Universidad Nacional Mayor de San Marcos
Universidad del Perú. Decana de América
Facultad de Ingeniería de Sistemas e Informática
Escuela Profesional de Ingeniería de Sistemas

Aplicación Web Basado en Minería de Datos usando la Técnica de Naive Bayes para la Predicción de la obesidad en edad infantil en los Hospitales Públicos de Lima

TESIS

Para optar el Título Profesional de Ingeniero de Sistemas

AUTOR

Nicole Emily BECERRA ROMERO

ASESOR

Mg. Ana María HUAYNA DUEÑAS

Lima, Perú

2023

Referencia bibliográfica

Becerra, N. (2023). *Aplicación Web Basado en Minería de Datos usando la Técnica de Naive Bayes para la Predicción de la obesidad en edad infantil en los Hospitales Públicos de Lima*. [Tesis de pregrado, Universidad Nacional Mayor de San Marcos, Facultad de Ingeniería de Sistemas e Informática, Escuela Profesional de Ingeniería de Sistemas]. Repositorio institucional Cybertesis UNMSM.

Metadatos complementarios autor/ asesor

Datos de autor	
Nombres y apellidos	Nicole Emily Becerra Romero
Tipo de documento de identidad	DNI
Número de documento de identidad	75310840
URL de ORCID	---------
Datos de asesor	
Nombres y apellidos	Ana María Huayna Dueñas
Tipo de documento de identidad	DNI
Número de documento de identidad	06017183
URL de ORCID	https://orcid.org/0000-0001-7726-8206
Datos del jurado	
Presidente del jurado	
Nombres y apellidos	Norberto Ulises Román Concha
Tipo de documento	DNI
Número de documento de identidad	08510560
Miembro del jurado 1	
Nombres y apellidos	Hugo Rafael Cordero Sánchez
Tipo de documento	DNI
Número de documento de identidad	40512428
Datos de investigación	
Línea de investigación	Sistemas Inteligentes

Grupo de investigación	No aplica
Agencia de financiamiento	Sin financiamiento
Ubicación geográfica de la investigación	Hospital José Agurto Tello - Chosica País: Perú Departamento: Lima Provincia: Lima Distrito: Lurigancho - Chosica Avenida: Arequipa 214 Latitud: -11.9342 Longitud: - 76.6934
Año o rango de años en que se realizó la investigación	2022
URL de disciplinas OCDE	Otras ingenierías y tecnologías https://purl.org/pe-repo/ocde/ford#2.11.02 Nutrición, Dietética https://purl.org/pe-repo/ocde/ford#3.03.04

UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS
FACULTAD DE INGENIERIA DE SISTEMAS E INFORMATICA
Escuela Profesional de Ingeniería de Sistemas

Acta Virtual de Sustentación de Tesis

Siendo las19:00 horas del día 15 de noviembre del año 2023, se reunieron virtualmente los docentes designados como miembros de Jurado de Tesis, presidido por el Ing. Hugo R. Cordero Sánchez, Lic. Norberto Ulises Román Concha (Miembro) y la Mg. Ana María Huayna Dueñas (Miembro Asesor), usando la plataforma Meet (https://meet.google.com/beu-asqh-puj), para la sustentación Virtual de la tesis Intitulada: **"Aplicación Web Basado en Minería de Datos usando la Técnica de Naive Bayes para la Predicción de la obesidad en edad infantil en los Hospitales Públicos de Lima",** de la Bachiller: **Nicole Emily Becerra Romero;** para obtener el Título Profesional de Ingeniero de Sistemas.

Acto seguido de la exposición de la Tesis, el Presidente invitó al Bachiller a responder las preguntas formuladas por los Miembros del Jurado.

El Bachiller, en el curso de sus intervenciones demostró pleno dominio del tema, al responder con acierto y fluidez las preguntas formuladas por los señores miembros del Jurado.

Finalmente habiéndose efectuado la calificación correspondiente por los miembros del Jurado, el bachiller obtuvo la nota de 15 **(** Quince **)**

A continuación, el presidente del Jurado Ing. Hugo R. Cordero Sánchez, declara a la Bachiller **Ingeniero de Sistemas.**

Siendo 19:58 horas, se levantó la sesión.

Ing. Hugo R. Cordero Sánchez
Presidente

Lic. Norberto Ulises Román Concha
Miembro

Mg Ana María Huayna Dueñas
Miembro Asesor

Yo **ANA MARÍA HUAYNA DUEÑAS** en mi condición de asesor acreditado con la **Resolución Directoral N°0002-EPIS-FISI-2020** de la **tesis**/monografía/informe de investigación/trabajo académico, cuyo título es **Aplicación Web Basado en Minería de Datos usando la Técnica de Naive Bayes para la Predicción de la obesidad en edad infantil en los Hospitales Públicos de Lima**, presentado por el **bachiller**/magíster/egresado/licenciado/estudiante **Nicole Emily Becerra Romero** para optar el grado/**título**/especialidad de **INGENIERO DE SISTEMAS**, CERTIFICO que se ha cumplido con lo establecido en la Directiva de Originalidad y de Similitud de Trabajos Académicos, de Investigación y Producción Intelectual. Según larevisión, análisis y evaluación mediante el software de similitud textual, el documentoevaluado cuenta con el porcentaje de **8 %** de similitud, nivel **PERMITIDO** para continuarcon los trámites correspondientes y para su **publicación en el repositorio institucional**

Se emite el presente certificado en cumplimiento de lo establecido en las normas vigentes, como uno de los requisitos para la obtención del grado/ título/ especialidad correspondiente.

Firma del Asesor .....

DNI: 06017183

Nombres y apellidos del asesor:
HUAYNA DUEÑAS, ANA MARIA

Dedicatoria

Dedico la presente tesis a mi madre, mi padre y mis hermanos Gonzalo y Rodrigo.

A mi madre por su apoyo constante y ánimos durante este proceso y a mi padre por haber velado por mi educación siempre.

Agradecimientos

Un gran agradecimiento a mi asesora Ing. Ana María Huayna Dueñas por su guía, paciencia y orientación.

Y al área de Crecimiento y Desarrollo del Hospital de Chosica por brindarme la información adecuada para construir este proyecto de Tesis.

INDICE GENERAL

INDICE DE FIGURAS

INDICE DE TABLAS

Resumen

Hoy en día, la obesidad infantil es una enfermedad que causa preocupación a nivel mundial debido a que es considerada la principal causa de enfermedades crónicas como la diabetes y otras afecciones al sistema respiratorio, además de ser el factor de riesgo más relevante para el COVID-19. Es por esto por lo que se han realizado distintos trabajos de investigación con el fin de predecir la obesidad, estos trabajados abordaron el tema usando distintas técnicas de minería de datos. El presente trabajo de investigación creó un modelo predictivo usando la técnica Naive Bayes bajo la metodología KDD para definir la probabilidad de que un niño va a padecer la enfermedad en algún momento de su vida, esperando obtener una exactitud mayor a 90%. El conjunto de datos utilizado para la implementación del modelo Naive Bayes estuvo compuesta por 770 historias clínicas y contó con 27 variables; esta información fue extraída del aplicativo e-Qhali. El conjunto de datos de para probar el modelo estuvo compuesto por 317 registros, se logró obtener 94.32% de exactitud y 95.23%; estos resultados comparados con los de otras técnicas como Regresión Logística, Random Forest y SVM fueron superiores.

Palabras Claves: Naive Bayes, Machine Learning, Obesidad Infantil, Metodología KDD.

Abstract

Today, childhood obesity is a disease that causes worldwide concern because it is considered the main cause of chronic diseases such as diabetes and other conditions of the respiratory system, as well as being the most relevant risk factor for COVID-19. 19. This is why different research papers have been carried out to predict the disease, these works addressed the issue using different data mining techniques. This research work creates a predictive model using the Naive Bayes technique under the KDD methodology to define the probability that a child will suffer from the disease at some point in her life, hoping to obtain an accuracy greater than 90%. The data set used for the implementation of the Naive Bayes model consisted of 770 medical records and had 27 variables; this information was extracted from the e-Qhali application. The data set to test the model consisted of 317 records, it was possible to obtain 94.32% accuracy and 95.23%; these results compared to those of other techniques such as Logistic Regression, Random Forest and SVM were superior.

Keywords: Naive Bayes, Machine Learning, Childhood Obesity, KDD Methodology.

1. Capítulo 1: Introducción

1.1. Antecedentes del Problema

Desde los años ochenta, los niños han tenido la inclinación por consumir alimentos distintos a los acostumbrados en la familia; desean comer la denominada comida "chatarra" en lugar de una alimentación balanceada; debido a este desorden alimenticio surgieron el sobrepeso y obesidad, actualmente denominados "los problemas más graves en el siglo XXI" según lo indica la Organización Mundial de la Salud [OMS] (2016). "La obesidad es la causante de 4.7 millones de muertes prematuras cada año a nivel mundial, siendo 4 veces más el número de muertos en accidentes de tránsito y 5 veces más el número de fallecidos por VIH." (Becerra Romero & Huayna Dueñas, 2022, pág. 90)

La obesidad altera el normal funcionamiento del sistema pulmonar e inmunitario causando una disminución en la función pulmonar; es por esto por lo que se considera el factor de riesgo más importante en pacientes con COVID-19. El Ministerio de Salud (MINSA-PERU, 2020) informó que el 85.5% de los pacientes fallecidos debido al COVID-19 en el 2020 sufrían de obesidad; Guija & Guija (2020) indicó que en New York las personas con obesidad que contrajeron COVID-19 eran 3.6 veces más propensos a ingresar a UCI y en Francia el riesgo de requerir ventilación mecánica fue de 7 veces más para las personas obesas en comparación con los pacientes de COVID-19 que tenían un rango de IMC saludable.

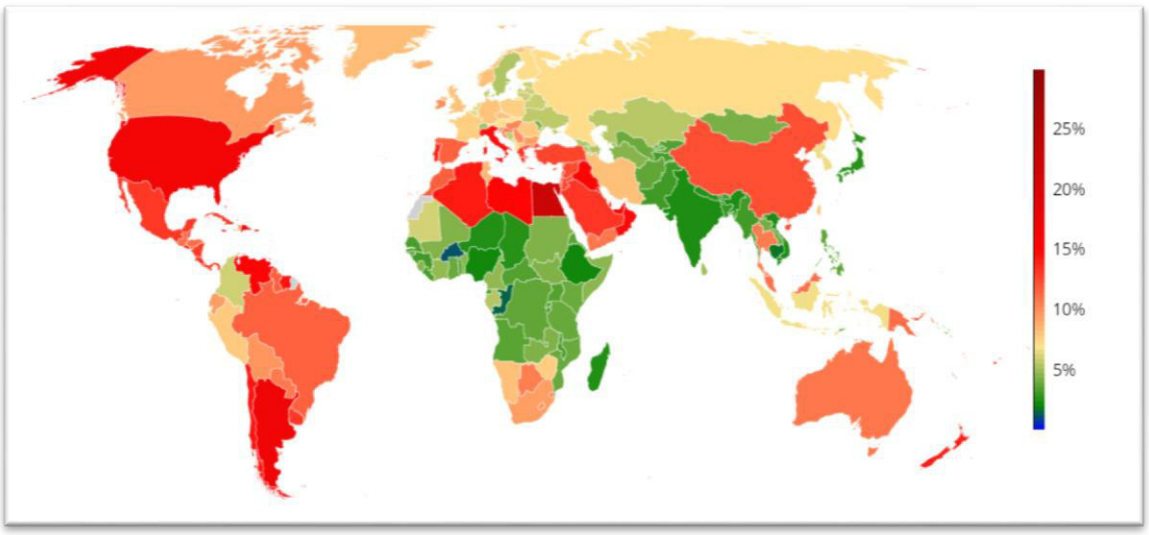
Según la OMS (2021), en el 2016 un total de 41 millones de niños sufrían de sobrepeso u obesidad; y en el continente africano la cantidad de niños con sobrepeso aumentó en un 50% en tan solo 16 años (2000 – 2016).

Las causantes de la obesidad son muy distintas en cada país, según las organizaciones de salud en África existe una ausencia de médicos que orienten a las familias con una dieta balanceada; además de existir una falta de acceso a alimentos saludables.

En la Figura 1, NCDRISC (2016) nos muestra los casos de obesidad infantil a nivel mundial en niñas menores de 17 años, podemos observar que Oceanía y países del norte de África cuentan con una gran cantidad de casos, esto evidencia la necesidad de generar estrategias para combatir la obesidad infantil y de ser posible, tomar acciones para poder prevenirla.

Figura 1.

Mapa de prevalencia del sobrepeso en niñas y adolescentes a nivel mundial



Fuente: *(NCDRISC, 2016)*

Como muestra la tabla 1., es América Latina y el Caribe la región con mayor prevalencia de obesidad en niños con 13.39%, seguida de Asia Central y Oceanía con porcentajes que también son elevados.

Tabla 1.

Prevalencia de obesidad en niños en edad de 5 a 12 años por región

Región	Porcentaje de prevalencia de obesidad
América Latina y el Caribe	13.39 %
Asia Central, Medio Oriente y África del Norte	12.91 %
Oceanía	11.05 %
Europa Central y Oriental	10.36 %
Asia Oriental y Sur oriental	9.69 %
Sur de Asia	2.96 %
África Sub-Sahariana	2.79 %

Fuente: *(NCDRISC, 2016)*

Con respeto a estos altos porcentajes de obesidad Senthilingam (2017) explica que uno de los factores principales de la obesidad en estas regiones es la "desigualdad de actividad", esta desigualdad es el resultado de restar los pasos diarios entre las personas que más caminan y las que menos actividad realizan en una determinada región; la diferencia de pasos es directamente proporcional a la tasa de obesidad. La aplicación móvil "Azumio Argus" recolectó la cantidad de pasos diarios que caminaban las personas en distintas partes del mundo y los países con mayor "desigualdad de actividad" fueron Estados Unidos, Australia; Arabia Saudita, Egipto y Canadá.

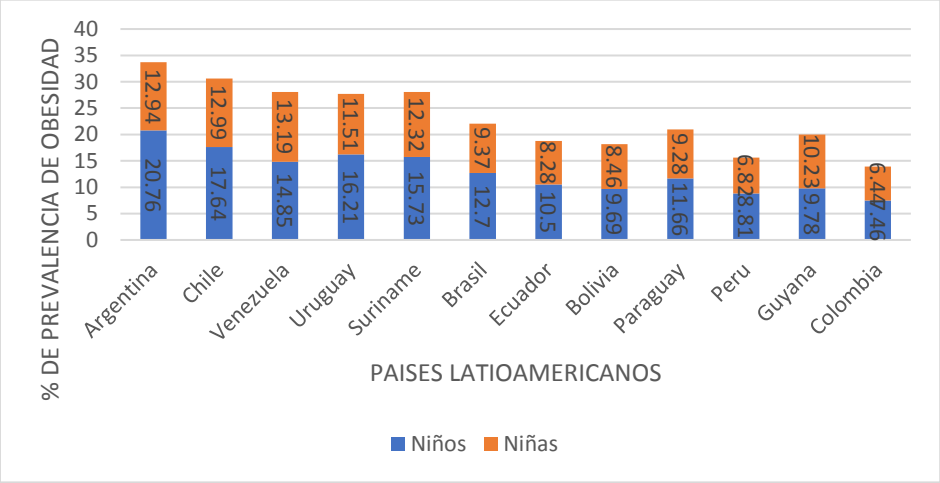
Analizando el problema en el continente americano, la figura 2 nos muestra los porcentajes de obesidad en niños menores de 12 años en el 2019, siendo Argentina el país con mayor cantidad de casos con un 13% de prevalencia en niñas y 20% en niños; muy por encima de nuestro país que se encuentra en décimo lugar en Sudamérica.

Según Orgaz (2019) indica que uno de los principales factores de obesidad infantil en nuestro continente es la industrialización de las ciudades, ya que con la tecnología la

población tiene una vida más sedentaria; otro factor importante es la falta de conocimiento de los padres de familia sobre el tema, es por ello que no acuden a especialistas en busca de orientación. La figura 3 nos muestra que en todos los países los porcentajes de sobrepeso están por encima del 24% y los porcentajes de obesidad no bajan de 7%.

Figura 2.

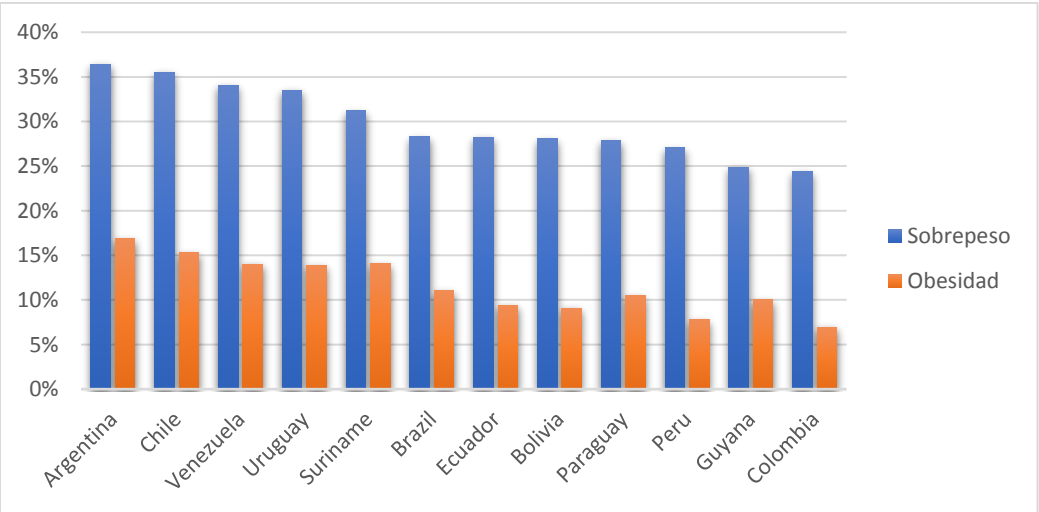
Porcentaje de obesidad en niños y niñas de 12 años en países de América Latina



Fuente: *(MINSA-PERU, 2020)*

Figura 3.

Porcentaje de sobrepeso y obesidad en niños y niñas de 12 años en países de América Latina



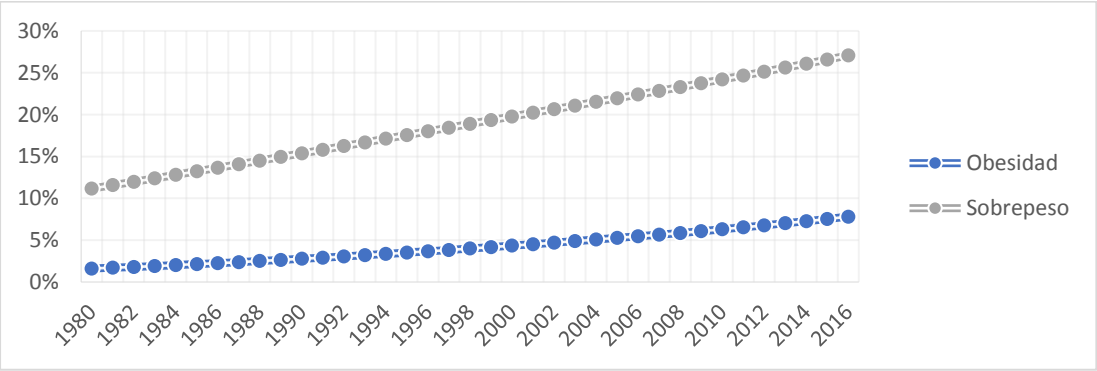
Fuente: *(MINSA-PERU, 2020)*

La razón de que Argentina ocupe el primer lugar en obesidad infantil según el Ministerio de Salud de Argentina (MINSA - ARGENTINA, 2017) es que solo el 17.6% consumen 5 porciones diarias de frutas y verduras, el 50% consumen 2 o más bebidas azucaradas por día y están expuestos a más de 60 anuncios publicitarios de alimentos no saludables diariamente. En Chile, que ocupa el segundo lugar en el gráfico, el Instituto de Políticas Públicas en Salud (IPSUSS, 2021) señala que una las principales razones para padecer obesidad es el país es la falta de actividad física en la población y los trastornos de sueño.

La situación en el Perú es similar a la de los demás países de Latinoamérica, en la Figura 4 podemos observar que en un rango de 16 años (1980 – 2016) la tasa de obesidad aumentó de 2% a 8% y el sobrepeso varió del 11% al 27% (NCDRISC, 2016). CENAN (2019) nos muestra en la figura 5 que la tasa de prevalencia de obesidad infantil aumento de 1.5% a 1.9% en 4 años (2015 – 2019), y que los departamentos con mayores porcentajes de prevalencia de sobrepeso y obesidad fueron Ayacucho, Lambayeque y La Libertad (figura 6).

Figura 4

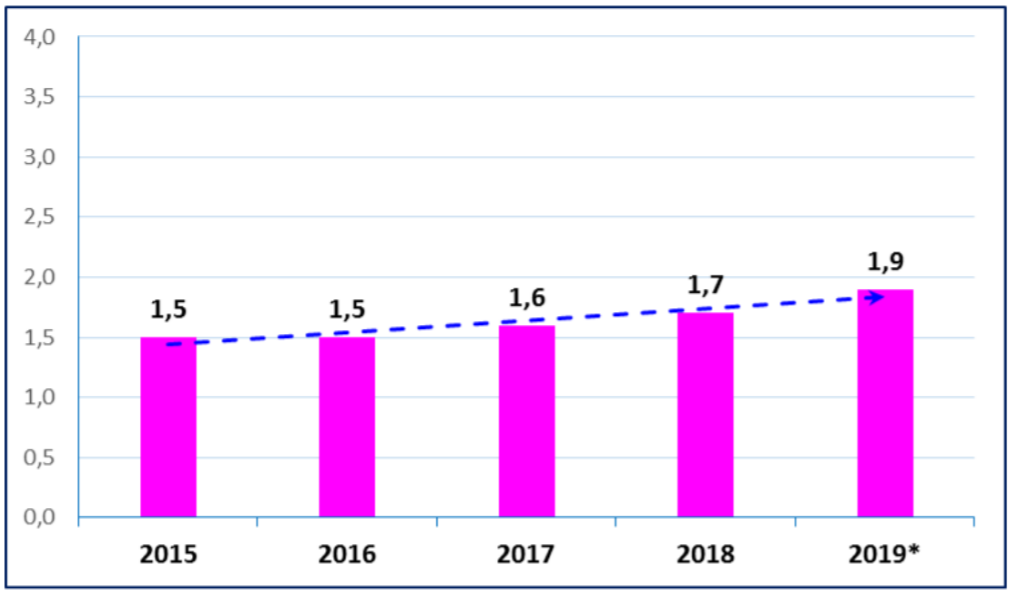
Variación de obesidad y sobrepeso en el Perú a lo largo de la historia



Fuente: *(NCDRISC, 2016)*

Figura 5.

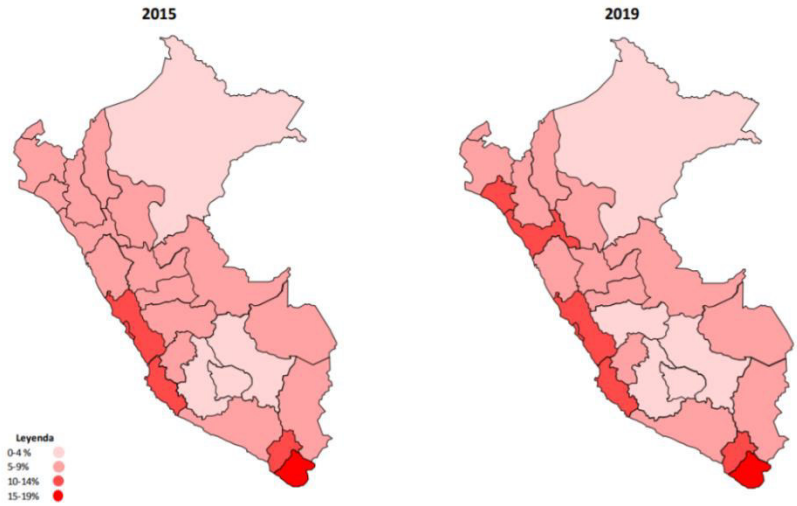
Evolución de la prevalencia de obesidad en niños menores de 5 años, del 2015 al 2019 en el Perú



Fuente: *(CENAN, 2019)*

Figura 6.

Porcentaje de peso (sobrepeso + obesidad) en niños menores de 5 años. Del 2015 al 2019



Fuente: *(CENAN, 2019)*

Actualmente existen gran cantidad de proyectos e iniciativas que tiene como objetivo el prevenir la obesidad infantil, una de ellas es Alva et al. (2020) que en su artículo "Propuesta de un modelo difuso para determinar sobrepeso y obesidad en niños y adolescentes" publicado en la revista de nutrición chilena, utilizaron lógica difusa para la creación de un modelo predictivo que mediante 81 reglas detectaba el sobrepeso y obesidad infantil con un 95.5% de precisión. Ticona (2018) en su tesis de pregrado llamada "Sistema para la predicción de obesidad en la adolescencia utilizando técnicas de minería de datos" que hizo en la Universidad Católica Santa María de Arequipa utilizó la técnica J48 para implementar un sistema predictivo con 94.39% de precisión en el proceso de validación.

En este estudio, se creó una aplicación web utilizando la técnica de Naive Bayes. Comparada con otras técnicas de minería de datos, esta técnica destaca por su eficiencia computacional. Requiere un conjunto de datos de entrenamiento más pequeño, y exhibe una solidez natural ante los datos faltantes y el ruido. Por lo tanto, Al-Aidaroos et al. (2018) consideran a Naive Bayes como una de las opciones más consistentes para apoyar las decisiones médicas. Según Langarizadeh y Moghbeli (2016) esta técnica actualmente es aplicada en muchas áreas de la medicina para diagnosticar tumores cerebrales, cáncer de próstata, glaucoma severo, demencia en pacientes con Parkinson y muchas enfermedades más.

1.2. Formulación del Problema

1.2.1 Problema Principal

Identificar con anticipación cual es la probabilidad de que un niño padezca obesidad es el principal problema del presente trabajo. Actualmente no existe un método aceptado y usado internacionalmente para predecir la obesidad infantil, debido al constante cambio de

peso y talla que experimentan los niños y niñas durante su crecimiento, por ello no se puede evitar las enfermedades crónicas (obesidad, males cardiacos) que estos puedan padecer en un futuro.

1.2.2 Problemas Específicos

- Problemas de salud ocasionados por la obesidad (obesidad, cardiacos, tabaquismo).
- Problemas de falta de capacitación al personal médico con respecto a la importancia de combatir la obesidad infantil.
- Problemas de falta de un sistema de información hospitalario que nos permita contar con el historial clínico del paciente para realizar un correcto seguimiento.
- Problema de falta de concientización de los padres al no conocer el estado nutricional de sus menores hijos.

