UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Facultad de Administración y Dirección de Empresas /UPV

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Faculty of Business Administration and Management

Investigation of factors affecting patient retention in dental clinics using machine learning techniques

Master's Thesis

Master's Degree in Business, Product and Service Management

AUTHOR: Serrano Amaya, Diego Alejandro

Tutor: Peiró Signes, Ángel

ACADEMIC YEAR: 2023/2024

Abstract

Patient retention is a key measure of organizational performance and is still one of the most significant challenges dental clinics face today. Sometimes, prospective patients never return after the first visit, forsaking the opportunity for a dental clinic to address their oral status and improve their overall well-being, quality of life, and self-confidence. Other times, patients abandon a prescribed dental treatment halfway, which can worsen the problem over time and increase the pain, discomfort, or other complications. For example, most periodontal diseases are complex, chronic, and progressive, requiring treatment control, evaluation, and maintenance over an extended duration of time.

From a business perspective, it is widely accepted that retaining existing customers, or patients in this case, is perceived as more cost-effective than acquiring new ones.

Consequently, knowledge and improvement in retention measures ought to be essential to take into consideration for dental care providers and patients alike.

This study is made possible through access to patient records obtained from a privately owned, non-franchised dental clinic located in Valencia, Spain.

The main objective of this investigation is to utilize machine learning algorithms to interpret a classification model that attempts to predict the likelihood of returning by a prospective dental patient who has initially contacted an oral healthcare clinic.

In addition to this primary objective, the study seeks to analyze the individual characteristics and historical behavior factors most correlated to the retention of the patients in the dataset.

Lastly, a comprehensive literature review will be conducted to examine the current state-of-the-art studies, findings, and strategies concerning patient retention in dental clinics.

Keywords:

Patient Retention, Dental Clinics, Machine Learning, Classification Models, Clinic Management.

Acknowledgments

I am deeply grateful to my beloved family for their support, to my colleagues, to my master's program teachers, and especially to Ángel for expertly structuring and guiding my ideas.

List of Abbreviations

AI	Artificial Intelligence
CRM	Customer Relationship Management
EU	European Union
FDI	Fédération Dentaire Internationale (World Dental Federation)
ML	Machine Learning
OECD	Organisation for Economic Co-operation and Development
SHAP	SHapley Additive exPlanations
U.S.	United States
US$	United States Dollar
WHO	World Health Organization
XGBoost	Extreme Gradient Boosting
ZIP Code	Zone Improvement Plan Code

List of Figures

List of Tables

Table of Contents

1. Introduction

1.1. Justification

In the current highly competitive landscape of dental services, where the offering of treatments is broad and varied, clinics should strive to distinguish themselves in the marketplace through innovative techniques and personalized patient care. This imperative for differentiation is underscored by the private oral healthcare sector's significant evolution in recent years, characterized by clinical and technological advances (Naamati-Schneider & Salvatore, 2022), as well as an emphasis on enhancing managerial practices and implementing effective patient retention strategies (Amano, 2023).

Despite its significance, patient retention remains one of the primary challenges private clinics and oral healthcare providers face worldwide. This fact is highlighted by (Maycher, 2023) in her discussion on the Dental Brief Podcast, where she states that “retention still stands as one of the most important challenges in the industry.” This matter is echoed by (Wright, 2024), who similarly mentions the critical importance of addressing patient retention in the dental industry. He initially focuses on patient volume before shifting his investigation to patient retention.

Acknowledging these challenges is crucial to understand that dentistry is not only a technically oriented profession, but also one that requires strong organizational, managerial, and business acumen. Therefore, private non-franchised dental clinics must adopt competitive business strategies, preferably including digital technologies, to improve both their business performance and patients’ experience (Naamati-Schneider & Salvatore, 2022).

On the other hand, it is important to emphasize that the sector is significantly shaped by the trends of digitalization and technology, to the extent that scholars are now introducing the

term “Data Dentistry” (Schwendicke & Krois, 2022, p. 21) to denote the integration of data-driven decision-making into oral healthcare practices. This concept emerges from the vast array of data that can be collected in the dental environment, which, according to Schwendicke & Krois (2022), can go "from the individual level (e.g., demographic, social, and clinical data obtained via records mining, clinical assessment, omics analyses, and real-time consumer data from wearables and tracking devices); setting level (e.g., geospatial, environmental, or provider-related data); and systems level (e.g., health insurance, regulatory, and legislative data)” (p. 23). Figure 1 illustrates the multitude of data that can be harvested from the dental workflow and made available for use.

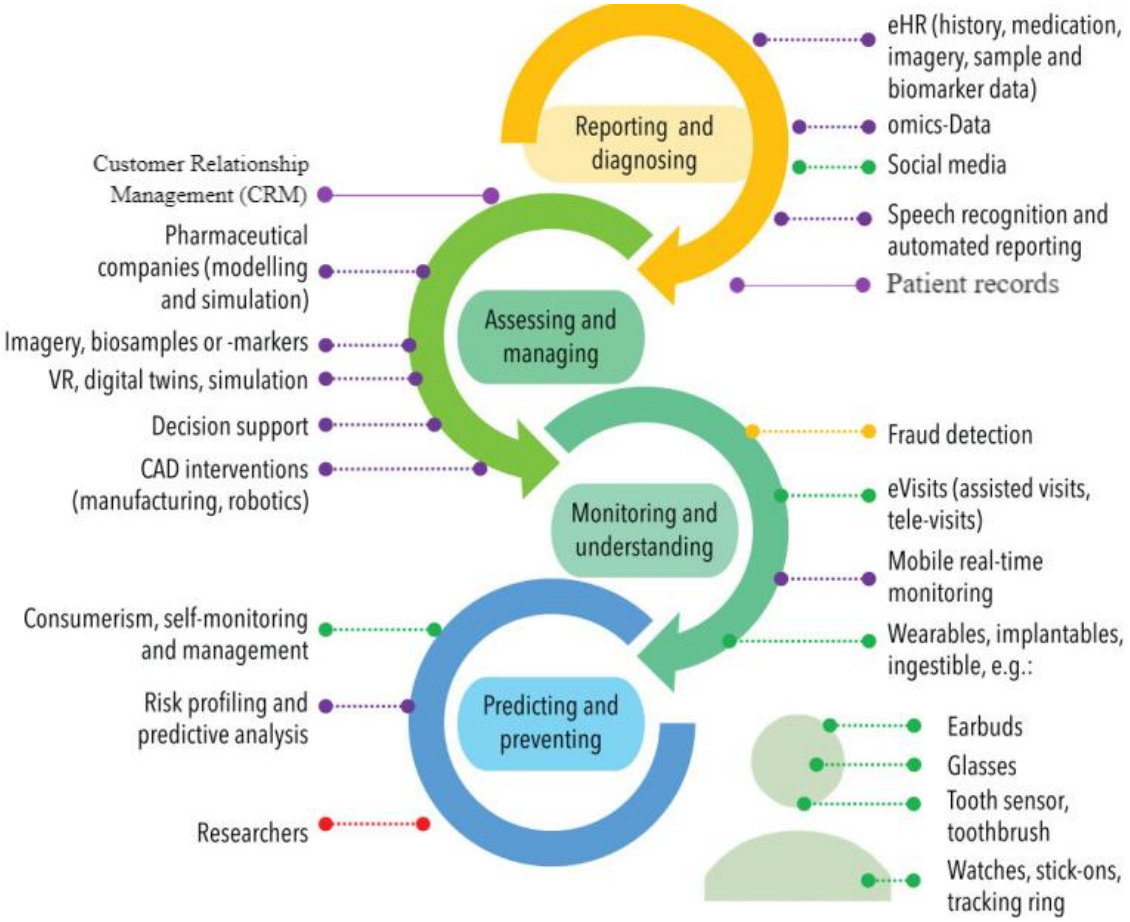

Figure 1. The data-driven clinical workflow. (purple: provider; green: patient; yellow: payer; red: researcher). Adapted from Schwendicke & Krois (2022).

Wright (2024) further emphasizes the value of leveraging data collected from diverse sources, such as social media and the office's software, as it provides a foundation for acquiring knowledge and enhancing patient recruitment and retention while monitoring and improving the quality of care provided in dental settings.

Considering the challenges and opportunities presented by digital advancements, this investigation aims to apply machine learning techniques to develop a model predicting the likelihood of patients returning to dental clinics based on various factors. By precisely forecasting patient return rates, the model provides actionable insights to refine patient retention strategies, leading to improved financial outcomes, and operational efficiencies.

1.2. Objectives

1.2.1. Main objective.

To utilize machine learning algorithms to develop and interpret a classification model predicting the likelihood of prospective dental patients returning after their initial contact and visit to an oral healthcare clinic.

1.2.2. Specific objectives.

- Analyze a dataset from a privately owned, non-franchised dental clinic to identify individual characteristics and historical behavior factors correlated with patient retention.
- Conduct a comprehensive literature review on current state-of-the-art studies, findings, and strategies related to patient retention in dental clinics.
- Provide insights and recommendations for improving patient retention strategies based on the predictive model and literature review findings.

2. Literature Review

2.1. The Importance of Oral Healthcare for Individuals and Society

Oral health is linked with overall health and quality of life. The World Health Organization (WHO) defined oral health as "a state of being free from mouth and facial pain, oral and throat cancer, oral infection and sores, periodontal disease, tooth decay, tooth loss, and other diseases and disorders that limit an individual's capacity in biting, chewing, smiling, speaking, and psychosocial well-being" (World Health Organization, 2012, p. 1). The latest WHO publication on oral health, *Global Oral Health Status Report: Towards universal health coverage for oral health by 2030*, expands on this definition. It accentuates that oral health is present when the orofacial structures of the human body function without pain, discomfort, or embarrassment, hence encompassing psychosocial dimensions such as self-confidence, well-being, and the ability to socialize (World Health Organization, 2022).

Addressing oral health needs requires a holistic approach throughout every stage of life, regardless of age or genetics. Maintaining optimal oral health demands good hygiene habits, consistent self-care practices, avoiding risk factors like high sugar consumption or smoking, and regular professional attention (Elflein, 2024). In other words, a healthy lifestyle plus a diligent adherence to preventive measures such as regular dental cleanings and check-ups can ensure individuals with healthier mouths and smiles.

The aforementioned WHO publication highlights that advancements in oral healthcare interventions and technologies have significantly improved treatment outcomes for those seeking professional care. According to the WHO (2022), "clinical oral healthcare procedures now effectively alleviate pain, discomfort, and infection caused by oral diseases, and they

help to restore patients' oral function and aesthetics, thereby improving their psychosocial well-being and health" (p. 60).

Despite these advancements in oral healthcare, conditions like dental caries, periodontal disease, oral cancer, and tooth loss still affect billions worldwide. James et al. (2019) report that oral disease has the most prevalent presence globally among all ages and sexes, with dental caries in permanent teeth ranked first versus 354 other diseases and injuries across 195 countries and territories. Furthermore, as time has passed, the WHO (2022) continues to underscore the significance of untreated dental caries in permanent teeth as the most common health condition and untreated caries in deciduous teeth as the single most common chronic childhood disease, affecting 514 million children worldwide.

The FDI World Dental Federation (FDI) (n.d.) highlights on their website how oral disease can "impact every aspect of life, including personal relationships and self-confidence. It can lead to significant pain, anxiety, disfigurement, acute and chronic infections, eating as well as sleep disruption and can result in social isolation, loss of work and school days, and impaired quality of life."

Further emphasizing the widespread impact, the *Oral Health Atlas 2nd edition*, published by the FDI (2015b), asserts that oral diseases significantly impact individuals, communities, society, health systems, and the economy. The Atlas also cites the World Health Organization, noting that "oral diseases are the fourth most expensive diseases to treat" (p. 56). This multifaceted impact underscores the importance of addressing oral health not only at an individual level but also at a societal and systemic level.

2.2. Challenges in the Oral Healthcare System

The attainment of Universal Health Coverage (UHC) by 2030, as outlined in Goal 3.8 of the Sustainable Development Goals (United Nations, 2015), represents a crucial milestone for global general health care. However, despite its inclusion on the agenda, essential oral health services remain largely overlooked and inaccessible in many countries. Treatment for oral health conditions continues to be expensive and is often excluded from Universal Health Coverage (World Health Organization, 2023).

This chapter explores the complex challenges confronting the oral healthcare system, which includes Government and Policy Support; Inequalities in Access to Oral Healthcare; Financial Barriers and Affordability; Emerging Trends and Technology Integration; and Workforce and Resource Allocation.

Government and Policy Support. Oral diseases, despite being the most prevalent health issues worldwide, receive inadequate attention and government support in countries with weak healthcare systems (FDI, 2015a). This neglect is reflected in the high spending on oral healthcare. For instance, annual spending on oral healthcare in the 27 European Union member states was estimated at €79 billion (annual average 2008–12), while the U.S. alone spent more than US$ 110 billion (FDI, 2015b).

The WHO reports that "the total direct expenditure for oral diseases among 194 countries amounted to US$ 387 billion or a global average of about US$ 50 per capita in 2019. This represents about 4.8% of global direct health expenditures" (WHO, 2022, p. 26).

Qin et al. (2022) compared Global Burden of Disease data from 1990 to 2019, finding that the "treatment rate of dental caries is increasing, indicating that the health care

system in various countries is constantly improving, and more people are getting dental treatment. However, the incidence of dental caries has not declined, which shows that prevention work is not good enough, and the coverage of prevention is not wide enough" (p. 12). This insight underscores the need for more robust government intervention and policy frameworks to support the enhancement of dental education, bolstering preventive measures, and improving oral health coverage.

Inequalities in Access to Oral Care. Persisting oral care inequalities and lack of access to oral healthcare perpetuate disparities, leaving many individuals without essential services. Limited or no access for rural, remote, or disadvantaged populations exacerbates these issues, resulting in untreated oral disease for large segments of society.

This concern is echoed by the FDI (2015a), which explains that "increasing privatization of oral health services in many countries, driven by reduced government spending, is likely to decrease the accessibility and universality of oral healthcare and may increase inequalities" (p. 16). This trend is further emphasized by the WHO (2022), stating that the predominance of private provision models and under-resourced public services further compound existing disparities.

For instance, in some African regions, "the dentist-to-population ratio is 1:150,000 or higher, whereas, in industrialized countries, there is one dentist per 5,000 people or more" (FDI, 2015a, p. 16). According to the latest report by the *Consejo General de Colegios Oficiales de Odontólogos y Estomatólogos de España*, the recommended ratio by the WHO is one dentist per 3,500 inhabitants, and Spain's ratio in 2022 was 1:1,171 (Consejo General de Colegios de Dentistas de España, 2022b).

The FDI (2015b) stresses that even with amplified public subsidies for dental care, expanded health insurance coverage, and significant availability of oral healthcare services, disparities will persist unless individuals from disadvantaged backgrounds understand the importance of good oral health and unless policy programs target the broader determinants of preventive health.

Financial Barriers and Affordability. Financial barriers pose significant challenges to accessing oral healthcare services worldwide. In many regions, dental treatment remains unaffordable, especially in low- and middle-income countries. Even in some high-income countries, large segments of the population face obstacles to oral healthcare access due to high treatment cost. As the FDI (2015b) emphasized, the affordability of oral care services is a major barrier since patient bear most of the treatment costs.

A comparison between Spain and other countries reveals striking differences in dental care financing. As shown in Figure 2, practically the entirety of dental expenditure in Spain is financed by patients' out-of-pocket expenses, with only 2% covered by public expenditure.

In contrast, the public healthcare system in France funds a significant amount of oral health treatments for the population. Individuals with top-up insurance have access to a limited range of dental prostheses, such as implants, crowns, or bridges. As stated by Buswell (2024), most dentists work within the public French healthcare system, and costs are reimbursed in the same way as other medical treatments. Most adults receive 70% reimbursement for dental charges, while children's checkups are fully reimbursed at 100%. However, orthodontic treatment is not covered under the

scheme, meaning that braces, often recommended for teenagers, are not included. However, (Pegon-Machat et al., 2016) discuss two major problems in the provision of oral healthcare in France: the high cost, making France the EU country with the highest health expenditure as a share of Gross Domestic Product (GDP) in 2008, and the exacerbation of oral health inequalities, partly due to limitations within the health insurance system. Figure 2 shows that France's take-up of dental care has the highest probability of visiting the dentist in the past 12 months among the selected OECD countries. However, this trend is predominantly seen among individuals in the upper end of the socioeconomic scale.

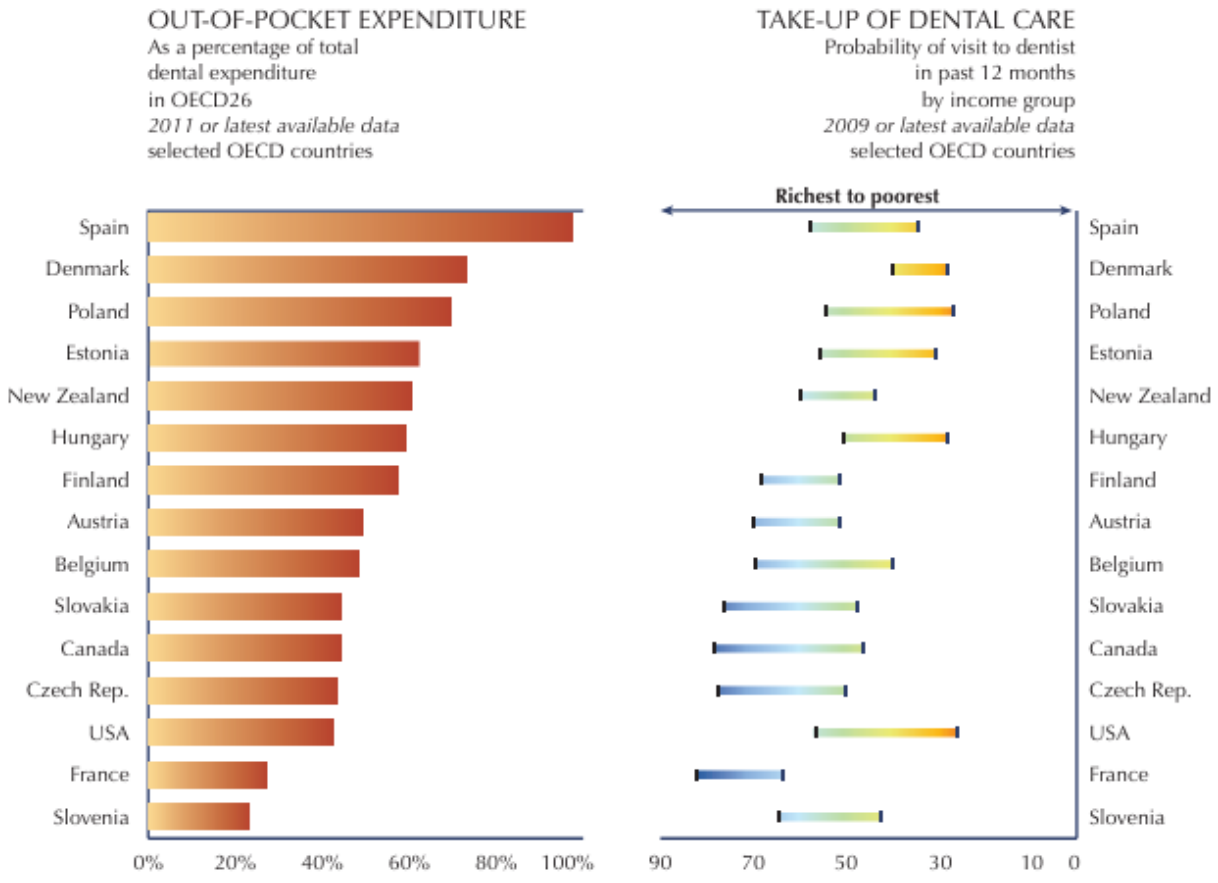

Figure 2. Comparison of Out-of-Pocket Expenditure and Take-Up of Dental Care in Selected OECD Countries. Taken from (FDI, 2015b).

In Germany, which is not visible in the figure, public sources cover essential dental services like routine check-ups, cleanings, and caries, and finance more than half of dental care spending for the most complex dental procedures, at 68% (Consejo General de Colegios de Dentistas de España, 2022a).

