DS4A FINAL PROJECT

IDENTIFICATION OF DIABETES AND SYPHILIS DIAGNOSIS FROM MEDICAL NOTES USING NATURAL LANGUAGE PROCESSING AND MACHINE LEARNING

TEAM 24

Jorge Alfredo Acevedo Ramos

Charic Daniel Farinango Cuervo

Juan Manuel Gómez González

David Santiago Garcia Chicangana

Daniel Montes Agudelo

Cristian Eduardo Prieto Triana

Gerson Steven Ruiz Lozano

March – July

Application link: www.doctordisease.org

1. INTRODUCTION

IQVIA is an American multinational company serving the combined industries of health information technology and clinical research. Based on data, advanced analytics, and expert insight, the company helps its customers in the healthcare industry to make smarter decisions and unleash new opportunities.

As part of its mission, IQVIA seeks to improve medical attention and treatments for patients. To contribute to this, the company is interested in leveraging Electronic Health Records (EHR) for further analysis and research. However, using this data brings challenges like inconsistencies in the records, data fragmentation, and missing values, among others affecting its quality. Consequently, our capstone project is oriented to address data wrangling and cleaning, to then use it to accurately identify patients with specific pathologies or health conditions from text elements. Specifically, we aim to:

- Identify and correct inputting errors in the Electronic Health Records, such as wrong codes or diagnosis, leading to inappropriate treatments.
- Extract relevant information from notes using natural language processing.
- Train a machine learning model to accurately predict possible diabetes or syphilis diagnoses from EHR analysis.

Human errors when writing information in the EHR analysis could influence the final diagnosis of a person and this represents a very serious problem. Therefore, addressing the mentioned issues is relevant and will enhance the final outcome when improving the wellbeing of people.

1.1. Context

1.1.1. Syphilis

Syphilis is a type of sexually transmitted inffection (STI) caused by the bacteria Treponema Pallidum (LaFond & Lukehart, 2006, 29). Efforts to control this disease have been impeded by the difficulty to correctly diagnose its symptoms (Ricco & Westby, 2020, 91), as this condition can have multiple different manifestations and so it is known as the “great imitator” (Brown & Frank, 2003, 283).

1.1.1.1. Epidemiology

More than 5 million cases of syphilis are reported each year, the majority diagnosed in low to middle-income countries (Ricco & Westby, 2020, 91). Its incidence reduced drastically with the advent of penicillin in the 1940’s, but rose

again in the 1980's with the emergence of the Human Immunodeficiency Virus, commonly known as HIV (Brown & Frank, 2003, 283). There is a larger incidence of this disease in men, which account for 90% of the cases, and for which 82% of those cases are accounted by men who have sex with men (Ricco & Westby, 2020, 91).

1.1.1.2. Stages of Syphilis

The disease mainly manifests in 5 stages, known as Primary, Secondary, Early latent, Late latent, and Tertiary, each with its own set of characteristics (LaFond & Lukehart, 2006, 30-32) (Ricco & Westby, 2020, 92).

- Primary Syphilis:
  - Has an average duration range of 10 to 90 days, with the appearance of chancres at the location of the initial infection with Treponema Pallidum.
- Secondary Syphilis:
  - Has an approximate duration of 1 to 3 months, with a series of different conditions like headaches, rashes throughout the body, regional lymphadenopathy, and many more symptoms.
- Early Latent Syphilis:
  - Appears after the primary and secondary stages and up to a year of no symptoms, where there is a possibility of recurrence of secondary symptoms.
- Late latent Syphilis
  - Occurs after more than one year without presenting symptoms
- Tertiary syphilis
  - Occurs months to years after not presenting any symptoms, and comes with illnesses like cardiovascular syphilis, neurological complications and late neurosyphilis (seizures, ataxia, visual and hearing loss, etc).

1.1.1.3. Diagnosis

Clinical evidence, like the prevalence of chancres and sores can be used as a basis for suspicion of carrying the disease (LaFond & Lukehart, 2006, 30-32). Common diagnosing techniques are the use of serological tests and the identification of the reponemes using direct fluorescent antibody staining (DFA) or dark field microscopy (DFM) (Goh, 2005, 449).

DFA is particularly specific when there is a chancre present and as such it is widely used in diagnosing primary syphilis (Goh, 2005, 449). However, the accuracy of the

test is highly dependent on the experience of the operator, the quantity of treponemes in the lesion, and the existence of non-pathologic treponemes in the lesions (Brown & Frank, 2003, 284). On the other hand, the diagnosis of secondary syphilis is usually performed using DFM, lesions in the genital mucosa, skin papules and through serological tests (Goh, 2005, 449). Finally, latent syphilis is diagnosed with the help of serological tests combined with the lack of clinical evidence of syphilis (Goh, 2005, 449).

1.1.1.4. Treatment

Penicillin is considered by the majority of countries and the World Health Organization as the mainstay treatment for patients not allergic to it (Goh, 2005, 450) (Brown & Frank, 2003, 284), and necessary for pregnant women and patients with neurosyphilis even if allergic to penicillin, following a period of desensitization (Brown & Frank, 2003, 284).

1.1.2. Diabetes

Diabetes is a group of illnesses which affect how the body uses glucose, or blood sugar, which is a necessary molecule for humans as it is the main source of energy for the body. The cause for this disease depends on the type, which are divided into Type 1, Type 2, and gestational diabetes (Mayo Clinic, 2020). The World Health Organization mentions that there is a wide range of complications related to this condition:

"*Diabetes of all types can lead to complications in many parts of the body and can increase the overall risk of dying prematurely. Possible complications include heart attack, stroke, kidney failure, leg amputation, vision loss and nerve damage. In pregnancy, poorly controlled diabetes increases the risk of fetal death and other complications.* (World Health Organization, 2016, 6)".

It is estimated that 30% of the people afflicted by this condition are undiagnosed in the United States (Deshpande et al., 2008, 1254), and approximately 25% are undiagnosed in Colombia (Scully, 2012, S2).

1.1.2.1. Epidemiology

It was estimated in 2014 that around 422 million adults live with diabetes, an almost 4-fold increase in the amount of people afflicted by this condition compared with 1980 (World Health Organization, 2016, 6), and has doubled in the past 20 years (Zimmet et al., 2014, 1).

Diabetes mainly affects low and middle income countries, representing more than 80% of the deaths related to this disease (Scully, 2012, S2).

1.1.2.2. Types of Diabetes

Two main types of diabetes are normally recognized, called Type 1 and 2, with gestational diabetes grouped with the main 2 types (Diabetes Canada Clinical Practice Guidelines Expert Committee et al., 2018, S10). Additional types caused by other miscellaneous causes like disease of the pancreas, drugs and/or chemicals tend to be grouped together on its own category as well (Deshpande et al., 2008, 1255).

- Type 1:
  - Caused by the destruction of the beta cells in the pancreas, mainly due to an autoimmune process, and causing an insulin deficiency (Deshpande et al., 2008, 1255).
  - Most commonly present in people younger than 25, and represents approximately 5 to 10% of all types of diabetes (Diabetes Canada Clinical Practice Guidelines Expert Committee et al., 2018, S10) (Burahmah et al., 2022).
  - Ketoacidosis is common in this type of diabetes (Diabetes Canada Clinical Practice Guidelines Expert Committee et al., 2018, S10).
- Type 2:
  - Occurs when the pancreas is not capable of producing enough insulin, or the body is not able to use appropriately the insulin that the pancreas produces (Diabetes Canada Clinical Practice Guidelines Expert Committee, 2018, S10).
  - Accounts for roughly 85% of the cases of diabetes (Forouhi & Wareham, 2010, 602).
- Gestational and miscellaneous cases:
  - Occurs when a person develops a glucose intolerance during pregnancy, and tends to resolve itself afterwards (Diabetes Canada Clinical Practice Guidelines Expert Committee et al., 2018, S10).
  - Account for 1% to 2% of the cases of diabetes (Harris, 1995, 1).

1.1.2.3. Diagnosis

Diagnosis is done with laboratory testing through venous samples. There are three main blood tests that can be performed, which are: fasting plasma glucose (FPG), 2-hour plasma glucose (2hPG) in an oral glucose tolerance test (OGTT), and glycated hemoglobin (A1C) (Diabetes Canada Clinical Practice Guidelines Expert Committee et al., 2018, S12). FPG is the fastest and easiest, but presents a high level of variability from day to day measurements, while also requiring fasting. A1C has a lower level of variability at a higher cost and is misleading in different medical

conditions like anemia and renal disease (Diabetes Canada Clinical Practice Guidelines Expert Committee et al., 2018, S12).

Finally, 2hPG in OGTT is considered to be unpalatable and is somewhat costly, while, as the name implies, taking 2 hours of the patient's time (Diabetes Canada Clinical Practice Guidelines Expert Committee et al., 2018, S12).

To differentiate between the two main types of diabetes, Burahman et al recommend that islet antibodies should be tested (Burahmah et al., 2022). If they are positive, it means that there is a high possibility of the disease being type 1. On the contrary, if the antibodies test returns negative, a serum C-peptide test while not fasting is performed, for which a lower value than a 300pmol per liters (L) would possibly indicate Type 1 as well. In the case that the C-peptide is higher than 600pmol/L, Type 2 diabetes should be considered. C-peptide values in the middle of both thresholds indicate an uncertain classification, and periodic retesting should be performed to correctly diagnose.

1.1.2.4. Treatment

There is currently no effective treatment for curing diabetes, however in recent years there has been some developments in preventing it, as well as the existence of techniques for managing it (Nathan, 2015, 1054).

Prevention involves immune manipulation for patients of type 1 diabetes, and changes in diet and exercise routines for type 2 diabetes (Nathan, 2015, 1054).

For managing diabetes of type 1, insulin replacement is usually performed, trying to maintain the patient's optimal level with the use of daily injections or specialized insulin pumps (Diabetes Canada Clinical Practice Guidelines Expert Committee et al., 2018, S81 - S82).

Managing type 2 diabetes involves changes in diet and an increase in the amount of physical activity that the patient has, but there has been a change in recent years to start a pharmacologic-based management practices in type 2 diabetic adults (Diabetes Canada Clinical Practice Guidelines Expert Committee et al., 2020, 576).

1.2. Justification

Syphilis is a disease which can present a multitude of different symptoms, for which the main indicator (chancres) is only present during the onset of the condition, i.e. Primary syphilis. Each year, 5 million people are diagnosed with this affliction, and its diagnosis can be difficult and can depend on a multitude of factors

such as the microbiologist's experience when analyzing a patient's sample to the quantity of treponema bacteria present in the lesion during the sampling process.

Similarly, diabetes is an incurable disease affecting more than 400 million people worldwide, where 25% of the Colombians who are afflicted by it may go undiagnosed until it's too late. Even though there is no cure for diabetes, it is still a disease which can be managed through the use of clinical treatments as well as diet and exercise, and for which the earlier a diagnosis is done, the better prognosis the patient will have.

Leveraging the use of medical notes taken by multiple examiners, combined with the different laboratory testing performed on a patient, could help raise a flag in order for another medical expert to give their informed second opinion to a prior diagnosis and prevent a further onset of these diseases.

2. DATASETS DESCRIPTION

A database with three (3) tables regarding anonymized electronic health records of diabetes and syphilis collected by different institutions located in many regions of Colombia was provided to fulfill the objectives of our project. These datasets are described below. It is relevant to mention that all the information contained in the tables and variables' names are in Spanish.

2.1. Sociodemographic information (sociodemografico.csv)

Table 1 contains sociodemographic information about patients included in the study. It contains 9,306 rows, i.e., 9,306 patients, and 7 columns:

- IDRecord: patient's ID number (*type:* INTEGER).
- Edad (Age): patient's age in years (*type:* INTEGER).
- Genero (Gender): patient's gender (*type:* STRING – 'Mujer' or 'Hombre').
- GrupoEtnico(Ethnic Group): patient's ethnic group (*type:* STRING – 'Mestizo', 'Negro, Mulato, Afrocolombiano o Afro descendiente', 'Blanco', 'Indígena', 'Palenquero de San Basilio', or 'Ninguno de los anteriores').
- AreaResidencial (Residential Area): patient's type of geographic area of residence (*type:* STRING – 'Zona Rural' or 'Zona Urbana').

- EstadoCivil (Marital Status): patient's marital status (*type:* STRING – 'Separado', 'Casado', 'Soltero', 'No reportado', 'Viudo/a', 'Unión libre', 'Desconocido', or 'Divorciado')
- TSangre (Blood type): patient's blood type (*type:* STRING – 'O+', 'A+', 'B+', 'O-', 'A-', 'B-', 'AB+', or 'AB-').

IDRecord	Edad	Genero	GrupoEtnico	AreaResidencial	EstadoCivil	TSangre
123011	75	Mujer	Mestizo	Zona Urbana	Viudo/a	NaN
104432	25	Hombre	Negro, Mulato, Afrocolombiano o Afro descendiente	Zona Urbana	Soltero	O+
54634	40	Hombre	Mestizo	Zona Urbana	Soltero	NaN
42004	30	Hombre	Mestizo	Zona Rural	Unión libre	A+
87402	30	Hombre	Mestizo	Zona Urbana	Desconocido	O+

Table 1: Sample of 5 rows from sociodemographic table

2.2. Laboratory tests information (laboratorios.csv)

Table containing information regarding laboratory tests that have been performed in the health institution to the patients in the study from September 12, 2001, to March 12, 2022. For one patient, there may be more than one test. Therefore, this table is much larger. It contains 189,643 entries and the following 5 columns[1]:

- IDRecord: patient's ID number (*type:* INTEGER).
- Codigo (Code): ID number of the test performed (*type:* INTEGER).
- Nombre (Name): name of the test performed (*type:* STRING).
- Fecha (Date): date when the test was performed (type: DATE – DD/MM/YYYY hh:mm).
- Valor (Value): value obtained as a result for the test performed (type: FLOAT).

[1]The initial type of the data received by Team 24 was of type "object", however the Schema shared by IQVIA indicates that these should be the expected types of the variables, and thus it means it is imperative to do data cleaning of the dataset.

2.3. Medical notes information (notas.csv)

Table containing notes from medical appointments of patients included in the study. This is the main table as it contains the Electronic Health Records (EHR). It has 140,227 entries and 5 columns:

- IDRecord: patient's ID number (*type:* INTEGER).
- Codigo (Code): ICD-10 (International Statistical Classification of Diseases and Related Health Problems - 10th Revision) code associated with the diagnosis (*type:* STRING).
- Nombre (Name): Name of the diagnosis.
- Tipo (Type): Type of diagnosis (*type:* STRING – 'Confirmado Nuevo', 'Confirmado Repetido', 'Impresión Diagnóstica').
- Plan (Plan): Medical record registered during the appointment (*type:* STRING).

3. EXPLORATORY DATA ANALYSIS (EDA)

Our methodology starts by analyzing the dataset handed by IQVIA, seeking to understand all the available variables and their relations to draw insights on the data that could lead us to design our solution. We started by analyzing the information contained in each table individually, identifying missing values, inconsistencies, and possible solutions for these issues, to then join the tables and start analyzing the relationships among variables and their potential to fulfill the objectives of our project.

3.1. Sociodemographic information (sociodemografico.csv)

As previously mentioned, the sociodemographic table contains information about 9306 patients. Without considering IDRecord, that is simply an identifier and does not provide valuable information, there are 6 variables:

- Edad (Age): It is the only numerical feature in this table. When joined with other tables, age might give us information on to which population groups can develop one of the conditions we are trying to look for.
- Genero (Gender): This feature may offer relevant information to understand diagnosis and maybe identify errors, as health diseases may develop and

evolve differently depending on gender. In fact, it is believed that males develop diabetes easier than females.

- GrupoEtnico (Ethnic group): Different diets and practices that each ethnic group has can facilitate developing diabetes or syphilis, so this can be useful.
- AreaResidencial (Residential area type): It may be relevant to check if people living in urban areas are more propense to have one of the diagnoses than people living in rural areas, or vice versa.
- EstadoCivil (Marital status): Patients that don't have a stable relationship are exposed to higher levels of Sexually Transmitted Diseases (STDs), so this could be an indicator for syphilis.
- TSangre (Blood Type): Through different studies, it has been shown that there exists a correlation between blood type and diabetes, so this is also relevant information.

All the above are hypotheses coming from the context of the data and we would need to perform further analysis, merging the information with other tables, to validate them. Before that, we analyze each variable as follows:

3.1.1. Numerical features analysis

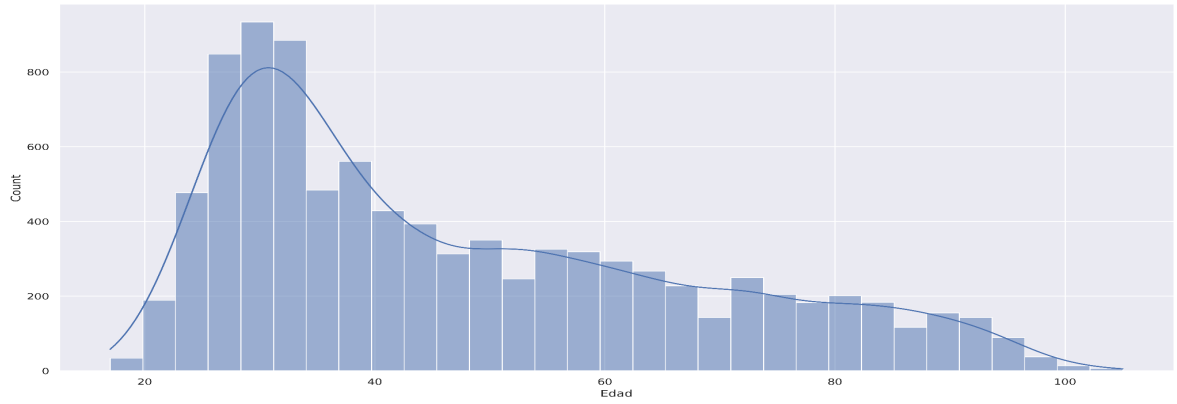


Figure 1. Histogram of patients' age.

'Edad' has an int-64 data type and there are no missing values i.e., the age of each patient is available in the dataset. The mean age of patients in the study is 47.50 years and the median is 41 years, whereas the standard deviation is 20.12. The youngest patient is 17 years, whereas the oldest patient registered is 105 years

old. The variable has a right-skewed distribution, with most patients between 25 and 35 years old.

3.1.2. Categorical features analysis

'Genero', 'GrupoEtnico', 'AreaResidencial', 'EstadoCivil', and 'TSangre' variables are all of type object. In addition to the categories for each variable, previously mentioned in Section 2, we find that 'EstadoCivil' and 'TSangre' have missing values for some patients. We try to dig into this in Section 3.1.3.

Figure 2 presents bar-plots reporting the count (and percentage) of patients for each category in the different variables. We can evidence that variables are highly imbalanced. Most patients in the study are males, identify themselves with the mestizo ethnic group, live in urban areas, are single, and have O+ as blood type. This could limit our prediction abilities in the future, and suggests it could be necessary to implement techniques to handle imbalanced data.

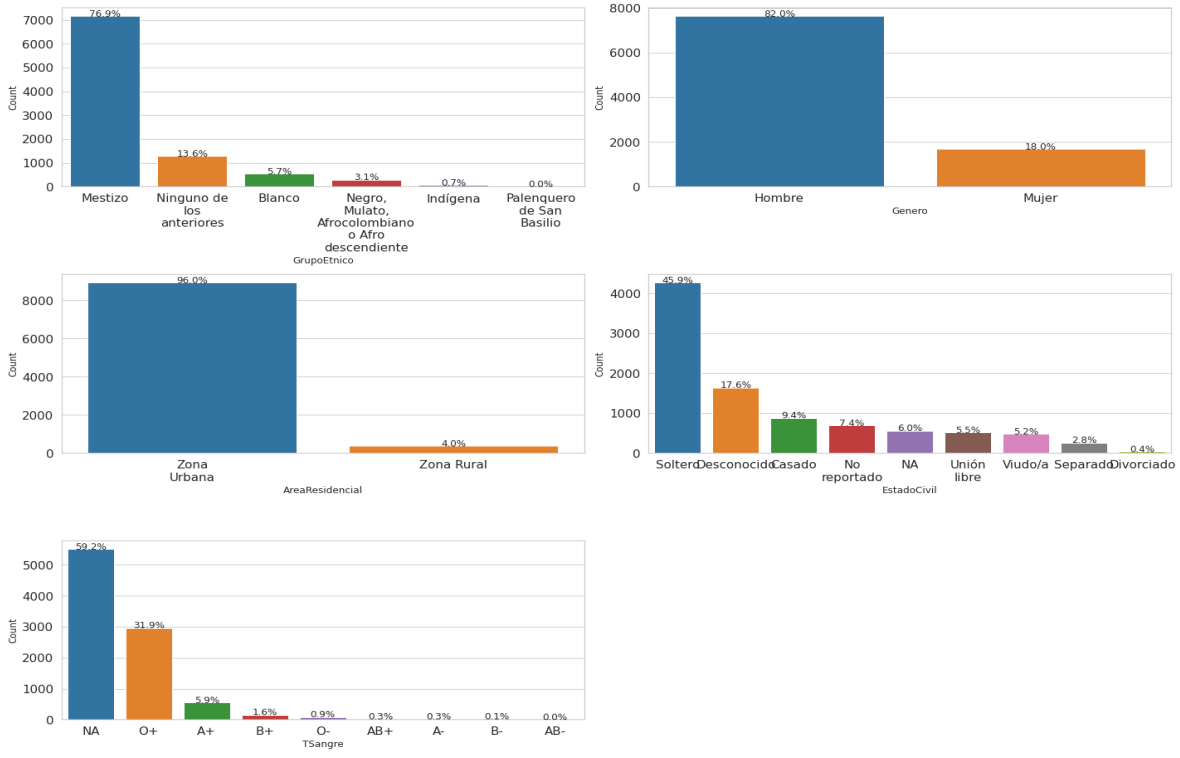


Figure 2. Bar plots for categorical features.

Moreover, we can see from the bar plots for 'EstadoCivil' and 'TSangre' that data is missing for an important proportion of the sample. In fact, blood type is missing for almost 60% of the patients (5505 of the 9306) and, although blood type probably

has a relationship with diabetes as previously stated, it may be difficult to use it in our analysis. On the other hand, there are 555 missing values for 'EstadoCivil'. However, there are two categories labeled as 'Desconocido' (Unknown) and 'NoReportado' (Not reported), that could also be understood as missing data. In total, there are 2327 (26.59%) patients for whom we do not have information regarding their marital status.

