

UNIVERSIDAD DE SONORA DIVISIÓN DE INGENIERÍA



POSGRADO EN INGENIERÍA INDUSTRIAL MAESTRÍA EN INGENIERÍA EN SISTEMAS Y TECNOLOGÍA

DESARROLLO DE ÍNDICE DE ESTADO DE SALUD EN
DERECHOHABIENTES DEL ISSSTESON MEDIANTE CARACTERÍSTICAS
BIOMÉTRICAS Y CONSULTAS A MÉDICO GENERAL

T E S I S

PRESENTADA POR

LYLA BERENICE MORALES VILLALBA

Desarrollada para cumplir con uno de los
requerimientos parciales para obtener
el grado de Maestra en Ingeniería

DIRECTORA DE TESIS
DRA. RAQUEL TORRES PERALTA

CODIRECTOR
DR. ALONSO PEREZ SOLTERO

HERMOSILLO, SONORA, MÉXICO.

DICIEMBRE 2020

Universidad de Sonora

Repositorio Institucional UNISON



**"El saber de mis hijos
hará mi grandeza"**



Excepto si se señala otra cosa, la licencia del ítem se describe como openAccess



Hermosillo, Sonora a 1 de diciembre de 2020


LYLA BERENICE MORALES VILLALBA

Con fundamento en el artículo 66, fracción III, del Reglamento de Estudios de Posgrado vigente, otorgamos a usted nuestra aprobación de la fase escrita del examen de grado, como requisito parcial para la obtención del Grado de Maestra en Ingeniería: Ingeniería en Sistemas y Tecnología.


Por tal motivo este jurado extiende su autorización para que se proceda a la impresión final del documento de tesis: **DESARROLLO DE ÍNDICE DE ESTADO DE SALUD EN DERECHOHABIENTES DEL ISSSTESON MEDIANTE CARACTERÍSTICAS BIOMÉTRICAS Y CONSULTAS A MÉDICO GENERAL** y posteriormente efectuar la fase oral del examen de grado.

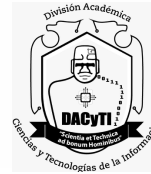
ATENTAMENTE


Dra. Raquel Torres Peralta
Directora de tesis y Presidente del jurado


Dr. Alonso Pérez Soltero
Codirector y Vocal del Jurado


Dr. Federico Miguel Cirett Galán
Secretario del Jurado


Dr. Mario Barceló Valenzuela
Vocal del Jurado



Villahermosa, Tabasco, México, a 30 de noviembre de 2020

LYLA BERENICE MORALES VILLALBA

Con fundamento en el artículo 66, fracción III, del Reglamento de Estudios de Posgrado de la Universidad de Sonora, otorgo a usted mi aprobación de la fase escrita del examen de grado, como requisito parcial para la obtención del Grado de Maestra en Ingeniería: Ingeniería en Sistemas y Tecnología.

Por tal motivo, como sinodal externo y vocal del jurado, extiendo mi autorización para que se proceda a la impresión final del documento de tesis: **DESARROLLO DE ÍNDICE DE ESTADO DE SALUD EN DERECHOHABIENTES DEL ISSSTESON MEDIANTE CARACTERÍSTICAS BIOMÉTRICAS Y CONSULTAS A MÉDICO GENERAL** y posteriormente efectuar la fase oral del examen de grado.

ATENTAMENTE

DR. MIGUEL ANTONIO WISTER OVANDO
UNIVERSIDAD JUÁREZ AUTÓNOMA DE TABASCO
Sinodal Externo y Vocal del Jurado

RESUMEN

El presente proyecto se desarrolla en ISSSTESON (Instituto de Seguridad y Servicios Sociales de los Trabajadores del Estado de Sonora), la cual se encuentra recabando datos desde hace años de las consultas realizadas a médicos generales por derechohabientes, diagnóstico de visita y características biométricas. ISSSTESON ha permitido el estudio de estos datos en los últimos años para mejorar el servicio (Llanes, 2018; Sanez, 2018), por lo que en el presente estudio se enfocó a las visitas frecuentes ya que se presentaban pacientes que asistían varias veces al año por distintos problemas, pero cada vez son atendidos por un médico diferente, de modo que podría no relacionar su padecimiento con los anteriores. Esta situación puede llevar a una persona a asistir hasta 60 consultas al año sin encontrar la causa subyacente.

El proyecto se desarrolló utilizando la metodología CRISP-DM (Metodología de proceso intersectorial para la minería de datos) (Lin, 2017) para segmentar pacientes de acuerdo al riesgo en la salud en derechohabientes según consultas a médico general, tomando en cuenta características biométricas, comorbilidades, frecuencia de visitas y reincidencia; y utilizando técnicas de minería de datos.

Con ello, se obtuvo con redes bayesianas la predicción de padecimientos; presentando correlación predictiva mayor al 20% en diez padecimientos y en promedio de los 38 padecimientos evaluados 16.75%. Se realizó segmentación de derechohabientes según número de comorbilidades para identificar aquellos con mayor riesgo. Para su uso diario, se desarrolló una plataforma tipo tablero donde se ingresa al derechohabiente durante su visita y aparecen avisos y categorías de riesgo de derechohabiente con semáforos para identificar rápidamente qué aspectos de su histórico médico se deberán considerar. Con ello, la plataforma busca asistir en el diagnóstico del derechohabiente y aportar información valiosa al departamento de medicina preventiva.

ABSTRACT

This research establishes a continued effort (Llanes, 2018; Sanez, 2018) with a public medical institute of utilizing electronic health records (EHR) to impact the institute efficiency. Data included general practitioner (GP) visits, diagnosis, and patient demographics. The main problem investigated is recurring visits to different GP with different diagnosis since the later GP may not relate historical visits with current assessments. This situation can lead a patient to attend up to 60 consultations a year without finding the underlying cause.

CRISP-DM (CRoss Industry Standard Process for Data Mining) (Lin, 2017) is the methodology used to segment and establish a health index using the data and to identify which techniques were adaptable with the data and the problem. As a result, 38 health conditions were analyzed with Bayesian network obtaining a predictive correlation greater than 20% in ten health conditions and an average of 16.75%. Segmentation by k-means identified the characteristics related with higher number of comorbidities and therefore with higher health risk.

An interactive dashboard was developed for GP's daily use. After entering the patient's code, critical health characteristics; from historical visits' behavior, patient demographics to past diagnosis were evaluated and presented in a single view. Risk characteristics shown in patient were highlighted and future health conditions with a high predictive probability were reported. With this, the dashboard seeks to assist in the diagnosis of the patient and to provide valuable information to the preventive medicine department.

AGRADECIMIENTOS

A mi esposo por el apoyo brindado en todo momento, tanto en aspectos académicos como personales.

A mis padres por siempre acompañarme y estar ahí para cualquier eventualidad, y a mi familia sanguínea y política que me apoyaron y animaron en esta etapa de mi vida.

A mi directora de tesis, Dra. Raquel Torres, por su disposición, por sus valiosas retroalimentaciones y cursos impartidos, por darme la oportunidad de participar en este proyecto de investigación con esta institución y por su apoyo moral durante este tiempo.

A mi codirector y coordinador del posgrado, Dr. Alonso Pérez, por su profesionalismo y eficiencia en todos los procesos académicos y administrativos.

Al ISSSTESON por confiarme los datos de la institución y a su comité de ética por su valiosa retroalimentación.

A mis profesores por ayudarme directa o indirectamente en la redacción del artículo y en mi conocimiento para desempeñarme mejor laboralmente.

A mis amigos y compañeros de maestría con quienes pude disfrutar estos dos años y trabajar en equipo.

Al Consejo Nacional de Ciencia y Tecnología (CONACYT) y al Programa de Fortalecimiento de la Calidad Educativa (PFCE) por su apoyo económico brindado en mi estudio de posgrado.

ÍNDICE GENERAL

RESUMEN	ii
ABSTRACT	iii
AGRADECIMIENTOS	iv
ÍNDICE GENERAL	v
ÍNDICE DE FIGURAS	viii
ÍNDICE DE TABLAS	xii
1. INTRODUCCIÓN.....	1
1.1. Presentación	1
1.2. Planteamiento del problema.....	2
1.2.1 Descripción del problema	2
1.2.2 Formulación del problema	3
1.3. Objetivo general	3
1.4. Objetivos específicos	3
1.5. Hipótesis	4
1.6. Alcances y delimitaciones	4
1.7. Justificación.....	4
2. MARCO DE REFERENCIA	6
2.1. Servicio médico.....	6
2.2. Análisis de datos y la salud.....	7
2.2.1 Uso de expediente clínico electrónico (ECE) para la investigación	8
2.2.2 Análisis de enfermedades	10
2.2.3 Algoritmos utilizados en investigación de expedientes clínicos	12

2.3.	Índice de estado de salud	14
2.4.	Metodología para análisis de datos.....	20
2.5.	Estudios previos.....	21
3.	METODOLOGÍA.....	24
3.1.	Inmersión en la información y en el instituto	25
3.2.	Análisis de los datos.....	27
3.3.	Identificación de herramientas	30
3.4.	Diseño de metodología	32
3.5.	Evaluación de la metodología	34
4.	IMPLEMENTACIÓN.....	36
4.1.	Inmersión en el instituto	36
4.1.1	Entender el negocio.....	36
4.1.2	Enfoque analítico	37
4.1.3	Requerimientos de datos.....	37
4.1.4	Recolección de datos	38
4.2.	Análisis de datos	38
4.2.1	Entendimiento de datos	39
4.2.2	Preparación de datos.....	63
4.3.	Identificación de herramientas	63
4.3.1	Análisis de algoritmos.....	64
4.3.2	Selección de algoritmos.....	64
4.4.	Diseño de metodología	66
4.4.1	Modelado.....	66
4.4.2	Diseño de plataforma.....	88

4.5.	Evaluación de metodología	93
4.5.1	Evaluar la metodología	93
4.5.2	Obtener retroalimentación	96
5.	CONCLUSIONES, RECOMENDACIONES Y TRABAJOS FUTUROS.....	97
5.1.	Conclusiones.....	97
5.2.	Recomendaciones.....	98
5.3.	Trabajos futuros	98
6.	REFERENCIAS	100
7.	ANEXOS.....	107
7.1.	Comorbilidad de Charlson.....	107
7.2.	Comorbilidad de Elixhauser	108
7.3.	Índice de fragilidad CSHA	110
7.4.	Plataforma de análisis exploratorio de datos.....	111
7.5.	Capítulos de ICD-10.....	113
7.6.	Grupos de categorías ICD-10	114
7.7.	Regresión logística en diagnósticos de interés	115
7.8.	Análisis de comorbilidades de Elixhauser según padecimiento de interés, edad y sexo.....	118
7.8.1	Relación general entre comorbilidades.....	119
7.8.2	Relación de comorbilidades por sexo	120
7.8.3	Relación de comorbilidades por edad.....	122
7.8.4	Relación de comorbilidades en derechohabientes con trastorno de mama	126
7.8.5	Conclusiones	127

ÍNDICE DE FIGURAS

Figura 2.1. Artículos publicados por PubMed con la palabra “EHR” desde el 2000	9
Figura 2.2. Ejemplo de red bayesiana (Sucar, 2006)	13
Figura 2.3. Relación entre comorbilidad y otros constructos	15
Figura 2.4. Supervivencia estimada a 10 años según puntos calculados de índice de comorbilidad de Charlson.....	17
Figura 2.5. Codificación de enfermedades según ICD-10.....	19
Figura 2.6. Diagrama de metodología de CRISP-DM	20
Figura 2.7. Diagrama de metodología CRISP-DM según IBM	21
Figura 3.1. Metodología propuesta.....	24
Figura 3.2. Inmersión en la información y en el instituto	25
Figura 3.3. Ejemplos de enfoques analíticos.....	27
Figura 3.4. Análisis de los datos.....	27
Figura 3.5. Cuarteto de Ascombe.....	28
Figura 3.6. Tipos de visualización de datos.....	29
Figura 3.7. Identificación de herramientas	31
Figura 3.8. Comparación de algunos algoritmos de análisis de conglomerados (Seif, 2019)	32
Figura 3.9. Diseño de metodología	33
Figura 3.10. Evaluación de la metodología	34
Figura 4.1. Visitas diagnóstico y generales según grupo de interés.....	42
Figura 4.2. Días entre visitas generales de pacientes con diagnóstico de trastorno de mama	43
Figura 4.3. Días entre visitas generales de pacientes con diagnóstico de diabetes..	43
Figura 4.4. Visitas por periodo por derechohabiente y número de derechohabientes registrados por periodo (derechohabientes con diabetes = 3,502, con trastorno de mama = 1,074)	44

Figura 4.5. Visitas de derechohabientes con mayores visitas acumuladas en el mes (cada punto es una visita) (a) visitas diagnóstico y generales, (b) visitas en periodos de visitas generales.....	46
Figura 4.6. Derechohabientes con número alto de visitas registradas pero diferentes comportamientos (a) visitas diagnóstico y generales, (b) visitas en periodos de visitas generales.....	47
Figura 4.7. Visitas con mediana de reincidencia menor a 15 días	48
Figura 4.8. Edades y géneros de pacientes con diabetes y trastorno de mama	49
Figura 4.9. Edades y géneros de pacientes con trastorno de mama	49
Figura 4.10. Edades y géneros de pacientes con diabetes	50
Figura 4.11. Interconsulta según grupo de derechohabientes (top 10)	51
Figura 4.12. Prescripción de receta durante visita según grupo de derechohabientes	52
Figura 4.13. Cantidad de recetas por visita según grupo de derechohabientes.....	52
Figura 4.14. Cantidad de medicamentos por visita según grupo de derechohabientes	54
Figura 4.15. Orden de clasificación de enfermedades ICD10	54
Figura 4.16. Capítulos de diagnósticos (ICD10) de derechohabientes con trastorno de mama (top 10).....	55
Figura 4.17. Categorías de diagnósticos (ICD10) de derechohabientes con trastorno de mama (top 10).....	56
Figura 4.18. Capítulos de diagnósticos (ICD10) de derechohabientes con diabetes (top 10).....	56
Figura 4.19. Categorías de diagnósticos (ICD10) de derechohabientes con diabetes (top 10).....	57
Figura 4.20. Trastornos identificados en derechohabientes según grupo de interés (Índice de comorbilidad de Elixhauser)	58
Figura 4.21. Trastornos identificados en derechohabientes según grupo de interés (índice de comorbilidad de Charlson).....	59

Figura 4.22. Comorbilidades de Elixhauser según grupo de edad y grupo de derechohabientes.....	60
Figura 4.23. Trastornos de índice de comorbilidad Elixhauser identificados según sexo y grupo de derechohabientes.....	61
Figura 4.24. Cantidad de medicinas prescrita en las visitas de derechohabientes con respecto a número de trastornos de índice de comorbilidad Elixhauser y grupo de derechohabientes.....	62
Figura 4.25. Trastornos presentados en derechohabientes según índice de Elixhauser	63
Figura 4.26. Evaluación de modelos para describir derechohabientes con diabetes	67
Figura 4.27. Evaluación de modelos para describir derechohabientes con trastorno de mama	68
Figura 4.28. Red bayesiana de trastorno de mama	70
Figura 4.29. Red bayesiana de diabetes.....	71
Figura 4.30. Red bayesiana de grupo de categorías (IDC-10).....	73
Figura 4.31. Diagnósticos que explican a la Diabetes Mellitus según análisis de redes bayesianas	75
Figura 4.32. Diagnósticos que explican la edad según análisis de redes bayesianas	75
Figura 4.33. Principales diagnósticos que afectan la edad según redes bayesianas	76
Figura 4.34. Principales diagnósticos que afectan el sexo según análisis de redes bayesianas	77
Figura 4.35. Principales diagnósticos que afectan las enfermedades infecciosas intestinales	77
Figura 4.36. Principales diagnósticos que afectan las enfermedades infecciosas intestinales	78
Figura 4.37. Principales diagnósticos que afectan las enfermedades hipertensivas.	78
Figura 4.38. Número óptimo de grupos.....	80
Figura 4.39. Agrupamiento por k-means entre 2 a 5 grupos	80
Figura 4.40. Comorbilidades de Elixhauser según la edad de derechohabientes	82

Figura 4.41. Relación entre edad y cantidad de medicamentos.....	83
Figura 4.42. Número de visitas en los últimos 30 días y distribución de visitas acumuladas.....	84
Figura 4.43. Aspectos que afectan el índice de estado de salud	86
Figura 4.44. Gráfico de visitas históricas de paciente con acumulados de visitas en el último mes.....	87
Figura 4.45. Características de evaluación de derechohabiente.....	88
Figura 4.46. Plataforma diseñada para el médico general	89
Figura 4.47. Índice de estado de salud de derechohabientes entre 0 a 100%.....	89
Figura 4.48. Índice de estado de salud y variables numéricas relacionadas.....	90
Figura 4.49. Variables no numéricas relacionadas con el índice de estado de salud y el número de exámenes.....	90
Figura 4.50. Visitas históricas a médico general	91
Figura 4.51. Número de comorbilidades identificadas en el derechohabiente según el índice de Elixhauser	91
Figura 4.52. Histórico de visitas de derechohabiente seleccionado	92
Figura 4.53. Predicción de enfermedades según red bayesiana con porcentaje de predicción.....	92
Figura 4.54. Plataforma diseñada con usuario de 100% de índice de salud.....	93
Figura 7.1. Lista de variables para el índice de fragilidad CSHA (Rockwood et al., 2005)	110
Figura 7.2. Interactive Data Exploratory Analysis and Reporting (IDEAR) (Guo, 2017)	112
Figura 7.3. Derechohabientes con al menos una comorbilidad de Elixhauser según género y grupo de edad	118

ÍNDICE DE TABLAS

Tabla 2.1. Funciones de expedientes clínicos electrónicos requeridos.....	9
Tabla 2.2. Uso de expedientes clínicos electrónicos en la investigación	10
Tabla 2.3. Índice de comorbilidad de Charlson	16
Tabla 2.4. Índice de comorbilidad de Elixhauser	18
Tabla 2.5. Modelos de evaluación de fragilidad para adultos mayores	19
Tabla 3.1. Ejemplo de características de datos por cada conjunto de datos relevante	29
Tabla 3.2. Ejemplo de reporte de modelo para presentar	33
Tabla 4.1. Registros de visitas totales de derechohabientes	40
Tabla 4.2. Diagnósticos para trastornos de mama	40
Tabla 4.3. Diagnósticos para diabetes	41
Tabla 4.4. Periodos de visitas generales a médico general	41
Tabla 4.5. Estadísticas descriptivas de derechohabientes con mayores visitas acumuladas en el mes	45
Tabla 4.6. Principales diagnósticos generales presentados por derechohabientes de ambos grupos.....	57
Tabla 4.7. Ejemplo de tabla de evaluación de grupos de categorías de IDC-10 de derechohabientes.....	67
Tabla 4.8. Ejemplo de tabla de evaluación de comorbilidades de derechohabientes	69
Tabla 4.9. Códigos de comorbilidades de índice de Elixhauser	71
Tabla 4.10. Ejemplo de tabla de evaluación de grupos de categorías de IDC-10 de derechohabientes.....	72
Tabla 4.11. Resultados de red bayesiana con “Hill-climbing” y TABU	74
Tabla 4.12. Ejemplo de datos para evaluar por k-means	78
Tabla 4.13. Ejemplo de datos normalizados para evaluar por k-means.....	79
Tabla 4.14. Grupos según k-means y centros de los grupos según cada variable ...	81
Tabla 4.15. Correlación entre medicamentos, recetas, edad y comorbilidades de Elixhauser.....	81

Tabla 4.16. Evaluación de riesgo según número de comorbilidades de Elixhauser..	85
Tabla 4.17. Evaluación de riesgo según grupo perteneciente del análisis de conglomerados.....	85
Tabla 7.1. Códigos de enfermedades para calcular el índice de comorbilidad de Charlson (Gasparini, 2019)	107
Tabla 7.2. Códigos de enfermedades para calcular el índice de comorbilidad de Elixhauser (Gasparini, 2019).....	109
Tabla 7.3. Capítulos de ICD-10.....	113
Tabla 7.4. Grupos de categorías ICD-10 más frecuentes en derechohabientes.....	114
Tabla 7.5. Matriz de confusión para regresión logística de derechohabientes con diabetes.....	116
Tabla 7.6. Matriz de confusión para regresión logística de derechohabientes con trastorno de mama.	117
Tabla 7.7. Combinaciones de comorbilidades de Elixhauser en derechohabientes	119
Tabla 7.8. Combinaciones de comorbilidades de Elixhauser en derechohabientes (mujeres)	121
Tabla 7.9. Combinaciones de comorbilidades de Elixhauser en derechohabientes (hombres)	122
Tabla 7.10. Grupos de edades para análisis de relación de comorbilidades	122
Tabla 7.11. Combinaciones de comorbilidades de Elixhauser en derechohabientes (15 – 39 años)	123
Tabla 7.12. Combinaciones de comorbilidades de Elixhauser en derechohabientes (40 – 49 años)	124
Tabla 7.13. Combinaciones de comorbilidades de Elixhauser en derechohabientes (50 – 64 años)	124
Tabla 7.14. Combinaciones de comorbilidades de Elixhauser en derechohabientes (65 años o más).....	125
Tabla 7.15. Combinaciones de comorbilidades de Elixhauser en derechohabientes (trastorno de mama).....	126

1. INTRODUCCIÓN

La recolección de datos se ha vuelto una tendencia en las organizaciones para que éstas puedan analizar su historia y definir hacia dónde deben de ir. Los datos recolectados son de gran importancia para una institución siempre que se puedan traducir en acciones para mejorarla; pudiendo ser en reducción de costos, toma de decisiones fundamentada, mejoría en el servicio, entre otros.

En este capítulo se presenta la situación de un instituto de salud pública del estado de Sonora, que a pesar de recolectar datos y utilizarlos para acciones de mejora, sigue en la búsqueda de nuevas formas de interpretar y utilizar la información para mejora continua en sus servicios y procesos. En esta tesis se pretende desarrollar una metodología que permita agrupar a los pacientes de acuerdo con su situación de salud con el fin de saber el estado de los pacientes, predecir la demanda de servicios y ajustar programas personalizados de medicina preventiva. A continuación, se presentan los objetivos, hipótesis, alcances y justificación de esta investigación, que es parte de un proyecto en el que se ha estado trabajando previamente.

1.1. Presentación

El ISSSTESON (Instituto de Seguridad y Servicios Sociales de los Trabajadores del Estado de Sonora) atiende a más de 180,000 derechohabientes presentando una mayor carga poblacional en mayores de 65 años (más de 21,761), teniendo este grupo de edad mayor demanda de servicio médico. ISSSTESON cuenta con 42 unidades médicas, 97 consultorios de médico general y 213 médicos generales o familiares (Gobierno del Estado de Sonora, 2019; ISSSTESON, 2017).

ISSSTESON se encuentra recabando datos desde hace años de las consultas realizadas a médicos generales por derechohabientes, diagnóstico de visita y características biométricas (análisis de sangre, peso, talla, y evolución de respuesta a tratamiento). Si bien se están recabando, esta información no se ha enfocado a la toma de decisiones con respecto a visitas frecuentes. Por ello, no se han analizado para

tener una visión del nivel de salud de los pacientes, quienes asisten varias veces al año por distintos problemas, pero cada vez son atendidos por un médico diferente, de modo que podría no relacionar un padecimiento con otro. Esta situación puede llevar a una persona a asistir hasta 60 consultas al año sin encontrar la causa subyacente.

Un objetivo estratégico de la institución es innovar el modelo de gestión de los servicios médicos bajo varias estrategias y líneas de acción; una de éstas siendo evaluar la prestación de servicio médico para la toma de decisiones efectiva con el fin de garantizar oportunidad y eficiencia. Soluciones de baja inversión y para aumento de eficiencia estarían alineadas a la situación financiera presentada en el 2016, con el fin de optimizar la utilización de recursos y presentar mejores opciones tanto para el médico como para el paciente, así como para la toma de decisiones a nivel ejecutivo (Gobierno del Estado de Sonora, 2019; ISSSTESON, 2017; 2019).

El no evaluar la frecuencia de visitas según diagnósticos ni alertar usuarios con mayor reincidencia pudiese conllevar a una mayor demanda de médicos generales. Esto pudiera impedir que la población agende citas en menor tiempo, presentar diagnósticos que no consideraron visitas anteriores del derechohabiente, gastos innecesarios de insumos, menor flexibilidad de tiempos para médicos generales, entre otros recursos que pudiesen disminuir al prevenir reincidencias (Secretaría de Salud, 2015; 2017).

1.2. Planteamiento del problema

En este apartado, se presentará la necesidad de la institución y el contexto, así como los objetivos, alcance y justificación.

1.2.1 Descripción del problema

Es necesario explorar maneras de utilizar los datos recabados por el ISSSTESON en busca de maneras de mejorar la eficiencia en los procesos, y asistir en el diagnóstico del médico general o apoyar en la toma de decisiones para fines administrativos o de

políticas en la atención a los derechohabientes para que además de recabar datos, se gestione el conocimiento para la mejora dinámica de la organización.

1.2.2 Formulación del problema

Se han presentado derechohabientes con alta reincidencia en visitas a médico general. Existe la necesidad de evaluar la frecuencia y características de población de visitas a médicos generales según diagnósticos, así como la de priorizar la atención y dar seguimiento adecuado a los usuarios con mayor riesgo en salud que fomente la reincidencia en atención médica.

1.3. Objetivo general

Elaborar una metodología para la evaluación de riesgo en la salud en derechohabientes del ISSSTESON según consultas a médico general, tomando en cuenta características biométricas, comorbilidades, frecuencias de visitas a médico general y reincidencia, utilizando técnicas de minería de datos, con el fin de realizar consultas más eficientes asistidas con información histórica, y reducir reincidencias; además de generar grupos de usuarios de acuerdo a las similitudes entre ellos para campañas de medicina preventiva dirigidas.

1.4. Objetivos específicos

- Analizar características de los registros para un pre-filtrado según pertinencia del estudio y evaluar el estado actual de la población
- Revisar y seleccionar técnicas de procesamiento de registros de visita a médico general para detectar derechohabientes con alta reincidencia de enfermedades o con un número de visitas a médico general más alto de lo esperado
- Desarrollar una metodología para calcular los atributos de salud, así como los métodos y criterios en el procesamiento de registros de visitas con técnicas seleccionadas y utilizarlo para la segmentación de derechohabientes
- Evaluar y validar modelo y resultados obtenidos de las técnicas seleccionadas.

1.5. Hipótesis

La evaluación de riesgo de nivel de salud de los derechohabientes de ISSSTESON podría ayudar a identificar usuarios que requieren mayor atención durante su consulta y asistir en el diagnóstico para prevenir reincidencias, así como crear grupos de pacientes para la mejora en campañas de medicina preventiva dirigida.

1.6. Alcances y delimitaciones

El análisis se realizó utilizando los datos que ISSSTESON proporcionó para su análisis con respecto a las consultas a médicos generales.

En este trabajo se procesaron los registros de la ciudad de Hermosillo, que es donde residen el 33% del estado (991,645 derechohabientes y 144 médicos generales) (Gobierno del Estado de Sonora, 2019).

Aunque los resultados son aplicables a los pacientes de la institución, la metodología es aplicable a cualquier institución de salud que lleve un registro de visitas con los datos analizados.

1.7. Justificación

Se consideran más de 180,000 derechohabientes del ISSSTESON que pudiesen verse beneficiados mediante la implementación de técnicas de análisis y minería de datos, así como de gestión del conocimiento en base al historial de consultas para mejorar el nivel de servicio a largo plazo. Los médicos al recibir al paciente podrían saber de su nivel actual de salud y tener una idea de la clasificación a la que pertenecen de acuerdo con sus padecimientos.

Además, la institución contará con un panorama del estado de salud de los inscritos a la institución, a su vez, de mejorar la eficiencia en las actividades al realizar decisiones mejor informadas a partir de los datos, asistiendo al equipo de medicina preventiva para encaminar sus recursos y campañas de manera selectiva y dirigida. Esta herramienta pudiera asistir en varias áreas, por ejemplo, detectar los casos en que

después de varias consultas la prescripción de medicamentos se incrementa a un mismo paciente por falta de conocimiento del médico sobre sus antecedentes. El impacto de la implementación de estas medidas también influiría de modo favorable en los costos de la institución, sin embargo, ese rubro queda fuera de la evaluación de este estudio. Finalmente, con la implementación de técnicas de inteligencia artificial y minería de datos se pretende mejorar la percepción del servicio y repercutir positivamente, como resultado de una mejor atención en la salud de los derechohabientes.

2. MARCO DE REFERENCIA

Este trabajo forma parte de un proyecto que abarca varios objetivos. Anteriormente se han presentado resultados de un primer enfoque y esta investigación pretende extender el alcance de los objetivos presentados en una primera entrega.

En la presente sección se presentan generalidades del servicio médico, cómo se ha llevado a cabo el análisis de los expedientes clínicos electrónicos (ECE), el análisis de enfermedades utilizando ECE, algoritmos utilizados en la investigación de ECE, índice de estado de salud y finalmente estudios previos.

2.1. Servicio médico

Una característica fundamental del servicio médico que impacta tanto la calidad como el costo es su eficiencia. En términos de eficiencia se pueden considerar dos conceptos: eficiencia asignativa y técnica. La eficiencia de asignación consiste en asignar recursos de tal manera que proporcione la combinación óptima de bienes y servicios para maximizar los beneficios para la sociedad. Mientras tanto, la eficiencia técnica consiste en utilizar la menor cantidad o la combinación correcta de insumos. En términos de salud, se pueden considerar otro término importante para el desarrollo del presente proyecto: eficiencia dinámica en salud, la cual busca asegurar la eficiencia mediante la innovación en el tiempo y mediante el aprendizaje (Cid, C., Báscolo, E. y Morales, C., 2017; Yip y Hafez, 2015).

La innovación y aprendizaje se relaciona con una organización en esta era del conocimiento, ya que es aquella que aprende, recuerda y actúa basado en la información disponible, conocimiento, y mejores prácticas. Los propósitos de la gestión del conocimiento son mejorar la eficiencia organizacional y aumentar la capacidad organizativa para innovar mediante el uso y reutilización del conocimiento (Dalkir, 2017).

2.2. Análisis de datos y la salud

La tecnología puede utilizarse en varias fases de la gestión del conocimiento; desde la generación de conocimiento, hasta el acceso del conocimiento que le permita realizar decisiones informadas de forma oportuna. Para este desarrollo, se estará enfocando en herramientas para la captura y creación de conocimiento, siendo una herramienta muy importante la minería de datos. Éste es un proceso en el que se extrae automáticamente información predictiva de grandes bases de datos basado en análisis estadístico (usualmente análisis de conglomerados). Utilizando una combinación de machine learning (aprendizaje automatizado), análisis estadístico, técnicas de modelado y bases de datos, la extracción de datos puede detectar patrones ocultos y relaciones entre los datos que permitan la predicción futura de resultados (Dalkir, 2017; Witten, Frank y Hall, 2011).

Machine learning o aprendizaje automatizado es un método de análisis de datos que automatiza la construcción de modelos analíticos. Éste espera que los sistemas aprendan de los datos, identifiquen patrones y tomen decisiones con la mínima intervención humana. Análisis de conglomerados busca agrupar observaciones con características en común (Witten, Frank y Hall, 2011).

Machine learning es una rama de la inteligencia artificial. Ésta última es de interés en el tema de salud de México para resolver las cuatro “P”: predictivo, preventivo, personalizado y participativo. A su vez, de forma generalizada se ha vuelto una tendencia gracias a que puede ser utilizada para asistir a médicos en el análisis de tendencias o alertas que puedan mejorar un diagnóstico y tratamiento más acertado (Celis, 2018; Witten, Frank y Hall, 2011).

Para su uso en el servicio médico, la inteligencia artificial explicable considera siete pilares: (1) transparencia (entendible por el usuario), (2) sentido de dominio, (3) coherencia, (4) parsimonia (tan simple como sea posible), (5) generalizable, (6) confianza/rendimiento, y (7) fidelidad. Según la confianza/rendimiento, se debe incluir el error esperado de las explicaciones presentadas por el algoritmo utilizado. Según la

fidelidad se debe de alinear la expectativa de la explicación con el modelo predictivo (Teredesai et al., 2018).

2.2.1 Uso de expediente clínico electrónico (ECE) para la investigación

El presente desarrollo utiliza como base los registros de visitas a médico general en la institución, siendo este no un caso innovador sino una fuente de información que ha estado utilizándose cada vez más para este tipo de estudios. Desde el 2009 ha despuntado la adopción de sistemas de expedientes clínicos electrónicos en hospitales privados de Estados Unidos, llegando en el 2015 a un 83.8%, mientras que otros países como Japón llegó a 34.2% en el 2014 y Corea del Sur a 58.1% en el 2015. Según encuestas a médicos del 2012 al 2015, la adopción ha sido cercana al 99% en Noruega, Suecia, Reino Unido, Países Bajos y Nueva Zelanda. Con ello, se ha habilitado su uso en estudios (Charles, Gabriel y Searcy, 2015; Kanakubo y Kharrazi, 2019; Schoen et al., 2012; The Commonwealth Fund, 2015).

Los expedientes clínicos electrónicos fueron clasificados como básico si presenta el mínimo, pero esencial para su funcionalidad, como básico, básico con notas clínicas y finalmente como exhaustivo. Éstos se presentan en la Tabla 2.1 (Charles, Gabriel y Searcy, 2015).

Funciones requeridas	Básico sin notas clínicas	Básico con notas clínicas	Exhaustivo
Información clínica electrónica			
Demografía del paciente			
Notas del médico			
Evaluación de enfermería			
Lista de problemas			
Lista de medicamentos			
Resúmenes de altas			
Directivas anticipadas			
Orden computarizada			
Informes de laboratorio			
Pruebas de radiología			
Medicamentos			
Solicitudes de consulta			
Órdenes de enfermería			
Gestión de resultados			
Ver reportes de laboratorio			
Ver reportes de radiología			

Funciones requeridas	Básico sin notas clínicas	Básico con notas clínicas	Exhaustivo
Ver imágenes de radiología			
Ver resultados de pruebas de diagnóstico			
Ver reporte de consulta			
Apoyo a decisiones			
Pautas clínicas			
Recordatorios clínicos			
Resultados de alergia a medicamentos			
Interacciones farmacológicas			
Interacciones de laboratorio de drogas			
Soporte de dosificación de drogas			

Tabla 2.1. Funciones de expedientes clínicos electrónicos requeridos

Los formatos más comunes que presentan los registros en orden de utilidad son los siguientes:

1. Estructurado y terminología codificada
2. Estructurada con códigos locales o sin codificación
3. Texto no estructurado legible por máquina (reportes de estudios, ...)
4. Texto no estructurado y escaneado (PDF, ...) (Kim et al., 2019)

Con la adopción de estos expedientes clínicos electrónicos en hospitales e incentivos en países como Estados Unidos para utilizar esta información, su uso en investigación y desarrollo ha aumentado significativamente en los últimos años como se puede ver en la Figura 2.1 (Kim et al., 2019; PubMed, 2019).

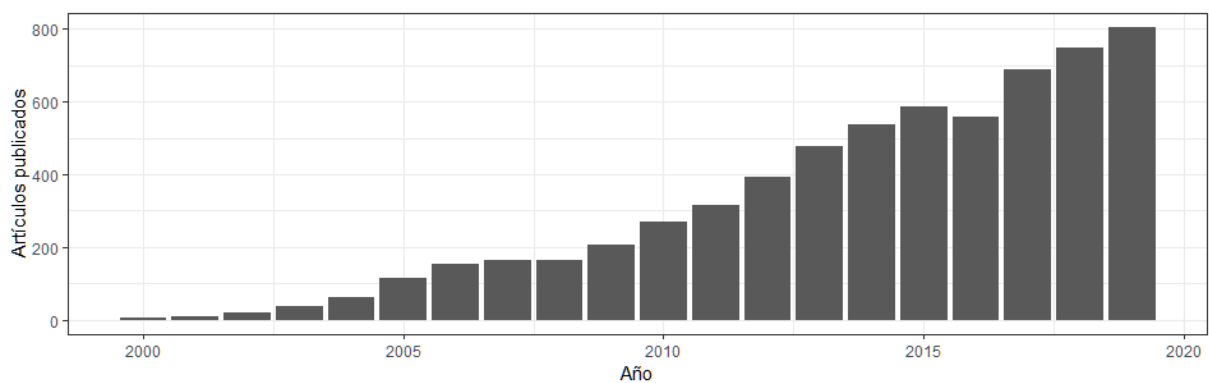


Figura 2.1. Artículos publicados por PubMed con la palabra "EHR" desde el 2000

En el 2007 la Asociación Americana de Informática Médica presentó un panel con recomendaciones de expertos sobre el uso de datos de salud, siendo entre ellas

promover los beneficios y riesgos de utilizar registros de salud hacia un rango más amplio de partes interesadas, crear una taxonomía para usos secundarios de los expedientes clínicos, y enfocar a la atención estatal y nacional de los posibles usos y aplicaciones (Safran et al., 2007).

2.2.2 Análisis de enfermedades

Las aplicaciones pueden llegar a ser infinitas y cada vez más confiables al registrar datos cada vez más completos y confiables. Algunas aplicaciones en investigación se presentan en la Tabla 2.2 (Cowie et al., 2017; Yadav et al., 2018).

Tipo	Ejemplo	Estado
Estudios observacionales	Utilización de la salud	Ampliamente utilizado y aceptado
	Utilización de drogas	
	Epidemiología (incidencia/prevalencia)	
	Historia natural	
	Factores de riesgo	
Vigilancia de seguridad	Vigilancia de seguridad tradicional posterior a la comercialización	Ampliamente utilizado y aceptado
	Vigilancia activa (como Sentinel)	Emergente
Investigación clínica	Generación de hipótesis	Aceptado
	Evaluaciones de viabilidad	
	Mejora de rendimiento, adherencia a las pautas	
	Reclutamiento de pacientes	Emergente
	Efectividad comparativa, evaluaciones de tecnología de salud	
	Ensayos pragmáticos	
	Aleatorización de punto de atención	
	Registro de ensayos aleatorios para probar nuevas intervenciones	Emergente / potencial
	Fuente de datos para llenar eCRF (eliminando o minimizando la necesidad de extracción / ingreso de datos)	
Punto final o verificación de evento adverso grave	Emergente / potencial	
Regulador	Vigilancia de seguridad, farmacovigilancia	Aceptado
	Nuevas indicaciones o autorización de comercialización	Potencial

Tabla 2.2. Uso de expedientes clínicos electrónicos en la investigación

En cuanto a regulación y seguridad, el proyecto EU-ADR compuesto por los países de Dinamarca, Italia, Países Bajos y Reino Unido puso a disposición de terceros ocho bases de datos para el análisis de ciertos eventos farmacológicos adversos utilizando técnicas de minería de datos, demostrando además la factibilidad de armonizar bases de datos de diferentes países (Trifiro et al., 2009). En Estados Unidos, la FDA

(Administración de Alimentos y Medicamentos) utiliza datos de varias fuentes (Sentinel, centros de salud, y otros departamentos) para apoyar en la investigación de seguridad de medicamentos posterior a su comercialización (Cowie et al., 2017; Kim et al., 2019). En México, a pesar de generar mucha información; ésta se encuentra desconectada entre los niveles y diferentes organizaciones. La iniciativa de consolidar de manera nacional los sistemas de información es esencial para las funciones clave; entre ellas, mejora continua de la calidad, atención personalizada y predicción de las necesidades cambiantes de salud. Portugal se presentó como ejemplo de cómo el uso de datos asistió en costos y calidad de numerosas áreas clínicas a partir de estos esfuerzos; como prescripción, cirugía ambulatoria y cuidado en enfermedades crónicas. Una de las iniciativas es IMSS Digital, el cual busca mejorar los expedientes clínicos al homogenizar las bases de datos (OECD, 2016; OPM y OMS, 2015).

A pesar de esto, debido a que los registros médicos no son diseñados para la investigación, por lo que convertir los datos en un formato puede ser tardado y costoso. El procesamiento natural de lenguaje contiene técnicas capaces de convertir la información en datos analizables. El uso de algoritmos utilizando aprendizaje automatizado para generar modelos predictivos en medicina ha aumentado considerablemente en los últimos años (Kim et al., 2019).

Otros desafíos utilizando expedientes clínicos electrónicos es la censura de datos (por ejemplo, observar al paciente en su periodo de tiempo cuando potencialmente podría tener eventos interesantes para el estudio fuera de este periodo y por tanto no disponible para el estudio), fragmentación (que el usuario visite varios proveedores de salud y por ello sólo se tenga información parcial de su historial médico), serie de tiempo irregular (debido a la irregularidad de las visitas de los pacientes), datos faltantes (los cuales pueden ser aquellos no registrados durante la visita, como recomendaciones del doctor al paciente con respecto a su estilo de vida), y sesgos (Yadav et al., 2018).

2.2.3 Algoritmos utilizados en investigación de expedientes clínicos

Los expedientes clínicos electrónicos (ECE) han sido analizados históricamente por varios algoritmos predictivos que serán brevemente explicados a continuación. Con respecto a los algoritmos seleccionados, véase el capítulo de Estudios Previos.

2.2.3.1 Métodos cuantitativos

Regresión

La regresión busca encontrar si hay relación entre dos tipos de variables: dependientes (la variable que se busca entender o predecir) e independientes (aquello que se sospecha tenga un impacto sobre la variable dependiente) (Gallo, 2015).

Análisis de conglomerados

Es una técnica de aprendizaje no supervisada conocida también como clustering que se encarga de agrupar un conjunto de objetos que son más similares entre ellos que en otros grupos. Se utiliza para muchas aplicaciones como en el reconocimiento de patrones, análisis de imágenes, recuperación de información, bioinformática, entre otros. La agrupación se puede hacer de dos tipos: dura y suave. La dura obliga a cada punto a clasificarse en un grupo, mientras que la suave permite que un punto pueda pertenecer a más de un grupo con un valor de probabilidad (Forsyth, 2018; Jaiswal, 2018).

