITÈCNICA DE VALÈNCIA

_ TELECOM ESCUELA TÉCNICA VLC SUPERIOR DE INGENIERÍA DE TELECOMUNICACIÓN

MMAES – MULTIMEDIA AD EXPOSURE SCALE: MEASURING SHORT TERM IMPACT OF AD EXPOSURE

Javier Martínez Tornero

Mentor: Andrej Košir

Resumen

El objetivo de este trabajo consiste en demostrar que el MMAES es una herramienta válida para medir los efectos a corto plazo de la exposición a anuncios en medios digitales. Dado que los formatos digitales son la forma predominante de medios de comunicación en la actualidad, con individuos expuestos a una variedad de información de diferentes plataformas y fuentes. Medir esta exposición presenta desafíos debido a la falta de consenso e investigación sistemática sobre métodos de medición válidos. Los enfoques tradicionales son propensos a la sobreinformación y pueden no ser adecuados para la exposición en línea. Para abordar esto, se ha desarrollado un nuevo instrumento llamado Escala de Exposición Publicitaria Multimedia (MMAES) para medir los efectos a corto plazo de la exposición a anuncios multimedia en línea. MMAES tiene como objetivo proporcionar mediciones más precisas de la exposición en este ámbito.

Para validar el MMAES, se ha recurrido a técnicas avanzadas de aprendizaje automático, utilizando Python como lenguaje de programación principal. Se hizo uso de la biblioteca scikit-learn, reconocida por su robustez y versatilidad en el análisis de datos. Esta herramienta proporcionó una variedad de algoritmos y métodos de preprocesamiento, permitiendo un ajuste y optimización detallados de los modelos. Se implementaron técnicas de validación cruzada y se evaluaron diversas métricas para garantizar la solidez de los hallazgos. Al analizar los resultados, no solo se observaron métricas estándar como precisión y recall, sino que también se dio especial atención a matrices de confusión y otros indicadores relevantes. Los análisis revelaron que el MMAES es una herramienta confiable y válida para medir la exposición a anuncios en medios digitales, subrayando su eficacia en diferentes contextos y escenarios.

Abstract

The objective of this work is to demonstrate that MMAES is a valid tool to measure the short-term effects of exposure to multimedia advertisements.

Multimedia is the predominant form of media today, with individuals exposed to a variety of information from different platforms and sources. Measuring multimedia exposure poses challenges due to the lack of consensus and systematic research on valid measurement methods. Traditional approaches are prone to over-reporting and may not be suitable for online exposure. To address this, a new instrument called the Multimedia Ad Exposure Scale (MMAES) has been developed to measure the short-term effects of online multimedia ad exposure. MMAES aims to provide more accurate measurements of multimedia exposure.

To establish the validity of MMAES, advanced machine learning techniques were employed, using Python as the primary programming language. The renowned scikit-learn library was utilized for its robustness and versatility in data analysis. This tool provided a range of algorithms and preprocessing methods, allowing for detailed tuning and optimization of the models. Cross-validation techniques were implemented, and various metrics were evaluated to ensure the robustness of the findings. When analyzing the results, not only were standard metrics such as precision and recall observed, but special attention was also given to confusion matrices and other relevant indicators. The analyses revealed that the MMAES is a reliable and valid tool for measuring exposure to advertisements in digital media, underscoring its effectiveness in various contexts and scenarios.

Table of contents

Chapter 1.

Introduction

1.1 Introduction to multimedia exposure

Multimedia refers to the integration of various communication media, including text, graphics, images, sound, video, and animation, all working cohesively to convey information. As the digital landscape has evolved, multimedia has transitioned from being a mere novelty to an essential tool in the realm of digital communication. It offers an interactive user experience that can be tailored to individual preferences, ensuring that information is not just presented, but is also engaging and resonates with the audience. For example, while a written text can delve deep into details and provide context, a video has the power to visually illustrate complex concepts, making them more accessible and relatable to a broader audience.

Multimedia exposure is more than just a passive interaction with content; it's an ongoing, dynamic experience where individuals engage with a myriad of communication media. In today's age, dominated by technology and an ever-evolving digital media landscape, the idea of consuming information through a singular channel seems almost archaic. Whether it's browsing a news website, streaming a television series, or engaging in lively debates on social media platforms, our interactions are seldom limited to one medium. Instead, we find ourselves navigating through an integrated and multifaceted digital experience, often without even realizing the complexity of our media consumption patterns.

The significance of multimedia exposure cannot be understated. It plays a pivotal role in shaping our learning processes and enhancing the efficacy of communication. By strategically combining different media forms, content creators can capture and sustain user attention, even in an environment rife with distractions. This is especially pertinent in sectors like marketing and advertising. In these domains, the power of multimedia is harnessed to craft campaigns that are not only memorable but also have a persuasive edge. Such campaigns leverage the strengths of different media forms, ensuring that the core message resonates with the target audience, ultimately influencing their purchasing decisions and brand perceptions.

1.2 Measurement of multimedia exposure

Measuring multimedia exposure, despite its prevalence, remains challenging due to a lack of systematic research on its validity and reliability. There's no agreed-upon method among researchers for measuring exposure. An analysis of over 200 studies revealed no universally accepted understanding of media exposure and its effects. Traditional methods, like frequency measures and consumer recall-based reports, often result in inaccuracies. The digital evolution has shifted advertising strategies, emphasizing targeted digital advertising. The fragmented media landscape, ever-changing digital platforms, and evolving consumer behaviors further complicate reliable measurement. Based on our review, no existing research offers a suitable psychometric tool for measuring short-term effects of online multimedia ad exposure. Hence, we've developed an instrument focusing on key aspects such as engagement, psychological reactance, attitude, awareness, memory, and purchase intention.

1.3 The problem of automated multimedia ad exposure by measuring signals

Traditionally, multimedia exposure was gauged through surveys, self-reports, and potential ecological measures. However, these methods often faced challenges in accuracy and comprehensiveness. With the advent of technology and advanced analytics, the measurement has evolved to capture real-time data, track user interactions, and even assess physiological responses to media content.

The MMAES (Multimedia Ad Exposure Study) takes a comprehensive approach to measure multimedia exposure. Instead of relying solely on self-reported data, the study integrates psychophysiological measurements. This means that beyond just asking participants about their media interactions, the study observes and records physiological reactions, such as heart rate, skin conductance, and eye movements, to gauge genuine engagement and response to multimedia content.

The MMAES study recognizes that in today's multifaceted digital environment, individuals often have simultaneous multimedia experiences. For instance, a person might watch a video while reading real-time comments and participating in a related social media discussion. Such integrated experiences are central to the MMAES's measurement approach.

Furthermore, the study introduces the MMAES instrument, a tool designed to capture key aspects of online multimedia ad exposure. This includes factors like engagement, psychological reactance, attitude, awareness, memory, and purchase intention.The connection between questionnaire information and physiological reactions lies in the ability to correlate behavioral and cognitive responses with physical reactions to multimedia ad exposure. By merging this data, we achieve a holistic view of individual reactions to ads. While the questionnaire may reflect conscious perception, physical reactions, such as heart rate variability or pupillary dilation, can unveil unconscious responses. The latter serve as objective indicators of emotional and cognitive activation. Analyzing both responses allows us to discern the coherence between what an individual expresses and what they truly feel. For instance, an individual might indicate that an ad did not affect them, but their physical reactions could reveal heightened emotional activation. In summary, by integrating questionnaire responses with physiological reactions, a multidimensional perspective of the experience with multimedia ads is obtained, enriching our understanding of advertising impact on the individual

Chapter 2.

State of art

2.1 Multimedia exposure and ads

In the digital age, multimedia exposure and targeted advertising have become increasingly significant. As consumers navigate through various online platforms, they are exposed to a plethora of multimedia content, including targeted advertisements. These ads, tailored based on user's behavior, preferences, and demographics, aim to capture attention and provoke a desired response, such as a purchase or a click-through.

Multimedia exposure refers to the extent to which consumers come into contact with multimedia content, including text, images, audio, and video across different platforms. This exposure can be passive, such as viewing a video, or active, such as interacting with an online game. The nature and level of multimedia exposure can significantly influence a consumer's perception and response to targeted ads.

Targeted advertising, on the other hand, is a marketing strategy that involves segmenting the audience based on certain characteristics and delivering personalized ads to each segment. This personalization can be based on various factors, including demographic information, browsing history, purchase behavior, and even psychographic factors like interests and attitudes. The goal is to deliver ads that are relevant and engaging to each individual, thereby increasing the likelihood of a positive response.

The intersection of multimedia exposure and targeted advertising is a dynamic space where marketers strive to optimize their strategies to reach and engage their target audience. By understanding the nuances of multimedia exposure and leveraging the power of targeted advertising, marketers can create more effective campaigns that resonate with their audience and drive desired outcomes.