3.1.3. Missing values analysis

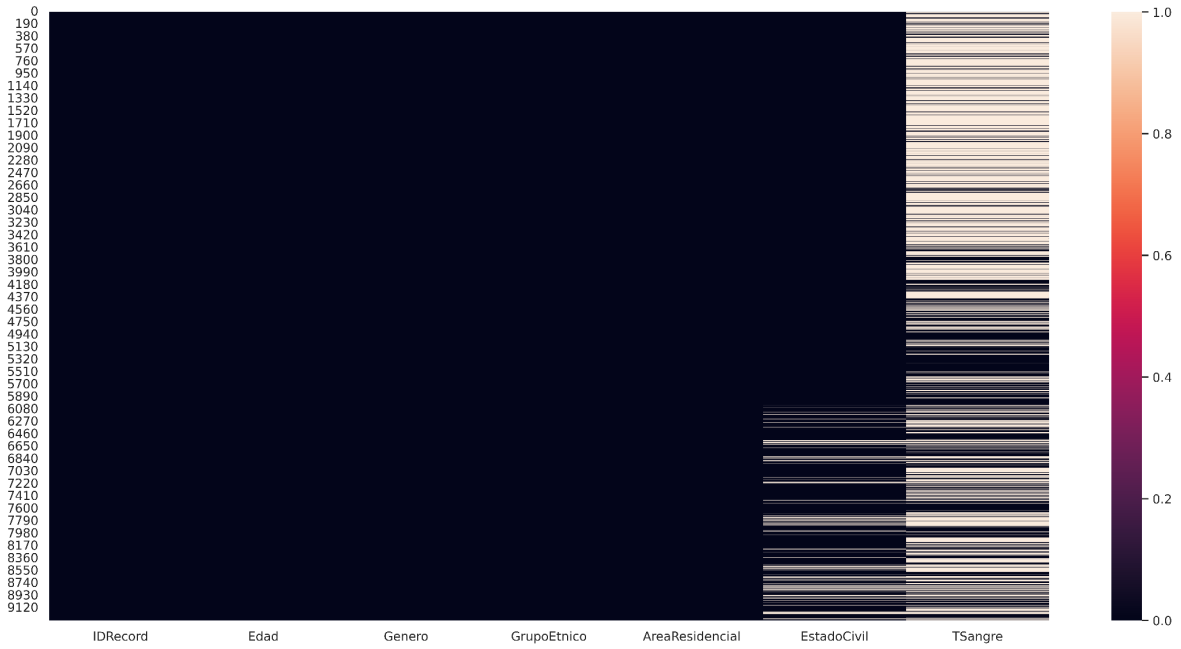


Figure 3. Missing values Heat map for Sociodemografico.

We proceed to analyze the relationship of 'EstadoCivil' and 'TSangre' with the remaining variables in the table, seeking to identify the type of missing values we are dealing with, i.e., missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR).

We start this analysis by making percent stacked bar plots for 'EstadoCivil' against each of the remaining variables, excepting 'Edad', the numerical feature, for which we make a box plot (see Figure 4). By this, we aimed to evidence if missing values are related to other variables, e.g., marital status could be unknown for young people. However, we fail to find relationships among variables and actually check that, at first sight, there are no inconsistencies and variable behaviors are expected, such as having a great proportion of widowers (or widows) for patients older than 75 years.

From the analysis, we can't either discriminate between 'Desconocido', 'No reportado', and missing value. Therefore, we decided to merge them together into one distinct category: 'NA'.

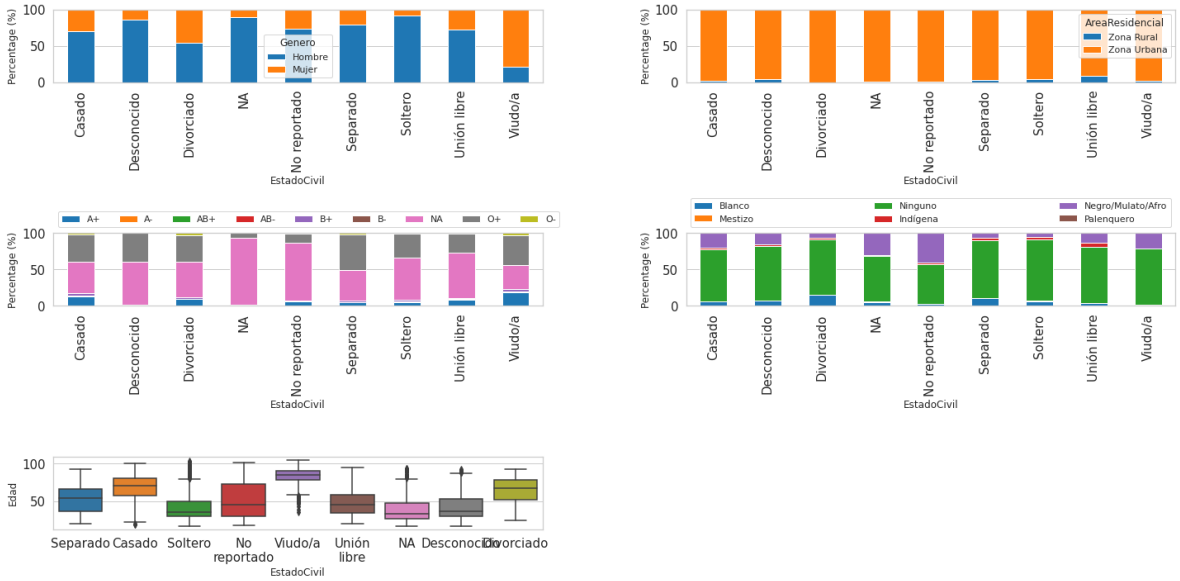


Figure 4. Relationships of marital status with the remaining variables.

A similar analysis was carried out for 'TSangre', also failing to find relationships between missing values and other variables in the table. As can be seen from Figure 5, there are no indications that variables behave distinctly when blood type is missing.

Both cases seem to be missing not at random (MNAR), i.e. the value of the variable that's missing is related to the reason it's missing. It may be related to people not comfortable giving the information or not remembering it, especially for blood type, as not everybody is aware of this information. Given that we did not find relationships with other variables and imputing with the mode would be highly inaccurate given the context of the data, in addition to introducing bias, we then decided to fill missing values with 'NA's for blood type.

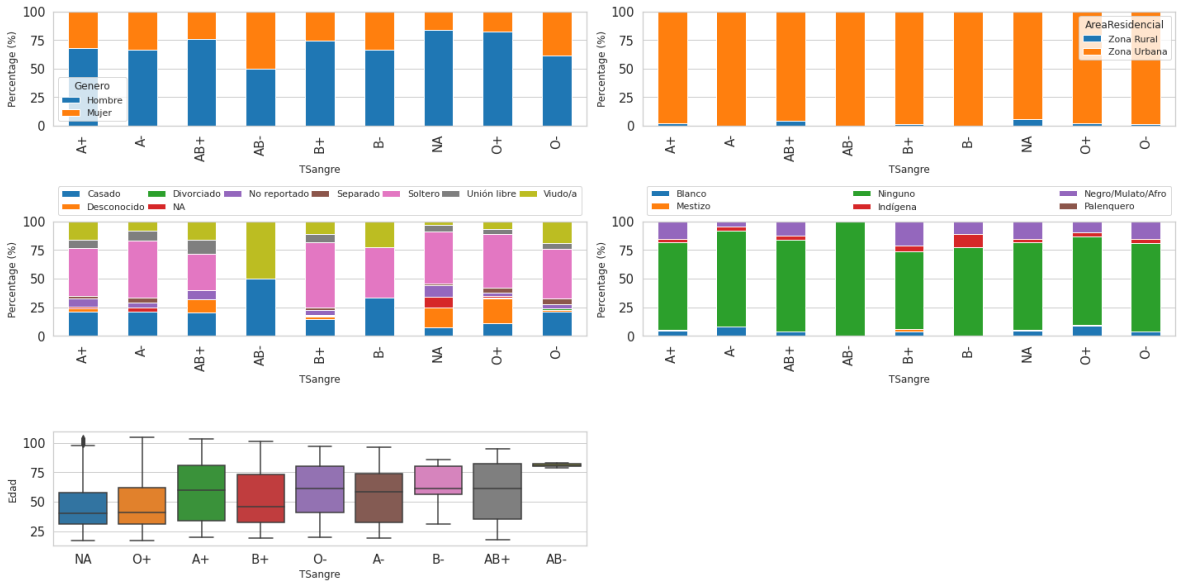


Figure 5. Relationships of blood type with the remaining variables.

3.2. Laboratory tests information (laboratorios.csv)

As mentioned before this dataset contains the information about the test performed and the value obtained in that analysis for each patient and type of test. The data in total contains 189,643 entries with 5 columns.

- IDRecord: patient's ID number (*type:* INTEGER).
- Codigo(Code): ID number of the test performed (*type:* INTEGER).
- Nombre (Name): name of the test performed (*type:* STRING).
- Fecha (Date): date when the test was performed (type: DATE – DD/MM/YYYY hh:mm).
- Valor (Value): value obtained as a result for the test performed (type: FLOAT).

3.2.1. Numerical features analysis

In the numerical values the dataset contains the attribute named IDRecord, that corresponds to the patient's ID number. Starting with the patient with the IDRecord 5 and the final IDRecord with 206307.

Another numerical value is the Date. In this case, we have that the laboratory tests have an initial date starting at September 12, 2001, and the final date is March 12, 2022. The date when most tests were taken was October 14, 2020.

If we look closely at the Date and glance at the days of the week in which the tests were taken we can see that these laboratory tests are made the seven days of the week, being saturday the day with most tests taken.

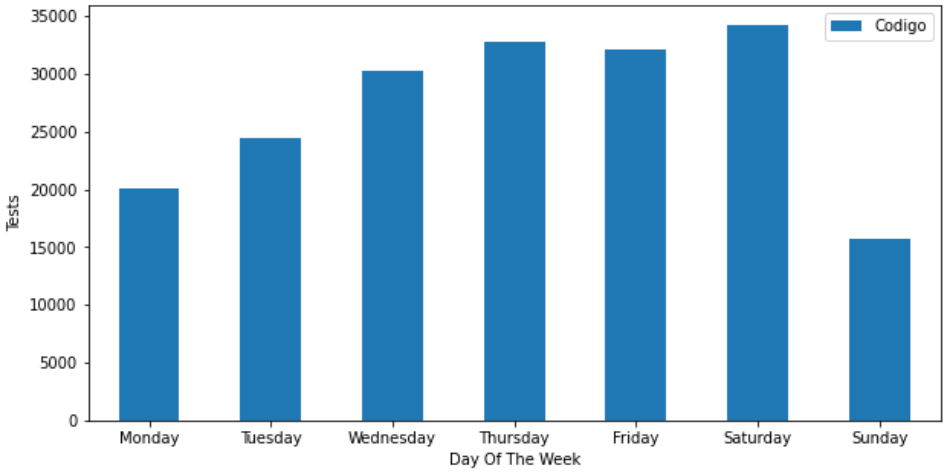


Figure 6. Tests by Day of the Week

Watching the hours that are in the dataset we can see that the hour with most tests taken is 0 hours, having another peak in the 13 hours. However, the difference is great as the number of tests at 0 hours is greater than 80,000, whereas at 13 the number is near 10,000. It is possible that most tests that apparently were taken at 0 hours correspond actually to tests where there was no information recorded regarding the exact time, but only the day.

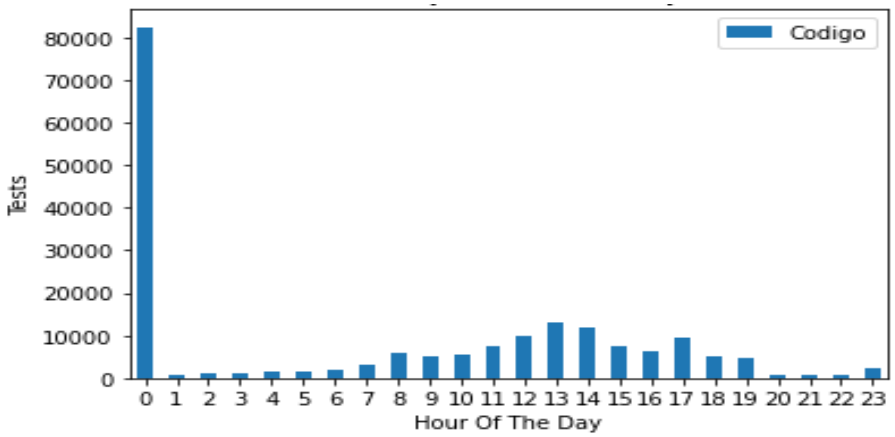


Figure 7. Tests by Hour of the day

One interesting idea was to identify the Period of time between each test. For that, the dataset was grouped by IDRecord to know all the tests associated with each user. The following table presents an example of a patient with their exams performed.

	IDRecord	Codigo	Nombre	Fecha	Valor
123318	42456	903815	COLESTEROL DE ALTA DENSIDAD (HDL)	2018-12-18	39.0 mg/dl
129521	42456	903816	COLESTEROL DE BAJA DENSIDAD (LDL) ENZIMÁTICO	2018-12-18	143.0 mg/dl
139536	42456	903818	COLESTEROL TOTAL	2018-12-18	244.6 mg/dl
160087	42456	906219	Hepatitis A, ANTICUERPOS TOTALES (ANTI HVA)	2018-12-18	Mayor 50 UI/L
177584	42456	903868	TRIGLICÉRIDOS	2018-12-18	381.0 mg/dl

Table 2: Sample of laboratories performed to patient 42456

The results of a rapid assessment revealed some interesting aspects:

- Several tests were performed on the same day. Besides, some users have only exams recorded for one day while others present groups of exams performed on different dates.
- On average, there are 1891 patients whose time between tests was zero days (table below). In addition, the right-hand figure presents the longest time period between tests recorded in the dataset, particularly the last record that means there were 3206 days of difference between the date of the tests.

	days_avg	count
618	0.0	1891
616	1.0	48
617	2.0	88
614	3.0	39
463	4.0	10
...	...	...
69	1340.0	1
64	1360.0	1
171	1416.0	1
175	1776.0	1
118	3206.0	1

Table 3: Count of cases for each average days of difference between tests

The following graph can explain better the information presented above. As we can see, there is a concentration of values between 0 and 500 in the axis X with a High peak caused by the count of records with zero days of difference.

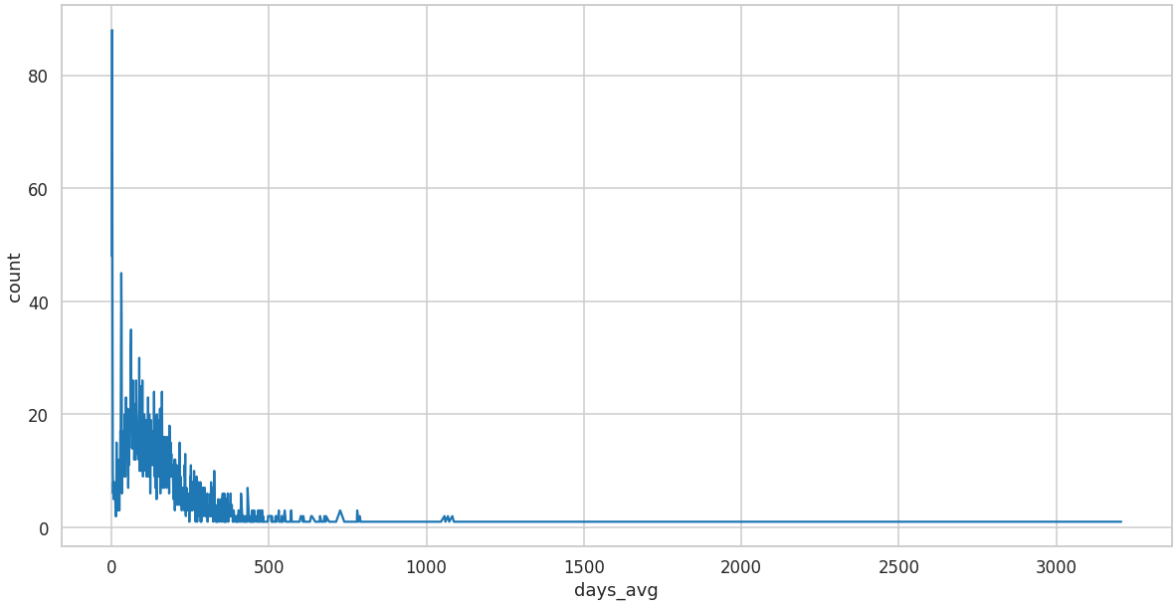


Figure 8. Count of records by average day difference

It is important to mention that the information below does not distinguish between the type of test. Taking this into account, a second analysis was performed considering the type.

	test	avg_days
0	CONSULTA DE CONTROL O DE SEGUIMIENTO POR ESPE...	0.0
110	LINFOCITOS T CD8 POR INMUNOFLUORESCENCIA	0.0
105	IRRIGACION O ENEMA TRANSANAL SOD	0.0
104	Histoplasma capsulatum, ANTICUERPOS POR EIA	0.0
103	Herpes II, ANTICUERPOS Ig M	0.0
...	...	...
45	DESHIDROGENASA LÁCTICA (LDH)	375.0
82	HORMONA ESTIMULANTE DEL TIROIDES (TSH) NEONATAL	397.0
59	FOSFATASA ÁCIDA	493.0
62	GAMMA GLUTAMIL TRANSFERASA (GGT)	494.0
173	Virus de Inmunodeficiencia Humana GENOTIPO	865.0

178 rows × 2 columns

Table 4: Average day of difference for each laboratory test

58 of the tests were performed with zero days of differences between each other. 38 presents a difference between 1 and 6 days (See the graph below).

Figure 9. Test with an average days of difference between 1 and 6 days

9 tests were performed with a difference between them of 7 and 14 days.

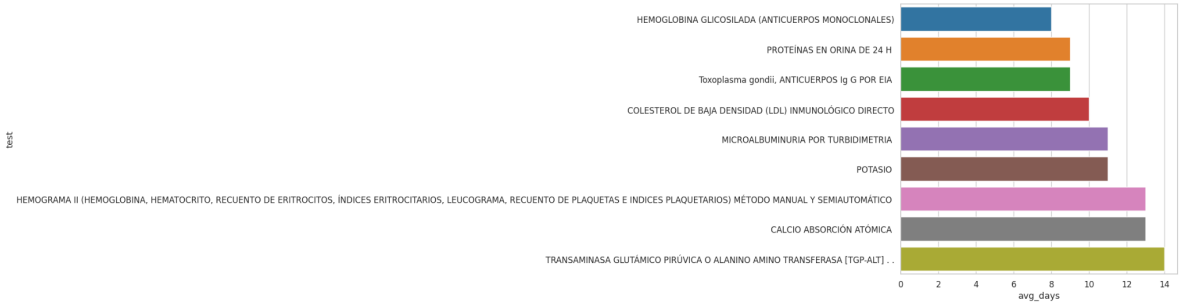


Figure 10. Tests with an average days of difference between 7 and 14 days

Finally, we can see in the following plot that 73 tests were performed with a difference between them of 15 or more days.

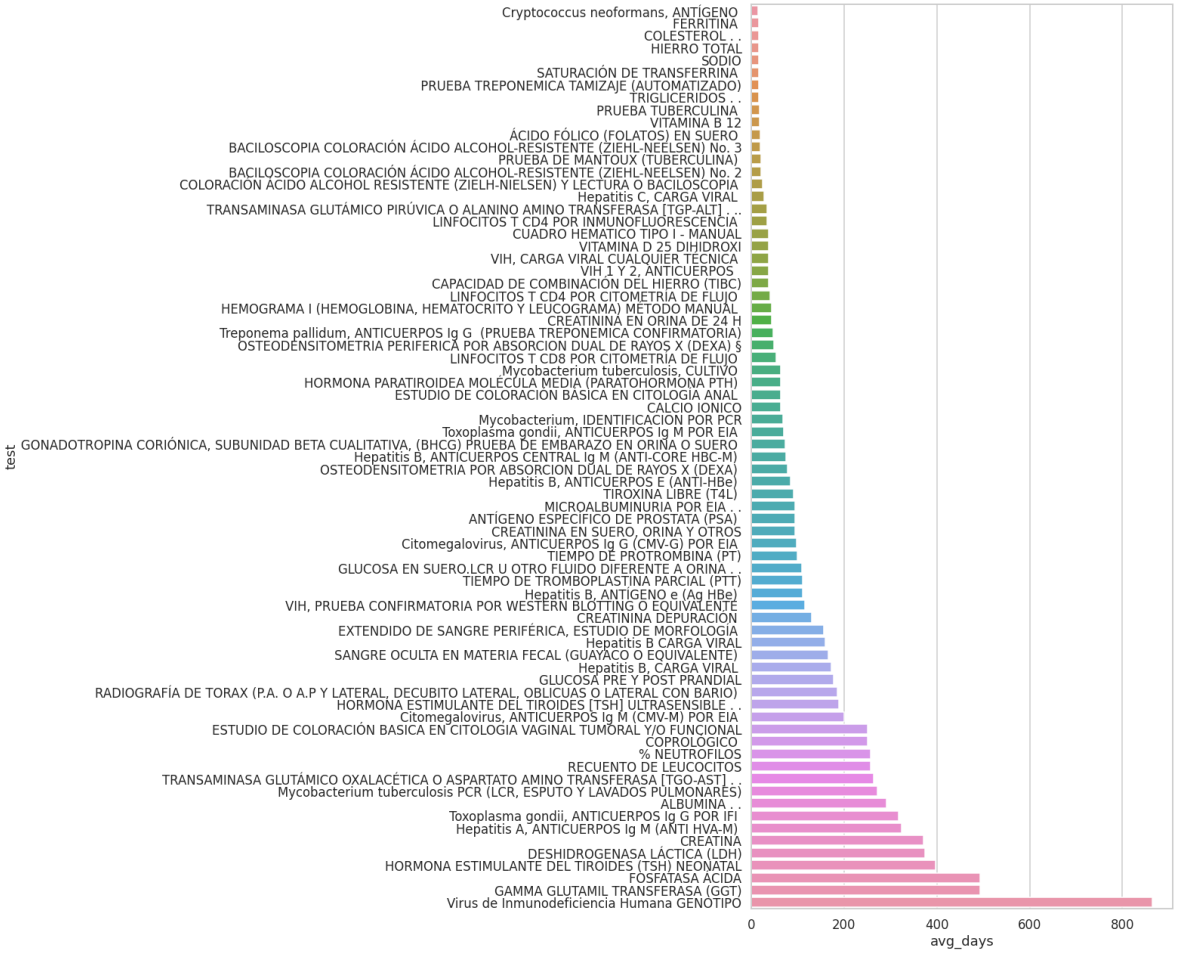


Figure 11. Tests with an average days of difference of 15 or more

3.2.2. Categorical features analysis

This dataset contains a categorical attribute called “Codigo” that contains the identificator of each laboratory test. We have a total of 180 alphanumeric codes. A general review of these values could suggest the need for applying cleaning tasks for some cases where values are not appropriate, e.g. '898002\xa0\xa0 ”.

