UNIVERSIDAD DEL VALLE DE GUATEMALA
Facultad de Ingeniería


# Detección de posible fraude por proveedores en reclamos de salud

Trabajo de graduación presentado por Jose Javier Jo Escobar para optar al grado académico de Licenciado en Ingeniería en Ciencias de la Computación y Tecnologías de la Información

Guatemala,

2020

UNIVERSIDAD DEL VALLE DE GUATEMALA
Facultad de Ingeniería


# Detección de posible fraude por proveedores en reclamos de salud

Trabajo de graduación presentado por Jose Javier Jo Escobar para optar al grado académico de Licenciado en Ingeniería en Ciencias de la Computación y Tecnologías de la Información

Guatemala,

2020

Vo.Bo.:

(f) _____

Ing. Roberto Gomez

Tribunal Examinador:

(f) _____

Ing. Roberto Gomez

(f) _____

MSc. Douglas Barrios

(f) _____

Ing. Tomas Gálvez

Fecha de aprobación: Guatemala, 14 de Diciembre de 2020.

# Prefacio

El mercado asegurador no siempre me llamó la atención, sin embargo, en el 2016 por motivos de salud tuve que visitar un hospital y hacer uso de mi seguro de salud personal. Ese día tuve que esperar mucho tiempo para autorizaran mi reclamo. Desde entonces, surgió un proyecto con el fin de agilizar los procesos realizados por las aseguradoras ya que se toman mucho tiempo en realizar. En ese momento estaba cursando Ingeniería de software 1 en la universidad por lo que propuse el tema para proyecto del curso. El proyecto fue desarrollado en los cursos de Ingeniería de software 1 y 2, Interacción Humano Computador y Administración de proyectos, pero lo detuve porque empecé a trabajar y ya no me quedaba tiempo para nada. Pero, cuando se presentó la oportunidad de darle seguimiento al proyecto como trabajo de graduación lo tomé con mucha emoción, ya que es un tema que me interesa mucho y podría ayudar tanto al sector privado cómo el público.

Con respecto al proyecto, de las últimas fases que tenía pensado elaborar era la digitalización de los formularios y la implementación de machine learning para la automatización de los procesos de autorización. Por lo que inicie a buscar la información necesaria para elaborar mi primer prototipo de automatización de procesos de reclamo, buscando ayuda en aseguradoras guatemaltecas. En la búsqueda de la información platique con varios ejecutivos y conocedores de la industria aseguradora identificando que el fraude es uno de los problemas que retrasan el proceso de autorización y aumentan los costos. Es por ello que me pareció una muy buena idea identificar e implementar un modelo de machine learning que detecte con la mayor precisión el posible fraude por proveedores en reclamos de salud.

Quiero agradecer a mis padres por apoyarme desde el principio en mi proceso académico, a mis hermanos por estar siempre a mi lado y a mi novia por motivarme y ayudarme siempre. También agradecerle a mi asesor el Ingeniero Roberto Gomez por transmitirme sus conocimientos en la industria aseguradora e incentivarme a realizar un trabajo de calidad. Este camino no fue fácil, pero estuvieron motivando y ayudando hasta donde sus alcances lo permitía. Se los agradezco mucho, esto va por ustedes.

# Prefacio

El mercado asegurador no siempre me llamó la atención, sin embargo, en el 2016 por motivos de salud tuve que visitar un hospital y hacer uso de mi seguro de salud personal. Ese día tuve que esperar mucho tiempo para autorizaran mi reclamo. Desde entonces, surgió un proyecto con el fin de agilizar los procesos realizados por las aseguradoras ya que se toman mucho tiempo en realizar. En ese momento estaba cursando Ingeniería de software 1 en la universidad por lo que propuse el tema para proyecto del curso. El proyecto fue desarrollado en los cursos de Ingeniería de software 1 y 2, Interacción Humano Computador y Administración de proyectos, pero lo detuve porque empecé a trabajar y ya no me quedaba tiempo para nada. Pero, cuando se presentó la oportunidad de darle seguimiento al proyecto como trabajo de graduación lo tomé con mucha emoción, ya que es un tema que me interesa mucho y podría ayudar tanto al sector privado cómo el público.

Con respecto al proyecto, de las últimas fases que tenía pensado elaborar era la digitalización de los formularios y la implementación de machine learning para la automatización de los procesos de autorización. Por lo que inicie a buscar la información necesaria para elaborar mi primer prototipo de automatización de procesos de reclamo, buscando ayuda en aseguradoras guatemaltecas. En la búsqueda de la información platique con varios ejecutivos y conocedores de la industria aseguradora identificando que el fraude es uno de los problemas que retrasan el proceso de autorización y aumentan los costos. Es por ello que me pareció una muy buena idea identificar e implementar un modelo de machine learning que detecte con la mayor precisión el posible fraude por proveedores en reclamos de salud.

Quiero agradecer a mis padres por apoyarme desde el principio en mi proceso académico, a mis hermanos por estar siempre a mi lado y a mi novia por motivarme y ayudarme siempre. También agradecerle a mi asesor el Ingeniero Roberto Gomez por transmitirme sus conocimientos en la industria aseguradora e incentivarme a realizar un trabajo de calidad. Este camino no fue fácil, pero estuvieron motivando y ayudando hasta donde sus alcances lo permitía. Se los agradezco mucho, esto va por ustedes.

# Índice

# Lista de figuras

# Lista de cuadros

# Resumen

Conforme pasa el tiempo, en EE.UU. los avances en tecnología y ciencias médicas mejoran la calidad de vida de la población. Esto implica la necesidad de programas como Medicare que ayuden en la administración de los altos costos asociados a la atención médica de calidad. Su problemática es la existencia de personas que cometen fraude para beneficio personal, lo que reduce la cobertura de Medicare para satisfacer de manera efectiva las necesidades de atención médica de la población que califica. En busca de reducir el fraude, los Centros de Servicios de Medicare y Medicaid (CMS) compartieron conjuntos de datos que abarcan diferentes partes del programa Medicare. En esta investigación, nos enfocamos en la implementación de ***machine learning*** para la detección de posible fraude por proveedores en reclamos de salud de Medicare. Para ello se utilizó el conjunto de datos DE-SynPUF de CMS que contiene: (1) Listado de beneficiarios del 2008 de Medicare, (2) Reclamos realizados por beneficiarios hospitalizados de 2008 al 2010, (3) Reclamos realizados por beneficiarios ambulatorios de 2008 al 2010. Además, se creó un cuarto conjunto de datos que es la combinación de los tres conjuntos de datos mencionados. Se realizó minería de datos en los cuatro conjuntos y el etiquetado de fraude de proveedores del mundo real utilizando la Lista de Personas y Entidades Excluidas (LEIE) de la Oficina de Inspectoría General. El resultado de la preparación de los datos fue utilizado para la creación y evaluación de cuatro algoritmos de ***machine learning***. Estos fueron evaluados con base a la exactitud, puntuación f1 y puntuación AUROC. Los algoritmos supervisados utilizaron cómo entrenamiento el conjunto de datos combinado con su etiquetado de fraude a diferencia de los no supervisados que utilizaron únicamente reclamos no fraudulentos. Los resultados finales de la investigación muestran que el algoritmo de regresión logística (LR) tuvo un desempeño consistente en la detección de posible fraude con una exactitud de  0.90, una puntuación AUROC 0.80 y una puntuación F1  0.59. Por lo tanto, dados los resultados se sugiere utilizar el conjunto de datos combinado para detectar comportamientos fraudulentos por proveedores de Medicare. También se recomienda agregar más registros fraudulentos en el entrenamiento para mejorar la precisión de predicción.

**PALABRAS CLAVE**— Aprendizaje Automático, Medicare, Regresion Logistica, Bosques Aleatorios, *Autoencoder*, Detección de fraude, Puntuación F1, AUROC, Exactitud.

# Abstract

Over time, advances in technology and medical sciences in the US improve the quality of life of the population. This implies the need for programs like Medicare to help manage the high costs associated with quality health care. Its problem is the existence of people who commit fraud for personal gain, reducing Medicare coverage to effectively meet the health care needs of the population that qualifies. In an effort to reduce fraud, the Centers for Medicare  Medicaid Services (CMS) shared data sets that span different parts of the Medicare program. In this research, we focused on the implementation of machine learning for the detection of possible fraud by providers in Medicare health claims. For this, the CMS DE-SynPUF dataset was used, which contains: (1) List of Medicare beneficiaries from 2008, (2) Claims made by hospitalized beneficiaries from 2008 to 2010, (3) Claims made by outpatient beneficiaries from 2008 to 2010. In addition, a fourth data set was created which is the combination of the three mentioned data sets. Data mining was performed on all four sets and real-world supplier fraud tagging using the List of Excluded Persons and Entities (LEIE) from the Office of Inspector General. The result of the data preparation was used for the creation and evaluation of four machine learning algorithms. They were evaluated based on accuracy, f1 score, and AUROC score. The supervised algorithms used the dataset combined with its fraud labeling as training, as opposed to the unsupervised ones that used only non-fraudulent claims. The final results of the investigation show that the logistic regression (LR) algorithm had a consistent performance in detecting possible fraud with an accuracy of  0.90, an AUROC score of  0.80 and an F1 score  0.59. Therefore, given the results, it is suggested to use the combined data set to detect fraudulent behavior by Medicare providers. It is also recommended to add more fraudulent records in training to improve prediction accuracy.

**KEYWORDS**— Machine Learning, Medicare, Logistic Regresion, Random Forest, Autoencoder, Fraud detection, F1-Score, AUROC, Accuracy.

# CAPÍTULO 1

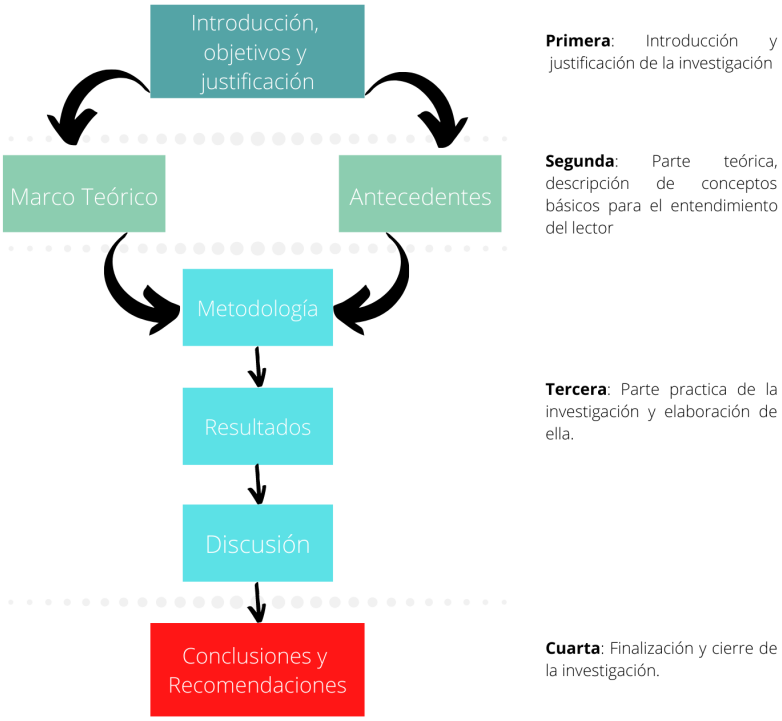
## Introducción

Actualmente el fraude es uno de los principales problemas que enfrentan los sistemas de seguros médicos y sanitarios, la RAE (2020) define fraude como: acción contraria a la verdad y a la rectitud, que perjudica a la persona contra quien se comete. Este es un problema al cual se enfrentan las aseguradoras debido a la falta de profesionalidad de algunos médicos y proveedores, impacta en los costos de operación generados por los reclamos de los asegurados en el rubro de la salud. Las aseguradoras se han dado la tarea de analizar numerosos registros para encontrar conductas posiblemente sospechosas y fraudulentas [46], este proceso actualmente es ineficiente ya que se realiza de manera manual [34,35].

Esta investigación tiene como objetivo proponer y desarrollar un modelo de *machine learning* que detecte de manera precisa el posible fraude realizado por proveedores en reclamos de salud. También busca utilizar la ingeniería de características ya que esta técnica mejora la precisión de predicción realizada por el modelo. Se busca desarrollar diferentes algoritmos de categorización realizando el preprocesamiento la información necesaria para su entrenamiento y evaluación, logrando mostrar su precisión en la detección de fraude. Finalmente se determinará el modelo que detectó con mayor precisión el posible fraude.

La metodología para alcanzar este objetivo es utilizada normalmente en proyectos de *machine learning*, esta permitirá desarrollar el modelo que se adapte a los requisitos de detección de posible fraude. Para una comprensión correcta de la investigación, esta se dividió en cuatro fases (Figura 1).

Figura 1: Estructura de investigación




Fuente: Elaboración propia

1. Primer fase: Representa la introducción y justificación de la razón de ser de la investigación.

2. Segunda fase: Se definieron los conceptos básicos necesarios para el entendimiento futuro de la investigación por medio del marco teórico, incluyendo la metodología utilizada para el desarrollo del modelo que detecte con la mayor precisión el posible fraude.

3. Tercera fase: Se definió la metodología a utilizar, se entrenaron y evaluaron cuatro modelos para la detección de posible fraude y se realizó la discusión de los resultados obtenidos. Para ello, se hizo uso de la siguiente metodología, la cual se divide en las siguientes etapas:

    - Definición del problema: Se definió el problema a resolver con machine learning.
    - Investigación: Se investigaron los modelos de machine learning candidatos para solucionar el problema de categorización definido en el paso anterior.
    - Minería de datos: Se realizó limpieza y análisis exploratorio del conjunto de datos.
    - Preprocesamiento de datos: Se formatearon y estandarizaron los datos, también se realizaron conversión de variables y separados los conjuntos de entrenamiento y prueba.
    - Implementación de modelo: Se definieron las bibliotecas y modelos a utilizar en posteriores fases.
    - Entrenamiento: Se realizó el entrenamiento de los modelos de regresión logística, ***random forest*** y ***autoencoder***.
    - Evaluación: Se evaluó el rendimiento de los modelos entrenados.

- Predicción: Se verificaron los verdaderos positivos y falsos positivos de la predicción realizada por cada modelo.

4. Cuarta fase: Se cierra con las conclusiones pertinentes a la investigación realizada y sus recomendaciones.

El modelo de *machine learning* que detectó con mayor precisión el posible fraude por proveedores es el de regresión logística, ya que tuvo un exactitud de 0.91 lo que nos indica que identificó con un 91 % de certeza los casos positivos correctamente identificados de todos los casos predichos. También tuvo una puntuación F1 de 0.59 la cual nos proporciona una mejor medida de los casos clasificados incorrectamente siendo crucial obtener una puntuación alta en un conjunto de datos desequilibrado.

# CAPÍTULO 2

## Antecedentes

A medida que avanza la tecnología y aumenta su uso, también lo hace la capacidad de aplicar la ciencia de datos y aprendizaje automático en *Big Data*, lo que puede mejorar el estado de los programas de seguro médico y de atención médica para que los pacientes reciban atención médica de calidad.

Los centros de servicios de Medicare y Medicaid (CMS) se unieron a este esfuerzo al publicar conjuntos de datos de Medicare de "*Big Data*" para ayudar a identificar el fraude dentro de Medicare [22]. Hay varios conjuntos de datos disponibles en el sitio web de los Centros de Servicios de Medicare y Medicaid [16].

Han habido una serie de estudios realizados, por grupos de investigación y otros, utilizando datos de archivos de uso público (PUF) de CMS para evaluar posibles actividades fraudulentas a través de la minería de datos y otros métodos de análisis. La gran mayoría de estos estudios utiliza datos de la Parte B el cual proporciona únicamente información sobre los reclamos y procedimientos realizados por un médico en un año determinado [3,5,6,8,22,26], sin tener en cuenta otras partes de Medicare al detectar conductas fraudulentas. Dentro de los sistemas de salud, o cualquier lugar donde se intercambie dinero, existe la oportunidad de que un mal actor manipule el proceso y desvíe fondos, lo que afectaría la eficiencia y eficacia del proceso de atención de salud de Medicare.

Existe información previa limitada donde se incrementa el riesgo de fraude dentro del sistema de Medicare, por lo que la elección de una sola parte de Medicare podría evitar el fraude cometido en otro lugar. En esta investigación, se enfoca en el procesamiento y etiquetado del conjunto DE-SynPUF de Medicare y el desempeño sobre la detección de fraude. Por lo tanto, generalmente se limitará la discusión a este pequeño grupo de trabajos que intentan identificar comportamientos fraudulentos utilizando múltiples conjuntos de datos de CMS. A partir de esta investigación, solo se encontraron dos trabajos [9,48] que entran en esa categoría.

**Análisis gráfico para la estimación del riesgo de fraude en la atención médica. (2016)**

Branting [9], utiliza el conjunto de datos de la Parte B (2012-2014), la parte D (2013) y LEIE. El conjunto de datos de la parte D proporciona información relacionada con los medicamentos recetados administrados bajo el programa de medicamentos recetados de Medicare. Los autores no mencionan específicamente cómo procesan previamente los datos o combinan la Parte B y la Parte D, pero toman atributos de los conjuntos de datos de la Parte B y la Parte D, y tratan los medicamentos y los códigos HCPCS de la misma manera. Emparejaron a 12,153 médicos fraudulentos utilizando el Identificador Nacional de Proveedores (NPI) [17] con su algoritmo de coincidencia de única identidad. Decidieron no distinguir entre las reglas/códigos de exclusión LEIE y en su lugar utilizaron a todos los médicos enumerados. No está claro si los autores tuvieron en cuenta las exenciones, las fechas de inicio de exclusión o la duración de la exclusión asociada durante su proceso de mapeo de etiquetas de fraude. Estos detalles son importantes para reducir las etiquetas de exclusión redundantes y superpuestas y para evaluar el rendimiento preciso de la detección de fraudes. Por lo tanto, debido a esta falta de claridad en la metodología de etiquetado de exclusión, los resultados de su estudio no se pueden reproducir de manera confiable y pueden ser difíciles de comparar con otras investigaciones. Desarrollaron un método para identificar comportamientos fraudulentos mediante la determinación del riesgo de fraude aplicando algoritmos de red a partir de grafos. Debido a la naturaleza altamente desequilibrada de los datos, los autores utilizaron una distribución de clases 50:50, reteniendo 12,000 proveedores excluidos mientras seleccionan al azar 12,000 proveedores no incluidos. Presentaron algunos grupos de algoritmos y determinaron sus resultados de detección de fraude basándose en los médicos fraudulentos del mundo real encontrados en el conjunto de datos LEIE. Un conjunto de algoritmos, que denotan cómo similitud Comportamiento-Vector, determina la similitud en el comportamiento de los médicos fraudulentos y no fraudulentos del mundo real, utilizando valores nominales cómo recetas de medicamentos y procedimientos médicos. Otro grupo de algoritmos conforma su propagación de riesgo, que utiliza la ubicación geoespacial conjunta para estimar la programación del riesgo de proveedores fraudulentos de atención médica.