2.2 Measurement of MM exposure

The traditional metrics of advertising exposure, such as viewability and frequency, are no longer sufficient. The evolution of online behavior necessitates a deeper understanding of consumer engagement that goes beyond these basic metrics. Modern digital platforms allow for tracking nuanced behaviors like browsing patterns, social media interactions, and even attitudes and purchase intentions. However, this digital transition presents challenges, especially in defining and consistently measuring online multimedia ad exposure. While some studies have ventured into this domain, many remain tethered to traditional marketing metrics or methods not tailored for the digital environment.

The Multimedia Ad Exposure Scale (MMAES) emerges as a solution to these challenges. Designed to holistically measure online multimedia ad exposure, the MMAES encompasses key aspects such as engagement, psychological reactance, attitude, awareness, memory, and purchase intention. Engagement, a multifaceted concept, is pivotal in the digital advertising landscape, while reactance underscores the potential backlash from aggressive advertising. Other elements like attitude, brand recall, and purchase intent are integral to comprehending online consumer behavior.

The creation of the MMAES questionnaire was methodical. It began with a conceptual phase rooted in literature reviews and expert consultations, notably with The Nielsen Company. This phase pinpointed essential dimensions for online multimedia ad exposure measurement. The subsequent operational phase crafted items for these dimensions, leveraging existing validated tools and integrating bespoke items based on expert insights. This questionnaire was then field-tested on young American adults (18-24 years) using the Clickworker platform. Post exposure to video content with embedded ads, participants tackled the 55-item questionnaire. Their feedback was instrumental in refining and validating the MMAES, ensuring its efficacy in measuring short-term online multimedia ad exposure.

Upon its completion, the MMAES comprised 65 items. However, the final version was streamlined to four core components with 31 items: Ad Engagement (AE), Reactance (RE), Awareness and Attitude (AA), and Purchase Intention (PI). AE evaluates attention and emotional responses to ads, RE gauges negative sentiments, AA measures brand likability, and PI assesses buying intent. During its refinement, 12 items were discarded due to various reasons, including low factor loadings and conceptual relevance. An Exploratory Factor Analysis validated the MMAES's structure, and component correlations revealed both related and independent dimensions. In essence, the MMAES offers a modern, comprehensive tool to navigate the intricate landscape of online multimedia advertising exposure.

2.3 Psycho physiological signals and sensors

In this research, the integration of psychophysiological signals and sensors with the questionnaire plays a pivotal role in comprehensively assessing the short-term effects of multimedia advertisements on individuals. Participants are exposed to multimedia advertisements in a controlled environment, during which sensors, including Empatica sensors measuring heart rate (HR), electrodermal activity (EDA), and the Accx signal (accelerometer data), as well as Tobii sensors tracking pupil diameter, are employed to monitor their physiological responses. Following this exposure, participants complete the MMAES (Multimedia Advertisement Experience Scale) questionnaire, designed to capture their conscious perceptions and responses to the advertisements. The data collected from these multiple sources are then harmoniously integrated and analyzed. This integration allows for a unique exploration of the correlation between participants' physiological reactions, reflective of unconscious responses, and their conscious perceptions as revealed in the questionnaire. By intertwining these data streams, we gain a multidimensional understanding of participants' experiences with multimedia advertisements, shedding light on both their emotional and cognitive reactions, thereby enriching our insights into the impact of advertising on individuals. This combined approach proves invaluable in unraveling the intricate dynamics between multimedia advertisements and the human psyche.

Sensors

Empatica and Tobii are two companies that specialize in producing advanced sensor technologies for specific applications. Empatica is known for its wearable physiological sensors designed for health and wellness monitoring, including sensors that measure Electrodermal Activity (EDA), Heart Rate (HR), and accelerometer data (Accx). These sensors provide valuable insights into emotional and physiological responses. On the other hand, Tobii specializes in eye-tracking technology, which is widely used in market research, user experience (UX) research, and psychological research to track eye movements and gaze patterns. These technologies contribute to a deeper understanding of human behavior and play a crucial role in studying multimedia ad exposure.

2.4 Classification algorithms

2.4..1 Basics

Classification is a supervised machine learning approach that focuses on predicting the correct label for a given input data point. This process involves training a model on a labeled dataset to understand the underlying patterns and relationships between input features and their corresponding labels. Once trained, the model can be evaluated on a separate test dataset to assess its performance before using it to make predictions on new, unseen data.

There are two types of classification, binary classification, if the problem has only two possible outcomes, and multiclass classification if there are more than two. In binary classification tasks, the objective is to classify input data into one of two exclusive categories. The training dataset for binary classification contains labels that are binary in nature, such as true/false, positive/negative, 0/1, spam/not spam, etc., depending on the specific problem being addressed. For instance, binary classification could be used to determine whether a given image represents a truck or a boat.

There are two main categories of machine learning classification algorithms: eager and lazy learners. Eager learners are algorithms that construct a model during the training phase before making predictions. While they invest more time in training to achieve better generalization, they require less time for making predictions. Examples of eager learners include Logistic Regression, Support Vector Machines, Decision Trees, and Artificial Neural Networks. The algorithm that is used in this project is the random forest classifier, an example of eager learner.

2.4..2 Random Forest Classifier

Random Forest Classifier is a powerful ensemble machine learning algorithm used for classification tasks. It leverages the strength of decision trees by constructing a forest of diverse trees through bootstrapping and feature selection. Each tree independently predicts the class of data points, and the final classification is determined by majority voting.

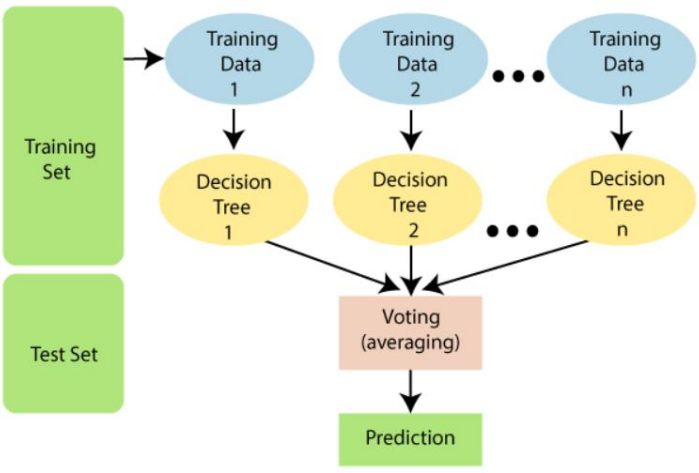

Figure 2.1: Random Forest algorithm

Since the random forest combines multiple trees to predict the class of the dataset, it is possible that some decision trees may predict the correct output, while others may not. But together, all the trees predict the correct output.

This approach leads to high predictive accuracy, reduced overfitting, and the ability to handle various classification problems. Random Forests are particularly valuable when dealing with

complex datasets, making them a popular choice across multiple domains, including healthcare, finance, natural language processing, and image recognition. They also provide insights into feature importance, aiding in feature selection and interpretation.

2.5 Evaluation

After finishing the development of our model, evaluating its performance is crucial.For evaluating a Classification model, the following methods can be used:

2.5..1 Confusion matrix

It is a performance measurement for machine learning classification problem where output can be two or more classes. It is a table with 4 different combinations of predicted and actual values. The confusion matrix itself is relatively simple to understand, but it can provide a lot of insights into the classification performance.

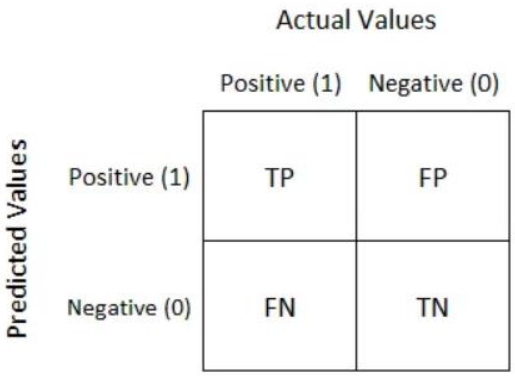

Figure 2.2: Confusion Matrix

True positive (TP): The number of true positives is the number of instances that were actually positive and were correctly classified as positive by the model.

True negative (TN): The number of true negatives is the number of instances that were actually negative and were correctly classified as negative by the model.

False positive (FP): The number of false positives is the number of instances that were actually negative but were incorrectly classified as positive by the model. This is also known as a "Type I error."

False negative (FN): The number of false negatives is the number of instances that were actually positive but were incorrectly classified as negative by the model. This is also known as a "Type II error."

The confusion matrix also helps computing other performance metrics. By considering precision, recall, and the F1-score together, you can gain a more comprehensive understanding of a model's performance, particularly in situations where one metric alone might not provide the full picture. The goal is to strike a balance between making accurate positive predictions (high precision) and capturing as many positive instances as possible (high recall) to achieve the best overall model performance.

