Modelo de Analítica para la Predicción del Riesgo de Cáncer de estómago en el Departamento del Cauca.

PROYECTO DE GRADO

Jorge Luis Pérez Narváez

Asesor
Javier Gustavo Diaz Cely, PhD
Director Maestría en Ciencia de Datos
Facultad de Ingeniería
Universidad Icesi

FACULTAD DE INGENIERÍA
MAESTRÍA EN GESTIÓN DE INFORMÁTICA Y TELECOMUNICACIONES
SANTIAGO DE CALI
2019

Modelo de Analítica para la Predicción del Riesgo a Cáncer de estómago en el Departamento del Cauca.

Jorge Luis Pérez Narváez

Trabajo de grado para optar al título de
Máster en Gestión de Informática y Telecomunicaciones

Asesor
Javier Gustavo Diaz Cely, PhD
Director Maestría en Ciencia de Datos
Facultad de Ingeniería
Universidad Icesi

FACULTAD DE INGENIERÍA
MAESTRÍA EN GESTIÓN DE INFORMÁTICA Y TELECOMUNICACIONES
SANTIAGO DE CALI
2019

CONTENIDO

RESUMEN 8

1. INTRODUCCIÓN 10
- Contexto y Antecedentes 10
- Planteamiento del Problema 12
- Objetivo General 15
- Objetivos Específicos 15
- Organización del Documento 15

2. ANTECEDENTES 17
- Marco Teórico 17
- Estado del Arte 21
  - Trabajos relacionados 23

3. MARCO INTERPRETATIVO Y METODOLÓGICO DE LA INVESTIGACIÓN 28

4. COMPRENSIÓN DEL NEGOCIO Y DE LA INFORMACIÓN 33
- Metodología 33
  - Fase I Comprensión del Negocio 33
  - Fase II Comprensión de los Datos 34
- Resultados 35
  - Fase I Comprensión del Negocio 35
  - Fase II Comprensión de los Datos 42

5. PREPARACIÓN DE LOS DATOS Y MODELADO 52
- Metodología 52
  - Fase III Preparación de los Datos 52
  - Fase IV Modelado 54
- Resultados 55
  - Fase III Preparación de los Datos 55
  - Fase IV Modelado 63

6. EVALUACION Y VALIDACIÓN 79
- Resultados 75
  - Fase V Evaluación del modelo 75

CONCLUSIONES Y FUTURO TRABAJO 83

BIBLIOGRAFÍA 85

ANEXOS 88

LISTA DE TABLAS

Tabla 1 Bases de datos consultadas 24
Tabla 2 Plan del proyecto 41
Tabla 3 Categorías de variables 44
Tabla 4 Variables anonimizadas 47
Tabla 5 Distribución de variables e instancias por Datasets 49
Tabla 6 Variables excluidas 56
Tabla 7 Identificadores únicos de variables 58
Tabla 8 Integración de Datasets Hogar-Adulto 61
Tabla 9 Integración de Datasets Hogar-Niño 62
Tabla 10 Algoritmos Priorizados 70
Tabla 11 Resultados de Hiperparámetros 73
Tabla 12 Matriz de confusión - Dataset Hogar-Adulto 76
Tabla 13 Matriz de confusión - Dataset Hogar-Niño 76

LISTA DE FIGURAS

Figura 1 Ciclo de vida metodología CRISP-DM 30
Figura 2 Fases de la Metodología CRISP-DM 32
Figura 3 Herramientas y Técnicas..... 42
Figura 4 Esquema Racional de los Dataset 48
Figura 5 Exploracion de datos..... 50
Figura 6 Codificación de características categóricas 60
Figura 7 Métricas de Evaluación 65
Figura 8 Ajuste de parámetros para el Modelo SVC 72

LISTA DE ANEXOS

Anexo 1 Glosario de Terminología de Minería de Datos

Anexo 2 Diccionario de Datos

Anexo 3 Colección de la Información

Anexo 4 Comandos de Programación

RESUMEN

El cáncer gástrico es uno de los más comunes a nivel mundial ocupando el quinto lugar de las neoplasias malignas. Adicionalmente, en el departamento del Cauca, representa la primera causa de muerte por cáncer y la cuarta a nivel general. La etiología del cáncer gástrico no es muy clara hasta ahora, pero se reconocen como factores de riesgo la infección por Helicobacter pylori, además de otros agentes relacionados como factores dietéticos asociados al consumo de alimentos con alto contenido de sal, consumo de alcohol, tabaquismo, obesidad, factores genéticos, entre otros. Aunque en la actualidad el sector salud sea uno de los principales generadores de información, con características diferentes a las tradicionales (semiestructurada o no estructurada, geo localizada, etc.) nos encontramos con que existen pocos estudios acerca del uso y aprovechamiento que las ciencias de la computación pueden aportar sobre los datos con el fin de predecir el riesgo de enfermedades de alto costo prevalentes en Colombia, como lo es el cáncer de estómago. En este contexto, se hace necesario que la tecnología contribuya al mejoramiento de la gestión de la información para alcanzar una mayor eficiencia y transparencia, facilitar la administración y el control de los recursos, y brindar información objetiva y oportuna para la toma de decisiones.

Es así, que el objetivo de este estudio fue construir un modelo de analítica para la predicción del riesgo a Cáncer de estómago en una población del Departamento del Cauca. Se utilizó una base de datos de 13741 registros asociados a más de 90 variables, organizados en tres sets de datos (Hogares, Adultos y Niños).

Para alcanzar los objetivos planteados se siguió la metodología CRISP-DM, la cual proporciona una descripción normalizada del ciclo de vida de un proyecto estándar de análisis de datos. Este proceso fue abordado en seis fases: Comprensión del negocio; Comprensión de los datos; Preparación de datos; Modelado; Evaluación y Validación. La construcción del modelo de analítica basado en los algoritmos *RandomForest* y *Gradient Boosting* permitió predecir

correctamente el 99,9% de Falsos negativos de los Datasets de pruebas, así como también logró identificar las variables con mayor importancia al riesgo de desarrollar cáncer de gástrico por la infección H.pylori como principal agente etiológico de la enfermedad.

Capítulo 1

Introducción

1.1. Contexto y Antecedentes

La información es el elemento básico principal en el proceso de adquisición, generación, gestión y transmisión del conocimiento (Díaz Pérez, de Liz Contreras, & Amador, 2009). Hoy en día hay la cantidad excesiva de información necesita ser estudiada, analizada y depurada para convertirla en el conocimiento indispensable y necesario para la búsqueda de soluciones reales en la toma de decisiones en los diferentes campos. Es así, que los sistemas de información se han convertido en pieza importante en el almacenamiento de la información, en donde la medición de la calidad de los datos almacenados cada vez es más rigurosa (Barragán Ocaña, 2009).

Desde la década de los 80, la llamada Gestión de Información, Gerencia de Información o Information Management (en inglés) ha ganado un espacio importante en la vida cotidiana en el marco de las instituciones en general y en particular en aquellas que tienen como misión el desarrollo de servicios y productos de información (López Noreña, 2010). Sin embargo, aunque existen diferentes modelos para la gestión del ciclo de la información, nos encontramos que aún persisten grandes volúmenes de datos que aunque permiten contar con un conjunto básico de indicadores y llevar a cabo ciertas acciones de mejoramiento, no presentan la información que permita su análisis integral que ayude a correlacionar variables que evidencien situaciones y agreguen valor a la gestión y la toma de decisiones.

Si bien es cierto, que mediante consultas simples sobre los datos se pueden obtener algunos resultados, a medida que crece la complejidad de las bases de datos y el número de registros, los resultados son cada vez más difíciles de ser interpretados y utilizados. Bajo esta mirada, nace la minería de datos (MD) que es la ciencia que estudia patrones en grandes bases de datos y emplea técnicas de la inteligencia artificial, la estadística o el aprendizaje automático para extraer dicha información y traducirla a resultados interpretables que permitan obtener relaciones existentes entre los mismos y dar beneficios de algún modo al negocio.

La MD es un intento de buscarle sentido a la explosión de información que actualmente puede ser almacenada. Así pues, para comenzar el proceso de minería de datos es importante partir de una base de datos o data warehouse (almacén de datos) que contenga la información que se quiere analizar, correctamente estructurada. La MD trata de sacar toda la información posible de los almacenes de datos, no se conforma sólo con la visualización de estos datos como podría pasar con las consultas simples, sino que trata de obtener resultados en cuanto a la relación que existe entre los mismos y cómo podrían dar beneficios de algún modo al negocio(Mitra & Acharya, 2005). Adicionalmente, la capacidad de la información comprende no sólo su acceso; ésta también comprende la conciencia de su existencia y las habilidades para explotarla bajo un proceso transversal e indispensable de medir la calidad de la información. Es en este ámbito, el Machine-learning (ML) ofrece un enfoque alternativo al modelado de predicción estándar que puede abordar las limitaciones actuales (Dietterich, 1998).

En los últimos años, la Minería de datos y el Machine Learning han alcanzado un auge importante en la gestión de la información, esto último, gracias a que el primero tiene una función exploratoria mientras que el segundo se focaliza la predicción (Bishop, 2006). Su fin es explorar y analizar las bases de datos disponibles para ayudar a la toma de decisiones. Permiten a su vez, la extracción

de la información existente en textos, así como crear sistemas inteligentes capaces de entenderlos. En este proceso se barren las bases de datos y se identifican modelos previamente escondidos, se lleva a cabo la reducción, transformación y clasificación de las bases de datos, que conlleve a una valoración de la información y finalmente a la conformación de la base del conocimiento que permita tomar decisiones en base a la información clasificada (Mitra & Acharya, 2005).

En la actualidad, el sector salud es uno de los principales generadores de información, con características diferentes a las tradicionales: es semiestructurada o no estructurada, geo localizada, se genera a altísimas velocidades y de forma masiva, y es naturalmente compleja. En este contexto, se hace necesario que la tecnología contribuya al mejoramiento de la gestión de la información para alcanzar una mayor eficiencia y transparencia, facilitar la administración y el control de los recursos, y brindar información objetiva y oportuna para la toma de decisiones. En la actualidad son muchas las áreas del sector salud en las que ha incursionado la minería de datos, un ejemplo de ello son aplicaciones complejas como el tratamiento, procesamiento y reconocimiento de imágenes en el cerebro o problemas del corazón (Vega, Rosano, López, Cendejas, & Ferreira, 2012). Sin embargo, existen pocos estudios acerca del uso de Ciencias de la Computación en la predicción del riesgo de enfermedades de alto costo prevalentes en Colombia, como lo es el cáncer de estómago.

Por lo anterior, se hace necesario desarrollar herramientas que permitan la integración de la información para mejorar el análisis e interpretación del riesgo en salud en la medida en que contribuyan a facilitar la toma de decisiones.

1.2. Planteamiento del Problema

En América Latina encontramos una serie de problemas en los datos y la información, especialmente sobre las condiciones de vida y eventos del proceso

Salud - Enfermedad que dificultan su integración y utilización. Cabe mencionar que las necesidades de información en salud cambian a lo largo del tiempo en relación a cambios en la situación social del espacio considerado. En las unidades de información se puede observar claramente la calidad heterogénea de los datos recolectados; dicha heterogeneidad se debe a diversas realidades locales con diferentes estructuras socioeconómicas, capacidades organizativas, capacidades administrativas y de prestación de servicios.

En el sector salud, la gestión de la información no está sólo en compartir los recursos tecnológicos, sino también en la apropiación de un modelo que permita interpretar la interacción entre información epidemiológica y la distribución geográfica y que además contribuya al empoderamiento social, el mejoramiento de la gestión del riesgo y a orientar la toma de decisiones en salud pública basado en la gestión del conocimiento (Martínez Moreno, 2016). Por lo anterior, se deben explorar enfoques que incorporen mejor los factores de riesgo múltiples y determinen relaciones más matizadas entre los factores de riesgo y los resultados.

Uno de los problemas de salud pública es la gestión del riesgo a partir de la información masiva de datos para disminuir el impacto de enfermedades de alto costo como es el cáncer, especialmente el cáncer de estómago. El cáncer gástrico (CG) es la quinta neoplasia maligna más común en el mundo y la segunda causa de muerte por cáncer anualmente, totalizando más de un millón de muertes por año, siendo el adenocarcinoma del estómago el tumor más frecuente (95%) (Jemal et al., 2011).

Colombia está entre los países de América Central y del Sur con mayores tasas en incidencia y mortalidad de cáncer gástrico (Brenner, Rothenbacher, & Arndt, 2009). El departamento del Cauca tiene la más alta incidencia de cáncer gástrico en Colombia, con una tasa anual de 42.5/100.000 habitantes para los hombres y 28.6/100.000 habitantes para las mujeres (Jemal et al., 2011). Adicionalmente se

diagnostica el cáncer gástrico en estados avanzados de la enfermedad, en aproximadamente el 93% de los casos, lo cual produce un impacto negativo en el pronóstico de vida de los pacientes a 5 años. Esta neoplasia es una de las pocas neoplasias malignas para la cual se ha establecido que agentes infecciosos como *Helicobacter pylori (H.pylori)* tienen un reconocido e importante rol etiológico. La infección por esta bacteria es la más común en el ser humano, afectando el 60% de la población en países desarrollados y 80% en países en vías de desarrollo. La infección por esta bacteria desempeña un papel importante en la génesis de la gastritis, úlcera péptica duodenal, úlcera péptica gástrica, cáncer gástrico y linfoma tipo MALT (Asaka, Takeda, Sugiyama, & Kato, 1997).

En relación al rol que puede jugar el *H. pylor*i dentro de la multifactoriedad etiopatogénica del cáncer gástrico, se han publicado muchas evidencias epidemiológicas (Parsonnet et al., 1991), por lo tanto, desde este punto de vista, no es difícil establecer conclusiones con sustento estadístico. Lo que generalmente es difícil explicar satisfactoriamente, es las relaciones entre el vínculo de *H. pylori*, cáncer gástrico y la epidemiología.

Estudios epidemiológicos han demostrado que además de la infección, existen otros factores de riesgo asociados, como lo son factores sociodemográficos, dieta, consumo de alcohol, saneamiento insuficiente, entre otros, Sin embargo, el comportamiento de la enfermedad (p.ej. prevalencia e incidencia) varía según la distribución geográfica de los factores de riesgo en la población. Es decir, el cáncer de estómago continúa siendo un reto en términos de coordinación de las acciones de salud pública, pues se desconoce con precisión cómo se distribuyen y comportan estos factores en las diversas poblaciones. Muchos estudios epidemiológicos han evaluado los factores de riesgo de la incidencia de cáncer gástrico, sin embargo solo unos pocos estudios han desarrollado modelos de predicción para el el desarrollo de esta patología.

1.3. Objetivo General

Construir un modelo de Analítica para la predicción del Riesgo a Cáncer de estómago en una población del Departamento del Cauca

1.4. Objetivos Específicos

1) Establecer un proceso de gestión analítica basado en la comprensión del negocio y de la información para la orientación del análisis de datos.

2) Generar un modelo de analítica a partir de una estructura de datos mediante los recursos de Información para la predicción del riesgo a desarrollar cáncer de estómago.

3) Validar e implementar el modelo de analítica construido para la predicción del riesgo a cáncer de estómago en la población de estudio.

