Diabetes Predictive Model Using the CRISP-DM Process for Disease Prevention Implementing Machine Learning

Modelo Predictivo de Diabetes Utilizando el Proceso CRISP-DM para la Prevención de la Enfermedad Implementando Aprendizaje Automatizado

DOI: https://doi.org/10.17981/cesta.04.02.2023.02

Scientific research article. Date of receipt: 14/11/2023, Date of acceptance: 14/12/2023.

Juan Montes-Bustamante
Universidad de Sucre. Sincelejo (Colombia)
juandavidmontesbustamante@gmail.com

How to cite:
J. Montes-Bustamante, "Diabetes Predictive Model using the CRISP-DM Process for Disease Prevention Implementing Machine Learning", *J. Comput. Electron. Sci.: Theory Appl.*, vol. 4, no. 2, pp. 17-37, 2023. https://doi.org/10.17981/cesta.04.02.2023.2

Abstract

Introduction— In the information age, data mining unveils valuable patterns and insights from large datasets, empowering informed decision-making.

Objective— This work aims to develop an advanced model for accurately predicting diabetes in its early stages.

Methodology— The work employes the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology as the primary approach for data analysis.

Results— The decision tree model stands out for its superior predictive capability compared to the KNN algorithm, effectively discriminating between diabetes cases and the absence of pathologies, achieving an 83.01% overall accuracy in the test partition and surpassing the KNN model in predictive capacity.

Conclusions— Data mining, especially decision tree models, emerges as a valuable alternative for early detection of diabetes in our dataset.

Keywords— Diabetes; Predictive model; CRISP-DM; Early detection; Data mining.

Resumen

Introducción— En la era de la información, la minería de datos revela patrones y conocimientos valiosos desde grandes conjuntos de datos, capacitando la toma de decisiones informada.

Objetivo— Este trabajo aspira a desarrollar un modelo avanzado que prevea con precisión la diabetes en sus primeras etapas.

Metodología— Este trabajo emplea la metodología CRISP-DM (Cross-Industry Standard Process for Data Mining) como enfoque principal para analizar datos.

Resultados— El modelo de árbol de decisión destaca por su capacidad predictiva superior al algoritmo KNN, logrando discriminar eficazmente entre casos de diabetes y la ausencia de patologías, con una precisión global del 83.01% en la partición de prueba, superando al modelo KNN en capacidad predictiva.

Conclusiones— La minería de datos, especialmente los modelos de árboles de decisión, emerge como una alternativa valiosa para la detección temprana de la diabetes en nuestro conjunto de datos.

Palabras clave— Diabetes; Modelo predictivo; CRISP-DM; Detección temprana; Minería de datos.

I. INTRODUCTION

Diabetes is a chronic disease that affects how the body converts food into energy. When someone has diabetes, their body either does not produce enough insulin or cannot use it effectively. Insulin is a hormone that acts like a key, allowing blood sugar to enter the body's cells for use as energy. As a result, excess sugar remains in the bloodstream, which can lead to serious health problems over time, including heart disease, vision loss, and kidney disease. Currently, there are three known types of diabetes: type 1 diabetes, type 2 diabetes, and gestational diabetes [1].

Diabetes Mellitus (DM) significantly impacts not only the health of those affected but also the healthcare system, social services, and society at large. This presents clinical, economic, and social challenges, highlighting the importance of assessing the disease's economic impact. Addressing this issue involves considering a range of methodological elements throughout the formulation, design, and execution of relevant studies.

The widespread impact of Diabetes Mellitus (DM), affecting both individual health and healthcare and social infrastructure, underscores the need to assess its economic impact thoroughly. However, this analysis must confront specific methodological challenges, from the conception of the study to its implementation. Careful consideration of these methodological elements is crucial to gain a complete and accurate understanding of the economic implications of DM. These implications encompass not only the direct economic impact of type 2 diabetes but also its broader effects on the healthcare system [2].

Type 2 diabetes is a chronic condition characterized by elevated blood glucose or sugar levels. Traditionally seen in individuals over 45 years of age, it is increasingly being diagnosed in children, adolescents, and young adults. In this form of diabetes, the body either does not produce enough insulin or does not use it efficiently. Insulin is vital for enabling glucose to enter cells and supply energy. However, in type 2 diabetes, insufficient insulin production or insulin resistance in cells leads to high blood sugar levels, intensifying the complications associated with the disease. The complexity of type 2 diabetes extends beyond metabolic regulation to affect various body systems. Furthermore, the emergence of this disease in younger populations raises additional concerns regarding their long-term health and quality of life, considering the inherent risks of diabetes [3].

Gestational diabetes is a condition that manifests in pregnant women and is characterized by carbohydrate intolerance, leading to varying levels of hyperglycemia. It typically develops and is identified during pregnancy. This disorder is linked to increased complications for the mother during pregnancy. It can have long-lasting effects on the fetus's health, extending from the neonatal period to adulthood. These complications underscore the importance of monitoring and managing gestational diabetes effectively [4].

Between 2000 and 2019, the Pan American Health Organization (PAHO) and the World Health Organization (WHO) conducted an extensive study in the Americas to construct a detailed profile of the disease burden with a specific focus on diabetes. The findings of this study were striking, underlining diabetes as a major catalyst for various significant health issues. This comprehensive research emphasizes the critical need for enhanced strategies and interventions to manage and mitigate the impact of diabetes on public health.

Diabetes has been identified as a critical factor contributing to several severe health complications, including vision loss, kidney failure, heart attacks, and lower limb amputation accidents. These findings highlight the substantial negative impact of diabetes on health, affecting not only the quality of life of those diagnosed but also placing a considerable strain on health systems and social resources [5].

In 2019, diabetes was recognized as the sixth leading cause of death in the Americas, responsible for an estimated 244,084 deaths directly attributable to the disease. Furthermore, it ranked as the second leading cause of Disability Adjusted Life Years (DALYs), which underscores the extensive and limiting complications that individuals with diabetes face throughout their lives. The disease's serious health consequences extend to cardiovascular complications, retinopathy, nephropathy, and non-traumatic amputations, further emphasizing the need for effective management and prevention strategies to mitigate these outcomes [5].

The methodology employed by modern systems in detecting diabetic retinopathy is based on interpreting specific patterns in retinal images, linking visual characteristics to the presence of the condition. This advanced analytical capability allows healthcare professionals to take proactive measures, offering timely preventive treatments to halt the development or progression of the disease [6]. The increasing reliance on these systems is a testament to the continuous advancements in medical technology.

To address the challenge of detecting diabetic retinopathy, sophisticated techniques leveraging artificial intelligence have been developed. These systems utilize complex algorithms and detailed image analysis to scrutinize retina photographs. They are capable of accurately identifying signs indicative of damage caused by diabetes. Such advanced technology underscores the significant progress in medical diagnostics and the potential for early and precise detection of diabetic retinopathy, thereby enhancing patient care and outcomes.

The methodology of these systems for detecting diabetic retinopathy relies on interpreting specific patterns in retinal images, linking visual characteristics to the presence of the condition. This advanced analytical capability equips healthcare professionals with the means to proactively intervene, offering timely preventive treatments to prevent the onset or progression of the disease [7].

In response to this health challenge, the continuous advancements in medical technology have spurred the development of innovative techniques underpinned by artificial intelligence for the early and precise detection of diabetic retinopathy. These systems employ sophisticated algorithms and detailed image analysis to evaluate retinal photographs thoroughly. Their ability to accurately identify signs of diabetes-related damage is a testament to the potential of AI in enhancing diagnostic accuracy and efficiency, thereby improving patient outcomes in the management of diabetic retinopathy.

At the heart of these systems methodology is the capability to interpret specific patterns in retinal images, drawing correlations between visual characteristics and the presence of diabetic retinopathy. This advanced analytical ability provides healthcare professionals with a means for proactive intervention, enabling them to administer early preventive treatments to halt the development or progression of the disease.

This pioneering approach simplifies the diagnostic process and significantly improves clinical outcomes and patient care. By facilitating the early detection of diabetic retinopathy, these systems pave the way for more effective therapeutic interventions. This underscores the profound impact of artificial intelligence in healthcare, marking a vital stride towards personalized and preventive care. Such breakthroughs exemplify technology's critical role in continually enhancing the visual health and overall well-being of individuals affected by diabetic retinopathy [7].

II. RELATED WORKS

Innovative and strategic initiatives have been undertaken to facilitate the early detection of diabetes. These efforts, underpinned by extensive research, aim to develop methods and technologies for the early identification of signs and risk factors associated with diabetes. The core idea behind these initiatives is to enhance diagnostic capabilities, equipping healthcare professionals and individuals at risk with practical tools for proactive intervention during the early stages of the disease.

The development and validation of artificial intelligence (AI) systems for detecting and monitoring diabetic retinopathy exemplify how technology can play a pivotal role in identifying and managing diabetes-related complications. These advanced systems, leveraging the predictive power of AI, represent a significant advancement in ophthalmic care for diabetic patients. They promise to improve patient outcomes and highlight the critical importance of technological innovation in combating diabetes and addressing its long-term health implications Principio del formulario[7].

Among the notable developments in this field is creating a predictive model utilizing the SAP Predictive Analytics tool. This model primarily aims to predict the diagnosis of type 2 diabetes mellitus. The successful implementation of this model is intended not just to streamline the diagnostic process for this chronic disease but also to provide critical insights about effective patient-centric interventions for healthcare providers, both in the public and private sectors [8].

Similarly, there have been efforts to develop a model based on artificial intelligence to assist clinical decision-making in the early detection of diabetes. For this purpose, a cross-sectional study was conducted using a dataset comprising vital information such as age, signs, and symptoms from diabetic and healthy individuals. Prior to model construction, the data underwent preprocessing techniques. The model itself was developed using fuzzy cognitive maps designed to emulate human reasoning and decision-making processes [9].

Additionally, some studies assess the impact of Machine Learning (ML) models using medical attributes to predict type 2 diabetes mellitus in elderly patients. These studies have developed and tested 13 different ML methods, encompassing classical models, neural networks, and ensemble models. The evaluation of these models was based on key performance metrics, including accuracy, precision, sensitivity, specificity, F1-score, misclassification rate, and the area under the curve (AUC), using both training and test datasets. Analysis revealed that the LightGBM model consistently outperformed others across all seven performance measures [10]. This finding from the project focusing on LightGBM is particularly significant.

