CLASSIFYING SMOKING URGES VIA MACHINE LEARNING

by

Antoine Dumortier

B.S. in Electrical and Computer Engineering, ENSEA, 2014

Submitted to the Graduate Faculty of

the Swanson School of Engineering in partial fulfillment

of the requirements for the degree of

Master of Science

University of Pittsburgh

2015

UNIVERSITY OF PITTSBURGH

SWANSON SCHOOL OF ENGINEERING

This thesis was presented

by

Antoine Dumortier

It was defended on

July 15, 2015

and approved by

Ervin Sejdić, Ph. D., Assistant Professor,

Department of Electrical and Computer Engineering

Amro El-Jaroudi, Ph. D., Assistant Professor,

Department of Electrical and Computer Engineering

Zhi-Hong Mao, Ph. D., Assistant Professor,

Department of Electrical and Computer Engineering and Bioengineering

Thesis Advisor: Ervin Sejdić, Ph. D., Assistant Professor,

Department of Electrical and Computer Engineering

CLASSIFYING SMOKING URGES VIA MACHINE LEARNING

Antoine Dumortier, M.S.

University of Pittsburgh, 2015

Smoking is the largest preventable cause of death and diseases in the developed world, and advances in modern electronics and machine learning can help us deliver therapies to smokers in novel ways. If a mobile device monitoring a smoker's situation could detect when the smoker is likely to have an urge to smoke, it would be helpful for optimizing the timing of real-time intervention. In this thesis, we examine different machine learning approaches to use situational features associated with having or not having urges to smoke during a quit attempt in order to accurately classify high-urge states. To test our machine learning approaches –specifically naive Bayes, discriminant analysis and decision tree learning methods– we used a dataset collected from over 300 participants who had recently initiated a quit attempt. The three classification approaches are evaluated observing sensitivity, specificity, accuracy and precision. The outcome of the analysis showed that algorithms based on feature selection make it possible to obtain high classification rates with only a few features selected from the entire dataset. The classification tree method outperformed the naive Bayes and discriminant analysis methods, with an accuracy of the classifications up to 86%.

Keywords: smoking urges, smoking cessation, machine learning, supervised learning, discriminant analysis classification, naive Bayes classification, decision tree classification, feature selection.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

I would like to thank Dr. Mahmoud El Nokali for giving me the chance to study at the University of Pittsburgh in the Swanson School of Engineering. I would also like to express my greatest gratitude to my advisor, Dr. Ervin Sejdić, who has given a welcoming attitude and useful guidance. Finally, I am very grateful to my colleagues and my family for their support and their help during my master studies.

1.0 INTRODUCTION

1.1 SMOKING AND TOBACCO USE

Centers for Disease Control and Prevention define a "smoker" as persons who declared smoking 100 or more cigarettes in their entire life and who, at the time they took part in a study or a survey, declared smoking every days or some days. According to this definition, there are 1.1 billion smokers in the world (15.4% of the world population), and this number is expected to increase to 1.6 billion over the next two decades [1]. Worldwide, tobacco use causes more than 5 million deaths per year [1]; that is to say, one person dies every six seconds from a tobacco related disease. In addition, there is approximatively 50% chance that a regular smoker will die precipitately from a complication due to smoking [2], and current trends show that tobacco and smoking activities will lead to the death of more than 8 million people annually by 2030 [1]. Tobacco smoking is a highly addictive practice due to the release of nicotine during the combustion of cigarettes [3], and it is generally the consequence of a nicotine dependence [4]. This addiction is currently the single largest preventable cause of death and diseases in the developed world [5]. Moreover, the overall mortality among both male and female smokers is about three times higher than that among similar people who never smoked [6]. Smoking is proven to induce several serious diseases such as lung cancer, coronary heart, or cardiovascular diseases [7], and meanwhile, smoke emitted by smokers near non-smokers is known to be dangerous and a potential cause for lung cancer [8] or heart attack [9]. Cigarette-smoking is also a source of problem for respiratory tract and represents a risk factor for infections, gastric ulcers, diabetes, and even fire-related or trauma-related injuries [10]. Diseases are generally caused by an excessive exposure to toxins present in tobacco smoke. The chart in Figure 1.1 presents the estimated average annual

number of smoking-attributable deaths in the United States during 2005 through 2009 by specific causes. Each year in the United States, cigarette smoking is responsible for the death of more than 440,000 Americans, including more than 9% related to exposure to secondhand smoke. As a result, the economic consequences for the country are critical. Smoking-related illness in the United States costs more than $300 billion each year, including at least $170 billion in direct medical care for adults and more than $156 billion in lost productivity [6, 11].

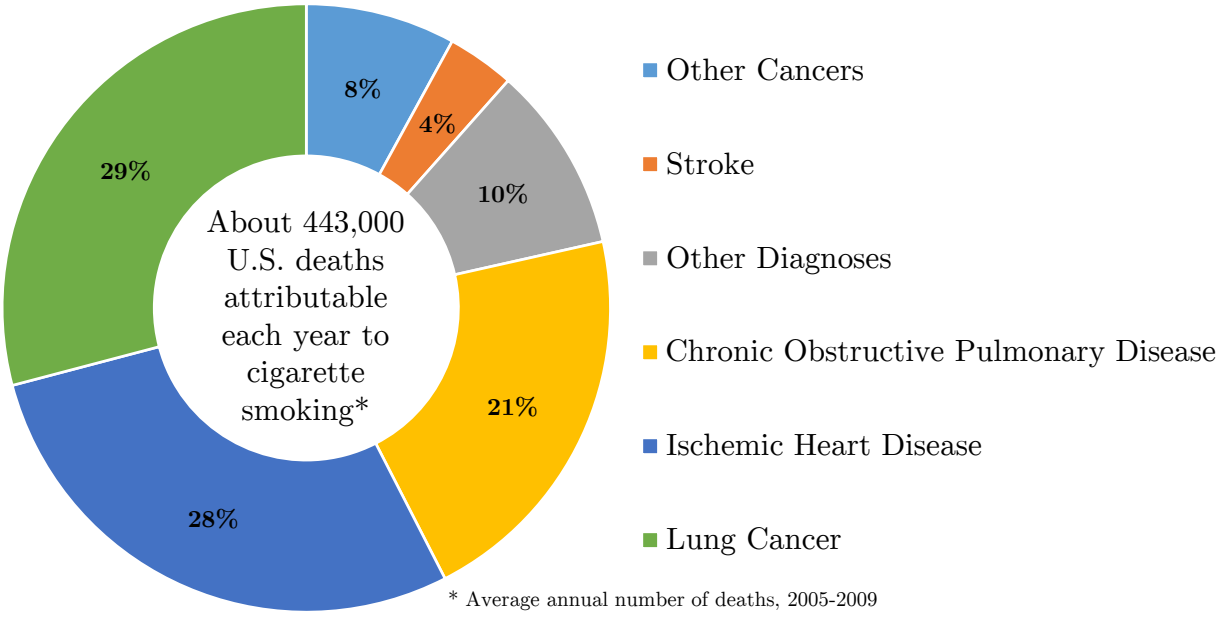


Figure 1.1: Annual Deaths Attributable to Cigarette Smoking in the United States from 2005 to 2009. Adapted from the U. S. Department of Health and Human Services (2014) [6].

In 2013, an estimated 18% (42.1 million) U.S. adults were current cigarette smokers [12] and 78.4% (33 million) of these adults smoked every day [13]. In average, there are more men smokers than women (20.5% versus 15.3%) and current cigarette smoking is higher among adults aged 18-24 years, 25-44 years, and 45-64 years than among those aged 65 years and older [13]. Besides, recent statistics showed that additional factors have to take into consideration when studying current cigarette smoking among adults in the United States: education, poverty status, geographical location, disability or limitation, sexual orientation...

In 2011, 68.8% of adult cigarette smokers wanted to stop smoking completely [14], and 42.7% had made a quit attempt in the past year [6]. Since 2002, the number of former smokers remains above the number of current smokers. The chart in Figure 1.2 gives more specific data about the percentage of adult who succeeded in stopping smoking for more than one day. The lowest percentage corresponds to adults aged 65 years or older, who are also the adults who smoke the less (9%) [13]. On the contrary, the three highest percentages (adults aged between 18 and 64 years) in Figure 1.2 coincide with the three categories of adults who smoke the most (20%) [13]. As explained in [15], highly addicted smokers make serious quit attempts, but they are able to maintain this situation for only a few hours. Each year, only 3% of smokers' attempts are successful, and unfortunately, the rate at which people become daily smokers approximatively corresponds to the quit rate. That partly explains why the preponderance of cigarette smoking has decreased very slowly over the past few years [16].

In the meantime, a lot of research studies have identified efficient actions that could contribute to put an end to tobacco use [17, 18, 19, 20]:

- Implementing smoke free laws.
- Increasing fundings for programs that control tobacco use.
- Raising tobacco prices.
- Reducing tobacco advertising and promotion.
- Providing support and helping smokers who are willing to quit.

1.2 SMOKING CESSATION PROCESS

It is clear and well known that rejection of smoking would produce significant benefits for public health, allowing the reduction of short-term health care costs of smoking [21]. In return, this costs' reduction improves the available resources for people that are willing to quit smoking. But one could wonder what benefits smokers could have if they quit smoking. Firstly, one of the most common observed benefit of smoking cessation is the increase of life expectancy [22]. Quitting smoking is known to reduce or eliminate the risks of having one of

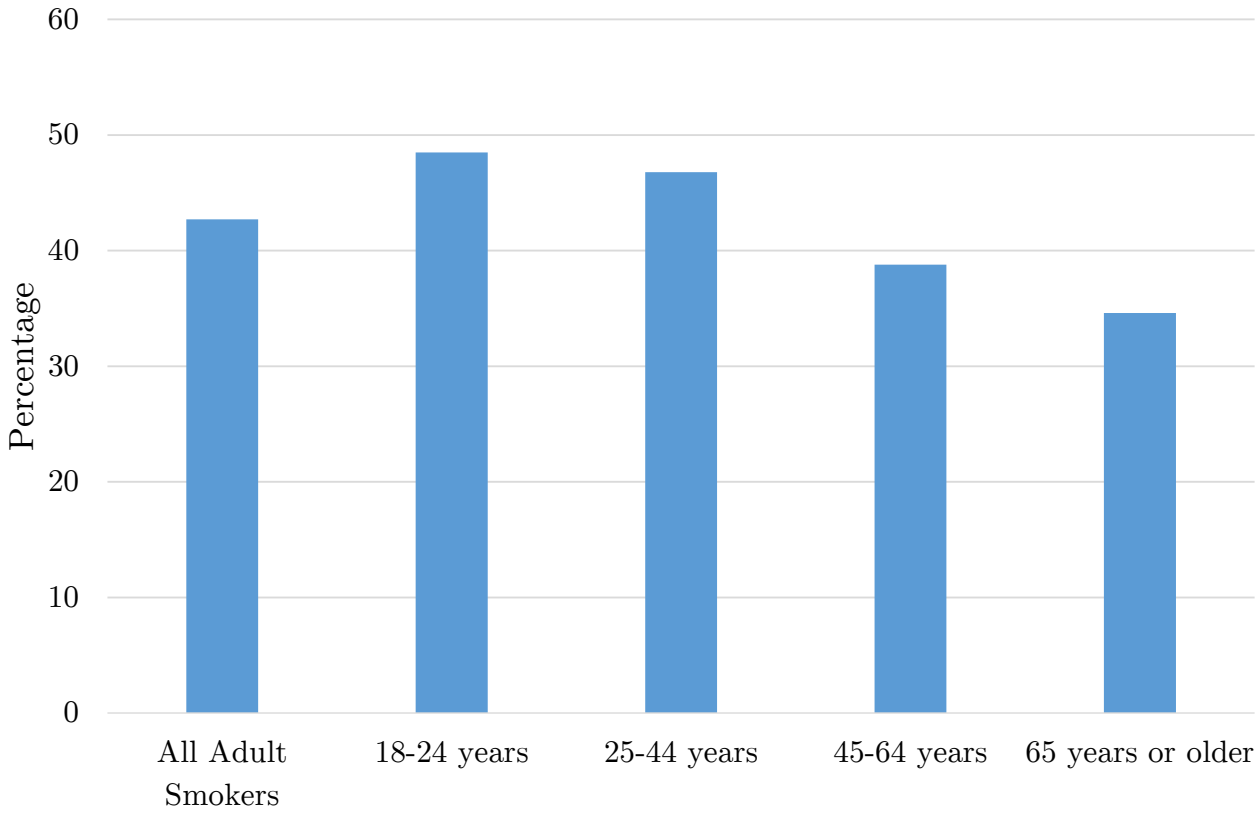


Figure 1.2: Percentage of adult daily cigarette smokers who stopped smoking for more than one day in 2010 because they were trying to quit. Adapted from the Center for Disease Control and Prevention (2011) [14].

the diseases presented in Figure 1.1. For example, the risk of dying from lung cancer, which is the most common cause of cancer death in both men and women, is 22 times higher among male smokers and 12 times higher among female smokers compared with people who have never smoked [22]. But more importantly, after 10 years of abstinence the risk of lung cancer is divided by three [22]. Similarly, risks of dying from a Coronary Heart Disease (CHD) or a stroke, which are respectively the first and third leading cause of death in the United States, are reduced by about half among ex-smokers after only one year of smoking abstinence, and declines significantly from that time [22]. More generally, many investigations have shown

that for subjects who took a break in smoking for a relatively long time, the health benefits increased and the risk of diseases decreased proportionally with the number of years since cessation [22, 23].

Other studies also showed that people can live considerably longer when they stop smoking, and this is not even related to the age at which they abandoned the cigarettes [24]. More particularly, results based on US-specific data showed that the majority of mortality from smoking could be avoided by quitting smoking at 35 years old [25]. Even smokers who quit when they are 65 or older appeared to expand their life by up to 3.7 years, compared to those who keep smoking [25]. All these findings support the emergency of emphasizing smoking cessation to all smokers, without considering their age or gender.

Abstinent smokers usually have to deal with frequent challenges to abstinence, which can appear as moments of intense craving [26, 27, 28], urge to smoke [26], or nicotine withdrawal [29]. The overall process of relapse is summarized in Figure 1.3. The situation in which a subject feels a strong desire of smoking is called *relapse*, and can be considered as the main step in order to achieve abstinence. When a crisis is cleared up without smoking, it is identified as a *temptation* episode. On the other hand, if a smoking situation happens after a relapse crisis, the step is named *lapse*. By themselves, lapses following relapse crisis could be riskless for the subject if would not lead to relapse, i.e., substantial resumption of smoking. Unfortunately, quit attempts are typically unsuccessful and end in failure [30]. In other words, most of the relapse crisis lead to smoking situations. Although behavioral and pharmacological medication can help subjects to quit [18, 31, 32], success rates remain modest and relapse is generally observed in between 95% and 99% of the unaccompanied quit attempt cases [33].

1.3 RESEARCH OBJECTIVES

Our first goal will be to carry out a comparative analysis of machine learning algorithms to determine if we are able to provide smoking urges classifications. Three different classifiers will be compared: naive Bayes, discriminant analysis, and classification tree. Two approaches

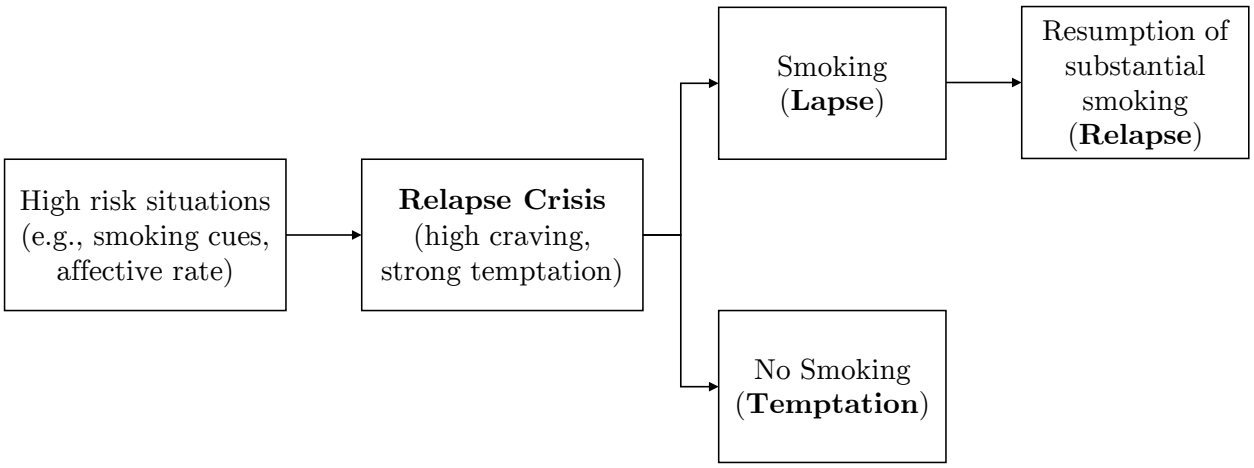


Figure 1.3: Schematic diagram of relapse crises, temptations, lapses, and relapse. Adapted from Shiffman (2005) [34].

will be considered with these three classifiers: the first one will be based on the whole dataset, while the second one will be based on a subset of eight contextual features. Then, implementing feature selection algorithms, the second goal is to extract the most relevant features (in a given dataset) that can provide the best classifications of urges to smoke. A sequential feature selection algorithm will be used to extract the features.