Redes bayesianas

Las redes bayesianas modelan la dependencia de variables de un fenómeno. Se pueden utilizar tanto para predecir, diagnosticar o clasificar (Sucar, 2006). En la Figura 2.2 se presenta la dependencia de tifoidea con respecto a la comida, y cómo ésta afecta en las reacciones, fiebre y dolor.

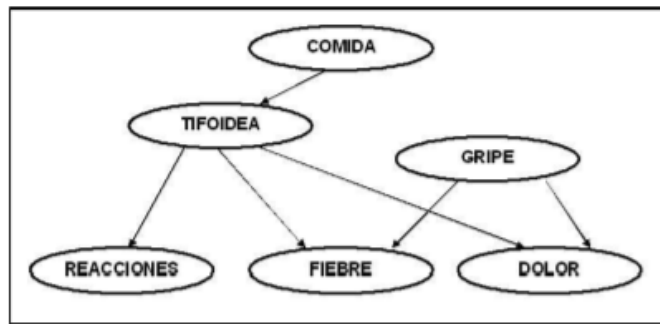


Figura 2.2. Ejemplo de red bayesiana (Sucar, 2006)

2.2.3.2 Métodos cualitativos

Métodos lineales

- Análisis discriminante lineal. Este algoritmo busca reducir las dimensiones (variables) de una base de datos mientras retiene el máximo de información posible (Maklin, 2019).
- Regresión logística. Este algoritmo busca analizar la influencia de variables sobre una respuesta binaria (Härdle y Simar, 2015).
- Modelo lineal generalizado. La regresión logística es un caso especial del modelo lineal generalizado. Este algoritmo busca analizar la influencia de variables sobre una respuesta, siendo en este caso respuesta categórica (Heiberger y Holland, 2015).

Métodos no lineales

- Máquina de soporte vectorial. Este algoritmo busca un hiperplano que divide un conjunto de datos en dos clases (Forsyth, 2018).
- K vecinos más cercanos. Este algoritmo identifica los puntos más cercanos a la localización que se desea clasificar y según la etiqueta de los vecinos se obtiene una aproximación de qué etiqueta se debe utilizar para el punto evaluado (Forsyth, 2018).

Conjuntos de árboles

Los árboles de decisión presentan una red de preguntas que forma una estructura parecida a un árbol donde las puntas de la red es la predicción y las reglas se presentan en las bifurcaciones (Nisbet et al., 2018).

Con respecto a la validación, se presentan los siguientes métodos.

Validación cruzada

Este tipo de validación implica separar datos en subconjuntos donde cada uno de ellos se usa como validación y entrenamiento ambos aleatorios (Nisbet et al., 2018). Primero se entrena el conjunto de entrenamiento y se evalúa con el conjunto de validación para cada iteración (donde cada iteración son conjuntos aleatorios y diferentes). Para evaluar se promedian los errores de todas las iteraciones. Con ello, se obtiene una estimación de cómo se va a comportar el algoritmo en un futuro (Forsyth, 2018).

Bootstrapping

Este tipo de validación genera N tablas nuevas del conjunto de entrenamiento con el mismo tamaño que el original y con reemplazo. El modelo se aplica en cada tabla y se genera la predicción. Después de algunas iteraciones del proceso se obtiene el promedio para calcular el error (Nisbet et al., 2018).

2.3. Índice de estado de salud

El término de estado de salud fue acuñado por Kaplan et al (1976) identificando la necesidad de agregar observaciones múltiples de diversas dimensiones en todo el espectro de disfunciones que las enfermedades y lesiones imponen a todos los miembros de la población. EPA (2018) indica que puede ser medida por diferentes factores: tasas de natalidad y mortalidad, esperanza de vida, calidad de vida, morbilidad por enfermedades específicas, factores de riesgo, uso de atención ambulatoria y de hospitalización, accesibilidad del personal de salud e instalaciones,

financiamiento de la atención médica, cobertura de seguro de salud y muchos otros factores. OECD (2017) mide el estado de salud percibida de 35 países, donde en algunos como México se encuestó utilizando la escala Likert.

La comorbilidad ha sido de interés desde hace décadas, ya que no sólo estudia una sola enfermedad, sino considera las relaciones entre la enfermedad de interés con otras enfermedades (Sarfati, Koczwara y Jackson, 2016). Se podría considerar la comorbilidad como un constructo relacionado a la severidad de la enfermedad y que afecta otros aspectos del paciente, como se muestra en la Figura 2.3 (Sarfati, Koczwara y Jackson, 2016).

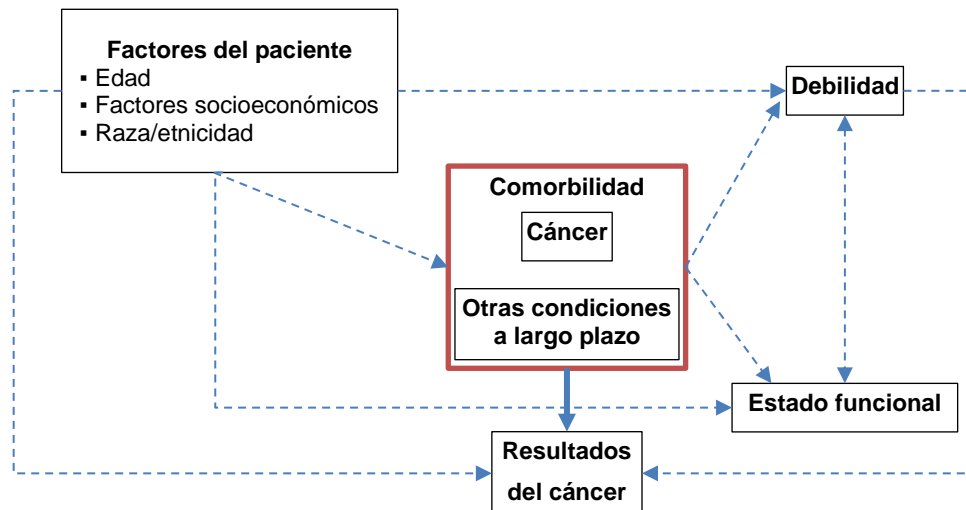


Figura 2.3. Relación entre comorbilidad y otros constructos

En una persona, la comorbilidad representa resultados adversos en su salud, menor calidad de vida, discapacidad, problemas psicológicos y aumento de mortalidad. Por otro lado, para los organismos de salud representa aumento de uso de servicios de emergencia, eventos farmacológicos adversos, polifarmacia, pruebas duplicadas y mala coordinación (Aronin, 2017).

En México, una prioridad es proveer una atención preventiva tanto para condiciones a largo plazo y comorbilidades, por lo que este índice llega a ser pertinente para la clasificación de pacientes (OECD, 2016).

El índice de comorbilidad de Charlson predice la supervivencia a 10 años en pacientes con múltiples comorbilidades y fue creado en 1987 (Daskivich et al., 2018; MDCalc, 2019; Rosas et al., 2017). En la Tabla 2.3 se presentan las comorbilidades que se utilizan para realizar el índice (MDCalc, 2019; Rosas et al., 2017). En Tabla 7.1 se presentan estas enfermedades y sus códigos de enfermedades correspondientes.

Comorbilidad	Calificación				
	< 50 0	50-59 +1	60-69 +2	70-79 +3	≥ 80 +4
Infarto de miocardio	No 0		Sí +1		
Insuficiencia cardíaca congestiva					
Enfermedad vascular periférica					
Accidente cerebrovascular o ataque isquémico transitorio					
Demencia					
Enfermedad pulmonar obstructiva crónica					
Enfermedad del tejido conectivo					
Enfermedad de úlcera péptica					
Enfermedad del hígado	No 0	Leve +1	Moderado a severo +3		
Diabetes	No/dieta controlada 0	Sin complicaciones +1	Daño en órganos +2		
Hemiplejía	No 0		Sí +2		
Enfermedad renal crónica moderada a severa					
Tumor sólido	No 0	Localizado +2	Metastásico +6		
Leucemia	No 0		Sí +2		
Linfoma					
SIDA	No 0		Sí +6		

Tabla 2.3. Índice de comorbilidad de Charlson

El índice contempla 26 puntos posibles, algunos de ellos excluyentes, con lo que el índice lleva el comportamiento de la Figura 2.4 según la puntuación. En éste se puede apreciar que el máximo de puntos que una persona puede tener para tener un porcentaje de supervivencia es de 6, permitiendo hasta 2 puntos para tener $\geq 90\%$ (MDCalc, 2019).

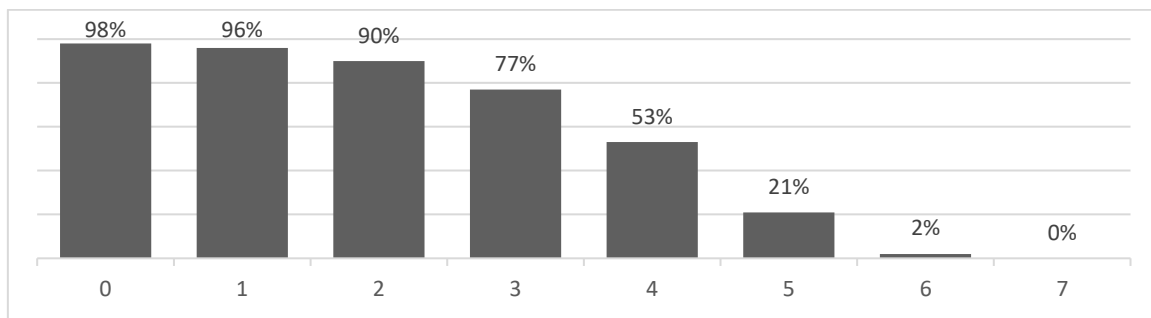


Figura 2.4. Supervivencia estimada a 10 años según puntos calculados de índice de comorbilidad de Charlson

Similar a este índice, se desarrolló en 1998 el índice de comorbilidad de Elixhauser; el cual, en vez de calcular 17 factores, se evalúan entre 30 y 31 factores (según el algoritmo utilizado). Durante el cálculo, las comorbilidades se asociaron con aumentos sustanciales en la duración de la hospitalización, cargos hospitalarios y mortalidad tanto para grupos de enfermedades heterogéneas como para las homogéneas. Se incluyeron aquellos predictores que comúnmente no se miden como trastornos mentales, abuso de drogas y alcohol, obesidad, coagulopatía, pérdida de peso y trastornos de líquidos y electrolitos (Elixhauser et al., 1998). En la Tabla 2.4, se presenta el índice con los pesos según los algoritmos utilizados. En el caso del algoritmo AHRQ, no incluye arritmia cardíaca por lo que se asigna un peso de cero (Gasparini, 2019). Los códigos de las enfermedades utilizadas para el algoritmo se encuentran en la Tabla 7.2.

Comorbilidad	Algoritmo AHRQ	Algoritmo van Walraven
Insuficiencia cardíaca congestiva	9	7
Arritmia cardíaca	0	5
Enfermedad valvular	0	-1
Trastornos de la circulación pulmonar.	6	4
Trastornos vasculares periféricos.	3	2
Hipertensión (combinada sin complicaciones y complicada)	-1	0
Parálisis	5	7
Otros trastornos neurológicos.	5	6
Enfermedad pulmonar crónica	3	3
Diabetes no complicada	0	0
Diabetes complicada	-3	0

Comorbilidad	Algoritmo AHRQ	Algoritmo van Walraven
Hipotiroidismo	0	0
Insuficiencia renal	6	5
Enfermedad del hígado	4	11
Úlcera péptica, excluyendo sangrado	0	0
SIDA / VIH	0	0
Linfoma	6	9
Cáncer metastásico	14	12
Tumor sólido sin metástasis	7	4
Artritis reumatoide / enfermedades vasculares del colágeno	0	0
Coagulopatía	11	3
Obesidad	-5	-4
Pérdida de peso	9	6
Trastornos de líquidos y electrolitos.	11	5
Anemia por pérdida de sangre	-3	-2
Anemia por deficiencia	-2	-2
Abuso de alcohol	-1	0
Abuso de drogas	-7	-7
Psicosis	-5	0
Depresión	-5	-3

Tabla 2.4. Índice de comorbilidad de Elixhauser

Ambos índices se podrían considerar adaptables a los formatos de expedientes clínicos electrónicos ya que se pueden calcular utilizando códigos de enfermedades internacionales como el ICD-10. Considerando que los historiales de consultas a médico general presentan diagnóstico de visita, se podría llegar a una evaluación del estado de salud o riesgo estimado que sería de utilidad para el estudio.

El ICD-10 fue creado desde 1900, siendo su décima edición en 1995. Éste último presenta 22 capítulos, éstos divididos en 271 grupos de categorías, y éstos a su vez en 2,030 categorías con 14,198 subcategorías.

- Capítulo se relaciona con la letra, siendo 22 en total
- Grupo de categoría se relaciona con los primeros tres caracteres y representan enfermedades similares entre ellas
- Categoría se relaciona con los primeros tres caracteres y presenta la enfermedad de manera generalizada

- Subcategoría se relaciona con el número después del decimal que presenta alguna especificación adicional a la enfermedad (CEMECE, 2014)

En la Figura 2.5 se presentan los diferentes niveles que se maneja el análisis. En la Tabla 7.3 se presentan los capítulos y en la Tabla 7.4 los grupos de ICD-10.

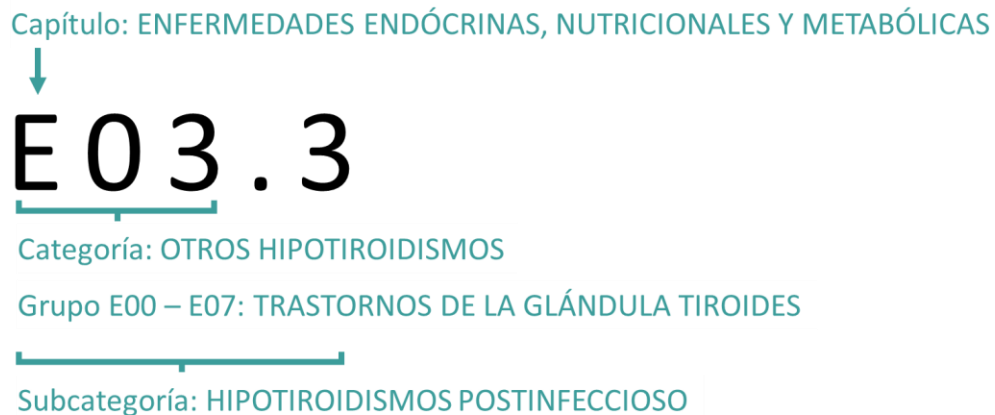


Figura 2.5. Codificación de enfermedades según ICD-10

Además de los índices de comorbilidad, existen otros predictores de resultados negativos para la salud como los modelos de evaluación de fragilidad para adultos mayores. En la Tabla 2.5 se presentan algunos de ellos con una breve descripción (Pijpers et al., 2012).

Modelo	Evaluación
<i>Edad y modalidad única</i>	
3MS	Estudio cognitivo para evaluar demencia (escritura, lectura, operaciones mentales, orientación temporal y espacial, agilidad mental, memoria, ...)
Escala de fragilidad CSHA	Evalúa cumulativamente desde atlético (activo enérgico, sin enfermedades) hasta gravemente frágil (dependiente de terceros) Véase Figura 7.1 para variables.
<i>Síndrome biológico</i>	
Índice de fragilidad CSH (estudio de salud cardiovascular)	Se evalúa por pérdida de peso, energía reducida, fuerza de agarre reducida, velocidad lenta para caminar y baja energía física.
Índice SOF (estudio de fracturas osteoporóticas)	Se evalúa por pérdida de peso, habilidad de pararse de la silla cinco veces sin usar brazos y energía reducida
<i>Definido por lugar</i>	
Diagnóstico para pacientes psico geriátricos	Se evalúa según edad, sexo, cuidador, IMC, enfermedad cardiovascular, movilidad, proceso y habilidades motoras, polifarmacia
Detección de riesgo de triaje	Se evalúa por deterioro cognitivo, dificultad para caminar, transferencia, cuidador, polifarmacia, hospitalización reciente
Identificación de adultos mayores en riesgo	Se evalúa por deterioro funcional pre-mórbido, deterioro funcional agudo, hospitalización, visión y memoria, ≥3 medicamentos
Indicativo variable de riesgo de colocación	Se evalúa por su situación de vida, estado funcional y cognitivo

Tabla 2.5. Modelos de evaluación de fragilidad para adultos mayores

2.4. Metodología para análisis de datos

Se revisaron algunas metodologías para análisis de datos, de los cuales se presenta CRISP-DM (Metodología de proceso intersectorial para la minería de datos). Esta metodología creada desde 1996 ha sido comprobada para guiar en el desarrollo de proyectos de minería de datos. En sus inicios se presentan en seis pasos como se muestra en la Figura 2.6 (Wirth y Hipp, 2000).

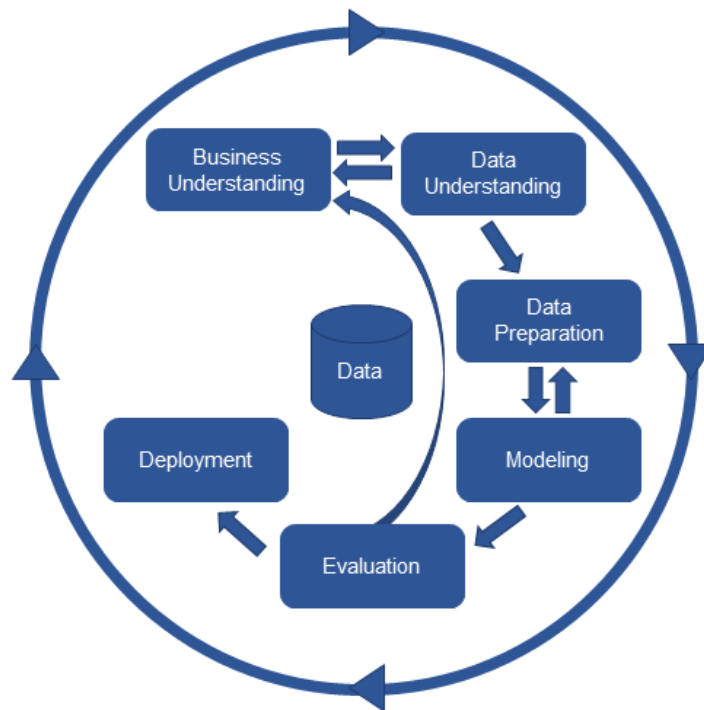


Figura 2.6. Diagrama de metodología de CRISP-DM

En IBM se dividieron los aspectos de los datos en pasos complementarios, los cuales se presentan en la Figura 2.7. Éstos se definen brevemente a continuación:

- 1) **Entender del negocio** – principalmente el objetivo a seguir y cómo se puede evaluar el éxito del proyecto;
- 2) **Enfoque analítico** – se expresa el problema en el contexto estadístico;
- 3) **Requerimientos de datos** – en base al problema determinado se identifica el formato y contenido necesario para su análisis;

- 4) **Recolección de datos** – tanto los disponibles como los que se requiere recolectar;
- 5) **Entendimiento de datos** – utilizando estadísticas descriptivas y visualización;
- 6) **Preparación de datos** – todo lo necesario para limpiar, transformar, combinar datos, entre otros;
- 7) **Modelado** – desde desarrollar modelos predictivos o descriptivos hasta el uso de múltiples algoritmos (suele ser iterativo);
- 8) **Evaluación** – enfocado a entender la calidad del modelo y asegurar que es significativo para el estudio;
- 9) **Despliegue** – en un ambiente productivo;
- 10) **Retroalimentación** – qué tan bien se comportó el modelo (Lin, 2017; Malik, Abdallah y Ala'raj, 2018; Wirth y Hipp, 2000).

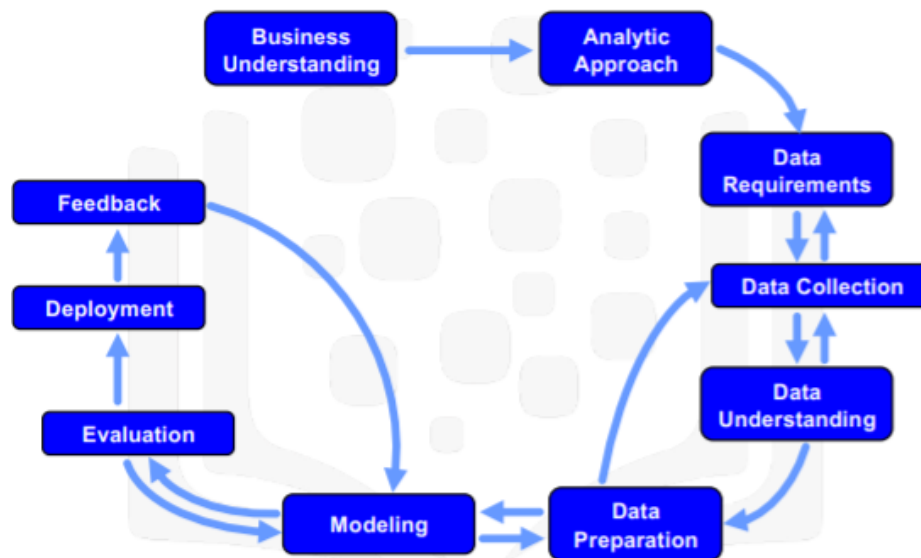


Figura 2.7. Diagrama de metodología CRISP-DM según IBM

2.5. Estudios previos

A continuación, se presentarán estudios similares que se han realizado en distintas instituciones de salud para predicción de readmisión a partir de esta información.

El primer estudio sobre predicción de readmisión se realizó en Estados Unidos utilizando una base de datos nacional de readmisiones, la cual utilizó estadística

descriptiva según características demográficas y características clínicas, índice de readmisión según subcategorías de características, y modelo de regresión logística para el análisis multivariado; mismo que predecía probabilidades de readmisión según su edad y otras variables significativas (sexo, condiciones crónicas, tipo de admisión, entre otras). Uno de los resultados más relevantes fue que los problemas mentales aparecían más frecuentemente y tenían un índice alto de readmisión (Berry et al., 2018).

Otro estudio similar realizado para Nueva Zelanda presenta un análisis de riesgo de readmisión, la cual utilizó información básica del paciente (raza, sexo, edad, tiempo internado, condición), tipo de hospital, entre otros. Para su análisis, los métodos que se utilizaron fueron la regresión logística, random forest, máquinas de vectores de soporte, y aprendizaje profundo. Las predicciones de readmisión tuvieron mejores rendimientos que los métodos revisados como literatura por los autores (Futoma, Morris y Lucas, 2015).

Utilizando ECE se investigó el riesgo de demencia en Korea para adultos mayores de 40 años, obteniendo factores de riesgo como edad; sexo femenino; bajo peso; hipertensión; trastorno psiquiátrico o neurológico comórbido; enfermedad cardiovascular, diabetes; fumador; y ejercicio. Con ello, se comprobó que puede ser utilizado efectivamente para predecir demencia y facilitar su identificación temprana en esa comunidad (Park et al., 2019).

En una revisión de estudios en PubMed del 2009 al 2014 utilizando expedientes clínicos electrónicos, de 107 artículos 12 buscaron predecir mortalidad, 9 readmisión, 9 algún tipo de cáncer, 4 paro cardíaco y traslado a la UCI, y 3 diabetes. Entre ellos, la misma base de datos del sistema de salud de Geisenger se presentó en diferentes publicaciones donde cada uno modeló la probabilidad de insuficiencia cardiaca, readmisión, ataque al corazón y diabetes. Con respecto a los algoritmos utilizados, la regresión lineal fue el algoritmo más utilizado (84), seguido de métodos bayesianos (11), árboles de decisión (10) y regresión regularizada (7). La validación utilizada más

común fue la división de la muestra (67), seguido por validación cruzada (21) y bootstrapping (9) (Goldstein et al., 2017).

Lix et al (2016) examinó los registros, variables demográficas y características de diagnóstico y tratamiento de pacientes con cáncer, donde se revisó la validez predictiva de los índices de comorbilidad de Charlson y Elixhauser.

En un estudio de casi 100 mil empleados mexicanos, se evaluó la utilidad de un índice de comorbilidad para segmentar e identificar áreas de oportunidad en la institución de salud para reducir costos en el servicio, e incluso predecir un costo promedio de los usuarios según su calificación y padecimientos (Garza y Ramírez, 2017).

Otros modelos relacionados son los del Grupo Clínico Ajustado que agrupa individuos según su edad, género y diagnósticos médicos que han sido recolectados en un periodo de tiempo (generalmente un año). El cual sirvió en un estudio para segmentar pacientes en grupos y predecir los costos que se presentarán en el siguiente periodo de tiempo para la organización (Reid et al., 2001).

Se han realizado anteriormente otros estudios con la base de datos proporcionada por ISSSTESON, pero han tenido otros fines y usuarios finales; como, por ejemplo, alertas de focos de epidemia utilizando información geográfica de derechohabientes, y detección de riesgo de obesidad y diabetes para asistir a campañas de medicina preventiva. Este último tiene un objetivo similar, pero el presente es de enfoque más abierto al no delimitarse a estas dos enfermedades (Llanes, 2018; Sanez, 2018).

3. METODOLOGÍA

La metodología se basó en CRISP-DM y se adaptó según las características específicas del proyecto (Lin, 2017; Malik, Abdallah y Ala'raj, 2018; Wirth y Hipp, 2000). Como se mencionó en el capítulo 1.3, el objetivo del proyecto es elaborar una metodología para la segmentación de pacientes de acuerdo al cálculo de índice de comorbilidad en derechohabientes según consultas a médico general, tomando en cuenta características biométricas, frecuencias de visitas a médico general y reincidencia, utilizando técnicas de minería de datos, con el fin de realizar consultas más eficientes asistidas con información histórica, y reducir reincidencias; además de generar grupos de usuarios de acuerdo a las similitudes entre ellos para campañas de medicina preventiva dirigidas. El proyecto es mayormente descriptivo y durante la implementación hubo un algoritmo con resultados predictivos (Hernández, 2014). En la Figura 3.1 se presenta la metodología propuesta (Lin, 2017; Malik, Abdallah y Ala'raj, 2018; Wirth y Hipp, 2000).

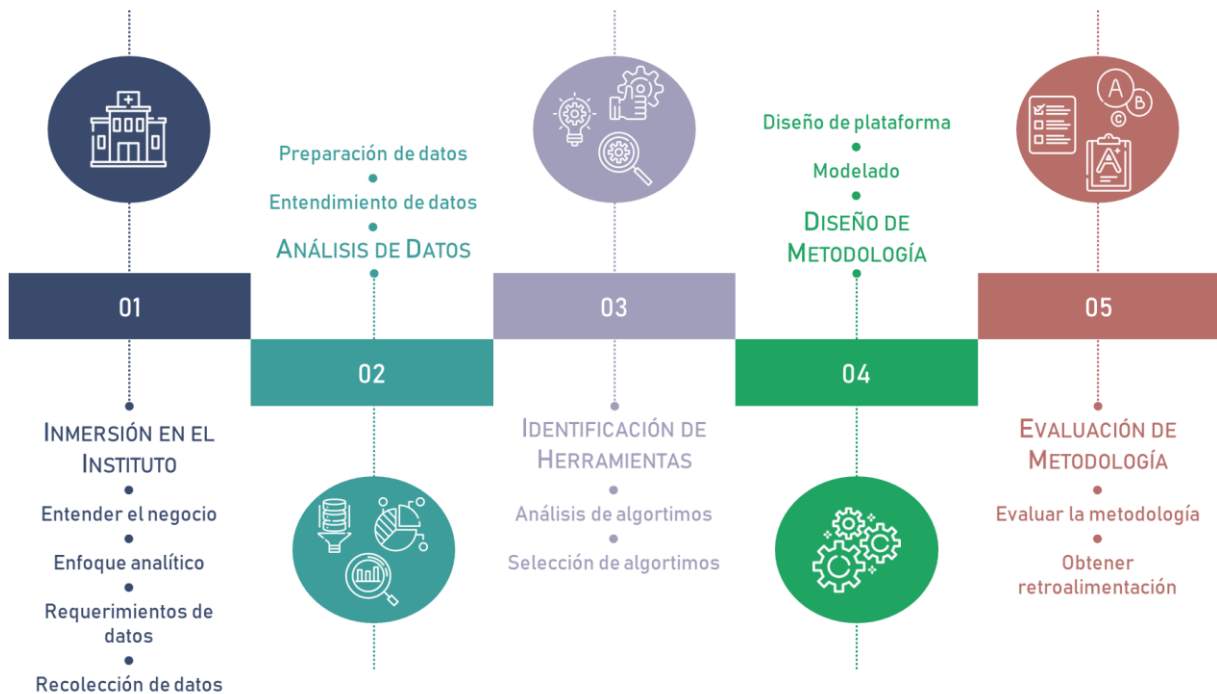


Figura 3.1. Metodología propuesta

Cada paso de cada etapa tiene la capacidad de retroalimentar pasos anteriores según se vaya desarrollando el proyecto; por ejemplo, durante la preparación de datos es posible que se requieran recolectar nuevos datos que ayuden a la transformación y análisis.

3.1. Inmersión en la información y en el instituto

Esta etapa consiste en cuatro pasos: entender el negocio, enfoque analítico, requerimientos de datos y recolección de datos, como se presenta en la Figura 3.2 y se presentan a continuación.



Figura 3.2. Inmersión en la información y en el instituto

Esta etapa comienza con **entender el negocio**, desde la problemática desde la perspectiva de la empresa y el comportamiento de sus usuarios. Esto puede realizarse a partir de entrevistas realizadas a personal de interés es posible obtener el propósito del proyecto (Lin, 2017).

Este paso de identificar el propósito es de gran relevancia para el proyecto ya que, de no detectar correctamente la problemática, es posible que el resto del proyecto sea incapaz de cumplir con las expectativas del cliente. Además, el cliente no siempre estará disponible para resolver dudas más adelante.

Un aspecto muy importante es realizar preguntas específicas, relevantes y sin ambigüedad; muchas veces las mismas preguntas te indican preliminarmente qué tipo de algoritmos se van a utilizar más adelante. Por ejemplo,

- ¿Qué tanto o cuántos? → regresión
- ¿Qué categoría? → clasificación
- ¿Qué grupo? → agrupamiento
- ¿Es esto raro? → detección de anomalías
- ¿Qué opción debería de llevarse a cabo? → recomendaciones (Martens et al., 2017)

En el caso de la problemática presentada, una pregunta de interés podría ser cómo agrupar pacientes de acuerdo con su situación o estado de salud siguiendo algunos criterios tales como el grupo de edad, género, tipos de enfermedades, recurrencia de una misma enfermedad, entre otros.

En caso de requerir métricas de éxito para el proyecto, estos deberán de ser específicos, medibles, alcanzables, relevantes y limitados en el tiempo (Martens et al., 2017).

A partir de ello, es necesario expresar el problema principal en un **enfoque analítico**; pudiendo ser preguntas de regresión al tratar de predecir periodos en el futuro, clasificación de usuarios con respecto a una pregunta definida, agrupamiento de usuarios con respecto a variables no definidas, recomendaciones o personalización con respecto a variables (como qué información mostrar a los usuarios que les sea pertinente), detección de usuarios fuera de la media (usuarios atípicos), entre otros. En la Figura 3.3 se ilustran algunos ejemplos (Lin, 2017).

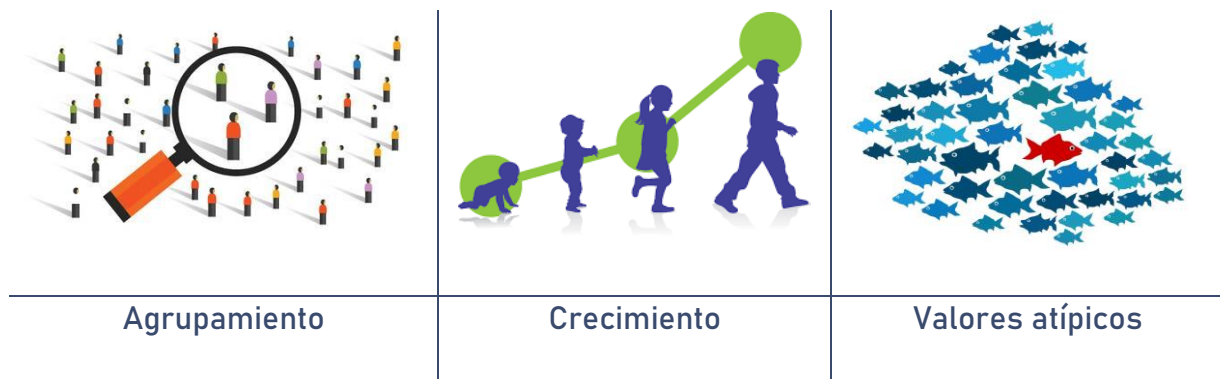


Figura 3.3. Ejemplos de enfoques analíticos

A partir de la expresión del problema, es posible saber **qué datos son requeridos** para resolver el problema. Los requerimientos pueden variar entre periodicidad, formato, definición, antigüedad, entre otros. Finalmente se llega a la **recolección de datos** disponibles según los requerimientos definidos en el punto anterior (Lin, 2017).

La recolección se puede llevar a cabo al tener acceso a las bases de datos propios de la organización, pero también pueden ser recolectados por medio de entrevistas, observaciones, encuestas, entre otros, según sean los requerimientos de datos y la factibilidad tanto en tiempo como accesibilidad de realizarse en el entorno deseado (Hernández, 2014; University of Leicester, 2019).

3.2. Análisis de los datos

Esta etapa consiste en dos pasos, los cuales son entendimiento de datos y preparación de datos, como se presentan en la Figura 3.4 y se desarrolla a continuación.



Figura 3.4. Análisis de los datos

Durante esta etapa, los datos recolectados pasan a ser entendidos por el analista. Este análisis es el inicial sobre los datos que permite conocer el comportamiento de los usuarios (lo cual puede ser un comportamiento esperado a partir de las entrevistas realizadas en el paso anterior).

Para **entender los datos** es necesario realizar en mancuerna tanto estadística descriptiva como visualización de los datos. La estadística descriptiva puede dar a entender comportamiento de datos que no siempre corresponde a toda la verdad del comportamiento. Por ejemplo, en la Figura 3.5 se presentan cuatro conjuntos de datos con los mismos valores de media, varianza, correlación entre ambos ejes y regresión lineal; pero con ayuda de la visualización es posible identificar cuál es el comportamiento real de los datos (Lin, 2017; MathWarehouse, 2019).

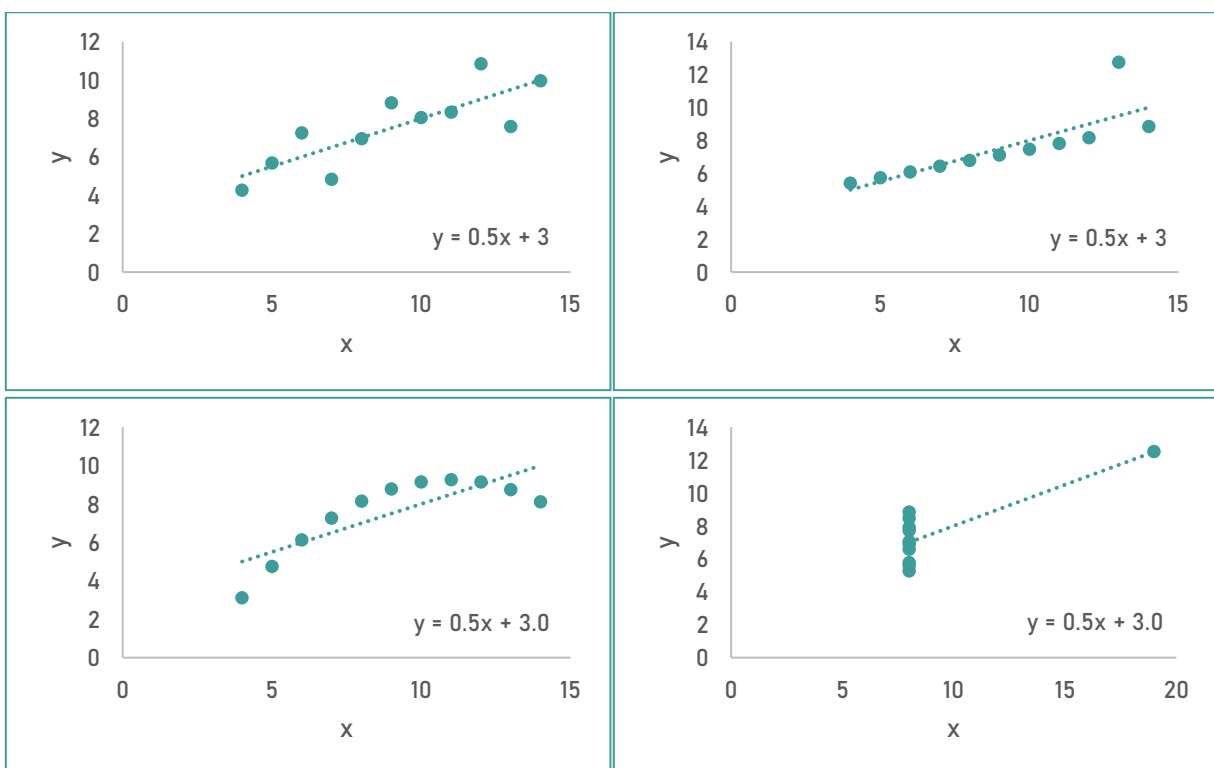


Figura 3.5. Cuarteto de Ascombe

También, se pueden identificar aquellas características de los datos recolectados que sean de pertinencia para el problema. En la Tabla 3.1 se presentan algunas

características de cada conjunto de datos relevante que pueden utilizarse para este paso, según sea aplicable (Patience, 2018).

Columnas	Nombres de columnas
Tamaño del conjunto de datos	Por ejemplo, (1000, 10)
Formato de cada columna	Cuáles de esos son numéricos, caracteres, entre otros.
Conteo	Conteo de resultados posibles por columna
Estadística descriptiva de cada columna	Máximo, mínimo, media, desviación estándar, entre otros.
Valores atípicos	Revisar valores en columnas numéricas que estén muy alejados del resto (por ejemplo, $3 \times \text{rango intercuartil}$)
Extracto de conjunto de datos	Pueden ser las primeras filas; censuradas o con datos modificados si es requerido.

Tabla 3.1. Ejemplo de características de datos por cada conjunto de datos relevante

En la Figura 3.6 se presentan algunos tipos de visualización de datos que pueden utilizarse para revisar cómo se comportan los conjuntos de datos analizados (Grosser, 2019).



Figura 3.6. Tipos de visualización de datos

Los conjuntos de datos más relevantes para esta investigación son los proporcionados por el personal de estadística de ISSSTESON, mismos que presentan registros del sistema de control desde enero de 2013 a marzo de 2017 conteniendo ID de paciente (mismo ID que ha sido modificado para asegurar el anonimato de cada derechohabiente), grupo de edad, sexo, diagnóstico, entre otras características recolectadas durante la visita a médico general.

Seguido de esto, se **preparan los datos** para su análisis. Este paso conlleva a lo siguiente:

- Seleccionar datos
- Limpiar datos
 - Valores faltantes o inválidos
 - Eliminar filas duplicadas
 - Aplicar el formato adecuado
- Construir datos
 - Atributos derivados (cuáles columnas fueron creadas, cómo se crearon y por qué)
 - Registros generados (cuáles filas fueron creadas, cómo se crearon y por qué)
- Combinar múltiples fuentes de datos
- Transformar los datos (Brown, 2019a; Lin, 2017)

Al final, se espera presentar una base de datos preparada para el modelado y un reporte describiendo el conjunto de datos (Brown, 2019a). En la Figura 7.2 se presenta una plataforma para explorar, visualizar y analizar conjuntos de datos antes de modelar (Guo, 2017).

3.3. Identificación de herramientas

Esta etapa consiste en dos pasos: análisis de algoritmos y selección de algoritmos. Éstos se presentan en la Figura 3.7 y se desarrolla a continuación.



Figura 3.7. Identificación de herramientas

A partir del análisis de las herramientas posibles durante la revisión bibliográfica, se realiza una identificación sistemática de las herramientas que puedan utilizarse para analizar la información disponible para evaluar el riesgo de salud de derechohabientes.

Para ello, se **analizan algoritmos** que sean de utilidad a partir de la problemática planteada y los datos disponibles que fueron preparados en el paso anterior. Este paso es iterativo y pueden ser algoritmos tanto predictivos como descriptivos. Algunos ejemplos que se vieron en el Capítulo 2.2.3 del marco de referencia fueron análisis de conglomerados, regresión y árboles de decisión (Lin, 2017).

A pesar de que desde el paso de enfoque analítico se tiene una idea de qué tipo de algoritmos se estarán utilizando para resolver la problemática, cada tipo de algoritmo presenta diferentes configuraciones dependiendo del tipo de caso que se está utilizando o incluso de los autores. En el caso de una regresión, en programas computacionales estándares como Excel se presentan regresiones exponenciales, lineales, logarítmicas, polinomiales (de varios órdenes), poder, media móvil (con diferentes periodos). Las opciones aumentan entre más complejo sea el comportamiento de las variables. En la Figura 3.8 se presenta una comparación de algunos algoritmos de análisis de conglomerado según el tipo de datos que es analizado.