The project assessing the efficacy of ML models for predicting type 2 diabetes mellitus using medical attributes holds direct relevance to the implementation phase of the present work. As both studies share the common objective of enhancing diabetes prediction, the insights and successful strategies derived from this research can offer valuable guidance for implementing the current project. These findings, particularly regarding the LightGBM model, provide essential learnings that could be instrumental in advancing predictive efforts in diabetes care.

The array of efforts dedicated to the early detection of diabetes reflects a solid commitment to enhancing public health. It showcases the positive impact that research and technology can have in preventing and managing this chronic disease. These initiatives significantly advance health promotion, contributing to developing a comprehensive and progressive approach to diabetes care [11]. In this context, it is crucial to emphasize the significance of such an approach.

Significantly, these endeavors extend beyond merely identifying diabetes at an early stage. They delve into more profound aspects, including implementing preventive and personalized measures tailored to the identified risk profiles. Integrating advanced technologies like artificial intelligence, data analytics, and other emerging tools has been pivotal in refining screening strategies. This holistic approach has enhanced the precision of diagnostic processes and markedly boosted the effectiveness of preventive interventions, thereby shaping a more efficient and targeted approach to diabetes care.

III. METHODOLOGY

Various methods are available for conducting predictive and descriptive data mining. The primary distinction between these methods lies in their application scope and specific procedural steps. Various methods exist within the realm of predictive mining, each characterized by its unique features and specific applications.

- The process begins with the Sampling stage, where a representative subset of the data is selected for analysis.
- The Exploration stage then follows, thoroughly examining the data to understand its structure and underlying trends.
- During the Modification stage, the data is manipulated and cleaned as necessary to prepare it for modeling.
- The Modeling stage involves applying algorithms to construct predictive models based on the data.
- Finally, the process culminates in the Assessment stage, where the effectiveness and accuracy of the developed models are evaluated and verified.

Each stage plays a crucial role in the SEMMA methodology, ensuring the resulting data mining process is comprehensive, accurate, and effective [12]. The Assessment stage, in particular, is integral for validating the reliability and applicability of the predictive models derived from the data mining process.

Another methodology, such as KDD (Knowledge Discovery in Databases), encompasses the full spectrum of database knowledge discovery processes. It starts with the initial data selection, followed by preprocessing and transformation, then proceeds to the data mining stage and concludes with evaluating the results. KDD is a holistic framework that identifies, interprets, and applies valuable knowledge extracted from large datasets. Each step in this process is crucial to the success of the overall knowledge discovery [12].

As a complement, TDSP (Team Data Science Process) provides an agile and structured framework for managing data science projects. It involves five phases, starting with project planning, followed by data acquisition and preparation, modeling, deployment, and finally, continuous monitoring and maintenance. TDSP fosters collaboration and iterative improvements among team members, making it well-suited to data science projects' dynamic and evolving nature. Its flexibility and structured approach enable teams to adapt effectively to various project demands and challenges [12].

Each of these methodologies, KDD and TDSP, offers a unique and systematic approach to data mining and data science, facilitating effective management and execution of projects in these fields [12].

In this context, CRISP-DM (Cross-Industry et al. for Data Mining) has been identified as the most fitting framework to achieve the specific objectives of this project, which is centered around developing a predictive model for diabetes prevention. This choice is influenced by the compatibility of CRISP-DM with this type of project and previous successful experiences utilizing this methodology in similar initiatives. Furthermore, CRISP-DM aligns well with the nature of the problem at hand, facilitating the identification of critical data patterns and constructing a robust predictive model [13].

Before implementing the CRISP-DM methodology, it is imperative to consider certain applied concepts. These concepts lay the theoretical and practical groundwork for developing the predictive diabetes model, ensuring a comprehensive and accurate understanding of the data involved. The key concepts that form the basis of this model are as follows:

- **Data Mining** refers to analyzing large datasets to uncover trends and patterns. It involves sorting through vast amounts of data to identify meaningful connections and insights that might not be apparent at first glance [14].
- **Data Mining Techniques:** These techniques fall into three primary categories: descriptive, predictive, and prescriptive. Descriptive mining focuses on summarizing and describing data to identify patterns and trends. Predictive mining, on the other hand, is utilized to forecast future behaviors or outcomes based on historical data. Prescriptive mining recommends specific actions or strategies based on descriptive and predictive mining insights. Together, these approaches provide a comprehensive framework for data analysis, enabling more informed decision-making and strategic planning [14], [15].
- **Knowledge Modeling or Model:** This stage involves the development of knowledge models utilizing previously collected data. These models can take various forms, such as classification or regression models,

designed to estimate or infer the value of a specific variable. The choice of model type depends on the problem being addressed and the kind of predictions or inferences that need to be made [16].

- **Dataset** refers to data collection used for training and testing machine learning models. A dataset typically includes a variety of data points, each consisting of several attributes or features. The dataset's quality, size, and relevance are critical factors in the effectiveness of the machine learning models developed from it. The dataset plays a central role in the ability of these models to learn, adapt, and make accurate predictions or decisions [17].
- **Machine Learning:** This is a subfield of artificial intelligence that concentrates on developing and analyzing statistical algorithms capable of learning from data. These algorithms are designed to generalize and perform tasks without being explicitly programmed. Machine learning enables systems to make predictions or decisions based on data, improving their performance on specific tasks over time through experience [18].
- **Overfitting:** This term describes a phenomenon in machine learning where a model becomes too closely fitted to the specific patterns of the training data. As a result, it may lose its ability to generalize and perform effectively on new, unseen data. Overfitting typically occurs when a model is excessively complex relative to the amount and variety of the data available, capturing noise or random fluctuations in the training data as if they were vital features [19].
- **Predictor Variables:** These are variables used in data analysis and machine learning to predict or determine the value of a target variable. Predictor variables can encompass various data types, including numerical data, categorical data, text, and others. They serve as input in predictive models and are essential in identifying patterns or relationships that can be used to make accurate predictions or estimations [20].
- **Target or Objective Variable:** This is the variable or characteristic that a predictive model aims to forecast or anticipate. It represents the main focus or outcome of the prediction. In machine learning models, the target variable is what the model is trained to predict using the insights derived from the predictor variables. It is central to the objectives of the analysis and is often the critical measure of interest in various data-driven tasks [21].
- **Supervised Learning:** This is a type of machine learning where the model is trained on a labeled dataset. It involves identifying patterns and making predictions based on behaviors or characteristics observed in previously stored (historical) data. In supervised learning, the model learns from the input data (predictor variables) and the corresponding output (target variable) to make future predictions [21].
- **Classification:** Classification is a form of supervised learning that uses categorical labels. It is applied when the output or target variable belongs to a finite set of categories or classes. For example, it is used to categorize email as 'spam' or 'not spam' based on the features of the email [22].
- **Regression:** Regression, another form of supervised learning, aims to establish a relationship between a set of predictor variables and a continuous target variable. It is used for predicting numeric values, such as the price of a house, based on its features like size, location, and age. Unlike classification, regression deals with continuous outcomes rather than discrete categories [22].

Upon establishing a foundational understanding of the key concepts, the subsequent step involves a detailed delineation of the various phases of the CRISP-DM (Cross-Industry et al. for Data Mining) model. This methodological framework offers a structured approach to data mining and developing predictive models. The CRISP-DM methodology comprises six stages, interconnected sequentially and cyclically, forming a dynamic process. This structure allows for continuous interactions and refinements between the phases, enhancing the overall effectiveness of the data mining process. The stages of the CRISP-DM methodology include:

1. Business Understanding: Defining the project objectives and requirements from a business perspective.
2. Data Understanding: Collecting initial data and familiarizing with the data to identify quality issues.
3. Data Preparation: Preprocessing and cleaning the data for analysis.
4. Modeling: Selecting and applying various modeling techniques.
5. Evaluation: Assessing the models to ensure they meet the business objectives.
6. Deployment: Implementing the model into the operational environment.

The relationship and flow among these stages are graphically represented in Fig. 1, which provides a visual guide to the sequence and interconnectivity of each phase within the CRISP-DM framework [16].

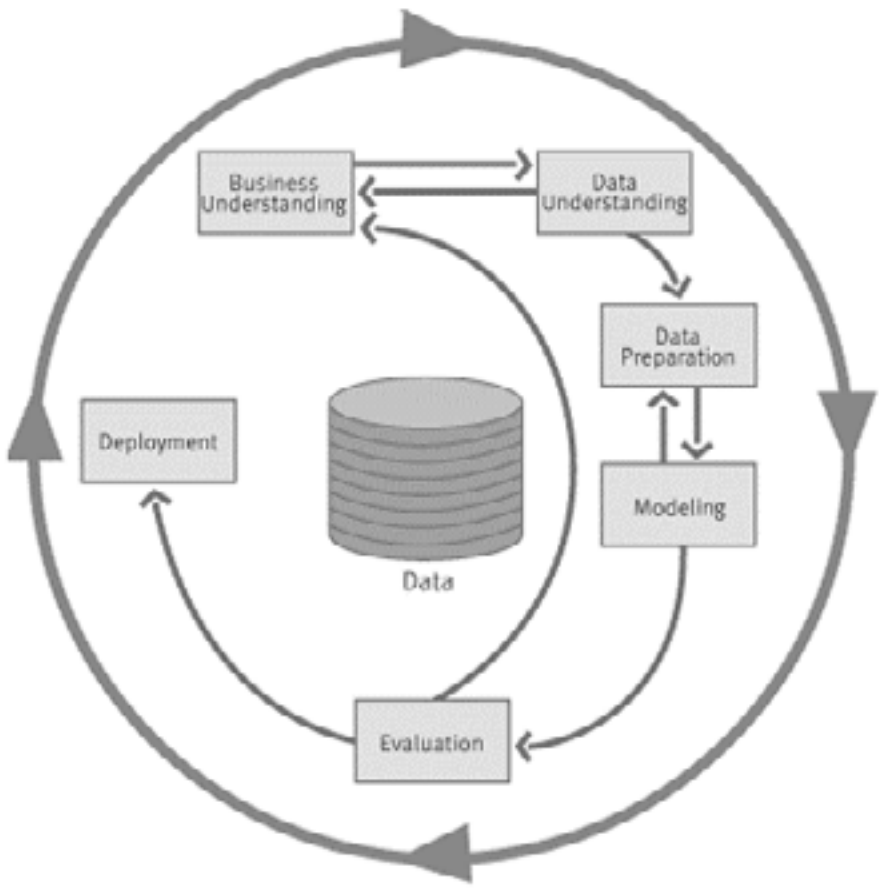

Fig. 1 CRISP-DM Methodology Cycle.
Source: https://anibalgoicochea.com/2009/08/11/crisp-dm-una-metodologia-para-proyectos-de-mineria-de-datos/

