INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Machine learning for the development and validation of predictive models for periodontitis

Ricardo Alexandre Gonçalves Maltez

Master's in Business Analytics,

Supervisor:
Professor Nuno Duarte Fialho Sanches Borges dos Santos,
Assistant Professor, Iscte - Instituto Universitário de Lisboa

Co-supervisor:
PhD João Tiago da Silva Botelho, Assistant Professor,
Egas Moniz School of Health & Science

October, 2023

BUSINESS
SCHOOL

Department of Quantitative Methods for Management and Economics

Machine learning for the development and validation of predictive models for periodontitis

Ricardo Alexandre Gonçalves Maltez

Master's in Business Analytics,

Supervisor:
Professor Nuno Duarte Fialho Sanches Borges dos Santos, Assistant Professor, Iscte - Instituto Universitário de Lisboa

Co-supervisor:
PhD João Tiago da Silva Botelho, Assistant Professor, Egas Moniz School of Health & Science

October, 2023

I dedicate this Master's thesis to all those who have always been there for me and have added value to my life.

Acknowledgments

Writing this dissertation represented one of the most important stages of my academic journey, and several people deserved my thanks.

I want to thank my advisor and Professor Nuno Santos for his support and encouraging me to do more and better.

I would also like to thank my co-supervisor and Professor João Botelho for sharing his knowledge throughout the project.

To my family, who have always supported and guided me when I needed it most and have always been by my side at every stage of my life, they deserve special recognition. I want to thank my parents and sister for their unconditional support and for teaching me the values I have today, making me the man I am. Without them, none of this would be possible.

To my girlfriend, who has always supported me and been by my side in good times and bad, always providing extra motivation.

To all the family and friends with whom I could share the best moments of my life, thank you also for being part of this beautiful journey.

Thank you all!

Abstract

This dissertation had as its main objective the development of a predictive model for periodontitis in Portugal, trying to understand its capacity.

In this study, the dataset used was provided by the authors of Almada-Seixal Periodontal Health Study (SoPHiAS) and consisted of 1,064 participants aged between 18 and 95 years. This dataset includes variables relating to sociodemographic, behavioral and medical characteristics.

This work adopted as a methodology a slight adaptation of Cross-Industry Standard Process for Data Mining (CRISP-DM), reporting the understanding of the problem, data understanding, data preparation, modeling, evaluation and finally, implementation.

A Logistic Regression analysis was applied to determine possible risk factors for periodontitis and proceed to create the classification model. This model included variables relating to years of smoking, diabetes, use of dentures, bruxism, gender, age and education.

Concerning the most relevant performance metrics, this model achieved values of 70.2% and 71.7% for sensitivity and precision, respectively.

The prevalence of periodontitis was 59.9% in the studied sample. Regarding the risk of periodontitis, it increased significantly with age, years of smoking, lower levels of education, the use of dentures, male gender and the presence of diabetes.

These results demonstrate the relationship between several factors and periodontal disease, helping the medical community to define prevention strategies for periodontitis.

Keywords: Periodontitis, Data Analysis, Predictive Analysis, Oral Health, Systematic Literature Review

JEL codes: I14, C00

Resumo

Esta dissertação teve como principal objetivo o desenvolvimento de um modelo preditivo de periodontite em Portugal, tentando compreender a sua capacidade.

Neste estudo, o conjunto de dados utilizado foi fornecido pelos autores do Almada-Seixal *Periodontal Health Study* (SoPHiAS) e consistiu em 1.064 participantes com idades entre 18 e 95 anos. Este conjunto de dados inclui variáveis relativas a características sociodemográficas, comportamentais e médicas.

Este trabalho adotou como metodologia uma ligeira adaptação do *Cross-Industry Standard Process for Data Mining (CRISP-DM)*, relatando a compreensão do problema, compreensão dos dados, preparação dos dados, modelagem, avaliação e por fim, implementação.

Uma análise de regressão logística foi aplicada para determinar possíveis fatores de risco para periodontite e proceder à criação do modelo de classificação. Este modelo incluiu variáveis relativas a anos de tabagismo, diabetes, uso de prótese dentária, bruxismo, sexo, idade e escolaridade.

Relativamente às métricas de desempenho mais relevantes, este modelo atingiu valores de 70,2% e 71,7% para sensibilidade e precisão, respetivamente.

A prevalência de periodontite foi de 59,9% na amostra estudada. Quanto ao risco de periodontite, este aumentou significativamente com a idade, anos de tabagismo, menores níveis de escolaridade, uso de prótese dentária, género masculino e presença de diabetes.

Estes resultados demonstram a relação entre vários fatores e a doença periodontal, ajudando a comunidade médica a definir estratégias de prevenção da periodontite.

Palavras-chave: Periodontite, Análise de Dados, Análise Preditiva, Saúde Oral, Revisão Sistemática da Literatura

Códigos JEL: I14, C00

Index

Index of Figures

Index of tables

List of Acronyms

AIC – Akaike's Information Criterion

ANNs – Artificial Neural Networks

AUC – Area Under the Curve

BMI – Body Mass Index

CHAID - Chi-squared Automatic Interaction Detection

CNN – Convolutional Neural Network

COPD – Chronic Obstructive Pulmonary Disease

CRISP-DM - Cross-Industry Standard Process for Data Mining

CVDs – Cardiovascular Diseases

EGAT – Electricity Generating Authority of Thailand

HbA1c - Glycemic control indicator

KNN – K-nearest Neighbors

LCI – Lower Confidence Interval

MLP – Multiplayer Perceptron Artificial Neural Network

NHANES – National Health and Nutrition Examination Survey

NPV – Negative Predictive Value

ORs – Odds-ratio

PCA – Principal Component Analysis

PPV – Positive Predictive Value

PRISMA – Preferred Reporting Items for Systematic Reviews and Meta-Analysis

PTL – Periodontitis Related Tooth Loss

RBNN – Radial Basis Function Neural Network

RFECV – Recursive Feature Elimination, Cross Validated

RMSE – Root-mean-squared-error

ROC – Receiver Operating Characteristic

R-PBL – Periodontal Bone Loss Based on Radiographs

SGD – Stochastic Gradient Descent

SLR – Systematic Literature Review

SoPHiAS – Study of Periodontal Health in Almada-Seixal

SVM – Support Vector Machine

UCI – Upper Confidence Interval

WoS – Web of Science

1. Introduction

1.1. Context and Motivation

The 2022 World Health Organization Global Oral Health Status Report reviewed recent data on major oral diseases, risk factors, health system challenges, and reform opportunities. Overall, this report concludes that global oral health status is alarming and requires urgent action. Approximately 45% of the world's population is affected by one or more untreated oral diseases. Periodontitis affects one out of two adults (Trindade et al., 2023), and its most advanced stage is the sixth most common disease worldwide (Kassebaum et al., 2014). Periodontitis is a chronic inflammatory disease caused by the tissues' microbiome supporting the teeth, namely, the gums, periodontal ligament, cementum, and alveolar bone (Kinane et al., 2017).

Beyond its local consequences in the mouth, periodontitis can cause bacterial dissemination throughout the body (literally, bacteria that cause this disease can colonize other organs) and a systemic inflammatory reaction. Therefore, periodontitis interferes with other systemic processes and is strongly associated with several systemic diseases, such as cardiovascular disease, rheumatoid arthritis, and Alzheimer's disease (Hajishengallis & Chavakis, 2021; Schenkein et al., 2020). According to a recent study from the Egas Moniz School of Health and Science, periodontal disease is strongly associated with twenty-three non-communicable diseases and five types of cancer (Botelho, Mascarenhas, et al., 2022).

The socio-economic impact is significant, with estimates for the United States and the European Union indicating direct and indirect losses of $154.06 billion and €158.64 billion, respectively (Botelho, Machado, et al., 2022). Therefore, the problem we aim to minimize is the progression of the disease and the serious problems that may arise later, since the early diagnosis of periodontitis is a key element for successful treatment, as the progression of the disease causes an irreversible loss of periodontal structures (Kinane et al., 2017). Therefore, the development of a predictive model for periodontitis can help prevent disease progression, detect cases of the disease, and control the risk of developing other systemic diseases.

In Portugal, predictive models have been explored in the context of periodontal health. In the year 2022, two studies were conducted within this scope: one aimed to estimate the geographical distribution of the prevalence and risk of periodontitis (Antunes et al., 2022), and the other aimed to explore the accuracy of self-report to predict the prevalence of periodontitis (Machado et al., 2022). The first study addressed the application of predictive models in

healthcare management, highlighting their potential to improve the efficiency of services and optimize clinical decision-making. The second article focused on using predictive models to predict the incidence of periodontitis in Portugal, which is crucial for planning and implementing public health measures. Thus, these studies highlight the potential of predictive models as valuable tools for preventing and managing diseases, including periodontitis. These predictive models can help healthcare professionals identify risk factors early, develop personalized treatment strategies, and promote better clinical outcomes for patients with periodontitis.

In conclusion, this study will significantly impact the process of preventing the disease or its development, as periodontitis is a very comprehensive disease, and there needs to be more knowledge and understanding of it in a large part of society. Consequently, a model that predicts the presence of the disease may help in early diagnosis and control the progress of the disease to the most severe stage.

1.2. Research Question

Considering this problem, it is important to define the research questions that will guide this dissertation clearly. Thus, the following question was posed:

- Can a predictive model predict periodontitis for the Portuguese population?

1.3. Objectives and Forms of Validation

Considering the previously defined research question, our main objective of the study is to develop a predictive model for periodontitis to provide knowledge about possible risk indicators for the disease to assist in preventing it and anticipating cases of it.

Subsequently, specific objectives were defined to help achieve the main objective and answer the research questions, as presented in Table 1.1.

Table 1.1. Objectives and forms of validation

Objective	Form of validation
Study the association between variables and the target	This objective is achieved after quality and bivariate analyses are carried out. The quality analysis is carried out at both record and column levels. On the other hand, bivariate analysis allows an understanding of the degree of association of variables with the target.
Identify risk indicators for periodontitis and understand which variables are most relevant	The objective is achieved after identifying the most influential predictors and interpreting the results of the created model. Validation is done through performance metrics calculated from the confusion matrix.

1.4. Contributions

Periodontitis has been associated with several diseases, as mentioned previously. Therefore, the integration of previously identified risk factors for periodontitis into a model would be helpful to aid in the diagnosis and prevention of the disease.

Machine learning has grown exponentially in relation to applying different techniques in different areas, including periodontology.

Therefore, the main contribution of this study is the development of a predictive model for periodontitis in the Portuguese population to help in the screening of periodontal disease. Furthermore, this study may broaden the horizons of new models and provide evidences as to which variables to consider or not as possible predictors.

1.5. Dissertation Structure

After presenting the introduction, this study is elaborated through four essential chapters, each of which plays a significant role for the investigation structure and content.

A systematic literature review in Chapter 2 is presented, and existing knowledge in the study area is summarized in depth. This chapter serves as the basis for the development of the study and provides a broad and detailed context of the relevant topics.

Chapter 3 details the methodology adopted based on the Cross-Industry Standard Process for Data Mining (CRISP-DM). This chapter covers six steps and describes all the procedures

performed in each of them: understanding the problem, understanding the data, data preparation, modeling, evaluation, and implementation.

The results obtained are illustrated in Chapter 4. Furthermore, the most important predictors and the effects of each variable on the target are highlighted. In Chapter 5, the study's conclusions are described, and the limitations, contributions, and recommendations for future studies are presented.

In the end, it is possible to observe the references and appendixes.

2. Literature Review

2.1. Protocol of the Systematic Literature Review

A systematic literature review (SLR) incorporates research that comprehensively overviews existing literature on a specific subject. This process should follow a detailed protocol to simplify the definition of review methods (Moher et al., 2009). Thus, this review synthesizes the literature on using predictive analytics in oral health to predict periodontal disease and identify associated factors.

A commonly used methodology for conducting an SLR consists of Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA), which is a very current methodology, updated in 2020, which helps in the detailed description of the entire systematic review process (Page et al., 2021). First, for the research and literature collection, the initial objectives and initial questions are defined, for which a search query and inclusion and exclusion criteria are created to select the relevant articles to answer the initial questions. Then, all selected articles are analyzed and summarized in terms of content and, finally, individually assessed to determine which articles are most suitable for the topic under investigation and which criteria are addressed in the literature.

Thus, using PRISMA helps ensure the SLR is conducted comprehensively and transparently, with a clear and accurate reporting of methods and findings.

2.1.1. Objectives and Research Questions

The main objective of an SLR is to identify and synthesize all available knowledge regarding a specific research question or subject rigorously and transparently.

This review provides a comprehensive overview of the current knowledge regarding applying predictive analytics to periodontal health. Thus, this review aims to identify and synthesize information on what has already been studied in periodontal health in terms of predictive analytics and its importance and identify which factors are most associated with periodontal disease.

Therefore, this research aims to answer the following central question: In the area of oral health, how does predictive analysis contribute to periodontal health?

Considering the topic, research objectives, and general research question, five specific research questions are raised:

Q1. What are the scopes and objectives of predictive analysis?

Q2. What are the data collection methodologies and characteristics that constitute the data?

Q3. What algorithms are used, and what are their performances?

Q4. Which variables have the greatest impact on periodontitis prediction?

Q5. What are the limitations and contributions of the research?

2.1.2. Article Selection Process

The strategy for the selection of articles for review consisted of a process composed of three phases: identification, screening, and inclusion.

In the identification phase, two scientific databases, Web of Science (WoS) and Scopus, were used. Subsequently, a query was made for the two chosen databases to select articles related to the study topic, resulting in 133 articles from WoS and 195 from Scopus: (periodontitis OR phyorrhoea) AND ("predictive analysis" OR prediction OR "predictive model").

In the second phase, related to the screening, several articles were excluded based on the defined exclusion criteria. Finally, in the last phase, the final selection of articles included in the SLR was made based on the inclusion criteria. The criteria used are listed in Table 2.1.

