MÁSTER EN BIG DATA Y ANALYTICS

Fraude en empresas de seguro

Entrega Final– 18/10/2022

GRUPO N°4

Autores:

Alejandro Alfageme

Hambert Cuba Del Carpio

Jimmy Javier Carhuancho

Salvador Chorro

THREE POINTS

THE DIGITAL BUSINESS SCHOOL

Fraude en empresas de Seguro

GRUPO N° 4

Autores:

- Alejandro Alfageme
- Hambert Cuba Del Carpio
- Jimmy Javier Carhuancho
- Salvador Chorro

Tutor Académico: David Monter

ENTREGA FINAL: 18/10/2022

RESUMEN EJECUTIVO

Según datos del INESE, compañía que agrupa profesionales del sector asegurador, "detrás del 1,60% de todos los expedientes que tramitaron las aseguradoras en 2021 había un fraude. Esta cifra es superior a la anotada en 2020 (1,18%) y, también, a los datos de 2019 (1,38%)[1]" (INESE, 2 de junio 2022), esto nos permite identificar un panorama en el cual el rubro de aseguradoras está siendo cada vez más afectado, lo cual impacta no solo la rentabilidad de las empresas y la economía, sino a la sociedad en general que busca los productos que estas ofrecen, ya que al incrementar estos casos de fraude los costos de las pólizas deben ser ajustados para sopesar dicho impacto.

Teniendo a bien esta situación donde este rubro de la economía presenta esta necesidad y con base en los conocimientos adquiridos en el desarrollo del Máster en Big Data y Analytics, proponemos mostrar que por medio de la aplicación de modelos de Machine Learning es posible identificar de forma oportuna casos de potencial fraude, lo cual permitiría que las empresas puedan ser proactivas en la identificación de "red flags" al momento en que se presentan las reclamaciones. Esto permitiría que indirectamente se proteja el interés de la sociedad en general a raíz de que las pólizas podrían mantenerse en costos razonables.

El proyecto desarrollado está enfocado en la construcción de un modelo predictivo supervisado, el cual permite identificar de forma proactiva un potencial fraude a partir de una fuente de información que contiene datos de pacientes ambulatorios, proveedores, reclamaciones, monto de reembolso solicitado, médicos tratantes, códigos de diagnóstico, historial de fraudes, entre otros. Como fuente se ha utilizado un Dataset de dominio público por medio de KAGGLE[2].

En el documento se proporciona el diseño e implementación de un mínimo producto viable del modelo predictivo, este podrá ser utilizado como referencia en caso se desee probar con información interna. Se anexan scripts de código, diagrama de arquitectura y proceso de extracción, transformación y carga de información, simulación y dashboard para mostrar su resultado.

Este trabajo toma mucha relevancia ya que está enfocado en la generación de valor de forma disruptiva en las compañías de seguro, a raíz de que la identificación proactiva de un potencial fraude se vuelve en un diferenciador y marca un hito importante en la ventaja competitiva para aquellas empresas que lo implementan. Es de recordar que información sobre este tema no se encuentra fácilmente o de acceso a dominio público, por lo cual esperamos abrir un sinfín de oportunidades para quienes se interesen en su aplicación.

[1] INESE. (2 de junio 2022). *Crecen los intentos de fraude al seguro y su coste económico*. Obtenido de https://www.inese.es/crecen-los-intentos-de-fraude-al-seguro-y-su-coste-economico/#:~:text=Detr%C3%A1s%20del%201%2C60%25%20de,2019%20(1%2C38%25).

[2] KAGGLE. (s.f.). Obtenido de https://www.kaggle.com/

INDICE

ILUSTRACIONES

TABLAS

INTRODUCCIÓN

Con base en los conocimientos adquiridos durante el Máster en Big Data y Analytics y de la experiencia en el sector de Seguros, se logró detectar la oportunidad que existía para volver el proceso legacy de identificación de fraudes que es reactivo, y que algunas empresas tienen de cara al manejo de las reclamaciones, a uno proactivo donde por medio de la aplicación de modelos de Machine Learning de forma anticipada se podrían detectar casos potenciales de fraude.

El producto ofrecido como salida de este proyecto permitiría proporcionar inteligencia y agilidad a las empresas en el tratamiento de reclamaciones e identificación de posibles fraudes, lo cual le habilita enfocar sus recursos en la atención de las solicitudes con justa causa, orientar el análisis riguroso de los casos potenciales y disminuir el desperdicio en costos y procesos operativos de la compañía. Esto aunado al beneficio intrínseco a la sociedad en general, ya que podría mantener su capacidad de adquirir protección ante riesgos eventuales respecto a su salud sin que esto afecte su capacidad de pago.

El proyecto utiliza como insumo el Dataset público "HEALTHCARE PROVIDER FRAUD DETECTION ANALYSIS" obtenido de KAGGLE[3], se presenta el contexto de la problemática, objetivos del proyecto, la propuesta de valor, viabilidad, planificación, diseño de la solución, e implementación de un producto mínimo viable funcional el cual podrá ser utilizado como referencia a los interesados.

El modelo de machine learning principalmente se enfoca en los casos de fraude asociados a pacientes ambulatorios los cuales no quedan internados al momento de la atención médica.

El resultado final del proyecto consiste en un modelo de machine learning el cual con base a una serie de datos input asociados a pacientes ambulatorios, reclamaciones, diagnósticos y otros, logra identificar los potenciales fraudes de forma predictiva.

[3] KAGGLE. (s.f.). *HEALTHCARE PROVIDER FRAUD DETECTION ANALYSIS*. Obtenido de https://www.kaggle.com/datasets/rohitrox/healthcare-provider-fraud-detection-analysis

1. DEFINICION DEL PROYECTO

Contexto y alcance

El fraude a coberturas asociadas a salud representa uno de los retos más importante tanto para las aseguradoras como para los sistemas de salud en general, a raíz de que cada suceso impacta negativamente en la sociedad ya que afecta no solo a las empresas sino que repercute en los costos y rigurosidad en el manejo de riesgos lo cual limita el acceso a las personas en general, ya que el riesgo y costes siempre son transferidos a las primas asociadas al servicio contratado por la sociedad en general (empresas y personas naturales).

Según datos proporcionados por Johns Hopkins HealthCare LLC, se estima que anualmente se pierden cerca de 60 mil millones de dólares debido al fraude y abuso en la atención médica[4] y la Comisión de Sentencias de Estados Unidos (United States Sentencing Commission - USSC)[5], indica que a pesar de que el número de fraudes ha venido decreciendo desde el año 2016 el costo medio en el año 2020 fue de $1,250,00 millones de dólares teniendo como principales estados afectados el de Florida y California.

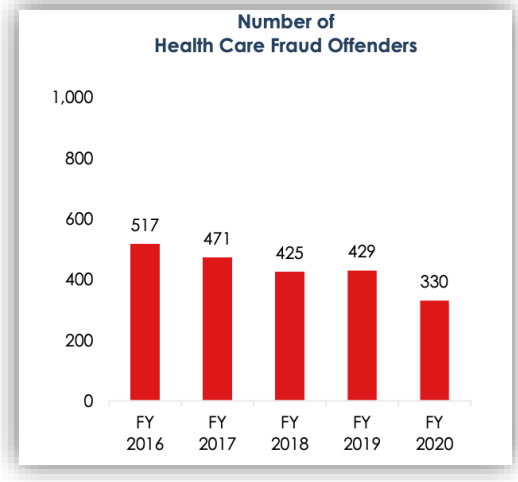


Ilustración 2 - USSC Total de fraudes 2016 - 2020

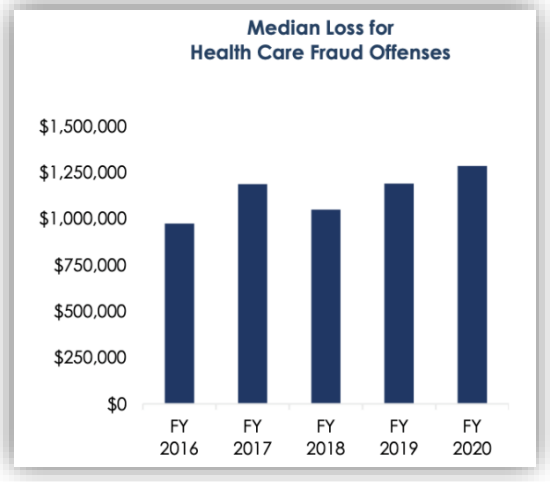


Ilustración 1 - USSC Costo medio de fraudes 2016 - 2020

[4] Johns Hopkins HealthCare LLC. (s.f.). *Health Care Fraud and Abuse*. Obtenido de https://www.hopkinsmedicine.org/johns_hopkins_healthcare/providers_physicians/health_care_fraud_and_abuse/

[5] United States Sentencing Commission. (s.f.). *HEALTH CARE FRAUD*. Obtenido de https://www.ussc.gov/research/quick-facts/health-carefraud#:~:text=(July%202021)%20In%20fiscal%20year,36.2%25%20since%20fiscal%20year%202016.&text=The%20USSC%20HelpLine%20assists%20practitioners%20in%20applying%20the%20guidelines

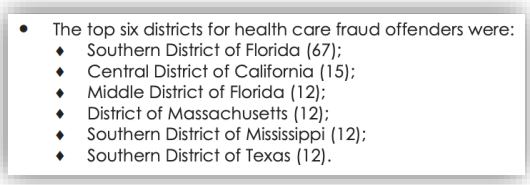


Ilustración 3 - USSC Top 6 distritos afectados

A raíz de que esta actividad puede ser llevada a cabo de forma individual o por medio de grupos organizados, buscando un lucro ilegitimo, aumenta la dificultad para su detección, sin embargo, por medio de la colaboración entre instituciones, uso de medios digitales, datos de clientes, transacciones y con el apoyo de nuevas tecnologías, es posible la creación de modelos que permitan su detección y prevención, lo cual habilita la posibilidad de actuar de forma anticipada y no solo reactiva como sucedía en el pasado.

Según la oficina de la fiscalía general de Texas[6] los fraudes pueden tomar muchas formas, sin embargo, las más comunes son:

- Fraude de facturación médica de seguros de salud, el cuál ocurre cuando un proveedor de atención médica o una persona engaña a una aseguradora para recibir un reembolso mayor
- Fraude de Medicare y Madicaid, es un fraude de facturación médica asociados a programas de beneficios del gobierno y se presenta cuando un proveedor de atención médica reclama reembolsos de estos programas a los que no tiene derecho para cobrar dinero de manera deshonesta
- Fraude de atención médica en el hogar, se da cuando las agencias de atención médica domiciliaria facturan a las aseguradoras, a los programas de beneficios del gobierno o a los pacientes confinados en el hogar por servicios innecesarios o por servicios que nunca se prestaron
- Fraude y abuso de drogas, dentro de este hay algunos fraudes y abuso de medicamentos como lo son: fraude de precios de medicamentos, fraude de medicamentos falsificados y abuso de desvío de medicamentos

Dada la diversidad de fraudes, el impacto que estos tienen en la sociedad y con base en aplicaciones de modelos de analítica desarrollados en Estados Unidos, los cuales por medio de análisis de datos se ha logrado pesquisar médicos que cometen fraude en el sistema de salud en dicho país, identificamos que se puede extrapolar el ejercicio a otros sistemas de salud del mundo, por lo que se plantea este proyecto de investigación el cuál sirva como guía base en dicho sentido. A

[6] KEN PAXTON ATTORNEY GENERAL of TEXAS. (s.f.). *Health Care Fraud and Abuse*. Obtenido de https://www.texasattorneygeneral.gov/consumer-protection/health-care/health-care-fraud-and-abuse#medicare

pesar de ello, los datos analizados en este trabajo corresponden a un dataset de carácter público de prestaciones de salud dentro del Sistema Nacional de Salud de Estados Unidos, país del cuál se facilita acceso a datos de este tipo.

Con esta investigación buscamos proponer un modelo que tome como base las mejores prácticas disponibles, alcance un nivel aceptable de predicción de riesgo de fraude para atenciones ambulatorias y que, al mismo tiempo, permita identificar la factibilidad de aplicar el mismo acercamiento para otros datasets independiente del sistema de salud del que provengan los datos.

2. OBJETIVOS

El objetivo principal para las empresas de seguros de salud es mitigar el fraude de manera oportuna y eficiente, a través de la aplicación en su flujo operativo de un modelo confiable de predicción de riesgo de fraude.

Esperamos que los hallazgos de esta investigación puedan ser aplicado a cualquier compañía aseguradora de manera costo eficiente, dentro del caso específico de riesgo de fraude de seguros de salud para prestaciones ambulatorias.

Para ello dentro de los objetivos específicos del proyecto se plantean los siguientes:

- Realizar una investigación bibliográfica respecto de las mejores prácticas y tendencias utilizadas a nivel mundial para enfrentar el problema de fraude en seguros de salud por medio de análisis de datos
- Contar con un modelo de predicción que tenga un nivel de acierto aceptable para un data set publico específico de registros de solicitudes de pago de seguros de salud dentro del sistema de salud pública de Estados Unidos (Medicare)
- Investigar la aplicabilidad de la metodología utilizada en otros escenarios y sistemas de salud por medio de investigación bibliográfica
- Establecer una metodología y proporcionar recomendaciones necesarias para enfrentar el diseño e implementación de un modelo predictivo de riesgo de fraude en el escenario de aseguradoras de salud.

Propuesta de valor

Hoy en día la analítica y el uso adecuado de los datos son una ventaja competitiva que muchas industrias han aprovechado y, tomando como base que la problemática de fraudes a seguros asociados al sector de salud es general y aplica a nivel mundial, se abre la posibilidad de potenciar el uso de estas herramientas las cuales permitan optimizar la gestión de riesgos y proveer un impacto positivo tanto a las empresas del sector, como a los usuarios, ya que estos se podrían beneficiar de precios más competitivos y accesibles.

La industria de salud es sin duda un sector que tiene un gran impacto económico y social en el mundo entero, y la contribución de la investigación bibliográfica y su aplicación a un modelo, así como la sistematización de las tendencias metodológicas y principales recomendaciones, esperamos sea de utilidad para contribuir a enfrentar esta problemática que afecta a muchas empresas de salud y finalmente impacta los planes de salud subiendo innecesariamente los costos asociados a la salud para la población en su conjunto, dando una ventaja

fundamental frente a empresas que no adopten el análisis de datos para enfrentar el fraude.

Como equipo contamos con una combinación de habilidades técnicas y de negocio que permiten tener una visión realista sobre la implementación, y real utilidad, así como de la eficiencia y oportunidad que entregarán nuestras recomendaciones metodológicas a la hora de implementar un modelo predictivo en cualquier aseguradora de salud.

3. VIABILIDAD DEL PROYECTO

Técnica

Se ha conseguido información con granularidad, volumen y calidad adecuada para el análisis que se pretende realizar, ya que contamos con un data set público, ampliamente utilizado y comentado para investigar el caso específico de seguros en salud. Este Dataset ha sido obtenido por medio de KAGGLE[7].

El objetivo por modelar está adecuadamente especificado y para el desarrollo se aplicarán diversas técnicas de modelamiento buscando obtener el resultado más eficiente y estable posible.

Respecto al Dataset, se especifica que se contemplan reclamaciones de pacientes que han sido hospitalizados, pacientes ambulatorios (es decir no requieren hospitalización) y detalles del beneficiario de cada proveedor, a continuación, se aplica un poco más sobre las características del Dataset:

- Pacientes que si han sido hospitalizados
    - Se brinda información sobre las reclamaciones presentadas por pacientes que son ingresados en los hospitales, también se proporciona detalles adicionales como sus fechas de admisión, alta, código de diagnóstico de admisión, etc.
- Datos de pacientes que no requieren hospitalización (ambulatorios)
    - Se brinda información sobre las reclamaciones presentadas por pacientes que se presentan a los hospitales y no requieren hospitalización, contiene datos de fechas del reclamo, montos, deducibles, código del procedimiento aplicado, etc.
- Datos de beneficiario
    - Estos contienen detalles del proceso de Know Your Customer (KYC) del beneficiario, como condiciones de salud, región a la que pertenece, fecha de nacimiento, fecha de defunción, género, etc.

Económica

La inversión en tiempo, contratación, infraestructura y posterior monitoreo justifica el ahorro en gasto proveniente de fraudes del negocio.

[7] KAGGLE. (s.f.). *HEALTHCARE PROVIDER FRAUD DETECTION ANALYSIS*. Obtenido de https://www.kaggle.com/datasets/rohitrox/healthcare-provider-fraud-detection-analysis

Es clave en este sentido las recomendaciones que levantemos tanto en relación al ajuste del modelo a una empresa particular, y a un sistema de salud en particular, como también respecto de la implementación de este dentro del flujo operativo de la empresa, que permita entregar un scoring de riesgo de manera oportuna, prediciendo un riesgo de fraude previo al desembolso de la compañía pero sin perjudicar la calidad del servicio de estas instituciones, que hoy por hoy se miden principalmente por la velocidad de pago a sus beneficiarios.

Nuestras estimaciones se basan en el hecho de que entre un 1% y un 3% de los ingresos se pierden por fraude. Por lo tanto, en una empresa de seguros de tamaño mediano que tiene una facturación de USD 500 millones anuales, estamos hablando de una recuperación de USD 5 millones al año. Un proyecto piloto con alcance acotado es posible construirlo bajo unos costos menores a EUR 25,000 y por lo tanto venderlo a un precio competitivo con cierta facilidad, dado que la generación de valor económico que este puede producir es alta, sin embargo, esta estimación la iremos madurando en el transcurso de la ejecución del proyecto a raíz de mejorar la estimación inicial.

Conclusión de viabilidad

Las buenas prácticas y casos de éxito en casos similares dan indicios de la viabilidad técnica y económica del desarrollo de un modelo de prevención de fraude en empresas de seguros de salud. Así mismo, creemos que el tiempo definido para este proyecto es suficiente para recoger las mejores prácticas y lograr sistematizar una serie de recomendaciones de diseño e implementación, que permitiría adaptar a una empresa de seguros en particular el modelo y la experiencia adquirida de manera eficaz.

4. PLANIFICACION

La creación de modelos de predicción de fraude toma como base el apoyo entre de un equipo multidisciplinario de roles, los cuales cada uno desde sus habilidades y experiencia colaboran para cumplir el objetivo planteado.

Dentro de los roles que participan en este tipo de proyectos se presentan:

- **Data Architect**: Encargado de la estrategia de arquitectura de Machine Learning y mantienen en sintonía a los científicos de datos, ingenieros de datos, analistas, equipo de operaciones y negocio para gobernar y escalar las iniciativas de Inteligencia Artificial[8]
- **Data Scientist**: Encargado de la extracción de insights de los activos de información y de las iniciativas de Big Data[9]
- **Data Analyst**: Encargados de convertir las preguntas de negocio en consultas a datos que permitan identificar como la operación del negocio, los objetivos estratégicos y sus datos se unen para generar valor, utilizan los datos y comunican sus hallazgos para el proceso de toma de decisiones[10]
- **Data Engineer**: Responsables de realizar la extracción de datos de los sistemas transaccionales y dejarlos disponibles en la plataforma analítica para que estos puedan ser explotados por los equipos de Data Analyst y Data Scientist[11]
- **Project Manager / Scrum Master**: Responsable de la efectividad del equipo desempeñando actividades como coaching, enfoque en valor, remover impedimentos y productividad[12]
- **Product Owner**: Responsable de la maximización del valor del producto a ser construido[13]

---

[8] GARTNER, Ashutosh Gupta. (19 de enero 2022). *What Are AI Architects and What Do They Do?*. Obtenido de https://www.gartner.com/en/articles/what-are-ai-architects-and-what-do-they-do

[9] GARTNER. (s.f.). Gartner Glossary - Data Scientist. Obtenido de https://www.gartner.com/en/information-technology/glossary/data-scientist#:~:text=The%20data%20scientist%20role%20is,fulfilled%20better%20as%20a%20team

[10] Future Learn. (s.f.). *Roles and Responsibilities in Data-Driven Organizations*. Obtenido de https://www.futurelearn.com/info/courses/financial-analysis-business-performance-data-driven-decision-making-sc/0/steps/177018

[11] Future Learn. (s.f.). *Roles and Responsibilities in Data-Driven Organizations*. Obtenido de https://www.futurelearn.com/info/courses/financial-analysis-business-performance-data-driven-decision-making-sc/0/steps/177018

[12] SCRUM.ORG. (s.f.). *What is a Scrum Master?*. Obtenido de https://www.scrum.org/resources/what-is-a-scrum-master

[13] SCRUM.ORG. (s.f.). *What is a Product Owner?*. Obtenido de https://www.scrum.org/resources/what-is-a-product-owner

Roles y tarifas por jornada

A continuación, se detallan los roles involucrados en la creación del modelo predictivo junto con su costo estimado por jornada:

Rol	Tarifa jornada	Tarifa hora
Data Architect	€ 360.00	€ 45.00
Data Scientist	€ 280.00	€ 35.00
Data Analyst	€ 280.00	€ 35.00
Data Engineer	€ 320.00	€ 40.00
Product Owner	€ 280.00	€ 35.00
Project Manager	€ 400.00	€ 50.00

Tabla 1 - Roles y tarifas por jornada

Tareas de planificación

Dentro de las actividades generales a ser aplicadas al momento de la construcción de un modelo predictivo, se contemplan como mínimo las que se especifican en la Tabla 2 - Actividades para construcción de Modelo de ML, de las cuales, dependiendo de las características de las fuentes de información, y el estado de la calidad de estas, se puede impactar positiva o negativamente al esfuerzo requerido para llevarlas a cabo:

Fase	Rol	Tarea	Detalle	Días	Importe	Fecha inicio
Requerimientos	Product Owner	Definición de requerimientos	Definición de requerimientos funcionales y no funcionales del modelo	2	€ 560.00	1/8/2022
Arquitectura	Data Architect	Definición de arquitectura	Definición de componentes	2	€ 720.00	3/8/2022
Arquitectura	Data Architect	Configuración de entorno y componentes	Configuración de componentes, máquinas virtuales, comunicación, seguridad y accesos	2	€ 720.00	5/8/2022
Construcción	Data Engineer	Construcción de ETL	Creación de Ingesta, limpieza inicial y transformación de datos, capas de datos (RAW, BRONZE)	5	€ 1,600.00	9/8/2022
Construcción	Data Engineer	Creación de capa Analítica	Capa GOLD	2	€ 640.00	16/8/2022
Modelo ML	Data Scientist	Exploratory Data Analysis (EDA)	Ejecución del análisis exploratorio de datos para identificación de reglas de limpieza a ser aplicados y selección de variables	2	€ 560.00	18/8/2022
Modelo ML	Data Scientist	Construcción de modelo predictivo	Construcción de modelo	15	€ 4,200.00	22/8/2022
Modelo ML	Data Scientist	Presentación de resultados	Preparación y presentación de insights	5	€ 1,400.00	12/9/2022
Implementación	Data Engineer	Inclusión de modelo en capa analítica	Implementación de modelo en capa analítica	2	€ 640.00	19/9/2022

Análisis (Visualización)	Data Analyst	Construcción de mockup	Diseño de análisis	3	€ 840.00	21/9/2022
Análisis (Visualización)	Data Analyst	Construcción de tablero	Construcción de visualizaciones, publicación y acceso	7	€ 1,960.00	26/9/2022

Tabla 2 - Actividades para construcción de Modelo de ML

Calendario

A continuación, se muestra un diagrama de Gantt de los tiempos estimados que puede llevar la ejecución de un proyecto de ML para identificación de fraudes.

