Pontificia Universidad Católica del Ecuador

Facultad de Ingeniería



TEMA:

Modelo de predicción de enfermedades ocupacionales utilizando técnicas de aprendizaje automático y análisis de datos.

AUTOR:

Priscila Margarita Guartatanga Guartatanga

TUTOR:

Msg. Christian Andrés Rueda Ayala

TRABAJO PREVIA A LA OBTENCIÓN DEL TÍTULO DE MAGISTER EN SISTEMAS DE INFORMACIÓN MENCIÓN DATA SCIENCE

Quito, febrero 2024

RESUMEN

El presente trabajo tiene como objetivo desarrollar un modelo de predicción de enfermedades ocupacionales mediante la utilización de técnicas de Aprendizaje Automático y análisis de datos. Este trabajo se centra en el análisis de datos de exámenes de sangre y diagnósticos médicos clasificados con la Codificación Internacional de Enfermedades CIE 10ª edición. El modelo pretende ser una herramienta que permita analizar los datos y generar conocimiento para la toma de decisiones.

Para el desarrollo de este trabajo se utilizó la metodología CRISP-DM, comprendiendo las necesidades del área de salud ocupacional y los datos de los que disponen, preparando los datos para finalmente generar y evaluar el modelo.

Los algoritmos que se utilizaron para este trabajo fueron Bosques Aleatorios y el Máquina de vectores de soporte (SVM). El desarrollo se realizó con Python utilizando Jupyter Notebook y las librerías Pandas, Sklearn, Numpy, etc.

Contenido

CAPÍTULO I: INTRODUCCIÓN 4
1. INTRODUCCIÓN 4
1.1. Antecedentes 4
1.2. Planteamiento del problema 4
1.3. Objetivos 5
1.4. Alcance 6
CAPITULO II: FUNDAMENTACIÓN TEÓRICA 6
2. MARCO TEÓRICO Y CONCEPTUAL 6
2.1. Salud Ocupacional 6
2.2. Control médico ocupacional 7
2.3. Exámenes de Laboratorio 7
2.4. Clasificación Internacional de Enfermedades CIE 8
2.5. Aprendizaje Automático 9
2.6. Métodos de Aprendizaje Automático 9
2.7. Metodología de Minería de Datos 10
CAPITULO III: MARCO METODOLÓGICO 12
3. Metodología 12
3.1. Comprensión del Negocio 13
3.2. Comprensión de Datos 15
3.3. Preparación y muestreo de datos 22
3.4. Modelado 29
3.4.1. Selección del modelo 29
3.4.2. Diseño de pruebas del modelo 29
3.4.3. Construcción y evaluación del modelo 29
3.5. Evaluación 32
4. Conclusiones y Recomendaciones 33
4.1. Conclusiones 33
4.2. Recomendaciones 34
BIBLIOGRAFÍA 36

Índice de Figuras

Figura 1:Diagrama Entidad - Relación 16
Figura 2: Distribución de los registros por Año. 18
Figura 3: Distribución de los registros por Cargo. 19
Figura 4: Distribución de los registros por Género 19
Figura 5: Distribución de los registros por Estado Civil. 20
Figura 6: Distribución de la variable Diagnósticos CIE-10. 20
Figura 7: Análisis de variables categóricas. 21
Figura 8: Datos de Exámenes de Sangre 25
Figura 9: Datos de Exámenes de sangre - Nuevo archivo. 25
Figura 11: Tratamiento variable PSA. 28
Figura 12: Tratamiento de variables, T3, T4 y TSH. 28
Figura 13: Tratamiento de variables KNN. 29
Figura 14:División dataset Bosques Aleatorios 30
Figura 15: Entrenamiento del modelo Bosques Aleatorios. 30
Figura 16: Entrenamiento del modelo Bosques Aleatorios mejores hiperparámetros. . 31
Figura 17: Entrenamiento del modelo SVM 31
Figura 18: Entrenamiento del modelo SVM mejores hiperparámetros. 32

Índice de Tablas:

Tabla 1: Descripción atributos de Exámenes médicos. 17
Tabla 2: Descripción de atributos de consultas médicas y diagnósticos 17
Tabla 3: Nuevos cargos. 23
Tabla 4:Tipos de examen unificados. 23
Tabla 5: Género 24
Tabla 6: Estado Civil. 24
Tabla 7: Descripción de atributos de datos integrados. 27
Tabla 8: Diagnósticos CIE-10 27

CAPÍTULO I: INTRODUCCIÓN

1. INTRODUCCIÓN

1.1. Antecedentes

La salud ocupacional se ha constituido en un tema prioritario para las empresas, la prevención de enfermedades y la promoción de un entorno de trabajo seguro y saludable son fundamentales para el bienestar y el desempeño eficiente de los empleados y trabajadores.

Las enfermedades ocupacionales representan un desafío en varios sectores como el industrial, comercial, de servicios, de la construcción, etc. debido a que los empleados y trabajadores están expuestos a diversos riesgos de seguridad tales como: biológicos, físicos, químicos, ergonómicos, y psicosociales por lo que es importante contar con herramientas que permitan predecir enfermedades en base a un análisis de datos existentes que permitan tomar decisiones informadas y realizar acciones preventivas promoviendo un entorno laboral saludable y productivo, fomentando un ambiente de prevención y de fortalecimiento de los sistemas de gestión de seguridad y salud ocupacional en beneficio de los empleados y trabajadores.

1.2. Planteamiento del problema

El problema principal que afrontan las áreas de Seguridad y Salud Ocupacional es la dificultad de implementar planes de medicina preventiva de manera efectiva, basados en un análisis adecuado de datos y diagnósticos médicos. La carencia de este análisis de datos dificulta la identificación temprana de enfermedades y la implementación de medidas preventivas, por lo que con el tiempo puede afectar la salud y el bienestar de sus empleados y/o trabajadores.

La falta de acciones medicas preventivas puede causar un impacto negativo en la calidad de vida de los empleados, disminución de la productividad y eficiencia en el trabajo. Además, las empresas pueden verse afectadas económicamente por compensaciones laborales y afectaciones en la planificación y la continuidad en el trabajo por ausentismo laboral prolongado o frecuente.

1.3. Objetivos

1.3.1. Objetivo General

El objetivo general es desarrollar un modelo de predicción de enfermedades ocupacionales que contribuya a la toma de decisiones informadas dentro del área de salud ocupacional. Los datos para analizar corresponden a exámenes de sangre y diagnósticos médicos de los periodos comprendidos entre los años 2019 al 2023.

1.3.2. Objetivos Específicos

- Realizar la extracción de datos de exámenes de sangre y diagnósticos médicos del correspondientes a los empleados activos de la empresa hasta la fecha de extracción.
- Analizar los datos relevantes para identificar patrones y tendencias que puedan estar asociadas con enfermedades ocupacionales.
- Seleccionar y aplicar técnicas de aprendizaje supervisado para generar un modelo de predicción de enfermedades.

1.4. Alcance

Para este estudio se dispone de los datos de exámenes de sangre y diagnósticos médicos de los empleados activos, correspondientes a los años 2019, 2020, 2021, 2022 y 2023. Por lo general el control médico se ejecuta a inicios de año, pero por una situación particular en el año 2022 el control médico se realizó a finales de año, distribuyéndose los datos entre el 2022 y 2023. Los diagnósticos médicos están catalogados con código CIE-10 (Clasificación Internacional de Enfermedades 10ª Edición).

Como política de seguridad ocupacional las empresas realizan exámenes y diagnósticos médicos pre ocupacionales para la selección del personal idóneo para el cargo y a partir de siguiente año de trabajo, el empleado/trabajador, se realiza el control médico ocupacional, así como también al finalizar su periodo laboral en la institución.

CAPITULO II: FUNDAMENTACIÓN TEÓRICA

2. MARCO TEÓRICO Y CONCEPTUAL

2.1. Salud Ocupacional

La salud ocupacional se enfoca en promover y mantener el bienestar físico, mental y social de los trabajadores. Sus objetivos principales se centran en el mantenimiento, promoción de la salud y la capacidad laboral, mejorar las condiciones y el entorno de trabajo además fomentar el desarrollo de una organización que incorpore sistemas de gestión eficaces, políticas y prácticas de seguridad y salud en el trabajo.

Dentro de la salud ocupacional se involucra disciplinas como la seguridad en el trabajo, la higiene industrial, ergonomía, psicología, medicina del trabajo, enfermería ocupacional, estadística, legislación laboral, terapia ocupacional, nutrición y promoción de la salud.

2.2. Control médico ocupacional

El control médico ocupacional es un conjunto de procedimientos y evaluaciones médica dirigidas a los empleados/trabajadores, con el objetivo de prevenir, detectar y tratar enfermedades relacionadas con el trabajo, este proceso es continuo e inicia con el control médico pre ocupacional, continua con el control periódico y al finalizar su relación laboral se realiza el control médico post ocupacional.

En el control médico ocupacional se pueden realizar evaluaciones clínicas, pruebas de laboratorio, exámenes de imagenología, exámenes visuales, evaluaciones ergonómicas, etc. La implementación efectiva de controles médicos ocupacionales beneficia tanto a los empleados como a los empleadores ya que promueve un ambiente laboral seguro y saludable y ayuda a reducir las enfermedades y accidentes laborales mejorando la productividad y el entorno laboral.

2.3. Exámenes de Laboratorio

Entre las evaluaciones que se realizan en el control médico están los exámenes de laboratorio tales como exámenes de sangre, orina, heces, y otros más específicos como los hormonales, etc.

La realización de un análisis de sangre es fundamental para determinar el estado de salud general de una persona, identificar patologías o agentes patógenos que puedan estar afectando su bienestar, así como para dar seguimiento de la evolución de las enfermedades y evaluar la eficacia de los tratamientos aplicados. Así mismo, es clave para verificar el correcto funcionamiento de órganos vitales como el hígado, riñones, tiroides, corazón, etc. y para la detección temprana de enfermedades como diabetes, anemia, cáncer y enfermedades infecciosas.

En un análisis de sangre se puede realizar hemogramas completos, análisis de bioquímica sanguínea/perfil metabólico, análisis de enzimas en sangre para evaluar la función hepática, pruebas de función renal, lípidos en sangre, pruebas de tiroides, niveles de azúcar, etc.

2.4. Clasificación Internacional de Enfermedades CIE

De acuerdo con la Organización Mundial de la Salud (OMS), la Clasificación Internacional de Enfermedades CIE, desempeña un papel importante a nivel global, ofreciendo información esencial sobre el alcance, los orígenes y las consecuencias de las enfermedades y causas de muertes a través de los datos recopilados y categorizados bajo esta clasificación.

Esta herramienta asigna a cada condición de salud una categoría específica y un código alfanumérico, lo que simplifica la organización de la información y apoya el análisis estadístico en áreas como la epidemiología, la administración de salud, entre otros aspectos. La recopilación y el análisis de estos datos son fundamentales para la planificación de servicios de salud, la gestión de la calidad y la seguridad, así como también la investigación en el ámbito sanitario. Además,

el uso de estos códigos para diagnosticar enfermedades estandariza la recogida de información, facilitando la investigación a gran escala.

2.5. Aprendizaje Automático

El aprendizaje automático es una rama de la inteligencia artificial (IA) y la informática, que utiliza datos y algoritmos para emular la manera en que los humanos aprenden, mejorando de manera gradual su precisión (Adams, 2020).

El aprendizaje automático surge como respuesta a la demanda de analizar y aprovechar grandes conjuntos de datos con el objetivo de predecir comportamientos y tendencias. Esto se logra a través del uso de algoritmos que no son secuencias de comandos creadas para convertir datos específicos de entrada en un resultado esperado.

2.6. Métodos de Aprendizaje Automático

2.6.1. Aprendizaje supervisado

El aprendizaje supervisado, los algoritmos se entrenan utilizando conjuntos de datos que ya han sido etiquetados, lo que les ayuda a clasificar o predecir resultados de manera más precisa. A lo largo del entrenamiento, el modelo ajusta sus parámetros como los pesos para alcanzar un ajuste óptimo con los datos de entrada y de esa manera evitar que el modelo se ajuste demasiado o muy poco a los datos.

2.6.2. Algoritmos de Aprendizaje supervisado

Bosques Aleatorios (Random Forest): este algoritmo se utiliza tanto para clasificación como para regresión y es útil para tratar con datos incompletos,

manteniendo su precisión aun si faltaran datos. Combina la salida de varios árboles de decisiones para alcanzar un resultado único.

Los Árboles de decisión son algoritmos de aprendizaje supervisado, que dividen los datos en grupos más pequeños basándose en decisiones sucesiva. El algoritmo parte de los datos de entrada y mediante una serie de preguntas o criterios de división que se aplican a las características de los datos se van construyendo el árbol de manera que los datos similares se agrupan. Los árboles de decisión pueden caer en el sobreajuste al intentar adaptarse perfectamente a los datos de entrenamiento, sin embargo, el uso de múltiples árboles en un Bosque Aleatorio mitiga ese riesgo ya que la combinación de árboles no correlacionados reduce el error de predicción y la varianza.

Máquina de vectores de soporte (SVM): representa un grupo de técnicas de aprendizaje supervisado aplicadas en clasificación, regresión y en la identificación de datos anómalos.

Entre sus beneficios se destacan su eficacia en ambientes de alta dimensionalidad, incluso cuando la cantidad de dimensiones supera el número de muestras. También son eficientes en términos de memoria ya que se basan en un subconjunto de los datos de entrenamiento, conocidos como vectores de soporte para su función de decisión. Además, son versátiles, permitiendo el uso de diversas funciones de núcleo, tanto estándar como personalizadas para adaptarse a diferentes necesidades de análisis. Sin embargo, la desventaja puede ser el sobreajuste como cuando el número de características es mucho mayor al número de muestras.

2.7. Metodología de Minería de Datos

2.7.1. CRISP-DM

CRISP DM (Cross-Industry Standard Process for Data Mining), es un modelo de referencia para orientar proyectos de minería de datos y como metodología describe las fases, tareas y la relación entre estas tareas. El modelo tiene seis etapas, interconectadas entre sí que indican la dependencia entre ellas, pero no necesariamente estrictas ya que los proyectos pueden fluir entre las fases según sea necesario para adaptar a los requerimientos surjan durante el desarrollo de un proyecto. Es un modelo flexible que puede ser personalizado fácilmente de acuerdo con las necesidades del modelo de minería de datos. A continuación, se describen las fases:

Comprensión del negocio

El objetivo de esta fase es definir el propósito empresarial, identificando el problema a resolver. Se establecen la metas, tareas y necesidades del proyecto desde un enfoque empresarial, los cuales se convierte en metas técnicas que se reflejan en un plan del proyecto.

Comprensión de los datos

Esta fase inicia con una recopilación inicial de los datos y se enfoca en actividades para lograr un entendimiento de estos, mediante la descripción y exploración de los datos, empleando análisis estadísticos básicos para evaluar problemas de calidad, obteniendo conocimientos preliminares y detectando segmentos de interés para generar hipótesis sobre patrones no evidentes.

Preparación de los datos

En la fase de preparación de datos se realizan las actividades para obtener el conjunto de datos final, para ello se realizan tareas de selección de tablas,

registros y atributos, así como la transformación, limpieza e integración de los datos.

Modelado

Una vez preparado los datos, se seleccionan, aplican y ajustan múltiples técnicas de modelado para calibrar sus parámetros a los valores óptimos. Pueden existir diversas técnicas aplicables a un mismo problema de minería de datos, cada una con requisitos específicos sobre la estructura de los datos, por lo que puede ser necesario regresar a la fase de preparación de datos para realizar los ajustes necesarios.

Evaluación

Antes de proceder con la implementación del modelo, es importante realizar la evaluación y revisar que el proceso de construcción del modelo cumpla con los objetivos del negocio, por lo que clave determinar si existe algún objetivo importante que no haya sigo considerado. En esta fase se debe tomar la decisión sobre el uso de los resultados de la minería de datos.

Despliegue

En esta fase se despliega el modelo en un ambiente de producción. Esta implementación se puede realizar mediante la creación de un servicio web, una API, un informe, etc.

CAPITULO III: MARCO METODOLÓGICO

3. Metodología

Para el desarrollo del modelo de predicción de enfermedades se basará en la metodología CRISP-DM. En este proceso se contemplarán las fases de

entendimiento del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación, el despliegue no será parte de este proyecto.