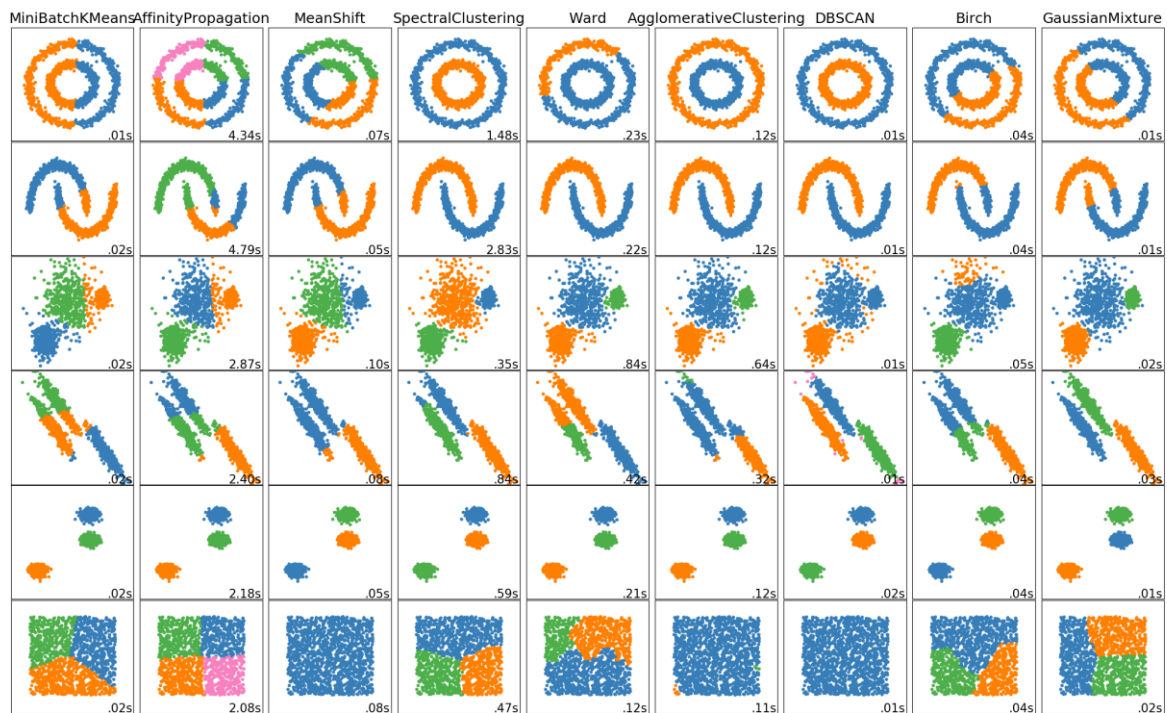


Figura 3.8. Comparación de algunos algoritmos de análisis de conglomerados (Seif, 2019)

A partir de los datos disponibles se **seleccionan los algoritmos** que se vayan a modelar (Lin, 2017). Para este paso se requiere reportar las técnicas de modelado que se estarán utilizando y los supuestos que se vayan a adoptar (de darse el caso) (Brown, 2019b). Uno de los principales algoritmos a evaluar será K-Means, que agrupa las muestras de acuerdo con su similitud o cercanía. Se encontrará la K óptima por medio de distintas técnicas hasta encontrar el mejor desempeño.

3.4. Diseño de metodología

Esta etapa consiste en dos pasos: modelado y diseño de plataforma (véase en la Figura 3.9).



Figura 3.9. Diseño de metodología

Antes de iniciar el modelado, se debe definir qué tipo de estudio se va a realizar para determinar qué tan bien se desempeña el modelo. En la sección 2.2.3 se presentan distintas formas de validación de los datos. Durante el **modelado** se debe registrar la configuración de parámetros; la descripción del modelo con respecto a las variables, interpretación y dificultades presentadas; y los modelos. Los modelos también deberán ser evaluados (entre ellos también en el caso de ser más de uno) según su uso o impacto de la problemática (Brown, 2019b). En la Tabla 2.1 se presenta un ejemplo de qué es lo que deberá llevar cada modelo que se esté utilizando en esta etapa (Patience, 2018).

Nombre	Nombre del modelo utilizado
Parámetros	Lista de parámetros y los valores utilizados justificando la elección de estos parámetros
Variables	Las variables que se están utilizando de entrada al modelo
Interpretación de resultados	Describir los resultados que presenta el modelo
Dificultades presentadas	Notas sobre utilizar este modelo y lo que haya acontecido durante la construcción de este
Modelo	El modelo producido por la herramienta utilizada

Tabla 3.2. Ejemplo de reporte de modelo para presentar

Debido a que el proyecto en sí va dedicado a usuarios, es necesario **diseñar la plataforma** con la que el usuario pueda interactuar con los modelos y ser capaz de utilizarla para una toma de decisiones oportuna. Este puede partir de qué tipo de datos le será de utilidad; desde una calificación, una tabla, una gráfica (y qué tipo de gráfica se apreciará mejor la información), entre otros. Con ello, la plataforma debe diseñarse

tanto para ser útil visualmente, la información más importante sea desplegada y permita una interacción intuitiva. Los atributos que tomarán parte de la representación de las muestras serán la edad, sexo, diagnóstico de enfermedades, entre otros.

3.5. Evaluación de la metodología

Esta última etapa consiste en dos pasos, los cuales son evaluar la metodología y obtener retroalimentación (véase Figura 3.10).

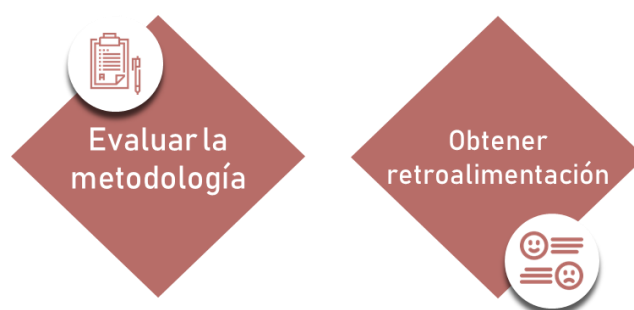


Figura 3.10. Evaluación de la metodología

En esta última etapa se evalúan los modelos según su uso en el contexto del problema y se probará el modelo en una aplicación práctica para determinar qué tan satisfactorio es el uso por el usuario. También se deberá evaluar el proceso según preocupaciones, pasos que necesitan mayor revisión, entre otros, para que de una forma continua se vaya mejorando la plataforma (Brown, 2019c). En esta etapa se presentarán resultados de la metodología y resultados preliminares al grupo de trabajo, incluyendo médicos y personal a cargo del proyecto por parte de la institución, a fin de recibir retroalimentación oportuna para realizar los ajustes correspondientes y asegurarse de seguir alineados a los intereses del usuario final.

A partir de la retroalimentación del usuario final, la plataforma y el modelo podrán adaptarse mejor al proceso y ser cada vez más intuitivo y eficiente con su interacción.

Finalmente, se deberán determinar algunas acciones alternativas que se pudieron haber realizado y la razón de que se haya decidido por lo realizado. A su vez, se deberá

de considerar acciones futuras o posibles mejoras para proyectos futuros de continuidad (Brown, 2019c).

4. IMPLEMENTACIÓN

La aplicación de la metodología presentada en el capítulo anterior se presenta a continuación; así como la retroalimentación obtenida en algunos pasos que afectó el proceso y aspectos que alinearon el análisis.

4.1. Inmersión en el instituto

Esta primera etapa de inmersión del instituto comenzó desde el periodo de enero a marzo de 2019 con el entendimiento del negocio y enfoque analítico para realizar el anteproyecto. El requerimiento de datos se realizó en los meses posteriores y la recolección concluyó en el semestre agosto – diciembre 2019, dando por terminada esta etapa.

4.1.1 Entender el negocio

El primer paso de entender el negocio se realizó por medio de visitas a personal vinculado para registrar qué es lo que les interesaba obtener y cuál es su problemática. Las entrevistas fueron abiertas y guiadas según las respuestas del entrevistado. Con ello, se identificaron algunas preguntas de interés para la investigación.

- ¿Cuáles son las causas de los diagnósticos de derechohabientes de ISSSTESON de acuerdo con historial clínico?
- ¿Cuáles son las técnicas o metodologías de procesamiento y análisis de datos que pudiesen asistir en el análisis de datos de consultas recabadas?
- ¿Qué enfermedades o características presentan un riesgo mayor de reincidencia en consultas posteriores menores a 30 días?
- ¿Qué manera sería la más sencilla para comunicar la información analizada con médicos generales para que sea accesible y fácilmente adoptable?

Además de estas preguntas, se encontraron dos grupos de derechohabientes de interés para la institución; siendo estos pacientes con trastornos de mama y pacientes

con diabetes. Por lo que el análisis se estará enfocando principalmente en ellos ya que de éstos se obtuvieron los datos de visitas principalmente.

4.1.2 Enfoque analítico

Este paso requirió expresar el problema principal desde un enfoque analítico, donde se obtuvo la problemática principal traducido a objetivo general e hipótesis.

Objetivo general: Elaborar una metodología para la segmentación de pacientes de acuerdo al cálculo de índice de comorbilidad en derechohabientes del ISSSTESON según consultas a médico general, tomando en cuenta características biométricas, frecuencias de visitas a médico general y reincidencia, utilizando técnicas de minería de datos, con el fin de realizar consultas más eficientes asistidas con información histórica, y reducir reincidencias; además de generar grupos de usuarios de acuerdo a las similitudes entre ellos para campañas de medicina preventiva dirigidas.

Éste habla de segmentación, por lo que se identifica la necesidad de identificar herramientas de agrupación como k-means.

Hipótesis: La elaboración de un indicador de nivel de salud de los derechohabientes de ISSSTESON podría ayudar a identificar usuarios que requieren mayor atención durante su consulta y asistir en el diagnóstico para prevenir reincidencias, así como crear grupos de pacientes para la mejora en campañas de medicina preventiva dirigida.

A su vez, habla de indicadores de nivel de salud por lo que se deberá de investigar indicadores existentes en la aplicación para segmentar a pacientes, como el índice de comorbilidad.

4.1.3 Requerimientos de datos

Como requerimiento principal fue la recolección de visitas a médico general. Para la codificación y cálculo de índices de comorbilidad se necesitó recolectar una base de datos de código de enfermedades.

4.1.4 Recolección de datos

Por parte de la institución se recibieron varias bases de datos:

- Visitas a médico general de pacientes sin diagnósticos:
 - 1 de enero – 1 de marzo de 2017
- Diagnósticos de pacientes de padecimientos relacionados a trastornos de mama según visitas a médico general:
 - 1 de enero de 2013 – 15 de enero de 2018
- Diagnósticos de pacientes de padecimientos relacionados con diabetes según visitas a médico general:
 - 2 de enero de 2013 – 31 de mayo de 2017
- Visitas a médico general de pacientes diagnosticados con trastornos de mama y/o diabetes en los siguientes periodos con diagnóstico:
 - 2 de enero – 7 de abril de 2015
 - 4 de enero – 5 de abril de 2016
 - 2 de enero – 1 de marzo de 2017

Los números de identificación de paciente fueron modificados para asegurar la anonimidad de los pacientes.

Para asistir al análisis,

- Se obtuvo la base de datos de código de enfermedades a 10 dígitos (ICD-10) creada por la Organización Mundial de la Salud (WHO) en español.

4.2. Análisis de datos

En esta etapa se realiza un análisis descriptivo de los datos y se preparan para el modelado. Se estarán presentando los resultados más relevantes para el estudio obtenidos de este análisis. A partir de esta etapa, se utilizó el lenguaje R en el entorno RStudio y sus distintas librerías (R Core Team, 2019; RStudio Team, 2019).

4.2.1 Entendimiento de datos

Para entender los datos se presentó un análisis descriptivo de los registros. Para ello, se analizó el total de visitas y derechohabientes, el género de los derechohabientes según la base de datos analizada, y el número de visitas según el año y base de datos analizada. El análisis descriptivo y gráficas fueron realizados principalmente en RStudio (R Core Team, 2019; RStudio Team, 2019) utilizando paqueterías disponibles y Microsoft Excel. La plataforma RStudio se utilizó en este paso debido a que es una plataforma abierta (cualquier institución médica lo puede descargar sin comprar, lo cual facilita su adopción) y está enfocado al estudio estadístico.

Cada base de datos consultada tenía las siguientes variables para cada visita (fila):

- hashAfilia (numérico) – identificador numérico del derechohabiente modificado para mantener su anonimato
- Fecha – presenta el año, mes y día de la consulta
- Grupo de edad (caracteres) – presenta rango de edad del derechohabiente (ver en la Figura 4.8, Figura 4.9 y Figura 4.10)
- Interconsulta (caracteres) – campo de medicina a la que se remite al derechohabiente durante la visita a médico general (ver Figura 4.11)
- Rayos X (binario) – si se presenta estudio de rayos x
- Receta (binario) – si se prescribe receta médica (ver Figura 4.12)
- Sexo (caracteres) del derechohabiente (ver en la Figura 4.8, Figura 4.9 y Figura 4.10)
- Cantidad de medicamentos (numérico) prescritos (ver Figura 4.14)
- Cantidad de recetas (numérico) generadas (ver Figura 4.13)
- Electrocardiograma (binario) - si se realiza electrocardiograma
- Examen de Mama (binario) – si se realiza examen de mama
- Diagnostico (caracteres) – diagnóstico descriptivo del paciente (ver Figura 4.17 y Figura 4.19)

- Sub Gpodia (caracteres) – capítulo del diagnóstico (ver Figura 4.16 y Figura 4.18)

El número de derechohabientes obtenido en cada base de datos registrada y número de visitas se presenta a continuación (se presentaron 44 derechohabientes con ambos diagnósticos). En el caso de diabetes, se presentó mayor número de visitas para un periodo de tiempo similar.

	Trastornos de mama	Diabetes	Ambos	Totales
Derechohabientes	1,074	3,502	44	4,620
Visitas diagnóstico	1,375	16,191	272	17,838
Visitas generales	3,173	7,925	381	11,479
Totales	4,548	24,116	653	29,317

Tabla 4.1. Registros de visitas totales de derechohabientes

Para identificar si una visita era clasificada como visita diagnóstico o general, se identificaron aquellos diagnósticos relacionados a trastornos de mama o diabetes para segmentarlos como diagnóstico. Se consideraron las visitas diagnóstico para trastornos de mama los siguientes códigos internacionales de enfermedades (Tabla 4.2).

ICD	Descripción
C50	Tumor maligno de la mama
D05	Carcinoma in situ de la mama
D24	Tumor benigno de la mama
D48.6	Tumor de comportamiento incierto o desconocido de otros sitios y de los no especificados: de la mama
N60	Displasia mamaria benigna
N62	Hipertrofia de la mama
N63	Masa no especificada en la mama
N64	Otros trastornos de la mama
R92	Hallazgos anormales en diagnóstico por imagen de la mama
Z12.3	Examen de pesquisa especial para tumores: de la mama
Z80.3	Historia familiar de tumor maligno: de mama

Tabla 4.2. Diagnósticos para trastornos de mama

Mientras tanto, se consideraron las visitas diagnóstico para diabetes los siguientes códigos (Tabla 4.3).

ICD	Descripción
E10	Diabetes mellitus insulino dependiente
E11	Diabetes mellitus no insulino dependiente
E12	Diabetes mellitus asociada con desnutrición
E13	Otras diabetes mellitus especificadas
E14	Diabetes mellitus, no especificada
E23.2	Hipofunción y otros trastornos de la glándula hipófisis: Diabetes insípida
P70.2	Trastornos transitorios del metabolismo de los carbohidratos específicos del feto y del recién nacido: Diabetes mellitus neonatal

Tabla 4.3. Diagnósticos para diabetes

Si bien ya se conoce el número de visitas de cada grupo de derechohabientes, es necesario registrar que cada periodo del año que se registraron visitas a médico general presentó diferentes días. Estos periodos se encuentran en la Tabla 4.4. Para el último año, sólo se presenta aproximadamente el 62% de los periodos anteriores, por lo que se espera una disminución relevante en visitas generales.

Año	Inicio	Fin	Días
2015	2 enero	7 abril	96
2016	4 enero	5 abril	93
2017	2 enero	1 marzo	59

Tabla 4.4. Periodos de visitas generales a médico general

El número de visitas registrada por cada periodo y según cada grupo de derechohabientes se presenta en la Figura 4.1. Como se vio en la Tabla 4.4, las observaciones abarcan los años del 2015 al 2017 (visitas generales). Las visitas diagnóstico (las que contemplan un diagnóstico incluido en la tabla Tabla 4.2 o Tabla 4.3 por parte del médico) abarcaron desde el 2013 hasta el 31 de mayo de 2017 para diabetes y hasta el 15 de enero de 2018 para trastornos de mama. Se puede apreciar una caída en visitas generales en el 2017 debido a la disminución de días de este periodo.

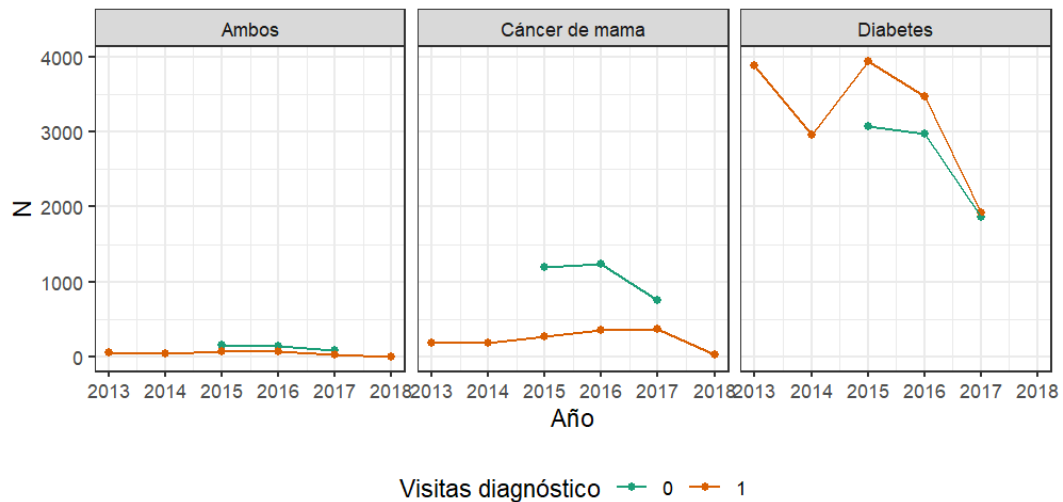


Figura 4.1. Visitas diagnóstico y generales según grupo de interés

Ahora, analizando el comportamiento de los derechohabientes por cada una de las visitas registradas se presenta a qué tantos días después de cada visita vuelven a asistir. Para el siguiente diagrama se calcularon sólo en los derechohabientes que hubieran tenido más de una visita por cada periodo y no se calculó el tiempo entre visitas de diferentes periodos. En la Figura 4.2, se presentan los derechohabientes con trastornos de mama. Las visitas son mayormente frecuentes en menos de 25 días, pero aparentemente va disminuyendo la probabilidad con una tendencia marcada según aumentan los días. Se puede apreciar en el 2017 una disminución en número de visitas debido a que se registró un mes menos que los otros dos periodos, pero se mantiene la tendencia independiente del año.

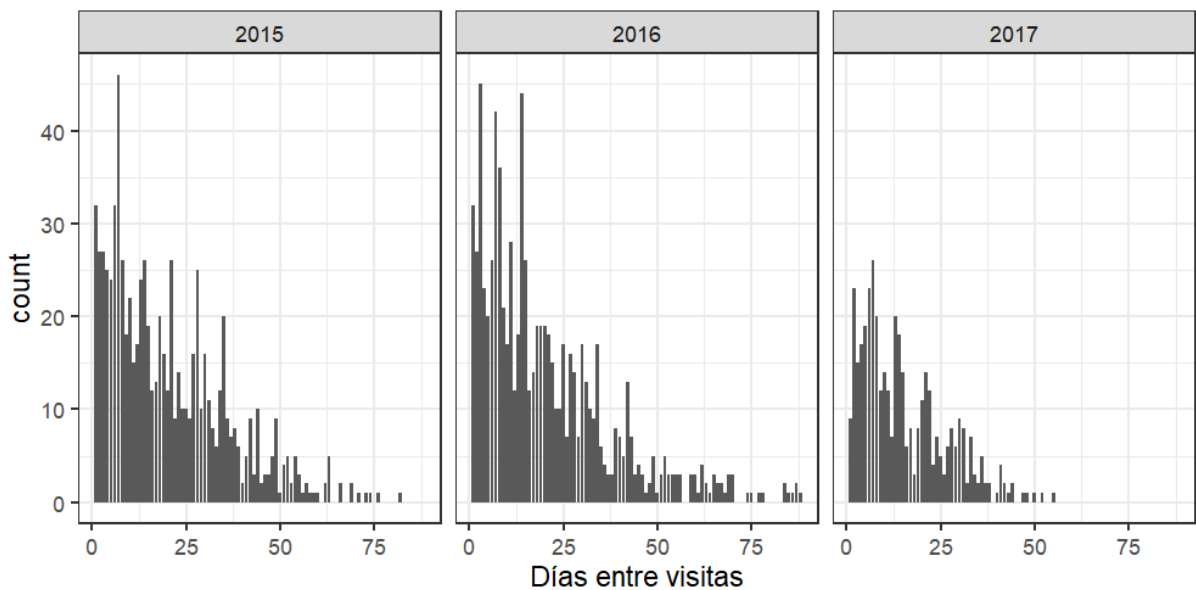


Figura 4.2. *Días entre visitas generales de pacientes con diagnóstico de trastorno de mama*

Con respecto a los derechohabientes con diabetes, se detectaron dos sesgos, uno a los pocos días que va disminuyendo y otro que hay un sesgo muy significativo al mes. Esto podría deberse a las visitas mensuales de seguimiento que podrían estarse llevando a cabo.

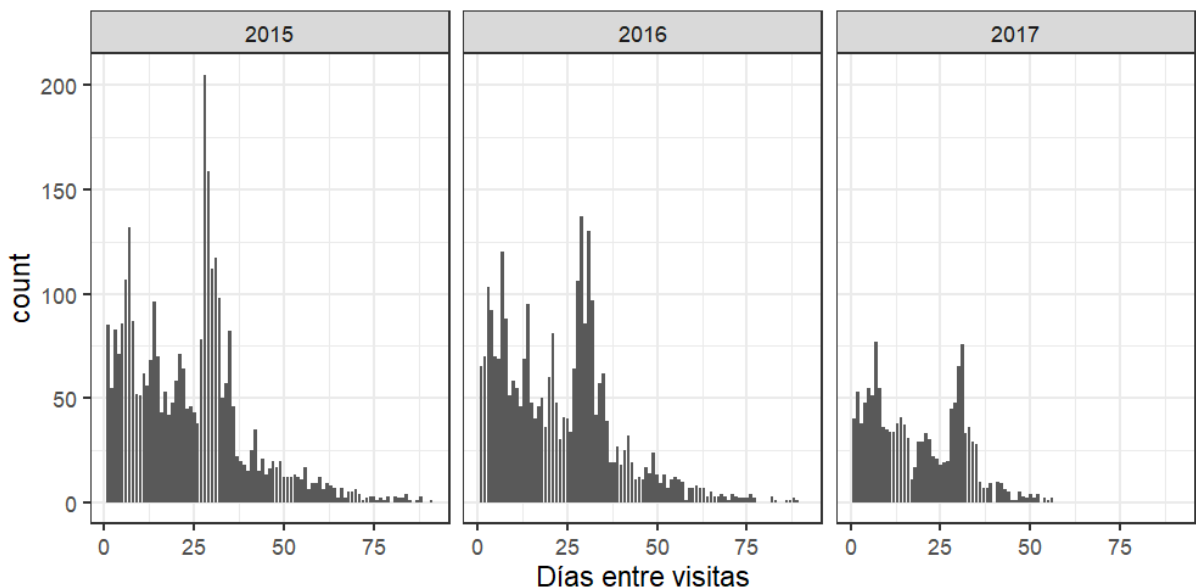


Figura 4.3. *Días entre visitas generales de pacientes con diagnóstico de diabetes*

Otra manera de ver las visitas de los derechohabientes es conociendo cuántas visitas realizó un solo derechohabiente por cada periodo. En cada columna de la Figura 4.4 se presenta un año y en cada fila un grupo de derechohabientes. Se puede apreciar que cada año diferente número de derechohabientes fueron registrados (los cuales se presentan al centro de cada gráfica con número). Se identifican derechohabientes con hasta 15 visitas registradas en los periodos de 90 días con una tendencia a la baja según el número de visitas, siendo mayormente entre una y tres visitas por periodo de los derechohabientes con diabetes, mientras que una a dos visitas en derechohabientes con trastorno de mama.

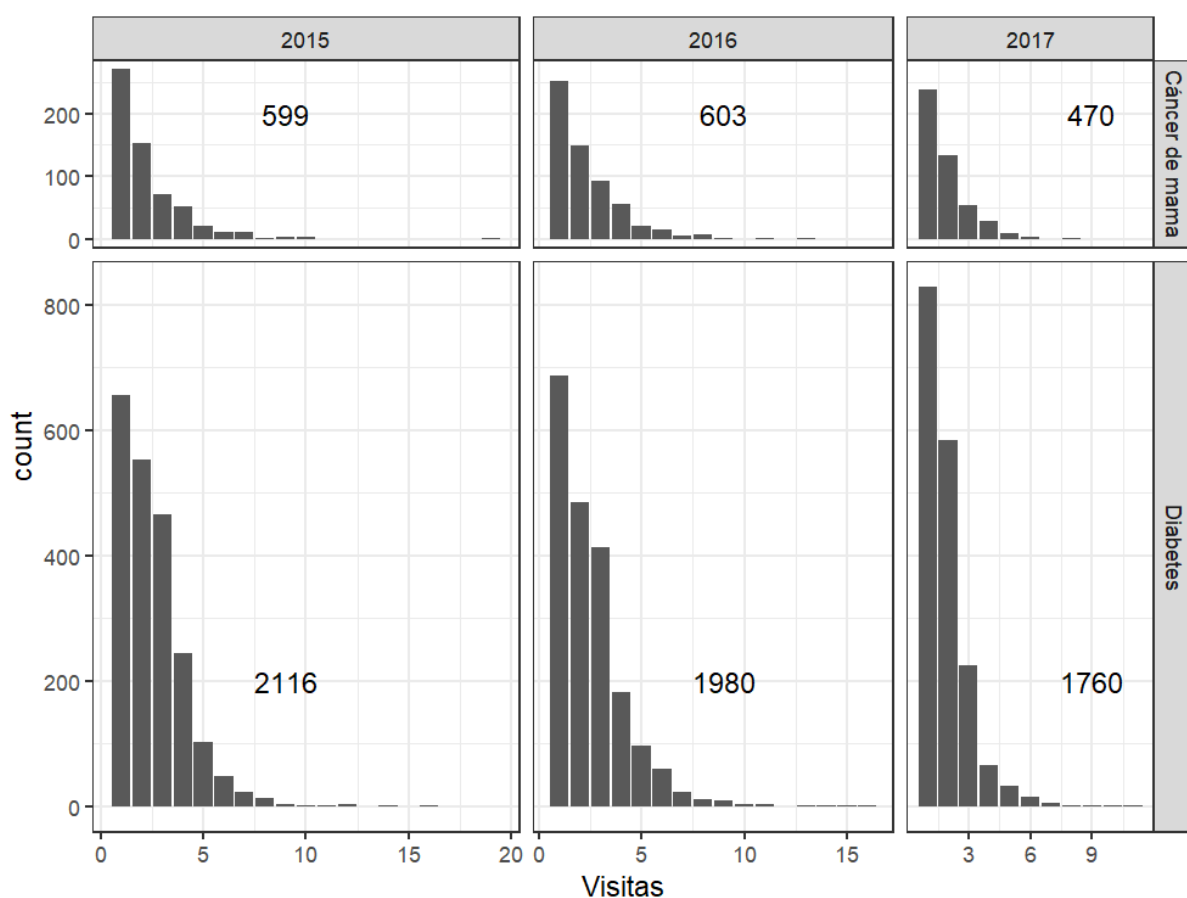


Figura 4.4. *Visitas por periodo por derechohabiente y número de derechohabientes registrados por periodo (derechohabientes con diabetes = 3,502, con trastorno de mama = 1,074)*

Para identificar qué derechohabientes presentaban un número de visitas elevado dentro de un periodo de tiempo (se seleccionaron 30 días), se comparó cada visita

registrada con la sumatoria de las visitas en el último mes del derechohabiente. Posteriormente, se calcularon estadísticas descriptivas de visitas acumuladas en el mes por cada derechohabiente. En la Tabla 4.5, se presentan los principales derechohabientes con mayor número de visitas acumuladas en el mes.

hashAfilia	mínimo	cuartil 25%	median	cuartil 75%	máximo	media	Desv. Estándar
2939276	1	3	5	6	11	4.94	2.51
7428174	1	2	3	6	11	4.33	3.11
2156874	1	1	2	7.5	10	4.1	3.44
5304081	1	2	3.5	5.75	9	4.07	2.56
3545774	1	1.5	2	4	8	2.97	1.96
7386774	1	1	3	5	8	3.35	2.32

Tabla 4.5. Estadísticas descriptivas de derechohabientes con mayores visitas acumuladas en el mes

Los principales tres derechohabientes con mayor número de visitas en los tres periodos se pueden visualizar en la Figura 4.5. Debido a que las visitas a médico general sólo se registraron durante el mes de enero a abril durante el 2015 y 2016, es donde se están presentando los mayores acumulados. Si bien en el derechohabiente 2156874 se registra un acumulado en el 2015, este comportamiento no se presentó durante el 2016 y 2017; mientras tanto en los otros dos derechohabientes, su comportamiento fue similar en dos de los tres años; acumulando una gran cantidad de visitas en poco tiempo. De esta manera, se podría identificar qué comportamiento podría considerarse normal para el derechohabiente y qué es atípico. Eso permite dar atención especializada a los pacientes según su comportamiento.

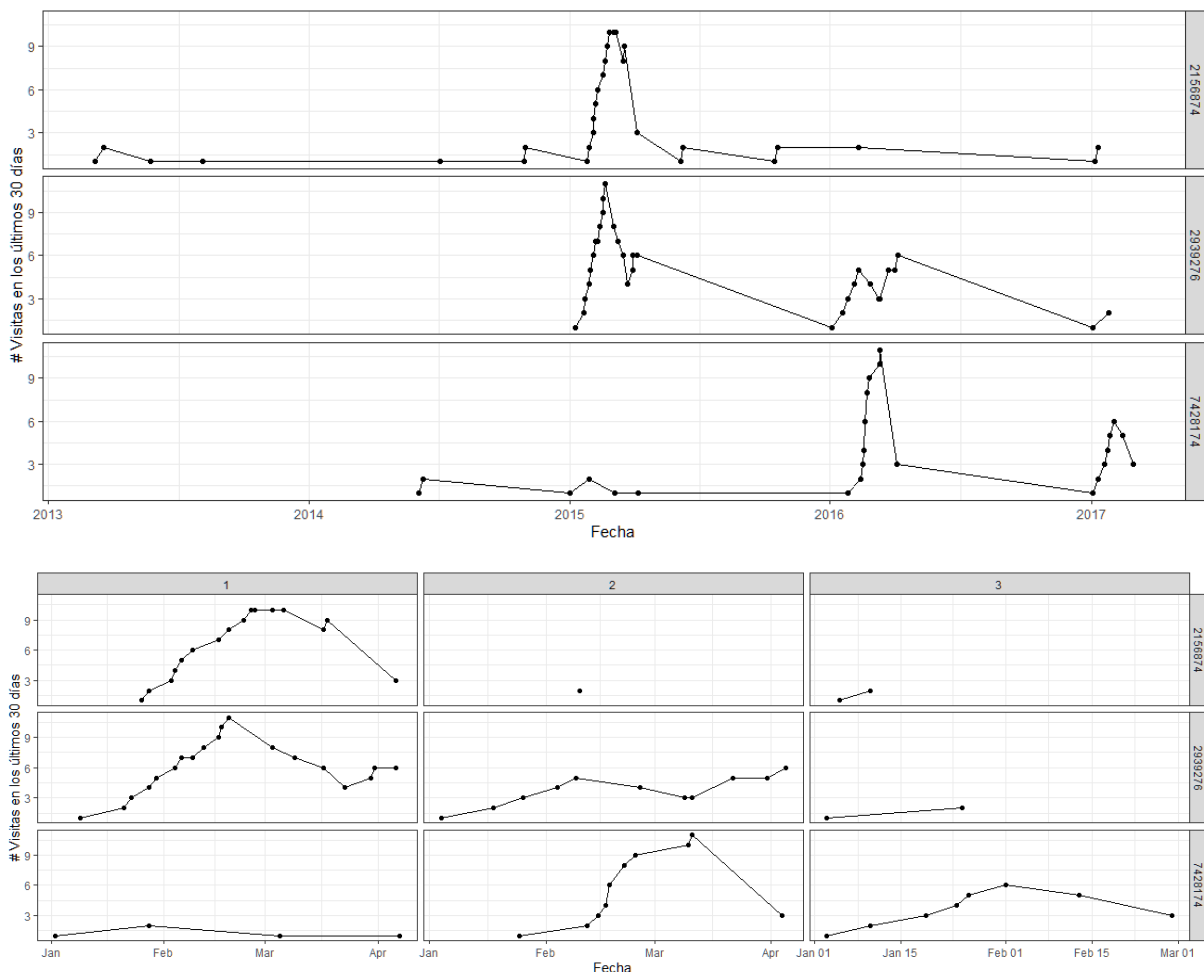


Figura 4.5. Visitas de derechohabientes con mayores visitas acumuladas en el mes (cada punto es una visita) (a) visitas diagnóstico y generales, (b) visitas en periodos de visitas generales

En la Figura 4.6 se presentan los comportamientos de tres derechohabientes con el mismo número de visitas totales (tanto de diagnóstico como generales), pero comportamientos distintos. En el derechohabiente 7097074 se identifica en el segundo periodo un incremento de visitas generales, mientras que los otros dos no presentan incremento de visitas, pero sí un número de visitas alto de visitas diagnóstico durante los tres años. Con esto, de los tres derechohabientes, el tercero presentaría una alerta con respecto a las visitas ya que presentó un incremento anormal.

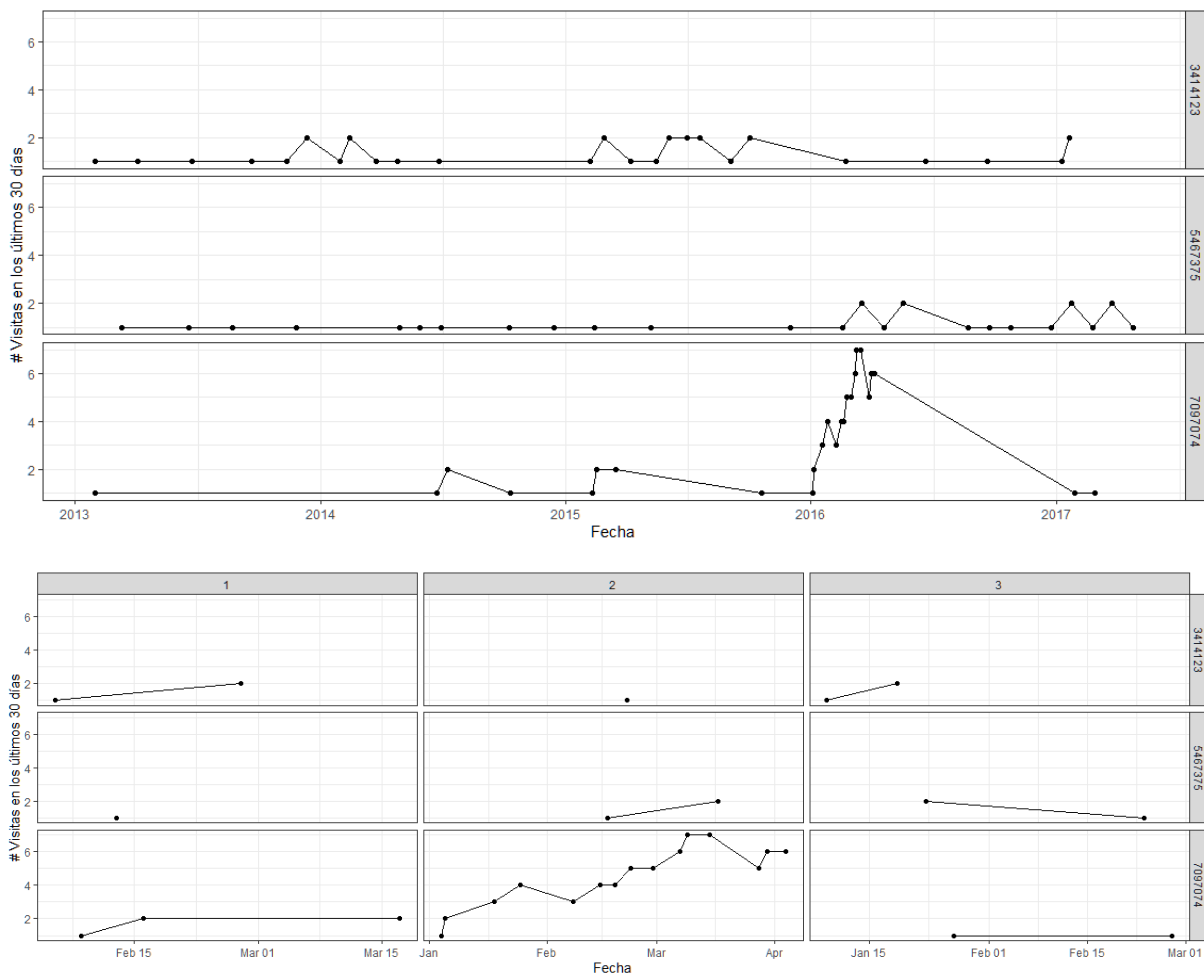


Figura 4.6. Derechohabientes con número alto de visitas registradas pero diferentes comportamientos (a) visitas diagnóstico y generales, (b) visitas en periodos de visitas generales

Con respecto a qué padecimientos presentaban una visita siguiente menor a 30 días, se identificaron aquellas que tuvieran una mediana de 15 días (Figura 4.7). De los padecimientos presentaron se identificaron 16 padecimientos clasificados según IDC-10 (código internacional de enfermedades). Los tres que presentaron menor mediana fueron dorsalgia (cervicalgia [M54.2], lumbago con ciática [M54.4] y lumbago no especificado [M54.5]), seguido de este está tos (R05) y examen de laboratorio (Z01.7). El que tiene más frecuencia se presenta en rojo, siendo obesidad (E66).

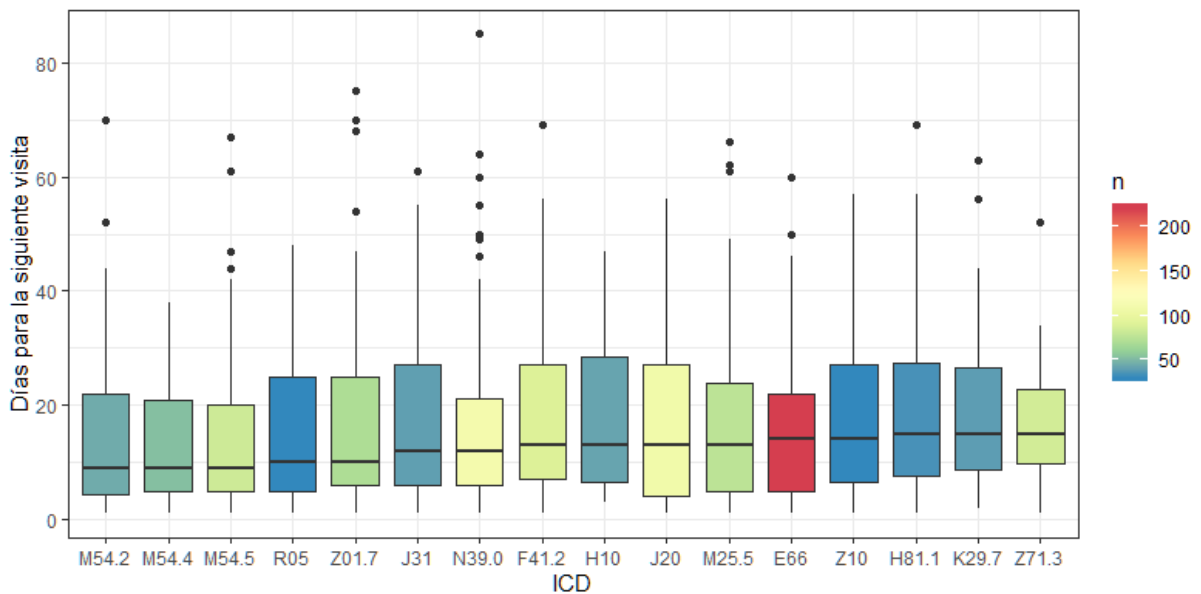


Figura 4.7. Visitas con mediana de reincidencia menor a 15 días

Para conocer más sobre los derechohabientes que presenta este análisis, se presenta a continuación el género y edad de cada grupo de derechohabientes. El grupo de edad se obtuvo de la última visita registrada de cada derechohabiente. El primer grupo de derechohabientes que se evalúa en la Figura 4.8 es el que presenta tanto el diagnóstico de trastorno de mama como el de diabetes. Debido al trastorno de mama, el género predominante es el femenino, y se presenta principalmente en pacientes de más de 40 años.