Oral diseases have a considerable impact in terms of treatment costs and productivity losses. The barriers to accessing oral healthcare result in productivity losses from untreated oral diseases, estimated at US$ 42 per capita, totaling around US$ 323 billion globally (Jevdjevic & Listl, 2022).

This discrepancy highlights the disparities in access to affordable oral healthcare across different healthcare systems and underscores the need for policy interventions to address these inequalities. Addressing these financial barriers is crucial to ensuring equitable access to oral healthcare for all individuals, regardless of socioeconomic status.

Emerging Trends and Technology Integration. Emerging trends and technological advancements are revolutionizing oral healthcare, offering new opportunities for improving access, efficiency, and outcomes.

The *Oral Health Atlas 2nd edition* highlights several key trends shaping the future of oral healthcare, including a "growing and aging population, workforce migration, dental tourism, the emergence of new educational models, the evolving distribution of tasks among members of the oral healthcare workforce, ongoing legislative actions targeting hazardous materials, and the increasing use of information and communication technologies (ICTs) in all segments of lives and occupations" (FDI, 2015b, p. 71).

One of the most significant trends in dentistry is the rise of digital technologies such as tele-dentistry, electronic health records (EHR), and artificial intelligence (AI) (Alauddin et al., 2021; Joda et al., 2021). Tele-dentistry enables remote consultations and follow-ups, reducing the need for in-person visits and thereby increasing access to care, especially in underserved areas. EHR systems streamline patient record management, enhancing the accuracy and accessibility of patient information across different healthcare providers. AI applications are being developed to assist in diagnostics, treatment planning, and even in performing routine dental procedures. These innovations not only improve the efficiency of dental practices but also hold the potential to reduce costs and improve patient outcomes.

Countries in economic transition are experiencing the highest rates of dental decay as rising incomes lead to increased risk exposure, such as unhealthy diets and tobacco consumption. These health systems often lack the infrastructure and population-wide preventive measures to combat this rising tide of oral diseases (FDI, 2015a). This is particularly concerning as oral diseases are often hidden and invisible, or they are accepted as an unavoidable consequence of life and aging. However, there is definite evidence that oral diseases are not inevitable but can be reduced or prevented through simple and effective measures at all stages of life, both at the individual and population levels (FDI, 2015b).

Despite the promising developments, several challenges remain. Unequal access to technology, data security concerns, and the need for significant investment in infrastructure and training are major hurdles that need to be addressed to fully realize the potential of these advancements.

Workforce and Resource Allocation. Ultimately, in addressing the array of challenges outlined in this chapter, it is imperative to recognize the critical issues affecting the dental workforce both within and outside traditional healthcare settings. The WHO (2022) identifies several concerns related to the global oral health workforce, including a low number of professionals in the field, unequal distribution of dentists within nations, skill imbalances, and insufficient management or planning roles within oral healthcare teams as significant obstacles to delivering effective oral healthcare services.

For example, in the U.S., 91% of active dentists worked in private practice settings in 2018, and by 2021, 46% of private practice dentists were in solo practice (Fellows et al., 2022). The WHO (2022) estimates that "the total oral health workforce amounts to nearly 4 million globally, comprising about 2.5 million dentists, 1.2 million dental assistants, and nearly 300,000 prosthetists and technicians. The global average density for dentists is 3.28 dentists per 10,000 population, for assistants is 1.88 per 10,000 population, and for technicians is 0.57 per 10,000 population" (p. 63).

Additionally, the FDI (2015b), in the most recent version of the *Oral Health Atlas*, estimated approximately "2 million oral health providers such as private clinics, and a burden of over 10 million diseases attributed solely to tooth decay and periodontal disease, resulting in a global average oral disease burden/provider ratio of around 5.3" (p. 60). These statistics indicate a clear insufficiency in the current workforce and distribution models to meet the increasing demands for oral healthcare.

With the majority of dentists concentrated in urban areas serving more affluent populations, rural and underserved communities are left without adequate access to

essential oral health services. Addressing these workforce challenges and promoting the equitable distribution of dental professionals are essential steps toward ensuring universal access to quality oral healthcare for all individuals. In the subsequent chapter of this study, the focus will be on exploring how some of these workforce dynamics impact patient care and retention strategies in private dental clinics.

2.3. Patient Retention in Private Dental Clinics

Private dental clinics operate as for-profit businesses within a highly competitive free market. Unlike dental school clinics and publicly funded facilities, these clinics are established and run with the primary intention of generating profit. Their financial viability relies entirely on the revenue generated from dental services, as they do not receive subsidies or government support.

Typically, private clinics use a fee-for-service model, where patients pay out-of-pocket or through private dental insurance plans. This model places private dental clinics in direct competition with each other and public counterparts. To remain competitive, private clinics must continually innovate, effectively manage patient retention strategies, and invest in marketing efforts to attract new patients and maintain a steady stream of clientele. Protecting and growing the patient base is essential for sustaining revenue and ensuring long-term business success.

Strategic Management and Marketing Challenges in Private Dental Clinics. In this competitive environment, private dental clinics must differentiate themselves through various means, including quality of care, patient experience, financing plans, technological advancements, and, no less importantly, effective marketing strategies.

However, many of these clinics, especially non-franchised ones, are predominantly established and managed by dentists and professionals within the oral healthcare domain, who may lack formal education in business administration, economics, and marketing (Naamati-Schneider & Salvatore, 2022).

As a result, smaller private dental clinics often struggle to implement effective business practices. They face significant challenges in developing the managerial and strategic expertise necessary to navigate the complexities of running a successful healthcare business. This situation compounds their difficulties in attracting and retaining patients. Additionally, as for-profit entities, these clinics are subject to market forces such as supply and demand, competition, pricing pressures, and consumer choice, making it essential for them to attract and retain patients to ensure profitability.

Another challenge impacting patient retention strategies is the ongoing debate about the optimal frequency of dental visits. The commonly recommended twice-a-year visits, often promoted by toothpaste advertisements, lacks substantial research support. According to Colgate's Global Scientific Communications (2023), this guideline may not be suitable for everyone; some individuals may need only one or two visits annually, while others require more frequent check-ups or treatments. Systematic reviews by Kay (1999) and Gussy et al. (2013) indicate that existing research on the ideal frequency of dental visits is insufficient to draw meaningful conclusions.

Therefore, it is the obligation of dental clinics to determine an optimal frequency of dental appointments tailored to each patient's individual needs and oral health status.

This approach requires careful consideration and consultation with patients to establish appropriate schedules for future check-ups, cleanings, or necessary treatments.

Leveraging Patient Retention for Private Dental Practice Growth. While there has yet to be a consensus on how often to visit the dentist or encourage patients to return, there is widespread agreement on the importance of retention over acquisition. In fact, retention is a vital measure of an organization's performance (Gruen et al., 2000).

For example, Kotler & Keller (2016) emphasize in the 15th edition of *Marketing Management* that "acquiring new customers can cost five times more than satisfying and retaining current ones" (p. 163). As cited by Sabbeh (2018), other authors assert that retaining existing customers is at least 5 to 25 times more cost-effective than acquiring new ones, depending on business domains. The reason for this is primarily due to the extensive marketing and advertising efforts required to attract and then acquire new customers, or patients in this case.

Moreover, studies show that a 5 percent reduction in the defection rate, which is the opposite of the retention rate, can increase profits by 25 percent to 85 percent, depending on the industry. The significant boost in profitability can be attributed to factors such as increased purchases from loyal customers, who are more likely to keep visiting for additional treatments; referrals, as satisfied patients recommend the clinic to others; and reduced operating costs, since the expenses associated with servicing and marketing to existing patients are generally lower than those required for acquiring new ones (Kotler & Keller, 2016).

Some strategies to improve retention and thereby ensure sustained growth and profitability for private dental practices include:

- **Reducing the defection rate:** Training employees to be knowledgeable and friendly, providing exceptional customer service to nurture strong connections with patients.
- **Increasing the relationship's longevity:** The more profoundly a patient is involved and listened to, the higher the likelihood that the patient will remain loyal to the clinic. Encouraging feedback and genuinely listening to patients' needs and pains can foster engagement with the brand and clinic.
- **Focusing concentrated efforts on high-priority patients**: Emergencies, urgent needs, or severe pain should be promptly addressed and given the highest priority in the clinic's care delivery. Thoughtful gestures to every patient, such as birthday greetings, small gifts, or special event invitations, can send them a strong positive signal.
- **Utilizing Social Media Analytics**: Having a social media strategy and using social media analytics throughout the patient's journey outperforms peers who choose to ignore or avoid them (Paul Isson, 2018, p. 151).

Given the scarcity of strategies specifically addressing patient retention in private dental clinics, insights related to retention have been adapted to patients and private dental practices. As private dental clinics strive to enhance patient experiences and foster long-term relationships, strategies such as reducing the defection rate, increasing relationship longevity, prioritizing specific individuals, and leveraging social media analytics are pivotal for any service provider organization.

The next section of this chapter explores the journey of patients of private dental clinics, where retention plays a crucial role in shaping their engagement with oral healthcare services.

Private dental clinic patients' journey. Understanding the patient journey in private dental clinics is important for identifying key touchpoints that influence overall patient experience. By meticulously analyzing each step of the journey, clinics can optimize operations, reduce unnecessary expenditures, and focus efforts on high-value activities.

A patient's journey encompasses interactions from initial contact and appointment scheduling to consultation, treatment, and follow-up care. Each stage presents opportunities for clinics to connect with their patients, meet their needs, and exceed expectations.

Sabbeh (2018), in her investigation, delineates two crucial phases before retention can be conceived: identification and attraction. The identification phase involves identifying and categorizing potential patient segments that are most likely to engage with and benefit from the clinic's services. Clinics use tactics such as market research and surveys to understand patient demographics, preferences, and needs of the local population. This insight helps in positioning the clinic effectively to attract the most suitable segment of patients for sustainable practice growth.

The attraction phase centers on capturing and engaging these potential patients through targeted marketing strategies. This includes creating compelling content, optimizing the website for search engines, maintaining active social media profiles, and leveraging positive online reviews to build a strong online presence. During this

phase, the clinic's focus shifts from internal operations to external outreach. Once a patient is drawn in and attends an initial appointment, they are considered acquired.

Some individuals may skip directly to the acquisition phase, actively seeking and contacting a clinic for attention due to urgent dental issues. It is the clinic's responsibility to effectively address the patient's concerns in a professional and caring manner.

According to Makarem & Coe (2014), successfully managing or resolving dental conditions enhances patient retention. Thus, retention strategies should prioritize disease control as an initial step.

The next critical phase is patient retention, which involves encouraging continuous care through regular check-ups, cleanings, and necessary treatments. Effective retention strategies focus on building strong relationships with patients, determining optimal visits frequencies, promptly addressing emergencies, swiftly responding to concerns and inquiries, and ensuring adherence to preventive measures and follow-up care. Integrating Customer Relationship Management (CRM) systems into clinic operations can support these efforts by automating patient communications, personalizing interactions, facilitating appointment scheduling and reminders, and tracking patient satisfaction and engagement levels. This integration of CRM systems allows proactive outreach, enhances efficiency, and facilitates the development of tailored care plans, ultimately improving patient retention (Rigby, 2017).

As Figure 3 illustrates, retention is the critical middle step between acquisition and continuation in the funnel of private dental clinic patients. It bridges the gap between initial patient acquisition and long-term engagement, marking the beginning of a

sustained relationship between the patient and the healthcare provider. Retention is vital because it transforms initial interactions into ongoing relationships, ultimately contributing to patient loyalty and continuity of care.

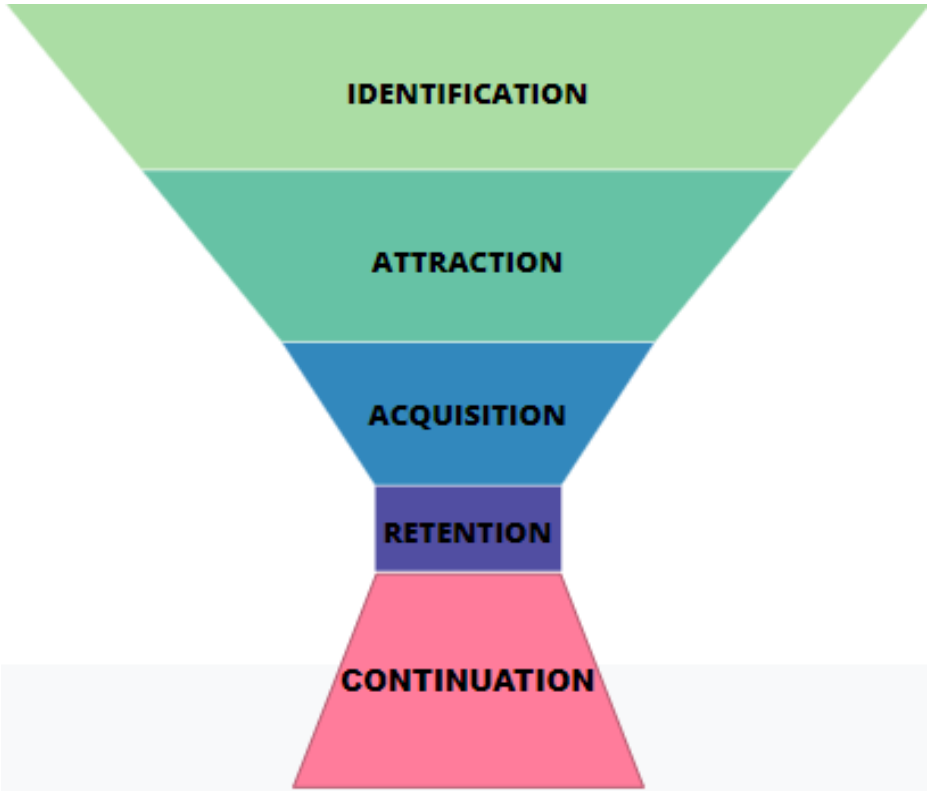

Figure 3. The private dental clinic patients' journey funnel.

Once a patient visits the clinic for their initial appointment, whether for a routine check-up or to address a specific dental concern, administrative tasks should also focus on ensuring their continued engagement with the clinic's services. This involves fostering a sense of trust, care, and professionalism that encourages patients to return for future appointments. Through regular communication, personalized care plans, and attentive follow-ups, clinics aim to create a positive patient experience that motivates ongoing engagement with their oral health journey.

Regrettably, retention is the narrowest phase of the patient journey funnel. Consequently, many patients do not advance to the continuation phase, which prevents them to maintain proper oral health and receive necessary care. Personal circumstances or external factors beyond their control, may lead some patients to abandon treatment, while others may miss the opportunity for the dental clinic to monitor their dental health, affecting their overall well-being.

Retaining existing patients proves cost-effective, contributes to a steady revenue stream, and supports community healthcare continuity and clinic growth. Understanding the factors influencing patient retention is fundamental. The next chapter delves into these factors in detail to provide valuable insights for oral healthcare providers.

2.4. Factors Influencing Patient Retention in Private Dental Practice

Having established the critical role of patient retention in private dental clinics, it is imperative to delve deeper into the various factors influencing this process. According to Han & Hyun (2015), perceived service quality, satisfaction, and trust are consistently recognized by researchers as crucial concepts that drive favorable intentions toward loyalty and post-purchase behaviors, thereby significantly affecting retention.

Exploring these concepts further, Onyeaso & Adalikwu (2008) identified a stable positive link between retention and perceived quality. They emphasize that customers' past experiences with perceived quality positively influence current retention levels. They also cite research showing that "customer satisfaction is a major driver of retention rate, and the latter is positively related to market share" (p. 55). Concurrently, Onyeaso & Adalikwu

(2008) explain that customers are retained due to their perception of quality and value in an organization's services, which contributes to the establishment of trust.

Szabó et al. (2023) investigated aspects of patients' dental care experiences that affect perceived satisfaction and loyalty to their dentists. They surveyed 1,121 patients and 77 dentists, and concluded that factors such as "location convenience, treatment quality, trust in dentists' decisions, visit frequency satisfaction, clear treatment explanations, dentist's interest in symptoms, patient-dental personnel attachment, and the dentist's knowledge of the patient and their medical records" (p. 1) significantly influence satisfaction and loyalty.

Makarem & Coe (2014) highlight that multiple components of the service encounter should be considered when assessing drivers of retention. These components can be grouped into three main areas: the service receiver; the service provider; and the context of the service encounter.

In the context of the service receiver, demographics such as age, education, and income are influential factors in retention. Referenced research by Makarem & Coe (2014) indicates that that older patients tend to rely less on information search and prioritize relationships and emotional connections, which positively influence retention. Conversely, higher income and education levels may be associated with lower retention rates due to a wider range of available options. For instance, private dental insurance can mitigate dental care costs and provide holders with various options for receiving oral healthcare. Additionally, dental phobia or anxiety can lead to significant distress (Appukuttan, 2016), resulting in avoidance of dental treatment and missed appointments, which negatively impacts retention (Hoffmann et al., 2022).

In the service provider context, employee interpersonal performance is crucial for ensuring patient retention. As a highly service-oriented discipline, dentistry requires providers to build trust and engage effectively with patients. In addition to delivering quality care and demonstrating professional competence, patients highly value clear explanations of treatments, active involvement in treatment decisions, and the dentist's attentiveness and empathy (Makarem & Coe, 2014). Strong interpersonal skills significantly enhance patient retention by fostering positive relationships and patient satisfaction.

Furthermore, the consistency of patient-provider interactions also influence retention. Regular and reliable communication nurtures a sense of commitment that supports retention, aligning with Grönroos' (2000) findings that past retention levels positively impact current retention, as cited by Onyeaso & Adalikwu (2008). Interestingly, patients previously retained are more likely to remain engaged and retained.

The context of the service encounter also plays a crucial role in retention. For instance, certain dental procedures, such as tooth implants, require the patient's active presence and consistent attendance. This shared responsibility for the outcome of the procedure significantly impacts patient retention.

Amano (2023) conducted a comprehensive examination of factors influencing retention in dental practice through a systematic literature review. He initially sorted and selected articles using specific keywords, covering 32 journal articles from 19 different countries. This process led to the identification of 18 influencing factors, which were categorized into six thematic groups or nodes. To gain further insights, Amano developed a survey to assess dental patients' perceptions of the significance of these factors for retention, using a Likert scale.

In addition to incorporating direct patient feedback, he analyzed contemporary academic research to determine which of the 18 factors had been most extensively studied, reflecting a strong academic and scientific approach on patient retention in dental practice.

Table 1 summarizes Amano's findings, detailing the categorized factors, their frequency of appearance in the selected journal articles, and the mean scores rated by patients on a 5-point Likert scale.

Table 1. Summary of Factors Influencing Retention in Dental Practice. Adapted from Amano (2023).

Category	**Factor**	Number of articles	Mean score ($\bar{X}$)
Relationship	Communication	18	4.47
	Relation	16	4.13
	Trust	12	4.70
	Empathy	9	4.30
		= 55	
Internal factors	Service quality	12	4.45
	Facilities	11	4.22
	Equipment	9	4.39
	Flexibility of appointment	6	4.25
	Accessibility	4	4.43
		= 42	

Professionalism	Time management	10	4.29
	Skillfulness	7	4.70
	Professional manner	7	4.46
	Treatment gentleness	7	4.40
		= 31	
Referral	Word of mouth	9	3.40
	Social Networking Service	3	2.66
		= 12	
Customer's value	Perceived value	10	4.39
		= 10	
Costs	Cost transparency	5	4.46
	Insurance	5	4.14
		= 10	

Note: This table summarizes the factors identified in Amano's (2023) study, categorizing them into thematic groups, indicating their frequency of appearance in selected journal articles, and presenting the mean vote by patients on a 5-point Likert scale.

Amano's (2023) investigation concluded that *skillfulness* and *trust* (See Table 1), were identified as the most important factors influencing patient retention, each receiving the highest mean scores for their perceived importance according to actual patients. *Skillfulness*, which reflects the professional competence expected by patients, was deemed essential for retention. Similarly, *trust* was acknowledged as a critical element in building long-term relationships and contributing to the success of small dental clinics. Other notable factors

included effective *communication* with the patient and *transparency* regarding treatment costs, both of were highlighted as significant contributors to patient retention.