Ilustración 4 - Gantt de actividades para construcción de Modelo de ML

Valoración económica

Tomando como base los roles y la estimación inicial de esfuerzo para llevar a cabo un proyecto de este tipo, a continuación, se muestra la valoración económica tomando como insumo inicialmente solo el costo de los roles involucrados, se deja por fuera los componentes/aplicaciones/servicios utilizados dado que estos dependerán de la solución de arquitectura que se utilice como referencia.

Grupo de actividades	Jornadas	Valoración
Modelo ML	**22**	**€ 6,160.00**
Exploratory Data Analysis (EDA)	2	€ 560.00
Construcción de modelo predictivo	15	€ 4,200.00
Presentación de resultados	5	€ 1,400.00
Análisis (Visualización)	**10**	**€ 2,800.00**
Construcción de mockup	3	€ 840.00
Construcción de tablero	7	€ 1,960.00
Construcción	**7**	**€ 2,240.00**
Construcción de ETL	5	€ 1,600.00
Creación de capa Analítica	2	€ 640.00
Arquitectura	**4**	**€ 1,440.00**
Definición de arquitectura	2	€ 720.00
Configuración de entorno y componentes	2	€ 720.00
Implementación	**2**	**€ 640.00**
Inclusión de modelo en capa analítica	2	€ 640.00
Requerimientos	**2**	**€ 560.00**
Definición de requerimientos	2	€ 560.00
Grand Total	**47**	**€ 13,840.00**

Tabla 3 - Valoración económica

Ahora, si tomamos como insumo preliminar la contratación del servicio del componente de Databricks en Microsoft Azure para realizar una prueba de concepto utilizando únicamente este componente, a continuación, se detalla el costo estimado[14] del consumo, tomando como premisa que se mantendrá como herramienta oficial dentro del ciclo de vida de las soluciones que se construyan:

- Este componente es un servicio en la nube que puede ser utilizado en nubes públicas como Amazon Web Services, Microsoft Azure y Google Cloud
- El servicio esta optimizado para la construcción y ejecución de ETLs y Modelos de Machine Learning (ML) utilizando tecnologías de BIG Data

[14] Para más información utilizar Azure Databricks Pricing: *https://azure.microsoft.com/en-us/pricing/details/databricks/*

- Costo estimado del servicio realizando la compra anticipada y comprometida de 1 y 3 años a junio 2020, incluyendo las siguientes características:
    - Workload: All-Purpose Compute
    - Tier: Standar
    - Región: North Europe
    - Currency: EUR

Plan de pre – compra	Databricks commit unit	Precio anual
1 año	25,000	€22,350.088
3 año	75,000	€65,623.663

Tabla 4 - Estimación de costos de componente Databricks

Sin embargo, en caso se requiera contemplar componentes de Data Factory (flujos de movimientos de datos de sistemas fuentes a Data Lake y ejecución de notebooks de Databricks para posteriormente almacenarlos en una Base de Datos), Storage, Base de Datos y Databricks (contemplando características básicas), se muestra a continuación un ejercicio de estimación de componentes[15] tomando como base región North Europe con requisitos mínimos para llevar a cabo una prueba de concepto, el cual puede variar en precios según proveedor de nube y características de componentes:

Service category	Service type	Description	Estimated monthly cost	Estimated upfront cost
Databases	Azure SQL Database	Single Database, vCore, RA-GRS Backup Storage, General Purpose, Serverless, Standard-series (Gen 5), Local Redundancy, 16 Billed vCores, 250 GB Storage, 0 GB Backup Storage	€ 40.14	€ 0.00
Storage	Storage Accounts	Data Lake Storage Gen2, Standard, LRS Redundancy, Hot Access Tier, Hierarchical Namespace File Structure, 1,000 GB Capacity - Pay as you go, Write operations: 4 MB x 10 operations, Read operations: 4 MB x 10 operations, 10 Iterative read operations, 10 Iterative write operations, 10 Other operations. 1,000 GB Data Retrieval, 1,000 GB Data Write, 1,000 GB Meta-data storage	€ 55.45	€ 0.00

[15] Información obtenida a junio 2022 de la calculadora de precios de Azure: *https://azure.microsoft.com/en-us/pricing/calculator/*

Analytics	Azure Databricks	All-Purpose Compute Workload, Standard Tier, 1 D8SV3 (8 vCPU(s), 32 GB RAM), 1 year reserved, 1.5 DBU x 730 Hours	€ 438.00	€ 2,457.00
Analytics	Azure Data Factory	Azure Data Factory V2 Type, Data Pipeline Service Type, Azure Integration Runtime: 1 Activity Run(s), 20 Data movement unit(s), 20 Pipeline activities, 0 Pipeline activities – External, Self-hosted Integration Runtime: 10 Activity Run(s), 10 Data movement unit(s), 10 Pipeline activities, 10 Pipeline activities – External,	€ 22.12	€ 0.00
Support	Support		€ 100	€ 0.00
	Licensing Program	Microsoft Customer Agreement (MCA)	N/A	N/A
	Total		€ 655.71	€ 2,457.00

Tabla 5 - Estimación de costos de componentes para prueba de concepto

5. INVESTIGACION SOBRE MODELOS

Con el objetivo de asegurar que si existen modelos / librerías que permitan llevar a cabo el objetivo del proyecto, se realizó una investigación sobre modelos disponibles de tal forma de asegurar el éxito en la ejecución del trabajo. Para esto se investigó varios modelos de Machine Learning (LM) y Deep Learning (DL) para construir un modelo eficiente de detección de fraude en seguros de salud. Los desafíos clave que se identifican en la construcción de un modelo eficiente de detección de fraude incluyen:

- Desequilibrio de datos: número sesgado de casos fraudulentos menores en comparación con los casos no fraudulentos.
- Selección del modelo de clasificación: uso de modelos apropiados de ML) y DL para identificar casos de fraude o no fraude.

Como parte de los hallazgos se identifica que como metodología para abordar estos desafíos se puede llevar a cabo las siguientes acciones:

- Desbalanceo: Existen varias técnicas para abordar el desbalanceo de datos utilizando tres técnicas de "sobre muestreo", o creación de datos sintéticos entre ellas están SMOTE, ADASYN y TGAN. Aquí, las muestras se generan sintéticamente para equilibrar las dos clases en el conjunto de datos (Fraude / No fraude).
- Modelo de clasificación: Existen varios modelos de clasificación utilizados, como Decision Tree (DT), Random Forest (RF), XGBoost, LightGBM y Gradient Boosting Machine (GBM) utilizados para este caso de negocio.

Nuestra metodología mezcla investigación bibliográfica para encontrar las mejores prácticas para abordar el problema de negocio, en el entendido que un mismo problema implica una similitud importante en las características de los datos.

En el trabajo realizado por Rohan Yashraj y otros[16] realizan un estudio comparativo de distintos modelos de ML y DL para detección de fraude en el contexto de seguros de salud. Los indicadores y resultados obtenidos en este estudio son los siguientes:

[16] ResearchGate, Rohan Yashraj Gupta, Pallav Kumar Baruah, Satya Sai Mudigonda (marzo 2021). *A Comparative Study of Using Various Machine Learning and Deep Learning-Based Fraud Detection Models For Universal Health Coverage Schemes*. Obtenido de https://www.researchgate.net/profile/Rohan-Gupta-5/publication/350132738_A_Comparative_Study_of_Using_Various_Machine_Learning_and_Deep_Learning-Based_Fraud_Detection_Models_For_Universal_Health_Coverage_Schemes/links/605d6e21a6fdccbfea085b79/A-Comparative-Study-of-Using-Various-Machine-Learning-and-Deep-Learning-Based-Fraud-Detection-Models-For-Universal-Health-Coverage-Schemes.pdf

Models			AUC-ROC	Recall	Specificity	Precision	Accuracy	F1 Score
Decision Tree	Baseline	M1	**0.9566**	0.9248	0.9885	0.9174	0.9808	0.9211
	SMOTE	M2	0.9534	0.9208	0.9860	0.9006	0.9781	0.9106
	ADASYN	M3	0.9508	0.9155	0.9862	0.9016	0.9776	0.9085
	TGANs	M4	0.9548	0.9214	0.9883	0.9155	0.9801	0.9185
Random Forest	Baseline	M5	0.9462	0.8947	0.9977	0.9818	0.9852	0.9362
	SMOTE	M6	0.9493	0.9027	0.9959	0.9682	0.9846	0.9343
	ADASYN	M7	0.9500	0.9057	0.9942	0.9556	0.9834	0.9300
	TGANs	M8	0.9460	0.8942	0.9977	0.9820	0.9852	0.9361
XGBoost	Baseline	M9	0.9307	0.8615	0.9999	0.9989	0.9831	0.9252
	SMOTE	M10	0.9458	0.8970	0.9945	0.9572	0.9826	0.9262
	ADASYN	M11	0.9270	**0.9835**	0.8705	0.5119	0.8842	0.6733
	TGANs	M12	0.9111	0.8223	**1.0000**	**1.0000**	0.9784	0.9025
LightGBM	Baseline	M13	0.9486	0.8977	0.9994	0.9952	**0.9871**	**0.9440**
	SMOTE	M14	0.9499	0.9014	0.9988	0.9905	0.9869	0.9438
	ADASYN	M15	0.9523	0.9105	0.9940	0.9547	0.9839	0.9320
	TGANs	M16	0.9482	0.8970	0.9994	0.9950	0.9870	0.9435
GBM	Baseline	M17	0.9425	0.8852	0.9997	0.9975	0.9858	0.9380
	SMOTE	M18	0.9451	0.8958	0.9945	0.9576	0.9825	0.9257
	ADASYN	M19	0.9288	0.9779	0.8796	0.5288	0.8916	0.6864
	TGANs	M20	0.9282	0.8566	0.9992	0.9992	0.9224	0.9224

Ilustración 5 - Resultado de evaluación de modelos ML y DL

Pudimos identificar que estos modelos se aplican en el contexto de fraude a nivel de cliente, mientras los datos que utilizaremos nosotros apuntan a fraude de proveedor. Los registros en ambos casos consisten en solicitudes de reembolso registrados por los sistemas. Para explicar un poco las diferencias, el fraude de proveedores de atención médica puede ser cometido por individuos (p. ej., médicos, doctores) o por organizaciones (p. ej., hospitales). A veces, el fraude del proveedor también puede involucrar a otros proveedores de servicios o personas (por ejemplo, pacientes). El fraude del cliente puede cometerse cuando el asegurado/consumidor tergiversa los hechos a con el propósito de obtener beneficios adicionales. Pueden trabajar en unión con proveedores de atención médica (p. ej., médicos).

6. DESARROLLO DEL PROYECTO

Diagrama de arquitectura

A continuación, se detalla el diagrama de la arquitectura que será utilizada para llevar a cabo el desarrollo proyecto, en esta se especifican las fuentes de información que se utilizaran como input, el componente seleccionado en el cual se realizará la integración, almacenamiento, acceso a datos, analítica y los roles involucrados.

Con el objetivo de simplificar la arquitectura de componentes y a raíz de que Databricks es una herramienta altamente valorada y robusta para el tratamiento de información, hemos decidido utilizarlo como pieza única para ejecutar todo el ciclo de creación de ETL y Modelo correspondiente, ya que provee todas las funcionalidades requeridas.

Si a futuro una empresa desea realizar un MVP utilizando esta arquitectura se recomienda que este se alimente con información proveniente de sus transaccionales como el CRM o base de datos de reclamaciones, con el objetivo de mostrar información relevante para sus análisis.

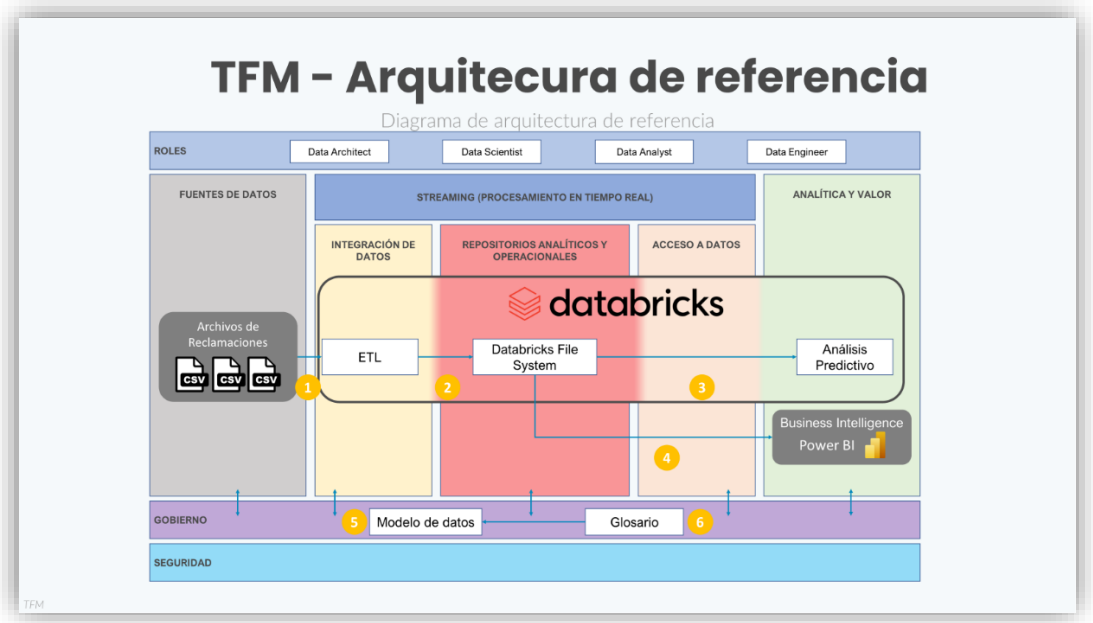


Ilustración 6 - Diagrama de Arquitectura

A continuación, se detallan los componentes de la arquitectura:

1. La función del componente de "**ETL**" es Extraer, por medio de un proceso batch, la información proveniente de la fuente de datos estructurada "**Archivos de Reclamaciones**", esto con el objetivo de Transformar los datos y Cargarlos en el repositorio analítico de data estructurada "**Databricks File System**", el cual sirve como repositorio de las tablas parquet. Este ETL crea las distintas capas de datos que posteriormente podrán servir de insumo para distintos objetivos de negocio, las capas que se crean son RAW, BRONZE y GOLD.

2. La función del repositorio "**Databricks File System**" es el almacenamiento estructurado de información de la fuente de datos "**Archivos de Reclamaciones**", así como también, de proveer las distintas capas de datos (RAW, BRONZE y GOLD) las cuales facilitan la explotación de información oficial y de calidad para los procesos analíticos que se desarrollarán.

3. La función del componente "**Análisis Predictivo**" es proveer la funcionalidad al "**Científico de Datos**" de configurar, construir y ejecutar modelos predictivos que potencien la generación de valor al negocio explotando la información disponible en el repositorio "**Databricks File System**". Este componente se desarrolla sobre la base de Notebooks ejecutándose dentro del Clúster de Databricks.

4. La función del componente de "**Business Intelligence**" es proveer la funcionalidad de autoservicio para el "**Analista de Negocio**", así como, facilitar la generación de analítica descriptiva y reporting tradicional a raíz del consumo del repositorio analítico construido por el "**Ingeniero de Datos**" en el "**Databricks File System**". La herramienta de visualización que se estará utilizando es Power BI.

5. La función del "**Modelo de Datos"** es proveer la funcionalidad de gestión del Inventario de Metadatos y el Modelamiento de datos respectivo, con el objetivo de proveer una capa semántica que facilite el consumo de la información a los diferentes roles que la utilizan/consultan/explotan.

6. La función del "**Glosario**" es servir como base de conocimiento respecto a los términos de negocio y su relación con las variables técnicas descritas en los Modelos de Datos, estos facilitarán el consumo de información a los diferentes roles establecidos en la arquitectura. El "**Modelo de datos**" provee la funcionalidad de gestión del Inventario de Metadatos y el Modelamiento de Datos respectivo, con el objetivo de proveer una capa semántica que facilite el consumo de la información a los diferentes roles que la utilizan / consultan / explotan.

Diccionario de datos

Para el proyecto se identifican tres datasets los cuales se describen a continuación:

- **Inpatientdata (Train_Inpatientdata.csv):** Pacientes que han sido hospitalizados, de los cuales a continuación se detalla los datos que son recolectados.

Campo	Descripción	Tipo de dato	Longitud	Obligatorio
BeneID	Id del beneficiario	Alfanumérico	20	Si
ClaimID	ID de solicitud	Alfanumérico	20	Si
ClaimStartDt	Fecha inicio de solicitud	Fecha		Si
ClaimEndDt	Fecha fin de solicitud	Fecha		Si
Provider	ID de proveedor	Alfanumérico	20	Si
InscClaimAmtReimbursed	Monto de solicitud reembolsado	Decimal	10,6	Si
AttendingPhysician	Médico tratante	Alfanumérico	20	Si
OperatingPhysician	Médico operador	Alfanumérico	20	No
OtherPhysician	Otro médico	Alfanumérico	20	No
AdmissionDt	Fecha de admisión	Fecha		Si
ClmAdmitDiagnosisCode	Código de diagnóstico admitido	Alfanumérico	20	Si
DeductibleAmtPaid	Monto deducible pagado	Decimal	10,6	Si
DischargeDt	Fecha de pago	Fecha		Si
DiagnosisGroupCode	Código de grupo diagnosticado	Alfanumérico	20	No
ClmDiagnosisCode_1	Código de diagnóstico 1	Alfanumérico	20	No
ClmDiagnosisCode_2	Código de diagnóstico 2	Alfanumérico	20	No
ClmDiagnosisCode_3	Código de diagnóstico 3	Alfanumérico	20	No
ClmDiagnosisCode_4	Código de diagnóstico 4	Alfanumérico	20	No
ClmDiagnosisCode_5	Código de diagnóstico 5	Alfanumérico	20	No
ClmDiagnosisCode_6	Código de diagnóstico 6	Alfanumérico	20	No
ClmDiagnosisCode_7	Código de diagnóstico 7	Alfanumérico	20	No
ClmDiagnosisCode_8	Código de diagnóstico 8	Alfanumérico	20	No
ClmDiagnosisCode_9	Código de diagnóstico 9	Alfanumérico	20	No
ClmDiagnosisCode_10	Código de diagnóstico 10	Alfanumérico	20	No
ClmProcedureCode_1	Código de procedimiento 1	Alfanumérico	20	No
ClmProcedureCode_2	Código de procedimiento 2	Alfanumérico	20	No
ClmProcedureCode_3	Código de procedimiento 3	Alfanumérico	20	No

ClmProcedureCode_4	Código de procedimiento 4	Alfanumérico	20	No
ClmProcedureCode_5	Código de procedimiento 5	Alfanumérico	20	No
ClmProcedureCode_6	Código de procedimiento 6	Alfanumérico	20	No

Tabla 6 - Diccionario de datos de Inpatientdata Dataset

- **Outpatientdata (Train_Outpatientdata-1542865627584.csv):** Datos de pacientes que no requieren hospitalización (ambulatorios), los cuales son el foco de estudio dado que es donde se materializa el riesgo de fraude a raíz de que se presentan reclamaciones de servicios no prestados o necesarios, de estos se recolectan los datos mostrados en la siguiente tabla.