Chapter 2 provides a presentation of the different notions used in this study. Then Chapter 3 describes the methodology for the different phases of the research. Before giving a presentation and technical information about the three classification methods used in this study, explanations about how the data selection is made and how the data are presented are given. Validation techniques, which are implemented to split the dataset between a training set, to create and train the model, and a testing set, to test the model, are then introduced. We also operate and compare feature selection algorithms in an attempt to exclude useless features and only select those which provide the best results. Chapter 4 presents the classification results and the final selected features, as well as the relevance of the feature selection. Implications for these results are overviewed in Chapter 5. Finally, conclusions and future directions for the study are provided in Chapter 6 followed by a list of references.

2.0 BACKGROUND

2.1 BRAIN MECHANISMS OF THE NICOTINE ADDICTION

Dependence to tobacco is the consequence of various interrelated factors (neurochemical, environmental and individual), and all of them have to be taken into consideration in order to find treatment for tobacco dependence [35, 36]. Inhalation of cigarette's smoke extracts the nicotine from tobacco in the cigarette. Smoke particles bears the nicotine into the lungs, before it is absorbed by the pulmonary venous circulation and penetrates the arterial circulation. Then the nicotine goes almost instantaneously (10 to 19 seconds) from the lungs to the brain. Nicotine levels then declines thanks to the absorption by peripheral tissues that allow its removal from the body [37].

One of the first effect is the activation of nicotinic cholinergic receptors located on specific neurons in the ventral tegmental area [10, 38]. This process allows the entry of cations such as sodium and calcium into the brain cells [39]. This is known to activate calcium channels, which are very important for a lot of cellular functions [40], allowing more calcium to enter in the cells. The stimulation of nicotinic cholinergic receptors, that allows neurons to access calcium, will release a variety neurotransmitters [39, 41] and change the practical state of the cells [40]. One of them, called dopamine, is responsible for reporting delightful experiences and is critical for the management of nicotine [42]. Studies conducted with animals like rats submitted results showing that the mesolimbic dopamine system is fundamental for the reinforcement of nicotine's effects [43]. It is indeed responsible for the projection of dopamine from the ventral tegmental area of the midbrain to the nucleus accumbens [37]. It is also associated with the regulation of other drugs abused by humans, such as cocaine [44]. Dopamine is delivered by nicotine in the amygdala, the nucleus accumbens and the prefrontal

cortex (Figure 2.1). Outside the nervous system, dopamine can be found in various part of the body (e.g., blood vessels, kidneys, pancreas, digestive systems...) where it operates as a local chemical agent [45].

Nicotine is also responsible for the augmentation of both glutamate and γ-aminobutyric acid (GABA) release [46, 3]. Glutamate assists the release of dopamine, while GABA restricts it so that is why a non-addictive subject manages to keep normal dopamine's level. In the case of long-term exposure to nicotine, some GABA receptors become insensitive. As a result, the GABA acid becomes ineffective while the glutamate acid excitation remains active, thereby improving the receptivity to nicotine.

Nicotine also plays an important role for cardiovascular functions, via the stimulation of the sympathetic nervous system [47]. This stimulation occurs after the activation of receptors present in the nervous-system, and is partially responsible for heart rate acceleration or the increase of blood pressure [48] in addition to the inducement of atherogenic genes coronary artery endothelial cells [49]. These effects of tobacco smoking on the cardiovascular system can brought to an end with phentolamine [37].

Cigarette smoke also contains other components that contribute to nicotine addiction. One of them, the monoamine oxydases, a family of enzymes that can be found in neurons [10], reduce the metabolism of dopamine and are proven to be partly responsible for the addictiveness of smoking [50, 51]. Besides, chronic exposure to nicotine generates tolerance and durable modifications to the brain of subjects (human or animal) participating in chronic treatments focused on active drugs [52]. It is called *neuroadaptation*, and it is responsible for some known issues of the nicotine [53] because it creates a brain condition that requires the nicotine level to be continuously maintained [54].

Neuroadaptation also describes the durable modifications provoked to the brain , Once the nicotine is removed from the body, withdrawal syndromes start to appear causing stress and anxiety. These symptoms are known to be very strong enticement to resume smoking [55].

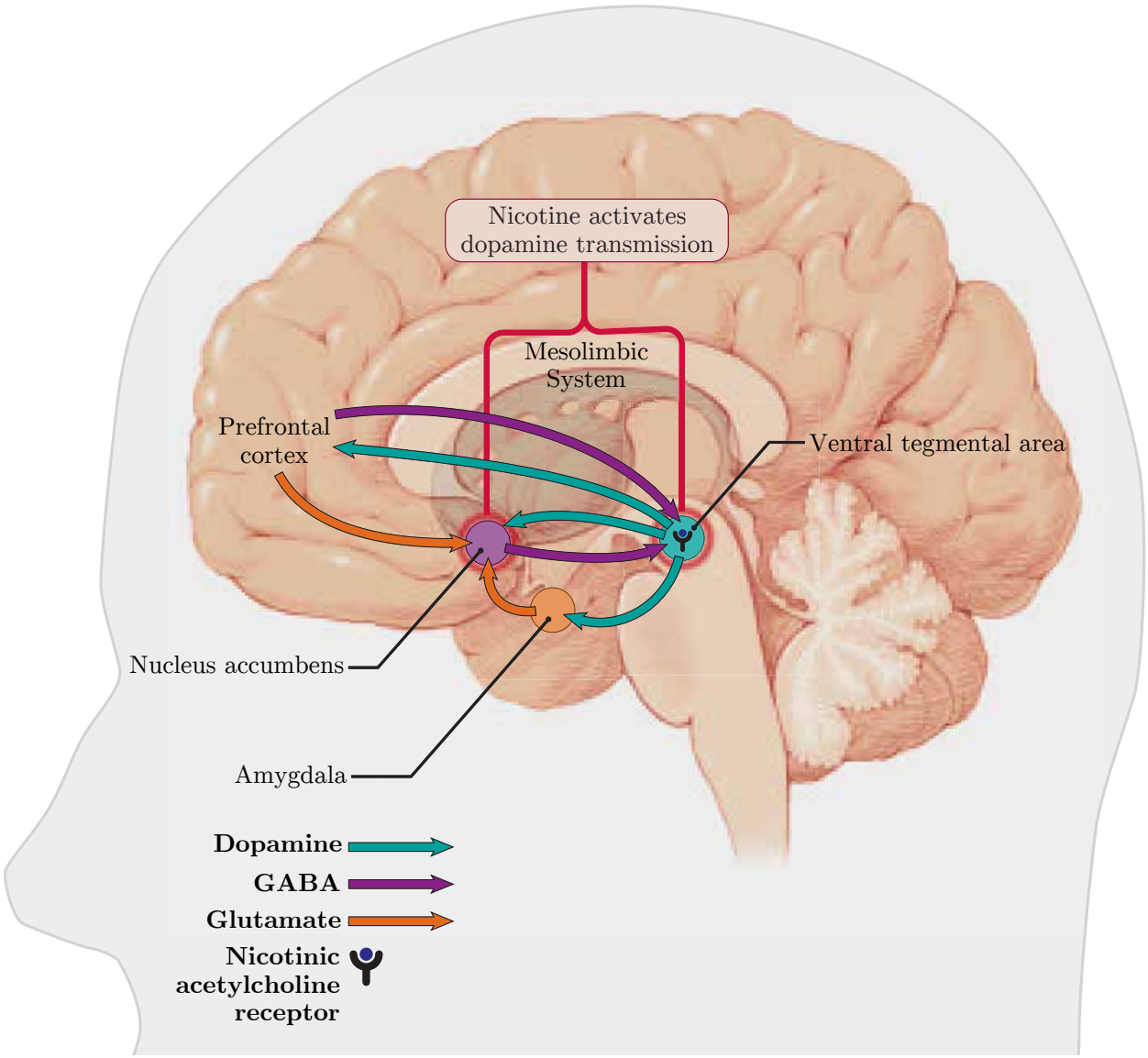


Figure 2.1: Areas in the brain involved in nicotine addiction. Nicotine activates receptors in the ventral tegmental area, resulting in dopamine release inside the mesolimbic area and the prefrontal cortex. Adapted from Le Foll (2007) [38].

2.2 CLINICAL ASPECTS OF NICOTINE ADDICTION

Nicotine is a psychoactive drug, that is to say it is a chemical substance that affect brain functions and results in alterations in perception, mood or consciousness [56]. Smokes generally use it to adjust their excitement and to temper their attitude because it is responsible for the activation of brain regions that are accountable for stress and anxiety decrement. Smoking also helps to improve the execution of certain tasks by improving the concentration and boosting the reaction time [10]. On the other hand, withdrawal symptoms like depression, anxiety, feelings of restlessness or frustration [57, 58] often occur. Anhedonia, defined as the inability to experience pleasure from activities usually found enjoyable [59], can also be listed as one of the psychoactive effects of nicotine [60].

The action of smoking a cigarette, which allow to provide nicotine to the brain, is caused by a strong desire of smoking, itself generated by stress [61], smoking cues [62], or desire to relieve withdrawal symptoms [63]. It provokes the activation of nicotinic cholinergic receptors (nAChRs), inducing the release of neurotransmitters (e.g., dopamine) occasioning pleasure feelings and temperament modulation. After being activated by nicotine, brain receptors become finally desensitized to it. The consequence of this desensitization is reduction of the pleasure acquired from smoking, and a short-term resistance to the effects of nicotine. Brain nicotine levels appear to decrease during the between cigarettes smoking interval of time, but also after quitting period. This leads to a reduction of neurotransmitters level, such as dopamine, and to withdrawal symptoms. In situations where nicotine is missing, nicotinic cholinergic receptors get their responsiveness to nicotine back [10]. This whole process describing the molecular and behavioral aspects of the nicotine addiction is presented in Figure 2.2.

Conditioning is assimilated with the expansion of smoking addiction, since former addicted smokers often experience urges of smoking after withdrawal symptoms dissipate [10]. After short periods of cessation, relapses are generally caused by smoking-related cues like particular environmental situations or emotions that remind old smokers the pleasant effects of nicotine. The fact that subjects associate well-known situations with nicotine represent a kind of conditioning, which has been demonstrated in [65] and [66]. These studies showed

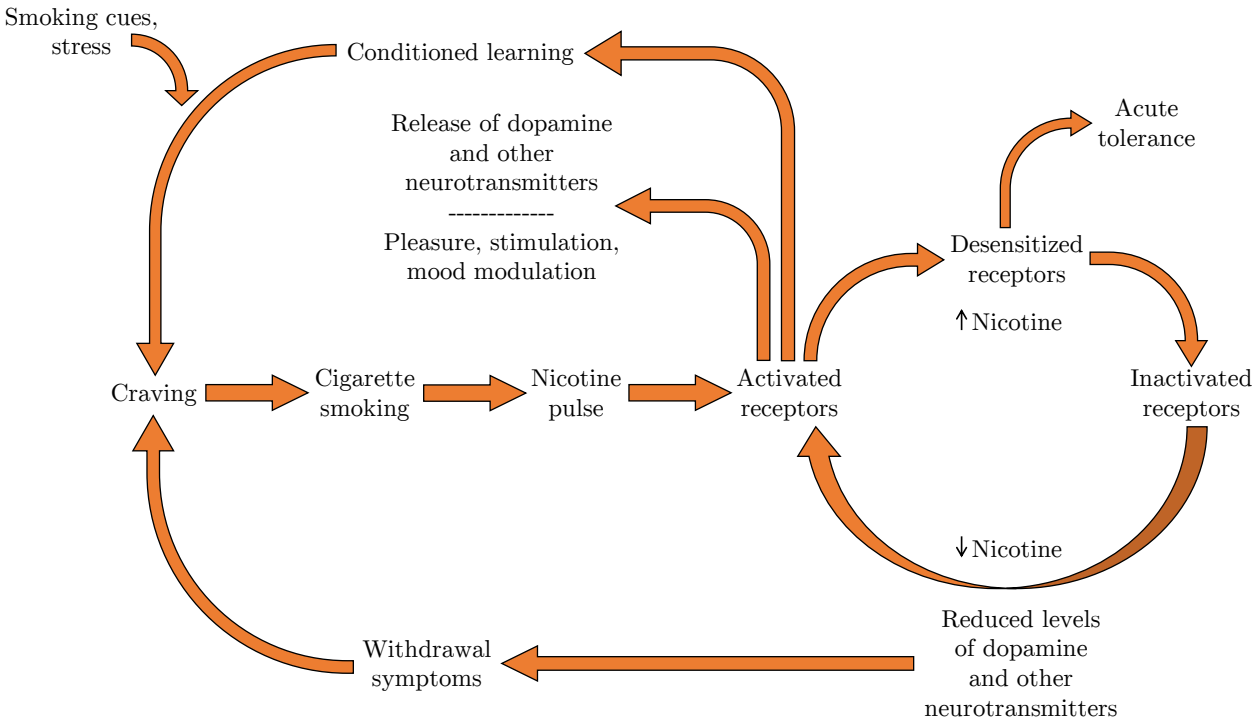


Figure 2.2: Molecular and behavioral aspects of nicotine addiction. Adapted from Dani and Heinemann (1996) [64].

that nicotine is highly responsible for the modification of the behavior of specific brain cells. Moreover, body reaction to conditioned stimuli can also be intensified by nicotine, leading to uncontrollable smoking [67].

This conditioning allows to keep up with a continuous desire of smoking. Repeated and pleasant situations like smoking a cigarette during lunch break with a friend or after dinner with an alcoholic drink are a strong cue for smoking urge [10]. In addition to these situations, characterization of smoking like the smell or the taste is a powerful factor that influence and reinforce the nicotine addiction [68, 69]. Sometimes, even bad feelings can trigger smoking: smokers can realize that going without cigarettes for a while make them bad-tempered, and that smoking one bring satisfaction and happiness back [10]. Furthermore, while cigarettes are responsible for the release of nicotine in the brain, they are designed with

added ingredients in order to strengthen their addictiveness [70] and additionally, billions of dollars are spent every year by tobacco companies on advertising and promotional expenses [71]. On the contrary, nicotine medication is conceived to carry nicotine slowly in the body, reducing the risks of damage [72].

As soon as nicotine becomes less available, smokers try to adjust their attitude towards smoking to counterbalance the change. They usually look after a regular consumption to adjust their level of nicotine and reach the expected effects [73]. Light smokers (i.e., people who smoke less than five cigarettes per day [10]) and occasional smokers generally use cigarettes for the pleasant effects of nicotine, so that is why they rarely suffer from withdrawal symptoms [74]. For this category of smokers, the main occasions of smoking are during particular activities like after eating or during drinking alcohol and besides, negative affect is a less likely cause of smoking need.

2.3 DATA COLLECTION METHODS

2.3.1 Current measurements of social behavior

Clinician reports are currently the most commonly used method to study affect, mood and interpersonal behavior in psychopharmacological studies [75]. They consist in the collection of information from clinicians who are close to the patients and have an accurate knowledge of their situation [75]. Several formats and rating scales, such as the Hamilton Rating Scale for Depression (HAM-D) [76], have been used to evaluate mood and social behavior. However, these approaches based on clinician's judgments present several disadvantages. The first one is that the data collection and/or the interpretation of the collected data are certainly affected by the closeness of the clinician with the study participant [75]. Limitations of clinician ratings have also been illustrated by Rosenhan's study [77], which showed that clinicians are sometimes unable to discern healthy from unhealthy patients. Finally, most individuals including clinicians who are conducting these studies usually have expectations and make suppositions before getting the results [78]. These assumptions can lead to biased analysis.