Regarding the articles reviewed by Amano (2023), the majority of the focus was on *communication*, *relational* aspects, *trust*, and *service quality* within dental practice. Conversely, dental patients rated *Social Networking Services* as the least important factor for retention, and it was also the least investigated by scholars among the 18 journals reviewed. Additionally, *Word of mouth* was identified as the second least important factor based on patient survey responses.

The insights gathered from these studies underscore the multifaceted nature of patient retention in dental practices, highlighting the critical roles of professional competence, effective communication, and trust-building. These factors are essential for dental clinics aiming to improve patient retention, enhance the overall patient experience, and achieve better oral healthcare outcomes.

2.5. Elevating Dental Practices with Technology and Machine Learning

This chapter explores how the integration of disruptive technologies and machine learning can revolutionize all aspects of dental operations, from diagnosis and treatment planning to patient care. It introduces key advancements and tools reshaping modern dental practices and highlights the benefits these innovations bring to oral healthcare and strategic clinic management.

In today's dental landscape, competitive private clinics are leveraging cutting-edge techniques, technologies, and data across every aspect of their workflow and environment. This integration enhances efficiency and accuracy in dental practices, enabling clinics to

predict patient behaviors, optimize patient retention strategies, and refine both operational and non-operational processes through advanced analytics. These tools lead to improved outcomes, increased patient satisfaction, and overall better care quality.

Machine Learning (ML), as illustrated in Figure 4, is a subset of Artificial Intelligence (AI), a term coined by Arthur Samuel in 1959. AI encompasses a broad range of techniques designed to enable machines to simulate human intelligence. Within this domain, ML focuses specifically on training algorithms to recognize intrinsic statistical patterns and make predictions based on data. Deep Learning (DL), a further specialization within ML, involves neural networks with many layers that can model complex patterns in large datasets. Figure 4 visually represents this hierarchical relationship, showing how DL is nested within ML, and ML is nested within AI, along with their respective definitions.

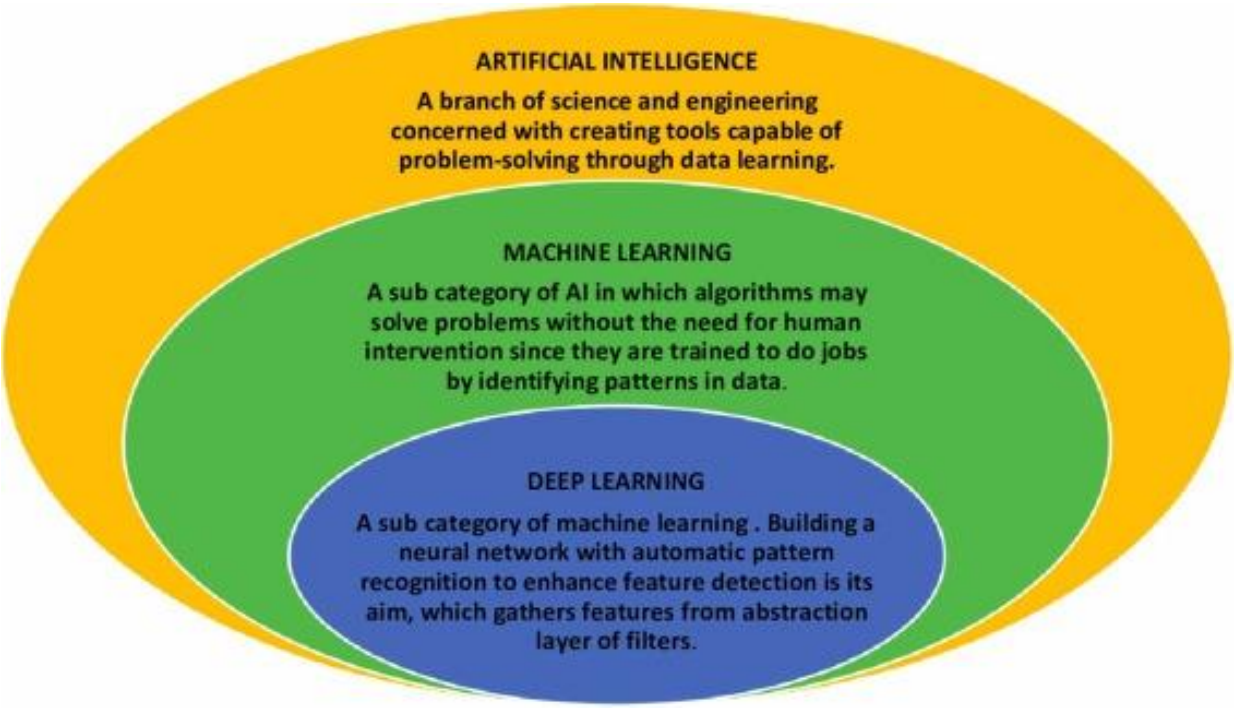

Figure 4. Hierarchical Relationship and Definitions of Artificial Intelligence, Machine Learning, and Deep Learning. Taken from (Vashisht et al., 2024).

The technological framework of AI, ML, and DL is increasingly adopted in oral healthcare. This integration is transforming the dental industry by enhancing diagnostic precision, optimizing treatment planning, and supporting clinical decision-making. According to Vashisht et al. (2024), these advanced technologies are becoming integral to nearly all areas of dental practice, driving improvements in patient care, workflow efficiency, and forecast predictions.

ML algorithms are categorized into supervised and unsupervised learning. Supervised learning includes classification models, which are fundamental for predicting patient retention by analyzing historical retention data. Additionally, feature learning, a specialized part of ML, improve these capabilities by extracting and selecting relevant features, which facilitates efficient and accurate classification tasks (Hastie et al., 2009).

A powerful application of ML is predictive analytics, which identifies the likelihood of future occurrences based on past data. Predictive models “assess risks and opportunities in customer-patient acquisition and retention strategies” (Paul Isson, 2018, p. 117). By leveraging these models, businesses can personalize marketing campaigns, forecast demand, offer targeted recommendations, and create tailored customer experiences. This strategic use of predictive analytics enhances customer satisfaction and improves retention rates (Bharadiya, 2023).

In oral healthcare, predictive analytics powered by AI and ML is increasingly used to assess patients' risk for developing dental conditions like tooth decay, gum disease, and oral cancer. These advanced tools analyze various factors, including lifestyle habits, medical history, and genetic predispositions, to help dentists identify patients who might benefit from preventive measures (Vashisht et al., 2024). Modern predictive analytics use a variety of ML algorithms

to forecast outcomes and make data-driven decisions. Foundational algorithms like Linear Regression and Logistic Regression (LR) offer straightforward data analysis techniques. Non-linear models such as Decision Trees (DT) and Support Vector Machines (SVM) capture more complex relationships in data. Additionally, advanced techniques like Random Forest (RF) and Extreme Gradient Boosting (XGBoost) are utilized to manage large datasets and enhance prediction accuracy through ensemble methods.

Kahurke (2023) highlights the use of ML algorithms in dentistry for refining prediction, diagnosis, prevention, and treatment planning. ML aids in analyzing extensive health data from diagnostic tools such as Intraoral periapical radiographs (IOPA), Orthopantomograms (OPG), X-rays, Cone Beam Computed Tomography (CBCT), Computed tomography scans (CT), RadioVisioGraphy (RVG), and 3D scans.

Figure 5 illustrates how data inputs from the aforementioned technologies, processed through AI, ML, and DL algorithms, can lead to improvements in treatment, diagnosis, and prognosis.

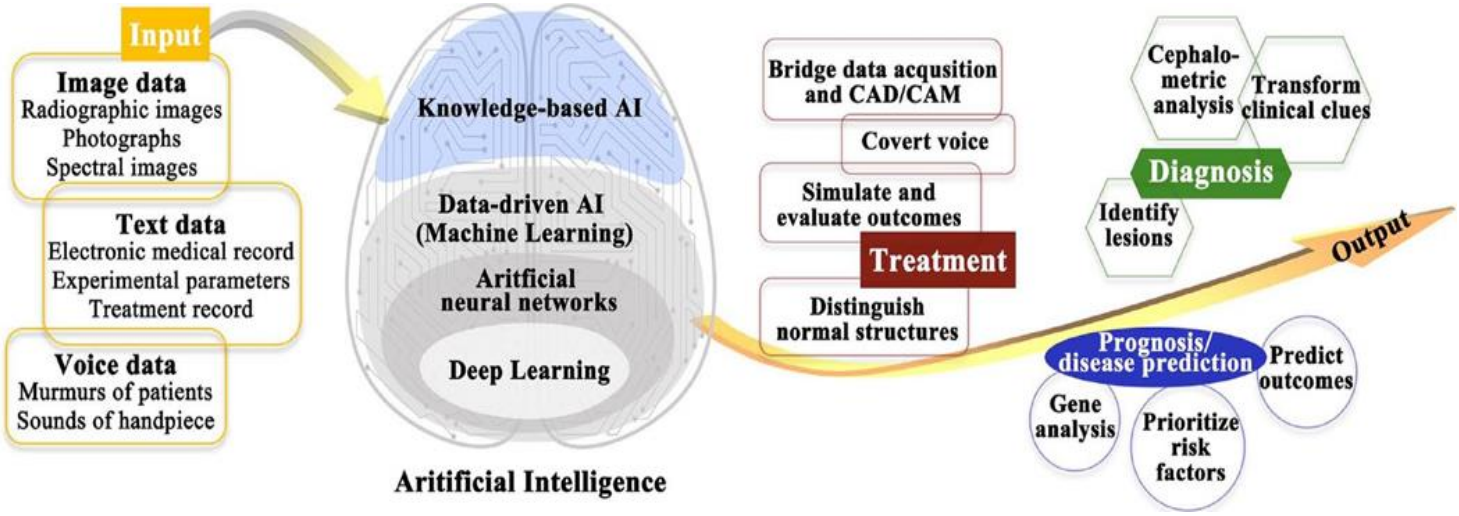

Figure 5. Hierarchical Overview and Major Dental Applications of Artificial Intelligence (AI), Machine Learning (ML), and Deep Learning (DL). Taken from (Shan et al., 2021).

A pivotal development in this area is the concept of "Data Dentistry," as introduced by Schwendicke & Krois (2022). This innovative approach integrates data-driven decision-making into oral healthcare practices. Remarkably, a bibliometric analysis of Artificial Intelligence in dentistry from 2000 to 2023 by Xie et al. (2024), recognized Schwendicke and Krois as the most prolific authors in this field.

"Data Dentistry" leverages vast data from diverse sources, including demographic, social, clinical, environmental, technological, and provider-related information. By effectively utilizing this wealth of data, dental clinics can significantly strengthen patient acquisition and retention strategies, and “eventually facilitate personalized, predictive, preventive, and participatory dentistry” (Schwendicke et al., 2020, p. 769).

In conclusion, adopting advanced technology and machine learning algorithms for decision-making in dental clinics significantly enhances daily operations and patient retention strategies. Implementing these technologies discussed earlier allows clinics to gain deeper insights into patient behavior, anticipate needs, preferences, and future actions, and tailor efforts based on data. This data-driven approach ensures favorable patient retention rates and long-term success in the competitive dental market. By leveraging data-based knowledge over subjective opinions and assumptions, clinics can more effectively focus their actions, resulting in improved patient retention and sustained growth.

3. Methodology

The primary aim of this study is to develop and interpret a classification model to predict the likelihood of prospective patients returning to a dental clinic after their initial contact and visit. To achieve this, various algorithms were evaluated, including both linear and non-linear methods, known for their simplicity, intuitiveness, and ease of interpretation.

The study prioritizes predictive performance over interpretability, acknowledging the trade-off between the two aspects. This decision stems from a thorough review of state-of-the-art methods for predicting patient return rates and retention, which revealed that machine learning ensemble algorithms are increasingly being used (Peiró-Signes et al., 2022) in dentistry (Schwendicke & Marazita, 2022) due to their high predictive performance (Sharma et al., 2022).

For the development and evaluation of the machine learning classifier, historical, demographic, and behavioral data from patients who contacted (n = 1,501) a dental clinic in Valencia, Spain, from December 2021 until December 2023 were utilized. Anonymization of this data was carried out by removing personally identifiable information, such as names and contact details, to ensure compliance with data protection regulations and ethical standards, thus safeguarding patient privacy and confidentiality.

This chapter will detail the methodology used in developing the classification model, including the selection and evaluation of algorithms, and various data handling techniques. It will cover the entire process from the initial sample description and data preprocessing — such as cleaning and transformation— to the selection of features and tuning of the model. Additionally, the chapter will explain how the model's performance was assessed and interpreted, presenting the findings and insights derived from the analysis.

3.1. Sample

For this investigation, a dataset comprising all contacts (n = 1,501) received by a dental clinic in Valencia, Spain, from December 2021 until December 2023 was extracted from the clinic's information system. The points of contact included telephone, website, and physical visits to the dental clinic; however, the dataset does not specify the contact channel used by each potential patient. To address missing information, the dataset was complemented with patient records kindly provided by the clinic's management.

The resulting dataset contains 17 variables that could be grouped into the following categories:

- Demographic Features: Variables such as age; gender; insurance coverage; and ZIP code.
- Contact Reasons and Dates: Variables including the reason for contacting the dental clinic; the date of the first contact; the date of the last contact; and the dates of the first and last visits (if applicable).
- Dental Care Details: Variables related to the dental prescription, including treatment type; category; and price.
- Patient Outcomes: Variables such as the total number of visits made to the dental clinic; the total number of dental treatments prescribed; the total number of treatments completed; the treatment completion percentage (calculated by dividing the total number of treatments completed by the total number of treatments prescribed); and whether the patient returned to the dental clinic after the first visit.

Figure 6 provides a visual representation of the patient return variable, also referred to as *Patient Retention Status*, which serves as the predicted outcome in this study.

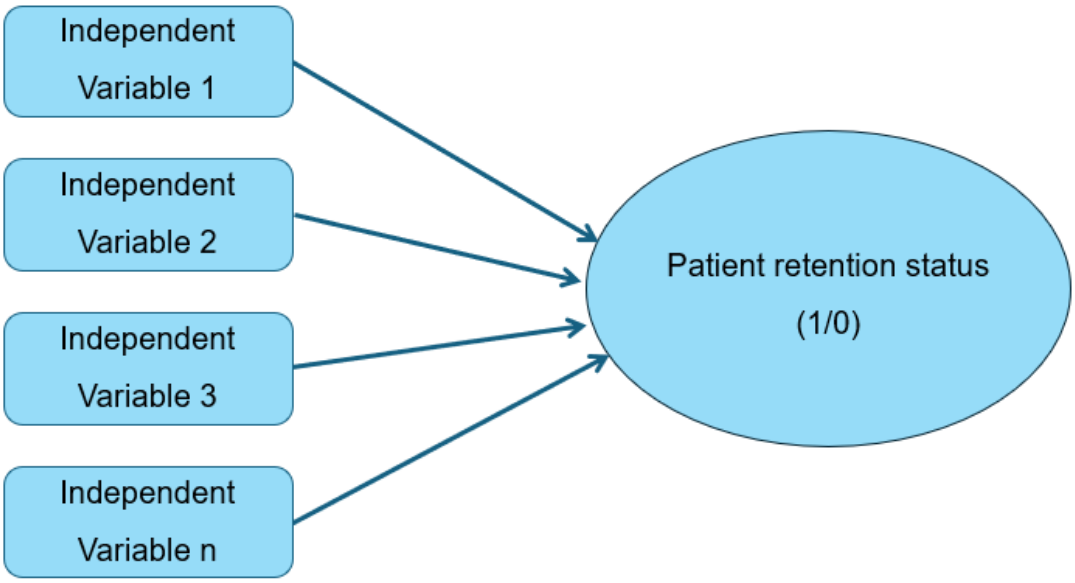

Figure 6. Visualization of Patient Retention Status Based on Variables.

Given that this investigation aims to develop a classification model for predicting whether patients will return to the clinic after their initial visit, it is important to understand how the '*Patient Retention Status*' variable was defined based on historical data from the dental clinic. Specifically, this involves clarifying how patients were categorized as retained or not retained based on their visit history.

Contacts made without a subsequent visit to the clinic were classified as not retained. Additionally, patients who visited the dental clinic once and either did not return for a follow-up visit within six months or did not make any further contact with the clinic during that period —regardless of the number of treatments received during their initial visit— are also categorized as not retained.

Moreover, new patients who visited the dental clinic only once and were registered between June 2023 and December 2023 were excluded from the dataset. This exclusion was necessary

due to the impossibility to provide them with a value in the *Patient Retention Status* given the limited time frame for evaluation.

Figure 7 illustrates the decision-making process for assigning values to the *Patient Retention Status* variable, showing how each contact received by the clinic was categorized as either retained or not.

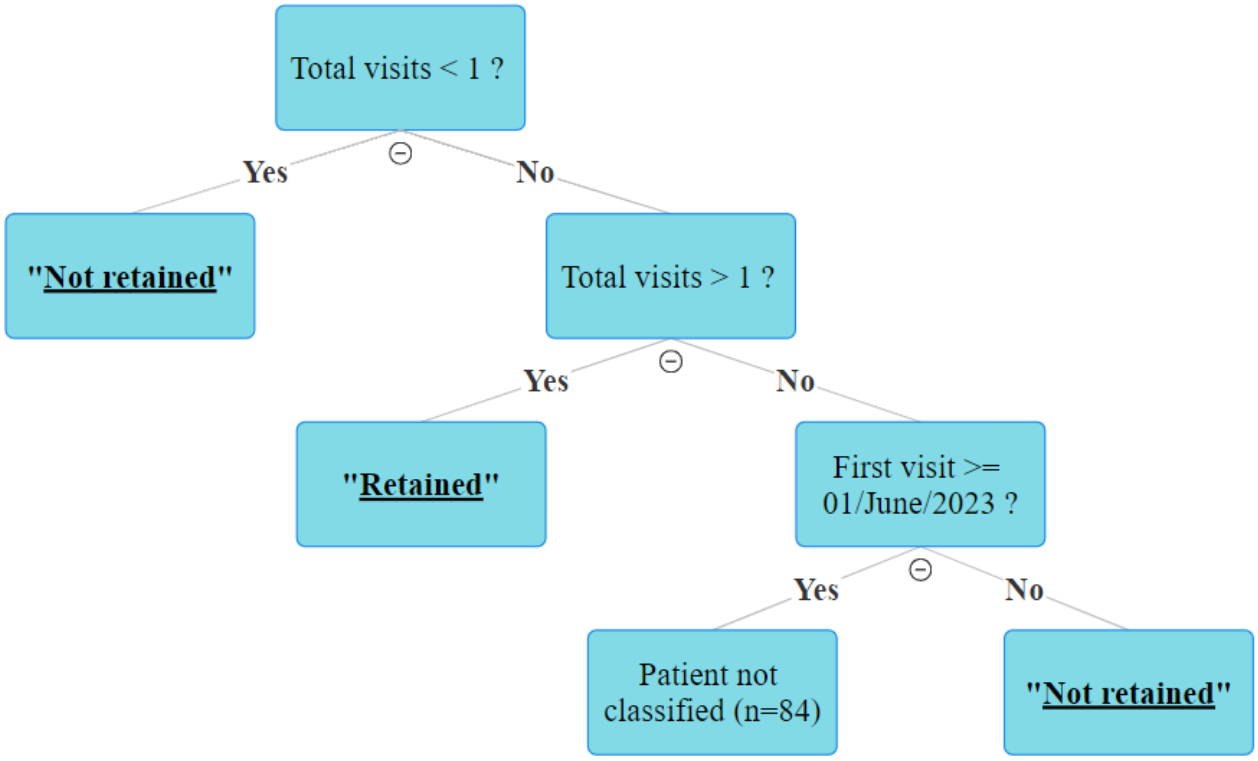

Figure 7. Decision-making process used to assign a value to the Retention Status.

Table 2 displays the distribution of each variable in the dataset, including its final type, absolute frequency, and relative frequency. This table provides a comprehensive overview of the data used in the classification model and aids in understanding the characteristics of the dataset.

Table 2. Distribution of Variables, Final Data Types, and Frequencies.