No **Campo**	**Descripción**	**Tipo de dato**	**Longitud**	**Obligatorio**
BeneID	Id del beneficiario	Alfanumérico	20	Si
ClaimID	ID de solicitud	Alfanumérico	20	Si
ClaimStartDt	Fecha inicio de solicitud	Fecha	6	Si
ClaimEndDt	Fecha fin de solicitud	Fecha	6	Si
Provider	ID de proveedor	Alfanumérico	20	Si
InscClaimAmtReimbursed	Monto de reembolso solicitado	Decimal	10,6	Si
AttendingPhysician	Médico tratante	Alfanumérico	20	Si
OperatingPhysician	Médico operador	Alfanumérico	20	No
OtherPhysician	Otro médico	Alfanumérico	20	No
ClmDiagnosisCode_1	Código de diagnóstico 1	Alfanumérico	20	No
ClmDiagnosisCode_2	Código de diagnóstico 2	Alfanumérico	20	No
ClmDiagnosisCode_3	Código de diagnóstico 3	Alfanumérico	20	No
ClmDiagnosisCode_4	Código de diagnóstico 4	Alfanumérico	20	No
ClmDiagnosisCode_5	Código de diagnóstico 5	Alfanumérico	20	No
ClmDiagnosisCode_6	Código de diagnóstico 6	Alfanumérico	20	No
ClmDiagnosisCode_7	Código de diagnóstico 7	Alfanumérico	20	No
ClmDiagnosisCode_8	Código de diagnóstico 8	Alfanumérico	20	No
ClmDiagnosisCode_9	Código de diagnóstico 9	Alfanumérico	20	No
ClmDiagnosisCode_10	Código de diagnóstico 10	Alfanumérico	20	No
ClmProcedureCode_1	Código de procedimiento 1	Alfanumérico	20	No
ClmProcedureCode_2	Código de procedimiento 2	Alfanumérico	20	No
ClmProcedureCode_3	Código de procedimiento 3	Alfanumérico	20	No
ClmProcedureCode_4	Código de procedimiento 4	Alfanumérico	20	No
ClmProcedureCode_5	Código de procedimiento 5	Alfanumérico	20	No
ClmProcedureCode_6	Código de procedimiento 6	Alfanumérico	20	No
DeductibleAmtPaid	Monto deducible pagado	Decimal	10,6	Si

ClmAdmitDiagnosisCode	Código de diagnóstico admitido	Alfanumérico	20	Si

Tabla 7 - Diccionario de datos de Outpatientdata Dataset

- **Beneficiary (Train_Beneficiarydata-1542865627584.csv):** Datos del beneficiario el cual recolecta información relacionada al proceso de Know Your Customer (KYC), del cual se recolectan los siguientes datos.

Campo	Descripción	Tipo de dato	Longitud	Obligatorio
BeneID	Id del beneficiario	Alfanumérico	20	Si
DOB	Fecha de nacimiento	Fecha		Si
DOD	Fecha de muerte	Fecha		No
Gender	Género	Numérico	1	Si
Race	Raza	Numérico	1	Si
RenalDiseaseIndicator	Indicador de enfermedad renal	Numérico	1	Si
State	Estado	Numérico	1	Si
County	País	Numérico	1	Si
NoOfMonths_PartACov		Numérico	1	Si
NoOfMonths_PartBCov		Numérico	1	Si
ChronicCond_Alzheimer	Condición de Alzheimer crónico	Numérico	1	Si
ChronicCond_Heartfailure	Condición de insuficiencia cardiaca	Numérico	1	Si
ChronicCond_KidneyDisease	Condición Enfermedad renal crónica	Numérico	1	Si
ChronicCond_Cancer	Condición crónica de cáncer	Numérico	1	Si
ChronicCond_ObstrPulmonary	Condición crónica de obstrucción pulmonar	Numérico	1	Si
ChronicCond_Depression	Condición depresión	Numérico	1	Si
ChronicCond_Diabetes	Condición crónica de diabetes	Numérico	1	Si
ChronicCond_IschemicHeart	Condición crónica de corazón Isquémico	Numérico	1	Si
ChronicCond_Osteoporasis	Condición crónica de osteoporosis	Numérico	1	Si
ChronicCond_rheumatoidarthritis	Condición Artritis reumatoide crónica	Numérico	1	Si
ChronicCond_stroke	Condición accidente vascular crónica	Numérico	1	Si
IPAnnualReimbursementAmt	Monto anual de Reembolso atención ambulatoria	Decimal	10,6	Si
IPAnnualDeductibleAmt	Monto anual de deducible atención ambulatoria	Decimal	10,6	Si
OPAnnualReimbursementAmt	Monto anual de Reembolso atención hospitalización	Decimal	10,6	Si
OPAnnualDeductibleAmt	Monto anual de deducible atención hospitalización	Decimal	10,6	Si

Tabla 8 - Diccionario de datos de Beneficiary Dataset

- **Train (Train.csv):** Nos muestra la lista de proveedores de los cuales se identifica un potencial fraude

Campo	Descripción	Tipo de dato	Longitud	Obligatorio
Provider	Código del proveedor	Alfanumérico	20	Si
PotentialFraud	Resultado si es un potencial fraude	Texto	2	Si

Tabla 9 - Diccionario de datos de Train Dataset

Data Profiling

La primera etapa para garantizar una adecuada calidad de la data utilizada es realizar un proceso de descubrimiento y entendimiento de la información presente en el data set, con el fin de entender el problema de negocio y sus fuentes de información, esto nos permite identificar los procesos de limpieza, transformación y carga que serán necesarios ejecutar.

Generalidades del Dataset Outpatientdata

Como se menciona en el apartado del Diccionario de Datos, el Dataset que contiene la información que es foco de estudio es el Outpatientdata, ya que contiene la información de los pacientes ambulatorios y de los cuales se presenta el riesgo de fraude, motivo por el cuál a continuación se describen sus características generales.

Outpatientdata	
Descripción	Datos de pacientes que no requieren hospitalización (ambulatorios)
Fuente	Kaggle
URL	https://www.kaggle.com/datasets/rohitrox/healthcare-provider-fraud-detection-analysis
Formato	CSV
Total de registros	517,737
Total de atributos	27

Tabla 10 - Data profiling Outpatientdata Dataset

Resumen de perfil de datos para atributos categóricos y numéricos

De forma general a continuación se ha elaborado un resumen de las estadísticas del set de datos Outpatientdata y posteriormente se amplía información sobre cada atributo:

Atributo	Min	Max	Promedio	Desviación estándar	Conteo	Cantidad de nulos	Cantidad de únicos
AttendingPhysician	PHY311001	PHY434158			517737	1396	74110
BeneID	BENE110000	BENE99999			517737	0	133980
ClaimEndDt	39810	40178	39990		517737	0	366
ClaimID	CLM110011	CLM82318			517737	0	517737
ClaimStartDt	39794	40178	39989		517737	0	385

ClmAdmitDiagnosisCode		V854			517737	0	3716
ClmDiagnosisCode_1	0010	V8909			517737	10453	10355
ClmDiagnosisCode_10	0389	V860			517737	516654	496
ClmDiagnosisCode_2	0039	V8803			517737	195380	5057
ClmDiagnosisCode_3	0060	V8802			517737	314480	4449
ClmDiagnosisCode_4	00845	V8802			517737	392141	3926
ClmDiagnosisCode_5	0059	V8801			517737	443393	3413
ClmDiagnosisCode_6	0059	V8801			517737	468981	2969
ClmDiagnosisCode_7	00845	V8801			517737	484776	2636
ClmDiagnosisCode_8	00845	V8801			517737	494825	2261
ClmDiagnosisCode_9	00845	V8801			517737	502899	1895
ClmProcedureCode_1	151	9999			517737	517575	81
ClmProcedureCode_2	1741	9982			517737	517701	23
ClmProcedureCode_3	2724	4401			517737	517733	5
ClmProcedureCode_4	311	7840			517737	517735	3
ClmProcedureCode_5					517737	517737	1
ClmProcedureCode_6					517737	517737	1
DeductibleAmtPaid	0	897	2,779233472	15,78583932	517737	0	16
InscClaimAmtReimbursed	0	102500	286,3347993	694,0343433	517737	0	342
OperatingPhysician	PHY311005	PHY434155			517737	427120	28533
OtherPhysician	PHY311001	PHY434158			517737	322691	44389
Provider	PRV51001	PRV57763			517737	0	5012
Test_Outpatientdata					517737	0	

Tabla 11 – Resumen Data profiling Outpatientdata Dataset

BeneID

Valores únicos duplicados y nulos en el ID de beneficiario

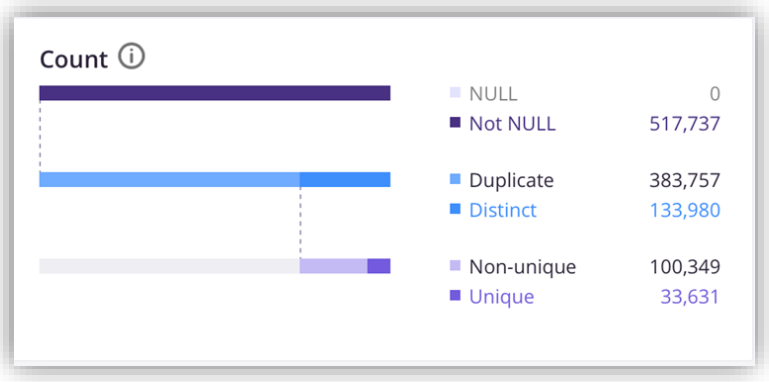


Ilustración 7 - Estadísticas atributo BeneID

Los beneficiarios totales en el data set son 133.980 de los cuales 33,631 solicitan un solo reembolso.

ClaimID

Cantidad de datos nulos, duplicados y únicos en el ID de solicitud de reembolso

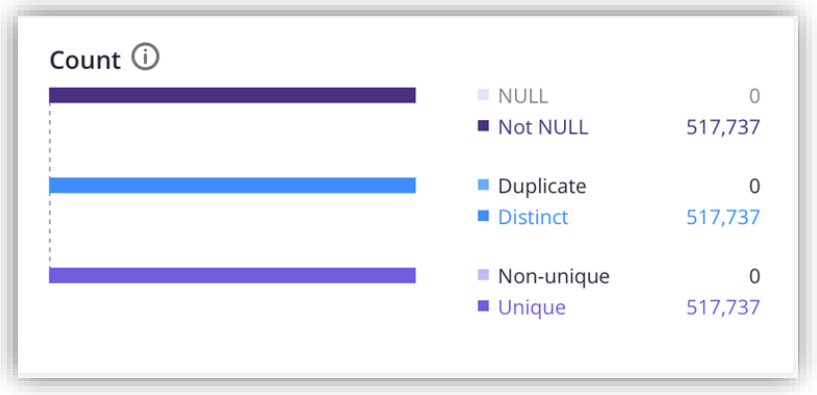


Ilustración 8 - Estadísticas atributo ClaimID

El id único de único de la tabla es el ID de reclamos o solicitudes de reembolso. Cada registro es una solicitud de reembolso única.

ClaimStartDt

Descripción de la distribución de fechas de tratamiento. Vemos que las solicitudes de reembolso se distribuyen de manera relativamente homogénea durante el período diciembre 2008 diciembre 2009.

Quantil	Fecha	Dias
0%	12/12/08	
10%	3/2/09	53
20%	10/3/09	35
30%	14/4/09	35
40%	19/5/09	35
50%	23/6/09	35
60%	29/7/09	36
70%	3/9/09	36
80%	11/10/09	38
90%	19/11/09	39
100%	31/12/09	42

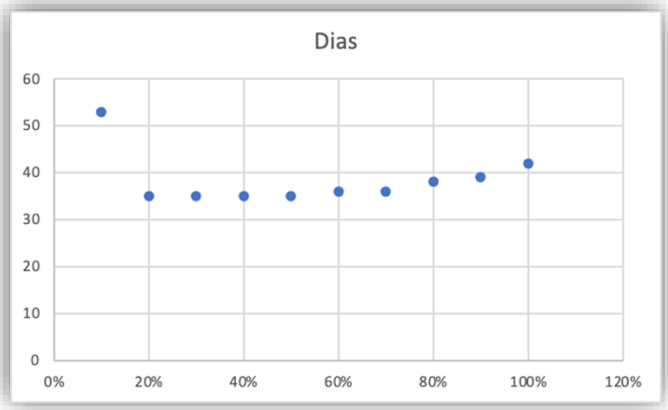


Ilustración 9 - Estadísticas atributo ClaimStartDt

Se observa una distribución homogénea durante todo el año en las solicitudes de reembolso, solo al menor entre los meses de invierno noviembre, diciembre, enero y febrero.

InscClaimAmtReimbursed

Descripción del monto efectivamente pagado por la compañía de seguros. El promedio es USD $286 con una desviación estándar de USD $694, y una máxima de USD $102.500. La mayoría de los reembolsos son bajo el promedio, la mediana es USD $80.

Ilustración 10 - Estadísticas atributo InscClaimAmtReimbursed

Distribuciones en deciles de los montos pagados por la compañía aseguradora.

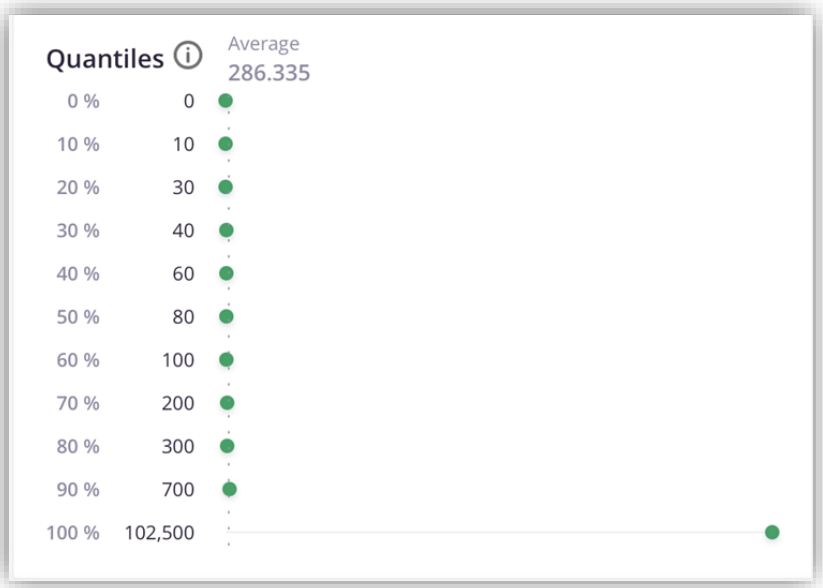


Ilustración 11 - Estadísticas atributo InscClaimAmtReimbursed

El 50% de los casos son por montos de reembolso pagados son bajo 80 dólares, y solo el 10% superior de casos son reembolsos por montos sobre 700 dólares. Casi el 80% de los casos está bajo la media de reembolsos.

AttendingPhysician

Descripción de variable médico tratante

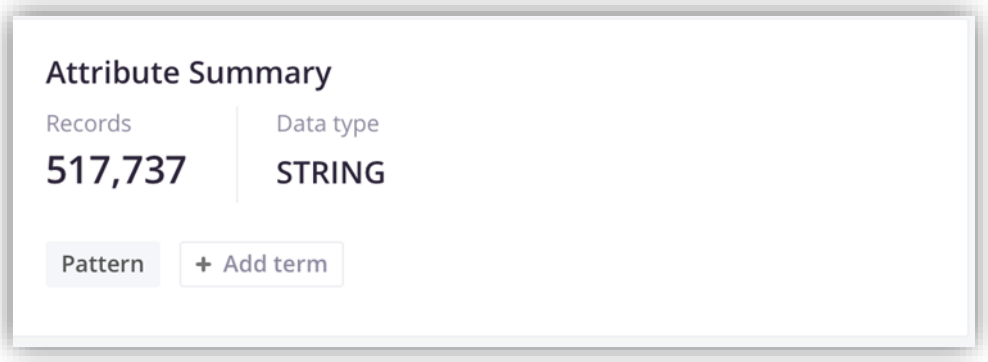


Ilustración 12 - Estadísticas atributo AttendingPhysician (totales)

Valores únicos, nulos y repetidos de médico tratante. Existen en la base de 517.737 solicitudes un total de 74.110 médicos distintos realizando el total de atenciones.

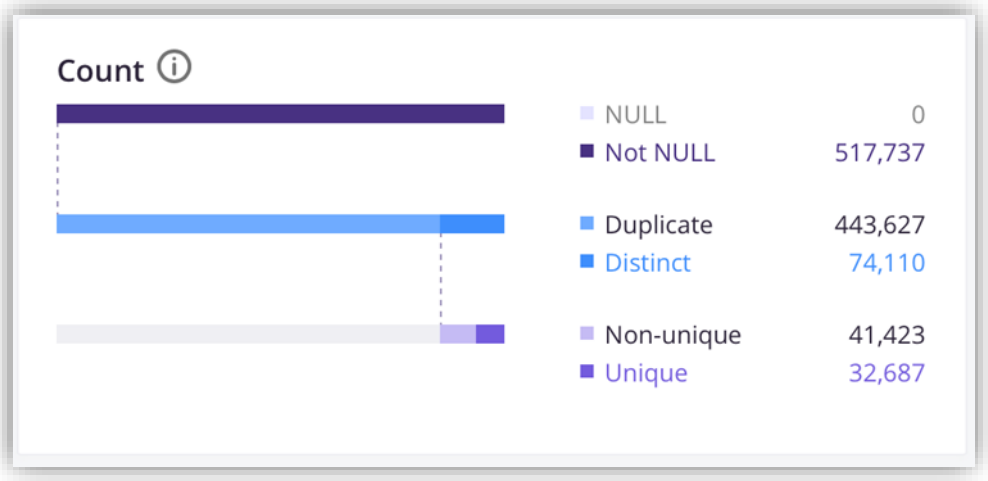
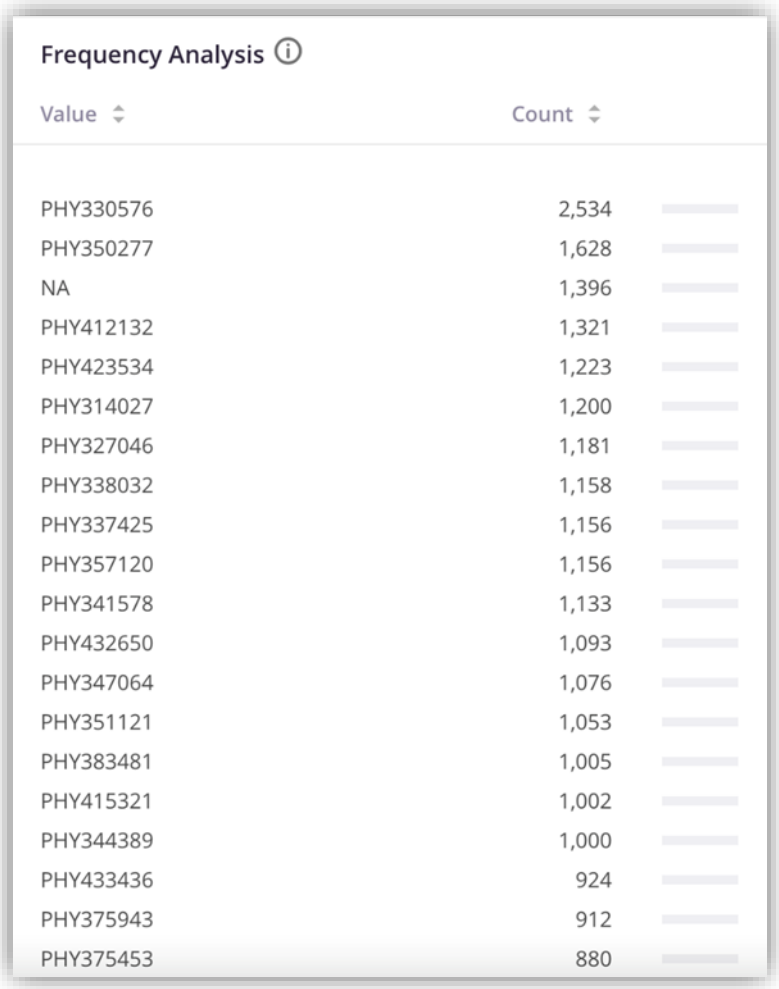


Ilustración 13 - Estadísticas atributo AttendingPhysician (general)



Frequency Analysis

Value	Count
PHY330576	2,534
PHY350277	1,628
NA	1,396
PHY412132	1,321
PHY423534	1,223
PHY314027	1,200
PHY327046	1,181
PHY338032	1,158
PHY337425	1,156
PHY357120	1,156
PHY341578	1,133
PHY432650	1,093
PHY347064	1,076
PHY351121	1,053
PHY383481	1,005
PHY415321	1,002
PHY344389	1,000
PHY433436	924
PHY375943	912
PHY375453	880

Ilustración 14 - Estadísticas atributo AttendingPysician (frecuencia)

Un 0,49% de los pacientes son atendidos por el doctor PHY330579, con la mayor concentración de casos, y 0,31% el siguiente doctor en la lista. No se observan saltos anómalos en la concentración de casos.