Siguiendo de manera secuencial las fases de esta metodología, se iniciará con la fase de Comprensión del Negocio, para esto se mantendrá un dialogo con un médico interno responsable del área de salud ocupacional donde se identificará la necesidad que tiene esta área en cuanto al análisis de datos debido a que como parte de la gestión del Talento Humano, anualmente se realizan los exámenes médicos ocupacionales al personal activo fijo y contratado que este laborando más de un año en la empresa y esta recopilación de datos se almacena en una base de datos correspondiente al sistema de Control Médico

Se continuará con la fase de Preparación de Datos, para esta fase se realizará la extracción de datos correspondientes de los exámenes de sangre y diagnósticos médicos codificados en CIE10, mediante Lenguaje SQL y se realizará un análisis de estos datos con el experto, para verificar la relación existente entre los valores de exámenes de sangre y los diagnósticos médicos. Además de manera técnica se analizará las estructuras para identificar los atributos que contienen los datos relevantes para realizar este estudio.

Para la tercera fase que corresponde a la Modelado, se aplicará algoritmos de aprendizaje supervisado.

3.1. Comprensión del Negocio

Es la fase inicial que se enfoca en comprender los objetivos y requisitos del proyecto desde una perspectiva comercial y luego convertir este conocimiento

en una definición del problema de minería de datos y un plan preliminar diseñado para lograr los objetivos.

3.1.1. Determinar los objetivos del negocio

El departamento de Salud Ocupacional lleva a cabo exámenes médicos anuales para el personal permanente y contratado con más de un año de servicio en la institución. Este chequeo integral abarca análisis de laboratorio, cuyos resultados son evaluados por un médico internista que emite diagnósticos utilizando la Clasificación Internacional de Enfermedades CIE-10. Aunque esta información se registra en el sistema de gestión médica y se almacena en una base de datos relacional, hasta ahora no se ha aprovechado de manera adecuada. La falta de análisis de estos datos ha impedido el desarrollo de estrategias que permita tomar acciones medicas preventivas.

Para este estudio se tomará como referencia los datos de exámenes de sangre y diagnósticos médicos del personal activo de los periodos comprendidos entre el año 2019 al 2023.

El objetivo principal es desarrollar un modelo de predicción que permita la identificación temprana de empleados y trabajadores en riesgo de desarrollar enfermedades específicas, enfocándose principalmente en afecciones detectables a través del análisis de exámenes de sangre como hiperlipidemias, diabetes, anemias, enfermedades hepáticas, gota, etc. Este modelo aspira ser una herramienta importante para áreas de salud ocupacional como apoyo en la toma de decisiones, permitiendo implementar estrategias preventivas y personalizadas que mejoren la calidad de vida de los empleados, al mismo tiempo reducir la incidencia y el impacto de estas enfermedades en el ambiente laboral.

Para el desarrollo de este trabajo se priorizará la protección y privacidad de los datos de salud de los trabajadores en todas las fases del proyecto, por lo que no se trabajará con datos personales que los identifiquen.

3.2. Comprensión de Datos

Esta fase se realizaron las siguientes actividades:

3.2.1. Recopilación de datos iniciales

Para este estudio se recopiló datos de exámenes de sangre y diagnósticos médicos desde año 2019 al 2023 del personal que hasta la fecha de extracción estaba activo. La extracción de los datos se realizó mediante la ejecución de sentencias SQL de la base de datos del sistema de control médico y se guardaron en archivos de extensión .csv.

En la siguiente imagen se visualiza las relaciones de las tablas de la base de datos:

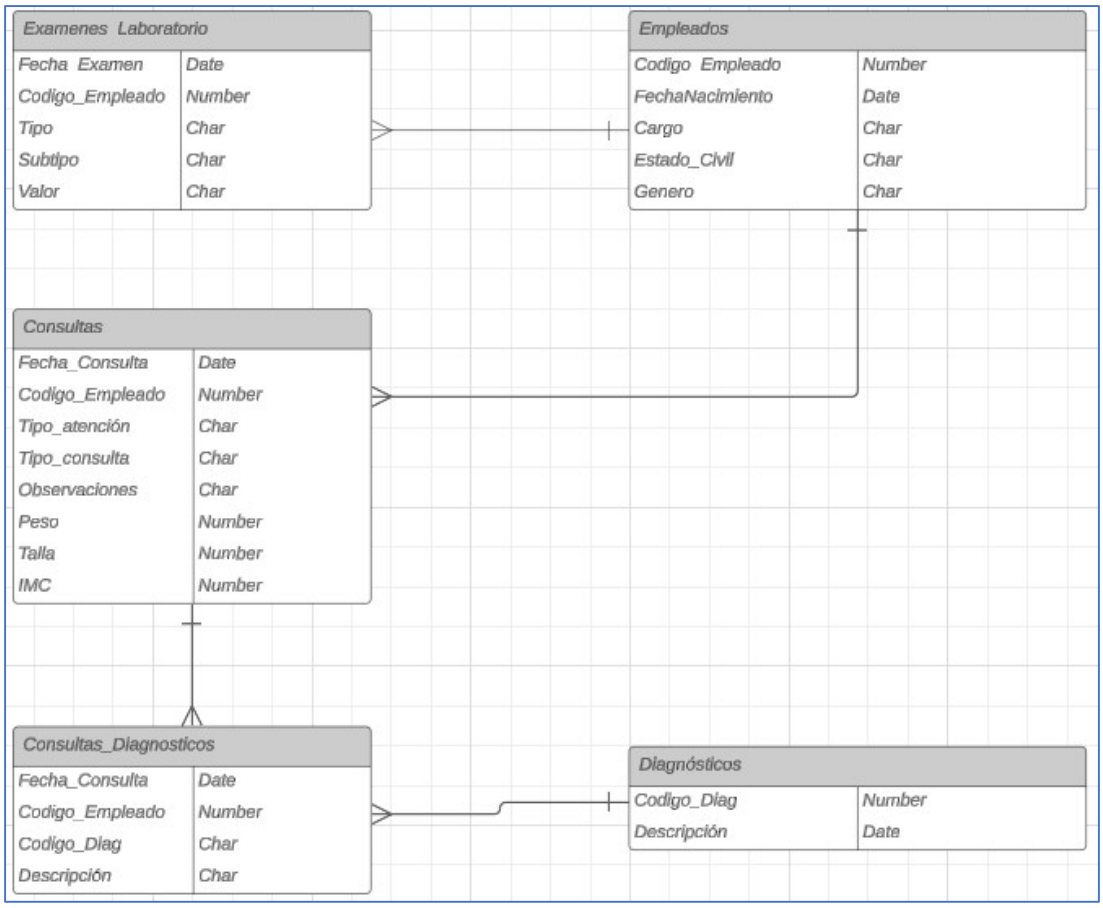


Figura 1:Diagrama Entidad - Relación

1.1.1. Descripción de los datos

El archivo que contiene los datos de exámenes médicos es el resultado de la relación entre las tablas de exámenes de sangre y la tabla de empleados. A continuación, se describen los atributos.

Atributos	Descripción
CODIGO_EMP	Código empleado o trabajador
FECHA_NACIMIENTO	Fecha de nacimiento
CARGO	Cargo de empleado o trabajador
ESTADOCIVIL	Soltero, Casado, Divorciado, Viudo, Unión libre
GENERO	Masculino, Femenino
FECHA_EXAMEN	Fecha de realización del examen
TIPO	Tipo de examen - Hematología
SUBTIPO	Subtipo examen de hematología

VALOR	Valor registrado por el laboratorio
REFERENCIA	Referencia del laboratorio
TIPO_SANGRE	Factor y tipo de sangre del empleado

Tabla 1: Descripción atributos de Exámenes médicos.

El archivo de datos de las consultas médicas y diagnósticos es el resultado de relación entre las tablas de empleados, consultas médicas y el catálogo de diagnósticos. A continuación, se detallan los atributos.

Atributos	**Descripción**
CODIGO_EMPLEADO	Código empleado o trabajador
FECHA _CONSULTA	Fecha de consulta médica
TIPO_ATENCION	Tipo de atención médica: Enfermedad, Ocupacional
TIPO _CONSULTA	Tipo de consulta: Primera, Subsecuente
PESO	Peso en Kilogramos del empleado o trabajador
TALLA	Talla en metros del empleado o trabajador
IMC	Índice de masa corporal en base al peso y talla
TEMPERATURA	Valor de la temperatura corporal
FRECUENCIA RESPIRATORIA	Número de veces que el paciente respira por minuto
TA_SISTOLICA	Presión sistólica, presión de la sangre en la arteria cuando se contrae el corazón.
TA_DIASTOLICA	Presión arterial diastólica, presión de la sangre en la arteria cuando el corazón se relaja entre latidos.
EVOLUCIÓN	Descripción de la evolución del paciente
INDICACIONES	Indicaciones farmacológicas y no farmacológicas
PRESCRIPCIONES	Prescripción médica
CODIGO_DIAGNOSTICO	Código del diagnóstico médico según la Clasificación Internacional de enfermedades.
DESCRIPCION	Descripción del diagnóstico médico

Tabla 2: Descripción de atributos de consultas médicas y diagnósticos

3.2.2. Exploración de datos

El archivo que contiene los datos de exámenes de laboratorio tiene 52.669 registros y el archivo que contiene las consultas médicas y diagnósticos tiene

9.087 registros. Luego de la integración de los datos en el archivo definitivo para el estudio se tiene 866 registros.

Se realiza un análisis gráfico de la distribución de las variables categóricas: AÑO, CARGO, GENERO, ESTADO CIVIL y DIAGNÓSTICOS CIE10.

En el siguiente gráfico se visualiza la distribución de datos por año de control médico, se puede observar que, en 2019, 2021 y 2023 existió concurrencia de los empleados a realizarse los exámenes del control médico anual, existiendo una disminución en el año 2020 debido a la falta de presencialidad por la pandemia del COVID-19. En el año 2022 se visualizan menos registros debido a un cambio en las fechas del proceso de control médico anual que inicia a finales del mes de diciembre y continua el proceso en los primeros meses de año 2023.

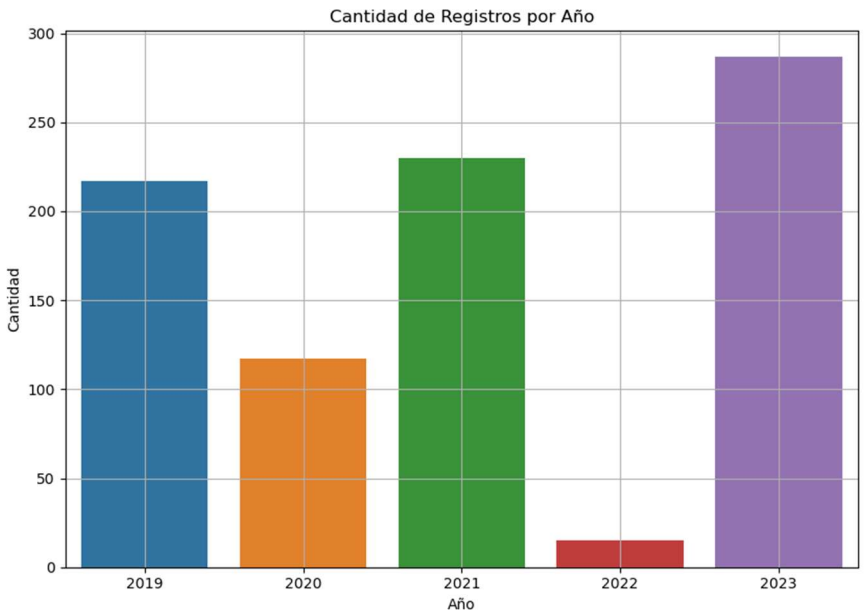


Figura 2: Distribución de los registros por Año.

En el gráfico por Cargo, indica la distribución del personal en la institución por distribución jerárquica por lo que el cargo Administrativo 3 y Operativo son los que prevalecen sobre los demás cargos.

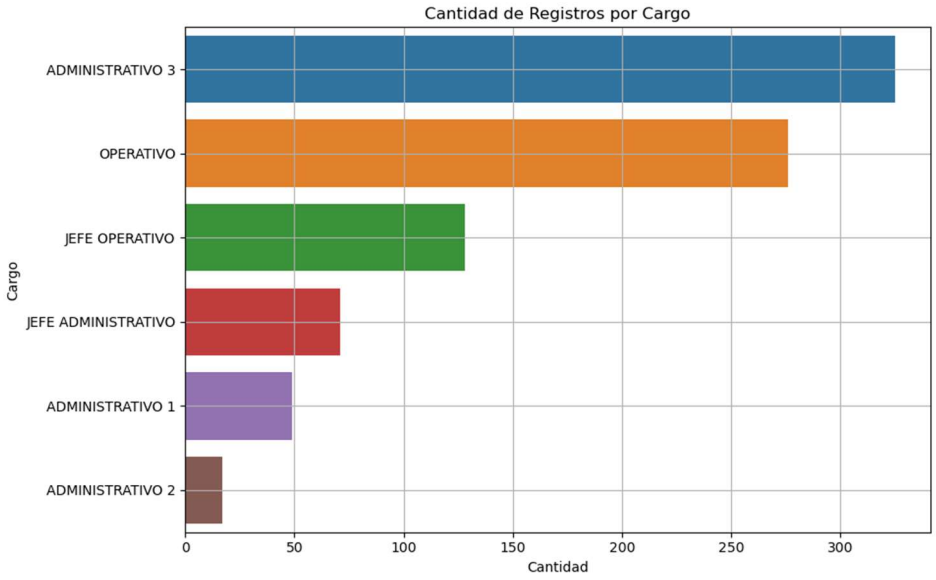


Figura 3: Distribución de los registros por Cargo.

En el siguiente gráfico por Género, prevalece el género masculino debido a la naturaleza del trabajo que realiza la institución lo que indica que existe una fuerza laboral masculina que prevalece sobre la femenina.

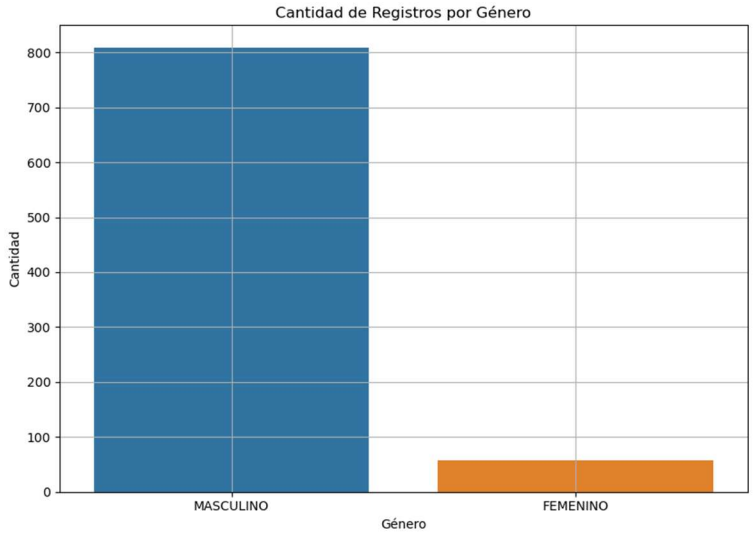


Figura 4: Distribución de los registros por Género.

El gráfico por Estado Civil muestra una mayoría de registros entre empleados y trabajadores casados y solteros, con un menor número de registros para divorciados, en unión libre y viudos.

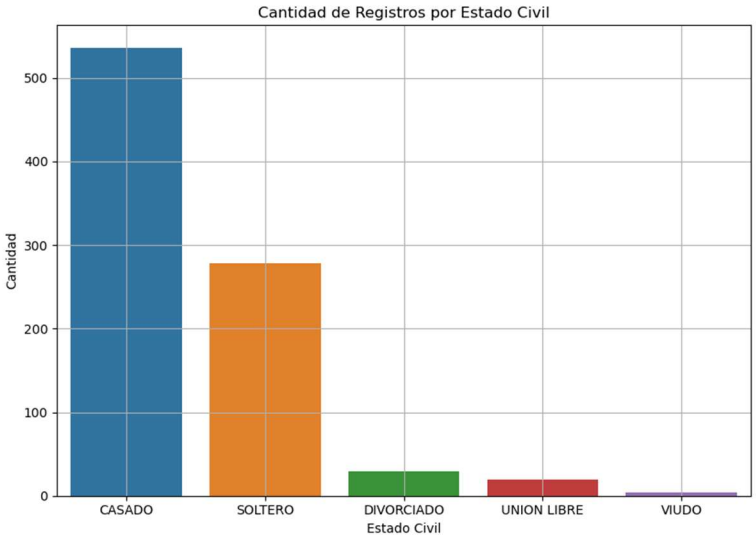


Figura 5: Distribución de los registros por Estado Civil.

