

**REPOSITORIO DE DATOS PARA LA TOMA DE DECISIONES
EN UNA INSTITUCIÓN DE SALUD**

ÁLVARO JOSÉ ORTEGA BRAVO

**Trabajo de grado para optar al título de
Ingeniero Biomédico**

Javier Enrique Camacho Cogollo

Profesor de Ingeniería Clínica

Estudiante de PhD. en Ingeniería



UNIVERSIDAD EIA

INGENIERÍA BIOMÉDICA

ENVIGADO

2018

AGRADECIMIENTO

Al director de trabajo de grado Javier Enrique Camacho Cogollo ya que sus indicaciones encaminaron este trabajo por la dirección correcta.

Al profesor Christian Lochmueller por su tiempo y disponibilidad para atender todas mis dudas para la ejecución de las actividades para la minería de datos y análisis de resultados.

Y por último a mi familia por todos sus esfuerzos por mantenerme estudiando en esta universidad y el apoyo incondicional durante toda la carrera, entendiendo cada situación adversa que se me pudo haber presentado.

CONTENIDO

	pág.
INTRODUCCIÓN	14
1. PRELIMINARES	15
1.1. Planteamiento del problema	15
1.2. Objetivos del proyecto.....	18
1.2.1. Objetivo General	18
1.2.2. Objetivos Específicos	18
1.3. Marco de referencia.....	18
1.3.1. Antecedentes	18
1.3.2. Salud digital o eHealth	21
1.3.3. Los historiales o registros médicos electrónicos integrados.....	22
1.3.4. Minería de datos o Data mining	25
1.3.5. Métodos de Minería de Datos	28
1.3.5.1. Árboles de decisión	28
1.3.5.2. Clustering	30
1.3.5.3. Razonamiento bayesiano	32
1.3.5.4. Redes neuronales	34
1.3.5.5. Reglas de asociación	37
1.3.5.6. Regresión lineal.....	39
1.3.5.7. Regresión logística	40
1.3.5.8. Series de tiempo	42
1.3.7. Business Intelligence.....	44
1.3.8. Análisis de datos en la salud.....	45
2. METODOLOGÍA	48
2.1. Caracterizar los modelos analíticos para repositorios de datos clínicos y técnicos que existen..	48

2.2.	Identificar las fuentes de datos clínicos y técnicos más apropiadas	48
2.3.	Diseñar un repositorio de datos clínicos y técnicos.....	49
2.4.	Evaluar la propuesta de repositorio de datos clínicos y técnicos	50
3.	PRESENTACIÓN Y DISCUSIÓN DE RESULTADOS.....	52
3.1.	SELECCIONAR LAS HERRAMIENTAS A UTILIZAR.....	52
3.1.1.	Almacenamiento y organización de los datos	52
3.1.2.	Diseño de la interfaz gráfica de usuario	56
3.1.3.	ETL y Minería de datos.....	58
3.2.	FUENTES DE DATOS CLÍNICOS Y TÉCNICOS	59
3.3.	ALMACENAMIENTO Y ORGANIZACIÓN DE LOS DATOS	62
3.4.	DISEÑO DE LA INTERFAZ GRÁFICA DE USUARIO	69
3.5.	ETL Y MINERÍA DE DATOS	74
3.6.	IMPLEMENTACION DEL METODO EN UNA INSTITUCION DE SALUD	81
3.6.1.	Fuentes de datos clínicos y técnicos.....	81
3.6.2.	Almacenamiento y organización de los datos	82
3.6.3.	Diseño de la interfaz gráfica de usuario	86
3.6.4.	ETL y Minería de datos.....	88
4.	CONCLUSIONES Y CONSIDERACIONES FINALES	108
	REFERENCIAS	110
	ANEXOS.....	115

LISTA DE TABLAS

	pág.
Tabla 1. <i>Atributos de los datos ofrecidos por entidades de salud (Entidad A & Entidad B, 2018)</i>	82
Tabla 2. <i>Tabla dinámica datos filtrados Cuenta de DESCRIPCION DX CIE 10. (Fuente propia, 2018)</i> . .	92

LISTA DE FIGURAS

	pág.
<i>Figura 1.</i> Representación árbol de decisión (Mitchell, 1997).....	29
<i>Figura 2.</i> Representación clusterización (Pacula, 2011).....	31
<i>Figura 3.</i> Ilustración modelo redes neuronales (Támez, 2016).....	36
<i>Figura 4.</i> Soporte y confianza en reglas de asociación (Monteserin, 2018).	38
<i>Figura 5.</i> Estructura regresión lineal (Santana, 2015).....	40
<i>Figura 6.</i> Estructura de un Cuadrante Mágico de Gartner (Gartner, Inc., 2018).	52
<i>Figura 7.</i> Cuadrante Mágico del 2017 para Sistemas Operativos de Gestión de Bases de Datos (Microsoft Corporation, 2017).	54
<i>Figura 8.</i> Cuadrante Mágico del 2016 para Sistemas Operativos de Gestión de Bases de Datos (Walker, 2016).....	54
<i>Figura 9.</i> Cuadrante Mágico del 2015 para Sistemas Operativos de Gestión de Bases de Datos (Naeem, 2015).....	55
<i>Figura 10.</i> Ejemplo GUI (Shakya, 2017).....	56
<i>Figura 11.</i> Entorno GUIDE de Matlab (MathWorks, s.f.).....	58
<i>Figura 12.</i> Fuente de datos (University of California, Irvine, 2018).	60
<i>Figura 13.</i> Crear una nueva base de datos (Fuente propia, 2018).....	63
<i>Figura 14.</i> Asignar nombre a la nueva base de datos (Fuente propia, 2018).....	63
<i>Figura 15.</i> Abrir el Asistente para importación y exportación de SQL Server (Fuente propia, 2018).	64
<i>Figura 16.</i> Definición del origen de datos para la importación (Fuente propia, 2018).....	64
<i>Figura 17.</i> Definición del destino de importación (Fuente propia, 2018).....	65
<i>Figura 18.</i> Selección de nombres para las tablas (Fuente propia, 2018).....	66
<i>Figura 19.</i> Tablas OpenSource (Fuente propia, 2018).....	66
<i>Figura 20.</i> Datos tabla diabetic_data (Fuente propia, 2018).....	67
<i>Figura 21.</i> Datos tabla admission_type (Fuente propia, 2018).	67
<i>Figura 22.</i> Datos tabla admission_source (Fuente propia, 2018).....	67

<i>Figura 23.</i> Datos tabla discharge_disposition (Fuente propia, 2018).	68
<i>Figura 24.</i> Creación de una nueva vista de base de datos (Fuente propia, 2018).	68
<i>Figura 25.</i> Datos organizados y almacenados en la vista opensource_view (Fuente propia, 2018).	69
<i>Figura 26.</i> Configuración origen de datos en Database Explorer App (Fuente propia, 2018).	70
<i>Figura 27.</i> Administrador de origen de datos ODBC (Fuente propia, 2018).	70
<i>Figura 28.</i> Selección de controlador para establecer el origen de datos (Fuente propia, 2018).	71
<i>Figura 29.</i> Configuración DSN de Microsoft SQL Service (Fuente propia, 2018).	71
<i>Figura 30.</i> Etapa de definición de parámetros para la búsqueda (Fuente propia, 2018).	72
<i>Figura 31.</i> Etapa de presentación de los resultados de la búsqueda (Fuente propia, 2018).	73
<i>Figura 32.</i> Aviso sin coincidencias (Fuente propia, 2018).	73
<i>Figura 33.</i> Aviso vacío (Fuente propia, 2018).	74
<i>Figura 34.</i> Datos vista ETL (Fuente propia, 2018).	76
<i>Figura 35.</i> Nuevo proyecto de Analysis Services (Fuente propia, 2018).	76
<i>Figura 36.</i> Nuevo origen de datos (Fuente propia, 2018).	77
<i>Figura 37.</i> Elección de base de datos para el origen de datos (Fuente propia, 2018).	77
<i>Figura 38.</i> Opciones de seguridad (Fuente propia, 2018).	78
<i>Figura 39.</i> Ingreso al Asistente de Vista de Origen de Datos (Fuente propia, 2018).	78
<i>Figura 40.</i> Elección de tablas para el modelo (Fuente propia, 2018).	79
<i>Figura 41.</i> Elección de método de minería de datos (Fuente propia, 2018).	79
<i>Figura 42.</i> Elección de variables de entrada, de salida y clave primaria (Fuente propia, 2018).	80
<i>Figura 43.</i> Inicio procesamiento del modelo (Fuente propia, 2018).	80
<i>Figura 44.</i> Datos importados ENTIDAD A (Fuente propia, 2018).	83
<i>Figura 45.</i> Datos importados ENTIDAD B (Fuente propia, 2018).	84
<i>Figura 46.</i> Vista entidadA_view creada (Fuente propia, 2018).	84
<i>Figura 47.</i> Datos entidadA_view (Fuente propia, 2018).	85
<i>Figura 48.</i> Vista entidadB_view creada (Fuente propia, 2018).	85

Figura 49. Datos entidadB_view (Fuente propia, 2018).	85
Figura 50. Ventana elección parámetros para búsqueda en Entidad A (Fuente propia, 2018).	86
Figura 51. Ventana elección parámetros para búsqueda en Entidad B (Fuente propia, 2018).	87
Figura 52. Ventana para la integración de varias fuentes (Fuente propia, 2018).	88
Figura 53. Crear Tabla dinámica de Excel (Fuente propia, 2018).	89
Figura 54. Selección de datos para la Tabla dinámica (Fuente propia, 2018).	89
Figura 55. Selección de campos para el análisis en la Tabla dinámica (Fuente propia, 2018).	90
Figura 56. Campos elegidos para análisis de diagnósticos (Fuente propia, 2018).	90
Figura 57. Crear un Gráfico dinámico (Fuente propia, 2018).	91
Figura 58. Grafica circular con análisis de Cuenta DESCRIPCION DX CIE 10 (Fuente propia, 2018)..	91
Figura 59. Grafica circular análisis de Cuenta DESCRIPCION DX CIE 10 con datos filtrados (Fuente propia, 2018)	92
Figura 60. Datos filtro_diagnosticos (Fuente propia, 2018).	93
Figura 61. Tabla de frecuencias con SQL (Fuente propia, 2018).	94
Figura 62. Gráfico de frecuencias diagnósticos Entidad B (Fuente propia, 2018).	95
Figura 63. Datos vista filtro_diabetes (Fuente propia, 2018).	96
Figura 64. Origen de datos Entidad A (Fuente propia, 2018).	96
Figura 65. Origen de datos Entidad B (Fuente propia, 2018).	96
Figura 66. Variables modelo Árbol de decisión Entidad A (Fuente propia, 2018)	97
Figura 67. Árbol de decisión Entidad A (Fuente propia, 2018).	98
Figura 68. Red de dependencias Árbol de decisión Entidad A (Fuente propia, 2018).	98
Figura 69. Lift chart árbol de decisión Entidad A (Fuente propia, 2018).	99
Figura 70. Variables para los modelos Entidad B (Fuente propia, 2018).	100
Figura 71. Árbol de decisión Entidad B (Fuente propia, 2018).	100
Figura 72. Regresión logística Entidad B (Fuente propia, 2018).	101
Figura 73. Redes neuronales Entidad B (Fuente propia, 2018).	101

Figura 74. Lift chart Árbol de decisión vs Regresión logística vs Redes neuronales (Fuente propia, 2018). 102

Figura 75. Error cuadrático medio Árbol de decisión Entidad B (Fuente propia, 2018)..... 102

Figura 76. Error cuadrático medio Redes neuronales Entidad B (Fuente propia, 2018). 102

Figura 77. Error cuadrático medio Regresión logística Entidad B (Fuente propia, 2018). 103

Figura 78. Variables modelo de clusterización Entidad B (Fuente propia, 2018). 103

Figura 79. Clusters pacientes Entidad (Fuente propia, 2018)..... 104

Figura 80. Lift chart cluster Entidad B (Fuente propia, 2018)..... 104

Figura 81. Perfiles de cada cluster (Fuente propia, 2018). 105

LISTA DE ANEXOS

ANEXO 1. Código query para crear la vista opensource_view	115
ANEXO 2. Código búsqueda en la vista opensource_view por medio de la GUI	115
ANEXO 3. Código tabla para presentación de resultados de la búsqueda	116
ANEXO 4. Código aviso sin coincidencias	117
ANEXO 5. Código aviso campo vacío.....	117
ANEXO 6. Código SQL query para crear la vista ETL	117
ANEXO 7. Fórmula para la clasificación de la edades	117
ANEXO 8. Código SQL query para crear la vista entidadA_view	118
ANEXO 9. Código SQL query para crear la vista entidadB_view	118
ANEXO 10. Nueva programación para la muestra de resultados de búsqueda	118
ANEXO 11. Programación ventana para definición de parámetros de búsqueda Entidad A.....	120
ANEXO 12. Programación ventana para definición de parámetros de búsqueda Entidad B.....	121
ANEXO 13. Programación para integración de distintas fuentes	121
ANEXO 14. SQL query para la vista filtro_diagnosticos	122
ANEXO 15. Query para determinación de frecuencias	122
ANEXO 16. SQL query para vista filtro_diabetes.....	123
ANEXO 17. Lista de características y sus descripciones en el conjunto de datos inicial. (Strack, y otros, 2014).....	123
ANEXO 18. Descripción id para Admission type. (University of California, Irvine, 2018).....	125
ANEXO 19. Descripción id para Discharge disposition. (University of California, Irvine, 2018).....	125
ANEXO 20. Descripción id para Admission source. (University of California, Irvine, 2018).....	126
ANEXO 21. Codificación ICD9 para diagnósticos (Strack, y otros, 2014).....	127

RESUMEN

En el presente trabajo se describe el proceso para la realización de una interfaz gráfica que cumpla la función de un repositorio y por medio de búsquedas muestre la información requerida proveniente de varias fuentes de datos, con el fin de lograr una integración de la información y evitar que esta se mantenga dispersa, logrando con esto facilitar el proceso de análisis de datos.

Para lograr los objetivos se recurrió primero a una investigación en la que se recopiló información sobre los métodos para el análisis de datos, luego de tener los conceptos claros se buscaron fuentes de datos (como bases de datos open source) que hicieran parte de un entrenamiento de la metodología propuesta en este trabajo. Al tener los datos, ellos son importados desde el entorno de MS SQL Server para su almacenamiento y organización. Luego, por medio de la aplicación Database Explorer de Matlab, conectar el entorno de Matlab con el servidor de SQL Server y establecer una comunicación con los datos para poder interactuar con ellos desde Matlab. El siguiente paso fue diseñar la interfaz gráfica de usuario en el entorno GUIDE de Matlab para crear un buscador para la base de datos, utilizando la conexión antes establecida. En paralelo a este proceso se usó la herramienta SQL Server Data Tools de Visual Studio en la que se hicieron los análisis de datos por medio de los métodos investigados en primera instancia, con los que se identificaron patrones que podrán ayudar en la toma de decisiones asistenciales.

Con esta metodología ya realizada, se acudió a entidades de salud locales, los cuales brindarán información de sus bases de datos para aplicar dicha metodología. Como resultado final del trabajo se construyó una interfaz gráfica capaz de hacer búsquedas en tres bases de datos distintas y un análisis de minería de datos con el que se proponen alternativas para mejorar el servicio prestado.

Palabras clave: Minería de datos, ETL, Interfaz Gráfica de Usuario (GUI).

ABSTRACT

In the present work we describe the process for the realization of a graphical interface that fulfills the function of a repository and through of queries it shows the required information coming from several data sources, in order to achieve an integration of the information and avoid that it remains dispersed, thus facilitating the process of data analysis.

To achieve the objectives, we first resorted to an investigation in which information was collected on the methods for data analysis, after having clear concepts, we searched for data sources (such as open source databases) that were part of a training of the methodology proposed in this work. By having the data, they are imported from the MS SQL Server environment for storage and organization. Then, through the Matlab Database Explorer application, connect the Matlab environment to MS SQL Server and establish a communication with the data to interact with them from Matlab. The next step is to design the graphical user interface in the GUIDE environment of Matlab to create a search engine for the database, using the connection established above. In parallel to this process, the SQL Server Data Tools of Visual Studio will be used in which the data analysis will be done through of the methods investigated in the first instance, with which patterns that will help in the decision-making process will be detected.

With this methodology already done, we will go to local hospitals that will provide information on their databases to apply this methodology. As a final result of the work, a graphical interface capable of searching in three different databases and a data mining analysis was proposed, with which alternatives are proposed to improve the service provided.

Keywords: Data mining, ETL, Graphical User Interface (GUI).

INTRODUCCIÓN

La prestación de servicios de salud es una tarea compleja. A nivel clínico, la atención personalizada se guía, en parte, por la historia clínica, el examen, los signos vitales y la evidencia. El diseño y operación de los sistemas de las organizaciones de salud, especialmente los departamentos de emergencia, es extremadamente complejo, principalmente debido a: la gran cantidad de recursos diferentes, como los equipos médicos y el personal asistencial involucrados en las actividades de atención, la incertidumbre resultante de estas actividades que ocurren en diferentes momentos y la clara probabilidad de necesitar de estos recursos simultáneamente.

Por lo tanto, es necesaria una herramienta que permita la organización de datos que se caracterizan por contener un alto volumen de información y por proporcionar información de una alta complejidad, para su posterior análisis con el que se ayude a la toma de decisiones administrativas dentro de las entidades de salud.

Para el desarrollo de esta herramienta se hace una revisión bibliográfica con la que caractericen los modelos para la analítica de datos. Posteriormente se identifican las fuentes de datos clínicos y técnicos más apropiadas. Luego se diseña una propuesta de repositorio de datos clínicos y técnicos, y se evaluará su desempeño con datos de una entidad de salud local.

La propuesta consiste en una interfaz gráfica de usuario en la que se integra la información contenida de diferentes bases de datos y por medio de esta aplicación se realizan búsquedas que permitan mostrar datos específicos de un gran volumen de datos. Además de un análisis de minería de datos que ayude a la toma de decisiones para contribuir al mejoramiento del servicio prestado por una entidad de salud.

1. PRELIMINARES

1.1. PLANTEAMIENTO DEL PROBLEMA

Una era de información abierta en salud ahora está en camino. Ya se ha experimentado una década de progreso en la digitalización de registros médicos, ya que las compañías farmacéuticas y otras organizaciones agregan años de datos de investigación y desarrollo en bases de datos electrónicas. El gobierno federal de los EE.UU, y otras partes interesadas públicas también han acelerado el avance hacia la transparencia al hacer que el sector sanitario en su conjunto pueda usar, buscar y poner en práctica décadas de datos almacenados (Groves, Kayyali, Knott, & Kuiken, 2013). En la actualidad, se están llevando a cabo implementaciones de tecnologías en salud potencialmente transformadoras a nivel internacional, a menudo con un impacto significativo en el gasto nacional. Inglaterra, por ejemplo, invirtió al menos £ 12.8 mil millones en un Programa Nacional de Tecnología de la Información para el Servicio Nacional de Salud, y el gobierno de Obama en los Estados Unidos se comprometió de manera similar a una inversión de eHealth de US \$ 38 mil millones cuidado de la salud (Catwell & Sheikh, 2009).

Los sistemas actuales de soporte de decisiones clínicas (CDS) ya incluyen capacidades computarizadas de ingreso de pedidos por parte del médico que analizan las entradas y las comparan con las guías médicas para alertar sobre posibles errores como las reacciones adversas a medicamentos. Al implementar estos sistemas, los proveedores pueden reducir potencialmente las reacciones adversas y reducir las tasas de error de tratamiento y las reclamaciones de responsabilidad, especialmente las derivadas de errores clínicos. Si bien todavía se deben superar obstáculos como la "fatiga de alerta" que puede anular los beneficios de las revisiones de utilización de medicamentos, los signos son prometedores. En un estudio realizado en una unidad de cuidados intensivos pediátricos en una gran área metropolitana de los EE. UU, una

herramienta del sistema CDS redujo las reacciones adversas a los medicamentos y los eventos en un 40 por ciento en solo dos meses (The Data Governance Institute, 2012).

La prestación de servicios de salud es una tarea compleja tanto a nivel individual como de población. A nivel clínico, la atención personalizada a las personas se guía, en parte, por la historia clínica, el examen, los signos vitales y la evidencia (Wyber, y otros, 2015). El diseño y operación de los sistemas de las organizaciones de salud, especialmente los departamentos de emergencia (ED), es extremadamente complejo, principalmente debido a: la gran cantidad de recursos diferentes involucrados en las actividades de atención, la incertidumbre resultante de estas actividades que ocurren en diferentes momentos y la clara probabilidad de necesitar recursos simultáneamente (Young, Eatock, Jahangirian, Naseer, & Lilford, 2009). Como resultado, los largos tiempos de espera del paciente y la sobrepoblación son problemas comunes en los EDs en todo el mundo (El-Zoghby, Farouk, & El-Kilany, 2016). Los ED también son uno de los departamentos hospitalarios más críticos para salvar vidas, además los servicios de ED tienen un mayor impacto político en la opinión del público en general con respecto a cómo se ejecutan los servicios de salud, en comparación con otros servicios. Después de todo, casi todos los habitantes, independientemente de su edad o estado de salud, lo usan en algún momento u otro. Por lo tanto, la falta de excelencia operativa suele definir la opinión del público en general sobre si los servicios de salud se administran de manera eficiente o no. Otro aspecto que cabe resaltar es el flujo de pacientes que emana de los ED, pues este determina las condiciones de operación de muchas unidades y salas en un hospital y, en consecuencia, también sus recursos y niveles de servicio. Abordar las mejoras de los ED y garantizar un proceso eficiente de toma de decisiones son, en consecuencia, asuntos muy importantes para los hospitales y el público en general (Uriarte, Zúñiga, Moris, & Ng, 2017). Estas razones motivan el uso de metodologías de para apoyar a los tomadores de decisiones en el diseño y la mejora de un ED eficiente.

Los datos recopilados para la investigación en Informática de la Salud se caracterizan por el volumen de información proveniente de grandes cantidades de registros almacenados para pacientes: por ejemplo, en algunos conjuntos de datos cada instancia es bastante grande (por ejemplo, conjuntos de datos que utilizan imágenes MRI o micromatrices genéticas para cada paciente), mientras que otros tienen un gran grupo para recopilar datos (tales como datos de redes sociales recopilados de una población). Estos datos también se caracterizan por el ingreso de nuevos datos a altas velocidades, que se pueden ver cuando se trata de monitorear eventos en tiempo real ya sea monitoreando la condición actual de un paciente a través de sensores médicos o intentando rastrear una epidemia a través de multitudes de publicaciones entrantes (por ejemplo, desde Twitter). Otra cualidad de los datos en salud es la variedad los conjuntos de datos con una gran cantidad de tipos variables de atributos independientes, conjuntos de datos recopilados de muchas fuentes (por ejemplo, los datos de consulta de búsqueda provienen de diferentes grupos de edad que usan un motor de búsqueda) o cualquier conjunto de datos que es complejo y necesita para ser visto en muchos niveles de datos en Informática de Salud (Herland, Khoshgoftaar, & Wald, 2014).

Basado en lo anterior, se deduce que es necesaria una herramienta que permita la organización de datos clínicos que se caracterizan por contener un alto volumen de información y que dicha información sea de una alta complejidad, para su posterior análisis con el que se ayude a la toma de decisiones administrativas dentro de las entidades de salud.

1.2. OBJETIVOS DEL PROYECTO

1.2.1. Objetivo General

Desarrollar un repositorio de datos clínicos y técnicos usando los datos de historia clínica y del área de laboratorio clínico, a partir de un esquema de selección e integración de diferentes fuentes para la identificación de patrones que sirvan en la toma de decisiones en una institución de salud

1.2.2. Objetivos Específicos

- Caracterizar los modelos analíticos para repositorios de datos clínicos y técnicos que existen.
- Identificar las fuentes de datos clínicos y técnicos más apropiadas.
- Diseñar un repositorio de datos clínicos y técnicos.
- Evaluar la propuesta de repositorio de datos clínicos y técnicos

1.3. MARCO DE REFERENCIA

1.3.1. Antecedentes

Se han descubierto las bondades de la integración de tecnologías de la información en el ámbito médico, más que todo en las entidades de alto nivel donde cuentan con recursos destinados para investigaciones en distintos campos que favorezcan al avance de la salud. Los estudios de mayor trascendencia se han llevado a cabo en instituciones ubicadas en Norteamérica, donde se dieron cuenta que tanto los pacientes como los procesos arrojan datos que pueden ser aprovechados para mejorar la calidad del servicio prestado.

Un ejemplo claro es la red INDEPTH1, la cual lanzó el primer repositorio de datos en línea que se especializa en exposición individual longitudinal y datos de mortalidad específica de causa de sistemas de vigilancia demográfica y de salud ubicados en países de ingresos bajos y

medianos, incluyendo África, Asia y Oceanía. Estas son regiones donde tales datos de alta calidad, particularmente longitudinales, son tradicionalmente muy difíciles de obtener. Cada conjunto de datos en el repositorio está documentado de acuerdo con estándares de metadatos aceptados internacionalmente por la Iniciativa de Documentación de Datos, que permite a los usuarios identificar y descargar rápidamente los datos que necesitan. Los identificadores de objetos digitales (DOI) se utilizan para que los conjuntos de datos sean citables y las versiones posteriores, cuando sea necesario, sean inequívocamente identificables (Sankoh, y otros, 2013).

Otro ejemplo es el Project Artemis desarrollado por el Instituto de Tecnología de la Universidad de Ontario en asociación con IBM, una plataforma altamente flexible que aprovecha el análisis de transmisión para monitorear recién nacidos en la unidad de cuidados intensivos neonatales de un hospital. Usando estas tecnologías, el hospital pudo predecir la aparición de infecciones nosocomiales 24 horas antes de que aparecieran los síntomas. El hospital también etiquetó todos los datos de series temporales que habían sido modificados por algoritmos de software. En caso de una demanda o investigación médica, el hospital consideró que tenía que producir tanto las lecturas originales como las modificadas. Además, el hospital estableció políticas para salvaguardar la información de salud protegida (Soares, 2012).

El autor Atul Gawande en un artículo para la revista *New Yorker*, describe cómo los cirujanos ortopédicos del Hospital Brigham and Women de Boston, basándose en su propia experiencia combinada con los datos extraídos de la investigación sobre una serie de factores críticos para el éxito de la cirugía de reemplazo de articulaciones, sistemáticamente estandarizaron la cirugía de reemplazo de articulación de rodilla, con un aumento resultante en resultados más exitosos y costos reducidos. Del mismo modo, el Sistema de Salud de la Universidad de Michigan estandarizó la administración de transfusiones de sangre, reduciendo la necesidad de transfusiones en un 31% y los gastos en \$ 200,000 al mes (Gawande, 2012).

El Departamento de Asuntos de Veteranos (VA) en los Estados Unidos ha demostrado con éxito varios programas de tecnología de la información de atención médica (HIT) y programas remotos de monitoreo de pacientes. El sistema de salud de la Administración de Veteranos generalmente supera al sector privado en seguir los procesos recomendados para el cuidado del paciente, cumplir con las guías clínicas y lograr mayores tasas de terapia farmacológica basada en la evidencia. Estos logros son en gran medida posibles gracias al marco de rendición de cuentas basado en el desempeño del VA y las prácticas de gestión de enfermedades que permiten los registros médicos electrónicos (EMR) y HIT (TechAmerica Foundation, 2012).

Kaiser Permanente, el consorcio integrado de atención administrada con sede en California, conectó datos clínicos y de costos para proporcionar un conjunto de datos crucial que condujo al descubrimiento de los efectos adversos del fármaco y la posterior retirada del medicamento Vioxx del mercado (McKinsey Global Institute, 2011).

Investigadores de la Facultad de Medicina Johns Hopkins descubrieron que podrían usar datos de Google Flu Trends (un agregador gratuito y públicamente disponible de términos de búsqueda relevantes) para predecir oleadas en visitas a la sala de emergencia relacionadas con la gripe una semana antes de que llegaran las advertencias de los Centros para Enfermedades Control y Prevención. Del mismo modo, las actualizaciones de Twitter fueron tan precisas como los informes oficiales para rastrear la propagación del cólera en Haití después del terremoto de enero de 2010; también fueron dos semanas antes (McAfee, 2012).

Los investigadores de IBM han ideado un programa prototipo que predice los resultados probables de los pacientes con diabetes, basándose en los datos de salud longitudinales de los pacientes y su asociación con médicos particulares, protocolos de gestión y relación con los promedios de gestión de la salud de la población (Hani Neuvirth, 2012).

Estos son algunos ejemplos del desarrollo que se ha venido dando en los años recientes, pero es una tendencia que según muchos expertos es el futuro de la salud, tanto para la gestión administrativa de las instituciones como para el tratamiento de enfermedades y su detección temprana con diagnósticos cada vez más acertados.

1.3.2. Salud digital o eHealth

El Plan de Acción sobre salud digital (o eHealth) de la Comisión Europea define salud digital como "la aplicación de tecnologías de la información y las comunicaciones en toda la gama de funciones que afectan al sector de la salud incluyendo productos, sistemas y servicios que van más allá de las aplicaciones simplemente basadas en Internet" (Gobierno de España: Ministerio de Energía, Turismo y Agenda Virtual, 2017).

Así, la salud digital puede ser entendida como la aplicación de Internet y otras tecnologías relacionadas en la industria de la salud para mejorar el acceso, la eficiencia, la eficacia y calidad de los procesos clínicos y empresariales utilizadas por las organizaciones de salud, médicos, pacientes y consumidores en un esfuerzo por mejorar el estado de salud de los pacientes (Fundación Vodafone España, Red.es & Gobierno de España: Ministerio de Energía, Turismo y Agenda Virtual, 2017).

En el pasado, el término "tecnología médica" a menudo se usaba para describir el conjunto de técnicas, medicamentos, equipos y procedimientos utilizados por los profesionales de la salud para brindar atención médica a las personas. Por lo tanto, las implementaciones de TIC se habrían considerado históricamente bajo este título (Office of Technology Assessment, 1976).

Ejemplo de esto son las intervenciones de telemedicina, es decir, la prestación de servicios de atención médica a larga distancia por medios tales como dispositivos de telemonitorización (por ejemplo, teleradiología y telecardiología), teleconsulta o incluso telecirugía (McLean & Sheikh,

2009). Numerosos sistemas de registros médicos electrónicos se utilizan, por ejemplo, para registrar los detalles de la salud de los pacientes y sus medicamentos como lo mencionan Garets y Davis en su publicación *Electronic Medical Records vs. Electronic Health Records: Yes, There Is a Difference*. Finalmente, hay muchos portales de salud que proporcionan un medio para acceder a registros médicos e información relacionada con la salud a través de una red segura, como Google Health, Microsoft HealthVault y HealthSpace del Servicio Nacional de Salud.

Todos estos sistemas se han definido como tecnologías eHealth, pero todos pueden tener fines muy diferentes y todos pueden estar dirigidos a grupos de usuarios muy diferentes. Por esta razón, se sugiere que cualquier definición de eHealth debe abarcar todo el espectro de las TIC, al mismo tiempo que se aprecia el contexto de uso y el valor que pueden aportar a la sociedad (Catwell & Sheikh, 2009).

Los repositorios de datos clínicos con miras a su posterior análisis se podrían categorizar dentro del gran conjunto de herramientas TIC que se enmarcan dentro del campo de acción del eHealth como alternativa para la mejora del desempeño de las instituciones de salud.

1.3.3. Los historiales o registros médicos electrónicos integrados

La Fundación Vodafone España junto con Red.es y el Ministerio de Energía, Turismo y Agenda Virtual del Gobierno de España, realizó una exhaustiva investigación en marzo de 2017 sobre el Big Data (del cual se hablará luego en este trabajo) y su impacto en el sector de la salud donde guarda un apartado en el que se explica las funcionalidades de la digitalización de los registros médicos.

Primeramente, se define a los historiales médicos electrónicos integrados en redes o registros electrónicos de salud como expedientes que pueden integrar la información clínica y administrativa de un paciente en redes de información sanitaria que cumplen los estándares de

interoperabilidad para poder ser utilizados y compartidos por profesionales autorizados dentro de más de una organización de salud.

Los sistemas de información hospitalaria suelen contemplar la información demográfica y general del paciente, la agenda médica y la ficha clínica del paciente. También, almacenan y organizan toda la información específica de los diagnósticos y tratamientos efectuados. La implementación de estos sistemas en una institución de salud permite el acceso expedito a la información de tratamiento y facilita al personal médico obtener un amplio conocimiento del estado del paciente. Estos sistemas son gestionados por los profesionales de la salud de un centro de asistencia sanitaria.

Además, los sistemas de información hospitalaria también permiten el control de los servicios prestados a los pacientes y sus costes asociados, así como el resto de información administrativa.

Algunas veces, estos registros electrónicos de salud pueden ser gestionados y compartidos por el propio paciente. En este caso se conocen como carpetas personales de salud o expedientes electrónicos del paciente.

A continuación, se presentan las funcionalidades más habituales de los registros médicos electrónicos. Tanto los registros electrónicos de salud (EHR, por sus siglas en inglés) gestionados por los profesionales médicos, como los registros personales de salud (PHR, por sus siglas en inglés), gestionados por los propios pacientes.

Registros electrónicos

Los registros electrónicos de salud están diseñados para registrar y compartir la información por los profesionales del sistema de salud. Un registro electrónico de la información completa sobre la salud del paciente permite a los profesionales dar la mejor atención posible, ya sea durante una consulta o en una emergencia médica, proporcionando la información que pueden

necesitar para evaluar la condición de salud del paciente (Gobierno de España: Ministerio de Energía, Turismo y Agenda Virtual, 2017).

De la misma manera, los registros electrónicos de salud agilizan la atención médica. Por ejemplo, en una situación de emergencia pueden proveer acceso instantáneo a la información sobre la historia de un paciente y tener información inmediata sobre alergias o medicamentos sin tener que esperar ninguna prueba necesaria para esclarecerlo (Gobierno de España: Ministerio de Energía, Turismo y Agenda Virtual, 2017).

