WHITEETH

Project III, Data Analysis

Carla Deveaud Sánchez
Jezabel Esbrí Rodríguez
Evgeny Grachev
Yassmina Jebbour Maamri
Manuel Rocamora Valenti

CONTENT INDEX.

1 Abstract. 3
2 Introduction. 3
2.1 Context and need. 3
2.2 Task and objective. 3
3 Materials and Methods. 4
3.1 Methodology used. 4
3.2 Technology used. 5
3.3 Data. 5
3.4 Data preparation. 6
3.5 Statistical analysis. 8
4 Results. 13
5 Discussion. 15
6 Legacy. 16
7 Conclusion. 17
8 Acknowledgements. 17
9 References. 18
10 Annexes. 19

FIGURES INDEX.

Figure 1. Methodology and steps followed (modified CRISP-DM). 4
Figure 2. Bar plot of the distribution of the intervention duration sorted by frequency. 7
Figure 3. Bar plots of the distribution of relevant variables. 9
Figure 4. Bar plots showing the differences between the means of implantology patients and peri-implantology patients. 10
Figure 5. Contribution of the individuals to the first and second dimensions of the PCA according to the type of intervention. 11
Figure 6. Bar plot of the contribution of the variables to the first dimension of the PCA. 11
Figure 7. Bar plot of the 10 most important features obtained with Random Forest. 12
Figure 8. Contribution of the main variables to the first and second dimensions of the PCA. ... 13
Figure 9. Scatterplot of real values versus predicted values with the Regression Tree. 14

1 ABSTRACT.

This study, in collaboration with the Master's program in Continuing Education in Periodontology and Peri-implantology at the University of Barcelona, focused on improving the efficiency of dentoalveolar surgery procedures, such as tooth extractions and implant placements, addressing challenges such as peri-implantitis. The main objectives were to optimize the management of the duration and difficulty of surgical interventions and to improve the assignment of operations to students according to their level of experience.

The results identified critical variables that influence the duration and complications of surgeries, enabling the implementation of two predictive models in a Streamlit application for clinical use. These models not only predict the time required for each procedure but also classify interventions by difficulty, improving the planning and training of both patients and students. The study demonstrates that data science can optimize resource management and patient care in clinical learning environments, significantly contributing to dental practice.

2 INTRODUCTION.

2.1 CONTEXT AND NEED.

This dental project arises from collaboration with the Master's program in Continuing Education in Periodontology and Peri-implantology at the University of Barcelona, where students carry out dentoalveolar surgery procedures under professional supervision. Dentoalveolar surgery is a subdiscipline of oral and maxillofacial surgery that focuses on interventions related to teeth and the alveolar bone in the oral cavity, such as tooth extractions and implant placements (Sociedad Española de Cirugía Oral y Maxilofacial y de Cabeza y Cuello, 2023).

Despite advances in surgical technique and materials used in dental implant placement, the possibility of complications such as peri-implantitis still persists. This inflammatory condition can lead to the loss of the implant and surrounding tissue, negatively impacting the patient's quality of life and potentially requiring additional interventions.

In addition to addressing patients' health issues, those responsible for the master's program seek to optimize the management of their procedures to improve the quality of care and make better use of clinic resources. This helps avoid waste of time and resources by accommodating more students in the master's program, aiming to reach the optimal number of students for good training and thus improve the quality of patient care.

2.2 TASK AND OBJECTIVE.

As previously mentioned, the director of the master's program, Dr. Rui Figueiredo, has expressed interest in improving its functioning. Therefore, the main objective is to optimize the management of clinical consultations through a Streamlit application, by analyzing the duration and potential complications of operations, in order to administer resources more efficiently.

To achieve the objective most accurately, the following subtasks have been carried out:

1. **Identify patient profiles with higher risk.** With this aim, we seek to determine the profile of patients at higher risk of experiencing postoperative complications. This will allow us to identify which types of procedures may be more delicate and require specialized handling.
2. **Predict the duration of the intervention** for each patient. Predictive models based on data will be developed to estimate the duration of the surgical intervention, which is closely related to its difficulty. This involves analyzing factors related to both the operations and the oral health of the patients.
3. **Assign operations to different students according to their experience** and the level of operative difficulty. Finally, by managing operations based on their duration and difficulty, they are assigned to students according to their level and in an optimal way to maximize clinic resources. This could allow, in the future, for more students to be accommodated in the clinic, ensuring optimal coverage of available treatment rooms.

As previously stated, this information will be available to students and clinic staff through an application. They will simply need to input crucial operation and patient data, and the application will automatically provide all the requested information.

3 Materials and Methods.

3.1 Methodology used.

Regarding the methodology used to carry out the work, CRISP-DM was employed, as it is one of the methodologies that best fits the typical steps of a data science project and is one of the most popular in the industry. Additionally, some additions were made to this typical cycle to better fit the context of this project. The steps followed are shown in *Figure 1*.

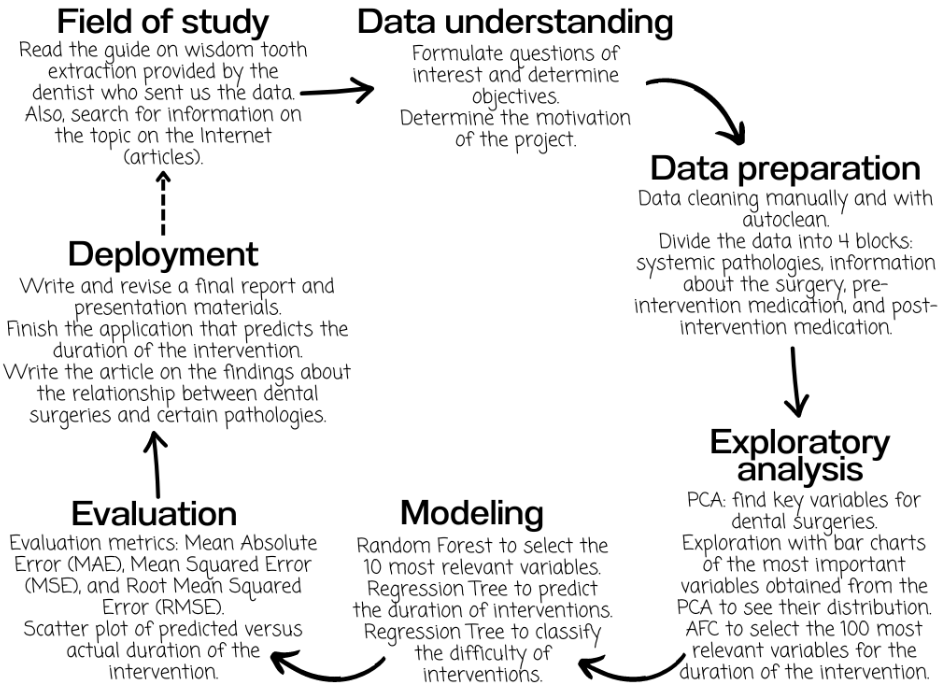

Field of study
Read the guide on wisdom tooth extraction provided by the dentist who sent us the data. Also, search for information on the topic on the Internet (articles).
Data understanding
Formulate questions of interest and determine objectives. Determine the motivation of the project.
Data preparation
Data cleaning manually and with autoclean. Divide the data into 4 blocks: systemic pathologies, information about the surgery, pre-intervention medication, and post-intervention medication.
Exploratory analysis
PCA: find key variables for dental surgeries. Exploration with bar charts of the most important variables obtained from the PCA to see their distribution. AFC to select the 100 most relevant variables for the duration of the intervention.
Modeling
Random Forest to select the 10 most relevant variables. Regression Tree to predict the duration of interventions. Regression Tree to classify the difficulty of interventions.
Evaluation
Evaluation metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). Scatter plot of predicted versus actual duration of the intervention.
Deployment
Write and revise a final report and presentation materials. Finish the application that predicts the duration of the intervention. Write the article on the findings about the relationship between dental surgeries and certain pathologies.