**Extracción de anomalías en conjunto de datos de Medicare utilizando utilizando el método de inducción de la regla del paciente (2017)**

Sadiq [48], utiliza los conjuntos de datos de 2014 CMS Parte B, Parte D y DMEPOS (utilizando solo los reclamos de proveedores de Florida) para encontrar anomalías que posiblemente apunten a comportamientos fraudulentos u otros comportamientos interesantes. Los autores no profundizan en cómo desarrollaron el preprocesamiento de sus conjuntos de datos. A partir de su estudio, podemos asumir que los autores utilizan, cómo mínimo, las siguientes características: NPI, género, ubicación (estado, ciudad, dirección , etc.), tipo, número de servicio, monto de cargo promedio enviado, monto promedio permitido en Medicare y la cantidad estándar promedio en Medicare. Tampoco está claro si utilizaron los conjuntos de datos juntos o por separado o qué atributos se utilizaron y cuáles no, lo que dificulta la reproducción de estos experimentos. Los autores determinan que cuando se trata de variables de pago, es mejor ir estado por estado, ya que los datos de cada estado pueden variar. Sin embargo, en esta investigación, encontramos que se pueden lograr buenos resultados utilizando datos de Medicare que abarcan todo EE.UU. El marco que emplean es el

método de búsqueda *bump hunting* en el método de inducción de reglas del paciente, que es un enfoque no supervisado que busca determinar anomalías máximas al detectar espacios en masas mayores del conjunto de datos. Explican que al aplicar su marco, pueden caracterizar el espacio de atributos de los conjuntos de datos de CMS ayudando a descubrir los eventos que provocan pérdidas financieras.

Se observó una serie de diferencias con estos dos estudios [48,17], incluidos los métodos de procesamiento de datos, el proceso de combinación de datos y las comparaciones realizadas entre los tres conjuntos de datos de Medicare, tanto individualmente cómo combinados. Proporcionamos una descripción detallada de los métodos de procesamiento de datos para el conjunto de datos de Medicare utilizado en esta investigación, así cómo el mapeo y la generación de etiquetas de fraude utilizando el conjunto de datos LEIE. Hasta donde se sabe, este es el primer estudio que utiliza el conjunto de datos De-SynPUF de Medicare, sin otros estudios relacionados conocidos.

La contribución de este estudio proporciona discusiones, experimentos y análisis exploratorios con el fin de mostrar los algoritmos más precisos para la detección de posibles proveedores fraudulentos de Medicare. Los pasos de procesamiento de los datos consisten en la imputación de los mismos, determinar qué variables (características del conjunto de datos) mantener, transformar los datos del nivel de procedimiento al nivel de proveedor a través de la agregación de características para que coincida con el nivel del conjunto de datos LEIE logrando así el mapeo de etiquetas de fraude y crear un conjunto de datos combinado. Hay que tomar en cuenta que las etiquetas de fraude se utilizan para evaluar el fraude aprovechando la información histórica de exclusión, así cómo los pagos realizados por Medicare a los proveedores actualmente excluidos.

# CAPÍTULO 3

## Justificación

La salud es uno de los factores más importantes para garantizar la calidad de vida del ser humano, desafortunadamente los altos costos de los servicios dejan a muchos pacientes con una atención médica limitada. En respuesta, el gobierno de EE.UU ha establecido y financiado programas, cómo Medicare [53], que brindan asistencia financiera para que las personas que califiquen reciban los servicios médicos necesarios [23]. Hay una serie de problemas a los que se enfrentan los sistemas de seguros médicos y sanitarios, cómo el aumento de la población o los malos proveedores (es decir, médicos/proveedores fraudulentos o potencialmente fraudulentos), lo que reduce los fondos asignados para estos programas.

Estados Unidos ha experimentado un crecimiento significativo en la población anciana (65 años o más), debido a la buena calidad de atención médica que ofrece. La población de la tercera edad aumentó un 35 % entre 2004 y 2018 en comparación con el 6.5 % de los estadounidenses menores de 65 años [1]. Debido a esto el gasto en salud de EE. UU. aumento con una tasa de crecimiento anualizada entre 1995 y 2015 del 4.0 % (ajustada por la inflación) [20]. Es de suponer que el gasto seguirá aumentando y con él la necesidad de un sistema sanitario eficiente y rentable. Un problema importante al que se enfrenta la asistencia sanitaria es el fraude, ya que actualmente no se cuenta con un sistema eficiente que permita controlar el despilfarro y el abuso, debido a esto no se logra reducir significativamente la tensión financiera [2].

Actualmente se analizan los registros manualmente por medio de auditores e investigadores que buscan entre numerosos datos encontrar conductas posiblemente sospechosas o fraudulentas [47]. Este proceso manual, con enormes cantidades de datos para filtrar, puede ser tedioso y muy ineficiente. Actualmente existen sistemas que utilizan el enfoque de la ciencia de datos y *machine learning* para detectar fraudes [15,16]. El volumen de información dentro de la atención médica continua aumentando en proporción a la población y a la recolección de datos de la misma. Es por ello que los registros médicos electrónicos (EHR) aumentan, lo que permite el uso de "*Big Data*".

Es fundamental detectar los fraudes, la Oficina Federal de Investigaciones (FBI) estima que representan, del 3 % al 10 % de los costos de atención médica [42], con un total que

oscila entre los $19 mil millones y $65 mil millones en pérdidas financieras por año. Medicare representa el 20 % de todo el gasto sanitario de EE.UU. [16], con la implementación de una metodología de detección de fraudes correcta, se estima la recuperación de los costos totales de $3.8 a $13 mil millones de dólares. Medicare es un seguro médico con subsidio federal y, por lo tanto, no es un mercado de seguros de salud que funcione de la misma manera que las compañías de seguros de salud privadas [41].

Hay dos sistemas de pago disponibles a través de Medicare: pago por servicio y *Medicare Advantage*. Para este estudio, nos enfocamos en los datos dentro del sistema de pago por servicio de Medicare, donde el proceso básico de reclamos consiste en que un médico (u otro proveedor de atención médica) realice uno o más procedimientos y luego envie un reclamo a Medicare para el pago, en lugar de facturar directamente al paciente. El segundo sistema de pago, *Medicare Advantage*, se obtiene a través de una empresa privada contratada con Medicare, donde la empresa privada gestiona las reclamaciones y los procesos de pago [25]. Se proporciona información adicional sobre el proceso de Medicare y el fraude a Medicare dentro de [4,18,40,53].

Se eligió el conjunto de datos DE-SynPUF de Medicare porque cubre una amplia gama de posibles reclamaciones de proveedores, la información presenta un formato similar a través de los años y está disponible públicamente. Además, el conjunto de datos DE-SynPUF contiene componentes claves del programa Medicare para la detección de fraude, está investigación proporciona una visión integral del fraude en el programa Medicare. La información proporcionada por este conjunto de datos incluye el monto promedio pagado por estos servicios y otros puntos de datos relacionados con los procedimientos realizados, los medicamentos administrados o los suministros emitidos. El último conjunto de datos examinado en la investigación es la lista de personas y entidades excluidas (LEIE) [33], proporcionado por la Oficina de Inspectoría General, que contiene entidades y médicos fraudulentos del mundo real.

Es importante determinar los posibles fraudes dentro del sistema, ya que la reducción de los costos operativos aumentan la cobertura, mejoran el nivel de servicio de los programas de salud permitiendo el acceso a la salud a más población y así mejorar su calidad de vida

# CAPÍTULO 4

# Objetivos

Se presentan los objetivos que se buscan alcanzar en el desarrollo del trabajo.

## 4.1. Objetivo general

Implementar el modelo de *machine learning* que detecte con mayor precisión el posible fraude por proveedores en reclamos de salud.

## 4.2. Objetivos específicos

- Implementar ingeniería de características para mejorar la precisión de la predicción y el reconocimiento de patrones de fraude en el conjunto de datos analizado.
- Desarrollar prototipos con diferentes modelos de *machine learning* identificando sus precisiones en la detección del posible fraude por proveedores en reclamos de salud.
- Identificar el modelo de *machine learning* con la mejor precisión para la detección del posible fraude por proveedores de reclamos de salud en el conjunto de datos analizado.

# CAPÍTULO 5

# Marco teórico

En esta sección, describimos los conjunto de datos de CMS que usamos (*Beneficiary Summary*, *Inpatient Claims* y *Outpatient Claims*). Además, se utiliza la metodología de procesamiento de datos para crear el conjunto de datos, incluido el procesamiento, el mapeo de etiquetas de fraude entre los conjuntos de datos de Medicare y el LEIE, y la codificación *one-hot* para variables categóricas. La información dentro de cada conjunto de datos se basa en una muestra de 5 % de los beneficiarios de Medicare del 2008 y sus reclamos del 2008, 2009 y 2010. Hay que tomar en cuenta que estos datos no forman parte de ninguna reclamación presentada a través del programa *Medicare Advantage* [54]. Dado que CMS registra toda la información de reclamaciones después de que se realizan los pagos [13,14,15], se asume que los datos de Medicare ya están depurados y son correctos. Tomemos en cuenta que NPI no se utiliza en el paso de minería de datos, sino más bien para la identificación y agregación. Además, se agregaron características cómo la edad de los beneficiarios basado en su fecha de muerte y nacimiento, y una variable que nos indica si el beneficiario está muerto o no.

## 5.1. Medicare

### 5.1.1. Descripción de conjunto de datos

DE-SynPUF: El conjunto de datos CMS 2008-2010 DE-SunPUF de Medicare contiene 5 tipos de archivos los cuales son: *Beneficiary Summary*, *Inpatient Claims* y *Outpatient Claims*, Claims y Prescription Drug Events del 2008 al 2010. Cabe mencionar que para el cumplimiento de los objetivos de esta investigación se utilizaron únicamente los siguientes archivos: *Beneficiary Summary*, *Inpatient Claims* y *Outpatient Claims*. Actualmente, estos conjuntos de datos están disponibles en el sitio web de CMS para los años calendario 2008 a 2010 [12]. El conjunto DE-SynPUF contiene varios archivos por año y abarcan varios años. El conjunto contiene datos sintetizados tomados de una muestra aleatoria del 5 % de los beneficiarios de Medicare en 2008 y sus reclamos de 2008 a 2010. A cada beneficiario sintético se le asignó una identificación única, DESYNPUF_ID, que se proporciona en cada archivo

para vincular los reclamos sintéticos con un beneficiario. Esta identificación de beneficiario no contiene información sobre el afiliado ni sobre los registros del paciente, y se proporciona únicamente con fines de referencia y procesamiento de datos. El conjunto de datos también proporciona información sólida de metadatos sobre los datos de reclamaciones de CMS que no han estado disponibles en el dominio público.

- **Demográfico:** Las estimaciones del conjunto DE-SynPUF acerca de las características demográficas cómo (fecha de nacimiento, fecha de muerte, sexo, grupo etnico, estado y condado) de la población beneficiaria coinciden con la frecuencia univariante de la población completa de beneficiarios inscritos en Medicare en cualquier momento durante el año 2008.
- **Clínico:** Para las variables clínicas cómo condiciones crónicas pueden proporcionar a la investigación límites sobre cuántos casos con una condición específica es probable que estén en los reclamos de Medicare, lo que podría ser utilizado para generar cálculos de potenciales solicitudes.
- **Económico/Financiero:** Para las variables económicas y financieras proporciona un límite inferior para la estimación real del costo para la población total de beneficiarios inscritos en Medicare en cualquier momento del año 2008 y los costos para el 2009 y 2010.
- **Modelado multivariado:** Cabe mencionar que fueron alteradas las relaciones dinámicas entre las variables (información demográfica, clínica, económica y del proveedor) para limitar el riesgo de reidentificación. Por lo tanto, los análisis de modelos multivariados deben interpretarse con precaución. Sin embargo, los programas y procedimientos empleados en el modelado multivariante funcionaran en los conjuntos de datos limitados o datos identificables de CMS antes del 2011.

*CMS Beneficiary Summary* DE-SynPUF

El resumen de beneficiarios contiene 25 variables. Para las variables disponibles en los archivos derivados, se mantuvo el mismo nombre de variable. Aunque las variables en el resumen de beneficiarios fueron imputadas y aproximadas, para la mayoría de variables, el formato de valores de los datos es el mismo que en los datos originales ( por ejemplo, los códigos de condado definidos son códigos de condado válidos). En las pocas excepciones se agrega "SP_" como prefijo al nombre de la variable original para distinguir aquellos elementos del conjunto de datos cuyos valores ya no representan los valores típicos o el formato del campo de datos original. Cada registro pertenece a un beneficiario sintético de Medicare. En el Cuadro 1 se puede ver un ejemplo de un beneficiario con BeneID = BENE11001 extraído del conjunto de datos del resumen de beneficiarios de 2008.

Cuadro 1: Muestra de *CMS Beneficiary Summary* DE-SynPUF

| | BeneID | DOB | DOD | Gender | Race | RenalDiseaseIndicator | State | County | NoOfMonths_PartACov |
|---|---|---|---|---|---|---|---|---|---|
| 0 | BENE11001 | 1943-01-01 | NaT | 1 | 1 | 0 | 39 | 230 | 12 |

*CMS Inpatient Claims* DE-SynPUF:

El conjunto de datos de *Inpatient Claims* proporciona información relacionada con los reclamos de salud realizados por beneficiarios hospitalizados de 2008 al 2010. También proporciona detalles adicionales como sus fechas de admisión y alta y el código de diagnóstico de admisión. Este conjunto de datos contiene reclamos de pacientes hospitalizados compuestos por una muestra del 5 % de los beneficiarios de Medicare. Hay variables demográficas relacionadas con los reclamos que se proporcionan en este conjunto de datos. Los reclamos fueron seleccionados de una muestra aleatoria del 5 % de los realizados por los beneficiarios incluidos en el conjunto de *Beneficiary Summary* durante los años 2008 al 2010. Cabe mencionar que se cuenta con 30 variables tanto numéricas cómo categóricas que conforman el conjunto de datos. En el Cuadro 2 se puede ver un ejemplo de un beneficiario con BeneID = BENE11001 extraído del conjunto de datos Inpatient Claim de 2008 al 2010.

Cuadro 2: Muestra de *CMS Inpatient Claims* DE-SynPUF

| | BeneID | ClaimID | ClaimStartDt | ClaimEndDt | Provider | InscClaimAmtReimbursed | AttendingPhysician |
|---|---|---|---|---|---|---|---|
| 0 | BENE11001 | CLM46614 | 2009-04-12 | 2009-04-18 | PRV55912 | 26000 | PHY390922 |

*CMS Outpatient Claims* DE-SynPUF:

El conjunto de datos de *Outpatient Claims* proporciona información relacionada con los reclamos de salud hechos por beneficiarios ambulatorios de Medicare que visitan hospitales y no son ingresados. Este conjunto de datos brinda información en donde cada registro son procedimientos realizados en reclamos ambulatorios incurridos. Los reclamos fueron realizados por el 5 % de los beneficiarios de Medicare, los cuales están incluidos en *Beneficiary Summary* durante los años 2008 al 2010. Cabe mencionar que se cuenta con 27 variables tanto numéricas cómo categóricas que conforman el conjunto de datos. En el Cuadro 3 se puede ver un ejemplo de un beneficiario con BeneID = BENE11002 extraído del conjunto de datos Outpatient Claim de 2008 al 2010.

Cuadro 3: Muestra de *CMS Outpatient Claims* DE-SynPUF

| | BeneID | ClaimID | ClaimStartDt | ClaimEndDt | Provider | InscClaimAmtReimbursed | AttendingPhysician |
|---|---|---|---|---|---|---|---|
| 0 | BENE11002 | CLM624349 | 2009-10-11 | 2009-10-11 | PRV56011 | 30 | PHY326117 |

LEIE:

Para evaluar con precisión el rendimiento de la detección de fraudes tal cómo aparece en la práctica del mundo real, es necesaria una fuente de datos que contenga médicos que hayan cometido fraudes en el mundo real. Por lo tanto, se empleó la Lista de personas y entidades excluidas (LEIE) [33], que contiene la siguiente información: motivo de exclusion, fecha de exclusión y fecha de restitución / renuncia de todos los médicos actuales que no son aptos para ejercer la medicina y, por lo tanto, están excluidos de la práctica en los Estados Unidos durante un periodo de tiempo determinado. Este conjunto de datos fue establecido y mantenido mensualmente por la Oficina de Inspectoría General (OIG) [44] de conformidad con las secciones 1128 y 1156 de la Ley de Seguridad Social [45]. La OIG tiene autoridad para excluir a personas y entidades de los programas de atención médica financiados con fondos federales, como Medicare. Desafortunadamente, el LEIE no es exhaustivo, ya que el 38 % de los proveedores con condenas por fraude continúan ejerciendo la medicina y el 21 % no fue suspendido de la práctica médica a pesar de sus condenas [46]. Además, el conjunto de datos LEIE solo contiene los valores de NPI para un pequeño porcentaje de médicos y entidades. En el Cuadro 4 se muestra un ejemplo de cuatro médicos diferentes y cómo se les representa dentro del LEIE, donde cualquier médico sin un NPI en la lista tiene un valor 0.

