



**UNIVERSIDAD DE CONCEPCIÓN
FACULTAD DE INGENIERÍA
DEPARTAMENTO DE INGENIERÍA ELECTRICA**



**VISUALIZADOR DE DATOS DE ATENCIONES AMBULATORIAS DEL
HOSPITAL GUILLERMO GRANT BENAVENTE DE CONCEPCIÓN**

POR

Alan Andrés Benavente Toro

Informe Final Memoria de Título presentada a la Facultad de Ingeniería de la Universidad de Concepción para optar al grado académico de Ingeniero/a Civil Biomédica

Profesores Guía:
Dr. Rosa Figueroa I.
Dr. Christopher Flores

Comisión:
Dr. Cesar García
Dr. Pamela Guevara

Octubre
Concepción
(Chile)

© 2023 Alan Benavente Toro

© 2023 Alan Benavente Toro

Ninguna parte de esta memoria puede reproducirse o transmitirse bajo ninguna forma o por ningún medio o procedimiento, sin permiso por escrito del autor.

Agradecimientos

A mi familia por su constante apoyo y cariño.

A mis amigos por las historias vividas.

Y a todos aquellos para quienes la vida nos ha dado distintos caminos, pero tendrán siempre un lugar en mis recuerdos y mi corazón.

Resumen

El presente trabajo tuvo como finalidad crear un tablero de visualización de datos utilizando información de una base de registros médicos electrónicos provistos por el hospital Guillermo Grant Benavente de Concepción, presentando la información en diferentes medios gráficos con el propósito de apoyar la toma de decisiones. Este visualizador se diseñó y desarrollo con una arquitectura que permitiera la escalabilidad del modelo propuesto y la incorporación de nuevos datos en el futuro.

Como fuente de información del tablero, se utilizaron tanto los campos estructurados como no estructurados de los registros médicos electrónicos. Para la utilización de los campos no estructurados de la base de datos fue necesario realizar un preprocesamiento del texto libre, su tokenización y posterior almacenamiento en un *data warehouse* creado para alimentar el tablero de visualización.

Para generar los diferentes elementos visuales incluidos en el tablero, se filtraron los registros para cada una de las 8 especialidades de atención ambulatoria. Los elementos visuales representan la siguiente información: i) distribución etaria de los pacientes, ii) distribución en serie temporal de las consultas médicas, iii) hábitos más comunes de los pacientes y iv) la relación de los hábitos y antecedentes de los pacientes con su posterior diagnóstico, entre otros.

Del análisis de los medios gráficos creados con el conjunto de datos disponible para este trabajo, se puede notar que mayoritariamente los hábitos reportados por los médicos en las distintas especialidades son los mismos. Lo anterior, puede tener su causa en distintos factores, tales como, un subreporte de un hábito o antecedente considerado de menor relevancia o bien a una fuerte correlación del hábito o antecedente con patologías de esa subespecialidad. Adicionalmente, se puede comentar que, del análisis visual de la información obtenida, es posible establecer un perfil de las características de los pacientes para cada especialidad. Finalmente se condujo una evaluación de la plataforma propuesta con ayuda de 3 evaluadores, los cuales coincidieron en estar muy satisfechos con el diseño y apariencia de la plataforma. Sin embargo, 2 de ellos acotaron que podrían mejorarse las descripciones para dar más claridad a la información presentada.

Abstract

The purpose of this work was to create a data visualization dashboard using information from an electronic medical records database provided by the Guillermo Grant Benavente Hospital in Concepción, presenting the information in different graphic media with the purpose of supporting decision making. This viewer was designed and developed with an architecture that would allow the scalability of the proposed model and the incorporation of new data in the future.

As a source of information for the dashboard, both structured and unstructured fields from the electronic medical records were used. In order to use the unstructured fields of the database, it was necessary to preprocess the free text, tokenize it and then separate it with the purpose of storing it in a data warehouse created to feed the visualization dashboard.

To generate the different visual elements included in the dashboard, records were filtered for each of the 8 outpatient specialties. The visual elements represent the following information: i) age distribution of patients, ii) time series distribution of medical consultations, iii) most common habits of patients and iv) the relationship of patients' habits and history with their subsequent diagnosis, among others.

From the analysis of the graphic media created with the dataset available for this work, it can be noted that the habits reported by physicians in the different specialties are mostly the same. This may be due to different factors, such as an underreporting of a habit or history considered of minor relevance or a strong correlation of the habit or history with pathologies of that subspecialty. Additionally, it can be commented that, from the visual analysis of the information obtained, it is possible to establish a profile of patient characteristics for each specialty.

Finally, an evaluation of the proposed platform was conducted with the help of 3 evaluators, all of whom agreed that they were very satisfied with the design and appearance of the platform. However, 2 of them noted that the descriptions could be improved to give more clarity to the information presented.

Tabla de Contenidos

CAPÍTULO 1. INTRODUCCIÓN.....	9
1.1. INTRODUCCIÓN GENERAL	9
1.2. OBJETIVOS.....	10
1.2.1 <i>Objetivo General</i>	10
1.2.2 <i>Objetivos Específicos</i>	10
1.3. ALCANCES Y LIMITACIONES.....	10
1.4. TEMARIO	10
CAPÍTULO 2. MARCO TEÓRICO	12
2.1. INTRODUCCIÓN.....	12
2.2. REVISIÓN BIBLIOGRÁFICA	12
2.2.1 <i>Aplicaciones del big data en salud</i>	12
2.2.2 <i>Visualización de datos en salud</i>	13
2.3. MARCO TEÓRICO.....	16
2.3.1 <i>Procesamiento de textos</i>	16
2.3.2 <i>Tableros</i>	16
2.3.3 <i>Visualización de datos</i>	17
2.3.4 <i>Registro médico electrónico</i>	17
2.3.5 <i>Business Intelligence</i>	18
2.4. DISCUSIÓN.....	20
CAPÍTULO 3. MATERIALES Y MÉTODOS.....	21
3.1. INTRODUCCIÓN.....	21
3.2. MATERIALES	21
3.2.1 <i>Herramientas de software</i>	22
3.3. MÉTODOS	23
3.3.1 <i>Procesamiento del texto</i>	23
3.3.2 <i>Creación del data warehouse</i>	23
3.3.3 <i>Construcción de medios gráficos</i>	23
3.3.4 <i>Encuesta de evaluación</i>	24
CAPÍTULO 4. RESULTADOS.....	26
4.1. INTRODUCCIÓN.....	26
4.2. RESULTADOS	26
4.3. EVALUACIÓN.....	28
CAPÍTULO 5. CONCLUSIONES.....	30
5.1. INTRODUCCIÓN.....	30
5.2. CONCLUSIONES	30
5.3. TRABAJO FUTURO	30
CAPÍTULO 6. GLOSARIO	31
CAPÍTULO 7. REFERENCIAS	32
ANEXO A. EXTRACTOS DE CÓDIGO.....	34
A.1 SCRIPT DE PREPROCESAMIENTO	34
A.2 SCRIPT DE CONEXIÓN A BASE DE DATOS.....	34

Lista de Tablas

Tabla 2.1 Categorización de métodos de visualización	18
Tabla 4.1 Edad promedio de los pacientes por especialidad.....	26
Tabla 4.2 Hábito más reportado por especialidad	27
Tabla 4.3 Antecedente más reportado por especialidad	27
Tabla 4.4 Resultados de la encuesta de evaluación.....	29

Lista de Figuras

Figura 2.1 Dashboard del Coronavirus Resource Center de la Jonhns Hopkins University, recuperado de [16].....	15
Figura 2.2 Tablero de KPIs anual para un centro de salud el cual muestra métricas relacionadas al número de citas para ese periodo, recuperado de [12]	15
Figura 3.1 Diagrama de componentes del modelo de visualización propuesto	22
Figura 3.2 Esquema de la base de datos utilizada como data warehouse	24
Figura 4.1 Tablero de métricas construido filtrado por especialidad de cardiología	27
Figura 4.2 Gráficos de relación de términos de antecedentes y hábitos con diagnósticos filtrado por especialidad cardiología adulto	28

Capítulo 1. Introducción

1.1. Introducción General

Durante la última década, al igual que muchas otras industrias, los servicios de salud han experimentado un rápido proceso de digitalización de la información, lo que ha llevado a su vez a una rápida expansión del campo de la tecnología de la información sanitaria (HIT) con enfoque en aumentar la calidad, eficiencia y seguridad de los servicios sanitarios [1]. Dentro de este nuevo campo de la HIT los registros médicos electrónicos (EMR) son el remplazo a las fichas clínicas de papel.

La adopción de los EMR abre una nueva ventana de oportunidad para análisis basado en datos, puesto que pone al alcance de los investigadores grandes volúmenes de información epidemiológica de la población. Sin embargo, estos análisis no están exentos de dificultades, ya que buena parte de la información relevante en estos EMR se encuentra en forma de texto libre para lo cual se hace necesario el uso de mecanismos de extracción de información mediante el uso de procesamiento del lenguaje natural (NLP) [2] en conjunto a otras técnicas como la visualización de datos, que permitan presentar la información extraída de forma resumida y comprensible para el ser humano.