Figure 1. Methodology and steps followed (modified CRISP-DM).

3.2 TECHNOLOGY USED.

During the development of this project, various technological challenges were faced due to the complexity of the database. The selection of specific tools and languages responded to both the project's needs and familiarity with available technologies, ensuring efficiency and effectiveness in each phase.

First, for project planning, The Gantt Project was used. This tool was chosen for being one of the most well-known and easy-to-understand planners. Additionally, the application offers predefined templates, simplifying time and task management, thus facilitating project organization and tracking.

For data cleaning, preparation, and exploration, the R language was used. This language is specialized in statistics and data analysis, making it ideal for these tasks. R has a large number of packages and libraries that were used, such as \`dplyr\`, \`factoextra\`, and \`ggplot2\`, allowing efficient data analysis and visualization.

Regarding modeling, Python was chosen due to its versatility and familiarity. Python is extremely useful for a wide range of applications, from web development to data analysis and machine learning. Powerful and widely used libraries in these fields were used, such as \`sklearn\` (DecisionTreeRegressor, mean_absolute_error, and mean_squared_error) and \`scipy.stats\`. For the development and execution of models in Python, Google Colab was used as it provides free computational resources, including GPUs, which is especially beneficial given the size of the database.

Finally, to turn the work into a functional product, Streamlit was explored and started to be used. This is an open-source Python library that allows creating interactive and custom web applications for data science and machine learning projects. The main advantage of Streamlit is its ease of use and focus on simplicity. It allows data scientists and developers to quickly and easily create interactive user interfaces using just a few lines of Python code. With Streamlit, visualizations, user controls, and machine learning models could be integrated into a data web intuitively.

The choice of these tools and languages was aimed at maximizing efficiency and effectiveness in each phase of the project, thus ensuring the quality and functionality of the final product.

3.3 DATA.

Regarding the data used, it was provided by the director of the master's program in Barcelona, Rui Figueiredo. The data consists of three datasets: one from 2016, another from 2017, and the last one from 2018. All of them resulted from a survey conducted among the students and professionals under the charge of Rui, and each dataset contains 936 variables.

Since the datasets have the same variables and there is no continuity between them, meaning there are no repeated patients, it was decided to consider all of them as cross-sectional data and concatenate them all by rows. This resulted in a larger dataset with 2106 observations, over 930 variables, and no time distinction, as it was not useful.

Additionally, as the number of variables is too high, the vast majority are binary, and a thematic distinction can be identified among them, they have been divided into 4 blocks considering what is desired to be done with them and what the main objective requires. These blocks are: systemic patient pathologies, which contain characteristics such as gender, date of birth, number of cigarettes per day if they smoke, if they consume alcohol or not, cholesterol levels, among others; intervention details, which contain variables such as its duration, tools used and their characteristics, type of intervention, and much more; pre-intervention medication, which are the medications that patients took before undergoing the intervention; and finally, post-intervention medication.

In terms of data protection, initial access will be restricted exclusively to clinical researchers. Personal data used in the study must be safeguarded in accordance with the provisions of the General Data Protection Regulation (EU) 2016/679 and Organic Law 3/2018 on Personal Data Protection and guarantee of digital rights, which adjusts Spanish legislation to the General Data Protection Regulation of the European Union. The patient's name should not be included in any study document, only a patient code will be assigned at the beginning of it. Information disclosure will be made exclusively in aggregated form, and under no circumstances will the identity of the participants be revealed in any publication or congress presentation derived from the results of this study.

3.4 DATA PREPARATION.

Initially, since the first subtask focuses on analyzing patients, the first 111 columns of the database were selected, which contain patient characteristics, pathologies, and medications.

Next, in order to generate bar charts representing the distribution of the variables of interest in the database, these variables were transformed into categorical ones, grouping the characteristics according to the following criteria.

First, the binary variables for each type of pathology were reduced to a column called "Patologia_1stemica" indicating the type of pathology the patient has. Similarly, the various medications were grouped into the variable "Medicacion_Actual2". Next, to gather the patients' ages, ranges of 10 years each were created in the "Edad" variable from the birth date; and smokers were grouped into the variable "nivel_de_fumador".

Finally, variables that are not of interest for this analysis were removed ("Fecha_Nacimiento", "Date_Create", "IP_Adress", "Operador", "Auxiliar", "Jefe_de_dia", "Fecha_intervencion", "No.fumador.a", "Exfumador.a", "Fumador.a", "Otras_Drogas", "Medicacion_actual", "Otro..especifique.", "edad", "Tipo_De_cirugia", "Otro..especifique._17"); and the variables "Patologia_1stemica" and "Medicacion_Actual2" were modified by grouping values that appear less than 5 times into the category "otros". The resulting variables are explained below:

- Gender: Groups different gender categories into one column.
- Systemic Pathology: Combines all systemic pathologies into a single category.
- Alcohol: Groups different types of alcohol consumption into one column.
- Current Medication: Combines all current medications into a single category.
- Type of Surgical Intervention: Summarizes different types of surgical interventions into one category.
- Age Category: Classifies ages into significant groups in 10-year ranges (between 10 and 20 years, between 20 and 30 years, and more).
- Smoker Level: Groups different levels of tobacco consumption into one column.

After the categorical exploratory analysis of the general database, a numerical analysis will be conducted by separating the database according to the type of operation (Oral Implantology and Peri-implant Surgery). In this database, a new category called “Género” has been incorporated, where the value 1 will be assigned for "Female" and the value 0 for "Male". Additionally, variables that are not of interest in the analysis are again removed. With the resulting database, the most significant differences between the variables of the two types of operations have been analyzed.

Finally, to perform an explanatory PCA, the previous numerical database is used, from which the variables "Systemic Pathology" and "Current Medication" have been removed since this information is already represented through the various recorded pathologies and medications. Additionally, medicines or diseases that appear in the database less than three times have been removed, which has contributed to improving the obtained results. In this way, we obtain a binary database for analysis with gender, smoker level or other drugs, and current diseases and medications.

For the second subtask, which is to predict the duration of the intervention, the variables selected with PCA as relevant are taken into account. Prior to the statistical analysis, the 100 most influential variables are selected from the entire database through another dimensionality reduction technique similar to PCA, a Multiple Correspondence Analysis (MCA). This would not only help to reduce the options to the essentials but also help to confirm if the variables obtained with PCA make sense. The result of this can be summarized by noting that many of the variables do match those of the PCA, such as current medication, alcohol consumption, type of surgical intervention, among others.