Table 2.1. Exclusion and inclusion criteria

Exclusion Criteria
• Documents that are not articles • Articles not in English • Duplicate articles • Articles outside the publication period 2016 to 2023 • Articles of journals with an impact factor lower than 4
Inclusion Criteria
• Articles that investigate prediction/risk/association models with periodontitis • Articles that indicate risk factors/indicators for periodontitis

After checking and analyzing the exclusion criteria, 56 articles were selected. However, in the final inclusion phase, the title and summary of each article were analyzed to verify whether the criteria were met. After applying all the inclusion criteria, 18 articles were selected for SLR according to the flowchart in Figure 2.1.

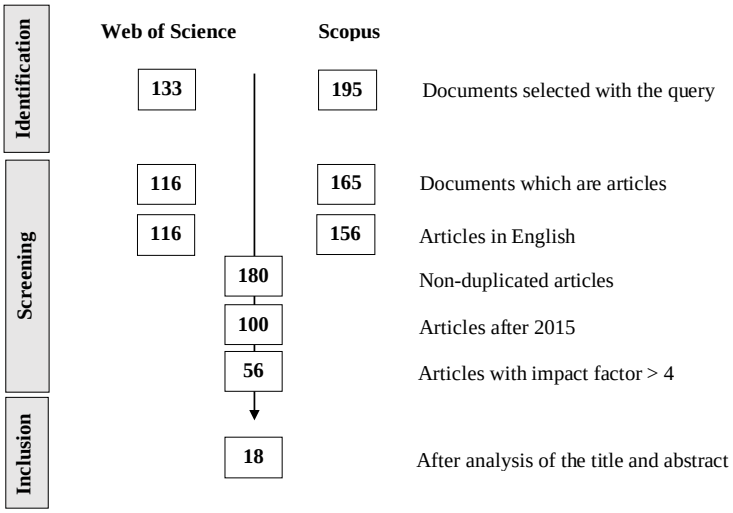

Figure 2.1. Article selection process

2.1.3. SLR Article Evaluation

Once the final set of selected articles was reached, they were thoroughly analyzed, and all relevant information was collected. This information was then archived and stored in an Excel file to summarize all information extracted from the articles. In this way, it was possible to find patterns and conclusions in the literature that helped improve this research.

Afterwards the information was structured, the quality criteria were defined for each research question (Table 2.2). These criteria were designed to make it possible to assess which articles best answered each question. Thus, for each article, a rating was given for the ten criteria, considering the context and objective of the research.

Table 2.2. Quality criteria defined for each specific research question

Specific question	ID	Quality criteria
Q1	**C1**	Does it clearly describe the context of the research?
	C2	Is the scope of the investigation periodontitis?
Q2	**C3**	Does it specify how the data was collected?
	C4	Does it say what kind of characteristics the data is composed of?
Q3	**C5**	Does it specify the models created in the modeling phase?
	C6	Does it say which performance indicators were analyzed and which were the most important?
Q4	**C7**	Does it mention which variables are included in the models?
	C8	Does it detail the profiles of people with periodontitis?
Q5	**C9**	Does it explain the limitations of the study?
	C10	Does it tell the contributions of the study?

The critical evaluation of full articles is carried out based on a checklist of questions, which are scored between 0 and 1, according to the response to the associated questions: score 1 for articles that answer the question thoroughly, 0.5 for those that answer partially, and 0 for those that do not answer the question at all.

2.2. Summary of the Contents of the Articles

Table 2.3 lists the selected articles and their characteristics. In addition, in recent years, there has been a significant amount of research on predictive analysis in oral health, both for the prediction of periodontitis and tooth loss. In terms of methods, several studies have investigated the effectiveness of self-reporting in predicting disease. However, some studies have analyzed the efficacy of radiographic-based methods, and others based on salivary biomarkers.

Table 2.3. Articles in SLR, published between 2016 and 2022.

ID	Year	Title	Authors	Journal	Q	Objective
1	2016	A Two-Domain Self-Report Measure of Periodontal Disease Has Good Accuracy for Periodontitis Screening in Dental School Outpatients	Chatzopoulos et al.	Journal of Periodontology	1	Evaluation of a two-domain self-report questionnaire for rapid periodontitis screening
2	2017	Prediction of Periodontitis Occurrence: Influence of Classification and Sociodemographic and General Health Information	Leite et al.	Journal of Periodontology	1	Investigating the accuracy of oral, systemic and socio-economic data in predicting the development of periodontal disease
3	2017	Natural history of periodontitis: Disease progression and tooth loss over 40 years	Ramseier et al.	Journal of Clinical Periodontology	1	To evaluate insertion and tooth loss over 40 years about untreated periodontitis
4	2018	Validation of multivariable models for predicting tooth loss in periodontitis patients	Schwendicke et al.	Journal of Clinical Periodontology	1	Validate six prediction models for tooth loss
5	2019	Development and validation of a predictive model for periodontitis using NHANES 2011-2012 data	Montero et al.	Journal of Clinical Periodontology	1	Develop and validate a predictive model for moderate to severe periodontitis in adults
6	2019	Adult Patient Risk Stratification Using a Risk Score for Periodontitis	Nobre et al.	Journal of Clinical Medicine	1	Estimate and evaluate a risk score for periodontitis prediction

ID	Year	Title	Authors	Journal	Q	Objective
7	2020	Accuracy of Panoramic Radiograph for Diagnosing Periodontitis Comparing to Clinical Examination	Machado et al.	Journal of Clinical Medicine	1	Explore the diagnostic accuracy of the radiograph-based PBL (R-PBL) method for periodontitis screening
8	2020	Validation of self-reported measures of periodontitis in a Spanish Population	Montero et al.	Journal of Periodontal Research	1	To assess the accuracy of self-reporting in predicting the prevalence of periodontitis
9	2020	Diagnostic accuracy of IL1β in saliva: The development of predictive models for estimating the probability of the occurrence of periodontitis in non-smokers and smokers	Arias-Bujanda et al.	Journal of Clinical Periodontology	1	Obtain models based on salivary indicators to predict the probability of periodontitis occurrence, differentiated by smoking habit
10	2021	Association, prediction, generalizability: Cross-center validity of predicting tooth loss in periodontitis patients	Schwendicke et al.	Journal of Dentistry	1	Predict tooth loss during supportive periodontal therapy
11	2021	An Interpretable Computer-Aided Diagnosis Method for Periodontitis from Panoramic Radiographs	Li et al.	Frontiers in Physiology	1	Predict the severity of periodontitis on panoramic dental radiographs
12	2021	Periodontitis is associated with cardiovascular diseases: A 13-year study	Tiensripojamarn et al.	Journal of Clinical Periodontology	1	Assess the association between periodontitis and the incidence of cardiovascular disease in Thai adults
13	2021	Accuracy of a 7-Item Patient-Reported Stand-Alone Tool for Periodontitis Screening	Sekundo et al.	Journal of Clinical Medicine	1	Evaluate the accuracy of a short seven-item tool for periodontitis screening based on patient report
14	2022	Self-Reported Measures of Periodontitis in a Portuguese Population: A Validation Study	Machado et al.	Journal of Personalized Medicine	2	Explore the accuracy of self-report in predicting the prevalence of periodontitis
15	2022	Associations between Periodontitis and COPD: An Artificial Intelligence-Based Analysis of NHANES III	Vollmer et al.	Journal of Clinical Medicine	1	Investigate possible associations between chronic obstructive pulmonary disease and periodontitis
16	2022	Periodontitis, age-related diseases and diabetes in an endocrinological outpatient setting (PARADIES): a cross-	Kabisch et al.	Acta Diabetologica	1	Assess predictors and risk indicators for periodontitis in patients with diabetes

ID	Year	Title	Authors	Journal	Q	Objective
		sectional analysis on predictive factors for periodontitis in a German outpatient facility				
17	2022	Geographical Distribution of Periodontitis Risk and Prevalence in Portugal Using Multivariable Data Mining and Modeling	Antunes et al.	International Journal of Environmental Research and Public Health	2	Estimate the geographical distribution of prevalence and risk of periodontitis using socio-demographic and economic data
18	2022	Systematic comparison of machine learning algorithms to develop and validate predictive models for periodontitis	Bashir et al.	Journal of Clinical Periodontology	1	Compare the validity of different machine learning algorithms for developing predictive models of periodontitis.

Note: Q – Quartile; COPD – Chronic Obstructive Pulmonary Disease

Regarding the context of the studies included in the SLR, Table 2.4 describes the scope of each article and the size, source, and country of data used.

Thirteen of the 18 articles focused on periodontitis (72%), the most studied area. However, there are also studies on tooth loss, a more advanced stage of the disease, and the association between the disease and other factors.

Table 2.4. Summary of the context of the study of SLR articles

ID	Scope	Sample	Source	Country
1	Periodontitis	535	Interview & Periodontal examination & Physical examination	Greece
2	Periodontitis	471	Oral Health Study (questionnaire & periodontal examination)	Brazil
3	Long-term attachment and PTL	75	Longitudinal cohort study	Sri Lanka
4	Tooth Loss	301	Database from ParoDat including periodontal therapy	Germany
5	Periodontitis	3017	NHANES	USA
6	Periodontitis	330	Prospective study of epidemiological surveillance of oral diseases	Portugal
7	Periodontitis	456	SoPHiAS & Panoramic dental X-ray	Portugal
8	Periodontitis	231	Di@bet.es (questionnaire & periodontal examination)	Spain
9	Periodontitis	141	Periodontal examination	Spain
10	Tooth Loss	897	Data from university centers	Germany

ID	Scope	Sample	Source	Country
11	Periodontitis	298 + 204	Suzhou data set & Zhongshan data set	China
12	Association between periodontitis and incidence of CVDs	1850	EGAT study	Thailand
13	Periodontitis	88	Questionnaire & Periodontal Examination	Germany
14	Periodontitis	103	Questionnaire & Periodontal Examination	Portugal
15	Association between periodontitis and COPD	15868	NHANES III	USA
16	Periodontitis	1641	Questionnaire & Medical file & Periodontal examination	Germany
17	Periodontitis	1064	Pordata & SoPHiAS	Portugal
18	Periodontitis	3453 + 3685	Taiwan study & NHANES	Taiwan and USA

Note: PTL – Periodontitis Related Tooth Loss; NHANES – National Health and Nutrition Examination Survey; SoPHiAS – Study of Periodontal Health in Almada-Seixal; CVDs – Cardiovascular Disease; EGAT – Electricity Generating Authority of Thailand; COPD – Chronic Obstructive Pulmonary Disease.

Table 2.5 lists the analysis techniques and evaluation metrics used in these studies. The most used analytical technique was logistic regression (67%). About the evaluation metrics analyzed, Sensitivity, Specificity, Receiver Operating Characteristic (ROC) and Area under the Curve (AUC) were the most frequently used.

Table 2.5. Summary of evaluation techniques and metrics used in SLR articles

ID	Technique used	Evaluation metrics
1	Logistic Regression	Sensitivity, Specificity, C-Statistic, PPV, NPV, ROC and AUC
2	Multivariable Binomial Regression	Sensitivity, Specificity, AUC and ROC
3	Logistic Regression and Markov chain modeling	Sensitivity, Specificity and ROC
4	Logistic Regression, Classification and regression trees and Decision-making chart	Sensitivity, Specificity, AUC and ROC
5	Logistic Regression	Sensitivity, Specificity, AUC, ROC and AIC
6	Logistic Regression	ROC and C-Statistic

ID	Technique used	Evaluation metrics
7	R-PBL	Sensitivity, Specificity, Accuracy, Precision, ROC and AUC
8	Logistic Regression	Sensitivity, Specificity, ROC and AUC
9	Logistic Regression	Accuracy, Sensitivity, Specificity, PPV, NPV, AUC and bc-AUC
10	Multivariable Linear Regression	Root-mean-squared-error (RMSE), Lower confidence interval (LCI), Upper confidence interval (UCI) and P-value
11	Deetal-Perio, SVM, Decision tree, Adaboost and CNN	Macro F1-score and Accuracy
12	The Cox proportional hazard regression model	Not specified
13	Logistic Regression	Sensitivity, Specificity, PPV, NPV, ROC and AUC
14	Logistic Regression	Sensitivity, Specificity, Accuracy, Precision, AUC, ROC and AIC
15	Logistic Regression, Random Forest Classifier, SGD, K-Nearest Neighbors, Decision Tree Classifier, Gaussian Naive Bayes, Support Vector Machines, CNN, MLP and RBNN	AUC and Accuracy
16	Logistic Regression	Sensitivity, Specificity, AUC, ROC and P-value
17	Linear Regression	Not specified
18	AdaBoost, ANNs, Decision trees, Gaussian process, KNN, Linear Support Vector Classification, Linear Discriminant Analysis, Logistic Regression, Random Forests and Naïve Bayes	Sensitivity, Specificity, PPV, NPV, Accuracy and AUC

Note: PPV – Positive Predictive Value; NPV – Negative Predictive Value; AIC – Akaike's Information Criterion; R-PBL – Periodontal Bone Loss Based on Radiographs; bc-AUC – Bias-corrected Area Under the Curve; RMSE – Root-mean-squared-error; LCI - Lower Confidence Interval; UCI – Upper Confidence Interval; SVM – Support Vector Machine; CNN – Convolutional Neural Network; SGD – Stochastic Gradient Descent; MLP – Multiplayer Perceptron Artificial Neural Network; RBNN – Radial Basis Function Neural Network; ANNs – Artificial Neural Networks; KNN – K-nearest Neighbours.

Table 2.6 summarizes the limitations and contributions of the 18 SLR articles.

The limitations are mostly related to the dataset (e.g., reduced-dimension dataset) or data collection (e.g., insufficient diversity in the data). As far as contributions are concerned, the studies can contribute, for example, to understanding the combination of variables to be used and to creating models for use by specialists to complement periodontal examination.