A frequency analysis of each code indicates that there is a small group of tests that are more employed than others. The 20% of the high-frequency tests employed are shown in Figure 9.

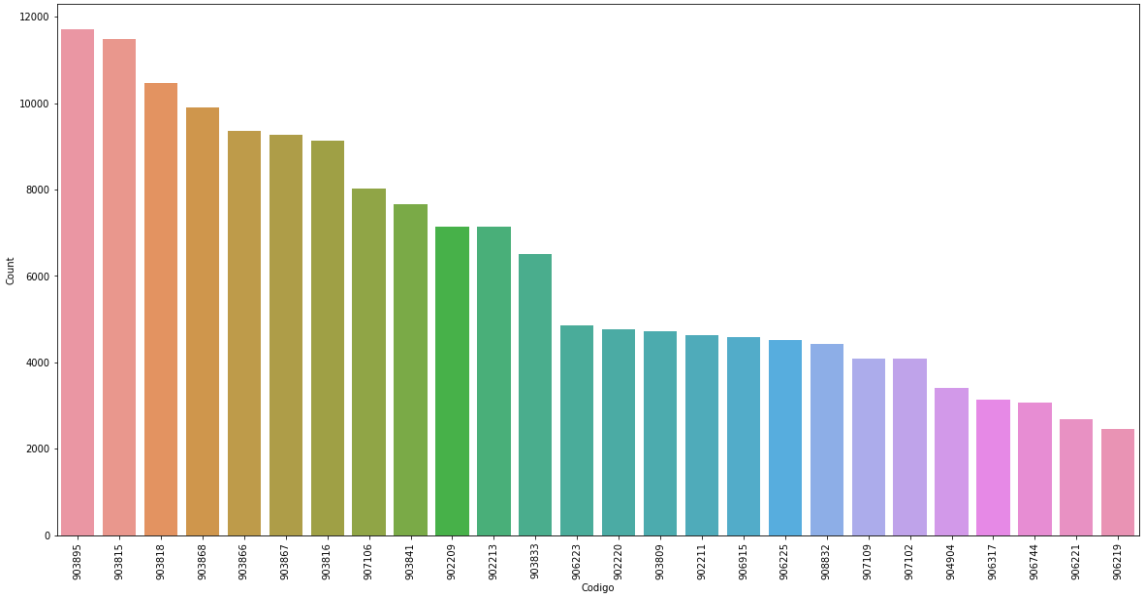


Figure 12. Laboratory codes by frequency

In relation to the nominal variable “Nombre” (name), the values present a different range of names for the laboratory tests. Almost all the names are in upper case and only a few ones in lower case. Table 5, presented below, shows some of the values of this attribute with their frequency and percentage respectively.

	quantity	percentage
CREATININA	11841	6.243837
COLESTEROL DE ALTA DENSIDAD (HDL)	11492	6.059807
COLESTEROL TOTAL	10462	5.516681
TRIGLICÉRIDOS	9905	5.222972
TRANSAMINASA GLUTÁMICOPIRÚVICA O ALANINO AMINO TRANSFERASA (TGP-ALT)	9353	4.931898
...	...	...
VIH 1 Y 2 ANTICUERPOS SEGUNDA PRUEBA	1	0.000527
prueba de Tropismo vírico	1	0.000527
XEROMAMOGRAFIA O MAMOGRAFIA, BILATERAL	1	0.000527
HEPATITIS A MAYORES DE 19 A	1	0.000527
HTLV-I Y II, ANTICUERPOS (ANTI HTLV-I) TOTALES CONFIRMATIVO	1	0.000527

Table 5. Values of Nombre variable

3.2.3. Missing values analysis

Value, which presents the value obtained as a result for the test performed, is the only variable with null entries.

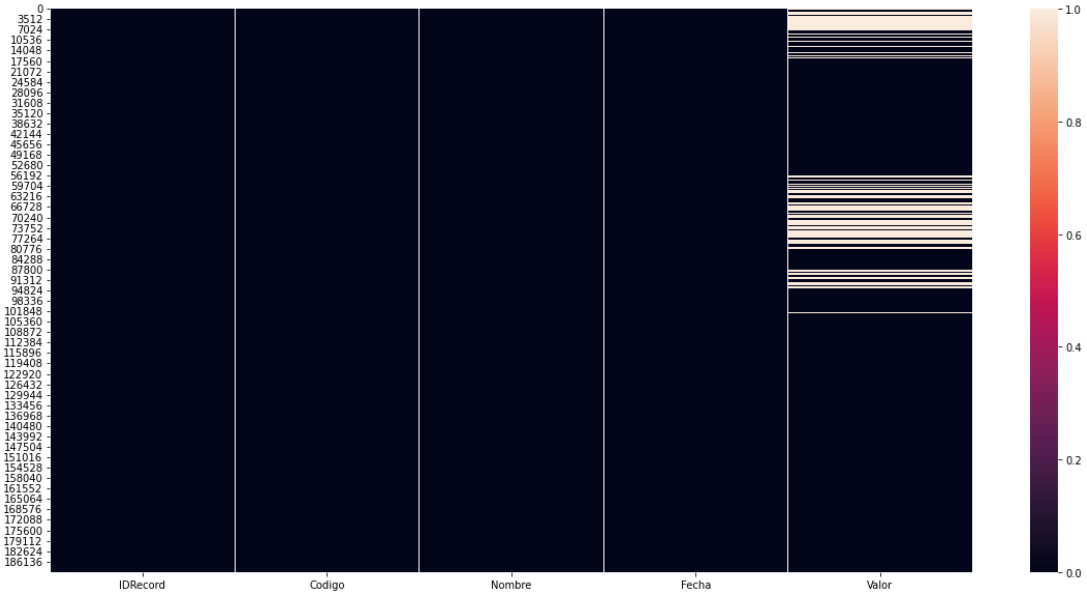


Figure 13. Missing values heat map for laboratories

We have a total of 27,889 entries with Value column null. That represents 14.07% of all the entries.

3.2.4. Duplicate values analysis

One of the assumptions was that each laboratory test name was associated with a unique code. However, the results indicate there are 178 unique names and 180 codes. There is a difference of two.

To find the cause of this difference, a table with the columns “Nombre” and “Codigo” was created, for which a pandas function was later applied in order to identify duplicates in the records. The result of this process was that repeated values were found, and can be seen in Table 6.

	Nombre	Codigo	Count
51	CREATININA	903895..	129
150	TRANSAMINASA GLUTÁMICO OXALACÉTICA O ASPARTATO...	60242	1

Table 6. Values duplicated

As we can see, there are two tests that are present in the column more than once. With this information, a search was executed to identify the values for further cleaning process. An interesting aspect about the duplicates was these test names have two different codes, which can be seen in Table 7 below.

	Nombre	Codigo	Count
139	TRANSAMINASA GLUTÁMICO OXALACÉTICA O ASPARTATO...	60241	1
150	TRANSAMINASA GLUTÁMICO OXALACÉTICA O ASPARTATO...	60242	1
0	CREATININA	903895	11712
51	CREATININA	903895..	129

Table 7. Duplicated names

One way to handle this is to edit the codes to be the same, after consulting with the company to confirm that they are indeed the same tests even though their codes are different.

3.3. Medical notes information (notas.csv)

As it was mentioned in section 2, this table contains notes from medical appointments of patients, i.e., it is where the EHR notes are saved, and hence requires a deeper natural language analysis. The table has 140,227 entries and 5 variables. IDRecord identifies the Record of a patient and allows to have a relationship between the tables. The other 4 features are the following:

- Codigo (Code): It allows to standardize the type of diagnosis that each patient has (in accordance with the ICD-10). Ultimately, it can serve as labels for a predictive model.
- Nombre (Name): Similar to Codigo, it allows us to comprehend the diagnosis of each person. Ultimately, it can serve as labels for a predictive model.
- Tipo (Type): This feature indicates the type of diagnosis, which allows us to know the stage in which the disease is. This might be helpful when looking for possible trends related to the passage from one stage to another, or the relation between a stage and other variables.
- Plan (Plan): The medical notes registered during appointments contain raw information about the state of the person. It might have data that can help to reinforce or predict a diagnosis. Nevertheless, as this information is entered

manually by the health professional, we expect some errors. Thus, it has to be inspected and cleaned.

All the above are ideas coming from the context of the data and we would need to perform further analysis, merging the information with other tables, to validate them. Before that, we analyze each variable as follows:

3.3.1. Numerical features analysis

The only feature that should be numeric according to the dataset description provided by IQVIA is IDRecord. After inspecting its type, it is shown as an object dtype, which indicates that it has non-numeric values in it. After searching for non-numerical values in this column we obtained the following table.

	IDRecord	Código	Nombre	Tipo	Plan
76968	SPECIFICADA	Confirmado Repetido	A/ PACIENTE CON DX DE INFECCION POR VIH HACE C...	NaN	NaN
85091	TE, NO ESPECIFICADA COMO PRECOZ O TARDIA	Confirmado Repetido	CONTINUA CON IGUAL MANEJO TARV. SE SOCLITA RP...	NaN	NaN
86209	ado Repetido	EMPEZÓ TAR EN DIC/15 SE REFORMULA IGUAL TAR ...	NaN	NaN	NaN

Table 8. Non-numerical IDRecords in table Notas

It can be seen that IDRecord has information from other features. We removed those elements because without IDRecord we can't relate the data to a specific patient, and thus it proves to be non-useful for us. Finally, we transform that column to numeric type.

3.3.2. Categorical features analysis

The three categorical features in this table are ‘Code’, ‘Name’ and ‘Type’. First, feature Code is described by the ICD-10 code system, where each disease is characterized by a unique number. In our case we have, in alphabetical order:

- A510: Primary genital syphilis
- A511: Primary anal syphilis
- A514: Other secondary syphilis
- A529: Late syphilis, unspecified
- A530: Latent syphilis, unspecified as early or late
- A539: Syphilis, unspecified

- E109: Type 1 diabetes mellitus
- E119: Type 2 diabetes mellitus
- E149: Unspecified diabetes mellitus

After we grouped the information by Code, we obtained one more unique code with one occurrence in the table: 'SIFILIS LATENTE, NO ESPECIFICADA COMO PRECOZ O TARDIA'. So, we had the different codes for the diseases plus what it seems to be an instance of feature Name, which shouldn't be here.

	IDRecord	Código	Nombre	Tipo	Plan
42708	30.0	SIFILIS LATENTE, NO ESPECIFICADA COMO PRECOZ O...	Confirmado Repetido	*Control por Trabajo Social según frecuencia y...	NaN

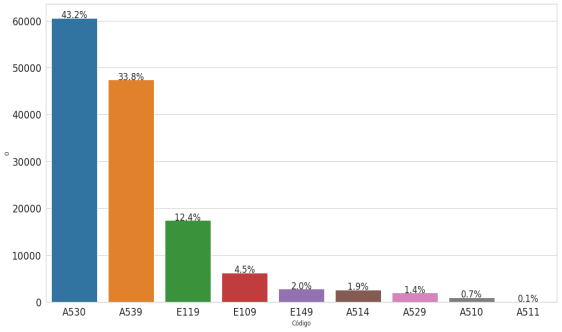
Table 9. Troubling record before applied solution

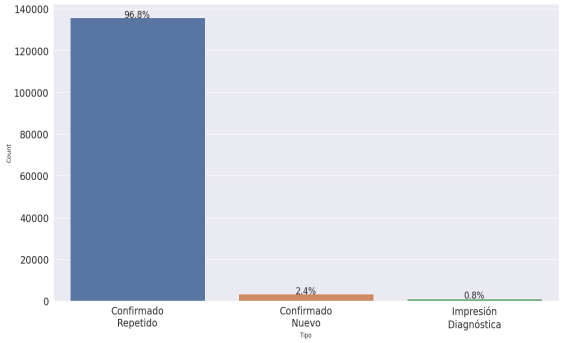
	IDRecord	Código	Nombre	Tipo	Plan
42708	30.0	A530	SIFILIS LATENTE, NO ESPECIFICADA COMO PRECOZ O...	Confirmado Repetido	*Control por Trabajo Social según frecuencia y...

Table 10. Troubling record after applied solution

Tables 9 and 10 show the specific row before and after we handle the issue. It can be observed that, using the ICD-10 code system and the diagnosis name, we were able to input the correct code and move all the columns to the right starting from Code.

Other than Code, Name and Type do not present issues on the data. Therefore, we proceed to plot their frequencies.





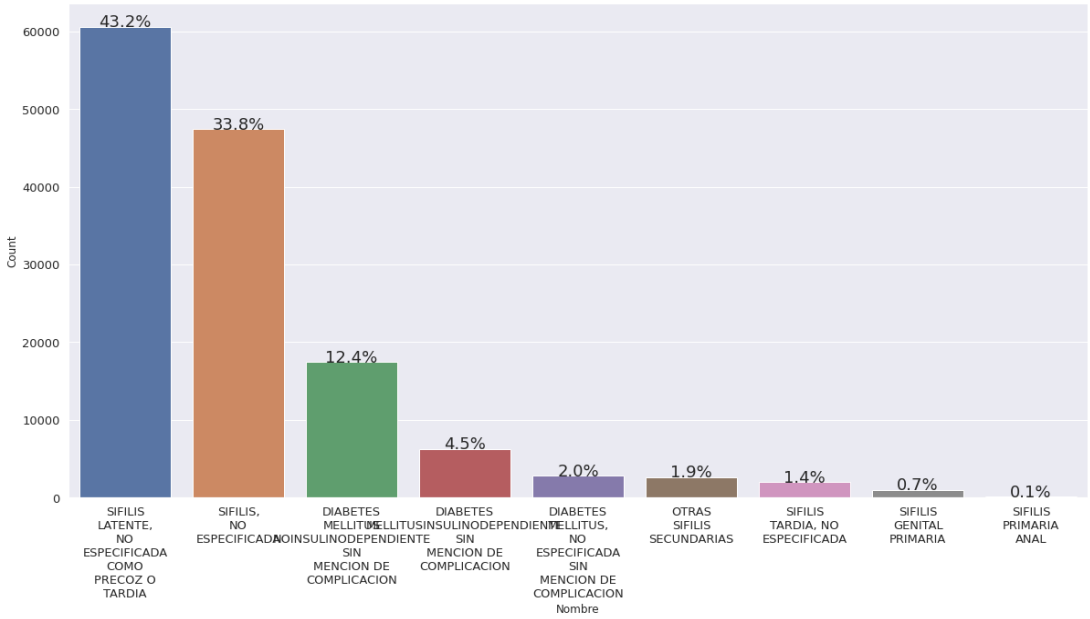


Figure 14. Frequencies for features Code, Type and Name

As it can be observed, Code and Name frequencies correspond to each other. This helps us to verify that the data is consistent. Moreover, we can also observe that the three top diagnosis are ‘Sifilis latente, no especificada como precoz o tardia’, ‘Sifilis, no especificada’ and ‘Diabetes mellitus no insulinodependiente sin mencion de complicacion’ with percentages of 43.22%, 33.82% and 12.44% respectively. That is 89.48% of the data gathered in three categories and almost 50% in one. Furthermore, we also analyzed the distribution of data for each of our main health conditions: Syphilis (A5) and Diabetes (E1).

	Count	Percentage
A5	113650	81.1%
E1	26525	18.9%

Table 11. Syphilis vs Diabetes percentages

As we can see from Table 11, around 80% of the dataset is related to Syphilis, while the remaining 20% relates to Diabetes. This reinforces the fact that the dataset is very imbalanced with respect to the target variable for prediction. It is possible we will need to find a way to balance the data to be able to correctly predict between the two target classes.

Finally, feature Type shows three categories:

- Impresión Diagnóstica: Medical impression, or the "educated guess" on the condition/disease of the patient.
- Confirmado Nuevo: The patient has just been confirmed to suffer from the disease present in Name and Code.
- Confirmado repetido: The patient had previously been diagnosed and is in a follow-up or getting further tests for their condition.

Figure 11 shows there is a huge imbalance in the data with respect to the type of medical diagnoses, where 96.82% of the data has type: ‘Confirmado Repetido’. This again will have to be handled with methods to balance the data.

Finally, when we looked at the relationship of feature Type per Diagnoses (Diabetes or Sifilis), we observed that the percentages are very similar, where Type ‘Confirmado Repetido’ has the bigger proportion in both diagnoses.

3.3.3. Missing values analysis

We begin the analysis by plotting a missing values matrix.

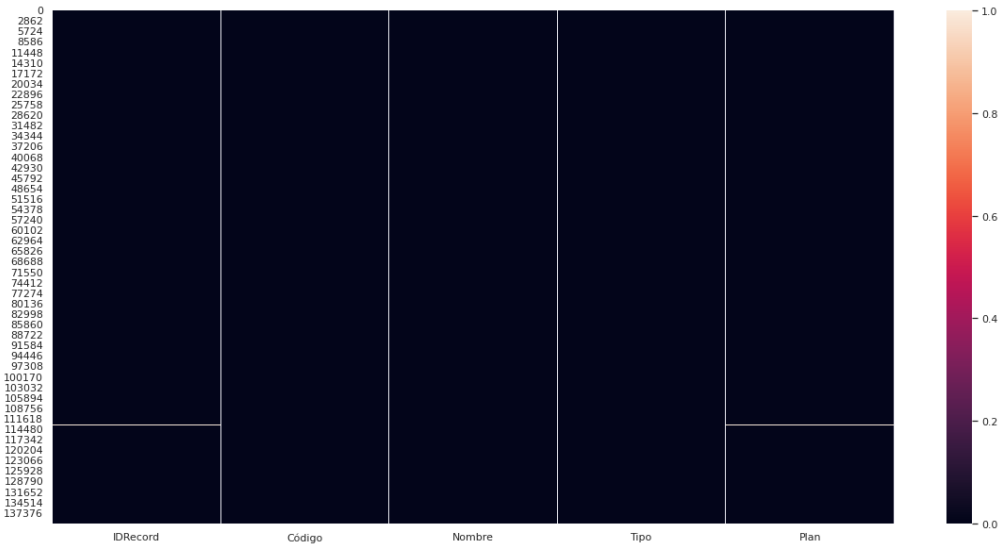


Figure 15. Missing values heat map for notes

As it can be noticed in Figure 15, there are few missing values for features IDRecord and Plan. Some possible solutions are:

- IDRecord: We can't replace IDRecord as that is a unique value used for joining together the information of the different datasets.

- Code/Name: Code/Name would be the feature we would be trying to predict on, so we can't fill those either. However, each Name uniquely corresponds to each Code, so we might be able to recover some missing values with that.

- Type, Plan: We can do simple inputting. We can replace the missing values from the null samples with the rest of the samples, in order to see if we can use the rest of the data found in the other datasets to predict the condition of the patient.

So, first, we take interest in the Code/Name features. Thus, we get the rows where these features are missing.

	IDRecord	Código	Nombre	Tipo	Plan
67822	espirar o tomar aire - Fiebre de difícil cont...	NaN	NaN	NaN	NaN
68962	ALIDADES DERIVADAS COMO PSICOLOGÍA, ODONTOLOGÍ...	NaN	NaN	NaN	NaN
70367	L TRATAMIENTO 4. ACUDIR A CITAS DE CONTROL EQ...	NaN	NaN	NaN	NaN
73613	/day for 10 days, ceftriaxone 500 mg IM in a s...	NaN	NaN	NaN	NaN
84051	SIÓN ARTERIAL ESTÁN POR ENCIMA DE 160/100. • ...	NaN	NaN	NaN	NaN
85907	n infectología ultima 11/2019 proximo en 05/20...	NaN	NaN	NaN	NaN
86209	ado Repetido	EMPEZÓ TAR EN DIC/15 SE REFORMULA IGUAL TAR ...	NaN	NaN	NaN
89127	EVIAMENTE SE EXPLICA CLARAMENTE CONDUCTA M...	NaN	NaN	NaN	NaN
89245	8 con TNF/FCT/EFV se educa acerca de efectos a...	NaN	NaN	NaN	NaN
90911	MOMENTO SE FORMULAN PRESERVATIVOS PARA PREVE...	NaN	NaN	NaN	NaN
106576	DE TRANSMISIÓN Y A DISMINUIR EL RIESGO DE TRA...	NaN	NaN	NaN	NaN
108230	2021. f) Paraclinicos control 01/2022 g) se ...	NaN	NaN	NaN	NaN
116216	ados hace una semana + anoscopia + biopsia -...	NaN	NaN	NaN	NaN
121294	8 TFG mdrd-4 (ml/min/1,73 m2) 104 ml/min. - V...	NaN	NaN	NaN	NaN
126567	titis b, influenza cepa 2021, neumococo preven...	NaN	NaN	NaN	NaN

Table 12. Missing values for Name/Code features

Table 12 shows that we are missing the same data for both Name and Code features, except for one (row 86209). At the same time, we can see that the samples that have NaN for Code and Name also have a problem with their IDRecord, making them unusable. After this, we decide to count the number of missing values per row again. Results are shown in Table 13.

	count
IDRecord	1
Código	48
Nombre	49
Tipo	51
Plan	110

	count
IDRecord	0
Código	0
Nombre	1
Tipo	3
Plan	61

Table 13. Missing values per column. Left: before dropping unusable missing values. Right: After dropping unusable missing values.

Next, we analyze the remaining missing values. The ones belonging to Name and Type (4 items) do not have IDRecords either, so they are unusable. For Plan, we analyze their relationship with other variables.

