LA EFICIENCIA DE MODELOS SUPERVISADOS (REGRESIÓN LOGISTICA, ÁRBOL DE DECISIÓN Y XGBOOST) EN LA DETECCIÓN DE FRAUDES EN PÓLIZAS DE SEGUROS VEHICULARES

JOSÉ AGUIRRE GONZÁLEZ

UNIVERSIDAD DE CUNDINAMARCA

FACULTAD DE EDUCACIÓN

LICENCIATURA EN MATEMÁTICAS

2023

LA EFICIENCIA DE MODELOS SUPERVISADOS (REGRESIÓN LOGISTICA, ÁRBOL DE DECISIÓN Y XGBOOST) EN LA DETECCIÓN DE FRAUDES EN PÓLIZAS DE SEGUROS VEHICULARES

JOSÉ AGUIRRE GONZÁLEZ

ASESOR: MG. JESÚS ANTONIO VILLARRAGA PALOMINO

TRABAJO DE GRADO PARA OPTAR POR EL TÍTULO DE LICENCIADO EN MATEMÁTICAS

UNIVERSIDAD DE CUNDINAMARCA

FUSAGASUGÁ, COLOMBIA

15 DE MAYO DE 2023

TABLA DE CONTENIDO

TABLA DE ILUSTRACIONES

Resumen

Esta investigación está basada en el análisis y evaluación de modelos que permitan la predicción de fraudes en las solicitudes de reclamación de pólizas vehiculares. Existen varios modelos utilizados en Machine Learning, en este caso se implementaron solo tres: regresión logística, Árbol de decisión y XGBoost. Se compararon la eficacia de cada uno de ellos mediante métricas específicas y la general la Curva ROC.

Palabras Claves: Fraudes, pólizas, reclamos de asegurados, Machine Learning, Regresión Logística, Árbol de Decisión y XGBoost

Abstract

This research is based on the analysis and evaluation of models that allow the prediction of fraud in the claim of insured vehicle policies. There are several models used in Machine Learning, in this case only three were implemented: Logistic Regression, Decision Tree and XGBoost. The effectiveness of each of them was compared using specific metrics and the general ROC Curve.

Keywords: Fraud, policy, insured claims Machine Learning, Logistic Regression, Decision Tree and XGBoost.

Introducción

Durante varios años atrás, el fraude ha venido creciendo rápidamente y creando infinidad de problemas como la economía en las entidades pertinentes, en esta investigación se estudian estafas que se ocasionan en la reclamación de pólizas vehiculares, que por medio de modelos de clasificación supervisada de Machine Learning se busca la detección de los casos donde se logre la minimización dichos problemas.

Esta investigación se realiza con el fin de revisar modelos que permite a las aseguradoras encontrar los diferentes fraudes que se presentan en el día a día. Por medio del machine learning se ha permitido estudiar diferentes campos, donde permite encontrar patrones que sirven como predicción y toma de decisiones.

Se realiza un análisis estadístico, con el fin de poder interpretar el comportamiento que tiene la solicitud en las reclamaciones de las pólizas de seguros vehiculares, la aplicación y desarrollo de los modelos nos permite evidenciar los fraudes que predice cada uno de estos. La eficacia es evaluada mediante la métrica general de la curva ROC.

Se realiza una revisión bibliográfica sobre el comportamiento del SOAT en Colombia, los actos que son categorizados como fraudes y modelos que han sido utilizados para la revisión de este tema, optando por analizar las bases de datos con Regresión Logística, Árbol de Decisión y XGBoost.

La metodología de esta investigación esta basada en las fases del Proceso estándar de la industria cruzada para la minería de datos (CRISP-DM), logrando entender el comportamiento del conjunto de los datos y por último la evaluación de cada uno de los modelos.

Se concluye cual es el modelo mas eficiente en la predicción de fraudes, mediante modelamiento en Jupyter Notebook, teniendo en cuenta las métricas que arroja cada modelo.

Planteamiento Del Problema

El fraude es un factor que se presenta en el día a día en los usuarios que adquieren su seguro vehicular, afectando directamente el bolsillo de los entes encargados; según Castellanos Heras (2021) el aumento en los casos de fraude detectados, han hecho que las compañías encargadas tomen medidas drásticas sobre el impacto negativo en el negocio de estos procesos ilícitos. Como falsas solicitudes en las reclamaciones implementadas por algunas personas, son unas de las mentiras empleadas para cometer dichos casos arbitrarios.

Según Dalhia de la O (2021) dentro de los entes encargados, los fraudes en seguros de autos se entienden como una serie de mentiras empleadas por una persona para cobrar una póliza de seguros. Los actos son ejecutados por sujetos que obtienen información personal de los clientes y la utilizan de manera incorrecta, adquiriendo los beneficios monetarios sin haber participado en los hechos de accidentes viales; Se puede decir, que estos actos se realizan con el objetivo de obtener ganancias a través de los falsos accidentes.

En los últimos años en Colombia aproximadamente un 37% de los casos donde se pretende cobrar las pólizas de seguros no estaban involucradas en los siniestros viales según FASECOLDA (2022). Miguel Gómez, presidente de Fasecolda, expresó que “el fraude es un delito y las compañías de seguros usan cada vez herramientas más sofisticadas como la inteligencia artificial para combatir y controlar este flagelo” dando así solución a las diferentes irregularidades que observan en dichos casos.

Algunos de los modelos estadísticos que han sido utilizados por diferentes compañías para la identificación de los casos donde se implementan fraudes, no han sido de gran ayuda; Debido a que lo avances no han sido solo tecnológicos, sino que también, las personas que

cometen los fraudes utilizan nuevas herramientas o engaños. Sin embargo, se han utilizado diversos modelos que controlan de una u otra forma algunos de los casos donde se presentan estafas, así mismo, minimizando dicha problemática.

De acuerdo con la anterior información y a la necesidad que se tienen de buscar modelos que permitan, la detección de comportamientos deshonestos que utilizan ciertas personas para cobrar las pólizas en los falsos accidentes viales, y en referencia a Badal Valero et al. (2020) donde indica que *"El reconocimiento de esta problemática implica la adopción de sistemas de prevención y la identificación de patrones de comportamiento de distintos aspectos que conciernen al asegurado en su conducta frente al siniestro para minimizar este tipo de anomalías que se presentan sin medida alguna".* Sin embargo, las empresas no cuentan con modelos efectivos que puedan dar solución definitiva a esta gran problemática que no solo se vive en Colombia sino a nivel mundial.

Actualmente en el ámbito de la analítica de datos hay diferentes modelos que se utilizan para el análisis, especialmente para lograr la identificación de los casos en que pueda presentar fraudes, pero sin mucho éxito en ninguno de ellos. Por esta razón nace la necesidad de analizar algunos modelos de clasificación supervisada para la detección de los fraudes y así observa cuál de ellos presentan un mayor grado de eficacia y que les permita a las entidades un mejor funcionamiento.

Pregunta problema:

¿Cuáles son los modelos de clasificación supervisada más pertinentes para la prevención de fraudes a seguros vehiculares?

Justificación

En el sector de las aseguradoras crece la necesidad de minimizar los fraudes que se presentan día a día por dichas personas que ven los accidentes como la oportunidad de generar ganancias gracias a los diferentes siniestros viales. Como lo indica Viteri Gutiérrez (2020) los defraudadores se han especializado en diferentes técnicas, herramientas y métodos que utilizan para la implementación de dichos fraudes, facilitando con mayor habilidad las estafas.

Según Martínez Mayorga (2017) la importancia que reciben las pólizas se debe a que cada persona que cuenta con un vehículo se le genera un aporte obligatorio que garantiza como mínimo la atención en urgencias en caso de algún accidente, esto a través del Fondo de Solidaridad y Garantía (FOSYGA) quien destina los recursos para la inversión de la salud. Por esta razón comienzan a crecer sin medida alguna las anomalías, como elevados costos médicos, altas comisiones por diferentes entidades de la salud, accidentes que nunca ocurrieron, incluso por pólizas falsas, esto viene ocurriendo desde hace un par de décadas por el simple hecho de no contar con la presencia de las autoridades locales que logren controlar dichas incoherencias.

En la revisión de los diferentes modelos de clasificación supervisada que son utilizados para el estudio de comportamientos sospechosos en las actividades fraudulentas de las pólizas de seguros vehiculares, se estudia la posibilidad de combatir estos inconvenientes y como lo resalta Corso (2009) en el análisis estadístico se utilizan técnicas de minería de datos que abordan la solución a problemas de predicción, clasificación y segmentación.

Según Ortiz y Guzmán (2021) el Machine Learning hace referencia a la detección sistemática de conductas y patrones (Algoritmos) significativos de un grupo de datos. Esto conllevando a una precisión para la identificación de fraudes en los seguros vehiculares, así,

evitando que ingresen datos fraudulentos en los registros de las reclamaciones por accidentes viales.

En esta investigación se estudian los siguientes modelos: Árbol de decisión, Regresión logística y XGBoost, con el fin de comparar la eficacia al momento de modelar cada uno ellos. Se realizará el análisis para la base de datos que se tiene sobre supuestos fraudes presentados con cada uno de estos modelos; queriendo obtener una efectividad en alguno de estos evaluado con una matriz de confusión y curva ROC.

Objetivos de investigación:

Objetivo General:

Evaluar modelos de clasificación supervisada (Regresión Logística, Árbol de decisión & XGBoost) para la identificación de posibles fraudes a las pólizas de seguros de vehículos.

Objetivos Específicos:

1. Realizar un análisis estadístico descriptivo bivariado de la base de datos para la identificación de patrones que muestren el comportamiento de fraude en pólizas de seguros de vehículos.
2. Aplicar los modelos de clasificación supervisada para la identificación de fraudes en seguros de vehículos.
3. Comparar la efectividad de los modelos de clasificación supervisada seleccionados a partir de una matriz de confusión y curva ROC (**curva de característica operativa del recepto).**

Marco Teórico

En este apartado se conocerá el concepto sobre fraude y del sistema que gira en torno al SOAT (Seguro obligatorio de accidentes de tránsito) y las diferentes anomalías que se presentan en los mismos, se estudia diferentes investigaciones referentes a este artículo para revisar los modelos más efectivos al momento de la detección de fraudes en las pólizas de seguros vehiculares.

El fraude es denominado un engaño que realizan terceros para el beneficio propio y afectando drásticamente a las aseguradoras, teniendo que pagar monetariamente por falsas solicitudes en los seguros. Como lo resalta Carmona y Londoño (2021) el fraude no es un tema fácil de resolver, debido a las múltiples modalidades que surgen por los nuevos estafadores y a la evolución que se ve cada día. Gracias a las nuevas herramientas y el progreso de la tecnología se ha ido utilizando el Machine Learning (que de ahora en adelante en el documento se denotara como ML) para la reducción de estas problemáticas que presentan las entidades competentes.

La directora Angela Huzgame de FASECOLDA en el año 2016, indica que el SOAT es un seguro de todos y para todos, de las personas depende hacerlo sostenible para que siga protegiendo en caso de algún accidente. En la ley 33 de 1986, la cual da origen al Código Nacional de Tránsito Terrestre, donde se incorpora el seguro en Colombia y se decreta como requisito obligatorio para todo vehículo automotor que transite por las diferentes vías del territorio nacional. La creación del SOAT nace con el fin de garantizar los recursos médicos que faciliten la atención integral para las víctimas de accidentes viales, beneficiando tanto a pasajeros, peatones y conductores de los vehículos que estén involucrados en dichos accidentes.