1.3. Justificación de la Investigación

La obesidad infantil es una enfermedad cuyas consecuencias son muy graves si no es tratada a tiempo, los niños que la padecen son más propensos a sufrir de obesidad en la adultez. Prevenir esta enfermedad a temprana edad contribuye en la reducción de enfermedades coronarias y cerebrovasculares según (Tupia Vargas, 2019)

De acuerdo con Liria (2012) la obesidad en niños y adolescentes trae consecuencias a corto y largo plazo, entre las principales están las alteraciones metabólicas, la diabetes tipo 2 y la tolerancia alterada a la glucosa. También describe que los niños que padezcan esta enfermedad pueden llegar a sufrir discriminación social, baja autoestima y discriminación. A largo plazo los niños que sufrieron obesidad serán adultos obesos que padecerán de enfermedades crónicas, las cuales aumentarán su tasa de mortalidad.

Son muchos los estudios acerca de los efectos de la obesidad en los menores de edad; como el de Cigarro et al. (2016) en el cual recopilaron información de trabajos que relacionaban las consecuencias de la obesidad con alteraciones en el sistema psicomotor en países Latinoamericanos, este trabajo fue aplicado a 284 niños entre 6 y 10 años con el objetivo de analizar la evolución de la edad cronológica con la edad motora general en los niños, el resultado de este análisis fue que los niños mostraron un retraso en el equilibrio, esquema corporal y organización espacial.

Nuestro país no es ajeno a este problema, Diario La República (2019) indicó que cada mes 21 niños son diagnosticados con diabetes en los hospitales de ESSALUD y el 40% de la población escolar está padeciendo de sobrepeso u obesidad. Según Manrique et al. (2015) en su estudio realizado en los hospitales Arzobispo Loayza y Cayetano Heredia, de 25 niños con diabetes mellitus 2 se obtuvo una media de IMC alto de más de 32.8 kg/m^2; y en otra investigación realizada por Medina (2014) en una escuela de Trujillo de un total de 25 niños con prehipertensión o hipertensión el 56% sufría de obesidad.

Para abordar este gran problema actualmente en el Perú existen centros como CENAN y el Instituto de Investigación Nutricional (IIN) que amparados por la ley 30021, "ley de promoción de la alimentación saludable" cumplen con brindar asistencia a personas que sufren este mal, y se destaca dentro de sus principales funciones las siguientes:

- Prestar labor asistencial de salud y nutrición a personas necesitadas, con énfasis en la prevención, como parte de las actividades de investigación.
- Realizar un monitoreo continuo de la situación alimentaria y nutricional, estableciendo y actualizando regularmente la información relevante. Además, llevar a cabo investigaciones con el propósito de determinar los indicadores e índices más apropiados para evaluar el estado de nutrición de la población de manera efectiva.

- Implementar, desarrollar, mantener y proponer los objetivos y lineamientos generales del Sistema de Gestión de Calidad en el campo de su competencia según la política de calidad del Instituto Nacional de Salud (INS)

El poder predecir la obesidad infantil haría más eficiente la labor de los médicos en el área de Nutrición en los Hospitales Públicos de Lima debido a que podrían identificar a los niños que tienen más riesgo de sufrir estos males. Implementar este proceso trae como consecuencia reducir la incertidumbre y el tiempo que tardan los profesionales de la salud en brindar su diagnóstico.

1.4. Alcances de la Investigación

- La aplicación web solo podrá ser utilizada por expertos debido a que es una herramienta de apoyo al médico.
- Los pacientes que sean evaluados por el modelo de predicción tienen que ser mayores a 5 años y contar de manera obligatoria con su cartilla de crecimiento y desarrollo otorgado por el MINSA para que el médico ingrese los datos exactos del paciente en la aplicación y así evitar posibles errores de predicción.
- La aplicación permite iniciar sesión en el sistema, crear el registro de un nuevo paciente, realizar predicciones y ver el historial de predicciones por paciente y doctor.

1.5. Limitaciones de la Investigación

a) Limitación Espacial

Esta investigación abarcara como locación espacial el área de Crecimiento y Desarrollo del Hospital José Agurto Tello de Chosica

b) Limitación Social

La aplicación web podrá ser utilizada por el medicó en niños que tengan entre 5 y 7 años

de edad, no es seguro indicar que el mismo resultado se cumpla para una edad mayor del paciente.

Debido a la pandemia, hubo poca afluencia de pacientes pediátricos en el hospital, por ello al momento de definir el conjunto de datos se trabajó con 770 pacientes.

c) Limitación Técnica

La aplicación web inicialmente cuenta con los módulos de registro de paciente, búsqueda del paciente y predicción del estado nutricional del paciente.

La web y el modelo predictivo están desarrollados con Python 3.7, y se utiliza un servidor heroku para el despliegue. El patrón de diseño es MVC y la base de datos utilizada es PostgreSQL.

1.6. Objetivos

1.6.1 Objetivo Principal

Implementar un modelo predictivo utilizando Naive Bayes para predecir la obesidad infantil en los Hospitales Públicos de Lima y desarrollar una aplicación web que nos permita interactuar con el modelo resultante.

1.6.2. Objetivos Específicos

- Recolectar los datos y seleccionar variables importantes para diagnosticar la obesidad infantil.
- Estudiar las distintas técnicas de minería de datos existentes que se puedan aplicar al problema para obtener una solución
- Identificar el modelo que nos brindara un mejore precisión en la predicción de obesidad infantil y evaluar su efectividad utilizando datos de pacientes.
- Desarrollar la aplicación web para interactuar con el modelo seleccionado

1.7. Propuesta

Crear una aplicación web que permita predecir la obesidad en edad infantil utilizando información de la historia clínica de un paciente, para prevenir y reducir la probabilidad de que un niño padezca esta enfermedad en el futuro.

1.8. Organización de la Tesis

Para fines de estructuración de la tesis, esta se organizó en siete capítulos, los cuales se mencionan a continuación:

- El Capítulo 1 abarca los antecedentes y definición del problema, así como también describe la justificación y el objetivo principal y los específicos.
- En el Capítulo 2 se desarrolla el marco teórico, donde se explica a detalle temas relacionados a la investigación
- El Capítulo 3 ahonda en las distintas técnicas de Minería de Datos existentes y abarca la revisión de trabajos aplicados a la predicción de la Obesidad Infantil.
- En el Capítulo 4 se desarrolla a profundidad la técnica Naive Bayes, sus características principales y se explica la metodología a utilizar en la creación del modelo predictivo
- En el Capítulo 5 se describe como es que se aplica cada fase de la metodología KDD a nuestra investigación, desde la definición y limpieza del conjunto de datos hasta la definición de las métricas que se emplearán.
- En el Capítulo 6 se define el diseño de la aplicación web, la arquitectura del sistema, la descripción de los casos de uso y el modelado de datos.
- En el Capítulo 7 se desarrolla la validación del sistema, en la cual se comparan los resultados obtenidos por el sistema contra el diagnóstico del experto con la finalidad de medir su grado de sensibilidad.

- En el Capítulo 8 se mencionan las conclusiones a las que se llegó con la realización de la tesis y los estudios a futuro que se pudiesen desarrollar a partir de esta

2. Capítulo 2: Marco Teórico

2.1. Minería de Datos

Según Novoseltseva (2021) la minería de datos analiza grandes volúmenes de información para encontrar patrones que permitan a las empresas a resolver problemas, visualizar oportunidades de negocio o reducir los riesgos. La minería de datos combina herramientas estadísticas con la gestión de bases de datos. Para obtener resultados oportunos se debe seguir un proceso estructurado de 6 pasos:

- Comprensión comercial: Trata de una compresión profunda del proyecto, el objetivo comercial y los criterios de éxito a considerar.
- Comprensión de datos: Se determina los datos que se utilizarán y se recopila la información de distintas fuentes.
- Preparación de Datos: Preparar los datos en un formato adecuado, solucionar problemas de calidad de datos.
- Modelado: Se utiliza los algoritmos de minería de datos para encontrar patrones dentro de la información.
- Evaluación: Se determina en cuanto aportará a los objetivos la información obtenida en el modelado.
- Despliegue: Colocar los resultados obtenidos a disposición de los analistas para la toma de decisiones.

Entre los beneficios que nos brinda la minería de datos esta la oportunidad de tomar decisiones rápidas y oportunas para el negocio; otros de los beneficios es que es fácil de implementar en sistemas antiguos o existentes de la empresa. Existen varias áreas donde la minería de datos es aplicable como por ejemplo el marketing mejorando la segmentación del mercado, los bancos para comprender mejor el riesgo financiero de un cliente dependiendo de

sus características, el comercio donde facilita las ventas cruzadas mediante una web y la medicina donde permite diagnósticos más precisos utilizando el historial médico de un paciente.

2.2. Control de Crecimiento y Desarrollo (CRED)

El Área de Control de Crecimiento y Desarrollo (CRED) debe estar implementada en todos los establecimientos médicos del Perú, este servicio es gratuito para niños y niñas que cuentan con el Seguro Integral de Salud y para los que carecen de algún otro seguro de salud. Cada establecimiento debe contar con personal capacitado para realizar el control de manera adecuada.

El proceso CRED según detalla el MINSA (2017) se divide en 4 procesos:

- Valoración: Implica la recopilación de datos mediante la observación, entrevistas, exploración física, exámenes de ayuda diagnostica, etc.
- Diagnóstico: Se analiza los datos recopilados en la fase de valoración, y en base a este análisis se diagnostica al niño o niña.
- Intervención: Se genera acciones a seguir para que sean aplicadas por el niño, familia o comunidad.
- Seguimiento: El médico realiza el acompañamiento de niño o niña durante todo el proceso de intervención.

La periodicidad de los controles CRED varía según la edad; se realiza cada mes hasta que el menor cumpla un año de edad, luego de evalúa bimestralmente hasta los dos años, trimestralmente hasta los tres años y semestralmente hasta los 5 años de edad. Esta información recopilada se registra en una cartilla de uso único del paciente.

2.3. Obesidad Infantil

Según Mayo Clinic (2019) la obesidad infantil es un problema de salud que afecta a niños y adolescentes. Esta afección causa baja autoestima en el niño y además provoca problemas como presión arterial, colesterol alto y diabetes. Existen varios factores de riesgo que aumentan la probabilidad de que un menor padezca obesidad, entre los principales están:

- Alimentación: El consumo regular de comida procesada y dulces pueden causar el aumento de peso.
- Falta de ejercicio: Una vida sedentaria y el no hacer ejercicio diariamente evita que el niño pueda perder peso.
- Factores familiares: En un hogar donde los padres sufren de obesidad, existe mayor probabilidad que el niño la padezca
- Factores psicológicos: Un niño que sufra de estrés personal o ansiedad canalizará sus emociones aumentando la cantidad de comida que consume

2.4. Aplicación Web

Una aplicación web es un programa informático que utiliza tecnología en línea para realizar una gran variedad de tareas diferentes, así la define O´brien (2021). Muchas aplicaciones se utilizan con fines de venta minorista en línea, sin embargo, pueden servir para todo tipo de propósitos diferentes, desde pedir comida para llevar, reservar vacaciones o un formulario médico línea.

Hay innumerables beneficios para las aplicaciones web. En particular, ayudan a reducir costos para empresas y usuarios individuales. Esto se debe a que requieren menos mantenimiento y también pueden tener requisitos más bajos para las computadoras de los usuarios (en términos de potencia de procesamiento, etc. Finalmente, otro beneficio de usar

aplicaciones web es que las actualizaciones son automáticas y, debido a que se aplican de manera centralizada, todos los usuarios deberían estar trabajando desde la misma versión.

Dentro del proceso de desarrollo de una aplicación web se deben considerar 6 criterios claves:

- Concepto: Al desarrollar una aplicación web, tener el concepto claro es uno de los requisitos previos. Este es el punto de partida para todos los que desarrollan una nueva aplicación. Debe tener una idea clara de por qué su aplicación es necesaria y por qué podría ser útil para los usuarios.
- Innovación**:** Otra consideración importante es la innovación, se debe desarrollar una aplicación web que sea útil y valiosa para los usuarios; pero también se debe tener en cuenta en que están trabajando sus competidores y en base a sus deficiencias aprovecharlas para entregar un mejor producto.
- Diseño: El diseño riguroso es vital para desarrollar una aplicación web exitosa. Al diseñar una aplicación, se debe priorizar la experiencia del usuario. La interfaz de usuario debe ser visualmente atractiva y fácil de entender incluso para los usuarios novatos.
- Desarrollo: Los desarrolladores deben tener una guía clara y acceso a las herramientas y scripts que necesitan. Ya sea un desarrollador back-end o front-end, debe tener especificaciones y objetivos precisos para trabajar.
- Entrega: La aplicación web debe ser lo suficientemente robusto como para soportar todas las solicitudes y así evitar algún inconveniente.

3. Capítulo 3: Estado del Arte

En el presente capítulo se realiza una revisión de la literatura en cuanto a artículos y tesis publicados en los últimos años que nos permitan estudiar la predicción de la obesidad infantil. Se lleva a cabo un análisis de los casos de éxito presentes en la literatura de manera cronológica, se describen los métodos que se utilizaron para afrontar el problema en estudio, se explica con detalle cómo se trabajó en cada caso de éxito, y finalmente se revisa las aplicaciones webs o software existentes en el mercado para predecir la obesidad y también las herramientas que nos permitan crear modelos predictivos.

3.1. Revisión de la Literatura

Umoh e Isong (2015) en su artículo llamado "Design Methodology of Fuzzy Expert System for the Diagnosis and Control of Obesity" presentado a la revista "Computer Engineering and Intelligent Systems" diseñaron un sistema al cual llamaron FESDMO, utilizaron reglas difusas de tipo Mandani para poder manejar los datos faltantes; este sistema facilitó el diagnóstico del estado nutricional de los pacientes para saber si se encontraban en un estado saludable , de sobrepeso o si sufrían de obesidad, finalmente el sistema fue validado en Matlab.

Suca, Cordova, Condori y Cayra (2016) de la Universidad Nacional de San Agustín de Arequipa – Perú (UNSA) presentaron su artículo titulado "Modelo difuso para la predicción de casos de obesidad empleando el árbol GFID3 generalizado", en este trabajo desarrollaron un modelo de clasificación que permitió diagnosticar la obesidad en niños y adolescentes en el rango de 6 a 17 años. Este modelo resultó con una exactitud de 76.13% y 83.65% para mujeres y hombres respectivamente.

Suca, Córdova, Condori, Cayra y Sulla. (2016) de la UNSA en su artículo "Comparación de Algoritmos de Clasificación para la Predicción de Casos de Obesidad

Infantil" estudiaron diversas técnicas de clasificación con la finalidad de comparar sus resultados al utilizar un mismo conjunto de datos; entre las técnicas estudiadas estuvieron Naive Bayes , Arboles de Decisión , Back Propagation, entre otros; como resultado se obtuvo que los Árboles de Decisión tuvieron una mejor precisión en comparación con el resto de métodos.

Morlán et al. (2017) , en su artículo del Boletín de la Sociedad de Pediatría de Aragón, La Rioja y Soria de España publicó un "Modelo estadístico para la prevención precoz de desarrollo de sobrepeso/obesidad en población infantil". La investigación logró predecir la obesidad infantil construyendo un modelo de regresión logística usando 14 variables como parámetros de ingreso; el estudio se realizó con 242 niños y los resultaron fueron óptimos ya que el modelo obtuvo una sensibilidad de 96.65% y 96.3% para niñas y niños respectivamente.

Sulla et al. (2018) publicaron el artículo "Application of the ANFIS Neuro-Fuzzy model for the classification of obesity in children and adolescents" en la 16° LACCEI, este trabajo clasificó la obesidad entre niños y adolescentes aplicando una red neuronal artificial basada en el sistema de inferencia difuso Takagi-Sugeno-Kang. Para la implementación se utilizó MATLAB ya que dentro de sus herramientas tenía módulos ya desarrollados para el entrenamiento y validación del modelo. Los resultados revelaron un 96.96% de exactitud en la clasificación.

Ticona (2018) presentó su tesis de pregrado titulada "Sistema para la predicción de Obesidad en la Adolescencia utilizando técnicas de minería de datos" en la Universidad Católica de Santa María (UCSM) de Arequipa – Perú. El método utilizado para el modelo fue el Árbol de Decisión J48, este se implementó en la herramienta Weka; al ser probado se

obtuvo una precisión de 94.39%; resultado que fue superior al de otras técnicas también evaluadas.

Hammond et al. (2019) público el artículo "Predicting childhood obesity using electronic health records and publicly available data" en la revista PLOS ONE Vol. 4°, en su trabajo utilizaron registros médicos con información historia de la madre y del niño para la implementación de modelos de Regresión Logística, Random Forest y Lasso, al comparar los modelos se concluyó que algunas de las características más importantes que influyen en la predicción de obesidad es el peso y talla al nacer, y se obtuvo un AUC de 81.7% para las niñas y del 76.1% para los niños.

Rossman et al. (2021), publicaron el artículo "Prediction of Childhood Obesity from Nationwide Health Records" en la revista "The Journal of Pediatrics" en el cual diseñaron un modelo predictivo capaz de identificar a los niños con un posible riesgo de obesidad utilizando registros médicos de los 2 primeros años de vida de cada menor; trabajaron con un total de 136 196 niños y el modelo implementado Gradiente Boosting arrojó un AUC de 80.4%.

3.2. Técnicas Previamente Aplicadas

3.2.1. Sistemas de Inferencia Borrosa

Diciembre (2017) define los sistemas de inferencia borrosa (SIB) como sistemas expertos que mapean un vector de entradas a una salida única utilizando lógica borrosa. En la figura 7. se muestra la arquitectura de un SIB mostrando los módulos que lo conforman y la forma en que se relacionan.

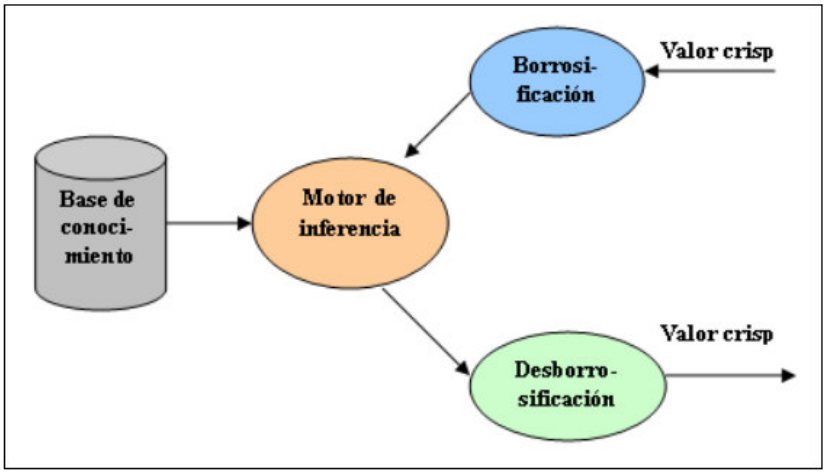
Maguiña (2010) indica que un SIB está compuesto por 4 módulos; el módulo de borrosificación recibe las variables de entrada y los convierte en conjuntos borrosos aplicando la función de borrosificación. La base de conocimiento almacena las reglas definidas por los

expertos, el motor de inferencia ejecuta las reglas sobre la información de entada y, por último, el módulo de desborrosificación se encarga de convertir a un valor numérico el conjunto borroso ejecutando el motor de inferencia.

Los sistemas de inferencia borrosa admiten datos imprecisos, los conceptos matemáticos dentro de razonamiento son muy simples; además son de fácil escalabilidad. Los dos tipos de SIB más usados son Inferencia de Mamdani y de Takagi-Sugeno-Kang (TSK).

Figura 7.

Arquitectura de un SIB.

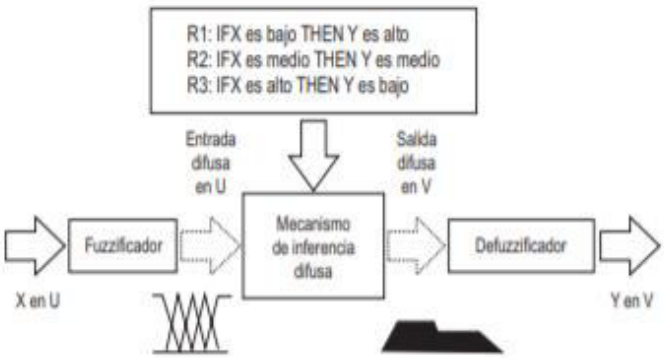


Fuente: *(Diciembre, 2017)*

Inferencia Mamdani

El método de Inferencia Mamdani es el más utilizado y fue propuesto por Ebrahim Mamdani en 1975. Según indica Diciembre (2017) este método se realiza en 4 pasos como se observa en la figura 8, el primero es la fuzzificacion de las variables de entrada, en el segundo paso se evalúan las reglas, luego se procede con la agregación de las salidas de las reglas y por último se realiza la defusificación.

Figura 8.
Arquitectura y módulos de un SIB.



Fuente: *(Maguiña, 2010)*

Inferencia de Takagi-Sugeno-Kang.

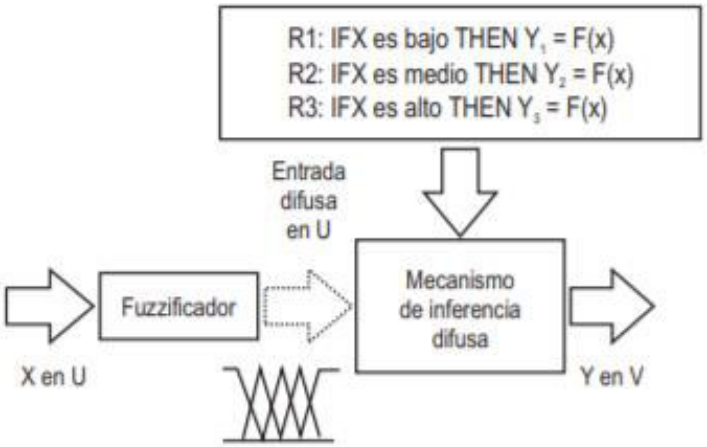
Diciembre (2017) narra que el método TSK fue definido por Takagi y Sugeno en 1985 como método alternativo al método Mandani. La principal diferencia es que este nuevo método tiene como consecuentes funciones lineales (Figura 9) y ya no es necesario el proceso de defuzzificacion ya que el resultado no es un conjunto difuso sino un conjunto de funciones lineales como se observa en la arquitectura (Figura 10).

Figura 9.
Reglas de Inferencia TSK.

$$R_1\text{: Si } x \text{ es } A_1 \wedge y \text{ es } B_1 \text{ entonces } z = f_1(x, y)$$
$$R_2\text{: Si } x \text{ es } A_2 \wedge y \text{ es } B_2 \text{ entonces } z = f_2(x, y)$$
$$\vdots$$
$$R_n\text{: Si } x \text{ es } A_n \wedge y \text{ es } B_n \text{ entonces } z = f_n(x, y)$$

Fuente: (Diciembre, 2017)

Figura 10.
Arquitectura de un SIB sin módulo de Defuzzificacion.



Fuente: (Maguiña, 2010)

3.2.2. Sistemas Neuro – Difusos

Los sistemas Neuro-difusos tienen origen en la inteligencia artificial, según (Chahuara, 2005) este tipo de sistema engloba un conjunto de técnicas que manejan de manera robusta información ruidosa e imprecisa. Las técnicas Neuro Difusas pueden ser: lógica difusa, redes neuronales, algoritmos genéticos, teoría del caos, etc., dichas técnicas pueden ser combinadas para obtener mejores resultados aprovechando las ventajas individuales.

Los sistemas Neuro-Difusos combinan la capacidad de aprendizaje de las RNA con el poder de interpretación lingüística de los sistemas de inferencia difusos, las características de ambos métodos se ven en la figura 11.

Figura 11 .
Características de Lógica Difusa y Redes Neuronales

Lógica Difusa	Redes Neuronales
Permite utilizar el conocimiento disponible para optimizar el sistema directamente.	No existe un método sencillo que permita modificar u optimizar la red, ya que esta se comporta como una "caja negra"
Permite describir el comportamiento de un sistema a partir de sentencias "si – entonces"	La selección del modelo apropiado de red y el algoritmo de entrenamiento requiere de mucha experiencia
Permite utilizar el conocimiento de un experto	Permite hallar soluciones a partir de un conjunto de datos
El conocimiento es estático	Son capaces de aprender y auto-adaptarse
Existen muchas aplicaciones comerciales	Su aplicación es mayormente académica
Permiten encontrar soluciones sencillas con menor tiempo de diseño	Requieren un enorme esfuerzo computacional

Fuente: (Chahuara, 2005)

3.2.3. Regresión Logística

La regresión logística es un algoritmo de clasificación de Machine Learning que se utiliza para predecir la probabilidad de una variable dependiente categórica. Según explica Amat (2016) la variable dependiente de una regresión es binaria y contiene datos como 0 y 1 que determinan la falla o éxito respectivamente, es decir que el modelo de regresión logística calcula la probabilidad de que Y=1 en función de X.

La regresión logística es una de las formas más populares de ajustar modelos para datos categóricos, especialmente para datos de respuesta binaria en el modelado de datos. Es el miembro más importante (y probablemente el más utilizado) de una clase de modelos llamados modelos lineales generalizados. A diferencia de la regresión lineal, la regresión

logística puede predecir directamente las probabilidades (valores que están restringidos al intervalo (0,1));

Este método preserva las probabilidades marginales de los datos de entrenamiento. Los coeficientes del modelo también proporcionan una pista de la importancia relativa de cada variable de entrada.

La regresión logística tiene muchas aplicaciones en el campo de la medicina, como el puntaje de gravedad de una lesión traumática (TRISS) muy utilizada para predecir la mortalidad de los pacientes lesionados; también se usa este modelo estadístico para predecir la probabilidad de un paciente de padecer alguna enfermedad. Para los distintos tipos de aplicaciones que se le da a la regresión logística es posible usar sus diversas variaciones:

- Regresión Logística Binaria: La respuesta categórica tiene solo 2 respuestas, por ejemplo, si un correo es spam o no.
- Regresión Logística Multinomial: Tiene 3 o más categorías sin ordenar, por ejemplo, que tipo de comida prefieres más (vegetariano, no vegetariano o vegano)
- Regresión Logística Ordinal: Tres o más categorías con ordenamiento, por ejemplo, clasificación de una canción del 1 al 5

En la siguiente tabla (Tabla 2) se describen ventajas y desventajas de la Regresión Logística respecto a otras técnicas de predicción.

Tabla 2.