Precision

Precision is a metric that aims to solve the question: out of all positive outcomes given by the model what percentage were correct? It is calculated using this expression:

$$Precision = \frac{TP}{TP + FP}$$

Recall

It can be considered as the flip of precision. It's aim is to answer the question: how good is the model at predicting positive outcomes? It is calculated using this expression:

$$Recall = \frac{TP}{TP + FN}$$

F1-score

The F1-score is a single metric that combines both precision and recall to provide a balanced view of a classification model's performance. It's especially useful for comparing models when both false positives and false negatives are important considerations.

$$F1 \cdot score = \frac{2 \times Recall \times Precision}{Recall + Precision}$$

It uses Harmonic Mean in place of Arithmetic Mean by punishing the extreme values more.

AUC-ROC curve

The AUC-ROC curve serves as a performance metric for classification tasks across different threshold configurations. ROC, a probability curve, signifies the level of separability in the model's predictions. It quantifies the model's ability to differentiate between classes: the higher the AUC, the more adept the model is at correctly identifying instances of class 0 as 0 and class 1 as 1. In analogy, a higher AUC implies that the model excels in distinguishing individuals with a medical condition from those without it, highlighting its discriminatory power.

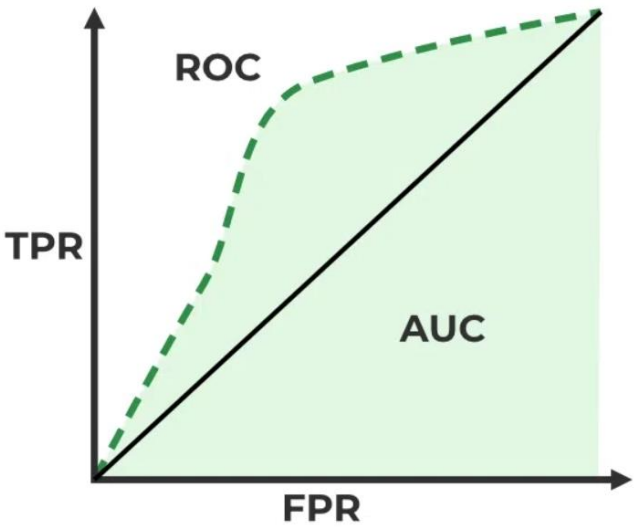

Figure 2.3: ROC curve

In order to know the model's performance based on AUC results, it's essential to understand that a highly effective model typically exhibits an AUC close to 1, indicating a strong separability measure. Conversely, a less effective model tends to yield an AUC close to 0, signifying poor separability. In essence, a low AUC suggests that the model is inversely predicting results, swapping 0s for 1s and vice versa. When the AUC hovers around 0.5, it indicates that the model lacks any discernible capacity for class separation.

2.5..2 Folding

Cross-validation, often referred to as folding, is a fundamental technique in machine learning used to assess and optimize model performance. It involves dividing a dataset into K subsets or folds, where each fold is used as a validation set while the remaining folds are used for training. This process is repeated K times, allowing for comprehensive evaluation and estimation of the model's generalization ability. By calculating summary statistics of performance metrics across iterations, such as mean and standard deviation, cross-validation helps in selecting the best-performing model and tuning hyperparameters. This technique ensures that machine learning models are robust and capable of making accurate predictions on unseen data, minimizing issues like overfitting and underfitting.

K-fold cross-validation is a data partitioning technique which splits an entire dataset into k groups. In the context of machine learning, a fold represents a set of rows in a dataset, and the term "k-folds" describes the number of groups into which the data is divided. For example, in a dataset with 20 rows, it can be divided into 2 folds with 10 rows each, 4 folds with 5 rows each, or 10 folds with 2 rows each. The process of k-fold cross-validation involves several steps:

- The entire dataset is randomly split into equally-sized, independent k-folds, ensuring that no rows are reused across folds.

- We use k-1 folds for model training, and once that model is complete, we test it using the remaining 1 fold to obtain a score of the model’s performance.

- This process is repeated k times, resulting in k models and performance scores.

- Lastly, we take the mean of the k number of scores to evaluate the model’s performance.

There are various forms of K-fold cross-validation, with this study employing "stratified k-fold," which aims to ensure that each fold maintains the same distribution of categorical values, such as class outcomes, to enhance the reliability of the assessment.

Stratified

Stratified K-fold cross-validation is an enhancement of the traditional K-fold cross-validation method, designed to address class imbalance in datasets. It partitions the dataset into 'K' subsets while preserving the original class distribution in each fold. In each iteration, one fold is used for validation, and the remaining folds form the training set. This process is repeated 'K' times, allowing for robust model evaluation and hyperparameter tuning. Stratified K-fold cross-validation ensures that each class is represented proportionately, making it a valuable technique for assessing machine learning models when dealing with imbalanced data, ultimately leading to more reliable performance estimates and model selection.

Chapter 3.

Materials and methods

In this section, we outline the materials used and the methodology employed to achieve the objectives of this study. The materials encompass both the tools utilized for data collection and the software for data processing and analysis. Additionally, the methodology describes the step-by-step approach taken to gather data, preprocess it, and apply machine learning techniques to derive meaningful insights. The combination of these materials and methods constitutes the foundation for conducting a comprehensive analysis of short-term effects of multimedia advertisements on digital platforms.

3.1 Test data

3.1..1 Subjects

The target user group for the observational study were young adults (ages 18-24) who are native English speakers living in the U.S. This age group represents the largest segment of digital natives who engage with online multimedia, particularly ad-supported video streaming.

A total of 360 participants (62.9% female) took part in the study, which lasted nine days. Responses from 60 participants were collected for each of the six multimedia combinations.

98.6% of participants were within the specified age range. Eleven participants were excluded (five were older than 24 years, and six failed the control tests), resulting in data from 349 participants being used for further analysis.

3.1..2 Sensor and signals

A combination of physiological sensors to capture participants' responses during exposure to multimedia advertisements. Two primary sensors, the Empatica E4 wristbands and Tobii eye-tracking devices, were employed to monitor specific physiological signals indicative of emotional and cognitive responses.

The Empatica E4 wristbands provided three key signals: Accelerometer (Accx), Heart Rate (HR), and Electrodermal Activity (EDA). The accelerometer data allowed us to detect participants' physical activity levels and movement patterns during advertisement viewing. The heart rate signal provided insights into participants' cardiovascular responses, reflecting changes in arousal and emotional engagement. The electrodermal activity signal, also known as skin conductance, offered information about participants' emotional responses and levels of physiological arousal.

On the other hand, the Tobii eye-tracking devices measured the diameter of participants' pupils. Pupillary dilation is a well-established indicator of cognitive load and attention, reflecting the cognitive processing effort required during advertisement exposure.

By utilizing these sensors and capturing signals such as Accx, HR, EDA, and pupil diameter, we were able to gather objective physiological data that complemented participants' subjective responses collected through the MMAES questionnaire. This comprehensive approach allowed us to gain deeper insights into participants' emotional and cognitive reactions to multimedia

advertisements and facilitated the integration of objective and subjective data for a more holistic analysis.

3.1..3 Data collecting and preprocessing

The data was collected from the participant's responses to the initial set of multimedia ad exposure items used in the observational study. First participants were given a brief description of the study's purpose and duration (less than 15 minutes). They were instructed to set an appropriate volume on their computer to fully experience the multimedia content and to pay attention to it. Then participants were presented with a scenario where they are at home watching videos in a relaxed manner when they come across a short video including an ad. They were informed that a control number would be presented at a certain point in the video, which they would need to remember for a control question later.

Study Phases

- Pre-questionnaire: Participants answered questions about demographics, technology experience, and mood assessment.

- Multimedia Phase: Participants watched one of the video/ad combinations, which were randomly assigned.

- Main Survey: Participants' responses were recorded in relation to the multimedia combination they were assigned in the previous phase. They responded to the initial set of 55 ad exposure items on a Likert scale. These responses were later used in the exploratory factor analysis.

In the post-survey participants were asked about their habits, assessed their current mood, and shared their overall impression of the entire survey in writing.

To detect and eliminate inattentive or unreliable participants, a series of control questions on content, consistency, and attention were asked. For instance, they were asked about the number shown in the video clip and the content of the commercial.