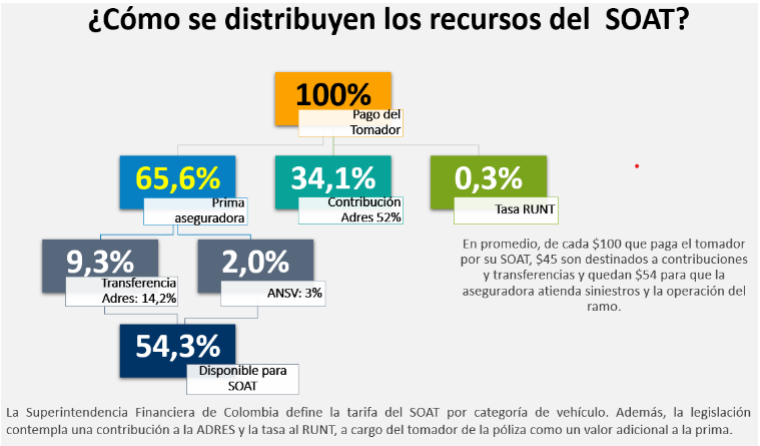
Según la federación de aseguradores colombianos (Fasecolda) informó que el año 2022 aproximadamente el 47% de los vehículos que están matriculados no cuenta con la póliza del

SOAT, siendo 16 millones de vehículos matriculados. Analizando cada región del país donde resaltan ciudades como Amazonas, Vichada y Arauca, de las cuales son las que menos cuenta con vehículos con el seguro obligatorio. También, indican que las motocicletas son los vehículos que menos cumplen con este reglamento.

El gobierno nacional a través de la superintendencia financiera colombiana (SFC) indica cual es el valor monetario que debe pagar cada ciudadano por la compra del SOAT, teniendo en cuenta las características de cada vehículo. Según FASECOLDA (2022) el monto total no es destinado a los seguros de riesgos catastróficos; los porcentajes son destinado para las aseguradoras, Runt, y el ADRES (Administradora de los Recursos del Sistema General de Seguridad Social en Salud). En la siguiente ilustración se puede observar cómo se distribuye el valor que realiza una persona por el pago de su SOAT.

Ilustración 1:

Distribución de recursos del SOAT



Nota. La Superintendencia Financiera de Colombia (SFC) contempla el valor que debe pagar cada vehículo, y la legislación colombiana define la distribución del dinero. Tomado de FASECOLDA (2022)

Dependiendo el accidente así es el monto que la aseguradora le indemniza a la víctima, estas tarifas basadas según el salario mínimo mensual legal vigente en Colombia. Algunos

montos que se logran cobrar son tan elevados que esto genera que diferentes personas presenten solicitudes de reclamos para generar algún dinero extra para sus bolsillos.

Por esta razón el fraude en Colombia y en todo el mundo crece drásticamente, afectando directamente a las aseguradoras; según la investigadora de la cámara técnica del SOAT De la Espriella (2022) indica que son diferentes componentes de frauden que se utilizan hoy en día, siendo el **No Accidente de Tránsito**, la causa más frecuente que se presenta en dichas anomalías; Las modalidades podrían variar dependiendo del monto que se presente cobrar.

Para lograr ver desde otro punto de vista las inconformidades que se presenta en las pólizas de seguros vehiculares, también, poder detectar dichas anomalías que presentan las aseguradoras, se utilizan diferentes mecanismos relacionados con el aprendizaje supervisado ML en inglés y sus diferentes modelos de clasificación supervisada.

La inteligencia artificial (IA) se hace más importante debido a tanta información recolectada durante muchos años, esto le genera a múltiples empresas entender como es el funcionamiento de sus empresas y así tomar medidas y decisiones que le sirven de mejoría. Para Rouhiainen (2018) la inteligencia artificial es la habilidad de los ordenadores para hacer actividades que normalmente requieren de inteligencia humana, siendo un poco más detallado el investigador, indica que la IA es la capacidad de las máquinas para usar algoritmos, aprender de base de datos y así tomar decisiones de lo aprendido, tal cual como lo hace un ser humano.

Dichos modelos logran la solución de problemas como el fraude, no solo en pólizas sino también en tarjetas de créditos, política, entre otras. Según García et al. (1998) los modelos de clasificación supervisada son procesos que se realizan para encontrar propiedades comunes entre un conjunto de datos y clasificar dentro de diferentes clases. La idea principal de esto modelos es

interpretar la descripción de la base de datos y así lograr un modelo que permita encontrar algoritmos basados en la información dada.

Para la identificación de los fraudes en las pólizas de seguros vehiculares se utilizan 3 modelos mencionado anteriormente, los cuales sirven para analizar la eficiencia de cada uno de ellos al momento de la detección de dichas irregularidades.

La regresión logista ha sido un modelo de ML que ha servido para observar el comportamiento en diferentes campos de la investigación, en los cuales se presentan fraudes para las aseguradoras o entidades encargadas de pago por diferentes causas o daños a sus clientes. Como lo indica Moreno Palenzuela (2018) la regresión logística es un método estadístico lo cual sirve para la clasificación binaria de los datos propuestos, esto beneficiando el análisis del comportamiento que tiene la base de datos. Con el estudio de este modelo se representa si el suceso ocurre o no, en este caso identifica si hay o no fraude en cada una de las solicitudes de las pólizas que se estudia.

Según Robles et al. (2020) para logar un buen funcionamiento del algoritmo, es importante realizar un preprocesamiento de la base de datos. Uno de estos procesos que se debe hacer y no solamente para este modelo, es hacer una limpieza de datos, también, eliminar variables por falta de fiabilidad o que no tienen importancia en el análisis; realizar los diferentes pasos mencionados anteriormente permite el rendimiento y la calidad de los sistemas predictivos.

Otro de los modelos que se utiliza para la predicción de los fraudes es **Árbol de Decisión**, también conocido en el ML como modelo supervisado de clasificación mediante particiones binarias recursivas. Este modelo se encuentra entre los más utilizados en la ciencia de datos, no sólo por su interpretabilidad y su performance (rendimiento) sino también por ser la base de los modelos más potentes utilizados en la actualidad, de acuerdo con Arana (2021).

Según Bogoya Contreras (2022) los árboles de decisión están compuestos por nodos los cuales corresponden a las variables de entrada, ramas que representan los posibles valores que puede tener las variables de entrada y una hoja la cual representa los posibles valores de salida. Este sistema en que se procesan los datos según el modelo facilita al científico a interpretar los datos de una forma más fácil y entendible.

Para las aseguradoras, entes encargados de las pólizas de seguros vehiculares han utilizado en los últimos años diferentes métodos para minimizar la corrupción que se presenta en ellas. Algunos métodos es el ML que por medio de modelos que han sido utilizado para la predicción y análisis de comportamientos de datos, permitiendo ver el tipo de fraude y el procedimiento que le da los estafadores en cada uno de ellos; así mismo, conociendo cada una de las estrategias que utilizan los impostores y evitando nuevos fraudes.

En otras palabras y como lo indica los investigadores Bouza y Agustín (2012) un árbol de decisión es una representación gráfica que permite determinar a partir de algún método que modela las tomas de decisiones por medio de reglas de fácil comprensión. En el proceso de la predicción el camino lleva a una evaluación de dichas reglas de las decisiones que se puedan tomar, al momento de hacer un análisis de los nodos (indican una decisión que se está tomando en el árbol y se representan con un cuadrado) se llega a la conclusión si hay representación de fraude.

El modelo **Extreme Gradient Boosting (XGBoost)** es una técnica que se basa en un procesamiento paralelo de los algoritmos y observaciones, por lo que obtiene una mayor rapidez y eficiencia al momento de su aplicación, esto según Castellanos Heras (2021) EL XGBoost fue desarrollado inicialmente por el profesor de ML Tianqi Chen, en la actualidad han ido interviniendo muchos más desarrolladores para la mejoría y eficacia del modelo.

Frutos Serrano (2021) define el XGBoost como una biblioteca que implementa dichos algoritmos que se basan en la implementación de modelos anteriores, teniendo en cuanta lo errores que otros modelos (Regresión Lineal) generan. Algunas que distinguen el modelo XGBoost es una penalización inteligente en la toma de decisión, tanto una reducción en los nudos de las hojas.

Los resultados precisos y la opción de manejar bases de datos con múltiples variables son una de las ventajas que define Espinosa Zúñiga (2020) para el modelo XGBoost, gracias a esto, en los campos donde se utilizan huellas digitales, seguridad financiera, aseguradoras, seguridad vial, etc. Puede beneficiarse y minimizar los daños y perjuicios que le puedan ocasionar los diferentes tipos de fraudes. El investigador también aclara algunas desventajas y como la más principal es que el modelo no funciona con tipo de datos cualitativos, sino que tienen que ser específicamente cuantitativos.

Con el estudio de las diferentes teorías mencionadas anteriormente se puede definir que los tres modelos a trabajar generan un porcentaje de mayor eficacia al momento de la predicción. También se tiene en cuenta las características de cada modelo para el proceso de modelaje, como bien indican algunos investigadores los datos o más bien las variables tiene que ser de tipo cuantitativo, así, generando una mayor efectividad en la predicción.

Antecedentes

Entre las muchas investigaciones que se han llevado a cabo para la detención de fraudes en diferentes campos de la economía mundial, especialmente en el contorno de las aseguradoras de pólizas de seguros vehiculares, que estudian por medio de aplicación de modelos de supervisión clasificada y pueden contrarrestar la problemática que se presentada en el día a día afectando directamente el bolsillo de estas entidades.

Belhadji et al. (2000) en su investigación realizada sobre un modelo para la detección de fraudes de seguros, indica que lograría evaluar el alcance el fraude de seguros vehiculares por medio sistemas expertos que le permitiera la revisión de dichos fraudes. Determinar indicadores que le convenga minimizar dichas anomalías para las entidades aseguradoras, revisando por varios factores que pueden intervenir en la investigación.

El manejo que lleven las aseguradoras sobre el seguro vehicular puede cambiar dependiendo la entidad. Según Ayuso et al. (1999) El funcionamiento del seguro dicta unas pautas que debe tener en cuenta el dueño del vehículo al momento de cobrar la indemnización o dicho en otras palabras cobrando el seguro en caso de algún accidente. En ocasiones los siniestros viales pueden afectar a terceros, esto es lo que puede ocasionar o presentar un porcentaje mayor en dichas anomalías; permitiendo cobrar tarifas elevadas que jamás se presentaron.

En el trascurso del tiempo la tecnología ha evolucionado de una manera valiosa, dando mejores oportunidades en los diferentes campos o disciplinas que ayudan en el bienestar del ser humano. Con esto, también ayudando en la creación de diferentes modelos probabilísticos que han servido para detectar situaciones de fraude como los que se presentan en las pólizas del SOAT. Citando de nuevo a Badal Valero et al. (2020) donde menciona que, en materia de las

técnicas para la detección de fraudes y en la implementación de algoritmos para la modelación y predicción ha utilizado modelos como regresión lineal, redes neuronales y modelos de elección discreta.