Other methods, such as self-reports questionnaires where information is collected in the form of global ratings [75], or laboratory tests where participants have to answer to one or two brief situations [75], have many advantages. They can be adapted in various situations, and this allows to measure a large spectrum of affects, moods and personal behaviors. However, the dominant disadvantage is that reconstructive processes regularly affect remembered information [75], leading in reduced results' accuracy. That is why methodologists and scientists presented support to methods depending on data collected closer in time to the occurrence of the considered events, allowing to reduce bias linked to memory and thoughts, and to improve the global accuracy [79].

2.3.2 An alternative: Ecological Momentary Assessment

Ecological Momentary Assessment (EMA) [80] can be characterized as an alternative method to static retrospective reports. Indeed, EMA allows a collection of real-time and real-word data informing about subjects' behavior and life experiences [81], and its main interest is that it presents the possibility to collect measure close in time to the experience [75, 82].

The EMA method can easily be applied to smoking cessation processes since it provides circumstances where subjects' behaviors can be evaluated without difficulty, facilitating the study of quitting and relapsing cases. It involves the collection of repeated sampling of subjects' real world mood, thoughts and state of mind at specific and random times during the day [81, 83], through completion of assessments in subject's daily routine using traditional diaries [84], palmtop computers [85] (Figure 2.3), or telephones [86].

These assessments are designed to capture not only the events associated with lapses, but also the mood, behavior and environmental situation in the hours and days preceding and following lapses [81]. That is why they allow to provide answers to a large variety of questions like: How much craving do smokers experience when they quit smoking? [27] Does craving intensity vary with individual characteristics, such as nicotine dependence? [87] Do situational factors affect lapse risk? [85]...

Figure 2.3: Example of an electronic diary. (Mark Richards, 1991)

In order to guarantee efficient results, EMA approaches are generally based on the following key rules [80, 88]:

- The "ecological" aspect of the method is justified by the fact that the data have to be collected in real-world environments.
- The "momentary" aspect imposes to obtain data from the subject's present or recent state. It is a real-time assessment and this allow to minimize the error and bias attributed to memory and consciousness.
- Instants of collection have to be strategically chosen in order to maximize the relevance with the area of study. For example, assessments can be suggested during craving period, but they can also be submitted at random times in order to provide a more global characterization of subjects' state.
- The last important rule is to ask patients to complete several assessments during short period of times, providing a comprehensive archetype of the patient's situation and behavior.

Researches based on EMA have been conducted to study relapse among groups of people who have started a smoking cessation process. In [89, 90], a subset of features previously associated with lapses is used to analyze how craving, emotion, and social environment impact on smoking rate [89]. In [90], self-reports of contextual variables were analyzed to examine correlates of craving when cigarettes were smoked. Results showed, for example, that craving was higher when cigarettes were smoked while eating or drinking, during activity, and early in the day. On the other hand, craving does not appear to be related with the location, alcohol, or caffeine. However, there is variability in the evidence regarding the degree to which different contextual features are associated with smoking risk. For example, during a quit attempt [91], self-reported temptation episodes (i.e., intense craving to smoke) were associated with negative moods, exposure to others smoking, and food, coffee, or alcohol consumption.

Like other methods, EMA has its disadvantages. Similarly to other self-report measures, there is no way to check the veracity of the information given by the participants, since a scientist is not required to collect the data. Besides, it is more consuming than meeting with a clinician at specified times [75].

2.4 MACHINE LEARNING

In a wide variety of fields, objects have to be categorized. To perform these tasks, several properties of the objects need to be known. A general example could be the classification of birds, where birds are assigned to their species. Different characteristics can be considered, such as their sounds, their wingspans, their weights, etc... Then, based on a set of recorded characteristics, a considered bird can be assigned to a specie. This whole process is called "machine learning", the set of characteristics is called "features" and the different classification labels are the "classes" [92, 93, 94, 95].

Machine learning, firstly introduced and defined in 1959 by Arthur Samuel, an American experts in computer gaming and artificial intelligence, can be defined as an area of study where computers have the capacity to learn from data without having recourse to explicit

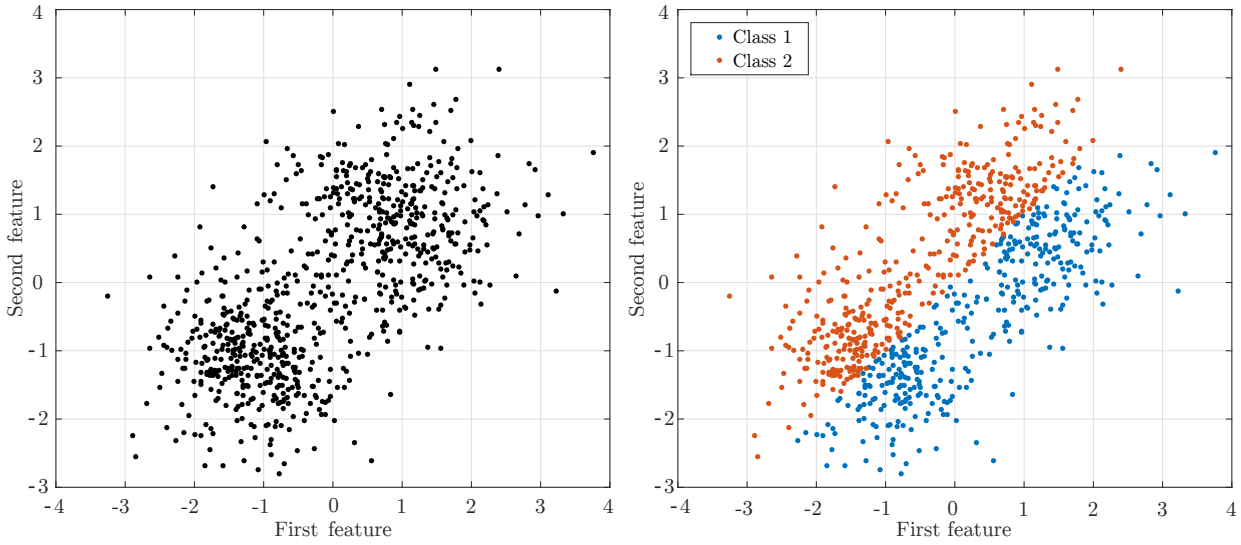
programmed instructions [96]. A training selection is used to learn a set of rules [97] in order to create a classifier [98] which will be able to make predictions. This subfield of computer science and statistics is nowadays one of the most reliable solutions to address problems of pattern recognition. Indeed, the field of pattern recognition is focused on the automatic recognition of data similarities in order to establish classifications in different categories [99].

Machine learning tasks can be separated into two major subfields, which depend on the nature of the available dataset. If the pattern recognition system has a labeled training data, that is to say each observation maps to a given class, the learning task is called supervised learning. On the other hand, the unsupervised learning task occurs when the dataset is unlabeled. Thus, the learning algorithm is required to find the different hidden patterns.

2.4.1 Unsupervised machine learning

Dealing with unsupervised learning always raises the issue of unsupervised classification, also called clustering. This term, which refers to the grouping task of unlabeled patterns into meaningful clusters [100], appears to be very useful in several classification studies, or grouping and machine-learning situations, including data mining, document retrieval, or image segmentation [100]. Nevertheless, a prior information about the data is often required to bring classifications to more successful conclusion.

The clustering task is used to partition the dataset into separated groups, and the main purpose of this classification is to reach a state where data in the same group share similarities, and data in different groups share differences. Figure 2.4b shows an example of a possible partition of the dataset presented in Figure 2.4a. Besides, it is often difficult to find an optimal method or discriminant criterion because the definitions of similarities or differences are not always clear, and also because there are many clustering techniques. The most famous clustering algorithms are the k-means [101], the fuzzy C-means [102], the information theory-based clustering [103], the expectation-minimization algorithm [104] and neural network models such as Kohonen's self-organizing maps [105].



(a) Randomly generated dataset.

(b) Classification, using the dataset on the left.

Figure 2.4: Example of unsupervised classification, using a k-means clustering method.

2.4.2 Supervised machine learning

The previous studies, as well as the recent advances in computational algorithms, has led us to believe that machine learning approaches can be useful in the process of smoking cessation. Mobile devices are nowadays increasingly utilized to provide assistance to smokers trying to quit (e.g., [106, 107]). However, very few of these systems collect data from the user in order to understand, in real-time, the context in which their urges to smoke occur [108], and that is why most of the time there is no way to be optimally responsive to the dynamic nature of the experiences smokers are facing with when they are trying to quit [109].

Because urges to smoke are contextually-dependent, one way to potentially enhance the power of using mobile devices to promote smoking cessation is to use machine learning algorithms, and more particularly supervised machine learning algorithms.

This research project deals with supervised learning (Definition 2.4.1, [110]) where the computer's goal is to map labeled input features to output classes [111]. The idea behind this concept, also known as pattern recognition, is to create algorithmic tools that are able to perform accurate predictions automatically.

Definition 2.4.1 (Supervised Learning)**.** Given a set of data $\mathcal{D} = \{(x_i, y_i), i = 1, \ldots, N\}$, the supervised learning task consist in "learning" the relationship between the input x and the output y, such that, when a new input x^* is provided, the new predicted output y^* is accurate. The meaning of "accuracy" can be specified using a loss function $L(y_{pred}, y_{true})$.

The main interest in supervised learning is to describe y when x is known, that is to say knowing the conditional probability $P(y|x, \mathcal{D})$. The word "supervised" then refers to a "supervisor" in charge of labeling the output y (often called "label") for each input x in the available dataset $\mathcal{D}$.

A supervised learning task usually has a design similar to the one shown in Figure 2.5. The first step consists in the data collection. It requires either an expert to pick the most descriptive attributes, or a selection methods to segregate the data [97]. However, as usually, the dataset includes missing and/or erroneous features [112] so that is why the second step is a data preparation (i.e., pre-processing) step[113]. Its purpose is to get a ready-to-use dataset containing the required features. The first step (data collection) is decisive to be able to perform efficient classification training (fourth step) and testing (fifth and last step). Then the classifier is chosen (third step) according to the features determined in the second step, and in agreement with the classification requirements. The training operation (fourth step) can use some or all the data to generate classifications, and finally, the process ends when the results are evaluated (fifth step) using statistical measures.

2.5 FEATURES EXTRACTION AND DIMENSIONALITY REDUCTION

Reducing the number of features is an important point in machine learning. The dimensionality of the datasets used to perform data mining algorithms has increased a lot since data can be acquired automatically, but not useful or correlated features can lead to overfitting when we deal with a large amount of data [114]. That is it is necessary to use features extraction and dimensionality reduction algorithms.

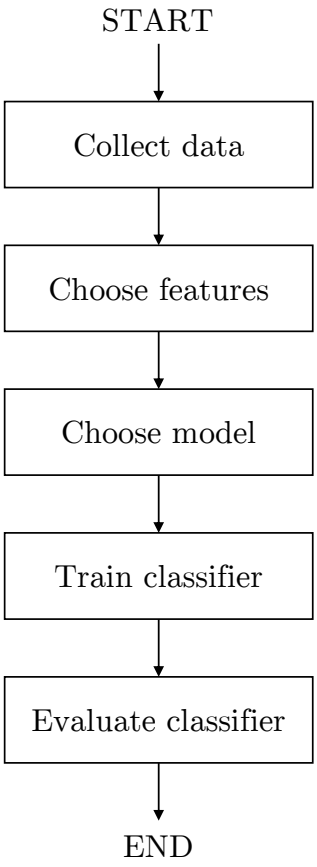


Figure 2.5: General cycle for a supervised learning task. Adapted from Duda et al. (2000) [92].

2.5.1 Principal component analysis

Principal Component Analysis (PCA) can be characterized as an unsupervised algorithm because it disregards class labels. Its main purpose is to decrease the dimensionality of a dataset constituted of numerous inter-connected variables [115]. Meanwhile, the algorithm has to retain as much as possible of the variation already existing in the dataset. This is accomplished by replacing the N original variables by a smaller number, M, of uncorrelated variables, called the principal components (PCs). These principal components are linear combinations of the initial variables, and they are ordered so that the first few retain the majority of the variation present in all of the original variables [115]. That is to say the

first principal component possesses the largest possible variance, and each succeeding one has the highest possible variance under the constraint that it is uncorrelated with (i.e., orthogonal to) the preceding components. The orthogonality principle is defined by the fact that the principal components are the eigenvectors of the covariance matrix. To perform this algorithm, the covariance matrix Σ of a given dataset $\mathbf{f} = [\mathbf{f}_1, \ldots, \mathbf{f}_n]$ is transformed into a diagonal matrix M [116] such that:

$$M = U^T \Sigma U \tag{2.1}$$

where M is the diagonalized matrix and U is the transformation allowing such a diagonalization.

The principal components are then found computing

$$Z = U^T [\mathbf{f}_i - \overline{\mathbf{f}}_i] \tag{2.2}$$

where $\overline{\mathbf{f}}_i$ is the mean of $\mathbf{f}_i$.

2.5.2 Linear discriminant analysis

In contrast to PCA, Linear Discriminant Analysis [92, 94] is a supervised algorithm that computes the directions (the linear discriminants) that will represent the axes maximizing the separation between multiple classes. Although one would find obvious that LDA could provide better results than PCA for a multi-class classification task with known class labels (i.e., a supervised learning task), it is not always the case. For example, comparisons between classification results for image recognition after using PCA or LDA highlighted that PCA tends to surpass LDA if the number of samples per class remains relatively small [117].

LDA is a well-known technique for dimension reduction and feature extraction, which has been widely used in applications such as image retrieval [118], face recognition [119], automatic speech recognition tasks [120, 121], or microarray data classification [122].

Classical LDA makes a projection of the data onto a lower-dimensional vector space in order to maximize the ratio of the between-class distance to the within-class distance and achieve maximum differentiation [123].

LDA has received a lot of extensions. One of them, called *PCA+LDA*, is particularly encountered in face recognition problems [119]. It generally applies an intermediate dimension reduction stage based on principal component analysis before before computing a linear discriminant analysis [123].

2.5.3 Feature selection

The purpose of feature selection is firstly to preserve the relevant [114] features and to get rid of irrelevant and redundant [114] features. [124] presents a definition of a *relevant* feature:

Definition 2.5.1 (Relevant Feature)**.** The feature $\mathbf{X_k}$ is relevant iff there exists some x_k and y_k such that

$$P(\mathbf{Y} = y_k | \mathbf{X_k} = x_k) \neq P(\mathbf{Y} = y_k)$$

Under this definition, $\mathbf{X_k}$ is relevant if knowing its value can change the prediction for the class $\mathbf{Y}$, that is to say if $\mathbf{Y}$ is dependent on $\mathbf{X_k}$. On the other hand, an *irrelevant* feature alter the learning process, whereas a *redundant* one does not add any improvement to the prediction [125]. However, it is important to notice that a variable useless or irrelevant by itself can still improve the algorithm performance when grouped with other variables [126]. Finally, variable or feature selection can be potentially beneficial in many points. It can allow to facilitate the visualization and the understanding of the data [114], and can also allow to reduce computation times and storage needs. However, its main purpose often remains to reduce the dataset's dimensionality in order to improve prediction performance [126].

2.6 CROSS VALIDATION

For the design of a practical classifier, we very often have to deal with only one dataset to perform the study. Thus, we use cross-validation to partition the initial dataset into two sub-dataset [127]. The first part, the training sample, is used to train the algorithm and build the classifier. On the other hand, the validation sample, which represents the remaining part of the data, is used to test and validate the classifier. The two main techniques used to perform cross-validation are the *leave-one-out* cross validation [128] and the *k-fold* cross-validation [129].

For the leave-one-out cross validation method, the training set is made up with the whole dataset except the first observation, and the validation set is the remaining record. The classifier is trained and tested using these datasets, and then the validation set is chosen to be the second observation. The training set is once again created with the entire dataset except this observation. This process is repeated until every record is used as a training set [129, 130, 131]. The procedure is illustrated in Figure 2.6. This method often provides the best results compared to other cross-validation methods, but it is computationally expensive since the classifier is trained and tested a number of times equals to the number of available observations [127].

The second approach, called k-fold cross-validation, randomly splits the dataset in k partitions of relatively equal sizes. k-1 partitions are used to create and train the model, and the remaining partition is used to validate the model. This process is repeated k times until the k partitions are used as a validation set [98, 127]. It can be noticed that when $k = n$, this method is rigorously the same as the leave-one-out technique. k-fold cross-validation is more computationally efficient because the classifier is only trained and tested k times. The variance of this algorithm is stable and is not impacted by the choice of k [98, 132]. Besides, 10-fold cross validation is frequently adopted [133].