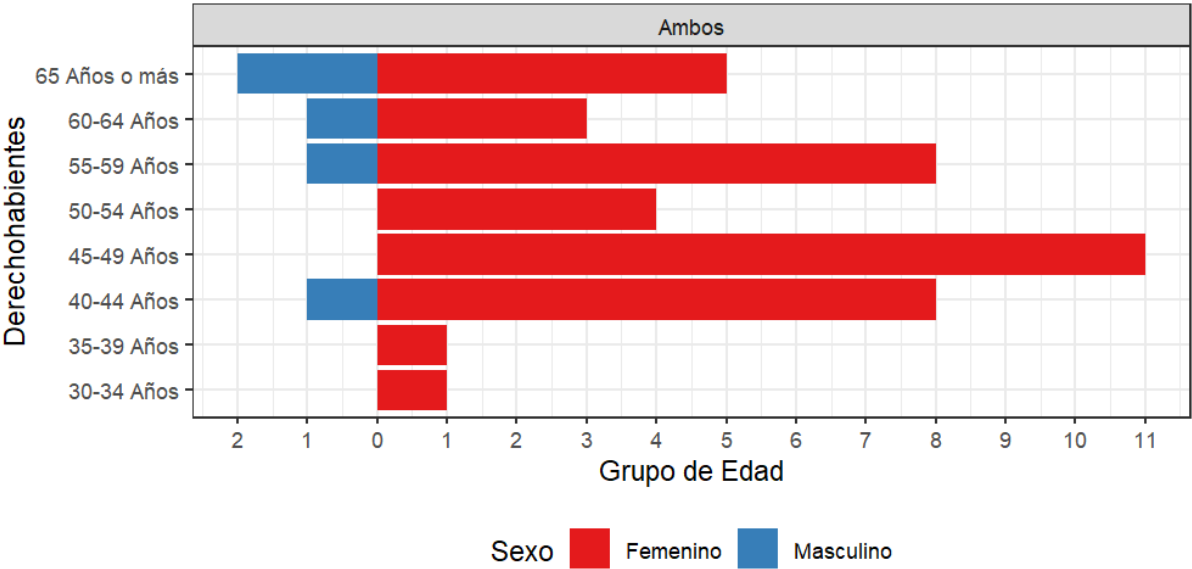


Figura 4.8. Edades y géneros de pacientes con diabetes y trastorno de mama

Siguiendo con el grupo de derechohabientes con trastorno de mama (Figura 4.9), se presentan mayormente derechohabientes mujeres entre 30 y 59 años.

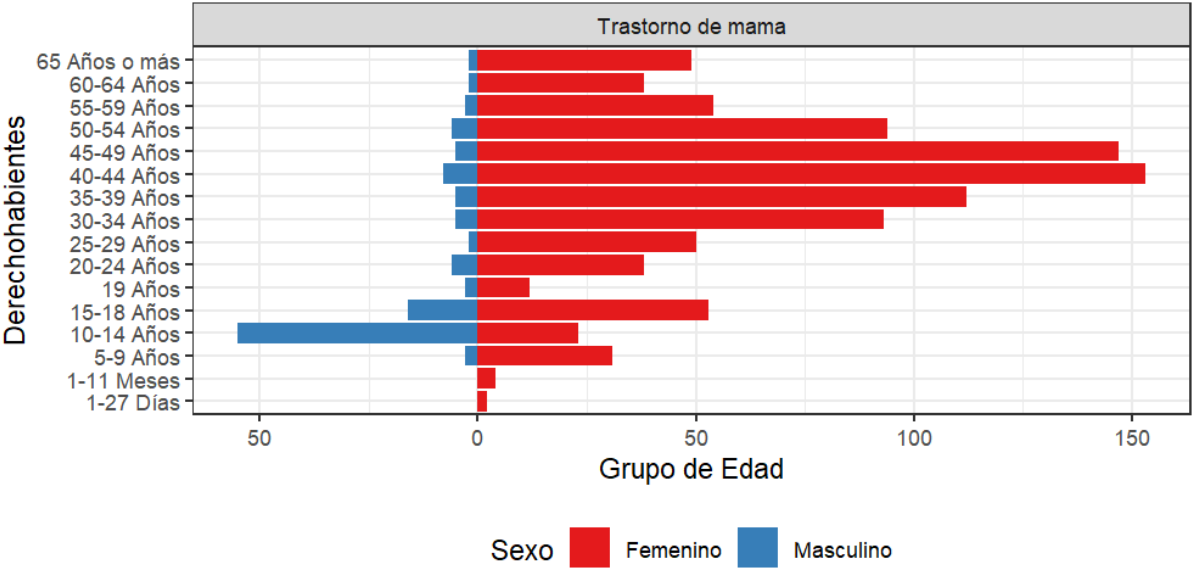


Figura 4.9. Edades y géneros de pacientes con trastorno de mama

Finalmente, los derechohabientes con diabetes (Figura 4.10) presentan ambos géneros de manera casi equivalente relación de grupo de edad muy similar entre ambos. Se presenta mayormente en derechohabientes con más de 65 años.

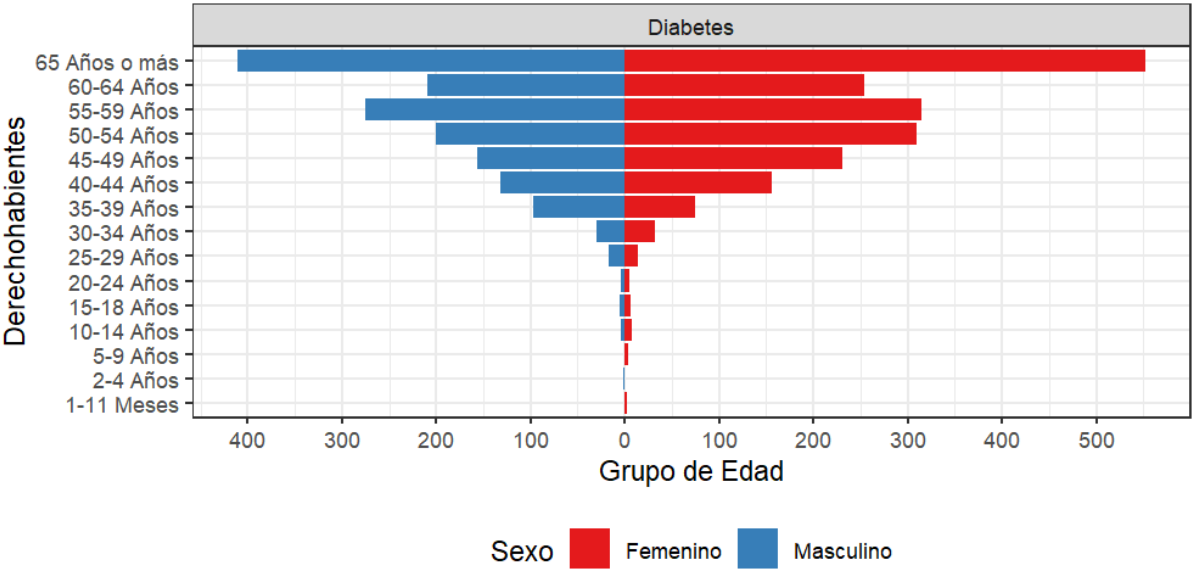


Figura 4.10. Edades y géneros de pacientes con diabetes

Seguido de las edades y géneros, se presentarán los otros campos que se registraron durante las visitas. Con respecto al campo de medicina a la que se remite al derechohabiente, en la Figura 4.11 se presentan los 10 principales, de éstos en primer lugar se encuentra la medicina interna, seguido de oftalmología y finalmente ginecología y obstetricia (éste último a diferencia presentado principalmente en trastorno de mama, mientras los demás liderados por diabetes).

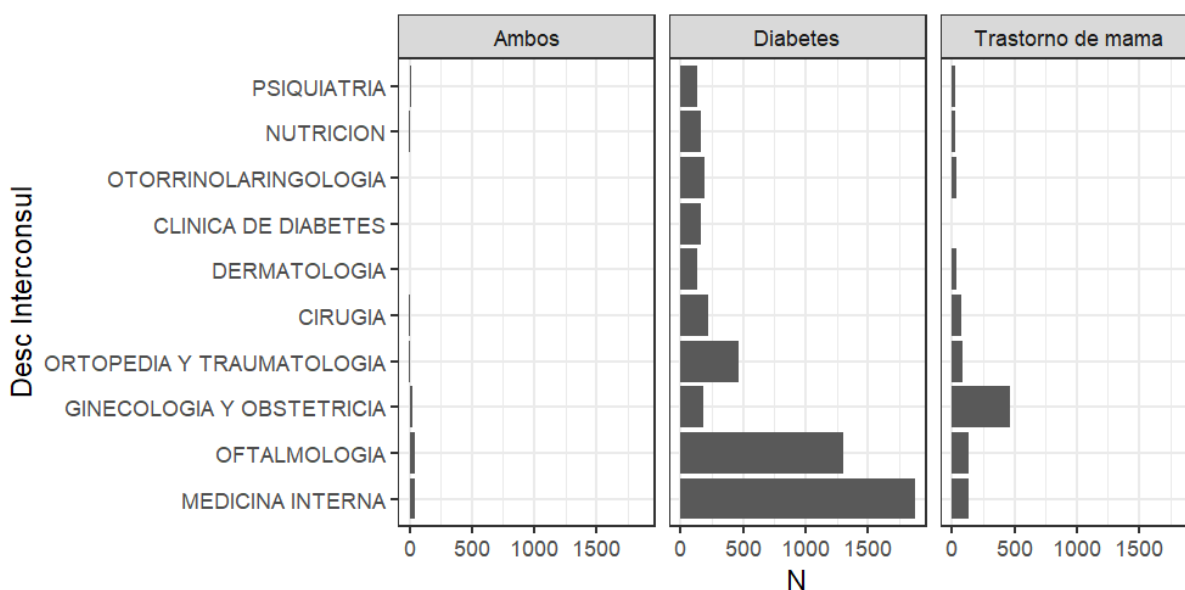


Figura 4.11. Interconsulta según grupo de derechohabientes (top 10)

Se registraron en las visitas tres tipos de estudios: rayos X, electrocardiograma y examen de mama, de los cuales principalmente predominó que no se presentaran ninguno de estos tres estudios durante las visitas.

Con respecto a recetas, se registra tanto si se prescribió alguna receta (Sí o No), consecuentemente cuántas recetas y cuántos medicamentos fueron prescritos; éstos dos últimos sólo en caso de que se responda que Sí a la prescripción de recetas. Éstos se presentarán a continuación.

En la Figura 4.12 se presenta la relación de visitas a médico general que presentaron recetas médicas. En los tres grupos de derechohabientes predominan las visitas con recetas, de las cuales, diabetes presenta una cantidad considerable de visitas con receta, mientras que trastorno de mama o ambos presentan conteos similares entre sí y no.

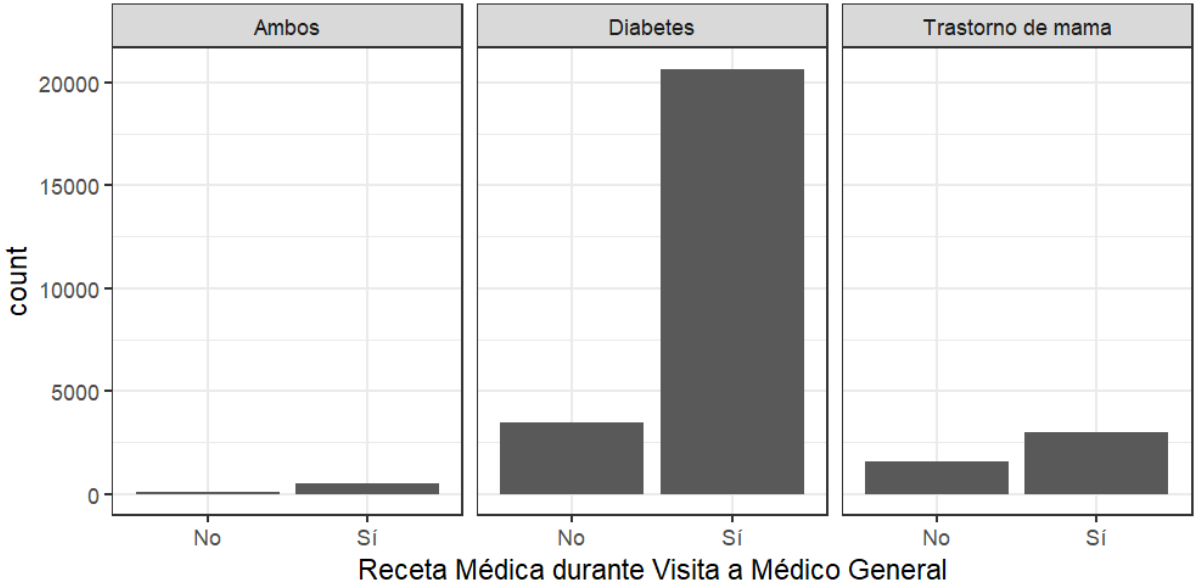


Figura 4.12. Prescripción de receta durante visita según grupo de derechohabientes

De éstos, en la Figura 4.13 se presenta la cantidad de recetas prescritas. En pacientes con diabetes, la mayoría se centran en una o dos recetas; mientras que en trastorno de mama y ambos predomina una receta médica.

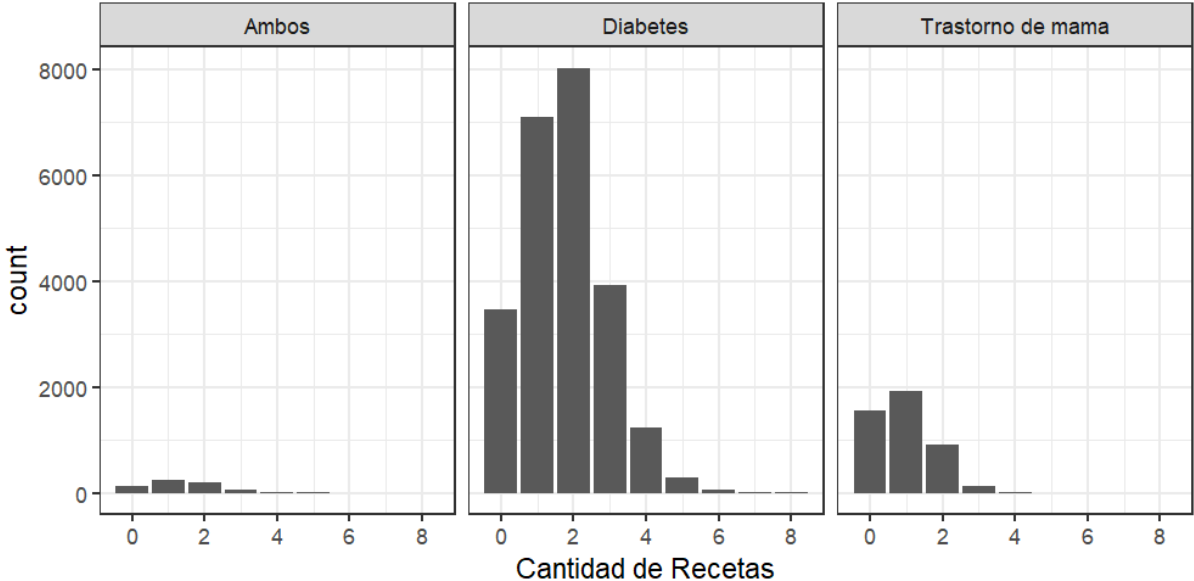


Figura 4.13. Cantidad de recetas por visita según grupo de derechohabientes

Finalmente, la cantidad de medicamentos que son prescritos se presentan en la Figura 4.14. En diabetes predominan 0, 4 y 5 medicamentos prescritos, mientras que en trastorno de mama predominan cero medicamentos.

Esta relación entre recetas y número de medicamentos con diferencias significativas entre ambos grupos de derechohabientes indica la oportunidad de implementar, en caso de que no esté implementado en la actualidad, de cálculo de costos en medicamentos por cada padecimiento. Esto podría ser que, a partir del diagnóstico principal del paciente, se identificaran todos los medicamentos que fueron prescritos al paciente y cuantificarlos de esa manera.

Aunque los tratamientos de cáncer resultan más costosos, el volumen de medicamentos y derechohabientes con diabetes pudiera superar los totales. A pesar de que no se cuenta con los montos e información a ese nivel (además que no es parte del alcance de esta investigación), el instituto pudiera hacer un análisis a fondo para determinar sus costos y tomar decisiones de inversión al momento de aplicar medidas preventivas para dichos padecimientos.

Entre estos cálculos se deberán incluir los medicamentos que no son propiamente para el tratamiento de la enfermedad, pero se administran como consecuencia de esta, por ejemplo, la administración de antibióticos o material de curación en la atención de heridas a causa de pie diabético. No se debe sólo sacar totales de los medicamentos exclusivos para el tratamiento de la enfermedad (insulina para diabéticos, quimioterapia para pacientes con cáncer) sino analizar el total de tratamientos por pacientes con cada diagnóstico.

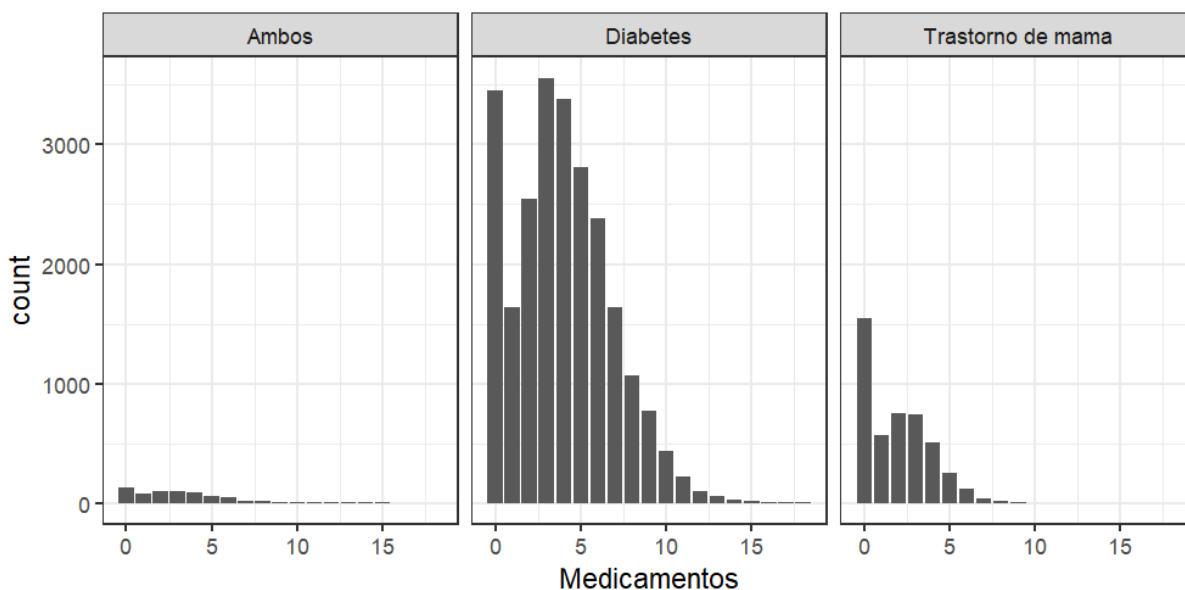


Figura 4.14. Cantidad de medicamentos por visita según grupo de derechohabientes

En las siguientes gráficas se presentarán los diagnósticos de derechohabientes partiendo de lo general a particular del diagnóstico. La clasificación se realiza según el código internacional de enfermedades (CIE10 o ICD10 por sus siglas en inglés). El orden se presenta en la Figura 4.15 que va desde lo general (arriba) a lo particular (abajo). En la Figura 2.5 se presenta un ejemplo de codificación y una descripción de cada uno.

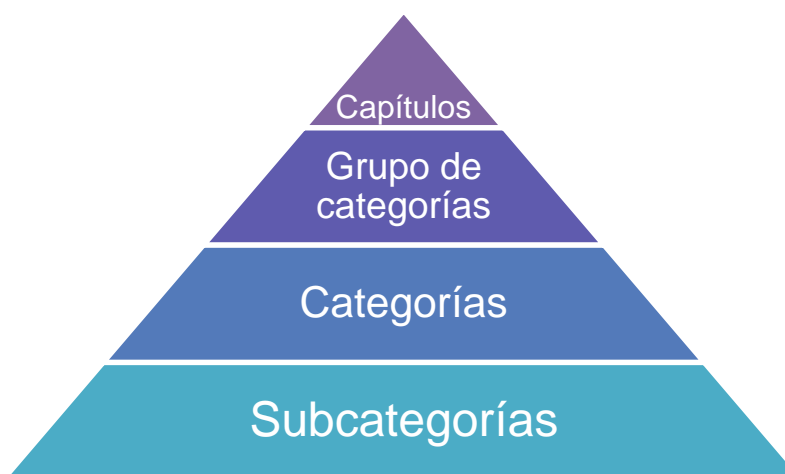


Figura 4.15. Orden de clasificación de enfermedades ICD10

En la Figura 4.16, se presentan los capítulos de diagnósticos de derechohabientes con trastorno de mama. Como es de esperarse, el principal son enfermedades del sistema genitourinario, seguido cercanamente de enfermedades de sistema respiratorio.

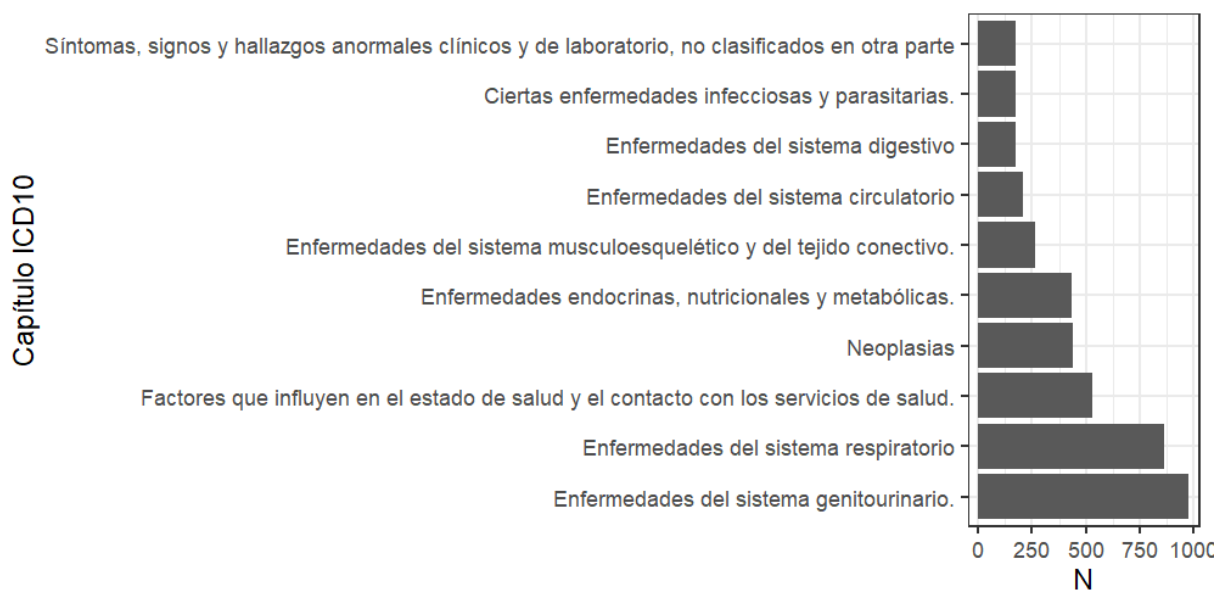


Figura 4.16. Capítulos de diagnósticos (ICD10) de derechohabientes con trastorno de mama (top 10)

La principal enfermedad del sistema genitourinario se presenta principalmente en Figura 4.17 como masa no especificada de la mama, mientras que la enfermedad del sistema respiratorio se presenta como resfriado común. Otros de los padecimientos no relacionados con resfriados y trastorno de mama fueron obesidad, hipotiroidismo, hipertensión y dorsalgia.

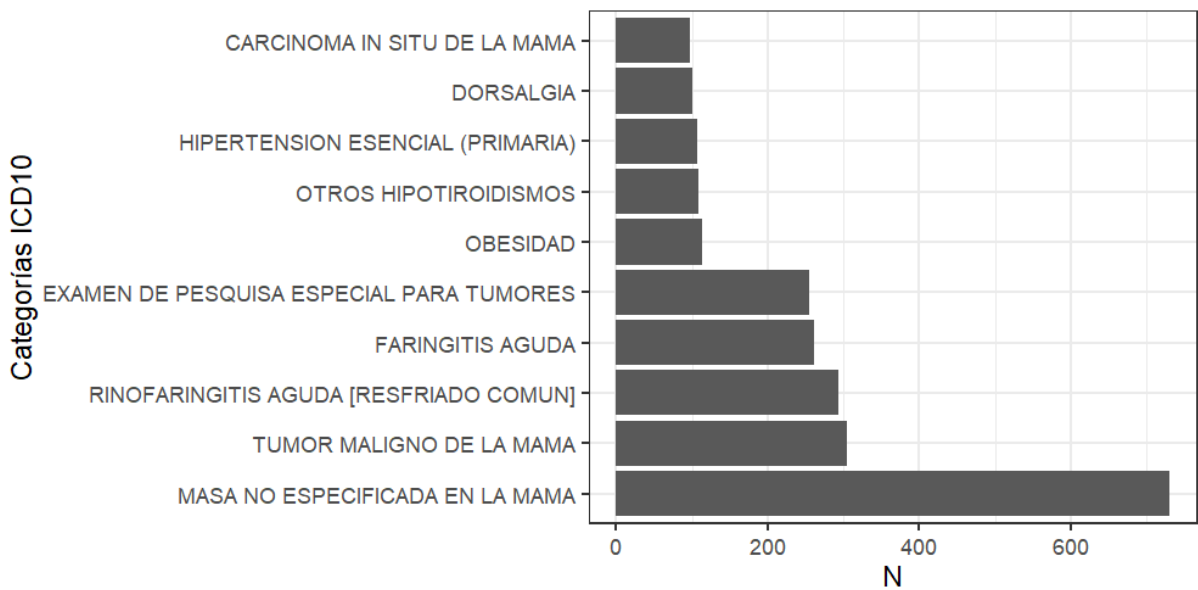


Figura 4.17. Categorías de diagnósticos (ICD10) de derechohabientes con trastorno de mama (top 10)

Por parte de los derechohabientes con diabetes, presentan un sesgo más significativo en las visitas por enfermedades endocrinas, nutricionales y metabólicas (Figura 4.18).

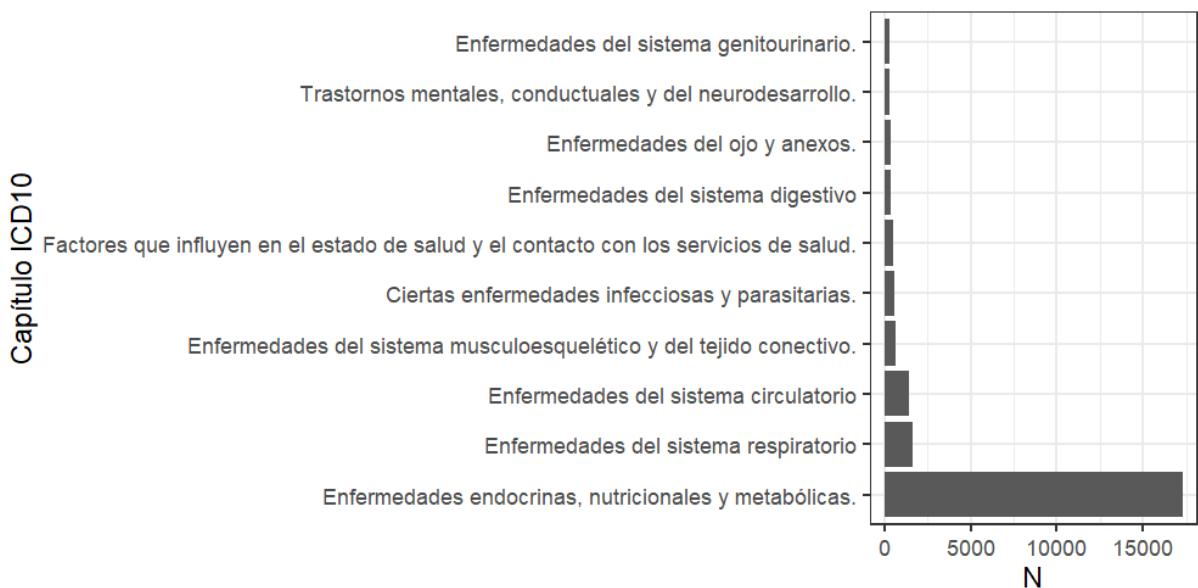


Figura 4.18. Capítulos de diagnósticos (ICD10) de derechohabientes con diabetes (top 10)

De manera más específica, en la Figura 4.19 se presenta las principales categorías de diabetes diagnosticadas por médico general; siendo esta diabetes mellitus no insulino dependiente. Otros de los diagnósticos que no son diabetes se presentan la hipofunción, hipertensión, resfriado esencial y obesidad.

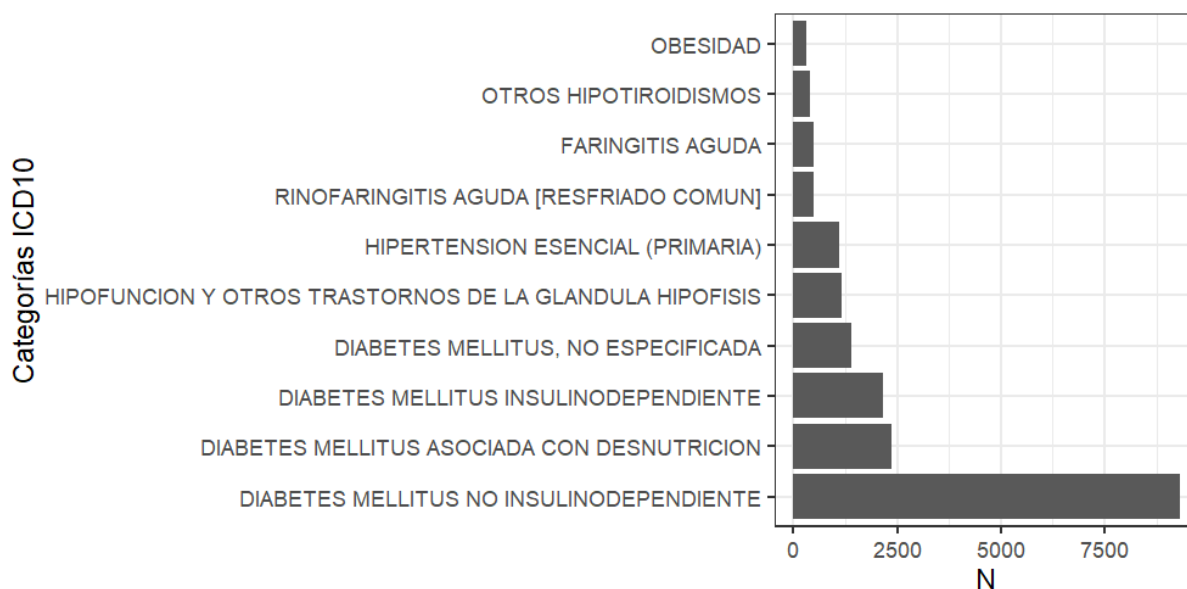


Figura 4.19. Categorías de diagnósticos (ICD10) de derechohabientes con diabetes (top 10)

Los diagnósticos que se presentaron que no fuese específicamente diabetes o algún trastorno de mama fueron los siguientes. De éstos, se presentó principalmente hipertensión, rinofaringitis y faringitis agudas con un porcentaje significativo en ambos grupos. Además, se presentan trastornos psicológicos en ambos grupos.

Diagnóstico	Diabetes	Trastorno de mama
Hipertensión esencial (primaria)	4.5%	1.8%
Rinofaringitis aguda (resfriado común)	2.0%	6.3%
Faringitis aguda	1.9%	5.2%
Hipotiroidismo, no especificado	1.5%	2.1%
Obesidad	1.2%	2.3%
Infección de vías urinarias, sitio no especificado	0.4%	1.9%
Diarrea y gastroenteritis en presunto origen infeccioso	0.8%	1.3%
Trastorno mixto de ansiedad y depresión	0.4%	1.0%
Trastorno de ansiedad, no especificado	0.3%	0.5%
Insuficiencia venosa (crónica) (periférica)	0.6%	1.1%

Tabla 4.6. Principales diagnósticos generales presentados por derechohabientes de ambos grupos

Para identificar qué padecimientos de los derechohabientes afectan su calidad de vida, los diagnósticos de cada derechohabiente fueron comparados con dos índices de comorbilidad. Comparando el número de trastornos (comorbilidades) identificados por ambos en la Figura 4.20, Elixhauser presentaba un mejor comportamiento al clasificar un número mayor de derechohabientes con uno o más trastorno y contando con más clasificaciones con número considerable de derechohabientes.

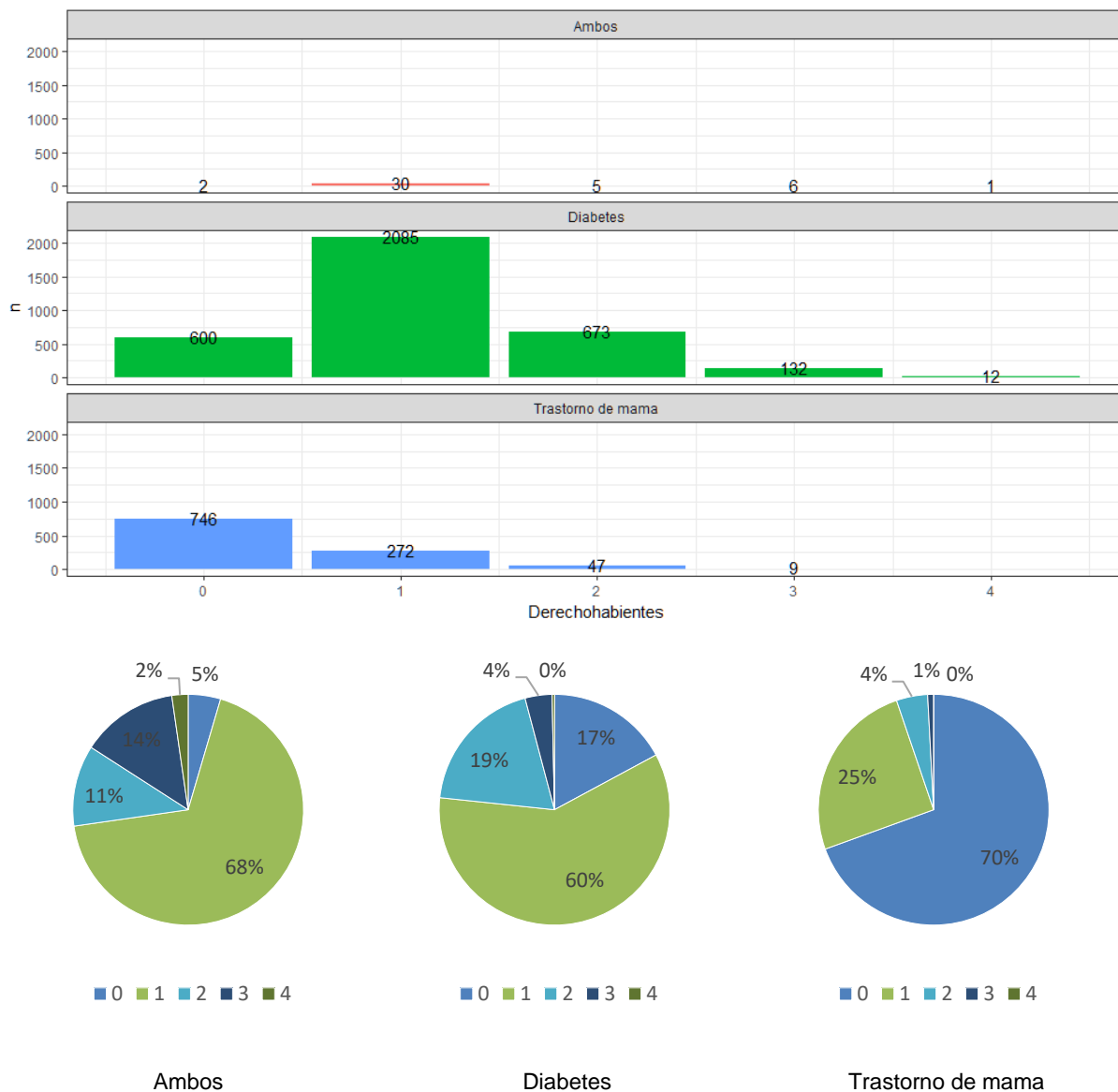


Figura 4.20. Trastornos identificados en derechohabientes según grupo de interés (Índice de comorbilidad de Elixhauser)

En comparación, revisando la Figura 4.21 con el índice de comorbilidad de Charlson, la mayoría de los derechohabientes que se identificaron con uno o más trastornos (comorbilidades) se presentan en gran medida en sólo un trastorno y un porcentaje muy bajo fue clasificado en dos o más trastornos.

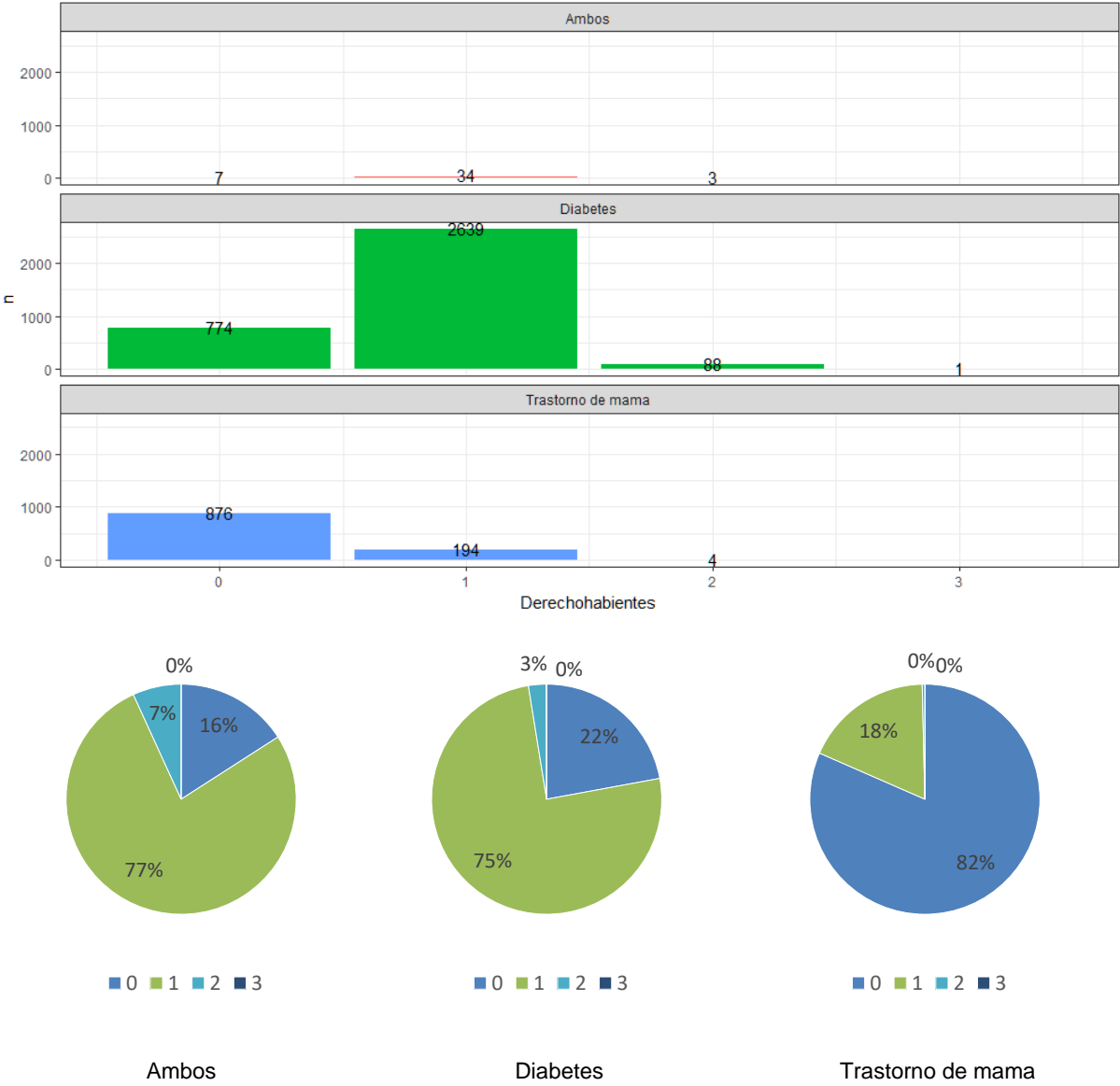


Figura 4.21. Trastornos identificados en derechohabientes según grupo de interés (índice de comorbilidad de Charlson)

Debido a que Elixhauser presentaba un índice con comorbilidades que se relacionaban a una cantidad mayor de derechohabientes, se estará analizando el índice con

respecto a varias características de los derechohabientes. En la Figura 4.22, se presenta una mayor frecuencia entre más grande su edad, con excepción a cero trastornos identificados que presenta una frecuencia total más homogénea entre un mayor grupo de edades. Con ello, se puede identificar que aquellos con cuatro comorbilidades se encuentra exclusivamente en derechohabientes con más de 45 años.