ClmDiagnosisCode_1

Un Valores únicos, distintos y nulos dentro de los diagnósticos principales

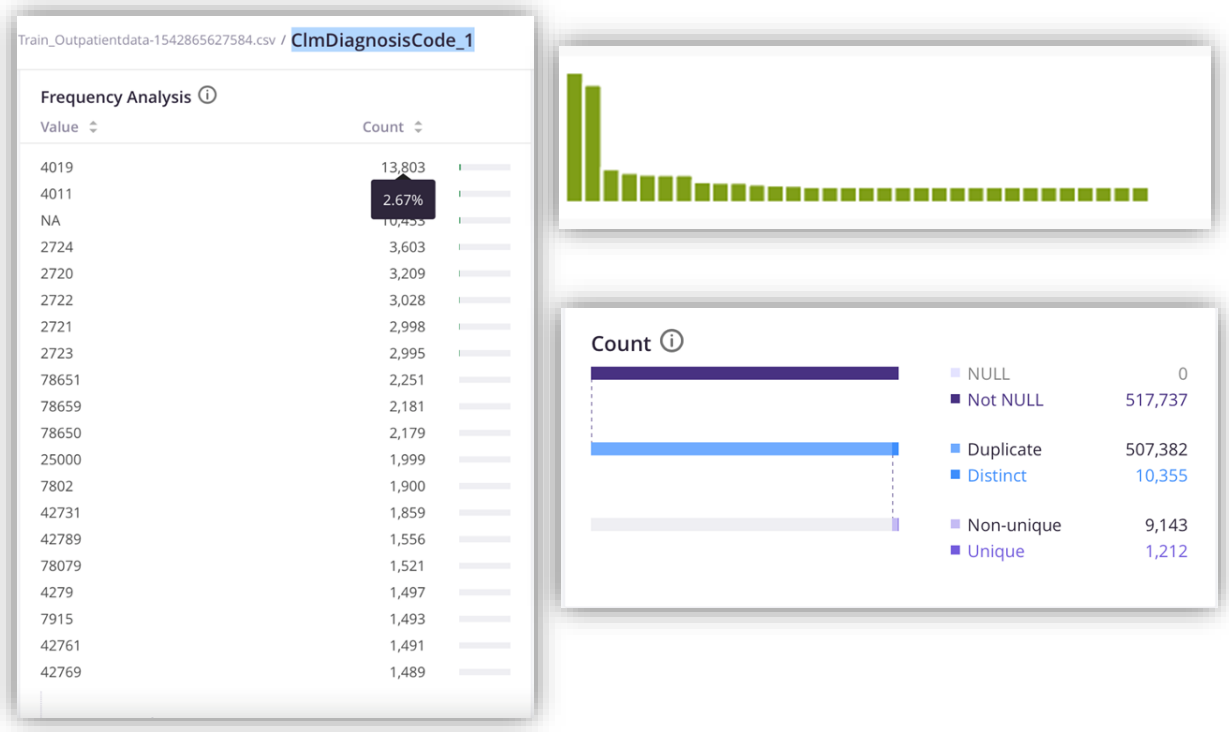


Ilustración 15 - Estadísticas atributo ClmDiagnosisCode_1

Diagnósticos 4019 y 4011 tienen preponderancia como principal diagnóstico, con 2,67% y 2,42% respectivamente sobre el total de casos.

Provider

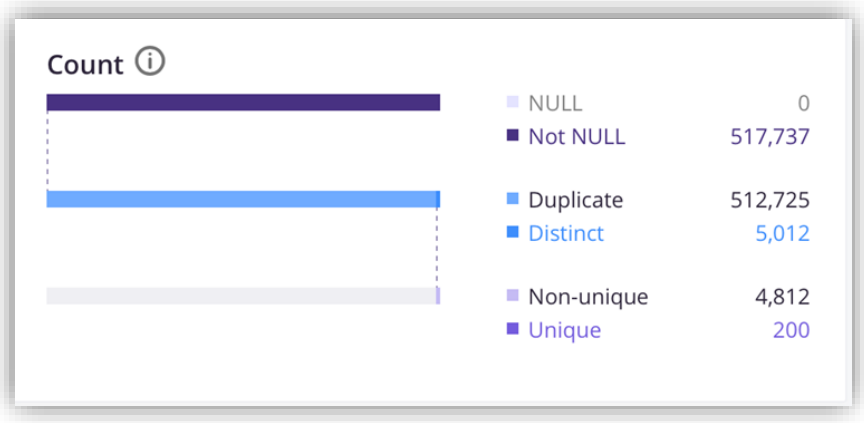


Ilustración 16 - Estadísticas atributo Provider

El data set cuenta con 512.725 solicitudes para un total de 5.012 centros de atención distintos. De ellos 200 tienen una sola solicitud dentro del data set.

Train_Outpatientdata-1542865627584.csv / **Provider**

Frequency Analysis

Value	Count
PRV51459	8,240
PRV53797	4,739
PRV51574	4,444
PRV53918	3,588
PRV54895	3,433
PRV55215	3,250
PRV56011	2,833
PRV52064	2,806
PRV55004	2,396
PRV57306	2,315
PRV52030	2,271
PRV52649	2,156
PRV52628	2,098
PRV51369	2,083
PRV51347	2,067
PRV56560	2,065
PRV55039	2,027
PRV57103	2,020
PRV54772	2,006
PRV51480	1,924

Ilustración 17 - Estadísticas atributo Provider (frecuencia)

1,59% de las solicitudes corresponden al prestador PRV51459 bastante superior al siguiente proveedor con 0,92% de las atenciones. Sin embargo, no se detectan outliers.

PotentialFraud

Frequency Analysis

Value	Count
No	328,343
Yes	189,394

Ilustración 18 - Estadísticas atributo PotentialFraud

36,6% de los casos están etiquetados como potencial riesgo de fraude en el data set completo

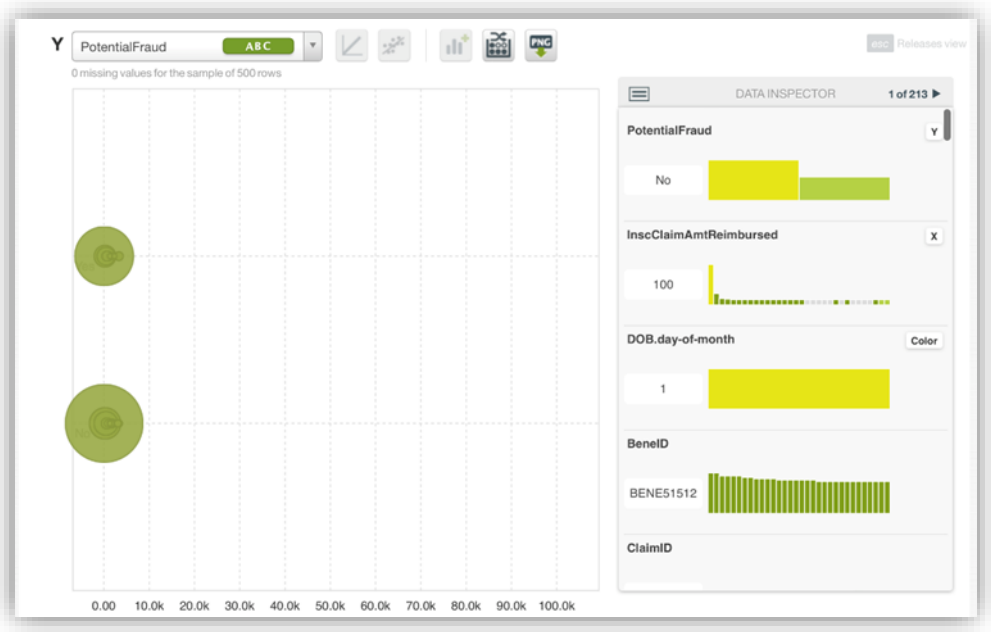


Ilustración 19 - Estadísticas atributo PotentialFraud (casos)

Se observa que los montos más altos están más presentes en los casos de no fraude.

Procesos de Data Quality

Posterior a la etapa de perfilamiento de datos, es necesario siempre identificar cuáles serán las reglas de calidad que deberán ser aplicadas durante la construcción del ETL para asegurar que el set de datos, insumo para el modelo, provee información relevante y ausente de errores que puedan impactar negativamente al modelo, como se ha mencionado anteriormente nos enfocaremos únicamente en el set de datos de Outpatientdata ya que es el foco de evaluación.

Regla general: Todos los campos relevantes siempre deberán contener datos, por lo que de encontrarse valores nulos se recomienda aplicar un valor por defecto usando las técnicas correspondientes, ver también la obligatoriedad de datos identificada en el apartado de "Diccionario de datos", el cuál ayudará a solucionar de raíz cualquier inconveniente con datos nulos o incompletos, acá debería entrar en acción el equipo de Data Governance para asegurar que sea resuelto cualquier inconveniente de calidad y que la solución sea lo más cercana al origen de datos.

Aparte de la regla general, con base en el perfilamiento de datos se han logrado identificar las siguientes consideraciones como reglas de calidad específicas:

Outpatientdata:

Campo	Descripción	Transformación
ClaimID	ID de solicitud	No aceptar duplicados
ClaimStartDt	Fecha inicio de solicitud	1980 <= ClaimStartDt <= fecha actual
ClaimEndDt	Fecha fin de solicitud	ClaimEndDt >= ClaimStartDt
InscClaimAmtReimbursed	Monto de solicitud reembolsado	>= 0
AttendingPhysician	Médico tratante	Not null
ClmDiagnosisCode_1	Código de diagnóstico 1	Not null
DeductibleAmtPaid	Monto deducible pagado	>=0

Tabla 12 - Reglas de calidad para dataset Outpatientdata

ETL

Diagrama de flujo

A continuación, se detalla el flujo general que llevará a cabo el proceso de ETL (extracción, transformación y carga de información), tanto al momento de construir el modelo (tomando como insumo la información de Kaggle) como al momento en que se realiza un deployment de este, el cuál en esa etapa productiva el ETL debería embeber el modelo construido en las etapas de "Organización de los datos" (capa harmonizada o bronce) en adelante para poder consumir la data nueva ya proveniente de los sistemas transaccionales de la compañía donde se haya implementado la solución.

Se debe recordar que todo modelo debe ser recalibrado o ajustado según la evolución de los datos, siempre y cuando las características o supuestos bajo los cuales fue creado hayan cambiado, por ello en la porción de la derecha de la siguiente ilustración, se especifican las etapas donde se deben ejecutar los ajustes correspondientes (ver pasos del 5 al 7).

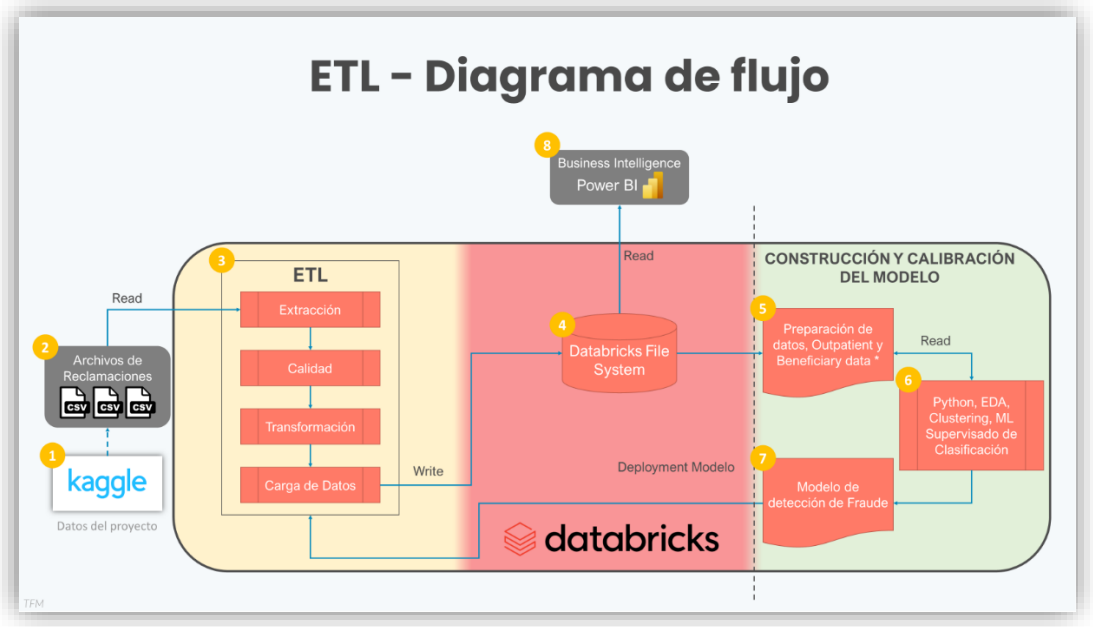


Ilustración 20 - Diagrama de flujo del ETL para construcción y mantenimiento del modelo de ML

1. Como fuente se utiliza el sistema de registros de pacientes y seguros de la compañía en la cual se registren todas las transacciones asociadas a las reclamaciones de los pacientes y se obtendrán todas las generales de los clientes, entre estos pueden estar: nombre, id, dirección, edad, sexo, etc. La información general de clientes y reclamaciones será almacenada en formato CSV para luego ser procesada y transformada según el modelo de datos que se necesita para el análisis, esto va a incluir cambiar el tipo de datos de las columnas, crear algunas nuevas según necesidad y

realizar limpieza de datos, lo cual genera un output principal para consumo en el modelo de clustering.

2. El input de información para el modelo en una fase inicial será tomado del repositorio “Kaggle” con datos alineados al estudio, esto servirá como plantilla para la puesta en producción de este en las diferentes empresas que lo usen de guía.
3. En el proceso ETL se ejecutarán las actividades de calidad de datos (limpieza y harmonización de datos) lo cual como output generará la tabla insumo para generación de modelo de clustering a ser analizado, entre las subtareas que el ETL realiza son:
    a. Extracción de datos:
        i. Configuración de lectura de archivos
        ii. Conexión a fuente de datos
        iii. Lectura de datos a plataforma analítica (Databricks)
    b. Actividades de Calidad de datos (limpieza y validación):
        i. Los códigos identificadores de paciente y médico, montos de reembolso, deducibles de pago, nombres y apellidos, numero de documento no pueden ser nulos o tienen que ser mayores que 0 en caso sean pagados.
        ii. Eliminar datos nulos o NA.
        iii. Validación de fechas de solicitud que estén comprendidos en un rango de años.
    c. Actividades de Transformación de datos:
        i. Las fechas se estandarizan a un formato uniforme (dd-mm-yyyy) que permita realizar consultas.
        ii. Convertir campos de condición del paciente (indicadores lógicos de verdadero o falso) a numéricos para evitar problemas en las consultas o redundancias.
        iii. Crear columnas calculadas de edad y tiempo de solicitud en días, entre otras columnas que se necesiten para el análisis.
    d. Carga de datos:
        i. Necesario poder monitorear esta actividad para detectar alguna caída y retroalimentar alguna mejora durante los pasos anteriores.

4. El resultado del proceso de ETL será depositado en el repositorio Databricks File System. Este repositorio será fuente de información para generar un modelo de entrenamiento que será analizado posteriormente, además va a recibir el resultado del modelo construido a raíz se ejecuten nuevas cargas de información con datos productivos, para que esta data enriquecida con la predicción pueda ser consumida por la capa de visualización, es decir es el principal repositorio de datos pre y post procesamiento.
5. La preparación de los datos consiste en realizar una selección de datos que serán input del modelo a ser construido, se preprocesa información para poder iniciar la fase construcción del modelo.
6. El análisis y evaluación será realizado principalmente vía Notebooks en Databricks apoyándose de librerías de tratamiento de alto volumen de datos, algoritmos de ML supervisados y de clasificación a ser ejecutados en Python.
7. Modelo de detección de Fraude es el output del procesamiento anterior, este resultado tiene que ser adaptado a un formato tabular para que nuevamente pase por un proceso ETL de validación y transformación de los datos e insertados al repositorio Databricks File System para su posterior consumo en la herramienta de visualización.
8. Por último, tenemos el dashboard de visualización, esto nos permite tener una visión general de los datos desde una perspectiva AS-IS en cómo han ingresado hasta mostrar el TO-BE de los datos con los hallazgos del modelo de Machine Learning (ML).

Estructuras y construcción del modelo

Con el objetivo de suplir las distintas necesidades de información que pueden surgir en el tiempo, el ETL se ha estructurado para que se creen las siguientes capas de información:

1. RAW: Esta capa contiene la información resultante de la ingesta de datos tal cual se recibe como input, sufre transformaciones y es fiel al contenido original.
2. BRONZE: Se realiza la limpieza de datos y las transformaciones necesarias (armonización de datos), como output se dejan estructuras con datos armonizados para uso de distintos fines (como de aplicaciones), previo a dejarlos disponibles en capa analítica.
3. GOLD: Esta capa contiene la información que quedará disponible para consumo de usuarios en capa analítica y uso de analítica avanzada, podrá ser consumida desde aplicaciones como Power BI para análisis descriptivos.

Nota: No se cuenta como capa de datos el espacio del Databricks File System donde se alojan los archivos Input (fuente). En este espacio se colocan los archivos "CSV" que contienen la información a ser procesados por el ETL. Esto según las necesidades de la empresa podrá ser sustituido por un sistema transaccional, FTP, Blob Storage, etc. Por el momento se alimenta con los archivos que se descargan de Kaggle.

A continuación, se proporcionan a nivel de diagramas las estructuras que se tienen por capa de datos:

Ilustración 21 - ETL - Estructuras por capa

En la siguiente imagen se muestran las estructuras Input y Output del modelo junto con un resumen de pasos generales que se aplican para construirlo:

Ilustración 22 - Modelo - Estructuras de Input, Output y pasos de construcción del Modelo

Nota: es importante tomar en cuenta que previo a ejecutar el ETL en Databricks, es necesario haber configurado el Clúster y adherir dicho poder de cómputo al

Notebook correspondiente. Para realizar esta actividad puede referirse al apartado de Anexo: Databricks.

7. OUTPUT DEL PROYECTO

Como se mencionó en el apartado de *investigación sobre modelos* , posterior a identificar la viabilidad del proyecto se procedió a tomar como base dicho insumo para proceder con la construcción de un modelo que aplicara las recomendaciones correspondientes, por lo que a continuación se detalla la información general del Producto Mínimo Viable construido

Resultado de Producto Mínimo Viable implementado

Características del modelo utilizando Extreme Gradient Boosting con XGBoost

Tomando como base la investigación expuesta anteriormente, nos inclinaremos por un modelo XGBoost, que es un modelo de clasificación de Gradient Boosting, de la familia de los árboles de decisiones, específicamente se clasifica como un tipo de Random Forest con Gradient Boosting.

De acuerdo con nuestros estudios descritos anteriormente, se recomendaba el uso de sobre muestreo para equilibrar los datos para los problemas de fraude. En nuestro caso no aplicamos esta recomendación ya que tenemos un alto porcentaje de proveedores con clasificación de potencial riesgo de fraude positivo, como se puede ver en el gráfico a continuación:

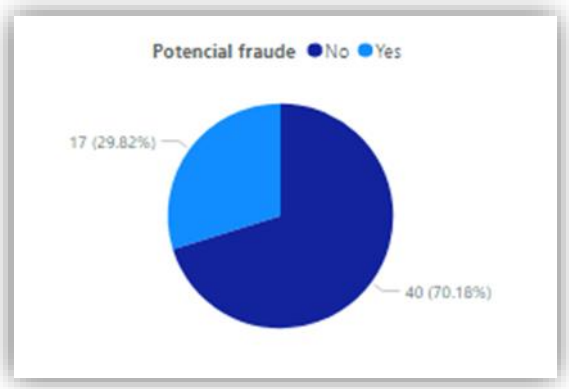


Ilustración 23 - Gráfico de porcentaje de Fraude

Trabajamos la compilación de códigos inicialmente en un jupyter notebook y luego en el compilador de Databricks. Las librerías utilizadas fueron las siguientes.