Variable – Data type	Absolute Frequency (Missing)	Relative Frequency
X1. Gender (n=1,501) – Categorical		
Female	831	55%
Male	670	45%
X2. Insurance coverage (n=1,501) – Categorical		
No insurance	1,244	83%
Private insurance	257	17%
X3. Age (n=930) – Categorical		
Missing	(571)	38%
Under 18 years	137	9%
18-29 years	126	9%
30-39 years	139	9%
40-49 years	244	16%
50-59 years	139	9%
60-69 years	78	5%
70 years and over	67	5%
X4. Location by ZIP code (n=822) – Categorical		
Missing	(679)	45%
Outside 46980	248	17%
46980 (Neighbor)	574	38%
X5. Contact reason (n=1,363) – Categorical		
Missing	(138)	9%
Checkup	935	62%
Specific Treatment	301	20%

Emergency	127	9%
X6. First contact (n=1,501) – Categorical		
Under 6 months ago (>= 01/07/2023)	154	10%
Between 6 and 12 months ago [01/01/2023 - 30/06/2023]	815	54%
Over a year ago (< 01/01/2023)	532	36%
X7. First visit (n=1,332) – Categorical		
Missing	(169)	11%
Under 6 months ago (>= 01/07/2023)	145	10%
Between 6 and 12 months ago [01/01/2023 - 30/06/2023]	737	49%
Over a year ago (< 01/01/2023)	450	30%
X8. Treatment type (n=1,301) – Categorical		
Missing	(200)	13%
Non-surgical treatments	1,048	70%
Surgical treatments	118	8%
Both surgical and non-surgical dental treatments	135	9%
X9. Treatment category (n=1,301) – Categorical		
Missing	(200)	13%
Preventive	513	34%
Orthodontic	158	11%
Restorative	300	20%
Cosmetic	64	4%
Other Surgical Procedures	58	4%
Other Dental Treatments	208	14%
X10. Treatment price (n=1,301) – Categorical		
Missing	(200)	13%
Very low (≤ 60 €)	352	24%
Low [61 € - 150 €]	210	14%
Moderate [151 € - 450 €]	329	22%

High [451 € - 1100 €]	110	7%
Very high (> 1101 €)	300	20%
X11. Last contact (n=1,501) – Categorical		
Under 6 months ago (>= 01/07/2023)	146	10%
Between 6 and 12 months ago [01/01/2023 - 30/06/2023]	225	15%
Over a year ago (< 01/01/2023)	1,130	75%
X12. Last visit (n=1,332) – Categorical		
Missing	(169)	11%
Under 6 months ago (>= 01/07/2023)	119	8%
Between 6 and 12 months ago [01/01/2023 - 30/06/2023]	187	12%
Over a year ago (< 01/01/2023)	1,026	69%
X13. Total visits (n=1,332) – Numerical		
None (0)	169	11%
# (Numerical elements)	1,332	89%
X14. Total dental treatments prescribed (n=1,301) – Numerical		
None (0)	200	13%
# (Numerical elements)	1,301	87%
X15. Total dental treatments completed (n=1,033) – Numerical		
None (0)	468	31%
# (Numerical elements)	1,033	69%
X16. Treatment completion percentage (n=1,033) – Numerical		
None (0%)	468	31%
# (Numerical elements)	1,033	69%
X17. Patient Retention Status (n=1,501) – Categorical		
Patient not retained	695	46%
Patient retained	806	54%

Note: Due to missing data, not all variables add up to the total sample size of 1,501. The variables X13, X14, X15, and X16 are continuous and the full list of data was not included to save space.

3.2. Data preprocessing

3.2.1. Data cleaning.

The data cleaning stage, which includes handling and imputing missing data to ensure the dataset is as clean as possible for analysis, was necessary due to the presence of missing values in some categorical variables used in this study.

The literature presents various methods for treating missing data, including discarding the observations with missing values or replacing the missing values with statistical measures like mean, median, or mode. The former approach was initially contemplated; however, this led to a significant data loss, reducing the dataset to 562 contacts from the original 1,501, which also compromised the model's training size.

Since discarding categorical predictors would lead to the loss of valuable insights into factors affecting patient return rates, an alternative approach was sought.

Considering the potential utilization of tree-based models in this study, a superior approach emerged. As proposed by Hastie et al. (2009), a pragmatic solution entails creating "a new category named 'missing' to accommodate missing values" (p. 311). See Table 2, which already has the category of missing values as *Missing* along with the other variables in the dataset.

3.2.2. Data transformation.

Two demographic features (*Age* and *Zip code*) were initially transformed from numerical to categorical forms. *Age* was categorized into seven distinct age groups, while the *ZIP code* was classified based on its proximity to the dental clinic's ZIP code, determining whether it was considered a neighboring ZIP code or not. This categorization aimed to capture

geographic proximity as a potential factor affecting patients' willingness to return to the dental clinic.

The contact and visit dates were also transformed into categorical values to simplify the analysis of their temporal attributes. Additionally, the price of the prescribed dental treatments was transformed from numerical form into categorical form using a Likert scale.

The *treatment type* variable was created as a categorical variable indicating whether the prescribed dental budget required surgery, did not require surgery, or required both types of treatment. Six common categories of treatments performed in dental clinics were created for the *treatment category* variable, with no specific order. Each entry in the dataset was placed in one of these categories. If a budget included more than one category, the value reflected the primary treatment the patient was seeking or the most expensive one.

Binomial features, such as gender (female/male), insurance coverage (yes/no), and patient retention status (retained/not retained), were all transformed into binary form to prepare the dataset for statistical and data analysis.

The final step in the data transformation process involved applying the one-hot encoding technique to further prepare the dataset for machine learning modeling. This technique "transforms each categorical variable with n categories into n dummy variables with a value 1 (hot) if the sample case belongs to the suggested category and 0 (cold) otherwise" (Peiró-Signes et al., 2022, p.6). Géron (2019) and others recommend this process to handle categorical variables with multiple categories where there is no inherent relationship between the categories or ordinality. One-hot encoding was applied to the following variables in this study: *Age, Zip code, Contact reason, First contact, Last contact, First visit, Last visit, Treatment category, Treatment type*, and *Treatment price*. This process generated a binary

attribute for each category within the selected variables (e.g., Age_1 (1/0), Age_2 (1/0), Treatment price_3 (1/0), etc.).

Upon completion, the transformed dataset resulted in a total of 53 variables, including four continuous variables: the *total number of visits made*; the *total number of dental treatments prescribed*; the *total number of treatments completed*; and the *treatment completion percentage*.

3.2.3. Algorithm selection.

This chapter aims to determine the machine learning algorithm with the highest classification efficiency, considering factors such as predictive accuracy, interpretability, and compatibility with the categorical variables in the dataset. The objective is to use the input variables to predict an outcome: whether a dental patient who initially contacted and visited a dental clinic will return or not, expressed as (1/0).

Python code was utilized to evaluate various models, ranging from traditional linear models to more complex ensemble methods. The versatility of Python, along with libraries such as pandas and scikit-learn, played an instrumental role throughout the methodology of this study. These tools facilitated the comparison and evaluation of the algorithms under consideration.

The algorithms discussed in the literature review, including Logistic Regression, Decision Trees, Support Vector Machines, Random Forest, and Extreme Gradient Boosting, were included in the analysis, alongside additional algorithms to ensure a comprehensive evaluation.

Initially, the study explored the feasibility of employing Logistic Regression, a well-established algorithm for predicting binary outcomes when the dependent variable is binary

(Makarem & Coe, 2014). However, rather than relying solely on Logistic Regression, the decision was made to comprehensively evaluate the performance of various algorithms and assess their suitability for the dataset, particularly in handling the categorical variables.

To evaluate the performance of the algorithms and estimate prediction error, cross-validation was employed, a commonly used method in machine learning (Hastie et al., 2009). As explained by Hastie et al. (2009), "To finesse the problem, this technique uses part of the available data to fit models and a different part to test it" (p. 241). This method partitions the dataset into K equal-sized folds while preserving the percentage of samples for each class in every fold. The scores from all the folds are then aggregated to provide an overall performance estimate of the algorithms tested.

Repeated stratified k-fold validation with multiple repeats was used to enhance the reliability of the evaluation process. Hastie et al. (2009) demonstrated that with K = 5, this approach exhibits lower variance, though bias may arise depending on the learning method's sensitivity to training set size. The classifier's performance improves as the training set size increases to 100 or 200 observations, suggesting that cross-validation would not be significantly biased.

Given the modest size of the dental clinic dataset (1,501 contacts), a 5-fold cross-validation approach with five repeats was chosen to ensure robust estimation of algorithms performance while balancing bias and variance. The evaluation metrics, including mean accuracy and standard deviation, provided valuable insights into each algorithm's predictive capabilities. These metrics served as the basis for identifying the most suitable machine learning algorithm for achieving the research objectives. Table 3 summarizes each algorithm's mean and standard deviation scores.

Table 3. Algorithm Performance over the Total Sample.

Algorithm	Mean accuracy (Std. Deviation)
Linear Discriminant Analysis (LDA)	0.748 (0.024)
Logistic Regression (LR)	0.769 (0.023)
Decision Trees (DT)	0.731 (0.023)
Quadratic Discriminant Analysis (QDA)	0.669 (0.015)
Gaussian Naïve Bayes (GNB)	0.666 (0.016)
Support Vector Machine (SVM)	0.790 (0.023)
K-Nearest Neighbors (KNN)	0.749 (0.024)
Bagged Decision Trees (BAG)	0.767 (0.021)
Random Forest (RF)	0.778 (0.018)
Extra Trees Classifier (ET)	0.745 (0.024)
Extreme Gradient Boosting (XGBoost)	0.770 (0.018)

Note: This scores were obtained using cross-validation in Python with the dental clinic's dataset. The standard deviation (Std. Deviation) represents the variability of scores across the folds.

After implementing a cross-validation approach with K=5 folds and five repeats on a dataset comprising 1,501 entries, and calculating the mean performance score and standard deviation for each algorithm (as shown in Table 3), it became evident that two non-linear algorithms, Logistic Regression (LR) and Support Vector Machine (SVM), demonstrated decent performance in the classification task. However, given the study's emphasis on predictive

performance over interpretability, the decision was made to pursue an ensemble algorithm for predicting the likelihood of a prospective dental patient returning to the clinic.

Although Random Forest (RF) demonstrated slightly better efficiency in predicting the outcome during cross-validation, it was determined that the Extreme Gradient Boosting (XGBoost) algorithm was more suitable for advancing the research objectives. As shown in Figure 8, XGBoost is at the forefront of tree-based algorithm evolution as of 2020 (Espinosa-Zúñiga, 2020). Widely recognized by data scientists, XGBoost consistently outperforms many other algorithms and provides superior customization and tuning capabilities (Chen & Guestrin, 2016; Espinosa-Zúñiga, 2020).

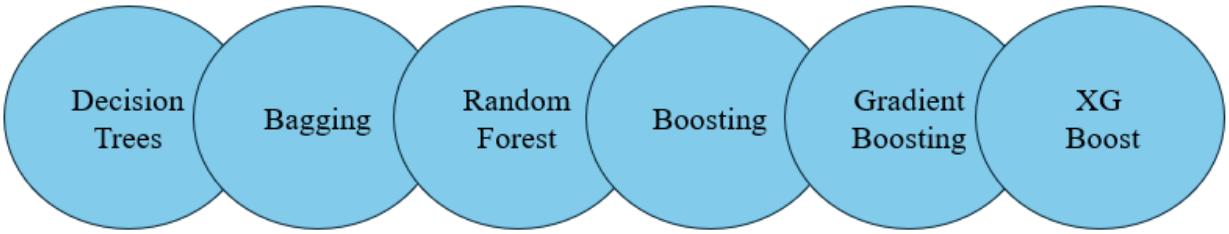

Figure 8. Evolution of Tree-Based Algorithms. Adapted from Espinosa-Zúñiga (2020).

XGBoost is a machine learning algorithm that constructs decision trees sequentially to predict the dependent variable. Each tree is evaluated by posing if-then-else true/false questions to determine the minimal number of splits needed to accurately predict outcomes (NVIDIA Corporation, n.d.). Subsequent trees in XGBoost are designed to rectify the errors of their predecessors trees through an iterative process, which helps reduce both bias and variance, thereby enhancing predictive accuracy (Kumar, 2023).

XGBoost also efficiently manages binary categorical data, making it well-suited for the dataset variables that were transformed into dummy categorical form in the previous chapter. Additionally, its extensive parameter customization allows for fine-tuning and improves performance without the need for further data transformations. This flexibility is advantageous for interpreting the impact of the variables within the model.

3.2.4. Feature selection.

Feature selection is the process of narrowing down input data dimensionality by selecting a relevant subset of features to focus on while ignoring the rest. This reduction in dataset dimensionality addresses two aspects: optimizing the learning process for precise classifiers and identifying the most significant features in the model, which may offer deeper insights into the classification problem (Nilsson et al., 2007).

To initiate feature selection, an Analysis of Variance (ANOVA) test was performed to assess the relationship between the target variable, *patient retention status*, and the four continuous variables. The ANOVA test produced the F and p values, which are critical for evaluating the statistical significance of the features. The analysis revealed that the feature *Total.visits* had a deterministic relationship with the target variable *patient retention status*. Since including this feature would not add new information and might introduce multicollinearity issues, it was excluded from the model, leaving it with 52 variables.

Subsequently, the performance of the selected algorithm, XGBoost, was assessed. The dataset was divided into training and testing sets using Python, with two-thirds of the data allocated for training and one-third reserved for validation.

With the default settings in the XGBoost library, the initial execution of the algorithm achieved an accuracy of 77.08%. The confusion matrix (see Figure 9) showed 107 True

Positives (TP), 32 False Positives (FP), 37 False Negatives (FN), and 125 True Negatives (TN), providing an overview of the model's classification performance in predicting whether a dental patient who initially contacted and visited a dental clinic would return or not.

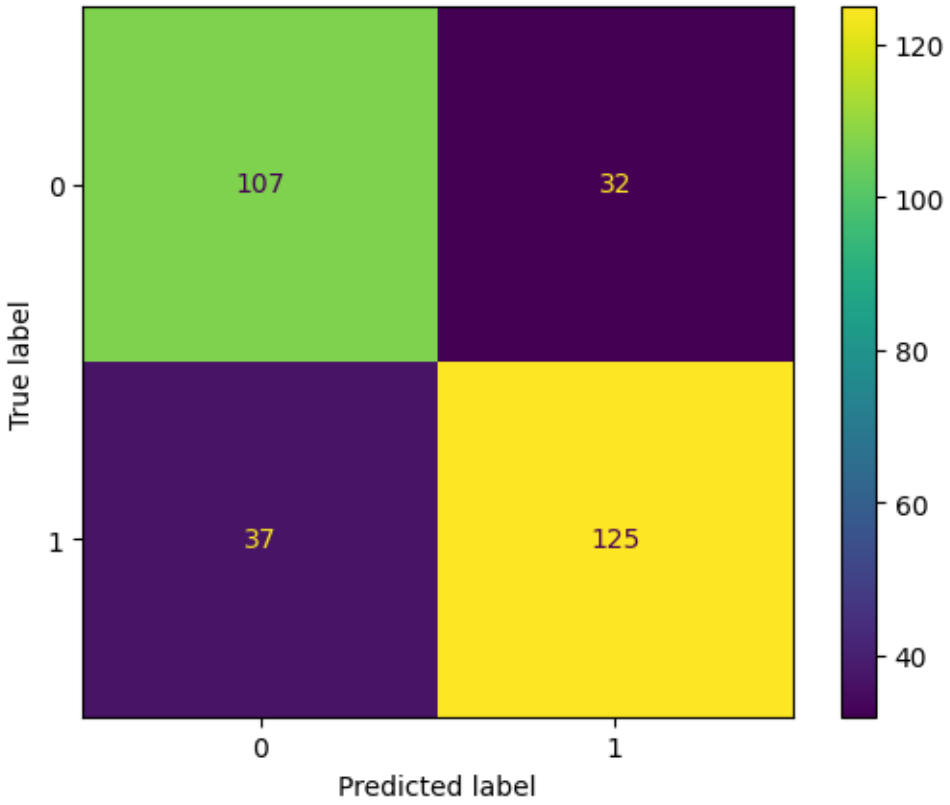

Figure 9. Confusion Matrix for the initial XGBoost Classification Model.

Despite efforts to optimize the dataset, the number of predictors remained substantial, particularly after applying one-hot encoding. Reducing the feature set to a manageable size is beneficial for improving model performance and interpretability. Various methods can achieve this, including wrapper methods, which involve a search process to identify the subset of predictors that yield the best results when included in the model. This iterative approach evaluates multiple feature combinations to find the optimal set that maximizes model performance (Kuhn & Johnson, 2013).

One effective wrapper feature selection method is BorutaShap, which combines the Boruta algorithm and SHapley Additive exPlanations (SHAP) values (Keany, 2020). Boruta identifies features that hold meaningful predictive value by iteratively discarding those deemed less statistically significant. The process begins by creating shadow features — duplicate, scrambled versions of the dataset's features— to serve as a reference point to distinguish between relevant and irrelevant features. Boruta then compares each feature's importance score to the maximum score of the shadow features, establishing a significance threshold (Kursa & Rudnicki, 2010). Features surpassing this threshold are retained, resulting in the minimal subset of the most influential predictors.

Meanwhile, SHAP assigns Shapley values to each feature. Shapley values, derived from cooperative game theory, quantify each feature's contribution to model predictions by measuring its impact on the model's output, thus aiding in the visual representation and interpretation of the model's performance (Rodríguez-Pérez & Bajorath, 2020).

In applying BorutaShap to the dental dataset, six attributes were identified as important, while 45 were deemed unimportant (excluding the predicted variable). This reduction in the number of features simplifies the model and makes it more manageable.

A second execution of the XGBoost algorithm, using only the six variables confirmed as important and the *patient retention status* variable as the predicted outcome, achieved an accuracy of 74.75%. Although this represents a slight decrease from the initial model's accuracy of 77.08%, the substantial reduction in variables —from 51 to 6—results in a model easier to understand and manage. The reduced complexity enhances interpretability, while the model's performance remains robust despite the smaller feature set. This simplified model continues to effectively predict the likelihood of patients returning to the dental clinic.

The final subset of features identified by BorutaShap can be confidently utilized for subsequent predictive modeling and tuning tasks, ensuring that only the most relevant and impactful variables are considered. Figure 10 illustrates the mean SHAP value of the features and their average impact on the model output, while Table 4 offers a detailed description of these features, including their SHAP importance.

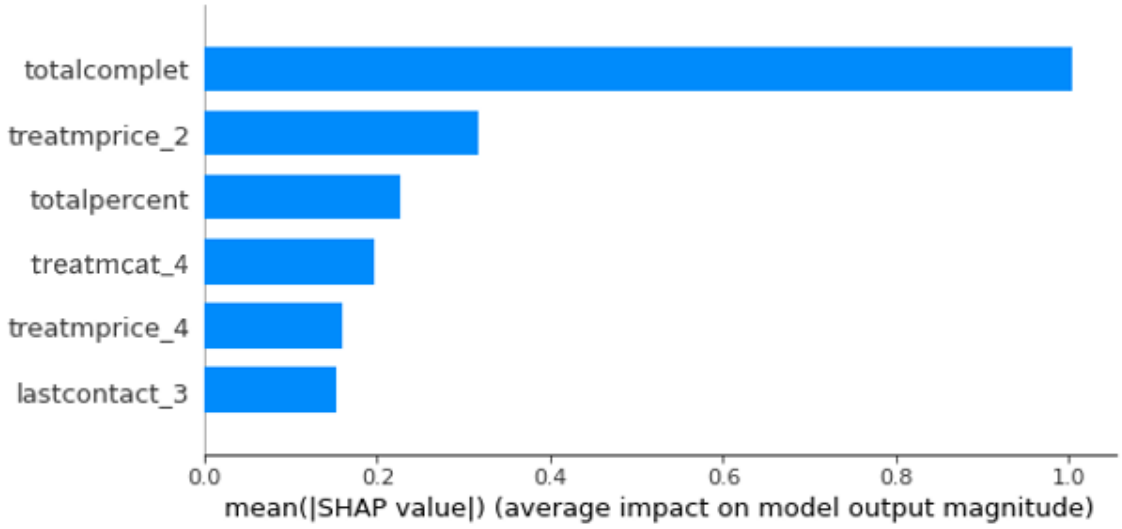

Figure 10. Features Mean SHAP Values and Average Impact on Model Output.