1.5. Organización del Documento

Este documento está estructurado en siete capítulos. Los primeros dos corresponden al contexto de investigación, problema, estado del arte y la teoría detrás de la minería de datos repasando algunos conceptos básicos y conocimientos previos. En el tercer capítulo se aborda el marco interpretativo general del estudio, para ello se introduce la metodología de minería de datos CRISP-DM, todo desde el punto de vista teórico listando y resumiendo cada una de sus fases. Los tres siguientes capítulos (cuatro, cinco y seis) corresponden al desarrollo de cada uno de los objetivos específicos respectivamente: *i)* Establecer un proceso de gestión analítica basado en la comprensión del negocio y de la información para la orientación del análisis de datos. *ii)* Generar un modelo de analítica a partir de una estructura de datos mediante los recursos de Información para la predicción del riesgo a desarrollar cáncer de estómago. *iii)* Validar e

implementar el modelo de analítica construido para la predicción del riesgo a cáncer de estómago en la población de estudio. En cada objetivo se desarrolla la metodología específica propuesta así como la presentación de los resultados asociados. Por último, el capítulo siete resume presentan las conclusiones del proyecto y futuras recomendaciones.

La metodología transversal del documento está basada estrictamente en cada una de las distintas fases de la metodología CRISP-DM, aplicada sobre el análisis de los datos almacenados en el marco de un proyecto relacionado con la infección por H.pylori como agente principal del riesgo a desarrollar cáncer de estómago en 8 municipios del departamento del Cauca. A partir de dicho análisis, se pretende encontrar el mejor modelo de analítica para la predicción del riesgo cáncer de estómago que permita dar respuestas a esta patología tan compleja.

Capítulo 2

Antecedentes

2.1. Marco Teórico

Minería de Datos

En muchos dominios, es común encontrar que el análisis de la información aún corresponde a un proceso manual soportado en técnicas estadísticas que proporcionan resúmenes y generan informes, sin embargo, tal enfoque ha cambiado como consecuencia de la necesidad de encontrar valor en los datos(Mitra & Acharya, 2005). La minería de datos (MD) deriva entonces, en un intento de buscarle sentido a la explosión de información que actualmente puede ser almacenada.

En primer lugar, cabe situar la minería de datos a partir de algunas definiciones que se han dado sobre la misma:

Minería de datos: "e*l proceso de análisis secundario de grandes bases de datos apunta en la búsqueda de relaciones insospechadas que son de interés o valor*"(Dua & Du, 2011); "P*roceso iterativo de extracción de patrones predictivos ocultos de grandes bases de datos, utilizando tecnologías de Inteligencia Artificial y técnicas de estadística".*(Bishop, 2006)

Por lo anterior, podemos referirnos a Data mining o minería de datos como el proceso de extracción de información potencialmente útil, implícita y previamente conocida a partir de los datos, donde la idea es construir sistemas capaces de examinar cuidadosamente en las bases de datos, buscando irregularidades o

patrones que de otra forma resultaría casi imposible de identificar (Jaramillo & Paz, 2015).

Existen diversas técnicas y métodos de minería de datos que permiten resolver diversas tareas:

Tareas

Una tarea se puede definir como un tipo de problema a ser resuelto por un algoritmo de minería de datos. Por lo tanto esto implica que cada tarea tiene sus propios requisitos y que la información que se obtiene empleando una tarea en concreto puede ser muy distinta a la obtenida si se emplea otra tarea diferente. Podemos dividir las tareas en dos tipos, predictivas o descriptivas. En las predictivas el objetivo es estimar valores futuros o desconocidos de algunas variables de interés a partir de otras variables independientes (variables predictivas). En el caso de las tareas descriptivas el objetivo es identificar patrones en los datos que los explican o resumen. Las tareas más importantes de la minería de datos para cada uno de los dos tipos anteriores son:

Predictivas Clasificación o discriminación (en estadística): La clasificación asume que hay un conjunto de objetos caracterizados por algún atributo o rasgo que pertenece a distintas clases. La etiqueta de clase es un valor discreto y es conocido para cada objeto. El objetivo de esta tarea es asignar la etiqueta de clase correcta a objetos nuevos y sin etiqueta dados los valores de sus atributos.

Estimación de probabilidad de clasificación: Es una generalización de la clasificación suave. El problema a resolver es el mismo que para la clasificación y clasificación suave. La diferencia está en que en esta tarea el resultado es un conjunto de probabilidades de que el objeto pertenezca a una clase u otra. Por ejemplo, si se quisiera clasificar entre varios medicamentos cuál es el mejor para una determinada patología, esta tarea proporcionaría la probabilidad de que sea cada uno de los medicamentos escogidos

Categorización: En esta tarea a diferencia de las clasificaciones donde a cada objeto le corresponde una y sólo una clase, a un objeto le puede corresponder n clases. Por ejemplo, dado un conjunto de documentos, asignar categorías de los temas que trata cada documento.

Regresión: Esta tarea es muy parecida a la clasificación ya que a cada elemento se le asigna únicamente un valor de salida, con la diferencia de que este valor de salida es un valor numérico, es decir, puede ser un valor entero o real. Un ejemplo muy sencillo sería estimar las ventas de un determinado producto para un determinado año.

Las técnicas de minería de datos y el machine learning han surgido a partir de sistemas de aprendizaje inductivo en computadoras, siendo los datos sobre los que se realiza la búsqueda de nuevo conocimiento, la principal diferencia entre ellos. En el caso tradicional de aprendizaje en computadoras (machine learning), se usa un conjunto de datos “pequeño” y cuidadosamente seleccionado para entrenar al sistema. Por el contrario, en la minería de datos se parte de una base de datos generalmente grande, en la que los datos han sido generados y almacenados para propósitos diferentes del aprendizaje con los mismos (A. L. Buczak & E. Guven, 2016).

Machine Learning

Esta subespecialidad de la informática se engloba en el campo de la inteligencia artificial y está relacionada con el diseño y desarrollo de algoritmos que permiten a las computadoras promover la acción en base a datos empíricos (Moine, Haedo, & Gordillo, n.d.). Su objetivo es aprender a reconocer automáticamente patrones complejos y tomar decisiones inteligentes basadas en datos.

Las técnicas de análisis de datos tienen en el machine learning un sólido apoyo para la generación de conocimiento. Si bien es cierto que estas técnicas explotan

la estadística y muchas otras áreas de las matemáticas, su ventaja no sólo radica en la velocidad y potencia de procesamiento tanto secuencial como paralela, sino en el potencial del aprendizaje automático para complementar otras técnicas de análisis de datos más tradicionales en la medida en que tienen la capacidad de volver a aprender a partir de la representación de los datos (Dua & Du, 2011).

En este punto es importante introducir la analítica como el uso de matematicas y estadistica para derivar el verdadero significado de los datos a fin de contribuir en el proceso de toma de decisiones sólidas basadas en los datos. Existen tres tipos de analítica: i) Analítica descriptiva, datos que expresan el *Qué* ha sucedido en el pasado pero no el *Porqué* o el *Qué* podría cambiar. ii) Analítica predictiva, abarca el *Qué* puede suceder a partir de interpretar datos históricos para modelar futuros resultados. iii) Analítica prescriptiva, el nivel más alto, en donde se identifica el *Qué* se debe hacer (Sebastiani, 2002).

Cabe resaltar el papel y nivel de compromiso que desempeñan los tres agentes principales (data, tipo de analítica y experto) en cualquier proceso de analítica. En el primero de tipo descriptivo, el proceso se basa en los datos y llega hasta el punto de representar el *Qué* ha sucedido y es el experto quien interpreta los datos para llegar a una acción de toma de decisiones (A. L. Buczak & E. Guven, 2016). En el tipo predictivo, los resultados expresan el *Qué* sucederá, con lo cual el experto tiene nuevas y mejoradas bases para una acción de toma de decisiones, pero es sólo hasta el último tipo de analítica, la analítica prescriptiva en la que los resultados del análisis interpretan el *Qué* se debe hacer, automatizando en un punto el proceso de toma de decisiones dejando sólo al experto el nivel de acción en el mismo.

Existen dos tipos principales de aprendizaje automático, el supervisado y el no supervisado (Zhu, Ghahramani, & Lafferty, n.d.). Como ejemplo del supervisado, podría ser el análisis de sentimientos en redes sociales que se basa en clasificar

las opiniones en favorables o desfavorables (puede haber más de dos tipos). Se parte de unas normas u órdenes que se han catalogado anteriormente a mano, para que el programa pueda generalizar a partir de ellos. Esto requiere un tratamiento del lenguaje natural y la posibilidad de aplicar cálculos estadísticos como la obtención de frecuencias y la elaboración de gráficos. En el caso del no supervisado, podemos pensar en la segmentación de clientes, agrupando los que tienen características comunes entre sí(Caruana & Niculescu-Mizil, 2006). De nuevo se utiliza la estadística para obtener esa medición de cercanía o distancia entre ellos, pero sin utilizar a una clasificación previa realizada por una persona.

Un último punto a tener en cuenta es la anonimización de la información sensible, aquella información personal privada de un individuo. En el contexto del análisis de la información mediante técnicas de machine learning, entre otros en los que se manejan cantidades ingentes de datos, la virtud del anonimato es clave. No solo es deseable, sino que debe ser obligatoria: no se pueden asociar los datos recogidos con personas físicas o jurídicas, de ahí que los datos anonimizados cobren mayor importancia.

2.2. Estado del Arte

Hoy en día, los datos no sólo no están restringidos a tuplas representadas únicamente con números o caracteres sino que además el avance tecnológico para la gestión de bases de datos ha hecho posible integrar diferentes estructuras y tipos de datos, tales como imagen, video, texto, y otros datos numéricos(Dua & Du, 2011). El campo de la Salud es un ejemplo de los generadores de información de este escenario, con historias clínicas, resultados, prescripciones, detalles en imágenes, atenciones etc. información que es potencialmente importante pero aún no ha sido descubierta ni articulada. En este punto y como un mecanismo para el análisis de los datos aparece el Machine-learning, que ofrece un enfoque alternativo al modelado de predicción estándar permitiendo abordar las

limitaciones actuales. El Machine-learning tiene potencial para transformar la medicina explotando mejor los "grandes datos" para el desarrollo de algoritmos a partir del estudio del reconocimiento de patrones y el aprendizaje computacional o inteligencia artificial. Esto se basa en una computadora para aprender todas las interacciones complejas y no lineales entre las variables al minimizar el error entre los resultados previstos y observados, además de mejorar potencialmente la predicción. El machine-learning puede identificar variables latentes, que son poco probable que se observen, pero podrían inferirse a partir de otras variables. Hasta la fecha, existen pocas investigaciones a gran escala que apliquen el aprendizaje automático para la evaluación pronóstica en la población general, utilizando datos clínicos de rutina.

A nivel mundial, el Sistema de Información en Salud (SIS), es un mecanismo de colecta, procesamiento, análisis y transmisión de la información necesaria para organizar y operar los servicios de salud. Lo anterior, junto con el gran desarrollo de la informática en los últimos años, ha permitido trabajar con volúmenes muy grandes de datos de información, para ser transmitidos sin dificultades.

En Estados Unidos se están adoptando rápidamente historias clínicas electrónicas, lo que ha aumentado drásticamente la cantidad de datos clínicos disponibles digitalmente pero sin generar resultados concretos de diagnósticos específicos. Una forma de abordar este problema ha sido la utilización de modelos predictivos mediante el uso de inteligencia artificial, lo que ha generado gran interés en los últimos años especialmente para el cuidado de la salud. Dichos modelos implementan sistemas basados en reglas o técnicas de regresión para determinar los factores de riesgo que pueden analizar los miembros o grupos poblacionales con el fin de facilitar la toma de decisiones acerca de qué cambios se pueden hacer para mejorar el nivel de atención médica que reciben las personas.

En Colombia, el uso de sistemas predictivos ha sido implementado principalmente en ciencias de la tierra (ej. Sistemas de información geográfica) pero su uso en las ciencias de la salud aún es escaso (Arce, n.d.). Hoy en día, dadas las poderosas herramientas analíticas con las que cuentan los sistemas expertos ej. Lógica difusa, algoritmos genéticos, etc, para evaluar diversos escenarios de decisión y validar ciertos parámetros, es posible abordar integralmente problemas prioritarios en salud como el cáncer.

Estas herramientas han permitido establecer con mayor precisión la relación que existe entre ciertos factores de riesgo de desarrollar una enfermedad, incluyendo dimensiones de la triada epidemiológica: persona (comunidad) – agente etiológico (exposición) – ambiente (locación). Dada la compleja estructura de los datos que conforma la triada, la aplicación de técnicas innovadoras de minería de datos y descubrimiento de conocimiento, permiten revelar relaciones y tendencias que son generalmente difíciles de descubrir en un formato tabular utilizado métodos de análisis estadístico tradicionales como la Regresión logística.

2.2.1. Trabajos relacionados

Las Ciencias de la Computación en la predicción del riesgo a cáncer de estómago

A medida que un área de investigación madura, a menudo hay un fuerte aumento en el número de informes y resultados disponibles, lo que hace importante resumir y proporcionar una visión general. Una tendencia general hacia la investigación basada en la evidencia, ha llevado a un mayor enfoque con nuevos métodos de investigación empíricos y sistemáticos.

La búsqueda de la información acerca de las contribuciones de las ciencias de la Computación en la predicción del riesgo de a cáncer de estómago fue desarrollada mediante la metodología de mapeo sistemático. Se realizó una cadena de

búsqueda de la siguiente manera: ("prediction model" OR "risk prediction model" OR "risk stratification" OR "risk assessment tool") AND (“classification” OR “algorithm” OR "data analysis" OR "big data" OR "machine learning" OR "artificial intelligence").

Debido a las restricciones de longitud de las bases de datos de Science Direct y Scopus (fue utilizado ("gastric cancer" OR "preneoplastic gastric lesions" OR "Helicobacter pylori") AND ("prediction model") AND (“algorithm” OR "data analysis" OR "machine learning").

Los campos de búsquedas incluidos fueron: Título, Abstract y Palabras claves.

I. Bases de datos consultadas

Tabla 1 Bases de datos consultadas

Base de datos	Artículos encontrados
PubMed	28
Science Direct	2
Scopus	11
IEEE Xplore	1

Los criterios de inclusión que se tuvieron en cuenta fueron: Artículos de revista o de conferencia, capítulos de libros y similares que:

- Consideren la predicción de riesgo de neoplásicas o predicción de riesgo de cáncer gástrico (enfoque preventivo).
- Utilicen para la predicción alguna herramienta, técnica o método estadístico o de las ciencias de la computación.

II. Resultados de la búsqueda

Se leyeron los títulos y abstracts de todos los artículos encontrados (42), después se seleccionaron aquellos que cumplen con los criterios de inclusión, y finalmente, se examinaron de forma completa estos últimos. A continuación se presenta una descripción corta de todos los artículos que cumplen con los criterios de inclusión (2), junto a las brechas de investigación identificadas.

- **Prediction Model for Gastric Cancer Incidence in Korean Population, 2015.**

El objetivo de este estudio fue desarrollar un modelo de predicción para la incidencia de cáncer gástrico basado en una gran cohorte poblacional en Corea(Eom et al., 2015). Con base en los datos de la Corporación Nacional de Seguros de Salud, se analizaron 10 factores de riesgo principales para el cáncer gástrico. Se utilizó el modelo de riesgos proporcionales de Cox para desarrollar modelos de predicción específicos para cada género. El rendimiento del modelo fue evaluado utilizando una cohorte independiente en términos de discriminación y calibración. La capacidad de discriminación se evaluó utilizando la C-Statistic de Harrell, y la calibración se evaluó utilizando gráficos de calibración y de pendiente.
Durante una mediana de seguimiento de 11,4 años, se observaron 19.465 (1,4%) y 5.579 (0,7%) casos de cáncer gástrico de nuevo desarrollo entre 1.372.424 hombres y 804.077 mujeres, respectivamente. Los modelos de predicción incluyeron la edad, el IMC, el historial familiar, la regularidad de las comidas, la preferencia de sal, el consumo de alcohol, el tabaquismo y la actividad física para los hombres; y la edad, el IMC, el historial de la familia, la preferencia de sal, el consumo de alcohol y el tabaquismo para las mujeres. Este modelo de predicción mostró una buena precisión y previsibilidad tanto en las cohortes de desarrollo como en las de validación (C-statistics: 0,764 para hombres, 0,706 para mujeres).