Once the study phase is complete and the data has been extracted from the participants, there begins the code writing phase. Responses to each video were stored in four separate Excel tables, which will be loaded into python for further anlysis. These tables are organized based on which of the four videos the participants watched. In each table, we have the participants' responses to each item, as well as their scores for the following factors: Ad Engagement (AE), Reactance (RE), Awareness and Attitude (AA) and Purchase Intention (PI).

uID	uID.1	Q2a	Q2b	Q2c	Q2d	Q2e	Q2f	Q2g	Q2h	Q2i	...	Q2p	Q2q	Q2r	Q2s	Q2t	AE	RE	AA	PI	BoxID
1.0	1.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	3.525	3.087	2.094	3.462	1
2.0	2.0	3.0	4.0	4.0	4.0	3.0	3.0	3.0	2.0	3.0	...	3.0	3.0	4.0	3.0	2.0	1.000	1.895	2.262	2.849	3
3.0	3.0	2.0	4.0	1.0	2.0	3.0	1.0	3.0	2.0	3.0	...	3.0	2.0	3.0	3.0	2.0	2.559	3.274	3.057	3.420	3
4.0	4.0	4.0	4.0	3.0	5.0	4.0	1.0	1.0	1.0	4.0	...	5.0	5.0	4.0	5.0	1.0	4.337	3.762	1.000	3.621	3
5.0	5.0	2.0	5.0	3.0	2.0	3.0	2.0	4.0	1.0	4.0	...	5.0	4.0	5.0	4.0	4.0	3.099	2.490	2.337	3.215	2

Figure 3.1: Answers and scores stored

The final scores for this components are done by calculating the average of the summed scores for the items in each component. As we can see in the figure above, the answers are stored for each user ID (the first column), the last four columns without the 'BoxID' column represent the four components we mentioned before. The value of this factors for each user is calculated by averaging the score for the previous answers.

In addition to the participant responses, raw data is collected through sensors used during the study. Similar to the Excel tables, the raw data is organized into four different folders, each

corresponding to one of the four videos. Within each of these folders, the data is further organized based on the type of sensor used: Empatica or Tobii.

- When the Empatica sensor is used, we find three different signals: Accx (Accelerometer X-axis), EDA (Electrodermal Activity) and HR (Heart Rate).

- When the Tobii sensor is used, we only have one signal: Diameter (Pupil Diameter).

Each signal that we have stored is assigned a unique index. This index corresponds to the user to whom the measurements belong. Each user has different measurements for each signal, and this index ensures that the data for each user is kept separate and identifiable.

Raw data files, which are presumably in CSV format, are loaded into Pandas Data Frames. The file paths are constructed based on various parameters such as user ID, sensor type, signal type, and advertisement string. (e.g., 'C1', 'C2', etc.). This step is essential to begin the data analysis. It involves reading the raw data files and converting them into a structured format (like a table) that Python can manipulate.

```
data_df = pd.read_csv(os.path.join(sig_path, file_name))
```

Figure 3.2: Code. Loading data.

Where ‘sig_path’ refers to the location of the file and ‘file_name’ to the exact name of the file.

If a file does not exist for a given user ID (indicating that data for that user is missing), NaN (Not a Number) values are assigned to the corresponding features for that user. This ensures that the dataset remains consistent in structure, even when some data is missing. It allows for easier handling of missing data in later analysis stages.

After loading data and assigning Nan values for those IDs which don’t have data, we have to create the features we are going to work with and structure them into a DataFrame. Feature engineering is a pivotal process in data preprocessing for machine learning, transforming raw data into a format more amenable to modeling. Especially in the context of sensor data, raw readings can be noisy and may contain extraneous information. Through feature engineering, such as calculating the mean or standard deviation of a signal, we can distill the most pertinent information from the data. For instance, the mean of a signal provides insight into its central tendency, offering the model a glimpse into the typical value of a signal under various conditions. Similarly, the standard deviation captures the variability of the signal, which can be indicative of specific states or conditions. By crafting these features, not only is the model's performance potentially enhanced, but the data also becomes more interpretable, aiding in understanding the underlying patterns and relationships. Other relevant metrics that are extracted and used in the DataFrame are skewness, kurtosis, and features based on the Gramian Angular Field (GAF) such as energy and mean.

In conclusion, after meticulously aggregating physiological signals from diverse sensors across multiple advertisement categories for a cohort of users, we've successfully constructed a comprehensive DataFrame. This DataFrame, derived from preprocessed CSV files, encapsulates a rich set of statistical attributes, including mean, standard deviation, skewness, kurtosis, and advanced features from the Gramian Angular Field transformation. By systematically structuring this data, we've laid a robust foundation for the subsequent phases of our analysis, ensuring that the raw data is now in an organized, accessible, and analyzable format.

3.1..4 Labelling and thresholding

Labeling and thresholding are foundational in data preparation for classification in machine learning. Labeling assigns a discrete category to each dataset instance, guiding the model's predictions. When the target variable is continuous but the task is classification-based, thresholding sets cut-off values to categorize the variable. In the code, a threshold T1 is employed:

values in y_cont below T1 are labeled as 0, and others as 1, converting the continuous variable into a binary label. The threshold choice, influenced by the data's distribution visualized in a histogram, is pivotal, affecting both model predictions and its interpretability.

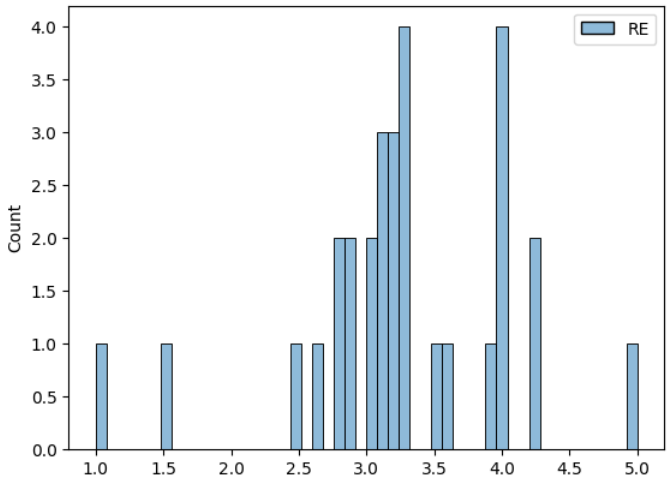

Figure 3.2: RE histogram

The figure 3.2 shows the histogram of the label RE and is used as a visual tool to help determine an appropriate threshold value for converting a continuous variable into binary labels (0 or 1). By examining the histogram, you can identify points where there is a natural separation in the data. These points are where the frequency of the data changes significantly, indicating different 'groups' or 'clusters' within the data. A good choice for this example can be selecting a threshold of value 3.4 which clearly separates the data into two groups.

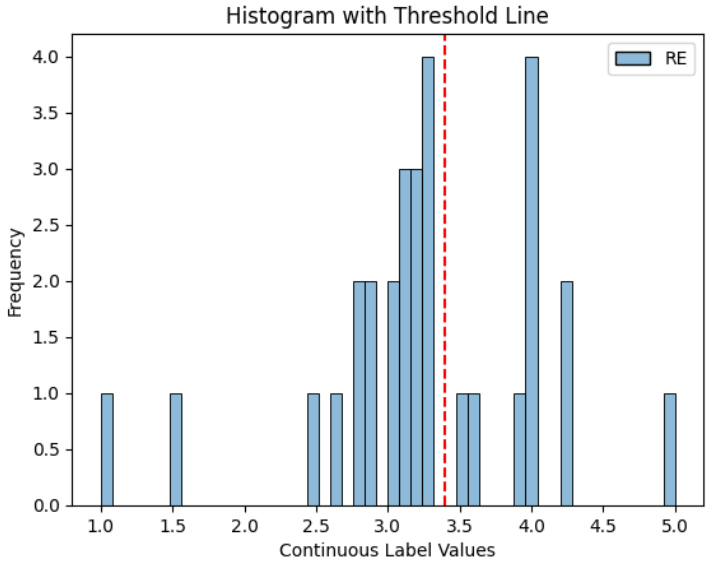

Figure 3.3: RE histogram threshold added

After setting a threshold based on the histogram, it is important to validate this choice. This could involve checking how well the chosen threshold separates the classes in practice, and potentially adjusting the threshold based on further analysis or domain knowledge.

The histogram, along with the chosen threshold, provides a visual and quantitative way to understand how the continuous data is being split into two categories. It allows you to see where

the majority of the data points lie relative to the threshold and how balanced or imbalanced the two resulting classes are.

3.2 Classification

3.2..1 Random forest classifier

In this analysis, a Random Forest Classifier is employed as the machine learning model. The Random Forest Classifier is an ensemble learning method that operates by constructing a multitude of decision trees during training and outputs the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. It is known for its flexibility and is widely used for classification tasks due to its ability to handle large data sets with higher dimensionality and its ability to model complex relationships. In this code, the Random Forest Classifier is not only used for classification but also for feature selection. During the feature selection step, the classifier computes the importance of each feature, which is a score indicating how useful or valuable each feature was in the construction of the decision trees within the model. Features with an importance score above a certain threshold (0.01 in this case) are retained for further analysis. This process of feature selection helps to reduce the dimensionality of the data, potentially improving the model's performance by focusing on the most informative features.