En este tipo de incoherencias se analiza si los diferentes casos, donde se puede presentar de que, si son anomalías o no, para ello los modelo constituye a la creación y revisión de los algoritmos que permitan revisar cada uno de los casos; según (Ameijeiras Sánchez et al. 2021) el reto que tiene cada desarrollador de modelos es la de construir un modelo predictivo, en la cual sea cada vez más flexible y adaptables a condiciones cambiantes con rapidez, permitiendo detectar eventos únicos o emergentes.

Los procesos que se llevan para la detección de los fraudes se basan en: determinar un conjunto de datos que sirve para la determinación, validación de la información tomada, y por último probar las variables significativas, dando así resultados confiables o llegado al caso de que no lo sean. Según Santamaria Ruiz (2006) la detección basada en reglas está basada en el análisis individual de cada caso, con la utilización de ML que es capaz de extraer patrones que permita comparar las bases de datos gigantes, esto permitiendo a los analistas el comprender y modelar de una manera más eficiente, ayudando a la toma de decisiones.

La necesidad de las aseguradoras de evitar algún tipo de fraude en las pólizas de seguros vehiculares conlleva a que utilicen diferentes estrategias en la eliminación de todas las rarezas que se presenten. Una de esas estrategias es el ML, que por medio de modelos de predicción y algoritmos permiten analizar el comportamiento en la minería de datos y así observar cómo se pueden presentar fraudes en las diferentes solicitudes en las aseguradoras.

En esta investigación se utiliza tres modelos, que por medio de su implementación individual se lograra evidenciar el comportamiento en las anomalías. También, se evaluará el

modelo con mejor rendimiento en la predicción y el descubrimiento de mayores patrones que utilizan los estafadores.

Según Castellanos Heras (2021) el análisis del fraude cada vez se extiende más en todos los campos de negocio, siendo las pólizas de seguros de automóviles las que poseen mayor número de herramienta para poder controlar los fraudes. El modelo que utiliza el investigador Castellanos es el árbol de decisión, lo implementa debido a que es una de las metodologías más utilizas en la inteligencia artificial y por su fácil interpretación.

Citando a los investigadores Ayuso et al. (1999) donde explica que la aplicación de los modelos de elección múltiple permite determinar en primer lugar cuáles son los indicadores para estudiar y así confirmar la sospecha de fraude en un expediente. Estos investigadores proporcionaron un modelo a las compañías lo cual le permitió identificar un tipo de comportamiento fraudulento, con esto facilitando el proceso de detección, el Árbol de decisión es un modelo que le permite realizar el respectivo análisis.

Patiño Espinoza (2014) opta por utilizar la Regresión Logística como modelos de predicción, debido a que son caracterizados por su sencilla aplicación e intuitiva interpretación del resultado que arroje el modelo. El investigador Patiño en su investigación implementa otros dos modelos que lo conllevan a lo mismo, pero no con la misma eficacia que lo hace el Árbol de Decisión con un porcentaje del 78% de favorabilidad.

La regresión logística consiste en la medición de la calidad del modelo en una sucesión para distintos puntos de corte entre 0 y 1, según Moreno Palenzuela (2018), basándose en esta breve definición, el investigador utiliza este modelo de regresión para su investigación. Aunque, explica muy breve que el gran problema que se puede tener al momento de aplicar el modelo es el elevado uso de recursos en relación con la cantidad de casos que se toma como entrenamiento.

Para los siguientes investigadores Carmona y Londoño (2021) en su proceso de investigación, indican que el mejor modelo con mayor porcentaje de eficacia es la regresión logística, aunque para ellos es importante tener en cuenta que este modelo presenta un gran número de “falsos positivos” y para la entidad a la cual hace el proceso de investigación es más favorable detectar los fraudes y no se vea afectado los ingresos de la compañía.

En la comparación y eficacia de modelos clasificados que resalta Frutos Serrano (2021) en su investigación, resalta que el modelo regresión lineal es más eficaz que el XGBoost, teniendo en cuenta que para este estudio no se implementara la Regresión Lineal; Resalta que el XGBoost no es tan eficaz debido a que el otro modelo realiza predicciones más precisas.

Según los antecedentes revisados, se puede evidenciar cuales modelos son los más utilizados e incluso los más afectivos al momento de la predicción. Llegando a la implementación de los modelos (Regresión logística, Árbol de Decisión y el XGBoost) para esta investigación. En la búsqueda del modelo más eficiente, se planea trabajar la misma base de datos para así Comparar la efectividad de los modelos de clasificación supervisada seleccionados a partir de una matriz de confusión y curva ROC, como lo indica uno de los objetivos específicos.

Metodología

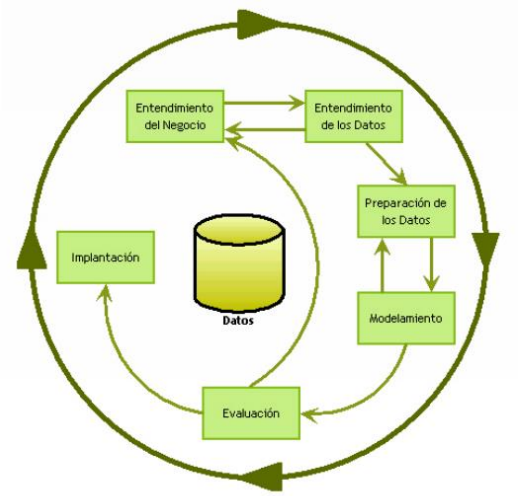
Para el desarrollo de esta investigación y queriendo dar respuesta al interrogante planteado, también dando cumplimiento con los objetivos, y basándonos en los paradigmas de investigación, se implementará una metodología de carácter cuantitativa; esto porque la investigación está basada en el análisis de variables, el cual permitirá llegar a la solución del problema. Haciendo un análisis experimental y manipulativo, teniendo en cuenta las creencias básicas según Guba y Lincoln (2002) sobre el proceso metodológico aplicado para esta investigación, lo cual está basado en el estudio y análisis de la realidad que se puede ver por medio de mediciones, teniendo un control e inferencia de otras investigaciones que aportan a este mismo estudio.

Con el propósito de desarrollar esta investigación se selecciona la base de datos acorde con la investigación, luego de esto se realiza una limpieza y procesamiento para que no vaya a ocurrir distorsiones en los posibles resultados, concluyendo con una modelación de los modelos de clasificación supervisada, mediante las herramientas e implementación del ML.

Para el desarrollo metodológico se implementará las fases que utiliza (Galán Cortina, 2015) en su proyecto investigativo basándose en el Proceso estándar de la industria cruzada para la minería de datos (CRISP-DM, siglas en ingles).

Ilustración 2:

Comportamiento de la metodología CRIP-DM



Nota. Secuencias de las fases en la metodología CRIP-DM. Tomado de (Galán Cortina, 2015)

Basado en lo mencionado anteriormente, las fases de la investigación se desarrollarán de la siguiente forma:

Fase 1:

Recolección y entendimiento de la base de datos, la plataforma de Kaggle cuenta con múltiples bases de datos, en este sitio web se encuentra las pólizas de los seguros vehiculares en la que presuntamente se presentan los fraudes. Filtrando en la búsqueda de la página con las palabras en inglés “Insurance Claims” que en el idioma español es Reclamación de Seguros y utilizando la base de datos del autor Shah (2018) para la respectiva investigación y análisis de este. Se escoge esta base datos de los miles que genera la página, por la razón de que se ve más completa, en el sentido de que no tiene ningún análisis y no le han hecho ningún recorte en las variables.

Fase 2:

Análisis de los datos, en esta fase nos permite ver el comportamiento de cada una de las variables de las bases de datos, así determinar cuáles son las que afectan con mayor porcentaje en dichos fraudes. Esta base de datos al descargarla cuenta con un total de 40 columnas por 1000 filas, de las cuales, hay variables que no genera mucha información relevante y se llegara el caso de no hacer ningún análisis con ellas.

Fase 3:

Recolección de bibliográfica, hacer una averiguación sobre lo que se ha hecho acerca de lo que se trabaja, revisando cuales modelos han servido para la solución de la problemática. En esta fase también se estudia cuáles de los modelos utilizados han sido más efectivos en otras investigaciones, así manejando los más viables para esta investigación.

Fase 4

Preparación de los datos, en esta fase implicaría una limpieza en la base de datos y un respectivo procesamiento a través de técnicas (Agrupamientos, selección de variables, transformación de variables, datos faltantes y hasta una reducción de datos), logrando esto se llevará a cabo la siguiente fase que será la de modelación.

Fase 5:

Modelación, en esta fase los datos se utilizan para construir y evaluar la eficiencia de los modelos de clasificación supervisada aplicados en el ML. Luego de realizado la modelación de cada uno de los modelos mencionados anteriormente se compara la efectividad a la hora de detectar los fraudes.

Fase 6:

Evaluación, Realizando las anteriores fases es necesario hacer una evaluación sobre los resultados que se obtiene sobre cada uno de los modelos aplicados en la investigación. Es necesario saber si en la base de datos analizada se presentan los dichos fraudes, es posible que nuevos hallazgos puedan servir para generar nuevas ideas a otros proyectos a explorar.

Aplicación

En el presente capitulo se describe la aplicación del proyecto teniendo en cuenta las fases anteriormente mencionadas.

Descripción de los datos:

La base de datos que se utilizó en la investigación es del autor Shah (2018) que se puede descargar de la plataforma Kaggle, filtrando la búsqueda en la página web con las palabras en inglés “Insurance Claims” (reclamos de aseguramiento). A continuación, se muestra en la tabla las variables e información de toda la base de datos.

Ilustración 3:

Tipo de variables

```
Data columns (total 40 columns):
 #   Column                       Non-Null Count  Dtype  
---  ------                       --------------  -----  
 0   months_as_customer           1000 non-null   int64  
 1   age                          1000 non-null   int64  
 2   policy_number                1000 non-null   int64  
 3   policy_bind_date             1000 non-null   object 
 4   policy_state                 1000 non-null   object 
 5   policy_csl                   1000 non-null   object 
 6   policy_deductable            1000 non-null   int64  
 7   policy_annual_premium        1000 non-null   float64
 8   umbrella_limit               1000 non-null   int64  
 9   insured_zip                  1000 non-null   int64  
 10  insured_sex                  1000 non-null   object 
 11  insured_education_level      1000 non-null   object 
 12  insured_occupation           1000 non-null   object 
 13  insured_hobbies              1000 non-null   object 
 14  insured_relationship         1000 non-null   object 
 15  capital-gains                1000 non-null   int64  
 16  capital-loss                 1000 non-null   int64  
 17  incident_date                1000 non-null   object 
 18  incident_type                1000 non-null   object 
 19  collision_type               1000 non-null   object 
 20  incident_severity            1000 non-null   object 
 21  authorities_contacted        1000 non-null   object 
 22  incident_state               1000 non-null   object 
 23  incident_city                1000 non-null   object 
 24  incident_location            1000 non-null   object 
 25  incident_hour_of_the_day     1000 non-null   int64  
 26  number_of_vehicles_involved  1000 non-null   int64  
 27  property_damage              1000 non-null   object 
 28  bodily_injuries              1000 non-null   int64  
 29  witnesses                    1000 non-null   int64  
 30  police_report_available      1000 non-null   object 
 31  total_claim_amount           1000 non-null   int64  
 32  injury_claim                 1000 non-null   int64  
 33  property_claim               1000 non-null   int64  
 34  vehicle_claim                1000 non-null   int64  
 35  auto_make                    1000 non-null   object 
 36  auto_model                   1000 non-null   object 
 37  auto_year                    1000 non-null   int64  
 38  fraud_reported               1000 non-null   object 
 39  _c39                         0 non-null      float64
dtypes: float64(2), int64(17), object(21)
```

Nota. Elaboración propia desde Jupyter Notebook

El tipo de cada variable es caracterizado por palabras como: Float64, Int64 y Object; para realizar el análisis con los modelos propuestos, es importa conocer toda la información de la Data, como se mencionó anteriormente los modelos trabajan con más eficacia si son de tipo cuantitativas. A continuación, una explicación a lo que hace referencia cada tipo de variable:

- Object: Se utiliza para representar cualquier tipo de objeto en Python, como cadenas de texto, listas, diccionarios, entre otros. Es el tipo de datos más genérico en Python y puede almacenar cualquier tipo de valor.
- Int64: Se utiliza para representar números enteros de 64 bits en Python. Este tipo de datos es útil cuando se necesitan valores enteros muy grandes o cuando se realizan operaciones matemáticas que requieren una alta precisión.
- Float64: Se utiliza para representar números de punto flotante de 64 bits en Python. Este tipo de datos es útil para representar valores decimales y se utiliza con frecuencia en cálculos científicos y matemáticos.