After this step, cleanups are applied, such as the removal of interventions with unspecified or incorrect durations. Then, the duration ranges are converted into numerical values by calculating the average of each range. For example, the range of 20 to 40 minutes is assigned the value of 30 minutes. The distribution sorted by frequency is shown as in *Figure 2*.

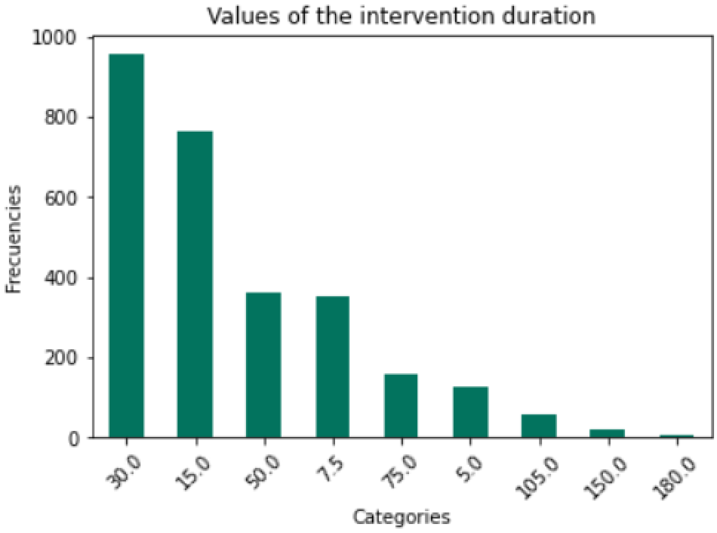


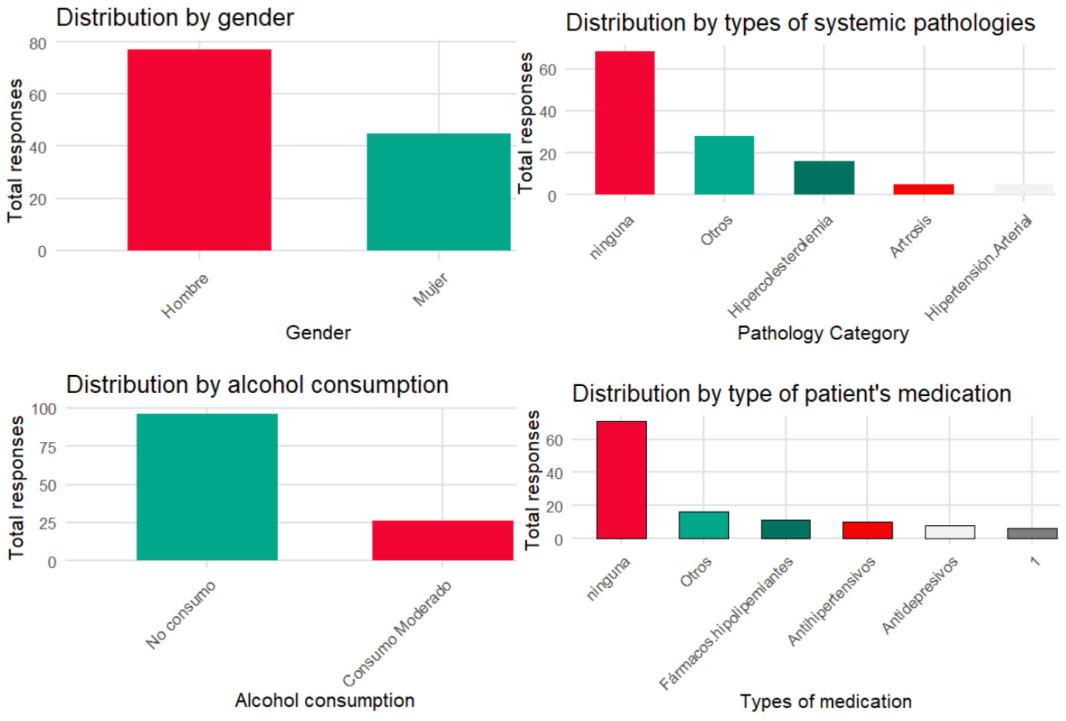
Figure 2. Bar plot of the distribution of the intervention duration sorted by frequency.

Next, Autoclean is applied, which is a function that automatically cleans the data and performs a process similar to one-hot encoding. Subsequently, a significant imbalance in the data was identified, especially in the most influential variable (Type of Surgical Intervention). To address this problem, various data rebalancing techniques were experimented with, including undersampling and oversampling, as well as the implementation of deep learning models and adjustments to hyperparameters. The strategy that had the most impact was the application of the Synthetic Minority Over-sampling Technique (SMOTE).

Therefore, it was decided to apply SMOTE (Synthetic Minority Over-sampling Technique) to the indicated variable, with the purpose of generating synthetic examples for the minority classes and balancing the distribution of this variable. Balancing this variable instead of the duration of the intervention (the study variable) provided better prediction results. SMOTE was specifically chosen because, unlike simply duplicating existing data, it creates new samples that represent linear combinations of the nearest neighbors of the minority classes, helping to maintain the diversity of the dataset. This is crucial to avoid overfitting and improve the model's ability to generalize. Moreover, SMOTE directly addresses the problem of imbalanced classes, allowing the model to perform better in predicting these classes. The improvement in the model's performance after applying SMOTE is noticeable, providing greater accuracy and stability in future model predictions.

3.5 Statistical analysis.

In the first instance, the results of the exploratory analysis will be presented in *Figure 3* to examine the data contained in the categorical database. There is a notable disparity in the amount of data between oral implantology surgery and other types of operations. Additionally, there is a predominance of males in the records, with the most frequent ages ranging between 60 and 80 years. Furthermore, there is a lower proportion of non-smokers compared to smokers. And the most common diseases identified are hypercholesterolemia, osteoarthritis, and hypertension.