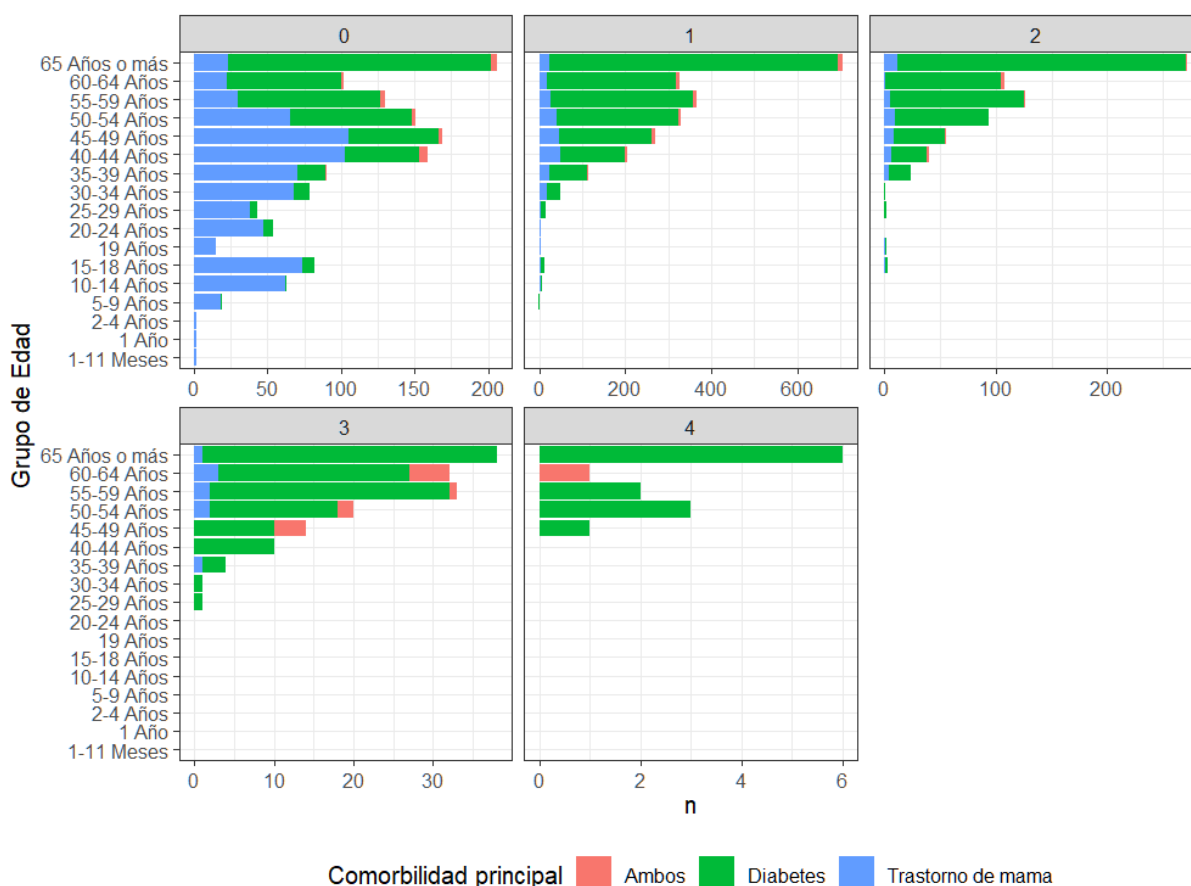


Figura 4.22. Comorbilidades de Elixhauser según grupo de edad y grupo de derechohabientes

Con respecto a los géneros que afectan en el número de comorbilidades en derechohabientes, en la Figura 4.23 se identifica que independientemente del género, se presenta principalmente una comorbilidad.

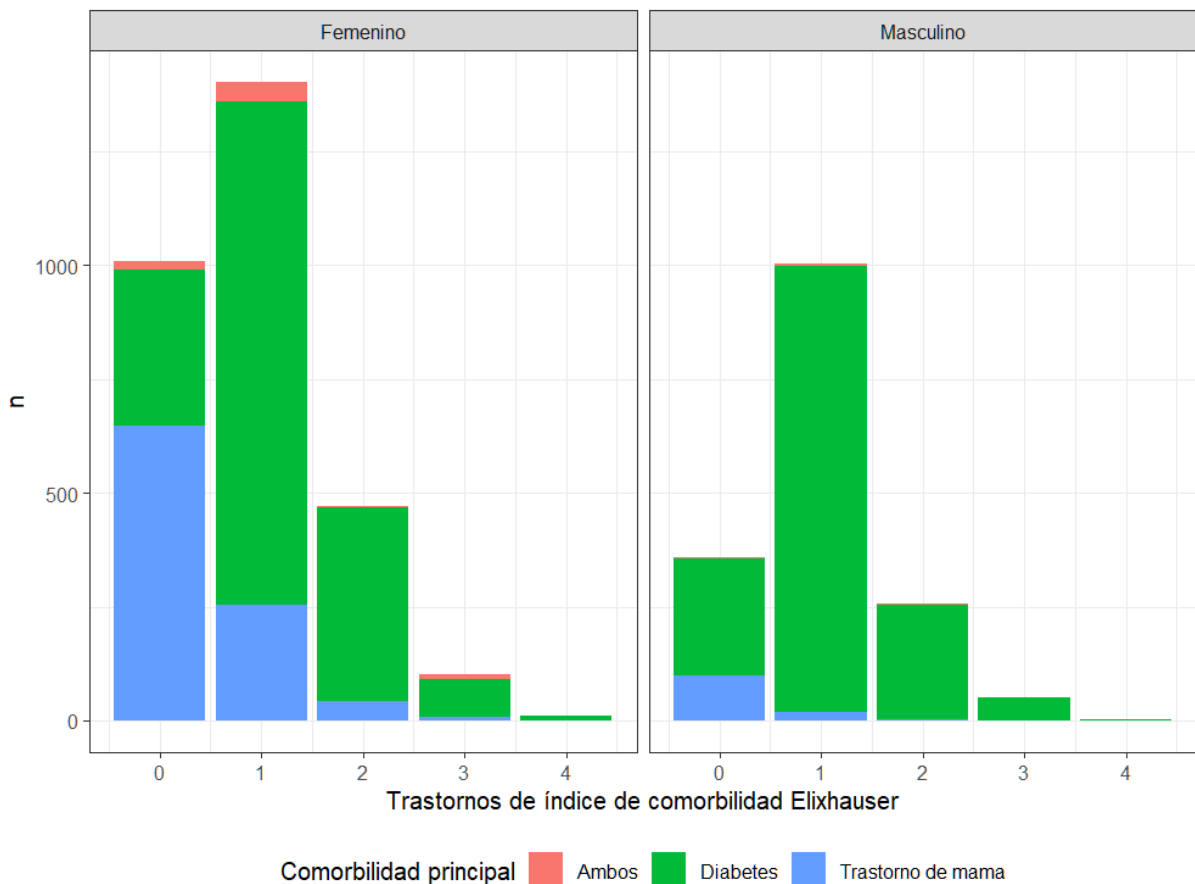


Figura 4.23. Trastornos de índice de comorbilidad Elixhauser identificados según sexo y grupo de derechohabientes

Revisando el número de comorbilidades según el grupo de derechohabientes y número de medicinas prescritas, la única diferencia significativa en la Figura 4.24 es que los derechohabientes con ambos (tanto trastorno de mama y diabetes) que presenta dos comorbilidades tiene una mediana de cinco medicinas prescritas durante sus visitas. En otro número de trastornos, las distribuciones con muy similares entre ellas por lo que posiblemente sea indiferente. Por otro lado, los pacientes con trastorno de mama presentan un cuartil de 25% constante en cero, lo cual se demostró en la Figura 4.14.

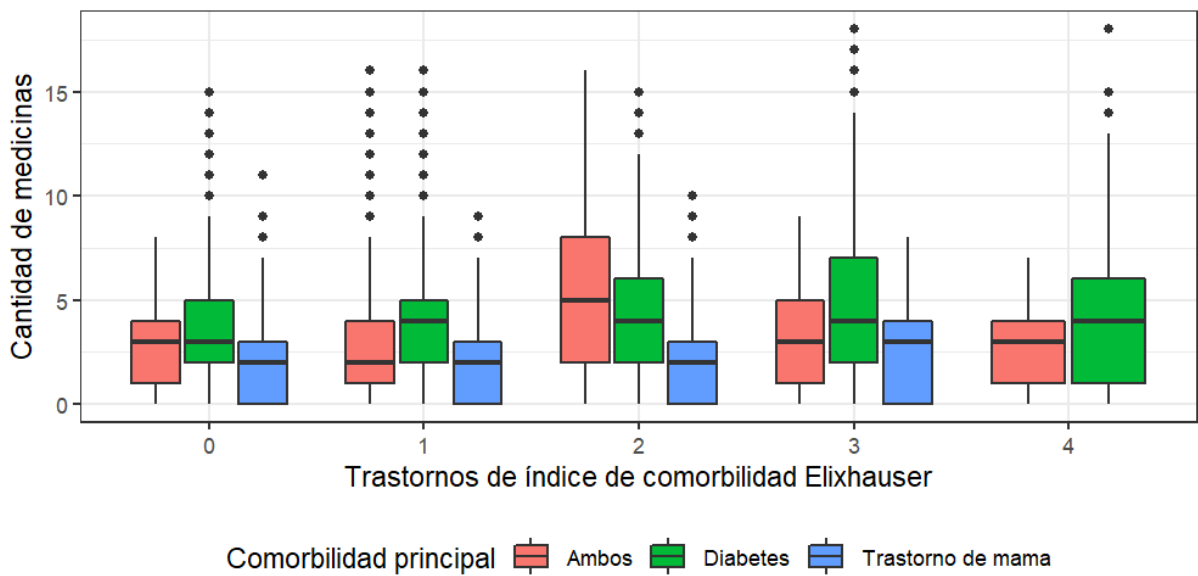


Figura 4.24. Cantidad de medicinas prescrita en las visitas de derechohabientes con respecto a número de trastornos de índice de comorbilidad Elixhauser y grupo de derechohabientes

Ahora, analizando qué comorbilidades son identificadas principalmente en los derechohabientes por parte del índice de Elixhauser; éstos se presentan en la Figura 4.25. De éstos, en trastorno de mama se presenta mayormente tumor sólido sin metastásis para los derechohabientes con trastorno de mama, mientras que diabetes no complicada por parte de los derechohabientes con diabetes; ambos como es de esperarse. Las otras comorbilidades presentadas en derechohabientes con trastorno de mama son hipotiroidismo, obesidad, hipertensión no complicada y depresión. Por parte de los derechohabientes con diabetes, las comorbilidades relacionadas son hipertensión, hipotiroidismo y obesidad.

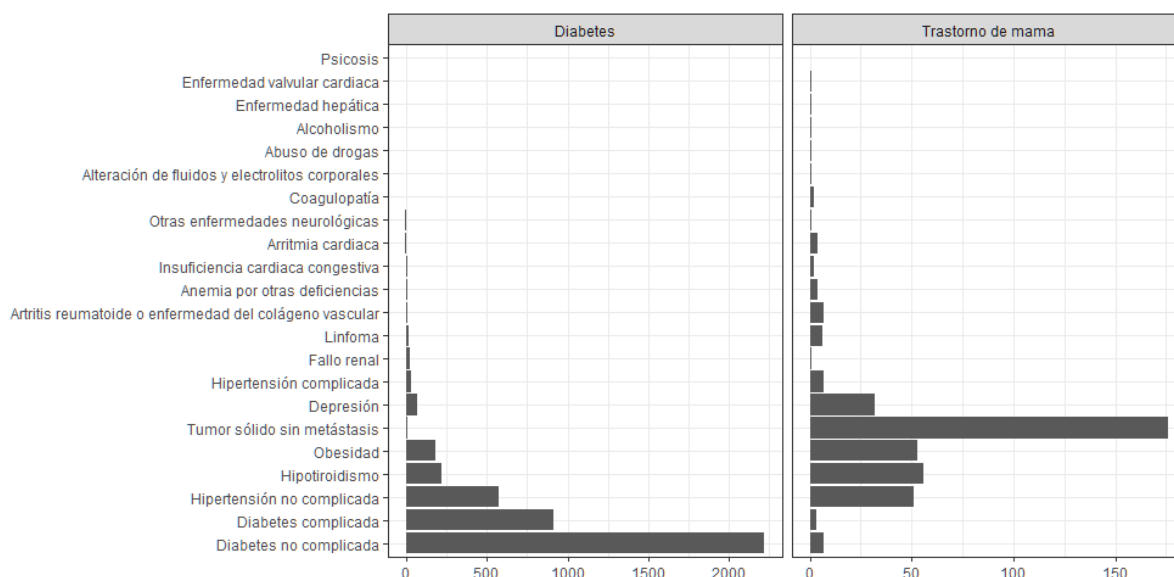


Figura 4.25. Trastornos presentados en derechohabientes según índice de Elixhauser

4.2.2 Preparación de datos

Un paso exhaustivo de preparación de datos fue el relacionar cada diagnóstico con el código de enfermedades (ICD-10). Esto fue realizado en Excel de manera automática para la mayoría de los registros y aproximadamente 30% de los diagnósticos fueron relacionados manualmente. Con el código ICD-10, fue posible conocer el capítulo, grupo de categoría, categoría y subcategoría de la enfermedad.

A partir del ICD-10, se calcularon los índices de comorbilidad de Elixhauser y Charlson para los derechohabientes utilizando la paquetería “Comorbidity” de R (Gasparini, 2019).

Se realizó análisis de tiempos entre visitas para cada periodo de diagnóstico (entre 59 a 93 días al año), además de análisis de visitas acumuladas en el último mes para cada registro de visita.

4.3. Identificación de herramientas

Esta etapa consiste en dos pasos: análisis de algoritmos y selección de algoritmos. Éstos serán desarrollados a continuación.

4.3.1 Análisis de algoritmos

En este paso se realiza un análisis de herramientas posibles para analizar la información disponible y generar con ello la evaluación de riesgo de salud de derechohabientes. Para ello, se identificaron los modelos de conglomerados, regresión, árboles de decisión y redes bayesianas. Lo importante de considerar, es que puede darse el caso de que se evalúen valores que no son numéricos; por lo que sería favorable poder utilizar una forma de clasificación que considere también factores.

- Redes bayesianas considera tanto variables discretas como continuas, considerando la dependencia entre las variables.
- Análisis de conglomerados realiza agrupación de variables numéricas según la distancia entre ellas.
- Regresión realiza la función de predecir o entender la relación de variables independientes con dependientes.
- Árboles de decisión realizan predicción utilizando bifurcaciones de variables tanto discretas como continuas, como redes bayesianas.

4.3.2 Selección de algoritmos

Para cada aplicación del problema se estuvieron eligiendo algunos algoritmos para para modelo, con el fin de evaluar su rendimiento y confiabilidad.

4.3.2.1 Identificación de algoritmos para factores relevantes

Una de las primeras preguntas que se presentaron en el estudio es si a partir de estos dos grupos distintos de derechohabientes (aquellos diagnosticados con diabetes contra aquellos con trastorno de mama), se podrían identificar aquellas variables que representarían, esto es, los atributos que definen las características de los registros de cada uno de los derechohabientes. Como, por ejemplo, el trastorno de mama es explicado por la variable de sexo en que principalmente las mujeres presentan este diagnóstico, o la diabetes es explicado por la variable de edad en que principalmente

se presenta este diagnóstico en adultos. Para ello, se realizó un estudio sobre cada tipo de derechohabiente para identificar qué algoritmos se podrían adaptar mejor para identificar variables relevantes. Estos métodos de clasificación de variables cualitativas se presentan a continuación:

- Métodos lineales
 - Análisis discriminante lineal (Amat, 2016)
 - Regresión logística (Fiuza-Pérez y Rodríguez-Pérez, 2000)
 - Modelo lineal generalizado (Hastie y Qian, 2014)
- Métodos no lineales
 - Máquina de soporte vectorial (Le, 2018)
 - K vecinos más cercanos
- Conjuntos de árboles
 - C5.0
 - Modelado de regresión generalizada

4.3.2.2 Redes bayesianas

Con respecto al modelo de redes bayesianas, se identificó su utilidad en relacionar los padecimientos y características biométricas de los pacientes con respecto a su dependencia. Para ello, se identificaron los siguientes algoritmos a evaluar:

- Hill-climbing (ascenso de colinas) que realiza una búsqueda iterativa hasta que el resultado no puede ser mejorado (Sucar, 2006)
- TABU(Scutari, 2020)

4.3.2.3 Análisis de conglomerados

Con respecto al análisis de conglomerados, se utilizó el algoritmo de k-means por Hartigan y Wong (1979).

4.4. Diseño de metodología

Esta etapa se realizó en dos pasos: modelado y diseño de plataforma. Cada uno de los pasos se presentan a continuación.

4.4.1 Modelado

Se implementó cada uno de los algoritmos para verificar funcionamiento con respecto al problema en cuestión. Para ello se utilizaron varias paqueterías de RStudio. Este programa se utilizó para el modelado ya que es una plataforma libre y enfocada al análisis estadístico.

4.4.1.1 Identificación de algoritmos para factores relevantes

Como se presentó en la selección de algoritmos, se están buscando aquellos algoritmos que funcionasen mejor para representar a los derechohabientes con diabetes y trastorno de mama (comparando unos de los otros debido a que son los dos grupos de derechohabientes que se tienen, pero la metodología podría aplicarse para derechohabientes más heterogéneos). Para ello, se utilizó la paquetería caret de R (Kuhn, 2019) a fin de evaluar estos modelos entre ellos e identificar cuáles tendrían mejor rendimiento. El primer estudio que se realizó fue sobre los derechohabientes diagnosticados con diabetes (que presenten el grupo de categoría según la IDC-10 Diabetes Mellitus). Los datos evaluados se presentan en la Tabla 4.7 donde el sexo se evalúa como 2 (masculino) y 1 (femenino), el grupo de edad se presenta como el promedio del grupo y el resto de las columnas presentan grupos de categorías de enfermedades según IDC-10. Los grupos de categoría evaluados fueron aquellos que fueran diagnosticados más de 100 veces en derechohabientes y presentan 0 si no fue diagnosticado con ello, o 1 si fue diagnosticado al menos una vez en el periodo de tiempo evaluado. La columna de interés, en el caso de la figura como diabetes mellitus, se evaluó como factor; a diferencia del resto que se presentó como numérico.

Sexo	Grupo Edad	X1	X11	X15	X22	X23	X33	...	X210
2	65	0	0	0	0	0	1	...	0
2	52	0	0	0	0	1	1	...	0

Sexo	Grupo Edad	X1	X11	X15	X22	X23	X33	...	X210
1	47	0	0	0	0	0	1	...	0
1	47	0	0	0	0	0	0	...	0
1	52	1	0	0	0	0	0	...	0
1	65	0	0	0	0	0	1	...	0

Tabla 4.7. Ejemplo de tabla de evaluación de grupos de categorías de IDC-10 de derechohabientes

La evaluación presentó validación cruzada 10 veces y tres repeticiones para proporcionar una estimación sólida, los resultados se presentan en la Figura 4.26. Los primeros cuatro algoritmos presentaron resultados de exactitud y kappa muy similares entre ellos. Por su simplicidad, se estaría utilizando la regresión logística ya que no presenta una diferencia significativa entre éste y los principales tres algoritmos.

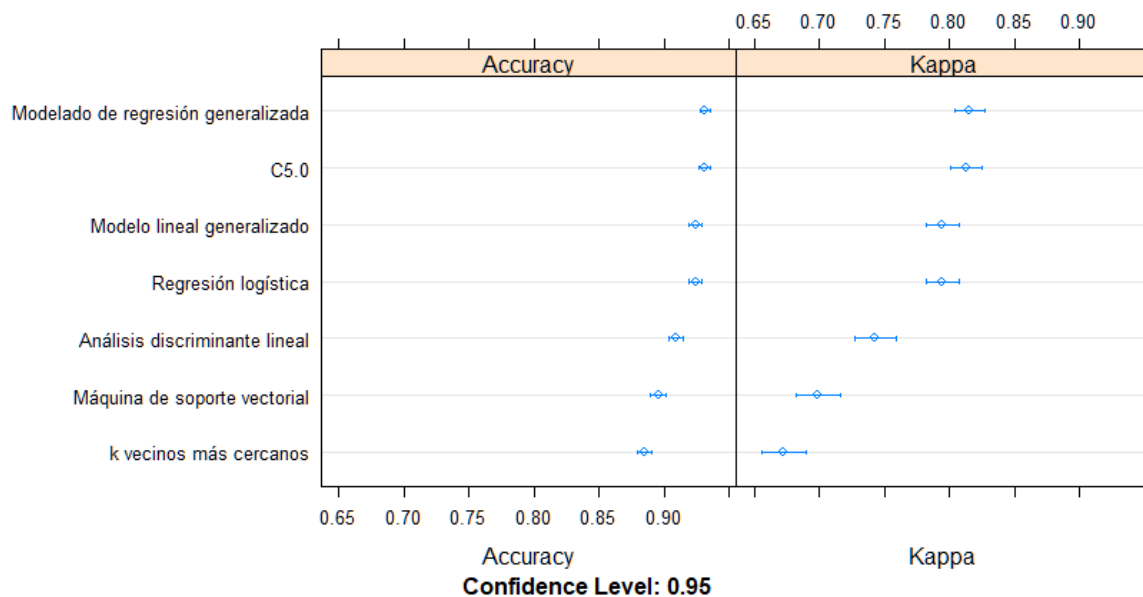


Figura 4.26. Evaluación de modelos para describir derechohabientes con diabetes

Se realizó el mismo estudio, pero en este caso evaluando a los derechohabientes con el grupo de categoría de “Trastornos de mama”. Éste, al igual que en derechohabientes con diabetes, presentó en los primeros cuatro algoritmos similares resultados de exactitud y coeficiente kappa (como se puede ver en la Figura 4.27). Por simplicidad, se selecciona de nuevo regresión logística.

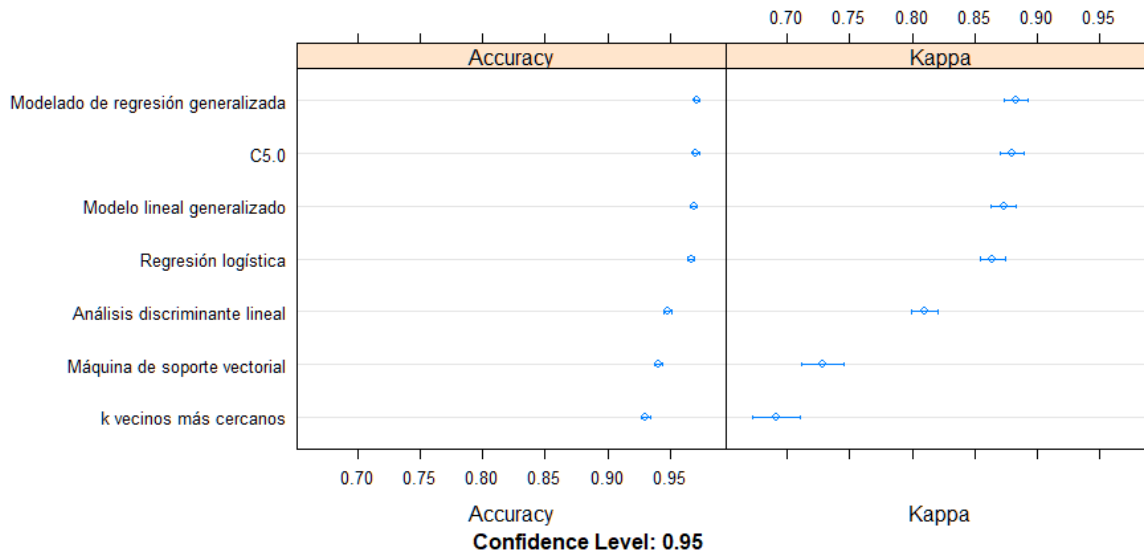


Figura 4.27. Evaluación de modelos para describir derechohabientes con trastorno de mama

Los resultados de la regresión logística para los diagnósticos de diabetes y trastorno de mama se presentan en anexos (7.7) ya que se presentó un sesgo en los resultados debido a las diferencias entre ambos grupos de derechohabientes. Se recomienda realizar este estudio en un grupo con padecimientos más heterogéneos, pero se presenta esta metodología por su potencial uso en los datos de la institución principalmente para el departamento de medicina preventiva.

4.4.1.2 Redes bayesianas

Con respecto a redes bayesianas, se probaron varios algoritmos para evaluar cuál presentaba mejor rendimiento. El objetivo era encontrar tanto la relación entre las características de los pacientes y sus diagnósticos, como la relación entre los diagnósticos. Primero, se probó utilizando los datos de derechohabientes con respecto a las comorbilidades presentes según el índice de Elixhauser. El análisis se presenta a continuación, pero no fue implementado en la plataforma.

Comorbilidades según índice de Elixhauser

El análisis de comorbilidades según redes bayesianas se realizó utilizando el algoritmo de “Hill climbing” en RStudio con la paquetería bnlearn (Scutari, 2010). Un ejemplo de cómo se presenta la información para analizar con redes bayesianas se presenta en la Tabla 4.8. En esta, cada fila es un derechohabiente y cada columna es una característica presentada como factor: sexo, grupo de edad, y el resto comorbilidades que se presentaron en uno o más derechohabientes.

Sexo	Gpo.Edad	chf	cevd	dementia	copd	rheumd	pudd	mld	diab	diabwc	rend	canc
Masculino	65 años o más	no	no	no	no	no	no	no	yes	no	no	no
Masculino	50-54 Años	no	no	no	no	no	no	no	yes	yes	no	no
Femenino	45-49 Años	no	no	no	no	no	no	no	no	no	no	no
Femenino	45-49 Años	no	no	no	no	no	no	no	no	yes	no	no
Femenino	65 años o más	no	no	no	no	no	no	no	yes	no	no	no
Masculino	65 años o más	no	no	no	no	no	no	no	yes	no	no	no

Tabla 4.8. Ejemplo de tabla de evaluación de comorbilidades de derechohabientes

Se realizó tanto para derechohabientes con diabetes como con trastorno de mama. En la Figura 4.28, se presentan los derechohabientes con trastorno de mama. Entre las relaciones que se presenta, hay una relación entre sexo con insuficiencia cardiaca congestiva y obesidad (pero hay que considerar que hay un sesgo al sexo femenino), relación entre enfermedad hepática y alcoholismo (100% los que tienen enfermedad hepática presentan alcoholismo y 0% de los que no tienen enfermedad hepática presentan alcoholismo), y relación entre hipotiroidismo con artritis reumatoide (6% de los que tienen hipotiroidismo tienen artritis).

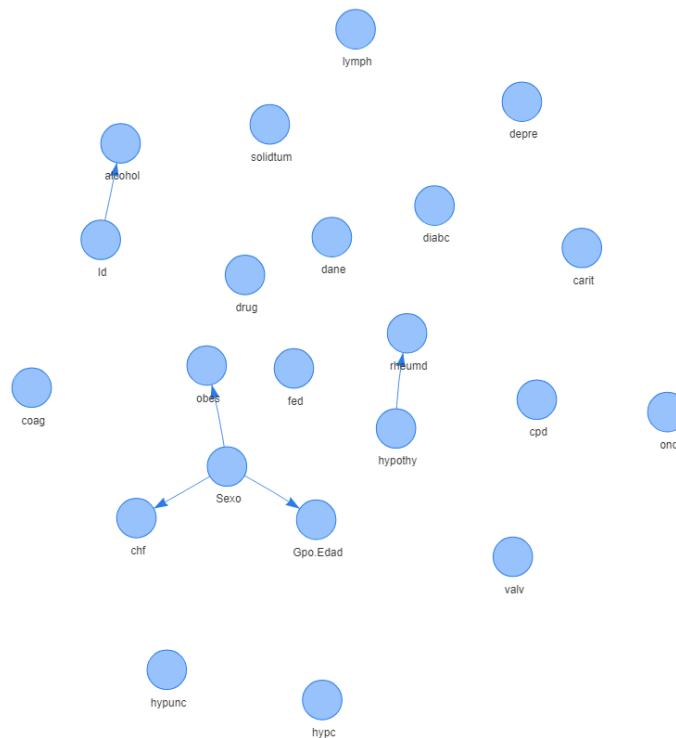


Figura 4.28. Red bayesiana de trastorno de mama

En la Figura 4.29 se presentan los derechohabientes con diabetes. Las relaciones que se presentaron fueron las siguientes: (1) sexo con arritmia cardiaca (100% de las mujeres no presentaron arritmia, de los hombres el 3%), (2) sexo con enfermedad pulmonar crónica (2% de las mujeres y 0.5% de los hombres), (3) sexo con hipotiroidismo (8.5% de las mujeres y 3% de los hombres), (4) arritmia cardiaca con hipertensión complicada (de los que presentaban arritmia, el 60% no presentaba hipertensión complicada, mientras que casi el 100% de los que no presentaban arritmia no presentaron hipertensión complicada), (5) hipertensión no complicada con obesidad (el 4% de los que no presentaban hipertensión no complicada presentaban obesidad, mientras que el 8% de los que presentaban hipertensión no complicada presentaban obesidad), (6) diabetes complicado con fallo renal (el 3% sin diabetes no presentaba fallo renal, mientras que el 1.6% que presentaba diabetes presentaba fallo renal), y (7) hipotiroidismo con hipertensión no complicada (el 15% sin hipotiroidismo presentaban hipertensión, mientras que el 25% con hipotiroidismo presentaba hipertensión).

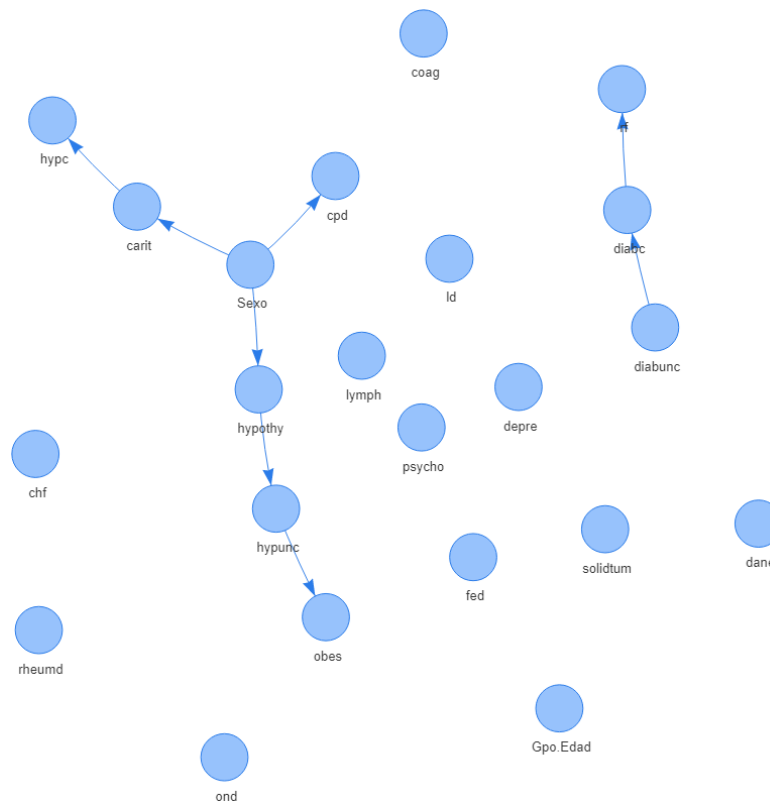


Figura 4.29. Red bayesiana de diabetes

chf	Insuficiencia cardiaca congestiva	aids	Enfermedad por VIH	ond	Otras enfermedades neurológicas
hypunc	Hipertensión no complicada	coag	Coagulopatía	rf	Fallo renal
diabunc	Diabetes no complicada	dane	Anemia por otras deficiencias	metacanc	Cáncer metastásico
pud	Úlcera péptica sin sangrado	valv	Enfermedad valvular cardiaca	wloss	Pérdida de peso
psycho	Psicosis	para	Parálisis	drug	Abuso de drogas
Blane	Anemia por pérdida de sangre	hypothy	Hipotiroidismo	Pvd	Enfermedad vascular periférica
depre	Depresión	lymph	Linfoma	Cpd	Enfermedad pulmonar crónica
carit	Arritmia cardiaca	obes	Obesidad	ld	Enfermedad hepática
hycp	Hipertensión complicada	alcohol	Alcoholismo	solidtum	Tumor sólido sin metástasis
diabc	Diabetes complicada	pcd	Enfermedad de la circulación pulmonar	fed	Alteración de fluidos y electrolitos corporales

rheumd Artritis reumatoide o enfermedad del colágeno vascular

Tabla 4.9. Códigos de comorbilidades de índice de Elixhauser

En el anexo 7.8 se presenta una continuación de análisis de combinaciones de comorbilidades según sexo, edad y diagnóstico de interés.

Grupos de categorías de enfermedades IDC-10

Después, las redes bayesianas se aplicaron entre los diferentes grupos de categorías de enfermedades IDC-10, sexo, grupo de derechohabiente y edad. Debido a que se registraron 953 subcategorías de diagnósticos presentadas en los derechohabientes, se optó por evaluar los grupos de categorías según la clasificación de ICD-10; de los cuales son 211 en total. De 150 grupos de categorías presentadas en derechohabientes, sólo 36 fueron diagnosticadas más de 100 veces; por lo que estos 36 grupos de categorías fueron los que se utilizaron para analizar los padecimientos de los derechohabientes.

Adicional a los padecimientos, los cuales se presentaron de forma binaria donde cada derechohabiente se presentó en una fila, se registró el sexo (1 como masculino y 2 como femenino) y según el grupo de edad se presentó la edad promedio. En la tabla 2 se presenta un ejemplo de cómo se ingresó la información al algoritmo. En ésta se puede apreciar en la primera columna el sexo, seguido de su grupo de edad y en las siguientes columnas el grupo de categoría evaluado (1 si lo padece, mientras que 0 si no). En total serían 36 columnas de cada padecimiento evaluado más los dos de características biométricas (sexo y grupo de edad). Un ejemplo de cómo se presentó esta información se encuentra en la Tabla 4.10.

Sexo	Grupo Edad	X1	X11	X15	X22	X23	X24	...	X210
2	65	0	0	0	0	0	1	...	0
2	52	0	0	0	0	1	1	...	0
1	47	0	0	0	0	0	1	...	0
1	47	0	0	0	0	0	0	...	0
1	52	1	0	0	0	0	0	...	0
1	65	0	0	0	0	0	1	...	0

Tabla 4.10. Ejemplo de tabla de evaluación de grupos de categorías de IDC-10 de derechohabientes. Las variables x1 a x210 (que son los grupos de categorías según ICD-10) se pueden revisar con su respectivo código en el anexo 7.6.

El análisis con redes bayesianas se realizó en RStudio con la paquetería “bnlearn” utilizando los algoritmos de “hill-climbing” y “TABU” (Scutari, 2010). Para validar los

resultados de cada columna, éste se realizó por medio de validación cruzada y correlación predictiva ya que se utiliza para evaluar redes bayesianas continuas (Scutari, 2017). Para esta validación, los valores pronosticados se calcularon promediando simulaciones de ponderación de probabilidad realizadas utilizando todas las columnas disponibles como evidencia (con excepción de la columna que se estaba evaluando). Como la variable es continua, el valor predicho es el valor esperado de la distribución condicional (Ness, 2019). Se utilizó una variación que permitiera evaluar columnas con pocos o sin padres (Scutari, 2017).

En la Figura 4.30 se presenta la red bayesiana de estas 38 variables.

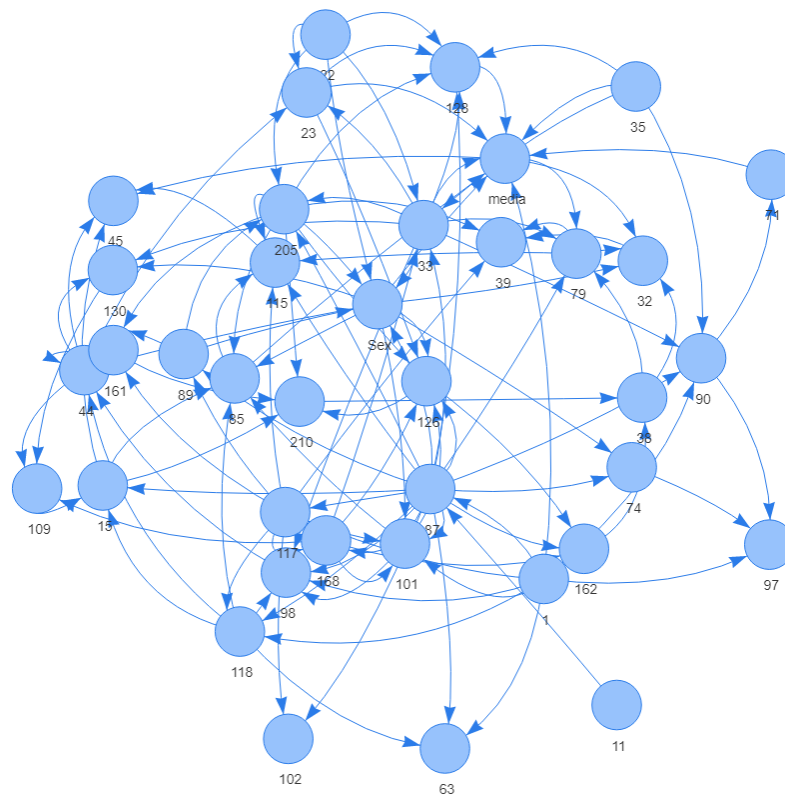


Figura 4.30. Red bayesiana de grupo de categorías (IDC-10)

En la Tabla 7.4 se presentan los grupos de categorías utilizadas para este estudio.

Como resultado de la validación, 10 variables presentaron una correlación superior al 20%, los cuales se presentan en la tabla Tabla 4.12 (en promedio de las 38 variables)

validadas se obtuvo una correlación de 16.75%). Se revisó el algoritmo TABU con el fin de verificar cuál tenía un mejor resultado de validación, presentando resultados muy similares y una correlación de 16.90% (una diferencia de 0.15%).

Diagnósticos y características biométricas	Correlación predictiva		Dif
	“hill climbing”	TABU	
Diabetes mellitus	0.80616	0.805901	0.000259
Trastornos de la mama	0.777412	0.776905	0.000507
Característica biométrica: Edad	0.689867	0.690102	-0.00023
Tumores malignos	0.439411	0.441348	-0.00194
Personas en contacto con los servicios de salud para investigación y exámenes	0.398917	0.397504	0.001413
Característica biométrica: Sexo	0.306319	0.306739	-0.00042
Infecciones agudas de las vías respiratorias superiores	0.285861	0.285991	-0.00013
Trastornos de otras glándulas endocrinas	0.243522	0.24422	-0.0007
Tumores in situ	0.2287	0.230357	-0.00166
Enfermedades hipertensivas	0.21395	0.212334	0.001616

Tabla 4.11. Resultados de red bayesiana con “Hill-climbing” y TABU

La más alta (diabetes mellitus) presentaba una correlación negativa (esto quiere decir que la presencia de estos diagnósticos excluye la presencia de diabetes) con cuatro diagnósticos, los cuales eran tumores malignos, trastornos de otras glándulas endocrinas, infecciones agudas en las vías respiratorias y síntomas y signos generales. En la Figura 4.31 se ilustra la relación entre la diabetes y los diagnósticos. El símbolo negativo explica que la ausencia del diagnóstico (por ejemplo, de tumores malignos) conlleva a la presencia de diabetes. El grosor de la línea presenta que los tumores malignos es un diagnóstico que afecta en mayor medida el pronóstico de diabetes. Al presentarse sólo 44 derechohabientes con diabetes y trastorno de mama, era de esperarse que los diagnósticos de trastorno de mama afectaran al pronóstico de diabetes.

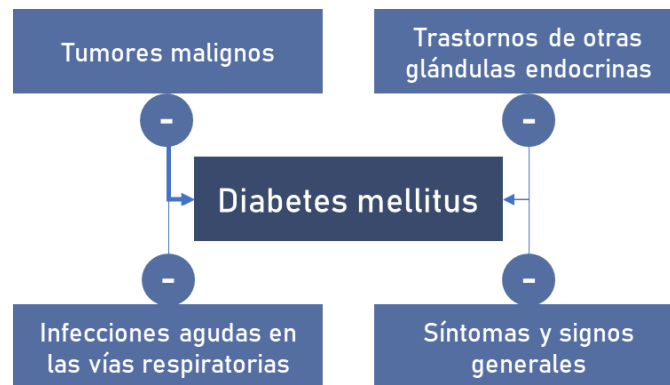


Figura 4.31. Diagnósticos que explican a la Diabetes Mellitus según análisis de redes bayesianas

La edad se veía afectada por siete diagnósticos y el sexo, siendo estas enfermedades infecciosas intestinales, tumores in situ, diabetes mellitus, trastornos de otras glándulas endocrinas, alteraciones de la visión y ceguera, enfermedades de las venas y de los vasos y ganglios linfáticos, y trastornos de la mama. De éstos, sólo enfermedades infecciosas intestinales y trastornos de la mama disminuyen la edad pronosticada. Por ello, se puede inferir que el resto de los trastornos se presentan en derechohabientes con mayor edad. En la siguiente figura se presentan los diagnósticos que explican la edad y qué tipo de relación presentan.

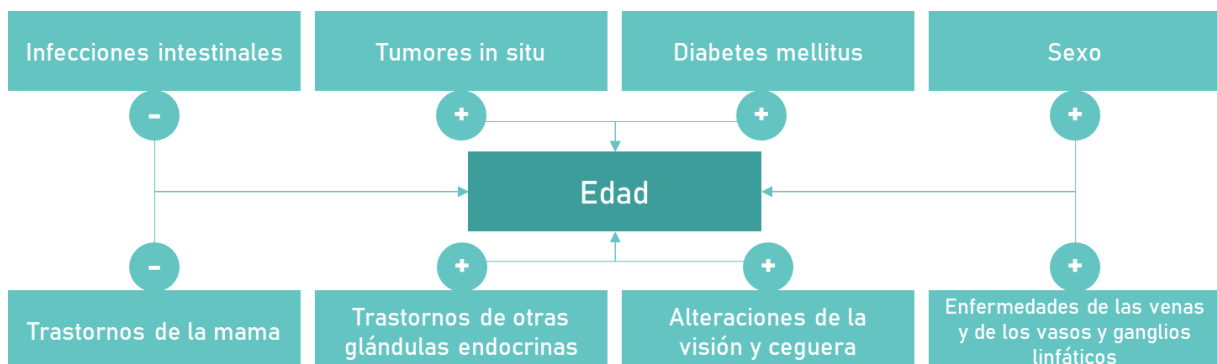


Figura 4.32. Diagnósticos que explican la edad según análisis de redes bayesianas

En la siguiente figura se presentan los diagnósticos que afectan principalmente la edad. En esto se puede apreciar que trastornos de la mama e infecciones intestinales presentan edades tanto en adultos como en la tercera edad; mientras los demás diagnósticos se presentan en mayor medida entre más avanzada sea la edad. La relación entre sexo y edad se puede apreciar a partir de la Figura 4.8.