Ventajas y desventajas de Regresión Logística

Ventajas	Desventajas
• No son necesarios grandes recursos computacionales. • Los resultados son interpretables. • Nos permite entender la importancia de las variables. • Produce probabilidades pronosticadas bien calibradas.	• No es uno de los algoritmos más potentes, debido a que puede ser superado por otros algoritmos más desarrollados. • No podemos resolver problemas no lineales con regresión logística, ya que su superficie de decisión es lineal. • Tiene alta dependencia de una presentación adecuada de sus datos. • Es un algoritmo conocido por su vulnerabilidad al sobreajuste.

Fuente: (Amat, 2016)

3.2.4. Naive Bayes

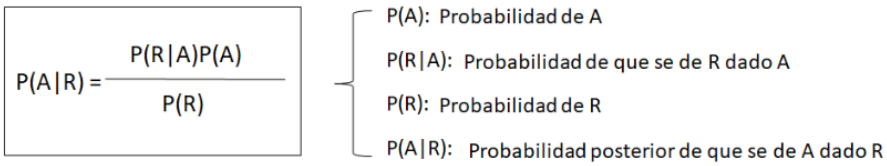
El clasificador naive bayes es uno de los clasificadores más utilizados por su simplicidad y rapidez, es una técnica de clasificación y predicción supervisada que construye modelos que predicen la probabilidad de posibles resultados basada en la técnica de clasificación estadística llamada “teorema de bayes”.

Según menciona Román (2019), Naive Bayes asume la independencia entre sí de las variables; es decir que, si existe alguna característica en el conjunto de datos, esta no tiene

relación alguna con la existencia de otra característica, la razón de que este modelo sea fácil de construir se debe a que nos proporciona una manera de calcular la probabilidad de que exista cierto evento dada la existencia de eventos anteriores como se observa en la figura 12.

Figura 12.

Cálculo de probabilidad Naive Bayes



Fuente: (Roman, 2019)

Naive bayes consta de unos pasos definidos, el primero es crear una tabla de frecuencias en base al conjunto de datos existente, luego se evalúa la probabilidad de que suceda cada evento para crear una tabla de probabilidades, después de esto se usa el cálculo de Naive Bayes para calcular la probabilidad posterior de cada clase y por último el resultado de la predicción será dado por la clase que haya obtenido la mayor probabilidad (Roman, 2019).

Según Pedamkar (2019), Naive Bayes es para la predicción en tiempo real ya que muy rápido a la hora de entrenar, también es usado en predicción de clases múltiples, clasificación de textos, análisis de opinión y filtrado de spam gracias a su regla de regla de independencia

Las ventajas y desventajas de implementar un modelo con el método Naive Bayes se muestra en la tabla 3.

Tabla 3.
Ventajas y desventajas de Naive Bayes

Ventajas	Desventajas
• Se obtiene buenos resultados en gran parte de los casos. • Es fácil de implementar. • Robusto para ejemplos ruidosos y atributos irrelevantes. • Maneja valores faltantes simplemente ignorando la instancia durante los cálculos. • El entrenamiento y predicción del modelo es veloz considerando la gran cantidad de información que se maneja.	• La función de independencia puede no cumplirse para algunos atributos. Se deben usar otras técnicas tales como redes de creencias bayesianas. • Debido a que se considera que las variables son independientes entre sí puede que no se refleje los datos tal como son en realidad

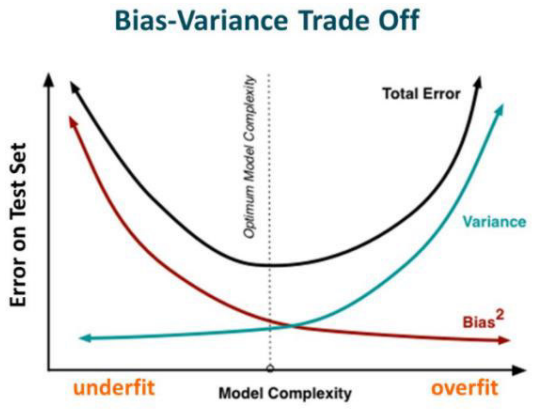
Fuente: (Yi-lai y otros, 2009)

3.2.5. Métodos de Ensamble de Modelos

Según Orellana (2018) los métodos ensambladores son una combinación de distintos modelos que al combinarlos nos dan un modelo ensamblado con mayor precisión y estabilidad ya que al combinar distintas técnicas que funcionan de manera diferentes los errores se reducen. Al igual que los demás modelos predictivos, un árbol de decisión ensamblado también tiene errores de sesgo y varianza. Depende mucho de la complejidad del problema para mantener un balance entre estos dos tipos de errores como vemos en la figura 13. A esto se le conoce como “trade-off” (equilibrio) entre errores de sesgo y varianza. El uso de ensambladores es una forma de aplicar este “trade-off”.

Figura 13.

Ejemplificación de Trade-off



Fuente: (Orellana, 2018)

Hay varias formas de construir estos modelos ensambladores, las principales son:

- **Bagging**

Según explica Heras (2019) el bagging es una combinación de modelos de machine learning, si bien se puede utilizar cualquier modelo, los árboles aleatorios son los más populares debido a su rapidez para construirse. La manera de conseguir que los errores se reduzcan es que cada modelo se entrene con subconjuntos del conjunto de entrenamiento.

Los resultados se combinan para problemas de clasificación y el resultado será la clase más votada como lo muestra la figura 14,

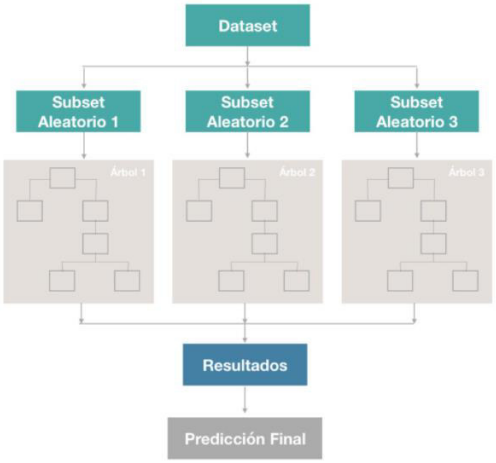
- **Boosting**

El Boosting según explica Gonzales (2020) es una técnica que se construye de manera secuencial, al principio se entrena todo el conjunto de datos con un modelo, y luego los modelos posteriores se construyen ajustando los valores de error residual del modelo inicial. De esta manera se da mayor peso a los registros que el modelo anterior clasificó

erróneamente. Luego de crear la secuencia de modelos (Figura 15), las predicciones son ordenadas por sus precisiones y estos resultados se combinan para obtener el resultado final. Los modelos más comunes de Boosting son XGBoost y ADABoost.

Figura 14.

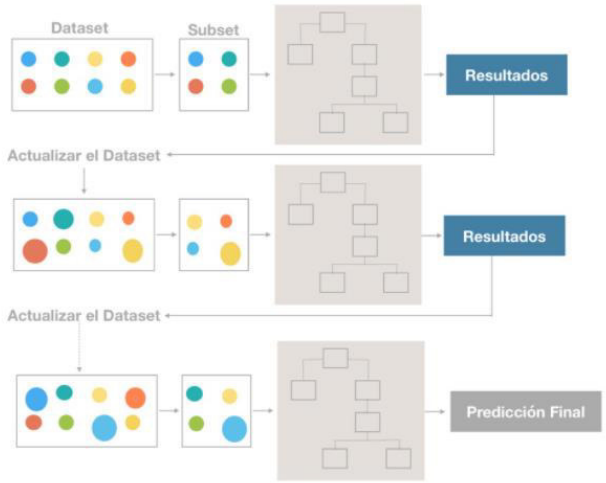
Arquitectura de un modelo Bagging



Fuente: (Gonzales, Métodos de Ensamble de Modelos, 2020)

Figura 15.

Arquitectura de un modelo Boosting.



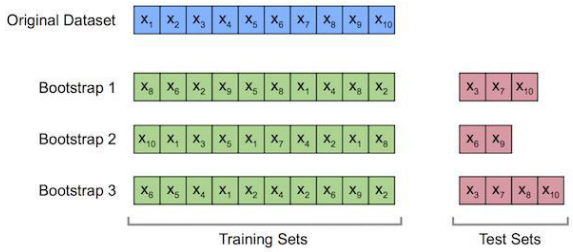
Fuente: (Gonzales, 2020)

Random Forest

El modelo Random Forest se forma en base a un conjunto de árboles de decisión, cada árbol es entrenado con un subconjunto de datos aleatorios extraídos de los datos originales utilizando bootstrapping (Figura 16). Cada árbol es entrenado con datos distintos (Figura 17), y las observaciones se distribuyen mediante bifurcaciones hasta alcanzar un nodo terminal. (Amat, 2022) . Luego de entrenar el modelo e ingresar información para clasificar, cada árbol de decisión vota por una clase y la clase que tengo el mayor número de votos (arboles) será el resultado

Figura 16.

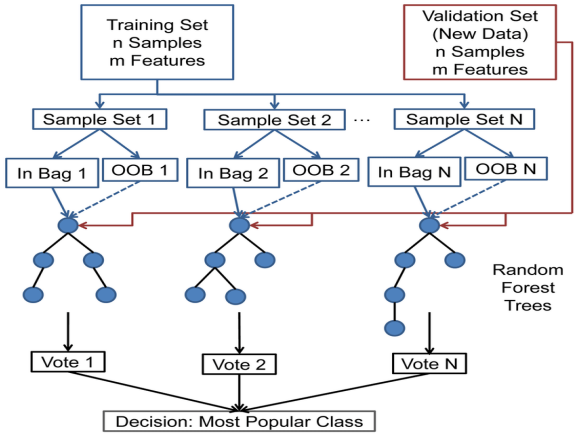
Entrenamiento mediante Bootstrapping



Fuente: (Orellana, 2018)

Figura 17.

Arquitectura de un modelo Random Forest



Fuente: (Orellana, 2018)

Las ventajas y desventajas de implementar un modelo Random Forest se describen en la tabla 4.

Tabla 4.

Ventajas y desventajas de Random Forest

Ventajas	Desventajas
• Se requiere muy poca preparación de los datos. • Gracias al método de reducción de dimensionalidad se puede manejar una gran cantidad de variables. • El modelo entrega como salida la importancia de cada variable. • Buen manejo de estimación de valores faltantes. • Se puede utilizar Random Forest como método no supervisado y para la eliminación de outsiders.	• Es difícil interpretar los resultados • Los resultados de las predicciones no son de tipo continuo, por ello es bueno para resolver problemas de clasificación más no de regresión. • Al ser un modelo compuesto por varios árboles, existe poco control por parte de los estadísticos

Fuente: (Orellana, 2018)

Orellana (2018) explica la importancia de la hiperparametrización en un modelo de Random Forest, esta es necesaria para ajustar bien las variables candidatas para cada ramificación. Dentro de los parámetros importantes se encuentran:

- ntree: Indica el n de árboles a utilizar en el modelo, el uso de una gran cantidad de árboles puede inducir al modelo en error
- mtry: Es la cantidad de variables que serán seleccionadas aleatoriamente para cada ramificación.

- sampsize: es la cantidad de registros con los cuales se entrenará el modelo y tiene un valor por defecto de 63.25%.
- maxnodes: Indica la máxima cantidad de nodos terminales que puede tener el modelo.

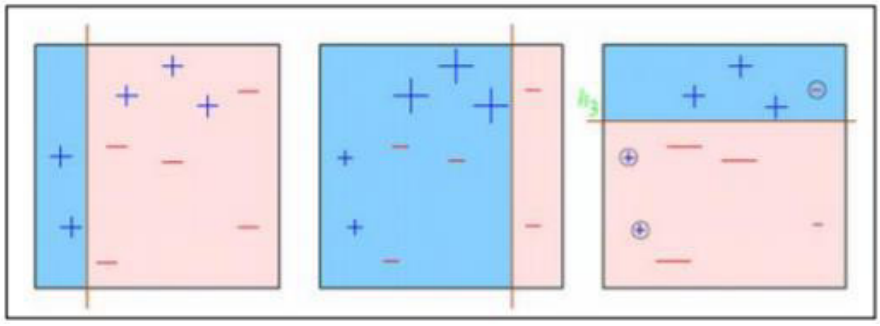
Gradient Boosting

Como ya lo definimos anteriormente, Gradient Boosting es uno de los métodos de Boosting más utilizados y uno de los métodos más potentes de Machine Learning. El significado de este método parte de dos palabras:

- Boosting (Aumentar): Esto implica que de manera sucesiva se va ensamblando modelos simples para obtener un modelo final con mejores resultados. La figura 18 en un ejemplo visual del Boosting nos muestra que a los resultados predichos correctamente en un modelo t se les asigna un peso menor y a los incorrectos un peso mayor, para que así en el siguiente modelo (t + 1) se trabaje con estos resultados y así sucesivamente hasta llegar a un modelo final con mayor porcentaje de efectividad como se muestra en la figura 19
- Gradiente (Gradiente): Se utiliza la función de Gradiente Descendiente para encontrar el mínimo de una función, iterativamente hasta encontrar el punto C = C', como se observa en la figura 20., en ese punto la gradiente es cero y se procede a finalizar el proceso iterativo, ya que cumple con la condición.

Figura 18.

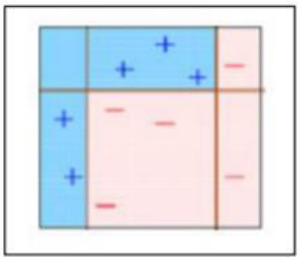
Ejemplo Visual de Boosting



Fuente: (Lopez, 2017)

Figura 19.

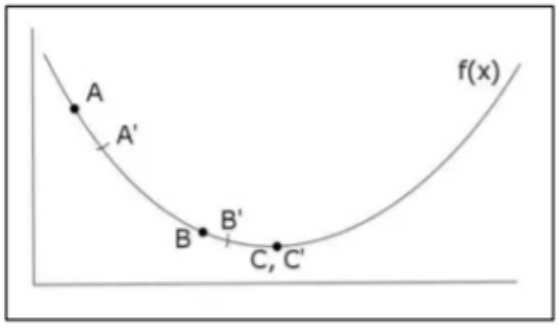
Modelo Final de Boosting



Fuente: (Lopez, 2017)

Figura 20.

Gradiente Descendiente



Fuente: (Lopez, 2017)

Gradient Boosting tiene 3 elementos principales, el primero es “Loss function”, la definición de esta función depende del problema que se quiera resolver, se pueden usar

errores cuadráticos o logistic. El otro componente principal es los “Weak Learners” que son los árboles de decisión simple que componen los modelos, a estos árboles se les restringe características como los nodos, las ramas para que se mantenga la simplicidad del modelo. Por último, es importante mencionar los “Additive Model” son los modelos añadidos que aparte de los árboles de decisión simple se van añadiendo cada iteración.

3.2.6. Árboles de Decisiones

El árbol de decisión es uno de los métodos supervisados más utilizados, su manera de representar el conocimiento es sencilla y fácil de interpretar. Su dominio de aplicación es amplio, se pueden utilizar para diagnóstico médico, control de calidad, etc.

Un árbol de decisión es un conjunto de condiciones jerárquicas, es así como la decisión final se puede explicar siguiendo las condiciones desde la raíz hasta alguna de sus hojas.

Los principales elementos de un árbol de decisión según Herrera (2022) son los nodos raíz que representa el nivel superior, las ramas representan las opciones disponibles para la toma de una decisión y el nodo hoja que representa el resultado para cada acción que fue tomada

Ferrero (2020) explica que al crear un árbol de decisión es de suma importancia la selección correcta de la variable que va a dividir el subconjunto de datos en subconjuntos más pequeños y existen las técnicas más comunes para realiza esta división se encuentran el índice de Gini, la Entropía y el error de clasificación.

Las ventajas y desventajas de implementar un Árbol de Decisión se describen en la tabla 5.

Tabla 5.

Ventajas y desventajas del Árbol de Decisión

Ventajas	Desventajas
• Se puede usar para variables predictoras continuas o categóricas. • La interpretación del resultado es directa e intuitiva. • Trabaja muy bien con datos ruidosos. • Es computacionalmente rápido. • Realiza automáticamente selección de variables.	• La selección de variables es segada hacia las variables con valores diferentes. • Se dificulta elegir el árbol óptimo. • Requieren un gran número de datos para asegurarse que la cantidad de las observaciones de los nodos sea significativa.

Fuente: *(Ferrero & López, 2021)*

ID3 (Inductive Decision Trees)

ID3 es un tipo de árbol de decisión que se caracteriza por el bueno manejo de múltiples casos repetidos en el conjunto de entrenamiento. El ID3 maneja una post-poda, para aquellos nodos de los cuales nacen muchas ramas y terminan en la misma clase, entonces se procede a reemplazar dicho nodo por un nodo hijo. Este algoritmo elije el mejor atributo mediante la entropía, eligiendo aquel que proporcione una mejor ganancia de información.

Método Árbol ID3 Fuzzificado

Suca et al (2016) define que “el algoritmo ID3 fuzzificado, es una extensión del algoritmo clásico ID3, como un método para construir un árbol de decisión. En la construcción recursiva del árbol, se adopta la estrategia ‘divide y vencerás’.” (p.17)

Los pasos propuestos por Begenova y Avdeenko (2018) para la construcción de un algoritmo ID3 son los siguientes:

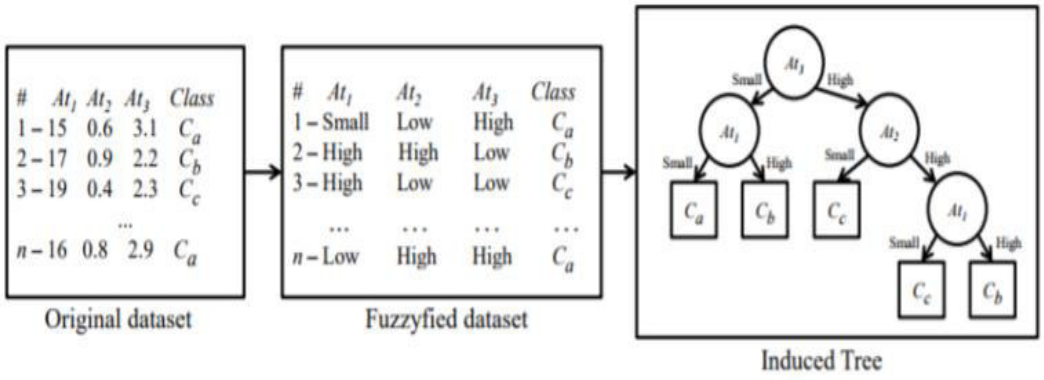
- Defina la base de datos difusa, es decir, la granulación difusa para los dominios de las características continuas.
- Reemplace los atributos continuos del conjunto de entrenamiento usando las etiquetas lingüísticas de los conjuntos difusos con mayor compatibilidad con los valores de entrada.
- Calcule la entropía y la ganancia de información de cada característica para dividir el conjunto de entrenamiento y defina los nodos de prueba del árbol hasta que se utilicen todas las características o se clasifiquen todos los ejemplos de capacitación.

En la figura 21 el primer bloque se ilustra un conjunto de datos con n ejemplos, tres atributos (At1, At2, At3) y un atributo de clase. La versión difusa de este conjunto de datos se presenta en el segundo bloque. Una vez fuzzificado se usa el conjunto de ejemplos para inducir el árbol de decisión final, ilustrado en el último bloque de la figura

Las fórmulas de entropía y ganancia de información siguen siendo las mismas para la versión clásica de un árbol de decisión.

Figura 21.

Pasos para construir un árbol de decisión difusa



Fuente: (Begenova & Avdeenko, 2018)

3.3. Casos de Éxito

3.3.1. Diseño de la Metodología de un sistema experto difuso para el diagnóstico y control de la obesidad

Umoh et al. (2015), Malasia en su artículo publicado en la revista "Computer Engineering and Intelligent System" diseñaron un sistema utilizando lógica difusa para diagnosticar y controlar el aumento de casos de obesidad.

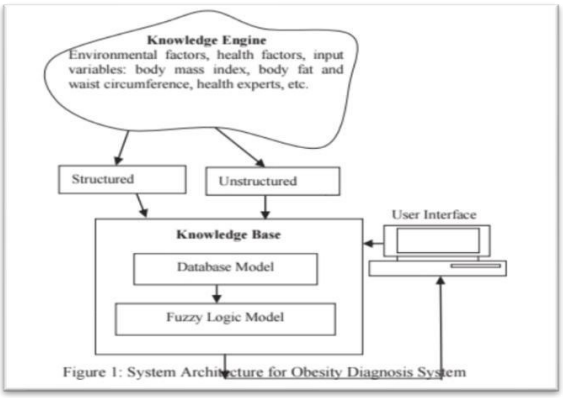
La arquitectura del sistema mencionado presentó componentes como la base de conocimiento, el motor de inferencia y la interfaz de usuario. Los datos se ingresaron a través del motor de conocimiento para luego ser procesados en la base de conocimiento utilizando la lógica difusa y finalmente el diagnóstico se mostraba mediante una interfaz de usuario. (figura 22).

El siguiente paso fue establecer el modelo de la base de datos que tuvo como principal objetivo guardar los datos más importantes del paciente, del médico y del reporte médico.

Las unidades de construcción principales para el sistema de diagnóstico fueron la unidad de fusificación, la unidad de inferencia difusa, la base de conocimientos y la unidad de defusificación.

Figura 22.

Arquitectura del Sistema



Fuente: (Umoh & Isong, 2015)

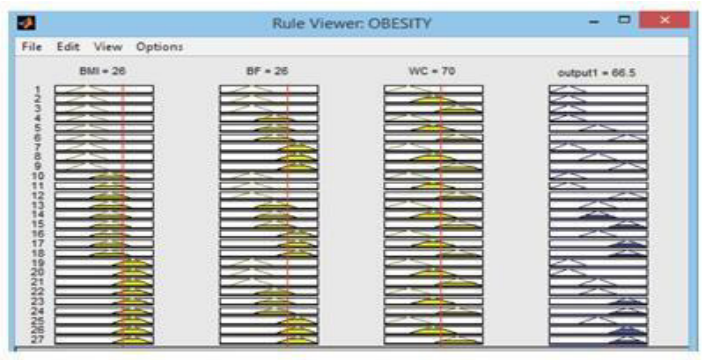
La primera unidad convertía la entrada nítida en el valor de membresía a través de las funciones triangulares, las entradas consideradas por el sistema fueron el índice de masa corporal, la grasa corporal y la circunferencia de la cintura. El IMC contó con 3 atributos que son: bajo, moderado y alto, la grasa corporal tuvo los atributos bajo, normal y alto y por último la circunferencia de la cintura se definió como pequeña, mediana y grande. En la unidad de inferencia difusa se utiliza las reglas de la base para comparar la entra obtenida con conocimiento factico de la base de datos. La base de conocimiento contuvo las 27 reglas definidas y los datos utilizados por el motor de inferencia y la unidad de defusificación se encargó de convertir la membresía en la salida nítida.

El sistema se implementó en Matlab y se utilizó su caja de herramientas de lógica difusa para desarrollar una simulación por computadora mostrando al usuario la interfaz y la inferencia difusa para ayudar a la decisión experimental para la mejor acción de control.

Finalmente, se procedió a realizar las pruebas con los valores de un paciente (ver figura 23) y como resultado el paciente tuvo un 66.5% de sobrepeso en el nivel de riesgo de la obesidad por ende el paciente es aconsejado por el médico para revisar su dieta y participar en el ejercicio físico activo.

Figura 23.

Construcción Gráfica



Fuente: (Umoh & Isong, 2015)

En conclusión, los autores demostraron que el presente trabajo podría ser utilizado para diagnosticar obesidad y evitar sus consecuencias con antelación tanto en hombres como en mujeres.

3.3.2. Modelo difuso para la predicción de casos de obesidad empleando el árbol GFID3 generalizado

Suca et al. (2016), Perú en su tesis de pregrado presentaron su trabajo titulado "Modelo difuso para la predicción de casos de obesidad empleando el árbol GFID3 generalizado", el modelo desarrollado con reglas buscó clasificar a hombres y mujeres de 6-17 años que sufran de obesidad.

Para realizar este modelo se consideró 4 etapas: la etapa de preprocesamiento de datos, la fusificación, la etapa de generación de reglas y la clasificación de resultados.

En la etapa inicial del preprocesamiento de datos se dispuso de registros médicos que contenían 14 variables, entre ellas se encontraban información de edad, sexo, IMC, medida de pliegues, etc. Junto con los expertos se definió los valores CRISP y los factores críticos utilizando la información ya obtenida anteriormente. Al existir una diferencia de rangos para el cálculo del IMC entre hombres y mujeres la clasificación se realizará de manera separada.

Luego del procesamiento de datos, la siguiente etapa es la fusificación donde es necesario obtener los rangos y clases para cada atributo. Luego que el experto tome los rangos en base a lo que indica la OMS, estos serán inputs del sistema Mandani. La función de pertenencia de forma trapezoidal fue la que más se ajustó al contexto.

Una vez que tenemos los datos fuzzificados, se procede a generar las reglas y para este caso se utilizó el algoritmo ID3 fuzzificado gracias a la velocidad con la que proporciona las reglas y su buen manejo de datos difusos. Después de haber definido el modelo difuso correctamente se entrenó el árbol ID3con 1190 registros y el resto de los registros fue utilizado para la prueba.

Finalmente, se realizaron las pruebas con 2,935 niños y 3,021 en niñas y la precisión que se obtuvo se muestra en porcentajes en las figuras 24 y 25 respectivamente.

Figura 24.

Precisión para niños

Niños	Precisión
Correctamente Clasificados	83.65
Incorrectamente Clasificados	16.35

Fuente: (Suca C. y otros, 2016)

Figura 25.

Precisión en Niñas

Niñas	Precisión
Correctamente Clasificados	76.13
Incorrectamente Clasificados	23.87

Fuente: (Suca C. y otros, 2016)

En conclusión, este trabajo demostró que el algoritmo GFID3 generalizado es más preciso respecto a otras variantes de ID3, con un porcentaje de precisión de 83.65% para los niños y 76.13% para las niñas.

3.3.3. Comparación De Algoritmos De Clasificación Para La Predicción De Casos De Obesidad Infantil

Suca C. , Córdova, Condori, Cayra, y Sulla (2016) Perú, presentaron en la UNSA su investigación "Comparación De Algoritmos De Clasificación Para La Predicción De Casos De Obesidad Infantil" con el objetivo de evaluar distintas técnicas de Machine Learning y comparar los resultados al clasificar si sufren de obesidad a niños de 6 a 17 años.

Este trabajo realizó una comparación entre algoritmos de clasificación (arboles de decisión, J48, SVM, C4.5, Back Propagation y Naive Bayes) para determinar cuál de estos es más adecuado para la predicción de obesidad infantil.