3.2..2 Cross-validation

After this process of feature selection, it continues with the cross-validation method that is used to evaluate the machine learning model. For each split of the data in the cross-validation process, the Random Forest Classifier is retrained on the training set (using the selected features) and is used to make predictions on the testing set. These predictions are in the form of probabilities, indicating the likelihood that a given data point belongs to a particular class.

In this case a k-fold cross-validation method is used, where the dataset is divided into 'k' equal-sized folds or subsets. The model is trained on k-1 of these folds and tested on the remaining one. This process is repeated k times, with each of the k folds used exactly once as the validation data. The k results from the folds can then be averaged to produce a single performance score, providing a more robust and comprehensive assessment of the model's performance.

Repeated Stratified K-Fold cross-validation is the method used in the code, it repeats the stratified k-fold cross-validation multiple times, ensuring that the class distribution is preserved in each fold and providing a more reliable estimate of model performance. This approach is particularly beneficial as it ensures that every observation from the original dataset has the chance of appearing in the training and test set, which is important for a reliable performance estimate.

3.2..3 ROC curve and AUC score

After training the Random Forest Classifier on a subset of the data (the training set), the model is used to predict the probabilities of the testing set instances belonging to a particular class. These predicted probabilities are then used to compute the ROC curve and the AUC score for each split in the cross-validation process.

The ROC curve is plotted with False Positive Rate (FPR) on the X-axis and True Positive Rate (TPR) on the Y-axis, for various threshold values. For each split of the data, an individual ROC curve is plotted, and the AUC is calculated as the area under this curve. The AUC provides a summary measure of the model's ability to distinguish between the two classes, independent of the threshold.

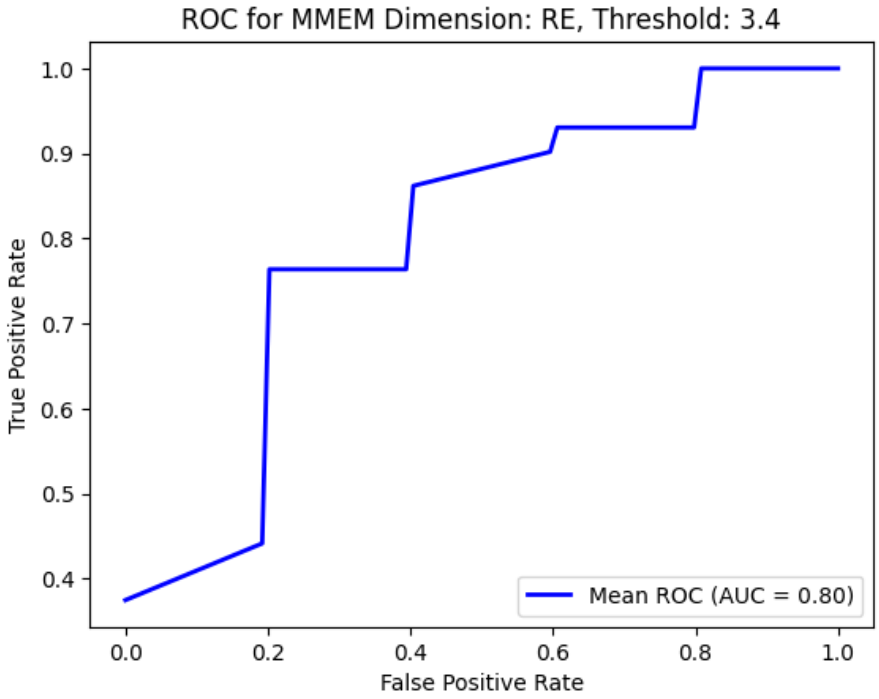

Figure 3.3: ROC curve

In the final step, the mean of the TPRs at each FPR value is calculated to produce a mean ROC curve, and the mean of these AUC values is computed to give an overall summary of the model's performance across all the splits.

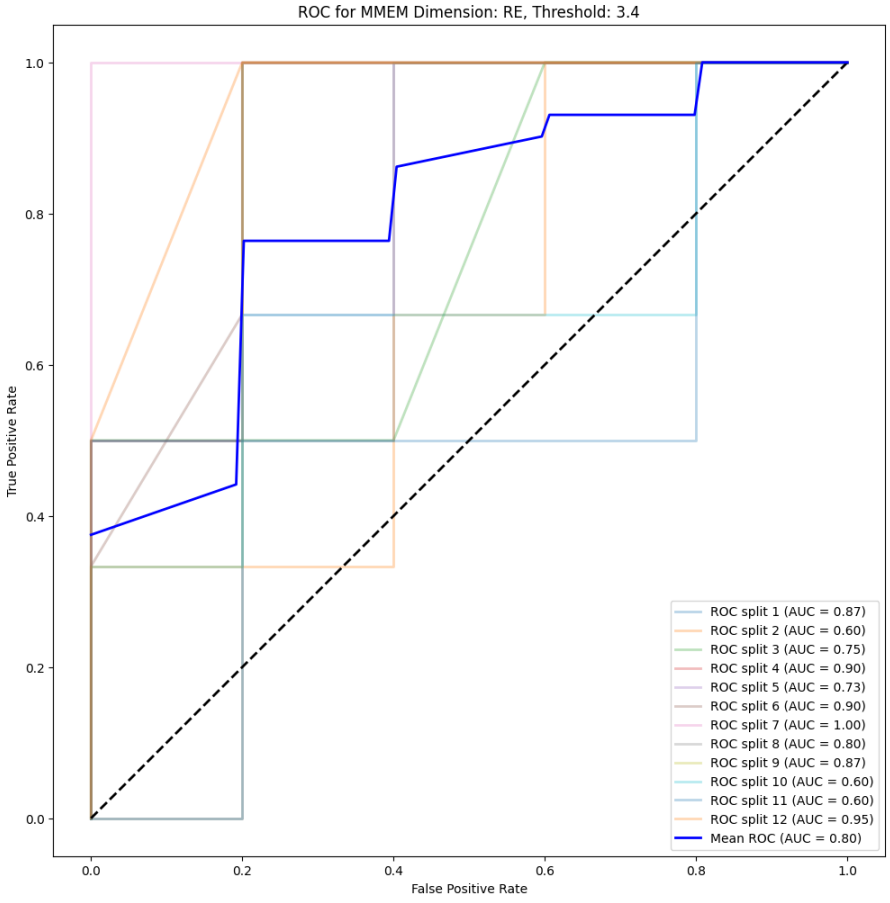

Figure 3.4: ROC curve of all splits

This process allows for a robust and comprehensive assessment of the model's classification performance, as it considers the performance across different splits of the data, thereby providing a more reliable estimate.

The Random Forest Classifier, with its ensemble of decision trees, is particularly well-suited for this task as it can model complex, non-linear relationships in the data, and is generally robust to overfitting, making it a strong choice for this analysis.

Chapter 4.

Experimental results

In this section, we delve into the outcomes of our analysis, shedding light on the performance and insights derived from our machine learning model. The results presented here are a culmination of rigorous data preprocessing, feature engineering, and model training, all aimed at understanding and predicting the specific dimensions: Awareness and Attitude (AA), Ad Engagement (AE), Reactance (RE), and Purchase Intention (PI). For each factor, we highlight the performance of our machine learning model in predicting specific dimensions through physiological signals. We'll present our findings in two main subsections, each focusing on distinct performance metrics:

4.1 Classification of Awareness and attitude

- **Overall confusion matrix, precision, recall, f1**

For the factor Awareness and Attitude the model correctly predicted 8 negative instances and 12 positive instances, while mistakenly identifying 6 negative instances as positive and missing 4 positive instances as we can see illustrated in figure 4.1.

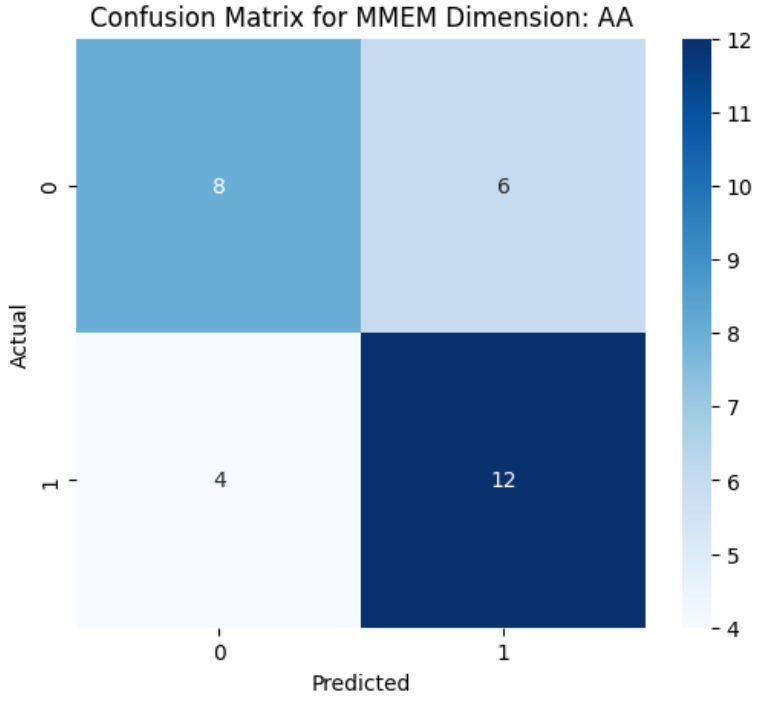

Figure 4.1: Confusion matrix for AA

This data from the confusion matrix translates to a precision of 66.67%, meaning 66.67% of its positive predictions were accurate. The model's recall is 75%, indicating it identified 75% of all true positive samples. The F1-score, which balances precision and recall, is 70.59%, suggesting a reasonably good balance between the two metrics.

