

Universidad Tecnológica de Pereira

FACULTAD DE INGENIERÍAS EEFC

Proyecto de Grado:

ALGORITMOS DE CLASIFICACIÓN Y REDES COMPLEJAS PARA LA MEJORA DE LA CALIDAD DE LOS DATOS EN SALUD

Proyecto presentada por Jhonatan Barrera para el grado de
Magister en Ingeniería de Sistemas y Computación
de la Universidad Tecnológica de Pereira

Supervisado por:

Ph.D. Rafael R. Rentería
Universidad Nacional de Colombia
Sede Bogotá

ALGORITMOS DE CLASIFICACIÓN Y REDES COMPLEJAS PARA LA MEJORA DE LA CALIDAD DE LOS DATOS EN SALUD

*CASO DE ESTUDIO: PROYECTO DE "DESARROLLO DE CAPACIDADES
CT+I PARA INVESTIGACIÓN Y SIMULACIÓN DE POLÍTICAS PÚBLICAS EN
SALUD Y SEGURIDAD SOCIAL EN EL DEPARTAMENTO DE RISARALDA"*

Ing. JHONATAN STIVENS BARRERA ORDOÑEZ

MAESTRÍA EN INGENIERÍA DE SISTEMAS Y COMPUTACIÓN
LÍNEA: INTELIGENCIA ARTIFICIAL - CIENCIA DE DATOS
Grupo de investigaciones en economía, gestión y tecnologías en salud
GIA - Grupo de investigación en inteligencia artificial

UNIVERSIDAD TECNOLÓGICA DE PEREIRA
FACULTAD DE INGENIERÍAS EEFC
PEREIRA, RISARALDA

7 de Agosto de 2018

*A mis padres por su cálido amor y permitirme siempre soñar,
a mi abuela por la alegría que siempre me da, eres mi ángel,
al amor de mi vida por su apoyo y paciencia en este camino!*

Agradecimientos

Gracias a mis padres, abuela, hermano, a mí pareja, son ellos mí familia y las personas que me dan fuerza, me inspiran a luchar por mis pasiones, romper toda barrera y hacer que cada esfuerzo valga la pena.

A mi supervisor Rafael Ricardo Rentería, en quien no solo encontré un excelente tutor sino un excelente ser humano y un gran amigo, el cual me ha dado aportes valiosos para el desarrollo de este trabajo. A el Ingeniero Jorge I. Rios por todas las etapas de formación en las que me ha acompañado y por su llamado a superarme constantemente en este apasionante mundo, e incentivarme siempre por aprender e investigar.

Agradezco de manera especial a la Gobernación de Risaralda por apostar en la formación de Talento Humano Calificado para el desarrollo de nuestro departamento y país, a la Fundación Salutia por abrirme las puertas de su organización y permitirme ser parte de este proyecto.

De igual forma agradezco a los maestros que han pasado por mi vida y han dejado su huella, Luz Angela Cardona, Luz Mery Castañeda, Mauricio A. Álvarez, Omar Ivan Trejos, Saulo Torres R., John H. Osorio, Hugo H. Morales, entre muchos otros seres excepcionales que acompañaron hasta ahora mi formación.

A mis compañeros y amigos con los que recorrí este proceso de formación. Inventors son uno de los mejores equipos con el que he trabajado. Son fantásticos!

Índice general

Agradecimientos	VII
Resumen	XIX
1. Introducción	1
1.1. Contexto	1
1.2. Problema Objeto de Estudio	2
1.2.1. Descripción	2
1.2.2. Pregunta de Investigación	2
1.3. Justificación	3
1.4. Objetivos y Alcance de la Investigación	4
1.4.1. Objetivo General	4
1.4.2. Objetivos Específicos	4
2. Marco Conceptual	5
2.1. Calidad en Datos	5
2.2. Tidy Data	5
2.3. Selección de Características	6
2.4. Clasificación	6
2.5. Clasificación Supervisada y No Supervisada	6
2.6. Machine Learning	7
2.7. Teoría de la Información	7
2.7.1. Elementos del Sistema de Comunicación	8
2.7.2. Entropía y Teoría de la Información	8
2.8. Redes Complejas	9
2.8.1. Conceptos Generales de Redes	10
3. Metodología	11
4. Fuentes de Información	13
4.1. Base de Datos Única de Afiliados (BDUA)	13
4.2. Registro Especial de Prestadores de Servicios de Salud (REPS)	14
4.3. Precios reportados de medicamentos	18
4.4. Eventos de vigilancia en salud pública	18
4.5. Registro Individual de Prestación de Servicios de Salud (RIPS)	19
4.6. Sistema de Información Hospitalaria (SIHO)	22

4.7. Información contable IPS publicas	23
4.8. Registro Único de Afiliados (RUAF)	24
4.9. Formulario Único Territorial (FUT)	26
5. Fuentes Seleccionadas y Estructura	27
5.1. Estructuración de la información presentada en las fuentes	27
5.2. Fuentes de información seleccionadas	28
5.2.1. BDUA	28
5.2.2. RIPS	29
5.2.3. CIE-10	29
6. Métricas de Rendimiento	33
6.1. Accuracy	33
6.2. Confusion Matrix	34
6.3. Precision	34
6.4. Recall	34
6.5. F-Score	35
7. Clasificadores	37
7.1. Random Forest	37
7.1.1. Esquema del algoritmo Random Forest	37
7.1.2. Observaciones	38
7.1.3. Resumen del algoritmo	38
7.2. Naïve Bayes	38
7.2.1. Teorema de Bayes	39
7.2.2. El clasificador Bayesiano Ingenuo	40
7.3. Decision Tree	40
8. Entrenamiento Clasificadores	43
8.1. Naïve Bayes	43
8.1.1. Gaussian Naïve Bayes	44
8.1.2. Bernoulli Naïve Bayes	45
8.1.3. Multinomial Naïve Bayes	48
8.2. Decision Tree	50
8.3. Random Forest	51
8.4. Análisis de rendimiento	53
9. Estructuración de la Red Compleja	55
9.1. Integración de RIPS y BDUA	57
9.2. Construcción de la red compleja	60
9.2.1. Herramienta de Análisis y Visualización	61
10. Conclusiones y Recomendaciones	63
10.1. Conclusiones	64
10.2. Recomendaciones	65

ÍNDICE GENERAL

XI

Bibliografía

66

Índice de figuras

8.1. Dispersión de las observaciones para las clases Hombre, Mujer y General en el dataset	43
8.2. Dispersión de las observaciones para las clases Hombre y Mujer en un rango de 0 a 1.	44
8.3. Modelo Naïve Bayes para las clases Hombre y Mujer	45
8.4. Métricas obtenidas de la clasificación con el <i>Gaussian Naïve Bayes</i>	46
8.5. Visualización de la clasificación con el <i>Gaussian Naïve Bayes</i>	46
8.6. Métricas obtenidas de la clasificación con el <i>Bernoulli Naïve Bayes</i>	47
8.7. Visualización de la clasificación con el <i>Bernoulli Naïve Bayes</i>	47
8.8. Métricas obtenidas de la clasificación con el <i>Multinomial Naïve Bayes</i>	48
8.9. Visualización de la clasificación con el <i>Multinomial Naïve Bayes</i>	49
8.10. Visualización de la clasificación con <i>Decision Tree</i> , profundidad 1, 2 y 3.	49
8.11. Métricas obtenidas de la clasificación con el <i>Decision Tree</i>	50
8.12. Visualización de la clasificación con el <i>Decision Tree</i>	51
8.13. Métricas obtenidas de la clasificación con el <i>Random Forest</i>	52
8.14. Visualización de la clasificación con el <i>Random Forest</i>	52
9.1. Muestra de registros RIPS en régimen contributivo.	56
9.2. Red para los 10 registros presentados en la Figura 9.1.	56
9.3. Frecuencia de ocurrencias para el Cuadro 9.1.	59
9.4. Frecuencia de ocurrencias para el Cuadro 9.2.	60
9.5. Diagnoses Data Analysis. Prototipo de herramienta para análisis y visualización de diagnósticos.	61

Índice de cuadros

4.1. Reporte BDU A	14
4.2. Prestadores de servicios de salud	15
4.3. Servicios de salud prestados por sede.	16
4.4. Capacidad instalada por sede	17
4.5. Precios reportados por medicamento.	18
4.6. Eventos reportados	19
4.7. Consultas	20
4.8. Procedimientos	21
4.9. Urgencias	21
4.10. Hospitalización	22
4.11. Información reportada por subsistema RUAF	25
6.1. Matriz de Confusión	34
8.1. Métricas de rendimiento para los clasificadores planteados.	53
9.1. Ocurrencias Negativas del CIE-10 en datos del régimen subsidiado . . .	58
9.2. Ocurrencias Negativas del CIE-10 en datos del régimen contributivo. . .	60

Lista de codigos

1.	Fragmento de código para clasificación con <i>Gaussian Naïve Bayes</i> . . .	44
2.	Fragmento de código para clasificación con <i>Bernoulli Naïve Bayes</i> . . .	45
3.	Fragmento de código para clasificación con <i>Multinomial Naïve Bayes</i> .	48
4.	Fragmento de código para clasificación con <i>Decision Tree</i>	50
5.	Fragmento de código para clasificación con <i>Random Forest</i>	51

Resumen

En los últimos años el incremento de los datos digitales se ha manifestado de manera vertiginosa, esta información ha sido almacenada por todo tipo de organización desde su puesta en operación, contiene la huella de su trayectoria, contemplando sus éxitos al igual que sus fracasos. Sin embargo, por mucho tiempo esta información permaneció en gran parte como un adorno en un viejo anaquel, desaprovechándose su capacidad en cuanto a la mejora de procesos, operaciones e incluso en la toma de decisiones. No se trata de tomar toda la información almacenada y que de manera mágica se encuentre la solución a una pregunta o resulten de igual forma tópicos que permitan encaminar los planes estratégicos al éxito de los objetivos propuestos, ya que esta información no viene predispuesta para suministrar inmediatamente los insumos requeridos en los análisis pertinentes, para esto la información debe ser depurada y verificada de manera que se disponga de una calidad adecuada en la misma, desde la cual sea confiable la información con la que se pretende realizar análisis. Los datos que se registran en el sistema de salud presentan estas mismas dificultades y riquezas.

Para la administración del Sistema de Salud Pública se cuenta con un amplio abanico de bases de datos dispuestas para registrar toda la información competente en su proceso. Se tomaron para el proyecto las bases de datos BDUA (Base de Datos Única de Afiliados), RIPS (Registros Individuales de Prestación de Servicios de Salud) y el CIE-10 (Clasificación Internacional de Enfermedades, en su décima revisión). Las dos primeras hacen referencia al compendio de los datos de todos los afiliados al sistema de salud y el registro requerido para el seguimiento al Sistema de Prestaciones de Salud. El CIE-10 constituye uno de los estándares internacionales usados para la morbilidad y mortalidad en el mundo. A partir de la información que suministra BDUA y RIPS se obtiene una sábana de datos que integra componentes indicados como relevantes al análisis que se desee realizar, esta sábana es el resultado de un proceso de validación realizado a través de una maquina clasificadora que integra la clasificación de enfermedades dispuesta en el CIE-10 y que pensando en la posterior toma de decisiones, suministra registros depurados que son almacenados y consecutivamente transformados en una red, desde la cual se obtiene los diagnósticos que más se presentan en la población y la interacción entre estos desde la perspectiva de un mismo paciente. Adicionalmente la sábana de datos, como se sugirió puede ser suministro para diversos análisis posteriores.

Capítulo 1

Introducción

1.1. Contexto

El Departamento de Risaralda, en el marco de su Plan Departamental de Desarrollo 2012-2015: “*Risaralda: unida, incluyente y con resultados*”, tomo acciones decisivas que condujeran al logro de sus metas de gobierno. Dentro de esta estrategia se ha propuesto fortalecer las capacidades territoriales en ciencia, tecnología e innovación (CT+I) que permitan inducir un impulso a la gestión en el sector salud y en particular el fortalecimiento de su nivel de rectoría “hacia la gobernanza del sistema de salud territorial para mejores resultado”.

Reconociendo las dificultades presentes en la operación del Sistema de Salud la Gobernación de Risaralda formulo una propuesta que permite afrontar de manera adecuada los problemas derivados de la disfuncionalidad del sistema, se propuso entonces el proyecto de investigación “Desarrollo de capacidades CT+I para investigación y simulación de políticas públicas en salud y seguridad social en el Departamento de Risaralda”, el cual es ejecutado en dos etapas de 36 meses cada una -iniciando en 2016 y terminando en 2020- y en el que se plantea como objetivo contribuir al incremento de la capacidad del Sistema de Salud en el Departamento, mediante la producción de “conocimiento científico aplicado” a través de actividades de desarrollo e investigación [I+D] que conlleven a la obtención de conocimiento y desarrollo de tecnologías, modelos e instrumentos que permitan producir, proveer y utilizar evidencia científica desde la que se propenda por incidir y potenciar la “estructura de decisiones” de sus agentes y que contribuyan con el logro de los fines del sistema de salud (seguridad social en salud, y salud pública), esperando que esto aporte al mejoramiento del bienestar y calidad de vida de las personas de nuestro territorio, con desarrollos escalables a nivel regional y nacional.

En el transcurso de la ejecución del proyecto antes mencionado, se han realizado trabajos que desembocan en la materialización del mismo, redundando en la formulación de protocolos y guías referenciadas en el plano tecnológico. La actividad con código [1506873] “*Gestión de administración de datos de todos los proyectos del programa de investigación*”, y que ha resultado en el documento titulado: “Protocolo para bases de

datos fuentes secundarias” [1], presenta las bases de datos y fuentes de información de carácter administrativo, económico y asistencial que han sido identificados de utilidad.

Es a partir de las necesidades planteadas por la Secretaria de Salud de Risaralda y teniendo como referente de fuentes de información implicadas en el proceso operativo y de gestión de la entidad las que se presentan en el documento “Protocolo para bases de datos fuentes secundarias” [1], que se propone el desarrollo de este proyecto, con la visión de aportar elementos relevantes para la toma de decisiones.

1.2. Problema Objeto de Estudio

1.2.1. Descripción

En el proyecto se han identificado necesidades y problemas concernientes a las fuentes de información, tales como:

- La existencia de múltiples fuentes de información, que contienen datos incompletos o a veces inexistentes, los cuales deben ser sujeto de validación y/o transformación antes de ser usados en un proceso de toma de decisiones.
- La necesidad de gestionar los datos a través de metodologías y herramientas tecnológicas que permitan clasificar, organizar, almacenar y utilizar adecuadamente la información involucrada en la toma de decisiones.
- Se hace necesario establecer y mantener un modelo metodológico para diseñar e implementar una Bodega de Datos (Data Warehouse) y bodegas de datos especializadas (Data Mart).

Dentro de estos requerimientos hay que destacar la importancia que tienen los datos a la hora de ser usados en la toma de decisiones, pues es a partir de estos que se puede obtener un conocimiento que conlleve a descubrir aspectos relevantes para la organización, en este caso, conocimiento para la formulación de políticas de salud pública.