En la siguiente figura se muestra la distribución de los registros por Diagnósticos CIE-10:

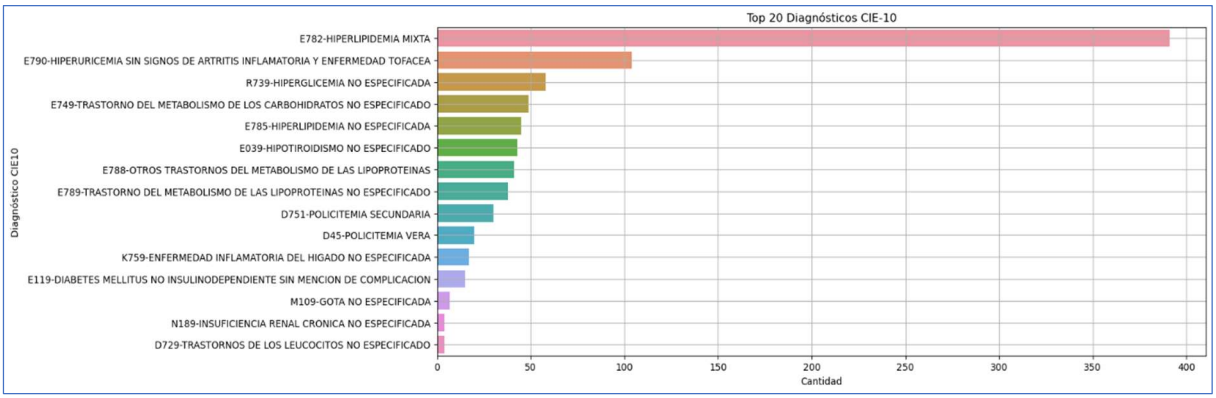


Figura 6: Distribución de la variable Diagnósticos CIE-10.

A continuación, se visualiza un análisis de las variables categóricas:

	CARGO	GENERO	ESTADOCIVIL	DIAGNOSTICO CIE10
count	866	866	866	866
unique	6	2	5	15
top	ADMINISTRATIVO 3	MASCULINO	CASADO	E782-HIPERLIPIDEMIA MIXTA
freq	325	809	536	391

Figura 7: Análisis de variables categóricas.

Se realiza un análisis de las variables categóricas obteniendo los siguientes resultados:

- *Cargo:* existen 6 categorías, siendo el cargo 'ADMINISTRATIVO 3', el más frecuente.
- *Género:* existen 2 categorías, siendo 'MASCULINO' el valor dominante.
- *Estado Civil:* existen 5 categorías, siendo el estado 'CASADO' el más común.
- *Diagnósticos CIE10:* se registran 15 códigos únicos de la Clasificación Internacional de Enfermedades con el código E782-HIPERLIPIDEMIA MIXTA, como el diagnóstico más frecuente.

3.2.3. Verificación de la calidad de los datos

Existen nombres de los subtipos de exámenes en letra minúscula por lo que se deberá convertir a mayúsculas para mantener estandarizados los datos categóricos.

Para mantener la confidencialidad de la información se deberá anonimizar los cargos de los empleados y trabajadores.

Se verifica que los datos de los tipos de exámenes de sangre están clasificados en varias categorías por lo que es necesario unificar en un solo tipo.

Los datos de Genero y Estado Civil están con nombres abreviados por lo que se deberá completar los nombres.

Luego de realizar la exploración inicial de los datos se detectó que existen valores erróneos en la talla y el peso del empleado, lo que afecta al cálculo del IMC (Índice de masa corporal), por lo que es necesario revisar el historial clínico de los empleados y realizar un nuevo cálculo del IMC.

Se debe realizar el cálculo de la edad del empleado con relación al año de control médico, debido a que se dispone como referencia la fecha de nacimiento.

El archivo de exámenes de sangre tiene los datos de subtipo de examen en filas por lo que es necesario realizar una transposición de filas en columnas para una mejor comprensión de los datos.

Al realizar la revisión de los datos, se verifica que no existe una relación directa entre los datos de exámenes médicos y consultas médicas - diagnósticos, por lo que será necesario realizar un cruce de información y obtener un nuevo archivo para el análisis.

3.3. Preparación y muestreo de datos

3.3.1. Selección de datos

Del archivo de Exámenes de hematología se escogieron los siguientes campos: CODIGO_EMP, FECHANACIMIENTO, CARGO, ESTADOCIVIL, GENERO, FECHA_EXAMEN, SUBTIPO, VALOR.

Del archivo de Consultas-Diagnósticos médicos, se escogieron los siguientes campos: CODIGO_EMP, FECHA_CONSULTA, TIPO_ATENCION,

TIPO_CONSULTA, PESO, TALLA, IMC, CODIGO_DIAGNOSTICO, DESCRIPCION.

1.1.2. Limpieza de los datos

Dentro de proceso de limpieza y transformación de datos se realiza las siguientes actividades:

Se corrigen faltas de ortografía y se estandariza a letra mayúscula los datos categóricos por ejemplo de los subtipos de exámenes de sangre.

Se generan 6 categorías genéricas de cargos para anonimizar los cargos originales y se reemplazan en el archivo original como se visualiza en la siguiente tabla:

NUEVOS CARGOS
ADMINISTRATIVO 1
ADMINISTRATIVO 2
ADMINISTRATIVO 3
JEFE ADMINISTRATIVO
JEFE OPERATIVO
OPERATIVO

Tabla 3: Nuevos cargos.

Se unifica en un solo tipo de examen, las diversas clasificaciones que contiene el archivo original de exámenes, como se indica en la siguiente tabla:

VARIOS TIPOS	TIPO UNICO
GGT	HEMATOLOGIA
HEMATOLOGIA CONTROL MEDICO	HEMATOLOGIA
HEMATOLOGIA TIROIDEA (T3, T4 Y TSH)	HEMATOLOGIA
HEMATOLOGIA_PSA	HEMATOLOGIA
HEMOGLOBINA GLICOSILADA A1C	HEMATOLOGIA
SOLICITUD_HEMATOLOGIA	HEMATOLOGIA

Tabla 4:Tipos de examen unificados.

A los datos de GENERO y ESTADO CIVIL se cambia las abreviaturas por nombres completos, como se muestra en las siguientes tablas:

GENERO	
Abreviatura	**Detalle**
M	MASCULINO
F	FEMENINO

Tabla 5: Género

ESTADO CIVIL	
Abreviatura	**Detalle**
S	SOLTERO
C	CASADO
D	DIVORCIADO
V	VIUDO
U	UNION LIBRE

Tabla 6: Estado Civil.

Se verifica los registros de talla y pesos del historial médico más próximo a la fecha de consulta de control médico ocupacional del empleado, se realiza la corrección de los datos y se recalcula el valor del IMC (Índice de Masa Corporal).

Se realiza el calculo de la edad del empleado tomando como referencia el año de nacimiento y el año del control médico para agregar la columna edad a la data.

El detalle del subtipo de examen de hematología por empleado se visualizaba en varias filas por lo que se realizó una transposición a columnas, como se puede ver en las siguientes imágenes.

DATOS DE EXAMENES DE HEMATOLOGIA										
CODIGO_EMP	AÑO	FECHA	CARGO	ESTADOCIVIL	GENERO	FECHANACIMIENTO	ANIO_NACIMIENTO	EDAD	SUBTIPO	VALOR
CE119XX	2019	20190513	ADMINISTRATIVO 1	CASADO	MASCULINO	19750903	1975	44	T4	1.07
CE119XX	2019	20190513	ADMINISTRATIVO 1	CASADO	MASCULINO	19750903	1975	44	T3	3.3
CE119XX	2019	20190513	ADMINISTRATIVO 1	CASADO	MASCULINO	19750903	1975	44	TSH	1.41
CE124XX	2019	20190513	ADMINISTRATIVO 3	SOLTERO	FEMENINO	19870125	1987	32	ERITROCITOS	5.27
CE124XX	2019	20190513	ADMINISTRATIVO 3	SOLTERO	FEMENINO	19870125	1987	32	HEMOGLOBINA	14.9
CE124XX	2019	20190513	ADMINISTRATIVO 3	SOLTERO	FEMENINO	19870125	1987	32	HEMATOCRITO	43.1
CE124XX	2019	20190513	ADMINISTRATIVO 3	SOLTERO	FEMENINO	19870125	1987	32	UREA	21
CE124XX	2019	20190513	ADMINISTRATIVO 3	SOLTERO	FEMENINO	19870125	1987	32	CREATININA	0.49
CE124XX	2019	20190513	ADMINISTRATIVO 3	SOLTERO	FEMENINO	19870125	1987	32	ERITROSEDIMENTACION	6
CE124XX	2019	20190513	ADMINISTRATIVO 3	SOLTERO	FEMENINO	19870125	1987	32	TGO	17
CE124XX	2019	20190513	ADMINISTRATIVO 3	SOLTERO	FEMENINO	19870125	1987	32	TGP	15
CE124XX	2019	20190513	ADMINISTRATIVO 3	SOLTERO	FEMENINO	19870125	1987	32	LEUCOCITOS	7.41
CE124XX	2019	20190513	ADMINISTRATIVO 3	SOLTERO	FEMENINO	19870125	1987	32	NEUTROFILOS	4.33
CE124XX	2019	20190513	ADMINISTRATIVO 3	SOLTERO	FEMENINO	19870125	1987	32	EOSINOFILOS	0.32
CE124XX	2019	20190513	ADMINISTRATIVO 3	SOLTERO	FEMENINO	19870125	1987	32	MONOCITOS	0.41
CE124XX	2019	20190513	ADMINISTRATIVO 3	SOLTERO	FEMENINO	19870125	1987	32	LINFOCITOS	2.33
CE124XX	2019	20190513	ADMINISTRATIVO 3	SOLTERO	FEMENINO	19870125	1987	32	PLAQUETAS	312
CE124XX	2019	20190513	ADMINISTRATIVO 3	SOLTERO	FEMENINO	19870125	1987	32	ACIDO URICO	4.4
CE124XX	2019	20190513	ADMINISTRATIVO 3	SOLTERO	FEMENINO	19870125	1987	32	TRIGLICERIDOS	135
CE124XX	2019	20190513	ADMINISTRATIVO 3	SOLTERO	FEMENINO	19870125	1987	32	COLESTEROL	152.4
CE124XX	2019	20190513	ADMINISTRATIVO 3	SOLTERO	FEMENINO	19870125	1987	32	LDL	72.1
CE124XX	2019	20190513	ADMINISTRATIVO 3	SOLTERO	FEMENINO	19870125	1987	32	HDL	53.3
CE124XX	2019	20190513	ADMINISTRATIVO 3	SOLTERO	FEMENINO	19870125	1987	32	GLICEMIA BASAL	86
CE124XX	2019	20190513	ADMINISTRATIVO 3	SOLTERO	FEMENINO	19870125	1987	32	MCV	82
CE124XX	2019	20190513	ADMINISTRATIVO 3	SOLTERO	FEMENINO	19870125	1987	32	MCH	28.3
CE124XX	2019	20190513	ADMINISTRATIVO 3	SOLTERO	FEMENINO	19870125	1987	32	MCHC	34.6
CE124XX	2019	20190513	ADMINISTRATIVO 3	SOLTERO	FEMENINO	19870125	1987	32	T4	1.15
CE124XX	2019	20190513	ADMINISTRATIVO 3	SOLTERO	FEMENINO	19870125	1987	32	T3	4.04
CE124XX	2019	20190513	ADMINISTRATIVO 3	SOLTERO	FEMENINO	19870125	1987	32	TSH	1.93
CE124XX	2019	20190513	ADMINISTRATIVO 3	SOLTERO	FEMENINO	19870125	1987	32	GGT	20
CE1245X	2019	20190514	ADMINISTRATIVO 3	UNION LIBRE	MASCULINO	19860620	1986	33	ERITROCITOS	5.38
CE1245X	2019	20190514	ADMINISTRATIVO 3	UNION LIBRE	MASCULINO	19860620	1986	33	HEMOGLOBINA	15.9
CE1245X	2019	20190514	ADMINISTRATIVO 3	UNION LIBRE	MASCULINO	19860620	1986	33	HEMATOCRITO	47.1
CE1245X	2019	20190514	ADMINISTRATIVO 3	UNION LIBRE	MASCULINO	19860620	1986	33	UREA	36
CE1245X	2019	20190514	ADMINISTRATIVO 3	UNION LIBRE	MASCULINO	19860620	1986	33	CREATININA	0.91

Figura 8: Datos de Exámenes de Sangre

DATOS DE EXAMENES DE HEMATOLOGIA (NUEVO ARCHIVO)																
CODIGO_EMP	AÑO	CARGO	GENERO	ESTADOCIVIL	EDAD	ACIDO URICO	COLESTEROL	CREATININA	EOSINOFILOS	ERITROCITOS	ERITROSEDIMENTACION	GGT	GLICEMIA BASAL	HDL	...	UREA
CE1136XX	2019	JEFE OPERATIVO	MASCULINO	CASADO	55	5.8	184	0.88	0.09	5.21	20	45	100	26.3	47.1	28
CE1136XX	2021	JEFE OPERATIVO	MASCULINO	CASADO	57	4.4	217	0.85	1	5.15	7	47	110.6	34	46.2	22.2
CE1136XX	2023	JEFE OPERATIVO	MASCULINO	CASADO	59	5.1	195.9	0.76	0.9	5.14	0	51	85	33.3	47.2	30
CE1136XX	2023	JEFE OPERATIVO	MASCULINO	CASADO	59	5.1	195.9	0.76	0.9	5.14	0	51	85	33.3	47.2	30
CE1137XX	2019	ADMINISTRATIVO 3	MASCULINO	CASADO	55	7.6	229	0.85	0.28	5.01	16	40	132	36.8	42.8	32
CE1137XX	2021	ADMINISTRATIVO 3	MASCULINO	CASADO	57	5.7	222	0.78	4.1	5.39	23	42	333.7	31	46	32.6
CE1137XX	2023	ADMINISTRATIVO 3	MASCULINO	CASADO	59	7.5	212	0.77	3.6	5.07	0	41	124	37.6	44.7	31
CE1137XX	2023	ADMINISTRATIVO 3	MASCULINO	CASADO	57	8.8	201.7	1.18	1.1	5.86	0	26	90	44.3	53.6	35
CE1137XX	2023	ADMINISTRATIVO 3	MASCULINO	CASADO	57	8.8	201.7	1.18	1.1	5.86	0	26	90	44.3	53.6	35
CE1137XX	2023	ADMINISTRATIVO 3	MASCULINO	CASADO	57	8.8	201.7	1.18	1.1	5.86	0	26	90	44.3	53.6	35
CE1138XX	2023	ADMINISTRATIVO 3	FEMENINO	CASADO	59	4.7	248.3	0.76	1.4	5.48	0	27	86	74.7	48.8	40
CE1139XX	2019	ADMINISTRATIVO 3	MASCULINO	CASADO	51	5.4	215.3	1.15	0.1	5.89	3	30	104	32.3	51.2	0
CE1139XX	2021	ADMINISTRATIVO 3	MASCULINO	CASADO	53	5.9	204	1.07	0.7	5.84	7	21	89.8	32	50.9	34.3
CE1139XX	2019	ADMINISTRATIVO 3	MASCULINO	CASADO	56	6.1	139.4	0.93	0.17	5.39	6	40	91	30.8	45.7	34
CE1139XX	2019	ADMINISTRATIVO 3	MASCULINO	CASADO	56	6.1	139.4	0.93	0.17	5.39	6	40	91	30.8	45.7	34
CE1139XX	2020	ADMINISTRATIVO 3	MASCULINO	CASADO	57	6	170	1.03	1.3	5.61	5	33	96	39	46.5	34.7
CE1139XX	2021	ADMINISTRATIVO 3	MASCULINO	CASADO	58	5.3	169	1.08	1.4	5.54	5	24	87.1	42	47.1	31.3
CE1139XX	2021	ADMINISTRATIVO 3	MASCULINO	CASADO	58	5.3	169	1.08	1.4	5.54	5	24	87.1	42	47.1	31.3
CE1139XX	2023	ADMINISTRATIVO 3	MASCULINO	CASADO	60	7.1	191	1.04	1.8	5.86	0	33	88	38.9	50.7	33
CE1139XX	2023	ADMINISTRATIVO 3	MASCULINO	CASADO	60	7.1	191	1.04	1.8	5.86	0	33	88	38.9	50.7	33
CE1141XX	2019	ADMINISTRATIVO 3	MASCULINO	CASADO	50	6.3	241.4	0.93	0.04	4.76	7	163	99	56.8	44.3	28
CE1141XX	2020	ADMINISTRATIVO 3	MASCULINO	CASADO	51	5.4	220	0.79	0.8	4.94	25	158	107.7	58	45.8	16.6
CE1141XX	2021	ADMINISTRATIVO 3	MASCULINO	CASADO	52	5.6	246	0.76	1.9	4.93	11	167	96.5	58	46.1	19.4
CE1141XX	2021	OPERATIVO	MASCULINO	CASADO	56	4.8	218	1.05	1.6	5.8	8	71	104.7	55	48.2	36.9
CE1141XX	2023	OPERATIVO	MASCULINO	CASADO	58	4.6	227.3	1.05	1.6	5.46	0	118	103	55.1	47.6	38
CE1141XX	2020	ADMINISTRATIVO 3	FEMENINO	CASADO	53	5.1	224	0.75	2	6.16	8	29	119.5	39	48	26
CE1141XX	2021	ADMINISTRATIVO 3	FEMENINO	CASADO	54	4.7	226	0.8	2	5.81	8	21	105.9	47	47.1	27.6
CE1141XX	2023	ADMINISTRATIVO 3	FEMENINO	CASADO	56	5.8	234.9	0.83	1.6	5.98	0	28	111	43.7	46.1	35
CE1141XX	2023	ADMINISTRATIVO 3	FEMENINO	CASADO	56	5.8	234.9	0.83	1.6	5.98	0	28	111	43.7	46.1	35

Figura 9: Datos de Exámenes de sangre - Nuevo archivo.