Cuadro 4: Muestra de LEIE

| Specialty | ... | Npi | ... | Excltype | Excldate | ... |
|---|---|---|---|---|---|---|
| GENERAL PRACTICE/FP | ... | 0 | ... | 1128b6 | 19770701 | ... |
| EMPLOYEE | ... | 0 | ... | 1128b6 | 19780124 | ... |
| GENERAL PRACTICE | ... | 1003016742 | ... | 1128a1 | 20170720 | ... |
| NURSE/NURSES AIDE | ... | 1003011644 | ... | 1128b4 | 20091220 | ... |

Fuente: [46]

El LEIE se agrega a nivel de proveedor y no tiene información específica sobre procedimientos, medicamentos o equipos relacionados con actividades fraudulentas. Existen diferentes categorías de exclusiones, basadas en la gravedad de la falta, descritas por varios números de reglas. No se utilizaron todas las exclusiones, sino que se filtraron los proveedores excluidos por reglas seleccionadas que indican que se cometió fraude [5]. El Cuadro 5 muestra los códigos que corresponden a las exclusiones de proveedores fraudulentos y la duración de la exclusión obligatoria. Se ha determinado que cualquier comportamiento antes y durante de la "fecha de finalización de la exclusión" de un médico constituye un fraude.

Cuadro 5: Reglas de LEIE que involucran fraude

| Número de regla | Descripción |
|---|---|
| 1128 (a) (1) | Condena por delitos relacionados con el programa |
| 1128 (a) (2) | Condena relacionada con abuso o negligencia del paciente |
| 1128 (a) (3) | Condena por delito grave relacionada con el fraude a la atención médica |
| 1128 (b) (4) | Revocación o suspensión de la licencia |
| 1128 (b) (7) | Fraude, comisiones ilegales y otras actividades prohibidas |
| 1128 (c) (3) (g) (i) | Condena por dos delitos de exclusión obligatoria 10 años |
| 1128 (c) (3) (g) (ii) | Condena de 3 delitos de exclusión obligatoria indefinida |

Fuente: [5]

## 5.2. *Big Data*

El término de *Big Data* no se acepta universalmente en toda la literatura [27,28,37,38,43], por lo que se hizo uso de una definición general dada por Demchenko en [19] que definen *Big Data* con cinco V: Volumen, Velocidad, Variedad, Veracidad y Valor. El volumen se refiere a grandes cantidades de datos, la velocidad se aplica al alto ritmo al que se generan nuevos datos, la variedad se refiere al nivel de complejidad de los datos (por ejemplo, la incorporación de datos de diferentes fuentes), la veracidad representa la autenticidad de los datos y valor implica que tan buena es la calidad de los datos en referencia a los resultados esperados. Los conjuntos de datos publicados por CMS exhiben muchas de estas cualidades de *Big Data*.

Dado a que el objetivo definido es un problema de categorización, se optó por buscar algoritmos de *machine learning* con especialidad en este tipo de problemas. Se determinó que es un problema de categorización, debido a que se busca detectar cuando un reclamo realizado por un proveedor puede ser fraudulento o normal. Es por ello que se utilizaron dos modelos de clasificación disponibles en la biblioteca de aprendizaje automático scikit-learn 0.23.2: regresión logística y bosque aleatorio (*random forest*). También se utilizó la red neuronal *Autoencoder* de la biblioteca Keras en su ajuste predeterminado y con dos capas ocultas. En esta sección, describimos brevemente a cada modelo y se anotará cualquier cambio que difiera de la configuración predeterminada.

## 5.3. *Machine Learning*

### 5.3.1. Regresión Logística (LR)

La regresión logística [51] es un tipo de análisis de regresión, utilizado para predecir el resultado de una variable categórica (esto quiere decir que una variable puede adoptar un número limitado de categorías) en función de las variables independientes o predictoras. LR usa una función sigmoidea (logística) para generar valores que pueden interpretarse cómo

probabilidades de clase. LR es similar a la regresión lineal pero utiliza una clase de hipótesis diferente para predecir la pertenencia a la clase [24,31,55]. La matriz enlazada se estableció para que coincida con la forma de los datos (número de clases y características) para que el algoritmo sepa el número de clases y características que contiene el conjunto de datos. El tamaño del vector límite es igual a 1 para la regresión binomial y no establece umbrales para la clasificación binaria.

### 5.3.2. *Random Forest (RF)*

El modelo *Random Forest* [50] es un método de aprendizaje por conjuntos que genera una gran cantidad de árboles. El valor categórico que aparece con mayor frecuencia entre los árboles es la categoría que se predice cómo resultado del modelo. Como método de aprendizaje por conjuntos, RF es una combinación de varios árboles predictores. Cada árbol dentro del bosque depende de los valores dictados por un vector aleatorio que se muestrea de forma independiente y donde cada árbol se distribuye por igual entre el bosque [50]. El conjunto de RF inserta la aleatoriedad en el proceso de entrenamiento que puede minimizar el sobreajuste y es bastante robusto para datos desequilibrados [7,29]. Construimos cada modelo de RF con 500 árboles, ya que en el desarrollo de la investigación se encontró el poco beneficio obtenido al utilizar más árboles. Los pesos de las clases se definieron en modo "balanceado" para utilizar los valores de 'y' para ajustar automáticamente los pesos de forma inversamente proporcional a las frecuencias de clase de los datos de entrada. Para la profundidad máxima del árbol se estableció en una profundidad máxima de 4.

### 5.3.3. Redes Neurales - *Autoencoder*:

El objetivo del *autoencoder* [21] es aprender la representación (*encoding*) para un conjunto de datos, normalmente para la reducción de dimensionalidad. De lado de la reducción, va aprendiendo en la reconstrucción, donde el *autoencoder* intenta regenerar a partir del *encoding* una representación lo más cercana posible a la entrada original, de ahí su nombre. Recientemente, el concepto del *autoencoder* ha sido utilizado en modelos generativos de datos. Para esta investigación se utilizó la misma técnica para aprender patrones en datos no fraudulentos y con ellos entrenar el modelo. Se utilizó el umbral de reconstrucción de error (*reconstruction error threshold*), para predecir la clase de datos objetivo.

## 5.4. Métricas de evaluación:

### 5.4.1. Matriz de confusión:

Según [36], para problemas de clasificación binaria, la evaluación de discriminantes es la mejor solución durante el entrenamiento de clasificación se puede definir según la matriz de confusión cómo se muestra en el Cuadro 6. La fila de la tabla representa la predicción de la clase, mientras que la columna representa el valor real de la clase.

Cuadro 6: Matriz de confusión para una clasificación binaria.

| | **Actual Positive Class** | **Actual Negative Class** |
|---|---|---|
| **Predicted Positive Class** | True positive (*tp*) | False negative (*fn*) |
| **Predicted Negative Class** | False positive (*fp*) | True negative (*tn*) |

Fuente: [36]

De esta matriz de confusión, tp y tn denotan el número de verdaderos positivos y negativos que se clasifican correctamente. Mientras tanto, fp y fn denota el número de falsos negativos y positivos mal clasificados. A partir de la matriz de confusión se pueden generar varias métricas cómo las siguientes: Exactitud, Sensibilidad, Especificidad, Precisión, Recuperación, Puntuación F1, etc.

### 5.4.2. Exactitud:

En general, la exactitud mide la proporción de predicciones correctas sobre el número total de instancias evaluadas [36]. Su fórmula es la siguiente:

Figura 2: Fórmula de exactitud

$$\frac{tp + tn}{tp + fp + tn + fn}$$

Fuente [36]

Donde;

- **Verdadero positivo (TP):**número de instancias positivas reales predichas correctamente como positivas.
- **Verdadero negativo (TN):** número de instancias negativas reales predichas correctamente como negativas.
- **Falso positivo (FP):** número de casos negativos clasificados incorrectamente como positivos.
- **Falso negativo (FN):** número de casos positivos asignados incorrectamente como negativos.

### 5.4.3. Sensibilidad:

Esta métrica se utiliza para medir la fracción de verdaderos positivos que están correctamente clasificados [36]. Su fórmula es la siguiente:

Figura 3: Fórmula de sensibilidad

$$\frac{tp}{tp + fn}$$

Fuente: [36]

### 5.4.4. Especificidad:

Esta métrica se utiliza para medir la fracción de verdaderos negativos que están correctamente clasificados [36]. Su fórmula es la siguiente:

Figura 4: Fórmula de especificidad

$$\frac{tn}{tn + fp}$$

Fuente: [36]

### 5.4.5. Precisión:

La precisión se usa para medir los verdaderos positivos que se predicen correctamente a partir de patrones totales predichos en una clase positiva [36]. Su fórmula es la siguiente:

Figura 5: Fórmula de precisión

$$\frac{tp}{tp + fp}$$

Fuente: [36]

### 5.4.6. Recuperación:

La precisión se usa para medir los verdaderos positivos que se predicen correctamente a partir de patrones totales predichos en una clase positiva [36]. Su fórmula es la siguiente:

Figura 6: Fórmula de recuperación

$$\frac{tp}{tp + tn}$$

Fuente: [36]

### 5.4.7. Puntuación F1:

Según [36], esta es la media armónica de precisión y recuperación y proporciona una mejor medida de los casos clasificados incorrectamente que la métrica de precisión. La puntuación F1 tiene un mejor desempeño que la precisión para problemas de clasificación binaria y es una mejor métrica cuando se tienen conjuntos de datos desequilibrados. También la puntuación F1 se usa cuando los falsos negativos y falsos positivos son importantes. Su fórmula es la siguiente:

Figura 7: Fórmula de puntuación F1

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

Fuente: [36]

### 5.4.8. Error Cuadrado Medio (MSE):

En general, el MSE mide la diferencia entre las predicciones y soluciones deseadas. Se requiere el menor valor de MSE para obtener un mejor entrenamiento supervisado [36]. El MSE se define a continuación:

Figura 8: Fórmula de MSE

$$MSE = \frac{1}{n}\sum_{j=1}^{n}(P_j - A_j)^2$$

Fuente: [36]

Donde Pj es el valor de predicción por la instancia j, Aj es el valor objetivo real de la instancia j y n es el número total de instancias. De manera similar a la precisión, la principal limitación del MSE es que no proporciona la información de compensación entre los datos clasificados.

### 5.4.9. Área bajo la curva ROC (AUROC):

AUROC es una de las métricas más populares, ya que refleja el desempeño general de un clasificador [36]. Para problemas de clasificación binaria, el valor AUROC se calcula de la siguiente manera:

Figura 9: Fórmula de AUROC

$$\frac{S_p - n_p(n_n + 1)/2}{n_p n_n}$$

Fuente: [36]

Donde Sp es la suma de todos los ejemplos positivos clasificados, mientras que np y nn muestran el número de ejemplos positivos y negativos respectivamente.

# CAPÍTULO 6

# Marco metodológico

El desarrollo del modelo de *machine learning* para la detección de posible fraude por proveedores en reclamos de salud se desarrolló en cuatro partes.

## 6.1. Primera parte

Presentación del tema y la justificación enfocada desde el punto de vista de impacto en costos.

## 6.2. Segunda parte: Teoría

Se presentan los conceptos básicos necesarios, y se definen los conjuntos de datos utilizados para el desarrollo del modelo de *machine learning* con el fin de detectar el posible fraude por proveedores en reclamos de salud. A continuación, se explican los algoritmos utilizados en la investigación para el cumplimiento del objetivo. También se añadieron los conocimientos descubiertos por otras investigaciones acerca de la detección de fraude por medio del uso de Inteligencia Artificial en los antecedentes.

## 6.3. Tercera parte: Desarrollo

Se propone la metodología a utilizar, se desarrollan los prototipos y se discuten los resultados obtenidos. En esta fase se utilizó como referencia la metodología de un proyecto de *machine learning* mostrada en [32] la cual propone las siguientes fases: (1) Definiciones y

terminologías. (2) Definición de Problema. (3) Recolección de datos. (4) Limpieza de datos. (5) Resumen estadístico. (6) Partición de datos. (7) Selección de modelo (8) Entrenamiento de modelo. (9) Implementación de modelo. El propósito de utilizar está metodología es lograr alcanzar un modelo que logre identificar posibles reclamos fraudulentos con la mayor precisión. Las etapas desarrolladas fueron las siguientes:

### 6.3.1. Definición del problema

Con base al problema a resolver se propone el uso de *machine learning*. Donde el objetivo principal de la investigación es la detección del posible comportamiento fraudulento en reclamos de atención médica, se encontró que los centros de servicios de Medicare y Medicaid (CMS) se unificaron al publicar conjuntos de datos de Medicare para ayudar a identificar el fraude dentro de Medicare. Por ello, se unificó el objetivo de la investigación con las necesidades de Medicare para la detección de fraudes en reclamos de salud.

El fraude en la atención médica es un crimen organizado que involucra a pares de proveedores, médicos y beneficiarios que actúan en conjunto para realizar reclamos fraudulentos. El análisis riguroso de los datos de Medicare ha resultado en que muchos médicos han participado en el fraude. Ellos adoptan formas en las que se utiliza un código de diagnóstico ambiguo para adoptar los procedimientos y medicamentos más costosos. Las compañías de seguros son las instituciones más vulnerables afectadas por estas malas prácticas. Por esta razón, las compañías de seguros pueden aumentar sus primas de seguro y, cómo resultado, la atención médica se vuelve más costosa día a día.

El fraude de la atención médica adopta muchas formas. Algunos de los tipos de fraude más comunes por parte de los proveedores son:

- Facturación por servicios no prestados.
- Reclamos duplicados por un mismo servicio.
- Tergiversación del servicio prestado.
- Cobro por un servicio más complejo o costoso del que realmente se prestó.
- Facturación de servicios cubiertos cuando el servicio realmente brindado no está cubierto.

Es por ello que, el objetivo de esta investigación es analizar si el uso de *machine learning* puede ayudar en la predicción de reclamos fraudulentos basados en historia de reclamos similares que fueron fraudulentos. Junto con esto, también se descubrirán importantes variables que son útiles para la detección del comportamiento de los proveedores potencialmente fraudulentos. Además, se estudiarán los patrones en las afirmaciones del proveedor para comprender el comportamiento futuro de los proveedores, con el fin de evitar este tipo de problemática.

### 6.3.2. Investigación

En esta etapa se llevó a cabo una exploración de modelos de machine learning que serían candidatos para dar solución al problema de categorización definido en la sección anterior. Una comprensión del problema, unificada con los datos disponibles, permitió un proceso de selección de algoritmos adecuados para lograr una primera aproximación de la solución. Se seleccionaron tres tipos de algoritmos para esta investigación, dos de aprendizaje supervisado y uno no supervisado. Los modelos de regresión logística y *random forest* fueron los seleccionados para el aprendizaje supervisado, ya que son utilizados comúnmente y brindan un rendimiento razonablemente bueno para este tipo de problema. La red neural *autoencoder* es un tipo de red neural que es utilizada para aprender patrones de datos eficientemente de manera no supervisada.

### 6.3.3. Minería de datos

Para esta etapa fue realizada la minería de los datos obtenidos de CMS, los cuales incluyen información de beneficiarios y sus reclamos de tres años. Este paso es crucial ya que permite una mejor efectividad y rendimiento para los modelos entrenados. El resultado de la solución propuesta es definida por los datos minados. Los datos obtenidos de CMS fueron examinados y analizados utilizando herramientas de visualización y métodos estadísticos. Entre las técnicas utilizadas está la codificación *one hot* ya que permite que la representación de datos categóricos sea más expresiva. Muchos algoritmos de *machine learning* no pueden trabajar directamente con datos categóricos por lo que las variables categóricas deben convertirse en números.

El análisis y exploración de los datos permitió asegurar que se cumplan los siguientes requisitos:

- Los datos de CMS son lo suficientemente diversos como para proporcionar una capacidad de predicción de los modelos seleccionados que se adapten a una variedad de escenarios posibles.
- Los datos de CMS son imparciales para garantizar que los modelos seleccionados puedan generalizarse adecuadamente durante la inferencia.
- Los datos de CMS son abundantes.

Como CMS es un programa gubernamental con controles de calidad transparentes y documentación detallada para cada uno de sus conjuntos de datos, se considera que estos datos son confiables, válidos y representativos de todas las afirmaciones conocidas de proveedores de Medicare. Además, el conjunto de datos de LEIE también se considera significativo, ya que contiene el mayor depósito conocido de proveedores médicos fraudulentos del mundo real en los Estados Unidos.

Cabe destacar que fue utilizado el lenguaje de programación Python para el desarrollo de esta y las siguientes secciones de la investigación.

### 6.3.4. Preprocesamiento de datos

La etapa de preprocesamiento de los datos se basó principalmente en los requisitos de entrada de los modelos seleccionados previamente. Se regresó a la etapa de investigación para verificar los parámetros y requisitos de entrada que requieren los algoritmos de aprendizaje supervisado cómo los no supervisados. El preprocesamiento transforma los datos sin procesar, en un formato que permite el entrenamiento exitoso de los modelos. Entre los pasos realizados están los siguientes:

- Reformateo de los datos.
- Conversión de variables.
- Estandarización.
- Separación del conjunto de datos (80 % entrenamiento y 20 % evaluación).

### 6.3.5. Implementación de modelo

Para la implementación de los modelos de *machine learning*, se utilizó la biblioteca scikit-learn 0.23.2 la cual proporciona una amplia variedad de herramientas cómo modelos de clasificación, regresión, clustering, preprocesamiento, etc. Para esta investigación, se utilizaron los modelos de aprendizaje supervisado: regresión logística y *random forest* de la biblioteca *scikit-learn*, ya que al ser modelos de clasificación, son de gran ayuda para clasificar los reclamos hechos por proveedores potencialmente fraudulentos. La decisión de elegir el algoritmo de regresión logística también agrega explicabilidad a las predicciones. El rendimiento del modelo de regresión logística muestra la linealidad entre las variables. Por otro lado, uno de los beneficios del *random forest*, es el poder de manejar grandes conjuntos de datos con una mayor dimensionalidad y puede manejar miles de variables de entrada identificando las más significativas. Además, el modelo da cómo resultado la importancia de cada variable, siendo una característica muy útil. Otro beneficio es que verifica la no linealidad entre las variables.

También se utiliza el modelo de red neural *autoencoder* de la biblioteca de *machine learning* Keras. El *autoencoder* es un modelo muy bueno para la detección de anomalías. Para cumplir nuestro objetivo el modelo se entrenará con registros no fraudulentos usando autoencoders. Mientras se reconstruyen los datos fraudulentos, se creará un error, llamado error de reconstrucción. Basándonos en la configuración del umbral de errores de reconstrucción, podemos predecir fácilmente el comportamiento fraudulento de proveedores en atención médica.