A. Business understanding

The initial stage of the CRISP-DM methodology focuses on gaining a comprehensive understanding of the problem's magnitude. This involves detailed analysis and clarity regarding the business objectives and challenges [16]. Pertinent to this project, the following key points are proposed:

- **Understanding the Business:** The primary goal of this project is to contribute to the prevention and early detection of diabetes while promoting overall health. The objective is to develop an accurate and reliable predictive model using the CRISP-DM methodology. This initiative aligns with the critical need for a preventive and proactive approach to diabetes management. By doing so, it aims to facilitate timely and effective interventions, thereby potentially improving health outcomes and reducing the long-term impact of diabetes.
- **Description of the Problem:** The primary issue addressed in this project is the enhancement of early-stage diabetes detection capabilities. Given the complexity of the disease and its significant impact on public health, a predictive approach is essential. This approach should utilize data related to risk factors, symptoms, and results from medical tests. Currently, the lack of precise tools for early diabetes detection often leads to delayed diagnoses and an increase in the prevalence of complications associated with the disease.
- **Purpose of Data Mining:** The overarching objective of data mining in this project is to conduct an in-depth analysis and understanding of data about diabetes. Specific goals include identifying significant patterns and relationships for predicting the disease. This involves applying data preparation techniques to ensure the quality and accuracy of the information and evaluating the performance of predictive models using key metrics. Data mining is a pivotal tool in developing a robust and precise model. This model aims to contribute to the early detection of diabetes, thereby enhancing public health outcomes and mitigating the broader impacts of the disease.

B. Understanding the data

The target variable in this study is designated as 'diabetes' and is categorized into two distinct groups: 'YES' and 'NO.' These categories are based on the diagnosis status of individuals within the pooled data set.

- Yes Group: This Group consists of individuals diagnosed with diabetes, as the data indicates. Their status in the dataset is marked as positive for the disease, reflecting the presence of diabetes in their medical and demographic history.
- No Group: This Group includes individuals in the diabetes prediction dataset who do not have diabetes, as indicated by their negative status in the dataset. These individuals have not been diagnosed with diabetes, according to the collected information, which sets them apart from those in the Yes Group.

The clear distinction between these two groups — Yes and No — is crucial for the study. It enables the exploration and identification of patterns in the data characteristic of individuals with diabetes (Yes Group) instead of those without the condition (No Group). Consequently, the variable "diabetes" serves as a vital component in distinguishing and analyzing the traits associated with the presence or absence of the disease among the study's participants.

C. Data preparation

Before commencing this study phase, becoming acquainted with the chosen work environment, Google Colab, is essential for Collaboratory. Developed by Google Research, Google Colab is a browser-based platform that allows users to write and execute Python code. This environment is particularly advantageous for machine learning, data analysis, and educational tasks. Additionally, Google Colab is an effective tool for developing new applications [23].

Upon familiarizing with the Google Colab environment, the next step involves importing essential libraries for the planned tasks. These libraries include:

- Pandas [24]: This library provides high-performance, user-friendly data structures and analysis tools. It is instrumental in handling and analyzing large datasets.
- Numpy [25]: Utilized for working with matrices or arrays, Numpy is essential for numerical computations.
- Matplotlib [26] and Seaborn [27]: Both libraries offer high-level interfaces for creating various statistical plots and graphs. They are crucial for visualizing data, which aids in understanding patterns and insights derived from the analysis.

Each of these libraries plays a vital role in the data analysis process, offering specialized functions and capabilities that enhance the efficiency and effectiveness of the work.

Additionally, the dataset used for diabetes prediction comprises an extensive collection of patient medical and demographic data coupled with their diabetes status (positive or negative). This dataset includes a variety of characteristics, encompassing age, gender, body mass index (BMI), presence of hypertension, history of heart disease, smoking habits, HbA1c levels, and blood glucose levels. These diverse data points provide a comprehensive view of each patient's health profile, which is crucial for accurate diabetes prediction. The detailed breakdown of this data is presented in Table 1 [28].

Tabla 1. Variables used in the Dataset.

Gender	Biological sex of the individual, which can affect susceptibility to diabetes. There are 3 categories: Male, Female and Other.
Age	Indicates the time elapsed since the individual's birth. The age range in the data set is 0.08 to 80 years.
Hypertension	A medical condition in which blood pressure in the arteries remains persistently elevated. A scale of 0 or 1 is used, where 0 indicates absence of hypertension and 1 indicates presence of hypertension.
Heart disease	Medical condition associated with an increased risk of developing diabetes. A scale of 0 or 1 is used, where 0 indicates absence of heart disease and 1 indicates presence of heart disease.
Smoking History	A risk factor for diabetes that can also aggravate associated complications. In our dataset, it is classified into 5 categories: not current, former, No Info, current, never and ever.
BMI (Body Mass Index)	A measure of body fat based on weight and height. Higher BMI values are associated with an increased risk of diabetes.
HbA1c_level	Measurement of a person's average blood glucose over the past 2-3 months. Higher levels indicate a higher risk of developing diabetes.
Blood_glucose_level	The amount of glucose in the bloodstream at any given time. Elevated blood glucose levels are a key indicator of diabetes.
Diabetes	Target variable to be predicted, with values of 1 indicating the presence of diabetes and 0 indicating the absence of diabetes.

Source: Author.

In this diabetes prediction dataset, an important aspect is the range of Body Mass Index (BMI) values, which spans from 10.16 to 71.55. The BMI values are categorized as follows: less than 18.5 indicates underweight, 18.5-24.9 is considered average weight, 25-29.9 is classified as overweight, and a BMI of 30 or more falls into the obesity category. Additionally, it's generally recognized that an HbA1c level of more than 6.5% indicates diabetes.Within the Google Colab environment, the provided dataset, which is in CSV (comma-separated va-

lues) format, will be imported. This file contains all the necessary data for analysis, including the predictor and target variables. The process of importing this dataset into Google Colab is depicted in Fig. 2. This step is crucial as it sets the stage for subsequent data processing and analysis.

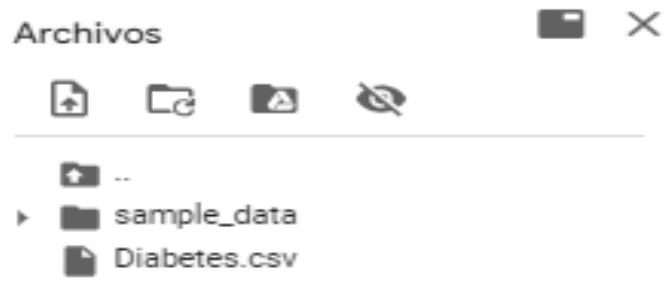

Fig. 2 Integration of the dataset to be used in Google Colab.
Source: Author

The Google Colab environment displays the first five records to facilitate a preliminary examination of the dataset's contents, thereby allowing for an initial data summarization [16]. This step is visually represented in Fig. 3.

	gender	age	hypertenYeson	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
0	Female	80.0	0	1	never	25.19	6.6	140	No
1	Female	54.0	0	0	No Info	27.32	6.6	80	No
2	Male	28.0	0	0	never	27.32	5.7	158	No
3	Female	36.0	0	0	current	23.45	5.0	155	No
4	Male	76.0	1	1	current	20.14	4.8	155	No