Esta base de datos cuenta originariamente con 40 columnas por 1000 filas, con variables de tipo cualitativas y cuantitativas. Como primer paso para el estudio de esta, se analiza la forma en que se encuentra conformada cada una de sus variables, esto con el fin de observar la información que nos genera cada una de ellas. La base de datos es tomada de EE.UU, y el estudio es enfocado en variables similares al del comportamiento en las pólizas de seguros vehiculares colombianas. En la tabla se muestra la información de cada variable, su traducción y lo que expresa cada una de ellas.

Tabla 1:

Base de datos con la traducción y definición de cada variable, se indica si es útil en el estudio o mejor eliminarla, con una breve justificación.

Variable	Traducción de Variable	Definición	¿Útil?		Justificación
			SI	NO	
months_as_customer	Meses como Cliente	Esta variable representa los meses que ha estado utilizando la póliza.	x		Todas las pólizas en general tienen una fecha en la cual muestran el día exacto de la vinculación
Age	Edad	Representa la edad del cliente	x		Es importante ya que refleja la edad de la persona víctima del accidente
policy_number	Numero de póliza	El número de póliza del seguro adquirido		x	El numero de la póliza no genera ningún dato relevante en el estudio
policy_bind_date	Fecha de la póliza	Indica la fecha en la cual se adquirió la póliza		x	la fecha de la póliza no genera ningún dato relevante en el estudio
policy_state	Estado de la póliza	Quiere indicar si la póliza está vigente o lo contrario vencida		x	El estado de la póliza no genera ningún dato relevante en el estudio
policy_csl	Limite único combinado	Representa el monto que puede pagar la póliza a terceros		x	Esta variable muestra el monto que la aseguradora pagó a los terceros, en el estudio no es relevante
policy_deductable	Deducible de póliza	EL valor que paga por la póliza	x		Este variable es importante, muestra el valor que paga las aseguradoras, se puede estudiar si el valor es muy grande en comparación a las demás solicitudes.
policy_annual_premium	prima anual de la póliza	Es el monto que debe pagarse para la protección de la aseguradora	x		Cada usuario paga un valor dependiendo el tipo de vehículo y tipo de seguro que desea adquirir.

umbrella_limit	limite seguro paraguas	este tipo de seguro es una cobertura adicional por encima de los limites básicos de tu seguro normal		x	Para el estudio no genera importancia, debido a que es proceso interno de EE.UU
insured_zip	PENDIENTE	PENDIENTE		x	Se determina a la eliminación de la variable
insured_sex	Sexo Asegurado	Define el género de la persona que solicito el seguro	x		La variable se puede filtrar dependiendo el género y se puede llegar a alguna conclusión
insured_education_level	Nivel educativo del asegurado	Indica cual es el máximo nivel del asegurado	x		La variable se puede filtrar dependiendo el nivel de educación del usuario y se puede llegar a alguna conclusión.
insured_occupation	ocupación del asegurado	Indica cual es la ocupación del asegurado	x		La variable se puede filtrar dependiendo la ocupación del usuario y se puede llegar a alguna conclusión
insured_hobbies	Hobbies del asegurado	El pasatiempo preferido del asegurado	x		La variable se puede filtrar dependiendo el hobby del usuario y se puede llegar a alguna conclusión
insured_relation_ship	Relación del asegurado	PENDIENTE	x		La variable se puede filtrar dependiendo la relación del usuario y se puede llegar a alguna conclusión
capital-gains	Ganancias del capital	Indica que valor de ganancia que recibe el cliente en caso de no haber ningún accidente	x		Se puede evidenciar que estos valores no pueden generar sospechas de fraudes
capital-loss	Perdidas del capital	Indica que valor le da en perdidas el cliente en caso de accidente	x		Se puede evidenciar que estos valores no pueden generar sospechas de fraudes

incident_date	Fecha del accidente	Indica el día en que ocasiono el siniestro		x	Por falta de información, se determina la eliminación de la variable.
incident_type	Tipo de Incidente	Indica como fue el accidente, si hubo más involucrados		x	Por falta de información, se determina la eliminación de la variable.
collision_type	Tipo de colisión. (Lateral, de frente)	Explica si la colisión fue de frente, lateral, etc.		x	Por falta de información, se determina la eliminación de la variable.
incident_severity	Gravedad del accidente	Indica que tan grave fue el accidente, (pérdida total) del vehículo	x		La variable muestra el suceso y gravedad de este, se estudia el caso.
authorities_contacted	Autoridades contactadas	Explica que autoridad se llama al momento del accidente	x		Cuando se llama a las autoridades se puede evidencia que si hubo accidente
incident_state	Estado donde ocurrió accidente	Indica en qué Estado sucedió el accidente		x	Información de EE.UU
incident_city	Ciudad donde ocurrió accidente	Indica en que ciudad sucedió el accidente		x	Información de EE.UU
incident_location	Dirección donde ocurrió accidente	Indica en que ubicación sucedió el accidente		x	Información de EE.UU
incident_hour_of_the_day	Hora en que ocurrió accidente	Indica la hora en que sucedió el accidente	x		Las pólizas deben contener la hora en que ocurrió el accidente, así se puede estudiar si la solicitud no corresponde a lo dicho
number_of_vehicles_involved	Número de vehículos involucrado en el accidente	Indica si hay más vehículos involucrados	x		Daños a terceros, es uno de los principales fraudes
property_damage	Daño a la propiedad	Indica si el accidente ocasiono daño a otra propiedad		x	Por falta de información, se determina la eliminación de la variable.

bodily_injuries	Lesiones corporales	indica que la póliza cubre las lesiones de terceros	x		Daños a terceros, es uno de los principales fraudes
witnesses	Testigos	Indica si hubo algún testigo en el accidente		x	Se pueden presentar falsos testimonios
police_report_available	Reporte de policía	Muestra el reporte del accidente		x	Se pueden presentar falsos testimonios si no hay reporte de policía
total_claim_amount	Importe total de reclamación	El valor que la aseguradora da al propietario	x		Las ganancias que puede obtener con la reclamación, es el objetivo principal de los fraudes
injury_claim	Demanda por lesiones	El valor que se paga luego de haber hecho una demanda para el pago de las lesiones	x		Las ganancias que puede obtener con la reclamación, es el objetivo principal de los fraudes
property_claim	Reclamo de propiedades	El valor que se le pagan a las propiedades en cada de hacer algún daño	x		Daños a terceros, es un de los principales fraudes
vehicle_claim	Reclamo vehicular	El valor que paga a los demás vehículos	x		Las ganancias que puede obtener con la reclamación, es el objetivo principal de los fraudes
auto_make	Marca del carro	Representa la marca del carro		x	
auto_model	Modelo del carro	Representa el modelo del carro		x	La marca o el modelo del carro no genera relevancia en el estudio
auto_year	Año del carro	Representa el año del carro		x	
fraud_reported	Reporte de Fraude	Indica si hay o no Fraude	x		Variable Respuesta
_c39	Sin Datos	Sin Datos		x	NO contiene información

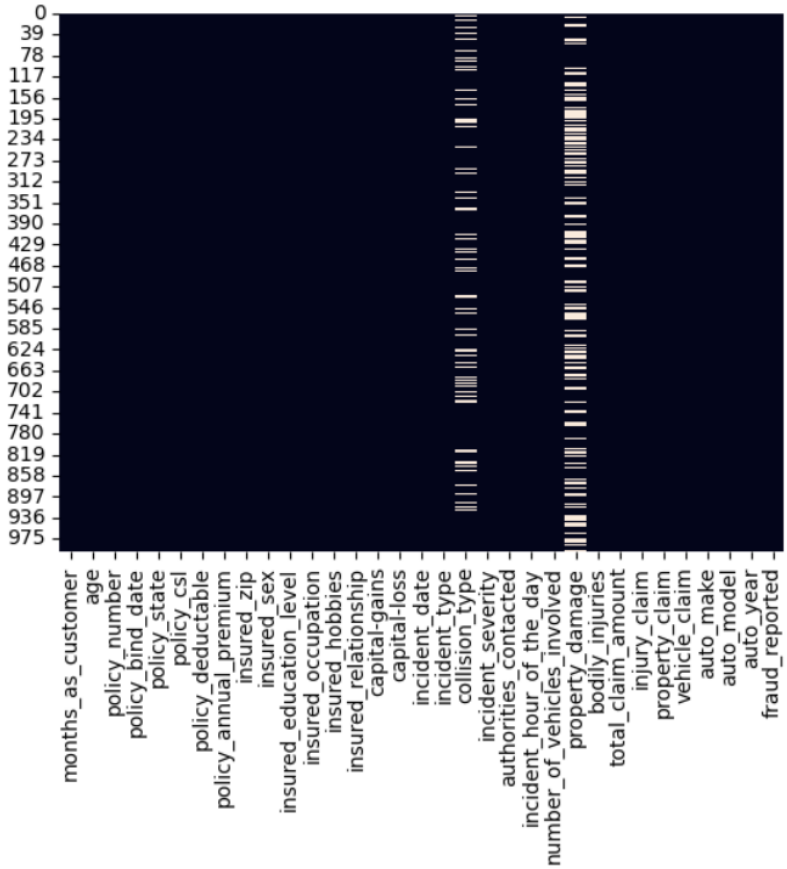
Nota. La tabla representa la traducción de cada una de las variables, en la justificación se argumenta la razón por la cual se elimina o estudia la variable. Elaboración propia.

Para el análisis de los datos se deben realizar diferentes procedimientos que permiten tener una data limpia de errores y así poder evaluar cada los modelos a estudiar.

Uno de estos procesos, es el observar los datos faltantes en las variables, que en primera instancia no arrojan ninguno, ya que están camuflados en la base de datos con '¿?', reemplazar estos signos de interrogación con casillas vacías se podrá evidenciar los datos que hacen falta en cada una de las variables. Como lo evidencia en la ilustración 4.

Ilustración 4:

Representación de datos faltantes.



Nota. La grafica muestra datos faltantes en cada una de las variables a estudiar. Elaboración propia desde Jupyter Notebook

Realizado el cambio, se puede observar datos faltantes en dos variables, la primera ‘Collision_Type’ y ‘Property_Damage’ con porcentajes de 18 (180 datos) y 37 (360 datos) respectivamente; como superan un 15% de valores ausentes, se determina la eliminación de las variables. En caso de que no supere el porcentaje mencionado, se puede realizar imputación de variable, dependiendo la técnica más adecuada para la variable.

