

**Predicting final user satisfaction
based on user experience data
using machine learning**

(機械学習を用いたユーザ体験データによる総合的満足度の推定)

March 2023

Kitti Koonsanit

Tokyo Metropolitan University

TABLE OF CONTENTS

1	Introduction.....	1
	1.1 Background of the Study	1
	1.2 Statement of the Problem	4
	1.3 Research Objectives	6
	1.4 Significance of the Research	6
	1.5 Organization	7
2	Literature Review	9
	2.1 Customer Relationship Management	9
	2.2 User Experience (UX)	10
	2.3 UX Evaluation Methods	11
	2.4 Order Effect in UX	16
	2.5 Classification Techniques.....	17
	2.6 Sampling Techniques	19
3	Predicting Final User Satisfaction Using Momentary UX Data and Machine Learning Techniques	21
	3.1 Introduction	21
	3.2 Proposed Method.....	21
	3.3 Experiments	22
	3.4 Results and Discussion	35
	3.5 Summary.....	42
4	Using Random Ordering in User Experience Testing to Predict Final User Satisfaction.....	45
	4.1 Introduction	45
	4.2 Proposed Method.....	48
	4.3 Experiment	50
	4.4 Results and Discussion	60

4.5 Summary.....	62
5 UX Evaluation in Practical Use	64
5.1 Introduction	64
5.2 Proposed Method in Practical Use	67
5.3 Experiments	70
5.4 Results and Discussion	76
5.5 Summary.....	77
6 Conclusions.....	79
6.1 Summary of the Study	79
6.2 Future Work.....	81
References.....	83
Acknowledgement.....	91

CHAPTER 1

INTRODUCTION

1.1 Background of the Study

This research study presented a framework to classify the final user satisfaction of products or services by the user experience (UX) data using machine learning. User experience (UX) refers to all aspects of how people interact with a product or service. UX emphasizes the experiential, affective, meaningful, and value aspects of human–computer interaction and product ownership but also includes a person’s perceptions of practical aspects such as utility, ease of use, and product or service efficiency. UX is highly context-dependent, subjective, dynamic, and random ordering [1–3], as it concerns an individual’s performance, feelings, and thoughts about the product or service. Moreover, these can change over time as circumstances change. The design processes of products and services are often evaluated using the comprehensive full user experience evaluation (UXE) method for time-continuous situations [4,5]. From the first to the final stage of usage, the user’s emotions and perceptions can change continuously through the receipt of multiple stimulatory experiences while using products or services. After usage, users are asked about their overall satisfaction as an indicator of final user satisfaction regarding one or more aspects of the product or service. Answers relating to final user satisfaction are often expressed on a scale that includes negative and positive values, ranging from –10 to +10 [6,7], with higher scores indicating higher satisfaction. In particular, final user satisfaction following their experiences has been considered an extremely important factor in users’ decisions about further use or recommending the products or services to other people [8]. However, final user satisfaction reported after use may be imprecise because it varies according to situations such as user activities, as shown in Figure 1-1.

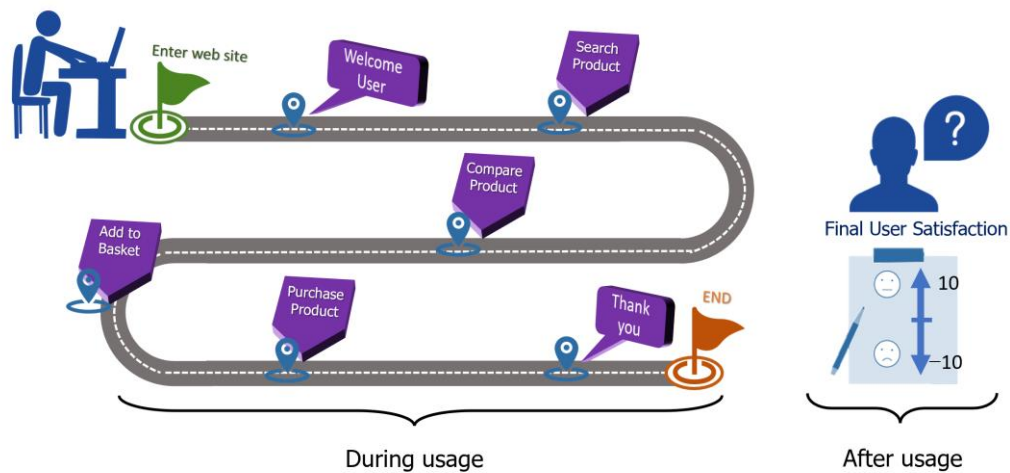


Figure 1-1. Evaluation of final user satisfaction after website usage.

Figure 1-1 shows several stages of usage, each of which may impact the user’s emotions and perceptions and, in turn, affect final user satisfaction. Roto et al. presented the User Experience White Paper, a document reporting Dagstuhl Seminar’s results on categorizing user experience from the viewpoint of time axis [9]. In that document, the importance of analyzing UX across time was underlined. There are four types of UX—anticipated, momentary, episodic, and cumulative (Figure 1-2)—each of which is defined based on usage time: (1) anticipated UX relates to the period before first use; (2) momentary UX relates to the period during usage; this type refers to any perceived change that occurs during the interaction, at the very moment [10]; (3) episodic UX relates to the period after usage; and (4) cumulative UX relates to the entire period, including from before first use, during usage, and after usage. The four types of UX can affect the final user satisfaction.

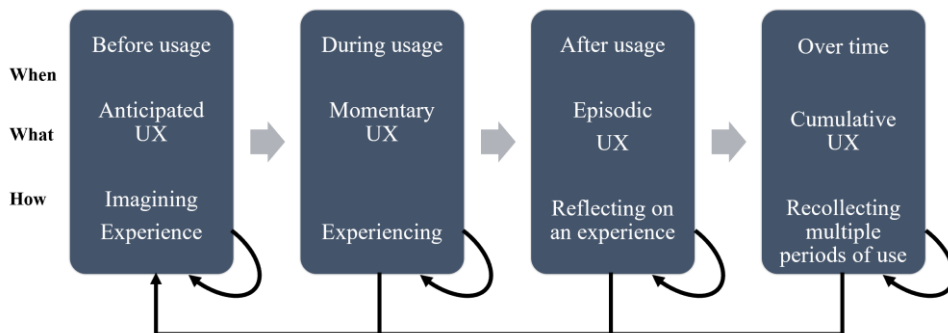


Figure 1-2. The four types of user experience (UX), adapted from Roto et al. (2011).

Previously, Kurosu et al. defined the meaning of user satisfaction as the vertical axes of satisfaction based on the UX graph being the same as user satisfaction [5,11]. Final user satisfaction means satisfaction after momentary UX. In the present study, this dissertation defined the meaning of final user satisfaction similar to Kurosu et al., and Marti et al. that is, user satisfaction after finished usage [5,11]. Based on this definition, the final user satisfaction is similar to episodic UX.

In the past decade, there have been many studies of episodic UX and cumulative UX, with most focusing on only one type [12]. Many studies have tried to estimate the satisfaction of users using various machine learning techniques [13–15] as shown in Table 1-1. Matsuda studied the satisfaction level of tourists during sightseeing by using the tourists' unconscious, natural actions [13]. They conducted experiments with 22 tourists in two different touristic areas in Germany and Japan. Their results confirmed the feasibility of estimating both the emotional status and satisfaction level of tourists. Cavalcante applied machine learning techniques including decision trees, support vector machines and ensemble learning to predict customer satisfaction from service data [14]. The results indicated that the development of an intelligent algorithm may assist in identifying customer satisfaction. Kumar presented a machine learning approach to analyze tweets to improve customer experience [15]. They found that a machine learning approach can provide better classifications for customer satisfaction in the airline industry. All of the aforementioned studies gathered data and measured satisfaction using episodic UX and cumulative UX from sensors or devices. However, during actual usage by customers, there are many external factors that can affect their satisfaction.

Table 1-1. Many studies have tried to estimate the satisfaction of users using various techniques.

Data Collection	X	Y	Predictive Method	No. of Samples	Proposed by
UX Type	UX Data	Target Prediction			
Momentary UX	Facial videos, eye and head motion, body movement	Satisfaction level of tourists	RNN-LSTM	22	Matsuda et al., 2018 [12].
Momentary UX	Text from user twitter	Customer satisfaction	NLP, CNN, SVM and ANN	120,766	Kumar et al., 2019 [15]
Momentary UX	Questionnaire and interview data	User satisfaction	Quantitative and Qualitative analysis	20	Lin Feng et al., 2019 [16]
Momentary UX	Self-Questionnaire	Final user satisfaction	SMOTE and SVM	50	This thesis (Kitti et al., 2021) [1]

RNN = Recurrent Neural Network; LSTM = Long Short-Term Memory;

CNN = Convolutional Neural Network; ANN = Artificial Neural Networks;

SMOTE = Synthetic Minority Oversampling Technique;

SVM = Support Vector Machine.

1.2 Statement of the Problem

A major problem with assessments of both episodic UX and cumulative UX, however, is that the graph or curve is recorded after the user has finished the task. Moreover, studies of both episodic UX and cumulative UX evaluation have employed the usage time to collect data, rather than using other methods to evaluate UX type. It has been pointed out that these two types of UX evaluation require participants'

dedication over time [17], with assessments typically spanning intervals ranging from a few days to a month or more. Therefore, this paper focuses on momentary UX and examines the emerging role of momentary UX in the context of final user satisfaction.

Momentary UX has been measured and evaluated by questionnaire (subjective evaluation) surveys, with question items for each step of the experience. However, comprehensive evaluation of subjective answers to these questionnaires is difficult because the conventional methods of analyzing subjective evaluation may not adequately relate to the momentary UX. Instead, the quality of conventional analytic methods is determined by the experts and is directly dependent on their level of expertise. Furthermore, multiple-evaluation comparisons may be difficult due to the variety of checklists used and difficulty in quantifying expert opinions [18].

Some previous studies have measured UX at each stage of usage [16,19]. Despite the relationships between momentary UX and episodic UX, various factors will interact intricately during the actual user experience, and the final satisfaction (episodic UX) will be determined from the accumulation of experiences at each stage. This view is supported by Sánchez-Adame [20], who writes that, as an example, the user might experience a strong, albeit temporary, negative reaction when evaluating momentary UX during usage, but when episodic UX is measured again after usage, the user may be more likely to prioritize good aspects over bad ones. These data are interesting because the evaluative judgment at each stage is related to overall final satisfaction with the product.

UX is subjective, relating to an individual's feelings and satisfaction. Expert evaluations of UX may lead to bias, and such opinions are not easily quantifiable. Humans are prone to many types of bias. Despite algorithms having their own challenges, machine learning algorithms may conceivably be capable of producing more fair, efficient, and bias-free outcomes than humans. This study aimed to predict final user satisfaction by combining momentary user experience data and machine learning techniques. My hypothesis is that machine learning will perform well on momentary user experience data in the prediction of final user satisfaction.

1.3 Research Objectives

This research aims to present a framework to predict the final user satisfaction of products or services by the user experience (UX) data using machine learning.

Main Objective

To propose a framework to predict final user satisfaction of products or services by user experience (UX) data using machine learning.

Specific Objectives

- (i) To predict the final user satisfaction from UX data during products or services usage
- (ii) To predict the final user satisfaction using random ordering in user experience testing

1.4 Significance of the Research

The invention of framework was developed as a user experience research tool to help UX researcher conducting UXE to obtain more accurate predictions of user satisfaction. This research is relevant and necessary because it directly supports UX researchers to understand final user satisfaction. Because final user satisfaction following their experiences has been considered an extremely important factor in users' decisions about further use or recommending the products or services to other people. The research could help to provide the appropriate final user satisfaction for UX designers and researchers. It could help businesses to achieve a better user satisfaction.

In the viewpoint of UX designers on this study, the use of this proposed framework based on machine learning could conceivably be capable of producing more fair, efficient, and bias-free outcomes in predicting final user satisfaction than humans. Because UX is subjective, relating to an individual's feelings and satisfaction. Expert evaluations of UX may lead to bias, and such opinions are not easily quantifiable. The UX data consist of many variables with differing dimensions, and analysis is therefore not easy.

Most UX designers are not easily to analysis those quantitative and qualitative data. This proposed framework will help UX designers predict the appropriate final user satisfaction for customers through machine learning approach. Because its results are reliable and unbiased in estimating final user satisfaction. This study will be useful to the related UX researchers and UX developers in User Experience testing.

The understanding of those aforementioned relationships could provide the best UX for customers, build a good brand image, and launch customer-centric marketing campaigns. It can help businesses to achieve a better user satisfaction and the goals of sustaining long-term competitive advantages. For example, in service industry, the product or service developers could discover the worst points of products or services at which a customer requires assistance during product or service usage. They need to ensure that customers can finish their transaction without difficulty in different usage situations.

As a result, the understanding of momentary UX could boost overall customer satisfaction as well as repeat purchase rate, maintain long-term sustainable customer satisfaction and achieve sustainability. It could help to understand customers better and thus enhance communication with stakeholders with regard to efficiency, performance, and sustainability of products or services.

1.5 Organization

The dissertation is composed of six chapters (Figure 1-3). Chapter 1 introduces the background, motivation, and the purpose of this research, and the structure of this dissertation. Chapter 2 describes the evaluation of UX and organizes related studies to explain the proposed framework. Chapter 3 describes experiments that predict final user satisfaction using momentary UX data and machine learning techniques. Chapter 4 describes experiments that predict final user satisfaction using random ordering in UX testing. Chapter 5 demonstrates the application of UX evaluation for some practical purposes. Chapter 6 presents the conclusion and future work of the research.

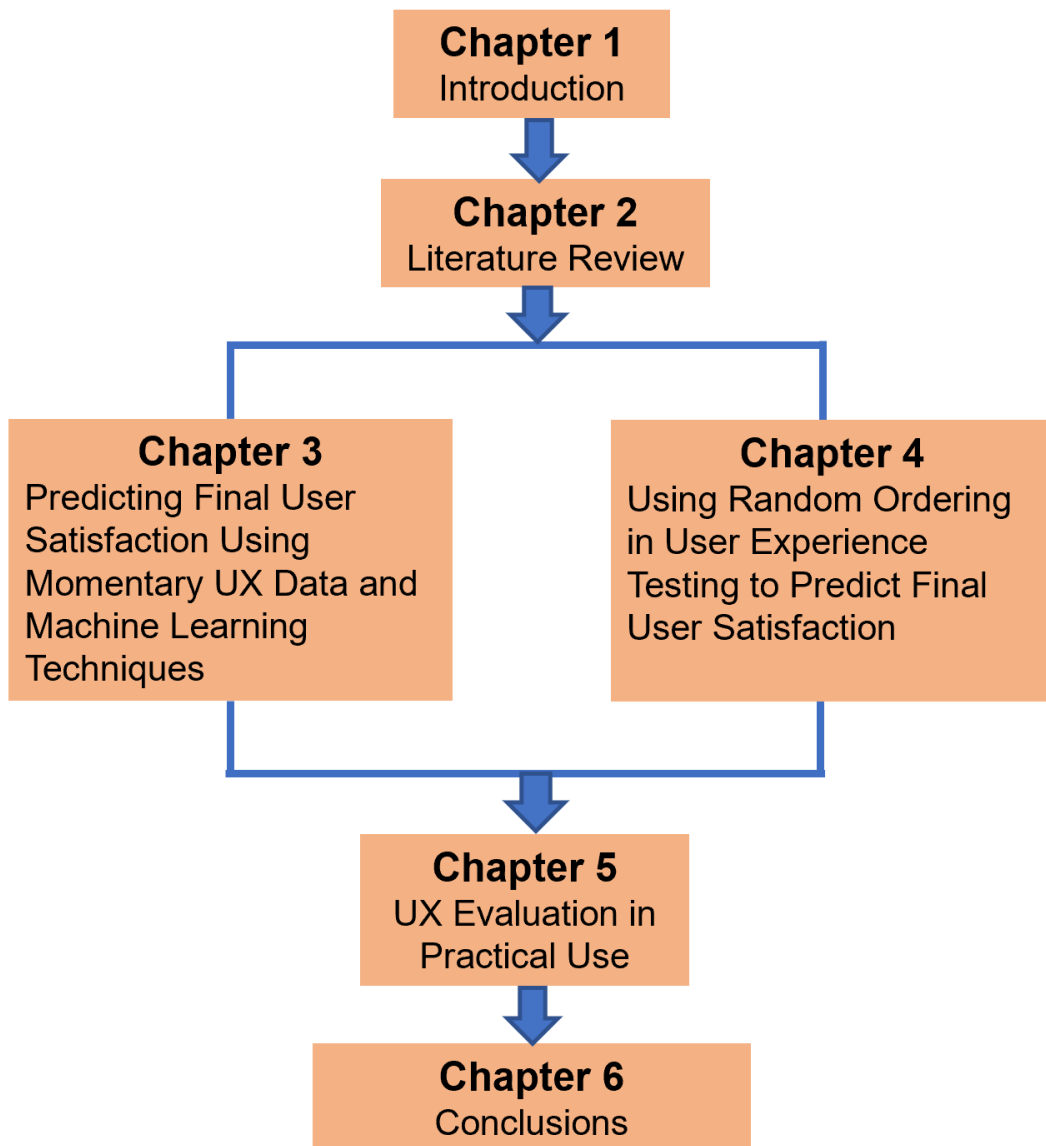


Figure 1-3. Relationship of each chapter

CHAPTER 2

LITERATURE REVIEW

This chapter aims to review the literature on the overall picture of the background of UX, UX evaluation, Machine Learning. It presents a discussion of relevant issues to the study regarding the prediction framework. Moreover, this chapter describe some background on UX evaluation methods, order effects in UX related to this work, and some details of general and relevant machine learning algorithms. In brief, this chapter is organized on the following sections: 2.1 Customer Relationship Management 2.2 User Experience 2.3 UX Evaluation Methods 2.4 Order Effect in UX 2.5 Classification Techniques and 2.6 Sampling Techniques.

2.1 Customer Relationship Management

Customer relationship management (CRM) is an approach to maintain positive customer relationships and to improve customer satisfaction [21]. This new management process is aimed at improving the business and customer relationships, strategically regarding the core enterprise business customers as an important resource, meeting customer needs through the improvement of customer service and in-depth analysis of the customer, so that enterprises can maximize customer satisfaction and loyalty, establish mutual long-term stable and trusted relationships, thereby maximizing customer lifetime value [22].

Furthermore, CRM provides data and information about customers, such as their feelings, shopping behavior, and product consumer habits, among others. These user data and information provide essential feedback from the customers' perspective, including their opinions, favors, preferences, and past experiences. The information thus obtained is used to improve communication with customers to create value and satisfaction [23]. CRM analytics can help facilitate better product or service decisions.

Some recent research reported that customer relationship capability and CRM technology in the service industry are important variables in building customer satisfaction post-purchase [24,25]. In other related work, Taufik proposed a method to

utilize user data from a CRM system. They developed an online analytical processing (OLAP)-based analytical CRM system to analyze customer data and classify it into two main segments, based on geography and demographics [26]. The benefit of his approach is that the analytical process can operate upon user data from various dimensional perspectives to quickly capture customer needs in real-time. This analytical CRM system can be easily accessed by managers to make decisions. A recent study presenting an approach to processing user data and information obtained from the CRM system to satisfy customers concluded that CRM plays a major role in increasing customer satisfaction. Thus, it improves both in-depth customer knowledge and higher customer satisfaction [21].

One highly interesting aspect of customer data from CRM is user experience [27]. Several articles regarding the measurement of UX to gauge satisfaction have been published [28]. In the modern digital world there are many methods to gather UX data via automation technologies, such as interactive responses, and online questionnaires [29]. In most approaches, UX is generally measured by a questionnaire or survey method. However, evaluation of the final user satisfaction with products and services using UX questionnaires has been considered challenging because it is difficult to measure the final user satisfaction. Due to differences in user experience for each user, both humans and computers have had difficulty in classifying these data for developing or improving products and services. Furthermore, although answers from UX questionnaires can provide abundant information about a range of feelings, their high complexity substantially increases the computational burden in interpretation for experts. In this context, a new approach integrates knowledge between UX obtained from CRM and intelligent systems with machine learning techniques. This approach can be of practical value for customer relationship management by improving understanding of user satisfaction [30,31].

2.2 User Experience (UX)

The term “user experience” refers to a person’s overall experience of interacting with a product or service [32]. UX covers not only direct interactions with the product, for example, but also how the resulting experience fits into the overall task completion process. Every interaction between the user and product or service is

factored into the overall user experience. As a result, final satisfaction with respect to these UX has been regarded as highly important in the users' decision to continue using or recommending the products or services to others [8].

2.3 UX Evaluation Methods

Many approaches to evaluating UX have been proposed [33], with studies proposing various methods and ways of categorizing the data. User experience evaluation (UXE) is intended to help the designer determine if a design effectively achieves its goals and what changes should be made for improvement. Some UXEs that might appear similar are, in fact, not. This dissertation decided to classify the many UXEs into five groups by considering "periods" of experience [33]. Methods are defined as uniquely applicable to a specific period, such as before, during, or after usage, and are sensitive to that period's characteristics, for example, momentary UXE, episodic UXE, and cumulative UXE. Accordingly, this study classified the UXE methods as follows:

- Before usage (prior to interacting with products/services);
- Momentary (a snapshot, e.g., perceptions, emotions);
- Single (a single episode in which a user explores design features to address a task goal);
- Typical test session (e.g., 100 min in which a user performs a specific task.)
- Long-term (e.g., interacting with products/services in everyday life).

When these five methods are applied, momentary evaluation is considered as short-term, and a reliable way to capture feelings and user experiences during usage. Although the short-term evaluation method may miss data between stages of user experience [19,34], it is one of the most reliable methods [35] because it records time-varying subjective experiences, reducing response biases and memory distortions. This is reflected in UX's dynamic nature in the longer term. During momentary use, users may experience various unexpected events during their interaction. As momentary data logs can be useful for UX evaluation, this study decided to use momentary evaluation for this research.

Most research on UX [33] has described changes in user experience over time. Examples include the UX Curve method [34], UX Graph method [5,11], and iScale method [36] as shown in Table 2-1.

Table 2-1. Summary of user experience evaluation through user experience (UX) curve, UX graph, and iScale.