- **Development and validation of a risk assessment tool for gastric cancer in a general Japanese population, 2018.**

El propósito de este estudio fue desarrollar y validar una herramienta para conocer el riesgo de desarrollar cáncer gástrico en la población japonesa(Iida et al., 2018). Se realizó seguimiento a un total de 2444 personas de 40 años o más durante 14 años a partir de 1988 (cohorte de derivación), y 3204 personas del mismo grupo de edad fueron seguidos durante 5 años a partir de 2002 (cohorte de validación). La ponderación (puntuación de riesgo) de cada factor de riesgo se determinó en función de los coeficientes de un modelo de riesgos proporcionales de Cox en la cohorte de derivación. La herramienta fue evaluada utilizando la C-Statistic y la prueba de Hosmer-Lemeshow en la cohorte de validación. Durante el seguimiento, el cáncer gástrico se desarrolló en 90 sujetos en la cohorte de derivación y en 35 sujetos en la cohorte de validación. En la cohorte de derivación, el modelo de predicción de riesgo para el cáncer gástrico se estableció utilizando factores de riesgo significativos: edad, sexo, la combinación del anticuerpo Helicobacter pylori el estado de fumador. La incidencia de cáncer gástrico aumentó significativamente a medida que aumentó la suma de las puntuaciones de riesgo (tendencia P <0,001). La herramienta de evaluación de riesgos se validó internamente y mostró una buena discriminación (C-Statistic = 0.76) y calibración (prueba de Hosmer-Lemeshow P = 0.43) en la cohorte de validación.

III. Brechas de investigación

No se considera la estimación del riesgo teniendo en cuenta la infección por *H.pylori* como principal agente etiológico de la enfermedad.

Se utilizan técnicas estadísticas tradicionales (Modelo de riesgos de Cox), a pesar de contar con una gran cantidad de datos.

Debido a que los modelos incluyen variables dependientes de la cultura nacional, como la alimentación, es necesario el desarrollo de un modelo de riesgo propio para Colombia y el Cauca, ajustado al contexto nacional y regional.

No realizan un análisis de factores de riesgo por género.

IV. Hipótesis

Teniendo en cuenta que la C-Statistic es equivalente al área bajo la curva ROC (entre el valor esté más cercano a 1, mejor), podría afirmarse que los resultados del modelo (C-statistics: 0.764 para hombres, 0.706 para mujeres para el primer artículo y C-statistics: 0.76 para el segundo) podrían mejorarse con técnicas de Machine Learning, ya que éstas permiten el descubrimiento automático de patrones que no son fácilmente visibles en los datos.

Capítulo 3

Marco Interpretativo y Metodológico de la Investigación

En este capítulo, se detalla una breve introducción a la metodología CRISP-DM, sus objetivos, fases y tareas de las que consta (en un alto nivel, pues el detalle de la metodología se aborda en los capítulos IV, V y VI), resumido de la guía del consorcio de empresas que propuso la metodología (Chapman et al., 2000).

Las técnicas de Data Science o Data Analytics, que tanto interés despiertan hoy en día, en realidad surgieron a comienzos de los 2000, cuando se usaba el término KDD (Knowledge Discovery in Databases) para referirse al (amplio) concepto de hallar conocimiento en los datos (Rodríguez León & García Lorenzo, 2016). En un intento de normalización de este proceso de descubrimiento de conocimiento, de forma similar a como se hace en ingeniería software para normalizar el proceso de desarrollo software, surgieron a finales de los 90 dos metodologías principales: CRISP-DM (Cross Industry Standard Process for Data Mining) y SEMMA (Sample, Explore, Modify, Model, and Assess). Ambas especifican las tareas a realizar en cada fase descrita por el proceso, asignando tareas concretas y definiendo lo que es deseable obtener tras cada fase (Wirth, 2000)(Jaramillo & Paz, 2015).

(Azevedo y Santos, 2008) compara ambas implementaciones y llega a la conclusión de que, aunque se puede establecer un paralelismo claro entre ellas,

CRISP-DM es más completo porque tiene en cuenta la aplicación al entorno de negocio de los resultados, y por ello es la que se adoptó popularmente (en encuestas realizadas en KDNuggets en 2002, 2004, 2007 y 2014 se comprobó que CRISP-DM era la principal metodología utilizada, 4 veces más que SEMMA), y es la que se ha considerado en el presente proyecto.

CRISP-DM (Cross Industry Standard Process for Data Mining) proporciona una descripción normalizada del ciclo de vida de un proyecto estándar de análisis de datos, de forma análoga a como se hace en la ingeniería del software con los modelos de ciclo de vida de desarrollo de software (Niakšu, 2014.). El modelo CRISP-DM cubre las fases de un proyecto, sus tareas respectivas, y las relaciones entre estas tareas. En este nivel de descripción no es posible identificar todas las relaciones; las relaciones podrían existir entre cualquier tarea según los objetivos, el contexto, y el interés del usuario sobre los datos.

La metodología CRISP-DM contempla el proceso de análisis de datos como un proyecto profesional, estableciendo así un contexto mucho más rico que influye en la elaboración de los modelos. Este contexto tiene en cuenta la existencia de un cliente que no es parte del equipo de desarrollo, así como el hecho de que el proyecto no sólo no acaba una vez se halla el modelo idóneo (ya que después se requiere un despliegue y un mantenimiento), sino que está relacionado con otros proyectos, y es preciso documentarlo de forma exhaustiva para que otros equipos de desarrollo utilicen el conocimiento adquirido y trabajen a partir de él.

El ciclo de vida del proyecto de minería de datos o analítica, consiste en seis fases mostradas en la siguiente figura:

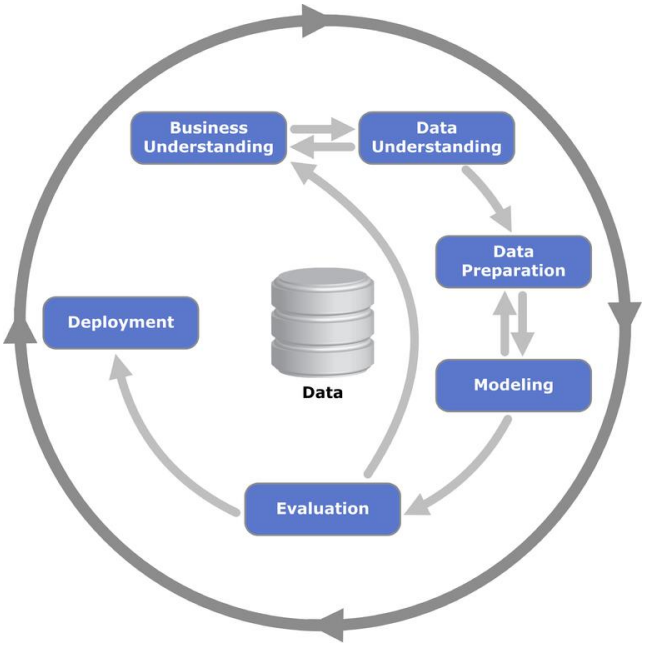

Figura 1 Ciclo de vida metodología CRISP-DM

La secuencia de las fases no es rígida: se permite movimiento hacia adelante y hacia atrás entre diferentes fases. El resultado de cada fase determina qué fase, o qué tarea particular de una fase, hay que hacer después. Las flechas indican las dependencias más importantes y frecuentes. El círculo externo en la figura simboliza la naturaleza cíclica de los proyectos de análisis de datos. El proyecto no se termina una vez que la solución se despliega. La información descubierta durante el proceso y la solución desplegada pueden producir nuevas iteraciones del modelo. Los procesos de análisis subsecuentes se beneficiarán de las experiencias previas.

A continuación se describe brevemente cada una de las fases.

Fase I. Business Understanding / Definición de necesidades del cliente (comprensión del negocio)

Esta fase inicial se enfoca en la comprensión de los objetivos de proyecto. Después se convierte este conocimiento de los datos en la definición de un

problema de minería de datos y en un plan preliminar diseñado para alcanzar los objetivos.

Fase II. Data Understanding / Estudio y comprensión de los datos

La fase de entendimiento de datos comienza con la colección de datos inicial y continúa con las actividades que permiten familiarizarse con los datos, identificar los problemas de calidad, descubrir conocimiento preliminar sobre los datos, y/o descubrir subconjuntos interesantes para formar hipótesis en cuanto a la información oculta.

Fase III. Data Preparation / Análisis de los datos y selección de características

La fase de preparación de datos cubre todas las actividades necesarias para construir el conjunto final de datos (los datos que se utilizarán en las herramientas de modelado) a partir de los datos en bruto iniciales. Las tareas incluyen la selección de tablas, registros y atributos, así como la transformación y la limpieza de datos para las herramientas que modelan.

Fase IV. Modeling / Modelado

En esta fase, se seleccionan y aplican las técnicas de modelado que sean pertinentes al problema (cuantas más mejor), y se calibran sus parámetros a valores óptimos. Típicamente hay varias técnicas para el mismo tipo de problema de minería de datos. Algunas técnicas tienen requerimientos específicos sobre la forma de los datos. Por lo tanto, casi siempre en cualquier proyecto se acaba volviendo a la fase de preparación de datos.

Fase V. Evaluation / Evaluación (obtención de resultados)

En esta etapa en el proyecto, se han construido uno o varios modelos que parecen alcanzar calidad suficiente desde la una perspectiva de análisis de datos.

Antes de proceder al despliegue final del modelo, es importante evaluarlo a fondo y revisar los pasos ejecutados para crearlo, comparar el modelo obtenido con los objetivos de negocio. Un objetivo clave es determinar si hay alguna cuestión importante de negocio que no haya sido considerada suficientemente. Al final de esta fase, se debería obtener una decisión sobre la aplicación de los resultados del proceso de análisis de datos.

Fase VI. Deployment. Despliegue (puesta en producción)

Generalmente, la creación del modelo no es el final del proyecto. Incluso si el objetivo del modelo es de aumentar el conocimiento de los datos, el conocimiento obtenido tendrá que organizarse y presentarse para que el cliente pueda usarlo. Dependiendo de los requisitos, la fase de desarrollo puede ser tan simple como la generación de un informe o tan compleja como la realización periódica y quizás automatizada de un proceso de análisis de datos en la organización.

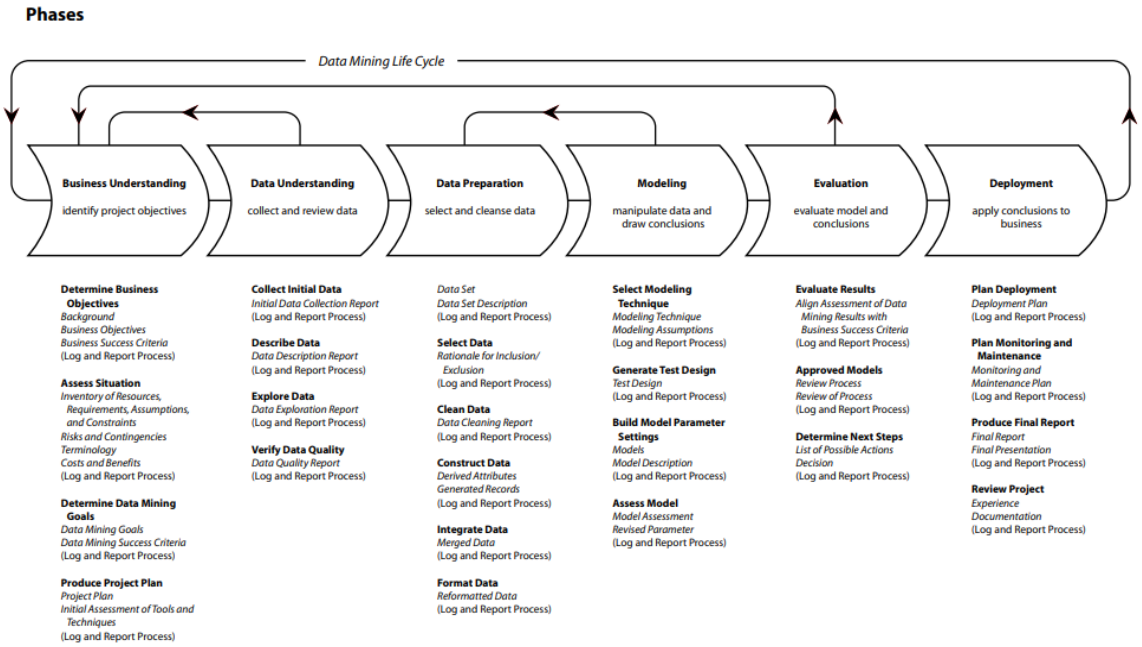

Figura 2 Fases de la Metodología CRISP-DM

Capítulo 4

Comprensión del Negocio y de la Información

Objetivo I

Establecer un proceso de gestión analítica basado en la comprensión del negocio y de la información para la orientación del análisis de datos.

En este capítulo se da inicio a las Fases iniciales del modelo (Fase I y II) de la metodología CRISP-DM con el fin de abordar el primer objetivo del proyecto. Se hizo énfasis en la comprensión del problema mediante la extracción, explotación y conocimiento preliminar de los datos, en un contexto de la salud, que permitiera la predicción del riesgo a cáncer de estómago en la población estudiada.

4.1. Metodología

4.1.1. Fase I Comprensión del Negocio

La comprensión del negocio o problema, es probablemente la más importante y aglutina las tareas de comprensión de los objetivos del negocio y requisitos del proyecto con el fin de convertirlos en objetivos técnicos que permiten definir el plan de proyecto. Esta primera fase fue abordada de la siguiente manera:

- Identificación de los objetivos del negocio: Se realizó un proceso de comprensión del contexto mediante sesiones dirigidas con expertos con el fin de determinar los objetivos del negocio y los criterios de éxito.

- Evaluación de la situación: Se determinaron los recursos claves para el inicio del proyecto tales como la disposición de datos, requerimientos, supuestos y restricciones asociados así como también los riesgos y contingencias a tener en cuenta para el análisis y futura construcción del modelo.

- Determinación de los objetivos de minería de datos: A partir de los objetivos de negocio, se definieron los objetivos de analítica en términos de inteligencia de negocios.

- Generación de un plan de proyecto: Se generó un plan para el desarrollo del proyecto de analítica para la construcción del modelo de predicción del riesgo a desarrollar cáncer de estómago, describiendo las fases, el alcance y el tiempo de ejecución.

4.1.2. Fase II Comprensión de los Datos

Comprendió la recolección inicial de los datos disponibles para el análisis, con el fin de establecer un primer contacto con el problema mediante la generación de tablas y gráficos para identificar las relaciones más evidentes de las variables, lo que permitió definir las primeras hipótesis. Esta segunda fase fue abordada de la siguiente manera:

- Recolección de los datos iniciales: Se recopilaron los datos a partir de una base de datos existente y se adecuaron para su posterior análisis.

- Descripción de los datos: Los datos fueron descritos en términos de cantidad, calidad, estado y disponibilidad. Se establecieron los diferentes sets de datos y el significado de sus variables.

- Exploración de los datos: Se aplicaron pruebas estadísticas básicas con el fin de establecer propiedades de los datos así como también diseñar una estructura general de los mismos. Por otra parte, la exploración también fue útil para buscar errores en los datos y dar forma a las tareas de transformación que tuvieron lugar en la *Fase III Preparación de los datos*.

- Verificación de calidad de los datos: Se evaluó la calidad de los datos para determinar la consistencia de las variables, cantidad y distribución de valores nulos y/o fuera de rango, con el ánimo de identificar incoherencias y asegurar la completitud y corrección de la información.