Table 4. Description and SHAP importance of Features Identified by BorutaShap.

Feature	SHAP importance	Description
1. Totalcomplet	1.003371	"*Totalcomplet*" is a discrete variable representing the total count of dental treatments completed by a patient since their first visit to the dental clinic, corresponding to the original **X15** variable.

2. Treatmprice_2	0.31641	"*Treatmprice_2*" is a category within the Treatment Price variable (**X10**), indicating that the prescribed treatments had a price of 60 euros or less.
3. Totalpercent	0.2259	"*Totalpercent*" is a continuous variable representing the percentage of treatment completion relative to the total prescribed, corresponding to the original **X16** variable.
4. Treatmcat_4	0.197084	"*Treatmcat_4*" is a category within the Treatment Category variable (**X9**) , indicating that the prescribed budget primarily included restorative treatments such as implants, prostheses, bridges, and crowns.
5. Treatmprice_4	0.15894	"*Treatmprice_4*" is a category within the Treatment Price variable (**X10**), indicating that the prescribed treatments had a price ranging from 151 euros and 450 euros.
6. Lastcontact_3	0.151735	*"Lastcontact_3"* is a category within the Last Contact variable (**X11**), indicating that the most recent recorded contact with the patient occurred more than a year ago, meaning before 01/01/2023.

Note: SHAP importance values represent the contribution of each feature to the model's predictions, with higher values indicating a greater influence on the model's output. The elements in parentheses (e.g., X10) correspond to the order of variables in the dataset as presented in Table 2.

3.3. Tuning the model

Achieving high performance in machine learning models hinges on several factors, with hyperparameter tuning being one of the most impactful. Hyperparameters are the vital settings that govern a model's behavior. Model tuning, which involves adjusting and determining the optimal learning task parameters, can significantly improve both the model's effectiveness and its predictive accuracy.

In addition to hyperparameter tuning, various other elements play important roles in influencing a model's performance. According to Ilemobayo et al. (2024), the quality of the data is paramount; it must accurately represent the problem domain to provide the model with the correct information to learn from. Data preprocessing steps, such as cleaning, normalization and feature selection, further enhance the model's performance. Equally important is the choice of the algorithm, as different algorithms have distinct strengths and are suited to different types of problems.

In this study, XGBoost was chosen for its high customizability and robust performance in classification tasks. The algorithm, built on the foundation of decision trees, offers a wide range of hyperparameters that facilitate fine-tuning. This capability allows for significant improvements in model performance beyond the default configurations (Kumar, 2023). Initially, the model achieved an accuracy of 77.08% without any hyperparameter optimization or additional refinement techniques. Following this, a revised model was developed using the six relevant attributes identified in the feature selection chapter. Although this updated model achieved a slightly lower accuracy of 74.75%, it benefited from reduced complexity and enhanced interpretability, all while maintaining strong performance with a smaller feature set.

With the revised model achieving an accuracy of 74.75%, there remains potential for further performance improvements through hyperparameter tuning. Given the binary nature of this investigation, which focuses on classifying patients as either returners or non-returners, the model was configured to use ′binary:logistic′ as its objective function. To evaluate the model's effectiveness, the Logarithmic Loss (′logloss′) metric was used, as it provides a measure of the accuracy of probabilistic predictions.

To enhance the model's performance, a grid search technique was employed to optimize the hyperparameters of the XGBoost algorithm. Grid search is described by Ilemobayo et al. (2024) as a "brute-force method" that exhaustively explores a predefined set of hyperparameters values to find the combination that yields the best performance (p. 391). This approach is more straightforward and easier to implement compared to random search.

The first hyperparameter adjusted was ′early_stopping_rounds′, which was set to 10 trees. This parameter controls how many additional trees the model will add before halting training if no improvement is observed in the evaluation metric. While XGBoost defaults to training up to 100 trees, early stopping helps prevent overfitting by ending the training process early when further tree additions no longer improve the model's performance.

The second hyperparameter considered was ′max_depth′, which determines the maximum depth of the trees in the XGBoost model. By default, XGBoost allows trees to grow without a specified limit, which can lead to excessively deep trees with many splits, increasing the risk of overfitting. The grid search identified that, among the proposed values, the default ′max_depth′ of 6 provided the best cross-validation score for this study. This outcome aligns with recommendations from Wade (2020), indicating that a ′max_depth′ of 6 strikes an effective balance between model complexity and performance.

The third hyperparameter, ′learning_rate′, also known as the shrinkage parameter, controls the contribution of each tree to the final prediction. By default, the XGBoost library sets ′learning_rate′ to 0.3. According to Wade (2020), a lower ′learning_rate′ helps prevent overfitting by reducing the size of the weights carried forward, resulting in a more stable and refined model. The grid search determined that a ′learning_rate′ of 0.1 was optimal for this model, demonstrating that reducing the default value improved performance and achieved better cross-validation results.

The fourth hyperparameter, ′min_child_weight′, dictates the minimum number of samples required in a node before it can be split into child nodes. The default value for min_child_weight in the XGBoost library is 1. A higher ′min_child_weight′ value helps ensure that only nodes with a sufficient number of samples are split, thereby mitigating the risk of overfitting. The grid search identified that a ′min_child_weight′ value of 5 was optimal for this model, which aligns with recommendations from Wade (2020).

The fifth and sixth hyperparameters, ′subsample′ and ′colsample_bytree′, were both optimized to a value of 0.8. By default, XGBoost sets both of these parameters to 1. The ′subsample′ parameter controls the fraction of training instances (rows) used for each boosting round. Setting ′subsample′ to 0.8 means that 80% of the training data is used for each iteration, which helps to mitigate overfitting by preventing the model from relying too heavily on any single subset of the data. Similarly, ′colsample_bytree′ specifies the fraction of features (columns) to be randomly selected for each tree. A value of 0.8 limits the number of features used for each tree, which helps reduce variance and further mitigate overfitting. Both parameters were optimized through grid search to enhance the model's ability to generalize to new data.

The final two hyperparameters considered were ′gamma′ and ′reg_alpha′. By default, both ′gamma′ and ′reg_alpha′ are set to 0 in XGBoost. According to Wade (2020) in *Hands-On Gradient Boosting with XGBoost and scikit-learn*, the ′gamma′ parameter sets a threshold that nodes must surpass before making further splits according to the loss metric. The grid search identified a ′gamma′ value of 0 as optimal, indicating that no additional regularization was needed for improving model performance within the tested range. Similarly, ′reg_alpha′ regulates the strength of regularization applied to the model's weights to help mitigate overfitting. The grid search determined that a ′reg_alpha′ value of 0 was optimal, suggesting that regularization did not enhance model performance within the tested range.

Certain hyperparameters were selected for adjustment in this study, while others, including ′n_estimators′, were maintained at their default values. Adjustments collectively contributed to an improvement in the model's performance, culminating in an accuracy score of 79.07%. Table 5 provides an overview of the tuning process, including the range of values tested for each hyperparameter, the optimal parameter values identified, and the corresponding accuracy evolution. These improvements underscore the effectiveness of the hyperparameter tuning process. Figure 11 illustrates the confusion matrix obtained with the optimized parameters.

Table 5. Hyperparameter Tuning Summary.

Step	Grid search range	Optimal parameter value	Accuracy
Base model	None	Default parameters	77.08%
Simplify model	None	Default parameters	74.75%

Tune ′max_depth′	[3, 4, 5, 6]	Default (max_depth = 6)	78.74%
Tune ′learning_rate′	[0.1, 0.2, 0.3]	Learning_rate = 0.1	78.74%
Tune ′Min_child_weight′	[1, 2, 3, 4, 5]	Min_child = 5	78.74%
Tune ′subsample′ and ′colsample_bytree′	[0.4, 0.5, 0.6, 0.7, 0.8]	Subsample = 0.8 Colsample_bytree = 0.8	79.07%
Tune ′gamma′	[0, 0.1, 0.2, 0.3, 0.4, 0.5]	Default (gamma = 0)	79.07%
Tune ′reg_alpha′	[0, 0.01, 0.02, 0.03]	Default (reg_alpha = 0)	79.07%

Note: Each subsequent hyperparameter adjustment incorporates the default parameters along with the newly tuned values from prior steps.

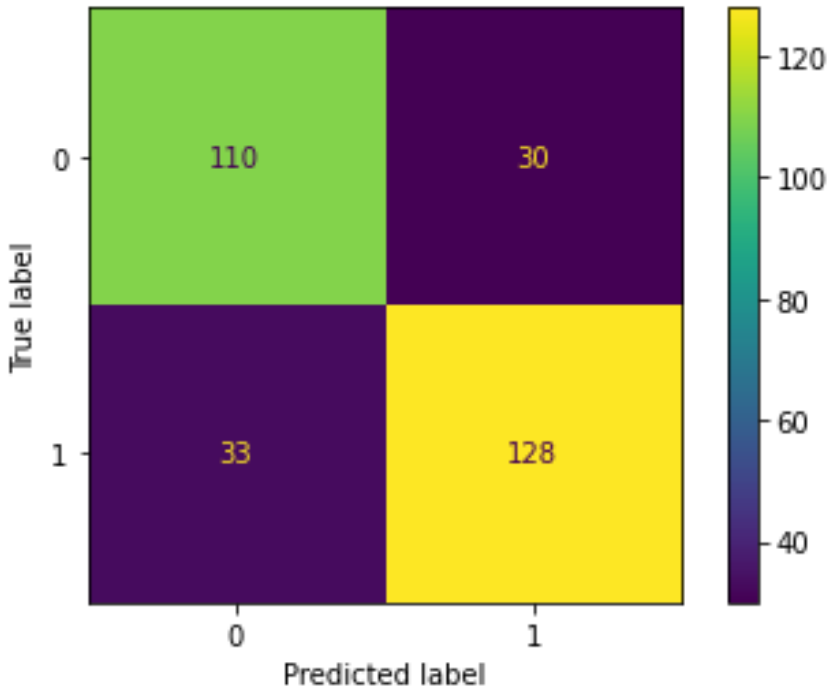

Figure 11. Confusion Matrix for the XGBoost Classification Model with Optimized Hyperparameters.

3.4. Interpreting the model

The complexity inherent in XGBoost, due to its ensemble nature, makes it challenging to understand the model's decision-making process. XGBoost combines multiple decision trees, which complicates the interpretations of how features contribute to the final output.

To tackle this interpretative challenge, the methodology employed in this investigation involved several key steps. Initially, the model's performance was evaluated through accuracy, which is the proportion of correctly predicted cases within the validation dataset. This metric provided an overview of the model's effectiveness in forecasting patient return behavior.

Subsequently, feature importance was then analyzed using the BorutaShap algorithm, which refined the model by retaining only the most impactful features while providing their Shapley importance values. Visualizations including the Tree Explainer, Summary Plot, Dependence Plot, Waterfall Plot, and Plot Tree were generated, each offering valuable insights into the model's predictive framework: The Tree Explainer clarifies individual feature contributions; the Summary Plot illustrates feature importance distributions; the Waterfall Plot details feature impacts on specific predictions; the Dependence Plot shows the relationship between a feature and the predicted outcome; and the Plot Tree reveals the decision rules used by the model.

These visualizations serve to bridge the gap created by XGBoost's complexity, offering a clearer understanding of how features influence predictions and enhancing the interpretability of the model. Detailed discussions and analyses of these visualizations will be presented in the following chapter.

4. Results

4.1. Model results

Upon finalizing the methodology, the optimized classification model achieved an accuracy of 79.07%, reflecting its overall effectiveness in predicting patient return behavior. The performance of the XGBoost model was evaluated using the confusion matrix and classification report.

The confusion matrix of the optimized model, presented in Figure 11, illustrates the distribution of predictions from a validation dataset of 301 instances. In the confusion matrix, the diagonal elements represent the correctly classified cases: True Positives (128 patients predicted to return who indeed do) and True Negatives (110 patients predicted not to return who indeed do not). Off-diagonal elements represent misclassifications: False Positives (30 patients predicted to return but do not) and False Negatives (33 patients predicted not to return but did). The accuracy metric is calculated as the proportion of correctly predicted cases out of the total number of cases (238/301 = 0.7907).

The classification report, detailed in Table 6, provides additional model performance metrics. Precision measures the proportion of correct positive predictions out of all the positive predictions made by the model. For patients predicted to return (Class 1), the precision is 0.81, meaning that 81% of these predictions were accurate. Recall evaluates the proportion of actual positive cases that the model successfully identified. For Class 1, the recall is 0.80, showing that 80% of actual returnees were correctly identified. The F1-score combines precision and recall into a single metric, balancing their trade-off. The F1-score of 0.80 for Class 1 reflects strong performance in both identifying returning patients and minimizing the misclassification of non-returning patients, demonstrating overall effectiveness in the

classification task. Finally, the "support" column indicates the number of actual instances of each class in the validation dataset, with Class 1 comprising 161 instances (53.5% of the dataset), highlighting the distribution of cases that the model evaluated.

In binary classification, a model is often considered effective if it surpasses the baseline accuracy of chance by a significant margin. According to general standards in the field, achieving an accuracy that is at least 25% higher than random chance is indicative of a strong model performance. The optimized XGBoost model proposed in this study, with an accuracy of 79.07%, exceeds this benchmark.

Table 6. Classification Model Performance Report.

Predicted	Precision	Recall	F1-score	Support
Class 0 (Non-returning Patients)	0.77	0.79	0.78	140
Class 1 (Returning Patients)	0.81	0.80	0.80	161
accuracy			0.79	301
macro avg	0.79	0.79	0.79	301
weighted avg	0.79	0.79	0.79	301

Note: The values shown in Table 6 represent percentages.

4.2. Feature importance results

The features selected by the BorutaShap algorithm, displayed in Figure 10, are sorted by increasing importance based on their average Shapley value. These values represent the global significance of each feature in the model, providing insight into their overall contribution to the predictive performance.

Evidently, the '***totalcomplet***' feature emerges as the most significant in predicting the likelihood of patients returning to a dental clinic after their initial contact and visit, with the highest Shapley value of 1.003371. This feature represents the total number of dental treatments a patient has completed since their first visit to the clinic. The high importance of '***totalcomplet***' suggests that the number of treatments a patient has completed is a strong indicator of their propensity to return, reflecting the potential value of maintaining ongoing treatment relationship with the oral healthcare provider.

Following this, the '***treatmprice_2***' and '***totalpercent***' features also exhibit substantial importance, with Shapley values of 0.31641 and 0.2259, respectively. '***Treatmprice_2***' refers to treatments priced at 60 euros or less, as indicated by its category within the Treatment Price variable. This suggests that lower-cost treatments are a significant factor influencing patient return behavior, possibly due to affordability and perceived value. '***Totalpercent***', which represents the percentage of treatments completed relative to the total prescribed, indicates that a higher completion rate is closely associated with patient retention. This reflects the critical role of overall treatment adherence in predicting whether patients will return for future visits.

Additionally, '***treatmcat_4***' and '***treatmprice_4***'—with Shapley values of 0.197084 and 0.15894, respectively—contribute to the model's predictive power. '***treatmcat_4***' represents

a specific category within the Treatment Category variable, indicating that the prescribed treatments primarily included restorative procedures such as implants, prostheses, bridges, and crowns. These treatments often require multiple visits to ensure successful completion, underscoring the importance of ongoing patient engagement. The score for '***treatmcat_4***' emphasizes the role of restorative treatments in driving patient return behavior.

On the other hand, '**treatmprice_4**' refers to a category within the Treatment Price variable, denoting treatments with a cost ranging from 151 to 450 euros. This mid-to-high price range signifies a considerable financial investment for patients, influencing their decision-making regarding follow-up care. The Shapley value for '**treatmprice_4**' highlights how treatment costs impact patient retention, where higher costs may encourage patients to return to complete their care. It could also suggest that patients who are willing to invest in mid-to-high priced treatments may place a high value on their oral health, which can positively influence their commitment to returning to the dental clinic for ongoing care. Together, these features illustrate that both the nature of the treatments and their associated costs play crucial roles in determining patient return rates, with specific treatment types and pricing structures significantly influencing patient decisions to continue their care at the clinic.

Lastly, '**lastcontact_3**', with a Shapley value of 0.151735, highlights the impact of the timing of the last recorded contact with the patient. This feature indicates that if the most recent contact with the patient occurred more than a year ago—prior to 01/01/2023—the likelihood of their returning to the clinic is influenced. When patients have not reached out to the clinic for an extended period, such as a year or more, it may signal a lapse in their ongoing dental care routine or an opportunity for the clinic to re-engage them. Patients who have not contacted the clinic recently might be less inclined to return, as they may feel less reminded

of the importance of their dental health. Conversely, after a significant time away, patients might feel an increased need for a check-up or other dental care, potentially motivating them to reconnect with the clinic. Therefore, understanding the timing of patient-initiated contact can inform strategies to re-engage patients and improve retention rates by addressing gaps in their dental care.

Following the analysis of individual feature importance, the next step is to evaluate the interactions between different features. This is effectively accomplished using the **Summary Plot**, illustrated in Figure 12. To gain deeper insights into the relationship between each feature and the model's predictions, we will also examine the individual **Dependence Plots** for each feature in parallel. Together, these visualizations offer a detailed understanding of how features interact with one another and their collective influence on the model's predictions.

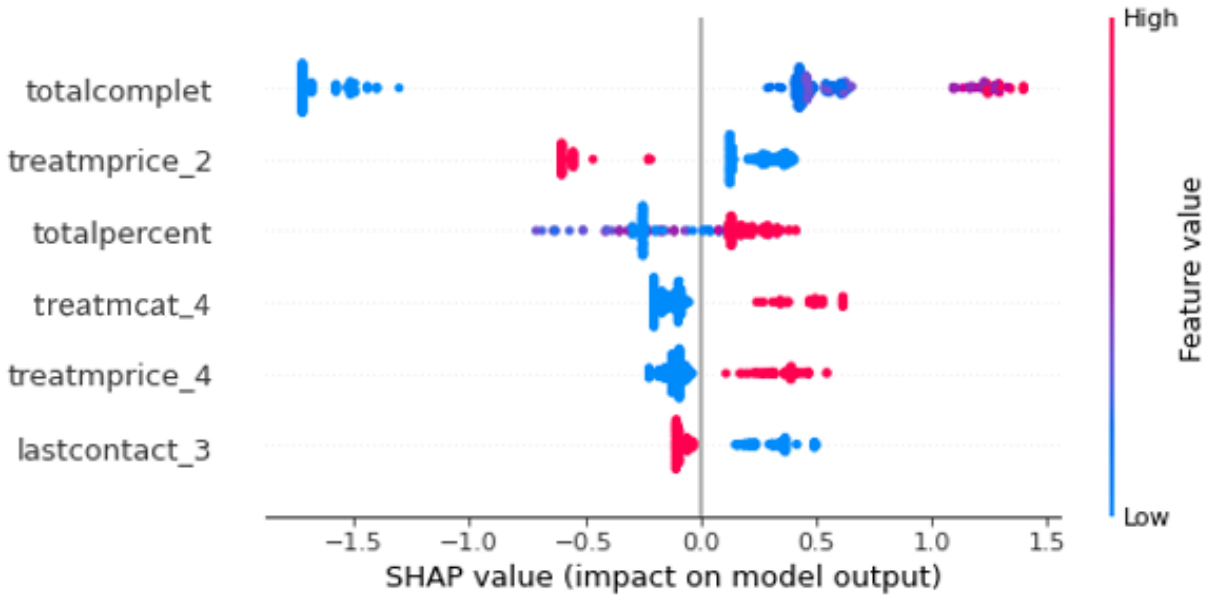

Figure 12. Shapley Summary Plot.

The Summary Plot from the BorutaShap algorithm offers a comprehensive view of the SHAP values' impact and distribution across features, illustrating their influence on the model's predictions. Each feature is ranked by importance on the y-axis, while the x-axis displays the SHAP values, reflecting the contribution of each feature. The color gradient, ranging from blue to red, indicates feature values, with red representing higher values and blue representing lower ones.

The '***totalcomplet***' feature, highlighted as the most significant by BorutaShap, exhibits a wide dispersion of SHAP values with a notable absence of dots in the middle range. This dispersion suggests that '***totalcomplet***' has a significant and variable impact on predictions. On the positive side of the plot, there are low values (blue dots), which tend to increase the prediction, and higher values (red dots) which continue to push the prediction higher. This shift from blue to red in the positive side of the x-axis underscores the feature's strong positive correlation with the prediction and highlights its complex interaction with other features.