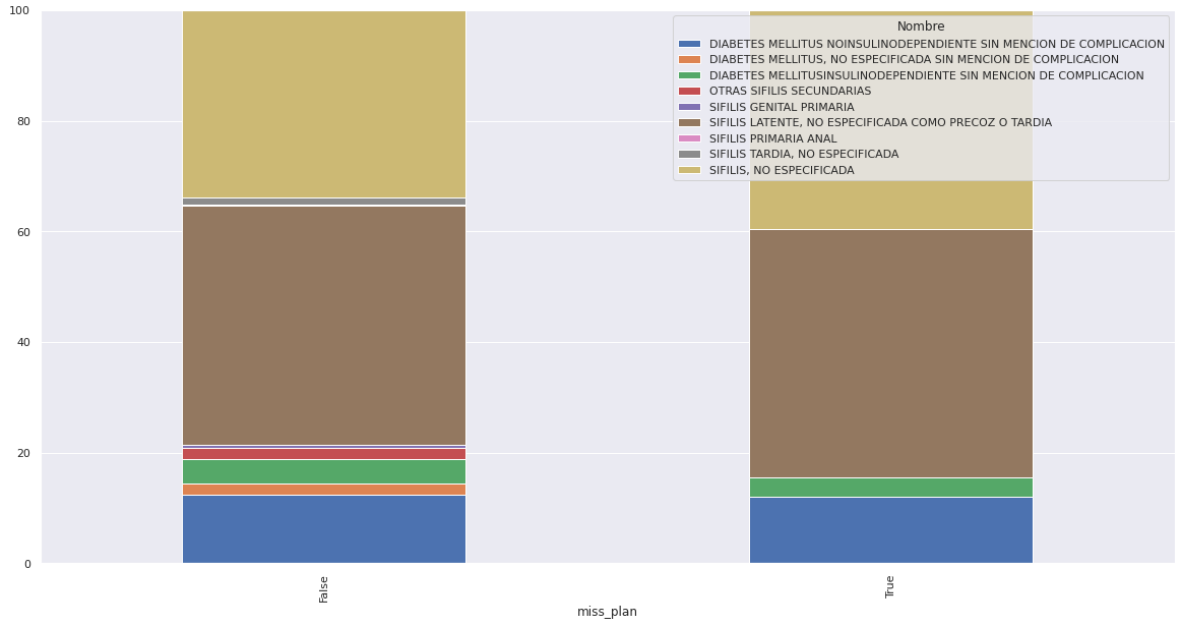


Figure 16. Missing and non missing values per diagnoses Name

After extracting the rows with missing values for Plan, we group them by Name and Type. As a result, we see that all of the rows have Type = Confirmado Repetido. At first sight, it could indicate that there is a relation between this category and the missing values. However, the proportion of this category in the dataset is close to 97%, so, it makes sense that this particular type is present in all of the missing values. Furthermore, we plot the relationship between feature Name and the missing/not missing values for Plan, but we saw that the proportions for missing and non missing values are similar. Therefore, we cannot identify a relationship other than it reflects the percentage per category in feature Name in the dataset.

In this case, we can only suppose the reason why the majority of null values have the plan description in the IDRecord feature and the rest of features as null. It may be a human error when passing information to be analyzed. In this sense, it would be a missing not at random (MNAR) type of missing values.

3.3.4. Text Analysis

3.3.4.1. Plan

Plan, as mentioned before, contains information related to the Electronic Health Records (EHR) of the patient, i.e. the Doctor's notes, in a String format. As the data is in Spanish, it can contain accented characters (e.g. "educación"), and thus it is ideal to remove these accents as they would be considered as a different word in a multitude of analysis and prediction algorithms.

Additionally, stopwords need to be removed to facilitate a better look at the available data.

A Word Cloud plot was obtained from the feature, in order to see what are some of the relevant words present in the data. This plot can be seen in Figure 17.

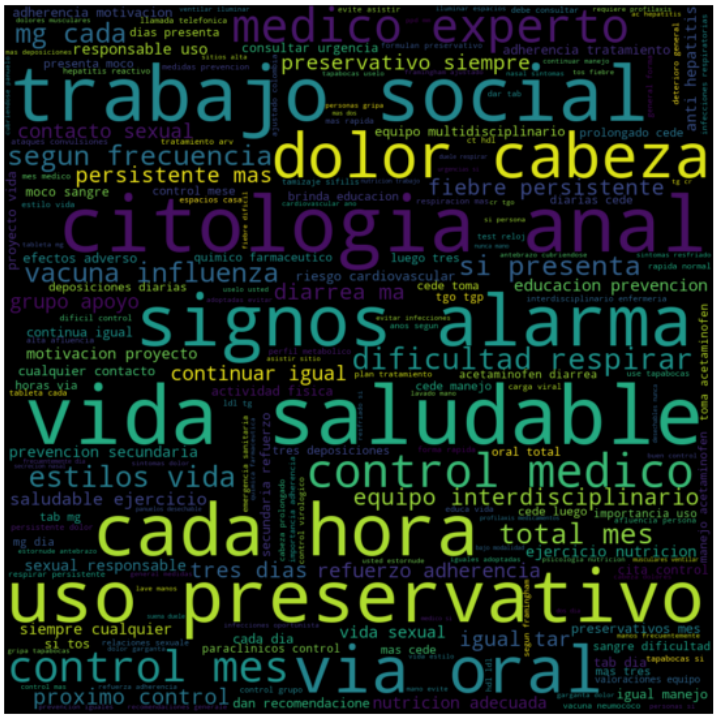


Figure 17. Word cloud of the Plan variable of the dataset, showing the most common words.

Some words of interest are medical control, signs of alarm, headache, anal cytology, and use of preservatives. These last two pairs of words can be attributed to Syphilis instead of diabetes, which further demonstrates the imbalance of the dataset for the two diseases that will be analyzed.

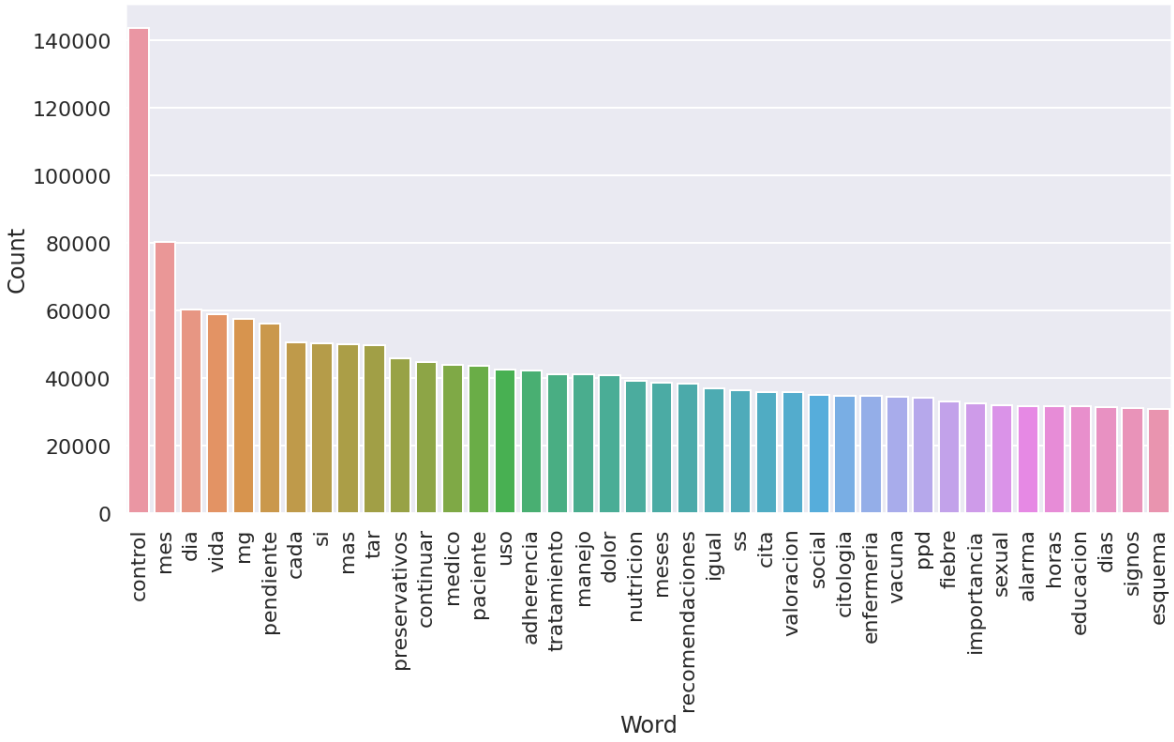


Figure 18. Barplot with word count for the top 50 words.

Figure 18 shows a barplot of the word count for the top 50 words in the dataset, discounting stopwords. It can be seen that control is a very common word found in the dataset, followed by date-related words (mes, meses, dia, etc).

When dividing the dataset by Code/disease and looking at the top 10 words of each one, as seen on Figure 19, the pattern is similar to the one present when looking at the full dataset, being “control” the most common word for the majority of the conditions. Pain, preservatives, and adherence were also relevant words for Syphilis, while date-related words as well as tablet and milligrams were stronger for diabetes, which could help us guide in the creation of new features based on this information in order to facilitate the prediction of the disease from the data.

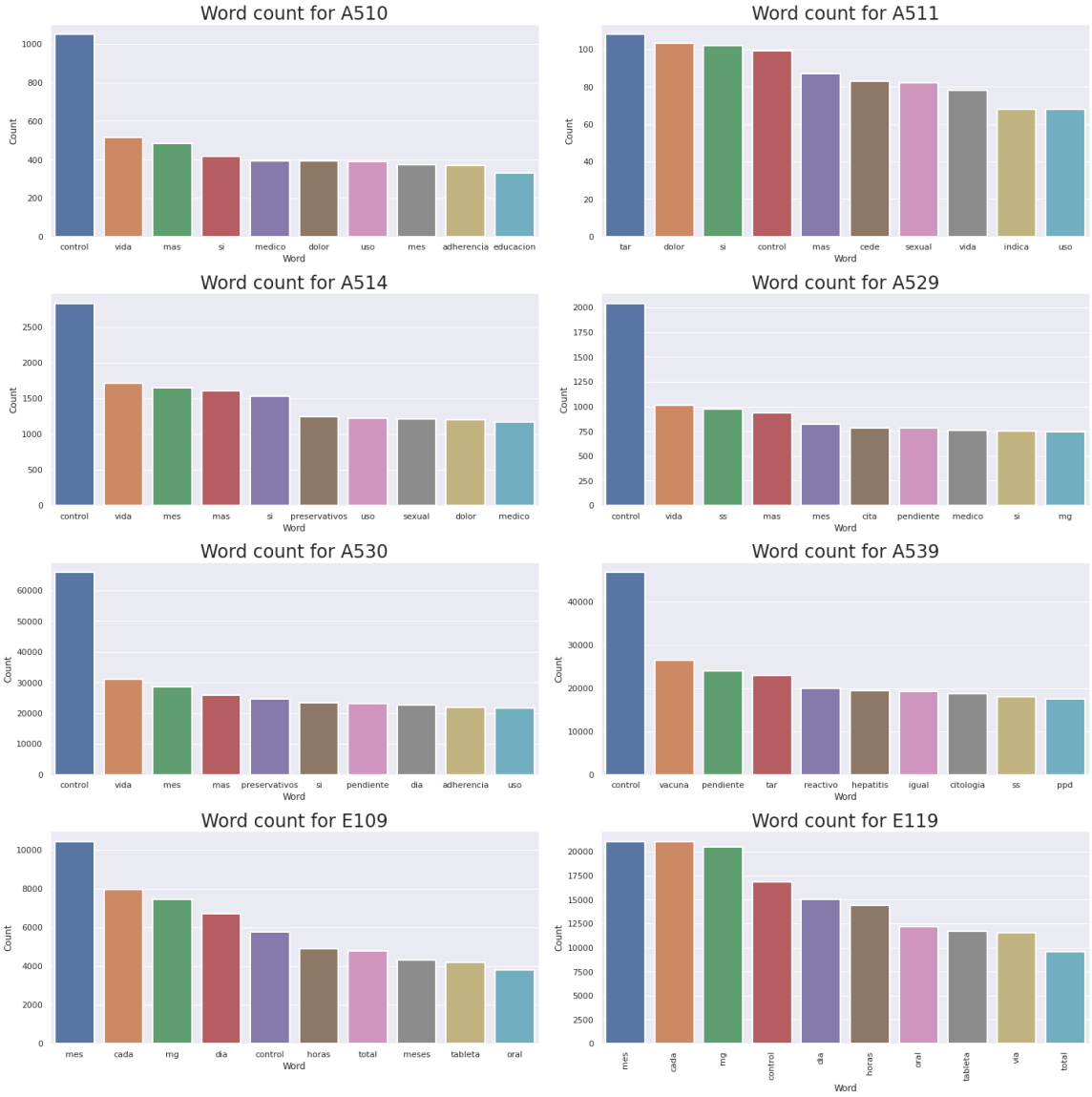


Figure 19. Word counts and percentages for the top 10 words of each disease

In addition to a word cloud, the term frequency–inverse document frequency (TF-IDF) statistic was calculated for the dataset after removing numbers and stopwords from the data. The resulting plot for the top 50 words can be seen in Figure 20.

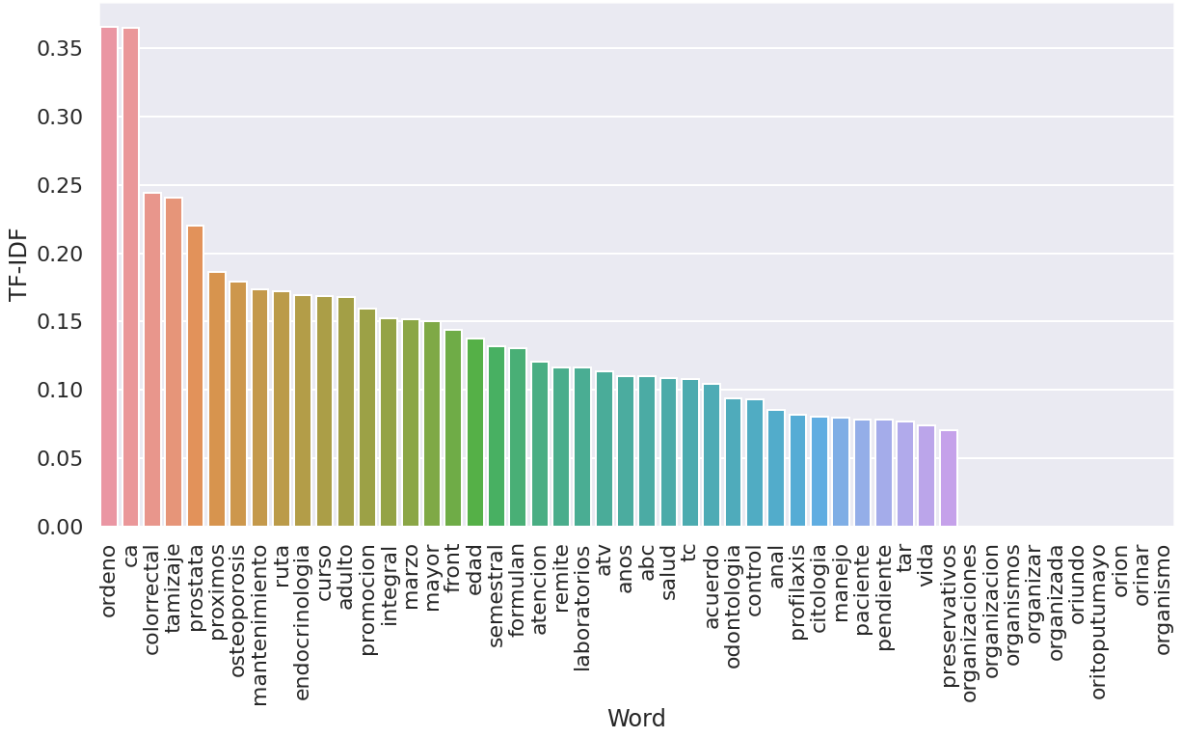


Figure 20. TF-IDF plot of the top 50 words of the Plan feature.

From this plot we can see that colorectal, endocrinology, age, prostate, etc., seem to have a higher relevance than the rest of the words in the dataset. In addition to this, it is also worth noting that there is a sudden drop in the TF-IDF value after the word “preservativos” (preservatives), which means there are only around 40 words that could be considered of high importance in the dataset according to this statistic. Some words, like CA, ATV, TC, TAR, seem to be acronyms, and should be further investigated to see if there is a way to obtain a better idea of what they represent, as it seems they relate to exams performed on the patient and thus could be relevant for our analysis

3.4. EDA between datasets

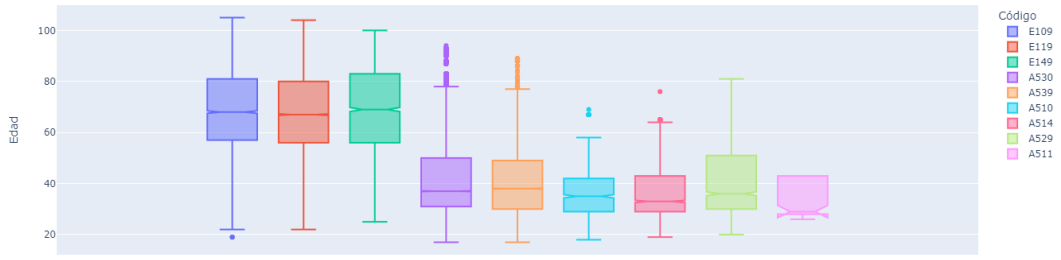
After analyzing each table and their columns, and cleaning some identified errors in the dataset, we proceeded to merge the information to obtain further insights.

3.4.1. Notes and Sociodemographic

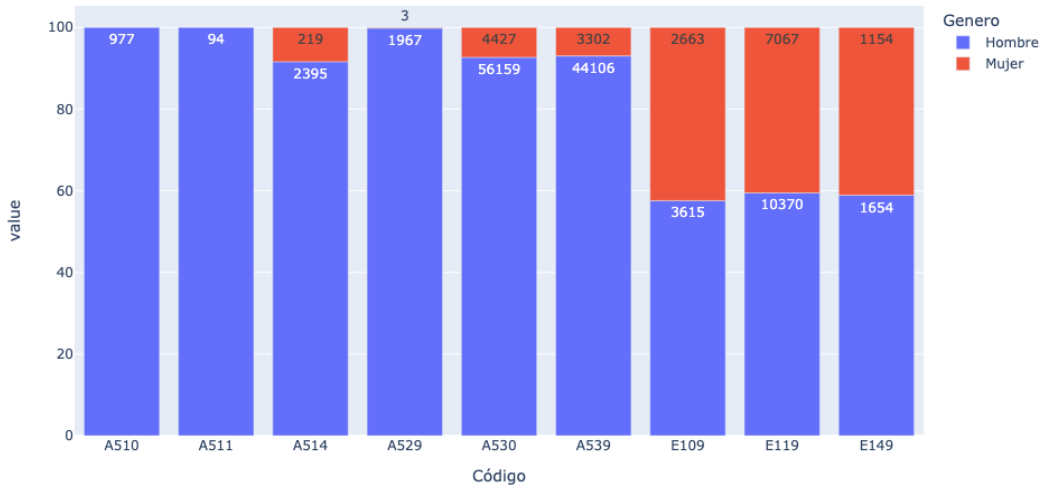
First, using the ID record, we merged the medical notes with the sociodemographic information. As our target variable is the code of the disease, it results relevant to analyze it against our categorical sociodemographic variables, such as gender,

ethnic group, residential area and marital status. Figures showing the frequency of cases for each disease depending on the categorical variables were plotted.

In section 3.1, we already identified that the dataset is highly unbalanced regarding gender as 80% of the 9306 patients are males. Similarly, in section 3.3 we identified that most observations correspond to syphilis diagnosis. From Figure 21 we can evidence that most diagnoses of syphilis are for young, male patients. One may think this is related to the fact that most patients in the dataset might be young males, but the interesting thing is that for diabetes there is balance regarding gender. 41% of observations (10,884) where the diagnosis is diabetes correspond to women. Meanwhile, there are only 7,948 records of syphilis diagnosis for women, although a total of 113,649 observations have a code beginning in 'A5'. Given this, when differentiating only between syphilis and diabetes, gender will probably be an important feature in a predictive model, and this correlates to what was found in our literature review (Ricco & Westby, 2020, 91).



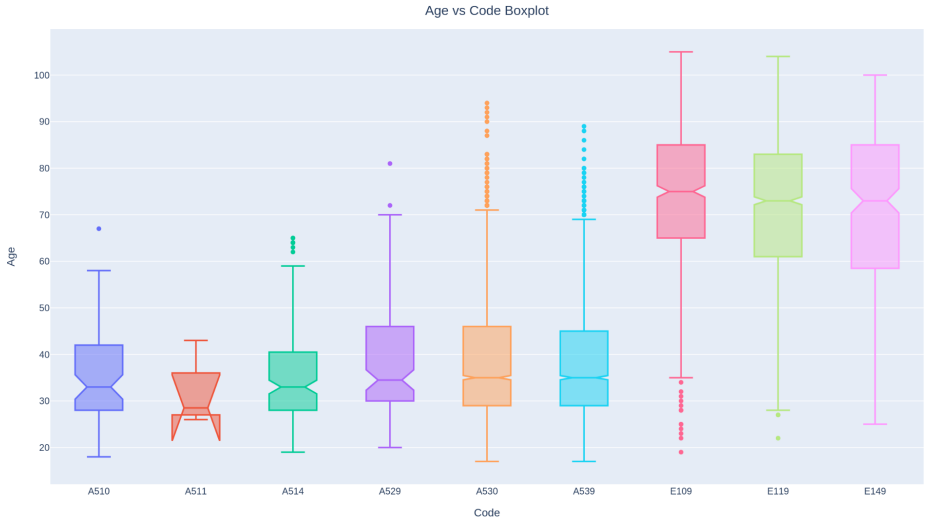
(a)



(b)

Figure 21. Disease distribution for Age (a) and gender (b), for all of the data

It should be noted from the previous figure that there could be multiple medical notes for the same patient. Therefore, it is relevant to group the data by patient and only keep one observation for each one, in order to further evidence the relationship between diagnoses and gender. Figure 22 confirms that almost all the patients diagnosed with syphilis are young males, whereas for diabetes we have a greater number of women, even surpassing males for type 1 diabetes mellitus.It can also be seen that the population that suffers from diabetes tends to be of a more advanced age.