Se utilizó la herramienta Weka para el análisis del conjunto de datos que estuvo compuesto por 5962 registros de niños entre los 6 y 17 años, estos registros fueron provistos por instituciones educativas de Brasil. Se seleccionó 12 parámetros para emplearlos en las predicciones, entre los principales esta la edad, el sexo, IMC, Atc, CP, PVC, Pliegue Tricipital y masa magra.

Weka es una herramienta que ya cuenta con algoritmos implementados dentro de su interfaz, por ello el proceso de validación para cada técnica se realizó usando esta herramienta y cargado la información en formato CSV.

Finalmente, luego de realizar las pruebas y obtener los resultados, se procedió a evaluar cual es el clasificador óptimo para la predicción de obesidad infantil. En la Tabla 7 se mostrarán el porcentaje de instancias correcta e incorrectamente clasificadas de cada uno de los métodos evaluados.

Tabla 6.

Comparación de resultados

	J48	Random Forest	SVM	NB	Red Neuronal
Correctas	97.23%	96.83%	95.72%	82.72%	97.08%
Incorrectas	2.77%	3.17%	4.27%	17.28%	2.925%

Fuente: (Suca C. y otros, 2016)

Los autores concluyeron en que las técnicas SVM, Redes Neuronales y los árboles de decisión obtuvieron una buena precisión, sensibilidad y especificidad, y también resaltaron la importación de realizar un buen tratamiento de datos antes de pasar la información por el modelo.

3.3.4. Modelo estadístico para la prevención precoz de desarrollo de sobrepeso/obesidad en población infantil.

En el año 2017, (Morlan y otros) en España llevaron a cabo la publicación de su artículo titulado "Modelo estadístico para la prevención precoz de desarrollo de sobrepeso/obesidad en población infantil". Este estudio se presentó en el Boletín de la Sociedad de Pediatría de Aragón, La Rioja y Soria. El objetivo principal del artículo fue anticipar la detección del riesgo de obesidad infantil en niños. Para lograr esto, se realizaron

investigaciones en un grupo de 242 niños. Utilizando 14 medidas antropométricas como base, se construyó un modelo de regresión logística con el propósito de identificar tempranamente la posibilidad de que un niño desarrolle obesidad.

Se realizó el estudio sobre una población sana de 122 niños y 120 niñas; de los cuales se recolectó 14 parámetros de medidas antropométricas (Tabla 8), estas medidas fueron recolectados desde el nacimiento al año de edad con una frecuencia trimestral, luego de manera semestral hasta los 2 años de edad y posteriormente anualmente hasta los 18 años.

Una vez seleccionados los parámetros fue necesario definir el sobrepeso, para el cual se usó el IMC y el perímetro abdominal, cuando ambos parámetros fueran mayores a la media más una desviación estándar, se considerarían como sobrepeso.

El modelo estadístico utilizado fue una regresión logística y fue implementado en MATLAB, el corte de edad para evaluar si un niño puede presentar o no obesidad fue la edad de 3 años.

Tabla 7.
Parámetros Antropométricos

Nombre del parámetro	
1. Peso(kg)	2. Talla(cm)
3. Talla sentada(cm)	4. Pliegue tricipital(mm)
5. Pliegue subescapular(mm)	6. Longitud
7. biacromial(cm)	8. Diámetro bicrestal(cm)
9. Perímetro del brazo(cm)	10. Perímetro cefálico(cm)
11. Diámetro biparietal(cm)	12. Diámetro frotooccipital(cm)
13. Perímetro torácico(cm)	14. Perímetro abdominal(cm)
15. IMC	

Fuente: (Morlan y otros, 2017)

Para reducir la cantidad de variables existentes se utilizó la técnica stepwise, esta técnica mide la significancia estadística de cada variable y decide si incluirla o no en el

modelo. Luego de haber definido las variables, se realizó la detección de outliers o datos atípicos.

Una vez se tuvo el conjunto de datos listo, se utilizó la técnica de validación cruzada para entrenar el modelo. Al existir registros de niños y niñas, se optó por crear modelos por separado, el modelo de regresión logística de las niñas entregó una sensibilidad de 96% y una especificidad de 92%, y el modelo para los varones tuvo como resultado una sensibilidad de 96% y especificidad de 94%; ambos tuvieron una calidad discriminatoria excelente bajo la curva ROC.

3.3.5. Aplicación del modelo Neuro-Difuso ANFIS para la clasificación de la obesidad en niños y adolescentes.

Sulla et al. (2018), Perú presentaron su artículo titulado "Aplicación del modelo Neuro-Difuso ANFIS para la clasificación de la obesidad en niños y adolescentes" en la conferencia LACCEI, esta investigación clasificó la obesidad en niños de 6 a 17 años usando lógica difusa y redes neuronales, para ello se utilizó el modelo ANFIS implementado en Matlab.

En la presente investigación se utilizaron como parámetros antropométricos el peso, la estatura e IMC para predecir la obesidad, en la niñez estos valores cambian constantemente debido al desarrollo en el cual se encuentran. El modelo se compuso en cuatro etapas, estas fueron el análisis de datos, el procesamiento de la información, la implementación del modelo neuro-difuso ANFIS y finalmente la evaluación de resultados.

Durante la etapa de comprensión de datos se analizó la información que se utilizó para el modelo, la base de datos tuvo 2938 registros y los parámetros fueron edad, peso, estatura e IMC, el resultado del modelo se guardó en la variable "CLASE"; el 75% de los registros se utilizó para entrenamiento y el 25% para prueba.

En la etapa de procesamiento de datos, para cada atributo de la etapa anterior se consideró tres conjuntos difusos (bajo, normal y alto), y para determinar si un valor pertenece a un conjunto difuso se tomó en cuenta los patrones de crecimiento de la OMS. Una vez definidos los percentiles P3, P15, P85 y P97 se procedió a definir las funciones trapezoidales para cada atributo.

En la siguiente etapa de clasificación se definió el sistema neuro-difuso, este consistió en un modelo difuso tradicional que cuenta con la capa de fuzzificacion, capa de reglas y capa de defuzzificacion (figura 26) con la diferencia de que cualquiera de las 3 capas puede representarse como una capa neuronal como nos muestra la figura 27.

Se definieron 27 reglas del tipo SI- ENTONCES, las salidas de estas neuronas estuvieron conectadas a la capa de defusificación, esta capa se encargó de evaluar las reglas y devolver un resultado entre 0 y 100 donde 0 representa 0% probabilidad de que sufra de obesidad y 100 % de probabilidad de padecer obesidad. Al terminar de diseñar el modelo se usó el algoritmo Retro-Propagation para el entrenamiento.

Finalmente, los autores obtuvieron un modelo neuro difuso que tuvo una exactitud de 96.96% y al compararlo con otras técnicas como redes neuronales, SVM y Naves fue superior en exactitud como muestra la figura 28.

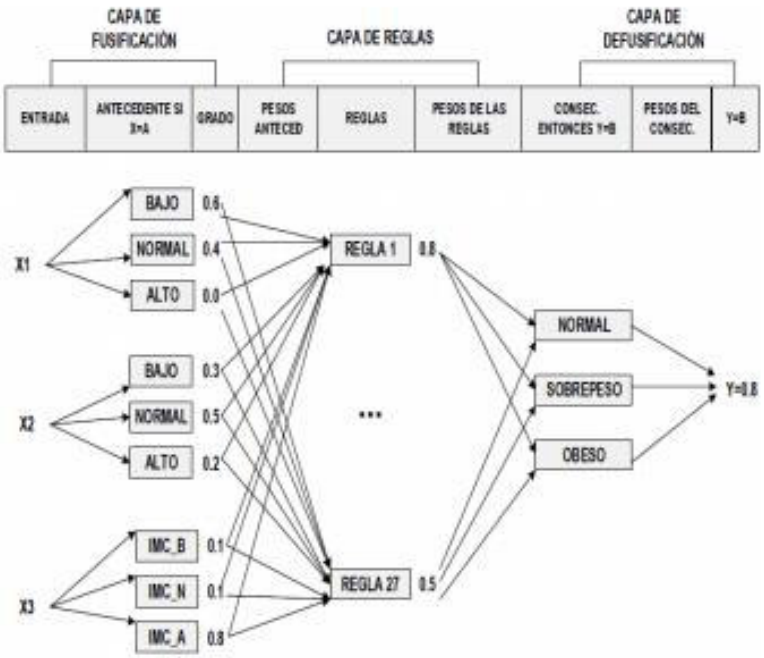
Figura 26.

Diagrama de bloques del Sistema Neuro-difuso

Fuente: (Sulla y otros, 2018)

Figura 27.

Estructura del Sistema Neuro-Difuso para obesidad



Fuente: (Sulla y otros, 2018)

Figura 28.

Exactitud para la clasificación de la Obesidad

Técnica	Exactitud
Red neuronal MLP	96.80%
SVM	96.28%
Naive Bayes	73.42%
Red Neuro-difusa	96.96%

Fuente: (Sulla y otros, 2018)

3.3.6. Predicting childhood obesity using electronic health records and publicly available data

Hammond et al. (2019) , Israel en su artículo publicado en la revista PLOS ONE Vol. 14° predijeron la obesidad infantil en niños y niñas, utilizando registros electrónicos con la

información clínica de los 2 primeros años de vida para predecir la obesidad a los cinco años utilizando modelos de clasificación binaria y regresión logística.

La obesidad infantil ha ido en aumento desde la década de 1970, por ello es importante definir una estrategia para predecir la obesidad a temprana edad, ya que un niño con obesidad puede dar como resultado un adulto con complicaciones médicas como problemas cardiacos o diabetes. Con esta premisa, la investigación utilizo variables de los primeros 24 meses de vida como se observa en la figura 29, entre las variables se encontraron la historia clínica de la madre antes del embarazo, durante el embarazo e información de peso y talla del niño al nacer y durante los siguientes meses, así como diagnóstico de algunas enfermedades (asma, alergias, infecciones, etc.)

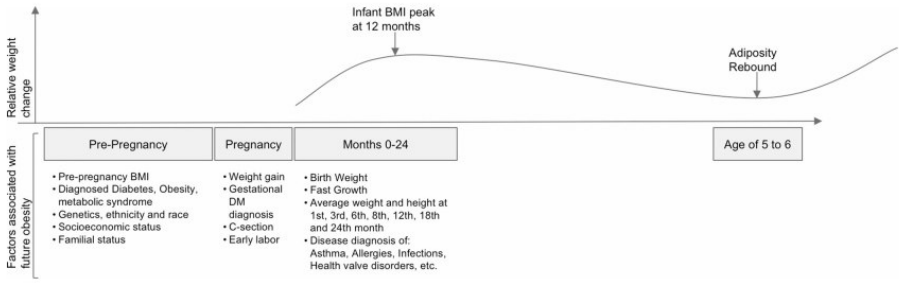
La base de registros clínicos contaba con un total de 52 945 registros a los cuales se tuvo que realizar varios filtros; el primero fue que tenía que existir información del IMC entre los 4.5 y 5.5 años, luego se tuvo que validar que el IMC sea válido (entre 10 y 40 kg/m^2), el tercer criterio fue que cada niño debía tener al menos una visita al médico los primeros 24 meses y, por último, tenía que contar con información de la madre. Al combinar los 3 criterios, el estudio se redujo a 3 449 registros.

Luego de tener la información de los registros médicos listos se procedió a normalizar las variables continúas restando la media de cada valor y dividiendo entre la desviación estándar, y se eligió modelos en 2 categorías para realizar las pruebas:

- Regresión Logística, Random Forest, y Gradient Boosting para predecir el resultado binario de obesidad (obeso/no obeso)
- LASSO, Random Forest y Gradient Boosting para pronosticar a los niños como si tuvieran obesidad cuando el valor es mayor al umbral de obesidad.

Figura 29.

Factores prenatales e infancia asociados con la obesidad

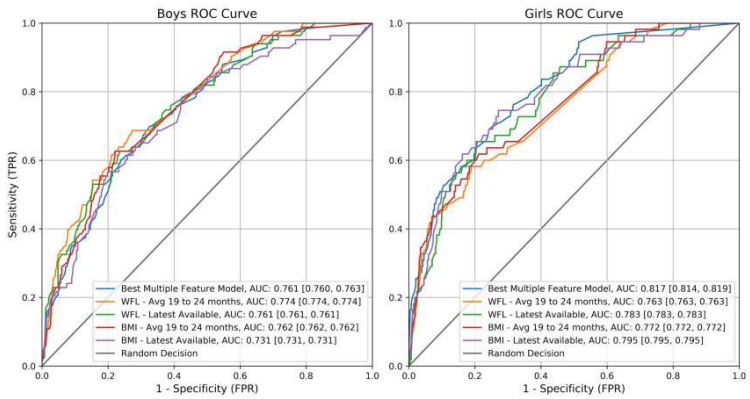


Fuente: (Hammond y otros, 2019)

En conclusión, los autores determinaron que las variables que predecían mejor la obesidad fueron el peso, la talla, el IMC entre 19 y 24 meses y la última medida de IMC registrada antes de los dos años; además solo la variable de diabetes gestacional dentro de todas las variables de diagnósticos tuvo una asociación significativa ($p < 0.001$) con obesidad a los 5 años. Por último, el modelo con mejor rendimiento para las niñas fue regresión LASSO con LASSO para la selección de variables con 81.7 % AUC y para los niños también fue LASSO el mejor modelo, sin selección de variables se obtuvo un 76.1 % de AUC (figura 30).

Figura 30.

Curvas ROC para el modelo de mejor rendimiento en comparación con las predicciones de características individuales



Fuente: (Hammond y otros, 2019)

3.3.7. Sistema para la predicción de obesidad en la adolescencia utilizando técnicas de minería de datos

Ticona (2018) Perú presentó su tesis de pregrado titulada “Sistema para la Predicción de Obesidad en la Adolescencia utilizando Técnicas de Minería De Datos”. Esta tesis tuvo como finalidad implementar un software que utilice un árbol de decisión que prediga la obesidad en base a la información ingresada por el usuario, se realizó un análisis exhaustivo para seleccionar el mejor algoritmo para reducir los patrones y tendencias que existen entre los datos.

Los datos que se procesaron en este trabajo fueron recolectados de distintas escuelas de Arequipa. La edad de los estudiantes osciló entre los 5 a 17 años y las variables seleccionadas fueron 17, luego de la limpieza de datos se eligieron 13 atributos (figura 31) de un total de 660 registros. Estos datos fueron cargados a la herramienta weka.

Figura 31.

Parámetros de modelo

Número de categoría	Parámetros
1	Edad
2	Sexo
3	¿Hace Deporte?
4	¿Fuma?
5	¿Desayuna todos los días?
6	Peso
7	Estatura parado
8	Estatura sentado
9	Circunferencia abdominal
10	Pico de Velocidad de Crecimiento (PVC)
11	Porcentaje grasa (%)
12	Masa grasa
13	Masa libre de grasa (Masa magra)

Fuente: (Ticona, 2018)

Luego de que la información se cargó a la herramienta, se seleccionó 4 métricas (precisión, sensibilidad, especificidad y curva ROC) para evaluar los distintos y seleccionar el

óptimo, se evaluó las técnicas J48, BayesNet, Naive Bayes, ForestPA y MPL. El modelo con mejor precisión fue la técnica J48 con un 92.8% de precisión, este modelo generó un árbol de decisión donde los parámetros con mayor importancia fueron “grasa magra”, “% de grasa”, “peso” y “género”.

Luego de seleccionar el modelo más optimo, se procedió a crear la interfaz de usuario con el software libre XAMPP que consiste en un sistema de gestión de bases de datos MySQL, el servidor web Apache y los intérpretes para lenguajes de script PHP y Perl.

En conclusión, los autores determinaron que el modelo con mayor precisión sobre los demás modelos fue el árbol de decisión J48 que obtuvo una precisión de 94.39%.

3.3.8. Prediction of Childhood Obesity from Nationwide Health Records

Rossman et al. (2021) en su artículo “Prediction of Childhood Obesity from Nationwide Health Records” publicado en la revista “The Journal of Pediatrics” diseñaron un modelo de predicción dirigido a identificar a los niños con alto riesgo de obesidad, prediciendo la obesidad a los 5-6 años de edad según los datos de los primeros 2 años de vida de 136,196 niños.

Los registros médicos con los que se trabajó fueron recolectados de un centro de atención integrada de Israel donde se atienden más de la mitad de la población del país. El conjunto de datos incluyó datos demográficos, medidas de peso y altura, diagnósticos clínicos y hospitalarios, medicamentos dispensados y pruebas de laboratorio de 2002 a 2018. Al principio fueron un total de 882 987 que al pasar por una serie de filtros terminan seleccionándose 108 416 registros para entrenar el modelo, y 27 780 para validar.

Luego de realizar la selección de registros, se procedió a crear nuevas variables a partir de las variables existentes que contenían información acerca de:

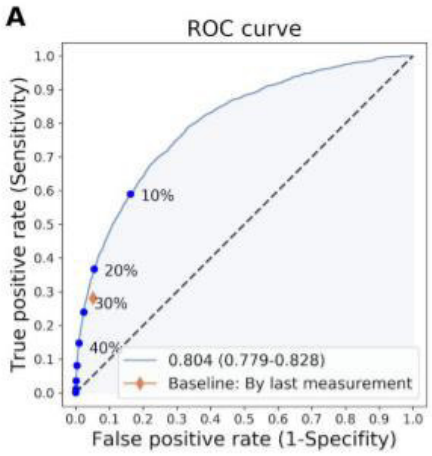
- Características del niño (sexo, ¿nació en Israel?, fecha de nacimiento, semana de nacimiento, información socioeconómica, tipo de embarazo)
- Características al nacer (peso, talla, tipo de parto)
- Medidas antropométricas
- Pruebas de laboratorio
- Diagnósticos clínicos
- Medicamentos administrados
- Características maternales
- Test de laboratorio durante el embarazo, etc.

El método propuesto para este trabajo de investigación fue Gradient Boosting, ya que, los árboles de decisión permiten capturar interacciones de características múltiples y no lineales, lo que puede ser importante para obtener un modelo de predicción preciso.

Finalmente, los autores obtuvieron un modelo cuya área bajo la curva tuvo un valor de 0,804 como se observa en la figura 32, el modelo fue implementado con la librería LightGMB de Python y los parámetros fueron ajustados mediante validación cruzada.

Figura 32.

Curva ROC del modelo



Fuente: (Rossman y otros, 2021)

3.4. Aplicaciones informáticas y herramientas de predicción

3.4.1. Aplicativos comerciales

A continuación, se describen distintas aplicaciones comerciales que tienen como fin controlar y predecir la obesidad infantil realizando un seguimiento al peso y talla del niño, así como su actividad física diaria.

Fitbit ACE2. Es una pulsera de actividad creada por la marca FITBIT pensada exclusivamente en niños de más de 6 años. Fitbit (s.f.) destaca que la pulsera incentiva a los niños mediante recordatorios divertidos a no estar demasiado tiempo sentados

Ante el gran problema que significa la obesidad infantil, la OMS recomienda que los niños realicen 60 minutos de actividad física diaria, por ellos Fitbit Ace 2 incluye objetivos de actividad de 1 hora. Otra importante característica de la pulsera es el control del sueño del niño mediante alarmas silenciosas, además permite hacer un análisis de la calidad del sueño del menor. Fitbit Ace2 tiene una aplicación para celulares Android que sincroniza la información de la pulsera, esta aplicación cuenta con dos tipos de vistas, la primera es la vista para padres donde se configura una cuenta familiar y se generan cuentas para los menores de edad para realizar seguimiento a las actividades del niño, la segunda vista es la vista protegida para niños donde los menores solo podrán ver sus insignias y cambiar los formatos de la pulsera.

Huawei Smart Scale. "Es una balanza inteligente con tecnología de análisis de impedancia bioeléctrica (BIA) y una capa de (ITO) sin metal para una mayor sensibilidad que le permite al usuario ver el peso corporal y la grasa desconectada" (Huawei, s.f.). Esta balanza mide la comprensión completa de tu condición física con 9 análisis de composición corporal en una detección, estos son: peso, % de grasa, IMC, grasa visceral, masa muscular, proteína, % de agua en el cuerpo, masa ósea y tasa metabólica basal.

Este dispositivo cuando con un reloj despertador inteligente para una medición más precisa, admite un reconocimiento automático de hasta 10 usuarios y cuenta con una aplicación para Android y iOS amigable al usuario que proporciona un informe de salud integral y sugerencias de salud personalizadas.

3.4.2. Herramientas utilizadas para desarrollar de sistemas de predicción

En esta sección se describen algunas de las herramientas para análisis predictivo más populares, todas estas con distinta interfaz hacia el usuario, pero con el mismo objetivo el cual es procesar un conjunto de datos para luego aplicar el algoritmo seleccionado y obtener nuestro análisis de predicción.

Weka. Es una herramienta de código abierto que se desarrolló en la Universidad de Wikato en Nueva Zelanda, es una herramienta muy popular y aproximadamente el 11% de usuarios de minería de datos realiza su análisis con Weka. Este sistema está escrito en Java y nos proporciona una interfaz para muchos algoritmos de aprendizaje en la etapa de pre y post procesamiento y para evaluar los resultados.

Morante (2018) indica que Weka contempla algoritmos de preprocesamientos de datos, técnicas de clasificación, clustering, evaluación de atributos y permite visualizar los resultados, entre sus principales ventajas se resalta la gran cantidad de algoritmos que contiene y su constante actualización y su mayor desventaja es la limitación de la memoria debido a que los datos deben cargarse completamente en la memoria principal

Scikit-learn. Es una biblioteca de Python utilizada para el aprendizaje automático. Más específicamente, es un conjunto de herramientas simples y eficientes para la minería de datos y el análisis de datos. Esta librería está construida sobre SciPy (Scientific Python) e incluye las librerías Numpy, pandas, SciPy, matplotlib, IPy y SymPy

Gonzales (2018) describe como principal ventaja de esta solución su accesibilidad y simplicidad: es fácil de usar incluso para principiantes y una excelente opción para tareas de análisis de datos más simples. Por otro lado, scikit-learn no es la mejor opción para el aprendizaje profundo. Scikit-learn es aplicada para implementar algoritmos de aprendizajes supervisados, no supervisados y validación cruzada.

RapidMiner. Es un conjunto de herramientas que proporciona una solución centralizada para crear, entregar y mantener análisis predictivos para empresas. Cuenta con una robusta interfaz gráfica de usuario que proporciona una forma poderosa de operar la plataforma. Según Zuckerman (2018) esta aplicación garantiza un proceso sin obstáculos desde el modelado hasta la implementación y también utiliza tecnología rica útil en proyectos para análisis de datos avanzados. RapidMiner no solo funciona para el análisis predictivo, sino también para la integración de aplicaciones, la integración de datos, el aprendizaje automático y la transformación. Esto permite que el sistema proporcione un enfoque unificado que dará a las empresas una ventaja para aumentar su eficiencia y productividad a través del aprendizaje acelerado y la mejora en la estandarización.

RapidMiner tiene compatibilidad con bases de datos como Oracle, MySQL, Excel, SPSS, Microsoft SQL Server, etc. Otra de sus características es la interfaz amigable que brinda de arrastrar y soltar para diseñar el proceso de análisis.

Matlab. Con frecuencia los equipos que desarrollas aplicaciones de análisis predictivo recurren a MATLAB, mediante esta herramienta es posible llevar a cabo análisis con distintos tipos de datos y se puede implementar en sistemas embebidos y a gran escala.

En Análisis Predictivo (s.f.) nos indican que las principales ventajas de utilizar MATLAB son:

- Admite formato de datos distintos, como sensores, imágenes, videos, telemetría y otros formatos en tiempo real
- Permite la conexión con interfaces de bases de datos ODBC/JDBC, y se pueden explorar los datos mediante Hadoop y Spark
- Se integra en los sistemas, clústeres y nubes empresariales.

4. Capítulo 4: Técnica de Naive Bayes

En el presente capítulo se justifica la técnica elegida apoyándonos en artículos realizados por instituciones que lo han aplicado y han obtenido resultados satisfactorios, luego se elabora un benchmarking de las técnicas predictivas basándonos en criterios, finalmente ejemplificamos un caso de estudio siguiendo las fases de Naive Bayes.

4.1. Justificación

Naive Bayes es una técnica que ha sido utilizada en muchas áreas de dominio en las que ha funcionado exitosamente dando excelentes resultados gracias a su gran versatilidad.

Un claro ejemplo de ello nos muestra Odei (2006), Reino Unido en su tesis de maestría de la Universidad de Bournemouth titulada "An Exploration of Classification prediction techniques in data mining: the insurance domain" en la cual exploró distintas técnicas de minería de datos a fin de identificar la que ofrezca el mejor rendimiento en la clasificación de clientes que respondían a envíos directos en la industria de seguros, como resultado Naive Bayes obtuvo 13.3% de precisión superando el 12.1% que obtuvo Redes Neuronales y el 11.4% de Regresión Logística; con respeto a la velocidad versus tiempo de entrenamiento Naive Bayes supero a los demás métodos.

Chaparro et al. (2015) en su artículo científico "Evaluación del clasificador Naive Bayes como herramienta de diagnóstico en Unidades de Cuidado Intensivo" presentado en la Revista de Tecnología - Colombia buscó resolver la incertidumbre para identificar el momento adecuado de retirar el ventilador mecánico a los pacientes de UCI, ya que si no se tomaba la decisión correcta se correría el riesgo de aumentar las probabilidades de muerte del 25% de pacientes. Para ello se utilizó una base de datos de 94 pacientes satisfactoriamente extubados y otros 38 que no, donde las variables de entrada se basaban en el flujo respiratorio de cada paciente (promedio de la serie del tiempo de expiración, el rango intercuartil del

tiempo de expiración, etc.), una vez implementado el método se procedió a la validación del sistema donde los resultados mostraron una exactitud del 78%, una sensibilidad de 75% y una especificidad de 74%.

Sharma et al. (2016) India en su artículo "Análisis de rendimiento de las técnicas de minería de datos para la Clasificación en Salud pública" presentado en la revista "International Journal of Innovative Research in Computer and Communication Engineering" realizaron un análisis de desempeño de distintas técnicas de clasificación y ayudó a encontrar la técnica con mejor rendimiento utilizando un determinado conjunto de datos, las técnicas estudiadas fueron KNN, Naive Bayes, Árbol de decisión y sus rendimientos se midieron en base a su exactitud. Los autores concluyeron que Naive Bayes era una técnica con mayor precisión y menor tasa de error en comparación con KNN y árboles de decisión ya que obtuvo una precisión de 77.94%.

Luego de haber mostrado los distintos trabajos en los que se aplicó Naive Bayes se realizó una comparación de los métodos utilizando atributos que permitan evaluar correctamente las características que cada método nos ofrece. Según Khan et al., (2016) los criterios que nos dan un resultado preciso se basan en la velocidad al obtener los resultados y los tipos de datos que puede manejar el método mientras que Garg et al. (2013) resalta el manejo de ruidos o datos corruptos.