La visualización de datos, es la representación visual de datos codificados usando posición, longitud, tamaño o color, entre otros atributos, para apoyar el descubrimiento y la comprensión de patrones. Relacionado a lo anterior se encuentran las técnicas de *visual analytics* las cuales combinan análisis computacional e interactivo de interfaces de usuario basadas en visualización para apoyar el razonamiento analítico y cognición humana [3].

En la última década diferentes industrias han integrado el análisis basado en datos en sus procesos de toma de decisiones, método conocido como toma de decisiones basada en datos (DDD), este método DDD ha permitido reemplazar la intuición o experiencia por datos empíricos como base la toma de decisiones, lo que ha conducido a mejores resultados de desempeño y procesos más fácilmente monitorizables, en el área de la salud en particular la integración de la metodología DDD suele traducirse en ahorro de recursos, reducción de tiempos de espera y mayores tasas de satisfacción del paciente. El presente trabajo utiliza información extraída de una base datos de EMR facilitados por el hospital regional Guillermo Grant Benavente (HGGB) de Concepción, con el objetivo de construir un visualizador de datos que sirva como herramienta de apoyo en la toma de decisiones y análisis dentro de dicho centro médico.

1.2. Objetivos

1.2.1 Objetivo General

Desarrollar un visualizador de datos de atención de pacientes ambulatorios del hospital Guillermo Grant Benavente de Concepción con información anonimizada proveniente de registros médicos electrónicos de atención ambulatoria.

1.2.2 Objetivos Específicos

- Creación de una base de datos con información textual obtenida de los registros médicos electrónicos de atención ambulatoria entre los años 2011-2012
- Seleccionar variables e indicadores a implementar en la visualización que puedan describir la información presente en el conjunto de datos para apoyo a la toma de decisiones.
- Desarrollo del visualizador de datos junto con el *data warehouse* utilizando una arquitectura escalable.
- Evaluación a nivel de laboratorio de las funcionalidades del visualizador.

1.3. Alcances y Limitaciones

- i. *Composición de la base de datos:* El presente trabajo utilizará EMR provenientes de atenciones médicas del Hospital Regional Guillermo Grant Benavente, las cuales fueron adquiridas durante el periodo 2011-2012 para 9 sub-especialidades médicas.

1.4. Temario

- (i) Capítulo 1: Introducción en forma general del trabajo, indicando los objetivos y alcances del proyecto.
- (ii) Capítulo 2: En esta sección se tiene una revisión bibliográfica sobre los trabajos de distintos grupos de investigadores, que aportan información sobre los métodos utilizados y experiencias al crear tableros de visualización de datos similares al presentado en este estudio. Además, en el segundo capítulo, se mostrará el marco teórico de los distintos conceptos necesarios para la construcción de un visualizador, haciendo énfasis en los tipos de tableros de datos y técnicas de visualización.

- (iii) Capítulo 3: Se realiza una descripción de la metodología utilizada para implementar la arquitectura del tablero de visualización, basándose principalmente en la información recabada en la revisión bibliográfica.
- (iv) Capítulo 4: Se entregan los distintos resultados obtenidos en la prueba de concepto del modelo de visualización propuesto y los resultados de la encuesta de evaluación.
- (v) Capítulo 5: Se relatan las conclusiones que se obtuvieron del trabajo realizado, así como recomendaciones de trabajos futuros en la línea de seguir madurando el modelo propuesto.

Capítulo 2. Marco Teórico

2.1. Introducción

En este capítulo se expondrán experiencias de otros investigadores implementando modelos de visualización de datos en salud. Por otra parte, se expondrán los componentes de un modelo de visualización de datos y los criterios a tener en cuenta a la hora de su implementación.

2.2. Revisión bibliográfica

Con el auge de las tecnologías de la información, el contar con una gran cantidad de datos para la toma de decisiones no representa un desafío técnico; el verdadero reto radica en lograr extraer la información relevante del universo de información disponible para aumentar la eficiencia del proceso de toma de decisiones. Dentro del abanico de métodos para el análisis de información se encuentran los métodos de visualización de datos, los cuales buscan representar los datos a través de diferentes métodos visuales con el fin de mostrar relaciones, patrones u otra clase de información relevante, de tal forma que un humano pueda interpretarla con facilidad. A continuación se revisarán algunos conceptos y las metodologías de visualización de datos más utilizadas y reportadas en la literatura.

2.2.1 Aplicaciones del big data en salud

El *big data* es el concepto que engloba todo conjuntos de datos o combinaciones de conjuntos de datos cuyo tamaño, complejidad y velocidad de crecimiento dificultan su captura, gestión, procesamiento o análisis mediante tecnologías y herramientas convencionales [4], tal como es el caso de los datos generados en salud. Actualmente existen diversas áreas de aplicación del *big data* en salud, estos sistemas buscan ayudar al personal clínico a tomar decisiones basados en datos empíricos en lugar de intuición u otros factores subjetivos, esto se logra analizando grandes volúmenes de información clínica con la finalidad de detectar patrones y relaciones. Por ejemplo, Obermeyer [5], utiliza algoritmos de aprendizaje automático para analizar los historiales médicos electrónicos de pacientes de la UCI con el fin de predecir la probabilidad de shock séptico. Mediante el análisis de patrones y factores de riesgo asociados con el shock séptico, el algoritmo proporciona señales de alerta temprana, mejorando los resultados atención y reduciendo el riesgo de shock séptico.

Otra área de aplicación de aplicación del *big data* en salud es el análisis predictivo, este tipo de análisis tiene por objetivo pronosticar posibles eventos en la salud de los pacientes, como desarrollo de patologías o tasa de éxito de tratamientos e identificar las poblaciones de alto riesgo determinando

los principales factores asociados a acontecimientos adversos, este tipo de análisis puede utilizarse posteriormente para desarrollar intervenciones y estrategias preventivas, según Rajkomar [6] este enfoque proactivo no solo mejora los resultados, sino que también reduce eficazmente los gastos sanitarios. Cutillo [7] señala que mediante la previsión del volumen de pacientes y la predicción de los resultados de los tratamientos, los proveedores de atención sanitaria pueden asignar recursos y optimizar los niveles de dotación de personal, garantizando una prestación eficiente de la atención. Esto no solo conlleva un ahorro de costes, sino que también mejora la satisfacción de los pacientes, ya que se reducen los tiempos de espera. Un ejemplo de ello lo presenta Lee [8] con un modelo predictivo que utiliza información clínica del paciente para estimar el riesgo de reingreso a hospitalización en pacientes diagnosticados con cáncer pancreático luego de ser dados de alta, ayudando a predecir la tasa futura de ocupación de camas en el centro clínico.

Entidades dedicadas al monitoreo de la salud pública también pueden beneficiarse de la incorporación de herramientas de *big data* en sus procesos de análisis, en este sentido estas herramientas pueden facilitar la detección precoz de brotes de enfermedades al permitir el seguimiento y análisis en tiempo real de grandes volúmenes de datos. Los sistemas de vigilancia tradicionales suelen depender de la notificación de enfermedades de declaración obligatoria por parte de los profesionales sanitarios, lo que puede provocar retrasos en la identificación y respuesta a los brotes. Sin embargo, el análisis de *big data* puede procesar grandes cantidades de información procedente de diversas fuentes, como las redes sociales, las búsquedas en Internet y los historiales médicos electrónicos, para identificar posibles grupos de síntomas o enfermedades. Esto puede permitir intervenciones oportunas para mitigar la propagación de enfermedades infecciosas y minimizar las tasas de morbilidad y mortalidad asociadas [9].

Como se mostró, el *big data* tiene a día de hoy un gran número de aplicaciones biomédicas, en este sentido el uso de herramientas como la visualización de datos puede revelar valiosa información con diversas aplicaciones. Senthilkumar [10] menciona algunas aplicaciones de herramientas de visualización en el campo de la biomedicina, tales como en compañías aseguradoras para la prevención de fraude, diseño de equipos médicos, desarrollo de fármacos y cuidado personalizado del paciente.

2.2.2 Visualización de datos en salud.

Para construir un tablero de visualización existen varias metodologías que pueden ser implementadas, Galuzio [11] hace una revisión de una serie de tableros de visualización orientados a

asistir en la toma de decisiones. Como resultado de la revisión, se estableció el método más utilizado para la construcción de tableros es el método de factores clave de rendimiento (KPI, por sus siglas en inglés). El trabajo de Pestana [12], nos muestra un ejemplo del uso de tableros de visualización para monitoreo y seguimiento de KPI. Este trabajo propone un tablero de KPI para monitorizar la productividad de un hospital con el fin de establecer parámetros que sirvan de guía a la hora de asignar los recursos del hospital de manera más eficiente. Otro método de visualización de datos, es el cuadro de mando integral (BSC, por sus siglas en inglés). Al respecto, Gonzalez-Sanchez [13] hace un análisis de la adopción del método BSC en España, Italia y Portugal en el área sanitaria y concluye que dicho método se encuentra en una fase temprana de adopción.