Una vez realizado este proceso se obtuvo un archivo con 2071 registros de exámenes de sangre.

1.1.3. Integración de los datos

Para obtener los datos definitivos para desarrollar este análisis, se realiza la integración de los datos en un solo archivo, para ello se realiza una fusión de los datos tomando en consideración el código del empleado, la fecha del examen, la fecha de la consulta subsecuente y el tipo de atención por control médico, obteniendo un nuevo archivo con 866 filas y 38 columnas, con los atributos que se describen a continuación:

Nro.	Atributos	Descripción
1	CODIGO_EMP	Código empleado o trabajador
2	AÑO	Año control médico
3	CARGO	Cargo de empleado o trabajador
4	GENERO	Soltero, Casado, Divorciado, Viudo, Unión libre
5	ESTADOCIVIL	Masculino, Femenino
6	EDAD	Edad del empleado con relación al año de control médico
7	DIAGNÓSTICO CIE10	Diagnóstico médico según la Clasificación Internacional de enfermedades
8	PESO	Peso en Kilogramos del empleado o trabajador
9	TALLA	Talla en metros del empleado o trabajador
10	IMC	Índice de masa corporal en base al peso y talla
11	ACIDO URICO	Bioquímicos
12	COLESTEROL	Bioquímicos
13	CREATININA	Bioquímicos
14	EOSINÓFILOS	Hemograma
15	ERITROCITOS	Hemograma
16	ERITROSEDIMENTACIÓN	Hemograma
17	GGT	Enzimas
18	GLICEMIA BASAL	Bioquímicos
19	HDL	Bioquímicos
20	HEMATOCRITO	Hemograma
21	HEMOGLOBINA	Hemograma
22	LDL	Bioquímicos
23	LEUCOCITOS	Hemograma
24	LINFOCITOS	Hemograma
25	MCH	Hemograma

26	MCHC	Hemograma
27	MCV	Hemograma
28	MONOCITOS	Hemograma
29	NEUTRÓFILOS	Hemograma
30	PSA	Inmunológicos
31	PLAQUETAS	Hemograma
32	T3	Hormonas
33	T4	Hormonas
34	TGO	Enzimas
35	TGP	Enzimas
36	TRIGLICERIDOS	Bioquímicos
37	TSH	Hormonas
38	UREA	Bioquímicos

Tabla 7: Descripción de atributos de datos integrados.

1.1.4. Formateo de datos

Se elimina las comas de la variable categórica DIAGNOSTICOS CIE10 debido a que contienen coma (,) en su texto y el archivo a procesar será delimitado por comas (CSV). En la siguiente tabla se visualizan los datos originales:

DIAGNOSTICO CIE10
D729-TRASTORNOS DE LOS LEUCOCITOS, NO ESPECIFICADO
E039-HIPOTIROIDISMO, NO ESPECIFICADO
E749-TRASTORNO DEL METABOLISMO DE LOS CARBOHIDRATOS, NO ESPECIFICADO
E789-TRASTORNO DEL METABOLISMO DE LAS LIPOPROTEINAS, NO ESPECIFICADO
K759-ENFERMEDAD INFLAMATORIA DEL HIGADO, NO ESPECIFICADA
M109-GOTA, NO ESPECIFICADA
N189-INSUFICIENCIA RENAL CRONICA, NO ESPECIFICADA
R739-HIPERGLICEMIA, NO ESPECIFICADA

Tabla 8: Diagnósticos CIE-10

Se mantiene la variable Eritrosedimentación con valores en 0 porque el rango de valores puede estar entre 0 a 22 mm/h para los hombres y de 0 a 29 mm/h para las mujeres.

En el caso de la variable PSA (Prueba de Antígeno prostático), al ser un valor medible solo para el género masculino, para el género femenino se realizo el tratamiento de datos combinando un indicador de genero con PSA, por lo que se creó una variable asignando un valor específico al género femenino de -1 para indicar que no es aplicable y mantener los valores de PSA para el género masculino, como se muestra en la siguiente figura:

	GENERO	PSA	PSA_Gender_Combined
0	MASCULINO	0.93	0.93
1	MASCULINO	0.27	0.27
2	MASCULINO	0.20	0.20
3	MASCULINO	0.20	0.20
4	MASCULINO	0.18	0.18

Figura 10: Tratamiento variable PSA.

Para el caso de valores en 0 de las variables T3, T4 Y TSH, se dio un tratamiento parecido al de la variable PSA, es decir se creo tres nuevas variables asignando un valor de -1, para indicar que la prueba no es aplicable o que no se realizó, las nuevas variables se muestran en la siguiente figura:

	T3	T3_Combined	T4	T4_Combined	TSH	TSH_Combined
0	3.32	3.32	0.93	0.93	4.17	4.17
1	2.39	2.39	1.00	1.00	6.41	6.41
2	3.01	3.01	1.31	1.31	0.39	0.39
3	3.01	3.01	1.31	1.31	0.39	0.39
4	0.00	-1.00	0.00	-1.00	0.00	-1.00

Figura 11: Tratamiento de variables, T3, T4 y TSH.

Para el tratamiento de los valores en 0 de las otras variables se opto por aplicar K Vecinos Cercanos (KNN) para preservar la mayor cantidad de datos, para ello utilizamos K=5 vecinos.

	ACIDO URICO	COLESTEROL	CREATININA	EOSINOFILOS	ERITROCITOS	GGT	GLICEMIA BASAL	HDL	HEMATOCRITO	HEMOGLOBINA	...	MCH	MCHC	MCV	MONOCI
0	5.8	184.0	0.88	0.09	5.21	45.0	100.0	26.3	47.1	15.7	...	30.1	33.3	90.0	
1	4.4	217.0	0.85	1.00	5.15	47.0	110.6	34.0	46.2	15.9	...	30.9	34.4	89.7	
2	5.1	195.9	0.76	0.90	5.14	51.0	85.0	33.3	47.2	16.1	...	31.3	34.1	92.0	
3	5.1	195.9	0.76	0.90	5.14	51.0	85.0	33.3	47.2	16.1	...	31.3	34.1	92.0	
4	7.6	229.0	0.85	0.28	5.01	40.0	132.0	36.8	42.8	15.3	...	30.5	35.7	85.0	

Figura 12: Tratamiento de variables KNN.

3.4. Modelado

El propósito de esta fase es la construcción de un modelo, empleando algoritmos de aprendizaje supervisado que permitan identificar patrones en los datos que ayuden a predecir enfermedades.

3.4.1. Selección del modelo

Para este estudio se ha seleccionado los algoritmos Bosques Aleatorios y Máquina de Vectores de Soporte SVM.

La variable por predecir (Y) es: ‘DIAGNOSTICO CIE10’,

3.4.2. Diseño de pruebas del modelo

Para evaluar el modelo, se dividió el conjunto de datos en dos partes: 80% para entrenamiento y 20% para prueba.

3.4.3. Construcción y evaluación del modelo

En esta fase se procedió a buscar los mejores hiperparámetros de cada modelo mediante técnicas de búsqueda y validación cruzada como RandomizedSearchCV con el fin de optimizar el rendimiento del modelo.

Bosques Aleatorios

Se divide el conjunto de datos con el 80% para entrenamiento y el 20% en pruebas:

```
# Separar las variables (X) y la variable objetivo (y)
X = df.drop('DIAGNOSTICO CIE10', axis=1)
y = df['DIAGNOSTICO CIE10']


# Dividiendo el dataset en conjuntos de entrenamiento y prueba nuevamente
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Verificando las dimensiones de los conjuntos de entrenamiento y prueba
X_train.shape, X_test.shape, y_train.shape, y_test.shape
```

```
((692, 31), (174, 31), (692,), (174,))
```

Figura 13:División dataset Bosques Aleatorios

Se crea y evalúa el modelo obteniendo una precisión del 0.477, lo que indica que el modelo predice correctamente el 47.7% de las instancias en el conjunto de prueba. Las métricas promedio indican que el modelo puede desempeñarse mejor en algunos casos que en otra debido al desequilibrio en el conjunto de datos ya que algunos diagnósticos tienen más muestras que otras.

```
Accuracy:  0.47701149425287354
```

	precision	recall	f1-score	support
D45-POLICITEMIA VERA	0.00	0.00	0.00	1
D729-TRASTORNOS DE LOS LEUCOCITOS NO ESPECIFICADO	0.00	0.00	0.00	1
D751-POLICITEMIA SECUNDARIA	0.33	0.20	0.25	5
E039-HIPOTIROIDISMO NO ESPECIFICADO	0.09	0.14	0.11	7
E119-DIABETES MELLITUS NO INSULINODEPENDIENTE SIN MENCION DE COMPLICACION	0.00	0.00	0.00	3
E749-TRASTORNO DEL METABOLISMO DE LOS CARBOHIDRATOS NO ESPECIFICADO	0.00	0.00	0.00	10
E782-HIPERLIPIDEMIA MIXTA	0.60	0.74	0.66	87
E785-HIPERLIPIDEMIA NO ESPECIFICADA	0.20	0.25	0.22	4
E788-OTROS TRASTORNOS DEL METABOLISMO DE LAS LIPOPROTEINAS	0.00	0.00	0.00	7
E789-TRASTORNO DEL METABOLISMO DE LAS LIPOPROTEINAS NO ESPECIFICADO	0.20	0.14	0.17	7
E790-HIPERURICEMIA SIN SIGNOS DE ARTRITIS INFLAMATORIA Y ENFERMEDAD TOFACEA	0.48	0.52	0.50	21
K759-ENFERMEDAD INFLAMATORIA DEL HIGADO NO ESPECIFICADA	0.00	0.00	0.00	2
M109-GOTA NO ESPECIFICADA	0.00	0.00	0.00	3
N189-INSUFICIENCIA RENAL CRONICA NO ESPECIFICADA	0.00	0.00	0.00	2
R739-HIPERGLICEMIA NO ESPECIFICADA	0.44	0.29	0.35	14
accuracy			0.48	174
macro avg	0.16	0.15	0.15	174
weighted avg	0.42	0.48	0.44	174

Figura 14: Entrenamiento del modelo Bosques Aleatorios.

Se entrena el modelo con los mejores hiperparámetros encontrados mediante la aplicación de la técnica RandomizedSearchCV que realiza una selección aleatoria de hiperparámetros a ser evaluados.

Se entrena nuevamente el modelo, obteniendo una precisión del 0.528, lo que indica que el modelo es capaz de predecir correctamente el 52.8% de las instancias.

```
print(accuracy_optimized)
print(classification_rep_optimized)
```

```
0.5287356321839081
                                                                    precision    recall  f1-score   support

                                                 D45-POLICITEMIA VERA       0.00      0.00      0.00         1
                                D729-TRASTORNOS DE LOS LEUCOCITOS NO ESPECIFICADO       0.00      0.00      0.00         1
                                                 D751-POLICITEMIA SECUNDARIA       1.00      0.20      0.33         5
                                            E039-HIPOTIROIDISMO NO ESPECIFICADO       0.11      0.14      0.12         7
  E119-DIABETES MELLITUS NO INSULINODEPENDIENTE SIN MENCION DE COMPLICACION       0.00      0.00      0.00         3
       E749-TRASTORNO DEL METABOLISMO DE LOS CARBOHIDRATOS NO ESPECIFICADO       0.00      0.00      0.00        10
                                                 E782-HIPERLIPIDEMIA MIXTA       0.57      0.89      0.70        87
                                            E785-HIPERLIPIDEMIA NO ESPECIFICADA       0.00      0.00      0.00         4
              E788-OTROS TRASTORNOS DEL METABOLISMO DE LAS LIPOPROTEINAS       0.00      0.00      0.00         7
       E789-TRASTORNO DEL METABOLISMO DE LAS LIPOPROTEINAS NO ESPECIFICADO       0.00      0.00      0.00         7
E790-HIPERURICEMIA SIN SIGNOS DE ARTRITIS INFLAMATORIA Y ENFERMEDAD TOFACEA       0.56      0.48      0.51        21
                  K759-ENFERMEDAD INFLAMATORIA DEL HIGADO NO ESPECIFICADA       0.00      0.00      0.00         2
                                                 M109-GOTA NO ESPECIFICADA       0.00      0.00      0.00         3
                        N189-INSUFICIENCIA RENAL CRONICA NO ESPECIFICADA       0.00      0.00      0.00         2
                                            R739-HIPERGLICEMIA NO ESPECIFICADA       0.38      0.21      0.27        14

                                                              accuracy                           0.53       174
                                                             macro avg       0.17      0.13      0.13       174
                                                          weighted avg       0.42      0.53      0.45       174
```

Figura 15: Entrenamiento del modelo Bosques Aleatorios mejores hiperparámetros.

Máquina de Vectores de Soporte (SVM)

Se divide el conjunto de datos con el 80% para entrenamiento y el 20% en pruebas, se ejecuta y evalúa el modelo obteniendo una exactitud de 0.4885, con lo que se puede observar que el modelo predice el 48.85% de las instancias en el conjunto de prueba.

```

# Calculando la exactitud del modelo
accuracy_svm = accuracy_score(y_test, y_pred_svm)
print(f"Exactitud del modelo SVM: {accuracy_svm:.4f}")
```

```
Exactitud del modelo SVM: 0.4885
```

Figura 16: Entrenamiento del modelo SVM

Se realiza una selección de los mejores hiperparámetros mediante la función RandomizedSearchCV, obteniendo como resultado la exactitud del modelo de un 50%.

```
# Evaluación del mejor modelo encontrado con el conjunto de prueba
best_model_svm = random_search_svm.best_estimator_
y_pred_optimized_svm = best_model_svm.predict(X_test)
optimized_accuracy_svm = accuracy_score(y_test, y_pred_optimized_svm)
print(f"Exactitud del modelo SVM optimizado: {optimized_accuracy_svm:.4f}")
```

```
Fitting 3 folds for each of 5 candidates, totalling 15 fits
C:\ProgramData\Anaconda3\lib\site-packages\sklearn\model_selection\_split.py:676: UserWarning: The least populated class in y h
as only 2 members, which is less than n_splits=3.
  warnings.warn(
Mejores hiperparámetros SVM: {'kernel': 'sigmoid', 'gamma': 'auto', 'C': 1}
Exactitud del modelo SVM optimizado: 0.5000
```

Figura 17: Entrenamiento del modelo SVM mejores hiperparámetros.

Se realizó una prueba imputando las tres clases con menos registros es decir los diagnósticos M109-GOTA NO ESPECIFICADA, N189-INSUFCIENCIA RENAL CRONICA NO ESPECIFICADA Y D729-TRASTORNOS DE LOS LEUCOCITOS NO ESPECIFICADOS y al aplicar el modelo de Bosques Aleatorio se obtuvo una exactitud del modelo 0.43 y en el modelo SVM una exactitud de 0.40, lo que indica que no mejoro su capacidad aún eliminando las muestras menos representativas.

3.5. Evaluación

El objetivo de negocio planteado en este proyecto se enfocó en la identificación temprana y precisa de riesgos de enfermedades en empleados y trabajadores, con la finalidad de poder realizar intervenciones preventivas, reduciendo el impacto en la salud del empleado, por lo que el criterio de éxito empresarial implicaría la generación de un modelo con alta precisión y capacidad predictiva.

En este caso la precisión de los modelos desarrollados fue de apenas el 52% para Bosques Aleatorios y 50% para SVM, lo que sitúa al modelo por debajo de lo que sería deseable para el criterio de éxito, por lo que se establece que no sería de utilidad práctica de acuerdo con los objetivos de negocio establecidos.

Para realizar este proyecto, se tuvo que ejecutar un tratamiento complejo de los datos por lo que se recomendaría analizar más a profundidad la calidad de los datos y seleccionar las características más representativas.