Approach.	Description
UX Curve [34]	UX Curve is a tool for drawing a timeline and a horizontal line that splits positive and negative experiences.
UX Graph [3,22]	UX Graph is a tool for drawing the degree of satisfaction on a time scale. It is an improved version of the conventional UX Curve.
iScale [36]	iScale is a tool for the backward-looking expression of long-term user experience data.

These three UXE methods involve self-reported measurements over time, whereby the participants report their feelings and emotions in the form of line graphs drawn by hand. However, these methods are not suitable for determining final user satisfaction because drawings are made after plotting each episodic event. This means that the UX curve and UX graph are drawn, and the points specified only after the task is finished, which makes the method time-consuming [5]. Furthermore, most UXE methods are used to describe only how user experience changes during usage. Untidy handwriting means that characters in text can be difficult to read [34], so that evaluation results may be difficult to analyze and interpret. Thus, iScale, UX curve, and UX graph fall short of the requirements for appropriate final user assessment.

As mentioned above, UX refers to all aspects of how people interact with a product or service. Many approaches have been proposed to evaluate UX. These methods are defined as being uniquely applicable to a specific period, namely before, during, or after usage. In 2011, Roto et al. published the User Experience White Paper, a document reporting Dagstuhl Seminar’s categorization of UX from the viewpoint of the time axis [9]. The document underlines the importance of analyzing UX across time, and describes four types of UX—anticipated, momentary, episodic, and cumulative—each of which is defined based on usage time. Anticipated UX relates to the period before the first use; momentary UX relates to the period during

usage and refers to any perceived change that occurs at the moment of interaction [10]; episodic UX relates to the period after usage; and cumulative UX relates to the entire period, from before the first use, during usage, and after usage. These four types of UX can affect final user satisfaction.

One way to conduct UXE is using a UX Curve, a method designed to facilitate the collection of past UX data [34,37]. The curve is used to help users retrospectively report how and why their experience with a product has changed over time, and allows researchers to visualize interactions with the product from the customer's point of view. The curve is drawn on a horizontal axis showing the time and activities engaged in during usage, and vertical axis showing the satisfaction level (positive or negative) during usage. The satisfaction level can fluctuate significantly depending on the time and order of activities performed, as shown in Figure 2-1.

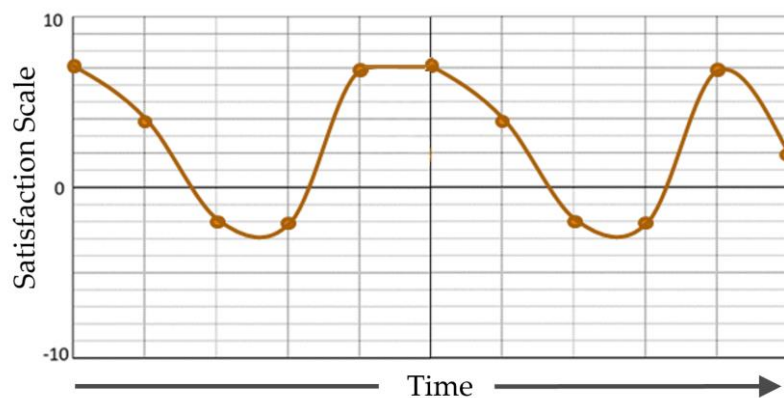


Figure 2-1. Example of a UX Curve

Another UXE method and descendent of the UX Curve is the UX Graph [5,11]. Designed by Kurosu et al., the method involves plotting the user's satisfaction as an intensive measure of UX on one graph. Kurosu developed software to enable users to easily depict their degree of satisfaction on a time scale. The graph can be drawn after the user enters "episodes" describing their experience and satisfaction rating. User satisfaction is represented on the vertical axis, and final user satisfaction is defined as satisfaction after momentary UX.

Several studies have measured UX to improve final user satisfaction [1,38–44]. In addition to the fact that momentary UX can affect episodic UX, various factors will interact in an intricate manner during the actual UX, and the final user satisfaction (episodic UX) will be determined from the accumulation of experiences. Sukamto et al. proposed an approach to enhance user satisfaction with a portal website by increasing the UX score in the User Experience Questionnaire (UEQ) [38]. The UEQ uses a questionnaire to measure users' subjective impression of their user experience of products. The UEQ is a semantic differential with 26 question items, with responses provided on a 7-point Likert scale from -3 (fully agree with negative term) to +3 (fully agree with positive term). Users will achieve a better score in the UEQ as they become more comfortable using the portal website. Based on the User-Centered Design method [38], the researchers assessed the initial and final usability score for a website using the UEQ. They showed that increasing the UX score through the UEQ in the development of a portal website could enhance customer satisfaction. In other related work, Pushparaja et al. examined the factors of UX that influence users' satisfaction when using a digital library website [39]. Through a literature review, the team found that attractiveness, efficiency, dependability, stimulation, and novelty were key factors of satisfaction. Meanwhile, Mominzada et al. reported the consequences of UX and its role in a gamified e-commerce platform [40]. With the aim to identify the effects of gamification on user experience and the related consequences, the researchers used an online survey questionnaire as the main instrument for data collection. They showed that UX positively affects final user satisfaction in a gamified e-commerce platform. The experiment was statistically tested and validated through a quantitative research approach.

Due to differences in UX among users, both humans and computers have had difficulty in classifying these data to develop or improve products and services. Furthermore, although responses to UX questionnaires can provide an abundance of information about a range of user feelings, the complexity of this data substantially increases the computational burden for experts. Machine learning techniques have recently been successfully used to make UX questionnaires easier to analyze and interpret. Many popular machine learning techniques have been used to analyze UX questionnaires, including support vector machine (SVM) classifiers, logistic

regression, decision trees, and neural networks (NNs). Several studies have used machine learning to gauge final user satisfaction [1] [41]. Research by Koonsanit et al. proposed an approach to predict final user satisfaction by combining momentary user experience data with machine learning techniques [1]. User satisfaction was measured while each user performed a fixed order of tasks on a product. The study reported that machine learning methods like SVM can accurately predict final user satisfaction and contribute to developing better products and services by analyzing UX. SVM with polynomial kernel had the highest cross-validation accuracy, at 93%. The findings suggest that machine learning could be useful for analyzing momentary UX data to predict final user satisfaction. In other related work, Nwakanma et al. proposed a method for classifying the quality of UX and predicting customer sentiment to improve service delivery [41]. They collected UX data using Google Forms and developed an improved logistic regression classifier to test, train and classify UX. The training accuracy of the proposed improved logistic regression was 97.67%, indicating the potential and capability of machine learning for analyzing a large database of sentiments or reviews and predicting customer sentiment. Machine learning approaches for UXE can thus be applied to industrial scenarios to evaluate users' perception of products and services.

In fact, Cong et al. proposed a machine learning-based iterative design approach to automate the prediction of user satisfaction called Smart PSS [44]. UX data in this study represented the subjective psychological feeling gained from the Smart PSS experience during use, and was attained by calculating the user satisfaction score collected on a 5-point Likert scale, with 5 points indicating extreme satisfaction and 1 point indicating extreme dissatisfaction. Overall user satisfaction with Smart PSS was determined after users had completed 10 tasks. Data were collected from 20 test users in this study. After the experiment, the collected data were reorganized and processed using several techniques including data cleansing, data integration, feature selection, and data augmentation, before creating classification models. After building three models using SVM, decision trees, and a NN and comparing their performances, the researchers showed that the NN was most accurate, with an overall accuracy of 90%. Meanwhile, models using SVM and decision trees had greater than 84% accuracy in the test set.

2.4 Order Effect in UX

In my literature review, this study found that one major challenge of UXE is the order effect. Keiningham et al. reported that the order of activities and users' satisfaction with the most recent task in the UX significantly affect final user satisfaction [42]. They showed that the weights of initial satisfaction and transaction-specific satisfaction decay geometrically with time. This causes user satisfaction with the most recent task to receive more weight and priority when determining final satisfaction. Thus, more recent transaction-specific satisfaction levels tend to have a greater influence on final satisfaction. In another interesting study, Min et al proposed the importance of task order in a customer complaint system [43]. In general, it is often believed that employees should first apologize to customers before listening to their complaints. To test this belief, Min et al. created a digital library site and invited participants to perform article searches to evaluate the website. During their task, the participants encountered a service failure in the form of a long wait-time caused by a slow response on the site. The researchers reported that participants with high expectations were more satisfied with a responsive apology (i.e., listen-and-then-apologize) than a preemptive apology (i.e., apologize-and-then-listen), with a mean satisfaction score of 4.64 and 3.85 ($p < .01$), respectively. Thus, a simple change in the sequential order of tasks such as apologizing and listening (listen-and-then-apologize) to complaints can significantly impact customer satisfaction and final user satisfaction.

The above-mentioned studies emphasize the challenges of UXE. The order of tasks in the UX and user's memory of their satisfaction with the most recent task significantly affect their final satisfaction. The sequence and order of actions, including whether they are fixed and random, performed on a product or service are also important. Accounting for the order or sequence of actions could improve predictions of final user satisfaction. In the present study, this study assumed that a change in the sequential order of tasks could significantly affect final user satisfaction.

2.5 Classification Techniques

The various UX data concerning momentary usage can be problematic when it comes to analysis. For example, UX data are not simply one-dimensional, and each questionnaire may have a different scoring range. This makes for characteristically tedious work that is considered repetitive by UX researchers; in particular, the use of human labor to explain and analyze user satisfaction is not optimal. Consequently, the field of feedback and user satisfaction from pilot product studies has shown little progress or improvement over time. Hence, machine learning methods that facilitate analysis and understanding of final satisfaction have long been sought [45].

There are two main benefits of using machine learning instead of solely relying on statistical analysis methods in predicting final user satisfaction.: Firstly, UX data from questionnaire consists of many variables with different dimensions, making it challenging to analyze using traditional statistical methods. Machine learning algorithms can handle complex data and extract useful insights from it. Secondly, machine learning algorithms are not susceptible to human biases, unlike expert evaluations. They provide an objective analysis of the data, reducing the risk of subjective interpretations.

The type of machine learning algorithms used in the present study were determined by multiple factors, ranging from the type of problem at hand to the type of output desired, including type and size of the data, available computational time, number of features, and observations in the data. All such factors are important when choosing an algorithm before conducting research. Many scholars hold the view that support vector machines (SVMs) [46] can efficiently perform non-linear classification when the correct kernel and an optimal set of parameters are used [47]. Recent research has suggested that SVMs can be used for classification as well as pattern recognition purposes, especially with speech and emotion data [48]. Furthermore, algorithms such as SVM, K-nearest neighbors (KNN) [49], and logistic regression [50] are easy to implement and run [51]. By contrast, neural networks with high convergence time require significant time to train the data.

This study chose seven machine learning algorithms as simple and easy-to-build classification models. This study compared these seven different methods including

polynomial kernel SVM, radial basis kernel SVM, linear kernel SVM, sigmoid kernel SVM, logistic regression [50], K-nearest neighbors [49], and multilayer perceptron [52] as shown in Table 2-2.

Table 2-2. Summary of classification techniques.

Approaches	Description
Support Vector Machine with Polynomial Kernel Function	The SVM algorithm uses the best hyperplane to separate n-dimensional space into classes. The learning of the hyperplane is processed by transforming the problem using Polynomial Function [50].
Support Vector Machine with Radial Basis Kernel Function	SVM models classify data by optimizing a hyperplane that separates the classes using Radial Basis Kernel Function [50].
Support Vector Machine with Linear Kernel Function	This classifier is formally defined by a separating line. The learning of the hyperplane is processed by transforming the problem using linear algebra [50].
Support Vector Machine with Sigmoid Kernel Function	SVM models process data points by drawing decision boundaries with the Sigmoid Kernel Function [50].
K-Nearest Neighbors	K-Nearest Neighbors uses the label of data points surrounding a target data point to define the class label by a majority vote of its neighbors [49].
Logistic Regression	Linear Regression is a technique to predict a continuous output value from a linear relationship. However, the output of Logistic Regression will provide a value between 0 and 1, a probability [50].
Multilayer Perceptron	A multilayer perceptron (MLP) is a technique to classify the target variable used for supervised learning. It is the same structure as a single layer perceptron with one or more hidden layers. [52].

According to previous research [1,15,44,48], machine learning algorithms like SVM [46] can accurately predict final user satisfaction based on momentary UX data. A classical machine learning model works through a simple premise: it learns from data with which it is fed. Collecting and feeding more data into a classical machine learning model leads to more training and more accurate predictions as it continually learns from the data. Previous studies have largely used supervised learning algorithms such as SVM for classification. One of the most popular classical machine learning algorithms in use today [53], SVM is particularly effective in high dimensional spaces. The effectiveness of an SVM depends upon the kernel function and parameters of the kernel.

2.6 Sampling Techniques

The number of data points plays an important factor in the creating of machine learning models. The issue of limited amounts of data has received considerable critical attention. Investigators have recently examined the effects of the sample size on machine learning algorithms. Although it may be hard to determine the exact number of data points that any given algorithm requires, some studies demonstrate that using small sample sizes for building classical machine learning model leads to better performance [54]. Other studies discuss the number of samples per class for small general datasets [55].

When dealing with small datasets, most researchers prefer to use classical machine learning algorithms instead of deep learning. There are a few reasons for this. Firstly, deep learning models often require a substantial amount of data for effective training, which can result in overfitting on small datasets. Secondly, classical machine learning algorithms, such as support vector machines, are often more interpretable and easier to debug, making them advantageous when working with small datasets. Lastly, training deep learning models can be computationally demanding and may not be feasible with limited computational resources, especially when handling small datasets.

A lack of sufficient data may lead to serious problems, such as an imbalanced distribution across classes [56]. Because many machine learning algorithms are designed to operate on the assumption of equal numbers of observations for each class, any imbalance can result in poor predictive performance, specifically for minority classes. To solve the problem of imbalance in distribution, this study covered a suite of data sampling techniques to generate alternative, synthetic data [57]. The sampling method is obtained from the creation of new data or a pre-existing original dataset, and then used to create a new classification model with the machine learning method. Different sampling techniques are available for imbalanced datasets [57–62].

One of the more practical oversampling methods for increasing the number of cases in dataset is SMOTE (Synthetic Minority Oversampling Technique) [57]. SMOTE is an oversampling technique where the synthetic samples are generated for the minority class. This oversampling technique works by generating new instances from existing minority classes. They do not increase the number of majority classes. The new instances are not simply duplicates of existing minority cases. Instead, the algorithm extracts feature space samples for each target class and its nearest neighbors. The algorithm then generates new examples by combining characteristics of the target case with characteristics of its neighbors. This method builds the new characteristics available to each minority class and enlarge the samples.

The aim of the using oversampling technique in this study is to balance a class distribution between the minority classes with the majority class in only making machine learning model process. Most machine learning algorithms work best when the number of samples and distribution in each class are about equal. This is because most algorithms are designed to maximize accuracy and reduce errors. However, if the data set in imbalance, the result will get a pretty high accuracy just by predicting the majority class, but it fails to capture the minority class.

This study used an oversampling method in step of the model-building process. The oversampling method was not used in the evaluation step of final user satisfaction. Thus, the result of the oversampling method doesn't affect in the evaluation step of final user satisfaction.

CHAPTER 3

PREDICTING FINAL USER SATISFACTION USING MOMENTARY UX DATA AND MACHINE LEARNING TECHNIQUES

This chapter presents the methodological approaches, and the proposed framework is explained. In brief, this chapter is organized on the following sections: 3.1 Introduction 3.2 Proposed Method 3.3 Experiments 3.4 Results and Discussion, and 3.5 Summary.

3.1 Introduction

User experience (UX) evaluation investigates how people feel about using products or services and is considered an important factor in the design process. However, there is no comprehensive UX evaluation method for time-continuous situations during the use of products or services. Because user experience changes over time, it is difficult to discern the relationship between momentary UX and episodic or cumulative UX, which is related to final user satisfaction. This research aimed to predict final user satisfaction by using momentary UX data and machine learning techniques. The participants were 50 and 25 university students who were asked to evaluate a service (Experiment I) or a product (Experiment II), respectively, during usage by answering a satisfaction survey. Responses were used to draw a customized UX curve. Participants were also asked to complete a final satisfaction questionnaire about the product or service. Momentary UX data and participant satisfaction scores were used to build machine learning models, and the experimental results were compared with those obtained using seven built machine learning models.

3.2 Proposed Method

In the UX approach, classification analytics-built models rely on momentary UX data to predict user satisfaction levels. My proposed framework aims to predict final user satisfaction guided by momentary UX data to answer satisfaction-related questions. The evaluation process workflow architecture is shown in Figure 3-1.

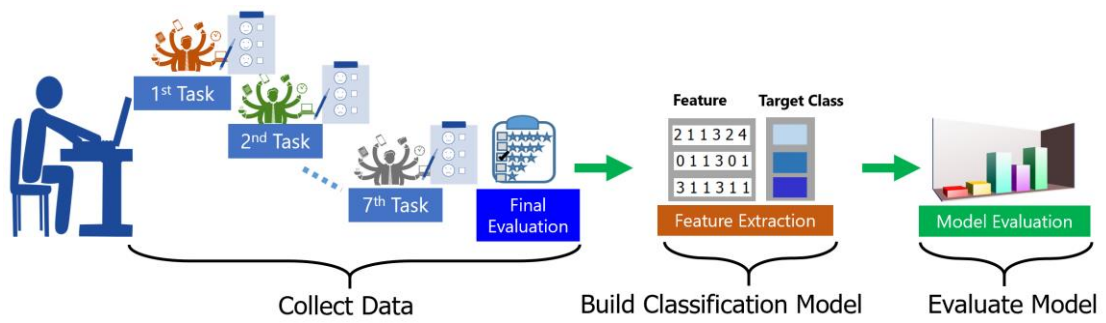


Figure 3-1. Workflow of my proposed evaluation process.

Our proposed framework was organized into three steps. First, data collection involved gathering and measuring information through satisfaction survey questions. Second, this study built a machine learning process to classify the final user satisfaction into different classes. To confirm the effectiveness of the proposed framework for a product and a service, two experiments were run, using different momentary UX data. Each experiment included momentary UX data from satisfaction survey questions representing changes in emotion. Experiment I included momentary UX data from visiting a travel agency site, a website service. Experiment II included UX usage data from Google Nest Mini [63], which is a smart AI speaker product. Finally, after the classification model was built, this study evaluated it using leave-one-out cross-validation and data splitting techniques.

3.3 Experiments

Two experiments were conducted: the first concerned use of a service in the form of a travel agency website; and the second concerned use of a product, namely Google Nest Mini.

3.3.1 Travel Agency Website (Service Group)

Fifty healthy university students aged 21 to 24 years were recruited as participants. The main reason for choosing people of this generation is that they typically have a better understanding of how to use products by themselves, with fewer gaps in relevant knowledge and education. This study used snowball sampling to recruit participants [64]. This is a network-based sampling method that starts with a convenience sample and incentivizes participants or respondents not only to participate in the survey themselves but also to ask their contacts in the target

population to participate. Snowball sampling is similar to peer-to-peer marketing, which is the best sampling method for new products or brands to reach new customers via word of mouth from one person to another [65]. For the main experiment, participants confirmed that they understood the procedure, and they responded to seven satisfaction survey questions concerning the travel agency website, as shown in Figure 3-2. Before they started the task, this study instructed them to use the agency website to find a place they wanted to visit once in their life. All participants appeared to perform the task attentively.

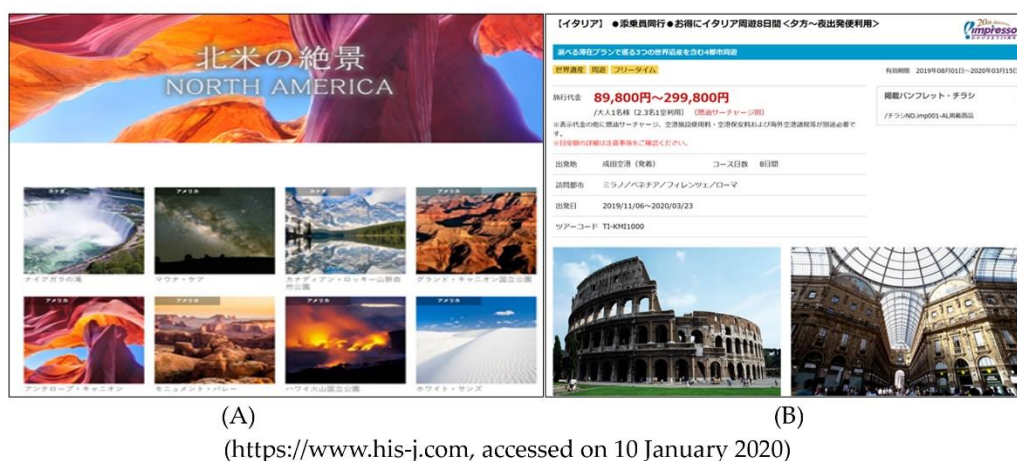


Figure 3-2. An example of a travel agency website interface (A) Attractive places (B) Tour information with a price range.

Ranges of rating scales can vary widely; for example, from 1–10 points to 1–100 points. One view is that overdetailed point scales may produce more variance. According to Spool (2015), enlarging a scale to see higher-resolution data may reveal that the data are meaningless [66]. Moreover, other evidence [67] suggests that a semantic differential scale may be appropriate for measuring satisfaction, with bipolar alternatives such as positive/negative, good/bad, helpful/unhelpful, and useless/valuable. These considerations led us to design and combine rating scales and the semantic differential for this evaluation. For momentary UX evaluation, this study used a 21-point scale that included negative values ranging from –1 to –10 and positive values ranging from 0 and +1 to +10 [7]. Adaption of the 21-point scale was done with reference to the UX graph form [11]. A new classification model was then

built using these data and the machine learning process. Finally, this study measured the classification model’s efficiency in terms of accuracy, precision, and recall.

Participants went through the six steps of their task in fixed order, completing the customized UX curve after each one (steps 1–6), as shown in Table 3-1 and the left side of Figure 3-3. This procedure is often implemented in actual service or product usage. Then, after completing the seventh step, they recorded “final satisfaction” based on several experiences, as shown in Figure 3-3. The seventh step was conducted for the study only and is not integral to the UX of the actual website itself. The data obtained in step 7 were used as a target variable for supervised learning. The right side of Figure 3-3 shows a final user satisfaction score of 4, based on a 21-point scale.