Metric	F1	Recall	Precision
Value	0.666667	0.75	0.705882352

Table 4.1: Results AA

- **ROC curve and AUC**

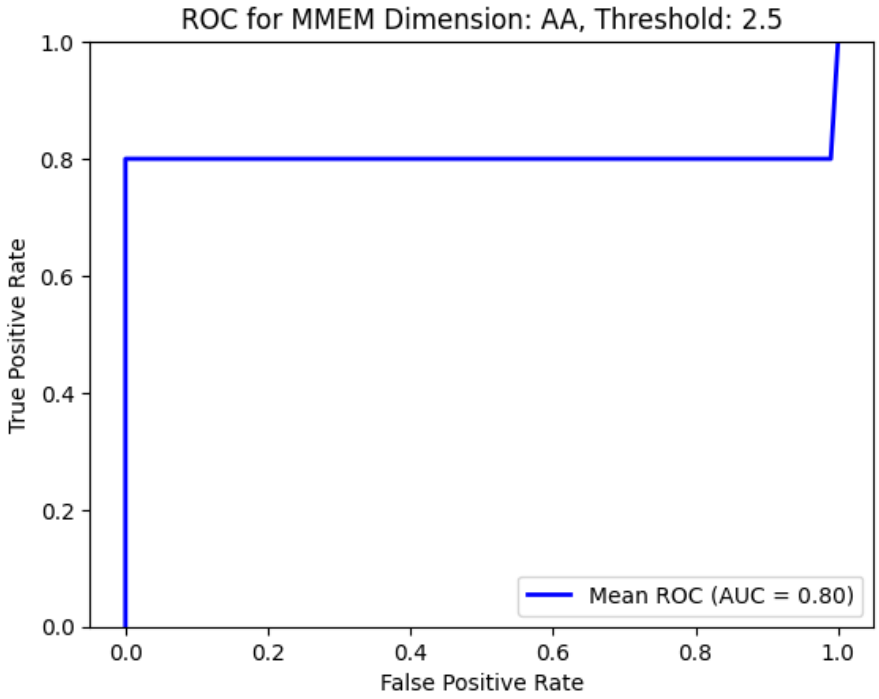

Figure 4.2: ROC curve for AA

An AUC of 0.8 on the ROC curve indicates that the model has an 80% probability of correctly distinguishing between positive and negative instances of awareness and attitude. This suggests a commendable predictive capability, as the model performs significantly better than random guessing (which would have an AUC of 0.5).

4.2 Classification of Reactance

- **Overall confusion matrix, precision, recall, f1**

For the Reactance factor, the confusion matrix with values 13, 3, 6, and 8 reveals the following: 13 True Negatives where the model accurately predicted the absence of reactance, 8 True Positives where reactance was correctly identified, 3 False Positives where the model incorrectly predicted the presence of reactance, and 6 False Negatives where it failed to detect actual reactance. This suggests that, while the model is fairly accurate in identifying non-reactance, it struggles with both overestimating (false positives) and underestimating (false negatives) true reactance instances.

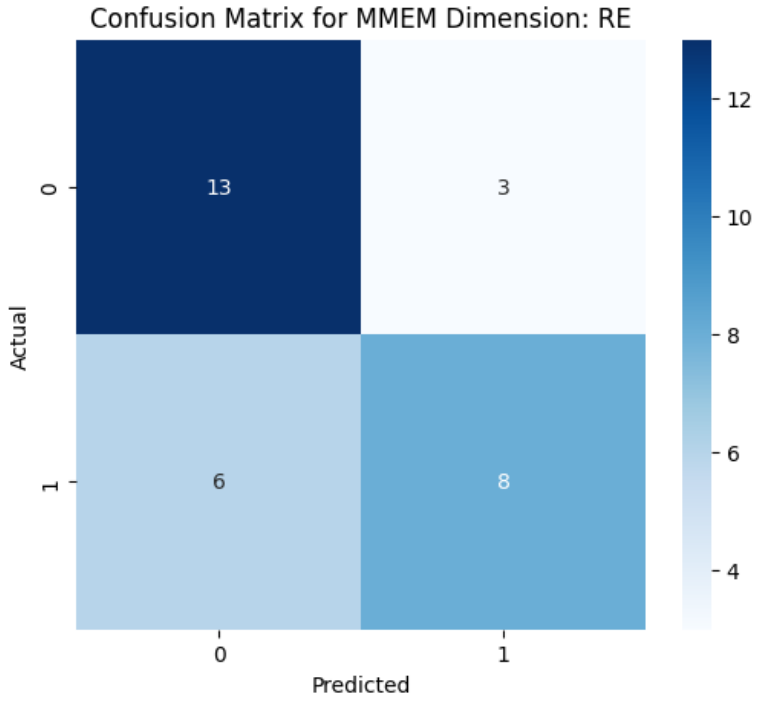

Figure 4.3: Confusion matrix for RE

Based on the provided confusion matrix for the Reactance factor, the model has a Precision of approximately 72.73%, indicating that about 72.73% of all instances predicted as reactance were correctly identified. The Recall stands at approximately 57.14%, meaning the model captured about 57.14% of all true reactance instances. The F1-Score, which balances precision and recall, is approximately 64.00%, suggesting a reasonable balance between the two metrics, but highlighting potential areas for improvement.

Metric	**F1**	**Recall**	**Precision**
Value	0.64	0.5714285714285714	0.7272727272727273

Table 4.2: Results RE

The performance metrics for the Reactance factor appear to be inferior compared to the preceding factor. Several underlying factors could account for this discrepancy. A primary consideration is the potential imbalance in the dataset. Establishing an appropriate threshold to binarize the 'y' signal in the code can be challenging, especially when the data distribution is skewed. This imbalance can significantly influence the model's predictive accuracy, leading to

suboptimal results. It's imperative to address such data irregularities to enhance the model's performance for the Reactance factor.

- **ROC curve and AUC**

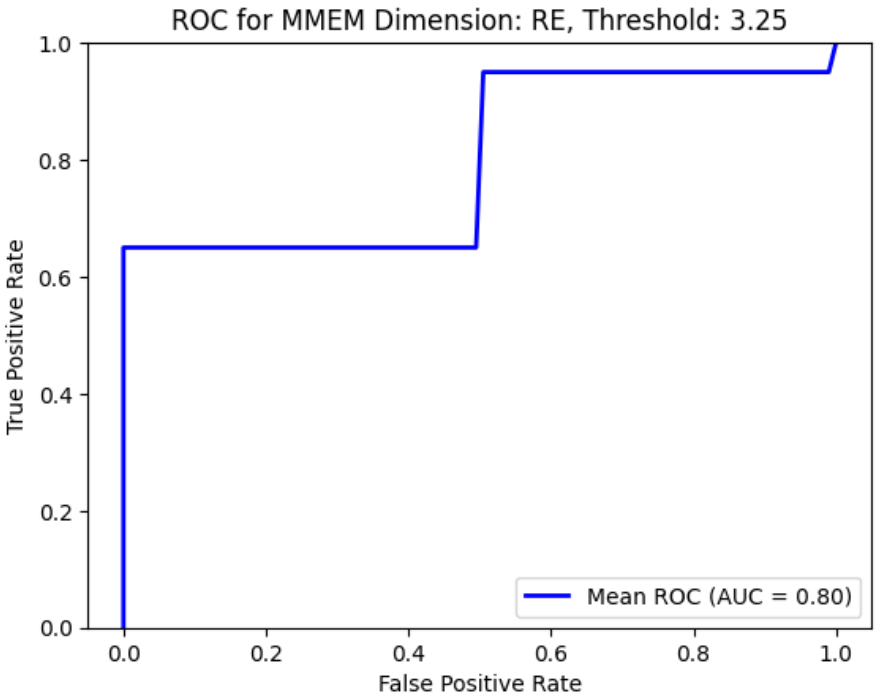

Figure 4.4: ROC curve for RE

For the Reactance factor, an AUC score of 0.8 on the ROC curve signifies a strong ability of the model to differentiate between the two classes (presence or absence of reactance). An AUC of 0.8 implies that there's an 80% chance that the model will correctly distinguish between a randomly selected positive instance and a randomly selected negative instance.