Fig. 3 Data integration. Source: Author

Fig. 3 presents the variables as outlined in Table 1. These variables are further scrutinized in detail for a deeper examination of the dataset, with the results showcased in Fig. 4.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 100000 entries, 0 to 99999
Data columns (total 9 columns):
 #   Column               Non-Null Count   Dtype
---  ------               --------------   -----
 0   gender               100000 non-null  object
 1   age                  100000 non-null  float64
 2   hypertension         100000 non-null  int64
 3   heart_disease        100000 non-null  int64
 4   smoking_history      100000 non-null  object
 5   bmi                  100000 non-null  float64
 6   HbA1c_level          100000 non-null  float64
 7   blood_glucose_level  100000 non-null  int64
 8   diabetes             100000 non-null  object
dtypes: float64(3), int64(3), object(3)
memory usage: 7.6+ MB
```

Fig. 4 Variable information.
Source: Author

In the depiction provided by Fig. 4, the dataset is revealed to encompass 100,000 entries. These entries identify age, BMI, HbA1c levels, and blood glucose levels as numerical variables. Additionally, the dataset includes categorical variables such as gender and diabetes status, along with other critical factors like heart disease and hypertension [22]. This comprehensive overview of the dataset is crucial for understanding the scope and nature of the data, setting the stage for subsequent analysis.

As part of advancing the CRISP-DM methodology, the current phase concentrates on eliminating irrelevant and redundant variables from the dataset. This step is critical to refining the dataset, ensuring it retains only those variables that provide significant information for the analysis. Variables that offer little value or are highly correlated with others are considered for exclusion. This process simplifies the model, enhancing its interpretability and efficiency [13]. In this case, gender, age, hypertension, heart disease, smoking history, BMI, hemoglobin A1C, and blood glucose levels are deemed necessary for diabetes prediction and are thus retained as they are not considered irrelevant or redundant.

The findings, which result from proceeding with the statistical description of the data, are presented in Fig. 5.

	age	hypertension	heart_disease	bmi	HbA1c_level	blood_glucose_level
count	100000.000000	100000.00000	100000.000000	100000.000000	100000.000000	100000.000000
mean	41.885856	0.07485	0.039420	27.320767	5.527507	138.058060
std	22.516840	0.26315	0.194593	6.636783	1.070672	40.708136
min	0.080000	0.00000	0.000000	10.010000	3.500000	80.000000
25%	24.000000	0.00000	0.000000	23.630000	4.800000	100.000000
50%	43.000000	0.00000	0.000000	27.320000	5.800000	140.000000
75%	60.000000	0.00000	0.000000	29.580000	6.200000	159.000000
max	80.000000	1.00000	1.000000	95.690000	9.000000	300.000000

Fig. 5 Statistical description of numerical variables.
Source: Author

The analysis reveals:

- The average age in the population is approximately 41.89 years, with a standard deviation of about 22.52 years.
- About 7.49% of the patients have hypertension, and roughly 3.94% suffer from heart disease.
- The mean BMI is around 27.32, with a standard deviation of about 6.64.
- The average HbA1c and blood glucose levels are approximately 5.53 and 138.06, respectively.

These statistics offer a comprehensive snapshot of the demographic and health-related characteristics of the population, underscoring the diversity and range of critical factors relevant to diabetes predictive analysis.

Further statistical exploration continues, focusing on the numerical variables. Representative graphical depictions of these statistics are shown in Fig. 6. This visual representation provides an intuitive understanding of the data distribution and key trends among the variables.

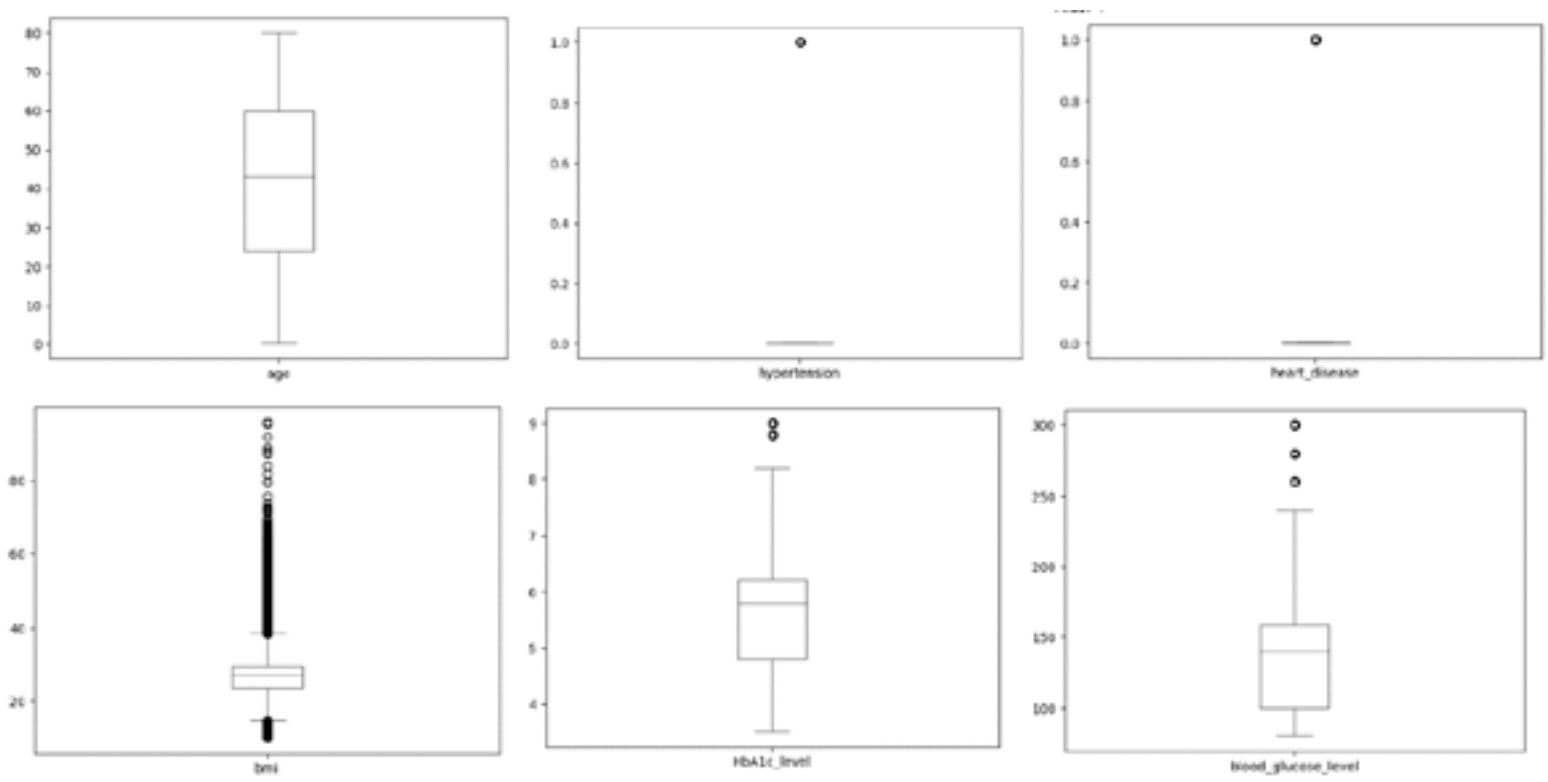

Fig. 6 Graphs for numerical variables.
Source: Author

To further enhance our understanding of the dataset, pie charts were used to visualize the distribution of categorical variables. These charts display the different categories as pie segments, providing a clear and intuitive representation of their proportions within the dataset. The results of this visualization are presented in Fig. 7.

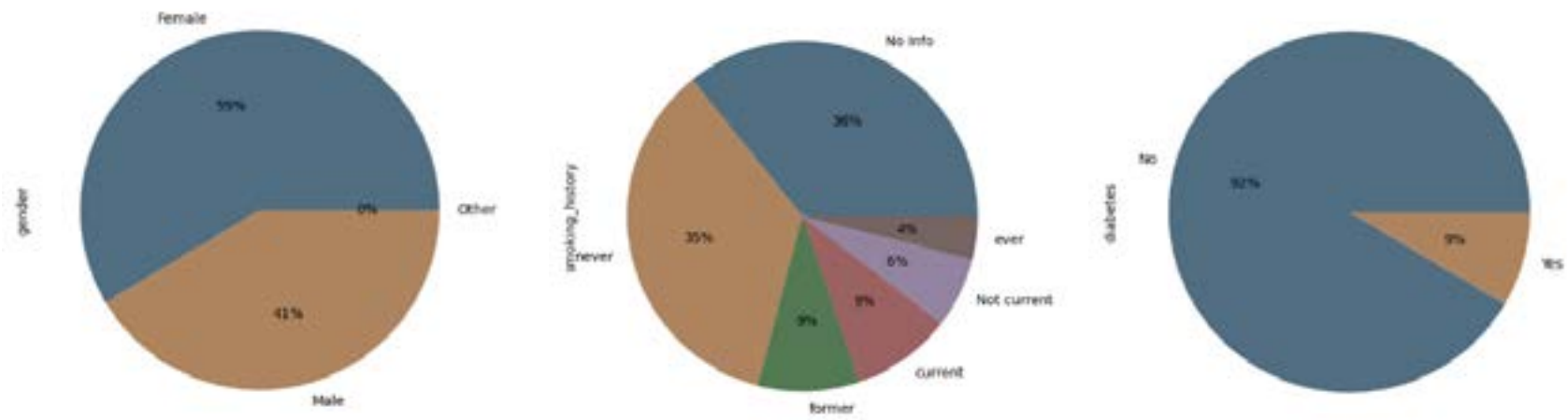

Fig 7. Graphs for categorical variables.
Source: Author

Next, the process of cleaning outlier data was undertaken. This step involved evaluating the data against the pre-established quality rules to determine the presence of outliers [10]. While outliers were identified in the dataset, their existence was not deemed detrimental to the analysis. This is because these outliers conform to the quality rules and do not constitute a significant portion of the dataset compared to the valuable data. Hence, their presence was accepted in the final dataset used for analysis.

Following the identification and handling of outliers, the next step involves cleaning null values from the dataset. This process, known as imputation, entails replacing missing values with estimated ones to maintain data integrity and completeness [29]. This aspect of data cleaning is corroborated by the information in Fig. 4, which indicates the absence of null data in any of the columns.

After addressing outlier and null data, the focus shifts to balancing the dataset [13]. his step is particularly crucial for classification analysis, given that the target variable is divided into two distinct groups, as previously mentioned [22]. Data balancing is necessary to mitigate the impact of significant disparities in each class's sample sizes. Such imbalances can adversely affect the performance of predictive models. The SMOTENC library, which is specifically designed for handling imbalanced datasets [30], is utilized to achieve a more balanced dataset.

Subsequently, an informative summary that outlines the structure and composition of the dataset is presented. The results of these steps, including the data balancing process and its impact on the dataset, are illustrated in Figs. 8 and 9.

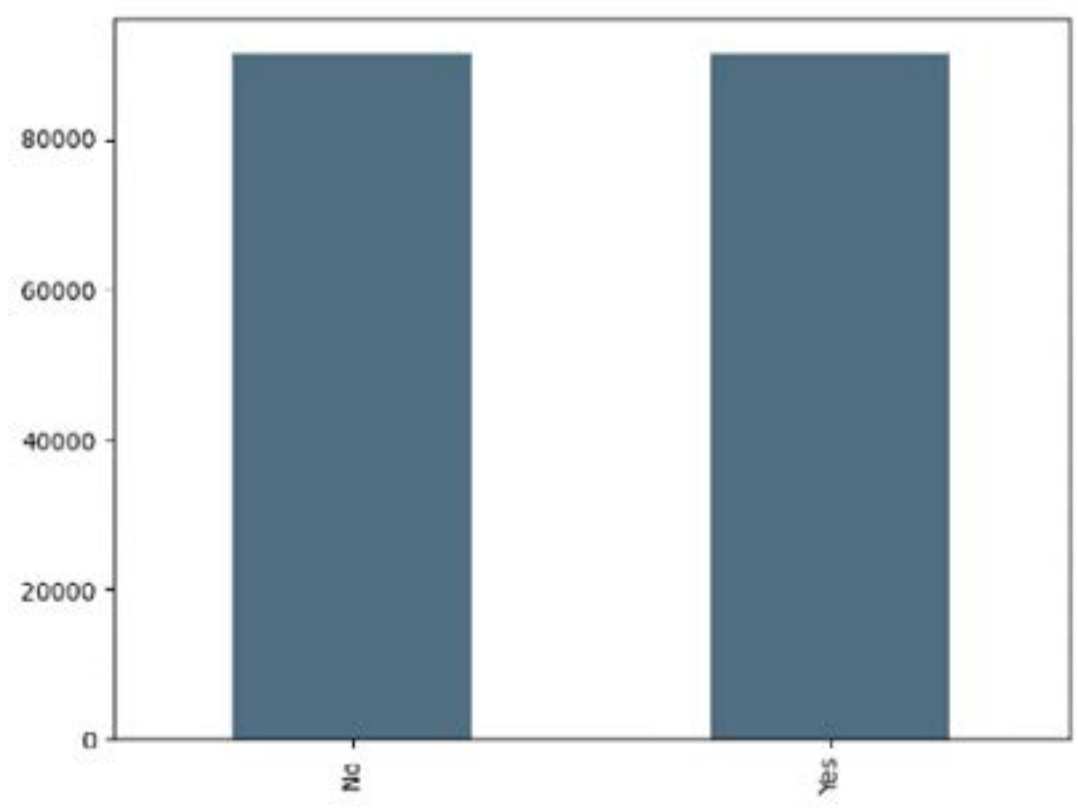

Fig. 8 Graph for the balanced target variable.
Source: Author

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 183000 entries, 0 to 182999
Data columns (total 16 columns):
 #   Column                      Non-Null Count   Dtype
---  ------                      --------------   -----
 0   age                         183000 non-null  float64
 1   hypertension                183000 non-null  int64
 2   heart_disease               183000 non-null  int64
 3   bmi                         183000 non-null  float64
 4   HbA1c_level                 183000 non-null  float64
 5   blood_glucose_level         183000 non-null  int64
 6   gender_Female               183000 non-null  uint8
 7   gender_Male                 183000 non-null  uint8
 8   gender_Other                183000 non-null  uint8
 9   smoking_history_No Info     183000 non-null  uint8
 10  smoking_history_Not current 183000 non-null  uint8
 11  smoking_history_current     183000 non-null  uint8
 12  smoking_history_ever        183000 non-null  uint8
 13  smoking_history_former      183000 non-null  uint8
 14  smoking_history_never       183000 non-null  uint8
 15  diabetes                    183000 non-null  int64
dtypes: float64(3), int64(4), uint8(9)
memory usage: 11.3 MB
```