1. Pandas
2. Numpy
3. Matplotlib.pyplot
4. Seaborn
5. Plotly.express
6. XGBoost
7. Sklearn.preprocessing
8. Matplotlib
9. Sklearn.model_selection
10. Sklearn.metrics
11. Sklearn

Las principales etapas para la construcción del modelo son las siguientes:

- **Administración de los null**
  Se eliminan todas las columnas que tienen más de un 20% null, mientras en las otras columnas se remplaza por el mean (media) de los valores distintos de null.
- **Remplazo de valores categóricas**
  Los valores categóricos se remplazan por números para poder ser interpretados por el modelo por medio de la función LabelEncoder.

```
In [16]: ## transformación del label encoding.
         df.head()
```

Out[16]:

	Provider	InscClaimAmtReimbursed	AttendingPhysician	ClmDiagnosisCode_1	DeductibleAmtPaid	Train-1542865627584.Provider	Train-1542865627584.PotentialFraud
0	3726	30	9159	7242	0	3726	1
1	3726	40	14065	6350	0	3726	1
2	3726	200	55907	7912	0	3726	1
3	3726	20	17416	1277	0	3726	1
4	3726	40	20943	10115	0	3726	1

Ilustración 24 - Reemplazo de valores categóricos

- **Creación de columna objetivo y Split**
  Se crea primero la columna objetivo y luego se realiza el Split del dataset. Las variables independientes son los datos de entrada del modelo, mientras que la variable resultante es la clasificación binaria Fraude y No Fraude.

Creación de la columna objetivo

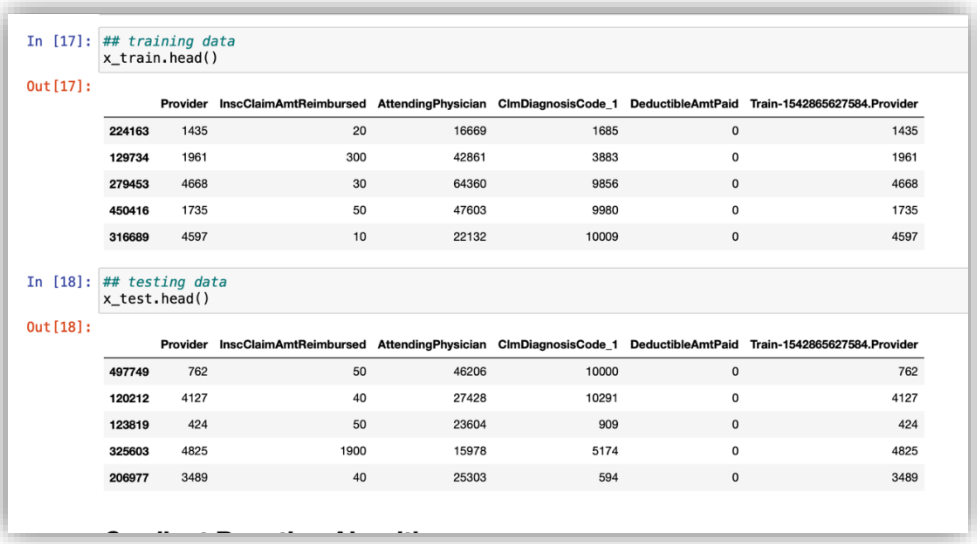
```
In [20]: ## Tomando la X como varaible independiente e y como variable objetivo
         X=df.drop(columns='Train-1542865627584.PotentialFraud')
         y=df['Train-1542865627584.PotentialFraud']
```

haciendo el split train test

```
In [21]: ### importación de comando para split el dataset
         from sklearn.model_selection import train_test_split

         ### split data 80% para el training del modelo y 20%  para testing del modelo
         x_train,x_test,y_train,y_test=train_test_split(X,y,test_size=0.2,random_state=0)
```

Ilustración 25 – Creación de columna objetivo y Split



```
In [17]: ## training data
         x_train.head()
```

Out[17]:

	Provider	InscClaimAmtReimbursed	AttendingPhysician	ClmDiagnosisCode_1	DeductibleAmtPaid	Train-1542865627584.Provider
224163	1435	20	16669	1685	0	1435
129734	1961	300	42861	3883	0	1961
279453	4668	30	64360	9856	0	4668
450416	1735	50	47603	9980	0	1735
316689	4597	10	22132	10009	0	4597

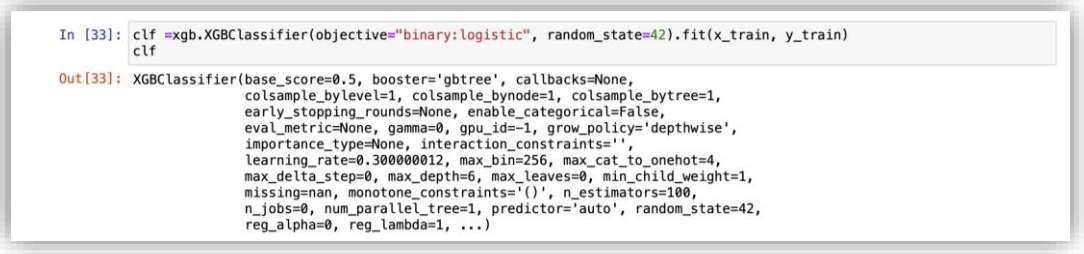
```
In [18]: ## testing data
         x_test.head()
```

Out[18]:

	Provider	InscClaimAmtReimbursed	AttendingPhysician	ClmDiagnosisCode_1	DeductibleAmtPaid	Train-1542865627584.Provider
497749	762	50	46206	10000	0	762
120212	4127	40	27428	10291	0	4127
123819	424	50	23604	909	0	424
325603	4825	1900	15978	5174	0	4825
206977	3489	40	25303	594	0	3489

Ilustración 26 – Training y test de datos

- **Entrenamiento de algoritmo de Gradient Boosting.**
  El modelo fue entrenado y genera la matriz de parámetros. Los parámetros están ajustados de manera automática y no se realizó ninguna intervención posterior de los parámetros. En el proceso de entrenamiento el modelo es provisto de data clasificada previamente como Fraude y No Fraude, con el fin de optimizar el *fit* de las variables independientes con la variable resultado.



```
In [33]: clf =xgb.XGBClassifier(objective="binary:logistic", random_state=42).fit(x_train, y_train)
         clf
```

Out[33]:

```
XGBClassifier(base_score=0.5, booster='gbtree', callbacks=None,
              colsample_bylevel=1, colsample_bynode=1, colsample_bytree=1,
              early_stopping_rounds=None, enable_categorical=False,
              eval_metric=None, gamma=0, gpu_id=-1, grow_policy='depthwise',
              importance_type=None, interaction_constraints='',
              learning_rate=0.300000012, max_bin=256, max_cat_to_onehot=4,
              max_delta_step=0, max_depth=6, max_leaves=0, min_child_weight=1,
              missing=nan, monotone_constraints='()', n_estimators=100,
              n_jobs=0, num_parallel_tree=1, predictor='auto', random_state=42,
              reg_alpha=0, reg_lambda=1, ...)
```

Ilustración 27 - Entrenamiento de algoritmo de Gradient Boosting

Resultado del Modelo

En este apartado mostraremos el Resultado del Modelo a raíz de haber realizado una serie de pruebas para evaluar su desempeño.

En primer lugar, revisamos el score de *clf* (modelo) tanto para data de entrenamiento como para data *unseen* con el fin de identificar sobre ajuste a los datos de entrenamiento. En este caso vemos que el score para el data set de prueba es mayor que el score del dataset de entrenamiento por lo que descartamos un sobre ajuste del modelo.

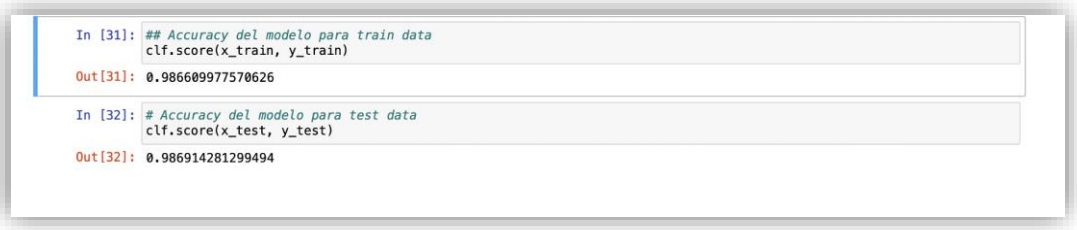


Ilustración 28 - Accuracy del modelo

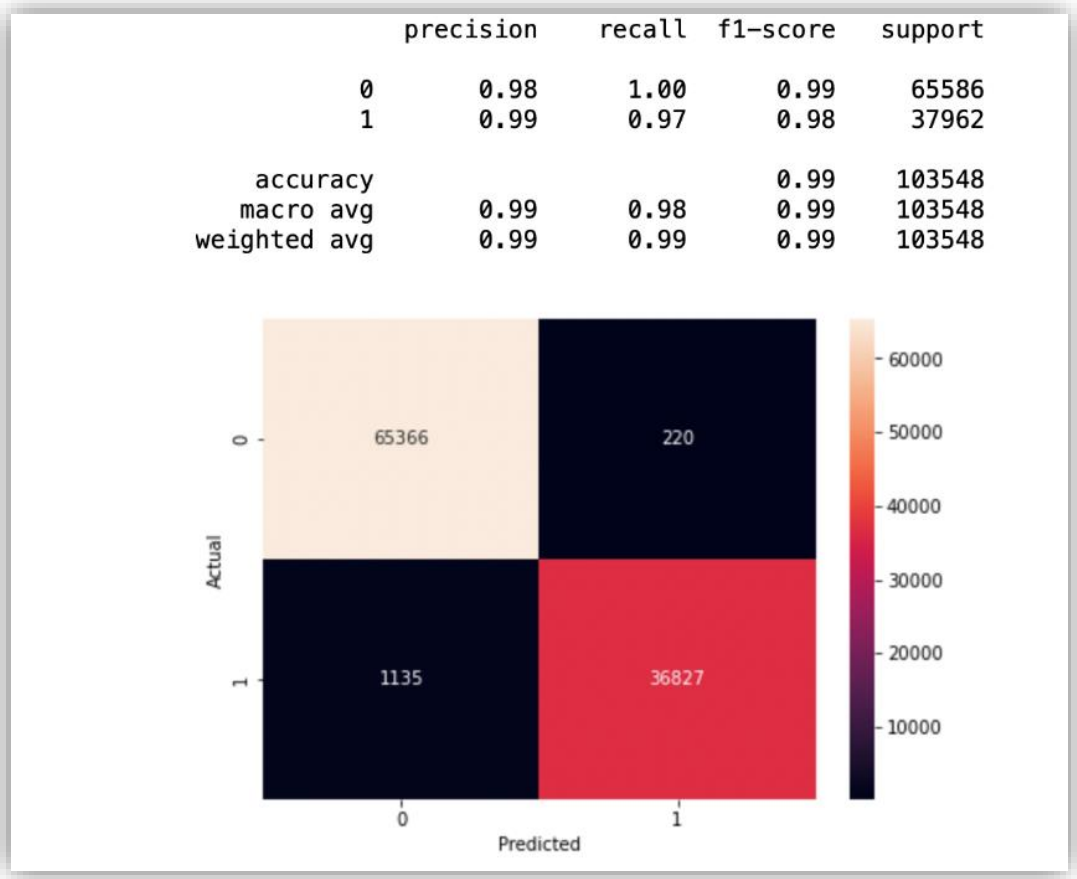


Ilustración 29 - Resultado del modelo

Curva ROC

```
In [39]: ### ROC
from sklearn import metrics
y_pred_proba =clf.predict_proba(x_test)[::,1]
fpr, tpr, _ = metrics.roc_curve(y_test,  y_pred_proba)

#crear curva ROC
plt.plot(fpr,tpr)
plt.ylabel('True Positive Rate')
plt.xlabel('False Positive Rate')
plt.show()
```

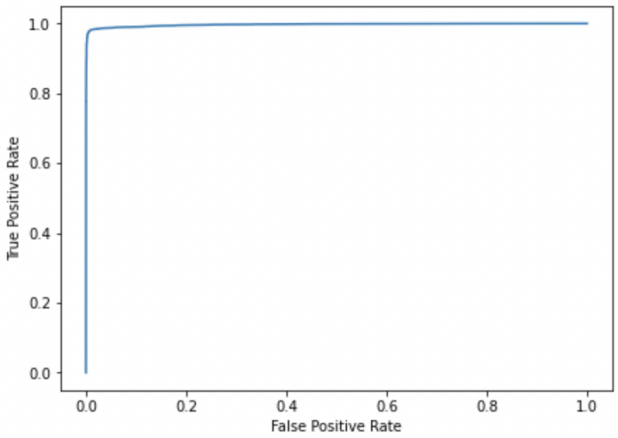


Ilustración 30 - ROC del modelo

Análisis de resultados en Power BI

Se ha construido un Dashboard (*para más información referirse al apartado de Anexo donde se comparten imágenes y explicación del tablero de Power BI construido*) en el cual se puede identificar:

- La proporción de casos que son identificados como fraude
- Top 10 de proveedores asociados a solicitudes
- Top 10 de médicos incluidos en solicitudes
- Top 10 de diagnósticos pagados
- Detalle de las solicitudes

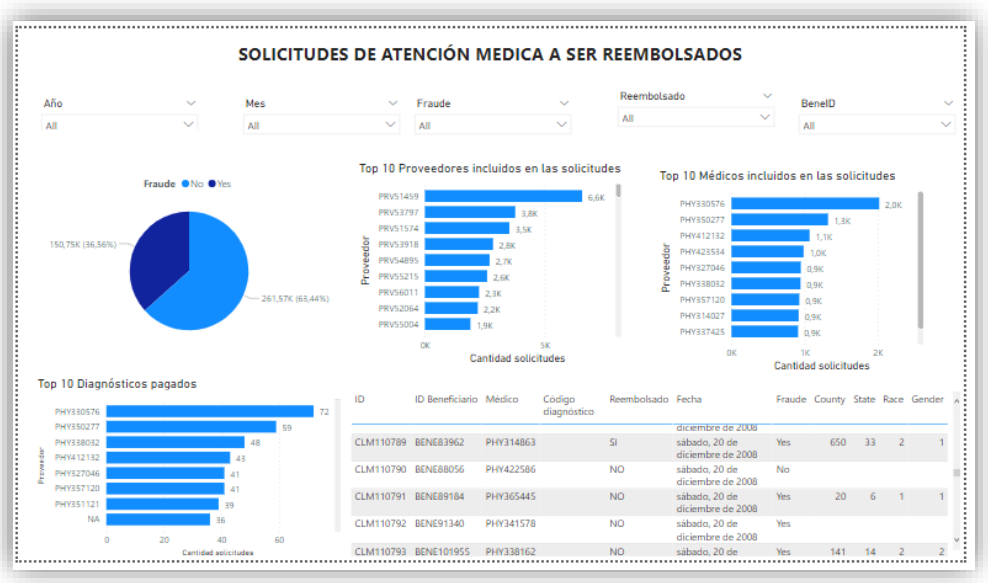


Ilustración 31 - Análisis de resultados en Power BI

8. CONCLUSIONES

A continuación, se listan las principales conclusiones obtenidas a raíz de la ejecución del proyecto de fin de máster:

- **A raíz de la ejecución del proyecto se demuestra que sí es posible construir un modelo predictivo con base en datos históricos sobre la clasificación de Fraude**

  Las empresas de seguro pueden tomar como base este proyecto para aplicar modelos predictivos que permitan identificar de forma oportuna que solicitudes pueden ser consideradas como potencial fraude para la compañía, únicamente deberían utilizar su data histórica y homologarla a los atributos que se utilizaron en el desarrollo de este proyecto, para posteriormente aplicarlo por medio de un ETL a toda información en tiempo real que reciban.

- **Es posible crear un modelo con alto desempeño general utilizando la librería XGBoost**

  El modelo construido muestra un buen desempeño general como se puede ver en el apartado Resultado del Modelo. Nuevamente la referencia para evaluar nuestros resultados es el artículo publicado por ResearchGate [17] mencionado en la bibliografía.

[17] ResearchGate, Rohan Yashraj Gupta, Pallav Kumar Baruah, Satya Sai Mudigonda (marzo 2021). *A Comparative Study of Using Various Machine Learning and Deep Learning-Based Fraud Detection Models For Universal Health Coverage Schemes*. Obtenido de https://www.researchgate.net/profile/Rohan-Gupta-5/publication/350132738_A_Comparative_Study_of_Using_Various_Machine_Learning_and_Deep_Learning-Based_Fraud_Detection_Models_For_Universal_Health_Coverage_Schemes/links/605d6e21a6fdccbfea085b79/A-Comparative-Study-of-Using-Various-Machine-Learning-and-Deep-Learning-Based-Fraud-Detection-Models-For-Universal-Health-Coverage-Schemes.pdf

TABLE 1 - PERFORMANCE METRICS OF FIVE MACHINE LEARNING MODELS USING BASELINE DATA AND OVERSAMPLED DATA

Models			AUC-ROC	Recall	Specificity	Precision	Accuracy	F1 Score
Decision Tree	Baseline	M1	**0.9566**	0.9248	0.9885	0.9174	0.9808	0.9211
	SMOTE	M2	0.9534	0.9208	0.9860	0.9006	0.9781	0.9106
	ADASYN	M3	0.9508	0.9155	0.9862	0.9016	0.9776	0.9085
	TGANs	M4	0.9548	0.9214	0.9883	0.9155	0.9801	0.9185
Random Forest	Baseline	M5	0.9462	0.8947	0.9977	0.9818	0.9852	0.9362
	SMOTE	M6	0.9493	0.9027	0.9959	0.9682	0.9846	0.9343
	ADASYN	M7	0.9500	0.9057	0.9942	0.9556	0.9834	0.9300
	TGANs	M8	0.9460	0.8942	0.9977	0.9820	0.9852	0.9361
XGBoost	Baseline	M9	0.9307	0.8615	0.9999	0.9989	0.9831	0.9252
	SMOTE	M10	0.9458	0.8970	0.9945	0.9572	0.9826	0.9262
	ADASYN	M11	0.9270	**0.9835**	0.8705	0.5119	0.8842	0.6733
	TGANs	M12	0.9111	0.8223	**1.0000**	**1.0000**	0.9784	0.9025
LightGBM	Baseline	M13	0.9486	0.8977	0.9994	0.9952	**0.9871**	**0.9440**
	SMOTE	M14	0.9499	0.9014	0.9988	0.9905	0.9869	0.9438
	ADASYN	M15	0.9523	0.9105	0.9940	0.9547	0.9839	0.9320
	TGANs	M16	0.9482	0.8970	0.9994	0.9950	0.9870	0.9435
GBM	Baseline	M17	0.9425	0.8852	0.9997	0.9975	0.9858	0.9380
	SMOTE	M18	0.9451	0.8958	0.9945	0.9576	0.9825	0.9257
	ADASYN	M19	0.9288	0.9779	0.8796	0.5288	0.8916	0.6864
	TGANs	M20	0.9282	0.8566	0.9992	0.9992	0.9224	0.9224

Ilustración 32 - Performance metrics of five ML models

- **El resultado de las variables de Specificity, Precision and Accuracy del modelo construido se encuentran por debajo del benchmark "TGANs XBoost"**

  Si bien las diferencias no son sustanciales, si calculamos los indicadores principales de desempeño vemos que tres indicadores están por debajo del benchmark "TGANs XBoost".

DESEMPEÑO DEL MODELO			Modelo	TGANs XBoost
Recall	TP / (TP+FN)	% de casos verdaderos detectados	0.9701	0.8223
Specificity	TN / (TN+FP)	% de casos falsos detectados	0.9966	1.0000
Precision	TP / (TP+FP)	% de predicciones positivas acertadas	0.9941	1.0000
Accuracy	(TP+TN) / (FP+FN+TN+TP)	% de predicciones acertadas	0.9869	0.9871

TP = True Positive
TN = True Negative
FP = False Positive
FN = False Negative

Ilustración 33 - Desempeño del modelo

- **El resultado de la variable Recall del modelo es superior al benchmark "TGANs XBoost"**

  En el indicador de sensibilidad (Recall), que por lo demás es el aspecto más importante al momento de evaluar un modelo de fraude, observamos un desempaño superior al benchmarck.

Este indicador representa la capacidad de predecir un fraude de entre los casos que ya clasificados como fraude en la data insumo. Porcentaje de verdaderos positivos detectados sobre el total de positivos reales.