Table 3-1. Details of the travel agency (service) task.

Steps	Directions
1st	Find where you want to visit once in your life from menu of website. Then, evaluate user satisfaction of website.
2nd	Find the country of interest. Then, evaluate user satisfaction of website.
3rd	Visit the homepage of the travel agency website. Then, evaluate user satisfaction of website.
4th	Review information and read comments on the travel agency website. Then, evaluate user satisfaction of website.
5th	Select a place tour in which you are interested. Then, evaluate user satisfaction of website.
6th	Select and then purchase a favorite tour. Then, evaluate user satisfaction of website.
7th	Evaluate your final user satisfaction with the travel agency website of website.

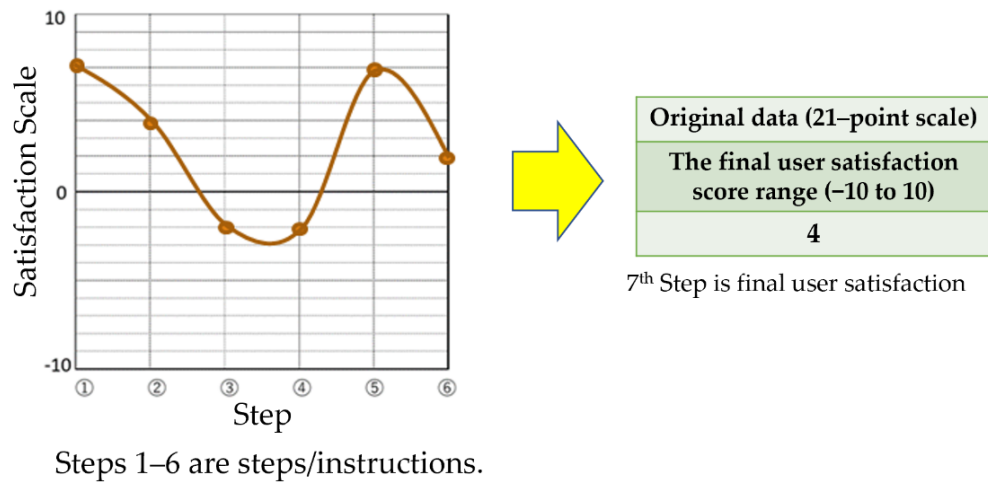


Figure 3-3. Customized UX curve for data collection.

Six satisfaction datapoints and one final satisfaction datapoint were obtained for each participant. The resulting dataset for building the model consisted of a 7×50 matrix (seven features, 50 participants). Because the original dataset revealed an accuracy score less than 0.50 when using my proposed framework, this study considered that results obtained using the 21-point scale were insufficiently accurate. Thus, this study scaled down, converting the original dataset into two datasets based on a seven-point scale and a five-point scale (see Table 3-2). Dataset I comprised seven classes (from -3 to 3), and Dataset II comprised five classes (from -2 to 2). After shrinking, in actual results, this study found that Dataset I comprised six classes due to zero samples in one class, while Dataset II still comprised five classes.

Table 3-2. Original data scale was reduced by scaling down to improve the predictive performance.

Meaning of Satisfaction Rating	Dataset I		Meaning of Satisfaction Rating	Dataset II	
	Original Data	After Shrinking		Original Data	After Shrinking
Extremely satisfied	10		Extremely satisfied	10	
	9	3		9	
	8			8	2
Satisfied	7		Satisfied	7	
	6	2		6	
	5			5	
Slightly satisfied	4		Satisfied	4	
	3			3	1
	2	1		2	
Neutral	1		Neutral	1	
	0	0		0	0
Slightly unsatisfied	-1		Unsatisfied	-1	
	-2	-1		-2	
	-3			-3	-1
Unsatisfied	-4		Unsatisfied	-4	
	-5	-2		-5	
	-6			-6	
Extremely unsatisfied	-7		Extremely unsatisfied	-7	
	-8			-8	-2
	-9	-3		-9	
	-10			-10	

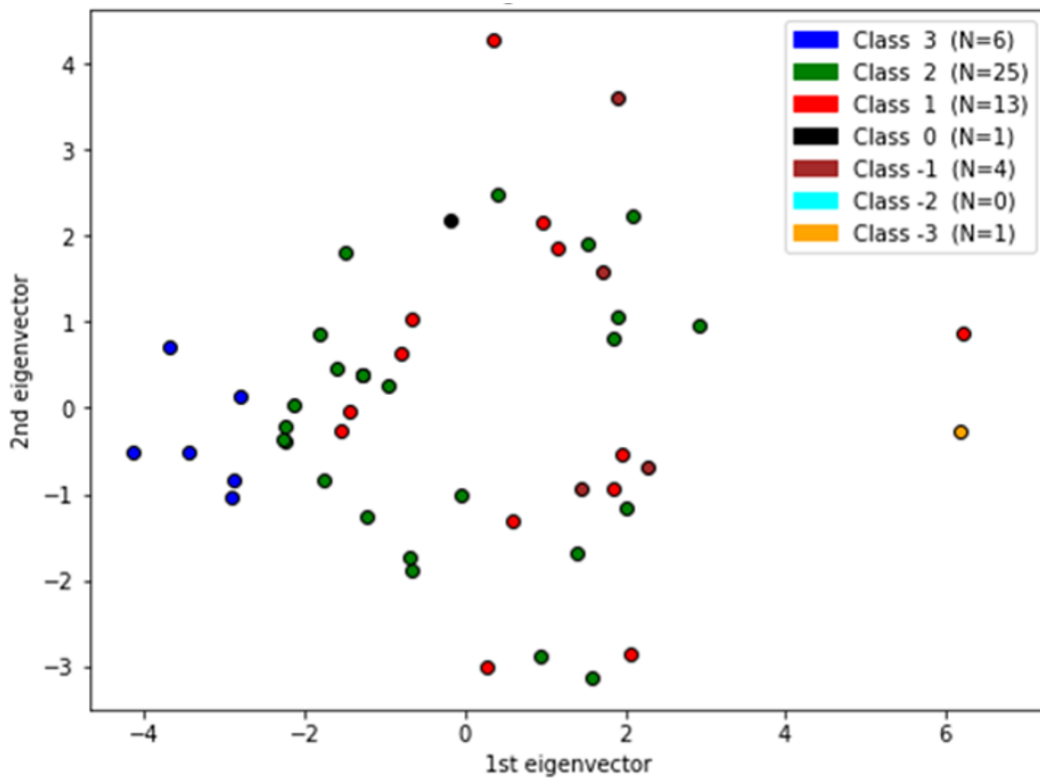


Figure 3-4. Example data points of Dataset I after shrinking.

Furthermore, in Dataset I, this study has already plotted data points between input variable and final satisfaction for representing a dataset. This plotting showed the relationship between momentary UX and final satisfaction as shown in Figure 3-4. This graph plot represents the scatter plot of 50 data points, in a 2-dimensional space, with a principal component analysis (PCA), where each point is color-coded based on its class assignment. The data points have been divided into 7 classes (level of final satisfaction), with the blue points representing target class 3, green blue points representing target class 2, red points representing target class 1, black points representing target class 0, dark red points representing target class -1, cyan points representing target class -2, and orange points representing target class -3. The X-axis represents the first eigenvector variable of PCA, and the Y-axis represents the second eigenvector variable of PCA. The different colors in the scatter plot represent the different classes, with each class having a unique label. This study found that that some points are close together within the same target class. This indicates that the data points in their classes are closer to each other than to the data points in other

classes. Although this graph plot helps to visualize the distribution of the data points and their patterns. It is not always easy to analyse the points and understand the relationships between them and target class. From this graph plot with many data points, it can be time-consuming and difficult to identify patterns and relationships manually. However, the use of machine learning algorithms can help in this process more quickly and accurately, by automating the analysis of the data points.

From Figure 3-4, this study found that the number of samples per class increased when the number of classes decreased. One advantage of shrinking is that the increased number of samples per class can be useful for building machine learning models. Before processing the dataset, this study used variance inflation factor (VIF) to check for multicollinearity of predictor variables (six answers about satisfaction score from the six-item questionnaire) where the dependent variable was final user satisfaction. VIF values exceeding 5 or 10 indicate problematic collinearity [68]. This study confirmed that all VIF values were under 5.

3.3.2 Google Nest Mini (Product Group)

Twenty-five university students aged 21–24 years were recruited as participants. In this experiment, the task was to remove the smart speaker (Google Nest Mini, as shown in Figure 3-5) from the box, set it up, and start using it through 12 steps in a fixed order. At the end of each step, the participants recorded their satisfaction on a form based on the customized UX curve. The data from the first experiment (service group) show good accuracy when rescaled in the form of a UX curve. Therefore, in this experiment, this study used a new form based on the customized UX curve with a 15-point scale ranging from -7 to $+7$. Their final user satisfaction for the product was recorded after the experimental task was completed.



Figure 3-5. Using Google Nest Mini.

The task assumed that a new smart speaker was purchased, removed from the box, set up, and made ready for use. The participants proceeded through each step while referring to the enclosed instructions. The 12 steps of the task are shown in Table 3-3.

Table 3-3. Details of the Google Nest Mini (product) task.

Steps	Directions
1st	Browse nest mini on Google Store.
2nd	Open the box, take out the smart speaker.
3rd	Read the instructions, turn on the smart speaker.
4th	Install the Google app on your smartphone, select an account.
5th	Connect apps and smart speakers using Wi-Fi connection with smartphone location information and router.
6th	Open a Wi-Fi connection between the smart speaker and router using the app.
7th	Follow the instructions in the app and using voice recognition on the smart speaker.
8th	Connect and set various setting services in the app.
9th	Play music on a smart speaker that has been set up.
10th	Set alarm timers with smart speakers.
11th	Listen to weather forecasts with smart speakers.
12th	Evaluate your final user satisfaction with the Google Nest Mini.

Eleven satisfaction datapoints and one final satisfaction datapoint were obtained for each participant, as shown in Figure 3-6. The resulting dataset for building the model consisted of a 12×25 matrix (12 features, 25 participants).

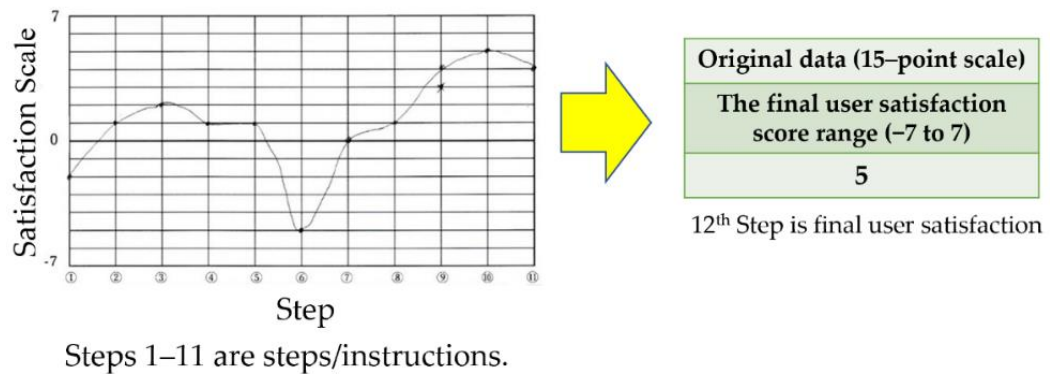


Figure 3-6. Example of a customized UX curve with participant’s final satisfaction.

The original dataset from the preliminary experiment showed an accuracy score less than 0.50 when using my proposed framework, indicating that sufficiently accurate results were not obtained using the 15-point scale. Thus, this study scaled down, converting the original dataset into two datasets, including a seven-point scale and a five-point scale (see Table 3-4). Dataset I comprised seven classes (from -3 to 3), and Dataset II comprised five classes (from -2 to 2). The actual results after shrinking showed three classes in Dataset I, and two classes in Dataset II due to zero samples in some classes.

Table 3-4. Original data scale was reduced by scaling down to improve the predictive performance.

Meaning of Satisfaction Rating	Dataset I		Meaning of Satisfaction Rating	Dataset II	
	Original Data	After Shrinking		Original Data	After Shrinking
Extremely satisfied	7	3	Extremely satisfied	7	2
	6			6	
	5			5	
Satisfied	4	2	Satisfied	4	1
	3			3	
Slightly satisfied	2	1	Satisfied	2	1
	1			1	
Neutral	0	0	Neutral	0	0
Slightly unsatisfied	-1	-1	Unsatisfied	-1	-1
	-2			-2	
	-3			-3	
Unsatisfied	-4	-2	Extremely unsatisfied	-4	-2
	-5			-5	
Extremely unsatisfied	-6	-3	Extremely unsatisfied	-6	-7
	-7			-7	

Before processing the dataset, this study used VIF to check multicollinearity for predictor variables (11 answers about satisfaction from the 11-item questionnaire), where the dependent variable was final user satisfaction. All VIF values were under 5.

In the current study, the first stage of the experiment to predict final user satisfaction using momentary UX was through the satisfaction survey form. For website evaluation, this study used a satisfaction survey form at the bottom of the webpage to be filled out after the completion of each task. For product evaluation, this study requested that the satisfaction survey be manually evaluated during product set up. This study found that it may not be easy to gather these satisfaction scores in an actual product evaluation situation. Future research could use other techniques for

product evaluation to collect momentary UX data, such as facial expression or gaze data.

3.3.3 Evaluation

Several studies have attempted to demonstrate that SVM and KNN algorithms can perform well with small datasets [69,70]. Thus, this study selected seven appropriate machine learning algorithms: SVM [46] including SVM with linear kernel, SVM with sigmoid kernel, SVM with RBF kernel, SVM with polynomial kernel, logistic regression [50], K-nearest neighbors (KNN) [49], and multilayer perceptron (MLP) [52]. Each model was trained by these algorithms using the datasets, and then classification models were tuned with various hyperparameters while evaluating machine learning models by a random search method to provide the best performance [71].

In Experiments I and II (Travel agency website and Google Nest Mini, respectively), an unequal distribution of classes within the datasets was observed; for example, the ratios of seven classes in Experiment I (class “-3”, class “-2”, class “-1”, class “0”, class “1”, class “2”, and class “3”) were 1, 0, 4, 1, 13, 25, and 6, respectively. As indicated in Section 2.5, multiple techniques exist for dealing with imbalanced sample distributions. Oversampling the minority class is one such approach used in data science [57]. This can be achieved by synthesizing new examples from the minority class in the training dataset prior to fitting a model. This can balance the class distribution and be highly effective for the created model. For example, in Experiment I, the number of samples increased from 50 to approximately 150. In Experiment II, the number of samples increased from 25 to approximately 36. By checking that the number of minority and majority class samples were equal, this study confirmed that the imbalance disappeared. The most basic method involves creating examples from the minority class; even though these examples add no new information to the model, they can be created by combining existing data. Thus, this study selected the synthetic minority oversampling technique, or SMOTE [57], based on results of the preliminary experiment.

One issue is that oversampling before performing cross-validation allows leakage from the test data into the training data. Because of the overlap between

training and test data, this can lead to an optimistic bias in performance evaluation, as shown in Figure 3-7. This is why this study used SMOTE oversampling techniques inside the cross-validation (CV) loop in the evaluation step. Oversampling inside the CV loop [72] is appropriate for revealing the model's true performance.

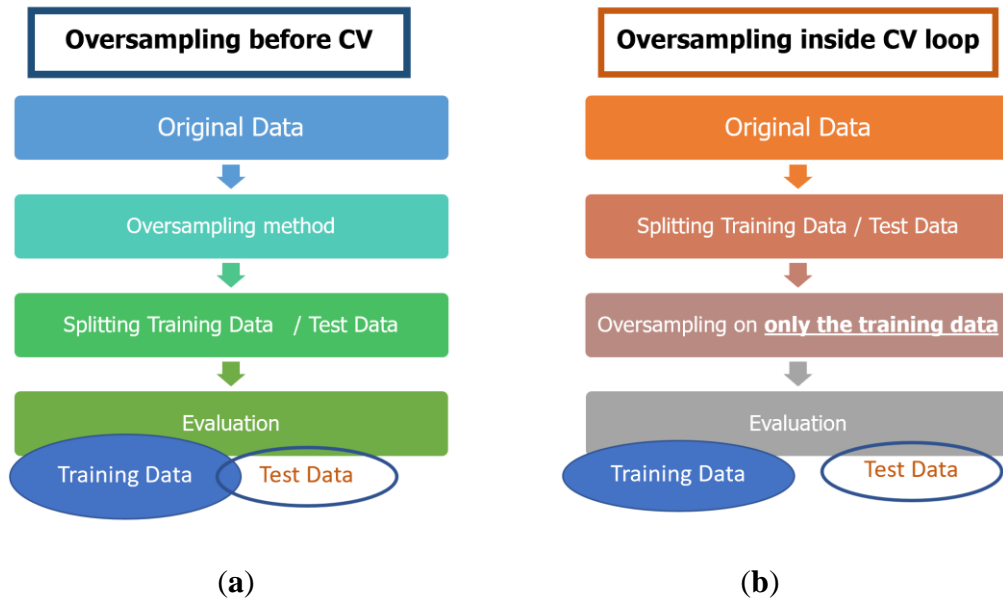


Figure 3-7. Comparison between (a) oversampling before the cross-validation loop and (b) oversampling inside the cross-validation loop.

In the evaluation step, two conventional methods were used to evaluate the performance of each classification model as follows: (1) leave-one-out cross-validation (LOOCV) [73], as shown in Figure 3-8 and (2) validation with training (80%)/test (20%) splitting by three indices: accuracy, recall, and precision, as shown in Figure 3-9 [74].

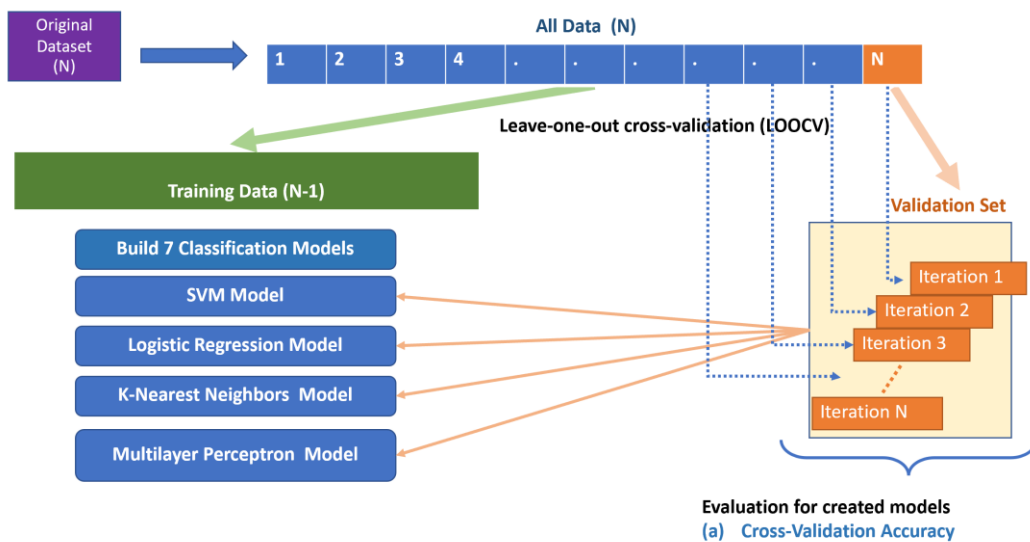


Figure 3-8. Evaluation workflow for created models using leave-one-out cross-validation (LOOCV).

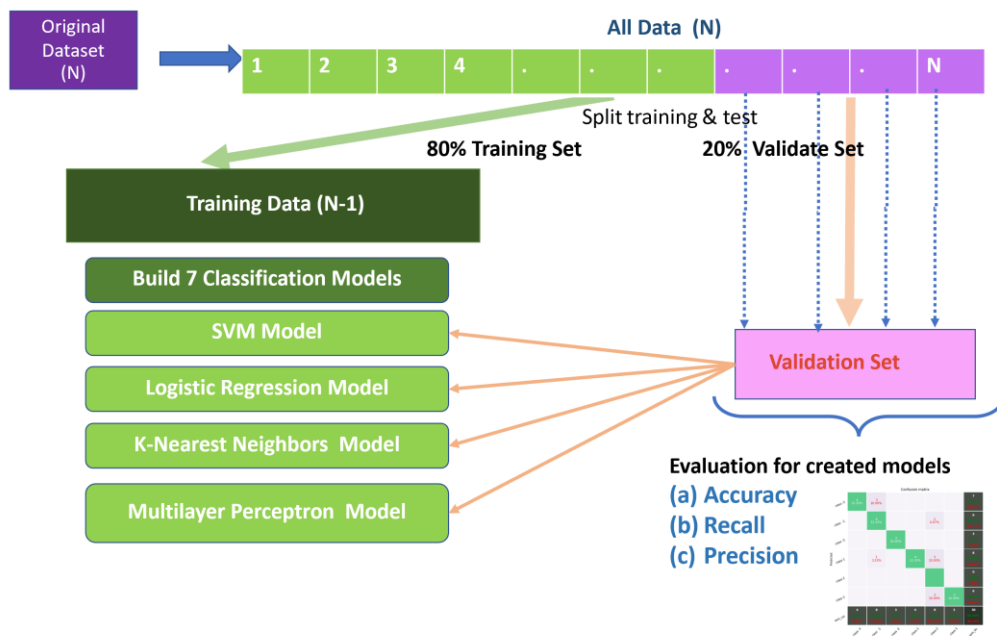


Figure 3-9. Evaluation workflow for created models using data splitting.

Moreover, performance can be measured using the percentage of accuracy observed in both data sets to conclude on the presence of overfitting. Overfitting is characterized by high accuracy for a classifier when evaluated on the training set but low accuracy when evaluated on a separate test set [75]. In my experiments, this study

confirmed that the accuracy score of the testing set was nearest when compared with that of the training set. Thus, all models were not overfitting.

In data science, data splitting according to an 80/20 ratio between the training set and test set provides the most practice for the machine learning model. The performance of each classification model is reported in terms of accuracy, precision, and recall. Accuracy is the most intuitive performance measure; namely, the ratio of correctly predicted observation to total observations. Precision is the ratio of correctly predicted positive observations to total predicted positive observations. Recall is the ratio of correctly predicted positive observations to all observations in the actual class.