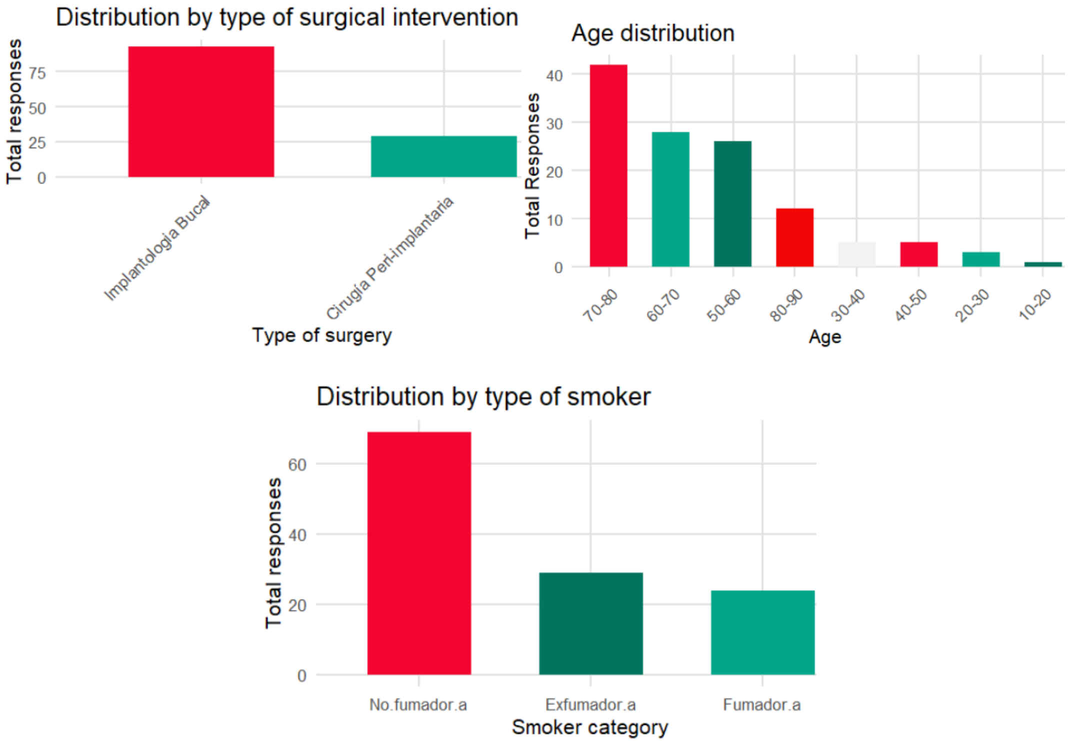


Figure 3. Bar plots of the distribution of relevant variables.

In addressing the first subtask, we generated the numerical database of patients described previously. This database includes two types of dentoalveolar procedures: oral implantology, which involves the placement of dental implants, and peri-implant surgery, which aims to address possible complications of peri-implantitis that arise after implantology.

Since the patients undergoing both types of operations (oral implantology and peri-implant surgery) are not related, it is impossible to use a model to predict whether an implantology procedure will result in peri-implantitis. However, it is feasible to analyze and compare the differences between patients who have experienced peri-implantitis and those who have received dental implants, even considering that the latter may also face peri-implantitis issues in the future. This analysis will allow us to identify whether there is significant variability in any specific variable between the two groups, which could indicate a relevant differentiating factor.

Initially, we chose to use a chi-square test to evaluate the differences between the same variable in each database. However, this method generally requires that the expected frequencies in each cell of the contingency table be at least 5. In our case, most variables did not meet this criterion, which could compromise the effectiveness of the test. Therefore, we decided to perform a difference of means between the variables of each database. This approach facilitates a direct and clearer comparison between the two groups, specifically focusing on the metric of interest (OpenAI, 2024). Additionally, this method is more intuitive and directly related to the research question, offering a more straightforward interpretation of the results.

Thus, after selecting the 10 variables with the most significant difference between their means among peri-implantitis patients and the entire population, we find that the characteristics that most differentiate the two types of patients are: "Fármacos.hipolipemiantes", "Genero", "Fumador.a", "Patologia_1stemica", "Hipercolesterolemia", "Hipertensión.Arterial", "Antihipertensivos", "Medicacion_Actual2", "Artrosis", and "Insulina". Later, we will analyze the meaning and relevance of these results in comparison with previous studies on the subject.

The *Figure 4* shows the difference in proportions of the most significant variables for the study.

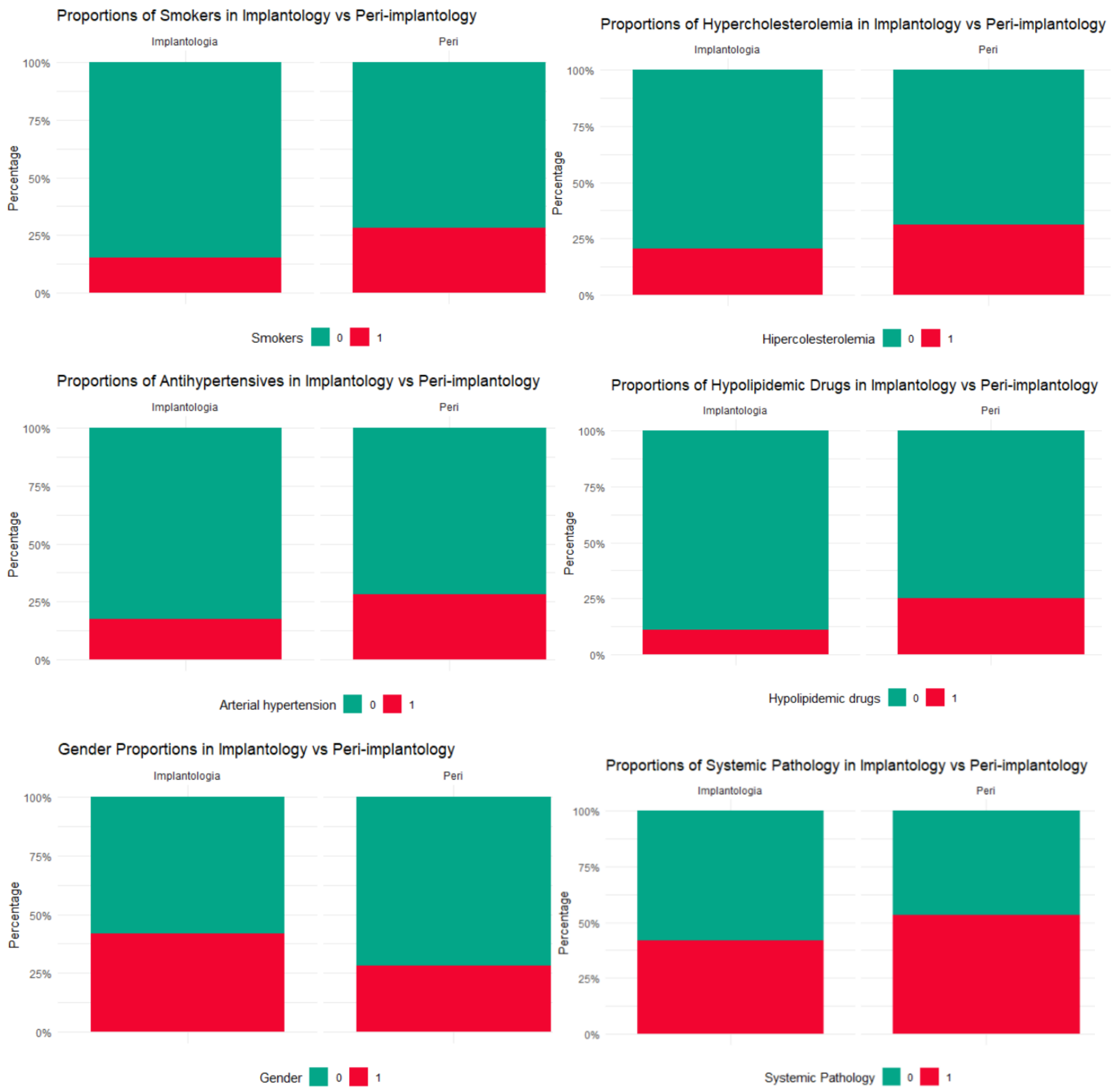


Figure 4. Bar plots showing the differences between the means of implantology patients and peri-implantology patients.