Como segundo paso, se realiza una eliminación de las variables que no generan importancia en el estudio, esto se hace antes de poder realizar el modelamiento con cada modelo trabajado; dejando como nueva de base la siguiente:

Ilustración 5:

Nueva base de datos

```
Data columns (total 24 columns):
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   Unnamed: 0                   1000 non-null   int64
 1   months_as_customer           1000 non-null   int64
 2   age                          1000 non-null   int64
 3   policy_deductable            1000 non-null   int64
 4   policy_annual_premium        1000 non-null   float64
 5   insured_sex                  1000 non-null   object
 6   insured_education_level      1000 non-null   object
 7   insured_occupation           1000 non-null   object
 8   insured_hobbies              1000 non-null   object
 9   insured_relationship         1000 non-null   object
 10  capital-gains                1000 non-null   int64
 11  capital-loss                 1000 non-null   int64
 12  incident_type                1000 non-null   object
 13  collision_type               1000 non-null   object
 14  incident_severity            1000 non-null   object
 15  authorities_contacted        1000 non-null   object
 16  incident_hour_of_the_day     1000 non-null   int64
 17  number_of_vehicles_involved  1000 non-null   int64
 18  bodily_injuries              1000 non-null   int64
 19  total_claim_amount           1000 non-null   int64
 20  injury_claim                 1000 non-null   int64
 21  property_claim               1000 non-null   int64
 22  vehicle_claim                1000 non-null   int64
 23  fraud_reported               1000 non-null   object
```

Nota. Nueva base de datos luego del procesamiento realizado a la misma. La fila 0 ‘Unnamed: 0’ es una variable que no genera información y al modificar la data nos queda reflejada. Con 23 variables se procede a realizar el modelamiento de cada modelo. Elaboración propia desde Jupyter Notebook

Análisis exploratorio

Una vez analizados los datos, se procede a realizar un análisis, generando y dando cumplimiento a uno de los objetivos de la investigación que es realizar un análisis bivariado de toda la base de datos, esto implica crear tablas de frecuencia o gráficos estadísticos; esto se realiza teniendo como fija la variable 'Fraud_Reported' (Reporte de Fraude), debido a que nos muestra si la solicito es relevada como fraude. A continuación, una de las gráficas que se lograron realizar, fue con la variable 'authorities_contacted' (Autoridades contactadas), 'Insured-Sex' (Sexo del asegurado), 'incident_hour_of_the_day' (Hora del accidente), y 'insured_hobbies' (Pasatiempo del asegurado).

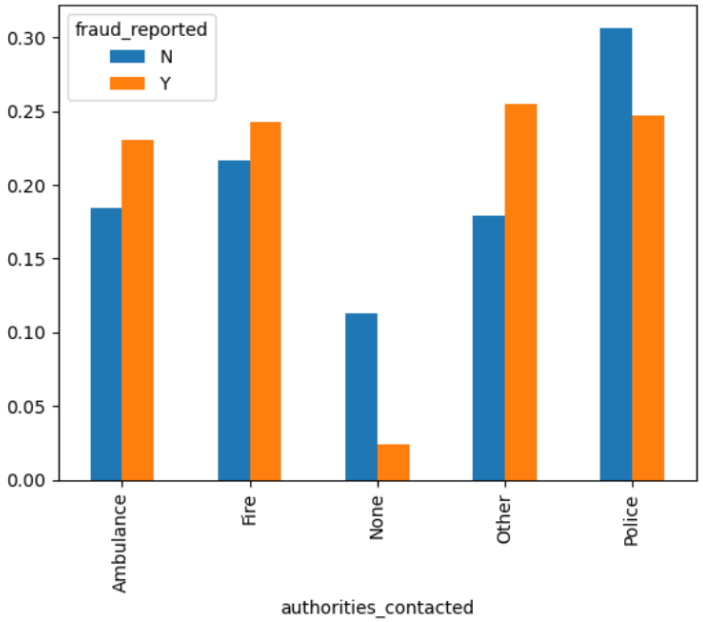
A continuación, en las gráficas que se presenta, las barras de color naranja serán las que reflejan el 'Fraude' y el color azul representa el 'No Fraude'. También se debe de tener en cuenta que el análisis bivariado se realizó, teniendo como fija la variable respuesta, en este caso 'Fraud_Reported'.

De la misma forma en la parte izquierda de las gráficas de barras, se encontrarán una tabla de proporcionalidad que van desde 0 a 1, que corresponde a los valores que están interpretado en el diagrama estadístico.

Ilustración 6:

Análisis bivariado, variables 'Authorities _Contacted' y 'Fraud_Reported'

fraud_reported	N	Y
authorities_contacted		
Ambulance	0.184595	0.230769
Fire	0.216467	0.242915
None	0.112882	0.024291
Other	0.179283	0.255061
Police	0.306773	0.246964



Nota. La grafica de barras es representa de acuerdo con la tabla de proporcionalidad y refleja el análisis bivariado entre la variable repuesta 'Fraud_Reported' y 'Authorities _Contacted'. Elaboración propia desde Jupyter Notebook

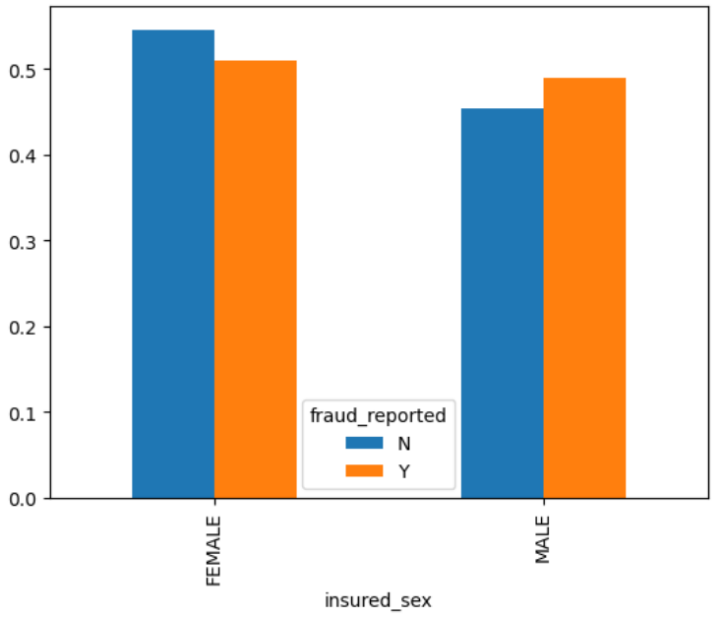
La variable "Autoridades contactadas" refleja 5 datos diferentes (Ambulancia, Bomberos, Ninguno, Policía y Otros), se puede observar que al contactar la policía es menos probable que las personas o clientes puedan cometer algún tipo de fraude.

En un momento que se comienza a realizar el análisis y al observar 'None' (Ninguno) como dato de la variable, se estimaba que el porcentaje de fraude fuera a ser mayor a los demás, en muchos de los casos debe de estar una autoridad presente para poder justificar el accidente ocasionado y al no presentarse ninguno se puede aprovechar para poder ejecutar los fraudes.

Ilustración 7:

Análisis bivariado, variables 'Insured_Sex' y 'Fraud_Reported'

fraud_reported	N	Y
insured_sex		
FEMALE	0.545817	0.510121
MALE	0.454183	0.489879



Nota. La grafica de barras es representa de acuerdo con la tabla de proporcionalidad y refleja el análisis bivariado entre la variable repuesta 'Fraud_Reported' y 'Insured_Sex'. Elaboración propia desde Jupyter Notebook

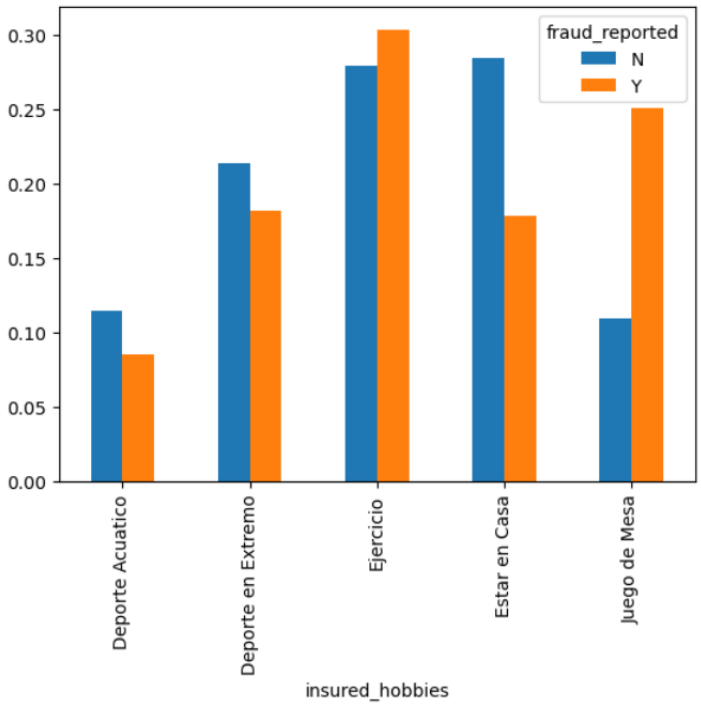
En la variable "Sexo del asegurado" se puede observar que la mujer muestra una proporción un poco mayor e en comparación a los hombres, lo que quiere decir es que los reportes que se solicitaron, los realizaron más las mujeres.

En la tabla de proporcionalidad, los números se mantienen en el mismo promedio y es muy relevante en el diagrama de barras. Las mujeres aun con mínimo número de diferencia se define como el tipo de cliente con mayor probabilidad de que realice algún tipo de fraude.

Ilustración 8:

Análisis bivariado, variables 'Insured_Hobbies' y 'Fraud_Reported'

fraud_reported	N	Y
insured_hobbies		
Deporte Acuatico	0.114210	0.085020
Deporte en Extremo	0.213811	0.182186
Ejercicio	0.278884	0.303644
Estar en Casa	0.284197	0.178138
Juego de Mesa	0.108898	0.251012



Nota. La grafica de barras es representa de acuerdo con la tabla de proporcionalidad y refleja el análisis bivariado entre la variable repuesta 'Fraud_Reported' y 'Insured_Hobbies'. Elaboración propia desde Jupyter Notebook

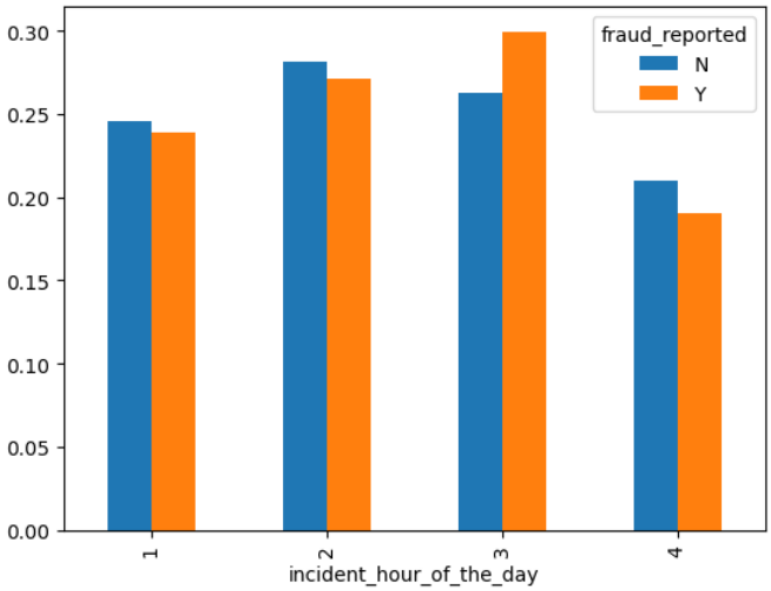
En la variable 'Insured_Hobbies', el Ejercicio se destaca por estar por encima de los demás datos; las persona que reportaron el pasatiempo como ejercicio, se encuentran con un mayor porcentaje de fraude.