Fig 9. Dataset with data balancing implementation.
Source: Author

The final step in the data preparation process is the dataset transformation, which must be tailored to the specific Machine Learning models chosen. For this experiment, the following models have been selected:

- The TREE model (Decision Trees) [31]
- The KNN (K-nearest neighbors) model [32]

D. Modeling

As previously mentioned, the transformations or modifications to the variables are dependent on the chosen model. Accordingly, the evaluation will commence with the TREE model. For this model, dummy variables are created for the categorical variables (as illustrated in Fig. 10), and the target variable is encoded using a Label Encoder (shown in Fig. 11). It is important to note that the TREE model does not require the normalization of numerical variables [31].

gender_Female	gender_Male	gender_Other
1	0	0
1	0	0
0	1	0
1	0	0
0	1	0

smoking_history_No Info	smoking_history_Not current	smoking_history_current
0	0	0
1	0	0
0	0	0
0	0	1
0	0	1

Fig. 10 Variables implementing dummies.
Source: Author

diabetes
0
0
0
0
0

Fig. 11 Target variable implementing Label Encoder.
Source: Author

Following the data preparation, a 70%-30% train-test split, also known as train-test Split, is employed to simulate the model's performance on unseen data. This approach involves dividing the dataset into two subsets: a "training set" used for training the model and a "testing set" for evaluating its performance [33]. This process is exemplified in Fig. 12.

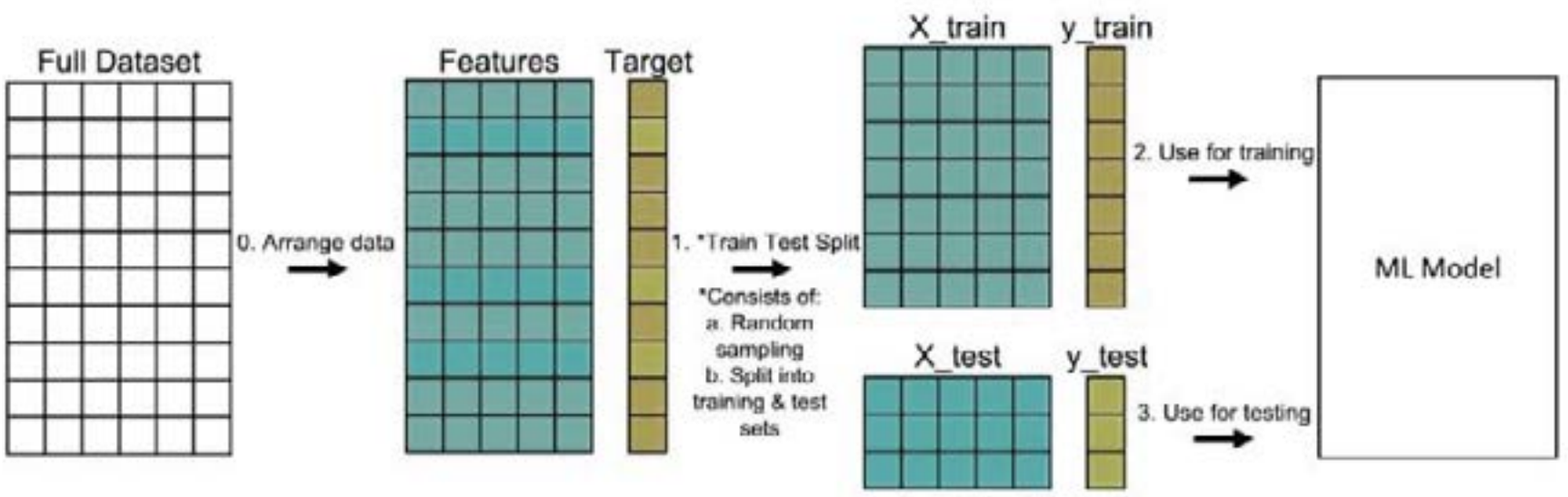

Fig. 12 Train Test Split Method.

Source: https://builtin.com/sites/www.builtin.com/files/styles/ckeditor_optimize/public/inline-images/1_train-test-split_0.jpg

The Decision Trees model uses the DecisionTreeClassifier library from Scikit-learn [34] to create the model with the training set. Additionally, the plot_tree library [35] from Scikit-learn [36] is employed to visualize the decision tree.

Once the decision tree model is established, the focus shifts to the KNN (K-nearest neighbors) model. For the KNN model, it is essential to normalize the numerical variables (as shown in Fig. 13) and convert categorical variables into dummies (illustrated in Fig. 14). The target variable also requires encoding using a label encoder (displayed in Fig. 15). It is important to note that since the 70%-30% validation has already been completed. Hypertension and HbA1c level variables are already within the 0 to 1 range; they do not require additional normalization.

	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level
77712	Female	0.299299	0	0	current	0.202031	0.400000	0.227273
9795	Female	0.199199	0	0	never	0.096055	0.181818	0.295455
75083	Female	0.411912	0	0	never	0.211134	0.472727	0.272727
135975	Female	0.699700	0	0	never	0.345093	0.447141	0.636364
165100	Female	0.034331	0	0	ever	0.300000	0.073327	0.295455

Fig. 13 Standardized numerical variables.

Source: Author

gender_Female	gender_Male	gender_Other	smoking_history_No Info	smoking_history_Not current
1	0	0	0	0
1	0	0	1	0
0	1	0	0	0
1	0	0	0	0
0	1	0	0	0

Fig. 14 Categorical variables in dummies.

Source: Author

diabetes	ger
0	
0	
0	
0	
0	

Fig. 15 Target variable with Label Encoder.

Source: Author

The next step involves training a K-Nearest Neighbors (KNN) classification model on 70% of the training data. For this purpose, the KNeighborsClassifier class from the Scikit-learn. Neighbors package is employed [40]. The model is configured with specific parameters, including the number of neighbors, and uses an Euclidean distance metric to create the model. After setting these parameters, the model is then fitted to the training data.

E. RESULTS

Regarding hyperparameter tuning, a focused approach is adopted, particularly regarding the depth of the decision tree. The goal is to maintain a relatively small tree depth to enhance the model's generalization ability. The hyperparameter values are adjusted to 2, 4, and 6 for experimentation to achieve this. Starting with smaller values is advisable as it often leads to better results. This strategy aims to optimize the model's performance and ensure more reliable generalization when applied to new, unseen data [37].

When working with smaller hyperparameter values in the decision tree model, the aim is to reduce the risk of overfitting. High precision in the model's predictions can often indicate overfitting [31], so a value of 2 is optimal for balancing accuracy and generalizability. Various tools are used to evaluate the model's performance, including a diagram (Fig. 16), a report table (Table 2) [38], a confusion matrix (Fig. 17), and a ROC curve (Fig. 18). These tools collectively provide a comprehensive view of the decision tree model's effectiveness.

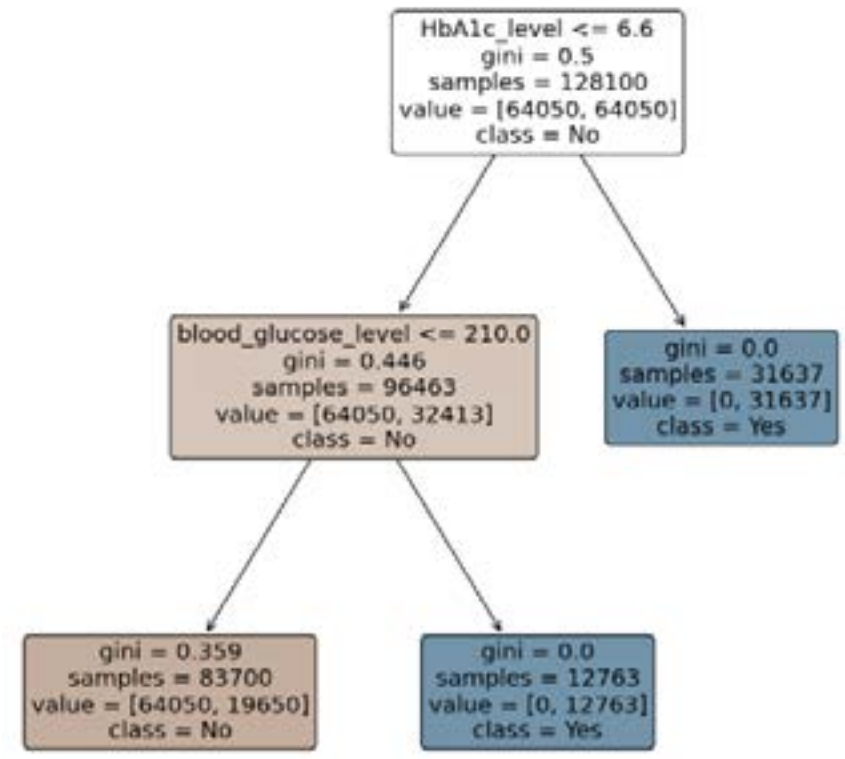

Fig. 16 Decision tree graph.
Source: Author

The CART (Classification and Regression Trees) method is employed to construct the decision tree. This technique builds a binary decision tree from the training dataset by recursively dividing the data into smaller subsets based on predictor variables. In this instance, the HbA1c_level is identified as the key variable from which the tree nodes are derived [39]. Various performance metrics are evaluated following the tree's construction, with the results presented in Table 2. This evaluation is crucial for assessing the model's accuracy, sensitivity, and specificity, ensuring its reliability and utility in diabetes prediction.