- **Databricks es un componente que permite reducir la arquitectura y enfocarse en la etapa de construcción y posterior consumo de información**

  El componente de Databricks al proveer capacidad de procesamiento y una flexibilidad respecto a lenguajes de programación, permitió acelerar los tiempos de desarrollo ya que en un mismo componente se pudo realizar el ciclo completo del ETL y proveer la capa de datos para la herramienta de visualización (en este caso Power BI)

- **Databricks provee una simplificación en configuración del ambiente acelerando los tiempos de puesta en producción**

  Al apalancarnos en Databricks logramos disminuir los tiempos de aprovisionamiento de la arquitectura y se simplifica la instalación de librerías para iniciar la etapa de construcción.

- **La calidad de información es crucial para el desarrollo de un proyecto de Machine Learning**

  Al momento de ejecutar la creación del modelo se logró identificar que la Calidad de Datos es fundamental en el tratamiento de información y es crucial que todo ETL contemple un apartado de armonización y limpieza de datos ya que no se puede asumir que los datos vendrán con la calidad requerida.

- **Las herramientas de visualización permiten construir y comunicar adecuadamente los resultados de los análisis**

  Al utilizar Power BI como herramienta de consumo se logró crear un producto mínimo viable que permite consumir la información procesada por un ETL, facilitando el ciclo de toma de decisiones y proveyendo un medio único para distribuir conocimiento dentro de los equipos de forma transversal.

9. PRINCIPALES APRENDIZAJES

La ejecución de este trabajo de fin de máster nos ha permitido identificar los siguientes aprendizajes, los cuales se volvieron en fortalezas para alcanzar los logros que obtuvimos, al superar los retos durante el desarrollo del trabajo:

- **Enfoque metodológico**

  Con el objetivo de llevar a cabo un modelo de machine learning se requiere de una elaboración metodológica que contemple siempre una investigación bibliográfica, esto para encontrar las mejores prácticas a la hora de abordar el problema específico de negocio, esto en el entendido que un problema implica una similitud importante en las características de los datos y del modelo a desarrollar.

- **Trabajo colaborativo comunidad Python**

  El conocimiento compartido por la comunidad Python con un amplio repositorio de librerías y paquetes con código pre-combinado que permite reducir el tiempo de código. XBoost Python Package[18] nos dio acceso al paquete de instalación, permitió la instalación de la librería y tuvimos acceso a la documentación asociada. Fue igualmente importante integrar el modelo en producción. Con la función "predict" el modelo utiliza los parámetros definidos para predecir nuevos resultados sin tener una variable objetivo. Para asociar a los datos originales con la nueva columna de resultados utilizamos un índice presente en el dataset de entrada que nos permitió incorporar la variable resultado en los análisis y Dashboard posteriores.

- **Uso de servicios y herramientas de procesamiento de datos**

  Como parte de la investigación y la metodología aplicada al proyecto, los datos públicos utilizados fueron importantes para poder estructurar un ETL adecuado, el cual durante las pruebas se evaluaron múltiples herramientas y servicios cloud, por lo que BigML da un panorama más intuitivo en el análisis de los múltiples algoritmos de Machine Learning. Esto nos da un roadmap más claro para el flujo del modelo y la implementación de una arquitectura. Con el análisis y resultados de los datos, Python nos provee de múltiples librerías para Machine Learning y la facilidad de transformarlos en conjunto de datos tipo Dataframe propio de la librería Pandas. Databricks nos permite simplificar la instalación de librerías, modelar los datos y transformarlos mediante scripts, así como también, provee la facilidad de consumo de datos hacia las herramientas de

[18] DMLC XGBoost. (s.f.). *Python Package Introduction*. Obtenido de https://xgboost.readthedocs.io/en/stable/python/python_intro.html#install-xgboost

visualización, todo esto facilitando que la arquitectura no sea tan compleja para construir un modelo.

- **Con el uso de herramientas de BIG Data en nube es posible la creación de productos de forma ágil**

  Los componentes de BIG Data en nube proveen mucha flexibilidad y agilidad para aprovisionar y crear soluciones de machine learning.

- **En las instituciones, es importante la implementación de criterios de calidad y de aplicación de un programa de Data Governance para asegurar la calidad del dato**

  Sin calidad de datos el flujo de toma de decisiones siempre se verá mermado, y con un impacto negativo, si no se aplican fases de validación previo a la construcción de modelos de machine learning.

10. SIGUIENTES PASOS

A continuación, se detallan las actividades que identificamos podrían ser ejecutas para seguir buscando una mejora continua en el objeto de estudio:

- **Enfoque no supervisado**

  Abordamos el problema de negocios como un problema de clasificación. Sería interesante evaluar otros enfoques no supervisados para, por una parte, tener mayor insight en la etapa de entendimiento de la data, y por último poder encontrar nuevos patrones de potencial fraude distintos a los entrenados en el modelo, ya que los modelos de clasificación se entrenan con base a casos de fraudes conocidos, donde podrían ocultarse comportamientos distintos.

- **Ajustar los parámetros del modelo**

  Con el fin de mejorar aún más el performance del modelo, es importante probar distintos ajustes de los parámetros para alterar positivamente el comportamiento de este. Adicionalmente, creemos que sería interesante comparar los resultados utilizando la técnica de "sobre muestreo", a pesar de que la data está equilibrada.

- **Data complementaria**

  Consideramos que data de contexto del proveedor puede entregar mayor riqueza tanto al modelo entrenado, como a la etapa de análisis e interpretación posterior. Algunos datos que sería de valor incorporar son: el tamaño de mercado, la distinción pública / privado, la ubicación geográfica, datos públicos asociados a las localidades, por ejemplo, criminalidad o cantidad de médicos per cápita, entre otros.

- **Prueba de concepto**

  Sería muy interesante, en posteriores trabajos, desarrollar al menos en calidad de prueba de concepto, la implementación en una organización específica de seguros de salud, con el fin de evaluar la capacidad de adaptación y ajustes necesarios de acuerdo con las variaciones regionales del caso de negocio, y el contexto regulatorio y organizacional particular.

- **Mejora continua**

  El modelo es capaz de consumir nuevos datos *unseen* (no vistos por el modelo en etapa de entrenamiento) y entregar resultados con base a los parámetros fijos del modelo ajustado. Una vez que el modelo esté en producción consideramos valioso que, con cierta periodicidad,

nuevos casos validados de fraude sean incorporados a la data de entrenamiento y testing para mejorar el resultado del modelo, e identificar potenciales nuevos esquemas de fraude no previstos en los parámetros anteriores.

- **Streaming de datos**

  Una vez que el modelo esté en producción sería de interés incorporar una arquitectura de streaming para que el negocio tenga información en tiempo real. Dada las características del caso de negocio, la clasificación de riesgo de fraude a nivel de cada transacción debe ser oportuna para lograr "atajar" pagos fraudulentos.

- **Modelo relacional de los datos**

  Establecer un modelo dimensional e implementarlo en un Data Warehouse el cuál permita realizar análisis descriptivos de la data y también que provea el insumo para el modelo, cada casuística de negocio es diferente, pero dimensiones como datos de pacientes, personal de salud siempre serán fijos. Esto servirá para tener más claro el almacenamiento de los datos y su consumo para la mejora de diferentes procesos ETL o de visualización.

- **Mejoras en el ETL**

  Respecto a mejoras en el ETL se podrían aplicar las que se detallan a continuación:

  - Capa Bronze para insumo del modelo
    - Se podrían aplicar campos calculados a raíz de los datos insumo, como, por ejemplo: calcular la edad del beneficiario
  - Armonización de datos
    - Se podría aplicar estandarización de formato de fecha que permita realizar consultas
  - Como output para capa de Análisis se podría crear un Modelo Dimensional para facilitar la analítica descriptiva para el negocio
  - Convertir campos de condición del paciente (indicadores lógicos de verdadero o falso) a numéricos para evitar problemas en las consultas o redundancias
  - Crear columnas de envejecimiento de solicitudes (tiempo que tienen abiertas previo a su cierre)
  - Crear dimensión de códigos únicos de diagnóstico y procedimiento de cada solicitud presentada

11. BIBLIOGRAFIA

- INESE. (2 de junio 2022). *Crecen los intentos de fraude al seguro y su coste económico*. Obtenido de

  https://www.inese.es/crecen-los-intentos-de-fraude-al-seguro-y-su-coste-economico/#:~:text=Detr%C3%A1s%20del%201%2C60%25%20de,2019%20(1%2C38%25)

- KAGGLE. (s.f.). Obtenido de

  https://www.kaggle.com/

- KAGGLE. (s.f.). *HEALTHCARE PROVIDER FRAUD DETECTION ANALYSIS*. Obtenido de

  https://www.kaggle.com/datasets/rohitrox/healthcare-provider-fraud-detection-analysis

- ResearchGate, Rohan Yashraj Gupta, Pallav Kumar Baruah, Satya Sai Mudigonda (marzo 2021). *A Comparative Study of Using Various Machine Learning and Deep Learning-Based Fraud Detection Models For Universal Health Coverage Schemes*. Obtenido de

  https://www.researchgate.net/profile/Rohan-Gupta-5/publication/350132738_A_Comparative_Study_of_Using_Various_Machine_Learning_and_Deep_Learning-Based_Fraud_Detection_Models_For_Universal_Health_Coverage_Schemes/links/605d6e21a6fdccbfea085b79/A-Comparative-Study-of-Using-Various-Machine-Learning-and-Deep-Learning-Based-Fraud-Detection-Models-For-Universal-Health-Coverage-Schemes.pdf

- GARTNER, Ashutosh Gupta. (19 de enero 2022). *What Are AI Architects and What Do They Do?*. Obtenido de

  https://www.gartner.com/en/articles/what-are-ai-architects-and-what-do-they-do

- GARTNER. (s.f.). *Gartner Glossary - Data Scientist*. Obtenido de

  https://www.gartner.com/en/information-technology/glossary/data-scientist#:~:text=The%20data%20scientist%20role%20is,fulfilled%20better%20as%20a%20team

- Future Learn. (s.f.). *Roles and Responsibilities in Data-Driven Organizations*. Obtenido de

https://www.futurelearn.com/info/courses/financial-analysis-business-performance-data-driven-decision-making-sc/0/steps/177018

- SCRUM.ORG. (s.f.). *What is a Scrum Master?*. Obtenido de

  https://www.scrum.org/resources/what-is-a-scrum-master

- SCRUM.ORG. (s.f.). *What is a Product Owner?*. Obtenido de

  https://www.scrum.org/resources/what-is-a-product-owner

- Johns Hopkins HealthCare LLC. (s.f.). *Health Care Fraud and Abuse*. Obtenido de

  https://www.hopkinsmedicine.org/johns_hopkins_healthcare/providers_physicians/health_care_fraud_and_abuse/

- United States Sentencing Commission. (s.f.). *HEALTH CARE FRAUD*. Obtenido de

  https://www.ussc.gov/research/quick-facts/health-carefraud#:~:text=(July%202021)%20In%20fiscal%20year,36.2%25%20since%20fiscal%20year%202016.&text=The%20USSC%20HelpLine%20assists%20practitioners%20in%20applying%20the%20guidelines

- KEN PAXTON ATTORNEY GENERAL of TEXAS. (s.f.). *Health Care Fraud and Abuse*. Obtenido de

  https://www.texasattorneygeneral.gov/consumer-protection/health-care/health-care-fraud-and-abuse#medicare

- DMLC XGBoost. (s.f.). *Python Package Introduction*. Obtenido de

  https://xgboost.readthedocs.io/en/stable/python/python_intro.html#install-xgboost

12. ANEXOS

A continuación, se presentan los componentes complementarios de la solución, los cuales proporcionan un mayor nivel de detalle a los artefactos que componen el producto final del proyecto.

Código

Se adjunta Notebook de Python en el siguiente archivo:

ETL_v2.ipynb

Y también se puede acceder por medio de este enlace:

https://databricks-prod-cloudfront.cloud.databricks.com/public/4027ec902e239c93eaaa8714f173bcfc/3255798924883686/4494598147557877/4091501915668071/latest.html

Configuración de componente Databricks

Como se especifica en la arquitectura, el componente de Big Data utilizado para construir y ejecutar el modelo de Machine Learning fue Databricks, del cual a continuación se detallan los conjuntos de pasos necesarios para poder configurar el ambiente y así poder llevar a cabo las tareas de construcción, evaluación y ejecución del modelo.

1. Configuración de cuenta en Databricks versión community
    a. Se ingresa a URL: https://community.cloud.databricks.com

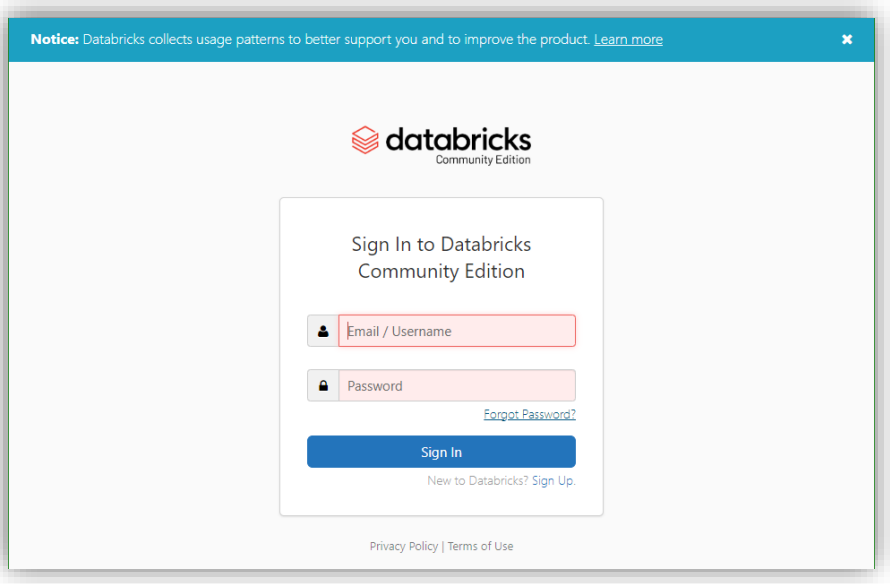


Ilustración 34 - Databricks Community

b. Se realiza registro de cuenta (Sign Up)

c. Se realiza Sign In con la cuenta creada en el paso b

d. Esto nos llevara a la página principal de Databricks, desde la cual se puede acceder a los módulos principales como Workspace (espacio de trabajo con notebooks), Data (file system), Compute (configuración de clústeres) y Workflows (administración de flujos de trabajo)

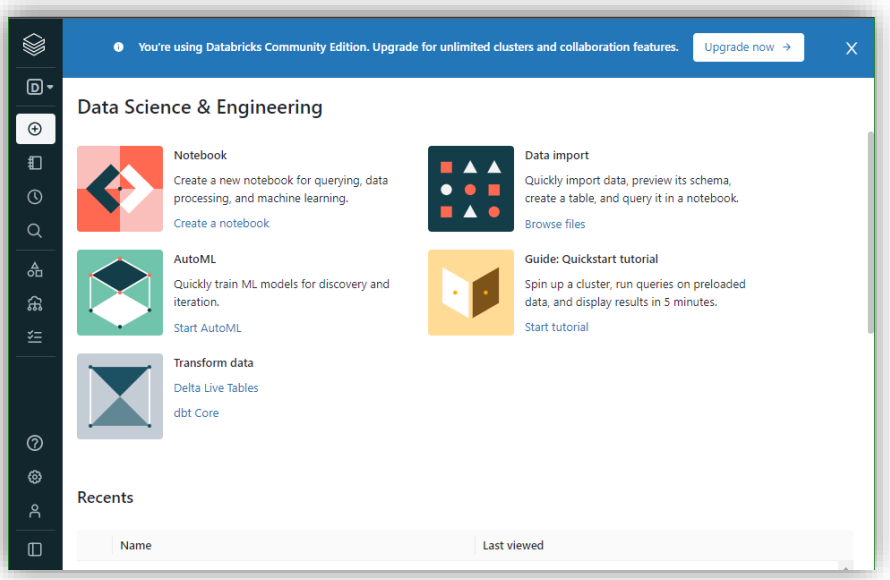


Ilustración 35 - Databricks página principal

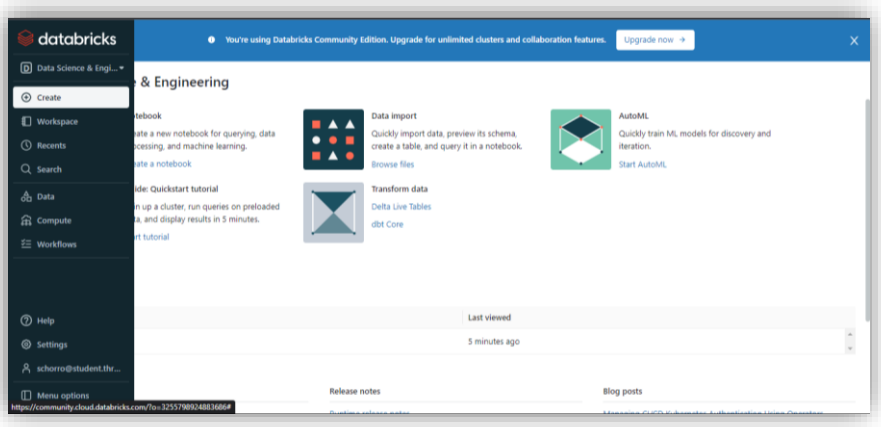


Ilustración 36 - Databricks menú principal

e. Se realiza configuración del Clúster de procesamiento (poder de cómputo)

   i. Se da clic en menú “Compute”

   ii. Se da clic en botón “Create Clúster”

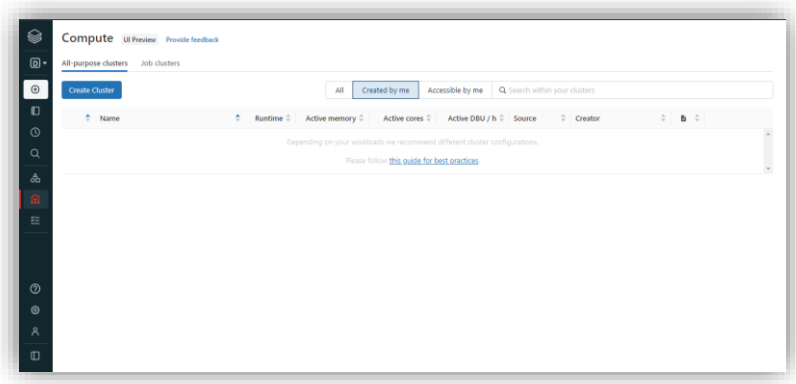


Ilustración 37 - Databricks creación de clúster

iii. Se procede con la configuración del Clúster asignando un nombre, la versión del runtime de Databricks, y opciones requeridas según se desee, posteriormente se da clic en botón “Create Clúster”. Para este proyecto se manejó la configuración que se muestra en la siguiente imagen.