Las funcionalidades de estos registros pueden clasificarse en distintas subdimensiones. Algunas de ellas son:

- Los registros de información e historia clínica incluyen, entre otros, el registro de la historia médica, los síntomas, los resultados de los tratamientos terapéuticos, las constantes vitales, las imágenes radiológicas, los parámetros médicos básicos, los test y pruebas o las razones de la cita médica.
- Los sistemas de ayuda al diagnóstico incluyen funcionalidades como los sistemas de ayuda al diagnóstico sobre contraindicaciones; el registro de las interacciones en medicamentos o guías clínicas y mejores prácticas.
- La gestión administrativa del paciente incluye aspectos como el registro de datos administrativos o de facturación.
- Los aspectos de apoyo farmacológico incluyen tanto los listados de fármacos como el registro de prescripciones (Gobierno de España: Ministerio de Energía, Turismo y Agenda Virtual, 2017).

Funcionalidades de los registros personales de salud (PHR)

Los registros personales de salud pueden ser vistos como la parte cliente de los registros electrónicos de salud. Estos pueden incluir tanto el resumen de la información e historia clínica del paciente como la administrativa, de facturación o información relacionada con los seguros médicos.

En este sentido, muchas de las posibilidades de registro de datos biológicos que permiten los llevables (wearables) se pueden incluir, en función del diseño del sistema, entre las funcionalidades de los registros personales de salud. Las principales funcionalidades del PHR se pueden clasificar en las siguientes subdimensiones:

- La información clínica, como la posibilidad del paciente de acceder a sus registros médicos; ver los resultados de sus pruebas; pedir renovaciones de prescripciones farmacéuticas o complementar sus propios registros médicos con otra información.
- La información administrativa incluye aspectos como la solicitud de citas o bien la solicitud de referencias médicas (Gobierno de España: Ministerio de Energía, Turismo y Agenda Virtual, 2017).

1.3.4. Minería de datos o Data mining

Pero de nada sirve guardar todos estos datos si no son analizados. La información que proporciona el análisis es lo que realmente le da el valor al hecho de almacenar todos esos datos porque finalmente esos datos son convertidos en conocimiento, ofrecen una perspectiva distinta del contexto abordado que facilita la toma de decisiones con miras a un mejoramiento de los procedimientos utilizados en una institución de cualquier tipo.

El método tradicional para convertir los datos en conocimiento se basa en el análisis y la interpretación manual. Por ejemplo, en la industria del cuidado de la salud, es común que los

especialistas analicen periódicamente las tendencias actuales y los cambios en los datos de cuidado de la salud, por ejemplo, trimestralmente. Luego, los especialistas proporcionan un informe que detalla el análisis a la organización patrocinadora de la atención de la salud; este informe se convierte en la base para la futura toma de decisiones y la planificación de la gestión de la atención de la salud (Fayyad, Piatetsky-Shapiro, & Smyth, 1996). Los registros médicos electrónicos integrados son la fuente principal para estos análisis, pues en ellos se almacena un gran volumen de información clínica y administrativa de pacientes en redes de información al interior de las entidades de salud.

Para muchas aplicaciones, esta forma de sondeo manual de un conjunto de datos es lenta, costosa y altamente subjetiva. De hecho, a medida que los volúmenes de datos crecen drásticamente, este tipo de análisis de datos manual se vuelve completamente impracticable en muchos dominios. Las bases de datos están aumentando de tamaño de dos maneras: (1) el número N de registros u objetos en la base de datos y (2) el número de campos o atributos a un objeto (Fayyad, Piatetsky-Shapiro, & Smyth, 1996).

Las empresas usan datos para obtener ventajas competitivas, aumentar la eficiencia y brindar servicios más valiosos a los clientes. Los datos que se capturan del entorno son la evidencia básica usada para construir teorías y modelos del universo en el que vivimos. Debido a que las computadoras han permitido a los humanos recopilar más datos de los que se pueden digerir, es natural recurrir a cálculos computacionales. Técnicas que permiten descubrir patrones y estructuras significativos a partir de volúmenes masivos de datos (Fayyad, Piatetsky-Shapiro, & Smyth, 1996). La minería de datos surge como una solución a esta necesidad que se presenta en la actual era de la información.

El data mining (minería de datos), es el conjunto de técnicas y tecnologías que permiten explorar grandes bases de datos, de manera automática o semiautomática, con el objetivo de

encontrar patrones repetitivos, tendencias o reglas que expliquen el comportamiento de los datos en un determinado contexto (Sinnexus Business Intelligence, s.f.).

Básicamente, el data mining surge para intentar ayudar a comprender el contenido de un repositorio de datos. Con este fin, hace uso de prácticas estadísticas y, en algunos casos, de algoritmos de búsqueda próximos a la Inteligencia Artificial y a las redes neuronales

El proceso común del data mining se suele componer de cuatro etapas principales:

- **Determinación de los objetivos.** Trata de la delimitación de los objetivos que el cliente desea bajo la orientación del especialista en data mining.
- **Preprocesamiento de los datos.** Se refiere a la selección, la limpieza, el enriquecimiento, la reducción y la transformación de las bases de datos. Esta etapa consume generalmente alrededor del setenta por ciento del tiempo total de un proyecto de data mining.
- **Determinación del modelo.** Se comienza realizando unos análisis estadísticos de los datos, y después se lleva a cabo una visualización gráfica de los mismos para tener una primera aproximación. Según los objetivos planteados y la tarea que debe llevarse a cabo, pueden utilizarse algoritmos desarrollados en diferentes áreas de la Inteligencia Artificial.
- **Análisis de los resultados.** Verifica si los resultados obtenidos son coherentes y los coteja con los obtenidos por los análisis estadísticos y de visualización gráfica. El cliente determina si son novedosos y si le aportan un nuevo conocimiento que le permita considerar sus decisiones. (Sinnexus Business Intelligence, s.f.).

Aunque también la aplicación ciega de métodos de minería de datos (correctamente criticada como dragado de datos en la literatura estadística) puede ser una actividad peligrosa, que fácilmente conduce al descubrimiento de patrones sin sentido e inválidos (Fayyad, Piatetsky-Shapiro, & Smyth, 1996).

1.3.5. Métodos de Minería de Datos

1.3.5.1. Árboles de decisión

Los árboles de decisión son actualmente uno de los métodos más populares utilizados para el modelado de datos. Tienen la ventaja de ser conceptualmente simples y se ha demostrado que funcionan bien en una variedad de problemas. Los árboles de decisión tienen muchos usos, como, por ejemplo, predecir un resultado probable, ayudar en el análisis de problemas y ayudar a tomar decisiones. Al formular y configurar árboles de decisión, se analizan y compilan los resultados de factores del mundo real, de modo que las características específicas de los factores anteriores y los resultados relacionados se utilizan para predecir los resultados de los factores futuros (EE.UU. Patente n° 10/406,836, 2004).

Desafortunadamente, para todos los árboles de decisión menos los más simples, la cantidad potencial de configuraciones de árbol puede ser enorme. Por ejemplo, se puede generar un árbol de decisiones para determinar si una persona tiene una expectativa de vida baja, media o alta. Los factores analizados pueden incluir, por ejemplo, si la persona es fumadora; la altura de la persona; el peso de la persona, el género de la persona y la ocupación de la persona. Como las ramas del árbol de decisión (cada una de las cuales representa un factor) pueden configurarse en muchas secuencias diferentes, el número de árboles de decisión potenciales aumenta rápidamente a medida que aumenta el número de factores. Además, hay muchas maneras diferentes de aprender árboles de decisión, por ejemplo, utilizando solo divisiones binarias, en lugar de aceptar cualquier cantidad de divisiones. (EE.UU. Patente n° 10/406,836, 2004).

Los árboles de decisión clasifican las instancias en el árbol desde la raíz hasta un nodo hoja, que proporciona la clasificación de la instancia. Cada nodo en el árbol especifica una prueba de algún atributo de la instancia, y cada rama que desciende de ese nodo corresponde a uno de los valores posibles para este atributo. Una instancia se clasifica comenzando en el nodo raíz del

árbol, probando el atributo especificado por este nodo, y luego bajando por la rama del árbol correspondiente al valor del atributo en el ejemplo dado. Este proceso se repite para el subárbol enraizado en el nuevo nodo (Mitchell, 1997).

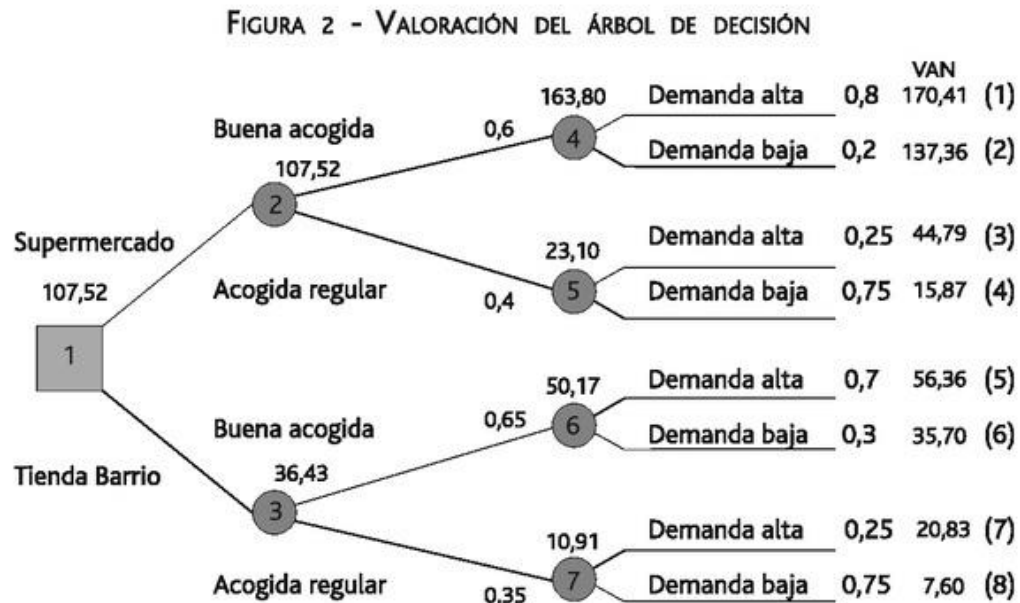


Figura 1. Representación árbol de decisión (Mitchell, 1997).

En general, los árboles de decisión representan una disyunción de conjunciones de restricciones en los valores de atributo de instancias. Cada camino desde la raíz del árbol hasta una hoja corresponde a una conjunción de pruebas de atributos, y el árbol mismo a una disyunción de estas conjunciones. Por ejemplo, el árbol de decisión que se muestra en la corresponde a la expresión:

El aprendizaje del árbol de decisiones generalmente se adapta mejor a los problemas con las siguientes características:

- Las instancias están representadas por pares de valor-atributo.
- La función objetivo tiene valores de salida discretos.

- Descripciones disyuntivas pueden ser requeridas
- Los datos de entrenamiento pueden contener errores.
- Los datos de entrenamiento pueden contener valores de atributo faltantes.

Se han encontrado muchos problemas prácticos que se ajustan a estas características. Por lo tanto, el aprendizaje del árbol de decisiones se ha aplicado a problemas tales como aprender a clasificar a los pacientes médicos por su enfermedad, mal funcionamiento de los equipos por su causa y solicitantes de préstamos por su probabilidad de impago en los pagos. Tales problemas, en los que la tarea es clasificar los ejemplos en uno de un conjunto discreto de categorías posibles, a menudo se denominan problemas de clasificación (Mitchell, 1997).

En ciertos casos, los datos disponibles pueden ser valores perdidos para algunos atributos. Por ejemplo, en un dominio médico en el que deseamos predecir el resultado del paciente según diversas pruebas de laboratorio, es posible que el resultado del análisis de sangre en la prueba de laboratorio esté disponible solo para un subconjunto de pacientes. En tales casos, es común estimar el valor del atributo faltante en base a otros ejemplos para los cuales este atributo tiene un valor conocido (Mitchell, 1997).

Los aspectos prácticos en el aprendizaje de árboles de decisión incluyen determinar qué tan profundamente crecer el árbol de decisión, manejar atributos continuos, transferir datos de entrenamiento con valores de atributo faltantes, manejar atributos con costos diferentes y mejorar la eficiencia computacional (Mitchell, 1997).

1.3.5.2. Clustering

El clustering es una técnica de Machine Learning que implica la agrupación de puntos de datos. Dado un conjunto de puntos de datos, podemos usar un algoritmo de agrupamiento para clasificar cada punto de datos en un grupo específico (Seif, 2018). La agrupación de un conjunto

particular de objetos se da en función de sus características, agregándolos de acuerdo con sus similitudes. En cuanto a la Minería de Datos, esta metodología divide los datos implementando un algoritmo de combinación específico, el más adecuado para el análisis de información deseado (Big Data Made Simple, 2015).

Dado que esta es una técnica de análisis de datos muy valiosa, tiene varias aplicaciones diferentes en el mundo de las ciencias. Cada gran conjunto de datos de información puede procesarse mediante este tipo de análisis, produciendo excelentes resultados con muchos tipos distintos de datos (Big Data Made Simple, 2015).

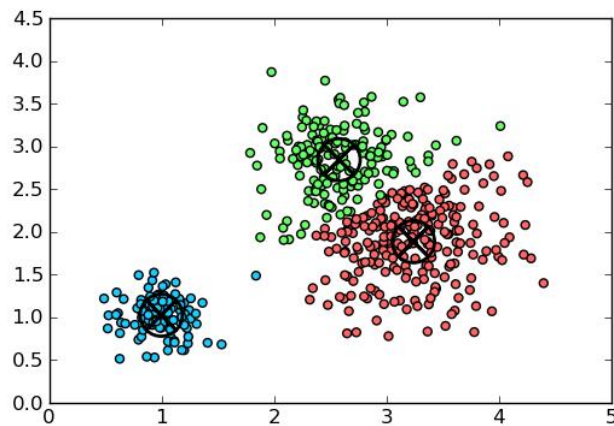


Figura 2. Representación clusterización (Pacula, 2011).

Dentro de las aplicaciones del análisis de clustering se encuentran las siguientes:

- El análisis de clustering se usa ampliamente en muchas aplicaciones, como la investigación de mercado, el reconocimiento de patrones, el análisis de datos y el procesamiento de imágenes.
- La agrupación en clústeres también puede ayudar a los especialistas en marketing a descubrir grupos distintos en su base de clientes. Y pueden caracterizar sus grupos de clientes en función de los patrones de compra.

- En el campo de la biología, se puede usar para derivar taxonomías de plantas y animales, categorizar genes con funcionalidades similares y obtener información sobre las estructuras inherentes a las poblaciones.
- La agrupación también ayuda a identificar áreas de uso de la tierra similar en una base de datos de observación de la tierra. También ayuda en la identificación de grupos de casas en una ciudad de acuerdo con el tipo de casa, el valor y la ubicación geográfica.
- La agrupación en clústeres también ayuda a clasificar documentos en la web para el descubrimiento de información.
- La agrupación en clústeres también se usa en aplicaciones de detección de valores atípicos, como la detección de fraudes con tarjetas de crédito.
- Como una función de minería de datos, el análisis de conglomerados sirve como una herramienta para obtener información sobre la distribución de datos para observar las características de cada grupo (Tutorials Point, s.f.).

1.3.5.3. Razonamiento bayesiano

El razonamiento bayesiano proporciona un enfoque probabilístico para la inferencia. Se basa en el supuesto de que las cantidades de interés se rigen por distribuciones de probabilidad y que las decisiones óptimas se pueden tomar razonando sobre estas probabilidades junto con los datos observados. Es importante el aprendizaje automático porque proporciona un enfoque cuantitativo para ponderar la evidencia que respalda hipótesis alternativas. El razonamiento bayesiano proporciona la base para algoritmos de aprendizaje que manipulan directamente las probabilidades, así como un marco para analizar el funcionamiento de otros algoritmos que no manipulan explícitamente las probabilidades (Mitchell, 1997).

Los métodos de aprendizaje Bayesianos son relevantes para nuestro estudio del aprendizaje automático por dos razones diferentes. En primer lugar, los algoritmos de aprendizaje Bayesiano que calculan probabilidades explícitas para hipótesis, como el clasificador ingenuo de Bayes, se encuentran entre los enfoques más prácticos para ciertos tipos de problemas de aprendizaje. Por ejemplo, Michie et al. (1994) proporcionan un estudio detallado que compara el clasificador ingenuo de Bayes con otros algoritmos de aprendizaje, incluidos los algoritmos de árbol de decisión y de red neuronal. Estos investigadores muestran que el clasificador ingenuo de Bayes es competitivo con estos otros algoritmos de aprendizaje en muchos casos y que en algunos casos supera a estos otros métodos (Mitchell, 1997).

La segunda razón por la que los métodos bayesianos son importantes para este estudio del Machine Learning es que proporcionan una perspectiva útil para comprender muchos algoritmos de aprendizaje que no manipulan explícitamente las probabilidades (Mitchell, 1997).

Las características de los métodos de aprendizaje bayesianos incluyen:

- Cada ejemplo de entrenamiento observado puede disminuir o aumentar gradualmente la probabilidad estimada de que una hipótesis sea correcta. Esto proporciona un enfoque de aprendizaje más flexible que los algoritmos que eliminan por completo una hipótesis si se considera que es inconsistente con un solo ejemplo.
- El conocimiento previo puede combinarse con datos observados para determinar la probabilidad final de una hipótesis. En el aprendizaje Bayesiano, el conocimiento previo se proporciona afirmando (1) una probabilidad previa para cada hipótesis candidata, y (2) una distribución de probabilidad sobre los datos observados para cada hipótesis posible. Los métodos bayesianos pueden acomodar hipótesis que hacen predicciones probabilísticas (por ejemplo, hipótesis tales como "este paciente con neumonía tiene un 93% de posibilidades de recuperación completa").

- Las nuevas instancias se pueden clasificar combinando las predicciones de múltiples hipótesis, ponderadas por sus probabilidades.
- Incluso en los casos en que los métodos Bayesianos demuestran ser computacionalmente intratables, pueden proporcionar un estándar de toma de decisiones óptimo contra el cual se pueden medir otros métodos prácticos.

Una dificultad práctica en la aplicación de métodos Bayesianos es que típicamente requieren un conocimiento inicial de muchas probabilidades. Cuando estas probabilidades no se conocen de antemano, a menudo se estiman basándose en el conocimiento previo, los datos disponibles previamente y las suposiciones sobre la forma de las distribuciones subyacentes. Una segunda dificultad práctica es el costo computacional significativo requerido para determinar la hipótesis óptima de Bayes en el caso general (lineal en el número de hipótesis del candidato). En ciertas situaciones especializadas, este costo computacional se puede reducir significativamente (Mitchell, 1997).

1.3.5.4. Redes neuronales

Los métodos de aprendizaje de la red neuronal proporcionan un enfoque robusto para aproximar las funciones objetivo de valor real, de valor discreto y de valor vectorial. Para ciertos tipos de problemas, como aprender a interpretar datos complejos de sensores del mundo real, las redes neuronales artificiales se encuentran entre los métodos de aprendizaje más efectivos actualmente conocidos. Por ejemplo, el algoritmo BACKPROPAGATION ha demostrado ser sorprendentemente exitoso en muchos problemas prácticos, como aprender a reconocer caracteres escritos a mano (LeCun et al., 1989), aprender a reconocer palabras habladas (Lang et al., 1990) y aprender a reconocer rostros (Cottrell 1990). Una encuesta de aplicaciones prácticas es proporcionada por Rumelhart et al. (1994).

El estudio de las redes neuronales artificiales (ANN, por sus siglas en inglés) se inspiró en parte en la observación de que los sistemas de aprendizaje biológico están formados por redes muy complejas de neuronas interconectadas. En una analogía aproximada, las redes neuronales artificiales se construyen a partir de un conjunto de unidades simples densamente interconectadas, donde cada unidad toma una cantidad de entradas de valores reales (posiblemente las salidas de otras unidades) y produce una sola salida de valor real (que puede convertirse en la entrada a muchas otras unidades) (Mitchell, 1997).

Para desarrollar una idea de esta analogía, consideremos algunos hechos de la neurobiología. El cerebro humano, por ejemplo, se estima que contiene una red densamente interconectada de aproximadamente 10^{11} neuronas, cada una conectada, en promedio, a 10^4 otras. La actividad neuronal normalmente se excita o inhibe a través de conexiones con otras neuronas. Se sabe que los tiempos de conmutación de neuronas más rápidos son del orden de 10^{-3} segundos, bastante lentos en comparación con las velocidades de cambio de computadora de 10^{-10} segundos. Sin embargo, los humanos pueden tomar decisiones sorprendentemente complejas, sorprendentemente rápidas. Por ejemplo, se requieren aproximadamente 10^{-1} segundos para reconocer visualmente a tu madre (Mitchell, 1997).

Mientras que las ANNs están poco motivadas por los sistemas neuronales biológicos, existen muchas complejidades para los sistemas neurales biológicos que no están modelados por las ANNs, y se sabe que muchas características de las ANNs que discutimos aquí son inconsistentes con los sistemas biológicos. Por ejemplo, consideramos aquí las ANNs cuyas unidades individuales arrojan un único valor constante, mientras que las neuronas biológicas generan una compleja serie temporal de picos (Mitchell, 1997).

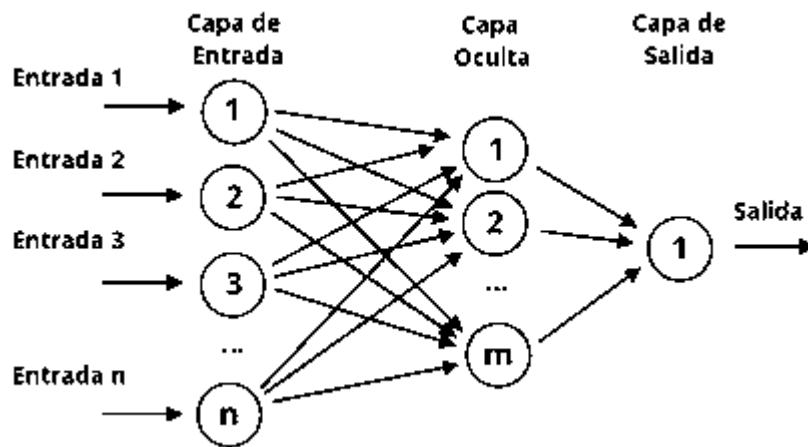


Figura 3. Ilustración modelo redes neuronales (Támez, 2016).

La figura anterior muestra cómo se mapea la imagen de una cámara montada hacia adelante en 960 entradas de red neuronal, que se alimentan a 4 unidades ocultas, conectadas a 30 unidades de salida. Las salidas de red codifican la dirección de manejo ordenada. La figura de la derecha muestra los valores de peso para una de las unidades ocultas en esta red. Los pesos de 30 x 32 en la unidad oculta se muestran en la matriz grande, con bloques blancos que indican positivo y negro que indican pesos negativos. Los pesos de esta unidad oculta a las 30 unidades de salida están representados por el bloque rectangular más pequeño directamente sobre el bloque grande. Como se puede ver a partir de estos pesos de salida, la activación de esta unidad oculta particular alienta un giro hacia la izquierda (Mitchell, 1997).

El aprendizaje ANN es adecuado para problemas en los que los datos de entrenamiento corresponden a datos de sensor complejos y ruidosos, como las entradas de cámaras y micrófonos. También es aplicable a problemas para los cuales se usan a menudo representaciones más simbólicas, como las tareas de aprendizaje del árbol de decisiones. En estos casos, la ANN y el aprendizaje del árbol de decisiones a menudo producen resultados de precisión comparable. El

algoritmo BACKPROPAGATION es la técnica de aprendizaje ANN más comúnmente utilizada.

Es apropiado para problemas con las siguientes características:

- Las instancias están representadas por muchos pares de valores-atributos.
- La salida de la función objetivo puede ser de valor discreto, de valor real o un vector de varios atributos con valores reales o discretos.
- Los ejemplos de entrenamiento pueden contener errores.
- Los tiempos de entrenamiento largos son aceptables.
- Es posible que se requiera una evaluación rápida de la función objetivo aprendida.
- La capacidad de los humanos para comprender la función del objetivo aprendido no es importante (Mitchell, 1997).

1.3.5.5. Reglas de asociación

Las reglas de asociación se derivan de un tipo de análisis que extrae información por coincidencias (Agrawal, 1993); el análisis permite descubrir correlaciones o coocurrencias en los sucesos de la base de datos por analizar y se formaliza en la obtención de reglas de tipo sí ... entonces ..., las cuales se convierten en un importante punto de apoyo para el descubrimiento de conocimiento a partir de la información analizada (Fayyad, Piatetsky-Shapiro, & Smyth, 1996).

Un ejemplo común del uso de las reglas de asociación es el análisis de la canasta familiar adquirida por los compradores, el cual es utilizado para encontrar asociaciones entre los diferentes productos comprados por los clientes; el descubrimiento de tales asociaciones puede ayudar a desarrollar estrategias de mercadeo orientadas al aumento de las ventas (Amaris & Rodríguez, 2003).

La minería de datos describe que el aprendizaje de reglas de asociación es un método popular y bien investigado para descubrir relaciones interesantes entre variables en grandes bases de

datos. Se pretende identificar reglas sólidas descubiertas en las bases de datos utilizando diferentes medidas de interés (A.KrishnaKumar, D.Amrita, & N.Swathi, 2013).

Entre los conjuntos de elementos en las bases de datos de transacciones, apunta a descubrir tendencias implicativas que pueden ser información valiosa para el tomador de decisiones. (A.KrishnaKumar, D.Amrita, & N.Swathi, 2013). En forma general las reglas de asociación poseen la estructura $X \rightarrow$ (entonces) Y , donde X e Y son conjuntos de ítems. X es denominado el antecedente de la regla e Y su consecuente. Además entran en juego los conceptos de soporte y confianza.. El soporte para X entonces Y es el porcentaje de las transacciones que contienen todos los ítems de X e Y , mientras que su confianza es el porcentaje de transacciones que contienen Y , entre las transacciones que contienen X (Monteserin, 2018).

- $\text{Soporte}(X \rightarrow Y) = \text{Prob}(X \cup Y) = \text{Soporte}(X \cup Y)$
- $\text{Confianza}(X \rightarrow Y) = \text{Prob}(Y / X) = \frac{\text{Soporte}(X \cup Y)}{\text{Soporte}(X)}$

Transacciones	
A B C	Soporte (A → C): 0,75 Confianza (A → C): 1
B C	
A C	
A C D	

Figura 4. Soporte y confianza en reglas de asociación (Monteserin, 2018).

Estas métricas nos ayudan a establecer el nivel de validez de la asociación creada y la cantidad de información ofrecida por ella, ya que una regla con bajo soporte indica que puede haber aparecido por casualidad y una regla con baja confianza nos dice que es probable que no exista relación entre antecedente y consecuente (Monteserin, 2018).

La implementación de reglas de asociación es una interesante opción en el proceso de selección de la técnica a utilizar para realizar minería de datos; su aplicación es fundamental para el descubrimiento de relaciones de asociación en grandes cantidades de datos, siendo útil en la

selección de una estrategia de mercadeo, procesos de toma de decisiones, manejo de negocios, etc. (Amaris & Rodríguez, 2003).

1.3.5.6. Regresión lineal

El análisis de regresión lineal ha sido ampliamente utilizado en los más diversos ámbitos para modelar la relación estructural entre una variable respuesta o dependiente y una o un conjunto de predictores o variables independientes o explicativas. En general, el objetivo final es utilizar dicho modelo para predecir, tan precisamente como sea posible, los valores de la variable respuesta frente a observaciones futuras de las variables explicativas. Estos modelos resultan de fácil construcción y, por sobre todo, de fácil interpretación, siendo quizás estas características las que han popularizado el análisis. Sin embargo, la bondad del modelo obtenido depende del cumplimiento de una serie de supuestos o condiciones sobre los que se basa su construcción, limitando su correcto uso a situaciones en las que tales condiciones se satisfagan. Hoy en día, no es difícil encontrar situaciones en las que las variables de interés se relacionan a través de complejas funciones no lineales que raramente puedan ser advertidas o sospechadas a priori para ser tenidas en cuenta en la postulación de los modelos. De allí surge la necesidad de contar con métodos que permitan modelar estas relaciones de manera más flexible en cuanto a los requerimientos que deben cumplirse (Dianda, Quaglino, & Pagura, 2016).

Si se quiere predecir una variable continua (numérica), puede usarse el modelo de regresión lineal, que consiste en crear una ecuación de la siguiente forma:

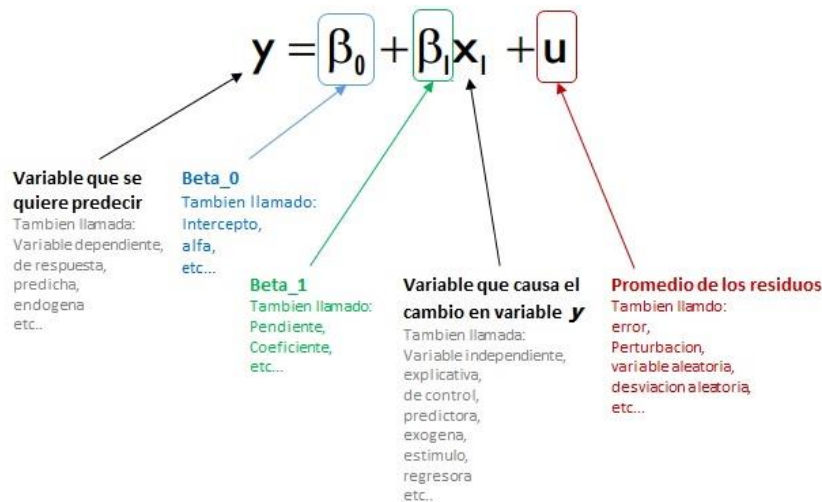


Figura 5. Estructura regresión lineal (Santana, 2015).

1.3.5.7. Regresión logística

Para predecir una variable binaria puede usarse el modelo de regresión logística, que consiste en una transformación de la regresión lineal.

La regresión logística es una de las herramientas estadísticas con mejor capacidad para el análisis de datos en investigación clínica y epidemiología, de ahí su amplia utilización (SEH-LELHA, 2001).

El objetivo primordial que resuelve esta técnica es el de modelar cómo influye en la probabilidad de aparición de un suceso, habitualmente dicotómico, la presencia o no de diversos factores y el valor o nivel de los mismos. También puede ser usada para estimar la probabilidad de aparición de cada una de las posibilidades de un suceso con más de dos categorías (politómico) (SEH-LELHA, 2001).

Este tipo de situaciones se aborda mediante técnicas de regresión. Sin embargo, la metodología de la regresión lineal no es aplicable ya que ahora la variable respuesta sólo presenta dos valores (nos centraremos en el caso dicotómico), por ejemplo, la presencia o ausencia de hipertensión (SEH-LELHA, 2001).

Si se clasifica el valor de la variable respuesta como 0 cuando no se presenta el suceso (ausencia de hipertensión) y con el valor 1 cuando sí está presente (paciente hipertenso), y buscamos cuantificar la posible relación entre la presencia de hipertensión y, por ejemplo, la cantidad media de sal consumida al día como posible factor de riesgo, podríamos caer en la tentación de utilizar una regresión lineal:

$$\text{Hipertensión} = a + b \cdot [\text{Consumo_sal}]$$

Ecuación 1

Y estimar, a partir de nuestros datos, por el procedimiento habitual de mínimos cuadrados, los coeficientes a y b de la ecuación. Sin embargo, y aunque esto es posible matemáticamente, nos conduce a la obtención de resultados absurdos, ya que cuando se calcule la función obtenida para diferentes valores de consumo de sal se obtendrá resultados que, en general, serán diferentes de 0 y 1, los únicos realmente posibles en este caso, ya que esa restricción no se impone en la regresión lineal, en la que la respuesta puede en principio tomar cualquier valor (SEH-LELHA, 2001).

Si se utiliza cómo variable dependiente la probabilidad p de que un paciente padezca hipertensión y construimos la siguiente función:

$$\ln \frac{p}{1-p}$$

Ecuación 2

Ahora sí tenemos una variable que puede tomar cualquier valor, por lo que podemos plantearnos el buscar para ella una ecuación de regresión tradicional:

$$\ln \frac{p}{1-p} = a + b \cdot [\text{consumo_sal}]$$

Ecuación 3

que se puede convertir con una pequeña manipulación algebraica en

$$\text{Pr. HTA} = \frac{1}{1 + e^{(-a - b \cdot [\text{consumo_sal}])}}$$

Ecuación 4

Y este es precisamente el tipo de ecuación que se conoce como modelo logístico, donde el número de factores puede ser más de uno, así en el exponente que figura en el denominador de la ecuación podríamos tener:

$$b1.\text{consumo_sal} + b2.\text{edad} + b3.\text{sexo} + b4.\text{fumador}$$

1.3.5.8. Series de tiempo

En el mundo existen diferentes tipos de fenómenos, estos pueden ser observados con diferentes dispositivos, pero cada uno de ellos es esencialmente observado en algún periodo de tiempo dando como resultado una serie de tiempo de este fenómeno (Rodríguez Elizalde, 2006).

Una serie de tiempo es un conjunto de valores en un periodo de tiempo, en la literatura existen diferentes definiciones. Algunos investigadores ven a las series de tiempo sólo como valores numéricos, otros son especificados en cada intervalo de tiempo, además de que las series de tiempo pueden ser continuas o discretas (Rodríguez Elizalde, 2006).

Una serie de tiempo es un conjunto de n valores $\{[t1, a1], [t2, a2], \dots, [tn, an]\}$. Los valores son identificados por puntos específicos bien definidos en el tiempo, en tal caso los valores pueden ser vistos como un vector $[a1, a2, \dots, an]$ (Rodríguez Elizalde, 2006).

El análisis de series de tiempo puede ser visto como la tarea de encontrar patrones en los datos y predictibilidad de valores. La detección de patrones puede incluir:

- Tendencias (Trend): El análisis de tendencias puede ser visto como cambios sistemáticos no repetitivos (lineales o no lineales) de algún valor sobre el tiempo. Un ejemplo podría ser el valor de una acción cuando continuamente sube de precio.