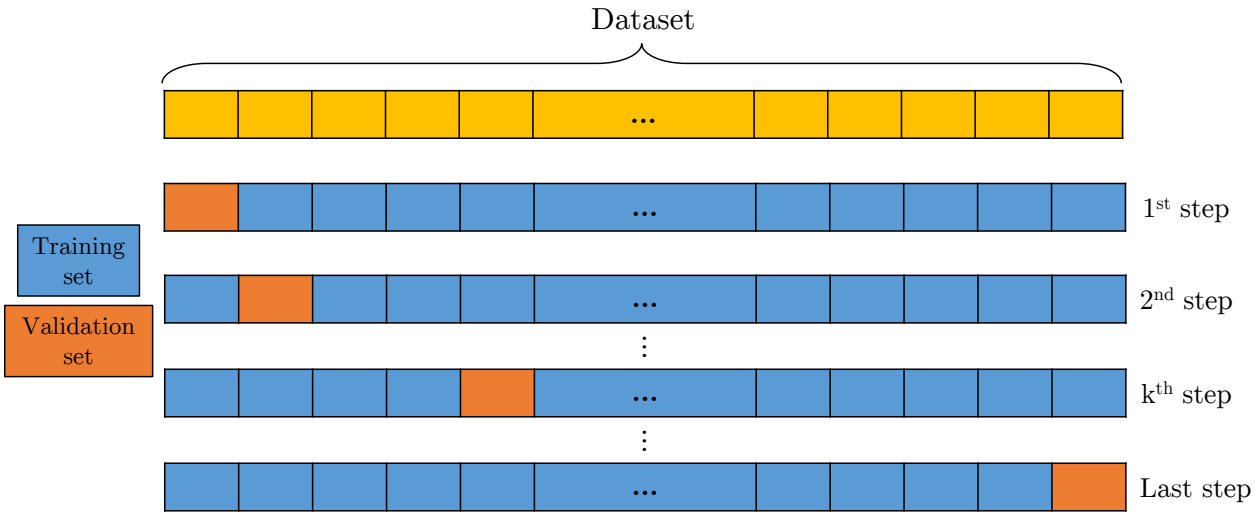


Figure 2.6: Illustration of the leave-one-out cross validation algorithm.

3.0 METHODOLOGY

3.1 DATA COLLECTION

The data were collected during the University of Pittsburgh Quit Study from 1990 to 1995. The purpose of this study, approved by the University of Pittsburgh Institutional Review Board, was to improve the comprehension of the factors causing lapses during smoking abstinence. 349 smokers seeking to quit smoking were recruited through a media advertising. Participants were firstly asked to stop smoking 17 days after the recruitment date, and they were asked to not change their smoking behavior during the first 17 days they were enrolled in the study. They were provided with one hour smoking cessation therapy session on day 1, 3, 8, 15, 17, 22, 29, 36 and 43 of the study.

In this study, smokers who had recently quit were asked to watch carefully their cigarette craving, nicotine withdrawal symptoms, mood, and activities over several weeks, using electronic diaries like the one presented in Figure 2.3. Since smoking episodes (lapses) represent our area of interest, subjects were asked to record any episodes of smoking as they happened. Then they were incited to answer brief questions about their craving, mood, and activities during the designated episode of smoking. Additionally, participants had to complete similar assessments at random times, five times a day for up to six weeks. The mobile device also offered questions evaluating the experience of the patient 24 hours after he or she confirmed smoking abstinence, and if the subject reported being in a highly tempting event or if he or she smoked, the device dispensed the appropriate protocol of contextual questions about participant's mood, location, and other environmental parameters.

The initial dataset consists in 69,143 reports from 248 unique subjects [34, 134]. Based on expert's specifications, a part of the dataset had to be discarded and the dataset used in this study finally consists in 41 parameters (features) and 29,959 environment reports from 248 unique subjects, organized as in Table 3.1. The original features are either nominal or numerical (discrete and continuous).

Table 3.1: Dataset used for the machine learning tasks. The real table has 29,959 rows and 41 columns.

Subject	**Feature #1**	**Feature #2**	...	**Feature #n**	**URGE (class)**
A	f_{A1}	f_{A2}	...	f_{An}	0 or 1
B	f_{B1}	f_{B2}	...	f_{Bn}	0 or 1
C	f_{C1}	f_{C2}	...	f_{Cn}	0 or 1
...	...	...	...	...	0 or 1

The smoking urges are evaluated according to the value of one discrete parameter, which represents the urge rating at any point in time. This variable was originally measured on a scale from 0 to 10: 0 is for the lowest smoking urge, while 10 is for an intense smoking urge. In order to simplify the classification and deal with a binary problem, the urge rating variable has been converted to a binary number such that:

$$\begin{cases} \text{URGE} = 0 \quad \text{if URGE} < 5 \quad \rightarrow \text{Negative cases} \\ \text{URGE} = 1 \quad \text{if URGE} \geqslant 5 \quad \rightarrow \text{Positive cases} \end{cases} \tag{3.1}$$

and the remaining nominal features have been transformed into numerical values.

Out of these 29,959 reports, 70% of them are attributed to class 0, while 30% of the reports are class 1. Our objective is to identify features associated with high urges (i.e., urge rating greater than or equal to 5).

3.2 CLASSIFICATION OF SMOKING URGES

Our unique dataset can be represented by the following form:

$$(\mathbf{F}, \mathbf{C}) = (\mathbf{F_1}, \mathbf{F_2}, \mathbf{F_3}, \ldots, \mathbf{F_k}, \ldots, \mathbf{F_N}, \mathbf{C}) \tag{3.2}$$

where $\mathbf{C}$ (binary column vector) is the target variable (the class) of the classifications (the urge rating). The matrix $(\mathbf{F_1}, \mathbf{F_2}, \mathbf{F_3}, \ldots, \mathbf{F_k}, \ldots, \mathbf{F_N})$ represents the features used by the classifier (each $\mathbf{F_k}$ is a column of $\mathbf{F}$). $\mathbf{F}$ can also be represented using a subject approach, $(\mathbf{f_1}, \mathbf{f_2}, \mathbf{f_3}, \ldots, \mathbf{f_k}, \ldots, \mathbf{f_N})^T$ where each $\mathbf{f_k}$ is a row of $\mathbf{F}$. Therefore, we have:

$$\mathbf{F} = \begin{pmatrix} f_{A1} & f_{A2} & \cdots & f_{Ak} & \ldots & f_{AN} \\ f_{B1} & f_{B2} & \cdots & f_{Bk} & \ldots & f_{BN} \\ f_{C1} & f_{C2} & \cdots & f_{Ck} & \ldots & f_{CN} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{pmatrix} = (\mathbf{F_1}, \mathbf{F_2}, \mathbf{F_3}, \ldots, \mathbf{F_k}, \ldots, \mathbf{F_N}) = \begin{pmatrix} \mathbf{f_A} \\ \mathbf{f_B} \\ \mathbf{f_C} \\ \vdots \end{pmatrix} \tag{3.3}$$

and

$$\mathbf{C} = \begin{pmatrix} c_A \\ c_B \\ c_C \\ \vdots \end{pmatrix}, \text{with } c_{k=A,B,\ldots} \in \{c_1 = 0, c_2 = 1\} \tag{3.4}$$

3.2.1 The naive Bayes classifier

The Bayesian decision theory is a statistical view of classification, based on the hypothesis that the problem is presented in probabilistic terms, and that all the relevant parameters are known [135, 92]. As in Equations 3.3 and 3.4, let define the kth observation vector $\mathbf{f_k}$ (the kth row in $\mathbf{F}$) and the set of classes $\{c_1, c_2\}$. One approach to perform the classification is to consider the posterior probability $P(c_i|\mathbf{f_k})$, where c_i represents the ith class. This method is

based on the Bayes rule [136, 137], which allows to compute the posterior probability given the prior probability $P(c_i)$ and the conditional probability $P(\mathbf{f_k}|c_i)$ such that

$$P(c_i|\mathbf{F} = \mathbf{f_k}) = \frac{P(\mathbf{F} = \mathbf{f_k}|c_i)P(c_i)}{P(\mathbf{F} = \mathbf{f_k})} \tag{3.5}$$

where

$$P(\mathbf{F} = \mathbf{f_k}) = \sum_{j=1}^{2} P(\mathbf{F} = \mathbf{f_k}|c_j)P(\mathbf{C} = c_j) \tag{3.6}$$

The notation of equation 3.5 may be simplified as:

$$P(c_i|\mathbf{f_k}) = P(c_i) \times \frac{P(\mathbf{f_k}|c_i)}{\sum_{j=1}^{2} P(\mathbf{f_k}|c_j)P(c_j)} \tag{3.7}$$

or also reformulated in English as:

$$P(c_i|\mathbf{f_k}) = \text{posterior} = \frac{\text{prior}(c_i) \times \text{likelihood}}{\text{prior}(\mathbf{f_k})} \tag{3.8}$$

$P(c_i)$ represents the *prior* probability of c_i, and describes how good is the hypothesis c_i, regardless of any data [138]. $P(\mathbf{f_k}|c_i)$ is the *likelihood* probability, or the probability of the observed data, given a particular hypothesis [138], and provides details about how well the hypothesis describes the data. The denominator $P(\mathbf{F} = \mathbf{f_k})$, also called *evidence* or *prior* probability of $\mathbf{f_k}$, acts as a normalizing factor [138]. Equations 3.7 and 3.8 can also be explained by a graph structure like the one presented in Figure 3.1. The class variable represents the root of the network, and each attribute has this class variable as a unique parent [139]. It can be noticed that the class is hidden for unsupervised learning problems.

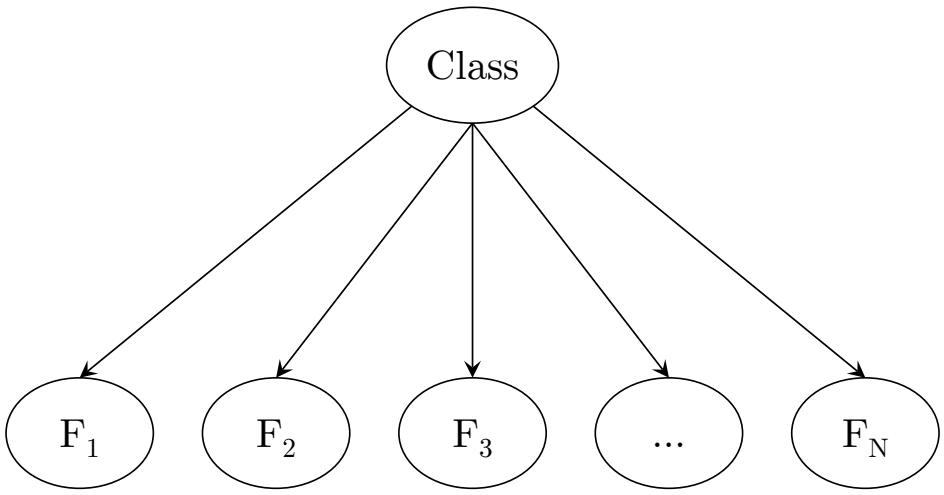


Figure 3.1: Structure of a naive Bayes Network. $F_1, \ldots, F_N$ symbolize the feature vectors. Adapted from Friedman et al. (1997) [139].

In order to compute the posterior probability $P(c_i|\mathbf{f_k})$, the conditional probability $P(\mathbf{f_k}|c_i)$ has to be known. Two cases are considered and are appropriate to most of the applications [92, 93]:

- The multivariate normal law assumption, where the conditional distribution of the ith class (i is equal to 0 or 1) is assumed to follow a multivariate Gaussian distribution with mean vector $\boldsymbol{\mu_i}$ and covariance matrix $\boldsymbol{\Sigma}_i$, of dimension $N \times N$, i.e., $(\mathbf{f_k}|c_i) \sim \mathcal{N}(\boldsymbol{\mu_i}, \boldsymbol{\Sigma}_i)$. Besides, this means the features are assumed to be continuous. The probability density function is then defined as [140]:

$$f(\mathbf{f_k}|c_i) = \frac{1}{\sqrt{(2\pi)^N |\boldsymbol{\Sigma}_i|}} \exp\left(-\frac{1}{2}(\mathbf{f_k} - \boldsymbol{\mu_i})^T \boldsymbol{\Sigma}_i^{-1} (\mathbf{f_k} - \boldsymbol{\mu_i})\right) \tag{3.9}$$

where $|\boldsymbol{\Sigma}_i|$ is the determinant of $\boldsymbol{\Sigma}_i$.

- The multivariate multinomial assumption, where the distribution $(\mathbf{f_k}|c_i)$ is assumed to follow a multinomial multivariate distribution. In this case the features are assumed to be discrete. The probability mass function is defined as [140]:

$$f(\mathbf{f_k}|c_i) = \frac{\left(\sum_{d=1}^{n} k_d\right)!}{\prod_{d=1}^{n} k_d!} \prod_{d=1}^{n} c_{id}^{k_d} \tag{3.10}$$

where $\mathbf{f_k} = [f_1, \ldots, f_N]$ is a feature vector and $f_1, \ldots, f_N$ represent the different feature values. If the dth feature is discrete, $f_d \in \{\phi_1, \ldots, \phi_n\}$ where $\phi_1, \ldots, \phi_n$ are the different values that f_d can take, and k_d is the frequency count of the value ϕ_d. The parameter vector for the class c_i is given as $\mathbf{c}_i = [c_{i1}, \ldots, c_{in}]$ where c_{id} is the probability that the discrete event d occurs in the ith class.

3.2.2 Discriminant analysis classifier

Similar to the naive Bayes classification method, the discriminant analysis assumes that the conditional probability density functions $P(\mathbf{f_k}|c_1 = 0)$ and $P(\mathbf{f_k}|c_2 = 1)$ are both normally distributed [141], with mean and covariance parameters $(\boldsymbol{\mu_1}, \boldsymbol{\Sigma_1})$ and $(\boldsymbol{\mu_2}, \boldsymbol{\Sigma_2})$. In other words, the fitting function, which is used to generate the classifier, estimates the parameters of a Gaussian distribution for each class.

More particularly, the Linear Discriminant analysis (LDA), also known as the Fisher Linear Discriminant (FLD), is a classifier that involves the projection of the data onto a line, and then perform the classification on this reduced one-dimensional space. The projection must be chosen in a way that maximizes the distance between the means of the classes, while minimizing the variance within each projected classes [142, 143]. Figure 3.2 illustrates the concept of an optimal class separation when a LDA is used.

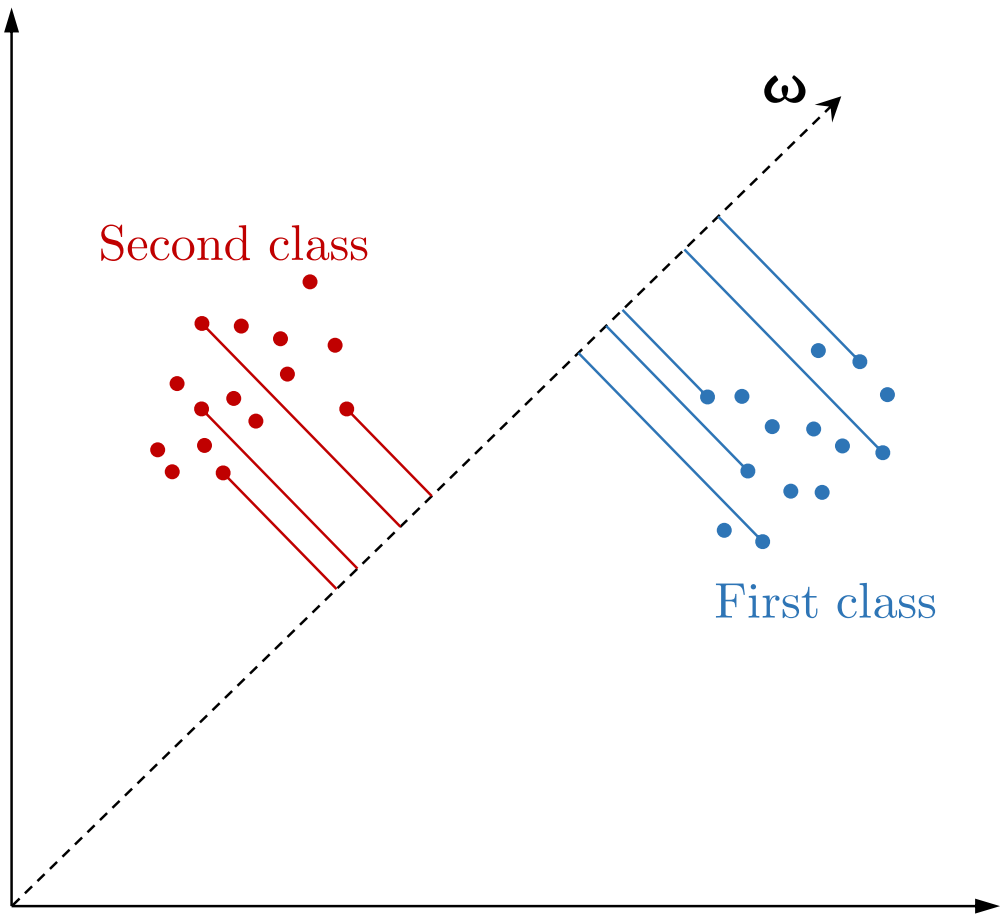


Figure 3.2: Two-dimensional illustration of a linear discriminant analysis approach: the direction w maximizes the within class and between class scatter.