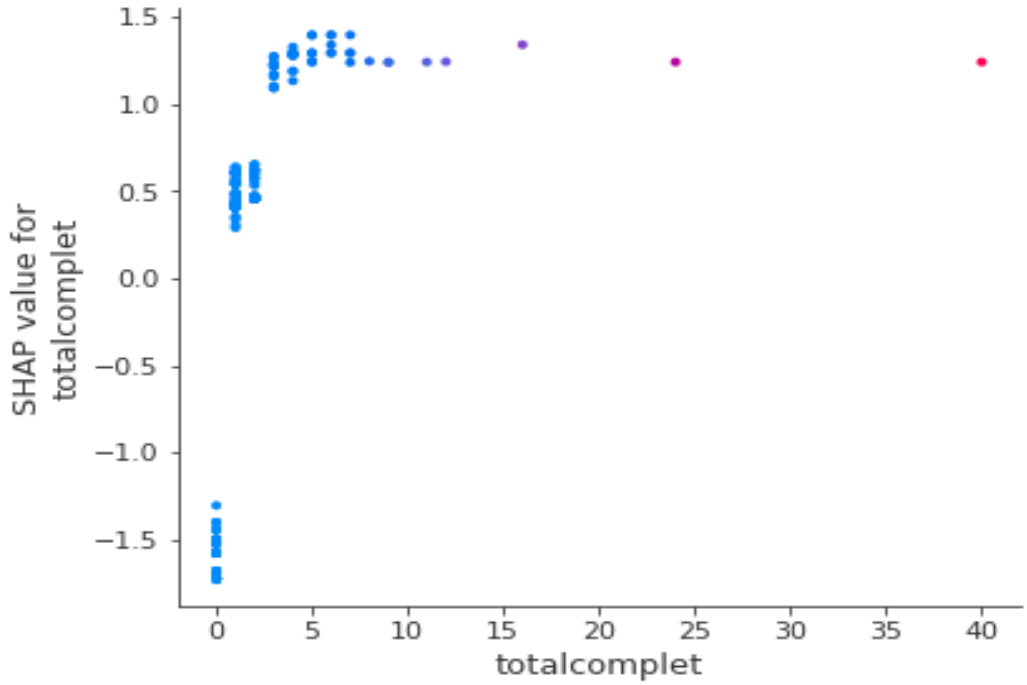

Figure 13. Dependence Plot for the 'totalcomplet' Feature.

Figure 13, the Dependence Plot for '***totalcomplet***', provides additional insights. The plot shows that the majority of dots are blue for '***totalcomplet***' values of 15 or less on the x-axis, indicating that lower percentages of '***totalcomplet***' affect the model's predictions. In the negative section of the plot, blue dots are consistently positioned at a '***totalcomplet***' value of 0, suggesting that values around this percentage may contribute to a decrease in the prediction. Notably, there is a single red dot after the feature hits 40, suggesting an outlier or a unique instance where a high value in the feature has an influential positive effect on the prediction. Additionally, there is one purple dot at 25 on the x-axis, indicating an overlap of blue and red dots—possibly signifying a transitional value where '***totalcomplet***' begins to have varying effects on predictions.

For the '***treatmprice_2***' feature, the Summary Plot (See Figure 12) shows an inverse color pattern compared to '***totalcomplet***'. The SHAP values for '***treatmprice_2***' start with red dots on the negative side of the x-axis and transition to blue dots on the positive side. This pattern indicates that higher treatment prices (closer to the higher end of the 0-60 euros range) are associated with lower SHAP values, meaning they decrease the likelihood of patients returning to the clinic. Conversely, lower treatment prices (closer to 0 euros) are linked to higher SHAP values, which increase the probability of return. The closer distribution of dots for '***treatmprice_2***' suggests that this feature has a more consistent and predictable impact on the model's predictions compared to '***totalcomplet***'. This pattern highlights that more affordable treatments are more likely to encourage patients to continue their care, while higher treatment costs may discourage return visits.

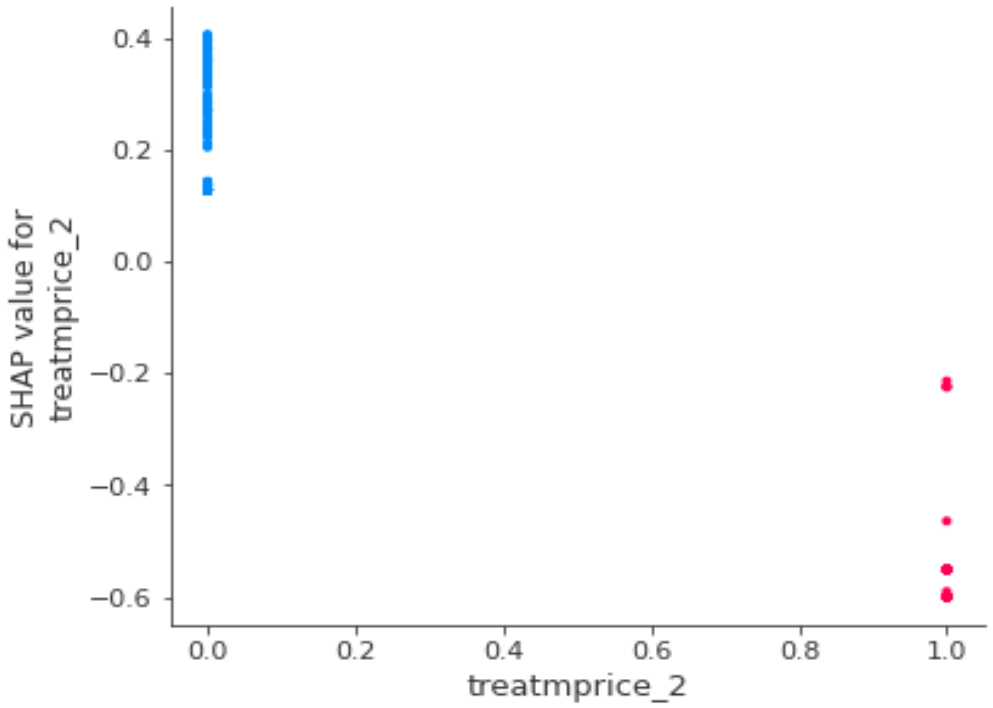

Figure 14. Dependence Plot for the 'treatmprice_2' Feature.

Figure 14 presents the Dependence Plot for '***treatmprice_2***', offering additional insights that align with the patterns observed above. The plot exhibits a binary nature, with dots placed in two distinct lines. The blue dots are on the positive side of SHAP values (ranging from 0.1 to 0.4) and are exclusively for values of 0 on the x-axis. This indicates that treatments priced at the lower end of the normalized range positively influence the likelihood of patients returning to the clinic. In contrast, the red dots appear on the negative side of SHAP values (ranging from -0.6 to -0.2), and are exclusively over the 1.0 value on the x-axis. This suggests that treatments priced at the higher end of the normalized range have a negative impact on the likelihood of return. The binary distribution of the dots highlights the clear distinction in how different pricing levels affect patient behavior, with lower prices encouraging return visits and higher prices discouraging them.

The analysis of the '***totalpercent***' feature in the Summary Plot (See Figure 12), reveals its dual and consistent impact on the model's predictions. The visualization shows that '***totalpercent***' influences the model's predictions across a SHAP value range of approximately -0.75 to 0.5. A concentration of blue dots on the negative side indicates that lower percentages of completed treatments are associated with a reduced likelihood of a positive prediction. Conversely, red dots on the positive side, particularly around the 0.1 mark on the x-axis, indicate that higher '***totalpercent***' values enhance the model's predictions. This suggests that as patients complete a larger percentage of their prescribed treatments, the likelihood of a favorable outcome increases. The strong presence of red dots emphasizes that surpassing certain treatment completion thresholds correlates with positive predictions, highlighting the importance of adherence to prescribed treatment plans.

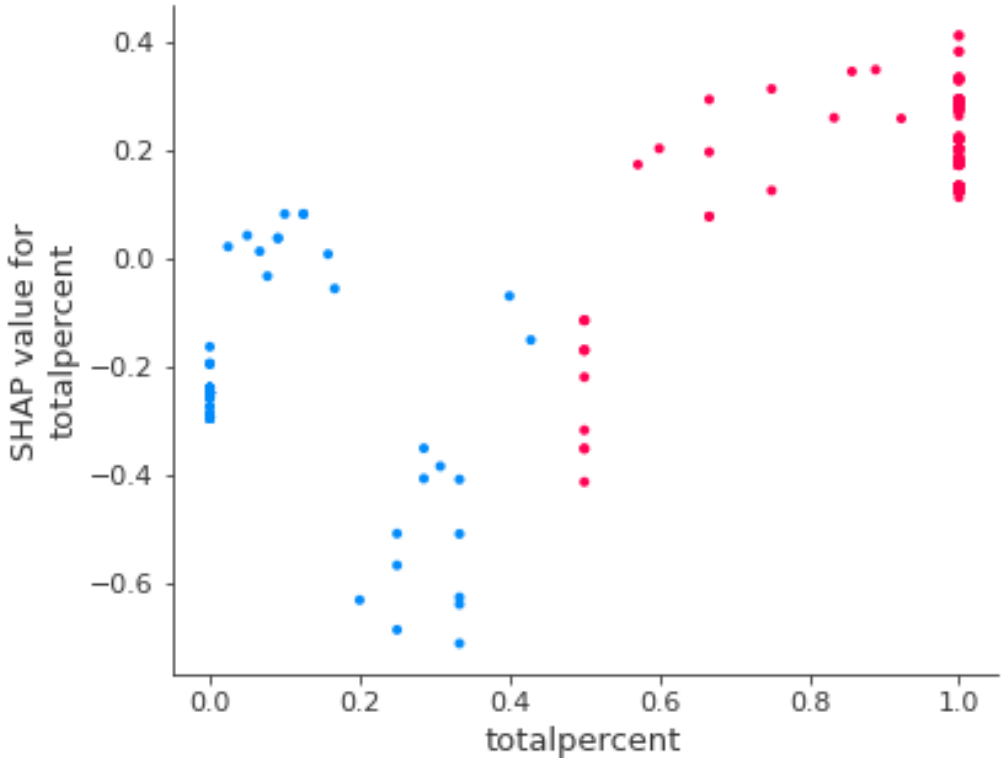

Figure 15. Dependence Plot for the 'totalpercent' Feature.

The Dependence Plot for the '***totalpercent***' feature, shown in Figure 15, provides significant insights into its role within the model. In the visualization, all blue dots are concentrated on the left side of the plot, corresponding to '***totalpercent***' values between 0 and 0.4. These blue dots are predominantly associated with negative SHAP values, suggesting that when the proportion of completed treatments is low, the model predicts a decreased likelihood of the desired outcome, such as patient returning or retention. In the other hand, all red dots are clustered on the right side of the plot, corresponding to '***totalpercent***' values between 0.5 and 1.0. These red dots are linked to positive SHAP values, indicating that as the proportion of completed treatments increases beyond 50%, the model predicts a higher likelihood of a positive outcome. Interestingly, at the exact midpoint of 0.5 on the x-axis, the dots still fall on the negative side of the SHAP values, along with the pattern seen in the blue cluster.

This distribution underscores a critical threshold in the '***totalpercent***' feature's influence: while lower completion rates diminish the likelihood of a positive outcome, surpassing the 50% completion mark significantly enhances it. This finding highlights the importance of patient adherence to treatment plans, as higher completion percentages are strongly correlated with more favorable predictions in the model, and subsequently patients returning.

Regarding the '***treatmcat_4***' and '***treatmprice_4'*** features, the Summary Plot (See Figure 12) reveals that both features exhibit a similar range of SHAP values, indicating that they have a comparable impact on the model's outcomes. The distribution of dots for both '***treatmcat_4***' and '***treatmprice_4***' shows the same color gradient from blue to red. This indicates that both features influence similarly the model's predictions: lower values (blue dots) have a lesser impact, while higher values (red dots) have a greater effect.

The similarities in SHAP values and distributions suggest a potential interaction between '***treatmcat_4***' (which denotes restorative dental treatments) and '***treatmprice_4***' (representing treatments priced between 151 and 450 euros). This interaction implies that changes in one feature may closely correspond to changes in the other, affecting the model's predictions in a related manner. The correlation likely arises from the relationship between the type of treatment and its cost, as certain restorative treatments fall within this price range. Moreover, the similarity between these features could indicate redundancy, meaning they provide overlapping information to the model. This redundancy suggests that consolidating or removing one of these features might simplify the model without significantly impacting its predictive power. The Dependence Plots for both '***treatmcat_4***' (Figure 16) and '***treatmprice_4***' (Figure 17) further corroborate these findings about the similarity of both features.

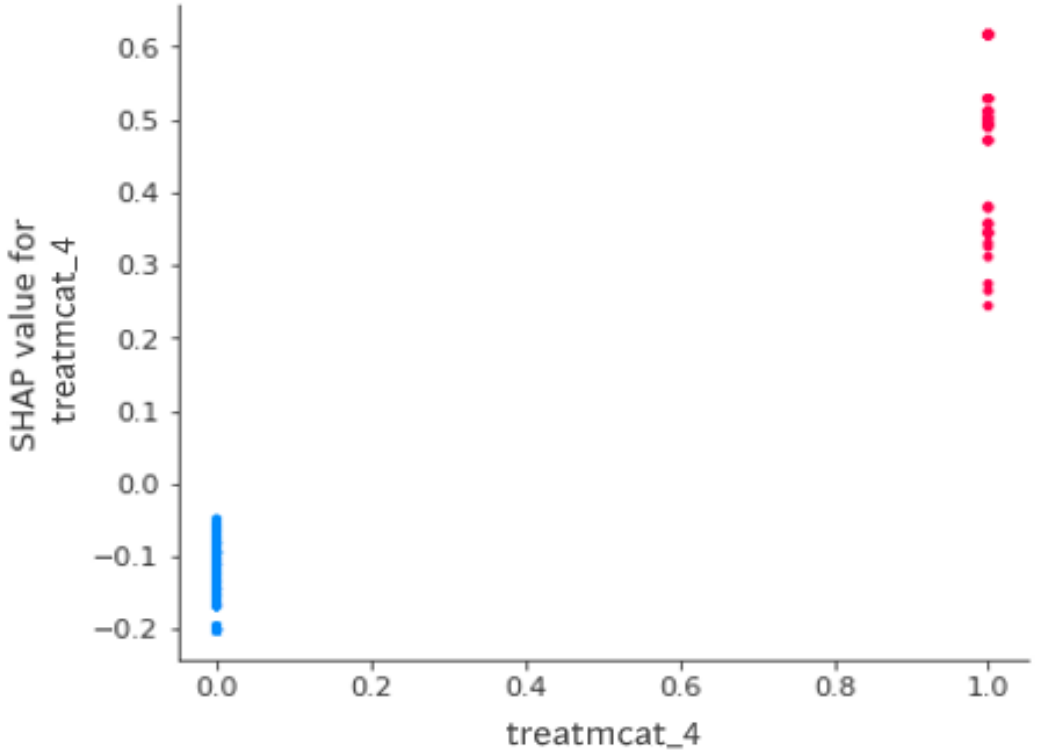

Figure 16. Dependence Plot for the 'treatmcat_4' Feature.

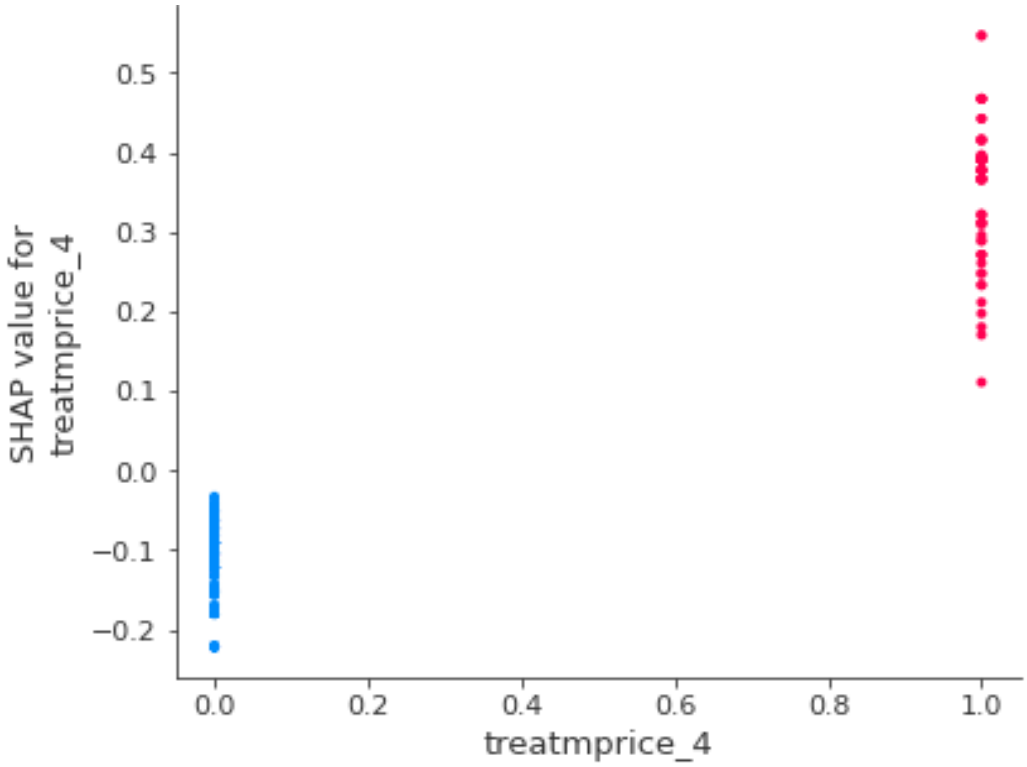

Figure 17. Dependence Plot for the 'treatmprice_4' Feature.

The Summary Plot for the '***lastcontact_3***' feature reveals a clear and concentrated pattern in its impact on the model's predictions (see Figure 12). The plot shows that the dots for '***lastcontact_3***' are agglomerated around specific SHAP values, with red dots on the negative side, primarily around -0.1 on the x-axis, indicating that higher values of the feature—contacts occurring more than a year ago—tend to decrease the likelihood of a positive outcome. The blue dots are situated on the positive side, ranging from 0.1 to 0.5 on the x-axis, signifying that more recent contacts have a positive impact on the model's predictions. The inverse color pattern shows a dichotomy: higher '***lastcontact_3***' values (red dots) correlate with negative impacts, while lower values (blue dots) are associated with positive impacts. The recency of contact is crucial, as longer intervals since the last contact (higher

'***lastcontact_3***' values) are linked to reduced predictions, whereas more recent contacts (lower '***lastcontact_3***' values) enhance predictions.

The dependence plot for this feature, shown in Figure 18, further confirms this pattern, showing that blue dots clustered around the 0 value on the x-axis with positive SHAP values, and red dots positioned above the 1 value on the x-axis with negative SHAP values. The binary nature of this feature reinforces this dichotomy, underscoring its significant role in shaping the model's predictions and emphasizing that recent interactions are more likely to yield favorable outcomes.