En conclusión, los resultados actuales indican que los modelos desarrollados no cumplen con los criterios de éxito planteados, sin embargo, la realización de este trabajo da un guía para futuros proyecto de minería de datos en este ámbito, orientado de alguna manera a enfocarse en una mejor comprensión de datos, selección de características más relevantes y tal vez en la aplicación de otras técnicas de modelado.

4. Conclusiones y Recomendaciones

4.1. Conclusiones

- Los datos de exámenes de sangre y los diagnósticos médicos obtenidos de la base de datos del sistema de control médico no están correctamente relacionados lo que dificulta realizar el proceso de integración de los datos.
- Se debe establecer políticas de registro de la información para mantener uniformidad y calidad de los datos, lo que ayudará a una mejor comprensión y preparación de los datos.

- El uso de la codificación CIE en los diagnósticos médicos ayuda a clasificar de mejor manera las enfermedades, manteniendo la uniformidad en el registro de los datos.
- Las características más relevantes en el desarrollo del modelo de bosques aleatorios fueron las relacionadas con valores metabólicos tales como colesterol, LDL y triglicéridos, además de otras características como ácido úrico y glicemia basal.
- Analizando los resultados de los dos modelos, se puede concluir que permiten predecir diagnósticos como hiperlipidemias, hiperuricemia e hiperglicemias, es decir enfermedades cardiovasculares y metabólicas comunes, pero no de una manera efectiva los demás diagnósticos.
- El desbalance de las clases (Diagnósticos) no permite realizar una predicción eficiente para las clases minoritarias por lo que para mejorar la capacidad de predicción del modelo se debería obtener más datos o analizar otras características relevantes.
- Aunque se realizaron pruebas con diferentes hiperparámetros los modelos no mejoraron la capacidad de predicción, lo que podría deberse a la calidad de datos obtenidos para este análisis.

4.2. Recomendaciones

- Realizar la integración en solo repositorio (ficha médica del paciente), tanto de los diagnósticos especializados como los diagnósticos del médico ocupacional, para evitar duplicidad de la información o selección de diagnósticos diferentes.

- Mantener uniformidad en el registro de valores de exámenes de laboratorio, generando un catálogo de valores referenciales tanto en unidades como en porcentajes.
- Es importante que tanto en el registro de diagnósticos médicos especializados y de exámenes de laboratorio se utilice la codificación CIE para mantener una clasificación ordenada de la información.
- Para un proyecto más completo de análisis se recomienda explotar los datos de imagenología, con lo que se puede generar nuevo conocimiento que permita tomar otras decisiones en el área de seguridad y salud ocupacional.

BIBLIOGRAFÍA

Adams, R. P. (2020). What is machine learning? *Machine Learning and AI for Healthcare*. https://doi.org/10.1007/978-1-4842-3799-1_3

Álvarez, F., & Faizal, E. (2012). *Salud ocupacional y su prevención*. Ediciones de la U.

Harrison, J. E., Weber, S., Jakob, R., & Chute, C. G. (2021). ICD-11: an international classification of diseases for the twenty-first century. *BMC Medical Informatics and Decision Making*, *21*(S6). https://doi.org/10.1186/s12911-021-01534-6

IBM. (n.d.-a). *¿Qué es el aprendizaje supervisado?* Ibm.com. Retrieved February 29, 2024, from https://www.ibm.com/es-es/topics/supervised-learning

IBM. (n.d.-b). *¿Qué es un bosque aleatorio?* Ibm.com. Retrieved February 29, 2024, from https://www.ibm.com/es-es/topics/random-forest

IBM. (n.d.-c). *What is machine learning?* Ibm.com. Retrieved February 29, 2024, from https://www.ibm.com/topics/machine-learning

IBM Documentation. (2021, August 17). Ibm.com. https://www.ibm.com/docs/es/spss-modeler/saas?topic=guide-introduction-crisp-dm

International Classification of Diseases (ICD). (n.d.). Who.int. Retrieved February 29, 2024, from

https://www.who.int/es/standards/classifications/classification-of-diseases

Ministerio de Salud Pública del Ecuador. (n.d.). Gob.Ec. Retrieved February 29, 2024, from https://www.salud.gob.ec/wp-content/uploads/2019/10/MANUAL-DE-POLITICAS-final.pdf

National Heart, Lung and Blood Institute. (n.d.). *Análisis de sangre*. NHLBI, NIH. Retrieved February 29, 2024, from https://www.nhlbi.nih.gov/es/salud/analisis-de-sangre

Occupational health. (n.d.). Who.int. Retrieved February 29, 2024, from https://www.who.int/health-topics/occupational-health

Rodrigo, J. A. (n.d.). *Máquinas de Vector Soporte (Support Vector Machines, SVMs)*. Cienciadedatos.net. Retrieved February 29, 2024, from https://cienciadedatos.net/documentos/34_maquinas_de_vector_soporte_support_vector_machines

Salud y Bienestar Laboral. (2019, January 25). *¿Qué es la salud laboral?* Salud Laboral y Discapacidad. https://saludlaboralydiscapacidad.org/salud-laboral/que-es/

scikit-learn developers. (n.d.). *1.4. Máquinas de Vectores de Soporte — documentación de scikit-learn - 0.24.1*. Github.Io. Retrieved February 29, 2024, from https://qu4nt.github.io/sklearn-doc-es/modules/svm.html

ANEXOS

Modelo de Bosques Aleatorios (Random Forest)

In [1]:
```
import pandas as pd
import numpy as np
from pandas import read_csv
from pandas.plotting import scatter_matrix
from matplotlib import pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.metrics import classification_report
from sklearn.metrics import confusion_matrix
from sklearn.metrics import accuracy_score
import seaborn as sns
import plotly as py
import plotly.graph_objs as go
from sklearn.cluster import KMeans
import warnings
import os
```

In [2]:
```
#Cargar los datos en un dataframe
csv_path = 'Examenes_Diagnosticos_Final.csv'
data = pd.read_csv(csv_path, sep=',')

```

In [3]:
```
data.head()
```

Out[3]:

	Unnamed: 0	CODIGO_EMP	AÑO	CARGO	GENERO	ESTADOCIVIL	EDAD	DIAGNOSTICO CIE1(
0	0	CE11365	2019	JEFE OPERATIVO	MASCULINO	CASADO	55	E782 HIPERLIPIDEMIA MIXTA
1	1	CE11365	2021	JEFE OPERATIVO	MASCULINO	CASADO	57	E782 HIPERLIPIDEMIA MIXTA
2	2	CE11365	2023	JEFE OPERATIVO	MASCULINO	CASADO	59	E039 HIPOTIROIDISMO NO ESPECIFICADO
3	3	CE11365	2023	JEFE OPERATIVO	MASCULINO	CASADO	59	E788-OTROS TRASTORNOS DE METABOLISMO DE LAS L.
4	4	CE11370	2019	ADMINISTRATIVO 3	MASCULINO	CASADO	55	R739 HIPERGLICEMIA NO ESPECIFICADA

5 rows × 39 columns

In [4]:
```
data = data.drop(columns=['Unnamed: 0'])
data_cleaned = data
```

In [5]:
```
#Revisión de la data final
data_cleaned.head()

```

Out[5]:

	CODIGO_EMP	AÑO	CARGO	GENERO	ESTADOCIVIL	EDAD	DIAGNOSTICO CIE10	PESO	T
0	CE11365	2019	JEFE OPERATIVO	MASCULINO	CASADO	55	E782-HIPERLIPIDEMIA MIXTA	67.5	
1	CE11365	2021	JEFE OPERATIVO	MASCULINO	CASADO	57	E782-HIPERLIPIDEMIA MIXTA	68.0	
2	CE11365	2023	JEFE OPERATIVO	MASCULINO	CASADO	59	E039-HIPOTIROIDISMO NO ESPECIFICADO	65.0	
3	CE11365	2023	JEFE OPERATIVO	MASCULINO	CASADO	59	E788-OTROS TRASTORNOS DEL METABOLISMO DE LAS L...	65.0	
4	CE11370	2019	ADMINISTRATIVO 3	MASCULINO	CASADO	55	R739-HIPERGLICEMIA NO ESPECIFICADA	90.3	

5 rows × 38 columns

In [6]:
```
#Tamaño de la data final
cleaned_size = data_cleaned.shape
cleaned_size
```

Out[6]:
```
(866, 38)
```

In [7]:

```
# Generando gráficos para la cantidad de registros por año, cargo, género y estad
plt.figure(figsize=(18, 12))

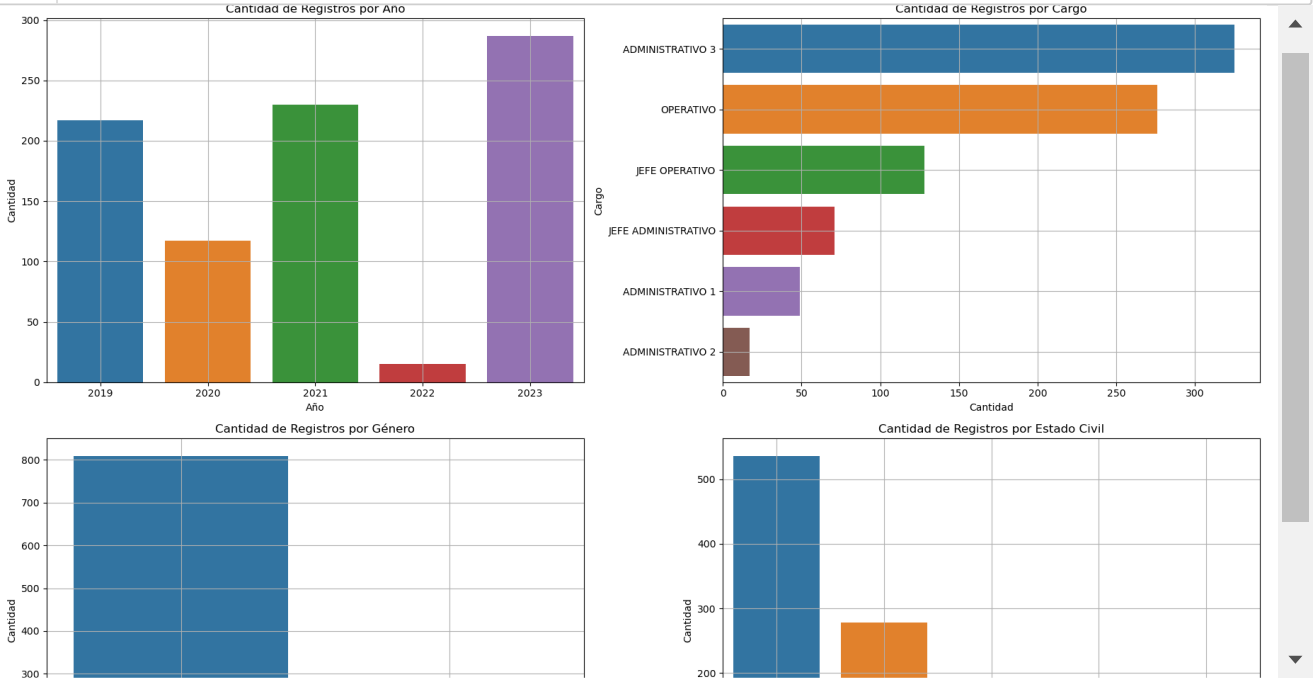
# Cantidad de registros por año
plt.subplot(2, 2, 1)
sns.countplot(x='AÑO', data=data_cleaned)
plt.title('Cantidad de Registros por Año')
plt.ylabel('Cantidad')
plt.xlabel('Año')
plt.grid(True)

# Cantidad de registros por cargo
plt.subplot(2, 2, 2)
sns.countplot(y='CARGO', data=data_cleaned, order=data_cleaned['CARGO'].value_cou
plt.title('Cantidad de Registros por Cargo')
plt.xlabel('Cantidad')
plt.ylabel('Cargo')
plt.grid(True)

# Cantidad de registros por género
plt.subplot(2, 2, 3)
sns.countplot(x='GENERO', data=data_cleaned)
plt.title('Cantidad de Registros por Género')
plt.xlabel('Género')
plt.ylabel('Cantidad')
plt.grid(True)

# Cantidad de registros por estado civil
plt.subplot(2, 2, 4)
sns.countplot(x='ESTADOCIVIL', data=data_cleaned, order=data_cleaned['ESTADOCIVIL
plt.title('Cantidad de Registros por Estado Civil')
plt.xlabel('Estado Civil')
plt.ylabel('Cantidad')
plt.grid(True)

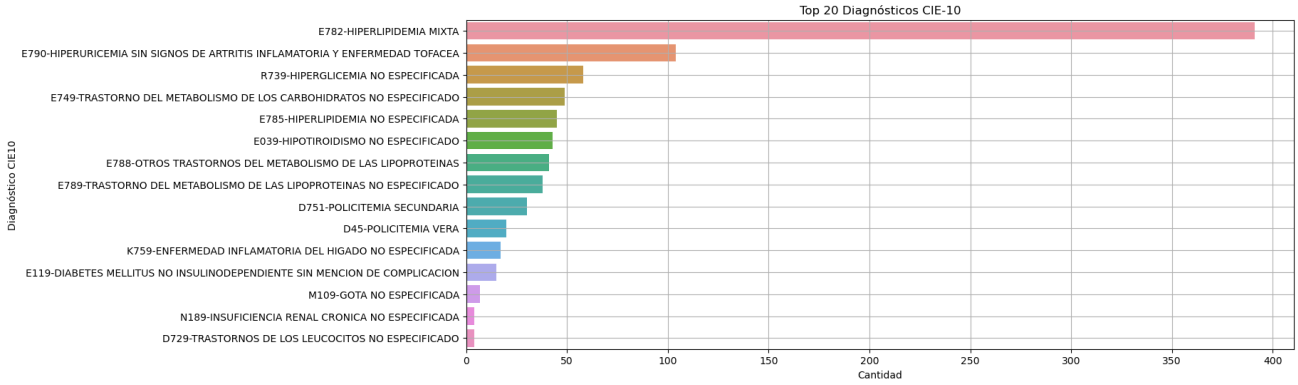
plt.tight_layout()
plt.show()
```



In [8]:

```
# Análisis de diagnósticos CIE-10
plt.figure(figsize=(15, 6))
diagnosis_counts = data_cleaned['DIAGNOSTICO CIE10'].value_counts().head(20)
sns.barplot(x=diagnosis_counts.values, y=diagnosis_counts.index)
plt.title('Top 20 Diagnósticos CIE-10')
plt.ylabel('Diagnóstico CIE10')
plt.xlabel('Cantidad')
plt.xticks(rotation=0)
plt.grid(True)
plt.show()

class_counts = data_cleaned['DIAGNOSTICO CIE10'].value_counts()
class_counts
```



Out[8]:

```
E782-HIPERLIPIDEMIA MIXTA                                                391
E790-HIPERURICEMIA SIN SIGNOS DE ARTRITIS INFLAMATORIA Y ENFERMEDAD TOFACEA    104
R739-HIPERGLICEMIA NO ESPECIFICADA                                        58
E749-TRASTORNO DEL METABOLISMO DE LOS CARBOHIDRATOS NO ESPECIFICADO       49
E785-HIPERLIPIDEMIA NO ESPECIFICADA                                       45
E039-HIPOTIROIDISMO NO ESPECIFICADO                                       43
E788-OTROS TRASTORNOS DEL METABOLISMO DE LAS LIPOPROTEINAS                41
E789-TRASTORNO DEL METABOLISMO DE LAS LIPOPROTEINAS NO ESPECIFICADO       38
D751-POLICITEMIA SECUNDARIA                                               30
D45-POLICITEMIA VERA                                                      20
K759-ENFERMEDAD INFLAMATORIA DEL HIGADO NO ESPECIFICADA                   17
E119-DIABETES MELLITUS NO INSULINODEPENDIENTE SIN MENCION DE COMPLICACION  15
M109-GOTA NO ESPECIFICADA                                                  7
N189-INSUFICIENCIA RENAL CRONICA NO ESPECIFICADA                           4
D729-TRASTORNOS DE LOS LEUCOCITOS NO ESPECIFICADO                          4
Name: DIAGNOSTICO CIE10, dtype: int64
```

In [9]:

```
# Análisis de variables categóricas
categorical_cols_cleaned = ['CARGO', 'GENERO', 'ESTADOCIVIL', 'DIAGNOSTICO CIE10'
categorical_data_cleaned = data_cleaned[categorical_cols_cleaned].describe(includ

categorical_data_cleaned
```

Out[9]:

	CARGO	GENERO	ESTADOCIVIL	DIAGNOSTICO CIE10
count	866	866	866	866
unique	6	2	5	15
top	ADMINISTRATIVO 3	MASCULINO	CASADO	E782-HIPERLIPIDEMIA MIXTA
freq	325	809	536	391