En cuanto al objetivo del análisis, en la construcción de tableros de visualización en salud podemos distinguir principalmente tres niveles agregación de la información, los cuales son nivel individual, institucional y nivel poblacional. A nivel poblacional el análisis realizado es de tipo epidemiológico y suele estar relacionado principalmente a factores geográficos entre otros. Santos [14] describe en su artículo el desarrollo de un tablero, en el que se utilizan diferentes técnicas de visualización geoespacial para determinar la importancia de distintos factores en la mortalidad por neumonía, por otro lado, Labrač [15] muestra de una metodología similar para analizar la accesibilidad a centros asistenciales primarios mediante el uso de técnicas de minería de datos para la extracción de información de las bases de datos de distintos centros ambulatorios con el fin de construir un mapa el cual categoriza el nivel de acceso de la población a los centros de atención primaria en cada zona de la región evaluada en el estudio, los dos tipos de tableros mencionados son del tipo estratégico y están orientados a la adopción de políticas públicas de largo plazo. Por otra parte, desde el comienzo de la pandemia de COVID-19 surge la necesidad de mantener un monitoreo constante de la evolución de la enfermedad, en respuesta a esta necesidad se crearon tableros de visualización operacionales con datos en tiempo real de información geoespacial a nivel global. Un ejemplo de esto es el Coronavirus Resource Center (CRC), creado por la universidad Johns Hopkins, el cual muestra una serie de estadísticas sobre COVID-19, tales como, casos activos y muertes el cual se observa en la figura 2.1.

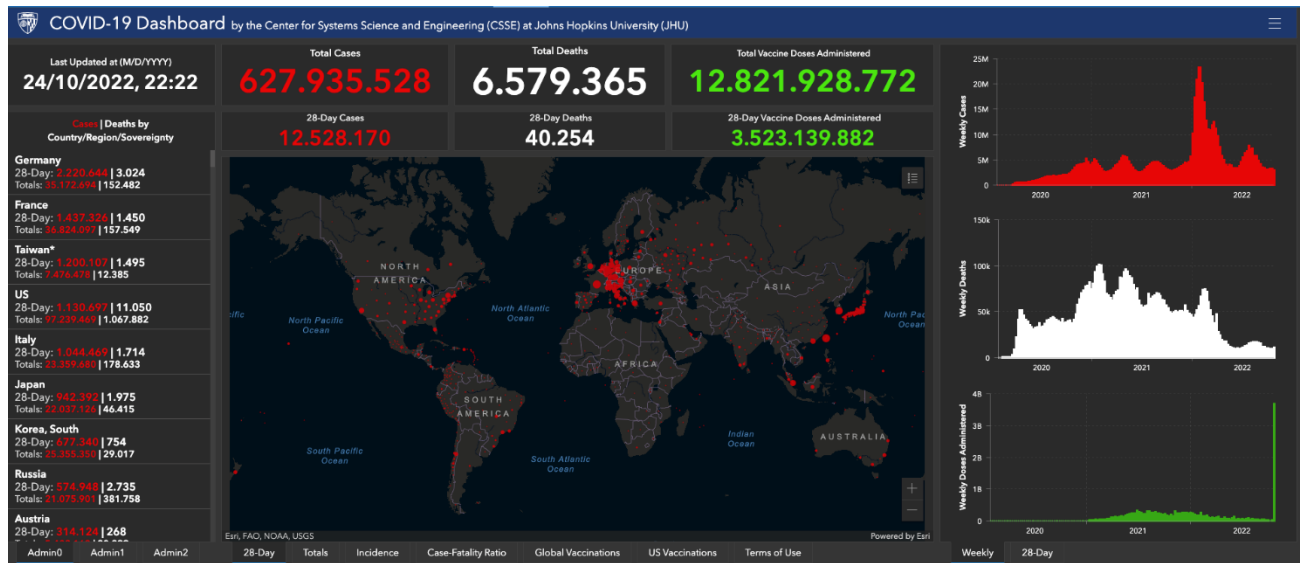


Figura 2.1 Dashboard del Coronavirus Resource Center de la Johns Hopkins University, recuperado de [16]

A nivel institucional los tableros de visualización pueden ser tanto operacionales como estratégico-tácticos. Estos tableros suelen estar orientados a la gestión de recursos y la búsqueda de eficiencias en los sistemas de la institución, como es el caso del tablero de Pestana [12] el cual se muestra en la figura 2.2.

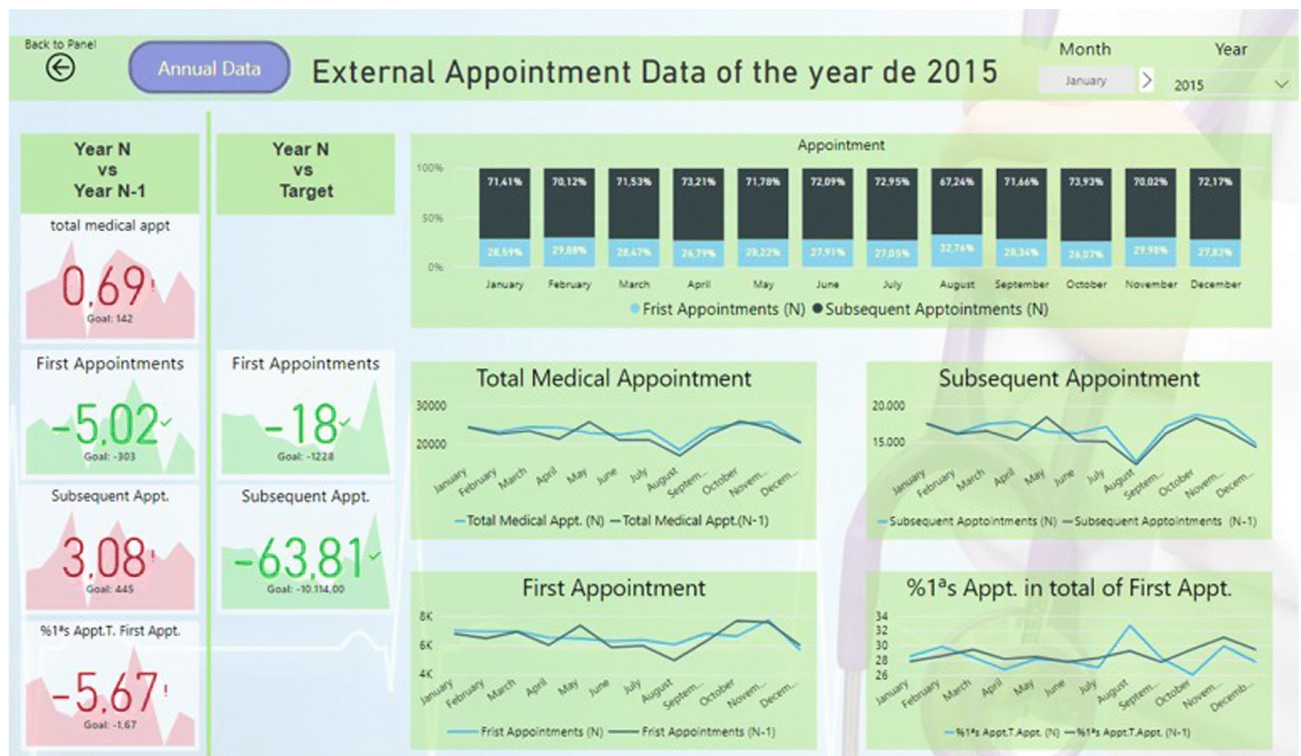


Figura 2.2 Tablero de KPIs anual para un centro de salud el cual muestra métricas relacionadas al número de citas para ese periodo, recuperado de [12]

A nivel de individual los tableros suelen ser más bien estratégicos. Estos tableros se centran en el análisis y evolución de variables fisiológicas, con el fin de pesquisar anomalías que puedan ser indicativas de alguna patología o puedan suponer un factor de riesgo. Estos tipo de tableros han experimentado un auge debido a la masificación de los wereables y el internet de las cosas (IoT) que permiten obtener más mediciones sobre parámetros fisiológicos del paciente. Ejemplos de lo anterior son los sistemas Apple Health o Google fit.