- Cíclicos: Aquí el comportamiento observado es cíclico.
- Periódicos: En este los patrones detectados se repiten en base a un periodo de tiempo, ya sea por año, mensual o día. Un ejemplo de ello es cuando los volúmenes de venta aumentan en la temporada navideña.
- Detección de anomalías (outliers): Para ayudar a encontrar patrones, la técnica de detección de anomalías, elimina mucho de los llamados falsos positivos (Rodríguez Elizalde, 2006).

1.3.6. Machine Learning

Machine Learning es una disciplina científica del ámbito de la Inteligencia Artificial que crea sistemas que aprenden automáticamente. Aprender en este contexto quiere decir identificar patrones complejos en millones de datos. La máquina que realmente aprende es un algoritmo que revisa los datos y es capaz de predecir comportamientos futuros. Automáticamente, también en este contexto, implica que estos sistemas se mejoran de forma autónoma con el tiempo, sin intervención humana (Gonzalez, 2014).

En Big Data, las herramientas de análisis tradicionales no son las más adecuadas para capturar el valor total que se puede obtener. El volumen de datos es demasiado grande para un análisis integral tradicional; es demasiado grande para que un analista pueda probar todas las hipótesis y obtenga todo el potencial valor subyacente en los datos. En este contexto, el aprendizaje automático es ideal para aprovechar las oportunidades ocultas en Big Data (Fundación Vodafone España, Red.es & Gobierno de España: Ministerio de Energía, Turismo y Agenda Virtual, 2017).

Machine Learning (junto a Big Data) obtiene más valor de las fuentes de datos, sobre todo si son de estructura heterogénea y de elevado volumen. Además, a diferencia de los análisis

tradicionales, Machine Learning se nutre de conjuntos de datos en constante crecimiento. Cuantos más datos se introducen en un sistema de aprendizaje automático, más puede aprender el algoritmo y obtener resultados de mayor calidad (Fundación Vodafone España, Red.es & Gobierno de España: Ministerio de Energía, Turismo y Agenda Virtual, 2017).

1.3.7. Business Intelligence

El data mining y el big data son importantes por todo lo expuesto anteriormente, en el sentido que ofrecen una gran cantidad de beneficios para el análisis de grandes volúmenes de datos, pero todos estos análisis son estudios vacíos mientras no se los lleve a una aplicación en el mundo real, porque como tal es solo información. Es aquí donde entra el Business Intelligence (BI), pues por medio de dicha información el BI puede generar escenarios, pronósticos y reportes que apoyen a la toma de decisiones, lo que se traduce en una ventaja competitiva. La clave para BI es la información y uno de sus mayores beneficios es la posibilidad de utilizarla en la toma de decisiones (Vallejos & Martínez, 2006).

El BI se puede definir como el proceso de analizar los bienes o datos acumulados en la empresa y extraer una cierta inteligencia o conocimiento de ellos. Dentro de la categoría de bienes se incluyen las bases de datos de clientes, información de la cadena de suministro, ventas personales y cualquier actividad de marketing o fuente de información relevante para la empresa.

BI apoya a los tomadores de decisiones con la información correcta, en el momento y lugar correcto, lo que les permite tomar mejores decisiones de negocios. La información adecuada en el lugar y momento adecuado incrementa efectividad de cualquier empresa (Vallejos & Martínez, 2006).

1.3.8. Análisis de datos en la salud

Al digitalizar y organizar de manera efectiva los datos, las organizaciones de atención médica que van desde consultorios de un solo médico hasta grandes redes hospitalarias y otras organizaciones de atención pueden obtener importantes beneficios. Estos pueden incluir la mejora de la calidad y la eficiencia de la prestación de servicios de salud, como la detección de enfermedades en etapas más tempranas, cuando pueden ser tratadas con mayor éxito (Cottle, y otros, 2013).

Los registros de salud electrónicos (EHR), junto con las nuevas herramientas de análisis, abren la puerta a la información de minería para identificar tendencias estadísticamente válidas y proporcionar evaluaciones basadas en la verdadera calidad de la atención (Cottle, y otros, 2013).

Los sensores electrónicos se utilizan cada vez más para monitorear marcadores bioquímicos clave, dichos sensores arrojan datos de pacientes individuales constantemente en tiempo real los cuales fluyen a los sistemas de análisis de datos. Los análisis de este tipo pueden alertar a individuos específicos y a sus proveedores sobre eventos potencialmente adversos (Cottle, y otros, 2013).

Aplicaciones del análisis de datos en la salud

El informe de la Fundación Vodafone España, Red.es y el Gobierno de España en el que se exponen los resultados de sus investigaciones hechas en cuanto al Big Data se exponen algunos ejemplos de aplicaciones que mejorarían el ámbito de la salud.

Investigación Clínica: La aplicación de la minería de datos al ámbito de la salud promete grandes efectos para los diferentes agentes implicados en la investigación clínica. Por un lado, los laboratorios de análisis clínicos pueden llegar a prestar sus servicios de mayor calidad, debido a la capacidad de optimizar los procesos involucrados en la prestación de dichos servicios. Por su parte, los laboratorios farmacéuticos pueden ver disminuir considerablemente el infra-diagnóstico

de todas aquellas patologías para las que dispone de opciones terapéuticas comercializadas. Aunque el principal beneficiado de la aplicación de la analítica de datos en salud y, concretamente, en la investigación clínica, es precisamente el paciente, que podrá obtener diagnósticos más rápidos y precisos. Dar uso a estos datos podría ayudar a salvar vidas, la auténtica y verdadera finalidad en la prestación de servicios sanitarios

Informática Clínica: La informática clínica se centra en la aplicación de la tecnología de la información en el ámbito de la atención de la salud. Incluye la investigación basada en la actividad, el análisis de la relación entre el diagnóstico principal del paciente y la causa subyacente de muerte y el almacenamiento de datos de EHR y otras fuentes (por ejemplo, datos electrofisiológicos [como EEG]) (Luo, Wu, Gopukumar, & Zhao, 2016).

Actualmente, las decisiones se toman principalmente sobre la información general que ha funcionado anteriormente, o sobre la base de lo que los expertos han encontrado para trabajar en el pasado. El sistema de salud puede adoptar nuevas formas que pueden ser más precisas, confiables y eficientes (Herland, Khoshgoftaar, & Wald, 2014).

Informática de salud pública: La salud pública tiene tres funciones principales: evaluación, desarrollo de políticas, y aseguramiento. La evaluación involucra principalmente la recopilación y el análisis de datos para rastrear y monitorear el estado de salud pública, proporcionando pruebas para la toma de decisiones y el desarrollo de políticas.

La informática de salud pública aplica la minería de datos y el análisis a los datos de la población, con el fin de obtener información médica. Los datos en Informática de Salud Pública provienen de la población, reunidos a partir de medios "tradicionales" (expertos u hospitales) o recopilados de la población (redes sociales) (Herland, Khoshgoftaar, & Wald, 2014).

Operativa clínica: Se puede definir la operativa clínica como el conjunto de decisiones estratégicas, tácticas y operativas sobre la planificación, gestión de los recursos disponibles y/o

gestión de departamentos con la finalidad de optimizar la calidad y la eficiencia de la atención sanitaria.

Las organizaciones sanitarias se enfrentan a nuevos modelos en los que es altamente probable que el análisis de datos clínicos, juegue un papel importante a la hora de prestar asistencia a los pacientes.

De esta manera, se puede conseguir una operativa clínica más efectiva y eficaz, proporcionando información en tiempo real a los técnicos, enfermeras y médicos. Un análisis riguroso de toda esta información concluirá en una mejor gestión de centros sanitarios y, consecuentemente en una mejor distribución del material sanitario y medicamentos.

En este aspecto, una herramienta concreta de análisis de datos, tiene diversas aplicaciones orientadas a la gestión:

- Mejora operativa
- Gestión financiera
- Planificación de recursos
- Inteligencia de procesos clínicos y operativos

En general, la aplicación de los aspectos analíticos derivada del Big Data en salud digital ofrece diversos beneficios a los responsables de gestión de los centros sanitarios:

- Visión global del estado de la organización.
- Seguimiento del cumplimiento de los objetivos estratégicos de la organización a través de indicadores y/o alertas.
- Repositorio único de datos, independientemente de los sistemas de información.
- Acceso directo a los datos sin peticiones al departamento de tecnología.

2. METODOLOGÍA

2.1. Caracterizar los modelos analíticos para repositorios de datos clínicos y técnicos que existen.

Se realizará una investigación para conocer los modelos y métodos que existen actualmente para el análisis de información que se encuentra organizada dentro de repositorios.

Específicamente, el análisis que se pretende llevar a cabo con este trabajo es uno de tipo descriptivo que permita detectar patrones de comportamiento y relaciones en los datos estudiados, ya que teniendo dichos patrones se llegaría a un diagnóstico más detallado sobre los factores que influyen en las labores al interior de la institución de salud, además de ofrecer información que contribuya a optimizar sus procesos, y así responder de manera más contundente ante las eventualidades que se presentan. Los métodos necesarios para hacer estos estudios se enmarcan dentro de la disciplina de la minería de datos

Por medio de una recopilación de información, se estudiarán las técnicas de minería de datos para así determinar cuál(es) modelo(s) se acomoda(n) a los datos clínicos y técnicos que se abordarán en este trabajo, teniendo en cuenta el tipo de información utilizada, la forma como se deben organizar los datos y qué información se brinda como resultado final en cada técnica estudiada.

La revisión bibliográfica se hará a partir de artículos científicos, páginas web y libros especialistas en la minería de datos, pero más que todo enfocado y aplicado en el contexto de instituciones de salud.

2.2. Identificar las fuentes de datos clínicos y técnicos más apropiadas

Luego de investigar las técnicas de minería de datos, se procederá a determinar las fuentes de las que provendrán los datos para hacer el respectivo análisis.

Con estos datos se quiere realizar una simulación, pues se buscarán datos que ofrezcan información de un escenario semejante al que se quiere abordar en este trabajo, para que al momento de enfrentar un caso específico ya se tenga la experiencia necesaria y los análisis diseñados estén a la altura.

Para identificar las posibles fuentes de datos se realizará una revisión bibliográfica sobre qué tipo de datos ofrecen información más relevante, indagando en artículos científicos que detallen estudios similares, páginas web y bases de datos con libre acceso donde se hallen datos que puedan ser muy bien aprovechados.

2.3. Diseñar un repositorio de datos clínicos y técnicos

Ya obtenidos los datos, se procederá a desarrollar el repositorio donde se integrarán los datos ya existentes junto con los datos que serán ingresados en el futuro. Para ello se diseñará una base de datos transaccional utilizando Microsoft SQL Server Management Studio, desde donde se importarán los datos en formato de hojas de cálculo de Excel (.xlsx) y formato texto (.txt) que fueron proporcionados por la institución de salud.

Teniendo en cuenta la estructura de los datos y la forma en como están organizados se creará una interfaz gráfica en el entorno GUIDE (Graphical User Interfaces Development Environment) de Matlab, la cual permitirá la visualización de los datos almacenados en la base de datos utilizando las opciones proporcionadas por la extensión Database Toolbox, también de Matlab, la cual permite integrar las instancias de SQL Server y las bases de datos asociadas a ellas con el entorno de Matlab. A través de búsquedas se podrá filtrar la información que se necesite extraer y se mostrarán los datos de forma más depurada y organizada.

Para proceder con el análisis de los datos, es necesario primeramente realizar un tratamiento de ETL (Extraer, Transformar and Cargar por sus siglas en inglés). El ETL que consiste en depurar la información mediante procesos de limpieza, de filtrado, división de una columna en

múltiples columnas y viceversa, uniendo datos de múltiples fuentes o mediante la transposición de filas y columnas. Todo esto con el fin de contar con datos más claros y mucho más comprensibles, ya que mucha de la información ingresada en las bases de datos se encuentra estructurada de tal forma que no puede ser interpretada por los softwares destinados a los procesos de minería de datos. El resultado final del ETL será un DataWarehouse en el que se encuentren dispuestos los datos con la información más relevante para lograr un análisis preciso.

Con el DataWarehouse construido se procederá a aplicar los métodos de minería de datos escogidos en la matriz desarrollada en la primera actividad utilizando las herramientas de Analissys Services de Visual Studio 2015, este análisis brindará información más detallada sobre los procesos llevados en el laboratorio clínico de la entidad dependiendo de los métodos escogidos.

2.4. Evaluar la propuesta de repositorio de datos clínicos y técnicos

A través de un trabajo de tesis doctoral, se estableció un convenio entre la Universidad EIA y una institución de salud en la que se acordó un intercambio de datos con información del área de laboratorio clínico e historia clínica de esa institución, con el fin de llegar a una optimización de los procesos llevados a cabo en estas áreas por medio de un análisis de minería de datos. Se pretende que con este trabajo de tesis de pregrado se apoye a dicho análisis de Big Data a través de los modelos de minería de datos que se utilizarán.

Para formalizar este intercambio de información se debe estructurar un contrato con un acuerdo de confidencialidad donde ambas partes se comprometan a mantener secreta la información de los datos y de los resultados obtenidos de la investigación. Posteriormente este acuerdo pasará a ser verificado por el Comité de Ética de la institución quienes darán el visto bueno al acuerdo pactado y tendrán la última palabra en cuanto a las condiciones estipuladas en el acuerdo.

Luego de los trámites pertinentes, se adquirirán los datos y se analizará la problemática en cuestión para contribuir por medio de esta investigación a que se preste un servicio de calidad a los pacientes que llegan cada día a la entidad, aplicando la metodología desarrollada en el simulacro previamente descrito.

3. PRESENTACIÓN Y DISCUSIÓN DE RESULTADOS

3.1. SELECCIONAR LAS HERRAMIENTAS A UTILIZAR

3.1.1. Almacenamiento y organización de los datos

Teniendo datos proporcionados por la institución de salud se procede a seleccionar el software en el que se iniciará su tratamiento y organización, un software en el que se puedan importar dichos datos y que además ofrezca un entorno amigable.

Para seleccionar un programa adecuado, se recurre a la información brindada por Gartner, Inc. A través de los recursos de Gartner Research, Gartner proporciona el análisis de investigación y el consejo para profesionales de las TIC (tecnologías de la información y la comunicación), empresas de tecnología y la comunidad de la inversión. La información de esos análisis es presentada en forma gráfica mediante los Cuadrantes Mágicos de Gartner. Un Cuadrante Mágico de Gartner es resultado final de la investigación en un mercado específico, que brinda una visión panorámica de las posiciones relativas de los competidores del mercado. Un Cuadrante Mágico ayuda a determinar rápidamente qué tan bien los proveedores de tecnología están ejecutando sus visiones declaradas y qué tan bien se están desempeñando en relación con la visión del mercado de Gartner (Gartner, Inc., 2018). En la figura 7 se muestra la estructura de un Cuadrante Mágico.



Figura 6. Estructura de un Cuadrante Mágico de Gartner (Gartner, Inc., 2018).

Un Cuadrante Mágico proporciona un posicionamiento gráfico competitivo de cuatro tipos de proveedores de tecnología dentro de un mismo mercado los cuales se describen de la siguiente forma: Los LEADERS se ejecutan bien contra su visión actual y están bien posicionados para el mañana. Los VISIONARIES entienden a dónde va el mercado o tienen una visión para cambiar las reglas del mercado, pero aún no se ejecutan bien. Los NICHE PLAYERS se enfocan con éxito en un segmento pequeño, o están desenfocados y no superan o superan a otros. Los CHALLENGERS se ejecutan bien hoy o pueden dominar un segmento grande, pero no demuestran una comprensión de la dirección del mercado. (Gartner, Inc., 2018)

Con los resultados presentados de esta forma se hace más comprensible los respectivos reportes de las investigaciones realizadas y de forma rápida se puede conocer los proveedores de tecnología competidores de un mercado y su capacidad para cumplir con lo que los usuarios finales requieren hoy y en el futuro, además de comprender cómo los proveedores de tecnología de un mercado están posicionados competitivamente y las estrategias que están utilizando para competir por el negocio del usuario final (Gartner, Inc., 2018). Teniendo así una comparación de las fortalezas y los desafíos de un proveedor de tecnología con nuestras necesidades específicas.

Sabiendo que Gartner es una fuente confiable y con credibilidad a nivel mundial para la elección de sistemas de información para determinadas funciones, acudimos al Cuadrante Mágico de Gartner para la categoría de Sistemas Operativos de Gestión de Bases de Datos, categoría en la que se muestran las herramientas existentes para administración de bases de datos operativas (DBMS) que admiten múltiples estructuras y tipos de datos, como XML, texto, audio, imágenes y contenido de video. Las herramientas DBMS también incluyen mecanismos para aislar los recursos de carga de trabajo y controlar varios parámetros de acceso de usuario final dentro de las instancias administradas de los datos (Zicardi, 2015).



Figura 7. Cuadrante Mágico del 2017 para Sistemas Operativos de Gestión de Bases de Datos (Microsoft Corporation, 2017).



Figura 8. Cuadrante Mágico del 2016 para Sistemas Operativos de Gestión de Bases de Datos (Walker, 2016).

Al observar los Cuadrantes Mágicos de los últimos tres años, se evidencia un liderazgo por parte de Microsoft en cada uno de ellos, no solo por estar encasillado dentro de los LEADERS sino que es el de mejor calificación en todos ellos de manera consecutiva, aunque también es importante mencionar que siempre se encuentra asediado por Oracle, siendo ambas compañías y sus sistemas para el manejo de bases de datos unas opciones muy atractivas para escoger.



Figura 9. Cuadrante Mágico del 2015 para Sistemas Operativos de Gestión de Bases de Datos (Naeem, 2015).

Los LEADERS generalmente demuestran una satisfacción constante del cliente y un sólido soporte al cliente. La mensajería, la I + D de productos y la entrega de LEADERS están en línea con el mercado actual y con las nuevas tendencias tanto en software como en tecnología de hardware (Zicardi, 2015).

Al comparar ambos servicios, se considera que la base de datos Oracle es mucho más compleja que MS SQL Server de Microsoft, ya que está destinado a organizaciones más grandes donde se necesita una gran base de datos. Si bien MS SQL Server ofrece una versión empresarial, ambos son ampliamente utilizados en todo el entorno empresarial. Pero cuál se considera superior es una cuestión de preferencia y para qué se utiliza esa base de datos en particular. MS SQL Server es ideal para principiantes de bases de datos, mientras que Oracle es óptimo para aquellos que administran grandes cantidades de datos (Lynch, 2017).

La elección finalmente se decanta por escoger los servicios de Microsoft, más que todo por la familiaridad de las herramientas Microsoft con respecto a las aplicaciones de Oracle y la tendencia que ha tenido la primera compañía por la integración de todos sus softwares desarrollados, en este caso con sus herramientas de Business Intelligence para el análisis de

datos, que es en última instancia el objetivo principal de este trabajo. Un aspecto que también favoreció la inclinación por Microsoft es la simplicidad del lenguaje Transact SQL para base de datos de MS SQL Server comparado con el Procedural Language SQL de Oracle Database.

3.1.2. Diseño de la interfaz gráfica de usuario

Una Interfaz gráfica de usuario (En inglés Graphic User Interface, también conocido con su acrónimo GUI) es un método para facilitar la interacción del usuario con el ordenador o la computadora a través de la utilización de un conjunto de imágenes y objetos pictóricos (iconos, ventanas...) además de texto. Surge como evolución de la línea de comandos de los primeros sistemas operativos y es pieza fundamental en un entorno gráfico.

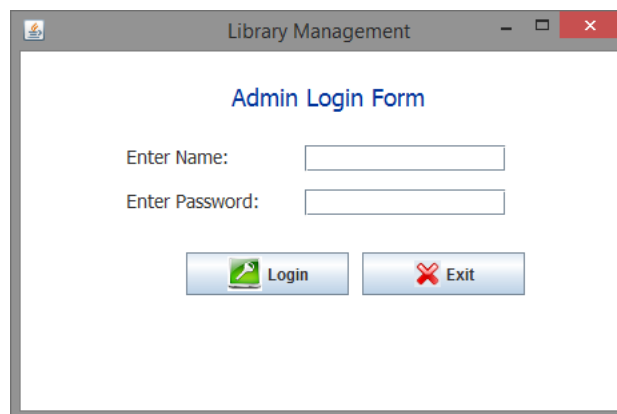


Figura 10. Ejemplo GUI (Shakya, 2017).

Existen principios relevantes para el diseño e implementación de Interfaces de usuario (IU) ya sean para IU graficas como para la web:

- **Autonomía:** La computadora, la IU y el entorno de trabajo deben estar a disposición del usuario. Se debe dar al usuario el ambiente flexible para que pueda aprender rápidamente a usar la aplicación.

- Percepción del Color: Aunque se utilicen convenciones de color en la IU, se deberían usar otros mecanismos secundarios para proveer la información a aquellos usuarios con problemas en la visualización de colores.
- Legibilidad: Para que la IU favorezca la usabilidad del sistema de software, la información que se exhiba en ella debe ser fácil de ubicar y leer. Es importante hacer clara la presentación visual (colocación /agrupación de objetos, evitar la presentación de excesiva información (González Sojo, s.f.).

Interfaz Gráfica de Usuario en Matlab

GUIDE es un entorno de programación visual disponible en Matlab para realizar y ejecutar programas que necesiten ingreso continuo de datos. Tiene las características básicas de todos los programas visuales como Visual Basic o Visual C++.

Una aplicación GUIDE consta de dos archivos: .m y .fig. El archivo .m es el que contiene el código con las correspondencias de los botones de control de la interfaz y el archivo .fig contiene los elementos gráficos.

Cada vez que se adicione un nuevo elemento a la interfaz gráfica, se genera automáticamente código en el archivo .m.

Para ejecutar una Interfaz Gráfica, si se ha etiquetado con el nombre curso.fig simplemente se ejecuta en la ventana de comandos >> curso. O haciendo click derecho en el m-file y seleccionando la opción RUN.

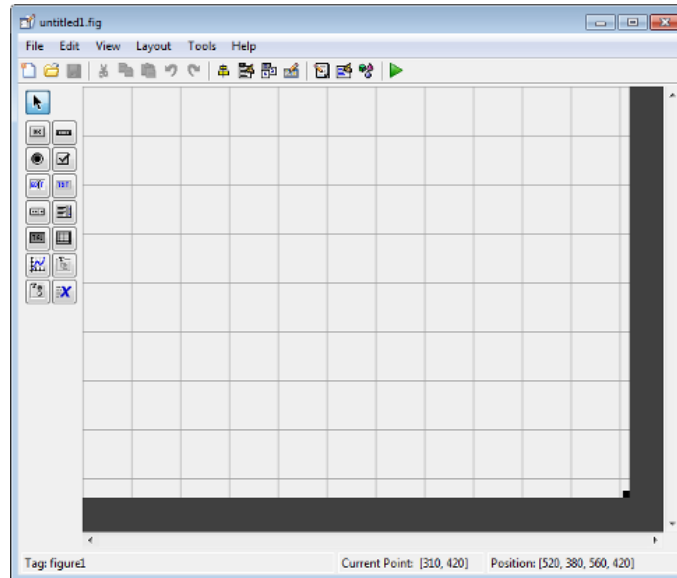


Figura 11. Entorno GUIDE de Matlab (MathWorks, s.f.).

3.1.3. ETL y Minería de datos

Se escoge la herramienta de Microsoft SQL Server Data Tools, ya que la integración de esta herramienta con MS SQL Server fue una de las razones para escoger la aplicación de Microsoft para cumplir la tarea de almacenamiento y tratamiento de los datos. Se utilizarán el recurso de Integration Services para el ETL. El SQL Server Integration Services Import/Export Wizard permite mover datos de un origen a destino sin modificar los datos del origen y permitiendo hacer iteraciones y cambios de información antes de llegar al destino dentro de tablas de ETL. Microsoft Integration Services es una plataforma para la creación de soluciones empresariales de transformaciones de datos e integración de datos. Integration Services sirve para resolver complejos problemas empresariales mediante la copia o descarga de archivos, la carga de almacenamientos de datos, la limpieza y minería de datos y la administración de datos y objetos de SQL Server (Microsoft, 2018).

Para implementar los modelos de minería de datos, se recurre a la herramienta de Microsoft Analysis Services. Microsoft SQL Server Analysis Services (SSAS) es una herramienta de

procesamiento analítico en línea (OLAP) y de minería de datos en Microsoft SQL Server. SSAS es utilizado como una herramienta por las organizaciones para analizar y dar sentido a la información que posiblemente se distribuye en múltiples bases de datos, o en tablas o archivos dispares. Analysis Services incluye un grupo de capacidades de OLAP y minería de datos y viene en dos tipos: Multidimensional y Tabular (Agile Design, s.f.). SSAS cuenta con los modelos de Árboles de decisión, Naive Bayes, Clusters, Redes Neuronales, Reglas de asociación, Regresión lineal y Regresión logística para el análisis de los datos ingresados en su herramienta.


3.2. FUENTES DE DATOS CLÍNICOS Y TÉCNICOS

Para realizar las pruebas del funcionamiento de la herramienta de almacenamiento de datos, mientras los datos de las instituciones de salud eran proporcionados, se hizo una revisión en bases de datos de libre acceso, hasta llegar al UCI Machine Learning Repository. Este repositorio es una colección de bases de datos, teorías de dominio y generadores de datos que utiliza la comunidad de Machine Learning para el análisis empírico de los algoritmos de aprendizaje automático. Desde su creación ha sido ampliamente utilizado por estudiantes, educadores e investigadores de todo el mundo como fuente principal de conjuntos de datos de aprendizaje automático (University of California, Irvine, 2018). Como una indicación del impacto del repositorio, se ha citado más de 1000 veces, lo que lo convierte en uno de los 100 más citados de toda la informática.

Buscando dentro de dicho repositorio se encuentran datos de todo tipo y organizados en distintas categorías siguiendo criterios específicos como el tipo de análisis realizado con los datos, el tipo de atributos en los métodos analíticos utilizados, el área de conocimiento al que pertenece el contexto de los datos, entre otros.

El criterio de depuración para escoger los datos que se iban a usar en este trabajo para hacer el entrenamiento pertinente a la hora de trabajar con MS SQL Server de Microsoft fue el área de conocimiento en el que eran usados los datos. El contexto deseado, es uno de tipo hospitalario en el que se puedan analizar los patrones de ciertas características de los pacientes, sin incurrir en fallas a la ética como por ejemplo que se mantenga protegida la identidad de los pacientes contenidos dentro de la base datos escogida, asignándole una variable ID para distinguir cada persona en la base de datos, como lo haría una institución de salud.

Fruto de la búsqueda se llega a parar en el siguiente conjunto de datos:



UCI
Machine Learning Repository
Center for Machine Learning and Intelligent Systems

Diabetes 130-US hospitals for years 1999-2008 Data Set

Download: [Data Folder](#), [Data Set Description](#)

Abstract: This data has been prepared to analyze factors related to readmission as well as other outcomes pertaining to patients with diabetes.

Data Set Characteristics:	Multivariate	Number of Instances:	100000	Area:	Life
Attribute Characteristics:	Integer	Number of Attributes:	55	Date Donated	2014-05-03
Associated Tasks:	Classification, Clustering	Missing Values?	Yes	Number of Web Hits:	206589

Figura 12. Fuente de datos (University of California, Irvine, 2018).

Este conjunto de datos representa 10 años (desde 1999 a 2008) de atención clínica en 130 hospitales de EE. UU. Incluye más de 50 características que representan los resultados de pacientes y hospitales (University of California, Irvine, 2018). La información se extrajo de la base de datos para los encuentros que cumplieron con los siguientes criterios.

- Es un encuentro hospitalario (ingreso hospitalario).

- Es un encuentro diabético, es decir, durante el encuentro se ingresó en el sistema cualquier tipo de diabetes como diagnóstico.
- La duración de la estancia fue de al menos 1 día y como máximo 14 días.
- Se realizaron pruebas de laboratorio durante el encuentro.
- Se administraron medicamentos durante el encuentro (Strack, y otros, 2014).

Los datos contienen atributos tales como número de paciente, raza, sexo, edad, tipo de ingreso, tiempo en el hospital, especialidad médica del médico de admisión, número de pruebas de laboratorio realizadas, resultado de la prueba de HbA1c, diagnóstico, número de medicamentos, medicamentos para la diabetes, número de pacientes ambulatorios, hospitalización y visitas de emergencia en el año anterior a la hospitalización, etc.

La descripción detallada de todos los atributos se proporciona en el ANEXO 17.

Estos datos fueron obtenidos de la base de datos Health Facts (Cerner Corporation, Kansas City, MO), un data warehouse nacional que recopila registros clínicos completos en todos los hospitales de los Estados Unidos. Health Facts es un programa voluntario que se ofrece a las organizaciones que utilizan el Cerner Electronic Health Record System. La base de datos contiene datos recopilados sistemáticamente de las historias clínicas electrónicas de las instituciones participantes e incluye datos de encuentros (de emergencia, ambulatorios e internos), especialidad del proveedor, datos demográficos (edad, sexo y raza), diagnósticos y procedimientos hospitalarios documentados por los códigos ICD-9- CM, datos de laboratorio, datos de farmacia, mortalidad hospitalaria y características del hospital (Strack, y otros, 2014). Todos los datos fueron deidentificados de conformidad con la Ley de Responsabilidad y Portabilidad del Seguro de Salud de 1996 antes de ser entregados a los investigadores. Este conjunto de datos ya ha sido utilizado previamente en proyectos de investigación, como el que se

describe en el artículo *“Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records”* publicado en la revista BioMed Research International, en el que se realizó el análisis de esta base de datos clínicos para proporcionar una evaluación clínica y encontrar futuras direcciones que podrían conducir a mejoras en la seguridad del paciente.

3.3. ALMACENAMIENTO Y ORGANIZACIÓN DE LOS DATOS

Se procedió con la descarga de los datos del UCI Machine Learning Repository. Con la descarga se obtuvo una carpeta con dos archivos, un archivo de valores separado por comas (.csv) que contiene los datos descritos anteriormente con las características detalladas en la Tabla 1 y el segundo archivo es una hoja de cálculo de Excel en la que se especifican los valores nominales a los que corresponden los enteros que están asignados en las columnas de las características Admission type, Discharge disposition y Admission source. En los anexos 18, 19 y 20 se muestra la información proporcionada por la hoja de cálculo de Excel organizada en tablas.

Además, se construyó una tabla con las enfermedades para los diagnósticos primarios, secundarios y terciarios según siguiendo la codificación como los tres primeros dígitos de ICD9 agrupándolos por el tipo de enfermedad, como se muestra en la tabla de anexo 21.

Al haber interactuado y conocido los datos obtenidos del repositorio de Machine Learning se continúa el proceso de organización y almacenamiento utilizando MS SQL Server, creando una nueva base de datos a partir de los datos descargados.

Para llevar a cabo esta tarea, primero se crea una nueva base de datos en SQL Server y se le asigna un nombre haciendo clic derecho sobre la opción Databases del Object Explorer y se elige New Database:

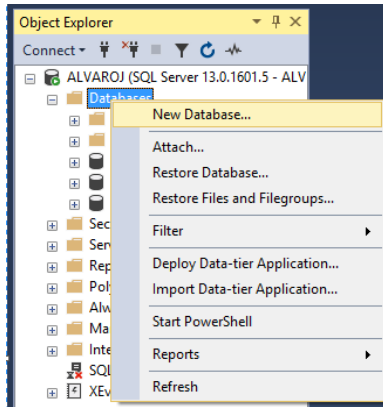


Figura 13. Crear una nueva base de datos (Fuente propia, 2018).

Luego en la ventana que se despliega se le asigna un nombre a la nueva base de datos, para este caso se le dio el nombre de “OpenSource”:

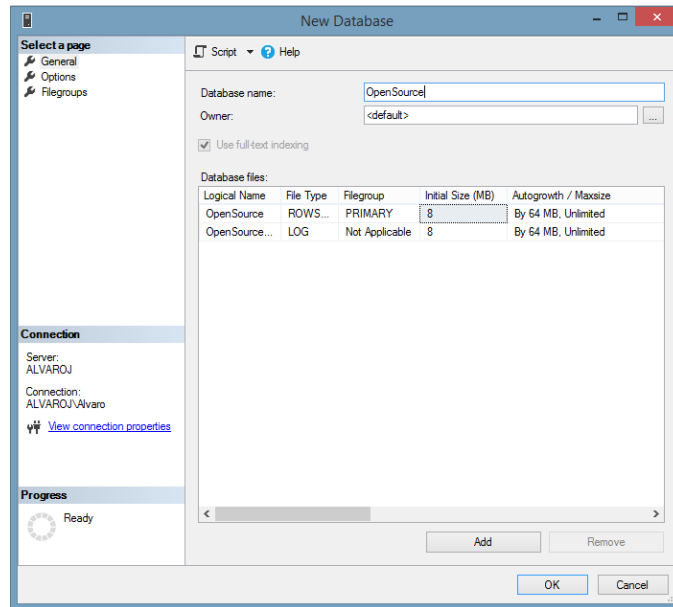


Figura 14. Asignar nombre a la nueva base de datos (Fuente propia, 2018).

Con la nueva base de datos creada, el paso a seguir fue crear las tablas de la base de datos con toda la información contenida en los archivos descargados. Para ello se acude al Asistente de importación y exportación de SQL Server, el cual es una herramienta que ofrece una forma sencilla de copiar datos de un origen a un destino:

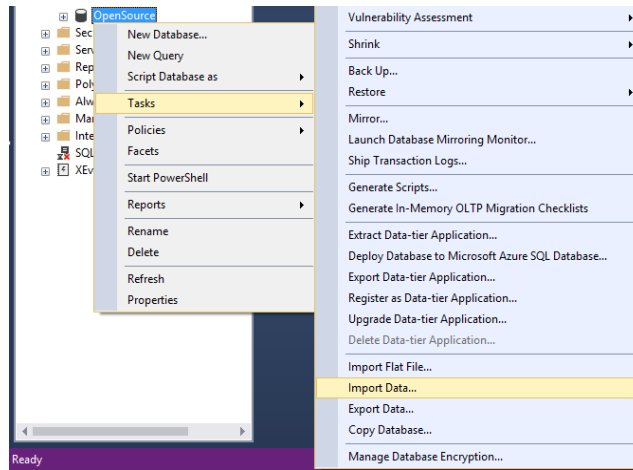


Figura 15. Abrir el Asistente para importación y exportación de SQL Server (Fuente propia, 2018).