In [10]:

```
# Reconfigurando el estilo de los gráficos y creando las visualizaciones con el d
sns.set(style="whitegrid")

# Creación de figuras para visualización de datos con data_cleaned
fig, axs = plt.subplots(2, 2, figsize=(14, 10))

# Histograma de Edad
sns.histplot(data_cleaned['EDAD'], bins=20, kde=True, color="skyblue", ax=axs[0,
axs[0, 0].set_title('Distribución de Edades')

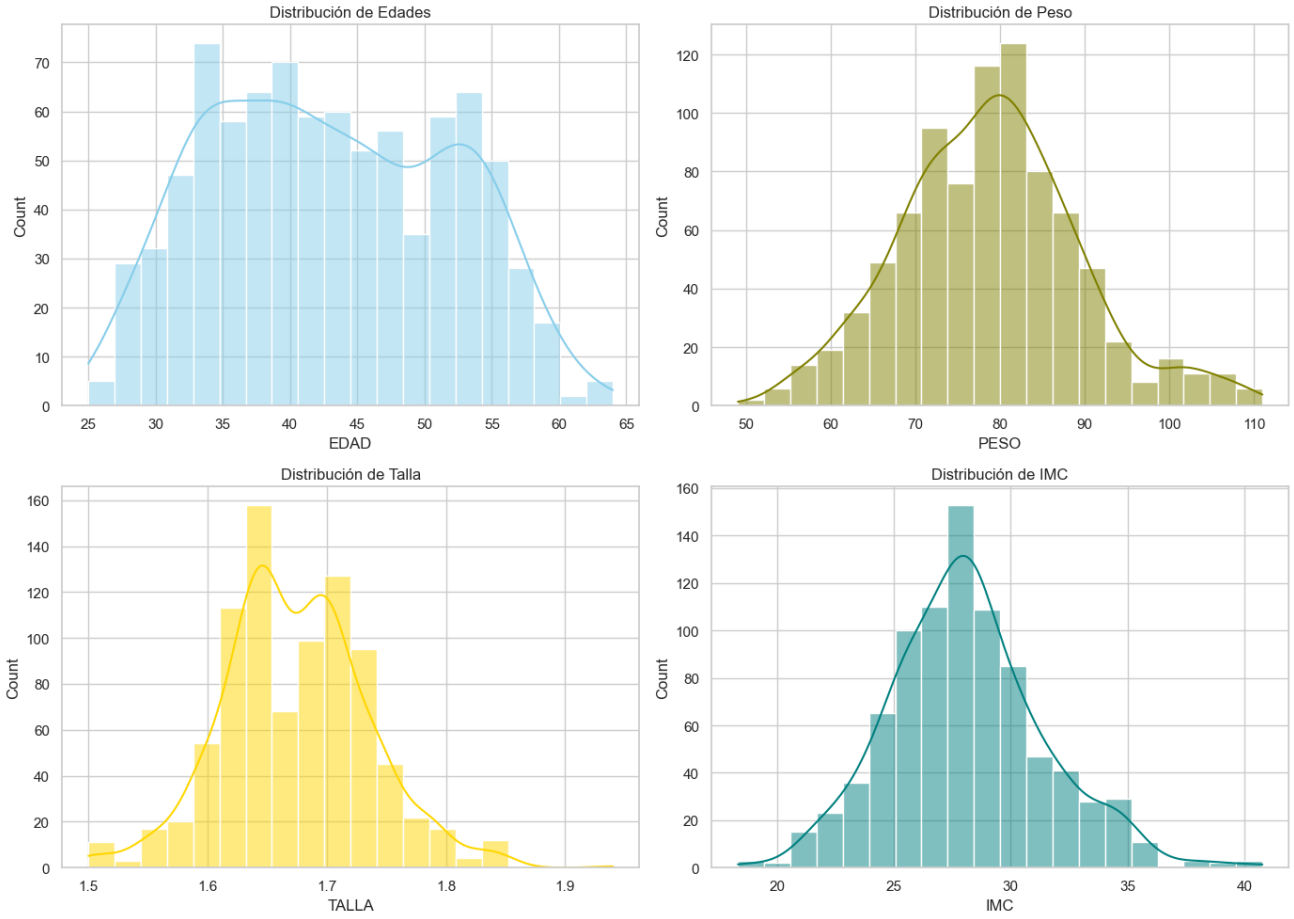
# Histograma de Peso
sns.histplot(data_cleaned['PESO'], bins=20, kde=True, color="olive", ax=axs[0, 1]
axs[0, 1].set_title('Distribución de Peso')

# Histograma de Talla
sns.histplot(data_cleaned['TALLA'], bins=20, kde=True, color="gold", ax=axs[1, 0]
axs[1, 0].set_title('Distribución de Talla')

# Histograma de IMC
sns.histplot(data_cleaned['IMC'], bins=20, kde=True, color="teal", ax=axs[1, 1])
axs[1, 1].set_title('Distribución de IMC')

plt.tight_layout()
plt.show()


```



Análisis de Variables Categóricas

CARGO: Hay 6 categorías únicas, siendo 'ADMINISTRATIVO 3' la más frecuente.

GENERO: Hay 2 categorías ('MASCULINO' y 'FEMENINO'), 'MASCULINO'

como la categoría dominante.

ESTADOCIVIL: Hay 5 categorías de estado civil, siendo 'CASADO' la más común.

Diagnósticos CIE10: Se registran 15 códigos únicos de la Clasificación Internacional de Enfermedades, ### con 'E782' (relacionado con

In [11]:
```
1 data.head()
```

Out[11]:

	CODIGO_EMP	AÑO	CARGO	GENERO	ESTADOCIVIL	EDAD	DIAGNOSTICO CIE10	PESO	T
0	CE11365	2019	JEFE OPERATIVO	MASCULINO	CASADO	55	E782-HIPERLIPIDEMIA MIXTA	67.5	
1	CE11365	2021	JEFE OPERATIVO	MASCULINO	CASADO	57	E782-HIPERLIPIDEMIA MIXTA	68.0	
2	CE11365	2023	JEFE OPERATIVO	MASCULINO	CASADO	59	E039-HIPOTIROIDISMO NO ESPECIFICADO	65.0	
3	CE11365	2023	JEFE OPERATIVO	MASCULINO	CASADO	59	E788-OTROS TRASTORNOS DEL METABOLISMO DE LAS L...	65.0	
4	CE11370	2019	ADMINISTRATIVO 3	MASCULINO	CASADO	55	R739-HIPERGLICEMIA NO ESPECIFICADA	90.3	

5 rows × 38 columns

In [12]:

```
# Realizando un análisis estadístico descriptivo del dataframe limpio
descriptive_stats = data_cleaned.describe()

# Añadiendo análisis de las variables categóricas
categorical_stats = data_cleaned.describe(include=['object'])

descriptive_stats, categorical_stats

```

Out[12]:

```
(               AÑO        EDAD        PESO       TALLA         IMC  \
count   866.000000  866.000000  866.000000  866.000000  866.000000
mean   2021.043880   42.911085   78.794815    1.673568   28.108256
std       1.575217    8.907628   10.669107    0.061121    3.341469
min    2019.000000   25.000000   49.000000    1.500000   18.330000
25%    2019.250000   35.000000   72.000000    1.630000   25.950000
50%    2021.000000   42.000000   79.000000    1.670000   27.990000
75%    2023.000000   51.000000   85.000000    1.710000   30.055000
max    2023.000000   64.000000  111.000000    1.940000   40.770000

       ACIDO URICO  COLESTEROL  CREATININA  EOSINOFILOS  ERITROCITOS  ...  \
count   866.000000  866.000000  866.000000   866.000000   866.000000  ...
mean      6.988799  220.201270    0.947921     2.130866     5.563176  ...
std      10.309935   44.611245    0.170510     2.250621     0.670487  ...
min       0.000000    0.000000    0.000000     0.000000     0.000000  ...
25%       5.600000  195.000000    0.860000     0.352500     5.302500  ...
50%       6.500000  221.000000    0.950000     1.700000     5.590000  ...
75%       7.400000  247.000000    1.040000     3.000000     5.860000  ...
max     229.000000  428.200000    1.650000    19.000000     7.540000  ...

       NEUTROFILOS         PSA   PLAQUETAS          T3          T4  \
count   866.000000  866.000000  866.000000  866.000000  866.000000
mean     39.419527    0.477485  276.781351    0.897829    0.330612
std      22.901915    0.771902   63.242373    1.479161    0.545750
min       0.000000    0.000000    0.000000    0.000000    0.000000
25%       5.287500    0.000000  242.000000    0.000000    0.000000
50%      48.500000    0.000000  274.000000    0.000000    0.000000
75%      56.000000    0.717500  314.000000    2.597500    0.950000
max      82.700000    5.500000  479.000000    4.720000    2.190000

              TGO         TGP  TRIGLICERIDOS         TSH        UREA
count  866.000000  866.000000     866.000000  866.000000  866.000000
mean    29.094457   44.145577     242.548637    0.948176   31.541686
std     13.436558   27.488486     150.543426    1.815739    8.317360
min      0.000000    0.000000       0.000000    0.000000    0.000000
25%     21.000000   25.850000     156.250000    0.000000   26.000000
50%     25.100000   38.000000     208.000000    0.000000   31.000000
75%     33.000000   53.225000     294.000000    1.367500   36.875000
max    138.000000  222.000000    1237.000000    9.120000   67.000000

[8 rows x 33 columns],
       CODIGO_EMP             CARGO     GENERO ESTADOCIVIL  \
count         866               866        866         866
unique        379                 6          2           5
top       CE11517  ADMINISTRATIVO 3  MASCULINO      CASADO
freq            7               325        809         536

               DIAGNOSTICO CIE10
count                        866
unique                        15
top     E782-HIPERLIPIDEMIA MIXTA
freq                         391  )
```

Estadísticas Descriptivas de Variables Numéricas

- **Edad**: Varía entre 25 y 64 años, con una media de aproximadamente 43 años.
- **Peso**: El peso promedio es de alrededor de 79 kg, con una variación de 49 a 111 kg.
- **Talla**: La talla media es de aproximadamente 1.67 m, variando entre 1.5 y 1.94 m.
- **IMC**: El IMC promedio es de 28.11, indicando una tendencia hacia el sobrepeso.
- Los parámetros de los exámenes de sangre (como colesterol, glicemia basal, triglicéridos) también muestran una amplia gama de valores.

Estadísticas Descriptivas de Variables Categóricas

- **Cargo**: Hay 6 cargos diferentes, siendo "ADMINISTRATIVO 3" el más común.
- **Género**: Mayoría de registros masculinos.
- **Estado Civil**: "CASADO" es el estado civil más común.
- **Diagnóstico CIE10**: Hay 15 diagnósticos diferentes, con "E782-HIPERLIPIDEMIA MIXTA" siendo el más frecuente.

In [13]:
```
# Crear una nueva columna que combine el género y los valores de PSA
# Asignamos -1 a las mujeres para indicar que el PSA no es aplicable
data_cleaned['PSA_Gender_Combined'] = data_cleaned.apply(lambda row: row['PSA'] i

# Mostrar las primeras filas para confirmar los cambios
data_cleaned[['GENERO', 'PSA', 'PSA_Gender_Combined']].head()
```

Out[13]:

	GENERO	PSA	PSA_Gender_Combined
0	MASCULINO	0.93	0.93
1	MASCULINO	0.27	0.27
2	MASCULINO	0.20	0.20
3	MASCULINO	0.20	0.20
4	MASCULINO	0.18	0.18

In [14]:
```
# Crear nuevas columnas para T3, T4 y TSH donde se asigna -1 a los valores de 0
data_cleaned['T3_Combined'] = data_cleaned['T3'].apply(lambda x: x if x != 0 else
data_cleaned['T4_Combined'] = data_cleaned['T4'].apply(lambda x: x if x != 0 else
data_cleaned['TSH_Combined'] = data_cleaned['TSH'].apply(lambda x: x if x != 0 el

# Mostrar las primeras filas para confirmar los cambios en estas columnas
data_cleaned[['T3', 'T3_Combined', 'T4', 'T4_Combined', 'TSH', 'TSH_Combined']].h
```

Out[14]:

	T3	T3_Combined	T4	T4_Combined	TSH	TSH_Combined
0	3.32	3.32	0.93	0.93	4.17	4.17
1	2.39	2.39	1.00	1.00	6.41	6.41
2	3.01	3.01	1.31	1.31	0.39	0.39
3	3.01	3.01	1.31	1.31	0.39	0.39
4	0.00	-1.00	0.00	-1.00	0.00	-1.00

In [15]:
```
1 data_cleaned.head()
```

Out[15]:

	CODIGO_EMP	AÑO	CARGO	GENERO	ESTADOCIVIL	EDAD	DIAGNOSTICO CIE10	PESO	T
0	CE11365	2019	JEFE OPERATIVO	MASCULINO	CASADO	55	E782-HIPERLIPIDEMIA MIXTA	67.5	
1	CE11365	2021	JEFE OPERATIVO	MASCULINO	CASADO	57	E782-HIPERLIPIDEMIA MIXTA	68.0	
2	CE11365	2023	JEFE OPERATIVO	MASCULINO	CASADO	59	E039-HIPOTIROIDISMO NO ESPECIFICADO	65.0	
3	CE11365	2023	JEFE OPERATIVO	MASCULINO	CASADO	59	E788-OTROS TRASTORNOS DEL METABOLISMO DE LAS L...	65.0	
4	CE11370	2019	ADMINISTRATIVO 3	MASCULINO	CASADO	55	R739-HIPERGLICEMIA NO ESPECIFICADA	90.3	

5 rows × 42 columns

In [16]:
```
1 #Borro las columnas: 'PSA', 'T3', 'T4', 'TSH'
2 data_cleaned = data_cleaned.drop(columns=['PSA', 'T3', 'T4', 'TSH'])
3 data_cleaned.head()
```

Out[16]:

	CODIGO_EMP	AÑO	CARGO	GENERO	ESTADOCIVIL	EDAD	DIAGNOSTICO CIE10	PESO	T
0	CE11365	2019	JEFE OPERATIVO	MASCULINO	CASADO	55	E782-HIPERLIPIDEMIA MIXTA	67.5	
1	CE11365	2021	JEFE OPERATIVO	MASCULINO	CASADO	57	E782-HIPERLIPIDEMIA MIXTA	68.0	
2	CE11365	2023	JEFE OPERATIVO	MASCULINO	CASADO	59	E039-HIPOTIROIDISMO NO ESPECIFICADO	65.0	
3	CE11365	2023	JEFE OPERATIVO	MASCULINO	CASADO	59	E788-OTROS TRASTORNOS DEL METABOLISMO DE LAS L...	65.0	
4	CE11370	2019	ADMINISTRATIVO 3	MASCULINO	CASADO	55	R739-HIPERGLICEMIA NO ESPECIFICADA	90.3	

5 rows × 38 columns

In [17]:

```python
import numpy as np
from sklearn.preprocessing import StandardScaler
from sklearn.impute import KNNImputer
import pandas as pd

# Especificando las columnas a utilizar
specific_columns = [
    "ACIDO URICO", "COLESTEROL", "CREATININA", "EOSINOFILOS", "ERITROCITOS",
    "GGT", "GLICEMIA BASAL", "HDL", "HEMATOCRITO", "HEMOGLOBINA", "LDL",
    "LEUCOCITOS", "LINFOCITOS", "MCH", "MCHC", "MCV", "MONOCITOS",
    "NEUTROFILOS", "PLAQUETAS", "TGO", "TGP", "TRIGLICERIDOS", "UREA"
]

# Preparar los datos para la imputación
data_for_imputation = data_cleaned[specific_columns].replace(0, np.nan)

# Normalizar los datos antes de aplicar K-NN
scaler = StandardScaler()
data_scaled = scaler.fit_transform(data_for_imputation)

# Imputación con K-NN
knn_imputer = KNNImputer(n_neighbors=5)
data_imputed_scaled = knn_imputer.fit_transform(data_scaled)

# Revertir la normalización
data_imputed = scaler.inverse_transform(data_imputed_scaled)

# Convertir los datos imputados de nuevo a un DataFrame
data_imputed_df = pd.DataFrame(data_imputed, columns=specific_columns)

# Mostrar las primeras filas del DataFrame imputado
data_imputed_df.head()
```

Out[17]:

	ACIDO URICO	COLESTEROL	CREATININA	EOSINOFILOS	ERITROCITOS	GGT	GLICEMIA BASAL	HDL	HEMATOCRI
0	5.8	184.0	0.88	0.09	5.21	45.0	100.0	26.3	4
1	4.4	217.0	0.85	1.00	5.15	47.0	110.6	34.0	4
2	5.1	195.9	0.76	0.90	5.14	51.0	85.0	33.3	4
3	5.1	195.9	0.76	0.90	5.14	51.0	85.0	33.3	4
4	7.6	229.0	0.85	0.28	5.01	40.0	132.0	36.8	4