4.2. Resultados

4.2.1. Fase I Comprensión del Negocio

En esta primera fase se determinaron los objetivos y requisitos del proyecto desde una perspectiva de negocio, para más adelante detallar los objetivos técnicos que conllevan a la generación del plan del proyecto. Los resultados obtenidos se detallan a continuación:

I. Determinación de los Objetivos del Negocio

El objetivo del proceso de análitica que se pretendió implementar es el de realizar predicciones fiables a partir de los datos disponibles en el marco del programa para la prevención de cáncer de estómago en el departamento del Cauca, mediante la comprensión de la infección bacteriana Helicobacter pylori (H.pylori); el agente etiológico más importante. Es importante conocer, que existen algunos factores relacionados con el riesgo a obtener la bacteria como condiciones

socioeconómicas, estilos de vida, hábitos alimenticios, condiciones de saneamiento y saneamiento entre otros.

Contexto del negocio

En el marco del proyecto titulado *"Enfermedades infecciosas emergentes"* ejecutado en 8 municipios del Departamento del Cauca, se priorizó la población a estudiar en cada una de las zonas rurales de los 8 municipios. El muestreo de la población fue de forma estratificado con base en las veredas rurales del municipio acorde con el número de viviendas proporcional al tamaño de la vereda. El tamaño de la muestra fue estimada con base en las fórmulas para estimar los valores poblaciones de los variables objeto. En este cálculo se tuvo en cuenta la variabilidad (desviación típica reportada en estudios previos realizados en el Departamento del Cauca) y el error de estimación que se asume con base a la desviación (75%, 50% o 25%), con una confianza del 95%. Paralelamente se realizó una encuesta semi-estructurada para la obtención de variables socioculturales sobre el uso y manejo del agua en la población, hábitos de higiene, previa firma del consentimiento informado.

Al final, se colectó una base de datos con 13741 registros aproximadamente, en la que se caracterizaron un total de 91 variables, que van desde dimensiones sociales, culturales y demográficas hasta evaluaciones epidemiológicas.
Con respecto a los datos, en el proyecto en mención se realizaron análisis de prevalencia de la infección emergente por la bacteria Helicobacter pylori (H.pylori) y factores asociados.

A nivel mundial, la infección por H. pylori, es una de las infecciones bacterianas más frecuentes. En los países en desarrollo, la mayoría de personas son infectadas en la infancia o preadolescencia y permanecen con la infección durante su vida, mientras que en los países desarrollados, parece que la infección se adquiere gradualmente con la edad La infección de H. pylori es asociada al

desarrollo de úlceras pépticas y cáncer. A nivel mundial, se estima que existen seis millones de casos nuevos de úlcera duodenal cada año y 900,000 casos de carcinoma gástrico. En Colombia se estima que la infección afecta a más del 60% de la poblaciones y en el Cauca la frecuencia de infección asciende al 82%, siendo mayor en población rural (estudios previos realizados por el Grupo GIGHA), lo cual se correlaciona con la alta prevalencia del cáncer gástrico, siendo esta la primera causa de mortalidad por cáncer según la Secretaria Departamental de Salud del Cauca.

Teniendo en cuenta la importancia en salud pública, de esta patología asociada a la infección por esta bacteria, se hacen necesarios modelos integrales de analítica que permitan identificar las verdaderas poblaciones a en la región basados en algoritmos de predicción.

Objetivos del Negocio

Teniendo en cuenta la compresión del problema, y el contexto se estableció como objetivo la construcción y validación de un modelo de analítica que permitiera realizar predicciones fiables para el riesgo a desarrollar Cáncer de estómago partiendo de los datos colectados para 13741 usuarios.

Por lo tanto, el objetivo desde una perspectiva de negocio fue: *Determinar los factores epidemiológicos, sociales y demográficos relacionados con la presencia de la bacteria H.pylori como factor principal de riesgo a desarrollar cáncer de estómago, en 8 municipios del departamento del Cauca.*

Criterios de éxito del negocio

Desde el punto de vista del negocio, se estableció como criterio de éxito la generación del modelo de analítica para la predicción del riesgo de la infección por H.pylori como agente potencial del desarrollo de Cáncer de estómago, teniendo en cuenta multifactorialidad de esta patología tan compleja.

El modelo busca un porcentaje elevado de fiabilidad basado en los datos analizados, de tal forma que se pueden identificar aquellas variables que tienen mayor incidencia en la infección por H.pylori.

II. Evaluación de la Situación

Recursos

El insumo principal correspondió a la fuente de los datos colectada en el marco de la ejecución del Proyecto mencionado previamente. La fuente de los datos se encontró representada en un formato *.xls* (formato de Excel).

El instrumento de recolección de los datos en el marco del proyecto mencionado, corresponde a la encuesta utilizada para el levantamiento de la información, en donde se representa en detalle las preguntas (variables) consignadas en la fuente de datos, categorizada en dimensiones de individuo y hogar para efectos de consulta.

El detalle del proceso de Colección de los datos se describen en el Anexo 3.

Requisitos, supuestos y restricciones

El uso de los datos personales de las personas encuestadas, correspondió a una restricción debido al uso y manejo de la información sensible según la Ley 1581 de 2012 sobre protección de datos personales, que complementa la regulación vigente para la protección del derecho fundamental que tienen todas las personas naturales a autorizar la información personal que es almacenada en bases de datos o archivos. Si bien es cierto que para cada una de las 13741 encuestas realizadas se contó con un consentimiento informado firmado por el encuestado, en donde se accede al tratamiento y uso de la información, por motivos de uso y protección de datos personales y ética profesional se decidió anonimizar esta información.

Esta restricción no representó una desventaja sobre la información que se analizó dado que las variables que incurrían en la restricción en mención, correspondían a variables de identificación tales como nombres, apellidos, número de contacto, entre otros, no relevantes para el objeto del estudio.

No se presentaron restricciones en términos de acceso a la información (contraseñas o permisos especiales).

Terminología

Ver Anexo 1: Glosario de terminología de análitica de datos.

Costos y Beneficios

La ejecución de este proyecto no generó ningún coste adicional en términos de recolección de información para ninguna de las partes dado que los datos objeto de este análisis fueron colectados en el marco de la ejecución del Proyecto mencionado previamente.

En términos de los beneficios, como se mencionó en apartados anteriores, el desarrollo de un modelo de analítica basado en técnicas de minería de datos e inteligencia de negocios, permitirá constituir una base del conocimiento para la priorización e identificación de variables que inciden en el riesgo a desarrollar cáncer de estómago.

Muchos estudios epidemiológicos han evaluado los factores de riesgo de la incidencia de cáncer gástrico, sin embargo, solo unos pocos estudios han desarrollado modelos de predicción, por lo tanto, este modelo permitió descubrir y tener una mejor comprensión de los patrones relacionados con la infección por H.pylori que no fueron visibles fácilmente con las técnicas estadísticas tradicionales debido a la reducida significación estadística. Adicionalmente, el riesgo de cáncer de estómago es una enfermedad multifactorial con distintas

dimensionalidades. Para resolver el problema de la alta dimensionalidad, generalmente se reduce la dimensión de los datos (número de características o variables), lo cual representa beneficios en términos de la eficiencia computacional y en la precisión del modelo, siendo la transformación y selección de características (en términos generales), las dos estrategias utilizadas para el propósito del conocimiento del dominio como parte del proceso de la toma de decisiones.

III. Determinación de Objetivos de Inteligencia de negocios

A continuación se definen los objetivos en términos de minería de datos priorizados para el contexto antes descrito:

- Identificar las variables sociodemográficas más incidentes en la infección por el H.pylori como factor de riesgo a desarrollar Cáncer de estómago.
- Identificar las variables epidemiológicas relacionadas con la infección por el H.pylori como factor de riesgo a desarrollar Cáncer de estómago.
- Predecir el riesgo a desarrollar Cáncer de estómago basado en la presencia del H.pylori a partir de los datos colectados.

Criterios de éxito de minería de datos

Desde el punto de vista de inteligencia de negocios, se estableció como criterio de éxito la posibilidad de identificar las variables que representan mayor incidencia en la infección de H.pylori para la predicción del riesgo a desarrollar cáncer de estómago basados en el modelo de analítica, con un elevado porcentaje de fiabilidad. Concretamente, se ha definido este porcentaje en un 60%. Cabe mencionar que el grado de fiabilidad lo determinó la línea base de los análisis estadísticos convencionales, por lo que este tema se abordará nuevamente en la *Fase V Evaluación*.

IV. Plan del Proyecto

Los apartados de la metodología detallados en cada uno de los capítulos 4, 5 y 6 corresponden al desarrollo del Plan del proyecto.

De igual forma, a continuación se describe de forma resumida las principales etapas del estudio:

Tabla 2 Plan del proyecto

Fase	Resumen	Tiempo
I. Comprensión del negocio	Comprensión de los objetivos y requisitos del proyecto desde una perspectiva de negocio, con el fin de convertirlos en objetivos técnicos y en un plan de proyecto.	1 semana
II. Comprensión de los Datos	Análisis de la estructura de los datos y la información de la base de datos. Ejecución de consultas para tener muestras representativas de los datos.	3 semanas
III. Preparación de los Datos	Preparación de los datos (selección, limpieza, conversión y formateo) para facilitar la minería de datos sobre ellos.	1 mes, 2 semanas
IV. Modelado	Elección de las técnicas de modelado y ejecución de las mismas sobre los datos.	1 mes, 1 semana
V. Evaluación	Análisis de los resultados obtenidos en la etapa anterior, si fuera necesario repetir la etapa de Modelado.	1 semana
VI. Implantación	Producción de informes con los resultados obtenidos en función de los objetivos de negocio y los criterios de éxito establecidos. Presentación de los resultados finales	1 semana

Evaluación inicial de Herramientas y técnicas

En la figura 3 se detallan las herramientas y técnicas utilizadas para llevar a cabo cada una de las fases contempladas en el plan del proyecto.

Figura 3 Herramientas y Técnicas

4.2.2. Fase II Comprensión de los Datos

En esta fase se lograron los siguientes resultados:

I. Recolectar los datos iniciales

Los datos se recibieron en formato *.xls* (Excel), separados en tres hojas correspondientes a los siguientes datasets: i) Hogar ii) Adulto iii) Niño.

Descripción General

i) Hogar

Se encontraron variables relacionadas con las características físicas de las viviendas, comportamientos del grupo familiar frente a los hábitos de saneamiento, higiene, tenencia de mascotas, formas de almacenamiento del agua para

consumo humano, hábitos y estilos de vida, antecedentes gástrico-clínico y por último, conocimientos acerca de la bacteria H.pylori y su relación con enfermedades gástricas como el cáncer de estómago.

ii) Adulto

Se encontraron variables relacionadas con la identificación de los cuidadores de los niños de cada una de los hogares. Adicionalmente variables sociodemográficas, variables de riesgo relacionadas con la infección por H.pylori como la tenencia de animales como posibles vectores de la bacteria, hábitos de aseo personal, antecedentes familiares de infección y síntomas clínicos.

iii) Niño

Se encontraron variables relacionadas con la presencia de de la infección por H.pylori en los niños, identificación personal, características sociodemográficas, características de talla y peso y síntomas clínicos gástricos.

El detalle de los Datasets antes mencionados, puede ser consultado en el Anexo 2 – Diccionario de Datos.

Categorías de Variables

Se definieron categorías y subcategorías para las variables de los tres datasets, tal y como se detalla a continuación.

Tabla 3 Categorías de variables

Categoría	Subcategoría	Descripción
Dataset Hogar		
Identificación	-	Información relacionada con la persona, de forma que la identifica o la hace identificable
Condiciones Sociodemográficas	-	Indicadores sociales, económicos y demográficos que permiten segmentar la población en grupos homogéneos y así definir al público objetivo de estudio
Ubicación	Coordenadas Geográficas	Datos de georeferenciación relacionados con la ubicación de la vivienda.
Condiciones de Vivienda	Características de vivienda	Materiales de construcción usados en la vivienda.
	Acceso a Servicios Públicos	Descripción de los servicios públicos en la vivienda.
	Saneamiento básico	Tecnología disponibles en la vivienda para la eliminación de excretas y aguas residuales
Condiciones de Hogar	Hacinamiento	Se describe de acuerdo con el número de personas por dormitorios en la vivienda.
	Animales	Presencia de animales de forma permanente en la vivienda.
Hábitos de Higiene personal	-	Lavado de manos, frecuencia de cepillado
Agua de Consumo	Prácticas de uso	Técnicas y tecnologías para la potabilización y almacenamiento del agua.
	Percepciones	Características organolépticas del agua de consumo percibidas en el hogar
	Características	condiciones de disponibilidad del agua de consumo
Antecedentes clínicos	-	Enfermedades previas

Familiares		diagnosticada en familiares hasta tercer grado de consanguinidad en el hogar
Conocimiento	-	Información acerca de la bacteria Helicobacter pylori en el hogar
Perfil tecnológico	-	Acceso a las tecnologías de información y comunicaciones en la vivienda.
Dataset Adulto		
Información General	Encuesta	Variables de descripción general de la encuesta como Código del encuestador, Código del Hogar asociado, entre otros.
	Errores encuesta	Variables re capturadas para la solución de errores en el diligenciamiento de encuestas pasadas.
Identificación	-	Información relacionada con la persona, de forma que la identifica o la hace identificable
Condiciones Sociodemográficas	-	Indicadores sociales, económicos y demográficos que permiten segmentar la población en grupos homogéneos y así definir al público objetivo de estudio
Antecedentes Familiares	-	Enfermedades previas diagnosticada en familiares hasta tercer grado de consanguinidad en el hogar
Condiciones de Hogar	Animales	Tenencia de animales por parte del adulto.
Hábitos de higiene	-	Lavado de manos, frecuencia de cepillado
Síntomas	-	Manifestaciones clínicas en el estómago del adulto.
Antecedentes Clínicos	-	Enfermedades previas diagnosticadas en familiares hasta tercer grado de consanguinidad en el adulto

Identificación de H.Pylori	-	Presencia de la bacteria H.pylori en el adulto.
Autogenerado por Dispositivo móvil	-	Variables autogeneradas por el dispositivo móvil (tablet) con el que se tomó la encuesta.
Dataset Niño		
Información General	Encuesta	Variables de descripción general de la encuesta como Código del encuestador, Código del Hogar asociado, entre otros.
Identificación	-	Información relacionada con la persona, de forma que la identifica o la hace identificable
Condiciones Sociodemográficas	-	Indicadores sociales, económicos y demográficos que permiten segmentar la población en grupos homogéneos y así definir al público objetivo de estudio
Antropometría	-	Medidas de talla y peso en el niño
Síntomas	-	Manifestaciones clínicas en el estómago en el niño.
Identificación de H.Pylori	-	Presencia de la bacteria H.pylori en el niño.
Autogenerado por Dispositivo móvil	-	Variables autogeneradas por el dispositivo móvil (tablet) con el que se tomó la encuesta.

La descripción detallada de las variables se encuentra registrada en el Anexo 2 – Diccionario de Datos.

Protección de Datos personales

Previo a realizar el proceso de descripción de datos y de acuerdo a la restricción de uso de información sensible mencionada en un apartado anterior, fue necesario realizar una tarea de anonimización de la información personal a todas las

variables pertenecientes a la categoría *"Identificación"* de los tres datasets. Este es el proceso por el cual se desvincula un dato de interés de un dato personal, hasta el punto que la identificación personal a partir del dato anonimizado no sea posible (incluso mediante la comparación cruzada de los datos con otras fuentes de información).