Another approach we considered to determine if there are significant variables that differentiate patients is the implementation of an explanatory principal component analysis (PCA) after merging the two databases. This method allows us not only to identify hidden patterns in the data but also to observe if any resulting dimension effectively separates the patient groups. However, it is important to recognize the inherent limitations of this approach. As mentioned earlier, some patients currently with implants could develop peri-implantitis in the future. This eventuality prevents a clear and definitive separation between the groups through PCA.

After performing the principal component analysis (PCA), we found that the first two dimensions explain up to 44.5% of the total variability in the data. By visualizing these results with a scatter plot, where the points represent the patients and are colored according to the type of procedure, we observe certain trends. Although patients with oral implantology are not clearly distinguished, which could be due to the presence of future peri-implantitis cases, the majority of peri-implantitis patients tend to separate along the first dimension as shown in *Figure 5*.

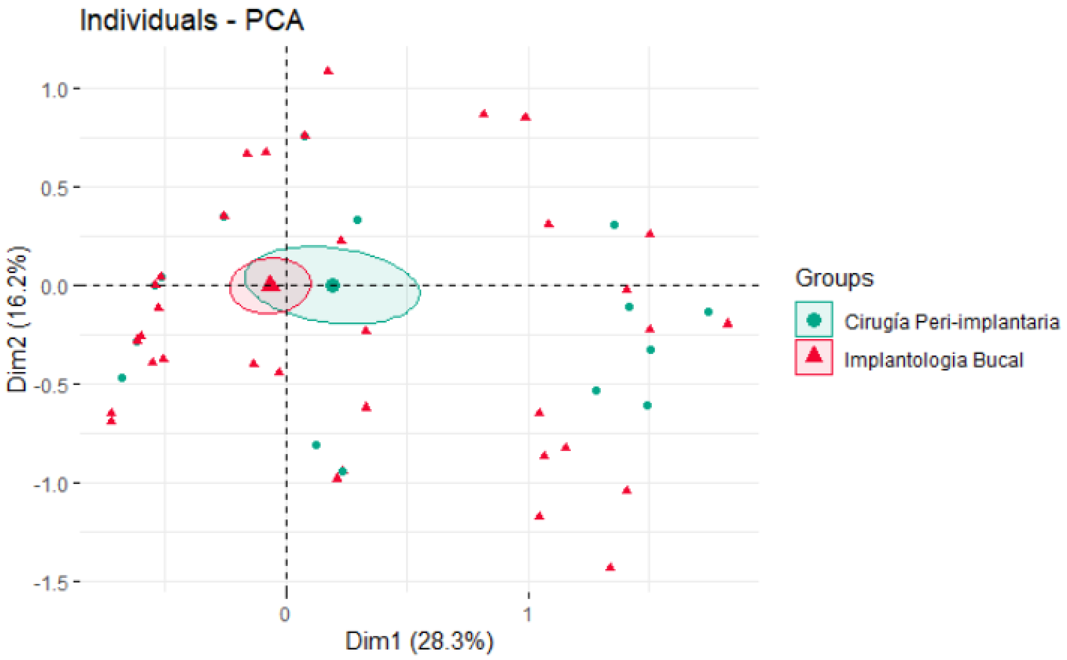


Figure 5. Contribution of the individuals to the first and second dimensions of the PCA according to the type of intervention.

Considering the limitations of our dataset, we found that the variables that most influence this first dimension coincide with those we previously identified as the most differentiating between peri-implantitis patients and oral implantology patients (see *Figure 6*). This finding reinforces the relevance of these variables in the characterization of the study groups.

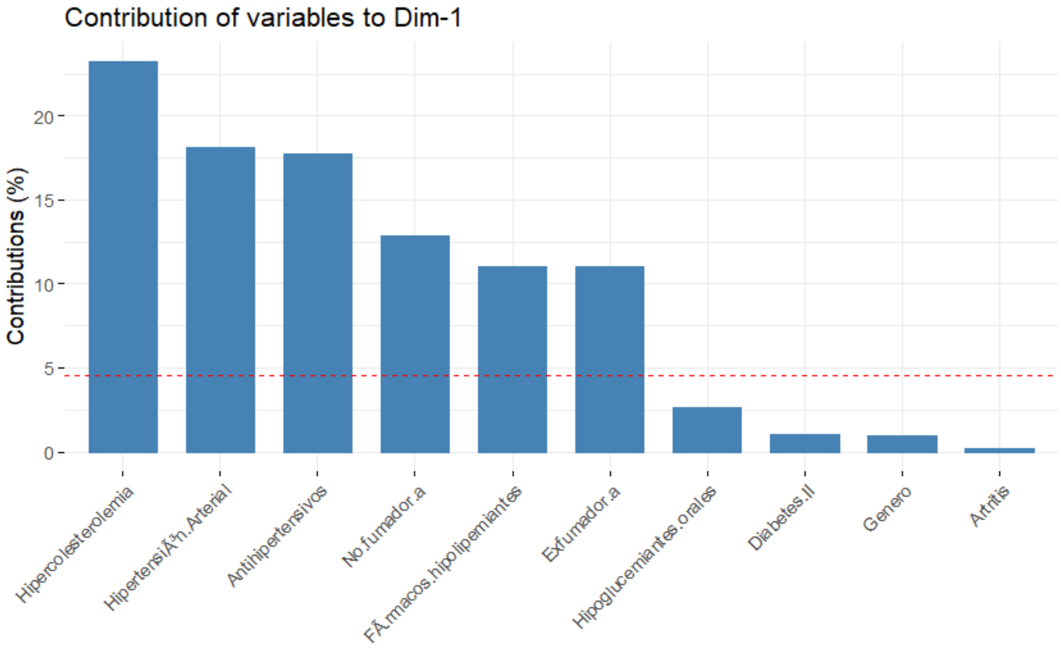


Figure 6. Bar plot of the contribution of the variables to the first dimension of the PCA.

Thus, it is observed that features such as Hypercholesterolemia, Antihypertensives, Arterial Hypertension, Smoker, Hypolipidemic Drugs, Insulin, and Type 2 Diabetes, which coincide in both analyses, stand out as the most relevant for distinguishing between patients with peri-implantitis.

To predict the intervention time, the Random Forest method is used to further reduce the predictor variables to those truly determinant in predicting the duration of the intervention. The 10 resulting variables from the Random Forest are shown in *Figure 7* and are chosen to analyze and predict durations with the help of the Regression Tree model, which returned the smallest error, as will be explained later.