3.4 Results and Discussion

3.4.1 Results from Experiment I: Service Usage

In Experiment I (travel agency website), this study measured the combination of oversampling techniques with the created machine learning models, including SVM with polynomial kernel, SVM with RBF kernel, SVM with linear kernel, SVM with sigmoid kernel, K-nearest neighbors, logistic regression, and multilayer perceptron techniques. This study then compared the performance between polynomial kernel SVM and polynomial kernel SVM with oversampling.

Table 3-5 shows a comparison of the classification models' performance for a combined synthetic minority oversampling technique (SMOTE) and machine learning techniques for two datasets (seven classes and five classes). Performance was measured by LOOCV and splitting the data into two subsets: training and test. The accuracy score can range between 0.00 and 1.00; a higher value indicates higher accuracy.

Table 3-5. Performance scores of created models with oversampling from travel agency website.

Scores	Dataset	SVM	SVM	SVM	SVM	KNN	LR	MLP	
		Poly	RBF	Linear	Sigmoid				
LOOCV	Cross-Validation Accuracy	I (7 Classes)	0.93	0.79	0.80	0.50	0.84	0.72	0.80
		II (5 Classes)	0.90	0.87	0.88	0.45	0.80	0.87	0.84
Split for training /test (80/20)	Accuracy	I (7 Classes)	0.87	0.60	0.73	0.33	0.73	0.60	0.67
		II (5 Classes)	0.93	0.93	0.86	0.54	0.86	0.89	0.93
Split for training /test (80/20)	Recall	I (7 Classes)	0.87	0.60	0.73	0.33	0.73	0.60	0.67
		II (5 Classes)	0.93	0.93	0.86	0.54	0.86	0.89	0.93
Split for training /test (80/20)	Precision	I (7 Classes)	0.90	0.64	0.85	0.21	0.75	0.70	0.65
		II (5 Classes)	0.96	0.95	0.87	0.42	0.88	0.90	0.93

SVM = support vector machine; Poly = polynomial kernel; LR = logistic regression; KNN = K-nearest neighbors; MLP = multilayer perceptron.

SVM with polynomial kernel using Dataset I (seven classes) had the highest LOOCV accuracy score (0.93). Moreover, each model was evaluated by splitting the training and test set techniques. SVM with polynomial kernel using Dataset I provided the highest accuracy (0.87), recall (0.87), and precision (0.90) scores.

SVM with the polynomial kernel using Dataset II (five classes) also showed the highest LOOCV (0.90). It also provided the highest scores for accuracy (0.93), recall (0.93), and precision (0.96).

Based on the results shown in Table 3-5, I then focused on SVM with the polynomial kernel. Table 3-6 summarizes the results of comparisons between polynomial kernel SVM and polynomial kernel SVM with oversampling into the cross-validation step. Polynomial kernel SVM with oversampling into the cross-validation step provided the highest cross-validation accuracies (0.93 and 0.90) on Datasets I and II, respectively. Moreover, the accuracy of polynomial kernel SVM

with oversampling into the cross-validation step was higher than for polynomial kernel SVM without oversampling.

Table 3-6. Comparison of performance between polynomial kernel SVM and polynomial kernel SVM with oversampling into the cross-validation step (travel agency website).

Model Performance		Dataset I: 7 Classes (7-Point Scale Data)		Dataset II: 5 Classes (5-Point Scale Data)	
		Polynomial Kernel SVM	Polynomial Kernel SVM with Oversampling into the Cross- Validation Step	Polynomial Kernel SVM	Polynomial Kernel SVM with Oversampling into the Cross- Validation Step
LOOCV	Cross- Validation Accuracy	0.48	0.93	0.72	0.90
Split for training/test (80/20)	Accuracy	0.40	0.87	0.70	0.93
	Recall	0.40	0.87	0.70	0.93
	Precision	0.65	0.90	0.61	0.96

Comparing classification results from two datasets differing in the number of classes revealed differences in accuracy scores between Datasets I (seven classes) and II (five classes). Overall, the accuracy with Dataset II was better than that with Dataset I, which suggests that the accuracy depends on the number of classes.

3.4.2 Results from Experiment II: Product Usage

In Experiment II (use of Google Nest Mini), the models were validated by LOOCV. The results show that the SVM with polynomial kernel model provided the highest accuracy (Table 3-7). The correct answer rate when using the SVM with polynomial kernel method was the highest, at 0.76, suggesting the high effectiveness of this proposed method. Moreover, comparison of the classification result from two datasets differing in the number of classes revealed differences in accuracy scores between Datasets I (seven classes) and II (five classes). Furthermore, the accuracy score with Dataset II was higher than that with Dataset I. Taken together, these results confirm that the proposed framework is feasible, and it is possible to predict final user satisfaction guided by momentary UX data to answer product-satisfaction-related questions.

Table 3-7. Performance scores of created models with oversampling from Google Nest Mini usage.

Scores	Dataset	SVM	SVM	SVM	SVM	KNN	LR	MLP	
		Poly	RBF	Linear	Sigmoid				
LOOCV	Cross-Validation Accuracy	I (7 Classes)	0.60	0.52	0.52	0.16	0.52	0.44	0.40
		II (5 Classes)	0.76	0.68	0.64	0.32	0.68	0.68	0.48
Split for training /test (80/20)	Accuracy	I (7 Classes)	0.88	0.80	0.80	0.20	0.40	0.20	0.60
		II (5 Classes)	0.86	0.60	0.80	0.40	0.40	0.40	0.60
Split for training /test (80/20)	Recall	I (7 Classes)	0.88	0.80	0.80	0.20	0.40	0.20	0.60
		II (5 Classes)	0.86	0.60	0.80	0.40	0.40	0.40	0.60
Split for training /test (80/20)	Precision	I (7 Classes)	0.92	0.87	0.80	0.60	0.37	0.20	0.67
		II (5 Classes)	0.89	0.60	0.85	0.53	0.53	0.53	0.87

SVM = support vector machine; Poly = polynomial kernel; LR = logistic regression;
 KNN = K-nearest neighbors;
 MLP = multilayer perceptron.

Based on the results shown in Table 3-7, this study focused on SVM with the polynomial kernel. Table 3-8 summarizes the results of comparisons between polynomial kernel SVM and polynomial kernel SVM with oversampling into the cross-validation step. Polynomial kernel SVM with oversampling into the cross-validation step provided the highest cross-validation accuracies (0.60 and 0.76) with Datasets I and II, respectively. Moreover, the accuracy of polynomial kernel SVM with oversampling into the cross-validation step was higher than that of polynomial kernel SVM without oversampling.

Table 3-8. Comparison of performance between polynomial kernel SVM and polynomial kernel SVM with oversampling into the cross-validation step (Google Nest Mini usage).

Model Performance		Dataset I: 7 Classes (7-Point Scale Data)		Dataset II: 5 Classes (5-Point Scale Data)	
Score		Polynomial Kernel SVM	Polynomial Kernel SVM with Oversampling into the Cross- Validation Step	Polynomial Kernel SVM	Polynomial Kernel SVM with Oversampling into the Cross- Validation Step
LOOCV	Cross- Validation Accuracy	0.52	0.60	0.60	0.76
Split for training/test (80/20)	Accuracy	0.60	0.88	0.60	0.86
	Recall	0.60	0.88	0.60	0.86
	Precision	0.50	0.92	0.80	0.89

When comparing the classification results from two datasets differing in the number of classes, differences in cross-validation accuracy between Datasets I (seven classes) and II (five classes) emerged. Overall, the cross-validation accuracy with

Dataset II was better than with Dataset I, which suggests that accuracy depends on the number of classes.

The present study was designed to predict final user satisfaction by machine learning techniques based on momentary UX Curve data. This study has several research implications, as discussed below.

3.4.3 Discussion I: Service Usage with Travel Agency Site

In the evaluation of service usage, performed with a travel agency website, this study found that the SVM with the polynomial kernel algorithm provided the highest cross-validation accuracy, at 0.93; all other algorithms scored lower, with the next-highest being KNN, at 0.84, and slightly higher than the rest. Thus, SVM and KNN appear to be good at predicting final user satisfaction. To test the performance of these machine learning methods, this study considered recall and precision on 20% testing and 80% training data. For travel agency website usage, SVM with polynomial kernel with both five and seven classes yielded the highest recall (0.87) and precision (0.90) among the seven candidate algorithms. Several previous studies have reported that recall and precision with imbalanced datasets may be poor [76] and lead to an optimistic bias in performance validation even after oversampling datasets [72]. In this study, however, recall and precision with oversampling resulted in data that were better than the original data, as shown in Tables 3-6 and 3-8. Two possible explanations for these improved results are, first, that this study optimized the machine learning model by finding the best parameters for the dataset, and second, this study performed oversampling during the cross-validation loop, which is the correct way to handle imbalanced data. Hence, to reveal the true performance of the model, it is appropriate that oversampling be conducted inside the cross-validation loop [72].

It is conceivable that the dimension of a dataset might be one factor influencing predictive performance. Some authors have reported that SVM and KNN might perform well on small datasets [69,70]. Moreover, in this study, accuracy was consistently higher with five classes (Dataset II) than that with seven classes (Dataset I), which suggests that accuracy depends on the number of classes. However, the

ability of SVM and KNN to predict final user satisfaction should be further examined using other kinds of services.

3.4.4 Discussion II: Product Usage with Google Nest Mini

To evaluate product usage, this study used a Google Nest Mini task and found that SVM with polynomial kernel with five classes had the highest cross-validation accuracy, at 0.76. Moreover, SVM with polynomial kernel with five classes had the highest recall (0.86) and precision (0.89) of the seven candidate algorithms. Again, SVM with polynomial kernel performs better when the dataset has few classes.

Comparing oversampling and no oversampling revealed that the former five classes resulted in high cross-validation accuracy, at more than 76%. In this context, it is noteworthy that the use of momentary UX during service usage and classification models had the highest predictive accuracy for final user satisfaction.

With regard to assessment of the use of momentary UX and classification models, the most interesting finding was that momentary UX and machine learning can predict final user satisfaction, which is important for users' decisions about further use or whether they recommend products or services to other people. One unanticipated finding was that polynomial kernel SVM with an oversampling technique achieved the best classification accuracy (more than 90%). These results match those of machine learning studies where polynomial kernel SVM also performed better with oversampling because a higher degree of polynomial kernel, which is one of the parameters of the SVM algorithm, allows a more flexible decision boundary [77].

In this investigation, the aim was to predict final user satisfaction using momentary UX data and machine learning techniques. The results show that the machine learning process can help in predicting final user satisfaction in at least two contexts: Experiment I, service usage, and Experiment II, product usage.

The strongest feature of the proposed method is that it is based on data supporting the idea of the relationship between UX time intervals: momentary UX might affect episodic UX (final user satisfaction). Due to practical constraints, the preliminary study did not extend to evaluations involving a wider variety of products or services,

and so this study is cautious about extrapolating to other situations. Nevertheless, the study has demonstrated significant relationships between momentary UX data and final user satisfaction, which is consistent with the argument of Feng and Wei (2019) that a first-time user experience is generally seen as a factor related to long-term user experience [16].

3.5 Summary

Customer relationship management is a tool to improve both the business and customer satisfaction with products or service [21]. It is generally accepted that CRM provides essential feedback and reflective data from the customers' perspective, including their opinions, preferences, and past UX regarding to products or services. These data and information are used to improve communication with customers to enhance value and satisfaction.

This study aimed to predict final user satisfaction using momentary UX data and machine learning techniques. The findings indicate that machine learning techniques such as polynomial kernel SVM can comprehend participants' momentary UX and make better predictions than six other machine learning algorithms concerning their final user satisfaction. Moreover, machine learning integrated with the oversampling technique yielded higher accuracy than that without oversampling. This technique integrated with the oversampling method could deal with imbalanced classes by synthesizing new samples and adjusting the class distribution of a data set.

The study was divided into two different experiments, the first concerning evaluation of a service (travel agency website), and the second concerning a product (Google Nest Mini). For service usage with the travel agency site, the results showed that SVM with polynomial kernel had the highest cross-validation accuracy, at 0.93. For product usage with Google Nest Mini, the results showed that SVM with polynomial kernel again had the highest cross-validation accuracy, at 0.76. The proposed method, therefore, shows promise for accurately predicting final user satisfaction using machine learning techniques; it facilitates classification and estimation of final user satisfaction based on momentary UX Curve data.

3.5.1 Theoretical Implications

Our study has contributed to knowledge in the field in various ways.

First, my contribution relates to the outcomes of UX. Data of time sequence questionnaire or UX curve is often difficult to understand and analyze. This study found the relationship between momentary UX and episodic or cumulative UX, which is related to final user satisfaction. My study indicates how understanding of momentary UX data can help determine the final user satisfaction during the changes of UX curve [34].

Second, machine learning like SVM could accurately predict final user satisfaction and contribute towards developing products and services by analyzing the UX obtained from CRM [21]. Hence, this study need to monitor the momentary UX carefully.

Third, combining and integrating machine learning and oversampling techniques could constitute a new approach for improving the predictive accuracy of final user satisfaction. This finding shows the relevance of considering UX in the analysis of customer satisfaction.

3.5.2 Practical Implications

The majority of businesses that consider adopting a CRM system are looking for a way to improve the quality and consistency of their relationships with customers and build customer loyalty. UX data from CRM has gradually become the main source of businesses' sustainable competitive advantage. In terms of the practical implication of this study, the result of my proposed method is that it is based on data supporting the idea of the relationship between UX time intervals: momentary UX might affect episodic UX (final user satisfaction). The understanding of those aforementioned relationships could provide the best UX for customers, build a good brand image, and launch customer-centric marketing campaigns. It can help businesses to achieve a better user satisfaction and the goals of sustaining long-term competitive advantages. For example, in service industry, the product or service developers could discover the worst points of products or services at which a customer requires assistance during product or service usage. They need to ensure that

customers can finish their transaction without difficulty in different usage situations. As a result, the understanding of momentary UX could boost overall customer satisfaction as well as repeat purchase rate, maintain long-term sustainable customer satisfaction and achieve sustainability. It could help to understand customers better and thus enhance communication with stakeholders with regard to efficiency, performance, and sustainability of products or services.

CHAPTER 4

USING RANDOM ORDERING IN USER EXPERIENCE TESTING TO PREDICT FINAL USER SATISFACTION

In this chapter, the researcher presents the using of random ordering of user tasks in user experience testing to predict final user satisfaction. In brief, this chapter is organized on the following sections: 4.1 Introduction 4.2 Proposed Method 4.3 Experiment 4.4 Results and Discussion, and 4.5 Summary.

4.1 Introduction

Final user satisfaction based on user experience (UX) is critically important for product evaluation and users' decision-making about whether or not to continue to use or recommend a product or service to others [1,78]. Predicting final user satisfaction is thus key for many use cases in product evaluation.

Many product developers have attempted to evaluate final user satisfaction by gathering and analyzing users' historical behavior. One example of this can be found in the analysis of consumer behavior on websites. With each visitor possessing their own unique website usage habits, product developers have aimed to decode visitor's usage and apply the data to predict final user satisfaction. However, extrapolating final user satisfaction from visitor's website usage can be difficult. Among the variety of evaluation methods used to assess final user satisfaction, each with its own advantages and drawbacks, one of the most significant methods is user satisfaction analysis. Previous studies have reported that user satisfaction analysis helps product developers identify how final user satisfaction affects decisions related to a product or service [1]. Traditionally, product developers have relied on collecting information on user satisfaction from questionnaire surveys administered from the first to the final stage of usage [29]. In the case of a website, final user satisfaction prediction is also based on the volume of activity on a website in the form of page views and hits, and the order of visits using the website's cookie technology and user logs [79], as shown in Figure 4-1.

User satisfaction analysis has been adopted to evaluate UX. Several studies have defined UX as any interaction a user has with a product or service [16,80]; for example, how the product looks, how its elements influence the user, how it makes them feel, and how they interact with it. Understanding momentary UX data can help determine final user satisfaction after usage. Recent studies have reported that UX is an important variable for building customer satisfaction post-purchase in the product industry [24,25]. All UX variables can affect a customer’s satisfaction [12]. In UX evaluation (UXE), users’ questionnaire responses are used to draw a customized UX Curve to represent changes in user satisfaction. After usage, the users are asked to complete a final questionnaire about their satisfaction with the product. Answers relating to final user satisfaction are often expressed on a scale of negative to positive values and are also often based on the order in which users interact with the product. Thus, user satisfaction assessments consider both the satisfaction score and order of users’ activities. Further, UXE is an important tool for improving the success of a product and forms a critical strategy by which to seek feedback from users with the goal of improving their final user satisfaction.

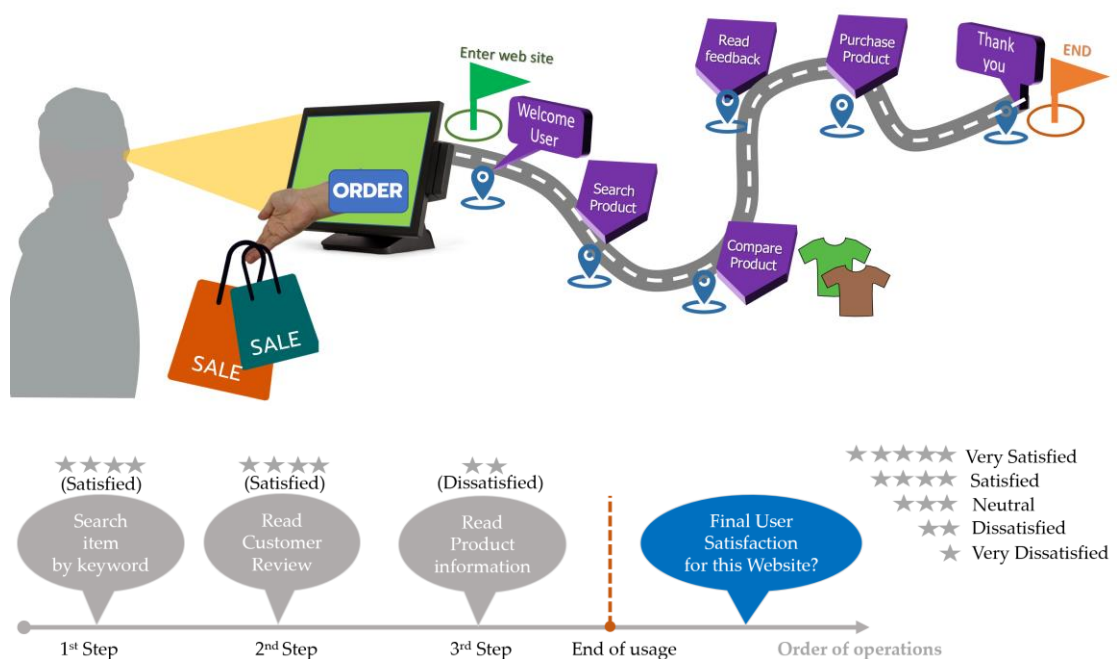


Figure 4-1. User satisfaction while users are on a shopping website.

To have a great experience using a product or service, customers need to be able to easily access all the essential functions they need. In addition to being aesthetically pleasing, a product should also be simple and easy to navigate. Poor product design can discourage a user from spending time on the product. Given the vastness of the product industry, users who have a bad experience using one product can easily find what they need with another, and will not waste their time on products with bad UX. In contrast, a good UX will lead customers to recommend the products or services to other people. As a consequence, it is now clear to companies in many business sectors that it is impossible to design products and services without contribution from representative users.

However, a major problem with UXE is that it does not necessarily reflect real UX in terms of the order in which users perform tasks on a product. In UXE, a UX designer must design an order of tasks for the user to perform and related questionnaires beforehand. To measure UX during UXE, users are asked to follow a set of sequential tasks and complete the questionnaire about their feelings related to the product or service in a fixed order. In reality, however, it is impossible to predict what the user will and wants to do. Furthermore, it is generally accepted that different users will choose a different set of sequential tasks to perform when using a product or service. This can clearly be seen when users visit e-commerce websites, where visitors can buy goods on any page on the site. While some visitors may navigate from the home page, others can do so from almost any other page on the site. Thus, in real life, the UX of a website involves randomly ordered tasks.

UXE based on the random ordering of tasks is challenging because of the difficulty related to classifying these data. Although a number of studies have conducted UXE of subjects, none have examined the sequence of actions in a UX that involves randomly ordered tasks. Further, to my knowledge, all available UXE approaches are based on a fixed sequential order of tasks, which means most UXE approaches are designed to examine actions performed in the same order, such as that reported in my previous research [1].

This chapter proposes a new approach using machine learning techniques to predict final user satisfaction based on UX related to randomly ordered tasks. This

chapter show that accounting for the sequence of actions may improve the prediction of final user satisfaction.

4.2 Proposed Method

In the previous research [1], this study designed an experiment with a fixed order of tasks in which participants were asked to complete all tasks, from the first to final task, in sequential order. For example, in an online shopping evaluation, each user had to visit webpages A, B, C, and D in that order. However, this is not reflective of the real-life UX, where users are free to use the product however they want. Some users may visit the web pages in different order to the specified A, C, B, and D, as shown in Figure 4-2 (right). This study called this case randomly ordered, as opposed to fix ordered.

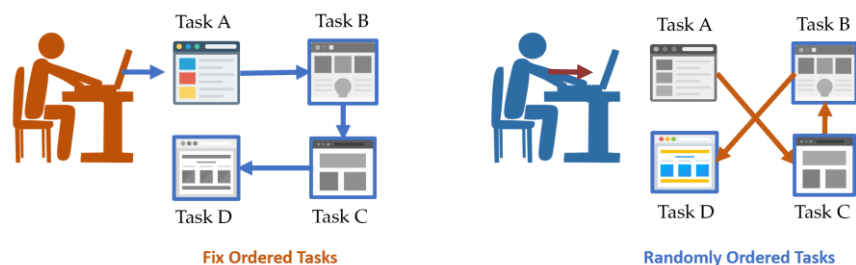


Figure 4-2. Comparison between fix ordered tasks and randomly ordered tasks performed when users use an online shopping website.