Tener en cuenta que para esta variable en un principio se minimizaron valores, esto con el fin se arroja un mejor reporte de este. Por ejemplo, había datos como: ajedrez, domino, video juegos; que se conectaron a un solo dato 'Juegos de Mesa'. Así mismo con los otros 4 componentes que completan la variable 'Insured_Hobbies'.

Ilustración 9:

Análisis bivariado, variables 'Incident_hour_of_the_day' y 'Fraud_Reported'

fraud_reported	N	Y
incident_hour_of_the_day		
1	0.245684	0.238866
2	0.281541	0.271255
3	0.262948	0.299595
4	0.209827	0.190283



Nota. La grafica de barras es representa de acuerdo con la tabla de proporcionalidad y refleja el análisis bivariado entre la variable repuesta 'Fraud_Reported' y 'Incident_Hour_Of_The_Day'. Elaboración propia desde Jupyter Notebook

Para la variable 'incident_hour_of_the_day' se realizó una clasificación en 4 etapas: primera '1' definido en el intervalo de tiempo [00:00,05:59], segunda '2' en el intervalo [06:00,11:59], tercera '3' en el intervalo [12:00,17:59] y por última fase, la cuarta '4' en el intervalo [18:00,23:59]. En el que el 3 'Tarde' representa un gran número de fraudes, cometidos en ese trascurso del día.

Luego de una sofisticada revisión de la base datos, se puede decir que están completos y cumplen con las condiciones de las pólizas colombianas. Para la realización de esta fase, primero se realiza una limpieza de las variables que no genera información relevante en el estudio; Variables como '_C39' que no contiene ningún dato, se considera variable no pertinente para el estudio.

Luego de una previa observación de los datos, se realiza una limpieza de las variables que no generan importancia en el estudio, esto se hace antes de poder realizar el modelamiento con cada modelo trabajado; en la *Ilustración 5* se puede observar las variables que son utilizadas en el procesamiento de los modelos.

La recategorización de las variables cualitativas permite que la programación analice solo variables de tipo cuantitativas, esto se logra por medio de las Variables Dummy, que es utilizada comúnmente en los análisis estadísticos y en la modelización de datos para representar características categóricas de una información y que permitiendo ser incluidas en los modelos estadísticos a trabajar. En el caso de la base de datos trabajada, características como Sexo del asegurado, Hobbies, Nivel educativo, entre otras. Se aplica Dummy para reclasificar las variables a tipo cuantitativo.

En algunas variables de tipo cuantitativo lo que se realizo fue la categorización, esto permitiendo un mejor análisis e interpretación de la misma. Para las fechas, se aplica la categorización con el fin de minimizar los datos de la variable y realizar de nuevo un análisis que pueda ser interpretado.

Los pasos realizados para lograr organizar la base de datos, genera un nuevo conjunto de datos con mejor visualización e interpretación. Dando así, continuidad a la siguiente fase de la investigación.

Modelos

Luego de tener una base de datos ordenada, se procede a realizar el análisis con cada uno de los modelos a trabajar. Esto con el fin de observar el rendimiento en cada uno de ellos y la eficiencia de los mismos.

Para la aplicación de los modelos, se divide la base de datos en dos partes, primera parte un 70% (700 datos) que corresponden a un nuevo conjunto del cual el modelo hará el estudio y aprenderá de los comportamientos que reflejan cada variable. Segunda parte, un 30% (300 datos) el cual corresponde al otro conjunto de datos, en donde, el modelo de lo aprendido hará la predicción.

Cada modelo arroja una predicción diferente y en esto se basa el objetivo de la investigación. Lo cual pretende analizar el comportamiento de cada uno y observar el más eficiente a la hora de encontrar las anomalías en las solicitudes de reclamaciones de las pólizas.

En cada uno de los modelos de muestra una tabla de clasificación, lo cual es útil para evaluar el rendimiento, analizar los errores de predicción y comprender cómo el modelo se desempeña en cada clase en particular. A continuación, se muestra el desarrollo de los 3 modelos de clasificación tipo Machine Learning:

Regresión logística

El informe de clasificación para el modelo de Regresión Logística es el siguiente:

Ilustración 10*:*

Clasificación modelo Regresión logística

	precision	recall	f1-score	support
0	0.85	0.92	0.89	222
1	0.70	0.55	0.62	78
accuracy			0.82	300
macro avg	0.78	0.74	0.75	300
weighted avg	0.81	0.82	0.82	300

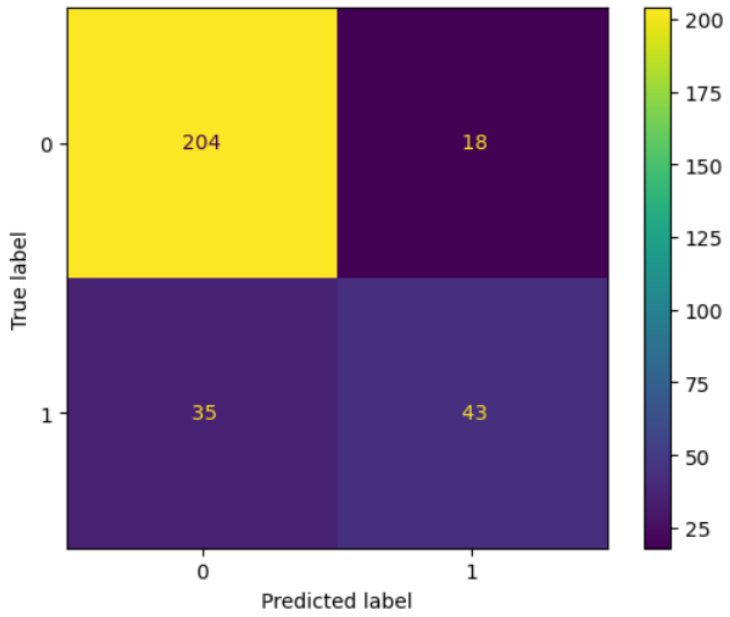
Nota. Esta tabla muestra la correspondencia entre las predicciones del modelo y las clases reales a las que pertenecen los ejemplos de prueba, con una exactitud de predicción de 82%. Elaboración propia desde Jupyter Notebook

El modelo arroja una exactitud de predicción de 82%, sin embargo, este resultado se debe de analizar con precaución ya que la basa de datos presentaba un desbalanceo importante. Por tal motivo se estudiarán métricas más específicas para la clase (fraude y no fraude).

Por otro lado, la sensibilidad del modelo (Recall) para la clase mayoritaria (no fraude) es del 92% y un 55% para ‘Fraude’.

Ilustración 11:

Matriz de confusión para el modelo Regresión Logística



Nota. La matriz de confusión está compuesta por dos partes: Casos reales (True Label) y Predicción del modelo (Predicted Label). La clase 'No Fraude' identificada con 0 y el 'Fraude' con 1. Elaboración propia desde Jupyter Notebook

De la anterior matriz de confusión se puede concluir que 204 de los datos son clasificados correctamente como 'No Fraude' por el modelo, se detecta 18 registros el modelo los clasifica como fraude cuando no lo son. Por otro lado, el modelo clasifica correctamente 43 casos de fraude de un total de 78 lo que indica que se equivoca 35 veces en decir que no es fraude cuando en realidad si lo es. En la ilustración 18 se muestra la curva ROC para el desempeño de este modelo de manera comparativa para los tres modelos aplicados.

Como conclusión a este modelo podemos identificar que la regresión logística es un modelo regular para clasificar casos de fraude ya que esta detectando un 55% un poco mas que lanzar una moneda al aire con una precisión del 70% para esta clase de fraude. En el caso

particular de los registros no fraudulentes, como ya se mencionaba anteriormente el modelo resulta muy útil por que detecta un 92% de casos con una precisión del 85%, sin embargo, para el estudio de negocio es más importante para la empresa de seguros los casos fraudulentos para activar los respectivos correctivos para mitigar el fraude de seguros.

Árbol de decisión.

El informe de clasificación para el modelo Árbol de Decisión es el siguiente:

Ilustración 12:

Clasificación del modelo Árbol de Decisión

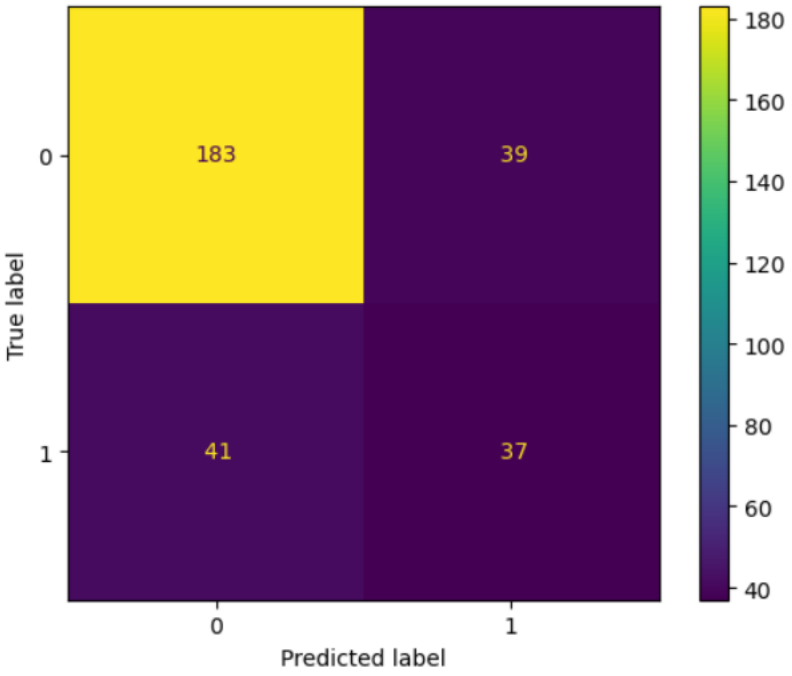
	precision	recall	f1-score	support
0	0.83	0.81	0.82	222
1	0.49	0.53	0.51	78
accuracy			0.73	300
macro avg	0.66	0.67	0.66	300
weighted avg	0.74	0.73	0.74	300

Nota. Esta tabla muestra la correspondencia entre las predicciones del modelo y las clases reales a las que pertenecen los ejemplos de prueba, con una exactitud de predicción de 73%. Elaboración propia desde Jupyter Notebook

El modelo proyecta una exactitud de predicción de 73%, sin embargo, este resultado se debe de revisar con precaución, ya que los modelos de Machine Learning pueden tener tendencia a sobre ajustarse a la clase mayoritaria (No Fraude) debido a su representación dominante en los datos. Por tal motivo se estudiarán métricas más específicas para la clase (fraude y no fraude). El modelo árbol de decisión, presenta una sensibilidad (recall) para los registros (fraude) es de 53% y para (no fraude) del 81%.