Ya dentro del asistente en el campo Data source se escoge la opción Flat file source (Fuente de archivo plano), y se escoge el archivo .csv para crear la tabla con los datos de los pacientes. Para crear las tablas con la descripción de las características Admission type, Discharge disposition y Admission source, se escoge como Data source la opción Microsoft Excel debido a que este es el formato en el que se encuentran guardados, y se busca la ubicación de la respectiva hoja de cálculo.

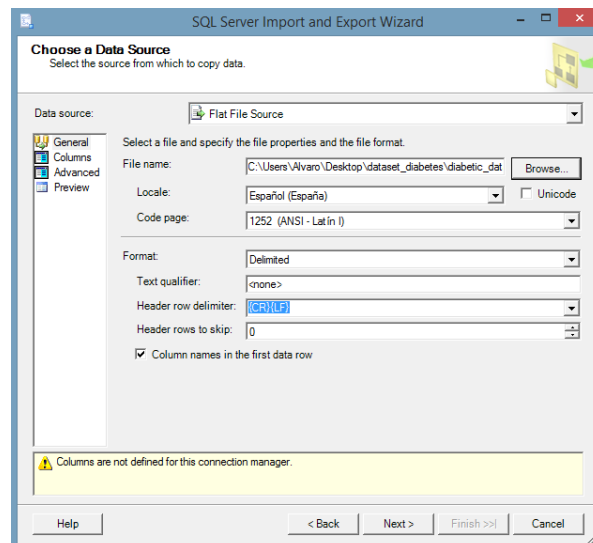


Figura 16. Definición del origen de datos para la importación (Fuente propia, 2018).

Después se escoge el destino de los datos escogido. En el campo Destination se selecciona la opción Microsoft OLE DB Provider for SQL Server y en el campo Database se elige la base de datos en la cual se importarán los datos, para este caso se escoge la base de datos OpenSource que ha sido creada,

El siguiente paso es definir los nombres de las tablas que contendrán los datos importados. Los nombres escogidos para las tablas son: “diabetic_data” para la tabla con la información de los pacientes y, “admission_source”, admission_type” y “discharge_disposition” para las tablas con las descripciones de las respectivas características, como se muestra en la Figura 19.

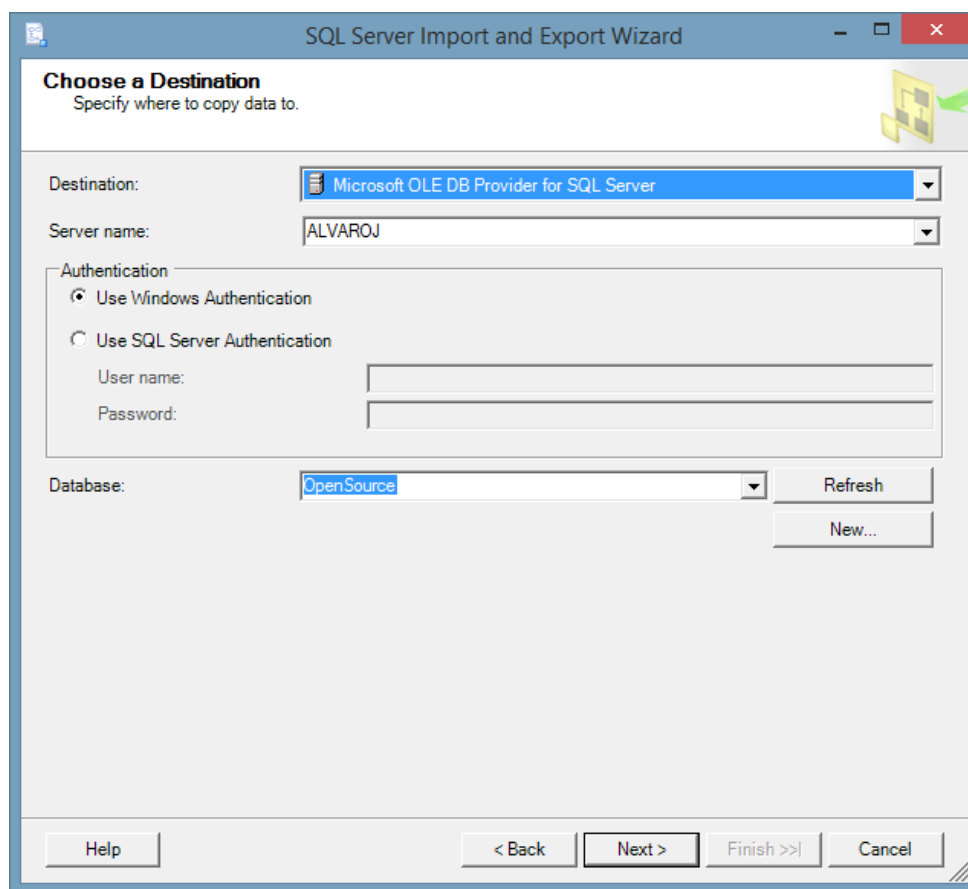


Figura 17. Definición del destino de importación (Fuente propia, 2018).

Para finalizar la importación de los datos se presiona el botón Next las veces que sea necesario hasta que aparece una ventana con el botón Finish. Se espera que se haga la importación de los

datos en las nuevas tablas creadas y que arroje un mensaje de que se ejecutaron todos los procesos de forma exitosa. Para salir del Asistente de importación y exportación de SQL Server se presiona el botón Close.

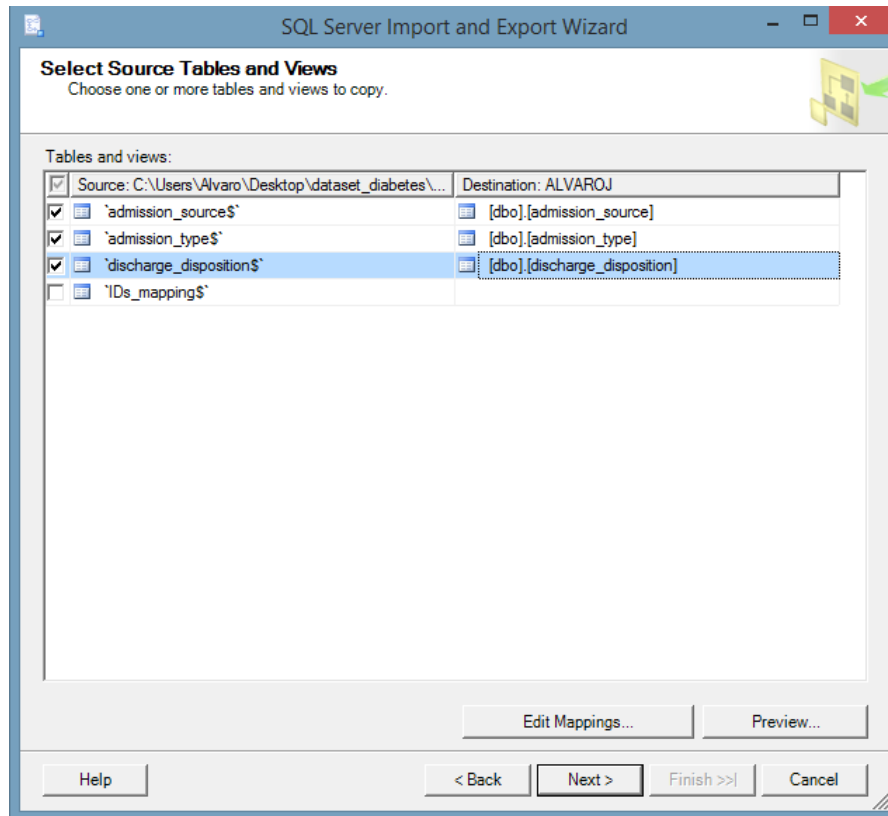


Figura 18. Selección de nombres para las tablas (Fuente propia, 2018).

Terminada la importación se tienen los datos organizados en cada una de las tablas creadas dentro de la base de datos OpenSource.

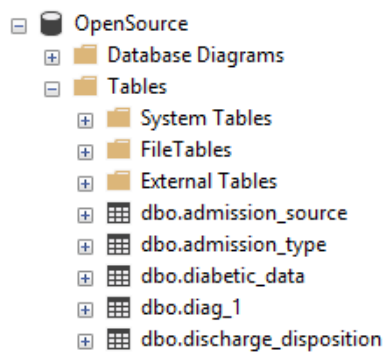


Figura 19. Tablas OpenSource (Fuente propia, 2018).

encounter_id	patient_nbr	race	gender	age	weight	admission_type_id	discharge_disposition_id	admission_source_id	time_in_hospital	payer_code	medical_specialty
1	2278392	8222157	Caucasian	Female	[0-10]	?	6	25	1	?	Pediatrics-Endoc
2	149190	55629189	Caucasian	Female	[10-20]	?	1	1	7	?	?
3	64410	86047875	AfricanAmerican	Female	[20-30]	?	1	1	7	?	?
4	500364	82442376	Caucasian	Male	[30-40]	?	1	1	7	?	?
5	16680	42519267	Caucasian	Male	[40-50]	?	1	1	7	?	?
6	35754	82637451	Caucasian	Male	[50-60]	?	2	1	2	3	?
7	55842	84259809	Caucasian	Male	[60-70]	?	3	1	2	4	?
8	63768	114882984	Caucasian	Male	[70-80]	?	1	1	7	5	?
9	12522	48330783	Caucasian	Female	[80-90]	?	2	1	4	13	?
10	15738	63555939	Caucasian	Female	[90-100]	?	3	3	4	?	InternalMedicine
11	28236	89869032	AfricanAmerican	Female	[40-50]	?	1	1	7	9	?
12	36900	77391171	AfricanAmerican	Male	[60-70]	?	2	1	4	7	?
13	40926	85504905	Caucasian	Female	[40-50]	?	1	3	7	7	Family/GeneralP
14	42570	77586282	Caucasian	Male	[80-90]	?	1	6	7	10	Family/GeneralP
15	62256	49726791	AfricanAmerican	Female	[60-70]	?	3	1	2	1	?
16	73578	86328819	AfricanAmerican	Male	[60-70]	?	1	3	7	12	?
17	77076	92519352	AfricanAmerican	Male	[50-60]	?	1	1	7	4	?
18	84222	108662661	Caucasian	Female	[50-60]	?	1	1	7	3	Cardiology
19	89682	107389323	AfricanAmerican	Male	[70-80]	?	1	1	7	5	?
20	148530	69422211	?	Male	[70-80]	?	3	6	2	6	?
21	150006	22864131	?	Female	[50-60]	?	2	1	4	2	?
22	150048	21239181	?	Male	[60-70]	?	2	1	4	2	?
23	182796	63000108	AfricanAmerican	Female	[70-80]	?	2	1	4	2	?

Query executed successfully. ALVAROJ (13.0 RTM) | ALVAROJ,Alvaro (57) | OpenSource | 00:00:27 | 101766 rows

Figura 20. Datos tabla diabetic_data (Fuente propia, 2018).

admission_type_id	description
1	Emergency
2	Urgent
3	Elective
4	Newborn
5	Not Available
6	NULL
7	Trauma Center
8	Not Mapped

Figura 21. Datos tabla admission_type (Fuente propia, 2018).

admission_source_id	description
1	Physician Referral
2	Clinic Referral
3	HMO Referral
4	Transfer from a hospital
5	Transfer from a Skilled Nursing Facility (SNF)
6	Transfer from another health care facility
7	Emergency Room
8	Court/Law Enforcement
9	Not Available
10	Transfer from critical access hospital
11	Normal Delivery
12	Premature Delivery

Figura 22. Datos tabla admission_source (Fuente propia, 2018).

discharge_disposition_id	description
1	Discharged to home
2	Discharged/transferred to another short term hospital
3	Discharged/transferred to SNF
4	Discharged/transferred to ICF
5	Discharged/transferred to another type of inpatient...
6	Discharged/transferred to home with home health ...
7	Left AMA
8	Discharged/transferred to home under care of Ho...
9	Admitted as an inpatient to this hospital
10	Neonate discharged to another hospital for neonat...
11	Expired
12	Still patient or expected to return for outpatient ser...

Figura 23. Datos tabla discharge_disposition (Fuente propia, 2018).

Para el análisis de minería de datos el último paso es agrupar y almacenar los datos de todas las tablas en una sola tabla. En SQL Server existen las vistas, las cuales son objetos en las bases de datos creados como resultado de una consulta o query. Para crear una vista de la base de datos el primer paso es crear un query con el que se puedan concentrar los datos de todas las tablas dentro del mismo query. El código de dicho query se puede encontrar en el apartado de anexos de este trabajo con el nombre de Anexo 1. Lo siguiente es crear la nueva vista en la base de datos deseada.

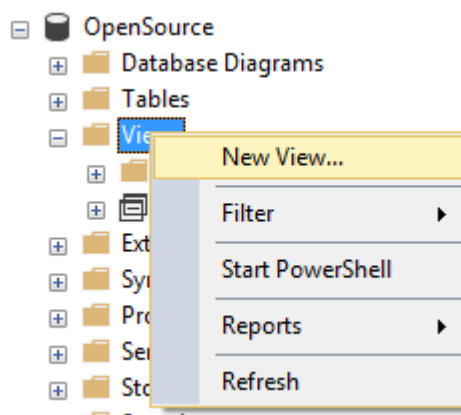


Figura 24. Creación de una nueva vista de base de datos (Fuente propia, 2018).

Luego de tener la vista creada se copia el código del query anteriormente definido en el área destinada para los códigos SQL de la vista, y como paso final se ejecuta la búsqueda a través del

método abreviado de SQL “Ctrl + R”, dando como resultado final los datos organizados y almacenados en una sola tabla listos para el análisis de minería de datos. Finalmente se guarda la vista y se le asigna un nombre (opensource_view fue el nombre escogido para este vista).

Encuentro	Paciente	Etnia	Genero	Edad	TipoAdmision	DescripcionAlta	OrigenAdmision	TiempoHospital	TipoPago	Espec	
1	2278392	8222157	Caucasian	Female	[0-10]	NULL	Not Mapped	Physician Referral	1	?	Pedia
2	149190	55629189	Caucasian	Female	[10-20]	Emergency	Discharged to home	Emergency Room	3	?	?
3	64410	86047875	AfricanAmerican	Female	[20-30]	Emergency	Discharged to home	Emergency Room	2	?	?
4	500364	82442376	Caucasian	Male	[30-40]	Emergency	Discharged to home	Emergency Room	2	?	?
5	16680	42519267	Caucasian	Male	[40-50]	Emergency	Discharged to home	Emergency Room	1	?	?
6	35754	82637451	Caucasian	Male	[50-60]	Urgent	Discharged to home	Clinic Referral	3	?	?
7	55842	84259809	Caucasian	Male	[60-70]	Elective	Discharged to home	Clinic Referral	4	?	?
8	63768	114882984	Caucasian	Male	[70-80]	Emergency	Discharged to home	Emergency Room	5	?	?
9	12522	48330783	Caucasian	Female	[80-90]	Urgent	Discharged to home	Transfer from a hospital	13	?	?
10	15738	63555939	Caucasian	Female	[90-100]	Elective	Discharged/transferred to SNF	Transfer from a hospital	12	?	Intern
11	28236	89869032	AfricanAmerican	Female	[40-50]	Emergency	Discharged to home	Emergency Room	9	?	?
12	36900	77391171	AfricanAmerican	Male	[60-70]	Urgent	Discharged to home	Transfer from a hospital	7	?	?
13	40926	85504905	Caucasian	Female	[40-50]	Emergency	Discharged/transferred to SNF	Emergency Room	7	?	Family
14	42570	77586282	Caucasian	Male	[80-90]	Emergency	Discharged/transferred to home with home health ...	Emergency Room	10	?	Family
15	62256	49726791	AfricanAmerican	Female	[60-70]	Elective	Discharged to home	Clinic Referral	1	?	?
16	73578	86328819	AfricanAmerican	Male	[60-70]	Emergency	Discharged/transferred to SNF	Emergency Room	12	?	?
17	77076	92519352	AfricanAmerican	Male	[50-60]	Emergency	Discharged to home	Emergency Room	4	?	?
18	84222	108662661	Caucasian	Female	[50-60]	Emergency	Discharged to home	Emergency Room	3	?	Cardic
19	89682	107389323	AfricanAmerican	Male	[70-80]	Emergency	Discharged to home	Emergency Room	5	?	?
20	148530	69422211	?	Male	[70-80]	Elective	Discharged/transferred to home with home health ...	Clinic Referral	6	?	?
21	150006	22864131	?	Female	[50-60]	Urgent	Discharged to home	Transfer from a hospital	2	?	?

Query executed successfully. | ALVAROJ (13.0 RTM) | ALVAROJ\Alvaro (53) | OpenSource | 00:00:28 | 101766 rows

Figura 25. Datos organizados y almacenados en la vista opensource_view (Fuente propia, 2018).

3.4. DISEÑO DE LA INTERFAZ GRÁFICA DE USUARIO

Al ser Matlab la herramienta escogida para el diseño de la Interfaz Gráfica de Usuario (GUI), se debe garantizar que se cuente con el entorno GUIDE para iniciar el respectivo proceso.

La función principal del GUI será la búsqueda de datos en la base de datos OpenSource, específicamente se pretende extraer la información de contenida dentro de la vista opensource_view, la cual es la que posee los datos de forma más organizada, con una mejor estructura en su disposición.

Antes de empezar con la creación de la GUI, se debe verificar que la versión de Matlab con la que se trabajará posea la Database Explorer App dentro de sus aplicaciones adicionales instaladas. Con la Database Explorer App se puede conectar el entorno de Matlab de manera rápida a una base de datos, explorar los datos de la base de datos e importar datos al Workspace de Matlab de manera visual.

Habiendo corroborado que las herramientas necesarias están en total disposición para comenzar el trabajo, el primer paso es crear y configurar la fuente de datos que se utilizara en Matlab, para ello se ingresa a la aplicación Database Explorer, se presiona el botón Configure Data Source y se elige la opción Configure ODBC Data Source, esto con el fin de establecer una conexión con el ODBC driver instalado en el equipo, ya que es el Microsoft ODBC Driver el cual proporciona conectividad nativa desde Windows a Microsoft SQL Server.

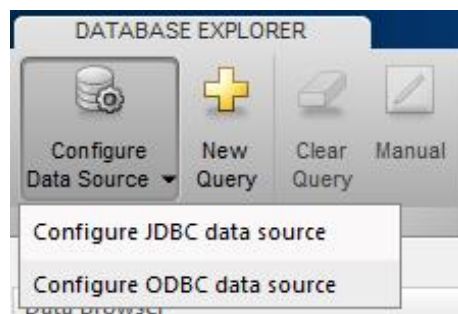


Figura 26. Configuración origen de datos en Database Explorer App (Fuente propia, 2018).

A continuación, aparecerá el Administrador de origen de datos ODBC. Aquí se ingresa a las opciones de la pestaña DNS de sistema y se da clic en Agregar.

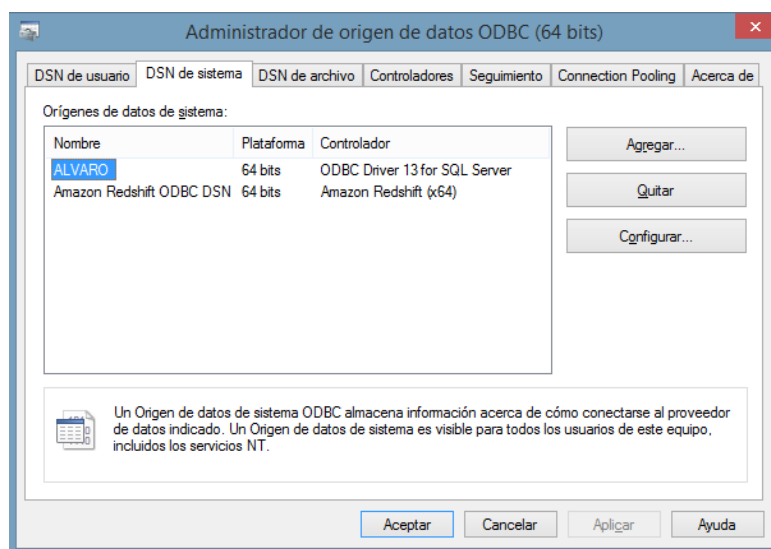


Figura 27. Administrador de origen de datos ODBC (Fuente propia, 2018).

Luego se selecciona el controlador para el que se desee establecer un origen de datos, para este caso se elige ODBC Driver for SQL Server y se da clic en Finalizar.

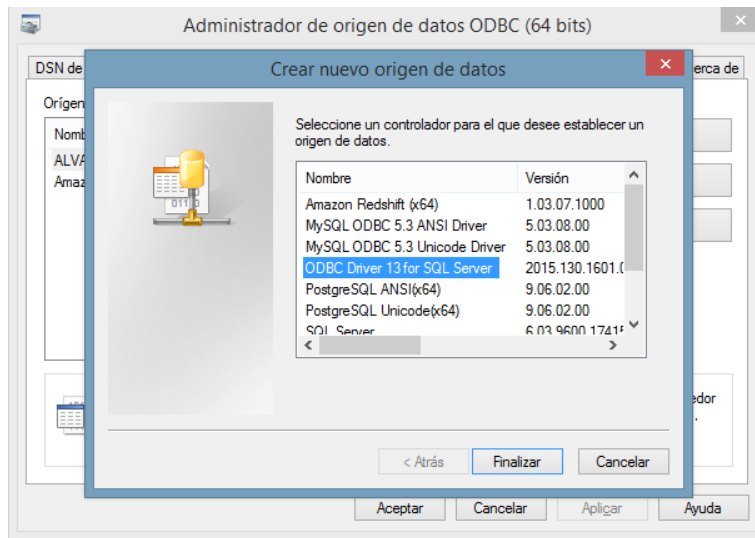


Figura 28. Selección de controlador para establecer el origen de datos (Fuente propia, 2018).

Para terminar con la configuración del origen de datos se le da un nombre y se especifica el servidor de SQL Server del que provendrán todos los datos, en esta ocasión al origen se le fue asignado el nombre de ALVARO. Se le da clic en Siguiente en esta ventana y todas las ventanas subsiguientes hasta dar clic en Finalizar.

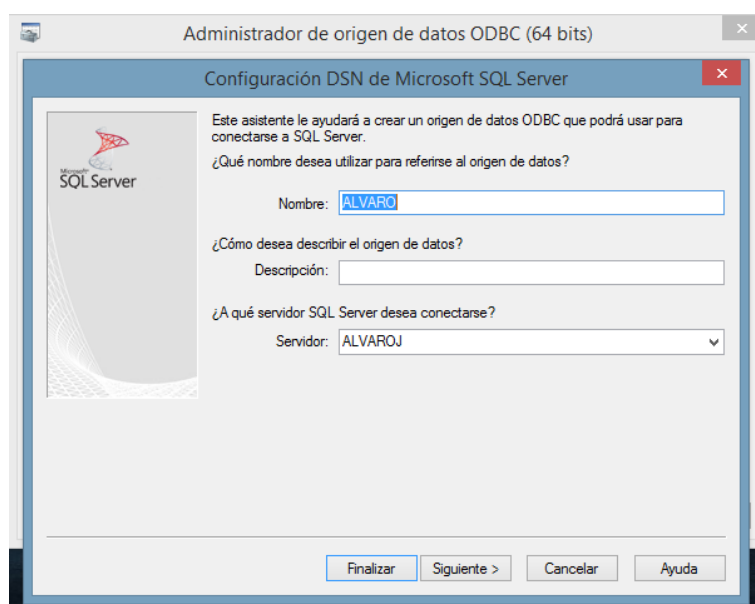


Figura 29. Configuración DSN de Microsoft SQL Service (Fuente propia, 2018).

Con el origen de datos plenamente configurado se procede a realizar la GUI utilizando el entorno GUIDE de Matlab. Se requiere un campo para determinar la característica que se usará como parámetro de búsqueda y su respectivo valor en la extracción de los datos. Además de una herramienta gráfica que muestre los resultados de la búsqueda realizada y que dichos resultados expongan la información de manera clara y entendible para los usuarios que interactúen con la interfaz.

Para la etapa de definición de parámetros de búsqueda se propone la interfaz ilustrada en la Figura 31, en la que se ofrecen las opciones del número con el que se encuentra registrado el Encuentro del paciente en el hospital y el número de identificación del paciente en la historia clínica de la institución de salud como características para definir los parámetros de la búsqueda, y un espacio en el que se digita el valor de la característica para la búsqueda (Código para la ejecución de la búsqueda en la vista `opensource_view` por medio de la GUI en el Anexo 2):

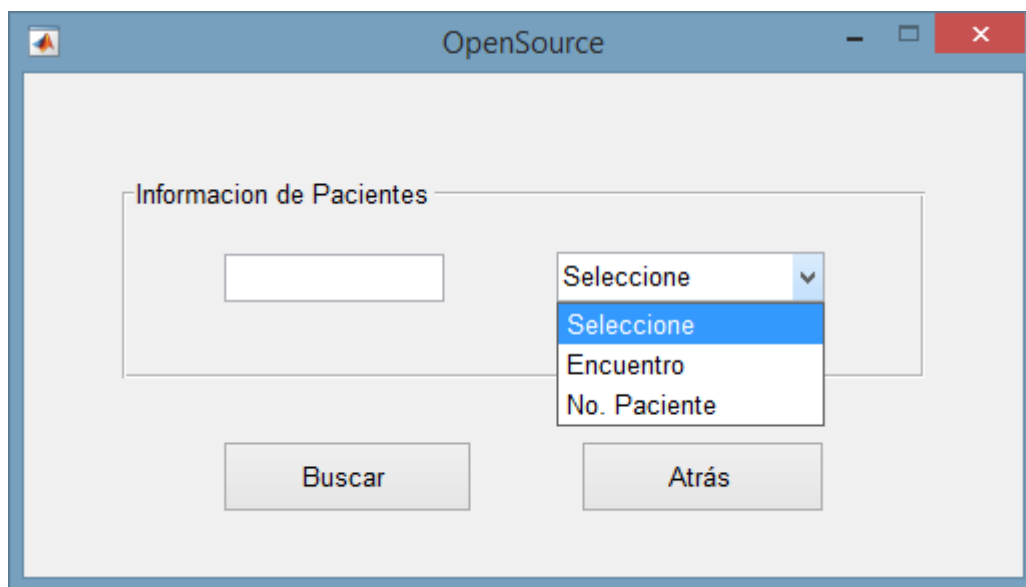


Figura 30. Etapa de definición de parámetros para la búsqueda (Fuente propia, 2018).

Para la etapa de presentación de los resultados de la búsqueda se propone una tabla en la que se muestre la información organizada por filas, debido a la gran cantidad de atributos de la vista

en donde son buscados los datos, además se da a opción de volver a la etapa de definición de parámetros para la búsqueda y otra función para guardar la información de la búsqueda en hojas de cálculo de Excel. En el Anexo 3 del trabajo se muestra el código para el funcionamiento de la tabla.

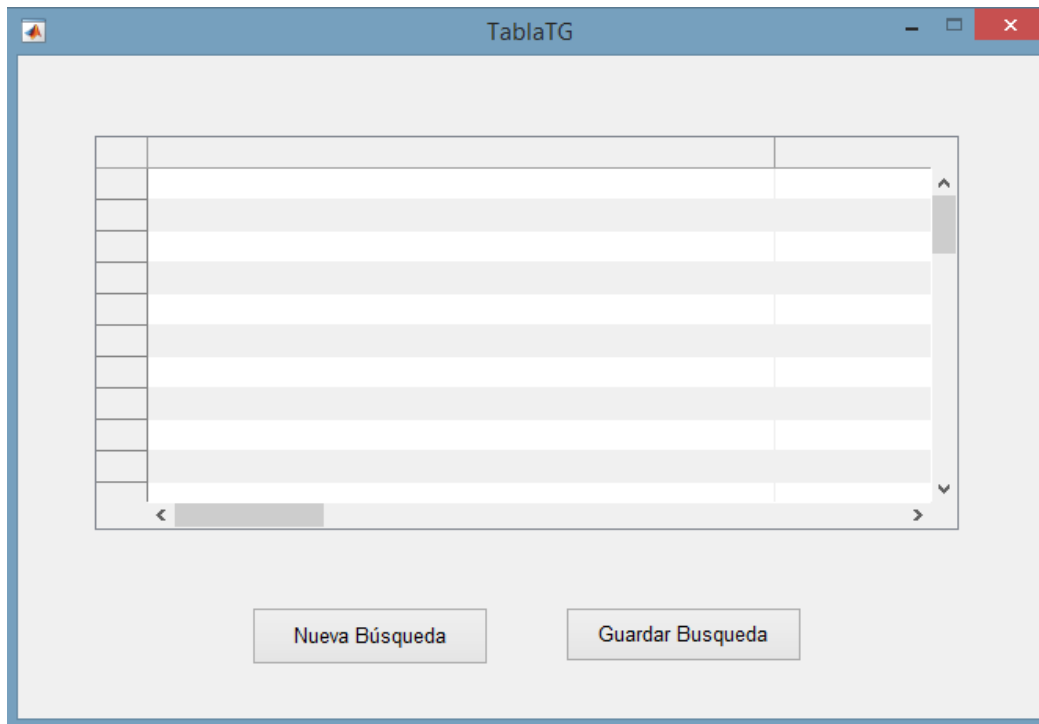


Figura 31. Etapa de presentación de los resultados de la búsqueda (Fuente propia, 2018).

En caso tal que se ingresen valores erróneos o datos que no sean encontrados en la base de datos, se desplegará un aviso que alerte al usuario de dicha situación. En el Anexo 4 se puede encontrar el código correspondiente a esta alerta.

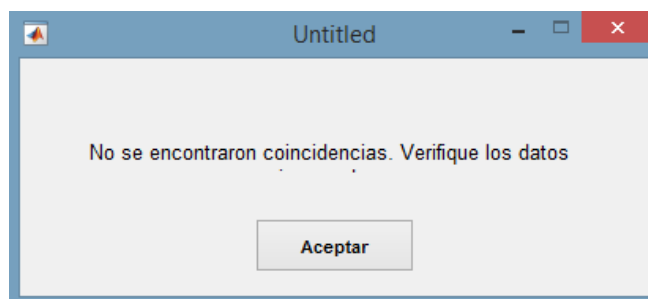


Figura 32. Aviso sin coincidencias (Fuente propia, 2018).

Dado el caso en el que algún campo este vacío y no se le haya ingresado ningún valor, saltará a la vista una alerta de que la información proporcionada a la interfaz no es suficiente para llevar a cabo la búsqueda. La información con el código para esta ventana se encuentra en el Anexo 5.

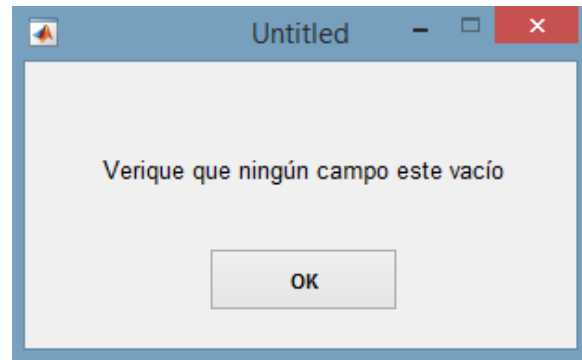


Figura 33. Aviso vacío (Fuente propia, 2018).

3.5. ETL Y MINERÍA DE DATOS

Con la información que ofrece esta base de datos se quiere determinar la probabilidad de que un paciente sea readmitido nuevamente en las instituciones de salud, analizando los resultados obtenidos en el examen para medidas de HbA1c y la edad del paciente.

El método de minería de datos propicio para determinar probabilidades de ocurrencia al relacionar dos o más variables de estudio es el modelo de regresión logística.

Para hacer este modelo se usará la herramienta de Analysis Services de Visual Studio, mediante la extensión de SQL Server Data Tools, debido a la facilidad de integración con la herramienta de MS SQL Server y la comunicación que se logra establecer entre estos dos entornos para el análisis de datos (uno que almacena los datos y el otro que permite su análisis).

Antes de hacer el modelo de regresión logística, en el artículo científico investigado (Strack, y otros, 2014) se describe un proceso de ETL en el que se trata de mejorar la calidad de los datos para el análisis de minería de datos que se realizará posteriormente, debido a que la base de datos original contiene información incompleta, redundante y ruidosa como se espera en cualquier información del mundo real. Hubo varias características que no se pudieron tratar directamente ya que tenían un alto porcentaje de valores faltantes. Estas características fueron Weight (97% de valores faltantes), Payer code (40%) y Medical speciality (47%). El atributo de Weight se consideró demasiado escaso y no se incluyó en un análisis posterior. El Payer code se eliminó porque tenía un alto porcentaje de valores faltantes y no se consideró relevante para el resultado. Se mantuvo el atributo de Medical speciality, agregando el valor "Missing" para tener en cuenta los valores faltantes.

El conjunto de datos preliminares contenía múltiples visitas de pacientes hospitalizados para algunos pacientes y las observaciones no se podían considerar estadísticamente independientes, un supuesto del modelo de regresión logística. Por lo tanto, se utilizó un solo encuentro por paciente; se consideró solo el primer encuentro para cada paciente como ingreso primario y determinamos si fueron readmitidos o no dentro de los 30 días. Además, se eliminaron todos los encuentros que resultaron en el alta hospitalaria o la muerte de un paciente, para evitar sesgar el análisis. Después de realizar las operaciones descritas anteriormente, quedaron 69.984 encuentros que constituyeron el conjunto de datos final para el análisis.