De manera ideal, los datos que se almacenan no deberían contener errores, sin embargo, esto no es lo que ocurre en la realidad, los datos pueden presentar diferentes tipos de problemas como lo son la duplicidad, incompletitud, conformidad, consistencia, entre otros. Los datos presentados en salud no son excepción a esta problemática.

1.2.2. Pregunta de Investigación

¿De qué manera se puede mejorar la calidad de los datos de diagnósticos provenientes del sistema de salud pública para la toma de decisiones?

1.3. Justificación

*"Datos, datos, datos. No puedo hacer ladrillos sin arcilla".
Sherlock Holmes.*

Según un estudio realizado por el Mckinsey Global Institute (MGI) y la McKinsey's Business Technology Office en el año 2011, la gestión y el análisis del ingente volumen de datos que se genera a cada momento es clave para que las organizaciones sean competitivas, crezca su productividad e innovación. Este estudio además destaca que una correcta utilización del Big Data supone un incremento de hasta 60 % del margen de rentabilidad de las organizaciones. De igual manera, el análisis de datos en el sistema de salud estadounidense ayudaría por ejemplo a incrementar la calidad y eficiencia, generando beneficios de más de 300 mil millones de dólares, de los cuales un 8 % corresponden a reducción del gasto sanitario [2].

El volumen y el detalle de la información que obtienen las organizaciones aumenta de manera vertiginosa, para el 2011 se habían creado suficientes datos digitales como para formar una pila de DVD que alcancen dos veces la distancia de la tierra a la luna [3]. Para el 2015 se esperaba que la información digital de los hospitales, provenientes mayoritariamente de los exámenes clínicos de diagnóstico por imagen, aumentara en 665 terabytes por día, y que esta información fuese útil para encontrar curas y salvar vidas [4].

Una investigación realizada en 2017 por la firma de análisis IDC¹ y patrocinada por Seagate², revela que en menos de una década el volumen de datos total a nivel mundial podría aumentar en un factor de 10 con respecto al año anterior, una estimación de 163 Zbyte³ para el año 2025 [5], de igual manera se estima que para ese año las empresas serán responsables del 60 % de la información a nivel mundial.

Las organizaciones cuentan con montañas de datos que sin ser comprensibles no son mas que observaciones aleatorias. Esta comprensión se logra al momento de obtener el conocimiento que contienen los datos. La generación de conocimiento es una herramienta que indiscutiblemente mejora la intención de analizar grandes fuentes de datos dispersos; para que este proceso sea exitoso se requiere del tratamiento de los mismos, de manera que se pueda establecer con cierta certidumbre la calidad de estos y la completitud de los mismos.

Se calcula que los científicos de datos pueden pasar entre el 50 y el 80 % de su tiempo preparando datos digitales antes de que estos se puedan explorar en búsqueda de piezas relevantes [6]. Para que los datos puedan ser útiles, se requiere de una validación de los

¹International Data Corporation es el principal proveedor mundial de inteligencia de mercado, servicios de asesoramiento y eventos para los mercados de tecnología de la información, telecomunicaciones y tecnología de consumo. www.idc.com/about

²Seagate Technology es un importante fabricante estadounidense de discos duros, fundado en 1979.

³Un Zbyte equivale a un trillón de gigabytes.

mismos.

Como ya se comento en este documento, el propósito de este trabajo esta centrado en mejorar la calidad de los datos diagnósticos registrados en el sistema de salud publica, y proporcionar a partir de estos, elementos o herramientas que permitan al tomador de decisiones comprender hitos que conlleven a establecer puntos críticos que se tengan en cuenta para la generación de nuevas políticas publicas en salud con el fin de prevenir y controlar las dinámicas de morbilidad de la población.

En función de lograr el propósito de este trabajo se requiere validar bajo algún parámetro los diagnósticos encontrados en el sistema de salud contra una fuente oficial. Se plantea realizar esta validación mediante la clasificación de los diagnósticos, lo que permitirá tener una base solida de datos diagnósticos de entrada con una calidad para el posterior estudio de la dinámica de morbilidad, a partir de la cual se pueden obtener hitos para la toma de decisiones, en este sentido, la teoría de redes complejas proporciona los elementos necesarios que permiten modelar la realidad de un sistema complejo.

1.4. Objetivos y Alcance de la Investigación

1.4.1. Objetivo General

Implementar algoritmos de clasificación y teoría de redes complejas para el tratamiento de la calidad de datos diagnósticos en salud.

1.4.2. Objetivos Específicos

- Identificar y analizar fuentes de información provenientes de los sistemas de salud pública.
- Estructurar los datos de las fuentes de información más importantes en la definición de la morbilidad del departamento.
- Implementar técnicas computacionales para el mejoramiento de la calidad de los datos.

Capítulo 2

Marco Conceptual

2.1. Calidad en Datos

Hoy en día son más las organizaciones que se unen a la estrategia de tomar sus decisiones en base al conocimiento derivado de los datos que tienen a disposición en sus DBs, mediante el enfoque llamado *Inteligencia de Negocio (Business Intelligence)*, y no a partir de la subjetividad que puede tener el juicio de los directivos. Dada la importancia de este proceso, se hace necesario que los datos analizados no contengan errores, o que por lo menos estos se presenten en la menor cantidad posible.

Los datos sucios en la toma de decisiones pueden llevar a conclusiones erróneas, que representan pérdidas en tiempo, presupuesto y credibilidad. La firma Gartner¹ afirmó en 2007 que los problemas en calidad de los datos llevan a costos importantes y pérdida de oportunidades, no obstante, las organizaciones ya estaban descubriendo el impacto que la calidad de los datos aporta a la generación de sus estrategias en ventas, marketing, proyección de presupuesto, procesos de producción, entre otras [7]

2.2. Tidy Data

¿Contienen nuestros “data set” información ordenada o desordenada?

Esta pregunta está estrechamente ligada con el problema de la preparación de los datos en el análisis de los mismos. Se conoce que el 80 % del tiempo invertido en un proyecto de Ciencia de Datos se emplea en la preparación de los datos [6], y el 20 % en el estudio o modelado.

La preparación de los datos incluye tratar con datos faltantes, valores registrados como 0, datos que difieren de los demás (valores atípicos). Estas inconsistencias podrían ser medidas de error o errores de transcripción, de manera que podría pensarse

¹Gartner es una organización referente en investigación y asesoría indispensable para la planificación crítica de las organizaciones del futuro www.gartner.com/en/about

en ignorarlos en el análisis puesto que puede que no aporten valor al trabajo que se quiera realizar, en caso contrario, podría tratarse de valores significativos, de manera que se deben conservar. Estos son los elementos esenciales que se tiene en cuenta a la hora de analizar datos, de igual manera existe un paso fundamentalmente diferente que transforma los datos para facilitar el procesamiento de estos a través de herramientas de aprendizaje de máquina y/o estadísticas.

En la propuesta de transformación de datos que propone Wickham [8] se definen dos tipos de conjuntos de datos, el conjunto de datos ordenados y que están estructurados en tablas donde las filas son observaciones y las columnas son variables. Existe una tercera condición en la que se dice que cada una de las tablas representa una unidad de observación. Cualquier conjunto de datos que no esté normalizado bajo estas tres condiciones es un conjunto desordenado.

2.3. Selección de Características

La selección de características nos permite reconocer los factores que son mas relevantes de una fuente de información. Se podría tener por ejemplo una base de datos (DB) con una alta dimensionalidad en la que se encuentren características irrelevantes, redundantes o engañosas, de manera que están incrementando innecesariamente el tamaño del espacio de búsqueda y dificultando procesos de aprendizaje de maquina. De manera que la selección de características puede ser usada para disminuir la complejidad y representar de manera simple los datos, al trabajar con las características que realmente son relevantes [9].

2.4. Clasificación

La clasificación es usada en el aprendizaje de máquina, tiene como fin estimar un modelo matemático que separa una entidad usando información que se suministra con anterioridad. La clasificación se puede realizar de diversas formas, entre ellas la clasificación de manera supervisada o no supervisada, aprendizaje por refuerzo, aprendizaje multi-tarea, entre otras.

2.5. Clasificación Supervisada y No Supervisada

Metodologías que se encargan de etiquetar las características relevantes de un entorno, para construir un umbral entre diferentes clases a identificar. La clasificación no supervisada en criterios donde no existe un “a priori” durante el entrenamiento, de manera que su codificación es compleja [10]. Por otra parte, en la clasificación supervisada se tiene un conocimiento previo de los datos de inicialización, por lo tanto, el proceso tiende a tener un tiempo de clasificación menor, puede haber errores de clasificación.

2.6. Machine Learning

El aprendizaje de máquina (ML) es una rama de la ciencia de la computación, que desarrolla algoritmos capaces de aprender automáticamente a través de experiencias o ejemplos. El sistema aprende de los cambios del ambiente y se adapta a este [11]. Cualquier metodología de ML debe seleccionar un modelo que sea candidato y estimar los parámetros del modelo usando datos disponibles y los algoritmos de aprendizaje. De esta manera se dice que el usuario escoge un modelo empírico el cual emplea un algoritmo de aprendizaje para estimar los parámetros del modelo [12].

2.7. Teoría de la Información

La teoría de la información conocida originalmente como teoría matemática de la comunicación fue propuesta a finales de la Segunda Guerra Mundial por Claude Shannon² y Warren Weaver³. Plantea las leyes matemáticas que dirigen el procesamiento y la transformación de la información, así como su transmisión y medición.

Los trabajos que permitieron el desarrollo de esta teoría se remontan a la década de 1910 con Andrey A. Márkov⁴, Ralph V. L. Hartley⁵ (1967), Alan Turing⁶ (1936). En 1948 Claude E. Shannon publica su “A Mathematical Theory of Communication”, quien junto a Warren Weaver publican en 1949 “The Mathematical Theory of Communication”.

Como indica Shannon[13] el problema fundamental de la comunicación es reproducir en un lugar exactamente el contenido de un mensaje que ha sido generado en otro lugar. Los mensajes están correlacionados acorde a algunos sistemas con ciertas entidades físicas o conceptuales, y el mensaje real es seleccionado de un conjunto de mensajes posibles. El sistema debe estar diseñado para operar cada una de las selecciones, no solo la que se elegirá, pues esto se desconoce en el momento del diseño.

Existen diversas definiciones para información, respecto a quien use el término o el contexto en que lo haga. Para nuestro trabajo interesa la información que es almacenada de manera digital, un ejemplo de esto se ve en el uso de DBs.

²C. E. Shannon fue un Ingeniero Electricista y Matemático, creador de la Teoría de la Información y padre de la Comunicación Digital.

³Weaver fue un matemático quien propuso ideas para la masificación de la teoría de Shannon.

⁴Andrey Markov fue un matemático que dedicó gran parte de su vida al estudio de procesos estocásticos, de su trabajo se conocen principalmente las cadenas y procesos de Markov.

⁵Hartley, R. V. L., “Transmission of Information”, Bell System Technical Journal, 1928, p.535

⁶Alan Turing introdujo el concepto de máquina de Turing en su trabajo “On computable numbers, with an application to the Entscheidungsproblem” publicado por la Sociedad Matemática de Londres en 1936.

2.7.1. Elementos del Sistema de Comunicación

El modelo matemático de comunicación de Shannon es un sistema compuesto principalmente por:[13]

- Fuente de información
Todo aquello que pueda emitir un mensaje o secuencia de mensajes es considerado como una fuente de información. Una DB podría ser una fuente y su contenido la información que se quiere transmitir.
- Transmisor
El transmisor opera sobre el mensaje para representar la información de manera adecuada a la transmisión que se dará sobre el canal.
- Canal
Es el medio mediante el cual se transmite la información desde el transmisor al receptor.
- Receptor
Realiza la operación inversa que el transmisor, de esta manera se recupera la información original.
- Destino
Es la "entidad" a quien va dirigida la información.

Shannon en su teoría define principalmente una medida cuantitativa de información, con la que se busca atacar problemas relacionados, como el teorema de la máxima capacidad de un canal. En su trabajo además propone la idea de convertir los datos a símbolos bi-estables 0 y 1, bits de información que se transmiten a través de un medio. Durante la transmisión se puede presentar errores, o ser introducido ruido, se busca reducir o corregir esto. Como se mencionó anteriormente, la información binaria recibida es representada de nuevo al medio para su entrega. El primer teorema presentado establece que la velocidad de la información está basada en su entropía.

El modelo de la teoría de la información busca determinar la forma más económica, rápida y segura de codificar un mensaje, evitando que la presencia de algún ruido inestabilice su transmisión. La cantidad de información perteneciente a un mensaje es un valor matemático definido y medible.

2.7.2. Entropía y Teoría de la Información

El termino entropía fue acuñado por Rudolf Clausius en 1865 en su trabajo "The Mechanical Theory of Heat". El concepto oculto sería sumamente importante para el desarrollo de la termodinámica y la mecánica estadística con el trabajo de J. W. Gibbs y L. Boltzmann a finales del siglo XIX. Los estudios de la entropía son una base fundamental en la teoría de la información de Shannon, existen de igual forma propiedades con respecto a la entropía de variables aleatorias debidas a A. N. Kolmogorov quien se

inspiró en trabajo de Shannon[14]. La teoría de la información indica que el nivel de la información de una fuente se puede medir a partir de la entropía de esta.

Dada una fuente de información M que emite mensajes, en donde para cada una existe una probabilidad de ocurrencia P_i que depende del mensaje, para codificar los mensajes de la fuente se usa la menor cantidad de bits para los mensajes más probables y mayor cantidad en caso contrario, de manera que el promedio de bits requeridos para la codificación de los mensajes es menor a la cantidad de bits promedio de los mensajes genuinos. Ahora, si L_i representa la longitud del mensaje codificado, se tiene que la longitud promedio de todos los mensajes codificados de la fuente de información (entropía de la fuente) es:

$$H = \sum_{i=0}^n P_i L_i \quad (2.1)$$

La entropía de la fuente determina la máxima comprensión de un conjunto de datos. El objetivo de la comprensión es encontrar los L_i en función de los P_i donde H es minimizado, se tiene entonces:

$$H = \sum_{i=0}^n P_i f(P_i) \quad (2.2)$$

Shannon demuestra que H es mínimo cuando $f(P_i) = \log_2(P_i)$, entonces:

$$H = - \sum_{i=0}^n P_i \log_2(P_i) \quad (2.3)$$

2.8. Redes Complejas

En los últimos años se ha visto como las redes del mundo han ido creciendo de manera compleja, al punto que el análisis clásico mediante la teoría de grafos es un proceso que se dificulta cada vez más. De esta manera se ha manifestado la necesidad de combinar los métodos de la teoría de grafos con técnicas matemáticas de otras disciplinas científicas como el aprendizaje automático, la estadística y la teoría de la información [15].

En la revolución de las tecnologías de la información y la comunicación, las innovaciones se han desarrollado no solo en los campos clásicos de la comunicación inalámbrica, la ingeniería informática, la bioinformática, así como en áreas nuevas como las redes de sensores inalámbricos (WSN), polvo inteligente, control de tráfico aéreo, interpretación de redes biológicas, entre otras. Todos estos problemas hacen referencia a sistemas complejos y optimización a gran escala, que requieren de una gran cantidad de esfuerzo

computacional no permitido.