Este procedimiento se aplicó a las siguientes variables:

Tabla 4 Variables anonimizadas

Hogar	**Adulto**	**Niño**
Primer nombre	Primer nombre	Cédula adulto
Segundo nombre	Segundo nombre	Primer nombre
Primer apellido	Primer apellido	Segundo nombre
Segundo apellido	Segundo apellido	Primer apellido
Fecha de nacimiento	Número de documento	Segundo apellido
Número de documento	Celular	Número de documento
Celular		Fecha de nacimiento

Informe de recopilación de datos

Como una percepción inicial se determinó que las variables que podrían estar más relacionadas con el riesgo a desarrollar cáncer de estómago, correspondían a las categorías: *Características de vivienda*, *Características de Hogar* y *Agua de consumo* para el dataset de *Hogar; Condiciones de Hogar* y *Síntomas* para el dataset de *Adulto;* y *Síntomas* para el dataset de Niño.

De la misma forma, se determinó aquellas variables no relevantes para el estudio y que a priori se podían excluir: Variables de la categorías *Conocimiento* y *Perfil tecnológico* para el dataset de *Hogar* y *Autogenerado por el dispositivo móvil* para los datasets Adulto y Niño.

Se determinó que existían datos suficientes para obtener conclusiones generales y/o realizar predicciones precisas, así como también variables suficientes para

completar el método de modelado contemplado para la consecución de los objetivos del proyecto.

II. Descripción de los datos

Los datos se encontraron registrados en tres datasets relacionados como se muestra en el siquiente esquema.

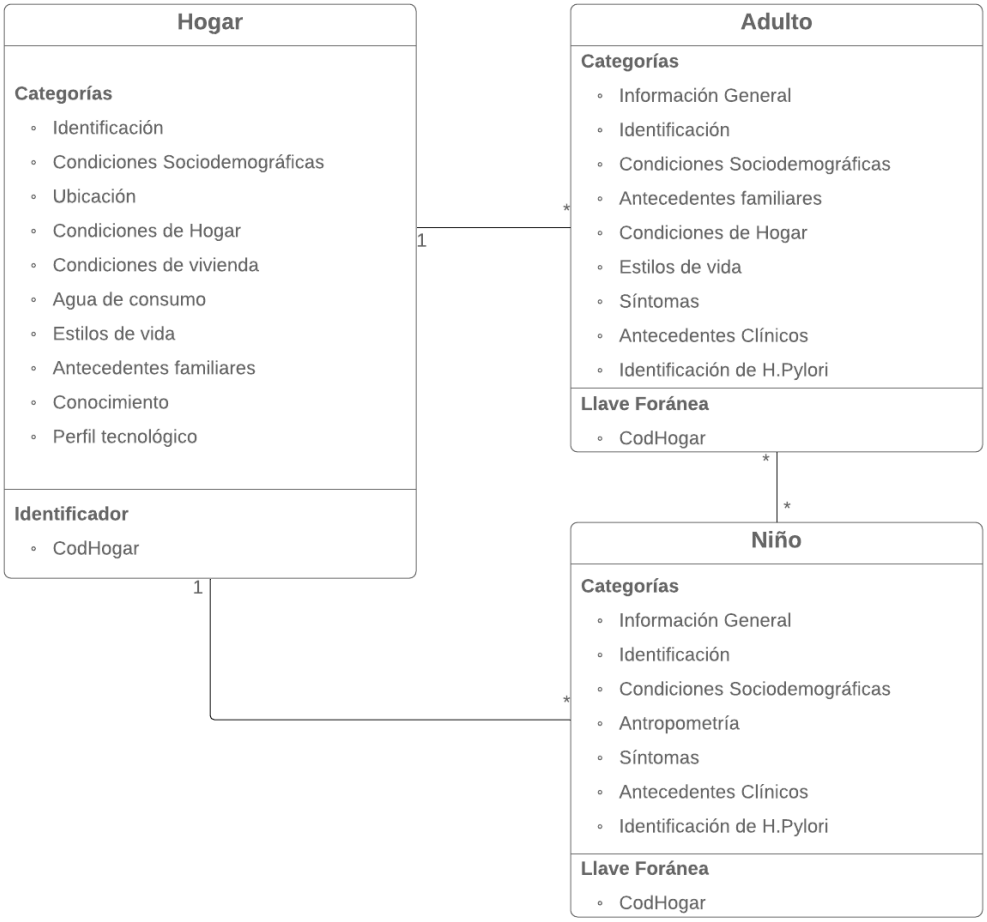

Figura 4 Esquema Racional de los Dataset

Cantidad de datos

Se contó con un total de 13741 registros de encuestas, correspondientes a 5472 adultos y 8269 niños en edades entre 1 y 12 años.

En la tabla 5 se describe la distribución de variables e instancias por Datasets.

Tabla 5 Distribución de variables e instancias por Datasets

Dataset	Número de variables	Número de Instancias
Hogar	91	7563
Adulto	57	5472
Niño	24	8269

El número de variables corresponde al número de columnas presentes en cada uno de los datasets. El número de instancias corresponde al número de filas presentes en cada uno de los datasets.

Tipos de valores

La mayoría de tipos de valor de las variables registradas en los tres datasets, son categóricos, pues corresponden a respuestas de preguntas de opciones múltiples de la encuesta.

Otros tipos de valor identificados en los datasets son numéricos y nominales, datos numéricos de una escala de intervalo o de razón y valores que representan categorías que no obedecen a una clasificación intrínseca.

Esquemas de codificación

No se encontraron esquemas de codificación especiales o que requirieran detalle específico. Cabe aclarar que para algunas variables categóricas se encontraron valores *NaN* correspondientes a datos vacíos para un tipo de categoría en específico. Lo anterior no se constituye como datos faltantes dada la estructura de las variables en los datasets.

III. Exploración de los datos

Una vez descritos los datos, se procedió a explorarlos, esto implicó aplicar pruebas estadísticas básicas, de forma que se pudiera determinar propiedades de los datos para la construcción de tablas de frecuencia y gráficos de distribución.

Este ejercicio se realizó con el uso de la herramienta 'Pandas', siguiendo los comandos detallados en el Anexo 4 - Comandos de Programación.

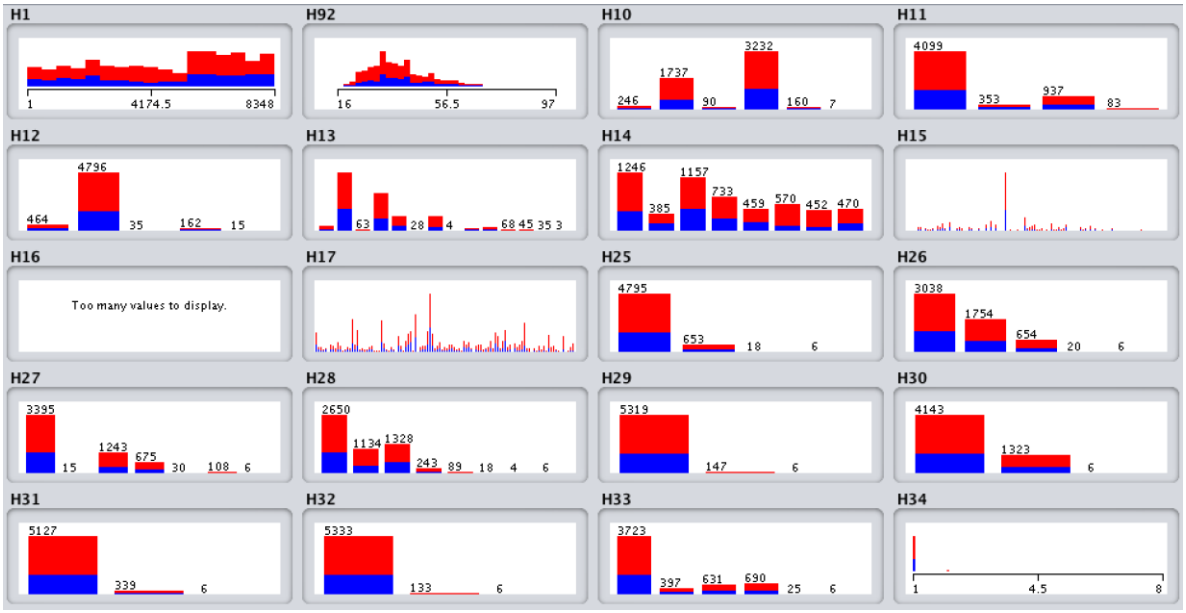

Figura 5 Exploracion de datos

De acuerdo a lo anterior se pudo determinar:

- **H92/Edad:** Los valores altos de la variable (superiores a 32 años aproximadamente) están relacionados con la Positividad de la infección por H.Pylori.

IV. Calidad de los datos

Después de realizada la exploración inicial de los datos, se pudo determinar la completitud de los mismos. Los datos cubrieron los casos requeridos para la

obtención de los resultados necesarios con el ánimo de cumplir con los objetivos del proyecto. No obstante, en función de los resultados de la exploración y verificación de calidad de datos, se presenta a continuación un breve informe con los aspectos más relevantes.

Datos perdidos: Se tuvieron en cuenta valores vacíos o codificados como sin respuesta (cómo *$null$, ? o 999*).
Errores de datos: Se analizaron errores tipográficos cometidos al introducir los datos.
Errores de mediciones: Se incluyeron datos que se introdujeron correctamente, pero basados en un esquema de mediciones incorrecto.
Incoherencias de codificación: Se incluyeron unidades no estándar de medidas o valores incoherentes, como el uso de *'M'* y *'masculino'* para expresar el género.
Metadatos erróneos: Se incluyeron errores entre el significado aparente de un campo incluido en un nombre o la definición del campo.

La solución a lo anterior, se abordó en la fase posterior de Limpieza de datos, detallada en el siguiente capítulo.

Capítulo 5

Preparación de los Datos y Modelado

Objetivo II

Generar un modelo de analítica a partir de una estructura de datos mediante los recursos de Información para la predicción del riesgo a desarrollar cáncer de estómago.

En este capítulo se abordan Fases intermedias del modelo (Fase III y IV) de la metodología CRISP-DM. Se realizó la preparación de los datos que posteriormente dio inicio a la construcción del modelo.

5.1. Metodología

5.1.1. Fase III Preparación de los Datos

La preparación de los datos corresponde a una de las fases más importantes para el desarrollo del modelo de analítica, y en este caso la fase en la que más tiempo se invirtió. En esta fase se prepararon los datos con el objetivo de adecuarlos a las técnicas de minería de datos definidas para emplear sobre ellos. Lo anterior, implicó seleccionar el subconjunto de datos a utilizar, limpiarlos para facilitar su interpretación, generar nuevas variables a partir de las existentes y darles el formato requerido por la tecnología y la herramienta de modelado.

Esta fase fue abordada de la siguiente manera:

- Selección de datos: En esta etapa se seleccionó el subconjunto de datos a analizar mediante técnicas de inclusión y exclusión de datos, a partir de los datasets adquiridos en la fase anterior, apoyándose en los criterios de calidad de datos en términos de coherencia para el estudio, completitud, corrección, volumen, entre otros.

- Limpieza de datos: Esta etapa incluyó la aplicación de distintas técnicas de optimización de calidad de los datasets en términos de corrección de errores en los datos, tareas de normalización, discretización de campos numéricos, tratamiento de datos perdidos, incoherencias de codificación, y metadatos ausentes o erróneos, con el ánimo de prepararlos para la fase posterior de modelación. Por lo anterior, el mayor tiempo de desarrollo se dedicó a la ejecución de esta etapa.

- Estructuración de datos: En esta etapa se aplicaron operaciones de derivación de atributos y generación de registros para la construcción de nuevas variables a partir de las ya existentes.

- Integración de datos: Esta etapa implicó la creación de nuevas estructuras de datos a partir de la fusión de los datasets previamente estructurados. Dicha integración se logró mediante la identificación de un atributo en común (código de hogar) que permitiera relacionar las variables de un dataset con otro. Lo anterior, permitió relacionar distintas categorías de variables para realizar un análisis optimizado de acuerdo a los objetivos del negocio.

- Formateo de datos: En esta etapa se realizaron transformaciones sintácticas a los datos sin modificar su significado, con la idea de permitir o facilitar las técnicas de analítica a utlizar en la fase posterior de Modelado.

5.1.2. Fase IV Modelado

La fase de Modelado se ejecutó en múltiples iteraciones. Se compilaron varios modelos utilizando los parámetros por defecto y posteriormente fueron realizados ajustes paramétricos necesarios con el ánimo de cumplir los criterios de éxito de inteligencia de negocios previamente establecidos. Previo a la construcción de dichos modelos, se determinaron los métodos de evaluación para los modelos en función de las características de los datos y de las características de precisión y fiabilidad que se querían lograr.

Esta etapa fue abordada de la siguiente manera:

- Selección de la técnica de modelado: En primer lugar, se seleccionó la técnica de modelado más apropiada para el proyecto de acuerdo al problema y al contexto del negocio, en función de criterios como: Disposición de datos adecuados, cumplimiento de requisitos del problema, tiempo adecuado para la obtención del modelo y coherencia con los objetivos de inteligencia de negocios.

- Generación del Plan de pruebas: Se determinó el proceso para evaluar la calidad y validez del modelo. Este proceso contempla tres factores clave: El particionamiento de los datos, las métricas de evaluación y las iteraciones a realizar. De acuerdo al contexto del problema que en este caso corresponde al riesgo de una enfermedad de alto impacto y teniendo en cuenta los objetivos de inteligencia de negocios definidos (Predecir el riesgo a desarrollar Cáncer de estómago basado en la presencia del H.pylori a partir de los datos colectados), las métricas de clasificación que se priorizaron para la evaluación fueron el Recall (sensibilidad) seguido por Accuracy (tasa de correctitud).

- Construcción del modelo: La generación del modelo constituyó a un proceso iterativo para lo cual se tuvo en cuenta la configuración de parámetros, los modelos reales producidos y descripciones de los resultados de los modelos.

- Evaluación del modelo: Se evaluaron los modelos con base en las métricas establecidas en el Plan de pruebas y se interpretaron de acuerdo a los criterios de éxito de inteligencia de negocio definidos. Cabe aclarar que esta evaluación está relacionada con los objetivos de inteligencia de negocios definidos anteriormente.

5.2. Resultados

5.2.1. Fase III Preparación de los Datos

Los resultados obtenidos en esta fase se detallan a continuación:

I. Selección de Datos

Un paso realmente importante en el aprendizaje automático es el análisis de variables y la ingeniería de características. Esto nos permitió obtener información valiosa de los datos e identificar las variables más importantes con el fin de definir la relevancia de mantenerlas o no previo a entrenar el modelo predictivo.

En términos del número de instancias (registros), se utilizaron todas las asociadas a cada Dataset que compone nuestra base de datos de estudio con el fin de aprovechar la mayor cantidad de información asociada a las variables definidas. Por otra parte, se prescindió de algunas variables dado que para cumplir con los objetivos de inteligencia de negocio definidos en la fase I, dichas variables no aportan información relevante que requiera ser estudiada. Este proceso no se debe confundir con el de reducción de dimensiones, si bien ambos procesos buscan reducir el número de atributos en nuestro Dataset, este último lo hace por

medio de la creación de nuevos atributos que son combinaciones de los anteriores, mientras que en el proceso de selección de atributos intentamos incluir y excluir los atributos prácticamente sin modificarlos.

El objetivo de la selección de atributos es triple: mejorar la capacidad predictiva de nuestro modelo proporcionando modelos predictivos más rápidos y eficientes, y proporcionar una mejor comprensión del proceso subyacente que generó los datos.