Table 2.6. Summary of limitations and contributions

ID	Limitations	Contributions
1	Data	Combining self-report variables and risk indicators for periodontitis in a single model
2	Data and Diagnostic Material	Evidence that environmental, systemic and local factors influence the disease
3	Data (Population)	Importance of early periodontitis treatment and smoking cessation under 30 years of age. Plaque, calculus, and gingivitis must be controlled to prevent disease progression, attachment loss, and tooth loss
4	Data and number of models tested	Validation of the models' performance and demonstration of the need for caution in their application
5	Data and External validation	Usability of a model combining cardiometabolic, demographic and lifestyle variables in primary medical care settings
6	Data	Development of a model to estimate a risk score for periodontitis that allows stratification into low, moderate and high risk
7	Level of inaccuracy	Support for the development of automated prediction systems for periodontitis
8	Diagnostic method and Data	A useful tool for screening patients at high risk of severe periodontitis
9	Data and recruitment	The quantification of IL1β levels in saliva has emerged as a very useful resource for identifying periodontitis in people who do not have systemic diseases
10	Data (selection bias), variables selection, inability to determine reasons for tooth loss and external validation	Even though covariates are significantly related to tooth loss, none of the models provided valuable predictions; Predictive value cannot be confused with association
11	Inability to interpret periodontitis severity and bias to identify the number teeth in patients with severe periodontitis	Development of a reliable and easy-to-understand approach for predicting the severity of periodontitis and a technique to number and segment teeth consisting of an innovative calibration algorithm
12	Data	Controlling risk factors such as periodontitis to prevent potentially fatal cardiovascular events can contribute to health promotion
13	Data and method limitations	Complementary analysis to periodontal examination
14	Data	Combining self-report measures with risk indicators shows predictive validity for the disease
15	Data	Demographic and oral health characteristics variables can be used as predictors of COPD cases using machine learning and deep learning algorithms when conducting a large-scale analysis
16	Data	Periodontitis in diabetics seems to be predicted by long-term complications and oral health-related quality of life
17	Data	A model to predict the risk of periodontitis in Portugal by municipality, allowing adequate oral health policy planning based on regional needs

ID	Limitations	Contributions
18	Data and External validation	Creation of a machine learning-based model with the potential to improve performance

2.3. Prediction of Periodontitis

Predictive analysis of oral health, especially in periodontal health, has been widely studied in different populations, and the inclusion of Machine Learning techniques in its development has been increasing. This is evident in Table 2.3, which demonstrates the diversity of the research conducted in this area.

In the German population, studies have been conducted to predict tooth loss, which represents the most advanced stage of periodontal disease (Schwendicke et al., 2018, 2021), concluding that age, smoking and the number of teeth before supportive periodontal therapy are significantly associated with this loss. Studies in other populations predict the degree of periodontitis based on radiographs (Li et al., 2021; Machado et al., 2020). Machado et al. (2020) concluded that the R-PBL method tested under the 2018 case definition is a reliable tool for screening periodontitis cases; however, this method is unable to accurately visualize the three-dimensional bone structure, leading to the fact that R-PBL cannot be considered as a definitive diagnostic tool. According to Li et al. (2021), the Deetal-Perio method not only outperforms other methods in segmenting and numbering teeth and predicting periodontitis but is also robust and generalizable to independent datasets, making Deetal-Perio a suitable method for screening and diagnosing periodontitis. Nevertheless, this method may have limitations in the analysis of radiographs with few teeth or abnormal shapes.

In addition, the development of predictive models based on salivary indicators has been studied (Arias-Bujanda et al., 2020), demonstrating that the diagnostic ability of salivary ILβ1 remains acceptable for differentiating between untreated and treated periodontitis.

Furthermore, focusing on the main topic of this study, in the literature, there is research in different populations that has explored the accuracy of self-report as a method to predict periodontitis (Chatzopoulos et al., 2016; Machado et al., 2022; Montero et al., 2020; Sekundo et al., 2021). Chatzopoulos et al. (2016) investigate, in the Greek population, the use of self-report to predict periodontitis and conclude that combining self-report measures from two domains (dentist-diagnosed and self-rated) increases the sensitivity of a predictive model, more concretely, the combination of the response to questions about the dental diagnosis of periodontal disease with bone loss and the feeling of loose or wobbly teeth. However, when age

and gender were added to the model, better performance was obtained, which was not the case for smoking, diabetic status, and body mass index (BMI). Moreover, Montero et al. (2020) used the self-report method combined with demographic and risk factors to evaluate the prevalence of severe periodontitis in the Spanish population based on different definitions of the disease. The results demonstrated that a combination of these indicators is helpful for estimating the prevalence of the disease. On the one hand, it is essential to highlight the answer to "*Do you think you may have gum disease?*" corresponds to the variable with the highest association with severe periodontitis. On the other hand, the most significant variables in the study were related to sociodemographic and behavioral indicators, precisely age and smoking. In addition, a study by Sekundo et al. (2021) in Germany aimed to evaluate the diagnostic accuracy of a screening tool based on patient self-report, measured through a risk score. Research suggests that this risk score is effective in screening for periodontitis. Moreover, it identifies several factors that are associated with the presence of periodontal inflammation, namely, advanced age, male gender, history of smoking, lower education, and previous occurrence of gingival bleeding or tooth mobility. Finally, in Portugal, Machado et al. (2022) assessed the validity of a self-report questionnaire with 13 questions to predict periodontitis. This study found more severe periodontitis in the older population, males, those with less education, predominantly non-smokers, and those with six or more lost teeth. Regarding self-reporting, the strongest associations were found in questions regarding gum disease, loose teeth, bone loss, and gum recession.

In summary, in various populations and backgrounds, self-reporting is used as an accurate method to predict periodontitis, combined with demographic and risk factors. In addition, factors such as age, gender, smoking, and educational level were consistently associated with periodontitis and gingival inflammation. However, among the studies described, there was no agreement regarding smoking, which may be linked to the characteristics of each sample studied.

Regarding other studies aiming to predict periodontitis in the United States, in the study by Montero et al. (2019), a predictive model for moderate to severe periodontitis in adults was developed and validated. In this study, regarding sociodemographic variables with the highest association with periodontitis, age stands out, followed by male gender, which is not the case for educational level and family income. Considering cardiometabolic risk indicators, smoking and the glycemic control indicator (HbA1c) were strongly associated, and no association was recorded for blood pressure, cholesterol, and periodontal condition. Consequently, this study

demonstrates that a model including age, gender, ethnicity, HbA1c, and smoking habits can be used to screen for periodontitis.

On the other hand, Bashir et al. (2022) compare the validity of different machine learning algorithms to develop and validate the models based on data from Taiwan and the USA. On the one hand, in the Taiwanese population, the profile more prone to disease corresponds to being older, male, less educated, former or current smoker, not drinking, having a higher BMI, higher waist circumference, systolic and diastolic blood pressure, not having visited the dentist in the previous year, having mobility in teeth, and not flossing. On the other hand, in the American population, the profile corresponding to older people, males, have a lower level of education, former or current smokers, do not drink, have a higher BMI, higher waist circumference, and systolic blood pressure, have not been to the dentist in the past year, have noticed mobile teeth, and do not floss.

Finally, Nobre et al. (2019) developed a risk model to identify individuals more likely to develop periodontal disease in Portugal. This study estimated a risk score for periodontitis prediction that allowed risk stratification (low-risk, moderate-risk, and high-risk). The risk indicators identified and included in the model were age, smoking, gingivitis, subgingival calculus, history of periodontitis, and fewer than two observations in the first year of follow-up, all of which were significant.

2.4. Risk Factors for Periodontitis

As already mentioned, periodontitis affects one in two adults (Trindade et al., 2023) and has recently been strongly associated with several non-communicable diseases and five types of cancer (Botelho, Mascarenhas, et al., 2022). Therefore, identifying risk factors associated with periodontitis is important, as early detection and effective treatment of the disease are essential for oral health and well-being.

Several studies have investigated the risk factors associated with periodontitis, and several factors have been identified as possible risk indicators for its development. These factors are listed in Table 2.7.

Table 2.7. Risk factors for periodontitis

Risk factor	References
Age	Bashir et al. (2022); Kabisch et al. (2022); Machado et al. (2022); Montero et al. (2019); Nobre et al. (2019); Sekundo et al. (2021)
Gender	Bashir et al. (2022); Chatzopoulos et al. (2016); Kabisch et al. (2022); Leite et al. (2017); Machado et al. (2022); Montero et al. (2019); Sekundo et al. (2021)
Ethnicity	Montero et al. (2019)
Smoking	Bashir et al. (2022); Kabisch et al. (2022); Montero et al. (2019, 2020); Nobre et al. (2019); Ramseier et al. (2017); Sekundo et al. (2021)
Self-reported gingival health	Leite et al. (2017); Machado et al. (2022); Nobre et al. (2019); Ramseier et al. (2017); Sekundo et al. (2021)
Oral hygiene habits	Bashir et al. (2022); Kabisch et al. (2022); Leite et al. (2017); Machado et al. (2022)
HbA1c	Montero et al. (2019)
Educational level	Bashir et al. (2022); Machado et al. (2022); Montero et al. (2020); Sekundo et al. (2021)

Age has consistently been identified as a risk factor for periodontitis. Studies have shown that the prevalence of the disease increases with advancing age (Bashir et al., 2022; Kabisch et al., 2022; Machado et al., 2022; Montero et al., 2019, 2020; Nobre et al., 2019; Sekundo et al., 2021). Gender also plays a role, with some studies indicating a higher susceptibility in males (Bashir et al., 2022; Kabisch et al., 2022; Leite et al., 2017; Montero et al., 2019; Sekundo et al., 2021).

Ethnicity may also be associated with periodontitis, and according to Montero et al. (2019), the prevalence of moderate to severe periodontitis was higher in African Americans and Hispanics, followed by Asian Americans, and lower in non-Hispanic whites.

Another well-established risk factor is smoking. Smoking is strongly associated with an increased risk of developing periodontitis (Bashir et al., 2022; Kabisch et al., 2022; Montero et al., 2019, 2020; Nobre et al., 2019; Ramseier et al., 2017; Sekundo et al., 2021). However, in a study conducted by Chatzopoulos et al. (2016), the addition of the variable regarding smoking to the model did not improve the model. As mentioned earlier, according to Machado et al. (2022), non-smokers are included in the profile as more susceptible to the disease.

Education may also have a role in the likelihood of having periodontitis. Individuals with lower educational levels have been associated with a higher risk of being diagnosed with

periodontitis (Bashir et al., 2022; Machado et al., 2022; Montero et al., 2020; Sekundo et al., 2021).

According to Montero et al. (2019), HbA1c, an indicator of diabetes control, is strongly associated with periodontitis. There is a bidirectional relationship between periodontitis and diabetes because inflammation is a central feature of both diseases.

In addition to sociodemographic and behavioral factors, specific oral health characteristics have also been identified as risk factors. The presence of gingivitis, calculus, bleeding, and insertion loss is associated with a high risk of periodontitis (Leite et al., 2017; Machado et al., 2022; Nobre et al., 2019; Ramseier et al., 2017; Sekundo et al., 2021).

Other risk factors include oral health habits and flossing. Patients with inadequate oral hygiene are at a higher risk of developing periodontitis (Bashir et al., 2022; Kabisch et al., 2022; Leite et al., 2017; Machado et al., 2022).

Finally, tooth loss has been identified as the final clinical outcome of periodontitis, particularly the most advanced condition of the disease. Tooth loss may indicate a history of previous periodontal disease, for this reason, the risk of being diagnosed with this condition is higher because it is a major clinical outcome expected in advanced periodontal bone loss (Machado et al., 2022; Montero et al., 2020; Schwendicke et al., 2021).

Understanding and recognizing these risk factors can assist in the implementation of preventive measures, early detection, and effective treatment strategies to improve oral health outcomes and the overall well-being of the population.

2.5. Methodologies and Techniques

In this section, the methodology and techniques of some articles selected in the literature review are addressed, as these sources were chosen because of their relevance and direct alignment with the specific objectives of the present study. The available literature covers a wide diversity of approaches; however, careful selection of these specific articles allows for a more in-depth and targeted analysis, providing a solid and oriented basis for the development of the predictive model in this research.

Montero et al. (2019) used descriptive statistical measures such as means and standard deviations to describe the characteristics of the sample. Candidate predictors were categorized and dichotomized. Next, the model was built using logistic regression, and to determine the ability of each model, candidate models were compared using ROC curves. Finally, the best-

performing model included the variables age, gender, ethnicity, HbA1c, and smoking habit, and obtained values of 0.773, 70.0%, and 67.6% for AUC, sensitivity, and specificity, respectively.

In the study by Chatzopoulos et al. (2016), logistic regression was used to build the model. The models were assessed in terms of their accuracy based on sensitivity, specificity, and c-statistics. The best-performing model, with only two self-report questions ("*Has a dentist ever told you that you have periodontal/gingival disease with bone loss?*" and "*Do you think your teeth are loose or wobbly?*") obtained a sensitivity of 80%, specificity of 82.5%, and a C-statistic of 0.83. However, with the addition of age and gender, these values were maximized to 82.1% for sensitivity, 82.2% for specificity, and 0.874 for the c-statistic, which was not verified with the inclusion of diabetes status, smoking, and BMI.