Los criterios elegidos para la comparación fueron los que menciona Paulino (2021) en su tesis:

- **Naturaleza de Datos:** Indica si el dato de entrada puede ser discreto, continuo o de ambos tipos.
- **Cantidad de datos:** Volumen de registros necesarios a para entrenar el modelo y obtener un buen performance.

- **Interpretabilidad:** Hace referencia a la capacidad de interpretar el modelo y entender las razones de sus resultados.
- **Velocidad:** Mide el tiempo maquina necesario para que el modelo sea entrenado y nos devuelva un resultado.
- **Precisión:** Es una métrica muy importante al evaluar el modelo, ya que nos indica un porcentaje de similitud entre los datos reales y las predicciones.
- **Manejo de Ruido:** Mide la capacidad del modelo para tratar datos faltantes o outliers.
- **Área de Dominio:** Representa qué tan limitada es la técnica respecto a la complejidad de los diferentes dominios de conocimiento en los cuales puede ser aplicada.
- **Nivel de complejidad:** Mide el tiempo y esfuerzo que tarda un data science en implementar y validar la técnica de machine learning.

En la tabla 9. Se muestra cada uno de los criterios junto con los valores numéricos que estos pueden tomar, cada valor numérico representa un puntaje, estos son de suma importancia en la evaluación de técnicas que se realizara.

Una vez descritos y asignados los valores, se coloca los puntajes a cada uno de los métodos en comparación, dependiendo del criterio que se está evaluando como se observa en la tabla 10. Por último, se realiza la sumatoria y asignación de los valores totales.

Tabla 8

Descripción de los Criterios de Evaluación

CRITERIO	VALOR	DESCRIPCION	PUNTAJE
Naturaleza de Datos	Continuo o Discreto	El método solo admite que el dato de entrada sea discreto o que sea variable.	1
	Ambos	El método es adaptable a ambos tipos de datos.	2
Cantidad de Datos de Entrenamiento	Alto	Se requiere de un alto número de datos de entrenamiento para alcanzar un alto rendimiento.	1
	Bajo	No se requiere de muchos datos de entrenamiento para alcanzar un alto rendimiento.	2
Interpretabilidad	No interpretable	El método no cuenta con un razonamiento simbólico y representación semántica.	1
	Compleja	El método es descriptivo y requiere cierta interpretación.	2
	Fácil	Se presentan los resultados de manera visual o al menos de manera que su interpretación sea muy clara.	3
Velocidad	Baja	El método requiere de un alto costo computacional, incluso si la cantidad de datos a manipular no es alta, por lo que los tiempos de respuesta son lentos.	1
	Alta	El método no siempre requiere de un alto costo computacional, dependiendo de la cantidad de datos a manipular, los tiempos de respuesta pueden ser incluso inmediatos.	2
Precisión	Baja	El método tiene una precisión ≤ 70%.	1
	Media	La precisión está entre 70% y 85%.	2
	Alta	El método tiene un indicador de precisión mayor a 85%	3
Manejo de Ruido	Bajo	El método tiende a cometer imprecisiones cuando recibe datos con alta presencia de ruido.	1
	Medio	Significa que la técnica es capaz de reconocer algunos datos ruidosos, sin embargo, no es constante y su tasa de precisión puede verse afectada notablemente.	2
	Alto	El método maneja eficientemente cualquier dato recibido, almacenado o editado sin que esto afecte la precisión del resultado.	3
Area de Dominio	Alta	Se requiere conocimiento completo del dominio que abarca el sistema.	1
	Media	Se requiere una delimitación en el dominio de trabajo para la construcción.	2
	Baja	La construcción no requiere ningún dominio de conocimiento.	3
Nivel de Complejidad	Alta	El método utiliza algoritmos complejos que requieren de una alta curva de aprendizaje y son difíciles de implementar si no se cuenta con experiencia previa.	1
	Media	La implementación de los algoritmos tiene una curva media de aprendizaje, no son complicados de implementar, pero si requieren experiencia previa para su buen entendimiento.	2
	Baja	Los algoritmos utilizados son sencillos e intuitivos, con una curva de aprendizaje alta, y no requieren experiencia previa para ser implementados.	3

Fuente: Elaboración Propia

Tabla 9
Benchmarking de las diferentes técnicas analizadas

CRITERIO/TECNICA	SISTEMAS NEURO–DIFUSOS	REGRESION LOGISTICA	NAIVE BAYES	RANDOM FOREST	GRADIENT BOOSTING	ID3	C4.5
Naturaleza de los Datos	2	1	2	2	2	2	2
Cantidad de datos de entrenamiento	1	2	2	1	1	1	1
Interpretabilidad	3	3	3	2	1	2	2
Velocidad	2	2	2	1	1	1	1
Precisión	2	2	2	3	3	2	3
Manejo de Ruido	2	1	2	3	3	2	2
Área de Dominio	1	2	2	3	3	3	3
Nivel de Complejidad	2	3	3	2	1	1	2
TOTAL	15	15	18	17	16	14	15

Fuente: Elaboración Propia

La técnica seleccionada con mayor puntaje en el cuadro comparativo es Naive Bayes con un total de 18 puntos, gracias a su manejo de valores faltantes y su robustez a atributos irrelevantes, además de la gran ventaja que tiene al no requerir gran cantidad de datos para su entrenamiento, criterio que es favorable para nuestra investigación.

En conclusión, considerando todos los resultados, es evidente que Naive Bayes es la técnica que será capaz de predecir con precisión la obesidad infantil en los hospitales públicos de Lima; esta técnica es rápida, robusta, y será la adecuada para implementar, ya que, es computacionalmente barata y los conjuntos de datos de la historia clínica de los niños son en su mayoría ruidosos.

4.2. Naive Bayes para predicción de Obesidad Infantil

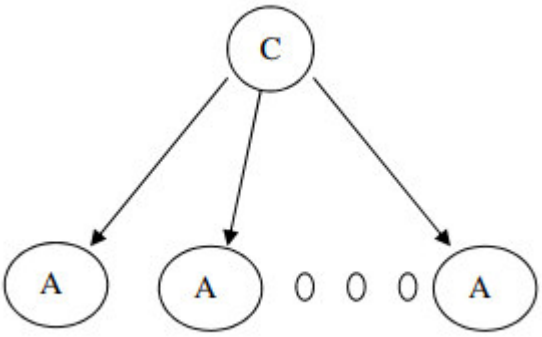
4.2.1. Definición

Alarcón (2015) define la técnica Naive Bayes como un clasificador Bayesiano, ya que esta asigna la clase con mayor probabilidad a cada observación; esta técnica trabaja bajo el fundamento de que todas las variables son independientes entre sí.

Sea A_1, A_2, …, A_n las variables que permiten predecir el valor de la clase C. La suposición de independencia asumida por el clasificador Naive Bayes da lugar a un modelo de red bayesiana con estructura simple descrita en la Figura 33. En él existe un único nodo raíz (la clase), y en la que todos los atributos son nodos hoja que tienen como único padre a la variable clase.

Figura 33 .

Estructura del Clasificador Naive Bayes



Fuente: (Alarcón, 2015)

Para el proceso de clasificación Naive Bayes se utiliza la ecuación (3.1) tomando en cuenta la suposición de que los atributos son condicionalmente independientes dada la clase, la probabilidad de que una observación o conjunto de atributos {a1, a2, ..., an} pertenezca a la clase "c" es

$$P\left(\frac{C=c}{a_1,a_2,\ldots,a_n}\right) = \frac{P(C=c)P(a_1/C=c)P(a_2/C=c)..P(a_3/C=c)}{P(a_1,a_2,\ldots,a_n)} \qquad (3.1)$$

con c1, c2, ..., cm esta ecuación se utilizará para la obtención de las probabilidades en la tarea de clasificación

4.2.2. Características Generales

A continuación, se muestra las principales ventajas y desventajas de implementar un modelo predictivo con Naive Bayes (ver Tabla 11)

Tabla 10.

Características Naive Bayes

VENTAJAS	DESVENTAJAS
• Funciona bien en la predicción de varias clases. • Necesita menos datos de entrenamiento. • Tiene un buen rendimiento en el caso de las variables categóricas de entrada en comparación con las variables numéricas. Para la variable numérica, se supone una distribución normal (curva de campana, que es una suposición fuerte).	• En caso no se observe una categoría de las clases predictivas en el conjunto de datos de prueba, esta categoría obtendrá un 0% de probabilidad y el modelo no podrá realizar la predicción, este caso se conoce como" Frecuencia Cero" • Naive Bayes trabaja asumiendo que las variables predictoras son independientes, aun cuando en la vida real es poco probable que un conjunto de datos contenga predictores 100% independientes.

Fuente: (Gonzalez, 2019)

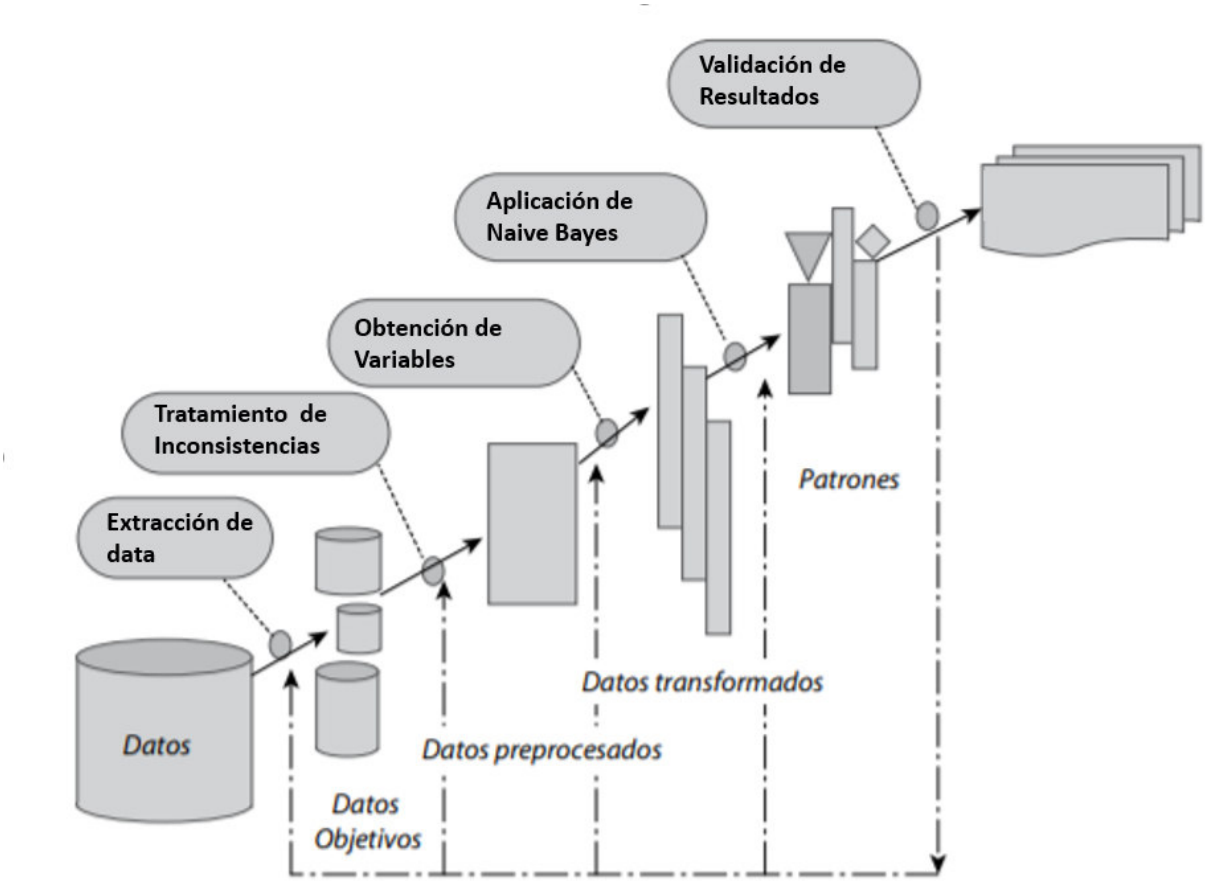
4.2.3. Metodología para Implementar Naive Bayes

Para la implementación de un modelo predictivo en el ámbito de medicina, es necesario comprender el problema y generar una serie de actividades que nos permitan lograr nuestro objetivo, por ello elegimos la metodología KDD (Knowlegde Discovery in Database) cuya finalidad es "la interpretación de patrones, modelos y un profundo análisis de la información" (ESAN, 2018) y es muy

usada en Minería de Datos. En la figura 34 se muestra las fases de la metodología KDD que desarrollaremos.

Figura 34 .

Representación Gráfica de la Metodología KDD



Fuente: Timarán (2016)

Fase 1 – Extracción de datos

Para esta primera fase de extracción de datos Timarán (2016) nos indica que es necesario haber definido previamente las metas del proceso, para luego poder crear un conjunto de datos objetivo sobre el cual se aplicará las siguientes fases de KDD, la selección de datos variará dependiendo de los objetivos que busque el negocio. Esta fase es crítica para

el éxito del modelo, ya que no recolectar una variable que contenga información importante puede reducir la precisión o efectividad del modelo predictivo.

Fase 2 – Tratamiento de Inconsistencias

En la etapa de tratamiento de inconsistencias Timarán (2016) recomienda analizar la calidad de los datos de la mano del experto. Mediante transformaciones se removerán datos ruidosos del conjunto de datos, también se eliminará datos duplicados y se aplicarán técnicas de reemplazo como la media, moda, mínimo o máximo para sustituir datos nulos.

Fase 3 – Obtención de Variables

Según indica Timarán (2016) la etapa de obtención de variables tiene como objetivo reducir el conjunto de datos horizontal como verticalmente. Se eliminan tuplas idénticas de registros para la reducción horizontal y la reducción vertical se centra en técnicas de eliminación de variables basándose en su significancia en el problema o excluyendo variables que contenga la misma información expresada de manera distinta (edad y fecha de nacimiento).

El desbalanceo de datos es muy común en conjuntos de registros médicos, debido a que en ciertas enfermedades es poco probable que el porcentaje de pacientes con el diagnóstico positivo sea mayor que el porcentaje con diagnóstico negativo, y este desbalanceo provoca que el modelo prediga mal la clase de menor porcentaje, por ello es importante utilizar algún algoritmo que balancee nuestros datos. Existen varios algoritmos, y en nuestro modelo utilizaremos el OverSampling que nos permitirá duplicar instancias de la clase minoritaria de manera aleatoria consiguiendo una paridad entre la cantidad de datos de las distintas clases.

Fase 4 – Aplicación de Naive Bayes

En la fase de aplicación de la técnica se elige la técnica apropiada de Minería de Datos, ya sea la clasificación, regresión o agrupación, según los objetivos que se haya planteado para la investigación nuestra tesis. "Una vez seleccionada la técnica, el siguiente paso es aplicarla a los datos ya seleccionados, limpiados y procesados. Es posible que el algoritmo se tenga que ejecutar varias veces para ajustar los parámetros que optimicen los resultados". (Landa, 2016)

4.2.4. Ejemplo usando la técnica Naive Bayes

Para poder ejemplificar las fases de la Técnica Naive Bayes usaremos un ejemplo extraído de Alarcón (2015), que trata sobre el diagnóstico de lentes de contacto.

- **Extracción de datos**

Se trabajará con datos clínicos de cada paciente recolectando información sobre su edad, si tiene lagrimeo o si padece de astigmatismo, con esos datos se predecirá si el paciente usará o no usará lentes y cuál será el tipo. Para demostrar el paso a paso de la implementación del algoritmo de Naive Bayes se utilizó un conjunto de datos que contiene 24 casos que se muestran en la tabla 12.

- **Preparación de Datos:**

Lo primero que debemos hacer es cargar nuestro conjunto de datos para el diagnóstico de lentes de contacto. Cada una de las variables posee una descripción y distintos valores los cuales se definen en la tabla 13.

Tabla 11.
Conjunto de datos del Diagnóstico de lentes de contacto

Paciente	Edad	Padecimiento	Astigmatismo	Lagrimeo	**Tipo de lentes**
1	Joven	Hipermétrope	Si	reducido	Ninguno
2	Joven	Hipermétrope	Si	normal	Duro
3	Joven	Hipermétrope	No	reducido	Ninguno
4	Joven	Hipermétrope	No	normal	Suave
5	Joven	Miope	Si	reducido	Ninguno
6	Joven	Miope	Si	normal	Duro
7	Joven	Miope	No	reducido	Ninguno
8	Joven	Miope	No	normal	Suave
9	Pre-Presbióptico	Hipermétrope	Si	reducido	Ninguno
10	Pre-Presbióptico	Hipermétrope	Si	normal	Ninguno
11	Pre-Presbióptico	Hipermétrope	No	reducido	Ninguno
12	Pre-Presbióptico	Hipermétrope	No	normal	Suave
13	Pre-Presbióptico	Miope	Si	reducido	Ninguno
14	Pre-Presbióptico	Miope	Si	normal	Duro
15	Pre-Presbióptico	Miope	No	reducido	Ninguno
16	Pre-Presbióptico	Miope	No	normal	Suave
17	Presbióptico	Hipermétrope	Si	reducido	Ninguno
18	Presbióptico	Hipermétrope	Si	normal	Ninguno
19	Presbióptico	Hipermétrope	No	reducido	Ninguno
20	Presbióptico	Hipermétrope	No	normal	Suave
21	Presbióptico	Miope	Si	reducido	Ninguno
22	Presbióptico	Miope	Si	normal	Duro
23	Presbióptico	Miope	No	reducido	Ninguno
24	Presbióptico	Miope	No	normal	Ninguno

Fuente: *(Alarcón, 2015)*

Tabla 12 .

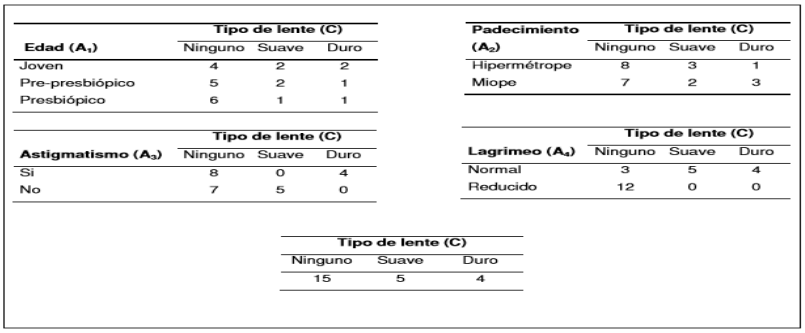
Descripción de las variables

Variables	Valores	Notación
A_1: Edad	Joven, Pre-Presbióptico, Presbióptico	A_1: a_{11}, a_{12}, a_{13}
A_2: Padecimiento	Hipermétrope, Miope	A_2: a_{21}, a_{22}
A_3: Astigmatismo	Sí, No	A_3: a_{31}, a_{32}
A_4: Lagrimeo	Normal, Reducido	A_4: a_{41}, a_{42}
C: Clase (Tipo de lente)	Ninguno, Suave, Duro	C: c_1, c_2, c_3

Fuente: *(Alarcón, 2015)*

Con los datos mostrados anteriormente realizamos un conteo, el resultado del conteo se muestra en la figura 35, qué está compuesta por 5 tablas de contingencia que indican la frecuencia de cada una de las variables

Figura 35.
Tablas de Contingencia de las variables



Edad (A_1)	Tipo de lente (C)		
	Ninguno	Suave	Duro
Joven	4	2	2
Pre-presbiópico	5	2	1
Presbiópico	6	1	1

Padecimiento (A_2)	Tipo de lente (C)		
	Ninguno	Suave	Duro
Hipermétrope	8	3	1
Miope	7	2	3

Astigmatismo (A_3)	Tipo de lente (C)		
	Ninguno	Suave	Duro
Si	8	0	4
No	7	5	0

Lagrimeo (A_4)	Tipo de lente (C)		
	Ninguno	Suave	Duro
Normal	3	5	4
Reducido	12	0	0

Tipo de lente (C)		
Ninguno	Suave	Duro
15	5	4

Fuente: (Alarcón, 2015)

- **Creación del Modelo:**

Luego de obtener el conteo de cada variable, calculamos las probabilidades utilizando el teorema de bayes, pero al aplicarlo al caso de padecer astigmatismo dado que se usa lente suave (en la tabla de conteo), la probabilidad asignada fue de 0 sobre 5, lo cual eliminaría la posibilidad de que se presente este caso, pero en circunstancias reales no se puede asegurar a

priori que nunca se presentará al consultorio una persona así. Por ello, para eliminar el problema presentado en la figura 35 como "Frecuencia Cero", se utilizó la corrección de Laplace que se define en la ecuación 3.2 (Alarcón, 2015).

$$P(ai/c) = \frac{n(ai,C=c)+1}{n(C=c)+ri} \tag{3.2}$$

Empleando la Corrección de Laplace a la distribución de probabilidad de la clase C, el resultado se muestra en la figura 36.

Figura 36.

Tabla de Distribuciones de las Probabilidades por Clase

$$P(c_1)=\frac{15+1}{24+3}=\frac{16}{27} \qquad P(c_2)=\frac{5+1}{24+3}=\frac{6}{27} \qquad P(c_3)=\frac{4+1}{24+3}=\frac{5}{27}$$

Fuente*:* (Alarcón, 2015)

De la misma manera, se aplica a cada una de probabilidades condicionales, calculando los valores para cada $P(A_i/C)$, cómo podemos observar en la figura 37.

Figura 37.

Figura de Distribuciones de Probabilidades Condicionales

C	$P(A_1/C)$		
C_1	5/18	6/18	7/18
C_2	3/8	3/8	2/8
C_3	3/7	2/7	2/7

C	$P(A_2/C)$	
C_1	9/17	8/17
C_2	4/7	3/7
C_3	2/6	4/6

C	$P(A_3/C)$	
C_1	9/17	8/17
C_2	1/7	6/7
C_3	5/6	1/6

C	$P(A_4/C)$	
C_1	4/17	13/17
C_2	6/7	1/7
C_3	5/6	1/6

Fuente: (Alarcón, 2015)

- **Evaluación del Modelo:**

En este módulo se procederá a realizar la predicción para un nuevo paciente que llegue al consultorio, utilizando al momento de la predicción los cálculos obtenidos en el módulo de entrenamiento. Los datos del nuevo paciente se muestran en la tabla 14.

Tabla 13 .

Datos del Nuevo Paciente

Paciente	Edad	Padecimiento	Astigmatismo	Lagrimeo	Tipo de Lente	Probabilidad
1	Joven (a_{11})	Hipermétrope (a_{21})	Si (a_{31})	Reducido (a_{42})		

Fuente: (Alarcón, 2015)

Entonces la probabilidad para esta combinación de atributos será reemplazada en la ecuación 3.3. donde α es una constante de proporcionalidad

$$P\left(\frac{C=c}{a_{11},a_{21},a_{31,}\alpha_{42}}\right) = \alpha P(C=c)P\left(\frac{a_{11}}{C}=c\right)P\left(\frac{a_{21}}{C}=c\right)P\left(\frac{a_{31}}{C}=c\right)P\left(\frac{a_{42}}{C}=c\right) \quad (3.3)$$

Luego hallamos las probabilidades para los valores de la clase C: c_1, c_2 y c_3

$$P(\frac{c_1}{a_{11},a_{21},a_{31,}\alpha_{42}}) = \alpha P(c_1)P(a_{11}/c_1)P(a_{21}/c_1)P(a_{31}/c_1)P(a_{42}/c_1)$$

$$P(\frac{c_2}{a_{11},a_{21},a_{31,}\alpha_{42}}) = \alpha P(c_2)P(a_{11}/c_2)P(a_{21}/c_2)P(a_{31}/c_2)P(a_{42}/c_2)$$

$$P(\frac{c_3}{a_{11},a_{21},a_{31,}\alpha_{42}}) = \alpha P(c_3)P(a_{11}/c_3)P(a_{21}/c_3)P(a_{31}/c_3)P(a_{42}/c_3)$$

Reemplazando, de las tablas de probabilidades

$$P\left(\frac{c_1}{a_{11}, a_{21}, a_{31,}\alpha_{42}}\right) = \alpha\frac{16}{27}\frac{5}{18}\frac{9}{17}\frac{9}{17}\frac{13}{17} = \alpha(0.03528)$$

$$P\left(\frac{c_2}{a_{11}, a_{21}, a_{31,}\alpha_{42}}\right) = \alpha\frac{6}{27}\frac{3}{8}\frac{4}{7}\frac{1}{7}\frac{1}{7} = \alpha(0.00097)$$

$$P\left(\frac{c_3}{a_{11}, a_{21}, a_{31,}\alpha_{42}}\right) = \alpha\frac{5}{27}\frac{3}{7}\frac{2}{6}\frac{5}{6}\frac{1}{6} = \alpha(0.00367)$$

Normalizando los valores anteriores, obtenemos

$$\alpha(0.03528 + 0.00097 + 0.00367) = 1\,;\;\; \alpha = 25.05$$

$$P\left(\frac{c_1}{a_{11}, a_{21}, a_{31,}\alpha_{42}}\right) = \alpha(0.03528) = 0.884$$

$$P\left(\frac{c_2}{a_{11}, a_{21}, a_{31,}\alpha_{42}}\right) = \alpha(0.00097) = 0.024$$

$$P\left(\frac{c_3}{a_{11}, a_{21}, a_{31,}\alpha_{42}}\right) = \alpha(0.00367) = 0.092$$

Luego,

$$P\left(\frac{Ninguno}{evidencias}\right) = 0.884$$

$$P\left(\frac{Lente\ suave}{evidencias}\right) = 0.024$$

$$P\left(\frac{Lente\ duro}{evidencias}\right) = 0.092$$

Finalmente, el clasificador Naive Bayes asignara al paciente la clase con mayor probabilidad, es decir el paciente será diagnosticado con que no es necesario el uso de lentes, ya que se tiene una mayor probabilidad (0.884).

5. Capítulo 5: Modelo propuesto para la predicción de Obesidad en Edad Infantil

En el presente capítulo se describe los pasos para realizar la implementación del modelo predictivo con la técnica Naive Bayes basándonos en la metodología KDD descrita en el capítulo anterior.

5.1. Fase 1 – Extracción de datos

El conjunto de datos utilizados proviene del aplicativo e-QHALI del MINSA, estos fueron recolectados de los registros de 4 centros de salud (Centro Materno Infantil Laura Rodríguez, Patibamba Baja, Rebelde Huayrana y Huancaquito Alto) de las consultas atendidas en el área de Crecimiento y Desarrollo de dichos establecimientos entre los años 2011 y 2021.