(a)

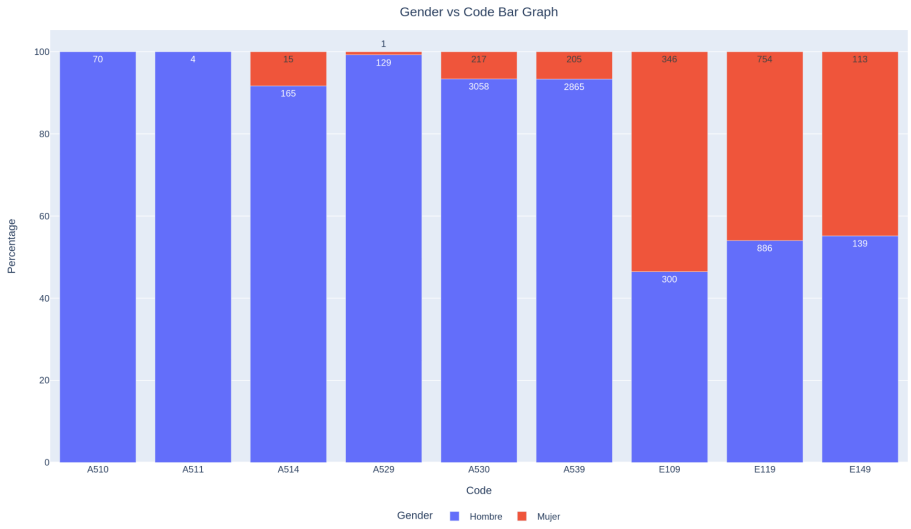


Figure 22. Disease distribution for Age (a) and gender (b), for unique patients

A similar analysis was conducted for ethnic group, residential area, and marital status, as shown in figures 23, 24, and 25, respectively. For ethnic groups, as it was already identified, most patients are mestizos hence it is expected that for all the diseases it is the prevalent group. However, it is interesting that for diabetes there is an important number of patients with unknown ethnic groups. Sadly, the fact that it is unknown does not allow us to further dig into this.

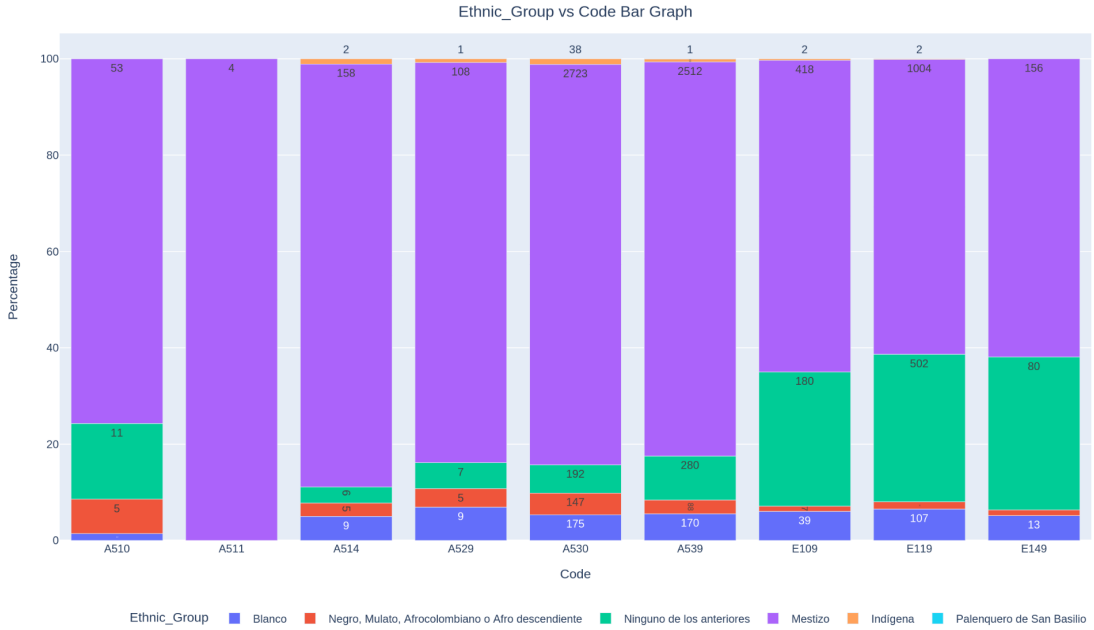


Figure 23. Disease frequency by ethnic group

For the variable residential area, there is not much information as most patients live in urban areas and there is a low proportion of patients in rural areas for each disease. On the other hand, there are relevant findings regarding marital status. Almost 50% of the patients in the dataset are single and this is the prevalent marital status for cases of syphilis, which again corroborates what was found in our literature review (Ricco & Westby, 2020, 91). However, for diabetes the most common marital status is marriage. Furthermore, there is an important number of widows with diabetes, whereas there are only 28 cases for syphilis. Like gender, this variable could be key to classify patients between syphilis and diabetes, at least in regards to this sample.

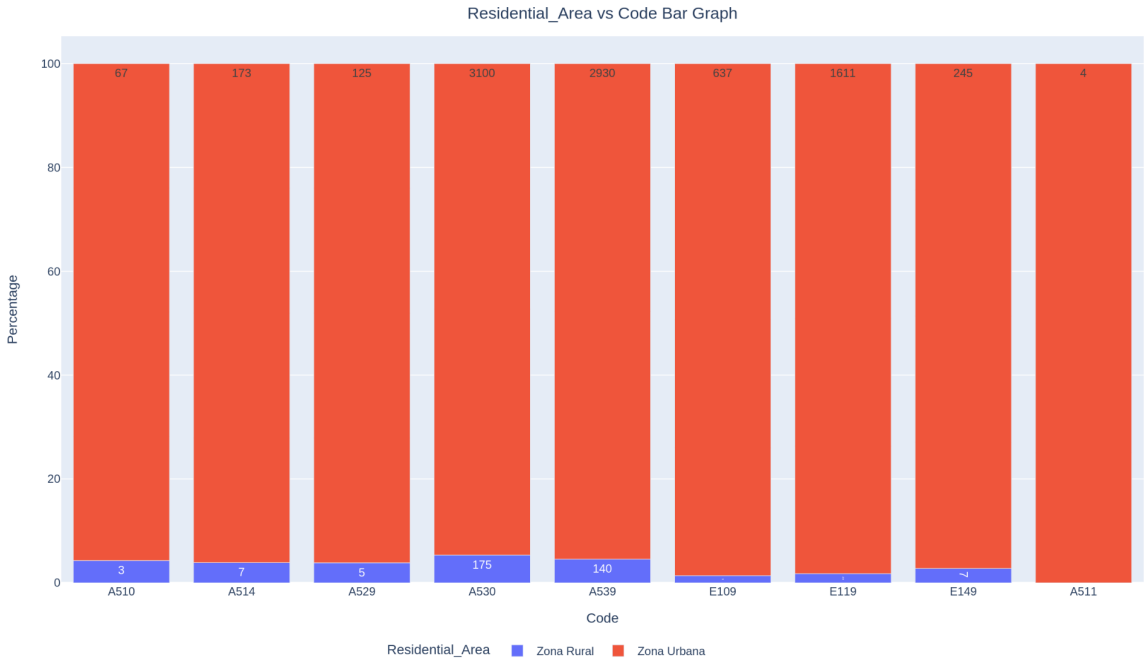


Figure 24. Disease frequency by residential area

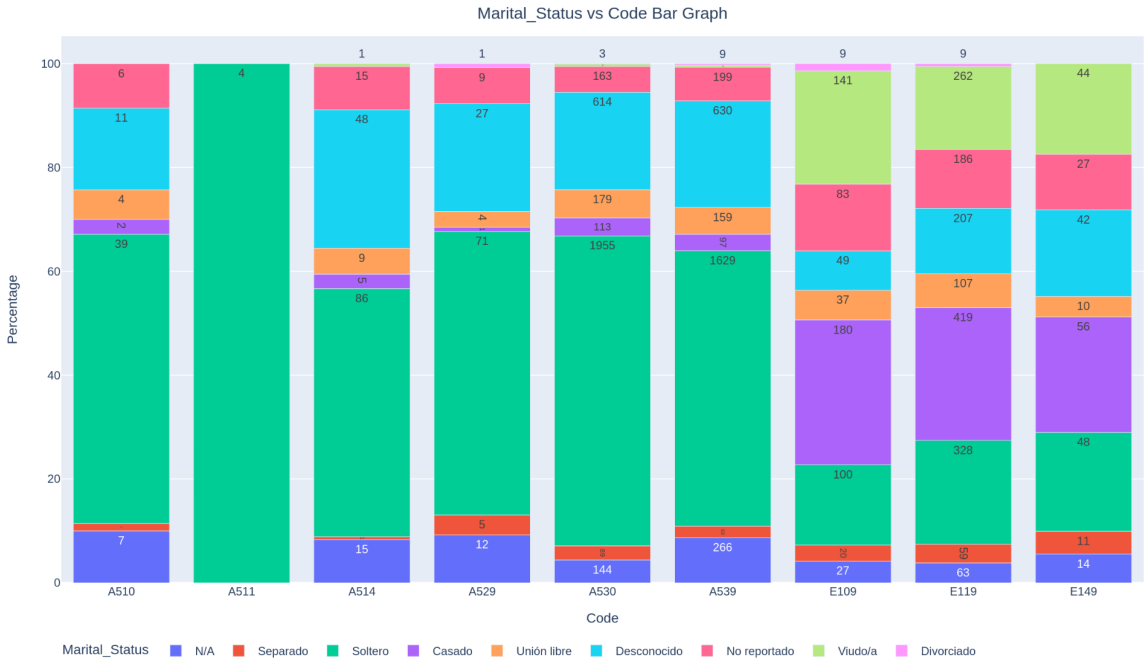


Figure 25. Disease frequency by marital status

3.4.2. Notes and Laboratories

We merged Laboratories and Notes data to analyze what information they hold. Each RecordID corresponds to the information of a person, so we used this key to merge them.

	IDRecord	Codigo	Nombre_labs	Fecha	Valor	Código	Nombre_notes	Tipo	Plan
0	95627	902045	TIEMPO DE PROTROMBINA (PT)	2022-02-22 18:43:00	NaN	A530	SIFILIS LATENTE, NO ESPECIFICADA COMO PRECOZ O...	Confirmado Repetido	VER ANALISIS
1	95627	902045	TIEMPO DE PROTROMBINA (PT)	2022-02-22 18:43:00	NaN	A530	SIFILIS LATENTE, NO ESPECIFICADA COMO PRECOZ O...	Confirmado Repetido	NOTIFICAR A TIEMPO LA FECHAS DE POSIBLE VIAJE ...
2	95627	902045	TIEMPO DE PROTROMBINA (PT)	2022-02-22 18:43:00	NaN	A530	SIFILIS LATENTE, NO ESPECIFICADA COMO PRECOZ O...	Confirmado Repetido	- TDF/FTC+DTG - CONTROL DE HEP C (EPCLUSA) -...
3	95627	902045	TIEMPO DE PROTROMBINA (PT)	2022-02-22 18:43:00	NaN	A530	SIFILIS LATENTE, NO ESPECIFICADA COMO PRECOZ O...	Confirmado Repetido	PLAN DE ENFERMERIA COMO SE DETECTA EL CANC...
4	95627	902045	TIEMPO DE PROTROMBINA (PT)	2022-02-22 18:43:00	NaN	A530	SIFILIS LATENTE, NO ESPECIFICADA COMO PRECOZ O...	Confirmado Repetido	Plan de tratamiento PROXIMO CONTROL CON: Me...
5	95627	902045	TIEMPO DE PROTROMBINA (PT)	2022-02-22 18:43:00	NaN	A530	SIFILIS LATENTE, NO ESPECIFICADA COMO PRECOZ O...	Confirmado Repetido	- TDF/FTC+DTG - CONTROL DE HEP C (EPCLUSA) -...
6	95627	902045	TIEMPO DE PROTROMBINA (PT)	2022-02-22 18:43:00	NaN	A530	SIFILIS LATENTE, NO ESPECIFICADA COMO PRECOZ O...	Confirmado Repetido	*Control por Trabajo Social segun frecuencia y...

Table 14. Laboratories and Notes tables merged

The new merged table contains 5,055,873 rows. Given that the IDRecord of 168 Notes and 16130 Laboratories tests did not correspond to each other, there are 16298 rows that are not included in the final table.

The Notes file contains 9270 unique IDRecords (unique people) and the Laboratories file has 5964 unique IDRecords (unique people). After merging the tables, the lost rows correspond to 12 unique IDRecords lost from Notes (0.13% of all people) and 3318 unique IDRecords lost from Laboratories (55.6% of all people), which gives a total of 3330 people who are not included in the final table (the not included 16298 rows correspond to them). It is clear that the loss of people coming from Laboratories is considerable. Thus, from the 9282 unique IDRecords from both tables, we work with 5962 unique patients.

Another remarkable finding is that the “Valor” column is the only one with null values. It has 727,516 null values which correspond to the 14.39% of the total rows.

When we look at one IDRecord information in both tables, we saw that the same ID may have multiple rows in Notes (different doctor appointments) and multiple rows in Laboratories (different tests). Laboratories contain the column Date which helps to place the lab results in time. However, Notes does not, which complicates our

matching between lab results and notes. We cannot be sure if a note mentions one or more lab results or if a lab result has one or more notes related.

3.4.2.1. Text processing for test names

There are many lab tests done for a particular diagnosis. After an initial inspection, we can see that there are many lab test names that are similar to each other. So, we inspected the Nombre column from the Laboratories. First we lower case the column, strip blank spaces at the start and end of sentences, and strip accents. Then, we use the Jaccard coefficient which computes a similarity measure between texts. This allows us to compare each row (test name) to the others in order to find similar names. After we returned the most similar names, we detected that the problematic names had common special characters. So, we search for these characters to return all of these terms. Results are shown on the left of Table 15.

30	*hemograma iv (hemoglobina, hematocrito, recue...
65	acido folico (folatos) en suero
147	acido folico [folatos] en suero . .
41	acido urico en suero u otros fluidos
36	albumina
129	albumina . .
146	albumina en orina de 24 h
35	albuminuria en orina parcial
75	baciloscopia coloracion acido alcohol-resisten...
71	baciloscopia coloracion acido alcohol-resisten...
99	citomegalovirus, anticuerpos ig g (cmv-g) por eia
110	citomegalovirus, anticuerpos ig m (cmv-m) por eia
6	colesterol de baja densidad (ldl) enzimatico
46	colesterol de baja densidad (ldl) inmunologico...
51	creatinina
0	creatinina
95	creatinina depuracion
74	creatinina en orina de 24 h
42	creatinina en orina parcial
87	estudio de coloracion basica en citologia anal

	name	Codigo	Count
152	acido folico (folatos) en suero	903505	1
65	acido folico (folatos) en suero	903105	60
125	albumina	501739	2
36	albumina	903803	973
0	creatinina	903895	11712
61	creatinina	903895..	129
98	hormona estimulante del tiroides (tsh) ultrase...	60281	8
21	hormona estimulante del tiroides (tsh) ultrase...	904904	3411
158	potasio	60198	1
34	potasio	903859	1101
5	transaminasa glutamico oxalacetica o aspartato...	903867	9269
136	transaminasa glutamico oxalacetica o aspartato...	60242	1
137	transaminasa glutamico oxalacetica o aspartato...	60241	1
83	transaminasa glutamico piruvica o alanino amin...	60232	18
66	transaminasa glutamico piruvica o alanino amin...	60231	58
68	trigliceridos	60101	52
3	trigliceridos	903868	9905

Table 15. Laboratory tests name cleaning. Left: before. Right: after.

Once these values were detected, we proceeded to clean them. We applied a Regex to replace special characters such as '[ ]' and periods, and we obtained the repeated names shown on the right of Table 15. There are repeated names with different codes. After asking our IQVIA contact, we defined that we would take the same names as the same tests even though they had different codes.

3.4.2.2. Test names and Diagnoses names

As an initial step, we aimed to see the relationship between the tests that were taken per diagnosis. To do so, we plotted a bubble diagram.

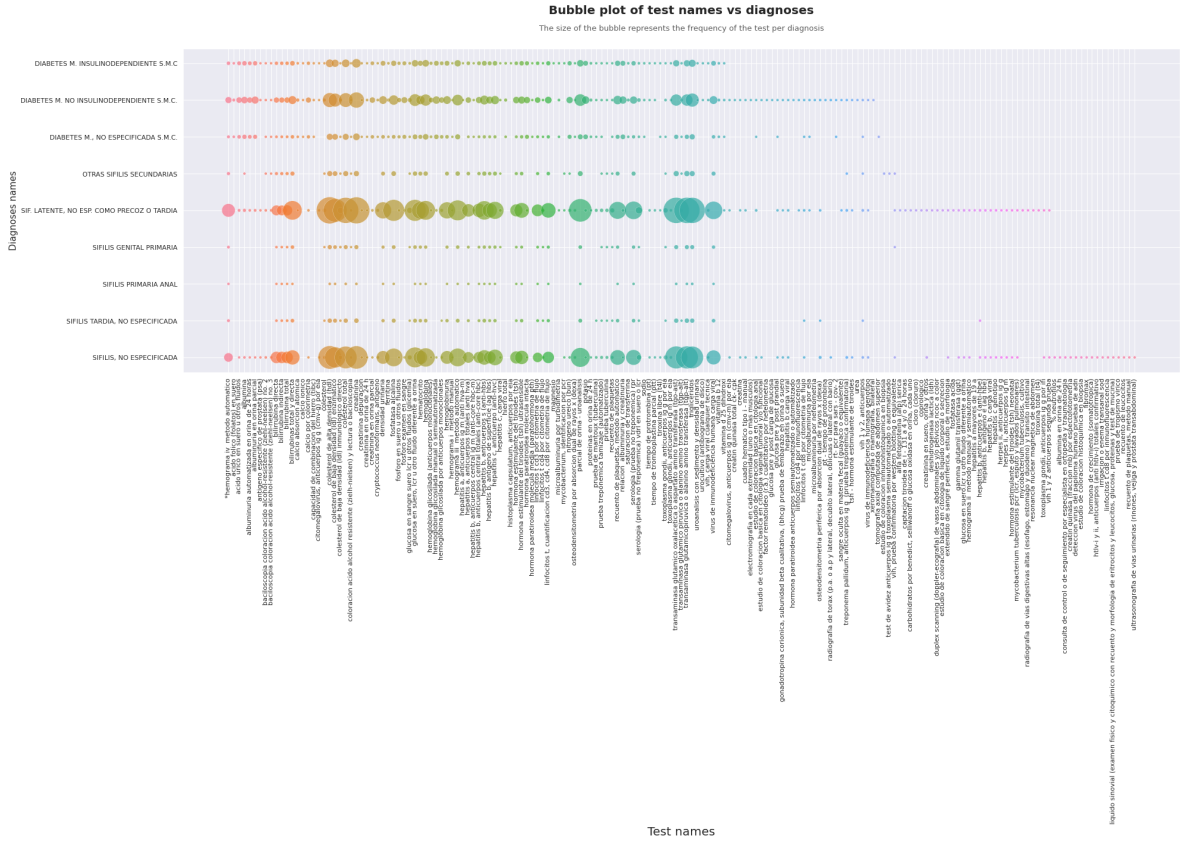


Figure 26. Bubble plot of test names vs diagnoses.

The previous plot let us see that 'Sifilis latente no espec. como precoz o tardía', 'Sifilis no especificada' and 'Diabetes No insulinodependiente' are the ones with more tests. However, this is due to the proportions of data that we have: 44.42%, 33.0% and 13.08% respectively. This means that around 90% of the data is concentrated in these three categories. If we divide the Notes categories in the generic categories of Diabetes and Sifilis, we can see that there are some tests that are not performed for Diabetes. Specifically, the tests shown after 'tomografia axial computada de abdomen superior' (including it).

Given that we have so many categories for Lab. names, and it is difficult to see any relationships between categories, we proceeded to group the categories for each feature into larger (meaningful) categories. This way we avoid losing data that might contain valuable patterns and, if we find a possible relationship between two larger categories, it will still clue us into looking into their subcategories in more detail.

The heuristic we followed to group the test names was: group the test names that have the same name with different variations. Figure 27 shows a subset of the

defined dictionary, where the grouped categories ended with the word ‘test’. Following the same idea, the diagnoses were grouped into the two bigger categories: ‘Diabetes’ and ‘Sifilis’.

```
dict_tests = {'% neutrofilos' : '% neutrofilos',
'acido folico (folatos) en suero' : 'acido folico (folatos) en suero',
'acido urico en suero u otros fluidos' : 'acido urico en suero u otros fluidos',
'albumina' : 'albumina test',
'albumina en orina de 24 h' : 'albumina test',
'albuminuria automatizada en orina de 24 horas' : 'albuminuria test',
'albuminuria en orina parcial' : 'albuminuria test',
'alfa fetoproteina (afp) serica' : 'alfa fetoproteina (afp) serica',
'antigeno especifico de prostata (psa)' : 'antigeno especifico de prostata (psa)',
'baciloscopia coloracion acido alcohol-resistente (ziehl-neelsen) no. 2' : 'baciloscopia test',
'baciloscopia coloracion acido alcohol-resistente (ziehl-neelsen) no. 3' : 'baciloscopia test',
'bilirrubina directa' : 'bilirrubina test',
'bilirrubina indirecta' : 'bilirrubina test',
'bilirrubina total' : 'bilirrubina test',
'bilirrubinas total y directa' : 'bilirrubina test',
'calcio absorcion atomica' : 'calcio test',
'calcio ionico' : 'calcio test',
'calcio por colorimetria' : 'calcio test',
'capacidad de combinacion del hierro (tibc)' : 'capacidad de combinacion del hierro (tibc)',
'captacion tiroidea de 1 - 131 a 4 y/ o 24 horas' : 'captacion tiroidea de 1 - 131 a 4 y/ o 24 horas',
'carbohidratos por benedict, sellwanoff o glucosa oxidasa en orina, cada uno' : 'carbohidratos por benedict,
'citomegalovirus, anticuerpos ig g (cmv-g) por eia' : 'citomegalovirus-anticuerpos test',
'citomegalovirus, anticuerpos ig m (cmv-m) por eia' : 'citomegalovirus-anticuerpos test',
'cloro (cloruro)' : 'cloro (cloruro)',
'colesterol' : 'colesterol test',
'colesterol de alta densidad (hdl)' : 'colesterol test',
'colesterol de baja densidad (ldl) enzimatico' : 'colesterol test',
'colesterol de baja densidad (ldl) inmunologico directo' : 'colesterol test',
'colesterol total' : 'colesterol test',
```

Figure 27. Subset of dictionary of test names.

As a result, the next plot shows the density distribution per diagnosis. It is normalized so that tests within each diagnosis sum one.