In a study conducted in Portugal by Machado et al. (2022), variables were categorized and dichotomized. Logistic regression analyses were performed to predict periodontal outcomes, and odds ratios (ORs) were determined. Four sets of variables that predicted periodontal outcomes were tested using multivariate binary logistic regression. These sets were (1) 13 self-reported oral health questions; (2) demographic and risk factors, including age, gender, education, smoking, diabetes, and tooth loss; (3) combined self-reported oral health questions and demographic/risk factors; and (4) selection of the most significant subset of predictor variables using the method of all possible equations. Predictive validity was assessed using the area under the curve (AUC), sensitivity, specificity, accuracy, precision, and Akaike's information criterion. The best model for the 2018 definition achieved values of 0.86 AUC, 88.9% sensitivity, 82.5% specificity, 86.4% accuracy, 88.9% precision and 87% AIC. For the 2012 definition, the values were 0.86 AUC, 96.8%, 75%, 88.3%, and 85.9%, respectively. As far as the variables are concerned, the model for the 2018 case included variables related to the issues of gum disease, gum treatment, and bone loss, while the 2012 case included variables related to the issues of gum disease and gum treatment and the tooth loss variable.

Montero et al. (2020) study followed a similar approach to the previous one, also using multivariate logistic regression. Four sets of predictor variables were also tested to predict the prevalence of the three periodontal outcomes. Predictive validity was assessed using metrics such as the area under the ROC curve (AUROCC), likelihood ratio, Hosmer-Lemeshow statistic, AIC, sensitivity, and specificity. In this case, the most effective model included variables related to periodontal disease as well as risk factors such as age, gender, smoking, and tooth loss. In terms of performance for predicting severe periodontitis, this model achieved values of 0.75 AUC, 75.2% sensitivity, 60.6% specificity, 243.5 AIC and 0.76 for the Hosmer-

Lemeshow statistic. Both studies highlighted the importance of combining self-reported and demographic/risk variables in predicting periodontal outcomes.

Finally, in the study by Bashir et al. (2022), the predictors in the datasets were extracted and subjected to pre-processing, which involved recursive feature elimination using cross-validation and dimensionality reduction. Then, 10 machine learning algorithms (AdaBoost, Artificial neural network, decision tree, Gaussian process, K-nearest neighbors, linear support vector classification, linear discriminant analysis, logistic regression, random forests, and Naïve Bayes) were applied to validate the models, both internally, through bootstrapping, and externally, using an alternative dataset of countries. The models were compared based on six performance metrics: AUC, accuracy, sensitivity, specificity, PPV, and NPV. Regarding Taiwanese data, the strongest performing algorithms after recursive feature elimination (RFECV) feature selection were random forests (AUC: 0.97, accuracy: 97.5%), followed by decision trees (AUC: 0.89, accuracy:89.3%). After principal component analysis (PCA), the best-performing algorithms were random forests (AUC: 0.99, accuracy: 99.3%), decision trees (AUC: 0.97, accuracy: 96.8%), and Gaussian process (AUC: 0.79, accuracy: 80.7%). Given the American data, the strongest performing algorithms after RFECV feature selection were K-nearest neighbors (AUC: 1.00, accuracy: 100.0%), followed by random forests (AUC: 0.98, accuracy: 98.1%), and decision trees (AUC: 0.86, accuracy: 86.2 %). After PCA, K-nearest neighbors (AUC: 1.00, accuracy: 100.0%), followed by random forests (AUC: 0.98, accuracy: 98.1%), decision trees (AUC: 0.94, accuracy: 94.1%), and the Gaussian process (AUC: 0.95, accuracy: 92.0%). In the external validation process, that is, testing the models on the counter population, all models experienced a drastic drop in their performance, with accuracy values between 50% and 60%.

In conclusion, these studies used various methodological approaches to investigate the association between variables and to predict periodontal outcomes. Descriptive statistics and machine learning algorithms were used to analyze the data and develop predictive models, with the most used analysis technique being logistic regression.

2.6. Evaluation of SLR Articles

Table 2.8 summarizes the evaluation carried out on 18 RSL articles. Therefore, it is possible to view in detail the classification given to each quality criterion for each article. The definition of the rating scale can be found in Section 2.1.3.

Table 2.8. Evaluation of SLR articles

ID	Criteria										Total
	1	2	3	4	5	6	7	8	9	10	
1	1	1	1	1	1	0.5	1	0,5	1	1	9
2	1	1	1	1	1	0.5	1	0	0.5	0.5	7.5
3	1	0	0.5	0.5	1	0.5	0.5	0	0.5	0.5	5
4	1	0	0.5	0.5	1	0.5	0.5	0	1	0.5	5.5
5	1	1	0.5	1	1	1	1	0	1	1	8.5
6	1	1	0	0.5	1	0.5	0.5	0.5	1	1	7
7	1	1	0.5	0.5	1	1	0.5	0	0.5	1	7
8	1	1	1	1	1	0.5	1	0.5	1	1	9
9	1	1	0.5	0.5	1	0.5	0.5	0	1	1	7
10	1	0	0.5	0.5	1	0.5	0.5	0.5	0.5	0.5	5.5
11	1	1	0.5	0.5	0.5	0.5	0.5	0	1	1	6.5
12	1	0.5	1	0.5	1	0	0.5	0	0.5	1	6
13	1	1	0.5	0.5	1	0.5	0.5	0	0.5	0.5	6.5
14	1	1	1	1	1	0.5	1	0.5	1	1	9
15	1	0.5	1	1	1	0.5	0.5	0	0.5	1	7.5
16	1	1	1	0.5	1	0.5	0.5	0.5	0.5	0.5	7
17	1	1	0.5	0.5	1	0	0.5	0	1	1	6.5
18	1	1	1	1	1	0.5	0.5	0	1	1	8
Total	18	14	12.5	12.5	17.5	9	11.5	3	14	15	

From Table 2.8, it can be observed that the most relevant articles correspond to numbers 1, 5, 8, 14, and 18 (Bashir et al., 2022; Chatzopoulos et al., 2016; Machado et al., 2022; Montero et al., 2019, 2020).

Regarding this set of articles, we realize that their focus is on the prediction of periodontitis and allows us to understand the relationship between some variables and periodontitis. Furthermore, they identify possible risk indicators for periodontal disease.

It is also possible to verify through Table 2.8 that the quality criteria with the highest scores correspond to questions 1, 5, and 10. These questions refer to the research context, the models created and the study's contributions, respectively.

In contrast, the questions with the worst classification are numbers 6 and 8, related to the performance indicators analyzed and the identification of profiles, respectively.

3. Methodology

The methodology used in this dissertation is the Cross-Industry Standard Process for Data Mining (CRISP-DM), which has emerged as a prominent methodological framework that guides the development of data mining projects and is mainly applied to studies in the areas of health and education (Pete et al., 2000). This methodology offers a flexible and iterative structure centered on the primary phases up to the construction and evaluation of the models (Schröer et al., 2021), offering a clear guide to the data mining process (Wirth, 2000).

Figure 3.1 shows the complete sequence that characterizes a data extraction project according to CRISP-DM. The process consists of six phases that are often interconnected, allowing a transition between them regardless of the direction (Pete et al., 2000). In this study, the first stage underwent a small adaptation to understand the problem, as we are facing a health problem related to periodontitis and not a business.

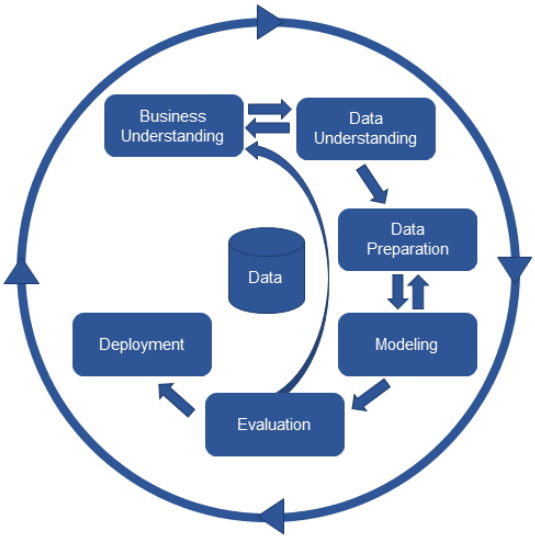

Figure 3.1. CRISP-DM process

3.1. Problem Understanding

The initial stage of each project was to understand the context. This step aims to identify the problem in question and establish the study's objectives. This phase also encompasses formulating a research plan that considers all factors that may impact the investigation, such as available resources, requirements, restrictions, risks, and unforeseen situations. In addition, the limitations, tools, and techniques available for research are considered.

In this investigation, the introduction and literature review chapters correspond to this phase, where a solid foundation is provided to understand the project's purpose and explore the relevant literature.

3.2. Data Understanding

In the second phase, data were analyzed, collected, and described. The data used corresponds to a set of structured data, more precisely, an Excel file related to Portugal. The dataset contained sociodemographic, behavioral, and medical information.

3.2.1. Portuguese Data Set

The data used in this study were provided by the authors of the Almada-Seixal Periodontal Health Study (SoPHiAS) (Botelho et al., 2019), and the purpose of this study was to analyze the prevalence and scope of periodontal diseases in adults in the southern region of the Lisbon Metropolitan Area in Portugal. This study selected the Almada-Seixal Health Centre Grouping as the set of health units studied and focused on individuals over 18 living in the municipalities of Almada and Seixal in Portugal. Finally, information was collected on sociodemographic, behavioral, and medical characteristics obtained through a self-reported questionnaire. The dataset is described in the following sections. The sample included 1,064 participants, and the prevalence of periodontitis was 59.9% (n = 637). The data dictionary can be seen in Appendix A.

3.2.1.1. Sociodemographic Characteristics

Regarding the sample's sociodemographic profile, as it is possible to verify in Appendix B, 58.0% were women, 38.8% were between 61 and 70 years old, 64.3% were married and 46.6% had high school education. Furthermore, it was also found that 52.2% were retired, 89.0% were Portuguese, 50.0% had no children, and 86.3% were Caucasian. The average age, weight, and height of the participants were 61 (±16) years, 73.3 (±14.4) kilograms and 1.64 (±0.09) m, respectively.

As can be seen from the age group variable, also in Appendix B, with increasing age, there is an increase in the prevalence of periodontitis, with the average age of individuals with the disease being 65 (±14) years. The disease is more prevalent in males, with 68.7% and 53.5% in men and women, respectively. In terms of educational level, there was a higher prevalence of the disease among the illiterate individuals (73.8%).

3.2.1.2. Behavioral Characteristics

Considering the variables related to the behavioral characteristics of the participants, described in Appendix C, it was possible to observe a predominance of non-smokers, representing 58.8% of the sample. However, the prevalence of this disease was higher among ex-smokers (71%) and smokers (68.3%).

Regarding behavioral characteristics related to oral hygiene habits, 52.6% brushed their teeth twice daily, 55.5% never performed interproximal hygiene and 64.9% did not use mouthwash. Concerning the last visit to the dentist, 44.9% had already been to the dentist more than two years ago. It can also be concluded that participants diagnosed with the disease had worse oral hygiene habits in terms of tooth brushing and interproximal hygiene and did not go to the dentist very often.

3.2.1.3 Medical Characteristics

As mentioned above, there is a need for greater public awareness about periodontitis, and it was found in the study sample that most respondents (81.9%) were not familiar with this disease (Appendix D).

Considering Appendix D, corresponding to the medical characteristics of all participants, it was observed that 99.3% were not pregnant, 45.1% had occasional dry mouth, 81.6% had no illness, 77% took medication, 77.3% did not take antibiotics in the last six months, 64.1% did not have a dental prosthesis and 53.5% did not have bruxism.

Regarding the group of people with periodontitis, there was a higher prevalence in people with occasional dry mouth (44.3%), people with hypertension (62.8%), people with hypercholesterolemia (55.1%), people with type II diabetes (22.9%) and people using medication (81.9%).

However, about diabetes, a risk indicator identified in the literature, there was a higher prevalence of periodontal disease in people with diabetes (74.0%).

3.3. Data Preparation

The data preparation phase plays a vital role in any investigation. In this phase, raw data are collected, organized, and processed to ensure they are ready for analysis. This can include cleaning data to remove missing values, normalizing formats, coding variables, categorizing

relevant data, and creating variables when necessary. Additionally, the variables to be used in the modeling phase were selected at this stage.

3.3.1. Cleaning, Transforming and Creating Variables

In this phase, errors and inconsistencies found in the data were cleaned and corrected. However, the questionnaire created and used by the authors of the Almada-Seixal Periodontal Health Study (SoPHiAS) (Botelho et al., 2019), before being applied in the data collection phase, was prepared over 6 to 7 months. Different questions and methods that would allow the desired variables to be collected were studied during this time. The authors completed a questionnaire based on the answers obtained from the participants. This preparation and the method used to complete the questionnaire made obtaining quality data possible without the need for exhaustive treatment. Thus, slight data cleaning and correction were performed to prepare the data for analysis by analyzing the quality of the data in the selected variables. Accordingly, missing values were assigned to variables that depended on the response of another, such as the type of diabetes (variable: diabetes_type) and diabetes control (variable: diabetes_control) variables, where the missing values corresponded to participants without diabetes and the value "none" was assigned.

Consequently, the same process was carried out for the variable relating to the type of denture (variable: typeofdenture), where missing data corresponded to people without dentures. Regarding the variable referring to interproximal hygiene (variable: interproximal_hygiene), the writing was corrected and checked to standardize the categorization, such as adding an accent mark to "Sometimes". Finally, a variable relating to the number of years smoked was created (variable: Smoke_years) based on two variables relating to two questions: "How many years have you smoked?" and "How many years did you smoke?".

3.3.2. Variable Selection

When choosing potential predictors for this study, a first approach based on the literature review was adopted. Initially, the variables in the data set recognized in the literature as possible risk indicators for periodontitis were selected (as detailed in Section 2.4).

After the first phase of variable selection, a bivariate analysis was carried out to complement the choice of variables and determine the exclusion or inclusion of the variables in the modeling phase according to the significance of the association. Two separate analyses were carried out following the chi-square analysis to assess the relationship between these

variables (which include quantitative, nominal, and ordinal qualitative variables) and the target variable (of a nominal qualitative nature).