Los sistemas complejos son el resultado de la interacción de una gran cantidad de entidades, más bien simples. De esta manera la complejidad radica en la arquitectura de las interacciones que puede ser descrita mediante una red, esta complejidad exige soluciones robustas y adaptables, que pueden desarrollarse a partir de modelos matemáticos que describen como las reglas locales originan el comportamiento global [16].

2.8.1. Conceptos Generales de Redes

Esta sub-sección está planteada a partir de [17]

Una red está definida por una 3-tupla $G = (V, E, f)$, donde V es un conjunto finito de nodos, E es el conjunto de aristas posibles en V , y f es un mapeo que asocia algunos elementos de E a un par de elementos de V , si $v_i \in V$ y $v_j \in V$ entonces $f : e_p \rightarrow [v_i, v_j]$ y $f : e_q \rightarrow [v_j, v_i]$. Una red ponderada se define remplazando el conjunto de aristas E por un conjunto de aristas ponderadas $W = \{w_1, w_2, \dots, w_m\}$ tal que $w_i \in R$. Entonces una red ponderada es definida por la 3-tupla $G = (V, W, f)$.

Una forma común de la topología de una red G es representarla mediante una matriz de adyacencia. Esta es una matriz cuadrada A cuyas entradas están definidas por:

$$A_{ij} = \begin{cases} 1 & \text{si } i, j \in E \\ 0 & \text{en otro caso} \end{cases} \quad (2.4)$$

Otra importante representación en forma de matriz de una red es mediante la matriz Laplaciana L , que es el análogo discreto del operador Laplaciano. Las entradas de esta matriz están definidas por:

$$L_{uv} = \begin{cases} -1 & \text{si } u, v \in E, \\ k_u & \text{si } u = v, \\ 0 & \text{en otro caso} \end{cases} \quad (2.5)$$

Se designa por Δ la matriz de incidencia de la red, que es una matriz $n * m$ cuyas filas y columnas representan los nodos y los bordes de la red, respectivamente, tal que:

$$\Delta_{ue} = \begin{cases} +1 & \text{si } e \in E \text{ entrante al nodo } u, \\ -1 & \text{si } e \in E \text{ es una salida del nodo } u, \\ 0 & \text{en otro caso} \end{cases} \quad (2.6)$$

Capítulo 3

Metodología

El desarrollo de este proyecto de grado se ha realizado de manera incremental, en primera instancia se identificaron las fuentes de información provenientes del sistema de salud pública, las cuales fueron desglosadas una a una para su comprensión y selección como recurso a trabajar durante la investigación. A partir de la revisión de las fuentes de información se seleccionan las fuentes que corresponden a BDUA y RIPS, desde las cuales finalmente se tiene una sábana de datos depurada como se comentara a continuación.

El cruce de estas dos fuentes de información por el campo dispuesto para la identificación de cada uno de los individuos fue el primer paso que se realizó, de este cruce se tiene una primera sábana de datos en la cual se evidencia las atenciones realizadas en el departamento tanto para los individuos que residen de manera permanente como los que fueron atendidos pero que corresponden a habitantes de otros departamentos, esta información es relevante para los tomadores de decisiones.

Los campos que se utilizaron de manera directa para la siguiente etapa de la investigación corresponden a: el *id* de identificación del individuo, el *sexo*, y los cuatro campos donde se registran los diagnósticos (uno principal y tres secundarios). La selección de estos campos se hizo con el fin de validar los registros de diagnósticos almacenados contra la Clasificación Internacional de Enfermedades en su décima revisión *CIE-10* respecto a las 3 categorías que se presentan en este, y que corresponden a diagnósticos para hombres, mujeres y diagnósticos presentados en ambos géneros.

Para realizar la validación de los códigos CIE-10 se construyeron máquinas de clasificación que fueron evaluadas con el fin de seleccionar la que obtenga un mejor desempeño clasificando los diagnósticos. Esta clasificación permite marcar los registros de la sábana de datos para indicar cuáles diagnósticos han sido mal registrados. En la etapa final se construye una red compleja a partir de estos diagnósticos depurados, desde la cual se puede calcular la fortaleza de los nodos (con qué persistencia se presenta un diagnóstico) y la interacción entre diagnósticos, flujos que dinamizan la emergencia de morbilidad y mortalidad en el departamento.

Los siguientes capítulos abordan el desarrollo de este trabajo, desde la identificación de las fuentes de información disponibles, la estructuración de esta información, el

entrenamiento de los clasificadores y su respectiva evaluación a través de métricas de rendimiento y finalmente la construcción de una red compleja a partir de los diagnósticos.

Capítulo 4

Fuentes de Información

A continuación, se presenta un resumen de la información de reporte regular consignada en el catálogo de información perteneciente al documento [1506873] *Gestión de administración de datos de todos los proyectos del programa de investigación que corresponde al Protocolo para bases de datos fuentes secundarias* [1].

4.1. Base de Datos Única de Afiliados (BDUA)

BDUA presenta los vínculos que tienen los individuos con las entidades de salud que permiten la afiliación al Sistema de Salud Colombiano. Se actualiza semanalmente a través de las Entidades Administradoras de Planes de Beneficios (EAPB). Según el numeral 1 del artículo 121 de la Ley 1438 de 2011 y en conformidad con el numeral 17 del artículo 6 del Decreto 2462 de 2013 son EAPB “las entidades promotoras de salud del régimen contributivo y subsidiado, las empresas solidarias, las asociaciones mutuales en sus actividades de salud, las cajas de compensación familiar en sus actividades de salud, las actividades de salud que realicen las aseguradoras, las entidades que administren planes voluntarios de salud, las administradoras de riesgos profesionales en sus actividades de salud, las entidades pertenecientes al régimen de excepción de salud y las universidades en sus actividades de salud”. Todas estas EAPB reportan a ADRES, encargado de la administración de la base de datos.

ADRES dentro de sus responsabilidades en la administración de la BDUA debe realizar validaciones y depuración de la información reportada, así como informar las inconsistencias encontradas a las respectivas entidades quienes deben realizar los ajustes necesarios. Los registros que superan este filtro inicial pasan a actualizar la BDUA y se genera un reporte de registros válidos para cada entidad.

<i>Campo</i>	<i>Descripción</i>
Régimen	Régimen al que pertenece la entidad (contributivo, subsidiado, régimen de excepción).
Año	Año en el que se realizó el reporte.
Mes	Mes en el que se realizó el reporte.

<i>Campo</i>	<i>Descripción</i>
Código Departamento	Código del departamento del afiliado (codificación DANE).
Nombre Departamento	Departamento del afiliado.
Código Municipio	Código del municipio del afiliado (codificación DANE).
Nombre Municipio	Municipio del afiliado.
Número de Afiliados	Número de afiliados reportados para el periodo consultado.
Porcentaje	Porcentaje de afiliados respecto a número total de afiliados del régimen respectivo para el periodo consultado.

Cuadro 4.1: Descripción de campos, reporte BDU A

Nota: Campos pertenecientes a consulta en reportes de afiliación del ADRES. Tomado de [1].

4.2. Registro Especial de Prestadores de Servicios de Salud (REPS)

REPS se origina a partir de la Ley 10 de 1990, la cual indica entre otras disposiciones, que “se debe llevar un registro especial de las personas que presten servicios de salud, y efectuar su control, inspección y vigilancia”. De la misma manera se indica como competencia del Estado definir las formas en que se deben prestar la asistencia pública y establecer un sistema único de habilitación, los prestadores de servicios de salud habilitados deben demostrar que cuentan con la capacidad necesaria para la prestación del o de los servicios a cargo.

De esta manera REPS presenta información detallada de la capacidad instalada y los servicios prestados por los distintos prestadores de servicios de salud. Se actualiza diariamente a través de las distintas entidades quienes reportan al Ministerio de Salud, encargado de la administración de la base de datos.

Los prestadores de servicio de salud se pueden clasificar en:

- I.P.S.
- Profesionales independientes.
- Transporte especial de pacientes.
- Empresas de objeto social diferente.

<i>Campo</i>	<i>Descripción</i>
Nombre Departamento	Departamento del prestador.
Nombre Municipio	Municipio del prestador.
Código Habilitación	Código de habilitación brindado por el Ministerio de Salud.
Nombre Prestador	Nombre del prestador.
NIT	NIT del prestador.
Clase Prestador	Clasificación del prestador según su tipo.
ESE	Indica si el prestador es una Empresa Social del Estado.
Dirección	Dirección del prestador.
Teléfono	Teléfono del prestador.
Fax	Fax del prestador.
Email	Correo electrónico del prestador.
Gerente	Nombre del gerente de la entidad.
Nivel	Nivel de atención del prestador (solo para públicos).
Carácter	Indica si el prestador es nacional, departamental, distrital, municipal o indígena (solo para públicos).
Acreditado	Indica si el prestador está acreditado ante el organismo encargado.
Habilitación	Indica si la entidad se encuentra actualmente habilitada para la prestación de servicios.
Fecha Radicación	Fecha de radicación de la habilitación del prestador.
Fecha Vencimiento	Fecha en la que se vence la habilitación.
Fecha Cierre	Fecha en la que el prestador dejó de funcionar.
Digito Verificación	Digito de verificación del RUT.
Clase Persona	Indica si el prestador es una persona natural o jurídica.
Naturaleza Jurídica	Indica si la naturaleza del prestador es pública, privada o mixta.

Cuadro 4.2: Descripción de campos, prestadores de servicios de salud

Nota: Campos pertenecientes a la base de datos de Prestadores de Servicios de Salud. Tomado de [1].

<i>Campo</i>	<i>Descripción</i>
Nombre Departamento	Departamento del prestador.
Nombre Municipio	Municipio del prestador.
Código Habilitación	Código de habilitación brindado por el Ministerio de Salud.
Numero Sede	Número de la sede del prestador.
Nombre Sede	Nombre de la sede del prestador.
Dirección	Dirección del prestador.
Teléfono	Teléfono del prestador.
Fax	Fax del prestador.
Email	Correo electrónico del prestador.
NIT	NIT del prestador.
Digito Verificación	Digito de verificación del RUT.
Clase de Persona	Indica si el prestador es una persona natural o jurídica.
Naturaleza Jurídica	Indica si la naturaleza del prestador es pública, privada o mixta.
Clase Prestador	Clasificación del prestador según su objeto.
ESE	Indica si el prestador es una Entidad Social del Estado.
Nivel	Nivel de atención del prestador.
Carácter	Indica si el prestador es nacional, departamental, distrital, municipal o indígena (solo para públicos).
Habilitación	Indica si la entidad se encuentra actualmente habilitada para la prestación de servicios.
Grupo Servicios	Grupo al cual pertenece el servicio prestado: Internación, quirúrgicos, consulta externa, urgencias, transporte asistencial, apoyo diagnóstico, protección específica y detección temprana, procesos y otros servicios.
Servicio	Servicio prestado por la entidad.
Modalidad	Forma en la que se presta el servicio: Intramural, extramural o telemedicina.
Complejidad	Complejidad del servicio prestado: Baja, media o alta.

Cuadro 4.3: Descripción de campos, servicios de salud prestado por sede.

Nota: Campos pertenecientes a la base de los Servicios prestados por sede para cada prestador. Tomado de [1].

<i>Campo</i>	<i>Descripción</i>
Nombre Departamento	Departamento del prestador.

4.2. REGISTRO ESPECIAL DE PRESTADORES DE SERVICIOS DE SALUD (REPS)17

<i>Campo</i>	<i>Descripción</i>
Nombre Municipio	Municipio del prestador.
Código Habilitación	Código de habilitación brindado por el Ministerio de Salud.
Código Habilitación Sede	Código de habilitación para la sede brindado por el Ministerio de Salud.
Numero Sede	Número de la sede del prestador.
Nombre Sede	Nombre de la sede del prestador.
NIT	NIT del prestador.
Digito Verificación	Digito de verificación del RUT.
Clase de Persona	Indica si el prestador es una persona natural o jurídica.
Naturaleza Jurídica	Indica si la naturaleza del prestador es pública, privada o mixta.
Clase Prestador	Clasificación del prestador según su objeto.
ESE	Indica si el prestador es una Entidad Social del Estado.
Nivel	Nivel de atención del prestador.
Carácter	Indica si el prestador es nacional, departamental, distrital, municipal o indígena (solo para públicos).
Habilitación	Indica si la entidad se encuentra actualmente habilitada para la prestación de servicios.
Grupo	Indica el tipo de capacidad instalada: Camas, salas ambulancias o apoyo terapéutico.
Código Concepto	Código asignado al concepto de la capacidad instalada.
Nombre Concepto	Concepto relacionado con el grupo de la capacidad instalada.
Cantidad	Cantidad de recursos disponibles para cada concepto.
Numero Sede Principal	Número de la sede principal del prestador.
Numero Placa	Número de la placa de la ambulancia.
Modalidad	Modalidad de transporte del vehículo que presta el servicio de ambulancia: Fluvial, terrestre aéreo o marítimo.
Modelo	Modelo del vehículo.

Cuadro 4.4: Descripción de campos, capacidad instalada por sede

Nota: Campos pertenecientes a la base de la capacidad instalada por las entidades para la atención de usuarios (recursos y espacio disponible). Tomado de [1].

4.3. Precios reportados de medicamentos

El sistema de información sobre precios de medicamentos (SISMED) recopila datos con los que se logre un control del comportamiento de los precios. Se actualiza de manera trimestral a partir de los datos suministrados por laboratorios, distribuidores mayoristas, EPS e IPS. Es administrado por el Ministerio de Salud.

<i>Campo</i>	<i>Descripción</i>
Medicamento	Nombre de venta del medicamento.
Presentación	Presentación en la que es vendido el medicamento.
Fabricación Nacional	Indica si el medicamento es nacional.
Código ATC	Código de la sustancia farmacéutica o medicamento de acuerdo con clasificación anatómica, terapéutica, química recomendado por la OMS.
ATC	Estructura química del medicamento según el código ATC.
Principio Activo	Sustancia principal del medicamento.
Vía de Administración	Forma en la cual se administra el medicamento.
POS	Indica si el medicamento está incluido en el Plan Obligatorio de Salud.
CUM	Código único de medicamentos asignado por el INVIMA.
Periodo de precios	Periodo en el cual fueron reportados los precios.
Valor Mínimo	Valores reportados por ventas en canal institucional, comercial por los laboratorios y los centros mayoristas, así como las compras de estos.
Valor Máximo	
Unidades	
Precio	

Cuadro 4.5: Descripción de campos, precios reportados por medicamento.

Nota: La información que corresponde a los precios de medicamentos tiene como fuente el SISMED. Tomado de [1].

4.4. Eventos de vigilancia en salud pública

Mediante el Sistema nacional de vigilancia en salud pública (SIVIGILA) se registran de manera semanal los eventos que pueden incidir en la salud de una comunidad. Las entidades encargadas de realizar estos registros son instituciones prestadoras o no de servicios de salud o personas naturales, quien administra esta fuente es el Instituto Nacional de Salud.

<i>Campo</i>	<i>Descripción</i>
Año	Año en el que se registró el reporte del evento.