El conjunto de datos inicial tuvo 244 421 registros, el siguiente paso fue realizar la selección de registros de manera que sean adecuados para la técnica que vamos a utilizar, estos filtros fueron:

1. El paciente cuenta con información de peso, talla y antecedentes maternos registrados (n = 3 779)
2. El paciente tiene que ser mayor a 4 años de edad (n = 18 793)
3. El paciente debe tener registradas al menos dos consultas de seguimiento que tengan los campos talla y pesos rellenados (n = 8 729)
4. Por último, es necesario que su edad en la última consulta de seguimiento sea mayor de 5 años, ya que esta información será utilizada como variable predictora dentro del modelo; y que la otra consulta sea en la edad menor a 4 años (n = 771)

Finalmente, el conjunto de datos luego de aplicarle los filtros mínimos recomendado por el médico cuenta con 780 registros, con los cuales trabajaremos para implementar el modelo.

5.2. Fase 2 – Tratamiento de Inconsistencias

El tratamiento de variables es la fase más crucial de la construcción de un modelo, por ello es importante comprender la naturaleza de la información con la que se trabajará. En esta fase se identificaron 27 variables que fueron extraídas de los registros filtrados en la anterior fase, estas describen los antecedentes clínicos de la madre, características del parto e información de peso y talla de las consultas de seguimiento del niño o niña; la tabla 15 muestra la definición de cada variable y el tipo de dato.

Un requisito para aplicar el modelo de Naive Bayes es la homogeneización de todas las variables, lo que significa que deben ser exclusivamente categóricas o continuas; en nuestro caso decidimos transformar todas las variables en formato continuo para capitalizar el potencial informativo de las variables numéricas, como el peso, la altura y los niveles de hemoglobina, en beneficio del modelo. Los detalles específicos de las variables que pasaron por una transformación se encuentran en la figura 38.

Después de completar la preparación y transformación de nuestro conjunto de datos, es crucial llevar a cabo una evaluación de las variables que contienen valores faltantes. Las variables con un porcentaje bajo de información disponible son eliminadas del conjunto, mientras que aquellas con solo algunos valores faltantes se completan utilizando el promedio de los datos existentes, agrupados según la variable "Peso para la edad gestacional". Una vez que hemos completado este proceso y tenemos el conjunto de datos listo, trabajamos en

colaboración con un experto para identificar y tratar los valores atípicos (outliers). Esto implica eliminar los registros que posiblemente han sido rellenados con información incorrecta bajo la guía y discernimiento del experto.

- **Id_nacimiento_talla:** Se observa que la talla promedio es 49cm, pero también existen casos de bebes que superan los 55cm.
- **Id_nacimiento_peso:** El peso promedio de los bebes recién nacido es 3200gr, pero existen un registro con un peso menor a 500gr, según el experto ese es un error al ingresas la información, ya que es imposible que se dé ese caso y se procede a eliminar el registro
- **Id_nacimiento_perimetro_toraxico**: El promedio del perímetro torácico al nacer es 34cm, pero existe un registro con perímetro torácico mayor a 47.5cm lo cual también se califica como error al ingresar la información, por lo cual se eliminará del dataset.
- **Id_nacimiento_edad_gestacional:** El promedio de la edad gestacional al nacer es de 39 semanas, en el dataset existe un registro con 32 semanas, al consultarlo con el experto nos indicó que se trata de un bebe prematuro.
- **Id_nacimiento_perimetro_cefálico:** La distribución del perímetro cefálico muestra un promedio de 34cm, por ello será necesario borrar los dos registros que tienen un perímetro superior a 44cm ya que el experto lo considera como información errónea.

Luego de haber reconocido los outliers de nuestro dataset que indican información errónea se procedió a eliminarlos, quedándonos registros para el entrenamiento y evaluación del modelo (Ver Anexo D).

Tabla 14 .

Lista de variables con definición

VARIABLE	DESCRIPCION	TIPO
	ANTECEDENTES	
Fecha de nacimiento	Fecha de nacimiento del niño	Fecha
# de embarazo	Numero de embarazo de la madre en la que nació	Continuo
# de atenciones prenatales	Numero de atenciones prenatales que tuvo la	Continuo
Lugar de atenciones	Centro médico donde se atendió	Discreto
	PARTO	
Condición del parto	El parto pudo ser Cesárea, Espontaneo,	Discreto
Sexo	El sexo del paciente puede ser "H" o "M"	Discreto
Lugar del parto	Lugar de atención del parto, puede ser Domicilio,	Discreto
Parto atendido por	Indica quien fue quien atendió el parto, si un	Discreto
Edad gestacional del	Semanas de embarazo antes de dar a luz	Continuo
Peso al nacer (gr)	Peso del niño al nacer	Continuo
Talla al nacer (cm)	Talla del niño al nacer	Continuo
Perímetro Cefálico al	Perímetro cefálico al nacer	Continuo
Peso para la edad	Puede ser Pequeño, Adecuado o Grande	Discreto
Perímetro torácico al	Perímetro torácico del niño al nacer	Continuo
Índice de apgar_1	Índice de APGAR en el 1'	Continuo
Índice de apgar_5	índice de APGAR en el 5'	Continuo
Enfermedad congénita	Indica si nació o no con una enfermedad	Booleano
Índice CPP	Indica si existió contacto piel a piel	Booleano
Índice AC	Indica si existió Alojamiento Conjunto al nacer	Booleano
Requirió hospitalización	Indica si nació o no con una enfermedad	Booleano
	POSTPARTO	
Fecha_Atencion_1	Fecha del primer control de crecimiento	Date
Peso_consulta_1	Peso del primer control de crecimiento	Continuo
Talla_consulta_1	Talla del primero control del crecimiento	Continuo
Resultado hemoglobina	Resultado de hemoglobina en alguna consulta	Continuo
Fecha_Atencion_2	Fecha del segundo control de crecimiento	Date
Peso_consulta_2	Peso del segundo control de crecimiento	Continuo
Talla_consulta_1	Talla del segundo control del crecimiento	Continuo

Fuente: Elaboración Propia

Figura 38
Variables transformadas para el modelo

CONDICION DEL PARTO	
INFORMACION	TRANSFORMACION
Espontaneo	0
Cesárea	1
Instrumentado	2
Otro	3

CONDICION DEL PARTO	
INFORMACION	TRANSFORMACION
Espontaneo	0
Cesárea	1
Instrumentado	2
Otro	3

ATENDIDO POR	
INFORMACION	TRANSFORMACION
Personal de Salud	0
Personal Técnico	1
Otro	2

ESTADO NUTRICIONAL	
INFORMACION	TRANSFORMACION
Bajo peso	0
Normal	1
Sobrepeso	2
Obesidad	3

LUGAR DEL PARTO	
INFORMACION	TRANSFORMACION
Establecimiento de Salud	0
Domicilio	1

PESO PARA EDAD GESTACIONAL	
INFORMACION	TRANSFORMACION
Pequeño	0
Grande	1
Adecuado	2

EDAD	
INFORMACION	TRANSFORMACION
H	0
M	1

Fuente: Elaboración Propia

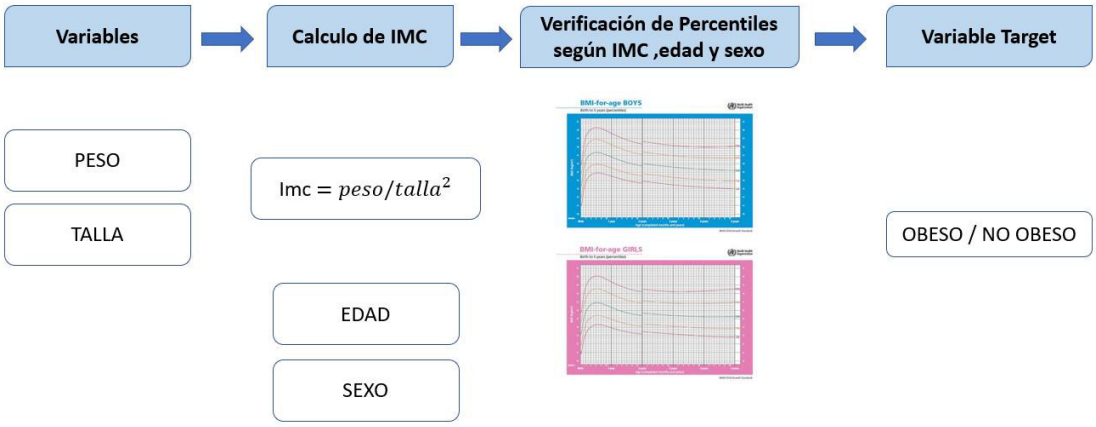
5.3. Fase 3 – Obtención de Variables

Después de establecer las variables en consulta con el experto, se generó una nueva variable llamada "TARGET". Esta variable se construyó usando la edad, el género, el peso y la altura del niño en la última evaluación con el propósito de clasificarlo como "OBESO" o "NO OBESO". El proceso para crear esta variable fue supervisado por un médico. Comenzó con el cálculo del Índice de Masa Corporal (IMC), como se muestra en la figura 39. Una vez calculado el IMC, se procedió a determinar la situación nutricional mediante la utilización de las tablas de percentiles proporcionadas por la Organización Mundial de la Salud (OMS)

(consulte el Anexo A y B). La clasificación dependió del peso y la edad del paciente en cuestión. (Becerra Romero & Huayna Dueñas, 2022)

Figura 39 .

Proceso para cálculo de la Variable Target



Fuente: Elaboración Propia

Luego de obtener la variable TARGET se procede a contar cuantos casos tenemos por cada clase (OBESO / NO OBESO), como se observa en la tabla 16, se obtuvo 46 pacientes en la clase OBESO y 725 en la clase NO OBESO.

Tabla 15.

Cantidad de Pacientes por Categoría

RESULTADO	CANTIDAD
Obeso	46
No Obeso	725

Fuente: Elaboración propia

Como podemos observar nuestro conjunto de datos final esta desbalanceado, solo el 6% de registros pertenece a la clase minoritaria, esta distribución de datos suele presentarse en

conjunto de datos de diagnóstico de enfermedades. Por ello procederemos a implementar el algoritmo OverSampling que nos permitirá balancear nuestros datos replicando instancias de la clase menor. Como se observa en la figura 40, la clase minoritaria paso de tener 46 registros a 725 registros luego de aplicar el algoritmo, este será nuestro conjunto de datos final con el cual entrenaremos nuestro modelo predictivo.

Figura 40.
Aplicación de algoritmo OverSampling

```
[101]: from collections import Counter
       from sklearn.datasets import make_classification
       from imblearn.over_sampling import RandomOverSampler
       # Define dataset
       X = pacientes.copy()
       del X['estado_nutricional']
       y = pacientes['estado_nutricional'].copy()
       # Calcula distribución de las clases
       print(Counter(y))
       # Define OverSampling
       oversample = RandomOverSampler(sampling_strategy='minority')
       # Aplica la transformación
       X_over, y_over = oversample.fit_resample(X, y)
       # Calcula la nueva distribución de clases
       print(Counter(y_over))

       Counter({1: 725, 0: 46})
       Counter({1: 725, 0: 725})
```

Fuente: Elaboración propia

5.4. Fase 4 – Aplicación de Naive Bayes

Para implementar el modelo Naive Bayes utilizamos la librería sklearn de Python, esta librería nos brinda módulos para la conversión de variables, separación de muestras para entrenamiento y validación del modelo, nos permite entrenar el modelo con los datos obtenidos y

también nos facilita el cálculo de métricas como la matriz de confusión y la precisión del modelo.

La librería Sklearn nos permite implementar 3 tipos distintos de algoritmo Naive Bayes:

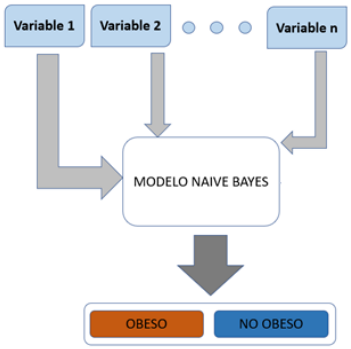
- Naive Bayes Gaussiano: Este se emplea cuando los valores predictores del modelo son continuos y se espera que sigan una distribución gaussiana.
- Naive Bayes Bernoulli: Se utiliza cuando los valores predictores son booleanos y se supone que siguen la distribución de Bernoulli.
- Naive Bayes Multinomial: Este clasificador se emplea con una distribución multinomial y se utiliza para la clasificación de textos.

Optamos por utilizar el clasificador Naive Bayes Gaussiano para nuestro modelo debido a que nuestros datos siguen una distribución normal.

Luego de seleccionar el modelo, el siguiente paso es definir los inputs o entradas para el proceso. Como se puede ver en la figura 41, contamos con 27 variables relacionadas con el paciente. Estas variables serán empleadas por el modelo Naive Bayes (Anexo C), que nos proporcionará una clase predicha para el niño o niña (Obeso o No Obeso) con la mayor probabilidad en función de las variables introducidas.

Figura 41.

Esquema del Modelo Naive Bayes



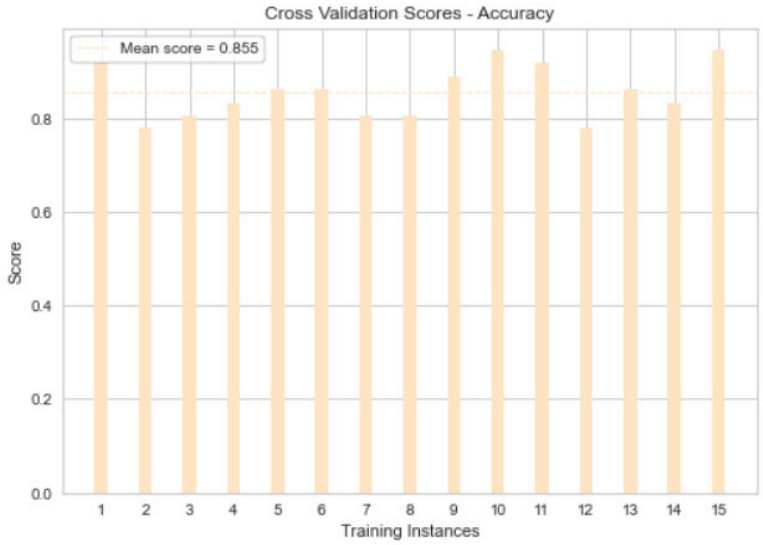
Fuente: Elaboración Propia

El siguiente paso fue importar las librerías que utilizaríamos para la implementación como se observa en las 3 primeras líneas del código, como ya teníamos los datos debidamente tratados en las fases anteriores, se procedió a dividirlos en subconjuntos de datos de entrenamiento (67%) y validación (33%).

```
1.  from sklearn. model_selection import train_test_split
2.  from sklearn. model_selection import cross_val_score, cross_validate.
3.  from sklearn. naive_bayes import GaussianNB.
4.  X_train, X_test, y_train, y_test = train_test_split (X, y, test_size=0.33)
5.  scoring = ['recall', 'precision','accuracy']
6.  clf = GaussianNB ()
7.  scores = cross_validate (clf, X_train, y_train, cv=15, scoring=scoring,
    return_estimator=True)
```

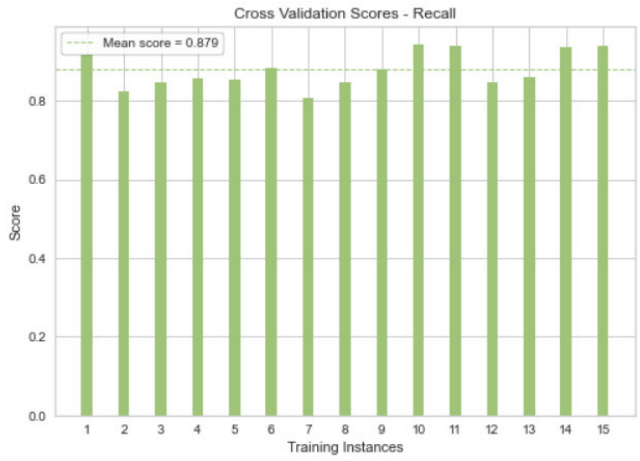
Luego de la definición del modelo, se procede a entrenarlo utilizando la técnica de Validación Cruzada para garantizar que los datos de entrenamiento y prueba sean independientes de la partición que se realice. Para nuestro caso se trabajó con 15 iteraciones, luego de ejecutar las iteraciones se obtuvo un promedio de 86% de accuracy (Figura 42), una media de 88% de recall (Figura 43), y un 97% de precisión (Figura 44).

Figura 42.
Score de Accuracy por cada iteración



Fuente: Elaboración Propia

Figura 43.
Score de Recall por cada iteración



Fuente: Elaboración Propia

Figura 44.
Score de Precisión por cada iteración

Fuente: Elaboración Propia

Después de completar el proceso de entrenamiento del modelo, este será guardado en un archivo denominado "finalized_model.sav". Este archivo nos permitirá cargar el modelo posteriormente, facilitando su conexión con la aplicación web que funcionará como interfaz. A través de esta aplicación, los expertos podrán realizar predicciones para sus nuevos pacientes de manera conveniente.

6. Capítulo 6: Desarrollo del Sistema para la Predicción de la Obesidad Infantil

En este capítulo se describen las características de la aplicación web de predicción de obesidad infantil, en donde se indica la arquitectura del sistema, el modelado de casos de uso y el modelo de la base de datos.

6.1. Descripción General del sistema

El sistema desarrollado tiene como principal función la predicción de obesidad infantil de los pacientes menores de 5 años de edad por parte de un doctor, y como funciones adicionales el permitir el ingreso de información de pacientes nuevos, además de mostrar un historial de predicciones por cada paciente para su respectivo seguimiento.

El alcance del Sistema abarca lo siguiente:

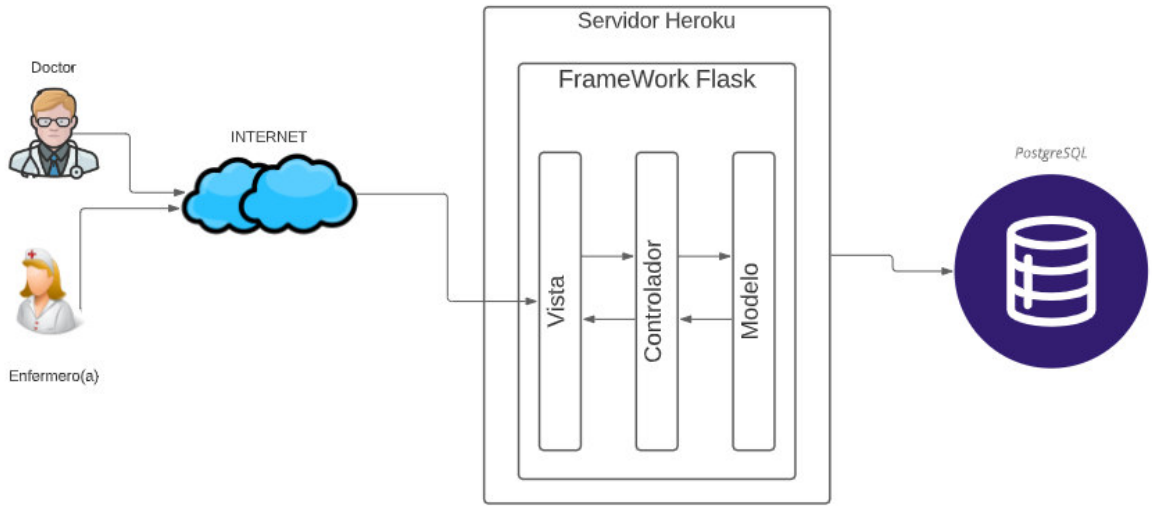
- Permitir el ingreso y actualización de los historiales médicos
- Predecir la obesidad en un paciente en edad infantil.
- Generar un registro de resultados de predicción por cada paciente.

6.2. Arquitectura del sistema

Al ser necesaria la creación de una aplicación web para consultar el modelo predictivo, se definió una arquitectura hibrida; por un lado se encuentra el patrón Cliente – Servidor donde el doctor accede a la web como cliente para realizar sus consultas a la aplicación que se encuentra desplegada en un servidor de Heroku y por otro lado está el patrón Modelo-Vista-Controlador ; este es utilizado por el framework Flask que se encuentra implementado en Python, el cual nos permite separar las 3 capas de la aplicación de una

manera más transparente al usuario y facilita que la aplicación sea fácilmente escalable en un futuro. Dentro de la capa Modelo, se utiliza la librería sklearn de Python para cargar y consultar el modelo predictivo. Por último, los datos son guardados en una base de datos PostgreSQL. La arquitectura se muestra en la figura 45.

Figura 45
Arquitectura del Sistema de Predicción



Fuente: Elaboración propia

6.3. Modelado de Casos de Uso del Sistema

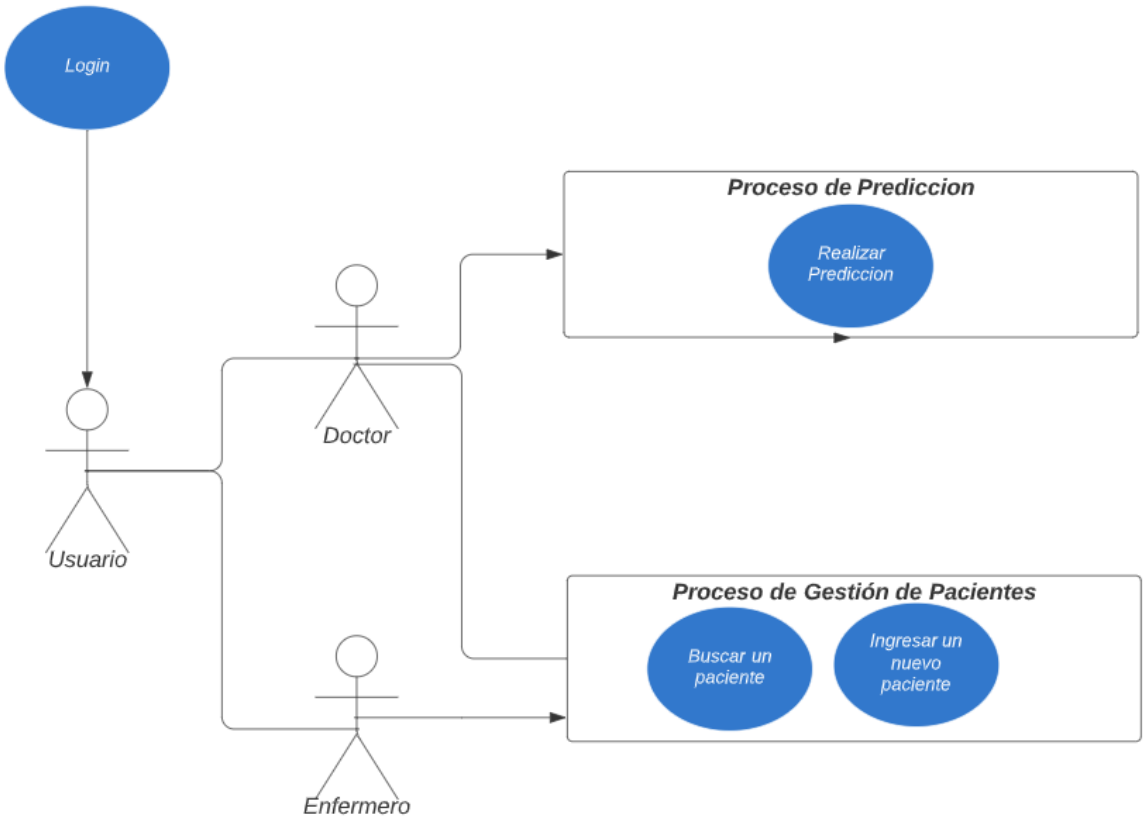
Para iniciar con el modelado del sistema fue necesario describir a los usuarios y los casos de uso. La aplicación web cuenta con 2 tipos de usuarios:

- El enfermero que es responsable de añadir información del paciente
- El doctor que se encarga de realizar la predicción del estado nutricional del paciente que se encuentra registrado, y también de realizar un seguimiento de la predicción hecha contra el peso actual del niño en cada consulta.

En la figura 46 podemos observar el diagrama de casos de uso agrupado por paquete. Los paquetes de casos de uso del sistema son:

- Proceso de Predicción
- Proceso de Gestión de Pacientes

Figura 46
Modelo de Casos de Uso



Fuente: Elaboración propia

A continuación, se procede a mostrar los casos de uso del sistema.

<table>
<tr><td>ID</td><td>CUS01</td></tr>
<tr><td>Caso de Uso</td><td>Login</td></tr>
<tr><td>Actor</td><td>Usuario</td></tr>
<tr><td>Descripción</td><td>Este caso de uso permite validar al usuario del sistema con el identificador y contraseña previamente asignados.</td></tr>
<tr><td>Precondición</td><td>Ninguno</td></tr>
<tr><td colspan="2">Flujo Principal</td></tr>
<tr><td colspan="2">1. El caso uso inicia cuando el usuario ingresa al sistema.

2. El sistema muestra un formulario con los siguientes elementos: Ingresar DNI, ingresar contraseña y el botón para el logeo.

3. El usuario ingresa el identificador, la contraseña y presiona el botón “Ingresar”.

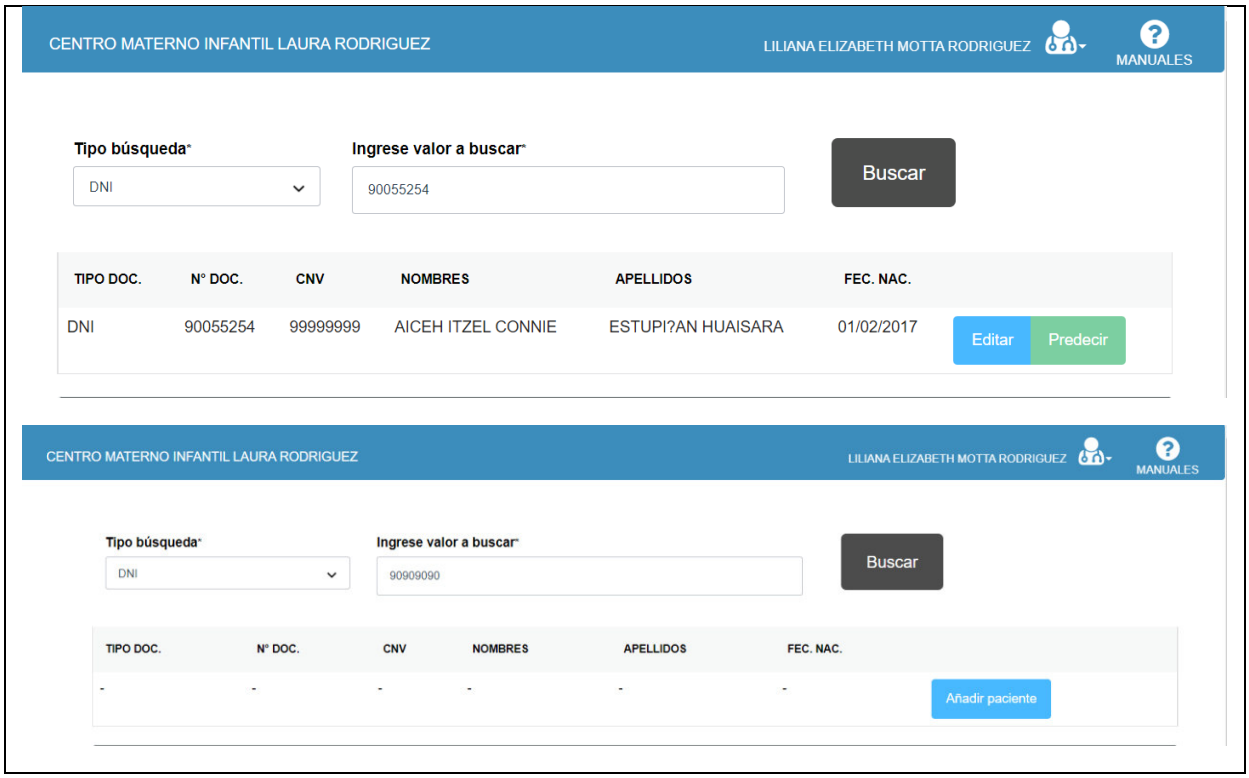
4. El sistema valida la existencia de datos ingresados en los campos del identificador y la contraseña.