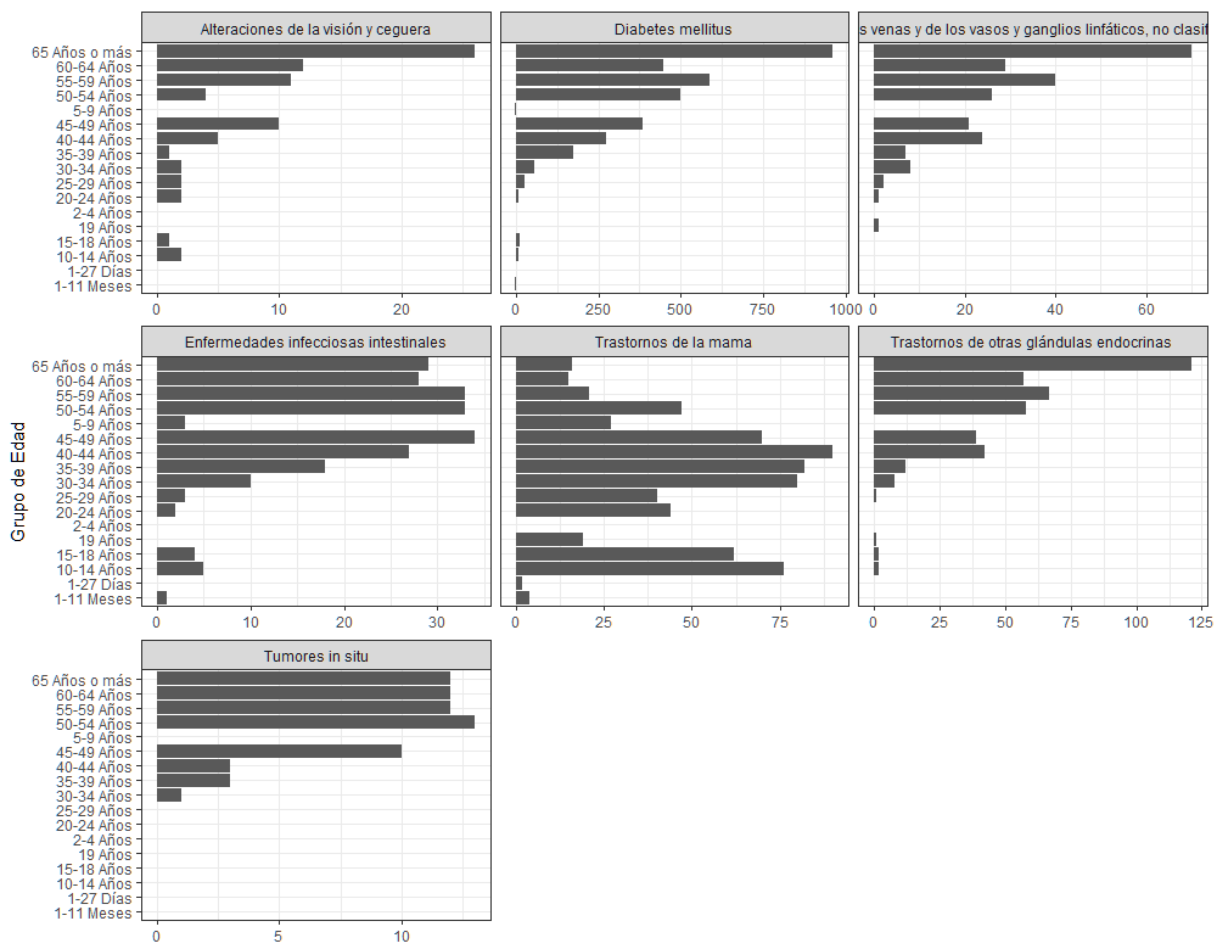


Figura 4.33. Principales diagnósticos que afectan la edad según redes bayesianas

El sexo fue afectado por cinco diagnósticos, siendo estos tumores malignos, diabetes mellitus, infecciones agudas de las vías respiratorias superiores, otras infecciones de las vías respiratorias superiores y personas en contacto con los servicios de salud para investigación y exámenes. Esto puede estar relacionado al sesgo de mujeres con trastorno de mama. En la siguiente figura, se presenta la relación de hombres y mujeres para cada diagnóstico, siendo principalmente el sexo femenino el que es diagnosticado.

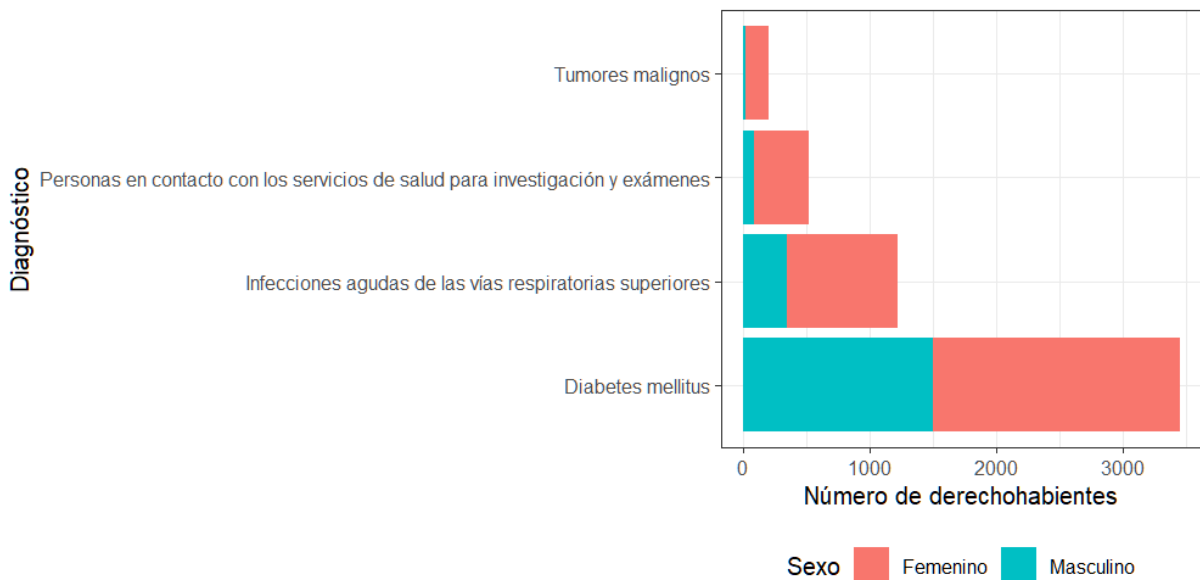


Figura 4.34. Principales diagnósticos que afectan el sexo según análisis de redes bayesianas

De los diagnósticos no relacionados con la muestra de derechohabientes obtenida:

- Infecciones agudas de las vías respiratorias superiores está relacionada con enfermedades infecciosas intestinales e infecciones virales caracterizadas por lesiones de la piel y de las membranas mucosas (ambas de forma positiva, por lo que de presentarse aumenta la probabilidad de presencia),

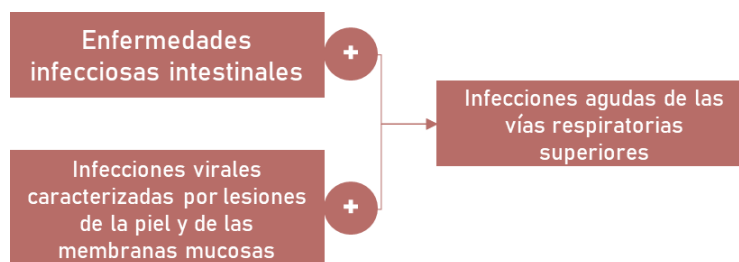


Figura 4.35. Principales diagnósticos que afectan las enfermedades infecciosas intestinales

- Trastornos de otras glándulas endocrinas está relacionada con cuatro variables: edad, los diagnósticos de trastornos de la mama, diabetes mellitus y otras enfermedades de las vías respiratorias superiores, y

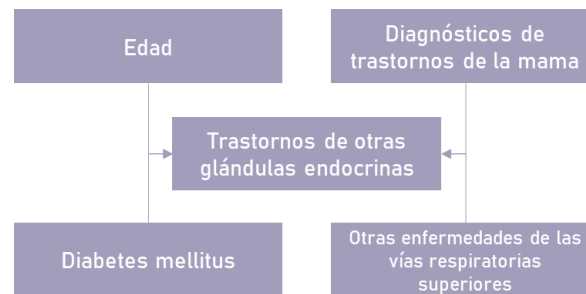


Figura 4.36. Principales diagnósticos que afectan las enfermedades infecciosas intestinales

- Enfermedades hipertensivas que está relacionada con la edad, diabetes mellitus, obesidad y otros tipos de hiperalimentación, e infecciones agudas de las vías respiratorias superiores.

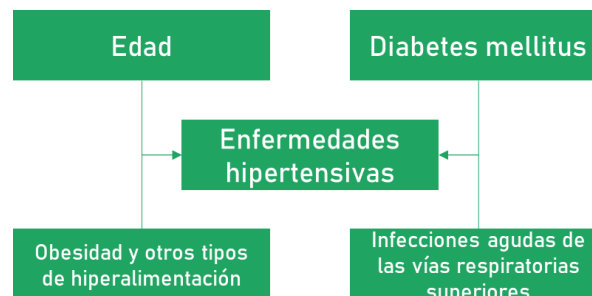


Figura 4.37. Principales diagnósticos que afectan las enfermedades hipertensivas

4.4.1.3 Análisis de conglomerados

Con respecto a análisis de conglomerados, se realizó el análisis con el algoritmo de k-means por Hartigan y Wong (1979) utilizando la paquetería stats en RStudio (R Core Team, 2019). Para ello, se utilizaron las siguientes variables por derechohabiente: promedio de cantidad de medicamentos de sus consultas, promedio de cantidad de recetas de sus consultas, sexo (1 si es femenino, 2 si es masculino), edad promedio de su grupo de edad y número de comorbilidades identificadas según el índice de Elixhauser. Un ejemplo de los datos se presenta en la Tabla 4.12, donde cada fila es un derechohabiente.

Medicamentos	Recetas	Sexo	Edad	Elixhauser
4.55	1.82	2	65	1
3	1.4	2	52	1
2	1.25	1	47	3
3.2	1.4	1	65	2

Tabla 4.12. Ejemplo de datos para evaluar por k-means

El número de recetas y de medicamentos se asume está relacionado con el estado de salud (una persona que toma mucho medicamento es menos sana que una que toma poco) pero no necesariamente es así, pues depende del tipo de medicamento y ese dato no se encuentra en la base de datos analizada. Esta muestra es solo para ilustrar el potencial de la técnica.

Debido a que el algoritmo evalúa distancias entre los datos, los rangos de los diferentes tipos de datos podrían sesgar los resultados según edad (ya que este presenta un mayor rango). Por ello, se normalizaron los datos para que todos estuviesen entre 0 y 1. Un ejemplo de cómo quedaron los datos se presenta en la Tabla 4.13.

Cantidad de medicamentos	Cantidad de recetas	Sexo	Edad	Elixhauser
0.303	0.303	1	1	0.25
0.2	0.233	1	0.798	0.25
0.133	0.208	0	0.721	0.75
0.213	0.233	0	1	0.5

Tabla 4.13. Ejemplo de datos normalizados para evaluar por k-means

Para conocer el número de grupos, se calculó el número óptimo de grupos. En la Figura 4.38 se presenta la gráfica la suma de cuadrados dentro del grupo el cual disminuye según se aumentan los grupos. El número de grupos se identificó como el “codo”, ya que la diferencia de suma de cuadrados deja de ser significativa. Para este caso, se podrían identificar que en 3 grupos es el óptimo.

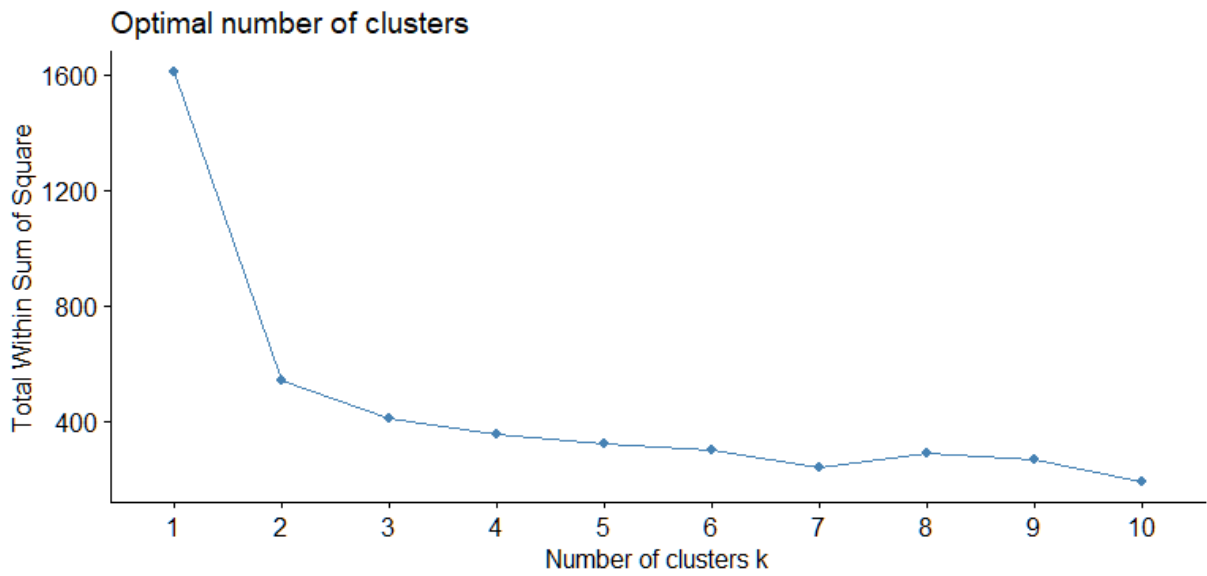


Figura 4.38. Número óptimo de grupos

Como demostración, se realizó el agrupamiento para 2, 3, 4 y 5 grupos, los cuales fueron graficados según sus variables más representativas. Éstas se presentan en la Figura 4.39.

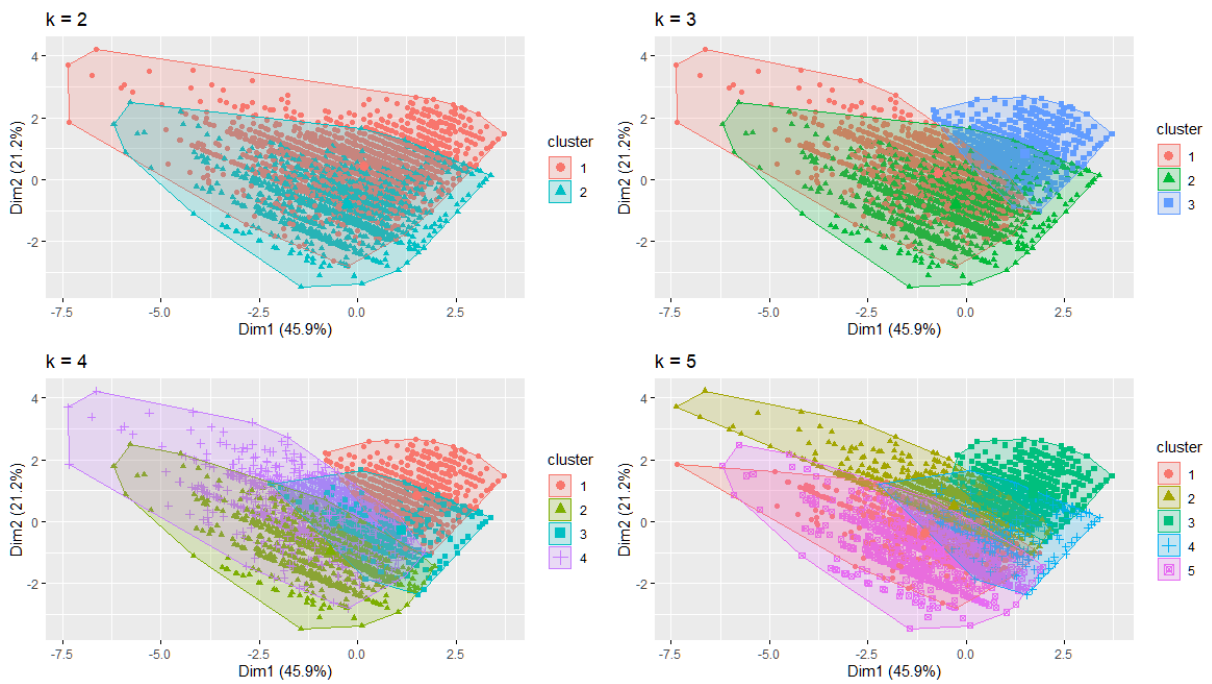


Figura 4.39. Agrupamiento por k-means entre 2 a 5 grupos

Los resultados de los grupos se presentan en la Tabla 4.14. En éste se puede apreciar que, según los promedios, el primer grupo presenta una mayor calificación de Elixhauser que afecta en el número de medicamentos, recetas y edad; al ser todas estos los mayores valores de los grupos; y además se presenta en mujeres. El segundo grupo se presenta en hombres con promedio de 53.5 años con una media de medicamentos y recetas, y un promedio de una comorbilidad de Elixhauser (siendo este en promedio 1). Finalmente, se presenta el grupo de mujeres con menor grupo de edad con los valores más bajos en promedios de medicamentos, recetas y comorbilidades de Elixhauser (siendo este último en promedio casi cero comorbilidades identificadas). Como se puede apreciar en la columna final, los tres grupos tienen un número similar de derechohabientes pertenecientes a cada grupo, siendo el menor de éstos el tercer grupo.

Grupo	Medicamentos	Recetas	Sexo	Edad	Elixhauser	n
1	3.80	1.69	Femenino	58.9	1.28	1802
2	3.21	1.44	Masculino	53.5	1.00	1665
3	1.72	0.84	Femenino	39.6	0.31	1153

Tabla 4.14. Grupos según k-means y centros de los grupos según cada variable

Debido a la aparente relación entre comorbilidades, medicamentos, recetas y edad, se procedió a evaluar la correlación entre ellas. En la Tabla 4.15 se presenta la correlación entre estas variables, de las cuales no se identificaron relaciones lineales significativas además de la esperada entre recetas y medicamentos (a mayor cantidad de medicamentos conlleva a una mayor cantidad de recetas).

	Medicamentos	Recetas	Edad	Elixhauser
Med		0.96	0.32	0.23
Rec	0.96		0.31	0.23
Edad	0.32	0.31		0.34
Elix	0.23	0.23	0.34	

Tabla 4.15. Correlación entre medicamentos, recetas, edad y comorbilidades de Elixhauser

En la Figura 4.22 se presenta la relación de edad con las comorbilidades de Elixhauser, mostrándose que a mayor edad se presenta hay una mayor densidad de

derechohabientes independientemente del número de comorbilidades identificadas. Viendo una comparación de las distribuciones como en la Figura 4.40, la mediana va elevándose cada vez que se aumenta el número de comorbilidades; manteniéndose estable en 2 y 4 comorbilidades (debido a que el último rango de edad es de 65 o más; por lo que se registra su media como 65). Que vaya disminuyendo la distribución según la edad, permite identificar que a mayor comorbilidad menos probable es que se presente en derechohabientes de menor edad. Como en el caso de los adultos mayores a 45 años, que es la única población que presenta cuatro comorbilidades. El análisis de comorbilidades de Elixhauser según grupo de edad, sexo y diagnóstico (si es parte de la población de diabetes o de cáncer de mama) se encuentra en el anexo.

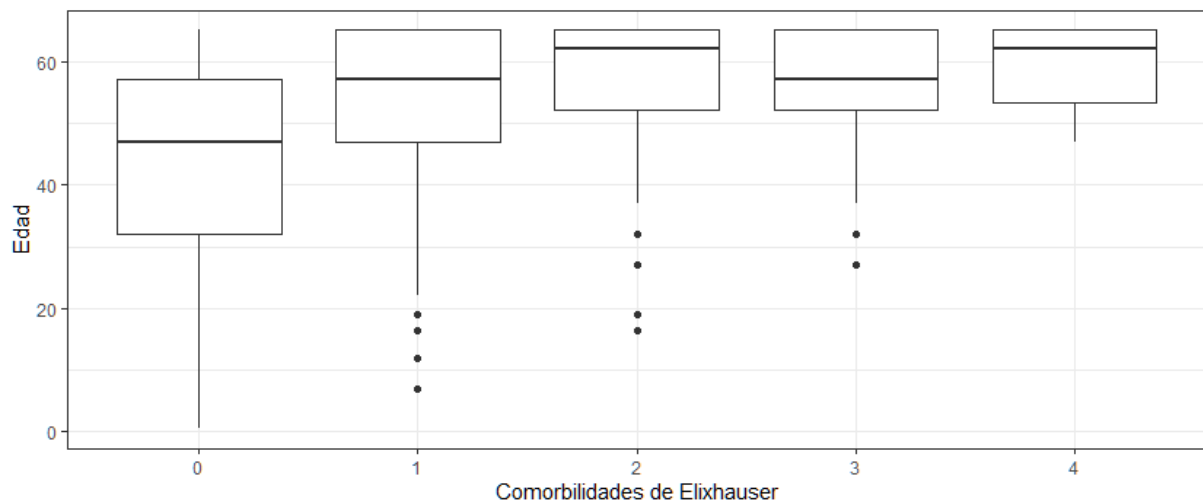


Figura 4.40. Comorbilidades de Elixhauser según la edad de derechohabientes

Con respecto a edad con medicamentos y recetas, se visualizó la edad con los medicamentos (ya que medicamentos presenta un rango mayor de valores como se aprecia en la Figura 4.14). En la Figura 4.41, se presenta que la distribución va subiendo según la edad promedio, siendo esta estable a partir de los 52 años.

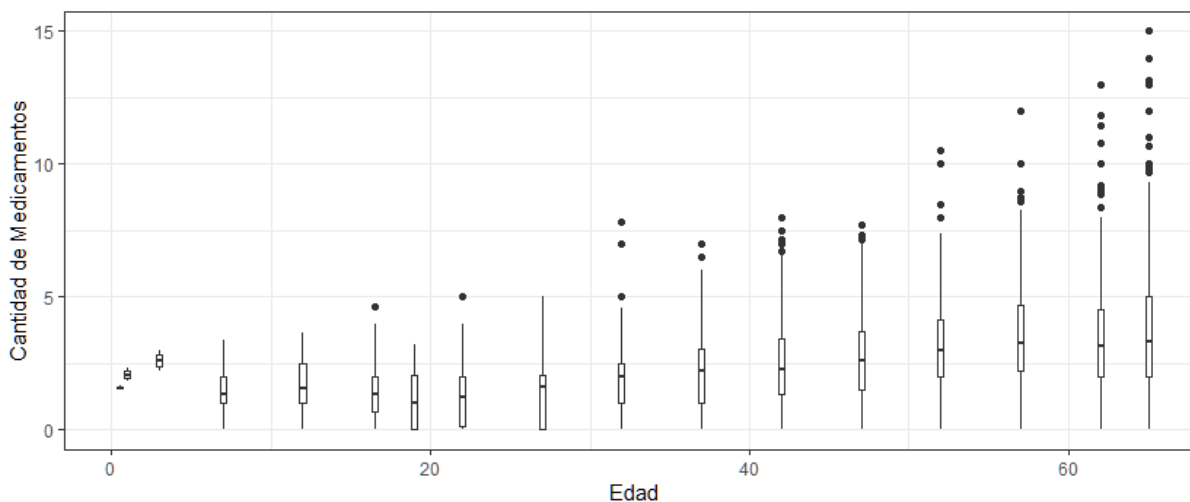


Figura 4.41. Relación entre edad y cantidad de medicamentos

4.4.1.4 Importación de resultados de algoritmos a plataforma

Para el modelado de la plataforma se ideó qué es lo que sería de interés para incluirse. Con los algoritmos desarrollados en la etapa anterior, se identificaron algunos de los puntos que pudiesen asistir al médico general a diagnosticar e identificar a derechohabientes según su evaluación de riesgo de salud que trate de reflejar el riesgo de una calidad de vida menor que el resto de la población por sus comorbilidades, pero que además considere las visitas anteriores, comorbilidades e histórico de visitas.

Cada uno de los apartados que se presentan en la plataforma diseñada se describen a continuación brevemente.

Índice de estado de salud

Con respecto al índice de estado de salud (siendo 0% que aparenta no presentar características de riesgo, mientras que 100% presenta todas las características de riesgo evaluadas), se le agregaron pesos preliminarmente de tres aspectos: (1) grupo de riesgo (obtenido por análisis de conglomerados), (2) frecuencia de visita (comparado con sus visitas históricas) y (3) número de comorbilidades (comorbilidades de Elixhauser).

Frecuencia de visitas en el último mes. Éste es evaluado según los cuartiles de 25% y mediana para identificar si está por arriba de su histórico personal. Los datos que son evaluados son los acumulados de las visitas durante el último mes, mismo que se calcula por cada visita. Un ejemplo se presenta en la Figura 4.5. Se evalúa en tres rangos: cero si es igual o menor del 25% [0%, 25%], 0.5 si está entre el 25% y el 50% (25%, 50%], y 1 si está sobre el 50% (50%, 100%] de visitas históricas acumuladas en un mes y calculada por cada visita.

Por ejemplo, en la siguiente figura se presenta el número de visitas históricas de un derechohabiente. Cada punto representa una visita, mientras que el eje vertical presenta cuántas visitas se han acumulado en los últimos treinta días de cada visita. A inicios del 2016 acumuló cuatro visitas en el mes, siendo éste su máximo histórico. En la distribución de la derecha se presenta la evaluación con respecto a los acumulados históricos que ha presentado, como en la última visita tiene un acumulado de una visita y es igual al 25% de sus visitas acumuladas, este derechohabiente no presentaría riesgo por parte de su comportamiento de visitas. Si tuviera más de dos visitas acumuladas, éste hubiera sido evaluado con el máximo riesgo en este apartado.

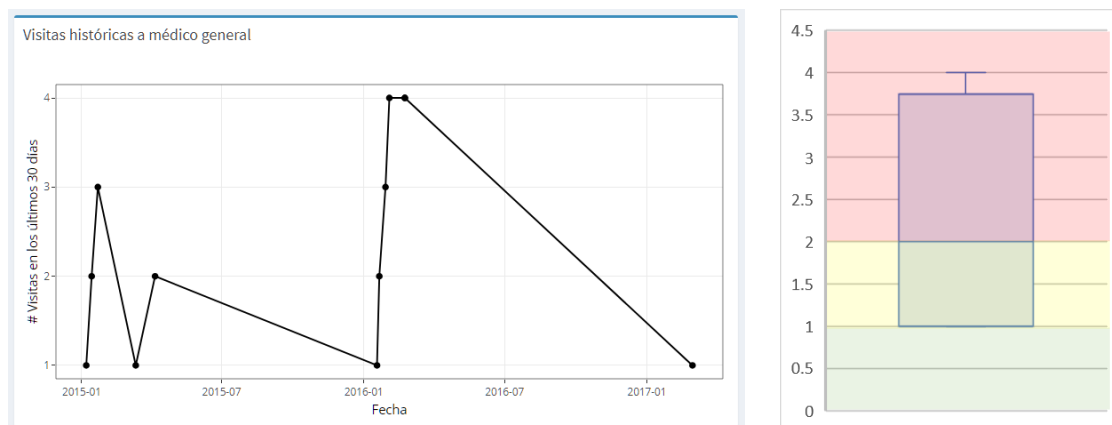


Figura 4.42. Número de visitas en los últimos 30 días y distribución de visitas acumuladas

Número de comorbilidades según el índice de Elixhauser. El número de comorbilidades en los derechohabientes evaluados presentan entre 0 y 4 comorbilidades, mismo que fue normalizado para presentar valores entre 0 y 1. Si

derechohabientes hubiesen presentado más de cuatro comorbilidades, se normalizaría según el número máximo de comorbilidades presentado de la población.

Número de comorbilidades	0	1	2	3	4
Evaluación de riesgo	0	0.25	0.5	0.75	1

Tabla 4.16. Evaluación de riesgo según número de comorbilidades de Elixhauser

Grupo perteneciente según la evaluación de k-means. Considerando la agrupación de derechohabientes según el número de medicinas, número de recetas, sexo, edad y número de comorbilidades; los derechohabientes fueron agrupados donde en el primer grupo se presentó un mayor promedio de comorbilidades según Elixhauser, mayor número de medicinas, mayor número de recetas y mayor edad. Éstos fueron disminuyendo consistentemente en los diferentes grupos. Estos grupos se presentan en la Tabla 4.14. Los grupos fueron normalizados para encontrarse entre el 0 y el 1.

Grupo según k-means	1	2	3
Evaluación de riesgo	1	0.5	0

Tabla 4.17. Evaluación de riesgo según grupo perteneciente del análisis de conglomerados

Cada aspecto de evaluación tiene un peso que puede ser ajustado a partir del uso periódico de la plataforma, mientras se mantenga su suma igual a 1. De esta manera, se presenta el riesgo de los derechohabientes del 0 (menor riesgo) a 100% (máximo riesgo).



Figura 4.43. Aspectos que afectan el índice de estado de salud

Visitas históricas

Se presentó el número de visitas en el último mes con respecto a la última visita, seguido de un histórico de sus visitas presentadas para que el médico pueda visualizar el comportamiento del derechohabiente. En la siguiente figura se puede ver un ejemplo de cómo se visualiza. Cada punto representa una visita y el eje vertical representa el acumulado de visitas del paciente en el último mes. Se puede visualizar que en su última visita acumulaba tres visitas en el último mes, lo cual serviría de alerta para el médico en general.

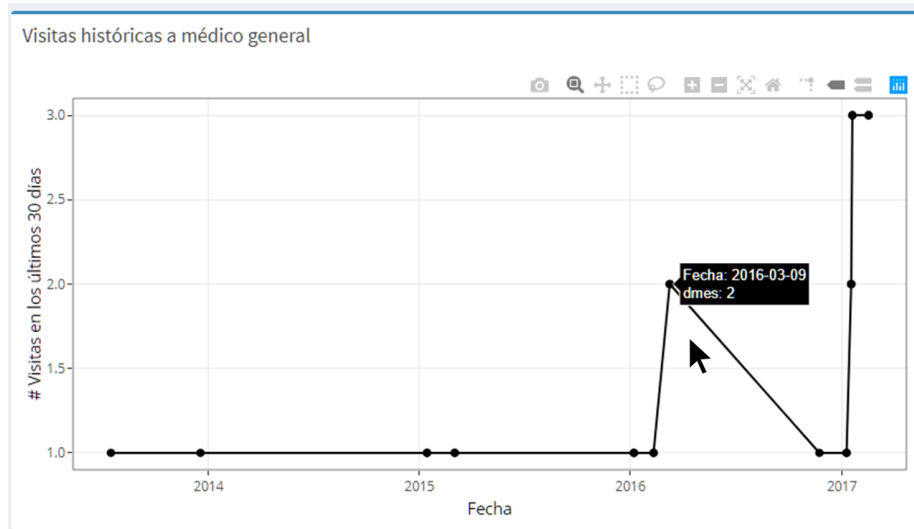


Figura 4.44. Gráfico de visitas históricas de paciente con acumulados de visitas en el último mes

Características de derechohabiente

Se presentan también el número de comorbilidades de Elixhauser (como número y lista de comorbilidades), el número de medicamentos promedio durante las visitas a médico general, número de recetas promedio, número de exámenes, grupo de edad y sexo del derechohabiente. En la siguiente figura se presenta cómo aparecen los datos, siendo numéricos y algunos con color semáforo (siendo estos: comorbilidades, medicamentos, recetas y grupo de edad). Tanto exámenes como sexo presentan colores que estándar. Esto no quiere decir que el sexo no haya sido una característica de riesgo, ya que éste forma parte del análisis de conglomerados para el grupo de riesgo y junto con medicamentos, recetas y comorbilidades localiza a los derechohabientes a un grupo según su riesgo.

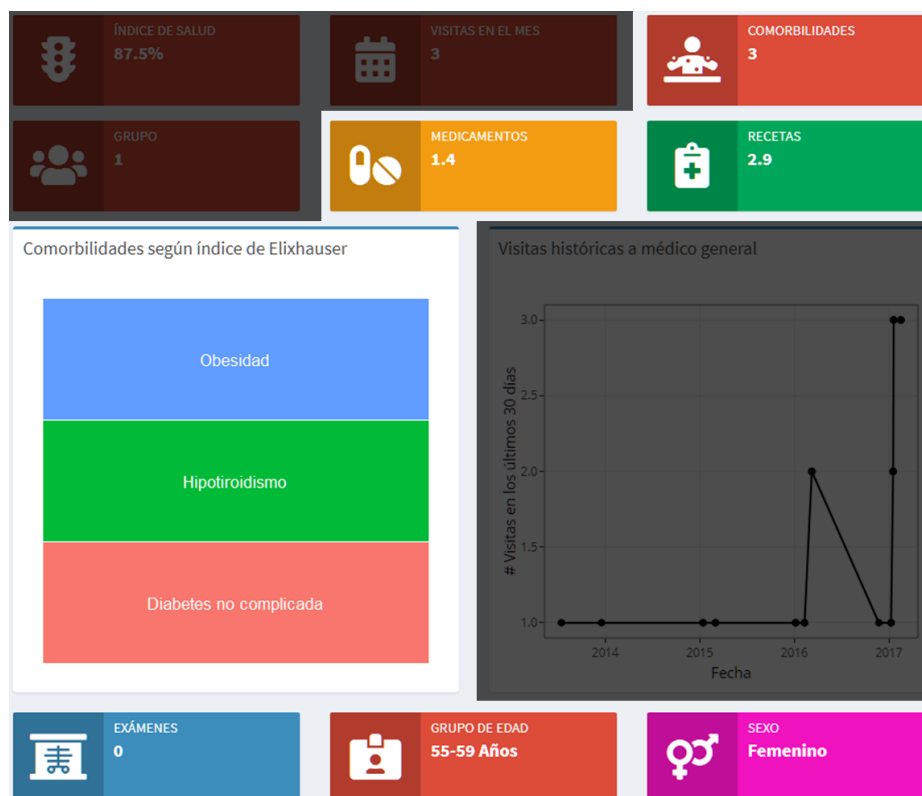


Figura 4.45. Características de evaluación de derechohabiente

Relación de enfermedades según redes bayesianas

Con respecto a la relación entre grupos de categorías de enfermedades presentadas en derechohabientes, se buscó la manera de incluir este en la plataforma. Se diseñó el sistema de mensajes en caso de que el usuario llegue a tener alguna predicción a una enfermedad con respecto a los resultados de redes bayesianas presentada en la Figura 4.30.

En diseño de plataforma se presenta cómo fueron presentados estos datos en la plataforma.

4.4.2 Diseño de plataforma

La plataforma diseñada se realizó en RStudio. Se utilizó RStudio principalmente porque es una plataforma abierta que permite el desarrollo de este tipo de plataformas. Además, se utilizó RStudio para la preparación de los datos y el análisis estadístico

realizado. Con ello, se facilita la adopción por parte de instituciones médicas al no complicar su implementación utilizando diferentes plataformas.

Cada una de las partes se presentan a continuación.

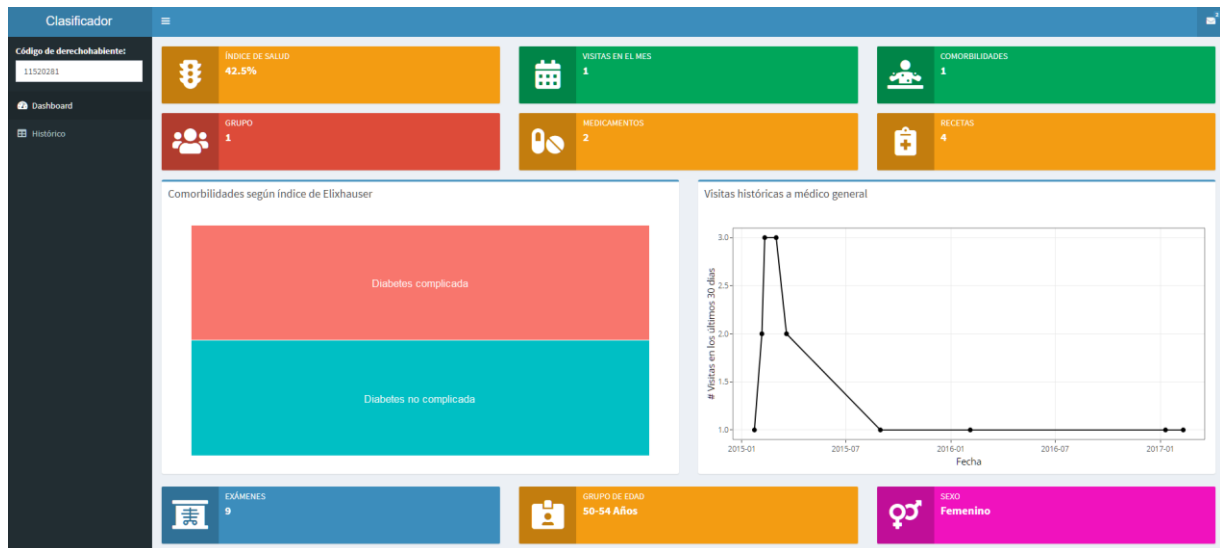


Figura 4.46. Plataforma diseñada para el médico general

Al entrar a la plataforma, se requiere ingresar el número de derechohabiente en la esquina izquierda. Ya ingresado, la plataforma se actualiza automáticamente. En el primer recuadro se presenta el índice de estado de salud en la plataforma que evalúa el grupo de riesgo perteneciente (según k-means), la frecuencia de visitas históricas y el número de comorbilidades según Elixhauser. Este se presenta entre 0 – 100% según el riesgo que presenta el derechohabiente. El color de este recuadro va cambiando según el porcentaje presentado.



Figura 4.47. Índice de estado de salud de derechohabientes entre 0 a 100%

Los otros indicadores que están relacionados con el índice de estado de salud directa o indirectamente se presentan alrededor. Las visitas en el mes consideran la última visita y la compara con el histórico del derechohabiente. El número de comorbilidades se presenta según hayan sido identificadas por el índice de Elixhauser, el grupo se

presenta según se identificó en el grupo por k-means, se presenta además el número de medicamentos y recetas promedio de visitas históricas. Éstas se presentan en la Figura 4.48.

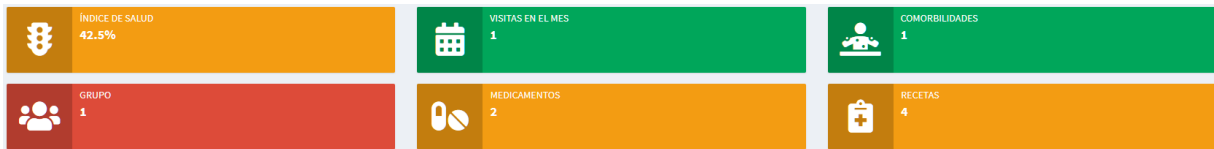


Figura 4.48. Índice de estado de salud y variables numéricas relacionadas

Otras de las variables relacionadas con el índice de salud son el grupo de edad y sexo, éstas se presentan en la parte de abajo donde sólo el grupo de edad presenta color semáforo. Para ayudar al médico general se incluye el número de exámenes realizados históricamente en el derechohabiente (pudiendo ser éstos examen de mama, electrocardiograma o rayos x). Éste se presenta en la Figura 4.49.



Figura 4.49. Variables no numéricas relacionadas con el índice de estado de salud y el número de exámenes

Para ayudar a visualizar el comportamiento de visitas, se presenta una gráfica con cada punto representando cada visita, mientras que el eje Y representa el número de visitas en el último mes (véase Figura 4.50). La gráfica es interactiva, por lo que permite seleccionar, visualizar la fecha de la visita, acercar o alejar.

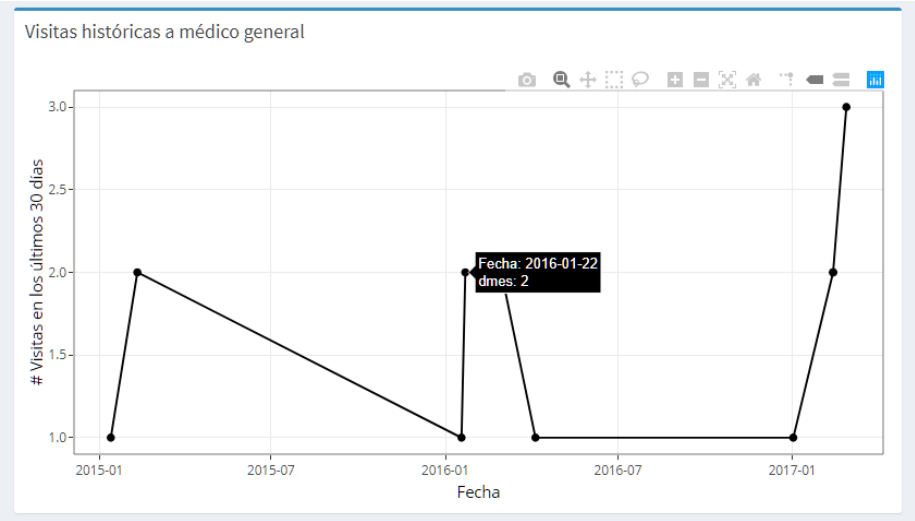


Figura 4.50. Visitas históricas a médico general

Debido a la importancia del índice de comorbilidad para el índice de salud, se presentan las comorbilidades según el índice de Elixhauser identificadas en el derechohabiente como se presenta en la Figura 4.51.