<i>Campo</i>	<i>Descripción</i>
Código Departamento	Código del departamento del afiliado (codificación DANE).
Nombre Departamento	Departamento del afiliado.
Nombre Municipio	Municipio del afiliado.
Evento	Nombre del evento reportado (estos están establecidos).
Semana	No de eventos reportados y confirmados por las entidades territoriales en la semana del año descrita (semana 1 a 53).
Total	No total de casos reportados para el evento.

Cuadro 4.6: Descripción de campos, eventos reportados.

Nota: Campos de reportes puestos a disposición del público. Tomado de [1].

4.5. Registro Individual de Prestación de Servicios de Salud (RIPS)

RIPS contiene el conjunto de datos mínimo y básico que el Sistema General de Seguridad Social en salud requiere para la regulación, control, soporte y dirección de la venta de servicios, que han sido estandarizados para las distintas entidades o instituciones involucradas en el sistema de salud.

RIPS está conformado por cuatro clases de datos, que aplican de acuerdo con el servicio de salud registrado: datos de identificación, del servicio, del motivo de la atención y datos del valor del servicio.

La base de datos del RIPS cuenta con cuatro tablas primarias en las que se registran los servicios de salud prestados de la siguiente manera:

<i>Campo</i>	<i>Descripción</i>
Código del prestador de servicios de salud	Código del prestador de servicios de salud.
No de la factura	Número de la factura.
Tipo de identificación	Tipo de identificación del usuario.
No de identificación	Número de identificación del usuario en el sistema.
Fecha de la consulta	Fecha en la que se realizó la consulta.
Código de la consulta	Código de la consulta según la especialidad.
Finalidad de la consulta	Razón de la consulta.
Causa externa	Causas externas de interés para vigilancia en salud pública.

<i>Campo</i>	<i>Descripción</i>
Código del diagnóstico principal	Código de la afección principal diagnosticada (codificación CIE-10).
Código del diagnóstico relacionado No.1	Otras afecciones o problemas relacionados con la afección principal. (codificación CIE-10).
Código del diagnóstico relacionado No.2	
Código del diagnóstico relacionado No.3	
Tipo de diagnóstico principal	Identificador para determinar si el diagnóstico es confirmado o presuntivo.
Valor de la consulta	Valor que el prestador cobra al pagador.
Valor de la cuota moderadora	Pago cobrado al usuario por el servicio.
Valor neto a pagar	Valor neto que el prestador cobrara al pagador.
Código del departamento	Código del departamento de residencia habitual.
Código del municipio	Código del municipio de residencia habitual.
Sexo	Sexo del usuario.
Edad	Edad del usuario.
Tipo de usuario	Régimen al que pertenece.

Cuadro 4.7: Descripción de campos, consultas.

Nota: Tomado de documento [1506873] Gestión de [1].

<i>Campo</i>	<i>Descripción</i>
Código del prestador de servicios de salud	Código del prestador de servicios de salud.
No de la factura	Número de la factura.
Tipo de identificación	Tipo de identificación del usuario.
No de identificación	Número de identificación del usuario en el sistema.
Fecha	Fecha del procedimiento.
Código del procedimiento	Código del procedimiento (CUPS).
Ámbito	Tipo de servicio en el que es atendido el usuario.
Finalidad	Finalidad o motivo por el cual se realiza el procedimiento.
Personal que atiende	Tipo de personal médico que atiende.
Diagnostico principal	Únicamente para procedimientos quirúrgicos.
Diagnostico relacionado	Únicamente para procedimientos quirúrgicos.
Complicación	Únicamente para procedimientos quirúrgicos y cuando se encuentra una condición adicional.
Valor del procedimiento	Valor que el prestador cobrara al pagador por el procedimiento.
Código del departamento	Código del departamento de residencia habitual.
Código del municipio	Código del municipio de residencia habitual.

<i>Campo</i>	<i>Descripción</i>
Sexo	Sexo del usuario.
Edad	Edad del usuario.
Tipo de usuario	Régimen al que pertenece.

Cuadro 4.8: Descripción de campos, procedimientos.

Nota: Tomado de [1].

<i>Campo</i>	<i>Descripción</i>
Código del prestador de servicios de salud	Código del prestador de servicios de salud.
No de la factura	Número de la factura.
Tipo de identificación	Tipo de identificación del usuario.
No de identificación	Número de identificación del usuario en el sistema.
Fecha	Fecha del evento
Causa externa	Causas externas de interés para vigilancia en salud pública.
Diagnostico a la salida	Causa por la que se justifica su estadía en observación.
Diagnostico relacionado 1	Diagnostico relacionado al principal (si existe).
Diagnostico relacionado 2	Segundo diagnostico relacionado al principal (si existe).
Diagnostico relacionado 3	Tercer diagnostico relacionado al principal (si existe).
Destino	Destino del usuario a la salida.
Estado	Registra si el paciente sale con vida o no de la observación.
Causa de muerte	Causa básica del deceso (según clasificación internacional).
Fecha de la salida	Fecha en la cual el paciente termina su estancia en observación.
Código del departamento	Código del departamento de residencia habitual.
Código del municipio	Código del municipio de residencia habitual.
Sexo	Sexo del usuario.
Edad	Edad del usuario.
Tipo de usuario	Régimen al que pertenece.

Cuadro 4.9: Descripción de campos, urgencias.

Nota: Tomado de [1].

<i>Campo</i>	<i>Descripción</i>
No de identificación	Número de identificación del usuario en el sistema.
Fecha ingreso	Fecha de ingreso a hospitalización.
Hora ingreso	Hora de ingreso a hospitalización.
Causa externa	Causas externas de interés para vigilancia en salud pública.
Diagnostico principal de ingreso	Diagnóstico de ingreso (puede ser presuntivo o no).
Diagnostico principal de egreso	Diagnostico principal de egreso confirmado.
Diagnostico relacionado 1 de egreso	Los diagnósticos relacionados son todos aquellos que hacen parte del estado de salud que justifico la estancia en el hospital.
Diagnostico relacionado 2 de egreso	
Diagnostico relacionado 3 de egreso	
Diagnóstico de complicación	Registrado si se presentó una complicación.
Fecha de egreso	Fecha en que el paciente deja el servicio de urgencias.
Hora de egreso	Hora en que el paciente deja la hospitalización.
Código del departamento	Código del departamento de residencia habitual.
Código del municipio	Código del municipio de residencia habitual.
Sexo	Sexo del usuario.
Edad	Edad del usuario.
Tipo de usuario	Régimen al que pertenece.

Cuadro 4.10: Descripción de campos, hospitalización.

Nota: Tomado de [1].

4.6. Sistema de Información Hospitalaria (SIHO)

SIHO alberga datos relacionados con variables contables, de presupuesto, financieras, de capacidad de instalada, recurso humano, producción de servicios y calidad. Esta información es reportada de manera trimestral por los prestadores de servicios de salud al Ministerio de Salud, quien es el encargado de su administración.

La información es reportada a través de los formularios de:

- Ingresos
- Gastos
- Facturación

- Cartera por deudor
- Pasivos
- Mecanismos de pago
- Balance general
- Estado de resultados
- Producción
- Calidad
- Capacidad instalada
- Recursos humanos
- Pasivo prestacional
- Ejecución presupuestal
- Contratación externa
- Infraestructura
- Procesos judiciales

4.7. Información contable IPS publicas

Actualizada trimestralmente con información de carácter financiero, económico, social y ambiental ante la Contaduría General de la Nación (CGN), es la base de datos referente a la información contable de las entidades públicas.

Concisamente la información reportada está relacionada con datos de los activos, pasivos, patrimonio, ingresos, gastos, costos de ventas y operaciones, y costos de producción de las entidades públicas. Los formularios usados para el reporte de información contable y financiera son:

- Saldos y Movimientos: información contable correspondiente al saldo inicial, movimientos débito y crédito, y saldo final, para las fechas de corte y por cada periodo definido.
- Operaciones Recíprocas: Saldos de las transacciones económicas y financieras realizadas entre entidades contables públicas.
- Notas de Carácter Específico: Aspectos que se refieren a situaciones particulares de las subcuentas.
- Notas de Carácter General: situaciones que contemplan los estados contables y los saldos de los reportes.

4.8. Registro Único de Afiliados (RUAF)

RUAF almacena los registros de los afiliados al Sistema Integral de Seguridad Social (Salud, Pensiones, Riesgos Profesionales), a Subsidio Familiar, a Cesantías, y de los beneficiarios de los programas prestados a través de la Red de Protección Social (Sena, ICBF, Acción social, entre otras). RUAF es administrado por el Ministerio de Salud.

La estructura de la información reportada se encuentra consignada en los anexos técnicos de la Resolución 1056 de 2015, que consiste en datos relacionados para los afiliados de los distintos subsistemas de protección social a excepción del subsistema de salud, del cual se obtiene la información del ADRES.

En el cuadro 4.11 muestra los valores que reporta cada entidad de los subsistemas.

<i>Tabla reportada</i>	<i>Subsistema</i>	<i>Seguridad Social en Pensiones</i>	<i>Cesantías</i>	<i>Seguridad Social en Riesgos Laborales</i>	<i>Seguridad Social en Pensiones. Entidades Pagadoras de Pensiones</i>	<i>Subsidio familiar</i>	<i>Asistencia Social y Subsistema de Parafiscales</i>
Afiliados Pensionados	X	X	X	X	X	X	
Subsidios Vinculados a programas de asistencia social					X	X	X
Novedades de actualización	X		X	X	X	X	X
Novedades de estado	X			X		X	
Inconsistencias	X		X	X	X	X	X

Cuadro 4.11: Información reportada por subsistema RUAF

Nota: Tomado de [1].

4.9. Formulario Único Territorial (FUT)

Con el objetivo de minimizar la cantidad de formularios que las entidades territoriales deben diligenciar para reportar a los organismos a nivel nacional se creó el Formulario Único Territorial. Esta información es reportada trimestralmente a través del sistema Consolidador de Hacienda e Información Financiera (CHIP) a cargo de la CGN.

Mediante el FUT se reporta la información de naturaleza organizacional, presupuestal, financiera, económica, geográfica, social y ambiental por parte de las entidades territoriales. Las categorías de información que se reportan son:

- Ingresos.
- Ingresos - Transferencias recibidas.
- Gastos de funcionamiento.
- Gastos de funcionamiento - Transferencias giradas.
- Gastos de inversión.
- Servicio de la deuda.
- Reservas.
- Ingresos ejecutados del SGR.
- Gastos ejecutados del SGR.
- Transferencias ejecutadas del SGR.
- Vigencias futuras.
- Cuentas por pagar.
- Ejecución presupuestal del Fondo de Salud.
- Tesorería Fondo de Salud.
- Registros presupuestales para el sector Agua potable y Saneamiento básico.
- Municipios descertificados.
- Regalías.
- Excedentes de liquides.

Capítulo 5

Fuentes Seleccionadas y Estructura

La información que nos es útil en el mundo se encuentra organizada y estructurada. La estructuración de la información supone ventajas tales como el manejo y el acceso fácil de los datos. Enciclopedias, guías, libros y muchos otros, son colecciones de datos que están organizados y estructurados para ser útiles bajo determinadas reglas ya establecidas. No siempre las reglas que se han definido pueden ser la mejor opción para tratar los datos en otro tipo de ámbito, por tanto, fuentes como las mencionadas anteriormente o las examinadas en este trabajo podrían ser más provechosas bajo esquemas de estructuración poco convencionales, pero que podrían permitir explotar patrones ocultos.

De acuerdo con los objetivos planteados en el presente trabajo de grado, se pretende en esta sección estructurar el conjunto de datos a partir del cual se desarrolla el trabajo propuesto, a fin de organizar estos datos de manera que sean utilizados eficientemente para el fin propuesto.

5.1. Estructuración de la información presentada en las fuentes

La información que se presenta en las diferentes fuentes del catálogo de información perteneciente al documento [1506873] [1] se puede agrupar en Datos Demográficos.

La demografía estudia la dimensión, la composición de la población y su evolución. Comprender los mecanismos de esta evolución, da origen a la metodología en específico [18].

Son cinco los aspectos en los que se centra la demografía:

- **Tamaño de la Población:** número de habitantes de un lugar en un momento determinado.
- **Distribución:** manera en que la población se encuentra dispersa en diferentes lugares del espacio geográfico en cierto momento.

- Composición: hace referencia al número de individuos por sexo, edad y otras categorías demográficas [19].
- Dinámica: nacimientos, defunciones, migración.
- Determinantes y consecuencias socio-económicas: características sociales y económicas adquiridas, que aparecen como causas y consecuencias de modificar características básicas de la dinámica demográfica.

El desarrollo del conocimiento demográfico se consigue mediante operaciones básicas como el conteo de la población sobre un territorio (a través de encuestas, censos, registros), el conteo de eventos que modifican la cantidad de la población sobre ese territorio (nacimientos, defunciones, migración). La evolución de la población está determinada entonces por la natalidad, mortalidad y migración. La información demográfica permite vincular el presente con el pasado y el futuro haciendo uso de procesos medibles.

Conocer el tamaño de la población, como se distribuye, de qué manera se moviliza, como se compone estructuralmente, que cambios hay en esa estructura y en su tamaño resulta ser de gran utilidad en la toma de decisiones, la planeación de políticas públicas en las cuales se quiere mejorar la calidad de vida de la población a partir de sus necesidades.

5.2. Fuentes de información seleccionadas

Conocer la dinámica de morbilidad y mortalidad puede permitir focalizar esfuerzos de las entidades en salud de manera tal que se prioricen elementos de atención y se distribuyan de manera optima los recursos. Es posible ver estas dinámicas a partir de la información registrada en las atenciones a los pacientes, de manera tal que RIPS proporciona un insumo valioso para esta tarea, al igual que lo hace BDUA desde la perspectiva de permitir agregar información, y visibilizar el origen de los pacientes. Otro elemento importante a revisar es la *Clasificación Internacional de Enfermedades*.

5.2.1. BDUA

La Base de Datos Única de Afiliados se constituye como:

“Una de las principales herramientas para el ejercicio de las funciones de dirección y regulación del Sistema General de Seguridad Social en salud, así como para el flujo de los recursos, su control y protección, de conformidad con las disposiciones legales y reglamentarias vigentes”. [20]

La estructura de los archivos maestro que contiene los datos remitidos por las entidades al ADRES se presentan en el Anexo Técnico de la Resolución 2232 de 2015 [21].

5.2.2. RIPS

El Sistema de Información de Prestación de Salud, es “el conjunto de datos mínimos y básicos que el Sistema General de Seguridad Social en Salud requiere para los procesos de dirección, regulación y control, como soporte de la venta de servicio, cuya denominación, estructura y características de ha unificado y estandarizado para todas las entidades”.

RIPS provee los datos mínimos y básicos requeridos para llevar el seguimiento al Sistema de Prestaciones de Salud en el SGSSS, en relación con el paquete obligatorio de servicios (POS y POSS). Los datos de este registro refieren la identificación del prestador del servicio de salud, el usuario que lo recibe, la prestación del servicio en mención, el motivo que origina el servicio: causa externa, diagnóstico.