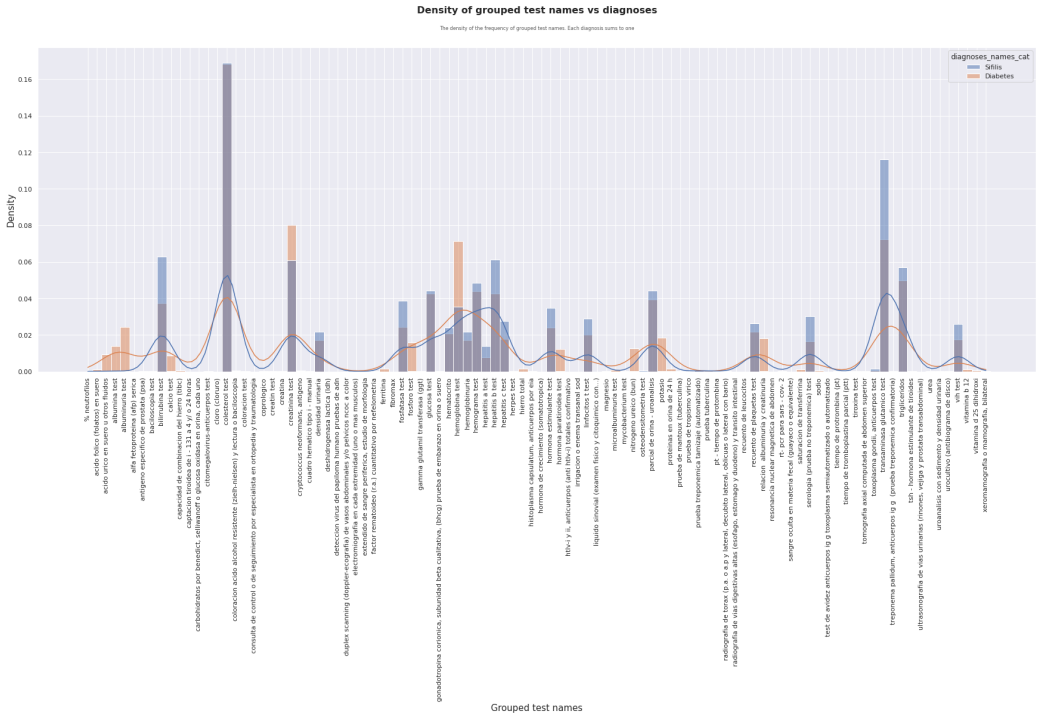


Figure 28. Density distribution per diagnosis.

We can see that there are more tests that are unique for Diabetes, such as albuminuria test or potasio. There are tests that are done for both diagnoses, however some of them have bigger frequency within Syphilis such as bilirrubina

test and transaminasa glutamico test. It is interesting to notice the differences within the kernel density estimates (kde's). There are two pikes where they differ the most, at the beginning of the plot and in the middle. In the middle we can see that the hemoglobina test and hepatitis B test skew the kde's in opposite directions.

Next, we will delve into the relationships between these categories.

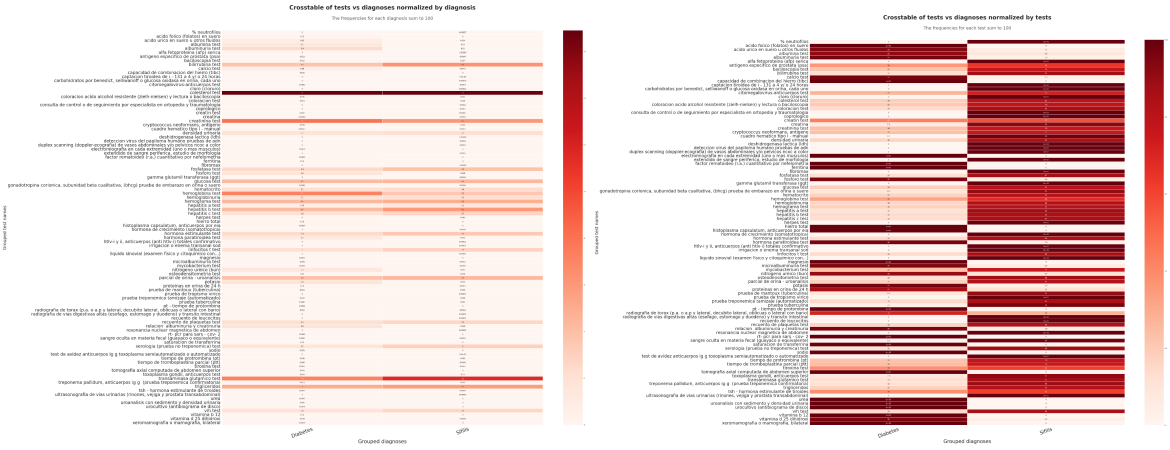


Figure 29. Cross tables of tests vs diagnoses. Left: normalized by diagnoses. Right: normalized by tests.

Cross table normalized by diagnoses: It allows us to bypass the unbalanced dataset regarding diagnoses because the percentages are given by Diabetes and Syphilis. In other words, we can see the "importance" of each test by each diagnosis. Table 16 shows the main tests, the ones that represent more than the 4% of all tests for both diagnoses. So, for Diabetes and Syphilis, it is more likely that people have taken these tests.

diagnoses_names_cat	Diabetes	Sifilis
tests_names_cat		
colesterol test	16.836439	16.901008
creatinina test	8.030761	6.088023
glucosa test	4.271277	4.438232
hemograma test	4.399174	4.850576
hepatitis b test	4.271794	6.122522
parcial de orina - uroanalisis	3.938581	4.430476
transaminasa glutamico test	7.228077	11.606624
trigliceridos	4.981781	5.686469

Table 16. Main tests

From these tests done for both diagnoses, we pay more attention to the ones that have a big difference in the importance of the correspondent diagnosis (big difference in percentages). For example, creatinina (8% in Diabetes vs 6.1% in

Syphilis), hepatitis B (4.3% vs 6.1%), trigliceridos (5% vs 5.7%) and transaminasa glutamico test (7.2% vs 12%).

Following the same reason, Table 17 shows the main tests (also the ones that represent more than the 4% of all tests) that are **u**nique between the two diagnoses and have more likelihood to occur in the diagnosis process. For Diabetes, a hemoglobina test and for Syphilis a bilirubina test and a fosfatasa test.

diagnoses_names_cat	Diabetes	Sifilis
tests_names_cat		
hemoglobina test	7.137032	3.527494
bilirrubina test	3.725935	6.272285
fosfatasa test	2.443561	3.878303

Table 16. Main unique tests

It is worth noticing that we have only compared names of tests. We should also take into account the values returned by the test results. Especially for those tests that are repeated and have similar percentages. For example in Cholesterol (17% vs 17%), glucosa test (4.3 vs 4.4) , hemograma test (4.4% vs 4.9%), and parcial de orina - uroanalisis test (3.9% vs 4.4%).

Cross table normalized by tests: In this case, each row shows the percentage of diagnoses that correspond to the test. However, around 80.85% of the data corresponds to Syphilis. Therefore, for tests that are done for both diagnoses, it is more likely that Syphilis has a bigger percentage. So, we should be careful when inspecting those tests. On the other hand, if there is a bigger percentage for Diabetes for a given test, it might be significant (given that there are far less diabetes cases). The table is also handy to see which tests are done only for Syphilis or for Diabetes. This can be observed for tests that have a 100% of occurrence for either of the diagnoses.

According to the table, we can see that there are 30 (more than 90% except for one of 87%) tests that are performed mostly for Diabetes. Specially, 18 are exclusive for Diabetes. For Syphilis, we see that there are 24 tests that happen only for this diagnosis.

From these unique tests, we are interested in the ones that have a bigger percentage of occurrence for each diagnosis. So, we compare the unique tests given by the cross table normalized by tests and search for the higher frequencies in the cross table normalized by diagnoses.

tests_names_cat	Diabetes	Sifilis
acido urico en suero u otros fluidos	0.916232	0.013897
albumina test	1.392619	0.030046
albuminuria test	2.429935	0.019867
calcio test	0.875871	0.007144
fosforo test	1.583587	0.017910
hormona paratiroidea test	1.225600	0.011671
nitrogeno ureico (bun)	1.267819	0.032297
potasio	1.857548	0.015023
relacion albuminuria y creatinuria	1.799019	0.016026

tests_names_cat	Diabetes	Sifilis
consulta de control o de seguimiento por especialista en ortopedia y traumatologia	0.000000	0.001248
coprologico	0.000000	0.001737
deteccion virus del papiloma humano pruebas de adn	0.000000	0.001370
duplex scanning (doppler-ecografia) de vasos abdominales y/o pelvicos ncoc a color	0.000000	0.001199
extendido de sangre periferica, estudio de morfologia	0.000000	0.001517
gamma glutamil transferasa (ggt)	0.000000	0.001786
herpes test	0.000000	0.003034
hormona de crecimiento (somatotropica)	0.000000	0.001590

Table 17. Most repeated unique tests. Left: Diabetes. Right: Syphilis

The left side of Table 17 shows the most repeated unique tests for Diabetes and the right side shows the most repeated unique tests for Syphilis.

As a result, we extracted the most requested tests for the diagnoses and the most repeated unique tests per diagnoses. We hypothesize that they are significant for diagnosing Diabetes or Syphilis. Figure 30 plots the main tests that we observed in the previous section.

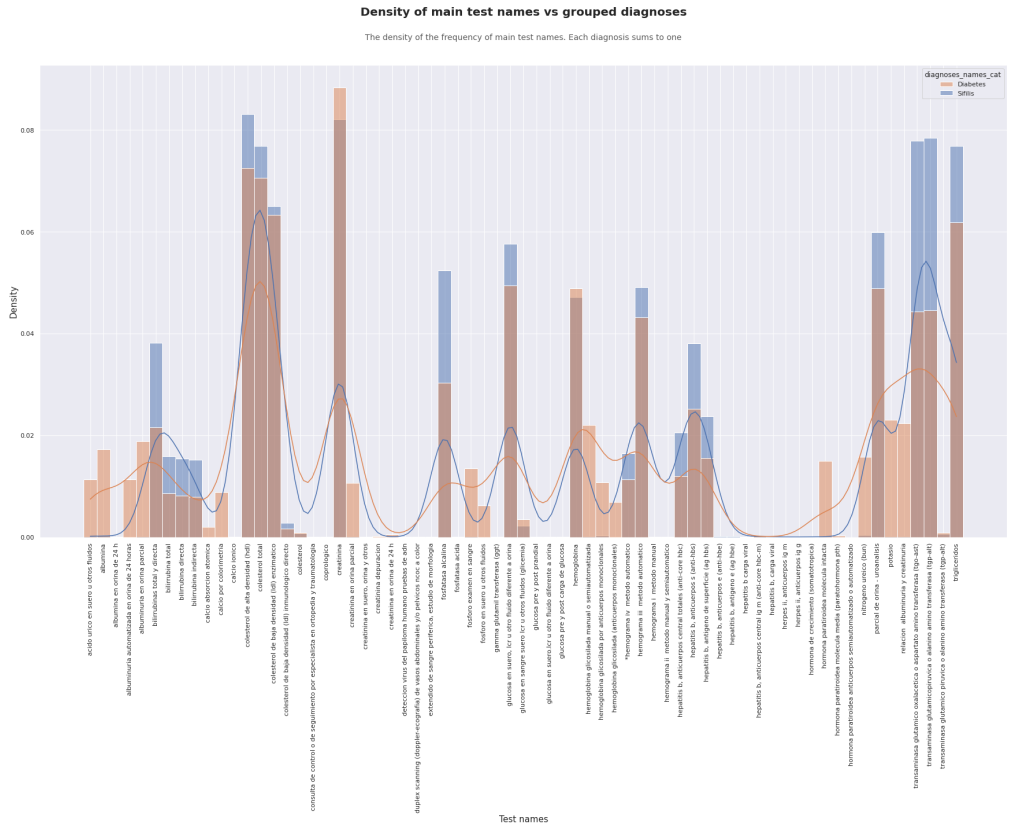


Figure 30. Density of main test names vs grouped diagnoses

3.5. Target feature analysis

As we previously saw, and as we can further see in Table 18, there is a strong imbalance in our dataset, where more than 75% of the data represents Syphilis-related diseases, and the remaining ~19% representing diabetes-related diseases.

		Count	Percentage
Nombre	Código		
SIFILIS LATENTE, NO ESPECIFICADA COMO PRECOZ O TARDIA	A530	60587	43.22
SIFILIS, NO ESPECIFICADA	A539	47408	33.82
DIABETES MELLITUS NOINSULINODEPENDIENTE SIN MENCION DE COMPLICACION	E119	17439	12.44
DIABETES MELLITUSINSULINODEPENDIENTE SIN MENCION DE COMPLICACION	E109	6278	4.48
DIABETES MELLITUS, NO ESPECIFICADA SIN MENCION DE COMPLICACION	E149	2808	2.00
OTRAS SIFILIS SECUNDARIAS	A514	2614	1.86
SIFILIS TARDIA, NO ESPECIFICADA	A529	1970	1.41
SIFILIS GENITAL PRIMARIA	A510	977	0.70
SIFILIS PRIMARIA ANAL	A511	94	0.07

Table 18. Target feature classes and percentages.

Due to this imbalance, we considered merging A510 and A511 with A514, as they all belong to the A51 Early syphilis ICD-10 denomination, indicating they share symptoms (World Health Organization, 2019). The equivalent Spanish name for this category is “Sífilis precoz” (Pan American Health Organization, 2019).

We also considered treating A529 as part of A539, in order to reduce the number of classes without losing track of the important subcategories we already have. With these merges, we end up with what can be seen in Table 19.

		Count	Percentage
Nombre	Código		
SIFILIS LATENTE, NO ESPECIFICADA COMO PRECOZ O TARDIA	A530	60587	43.22
SIFILIS, NO ESPECIFICADA	A539	49378	35.23
DIABETES MELLITUS NOINSULINODEPENDIENTE SIN MENCION DE COMPLICACION	E119	17439	12.44
DIABETES MELLITUSINSULINODEPENDIENTE SIN MENCION DE COMPLICACION	E109	6278	4.48
SIFILIS PRECOZ	A51	3685	2.63
DIABETES MELLITUS, NO ESPECIFICADA SIN MENCION DE COMPLICACION	E149	2808	2.00

Table 19 Target feature classes and percentages after merging A510, A511 and A514.

Nonetheless, this was not effective in our models and were discarded, opting instead for using data balancing

4. FEATURE ENGINEERING

4.1. Text-based feature engineering

4.1.1. Syphilis

According to the CDC, primary syphilis is characterized by a chancre mark where the disease enters the body (Centers for Disease Control and Prevention (CDC), n.d.). There is also a possibility of having extra sores in your body, but there does not seem to be any difference per se in the development of its condition based on where the disease started. Another clear indication of syphilis is Saber shin (pierna/tibia en sable).

There is also a reduction in cognitive abilities for patients who have been suffering of syphilis for some time, and this can be tested for using a simple test called the Clock Drawing Test (Government of British Columbia, Canada, n.d.), or Test del Reloj in Spanish (Organización Sanitas, n.d.).

Additionally, a main characteristic of primary syphilis seems to be chancres, as well as sores for both Primary and Secondary Syphilis, making a case for creating a new numerical variable called "chancres". Another main characteristic of Syphilis is the push to use preservatives in order to reduce the possibility of other people being infected as well, which could help differentiate between Syphilis and Diabetes (Brown & Frank, 2003, 283).

Other related keywords we tried based on our information review were genital, skin (lesions), headache, HIV, serology, hepatitis, and specific tests performed on the patients (Brown & Frank, 2003, 283-290) (Ricco & Westby, 2020, 91-98).

Based on these words, we created new variables to input into our dataset with the purpose of aiding our prediction model. These features were added through the representation of the number of times each of these words appeared in the text for a given sample.

4.1.2. Diabetes

For diabetes, we tried adding insulin and glucose as words of interest. Ketoacidosis is another relevant word which we can separate into keto and acido to see if we can capture more information.

Other keywords associated with diabetes are: obesity, carbohydrates, overweight, polyphagia, polydipsia, polyurea. These words were also created as new variables and added to the feature list for predicting with the prediction model.

4.2. Laboratory data feature engineering

As the lab results are not constant for each patient and thus we cannot create a consistent time series dataset, we tried to extract as much information about the lab tests that we could. For this, we decided to include the maximum and average difference in dates between the lab tests performed on each patient. Additionally, we also included the top lab test performed per patient, as well as the number of times that lab test was performed. Finally, we also added the first and last exam date on the dataset for each patient, as an integer starting from The Unix epoch, as well as the difference in days between these two variables.

Similarly to what was done in 4.1 Text-based Feature Engineering, we used relevant keywords for looking for specific laboratory tests performed on the patients, using terminology found for patients that suffer from syphilis (Brown & Frank, 2003, 284) (Goh, 2005, 450) (LaFond & Lukehart, 2006, 30) and diabetes (Diabetes Canada Clinical Practice Guidelines Expert Committee, 2018) (Nathan, 2015, 1052-1053).

Examples of the keywords used to look for specific tests are HIV, glucose, lymphocytes, CD3, CD4, and CD8. The complete list of the keywords we looked for can be seen in Table 19, with the results of those counts in Fig. 31.

Keywords	Feature name	Notes
`hepatitis | hepat | glutamic | bilirrub`	`'liver_damage'`	Related to liver damage
`hemo | hema`	`'hematic_info'`	Related to blood factors
`bacilo | bacter | colora | gram | tinc`	`'bacterias'`	Related to bacteria coloration/tinction, i.e. presence of bacterias
`tiroi | protro | tirox`	`'hormones'`	Related to hormones
`herpes | tuberc`	`'other_diseases'`	Tests for other conditions

album | creat | ureico | urico | uro | orina	'kidney_damage'	Related to kidney damage
colest | trigli | plaqu | protrom | trombo	'heart_damage'	Related to heart damage
calcio | fofs | pot	'minerals'	Related to mineral content
leuco | linfo | cd3 | cd4 | cd8 | anticuerpo | antigen | neutrof	'white_cells'	Related to white cells
deficiencia | vih | immuno	'vih'	Related to HIV
ayun | gluco | glico	'diabetes_tests'	Related to diabetes tests
trepo | anal | virus | viral	'syphilis_tests'	Related to syphilis tests

Figure 31. Boxplot for the different keyword counts versus the disease class

5. FRONTEND

A web page was designed by the team, with access to a dashboard with general information about our data, and a form that allows users to input data and get back a prediction for their condition using a pre-trained ML model, further explained in Section 7. The frontend was built using Javascript, specifically using the React framework, which in turn allows the web page to be responsive and interactive. The frontend includes three main pages: the Homepage, the Medical predictions page and the Dashboard.

5.1. Homepage

In the homepage, the project description is presented, mentioning the components of our project, its scope, the datasets used, the members of the team, etc. Appendix 10.1 shows a mockup of this homepage.

5.2. Medical predictions page

In this page, a user can fill out a form for which in return the web page will make an API call to the backend and get back a prediction made from the data supplied by the user. The data to provide is based on the features in the datasets, and then we perform the same feature engineering to the supplied sample according to the model training, validation and testing. The model will then show the user the predicted outcome for the given data, with the ICD-10 code and the name of the predicted diagnosis. A mockup of how this section looks can be found in Appendix 10.2.

5.3. Dashboard page

Finally, a dashboard was also designed, including some of the results obtained in the Exploratory Data Analysis presented in Section 3. The dashboard mainly contains the multivariate analysis between our different classes and the data in our datasets, e.g. the ratio for the different marital statuses present in our data against the disease codes. A mockup of this section can be found in Appendix 10.3.

6. API ARCHITECTURE

Following the requirements from IQVIA, we built and deployed an API that can be used from both the frontend and another client if necessary, in order to only obtain the prediction from the model without going through a frontend.

To accomplish this we used the Django framework, which follows the model–template–views (MTV) architectural pattern. With the help of this framework we made a RESTful API in order to be able to communicate the results of the model in a user friendly format, i.e. in a json format. It is worth mentioning that a RESTful API uses HTTP requests to access and use data to GET, PUT, POST and DELETE, i.e., reading, updating, creating, and deleting operations concerning resources, which in this case correspond to the ones related to the data model built.

To access these resources either by the frontend or other types of clients we defined different endpoints that can be identified by the uniform resource identifier (URI).

In summary, our REST API, which has different endpoints, can be requested by the frontend or a different client and will give a response in a json format. An example of this process can be seen in Figure 32.

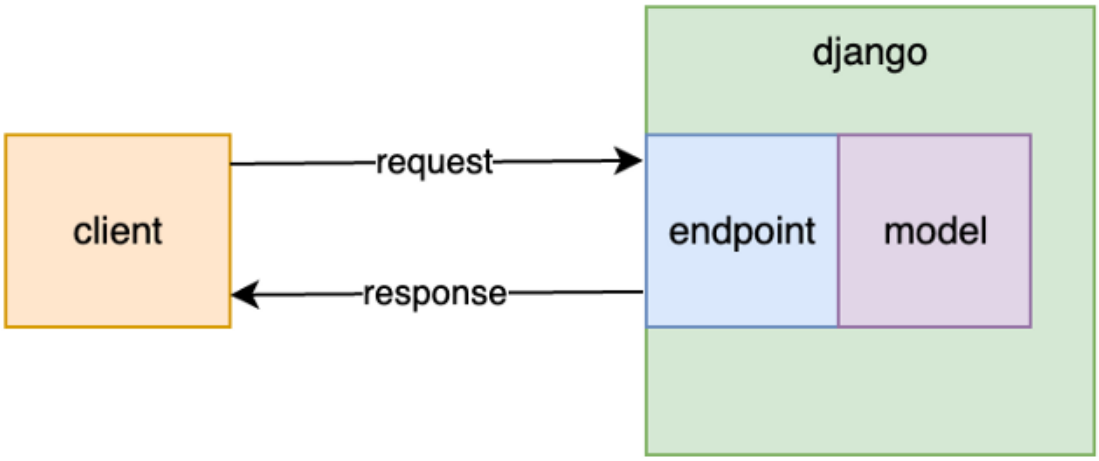


Figure 32. API Architecture.

6.1. First endpoint

As a first endpoint we have decided to implement a post method that will have as entries the next attributes: Age, Gender, Ethnic Group, Residential Area, Marital Status, Blood Type, Type and Plan. Additionally, it will receive all the information about the laboratory test of the patient as the name of the test, value and the date it was performed.

The URI format defined to call this endpoint is https://localhost:8000/api/post.