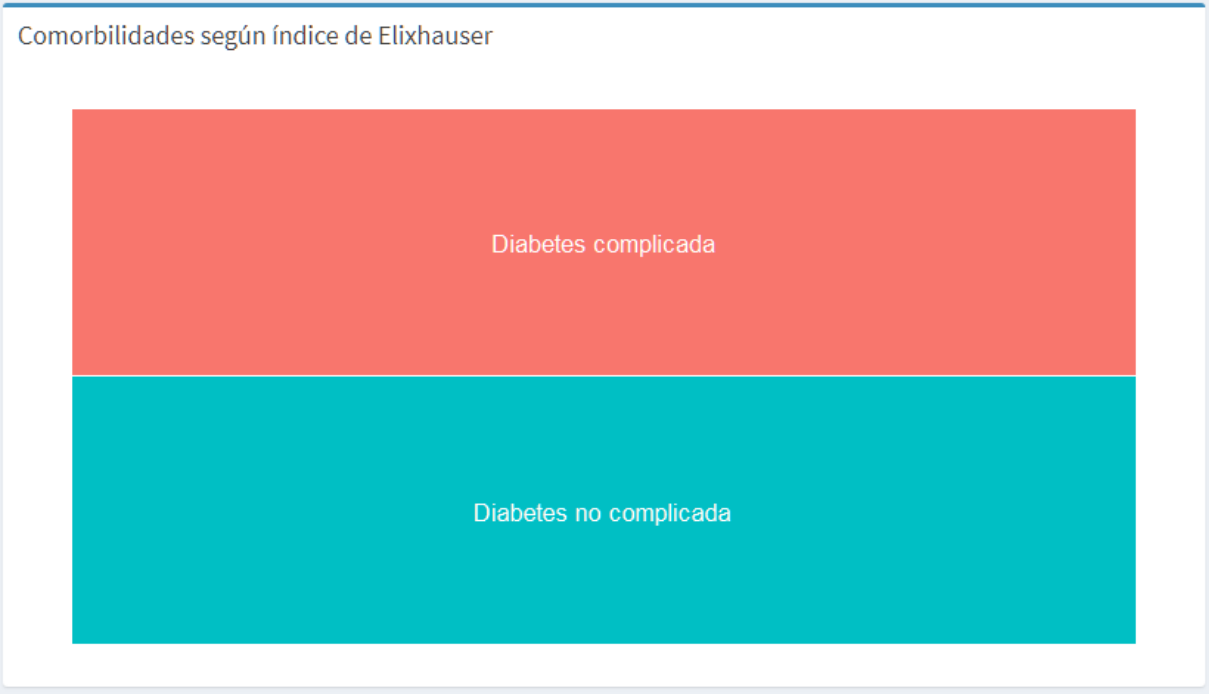


Figura 4.51. Número de comorbilidades identificadas en el derechohabiente según el índice de Elixhauser

Para asistir la visualización de número de medicinas, recetas, exámenes médicos, entre otros, se incluye la tabla del derechohabiente, como se presenta en la Figura 4.52.

	Fecha	ICD10	Receta	Cant.Med	Cant.Rece	Electro	Exam.Mama	Rayosx
1	2015-03-02	FRACTURA DEL PERONE SOLAMENTE	No	0	0	No	No	No
2	2016-02-04	DIABETES MELLITUS NO INSULINODEPENDIENTE	SI	9	4	SI	SI	SI
3	2015-02-05	DIABETES MELLITUS NO INSULINODEPENDIENTE	SI	5	2	No	SI	SI
4	2017-01-10	BRONQUITIS AGUDA	SI	5	3	No	No	No
5	2015-02-10	GASTRITIS, NO ESPECIFICADA	SI	2	1	No	No	SI
6	2017-02-10	DIABETES MELLITUS NO INSULINODEPENDIENTE	SI	7	3	No	SI	No
7	2015-03-20	FRACTURA DEL PERONE SOLAMENTE	SI	3	2	No	No	SI
8	2015-01-23	DIABETES MELLITUS INSULINODEPENDIENTE, CON COMPLICACIONES NEUROLOGICAS	SI	5	3	No	No	SI
9	2015-09-31	DIABETES MELLITUS NO INSULINODEPENDIENTE, SIN MENCION DE COMPLICACION	No	0	0	No	No	No

Figura 4.52. Histórico de visitas de derechohabiente seleccionado

Finalmente, incluyendo los resultados de la red bayesiana de grupos de categorías de padecimientos presentados en los derechohabientes. Se presenta un filtrado de diagnósticos que se predicen y el porcentaje de predicción en la esquina superior derecha de la plataforma, como se presenta en la Figura 4.53.



Figura 4.53. Predicción de enfermedades según red bayesiana con porcentaje de predicción

De esta manera, se visualiza en una página los aspectos más relevantes de su histórico de visitas y permite al médico general identificar qué partes están asistiendo en el índice calculado. En la Figura 4.54, se presenta la plataforma con un derechohabiente con 100% de riesgo según su índice de estado de salud.

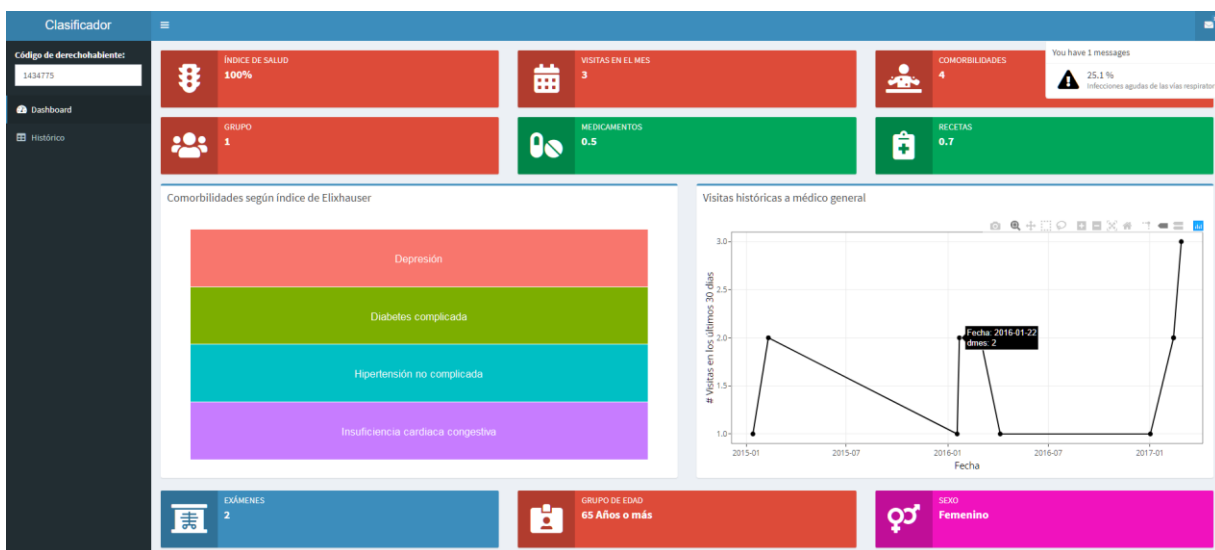


Figura 4.54. Plataforma diseñada con usuario de 100% de índice de salud

4.5. Evaluación de metodología

En este último paso de la implementación, se evalúan los modelos y plataforma según el contexto del problema y se prueba para determinar qué tan satisfactorio es el uso por el usuario. Esto se realiza en dos partes; evaluación de la metodología a partir de los objetivos, hipótesis y preguntas de investigación, mientras que la retroalimentación se enfoca en la satisfacción del usuario final.

4.5.1 Evaluar la metodología

Esta evaluación se realiza con respecto al objetivo general, hipótesis y preguntas de investigación.

4.5.1.1 Objetivos

Elaborar una metodología para la evaluación de riesgo en la salud en derechohabientes del ISSSTESON según consultas a médico general, tomando en cuenta características biométricas, comorbilidades, frecuencias de visitas a médico general y reincidencia, utilizando técnicas de minería de datos, con el fin de realizar consultas más eficientes asistidas con información histórica, y reducir reincidencias;

además de generar grupos de usuarios de acuerdo a las similitudes entre ellos para campañas de medicina preventiva dirigidas.

La plataforma presenta información histórica en forma de *dashboard*, permitiendo que se identifiquen las características críticas del usuario en forma de semáforo y de forma rápida pueda asistir al médico general en caso de que haya una característica crítica de riesgo de salud en el derechohabiente; además, una de las características que analiza la plataforma son las reincidencias por lo que se identifica el usuario que ha realizado un número anormal de visitas (con respecto a su histórico, no según la población), para así evitar que siga en aumento. Finalmente, se realizó una segmentación de población con respecto a varias características biométricas que podrían utilizarse para campañas de medicina preventiva. Con esto, se considera como satisfactoria la plataforma para cumplir este objetivo.

4.5.1.2 Hipótesis

La evaluación de riesgo de nivel de salud de los derechohabientes de ISSSTESON podría ayudar a identificar usuarios que requieren mayor atención durante su consulta y asistir en el diagnóstico para prevenir reincidencias, así como crear grupos de pacientes para la mejora en campañas de medicina preventiva dirigida.

La plataforma permite identificar usuarios según sus características biométricas con respecto a la presencia de padecimientos que afecten su calidad de vida, revisa su histórico de visitas para identificar tendencias anormales de visitas en el último mes, identifica si sus padecimientos históricos podrían causar otro padecimiento (según registros de pacientes), entre otros, utilizando una visualización rápida de una sola página y colores tipo semáforo para permitir que el médico general se enfoque en aquellos aspectos críticos del paciente. La plataforma habilita visualizar rápidamente el histórico evaluando automáticamente varios aspectos del derechohabiente, mientras que los modelos utilizados en la metodología permiten segmentar grupos de pacientes.

4.5.1.3 Preguntas de investigación

- *¿Cuáles son las causas de los diagnósticos de derechohabientes de ISSSTESON de acuerdo con historial clínico?*

Se revisaron los padecimientos y frecuencia utilizando código internacional de enfermedades (IDC-10) para ambos grupos de derechohabientes, revisando las causas y posibles relaciones entre los padecimientos presentados.

- *¿Cuáles son las técnicas o metodologías de procesamiento y análisis de datos que pudiesen asistir en el análisis de datos de consultas recabadas?*

Se identificaron algunos modelos como redes bayesianas para analizar relación entre padecimientos y características biométricas, clasificación (como k-means) para segmentar derechohabientes y otros métodos (lineales, no lineales y conjuntos de árboles) para comprender un padecimiento específico, encontrando a la regresión logística como un método eficiente para explicar padecimientos.

- *¿Qué enfermedades o características presentan un riesgo mayor de reincidencia en consultas posteriores menores a 30 días?*

Se identificó durante el entendimiento de datos qué padecimientos presentaban mayor probabilidad de que el paciente visitara en menos de 30 días, con medianas históricas menores a 15 días.

- *¿Qué manera sería la más sencilla para comunicar la información analizada con médicos generales para que sea accesible y fácilmente adoptable?*

La manera más sencilla para que sea accesible y entendible por el médico general fue una plataforma *dashboard* donde se presente la mayoría de la información en una sola página con semáforos para asistir a identificar qué características de derechohabientes son más críticas para considerar durante la visita.

4.5.2 Obtener retroalimentación

La retroalimentación fue realizada en presentaciones periódicas con el comité de ética y en reuniones con médico de la institución que pudiese identificar según su experiencia qué características le parecían más interesantes y cómo pudiese cambiar para ser más práctico para su uso diario y adopción en la institución.

5. CONCLUSIONES, RECOMENDACIONES Y TRABAJOS FUTUROS

Durante el desarrollo del proyecto, se documentaron conclusiones, recomendaciones y trabajos futuros relacionados a distintos puntos del desarrollo. Éstas se presentan a continuación.

5.1. Conclusiones

La implementación de herramientas de inteligencia artificial en datos de visitas y características biométricas de los pacientes se ha aplicado anteriormente en la institución con otros enfoques y ha sido tendencia en el sector salud internacionalmente.

Se han identificado varias variables que se ven relacionadas con los padecimientos que presenta el derechohabiente identificados por el índice de Elixhauser. Con ello, aumenta la confiabilidad de utilizar este índice para categorizar a los derechohabientes según su índice de estado de salud utilizándolo como principal identificador.

Otra de las variables de interés para clasificar al derechohabiente es identificar su comportamiento de visitas y clasificarlo según su histórico de visitas, en vez de con respecto a los demás derechohabientes. Esto se realizó ya que algunos pacientes podrían ir mensualmente de forma normal para darle seguimiento a su padecimiento (pudiendo ser diabetes), y no necesariamente por ello se debería de considerar de mayor riesgo. Por ello, evaluar las visitas históricas del paciente con respecto a su comportamiento histórico para identificar anomalías previene que se presente falsamente alerta de riesgo por alta reincidencia de visitas.

De esta manera, el médico general puede verificar en un vistazo las características relevantes del derechohabiente y éste a su vez asistir en el diagnóstico del derechohabiente considerando sus padecimientos anteriores más relevantes en su calidad de vida. Con ello, no sólo se espera que se genere información valiosa para

los médicos generales que podrán utilizarlos durante la atención diaria, pero a su vez la información histórica podrá aportar relaciones interesantes y valiosas para el departamento de medicina preventiva.

5.2. Recomendaciones

La metodología presentada no fue desarrollada exclusivamente para derechohabientes con trastorno de mama y/o diabetes, sino que esto se desarrolló alrededor de estos grupos de derechohabientes debido al interés de la institución de analizar estos grupos. Aún así, esta metodología se desarrolló para atender a derechohabientes independientemente de sus padecimientos; por lo que implementarlo en derechohabientes con padecimientos más heterogéneos permitiría obtener resultados que sean representativos para la población.

Se recomienda que, si se va a implementar en una institución, se considere los cambios en la infraestructura que implicaría que los médicos generales tuviesen acceso local con la plataforma. Además, programar los algoritmos que se actualicen cada cierto tiempo para permitir que continúen representando a la población, y cambiando junto con ella. Esto sería principalmente importante dado que, en cambios radicales en la población, como en la presencia de una pandemia, posiblemente los datos históricos dejen de representar a la población con la misma confianza y afectaría la confiabilidad de la plataforma.

5.3. Trabajos futuros

Como trabajo futuro se recomienda que se implemente esta metodología en datos más recientes u otra institución de salud para enriquecer los resultados obtenidos. Si bien la metodología puede ser la misma, otros algoritmos podrían ser más eficientes en predecir y comprender variables en el caso de que se analicen grupos de derechohabientes heterogéneos y representativos a la población; por lo que la mejora continua de este estudio siempre es posible y, como se mencionó en recomendaciones, necesaria para que la plataforma siga siendo confiable.

Debido a la actual situación de salud debido a la pandemia, la propuesta metodológica que se presenta en esta investigación podría detectar características de riesgo y comorbilidades en derechohabientes que presenten complicaciones al contraer COVID-19. Con ello, se pueden dirigir estrategias de campañas por el departamento de medicina preventiva en pacientes de riesgo.

6. REFERENCIAS

Amat, J., 2016. Análisis discriminante lineal (LDA) y Análisis discriminante cuadrático (QDA). *RPubs*. Disponible en: https://rpubs.com/Joaquin_AR/233932 [Consultado enero 28, 2020].

Aronin, T., 2017. Impact of comorbidity and multimorbidity on primary care practice. *BioMed Central Ltd.* Disponible en: <https://www.biomedcentral.com/collections/MultimorbidityInPrimaryCare> [Consultado septiembre 11, 2019].

Berry, J.G. et al., 2018. Age trends in 30 day hospital readmissions: US national retrospective analysis. *BMJ*, p.k497.

Brown, M.S., 2019a. Phase 3 of the CRISP-DM Process Model: Data Preparation. *Dummies*. Disponible en: <https://www.dummies.com/programming/big-data/phase-3-of-the-crisp-dm-process-model-data-preparation/> [Consultado octubre 22, 2019].

Brown, M.S., 2019b. Phase 4 of the CRISP-DM Process Model: Modeling. *Dummies*. Disponible en: <https://www.dummies.com/programming/big-data/phase-4-of-the-crisp-dm-process-model-modeling/> [Consultado octubre 22, 2019].

Brown, M.S., 2019c. Phase 5 of the CRISP-DM Process Model: Evaluation. *Dummies*. Disponible en: <https://www.dummies.com/programming/big-data/phase-5-of-the-crisp-dm-process-model-evaluation/> [Consultado octubre 22, 2019].

Celis, F., 2018. Inteligencia artificial y realidad virtual, oportunidades para la salud • Forbes México. *Forbes México*. Disponible en: <https://www.forbes.com.mx/inteligencia-artificial-y-realidad-virtual-oportunidades-para-la-salud/> [Consultado septiembre 6, 2019].

CEMECE, 2014. Metodología para la elaboración y validación de GPC - Utilización de la CIE -10 y CIE-9-MC. Disponible en: http://www.cenetec.salud.gob.mx/descargas/gpc/presentaciones_2014/Taller_de_clasificacixn_de_enfermedades_e_intervenciones._CIE-10_y_CIE-9MC._CEMECE._Instr._Rodolfo_Alanxs_Fuentes.pdf [Consultado octubre 27, 2019].

Charles, D., Gabriel, M. y Searcy, T., 2015. Adoption of Electronic Health Record Systems among U.S. Non-Federal Acute Care Hospitals: 2008-2014. , p.10.

Cid, C., Báscolo, E. y Morales, C., 2017. Eficiencia para la salud universal. Disponible en: <https://www.paho.org/hq/dmdocuments/2017/ppt-eficiencia-salud-universal-cid-feb-2017.pdf> [Consultado septiembre 5, 2019].

Cowie, M.R. et al., 2017. Electronic health records to facilitate clinical research. *Clinical Research in Cardiology*, 106(1), pp.1–9.

Dalkir, K., 2017. *Knowledge Management in Theory and Practice* 3a ed., The MIT Press.

Daskivich, T.J. et al., 2018. Electronic Health Record Problem Lists: Accurate Enough for Risk Adjustment? *The American Journal of Managed Care*, 24(1), p.6.

Elixhauser, A. et al., 1998. Comorbidity Measures for Use with Administrative Data. *Medical Care*, 36(1), pp.8–27.

EPA, 2018. Health Status | EPA's Report on the Environment (ROE) | US EPA. *United States Environment Protection Agency*. Disponible en: <https://www.epa.gov/report-environment/health-status> [Consultado marzo 9, 2019].

Fiuza-Pérez, M.D. y Rodríguez-Pérez, J.C., 2000. La regresión logística: una herramienta versátil. *Nefrología*, 20(6), pp.495–500.

Forsyth, D., 2018. *Probability and Statistics for Computer Science*, Cham: Springer International Publishing. Disponible en: <http://link.springer.com/10.1007/978-3-319-64410-3> [Consultado abril 29, 2020].

Futoma, J., Morris, J. y Lucas, J., 2015. A comparison of models for predicting early hospital readmissions - ScienceDirect. *Journal of Biomedical Informatics*, 56, pp.229–238.

Gallo, A., 2015. A Refresher on Regression Analysis. *Harvard Business Review*. Disponible en: <https://hbr.org/2015/11/a-refresher-on-regression-analysis> [Consultado septiembre 16, 2019].

Garza, C. y Ramírez, M., 2017. El Impacto Económico De La Comorbilidad En Un Sistema De Salud Privado En México. *Value in Health*, 20(9), p.A904.

Gasparini, A., 2019. Comorbidity scores. *The Comprehensive R Archive Network*. Disponible en: <https://cran.r-project.org/web/packages/comorbidity/vignettes/comorbiditiescores.html>.

Gobierno del Estado de Sonora, 2019. Anuario Estadístico 2018. *Instituto de Seguridad y Servicios Sociales de los Trabajadores del Estado de Sonora*. Disponible en: <http://www.isssteson.gob.mx/images/subdirecciones/medica/AnuarioEstadistico2018.pdf> [Consultado enero 30, 2019].

Goldstein, B.A. et al., 2017. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *Journal of the American Medical Informatics Association*, 24(1), pp.198–208.

- Grosser, Z., 2019. Accessible Colors for Data Visualization. *Medium*. Disponible en: <https://medium.com/@zachgrosser/accessible-colors-for-data-visualization-2ad64ac4ee7e> [Consultado octubre 27, 2019].
- Guo, W., 2017. Team data science process IDEAR instructions. *GitHub*. Disponible en: <https://github.com/Azure/Azure-TDSP-Utilities> [Consultado octubre 22, 2019].
- Härdle, W.K. y Simar, L., 2015. *Applied Multivariate Statistical Analysis*, Berlin, Heidelberg: Springer Berlin Heidelberg. Disponible en: <http://link.springer.com/10.1007/978-3-662-45171-7> [Consultado abril 29, 2020].
- Hartigan, J.A. y Wong, M.A., 1979. Algorithm AS 136: A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1), pp.100–108.
- Hastie, T. y Qian, J., 2014. Glmnet Vignette. *Stanford University*. Disponible en: https://web.stanford.edu/~hastie/glmnet/glmnet_alpha.html [Consultado enero 28, 2020].
- Heiberger, R.M. y Holland, B., 2015. *Statistical Analysis and Data Display*, New York, NY: Springer New York. Disponible en: <http://link.springer.com/10.1007/978-1-4939-2122-5> [Consultado abril 29, 2020].
- Hernández, R., 2014. *Metodología de la Investigación* 6a ed., México D.F.: McGraw Hill.
- ISSSTESON, 2019. Estado de Situación Financiera ISSSTESON. *Instituto de Seguridad y Servicios Sociales de los Trabajadores del Estado de Sonora*. Disponible en: <http://www.isssteson.gob.mx/images/contabilidadgubernamental/646a3d7f.pdf> [Consultado enero 30, 2019].
- ISSSTESON, 2017. Programa Institucional de Desarrollo ISSSTESON 2016-2021. Disponible en: <http://www.isssteson.gob.mx/index.php/informeanual> [Consultado enero 30, 2019].
- Jaiswal, S., 2018. K-Means Clustering in R Tutorial. *DataCamp Community*. Disponible en: <https://www.datacamp.com/community/tutorials/k-means-clustering-r> [Consultado septiembre 16, 2019].
- Kanakubo, T. y Kharrazi, H., 2019. Comparing the Trends of Electronic Health Record Adoption Among Hospitals of the United States and Japan. *Journal of Medical Systems*, 43(7), p.224.
- Kaplan, R.M., Bush, J.W. y Berry, C.C., 1976. Health status: types of validity and the index of well-being. *Health Serv Research*, 11(4), pp.478–507.

Kim, E. et al., 2019. The Evolving Use of Electronic Health Records (EHR) for Research. *Seminars in Radiation Oncology*, 29(4), pp.354–361.

Kuhn, M., 2019. *caret: Classification and regression training*, Disponible en: <https://CRAN.R-project.org/package=caret>.

Le, J., 2018. Support Vector Machines in R. *DataCamp Community*. Disponible en: <https://www.datacamp.com/community/tutorials/support-vector-machines-r> [Consultado enero 28, 2020].

Lin, P., 2017. The Data Science Process. Disponible en: [https://www-01.ibm.com/events/wwc/grp/grp304.nsf/vLookupPDFs/Polong%20Lin%20Presentation/\\$file/Polong%20Lin%20Presentation.pdf](https://www-01.ibm.com/events/wwc/grp/grp304.nsf/vLookupPDFs/Polong%20Lin%20Presentation/$file/Polong%20Lin%20Presentation.pdf) [Consultado octubre 21, 2019].

Lix, L. et al., 2016. *Cancer data linkage in Manitoba: expanding the infrastructure for research* 1a ed., Manitoba, Canada: Manitoba Centre for Health Policy.

Llanes, M.L., 2018. *Una metodología para la detección de brotes epidemiológicos de rápida propagación utilizando minería de datos y análisis de georreferenciación*. Hermosillo, Sonora, México: Universidad de Sonora.

Maklin, C., 2019. Linear Discriminant Analysis In Python. *Medium*. Disponible en: <https://towardsdatascience.com/linear-discriminant-analysis-in-python-76b8b17817c2> [Consultado abril 29, 2020].

Malik, M.M., Abdallah, S. y Ala'raj, M., 2018. Data mining and predictive analytics applications for the delivery of healthcare services: a systematic literature review. *Annals of Operations Research*, 270(1), pp.287–312.

Martens, J. et al., 2017. Business understanding in Team Data Science Process. Disponible en: <https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/lifecycle-business-understanding> [Consultado octubre 22, 2019].

MathWarehouse, 2019. What is Anscombe's Quarter in Statistics. *Interactive math activities, demonstrations, lessons with definitions and examples, worksheets, interactive activities and other resources*. Disponible en: <https://www.mathwarehouse.com/statistics/what-is-anscombes-quartet.php> [Consultado octubre 22, 2019].

MDCalc, 2019. Charlson Comorbidity Index (CCI). *MDCalc*. Disponible en: <https://www.mdcalc.com/charlson-comorbidity-index-cci> [Consultado septiembre 11, 2019].

Ness, R., 2019. *Lecture Notes for Causality in Machine Learning*, Disponible en: <https://bookdown.org/robertness/causalml/docs/tutorial-probabilistic-modeling-with-bayesian-networks-and-bnlearn.html> [Consultado diciembre 4, 2019].

Nisbet, R. et al., 2018. *Handbook of statistical analysis and data mining applications* Second edition., London: Academic Press.

OECD, 2016. *Estudios de la OCDE sobre los sistemas de salud: México*, OECD. Disponible en: https://www.oecd-ilibrary.org/social-issues-migration-health/oecd-reviews-of-health-systems-mexico-2016_9789264230491-en [Consultado septiembre 16, 2019].

OECD, 2017. Health Status. *Organisation for Economic Co-operation and Development*. Disponible en: <https://stats.oecd.org/index.aspx?queryid=24879#> [Consultado marzo 9, 2019].

OPM y OMS, 2015. Estrategia de cooperación de la organización panamericana de la salud organización mundial de la salud con México 2015-2018. *Pan American Health Organization / World Health Organization*. Disponible en: https://www.paho.org/mex/index.php?option=com_docman&view=download&category_slug=technical-documentation&alias=1053-ccs-mex-1518-web0&Itemid=493 [Consultado septiembre 16, 2019].

Park, K.M. et al., 2019. Population-based dementia prediction model using Korean public health examination data: A cohort study P. Werner, ed. *PLOS ONE*, 14(2), p.e0211957.

Patience, G., 2018. Applying the CRISP-DM Method to a Business Problem. *Collection of Resources useful for Data Science projects*. Disponible en: https://github.com/patiegm/Datasci_Resources [Consultado octubre 27, 2019].

Pijpers, E. et al., 2012. The frailty dilemma. Review of the predictive accuracy of major frailty scores. *European Journal of Internal Medicine*, 23(2), pp.118–123.

PubMed, 2019. ehr - PubMed - NCBI. *PubMed.gov*. Disponible en: <https://www.ncbi.nlm.nih.gov/pubmed/?term=ehr> [Consultado septiembre 6, 2019].

R Core Team, 2019. R: A language and environment for statistical computing. *R Foundation for Statistical Computing*. Disponible en: <https://www.r-project.org/> [Consultado noviembre 13, 2019].

Reid, R.J. et al., 2001. Performance of the ACG case-mix system in two Canadian provinces. *Medical Care*, 39(1), pp.86–99.

Rockwood, K. et al., 2005. A global clinical measure of fitness and frailty in elderly people. *CMAJ: Canadian Medical Association Journal*, 173(5), pp.489–495.

Rosas, O. et al., 2017. Evaluación de la comorbilidad en el adulto mayor. *Revista Médica del Instituto Mexicano del Seguro Social*, 49(2), pp.153–162.

RStudio Team, 2019. *RStudio: Integrated Development for R*, RStudio, Inc. Disponible en: <https://rstudio.com/>.

Safran, C. et al., 2007. Toward a National Framework for the Secondary Use of Health Data: An American Medical Informatics Association White Paper. *Journal of the American Medical Informatics Association*, 14(1), pp.1–9.

Sanez, E.A., 2018. *Minería de datos para una estrategia de medicina preventiva más robusta en una institución de salud pública del estado de Sonora*. Hermosillo, Sonora, México: Universidad de Sonora.

Sarfati, D., Koczwara, B. y Jackson, C., 2016. The impact of comorbidity on cancer and its treatment. *CA: A Cancer Journal for Clinicians*, 66(4), pp.337–350.

Schoen, C. et al., 2012. A Survey Of Primary Care Doctors In Ten Countries Shows Progress In Use Of Health Information Technology, Less In Other Areas. *Health Affairs*, 31(12), pp.2805–2816.

Scutari, M., 2017. bnlearn - Bayesian networks and cross-validation. *bnlearn - an R package for Bayesian network learning and inference*. Disponible en: <http://www.bnlearn.com/examples/xval/> [Consultado diciembre 4, 2019].

Scutari, M., 2010. Learning bayesian networks with the bnlearn R package. *Journal of Statistical Software*, 35(3), pp.1–22.

Scutari, M., 2020. Structure learning algorithms. *bnlearn - an R package for Bayesian network learning and inference*. Disponible en: <https://www.bnlearn.com/documentation/man/structure.learning.html> [Consultado enero 17, 2020].

Secretaría de Salud, 2015. *La calidad de la atención a la salud en México a través de sus instituciones* 2a ed., México D.F.: Biblioteca Mexicana del Conocimiento. Disponible en: https://www.gob.mx/cms/uploads/attachment/file/60110/libro_03.pdf [Consultado septiembre 6, 2019].

Secretaría de Salud, 2017. *Modelo de gestión de calidad en salud*, México D.F.

Seif, G., 2019. The 5 Clustering Algorithms Data Scientists Need to Know. *Medium*. Disponible en: <https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68> [Consultado octubre 22, 2019].

Sucar, L.E., 2006. Redes bayesianas. Disponible en: <https://ccc.inaoep.mx/~esucar/Clases-mgp/caprb.pdf> [Consultado noviembre 20, 2019].

Teredesai, A. et al., 2018. Explainable Models for Healthcare AI. *ACM Learning Webinar*, p.48.

The Commonwealth Fund, 2015. 2015 International Survey of Primary Care Physicians: International Health Care System Profiles. *The Commonwealth Fund*. Disponible en: <https://international.commonwealthfund.org/data/2015/> [Consultado septiembre 11, 2019].

Trifiro, G. et al., 2009. The EU-ADR project: preliminary results and perspective. *Studies in Health Technology and Informatics*, 148, pp.43–49.

University of Leicester, 2019. 3. Methods of collecting qualitative data. *University of Leicester*. Disponible en: https://www.le.ac.uk/oerresources/lill/fdmvco/module9/page_54.htm [Consultado octubre 28, 2019].

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

Witten, I., Frank, M. y Hall, M.A., 2011. *Data Mining* 3a ed., Morgan Kaufmann. Disponible en: <https://www.amazon.com/Data-Mining-Practical-Techniques-Management/dp/0123748569>.

Yadav, P. et al., 2018. Mining Electronic Health Records (EHRs): A Survey. *ACM Computing Surveys*, 50(6), pp.1–40.

Yip, W. y Hafez, R., 2015. Reforms for improving the efficiency of health systems: lessons from 10 country cases. *Health Systems Governance & Financing*, p.35.

7. ANEXOS

7.1. Comorbilidad de Charlson

Enfermedad	Código ICD-9-CM	Código ICD-10
Infarto de miocardio	410.x, 412.x	I21.x, I22.x, I25.2
Insuficiencia cardíaca congestiva	398.91, 402.01, 402.11, 402.91, 404.01, 404.03, 404.11, 404.13, 404.91, 404.93, 425.4 - 425.9, 428.x	I09.9, I11.0, I13.0, I13.2, I25.5, I42.0, I42.5 - I42.9, I43.x, I50.x, P29.0
Enfermedad vascular periférica	093.0, 437.3, 440.x, 441.x, 443.1 - 443.9, 47.1, 557.1, 557.9, V43.4	I70.x, I71.x, I73.1, I73.8, I73.9, I77.1, I79.0, I79.2, K55.1, K55.8, K55.9, Z95.8, Z95.9
Enfermedad cerebrovascular	362.34, 430.x - 438.x	G45.x, G46.x, H34.0, I60.x - I69.x
Demencia	290.x, 294.1, 331.2	F00.x - F03.x, F05.1, G30.x, G31.1
Enfermedad pulmonar crónica	416.8, 416.9, 490.x - 505.x, 506.4, 508.1, 508.8	I27.8, I27.9, J40.x - J47.x, J60.x - J67.x, J68.4, J70.1, J70.3
Enfermedad reumática	446.5, 710.0 - 710.4, 714.0 - 714.2, 714.8, 725.x	M05.x, M06.x, M31.5, M32.x - M34.x, M35.1, M35.3, M36.0
La enfermedad de úlcera péptica	531.x - 534.x	K25.x - K28.x
Enfermedad hepática leve	070.22, 070.23, 070.32, 070.33, 070.44, 070.54, 070.6, 070.9, 570.x, 571.x, 573.3, 573.4, 573.8, 573.9, V42.7	B18.x, K70.0 - K70.3, K70.9, K71.3 - K71.5, K71.7, K73.x, K74.x, K76.0, K76.2 - K76.4, K76.8, K76.9, Z94.4
Diabetes sin complicaciones crónicas	250.0 - 250.3, 250.8, 250.9	E10.0, E10.1, E10.6, E10.8, E10.9, E11.0, E11.1, E11.6, E11.8, E11.9, E12.0, E12.1, E12.6, E12.8, E12.9, E13.0, E13.1, E13.6, E13.8, E13.9, E14.0, E14.1, E14.6, E14.8, E14.9
Diabetes con complicación crónica	250.4 - 250.7	E10.2 - E10.5, E10.7, E11.2 - E11.5, E11.7, E12.2 - E12.5, E12.7, E13.2 - E13.5, E13.7, E14.2 - E14.5, E14.7
Hemiplejía o paraplejía	334.1, 342.x, 343.x, 344.0 - 344.6, 344.9	G04.1, G11.4, G80.1, G80.2, G81.x, G82.x, G83.0 - G83.4, G83.9
Enfermedad renal	403.01, 403.11, 403.91, 404.02, 404.03, 404.12, 404.13, 404.92, 404.93, 582.x, 583.0 - 583.7, 585.x, 586.x, 588.0, V42.0, V45.1, V56.x	I12.0, I13.1, N03.2 - N03.7, N05.2 - N05.7, N18.x, N19.x, N25.0, Z49.0 - Z49.2, Z94.0, Z99.2
Cualquier tumor maligno, incluido el linfoma y la leucemia, excepto la neoplasia maligna de la piel	140.x - 172.x, 174.x - 195.8, 200.x - 208.x, 238.6	C00.x - C26.x, C30.x - C34.x, C37.x - C41.x, C43.x, C45.x - C58.x, C60.x - C76.x, C81.x - C85.x, C88.x, C90.x - C97.x
Enfermedad hepática moderada o grave	456.0 - 456.2, 572.2 - 572.8	I85.0, I85.9, I86.4, I98.2, K70.4, K71.1, K72.1, K72.9, K76.5, K76.6, K76.7
Tumor sólido metastásico	196.x - 199.x	C77.x - C80.x
SIDA / VIH	042.x - 044.x	B20.x - B22.x, B24.x

Tabla 7.1. Códigos de enfermedades para calcular el índice de comorbilidad de Charlson (Gasparini, 2019)

7.2. Comorbilidad de Elixhauser

Enfermedad	Código ICD-9-CM	Código ICD-10
Insuficiencia cardíaca congestiva	398.91, 402.01, 402.11, 402.91, 404.01, 404.03, 404.11, 404.13, 404.91, 404.93, 425.4 - 425.9, 428.x	I09.9, I11.0, I13.0, I13.2, I25.5, I42.0, I42.5 - I42.9, I43.x, I50.x, P29.0
Arritmia cardíaca	426.0, 426.13, 426.7, 426.9, 426.10, 426.12, 427.0 - 427.4, 427.6 - 427.9, 785.0, 996.01, 996.04, V45.0, V53.3	I44.1 - I44.3, I45.6, I45.9, I47.x - I49.x, R00.0, R00.1, R00.8, T82.1, Z45.0, Z95.0
Enfermedad valvular	093.2, 394.x - 397.x, 424.x, 746.3 - 746.6, V42.2, V43.3	A52.0, I05.x - I08.x, I09.1, I09.8, I34.x - I39.x, Q23.0 - Q23.3, Z95.2 - Z95.4
Trastornos de la circulación pulmonar.	415.0, 415.1, 416.x, 417.0, 417.8, 417.9	I26.x, I27.x, I28.0, I28.8, I28.9
Trastornos vasculares periféricos.	093.0, 437.3, 440.x, 441.x, 443.1 - 443.9, 447.1, 557.1, 557.9, V43.4	I70.x, I71.x, I73.1, I73.8, I73.9, I77.1, I79.0, I79.2, K55.1, K55.8, K55.9, Z95.8, Z95.9
Hipertensión, sin complicaciones	401.x	I10.x
Hipertensión complicada	402.x - 405.x	I11.x - I13.x, I15.x
Parálisis	334.1, 342.x, 343.x, 344.0 - 344.6, 344.9	G04.1, G11.4, G80.1, G80.2, G81.x, G82.x, G83.0 - G83.4, G83.9
Otros trastornos neurológicos.	331.9, 332.0, 332.1, 333.4, 333.5, 333.92, 334.x - 335.x, 336.2, 340.x, 341.x, 345.x, 348.1, 348.3, 780.3, 784.3	G10.x - G13.x, G20.x - G22.x, G25.4, G25.5, G31.2, G31.8, G31.9, G32.x, G35.x - G37.x, G40.x, G41.x, G93.1, G93.4, R47.0, R56.x
Enfermedad pulmonar crónica	416.8, 416.9, 490.x - 505.x, 506.4, 508.1, 508.8	I27.8, I27.9, J40.x - J47.x, J60.x - J67.x, J68.4, J70.1, J70.3
Diabetes no complicada	250.0 - 250.3	E10.0, E10.1, E10.9, E11.0, E11.1, E11.9, E12.0, E12.1, E12.9, E13.0, E13.1, E13.9, E14.0, E14.1, E14.9
Diabetes complicada	250.4 - 250.9	E10.2 - E10.8, E11.2 - E11.8, E12.2 - E12.8, E13.2 - E13.8, E14.2 - E14.8
Hipotiroidismo	240.9, 243.x, 244.x, 246.1, 246.8	E00.x - E03.x, E89.0
Insuficiencia renal	403.01, 403.11, 403.91, 404.02, 404.03, 404.12, 404.13, 404.92, 404.93, 585.x, 586.x, 588.0, V42.0, V45.1, V56.x	I12.0, I13.1, N18.x, N19.x, N25.0, Z49.0 - Z49.2, Z94.0, Z99.2
Enfermedad del hígado	070.22, 070.23, 070.32, 070.33, 070.44, 070.54, 070.6, 070.9, 456.0 - 456.2, 570.x, 571.x, 572.2 - 572.8, 573.3, 573.4, 573.8, 573.9, V42.7	B18.x, I85.x, I86.4, I98.2, K70.x, K71.1, K71.3 - K71.5, K71.7, K72.x - K74.x, K76.0, K76.2 - K76.9, Z94.4
Úlcera péptica, excluyendo sangrado	531.7, 531.9, 532.7, 532.9, 533.7, 533.9, 534.7, 534.9	K25.7, K25.9, K26.7, K26.9, K27.7, K27.9, K28.7, K28.9
SIDA / VIH	042.x - 044.x	B20.x - B22.x, B24.x
Linfoma	200.x - 202.x, 203.0, 238.6	C81.x - C85.x, C88.x, C96.x, C90.0, C90.2
Cáncer metastásico	196.x - 199.x	C77.x - C80.x
Tumor sólido sin metástasis	140.x - 172.x, 174.x - 195.x	C00.x - C26.x, C30.x - C34.x, C37.x - C41.x, C43.x, C45.x - C58.x, C60.x - C76.x, C97.x

Enfermedad	Código ICD-9-CM	Código ICD-10
Artritis reumatoide / enfermedades vasculares del colágeno	446.x, 701.0, 710.0 - 710.4, 710.8, 710.9, 711.2, 714.x, 719.3, 720.x, 725.x, 728.5, 728.89, 729.30	L94.0, L94.1, L94.3, M05.x, M06.x, M08.x, M12.0, M12.3, M30.x, M31.0 - M31.3, M32.x - M35.x, M45.x, M46.1, M46.8, M46.9
Coagulopatía	286.x, 287.1, 287.3 - 287.5	D65 - D68.x, D69.1, D69.3 - D69.6
Obesidad	278	E66.x
Pérdida de peso	260.x - 263.x, 783.2, 799.4	E40.x - E46.x, R63.4, R64
Trastornos de líquidos y electrolitos.	253.6, 276.x	E22.2, E86.x, E87.x
Anemia por pérdida de sangre	280	D50.0
Anemia por deficiencia	280.1 - 280.9, 281.x	D50.8, D50.9, D51.x - D53.x
Abuso de alcohol	265.2, 291.1 - 291.3, 291.5 - 291.9, 303.0, 303.9, 305.0, 357.5, 425.5, 535.3, 571.0 - 571.3, 980.x, V11.3	F10, E52, G62.1, I42.6, K29.2, K70.0, K70.3, K70.9, T51.x, Z50.2, Z71.4, Z72.1
Abuso de drogas	292.x, 304.x, 305.2 - 305.9, V65.42	F11.x - F16.x, F18.x, F19.x, Z71.5, Z72.2
Psicosis	293.8, 295.x, 296.04, 296.14, 296.44, 296.54, 297.x, 298.x	F20.x, F22.x - F25.x, F28.x, F29.x, F30.2, F31.2, F31.5
Depresión	296.2, 296.3, 296.5, 300.4, 309.x, 311	F20.4, F31.3 - F31.5, F32.x, F33.x, F34.1, F41.2, F43.2

Tabla 7.2. Códigos de enfermedades para calcular el índice de comorbilidad de Elixhauser (Gasparini, 2019)

7.3. Índice de fragilidad CSHA

Appendix 1: List of variables used by the Canadian Study of Health and Aging to construct the 70-item CSHA Frailty Index		
• Changes in everyday activities	• Mood problems	• Seizures, partial complex
• Head and neck problems	• Feeling sad, blue, depressed	• Seizures, generalized
• Poor muscle tone in neck	• History of depressed mood	• Syncope or blackouts
• Bradykinesia, facial	• Tiredness all the time	• Headache
• Problems getting dressed	• Depression (clinical impression)	• Cerebrovascular problems
• Problems with bathing	• Sleep changes	• History of stroke
• Problems carrying out personal grooming	• Restlessness	• History of diabetes mellitus
• Urinary incontinence	• Memory changes	• Arterial hypertension
• Toileting problems	• Short-term memory impairment	• Peripheral pulses
• Bulk difficulties	• Long-term memory impairment	• Cardiac problems
• Rectal problems	• Changes in general mental functioning	• Myocardial infarction
• Gastrointestinal problems	• Onset of cognitive symptoms	• Arrhythmia
• Problems cooking	• Clouding or delirium	• Congestive heart failure
• Sucking problems	• Paranoid features	• Lung problems
• Problems going out alone	• History relevant to cognitive impairment or loss	• Respiratory problems
• Impaired mobility	• Family history relevant to cognitive impairment or loss	• History of thyroid disease
• Musculoskeletal problems	• Impaired vibration	• Thyroid problems
• Bradykinesia of the limbs	• Tremor at rest	• Skin problems
• Poor muscle tone in limbs	• Postural tremor	• Malignant disease
• Poor limb coordination	• Intention tremor	• Breast problems
• Poor coordination, trunk	• History of Parkinson's disease	• Abdominal problems
• Poor standing posture	• Family history of degenerative disease	• Presence of snout reflex
• Irregular gait pattern		• Presence of the palmomental reflex
• Falls		• Other medical history

Figura 7.1. Lista de variables para el índice de fragilidad CSHA (Rockwood et al., 2005)

7.4. Plataforma de análisis exploratorio de datos

- 1 Task Summary
- 2 Data Summary
- 3 Deep Dive into Individual Variables
- 3.1 Show detailed statistics of each variable
- 3.2 Visualize the target variable
- 3.3 Visualize the numerical variables
- 3.4 Visualize the categorical variables
- 4 Investigations on Multiple Variable Interactions
- 5 Final Report

3 Deep Dive into Individual Variables

3.1 Show detailed statistics of each variable

Show Variables:

as.data.frame(data)

10 Variables	32501 Observations
age	n missing unique Info Mean .05 .18 .25 .58 .75 .90 .95
32501	0 73 1 38.58 19 22 28 37 48 58 63

lowest : 17 18 19 20 21, highest: 85 86 87 88 90

workclass

n missing unique
32501 0 9

7 (1836, 6%), Federal-gov (949, 3%), Local-gov (2069, 6%), Never-worked (7, 0%)
 Private (22096, 76%), Self-emp-inc (1116, 3%), Self-emp-not-inc (2541, 8%)
 State-gov (1296, 4%), Without-pay (14, 0%)

fnlwtgt

n missing unique Info Mean .05 .18 .25 .58 .75 .90 .95
32501 0 21648 1 189378 39468 63716 117827 178356 237851 329054 379682

lowest : 12285 13769 14878 18827 19214, highest: 1220583 1268339 1366120 1455435 1484705

- 1 Task Summary
- 2 Data Summary
- 3 Deep Dive into Individual Variables
- 3.1 Show detailed statistics of each variable
- 3.2 Visualize the target variable
- 3.3 Visualize the numerical variables
- 3.4 Visualize the categorical variables
- 4 Investigations on Multiple Variable Interactions
- 5 Final Report

3.2 Visualize the target variable

Target:

Bar Plot of label.IsOver50K.