El registro RIPS facilita las relaciones de carácter comercial entre las entidades administradoras (pagadoras) y las entidades o profesionales independientes que prestan los servicios, mediante el correspondiente detalle de venta del servicio mediante una estructura estándar que facilita la comunicación independiente de las soluciones informáticas usadas por los prestadores.

La estructura de los datos básicos que las entidades administradoras de planes de beneficios deben reportar sobre la prestación individual de servicios de salud se presenta en el Anexo Técnico de la Resolución 3374 de 2000 [22].

5.2.3. CIE-10

El catálogo de patologías de la Dirección de Demografía y Epidemiología del Ministerio de Salud está basado en la décima revisión de la Clasificación Estadística Internacional de Enfermedades y Problemas Relacionados con la Salud, denominado como CIE-10. En la actualización del catálogo de patologías realizado por parte de la Dirección de Demografía y Epidemiología a los tres días del mes de Enero de 2018 [23], se presenta 12545 códigos. De estos, 2394 corresponden a códigos que aparecen en determinados grupos de edad, se contabilizan en total 56 de estos grupos. Los restantes 10151 códigos pertenecen a diagnósticos que se pueden presentar indiferentemente de la edad del paciente.

El interés en el CIE-10 radica en su estandarización de los diagnósticos, los cuales como ya se mencionó se enriquecen con la característica de la edad, además de su asociación con cada uno de los sexos en los que se pueden presentar, de manera que a partir de esta última característica podemos obtener tres grandes grupos.

A continuación se realizan las observaciones con las cuales se describe la asociación entre las características código y grupo de edad con relación al sexo:

- El DataSet que se construirá a partir del CIE-10 contiene 12545 códigos.

- Se posee 10151 códigos restringidos por edad, y 2394 códigos libres de un rango de edad.
- Se identifican tres grandes grupos, que se denominan como General, Hombre, Mujer, en donde su participación dentro del dataset es de 11661, 117, 767 códigos respectivamente¹.
- En el grupo General hay 1786 códigos que están restringidos a un rango de edad, los restantes 9875 códigos están libres de esta restricción.
- En el grupo Hombre hay 6 códigos que están restringidos a un rango de edad, los restantes 111 códigos están libres de esta restricción.
- En el grupo Mujer hay 12 códigos que están restringidos a un rango de edad, los restantes 755 códigos están libres de esta restricción.

De esta manera se propone entrenar las diversas máquinas de clasificación teniendo como características el código CIE-10 de cuatro dígitos (el cual es único para cada diagnóstico).

La clasificación que se plantea es supervisada, teniendo como objetivo el sexo en los diagnósticos. De la descripción del dataset realizado con anterioridad se tiene que existen tres clases, General, Hombre, y Mujer, las cuales etiquetan los datos. Estas clases claramente no son balanceadas, esto quiere decir que el número de “observaciones” o registros que se tienen para cada una de las clases no es el mismo.

Esto representa en clasificación supervisada un problema cuando se trabaja con algoritmos como el *Random Forest*, pues es sensible a las proporciones de las diferentes clases. En consecuencia, este tipo de algoritmos tiende a favorecer la clase con una mayor cantidad de observaciones, es decir, la clase con mayor proporción, de manera que se pueden obtener métricas de resultado sesgadas. Si el interés se concentrara en la clasificación correcta de la clase minoritaria, dado que estos algoritmos buscan minimizar la tasa de error global en vez de prestar atención a la clase minoritaria, la predicción exacta fallara si no se obtiene la información necesaria.

El problema de tener dataset no balanceados en clasificación supervisada se suele abordar mediante métodos de muestreo los cuales se puede agrupar en: submuestreo, sobremuestreo, generación de datos sintéticos, y aprendizaje sensible al costo. Mediante estos métodos se realizan modificaciones a la proporción en las clases y el tamaño original del dataset. Otro de los métodos usados es la recolección de más datos (observaciones), que permitan balancear las clases, sin embargo para el caso de estudio como se ve a continuación esto no es posible.

¹Estos grupos serán posteriormente referenciados como clases cuando se hable de las maquinas clasificadoras.

Los métodos mencionados para tratar el problema de dataset desbalanceados no es posible de aplicar, pues el desbalance que se presenta en el dataset es propio de la naturaleza de los datos, para cada una de las clases se tienen registros únicos, con características bien definidas. Es de esperar que la clase General se lleve consigo todas las predicciones. Nos concentraremos en los datos de la clase Hombre, Mujer, pues finalmente son estos los que interesan al desarrollo de este trabajo (un diagnóstico que está etiquetado como General, se asume entonces de manera trivial como bien etiquetado). Seguirá existiendo el desbalanceo de las clases, sin embargo en este caso será por una brecha menor.

Capítulo 6

Métricas de Rendimiento

6.1. Accuracy

En la mayoría de los algoritmos de clasificación se calcula la exactitud en función del porcentaje de observaciones bien clasificadas.

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (6.1)$$

$$error_{rate} = 1 - accuracy = \frac{FP + FN}{TP + TN + FP + FN} \quad (6.2)$$

Donde:

TP : Observaciones de la Clase 1 correctamente predichas.

TN : Observaciones de la Clase 2 correctamente predichas.

FP : Observaciones de la Clase 2 erróneamente predichas.

FN : Observaciones de la Clase 1 erróneamente predichas.

Cuando se trabajan con clases no balanceadas, esta métrica entrega resultados engañosos, las clases minoritarias tienen un peso mínimo en la exactitud general.

Este problema es conocido como la paradoja de la exactitud, se trata del caso en donde la medida de la exactitud entrega un valor excelente (mayor a 90 %), pero realmente ese valor representa la clase mayoritaria.

	Clase 1	Clase 2
Clase 1	TP	FN
Clase 2	FP	TN

Cuadro 6.1: Matriz de Confusión

6.2. Confusion Matrix

Para presentar de manera clara y sin equivocaciones los resultados de predicción de un clasificador se puede usar la matriz de confusión o tabla de contingencia. En un problema bi-clase, la matriz consta de 2 filas y 2 columnas como se muestra en el Cuadro 6.1.

Las etiquetas superiores hacen referencia a las clases esperadas, en el lateral se encuentra las etiquetas de las clases predichas. Cada celda contendrá las predicciones hechas por el clasificador. Una clasificación perfecta contendría en TP y TN la totalidad de las predicciones esperadas, cuando esto no ocurre, los valores mal clasificados se presentan en FN y FP .

Se tiene entonces que la matriz de confusión es un desglose de las predicciones en una matriz que muestra las *predicciones positivas* en su diagonal principal, y las *predicciones negativas* en el resto de sus celdas.

6.3. Precision

La precisión o valor predictivo positivo (PPV) es una medida de corrección lograda en la predicción positiva, es decir, observaciones etiquetadas positivas cuando realmente son positivas, se puede considerar como una medida de exactitud de un clasificador. Una baja precisión puede indicar una cantidad alta de falsos positivos.

$$precision = \frac{TP}{TP + FP} \quad (6.3)$$

6.4. Recall

La sensibilidad o Tasa positiva real es una medida de la integridad de un clasificador. Se define como el número de verdaderos positivos dividido entre el número de verdaderos positivos más el número de falsos positivos.

$$recall = \frac{TP}{TP + FN} \quad (6.4)$$

6.5. F-Score

Esta medida da como resultado el equilibrio entre las medidas *precision* y *recall*, se considera como la medida de la efectividad de la clasificación en términos de la proporción y de la importancia del peso en cualquier *recall* o *precision* determinado por un coeficiente β . Este coeficiente es usualmente fijado en 1.

$$f_{score} = \frac{(1 + \beta)^2 * recall * precision}{\beta^2 * recall * precision} \quad (6.5)$$

Capítulo 7

Clasificadores

7.1. Random Forest

Random Forest es un algoritmo de clasificación y regresión ampliamente usado gracias a su buen rendimiento en set de datos x -dimensionales. En su diseño *Random Forest* integra las técnicas de *Decision Tree*, *Bagging*, y *Random Subspace*. [24]

7.1.1. Esquema del algoritmo Random Forest

- De manera aleatoria se crea el conjunto de datos de entrenamiento (selección con reemplazo) de igual tamaño que el conjunto de datos original. Al seleccionar por azar con reemplazo no todos los datos del conjunto inicial estarán en el conjunto de entrenamiento.
- Los datos que no hacen parte del conjunto de entrenamiento forman el conjunto de validación o *out of bag data (OOB data)*.
- Para cada punto de división del árbol o nodo, la búsqueda de la mejor variable que divide los datos no se hace sobre todas las variables sino, sobre un subconjunto, m , de las mismas. La selección del subconjunto de variables se realiza aleatoriamente.
- Se busca la mejor división de los datos de entrenamiento teniendo en cuenta solo las m variables seleccionadas. Para esto se implementa una función objetivo, generalmente se usa la entropía o el índice de Gini.
- Las indicaciones anteriores son repetidas n veces, de manera que se tenga un conjunto de árboles de decisión entrenados sobre diferentes conjuntos de datos y atributos.
- Una vez el algoritmo este entrenado, la evaluación de cada nueva entrada es realizada con el conjunto de árboles. La clasificación es realizada por el voto mayoritario del conjunto de árboles, en el caso de la regresión por el valor promedio de los resultados.

7.1.2. Observaciones

Los datos OOB se usan para determinar la impureza en los nodos terminales. La suma de estas impurezas determina la impureza del árbol.

Cada registro del conjunto original de datos esta in bag para algunos árboles del *random forest*, y en *out of the bag* para otros árboles. Probablemente cualquier par de registros no comparten un patrón idéntico sobre que arboles están *in bag* y cuales en *out of the bag*.

Para medir el error de *random forest* se utiliza con frecuencia el *out-the-bag error*. Para cada árbol se usa el conjunto de objetos que no han sido seleccionados por su muestra *bootstrap* de entrenamiento para ser clasificados por dicho árbol.

Importancia de Gini: Cada vez que se realiza una división de un nodo en una variable m , el criterio de impureza de *Gini* para los nodos descendientes es menor que el de su padre. Sumar las disminuciones de *Gini* para cada variable individual sobre todos los árboles en el bosque proporciona una rápida variable que frecuentemente es consistente con la medida de importancia de la permutación.

7.1.3. Resumen del algoritmo

Random Forest es un algoritmo predictivo que usa *Bagging* para combinar diferentes árboles, en donde cada uno de estos se construye con observaciones y variables aleatorias.

De manera resumida sigue el siguiente esquema:

1. Seleccionar individuos aleatoriamente (muestreo con reemplazo) para crear diferentes conjuntos de datos.
2. Construye un árbol de decisión con cada conjunto de datos, de esta manera se obtienen diferentes árboles, cada conjunto contiene individuos y variables diferentes para cada nodo.
3. Al crear los árboles se eligen variables en cada nodo del árbol de manera aleatoria, dejando crecer el árbol en profundidad.
4. Predice datos nuevos usando voto mayoritario, donde se clasifica como positivo si la mayoría de los árboles predicen la observación como positiva.

7.2. Naïve Bayes

Naïve Bayes es una técnica de clasificación y predicción supervisada que construye modelos probabilísticos y que está basado en el Teorema de Bayes [25] y que tiene como

premisa la independencia de los datos [26, 27]. Se trata de una técnica supervisada dado que se requiere de ejemplos clasificados previamente para su operación.

El Teorema de Bayes en términos generales expresa la probabilidad de que ocurra un evento, conociendo que también se da otro evento. La estadística Bayesiana es usada para calcular estimaciones basadas en conocimiento subjetivo previo. Las implementaciones de este teorema se adaptan con el uso, y permiten combinar datos provenientes de diversas fuentes y expresarlos en el grado de probabilidad.

7.2.1. Teorema de Bayes

Bayes [25] estudio la relación intrínseca que hay entre la probabilidad, la predicción y el progreso científico. El trabajo realizado por Bayes se centró esencialmente en la manera en que se deben formular las creencias probabilísticas sobre el mundo cuando se encuentran nuevas evidencias. El argumento de este teorema es que aprendemos sobre el mundo por medio de la aproximación, de manera que nos acercamos a la verdad cada vez más en la medida en que encontramos evidencias.

Bayes expresa su teorema como:

$$P(Q|T) = \frac{P(T|Q)P(Q)}{P(T)} \quad (7.1)$$

$P(Q)$ es el *a priori*, o la manera en que se suministra al modelo el conocimiento previo sobre los valores que puede tomar la hipótesis Q . Cuando no se tiene conocimiento anticipado se puede usar *a prioris* que asignan probabilidad igual a todos los valores de Q , otra alternativa es asignar *a prioris* que restrinjan los resultados a rangos razonables.

$P(T|Q)$ se conoce como el *likelihood*, manera de incluir datos en el análisis. Esta expresión matemática especifica la aprobación de los datos. El *likelihood* tiene más peso en los resultados a medida que los datos aumentan, *likelihood* se asemeja a una probabilidad pero no lo es; el *likelihood* de una hipótesis Q , dados los datos T es proporcional a la probabilidad de obtener T dado que Q es verdadera.

$P(T)$ es la evidencia, la probabilidad de observar los datos T promediado sobre todas las posibles hipótesis Q . De manera general, la evidencia puede verse como una constante de normalización, que se puede omitir sin perder demasiada generalidad.

$P(Q|T)$ es el *a posteriori*, la distribución de probabilidad final para la hipótesis. Es el resultado lógico al haber procesado un conjunto de datos, un *likelihood* y un *a priori*. Puede verse como la actualización del *a priori* luego de agregar los datos adicionales.

El teorema de Bayes en su forma más básica es una expresión algebraica con tres variables conocidas y una incógnita, usa probabilidades condicionales, y determina la probabilidad de que una hipótesis Q sea verdadera si ha sucedido algún evento T .

7.2.2. El clasificador Bayesiano Ingenuo

Esta técnica de clasificación supervisada basada en el Teorema de Bayes asume que hay independencia entre los atributos [26] [27], esto es, la aparición de una característica en una clase no está relacionada con la presencia de cualquier otra característica. Incluso, si alguna característica depende de otra, todas contribuyen de manera independiente a la probabilidad *a posteriori*. Se llama ingenuo precisamente por asumir la independencia en los atributos, algo que en la realidad no se suele dar.

Este clasificador se ha usado en escenarios como:

- Análisis de sentimientos: análisis de los tweets, revisiones y comentarios. [28]
- Clasificación de texto: conocido de manera exitosa en la clasificación de texto, determinación de si un texto pertenece a una o varias categorías. [29]
- Detección de spam: ejemplo de clasificación de texto en correo electrónico.

Debido a las suposiciones que hace el clasificador bayesiano ingenuo sobre los datos, generalmente no funciona para modelos complejos.

Ventajas:

- Rápidos en entrenamiento y predicción.
- Proporción de predicción probabilística directa.
- Fáciles de interpretar.
- Pocos parámetros a optimizar.

Dicho esto, este clasificador es una buena opción de modelo de clasificación inicial. Si se obtienen buenos resultados, se tiene un clasificador rápido, en caso contrario se tiene una base con que explorar modelos más robustos.