### 6.3.6. Entrenamiento

Los datos preprocesados en etapas anteriores fueron utilizados para la etapa de entrenamiento. El entrenamiento de los modelos, implicó pasar los datos refinados a través del modelo para crear un modelo que logre realizar correctamente su objetivo.

Este proceso implicó pasar iterativamente mini lotes de los datos de entrenamiento a través del modelo durante una cantidad específica de épocas. Durante las primeras etapas, el rendimiento y precisión de los modelos fueron poco impresionantes. Pero a medida que el modelo entrena y realiza predicciones, también compara los valores pronosticados con el valor objetivo. En cada iteración de entrenamiento el modelo comienza a mejorar y aumentar su rendimiento en la tarea para la cual está siendo diseñado e implementado.

Cabe destacar que antes que el entrenamiento inicie, se establecen los parámetros que guiarán la efectividad de nuestra etapa de entrenamiento en el modelo. Al realizar el entrenamiento es vital registrar las métricas de cada proceso de entrenamiento. Las métricas más importantes recaudadas son las siguientes:

- Exactitud.
- AUROC.
- Puntuación F1.

Para clasificar y visualizar las métricas del entrenamiento y etapas posteriores, se utilizaron las disponibles en la biblioteca *sklearn*. Al visualizar las métricas de entrenamiento, fue posible identificar algunos problemas comunes de entrenamiento de modelos de machine learning, como el sobreajuste y el sub-ajuste.

### 6.3.7. Evaluación

En esta etapa de evaluación, se toma el modelo previamente entrenado para evaluar su desempeño. La evaluación consiste en utilizar una partición de los datos refinados, generalmente conocidos cómo "datos de prueba". Los datos de prueba no han sido vistos por el modelo durante el entrenamiento. Cabe destacar que también son representaciones de datos que se espera encontrar en escenarios prácticos. Entre las métricas utilizadas para la evaluación se encuentran las siguientes:

- Matriz de confusión (matriz de error).
- Exactitud.
- AUROC.
- Puntuación F1.

### 6.3.8. Predicción

En esta etapa, se utiliza una pequeña porción de datos de prueba para realizar predicciones de posible fraude. Fueron utilizados 1,353 reclamos de atención médica realizados por diferentes proveedores, para está etapa es vital el uso de la matriz de confusión ya que da una mejor idea de cómo están prediciendo los algoritmos de clasificación entrenados anteriormente. Se evalúa realizando un conteo de los aciertos y errores de los reclamos con posible

fraude, de esa manera se puede comprobar si el algoritmo está clasificando erróneamente el fraude y en qué medida.

## 6.4. Cuarta parte: Análisis de resultados

Para esta etapa se cierra con el análisis de los resultados, conclusiones y recomendaciones para la continuación de la investigación.

# CAPÍTULO 7

# Resultados

## 7.1. Minería de datos

Los datos utilizados para la minería de datos fueron recaudados de dos fuentes mencionadas en el marco teórico de la presente investigación. El conjunto de datos CMS DE-SynPUF del 2008 al 2010 está compuesto por cinco archivos, de los cuales únicamente fueron utilizados tres, ya que son los que mayor aporte generan al objetivo de la investigación. Los archivos recolectados y minados fueron: *CMS Beneficiary Summary, CMS Inpatient Claims y CMS Outpatient Claims*. También forma parte del conjunto de datos minados de la lista de personas y entidades excluidas (LEIE). Para una mejor comprensión de los conjuntos de datos utilizados para la etapa de minería de datos visitar la sección de marco teórico.

En el Cuadro 7 podemos observar la estructura de los datos antes de la división del entrenamiento y prueba.

Cuadro 7: Resumen de conjunto de datos

| Conjunto de datos | Descripción | Cantidad de Registros | Cantidad de Variables |
|---|---|---|---|
| Target | Listado de proveedores fraudulentos y no fraudulentos del LEIE | 6763 | 2 |
| BeneficiaryData | Resumen de beneficiarios pertenecientes a Medicare | 202,524 | 25 |
| InpatientData | Reclamos presentados por pacientes que fueron hospitalizados. | 50,025 | 30 |
| OutpatientData | Reclamos presentados por pacientes que visitan hospitales ambulatoriamente. | 643,578 | 27 |

Fuente: Elaboración propia

Inicialmente, se determinó si había proveedores duplicados para los datos de entrenamiento. Posteriormente se analizó la estructura de los beneficiarios, revisando si existían valores faltantes en cada columna del conjunto de datos de beneficiarios. En el Cuadro 8 se puede observar los tipos de datos que tiene cada columna del conjunto de datos de beneficiarios.

Cuadro 8: Tipos de datos del conjunto de datos de beneficiarios

| Variable | Dtype |
|---|---|
| BeneID | Categórica |
| DOB | Categórica |
| DOD | Categórica |
| Gender | Int64 |
| Race | Int64 |
| RenalDiseaseIndicator | Categórica |
| State | Int64 |
| County | Int64 |
| NoOfMonths_PartACov | Int64 |
| NoOfMonths_PartBCov | Int64 |
| ChronicCond_Alzheimer | Int64 |
| ChronicCond_HeartFailure | Int64 |
| ChronicCond_KidneyDisease | Int64 |
| ChronicCond_Cancer | Int64 |
| ChronicCond_ObstrPulmonary | Int64 |
| ChronicCond_Depression | Int64 |
| ChronicCond_Diabetes | Int64 |
| ChronicCond_IschemicalHeart | Int64 |
| ChronicCond_Osteoporasis | Int64 |
| ChronicCond_Rheumatoidarthritis | Int64 |
| ChronicCond_Stroke | Int64 |
| IPAnnualReimbursementAmt | Int64 |
| IPAnnualDeductibleAmt | Int64 |
| OPAnnualReimbursementAmt | Int64 |
| OPAnnuaklDeductibleAmt | Int64 |

Fuente: Elaboración propia

Fueron reemplazados los números 2 por 0 para las condiciones crónicas, eso significa que cuando no es una condición crónica sería '0' y '1' cuando lo sea. Se agregó la edad de las personas basado en su fecha de nacimiento y fecha de muerte (DOB y DOD). Se identificó que el último valor de DOD es 2009-12-01, eso significa que los datos de los beneficiarios son del año 2009. Entonces se calculó la edad de los beneficiarios para el año 2009. Se agregó

una nueva columna de tipo bandera llamada '*WhetherDead*' con los valores del DOD para saber si el beneficiario ha fallecido o no. Si el valor del 'WhetherDead' es 1 es porque el beneficiario está muerto y si su valor es 0 es porque aún vive.

Para el conjunto de datos de pacientes hospitalizados se inició revisando sus valores faltantes, para posteriormente iniciar a comprender su estructura. En el Cuadro 9 se encuentra un resumen de los tipos de datos que contiene el conjunto de pacientes hospitalizados.

Cuadro 9: Tipos de datos del conjunto de datos de pacientes hospitalizados.

| Variable | Dtype |
|---|---|
| BeneID | Categórica |
| ClaimID | Categórica |
| ClaimStartDt | Categórica |
| ClaimEndDt | Categórica |
| Provider | Categórica |
| InscClaimAmtReimbursed | Int64 |
| AttendingPhysician | Categórica |
| OperatingPhysician | Categórica |
| OtherPhysician | Categórica |
| AdmissionDt | Categórica |
| ClmAdmitDiagnosisCode | Categórica |
| DeductibleAmtPaid | Float64 |
| DischargeDt | Categórica |
| DiagnosisGroupCode | Categórica |
| ClmDiagnosisCode_1 | Categórica |
| ClmDiagnosisCode_2 | Categórica |
| ClmDiagnosisCode_3 | Categórica |
| ClmDiagnosisCode_4 | Categórica |
| ClmDiagnosisCode_5 | Categórica |
| ClmDiagnosisCode_6 | Categórica |
| ClmDiagnosisCode_7 | Categórica |
| ClmDiagnosisCode_8 | Categórica |
| ClmDiagnosisCode_9 | Categórica |
| ClmDiagnosisCode_10 | Categórica |
| ClmProcedureCode_1 | Float64 |
| ClmProcedureCode_2 | Float64 |
| ClmProcedureCode_3 | Float64 |
| ClmProcedureCode_4 | Float64 |
| ClmProcedureCode_5 | Float64 |
| ClmProcedureCode_6 | Float64 |

Fuente: Elaboración propia

Para el conjunto de datos de pacientes hospitalizados se creó una columna para indicar la cantidad de días que el paciente fue admitido para hospitalización. El nombre de la nueva

columna es '*AdmitForDays*'. Para su cálculo se utilizó la diferencia entre la fecha de alta y la fecha de admisión (*AdmissionDt, DischargeDt*). Si el paciente fue admitido por un día, se le asignó el valor 1. Posteriormente se revisaron los valores mínimos y máximos para la columna creada '*AdmitForDays*' para el conjunto de entrenamiento cómo el de prueba, y se obtuvo un valor mínimo de 1 y un valor máximo de 36 días de hospitalización.

En el conjunto de datos de pacientes externos o ambulatorios también se revisaron los valores faltantes y posteriormente se analizó su estructura.

En el Cuadro 10 se encuentra un resumen de los tipos de datos que contiene el conjunto de pacientes ambulatorios.

Cuadro 10: Tipos de datos del conjunto de datos de pacientes ambulatorios.

| Variable | Dtype |
|---|---|
| BeneID | Categórica |
| ClaimID | Categórica |
| ClaimStartDt | Categórica |
| ClaimEndDt | Categórica |
| Provider | Categórica |
| InscClaimAmtReimbursed | Int64 |
| AttendingPhysician | Categórica |
| OperatingPhysician | Categórica |
| OtherPhysician | Categórica |
| DeductibleAmtPaid | Float64 |
| ClmDiagnosisCode_1 | Categórica |
| ClmDiagnosisCode_2 | Categórica |
| ClmDiagnosisCode_3 | Categórica |
| ClmDiagnosisCode_4 | Categórica |
| ClmDiagnosisCode_5 | Categórica |
| ClmDiagnosisCode_6 | Categórica |
| ClmDiagnosisCode_7 | Categórica |
| ClmDiagnosisCode_8 | Categórica |
| ClmDiagnosisCode_9 | Categórica |
| ClmDiagnosisCode_10 | Categórica |
| ClmProcedureCode_1 | Float64 |
| ClmProcedureCode_2 | Float64 |
| ClmProcedureCode_3 | Float64 |
| ClmProcedureCode_4 | Float64 |
| ClmProcedureCode_5 | Float64 |
| ClmProcedureCode_6 | Float64 |
| ClmAdmitDiagnosisCode | Categórica |

Fuente: Elaboración propia

Se observó que las columnas de los pacientes hospitalizados y ambulantes, son similares, entonces se combinaron los conjuntos de datos basándose en sus campos similares utilizando

un *Outer Join*. Se utilizaron todas las columnas de los datos de pacientes ambulantes ya que se busca hacer una unión y no duplicar las columnas de ambos conjuntos. Al combinar los datos de pacientes hospitalizados y ambulantes, se creó un conjunto de datos de todos los reclamos realizados. Finalmente se combinó los detalles de los beneficiarios con todos sus reclamos. Para ello se utilizó la variable 'BeneID' cómo clave de unión para realizar un *Inner Join*.

El conjunto de datos de LEIE se une al conjunto detallado de todos los pacientes utilizando la columna '*Provider*' cómo la llave de unión. Luego de la unión, creamos un conjunto de datos con una característica de exclusión, el cual sería el atributo categórico final discutido en las secciones anteriores, que indica instancias de fraude o no fraude. A través de este proceso, estamos contabilizando dos tipos de comportamiento fraudulento: (1) comportamiento fraudulento real y (2) pagos hechos por medicare basados en presentaciones de proveedores excluidos, donde ambos drenan fondos de Medicare de manera inapropiada. El conjunto de datos final incluye todos los proveedores excluidos conocidos marcados mediante la función de exclusión categórica.

El Cuadro 11 muestra la distribución porcentual de fraude y no fraude dentro del conjunto de datos.

Cuadro 11: Distribución porcentual del fraude potencial

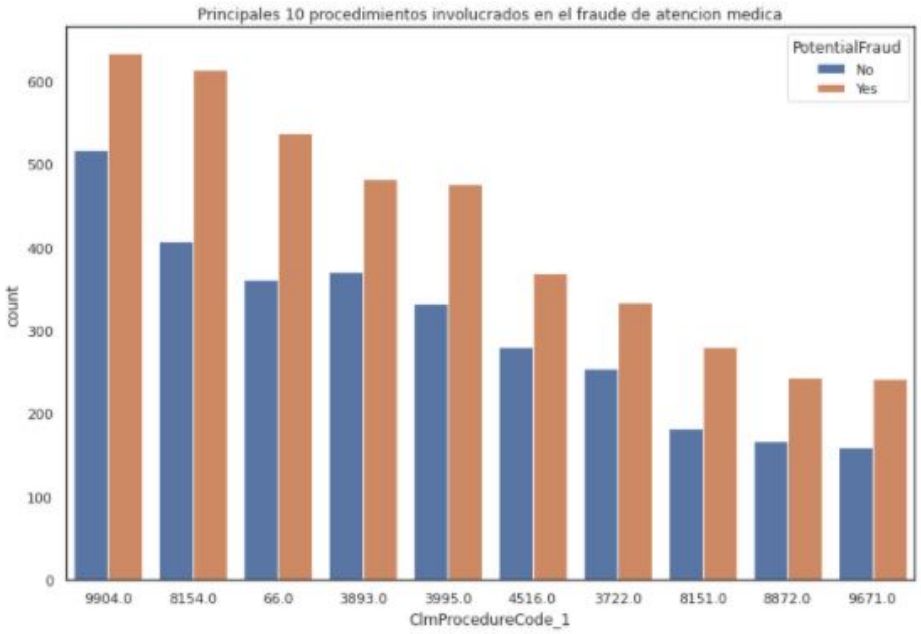
| Clase | Porcentaje de fraude potencial |
|---|---|
| Fraude | 38.12% |
| No Fraude | 61.87% |

Fuente: Elaboración propia

Se procedió a revisar la frecuencia de las transacciones fraudulentas y no fraudulentas en los datos de entrenamiento, y se obtuvo que el 9.353 % de los reclamos son potencialmente fraudes y el 90.646 % no son reclamos fraudulentos. Por lo tanto, se debe obtener información sobre la cantidad de reclamos y los montos involucrados de los siguientes casos: Beneficiario, Beneficiario + Médico, Médico, Diagnostico, Procedimiento, etc. Se detectó que en todo el conjunto de datos de entrenamiento el 38.12 % de los proveedores son fraudulentos y 61.87 % no lo son.

Para mejorar nuestra visión del conjunto de datos se procedió a identificar los principales procedimientos, diagnósticos y médicos, involucrados en el fraude de atención médica. En la Figura 10 podemos observar los 10 principales procedimientos involucrados en el fraude de atención médica.

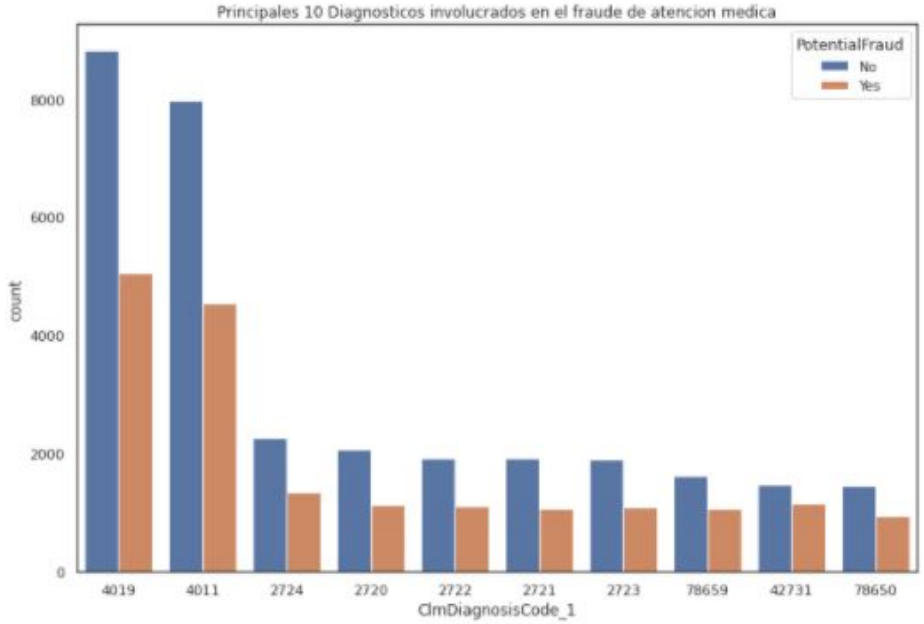
Figura 10: Principales 10 procedimientos involucrados en fraude de atención médica.




Fuente: Elaboración propia

De la Figura 10, se infiere que el procedimiento 9904, 8157 y 66 son los principales 3 involucrados ( en términos de dinero involucrado). El recuento de la distribución de reclamos fraudulentos y no fraudulentos muestra transacciones sospechosas involucradas en ellos.