The first analysis aimed to measure the association between quantitative variables and the target. In this case, the ETA (coefficient of nonlinear association) measure was used to assess the relationships between variables. The second analysis aimed to assess the association between the qualitative variables and the target. A measure of association based on the chi-square test of independence, Cramer's V, was used for this analysis.

To interpret the values obtained, the relationship scale defined was as follows: values between 0 and 0.2 represent a very weak relationship, between 0.2 and 0.4 indicates a weak relationship, between 0.4 and 0.7 indicates a moderate relationship, between 0.7 and 0.9 denotes a strong relationship, and between 0.9 and 1 signals a very strong relationship (Laureano, 2020). The measures of association between the variables and the objective can be seen in Appendix E for more detailed visualization. After analyzing the Appendix E, only very weak and weak associations were found.

Finally, 30 variables were selected for the modeling phase. These variables are shown in Table 3.1, divided by sociodemographic, behavioral and medical characteristics.

Table 3.1. Variables selected for the modeling phase

Characteristics	**Variables**
Sociodemographic	Age_group, Marital_status, Employment_status, gender, education, age
Behavioral	Smoking_habit, Smoke_years, Brushing_times_per_day, Interproximal_hygiene, How_many_times_last_7days, Last_dental_visit
Medical	Nodisease, Hypertension, Hypercholesterolemia, Heartdisease, Asthma, Allergies, Anemia, Diabetes, Diabetes_type, Diabetes_control, Medication, Antibiotic_last_6_months, Anti-chol, Antir TG, Anti-hyperglicemia, Denture, Typeofdenture, bruxism_yn

When analyzing the final set of variables, it is worth highlighting the consistency of some of them with the results reported in the literature. The variables related to age were found to be statistically relevant variables and candidates for possible predictors, which aligns with the different studies (Bashir et al., 2022; Kabisch et al., 2022; Machado et al., 2022; Montero et al., 2019; Nobre et al., 2019; Sekundo et al., 2021). Gender was also statistically significant, in line

with some studies (Bashir et al., 2022; Chatzopoulos et al., 2016; Kabisch et al., 2022; Leite et al., 2017; Machado et al., 2022; Montero et al., 2019; Sekundo et al., 2021). Education also emerged as a potential predictor, in line with several studies (Bashir et al., 2022; Machado et al., 2022; Montero et al., 2020; Sekundo et al., 2021). Smoking, also in line with the literature, emerged as a possible candidate predictor (Bashir et al., 2022; Kabisch et al., 2022; Montero et al., 2019, 2020; Nobre et al., 2019; Ramseier et al., 2017; Sekundo et al., 2021). As for the variable related to diabetes, it was also statistically significant, in agreement with the literature (Montero et al., 2019).

3.4. Modeling

Following the initial objectives of the study, the modeling techniques to be used were chosen. We established an evaluation plan and developed various models for subsequent analysis and revision, initiating an interactive process in concurrence with an evaluation phase.

Supervised analytical techniques were used in this study. As a result, supervised classification approaches were applied to identify profiles of individuals with periodontal disease and determine the most important predictors since classification is a prevalent task and is applied in various areas, including medicine, to classify whether a person has the disease (Larose & Larose, 2015).

In the classification process, we usually target a categorical variable, which in this study corresponds to periodontitis (variable called perio_dicom), dichotomized into having (1) or not having the disease (0), and a combination of input variables, in this case, selected and analyzed previously in Section 3.3.1. Regarding the approach adopted for dividing the data, the procedure chosen was holdout, in which 70% of the randomly selected data were used for the training set and 30% for validating the model (Quinn, 2020). First, in the training phase, a set of data is analyzed that includes the selected predictors and the target, where the algorithm detects patterns associated with having or not having periodontitis. Subsequently, new records are analyzed in the test phase, and based on the training, the respective classifications are predicted (Larose & Larose, 2015).

The dataset was balanced before estimating the models. This was because the number of people without periodontitis was lower than that of people with periodontitis, which increased the weight of the minority class and made it possible to balance the weights of the two classes. Subsequently, two algorithms were selected for use in the modeling phase:

- **CHAID** stands for Chi-Square Automatic Interaction Detection and is a trendy technique based on Pearson's chi-square statistical significance test (Quinn, 2020). It was also because one of the most relevant factors of decision trees is related to the construction of decision rules that lead to high interpretability (Larose & Larose, 2015). This model was created following the tree's growth, and all variables in Table 3.1 were included.
- **Logistic Regression** consists of trying to estimate the probability of occurrence of the outcome of a particular category of the dependent variable based on several independent variables (Quinn, 2020). Logistic regression was chosen because it corresponds to the technique most used in the literature in different studies (Arias-Bujanda et al., 2020; Chatzopoulos et al., 2016; Kabisch et al., 2022; Machado et al., 2022; Montero et al., 2019, 2020; Nobre et al., 2019; Sekundo et al., 2021). Regarding logistic regression, this also included all variables in Table *3.1* and was carried out using the stepwise method.

3.5. Evaluation

The evaluation phase corresponds to a process related to analyzing several specific performance specifications derived from the confusion matrix.

In Table 3.2, a confusion matrix can be observed. This matrix presents the number of records correctly and incorrectly classified for each class, where the rows are the actual classifications and the columns are the predictions (Berthold et al., 2020).

Table 3.2. Confusion matrix

		Prediction	
		0	1
Real	0	True Negatives (TN)	False Positives (FP)
	1	False Negatives (FN)	True Positives (TP)

The metrics analyzed to evaluate the performance of the models are mentioned below, referring to their importance (Berthold et al., 2020).

Accuracy and Overall Error Rate

Accuracy (1) is a metric that evaluates the overall performance of a model, indicating the proportion of correct classifications performed. The closer the result is to 1, the more similar the predictions are to the actual values. Its formula is given by:

$$\text{Accuracy} = \frac{TN+TP}{TN+FN+FP+TP} \quad (1)$$

The overall error rate (2) measures the proportion of incorrect classifications. This metric is given by:

$$\text{Overall error rate} = 1 - \text{Accuracy} = \frac{FN+FP}{TN+FN+FP+TP} \quad (2)$$

Sensitivity and Specificity

The sensitivity (3) demonstrates the model's ability to classify records as positive. This measurement is given by:

$$\text{Sensitivity} = \frac{TP}{TP+FN} \quad (3)$$

Specificity (4) reflects a model's ability to classify records as negative and this metric is given by:

$$\text{Specificity} = \frac{TN}{TN+FP} \quad (4)$$

Precision

Regarding precision (5), this metric gives us the perception of among all the positive class classifications that the model makes and how many are correct. This calculation is given by:

$$\text{Precision} = \frac{TP}{TP+FP} \quad (5)$$

F1-score

Finally, the F1-score measures the harmonic mean of precision and sensitivity. This measure incorporates precision and sensitivity into a single metric to understand the model better. This measure is given by:

$$\text{F1-score} = 2\,x\,\frac{\text{Precision x Sensitivity}}{\text{Precision+Sensitivity}} \quad (6)$$

In addition to these measurements, the AUC was analyzed. This measure is based on the principle that the larger the area under the ROC curve, the better the model performs (Berthold et al., 2020). The ROC curve consists of a graphical representation that attempts to illustrate the relationship between the rate at which a model accurately predicts the true positive result (sensitivity) and the rate of false positives it classifies (1-Specificity) (Quinn, 2020). This measure indicates that the larger the area under the curve, the better is the model (Berthold et al., 2020).

3.5.1 Models Evaluation

At this stage, the created models were analyzed based on all the metrics in the training and test sets to understand their behavior and how they could be improved.

After this analysis, the best model was selected between the CHAID algorithm and Logistic Regression according to the performance metrics shown in Table 3.3. It is also possible to analyze the variables included in each model.

Table 3.3. Variables and metrics of each model

Metrics / Models	**CHAID**	**Logistic Regression**
Variables		
Smoking_habit	X	
Bruxism_yn	X	X
Gender	X	X
Age	X	X
Education1	X	X
Smoke_years		X
Diabetes		X
Denture		X
Training		
Accuracy	69.2%	68.5%
Sensitivity	77.1%	70.6%

Metrics / Models	CHAID	Logistic Regression
Training		
Specificity	57.4%	65.4%
Precision	73.2%	75.5%
F1-score	75.1%	73.0%
AUC	0.721	0.75
Test		
Accuracy	66.9%	65.9%
Sensitivity	78.2%	70.2%
Specificity	50.4%	59.7%
Precision	69.7%	71.7%
F1-score	73.7%	71.0%
AUC	0.69	0.728

Table 3.3 shows that the behaviors of both models are similar, both in the training and test sets, which demonstrates the absence of overfitting.

It can also be seen from the sensitivity and specificity that both models are better at classifying people with the disease than those without, which may be related to the fact that the number of people with periodontitis is higher than that of the group without periodontitis.

Finally, to select the best model, the models considered were analyzed to verify their capabilities. Furthermore, a more detailed analysis was conducted of the most relevant metrics according to the study's objective. Thus, since predicting periodontitis is the key point of the study, sensitivity and accuracy were considered crucial for understanding the model's ability to classify people with periodontitis. In addition, the AUC was important in determining the final model, as it indicated the model's prediction performance.

In conclusion, Logistic Regression was selected as the best model for this research because it has the best overall performance indicators.

With the best performance corresponding to logistic regression, the model's structural equation (7) is demonstrated below:

$$P(Y_i = 1|X_i) = \frac{e^{(\alpha+\beta 1X1_i+\beta 2X2_i+\beta 3X3_i+\beta 4X4_i+\beta 5X5_i+\beta 6X6_i+\beta 7X7_i+v_i)}}{1 + e^{(\alpha+\beta 1X1_i+\beta 2X2_i+\beta 3X3_i+\beta 4X4_i+\beta 5X5_i+\beta 6X6_i+\beta 7X7_i+v_i)}} \quad (7)$$

However, after estimating the coefficients, the estimated equation (8) is shown below:

$$P(Y_i = 1|X_i) = \frac{e^{(-2.506+0.037\text{Smoke\_years}_i-0.572(\text{Diabetes=0})_i-0.348(\text{Denture=0})_i+0.742(\text{bruxism\_yn=0})_i-0.427(\text{gender=female})_i+0.035\text{age}_i+0.730(\text{education1=Elementary})_i+0.837(\text{education1=High School})_i+0.691(\text{education1=Illiterate})_i)}}{1+e^{(-2.506+0.037\text{Smoke\_years}_i-0.572(\text{Diabetes=0})_i-0.348(\text{Denture=0})_i+0.742(\text{bruxism\_yn=0})_i-0.427(\text{gender=female})_i+0.035\text{age}_i+0.730(\text{education1=Elementary})_i+0.837(\text{education1=High School})_i+0.691(\text{education1=Illiterate})_i)}} \quad (8)$$

3.6. Deployment

In the final phase, the so-called implementation phase, it is vital to summarize the conclusions and contributions at a scientific level. The following sections discuss these points. First, the results are broken down and discussed in detail. Subsequently, a complete summary of the work carried out is provided, including the contributions and limitations of the study and recommendations for future studies. In short, the implementation culminated in the completion and writing of this dissertation and its presentation.

However, although outside the scope of this study, the implementation could also culminate in integrating the developed model into clinical dental software or an interface, which could assist dentists in screening for periodontitis.

4. Results and Discussion

Regarding the predictive model developed in this study for periodontitis in Portugal, substantial results clearly show its predictive capacity and clinical applicability. Furthermore, the model's performance metrics, more precisely precision, sensitivity, specificity and accuracy, translate the model's effectiveness in distinguishing individuals with and without periodontitis.

With a precision of 71.7%, the model indicated considerable ability to identify periodontitis cases in this sample correctly. Concerning sensitivity, with a value of 70.2%, the model effectively recognized 70.2% of real cases of periodontitis. Regarding specificity, the model correctly identified 59.7% of cases without periodontitis among individuals who do not have the condition, indicating that it can be improved to identify negative cases more accurately. Finally, regarding accuracy, the model's accuracy was calculated at 65.9%, indicating reasonable performance, but it could be improved to increase the assertiveness of the model.

Focusing on the results of studies that used the logistic regression technique and comparing the results regarding the performance metrics obtained by the model created, it appears that concerning sensitivity, this presented a value slightly higher than that of the study by Montero et al. (2019), which obtained 70%. The same did not occur compared to the studies by Chatzopoulos et al. (2016), Machado et al. (2022) and Montero et al. (2020), who achieved 82.1%, 88.9% and 75.2%, respectively. As for specificity, it presented a lower value than all the previously mentioned studies. These divergences may be associated with three main reasons. Firstly, these studies are applied to different populations and contexts, which is one of the main conditions for generalizing the models. Another reason may be related to the inclusion of self-reported questions, given that the models that included these variables presented higher values in terms of metrics (Chatzopoulos et al., 2016; Machado et al., 2022; Montero et al., 2020). Finally, another reason may be associated with having different variables in the models, which may lead to discrepant results.

Therefore, an important step in evaluating a predictive model is undoubtedly the analysis of performance metrics.

In the results of this study, sensitivity and precision presented considerable values, unlike specificity and accuracy, which demonstrated room for improvement.

Concerning this study aiming to predict periodontitis, two metrics were highlighted as the most relevant. Sensitivity is focused on evaluating the model's ability to identify individuals with periodontitis, minimizing false negatives, an essential factor in the medical field. To assess the reliability of the model, precision was considered.

Thus, these metrics are important for a better perception of the model's usefulness, as they demonstrate how accurately it can identify individuals with periodontitis.

Continuous optimization of the model can be performed to improve its ability to discriminate between periodontitis cases and non-cases, thus contributing to an even more effective model for preventing and treating periodontitis in Portugal.

In Figure 4.1, the importance of predictors included in the final model is graphed.