As shown in Figure 4-2, this chapter used randomly ordered tasks. This chapter had no control over when participants would visit any particular page. While some may start from the first page, others may start at a different page on the website, as shown in Figure 4-3.

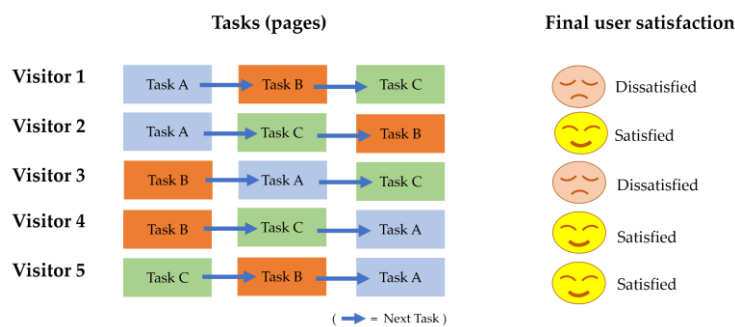


Figure 4-3. Example of randomly ordered tasks when users use an online shopping website.

In the UX approach, classification analytics-built models rely on random ordered UX data to predict user satisfaction levels. This proposed framework aims to predict final user satisfaction guided by randomly ordered UX data to answer satisfaction related questions. The evaluation process workflow architecture is shown in Figure 4-4.

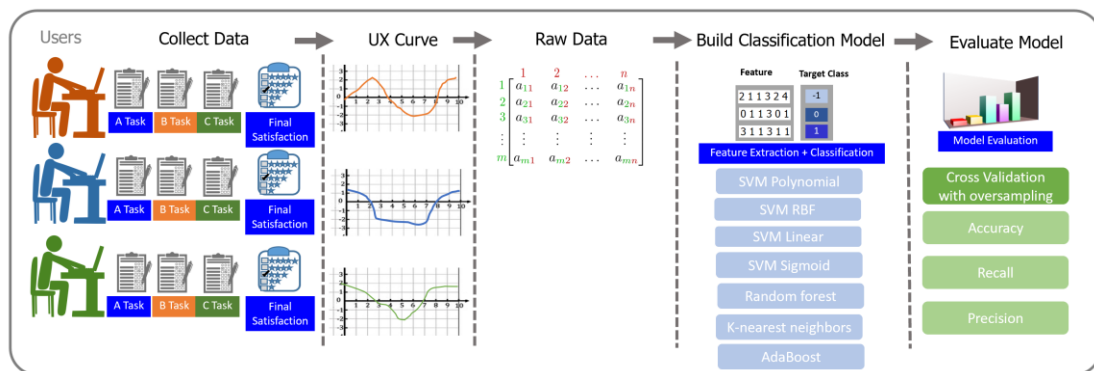


Figure 4-4. Workflow of my proposed evaluation process.

This proposed framework was organized into three main steps. First, UX data were collected by gathering information and calculating scores in satisfaction survey questionnaires completed by users who performed randomly ordered tasks. Second, this proposed framework built a machine learning framework to classify final user satisfaction into different classes. To confirm its effectiveness, the proposed framework was applied to an experiment using randomly ordered UX data from users' visits to a travel agency website. Randomly ordered UX data from the satisfaction survey questionnaire were used to represent changes in emotion. Finally, the classification model was evaluated using leave-one-out cross-validation and data splitting techniques.

In this proposed framework, this chapter wanted to compare the use of randomly ordered UX data between Dataset I, which did not account for the actual task order, and Dataset II, which did, under my specific conditions. Consequently, this study needed to collect and generate own original UX dataset. Details of the framework are explained in Section 4.3.2.

4.3 Experiment

In the previous study, the study collected responses to a questionnaire about users' feelings related to a product or service while they performed fixed ordered tasks [1]. In the present study, this study used this data to simulate UX after shuffling the original order of tasks performed in the fixed order experiment. The aim of these preliminary experiments was to compare the real-life results of the fixed order task with the simulated results of the shuffled order task.

Next, this study performed the main experiment, in which users conducted all tasks in random order. The actual order or sequence of actions was recorded during usage.

The preliminary and main experiments are described in detail below.

4.3.1 Preliminary Experiments

This study conducted preliminary experiments to determine the importance of task order in UX and how it can affect the final user satisfaction of customers while they are performing tasks on a website or product. To do this, this study simulated UX after shuffling the order of tasks performed in my previous experiments [1]. The participants of the original experiments were university students who were asked to evaluate a travel agency website (Preliminary Experiment I) or Google Nest Mini smart speaker (Preliminary Experiment II) during usage by completing a satisfaction survey. Responses were used to draw a customized UX curve. Participants were also asked to complete a final satisfaction questionnaire about the product or service.

4.3.1.1 Preliminary Experiment I: Travel agency website

In preliminary experiment I, this study used data from a study in which participants were asked to evaluate a travel agency website [1]. The aim of the prior study was to predict final user satisfaction based on the satisfaction score given by users while using the website. Participants had to respond to seven satisfaction survey questions concerning the travel agency website. Before starting the task, the details were told them that their goal was to use the website to find a place they wanted to visit once in their lifetime. All participants appeared to perform the task attentively.

Fifty participants completed the six steps required to achieve the goal in fixed order, allowing us to produce a customized UX curve depicting each one (steps 1–6). This procedure is often implemented in actual service or product usage to obtain the momentary UX, as shown in Figure 4-5. After completing the seventh step, the participants recorded their final satisfaction based on several experiences. The seventh step was conducted to obtain final user satisfaction as an indicator of episodic UX for this experiment only. The final satisfaction data obtained in step 7 were used as the target class variable for supervised learning. In this study, the number of classes was five.

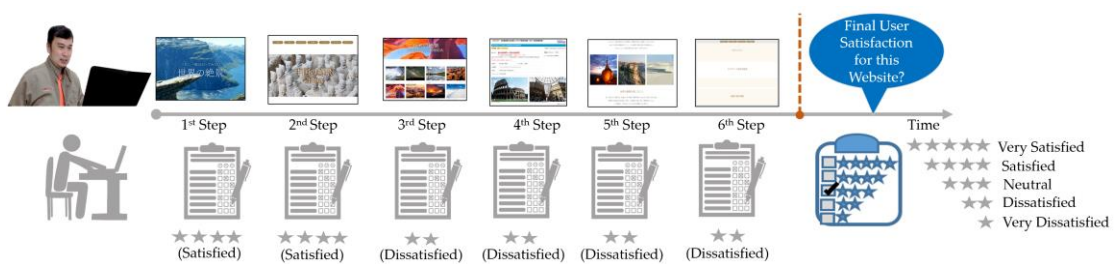


Figure 4-5. Evaluating user satisfaction while using a travel agency website.

As the aim of preliminary experiment I was to study the usefulness of sequential ordering for identifying factors predictive of final user satisfaction, this study created two different datasets with which to build my machine learning models (Figure 4-6). Dataset W1 contained UX data that was simulated based on a shuffled order of tasks performed by participants in the real-life experiment reported in the previous study [1]. Meanwhile, Dataset W2 contained the original UX data that was obtained based on the actual order of the tasks performed on the travel agency website. This dataset had the same structure in terms of the fixed order of tasks as that reported in my previous research [1].

Dataset W1: Shuffled order Shuffled order of tasks (Satisfaction scores were shuffled)					Dataset W2: Fixed Order Actual order of tasks (Satisfaction scores were sorted by actual order of fixed tasks)							
	Shuffled order of tasks					Actual order of tasks						
User01	-1	2	0	...	User01	2	→	-1	→	0	→	...
User02	-1	0	1	...	User02	-1	→	1	→	0	→	...
User03	-2	0	-1	...	User03	0	→	-2	→	-1	→	...

(→ = Next Task)

Figure 4-6. Example of the structure in each dataset.

In the evaluation step, this study used polynomial SVM and leave-one-out cross-validation (LOOCV) to evaluate the performance of each dataset in the prediction of final user satisfaction. Table 4-1 shows the models' performance for the two datasets.

Table 4-1. Performance of prediction models of final user satisfaction for a travel agency website.

Leave-one-out cross-validation (LOOCV)	Dataset W1 (Shuffled ordered of tasks)	Dataset W2 (Actual order of tasks)
Cross validation accuracy without oversampling	0.48	0.72
Cross validation accuracy with oversampling (SMOTEN)	0.58	0.90

The accuracy score ranges from 0.00 to 1.00; a higher value indicates higher accuracy.

4.3.1.2 Preliminary Experiment II: Google Nest Mini

In preliminary experiment II, this study used data from a prior study in which participants evaluated the Google Nest Mini smart speaker [1]. The aim of the prior study was to predict final user satisfaction based on the satisfaction score given by users while they used the product. Participants had to respond to twelve satisfaction survey questions concerning the Google Nest Mini.

Twenty-five university students aged 21–24 years were recruited as participants. The task assumed that the participants had purchased the new smart speaker and removed it from its box. After removing the smart speaker from its box, the participants were required to set it up and start using it by performing 11 steps in fixed order, as shown in Figure 4-7. The participants performed each step by referring to the enclosed instructions. At the end of each step, the participants recorded their satisfaction, which was then used to draw a customized UX curve. Their satisfaction with the product after completing each of the 11 steps was used as an indicator of momentary UX. The final user satisfaction obtained after completing all 11 steps was

used to indicate episodic UX and as the target class variable for supervised learning. In this study, the number of classes was five.

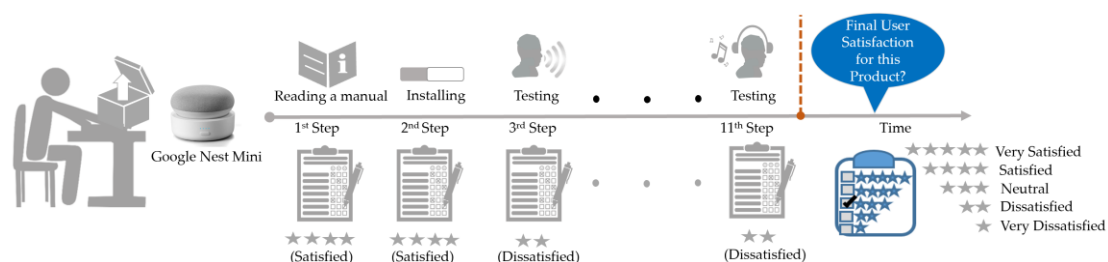


Figure 4-7. Evaluating user satisfaction while users set up the Google Nest Mini.

In preliminary experiment II, this study created two datasets (P1 and P2), as I did in preliminary experiment I, with which to build my machine learning models. Dataset P1 contained UX data that was simulated after shuffling the original order of tasks performed in the real-life experiment. Meanwhile, Dataset P2 contained the original UX data that was obtained based on the actual order of tasks performed on the Google Nest Mini in my previous research [1].

In the evaluation step, this study used SVM polynomial and LOOCV to evaluate the performance of each dataset in the prediction of final user satisfaction. Table 4-2 shows the models' prediction performance for the two datasets.

Table 4-2. Performance of prediction models of final user satisfaction for the Google Nest Mini smart speaker.

Leave-one-out cross-validation (LOOCV)	Dataset P1 (Shuffled ordered of tasks)	Dataset P2 (Actual order of tasks)
Cross validation accuracy without oversampling	0.56	0.60
Cross validation accuracy with oversampling (SMOTEN)	0.64	0.76

The accuracy score ranged from 0.00 to 1.00; a higher value indicates higher accuracy.

As noted above, this study wanted to understand the importance of task order while a user uses a product or service. Thus, after the evaluation step, this study compared the two datasets. My comparison demonstrated that the sequence or order of actions performed on a product or service, namely whether they were the actual fixed order of tasks or a shuffled order of the tasks, is important. From preliminary experiment I and II, this study found that Dataset W2 and P2, which contained the original UX data obtained based on the actual order of tasks performed by participants, produced the most accurate predictions of final user satisfaction, at 0.90 and 0.76, respectively. Integration of the actual order of actions into a dataset can thus affect a model's prediction of final user satisfaction.

The results of preliminary study I and II indicate that a change in the sequential order of tasks can significantly affect final user satisfaction. In addition, the actual order of tasks in the UX can significantly impact the prediction of final user satisfaction.

4.3.2 Main Experiment

In my preliminary studies, this study found that changing the sequential order of tasks can significantly affect the prediction of final user satisfaction. In my main experiment, this study aimed to demonstrate a new approach that uses sequential task order to predict final user satisfaction based on UX related to randomly ordered tasks. This study wanted to determine whether accounting for the sequence of actions can improve prediction of final user satisfaction.

Sixty university students aged 20-25 years were enrolled in this experiment. Before starting, this study explained that the participants' task was to find a once-in-a-lifetime trip they wanted to go on using the menu links on a travel agency's webpage. This website was created as a virtual service for my experiment only. The experiment involved three main tasks (A: finding a tour, B: finding a hotel, C: reviewing the information) as shown in Figure 4-8 and allowed participants to test the website by selecting from the available menus. Participants were not limited in the amount of time they had to complete the tasks. The participants took on average around 17 minutes to complete all tasks (A+B+C) attentively.

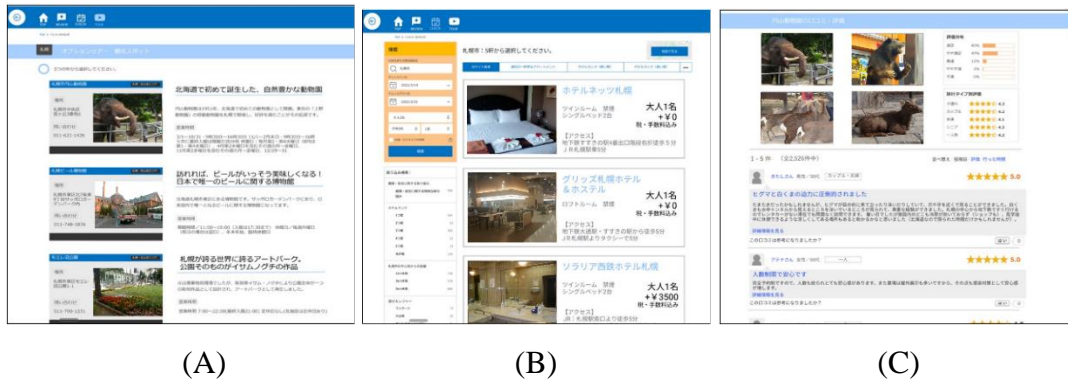


Figure 4-8. Examples of the interface of the travel agency website. (A) Task A: Tour finding; (B) task B: Hotel finding; (C) task C: Information review.

Each main task was further divided into three subtasks to obtain more specific data for building my machine learning models. The subtasks were a list of fix ordered tasks. For example, Task B consisted of subtask B1, subtask B2 and subtask B3. As each main task comprised 3 subtasks, the total number of subtasks performed by each participant was nine. The participants had to record their satisfaction after completing each subtask (they had to evaluate 9 subtasks). Once they had completed all tasks, they recorded their final user satisfaction, as shown in Table 4-3.

Table 4-3. Each main task was divided into three subtasks.

Main Task A: Finding a tour	Main Task B: Finding a hotel	Main Task C: Reviewing info
Subtask A1: view tours	Subtask B1: view hotels	Subtask C1: read trip reviews
Subtask A2: read tour details	Subtask B2: read hotel details	Subtask C2: read tour reviews
Subtask A3: compare and book a tour	Subtask B3: compare and book a hotel	Subtask C3: read hotel reviews

They were requested to record the order of their activities and their satisfaction during and after they had completed all tasks. Six groups comprising 10 participants each were assigned a different set of ordered main tasks to perform. A balanced

distribution of participants in each group is expected to reduce inequivalence across variables when building the model.

As participants completed the three main tasks in random order, a customized UX Curve was progressively constructed to record when Tasks A, B, and C were performed and the participants' level of satisfaction at each point, as shown in Figure 4-9. This procedure is often implemented in visits to travel websites in participant experiments. After completing the three main tasks, participants recorded their final satisfaction based on their experiences. It should be noted that final user satisfaction after the three main tasks was evaluated for the study only and did not affect the UX Curve for the website itself. The obtained final user satisfaction data were used as a target class variable for supervised learning. For UXE, this study used a 5-point scale that ranged from -2 to +2 (Figure 4-9, right) based on a previously reported UX graph template [11]. Thus, the number of classes used in this study was five.

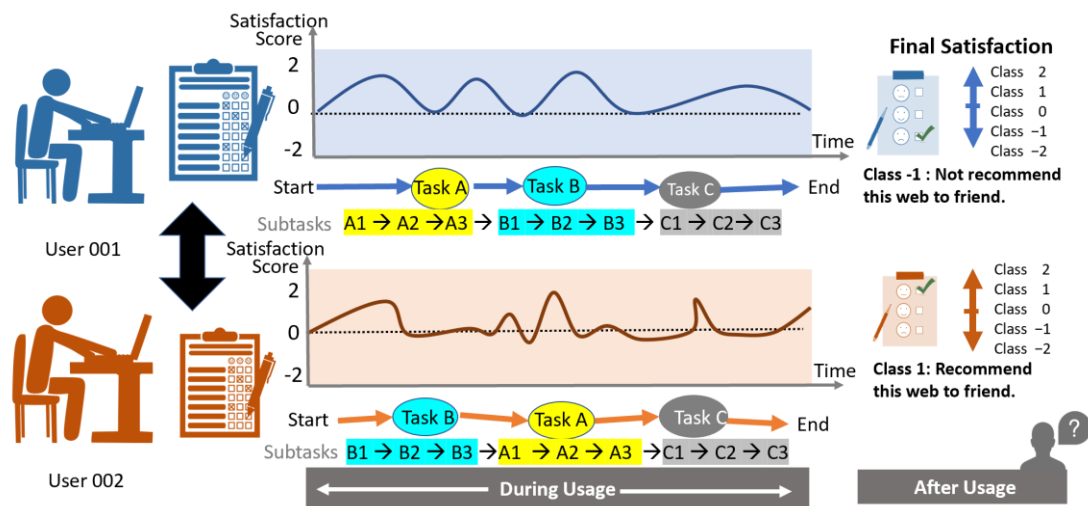


Figure 4-9. Example of a UX curve based on randomly order tasks.

4.3.3 Dataset Structure

An essential phase in my approach involved preparing the data for building machine learning models, in which the UX data were transformed into a feature matrix. As this study wanted to study the usefulness of sequential ordering for identifying factors predictive of final user satisfaction, this study created two different datasets. Dataset I contained UX data which did not account for the actual order of the

tasks performed on the travel agency website, while Dataset II contained a randomly ordered UX in the actual task order. Moreover, Dataset II accounted for the actual order of data as it comprised satisfaction score data that was sorted by the actual order of tasks performed. For example, Figure 4-10 shows that, in Dataset II, the first user (User01) gave a satisfaction score of 0 after completing task B1. They then gave a satisfaction score of 1 after completing task B2, and then a satisfaction score of 0 after completing task B3. In my study, this study compared the prediction results between Dataset I, which did not account for the actual task order, and Dataset II, which did.

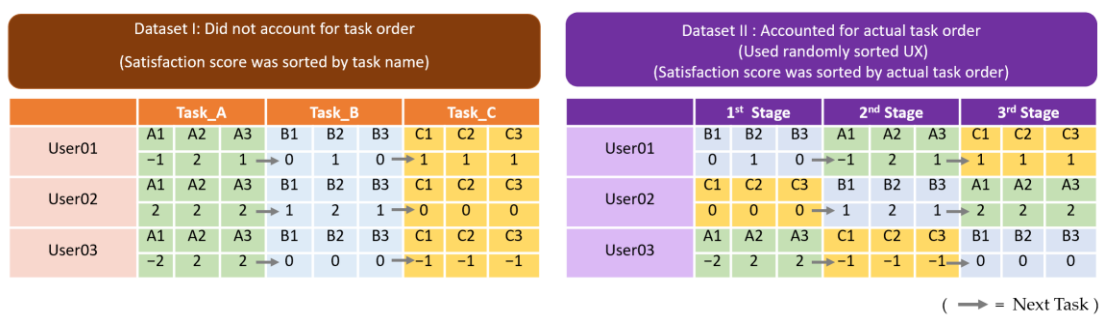
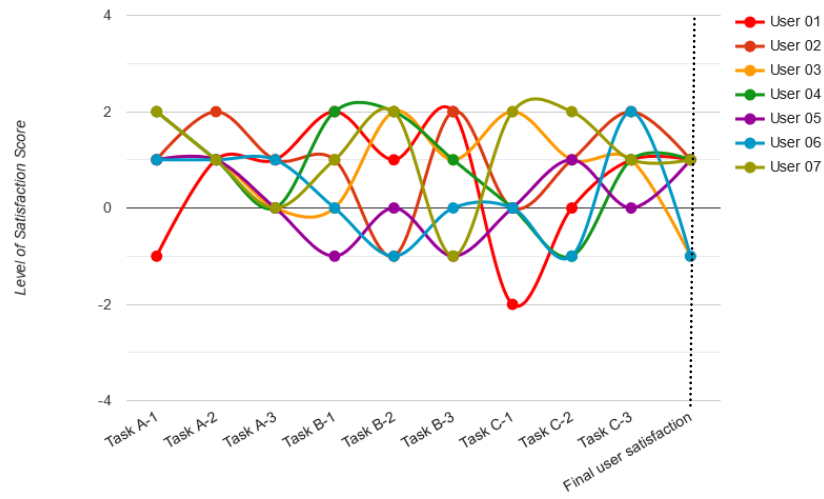


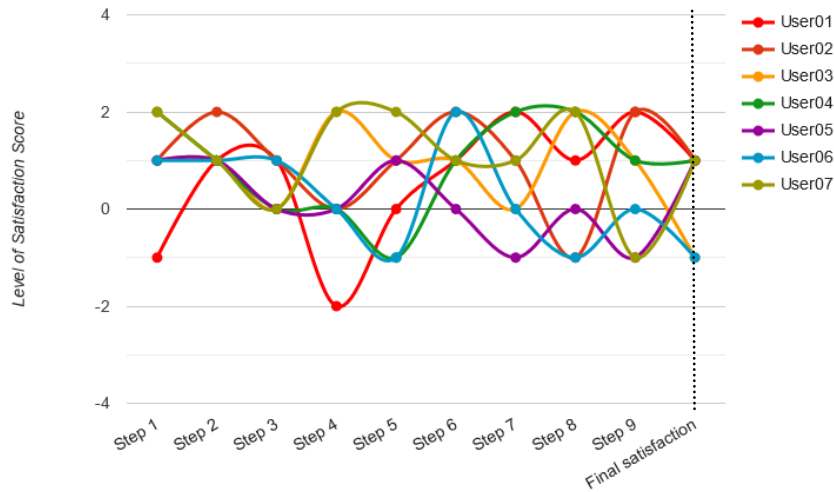
Figure 4-10. Two datasets were created based on task order.