Let consider the binary class classification case, with $\mathbf{f} = [\mathbf{f_1} \ldots, \mathbf{f_n}]$ a set of n multidimensional samples. $n = n_1$ for the first class c_1 and $n = n_2$ for the second class c_2. The sample mean μ_i for the ith class is given by [92]:

$$\boldsymbol{\mu_i} = \frac{1}{n_i} \sum_{\mathbf{f}_i \in c_i} \mathbf{f}_i \tag{3.11}$$

The direction $\boldsymbol{\omega}$ maximizing the space between the different $\boldsymbol{\mu_i}$ is the $\boldsymbol{\omega}$ that maximizes the function

$$J(\boldsymbol{\omega}) = \frac{\boldsymbol{\omega}^T \mathbf{S}_B \boldsymbol{\omega}}{\boldsymbol{\omega}^T \mathbf{S}_W \boldsymbol{\omega}} \tag{3.12}$$

where $\mathbf{S}_W$ is the within-class scatter matrix [142, 94] defined by

$$\mathbf{S}_W = \mathbf{S}_1 + \mathbf{S}_2, \text{with } \mathbf{S}_i = \left\{ \sum_{\mathbf{f}_k \in c_k} (\mathbf{f}_k - \boldsymbol{\mu}_i)(\mathbf{f}_k - \boldsymbol{\mu}_i)^T, i \in \{1, 2\} \right\} \tag{3.13}$$

and $\mathbf{S}_B$ is the between-class scatter matrix, defined as

$$\mathbf{S}_B = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \tag{3.14}$$

The Fisher linear discriminant, designated as the linear function maximizing the between-class scatter within-class scatter ratio is given by:

$$\boldsymbol{\omega} = \mathbf{S}_W^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \tag{3.15}$$

Then, if we assume that the samples follow multivariate Gaussian distributions with equal covariance matrix $\boldsymbol{\Sigma}$, the classification threshold can be determined and the classifier can be built. Finally, the multivariate multinomial estimate can be done using the following approaches:

- A *linear discriminant*, where the decision surface is an hyperplane and the covariance is estimated using the pooled approach.
- A *quadratic discriminant*, where the the decision surface is a paraboloid and the multivariate multinomial laws estimates are stratified by class.
- A *Mahalanobis distance discriminant*, where the Mahalanobis distance is used to estimate the covariance matrix.

3.2.3 Decision tree learning

A decision tree is a structure used to predict the response (the class) to inputs. This method, regularly used in machine learning theory [144], is based on the construction of a binary tree where the nodes represent the tests made on the inputs. The results of the different tests give the direction to follow in the tree. Finally, the prediction can be read when a leaf node is reached [145]. In this thesis we adopted classification trees, which provide binary (nominal) classification such as *true* (1) or *false* (0). These trees are very useful because they can provide easy to understand predictions in difficult conditions when many variables are present. Figure 3.3 shows an example of a classification tree with two classes ($C = 1$ or 2) and two features $\mathbf{F}_i$ ($i = 1, 2$). As can be seen, an advantage of the tree structure is its capability to deal with an important number of input variables, whereas a plot is limited to two (2 dimensions) or three (3 dimensions) input variables [146]. Further techniques, called *ensemble* methods [147], can also build more than one decision tree.

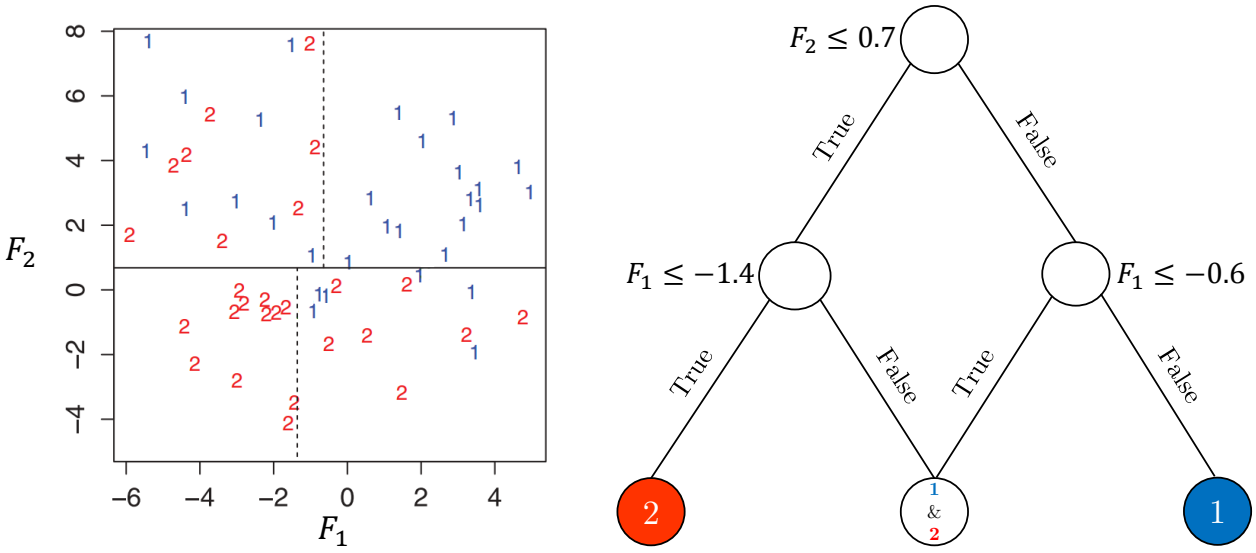


Figure 3.3: Partitions (left) and decision tree structure (right) for a classification tree model with two classes (1 and 2). Adapted from Loh (2011) [146].

3.3 VALIDATION METHODS

In order to evaluate and compare the efficiency of the classifiers, several parameters are defined. Based on the outcome of a binary classification test, we compute the four following fundamental parameters [148]:

- True positives (**TP**): number of smoking urges correctly identified as smoking urges.
- True negatives (**TN**): number of lacks of smoking urges correctly identified as lacks of smoking urges.
- False positives (**FP**): number of lacks of smoking urges incorrectly identified as smoking urges.
- False negatives (**FN**): number of smoking urges incorrectly identified as lacks of smoking urges.

  They can often be visualized in a *confusion matrix* [148] (Table 3.2).

Table 3.2: Confusion matrix

		CLASSIFIED CONDITION	
		Positive	**Negative**
ACTUAL CONDITION	**Positive**	*True Positive*	*False Negative*
	Negative	*False Positive*	*True Negative*

Each column of the matrix represents the instances in a classified class, while each row represents the instances in an actual class. The diagonal of the confusion matrix represents the correct classifications whereas the results out of the diagonal are the misclassifications.

Then, these four parameters are used to compute the following metrics [148]:

- the **True Positive Rate** (TPR) (or **Sensitivity**):

$$TPR = \frac{TP}{TP + FN} \tag{3.16}$$

  It describes the test's capability to identify a class correctly, and measures the proportion of positive subjects correctly identified as such.

- the **True Negative Rate** (TNR) (or **Specificity**):

$$TNR = \frac{TN}{FP + TN} \tag{3.17}$$

It describes the test's capability to eliminate a class correctly, and measures the proportion of negative subjects correctly identified as such.

- the **Accuracy** (ACC):

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.18}$$

It represents the proportion of true results, both TP and TN, in the whole population. The accuracy is also closely related to the misclassification rate ρ, which can be defined as:

$$\rho = 1 - ACC = 1 - \frac{TP + TN}{TP + TN + FP + FN} = \frac{FP + FN}{TP + TN + FP + FN} \tag{3.19}$$

- the **Positive Prediction Value** (PPV) (or **Precision**):

$$PPV = \frac{TP}{TP + FP} \tag{3.20}$$

It represents the proportion of true positives among all the positive results (both TP and FP).

These four metrics will be used to compare the outcomes of three classification analyses:

- The first analysis is based on the whole dataset in order to understand the accuracies of the algorithms when the entire dataset is utilized.
- The second analysis is based on eight clinically relevant and potentially actionable features. These specific features have been shown in previous studies to be related to the probability of lapse and/or relapse in a smoking cessation attempt [89, 90]. These features and their evaluation scales are presented in Table 3.3.
- The third and last classification is based on the selected features determined by the feature selection algorithm.

Table 3.3: Previously-identified features

Previously-identified features	**Evaluation scale**
Day of week	1 to 7, with 1 = Monday
Availability of cigarettes	Yes or No
Alcohol consumption	Yes or No
Confidence in ability to resist smoking	NO!!, no??, yes?? or YES!!
Location of the subject	Home, Work-place, Other's home, Bar/Restaurant, Vehicle, Outside, Other
Subject's mood	Very bad, Bad, Neutral, Good, Very good
Presence of people smoking near the subject	Yes or No
Is the weekday a weekend day?	Yes or No

3.4 FEATURE SELECTION

The feature selection was realized using a sequential feature selection algorithm. Its purpose is to optimize a criterion (e.g., classification accuracy, misclassification rate, specificity,...) related to the classification problem. The sequential feature selection uses three distinct sub-algorithms:

1. the search algorithm, that looks for the feature subset optimizing the criterion.
2. the evaluation algorithm, which evaluates the chosen criterion.
3. the performance function algorithm, which is in our case one of the three classifiers used in this study (i.e., naive Bayes, discriminant analysis, or classification tree).

In this thesis, the chosen optimization criterion is the misclassification rate given in Equation 3.19. The search algorithm starts with none included features, and then evaluates the chosen criterion for all the feature subsets containing a single feature and retains the subset resulting in the minimal misclassification rate. The criterion is evaluated using the leave-one-out cross validation method. The size of the tested subset is then incremented

and the same process is repeated on these subsets. The algorithm stops when there is no improvement in the classification, or when the size of the features' subset reaches a desired threshold. Figure 3.4 presents a flowchart implementation of the feature selection algorithm.

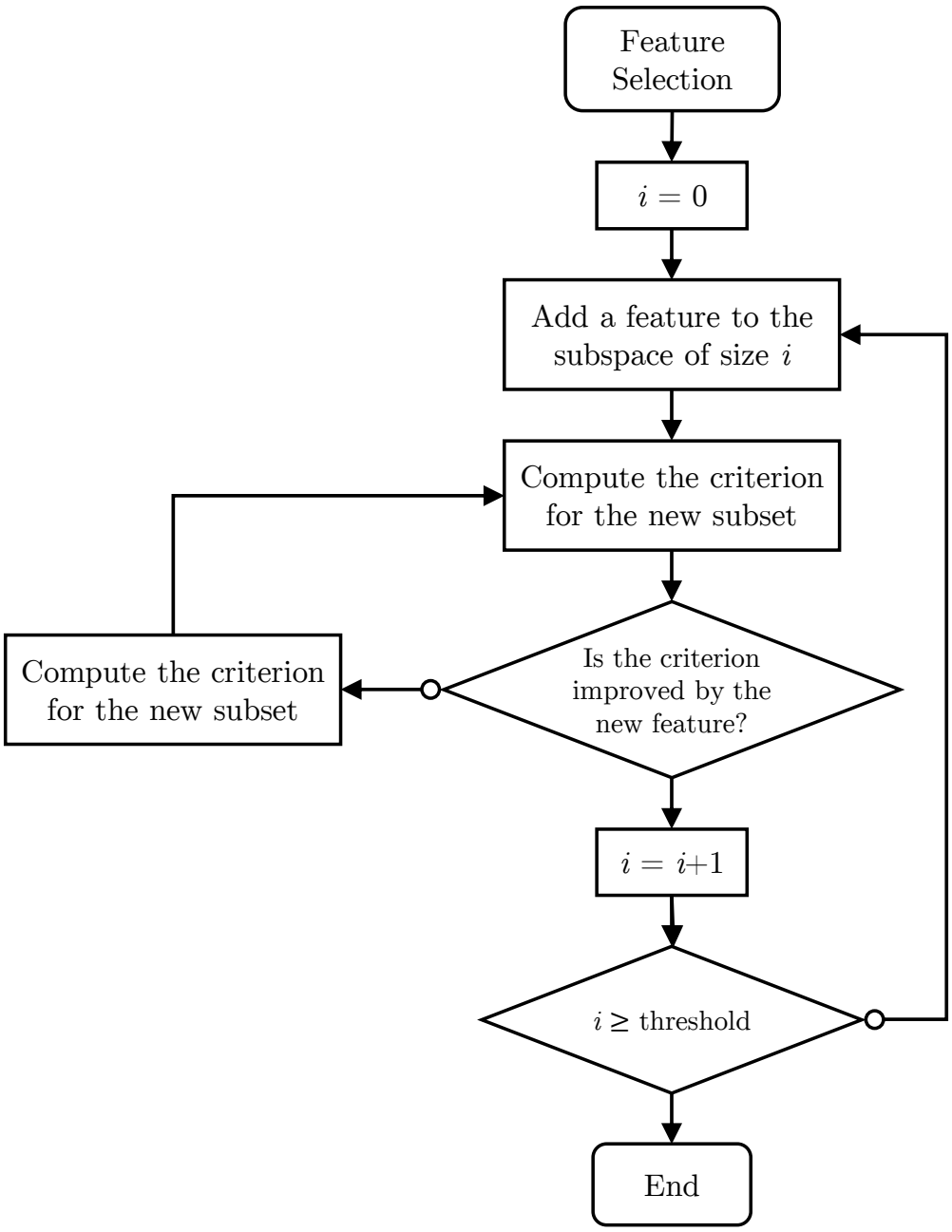


Figure 3.4: Feature selection flowchart structure.

We also use 10-fold cross-validation to test our algorithms [127] (Figure 3.5), since a leave-one-out cross-validation algorithm, which is in general more efficient, would take more than one year to compute according to the number of observations.

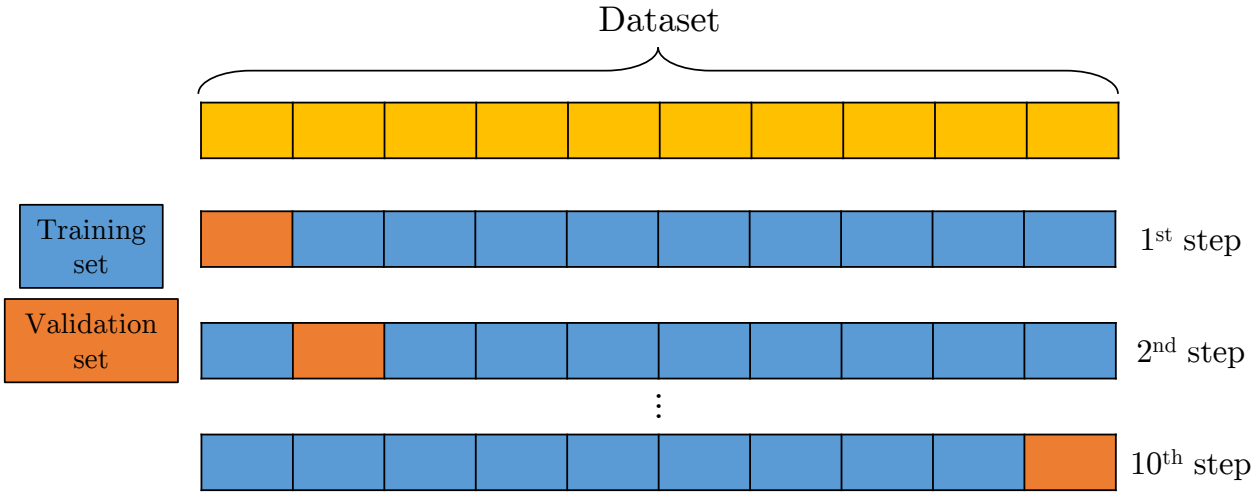


Figure 3.5: 10-fold cross-validation algorithm.