Figura 11: Principales 10 diagnósticos involucrados en el fraude de atención médica

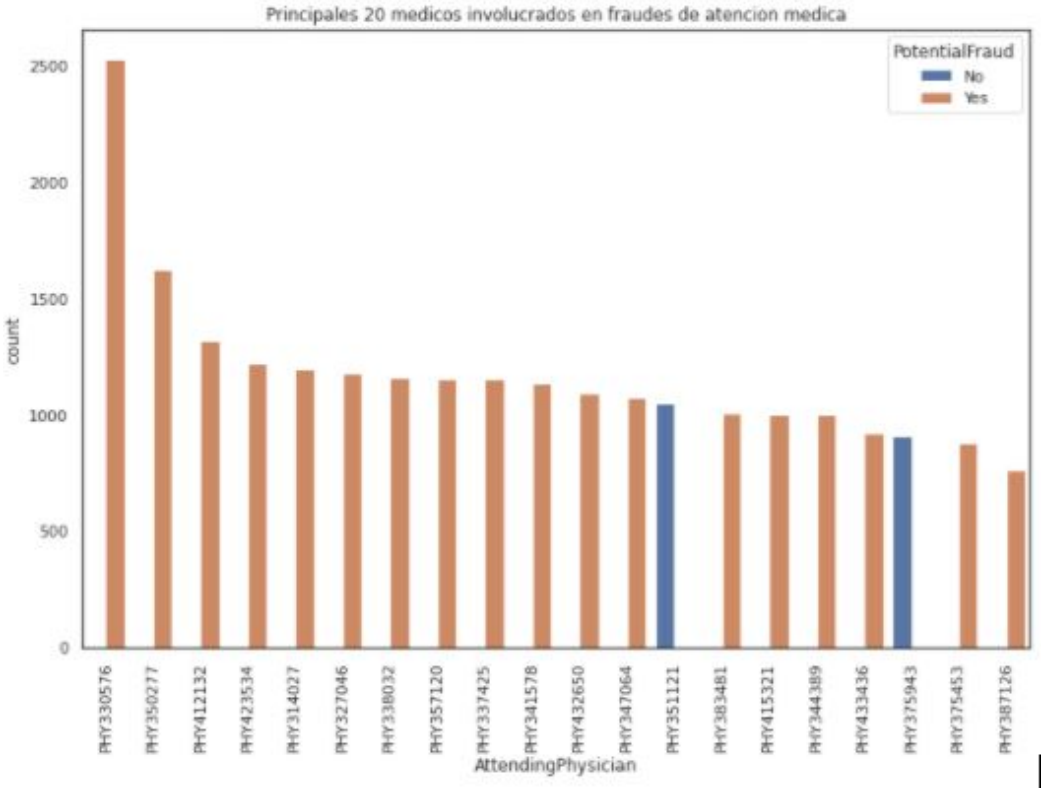



Fuente: Elaboración propia

De la Figura 11, se infiere que los principales diagnósticos involucrados en fraude son el 4019, 4011 y 2724 (en términos de dinero involucrado). El recuento de la distribución de diagnósticos fraudulentos y no fraudulentos muestra transacciones sospechosas involucradas en ellos.

En la Figura 12 tenemos un recuento de los 20 principales médicos involucrados en fraudes de atención médica.

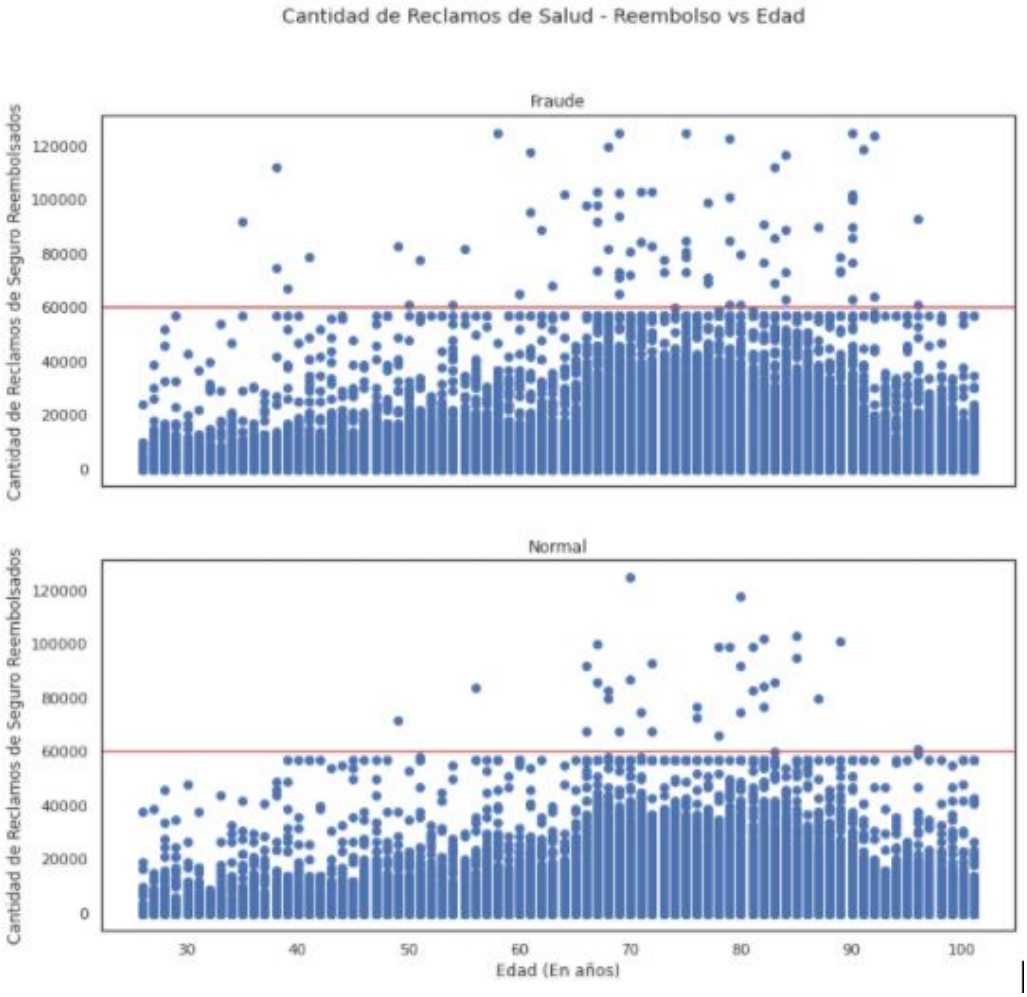
Figura 12: Principales 20 médicos involucrados en fraudes de atención médica




Fuente: Elaboración propia

Se observa que en la Figura 12, el recuento de participación de los médicos tratantes y la naturaleza del proveedor con el que están trabajando, identificando si es fraudulento o no. Ahora categorizamos los reclamos fraudulentos y no fraudulentos basándonos en la cantidad de reclamos de salud. En la Figura 13 podemos ver los reclamos de salud comparando reembolso y edad.

Figura 13: Principales 20 médicos involucrados en fraudes de atención médica




Fuente: Elaboración propia

De la Figura 13, vemos que la ocurrencia de casos potencialmente fraudulentos es más frecuente en los grupos de menor edad (30-70 años) en comparación de los grupos de edad avanzada (70+ años). La edad es una característica importante para diferenciar el comportamiento fraudulento y no fraudulento.

Dentro del análisis exploratorio realizado, se hizo uso de feature engineering o ingeniería de características [52] para mejorar el rendimiento de los algoritmos de *machine learning* seleccionados. Al agregar datos del conjunto de entrenamiento al conjunto de prueba nos ayudará a obtener una mejor puntuación en las pruebas, ya que podemos observar que no todos los niveles de variables están presentes en los datos de prueba en comparación con los de entrenamiento. Entonces, agregamos datos del conjunto de entrenamiento al de pruebas para sacar nuevas características y usar únicamente los datos de prueba para evaluar los resultados. Luego de combinar los conjuntos de datos de prueba y entrenamiento, verificamos que el primer registro haya sido agregado al conjunto de pruebas correctamente, para empezar a encontrar características agrupadas según las columnas del conjunto de datos.

Además de las exploraciones y visualizaciones básicas, se utilizarán métodos para identificar pistas de fraude y abuso. Uno de esos métodos sencillos es la agrupación basada en similitud. En este método básicamente se agrupan todos los registros por *ProcedureCodes, DiagnosisCodes y Provider*. Un ejemplo sería si tenemos un conjunto de datos con códigos de procedimiento solo para el procedimiento X, luego agrupamos y verificamos las cantidades promedio involucradas en cada nivel de procedimiento y analizamos su comportamiento.

Para una mejor comprensión de la agrupación de características compartidas ver el Anexo 1. Se agruparon características compartidas para las siguientes columnas del conjunto de datos: *Provider, BeneID, OtherPhysician, AttendingPhysician, DiagnosisGroupCode,ClmAdmitDiagnosisCode,ClmProcedureCode_1,ClmProcedureCode_2,ClmProcedureCode_3 ,ClmDiagnosisCode_1, ClmDiagnosisCode_2, ClmDiagnosisCode_3.*

Los reclamos son presentados por el proveedor, por lo que el fraude puede ser crimen organizado. Por lo tanto, verificamos el recuento de reclamos presentados por los proveedores y cuando los pares cómo: *Provider + BeneID, Provider + AttendingPhysician, Provider + ClmAdmitDiagnosisCode, Provider + ClmProcedureCode_1, Provider + ClmDiagnosisCode_1* esten juntos.

Al verificar las tuplas de los reclamos presentados por los proveedores en las diferentes combinaciones, podemos observar que si solo tomamos los primeros 2 caracteres del código de diagnóstico con el fin de agruparlos, podemos terminar creando una matriz dispersa bastante grande, ya que cada columna o 'código' genera más de 120 columnas ficticias. Esto aumentará el tiempo de cálculo y perdería explicabilidad.

## 7.2. Preprocesamiento de datos

Para el preprocesamiento del conjunto de datos resultante de la sección anterior se inició agregando ceros a las columnas numéricas. Posteriormente se eliminó las columnas innecesarias ya que en la sección anterior agrupamos en función de estas columnas y se obtuvo la mayor cantidad de información posible de ellas. También se realizó la conversión de tipos de las variables género y raza a categóricas. Se crearon *dummies* para las columnas categóricas y se convirtieron los valores de respuesta a 1 y 0, donde '1' significa 'Si' y '0' significa 'No'.

Se hizo la selección de los datos relacionados con el conjunto de prueba del conjunto de datos donde estaban combinados. Con ello se eliminaron los datos de entrenamiento agregados al conjunto de prueba. Se agregaron reclamos de proveedores únicos para los conjuntos de entrenamiento y prueba. También se dividieron los datos de entrenamiento para validación, para ello separamos el objetivo y los proveedores de las variables independientes, para ello se creó una variable y. Posteriormente estandarizamos mediante el uso de la función *StandardScaler* para la transformar los valores a su forma z, donde el 99.7 % de los valores oscilan entre -3 y 3. Por último, dividimos los datos en entrenamiento y validación utilizando el parámetro '*stratify* = y' el cual se asegura de que la distribución sea equitativa entre sí, tanto para el entrenamiento cómo en la validación.

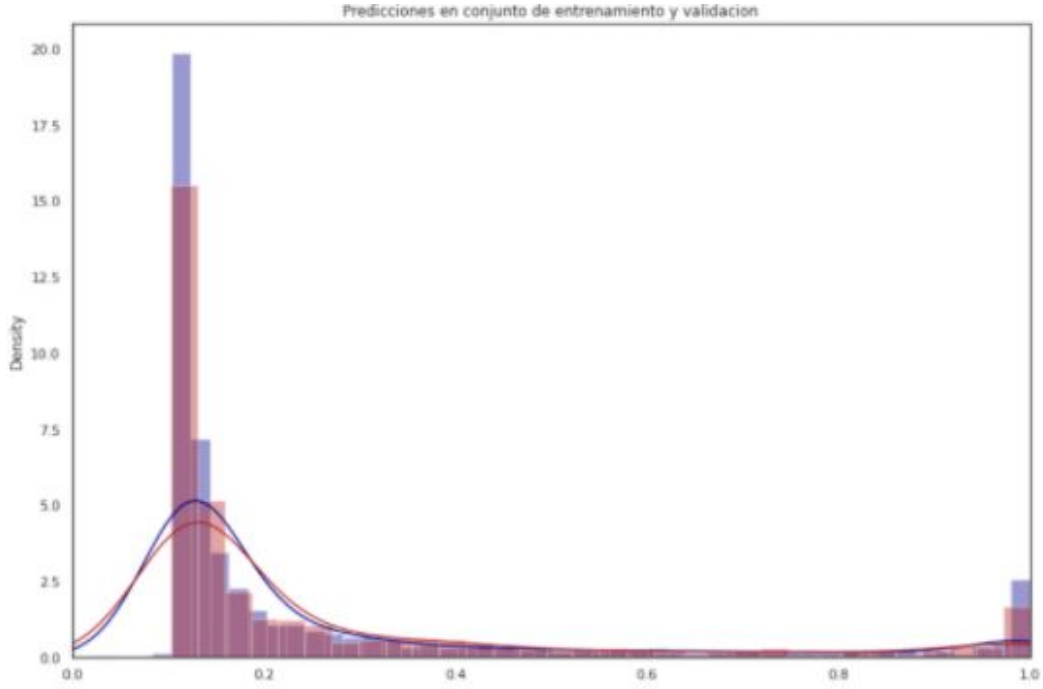
Finalmente para el preprocesamiento se separaron dos conjuntos de datos, 80 % para entrenamiento y 20 % para pruebas y evaluación.

## 7.3. Implementación del modelo

### 7.3.1. Regresión Logística (RL)

**Regresión Logística: Entrenamiento** Para el entrenamiento del modelo de regresión logística se inició con el uso de la función *LogisticRregressionCV* de la biblioteca *sklearn.linear_model*. Esta función utiliza la validación cruzada estratificada de *k-fold* al evaluar nuestros modelos, donde k = 10. La estratificación asegura que todos los pliegues tengan una representación de clase que coincida con la proporción de los datos originales, lo cual es importante cuando se trata de datos en gran parte desequilibrados. Los datos de entrenamiento se dividen uniformemente en diez partes, donde se usarán cuatro veces para entrenar el modelo y el doblez restante prueba el modelo. Este proceso se repite 10 veces, lo que permite que cada pliegue sea una oportunidad como pliegue de prueba, lo que garantiza que todo el conjunto de datos se aproveche por completo para el entrenamiento y la validación. Los datos de entrada tendrían una forma cómo '*n_samples / (n_classes * np.bincount(y))*'. Posteriormente se predijo la probabilidad de 1 y 0 para los conjuntos de entrenamiento y validación. De la Figura 14 se infiere el rendimiento del modelo en el conjunto de entrenamiento y validación.

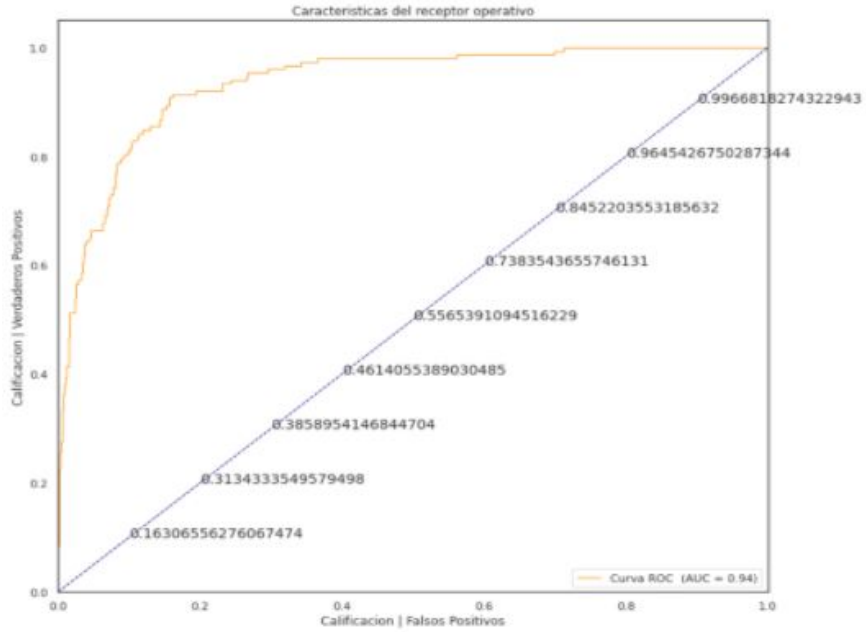
Figura 14: Predicciones en conjunto de entrenamiento y validación.




Fuente: Elaboración propia

En la Figura 15 se muestra la curva de características del receptor operativo (ROC) dando a conocer la calificación de los verdaderos positivos y falsos negativos.

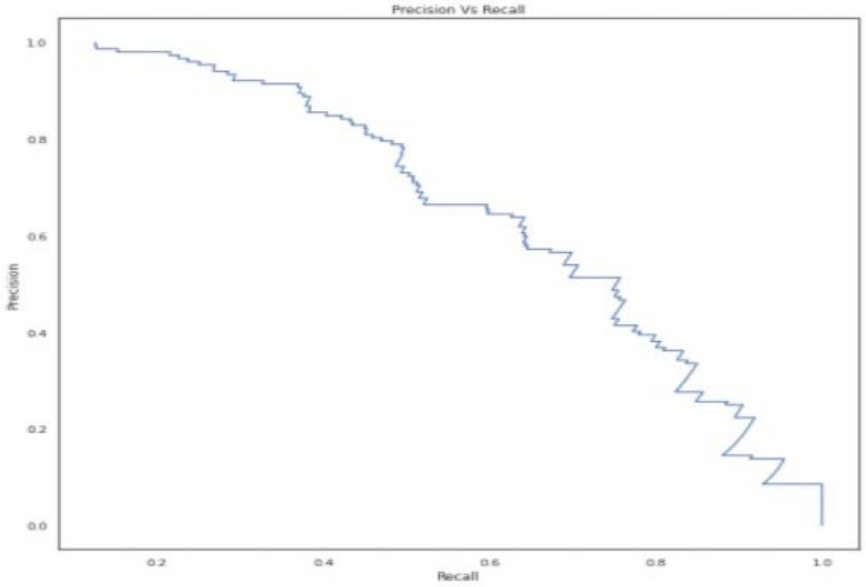
Figura 15: Curva ROC




Fuente: Elaboración propia

Del gráfico anterior se infiere que el modelo de regresión logística obtuvo un área bajo la curva ROC del 0.9357. Dados los resultados anteriores, se procedió a graficar la curva generada entre precisión y recuperación del modelo de regresión logística. En la Figura 16 se puede ver la curva generada en el proceso de entrenamiento utilizando en el eje x la recuperación del modelo y en el eje y la precisión que obtuvo.