Tabla 2. Decision Tree Report.

	Precision	Recall	F1-score	Support
No	0.77	1.00	0.87	27450
Yes	1.00	0.70	0.82	27450
accuracy			0.85	54900
macro avg	0.88	0.85	0.84	54900
Weighted avg	0.88	0.85	0.84	54900

Source: Author

The obtained metrics demonstrate the effectiveness of the model. Precision indicates the proportion of correct positive predictions relative to the total positive predictions made. Recall reflects the proportion of actual positive cases that were correctly identified. The F1-score is a metric combining precision and recall for a more comprehensive evaluation. For the "No" class, the model achieves a precision of 77%, a recall of 100%, and an F1-score of 87%. In the case of the "Yes" class, the precision is 100%, recall is 70%, and the F1-score is 82%. These figures are crucial in assessing the model's overall performance, which boasts an accuracy of 85%. This indicates the model's strong capability to classify the classes accurately.

The confusion matrix [39] provides additional insights, as illustrated in Fig. 17. In the matrix, position (1,1) signifies the instances correctly classified as "No," amounting to 27,450 cases. Position (2,2) indicates the instances correctly identified as "Yes," totaling 19,082. Notably, position (1,2) confirms that there were no false positives, meaning no instances were incorrectly classified as "Yes" when they were "No." However, position (2,1) reveals 8,368 false negatives, where some "Yes" instances were erroneously classified as "No."

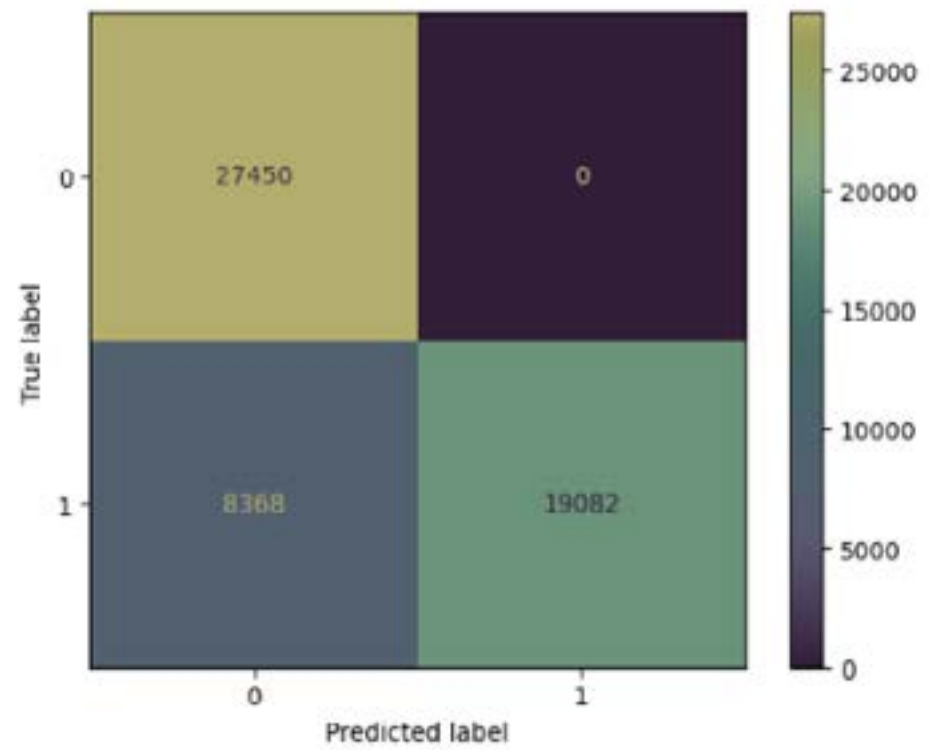

Fig. 17 Confusion matrix for the decision tree.
Source: Author

To further augment this analysis, the ROC (Receiver Operating Characteristic) curve is examined, as shown in Fig. 18. The ROC curve is a graphical representation that assesses the diagnostic ability of a binary classifier system, providing a visual summary of the trade-off between true positive rate and false positive rate at various thresholds.

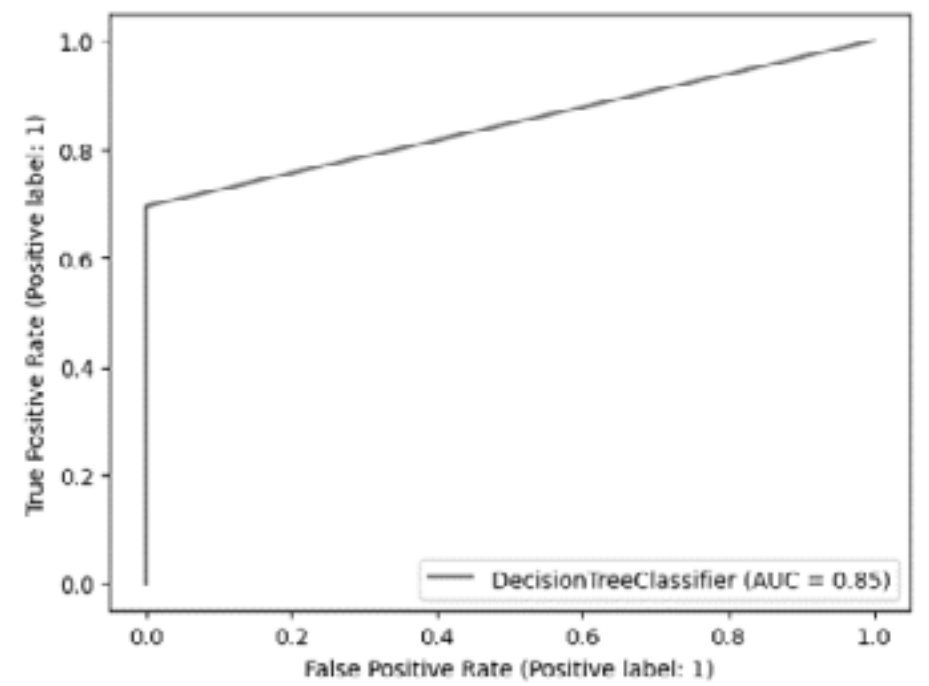

Fig. 18 ROC curve for the decision tree.
Source: Author

The ROC curve, along with its associated metric, the Area Under the Curve (AUC), is a crucial tool for assessing the performance of a classification model. These measures evaluate the model's ability to accurately distinguish between positive and negative classes at various decision thresholds. An AUC value of 0.85 signifies that the model can discriminate between the classes, suggesting its effectiveness in the classification task at hand.

After completing the analysis with the Tree model, attention shifts to the KNN (K-Nearest Neighbors) model [32]. For this model, it is essential to consider the specific transformations required for the variables. The KNN model has its own set of prerequisites regarding data preparation, especially regarding how the variables are treated and presented to the model. This step is vital to ensure the model's optimal performance and accuracy in the classification process.

For experimental purposes, adjustments were made to the KNN model by varying the parameters representing the nearest neighbors. Specifically, 1, 3, and 5 values were tested to gain insights into how the model's performance might change with different configurations [41]. This approach helps understand the model's sensitivity to the number of neighbors considered and its impact on classification accuracy.

Through the experimentation process, it was observed that variations in the number of neighbors did not significantly alter the model's performance. Consequently, adhering to the principle of minimal complexity, the choice was made to use the smallest number of neighbors, which in this case is 1. This decision is based on the understanding that a simpler model is often more robust and less prone to overfitting while maintaining effectiveness. The results and performance metrics of the KNN model with this configuration are detailed in Table 3. This table provides a comprehensive view of the model's classification accuracy and other key performance indicators, aiding in evaluating its efficacy.

Tabla 3. KNN Model Report.

	Precision	Recall	F1-score	Support
No	0.50	0.65	0.56	27369
Yes	0.50	0.34	0.41	27531
Accuracy			0.50	54900
Macro avg	0.50	0.50	0.48	54900
Weighted avg	0.50	0.50	0.48	54900

Source: Author

In this instance, the KNN model exhibits an accuracy rate of 50%, indicating that it correctly predicts the class of an instance (whether "Yes" or "No" for diabetes) half the time. This level of accuracy reflects the model's overall predictive capability across both classes.

The recall metric, which assesses the model's ability to identify all relevant instances of a particular class correctly, reveals a disparity between the two classes. For the "No" class, the recall is relatively higher at 65%, suggesting that the model is more effective at identifying instances where diabetes is absent. Conversely, for the "Yes" class, the recall is notably lower at 34%, indicating that the model is less adept at correctly identifying instances where diabetes is present.

To further understand these results, the confusion matrix is analyzed, as depicted in Fig. 19. The confusion matrix offers a detailed breakdown of the model's predictions, showing the number of true positives, true negatives, false positives, and false negatives. This visual representation aids in comprehensively evaluating the model's performance, especially in terms of its ability to distinguish between the two classes.