7.3. Decision Tree

Los *Decision Tree* están compuestos por una serie de nodos, inicia con un solo nudo padre y se extiende por una cantidad n de nodos hoja, las cuales están representando las las categorías que el árbol puede clasificar. Puede verse como un diagrama de flujo, el cual inicia en un nodo raíz y finaliza con una decisión tomada en las hojas.

El nodo raíz representa a toda la población que se analiza, desde este nodo se crean subgrupos de acuerdo a un conjunto de características, estos a su vez se dividen en nodos de decisión, para en algún momento llegar a un nodo terminal. Se realiza un proceso de poda como parte del proceso de la obtención de nodos terminales eliminando subnodos del nodo padre [30]. *Decision Tree* es un algoritmo de uso popular:

- Poder explicativo:
El resultado del *Decision Tree* se puede interpretar sin conocimiento profundo estadístico o matemático.
- Análisis de datos exploratorios:
Decision Tree permite a los analistas identificar variables significativas y relaciones entre variables.
- Mínima limpieza de datos:
Decision Tree es resistente a valores atípicos y faltantes, de esta manera requiere menos limpieza de datos que otros algoritmos.
- Cualquier tipo de datos:
Decision Tree puede hacer clasificaciones basadas en variables numéricas y categóricas.
- No paramétrico:
Decision Tree es no paramétrico, a diferencia de las redes neuronales que procesan los datos de entrada transformados en un tensor, a través de la multiplicación de tensores utilizando un gran número de parámetros.

Desventajas:

- Overfitting:
Over fitting es un defecto común de los *Decision Tree*. Establecer restricciones en los parámetros del modelo (limitación de profundidad) y simplificar el modelo mediante la poda son dos formas de regularizar un *Decision Tree* y mejorar su capacidad de generalización en el conjunto de prueba.
- Predicción de variables continuas:
Aunque los árboles de decisión pueden procesar datos numéricos continuos, no son una forma práctica de predecir dichos valores, las predicciones del árbol de decisiones deben separarse en categorías discretas, lo que resulta en una pérdida de información al aplicar el modelo a valores continuos.
- Características fuertes de Ingeniería:
Los *Decision Tree* requieren una gran ingeniería de características. Cuando se trata de datos no estructurados, el resultado resulta en subóptimos.

Capítulo 8

Entrenamiento Clasificadores

En esta sección se presenta la implementación y resultados para los clasificadores Naïve Bayes (Gaussiano, Bernoulli, Multinomial), Decision Tree, y Random Forest.

Inicialmente se cuenta en el set de datos con 3 clases no balanceadas, la Figura 8.1 presenta la dispersión de los datos para estas clases, Hombres, Mujeres, y General. Como se comentó anteriormente este problema resulta en un sesgo de clasificación hacia la clase mayoritaria.

En los clasificadores siguientes se trabajo con las clases 1 y 2 que corresponden respectivamente a Hombres y Mujeres, y que se pueden ver en la Figura 8.2.

8.1. Naïve Bayes

Los métodos de Naïve Bayes son un conjunto de algoritmos de aprendizaje supervisado que se basan en la aplicación del Teorema de Bayes [25], con una suposición ingenua de independencia entre cada característica [31].

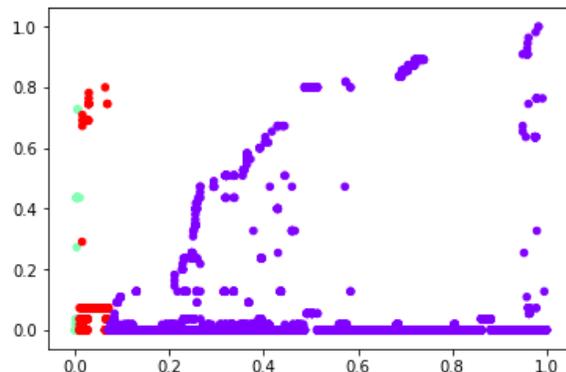


Figura 8.1: Dispersión de las observaciones para las clases Hombre, Mujer y General en el dataset

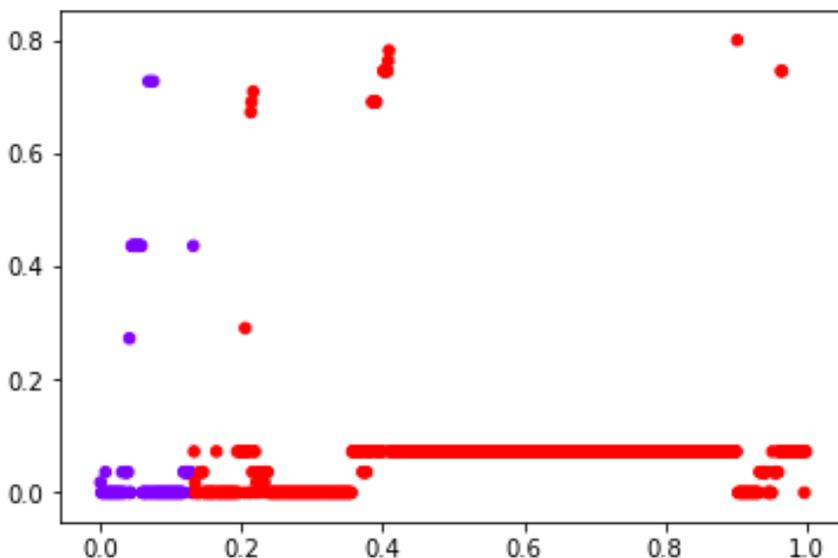


Figura 8.2: Dispersión de las observaciones para las clases Hombre y Mujer en un rango de 0 a 1.

Veremos a continuación tres variantes para este método.

8.1.1. Gaussian Naïve Bayes

```

1  # fit a Naive Bayes model to the data
2  modelGNB.fit(data, target)
3
4  # make predictions
5  predicted = modelGNB.predict(data)
6
7  # summarize the fit of the model
8  accuracy = accuracy_score(expected, predicted)
9  print 'accuracy: ' + str(accuracy)
10 print '\nreport of classification: \n' +
    ↪ classification_report(expected, predicted)
11 print '\nconfusion matrix: '
12 print pd.DataFrame(
13     confusion_matrix(expected, predicted),
14     columns=['Prediccion Hombres', 'Prediccion Mujeres'],
15     index=['Hombres', 'Mujeres']
16 )

```

Codigo 1: Fragmento de código para clasificación con *Gaussian Naïve Bayes*

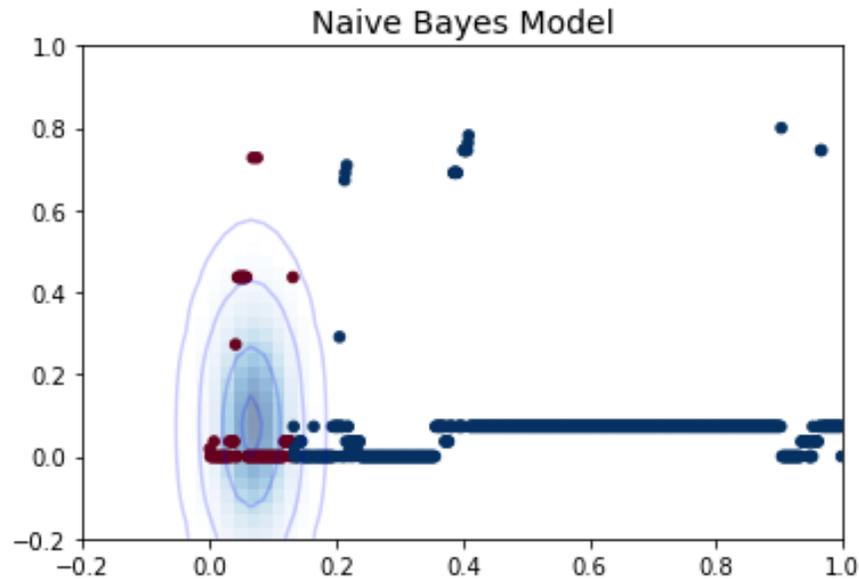


Figura 8.3: Modelo Naïve Bayes para las clases Hombre y Mujer

8.1.2. Bernoulli Naïve Bayes

```

1  # fit a Naive Bayes model to the data
2  modelBNB.fit(data, target)
3
4  # make predictions
5  predicted = modelBNB.predict(data)
6
7  # summarize the fit of the model
8  accuracy = accuracy_score(expected, predicted)
9  print 'accuracy: ' + str(accuracy)
10 print '\nreport of classification: \n' +
    ↪ classification_report(expected, predicted)
11 print pd.DataFrame(
12     confusion_matrix(expected, predicted),
13     columns=['Prediccion Hombres', 'Prediccion Mujeres'],
14     index=['Hombres', 'Mujeres']
15 )

```

Código 2: Fragmento de código para clasificación con *Bernoulli Naïve Bayes*

```

accuracy: 0.9898190045248869

report of classification:
      precision    recall  f1-score   support

     1         0.97     0.95     0.96     117
     2         0.99     1.00     0.99     767

avg / total         0.99     0.99     0.99     884

confusion matrix:
      Prediccion Hombres  Prediccion Mujeres
Hombres         111         6
Mujeres          3         764

```

Figura 8.4: Métricas obtenidas de la clasificación con el *Gaussian Naïve Bayes*

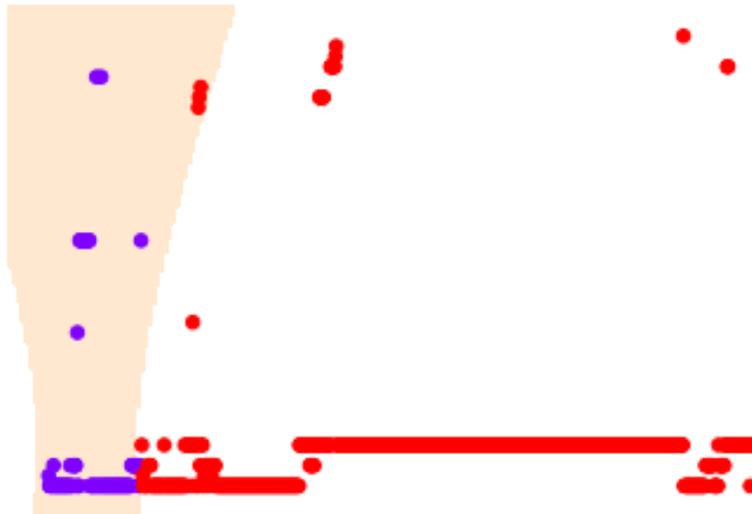


Figura 8.5: Visualización de la clasificación con el *Gaussian Naïve Bayes*

```

accuracy: 0.8676470588235294

report of classification:
      precision    recall  f1-score   support

     1         0.00      0.00      0.00     117
     2         0.87      1.00      0.93     767

avg / total         0.75      0.87      0.81     884

      Prediccion Hombres  Prediccion Mujeres
Hombres                0                117
Mujeres                0                767

```

Figura 8.6: Métricas obtenidas de la clasificación con el *Bernoulli Naïve Bayes*

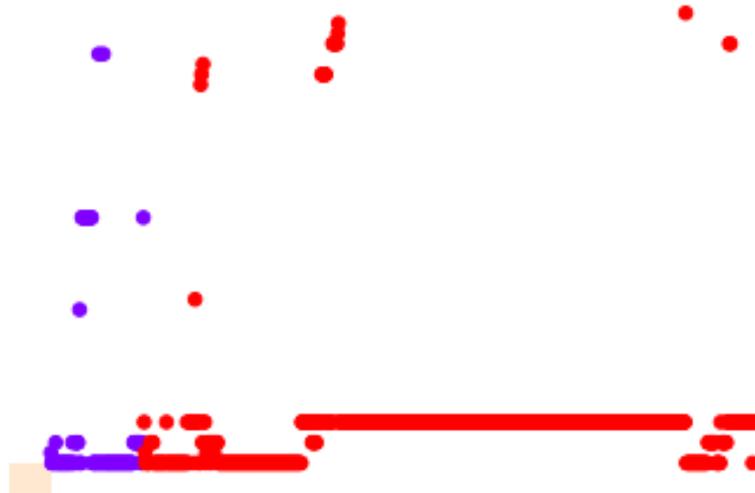


Figura 8.7: Visualización de la clasificación con el *Bernoulli Naïve Bayes*

```

accuracy: 0.8676470588235294

report of classification:
      precision    recall  f1-score   support

     1         0.00         0.00         0.00         117
     2         0.87         1.00         0.93         767

avg / total         0.75         0.87         0.81         884

confusion matrix:
      Prediccion Hombres  Prediccion Mujeres
Hombres           0           117
Mujeres           0           767

```

Figura 8.8: Métricas obtenidas de la clasificación con el *Multinomial Naïve Bayes*

8.1.3. Multinomial Naïve Bayes

```

1  # fit a Naive Bayes model to the data
2  modelMNB.fit(data, target)
3
4  # make predictions
5  predicted = modelMNB.predict(data)
6
7  # summarize the fit of the model
8  accuracy = accuracy_score(expected, predicted)
9  print 'accuracy: ' + str(accuracy)
10 print '\nreport of classification: \n' +
    ↪ classification_report(expected, predicted)
11 print '\nconfusion matrix: '
12 print pd.DataFrame(
13     confusion_matrix(expected, predicted),
14     columns=['Prediccion Hombres', 'Prediccion Mujeres'],
15     index=['Hombres', 'Mujeres']
16 )

```

Código 3: Fragmento de código para clasificación con *Multinomial Naïve Bayes*



Figura 8.9: Visualización de la clasificación con el *Multinomial Naïve Bayes*



Figura 8.10: Visualización de la clasificación con *Decision Tree*, profundidad 1, 2 y 3.

```

accuracy: 1.0

report of classification:
              precision    recall  f1-score   support

     1         1.00      1.00      1.00     117
     2         1.00      1.00      1.00    767

avg / total         1.00      1.00      1.00     884

confusion matrix:
              Prediccion Hombres  Prediccion Mujeres
Hombres              117              0
Mujeres              0              767

```

Figura 8.11: Métricas obtenidas de la clasificación con el *Decision Tree*

8.2. Decision Tree

```

1  # fit a Random Forest model to the data
2  modelDecisionTree.fit(data, target)
3
4  # make predictions
5  predicted = modelDecisionTree.predict(data)
6
7  # summarize the fit of the model
8  accuracy = accuracy_score(expected, predicted)
9  print 'accuracy: ' + str(accuracy)
10 print '\nreport of classification: \n' +
    ↪ classification_report(expected, predicted)
11 print '\nconfusion matrix: '
12 print pd.DataFrame(
13     confusion_matrix(expected, predicted),
14     columns=['Prediccion Hombres', 'Prediccion Mujeres'],
15     index=['Hombres', 'Mujeres']
16 )