From Figure 4-10, the input variables that this study give to machine learning models are called features. Each column in dataset constitutes a feature that is an input to a machine learning model. Regarding different of input features between two dataset, Dataset I used satisfaction scores in the fixed order as features. On the other hand, Dataset II used satisfaction scores which sorted in the actual order of user tasks as features.

Additionally, this study illustrated the UX Curve which demonstrates the correlation between satisfaction scores and overall user satisfaction, as depicted in Figure 4-11. However, it may not be straightforward to analyze these satisfaction scores from the UX Curve in a real-world product evaluation scenario.



(A)



(B)

Figure 4-11. An example of 7 UX Curves on (A) Dataset I which did not account for the actual order of the user tasks and (B) Dataset II which account for the actual order of the user tasks.

4.3.4 Building Classification Models

As mentioned above, several studies have used classification algorithms to predict final user satisfaction [1,38–41,44]. this study tested four SVM algorithms [46], namely SVM linear, SVM sigmoid, SVM RBF and SVM polynomial; and three other methods namely random forest, KNN, and AdaBoost. this study created and

trained these classification models using two datasets: a training set and test set, before comparing their performance to identify the best model for predicting final user satisfaction.

Because the original dataset, which was based on the 5-point scale, produced an accuracy score of less than 0.50, this study concluded that the results obtained using the 5-point scale were insufficiently accurate. Thus, this study converted the 5-point scale (-2, -1, 0, +1, +2) to a 3-point scale (-1, 0, +1). To do this, this study first merged the -2 and -1 point classes into a single “-1 point” class and grouped the +2 and +1 point classes into a “+1 point” class. The final 3-point scale comprised the target classes -1, 0, and +1 points. One advantage of rescaling the 5-point scale to a 3-point scale is the increased number of samples per class, which can be useful for building machine learning models.

In the main experiment, this study noted an unequal distribution of classes within Dataset I and Dataset II. Generating data for the minority class, defined as that with a smallest sample size of all the classes, is a challenging problem for training in machine learning. One way to solve this problem is to oversample samples in the minority class. To do this, this study used the SMOTEN [57] oversampling technique in the evaluation step. As indicated in Section 2.5, multiple techniques exist for dealing with imbalanced sample distributions. Oversampling the minority class is one such approach used in data science [57].

4.3.5 Model Evaluation

In the evaluation step, this study measured each classification model’s efficiency in terms of accuracy, precision, and recall. The conventional leave-one-out cross-validation method was used to evaluate performance [73], in which one set of data is left out of the training set. For example, of the original data from 10 users, those from 9 were used to train the model, and one was used for validation. The leave-one-out cross-validation (LOOCV) procedure is appropriate for small datasets because its results are reliable and unbiased in estimating model performance [81].

4.4 Results and Discussion

We conducted experiments to test my hypothesis that machine learning models that account for the sequential order of tasks performed when using a product or service produce more accurate predictions of final user satisfaction than those that do not. By comparing the predictive performance of Datasets I and II using seven machine learning algorithms, this study showed that, indeed, sequential ordering was important for prediction accuracy. Among the models tested, this study found that the machine learning classification model produced the greatest accuracy, as shown in Table 4-4.

Table 4-4. Performance of test models with SMOTEN oversampling.

Scores		Dataset	Random Forest	KNN	SVM Poly	SVM Linear	SVM RBF	SVM Sigmoid	AdaBoost
Cross-Validation Accuracy	LOOCV	Dataset I	0.68	0.61	0.68	0.60	0.75	0.46	0.61
		Dataset II	0.70	0.71	0.76	0.76	0.76	0.61	0.70
Accuracy	Split for training/test (80/20)	Dataset I	0.83	0.90	0.93	0.83	0.93	0.57	0.87
		Dataset II	0.83	0.83	0.97	0.83	0.93	0.70	0.73
Precision	training/test (80/20)	Dataset I	0.85	0.92	0.93	0.87	0.94	0.54	0.89
		Dataset II	0.88	0.85	0.97	0.88	0.94	0.84	0.80
Recall	training/test (80/20)	Dataset I	0.83	0.90	0.93	0.83	0.93	0.57	0.87
		Dataset II	0.85	0.83	0.97	0.83	0.93	0.70	0.73

SVM = support vector machine; Poly = polynomial kernel; KNN = K-nearest neighbors, RBF = radial basis function

Dataset I = Did not account for task order

Dataset II = Accounted the actual task order

Accuracy score values were between 0.00 to 1.00, with a higher value indicating higher accuracy.

4.4.1 Accounting for Actual Task Order in Randomly Ordered UX

In LOOCV, the highest leave-one-out cross-validation accuracy was obtained for Dataset II at 76%, which was significantly higher than that for Dataset I at 68%. This

difference is thought to be partly due to the difference in the structure of the datasets in terms of the task order considered in the data, as shown in Figure 4-10. Moreover, this study compared the prediction results obtained using the split validation technique between the two datasets and found that Dataset II demonstrated significantly better performance, producing the highest accuracy of 97% compared to 93% for Dataset I. Thus, my initial findings showed that evaluation based on Dataset II, which accounted for the actual task order, may be better for predicting satisfaction levels when estimating final user satisfaction. Therefore, this study confirmed that task order in UXE may directly affect final user satisfaction.

4.4.2 Machine Learning Algorithms

The best machine learning algorithm with which to predict final user satisfaction for the travel agency website was SVM, which had a cross-validation accuracy of 76% as shown in Table 4-4. this study propose that polynomial kernel SVM may be most accurate for the prediction of satisfaction level based on randomly ordered tasks because the higher degree polynomial kernel in the SVM algorithm allows for a more flexible decision boundary [77]. In this case, the polynomial kernel had three parameters (offset, scaling, degree), which are relatively easy to fine tune to obtain classification results with the highest accuracy.

As mentioned above, this study used classification algorithms to predict final user satisfaction. this study evaluated my machine learning model using two types of cross-validation techniques, namely LOOCV and train-test splitting for evaluating machine learning algorithms. However, this study found that the train-test splitting procedure is not appropriate when the dataset available is small [82]. The reason is that when the dataset is split into train and test sets, there will not be enough data in the training dataset for the model to learn an effective mapping of inputs to outputs. There will also not be enough data in the test set to effectively evaluate the model performance. The estimated performance could be overly optimistic (good) or overly pessimistic (bad) [82].

According to my original dataset, it is small dataset and insufficient data. When the dataset used for model building and evaluation is small, the LOOCV approach is recommended for model evaluation [81,83]. A suitable alternate model evaluation

procedure could be the LOOCV procedure. this study found that LOOCV is appropriate to use for small dataset in evaluating machine learning algorithms [84]. The greatest advantage of LOOCV procedure is that it doesn't waste much data. LOOCV used only one sample from the whole dataset as a test set, whereas the rest is the training set. Moreover, the advantage of LOOCV over Random Selection is zero randomness [85]. Besides, the bias will also be lower as the model is trained on the entire dataset, which consequently will not overestimate the test error rate. Thus, in my experiment, this study focused results from cross-validation accuracy based on LOOCV technique in order to evaluate model performance comprehensively.

4.5 Summary

This chapter presents a new approach using machine learning techniques to predict final user satisfaction based on UX related to randomly ordered tasks. The main experiment confirmed the effectiveness of accounting for task order when predicting user satisfaction as an indicator of UX. Use of Dataset II, which accounted for the actual order of the tasks performed by users, in the measurement of satisfaction provided the highest cross-validation accuracy of 76% when compared to Dataset I, which did not account for task order. Further, this study showed that polynomial kernel SVM produced the most accurate predictions among the machine learning methods tested.

The main finding of my study was that accounting for the actual order or sequence of actions can improve predictions of final user satisfaction. Given that the UX in real life involves randomly ordered tasks, my proposed method reflects the real-world setting. this study expect that my findings will help other researchers conducting UXE obtain more accurate predictions of user satisfaction. My study contributes to knowledge in the field in various ways. First, my contribution relates to the outcomes of UX. Data on randomly ordered tasks or random ordering is often difficult to understand and analyze. this study identified a relationship between the order or sequence of actions and episodic or cumulative UX, which is related to final user satisfaction. My study shows that understanding the order or sequence of actions as the UX curve changes [34,42,43] can help determine final user satisfaction. Second, machine learning algorithms such as SVM can be used to accurately predict

final user satisfaction and contribute to developing better products through analysis of randomly ordered UX. Hence, it is important to carefully monitor randomly ordered UX. Third, accounting for the actual order or sequence of actions performed by users of a product or service could constitute a new approach for improving the predictive accuracy of final user satisfaction.

In summary, the study highlights the importance of considering the sequence of actions and order in determining user satisfaction in UX. Furthermore, the concept of order is also relevant in recommendation systems, where the sequence and order of recommendations can affect user satisfaction. By considering the order and sequence of recommendations, recommendation systems can aim to provide a more enjoyable experience and increase the likelihood of serendipitous discoveries.

CHAPTER 5

UX EVALUATION IN PRACTICAL USE

This chapter addresses the application of UX evaluation in practical use which followed my proposed framework. Regarding to during actual usage by user, soft biometric data such as gender, age and facial expression can be used as the essential data for the user satisfaction analysis. In this research, this study assume that the facial expression is essential in physical expressions and can be used as the accurate satisfaction data. In brief, this chapter is organized on the following sections: 5.1 Introduction 5.2 Proposed Method in Practical Use 5.3 Experiments 5.4 Results and Discussion, and 5.5 Summary.

5.1 Introduction

In the current design process for products and services, UX (user experience) has been widely used and has been considered to have the strong relationship between user satisfaction. The user satisfaction provides essential feedback and reflective data from the customers' aspects such as their opinions, favors, preferences and past experiences. However, due to the difference in the user experience for each user, both human and computers has had difficulty in classifying this valuable data for developing or improving the products and services.

Several papers and articles regarding the measurement of UX as the satisfaction have been published. However, in the most approaches, UX was measured by questionnaire or survey collection method, which may lead to bias and a lack of exact feeling data of the target users owing to exaggeration, embarrassment and forgetting.

On the other hand, soft biometric data such as gender, age and facial expression can be used as the essential data for the user satisfaction analysis from the viewpoint of objective evaluation.

One of significant UX evaluation is the analysis of facial expression data. Facial expressions are relatively easy to collect and analyze in real-time, which can provide valuable information for real-time monitoring of customer satisfaction. It is real-time

method and more convenient compared to other method like questionnaire and survey method. Facial expression data is promoted as UX data that makes UX evaluation more convenient in data collection step. Instead of having to bother entering some answers into questionnaire form. Facial expressions of emotion are probably the most important signal of the face because they provide about people's personalities, emotions, motivations, or intent. They are extremely important to UX research about user satisfaction during product usage. Entire body reflects emotions. Body is governed by biological algorithms that determine how body reacts, especially the facial expressions. Thus, it can be used the facial expression data as the accurate satisfaction data. In this research, this study assume that the facial expression is essential in physical expressions and can be used as the accurate satisfaction data. It may be possible to capture the user's facial expression during the particular use of products or services without users' consciousness.

This study aimed to propose a framework to classify the user satisfaction of products or services by the facial expression recognition and machine learning. The classification of the user satisfaction is one of the first steps of customer segmentation. The customers are partitioned into groups that represent the relevant users. Thereby, the designers can appropriately develop or improve the products and services.

5.1.1 UX and User Satisfaction

The user satisfaction has been widely adopted as a subjective evaluation of UX [86]. User satisfaction information can assist designers in developing or improving their products and services in the design processes [87]. In the previous researches on UX or user satisfaction, there are several approaches using machine learning algorithms.

Asil Oztekin et al. [88] proposed a new machine learning based evaluation method to evaluate the usability of eLearning systems. In their research, three machine learning methods (support vector machines, neural networks, and decision trees) and multiple linear regression were used to develop prediction or classification models in order to discover the underlying relationship between the overall eLearning system usability and its predictor factors.

A predicting user satisfaction method with intelligent assistants was proposed by Julia Kiseleva et al. [86]. The intelligent assistants allow the user to use the interaction signals including voice commands and physical gestures. In their research, they analyzed the user satisfaction with three machine learning methods, including Gradient Boosted Decision Trees (GBDT), logistic regression, and support vector machines (SVM). The 10-fold cross-validation was conducted for the model evaluation. For each experiment, they reported the overall accuracy (ACC), precision (P), recall (R), F1 score (F1) and Area Under the Curve (AUC).

Kazi Md Munim [89] proposed a web-based tool for UX evaluation using the measurement of user facial expression. The tool can detect the gender and emotions of users such as engagement, valence, contempt, surprise, anger, sadness, disgust, fear and joy. For the evaluation of the tool, the results by the tool of the facial expression recognition were compared with the results by human judgement. They proposed the web-based tool for the facial expression recognition, but the UX evaluation part is not sufficiently verified.

5.1.2 Facial Expression

The facial expressions are the most visible and expressive in all the channels for human communication. Researches in psychology demonstrated that the facial expressions showed reliable correlation with self-reported emotions and with physiological measurement of human emotion [90–92]. The facial expressions are visual displays of different small muscle movements in the face and are also used to infer a person's discrete emotional state (e.g., happiness, anger) [93].

In recent years, both image processing and machine learning algorithms have been widely used for the effective automated facial expression recognition systems. It can be noticed that deep learning approaches were used and widely applied classifiers in facial expression recognition systems [94,95]. There are several open-source facial expression recognitions using machine learning algorithms as follows:

CNN Emotion Classification: Octavio Arriaga [96] proposed the real-time convolutional neural networks for the emotion and gender classification, and it was implemented by Pether Cunha [97]. The result of accuracy score was reported as 66%.

Emotion-recognition project: Omar Ayman [98] proposed the emotion-recognition project. The result of accuracy score was also reported as 66%.

Deep Convolutional Neural Networks: Emotion recognition using deep convolutional neural networks was proposed by Correa Enrique [99], and was implemented by Balaji Atul [100]. The result of accuracy score was reported as 63.2%.

Facial-Expression-Keras: The purpose of this project is to recognize facial expression from video streaming by using deep learning [101]. The result of accuracy score was reported as 84%. The expressions were divided into seven classes: angry, disgusted, neutral, sad, happy, surprised, and fear.

5.1.3 Machine Learning

The emotional state is possible to measure in real time by the facial expression recognition systems. Although the emotional state provides sufficient information, it is quite difficult for human to interpret or classify the user satisfaction using the time change of emotional state. Hence, the classification algorithm using the machine learning is an essential component of user satisfaction analysis.

In the previous researches, the supervised learning algorithms such as Support Vector Machine (SVM) [46], Logistics Regression (LS) [50], K nearest neighbor (KNN) [102] and Multi-Layer Perceptron (MLP) [103] were mainly used for the classification.

5.2 Proposed Method in Practical Use

The procedure of the proposed framework is shown in Figure 5-1. The proposed framework consists of the three main steps. First, the data of facial expression, gender and age are collected during the use of products or services, and the target products or services are also experimentally evaluated as the user satisfaction after the use of them. Second, classification models are built by machine learning algorithms using the data of facial expression, gender, age and user satisfaction. Finally, the model evaluation was employed to verify the accuracy of the model. After making the classification model, it is possible to classify the user satisfaction only from the data of facial expression, gender and age.

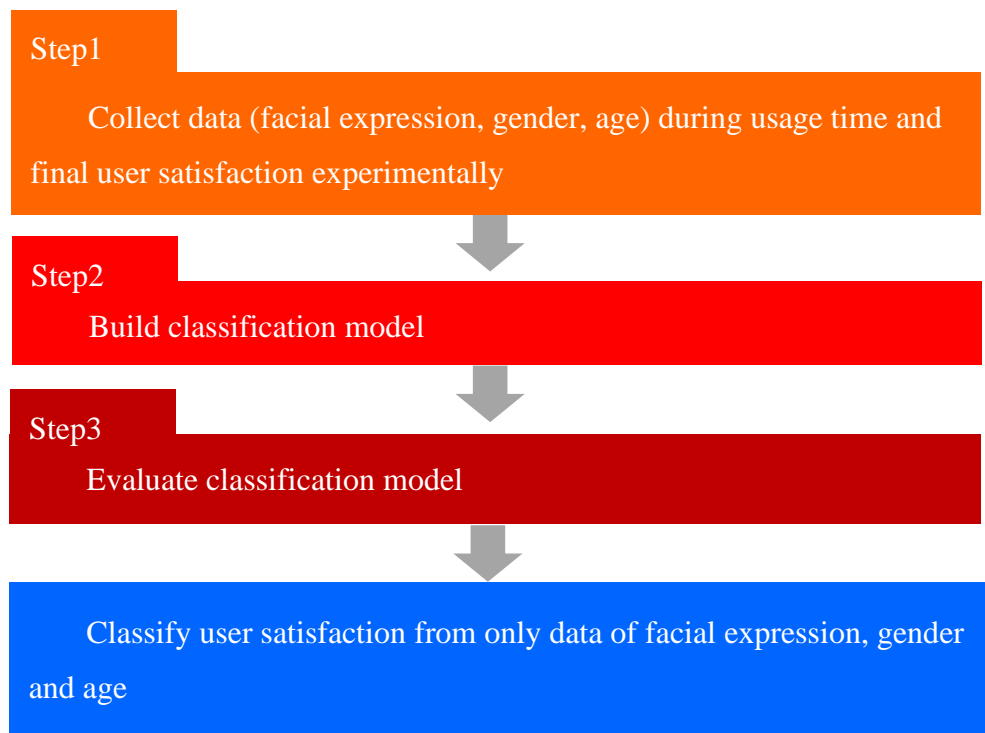


Figure 5-1. Procedure of proposed framework.

5.2.1 Collection data

Firstly, the facial images of each user are continuously captured by a video camera or a webcam while they were using products or services. As the extra information, the gender and age of the user is obtained. After the use of the products or services, the user evaluates them as the user satisfaction with 5-point scale (1 to 5 star) experimentally. Table 5-1 showed the user ID, gender, age, the time change of user facial expression during usage time of products or services, and the user satisfaction.

Before building the classification model, this study need to transform the emotional state to the numerical value that is suitable for the scheme of machine learning. The facial images were processed by the facial expression recognition system, and the numeric array of emotional states was generated. The user's gender, age and the user satisfaction are also transformed into the numerical value. Then, the dataset that is suitable for the scheme of machine learning is prepared in the table format, as shown in Table 5-2. In Table 5-2, three kinds of the facial expression is

shown as an example. The kind of the facial expression is depending on the output of the facial expression recognition system.

The primary key affecting the facial expression recognition accuracy is how to select the appropriate classifier that can successfully classify the facial expressions. Currently, this study adopted the Facial-Expression-Keras [101] for the facial expression recognition.

Table 5-1. Data Preparation













User ID, gender, age	Time changes of facial expression during usage time					Satisfaction
001, M, 25					...	★★
002 F, 23					...	★★★
003 M, 24					...	★

Table 5-2. Dataset for Machine Learning

User ID	M/F	Age	Scoring of facial expression during usage time [%]							Class label
001	0	25	Happy	22	23	19	18	26	23	3
			Sad	1	5	4	4	3	3	
			Neutral	5	2	1	4	5	6	
002	1	23	Happy	40	38	35	37	34	33	2
			Sad	4	4	3	3	40	2	
			Neutral	1	4	5	6	1	2	
003	0	24	Happy	19	12	10	9	11	13	1
			Sad	30	29	38	41	42	39	
			Neutral	5	6	1	2	3	2	

5.2.2 Classification Model

In the next step, the classification model is built using the prepared dataset. this study will try to use Support Vector Machine (SVM) [46], Logistics Regression (LS) [50], K nearest neighbor (KNN) [102] and Multi-Layer Perceptron (MLP) [103] as the candidate classification algorithm in the experimentation. this study will compare them and chose a suitable algorithm.

5.2.3 Model Evaluation

Finally, the model evaluation was employed to verify the accuracy of the built classification model. Researchers also want a more accurate estimate of the accuracy of the best model on unseen data by evaluating it on actual unseen data. In this step, the model evaluation is reported as the result of each classification algorithm, and a confusion matrix [104] is used to describe the performance of a classification model using a set of test data. The performances of each classification model are reported by accuracy score, precision score, and recall score, respectively.

5.2.4 Practical Use

After making the classification model, it is possible to classify the user satisfaction only from the data of facial expression, gender and age without the user's consciousness. Currently, the data of gender and age are added in the dataset, however, this study will verify that the only data of facial expression can be used for the classification of the user satisfaction.

As an example of the practical use of my proposed framework, it can be applied to the evaluation of movies at theater (for multiple people at the same time) or online (for one user), usability evaluation of products, evaluation for e-learning and service evaluation in museum, aquarium and exhibition hall.

5.3 Experiments

5.3.1 Facial Expression Recognition

The facial expression recognition [101] was demonstrated as a preliminary experiment using Python to develop on Google Colab environment, and was tested by the five test participants. Each facial expression was categorized into seven categories

(1=Angry, 2=Disgust, 3=Fear, 4=Happy, 5=Sad, 6=Surprise, 7=Neutral). The score of each facial expression is the prediction percentage from 0 to 100 representing the emotional level of the input facial image.

In order to demonstrate the facial expression recognition, the input images of various facial expression by the five test participants have been used. The facial expression recognition automatically processed and reported the results as the emotional state of the users. Some examples of the facial expression images and the predicted facial scores of the five test participants are shown in Figure 5-2. Three facial expression (happy, surprise, and disgust) is shown as an example.
















Volunteer	1 st	2 nd	3 rd	4 th	5 th
Happy					
Score	18%	20%	22%	16%	13%
Surprise					
Score	21%	23%	28%	26%	15%
Disgust					
Score	24%	15%	25%	20%	14%

Figure 5-2. Sample facial expression recognition from five volunteers.