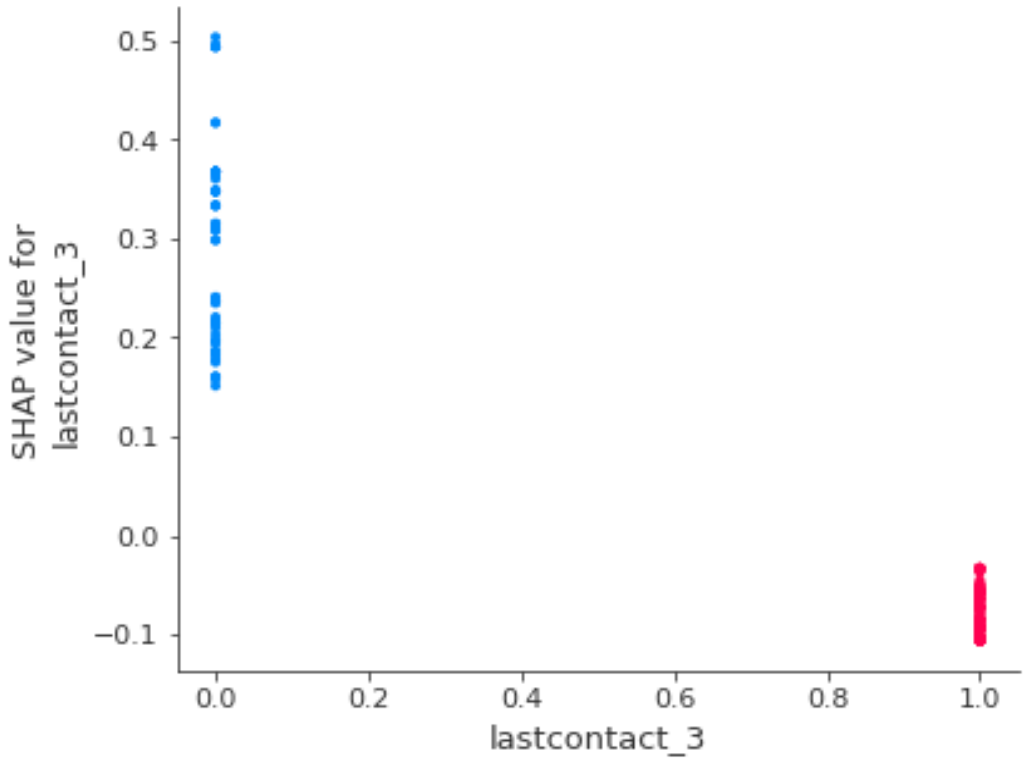

Figure 18. Dependence Plot for the 'lastcontact_3' Feature.

Further insight into the model's decision-making process is provided by the Waterfall Plots generated by BorutaShap. Randomly selected, Case #28, as illustrated in Figure 19, demonstrates how individual features contribute to the final prediction.

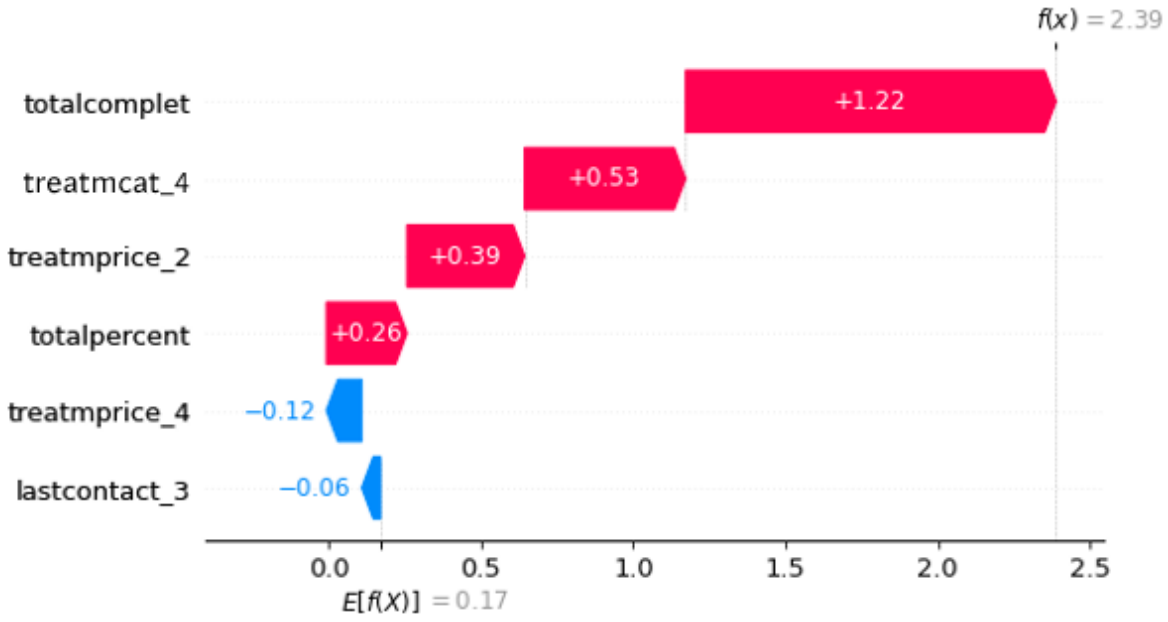

Figure 19. Waterfall Plot for single predictions (Case #28).

The plot for Case #28 begins from the expected baseline value of 0.17 and reaches a significantly higher prediction of 2.39. Positive contributors include '***totalcomplet***', which adds approximately +1.22, followed by '***treatmcat_4***' (+0.53), '***treatmprice_2***' (+0.39), and '***totalpercent***' (+0.26). These features collectively elevate the prediction above the expected mean. Conversely, '***treatmprice_4***' and '***lastcontact_3***' have negative contributions, slightly reducing the prediction by -0.12 and -0.06, respectively.

This breakdown for Case #28 highlights how higher values of '***totalcomplet***' and similar features enhance the predicted outcome, while '***treatmprice_4***' and '***lastcontact_3***' contribute modestly to lowering it. Knowing which features contribute positively or negatively can help come up with actionable insights, enabling to tailor follow-up strategies.

For instance, if '***totalcomplet***' is a major positive contributor, ensuring patients continue and complete their prescribed treatments might increase their likelihood of them returning for additional oral care.

Building on these insights, it is worth considering a case with a different outcome, such as Case #5, which was randomly selected and is illustrated in the next Waterfall Plot (Figure 20). The plot begins with the same expected baseline value of 0.17, but the final prediction for Case #5 is notably lower, at -0.179. This decrease is largely driven by the negative contributions of specific features. '***Treatmprice_2***' has the most substantial negative impact, subtracting approximately -0.6 from the prediction, indicating that higher values of this feature significantly lower the predicted outcome. Additionally, features like '***treatmprice_4***', '***treatmcat_4***', and '***lastcontact_3***' contribute further to the reduction, with respective negative impacts of -0.12, -0.1, and -0.09. This time again, '***totalcomplet***' and '***totalpercent***' have positive impacts, adding +0.43 and +0.13 to the prediction, but these are insufficient to counterbalance the substantial negative contributions.

The negative prediction for Case #5 stresses the importance of understanding which factors can lead to less favorable outcomes. By identifying that higher treatment prices ('***treatmprice_2***') and longer intervals since the last contact of the patient ('***lastcontact_3***') are associated with a decrease in predicted success, actionable strategies can be developed to address these issues. For instance, higher treatment prices may discourage some patients or create concerns about the cost-effectiveness of treatments, then offering transparent information and support can help address these concerns.

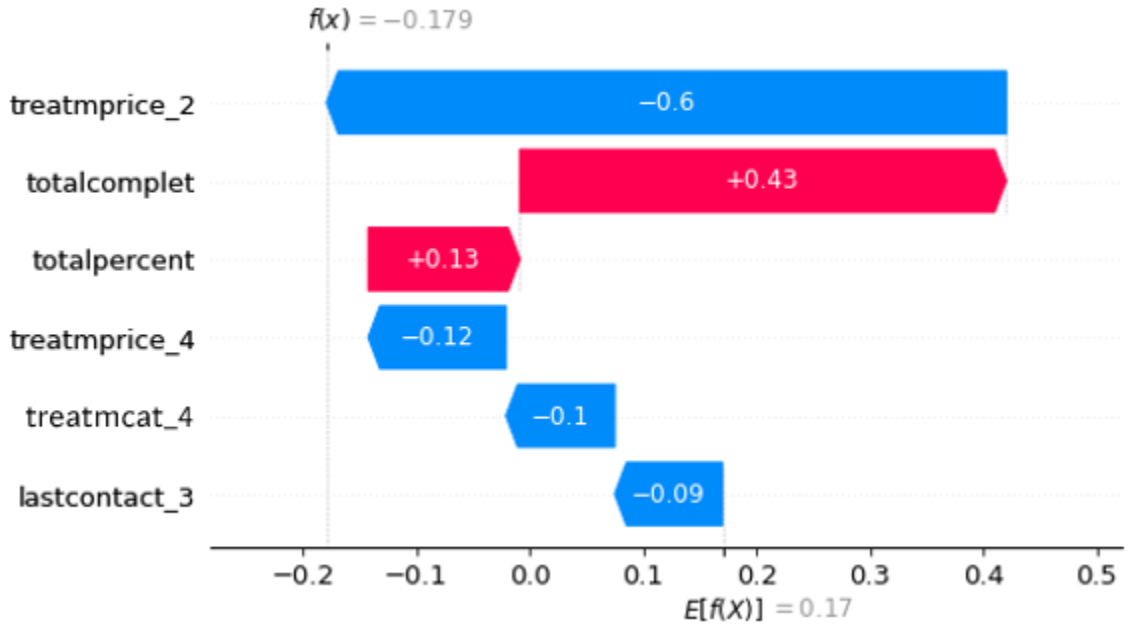

Figure 20. Waterfall Plot for single predictions (Case #5).

The contrasting predictions for Case #28, and Case #5 as depicted in the Waterfall Plots (Figure 19 and Figure 20, respectively), reveal how differently feature values can influence the model's output. In Case #28, the final prediction significantly exceeds the expected baseline due to positive contributions from features such as '***totalcomplet***' and '***treatmcat_4***'. In contrast, Case #5 shows a notably lower prediction, primarily driven by negative contributions from '***treatmprice_2***' and other features. This disparity underscores the model's sensitivity to varying feature values and their cumulative effects. The insights gained from these cases highlight that while certain features may boost predictions in some scenarios, their impact can be diminished by other factors in different cases. This variability emphasizes the need for a nuanced understanding of how individual feature contributions shape predictions, suggesting that actionable strategies must be tailored to address specific factors influencing each case.

To explore and visualize feature contributions, the Plot Tree generated by the BorutaShap algorithm, shown in Figure 21, offers a detailed view of how features impact the model's predictions through a decision tree framework. It provides a step-by-step breakdown of the decision-making process, illustrating how data points are classified based on feature values.

The diagram starts with the root node, which tests the condition ***´totalpercent´*** <= 0.012, splitting the data into two branches based on whether this condition is met. Internal nodes further refine these branches by evaluating additional conditions, such as ***´treatmcat_4´*** <= 0.5 and ***´treatmprice_2´*** <= 0.5. Key metrics at these nodes include the Gini index, which measures data impurity, the number of samples at each node, the distribution of data points across different classes, and the predominant class prediction.

At the end of the branches, leaf nodes display the final classification, showing the majority class and the distribution of data points within it. To predict the outcome for a new data point, the path is traced from the root node through branches based on feature values until reaching a leaf node for the final prediction. Features closer to the root are more influential, causing larger splits in the data, while class probabilities at each node show the likelihood of each class given the conditions up to that point.

Overall, the Decision Tree Plot offers a clear visualization of how individual features impact model predictions, illustrating the sequence of decisions and their effects. In summary, the Tree Explainer, Summary Plot, Dependence Plot, Waterfall Plots, and Plot Tree collectively provide a thorough understanding of the model's performance and feature contributions. These visualizations help address the challenges of interpreting the XGBoost ensemble algorithm, offering deeper insights into feature influence and setting the stage for exploring the implications of these findings in the following chapter.

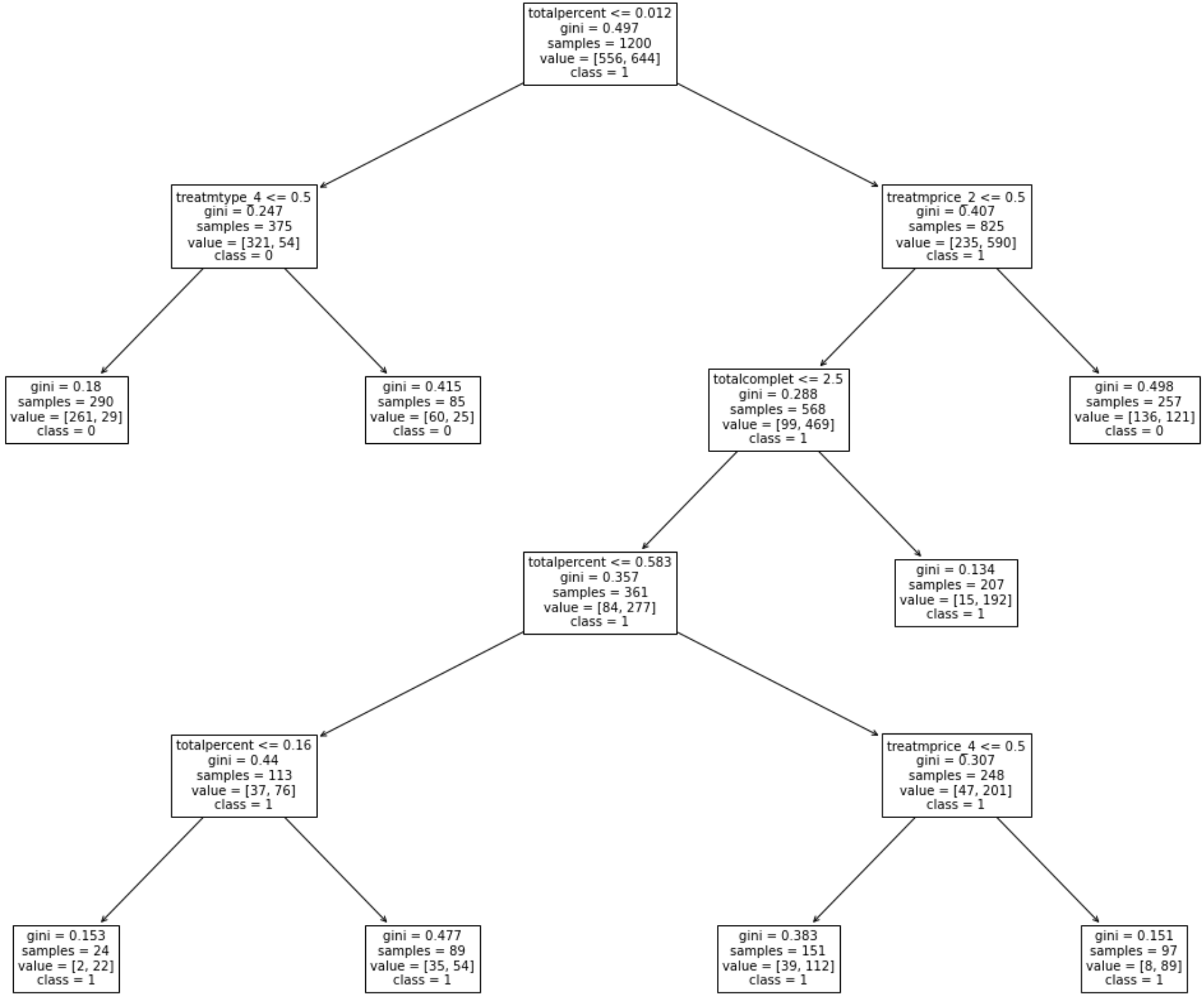

Figure 21. Decision Tree Plot.

5. Discussion

This chapter integrates the results obtained from the predictive model and its features with insights gathered from the literature review to explore their real-life implications for private dental clinics. The primary objective of this research was to develop and interpret a classification model predicting patient returning to dental clinics after their initial contact and visit. While the model achieved an accuracy of 79.07%, the focus now shifts to examining the significance of the features within the model and relating these findings to established theories and practices in patient retention.

Among the features analyzed, '***totalcomplet***,' which represents the total number of dental treatments completed by a patient, emerged as the most significant predictor, showing the highest Shapley value. This indicates that the number of treatments completed plays a crucial role in forecasting patient return rates.

The Dependence Plot for '***totalcomplet***' reveals substantial variability in SHAP values, suggesting that the feature's influence on predictions varies depending on the number of treatments completed. Specifically, lower '***totalcomplet***' values generally correlate with decreased predictions of patient return, while higher values, correlate with an increased likelihood of return.

These findings resonate with existing research on patient retention. For example, Han & Hyun (2015) highlight the importance of perceived service quality, satisfaction, and trust in driving patient loyalty and retention. The positive correlation between '***totalcomplet***' and patient retention supports these concepts, suggesting that as patients complete more treatments, their satisfaction and trust in the clinic likely increase, reinforcing their likelihood to return. Similarly, Onyeaso & Adalikwu (2008) note that positive past experiences with

service quality significantly enhance retention rates. This perspective aligns with the observed relationship between '***totalcomplet***' and retention, reflecting how a patient's continued engagement with the clinic through completed treatments strengthens their relationship with the clinic and fosters greater trust, reliability, and continued oral care.

The second feature, '***treatmprice_2***', representing treatments priced at 60 euros or less, shows significant importance with a Shapley value of 0.31641. This feature's importance emphasizes how affordability affects patient return behavior. For the sampled population, lower treatment prices are associated with a higher likelihood of returning.

The Summary Plot for '***treatmprice_2***' reveals an inverse relationship with patient retention. Higher treatment prices within the 0-60 euro range are associated with lower SHAP values, indicating a negative impact on the probability of patient return. Conversely, more affordable treatments, particularly those priced closer to the lower end of this range, show higher SHAP values which correlate with an increased likelihood of patients returning.

These findings align with existing literature on dental care affordability. As emphasized by the FDI (2015b), The high out-of-pocket costs for dental treatments present a significant barrier to patient care. In Spain, where dental expenses are predominantly covered by patients themselves, affordability remains a major concern. The pattern observed with '***treatmprice_2***' reflects this issue, suggesting that lower treatment costs could alleviate some of the financial barriers and enhance the population to visit or revisit their dental care providers.

In practical terms, private dental clinics might consider implementing pricing strategies that address affordability concerns. By offering more competitively priced treatments or financial incentives for returning patients, clinics can potentially improve their patient retention rates.

The feature '***totalpercent***,' which represents the percentage of prescribed treatments completed by patients, demonstrates a substantial role in predicting patient return behavior. The analysis shows a clear link between treatment completion rates and the likelihood of a positive outcome. Specifically, the Dependence Plot reveals that lower completion percentages are associated with decreased predictions, indicating that patients who complete fewer treatments are less likely to return. Conversely, higher completion percentages—especially those above the 50% mark—are linked to improved predictions, suggesting that patients who complete a larger portion of their treatments are more likely to return.

These results align with the existing literature on patient retention and treatment completion. Makarem & Coe (2014) argue that effective management and resolution of dental conditions are vital for improving patient retention. They emphasize that strategies aimed at increase patient loyalty and retention should be prioritize achieving successful treatment outcomes. Similarly, Amano (2023) highlights the importance of professional skillfulness, which includes promoting adherence to prescribed treatment plans, as a factor influencing patient's willingness to return. According to Amano, the perceived competence of dental professionals, as reflected in patients' adherence to treatment recommendations, is vital for maintaining patient trust and encouraging treatment completion and future return visits. Additionally Szabó et al. (2023), suggest that a dentist's knowledge and genuine interest in symptoms can lead to a more transparent and comprehensive prescription of treatments. This transparency helps patients better understand their oral health status and the necessity of completing prescribed treatments to effectively manage or resolve their conditions.

The feature '***treatmcat_4***,' which represents a category within the Treatment Category variable, also contributes to the model's predictive power, with a Shapley value of 0.197084.

This feature specifically refers to restorative procedures such as implants, prostheses, bridges, and crowns. The importance of '***treatmcat_4***' in predicting patient return behavior highlights the unique nature of these treatments, which often require multiple visits to ensure successful completion. The necessity of ongoing patient engagement for these complex procedures underscores their role in influencing retention rates.

This finding aligns with insights from Makarem & Coe (2014), who state that the context of the service, particularly the nature of the dental procedures involved, plays a crucial role in patient retention. Restorative treatments, by their very nature, demand a higher level of commitment and consistency from patients due to the extended treatment timeline and the need for multiple appointments. The shared responsibility between the patient and the dental professional in ensuring the success of these treatments significantly impacts retention.

Furthermore, the requirement for restorative procedures often indicates more advanced dental issues, necessitating a comprehensive treatment plan. The more extensive nature of these treatments means that patients are likely to return to the clinic multiple times to complete the prescribed care. This repeated engagement may also foster a stronger relationship between the patient and the dental provider, further enhancing retention.

The feature '***treatmprice_4***,' with a Shapley value of 0.15894, represents treatments within the 151 to 450 euro price range. This mid-to-high price category is indicative of a significant financial investment by patients, which appears to play a notable role in influencing their decision to return for follow-up care.