Ilustración 13:

Matriz de confusión modelo Árbol de Decisión



Nota. La matriz de confusión está compuesta por dos partes: Casos reales (True Label) y Predicción del modelo (Predicted Label). La clase 'No Fraude' identificada con 0 y el 'Fraude' con 1. Elaboración propia desde Jupyter Notebook

En la matriz de confusión que arroja el modelo, se puede concluir que 183 de los datos son clasificados correctamente como 'No Fraude', también se detecta que 39 de los registros son erróneos en la clasificación por el modelo. Por otra parte, el algoritmo clasifica correctamente 37 datos como fraudes de un total del 78, lo que indica que se equivoca 41 oportunidades en decir que no es fraude cuando en realidad si lo son. En la ilustración 17 se muestra la curva ROC para el desempeño de este modelo de manera comparativa para los tres modelos aplicados.

Como observación del modelo, se puede identificar que el Árbol de Decisión, es regular en la predicción de 'Fraudes' con tan solo un 53% de casos con una precisión de 49%. Por otro lado, en la parte de 'No Fraudes', el modelo no es tan eficaz por que detecta un 81% de casos

con una precisión del 83%, el valor entre más cerca al cien por ciento es más útil en la predicción de las clases. Para las entidades encargadas en temas de las pólizas, no serviría de gran ayuda el modelo, ya que no hay una predicción tan eficaz en los casos de fraudulentos.

XGBoost

El informe de clasificación para el modelo XGBoost es el siguiente:

Ilustración 14:

Clasificación del modelo XGBoost

	precision	recall	f1-score	support
0	0.91	0.95	0.93	222
1	0.84	0.72	0.77	78
accuracy			0.89	300
macro avg	0.87	0.83	0.85	300
weighted avg	0.89	0.89	0.89	300

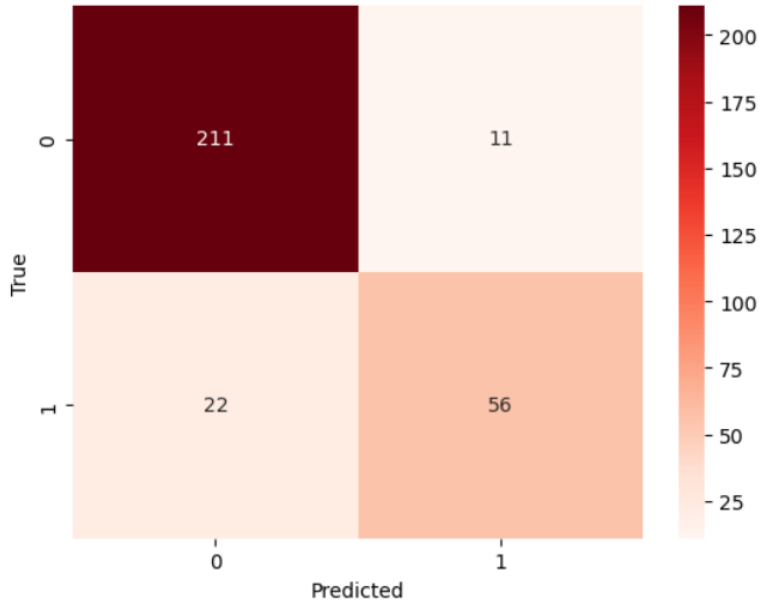
Nota. Esta tabla muestra la correspondencia entre las predicciones del modelo y las clases reales a las que pertenecen los ejemplos de prueba, con una exactitud de predicción de 89%. Elaboración propia desde Jupyter Notebook

El modelo arroja una exactitud de predicción de 89%, sin embargo, este resultado se debe de analizar con precaución porque el modelo puede mostrar una alta precisión general, puede tener un rendimiento deficiente en la clasificación de la clase minoritaria (Fraude) que en realidad es de mayor interés. Por tal motivo se estudiarán métricas más específicas para la clase (‘Fraude’ y ‘No Fraude’).

Por otro lado, la sensibilidad del modelo (Recall) para la clase mayoritaria (no fraude) es del 95% y un 75% para ‘Fraude’, esto generando un buen desempeño para la minimización de los casos fraudulentos.

Ilustración 15:

Matriz de confusión modelo XGBoost



Nota. La matriz de confusión está compuesta por dos partes: Casos reales (True Label) y Predicción del modelo (Predicted Label). La clase 'No Fraude' identificada con 0 y el 'Fraude' con 1. Elaboración propia desde Jupyter Notebook

Para la matriz de confusión de este modelo, se puede concluir que 211 de los registros son clasificados correctamente como 'No fraude' por el algoritmo, tan solo detecta 11 casos en el que se clasifican incorrectamente. Por otra parte, el modelo predice correctamente 56 registros como fraude de un total de 78, lo que muestra que se equivoco san solo 22 ocasiones en decir que no era fraude cuando en realidad si lo era. En la ilustración 17 se muestra la curva ROC para el desempeño de este modelo de manera comparativa para los tres modelos aplicados.

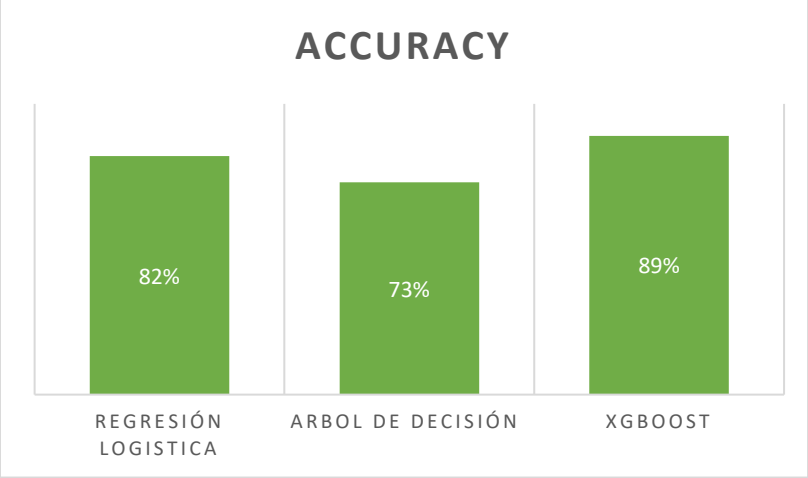
Como observación del modelo, podemos afirmar que el XGBoost es bueno en la predicción. Tanto para los fraudes y no fraudes con porcentaje mayores a 75, incluso llegando casi al 100% en la predicción de los no fraudulentos. La modelo sería de gran ayuda para las aseguradoras, ya que podría minimizar por completo las anomalías que se presentan en las solicitudes de reclamación de las pólizas.

Análisis Comparativo

Luego del análisis individual que se realizo a cada modelo, se puede hacer una comparación para determinar el más eficaz en la predicción de fraudes en la reclamación de pólizas. También se estudia cada uno de los modelos mediante la métrica general de la Curva Roc, la cual son herramientas comunes para evaluar y comparar el rendimiento de clasificadores binarios en Machine Learning.

***Ilustración 16*:**

Comparativo del Accuracy

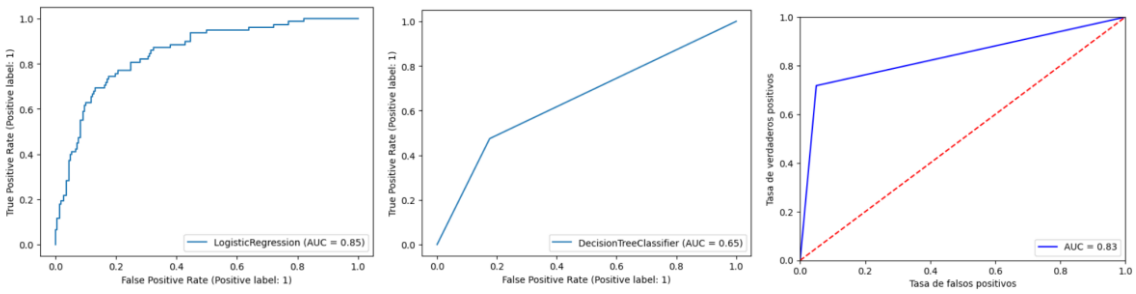


Nota. El Accuracy refleja la exactitud de cada modelo en la predicción. Los porcentajes corresponden al detalle individual de cada modelo. Elaboración propia, fuente Excel.

En la explicación detallada que se realizo anteriormente, con las métricas específicas sobre cada uno de los modelos, se deja evidenciado que el XGBoost presenta una mejor exactitud (Accuracy) de predicción en comparación a los otros dos modelos. En otras palabras, el Accuracy mide la fracción de casos clasificados correctamente por el modelo; en este caso, los procesos que son ‘Fraudes’ o ‘No Fraudes’. A continuación, se muestras las Curvas Roc para cada modelo:

Ilustración 17:

Curvas ROC, Modelos: Regresión Logística, Árbol de Decisión y XGBoost



Nota. La curva Roc, en el eje X representa la tasa de falsos positivos (FPR), mientras que el eje Y representa la tasa de verdaderos positivos (TPR) o sensibilidad. Los tres gráficos pertenecen a los modelos que están en el siguiente orden, de izquierda a derecha: Regresión Logística, Árbol de Decisión y XGBoost. Elaboración propia, fuente Jupyter Notebook.

Para evaluar la forma de la curva, primero se analiza la forma que toma la curva ROC. Una curva que se acerca a la esquina superior izquierda del gráfico indica un mejor rendimiento del clasificador. Cuanto más alejada esté la curva de la diagonal, mejor será la capacidad del modelo para distinguir entre clases positivas y negativas.

En el caso de los tres modelos, en cada uno de ellos están por encima de la diagonal, acercándose a 1 en el eje Y, dando como respuesta a que es mayor el número de tasa de verdaderos positivos en comparación a los falsos positivos.

Como se observa en la **Ilustración 17,** el modelo que está más cerca a 1 en el eje Y es la regresión logística quien presenta un mayor grado con un 85 en el AUC (Area Under the Curve) área bajo la curva, traducido del inglés.

Conclusiones

Se realizó un análisis estadístico descriptivo de la base de datos, donde se logro evidencia de patrones que permitieron observar la detección de fraudes en las solicitudes de reclamación. En este proceso eliminaron variables, que permiten un mejor estudio en cada uno de los modelos.

La aplicación de cada uno de los modelos lleva a encontrar porcentajes de Accuracy (Exactitud) en la predicción de las clases. Las métricas especificas como la mencionada anteriormente evalúan teniendo en cuenta factores como el desbalanceo de la base de datos; esto debido a que hay muchos más datos detectados como ‘No Fraude’ y el modelo aprende más de clases mayoritarias.

La comparación de los modelos mediante la métrica general de la curva Roc arroja un resultado diferente, favoreciendo al modelo de Regresión Logística en comparación a la evaluación por métricas especificas que resalta el modelo XGBoost como el mas eficaz en la predicción de casos fraudulentos.