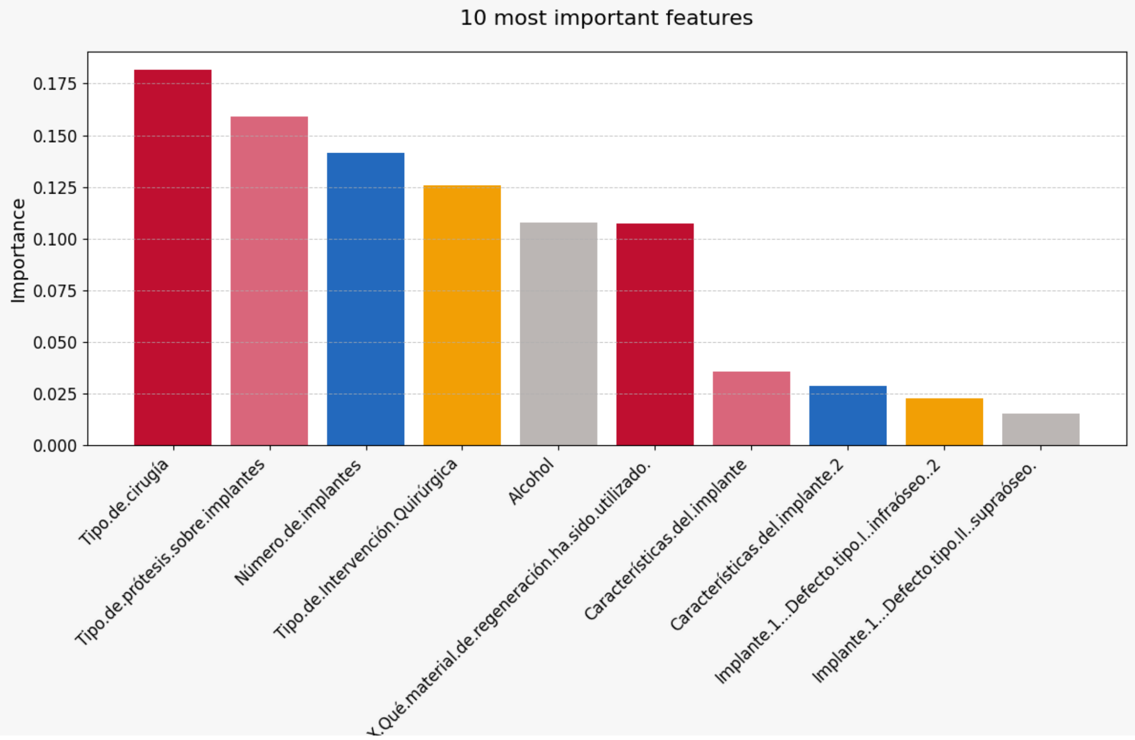


Figure 7. Bar plot of the 10 most important features obtained with Random Forest.

In the continuation of the study on estimating surgical intervention time, a third subtask was undertaken: developing a model to classify the difficulty of surgical interventions. This model is based on the selection of critical variables that directly influence the surgical process, provided by the results of the study, and approved by Dr. Rui Figueiredo.

The selected variables include patient pathologies, types of operations, patient gender, and the duration of the intervention, the latter being previously calculated using the developed intervention time prediction model. To classify the operations, a weighted scoring method was implemented. Following the criteria of the American Society of Anesthesiologists (ASA) for pathological variables and assigning a higher score to longer interventions, a points system was established to reflect the complexity and risk associated with each intervention (Doyle, Hendrix, & Garmon, 2023).

With these scores, the quartiles of the data distribution were calculated, assigning a difficulty category to each range: Very Easy, Easy, Moderate, and Difficult. This quantitative classification allowed the interventions to be structured into four clear levels of difficulty, thus facilitating resource management and medical staff preparation.

To validate the effectiveness of the classification model, a Random Forest (RF) model was fitted using these categories as target variables. The result, which will be discussed later, was an almost perfect fit, indicating that the model can accurately replicate and predict the difficulty classification based on the established parameters. However, it is important to acknowledge that this high degree of fit may be indicative of overfitting, as the model was specifically designed for the conditions of the current dataset. This specificity may limit the model's generalization to other clinical contexts.

This classification approach not only provides a practical tool for preoperative evaluation but also contributes to the academic field with an innovative method for managing and anticipating needs in the surgical environment.

4 RESULTS.

After concluding the analyses mentioned for subtask 1, we identified that the most significant variables in differentiating patients with peri-implantitis are hypercholesterolemia, antihypertensives, arterial hypertension, lipid-lowering drugs, insulin, type 2 diabetes, and smoking status. The analysis using the biplot of the variables contributing to the first dimension shown in *Figure 8* corroborates the influence of these variables in the separation of the patients, and also suggests that not smoking can help differentiate them as well.

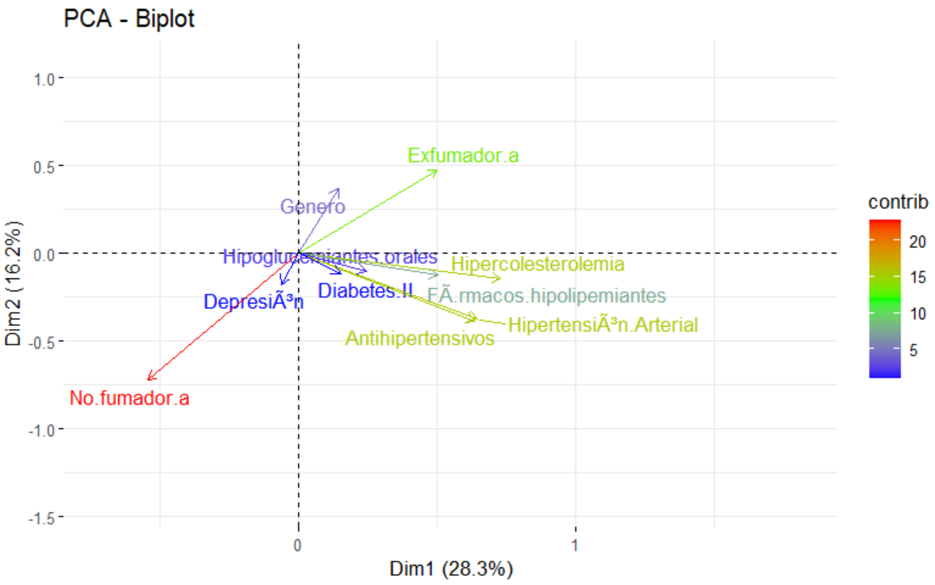


Figure 8. Contribution of the main variables to the first and second dimensions of the PCA.

Additionally, intrinsic relationships are observed within our data, derived from their characteristics. To understand the relevance of the variables identified as important, it is crucial to consider certain aspects. There is a notable coexistence of diseases such as arterial hypertension, hypercholesterolemia, and diabetes, along with the associated medications for their treatment (antihypertensives, lipid-lowering drugs, and insulin, respectively). These three diseases have a close medical relationship. Therefore, it can be inferred that patients affected by arterial hypertension, hypercholesterolemia, or diabetes, as well as smokers, represent a higher-risk group in surgical interventions and may require closer monitoring (see Annex 1 for more information).

The second task, predicting the duration of the interventions, was addressed using a Regression Tree model enhanced with the Synthetic Minority Over-sampling Technique (SMOTE) to balance the data. We present the evaluation metrics for two iterations of the model, emphasizing the importance of considering both RMSE and MAE.