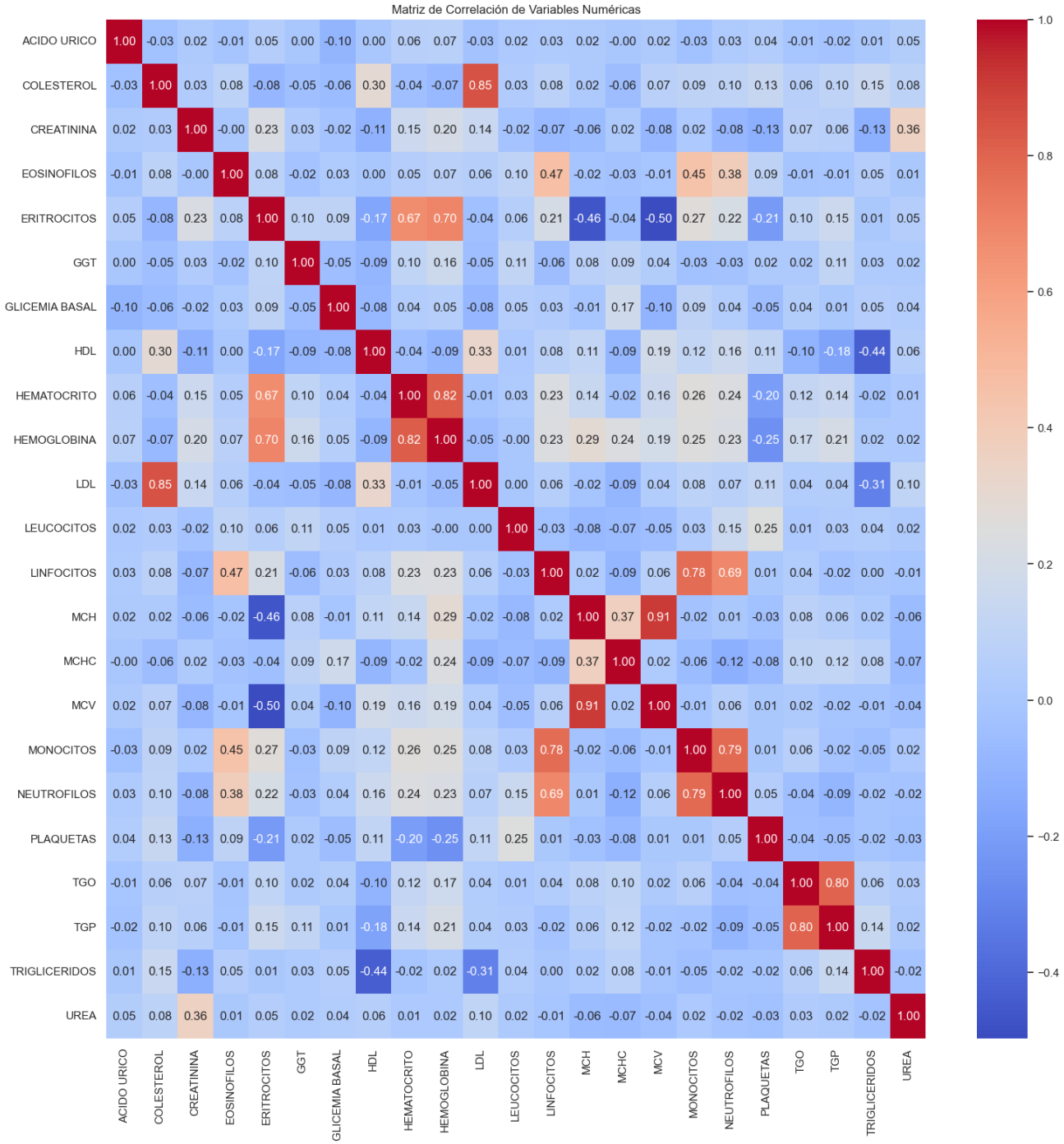
5 rows × 23 columns

In [18]:

```
# Análisis de correlación para variables numéricas
correlation_matrix_imp = data_imputed_df.select_dtypes(include=['float64', 'int64

# Visualización de la matriz de correlación
plt.figure(figsize=(18, 18))
sns.heatmap(correlation_matrix_imp, annot=True, cmap='coolwarm', fmt=".2f")
plt.title("Matriz de Correlación de Variables Numéricas")
plt.show()

```



In [19]:

```
# Identificar pares de variables con alta correlación
# Consideraremos una correlación alta si el valor absoluto es mayor que 0.7

# Crear una máscara para identificar pares de variables diferentes
mask = np.triu(np.ones_like(correlation_matrix_imp, dtype=bool))

# Encontrar pares de correlación que cumplen el criterio
high_corr = correlation_matrix_imp.mask(mask).abs() > 0.7
correlated_pairs = [(var1, var2) for var1 in high_corr.columns for var2 in high_c

correlated_pairs
```

Out[19]:

```
[('COLESTEROL', 'LDL'),
 ('HEMATOCRITO', 'HEMOGLOBINA'),
 ('LINFOCITOS', 'MONOCITOS'),
 ('MCH', 'MCV'),
 ('MONOCITOS', 'NEUTROFILOS'),
 ('TGO', 'TGP')]
```

In [20]:

```
# Identificar las columnas categóricas
categorical_columns = data_cleaned.select_dtypes(include=['object']).columns

# Extraer las variables categóricas
categorical_data = data_cleaned[categorical_columns]

categorical_data

```

Out[20]:

	CODIGO_EMP	CARGO	GENERO	ESTADOCIVIL	DIAGNOSTICO CIE10
0	CE11365	JEFE OPERATIVO	MASCULINO	CASADO	E782-HIPERLIPIDEMIA MIXTA
1	CE11365	JEFE OPERATIVO	MASCULINO	CASADO	E782-HIPERLIPIDEMIA MIXTA
2	CE11365	JEFE OPERATIVO	MASCULINO	CASADO	E039-HIPOTIROIDISMO NO ESPECIFICADO
3	CE11365	JEFE OPERATIVO	MASCULINO	CASADO	E788-OTROS TRASTORNOS DEL METABOLISMO DE LAS L...
4	CE11370	ADMINISTRATIVO 3	MASCULINO	CASADO	R739-HIPERGLICEMIA NO ESPECIFICADA
...	...	...	...	...	...
861	CE13007	OPERATIVO	MASCULINO	SOLTERO	E782-HIPERLIPIDEMIA MIXTA
862	CE13026	OPERATIVO	MASCULINO	SOLTERO	E782-HIPERLIPIDEMIA MIXTA
863	CE13029	OPERATIVO	MASCULINO	SOLTERO	E782-HIPERLIPIDEMIA MIXTA
864	CE13032	ADMINISTRATIVO 3	MASCULINO	SOLTERO	E789-TRASTORNO DEL METABOLISMO DE LAS LIPOPROT...
865	CE13066	JEFE ADMINISTRATIVO	MASCULINO	CASADO	E782-HIPERLIPIDEMIA MIXTA

866 rows × 5 columns

In [21]:
```
# Identificar y extraer las columnas categóricas y PSA_Gender_Combined
categorical_columns = data_cleaned.select_dtypes(include=['object']).columns
categorical_and_psa_data = data_cleaned[list(categorical_columns) + ['AÑO', 'PSA_
                                                                     'TSH_Combined

# Unir las variables categóricas y PSA_Gender_Combined con data_imputed_df
final_dataframe = pd.concat([categorical_and_psa_data.reset_index(drop=True), dat
```

In [22]:
```
final_dataframe.head()
```

Out[22]:

	CODIGO_EMP	CARGO	GENERO	ESTADOCIVIL	DIAGNOSTICO CIE10	AÑO	PSA_Gender_Cor
0	CE11365	JEFE OPERATIVO	MASCULINO	CASADO	E782-HIPERLIPIDEMIA MIXTA	2019	
1	CE11365	JEFE OPERATIVO	MASCULINO	CASADO	E782-HIPERLIPIDEMIA MIXTA	2021	
2	CE11365	JEFE OPERATIVO	MASCULINO	CASADO	E039-HIPOTIROIDISMO NO ESPECIFICADO	2023	
3	CE11365	JEFE OPERATIVO	MASCULINO	CASADO	E788-OTROS TRASTORNOS DEL METABOLISMO DE LAS L...	2023	
4	CE11370	ADMINISTRATIVO 3	MASCULINO	CASADO	R739-HIPERGLICEMIA NO ESPECIFICADA	2019	

5 rows × 34 columns

In [24]:
```
#Para exportar la data lista a un archivo .csv
#final_dataframe.to_csv('Examenes_Diagnosticos_7_03022024_jup_crisp.csv') #data l
```

In [23]:
```
final_dataframe = final_dataframe.drop(columns=['CODIGO_EMP', 'AÑO'])
final_dataframe.head()
```

Out[23]:

	CARGO	GENERO	ESTADOCIVIL	DIAGNOSTICO CIE10	PSA_Gender_Combined	T3_Combined
0	JEFE OPERATIVO	MASCULINO	CASADO	E782-HIPERLIPIDEMIA MIXTA	0.93	3.32
1	JEFE OPERATIVO	MASCULINO	CASADO	E782-HIPERLIPIDEMIA MIXTA	0.27	2.39
2	JEFE OPERATIVO	MASCULINO	CASADO	E039-HIPOTIROIDISMO NO ESPECIFICADO	0.20	3.01
3	JEFE OPERATIVO	MASCULINO	CASADO	E788-OTROS TRASTORNOS DEL METABOLISMO DE LAS L...	0.20	3.01
4	ADMINISTRATIVO 3	MASCULINO	CASADO	R739-HIPERGLICEMIA NO ESPECIFICADA	0.18	-1.00

5 rows × 32 columns

In [24]:

```
# Importando LabelEncoder de sklearn.preprocessing
from sklearn.preprocessing import LabelEncoder

# Codificación de variables categóricas (excepto la variable objetivo CIE10)
label_encoders = {}
for column in final_dataframe.columns:
    if final_dataframe[column].dtype == 'object' and column != 'DIAGNOSTICO CIE10
        le = LabelEncoder()
        final_dataframe[column] = le.fit_transform(final_dataframe[column])
        label_encoders[column] = le

final_dataframe.head()
```

Out[24]:

	CARGO	GENERO	ESTADOCIVIL	DIAGNOSTICO CIE10	PSA_Gender_Combined	T3_Combined	T4_Combin
0	4	1	0	E782-HIPERLIPIDEMIA MIXTA	0.93	3.32	0.
1	4	1	0	E782-HIPERLIPIDEMIA MIXTA	0.27	2.39	1.
2	4	1	0	E039-HIPOTIROIDISMO NO ESPECIFICADO	0.20	3.01	1.
3	4	1	0	E788-OTROS TRASTORNOS DEL METABOLISMO DE LAS L...	0.20	3.01	1.
4	2	1	0	R739-HIPERGLICEMIA NO ESPECIFICADA	0.18	-1.00	-1.

5 rows × 32 columns

In [25]:
```
# Asignar el dataframe final a un nuevo dataframe llamado 'df'
df = final_dataframe

# Verificar las primeras filas del nuevo dataframe 'df'
df.head()

```

Out[25]:

	CARGO	GENERO	ESTADOCIVIL	DIAGNOSTICO CIE10	PSA_Gender_Combined	T3_Combined	T4_Combin
0	4	1	0	E782-HIPERLIPIDEMIA MIXTA	0.93	3.32	0.
1	4	1	0	E782-HIPERLIPIDEMIA MIXTA	0.27	2.39	1.
2	4	1	0	E039-HIPOTIROIDISMO NO ESPECIFICADO	0.20	3.01	1.
3	4	1	0	E788-OTROS TRASTORNOS DEL METABOLISMO DE LAS L...	0.20	3.01	1.
4	2	1	0	R739-HIPERGLICEMIA NO ESPECIFICADA	0.18	-1.00	-1.

5 rows × 32 columns

Modelo Bosques Aleatorios

In [27]:
```
# Separar las variables (X) y la variable objetivo (y)
X = df.drop('DIAGNOSTICO CIE10', axis=1)
y = df['DIAGNOSTICO CIE10']


# Dividiendo el dataset en conjuntos de entrenamiento y prueba nuevamente
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_s

# Verificando las dimensiones de los conjuntos de entrenamiento y prueba
X_train.shape, X_test.shape, y_train.shape, y_test.shape
```

Out[27]:
```
((692, 31), (174, 31), (692,), (174,))
```

In [28]:

```
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import classification_report, accuracy_score

# Creando el modelo de Random Forest
rf_model = RandomForestClassifier(random_state=42)

# Entrenando el modelo con el conjunto de entrenamiento
rf_model.fit(X_train, y_train)

# Realizando predicciones en el conjunto de prueba
y_pred = rf_model.predict(X_test)

# Evaluando el modelo
accuracy = accuracy_score(y_test, y_pred)
classification_rep = classification_report(y_test, y_pred)

accuracy, classification_rep

print('Accuracy: ', accuracy)
print(classification_rep)

```

```
Accuracy:  0.47701149425287354
                                                                                                                precisi
on    recall  f1-score   support

                                                                                                  D45-POLICITEMIA VERA       0.
00      0.00      0.00         1
                                                                                 D729-TRASTORNOS DE LOS LEUCOCITOS NO ESPECIFICADO       0.
00      0.00      0.00         1
                                                                                                 D751-POLICITEMIA SECUNDARIA       0.
33      0.20      0.25         5
                                                                                              E039-HIPOTIROIDISMO NO ESPECIFICADO       0.
09      0.14      0.11         7
  E119-DIABETES MELLITUS NO INSULINODEPENDIENTE SIN MENCION DE COMPLICACION       0.
00      0.00      0.00         3
        E749-TRASTORNO DEL METABOLISMO DE LOS CARBOHIDRATOS NO ESPECIFICADO       0.
00      0.00      0.00        10
                                                                                                     E782-HIPERLIPIDEMIA MIXTA       0.
60      0.74      0.66        87
                                                                                              E785-HIPERLIPIDEMIA NO ESPECIFICADA       0.
20      0.25      0.22         4
                                E788-OTROS TRASTORNOS DEL METABOLISMO DE LAS LIPOPROTEINAS       0.
00      0.00      0.00         7
        E789-TRASTORNO DEL METABOLISMO DE LAS LIPOPROTEINAS NO ESPECIFICADO       0.
20      0.14      0.17         7
E790-HIPERURICEMIA SIN SIGNOS DE ARTRITIS INFLAMATORIA Y ENFERMEDAD TOFACEA       0.
48      0.52      0.50        21
                                  K759-ENFERMEDAD INFLAMATORIA DEL HIGADO NO ESPECIFICADA       0.
00      0.00      0.00         2
                                                                                                     M109-GOTA NO ESPECIFICADA       0.
00      0.00      0.00         3
                                                                                 N189-INSUFICIENCIA RENAL CRONICA NO ESPECIFICADA       0.
00      0.00      0.00         2
                                                                                              R739-HIPERGLICEMIA NO ESPECIFICADA       0.
44      0.29      0.35        14

                                                                                                                accuracy
0.48       174
                                                                                                            macro avg       0.
16      0.15      0.15       174
                                                                                                         weighted avg       0.
42      0.48      0.44       174
```

```
C:\ProgramData\Anaconda3\lib\site-packages\sklearn\metrics\_classification.py:1318:
UndefinedMetricWarning: Precision and F-score are ill-defined and being set to 0.0 i
n labels with no predicted samples. Use `zero_division` parameter to control this be
havior.
  _warn_prf(average, modifier, msg_start, len(result))
C:\ProgramData\Anaconda3\lib\site-packages\sklearn\metrics\_classification.py:1318:
UndefinedMetricWarning: Precision and F-score are ill-defined and being set to 0.0 i
n labels with no predicted samples. Use `zero_division` parameter to control this be
havior.
  _warn_prf(average, modifier, msg_start, len(result))
C:\ProgramData\Anaconda3\lib\site-packages\sklearn\metrics\_classification.py:1318:
UndefinedMetricWarning: Precision and F-score are ill-defined and being set to 0.0 i
n labels with no predicted samples. Use `zero_division` parameter to control this be
havior.
  _warn_prf(average, modifier, msg_start, len(result))
```

Type *Markdown* and LaTeX: α^2

RandomizedSearchCV

In [29]:

```python
from sklearn.model_selection import RandomizedSearchCV
import numpy as np

param_dist = {
    'n_estimators': [50, 100, 150, 500],
    'max_features': ['auto', 'sqrt'],
    'max_depth': [4, 6, 8, 50],
    'criterion': ['gini', 'entropy'],
    'min_samples_split': [2, 5],
    'min_samples_leaf': [1, 2, 4],
    'bootstrap': [True, False]
}

# Creando el modelo de Random Forest para el ajuste de hiperparámetros
rf = RandomForestClassifier(random_state=42)

# Configurando la búsqueda aleatoria con validación cruzada
random_search = RandomizedSearchCV(rf, param_distributions=param_dist, n_iter=10,

# Entrenando el modelo con diferentes hiperparámetros
random_search.fit(X_train, y_train)

# Mejores parámetros encontrados
random_search.best_params_