Once this post method is called, the API will send these attributes as a json format to the backend which will use the ML model to preprocess the data. The server will then respond with the type of diabetes or syphilis the model predicts the patient has based on the attributes sent, in addition to a 200 http response. If the query is incorrect, a response in the 400's range for access errors or 500's for server problems, with the details of the error, will be returned.

6.2. Dashboard endpoints

The following endpoints that were implemented were get methods that will return the necessary information to build the Dashboard described in 5.3. This information will be a summary of the relationship between the different sociodemographic, lab and EHR features and the disease types. We decided to not build a database in our system as that was a possible privacy issue due to the nature of our data, as it has a text field with possibly delicate information. Even though IQVIA performed an excellent job at removing sensitive information, there is no guarantee that the data is safe to be shared online.

The URI format defined to call these endpoints are:

- https://localhost:8000/api/socio_economics, for the sociodemographic data
- https://localhost:8000/api/laboratory, for the laboratory data
- https://localhost:8000/api/notes, for the notes data

Once this get method is called, the api will send the data as a json format to the frontend in addition to an http response of 200 for a correct access. Similar to the first endpoint, it will return an http error in the 400's or 500's depending on the error and issues with the call.

7. MACHINE LEARNING (ML) MODELS

A machine learning pipeline based on Scikit-learns pipeline approach was developed (Scikit-learn, 2022). This allowed us to test different models and combinations of inputters, scalers, feature selection algorithms, and estimators, as well as different parameters that can be passed to the already mentioned methods, using Scikit-learn's RandomizedSearchCV, which randomly selects hyperparameters from a given hyperparameter feature space and computes their K-fold cross-validation results based on selected metrics. The best performing model is then retrained with all of the train data, and scored against test data unseen by the model.

7.1. Methodology

The data was preprocessed and merged based on their respective IDRecord. The feature engineering techniques mentioned in *4. FEATURE ENGINEERING* were applied to the data as part of the preprocessing step. A train-test split of 80% was used to train and validate the models, with the remainder used to test its performance. Class merging, as described in Target Feature Analysis, was performed in order to try to correct the class imbalance present in the data and thus increase the performance of the model, however this was not very effective and was replaced with undersampling and oversampling techniques to the training component of the dataset instead. The remaining test data of the train-test split was not oversampled, which then was used to calculate a classification report, containing the precision, recall and F1-Score, and a confusion matrix por the true vs predicted classes of the test data.

7.2. Traditional ML models

A model pipeline, consisting of a preprocessing step, feature selection step, and finally an estimator step, was created using Scikit-learn's pipeline approach. An example of one of these model pipelines can be seen in Figure 32.

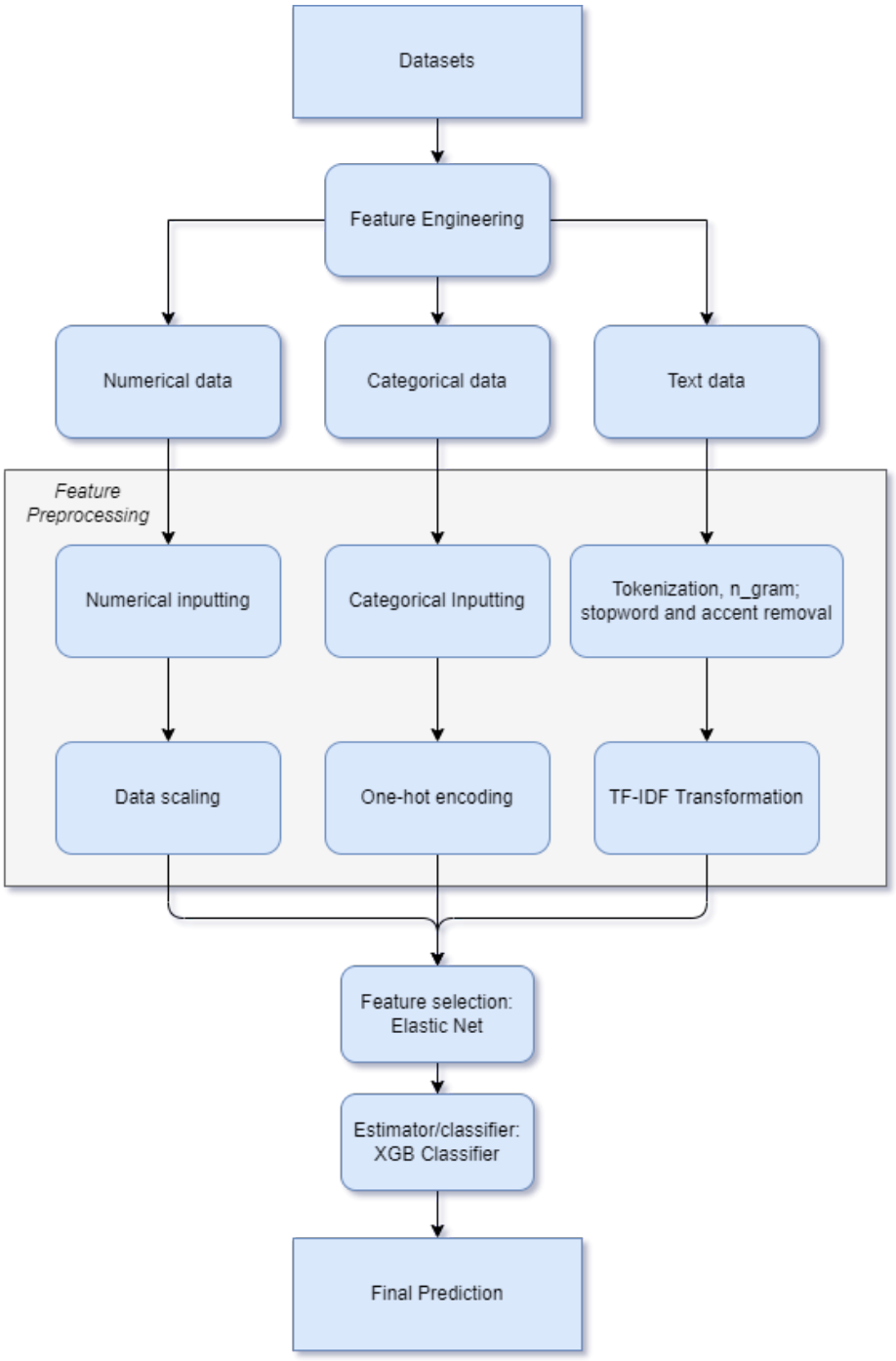


Figure 32. Example of an ML pipeline.

7.2.1. Preprocessor

The preprocessing step contains three transformers, one for categorical, one for numerical, and one for textual data. For the numerical data, a data imputer was used in order to convert missing numeric data with a specific technique, followed by a scaler to normalize the numerical data.

For the categorical data, a missing data inputter that replaces missing data with 'unavailable', followed by one-hot encoding was used in order to normalize this type of data.

Finally, for the text preprocessing, tokenization, stop-word removal, accent stripping, n_gram conversion, and finally TF-IDF was performed on the resulting n_grams.

Different inputting and scaling techniques, as well as different sets of n-grams for the text data were tested, and a summary of all the combinations tried can be seen in Table 20.

Type of variable	Type of preprocessor	Parameter tested
Numerical	Inputter	NaN replacement with the mean
		NaN replacement with the median
		NaN replacement with the mode
		K-Nearest Neighbors inputter
	Scaler	Normalizer[2]
		Standardizer
		Robust Scaler[3]
		MinMax scaler
Text	Tokenizer	1 to 1 n-grams
		1 to 2 n-grams
		1 to 3 n-grams
		2 to 3 n-grams

[2] Based on the technique described in: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.Normalizer.html#sklearn.preprocessing.Normalizer

[3] Based on the technique described in: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.RobustScaler.html

		3 to 3 n-grams
		4 to 4 n-grams
	TF-IDF	With l1 norm
		With l2 norm

Table 20. List of preprocessors and parameters tested on the preprocessing step.

7.2.2. Feature selector

Feature selection was also performed in the data, using an ElasticNet, a Ridge, or a Variance Threshold approach to find the best performing features of the dataset, and discard the non-performing ones. Variance Threshold, as the name implies, works by removing features with a low variance in the data. At the same time, we can take advantage of Elastic Net's and Ridge's intrinsic regularization approach to find the most important features that allow us to correctly predict a class.

7.2.3. Estimator/classifier

An XGB classifier, as well as a Support Vector Machine (SVM), were used for the classification of the labels of our samples. Due to training time constraints, a maximum tree depth of 15 levels was used in the model. Each one of the estimators were trained with multiple different hyperparameters, to try to find the best performing model. A 5-fold stratified cross-validation approach was used to test the performance of each of the models, and they were scored based on their F-1 Score to find the best performing model. Table 21 contains a list of the different hyperparameters tested on the pipeline.

Estimator	Hyperparameter name	Ranges/values tested
SVM	C	0 to 2
	kernel	linear, poly, rbf, sigmoid
	gamma	auto or scale
	class_weight	balanced
	coef0	0 to 2
	degree	1 to 5
XGB Classifier	n_estimators	1 to 200
	max_depth	2 to 10, and maximum

	eta	0.01 to 0.5
	min_child_weight	0.5 to 20
	gamma	0 to 1
	subsample	0.1 to 1
	colsample_bytree	0.2 to 1
	reg_lambda	0 to 12
	reg_alpha	0 to 12

Table 21. List of estimators and parameters tested on the classification step.

7.3. Deep Learning models

For preprocessing the data for this model, the same approach that was described in 7.2.1 was used. The pipeline was trained with hyperparameter tuning, and the resulting preprocessor was extracted to be used for preprocessing the data for the Deep Learning model.

A Deep Learning model was also developed, comprising two branches, one for the numerical and categorical data and the other for the text data. For the non-text branch, an overlapping set of dense and dropout layers was used, consisting of 3 dense and 2 dropout layers. The text branch consists of an embedded layer, for which Google's nnlm-es-dim128-with-normalization embed layer was used[4], followed by 2 dense layers with a dropout layer following each of the dense layers.

Afterwards, a concatenation layer merges the neurons of both of the branches into one, followed by a dense, a dropout, and another dense layer, and ends up on a fully connected dense layer that works as the output layer. All of the dropout layers used a dropout value of 0.5, and the dense layers used a relu activation function. The model was optimized using Adam, with a categorical cross entropy loss function used for both the loss and validation of the model. Figure 34 shows a representation of the model described before.

[4] A description of this layer can be found in https://tfhub.dev/google/nnlm-es-dim128-with-normalization/2

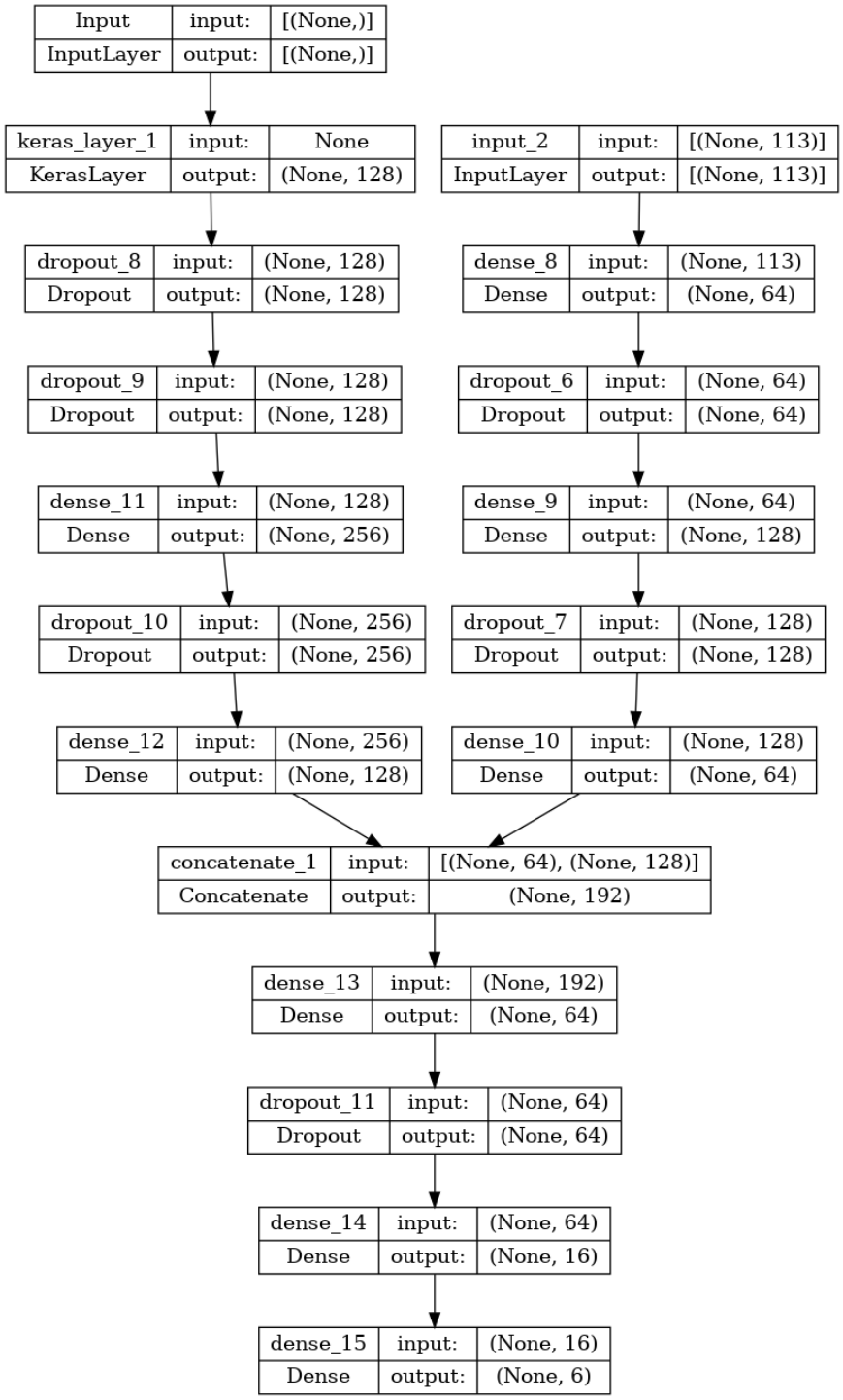


Figure 34. Example of a Deep Learning model trained.

8. RESULTS AND DISCUSSION

8.1. Class merging

As can be seen on Table 22, with the model described in 7. MACHINE LEARNING (ML) MODELS, we can obtain an average F1-Score for the test dataset of 65%, and Figure 35 shows the classification matrix of our model. Additionally, Table 23 shows the classification report for our Deep Learning model. These results seem to indicate that merging the classes does allow for a good performance of the model, yet we can also try using over and undersampling in order to see if we can improve the recall capacity of our system, which is somewhat low for E109, E149, and A51, and which would allow us to still discern between the different classes that would be under the consolidated A51.

	precision	recall	f1-score	support
A51	0.95	0.63	0.76	2948.00
A530	0.89	0.92	0.90	48469.00
A539	0.89	0.89	0.89	39503.00
E109	0.83	0.68	0.75	5022.00
E119	0.83	0.89	0.86	13950.00
E149	0.91	0.50	0.65	2246.00
accuracy	0.88	0.88	0.88	0.88
macro avg	0.88	0.75	0.80	112138.00
weighted avg	0.88	0.88	0.88	112138.00

Table 22. Classification Report for the merged data, using the best performing ML model.

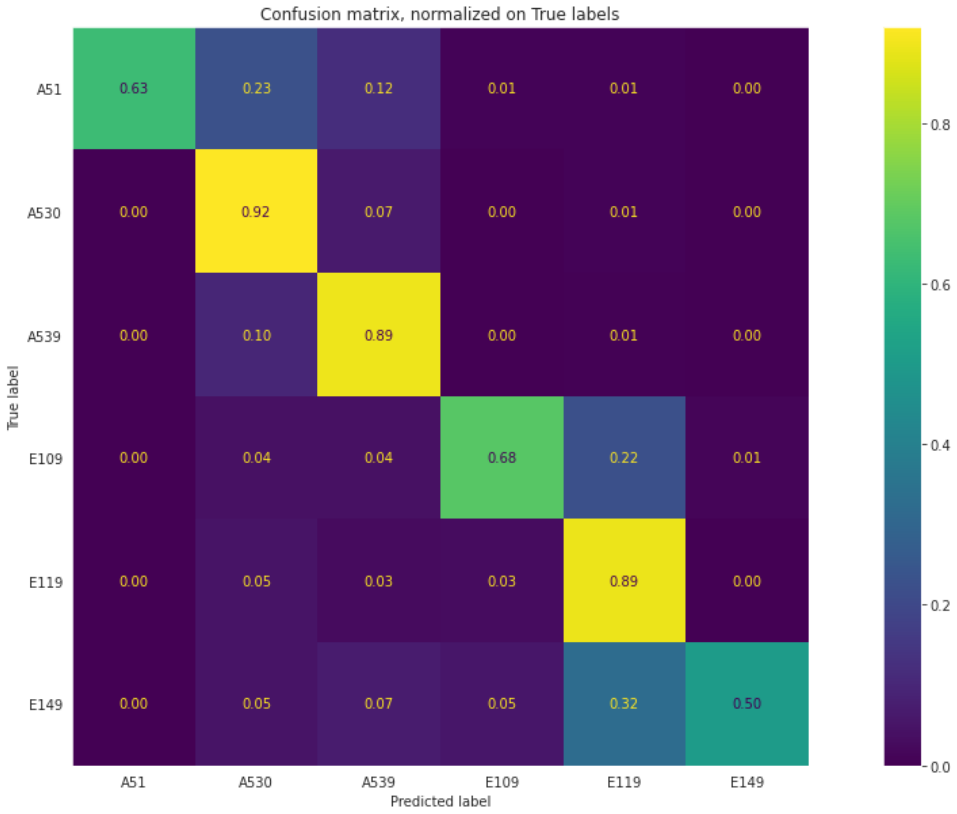


Figure 35. Classification Matrix for the merged data, normalized on the true labels, using the best performing ML model.

	precision	recall	f1-score	support
A51	0.04	0.00	0.00	2948.00
A530	0.66	0.83	0.74	48469.00
A539	0.73	0.62	0.67	39503.00
E109	0.59	0.36	0.45	5022.00
E119	0.65	0.68	0.66	13950.00
E149	0.00	0.00	0.00	2246.00
accuracy	0.68	0.68	0.68	0.68
macro avg	0.44	0.42	0.42	112138.00
weighted avg	0.65	0.68	0.66	112138.00

Table 23. Classification Report for the merged data, using the Deep Learning model.

8.2. Over and Under sampling

Over and undersampling were then tested as methods for improving the performance of the model, as even though it would be repeated data, it forces the model to see more samples of the minority class, similarly to increasing the weights of the underrepresented samples. Table 24 and Figure 36 show the classification report and confusion matrix, respectively, for undersampling. Additionally, Table 25 and Figure 37 are the classification report and confusion matrix for the best performing model after applying oversampling to the train and validation data. As can be seen on the classification reports, undersampling did not perform well on our test dataset, but there was a 10% increase on the macro accuracy of the model for the case of oversampling, with a high increase in the recall and thus F-1 score of the underrepresented classes.

	precision	recall	f1-score	support
A510	0.03	0.55	0.06	782.00
A511	0.04	0.99	0.09	75.00
A514	0.04	0.31	0.07	2091.00
A529	0.03	0.30	0.06	1576.00
A530	0.54	0.17	0.26	48469.00
A539	0.44	0.20	0.28	37927.00
E109	0.19	0.34	0.25	5022.00
E119	0.29	0.31	0.30	13950.00
E149	0.10	0.37	0.16	2246.00
accuracy	0.22	0.22	0.22	0.22
macro avg	0.19	0.39	0.17	112138.00
weighted avg	0.43	0.22	0.26	112138.00

Table 24. Classification Report for the undersampled train dataset.

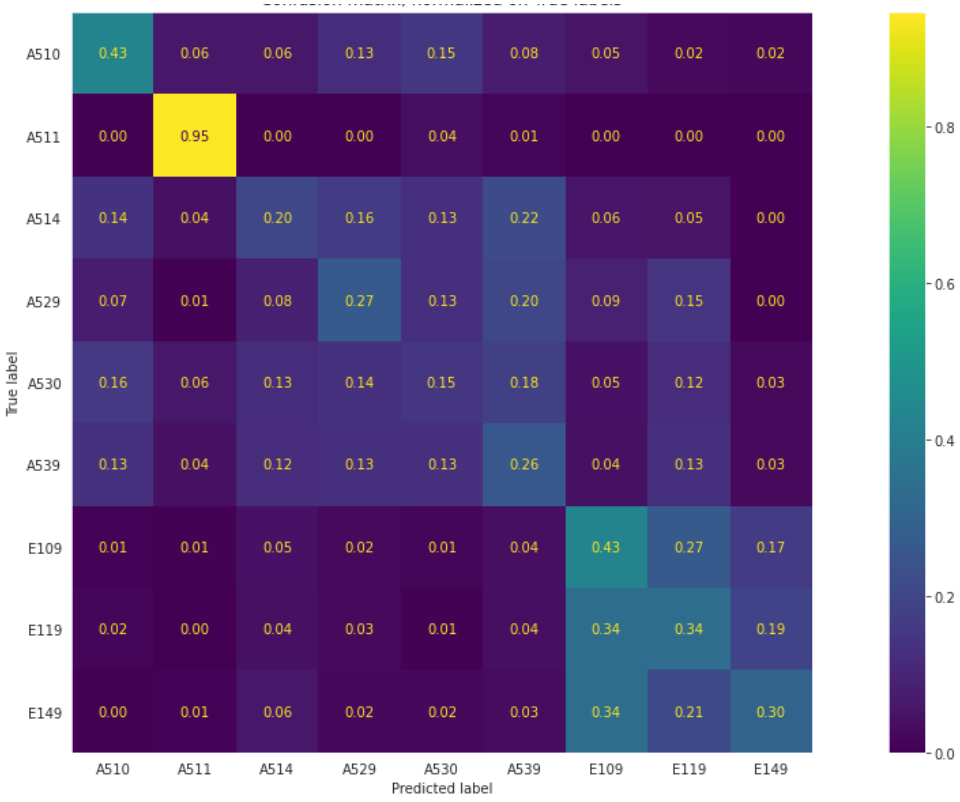


Figure 36. Confusion Matrix, normalized on the True labels, for the undersampled train dataset.

	precision	recall	f1-score	support
A510	0.84	0.84	0.84	782.00
A511	0.83	0.99	0.90	75.00
A514	0.62	0.84	0.71	2091.00
A529	0.76	0.78	0.77	1576.00
A530	0.91	0.90	0.90	48469.00
A539	0.90	0.88	0.89	37927.00
E109	0.76	0.86	0.81	5022.00
E119	0.86	0.89	0.88	13950.00
E149	0.82	0.81	0.81	2246.00
accuracy	0.88	0.88	0.88	0.88
macro avg	0.81	0.87	0.84	112138.00
weighted avg	0.89	0.88	0.88	112138.00

Table 25. Classification Report for the oversampled train dataset.