First Iteration of the Regression Tree before SMOTE:

- Mean Squared Error (MSE): 407.52
- Root Mean Squared Error (RMSE): 20.19
- Mean Absolute Error (MAE): 14.74

Second Iteration of the Regression Tree after SMOTE:

- Mean Squared Error (MSE): 374.26
- Root Mean Squared Error (RMSE): 19.35
- Mean Absolute Error (MAE): 11.84

MSE and RMSE are common metrics for evaluating regression models; however, both are especially sensitive to outliers because they square the differences between predicted and actual values. In contexts where the data distribution includes extreme values or the range of the target variable is very wide, RMSE can give a misleading impression of poor model performance as it magnifies the impact of large errors.

A graph is attached in *Figure 9* to visualize the results of the predictions with the Regression Tree model (see Annex 2 for more information on model preparation and training).

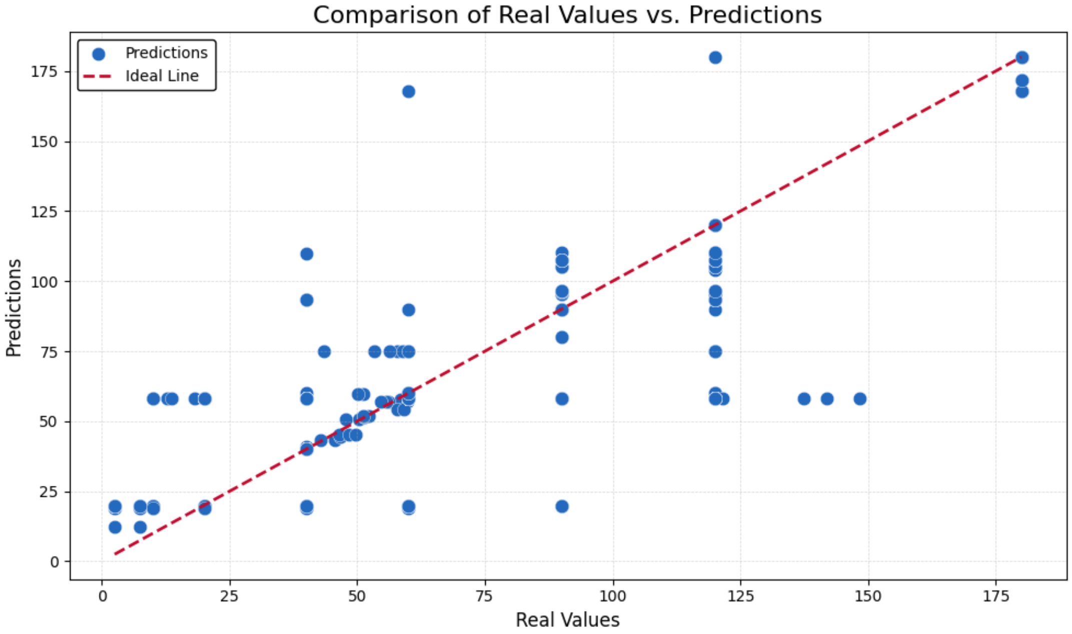


Figure 9. Scatterplot of real values versus predicted values with the Regression Tree.

On the other hand, MAE provides a more intuitive measure and is less sensitive to outliers, as it is based on the arithmetic mean of absolute errors. This makes it more representative of the typical errors in prediction, providing a better insight into the typical performance of the model in everyday situations. For this reason, although RMSE can be useful for highlighting issues in the presence of outliers, MAE offers a more robust and balanced perspective for evaluating the accuracy of predictions in our study.

The third task, classifying the difficulty of surgical interventions, employed a Random Forest model to classify the difficulty of surgical interventions with the following performance metrics:

Model Accuracy: 97.15%

Classification Metrics:

- Difficult: precision 0.88, recall 0.99, f1-score 0.93.
- Easy: precision 0.99, recall 0.96, f1-score 0.97.
- Moderate: precision 0.98, recall 0.95, f1-score 0.96.
- Very Easy: precision 0.98, recall 0.99, f1-score 0.99.

The overall accuracy and f1-scores of the Random Forest model highlight its ability to accurately classify surgical interventions into different levels of difficulty, particularly excelling in identifying challenging cases. However, as it was mentioned before, this model is specifically optimized for our dataset, limiting its applicability to other contexts without additional adjustments. Despite this limitation, the results confirm that the model is effective and valuable for planning and medical training within our scope of study.

5 DISCUSSION.

The three main diseases share a close medical relationship. Therefore, it can be inferred that patients affected by arterial hypertension, hypercholesterolemia, or diabetes, as well as smokers, represent a higher-risk group in surgical interventions and may require closer monitoring.

Numerous articles and scientific studies have investigated peri-implantitis and its causes to prevent it, highlighting the increasing importance of this disease, which is becoming more common and less understood (DENTAID; Sociedades Científicas; Colegios de Higienistas, 2020). These studies provide interesting and varied information, but they agree that both smoking and diabetes are significant risk factors for peri-implantitis. Smoking alters various aspects of the host's innate and adaptive immunity (Martínez Gómez, Hernández Andara, Quevedo Piña, Ortega Pertuz, & Chong, 2023), as nicotine is known to impair protein synthesis and affect the adhesion capacity of gingival fibroblasts (VERICAT - Instituto de Informacion, 2023). Regarding diabetes, elevated blood glucose levels lead to the production of advanced glycation end-products (Martínez Gómez, Hernández Andara, Quevedo Piña, Ortega Pertuz, & Chong, 2023), resulting in a deficient immune response and healing processes in patients, making them at-risk during surgical treatment (López Muñoz, 2020).

Moreover, additional information related to our results regarding hypercholesterolemia and hypertension can be observed. A study evaluated the influence of cholesterol on peri-implant Marginal Bone Level (MBL) by measuring lipid levels in patients' blood at the beginning and during follow-up. This study concluded that total cholesterol is one of the variables that most influences MBL (De Angelis, et al., 2023). However, the study had limitations due to the lack of a sufficiently large sample, a limitation similar to the results of the present study.

Finally, an article observed a trend towards a higher prevalence of peri-implantitis in the group with cardiovascular disease (Wang, et al., 2021). However, the study is inconclusive regarding the relationship with arterial hypertension due to confounding factors, suggesting that the true relationship could be with diabetes, which in turn contributes to hypertension. To confirm this theory, additional studies verifying the hypothesis that arterial hypertension directly influences peri-implantitis and is not just a confounding factor would be necessary. Again, a larger dataset would be required.

Furthermore, it has been demonstrated how predicting the duration of dental surgical interventions and classifying operations by their difficulty level is of great value for clinical management. Using advanced statistical techniques such as Random Forest and Multiple

Correspondence Analysis (MCA), the most influential variables affecting intervention times were identified. Among these, factors such as tooth position and specific procedural characteristics played a crucial role. These elements are fundamental since, for example, accessibility and complexity associated with different tooth positions can significantly vary, influencing the time required to complete the intervention.

The precision of the model, with a mean absolute error of 14 minutes reduced to 11 minutes, was considered satisfactory by the collaborating professional, Dr. Rui Figueiredo. The relevance of this level of precision lies in its practical application; it does not seek minute-by-minute accuracy but rather categorizes interventions into approximate durations (short, medium, long) for effective operational planning and resource allocation.