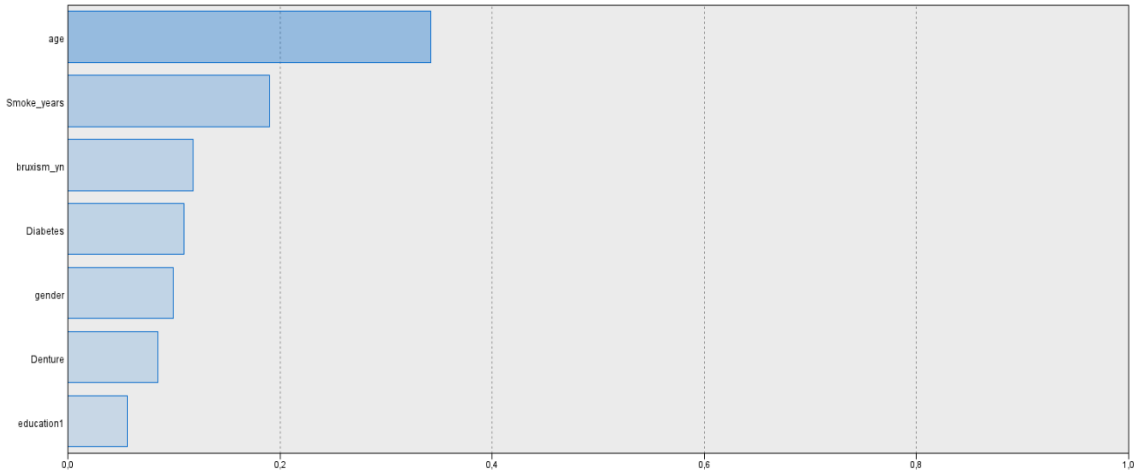

Figure 4.1. Importance of the variables in the Logistic Regression model

Obtained from IBM SPSS Modeler, the Predictor Importance graph analysis provides essential information about the variables or predictors that have the most significant impact on predicting a model's outcome. In this case, we check the presence of two highlighted variables. Age is the most influential predictor in the model, with an importance value of 0.34. This indicates that age plays a significant role in predicting the target (periodontitis in this case). Generally speaking, the more impact a variable has on a prediction, the more significant it is. Thus, according to the model, age is a critical factor in determining the risk of developing periodontitis in Portugal, as evidenced in the literature in Section 2.4. As for the second most important variable, this corresponds to Years of Smoking. Therefore, years of smoking also have a considerable impact, but slightly less than age, with a predictor importance score of 0.19. Consequently, the number of years of smoking also plays an important role in predicting periodontitis, meaning that smoking history is a relevant risk factor for periodontitis, in line with the literature that identifies this habit as an indicator of risk for periodontal disease (Section 2.4).

These results highlight the importance of these two predictors for predicting periodontitis based on the data and model in question. Age and years of smoking are critical variables that

must be considered when assessing the risk and developing prevention and treatment strategies for periodontitis in Portugal.

In Table 4.1, important information regarding the model's estimated coefficients is presented.

Table 4.1. Summary of the effect of each predictor

Target = 1	β	Std.Error	Wald	df	Sig.	Exp(β)	95% Confidence Interval for Exp(β)	
							Lower Bound	Upper Bound
Intercept	-2.506	0.540	21.516	1	<.001			
Smoke_years	0.037	0.006	35.692	1	<.001	1.037	1.025	1.050
Diabetes = 0	-0.572	0.219	6.818	1	0.009	0.565	0.368	0.867
Diabetes = 1	0			0				
Denture = 0	-0.348	0.170	4.165	1	0.041	0.706	0.506	0.986
Denture = 1	0			0				
Bruxism_yn = 0	0.742	0.160	21.583	1	<.001	2.099	1.535	2.871
Bruxism_yn = 1	0			0				
Gender = Female	-0.427	0.171	6.258	1	0.012	0.653	0.467	0.912
Gender = Male	0			0				
age	0.035	0.006	33.573	1	<.001	1.036	1.023	1.048
Education1 = Elementary	0.730	0.293	6.223	1	0.013	2.076	1.169	3.685
Education1 = High School	0.837	0.269	9.676	1	0.002	2.308	1.363	3.911
Education1 = Illiterate	0.691	0.487	2.016	1	0.156	1.996	0.769	5.180
Education1 = University	0			0				

Logistic regression analyses use p-values to determine whether there is a statistically significant relationship between an independent variable and a dependent variable. The relationship is statistically significant if the p-value is below the significance level (0.05). As a result, it was determined by analyzing the p-values associated with each coefficient that all variables had less than 0.05 p-values except the illiterate category of the education variable. This demonstrates that all variables, besides the illiterate category, are statistically significant in predicting periodontitis.

As for the coefficients (β), these represent the natural logarithm of the probability or Exp(β) (odds-ratio) that the dependent variable occurs in relation to non-occurrence, and then the value of Exp(β) translates into how many times the variable multiplies or reduces the chances of having periodontitis.

Therefore, about the years of smoking, with values of 0.037 and 1.037 for β and Exp(β), respectively, it appears that the longer the smoking period, the greater the probability of periodontitis occurring, which multiplies the chances of having the disease 1,037 times. This result is corroborated by previous studies that have documented a relationship between smoking habits and an increased risk of periodontal disease (Bashir et al., 2022; Kabisch et al., 2022; Montero et al., 2019, 2020; Nobre et al., 2019; Ramseier et al., 2017; Sekundo et al., 2021).

Concerning the diabetes variable, regarding having or not having diabetes, it can be concluded that when the diabetes variable is equal to no (0), compared to yes (1), the probability of having periodontitis is lower, reducing the chances of having periodontitis by 0.565 times, as the values for β and Exp(β) are -0.572 and 0.565, respectively. These effects also align with the literature, which indicates an association between diabetes and periodontitis, as concluded in a following study (Montero et al., 2019).

As for the variable relating to having or not, in the case of dentures equal to no (0), this shows values of -0.348 and 0.041 for β and Exp(β), respectively, which shows a lower probability of having periodontitis, reducing the chances of occurrence by 0.041. This conclusion may be related to the fact that people with dentures have experienced tooth loss, which is the most advanced stage of the disease, so this information is corroborated by the literature (Machado et al., 2022; Montero et al., 2020; Schwendicke et al., 2021), which indicates that tooth loss is a risk indicator for periodontitis, which may indicate a history of periodontitis.

Bruxism has also been identified in this study as a variable associated with periodontitis. When the bruxism variable is equal to no (0), compared to yes (1), the chances of having periodontitis are higher, increasing these chances by 2.099, with values of 0.742 for β and 2.099 for Exp(β). This results in an unexpected result since there has been no previous association since then, according to the literature.

For females, the values were -0.427 for β and 0.653 for Exp(β). Consequently, when the gender corresponds to females compared to males, the chances of having periodontitis are lower, reducing the chances of having periodontitis by 0.653. This result is corroborated by various studies in the literature that indicate a prevalence of periodontitis in males (Bashir et al., 2022; Kabisch et al., 2022; Leite et al., 2017; Montero et al., 2019; Sekundo et al., 2021).

In the case of age, the values of β and Exp(β) are 0.035 and 1.036, respectively. These values indicate that with advancing age, the probability of having periodontitis increases, multiplying the hypothesis by 1.036. In most studies in the literature (Bashir et al., 2022; Kabisch et al., 2022; Machado et al., 2022; Montero et al., 2019, 2020; Nobre et al., 2019; Sekundo et al., 2021), age has been shown to have a very large impact on the occurrence of periodontitis, indicating the same effect observed in this study.

Finally, analyzing the education variable, the elementary and high school education categories, compared to university, increase the probability of periodontitis, multiplying it by 2,076 and 2,308, respectively. These results, if we consider the university category to be the highest level of education, are corroborated by the literature, which reports that people with a lower level of education have a higher risk of periodontitis (Bashir et al., 2022; Machado et al., 2022; Montero et al., 2020; Sekundo et al., 2021).

5. Conclusions and Recommendations

5.1. Conclusion

The main objective of this study was to develop a predictive model for Portugal based on statistically significant predictors to correctly and accurately identify people with periodontitis and provide information so that prevention and combat measures can be applied to this condition by part of the medical community.

The final model chosen corresponded to a Logistic Regression, which included variables relating to years of smoking, age, bruxism, diabetes, gender, dentures and education, identified as risk indicators.

Regarding the presence of periodontitis, this model indicated that this risk increases significantly, considering several factors, such as years of smoking, the presence of diabetes, the use of dentures, the absence of bruxism, male gender, age and lower education levels. It is worth noting that in this study, bruxism was identified as a possible risk indicator, and until then, no association between it and periodontal disease had ever been reported in the literature.

Considering the various performance metrics, in the test set, the model presented values for accuracy of 65.9%, sensitivity of 70.2%, specificity of 59.7%, accuracy of 71.7%, F1 score of 71.0% and an AUC of 0.728, which demonstrates considerable overall performance by the main objective of the study.

Overall, we developed a model with moderate accuracy to predict the presence of periodontitis that can assist in identifying risk groups based on the previously mentioned variables. This model may pave the way for future studies to validate this model in other regions and test new modeling strategies for personalized prevention and treatment strategies.

5.2. Limitations

Although this investigation has several strengths, it also has some limitations.

One of the main limitations of the study is that this study has an unbalanced sample in terms of periodontitis cases and cases without the condition, which could lead to bias in the results.

Another area for improvement is the sample size used in the study, as this may affect the capacity of the model in the testing phase and limit its generalization to a wider population.

In addition, the problem of omitted variables related to periodontal indicators can lead to lower model performance.

Finally, this study is valid for the Portuguese population studied. Nonetheless, external validation in other populations is needed to understand the power of generalization and compare risk indicators in different contexts.

5.3. Future Studies

My primary suggestion would be to repeat the study with a representative sample of the universe studied, given that the quality of the data and the representativeness of the sample are elemental to the validity of the results. Periodontitis research must be conducted with samples spanning different demographic groups, age groups, and medical histories to obtain a completer and more accurate picture of this oral condition.

A second concerns the fact that there is a continuous evolution of scientific knowledge. Therefore, it is essential to include variables highlighted by the medical community as possible risk indicators. Medicine and dentistry are constantly discovering new relationships between factors that may play an important role in periodontitis. Therefore, any study on the subject must continually evolve, adapting to the latest discoveries to better understand and prevent this disease.

It would also be interesting to apply the methodology of this study in other contexts for later comparison of risk factors between different geographies and contexts.

Finally, as previously mentioned, it would be interesting to develop a tool or interface based on the created model, enabling dentists to use it as an aid in predicting periodontitis. A tool of this type would be a valuable contribution to dentistry, as it could be designed to analyze a patient's individual risk variables, allowing dentists to take proactive measures to prevent or treat periodontitis, thereby improving the oral health of their patients. patients.

In conclusion, research on periodontitis must evolve with representative samples, covering new risk variables and exploring the possibility of creating personalized tools to help dentists predict and prevent this disease more effectively. These advances can lead to a significant improvement in oral health and quality of life, as well as reduce socioeconomic impact.

References

Antunes, A., Botelho, J., Mendes, J. J., Delgado, A. S., Machado, V., & Proença, L. (2022). Geographical Distribution of Periodontitis Risk and Prevalence in Portugal Using Multivariable Data Mining and Modeling. *International Journal of Environmental Research and Public Health*, *19*(20). https://doi.org/10.3390/ijerph192013634

Arias-Bujanda, N., Regueira-Iglesias, A., Blanco-Pintos, T., Alonso-Sampedro, M., Relvas, M., González-Peteiro, M. M., Balsa-Castro, C., & Tomás, I. (2020). Diagnostic accuracy of IL1β in saliva: The development of predictive models for estimating the probability of the occurrence of periodontitis in non-smokers and smokers. *Journal of Clinical Periodontology*, *47*(6). https://doi.org/10.1111/jcpe.13285

Bashir, N. Z., Rahman, Z., & Chen, S. L. (2022). Systematic comparison of machine learning algorithms to develop and validate predictive models for periodontitis. *Journal of Clinical Periodontology*, *49*(10), 958–969. https://doi.org/10.1111/jcpe.13692

Berthold, M. R., Borgelt, C., Höppner, F., Klawonn, F., & Silipo, R. (2020). *Guide to Intelligent Data Science*. Springer International Publishing. https://doi.org/10.1007/978-3-030-45574-3

Botelho, J., Machado, V., Leira, Y., Proença, L., Chambrone, L., & Mendes, J. J. (2022). Economic burden of periodontitis in the United States and Europe: An updated estimation. *Journal of Periodontology*, *93*(3). https://doi.org/10.1002/JPER.21-0111

Botelho, J., Machado, V., Proença, L., Alves, R., Cavacas, M. A., Amaro, L., & Mendes, J. J. (2019). Study of Periodontal Health in Almada-Seixal (SoPHiAS): a cross-sectional study in the Lisbon Metropolitan Area. *Scientific Reports*, *9*(1). https://doi.org/10.1038/s41598-019-52116-6

Botelho, J., Mascarenhas, P., Viana, J., Proença, L., Orlandi, M., Leira, Y., Chambrone, L., Mendes, J. J., & Machado, V. (2022). An umbrella review of the evidence linking oral health and systemic noncommunicable diseases. *Nature Communications*, *13*(1), 7614. https://doi.org/10.1038/s41467-022-35337-8

Chatzopoulos, G. S., Tsalikis, L., Konstantinidis, A., & Kotsakis, G. A. (2016). A Two-Domain Self-Report Measure of Periodontal Disease Has Good Accuracy for Periodontitis Screening in Dental School Outpatients. *Journal of Periodontology*, *87*(10). https://doi.org/10.1902/jop.2016.160043

Hajishengallis, G., & Chavakis, T. (2021). Local and systemic mechanisms linking periodontal disease and inflammatory comorbidities. In *Nature Reviews Immunology* (Vol. 21, Issue 7). https://doi.org/10.1038/s41577-020-00488-6

Kabisch, S., Hedemann, O. S., & Pfeiffer, A. F. H. (2022). Periodontitis, age-related diseases and diabetes in an endocrinological outpatient setting (PARADIES): a cross-sectional analysis on predictive factors for periodontitis in a German outpatient facility. *Acta Diabetologica*, *59*(5). https://doi.org/10.1007/s00592-021-01838-z

Kassebaum, N. J., Bernabé, E., Dahiya, M., Bhandari, B., Murray, C. J. L., & Marcenes, W. (2014). Global Burden of Severe Periodontitis in 1990-2010. *Journal of Dental Research*, *93*(11). https://doi.org/10.1177/0022034514552491

Kinane, D. F., Stathopoulou, P. G., & Papapanou, P. N. (2017). Periodontal diseases. In *Nature Reviews Disease Primers* (Vol. 3). https://doi.org/10.1038/nrdp.2017.38

Larose, D. T., & Larose, C. D. (2015). Data Mining and Predictive Analytics (Wiley Series on Methods and Applications in Data Mining). *Wiley Series*.