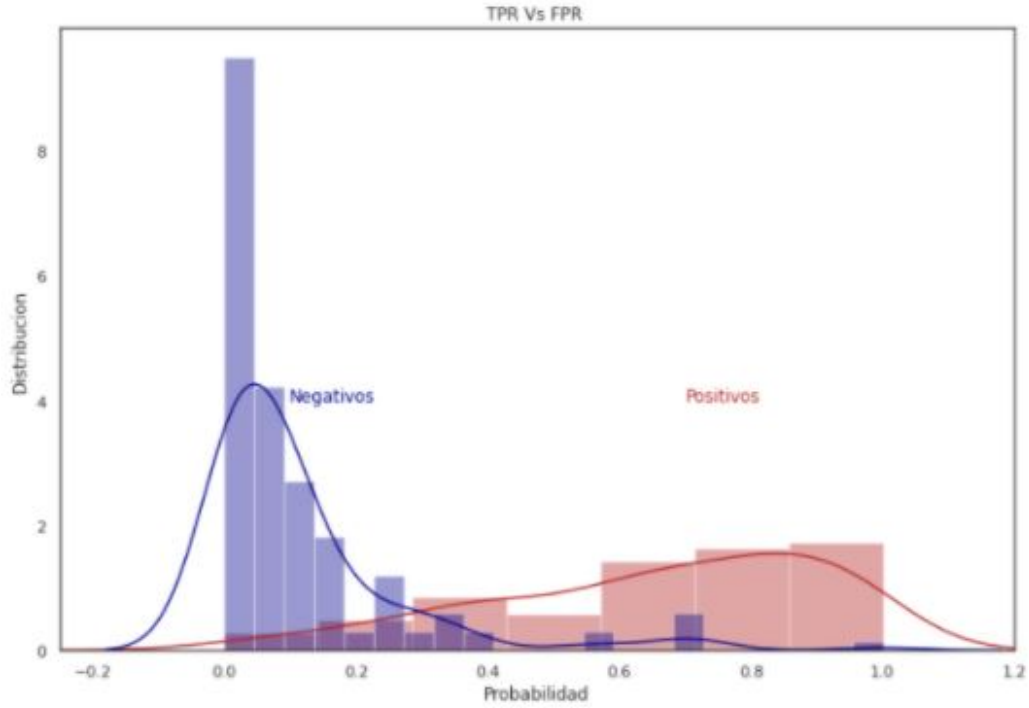
Figura 16: Regresión logística | Precisión Vs Recuperación




Fuente: Elaboración propia

Finalmente se graficó la distribución de los verdaderos positivos (TP) y los falsos positivos (FP) agrupándolos en por su probabilidad de ocurrencia. En la Figura 17 se muestra la distribución agrupada por probabilidad de ocurrencia de los TPR y FPR.

Figura 17: Regresión logística | TPR Vs FPR




Fuente: Elaboración propia

**Regresión logística: Evaluación**

La regresión logística elige la clase que tiene la mayor probabilidad. En nuestro caso la variable respuesta es binaria (fraude, no fraude), el umbral sería de 0.5 donde P(y=0) > P(y=1). Básicamente, es cambiar el umbral de clasificación de 50-50 a un nivel de compensación apropiado. Normalmente se puede optimizar generando una curva de la métrica de evaluación (por ejemplo, la medida F). La limitación es que haciendo este tipo de concesiones absolutas, cualquier modificación en el límite, a su vez puede reducir la precisión de la otra clase. Para la evaluación del modelo de regresión logística se modificó el umbral de clasificación ajustando su probabilidad a 0.60. En el Cuadro 12 se observa que la matriz de confusión para el conjunto de entrenamiento y en el Cuadro 13 la matriz de confusión del conjunto de validación.

Cuadro 12: Regresión logística | Matriz de confusión conjunto de entrenamiento

| | | Predicted Class | |
|---|---|---|---|
| | | No Fraude | Fraude |
| True Class | No Fraude | 270 | 84 |
| | Fraude | 210 | 3223 |

Fuente: Elaboración propia

Cuadro 13: Regresión logística | Matriz de confusión conjunto de validación

| | | Predicted Class | |
|---|---|---|---|
| | | No Fraude | Fraude |
| True Class | No Fraude | 103 | 49 |
| | Fraude | 93 | 1378 |

Fuente: Elaboración propia

En el Cuadro 14 se encuentran la interpretación de las métricas de rendimiento del modelo de regresión logística para la etapa de evaluación en comparación con el rendimiento en el entrenamiento.

Cuadro 14: Rendimiento de modelo para el entrenamiento y validación.

| | **Entrenamiento** | **Validación** |
|---|---|---|
| **Precisión** | 0.9223 | 0.9125 |
| **Sensibilidad** | 0.7627 | 0.6776 |
| **Especificidad** | 0.9388 | 0.9367 |
| **AUC** | 0.9357 | 0.8072 |
| **Puntuación F1** | 0.6474 | 0.5919 |
| **Kappa Puntuación** | 0.5438 | |

Fuente: Elaboración propia

### Regresión logística: Predicción

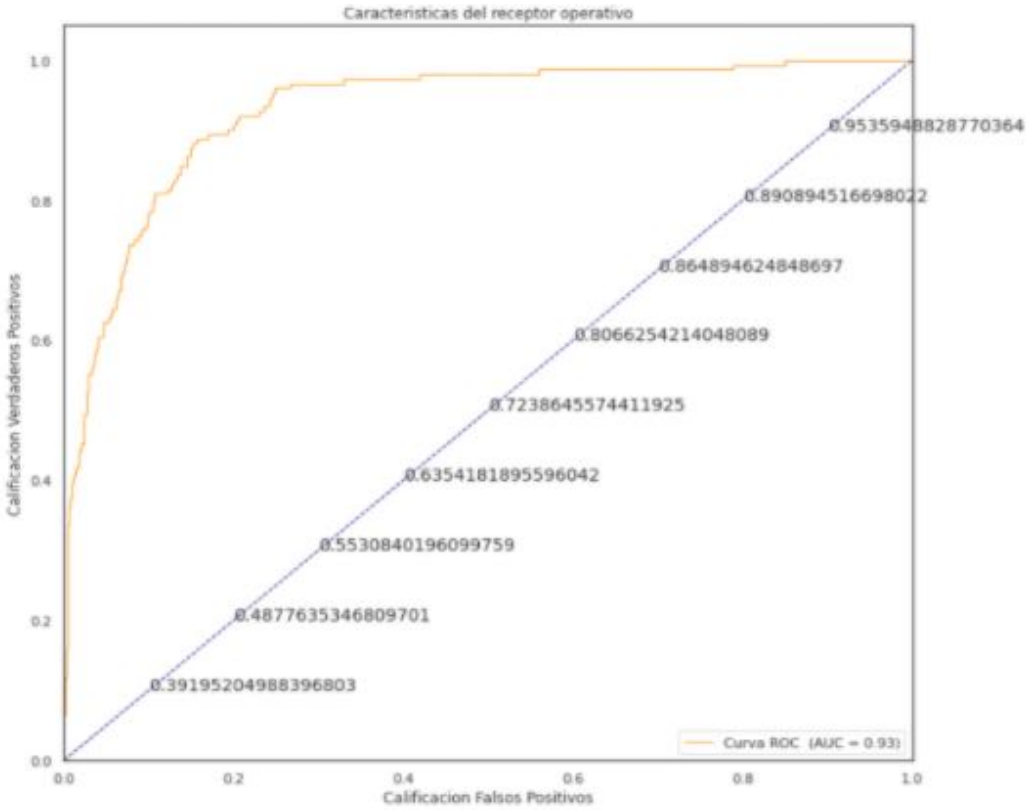
Para esta sección se estará utilizando el conjunto de prueba para realizar predicciones de posible fraude. Cabe mencionar que se utilizó un umbral mayor a 0.60 y se definió una variable booleana para representar cuando un reclamo tiene potencial a ser fraudulento. Posteriormente se reemplazaron los valores 0 y 1 por Sí y No, donde 0 significa No y 1 significa Si. Finalmente se realizó la predicción del modelo de regresión logística para el conjunto de datos de prueba. De 1,353 reclamos el modelo predijo que 1,182 reclamos no eran fraudulentos y 171 casos eran potencialmente fraudulentos. Para ver más a detalle los

casos etiquetados cómo fraudulentos y no fraudulentos, revisar el anexo 2 el cual contiene los reclamos categorizados por el modelo de regresión logística.

### 7.3.2. *Random Forest* (RF)

***Random Forest*: Entrenamiento** Para el entrenamiento del modelo de clasificación Random Forest se utilizó el proporcionado por la biblioteca de modelos de *machine learning sklearn* con el nombre *RandomForestClassifier*. Los parámetros que fueron configurados para el uso del algoritmo son los siguientes: *n_ estimators* = 500, *class_ weight = 'balanced'*, *random_ state = 123, max_ depth =4*. El parámetro *max_ depth* define la profundidad que tendrán los árboles generados, en nuestro caso su tamaño máximo será de 4. Posteriormente se ajustó el modelo con el conjunto de entrenamiento para iniciar el entrenamiento. En la Figura 18 se muestra la curva de características del receptor operativo (ROC) mostrando la calificación de los verdaderos positivos y falsos negativos en el conjunto de entrenamiento.

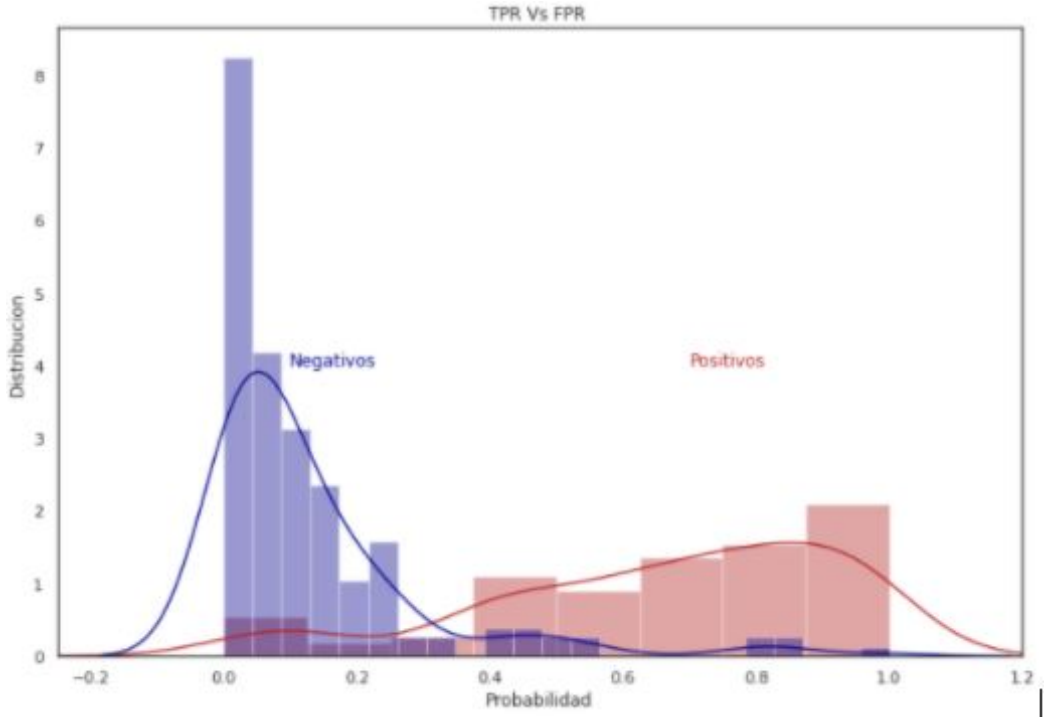
Figura 18: *Random forest* | TPR Vs FPR




Fuente: Elaboración propia

En el gráfico anterior se infiere que el valor AUC del RF fue de 0.93 para el conjunto de entrenamiento. Dados los resultados anteriores, se procedió a graficar la distribución que tuvieron los TP y FP agrupados por su probabilidad de ocurrencia. En la Figura 19 podemos ver las distribuciones agrupadas de los TP y FP.

Figura 19: TPR vs FPR RF




Fuente: Elaboración propia

### *Random Forest*: Evaluación

Para la evaluación del modelo RF se ajustó el umbral a 0.5 tanto en el conjunto de entrenamiento cómo en el conjunto de prueba. También se generaron matrices de confusión para los conjuntos de entrenamiento y prueba con medidores de rendimiento cómo; precisión, sensibilidad, especificidad, AUC y puntuación F1. En el Cuadro 15 se muestra la matriz de confusión para el conjunto de entrenamiento y en el Cuadro 16 la matriz de confusión del conjunto de validación.

Cuadro 15: RF | Matriz de confusión conjunto de entrenamiento.

| | | Predicted Class | |
|---|---|---|---|
| | | No Fraude | Fraude |
| True Class | No Fraude | 320 | 34 |
| | Fraude | 388 | 3045 |

Fuente: Elaboración propia

Cuadro 16: RF | Matriz de confusión conjunto de validación

| | | Predicted Class | |
|---|---|---|---|
| | | No Fraude | Fraude |
| True Class | No Fraude | 124 | 28 |
| | Fraude | 181 | 1290 |

Fuente: Elaboración propia

En el Cuadro 17 se encuentran la interpretación de las métricas de rendimiento del modelo de *random forest* para la etapa de evaluación en comparación con el rendimiento en el entrenamiento.

Cuadro 17: Rendimiento de modelo para el entrenamiento y validación.

| | **Entrenamiento** | **Validación** |
|---|---|---|
| **Precisión** | 0.8885 | 0.8712 |
| **Sensibilidad** | 0.9039 | 0.8157 |
| **Especificidad** | 0.8869 | 0.8769 |
| **AUC** | 0.9311 | 0.8463 |
| **F1-Score** | 0.6026 | 0.5426 |
| **Kappa Score** | 0.4773 | |

Fuente: Elaboración propia

Finalmente, se identificaron las principales 20 características del conjunto de datos que afectan el rendimiento del modelo RF realizando una puntuación por variable. En el Cuadro 18 se muestran las 20 variables y su importancia para el rendimiento del modelo.

Cuadro 18: Principales 20 características que afectan el modelo RF y su puntuación de importancia

| Variable | Importancia |
|---|---|
| PerProviderAvg_InscClaimAmtReimbursed | 0.08 |
| InscClaimAmtReimbursed | 0.07 |
| PerAttendingPhysicianAvg_InscClaimAmtReimbursed | 0.07 |
| PerOperatingPhysicianAvg_InscClaimAmtReimbursed | 0.06 |
| PerClmAdmitDiagnosisCodeAvg_InscClaimAmtReimbursed | 0.04 |
| PerClmAdmitDiagnosisCodeAvg_DeductibleAmtPaid | 0.04 |
| PerClmDiagnosisCode_1Avg_DeductibleAmtPaid | 0.04 |
| PerOperatingPhysicianAvg_IPAnnualReimbursementAmt | 0.03 |
| ClmCount_Provider_ClmDiagnosisCode_7 | 0.03 |
| ClmCount_Provider_ClmDiagnosisCode_8 | 0.03 |
| ClmCount_Provider_ClmDiagnosisCode_9 | 0.03 |
| DeductibleAmtPaid | 0.02 |
| AdmitForDays | 0.02 |
| PerProviderAvg_DeductibleAmtPaid | 0.02 |
| PerAttendingPhysicianAvg_DeductibleAmtPaid | 0.02 |

Fuente: Elaboración propia

#### *Random Forest*: Predicción

En esta sección fue utilizado el conjunto de datos de prueba para realizar las predicciones de reclamos posiblemente fraudulentos. Fue definida una variable booleana que representa cuando un reclamo tiene potencial de ser fraudulento y cuando no. Los valores booleanos fueron reemplazados, donde 0 significa No y 1 significa Sí. Finalmente, el modelo realizó la predicción de 1353 reclamos del conjunto de prueba, categorizando 1,094 reclamos cómo no fraudulentos y 259 cómo fraudulentos. En el Anexo 3 se puede observar los reclamos categorizados por el modelo de random forest.

### 7.3.3. Red Neural *Autoencoder*

#### Red Neural *Autoencoder*: Entrenamiento

Antes de iniciar con el entrenamiento de la red neural *autoencoder*, se realizará un análisis de componentes principales (PCA). Primero se utilizará la función *Standard Scaler* para escalar los conjuntos de datos de entrenamiento y prueba. Luego utilizamos la varianza máxima del paquete PCA incluido en la biblioteca *sklearn*. Se toman 29 componentes del PCA y se realiza una explicación de la varianza dada por el PCA. Transformamos los conjuntos de datos de entrenamiento y prueba basándonos en los componentes del entrenamiento y posteriormente lo convertimos en un *dataframe*. Finalmente, agregamos nuestra variable objetivo a los datos de entrenamiento.

Después de realizar el análisis PCA ya podemos iniciar con nuestro entrenamiento. Como primer paso convertimos nuestros conjuntos de datos en arreglos y dividimos los datos en entrenamiento y prueba. Al tener nuestros arreglos divididos, vamos a utilizar los registros no fraudulentos para entrenar el *autoencoder* y luego verificar el umbral de reconstrucción de error en los registros fraudulentos. Para ello vamos a separar los registros fraudulentos del conjunto de entrenamiento y agregamos los registros fraudulentos del conjunto de entrenamiento al conjunto de prueba. Separamos el 20 % de los registros aleatoriamente en prueba y evaluación. Posteriormente separamos la variable independiente y nuestra variable clasificadora para luego expandir las dimensiones de nuestra variable clasificadora y concatenarla después.

Para la construcción de la red *autoencoder*, primero definimos la capa de entrada que tendrá el mismo numero de elementos que cada ejemplo del conjunto de entrenamiento, siendo 29 neuronas. En la segunda capa de codificación tenemos 15 neuronas utilizando cómo función de activación RELU. Finalmente tenemos la tercera capa de reconstrucción con 29 neuronas siendo el mismo tamaño de la capa inicial. Para el entrenamiento se utilizó el método de adam y cómo función de pérdida utilizaremos la métrica del error cuadrático medio ya que cada neurona puede ser un modelo de regresión lineal Definimos el número de épocas o iteraciones en 100 y un tamaño de 32 para los lotes (o numero de registros de entrenamiento utilizados en una iteración). Se decidieron esas configuraciones después de realizar pruebas con diferentes combinaciones ya que se buscaba identificar la que mejores resultados brindara. Como función de pérdida utilizaremos la métrica del error cuadrático medio ya que cada neurona puede ser un modelo de regresión lineal. Finalmente, entrenamos el modelo y guardamos su historial. Para ver las predicciones realizadas por el modelo, se separaron los registros fraudulentos y los no fraudulentos para recuperar sus errores por separado. Para un mejor entendimiento de los ajustes realizados para la recuperación de error y categorización de los reclamos fraudulentos y no fraudulentos revisar código. En el Cuadro 19 podemos observar la matriz de confusión para el entrenamiento y en el Cuadro 20 podemos ver los resultados obtenidos en el entrenamiento de la red neural *autoencoder*.