Las variables que no se tuvieron en cuenta se detallan a continuación:

Tabla 6 Variables excluidas

Categoría	Variables	Razón de la exclusión en el estudio
Dataset Hogar		
Identificación	Primer nombre Segundo nombre Primer apellido Segundo apellido Fecha de nacimiento Número de documento Celular	Información personal anonimizada por razones de estándares de Protección de Datos personales.
Ubicación	Latitud (N) Minutos, Grados, Segundos Longitud (W) Minutos, Grados, Segundos Altitud (msnm) Georeferenciados	Datos de georeferenciación relacionados con la ubicación de la vivienda.
Conocimiento		Variables de caracterización del conocimiento de los habitantes del hogar con respecto a la infección por H.pylori y factores asociados.
Perfil tecnológico		Variables de caracterización

		del perfil tecnológico de los habitantes del hogar.
Dataset Adulto y Dataset Niño		
Información General	Encuestador CodHogar Errores nombres Errores cédula Errores fecha nac. Errores celular	Variables de información general de la encuesta como Código del encuestador, Código del Hogar asociado, etc.
Identificación	Primer nombre Segundo nombre Primer apellido Segundo apellido Número de documento Celular	Información personal anonimizada por razones de estándares de Protección de Datos personales.
Autogenerado por dispositivo móvil	Id Submission time Version device Version survey	Variables autogeneradas por el dispositivo móvil (tablet) con el que se tomó la encuesta.

II. Limpieza de Datos

Previo a la construcción del modelo, se hizo necesario realizar un proceso de limpieza de datos que implicó observar más de cerca los problemas en los datos previamente seleccionados.

A continuación se detallan los procedimientos de limpieza de datos llevados a cabo en términos de datos perdidos, errores de datos y codificación:

Datos perdidos

No se registraron datos perdidos dentro de las variables seleccionadas y para los que se hiciese necesario realizar un proceso de completar los datasets.

Errores de datos

Se hizo necesario realizar correcciones en datos de tipo Fecha que se representaban valores erróneos, por ejemplo: 19490 se reemplazó por 1949.

Codificación

- **Limpieza de variables de tipo fecha**

Se realizó un proceso para dar un formato correcto a las variables de tipo fecha, dado que era frecuente encontrar fechas de tipo 12/03/1967 o 1989-2-26. El formato estándar utilizado para la gestión de fechas dentro del dataset fue dd/mm/aaaa.

- **Reemplazo de nombres de variables por identificadores únicos.**

Se reemplazaron los nombres de las variables por identificadores únicos tal y como se presentan en el Anexo 2 - Diccionario de Datos.

Tabla 7 Identificadores únicos de variables

Código	Nombre de la variable
Dataset Hogar	
H01	Código de hogar
H02	Primer nombre
...	
H30	¿La unidad de vivienda cuenta con servicio de alcantarillado?
...	
Dataset Adulto	
...	
A37	Frecuencia dolor de estómago
...	
A41	Frecuencia vómito

Dataset Niño	
...	
N14	Antropometría/peso
...	
N18	Resultado H.pylori

Lo anterior, permitió realizar una mejor gestión de las variables en los análisis por realizar. Se decidió seguir la siguiente nomenclatura: Para el Dataset Hogar se utilizó la letra H seguido de un consecutivo entre 1 y 91; Para el Dataset Adulto se utilizó la letra A seguido de un consecutivo entre 1 y 57; Para el Dataset Niño se utilizó la letra N seguido de un consecutivo entre 1 y 24. De esta forma, cada variable del Dataset se encuentra mapeada a un identificador único. Para la construcción de nuevas variables, se continuó con la nomenclatura antes definida, así las cosas, por ejemplo para la variable Edad se definió el código H92 (proceso detallado en la siguiente sección de Construcción de datos).

- **Codificación de características categóricas o de clases nominales**

Previo a iniciar a trabajar con algoritmos de aprendizaje automático, se hizo necesario representar las variables no numéricas (es decir, características categóricas como Sexo, Nivel educativo, Tipo de vivienda, entre otras) en un formato numérico que un algoritmo pudiera "entender". Este proceso se denomina codificación de características.

Existen muchas técnicas de codificación, la utilizada para este proceso se denomina “Variables dummy”, que transforma la información de la variable en valores de 0 o 1 para indicar la ausencia o presencia del dato.
Este ejercicio se realizó con el uso de la herramienta ‘Pandas’, siguiendo los comandos detallados en el Anexo 4 - Comandos de Programación.

La Figura (Nombre figura) representa el proceso de codificación para la variable Sexo (A17) del Dataset Adulto.

Id	A17
1	Masculino
2	Femenino
3	Femenino

→

Id	A17_Masculino	A17 Femenino
1	1	0
2	0	1
3	0	1

Figura 6 Codificación de características categóricas

El detalle de codificación para todas las variables categóricas de los tres Datasets puede ser consultado en el Anexo 2 - Diccionario de Datos.

III. Construcción de los Datos

- **Creación de variables auto calculadas a partir de variables existentes**

Como parte de los datasets Adulto y Niño se registró la Fecha de nacimiento del encuestado. A partir de esta variable se calculó la Edad, variable indispensable para los objetivos de minería de datos, por sugerencia del experto.

Una vez calculada la variable Edad, como se mencionó en un apartado anterior, se omitió la variable Fecha de nacimiento de los datos dado que hacía parte de la categoría Identificación, priorizada como una de las categorías a anonimizar dentro de los datasets por motivos de Uso y manejo de información personal.

IV. Integración de los Datos

De acuerdo con los objetivos del negocio y de minería de datos definidos en el capítulo anterior, la variable de interés a predecir es la positividad o negatividad de H.pylori. Esta variable se encuentra registrada en los datasets Adulto y Niño. No obstante, lo que se buscó, fue enriquecer el análisis de los datos mediante la

integración de las variables presentes en el dataset de Hogar, pues es en este último en donde se encontraban registradas variables de Saneamiento, Hacinamiento, Condiciones de vivienda, etc que se hacía necesario analizar en conjunto con el ánimo de evaluar la posible relación de las mismas con el factor clave de presencia de H.Pylori que en últimas, deriva en el desarrollo o no de Cáncer de estómago.

Para lo anterior, se hizo necesario realizar un proceso de integración de datos que permitiera unir los datasets Hogar-Adulto y Hogar-Niño. Esto fue posible, relacionando los códigos únicos de Hogar presentes en los datasets Adulto y Niño, con los valores registrados en la variable 'CodHogar' del dataset Hogar.

Fusión de Datasets

El resultado de la integración de los Dataset se resume en la creación de dos nuevos datasets de datos detallados a continuación:

- **Dataset Hogar Adulto**

Se relacionó la variable A3 (g_inf_general/hogar) del Dataset Adulto con la variable H1 (CodHogar) del Dataset Hogar para lograr la integración de los dos Datasets.

La variable decisión (Identificación de H.Pylori) se encontró mapeada en la variable A51.

Tabla 8 Integración de Datasets Hogar-Adulto

Número de variables	Número de Instancias	Identificación de H.Pylori	
73	5472	**Positivo**	3595 **(65.69%)**
		Negativo	1877 **(34.31%)**

Categorías	
Condiciones Sociodemográficas	Antecedentes familiares
Condiciones de vivienda	Hábitos de Higiene
Condiciones de Hogar	Síntomas pépticos
Hábitos saludables	Antecedentes clínicos
Agua de consumo	

- **Dataset Hogar Niño**

Se relacionó la variable N3 *(g_inf_general/hogar)* del Dataset Niño con la variable H1 *(CodHogar)* del Dataset Hogar para lograr la integración de los dos Datasets.
La variable decisión (Identificación de H.Pylori) se encontró mapeada en la variable N18.

Tabla 9 Integración de Datasets Hogar-Niño

Número de variables	Número de Instancias	Identificación de H.Pylori	
54	8267	**Positivo**	4911 **(59.40%)**
		Negativo	3356 **(40.60%)**

Categorías	
Condiciones Sociodemográficas	Agua de consumo
Condiciones de vivienda	Antecedentes familiares
Condiciones de Hogar	Antropometría
Hábitos saludables	Síntomas

V. Formateo de los Datos

Como paso final previo a la construcción del modelo, fue útil comprobar si las técnicas de análisis a utilizar requerían de un formato concreto para la clasificación de algunas variables.

Gracias a la codificación de características categóricas detallada en un apartado anterior, no fue necesario aplicar un proceso adicional de formateo de los datos. Tampoco se hizo necesario aplicar algún proceso para reordenar los datos (instancias) de los datasets, así como tampoco la reordenación de categorías para variables en particular.

5.2.2. Fase IV Modelado

I. Selección de la técnica de modelado

Dado que el modelo que se pretendió implementar correspondía a un modelo de predicción, dentro de área de Machine learning encontramos dos tipos de aprendizaje automático (Supervisado y No supervisado), cada uno con unas técnicas de modelado propias (Clasificación, regresión, clustering, entre otras) cuya selección y utilización varía en función de las características de los datos y de los objetivos del ejercicio.

En primer lugar se hizo necesario definir la técnica de aprendizaje automático a seguir, seguido por los tipos de algoritmos a probar sobre los datasets. Para lo primero se tuvo en cuenta la estructura de los datasets, esto es, que a partir de la "Integración de datos" realizada en la fase anterior, cada una de las instancias de los datasets resultantes (Hogar-Adulto y Hogar-Niño) contaba con una etiqueta de clase, en este caso Infección por H.pylori (positivo o negativo), lo cual se ajustaba a las técnicas de Aprendizaje supervisado. Para lo segundo y dentro de las técnicas de Aprendizaje supervisado, se optó por algoritmos de Clasificación; El uso de estos últimos se hizo necesario dado que la variable que se quiere predecir corresponde a una variable categórica (positivo o negativo), y no a una variable de valores continuos como lo sería entonces el caso de algoritmos de Regresión.

Por lo anterior y teniendo en cuenta los objetivos de inteligencia de negocio definidos, los factores relacionados con la estructura de los datasets y el volumen

de los datos, la técnica de modelado escogida correspondió al uso de algoritmos de Clasificación de Aprendizaje supervisado.

II. Generación del Plan de pruebas

Para la definición del Plan de pruebas se tuvieron en cuenta los siguientes factores:

Métricas de evaluación

Las métricas de evaluación definidas para la evaluación de los modelos, corresponden a las métricas de los algoritmos de Clasificación. En el caso de modelos de clasificación con fines médicos se debe maximizar el número de positivos verdaderos y minimizar el número de falsos negativos, ya que en dicho contexto es mejor identificar un falso positivo, pues se podría desestimar dicho falso positivo con tests o pruebas clínicas adicionales, por otra parte un falso negativo puede ocasionar que una condición médica siga sin tratamiento oportuno lo que conlleva un impacto negativo en la salud del paciente. Dicho lo anterior, la métrica más adecuada para validar estas condiciones es el “Recall”.

- Recall o sensibilidad: Indica la proporción de todos los positivos reales que el modelo es capaz de identificar. Se priorizó esta medida debido al interés de reducir al máximo los “Falsos Negativos”, es decir aquellos individuos que el modelo predice como Negativos cuando en la realidad son Positivos.

- Accuracy o tasa de correctitud: Corresponde a la fracción de predicciones que el modelo realizó correctamente.

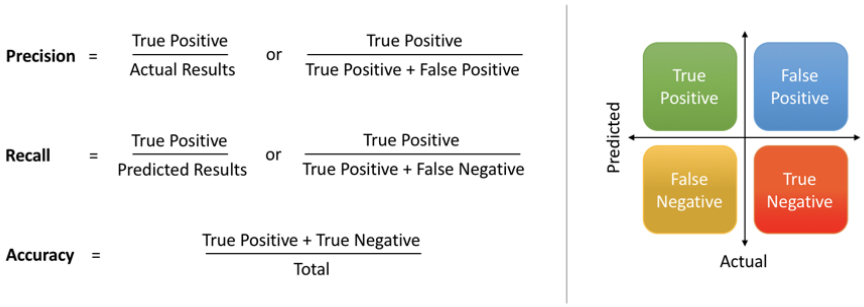

Figura 7 Métricas de Evaluación

Algoritmos e Iteraciones

A priori, los algoritmos a utilizar para la generación de los modelos fueron:

- `LogisticRegression`
- `KNeighborsClassifier`
- `SVC`
- `NuSVC`
- `DecisionTreeClassifier`
- `RandomForestClassifier`
- `AdaBoostClassifier`
- `GradientBoostingClassifier`
- `GaussianNB`

En una primera iteración, para cada algoritmo se construyó un modelo de predicción con los parámetros por defecto definidos y se evaluaron sus resultados. Posteriormente y de acuerdo a la evaluación del modelo con respecto a las métricas definidas, se seleccionaron los 3 mejores algoritmos para los que, en una segunda iteración, se realizó un proceso exhaustivo de ajuste de parámetros con el fin de mejorar la calidad del modelo.

Particionamiento de los datos y validación

De acuerdo al método de ejecución definido en el apartado anterior, para la primera iteración los datasets se particionaron en 2 conjuntos: Uno de entrenamiento y uno de pruebas. El primero corresponde a los datos que se utilizaron para generar/entrenar los modelos, y el segundo a los datos que emplearon para realizar las pruebas y medir la calidad de los modelos.

Para la segunda iteración (una vez seleccionados los 3 mejores algoritmos) se utilizó la técnica de “Validación cruzada”. La Validación cruzada es muy similar al particionamiento de entrenamientos/pruebas, pero se aplica a más subconjuntos.

Existen muchos métodos de Validación cruzada, el utilizado en este caso correspondió al método de validación K-Fold. K-Fold consiste en dividir los datos en *K* particiones del mismo tamaño. Para cada partición *i*, el modelo es entrenado con las restantes *K-1* particiones, y evaluado en la propia partición *i*. La puntuación final es la media de las *K* puntuaciones obtenidas.

III. Construcción del modelo

Cabe recordar que como resultado de la fase previa, se crearon dos nuevos Datasets (Hogar-Adulto y Hogar-Niño) producto de la integración de los sets de datos originales (Hogar, Adulto y Niño). Dicho esto, el proceso de construcción de modelos se realizó tanto para el Dataset Hogar-Adulto, como para el Dataset Hogar-Niño.

Se procedió a construir los distintos modelos sobre los conjuntos de datos de entrenamiento. A continuación se detallan los procesos de construcción de los distintos modelos, los ajustes de parámetros realizados, los modelos seleccionados, su descripción y su evaluación:

Este ejercicio se realizó con el uso de la herramienta ‘Pandas’, siguiendo los comandos detallados en el Anexo 4 - Comandos de Programación.