4.0 RESULTS

4.1 CLASSIFICATION OF SMOKING URGES

Table 4.1 presents the different selected features that resulted from the feature selection algorithm. Respectively four, eleven, and four features were selected for the naive Bayes, discriminant analysis, and classification tree algorithm. One feature, the day of week, appears in both the previously-identified and the classification tree selected features. Two features, the day of the study and the tense level (Does the subject feel tense?), have been selected with the three classification methods. Two others, the energy or arousal level (Does the subject feel energetic?) and the restlessness level (Does the subject feel restless?), have been selected with two classification methods out of the three studied in this thesis.

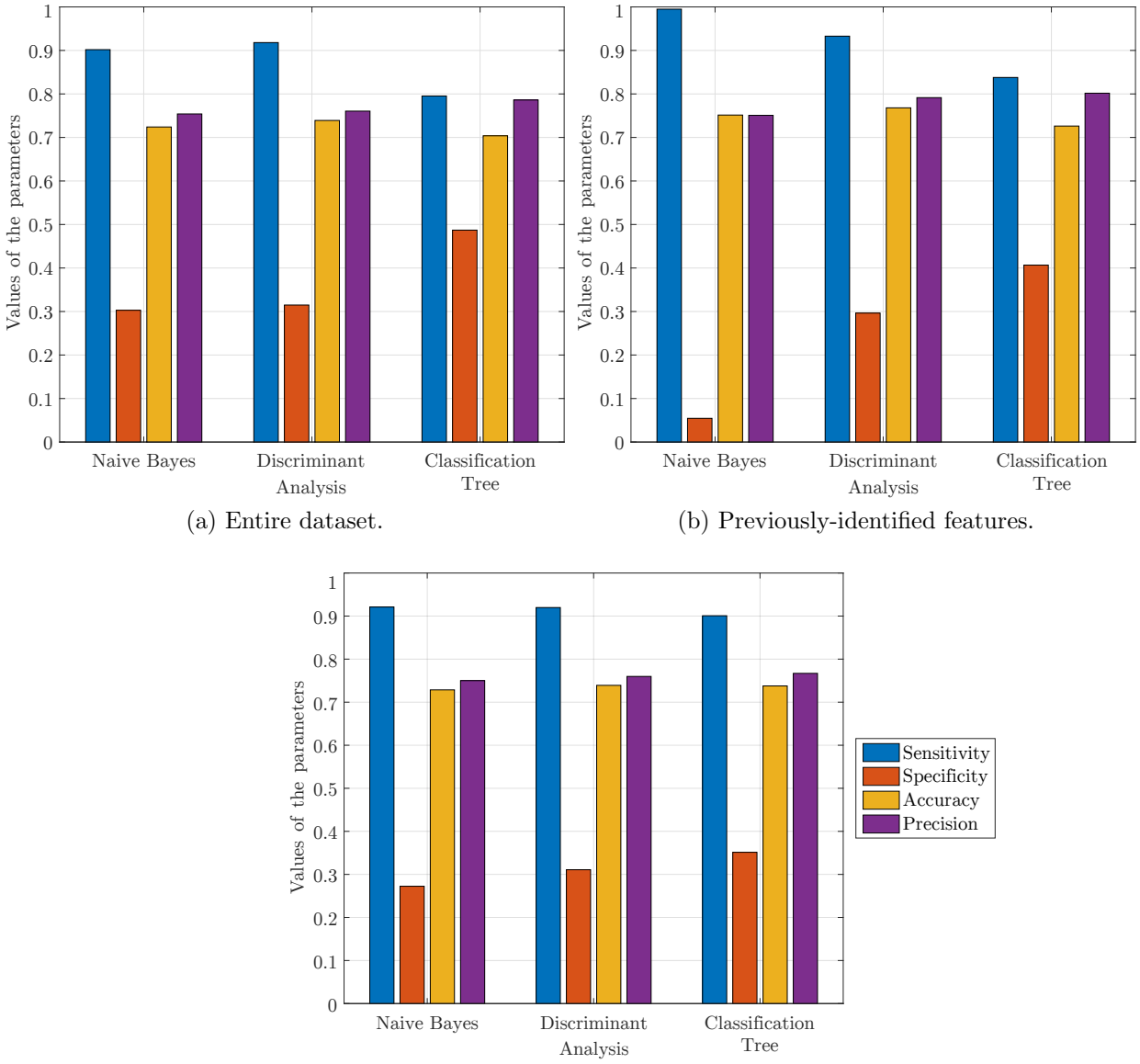
Figure 4.1 shows the classification results for the three different methods, based on the entire input dataset, the previously identified features, and the selected features. When the entire dataset is used (Figure 4.1a), the classification tree model has an average accuracy of 69.3% for the four metrics, but the sensitivity appears to be less than 80%. For the naive Bayes and discriminant analysis classification methods, the average accuracy is respectively 67% and 68.3% (Table 4.2).

Using respectively the previously-identified features (Table 3.3, Figure 4.1b) and then a selected subset (Table 4.1, Figure 4.1c), the metrics' mean is 63.8% and 66.8% for the naive Bayes classifier, 69.7% and 68.2% for the discriminant analysis classifier, and 69.3% and 68.9% for the classification tree classifier. All these results are summarized in Table 4.2.

Besides, as can be seen on Figure 4.1, none of the feature selection algorithms succeeded in providing both higher sensitivity and higher specificity, compared to the case where the whole dataset was used.

Table 4.1: Selected features

Selected Features		
Naive Bayes	**Discriminant Analysis**	**Classification Tree**
Day of the study	Day of the study	Day of the study
Feeling tense?	Feeling tense?	Feeling tense?
Feeling energetic?	Feeling energetic?	Day of week
Interacting with others?	Feeling restless?	Feeling restless?
	Is the subject alone?	
	What is the subject's arousal level?	
	Drinking coffee?	
	Feeling contented?	
	Inactive?	
	Feeling miserable?	
	Feeling sad?	



(a) Entire dataset.

(b) Previously-identified features.

(c) Selected features.

Figure 4.1: Comparison of three classification methods with different datasets.

Moreover, it is important to note that the results obtained with the feature selection algorithm are dependent on the selected features themselves, but are independent on the number of selected features. The purpose of the feature selection algorithm is to preserve the relevant features [114] and to get rid of irrelevant and redundant features [125]. That is why a variable useless or irrelevant by itself can still improve the algorithm performance when grouped with other variables [126].

Table 4.2: Average results (sensitivity, specificity, accuracy, and precision) of the three classification methods with different datasets.

	Average Results (%)		
	Entire dataset	**Prev. ident. features**	**Selected features**
Naive Bayes	67.1	63.8	66.8
Disc. Analysis	68.3	69.7	68.2
Classification Tree	69.3	69.3	68.9

The two first techniques (naive Bayes and discriminant analysis) present almost no significant differences when the feature selection algorithm is used (Figure 4.1c): with both methods, sensitivity is slightly higher but specificity is lower than in cases without feature selection (Figure 4.1a). For the classification tree model, sensitivity is 10.6% higher but specificity is 13.6% lower than in the classification tree without feature selection case, and they are respectively 6.3% higher and 5.5% lower than in the classification tree with the previously-identified features case. The previously-selected and the selected features seem to increase the sensitivity, at the cost of a decreased specificity. In other words, the test's capability to identify a class correctly is high, but this involves a lower capability to eliminate a class correctly. For example, the sensitivity in the naive Bayes with feature selection case is 2% higher than in the without feature selection case, but the specificity is 3% lower.

4.2 RELEVANCE OF THE FEATURE SELECTION ALGORITHM

To evaluate the relevance of the feature selection algorithm and to emphasize its efficiency, we display:

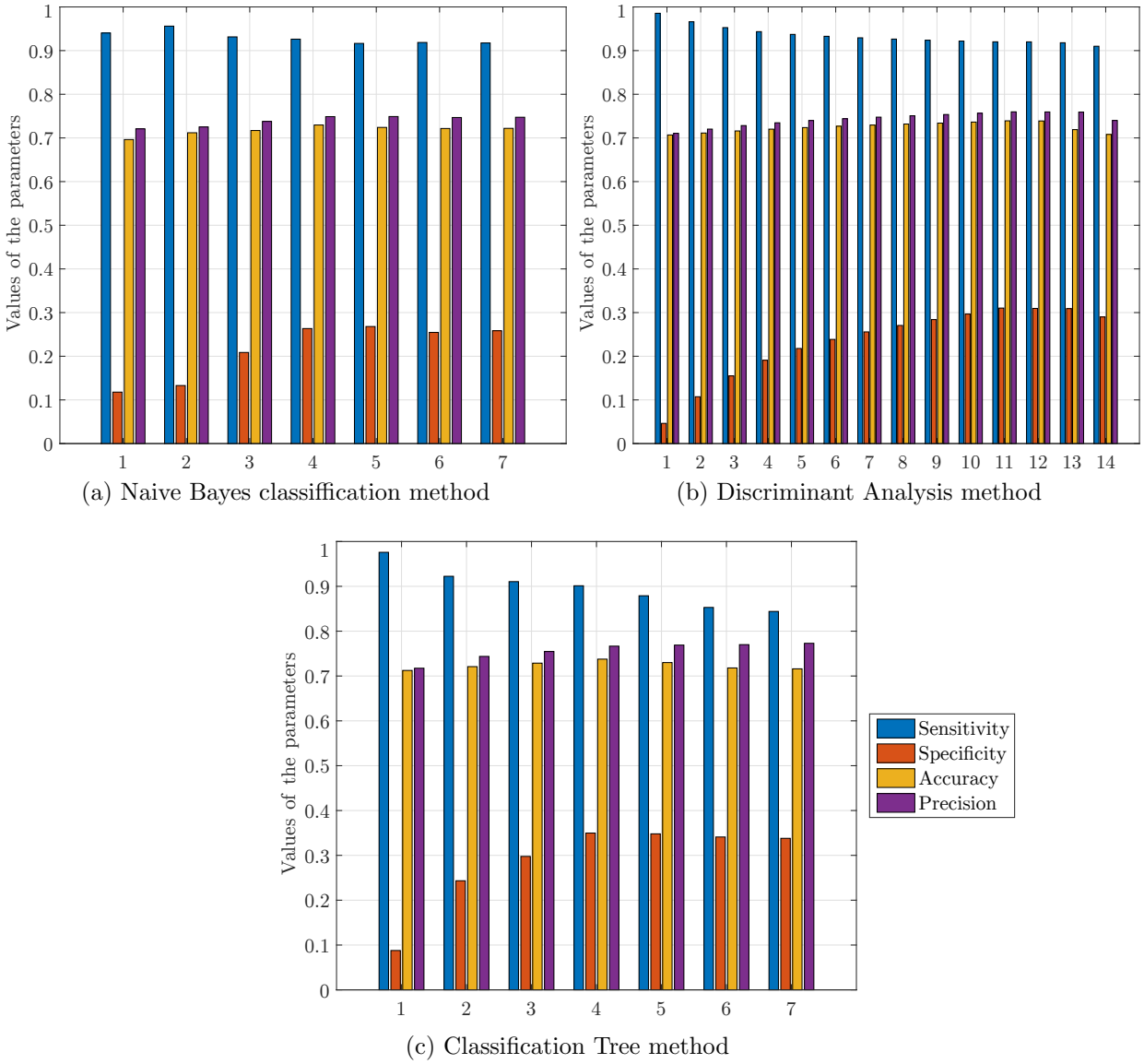
- the results when a different number of features is used for the classification (Figure 4.2);
- the evolution of the misclassification rates of the three classifiers versus a number of retained features (Figure 4.3).

On the first one (Figure 4.2), the number of features used for the classification appears on the X-axis. The three sub-figures have been computed as following:

- When the number of features is less than or equal to the number of selected features (four for the naive Bayes and classification tree classifiers, and eleven for the discriminant analysis classifier), the features are exclusively selected among the features presented in Table 4.1 and all the possible combinations of features are covered. There are 15 combinations for the naive Bayes and classification tree cases, and 2047 for the discriminant analysis method. In other words, each kth group of four bar plots presented in Figure 4.2 provides the results when k features are used for the classification. The displayed results represent the mean value for each possible combination of features.

- When the number of features is greater than the number of selected features (four for the naive Bayes and classification tree classifiers, and eleven for the discriminant analysis classifier), one, two and three randomly chosen features have been added 20 times to the dataset already made up with the selected features. In other words, the results have been obtained computing 20 permutations of five, six and seven features for the naive Bayes and classification tree classifiers, and 20 permutations of 12, 13 and 14 features for the discriminant analysis classifier.

Figure 4.2 shows that the sensitivity seems to be slightly decreasing while the number of features is increasing. That means the test's capability to identify a class correctly lowers with the number of features. However, it remains greater than 90% for the naive Bayes and discriminant analysis methods, and above 84% for the classification tree method. Negatively, the three other metrics (specificity, accuracy and precision) increase with the number of features when this one goes from one to the final number of selected features (i.e., four for the naive Bayes and classification tree classifiers, and 11 for the discriminant analysis classifier). The maximum values seem to be reached with the exact number of selected features, which is consistent with the purpose of the feature selection algorithm. Up to this



(a) Naive Bayes classiffication method

(b) Discriminant Analysis method

(c) Classification Tree method

Figure 4.2: Comparison of the number of features versus the performance of each classifier.

maximum, the evolution of accuracy and precision is hardly notable but the increase of the specificity is very significant: 15% for the naive Bayes method and 26% for the discriminant analysis and classification tree ones. Finally, specificity, accuracy and precision start to decrease again after the maximum was reached.

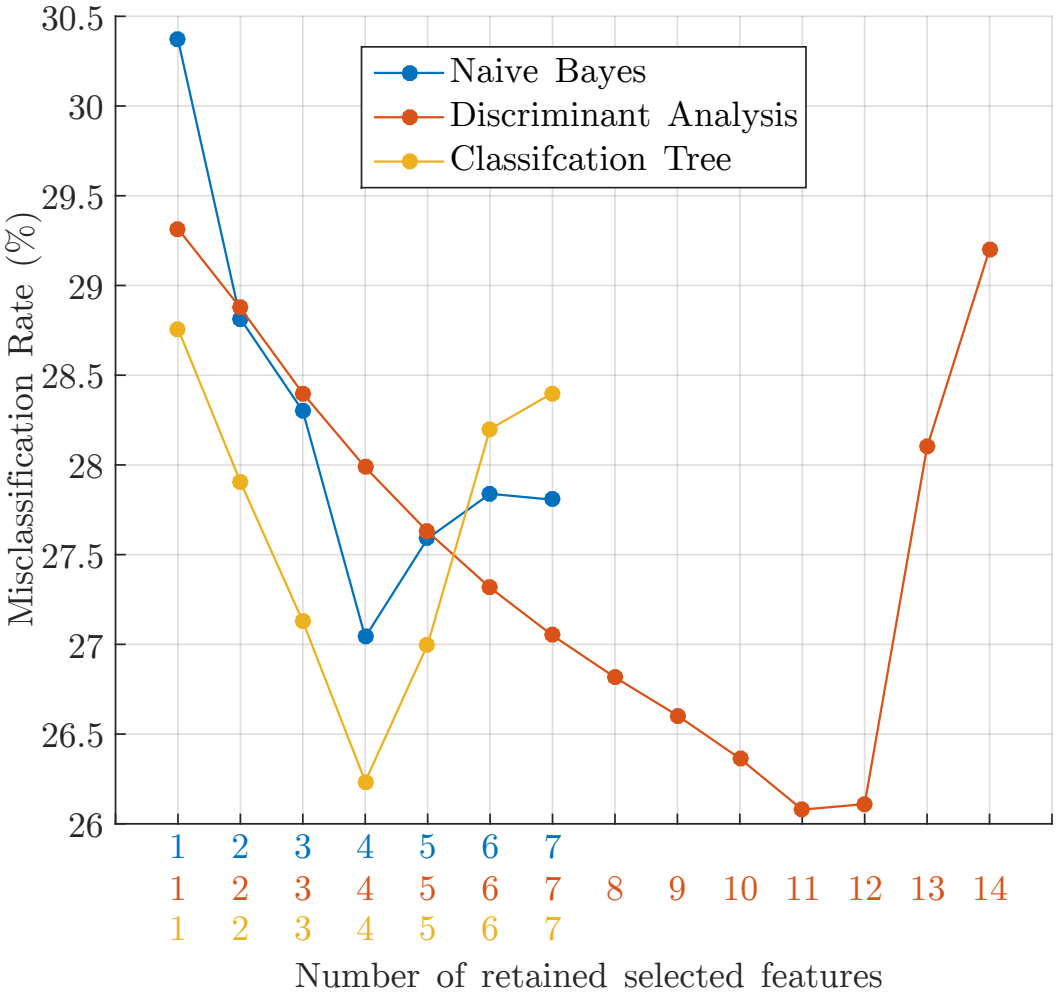


Figure 4.3: Urges to smoke misclassification rates versus the number of retained selected features.