Pie Chart of label.IsOver50K.

- 1 Task Summary
- 2 Data Summary
- 3 Deep Dive into Individual Variables
- 3.1 Show detailed statistics of each variable
- 3.2 Visualize the target variable
- 3.3 Visualize the numerical variables
- 3.4 Visualize the categorical variables
- 4 Investigations on Multiple Variable Interactions
- 5 Final Report

3.3 Visualize the numerical variables

You can select the variable from the drop list.

Histogram of age

Density Plot of age

Normality test of age rejected (p-value=0)

QQ Plot of age

Boxplot of age

- 1 Task Summary
- 2 Data Summary
- 3 Deep Dive into Individual Variables
- 3.1 Show detailed statistics of each variable
- 3.2 Visualize the target variable
- 3.3 Visualize the numerical variables
- 3.4 Visualize the categorical variables
- 4 Investigations on Multiple Variable Interactions
- 5 Final Report

3.4 Visualize the categorical variables

You can select the variable from the drop list.

Categorical Variable:

Bar Plot of workclass

Pie Chart of workclass

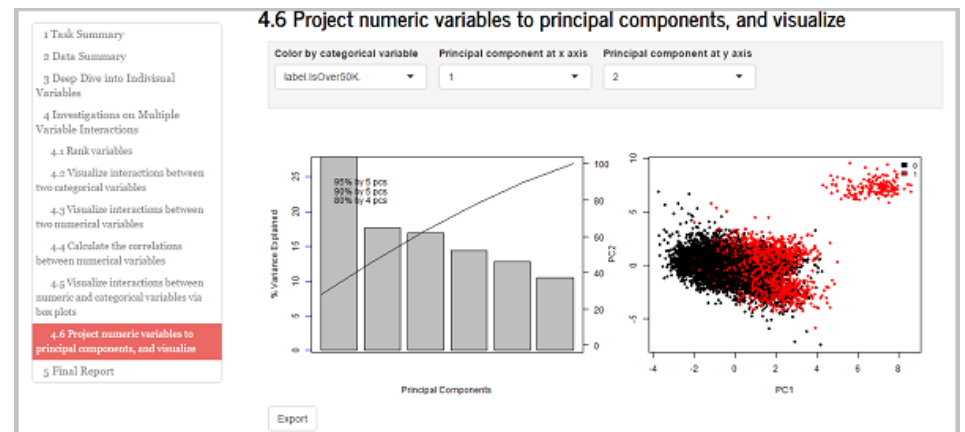
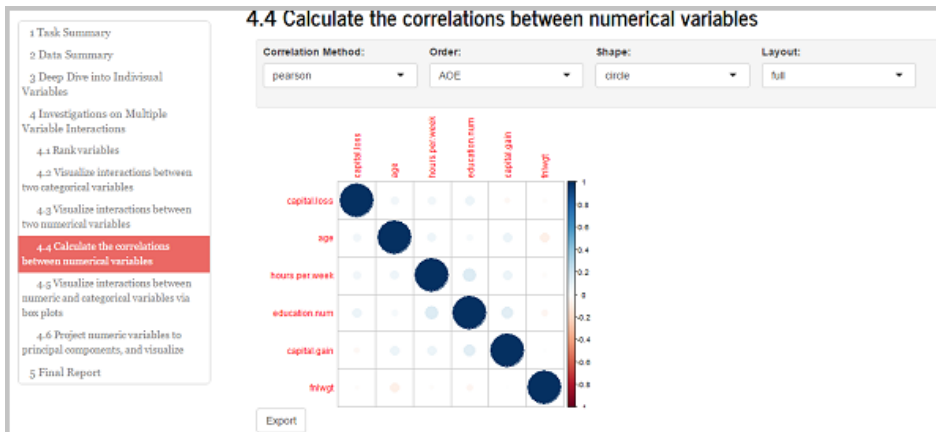
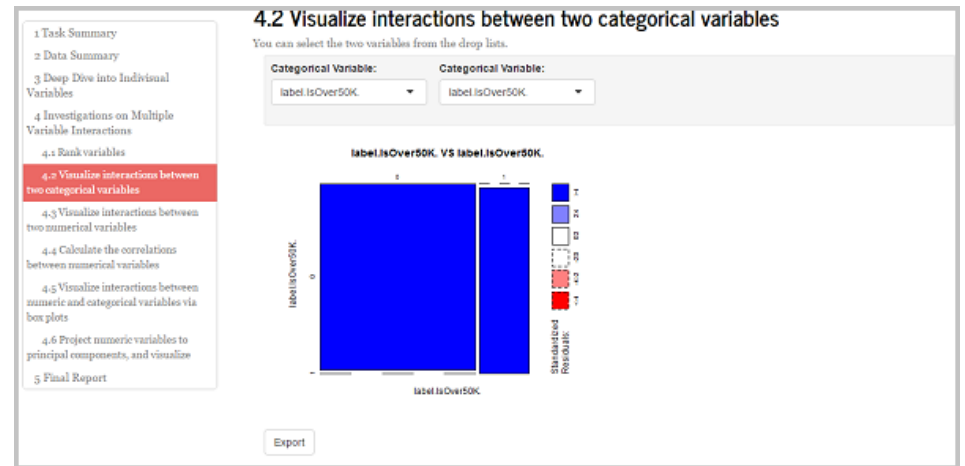
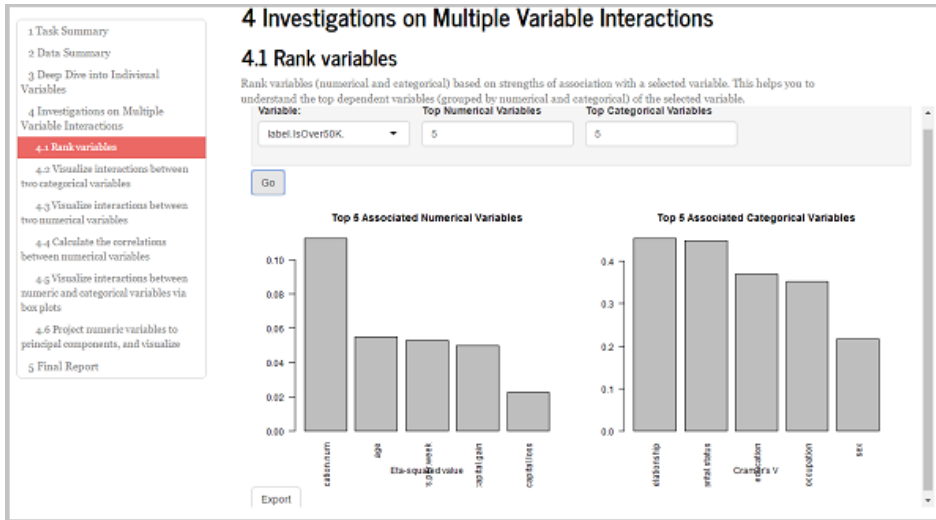


Figure 7.2. Interactive Data Exploratory Analysis and Reporting (IDEAR) (Guo, 2017)

7.5. Capítulos de ICD-10

Capítulo	Rango	Descripción
1	A00–B99	Ciertas enfermedades infecciosas y parasitarias
2	C00–D48	Tumores [neoplasias]
3	D50–D89	Enfermedades de la sangre y de los órganos hematopoyéticos, y ciertos trastornos que afectan el mecanismo de la inmunidad
4	E00–E90	Enfermedades endocrinas, nutricionales y metabólicas
5	F00–F99	Trastornos mentales y del comportamiento
6	G00–G99	Enfermedades del sistema nervioso
7	H00–H59	Enfermedades del ojo y sus anexos
8	H60–H95	Enfermedades del oído y de la apófisis mastoides
9	I00–I99	Enfermedades del sistema circulatorio
10	J00–J99	Enfermedades del sistema respiratorio
11	K00–K93	Enfermedades del sistema digestivo
12	L00–L99	Enfermedades de la piel y del tejido subcutáneo
13	M00–M99	Enfermedades del sistema osteomuscular y del tejido conjuntivo
14	N00–N99	Enfermedades del sistema genitourinario
15	O00–O99	Embarazo, parto y puerperio
16	P00–P96	Ciertas afecciones originadas en el período perinatal
17	Q00–Q99	Malformaciones congénitas, deformidades y anomalías cromosómicas
18	R00–R99	Síntomas, signos y hallazgos anormales clínicos y de laboratorio, no clasificados en otra parte
19	S00–T98	Traumatismos, envenenamientos y algunas otras consecuencias de causas externas
20	V01–Y98	Causas externas de morbilidad y de mortalidad
21	Z00–Z99	Factores que influyen en el estado de salud y contacto con los servicios de salud
22	U00–U99	Códigos para propósitos especiales

Tabla 7.3. Capítulos de ICD-10

7.6. Grupos de categorías ICD-10

Grupo de categoría	ID
Diabetes mellitus	33
Infecciones agudas de las vías respiratorias superiores	87
Enfermedades hipertensivas	79
Trastornos de otras glándulas endocrinas	35
Trastornos de la mama	128
Personas en contacto con los servicios de salud para investigación y exámenes	205
Trastornos de la glándula tiroides	32
Tumores malignos	22
Obesidad y otros tipos de hiperalimentación	38
Dorsopatías	117
Trastornos neuróticos, trastornos relacionados con el estrés y trastornos somatomorfos	44
Otras enfermedades de las vías respiratorias superiores	90
Artropatías	115
Enfermedades de las venas y de los vasos y ganglios linfáticos, no clasificadas en otra parte	85
Personas en contacto con los servicios de salud por otras circunstancias	210
Enfermedades infecciosas intestinales	1
Trastornos metabólicos	39
Micosis	15
Trastornos de los tejidos blandos	118
Otras enfermedades del sistema urinario	126
Enfermedades del esófago, del estómago y del duodeno	98
Otras infecciones agudas de las vías respiratorias inferiores	89
Trastornos de la conjuntiva	63
Tumores in situ	23
Enteritis y colitis no infecciosas	101
Dermatitis y eczema	109
Infecciones virales caracterizadas por lesiones de la piel y de las membranas mucosas	11
Síndromes del comportamiento asociados con alteraciones fisiológicas y factores físicos	45
Síntomas y signos generales	168
Enfermedades de la cavidad bucal, de las glándulas salivales y de los maxilares	97
Síntomas y signos que involucran los sistemas circulatorio y respiratorio	161
Alteraciones de la visión y ceguera	71
Enfermedades del oído medio y de la mastoides	74
Otras enfermedades de los intestinos	102
Síntomas y signos que involucran el sistema digestivo y el abdomen	162
Trastornos no inflamatorios de los órganos genitales femeninos	130

Tabla 7.4. Grupos de categorías ICD-10 más frecuentes en derechohabientes

7.7. Regresión logística en diagnósticos de interés

Para aplicar la regresión logística, se tomó una muestra del 80% de las observaciones (3688 derechohabientes) y de prueba 923 derechohabientes. Se utilizó la misma base de datos, similar a la Tabla 4.7. De éstos, los factores más relevantes para los derechohabientes de diabetes fue el sexo y 5 grupo de categorías (tumores malignos [22], tumores in situ [23], trastornos de otras glándulas endocrinas [35], trastornos de la mama [128] y personas en contacto con los servicios de salud para investigación y exámenes [205]). Éstos se encuentran marcados a continuación.

```
call: glm (formula = formula, family = "binomial", data = data_train)
```

Coefficients:

(Intercept)		Sex	Edad	`1`	`11`	`15`	
-1.327040		1.136139	0.066925	1.052315	-0.134236	0.987639	
	`22`	`23`	`32`	`35`	`38`	`39`	`44`
-4.409772	-4.010604		0.213618	-2.890536	0.883148	0.186762	-0.079363
	`45`	`63`	`71`	`74`	`79`	`85`	`87`
-0.119501	-0.069478	-0.796706		0.439728	0.512081	0.130080	-0.08679
	`89`	`90`	`97`	`98`	`101`	`102`	`109`
0.570215	-0.221600	0.255603	-0.495928	-0.317543	-0.355096		0.110235
	`115`	`117`	`118`	`126`	`128`	`130`	`161`
0.009969	0.020468	-0.248067	-0.205886	-5.390236	-0.572220		-0.379038
	`162`	`168`	`205`	`210`			
0.702512	-0.222775	-3.241982	-0.099760				

Degrees of Freedom: 3687 Total (i.e. Null); 3650 Residual

Null Deviance: 3703

Residual Deviance: 1427 AIC: 1503

La mayoría de los factores que descartaban la diabetes se relacionan directamente con el trastorno de mama, por lo que se recomienda realizar de nueva cuenta el estudio en una base de derechohabientes más heterogénea para identificar factores propios de la diabetes.

Para evaluar la eficiencia del estudio sobre los datos evaluados, se realizó una matriz de confusión, ésta se presenta en la Tabla 7.5. A partir de esta, se calcula un 94.04% de exactitud $\left(\frac{\text{Verdadero positivo} + \text{Verdadero negativo}}{\text{Total de observaciones}}\right)$.

		Predicción	
		Positivos	Negativos
Observación	Positivos	Verdadero positivo	Falso negativo
		395	29
	Negativos	Falso positivo	Verdadero negativo
		26	473

Tabla 7.5. Matriz de confusión para regresión logística de derechohabientes con diabetes

En esta matriz de confusión, se entiende que los valores falso positivo (que lo clasifique como paciente con diabetes cuando en realidad no lo padece) y falso negativo (que lo clasifique como paciente sin diabetes cuando en realidad tiene diabetes) son muy bajos en comparación con los derechohabientes que fueron correctamente clasificados como pacientes con o sin diabetes.

Para los derechohabientes con “trastornos de mama”, las variables que mejor explicaban el trastorno fueron los grupos de categoría de tumores malignos [22], tumores in situ [23], diabetes mellitus [33], trastornos de otras glándulas endocrinas [35] y personas en contacto con los servicios de salud para investigación y exámenes [205]. Las variables presentadas son las que se esperaban para los pacientes de trastorno de mama (y considerando que se tienen visitas de pacientes con ambos padecimientos [diabetes y/o trastorno de mama], los trastornos relacionados a una se vuelve trastorno excluyente de la otra). Por tal, con este análisis no se presentan hallazgos.

```
Call: glm(formula = formula, family = "binomial", data = data_train)
```

Coefficients:

```
(Intercept)      Sex      edad      `1`      `11`      `15`
  5.46683      -0.49136     -0.06608     0.89813     -0.48263     0.59658

`22`      `23`      `32`      `33`      `35`      `38`
-3.92586    -2.47195     0.39522    -6.17645    -4.68604     0.55839

`39`      `44`      `45`      `63`      `71`      `74`
0.77158     0.25244     1.74097    -0.22440    -0.66753     1.20820

`79`      `85`      `87`      `89`      `90`      `97`
-0.13279     0.37692     0.02797     0.10469     0.16224    -0.10736

`98`      `101`      `102`      `109`      `115`      `117`
-0.01339     0.23711    -0.86062     0.03140     0.01028     0.39764

`118`      `126`      `130`      `161`      `162`      `168`
```

-0.08777 -1.38787 0.27194 -0.57925 1.00845 0.82117
 `205` `210`
 -3.08524 -0.27031

Degrees of Freedom: 3687 Total (i.e. Null); 3650 Residual
 Null Deviance: 2442
 Residual Deviance: 726.3 AIC: 802.3

Con respecto a la validación, éste presentó una exactitud de 93.28%. Se presenta a continuación su matriz de confusión.

		Predicción	
		Positivos	Negativos
Observación	Positivos	Verdadero positivo 587	Falso negativo 20
	Negativos	Falso positivo 42	Verdadero negativo 274

Tabla 7.6. Matriz de confusión para regresión logística de derechohabientes con trastorno de mama.

En esta matriz de confusión, se entiende que los valores falso positivo (que lo clasifique como paciente con trastorno de mama cuando en realidad no lo padece) y falso negativo (que lo clasifique como paciente sin trastorno de mama cuando en realidad sí lo padece) son muy bajos en comparación con los derechohabientes que fueron correctamente clasificados como pacientes con o sin trastorno de mama.

En ambos casos, el hecho de que los pacientes evaluados hayan sido pacientes con diabetes y/o trastorno de mama, hizo que los resultados fueran muy claros que pertenecen a un grupo cuando no presentan los diagnósticos del otro grupo. Se presentan los resultados del análisis como ejemplo de metodología que se puede utilizar para encontrar aquellos padecimientos que explican un padecimiento y encontrar aquellos padecimientos que no son generalmente esperados. Para ello, se necesitará una base de datos de diagnósticos en pacientes con padecimientos más heterogénea.

7.8. Análisis de comorbilidades de Elixhauser según padecimiento de interés, edad y sexo

Se presenta el análisis realizado de comorbilidades de Elixhauser que afectan principalmente a ciertos grupos de personas; siendo esto según grupo de derechohabiente (si es parte del grupo de pacientes con diabetes o del grupo con trastorno de mama), grupo de edad y sexo. Este análisis se desarrolla a partir de los resultados del análisis de conglomerados (4.4.1.2). Se presenta el análisis de 3,245 derechohabientes de los cuales 2,354 (72.5%) presentan sólo una comorbilidad, 731 (22.5%) dos comorbilidades, 146 (4.5%) tres comorbilidades y 14 (0.4%) cuatro comorbilidades. De estos, el 89% son pacientes con diabetes y el resto pacientes con trastorno de mama (10%) o ambos diagnósticos (1%). El 60% fueron mujeres y 40% hombres. En la siguiente figura se presenta la relación de comorbilidades según grupo de edad y género. La mayor densidad de pacientes con comorbilidades se mantuvo consistente en los derechohabientes con mayor edad.

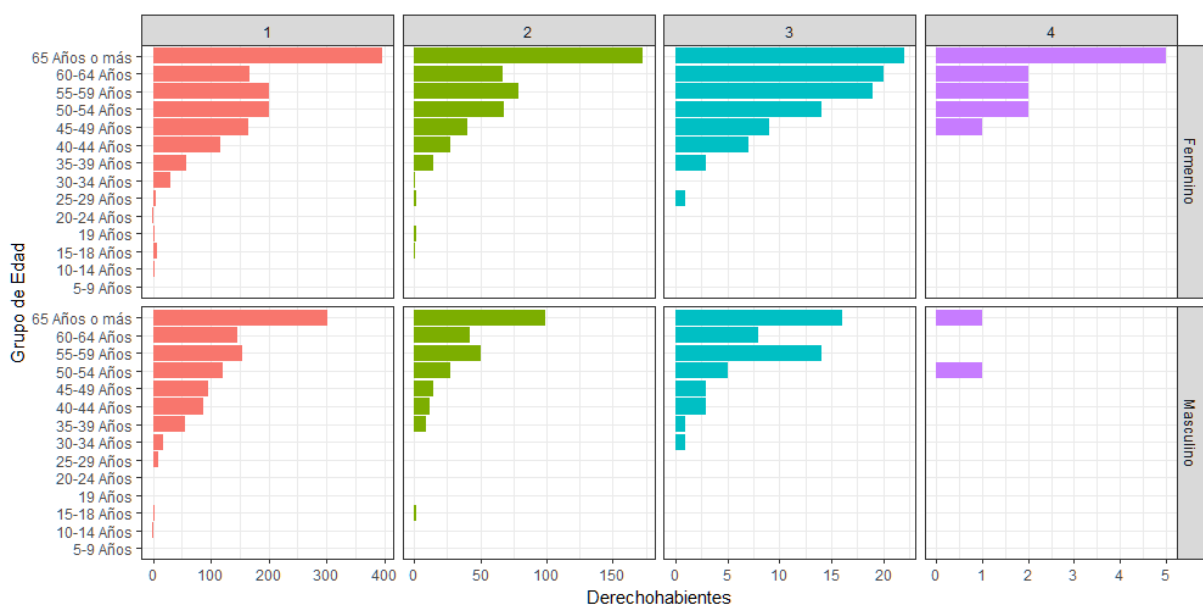


Figura 7.3. Derechohabientes con al menos una comorbilidad de Elixhauser según género y grupo de edad

El análisis que busca las relaciones causales entre las comorbilidades de Elixhauser se presenta en el capítulo 4.3.2.2 Redes bayesianas.

7.8.1 Relación general entre comorbilidades

Adicional al análisis con redes bayesianas en el capítulo 4.3.2.2, se realizó un estudio de canasta de mercado para obtener una lista de qué comorbilidades afectaban entre ellas y la frecuencia. En la siguiente tabla, se presentan las combinaciones de comorbilidades de Elixhauser presentes en derechohabientes. Debido a que el 89% de los pacientes son del grupo de diabetes, éste está presente en la mayoría de las combinaciones. Las comorbilidades más frecuentemente encontradas con la diabetes son hipertensión, hipotiroidismo, obesidad y depresión.

Combinaciones de comorbilidades de Elixhauser	Derechohabientes
Hipertensión & Diabetes	474
Hipotiroidismo & Diabetes	170
Obesidad & Diabetes	146
Hipotiroidismo & Hipertensión	64
Depresión & Diabetes	55
Obesidad & Hipertensión	51
Hipertensión, Hipotiroidismo & Diabetes	46
Hipertensión, Obesidad & Diabetes	42
Obesidad & Hipotiroidismo	27
Fallo renal & Diabetes	22
Depresión & Hipertensión	21
Hipotiroidismo, Obesidad & Diabetes	18
Tumor sólido sin metástasis & Diabetes	15
Tumor sólido sin metástasis & Hipotiroidismo	14
Depresión, Hipertensión & Diabetes	13
Linfoma & Diabetes	11
Fallo renal & Hipertensión	9
Tumor sólido sin metástasis & Obesidad	9
Insuficiencia cardiaca congestiva & Diabetes	8
Artritis reumatoide o enfermedad del colágeno vascular & Diabetes	8
Depresión & Hipotiroidismo	8
Fallo renal, Hipertensión & Diabetes	8
Hipotiroidismo, Obesidad & Hipertensión	8
Depresión, Hipotiroidismo & Diabetes	6
Hipertensión, Hipotiroidismo, Obesidad & Diabetes	6
Anemia por otras deficiencias & Diabetes	5
Otras enfermedades neurológicas & Diabetes	4
Artritis reumatoide o enfermedad del colágeno vascular & Hipotiroidismo	4
Linfoma & Hipotiroidismo	4
Linfoma & Hipertensión	4
Tumor sólido sin metástasis & Depresión	4

Tabla 7.7. Combinaciones de comorbilidades de Elixhauser en derechohabientes

En los siguientes apartados, se presentarán las combinaciones según diferentes segmentos de derechohabientes; sexo y edad. Al final, se presentarán combinaciones del 10% de derechohabientes que no pertenece al grupo de diabetes.

7.8.2 Relación de comorbilidades por sexo

La relación de derechohabientes según sexo es 60% mujeres y 40% hombres.

7.8.2.1 Mujeres

Se analizaron las comorbilidades de 1938 derechohabientes mujeres, descartando en los resultados 1356 que sólo presentaban una comorbilidad (resultando en 582 derechohabientes con dos o más comorbilidades). Las primeras cinco combinaciones se mantuvieron constantes con los resultados generales.

Mujeres	
Combinaciones de comorbilidades de Elixhauser	Derechohabientes
Diabetes & Hipertensión	265
Diabetes & Hipotiroidismo	132
Diabetes & Obesidad	94
Hipertensión & Hipotiroidismo	48
Diabetes & Depresión	39
Diabetes, Hipertensión & Hipotiroidismo	32
Hipertensión & Obesidad	26
Hipotiroidismo & Obesidad	24
Diabetes, Hipertensión & Obesidad	20
Diabetes & Fallo renal	17
Hipertensión & Depresión	16
Diabetes, Hipotiroidismo & Obesidad	15
Hipertensión & Tumor sólido sin metástasis	14
Hipotiroidismo & Tumor sólido sin metástasis	13
Diabetes & Tumor sólido sin metástasis	11
Fallo renal & Hipertensión	9
Tumor sólido sin metástasis & Obesidad	9
Depresión & Hipotiroidismo	8
Fallo renal, Hipertensión & Diabetes	8
Depresión, Hipertensión & Diabetes	8
Linfoma & Diabetes	7
Artritis reumatoide o enfermedad del colágeno vascular & Diabetes	6
Depresión, Hipotiroidismo & Diabetes	6
Hipotiroidismo, Obesidad & Hipertensión	6
Otras enfermedades neurológicas & Diabetes	4
Insuficiencia cardiaca congestiva & Diabetes	4
Artritis reumatoide o enfermedad del colágeno vascular & Hipotiroidismo	4

Mujeres	
Combinaciones de comorbilidades de Elixhauser	Derechohabientes
Linfoma & Hipotiroidismo	4
Hipertensión, Hipotiroidismo, Obesidad & Diabetes	4
Alteración de fluidos y electrolitos corporales & Diabetes	3
Fallo renal & Hipotiroidismo	3
Depresión & Tumor sólido sin metástasis	3
Depresión & Obesidad	3
Fallo renal, Hipotiroidismo & Diabetes	3
Anemia por otras deficiencias & Diabetes	2
Artritis reumatoide o enfermedad del colágeno vascular & Hipertensión	2
Linfoma & Hipertensión	2
Hipotiroidismo, Linfoma & Diabetes	2
Fallo renal, Hipotiroidismo & Hipertensión	2
Depresión, Hipotiroidismo & Hipertensión	2
Obesidad, Tumor sólido sin metástasis & Hipertensión	2
Hipotiroidismo, Tumor sólido sin metástasis & Hipertensión	2
Hipotiroidismo, Tumor sólido sin metástasis & Diabetes	2
Fallo renal, Hipertensión, Hipotiroidismo & Diabetes	2

Tabla 7.8. Combinaciones de comorbilidades de Elixhauser en derechohabientes (mujeres)

7.8.2.2 Hombres

Se analizaron las comorbilidades de 1307 derechohabientes hombres, descartando en los resultados 998 que sólo presentaban una comorbilidad (resultando en 309 derechohabientes con dos o más comorbilidades). A diferencia de los resultados generales y en mujeres, sólo cinco derechohabientes hombres con múltiples comorbilidades fueron diagnosticados con depresión (siendo la combinación de comorbilidades: depresión, hipertensión & diabetes). Los primeros tres lugares presentan las mismas combinaciones que en resultados generales, pero en diferente orden.

Hombres	
Combinaciones de comorbilidades de Elixhauser	Derechohabientes
Diabetes & Hipertensión	209
Diabetes & Obesidad	52
Diabetes & Hipotiroidismo	38
Hipertensión & Obesidad	25
Diabetes, Hipertensión & Obesidad	22
Hipertensión & Hipotiroidismo	16
Diabetes, Hipertensión & Hipotiroidismo	14
Fallo renal & Diabetes	5
Tumor sólido sin metástasis & Diabetes	5
Depresión & Hipertensión	5
Depresión, Hipertensión & Diabetes	5
Linfoma & Diabetes	4

Hombres	
Combinaciones de comorbilidades de Elixhauser	Derechohabientes
Insuficiencia cardiaca congestiva & Diabetes	4
Coagulopatía & Diabetes	3
Anemia por otras deficiencias & Diabetes	3
Arritmia cardiaca & Diabetes	3
Hipotiroidismo & Obesidad	3
Hipotiroidismo, Obesidad & Diabetes	3
Artritis reumatoide o enfermedad del colágeno vascular & Diabetes	2
Coagulopatía & Hipertensión	2
Anemia por otras deficiencias & Hipertensión	2
Arritmia cardiaca & Hipertensión	2
Linfoma & Hipertensión	2
Insuficiencia cardiaca congestiva & Hipertensión	2
Coagulopatía, Hipertensión & Diabetes	2
Anemia por otras deficiencias, Hipertensión & Diabetes	2
Hipotiroidismo, Obesidad & Hipertensión	2
Hipertensión, Hipotiroidismo, Obesidad & Diabetes	2

Tabla 7.9. Combinaciones de comorbilidades de Elixhauser en derechohabientes (hombres)

7.8.3 Relación de comorbilidades por edad

Sólo 891 derechohabientes presentan más de una comorbilidad, de estos según el rango de edad. Debido a que la población joven cuenta con pocos derechohabientes para el análisis, éste se dividió en cuatro grupos. Éstos se encuentran a continuación.

Edad	Derechohabientes	Grupo	%	Nombre
15-18 Años	3	1	4.3%	15 – 39 Años
19 Años	2			
25-29 Años	3			
30-34 Años	2			
35-39 Años	28	2	26.3%	40 – 49 Años
40-44 Años	49			
45-49 Años	68			
50-54 Años	117	3	34.0%	55 – 64 Años
55-59 Años	164			
60-64 Años	139	4	35.5%	65 Años o más
65 Años o más	316			
Total	891			

Tabla 7.10. Grupos de edades para análisis de relación de comorbilidades

7.8.3.1 Grupo 1: 15 a 39 años

Se analizaron 38 derechohabientes con múltiples comorbilidades. En la siguiente tabla se presentan las combinaciones de comorbilidades que presentaron. Las combinaciones con diabetes, obesidad e hipertensión prevalecen en los primeros lugares como en las combinaciones generales.

15 a 39 años	
Combinaciones de comorbilidades de Elixhauser	Derechohabientes
Diabetes & Obesidad	14
Diabetes & Hipertensión	9
Diabetes & Depresión	5
Obesidad & Hipertensión	4
Diabetes & Hipotiroidismo	3
Diabetes, Obesidad & Hipertensión	3
Obesidad & Hipotiroidismo	2
Obesidad & Tumor sólido sin metástasis	2
Tumor sólido sin metástasis & Abuso de drogas	1
Depresión & Alteración de fluidos y electrolitos corporales	1
Diabetes & Alteración de fluidos y electrolitos corporales	1
Anemia por otras deficiencias & Depresión	1
Hipotiroidismo & Hipertensión	1
Hipotiroidismo & Tumor sólido sin metástasis	1
Depresión, Diabetes & Alteración de fluidos y electrolitos corporales	1
Hipertensión, Obesidad & Hipotiroidismo	1

Tabla 7.11. Combinaciones de comorbilidades de Elixhauser en derechohabientes (15 – 39 años)

7.8.3.2 Grupo 2: 40 – 49 Años

Se analizaron 117 derechohabientes con múltiples comorbilidades. En la siguiente tabla se presentan las combinaciones de comorbilidades que presentaron. De éstas, se omitieron 16 combinaciones que sólo se presentaron una vez. Los primeros lugares en combinaciones concuerdan con los presentados en el general.

40 a 49 años	
Combinaciones de comorbilidades de Elixhauser	Derechohabientes
Diabetes & Hipertensión	49
Diabetes & Obesidad	31
Diabetes & Hipotiroidismo	19
Hipertensión & Obesidad	10
Diabetes, Hipertensión & Obesidad	7
Hipertensión & Depresión	6
Obesidad & Hipotiroidismo	5
Hipotiroidismo & Hipertensión	4
Diabetes, Hipertensión & Depresión	4
Linfoma & Diabetes	3

40 a 49 años	
Combinaciones de comorbilidades de Elixhauser	Derechohabientes
Depresión & Obesidad	3
Tumor sólido sin metástasis & Hipertensión	3
Diabetes, Obesidad & Hipotiroidismo	3
Depresión & Tumor sólido sin metástasis	2
Depresión & Hipotiroidismo	2
Tumor sólido sin metástasis & Hipotiroidismo	2
Diabetes, Hipertensión & Hipotiroidismo	2

Tabla 7.12. Combinaciones de comorbilidades de Elixhauser en derechohabientes (40 – 49 años)

7.8.3.3 Grupo 3: 50 – 64 años

Se analizaron 281 derechohabientes con múltiples comorbilidades. En la siguiente tabla se presentan las combinaciones de comorbilidades que presentaron. De éstas, se omitieron 30 combinaciones que sólo se presentaron una vez y 8 combinaciones que se presentaron dos veces. Los primeros lugares concuerdan con los resultados en combinaciones generales.

50 a 64 años	
Combinaciones de comorbilidades de Elixhauser	Derechohabientes
Diabetes & Hipertensión	151
Diabetes & Obesidad	57
Diabetes & Hipotiroidismo	48
Hipertensión & Obesidad	23
Diabetes & Depresión	21
Hipertensión & Hipotiroidismo	20
Diabetes, Hipertensión & Obesidad	20
Diabetes, Hipertensión & Hipotiroidismo	15
Hipotiroidismo & Obesidad	11
Hipertensión & Depresión	9
Diabetes, Hipotiroidismo & Obesidad	7
Tumor sólido sin metástasis & Hipotiroidismo	6
Diabetes & Tumor sólido sin metástasis	6
Diabetes, Hipertensión & Depresión	5
Diabetes & Artritis reumatoide o enfermedad del colágeno vascular	4
Linfoma & Diabetes	3
Tumor sólido sin metástasis & Obesidad	3
Hipotiroidismo, Obesidad & Hipertensión	3
Hipertensión, Hipotiroidismo & Obesidad	3
Diabetes, Hipertensión, Hipotiroidismo & Obesidad	3

Tabla 7.13. Combinaciones de comorbilidades de Elixhauser en derechohabientes (50 – 64 años)

7.8.3.4 Grupo 4: 65 años o más

Se analizaron 316 derechohabientes con múltiples comorbilidades. En la siguiente tabla se presentan las combinaciones de comorbilidades que presentaron. Los primeros nueve lugares concuerdan con los resultados generales.

65 años o más	
Combinaciones de comorbilidades de Elixhauser	Derechohabientes
Diabetes & Hipertensión	187
Diabetes & Hipotiroidismo	71
Hipertensión & Hipotiroidismo	27
Diabetes & Obesidad	23
Diabetes, Hipertensión & Hipotiroidismo	20
Diabetes & Fallo renal	15
Diabetes & Depresión	12
Hipertensión & Obesidad	9
Diabetes, Hipertensión & Obesidad	8
Anemia por otras deficiencias & Diabetes	5
Insuficiencia cardíaca congestiva & Diabetes	5
Fallo renal & Hipertensión	5
Obesidad & Hipotiroidismo	5
Diabetes, Hipotiroidismo & Obesidad	5
Tumor sólido sin metástasis & Hipertensión	4
Diabetes, Hipertensión & Fallo renal	4
Otras enfermedades neurológicas & Diabetes	3
Artritis reumatoide o enfermedad del colágeno vascular & Diabetes	3
Linfoma & Diabetes	3
Depresión & Hipertensión	3
Hipotiroidismo, Obesidad & Hipertensión	3
Diabetes, Hipertensión, Hipotiroidismo & Obesidad	3
Arritmia cardíaca & Diabetes	2
Anemia por otras deficiencias & Hipertensión	2
Insuficiencia cardíaca congestiva & Hipertensión	2
Depresión & Hipotiroidismo	2
Tumor sólido sin metástasis & Hipotiroidismo	2
Fallo renal & Hipotiroidismo	2
Anemia por otras deficiencias, Hipertensión & Diabetes	2
Depresión, Hipotiroidismo & Diabetes	2
Depresión, Hipertensión & Diabetes	2
Diabetes, Hipotiroidismo & Fallo renal	2

Tabla 7.14. Combinaciones de comorbilidades de Elixhauser en derechohabientes (65 años o más)

7.8.4 Relación de comorbilidades en derechohabientes con trastorno de mama

En todas las categorías anteriores prevalecieron las comorbilidades de diabetes, hipertensión, hipotiroidismo y obesidad; siendo estas esperadas en la población con diabetes. Por ello, se analiza a continuación a los 312 derechohabientes con trastorno de mama, de los cuales sólo 56 presentan más de una comorbilidad de Elixhauser diagnosticada. Este análisis sería del 6.3% de la población de derechohabientes con más de una comorbilidad. Los derechohabientes presentaron principalmente combinaciones con tumor sólido sin metástasis. Esta comorbilidad estuvo acompañada de hipertensión, hipotiroidismo y obesidad.

Combinaciones de comorbilidades de Elixhauser	Derechohabientes
Tumor sólido sin metástasis & Hipertensión	12
Tumor sólido sin metástasis & Hipotiroidismo	12
Tumor sólido sin metástasis & Obesidad	8
Hipotiroidismo & Obesidad	6
Hipotiroidismo & Hipertensión	6
Hipertensión & Obesidad	4
Hipotiroidismo & Artritis reumatoide o enfermedad del colágeno vascular	3
Hipertensión & Depresión	3
Tumor sólido sin metástasis & Depresión	3
Obesidad & Depresión	2
Hipertensión, Tumor sólido sin metástasis & Obesidad	2
Abuso de drogas & Tumor sólido sin metástasis	1
Insuficiencia cardíaca congestiva & Hipertensión	1
Alcoholismo & Enfermedad hepática	1
Alcoholismo & Tumor sólido sin metástasis	1
Enfermedad hepática & Tumor sólido sin metástasis	1
Anemia por otras deficiencias & Depresión	1
Linfoma & Tumor sólido sin metástasis	1
Artritis reumatoide o enfermedad del colágeno vascular & Hipertensión	1
Artritis reumatoide o enfermedad del colágeno vascular & Tumor sólido sin metástasis	1
Depresión & Hipotiroidismo	1
Alcoholismo, Enfermedad hepática & Tumor sólido sin metástasis	1
Artritis reumatoide o enfermedad del colágeno vascular, Hipotiroidismo & Tumor sólido sin metástasis	1
Depresión, Hipertensión & Tumor sólido sin metástasis	1
Hipertensión, Obesidad & Hipotiroidismo	1
Hipotiroidismo, Obesidad & Tumor sólido sin metástasis	1
Hipotiroidismo, Tumor sólido sin metástasis & Hipertensión	1

Tabla 7.15. Combinaciones de comorbilidades de Elixhauser en derechohabientes (trastorno de mama)

7.8.5 Conclusiones

Se recomienda realizar este estudio con población más heterogénea para identificar cuáles combinaciones de comorbilidades son exclusivas o principales para cada diagnóstico de interés (más allá de pacientes con diabetes o trastorno de mama) y en una población más grande que sea representativa para cada diagnóstico en la entidad. Tanto los pacientes del grupo de diabetes como los pacientes del grupo de trastorno de mama presentaron las comorbilidades de hipertensión, hipotiroidismo y obesidad. Estas comorbilidades se mantuvieron en los primeros lugares independientemente del género y edad. La comorbilidad que prevaleció a uno de los grupos fue la depresión, que fue mayormente diagnosticada en mujeres.