5.3.2 Procedure Confirmation for Getting Dataset

The second preliminary experiment was conducted by one test participant to confirm the procedure for getting the dataset for the machine learning.

As shown in Figure 5-3, the test participant was asked to watch a comedy movie (104 seconds), at which time facial images of the test participant were taken by a webcam and the facial expression recognition [101] was performed. It is assumed that the movie itself is evaluated, or the movie is evaluated as the user experience.

The facial expression recognition continuously calculated the seven facial expressions based on the input facial images, as shown in Figure 5-4. Although the frame rate of original facial expression data is 30 fps, the downsampling technique with 30 fps averaging filter was conducted for the noise cancelling. The score range of facial expression on the vertical axis is 0 to 100 percent and the range of time on the horizontal axis is 0 to 104 seconds of the comedy movie.

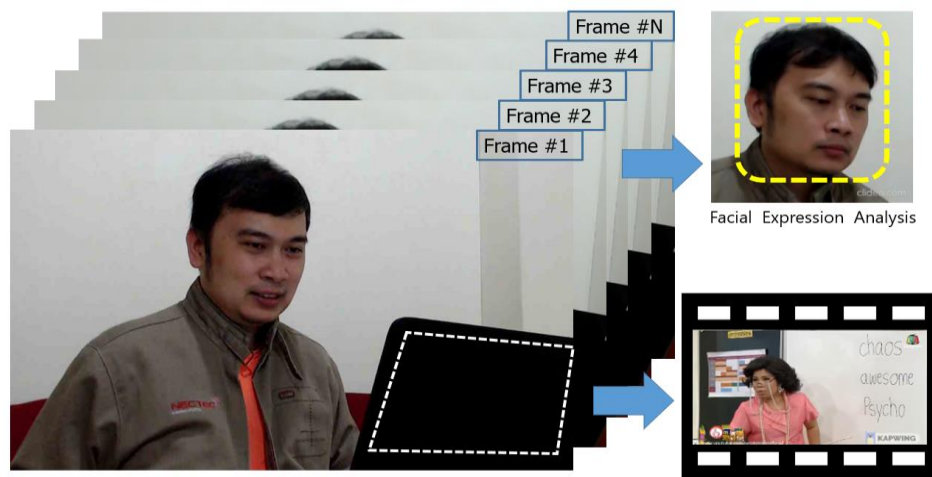


Figure 5-3. Facial expression recognition during watching comedy movie.

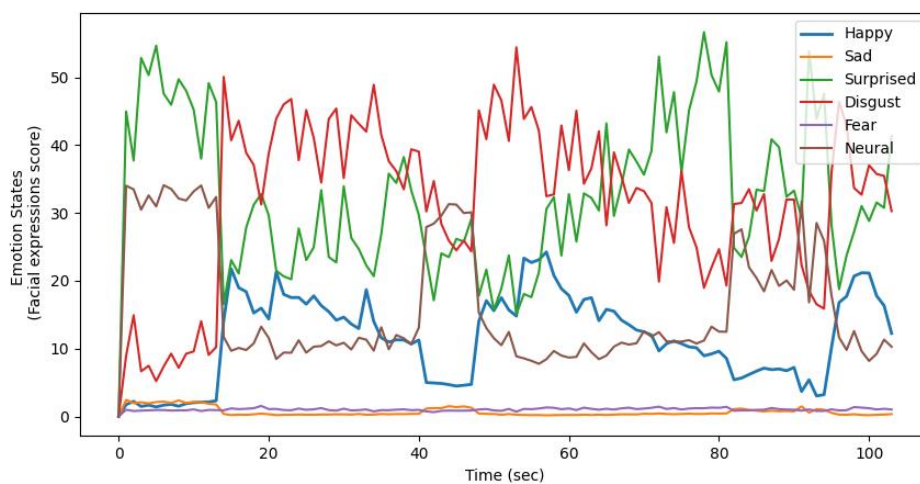


Figure 5-4. Time changes of facial expression score.

Each line shows the facial expression changes of happiness, sadness, surprise, disgust, fear, and neutral expression respectively during the watching 104 second movie. For the happiness line, it shows that the line increased dramatically during the

watching comedy movie. In contrast, both sadness and fearful lines were almost unchanged. In conclusion, it was noticeable that the happiness line changed over time while the test participant was watching the comedy movie.

After watching the comedy movie, as shown in Figure 5-5, the test participant experimentally answered the questionnaire about his user satisfaction score with a 5-point scale (1 to 5 star).

Satisfaction Survey

Name: Katt Test Date: July 20, 2020

Movie name: Learning English with Thai Teacher

How like are you to give star and recommend this movie to a friend or colleague?

★★★★★

★★★★

★★★

★★

★

Figure 5-5. User satisfaction after watching movie.

Finally, this study could confirm the procedure that the facial expression scores, gender and age (attributes) and the user satisfaction score (target class) were able to be prepared to build a classification model using machine learning for classifying the user satisfaction.

Additionally, this study illustrated the examples of facial expression curve which demonstrates the changes between facial expression and time, as depicted in Figure 5-6. However, it may not be straightforward to analyze these facial expression curve in a real-world product evaluation scenario.

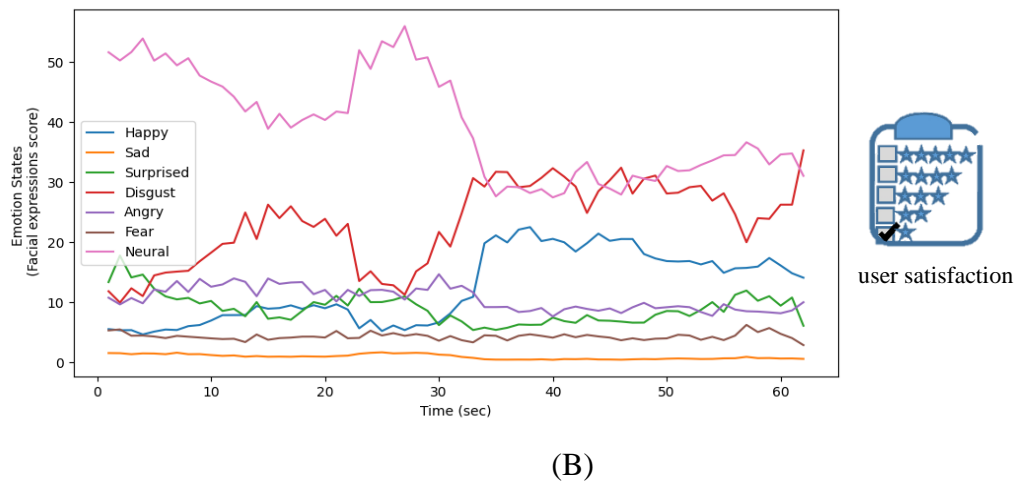
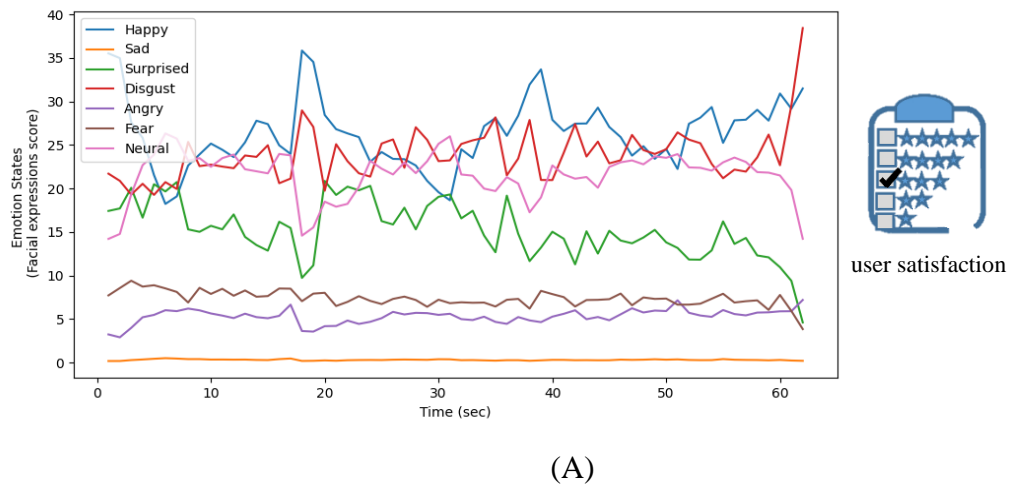


Figure 5-6. Examples of facial expression curve when (A) user satisfaction is 3 of 5-point scale and (B) user satisfaction is 1 of 5-point scale.

5.3.3 Confirmation for Effectiveness of Classification Framework

The third preliminary experiment was conducted by a test participant to confirm the effectiveness of building the classification model. The methodology of this experiment is shown in Figure 5-7. During a comedy movie (seventy-two minutes) watching, the facial expressions of the participant were recorded by a front-facing camera simultaneously. Furthermore, the test participant experimentally needed to answer the questionnaire about his satisfaction on a 5-point scale (1 to 5 star) at every one minute during the movie watching.

The 72-minute recorded facial expressions video was divided into 1-minute video clips. The seventy-two video clips were automatically processed by the facial

expression recognition system, and the results were reported as seventy-two facial expression data. The seventy-two facial expression data and satisfaction scores were used to build machine learning models.

This study wants to predict satisfaction score based on UX curve by facial Expression data. However, this study is preliminary experiment for testing the proposed framework. This study wants to test a proposed framework with facial expression data in predicting the satisfaction score every one min. However, further research should increase a period of time in predicting user satisfaction.

In this experiment, this study attempted to use also the SVM-SMOTE oversampling technique [60] to reduce the class-imbalance problem.



Figure 5-7. Overall the third preliminary experiment

For the model evaluation, seven machine learning methods were compared, namely K Nearest Neighbor (KNN) [49], Support Vector Machine (SVM) [46], SVM with sigmoid kernel, SVM with linear kernel, SVM with polynomial kernel, SVM with radial bias, Logistics Regression [50], and Neural Net [52]. Machine learning models were built to classify the satisfaction score from the only facial expression data.

5.4 Results and Discussion

The use of physiological data (facial expression data) was conducting as preliminary experiment. The use of facial expression data is challenges in practical use of this research, some problems were found in this study. During data acquisition process, a major problem with video data is noise of image. Image noise is undesired fluctuations of color or luminance that obscure detail during video recording. Noise data may affect facial expression curve as input of machine learning. Noise data characteristics could affect the predictive accuracy of predictive models. In future work, the further study needs to control the video environments.

After finishing the classification model building, each model was validated by leave-one-out cross-validation. The results show that the combination of SVM-SMOTE oversampling and SVM with polynomial kernel provided the best accuracy. The best cross-validation accuracy was 86%, as shown in Figure 5-8.

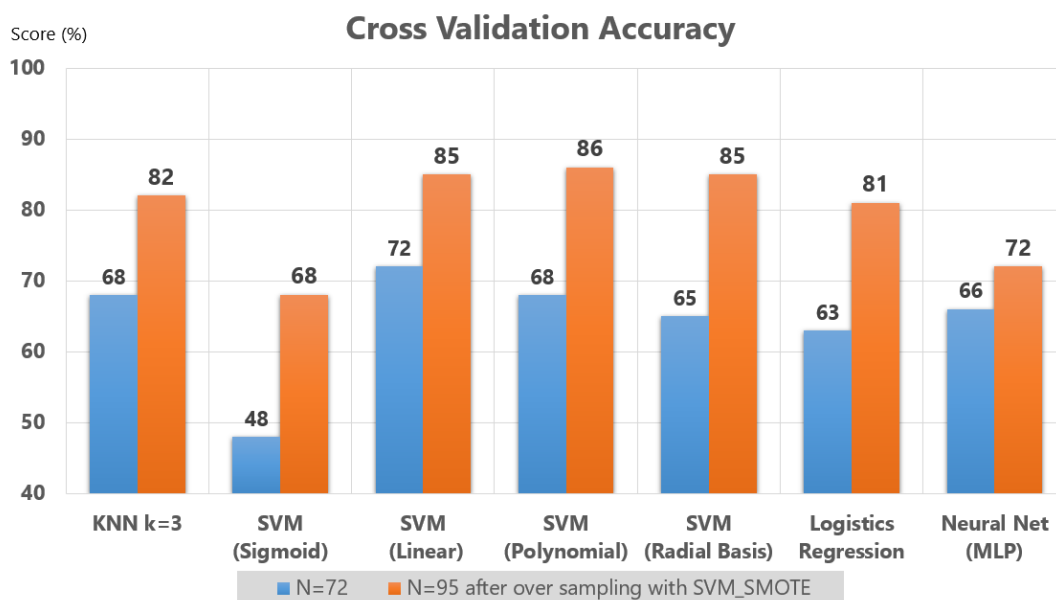


Figure 5-8. Cross-validation accuracy of each method.

From the results of the third preliminary experiment, it was confirmed that the proposed framework can be created successfully and possible to use this framework to

classify the user satisfaction of products or services by the facial expression recognition and machine learning.

5.5 Summary

This chapter proposed the application of UX evaluation in practical use which followed my proposed framework using facial expression recognition and machine learning. The proposed framework consists of three main steps. First, the data of facial expression, gender, age, and the user satisfaction are experimentally collected. Second, classification models are built by machine learning algorithms using the data. Finally, the model evaluation is employed to verify the accuracy of the model. After making the classification model, it could classify the user satisfaction from the data of facial expression, gender, and age without the user's consciousness.

In this chapter, the preliminary experiments were conducted. First, the facial expression recognition was demonstrated and confirmed the results of the facial expression recognition. The second preliminary experiment was also conducted to confirm the procedure for getting the dataset for the machine learning. The third preliminary experiment was conducted to confirm the framework's effectiveness, and the experimental results suggested that my proposed method can classify the user satisfaction practically.

In summary, facial expression data can provide a valuable and accurate measure of user satisfaction and can help to provide a better understanding of customer emotions and experiences with products and services. It can be used to identify patterns and trends in user satisfaction. Moreover, facial expression data can also be used in combination with other forms of user data, such as usability testing and surveys, to provide a more comprehensive understanding of customer satisfaction. By combining multiple forms of data, researchers can gain a more holistic view of the user experience and can identify areas of improvement that may not have been apparent with a single data source. Other forms of data that can be used in combination with facial expression data include gesture data, brain wave data, audio data, and eye-tracking data.

The classification model can be applied to various evaluation of products and services. It is considered that this framework is also useful for long term measurement of UX. It is allowed for us to observe the changes of users feeling toward their products or services.

CHAPTER 6

CONCLUSIONS

This chapter consists of two sections. The first section presents the summary of the study. In the second section, the future work to serve demands of UX evaluation are proposed.

6.1 Summary of the Study

This dissertation addressed the problems of UX evaluation and evaluated the final satisfaction for UX evaluation. UX is subjective, relating to an individual's feelings and satisfaction. Expert evaluations of UX may lead to bias, and such opinions are not easily quantifiable. Humans are prone to many types of bias. Despite algorithms having their own challenges, machine learning algorithms may conceivably be capable of producing more fair, efficient, and bias-free outcomes than humans. A proposed framework has been proposed to predict final user satisfaction by user experience data using machine learning techniques.

The contents described in each chapter and the advantages of the proposed method are summarized as follows.

In chapter 1 and 2, the introduction, background and the issues of UX evaluation for predicting final user satisfaction are described and discussed.

Chapter 3 described a framework to predict final user satisfaction by using momentary UX data and machine learning techniques. The participants were 50 and 25 university students who were asked to evaluate a service (Experiment I) or a product (Experiment II), respectively, during usage by answering a satisfaction survey. Responses were used to draw a customized UX curve. Participants were also asked to complete a final satisfaction questionnaire about the product or service. Momentary UX data and participant satisfaction scores were used to build machine learning models, and the experimental results were compared with those obtained using seven built machine learning models. This study shows that participants' momentary UX can be understood using a support vector machine (SVM) with a

polynomial kernel and that momentary UX can be used to make more accurate predictions about final user satisfaction regarding product and service usage.

Chapter 4 focused on the using random ordering of user tasks in user experience testing to predict final user satisfaction. In user experience evaluation (UXE), it is generally accepted that the order in which users perform tasks when using a product is often random rather than fixed. UXE based on these so-called randomly ordered tasks is challenging. Although several articles have been published on UXE, none have proposed a technique to evaluate the significance of randomly ordered tasks. In this study, this study propose a new approach to predict final user satisfaction based on UX related to randomly ordered tasks. this study aimed to study the importance of task order in the UX. In the main experiment, 60 participants completed questionnaires about satisfaction while performing a series of tasks on a travel agency website. Among the machine learning models tested, this study found that accounting for the order or sequence of actions actually performed by users in a support vector machine (SVM) algorithm with a polynomial kernel produced the most accurate LOOCV of final user satisfaction (76%). These findings indicate that some machine learning techniques can comprehend participants' randomly ordered UX data. Moreover, using random ordering, which accounts for the actual order of actions performed by users, can significantly impact the prediction of final user satisfaction.

Chapter 5 addressed the application of UX evaluation in practical use which followed my proposed framework. In the most approaches, UX was measured by questionnaire or survey collection method, which may lead to bias and a lack of exact feeling data of the target users. On the other hand, soft biometric data such as gender, age and facial expression can be used as the essential data for the user satisfaction analysis. In this research, this study assume that the facial expression is essential in physical expressions and can be used as the accurate satisfaction data. It may be possible to capture the user's facial expression during the particular use of products or services without users' consciousness. This study aimed to propose a framework to classify the user satisfaction of products or services by the facial expression recognition and machine learning. The proposed framework consists of the three main steps. First, the data of facial expression, gender, age and the user satisfaction are

experimentally collected. Second, classification models are built by machine learning algorithms using the data. Finally, the model evaluation is employed to verify the accuracy of the model. After making the classification model, it can classify the user satisfaction from the data of facial expression, gender and age.

6.2 Future Work

Our study confirmed the relationships between UX data and final user satisfaction from evaluations involving products and services. In experiments, the participants were university students. However, the only student group from participants is one limitation of this study. Further validation of the methods requires studies with other groups. Future research should confirm these initial findings by using random ordering in the UX to assess underlying factors for predicting final user satisfaction with other products.

Regarding subjective data in Chapter 3 and 4, the user experience research was limited by only quantitative data for predicting final satisfaction. However, quantitative data is not descriptive, it might sometimes be difficult to make decisions based solely on the collected information. Using quantitative data in an investigation is one of the preliminary strategies to guarantee reliable results that allow better decisions. Because bias in results is dependent on the question types included to collect quantitative data. The researcher's knowledge of questions and the objective of research are exceedingly important while collecting quantitative data. Future research is needed to use a mixed method based on the strengths of both quantitative and qualitative methods to produce a more inclusive and expansive understanding of predictive results.

The small number of participants and small number of tasks in the main experiment are also the limitations of this study. Further validation of the methods is needed using a suitable number of samples. In general, it is recommended to have a minimum of 500 to 1,000 data points overall and at least 50 to 100 data points per feature in the machine learning model [105]. Regarding future work, it may also be possible to introduce other measures and features such as eye movement data and operation time data, as well as using a feature selection approach to enhance the predictive performance reported in this study. Furthermore, regarding using predicting

final satisfaction from facial expression data, it may use a classification model of facial expression recognition which has already been pretrained. It could improve accuracy in predicting final satisfaction. Additionally, in the prediction of final satisfaction from facial expression data, a pretrained classification model of facial expression recognition could be utilized to improve accuracy in predicting final satisfaction from facial expression data. It is preferable to use a pretrained classification model of facial expression recognition.

REFERENCES

1. Koonsanit, K.; Nishiuchi, N. Predicting Final User Satisfaction Using Momentary UX Data and Machine Learning Techniques. *J. Theor. Appl. Electron. Commer. Res.* **2021**, *16*, 3136–3156. 10.3390/jtaer16070171.
2. Koonsanit, K.; Hiruma, D.; Yem, V.; Nishiuchi, N. Using Random Ordering in User Experience Testing to Predict Final User Satisfaction. *Informatics* **2022**, *9*, 85. 10.3390/informatics9040085.
3. Law, E.L.-C.; van Schaik, P.; Roto, V. Attitudes towards user experience (UX) measurement. *Int. J. Hum.-Comput. Stud.* **2014**, *72*, 526–541. 10.1016/j.ijhcs.2013.09.006.
4. Balasubramoniam, V.; Tungatkar, N. Study of user experience (UX) and UX evaluation methods. *Int. J. Adv. Res. Comput. Eng. Technol. IJAR CET* **2013**, *2*, 1214–1219.
5. Kurosu, M.; Hashizume, A.; Ueno, Y.; Tomida, T.; Suzuki, H. UX Graph and ERM as Tools for Measuring Kansei Experience. In Proceedings of the 18th International Conference on Human-Computer Interaction. Theory, Design, Development and Practice, Toronto, ON, Canada, 17-22 July 2016; Springer-Verlag: Berlin, Heidelberg; Vol. 9731, pp. 331–339. 10.1007/978-3-319-39510-4_31.
6. Barari, M.; Ross, M.; Surachartkumtonkun, J. Negative and positive customer shopping experience in an online context. *J. Retail. Consum. Serv.* **2020**, *53*, 101985.
7. Vall-Llosera, L.; Linares-Mustarós, S.; Bikfalvi, A.; Coenders, G. A Comparative Assessment of Graphic and 0–10 Rating Scales Used to Measure Entrepreneurial Competences. *Axioms* **2020**, *9*, 21. 10.3390/axioms9010021.
8. Carneiro, J.; Santos, R.; Marreiros, G.; Novais, P. Understanding decision quality through satisfaction. In Proceedings of the International Conference on Practical Applications of Agents and Multi-Agent Systems, Salamanca, Spain, 4–6 June 2014; Springer, 2014; pp. 368–377.
9. Roto, V.; Law, E.; Vermeeren, A.; Hoonhout, J. User experience white paper: Bringing clarity to the concept of user experience. In Proceedings of the Dagstuhl Seminar on Demarcating User Experience, Wadern, Germany, 15–17 September 2011; p. 12.
10. Marti, P.; Iacono, I. Anticipated, momentary, episodic, remembered: the many facets of User eXperience. In Proceedings of the 2016 Federated Conference on Computer Science and Information Systems (FedCSIS), Gdansk, Poland, 11 - 14 September 2016; IEEE: Gdansk, Poland, 2016; pp. 1647–1655.
11. Hashizume, A.; Kurosu, M. UX Graph Tool for Evaluating the User Satisfaction. *Int. J. Comput. Sci. Issue* **2016**, *13*, 86–93. 10.20943/01201605.8693.
12. Badran, O.; Al-Haddad, S. The Impact of Software User Experience on Customer Satisfaction. *J. Manag. Inf. Decis. Sci.* **2018**, *21*, 1–20.