Figure 4.3 displays the evolution of the misclassification rate of each classification method versus the number of selected features. The X-axis on Figure 4.3 has the same meaning as the one presented in Figure 4.2. Similarly to the results presented in Figure 4.2, the misclassification rates for each classifier are minimum when the exact number of selected features is used: four for the naive Bayes and classification tree classifiers, and 11 for the discriminant analysis. For all three classifiers, the misclassification rate is decreasing on the left side of the number of selected features, and it is increasing on the right side. The discriminant analysis is the method that reaches the lowest misclassification rate: 26.1%.

These results highlight the behavior of the feature selection algorithm. It is actually designed to determine the exact number of features that provide the best classification results. That is why the classification results decrease if some features are removed from or added to the selected dataset. Besides, among the different selected features presented in Table 4.1, it is not possible to identify one particular feature that is better than the remaining one, as these are fairly nonlinear feature selection process.

In addition to the improvement of the global prediction's accuracy that can be observed with the subset of features, a faster computation time has also been noticed. Algorithms with feature selection are approximately ten times faster than algorithms based on the entire dataset, and this aspect can be relevant in the case where some analysis has to be run on a less-powerful platform.

5.0 DISCUSSION

In this study, we considered three machine learning methods to classify situations with strong smoking urges. First of all, without using a feature selection algorithm, the presented results showed that the urge rating can be accurately classified into two states: a high smoking urge versus little or no urge. The naive Bayes method and discriminant analysis have the highest sensitivity ($\approx 90\%$), but the lowest specificity ($\approx 30\%$). On the contrary, the classification tree's results are more grouped together. This method provides, on average, the most accurate classifications when the entire dataset is used.

However, evaluating hundreds of parameters several times a day can be very time consuming, and lead to unnecessary rejection of potential therapies to be delivered via modern electronic devices such as smartphones or tablets. Therefore, it is beneficial to find a reduced set of features that can be used to classify the urge rating. The selected features appear to be dominated by measures of emotional state, especially in the discriminant analysis. It is known that negative emotional states are related to craving and lapsing [149, 26], but it is interesting to see that algorithms have extracted multiple (up to six for the discriminant analysis) indicators of emotional state. Furthermore, the meaning of this selection is that each feature makes an incremental contribution to the final classification since a feature is selected by the algorithm when its selection involves a decrease of the misclassification rate. Coffee consumption has also been associated with smoking and was selected in the discriminant analysis approach. It is interesting to notice that the feature related to drinking alcohol is not taken into consideration, as studies have shown that the effect of alcohol drinking on urges to smoke is stronger than coffee [26]. Finally, the feature that specifies the day of the study has probably been selected because craving decreases as abstinence progresses [27]. Regarding the one that specifies the week day ("Day of week"), it could differentiate

a smoking activity happening during week days from a smoking activity happening during weekends.

Feature selection algorithms have respectively selected four and eleven features for the classification tree method and the discriminant analysis. These two classifiers resulted in almost the same mean ($\approx$ 68%) for the four metrics evaluated in this study (sensitivity, specificity, accuracy and precision), and the first provides the highest specificity (35.1% against 31.1%) while the second has the highest sensitivity (92% against 90.1%) and the lowest misclassification rate (26.1%). These results also highlight that the algorithms are better at correctly classifying the true presence of an urge rather than the true absence of one. This is valuable for the classifications because it is preferable to inaccurately assume an urge is present (and risk unnecessary urge query) than miss a real urge (and risk not intervening to prevent smoking). Finally, according to these results, one of these two classifiers could be finally chosen to make the classifications.

6.0 CONCLUSIONS AND FUTURE WORK

6.1 CONCLUSIONS

In this thesis, we considered three different supervised machine learning algorithms to obtain classifications of smoking urges. Firstly, it has been shown that the three classifications methods can each classify urge ratings with reasonable sensitivity and specificity. In this thesis, three different classification methods have been considered for detecting smoking urges. First, it has been shown that the three classifiers can each classify urge ratings with reasonable sensitivity and specificity. Then, combined with a feature selection algorithm, the classifier based on a discriminant analysis method extracted eleven features, while methods based on naive Bayes classifier or classification tree chose four features. These feature selection algorithm enabled us to obtain a sensitivity and specificity of respectively 90% and 35%, showing that smoking urges can be accurately predicted with a reduced dataset.

6.2 FUTURE DIRECTIONS

In a future project, the selected features could be implemented in a personal mobile application in order to assist a subject in his/her smoking cessation process. The algorithm would be able to estimate his/her urge rating, which could allow the application to provide more adaptive and personalized support. According to the chosen method, as few as four have to be reported by a subject at each data collection to facilitate a highly sensitive classification of high-urge state. A real time data collection would also allow to compute the different algorithms with recent observations, updated by the patient as soon as he or she needs an urge classification.

However, there is an important clinical difference between knowing a high urge is likely to occur now (i.e., classification) and knowing a high urge is likely to occur next (i.e., prediction). While these analyses showed that urges can be correctly classified, the advanced machine learning algorithms adopted in this study could be reused, with new suitable time variables, in order to predict both smoking urges and time occurrences of these events. A possible approach to this problem would be to predict smoking urges and approximate time intervals when these urges occur.

BIBLIOGRAPHY

[1] World Health Organization, "WHO report on the global tobacco epidemic, 2011: warning about the dangers of tobacco: executive summary," 2011.

[2] R. Doll, R. Peto, J. Boreham, and I. Sutherland, "Mortality in relation to smoking: 50 years' observations on male british doctors," *British Medical Journal*, vol. 328, no. 7455, p. 1519, 2004.

[3] H. D. Mansvelder and D. S. McGehee, "Cellular and synaptic mechanisms of nicotine addiction," *Journal of Neurobiology*, vol. 53, no. 4, pp. 606–617, 2002.

[4] U. S. Department of Health and Human Services, "The health consequences of smoking: Nicotine addiction, a report of the surgeon general," *Washington, DC: US Government Printing Office*, 1988.

[5] Centers for Disease Control, "Cigarette smoking among adults–United States, 1990," *Morbidity and Mortality Weekly Report*, vol. 41, no. 20, p. 354, 1992.

[6] U.S. Department of Health and Human Services, "The health consequences of smoking-50 years of progress: A report of the surgeon general," *US Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health*, vol. 17, 2014.

[7] N. R. Anthonisen, M. A. Skeans, R. A. Wise, J. Manfreda, R. E. Kanner, and J. E. Connett, "The effects of a smoking cessation intervention on 14.5-year mortality: a randomized clinical trial," *Annals of Internal Medicine*, vol. 142, no. 4, pp. 233–239, 2005.

[8] A. K. Hackshaw, M. R. Law, and N. J. Wald, "The accumulated evidence on lung cancer and environmental tobacco smoke," *British Medical Journal*, vol. 315, no. 7114, pp. 980–988, 1997.

[9] J. Barnoya and S. A. Glantz, "Cardiovascular effects of secondhand smoke nearly as large as smoking," *Circulation*, vol. 111, no. 20, pp. 2684–2698, 2005.

[10] N. L. Benowitz, "Nicotine addiction," *The New England Journal of Medicine*, vol. 362, no. 24, p. 2295, 2010.

[11] X. Xu, E. E. Bishop, S. M. Kennedy, S. A. Simpson, and T. F. Pechacek, "Annual healthcare spending attributable to cigarette smoking: An update," *American Journal of Preventive Medicine*, 2014.

[12] I. T. Agaku, B. A. King, and S. R. Dube, "Current cigarette smoking among adults-United States, 20052012," *Morbidity and Mortality Weekly Report*, vol. 63, no. 2, pp. 29–34, 2014.

[13] A. Jamal, I. T. Agaku, E. O'Connor, B. A. King, J. B. Kenemer, and L. Neff, "Current cigarette smoking among adults-united states, 2005-2013," *Morbidity and Mortality Weekly Report*, vol. 63, no. 47, pp. 1108–1112, 2014.

[14] Centers for Disease Control and Prevention, "Quitting smoking among adults-United States, 2001-2010," *Morbidity and Mortality Weekly Report*, vol. 60, no. 44, p. 1513, 2011.

[15] S. Shiffman, D. M. Scharf, W. G. Shadel, C. J. Gwaltney, Q. Dang, S. M. Paton, and D. B. Clark, "Analyzing milestones in smoking cessation: illustration in a nicotine patch trial in adult smokers," *Journal of Consulting and Clinical Psychology*, vol. 74, no. 2, p. 276, 2006.

[16] Centers for Disease Contol and Prevention, "Cigarette smoking among adults and trends in smoking cessation-United States, 2008," *Morbidity and Mortality Weekly Report*, vol. 58, no. 44, p. 1227, 2009.

[17] I. of Medicine, "Ending the tobacco problem: A blueprint for the nation," 2007.

[18] M. Fiore, "Treating tobacco use and dependence; 2008 guideline," 2000.

[19] World Health Organization, "WHO report on the global tobacco epidemic, 2008: the mpower package," 2008.

[20] Centers for Disease Control and Prevention, Office on Smoking and Health, "Best practices for comprehensive tobacco control programs," *Atlanta: Centers for Disease Control and Prevention*, pp. 1–11, 2007.

[21] J. J. Barendregt, L. Bonneux, and P. J. Van Der Maas, "The health care costs of smoking," *New England Journal of Medicine*, vol. 337, no. 15, pp. 1052–1057, 1997.

[22] U.S. Surgeon General, "The health benefits of smoking cessation," *Washington: Department of Health and Human Services*, 1990.

[23] M. J. Thun and C. W. Heath, "Changes in mortality from smoking in two american cancer society prospective studies since 1959," *Preventive Medicine*, vol. 26, no. 4, pp. 422–426, 1997.

[24] J. M. Samet, "The health benefits of smoking cessation." *The Medical Clinics of North America*, vol. 76, no. 2, pp. 399–414, 1992.

[25] D. H. Taylor Jr, V. Hasselblad, S. J. Henley, M. J. Thun, and F. A. Sloan, "Benefits of smoking cessation for longevity," *American Journal of Public Health*, vol. 92, no. 6, pp. 990–996, 2002.

[26] S. Shiffman, "Relapse following smoking cessation: a situational analysis," *Journal of Consulting and Clinical Psychology*, vol. 50, no. 1, p. 71, 1982.

[27] S. Shiffman, J. B. Engberg, J. A. Paty, W. G. Perz, M. Gnys, J. D. Kassel, and M. Hickcox, "A day at a time: predicting smoking lapse from daily urge," *Journal of Abnormal Psychology*, vol. 106, no. 1, p. 104, 1997.

[28] S. Shiffman, L. Read, J. Maltese, D. Rapkin, and M. E. Jarvik, "Preventing relapse in ex-smokers: A self-management approach," *Relapse Prevention: Maintenance Strategies in the Treatment of Addictive Behaviors*, pp. 472–520, 1985.

[29] T. M. Piasecki, M. C. Fiore, and T. B. Baker, "Profiles in discouragement: two studies of variability in the time course of smoking withdrawal symptoms," *Journal of Abnormal Psychology*, vol. 107, no. 2, p. 238, 1998.

[30] S. L. Kenford, M. C. Fiore, D. E. Jorenby, S. S. Smith, D. Wetter, and T. B. Baker, "Predicting smoking cessation: who will quit with and without the nicotine patch," *Journal of the American Medical Association*, vol. 271, no. 8, pp. 589–594, 1994.

[31] C. Silagy, D. Mant, G. Fowler, and M. Lodge, "Meta-analysis on efficacy of nicotine replacement therapies in smoking cessation," *The Lancet*, vol. 343, no. 8890, pp. 139–142, 1994.

[32] C. Viswesvaran and F. L. Schmidt, "A meta-analytic comparison of the effectiveness of smoking cessation methods," *Journal of Applied Psychology*, vol. 77, no. 4, p. 554, 1992.

[33] M. Jarvis, "Epidemiology of cigarette smoking and cessation," *Journal of Clinical Psychiatry*, vol. 18, pp. 6–11, 2003.

[34] S. Shiffman, "Dynamic influences on smoking relapse process," *Journal of Personality*, vol. 73, no. 6, pp. 1715–1748, 2005.

[35] Tobacco Advisory Group of the Royal College of Physicians, "Nicotine addiction in britain: a report of the tobacco advisory group of the royal college of physicians." Royal College of Physicians of London, 2000.

[36] World Health Organization, "Policy recommendations on smoking cessation and treatment of tobacco dependence," in *Policy recommendations on smoking cessation and treatment of tobacco dependence*. WHO, 2004.

[37] N. L. Benowitz, "Pharmacology of nicotine: addiction and therapeutics," *Annual Review of Pharmacology and Toxicology*, vol. 36, no. 1, pp. 597–613, 1996.

[38] B. Le Foll and T. P. George, "Treatment of tobacco dependence: integrating recent progress into practice," *Canadian Medical Association Journal*, vol. 177, no. 11, pp. 1373–1380, 2007.

[39] F. Dajas-Bailador and S. Wonnacott, "Nicotinic acetylcholine receptors and the regulation of neuronal signalling," *Trends in Pharmacological Sciences*, vol. 25, no. 6, pp. 317–324, 2004.

[40] H. Reuter, "Calcium channel modulation by neurotransmitters, enzymes and drugs," *Nature*, vol. 301, no. 5901, pp. 569–574, 1982.

[41] S. Wonnacott, "Presynaptic nicotinic ach receptors," *Trends in Neurosciences*, vol. 20, no. 2, pp. 92–98, 1997.

[42] E. J. Nestler, "Is there a common molecular pathway for addiction?" *Nature Neuroscience*, vol. 8, no. 11, pp. 1445–1449, 2005.

[43] W. A. Corrigall, K. M. Coen, and K. L. Adamson, "Self-administered nicotine activates the mesolimbic dopamine system through the ventral tegmental area," *Brain Research*, vol. 653, no. 1, pp. 278–284, 1994.

[44] G. Di Chiara and A. Imperato, "Drugs abused by humans preferentially increase synaptic dopamine concentrations in the mesolimbic system of freely moving rats," *Proceedings of the National Academy of Sciences*, vol. 85, no. 14, pp. 5274–5278, 1988.

[45] W. Schultz, "Dopamine neurons and their role in reward mechanisms," *Current Opinion in Neurobiology*, vol. 7, no. 2, pp. 191–197, 1997.

[46] H. D. Mansvelder and D. S. McGehee, "Long-term potentiation of excitatory inputs to brain reward areas by nicotine," *Neuron*, vol. 27, no. 2, pp. 349–357, 2000.

[47] T. Yoshida, N. Sakane, T. Umekawa, and M. Kondo, "Effect of nicotine on sympathetic nervous system activity of mice subjected to immobilization stress," *Physiology & Behavior*, vol. 55, no. 1, pp. 53–57, 1994.

[48] M. Sabha, J. E. Tanus-Santos, J. C. Y. Toledo, M. Cittadino, J. C. Rocha, and H. Moreno, "Transdermal nicotine mimics the smoking-induced endothelial dysfunction*," *Clinical Pharmacology & Therapeutics*, vol. 68, no. 2, pp. 167–174, 2000.

[49] S. Zhang, I. Day, and S. Ye, "Nicotine induced changes in gene expression by human coronary artery endothelial cells," *Atherosclerosis*, vol. 154, no. 2, pp. 277–283, 2001.

[50] A. Lewis, J. Miller, and R. Lea, "Monoamine oxidase and tobacco dependence," *Neurotoxicology*, vol. 28, no. 1, pp. 182–195, 2007.

[51] J. S. Fowler, J. Logan, G.-J. Wang, and N. D. Volkow, "Monoamine oxidase and cigarette smoking," *Neurotoxicology*, vol. 24, no. 1, pp. 75–82, 2003.