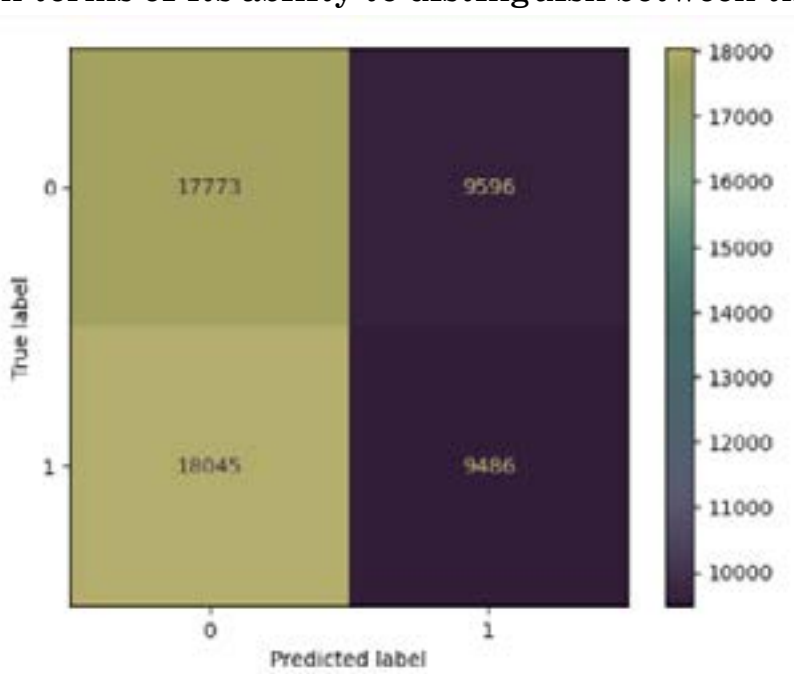

Fig. 19 Confusion matrix for the KNN model.

Source: Author

In the confusion matrix, displayed in position (1,1), a total of 17,773 instances were correctly identified as "No" (true negatives), demonstrating the model's effectiveness in correctly classifying cases without diabetes. Conversely, in position (2,2), the model correctly classified 9,486 instances as "Yes" (true positives), indicating its capacity to identify actual cases of diabetes accurately.

However, the model also produced a notable number of false classifications. Specifically, position (1,2) shows that there were 9,596 instances erroneously classified as "Yes" (false positives), where the model incorrectly predicted diabetes. Similarly, in position (2,1), there were 18,045 instances misclassified as "No" (false negatives), where the model failed to identify actual cases of diabetes.

The ROC curve for the model is examined to supplement this analysis, as shown in Fig. 20. The ROC curve provides a graphical representation of the model's diagnostic ability, illustrating the trade-off between the true positive rate and the false positive rate at various threshold settings. This visualization is instrumental in assessing the overall predictive performance of the KNN model.

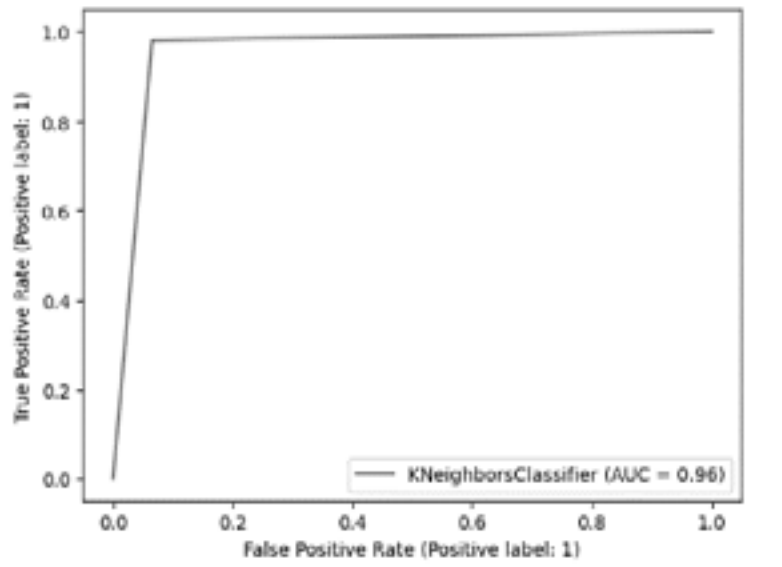

Fig. 20 ROC curve of the KNN model.
Source: Author

The AUC value of 0.96 indicates the model possesses a high discriminatory power, effectively distinguishing between positive (diabetic) and negative (non-diabetic) instances. An AUC value close to 1 indicates excellent model performance, suggesting that the model can make accurate classifications. Conversely, an AUC value around 0.5 would imply that the model's performance is no better than random chance, highlighting the significance of achieving a high AUC value in predictive modeling.

Upon completing the evaluation and fine-tuning of the model, the final step involves exporting it for implementation in other domains or applications. This exportation process makes the model available for broader use, allowing its predictive capabilities to be leveraged in different settings and potentially contributing to more effective diabetes management strategies in diverse contexts.

E. Deployment

In this project phase, the developed models — Tree and KNN — will be tested with new data to simulate their performance in a real-world, company-like environment [42]. This step is crucial for assessing how the models behave when exposed to data they have not encountered during the training phase.

The first step involves importing the new dataset to be used for predictions. This process and the contents of the dataset are illustrated in Figs. 21 and 22, respectively.

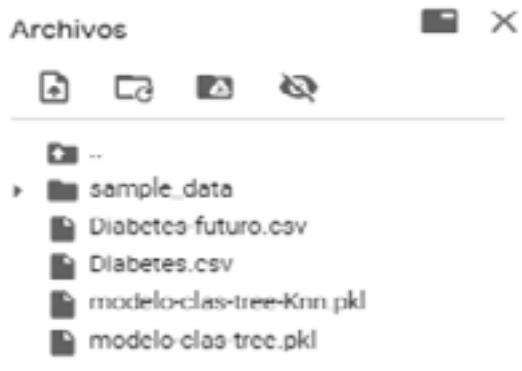

Fig 21. Dataset to predict imported into Google Colab.
Source: Author

	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes_real
0	Male	76.0	1	1	current	20.14	4.8	155	No
1	Male	42.0	0	0	never	33.64	4.8	145	No
2	Female	79.0	0	0	No Info	23.86	5.7	85	No
3	Female	20.0	0	0	never	27.32	6.6	85	No
4	Male	28.0	0	0	never	27.32	5.7	158	No
5	Female	44.0	0	0	never	19.31	6.5	200	Yes
6	Male	67.0	0	1	not current	27.32	6.5	200	Yes
7	Male	50.0	1	0	current	27.32	5.7	260	Yes
8	Male	73.0	0	0	former	25.91	9.0	160	Yes
9	Female	53.0	0	0	former	27.32	7.0	159	Yes

Fig 22. Dataset records to be predicted.
Source: Author

Following the import, the next step is to make predictions using the models. The results of these predictions are depicted in Fig. 23.

	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes_real	Prediccion Tree
0	Male	76.0	1	1	current	20.14	4.8	155	No	No
1	Male	42.0	0	0	never	33.64	4.8	145	No	No
2	Female	79.0	0	0	No Info	23.86	5.7	85	No	No
3	Female	20.0	0	0	never	27.32	6.6	85	No	No
4	Male	28.0	0	0	never	27.32	5.7	158	No	No
5	Female	44.0	0	0	never	19.31	6.5	200	Yes	No
6	Male	67.0	0	1	not current	27.32	6.5	200	Yes	No
7	Male	50.0	1	0	current	27.32	5.7	260	Yes	Yes
8	Male	73.0	0	0	former	25.91	9.0	160	Yes	Yes
9	Female	53.0	0	0	former	27.32	7.0	159	Yes	Yes

Fig. 23 Tree model prediction.
Source: Author

When using the KNN model for predictions, the dataset undergoes further preprocessing. This includes encoding categorical variables into dummies, as shown in Fig. 24, and normalizing the numerical variables, which is illustrated in Fig. 25. It is important to note that applying Label Encoder to the target variable is not necessary for this model, as no modifications are made to it [43].

gender_Female	gender_Male	smoking_history_No Info	smoking_history_current	smoki
0	1	0	1	
0	1	0	0	
1	0	1	0	
1	0	0	0	
0	1	0	0	

Fig. 24 Dummies for categorical variables.
Source: Author

e	bmi	HbA1c_level	blood_glucose_level	d
1	0.057920	0.000000	0.400000	
0	1.000000	0.000000	0.342857	
0	0.317516	0.214286	0.000000	
0	0.558967	0.428571	0.000000	
0	0.558967	0.214286	0.417143	

Fig. 25 Normalization of categorical variables.
Source: Author

For the final aspect of the study, the prediction of diabetes using the KNN model is presented, with the corresponding result displayed in Fig. 26.

	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes_real	Prediccion Knn
0	Male	76.0	1	1	current	20.14	4.8	155	No	No
1	Male	42.0	0	0	never	33.64	4.8	145	No	No
2	Female	79.0	0	0	No Info	23.86	5.7	85	No	No
3	Female	20.0	0	0	never	27.32	6.6	85	No	No
4	Male	28.0	0	0	never	27.32	5.7	158	No	No
5	Female	44.0	0	0	never	19.31	6.5	200	Yes	No
6	Male	67.0	0	1	not current	27.32	6.5	200	Yes	No
7	Male	50.0	1	0	current	27.32	5.7	260	Yes	No
8	Male	73.0	0	0	former	25.91	9.0	160	Yes	No
9	Female	53.0	0	0	former	27.32	7.0	159	Yes	No

Fig. 26 KNN Prediction.
Source: Author

IV. CONCLUSIONS

The thorough data collection provides an extensive understanding of the studied phenomenon. Each stage of the CRISP-DM process is crucial, from gaining an initial understanding of the problem to the final evaluation of the implemented models.

In analyzing the predictions made by the decision tree model, there is a noticeable alignment between the model's predictions ("Prediction Tree") and the actual diabetes status ("diabetes_actual"), reflecting a satisfactory level of performance. Nevertheless, there are instances where the model's predictions diverge from the actual condition, highlighting opportunities for further refinement. Factors such as "smoking_history" and "blood_glucose_level" emerge as significant influencers in the model's predictions, pointing toward areas that may require adjustment to enhance accuracy. A detailed examination of these mismatches is vital for improving the model's accuracy and reliability, strengthening its overall predictive capability.

The Tree model demonstrates an overall accuracy of 50%, which indicates a limited capacity for making accurate predictions. The macro and weighted average metrics underscore the model's challenges in effectively classifying instances into the "No" and "Yes" classes. These metrics suggest that while the model may perform reasonably in some respects, there is significant room for improvement, particularly in achieving a more balanced classification performance across both classes.