4.3 Classification of Purchase Intention

- **Overall confusion matrix, precision, recall, f1**

Analyzing the results for purchase intention, we find that the model has correctly spotted 10 instances where the intent was lacking and 12 where it was evident. On 4 occasions, it anticipated purchase intent where none existed, and conversely, it was oblivious to 4 instances brimming with intent. This pattern suggests that while the model grasps the broader strokes of purchase intention, there's a subtlety to this factor that might be eluding its current configuration. In other words, while the model is generally performing well, there might be specific patterns, intricacies, or unique characteristics of the data related to purchase intention that the model is missing or not fully understanding.

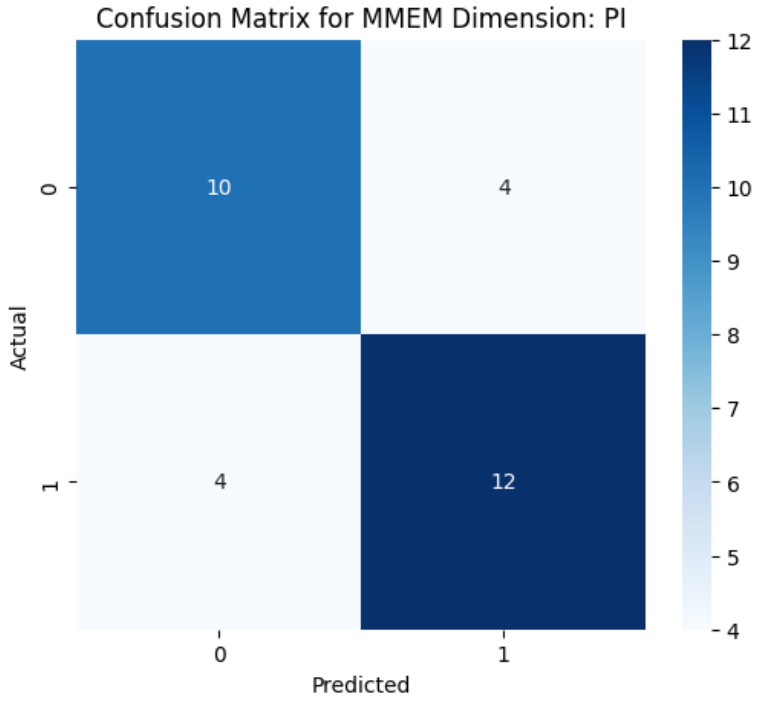

Figure 4.5: Confusion matrix for PI

The Precision, which gauges the accuracy of positive predictions, stands at 0.75, indicating that 75% of the instances predicted as having purchase intention were correct. The Recall, representing the proportion of actual positive instances that were accurately identified, also registers at 0.75, suggesting that the model captured 75% of all genuine purchase intentions. Balancing these two metrics, the F1-Score, a harmonic mean of Precision and Recall, is 0.75, reinforcing the model's consistent performance in both precision and recall aspects for the purchase intention factor.

Metric	**F1**	**Recall**	**Precision**
Value	0.75	0.75	0.75

Table 4.3: Results PI

- **ROC curve and AUC**

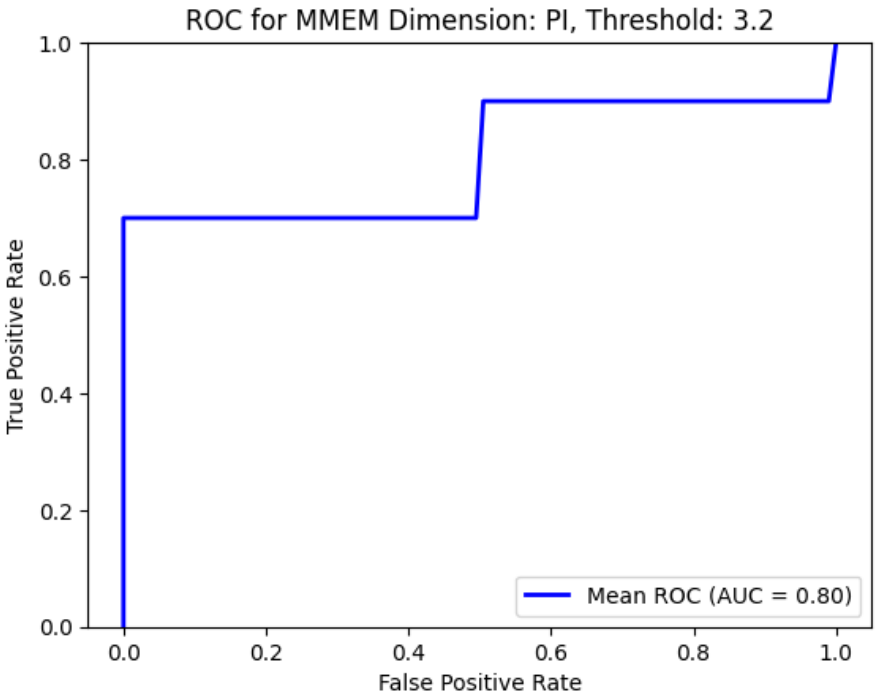

Figure 4.6: ROC curve for PI

For the purchase intention factor, the ROC curve, a graphical representation of a model's true positive rate against its false positive rate, boasts an AUC score of 0.8. This AUC score, standing for Area Under the Curve, is a metric that quantifies the overall ability of the model to distinguish between the classes. An AUC of 0.8 indicates that there's an 80% chance that the model will correctly differentiate between a randomly selected positive instance and a randomly selected negative instance. In essence, this score suggests that the model has a commendable discriminative power for the purchase intention factor, showcasing its proficiency in distinguishing between those with and without the intent to purchase, though there remains room for enhancement.

4.4 Classification of Ad Engagement

- **Overall confusion matrix, precision, recall, f1**

For the ad engagement factor, the model exhibited a keen sense in recognizing 16 instances with genuine engagement and 5 without it. However, it faced challenges, mistakenly anticipating engagement in 7 instances and overlooking it in 2 others. This pattern suggests that while the model has a grasp on identifying true engagement, it occasionally misinterprets certain scenarios, potentially overestimating the level of engagement in some cases.

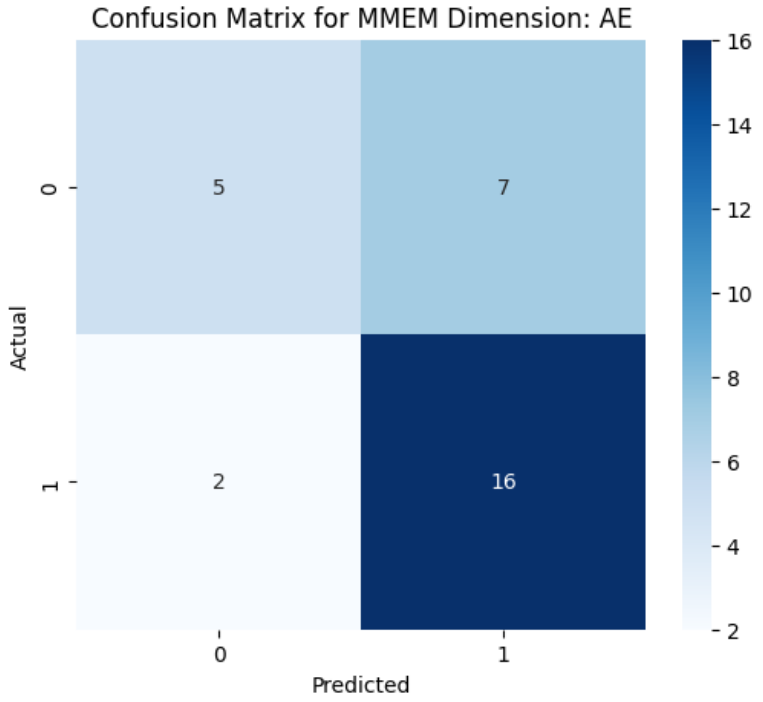

Figure 4.7: Confusion matrix for AE

From the confusion matrix we get that Precision is approximately 0.6957, indicating that about 69.57% of the instances predicted as having ad engagement were accurate. The Recall stands at approximately 0.8889, suggesting the model correctly identified 88.89% of all genuine ad engagements. Harmonizing these metrics, the F1-Score, which balances precision and recall, is approximately 0.7805, reflecting the model's overall effectiveness in predicting ad engagement.