```

```
C:\ProgramData\Anaconda3\lib\site-packages\sklearn\model_selection\_split.py:676: Us
erWarning: The least populated class in y has only 2 members, which is less than n_s
plits=5.
  warnings.warn(
```

Out[29]:

```
{'n_estimators': 50,
 'min_samples_split': 5,
 'min_samples_leaf': 4,
 'max_features': 'sqrt',
 'max_depth': 6,
 'criterion': 'entropy',
 'bootstrap': False}
```

In [30]:

```python
# Entrenando el modelo Random Forest con los mejores hiperparámetros encontrados
rf_optimized = RandomForestClassifier(
    n_estimators=50,
    min_samples_split = 5,
    min_samples_leaf = 4,
    max_features='sqrt',
    max_depth=6,
    bootstrap = False
)

rf_optimized.fit(X_train, y_train)

# Realizando predicciones en el conjunto de prueba
y_pred_optimized = rf_optimized.predict(X_test)

# Evaluando el modelo optimizado
accuracy_optimized = accuracy_score(y_test, y_pred_optimized)
classification_rep_optimized = classification_report(y_test, y_pred_optimized)

accuracy_optimized, classification_rep_optimized

print(accuracy_optimized)
print(classification_rep_optimized)
```

```
0.5229885057471264
                                                                                  precisi
on    recall  f1-score   support

                                                               D45-POLICITEMIA VERA       0.
00      0.00      0.00         1
                          D729-TRASTORNOS DE LOS LEUCOCITOS NO ESPECIFICADO       0.
00      0.00      0.00         1
                                                        D751-POLICITEMIA SECUNDARIA       1.
00      0.20      0.33         5
                                                E039-HIPOTIROIDISMO NO ESPECIFICADO       0.
00      0.00      0.00         7
  E119-DIABETES MELLITUS NO INSULINODEPENDIENTE SIN MENCION DE COMPLICACION       0.
00      0.00      0.00         3
      E749-TRASTORNO DEL METABOLISMO DE LOS CARBOHIDRATOS NO ESPECIFICADO       0.
00      0.00      0.00        10
                                                          E782-HIPERLIPIDEMIA MIXTA       0.
58      0.90      0.70        87
                                                E785-HIPERLIPIDEMIA NO ESPECIFICADA       0.
00      0.00      0.00         4
                       E788-OTROS TRASTORNOS DEL METABOLISMO DE LAS LIPOPROTEINAS       0.
00      0.00      0.00         7
      E789-TRASTORNO DEL METABOLISMO DE LAS LIPOPROTEINAS NO ESPECIFICADO       0.
00      0.00      0.00         7
E790-HIPERURICEMIA SIN SIGNOS DE ARTRITIS INFLAMATORIA Y ENFERMEDAD TOFACEA       0.
50      0.48      0.49        21
                    K759-ENFERMEDAD INFLAMATORIA DEL HIGADO NO ESPECIFICADA       0.
00      0.00      0.00         2
                                                          M109-GOTA NO ESPECIFICADA       0.
00      0.00      0.00         3
                           N189-INSUFICIENCIA RENAL CRONICA NO ESPECIFICADA       0.
00      0.00      0.00         2
                                                 R739-HIPERGLICEMIA NO ESPECIFICADA       0.
40      0.14      0.21        14

                                                                           accuracy
0.52       174
                                                                          macro avg       0.
17      0.11      0.12       174
                                                                       weighted avg       0.
41      0.52      0.44       174

C:\ProgramData\Anaconda3\lib\site-packages\sklearn\metrics\_classification.py:1318:
UndefinedMetricWarning: Precision and F-score are ill-defined and being set to 0.0 i
n labels with no predicted samples. Use `zero_division` parameter to control this be
havior.
  _warn_prf(average, modifier, msg_start, len(result))
C:\ProgramData\Anaconda3\lib\site-packages\sklearn\metrics\_classification.py:1318:
UndefinedMetricWarning: Precision and F-score are ill-defined and being set to 0.0 i
n labels with no predicted samples. Use `zero_division` parameter to control this be
havior.
  _warn_prf(average, modifier, msg_start, len(result))
C:\ProgramData\Anaconda3\lib\site-packages\sklearn\metrics\_classification.py:1318:
UndefinedMetricWarning: Precision and F-score are ill-defined and being set to 0.0 i
n labels with no predicted samples. Use `zero_division` parameter to control this be
havior.
  _warn_prf(average, modifier, msg_start, len(result))
```

Type *Markdown* and LaTeX: α^2

In [37]:

```python
import matplotlib.pyplot as plt
import seaborn as sns

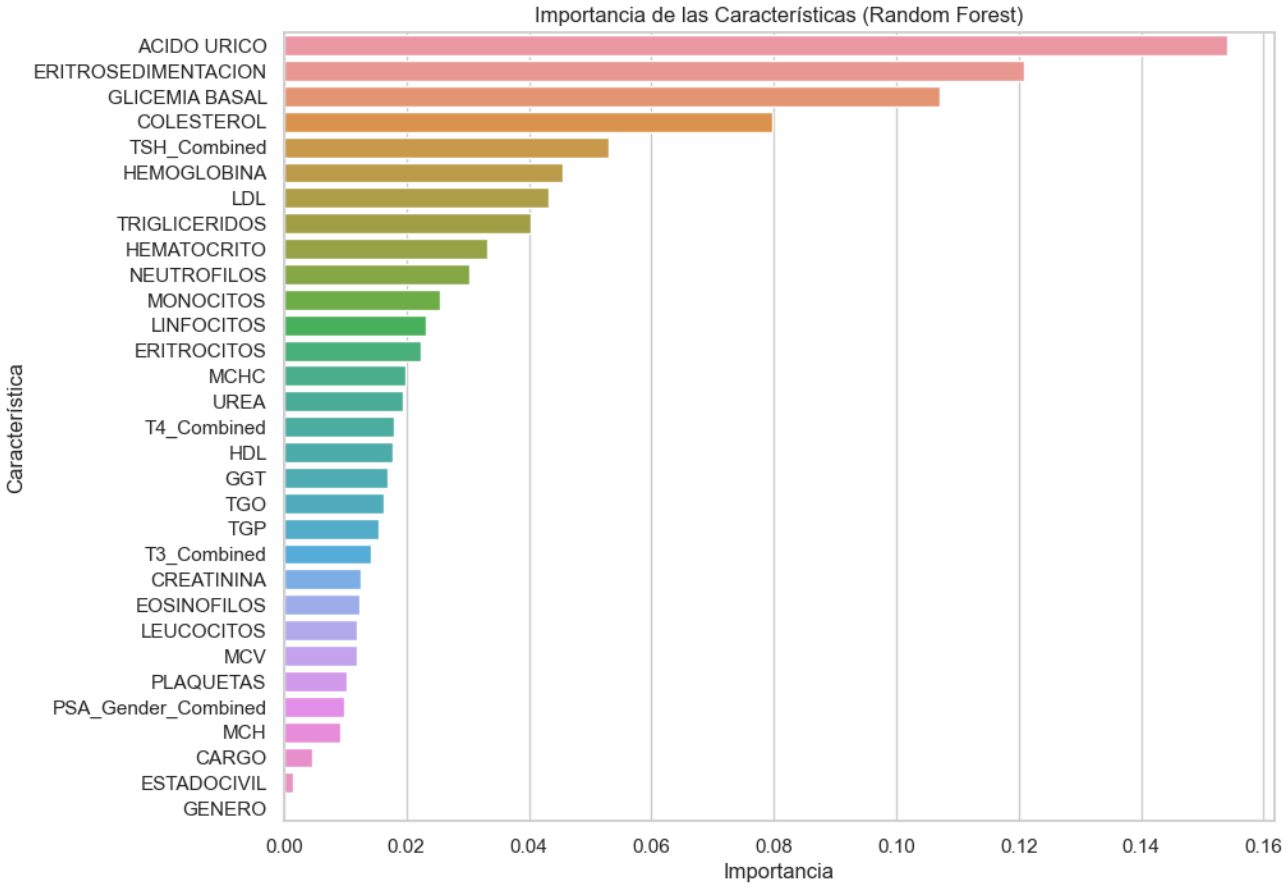
# Obteniendo las importancias de las características
feature_importances = rf_optimized.feature_importances_

# Creando un DataFrame para visualizar las importancias
features_df = pd.DataFrame({
    'Feature': X.columns,
    'Importance': feature_importances
})

# Ordenando las características por importancia
features_df = features_df.sort_values(by='Importance', ascending=False)

# Visualización
plt.figure(figsize=(10, 8))
sns.barplot(x='Importance', y='Feature', data=features_df)
plt.title('Importancia de las Características (Random Forest)')
plt.xlabel('Importancia')
plt.ylabel('Característica')
plt.show()

```



In [ ]:

```python

```

Modelo Máquina de Vectores de Soporte (SVM)

In [26]:

```
from sklearn.preprocessing import StandardScaler
import pandas as pd

# Separar características (X) y la variable objetivo (y)
X = df.drop('DIAGNOSTICO CIE10', axis=1)
y = df['DIAGNOSTICO CIE10']

# Identificar las columnas numéricas
num_columns = X.select_dtypes(include=['float64', 'int64']).columns

# Estandarizar solo las características numéricas
scaler = StandardScaler()
X_scaled = X.copy()
X_scaled[num_columns] = scaler.fit_transform(X[num_columns])

# Mostrar las primeras filas de las características estandarizadas
pd.DataFrame(X_scaled, columns=X.columns).head()

```

Out[26]:

	CARGO	GENERO	ESTADOCIVIL	PSA_Gender_Combined	T3_Combined	T4_Combined	TSH_Combined	EI
0	4	1	0	0.611081	1.638885	1.339542	1.788783	
1	4	1	0	-0.167013	1.154438	1.410396	2.804251	
2	4	1	0	-0.249538	1.477403	1.724176	0.075182	
3	4	1	0	-0.249538	1.477403	1.724176	0.075182	
4	2	1	0	-0.273117	-0.611451	-0.613990	-0.554952	

5 rows × 31 columns

In [26]:

```python
from sklearn.model_selection import train_test_split
from sklearn.svm import SVC
from sklearn.metrics import classification_report, accuracy_score

# Dividir el conjunto de datos en entrenamiento y prueba
X_train, X_test, y_train, y_test = train_test_split(X_scaled, y, test_size=0.2, rand

# Crear y entrenar el modelo SVM
svm_model = SVC(kernel='linear', decision_function_shape='ovr', random_state=42)
svm_model.fit(X_train, y_train)

# Predicciones en el conjunto de prueba
y_pred = svm_model.predict(X_test)

# Evaluación del modelo
accuracy = accuracy_score(y_test, y_pred)
classification_rep = classification_report(y_test, y_pred)

print(accuracy)
print(classification_rep)

```

```
0.4885057471264368
                                                                                precision
recall  f1-score   support

                                                                D45-POLICITEMIA VERA       0.00
0.00      0.00         1
                                  D729-TRASTORNOS DE LOS LEUCOCITOS, NO ESPECIFICADO       0.00
0.00      0.00         1
                                                        D751-POLICITEMIA SECUNDARIA       0.33
0.20      0.25         5
                                                    E039-HIPOTIROIDISMO, NO ESPECIFICADO       0.27
0.57      0.36         7
  E119-DIABETES MELLITUS NO INSULINODEPENDIENTE SIN MENCION DE COMPLICACION       0.67
0.67      0.67         3
       E749-TRASTORNO DEL METABOLISMO DE LOS CARBOHIDRATOS, NO ESPECIFICADO       0.00
0.00      0.00        10
                                                       E782-HIPERLIPIDEMIA MIXTA       0.60
0.78      0.68        87
                                                    E785-HIPERLIPIDEMIA NO ESPECIFICADA       0.00
0.00      0.00         4
                  E788-OTROS TRASTORNOS DEL METABOLISMO DE LAS LIPOPROTEINAS       0.25
0.14      0.18         7
       E789-TRASTORNO DEL METABOLISMO DE LAS LIPOPROTEINAS, NO ESPECIFICADO       0.20
0.14      0.17         7
E790-HIPERURICEMIA SIN SIGNOS DE ARTRITIS INFLAMATORIA Y ENFERMEDAD TOFACEA       0.57
0.19      0.29        21
                    K759-ENFERMEDAD INFLAMATORIA DEL HIGADO, NO ESPECIFICADA       0.00
0.00      0.00         2
                                                       M109-GOTA, NO ESPECIFICADA       0.00
0.00      0.00         3
                       N189-INSUFICIENCIA RENAL CRONICA, NO ESPECIFICADA       0.00
0.00      0.00         2
                                                    R739-HIPERGLICEMIA, NO ESPECIFICADA       0.40
0.29      0.33        14


                                                                     accuracy
0.49       174
                                                                    macro avg       0.22
0.20      0.19       174
                                                                 weighted avg       0.45
0.49      0.45       174
```

```
C:\ProgramData\Anaconda3\lib\site-packages\sklearn\metrics\_classification.py:1318: Und
efinedMetricWarning: Precision and F-score are ill-defined and being set to 0.0 in labe
ls with no predicted samples. Use `zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, msg_start, len(result))
C:\ProgramData\Anaconda3\lib\site-packages\sklearn\metrics\_classification.py:1318: Und
efinedMetricWarning: Precision and F-score are ill-defined and being set to 0.0 in labe
ls with no predicted samples. Use `zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, msg_start, len(result))
C:\ProgramData\Anaconda3\lib\site-packages\sklearn\metrics\_classification.py:1318: Und
efinedMetricWarning: Precision and F-score are ill-defined and being set to 0.0 in labe
ls with no predicted samples. Use `zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, msg_start, len(result))
```

In [27]:

```python
from sklearn.svm import SVC
from sklearn.metrics import accuracy_score, classification_report

# Dividir el conjunto de datos en entrenamiento y prueba
X_train, X_test, y_train, y_test = train_test_split(X_scaled, y, test_size=0.2, rand


# Instanciando el modelo de SVM con parámetros iniciales
svm_model = SVC(kernel='linear', C=1.0, random_state=42)

# Entrenando el modelo con los datos de entrenamiento
svm_model.fit(X_train, y_train)

# Haciendo predicciones con el conjunto de prueba
y_pred_svm = svm_model.predict(X_test)

# Calculando la exactitud del modelo
accuracy_svm = accuracy_score(y_test, y_pred_svm)
print(f"Exactitud del modelo SVM: {accuracy_svm:.4f}")
```

```
Exactitud del modelo SVM: 0.4885
```

In [29]:

```python
from sklearn.model_selection import RandomizedSearchCV

# Definición del espacio de hiperparámetros para explorar
param_distributions_svm = {
    'C': [0.1, 1, 10, 100],  # Regularización
    'gamma': ['scale', 'auto'],  # Coeficiente para kernels 'rbf', 'poly' y 'sigmoid'
    'kernel': ['linear', 'poly', 'rbf', 'sigmoid']  # Tipo de kernel
}

# Configuración de RandomizedSearchCV
random_search_svm = RandomizedSearchCV(estimator=SVC(random_state=42),
                                       param_distributions=param_distributions_svm,
                                       n_iter=5,
                                       cv=3,
                                       verbose=2,
                                       random_state=42,
                                       n_jobs=-1)

# Ejecución de la búsqueda en el conjunto de entrenamiento
random_search_svm.fit(X_train, y_train)

# Mejores hiperparámetros encontrados
best_params_svm = random_search_svm.best_params_
print("Mejores hiperparámetros SVM:", best_params_svm)

# Evaluación del mejor modelo encontrado con el conjunto de prueba
best_model_svm = random_search_svm.best_estimator_
y_pred_optimized_svm = best_model_svm.predict(X_test)
optimized_accuracy_svm = accuracy_score(y_test, y_pred_optimized_svm)
print(f"Exactitud del modelo SVM optimizado: {optimized_accuracy_svm:.4f}")
```

```
Fitting 3 folds for each of 5 candidates, totalling 15 fits

C:\ProgramData\Anaconda3\lib\site-packages\sklearn\model_selection\_split.py:676: UserW
arning: The least populated class in y has only 2 members, which is less than n_splits=
3.
  warnings.warn(

Mejores hiperparámetros SVM: {'kernel': 'sigmoid', 'gamma': 'auto', 'C': 1}
Exactitud del modelo SVM optimizado: 0.5000
```

In [ ]:

```python

```