Los errores de clasificación asimétricos que se evidenciaron en esta investigación se deben a que el Accuracy se calcula considerando todas las clases por igual, sin tener en cuenta las tasas de falsos positivos y falsos negativos. Sin embargo, la curva ROC muestra cómo varían estas tasas a medida que se ajusta el umbral de decisión. Si el clasificador tiene un alto TPR (tasa de verdaderos positivos) pero también un alto FPR (tasa de falsos positivos), es posible que el Accuracy sea alto pero la curva ROC indique un rendimiento deficiente en términos de equilibrio entre los errores de clasificación.

Anexos

Anexo 1:

Código: importación de librerías a Jupyter Notebook

```
In [1]: ##LIBRERIAS##

        import pandas as pd
        import seaborn as sns
        import numpy as np
        import seaborn as sns
        import matplotlib.pyplot as plt
        import matplotlib.pyplot as plt
```

Anexo 2:

Código: cargar base de datos a Jupyter Notebook

```
In [2]: seguros=pd.read_csv('insurance_claims.csv')
        seguros.columns

Out[2]: Index(['months_as_customer', 'age', 'policy_number', 'policy_bind_date',
               'policy_state', 'policy_csl', 'policy_deductable',
               'policy_annual_premium', 'umbrella_limit', 'insured_zip', 'insured_sex',
               'insured_education_level', 'insured_occupation', 'insured_hobbies',
               'insured_relationship', 'capital-gains', 'capital-loss',
               'incident_date', 'incident_type', 'collision_type', 'incident_severity',
               'authorities_contacted', 'incident_state', 'incident_city',
               'incident_location', 'incident_hour_of_the_day',
               'number_of_vehicles_involved', 'property_damage', 'bodily_injuries',
               'witnesses', 'police_report_available', 'total_claim_amount',
               'injury_claim', 'property_claim', 'vehicle_claim', 'auto_make',
               'auto_model', 'auto_year', 'fraud_reported', '_c39'],
              dtype='object')
```

Anexo 3:

Código: Análisis bivariado dos variables

```
In [17]: tabla4=pd.crosstab(seguros.authorities_contacted,seguros.fraud_reported, normalize="columns")
         tabla4.plot(kind="bar")
```

Anexo 4:

Código: revisar información de variable

```
In [16]: seguros["vehicle_claim"]
```

Anexo 5:

Código: Categorización de los datos en variables.

```
In [20]: seguros['incident_hour_of_the_day']=seguros['incident_hour_of_the_day'].replace([1],"Madrugada")
         seguros['incident_hour_of_the_day']=seguros['incident_hour_of_the_day'].replace([2],"Mañana")
         seguros['incident_hour_of_the_day']=seguros['incident_hour_of_the_day'].replace([3],"Tarde")
         seguros['incident_hour_of_the_day']=seguros['incident_hour_of_the_day'].replace([4],"Noche")
         seguros['incident_hour_of_the_day'].value_counts()
```

Anexo 6:

Código: Reemplazar datos en variables

```
In [9]: seguro_1.replace('?', np.nan, inplace = True)
```

```
In [10]: sns.heatmap(seguro_1.isnull(), cbar=False)
```

Anexo 7:

Código: Imputación de variables

```
In [14]: seguro_2['collision_type'].fillna(seguro_2['collision_type'].mode()[0], inplace=True)
         print("Valores perdidos collision_type: " +
               str(seguro_2['collision_type'].isnull().sum()))
```

```
Valores perdidos collision_type: 0
```

Anexo 8:

Código: Eliminación de variables

```
seguro_1=seguros.drop(columns=["_c39","umbrella_limit","incident_state","incident_city","incident_location","witnesses",
                               "police_report_available"])
```

Anexo 9:

Código: Escalamiento de variables

```
from sklearn.preprocessing import StandardScaler
# crear un objeto Scaler y ajustarlo a los datos
scaler = StandardScaler()
scaler.fit(numericas)
X_scaled = scaler.transform(numericas)
```

Anexo 10:

Código: Tabla de escalamiento de variable

```
df_1=pd.DataFrame(X_scaled)
df_1
```

Anexo 11:

Código: Separación de conjunto de datos: Entrenamiento y Testeo

```
X_train, X_test, y_train, y_test = train_test_split( df_ultima, seguros_y, test_size = 0.3, random_state = 1234, shuffle=True)
```

Anexo 12:

Código: Creación modelo Regresión Logistica

```
# Entrenamos un algoritmo basado en regresión logística
from sklearn.linear_model import LogisticRegression

lr = LogisticRegression()
lr.fit(X_train, y_train)
```

```
C:\Users\jmag0\anaconda3\lib\site-packages\sklearn\utils\validation.py:1688: FutureWarning: Feature names only support names th
at are all strings. Got feature names with dtypes: ['int', 'str']. An error will be raised in 1.2.
  warnings.warn(
```

Anexo 13:

Código: Creación modelo Árbol de decisión

```
from sklearn.tree import DecisionTreeClassifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score
```

```
abd = DecisionTreeClassifier()
abd.fit(X_train, y_train)
```

Anexo 14:

Código: Creación modelo XGBoost

```
import xgboost as xgb
```

```
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score
dtrain = xgb.DMatrix(X_test, label=y_test)
dtest = xgb.DMatrix(X_test)
params = {
    'max_depth': 3,
    'eta': 0.1,
    'objective': 'binary:logistic',
    'eval_metric': 'logloss'}
xgboost = xgb.train(params, dtrain)
```

Anexo 15:

Código: Tabla de clasificación de los modelos

```
from sklearn.metrics import classification_report
print(classification_report(y_test, y_pred))
```

Anexo 16:

Código: Matriz de confusión para los modelos

```
from sklearn.metrics import plot_confusion_matrix

plot_confusion_matrix(lr, X_test, y_test, values_format='3g')
```

Anexo 17:

Código: Curva ROC para los modelos

```
from sklearn.metrics import plot_roc_curve

plot_roc_curve(lr, X_test, y_test)
```

Bibliografía

Ameijeiras Sánchez, D., Valdés Suárez, O., & González Diez, H. (2021). Algoritmos de detección de anomalías con redes profundas. Revisión para detección de fraudes bancarios. *Revista Cubana de Ciencias Informáticas* , 244-264.

Arana, C. (2021). Redes neuronales recurrentes: análisis de los modelos especializados en datos secuenciales. *Econstor Make Your Publications Visible*, 1-23.

Ayuso, M., Montserrat, G., & Artís, M. (1999). *Técnicas cuantitativas para la detección del fraude en el seguro del automóvil.* Barcelona, España: Trabajo de grado, Universidad de Barcelona.

Badal Valero, E., Sanjuán Díaz, A., & y Segura Gisbert, J. (2020). *Algoritmos de machine learning para la detección del fraude en el seguro de automóviles.* Valencia, España: Trabajo de grado, Universidad de Valencia.

Belhadji, B., Dionne, G., & Tarkhani, F. (2000). A Model for the Detection of Insurance Fraud. *The International Association for the Study of Insurance Economics direccion*, 517-538.

Bogoya Contreras, S. A. (2022). *Detección de fraude en afiliaciones a través de un modelo de clasificación de machine learning en una aseguradora de riesgos laborales en Colombia.* Bogota, D.C. Colombia: Trabajo de grado, Fundación Universitaria Konrad Lorenz.

Bouza, C., & Agustín, S. (2012). *La minería de datos: arboles de decisión y su aplicación en estudios médicos.* Habana, Cuba: Trabajo de grado, Universidad de La Habana, Cuba & Universidad Autónoma de Guerrero, México.

Carmona Mora, M., & Londoño Morales, L. M. (2021). *Modelos de machine learning para la detección de fraude financiero.* Medellín, Colombia: Trabajo de grado, Universidad de Antioquia.

Castellanos Heras, G. (2021). *Detección del fraude en seguros de automóvil mediante técnicas de Machine Learning.* Madrid, Españna: Trabajo de grado, Universidad Carlos III de Madrid.

Corso, C. L. (2009). *Aplicación de Algoritmos de clasificación supervisada usando Weka.* Cordoba, Colombia: Trabajo de grado, Universiad Tecnológica Nacional.

Dalhia, d. l. (2021). *¿Cuales son los fraudes en seguros de autos y como evitarlos?* Obtenido de Agentes Digitales: https://www.elespectador.com/autos/5-recomendaciones-para-evitar-fraudes-en-polizas-de-seguros/

De la Espriella, C. (2022). Cuantificación del Fraude en SOAT. *Fasecolda*, 63-67.

Espinosa Zúñiga, J. J. (2020). *Aplicación de metodología CRISP-DM para segmentación geográfica de una base de datos pública.* México: Trabajo de grado, Ingeniería Investigación y Tecnología.

FASECOLDA. (2022). *Fasecolda*. Obtenido de Fasecoldad SOAT: https://fasecolda.com/ramos/soat/tarifas-y-coberturas/tarifas-comerciales/

Frutos Serrano, S. (2021). *Comparación entre XGBoost y Regresión Lineal Múltiple para la predicción de la evolución del precio de las acciones.* Madrid, España: Trabajo de grado, Universidad Complutense de Madrid.

Galán Cortina, V. (2015). *Aplicación de la metodología Crisp-DM a un proyecto de minería de datos en el entorno universitario.* Madrid, España: Trabajo de grado, Universidad Carlos III de Madrid.

García, A., Martínez, G., Núñez, G., & Guzmán, A. (1998). *Clasificación supervisada inducción de árboles de decisión, algoritmo K- D.* Ciudad de México, México: Trabajo de grado, Instituto Politécnico Nacional.

Guba, E., & Lincoln, Y. (2002). Paradigmas en competencia en la investigacion cualitativa. *Denmanc y Haro*, 113-145.

Martínez Mayorga, S. (2017). *EL problema del fraude en el sistema SOAT. Estudio del caso colombiano 1988 - 2016.* Bogotá D.C: Trabajo de grado, Escuela Colombiana de ingeniería Julio Garavito.

Moreno Palenzuela, J. (2018). *Regresión logística basada en distancias para detección de fraude en el IRPF.* Madrid España: Trabajo de grado, Universidad Politécnica de Madrid.

Ortiz Jones, C. V., & Guzmán Seraquive, J. E. (2021). Análisis de las técnicas de machine learning aplicadas en la detección de fraudes bancarios. *Revista ciencia y tecnología*, 114-122.

Patiño Espinoza, V. (2014). *Modelo de detección de fraude en clientes del servicio de agua potable de una empresa sanitaria.* Santiago de Chile: Trabajo de grado, Universidad de Chile.

Robles Velasco, A., Cortés, P., Muñuzuri, J., & Barbadilla, M. (2020). Aplicación de la regresión logística para la predicción de roturas de tuberías en redes de abastecimiento de agua. *Dirección y Organización, ,* 78-85.

Rouhiainen, L. (2018). *Inteligencia artificial: 101 cosas que debes saber hoy sobre nuestro futuro.* Barcelona, España: Editorial Planeta, S.A.

Santamaria Ruiz, W. (2006). *Técnicas de minería de datos aplicadas en la detección de fraude: estado del arte.* Bogota, Colombia: Trabajo de grado, Universidad Nacional de Colombia.

Shah, B. (2018). *Kaggle*. Obtenido de Auto Insurance Claims Data: https://www.kaggle.com/datasets/buntyshah/auto-insurance-claims-data

Viteri Gutiérrez, J. F. (2020). *Marco modelo en la prevención del fraude en el ramo de automóviles en el sector asegurador de la ciudad de Bogotá.* Bogotá, D.C: Trabajo de grado, Universidad de Santo Tomas.