```

Código 4: Fragmento de código para clasificación con *Decision Tree*

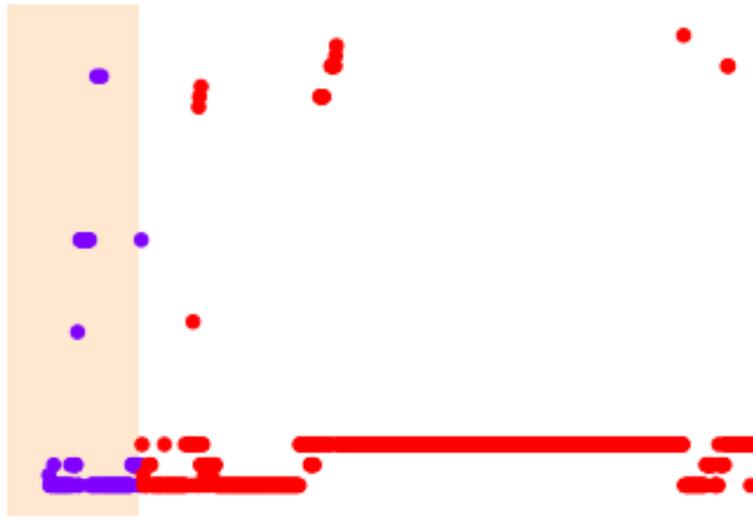


Figura 8.12: Visualización de la clasificación con el *Decision Tree*

8.3. Random Forest

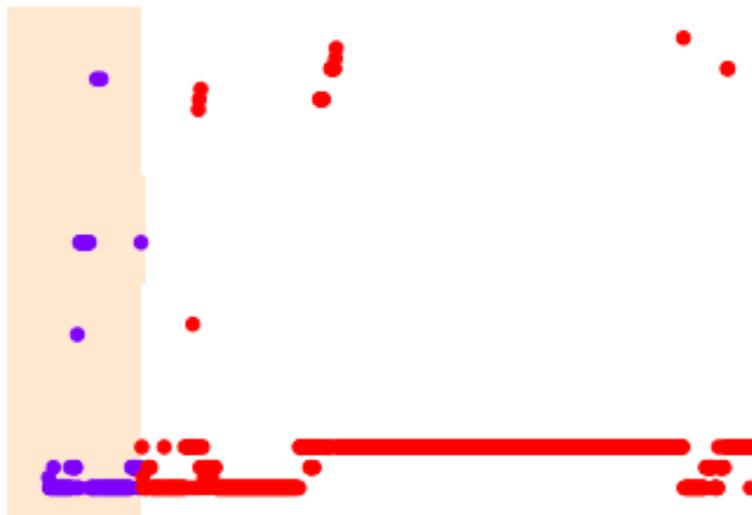
```

1  # fit a Random Forest model to the data
2  modelRandomForest.fit(data, target)
3
4  # make predictions
5  predicted = modelRandomForest.predict(data)
6
7  # summarize the fit of the model
8  accuracy = accuracy_score(expected, predicted)
9  print 'accuracy: ' + str(accuracy)
10 print '\nreport of classification: \n' +
    ↪ classification_report(expected, predicted)
11 print '\nconfusion matrix: '
12 print pd.DataFrame(
13     confusion_matrix(expected, predicted),
14     columns=['Prediccion Hombres', 'Prediccion Mujeres'],
15     index=['Hombres', 'Mujeres']
16 )

```

Codigo 5: Fragmento de código para clasificación con *Random Forest*

```
accuracy: 1.0  
  
report of classification:  
      precision    recall  f1-score   support  
  
     1         1.00      1.00      1.00     117  
     2         1.00      1.00      1.00     767  
  
avg / total         1.00      1.00      1.00     884  
  
confusion matrix:  
      Prediccion Hombres  Prediccion Mujeres  
Hombres           117           0  
Mujeres            0           767
```

Figura 8.13: Métricas obtenidas de la clasificación con el *Random Forest*Figura 8.14: Visualización de la clasificación con el *Random Forest*

8.4. Análisis de rendimiento

Para cada uno de los clasificadores anteriores se tomaron las respectivas métricas de desempeño, las cuales se agrupan en el Cuadro 8.1.

	Clase	Accuracy	Precision	Recall	F-Score	Confusion Matrix	
Gaussian Naive Bayes	Hombres	0.9898	0.97	0.95	0.96	111	6
	Mujeres		0.99	1.00	0.99	3	764
Bernouli Naive Bayes	Hombres	0.8676	0.00	0.00	0.00	0	117
	Mujeres		0.87	1.00	0.93	0	767
Multinomial Naive Bayes	Hombres	0.8676	0.00	0.00	0.00	0	117
	Mujeres		0.87	1.00	0.93	0	767
Decision Tree	Hombres	1.0	1.00	1.00	1.00	117	0
	Mujeres		1.00	1.00	1.00	0	767
Random Forest	Hombres	1.0	1.00	1.00	1.00	117	0
	Mujeres		1.00	1.00	1.00	0	767

Cuadro 8.1: Métricas de rendimiento para los clasificadores planteados.

Según los valores registrados por la métrica accuracy todos los clasificadores planteados tienen un buen resultado, en especial Decision Tree, Random Forest y Gaussian Naive Bayes, siendo este último el que tienen una tasa de error menor a los otros dos clasificadores que usan Naïve Bayes.

Es de notar que a pesar de que los clasificadores Bernouli Naïve Bayes y Multinomial Naïve Bayes tienen un accuracy mayor al 86 %, para estos clasificadores este valor está sesgado hacia la clase mayoritaria, lo cual se puede ver a través de las otras métricas, y de manera inequívoca se comprueba con la matriz de confusión. Con esto se descarta totalmente estos clasificadores como máquinas para la clasificación de los diagnósticos.

Los valores de todas las métricas para los clasificadores Decision Tree y Random Forest alcanzan el resultado máximo, lo cual nuevamente se evidencia con la matriz de confusión, de tal manera que son estos los clasificadores que resultan en las máquinas a usar para la clasificación de los diagnósticos. Dado que a las máquinas no ingresan registros nuevos, y que los clasificadores triunfantes presentan resultados parejos, es indiferente el uso de alguno para la clasificación.

Capítulo 9

Estructuración de la Red Compleja

La base de datos de RIPS contiene el registro de todos de los servicios que se han prestado por paciente. De cada uno de estos registros es de interés los campos que corresponden a los diagnósticos que se pueden registrar durante una consulta, y que corresponden a un diagnostico principal y tres diagnósticos secundarios que pueden ser o no registrados.

Para la construcción de la Red Compleja se puede tratar cada diagnostico como un nodo dentro del grafo, abra relación (arista) entre nodos siempre y cuando correspondan a un mismo paciente. La interacción entre los nodos de un mismo paciente estará dada sobre las siguientes reglas:

- Por cada registro el diagnostico principal tendrá una arista con cada uno de sus diagnósticos secundarios.
- Diagnósticos principales de un paciente tienen aristas consecutivas.
- Los diagnósticos secundarios tienen aristas entre sí.
- Diagnósticos principales tienen arista con cada uno de los diagnósticos secundarios del registro al que antecede.

En el caso de que en un mismo registro se encuentre como diagnostico secundario pero el mismo código que el diagnostico principal, la ejecución continúa en el siguiente registro, omitiendo los diagnósticos secundarios del registro antecesor.

Ejemplo: En la Figura 9.1 se muestran 10 registros de diagnósticos que se asumen pertenecen a un mismo individuo.

Se obtiene a partir de la información suministrada por esta tabla los nodos del grafo y las respectivas aristas en forma de una lista que contiene tuplas por cada relación de diagnósticos (construidas según las reglas descritas) de la siguiente manera:

[("R309", "K021"), ("R309", "K003"), ("K021", "J46X"), ("K021", "K003"), ("J46X", "K003"), ("J46X", "J189"), ("J189", "C448"), ("J189", "K076"), ("C448", "J00X"), ("C448",

fecha_cons	cod_diagnp	cod_diagnr1	cod_diagnr2	cod_diagnr3
2014-09-23	R309	NaN	NaN	NaN
2014-09-25	K021	K003	K003	NaN
2015-01-24	J46X	NaN	NaN	NaN
2015-01-26	J189	NaN	NaN	NaN
2015-02-25	C448	K076	NaN	NaN
2015-07-23	J00X	J450	N760	NaN
2015-08-20	F150	T659	NaN	NaN
2015-08-22	F320	NaN	NaN	NaN
2015-12-09	J00X	J039	NaN	NaN
2016-04-14	D649	E039	NaN	NaN

Figura 9.1: Muestra de registros RIPS en régimen contributivo.

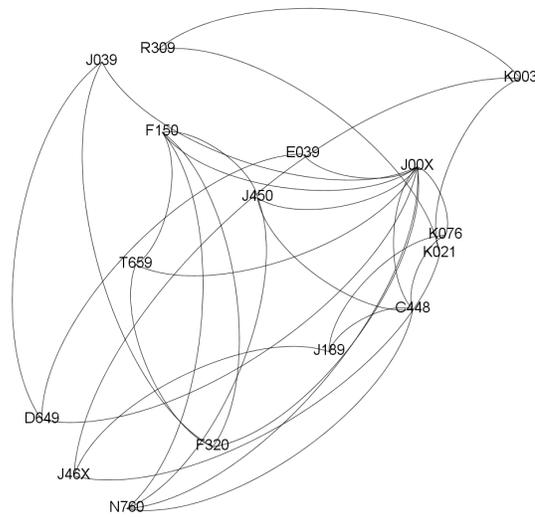


Figura 9.2: Red para los 10 registros presentados en la Figura 9.1.

["K076"), ("J00X", "K076"), ("C448", "J450"), ("C448", "N760"), ("J450", "N760"), ("J00X", "F150"), ("J00X", "J450"), ("F150", "J450"), ("J00X", "N760"), ("F150", "N760"), ("J00X", "T659"), ("F150", "F320"), ("F150", "T659"), ("F320", "T659"), ("F320", "J00X"), ("F320", "J039"), ("J00X", "D649"), ("J00X", "J039"), ("D649", "J039"), ("J00X", "E039"), ("D649", "E039")]

A partir de esta lista se obtiene la red de la Figura 9.2.

Desde esta red se puede analizar la dinámica de morbilidad de los individuos, así como detectar los diagnósticos que están conectados fuertemente, diagnósticos que tienen una alta relación con otros diagnósticos, y que pueden representar los puntos focales para la toma de decisiones. Para que este análisis pueda aportar a la toma de decisiones en Salud Pública, los registros con los que se realiza la construcción de la red deben cumplir con un criterio de calidad en cuanto a los diagnósticos que se tienen, de esta manera antes de construir la red los diagnósticos se someten a una evaluación a través de la máquina de clasificación construida para este fin.

9.1. Integración de RIPS y BDUA

El cruce entre la información suministrados para RIPS y BDUA se realiza mediante el campo de identificación común entre ambos, *identifica* para BDUA y *num_ident* para RIPS. Se genera una sabana de datos de 7'379.142 registros, de los cuales 4'732.309 corresponden a registros masculinos y 2'626.833 registros femeninos. Esta sabana nos permite realizar la validación de los diagnósticos por medio del clasificador *Random Forest*.

Como resultado de la clasificación sobre esta data se encuentra 1336 registros con diagnósticos incorrectos respecto a lo estipulado en el CIE-10. El Cuadro 9.1 resume el numero de ocurrencias por cada código diagnostico.

Códigos CIE-10	Ocurrencias Negativas
Z359	226
Z349	147
Z340	61
Z321, Z348	60
N760	50
N939	38
Z358	36
N771	30
Z392, O200	23
N911	22
N47X	21
Z390	20
Z391	19
Z305	18
N72X, N912	17
Z320	16
N761, B373	15
Z014, N739, Z356, N946	14
Z352, N459, Z124	12
O800	11
O234	10
N40X, N870	9
N926	8
O471, O470	7
O233, O429	6
N96X, O239, N938, N768, O420, N944, N511	5

Códigos CIE-10	Ocurrencias Negativas
O479, O260, Q523, N910, N489, N748, N925, O623, N762	4
N751, N481, O009, Z357, O235, N86X, N945, Q525, S312, N979, N921, O16X, O809, N512, O367, O034, O48X	3
C538, N482, O365, O912, O839, O838, O25X, N922, N850, O13X, O261, I861, Z975, N832, O231, E282, N838, O000	2
N764, N839, N430, O369, N879, N974, N872, Z301, F524, O269, N915, O335, N812, O620, N738, O730, N719, O688, Q531, N843, N410, O410, Q529, O325, N871, O140, N816, O926, N959, D400, N891, N900, N508, C61X, Z368, N819, B260, Y763, O829, O719, N941, Q510, O064, E298, D299, N484, N951, O209, O021, N856, D259, N898, O983, N920, N949, O360, Q501, O358, N811, O019, O049, N433, F539, D397, D251, N418, Q530, O100, N942, N841, D401, D293, Q524, N434, O044, N432, O039, N907, N499	1

Cuadro 9.1: Ocurrencias Negativas del CIE-10 en datos del régimen subsidiado

El mismo ejercicio realizado sobre la BDUA régimen contributivo y RIPS, nos arroja una sábana de datos de 4'694.171 registros, en esta caso 1'839.615 corresponden a registros masculinos y los restantes 2'853.640 registros son femeninos.

De la predicción sobre estos datos se obtienen 75.440 registros están mal registrados respecto a lo establecido por el CIE-10, se presenta en el Cuadro 9.2 el número de

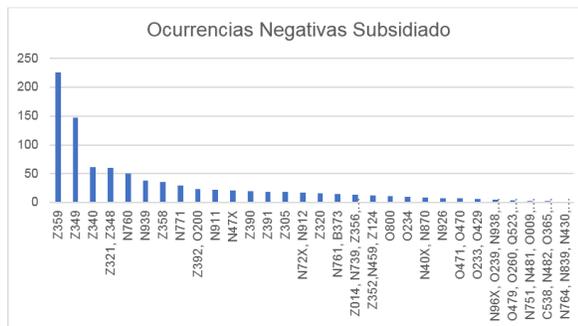


Figura 9.3: Frecuencia de ocurrencias para el Cuadro 9.1.

ocurrencias para cada código.

Código CIE-10	Ocurrencias Negativas
Z359	139
Z349	64
N760	41
Z321	34
Z348, Z340	24
N771	19
N40X	18
Z305	17
N939	16
O479	14
Z014, Z124, N912	11
N47X, O470, N979, O800	10
Z391, Z392	9
N739, O200, E282	8
N761, O039	7
Z358, O234, Z390, Z320	6
N911, Z356, O471, N511	5
N910, N926, B373, O429, O342, N870, O000, N72X, D259, O034, O809, O16X	4
N879, N512, N841, N459, O420, N410, N832, Z352, O926	3
O926, I861, O367, O839, O838, O623, N951, N481, O235, Z125, N750, N945, C538, N946, N768, N872, N710, N925, Q531, O244, N948, N830	2

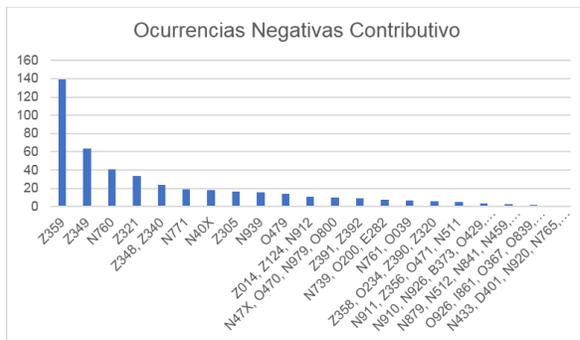


Figura 9.4: Frecuencia de ocurrencias para el Cuadro 9.2.