Cuadro 19: Autoencoder | Matriz de confusión conjunto de entrenamiento

| | | Predicted Class | |
|---|---|---|---|
| | | No Fraude | Fraude |
| True Class | No Fraude | 350 | 57 |
| | Fraude | 412 | 376 |

Fuente: Elaboración propia

Cuadro 20: Rendimiento del *autoencoder* para el entrenamiento.

| | Entrenamiento |
|---|---|
| **Recuperación** | 0.8599 |
| **Precision** | 0.4593 |
| **Exactitud** | 0.6075 |
| **F1-Score** | 0.5988 |

Fuente: Elaboración propia

**Red Neural *Autoencoder*: Evaluación** Para la evaluación o validación del autoencoder únicamente se realizaron las predicciones en el conjunto de datos de prueba. En el Cuadro 21 se puede observar la matriz de confusión para el conjunto de pruebas. En el Cuadro 22 podemos ver el rendimiento del autoencoder para el conjunto de pruebas.

Cuadro 21: *Autoencoder* | Matriz de confusión conjunto de entrenamiento.

| | | Predicted Class | |
|---|---|---|---|
| | | No Fraude | Fraude |
| True Class | No Fraude | 84 | 15 |
| | Fraude | 108 | 92 |

Fuente: Elaboración propia

Cuadro 22: Rendimiento del *autoencoder* para el entrenamiento.

| | Validation |
|---|---|
| **Recuperación** | 0.8484 |
| **Precision** | 0.4375 |
| **Exactitud** | 0.5886 |
| **F1-Score** | 0.5773 |

Fuente: Elaboración propia

**Red Neural *Autoencoder*: Predicción** Para la evaluación de predicción del *autoencoder* fue realizada en un conjunto de datos no vistos por el modelo. En el Cuadro 23 se encuentran los resultados obtenidos con el modelo de mejor rendimiento. Cabe mencionar que 1,353 reclamos evaluados, el modelo realizó una predicción de 759 reclamos fraudulentos y 594 reclamos no fraudulentos.

Cuadro 23: *Autoencoder* | Rendimiento del *autoencoder* en la predicción de fraude potencial.

| | Prediction |
|---|---|
| **Recuperación** | 0.8888 |
| **Precision** | 0.2002 |
| **Exactitud** | 0.5373 |
| **F1-Score** | 0.3268 |
| **Kappa** | 0.3848 |

Fuente: Elaboración propia

En el Cuadro 24 se muestran 10 proveedores categorizados por el *autoencoder* se establecen cómo potencialmente fraudulentos.

Cuadro 24: Listado de 10 predicciones categorizadas por el *autoencoder*

| Provider | PotentialFraud |
|---|---|
| PRV51002 | YES |
| PRV51006 | NO |
| PRV51009 | NO |
| PRV51010 | NO |
| PRV51018 | NO |
| PRV51019 | YES |
| PRV51020 | NO |
| PRV51022 | NO |
| PRV51028 | NO |
| PRV51033 | YES |

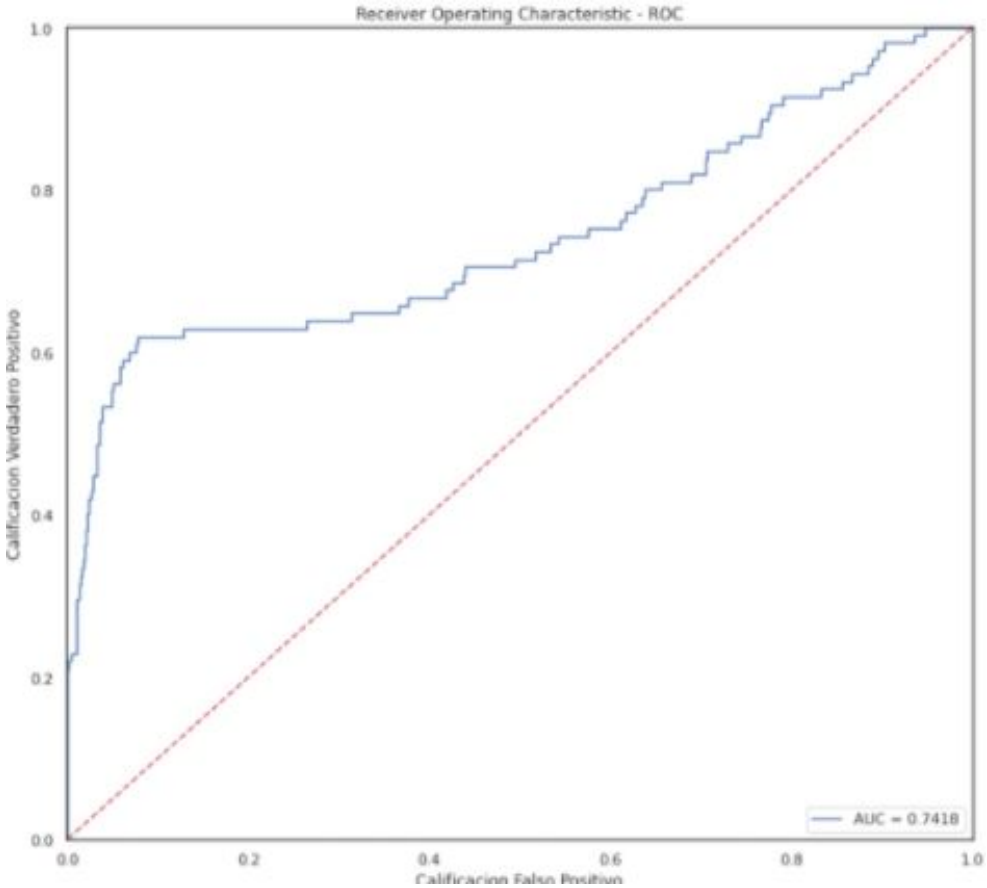
Fuente: Elaboración propia

### 7.3.4. Red Neural *Autoencoder* (Dos capdas ocultas)

Con el objetivo de encontrar el modelo que detecte con la mayor precisión el posible fraude por proveedores en reclamos de salud se decidió agregar dos capas ocultas al modelo neural. Para la construcción de la nueva red *autoencoder*, definimos la capa de entrada que tendrá el mismo número de elementos que cada ejemplo del conjunto de entrenamiento, siendo 29 neuronas. En la segunda capa de codificación tenemos 14 neuronas utilizando la tangente hiperbólica cómo función de activación. En la tercera capa de codificación tenemos 7 neuronas utilizando cómo función de activación RELU. Para la cuarta capa se inicia la decodificación con 7 neuronas utilizando cómo función de activación la tangente hiperbólica. En la quinta capa se tienen 14 neuronas utilizando RELU cómo función de activación. Finalmente tenemos nuestra última capa de reconstrucción con 29 neuronas. Posteriormente construimos el autoencoder donde basta especificar la capa de entrada y la capa del deco-

dificador. Para el entrenamiento se utilizó el método de adam y como función de pérdida utilizaremos la métrica del error cuadrático medio ya que cada neurona puede ser un modelo de regresión lineal. Por último configuramos un checkpointer donde almacenaremos el modelo entrenado. En la Figura 20 se demuestra la curva ROC-AUC donde determina la calificación de los TP y FP.

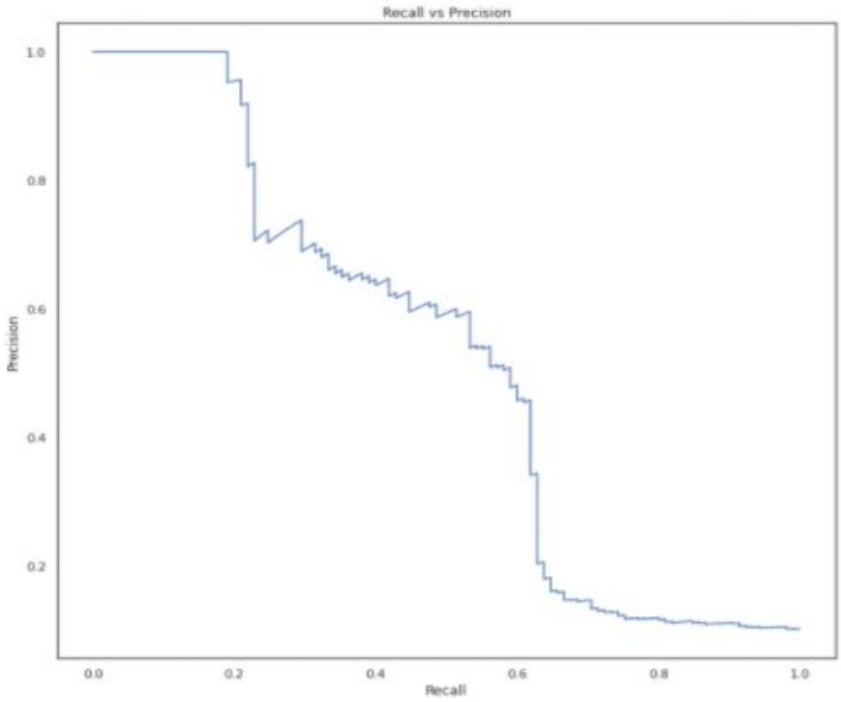
Figura 20: Curva ROC-AUC: *Autoencoder* 2 Capas Oculta




Fuente: Elaboración propia

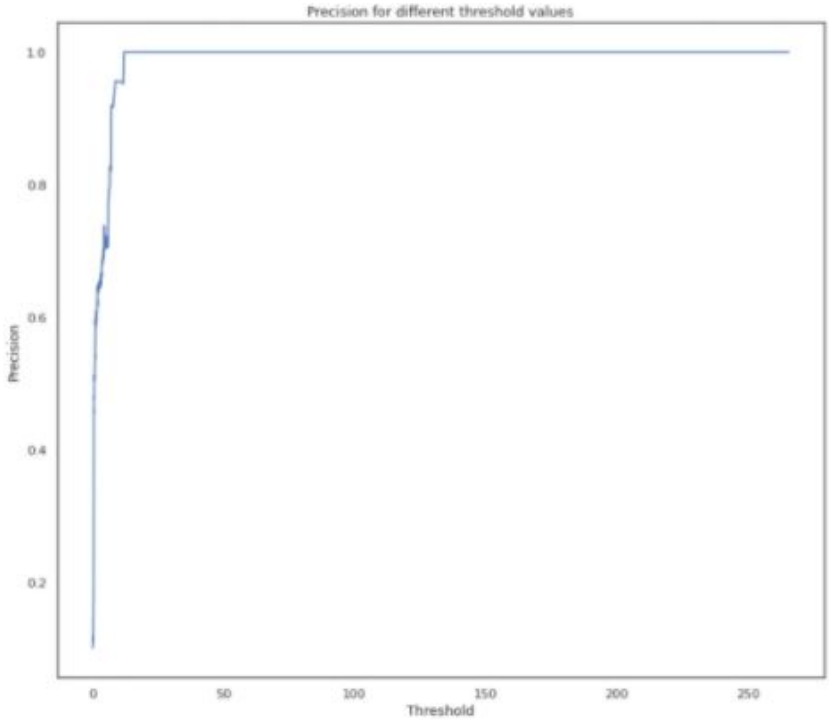
Del gráfico anterior se infiere que la red autoencoder con dos capas ocultas tuvo una calificación AUC de 0.7418. Dado los resultados anteriores se procedió a graficar la curva de precisión-recuperación para resumir el equilibrio entre la tasa de TP y el los valores predictivos positivos del modelo que utiliza diferentes umbrales de probabilidad. En la Figura 21 se demuestra la curva generada de la precisión-recuperación.

Figura 21: Curva recuperación vs precision



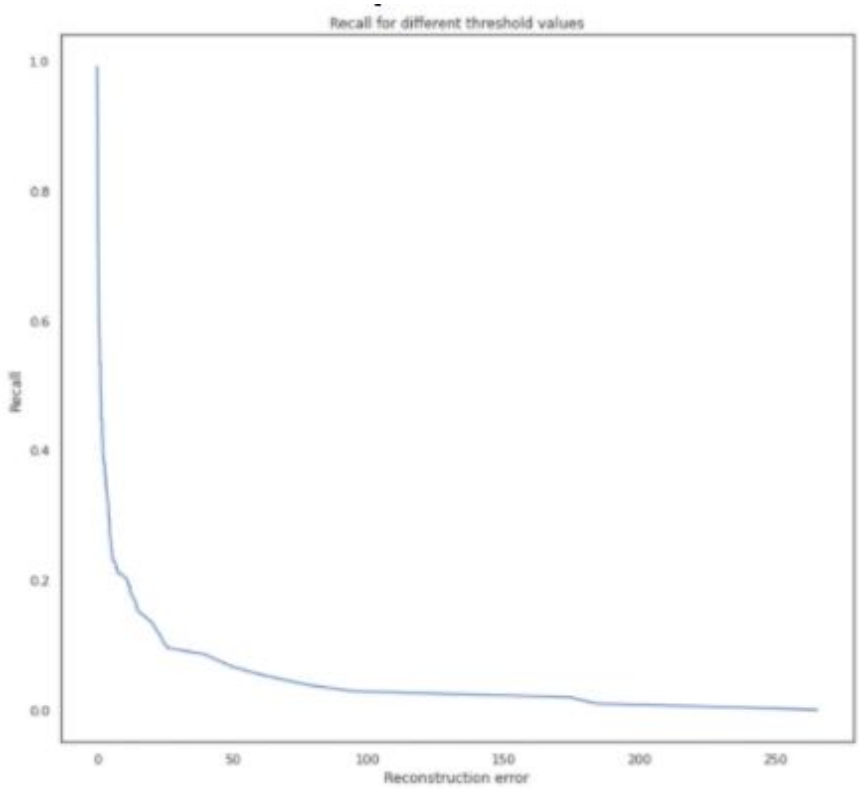

Fuente: Elaboración propia

Figura 22: Precisión para diferentes valores de umbral




Fuente: Elaboración propia

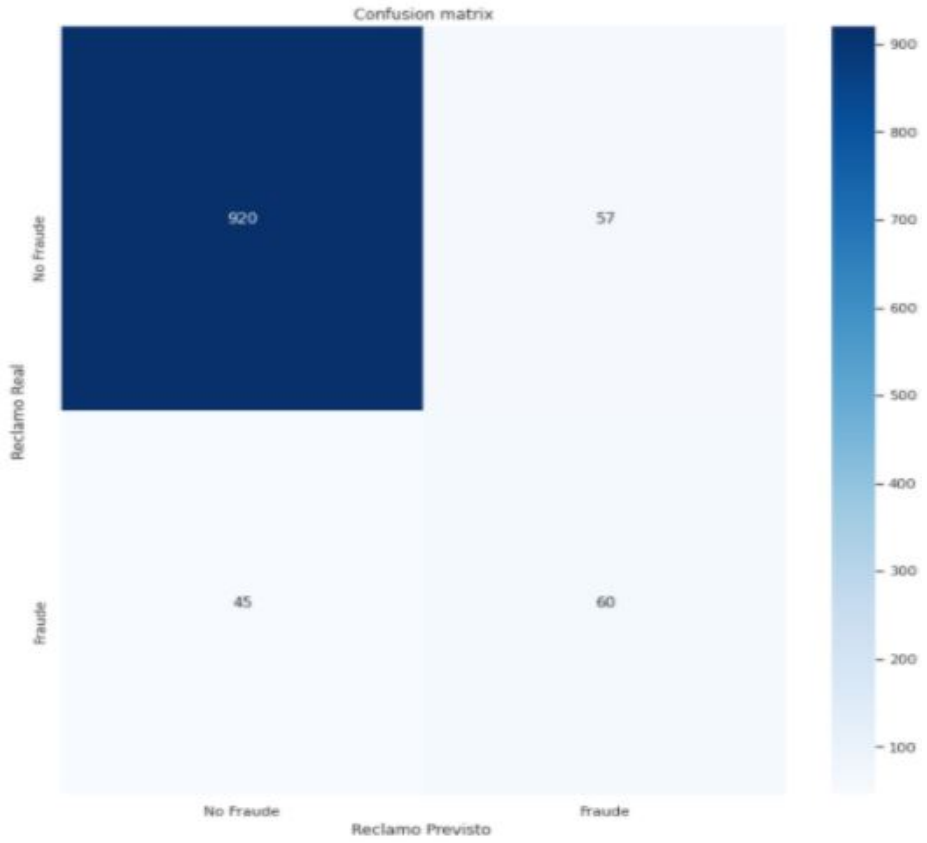
Figura 23: Recuperación para diferentes valores de umbral




Fuente: Elaboración propia

En la Figura 22 tenemos el comportamiento de la precisión para diferentes valores de umbrales. Al igual que la Figura 23 se muestra el comportamiento para la recuperación con diferentes valores de umbral.

Figura 24: Matriz de confusión para predicción de fraude potencial.




Fuente: Elaboración propia

Cuadro 25: Rendimiento del *autoencoder* con dos capas ocultas en la predicción de fraude potencial.

| | Prediction |
|---|---|
| **Exactitud** | 0.9057 |
| **Sensitivity** | 0.5714 |
| **Specificity** | 0.9416 |
| **Kappa Value** | 0.4881 |
| **AUC** | 0.7565 |
| **F1-Score** | 0.5705 |

Fuente: Elaboración propia

Con tan solo 2 capas ocultas y 100 épocas se logró un F1-Score de 0.57. El modelo parece detectar varias transacciones fraudulentas. La cantidad de transacciones no fraudulentas clasificadas cómo fraudulentas es alta. Como recomendación, se podría establecer un umbral para crear un equilibrio entre la predicción de transacciones fraudulentas y no fraudulentas. También agregar más datos al modelo y entrenarlo mejorará el rendimiento de la detección de nuevos patrones de fraude ayudando a comprender mejor el comportamiento de los proveedores fraudulentos.

# CAPÍTULO 8

## Discusión de resultados

Cuadro 26: Rendimiento final de los algoritmos en la predicción de fraude potencial.

| | ~Exactitud | ~AUROC | ~Puntuación F1 |
|---|---|---|---|
| **Regresión Logística** | 0.9125 | 0.8072 | 0.5919 |
| **Random Forest** | 0.8712 | 0.8463 | 0.5426 |
| **Autoencoder** | 0.5373 | 0.6451 | 0.3268 |
| **Autoencoder (2 capas ocultas)** | 0.9057 | 0.7565 | 0.5705 |

Fuente: Elaboración propia

Las aseguradoras realizan análisis de cada uno de sus reclamos en busca de detectar conductas sospechosas [47]. Este tipo de conductas pueden llevar al descubrimiento de diferentes tipos de fraude cómo; facturación de servicios no prestados, reclamos duplicados, cobros por servicios más costosos de lo que realmente son, etc. La investigación que conlleva la filtración de grandes cantidades de reclamos realizada por auditores e investigadores de forma manual, puede llegar a ser tediosa y muy ineficiente. Actualmente existen sistemas que utilizan el enfoque de la ciencia de datos y aprendizaje automático para la detección de fraudes [34,35].