13. Matsuda, Y.; Fedotov, D.; Takahashi, Y.; Arakawa, Y.; Yasumoto, K.; Minker, W. EmoTour: Estimating Emotion and Satisfaction of Users Based on Behavioral Cues and Audiovisual Data. *Sensors* **2018**, *18*, 3978. 10.3390/s18113978.
14. Cavalcante Siebert, L.; Bianchi Filho, J.F.; Silva Júnior, E.J. da; Kazumi Yamakawa, E.; Catapan, A. Predicting customer satisfaction for distribution companies using machine learning. *Int. J. Energy Sect. Manag.* **2019**, *15*, 743–764. 10.1108/IJESM-10-2018-0007.
15. Kumar, S.; Zymbler, M. A machine learning approach to analyze customer satisfaction from airline tweets. *J. Big Data* **2019**, *6*, 62. 10.1186/s40537-019-0224-1.
16. Feng, L.; Wei, W. An empirical study on user experience evaluation and identification of critical UX issues. *Sustainability* **2019**, *11*, 2432.
17. Bolger, N.; Davis, A.; Rafaeli, E. Diary Methods: Capturing Life as it is Lived. *Annu. Rev. Psychol.* **2003**, *54*, 579–616. 10.1146/annurev.psych.54.101601.145030.
18. Klotins, E. *Usability and user experience: measurement model*; School of Computing, Blekinge Institute of Technology: Karlskrona, Sweden, 2011;
19. Körber, M.; Bengler, K. Measurement of momentary user experience in an automotive context. In Proceedings of the Proceedings of the 5th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, Eindhoven, The Netherlands, 28–30 October 2013; pp. 194–201.
20. Sánchez-Adame, L.M.; Urquiza-Yllescas, J.F.; Mendoza, S. Measuring Anticipated and Episodic UX of Tasks in Social Networks. *Appl. Sci.* **2020**, *10*, 8199. 10.3390/app10228199.
21. Hassan, R.S.; Nawaz, A.; Lashari, M.N.; Zafar, F. Effect of Customer Relationship Management on Customer Satisfaction. *Procedia Econ. Finance* **2015**, *23*, 563–567. 10.1016/S2212-5671(15)00513-4.
22. Tao, F. Customer Relationship Management based on Increasing Customer Satisfaction. *Int. J. Bus. Soc. Sci.* **2014**, *5*, 255. 10.30845/ijbss.
23. Sulaiman; Musnadi, S. Customer Relationship Management, Customer Satisfaction and Its Impact on Customer Loyalty.; SCITEPRESS, 2018; Vol. 2, pp. 692–698. 10.5220/0008892606920698.
24. Angamuthu, B. Impact of customer relationship management on customer satisfaction and its role towards customer loyalty and retention practices in the hotel sector. *BVIMSR's J. Manag. Res.* **2015**, *7*, 43–52.
25. Rahimi, R.; Kozak, M. Impact of Customer Relationship Management on Customer Satisfaction: The Case of a Budget Hotel Chain. *J. Travel Tour. Mark.* **2017**, *34*, 40–51. 10.1080/10548408.2015.1130108.
26. Taufik, M.; Renaldi, F.; Umbara, F.R. Implementing Online Analytical Processing in Hotel Customer Relationship Management. *IOP Conf. Ser.: Mater. Sci. Eng.* **2021**, *1115*, 012040. 10.1088/1757-899X/1115/1/012040.

27. Gharaibeh, N.K. Enhancing crm business intelligence applications by web user experience model. *Int. J. Adv. Comput. Sci. Appl.* **2015**, *6*, 1–6.
28. Koonsanit, K.; Nishiuchi, N. Classification of User Satisfaction Using Facial Expression Recognition and Machine Learning. In Proceedings of the 2020 IEEE REGION 10 CONFERENCE (TENCON); 2020; pp. 561–566. 10.1109/TENCON50793.2020.9293912.
29. Kaul, D. Customer Relationship Management (CRM), Customer Satisfaction and Customer Lifetime Value in Retail. *Rev. Prof. Manag.- J. New Delhi Inst. Manag.* **2017**, *15*, 55. 10.20968/rpm/2017/v15/i2/163914.
30. Shankar, V.; Winer, R.S. When customer relationship management meets data mining. *J. Interact. Mark.* **2006**, *20*, 2–4. 10.1002/dir.20062.
31. Libai, B.; Bart, Y.; Gensler, S.; Hofacker, C.F.; Kaplan, A.; Kötterheinrich, K.; Kroll, E.B. Brave New World? On AI and the Management of Customer Relationships. *J. Interact. Mark.* **2020**, *51*, 44–56. 10.1016/j.intmar.2020.04.002.
32. Nielsen Norman Group The Definition of User Experience (UX) Available online: <https://www.nngroup.com/articles/definition-user-experience/> (accessed on Jun 17, 2021).
33. Vermeeren, A.P.; Law, E.L.-C.; Roto, V.; Obrist, M.; Hoonhout, J.; Väänänen-Vainio-Mattila, K. User experience evaluation methods: current state and development needs. In Proceedings of the 6th Nordic conference on human-computer interaction: Extending boundaries, Reykjavik, Iceland, 16–20 October 2010; pp. 521–530.
34. Kujala, S.; Roto, V.; Väänänen-Vainio-Mattila, K.; Karapanos, E.; Sinnelä, A. UX Curve: A method for evaluating long-term user experience. *Interact. Comput.* **2011**, *23*, 473–483. 10.1016/j.intcom.2011.06.005.
35. Schilling, K.; Applegate, R. Best methods for evaluating educational impact: a comparison of the efficacy of commonly used measures of library instruction. *J. Med. Libr. Assoc. JMLA* **2012**, *100*, 258–269. 10.3163/1536-5050.100.4.007.
36. Karapanos, E.; Martens, J.; Hassenzahl, M. On the retrospective assessment of users' experiences over time: memory or actuality? In Proceedings of the CHI '10: CHI Conference on Human Factors in Computing Systems, Atlanta, GA, USA, 10–15 April 2010; p. 4080. 10.1145/1753846.1754105.
37. Kujala, S.; Roto, V.; Väänänen, K.; Karapanos, E.; Sinnelä, A. Guidelines how to use the UX Curve method 2013.
38. Sukamto, R.A.; Wibisono, Y.; Agitya, D.G. Enhancing The User Experience of Portal Website using User-Centered Design Method. In Proceedings of the 2020 6th International Conference on Science in Information Technology (ICSITech), Palu, Indonesia, 21-22 October 2020; IEEE: Palu, Indonesia, 2020; pp. 171–175. 10.1109/ICSITech49800.2020.9392044.
39. Pushparaja, V.; Yusoff, R.C.M.; Maarop, N.; Shariff, S.A.; Zainuddin, N.M. User Experience Factors that Influence Users' Satisfaction of Using Digital Library. *Open Int. J. Inform.* **2021**, *9*, 28–36.

40. Mominzada, T.; Abd Rozan, M.Z.B.; Aziz, N.A. Consequences Of User Experience in A Gamified E-Commerce Platform. *Int. J. Electron. Commer. Stud.* **2021**, *13*, 113–136.
41. Nwakanma, C.I.; Hossain, M.S.; Lee, J.-M.; Kim, D.-S. Towards machine learning based analysis of quality of user experience (QoUE). *Int. J. Mach. Learn. Comput.* **2020**, *10*, 752–758.
42. Keiningham, T.L.; Aksoy, L.; Malthouse, E.C.; Lariviere, B.; Buoye, A. The cumulative effect of satisfaction with discrete transactions on share of wallet. *J. Serv. Manag.* **2014**, *3*, 310–333. <https://doi.org/10.1108/JOSM-08-2012-0163>.
43. Min, K.S.; Jung, J.M.; Ryu, K.; Haugtvedt, C.; Mahesh, S.; Overton, J. Timing of apology after service failure: the moderating role of future interaction expectation on customer satisfaction. *Mark. Lett.* **2020**, *31*, 217–230. [10.1007/s11002-020-09522-y](https://doi.org/10.1007/s11002-020-09522-y).
44. Cong, J.; Zheng, P.; Bian, Y.; Chen, C.-H.; Li, J.; Li, X. A machine learning-based iterative design approach to automate user satisfaction degree prediction in smart product-service system. *Comput. Ind. Eng.* **2022**, *165*, 107939. [10.1016/j.cie.2022.107939](https://doi.org/10.1016/j.cie.2022.107939).
45. Wawre, S.V.; Deshmukh, S.N. Sentiment classification using machine learning techniques. *Int. J. Sci. Res. IJSR* **2016**, *5*, 819–821.
46. Schölkopf, B.; Smola, A.J.; Williamson, R.C.; Bartlett, P.L. New support vector algorithms. *Neural Comput.* **2000**, *12*, 1207–1245.
47. Senechal, T.; McDuff, D.; Kaliouby, R. Facial action unit detection using active learning and an efficient non-linear kernel approximation. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Santiago, Chile, 7–13 December 2015; pp. 10–18.
48. Torres-Valencia, C.; Álvarez-López, M.; Orozco-Gutiérrez, Á. SVM-based feature selection methods for emotion recognition from multimodal data. *J. Multimodal User Interfaces* **2017**, *11*, 9–23.
49. Zhang, M.-L.; Zhou, Z.-H. A k-nearest neighbor based algorithm for multi-label classification. In Proceedings of the 2005 IEEE international conference on granular computing, Beijing, China, 25–27 July 2005; IEEE; Vol. 2, pp. 718–721.
50. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
51. Kinha, Y. An easy guide to choose the right Machine Learning algorithm. *KDnuggets* 2020.
52. Haykin, S.S. *Neural networks and learning machines*; New York: Prentice Hall, 2009;
53. Li, Z.; Tian, Z.G.; Wang, J.W.; Wang, W.M. Extraction of affective responses from customer reviews: an opinion mining and machine learning approach. *Int. J. Comput. Integr. Manuf.* **2020**, *33*, 670–685.

54. Khondoker, M.; Dobson, R.; Skirrow, C.; Simmons, A.; Stahl, D. A comparison of machine learning methods for classification using simulation with multiple real data examples from mental health studies. *Stat. Methods Med. Res.* **2016**, *25*, 1804–1823. 10.1177/0962280213502437.
55. Beleites, C.; Neugebauer, U.; Bocklitz, T.; Krafft, C.; Popp, J. Sample size planning for classification models. *Anal. Chim. Acta* **2013**, *760*, 25–33. 10.1016/j.aca.2012.11.007.
56. Oppong, S.H. The problem of sampling in qualitative research. *Asian J. Manag. Sci. Educ.* **2013**, *2*, 202–210.
57. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357.
58. Longadge, R.; Dongre, S. Class imbalance problem in data mining review. *Int. J. Comput. Sci. Netw.* **2013**, *2*, 1–6.
59. Lemaître, G.; Nogueira, F.; Aridas, C.K. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *J. Mach. Learn. Res.* **2017**, *18*, 559–563.
60. Nguyen, H.M.; Cooper, E.W.; Kamei, K. Borderline over-sampling for imbalanced data classification. *Int. J. Knowl. Eng. Soft Data Paradig.* **2011**, *3*, 4–21.
61. He, H.; Bai, Y.; Garcia, E.A.; Li, S. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In Proceedings of the 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence), Hong Kong, China, 1–8 June 2008; IEEE, 2008; pp. 1322–1328.
62. Batista, G.E.; Prati, R.C.; Monard, M.C. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explor. Newsl.* **2004**, *6*, 20–29.
63. Akinbi, A.; Berry, T. Forensic Investigation of Google Assistant. *SN Comput. Sci.* **2020**, *1*, 272. 10.1007/s42979-020-00285-x.
64. Stover-Wright, E. Snowball Sampling: An Alternate Approach to Obtaining Consumer Satisfaction Responses. *J. Rehabil. Adm.* **2013**, *37*.
65. Precourt, G. What Do We Know About Peer-to-Peer Marketing? *J. Advert. Res.* **2014**, 124–125. 10.2501/JAR-54-2-124-12.
66. Spool, J. Is Design Metrically Opposed? Available online: https://www.uie.com/wp-assets/transcripts/is_design_metrically_opposed.html (accessed on Jun 21, 2021).
67. DJokić, I. The Use of Semantic Differential in Function of Measuring Image of the Company. *Econ. Anal.* **2017**, 50–61.
68. James, G.; Witten, D.; Hastie, T.; Tibshirani, R. Linear Regression. In *An Introduction to Statistical Learning: with Applications in R*; James, G., Witten, D., Hastie, T., Tibshirani, R., Eds.; Springer Texts in Statistics; Springer: New York, NY, 2013; pp. 59–126 ISBN 978-1-4614-7138-7. 10.1007/978-1-4614-7138-7_3.

69. Meng, M.; Zhao, C. Application of support vector machines to a small-sample prediction. *Adv Pet Explor Dev* **2015**, *10*, 72–75.
70. Raikwal, J.S.; Saxena, K. Performance evaluation of SVM and k-nearest neighbor algorithm over medical data set. *Int. J. Comput. Appl.* **2012**, *50*, 2.
71. Paper, D. Scikit-Learn Classifier Tuning from Simple Training Sets. In *Hands-on Scikit-Learn for Machine Learning Applications*; Apress, Berkeley, CA: Berkeley, 2020; pp. 137–163 ISBN 978-1-4842-5373-1.
72. Santos, M.S.; Soares, J.P.; Abreu, P.H.; Araujo, H.; Santos, J. Cross-Validation for Imbalanced Datasets: Avoiding Overoptimistic and Overfitting Approaches. *IEEE Comput. Intell. Mag.* **2018**, *13*, 59–76. 10.1109/MCI.2018.2866730.
73. Ng, A.Y. Preventing" overfitting" of cross-validation data. In Proceedings of the ICML, Nashville, TN, USA, 8-12 July 1997; Carnegie Mellon University: Pittsburgh, PA, USA, 1997; Vol. 97, pp. 245–253.
74. Raschka, S. Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning. *ArXiv E-Prints* **2018**, 8–9.
75. Subramanian, J.; Simon, R. Overfitting in prediction models—is it a problem only in high dimensions? *Contemp. Clin. Trials* **2013**, *36*, 636–641.
76. Cahyana, N.; Khomsah, S.; Aribowo, A.S. Improving Imbalanced Dataset Classification Using Oversampling and Gradient Boosting. In Proceedings of the 2019 5th International Conference on Science in Information Technology (ICSITech), Yogyakarta, Indonesia, 23–24 October 2019; pp. 217–222. 10.1109/ICSITech46713.2019.8987499.
77. Ben-Hur, A.; Weston, J. A user’s guide to support vector machines. In *Data mining techniques for the life sciences*; Springer: Berlin/Heidelberg, Germany, 2010; pp. 223–239.
78. Borsci, S.; Federici, S.; Bacci, S.; Gnaldi, M.; Bartolucci, F. Assessing user satisfaction in the era of user experience: Comparison of the SUS, UMUX, and UMUX-LITE as a function of product experience. *Int. J. Hum.-Comput. Interact.* **2015**, *31*, 484–495.
79. Bujlow, T.; Carela-Español, V.; Sole-Pareta, J.; Barlet-Ros, P. A survey on web tracking: Mechanisms, implications, and defenses. *Proc. IEEE* **2017**, *105*, 1476–1510. 10.1109/JPROC.2016.2637878.
80. Hussain, J.; Ali Khan, W.; Hur, T.; Muhammad Bilal, H.S.; Bang, J.; Ul Hassan, A.; Afzal, M.; Lee, S. A Multimodal Deep Log-Based User Experience (UX) Platform for UX Evaluation. *Sensors* **2018**, *18*, 1622. 10.3390/s18051622.
81. Yadav, S.; Shukla, S. Analysis of k-fold cross-validation over hold-out validation on colossal datasets for quality classification. In Proceedings of the 2016 IEEE 6th International conference on advanced computing (IACC), Bhimavaram, India, 27-28 February 2016; IEEE: Bhimavaram, Andhra Pradesh, India, 2016; pp. 78–83. 10.1109/IACC.2016.25.
82. Brownlee, J. Train-Test Split for Evaluating Machine Learning Algorithms. *MachineLearningMastery.com* 2020.

83. Maleki, F.; Muthukrishnan, N.; Ovens, K.; Reinhold, C.; Forghani, R. Machine Learning Algorithm Validation: From Essentials to Advanced Applications and Implications for Regulatory Certification and Deployment. *Neuroimaging Clin.* **2020**, *30*, 433–445. 10.1016/j.nic.2020.08.004.
84. Brownlee, J. LOOCV for Evaluating Machine Learning Algorithms. *MachineLearningMastery.com* 2020.
85. James, G.; Witten, D.; Hastie, T.; Tibshirani, R. Resampling Methods. In *An Introduction to Statistical Learning: with Applications in R*; James, G., Witten, D., Hastie, T., Tibshirani, R., Eds.; Springer Texts in Statistics; Springer US: New York, NY, 2021; pp. 197–223 ISBN 978-1-07-161418-1. 10.1007/978-1-0716-1418-1_5.
86. Kiseleva, J.; Williams, K.; Awadallah, A.H.; Crook, A.; Zitouni, I.; Anastasakos, T. Predicting User Satisfaction with Intelligent Assistants. **2016**.
87. Seyda Serdarasan; Erkan Isikli *Engineering Education Trends in the Digital Era*; IGI Global, 2020; ISBN 978-1-79982-564-7.
88. Oztekin, A.; Delen, D.; Turkyilmaz, A.; Zaim, S. A machine learning-based usability evaluation method for eLearning systems. *Decis. Support Syst.* **2013**, *56*, 63–73.
89. Munim, K.M.; Islam, I.; Khatun, M.; Karim, M.M.; Islam, M.N. Towards developing a tool for UX evaluation using facial expression. In Proceedings of the 2017 3rd International Conference on Electrical Information and Communication Technology (EICT); IEEE, 2017; pp. 1–6.
90. Davidson, R.J.; Ekman, P.; Saron, C.D.; Senulis, J.A.; Friesen, W.V. Approach-withdrawal and cerebral asymmetry: emotional expression and brain physiology. *I. J. Pers. Soc. Psychol.* **1990**, *58*, 330–341.
91. Branco, P. Usability Indicators-In Your Face [Online], Procs. In Proceedings of the Computer Human Interaction; 2006.
92. Keltner, D. Signs of appeasement: Evidence for the distinct displays of embarrassment, amusement, and shame. *J. Pers. Soc. Psychol.* **1995**, *68*, 441–454. 10.1037/0022-3514.68.3.441.
93. Harley, J.M. Measuring Emotions: A Survey of Cutting Edge Methodologies Used in Computer-Based Learning Environment Research. In *Emotions, Technology, Design, and Learning*; Tettegah, S.Y., Gartmeier, M., Eds.; Emotions and Technology; Academic Press: San Diego, 2016; pp. 89–114 ISBN 978-0-12-801856-9. 10.1016/B978-0-12-801856-9.00005-0.
94. Huang, Y.; Chen, F.; Lv, S.; Wang, X. Facial Expression Recognition: A Survey. *Symmetry* **2019**, *11*, 1189. 10.3390/sym11101189.
95. Li, S.; Deng, W. Deep Facial Expression Recognition: A Survey. *ArXiv180408348 Cs* **2018**.
96. Arriaga, O.; Valdenegro-Toro, M.; Plöger, P. Real-time convolutional neural networks for emotion and gender classification. *ArXiv Prepr. ArXiv171007557* **2017**.

97. Cunha, P. petercunha/Emotion 2020.
98. Ayman, O. omar178/Emotion-recognition 2020.
99. Correa, E.; Jonker, A.; Ozo, M.; Stolk, R. Emotion recognition using deep convolutional neural networks. *Tech Rep. IN4015* **2016**.
100. Balaji, A. atulapra/Emotion-detection 2020.
101. Akcora, E. ezgiakcora/Facial-Expression-Keras 2020.
102. Jyoti, E.; Walia, E.A.S. “A Review on Recommendation System and Web Usage Data Mining using K-Nearest Neighbor (KNN) method. *Int. Res. J. Eng. Technol. IRJET* **2017**, 4, 2931–2934.
103. Pal, S.K.; Mitra, S. Multilayer perceptron, fuzzy sets, and classification. *IEEE Trans. Neural Netw.* **1992**, 3, 683–697. 10.1109/72.159058.
104. Tharwat, A. Classification assessment methods. *Appl. Comput. Inform.* **2018**. 10.1016/j.aci.2018.08.003.
105. Géron, A. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*; 2nd edition.; O’Reilly Media: Beijing Boston Farnham Sebastopol Tokyo, 2019; ISBN 978-1-4920-3264-9.

ACKNOWLEDGEMENT

I would like to express my sincerest gratitude to my best supervisor, Professor Nobuyuki Nishiuchi, for his invaluable guidance, support, and encouragement throughout my PhD journey. I am deeply grateful for the warm welcome he provided during my research and for facilitating my transition to life in Tokyo.

I also extend my appreciation to the members of my research committee: Professor Kentaro Go, Professor Yasufumi Takama, and Associate Professor Dr. Takao Fukui, for their insightful comments and suggestions. I am grateful to all lab members for their support and for helping me settle into life in Tokyo Metropolitan University. I am particularly thankful to Assistant Professor Vibol Yem for his valuable suggestions.

Finally, I would like to dedicate this dissertation to my parents and older brother, who have provided me with unwavering support and encouragement. Their support has been invaluable to me throughout this journey.

I acknowledge Tokyo Human Resources Fund for City Diplomacy scholarship established by the Tokyo Metropolitan Government for granting me with a full scholarship at the Tokyo Metropolitan University, Hino campus. Without this support, it would not be possible to pursue my Ph.D. study. Additionally, I express my gratitude for the financial support received from the JSPS KAKENHI grant (JP20K12511) and the local-5G project of the Tokyo Metropolitan University, which supported the work presented in this thesis.