Modelos

Como parte de la primera iteración, se procedió a construir los modelos. El resultado se presenta a continuación:

- **Dataset Hogar-Adulto**

1. LogisticRegression

```
LogisticRegression
****Results****
Accuracy: 63.5810%
              precision    recall  f1-score   support

           0       0.48      0.34      0.40       578
           1       0.69      0.80      0.74      1064

   micro avg       0.64      0.64      0.64      1642
   macro avg       0.58      0.57      0.57      1642
weighted avg       0.61      0.64      0.62      1642

Log Loss: 0.8434163778868042
```

2. KNeighborsClassifier

```
KNeighborsClassifier
****Results****
Accuracy: 62.6066%
              precision    recall  f1-score   support

           0       0.46      0.37      0.41       578
           1       0.69      0.77      0.73      1064

   micro avg       0.63      0.63      0.63      1642
   macro avg       0.58      0.57      0.57      1642
weighted avg       0.61      0.63      0.61      1642

Log Loss: 0.9491934159826468
```

3. SVC

```
SVC
****Results****
Accuracy: 64.7990%
              precision    recall  f1-score   support

           0       0.00      0.00      0.00       578
           1       0.65      1.00      0.79      1064

   micro avg       0.65      0.65      0.65      1642
   macro avg       0.32      0.50      0.39      1642
weighted avg       0.42      0.65      0.51      1642

Log Loss: 0.6345552571953936
```

4. NuSVC

```
NuSVC
****Results****
Accuracy: 64.4336%
              precision    recall  f1-score   support

           0       0.49      0.31      0.38       578
           1       0.69      0.82      0.75      1064

   micro avg       0.64      0.64      0.64      1642
   macro avg       0.59      0.57      0.57      1642
weighted avg       0.62      0.64      0.62      1642

Log Loss: 0.6310253429072727
```

6. RandomForestClassifier

```
RandomForestClassifier
****Results****
Accuracy: 62.1803%
              precision    recall  f1-score   support

           0       0.45      0.35      0.40       578
           1       0.69      0.77      0.72      1064

   micro avg       0.62      0.62      0.62      1642
   macro avg       0.57      0.56      0.56      1642
weighted avg       0.60      0.62      0.61      1642

Log Loss: 0.9010040817051187
```

7. AdaBoostClassifier

```
AdaBoostClassifier
****Results****
Accuracy: 65.6516%
              precision    recall  f1-score   support

           0       0.53      0.23      0.32       578
           1       0.68      0.89      0.77      1064

   micro avg       0.66      0.66      0.66      1642
   macro avg       0.60      0.56      0.55      1642
weighted avg       0.63      0.66      0.61      1642

Log Loss: 0.6844162816014976
```

8. GradientBoostingClassifier

```
GradientBoostingClassifier
****Results****
Accuracy: 65.3471%
              precision    recall  f1-score   support

           0       0.54      0.12      0.19       578
           1       0.66      0.94      0.78      1064

   micro avg       0.65      0.65      0.65      1642
   macro avg       0.60      0.53      0.49      1642
weighted avg       0.62      0.65      0.57      1642

Log Loss: 0.61876125916315
```

9. GaussianNB

```
GaussianNB
****Results****
Accuracy: 39.5859%
              precision    recall  f1-score   support

           0       0.36      0.94      0.52       578
           1       0.75      0.10      0.18      1064

   micro avg       0.40      0.40      0.40      1642
   macro avg       0.55      0.52      0.35      1642
weighted avg       0.61      0.40      0.30      1642

Log Loss: 20.859557045567424
```

5. DecisionTreeClassifier

```
DecisionTreeClassifier
****Results****
Accuracy: 58.7698%
              precision    recall  f1-score   support

           0       0.41      0.41      0.41       578
           1       0.68      0.69      0.68      1064

   micro avg       0.59      0.59      0.59      1642
   macro avg       0.55      0.55      0.55      1642
weighted avg       0.59      0.59      0.59      1642

Log Loss: 14.240409025185466
```

- **Dataset Hogar-Niño**

1. LogisticRegression

```
LogisticRegression
****Results****
Accuracy: 61.1447%
              precision    recall  f1-score   support

           0       0.51      0.38      0.43       976
           1       0.65      0.76      0.70      1505

   micro avg       0.61      0.61      0.61      2481
   macro avg       0.58      0.57      0.57      2481
weighted avg       0.60      0.61      0.60      2481

Log Loss: 0.6797442165622203
```

2. KNeighborsClassifier

```
KNeighborsClassifier
****Results****
Accuracy: 54.4942%
              precision    recall  f1-score   support

           0       0.43      0.49      0.46       976
           1       0.64      0.58      0.61      1505

   micro avg       0.54      0.54      0.54      2481
   macro avg       0.53      0.54      0.53      2481
weighted avg       0.56      0.54      0.55      2481

Log Loss: 0.809067807906377
```

3. SVC

```
SVC
****Results****
Accuracy: 60.6610%
              precision    recall  f1-score   support

           0       0.00      0.00      0.00       976
           1       0.61      1.00      0.76      1505

   micro avg       0.61      0.61      0.61      2481
   macro avg       0.30      0.50      0.38      2481
weighted avg       0.37      0.61      0.46      2481

Log Loss: 0.6607217895465263
```

6. RandomForestClassifier

```
RandomForestClassifier
****Results****
Accuracy: 59.6534%
              precision    recall  f1-score   support

           0       0.49      0.51      0.50       976
           1       0.67      0.65      0.66      1505

   micro avg       0.60      0.60      0.60      2481
   macro avg       0.58      0.58      0.58      2481
weighted avg       0.60      0.60      0.60      2481

Log Loss: 0.9643906276621121
```

7. AdaBoostClassifier

```
AdaBoostClassifier
****Results****
Accuracy: 61.4268%
              precision    recall  f1-score   support

           0       0.51      0.35      0.41       976
           1       0.65      0.79      0.71      1505

   micro avg       0.61      0.61      0.61      2481
   macro avg       0.58      0.57      0.56      2481
weighted avg       0.60      0.61      0.60      2481

Log Loss: 0.6909056063382352
```

8. GradientBoostingClassifier

```
GradientBoostingClassifier
****Results****
Accuracy: 62.5554%
              precision    recall  f1-score   support

           0       0.54      0.30      0.38       976
           1       0.65      0.84      0.73      1505

   micro avg       0.63      0.63      0.63      2481
   macro avg       0.60      0.57      0.56      2481
weighted avg       0.61      0.63      0.59      2481

Log Loss: 0.6457979255977274
```

4. NuSVC

```
NuSVC
****Results****
Accuracy: 59.0085%
              precision    recall  f1-score   support

           0       0.48      0.44      0.46       976
           1       0.65      0.69      0.67      1505

   micro avg       0.59      0.59      0.59      2481
   macro avg       0.57      0.56      0.56      2481
weighted avg       0.58      0.59      0.59      2481

Log Loss: 0.6597247203966062
```

9. GaussianNB

```
GaussianNB
****Results****
Accuracy: 60.3789%
              precision    recall  f1-score   support

           0       0.48      0.10      0.17       976
           1       0.61      0.93      0.74      1505

   micro avg       0.60      0.60      0.60      2481
   macro avg       0.55      0.52      0.45      2481
weighted avg       0.56      0.60      0.51      2481

Log Loss: 12.065795191827082
```

5. DecisionTreeClassifier

```
DecisionTreeClassifier
****Results****
Accuracy: 56.1870%
              precision    recall  f1-score   support

           0       0.44      0.45      0.45       976
           1       0.64      0.63      0.64      1505

   micro avg       0.56      0.56      0.56      2481
   macro avg       0.54      0.54      0.54      2481
weighted avg       0.56      0.56      0.56      2481

Log Loss: 15.132466723606575
```

Descripción de los Modelos

Para cada modelo se calculó un *Accuracy* que describe la precisión general del mismo con respecto al objetivo de predicción definido, del mismo modo también se calcularon las métricas de *Precisión, Recall, F1 Score* and *Support* con respecto a la variable de predicción (Infección por H.pylori) mapeada de la siguiente forma:

- 0: Negativo para H.pylori
- 1: Positivo para H.pylori.

Los valores de Recall para la estiqueta ‘0’ (Negativo), representan la precisión de los modelos para detectar los Falsos Positivos. Los valores de Log loss miden el rendimiento del modelo en la medida en que tiene en cuenta la incertidumbre de la predicción en función de cuánto varía de la etiqueta real, es decir el Log loss aumenta a medida que el valor predicho difiere del valor real, por lo que el objetivo de nuestros modelos fue el minimizar este valor.

Ajuste de parámetros

Como parte de la segunda iteración y de acuerdo a los resultados obtenidos para los 10 modelos, se determinó que los modelos #3 SVC, #6 RandomForest y #8 GradientBoosting, presentaban los mejores resultados en términos de equilibrio entre las métricas de *Recall* y *Accuracy* definidas en el Plan de pruebas, comparados con los otros modelos.

Tabla 10 Algoritmos Priorizados

Dataset Hogar-Adulto		
SVC *Accuracy:* 64.75 *Recall*: 100%	**Random Forest** *Accuracy:* 62.18 *Recall:* 77%	**Gradient Boosting** *Accuracy:* 65.35 *Recall:* 95%
Dataset Hogar-Niño		
SVC *Accuracy:* 64.75 *Recall*: 100%	**Random Forest** *Accuracy:* 62.18 *Recall:* 77%	**Gradient Boosting** *Accuracy:* 65.35 *Recall:* 95%

- **Grid Search o Búsqueda de grillas**

Para los tres modelos seleccionados se realizó un proceso de ajuste de parámetros, aquellos que rigen el propio proceso de entrenamiento con el fin de producir mejores resultados en términos de calidad de los modelos.

El objetivo de la exploración de los parámetros es buscar entre diversas configuraciones de hiper parámetros hasta dar con aquella que tenga como resultado un rendimiento óptimo. Cada algoritmo tiene sus propios parámetros a ajustar, luego como es de esperar, es un proceso arduo e iterativo de búsqueda de mejores combinaciones para obtener el mayor poder predictivo posible.

Para optimizar esta búsqueda de mejores parámetros, se utilizó la técnica de Grid Search o Búsqueda de cuadrícula, que consiste en una búsqueda exhaustiva a

través de un subconjunto especificado manualmente del espacio de los parámetros para cada algoritmo de aprendizaje.

Para esto:

- Se estableció la matriz de parámetros a evaluar. Esto se hizo creando un diccionario de todos los parámetros y su conjunto de valores correspondientes deseado para cada algoritmo.

- Se estableció el número de subdivisiones de datos, el estado aleatorio y un método de puntuación.

- Se definió un objeto “K-Fold”con el número de subdivisiones seleccionado.

Ejemplo para el algoritmo #1 SVC

```
# Hyper parámetros
svc_classifier = SVC(random_state=42)

# Matriz de parámetros
C = [0.5, 1.0, 1.5]
kernel = ['rbf', 'poly', 'sigmoid']
gamma = ['auto', 'scale']
param_grid = dict(C=C, kernel=kernel, gamma=gamma)

# K-Fold
kfold = KFold(n_splits=5, shuffle=True, random_state=42)

# Resultados
Starting Grid Search With Params: {'C': [0.5, 1.0, 1.5], 'kernel': ['rbf', 'poly',
'sigmoid'], 'gamma': ['auto', 'scale']}

Best: 0.972306 using {'C': 0.5, 'gamma': 'scale', 'kernel': 'poly'}

0.968513 (0.005597) with: {'C': 0.5, 'gamma': 'auto', 'kernel': 'rbf'}
0.971901 (0.005382) with: {'C': 0.5, 'gamma': 'auto', 'kernel': 'poly'}
0.971575 (0.005712) with: {'C': 0.5, 'gamma': 'auto', 'kernel': 'sigmoid'}
0.968513 (0.005597) with: {'C': 0.5, 'gamma': 'scale', 'kernel': 'rbf'}
0.972306 (0.004599) with: {'C': 0.5, 'gamma': 'scale', 'kernel': 'poly'}
0.971582 (0.004954) with: {'C': 0.5, 'gamma': 'scale', 'kernel': 'sigmoid'}
0.900775 (0.005875) with: {'C': 1.0, 'gamma': 'auto', 'kernel': 'rbf'}
0.956958 (0.007840) with: {'C': 1.0, 'gamma': 'auto', 'kernel': 'poly'}
0.914601 (0.006232) with: {'C': 1.0, 'gamma': 'auto', 'kernel': 'sigmoid'}
```

```
0.900775 (0.005875) with: {'C': 1.0, 'gamma': 'scale', 'kernel': 'rbf'}
0.957364 (0.007476) with: {'C': 1.0, 'gamma': 'scale', 'kernel': 'poly'}
0.914202 (0.006417) with: {'C': 1.0, 'gamma': 'scale', 'kernel': 'sigmoid'}
0.878600 (0.006416) with: {'C': 1.5, 'gamma': 'auto', 'kernel': 'rbf'}
0.945131 (0.009106) with: {'C': 1.5, 'gamma': 'auto', 'kernel': 'poly'}
0.884497 (0.009657) with: {'C': 1.5, 'gamma': 'auto', 'kernel': 'sigmoid'}
0.878600 (0.006416) with: {'C': 1.5, 'gamma': 'scale', 'kernel': 'rbf'}
0.945514 (0.009351) with: {'C': 1.5, 'gamma': 'scale', 'kernel': 'poly'}
0.885703 (0.008108) with: {'C': 1.5, 'gamma': 'scale', 'kernel': 'sigmoid'}
```

Este ejercicio se realizó con el uso de la herramienta 'Pandas', siguiendo los comandos detallados en el Anexo 4 - Comandos de Programación.

Adicional a lo anterior, se graficaron los resultados para facilitar su visualización:

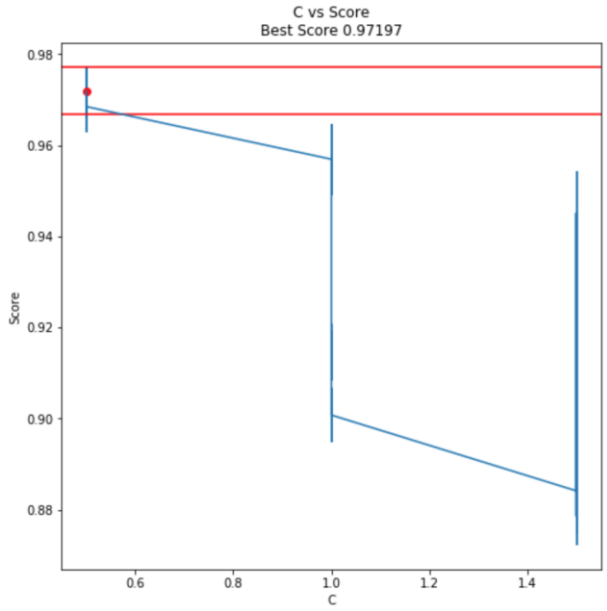

Figura 8 Ajuste de parámetros para el Modelo SVC

De acuerdo a lo anterior se identificó la mejor combinación de parámetros para el algoritmo.

Esta técnica permitió identificar un conjunto de hiper parámetros que se ajustaba de mejor forma a los algoritmos seleccionados. Los resultados se detallan a continuación:

Tabla 11 Resultados de Hiperparámetros

Dataset Hogar-Adulto			
	Algoritmo		
Parámetro / Valor	SVC	Gradient Boosting	Random Forest
	C = 0.2	learning_rate = 0.01	n_estimators = 10
	kernel = rbf	max_depth = 3	max_depth = 3
	gamma = auto	max_features = log2	max_features = auto
Dataset Hogar-Niño			
	Algoritmo		
Parámetro / Valor	SVC	Gradient Boosting	Random Forest
	C = 0.2	learning_rate = 0.01	n_estimators = 200
	kernel = rbf	max_depth = 3	max_depth = 3
	gamma = auto	max_features = log2	max_features = log2

IV. Evaluación del modelo

Como se detalló en el Plan de pruebas, se utilizó la técnica de K-Fold con el fin de evaluar los modelos seleccionados. Los resultados se presentan a continuación:

- **Dataset Hogar-Adulto**

SVC

```
Fold 5 of 5
****Results****
Accuracy: 65.7221%
              precision    recall  f1-score   support

           0       0.00      0.00      0.00       375
           1       0.66      1.00      0.79       719

   micro avg       0.66      0.66      0.66      1094
   macro avg       0.33      0.50      0.40      1094
weighted avg       0.43      0.66      0.52      1094
```

RandomForestClassifier

```
Fold 5 of 5
****Results****
Accuracy: 65.7221%
              precision    recall  f1-score   support

           0       0.00      0.00      0.00       375
           1       0.66      1.00      0.79       719

   micro avg       0.66      0.66      0.66      1094
   macro avg       0.33      0.50      0.40      1094
weighted avg       0.43      0.66      0.52      1094
```

GradientBoostingClassifier

```
Fold 5 of 5
****Results****
Accuracy: 65.7221%
              precision    recall  f1-score   support

           0       0.00      0.00      0.00       375
           1       0.66      1.00      0.79       719

   micro avg       0.66      0.66      0.66      1094
   macro avg       0.33      0.50      0.40      1094
weighted avg       0.43      0.66      0.52      1094
```

- **Dataset Hogar-Niño**

SVC

```
Fold 5 of 5
****Results****
Accuracy: 59.7701%
              precision    recall  f1-score   support

           0       0.65      0.02      0.04       671
           1       0.60      0.99      0.75       982

   micro avg       0.60      0.60      0.60      1653
   macro avg       0.62      0.51      0.39      1653
weighted avg       0.62      0.60      0.46      1653
```

RandomForestClassifier

```
Fold 5 of 5
****Results****
Accuracy: 59.4071%
              precision    recall  f1-score   support

           0       0.00      0.00      0.00       671
           1       0.59      1.00      0.75       982

   micro avg       0.59      0.59      0.59      1653
   macro avg       0.30      0.50      0.37      1653
weighted avg       0.35      0.59      0.44      1653
```

GradientBoostingClassifier

```
Fold 5 of 5
****Results****
Accuracy: 59.4071%
              precision    recall  f1-score   support

           0       0.00      0.00      0.00       671
           1       0.59      1.00      0.75       982

   micro avg       0.59      0.59      0.59      1653
   macro avg       0.30      0.50      0.37      1653
weighted avg       0.35      0.59      0.44      1653
```

Cuando entrenamos nuestros modelos con el conjunto de datos de entrada se hizo posible que el algoritmo fuese capaz de generalizar un concepto (positividad para H.pylori) de forma que al consultarle por un nuevo conjunto de datos desconocido éste fuese capaz de sintetizarlo, comprenderlo y devolvernos un resultado fiable dada su capacidad de generalización.