Figure 37. Confusion Matrix, normalized on the True labels, for the oversampled train dataset.

As can be seen in Figure 37, even after oversampling the minority classes the model still tends to have a bias for predicting the majority classes, i.e. A530 and A539, especially when the sample being predicted is A510. This could potentially be explained as well by the fact that A530 is unspecified syphilis, which might contain multiple cases of different types of syphilis that weren't correctly diagnosed, and thus the data might be contaminated.

The final best performing model was an XGBClassifier with an Elastic Net feature selector. Figure 38 shows an example of one of the trees, and Figure 39 shows the top 30 most important features for the dataset. As can be seen in the last figure,the model heavily relies on text features, and we can see that some words considered important by the model are clearly relevant for distinguishing between the different subclasses, like ritonavir (used for treating HIV/AIDS) and nutrition. Additionally, we can see that engineered features, like top lab name and other_diseases_count (which is the number of other diseases prescribed for the patient, e.g. for tuberculosis or hepatitis) are also helpful for discerning between the different subclasses in the system.

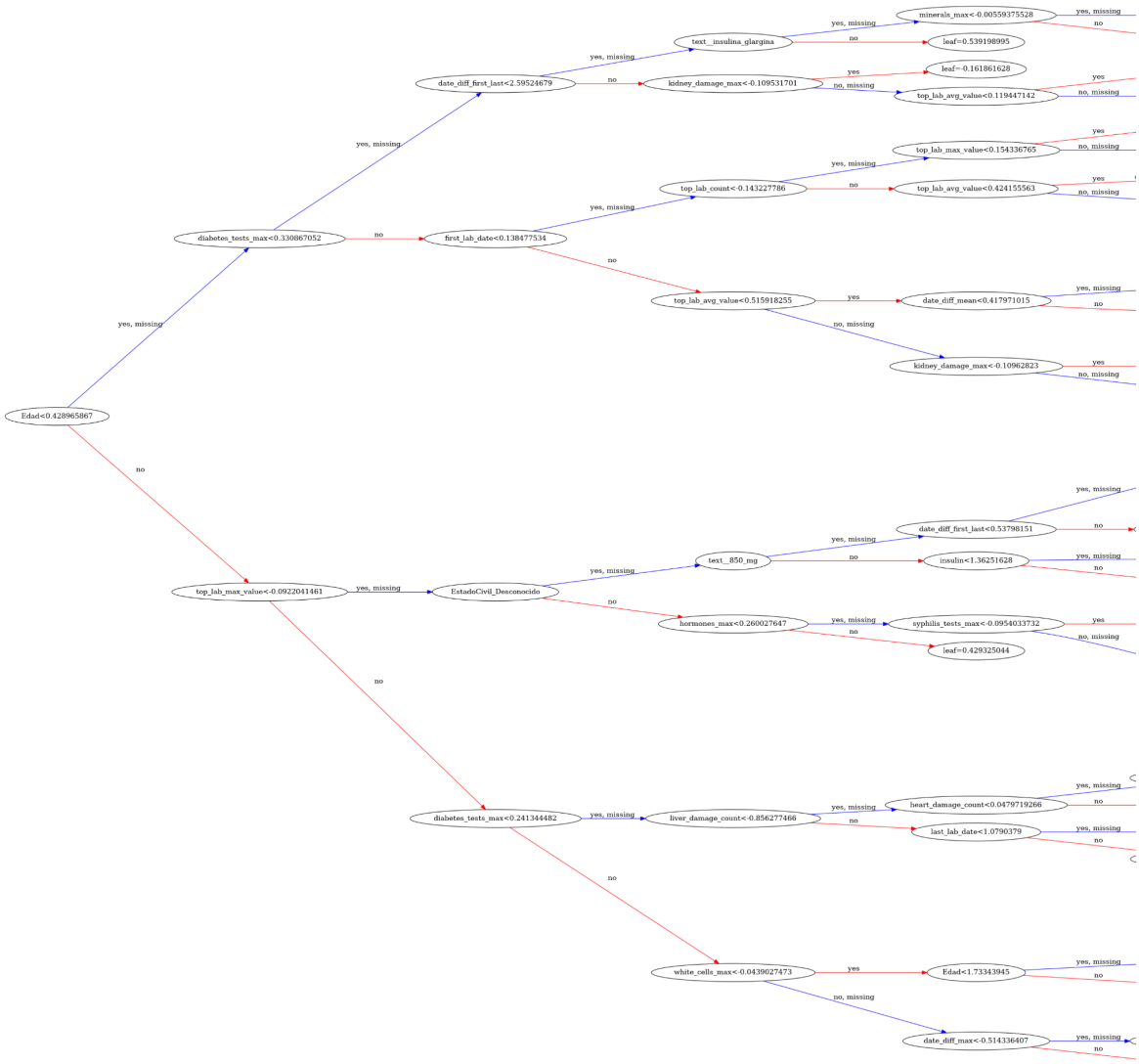


Figure 38. Example of one of the XGB decision trees of the ML model.

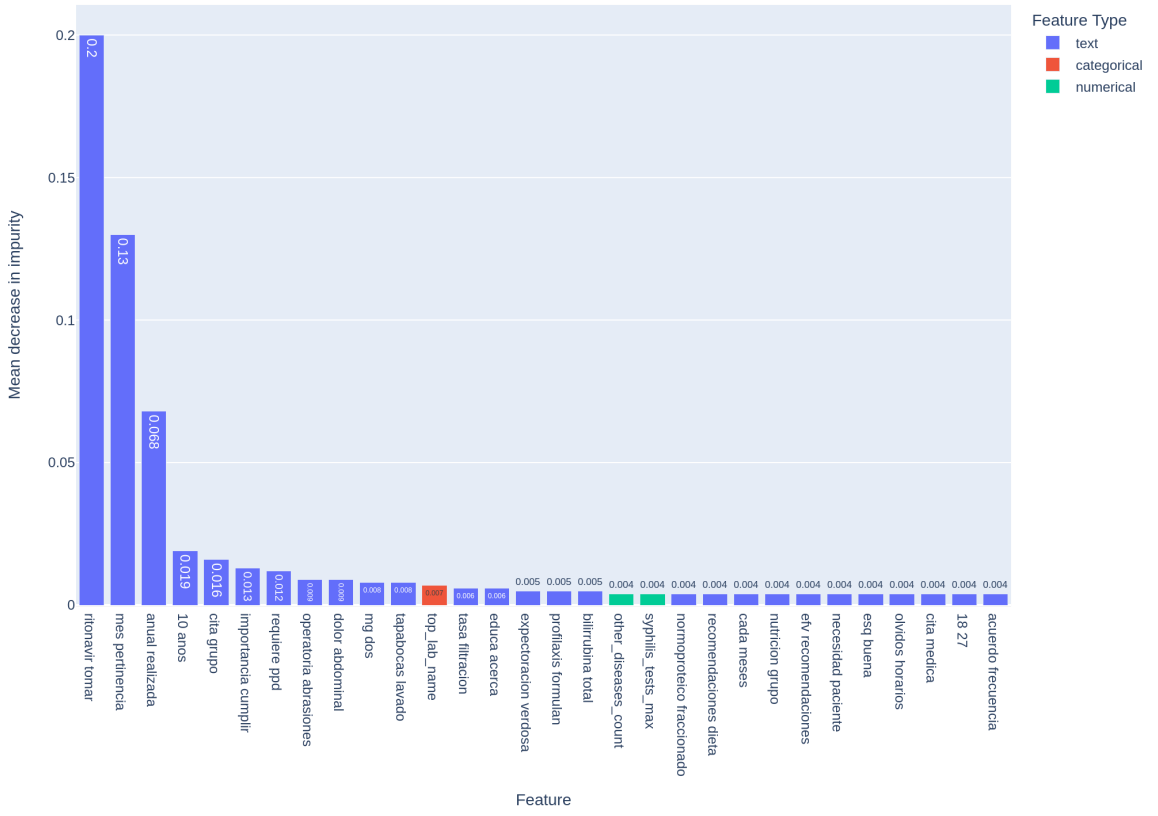


Figure 39. Feature importance to the model, ranked by their mean decrease in impurity of the decision tree.

9. CONCLUSIONS

In our quest toward cleaning Electronic Health Records and leveraging these resources to extract relevant information and predict specific pathologies or health conditions, this project shows the validity of using natural language processing and machine learning techniques as a model with 88% of F-score was obtained to distinguish between syphilis and diabetes diagnoses from medical notes. Furthermore, sociodemographic data like age, marital status, and gender, combined with EHRs are proven to be effective to differentiate between patients suffering from these diseases, whereas laboratory tests are also a powerful data source for helping in inter-class differentiation as well as intra-class differentiation, as shown in the EDA step of the project.

In addition to distinguishing between syphilis and diabetes with a high level of certainty, the model obtained can classify between the different types of syphilis and diabetes with the dataset supplied by IQVIA. Nonetheless, more samples containing data related to the minority class would greatly help increase the performance of the model.

Deep learning approaches did not outperform an XGBoost classifier with the data and features we had and developed for this iteration of the project. Whereas, undersampling did not work well with the data type we had available, oversampling permitted an increase in the recall capabilities of our model of approximately 10%. This technique was very important as one of the main challenges of the project was related to imbalanced data.

10. FUTURE WORK

Lemmatization, even though it is really slow to compute, could be performed on the dataset to reduce the feature space while still keeping the main features of the text input. Due to time constraints, only a total of around 30 models were trained through hyperparameter tuning. More models should be trained to guarantee that a higher percentage of the hyperparameter feature space is being covered and thus increase the certainty that the best parameters are being selected for the model training. Additionally, a deeper level for the XGBoost classifier should also be tested, as due to time constraints it was limited to only 15 levels, and it might not be taking into account the full extent of the features available for training.

The use of Recurrent Neural Networks, e.g. Long Short Term Memory layers should be considered for the model after the embedding layer, as this could also help the performance of the Deep Learning model. The use of more sophisticated Deep Learning models, such as BERT[5], or its spanish counterpart BETO[6] could also be considered.

Finally, further work should be done to see if the current pipeline built in this project can be applied to a wider range of diseases and classes, or if more data, e.g. body measurements like blood pressure and weight would be needed to further keep or improve the performance of the model with a larger set of prediction classes.

[5] A description of this model can be foun in https://github.com/google-research/bert/blob/master/multilingual.md
[6] An example of it can be found in https://github.com/dccuchile/beto

11. REFERENCES

Brown, D. L., & Frank, J. E. (2003). Diagnosis and management of syphilis. *American family physician*, *68*(2), 283-290. https://www.aafp.org/pubs/afp/issues/2003/0715/p283.html

Burahmah, J., Zheng, D., & Leslie, R. D. (2022). Adult-onset type 1 diabetes: A changing perspective. *European Journal of Internal Medicine*. 10.1016/j.ejim.2022.06.003

Centers for Disease Control and Prevention (CDC). (n.d.). *STD Facts - Syphilis (Detailed)*. Centers for Disease Control and Prevention. Retrieved June 27, 2022, from https://www.cdc.gov/std/syphilis/stdfact-syphilis-detailed.htm

Deshpande, A. D., Harries-Hayes, M., & Schootman, M. (2008, November). Epidemiology of Diabetes and Diabetes-Related Complications. *Physical Therapy*, *88*(11), 1254 - 1264. https://watermark.silverchair.com/ptj1254.pdf?token=AQECAHi208BE49Ooan9kkhW_Ercy7Dm3ZL_9Cf3qfKAc485ysgAAAsEwggK9BgkqhkiG9w0BBwagggKuMIICqgIBADCCAqMGCSqGSIb3DQEHATAeBglghkgBZQMEAS4wEQQMXetxIQBpfK-kR45BAgEQgIICdHjyIl8Do8nkutFp3KjQaxR7en7VfeJ374CrfljoAni_U4ya

Diabetes Canada Clinical Practice Guidelines Expert Committee. (2018). Diabetes Canada 2018 Clinical Practice Guidelines for the Prevention and Management of Diabetes in Canada. *Canadian Journal of Diabetes*, *42*(Suppl 1), S1 - S325. http://guidelines.diabetes.ca/cpg

Diabetes Canada Clinical Practice Guidelines Expert Committee, Lipscombe, L., Butalia, S., Dasgupta, K., Eurich, D. T., MacCallum, L., Shah, B. R., Simpson, S., & Senior, P. A. (2020). Pharmacologic Glycemic Management of Type 2 Diabetes in Adults: 2020 Update. *Canadian Journal of Diabetes*, *44*, 575 - 591. 10.1016/j.jcjd.2020.08.001

Diabetes Canada Clinical Practice Guidelines Expert Committee, McGibbon, A., Adams, L., Ingersoll, K., Kader, T., & Tugwell, B. (2018). Glycemic Management in Adults With Type 1 Diabetes. *Canadian Journal of Diabetes*, *42*, S80 - S87. 10.1016/j.jcjd.2017.10.012

Diabetes Canada Clinical Practice Guidelines Expert Committee, Punthakee, Z., Goldenberg, R., & Katz, P. (2018). Definition, Classification and Diagnosis of Diabetes, Prediabetes and Metabolic Syndrome. *Canadian Journal of Diabetes*, *42*, S10 - S15. http://guidelines.diabetes.ca/docs/cpg/Ch3-Definition-Classification-and-Diagnosis-of-Diabetes-Prediabetes-and-Metabolic-Syndrome.pdf

Forouhi, N. G., & Wareham, N. J. (2010). Epidemiology of diabetes. *Medicine*, *38*(11), 602 - 606. https://www.sciencedirect.com/science/article/abs/pii/S1357303910002082

Goh, B. T. (2005). Syphilis in adults. *Sexually Transmitted Infections*, *81*(6), 448-452. 10.1136/sti.2005.015875

Government of British Columbia, Canada. (n.d.). *Clock Drawing Test (CDT)*. Gov.bc.ca. Retrieved June 27, 2022, from

https://www2.gov.bc.ca/assets/gov/health/practitioner-pro/bc-guidelines/cogimp-clock-drawing-test.pdf

Harris, M. I. (1995). *Diabetes in America, 2nd Edition* (2nd ed.). National Diabetes Data Group 1995. https://books.google.ca/books?id=hcRrAAAAMAAJ

LaFond, R. E., & Lukehart, S. A. (2006). Biological Basis for Syphilis. *Clinical Microbiology Reviews*, *19*(1), 29-49. 10.1128/CMR.19.1.29-49.2006

Mayo Clinic. (2020, October 30). *Diabetes - Symptoms and causes*. Mayo Clinic. Retrieved June 28, 2022, from https://www.mayoclinic.org/diseases-conditions/diabetes/symptoms-causes/syc-20371444

Nathan, D. M. (2015, September 8). Diabetes: Advances in Diagnosis and Treatment. *Journal of the American Medical Association*, *314*(10), 1052-1062. 10.1001/jama.2015.9536

Organización Sanitas. (n.d.). *Test del reloj - Demencias*. Sanitas. Retrieved June 27, 2022, from https://www.sanitas.es/sanitas/seguros/es/particulares/biblioteca-de-salud/tercera-edad/demencias/test-reloj.html

Pan American Health Organization. (2019). *Enfermedades infecciosas y parasitarias*. Ciertas enfermedades infecciosas y parasitarias (A00–B99). Retrieved June 28, 2022, from http://ais.paho.org/classifications/chapters/CAP01.html?zoom_highlight=a51

Ricco, J., & Westby, A. (2020). Syphilis: Far from Ancient History. *American family physician*, *102*(2), 91-98. https://www.aafp.org/pubs/afp/issues/2020/0715/p91.html

Scikit-learn. (2022). *6.1. Pipelines and composite estimators — scikit-learn 1.1.1 documentation*. Scikit-learn. Retrieved June 28, 2022, from https://scikit-learn.org/stable/modules/compose.html

Scully, T. (2012, May 12). Diabetes in Numbers. *Nature*, *485*, S2 - S3. https://www.nature.com/articles/485S2a.pdf?origin=ppub

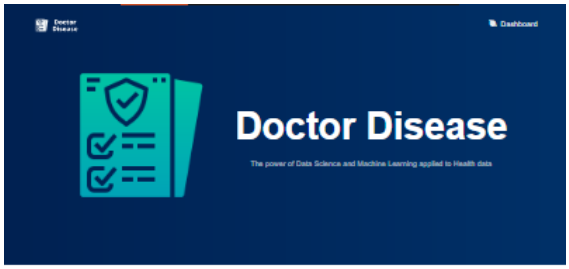
World Health Organization. (2016). *Global Report on Diabetes* (G. Roglic, Ed.). World Health Organization. https://apps.who.int/iris/bitstream/handle/10665/204871/9?sequence=1

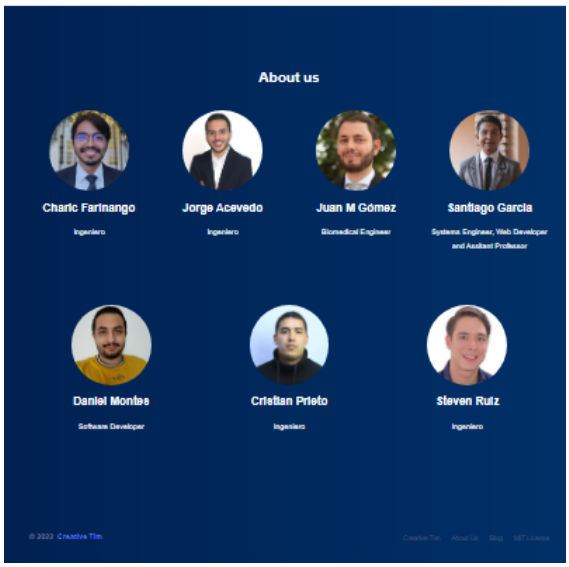
World Health Organization. (2019). *ICD-10 Version:2019*. ICD-10 Version:2019. Retrieved June 28, 2022, from https://icd.who.int/browse10/2019/en#/A51

Zimmet, P. Z., Magliano, D. J., Herman, W. H., & Shaw, J. E. (2014). Diabetes: a 21st century challenge. *The Lancet Diabetes & Endocrinology*, *2*(1), 56 - 64. 10.1016/S2213-8587(13)70112-8

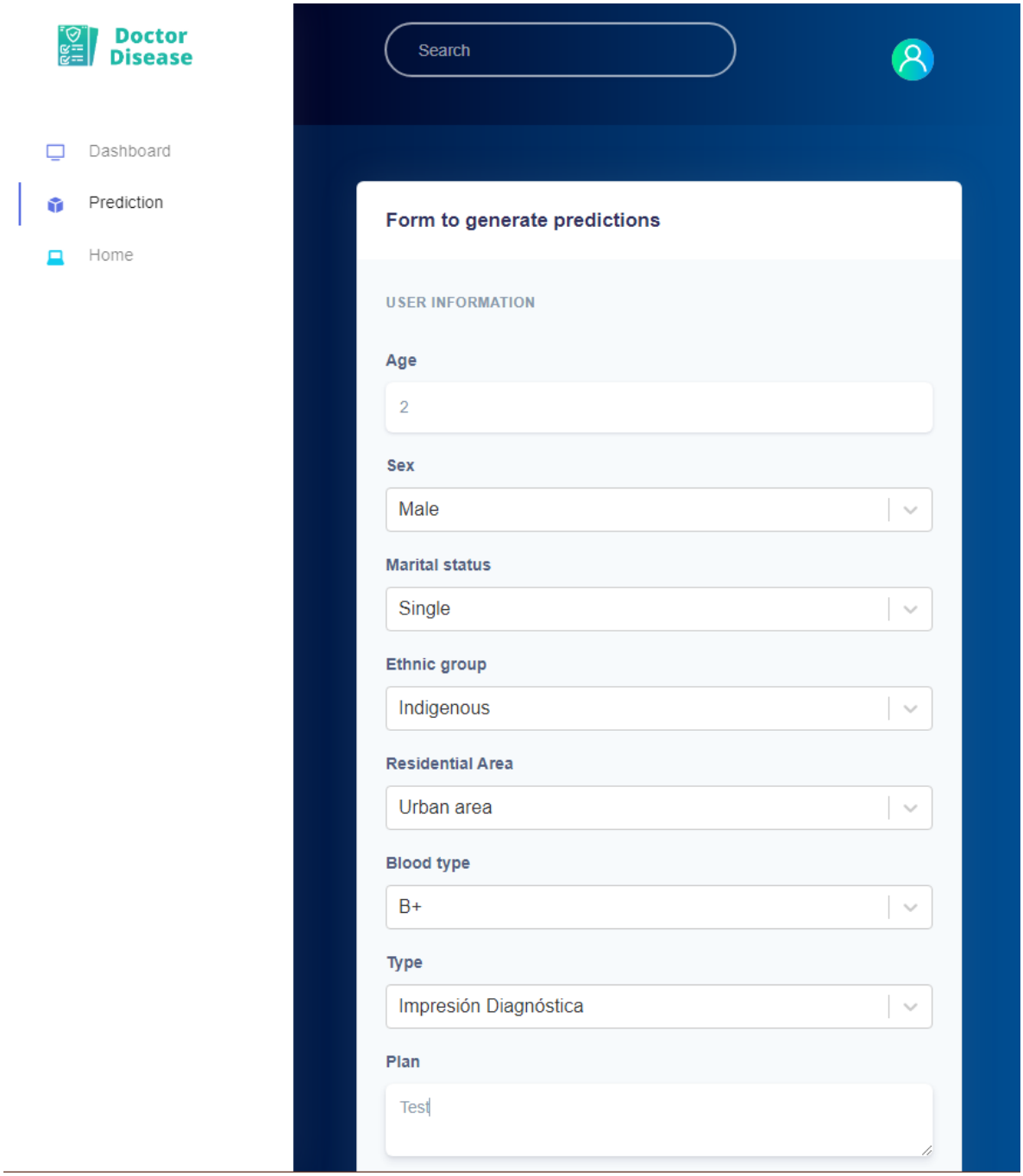
12. APPENDIX

12.1. Homepage Mockup





12.2. Data Prediction Page Mockup



12.3. Dashboard Page Mockup