This willingness to invest in more expensive treatments may reflect a heightened awareness and prioritization of dental care, which can positively influence their commitment to completing their treatment plans. Patients who allocate substantial resources to their oral

health are likely to be more motivated to see their treatments through to completion, enhancing their likelihood of returning to the clinic. Moreover, the financial commitment associated with ‘treatmprice_4’ could also suggest a level of trust in the dental provider's expertise and the perceived value of the treatments offered. When patients decide to undergo mid-to-high priced procedures, it may indicate a strong belief in the quality and necessity of the care they are receiving, which further reinforces their likelihood of returning for ongoing treatment and follow-up care.

The feature ‘***lastcontact_3***,’ emphasizes the critical role of patient-clinic interaction frequency in predicting patient retention. This feature represents the time elapsed since a patient's last contact with the clinic, with higher values indicating a longer time since the last interaction. The model’s analysis reveals a clear pattern: as the time since last contact increases, the likelihood of patient retention decreases, as shown by the concentration of red dots on the negative side of the SHAP value spectrum. Specifically, contacts occurring more than a year ago are associated with a lower probability of the patient returning, likely due to diminished engagement and the fading importance of regular dental care in the patient's mind. Conversely, the presence of blue dots on the positive side, representing more recent contacts, underscores the positive impact that timely and regular interactions have on patient retention. Patients who have recently engaged with the clinic are more likely to return, which highlights the importance of maintaining a consistent communication strategy. These findings suggest that patients who feel a stronger connection to their dental care provider—reinforced through regular follow-ups and reminders—are more inclined to continue their treatment plans and adhere to recommended check-ups.

Regular and reliable communication not only reinforces the importance of ongoing dental care but also nurtures a sense of commitment and trust between the patient and the provider. This aligns with Gronroos (2000) findings that effective relationship management, including consistent communication, is vital for long-term customer retention.

Amano (2023) also highlights effective communication as a significant contributor to patient retention, emphasizing that transparent and consistent interactions from dental professionals helps build trust and reliance with the clinic's services. These findings suggest that dental clinics should prioritize proactive and personalized communication strategies to maintain strong patient relationships, particularly with those who have been absent for an extended period.

The exploration of these findings influencing patient retention, such as treatment completion rates, cost categories, treatment types, and the recency of patient contact, provides a comprehensive understanding of the factors that drive patient loyalty and return behavior to dental clinic settings. By integrating these findings with insights from existing literature, we gain a nuanced perspective on the multifaceted nature of patient retention and the importance of customized strategies that address these specific elements.

6. Conclusion

This investigation aimed to develop and interpret a predictive classification model for patient retention in a dental clinic, identifying key factors that influence whether patients return for follow-up visits. The model, which achieved an accuracy of 79.07%, highlighted several significant predictors, including the total number of treatments completed, treatment price categories, the percentage of prescribed treatments completed, the type of treatment, and the recency of the last contact with the clinic.

Among the findings, treatment completion emerged as a critical factor, with higher percentages of completed treatments and a greater number of total treatments completed significantly predicting patient retention. Additionally, the role of affordability is significant, as lower-priced treatments were strongly associated with higher patient return rates. The type of treatment, particularly restorative procedures requiring multiple visits, was also crucial, indicating that the nature of the dental issue influences whether patients return. Moreover, the recency of the last contact with the clinic was a key determinant, with more recent interactions positively correlated with the likelihood of a patient returning.

These findings contribute to a broader understanding of patient retention in dental care, emphasizing the need to balance technical quality, cost, and communication to build stronger patient relationships. Future research could extend these insights by applying similar predictive models in different contexts and demographic groups, potentially broadening the understanding of factors that influence patient retention. Additionally, exploring other dimensions, such as psychological influences or the impact of social media, could offer a more comprehensive view of patient behaviors and motivations.

7. References

Alauddin, M. S., Baharuddin, A. S., & Mohd Ghazali, M. I. (2021). The Modern and Digital Transformation of Oral Health Care: A Mini Review. *Healthcare*, *9*(2), 118. https://doi.org/10.3390/healthcare9020118.

Amano, K. (2023). Factors Influencing Customer Retention and Loyalty in Dental Practice in the United States. *Westcliff International Journal of Applied Research*, *7*(1), 5–18. https://doi.org/10.47670/wuwijar202371KA.

Appukuttan, D. (2016). Strategies to manage patients with dental anxiety and dental phobia: literature review. *Clinical, Cosmetic and Investigational Dentistry*, 35. https://doi.org/10.2147/CCIDE.S63626.

Bharadiya, J. P. (2023). Machine Learning and AI in Business Intelligence: Trends and Opportunities. *International Journal of Computer (IJC)*, *48*(1). https://ijcjournal.org/index.php/InternationalJournalOfComputer/article/view/2087.

Buswell, G. (2024). *The French healthcare system*. Expatica. https://www.expatica.com/fr/healthcare/healthcare-basics/a-guide-to-the-french-healthcare-system-101166/#dentists.

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785.

Colgate Global Scientific Communications. (2023). *How Often Should You Go To The Dentist?* https://www.colgate.com/en-us/oral-health/dental-visits/how-often-should-you-go-to-the-dentist.

Consejo General de Colegios de Dentistas de España. (2022a). *La Salud Bucodental en la Union Europea*. Grupo ICM de comunicación. https://consejodentistas.es/wp-content/uploads/2023/05/La-Salud-Bucodental-en-la-UE.pdf.

Consejo General de Colegios de Dentistas de España. (2022b). *Los Dentistas en España: Análisis de situación*. https://odontologia.ugr.es/sites/centros/odontologia/public/ficheros/analisis-dentistas-espa%C3%B1a.pdf.

Elflein, J. (2024). *Oral health and dental care in the U.S. - Statistics & Facts*. https://www.statista.com/topics/3944/oral-health-and-dental-care-in-the-us/#topicOverview.

Espinosa-Zúñiga, J. J. (2020). Aplicación de algoritmos Random Forest y XGBoost en una base de solicitudes de tarjetas de crédito. *Ingeniería Investigación y Tecnología*, *21*(3), 1–16. https://doi.org/10.22201/fi.25940732e.2020.21.3.022.

FDI World Dental Federation. (n.d.). *About oral health* . Retrieved June 2, 2024, from https://www.fdiworlddental.org/key-facts-about-oral-health.

FDI World Dental Federation. (2015a). *Oral Health Worldwide: A report by FDI World Dental Federation*. https://www.fdiworlddental.org/sites/default/files/2020-11/2015_wohd-whitepaper-oral_health_worldwide.pdf.

FDI World Dental Federation. (2015b). *The Challenge of Oral Disease: A Call for Global Action. The Oral Health Atlas.* (2nd ed.). FDI World Dental Federation.

https://www.fdiworlddental.org/sites/default/files/2021-03/complete_oh_atlas-2_0.pdf.

Fellows, J. L., Atchison, K. A., Chaffin, J., Chávez, E. M., & Tinanoff, N. (2022). Oral Health in America. *The Journal of the American Dental Association*, *153*(7), 601–609. https://doi.org/10.1016/j.adaj.2022.04.002.

Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems* (2nd ed.). O'Reilly Media. https://powerunit-ju.com/wp-content/uploads/2021/04/Aurelien-Geron-Hands-On-Machine-Learning-with-Scikit-Learn-Keras-and-Tensorflow_-Concepts-Tools-and-Techniques-to-Build-Intelligent-Systems-OReilly-Media-2019.pdf.

Gronroos, C. (2000). *Service management and marketing: a customer relationship management approach* (2nd ed.). Wiley. https://tashfeen.pbworks.com/f/Book%204%20-%20Service%20Management%20and%20Marketing.pdf.

Gruen, T. W., Summers, J. O., & Acito, F. (2000). Relationship Marketing Activities, Commitment, and Membership Behaviors in Professional Associations. *Journal of Marketing*, *64*(3), 34–49. https://doi.org/10.1509/jmkg.64.3.34.18030.

Gussy, M. G., Bracksley, S. A., & Boxall, A. (2013). *How often should you have dental visits?* https://ahha.asn.au/wp-content/uploads/2024/04/20130627-Deeble-Institute-Evidence-brief-Dental-visit-frequency.pdf.

Han, H., & Hyun, S. S. (2015). Customer retention in the medical tourism industry: Impact of quality, satisfaction, trust, and price reasonableness. *Tourism Management*, *46*, 20–29. https://doi.org/10.1016/j.tourman.2014.06.003.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). Springer. https://www.sas.upenn.edu/~fdiebold/NoHesitations/BookAdvanced.pdf.

Hoffmann, B., Erwood, K., Ncomanzi, S., Fischer, V., O'Brien, D., & Lee, A. (2022). Management strategies for adult patients with dental anxiety in the dental clinic: a systematic review. *Australian Dental Journal*, *67*(S1). https://doi.org/10.1111/adj.12926.

Ilemobayo, J., Durodola, O., Alade, O., J Awotunde, O., T Olanrewaju, A., Falana, O., Ogungbire, A., Osinuga, A., Ogunbiyi, D., Ifeanyi, A., E Odezuligbo, I., & E Edu, O. (2024). Hyperparameter Tuning in Machine Learning: A Comprehensive Review. *Journal of Engineering Research and Reports*, *26*(6), 388–395. https://doi.org/10.9734/jerr/2024/v26i61188.

James, S. L., Abate, D., Abate, K. H., Abay, S. M., Abbafati, C., Abbasi, N., Abbastabar, H., Abd-Allah, F., Abdela, J., Abdelalim, A., Abdollahpour, I., Abdulkader, R. S., Abebe, Z., Abera, S. F., Abil, O. Z., Abraha, H. N., Abu-Raddad, L. J., Abu-Rmeileh, N. M. E., Accrombessi, M. M. K., … Murray, C. J. L. (2019). Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017.

The Lancet, *392*(10159), 1789–1858. https://doi.org/10.1016/S0140-6736(18)32279-7.

Jevdjevic, M., & Listl, S. (2022). *Economic impacts of oral diseases in 2019 - data for 194 countries*. https://heidata.uni-heidelberg.de/dataset.xhtml?persistentId=doi:10.11588/data/JGJKK0.

Joda, T., Yeung, A. W. K., Hung, K., Zitzmann, N. U., & Bornstein, M. M. (2021). Disruptive Innovation in Dentistry: What It Is and What Could Be Next. *Journal of Dental Research*, *100*(5), 448–453. https://doi.org/10.1177/0022034520978774.

Kahurke, S. (2023). Artificial Intelligence Algorithms and Techniques for Dentistry. *2023 1st International Conference on Cognitive Computing and Engineering Education (ICCCEE)*, 1–4. https://doi.org/10.1109/ICCCEE55951.2023.10424481.

Kay, E. J. (1999). How often should we go to the dentist? *BMJ*, *319*(7204), 204–205. https://doi.org/10.1136/bmj.319.7204.204.

Keany, E. (2020). *BorutaShap : A wrapper feature selection method which combines the Boruta feature selection algorithm with Shapley values*. Zenodo. https://doi.org/10.5281/zenodo.4247618.

Kotler, P., & Keller, K. L. (2016). *Marketing Management* (15th ed.). Pearson Global Editions. https://www.ingebook.com/ib/NPcd/IB_BooksVis?cod_primaria=1000187&codigo_libro=10922.

Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*. Springer . https://doi.org/10.1007/978-1-4614-6849-3.

Kumar, A. (2023). *Random Forest vs XGBoost: Which One to Use? Examples*. Analytics Yogi. https://vitalflux.com/random-forest-vs-xgboost-which-one-to-use/.

Kursa, M. B., & Rudnicki, W. R. (2010). Feature Selection with the Boruta Package. *Journal of Statistical Software*, *36*(11). https://doi.org/10.18637/jss.v036.i11.

Makarem, S. C., & Coe, J. M. (2014). Patient Retention at Dental School Clinics: A Marketing Perspective. *Journal of Dental Education*, *78*(11), 1513–1520. https://doi.org/10.1002/j.0022-0337.2014.78.11.tb05826.x.

Maycher, G. (2023). *Dental Strategies for Retention, Revenue, and Patient Outcomes*. The Dental Brief Podcast. https://www.dentaltown.com/blog/post/18903/dental-strategies-for-retention-revenue-and-patient-outcomes-gabriele-maycher-the-dental-brief-212.

Naamati-Schneider, L., & Salvatore, F. P. (2022). *Digital Transformation in Private Dental Clinics* (pp. 201–218). Palgrave Macmillan. https://doi.org/10.1007/978-3-031-07769-2_10.

Nilsson, R., Peña, Jose. M., Bjorkegren, J., & Tegnér, J. (2007). Consistent Feature Selection for Pattern Recognition in Polynomial Time. *Journal of Machine Learning Research*, *8*, 589–612. https://www.jmlr.org/papers/volume8/nilsson07a/nilsson07a.pdf.

NVIDIA Corporation. (n.d.). *XGBoost*. Retrieved June 30, 2024, from https://www.nvidia.com/en-us/glossary/xgboost/.

Onyeaso, G., & Adalikwu, C. (2008). An Empirical Test of Customer Retention-Perceived Quality Link: Strategic Management Implications. *Journal of Business Strategies*, *25*(1), 53–71. https://doi.org/10.54155/jbs.25.1.53-71.

Paul Isson, J. (2018). *Unstructured Data Analytics: How to Improve Customer Acquisition, Customer Retention, and Fraud Detection and Prevention*. Wiley. https://doi.org/10.1002/9781119378846.

Pegon-Machat, E., Faulks, D., Eaton, K. A., Widström, E., Hugues, P., & Tubert-Jeannin, S. (2016). The healthcare system and the provision of oral healthcare in EU Member States: France. *British Dental Journal*, *220*(4), 197–203. https://doi.org/10.1038/sj.bdj.2016.138.

Peiró-Signes, Á., Segarra-Oña, M., Trull-Domínguez, Ó., & Sánchez-Planelles, J. (2022). Exposing the ideal combination of endogenous–exogenous drivers for companies' ecoinnovative orientation: Results from machine-learning methods. *Socio-Economic Planning Sciences*, *79*. https://doi.org/10.1016/j.seps.2021.101145.

Qin, X., Zi, H., & Zeng, X. (2022). Changes in the global burden of untreated dental caries from 1990 to 2019: A systematic analysis for the Global Burden of Disease study. *Heliyon*, *8*(9), e10714. https://doi.org/10.1016/j.heliyon.2022.e10714.

Rigby, D. K. (2017). *Management Tools 2017: An executive's guide*. https://www.bain.com/contentassets/b03332ae288d49769485ee490ad9d267/bain_book_management_tools_2017.pdf.

Rodríguez-Pérez, R., & Bajorath, J. (2020). Interpretation of machine learning models using shapley values: application to compound potency and multi-target activity

predictions. *Journal of Computer-Aided Molecular Design*, *34*(10), 1013–1026. https://doi.org/10.1007/s10822-020-00314-0.

Sabbeh, S. F. (2018). Machine-Learning Techniques for Customer Retention: A Comparative Study. *International Journal of Advanced Computer Science and Applications*, *9*(2). https://doi.org/10.14569/IJACSA.2018.090238.

Schwendicke, F., & Krois, J. (2022). Data Dentistry: How Data Are Changing Clinical Care and Research. *Journal of Dental Research*, *101*(1), 21–29. https://doi.org/10.1177/00220345211020265.

Schwendicke, F., & Marazita, M. L. (2022). Data-Driven Dental, Oral, and Craniofacial Analytics: Here to Stay. *Journal of Dental Research*, *101*(11), 1255–1257. https://doi.org/10.1177/00220345221120564.

Schwendicke, F., Samek, W., & Krois, J. (2020). Artificial Intelligence in Dentistry: Chances and Challenges. *Journal of Dental Research*, *99*(7), 769–774. https://doi.org/10.1177/0022034520915714.

Sharma, A., Gupta, D., Nayak, N., Singh, D., & Verma, A. (2022). Prediction of Customer Retention Rate Employing Machine Learning Techniques. *2022 1st International Conference on Informatics (ICI)*, 103–107. https://doi.org/10.1109/ICI53355.2022.9786903.

Szabó, R. M., Buzás, N., Braunitzer, G., Shedlin, M. G., & Antal, M. Á. (2023). Factors Influencing Patient Satisfaction and Loyalty as Perceived by Dentists and Their Patients. *Dentistry Journal*, *11*(9), 203. https://doi.org/10.3390/dj11090203.

United Nations. (2015). *SDG Resource Document: Targets Overview*. https://sdgs.un.org/sites/default/files/2020-09/SDG%20Resource%20Document_Targets%20Overview.pdf.

Vashisht, R., Sharma, A., Kiran, T., Jolly, S. S., Brar, P. K., & Puri, J. V. (2024). Artificial intelligence in dentistry — A scoping review. *Journal of Oral and Maxillofacial Surgery, Medicine, and Pathology*, *36*(4), 579–592. https://doi.org/10.1016/j.ajoms.2024.04.009.

Wade, C. (2020). *Hands-On Gradient Boosting with XGBoost and scikit-learn*. Packt Publishing. https://learning.oreilly.com/library/view/hands-on-gradient-boosting/9781839218354/.

World Health Organization. (2012, April). *Oral Health. Fact sheet no.318*. www.who.int/mediacentre/factsheets/fs318/en/index.htm.

World Health Organization. (2022). *Global oral health status report: towards universal health coverage for oral health by 2030*. https://www.paho.org/en/documents/global-oral-health-status-report-towards-universal-health-coverage-oral-health-2030.

World Health Organization. (2023). *Oral health*. https://www.who.int/news-room/fact-sheets/detail/oral-health.

Wright, J. T. (2024). Critical challenges facing dentistry. *The Journal of the American Dental Association*, *155*(1), 1–2. https://doi.org/10.1016/j.adaj.2023.11.001.

Xie, B., Xu, D., Zou, X.-Q., Lu, M.-J., Peng, X.-L., & Wen, X.-J. (2024). Artificial intelligence in dentistry: A bibliometric analysis from 2000 to 2023. *Journal of Dental Sciences*, *19*(3), 1722–1733. https://doi.org/10.1016/j.jds.2023.10.025.

ANEXO I. RELACIÓN DEL TRABAJO CON LOS OBJETIVOS DE DESARROLLO SOSTENIBLE DE LA AGENDA 2030

Anexo al Trabajo de Fin de Grado y Trabajo de Fin de Máster: Relación del trabajo con los Objetivos de Desarrollo Sostenible de la agenda 2030

Grado de relación del trabajo con los Objetivos de Desarrollo Sostenible (ODS).

Objetivos de Desarrollo Sostenibles	**Alto**	**Medio**	**Bajo**	**No Procede**
ODS 1. **Fin de la pobreza.**				X
ODS 2. **Hambre cero.**				X
ODS 3. **Salud y bienestar.**	X			
ODS 4. **Educación de calidad.**				X
ODS 5. **Igualdad de género.**				X
ODS 6. **Agua limpia y saneamiento.**				X
ODS 7. **Energía asequible y no contaminante.**				X
ODS 8. **Trabajo decente y crecimiento económico.**				X
ODS 9. **Industria, innovación e infraestructuras.**		X		
ODS 10. **Reducción de las desigualdades**.			X	
ODS 11. **Ciudades y comunidades sostenibles**.				X
ODS 12. **Producción y consumo responsables.**				X
ODS 13. **Acción por el clima.**				X
ODS 14. **Vida submarina.**				X
ODS 15. **Vida de ecosistemas terrestres.**				X
ODS 16. **Paz, justicia e instituciones sólidas.**				X
ODS 17. **Alianzas para lograr objetivos.**			X	

Descripción de la alineación del TFG/TFM con los ODS con un grado de relación más alto.

***Utilice tantas páginas como sea necesario.

Editor: Secretaria General / UPV · D.L.: V-5092-2006 · ISSN: 1887-2298 · www.upv.es/secgen

DIEGO ALEJANDRO| SERRANO|AMAYA
Firmado digitalmente por DIEGO ALEJANDRO| SERRANO|AMAYA
Fecha: 2024.08.30 18:53:41 +02'00'
Firmat digitalment per UNIVERSITAT POLITECNICA DE VALENCIA
Data: 27/07/2022 18:54:40 CEST

Id: UPV-GENFirma-128544 Cod. Verificació: Q0W97FCA0C7FRQ01