2.3. Marco Teórico

En la siguiente sección se presentan los conceptos más relevantes relacionados a la construcción de un tablero de visualización de datos en salud:

2.3.1 Procesamiento de textos

A continuación se definen algunos conceptos relacionados al procesamiento de texto los cuales se utilizan para la extracción de información de los campos no estructurados de los EMR:

- **Tokenización:** es el proceso de dividir un texto en unidades llamadas tokens. Los tokens pueden ser palabras sueltas, frases o incluso oraciones enteras. En el proceso de tokenización pueden descartarse algunos caracteres, como los signos de puntuación [17].
- **Lematización:** Lematización es una técnica en la recuperación de datos en los sistemas de información (RDSI), esta técnica sirve para reducir variantes morfológicas de la formas de una palabra a raíces comunes o lexemas; con el fin de mejorar la habilidad de los motores de búsqueda para mejorar las consultas en documentos. Básicamente, este consiste en remover el plural, el tiempo, o los atributos finales de la palabra [18].
- **Normalización:** es el proceso en el cual se eliminan caracteres especiales de un texto, tales como tildes u otros caracteres no alfanuméricos.

2.3.2 Tableros

Un tablero es una forma de mostrar varios tipos de datos visuales en un solo lugar. Por lo general, un tablero está destinado a transmitir información diferente pero relacionada, en una forma fácil de interpretar [19]. Según el propósito específico existen tres tipos de tableros: operacional, estratégico-ejecutivo y analítico [9]. A continuación, se describe brevemente cada uno:

- **Tablero operacional:** Prioriza el monitoreo más que el análisis y la administración. Muestra datos que facilitan la parte operativa de un negocio, permitiendo realizar seguimiento de los

procesos y sectores de la organización, con la finalidad de tomar decisiones que aumenten el rendimiento de la institución.

- Tablero Estratégico-Ejecutivo: Proveen los KPI para lograr los objetivos de ejecución, alineando los grupos de trabajo hacia una misma dirección.
- Tablero Analíticos: Comparan el desempeño de un área o proyectos, los planes, pronósticos o resultados.

2.3.3 Visualización de datos

La visualización de datos es el proceso de transformación de la información, para generar representaciones visuales de manera que mejore el razonamiento analítico.

Existen diferentes técnicas de visualización de datos, [20] las clasifica según su propósito comunicativo en las siguientes categorías:

- Comparar categorías: facilita la comparación en términos absolutos o relativos de dos o más valores categóricos.
- Mostrar jerarquías y relaciones de una parte por el todo: provee un desglose de distintas categorías en relación a una población o como elementos constitutivos de estructuras jerárquicas.
- Mostrar cambios a lo largo del tiempo: Se vale del use de series temporales para establecer tendencias y patrones en los valores a lo largo del tiempo.
- Graficar conexiones y relaciones: Evalúa la asociación, distribución y patrones que existen entre distintas variables de un conjunto de datos.
- Mapeo de datos geospaciales: presenta conjuntos de datos con propiedades geospaciales a través de diferentes marcos de mapeo.

La elección de la técnica de visualización dependerá en buena medida del propósito comunicativo del caso en particular, la tabla 2.1 muestra una lista de las técnicas de visualización más comunes y el propósito comunicativo al que mejor se ajustan.

2.3.4 Registro medico electrónico

El registro médico electrónico o EMR por sus siglas en inglés, es un expediente en formato digital de los datos de salud propios de un individuo, que incluye el estado de la salud en distintos momentos de la vida (del individuo o del grupo) y las acciones que se desarrollaron para determinarlo y/o modificarlo. [21]

Tabla 2.1 Categorización de métodos de visualización

<i>Método de visualización</i>	<i>Propósito comunicativo</i>
grafica de puntos	comparación de categorías
gráfico de barras	comparación de categorías
gráfico de barra flotante	comparación de categorías
Histograma	comparación de categorías
gráfico de pendiente	comparación de categorías
nube de palabras	comparación de categorías
gráfico de torta	mostrar jerarquías
gráfico de barras apiladas	mostrar jerarquías
mapa de árbol	mostrar jerarquías
burbujas de jerarquía	mostrar jerarquías
jerarquía de árbol	mostrar jerarquías
gráfico de línea	cambios a lo largo del tiempo
gráfico de área	cambios a lo largo del tiempo
gráfico de áreas apiladas	cambios a lo largo del tiempo
mapa de flujo	cambios a lo largo del tiempo
gráfico de dispersión	Graficar conexiones y relaciones
mapa de calor	Graficar conexiones y relaciones
diagrama de red	Graficar conexiones y relaciones
mapa coroplético	Mapeo de datos geoespaciales
mapa isarítmico	Mapeo de datos geoespaciales
cartograma	Mapeo de datos geoespaciales

2.3.5 Business Intelligence

Business intelligence (BI es un conjunto de estrategia y herramientas que se alimenta de datos de negocios y presentados en la forma de paneles, tablas y gráficos de forma amigable para el usuario, con el fin de permitir a la empresa tomar decisiones basadas en evidencia [22]. El BI ha cobrado importancia en los últimos años, debido a que diferentes organizaciones, tanto públicas como privadas han adoptado este tipo de estrategias en sus procesos de toma de decisiones con un buen grado de éxito. Según Martin [23] hay cuatro niveles de consideración al evaluar las capacidades de BI:

- Informes: Un proceso automatizado que captura datos y produce un informe para el usuario final.
- Tablero y tarjetas de puntuación: las tarjetas de puntuación muestran el progreso consecutivo a lo largo del tiempo hacia metas y objetivos futuros. Los tableros indican el estado en un momento específico.

- **Análisis:** En este nivel, BI brinda información consolidada, para análisis tendencias con el fin de desarrollar nuevos conocimientos basados en datos sin procesar.
- *Analytics:* El descubrimiento y la comunicación de patrones significativos en los datos.

Por otra parte, Negash [24] señala una serie de aspectos a considerar en la construcción de un modelo de BI proactivo, estos son:

- *Data warehousing* en tiempo real: Es un almacén electrónico donde generalmente una empresa u organización mantiene una gran cantidad de información, el que debe ser accesible en tiempo real.
- Minería de datos: proceso mediante el que se intenta descubrir patrones o tendencias sobre un gran volumen de datos.
- Detección automática de anomalías y excepciones: El modelo debe ser capaz de detectar tendencias que se salen de la norma.
- Alertas proactivas con determinación automática de destinatarios: Relacionado al punto anterior, en caso de tener múltiples partes interesadas o *stakeholders*, el proceso debe ser capaz de determinar cuál o cuáles de todos ellos están interesados en determinados comportamientos dentro del proceso.
- Aprendizaje automático y refinamiento: Implica la utilización de algoritmos y técnicas computacionales para examinar y extraer datos relevantes de extensas cantidades de texto. En una primera etapa, se recopilan y preparan los datos, los cuales pueden contener documentos, publicaciones o mensajes. Posteriormente, se implementan algoritmos de Aprendizaje Automático para entrenar modelos con la capacidad de identificar patrones y relaciones en los textos.
- Sistemas de información geográficos: Un modelo de BI completo debe considerar información sobre la ubicación geográfica donde la información es recabada.
- Visualización de datos: Es el último paso del proceso de construcción de un modelo de BI, alude a la representación gráfica y visual de información y patrones extraídos de conjuntos de datos. Esta estrategia utiliza recursos gráficos y otras representaciones visuales con el objeto de realzar tendencias, relaciones y atributos cruciales de los datos.

2.4. Discusión

Con la información recopilada en el análisis bibliográfico es posible establecer que los tableros de visualización de datos son una herramienta útil para los encargados de toma de decisiones en distintas industrias. En cuanto a la industria de la salud, su uso va más allá de ser una herramienta para la toma de decisiones estratégico-tácticas, ya que el poder analizar los datos generados por una institución de salud, permiten conocer patrones epidemiológicos y a la vez permiten tomar decisiones que afecten la atención, basándose en evidencia.

Por todo lo anterior se considera que un tablero de visualización de datos a partir de los EMR del HGGB puede ser una herramienta que complemente la toma de decisiones administrativas del centro de salud en pos de mejorar la eficiencia y calidad de la atención brindada.

Capítulo 3. Materiales y Métodos

3.1. Introducción

En el presente capítulo se describirá la base de datos de EMR de pacientes del HGGB en las distintas subespecialidades analizadas, haciendo una descripción de los campos que componen los EMR, también se describirán las herramientas de software utilizadas en la construcción del tablero de métricas así como el proceso de preprocesamiento de los datos no estructurados de los EMR.

3.2. Materiales

Para este estudio, se consideró un total de 268.993 EMRs provenientes de atenciones médicas realizadas en el HGGB durante los años 2011 a 2012. Estos documentos pertenecen a atenciones médicas relacionadas a 9 sub-especialidades: i) broncopulmonar, ii) cardiología, iii) cirugía, iv) cirugía de mamas, v) endocrinología, vi) gastroenterología, vii) medicina interna, viii) reumatología.