5. El sistema:

• Busca el DNI de usuario.
• Compara la contraseña ingresada con la del identificador encontrado.
• Lee el perfil asociado al identificador del usuario.
• Asigna los permisos para ingresar al sistema.
• Muestra la pantalla de menú inicial según el perfil del usuario.

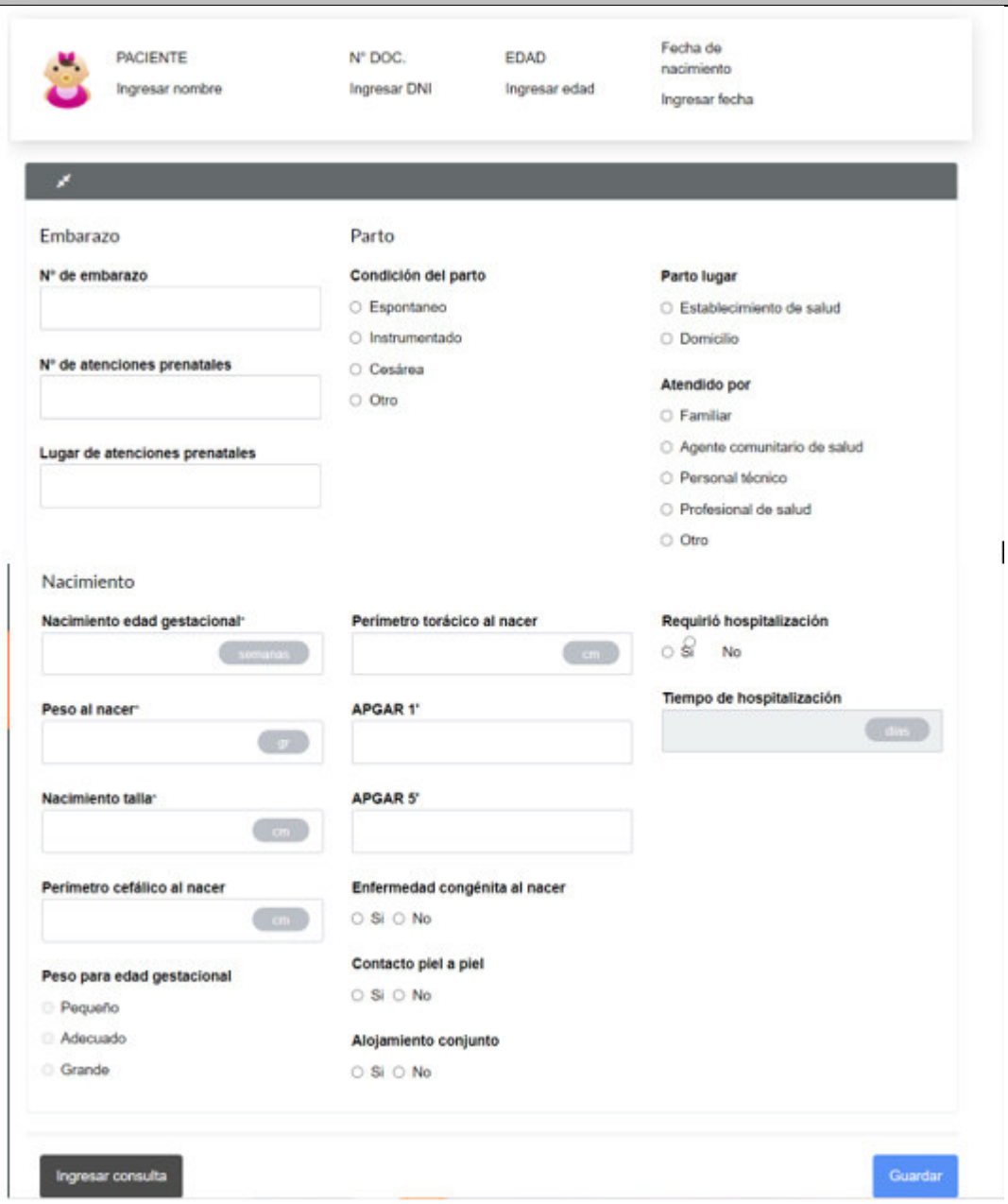
6. El usuario ingresa al sistema con el perfil asignado y el caso de uso finaliza</td></tr>
<tr><td>Post condición:</td><td>El usuario ha sido validado por el sistema y puede acceder al sistema con el perfil asociado.</td></tr>
<tr><td>Flujo Alterno</td><td>“Identificador o contraseña errónea”</td></tr>
<tr><td colspan="2">1. En el paso 5 el sistema encuentra que los datos del login son incorrectos, el sistema muestra el siguiente mensaje: “Identificador y/o Contraseña incorrecto” y el caso de uso continua el paso 2.</td></tr>
<tr><td colspan="2">Prototipo</td></tr>
</table>

ID	CUS02
Caso de Uso	Buscar paciente
Actor	Doctor o Enfermero
Descripción	Este caso de uso se ingresará al sistema un paciente que aún no se encuentre registrado.
Precondición	Que el enfermero este logeado correctamente.
Flujo Principal	
1. El caso uso inicia cuando el usuario ya está autentificado en el sistema e ingresa un número de DNI para validar la existencia en el sistema. 2. Dar clic al botón buscar 3. El sistema valida si el DNI ingresado se encuentra en la base 4. El sistema muestra el resultado de la búsqueda, con información del paciente y el botón para ejecutar la predicción	
Post condición:	Se muestra en pantalla el resultado de la búsqueda
Flujo Alterno	"No se encuentra registrado"
1. En el paso 4 el sistema no encuentra ningún registro para el paciente, por ello muestra la posibilidad de ingresar a un paciente nuevo, esto se revisará en el CUS03.	
Prototipo	



ID	CUS03
Caso de Uso	Ingresar nuevo paciente
Actor	Enfermero
Descripción	En Este caso de uso se ingresará al sistema un nuevo paciente
Precondición	Que el enfermero este logeado correctamente y se haya verificado que el paciente no existe.
Flujo Principal	
1. El caso uso inicia cuando el usuario ya está autentificado en el sistema, ya busco al paciente y el resultado fue sin registros. 2. El usuario da clic en el botón “Agregar paciente” 3. El sistema muestra el formulario con todos los campos a llenar 4. El usuario llena los campos y le da clic en guardar. 5. El sistema guarda la información. 6. El usuario da clic en “Agregar consulta” para agregar peso y talla de las consultas de	

desarrollo. Se repetirá el proceso para la cantidad de consultas que necesite llenar.	
Post condición:	El nuevo paciente se encuentra registrado
Prototipo	



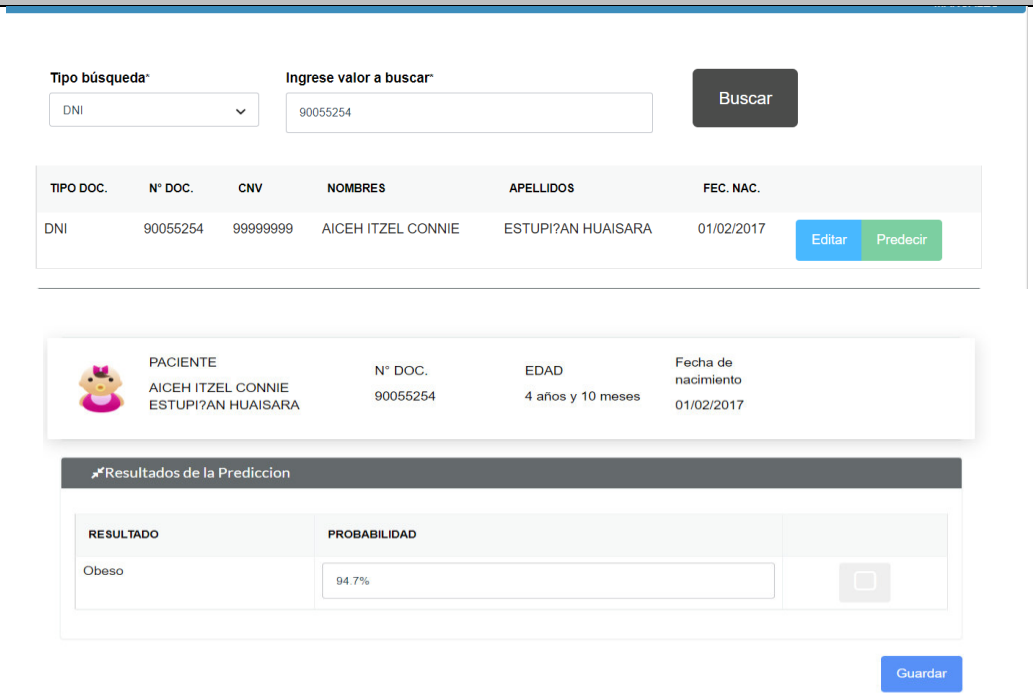
ID	CUS04
Caso de Uso	Realizar la predicción
Actor	Doctor
Descripción	En este caso de uso el doctor seleccionara un paciente para realizar la predicción
Precondición	Que el enfermero este logeado correctamente y se haya verificado que el paciente existe.

Flujo Principal

1. El caso uso inicia cuando el usuario ya está autentificado en el sistema, ya busco al paciente y el paciente existe.
2. El usuario da clic en el botón “Predecir”
3. El sistema muestra el resultado de la predicción, junto con la probabilidad
4. El usuario le da clic a guardar predicción.
5. El sistema guarda la información.

Post condición:	El paciente cuenta con una predicción registrada

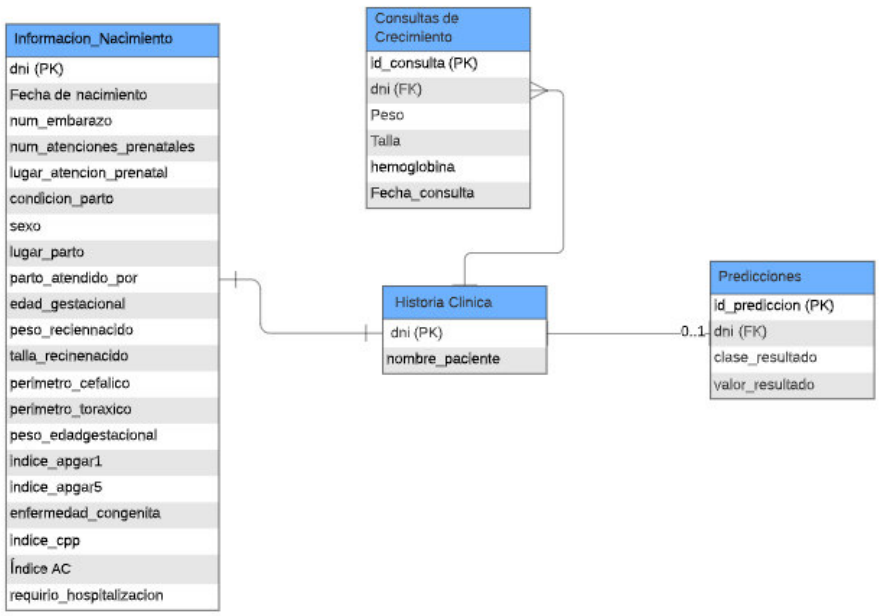
Prototipo



6.4. Modelo de datos

El modelo de base de datos que se muestra en la figura 47 es el que utilizamos para almacenar los datos de los pacientes, la información de nacimiento y el historial de predicciones.

Figura 47
Modelo de base de Datos del Sistema



Fuente: Elaboración Propia

7. Capítulo 7: Análisis de Resultados

En este capítulo se evalúa el modelo predictivo que fue construido en el capítulo anterior, se realiza el análisis de las métricas que mejor evalúen nuestro caso de estudio y se compara los resultados del modelo con el de otras técnicas.

7.1. Cálculo de Métricas

El modelo propuesto se evaluó sobre un conjunto de datos de 317 registros, basándonos en el tipo de variables que debíamos predecir (Obeso/No Obeso) se determinó utilizar las métricas de matriz de confusión, precisión, exactitud y sensibilidad; estas eran las más apropiadas debido a que nuestro target era una variable discreta y con un conjunto finito para clasificar.

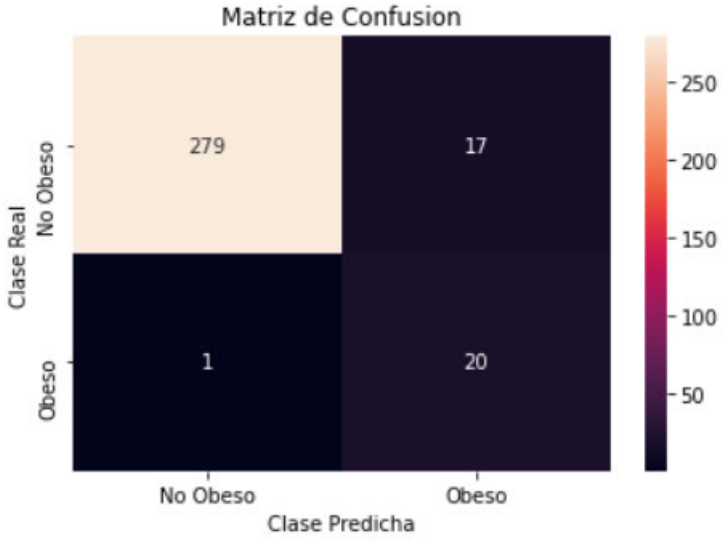
A) Matriz de Confusión

La matriz de confusión se calcula comparando el target de prueba contra el resultado de la predicción con la finalidad de mostrar de forma explícita cuando una clase es confundida con otra, para los 317 registros con los que se realizó la prueba se graficó la matriz de la figura 48, esta se obtiene separando los resultados en 4 categorías:

- Verdaderos negativos (VN): Son aquellos casos en los que el paciente no es obeso, y el modelo lo predice como no obeso (279 casos)
- Falsos positivos (FP): Son los casos en donde el paciente no es obeso y el modelo lo clasifica como obeso (17 casos)
- Falsos Negativos (FN): Son los casos en los que el paciente es Obeso y el modelo lo predice como no obeso (1 caso)

- Verdaderos Positivos (VP): Son los casos en los que el paciente es obeso y el modelo lo predice como obeso (20 casos)

Figura 48
Matriz de Confusión NB



Fuente: Elaboración propia

B) Exactitud o Accuracy

La exactitud de un modelo calcula el porcentaje de cuantas clases predijeron correctamente y se calcula sumando los 254 casos verdaderos positivos más los 13 casos verdaderos negativos y dividiendo entre los 317 casos en total, utilizando esta fórmula la precisión del modelo resulto 71.6%

$$Exactitud = \frac{(279 + 20\ )}{317} = 94.32\%$$

La precisión del modelo no fue muy alta, y esto es debido a que nuestro conjunto de datos está desbalanceado por la naturaleza del problema, esto quiere decir que la clase “No Obeso” es mucho más grande que la clase “Obeso” ya que es común que existan pocos

registros de niños obesos en un hospital en comparación de la cantidad de niños sin obesidad, lo que genera que el modelo generalice.

C) Sensibilidad

La sensibilidad se calcula dividiendo la cantidad de verdaderos positivos entre el total de casos en los que el resultado esperado era positivo. Esta es la mejor métrica para evaluar nuestro modelo ya que representa la capacidad que tiene para predecir correctamente la clase con menor cantidad de casos.

$$Sensibilidad = \frac{20}{20 + 1} = 95.23\%$$

D) Precisión

El cálculo de la precisión de la clase "no obeso", se calcula dividiendo los casos que son verdaderos negativos entre la suma de los casos verdaderos negativos más los falsos negativos.

$$Precisión = \frac{279}{1 + 279} = 99.6\%$$

7.2. Validación y comparación de técnicas

Se utilizaron las técnicas KNN, Regresión Logística, XGB y Random Forest para evaluar y contrastar sus resultados contra el resultado del modelo propuesto. Dado que nuestro conjunto de datos no es muy grande, se realizó la prueba con técnica de validación cruzada, que nos permitió ejecutar múltiples iteraciones con varios subconjuntos de datos. Estos métodos se implementaron en Python utilizando la biblioteca sklearn.

Para la técnica de regresión logística se definió 2 parámetros, el primero fue “C” que es una variable de control de la regularización y se le asigno 1, el segundo parámetro fue la penalidad al cual se le asigno L2 (Rigde Regresión), al evaluar la técnica se obtuvo una exactitud de 92.2% y una precisión de 93.3% como se observa en la tabla 18.

Tabla 16
Validación de Regresión Logística

REGRESIÓN LOGISTICA		
Parámetros		
C	Penalty	
1	L2	
Resultados		
Exactitud	Precisión	Sensibilidad
92.2%	93.3%	98.6%

Fuente: Elaboración propia

La siguiente técnica en evaluar fue Random Forest a la cual se le asigno 6 como máxima profundidad del árbol, el número de estimadores fue 100 y el valor mínimo para dividir un nodo fue 2. Con estos parámetros, RF obtuvo un 93.1% de Exactitud y 93.4% de Precisión como se muestra en la tabla 19.

Tabla 17
Validación de Random Forest

RANDOM FOREST		
Parámetros		
max_depth	min_samples_split	n_estimators
100	2	100
Resultados		
Exactitud	Precisión	Sensibilidad

93.1%	93.4%	99.5%

Fuente: Elaboración Propia

Al momento de evaluar la técnica KNN, se definió los parámetros que se observan en la tabla 20, el número de vecinos fue 5, para el tamaño de la hoja se mantuvo el valor por defecto y para el parámetro "P" se le asigno 2 que es equivalente a usar una distancia euclidiana. Se obtuvo como resultados una exactitud de 93.12% y una precisión de 93.8%.

Tabla 18
Validación de KNN

KNN		
Parámetros		
n_neighbors	leaf_size	P
5	30	2
Resultados		
Exactitud	Precisión	Sensibilidad
93.1%	93.8%	96%

Fuente: Elaboración propia

Por último, al evaluar la técnica XGB con el parámetro learning_rate igual a 0.3, la máxima profundidad igual a 6 y el booster definido como gbtree ya que cada iteración está basada en árboles. Con estos parámetros, XGB obtuvo un 92.7% de Exactitud y 97.08% de Precisión como se muestra en la tabla 21.

Tabla 19
Validación de XGB

XGB		
Parámetros		
Booster	Learning_rate	Max_depth

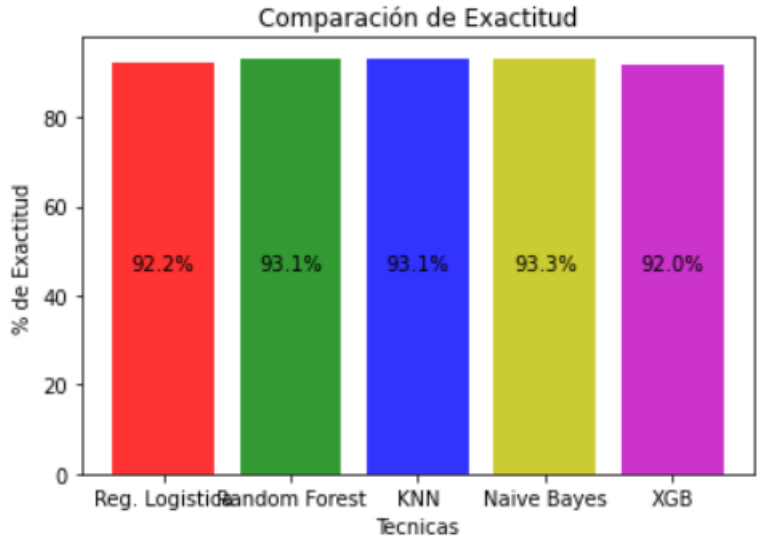
Gbtree	0.30	6
Resultados		
Exactitud	Precisión	Sensibilidad
92.7%	97.08%	98.1%

Fuente: Elaboración Propia

Después de llevar a cabo la evaluación de los cinco modelos, se procedió a comparar las métricas correspondientes a cada técnica. En la figura 49 se puede observar que la técnica que logró la mayor precisión fue Naive Bayes, con un 93.3%, seguida de cerca por Random Forest con un 93.1%. Esto indica que Naive Bayes obtuvo la mayor cantidad de predicciones correctas.

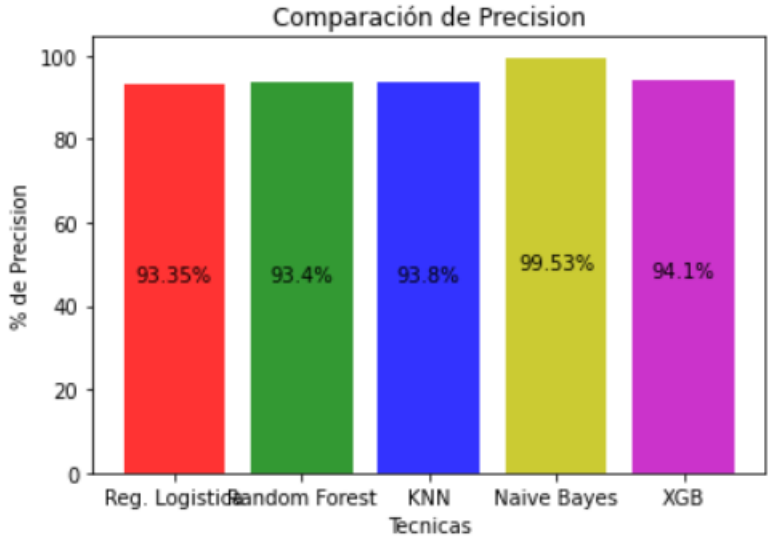
También se llevó a cabo una comparación en términos de precisión entre los modelos. En este aspecto, Naive Bayes sobresalió con un valor de 99.53% para la clase "No obeso". Las demás técnicas obtuvieron resultados bastante similares, como se presenta en la figura 50.

Figura 49
Comparación de Exactitud entre técnicas



Fuente: Elaboración Propia

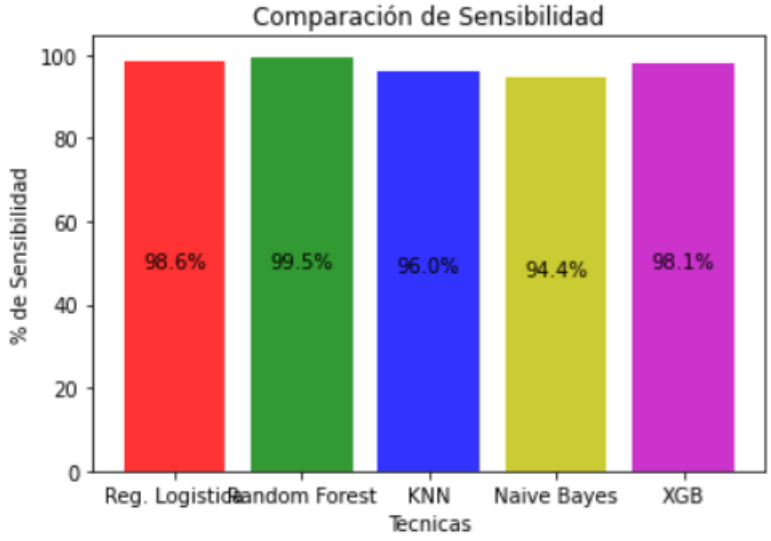
Figura 50
Comparación de Precisión entre técnicas



Fuente: Elaboración Propia

Finalmente, al comparar la sensibilidad de los modelos Naive Bayes obtuvo un 94.4% (ver figura 51).