Laureano, R. (2020). *Testes de Hipóteses e Regressão - O meu manual de consulta rápida* (Edições Sílabo, Ed.).

Leite, F. R. M., Peres, K. G., Do, L. G., Demarco, F. F., & Peres, M. A. A. (2017). Prediction of Periodontitis Occurrence: Influence of Classification and Sociodemographic and General Health Information. *Journal of Periodontology*, *88*(8). https://doi.org/10.1902/jop.2017.160607

Li, H., Zhou, J., Zhou, Y., Chen, Q., She, Y., Gao, F., Xu, Y., Chen, J., & Gao, X. (2021). An Interpretable Computer-Aided Diagnosis Method for Periodontitis From Panoramic Radiographs. *Frontiers in Physiology*, *12*. https://doi.org/10.3389/fphys.2021.655556

Machado, V., Lyra, P., Santos, C., Proença, L., Mendes, J. J., & Botelho, J. (2022). Self-Reported Measures of Periodontitis in a Portuguese Population: A Validation Study. *Journal of Personalized Medicine*, *12*(8). https://doi.org/10.3390/jpm12081315

Machado, V., Proença, L., Morgado, M., Mendes, J. J., & Botelho, J. (2020). Accuracy of panoramic radiograph for diagnosing periodontitis comparing to clinical examination. *Journal of Clinical Medicine*, *9*(7). https://doi.org/10.3390/JCM9072313

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Journal of Clinical Epidemiology*, *62*(10). https://doi.org/10.1016/j.jclinepi.2009.06.005

Montero, E., Herrera, D., Sanz, M., Dhir, S., Van Dyke, T., & Sima, C. (2019). Development and validation of a predictive model for periodontitis using NHANES 2011–2012 data. *Journal of Clinical Periodontology*, *46*(4). https://doi.org/10.1111/jcpe.13098

Montero, E., La Rosa, M., Montanya, E., Calle-Pascual, A. L., Genco, R. J., Sanz, M., & Herrera, D. (2020). Validation of self-reported measures of periodontitis in a Spanish Population. *Journal of Periodontal Research*, *55*(3). https://doi.org/10.1111/jre.12724

Nobre, M. D. A., Ferro, A., & Maló, P. (2019). Adult patient risk stratification using a risk score for periodontitis. *Journal of Clinical Medicine*, *8*(3). https://doi.org/10.3390/jcm8030307

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. In *The BMJ* (Vol. 372). https://doi.org/10.1136/bmj.n71

Pete, C., Julian, C., Randy, K., Thomas, K., Thomas, R., Colin, S., & Wirth, R. (2000). Crisp-Dm 1.0. *CRISP-DM Consortium*.

Quinn, J. (2020). *The Insiders' Guide To Predictive Analytics*. Smart Vision Europe.

Ramseier, C. A., Anerud, A., Dulac, M., Lulic, M., Cullinan, M. P., Seymour, G. J., Faddy, M. J., Bürgin, W., Schätzle, M., & Lang, N. P. (2017). Natural history of periodontitis: Disease progression and tooth loss over 40 years. *Journal of Clinical Periodontology*, *44*(12). https://doi.org/10.1111/jcpe.12782

Schenkein, H. A., Papapanou, P. N., Genco, R., & Sanz, M. (2020). Mechanisms underlying the association between periodontitis and atherosclerotic disease. In *Periodontology 2000* (Vol. 83, Issue 1). https://doi.org/10.1111/prd.12304

Schröer, C., Kruse, F., & Gómez, J. M. (2021). A systematic literature review on applying CRISP-DM process model. *Procedia Computer Science*, *181*. https://doi.org/10.1016/j.procs.2021.01.199

Schwendicke, F., Arsiwala, L. T., Krois, J., Bäumer, A., Pretzl, B., Eickholz, P., Petsos, H., Kocher, T., Holtfreter, B., & Graetz, C. (2021). Association, prediction, generalizability: Cross-center validity of predicting tooth loss in periodontitis patients. *Journal of Dentistry*, *109*. https://doi.org/10.1016/j.jdent.2021.103662

Schwendicke, F., Schmietendorf, E., Plaumann, A., Sälzer, S., Dörfer, C. E., & Graetz, C. (2018). Validation of multivariable models for predicting tooth loss in periodontitis patients. *Journal of Clinical Periodontology*, *45*(6). https://doi.org/10.1111/jcpe.12900

Sekundo, C., Bölk, T., Kalmus, O., & Listl, S. (2021). Accuracy of a 7-item patient-reported stand-alone tool for periodontitis screening. *Journal of Clinical Medicine*, *10*(2). https://doi.org/10.3390/jcm10020287

Trindade, D., Carvalho, R., Machado, V., Chambrone, L., Mendes, J. J., & Botelho, J. (2023). Prevalence of periodontitis in dentate people between 2011 and 2020: A systematic review and meta-analysis of epidemiological studies. In *Journal of Clinical Periodontology*. https://doi.org/10.1111/jcpe.13769

Wirth, R. (2000). CRISP-DM : Towards a Standard Process Model for Data Mining. *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining*, *24959*.

Appendixes

Appendix A. Portugal data dictionary

Variable	Type	Description
Age_group	OQ	Age group to which the person belongs
Marital_status	NQ	Person's marital status
Nationality	NQ	Person's nationality
Employment_status	NQ	Person's employment status
Pregnant	NQ	If the person has already been pregnant
Pregnancy_weeks	DQ	How many weeks pregnant
Previously_pregnant	NQ	If they have been pregnant before
Spontaneous_miscarriged	NQ	If the person has ever had a miscarriage
Children	DQ	Number of children
Premature_birth	NQ	If any were born prematurely
Frequency dry mouth	OQ	How often does the person feel dry mouth
Weight	CQ	The person's weight
Height	CQ	The person's height
Smoking_habit	NQ	Whether or not you smoke (or have ever smoked)
Smoked_years	CQ	For how many years did the person smoke?
Years_smoker	CQ	How many years has the person smoked?
Nodisease	NQ	Do you have any disease or not
Hypertension	NQ	Does or does not have the disease
Hypercholesterolemia	NQ	Does or does not have the disease
Heartdisease	NQ	Does or does not have the disease
Kidneydisease	NQ	Does or does not have the disease
Asthma	NQ	Does or does not have the disease
Tyroid	NQ	Does or does not have the disease
Stomachorduodenalulcer	NQ	Does or does not have the disease
Aggressiveperiodontitis	NQ	Does or does not have the disease
HIV	NQ	Does or does not have the disease

Variable	Type	Description
Allergies	NQ	Does or does not have the disease
Other	NQ	Does or does not have the disease
Osteoporosis	NQ	Does or does not have the disease
Stroke	NQ	Does or does not have the disease
Infarct	NQ	Does or does not have the disease
Depression	NQ	Does or does not have the disease
Anxiety	NQ	Does or does not have the disease
RheumatoidArthritis	NQ	Does or does not have the disease
prostateenlargement	NQ	Does or does not have the disease
Sleepingapnea	NQ	Does or does not have the disease
Cancer	NQ	Does or does not have the disease
Chronicgastritis	NQ	Does or does not have the disease
Fibromyalgia	NQ	Does or does not have the disease
COPD	NQ	Does or does not have the disease
Systemiclupuserythematosus	NQ	Does or does not have the disease
Hyperuricemia	NQ	Does or does not have the disease
Anemia	NQ	Does or does not have the disease
Diabetes	NQ	Does or does not have the disease
Diabetes_type	OQ	What type of diabetes
Diabetes_control	NQ	How you manage your diabetes
Medication	NQ	If the person takes medication
Antibiotic_last_6_months	NQ	If the person has taken antibiotics in the last 6 months
Medicaçãoregulardosedia	NQ	What medication does the person take regularly
Anti-chol	NQ	If the person takes cholesterol medication
Antir TG	NQ	If the person takes triglyceride medication
Anti-hyperglycemia	NQ	If the person takes hyperglycemia medication
Ever_did_periodontal_treatment	NQ	If the person has already had periodontal treatment

Variable	Type	Description
Denture	NQ	If the person has denture
Typeofdenture	NQ	What type of denture
Brushing_times_per_day	DQ	Number of times you brush your teeth a day
Interproximal_hygiene	NQ	If the person performs interproximal hygiene
How_many_times_last_7days	DQ	How many times have you performed interproximal hygiene in the last 7 days?
Mouthwash	NQ	If the person uses mouthwash
Last_dental_visit	OQ	When was the person's last visit to the dentist?
Do_you_know_what_is_periodontal_disease	NQ	If the person knows what periodontal disease is
Brushing_hand	NQ	Which hand does the person brush their teeth with
Perio_dicom	NQ	If the person has periodontitis
Bruxism_yn	NQ	If the person has bruxism
Age	DQ	How old the person is
Gender	NQ	The person's gender
Race	NQ	The person's ethnicity
Education1	OQ	The person's level of education
BMI	CQ	The person’s body mass index
Systolic	DQ	Systolic blood pressure
Diastolic	DQ	Diastolic blood pressure
Pulse	DQ	Pulse rate

Note: NQ – Nominal qualitative variable; OQ – Ordinal qualitative variable; DQ – Discrete quantitative variable; CQ – Continuous quantitative variable.

Appendix B. Description of sociodemographic data

	Periodontitis n (%)	Without n (%)	Total n (%)
Categorical variables			
Gender ***			
Male	307 (48.2)	140 (32.8)	447 (42.0)
Female	330 (51.8)	287 (67.2)	617 (58.0)
Age group (years) ***			
18-30	11 (1.7)	51 (11.9)	62 (5.8)
31-40	26 (4.1)	49 (11.5)	75 (7.0)
41-50	63 (9.9)	73 (17.1)	136 (12.8)
51-60	82 (12.9)	55 (12.9)	137 (12.9)
61-70	220 (34.5)	108 (25.3)	328 (30.8)
71-80	170 (26.7)	74 (17.3)	244 (22.9)
>80	65 (10.2)	17 (4.0)	82 (7.7)
Marital status ***			
Single	66 (10.4)	104 (24.4)	170 (16.0)
Married/Union of fact	422 (66.2)	262 (61.4)	684 (64.3)
Divorced	70 (11.0)	33 (7.7)	103 (9.7)
Widowed	79 (12.4)	28 (6.6)	107 (10.1)
Educational level ***			
Illiterate	31 (4.7)	11 (2.6)	42 (4.0)
Elementary	276 (43.3)	134 (31.4)	410 (38.5)
High School	287 (45.1)	209 (48.9)	496 (46.6)
University	43 (6.8)	73 (17.1)	116 (10.9)
Employment status ***			
Student	1 (0.2)	18 (4.2)	19 (1.8)
Employed	162 (25.4)	165 (38.6)	327 (30.7)
Unemployed	84 (13.2)	79 (18.5)	163 (15.3)
Retired	390 (61.2)	165 (38.6)	555 (52.2)
Nationality			
Angola	4 (0.6)	7 (1.6)	11 (1.0)
Angola/Portugal	4 (0.6)	4 (0.9)	8 (0.8)

	Periodontitis n (%)	Without n (%)	Total n (%)
Belgium	0 (0)	1 (0.2)	1 (0.1)
Brazil	11 (1.7)	11 (2.6)	22 (2.1)
Brazil/Portugal	3 (0.5)	5 (1.2)	8 (0.8)
Cape Verde	14 (2.2)	11 (2.6)	25 (2.3)
Cape Verde/Portugal	5 (0.8)	2 (0.5)	7 (0.7)
France/Portugal	0 (0)	1 (0.2)	1 (0.1)
Gabon	0 (0)	1 (0.2)	1 (0.1)
Guinea-Bissau	2 (0.3)	1 (0.2)	3 (0.3)
India/Portugal	0 (0)	2 (0.5)	2 (0.2)
Moldova/Portugal	1 (0.2)	0 (0)	1 (0.1)
Mozambique	0 (0)	2 (0.5)	2 (0.2)
Mozambique/Portugal	1 (0.2)	0 (0)	1 (0.1)
Poland	0 (0)	1 (0.2)	1 (0.1)
Portugal	580 (91.1)	367 (85.9)	947 (89.0)
Portugal/USA	1 (0.2)	0 (0)	1 (0.1)
São Tomé Island	4 (0.6)	9 (2.1)	13 (1.2)
São Tomé Island/Portugal	2 (0.3)	2 (0.5)	4 (0.4)
South Africa/Portugal	1 (0.2)	0 (0)	1 (0.1)
Sudan	1 (0.2)	0 (0)	1 (0.1)
Ukraine/Portugal	1 (0.2)	0 (0)	1 (0.1)
United Kingdom	1 (0.2)	0 (0)	1 (0.1)
Venezuela/Portugal	1 (0.2)	0 (0)	1 (0.1)
Children			
0	336 (52.7)	196 (45.9)	532 (50.0)
1	98 (15.4)	79 (18.5)	177 (16.6)
2	132 (20.7)	99 (23.2)	231 (21.7)
3	36 (5.7)	26 (6.1)	62 (5.8)
4	17 (2.7)	11 (2.6)	28 (2.6)
5	9 (1.4)	10 (2.3)	19 (1.8)
6	8 (1.3)	6 (1.4)	14 (1.3)
8	1 (0.2)	0 (0)	1 (0.1)
Race			