Cada uno de estos documentos, esta subdividido en campos estructurados y no estructurados, que corresponden a la interacción narrativa entre médicos y pacientes realizada en cada una de las consultas. Los campos estructurados corresponden a opciones dentro del EMR donde el médico puede informar sobre determinados hábitos o antecedentes mórbidos del paciente. Estos son datos categóricos que se utilizan, para indicar presencia o ausencia de dichos antecedentes. Si bien estos campos estructurados, permiten informar las condiciones de forma no ambigua, imposibilitan al médico de entregar mayor detalle de la condición del paciente. Los campos no estructurados corresponden a información recabada por el médico durante la atención y que es reportada en la forma de lenguaje escrito narrativo. Los campos no estructurados que serán utilizados en este trabajo son:

- Anamnesis: Corresponde a toda la información relatada por el paciente hacia el personal médico. En este campo principalmente se encuentran síntomas y conductas informadas por el paciente.
- Indicaciones: Corresponden a indicaciones que da el personal médico al paciente de acuerdo con las acciones necesarias que este debe efectuar para tratar su condición.
- Diagnóstico: Corresponde al registro de las condiciones y/o enfermedades que el personal médico pudo constatar en el paciente.

3.2.1 Herramientas de software

Para el desarrollo del tablero de visualización de datos se utilizaron principalmente 3 herramientas de software distintas, cada una para cada componente del proceso, PostgreSQL como gestor de base de datos, la librería Natural Language Toolkit (NLTK) de Python para el procesamiento del texto y Apache Superset como interfaz visual del proyecto.

PostgreSQL es un sistema de gestión de bases de datos relacional de código abierto y altamente escalable. Se utiliza para almacenar, organizar y manipular grandes cantidades de datos de manera eficiente y segura. PostgreSQL proporciona una amplia gama de características y funcionalidades avanzadas, como soporte para consultas complejas, transacciones ACID, integridad referencial, replicación, particionamiento de datos y muchas más.

NLTK es una librería de Python que se utiliza para el procesamiento de NLP. Proporciona un conjunto de herramientas y recursos que permiten a los desarrolladores trabajar con texto y realizar tareas relacionadas con el procesamiento del lenguaje humano. La librería NLTK ofrece una amplia gama de funcionalidades, como tokenización, etiquetado gramatical, lematización, análisis de sentimientos, extracción de información, entre otros. Además, proporciona acceso a diversos corpus y modelos lingüísticos pre-entrenados que ayudan en el desarrollo de aplicaciones de NLP.

Apache Superset es una aplicación nativa de la nube de software de código abierto para la exploración y visualización de datos, capaz de manejar datos a escala de petabytes. Esta herramienta compondrá la interfaz de usuario a utilizar para el visualizador de datos. Las características que impulsaron la elección de esta herramienta son la capacidad de integrar de forma nativa una amplia variedad de bases de datos relaciones, entre ellas el sistema PostgreSQL elegido para el *data warehouse*, así como sus capacidades para generar componentes de visualización, tanto por medio de la interfaz gráfica como por consultas SQL. Apache Superset resultase bastante versátil y amigable para generar graficas que requieran del cruce de información o cálculo de funciones más complejas.

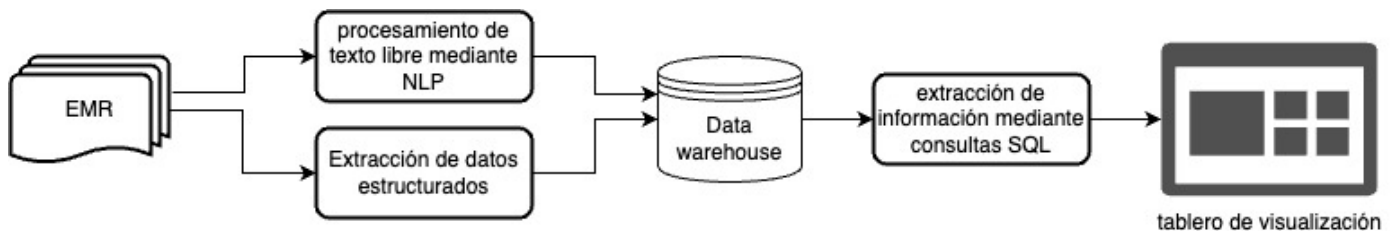


Figura 3.1 Diagrama de componentes del modelo de visualización propuesto

3.3. Métodos

Para lograr el objetivo de implementar un tablero de visualización se construyó un sistema separado en 3 componentes principales tal como se muestra en la figura 3.1. los cuales son procesamiento y extracción, el *data warehouse* y la plataforma o tablero de visualización.

3.3.1 Procesamiento del texto

Se tomó cada campo de texto libre de los EMR y se realizó un preprocesado del contenido, que consideró los siguientes pasos:

- **Tokenización:** Se dividió el texto extraído en tokens que representaban una palabra del texto cada uno.
- **Eliminación de caracteres especiales:** Se eliminaron del texto procesado todos los caracteres no alfanuméricos contenidos en ellos, tales como puntos, comas, signos de exclamación, etc.
- **Lemantización:** Cada palabra fue reducida a su lema base para favorecer la agrupación de términos de la misma familia de palabras.
- **Normalización:** También con el fin de favorecer la agrupación de términos se reemplazaron todas las letras mayúsculas presentes en el texto por su equivalente en minúscula.

3.3.2 Creación del data warehouse

Con la finalidad de almacenar la información procesada en el apartado anterior, se creó una base de datos que contiene una tabla principal, en la cual se encuentran los campos estructurados de los EMR, tales como edad, fecha de atención, sexo, etc. y una tabla por cada campo no estructurado de los EMR que contiene los tokens presentes en dicho campo. Estas tablas de tokens se relacionaron con la tabla principal utilizando un id de registro ficticio agregado a cada EMR, previo a su separación en tokens. La estructura de la base de datos se puede visualizar en la figura 3.2.

3.3.3 Construcción de medios gráficos

Finalmente se utilizó la información contenida en el *data warehouse*, mencionado en el apartado anterior, para alimentar una serie de medios gráficos que componen el tablero por especialidad. Este

tablero se observa en la figura 4.1. y la figura 4.2. Los gráficos que se muestran en el tablero son los siguientes:

- Cantidad de pacientes atendidos en el periodo de tiempo consultado
- Edad promedio de los pacientes
- 10 factores de riesgo más comunes
- Proporción de genero de los pacientes
- 10 diagnósticos más comunes de la especialidad
- Serie temporal de la distribución de las atenciones ambulatorias en el rango de tiempo consultado
- Proporción de los principales hábitos relacionados a los diagnósticos más comunes
- Proporción de los principales antecedentes relacionados a los diagnósticos más comunes

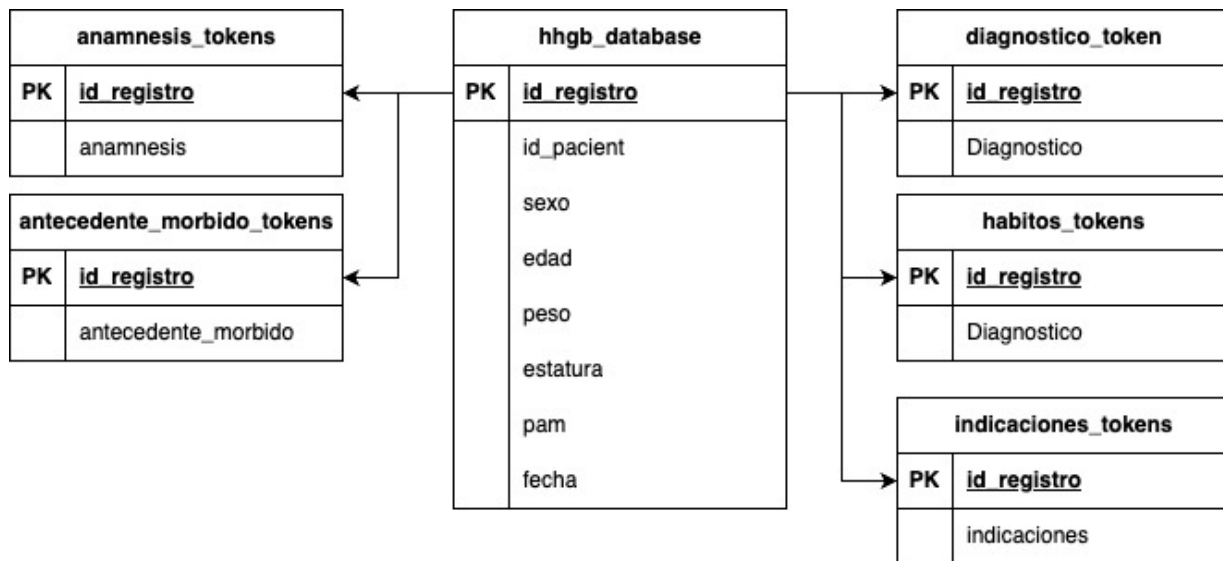


Figura 3.2 Esquema de la base de datos utilizada como data warehouse

3.3.4 Encuesta de evaluación

Para validar la funcionalidad de la herramienta de visualización creada se condujo una encuesta de evaluación la cual constó de las siguientes preguntas.