Las variables elegidas para controlar la gravedad demográfica y de la enfermedad del paciente fueron el sexo, la edad, la raza, la fuente de ingreso, la disposición para el alta, el diagnóstico primario, la especialidad médica del médico de admisión y el tiempo pasado en el hospital.

Todo este proceso se hace con los datos de la vista `opensource_view` con la creación de una nueva vista llamada ETL usando el SQL query que se indica en el Anexo 6.

Encuentro	Paciente	Etnia	Genero	Edad	DescripcionAta	OrigenAdmision	EspecialidadMedica	TiempoHospital	Diagnost	
1	222424002	45505143	Caucasian	Female	[80-90]	Discharged/transferred to SNF	Physician Referral	Missing	2	Digestiva
2	222432942	79956882	Caucasian	Female	[40-50]	Discharged to home	Emergency Room	Missing	1	Respirato
3	222437862	7814583	AfricanAmerican	Female	[30-40]	Discharged/transferred to home with home health ...	Physician Referral	Cardiology	13	Circulatori
4	222444726	74310516	Caucasian	Male	[70-80]	Discharged to home	Emergency Room	Missing	2	Otra
5	222452736	100877274	Caucasian	Male	[60-70]	Discharged to home	Physician Referral	Cardiology	1	Circulatori
6	222454092	90975843	Caucasian	Male	[80-90]	Discharged to home	Emergency Room	Missing	4	Genitourin
7	222475320	56244636	AfricanAmerican	Male	[70-80]	Discharged/transferred to SNF	Emergency Room	Missing	7	Circulatori
8	222488976	114113088	Caucasian	Male	[70-80]	Discharged to home	Emergency Room	Missing	3	Respirato
9	222505818	39144843	AfricanAmerican	Female	[60-70]	Discharged to home	Physician Referral	Missing	6	Genitourin
10	222512370	42026634	Caucasian	Male	[80-90]	Discharged/transferred to SNF	Emergency Room	Missing	11	Circulatori
11	222512406	72317520	Caucasian	Female	[10-20]	Discharged to home	Emergency Room	Missing	2	Diabetes
12	222515058	62856504	Caucasian	Female	[60-70]	Discharged to home	Emergency Room	Missing	2	Neoplasr
13	222525672	75829032	Hispanic	Female	[70-80]	Discharged to home	Emergency Room	Missing	2	Neoplasr
14	222530370	38086830	?	Male	[80-90]	Discharged/transferred to home with home health ...	Emergency Room	Missing	7	Respirato
15	222530658	88769880	Caucasian	Female	[60-70]	Discharged/transferred to SNF	Physician Referral	Pulmonology	14	Respirato
16	222537414	84355731	Caucasian	Male	[50-60]	Discharged/transferred to home with home health ...	Emergency Room	Missing	4	Circulatori
17	222539562	57906828	AfricanAmerican	Female	[30-40]	Discharged/transferred to home with home health ...	Emergency Room	Missing	2	Digestiva
18	222542574	94915044	Caucasian	Female	[90-100]	Discharged/transferred to SNF	Emergency Room	Missing	3	Neoplasr
19	222550596	67334355	?	Female	[80-90]	Discharged/transferred to home with home health ...	Emergency Room	Missing	2	Respirato
20	128456814	89015310	Caucasian	Female	[80-90]	Discharged/transferred to ICF	Emergency Room	Missing	4	Circulatori
21	128457168	36122508	Hispanic	Female	[60-70]	Discharged to home	Emergency Room	Emergency/Trauma	12	Digestiva
22	128458638	87121692	Caucasian	Male	[80-90]	Discharged/transferred to SNF	Emergency Room	InternalMedicine	4	Circulatori
23	128460708	79143579	Caucasian	Female	[80-90]	Discharged/transferred to home with home health ...	Emergency Room	Missing	3	Respirato
24	128460870	41456448	Caucasian	Female	[60-70]	Discharged to home	Emergency Room	Missing	4	Digestiva

Figura 34. Datos vista ETL (Fuente propia, 2018)

Finalizado el ETL se continúa el análisis de regresión logística. Para llevarlo a cabo es necesario en primera instancia crear un nuevo proyecto de Analysis Services Multidimensional y Minería de datos, eligiendo el nombre para el proyecto y la ubicación de la carpeta donde se guardará el proyecto.

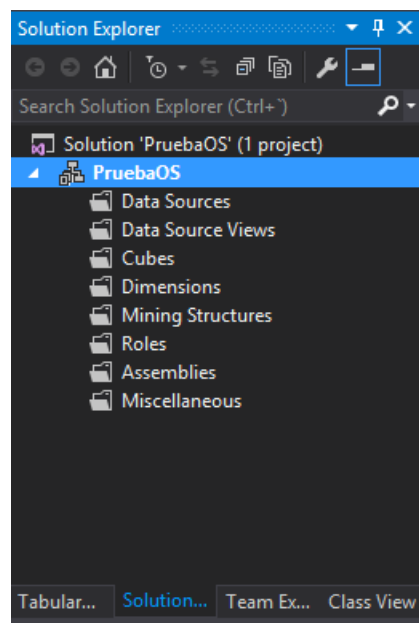


Figura 35. Nuevo proyecto de Analysis Services (Fuente propia, 2018).

Primero se especifica el origen de datos, en donde se escoge la base de datos de la que se extraerá toda la información para procesar los modelos correspondientes.

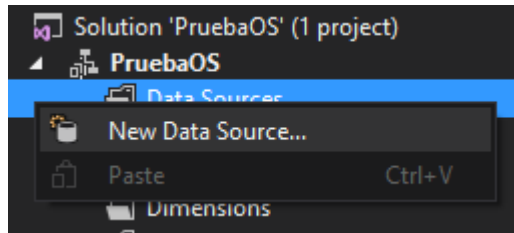


Figura 36. Nuevo origen de datos (Fuente propia, 2018).

Al presionar New Data Source se despliega la ventana del Asistente de Origen de Datos. En ella se escoge la base de datos, se le da el nombre al origen de datos y se establece un usuario y contraseña para controlar el ingreso a la información que resulta luego de la minería de datos.

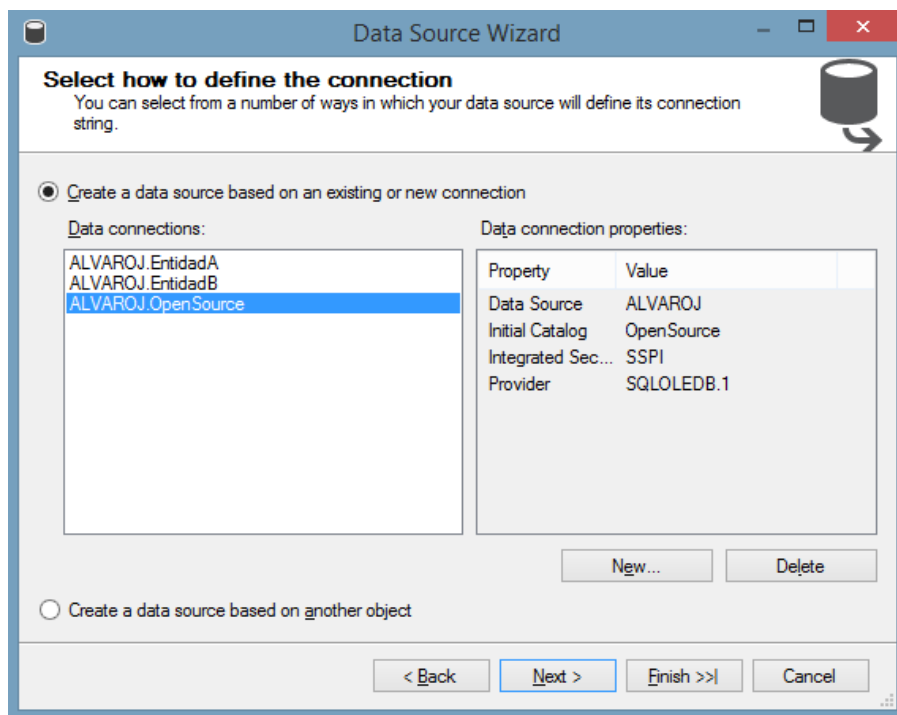


Figura 37. Elección de base de datos para el origen de datos (Fuente propia, 2018).

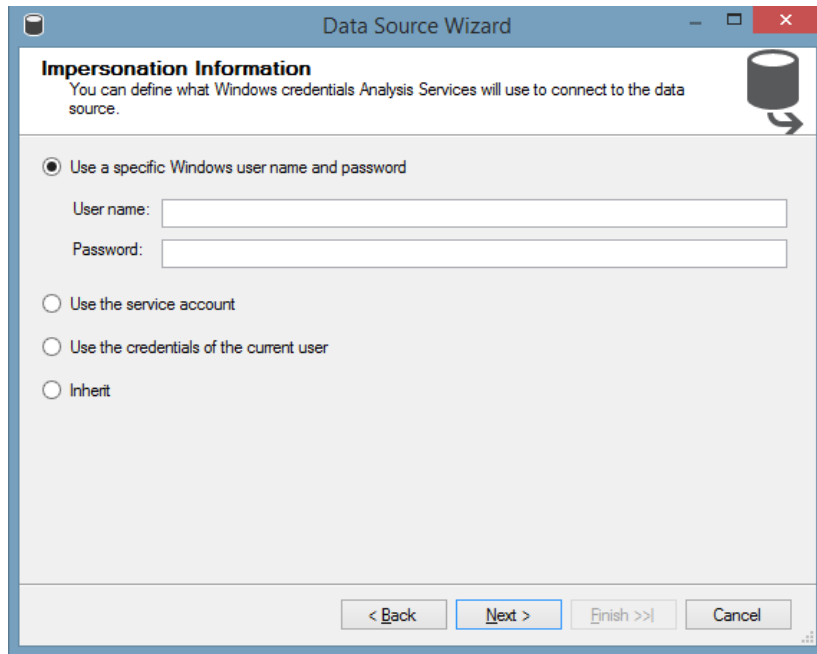


Figura 38. Opciones de seguridad (Fuente propia, 2018).

Lo siguiente es elegir la(s) tabla(s) de la base de datos que se usarán para el modelo, mediante la elección de las vistas del origen de datos. Para ello se ingresa al Asistente de Vistas de Origen de Datos, dando clic derecho sobre Data Source Views y presionando la opción New Data Source View.

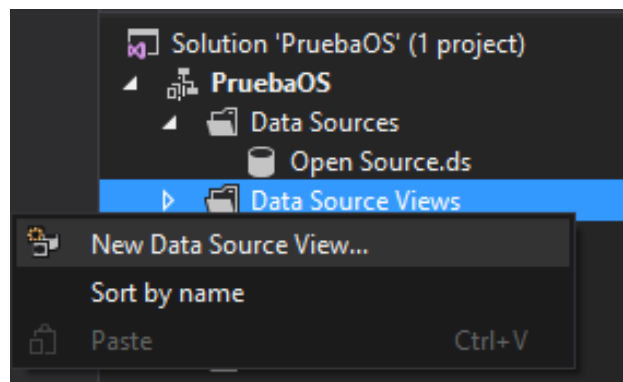


Figura 39. Ingreso al Asistente de Vista de Origen de Datos (Fuente propia, 2018).

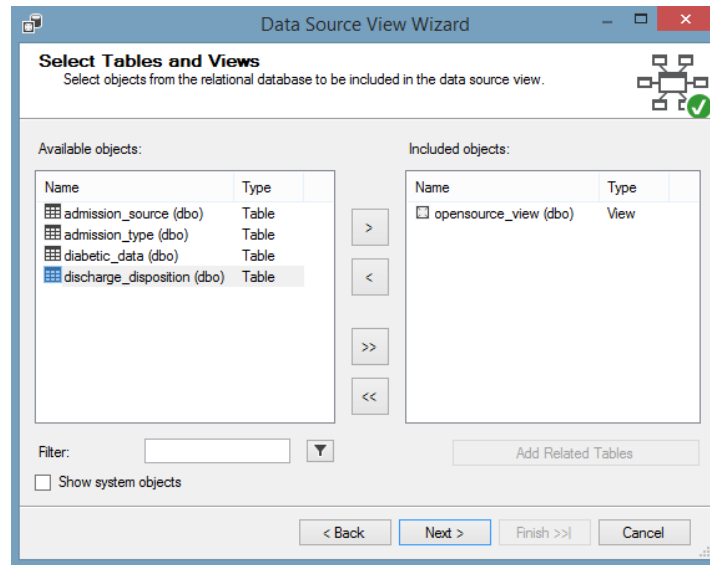


Figura 40. Elección de tablas para el modelo (Fuente propia, 2018).

Después de elegidas las tablas para el modelo, se procede a crear una estructura de minería de datos, con el Asistente de Minería de Datos. Allí se escoge cual método de minería de datos se utilizará y se determinan las variables de entrada y la variable a predecir, junto con la variable que tiene la característica de la clave primaria.

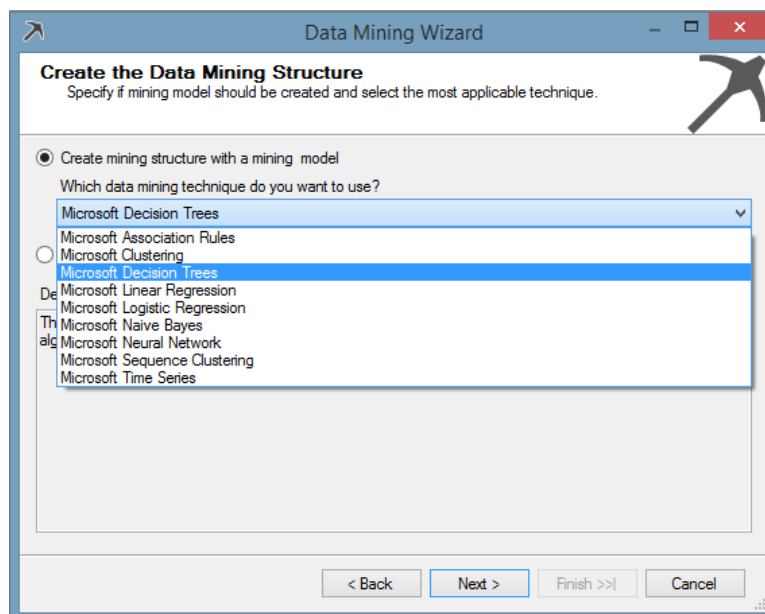


Figura 41. Elección de método de minería de datos (Fuente propia, 2018).

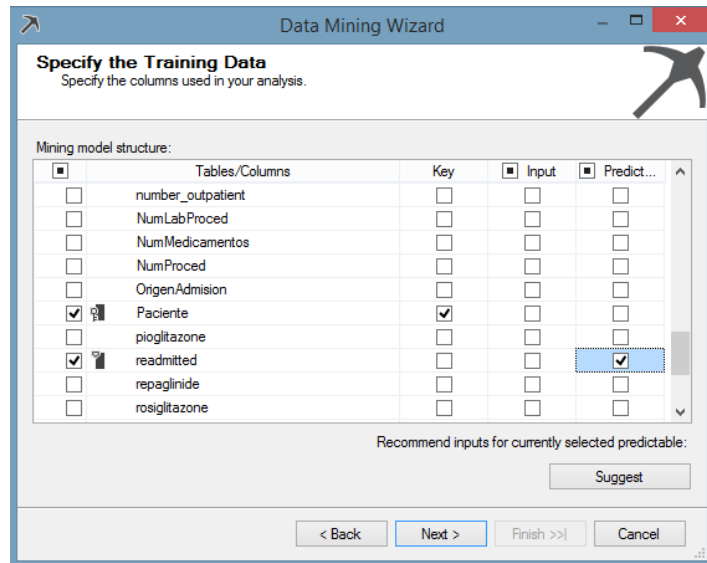


Figura 42. Elección de variables de entrada, de salida y clave primaria (Fuente propia, 2018).

Se escoge el nombre del modelo de minería de datos y se le da la opción Procesar a la estructura.

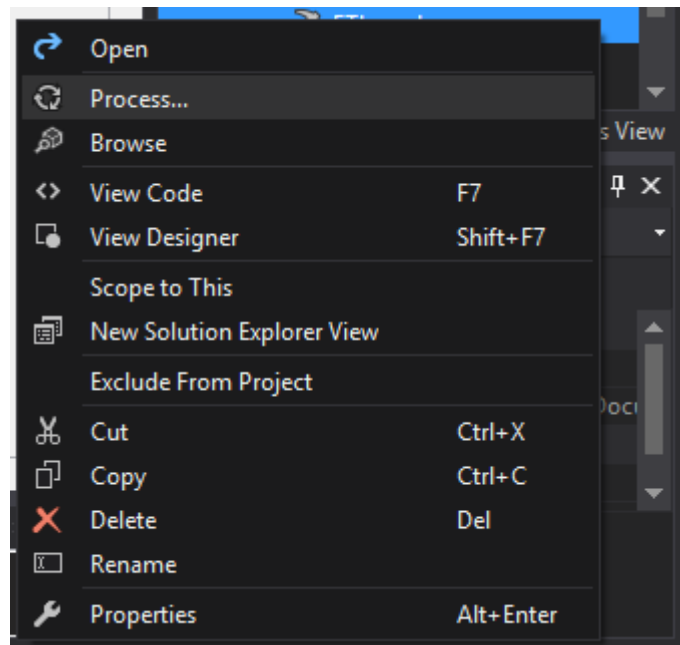


Figura 43. Inicio procesamiento del modelo (Fuente propia, 2018).

3.6. IMPLEMENTACION DEL METODO EN UNA INSTITUCION DE SALUD

3.6.1. Fuentes de datos clínicos y técnicos

Luego de meses de conversaciones con algunas entidades de salud del Valle de Aburrá se llegó a un acuerdo con dos de ellas, en el que ofrecían parte de la información de sus bases de datos de pacientes y procesos llevados a cabo al interior de su organismo.

La primera entidad, a la que por motivos de confidencialidad se mencionará en este trabajo como la ENTIDAD A, dio acceso a los datos que están almacenados en su base de datos de pacientes diabéticos. Estos datos fueron proporcionados bajo el formato de hojas de cálculo de Excel, separando los datos por las sedes de la entidad (una hoja de Excel por cada sede). En ello se contenía información de los pacientes que ingresaron a sus instalaciones durante el periodo que va desde diciembre de 2015 hasta agosto de 2018. En la primera columna de la Tabla 1 se nombran los atributos contenidos en los datos ofrecidos por la ENTIDAD A.

La segunda entidad, a la que se llamará ENTIDAD B, entregó datos en los que había información de los egresos de pacientes ocurridos desde enero de 2016 hasta diciembre de 2017, además de los resultados de los exámenes de laboratorio que le fueron practicados a los pacientes que estuvieron hospitalizados en este mismo periodo. Los egresos fueron entregados en hojas de cálculo de Excel, separados en hojas de Excel por periodos anuales, incluyendo una tabla con los códigos de los municipios de procedencia de los pacientes relacionados con los nombres de los municipios que representan, lo mismo para los diagnósticos. La información de los exámenes de laboratorio para los pacientes hospitalizados de la ENTIDAD B se entregaron en formato de bloc de notas, debido a que el volumen de estos era mucho mayor, igualmente separados por periodos anuales (un bloc de notas para la información de cada año); un bloc de notas con los códigos de los exámenes relacionados con los nombres de los exámenes que representan también fue

incluido. En la segunda y tercera columna de la Tabla 1 se muestran las características de los datos de la ENTIDAD B.

Tabla 1. *Atributos de los datos ofrecidos por entidades de salud (Entidad A & Entidad B, 2018).*

Entidad A	Egresos entidad B	Exámenes pacientes hospitalizados Entidad B
Sede	Historia	Historia
Diagnostico	Ingreso	Orden de trabajo
Descripción dx cie 10	Fecha de nacimiento	Ingreso
Consecutivo historia clínica	Edad	Código examen
Nro. de ingreso	Sexo	Descripción
Identificación	Municipio de residencia	Resultado
Nombre	Código diagnóstico	Valor mínimo posible
Edad	Principal / Secundario	Valor máximo posible
Sexo	Fecha de egreso	
Municipio		
Departamento		
Estado del paciente		
Servicio de ingreso		
Aseguradora		
Estado civil		
Ocupación		
Fecha de ingreso		

Los datos proporcionados por ambas entidades serán importados al servidor de MS SQL Server, ya que allí se ofrecen mayores facilidades para su tratamiento y organización para el posterior análisis de minería de datos y así encontrar patrones que permitan tomar decisiones para optimizar los procesos llevados a cabo dentro de ellas.

3.6.2. Almacenamiento y organización de los datos

El proceso para la importación de los datos desde el servidor de MS SQL Server es prácticamente el mismo que fue practicado para los datos que se obtuvieron del UCI Machine Learning Repository, tanto para los datos contenidos en los blocs de notas, como para lo que están organizados en las hojas de cálculo de Excel.

Aunque antes de proceder con el proceso de importación, se quiso hacer un tratamiento de los datos algo sencillo desde el entorno de Excel a los datos para los que fue posible hacerlo.

Este sencillo tratamiento es básicamente cambiar eliminar las columnas en las que se encuentran los atributos de Nombre e Identificación de los pacientes para los datos de la ENTIDAD A y clasificar las edades en rangos de edades (encasillándolas en intervalos de cada 10 años) para los datos de ambas fuentes. La razones que llevaron a la eliminación de las columnas Nombre e Identificación son más que todo de índole ético, para la protección de la identidad de los pacientes que acuden a las instituciones de salud, además de que la información ofrecida por esos atributos es irrelevante para los análisis de minería de datos, su función podría de la de una clave primaria en la base de datos para identificar de forma única a cada fila de una tabla, pero esta función ya se puede suplir con el número de Historia clínica. El motivo de clasificación de las edades es que, al haber tantos valores distintos para la variable Edad, se podría generar una especie de ruido a la hora de hacer los análisis de minería de datos, problema que se resuelve al organizar todos los valores de edad en intervalos, disminuyendo el número de valores posibles que puede tomar la variable Edad, procedimiento que igualmente fue realizado para la organización de los datos conseguidos del UCI Machine Learning Repository. La fórmula usada la clasificación de las edades se exhibe en el Anexo 7.

Teniendo organizados los datos en las hojas de Excel se tiene vía libre para continuar con la importación de los datos desde el servidor de MS SQL Server.

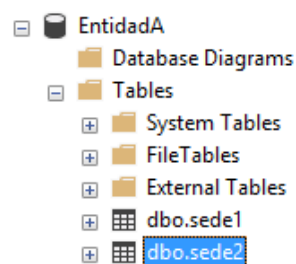


Figura 44. Datos importados ENTIDAD A (Fuente propia, 2018).

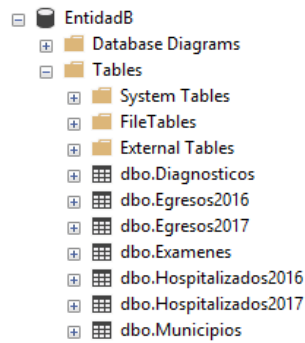


Figura 45. Datos importados ENTIDAD B (Fuente propia, 2018).

Como etapa final organizan los datos en una sola tabla por medio de las vistas al igual que con los datos de la fuente encontrada en internet, para cada base de datos se hace una vista distinta. Los códigos correspondientes a los queries realizados para crear las vistas entidadA_view (2224 datos) y entidadB_view (317749 datos) para las bases de datos de la ENTIDAD A y la ENTIDAD B son mostrados en los Anexos 8 y 9, respectivamente.

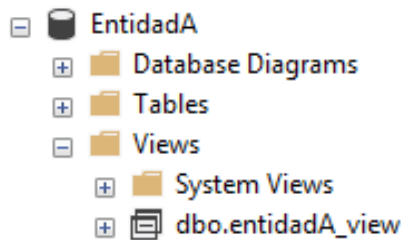


Figura 46. Vista entidadA_view creada (Fuente propia, 2018).

	CONSECUTIVO HISTORIA CLINICA	NRO DE INGRESO	RANGO DE EDAD	SEXO	MUNICIPIO	DESCRIPCION DX CIE 10	ESTADO
1	1545	2	0-10	M	ENVIGADO (ANTIOQUIA)	OTRAS DIABETES MELLITUS ESPECIFICADAS CON CETOACI...	ALTA
2	3301	1	60-70	M	MEDELLIN (ANTIOQUIA)	DIABETES MELLITUS NO ESPECIFICADA SIN MENCION DE C...	REMISIO
3	10697	32	80-90	F	MEDELLIN (ANTIOQUIA)	OTRAS DIABETES MELLITUS ESPECIFICADAS SIN MENCION ...	ALTA
4	26499	13	70-80	F	MEDELLIN (ANTIOQUIA)	DIABETES MELLITUS INSULINODEPENDIENTE SIN MENCION...	ALTA
5	26901	2	80-90	M	MEDELLIN (ANTIOQUIA)	DIABETES MELLITUS NO ESPECIFICADA SIN MENCION DE C...	ALTA
6	26901	3	80-90	M	MEDELLIN (ANTIOQUIA)	DIABETES MELLITUS NO ESPECIFICADA SIN MENCION DE C...	ALTA
7	26987	17	80-90	F	MEDELLIN (ANTIOQUIA)	DIABETES MELLITUS NO INSULINODEPENDIENTE SIN MENC...	ALTA
8	26987	18	80-90	F	MEDELLIN (ANTIOQUIA)	DIABETES MELLITUS NO INSULINODEPENDIENTE SIN MENC...	ALTA
9	26987	24	80-90	F	MEDELLIN (ANTIOQUIA)	DIABETES MELLITUS NO INSULINODEPENDIENTE SIN MENC...	ALTA
10	34064	30	70-80	F	MEDELLIN (ANTIOQUIA)	DIABETES MELLITUS INSULINODEPENDIENTE CON COMPLIC...	ALTA
11	34064	31	70-80	F	MEDELLIN (ANTIOQUIA)	DIABETES MELLITUS INSULINODEPENDIENTE CON COMPLIC...	ALTA
12	34064	32	70-80	F	MEDELLIN (ANTIOQUIA)	DIABETES MELLITUS INSULINODEPENDIENTE CON COMPLIC...	ALTA
13	34064	33	70-80	F	MEDELLIN (ANTIOQUIA)	DIABETES MELLITUS INSULINODEPENDIENTE CON COMPLIC...	ALTA
14	34064	35	70-80	F	MEDELLIN (ANTIOQUIA)	DIABETES MELLITUS INSULINODEPENDIENTE CON COMPLIC...	ALTA
15	34299	10	70-80	F	MEDELLIN (ANTIOQUIA)	DIABETES MELLITUS INSULINODEPENDIENTE SIN MENCION...	ALTA
16	34299	11	70-80	F	MEDELLIN (ANTIOQUIA)	DIABETES MELLITUS INSULINODEPENDIENTE SIN MENCION...	ALTA
17	34299	12	70-80	F	MEDELLIN (ANTIOQUIA)	DIABETES MELLITUS INSULINODEPENDIENTE SIN MENCION...	ALTA
18	34299	13	70-80	F	MEDELLIN (ANTIOQUIA)	DIABETES MELLITUS INSULINODEPENDIENTE SIN MENCION...	ALTA
19	34299	14	70-80	F	MEDELLIN (ANTIOQUIA)	DIABETES MELLITUS INSULINODEPENDIENTE SIN MENCION...	ALTA
20	34299	15	70-80	F	MEDELLIN (ANTIOQUIA)	DIABETES MELLITUS INSULINODEPENDIENTE SIN MENCION...	ALTA
21	34299	19	70-80	F	MEDELLIN (ANTIOQUIA)	DIABETES MELLITUS INSULINODEPENDIENTE SIN MENCION...	ALTA
22	46275	11	70-80	F	MEDELLIN (ANTIOQUIA)	DIABETES MELLITUS INSULINODEPENDIENTE SIN MENCION...	ALTA
23	52157	16	70-80	F	MEDELLIN (ANTIOQUIA)	DIABETES MELLITUS NO INSULINODEPENDIENTE SIN MENC...	ALTA

Query executed successfully. ALVAROJ (13.0 RTM) ALVAROJ\Alvaro (53) EntidadA 00:00:31 2224 rows

Figura 47. Datos entidadA_view (Fuente propia, 2018).

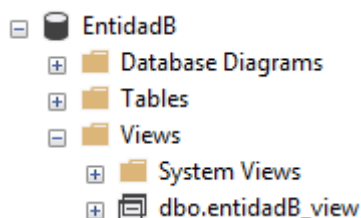


Figura 48. Vista entidadB_view creada (Fuente propia, 2018).

	HISTORIA	ORDEN DE TRABAJO	INGRESO	RANGO DE EDAD	SEXO	MUNICIPIO	DIAGNOSTICO	Principal / Secundario	EXAMÉ
1	100024	2504137	2	50-60	F	LA ESTRELLA	DIABETES MELLITUS NO INSULINODEPENDIENTE SIN MEN...	S	HEMO
2	100024	2504137	2	50-60	F	LA ESTRELLA	DIABETES MELLITUS NO INSULINODEPENDIENTE SIN MEN...	S	HEMO
3	100024	2504137	2	50-60	F	LA ESTRELLA	DIABETES MELLITUS NO INSULINODEPENDIENTE SIN MEN...	S	HEMO
4	100024	2504137	2	50-60	F	LA ESTRELLA	DIABETES MELLITUS NO INSULINODEPENDIENTE SIN MEN...	S	HEMO
5	100024	2504137	2	50-60	F	LA ESTRELLA	DIABETES MELLITUS NO INSULINODEPENDIENTE SIN MEN...	S	HEMO
6	100024	2504137	2	50-60	F	LA ESTRELLA	DIABETES MELLITUS NO INSULINODEPENDIENTE SIN MEN...	S	HEMO
7	100024	2504137	2	50-60	F	LA ESTRELLA	DIABETES MELLITUS NO INSULINODEPENDIENTE SIN MEN...	S	SODIC
8	100024	2504137	2	50-60	F	LA ESTRELLA	DIABETES MELLITUS NO INSULINODEPENDIENTE SIN MEN...	S	TIEMF
9	100024	2504137	2	50-60	F	LA ESTRELLA	DIABETES MELLITUS NO INSULINODEPENDIENTE SIN MEN...	S	TSH T
10	100115	2474338	41	70-80	F	MEDELLIN	DIABETES MELLITUS NO INSULINODEPENDIENTE SIN MEN...	P	CITOG
11	100115	2474338	41	70-80	F	MEDELLIN	DIABETES MELLITUS NO INSULINODEPENDIENTE SIN MEN...	P	CITOG
12	100115	2474338	41	70-80	F	MEDELLIN	DIABETES MELLITUS NO INSULINODEPENDIENTE SIN MEN...	P	CITOG
13	100115	2474338	41	70-80	F	MEDELLIN	DIABETES MELLITUS NO INSULINODEPENDIENTE SIN MEN...	P	CITOG
14	100115	2474338	41	70-80	F	MEDELLIN	DIABETES MELLITUS NO INSULINODEPENDIENTE SIN MEN...	P	CITOG
15	100115	2474338	41	70-80	F	MEDELLIN	DIABETES MELLITUS NO INSULINODEPENDIENTE SIN MEN...	P	CREA'
16	100115	2474338	41	70-80	F	MEDELLIN	DIABETES MELLITUS NO INSULINODEPENDIENTE SIN MEN...	P	HEMO
17	100115	2474338	41	70-80	F	MEDELLIN	DIABETES MELLITUS NO INSULINODEPENDIENTE SIN MEN...	P	HEMO
18	100115	2474338	41	70-80	F	MEDELLIN	DIABETES MELLITUS NO INSULINODEPENDIENTE SIN MEN...	P	HEMO
19	100115	2474338	41	70-80	F	MEDELLIN	DIABETES MELLITUS NO INSULINODEPENDIENTE SIN MEN...	P	HEMO
20	100115	2474338	41	70-80	F	MEDELLIN	DIABETES MELLITUS NO INSULINODEPENDIENTE SIN MEN...	P	HEMO
21	100115	2474338	41	70-80	F	MEDELLIN	DIABETES MELLITUS NO INSULINODEPENDIENTE SIN MEN...	P	HEMO
22	100115	2474338	41	70-80	F	MEDELLIN	DIABETES MELLITUS NO INSULINODEPENDIENTE SIN MEN...	P	HEMO
23	100115	2474338	41	70-80	F	MEDELLIN	DIABETES MELLITUS NO INSULINODEPENDIENTE SIN MEN...	P	PROT
24	100115	2474338	41	70-80	F	MEDELLIN	DIABETES MELLITUS NO INSULINODEPENDIENTE SIN MEN...	S	CITOG

Query executed successfully. ALVAROJ (13.0 RTM) ALVAROJ\Alvaro (54) EntidadB 00:00:54 317749 rows

Figura 49. Datos entidadB_view (Fuente propia, 2018).

3.6.3. Diseño de la interfaz gráfica de usuario

Al tener los datos organizados en las vistas, se puede continuar con creación de la GUI que realice las búsquedas pertinentes en las vistas entidadA_view y entidadB_view.

Basado en la GUI creada para las búsquedas en la vista opensource_view se crea una figura de Matlab para la definición de los parámetros de búsqueda para cada base de datos (EntidadA y EntidadB fueron os nombres escogidos para las figuras) y para presentar los datos se pretende utilizar la figura TablaTG ya creada previamente en la que se arrojan los resultados de la búsqueda en una tabla, aunque si es necesario realizar cambio en la programación de la figura para que reciba los nuevos datos (cabe mencionar que se le adiciona una función para guardar las búsquedas), cambios que son mostrados en los códigos del Anexo 10. En los Anexos 11 y 12 se encuentran los códigos necesarios para la programación del funcionamiento de las figuras EntidadA y EntidadB, respectivamente.