Figura 51
Comparación de Sensibilidad entre técnicas



Fuente: Elaboración Propia

8. Capítulo 8: Conclusiones y Trabajos Futuros

8.1. Conclusiones

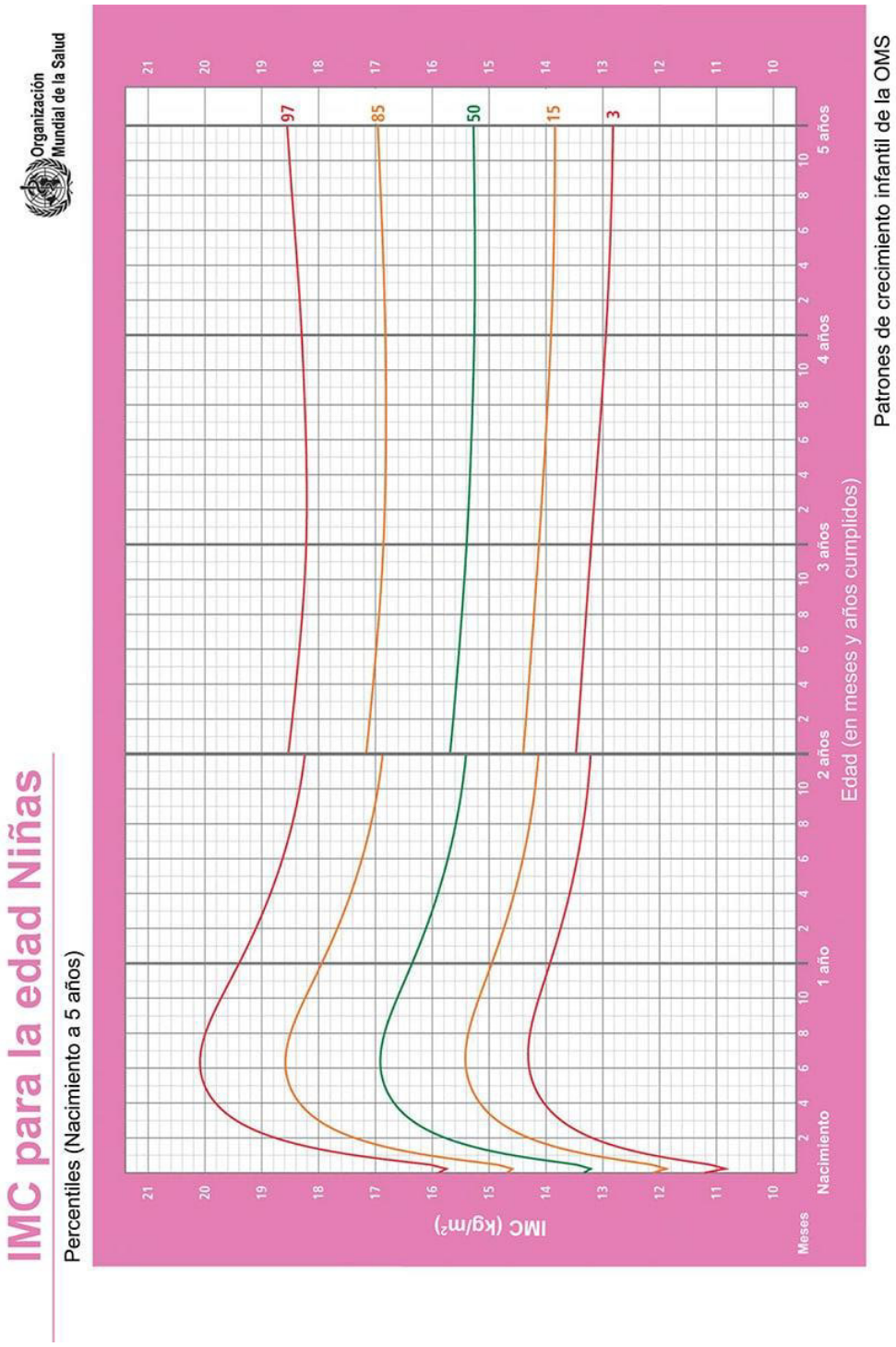
- La presente investigación utilizó 771 registros de historias clínicas extraídas del aplicativo e-Qhali del Ministerio de Salud, este conjunto de datos estuvo constituido por 27 variables medicas de cada paciente, al evaluar estas variables con el modelo predictivo implementado se concluyó que las variables más significativas usadas por el modelo para dar su predicción fueron "edad gestacional al nacer", "peso en la primera consulta de seguimiento" y "requiere hospitalización"
- La literatura existente para predecir la obesidad infantil fue amplia, se revisó 8 casos de existo en los que se utilizaron distintas técnicas para abordar el problema, desde Sistemas Neuro – difusos, modelos de clasificación y modelos híbridos. Entre todos los métodos estudiados, Naive Bayes fue la técnica que seleccionamos debido a su fácil interpretabilidad, y su capacidad para trabajar con datos continuos y discretos.
- Se desarrolló la aplicación web con los módulos necesarios para que el médico se registre, ingrese la información del niño y así pueda obtener el resultado de la predicción del menor (obeso/no obeso) y la probabilidad respectiva.
- Por último, se implementaron y entrenaron en Python las técnicas KNN, Regresión Logística, Random Forest y XGB, para compararlas con los resultados de Naive Bayes, nuestro modelo arrojó una exactitud de 93% y una precisión de 99.5%; siendo superior a las demás técnicas.

8.2. Trabajos Futuros

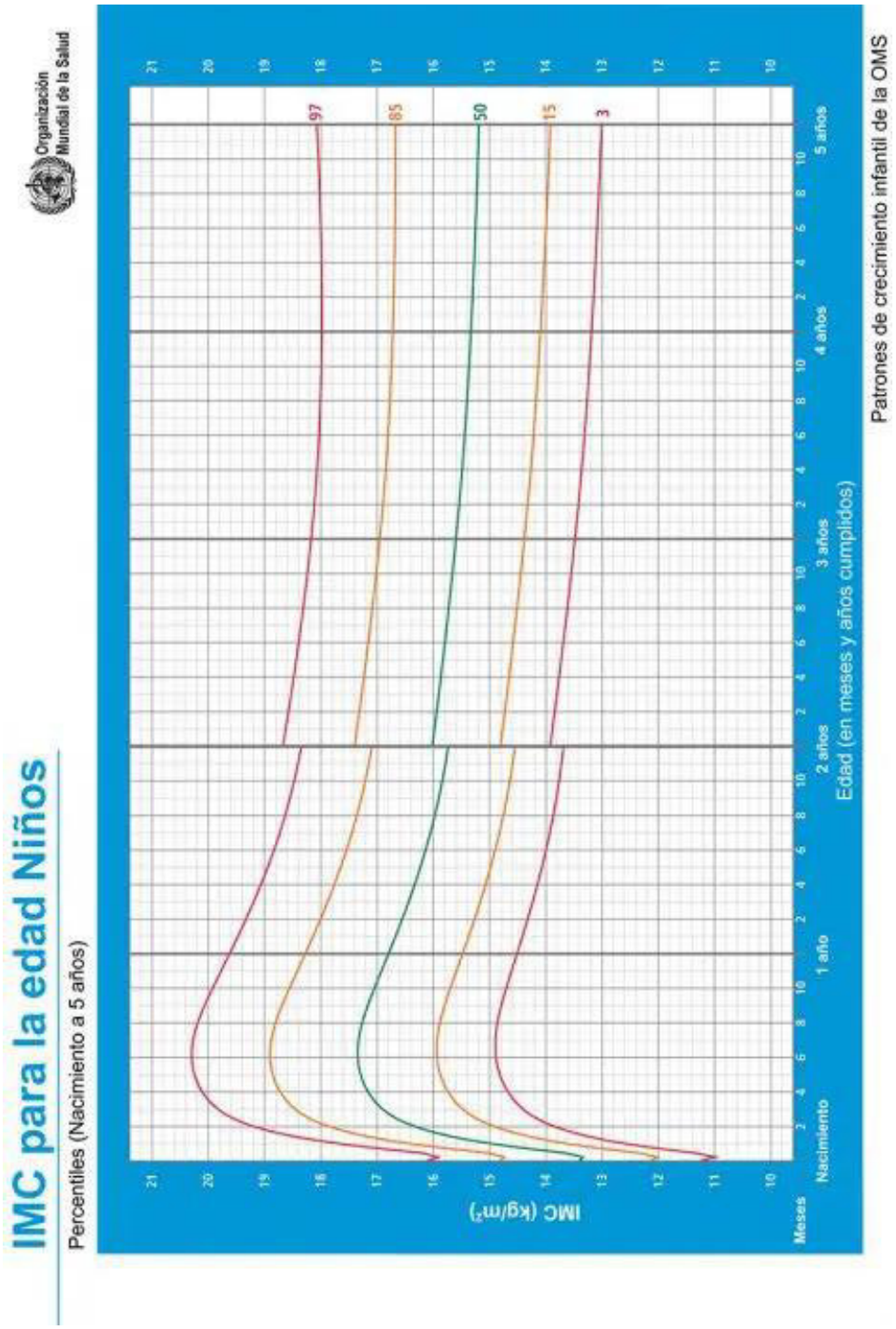
- Añadir el modelo predictivo Naive Bayes como una nueva funcionalidad dentro del Sistema de Nutrición de los Hospitales a nivel nacional.
- Construir un modelo hibrido utilizando Naive Bayes y Algoritmos Genéticos para evaluar si se obtiene una mejor precisión a la ya obtenida, esto guiándonos de Bin Muhamad et al. (2012) que obtuvo un buen resultado para este hibrido con un conjunto de datos distinto.
- Volver a entrenar el modelo con una mayor cantidad de registros, y buscar más información de la historia clínica como por ejemplo los antecedentes de la madre para validar si al incluir estos datos puede mejorar la precisión ya obtenida.

9. Anexos

Anexo A. Cartilla de IMC para niñas de 0 a 5 años



Anexo B .Cartilla de IMC para niños de 0 a 5 años



Anexo C. Código para Implementar Naive Bayes en Python

```
from sklearn.model_selection import train_test_split
from sklearn.model_selection import cross_val_score,cross_validate
from sklearn.naive_bayes import GaussianNB
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.33)
scoring = ['recall', 'precision','accuracy']
clf = GaussianNB()
scores = cross_validate (clf, X_train, y_train, cv=15,scoring=scoring,return_estimator=True)
```

Anexo D. Vista de Variables Definidas

sexo	id_numero_embazaro	id_numero_atenciones_perinatales	div_id_parto_condicion	div_id_parto_lugar	div_id_parto_atendido_por	id_nacimiento_edad_gestacional	id_nacimiento_peso	id_nacimiento_talla	id_nacimiento_perimetro_cefalico	div_id_peso_para_edad_gestacional	id_nacimiento_perimetro_toracico	id_nacimiento_apgar_1	id_nacimiento_apgar_5	div_id_enfermedad_congenita	div_id_cpp	div_id_ac	div_id_nacimiento_hospitalizacion	id_resultado	edad_en_atencion_anios	id_peso_1	id_talla_1	estado_nutricional
girl	2	7	E	ES	PS	40	3250	49	34	A	34	8	9	False	True	True	False	13.4	3.8	14900	97.9	Normal
girl	2	6	E	ES	PS	38	2800	49	34	A	34	9	9	False	True	True	False	11.6	3.6	12800	92.5	Normal
boy	2	8	C	ES	PS	40	3200	49	33	A	32.5	8	9	False	True	True	False	14.5	4	13500	95.9	Normal
boy	1	6	E	ES	PS	40	3320	48	32	A	33	8	9	False	True	True	False		3.9	13350	92.7	Normal
boy	2	8	E	ES	PS	39	3280	50	34.5	A	33	8	9	False	True	True	False	12	3.8	1600	94	Sobrepeso
boy	6	11	E	ES	PS	40	3800	50	35.5	A	36	9	10	False	True	True	False	12.7	3.6	15500	98.8	Normal
boy	3	7	E	ES	PS	37	2600	47.7	32	A	31	8	9	False	True	True	False	13	3.6	14700	93	Normal
boy	5	7	E	ES	PS	36	3700	53	33	G	34	8	9	False	True	True	False	14.6	3.6	16300	96.1	Sobrepeso
girl	4	7	E	ES	PS	38	2800	49.5	33	A	34	8	9	False	True	True	False	13.9	3.7	14400	95.4	Normal
girl	1	5	E	ES	PS	39	3520	49.2	34	A	33.5	8	9	False	False	True	True	13	3.6	16200	95.1	Normal
girl	2	9	E	ES	PS	39	3100	49.5	35	A	34	8	9	False	True	False	False	12.8	3.9	16000	98.3	Normal
boy	4	7	E	ES	PS	40	3440	53.5	35.5	A	34	9	9	False	True	True	False	13.5	3.8	15300	96.7	Normal
boy	5	6	E	ES	PS	37	1990	43	32.92222222	P	32.85283019	7.909090909	9.196428571	False	True	True	True	11	3.5	10940	85.1	Normal
boy	1	11	C	ES	PS	38	4600	53	37	G	33.68181818	9	10	False	True	True	False		4	19000	113	Obesidad
girl	1	9	C	ES	PS	39	4800	55	32	G	32	8	9	False	True	True	False	14.6	3.8	16000	96.7	Normal
girl	1	7.531914894	E	ES	PS	39	2350	45.5	32.92222222	P	32.85283019	7	9	False	True	True	False	13.7	4	15700	97.4	Normal
girl	12	7	E	ES	PS	39	3100	49	33	A	34	8	9	False	True	True	False	13.5	3.6	14000	90.8	Normal
girl	4	7	E	ES	PS	40	3070	50	34	A	34	9	10	False	True	True	False	14	3.6	12800	95.5	Normal
boy	1	6	E	ES	PS	41	3520	51	33	A	32	9	10	False	True	True	False	11.7	3.6	15100	98.2	Normal
boy	2	8	E	ES	PS	38	3490	49.2	36	A	34	8.210696921	9.277597403	False	True	True	False	13.1	3.9	13100	95.4	Normal
boy	2	9	E	D	PS	40	3450	50	35.5	A	35	9	9	False	True	True	False	14.4	3.8	14800	101.5	Normal
boy	2	8	E	ES	PS	39	3500	50.1	34	A	33	8	9	False	True	True	False	12.7	3.6	16500	95	Obesidad
boy	1	8	E	ES	PS	38	2710	44.5	32.5	A	31.7	9	10	False	True	True	False	14.2	3.6	12700	92	Normal

10. Referencias Bibliográficas

Al-Aidaroos, K., Bakar, A., & Othman, Z. (2018). Medical data classification with Naive Bayes approach. *Information Technology Journal*, 1166-1174.

Alarcón, C. (2015). *Optimización del Clasificador "Naive Bayes" usando Árbol de Decision C4.5 [Tesis de maestría]*. Universidad Nacional Mayor de San Marcos.

Alva, L., Laria, J., Ibarra, S., Castán, J., & Terán, J. (2020). A proposed diffuse model to determine overweight and obesity in children and adolescents. *Revista Chilena de Nutrición*, 545-551.

Álvarez-Dongo, D., Sánchez-Abanto, J., Gómez-Guizado, G., & Tarqui-Mamani, C. (2012). Sobrepeso y obesidad: prevalencia y determinantes sociales del exceso de peso en la población peruana. *Revista Peruana de Medicina Experimental y Salud Publica, 29*(3), 303-313. http://www.scielo.org.pe/scielo.php?script=sci_arttext&pid=S2304-51322017000400012

Amat, J. (Agosto de 2016). *Regresión logística simple y múltiple. Ciencia de datos.* https://www.cienciadedatos.net/documentos/27_regresion_logistica_simple_y_multiple

Amat, J. (10 de 2022). *Random Forest python.* https://www.cienciadedatos.net/documentos/py08_random_forest_python.html

Begenova, S., & Avdeenko, T. (2018). Building of fuzzy decision trees using ID3 algorithm.

Bin Muhamad Adnan, M. H., Husain, W., & Abdul Rashid, N. (2012). A hybrid approach using Naïve Bayes and Genetic Algorithm for childhood obesity prediction. *ICCISci*, 281-285.

CENAN. (2019). *Estado Nutricional de Niños Peruanos menores de 5 años 2019*. https://web.ins.gob.pe/sites/default/files/Archivos/cenan/van/sala_nutricional/sala_1/2020/sala_situacional_estado_nutricional_ninos_menores_de_5_anos_sien-his_2019.pdf

Chahuara, J. (2005). *Control Neuro – Difuso aplicado a una Grúa Torre [Tesis de Pregrado].* Universidad Nacional Mayor de San Marcos, Lima.

Chaparro, J., Giraldo, B., & Rondón, S. (2015). Evaluación del clasificador Naïve Bayes como herramienta de diagnóstico en Unidades de Cuidado Intensivo. . *Revista de Tecnología, 12*. https://doi.org/10.18270/rt.v12i2.774

Cigarro, I., Sarqui, C., & Zapata-Lamana, R. (2016). Efectos del sedentarismo y obesidad en el desarrollo psicomotor en niños y niñas: Una revisión de la actualidad latinoamericana. *Universidad y Salud*, 156-169.

Diario La Republica. (17 de 10 de 2019). *4 de cada 10 escolares sufren de sobre peso y obesidad en el Perú*. https://larepublica.pe/sociedad/2019/10/17/minsa-4-de-cada-10-escolares-sufren-de-sobre-peso-y-obesidad-en-el-peru-obesidad-anemia/

Diciembre, S. (2017). *Sistemas de Control con Lógica Difusa: Métodos de Mamdani y de Takagi-Sugeno-Kang (TSK).* http://repositori.uji.es/xmlui/bitstream/handle/10234/173788/TFG_2017_DiciembreSanahuja_Samuel.pdf?sequence=1

ESAN. (9 de Agosto de 2018). *Minería de datos: ¿en qué consiste el knowledge discovery in databases?* https://www.esan.edu.pe/apuntes-empresariales/2018/08/mineria-de-datos-en-que-consiste-el-knowledge-discovery-in-databases/

Ferrero, R. (05 de 2020). *Qué son los árboles de decisión y para qué sirven*. Máxima Formación: https://www.maximaformacion.es/blog-dat/que-son-los-arboles-de-decision-y-para-que-sirven

Ferrero, R., & López, J. (22 de Febrero de 2021). *Qué son los árboles de decisión y para qué sirven*. Máxima Información: https://www.maximaformacion.es/blog-dat/que-son-los-arboles-de-decision-y-para-que-sirven/#id4

Fitbit. (s.f.). https://www.fitbit.com

Garg, S., Singh, A., Sarje, A., & Peddoju, S. K. (2013). Behaviour analysis of machine learning algorithms for detecting P2P botnets. *15th International Conference on Advanced Computing Technologies*, 1-4.

Gonzales, L. (2 de Noviembre de 2018). Introducción a la librería Scikit Learn de Python:

https://aprendeia.com/libreria-scikit-learn-de-python/
Gonzales, L. (20 de Agosto de 2020). *Métodos de Ensamble de Modelos*. https://aprendeia.com/metodos-de-ensamble-de-modelos-machine-learning-ensemble-methods-en-espanol/
Growin. (s.f.). https://www.growin.online/es/
Guija, E., & Guija, H. (2020). *La obesidad como factores de riesgo para COVID-19.* https://medicina.usmp.edu.pe/images/noticias_eventos/2020/Obsidad-covid19.pdf
Hammond, R., Athanasiadou, R., Curado, S., Aphinyanaphongs, Y., Abrams, C., Messito, M., . . . Elbel, B. (2019). Predicting childhood obesity using electronic health records and publicly available data. *PLoS One*.
Heras, J. M. (31 de 05 de 2019). *Ensembles: voting, bagging, boosting, stacking*. https://www.iartificial.net/ensembles-voting-bagging-boosting-stacking/
Herrera, J. F. (11 de 07 de 2022). *Árbol de decisiones: ejemplos de ventajas y pasos a seguir*. LeanConstructionMexi: https://www.leanconstructionmexico.com.mx/post/%C3%A1rbol-de-decisiones-ejemplos-de-ventajas-y-pasos-a-seguir
Huawei. (s.f.). https://consumer.huawei.com/pe/accessories/smart-scale/
IPSUSS. (25 de 03 de 2021). *Niveles de obesidad en escolares aumentan durante la pandemia*. https://www.ipsuss.cl/investigacion/niveles-de-obesidad-en-escolares-aumentan-durante-la-pandemia
Landa, J. (19 de Febrero de 2016). *Tratamiento de Datos*. https://fcojlanda.me/es/ciencia-de-los-datos/kdd-y-mineria-de-datos-espanol/
Langarizadeh, M., & Moghbeli, F. (2016). Applying Naive Bayesian Networks to Disease Prediction: a Systematic Review. *Acta Informática Médica*, 364-369.
Liria, R. (2012). Consecuencias de la obesidad en el niño y el adolescente: un problema que requiere atención. *Revista Peruana de Medicina Experimental y Salud Publica, 29*(3), 357-360. http://www.scielo.org.pe/scielo.php?script=sci_arttext&pid=S1726-46342012000300010&lng=es&tlng=es
Londoño, C., Sánchez, C., Tovar, G., & Barbosa, N. (2015). *Sobrepeso en escolares: prevalencia, factores protectores y de riesgo en Bogotá*. https://tel.archives-ouvertes.fr/tel-01127332/file/2015PA113002_annexe.PDF
Lopez, P. (2017). *Desarrollador de productos basados en Extreme Gradient Boosting (Tesis de maestria).* Universidad Oberta de Cataluña, España.
Maguiña, R. (2010). *Sistemas de inferencia basados en Lógica Borrosa: Fundamentos y caso de estudio.* Universidad Nacional Mayor de San Marcos.
Manrique, H., Aro, P., & y Pinto, M. (2015). Diabetes tipo 2 en niños Serie de casos. *Revisa Medica Herediana*, 5-9.
Mayo Clinic. (5 de Febrero de 2019). *Obesidad infantil - Síntomas Y causas.* https://www.mayoclinic.org/es-es/diseases-conditions/childhood-obesity/symptoms-causes/syc-20354827
Medina, H. (2014). *El sobrepeso – obesidad como factor de riesgo para valores elevados de presión arterial en niños de 5 a 10 años de edad. (Tesis de Segunda Especialización).* Universidad Nacional de Trujillo, Trujillo.
Ministerio de Salud del Perú. (2017). *Norma Técnica de Salud para el Control del Crecimiento y Desarrollo de la niña y el niño menor de 5 años.* https://www.saludarequipa.gob.pe/archivos/cred/NORMATIVA%20CRED.pdf
MINSA - ARGENTINA. (2017). *Alimentación Saludable, Sobrepeso y Obesidad en Argentina*. http://www.msal.gob.ar/images/stories/ryc/graficos/0000001137cnt-2017-09_cuadernillo-obesidad.pdf
MINSA-PERU. (2020). *El 85% de pacientes fallecidos con comorbilidades por Covid-19 padecían obesidad.[Nota de Prensa].* https://www.gob.pe/institucion/minsa/noticias/286005-el-85-5-de-pacientes-fallecidos-con-comorbilidades-por-covid-19-padecian-obesidad
Morlan, L., de Arriba, A., de Francisco, R., Martínez, I., de Francisco, M., Pascual, J., . . . Ferrández, Á.

(2017). Modelo estadístico para la prevención precoz de desarrollo de sobrepeso/obesidad en población infantil. *Boletín de la Sociedad de Pediatría de Aragón, La Rioja y Soria*, 73-80.

NCDRISC. (2016). *Data Visualisations.* Recuperado de http://ncdrisc.org/data-visualisations.html

Novoseltseva, E. (27 de Abril de 2021). *Apiumhub*. https://apiumhub.com/tech-blog-barcelona/data-mining-use-cases/

O´brien, S. (19 de 01 de 2021). *Web Application*. RingCentral: https://www.ringcentral.co.uk/gb/en/blog/definitions/web-application/

Odei, S. (2006). *An Exploration of Classification prediction techniques in data mining: the insurance domain.* Bournemouth University.

Orellana, J. (12 de Noviembre de 2018). *Arboles de decision y Random Forest*. Bookdown.org: https://bookdown.org/content/2031/ensambladores-random-forest-parte-i.html

Organización Mundial de la Salud. (2016). *Establecimiento de áreas de acción prioritarias para la prevencion de la Obesidad Infantil [Archivo PDF].* https://apps.who.int/iris/bitstream/handle/10665/250750/9789243503271-spa.pdf;sequence=1

Organización Mundial de la Salud. (6 de 9 de 2021). *Who.int*. https://www.who.int/es/news-room/fact-sheets/detail/obesity-and-overweight

Orgaz, C. (14 de 05 de 2019). *Los países de América Latina donde más ha crecido la obesidad*. https://www.bbc.com/mundo/noticias-america-latina-48258937

Paulino Flores, L. A. (2021). *Sistema experto probabilístico basado en redes bayesianas para la predicción de riesgo de cáncer cervical.* Tesis de Pregado, Universidad Nacional Mayor de San Marcos.

Pedamkar, P. (14 de 05 de 2019). *Naive Bayes Algorithm*. EDUCBA: https://www.educba.com/naive-bayes-algorithm/

Roman, V. (25 de Abril de 2019). *Algoritmos Naive Bayes: Fundamentos e Implementación [Publicación de blog]*. https://medium.com/datos-y-ciencia/algoritmos-naive-bayes-fudamentos-e-implementaci%C3%B3n-4bcb24b307f

Rossman, H., Shilo, S., Barbash-Hazan, S., Shalom, N., Hadar, E., D. Balicer, R., . . . Segal, E. (2021). Prediction of Childhood Obesity from Nationwide Health Records. *The Journal of Pediatrics*, 132-140.

Senthilingam, M. (14 de 07 de 2017). *Estos son los paises mas obesos del Mundo*. CNN: https://cnnespanol.cnn.com/2017/07/14/estos-son-los-paises-mas-obesos-del-mundo/

Sharma, A., Sharma, T., & Mansotra, V. (2016). Performance Analysis of Data Mining Classification Techniques on Public Health Care Data. *International Journal of Innovative Research in Computer and Communication Engineering, 4*.

Suca, C., Córdova, A., Condori, A., & Cayra, J. (2016). Modelo difuso para la predicción de casos de obesidad empleando el árbol GFID3 generalizado. *Research in Computing Science*, 9-22. https://www.rcs.cic.ipn.mx/rcs/2016_113/Modelo%20difuso%20para%20la%20prediccion%20de%20casos%20de%20obesidad%20empleando%20el%20arbol%20GFID3%20generalizado.pdf

Suca, C., Córdova, A., Condori, A., Cayra, J., & Sulla, J. (2016). Comparación de Algoritmos de Clasificación para la Predicción de Casos de Obesidad Infantil. https://www.researchgate.net/profile/Abel-Condori-Castro/publication/301567339_COMPARACION_DE_ALGORITMOS_DE_CLASIFICACION_PARA_LA_PREDICCION_DE_CASOS_DE_OBESIDAD_INFANTIL/links/571a985c08aee3ddc568f97d/COMPARACION-DE-ALGORITMOS-DE-CLASIFICACION-PARA-LA-PR

Sulla, J., Soto, C., Cardenas, R., Huancco, L., & Alfara, L. (2018). Application of the ANFIS Neuro-Fuzzy model for the classification of obesity in children and adolescents. *16° LACCEI International Multi - Conference for Engineering Education and Technology.* Arequipa. https://www.laccei.org/LACCEI2018-Lima/full_papers/FP53.pdf

Ticona, M. (2018). *Sistema para la predicción de obesidad en la adolescencia utilizando técnicas de minería de datos [Tesis de Bachiller].* Universidad Catolica de Santa María.

Timarán Pereira, S., Hernández Arteaga, I., Caicedo Zambrano, S., Hidalgo Troya, A., & Alvarado Pérez,

J. (2016). *Descubrimiento de patrones de desempeño académico con árboles de decisión en las competencias genéricas de la formación profesional.* https://doi.org/http://dx.doi.org /10.16925/9789587600490

Timarán, S., Hernandez, I., Caicedo, S., Hidalgo, A., & Alvarado, J. (2016). *Descubrimiento de Patrones de Desempeño Academico.* https://doi.org/http://dx.doi.org/10.16925/9789587600490

Umoh, U., & Isong, E. (2015). Design Methodology of Fuzzy Expert System for the Diagnosis and Control of Obesity. *Computer Engineering and Intelligent Systems*.

Yi-lai, C., Tao, W., Ben-sheng, W., & Zhou-jun, L. (2009). A Survey of Fuzzy Decision Tree Classifier. *1*(2), 149-159. https://doi.org/10.1007/s12543-009-0012-2