1. ¿Tiene usted experiencia utilizando herramientas de visualización? (ej. PowerBI, Tableau, etc.)
2. En una escala del 1 al 10, ¿qué tan fácil es navegar y encontrar la información que necesitas en el dashboard?
3. ¿Qué aspecto del dashboard modificarías para mejorar su navegabilidad?
4. En una escala del 1 al 10, ¿qué tan comprensible es la información presentada en el dashboard?
5. ¿Qué aspecto del dashboard modificarías para hacer la información presentada más comprensible?
6. ¿Hay algún problema o error que hayas experimentado al utilizar el dashboard?
7. ¿Hay alguna información adicional que te gustaría ver en el dashboard?
8. ¿Qué aspectos del dashboard crees que podrían mejorarse?
9. ¿Qué tan satisfecho estás con la apariencia y diseño del dashboard? (muy insatisfecho, insatisfecho, neutro, satisfecho, muy satisfecho).
10. ¿Qué características encuentras más útiles en el dashboard?

Capítulo 4. Resultados

4.1. Introducción

En este capítulo se describen los resultados obtenidos en el procesamiento de los datos que llevó a la construcción de los tableros de visualización descritos en el capítulo anterior, además se mencionan las particularidades que se pudieron observar al analizar los diferentes gráficos generados y la información contenida en estos.

4.2. Resultados

De los gráficos generados para cada especialidad se pudo establecer una edad promedio de los pacientes de cada subespecialidad el cual se puede ver en la tabla 4.1. De esta información, se pudo observar que la mayoría de subespecialidades excepto endocrinología adulto, atienden en su mayor parte pacientes por sobre los 55 años de edad.

Por otra parte, los antecedentes y hábitos más reportados por los médicos son respectivamente la hipertensión y el tabaquismo, como se puede apreciar en la tabla 4.3 y tabla 4.2 , respectivamente.

Los antecedentes recopilados del análisis de los gráficos generados sugieren que en su mayoría los médicos de distintas especialidades reportan los mismos antecedentes y hábitos presentes en los pacientes. Esto puede deberse a una alta incidencia de estos factores en la población o un subreporte de otros factores que no se toman en cuenta a la hora de escribir las fichas clínicas de los pacientes, sin embargo, cualquiera de estas hipótesis debe ser corroborada con un análisis más profundo.

Tabla 4.1 Edad promedio de los pacientes por especialidad

<i>Subespecialidad</i>	<i>Edad promedio del paciente</i>
broncopulmonar adulto	62
cardiología adulto	67
cirugía adulto	57
cirugía de mamas	57
endocrinología adulto	50
gastroenterología adulto	56
medicina interna	54
reumatología	55

Tabla 4.2 Hábito más reportado por especialidad

<i>Subespecialidad</i>	<i>Habito reportado</i>	<i>Cantidad de apariciones</i>
broncopulmonar adulto	tabac	5773
cardiología adulto	tabac	5743
cirugía adulto	tabac	1766
cirugía de mamas	tabac	672
endocrinología adulto	tabac	2585
gastroenterología adulto	tabac	1288
medicina interna	tabac	294
reumatología	tabac	1155

Tabla 4.3 Antecedente más reportado por especialidad

<i>subespecialidad</i>	<i>Antecedente reportado</i>	<i>Cantidad de apariciones</i>
broncopulmonar adulto	hipertensión	6280
cardiología adulto	hipertensión	3963
cirugía adulto	hipertension	6734
cirugía de mamas	hipertension	3130
endocrinología adulto	diabet	7923
gastroenterología adulto	hipertensión	3471
medicina interna	hipertensión	3398
reumatología	hipertensión	4607

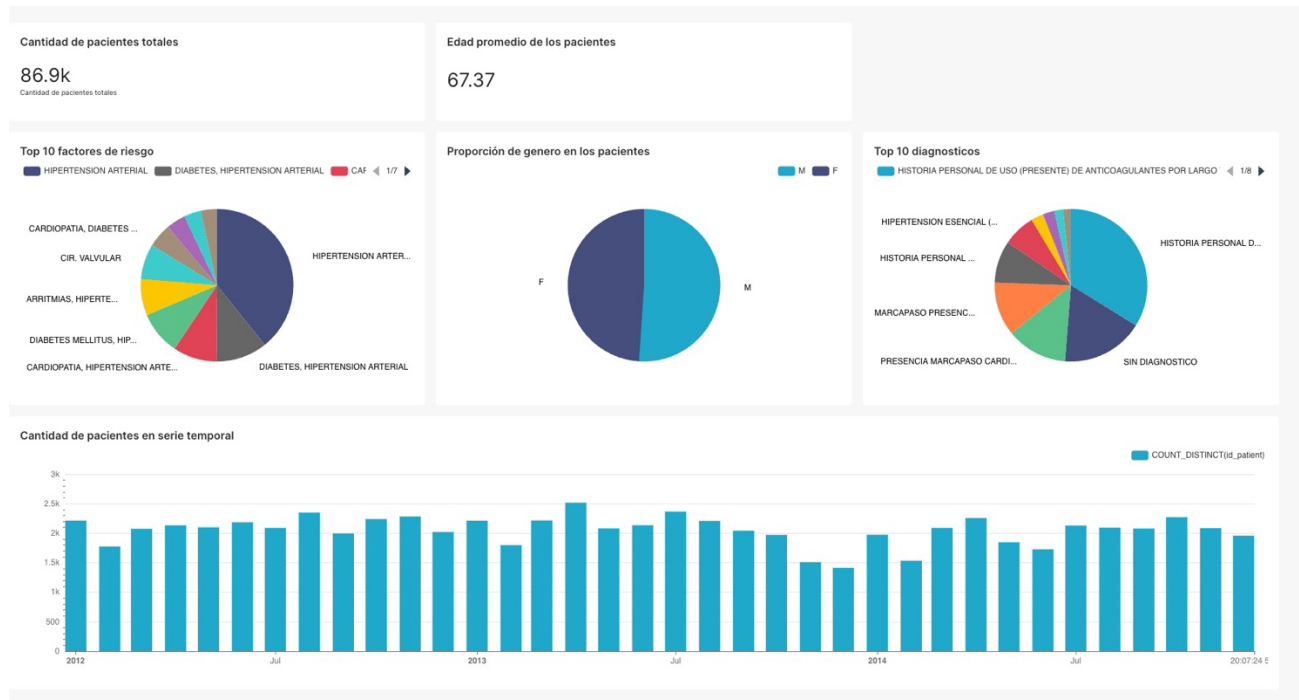


Figura 4.1 Tablero de métricas construido filtrado por especialidad de cardiología

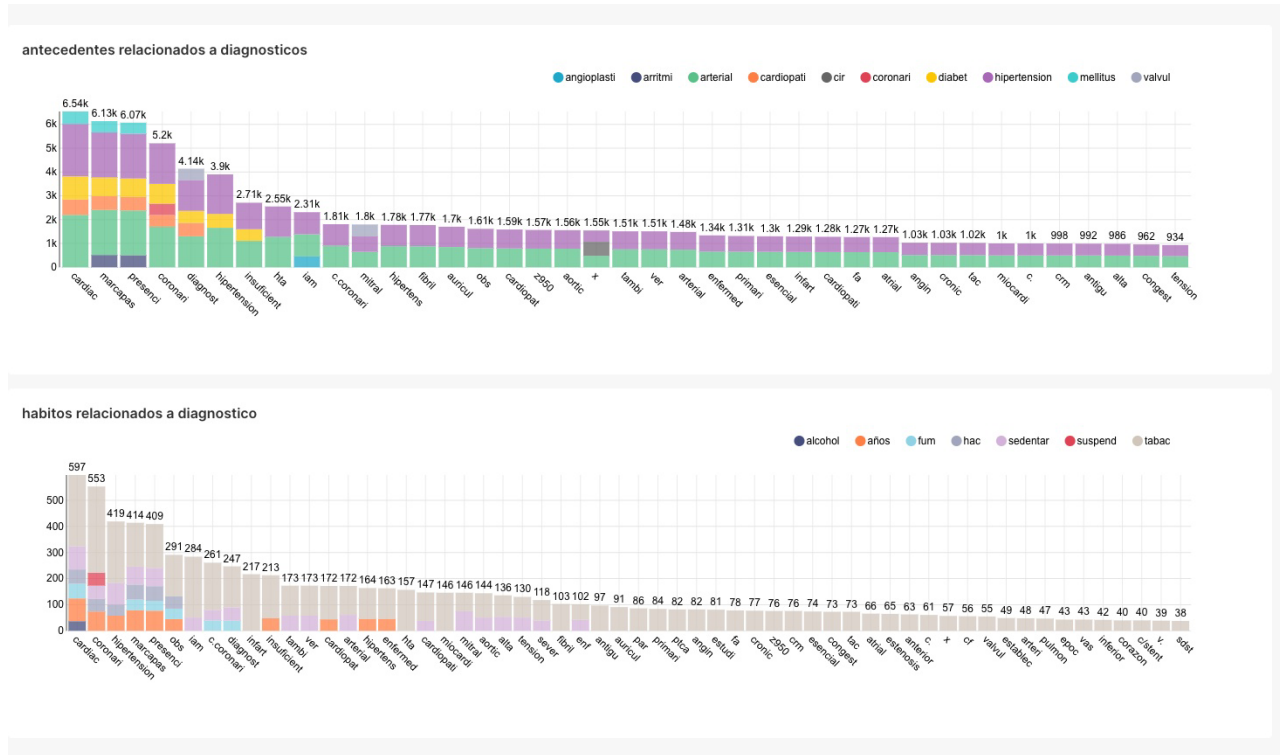


Figura 4.2 Gráficos de relación de términos de antecedentes y hábitos con diagnósticos filtrado por especialidad cardiología adulto

4.3. Evaluación

Para validar el modelo propuesto este se presentó a 3 evaluadores todos ellos estudiantes de la carrera de ingeniería civil biomédica, a los cuales se les dio una breve introducción del proyecto y una descripción de la información que contienen los EMR de la base de datos, luego de ello se les presentó la plataforma la cual pudieron navegar libremente con la finalidad de conducir una encuesta de evaluación finalizada la prueba cuyos resultados se muestran en la tabla 4-4.