	Periodontitis n (%)	Without n (%)	Total n (%)
Caucasian	558 (87.6)	360 (84.3)	918 (86.3)
Black	75 (11.8)	60 (14.1)	135 (12.7)
Mongoloid (Asian)	4 (0.6)	7 (1.6)	11 (1.0)

	Periodontitis	Without	Total
Age ***	65 (±14)	55 (±18)	61 (±16)
Weight	73.9 (±14.2)	72.3 (±14.7)	73.3 (±14.4)
Height	1.64 (±0.09)	1.63 (±0.09)	1.64 (±0.09)

Appendix C. Description of Behavioral data

	Periodontitis n (%)	Without n (%)	Total n (%)
Smoking status ***			
Non-smoker	330 (51,8)	296 (69.3)	626 (58,8)
Ex-smoker	208 (32,7)	85 (19.9)	293 (27,5)
Smoker	99 (15,5)	46 (10.8)	145 (13,6)
Brushing times per day **			
0	24 (3.8)	7 (1.6)	31 (2.9)
1	195 (30.6)	107 (25.1)	302 (28.4)
2	328 (51.5)	232 (54.3)	560 (52.6)
3	83 (13.0)	68 (15.9)	151 (14.2)
4	7 (1.1)	9 (2.1)	16 (1.5)
5	0 (0)	3 (0.7)	3 (0.3)
6	0 (0)	1 (0.2)	1 (0.1)
Interproximal hygiene ***			
Never	396 (62.2)	194 (45.4)	590 (55.5)
No	68 (10.7)	60 (14.1)	128 (12.0)
Sometimes	80 (12.6)	81 (19.0)	161 (15.1)
Yes	93 (14.6)	92 (21.5)	185 (17.4)

	Periodontitis n (%)	Without n (%)	Total n (%)
How many times last 7 days ***			
0	518 (81.3)	300 (70.3)	818 (76.9)
1	13 (2.0)	18 (4.2)	31 (2.9)
2	24 (3.8)	15 (3.5)	39 (3.7)
3	10 (1.6)	20 (4.7)	30 (2.8)
4	6 (0.9)	11 (2.6)	17 (1.6)
5	8 (1.3)	5 (1.2)	13 (1.2)
6	2 (0.3)	1 (0.2)	3 (0.3)
7	56 (8.8)	57 (13.3)	113 (10.6)
Mouthwash			
No	408 (64.1)	283 (66.3)	691 (64.9)
Yes	229 (35.9)	144 (33.7)	373 (35.1)
Last dental visit ***			
Never	6 (0.9)	11 (2.6)	17 (1.6)
<6 months	154 (24.2)	146 (34.2)	300 (28.2)
6-12 months	88 (13.8)	63 (14.8)	151 (14.2)
12-24 months	70 (11.0)	48 (11.2)	118 (11.1)
>2 years	319 (50.1)	159 (37.2)	478 (44.9)
Brushing hand			
Left	18 (2.8)	13 (3.0)	31 (2.9)
Right	619 (97.2)	414 (97.0)	1033(97.1)

	Periodontitis	Without	Total
Smoke_years ***	5.56 (±11.24)	12.74 (±17.02)	9.86 (±15.37)

Appendix D. Description of Medical data

	Periodontitis n (%)	Without n (%)	Total n (%)
Categorical variables			
Pregnant			
No	634 (99.5)	423 (99.1)	1057 (99.3)
Yes	3 (0.5)	4 (0.9)	7 (0.7)
Pregnancy weeks			
0	634 (99.5)	423 (99.1)	1057 (99.3)
2	0 (0)	1 (0.2)	1 (0.1)
14	0 (0)	1 (0.2)	1 (0.1)
15	1 (0.2)	1 (0.2)	2 (0.2)
19	1 (0.2)	0 (0)	1 (0.1)
23	1 (0.2)	0 (0)	1 (0.1)
32	0 (0)	1 (0.2)	1 (0.1)
Previously pregnant **			
No	331 (52.0)	193 (45.2)	524 (49.2)
Yes	306 (48.0)	234 (54.8)	540 (50.8)
Spontaneous miscarriged			
No	558 (87.6)	362 (84.8)	920 (86.5)
Yes	79 (12.4)	65 (15.2)	144 (13.5)
Premature birth			
No	609 (95.6)	404 (94.6)	1013 (95.2)
Yes	28 (4.4)	23 (5.4)	51 (4.8)
Frequency Dry mouth			
Never	241 (37.8)	170 (39.8)	411 (38.6)
Occasionally	282 (44.3)	198 (46.4)	480 (45.1)
Frequently	103 (16.2)	49 (11.5)	152 (14.3)
Always	11 (1.7)	10 (2.3)	21 (2.0)
Nodisease ***			
No	547 (85.9)	321 (75.2)	868 (81.6)
Yes	90 (14.1)	106 (24.8)	196 (18.4)
RheumatoidArthritis			

	Periodontitis n (%)	Without n (%)	Total n (%)
No	631 (99.1)	426 (99.8)	1057 (99.3)
Yes	6 (0.9)	1 (0.2)	7 (0.7)
Prostateenlargement *			
No	599 (94.0)	412 (96.5)	1011 (95.0)
Yes	38 (6.0)	15 (3.5)	53 (5.0)
Sleepingapnea			
No	628 (98.6)	420 (98.4)	1048 (98.5)
Yes	9 (1.4)	7 (1.6)	16 (1.5)
Cancer			
No	609 (95.6)	415 (97.2)	1024 (96.2)
Yes	28 (4.4)	12 (2.8)	40 (3.8)
Chronicgastritis			
No	635 (99.7)	424 (99.3)	1059 (99.5)
Yes	2 (0.3)	3 (0.7)	5 (0.5)
Fibromyalgia			
No	629 (98.7)	421 (98.6)	1050 (98.7)
Yes	8 (1.3)	6 (1.4)	14 (1.3)
COPD			
No	633 (99.4)	427 (100)	1060 (99.6)
Yes	4 (0.6)	0 (0)	4 (0.4)
Systemiclupuserythematosus			
No	636 (99.8)	424 (99.3)	1060 (99.6)
Yes	1 (0.2)	3 (0.7)	4 (0.4)
Hyperuricemia			
No	628 (98.6)	423 (99.1)	1051 (98.8)
Yes	9 (1.4)	4 (0.9)	13 (1.2)
Hypertension * **			
No	237 (37.2)	239 (56.0)	476 (44.7)
Yes	400 (62.8)	188 (44.0)	588 (55.3)
Hypercholesterolemia * **			

	Periodontitis n (%)	Without n (%)	Total n (%)
No	286 (44.9)	260 (60.9)	546 (51.3)
Yes	351 (55.1)	167 (39.1)	518 (48.7)
Heartdisease ***			
No	507 (79.6)	367 (85.9)	874 (82.1)
Yes	130 (20.4)	60 (14.1)	190 (17.9)
Kidneydisease			
No	637 (100)	427 (100)	1064 (100)
Yes	0 (0)	0 (0)	0 (0)
Asthma **			
No	621 (97.5)	405 (94.8)	1026 (96.4)
Yes	16 (2.5)	22 (5.2)	38 (3.6)
Tyroid			
No	583 (91.5)	387 (90.6)	970 (91.2)
Yes	54 (8.5)	40 (9.4)	94 (8.8)
Stomachorduodenalulcer			
No	604 (94.8)	408 (95.6)	1012 (95.1)
Yes	33 (5.2)	19 (4.4)	52 (4.9)
Agressiveperiodontitis			
No	637 (100)	427 (100)	1064 (100)
Yes	0 (0)	0 (0)	0 (0)
HIV			
No	637 (100)	427 (100)	1064 (100)
Yes	0 (0)	0 (0)	0 (0)
Allergies ***			
No	617 (96.9)	394 (92.3)	1011 (95.0)
Yes	20 (3.1)	33 (7.7)	53 (5.0)
Other			
No	516 (81.0)	337 (78.9)	853 (80.2)
Yes	121 (19.0)	90 (21.1)	211 (19.8)
Osteoporosis *			
No	602 (94.5)	413 (96.7)	1015 (95.4)
Yes	35 (5.5)	14 (3.3)	49 (4.6)

	Periodontitis n (%)	Without n (%)	Total n (%)
Stroke *			
No	629 (98.7)	426 (99.8)	1055 (99.2)
Yes	8 (1.3)	1 (0.2)	9 (0.8)
Infarct			
No	630 (98.9)	425 (99.5)	1055 (99.2)
Yes	7 (1.1)	2 (0.5)	9 (0.8)
Depression *			
No	594 (93.2)	385 (90.2)	979 (92.0)
Yes	43 (6.8)	42 (9.8)	85 (8.0)
Anxiety			
No	615 (96.5)	415 (97.2)	1030 (96.8)
Yes	22 (3.5)	12 (2.8)	34 (3.2)
Anemia **			
No	634 (99.5)	419 (98.1)	1053 (99.0)
Yes	3 (0.5)	8 (1.9)	11 (1.0)
Diabetes ***			
No	486 (76.3)	374 (87.6)	860 (80.8)
Yes	151 (23.7)	53 (12.4)	204 (19.2)
Diabetes type ***			
None	486 (76.3)	374 (87.6)	860 (80.8)
Type I	5 (0.8)	6 (1.4)	11 (1.0)
Type II	146 (22.9)	47 (11.0)	193 (18.1)
Diabetes control ***			
Food	2 (0.3)	1 (0.2)	3 (0.3)
Insulin	17 (2.7)	8 (1.9)	25 (2.3)
Insulin, Medication	3 (0.5)	0 (0)	3 (0.3)
Medication	125 (19.5)	42 (9.8)	166 (15.6)
Medication, Insulin	5 (0.8)	2 (0.5)	7 (0.7)
None	486 (76.3)	374 (87.6)	860 (80.8)
Medication ***			
No	115 (18.1)	127 (29.7)	242 (22.7)
Yes	522 (81.9)	300 (70.3)	822 (77.3)

	Periodontitis n (%)	Without n (%)	Total n (%)
Antibiotic last 6 months **			
No	506 (79.4)	311 (72.8)	817 (76.8)
Yes	131 (20.6)	116 (27.2)	247 (23.2)
Anti-chol ***			
No	353 (55.4)	299 (70.0)	652 (61.3)
Yes	284 (44.6)	128 (30.0)	412 (38.7)
Antir TG ***			
No	353 (55.4)	299 (70.0)	652 (61.3)
Yes	284 (44.6)	128 (30.0)	412 (38.7)
Anti-hyperglicemia ***			
No	504 (79.1)	380 (89.0)	884 (83.1)
Yes	133 (20.9)	47 (11.0)	180 (16.9)
Ever did periodontal treatment			
No	621 (97.5)	420 (98.4)	1041 (97.8)
Yes	16 (2.5)	7 (1.6)	23 (2.2)
Denture ***			
No	368 (57.8)	314 (73.5)	682 (64.1)
Yes	269 (42.2)	113 (26.5)	382 (35.9)
Type of denture ***			
None	367 (57.6)	314 (73.5)	681 (64.0)
Other	6 (0.9)	9 (2.1)	15 (1.4)
Acrylic removable denture	191 (30.0)	64 (15.0)	255 (24.0)
Acrylic removable prosthesis + Skeletal removable prosthesis	0 (0)	1 (0.2)	1 (0.1)
Removable skeletal prosthesis	70 (11.0)	37 (8.7)	107 (10.1)
Flexible removable prosthesis	3 (0.5)	2 (0.5)	5 (0.5)

	Periodontitis n (%)	Without n (%)	Total n (%)
Do you kow what is periodontal disease?			
No	531 (83.4)	340 (79.6)	871 (81.9)
Yes	106 (16.6)	87 (20.4)	193 (18.1)
Bruxism ***			
No	370 (58.1)	199 (46.6)	569 (53.5)
Yes	267 (41.9)	228 (53.4)	495 (46.5)

	Periodontitis	Without	Total
BMI	27.5 (±4.7)	27.1 (±4.9)	27.3 (±4.8)
Systolic	136 (±20)	129 (±20)	134 (±21)
Diastolic	79 (±14)	78 (±13)	79 (±14)
Pulse	76 (±12)	75 (±12)	76 (±12)

Note: (*) – p-value < 10%, (**) – p-value < 5%, (***) – p-value < 1%.

Appendix E. Association measures

Variable	Type	Association measures
Age_group	OQ	0.315
Marital_status	NQ	0.203
Employment_status	NQ	0.252
Smoking_habit	NQ	0.175
Smoke_years	OQ	0.316
Nodisease	NQ	0.135
Hypertension	NQ	0.185
Hypercholesterolemia	NQ	0.157
Heartdisease	NQ	0.081
Asthma	NQ	0.070
Allergies	NQ	0.103
Anemia	NQ	0.068
Diabetes	NQ	0.141
Diabetes_type	OQ	0.153

Variable	Type	Association measures
Diabetes_control	NQ	0.145
Medication	NQ	0.137
Antibiotic_last_6_months	NQ	0.077
Anti-chol	NQ	0.147
Antir TG	NQ	0.147
Anti-hyperglycemia	NQ	0.129
Denture	NQ	0.161
Typeofdenture	NQ	0.193
Brushing_times_per_day	DQ	0.124
Interproximal_hygiene	OQ	0.166
How_many_times_last_7days	DQ	0.158
Last_dental_visit	OQ	0.147
Bruxism_yn	NQ	0.113
Gender	NQ	0.153
Education1	OQ	0.191
Age	DQ	0.394