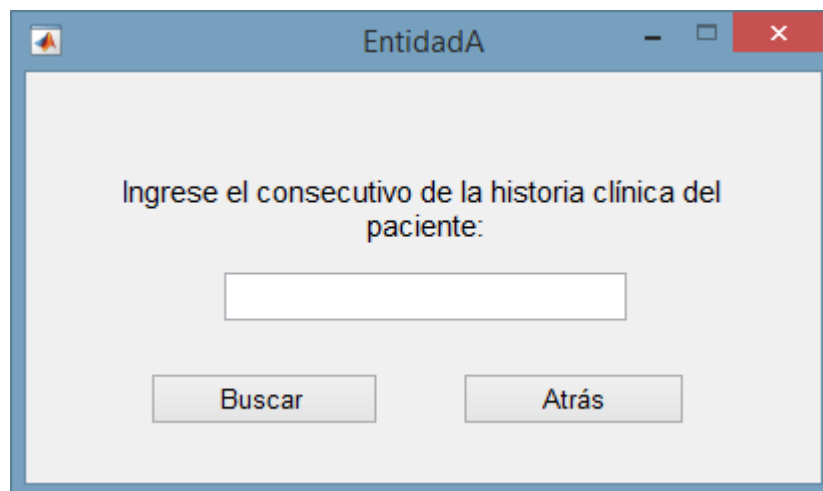


Figura 50. Ventana elección parámetros para búsqueda en Entidad A (Fuente propia, 2018).

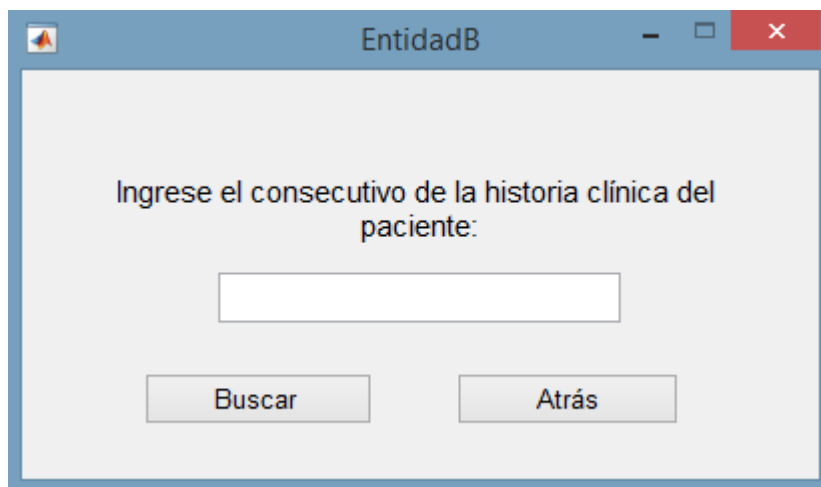


Figura 51. Ventana elección parámetros para búsqueda en Entidad B (Fuente propia, 2018).

Siguiendo la idea de integración de varias fuentes, se construye una ventana en la que se especifique la fuente en la que se desea realizar la búsqueda de datos, teniendo como opciones la ENTIDAD A, la ENTIDAD B y la base de datos OpenSource, unificando de esta forma las búsquedas dentro de una misma aplicación, en la que sin ningún problema pueden ser añadidas otras fuentes de datos, como por ejemplo otras entidades de salud y así logrando la creación de un repositorio donde se pueda encontrar de manera fácil y rápida cierta información de sus pacientes, existiendo la posibilidad de buscar información de pacientes que posean otras características además de los diabéticos o tal vez incluso establecer una aplicación que haga búsquedas de pacientes por áreas de servicio asistencial de la institución de salud. La programación usada para la labor de la ventana de integración de fuentes de la aplicación se describe en el Anexo 13.

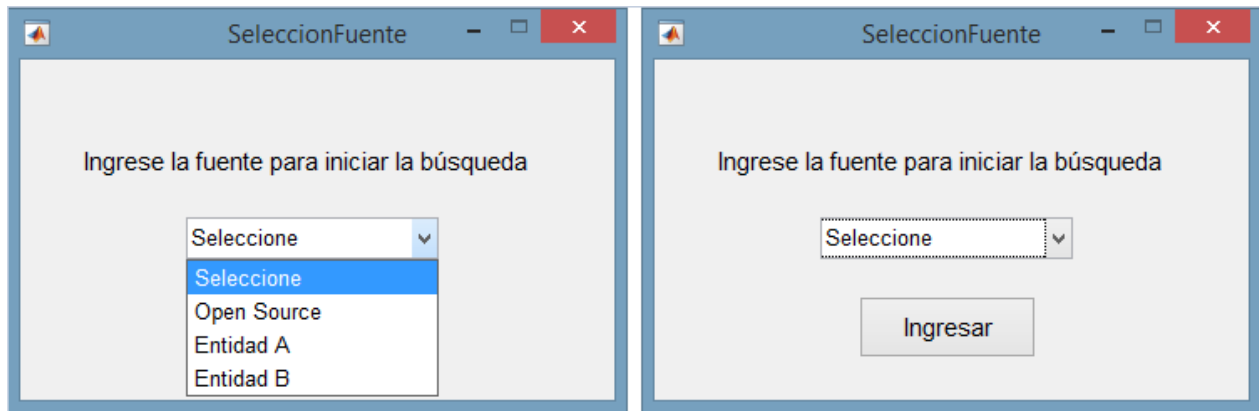


Figura 52. Ventana para la integración de varias fuentes (Fuente propia, 2018).

3.6.4. ETL y Minería de datos

Con el repositorio de datos ya finalizado, el paso a seguir es el análisis de los datos almacenados, utilizando las herramientas de análisis igualmente descritas con anterioridad y siguiendo un proceso muy parecido al realizado con los datos descargados del UCI Machine Learning Repository.

Pero antes de usar los datos para su análisis con la minería de datos, es necesario realizar un análisis descriptivo de ellos para evitar el uso de datos que estén sesgados o datos que no ofrezcan ninguna información relevante para un análisis de minería de datos y que al fin y al cabo solo generan ruido, afectando de manera negativa los resultados arrojados por los modelos utilizados.

Para hacer el análisis descriptivo Microsoft Excel ofrece muchas facilidades a través de las Tablas dinámicas, por ello se hace necesario nuevamente el uso del Asistente de importación y exportación de SQL Server, en este caso para hacer el proceso inverso al que se hizo con los datos en el ítem de Almacenamiento y Organización de los datos. Acá se transfieren los de datos desde el servidor de SQL Server a las hojas de cálculos de Excel usando la opción de exportación. El proceso es el mismo al de importación, solo que se escoge la opción Export Data

en lugar de Import Data, mientras que el origen sería Microsoft OLE DB Provider for SQL Server y el destino Microsoft Excel.

Habiendo exportado los datos de la ENTIDAD A al entorno de Excel, lo primero que se quiere ver es el número de valores posibles que puede tomar la variable diagnóstico, ya que es la principal característica que comparten los pacientes cuya información está contenida en esta base de datos (pacientes que han sido diagnosticados con algún tipo de diabetes). Usando una Tabla dinámica de Excel podemos obtener un análisis de frecuencia de datos de manera rápida.

Para crear una Tabla dinámica de Excel, hay que dirigirse al menú Insertar y allí se escoge la opción Tabla dinámica.

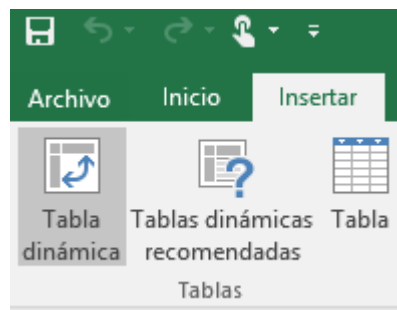


Figura 53. Crear Tabla dinámica de Excel (Fuente propia, 2018).

Luego en la ventana para la creación de la Tabla dinámica se selecciona el rango en que se encuentran los datos que se analizarán, en este se escoge toda la hoja en la que están los datos.

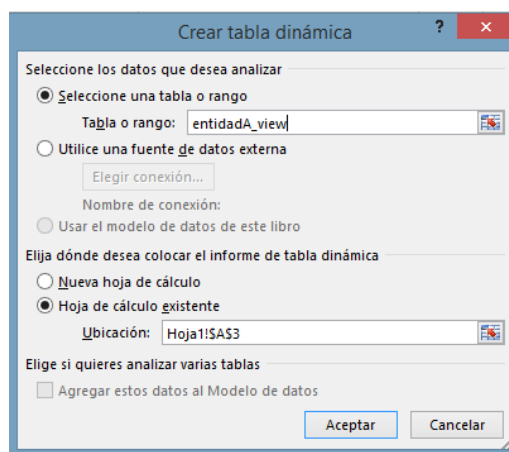


Figura 54. Selección de datos para la Tabla dinámica (Fuente propia, 2018).

A continuación, se escogen los campos de la tabla que se quieren analizar.

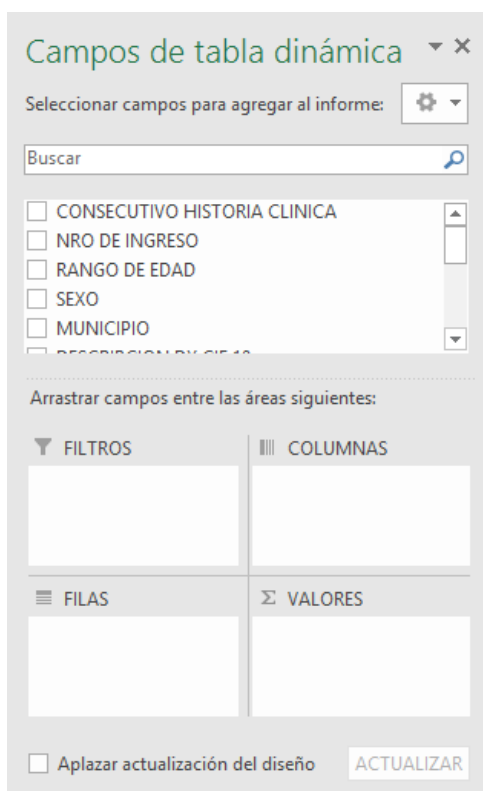


Figura 55. Selección de campos para el análisis en la Tabla dinámica (Fuente propia, 2018).

En este caso escogemos el campo que posee los diagnósticos de diabetes, arrastrando dicho campo al área de FILAS y al área de VALORES, en esta última área se escoge la operación cuenta.

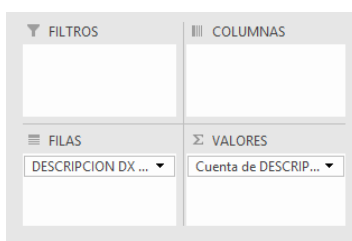


Figura 56. Campos elegidos para análisis de diagnósticos (Fuente propia, 2018).

Para observar los resultados de formas más ilustrativa, existe la opción Gráficas dinámicas en el menú de Herramientas de Tabla dinámica. Se elige un gráfico circular para ver los resultados, como se muestra en la Figura 56.

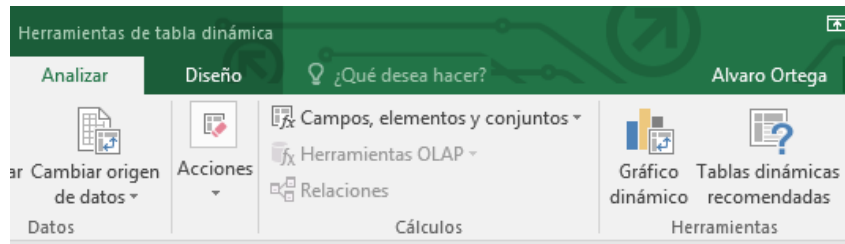


Figura 57. Crear un Gráfico dinámico (Fuente propia, 2018).

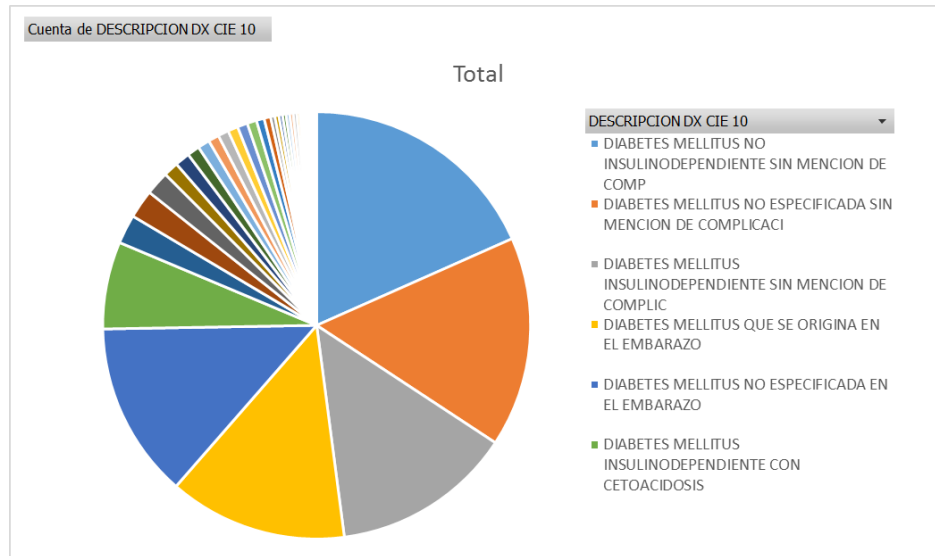


Figura 58. Grafica circular con análisis de Cuenta DESCRIPCION DX CIE 10 (Fuente propia, 2018).

Se observa que esta variable, que muy posiblemente será la variable de predicción puede tomar muchos valores distintos y que además muchos de los posibles valores tienen una frecuencia demasiado baja, por tanto, se requiere eliminar algunos campos que no ofrecen información relevante para los modelos y que más bien deterioran su calidad.

El filtrado de los datos se hace con los valores que tienen mayor frecuencia, aunque se observa que estos no tienen una frecuencia importante como para uso en un modelo de minería de datos (generalmente se necesita una frecuencia aproximadamente mínima de 1000 datos para que un valor de una variable ofrezca información relevante).

Tabla 2. Tabla dinámica datos filtrados Cuenta de DESCRIPCION DX CIE 10 (Fuente propia, 2018).

Cuenta de DESCRIPCION DX CIE 10	
DESCRIPCION DX CIE 10	Total
DIABETES MELLITUS NO INSULINODEPENDIENTE SIN MENCION DE COMP	408
DIABETES MELLITUS NO ESPECIFICADA SIN MENCION DE COMPLICACI	354
DIABETES MELLITUS INSULINODEPENDIENTE SIN MENCION DE COMPLIC	304
DIABETES MELLITUS QUE SE ORIGINA EN EL EMBARAZO	300
DIABETES MELLITUS NO ESPECIFICADA EN EL EMBARAZO	296
DIABETES MELLITUS INSULINODEPENDIENTE CON CETOACIDOSIS	147
Total general	1809

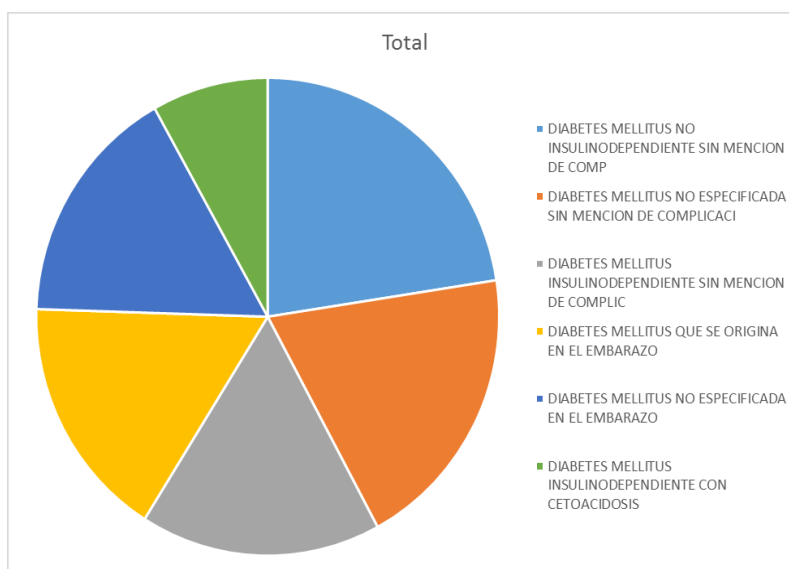


Figura 59. Gráfica circular análisis de Cuenta DESCRIPCION DX CIE 10 con datos filtrados (Fuente propia, 2018)

Teniendo esta información se hace el respectivo filtrado pero en la base de datos en el servidor de SQL Server , creando una nueva vistas en la que solo se incluyan los campos que contienen los diagnósticos escogidos en el filtrado realizado en la Tabla dinámica de Excel.

A la nueva vista creada se le asigna el nombre de filtro_diagnosticos. El código de la búsqueda de SQL realizado para crear la vista se encuentra en el Anexo 14 de este trabajo.

	CONSECUTIVO HISTORIA CLINICA	NRO DE INGRESO	RANGO DE EDAD	SEXO	MUNICIPIO	DESCRIPCION DX CIE 10	ESTADO
1	3301	1	60-70	M	MEDELLIN (ANTIOQUIA)	DIABETES MELLITUS NO ESPECIFICADA SIN MENCION DE C...	REMISION
2	26499	13	70-80	F	MEDELLIN (ANTIOQUIA)	DIABETES MELLITUS INSULINODEPENDIENTE SIN MENCIO...	ALTA
3	26901	2	80-90	M	MEDELLIN (ANTIOQUIA)	DIABETES MELLITUS NO ESPECIFICADA SIN MENCION DE C...	ALTA
4	26901	3	80-90	M	MEDELLIN (ANTIOQUIA)	DIABETES MELLITUS NO ESPECIFICADA SIN MENCION DE C...	ALTA
5	26987	17	80-90	F	MEDELLIN (ANTIOQUIA)	DIABETES MELLITUS NO INSULINODEPENDIENTE SIN MEN...	ALTA
6	26987	18	80-90	F	MEDELLIN (ANTIOQUIA)	DIABETES MELLITUS NO INSULINODEPENDIENTE SIN MEN...	ALTA
7	26987	24	80-90	F	MEDELLIN (ANTIOQUIA)	DIABETES MELLITUS NO INSULINODEPENDIENTE SIN MEN...	ALTA
8	34299	10	70-80	F	MEDELLIN (ANTIOQUIA)	DIABETES MELLITUS INSULINODEPENDIENTE SIN MENCIO...	ALTA
9	34299	11	70-80	F	MEDELLIN (ANTIOQUIA)	DIABETES MELLITUS INSULINODEPENDIENTE SIN MENCIO...	ALTA
10	34299	12	70-80	F	MEDELLIN (ANTIOQUIA)	DIABETES MELLITUS INSULINODEPENDIENTE SIN MENCIO...	ALTA
11	34299	13	70-80	F	MEDELLIN (ANTIOQUIA)	DIABETES MELLITUS INSULINODEPENDIENTE SIN MENCIO...	ALTA
12	34299	14	70-80	F	MEDELLIN (ANTIOQUIA)	DIABETES MELLITUS INSULINODEPENDIENTE SIN MENCIO...	ALTA
13	34299	15	70-80	F	MEDELLIN (ANTIOQUIA)	DIABETES MELLITUS INSULINODEPENDIENTE SIN MENCIO...	ALTA
14	34299	19	70-80	F	MEDELLIN (ANTIOQUIA)	DIABETES MELLITUS INSULINODEPENDIENTE SIN MENCIO...	ALTA
15	46275	11	70-80	F	MEDELLIN (ANTIOQUIA)	DIABETES MELLITUS INSULINODEPENDIENTE SIN MENCIO...	ALTA
16	52157	16	70-80	F	MEDELLIN (ANTIOQUIA)	DIABETES MELLITUS NO INSULINODEPENDIENTE SIN MEN...	ALTA
17	68423	12	60-70	M	MEDELLIN (ANTIOQUIA)	DIABETES MELLITUS INSULINODEPENDIENTE SIN MENCIO...	ALTA
18	80929	21	80-90	F	MEDELLIN (ANTIOQUIA)	DIABETES MELLITUS NO ESPECIFICADA SIN MENCION DE C...	ALTA
19	91717	22	60-70	F	MEDELLIN (ANTIOQUIA)	DIABETES MELLITUS NO INSULINODEPENDIENTE SIN MEN...	ALTA
20	91717	23	60-70	F	MEDELLIN (ANTIOQUIA)	DIABETES MELLITUS NO INSULINODEPENDIENTE SIN MEN...	ALTA
21	142653	10	80-90	F	MEDELLIN (ANTIOQUIA)	DIABETES MELLITUS NO ESPECIFICADA SIN MENCION DE C...	ALTA
22	142653	11	80-90	F	MEDELLIN (ANTIOQUIA)	DIABETES MELLITUS NO ESPECIFICADA SIN MENCION DE C...	ALTA
23	153967	31	90-100	F	MEDELLIN (ANTIOQUIA)	DIABETES MELLITUS NO ESPECIFICADA SIN MENCION DE C...	ALTA
24	159998	5	60-70	F	MEDELLIN (ANTIOQUIA)	DIABETES MELLITUS NO ESPECIFICADA SIN MENCION DE C...	ALTA

Query executed successfully. ALVAROJ (13.0 RTM) | ALVAROJ\Alvaro (56) | EntidadA | 00:00:00 | 1809 rows

Figura 60. Datos filtro_diagnosticos (Fuente propia, 2018).

Para los datos de la ENTIDAD B se pretende hacer el mismo análisis, pero existe un inconveniente, la cantidad de datos es mayor a la soportada por una hoja de cálculo de Excel, así que si se hace el análisis con una Tabla dinámica se daría la información de manera incompleta. La solución a este problema es realizar un query en SQL en el que se haga una cuenta de la cantidad de datos por cada posible valor que pueda tomar la variable que almacena los diagnósticos (en el Anexo 15 se indica el código para realizar esta búsqueda).

El resultado de esta búsqueda es copiado a una hoja de Excel para luego ser graficado y obtener una presentación de resultados de forma más dinámica para la tabla de frecuencias.

	DIAGNOSTICO	FREQ
1	DIABETES MELLITUS NO INSULINODEPENDIENTE SIN MENCIO...	164387
2	DIABETES MELLITUS NO INSULINODEPENDIENTE CON COM...	56406
3	DIABETES MELLITUS, NO ESPECIFICADA SIN MENCION DE C...	21499
4	DIABETES MELLITUS INSULINODEPENDIENTE SIN MENCION ...	15562
5	DIABETES MELLITUS INSULINODEPENDIENTE CON CETOACI...	8447
6	DIABETES MELLITUS INSULINODEPENDIENTE CON COMPLIC...	5851
7	DIABETES MELLITUS INSULINODEPENDIENTE CON OTRAS C...	5243
8	DIABETES MELLITUS NO INSULINODEPENDIENTE CON CETO...	5203
9	DIABETES INSIPIDA	4652
10	DIABETES MELLITUS, NO ESPECIFICADA CON CETOACIDOSIS	3526
11	DIABETES MELLITUS INSULINODEPENDIENTE CON COMPLIC...	3416
12	DIABETES MELLITUS INSULINODEPENDIENTE CON COMPLIC...	2739
13	DIABETES MELLITUS INSULINODEPENDIENTE CON COMPLIC...	2492
14	DIABETES MELLITUS QUE SE ORIGINA EN EL EMBARAZO	1921
15	DIABETES MELLITUS NO INSULINODEPENDIENTE CON OTRA...	1804
16	DIABETES MELLITUS ASOCIADA CON DESNUTRICION CON C...	1767
17	OTRAS DIABETES MELLITUS ESPECIFICADAS CON CETOACID...	1670
18	DIABETES MELLITUS, NO ESPECIFICADA CON COMPLICACIO...	1307
19	DIABETES MELLITUS, NO ESPECIFICADA CON OTRAS COMPL...	1251
20	DIABETES MELLITUS, NO ESPECIFICADA CON COMPLICACIO...	1111
21	TRASTORNOS GLOMERULARES EN DIABETES MELLITUS (E1...	1055
22	DIABETES MELLITUS ASOCIADA CON DESNUTRICION SIN ME...	942
23	OTRAS DIABETES MELLITUS ESPECIFICADAS SIN MENCION ...	930
24	DIABETES MELLITUS NO ESPECIFICADA, EN EL EMBARAZO	679

Figura 61. Tabla de frecuencias con SQL (Fuente propia, 2018).

Se observa que del total de datos (317749 datos) cerca del 70% pertenecen solo a dos valores (DIABETES MELLITUS NO INSULINODEPENDIENTE SIN MENCION DE COMP con 164387 datos y DIABETES MELLITUS NO INSULINODEPENDIENTE CON COMPLICACIONES con 56406 datos), por tanto, se decide hacer un filtrado de la vista entidadB_view en el que solo se tenga información de los dos valores más importantes.

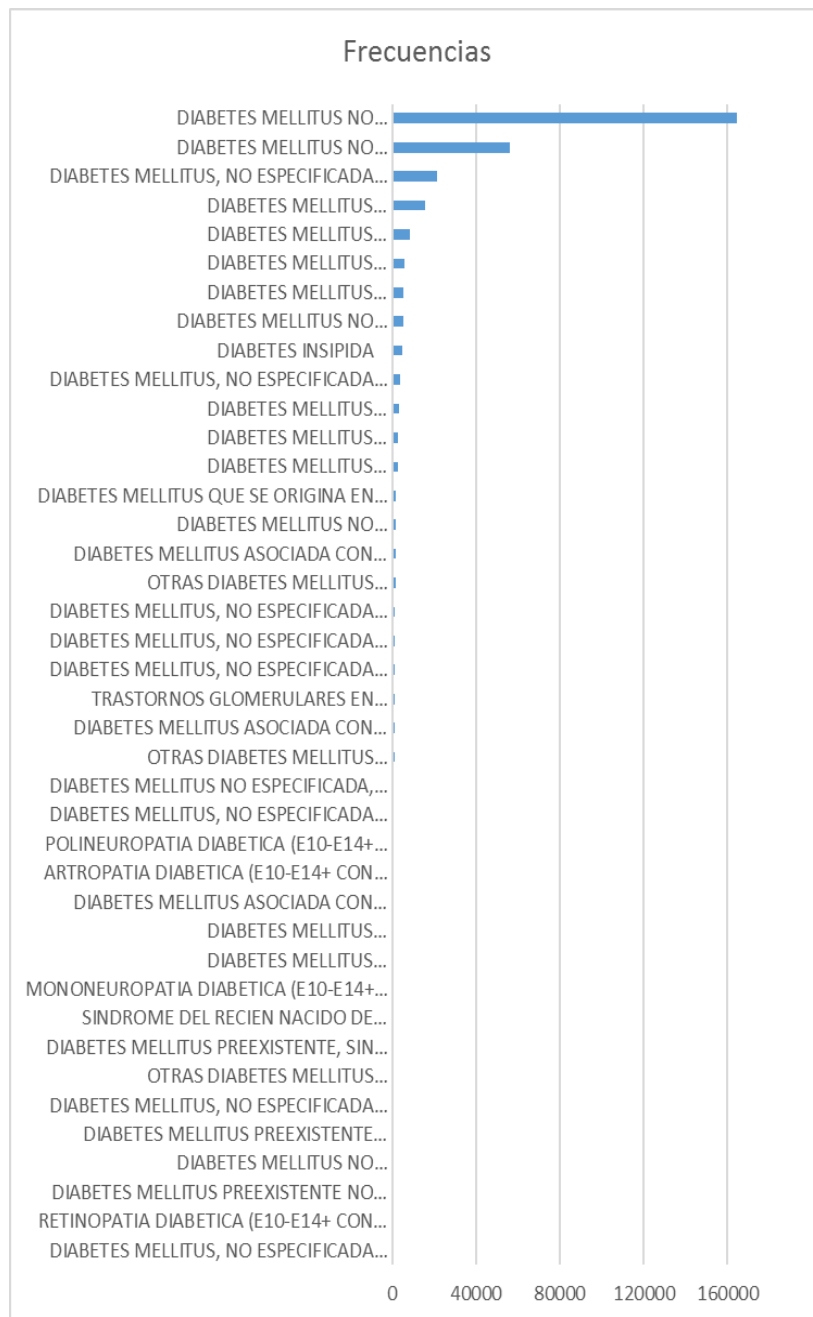


Figura 62. Gráfico de frecuencias diagnósticos Entidad B (Fuente propia, 2018).

A la nueva vista creada se le da el nombre de filtro_diabetes y el código para la búsqueda correspondiente se da a conocer en el Anexo 16.

HISTORIA	ORDEN DE TRABAJO	INGRESO	RANGO DE EDAD	SEXO	MUNICIPIO	DIAGNOSTICO	Principal / Secundario	EXAMEI	
1	1022	2304029	44	70-80	F	MEDELLIN	DIABETES MELLITUS NO INSULINODEPENDIENTE CON COM...	S	CITOQL
2	1022	2304029	44	70-80	F	MEDELLIN	DIABETES MELLITUS NO INSULINODEPENDIENTE CON COM...	S	CITOQL
3	1022	2304029	44	70-80	F	MEDELLIN	DIABETES MELLITUS NO INSULINODEPENDIENTE CON COM...	S	CITOQL
4	1022	2304029	44	70-80	F	MEDELLIN	DIABETES MELLITUS NO INSULINODEPENDIENTE CON COM...	S	CITOQL
5	1022	2304029	44	70-80	F	MEDELLIN	DIABETES MELLITUS NO INSULINODEPENDIENTE CON COM...	S	CREATI
6	1022	2304029	44	70-80	F	MEDELLIN	DIABETES MELLITUS NO INSULINODEPENDIENTE CON COM...	S	HEMOL
7	1022	2304029	44	70-80	F	MEDELLIN	DIABETES MELLITUS NO INSULINODEPENDIENTE CON COM...	S	HEMOL
8	1022	2304029	44	70-80	F	MEDELLIN	DIABETES MELLITUS NO INSULINODEPENDIENTE CON COM...	S	HEMOL
9	1022	2304029	44	70-80	F	MEDELLIN	DIABETES MELLITUS NO INSULINODEPENDIENTE CON COM...	S	HEMOL
10	1022	2304029	44	70-80	F	MEDELLIN	DIABETES MELLITUS NO INSULINODEPENDIENTE CON COM...	S	HEMOL
11	1022	2304029	44	70-80	F	MEDELLIN	DIABETES MELLITUS NO INSULINODEPENDIENTE CON COM...	S	MAGNE
12	1022	2304029	44	70-80	F	MEDELLIN	DIABETES MELLITUS NO INSULINODEPENDIENTE CON COM...	S	TIEMPC
13	1022	2304029	44	70-80	F	MEDELLIN	DIABETES MELLITUS NO INSULINODEPENDIENTE CON COM...	S	TIEMPC
14	1022	2304064	44	70-80	F	MEDELLIN	DIABETES MELLITUS NO INSULINODEPENDIENTE CON COM...	S	TROPO
15	1022	2304029	44	70-80	F	MEDELLIN	DIABETES MELLITUS NO INSULINODEPENDIENTE CON COM...	S	CITOQL
16	1022	2304029	44	70-80	F	MEDELLIN	DIABETES MELLITUS NO INSULINODEPENDIENTE CON COM...	S	CREATI
17	1022	2304029	44	70-80	F	MEDELLIN	DIABETES MELLITUS NO INSULINODEPENDIENTE CON COM...	S	HEMOL
18	1022	2304029	44	70-80	F	MEDELLIN	DIABETES MELLITUS NO INSULINODEPENDIENTE CON COM...	S	HEMOL
19	1022	2304029	44	70-80	F	MEDELLIN	DIABETES MELLITUS NO INSULINODEPENDIENTE CON COM...	S	HEMOL
20	1022	2304029	44	70-80	F	MEDELLIN	DIABETES MELLITUS NO INSULINODEPENDIENTE CON COM...	S	HEMOL
21	1022	2304029	44	70-80	F	MEDELLIN	DIABETES MELLITUS NO INSULINODEPENDIENTE CON COM...	S	HEMOL
22	1022	2304029	44	70-80	F	MEDELLIN	DIABETES MELLITUS NO INSULINODEPENDIENTE CON COM...	S	HEMOL
23	1022	2304029	44	70-80	F	MEDELLIN	DIABETES MELLITUS NO INSULINODEPENDIENTE CON COM...	S	TIEMPC
24	1022	2373891	48	70-80	F	MEDELLIN	DIABETES MELLITUS NO INSULINODEPENDIENTE CON COM...	S	CREATI

Query executed successfully. | ALVAROJ (13.0 RTM) | ALVAROJA\Alvaro (54) | EntidadB | 00:00:30 | 220793 rows

Figura 63. Datos vista filtro_diabetes (Fuente propia, 2018).

Con los datos ya organizados se puede continuar con el análisis de minería de datos, para ello se sigue el mismo proceso realizado en la sección de ETL y Minería de datos para los datos recopilados del UCI Machine Learning Repository. Los cambios necesarios para ejecutar los modelos de minería para los datos de la ENTIDAD A y la ENTIDAD B se dan a la hora de elegir el Origen de Datos y la Vista del Origen de Datos. Para la primera se selecciona como Origen de Datos la base de datos EntidadA y la Vista de Origen de Datos sería la vista filtro_diagnosticos. Para la segunda el Origen de Datos a escoger sería la base de datos EntidadB y la Vista de Origen de Datos es la vista filtro_diabetes.

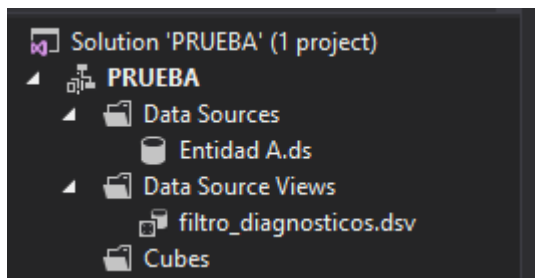


Figura 64. Origen de datos Entidad A (Fuente propia, 2018).

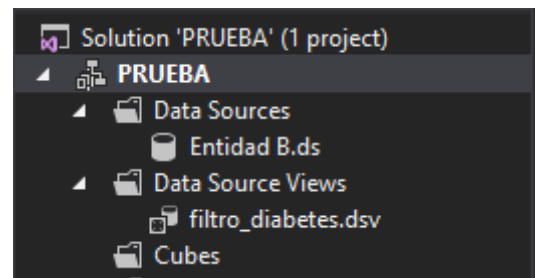


Figura 65. Origen de datos Entidad B (Fuente propia, 2018).