Dos de los tres evaluadores declararon haber tenido algún nivel de experiencia previa en softwares de visualización (Power BI, Tableau u otros), Al ser consultados por la apariencia y diseño en general del dashboard los 3 evaluadores señalaron estar muy satisfechos, por otra parte al preguntarles cual es el principal punto de mejora dos de ellos señalaron que podrían mejorarse las descripciones de la información presentada, a lo que uno de los evaluadores acotó que sería de utilidad contar con un cuadro contextual que explicara la información en los gráficos.

Tabla 4.4 Resultados de la encuesta de evaluación

<i>Nro. de pregunta</i>	<i>Evaluador 1</i>	<i>Evaluador 2</i>	<i>Evaluador 3</i>
1	No	Si	Si
2	8	9	8
3	Nada	Podría mejorarse el acceso a los filtros	Nada
4	7	7	8
5	Agregaría una descripción de los grafico	Podrían mejorarse algunos títulos	Se podría añadir una breve descripción
6	No	No	No
7	Un gráfico para hacer seguimiento a las atenciones de un paciente	No más información pero si otros tipos de gráficos más llamativos	Un gráfico de diagnósticos segmentado por edad
8	Agregar descripciones	Títulos más descriptivos	Gráficos más interactivos
9	Muy satisfecho	Muy satisfecho	Muy satisfecho
10	Los filtros para los gráficos son bastante útiles	Tener usuarios con distintos permisos	Los filtros están bien escogidos

Capítulo 5. Conclusiones

5.1. Introducción

En este último capítulo se llevará a cabo un análisis del proceso realizado para la construcción de un visualizador de datos basado en registros médicos electrónicos y se propondrán guías para el mejoramiento del modelo propuesto.

5.2. Conclusiones

El tablero propuesto por especialidad tiene como objetivo establecer un perfil del paciente de cada especialidad, que permita reconocer los principales factores que puedan estar correlacionados a ciertas condiciones médicas en la población en estudio. El tablero implementado permitirá establecer patrones que permitan generar estrategias para prevenir condiciones clínicas en los pacientes. En la elaboración de los tableros, se pudo observar que indicadores como los antecedentes mórbidos de los pacientes varían poco entre especialidades, al menos en los factores más reportados por los médicos, sin embargo, un número significativo de pacientes no presentan información sobre antecedentes mórbidos. Lo anterior puede deberse a que otros factores menos comunes se estén dejando fuera de los EMR.

La principal dificultad ha sido el extenso y variado uso de abreviaturas por parte de los médicos en el proceso de registro de información, ya que no presentan en general un estándar y varían en su utilización de especialidad en especialidad. El agregar procesos de procesamiento de datos capaces de extraer esta información relevante podría aumentar significativamente la cantidad de información extraída de la base de datos, enriqueciendo a su vez los tableros de información desarrollados en este trabajo y haciendo más evidentes ciertas correlaciones en la información.

5.3. Trabajo Futuro

Para mejorar la calidad de la información mostrada por el tablero de visualización un posible enfoque sería trabajar en la agrupación y etiquetado de términos relacionados en la base de datos, como por ejemplo las palabras “fumar” y “tabaquismo”. Otro punto a considerar son las abreviaturas propias del lenguaje médico presentes en buena parte de los registros analizados, una correcta interpretación de esta información podría mejorar el análisis visual significativamente.

Capítulo 6. Glosario

ACID	: Atomicidad, consistencia, aislamiento y durabilidad (Atomicity, Consistency, Isolation and Durability)
BI	: Inteligencia de negocio (Business Intelligence)
BSC	: Cuadro de mando integral (Balanced scorecard)
CRC	: Coronavirus Resource Center
EMR	: Registros médicos electrónicos (Electronic Medical Records)
ETL	: extracción, transformación y carga (Extraction, transformation, load)
HGGB	: Hospital Guillermo Grant Benavente
HIT	: Tecnología de la información sanitaria (Health information technologies)
KPI	: Indicador clave de desempeño (Key performance indicator)
NLP	: Procesamiento natural del lenguaje (Natural language processing)

Capítulo 7. Referencias

- [1] Y. Alotaibi y F. Federico, “The impact of health information technology on patient safety,” *Saudi Medical Journal*, vol. 38, n° 12, pp. 1173-1180, 2017. [En línea]. Disponible en: <https://doi.org/10.15537/smj.2017.12.20631>
- [2] W. Sun , Z. Cai, Y. Li, F. Liu, S. Fang y G. Wang, “Data processing and text mining technologies on electronic medical records: a review,” *Journal of Healthcare Engineering*, vol. 2018, p. 9, 2018. [En línea]. Disponible en: <https://doi.org/10.1155/2018/4302425>
- [3] D. Keim, J. Kohlhammer, G. Ellis y F. Mannsmann, *Mastering the information age. Solving problems with visual analytics*, Goslar, 2010, p. 10. [En línea] Disponible en: <https://doi.org/10.2312/14803>
- [4] powerdata, “Big Data: ¿En qué consiste? Su importancia, desafíos y gobernabilidad,” [En línea]. Available: <https://www.powerdata.es/big-data>. [Último acceso: 2022 diciembre 10].
- [5] Z. Obermeyer y E. Emanuel, “Predicting the future - Big data, machine learning and clinical medicine,” *New England Journal of Medicine*, vol. 375, n° 13, pp. 1216-1219, 2016. [En línea]. Disponible en: <https://doi.org/10.1056/NEJMp1606181>
- [6] A. Rajkomar, J. Dean y I. Kohane, “Machine learning in medicine,” *New England Journal of medicine*, n° 380, pp. 1347-1358, 2018. [En línea]. Disponible en: <https://doi.org/10.1056/NEJMra1814259>
- [7] C. Cutillo, K. Sharma, L. Foschini, S. Kundu, M. Mackintosh y K. Mandl, “Machine intelligence in healthcare-- prespectives on trustworthiness, explainability, usability and transparency,” *Neurocomputing*, n° 349, pp. 127-136, 2019. [En línea]. Disponible en: <https://doi.org/10.1038/s41746-020-0254-2>
- [8] K. Lee, S. Haneuse, D. Schrag y F. Dominici, “Bayesian Semi-parametric analysis of semi-compelling risk data: investigating hospital readmission after pancreatic cancer diagnosis,” vol. 64, n° 2, pp. 253-273, 2015. [En línea]. Disponible en: <https://doi.org/10.1111/rssc.12078>
- [9] S. Bansal y G. Chowell, “Big Data for Infectious Disease Surveillance and Modeling,” *Journal of Infectious Diseases*, vol. 214, pp. 375-379, 2016. [En línea] disponible en: <https://doi.org/10.1093/infdis/jiw400>
- [10] S. Senthilkumar, K. R. Bharatendara y A. M. Amruta , “Big Data in Healthcare Management: A Review of Literature,” *American Journal of Theoretical and Applied Business*, pp. 57-69 , 2018. [En línea] disponible en: <https://doi.org/10.11648/j.ajtab.20180402.14>
- [11] D. A. G. Frazão, T. S. Antunes da Costa, T. D. Oliveira de Araújo y C. G. Resque dos Santos, “2021 25th International Conference Information Visualisation,” de A brief review of dashboard visualizations employed to support management or business decisions, 2021. [En línea] disponible en: <https://doi.org/10.1109/IV53921.2021.00025>
- [12] M. Pestana, R. Pereira y S. Moro, “Improving Health Care Management in Hospitals Through a Productivity Dashboard,” *Journal of Medical Systems*, vol. 44, n° 47, 2020. [En línea] disponible en: <https://doi.org/10.1007/s10916-020-01546-1>
- [13] M. B. Gonzalez-Sanchez, L. Broccardo y A. M. Martins Pires, “The use and design of the BSC in the health caresector: A systematic literature review for Italy,Spain, and Portugal,” *Int J Health Plann Mgmt*, vol. 33, p. 6– 30, 2018. [En línea] disponible en: <https://doi.org/10.1002/hpm.2415>