Matrices de Confusión

Se utilizó la técnica de Matrices de confusión con el fin de evaluar el rendimiento del algoritmo de los modelos construidos. Una matriz de confusión es comúnmente utilizada para evaluar un clasificador en base a un conjunto de datos de prueba para los cuales se conocen los valores reales. Los resultados se presentan a continuación:

Tabla 12 Matriz de confusión - Dataset Hogar-Adulto

Dataset Hogar-Adulto		H.pylori Predicho	
		No	Si
H.pylori Real	No	VN = 3	FP = 373
	Si	FN = 1	VP = 718

Tabla 13 Matriz de confusión - Dataset Hogar-Niño

Dataset Hogar-Niño		H.pylori Predicho	
		No	Si
H.pylori Real	No	VN = 20	FP = 651
	Si	FN = 11	VP = 971

En el lado izquierdo, se representa la clase Real para H.pylori (etiquetada como *SÍ o NO*), mientras que la parte superior indica la clase que se está prediciendo (nuevamente *SÍ o NO*).

Lo anterior, representa que para el Dataset Hogar-Adulto, el modelo fue capaz de identificar 718 (99.86%) instancias positivas *(Verdaderos positivos)* y 3 instancias negativas *(Verdaderos negativos)* de forma correcta. Por otra parte, falló en 373

instancias negativas reales *(Falsos positivos)* y en 1 instancia positiva real *(Falsos negativos)*.

Para el Dataset Hogar-Niño, el modelo fue capaz de identificar 971 (98.87%) instancias positivas *(Verdaderos positivos)* y 20 instancias negativas *(Verdaderos negativos)* de forma correcta, así como 651 instancias negativas reales *(Falsos positivos)* y 11 instancias positivas reales *(Falsos negativos)* de forma incorrecta.

Dicho lo anterior y dado que la métrica que nos interesa priorizar es el *Recall,* cualquiera de los 3 modelos construidos para el Dataset Hogar-Adulto nos permitiría identificar correctamente el 99,87% de los Verdaderos Positivos, con lo cual se obtuvo una tasa baja de Falsos Negativos (0.13%) predichos, conforme con los objetivos de inteligencia de negocios definidos.

Para el Dataset Hogar-Niño los modelos *RandomForest* y *Gradient Boosting* permitirían identificar correctamente el 98,87% de los Verdaderos Positivos; De igual manera que para el Dataset Hogar-Adulto, se logró una tasa baja de Falsos Negativos (1.13%) predichos.

Sobreajuste

Uno de los conceptos más importantes en Machine Learning es el *overfitting* o *sobreajuste* del modelo. De acuerdo con los resultados obtenidos en la segunda iteración, podemos decir que nuestros modelos pueden encontrarse sobreajustados. Comprender cómo un modelo se ajusta a los datos es muy importante para entender las causas de baja o alta precisión en las predicciones. Un modelo va a estar sobreajustado cuando vemos que se desempeña bien con los datos de entrenamiento, pero su precisión es notablemente más baja con los datos de evaluación; esto se debe a que el modelo ha memorizado los datos que ha visto y no puede generalizar correctamente las reglas para predecir los datos que no ha visto. De aquí la importancia de dividir el Dataset en dos conjuntos de

datos distintos (entrenamiento y pruebas) de la forma como se detalló en el plan de pruebas para posteriormente construir los modelos.

En términos generales, el sobreajuste en este caso en particular está relacionado con la complejidad de los algoritmos dada la cantidad de variables utilizadas en la construcción de los modelos (mayor tendencia a sobreajustarse a los datos) ya que se cuenta con mayor flexibilidad para realizar las predicciones haciendo que los patrones que encuentre, estén relacionados con el ruido (pequeños errores aleatorios) en los datos y no con la verdadera señal o relación subyacente.

No existe una regla general para establecer cuál es el nivel ideal de complejidad que le podemos otorgar a nuestro modelo sin caer en el sobreajuste, no obstante podemos valernos de algunas herramientas analíticas para intentar entender cómo el modelo se ajusta a los datos y reconocer el sobreajuste.

Capítulo 6

Evaluación y Validación

Objetivo III

Validar e implementar el modelo de analítica construido para la predicción del riesgo a cáncer de estómago en la población de estudio.

6.1. Resultados

6.1.1. Fase V Evaluación del modelo

En esta fase se evaluó el modelo teniendo en cuenta el cumplimiento de los criterios de éxito del problema así como la fiabilidad definida en los criterios de éxito de minería de datos. Como se detalló en el capítulo anterior, fue utilizada la herramienta de *matrices de confusión* para la interpretación de los resultados de los distintos modelos.

I. Evaluación de resultados

En esta etapa se evaluó el modelo y se valoraron los resultados en relación al objetivo del negocio: *Determinar los factores epidemiológicos, sociales y demográficos relacionados con la presencia de la bacteria H. pylori como factor principal de riesgo a desarrollar cáncer de estómago, en 8 municipios del departamento del Cauca.*

De acuerdo con los resultados obtenidos en los distintos modelos, se calculó la importancia de las variables evaluadas con relación a la variable decisión (positividad H.pylori) en función del entrenamiento del modelo. El resultado de este proceso nos permitió identificar aquellas variables que tenían mayor peso para la identificación del riesgo a cáncer de estómago por la infección de H.pylori.

La importancia de las variables se calculó con los modelos *Gradient Boosting* y *RandomForest*; Los resultados se presentan a continuación:

```
# Valores ponderados de 0 a 1

H11_AFRO           (0.070)
H14_POPAYÁN        (0.064)
A39_NUNCA          (0.058)
H92                (0.049)
H14_SAN SEBASTIÁN (0.044)
A39_1_POR_MES      (0.040)
H12_NAN            (0.040)
H14_EL TAMBO       (0.039)
A3                 (0.033)
...
```

El detalle del ejercicio de 'Importancia de variables' puede ser consultado en el Anexo 4 – Comandos de programación.

De acuerdo a lo anterior y con el apoyo del experto se interpretaron las variables de importancia de acuerdo a su categoría como se mencionan a continuación:

Condiciones sociodemográficas

Edad: La exposición crónica del H.pylori en el tiempo parece estar relacionada con el aumento de edad (Entre mayor edad, mayor tasa de infección en la población).
Grupo étnico afro descendientes con mayor riesgo a infección por H.Pylori.

Los municipios con mayor tasa poblacional como Popayán, El Tambo y San Sebastián parecen estar asociados con mayor tasa de prevalencia de la infección.

Síntomas dispépticos

La ausencia de síntomas como tos seca, reflujo gástrico y náuseas parecen no estar asociados como manifestaciones clínicas relacionadas con la infección por H.Pylori.

Hábitos saludables

Los hábitos saludables de limpieza como lavado de manos, lavado de dientes y el baño frecuente de los miembros de la familia parecen relacionarse con el riesgo a infección por H.Pylori como factor etiológico del cáncer de estómago.

Condiciones de Hogar

El hacinamiento dado por el número de personas que habitan en la vivienda y que comparten habitaciones o dormitorios en mayores proporciones, se encuentra asociado a la presencia de H.pylori en la población.

Condiciones de vivienda

La falta de servicio de alcantarillado en los hogares está relacionado con la disposición de aguas residuales, lo cual podría constituir rutas de contaminación fecal asociado a la presencia de H.pylori.

Saneamiento

La ausencia de saneamiento básico mediante la ubicación externa del hogar del servicio sanitario puede ser una fuente de contaminación relacionada con la presencia del H.pylori en el núcleo familiar.

Prácticas de uso y manejo del Agua

Los recipientes de plástico podrían ser un reservorio importante de H.pylori que facilite su transmisión como principal factor de riesgo al desarrollo de cáncer de estómago.

7. CONCLUSIONES Y FUTURO TRABAJO

La comprensión del negocio se convirtió en una de las fases más importantes previo a realizar los procesos de analítica, dado que con esta fue posible desarrollar unas bases más estructuradas frente al contexto de la problemática en salud, esclareciendo la exploración de la información mediante la caracterización, descripción y verificación de los datos en aras de construir y validar el modelo de predicción.

La identificación de métricas de evaluación permitieron validar y priorizar los mejores modelos de predicción mediante un mayor Recall o sensibilidad así como una tasa elevada de correctitud o accuracy, con el fin de ajustar los parámetros buscando aumentar la calidad y fiabilidad de los mismos.

Los modelos de predicción validados RandomForest y Gradient Boosting, para la identificación del riesgo a desarrollar cáncer gástrico mediante la predicción del H.pylori nos permitieron identificar correctamente aproximadamente el 99,9% de los Falsos Negativos, es decir aquellos individuos que el modelo predice como Negativos cuando en la realidad son Positivos.

A partir de la validación de los modelos RandomForest y Gradient Boosting, se logró determinar las variables con mayor importancia al riesgo de desarrollar la infección por H.pylori como son las condiciones sociodemográficas (edad, grupo étnico procedencia), ausencia de hábitos saludables (limpieza personal), condiciones de hogar (hacinamiento como factor favorable para la transmisión de la bacteria), y la falta de saneamiento como parte de las condiciones de vivienda. Por último las prácticas de uso y manejo del agua, como el almacenamiento de agua de consumo en recipientes plásticos como principales reservorios de la bacteria.

Pese a que la fuente de información contaba con un número elevado de instancias y variables colectadas lo que permitió la construcción de los modelos, es importante aclarar que se deben incentivar los esfuerzos para el levantamiento de una mayor cantidad de información con el fin de mejorar el entrenamiento y aprendizaje de futuros modelos que permitan predecir de forma más precisa, viable y confiable.

BIBLIOGRAFÍA

A. L. Buczak, & E. Guven. (2016). A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection. IEEE Communications Surveys Tutorials, 18(2), 1153–1176. https://doi.org/10.1109/COMST.2015.2494502

Arce, F. (n.d.). Big Data y Salud, un paradigma aplicado. 84.

Asaka, M., Takeda, H., Sugiyama, T., & Kato, M. (1997). What role does Helicobacter pylori play in gastric cancer? Gastroenterology, 113(6), S56–S60.

Barragán Ocaña, A. (2009). Aproximación a una taxonomía de modelos de gestión del conocimiento.

Bishop, C. M. (2006). Pattern recognition and machine learning. springer.

Brenner, H., Rothenbacher, D., & Arndt, V. (2009). Epidemiology of stomach cancer. Methods in Molecular Biology (Clifton, N.J.), 472, 467–477. https://doi.org/10.1007/978-1-60327-492-0_23

Caruana, R., & Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. Proceedings of the 23rd International Conference on Machine Learning, 161–168. Pittsburgh, Pennsylvania, USA: ACM.

Díaz Pérez, M., de Liz Contreras, Y., & Amador, S. R. (2009). El factor humano como elemento dinamizador del proceso empresarial en la gestión de la información y conocimiento. Revista Cubana de Información En Ciencias de La Salud (ACIMED), 20(5), 42–55.

Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. Neural Computation, 10(7), 1895–1923.

Dua, S., & Du, X. (2011). Data Mining and Machine Learning in Cybersecurity. Auerbach Publications.

Eom, B. W., Joo, J., Kim, S., Shin, A., Yang, H.-R., Park, J., … Nam, B.-H. (2015). Prediction Model for Gastric Cancer Incidence in Korean Population. PLOS ONE, 10(7), e0132613. https://doi.org/10.1371/journal.pone.0132613

Iida, M., Ikeda, F., Hata, J., Hirakawa, Y., Ohara, T., Mukai, N., … Ninomiya, T. (2018). Development and validation of a risk assessment tool for gastric cancer in

a general Japanese population. Gastric Cancer, 21(3), 383–390. https://doi.org/10.1007/s10120-017-0768-8

Jaramillo, A., & Paz, H. (2015). Aplicación de Técnicas de Minería de Datos para Determinar las Interacciones de los Estudiantes en un Entorno Virtual de Aprendizaje (Vol. 28).

Jemal, A., Bray, F., Center, M. M., Ferlay, J., Ward, E., & Forman, D. (2011). Global cancer statistics. CA: A Cancer Journal for Clinicians, 61(2), 69–90.

López Noreña, G. (2010). Sobre las sociedades de la información y la del conocimiento: críticas a las llamadas ciudades del conocimiento latinoamericanas desde el paradigma ecológico. Edición electrónica gratuita.

Martínez Moreno, L. P. (2016). Plan estratégico para el mejoramiento de los servicios médicos de la clínica hospital Kennedy del cantón Buena Fe, periodo 2015-2019 (B.S. thesis). Quevedo: UTEQ.

Mitra, S., & Acharya, T. (2005). Data mining: multimedia, soft computing, and bioinformatics. John Wiley & Sons.

Moine, I. J. M., Haedo, D. A. S., & Gordillo, D. S. (n.d.). Estudio comparativo de metodologías para minería de datos. 4.

Niakšu, O. (n.d.). CRISP Data Mining Methodology Extension for Medical Domain. 18.

Parsonnet, J., Friedman, G. D., Vandersteen, D. P., Chang, Y., Vogelman, J. H., Orentreich, N., & Sibley, R. K. (1991). Helicobacter pylori Infection and the Risk of Gastric Carcinoma. New England Journal of Medicine, 325(16), 1127–1131.

Rodríguez León, C., & García Lorenzo, M. M. (2016). Adecuación a metodología de minería de datos para aplicar a problemas no supervisados tipo atributo-valor. Revista Universidad y Sociedad, 8(4), 43–53.

Sebastiani, F. (2002). Machine learning in automated text categorization. ACM Comput. Surv., 34(1), 1–47.

Vega, C. A., Rosano, G., López, J. M., Cendejas, J. L., & Ferreira, H. (2012). Data Mining Aplicado a la Predicción y Tratamiento de Enfermedades. CISCI (Conferencia Iberoamericana En Sistemas, Cibernética e Informática).

Wirth, R. (2000). CRISP-DM: Towards a standard process model for data mining. Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining, 29–39.

Zhu, X., Ghahramani, Z., & Lafferty, J. (n.d.). Semi-Supervised Learning Using Gaussian Fields and Harmonic Functions. 8.

ANEXOS