Metric	**F1**	**Recall**	**Precision**
Value	0.7804878048780488	0.8888888888888888	0.6956521739130435

Table 4.4: Results AE

- **ROC curve and AUC**

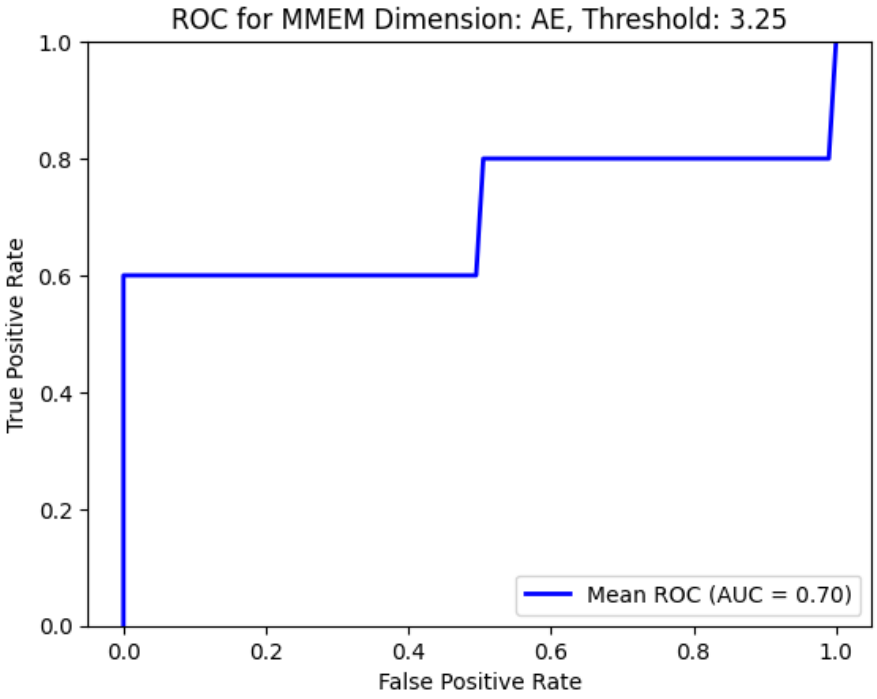

Figure 4.8: ROC curve for AE

For the ad engagement factor, an AUC score of 0.7 suggests that the model has a moderate ability to differentiate between engaged and non-engaged instances. While this indicates a fair level of performance, it's not optimal. An AUC closer to 1.0 would be ideal, so a score of 0.7 highlights areas where the model's predictive capability could be improved. In essence, while the model shows promise in discerning ad engagement, there's room for refinement.

Chapter 5.

Discussion and conclusion

Throughout this research, the primary objectives have been to demonstrate the reliability and consistency of MMAES as a tool, aiming to measure the short-term effects of advertisements on digital platforms. Participants were exposed to multimedia advertisements in a controlled environment, with devices monitoring physiological responses. They then completed the MMAES questionnaire to record their conscious perceptions. This data was consolidated into a DataFrame using Python, processed, and subjected to machine learning modeling using the RandomForest algorithm.

5.1 General observations

The results obtained reflect the effectiveness of MMAES as a tool to measure the short-term effects of advertisements on digital platforms. Key metrics showed precision ranging between 66.67% and 75%, recall between 57% and 88%, and an F1-Score that balanced these two metrics, with values ranging from 64% to 78%. Additionally, the AUC, indicative of the model's ability to distinguish between classes, recorded values between 0.7 and 0.8. These data underscore the model's ability to predict reactions to multimedia advertisements and its overall performance in terms of precision, recall, and class discrimination. By integrating questionnaire responses with physiological reactions, a multidimensional perspective of the experience in front of multimedia advertisements is obtained, enriching our understanding of the advertising impact on the individual. In summary, this study demonstrates the potential and relevance of combining traditional methodologies with modern machine learning techniques to better analyze and understand audience reactions to multimedia advertisements.

5.2 Limitations and future directions

- **Limitations:** While this study provides valuable insights into the short-term effects of multimedia advertisements on digital platforms, it is essential to acknowledge its limitations. The sample size, although diverse, might not represent the broader population, and thus, generalizability to other demographics should be approached with caution. Additionally, while physiological devices offer objective data, they might not capture the full spectrum of human emotional responses, as emotions are complex and multidimensional.
  During the development of the MMAES methodology, a significant limitation was encountered in the availability of data from the Empatica and Tobii sensors. This resulted in the loss of information due to fragmented datasets, where some users lacked data from one sensor while others lacked data from the other sensor. This limitation emphasizes the challenges inherent in collecting comprehensive physiological data and highlights the need for more robust data collection strategies in future research endeavors.

- **Future directions:** Future research could focus on expanding the sample size or exploring other physiological measures to gain a more comprehensive understanding. Additionally, integrating other machine learning algorithms or deep learning techniques might enhance the predictive accuracy of the model.
  Upcoming studies will concentrate on assessing the resilience and broad applicability of MMAES, as well as enhancing its psychometric framework. We intend to administer multiple evaluations of MMAES in various environments and scenarios, such as on mobile platforms, and among diverse consumer demographics. We will employ confirmatory factor analysis and other measures of external validity to determine the instrument's widespread relevance.

5.3 Implications for the Industry

The findings of this study have significant implications for the advertising industry. By understanding audience reactions in real-time, advertisers can tailor their campaigns more effectively, ensuring a higher return on investment. The integration of physiological responses with conscious feedback provides a holistic view, allowing for more nuanced and targeted advertising strategies.

MMAES can be utilized in its entirety or by choosing specific components that align with a particular scenario. When used fully, it takes under 15 minutes to finish MMAES. For research involving several cycles, each iteration or task should have its own separate MMAES assessment. In situations where only specific facets of online ad viewing are being examined, it's more beneficial to pick only the pertinent components. This approach helps keep participants engaged and minimizes weariness.

References

1. Han Y, Kim H gon, Lan T. The Impact of Multimedia Video Marketing on Consumer Psychology and Behavior. Mobile Information Systems. 10 de agosto de 2022;2022:1-8.

2. Aslan Oguz E, Kosir A. An Online Crowdsourcing Experiment to Model the Effects of a Commercial on a User's Consumption Behavior. En: 2020 4th International Conference on Computer Science and Artificial Intelligence [Internet]. Zhuhai China: ACM; 2020 [citado 29 de septiembre de 2023]. p. 15-23. Disponible en: https://dl.acm.org/doi/10.1145/3445815.3445818

3. Strle G, Košir A, Burnik U. Physiological Signals and Affect as Predictors of Advertising Engagement. Sensors. 3 de agosto de 2023;23(15):6916.

4. Haapalainen Ferreira E, Kim S, Forlizzi J, Dey A. Psycho-Physiological Measures for Assessing Cognitive Load. UbiComp'10 - Proceedings of the 2010 ACM Conference on Ubiquitous Computing. 2010. 301 p.

5. Aslan Oğuz E, Strle G, Košir A. Multimedia ad exposure scale: measuring short-term impact of online ad exposure. Multimed Tools Appl [Internet]. 29 de marzo de 2023 [citado 25 de septiembre de 2023]; Disponible en: https://doi.org/10.1007/s11042-023-14401-5

6. de Vreese CH, Neijens P. Measuring Media Exposure in a Changing Communications Environment. Communication Methods and Measures. 2 de abril de 2016;10(2-3):69-80.

7. www.javatpoint.com [Internet]. [citado 25 de septiembre de 2023]. Machine Learning Random Forest Algorithm - Javatpoint. Disponible en: https://www.javatpoint.com/machine-learning-random-forest-algorithm

8. Guide to AUC ROC Curve in Machine Learning [Internet]. GeeksforGeeks. 2020 [citado 25 de septiembre de 2023]. Disponible en: https://www.geeksforgeeks.org/auc-roc-curve/

9. Understanding Confusion Matrix | by Sarang Narkhede | Towards Data Science [Internet]. [citado 25 de septiembre de 2023]. Disponible en: https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62

10. Narkhede S. Medium. 2021 [citado 25 de septiembre de 2023]. Understanding AUC - ROC Curve. Disponible en: https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5

11. Classification in Machine Learning: A Guide for Beginners [Internet]. [citado 25 de septiembre de 2023]. Disponible en: https://www.datacamp.com/blog/classification-machine-learning

12. Classification Algorithm in Machine Learning - Javatpoint [Internet]. [citado 25 de septiembre de 2023]. Disponible en: https://www.javatpoint.com/classification-algorithm-in-machine-learning

13. Brownlee J. A Gentle Introduction to k-fold Cross-Validation [Internet]. MachineLearningMastery.com. 2018 [citado 25 de septiembre de 2023]. Disponible en: https://machinelearningmastery.com/k-fold-cross-validation/