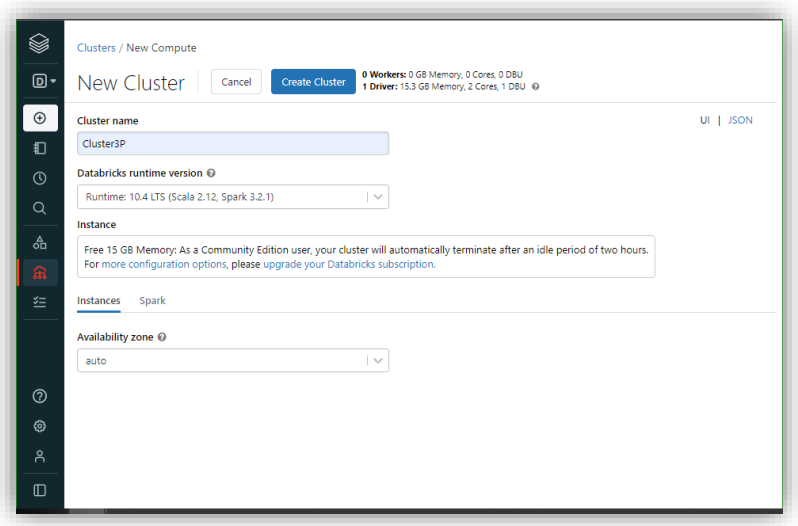


Ilustración 38 - Databricks configuración de clúster

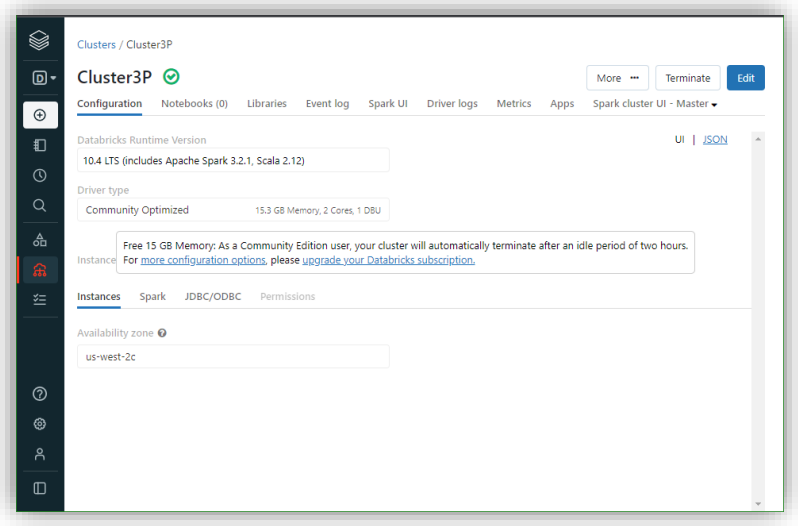


Ilustración 39 - Databricks confirmación de clúster configurado

f. Se realiza instalación de librería XGBOOST

i. Se da clic sobre el enlace del nombre del Clúster

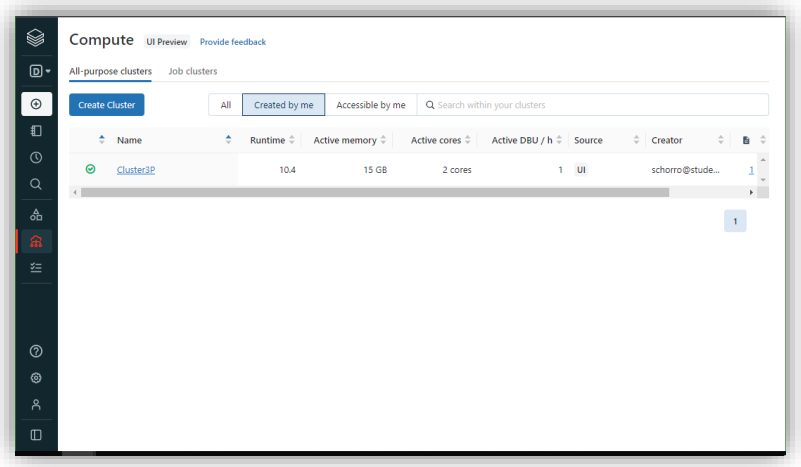


Ilustración 40 - Databricks selección de clúster

ii. Se selecciona el tab de Libraries

iii. Se da clic sobre el botón “Install New”

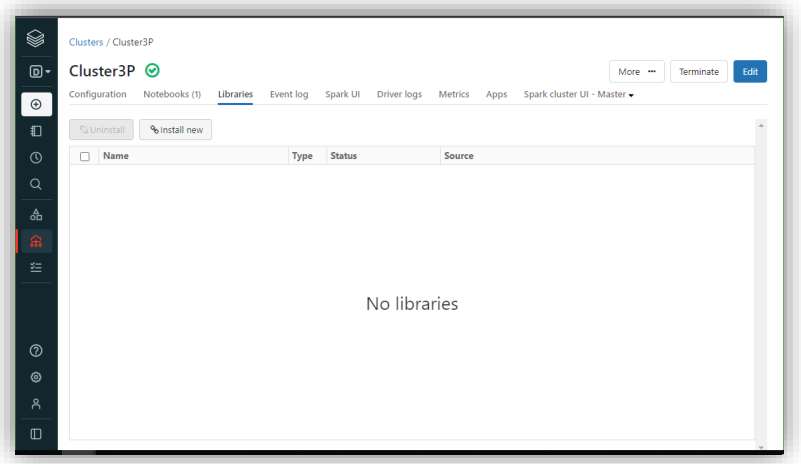


Ilustración 41 - Databricks instalación de librerías

iv. Se procede a indicar el nombre del paquete que se va a instalar, en este caso “xgboost”, que es la librería utilizada para la creación del modelo que se construye como producto de este proyecto, y se da clic en el botón “Install”

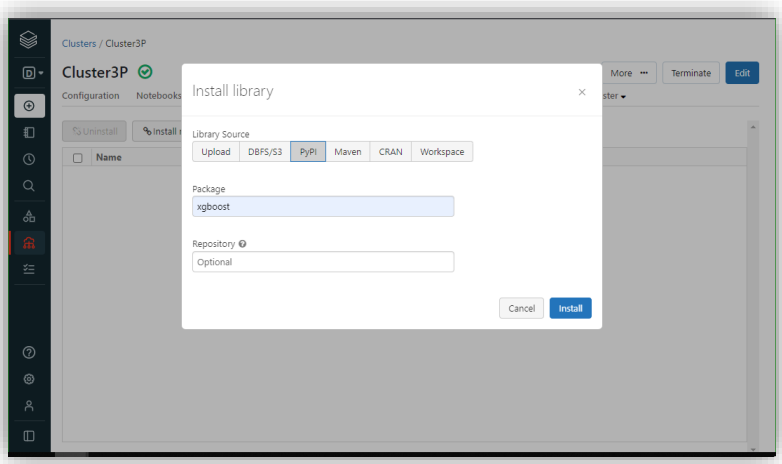


Ilustración 42 - Databricks instalación de librería XGBOOST

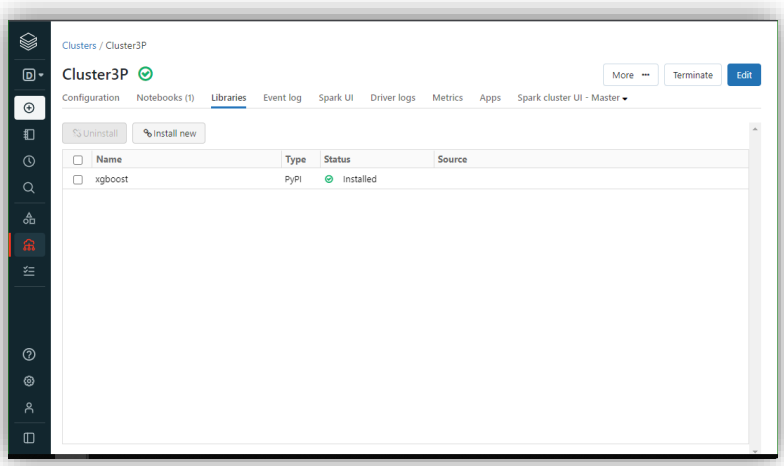


Ilustración 43 - Databricks confirmación de librería instalada

g. Se crea Notebook y se asocia al Clúster creado en paso “e”

i. Creación de Notebook

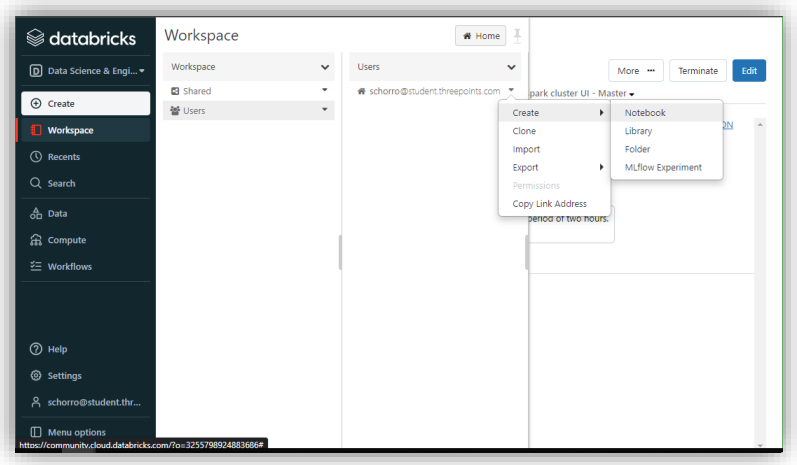


Ilustración 44 - Databricks creación de Notebook

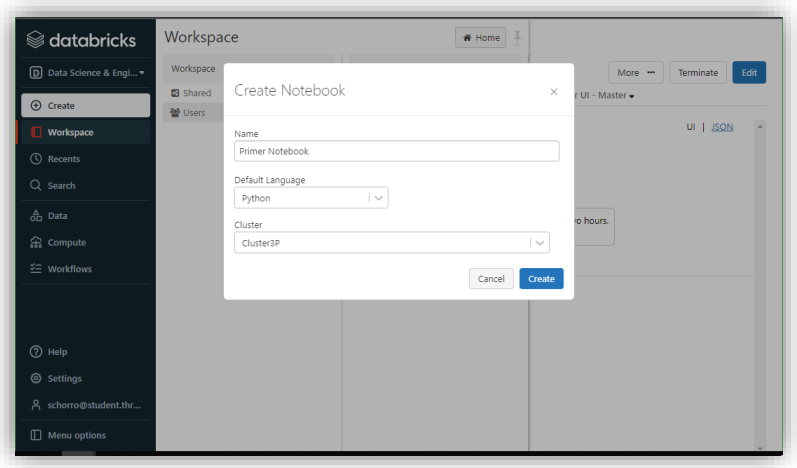


Ilustración 45 - Databricks asignación de nombre al Notebook

ii. Se da inicio con la etapa de construcción teniendo las opciones de utilizar los lenguajes: Markdown, Python, SQL, Scala y R

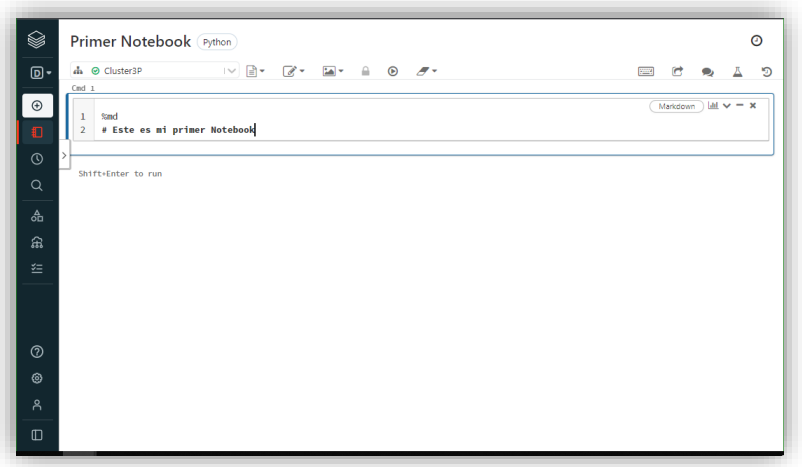


Ilustración 46 - Databricks con Notebook creado

Tablero en Power BI

Conexión a Databricks

Power BI presenta un conector para poder consumir información desde Databricks por lo que se decidió utilizarlo para poder realizar lectura y análisis del output del ETL posterior a la ejecución del Modelo.

En los siguientes pasos se muestran imágenes de cómo conectar Power BI con Databricks

1. Databricks
    a. En la herramienta se selecciona el menú de compute
    b. Se da clic en el nombre del Clúster que se haya configurado, este debe estar encendido.
    c. Se selecciona la viñeta de JDBC/ODBC como se muestra en la imagen

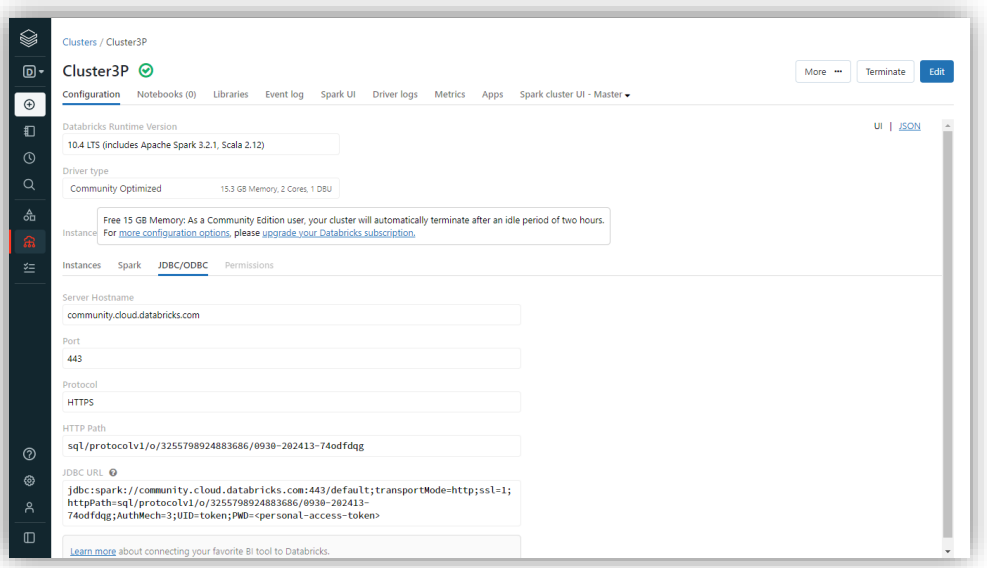


Ilustración 47 - Databricks - Conexión de JDBC/ODBC

d. Se identifican los valores asociados a **Server Hostname** y **HTTP Path**, los cuales se utilizan para configurar la conexión desde Power BI

2. Power BI

a. Se selecciona el menú de **Get data**

b. Se da clic en More

c. Se selecciona el conector Databricks (a este momento se mantiene en versión Beta) y se da clic en Connect

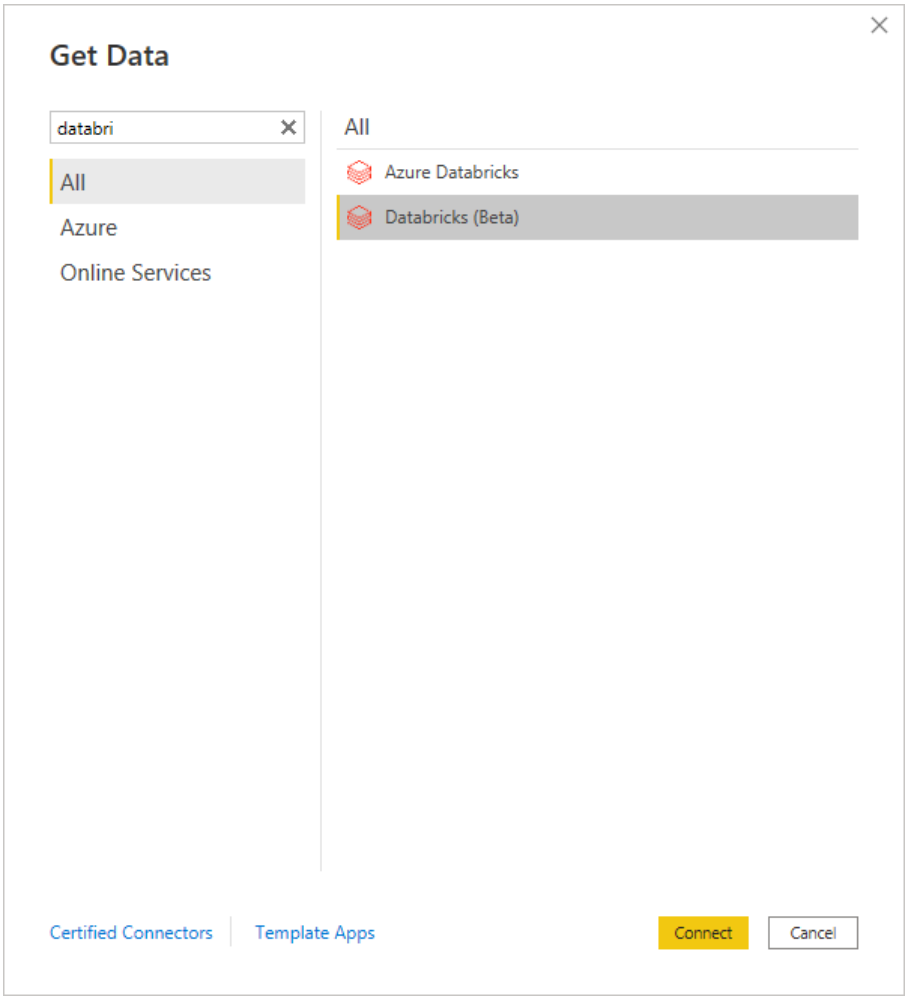


Ilustración 48 - Power BI menú Get Data

d. Se ingresan los parámetros correspondientes a la conexión JDBC/ODBC de Databricks tal cual se muestra en la siguiente imagen:

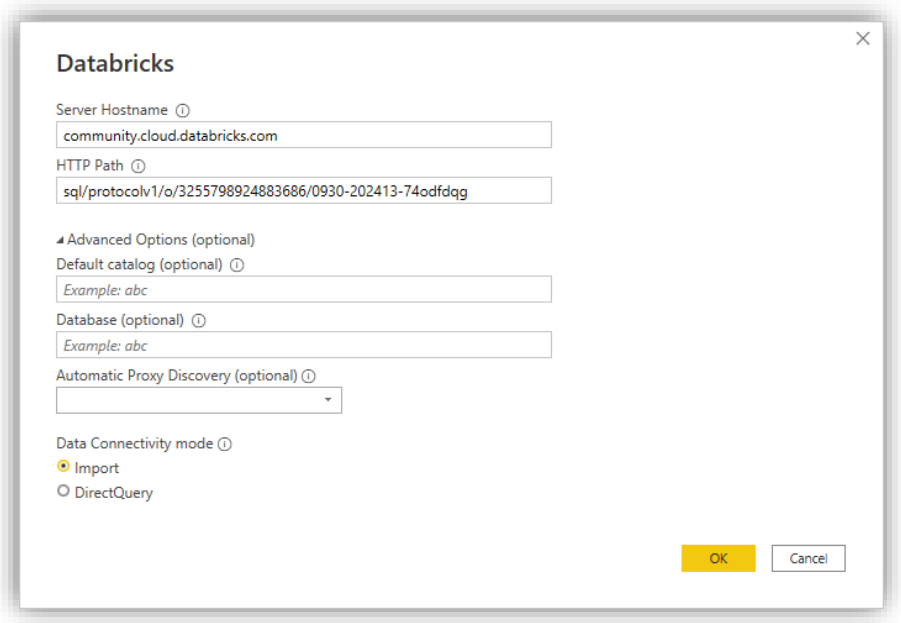


Ilustración 49 – Power BI Configuración de conexión

e. Se ingresa usuario y contraseña de Databricks

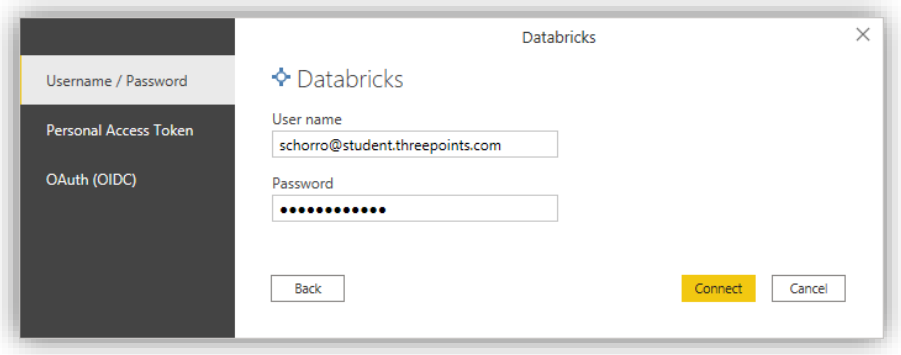


Ilustración 50 - Power BI Ingreso de credenciales de Databricks

f. Se seleccionan las tablas que se desean importar al modelo de Power BI

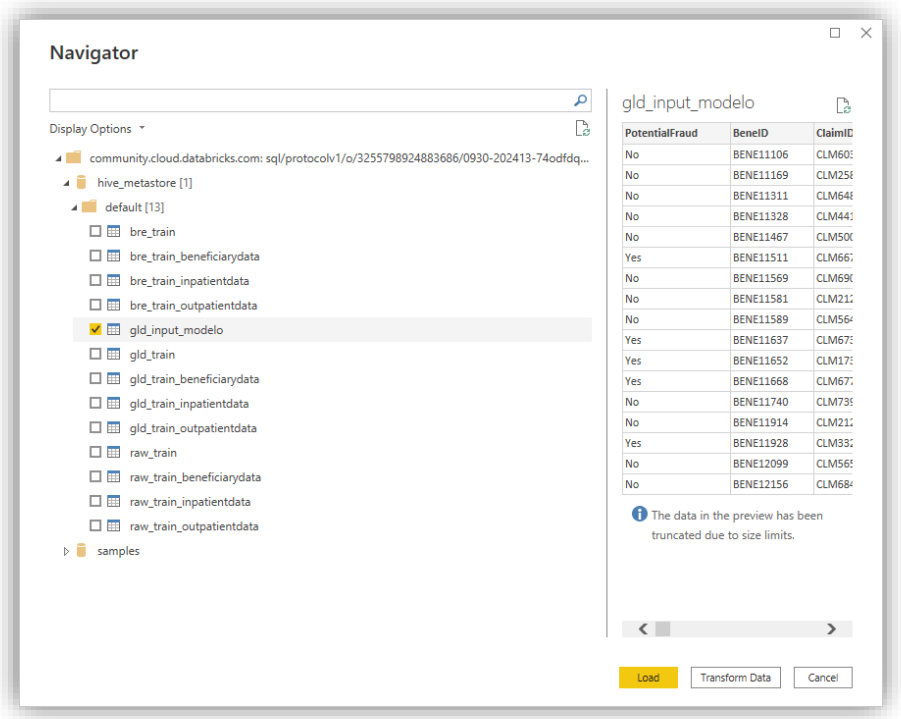


Ilustración 51 - Power BI Importación de tablas desde Databricks

Dashboard construido "Solicitudes de atención médica a ser reembolsados – enfoque volumétrico"

En este apartado se explicará el dashboard que permite visualizar los resultados después de aplicar el modelo de ML y generada la clasificación que determina si las solicitudes son fraudulentas. Dicho dashboard fue construido en Power BI el cual se alimenta de la salida del ETL posterior a la ejecución del modelo. Se presentará dos hojas en las que se visualizan las estadísticas de todas las solicitudes, pero se hará énfasis en el análisis de las solicitudes consideradas "fraudulentas" y desde el punto de vista volumétrico (cantidad de solicitudes). Las hojas (visualizaciones) tienen dos enfoques; el primero será en referencia a la cantidad de solicitudes fraudulentas de los asegurados y la segunda sobre los montos desembolsados (defraudación) a la aseguradora, los datos se analizan básicamente en tres categorías: Los proveedores que se encuentran en cada solicitud, los médicos tratantes y los diagnósticos aceptados.