- [14] M. Y. Santos, A. Carvalheira y A. Teles de Araujo, "A data-driven analytics approach in the study of pneumonia's fatalities," de 2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA), Paris, 2015. [En línea] disponible en: <https://doi.org/10.1109/DSAA.2015.7344868>
- [15] N. Lavrač, M. Bohanec y A. Pur , "Data mining and visualization for decision support and modeling of public health-care resources," *Journal of Biomedical Informatics*, vol. 40, pp. 438-447, 2007. [En línea] disponible en: <https://doi.org/10.1016/j.jbi.2006.10.003>
- [16] Johns Hopkins University & Medicine, «Coronavirus Resource Center,» 2023. [En línea]. Available: <https://coronavirus.jhu.edu/map.html>. [Último acceso: 19 Octubre 2023].
- [17] Great Learning Education Services, «Great Learning,» 22 Marzo 2022. [En línea]. Available: <https://www.mygreatlearning.com/blog/tokenization>. [Último acceso: 19 Octubre 2023].
- [18] J. Peinado, «Lematización para palabras médicas complejas: Implementación de un algoritmo en LISP.,» *Revista Medica Herediana*, vol. 14, n° 4, pp. 224-229, 2003.
- [19] Tableau, "What is a dashboard? A complete overview," [En línea]. disponible en: <https://www.tableau.com/learn/articles/dashboards/what-is>. [Último acceso: 25 octubre 2022].
- [20] A. Kirk, *Data Visualization: a Successful Design Process : a successful design process*, Packt Publishing, Limited, 2012.
- [21] Organización Panamericana de la salud, "Registros Médicos Electrónicos," [En línea]. disponible en: <https://www3.paho.org/relacsis/>. [Último acceso: 10 diciembre 2022].
- [22] IBM, "ibm.com," [En línea]. disponible en: <https://www.ibm.com/cl-es/topics/business-intelligence>. [Último acceso: 1 diciembre 2022].
- [23] Martin, "Better decision making with proper business intelligence," 2011. [En línea]. disponible en: <https://www.kearney.com/>. [Último acceso: 1 diciembre 2022].
- [24] S. Negash, "Business Intelligence," *Communications of the association of information systems*, vol. 13, 2004. [En línea] disponible en: <https://doi.org/10.17705/1CAIS.01315>
- [25] H. Kerzner, *Project management metrics, KPIs, and dashboards: A guide to measuring and monitoring project performance*, 2013. [En línea] disponible en: DOI:10.1002/9781119427599

Anexo A. Extractos de código

A continuación se presentan los scripts utilizados para el preprocesamiento de los campos de texto libre de la base de datos y conexión con las bases de datos.

A.1 Script de preprocesamiento

```
import nltk
from nltk import SnowballStemmer
from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords
import string

nltk.download('punkt')
nltk.download('stopwords')
nltk.download('wordnet')

# Pasos de preprocesamiento
def preprocess_text(text):
    # Tokenización
    tokens = word_tokenize(text)

    # Eliminación de puntuación y caracteres especiales
    tokens = [token for token in tokens if token not in string.punctuation]

    # Eliminación de números
    tokens = [token for token in tokens if not token.isdigit()]

    # Eliminación de stopwords
    stop_words = set(stopwords.words('spanish'))
    tokens = [token for token in tokens if token.lower() not in stop_words]

    # Lematización
    stemmer = SnowballStemmer('spanish')
    tokens = [stemmer.stem(token) for token in tokens]

    # Normalización de texto
    tokens = [token.lower() for token in tokens]

    return tokens
```

A.2 Script de conexión a base de datos

```
import psycopg2

from preprocessing_service import preprocess_text

def insert_preprocessed_column(target_colum, record_list):
    # Conexión a la base de datos
    conn = psycopg2.connect(database="mt", user="abenavente", password="",
host="localhost", port="5432")
    cursor = conn.cursor()
    query = f"INSERT INTO hhgb_database_preprocessed_text ({target_colum},
```

```

id_registro) VALUES (%s, %s) " \
    f"ON CONFLICT (id_registro) DO UPDATE SET {target_colum} = %s"

# Ejecutar la inserción masiva
cursor.executemany(query, record_list)

# Confirmar los cambios en la base de datos
conn.commit()

# Cierre de la conexión a la base de datos
cursor.close()
conn.close()

def insert_tokens_column(target_colum, record_list):
    # Conexión a la base de datos
    conn = psycopg2.connect(database="mt", user="abenavente", password="",
host="localhost", port="5432")
    cursor = conn.cursor()
    query = f"INSERT INTO {target_colum}_tokens ({target_colum}, id_registro)
VALUES (%s, %s) " \
    f"ON CONFLICT ({target_colum}, id_registro) DO NOTHING"

    # Ejecutar la inserción masiva
    cursor.executemany(query, record_list)

    # Confirmar los cambios en la base de datos
    conn.commit()

    # Cierre de la conexión a la base de datos
    cursor.close()
    conn.close()

def preprocess_column(target_colum):
    # Conexión a la base de datos
    conn = psycopg2.connect(database="mt", user="abenavente", password="",
host="localhost", port="5432")
    cursor = conn.cursor()

    # Consulta de los registros
    cursor.execute(f"SELECT {target_colum}, id_registro FROM hhgb_database;")
    rows = cursor.fetchall()

    # Preprocesamiento de cada registro
    preprocessed_record_list = []
    for row in rows:
        record_text = row[0]
        if record_text is not None:
            preprocessed_text = preprocess_text(record_text)
            for text in preprocessed_text:
                preprocessed_record_list.append((text, row[1]))

    # Inserción de la lista de registros en tabla destino
    # insert_preprocessed_column(target_colum=target_colum,
record_list=preprocessed_record_list)
    insert_tokens_column(target_colum=target_colum,
record_list=preprocessed_record_list)

```

```
# Cierre de la conexión a la base de datos  
cursor.close()  
conn.close()
```

UNIVERSIDAD DE CONCEPCION – FACULTAD DE INGENIERIA

RESUMEN DE MEMORIA DE TITULO

Departamento : Departamento de Ingeniería Eléctrica
Carrera : Ingeniería Civil Biomédica
Nombre del memorista : Alan Andrés Benavente Toro
Título de la memoria : Visualizador de datos de atenciones ambulatorias del Hospital Guillermo Grant Benavente de Concepción
Fecha de la presentación oral : 20 de octubre de 2023

Profesor(es) Guía : Rosa Figueroa, Christopher Flores
Profesor(es) Revisor(es) : Cesar García, Pamela Guevara
Concepto :
Calificación :

Resumen

El presente trabajo tuvo como finalidad crear un tablero de visualización de datos utilizando información de una base de registros médicos electrónicos provistos por el hospital Guillermo Grant Benavente de Concepción, presentando la información en diferentes medios gráficos con el propósito de apoyar la toma de decisiones. Este visualizador se diseñó y desarrollo con una arquitectura que permitiera la escalabilidad del modelo propuesto y la incorporación de nuevos datos en el futuro.

Como fuente de información del tablero, se utilizaron tanto los campos estructurados como no estructurados de los registros médicos electrónicos. Para la utilización de los campos no estructurados de la base de datos fue necesario realizar un preprocesamiento del texto libre, su tokenización y posterior separación en unigramas, con la finalidad de ser almacenada en un data warehouse creado para alimentar el tablero de visualización.

Para generar los diferentes elementos visuales incluidos en el tablero, se filtraron los registros para cada una de las 8 especialidades de atención ambulatoria. Los elementos visuales representan la siguiente información: i) distribución etaria de los pacientes, ii) distribución en serie temporal de las consultas médicas, iii) hábitos más comunes de los pacientes y iv) la relación de los hábitos y antecedentes de los pacientes con su posterior diagnóstico, entre otros.

Del análisis de los medios gráficos creados con el set de datos disponible para este trabajo, se puede notar que mayoritariamente los hábitos reportados por los médicos en las distintas especialidades son los mismos. Lo anterior, puede tener su causa en distintos factores, tales como, un subreporte de un hábito o antecedente considerado de menor relevancia o bien a una fuerte correlación del hábito o antecedente con patologías de esa subespecialidad. Adicionalmente, se puede comentar que, del análisis visual de la información obtenida, que es posible establecer un perfil de las características de los pacientes para cada especialidad. Finalmente se condujo una evaluación de la plataforma propuesta con ayuda de 3 evaluadores, los cuales coincidieron en estar muy satisfechos con el diseño y apariencia de la plataforma. Sin embargo, 2 de ellos acotaron que podrían mejorarse las descripciones para dar más claridad a la información presentada.