Regarding the KNN model, the results obtained during the experimentation phase reveal a constrained level of effectiveness, particularly in distinguishing between the "No" and "Yes" classes. The model's limited predictive proficiency is evidenced by its low accuracy and the pronounced imbalance in identifying instances across different classes. This situation necessitates considering adjustments or exploring alternative modeling strategies to enhance the model's performance.

These outcomes underscore the critical need for ongoing review and enhancement of the model's quality to achieve a more balanced and accurate predictive performance. The occurrence of false positives and false negatives indicates the model's struggle to maintain an even classification capability for both classes. Moreover, the lower f1-score reflects the challenges in striking an optimal balance between precision and recall. These insights are crucial for guiding future improvements and adjustments in the model, aiming for a more robust and reliable predictive tool.

Funding: This research has been developed with our own resources.
CRediT authorship contribution statement: Juan Montes-Bustamente - Writing: review & editing, Writing: original draft, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization.
Conflict of interest: The authors declare that they have no conflict of interest in reporting on this study.

V. ANNEXES

Code created in Google Colab:
https://colab.research.google.com/drive/14MnBhWajDjxOp35oUUmsOMSlEsF97v7_?usp=sharing

REFERENCES

[1] «¿Qué es la diabetes? | Información Básica | Diabetes | CDC». [En línea]. Disponible en: https://www.cdc.gov/diabetes/spanish/basics/diabetes.html

[2] J. Oliva Moreno y L. M. Peña Longobardo, «Impacto económico de la diabetes mellitus», *Diabetes Práctica*, vol. 13, n.º 1, 2022, doi: 10.52102/diabet/pract/2022.1/art3.

[3] «La diabetes tipo 2 | Spanish | Diabetes | CDC». Disponible en: https://www.cdc.gov/diabetes/spanish/basics/type2.html

[4] P. Vigil-De Gracia, J. Olmedo, P. Vigil-De Gracia, y J. Olmedo, «Diabetes gestacional: conceptos actuales», *Ginecol. Obstet. México*, vol. 85, n.º 6, pp. 380-390, 2017, [En línea]. Disponible en: http://www.scielo.org.mx/scielo.php?script=sci_abstract&pid=S0300-90412017000600380&lng=es&nrm=iso&tlng=es

[5] «Diabetes - PAHO/WHO | Pan American Health Organization». [En línea]. Disponible en: https://www.paho.org/en/topics/diabetes

[6] R. Simó y C. Hernández, «Tratamiento de la diabetes mellitus: objetivos generales y manejo en la práctica clínica», *Rev. Esp. Cardiol.*, vol. 55, n.º 8, pp. 845-860, ago. 2002, [En línea]. Disponible en: http://www.revespcardiol.org/es-tratamiento-diabetes-mellitus-objetivos-generales-articulo-13035236

[7] I. L. Fe, «La IA como herramienta para detectar y controlar la diabetes y la retinopatía diabética | Blog», IIS La Fe. [En línea]. Disponible en: https://www.iislafe.es/es/sociedad/blog/5/la-ia-como-herramienta-para-detectar-y-controlar-la-diabetes-y-la-retinopatia-diabetica

[8] D. A. Ordóñez Barrios, «Modelo Predictivo para el diagnóstico de la Diabetes Mellitus Tipo 2 soportado por SAP Predictive Analytics», Licenciatura, Universidad Peruana de Ciencias Aplicadas, Lima, 2018. doi: 10.19083/tesis/624417.

[9] W. Hoyos, K. Hoyos, y R. Ruíz, «Modelo de inteligencia artificial para la detección temprana de diabetes», *Biomédica*, vol. 43, n.º Sp. 2, Art. n.º Sp. 2, nov. 2023, doi: 10.7705/biomedica.7147.

[10] J. Y. Rosales Malpartida, «Predicción de diabetes mellitus tipo 2 utilizando atributos médicos del Policlínico Leo SAC de San Juan de Lurigancho mediante el enfoque de Machine Learning», *TecnoHumanismo*, vol. 2, n.º 4, pp. 1-19, 2022. [En línea]. Disponible en: https://dialnet.unirioja.es/servlet/articulo?codigo=8510612

[11] «La detección precoz y las tecnologías aplicadas, claves en la atención a la diabetes», Junta de Andalucía. [En línea]. Disponible en: https://www.juntadeandalucia.es/presidencia/portavoz/salud/166358/ConsejeriadeSalud/prevencion/deteccionprecoz/diabetes/complicaciones/tecnologias

[12] «Guía de CRISP-DM de IBM SPSS Modeler».

[13] A. anibal goicochea, «CRISP-DM, Una metodología para proyectos de Minería de Datos», aníbal goicochea. [En línea]. Disponible en: https://anibalgoicochea.com/2009/08/11/crisp-dm-una-metodologia-para-proyectos-de-mineria-de-datos/

[14] M. Coppola, «Qué es la minería de datos: conceptos, técnicas y ejemplos». [En línea]. Disponible en: https://blog.hubspot.es/marketing/mineria-datos

[15] J. Herrera Herbert, *Introducción a la Minería. Vol. I: Conceptos, tecnologías y procesos*, 2.ª ed. Madrid: Universidad Politécnica de Madrid. Escuela Técnica Superior de Ingenieros de Minas y Energía, 2017. doi: 10.20868/UPM.book.63396.

[16] D. Á. Gil, «Metodología CRISP-DM - Adictos al trabajo Tutoriales», Adictos al trabajo. [En línea]. Disponible en: https://www.adictosaltrabajo.com/2021/01/14/metodologia-crisp-dm/

[17] «Cómo preparar un conjunto de datos para machine learning y análisis | datos.gob.es». [En línea]. Disponible en: https://datos.gob.es/es/blog/como-preparar-un-conjunto-de-datos-para-machine-learning-y-analisis

[18] «¿Qué es el machine learning? - Explicación sobre el machine learning empresarial - AWS», Amazon Web Services, Inc. [En línea]. Disponible en: https://aws.amazon.com/es/what-is/machine-learning/

[19] «¿Qué es el sobreajuste? - Explicación del sobreajuste en machine learning - AWS», Amazon Web Services, Inc. [En línea]. Disponible en: https://aws.amazon.com/es/what-is/overfitting/

[20] «Opcional - Unidad de aprendizaje - Actualización de Machine Learning 101», AI Planet (formerly DPhi). [En línea]. Disponible en: https://aiplanet.com/getting-started-with-deep-learning-es/fundamentos-de-deep-learning-y-redes-neuronales/1775/opcional-unidad-de-aprendizaje-actualizacion-de-machine-learning-101

[21] «Conceptos básicos de Machine Learning – Cleverdata». [En línea]. Disponible en: https://cleverdata.io/conceptos-basicos-machine-learning/

[22] «Machine Learning: Algoritmos de clasificación y regresión», The Black Box Lab. [En línea]. Disponible en: https://theblackboxlab.com/2022/05/06/machine-learning-diferencias-entre-algoritmos-clasificacion-regresion/

[23] «colab.google», colab.google. [En línea]. Disponible en: http://0.0.0.0:8080/

[24] «pandas - Python Data Analysis Library».[En línea]. Disponible en: https://pandas.pydata.org/

[25] «NumPy». [En línea]. Disponible en: https://numpy.org/

[26] «Matplotlib — Visualization with Python». [En línea]. Disponible en: https://matplotlib.org/

[27] M. Waskom, «seaborn: statistical data visualization», *J. Open Source Softw.*, vol. 6, n.º 60, p. 3021, abr. 2021, doi: 10.21105/joss.03021.

[28] «Diabetes prediction dataset». [En línea]. Disponible en: https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset

[29] «Guía completa para el Manejo de Datos Faltantes», Codificando Bits. [En línea]. Disponible en: https://www.codificandobits.com/blog/manejo-datos-faltantes/

[30] «SMOTENC — Version 0.11.0». [En línea]. Disponible en: https://imbalanced-learn.org/stable/references/generated/imblearn.over_sampling.SMOTENC.html

[31] R. KeepCoding, «Árboles de decisión sobre series temporales». [En línea]. Disponible en: https://keepcoding.io/blog/arboles-de-decision-sobre-series-temporales/

[32] «¿Qué es el algoritmo de k vecinos más cercanos? | IBM». [En línea]. Disponible en: https://www.ibm.com/es-es/topics/knn

[33] «Train Test Split: What it Means and How to Use It | Built In». [En línea]. Disponible en: https://builtin.com/data-science/train-test-split

[34] «sklearn.tree.DecisionTreeClassifier», scikit-learn. [En línea]. Disponible en: https://scikit-learn/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html

[35] «sklearn.tree.plot_tree», scikit-learn. [En línea]. Disponible en: https://scikit-learn/stable/modules/generated/sklearn.tree.plot_tree.html

[36] «scikit-learn: machine learning in Python — scikit-learn 1.3.2 documentation». [En línea]. Disponible en: https://scikit-learn.org/stable/

[37] R. KeepCoding, «2 características de los boosted trees: hiperparámetros e interpretabilidad». [En línea]. Disponible en: https://keepcoding.io/blog/2-caracteristicas-de-los-boosted-trees/

[38] «Crea un Arbol de Decisión en Python | Aprende Machine Learning». [En línea]. Disponible en: https://www.aprendemachinelearning.com/arbol-de-decision-en-python-clasificacion-y-prediccion/

[39] J. I. B. Arce, «La matriz de confusión y sus métricas – Inteligencia Artificial –», Juan Barrios. [En línea]. Disponible en: https://www.juanbarrios.com/la-matriz-de-confusion-y-sus-metricas/

[40] «sklearn.neighbors.KNeighborsClassifier», scikit-learn. [En línea]. Disponible en: https://scikit-learn/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html

[41] «Algoritmok-NearestNeighbor|AprendeMachineLearning». [En línea]. Disponible en: https://www.aprendemachinelearning.com/clasificar-con-k-nearest-neighbor-ejemplo-en-python/

[42] «IBM Documentation». [En línea]. Disponible en: https://www.ibm.com/docs/es/spss-modeler/saas?topic=deployment-overview

[43] «Categorical Features Encoding in Decision Trees and KNN». [En línea]. Disponible en: https://www.linkedin.com/pulse/categorical-features-encoding-decision-trees-knn-sravan-malla-

Juan Montes-Bustamante, an industrial electronics engineering student at the University of Sucre (Colombia), focuses his interests on developing predictive models and applying data mining to the health sector.