Código CIE-10	Ocurrencias Negativas
N433, D401, N920, N765, O100, N938, O239, N411, Z363, N942, D27X, M810, N700, O260, O441, N891, N952, N943, D299, O983, N914, D397, N766, Z364, C61X, O829, O911, N949, N941, O620, O261, Q501, O212, O368, N915, O359, O741, O059, O719, O231, O149, F524, S312, Z33X, O48X, N751, O912, N922, N96X, O049, N871, N921, N992, N429, O411	1

Cuadro 9.2: Ocurrencias Negativas del CIE-10 en datos del régimen contributivo.

9.2. Construcción de la red compleja

Para la construcción del grafo se tomaron como insumo de datos los registros del BDU A y RIPS, se toman solo los registros que fueron marcados como aptos por las maquinas construidas para la clasificación de estos registros.

Estas dos fuentes de información se cruzan de tal manera que se obtiene una sábana de datos la cual es ordenada en primera instancia por identificación y luego por fecha de atención, de tal forma que se tiene un registro histórico de los individuos de manera ordenada, a partir del cual se procede a relacionar los diagnósticos por individuo como previamente se estableció en las reglas definidas para la construcción de la red.

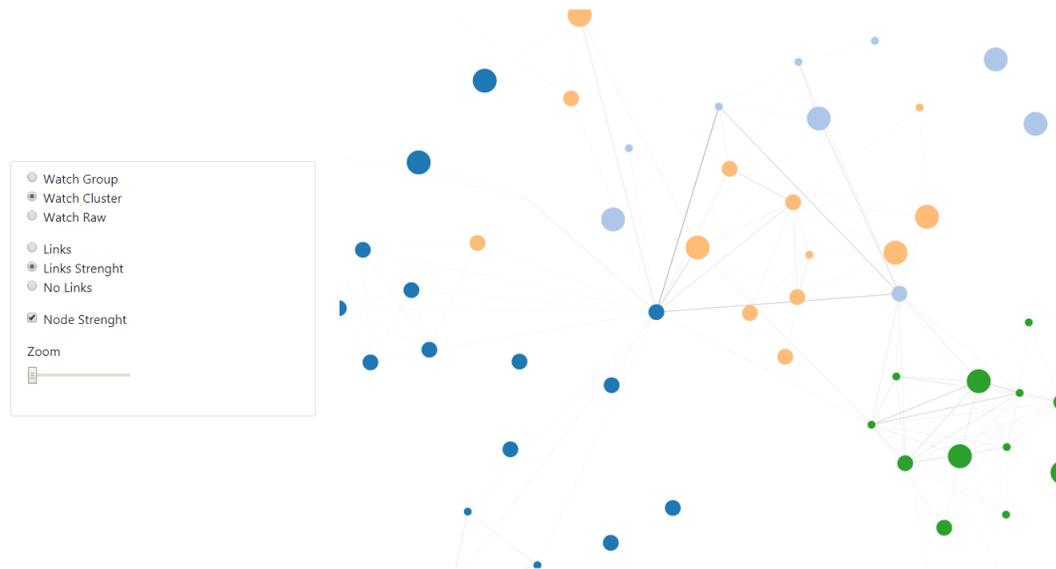


Figura 9.5: Diagnoses Data Analysis. Prototipo de herramienta para análisis y visualización de diagnósticos.

Finalmente lo que se tiene es un archivo de texto plano especializado para la construcción de grafos, que puede ser visualizado con herramientas como Gephi, D3.

9.2.1. Herramienta de Análisis y Visualización

Si bien no hace parte del objetivo la construcción de una herramienta de análisis y visualización de los elementos planteados en este trabajo, se presenta un prototipo de una herramienta denominada “*Diagnoses Data Analysis*” (Figura 9.5) y que permite a través de unas sencillas interacciones percibir de manera visual las dinámicas presentes en un conjunto limitado de diagnósticos que se modelan efectivamente mediante una red compleja, siguiendo las reglas establecida para tal propósito.

Se invita al lector a que interactúe con esta herramienta en la siguiente dirección: <https://jhonatanbarrera.github.io/DiagnosesDataAnalysis/>

Capítulo 10

Conclusiones y Recomendaciones

10.1. Conclusiones

- Los datos han dejado de ser un recurso escaso en todas las organizaciones, en donde hoy ocupan una posición de recurso fundamental, renovable y cada vez más abundante, el cual llevado a las instancias adecuadas genera un activo muy valioso para las organizaciones.
- Con el presente trabajo se evidencia que a pesar de existir mallas de validación para los datos reportados en salud, estas no contemplan todas las validaciones necesarias que deberían existir para que se garantice que los registros en RIPS representan la realidad de los diagnósticos sobre la población.
- A través de la maquina de clasificación se permite hacer un rastreo rápido que evidencia diagnósticos registrados en RIPS y que no corresponden al sexo original del individuo atendido. De esta manera se proporciona un filtro de entrada, que garantiza la calidad de los registros diagnósticos de acuerdo a la clasificación del CIE-10, la cual es insumo para el proceso de modelamiento de morbilidad por medio de redes complejas.
- Si bien los RIPS se crearon para llevar procesos de dirección, regulación y control, las inconsistencias que se señalan en sus registros hacen que estas labores se vean sesgadas y/o que se pierda la capacidad inicial con el que fue creado.
- La integración de las fuentes de información RIPS y BDUA permite obtener una sabana de datos en la que se tiene los registros completos de los individuos atendidos, y desde la que es posible evidenciar las atenciones correspondientes a los habitantes del departamento, así como los que se encuentran de paso. Esta información es útil para el tomador de decisiones al momento de realizar sus planes de atención en salud, y para la construcción de las dinámicas de morbilidad.
- Las redes complejas permiten de buena manera el análisis de los diagnósticos como un sistema complejo, desde el cual es posible encontrar por ejemplo flujos dinámicos, diagnósticos prioritarios, entre otros. Información pertinente para la gobernanza a la hora de construir mejores programas de oferta de servicios de salud que den prevención y corrección en los casos que sea posible a la situación de morbilidad presente en la población, a la hora de asignar recursos y en general adoptar medidas para política de salud.
- Este trabajo presenta elementos de algoritmos de clasificación y redes complejas que aportan al sistema de salud en su gobernanza y que permite ver de manera simple la realidad de morbilidad en su población.
- La explotación de datos tienen impactos económicos y sociales. Genera ingresos adicionales al mejorar por ejemplo procesos en las organizaciones, propende en el caso de este trabajo por la mejora de la calidad de vida de las personas a través de la generación de insumos valiosos para la generación de política publica en salud.

10.2. Recomendaciones

- Es conveniente revisar la manera como se establecen los mecanismos de regulación en salud, pues es a partir de estos que se recolecta la información base para la generación de política pública que debería de incidir en la forma como se asignan los recursos y se controla el gasto en salud.
- Es importante especificar estándares que permitan la generación de datos confiables desde la fuente, en la que será necesario como parte fundamental una política de datos, que ambiente una cultura de datos, un marco de especificación que la adopte y un capital humano calificado que la mantenga.
- Se podría integrar al presente trabajo otras fuentes de información como la proporcionada por los eventos de vigilancia en salud pública que den robustez al sistema complejo, y que permita tener un panorama más amplio de la atención en salud prestada.

Bibliografía

- [1] Salutia. *Protocolo para bases de datos fuentes secundarias*. Inf. téc. Salutia, oct. de 2017.
- [2] James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh y Angela Hung Byers. *Big data: The next frontier for innovation, competition, and productivity*. Inf. téc. McKinsey Global Institute, jun. de 2011. URL: https://www.mckinsey.com/~media/McKinsey/Business%20Functions/McKinsey%20Digital/Our%20Insights/Big%20data%20The%20next%20frontier%20for%20innovation/MGI_big_data_full_report.ashx.
- [3] Ben Golub. “Storage in a Big Data World”. En: *Computerworld* (ago. de 2011). URL: <https://www.computerworld.com/article/2470856/infrastructure-management/big-data-smaq-down.html>.
- [4] AGFA HealthCare. “Digital Imaging in the Cloud”. En: *THERE* 12 (2012), pág. 16. URL: http://www.agfahealthcare.com/he/global/en/binaries/THERE_12_tcm541-95647.pdf.
- [5] Reinsel David, Gantz John y Rydning John. *Data Age 2025: The Evolution of Data to Life-Critical Don't Focus on Big Data; Focus on the Data That's Big*. Inf. téc. IDC, mayo de 2017. URL: <https://www.seagate.com/www-content/our-story/trends/files/Seagate-WP-DataAge2025-March-2017.pdf>.
- [6] Steve Lohr. “For Big-Data Scientists, ‘Janitor Work’ Is Key Hurdle to Insights”. En: *The New York Times* (2014).
- [7] Gartner. *Dirty Data’ is a Business Problem, Not an IT Problem, Says Gartner*. Inf. téc. Gartner, mar. de 2007. URL: <https://www.gartner.com/newsroom/id/501733>.
- [8] Hadley Wickham. “Tidy data”. En: *The Journal of Statistical Software* 59 (sep. de 2014). DOI: [10.18637/jss.v059.i10](https://doi.org/10.18637/jss.v059.i10).
- [9] S. Khalid, T. Khalil y S. Nasreen. “A survey of feature selection and feature extraction techniques in machine learning”. En: *2014 Science and Information Conference*. Ago. de 2014. DOI: [10.1109/SAI.2014.6918213](https://doi.org/10.1109/SAI.2014.6918213).
- [10] C. Wu, J. Zhang, O. Sener, B. Selman, S. Savarese y A. Saxena. “Watch-n-Patch: Unsupervised Learning of Actions and Relations”. En: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40.2 (feb. de 2018), pág. 467. DOI: [10.1109/TPAMI.2017.2679054](https://doi.org/10.1109/TPAMI.2017.2679054).

- [11] D. Vidhate y P. Kulkarni. “Cooperative Machine Learning with Information Fusion for Dynamic Decision Making in Diagnostic Applications”. En: *2012 International Conference on Advances in Mobile Network, Communication and Its Applications*. 2012, págs. 70,74. DOI: [10.1109/MNCApps.2012.19](https://doi.org/10.1109/MNCApps.2012.19).
- [12] Y. Jin y B. Sendhoff. “Pareto-Based Multiobjective Machine Learning: An Overview and Case Studies”. En: *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 38.3 (2008), págs. 397-415. DOI: [10.1109/TSMCC.2008.919172](https://doi.org/10.1109/TSMCC.2008.919172).
- [13] C. E. Shannon. “A Mathematical Theory of Communication”. En: *The Bell System Technical Journal* 17.3 (jul. de 1948), págs. 379-423. DOI: <https://doi.org/10.1002/j.1538-7305.1948.tb01338.x>.
- [14] Yves Coudène. *Ergodic Theory and Dynamical Systems*. Ed. por Springer. 2016.
- [15] Matthias Dehmer. *Structural Analysis of Complex Networks*. Ed. por Birkhäuser. Birkhäuser Basel, 2011. DOI: [10.1007/978-0-8176-4789-6](https://doi.org/10.1007/978-0-8176-4789-6).
- [16] P.R. Kumar, Martin J. Wainwright y Riccardo Zecchina. *Mathematical Foundations of Complex Networked Information Systems*. Springer, 2009. DOI: [10.1007/978-3-319-16967-5](https://doi.org/10.1007/978-3-319-16967-5).
- [17] Ernesto Estrada. “Evolutionary Equations with Applications in natural Sciences”. En: Springer, Cham, 2014. Cap. Introduction to Complex Networks: Structure and Dynamics, págs. 93-131. DOI: doi.org/10.1007/978-3-319-11322-7_3.
- [18] Jacques Vallin. *La Demografía*. 41. CELADE, oct. de 1994. URL: <http://archivo.cepal.org/pdfs/1994/S9400508.pdf>.
- [19] Henry S. Shryock. *The Methods and Materials of Demography (Studies in Population)*. Ed. por Jacob Siegel. Academic Press, 1976. 577 págs. DOI: doi.org/10.1016/C2009-0-03142-0. URL: <https://www.elsevier.com/books/the-methods-and-materials-of-demography/siegel/978-0-12-641150-8>.
- [20] Ministerio de Salud y Protección Social. *Resolución Numero 3221 de 2007*. Inf. téc. Ministerio de Salud y Protección Social, sep. de 2007. URL: https://www.minsalud.gov.co/sites/rid/Lists/BibliotecaDigital/RIDE/DE/DIJ/Resoluci%C3%B3n_3221_de_2007.pdf.
- [21] Ministerio de Salud y Protección Social. *Resolución Numero 2232 de 2015*. Inf. téc. Ministerio de Salud y Protección Social, jun. de 2015. URL: <https://www.minsalud.gov.co/sites/rid/Lists/BibliotecaDigital/RIDE/DE/DIJ/resolucion-2232-de-2015.pdf>.
- [22] Ministerio de Salud y Protección Social. *Resolución Numero 3374 de 2000*. Inf. téc. Ministerio de Salud y Protección Social, dic. de 2000. URL: https://www.minsalud.gov.co/sites/rid/Lists/BibliotecaDigital/RIDE/DE/DIJ/Resoluci%C3%B3n_3374_de_2000.pdf.

- [23] Ministerio de Salud y Protección Social. *Catálogo de patologías - Tabla de CIE-10*. Inf. téc. Ministerio de Salud y Protección Social, ene. de 2018. URL: <https://www.minsalud.gov.co/sites/rid/Lists/BibliotecaDigital/RIDE/DE/OT/tabla-cie-10.zip>.
- [24] Leo Breiman. “Random Forests”. En: *Machine Learning* (2001), págs. 5-32.
- [25] Mr. Bayes y Mr. Price. “An Essay towards Solving a Problem in the Doctrine of Chances”. En: *Philosophical Transactions of the Royal Society of London* 53.0 (1763), págs. 370-418. DOI: [10.1098/rstl.1763.0053](https://doi.org/10.1098/rstl.1763.0053).
- [26] Musa Mammadov y Sona Taheri. “Structure Learning of Bayesian Networks Using a New Unrestricted Dependency Algorithm”. En: (ene. de 2012).
- [27] Laszlo Toth, Andras Kocsor y Janos Csirik. “On naive Bayes in speech recognition”. En: *International Journal of Applied Mathematics and Computer Science* 15.2 (2005), 287-294. ISSN: 1641-876X.
- [28] Daniel Jurafsky y James Martin. “Speech and Language Processing”. En: 2017. Cap. Naive Bayes and Sentiment.
- [29] Juan Pablo Cárdenas, Gastón Olivares y Rodrigo Alfaro. “Automatic text classification using words networks”. En: *Revista Signos. Estudios de Lingüística* (2014).
- [30] YY Song e Y Lu. “Decision tree methods: applications for classification and prediction”. En: *Shanghai Arch Psychiatry* (abr. de 2015), pág. 130. DOI: [10.11919/j.issn.1002-0829.215044](https://doi.org/10.11919/j.issn.1002-0829.215044).
- [31] Musa Mammadov y Sona Taheri. “Structure Learning of Bayesian Networks Using a New Unrestricted Dependency Algorithm”. En: (2012).