Los primeros modelos construidos se hicieron con el fin de determinar cuales eran las variables que presentan mayor dependencia con la variable de salida, que para estos estudios será el diagnóstico de diabetes en los pacientes que acuden a cada una de las entidades.

Para la ENTIDAD A se construyó un modelo de Árboles de decisión en el que se colocaron todas las variables como variables de entrada para el modelo y a medida que se llevaba a cabo el analisis de los resultados se van ignorando variables que de forma subjetiva se decide que sus características no ofrecen informacion relevante o que pueda incidir sobre la variable de salida.

Structure ↑	DecTree
	Microsoft Decision Trees
ASEGURADORA	Ignore
CONSECUTIVO HISTORIA C...	Key
DESCRIPCION DX CIE 10	PredictOnly
ESTADO CIVIL	Ignore
ESTADO	Ignore
FECHA DE INGRESO	Input
MUNICIPIO	Input
OCUPACION	Input
RANGO DE EDAD	Input
SERVICIO DE INGRESO	Input
SEXO	Input

Figura 66. Variables modelo Árbol de decisión Entidad A (Fuente propia, 2018)

Analizando el resultado final del modelo de árbol de desición y la red de dependencias, se observa que las variables que mas influyen en la variable de salida son la edad y el servicio de ingreso, por tanto serían las variables que se pretenden usar como entradas para los demas modelo.

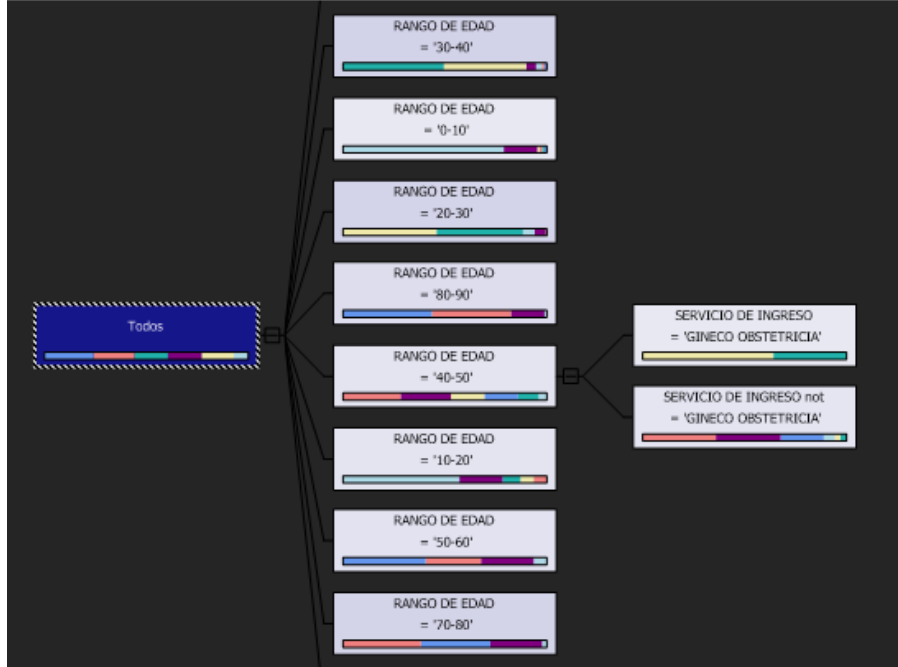


Figura 67. Árbol de decisión Entidad A (Fuente propia, 2018)

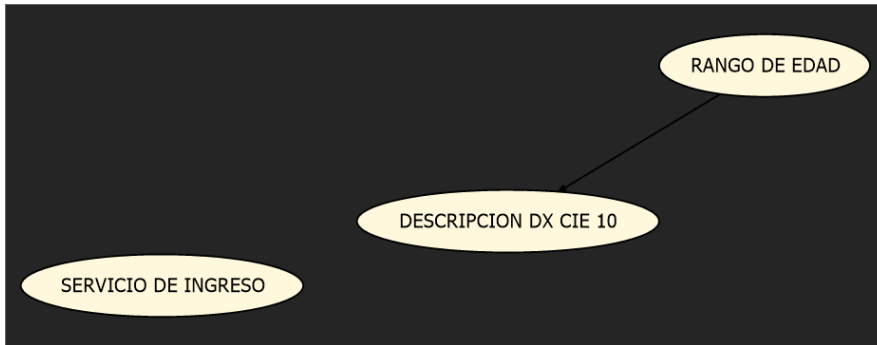


Figura 68. Red de dependencias Árbol de decisión Entidad A (Fuente propia, 2018)

Pero al observar el resultado de la precisión del modelo realizado por Analysis Services se ve un puntaje bajo de 0,47. Este resultado evidencia que es un modelo muy ambiguo, con una precisión del 47%, muy semejante a la probabilidad de acertar el lanzamiento de una moneda.

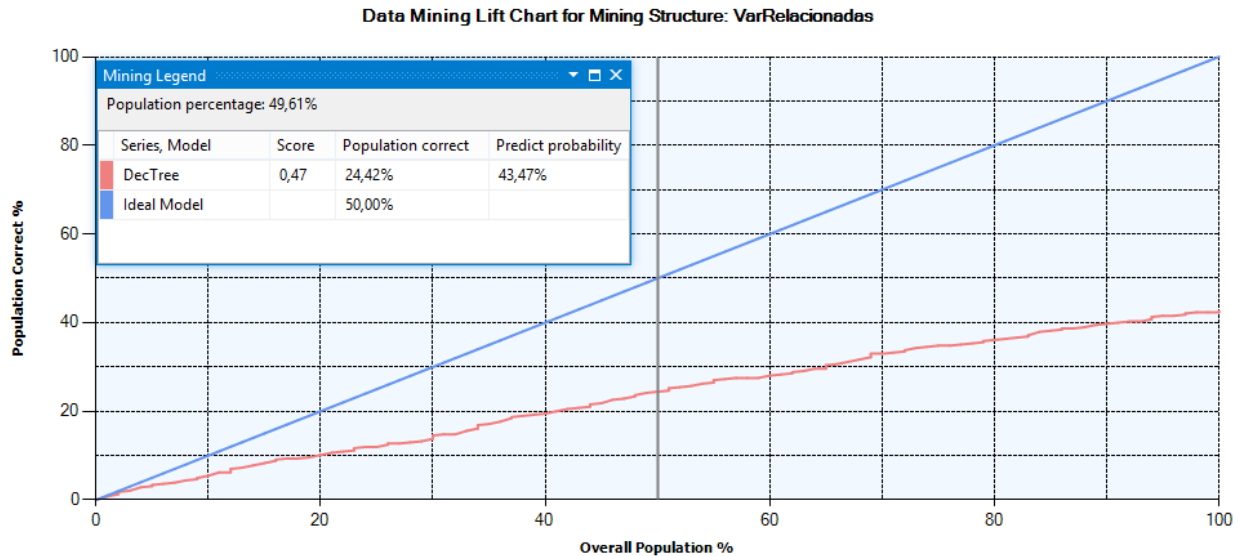


Figura 69. Lift chart árbol de decisión Entidad A (Fuente propia, 2018)

Por tanto se llega a la conclusión de que este modelo no ofrece mucha información. La razón que se da para explicar esta situación es la mencionada anteriormente donde al ver la **¡Error! No se encuentra el origen de la referencia.** Se ven frecuencias muy bajas para un modelo de minería de datos para los posibles valores que puede tomar la variable de salida. El algoritmo de Analysis Services para evaluar la precisión de un modelo utiliza un porcentaje de los datos para crear el modelo y el porcentaje restante lo utiliza para hacer predicciones, generalmente es una relación de 70/30 respectivamente, por tanto al tener pocos datos no se cuenta con una cantidad de datos suficiente para hacer las respectivas predicciones con el modelo creado.

Para los datos de la Entidad B se hace el mismo proceso para el análisis de la dependencias entre las variables de entrada con la variable de salida construyendo un modelo de árbol de decisión, y adicionando un modelo de regresión logística porque la variable de salida (Diagnosticos de diabetes) solo tiene 2 posibles valores, por esta razón se podría decir que es una variable booleana, y como se encontró en la caracterización de modelos de minería de datos, la regresión logística es una de las técnicas ideales a la hora de trabajar con variables booleanas.

También se hace una comparación con un modelo de redes neuronales, ya que en Analysis Services el algoritmo de regresión logística es un caso particular de redes neuronales, sería atractivo observar el comportamiento de las redes neuronales en este contexto.

Structure ↑	DecTreeB	NeurNetw	RegLog
	Microsoft_Decision_Trees	Microsoft_Neural_Network	Microsoft_Logistic_Regression
DESCRIPCION	Input	Input	Input
DIAGNOSTICO	PredictOnly	PredictOnly	PredictOnly
EXAMEN	Ignore	Ignore	Ignore
HISTORIA	Key	Key	Key
INGRESO	Ignore	Ignore	Ignore
MAX	Ignore	Ignore	Ignore
MIN	Ignore	Ignore	Ignore
MUNICIPIO	Input	Input	Input
ORDEN DE TRABAJO	Ignore	Ignore	Ignore
Principal Secundario	Ignore	Ignore	Ignore
RANGO DE EDAD	Input	Input	Input
RESULTADO	Ignore	Ignore	Ignore
SEXO	Input	Input	Input

Figura 70. Variables para los modelos Entidad B (Fuente propia, 2018)

Al analizar los resultados según el árbol de decisión las variables más dependientes del valor del diagnóstico son el sexo del paciente y su edad. Según la regresión logística son el municipio y la descripción del analito que se estudia en los exámenes de laboratorio clínico, mismas variables arrojadas por el modelo de redes neuronales, pero relaciona diferentes estados de dichas variables.



Figura 71. Árbol de decisión Entidad B (Fuente propia, 2018)

Attribute	Value	Favors DIABETES MELLITUS...	Favors DIABETES MELLITUS...
MUNICIPIO	NO APLICA		
DESCRIPCION	ANTIGENO E		
MUNICIPIO	CARTAGENA		
DESCRIPCION	FOSFORO SERICO		
MUNICIPIO	SALGAR		
MUNICIPIO	AMAGA		
DESCRIPCION	BILIRRUBINA DIRECTA		
MUNICIPIO	GIRARDOTA		
DESCRIPCION	ÃY-HIDROXI BUTIRATO EN SANGRE		
MUNICIPIO	RETIRO		

Figura 72. Regresión logística Entidad B (Fuente propia, 2018).

Attribute	Value	Favors DIABETES M...	Favors DIABETES M...
MUNICIPIO	AMAGA		
DESCRIPCION	B. GRAM. NEG		
DESCRIPCION	Ph DE LA ORINA		
MUNICIPIO	GUARNE		
DESCRIPCION	ANTIGENO DE DENGUE		
DESCRIPCION	EOSINOFILOS %		
DESCRIPCION	FOSFATASAS ALCALINAS		
MUNICIPIO	LA CEJA		
DESCRIPCION	REL T.P.T Pac/ T.P.T Cont		
MUNICIPIO	CARMEN DE VIBORAL		
MUNICIPIO	GIRARDOTA		
DESCRIPCION	BASOFILOS %		
DESCRIPCION	HEMOGLOBINA GLICADA (HbA1c)		

Figura 73. Redes neuronales Entidad B (Fuente propia, 2018).

Luego al observar la precisión de los modelos viendo el Lift chart de la comparación entre ellos se ve un comportamiento similar entre ellos con una precisión de más del 70%. También se miran los resultados de la tabla de Croos Validation para saber cuánto fue el Error cuadrático medio para cada uno de los modelos, obteniendo resultados del 27% de error.

Data Mining Lift Chart for Mining Structure: VarRelacionadas

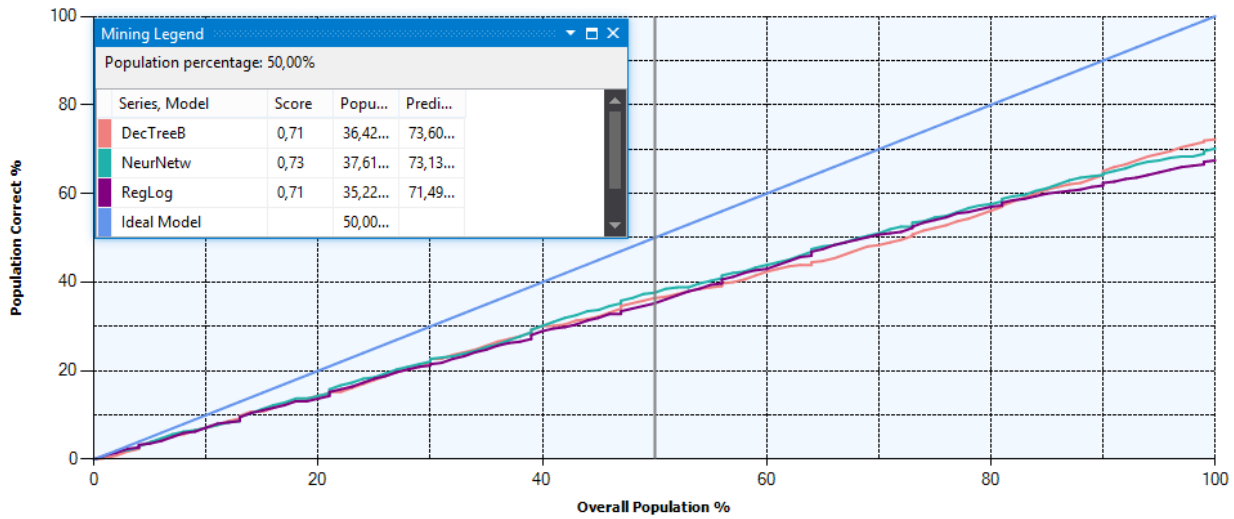


Figura 74. Lift chart Árbol de decisión vs Regresión logística vs Redes neuronales (Fuente propia, 2018).

DecTreeB				
Partition Index	Partition Size	Test	Measure	Value
1	78	Likelihood	Root Mean Square Error	0,2745
2	79	Likelihood	Root Mean Square Error	0,2758
3	79	Likelihood	Root Mean Square Error	0,2731
4	78	Likelihood	Root Mean Square Error	0,2777
5	78	Likelihood	Root Mean Square Error	0,2751
6	78	Likelihood	Root Mean Square Error	0,2755
7	78	Likelihood	Root Mean Square Error	0,2762
8	78	Likelihood	Root Mean Square Error	0,2749
9	78	Likelihood	Root Mean Square Error	0,2749
10	78	Likelihood	Root Mean Square Error	0,274
			Average	0,2752
			Standard Deviation	0,0012

Figura 75. Error cuadrático medio Árbol de decisión Entidad B (Fuente propia, 2018).

NeurNetw				
Partition Index	Partition Size	Test	Measure	Value
1	78	Likelihood	Root Mean Square Error	0,2628
2	79	Likelihood	Root Mean Square Error	0,2542
3	79	Likelihood	Root Mean Square Error	0,3011
4	78	Likelihood	Root Mean Square Error	0,238
5	78	Likelihood	Root Mean Square Error	0,2458
6	78	Likelihood	Root Mean Square Error	0,2699
7	78	Likelihood	Root Mean Square Error	0,2935
8	78	Likelihood	Root Mean Square Error	0,2849
9	78	Likelihood	Root Mean Square Error	0,265
10	78	Likelihood	Root Mean Square Error	0,2901
			Average	0,2705
			Standard Deviation	0,0202

Figura 76. Error cuadrático medio Redes neuronales Entidad B (Fuente propia, 2018).

RegLog				
Partition Index	Partition Size	Test	Measure	Value
1	78	Likelihood	Root Mean Square Error	0,2817
2	79	Likelihood	Root Mean Square Error	0,283
3	79	Likelihood	Root Mean Square Error	0,3164
4	78	Likelihood	Root Mean Square Error	0,2523
5	78	Likelihood	Root Mean Square Error	0,2742
6	78	Likelihood	Root Mean Square Error	0,2385
7	78	Likelihood	Root Mean Square Error	0,2594
8	78	Likelihood	Root Mean Square Error	0,2736
9	78	Likelihood	Root Mean Square Error	0,3172
10	78	Likelihood	Root Mean Square Error	0,2786
			Average	0,2775
			Standard Deviation	0,0238

Figura 77. Error cuadrático medio Regresión logística Entidad B (Fuente propia, 2018).

Como se tiene que los tres modelos tienen un rendimiento semejante con valores de precisión aceptables se prosigue a realizar un modelo de Clusterización en el que se creen grupos de pacientes con características compartidas, teniendo en cuenta las variables relacionadas que se hallaron en el análisis del árbol de decisión, redes neuronales y regresión logística. Con base en ello proponer alternativas para el mejoramiento del servicio prestado en la Entidad B.

Structure ↑	Cluster
	Microsoft_Clustering
DESCRIPCION	Input
DIAGNOSTICO	PredictOnly
HISTORIA	Key
MUNICIPIO	Input
RANGO DE EDAD	Input
SEXO	Input

Figura 78. Variables modelo de clusterización Entidad B (Fuente propia, 2018).

El modelo arrojó cinco clusters de pacientes, cada uno con su particularidad en cuanto a características de exámenes de laboratorio al que se sometieron, el municipio de residencia y sexo al que pertenecen. En los clusters 1 y 2 están clasificados el mayor número de pacientes, simbolizado en el color oscuro en el que son representados en la gráfica.

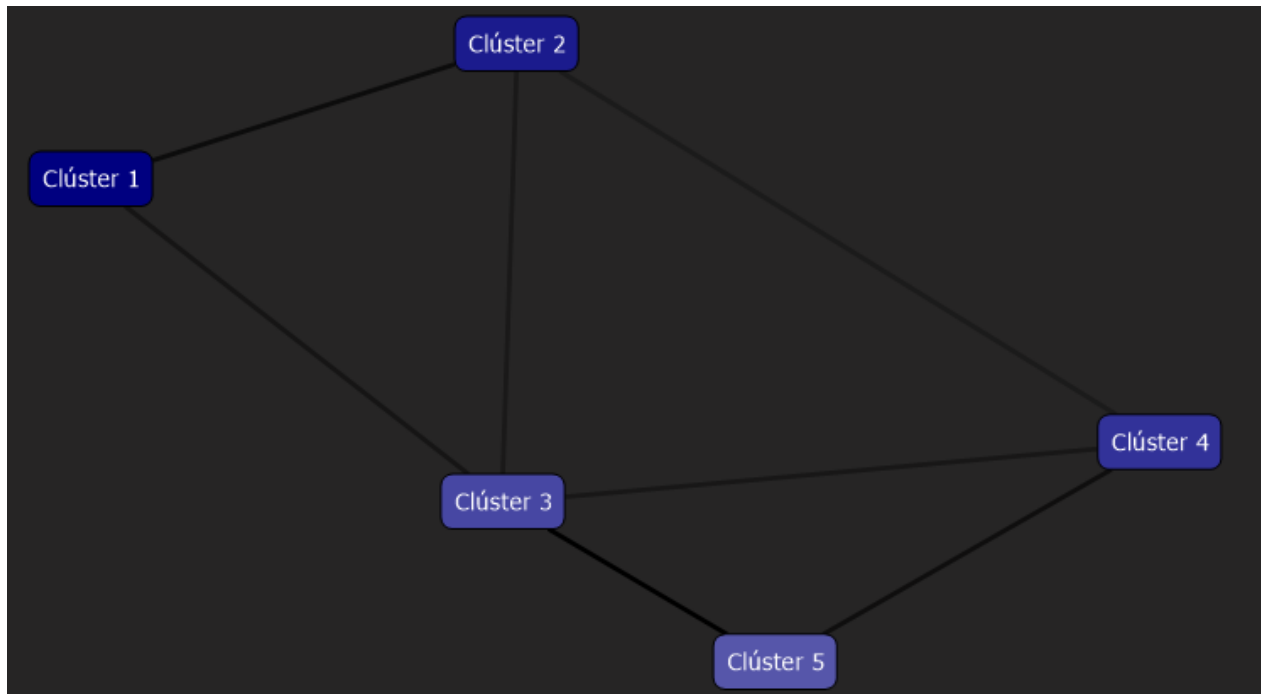


Figura 79. Clusters pacientes Entidad (Fuente propia, 2018).

Detallando el modelo se puede decir que ofreció buenos resultados, con una precisión del 75%, así que existe vía libre para seguir al análisis de cada cluster para realizar una proposición para el proceso de gestión dentro de la entidad.

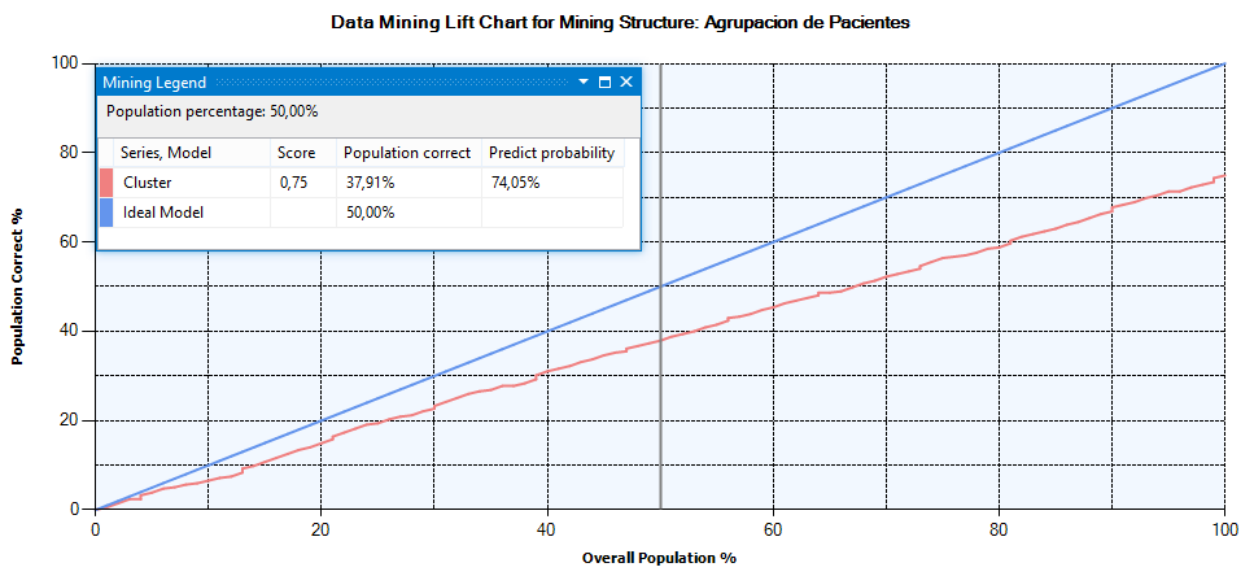


Figura 80. Lift chart cluster Entidad B (Fuente propia, 2018).



Figura 81. Perfiles de cada cluster (Fuente propia, 2018).

Como primera observación que salta a la vista de los clusters encontrados es el rango de edad al que pertenecen los pacientes, pues casi que la totalidad de los pacientes diabéticos se encuentran en un rango de edad de 50 años a 90 años de edad. Esta información no aporta nuevo conocimiento como tal, ya que es comúnmente conocido que las personas de mayor edad son las que poseen mayor tendencia a padecer diabetes. Pero confirmando que esta tendencia se cumple en los pacientes atendidos por la ENTIDAD B, se pueden tomar medidas buscando una mejora en la manera como se atienden los casos de pacientes diabéticos.

La función de los métodos de Data Mining no es propiamente la toma de decisiones sino entregar la información que está contenida en los datos almacenados de forma que se puedan identificar patrones y tendencias, para posteriormente con estudios de Business Intelligence crear las respectivas pautas que contribuyan a definir las acciones vayan acorde con los deseos y

objetivos de la organización teniendo en cuenta la información arrojada en el proceso de Data Mining.

A pesar de ello en este trabajo se menciona que los datos solo adquieren valor cuando se toman decisiones con miras a contribuir con un propósito determinado, es allí cuando los datos pasan de ser un montón de kilobytes a herramientas que ofrecen la posibilidad de lograr propósitos tangibles. Por esta razón se proponen algunas soluciones usando los resultados obtenidos en la metodología descrita.

Trayendo de vuelta la naturaleza de la edad de los pacientes diabéticos que son atendidos en la ENTIDAD B y de cómo teniendo este dato se puede contribuir al incremento de la calidad del servicio que se les presta, se plantea la opción de cambiar aspectos del ambiente de una persona hospitalizada, queriendo con ello que el paciente se sienta más a gusto y su estancia en las instalaciones de la ENTIDAD B sea de cierta forma algo más agradable. Si la norma lo permite, se propone hacer una adecuación de tipo estética o de cualquier otro tipo para ofrecer un ambiente algo más cómodo para personas de la tercera edad o que están próximos a estar catalogados dentro de este grupo social. Aspectos como el ruido de las alarmas de algunos equipos pueden perturbar la tranquilidad de los pacientes, para atender este aspecto en específico se pueden asignar unos equipos especiales para este tipo de pacientes, a los cuales se les disminuya el volumen de las alarmas y por medio de una red de conexiones por telemetría, dichos equipos envíen la señal de alarma a una estación central donde las personas encargadas del cuidado de los pacientes siempre sean advertidas de una posible eventualidad sin la necesidad de perturbar a los demás pacientes.

Otra observación que da pie a que se pueda tomar medidas es la relación entre el rango de edad de los pacientes en cada cluster y el examen de laboratorio clínico que más se practican lo pacientes de dicho cluster. Haciendo un conteo de la cantidad de paciente diabéticos agrupados por rangos de edad, se puede hacer un estimado de la cantidad de reactivos químicos necesarios para cumplir con la demanda de exámenes de laboratorio y hacer una mejor gestión de ellos, evitando sobrecostos debido a la compra excesiva de los implementos requeridos a la hora de llevar a cabo los exámenes de laboratorio para los pacientes hospitalizados o por el contrario si se le pide al proveedor una cantidad inferior a la que se necesita para cumplir la demanda para la prestación del servicio de laboratorio clínico, se puede caer en costos adicionales debido al gasto de transportes y distribución de los suministros. Porque una institución de salud además de ser una entidad cuyo objetivo principal es velar por la salud de los pacientes que acuden a ella, también es una entidad que busca generar los ingresos necesarios para al menos mantener su operación, y la disminución y/o optimización de los costos es una de las estrategias más lógicas que pueden ser utilizadas dentro de cualquier organización, en este caso gestionando de una mejor forma los consumibles necesarios para realizar las diferentes pruebas de laboratorio que les son practicadas a los pacientes diabéticos.

4. CONCLUSIONES Y CONSIDERACIONES FINALES

Se cumplió el objetivo del trabajo el cual consistía en entregar una herramienta que cumpla la función de un repositorio de datos usando los datos de historia clínica y del área de laboratorio clínico de distintas fuentes y que por medio de dicha herramienta se lograra la integración de todas ellas, y con ese objetivo cumplido proceder al análisis de los datos contenidos en el repositorio para poder identificar patrones que sirvan en la toma de decisiones en una institución de salud con miras al mejoramiento del servicio prestado.

La investigación realizada sobre los métodos de análisis de datos fue la parte más importante de todo el proceso, debido a que al iniciar este trabajo los conocimientos en el área de la analítica de datos era poca o casi nula. El entender, o tener al menos una perspectiva general, de todo el panorama del proceso necesario para llevar a cabo una minería de datos logró dar una visión más clara acerca de la forma de proceder en miras al cumplimiento de todos los objetivos planteados en el trabajo.

Con los datos obtenidos en la plataforma de libre acceso se obtuvo un entrenamiento adecuado en todas las etapas del proceso, para que a la hora de enfrentar los datos de un contexto local ya se tuviera cierta destreza para cumplir con el proceso de ETL, que generalmente para completar un análisis de minería de datos se necesita invertir cerca del 80% del todo el tiempo en esta etapa, sin dejar de lado la labor requerida en la aplicación de los modelos.

El rendimiento de la interfaz gráfica de usuario para el diseño del repositorio es satisfactorio, cumple con su función principal que es la búsqueda de la información que se encuentra contenida en cada una de las bases de datos integradas a la herramienta. Lo más destacable además del correcto funcionamiento es la escalabilidad con la que cuenta la interfaz, ya que su diseño se encuentra abierto a la posibilidad de agregarle más opciones, por ejemplo, al ingreso de

parámetros de búsquedas adicionales, así como agregar más fuentes que lleguen para integrarse a las actuales.

En cuanto al apartado del análisis de datos, se encontraron patrones en los datos que pueden ser usados para tomar decisiones para mejorar de cierta forma la calidad del servicio prestado por la ENTIDAD B, caso que no ocurrió para los datos de la ENTIDAD A. Aunque la etapa de tomar decisiones no hace parte de la minería de datos, sino que va después en la etapa de Business Intelligence, se intentó proponer alternativas que pueden afectar positivamente el aspecto financiero de la institución y al bienestar de los pacientes diabéticos que se encuentren remitidos en el área de hospitalización.

Todavía quedan muchas cosas por aprender de la analítica de datos, es un campo que ha tenido mayor auge en el área empresarial y comercial con el objetivo de optimizar procesos, disminuir costo, entre otras funcionalidades que si son aplicadas al área de la salud lograría una gran contribución para garantizar que se preste un servicio de calidad, todo esto encaminado al Big Data que es tan utilizado a nivel mundial.

REFERENCIAS

- The Data Governance Institute. (21 de Diciembre de 2012). *The Data Governance Institute Web*. Obtenido de <http://www.datagovernance.com/>
- A.KrishnaKumar, D.Amrita, & N.Swathi, P. (Abril de 2013). Mining Association Rules between Sets of Items in. *International Journal of Science and Modern Engineering (IJISME)*, 1(5).
- Accenture. (2015). *Servicios de Asesoramiento de Salud – Accenture*. Obtenido de https://www.accenture.com/_acnmedia/Accenture/Conversion-Assets/DotCom/Documents/Local/es-es/PDF_5/Accenture-El-Acceso-A-La-Historia-Clinica-Electronica.pdf
- Agile Design. (s.f.). *SSAS Entity Framework Provider - Product Description*. Obtenido de <http://www.agiledesignllc.com/Products>
- Alvarez, R. C. (2002). The promise of e-Health – a Canadian perspective. *BioMed Central*.
- Amaris, M. E., & Rodríguez, J. E. (2003). La contribución de las reglas de asociación a la minería de datos. *Tecnura*, 7(13), 94-109.
- American College of Physicians. (2008). *Position Paper. American College of Physicians*.
- González Sojo, D. (n.d.). Interfaz Grafica de Usuario. In Ampliacion de Simulador de Red. Iinterfaz Grafica del QSIM. Obtenido de <http://bibing.us.es/proyectos/abreproy/11300/fichero/PROYECTO%252FCapitulo3.pdf>
- Big Data Made Simple*. (1ero de Abril de 2015). Obtenido de <http://bigdata-madesimple.com/what-is-clustering-in-data-mining/>
- Catwell, L., & Sheikh, A. (18 de Agosto de 2009). Evaluating eHealth Interventions: The Need for Continuous Systemic Evaluation. *PLoS Medicine*, 8(6). Obtenido de <http://journals.plos.org/plosmedicine/article/file?id=10.1371/journal.pmed.1000126&type=printable>
- Comisión Europea. (2012). *eHealth Action Plan 2012-2020-Innovative healthcare for the 21st century*. Bruselas: Comisión Europea.
- Commission of the European Communities. (2004). *e-Health – making healthcare better for European citizens: An action plan for a European e-Health Area*. Bruselas.
- Commission of the European Communities. (2008). *Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions on Telemedicine for the benefit of patients, healthcare systems and society*. Bruselas.
- Cottle, M., Hoover, W., Kanwal, S., Kohn, M., Strome, T., & Treister, N. W. (2013). Transforming Health Care Through Big Data: Strategies for leveraging big data in the health care industry. *Institute for Health Technology Transformation*, <http://ihealthtran.com/big-data-in-healthcare>.