[52] E. L. Ochoa, L. Li, and M. G. McNamee, "Desensitization of central cholinergic mechanisms and neuroadaptation to nicotine," *Molecular Neurobiology*, vol. 4, no. 3-4, pp. 251–287, 1990.

[53] H. Wang and X. Sun, "Desensitized nicotinic receptors in brain," *Brain Research Reviews*, vol. 48, no. 3, pp. 420–437, 2005.

[54] M. De Biasi and J. A. Dani, "Reward, addiction, withdrawal to nicotine," *Annual Review of Neuroscience*, vol. 34, p. 105, 2011.

[55] M. Le Moal and G. F. Koob, "Drug addiction: pathways to the disease and pathophysiological perspectives," *European Neuropsychopharmacology*, vol. 17, no. 6, pp. 377–393, 2007.

[56] D. Venes, *Taber's cyclopedic medical dictionary*. FA Davis, 2013.

[57] T. P. George and A. H. Weinberger, "Nicotine and tobacco," *The American Psychiatric Publishing Textbook of Substance Abuse Treatment*, p. 201, 2008.

[58] J. R. Hughes and D. Hatsukami, "Signs and symptoms of tobacco withdrawal," *Archives of General Psychiatry*, vol. 43, no. 3, pp. 289–294, 1986.

[59] M. T. Treadway and D. H. Zald, "Reconsidering anhedonia in depression: lessons from translational neuroscience," *Neuroscience and Biobehavioral Reviews*, vol. 35, no. 3, pp. 537–555, 2011.

[60] G. F. Koob and M. Le Moal, "Drug abuse: hedonic homeostatic dysregulation," *Science*, vol. 278, no. 5335, pp. 52–58, 1997.

[61] K. A. Perkins and J. E. Grobe, "Increased desire to smoke during acute stress," *British Journal of Addiction*, vol. 87, no. 7, pp. 1037–1040, 1992.

[62] B. Bradley, M. Field, K. Mogg, and J. De Houwer, "Attentional and evaluative biases for smoking cues in nicotine dependence: component processes of biases in visual orienting," *Behavioural Pharmacology*, vol. 15, no. 1, pp. 29–36, 2004.

[63] N. Breslau, M. M. Kilbey, and P. Andreski, "Nicotine withdrawal symptoms and psychiatric disorders: findings from an epidemiologic study of young adults," *The American Journal of Psychiatry*, 1992.

[64] J. A. Dani and S. Heinemann, "Molecular and cellular aspects of nicotine abuse," *Neuron*, vol. 16, no. 5, pp. 905–908, 1996.

[65] J. A. Kauer and R. C. Malenka, "Synaptic plasticity and addiction," *Nature Reviews Neuroscience*, vol. 8, no. 11, pp. 844–858, 2007.

[66] T. J. Gould and J. A. Davis, "Associative learning, the hippocampus, and nicotine addiction," *Current drug abuse reviews*, vol. 1, no. 1, pp. 9–19, 2008.

[67] P. Olausson, J. D. Jentsch, and J. R. Taylor, "Repeated nicotine exposure enhances responding with conditioned reinforcement," *Psychopharmacology*, vol. 173, no. 1-2, pp. 98–104, 2004.

[68] J. E. Rose, F. M. Behm, and E. D. Levin, "Role of nicotine dose and sensory cues in the regulation of smoke intake," *Pharmacology Biochemistry and Behavior*, vol. 44, no. 4, pp. 891–900, 1993.

[69] J. E. Rose, F. M. Behm, E. C. Westman, and M. Johnson, "Dissociating nicotine and nonnicotine components of cigarette smoking," *Pharmacology Biochemistry and Behavior*, vol. 67, no. 1, pp. 71–81, 2000.

[70] J. E. Henningfield, N. Benowitz, G. Connolly, R. Davis, N. Gray, M. Myers, and M. Zeller, "Reducing tobacco addiction through tobacco product regulation," *Tobacco Control*, vol. 13, no. 2, pp. 132–135, 2004.

[71] F. T. Commission, "Federal trade commission cigarette report for 2012," *Tobacco Control*, 2012.

[72] E. J. Houtsmuller, J. E. Henningfield, and M. L. Stitzer, "Subjective effects of the nicotine lozenge: assessment of abuse liability," *Psychopharmacology*, vol. 167, no. 1, pp. 20–27, 2003.

[73] N. Benowitz, "Cigarette smoking and nicotine addiction." *The Medical Clinics of North America*, vol. 76, no. 2, pp. 415–437, 1992.

[74] S. Shiffman, "Light and intermittent smokers: background and perspective," *Nicotine & Tobacco Research*, p. ntn020, 2009.

[75] D. S. Moskowitz and S. N. Young, "Ecological momentary assessment: what it is and why it is a method of the future in clinical psychopharmacology," *Journal of Psychiatry and Neuroscience*, vol. 31, no. 1, p. 13, 2006.

[76] M. Hamilton, "Rating depressive patients." *Journal of Clinical Psychiatry*, 1980.

[77] D. L. Rosenhan, "On being sane in insane places," *Science*, vol. 179, no. 4070, pp. 250–258, 1973.

[78] D. A. Kenny and J. S. Berman, "Statistical approaches to the correction of correlational bias." *Psychological Bulletin*, vol. 88, no. 2, p. 288, 1980.

[79] L. Wheeler and H. T. Reis, "Self-recording of everyday life events: origins, types, and uses," *Journal of Personality*, vol. 59, no. 3, pp. 339–354, 1991.

[80] A. A. Stone and S. Shiffman, "Ecological momentary assessment (EMA) in behavorial medicine." *Annals of Behavioral Medicine*, 1994.

[81] S. Shiffman, A. A. Stone, and M. R. Hufford, "Ecological momentary assessment," *Annual Review of Clinical Psychology*, vol. 4, pp. 1–32, 2008.

[82] G. Affleck, A. Zautra, H. Tennen, and S. Armeli, "Multilevel daily process designs for consulting and clinical psychology: a preface for the perplexed." *Journal of Consulting and Clinical Psychology*, vol. 67, no. 5, p. 746, 1999.

[83] S. Shiffman, "Ecological momentary assessment (EMA) in studies of substance use." *Psychological Assessment*, vol. 21, no. 4, p. 486, 2009.

[84] A. S. Green, E. Rafaeli, N. Bolger, P. E. Shrout, and H. T. Reis, "Paper or plastic? Data equivalence in paper and electronic diaries," *Psychological Methods*, vol. 11, no. 1, p. 87, 2006.

[85] S. Shiffman, J. A. Paty, M. Gnys, J. A. Kassel, and M. Hickcox, "First lapses to smoking: within-subjects analysis of real-time reports," *Journal of Consulting and Clinical Psychology*, vol. 64, no. 2, p. 366, 1996.

[86] M. Perrine, J. C. Mundt, J. S. Searles, and L. S. Lester, "Validation of daily self-reported alcohol consumption using interactive voice response (IVR) technology," *Journal of Studies on Alcohol and Drugs*, vol. 56, no. 5, p. 487, 1995.

[87] S. Shiffman and A. J. Waters, "Negative affect and smoking lapses: a prospective analysis," *Journal of Consulting and Clinical Psychology*, vol. 72, no. 2, p. 192, 2004.

[88] A. A. Stone, S. Shiffman, A. A. Atienza, and L. Nebeling, "The science of real-time data capture," *New York*, 2007.

[89] S. Shiffman and S. L. Rathbun, "Point process analyses of variations in smoking rate by setting, mood, gender, and dependence," *Psychology of Addictive Behaviors*, vol. 25, no. 3, p. 501, 2011.

[90] M. S. Dunbar, D. Scharf, T. Kirchner, and S. Shiffman, "Do smokers crave cigarettes in some smoking situations more than others? Situational correlates of craving when smoking," *Nicotine and Tobacco Research*, vol. 12, no. 3, pp. 226–234, 2010.

[91] S. Shiffman, M. Gnys, T. J. Richards, J. A. Paty, M. Hickcox, and J. D. Kassel, "Temptations to smoke after quitting: a comparison of lapsers and maintainers." *Health Psychology*, vol. 15, no. 6, p. 455, 1996.

[92] R. Duda, P. Hart, and D. Stork, *Pattern Classification*. Wiley, 2000.

[93] A. R. Webb, *Statistical pattern recognition*. John Wiley and Sons, 2003.

[94] K. Fukunaga, *Introduction to statistical pattern recognition*. Academic press, 2013.

[95] B. Efron, "Estimating the error rate of a prediction rule: improvement on cross-validation," *Journal of the American Statistical Association*, vol. 78, no. 382, pp. 316–331, 1983.

[96] P. Simon, *Too Big to Ignore: The Business Case for Big Data*. John Wiley and Sons, 2013.

[97] S. B. Kotsiantis, I. Zaharakis, and P. Pintelas, *Supervised machine learning: A review of classification techniques*. IOS Press, 2007.

[98] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," vol. 14, no. 2, 1995, pp. 1137–1145.

[99] C. M. Bishop, *Pattern recognition and machine learning*. New York: springer, 2006, vol. 1.

[100] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," *Association for Computing Machinery Computing Surveys*, vol. 31, no. 3, pp. 264–323, 1999.

[101] J. MacQueen *et al.*, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, no. 14. Oakland, CA, USA., 1967, pp. 281–297.

[102] J. C. Bezdek, *Pattern recognition with fuzzy objective function algorithms*. Kluwer Academic Publishers, 1981.

[103] S. J. Roberts, R. Everson, and I. Rezek, "Maximum certainty data partitioning," *Pattern Recognition*, vol. 33, no. 5, pp. 833–839, 2000.

[104] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society*, pp. 1–38, 1977.

[105] T. Kohonen, "Self-organizing maps," *Springer Series in Information Sciences*, vol. 30, 1995.

[106] H. Brendryen, F. Drozd, and P. Kraft, "A digital smoking cessation program delivered through internet and cell phone without nicotine replacement (happy ending): randomized controlled trial," *Journal of Medical Internet Research*, vol. 10, no. 5, 2008.

[107] A. Rodgers, T. Corbett, D. Bramley, T. Riddell, M. Wills, R.-B. Lin, and M. Jones, "Do u smoke after txt? Results of a randomised trial of smoking cessation using mobile phone text messaging," *Tobacco Control*, vol. 14, no. 4, pp. 255–261, 2005.

[108] K. E. Heron and J. M. Smyth, "Ecological momentary interventions: incorporating mobile technology into psychosocial and health behaviour treatments," *British Journal of Health Psychology*, vol. 15, no. 1, pp. 1–39, 2010.

[109] W. T. Riley, D. E. Rivera, A. A. Atienza, W. Nilsen, S. M. Allison, and R. Mermelstein, "Health behavior models in the age of mobile interventions: are our theories up to the task?" *Translational Behavioral Medicine*, vol. 1, no. 1, pp. 53–71, 2011.

[110] D. Barber, *Bayesian reasoning and machine learning.* Cambridge University Press, 2012.

[111] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of machine learning.* MIT Press, 2012.

[112] S. Zhang, C. Zhang, and Q. Yang, "Data preparation for data mining," *Applied Artificial Intelligence*, vol. 17, no. 5-6, pp. 375–381, 2003.

[113] G. Batista and M. C. Monard, "An analysis of four missing data treatment methods for supervised learning," *Applied Artificial Intelligence*, vol. 17, no. 5-6, pp. 519–533, 2003.

[114] J. Tang, S. Alelyani, and H. Liu, "Feature selection for classification: A review," *Data Classification: Algorithms and Applications*, 2014.

[115] I. Jolliffe, *Principal component analysis.* Wiley Online Library, 2002.

[116] J. E. Jackson, *A user's guide to principal components.* John Wiley and Sons, 2005, vol. 587.

[117] A. M. Martínez and A. C. Kak, "PCA versus LDA," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 2, pp. 228–233, 2001.

[118] D. L. Swets and J. J. Weng, "Using discriminant eigenfeatures for image retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 8, pp. 831–836, 1996.

[119] P. N. Belhumeur, J. P. Hespanha, and D. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711–720, 1997.

[120] P. F. Brown, *The acoustic-modeling problem in automatic speech recognition.* Computer Science Department, Carnegie-Mellon University, 1987.

[121] G. R. Doddington, "Phonetically sensitive discriminants for improved speech recognition," in *International Conference on Acoustics, Speech, and Signal Processing.* IEEE, 1989, pp. 556–559.

[122] S. Dudoit, J. Fridlyand, and T. P. Speed, "Comparison of discrimination methods for the classification of tumors using gene expression data," *Journal of the American Statistical Association*, vol. 97, no. 457, pp. 77–87, 2002.

[123] J. Ye, R. Janardan, and Q. Li, "Two-dimensional linear discriminant analysis," in *Advances in Neural Information Processing Systems*, 2004, pp. 1569–1576.

[124] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial Intelligence*, vol. 97, no. 12, pp. 273–324, 1997.

[125] M. Dash and H. Liu, "Feature selection for classification," *Intelligent Data Analysis*, vol. 1, no. 3, pp. 131–156, 1997.

[126] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *The Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[127] S. Arlot and A. Celisse, "A survey of cross-validation procedures for model selection," *Statistics Surveys*, vol. 4, pp. 40–79, 2010.

[128] J. Shao, "Linear model selection by cross-validation," *Journal of The American Statistical Association*, vol. 88, no. 422, pp. 486–494, 1993.

[129] S. Geisser, "The predictive sample reuse method with applications," *Journal of The American Statistical Association*, vol. 70, no. 350, pp. 320–328, 1975.

[130] M. Stone, "Cross-validatory choice and assessment of statistical predictions," *Journal of the Royal Statistical Society*, pp. 111–147, 1974.

[131] D. M. Allen, "The relationship between variable selection and data agumentation and a method for prediction," *Technometrics*, vol. 16, no. 1, pp. 125–127, 1974.

[132] Y. Bengio and Y. Grandvalet, "No unbiased estimator of the variance of k-fold cross-validation," *The Journal of Machine Learning Research*, vol. 5, pp. 1089–1105, 2004.

[133] G. McLachlan, K.-A. Do, and C. Ambroise, *Analyzing microarray gene expression data*. John Wiley and Sons, 2005, vol. 422.

[134] K. A. O'Connell, J. E. Schwartz, and S. Shiffman, "Do resisted temptations during smoking cessation deplete or augment self-control resources?" *Psychology of Addictive Behaviors*, vol. 22, no. 4, p. 486, 2008.

[135] E. Sejdić, C. M. Steele, and T. Chau, "Classification of penetration-aspiration versus healthy swallows using dual-axis swallowing accelerometry signals in dysphagic subjects," *IEEE Transactions on Bio-Medical Engineering*, vol. 60, no. 7, pp. 1859–1866, 2013.

[136] D. D. Lewis, *Naive (Bayes) at forty: The independence assumption in information retrieval*. Springer, 1998.

[137] T. M. Mitchell, *Machine Learning*. McGraw Hill, 1997.

[138] S. Goldwater, T. L. Griffiths, and M. Johnson, "A Bayesian framework for word segmentation: Exploring the effects of context," *Cognition*, vol. 112, no. 1, pp. 21–54, 2009.

[139] N. Friedman, D. Geiger, and M. Goldszmidt, "Bayesian network classifiers," *Machine Learning*, vol. 29, no. 2-3, pp. 131–163, 1997.

[140] A. L. Garcia, *Probability, Statistics, and Random Processes for Electrical Engineering.* Prentice Hall, 2008.

[141] W. N. Venables and B. D. Ripley, *Modern applied statistics with S.* Springer, 2002.

[142] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, vol. 7, no. 2, pp. 179–188, 1936.

[143] E. Alexandre-Cortizo, M. Rosa-Zurera, and F. Lopez-Ferreras, "Application of Fisher linear discriminant analysis to speech/music classification," 2005.

[144] L. Rokach, *Data mining with decision trees: theory and applications*, ser. Series in machine perception and artificial intelligence. World Scientific, 2007.

[145] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and regression trees.* CRC press, 1984.

[146] W.-Y. Loh, "Classification and regression trees," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 1, no. 1, pp. 14–23, 2011.

[147] R. Maclin and D. Opitz, "Popular ensemble methods: An empirical study," *Journal of Artificial Intelligence Research*, vol. 11, pp. 169–198, 2011.

[148] S. V. Stehman, "Selecting and interpreting measures of thematic classification accuracy," *Remote Sensing of Environment*, vol. 62, no. 1, pp. 77–89, 1997.

[149] G. A. Marlatt and J. Gordon, *Relapse prevention: Maintenance strategies in the treatment of addictive behaviors.* Guilford Press, 1985.