Additionally, implementing findings from the analysis of complications in operations into a Streamlit application provides significant added value. This digital tool allows clinical operators to input specific patient and procedure data, receiving as output the estimated intervention duration and the classification of operative difficulty they will face. This facilitates daily planning, adjusts room occupancy time expectations, and improves efficiency in assigning students for different procedure types based on their level of experience.

Currently, the university clinic schedules all patients with a standard allocation of one hour per procedure, a practice that the model significantly optimizes. By more accurately predicting the duration of each intervention, schedules can be adjusted more effectively, avoiding inefficiencies and downtime, and allowing for a more precise allocation of resources.

This approach not only improves the logistics and operational efficiency of the clinic but also ensures that dental students can gain experience in interventions that align with their level of skills and knowledge. Additionally, by maximizing the use of available resources, the clinic can potentially increase the number of patients treated and, consequently, expand educational opportunities for more students.

6 LEGACY.

Regarding the conclusions drawn from the information on patients susceptible to peri-implantitis, the drafting of a scientific article for publication in an academic journal is proposed. To achieve this, suitable journals have been identified in the Scimago Journal & Country Rank, such as the International Journal of Oral Science, which publishes research on all aspects of oral science and related interdisciplinary fields (Sichuan University Press, n.d.). Additionally, the developed code could be reused with larger data samples in the future to obtain more conclusive results.

Furthermore, the website developed with Streamlit leverages the findings of a study to enhance dental interventions. It offers functionalities to predict the duration of surgeries based on specific data, aiding in more accurate time estimation. It also analyzes the complexity of operations using patients' pathological information, classifying the difficulty of interventions for better planning and preparation. The platform provides details on the study's findings, allowing users to examine how dental surgeries relate to various systemic pathologies. This tool not only improves the interpretation and practical application of data but also promotes progress in the fields of dentistry and data science. The website can be accessed through this link.

Lastly, this project has a direct impact on SDG 3 (Good Health and Well-being) by focusing on improving outcomes in dentoalveolar surgery and strengthening clinical management in the dental sector. By identifying and addressing risk factors in this type of surgery, postoperative complications can be minimized, ensuring safer and more effective dental treatment for patients. Additionally, by optimizing clinical management, access to dental services is improved, wait times are reduced, and efficiency in care is increased. This comprehensive approach not only promotes oral health but also contributes to overall well-being by providing quality and accessible medical care to individuals.

7 CONCLUSION.

This study, conducted in collaboration with the Master's program in Periodontology and Peri-Implantology at the University of Barcelona, utilized advanced data analysis to optimize clinical management and training in dentoalveolar surgery. Two main goals were established: identifying patient profiles at high risk of postoperative complications and developing predictive models to estimate the duration and difficulty of surgical interventions.

The analysis of high-risk patients allowed for anticipating specific needs for specialized management and adjusting intervention protocols to minimize risks for patients with arterial hypertension, type 2 diabetes, hypercholesterolemia, or who are smokers. Concurrently, predictive models were implemented in a Streamlit platform, improving resource allocation and clinical activity scheduling with a reduced mean absolute error precision from 14 to 11 minutes.

These advancements could significantly enhance operational efficiency and the educational experience of students within the specific clinic, although the model is highly tailored to this particular facility and is not designed for global application without considerable adaptations.

It is important to note that with a larger and more comprehensive dataset, more robust conclusions could be drawn in the future and useful applications could be developed for a wide variety of dental clinics.

8 ACKNOWLEDGEMENTS.

We would like to extend our heartfelt gratitude to the Master's program in Periodontics and Peri-implantology at the University of Barcelona for their invaluable support. Special thanks go to Dr. Rui Figueiredo, the director of the program, for generously providing us with the dental data from his students' practices. His contribution has been instrumental in advancing our research and understanding in the field.

9 REFERENCES.

De Angelis, P., Rella, E., Manicone, P. F., Gasparini, G., Giovannini, V., Liguori, M. G., . . . D'Addona, A. (2023, Abril 26). *NIH - National Library of Medicine. National Center for Biotechnology Information.* Retrieved from NIH - National Library of Medicine. National Center for Biotechnology Information.: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10241576/

DENTAID; Sociedades Científicas; Colegios de Higienistas. (2020, Enero 9). *DENTAID Profesional - Plataforma para profesionales de la Salud Bucal*. Retrieved from DENTAID Profesional - Plataforma para profesionales de la Salud Bucal: https://www.dentaid.es/pro/dentaidExpertise/1901/las-enfermedades-periimplantarias-un-problema-creciente-en-espana

Doyle, D. J., Hendrix, J. M., & Garmon, E. H. (2023, Agosto 17). *NIH - National Library of Medicine. National Center for Biotechnology Information*. Retrieved from NIH - National Library of Medicine. National Center for Biotechnology Information: https://www.ncbi.nlm.nih.gov/books/NBK441940/

López Muñoz, E. M. (2020). *Enfermedad Periimplantaria en Pacientes Periodontales.* Sevilla: Universidad de Sevilla. Retrieved from https://idus.us.es/bitstream/handle/11441/105381/1/Enfermedad%20periimplantaria%20en%20pacientes%20periodontales.pdf?sequence=

Martínez Gómez, J. C., Hernández Andara, A., Quevedo Piña, M., Ortega Pertuz, A. I., & Chong, M. L. (2023, Diciembre 26). *NIH - National Library of Medicine. National Center for Biotechnology Information*. Retrieved from NIH - National Library of Medicine. National Center for Biotechnology Information: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10880694/#B30

OpenAI. (2024). *ChatGPT (versión GPT-4)*. Retrieved from ChatGPT (versión GPT-4): https://www.openai.com/chatgpt

Sichuan University Press. (n.d.). *SJR - Scimago Journal & Country Rank*. Retrieved from SJR - Scimago Journal & Country Rank: https://www.scimagojr.com/journalsearch.php?q=19700180533&tip=sid&clean=0

Sociedad Española de Cirugía Oral y Maxilofacial y de Cabeza y Cuello. (2023, Diciembre 8). *SECOM CyC - Sociedad Española de Cirugía Oral y Maxilofacial y de Cabeza y Cuello*. Retrieved from SECOM CyC - Sociedad Española de Cirugía Oral y Maxilofacial y de Cabeza y Cuello: https://www.secomcyc.org/area-paciente/que-podemos-hacer-por-ti/cirugia-dentoalveolar/

VERICAT - Instituto de Informacion. (2023, Febrero 10). *VERICAT - Instituto de Informacion*. Retrieved from VERICAT - Instituto de Informacion: https://vericatinstitutodeformacion.com/blog/periimplantitis/

Wang, I.-C., Ou, A., Johnston, J., Giannobile, W. V., Yang, B., Fenno, C., & Wang, H.-L. (2021). Association between peri-implantitis and cardiovascular diseases: A case-control study. *Journal of Periodontology*, 1-11.

10 ANNEXES.

10.1 ANNEX 1.

It is the document named “Annex 1” and it contains everything related to the first subtask, which involved extracting relevant variables using PCA, analyzing variability, and calculating means.

10.2 ANNEX 2.

It is the document titled “Annex 2”, which contains the code for implementing the SMOTE method, Random Forest, and Regression Tree.