En la siguiente imagen se puede visualizar el dashboard que consta de tres secciones con distintos objetivos: Filtros, gráficos y tabla de solicitudes.

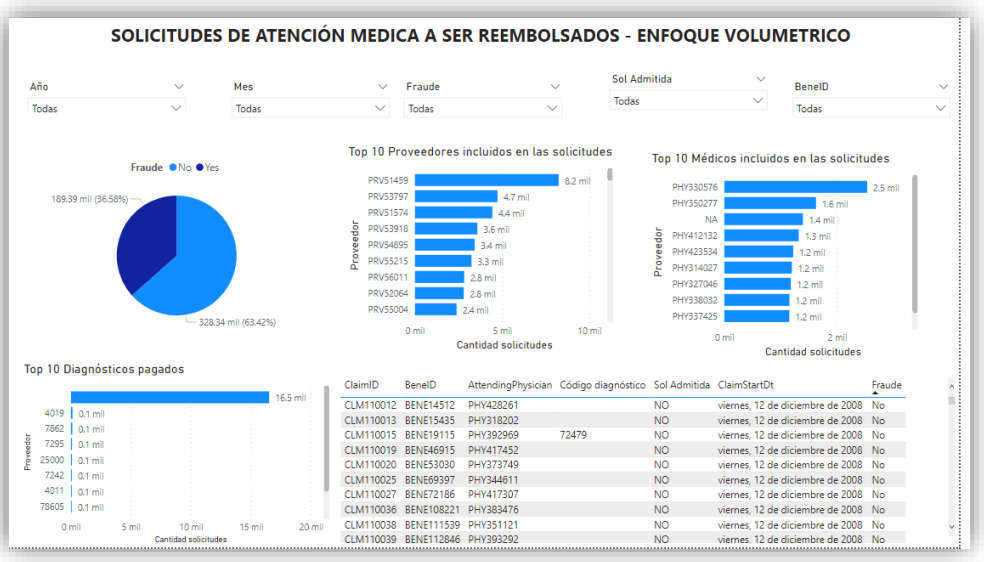


Ilustración 52 - Dashboard "Solicitudes de atención médica a ser reembolsadas - enfoque volumétrico"

Filtros

En la *Ilustración 53 - Filtros en Dashboard* se muestran los filtros que servirán para enfocar nuestra atención en un subconjunto de datos, se han dejado disponibles los siguientes:

- Año: Se refiere al año en el que fue ingresada la solicitud.
- Mes: Es el mes en el que se ingresó la solicitud.
- Fraude: Es el dato que indica si se considera como fraude dicha solicitud, este dato es el resultado de la evaluación del modelo de los datos de la solicitud de reembolso ingresado.
- Sol Admitida: Este dato nos indica si la solicitud de reembolso fue admitida.
- BeneID: ID del beneficiario.

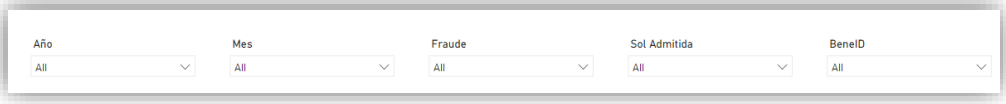


Ilustración 53 - Filtros en Dashboard

Gráficos

Los gráficos nos permitirán ver la distribución de las solicitudes tanto en las categorías como en su distribución en el tiempo.

En la siguiente imagen podemos apreciar la distribución de las solicitudes por la clasificación que da el modelo de machine learning, si se considera como una solicitud fraudulenta.

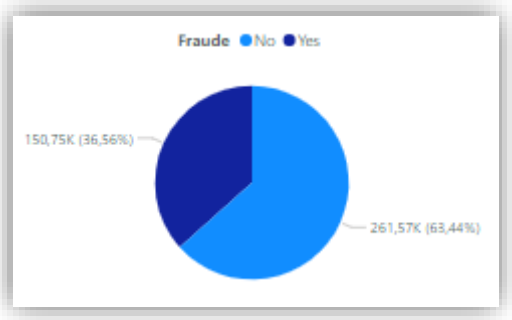


Ilustración 54 - Distribución de solicitudes consideradas fraudulentas y reales

A partir de este punto el análisis se realiza sobre las solicitudes consideradas fraudulentas y que la aseguradora hizo el desembolso; en la siguiente imagen, podemos apreciar el top de 10 proveedores en el cual fueron atendidos los asegurados que presentaron solicitud de reembolso, de esta forma podemos visualizar que proveedores atienden el mayor volumen de solicitudes.

Ilustración 55 - Top 10 proveedores incluidos en las solicitudes

En la siguiente imagen podemos visualizar el top de los 10 médicos responsables que trataron a los asegurados que presentaron solicitudes de reembolso. Con esta información podemos tener un listado de médicos a los cuales hacer revisiones más exhaustivas de sus atenciones.

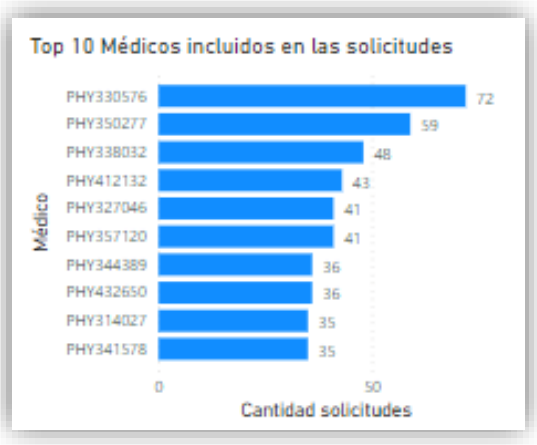


Ilustración 56 - Top 10 médicos incluidos en las solicitudes

En la siguiente imagen podemos visualizar el top de los 10 diagnósticos más recurrentes presentados en las solicitudes de reembolso. En ella podemos apreciar que dentro de las solicitudes que fueron reembolsadas, el grupo que mayor reembolso generó fueron los que no tuvieron un diagnóstico registrado; y si hacemos drill down, a su vez, el tablero permite identificar el Top 10 de los médicos tratantes que no registraron un diagnóstico en la atención realizada al asegurado.

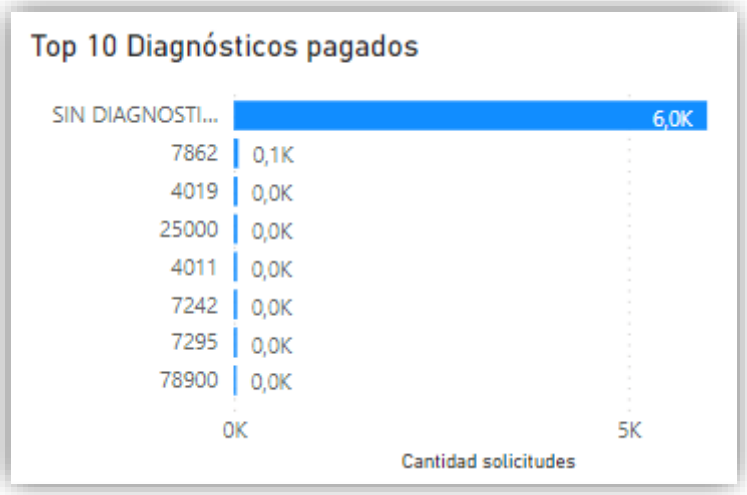


Ilustración 57 - Top 10 diagnósticos incluidos en las solicitudes

Ilustración 58 - Top 10 médicos que no registraron diagnóstico en solicitudes reembolsadas

En la siguiente imagen, podemos apreciar una tabla que nos permitirá ver cuáles son los principales datos de las solicitudes a fin de analizar los reembolsos realizados y las características de las solicitudes filtradas. Esta tabla muestra: código de reclamo, id del beneficiario, médico principal tratante, código de diagnóstico, si la solicitud fue reembolsada, fecha de ingreso de la solicitud, si se considera fraudulenta y datos del beneficiario como el país, estado, raza y género.

ID	ID Beneficiario	Médico	Código diagnóstico	Reembolsado	Fecha	Fraude	County	State	Race	Gender
CLM168936	BENE72695	PHY381249		SI	sábado, 31 de enero de 2009	Yes	100	10	1	2
CLM169091	BENE96660	PHY407877		SI	sábado, 31 de enero de 2009	Yes				
CLM169200	BENE113322	PHY329600		SI	sábado, 31 de enero de 2009	Yes	630	10	2	2
CLM169249	BENE120188	PHY384216		SI	sábado, 31 de enero de 2009	Yes				
CLM169362	BENE138289	PHY326000		SI	sábado, 31 de enero	Yes	440	49	2	1

Ilustración 59 - Detalle de solicitudes

Dashboard construido "Solicitudes de atención médica a ser reembolsados – enfoque económico"

A diferencia del dashboard anterior, en este presentaremos la información desde el punto de vista económico orientado a la pérdida por los desembolsos de solicitudes de atención fraudulentas, los datos también se analizarán en tres categorías: Los proveedores que se encuentran en cada solicitud, los médicos tratantes y los diagnósticos aceptados.

En la siguiente imagen se puede visualizar el dashboard que consta de tres secciones con distintos objetivos: Filtros, gráficos y tabla de solicitudes.

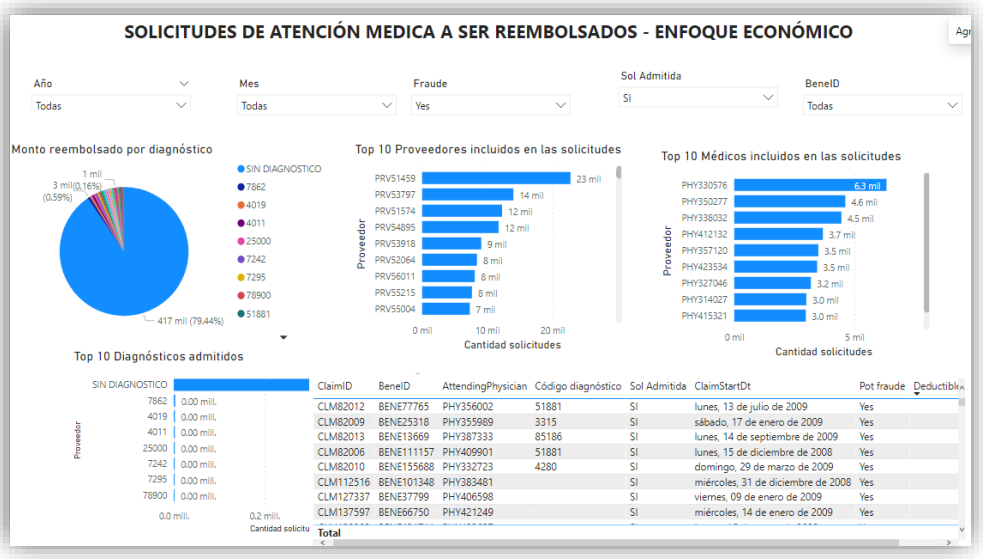


Ilustración 60 - Dashboard "Solicitudes de atención médica a ser reembolsadas - enfoque económico"

Filtros

Como los filtros son los mismos no los describiremos en esta sección.

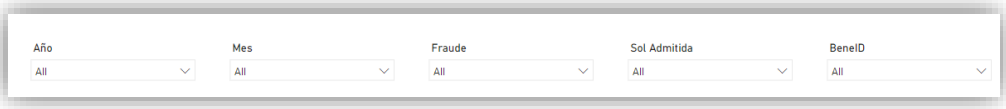


Ilustración 61 - Filtros en Dashboard

Gráficos

Los gráficos nos permitirán ver como las solicitudes fraudulentas afectan económicamente a la aseguradora y ayudará a analizar y definir controles para evitar ser sorprendidos.

En la siguiente imagen podemos apreciar los montos que fueron desembolsados por la aseguradora y de ello se aprecia que el 79.44% fueron por atenciones sin un diagnóstico.

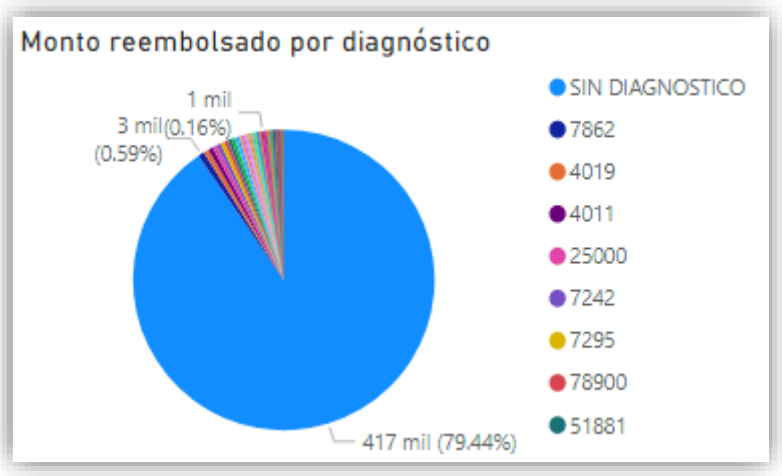


Ilustración 62 - Distribución de reembolsos por diagnóstico de solicitudes fraudulentas

en la siguiente imagen, podemos apreciar los reembolsos que se realizaron clasificados por los proveedores en orden descendente, de esta forma podemos saber desde que proveedor se reciben las solicitudes fraudulentas que causaron mayores pérdidas.

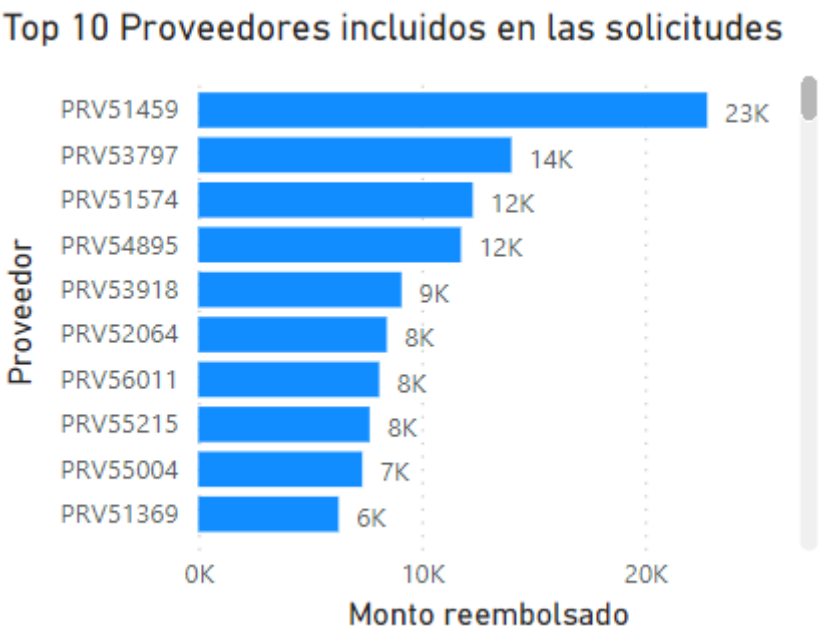


Ilustración 63 - Top 10 proveedores con mayores reembolsos

En la siguiente imagen podemos visualizar el top de los 10 médicos responsables que trataron a los asegurados y que generaron las mayores pérdidas para la aseguradora.

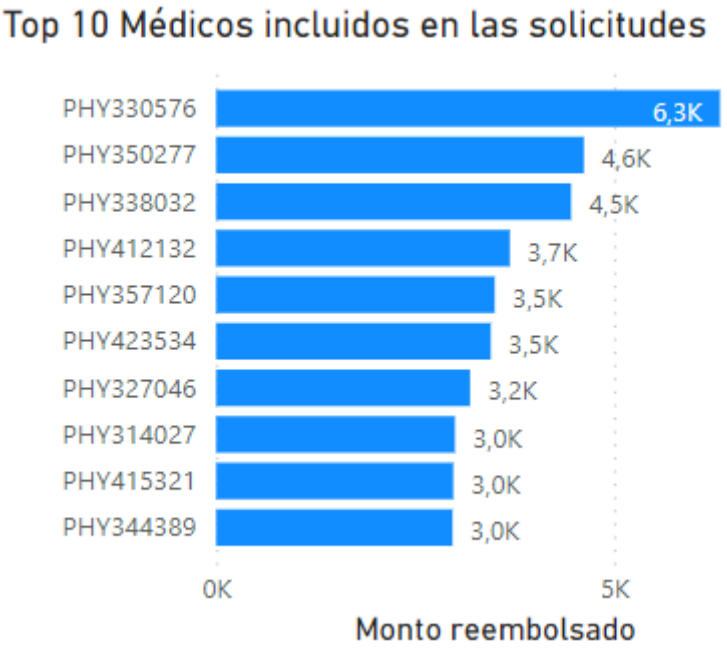


Ilustración 64 - Top 10 médicos incluidos en las solicitudes mayores pérdidas

En las figuras anteriores vimos que el 79.44% de los reembolsos no tienen un diagnóstico, en esta imagen podemos ver las pérdidas que provienen de estas

solicitudes sin diagnóstico que ascienden a 0.42 Millones, y si entramos al detalle podemos ver el top de los 10 médicos que no incluyeron diagnóstico y que generaron los mayores reembolsos.

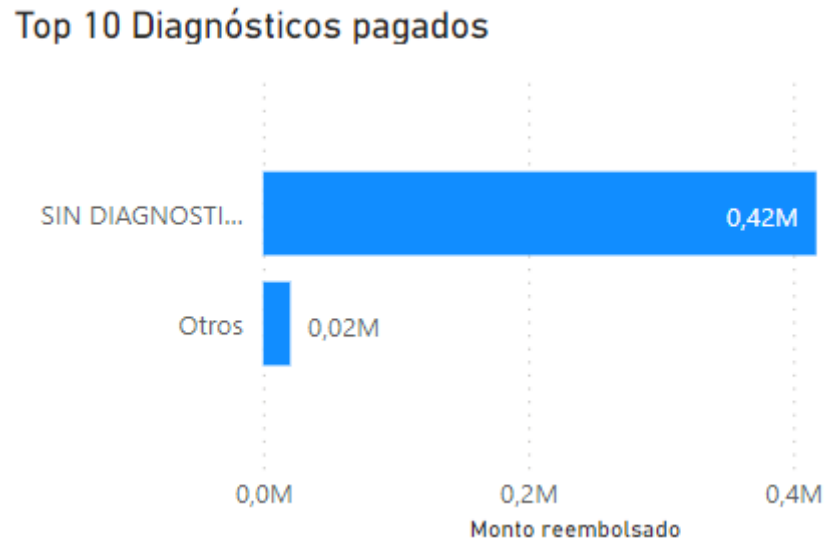


Ilustración 65 - Top 10 diagnósticos incluidos en las solicitudes con mayores reembolsos

Ilustración 66 - Top 10 médicos que no registraron diagnóstico en solicitudes de mayores reembolsos

Finalmente vemos la tabla con los datos de las solicitudes que se están visualizando en los distintos filtros a medida que fuimos navegando en los gráficos, esta tabla es la misma mencionada anteriormente por ello no mencionaremos mayor detalle.

ID	ID Beneficiario	Médico	Código diagnóstico	Reembolsado	Fecha	Fraude	County	State	Race	Gender
CLM168936	BENE72695	PHY381249		SI	sábado, 31 de enero de 2009	Yes	100	10	1	2
CLM169091	BENE96660	PHY407877		SI	sábado, 31 de enero de 2009	Yes				
CLM169200	BENE113322	PHY329600		SI	sábado, 31 de enero de 2009	Yes	630	10	2	2
CLM169249	BENE120188	PHY384216		SI	sábado, 31 de enero de 2009	Yes				
CLM169362	BENE138289	PHY326000		SI	sábado, 31 de enero	Yes	440	49	2	1

Ilustración 67 - Detalle de solicitudes