Para brindar una solución a la problemática mencionada anteriormente, se hizo uso del conjunto de datos DE-SynPUF proporcionados por CMS los cuales contienen información de beneficiarios de Medicare del año 2008 y sus reclamos hechos del 2008 al 2010. También se hizo uso de listado de personas y entidades excluidas (LEIE) el cual proporciona datos de médicos y proveedores que han cometido fraudes en el mundo real. Por medio de la minería de datos, se unificó la información de los beneficiarios con los reclamos realizados y

se etiquetaron los reclamos con posibilidad de fraude utilizando el listado de proveedores fraudulentos LEIE. El análisis y exploración del conjunto de datos detallado nos permite asegurar que la combinación de los datos de CMS y LEIE son lo suficientemente diversos para garantizar que la predicción de los modelos seleccionados se adapten a varios escenarios posibles. También son abundantes e imparciales, garantizando que los modelos puedan generalizarse adecuadamente durante la inferencia.

Para la implementación fue necesario realizar un preprocesamiento de los datos basándonos en los requisitos de entrada de los modelos seleccionados. Entre los cambios realizados están la conversión de variables categóricas a numéricas, reformateo de los datos y estandarización de los mismo. Para estimar el rendimiento de los modelos de predicción se dividió el conjunto de datos en 80 % para entrenamiento y 20 % para evaluación. El conjunto de datos de entrenamiento es utilizado para ajustar el modelo y "enseñarle" qué estructura tiene un reclamo fraudulento y uno normal. El conjunto de prueba es utilizado para hacer predicciones y hacer una comparación con los valores esperados. El objetivo es estimar el rendimiento de los modelos sobre nuevos datos que no fueron usados para el entrenamiento.

Los algoritmos de categorización utilizados fueron; regresión logística, *random forest, autoencoder y autoencoder* con dos capas ocultas. Luego del entrenamiento y evaluación de todos los algoritmos, se determinó que el algoritmo que mejor precisión tuvo para la detección de reclamos por proveedores fraudulentos fue el de regresión logística. Este algoritmo obtuvo una exactitud de 0.91 el cual nos indica que detectó con un 91 % de precisión los casos positivos correctamente identificados de todos los casos predichos. También tuvo una puntuación F1 de 0.59 dándonos una medición de los casos clasificados incorrectamente siendo de gran ayuda una puntuación alta en un conjunto de datos desequilibrado cómo el utilizado. Finalmente la puntuación AUROC obtenida por el algoritmo fue de 0.81, esto nos indica la capacidad del algoritmo para distinguir entre reclamos fraudulentos y normales.

El resultado obtenido por esta investigación ayudará a la detección del posible fraude cometido por proveedores en reclamos de salud. La detección temprana de los reclamos fraudulentos permitirá reducir los costos operativos por atención médica, aumentando la cobertura y mejora del nivel de servicio proporcionado por los programas de salud. Esto permitirá el acceso a la salud a más población dándoles una mejor calidad de vida.
=

# CAPÍTULO 9

# Conclusiones

1. Los modelos implementados tuvieron un desempeño consistente con una exactitud de 0.90, una puntuación AUROC de 0.80 y una puntuación F1 de  0.59.

2. La implementación de ingeniería de características mejoró significativamente la precisión en predicción de los algoritmos implementados.

3. El uso de técnicas de minería de datos cómo el análisis exploratorio permitió el reconocimiento de patrones de fraude en el conjunto de datos combinado de CMS y LEIE.

4. Se identificó que el modelo de *machine learning* con la mejor precisión en detección del posible fraude por proveedores en reclamos de salud dado el conjunto de datos analizados fue el de regresión logística.

# CAPÍTULO 10

## Recomendaciones

1. Agregar más datos de fraude al conjunto de datos de entrenamiento ayudará a predecir con mayor precisión el comportamiento fraudulento invisible poco a poco.
2. El uso de *Ensemble Methods* con ajuste de parámetros puede mejorar el rendimiento de los modelos.
3. La investigación realizada ayudará a predecir el fraude de proveedores, lo que será útil para que las compañías de seguros examinen minuciosamente los reclamos hechos por los proveedores categorizados con posible fraude.
4. Una mejora adicional en la investigación ayudará a las aseguradoras guatemaltecas tanto públicas como privadas a tomar decisiones contra los proveedores fraudulentos de salud y ayudará a enmendar las reglas y regulaciones en este campo.
5. La implementación de los modelos ayudará a detectar redes de médicos, proveedores y beneficiarios fraudulentos.
6. Esta investigación ayudará a mejorar la salud y economía al reducir la inflación causada por los fraudes, lo que reducirá las primas de seguros, lo que ciertamente no hará que la salud se convierta en algo costoso.

# CAPÍTULO 11

## Bibliografía


1. *Administration for Community Living: Profile of older Americans.* 2019. `https://acl.gov/aging-and-disability-in-america/data-and-research/profile-older-americans.` Visitado el 19 de Octubre 2020.

2. Atena. *The facts about rising health care costs.* `http://www.aetna.com/news/2010/CostDrivers.pdf.Visitadoel19deOctubre2020.`

3. Bauder RA, Khoshgoftaar TM, Richter A, Herland M. *Predicting medical provider specialties to detect anomalous insurance claims.* In: 2016 IEEE 28th international conference on tools with artificial intelligence (ICTAI). 2016. p. 784–90.

4. Bauder RA, Khoshgoftaar TM, Seliya N. *A survey on the state of healthcare upcoding fraud analysis and detection. Health Serv Outcomes Res Methodology.* 2017;17(1):31–55

5. Bauder RA, Khoshgoftaar TM. *A novel method for fraudulent medicare claims detection from expected payment deviations (application paper).* In: 2016 IEEE 17th international conference on information reuse and integration (IRI). 2016. p. 11–9.

6. Bauder RA, Khoshgoftaar TM. *A probabilistic programming approach for outlier detection in healthcare claims.* In: 2016 15th IEEE international conference on machine learning and applications (ICMLA). 2016. p. 347–54.

7. Bauder RA, Khoshgoftaar TM. *Medicare fraud detection using random forest with class imbalanced big data.* In: 2018 IEEE 19th international conference on information reuse and integration (IRI). 2018. p. 80–87.

8. Bauder RA, Khoshgoftaar TM. *The detection of medicare fraud using machine learning methods with excluded provider labels.* In: FLAIRS conference. 2018. p. 404–9.

9. Branting LK, Reeder F, Gold J, Champney T. *Graph analytics for healthcare fraud risk estimation.* In: 2016 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM). 2016. p. 845–51.

10. *Centers for Medicare  Medicaid Services. Medicare fraud  abuse: prevention, detection, and reporting. 2015.* https://www.cms.gov/Outreach-and-Education/Medicare-Learning-Network-MLN/MLNProducts/downloads/fraud_and_abuse.pdf. Visitado el 19 de Octubre 2020.

11. Chandola V, Sukumar SR, Schryver JC. *Knowledge discovery from massive healthcare claims data.* In: Proceedings of the 19th ACM SIGKDD international conference on knowledge discovery and data mining. ACM; 2013. p. 1312–20.

12. *CMS 2008-2010 Data Entrepreneurs Synthetic PUF.* https://www.cms.gov/Research-Statistics-Data-and-Systems/Downloadable-Public-Use-Files/SynPUFs/DESample01. Visitado el 20 de Octubre 2020.

13. CMS Office of Enterprise Data and Analytics. *Medicare Fee-For Service Provider Utilization  Payment Data Part D prescriber public use fle: a methodological overview.* https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Downloads/Prescriber_Methods.pdf. Visitado el 20 de Octubre 2020

14. CMS Office of Enterprise Data and Analytics. *Medicare Fee-For-Service Provider Utilization  Payment Data Physician and Other Supplier.* https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Downloads/Medicare-Physician-and-Other-Supplier-PUF-Methodology.pdf

15. CMS Office of Enterprise Data and Analytics. *Medicare Fee-For-Service Provider Utilization  Payment Data Referring durable medical equipment, prosthetics, orthotics and supplies public use file: a methodological overview.* https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Downloads/DME_Methodology.pdf. Visitado el 20 de Octubre 2020

16. *CMS: Center for Medicare and Medicaid Services.* https://www.cms.gov/ . Visitado el 19 de Octubre 2020.

17. *CMS: National Provider Identifier Standard (NPI).* https://www.cms.gov/Regulations-and-Guidance/AdministrativeSimplifcation/NationalProvIdentStand/ Visitado el 19 de Octubre 2020.

18. *CMS: Research, Statistics, Data, and Systems.* https://www.cms.gov/research-statistics-data-and-systems/researchstatistics-data-and-systems.html . Visitado el 19 de Octubre 2020.

19. Demchenko Y, Zhao Z, Grosso P, Wibisono A, De Laat C. *Addressing big data challenges for scientific data infrastructure.* In: 2012 IEEE 4th international conference on cloud computing technology and science (CloudCom). 2012. p.614–7.

20. Dieleman JL, Squires E, Bui AL, Campbell M, Chapin A, Hamavid H, Horst C, Li Z, Matyasz T, Reynolds A. *Factors associated with increases in us health care spending,* 1996–2013. Jama. 2017;318(17):1668–78.

21. F. Chollet, *"Building Autoencoders in Keras"*, The Keras Blog, May 2016, https://blog.keras.io/building-autoencoders-in-keras.html. Visitado el 19 de Octubre 2020.

22. Feldman K, Chawla NV. *Does medical school training relate to practice? evidence from big data.* Big Data. 2015;3(2):103–13.

23. Feldstein M. *Balancing the goals of health care provision and financing.* Health Affairs. 2006;25(6):1603–11.

24. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. *The weka data mining software: an update.* ACM SIGKDD Explor Newslett. 2009;11(1):10–8.

25. Henry J. Kaiser family foundation: *Medicare advantage* https://www.kff.org/medicare/fact-sheet/medicare-advantage/ . Visitado el 19 de Octubre 2020.

26. Herland M, Bauder RA, Khoshgoftaar TM. *Medical provider specialty predictions for the detection of anomalous Medicare insurance claims.* In: 2017 IEEE 18th international conference information reuse and integration (IRI). 2017. p. 579–88

27. Herland M, Khoshgoftaar TM, Wald R. *A review of data mining using big data in health informatics.* J Big Data. 2014;1(1):2.

28. John Walker S. *Big data: a revolution that will transform how we live, work, and think.* New York: Taylor  Francis; 2014.

29. Khoshgoftaar TM, Golawala M, Van Hulse J. *An empirical study of learning from imbalanced data using random forest.* In: ICTAI 2007. 19th IEEE international conference on tools with artificial intelligence, 2007, vol. 2. 2007. p. 310–7.

30. Khurjekar N, Chou C-A, Khasawneh MT. *Detection of fraudulent claims using hierarchical cluster analysis.* In: Proceedings IIE annual conference. Institute of Industrial and Systems Engineers (IISE). 2015. p. 2388.

31. Le Cessie S, Van Houwelingen JC. *Ridge estimators in logistic regression.* Appl Stat. 1992;41:191–201.

32. Lehr, D., Ohm, P. (2017). *Playing with the data: what legal scholars should learn about machine learning.* UCDL Rev., 51, 653.

33. *LEIE: Office of Inspector General LEIE Downloadable Databases* https://healthdata.gov/dataset/office-inspector-general-list-excluded-individuals-and-entities Visitado el 19 de Octubre 2020

34. Li J, Huang K-Y, Jin J, Shi J. *A survey on statistical methods for health care fraud detection.* Health Care Manag Sci. 2008;11(3):275–87.

35. Luo, Jake, et al. *"Big Data Application in Biomedical Research and Health Care: A Literature Review."* Biomedical Informatics Insights, Libertas Academica, 19 Jan. 2016, www.ncbi.nlm.nih.gov/pmc/articles/PMC4720168/.

36. M, Hossin, and Sulaiman M.n. *"A Review on Evaluation Metrics for Data Classification Evaluations."* International Journal of Data Mining  Knowledge Management Process, vol. 5, no. 2, 2015, pp. 01–11., doi:10.5121/ijdkp.2015.5201.

37. Manyika J, Chui M, Brown B, Bughin J, Dobbs R, Roxburgh C, Byers AH. *Big data: the next frontier for innovation, competition, and productivity.* New York: McKinsey Global Institute; 2011.

38. McAfee A, Brynjolfsson E, Davenport TH, Patil D, Barton D. *Big Data: the management revolution.* Harvard Bus Rev. 2012;90(10):60–8.

39. Medicare Fraud Strike Force. *Health Care Fraud Strike Force.* https://www.justice.gov/criminal-fraud/strike-force-operations. Visitado el 19 de Octubre 2020.

40. Medicare.gov. *What's medicare.* https://www.medicare.gov/sign-up-change-plans/decide-how-to-get-medicare/whats-medicare/what-is-medicare.html. Visitado el 19 de Octubre 2020

41. Medicare: *US Medicare Program.* https://www.medicare.gov . Visitado el 19 de Octubre 2020.

42. Morris L. *Combating fraud in health care: an essential component of any cost containment strategy.* Health Affairs. 2009;28(5):1351–6.

43. Ohlhorst FJ. *Big Data analytics: turning Big Data into big money.* New York: Wiley; 2012.

44. *OIG: Office of Inspector General Exclusion Authorities US Department of Health and Human Services.* https://oig.hhs.gov/. Visitado el 20 de Octubre 2020

45. *OIG: Office of Inspector General Exclusion Authorities.* https://oig.hhs.gov/exclusions/index.asp. Visitado el 20 de Octubre 2020

46. Pande V, Maas W. *Physician medicare fraud: characteristics and consequences.* Int J Pharm Healthcare Market. 2013;7(1):8–33.

47. Rashidian A, Joudaki H, Vian T. *No evidence of the effect of the interventions to combat health care fraud and abuse:* a systematic review of literature. PLoS ONE. 2012;7(8):41988.

48. Sadiq S, Tao Y, Yan Y, Shyu M-L. *Mining anomalies in medicare big data using patient rule induction method.* In: 2017 IEEE third international conference on multimedia Big Data (BigMM). 2017. p. 185–92.

49. Pedregosa, F., Varoquaux, Ga.[el], Gramfort, A., Michel, V., Thirion, B., Grisel, O., others. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*(Oct), 2825–2830. Visitado el 19 de Octubre 2020.

50. Sklearn.ensemble.RandomForestClassifier. (n.d.). Retrieved from https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.h ml

51. Sklearn.linear_model.LogisticRegression. (n.d.). Retrieved from https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.h ml

52. Turner, C. Wolf, Alexander Fuggetta, Alfonso Lavazza, Luigi. (1998). *Feature Engineering.*

53. U.S. Government, U.S. Centers for Medicare  Medicaid Services. *The Official U.S. Government Site for Medicare.* https://www.medicare.gov/ . Visitado el 19 Octubre 2020

54. US, Medicare Payment Advisory Commission. *Report to the Congress, Medicare Payment Policy.* Washington D.C.: Medicare Payment Advisory Commission; 2007

55. Witten IH, Frank E, Hall MA, Pal CJ. *Data Mining: practical machine learning tools and techniques.* Burlington: Morgan Kaufmann; 2016.

# CAPÍTULO 12

# Anexos

## 12.1. Repositorio

A continuación se adjunta el link al repositorio utilizado para el desarrollo de los algoritmos de *Machine Learning* utilizados para la investigación de "Detección de posible fraude por proveedores en reclamos de salud ".

**Link a repositorio de Github**

https://github.com/josejo911/Tesis-Deteccion-de-Fraude

# CAPÍTULO 13

## Glosario

- **Hiperparámetros:** son valores que se definen antes de que comience el entrenamiento de la red; se inicializan para ayudar a dirigir la red hacia un resultado de entrenamiento positivo. Su efecto está en la máquina / algoritmo de aprendizaje profundo, pero no son afectados por el algoritmo. Sus valores no cambian durante el entrenamiento. Un ejemplo de hiperparámetros son los valores de regularización, las tasas de aprendizaje, el número de capas, etc.
- **Matriz de confusión (matriz de error):** proporciona una ilustración visual del número de coincidencias o desajustes de la anotación de la verdad fundamental a los resultados del clasificador. Una matriz de confusión generalmente se estructura en forma de tabla, donde las filas se llenan con los resultados de observación de la verdad fundamental, y las columnas se llenan con los resultados de inferencia del clasificador.
- **Parámetro de red:** estos son componentes de nuestra red que no se inicializan manualmente. Son valores integrados que la red manipula directamente. Un ejemplo de un parámetro de red son los pesos internos de la red.
- **Recuperación de precisión:** Estas son métricas de rendimiento que se utilizan para evaluar algoritmos de clasificación, sistemas de búsqueda visual y más. Utilizando el ejemplo de evaluación de un sistema de búsqueda visual (encontrar imágenes similares basadas en una imagen de consulta), la precisión captura la cantidad de resultados devueltos que son relevantes, mientras que la recuperación captura la cantidad de resultados relevantes en su conjunto de datos que se devuelven.
- **Sub-ajuste (*Underfitting*):** esto ocurre cuando un algoritmo de aprendizaje automático no puede aprender los patrones en un conjunto de datos. La falta de adaptación se puede solucionar mediante el uso de un mejor algoritmo o modelo que sea más adecuado para la tarea. La adaptación insuficiente también se puede ajustar al reconocer más características dentro de los datos y presentarlas al algoritmo.

- **Sobreajuste *(Overfitting)*:** este problema implica que el algoritmo predice nuevas instancias de patrones que se le presentan, basándose demasiado en instancias de patrones que observó durante el entrenamiento. Esto puede hacer que el algoritmo de aprendizaje automático no se generalice con precisión a datos invisibles. El sobreajuste puede ocurrir si los datos de entrenamiento no representan con precisión la distribución de los datos de la prueba. El sobreajuste se puede solucionar reduciendo el número de características en los datos de entrenamiento y reduciendo la complejidad de la red a través de varias técnicas.