- Dianda, D., Quaglino, M., & Pagura, J. (Noviembre de 2016). *Facultad de Ciencias Económicas y Estadística: Universidad Nacional de Rosario*. Obtenido de https://www.fcecon.unr.edu.ar/web-nueva/sites/default/files/u16/Decimocuartas/dianda-quaglino-pagura_metodos_predictivos_de_data_mining.pdf
- El-Zoghby, J., Farouk, H. A., & El-Kilany, K. S. (2016). An Integrated Framework for Optimization of Resources in Emergency Departments. *6th International Conference on Industrial Engineering and Operations Management* (págs. 1621-1632). Kuala Lumpur: IEOM.
- Fayyad, U., Piatetsky-Shapiro, G., & Smyth, a. P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, 17(3).
- Forbes. (28 de octubre de 2018). *Forbes Media LLC. All Rights Reserved*. Obtenido de <https://www.forbes.com/sites/gartnergroup/#57c02e1c6deb>
- Fuente propia. (2018).
- Fundación Vodafone España, Red.es & Gobierno de España: Ministerio de Energía, Turismo y Agenda Virtual. (Marzo de 2017). *Informe Big Data en Salud Digital*. Obtenido de Ontsi-Red.es: <http://www.ontsi.red.es/ontsi/es/informe-big-data-en-salud-digital>
- Garets, D., & Davis, M. (2006). *Electronic Medical Records vs. Electronic Health Records: Yes, There Is a Difference*. Chicago: A HIMSS Analytics White Paper.
- Gartner, Inc. (28 de octubre de 2018). *Gartner®*. Obtenido de <https://www.gartner.com/en/research/methodologies/magic-quadrants-research>
- Gawande, A. (13 de Agosto de 2012). *Annals of Health Care*. Obtenido de The New Yorker: http://www.newyorker.com/reporting/2012/08/13/120813fa_fact_
- Glade. (29 de octubre de 2018). *Glade.gnome.org*. Obtenido de <https://glade.gnome.org/>
- Gonzalez, A. (1 de Julio de 2014). *Clever Data*. Obtenido de <http://cleverdata.io/que-es-machine-learning-big-data/>
- Google. (s.f.). *Google Health*. Obtenido de <https://www.google.com/accounts/ServiceLogin?service=health&nui=1&continue=https%3A%2F%2Fwww.google.com%2Fhealth%2Fp%2F&followup=https%3A%2F%2Fwww.google.com%2Fhealth%2Fp%2F&rm=hide>
- Groves, P., Kayyali, B., Knott, D., & Kuiken, S. V. (2013). The 'big data' revolution in healthcare: Accelerating value and innovation.
- Hani Neuvirth, M. O.-F. (2012). Toward Personalized Care Management of Patients at Risk--the Diabetes Case Study.
- Herland, M., Khoshgoftaar, T. M., & Wald, R. (2014). A review of data mining using big data in health informatics. *Journal of Big Data*.
- LEADA. (2015). *The Data Analytics Handbook: Big Data Edition*. Obtenido de <https://www.teamleada.com/handbook>

- Luo, J., Wu, M., Gopukumar, D., & Zhao, Y. (2016). Big Data Application in Biomedical Research and Health Care: A Literature Review. *Biomedical Informatics Insights*, 8, 1-10.
- Lynch, O. (6 de julio de 2017). *Cibrary*. Obtenido de <https://www.cybrary.it/0p3n/database-differences-microsoft-sql-server-vs-oracle-database/>
- MathWorks®. (s.f.). Obtenido de https://la.mathworks.com/help/matlab/creating_guis/about-the-simple-guide-gui-example.html
- McAfee, A. a. (2012). Big Data: The Management Revolution. *Harvard Business Review*, 61-68.
- McKinsey Global Institute. (2011). *Big data: The next frontier for innovation, competition, and productivity*. McKinsey & Company.
- McLean, S., & Sheikh, A. (2009). Does telehealthcare offer a patient-centred way forward for the community-based management of long-term respiratory disease? *Primary Care Respiratory Journal*, 3(18), 125-126.
- Microsoft. (5 de julio de 2018). *SQL Server Integration Services*. Obtenido de <https://docs.microsoft.com/es-es/sql/integration-services/sql-server-integration-services?view=sql-server-2017>
- Microsoft Corporation. (3 de noviembre de 2017). *SQL Server Blog*. Obtenido de <https://cloudblogs.microsoft.com/sqlserver/2017/11/03/three-years-in-a-row-microsoft-is-a-leader-in-the-odbms-magic-quadrant/>
- Microsoft. (s.f.). *HealthVault*. Obtenido de <http://www.healthvault.com/>
- Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill Science/Engineering/Math.
- Monegan, B. (24 de Abril de 2012). Federal Medicare Fraud Detection at an all-time high. *Health care IT News*. Obtenido de <http://www.healthcareitnews.com/news/federal-medicare-fraud-prevention-all-time-high>
- Monteserin, A. (13 de Abril de 2018). *Facultad de Ciencias Exactas: Universidad Nacional del Centro de la Provincia de Buenos Aires*. Obtenido de http://www.exa.unicen.edu.ar/catedras/optia/public_html/2018%20Reglas%20de%20asociaci%C3%B3n.pdf
- Naeem, M. (octubre de 2015). *HnHSol*. Obtenido de <http://hnhsol.blogspot.com/2015/11/garter-magic-quadrant-for-operational.html>
- NHS. (s.f.). *HealthSpace*. Obtenido de <https://www.healthspace.nhs.uk/visitor/default.aspx>
- Office of Technology Assessment. (1976). Development of Medical Technology: Opportunities for Assessment. *United States Congress Office of Technology Assessment*.
- Pagliari, C., Detmer, D., & Singleton, P. (2007). *Electronic personal health records: emergence and implications for the UK*. Londres: Nuffield Trust.
- Rodríguez Elizalde, J. M. (2006). *Calsificación de series de tiempo por minería de datos*. Mexico D. F.: Insitituto Politécnico Nacional: Centro de Investigación en Computación.

- Sankoh, O., Herbst, A. J., Juvekar, S., Tollman, S., Byass, P., & Tanner, M. (2013). INDEPTH launches a data repository and INDEPTHStats. *The Lancet Global Health*.
- Santana, E. (1 de Abril de 2015). *Data Mining con R*. Obtenido de <http://apuntes-r.blogspot.com/2015/04/regresion-lineal-simple.html>
- SEH-LELHA. (21 de enero de 2001). *Sociedad Española de Hipertensión: Liga Española para la Lucha contra la Hipertensión Arterial*. Obtenido de <https://www.seh-lelha.org/la-regresion-logistica/>
- Seif, G. (5 de Febrero de 2018). *Towards Data Science*. Obtenido de <https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68>
- Shakya, K. M. (30 de enero de 2017). *Quora.com*. Obtenido de <https://www.quora.com/How-do-I-make-a-GUI-for-a-Java-program>
- Silber, D. (2003). The Case for eHealth. *European Commission's first high-level conference on eHealth*. European Institute of Public Administration.
- Sinnexus Business Intelligence*. (s.f.). Obtenido de http://www.sinnexus.com/business_intelligence/datamining.aspx
- Smith, L. &. (2004). *EE.UU. Patente n° 10/406,836*.
- Soares, S. (13 de Junio de 2012). *IBM Data Mag*. Obtenido de <http://ibmdatamag.com/2012/06/a-framework-that-focuses-on-thedata-in-big-data-governance/>
- Strack, B., DeShazo, J. P., Gennings, C., Olmo, J. L., Ventura, S., Cios, K. J., & Clore, J. N. (2014). Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records. *BioMed Research International*, 11.
- Támez, M. (12 de mayo de 2016). *Monitor Educativo*. Obtenido de <https://monitor.iiiipe.edu.mx/notas/%C2%BFpor-qu%C3%A9-las-redes-neuronales-artificiales-son-el-futuro>
- TechAmerica Foundation. (2012). *Demystifying Big Data: A Practical Guide to Transforming the Business of Government*. Washington, D.C.: TechAmerica Foundation.
- Tutorials Point*. (s.f.). Obtenido de https://www.tutorialspoint.com/data_mining/dm_cluster_analysis.htm
- University of California, Irvine. (15 de noviembre de 2018). Obtenido de <https://archive.ics.uci.edu/ml/about.html>
- University of California, Irvine. (18 de 11 de 2018). *Diabetes 130-US hospitals for years 1999-2008 Data Set*. Obtenido de <https://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008#>
- Uriarte, A. G., Zúñiga, E. R., Moris, M. U., & Ng, A. H. (2017). How can decision makers be supported in the improvement of an emergency department? A simulation, optimization and data mining approach. *Operations Research for Health Care*, 102-122.

- Vallejos, S. J., & Martínez, D. L. (2006). *Trabajo de Adscripción: Minería de Datos*. Corrientes, Argentina: Universidad Nacional del Nordeste: Facultad de Ciencias Exactas, Naturales y Agrimensura.
- Walker, M. (13 de noviembre de 2016). *Data Science Association*. Obtenido de <http://www.datascienceassn.org/content/operational-database-management-systems-dbms-magic-quadrant-2016>
- Wyber, R., Vaillancourt, S., Perry, W., Mannava, P., Folaranmi, T., & Celi, L. A. (2015). Big data in global health: improving health in low- and. *Bulletin of the World Health*, 203-208.
- Young, T., Eatock, J., Jahangirian, M., Naseer, A., & Lilford, R. (2009). Three critical challenges for modeling and simulation in healthcare. *Winter Simulation Conference*, (págs. 1823-1830). Austin, TX, USA.
- Zicardi, R. (16 de octubre de 2015). *ODBMS.org*. Obtenido de <http://www.odbms.org/2015/10/gartner-magic-quadrant-for-operational-dbms/>

ANEXOS

ANEXO 1. Código query para crear la vista opensource_view

```
SELECT      dbo.diabetic_data.encounter_id AS Encuentro, dbo.diabetic_data.patient_nbr
AS Paciente, dbo.diabetic_data.race AS Etnia, dbo.diabetic_data.gender AS
Genero, dbo.diabetic_data.age AS Edad, dbo.admission_type.description AS
TipoAdmision, dbo.discharge_disposition.description AS DescripcionAlta,
dbo.admission_source.description AS OrigenAdmision,
dbo.diabetic_data.time_in_hospital AS TiempoHospital,
dbo.diabetic_data.payer_code AS TipoPago,
dbo.diabetic_data.medical_specialty AS EspecilidadMedica,
dbo.diabetic_data.num_lab_procedures AS NumLabProced,
dbo.diabetic_data.num_procedures AS NumProced,
dbo.diabetic_data.num_medications AS NumMedicamentos,
dbo.diabetic_data.number_outpatient, dbo.diabetic_data.number_emergency,
dbo.diabetic_data.number_inpatient, dbo.diag_1.ENFERMEDAD AS diag_1,
dbo.diabetic_data.diag_2, dbo.diabetic_data.diag_3,
dbo.diabetic_data.number_diagnoses, dbo.diabetic_data.max_glu_serum,
dbo.diabetic_data.A1Cresult, dbo.diabetic_data.metformin,
dbo.diabetic_data.repaglinide, dbo.diabetic_data.nateglinide,
dbo.diabetic_data.chlorpropamide, dbo.diabetic_data.glimepiride,
dbo.diabetic_data.acetohexamide, dbo.diabetic_data.glipizide,
dbo.diabetic_data.glyburide, dbo.diabetic_data.tolbutamide,
dbo.diabetic_data.pioglitazone, dbo.diabetic_data.rosiglitazone,
dbo.diabetic_data.acarbose, dbo.diabetic_data.miglitol,
dbo.diabetic_data.troglitazone, dbo.diabetic_data.tolazamide,
dbo.diabetic_data.examide, dbo.diabetic_data.citoglipton,
dbo.diabetic_data.insulin, dbo.diabetic_data.[glyburide-metformin],
dbo.diabetic_data.[glipizide-metformin], dbo.diabetic_data.[glimepiride-
pioglitazone], dbo.diabetic_data.[metformin-rosiglitazone],
dbo.diabetic_data.[metformin-pioglitazone], dbo.diabetic_data.change,
dbo.diabetic_data.diabetesMed, dbo.diabetic_data.readmitted
FROM        dbo.diabetic_data

INNER JOIN  dbo.admission_source ON dbo.admission_source.admission_source_id =
dbo.diabetic_data.admission_source_id
INNER JOIN  dbo.admission_type ON dbo.admission_type.admission_type_id =
dbo.diabetic_data.admission_type_id
INNER JOIN  dbo.discharge_disposition ON
dbo.discharge_disposition.discharge_disposition_id =
dbo.diabetic_data.discharge_disposition_id
INNER JOIN  dbo.diag_1 ON dbo.diag_1.[CODIGOS ICD9] = dbo.diabetic_data.diag_1
```

ANEXO 2. Código búsqueda en la vista opensource_view por medio de la GUI

```
function pushbutton1_Callback(hObject, eventdata, handles)

cond = string(get(handles.edit1, 'String'));
col = get(handles.popupmenu1, 'Value');

save varlop cond
save var2op col
```

```

if col == 2
    query='SELECT * FROM OpenSource.dbo.opensource_view WHERE Encuentro = ' +
cond;
    save varop query
end

if col == 3
    query='SELECT * FROM OpenSource.dbo.opensource_view WHERE Paciente = ' +
cond;
    save varop query
end

close
TablaTG

```

ANEXO 3. Código tabla para presentación de resultados de la búsqueda

```

load varop query
load varlop cond
load var2op col

if cond == "" || col == 1
    close
    AvisoVacioTG
else
    conn = database('ALVARO', '', '');

    curs = exec(conn, [query]);
    curs = fetch(curs);
    data = curs.Data

    if size(data)== 1
        close
        AvisoSinCoincidencias
    else
        c=table2cell(data);
        final=transpose(c);

        nombres={'Encuentro', 'Paciente', 'Etnia', 'Genero', 'Edad', 'Tipo de
Admision', 'Descripcion Alta', 'Origen de Admision', 'Tiempo en
Hospital', 'Tipo de Pago', 'Especilidad Medica', 'Numero de Proc de
Laboratorio', 'Numero de Procedimientos', 'Numero de Medicamentos',
'Numero Paciente Ambulatorio', 'Numero Emergencia', 'Numero Paciente
Hospitalizado', 'Diagnostico 1', 'Diagnostico 2', 'Diagnostico 3',
'Numero de Diagnosticos', 'max_glu_serum', 'Resultado AlC',
'Metformina', 'Repaglinida', 'Nateglinida', 'Clorpropamida',
'Glimepirida', 'Acetohexamida', 'Glipizide', 'Gliburida',
'Tolbutamida', 'Pioglitazone', 'Rosiglitazone', 'Acarbosa',
'Miglitol', 'Troglitazone', 'Tolazamida', 'Examida', 'Sitagliptina',
'Insulina', 'Gliburida-metformina', 'Glipizida-metformina',
'Glimepirida-mioglitazona', 'Metformina-rosiglitazona', 'Metformina-
pioglitazona', 'Cambio', 'Diabetes Med', 'Readmitido'};

        set(handles.uitable1, 'data', final);
    end
end

```

```

        set(handles.uitable1, 'RowName', nombres);
        set(handles.uitable1, 'ColumnWidth', {400 400 400 400});

        close(curs)
        close(conn)
        end
    end
end

```

```
function pushbutton1_Callback(hObject, eventdata, handles)
```

```

    close
    OpenSource;

```

ANEXO 4. Código aviso sin coincidencias

```
function pushbutton1_Callback(hObject, eventdata, handles)
```

```

close
OpenSource;

```

ANEXO 5. Código aviso campo vacío

```
function pushbutton1_Callback(hObject, eventdata, handles)
```

```

close
OpenSource;

```

ANEXO 6. Código SQL query para crear la vista ETL

```

SELECT      dbo.Opensource_view.Encuentro, dbo.Opensource_view.Paciente,
            dbo.Opensource_view.Etnia, dbo.Opensource_view.Genero,
            dbo.Opensource_view.Edad, dbo.Opensource_view.DescripcionAlta,
            dbo.Opensource_view.OrigenAdmision, REPLACE(Opensource_view.EspecilidadMedica
            , '?', 'Missing') AS EspecialidadMedica, dbo.Opensource_view.TiempoHospital,
            dbo.Opensource_view.diag_1 AS DiagnosticoPrimario,
            dbo.Opensource_view.A1Cresult, dbo.Opensource_view.readmitted AS Readmision

FROM        dbo.Opensource_view

INNER JOIN  dbo.primer_encuentro ON dbo.Opensource_view.Encuentro =
            dbo.primer_encuentro.PrimerEncuentro
INNER JOIN  dbo.filtroDA ON dbo.filtroDA.DescripcionAlta =
            dbo.Opensource_view.DescripcionAlta

```

ANEXO 7. Fórmula para la clasificación de la edades

```

=SI(EDAD <= 10; "0-10"; SI(Y(EDAD <=20; EDAD > 10); "10-20";
SI( Y(EDAD <= 30; EDAD > 20); "20-30"; SI( Y(EDAD <= 40; EDAD >30); "30-40";
SI( Y(EDAD <= 50; EDAD > 40); "40-50"; SI( Y(EDAD <= 60; EDAD >50); "50-60";
SI( Y(EDAD <= 70; EDAD > 60); "60-70"; SI( Y(EDAD <= 80; EDAD >70); "70-80";

```

```
SI( Y(EDAD <= 90; EDAD > 80); "80-90"; SI( Y(EDAD <= 100; EDAD > 90); "90-100";
">100")))))))))))
```

ANEXO 8. Código SQL query para crear la vista entidadA_view

```
SELECT      [CONSECUTIVO HISTORIA CLINICA], [NRO DE INGRESO], [RANGO DE EDAD], [SEXO],
[MUNICIPIO], [DESCRIPCION DX CIE 10], [ESTADO], REPLACE([SERVICIO DE
INGRESO], ' CENTRO', '') AS [SERVICIO DE INGRESO], [ASEGURADORA], [ESTADO
CIVIL], [OCUPACION], [FECHA DE INGRESO]
FROM        [EntidadA].[dbo].[centro]

UNION

SELECT      [CONSECUTIVO HISTORIA CLINICA], [NRO DE INGRESO], [RANGO DE EDAD], [SEXO],
[MUNICIPIO], [DESCRIPCION DX CIE 10], [ESTADO], REPLACE([SERVICIO DE
INGRESO], ' TESORO', '') AS [SERVICIO DE INGRESO], [ASEGURADORA], [ESTADO
CIVIL], [OCUPACION], [FECHA DE INGRESO]
FROM        [EntidadA].[dbo].[tesoro]
```

ANEXO 9. Código SQL query para crear la vista entidadB_view

```
SELECT      Egresos2016.Historia AS HISTORIA, [ORDEN DE TRABAJO],
[Hospitalizados2016].INGRESO, [Rango de Edad] AS [RANGO DE EDAD], [Sexo] AS
SEXO, [Descr_Mun] AS MUNICIPIO, [Descr_Diag] AS DIAGNOSTICO, [Principal /
Secundario], [DESCRIP_EXAMEN] AS EXAMEN, [DESCRIPCIÓN ] AS DESCRIPCION,
[RESULTADO], [MIN], [MAX]
FROM        Hospitalizados2016
INNER JOIN  Egresos2016 ON Hospitalizados2016.HISTORIA = Egresos2016.Historia
INNER JOIN  Diagnosticos ON Diagnosticos.Cod_Diag = Egresos2016.[Codigo diagnóstico]
INNER JOIN  Municipios ON Municipios.Cod_Mun = Egresos2016.[Municipio de residencia]
INNER JOIN  Examenes ON Examenes.[COD_EXAMEN] = Hospitalizados2016.COD_EXAMEN

UNION

SELECT      Egresos2017.Historia AS HISTORIA, [ORDEN DE TRABAJO], [Hospitalizados2017].
INGRESO, [Rango de Edad] AS [RANGO DE EDAD], [Sexo] AS SEXO, [Descr_Mun] AS
MUNICIPIO, [Descr_Diag] AS DIAGNOSTICO, Principal / Secundario],
[DESCRIP_EXAMEN] AS EXAMEN, [DESCRIPCIÓN ] AS DESCRIPCION, [RESULTADO],
[MIN], [MAX]
FROM        Hospitalizados2017
INNER JOIN  Egresos2017 ON Hospitalizados2017.HISTORIA = Egresos2017.Historia
INNER JOIN  Diagnosticos ON Diagnosticos.Cod_Diag = Egresos2017.[Codigo diagnóstico]
INNER JOIN  Municipios ON Municipios.Cod_Mun = Egresos2017.[Municipio de residencia]
INNER JOIN  Examenes ON Examenes.[COD_EXAMEN] = Hospitalizados2017.COD_EXAMEN
```

ANEXO 10. Nueva programación para la muestra de resultados de búsqueda

```
load fuenteg fuente

if fuente == 2
```

```

load varop query
load varlop cond
load var2op col

if cond == "" || col == 1
    close
    AvisoVacioTG
else
    conn = database('ALVARO', '', '');

    curs = exec(conn, [query]);
    curs = fetch(curs);
    data = curs.Data

    if size(data) == 1
        close
        AvisoSinCoincidencias
    else
        c=table2cell(data);
        final=transpose(c);

nombres={'Encuentro', 'Paciente', 'Etnia', 'Genero', 'Edad', 'Tipo de
Admision', 'Descripcion Alta', 'Origen de Admision', 'Tiempo en
Hospital', 'Tipo de Pago', 'Especilidad Medica', 'Numero de Proc de
Laboratorio', 'Numero de Procedimientos', 'Numero de Medicamentos',
'Numero Paciente Ambulatorio', 'Numero Emergencia', 'Numero Paciente
Hospitalizado', 'Diagnostico 1', 'Diagnostico 2', 'Diagnostico 3',
'Numero de Diagnosticos', 'max_glu_serum', 'Resultado AlC',
'Metformina', 'Repaglinida', 'Nateglinida', 'Clorpropamida',
'Glimepirida', 'Acetohexamida', 'Glipizide', 'Gliburida',
'Tolbutamida', 'Pioglitazone', 'Rosiglitazone', 'Acarbosa',
'Miglitol', 'Troglitazone', 'Tolazamida', 'Examida', 'Sitagliptina',
'Insulina', 'Gliburida-metformina', 'Glipizida-metformina',
'Glimepirida-mioglitazona', 'Metformina-rosiglitazona', 'Metformina-
pioglitazona', 'Cambio', 'Diabetes Med', 'Readmitido'};

        set(handles.uitable1, 'data', final);
        set(handles.uitable1, 'RowName', nombres);
        set(handles.uitable1, 'ColumnWidth', {400 400 400 400});

        close(curs)
        close(conn)
    end
end
end

if fuente == 3
    load varea query

    conn = database('ALVARO', '', '');

    curs = exec(conn, [query]);
    curs = fetch(curs);
    data = curs.Data

```

```

if size(data)== 1
    close
    AvisoSinCoincidencias
else
    final=table2cell(data);

    nombres={'CONSECUTIVO HISTORIA CLINICA', 'NRO DE INGRESO', 'RANGO
DE EDAD', 'SEXO', 'MUNICIPIO DE RESIDENCIA', 'DESCRIPCION DX CIE
10', 'ESTADO', 'SERVICIO DE INGRESO', 'ASEGURADORA', 'ESTADO
CIVIL', 'OCUPACION', 'FECHA DE INGRESO'};

    set(handles.uitable1, 'data', final);
    set(handles.uitable1, 'ColumnName', nombres);
    set(handles.uitable1, 'ColumnWidth', {'Auto' 'Auto' 'Auto' 'Auto'
'Auto' 550 'Auto' 'Auto' 300 'Auto' 'Auto' 'Auto'});

    close(curs)
    close(conn)
end
end

if fuente == 4
    load vareb query

    conn = database('ALVARO', '', '');

    curs = exec(conn, [query]);
    curs = fetch(curs);
    data = curs.Data

    if size(data)== 1
        close
        AvisoSinCoincidencias
    else
        final=table2cell(data);

        nombres={'HISTORIA', 'ORDEN DE TRABAJO', 'INGRESO', 'RANGO DE
EDAD', 'SEXO', 'MUNICIPIO', 'DIAGNOSTICO', 'Principal /
Secundario', 'EXAMEN', 'DESCRIPCION', 'RESULTADO', 'MIN', 'MAX'};

        set(handles.uitable1, 'data', final);
        set(handles.uitable1, 'ColumnName', nombres);
        set(handles.uitable1, 'ColumnWidth', {'Auto' 'Auto' 'Auto' 'Auto'
'Auto' 100 550 'Auto' 400 400 'Auto' 'Auto' 'Auto'});

        close(curs)
        close(conn)
    end
end
end

```

ANEXO 11. Programación ventana para definición de parámetros de búsqueda Entidad A

```
function pushbutton1_Callback(hObject, eventdata, handles)
```



```
cond = string(get(handles.edit2, 'String'));
query='SELECT * FROM EntidadA.dbo.entidadA_view WHERE [CONSECUTIVO
HISTORIA CLINICA] = ' + cond;
```

```
save varea query
```

```
if cond == ""
    abrir='AvisoVacioTG';
else
    abrir='TablaTG';
end
```

```
close
run(abrir);
```

```
function pushbutton2_Callback(hObject, eventdata, handles)
```

```
close
SeleccionFuente
```

ANEXO 12. Programación ventana para definición de parámetros de búsqueda Entidad B

```
function pushbutton1_Callback(hObject, eventdata, handles)
```

```
cond = string(get(handles.edit1, 'String'));
query='SELECT * FROM EntidadB.dbo.entidadB_view WHERE HISTORIA = ' +
cond;
```

```
save vacioeb cond
save vareb query
```

```
if cond == ""
    abrir='AvisoVacioTG';
else
    abrir='TablaTG';
end
```

```
close
run(abrir);
```

```
function pushbutton2_Callback(hObject, eventdata, handles)
```

```
close
SeleccionFuente
```

ANEXO 13. Programación para integración de distintas fuentes

```
function pushbutton1_Callback(hObject, eventdata, handles)
```

```
fuentes = get(handles.popupmenu1, 'Value');
save fuentes fuentes
```

```

if fuente == 1
  close
  abrir = 'AvisoVacioTG';
end

if fuente == 2
  close
  abrir = 'OpenSource';
end

if fuente == 3
  close
  abrir = 'EntidadA';
end

if fuente == 4
  close
  abrir = 'EntidadB';
end
close
run(abrir);

```

ANEXO 14. SQL query para la vista filtro_diagnosticos

```

SELECT *
FROM      EntidadA.dbo.entidadA_view
WHERE [DESCRIPCION DX CIE 10] = 'DIABETES MELLITUS NO INSULINODEPENDIENTE SIN MENCION DE
COMP'
UNION

SELECT *
FROM      EntidadA.dbo.entidadA_view
WHERE [DESCRIPCION DX CIE 10] = 'DIABETES MELLITUS NO ESPECIFICADA SIN MENCION DE
COMPLICACI'
UNION

SELECT *
FROM      EntidadA.dbo.entidadA_view
WHERE [DESCRIPCION DX CIE 10] = 'DIABETES MELLITUS INSULINODEPENDIENTE SIN MENCION DE
COMPLIC'
UNION

SELECT *
FROM      EntidadA.dbo.entidadA_view
WHERE [DESCRIPCION DX CIE 10] = 'DIABETES MELLITUS QUE SE ORIGINA EN EL EMBARAZO'
UNION

SELECT *
FROM      EntidadA.dbo.entidadA_view
WHERE [DESCRIPCION DX CIE 10] = 'DIABETES MELLITUS NO ESPECIFICADA EN EL EMBARAZO'
UNION

SELECT *
FROM      EntidadA.dbo.entidadA_view
WHERE [DESCRIPCION DX CIE 10] = 'DIABETES MELLITUS INSULINODEPENDIENTE CON CETOACIDOSIS'

```

ANEXO 15. Query para determinación de frecuencias

```
SELECT DIAGNOSTICO, COUNT(*) AS FREQ FROM EntidadB.dbo.entidadB_view
GROUP BY DIAGNOSTICO
ORDER BY FREQ DESC
```

ANEXO 16. SQL query para vista filtro_diabetes

```
SELECT *
FROM EntidadB.dbo.entidadB_view
WHERE DIAGNOSTICO = 'DIABETES MELLITUS NO INSULINODEPENDIENTE CON COMPLICACIONES'
```

UNION

```
SELECT *
FROM EntidadB.dbo.entidadB_view
WHERE DIAGNOSTICO = 'DIABETES MELLITUS NO INSULINODEPENDIENTE SIN MENCIÓN DE COMP'
```

ANEXO 17. Lista de características y sus descripciones en el conjunto de datos inicial. (Strack, y otros, 2014)

Nombre de la característica	Tipo	Descripción y valores	% valores faltantes
Encounter ID	Numérico	Identificador único de un encuentro	0%
Patient number	Numérico	Identificador único de un paciente.	0%
Race	Nominal	Valores: Caucasian, Asian, African American, Hispanic, and other	2%
Gender	Nominal	Valores: male, female, and unknown/invalid	0%
Age	Nominal	Agrupados en intervalos de 10 años: [0, 10), [10, 20), ..., [90, 100)	0%
Weight	Numérico	Peso en libras	97%
Admission type	Nominal	Identificador de entero correspondiente a 9 valores distintos, por ejemplo, emergencia, urgente, electivo, recién nacido y no disponible	0%
Discharge disposition	Nominal	Identificador de entero correspondiente a 29 valores distintos, por ejemplo, descargado a domicilio, caducado y no disponible	0%
Admission source	Nominal	Identificador de entero correspondiente a 21 valores distintos, por ejemplo, derivación de un médico, sala de emergencias y traslado de un hospital	0%
Time in hospital	Numérico	Número entero de días entre la admisión y el alta.	0%
Payer code	Nominal	Identificador de enteros correspondiente a 23 valores distintos, por ejemplo, Blue Cross/Blue Shield, Medicare, and self-pay	52%
Medical specialty	Nominal	Identificador de una especialidad del médico de admisión, que corresponde a 84 valores distintos, por ejemplo, cardiology, internal medicine, family/general practice, and surgeon	53%
Number of lab	Numérico		0%

procedures		Número de pruebas de laboratorio realizadas durante el encuentro.	
Number of procedures	Numérico	Número de procedimientos (distintos de las pruebas de laboratorio) realizados durante el encuentro	0%
Number of medications	Numérico	Número de nombres genéricos distintos administrados durante el encuentro.	0%
Number of outpatient visits	Numérico	Número de visitas ambulatorias del paciente en el año anterior al encuentro.	0%
Number of emergency visits	Numérico	Número de visitas de urgencia del paciente en el año anterior al encuentro.	0%
Number of inpatient visits	Numérico	Número de visitas hospitalarias del paciente en el año anterior al encuentro.	0%
Diagnosis 1	Nominal	El diagnóstico primario (codificado como los tres primeros dígitos de ICD9); 848 valores distintos	0%
Diagnosis 2	Nominal	Diagnóstico secundario (codificado como los primeros tres dígitos de ICD9); 923 valores distintos	0%
Diagnosis 3	Nominal	Diagnóstico secundario adicional (codificado como los tres primeros dígitos de ICD9); 954 valores distintos	1%
Number of diagnoses	Numérico	Número de diagnósticos ingresados al sistema.	0%
Glucose serum test result	Nominal	Indica el rango del resultado o si la prueba no fue tomada. Valores: "> 200", "> 300", "normal" y "none" si no se miden	0%
A1c test result	Nominal	Indica el rango del resultado o si la prueba no fue tomada. Valores: "> 8" si el resultado fue superior al 8%, "> 7" si el resultado fue superior al 7% pero menor al 8%, "normal" si el resultado fue inferior al 7%, y "none" si no es medido	0%
Change of medications	Nominal	Indica si hubo un cambio en los medicamentos para la diabetes (ya sea dosis o nombre genérico). Valores: "change" y "no change"	0%
Diabetes medications	Nominal	Indica si hubo algún medicamento para la diabetes prescrito. Valores: "yes" y "no"	0%
24 features for medications	Nominal	Para los nombres genéricos en inglés: metformin, repaglinide, nateglinide, chlorpropamide, glimepiride, acetohexamide, glipizide, glyburide, tolbutamide, pioglitazone, rosiglitazone, acarbose, miglitol, troglitazone, tolazamide, examide, sitagliptin, insulin, glyburide-metformin, glipizide-metformin, glimepiride-pioglitazone, metformin-rosiglitazone, and metformin-pioglitazone, la característica indica si el medicamento fue recetado o si hubo un cambio en la dosis. Valores: "up" si la dosis aumentó durante el encuentro, "down" si se redujo, "steady" si la dosis no cambió, and "no" si el medicamento no fue recetado	0%

Readmitted	Nominal	Días de reingreso hospitalario. Valores: "<30" si el paciente fue readmitido en menos de 30 días, "> 30" si fue en más de 30 días, y "No" para el registro de readmisión.	0%
------------	---------	---	----

ANEXO 18. Descripción id para Admission type. (University of California, Irvine, 2018)

Admission type id	Description
1	Emergency
2	Urgent
3	Elective
4	Newborn
5	Not Available
6	NULL
7	Trauma Center
8	Not Mapped

ANEXO 19. Descripción id para Discharge disposition. (University of California, Irvine, 2018)

Discharge disposition id	Description
1	Discharged to home
2	Discharged/transferred to another short term hospital
3	Discharged/transferred to SNF
4	Discharged/transferred to ICF
5	Discharged/transferred to another type of inpatient care institution
6	Discharged/transferred to home with home health service
7	Left AMA
8	Discharged/transferred to home under care of Home IV provider
9	Admitted as an inpatient to this hospital
10	Neonate discharged to another hospital for neonatal aftercare
11	Expired
12	Still patient or expected to return for outpatient services
13	Hospice / home
14	Hospice / medical facility
15	Discharged/transferred within this institution to Medicare approved swing bed
16	Discharged/transferred/referred another institution for outpatient services
17	Discharged/transferred/referred to this institution for outpatient services
18	NULL
19	Expired at home. Medicaid only, hospice.
20	Expired in a medical facility. Medicaid only, hospice.
21	Expired, place unknown. Medicaid only, hospice.
22	Discharged/transferred to another rehab fac including rehab units of a

	hospital .
23	Discharged/transferred to a long term care hospital.
24	Discharged/transferred to a nursing facility certified under Medicaid but not certified under Medicare.
25	Not Mapped
26	Unknown/Invalid
27	Discharged/transferred to a federal health care facility.
28	Discharged/transferred/referred to a psychiatric hospital of psychiatric distinct part unit of a hospital
29	Discharged/transferred to a Critical Access Hospital (CAH).
30	Discharged/transferred to another Type of Health Care Institution not Defined Elsewhere

ANEXO 20. Descripción id para Admission source. (University of California, Irvine, 2018)

Admission Source id	Description
1	Physician Referral
2	Clinic Referral
3	HMO Referral
4	Transfer from a hospital
5	Transfer from a Skilled Nursing Facility (SNF)
6	Transfer from another health care facility
7	Emergency Room
8	Court/Law Enforcement
9	Not Available
10	Transfer from critical access hospital
11	Normal Delivery
12	Premature Delivery
13	Sick Baby
14	Extramural Birth
15	Not Available
17	NULL
18	Transfer From Another Home Health Agency
19	Readmission to Same Home Health Agency
20	Not Mapped
21	Unknown/Invalid
22	Transfer from hospital inpt/same fac reslt in a sep claim
23	Born inside this hospital
24	Born outside this hospital
25	Transfer from Ambulatory Surgery Center
26	Transfer from Hospice

ANEXO 21. Codificación ICD9 para diagnósticos (Strack, y otros, 2014).

Enfermedad	Códigos ICD9
Circulatoria	390–459, 785
Respiratoria	460–519, 786
Digestiva	520–579, 787
Diabetes	250.xx
Herida	800–999
Musculoesqueletica	710–739
Genitourinaria	580–629, 788
Neoplasmas	140–239
Otra	780, 781, 784, 790–799 240–289, sin 250 680–709, 782 001–139, 290–319 E–V 320–389 630–679 740–759