Entrega #2

Integrantes

Giovani Steven Cardona Marín

Jhon Alexander Botero Gómez

Proyecto

Allstate Claims Severity

Competición de kaggel

Semestre

2023-2

Introducción

En el siguiente proyecto, se desarrollará un modelo predictivo cuyo objetivo es predecir la severidad de los reclamos a la aseguradora Allstate.

Para este propósito, se empleará un conjunto de datos proporcionado por la compañía, el cual consta de 188,318 filas y 131 columnas. Estas columnas incluyen datos categóricos, datos numéricos y la cantidad de pérdida en siniestros (accidentes o daños a bienes asegurados).

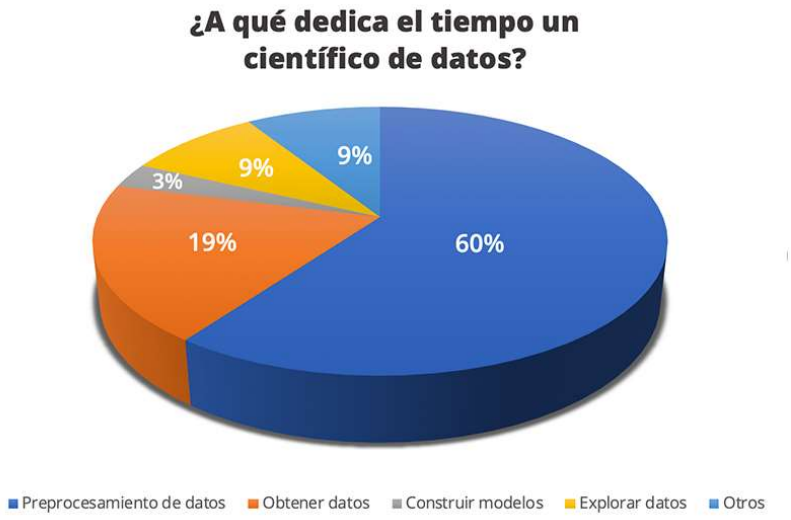
Experimentación

La experimentación comenzó abordando el problema mediante la simulación de la eliminación del 5% de los datos de tres columnas seleccionadas al azar, con el fin de garantizar la aleatoriedad del modelo. Una vez identificadas las columnas, se procedió a la eliminación de los datos. Sin embargo, surgió un desafío al intentar llenar los valores faltantes.

Este desafío radicaba en la naturaleza de los datos categóricos, ya que, al llenarlos, podríamos introducir un sesgo en el modelo al realizar elecciones arbitrarias, dado que el significado de cada variable estaba enmascarado para proteger información sensible contenida en el conjunto de datos y salvaguardar la privacidad de la empresa. Esta limitación restringió nuestras opciones para el análisis y limpieza de los datos.

Ante esta limitación, decidimos enfocarnos en las columnas de tipo numérico para simular una forma de llenar los datos que no afectara significativamente el modelo predictivo. Dada la cantidad de datos disponibles y la escasa información sobre las categorías, optamos por utilizar la media de la respectiva columna. Esta elección se basó en la dificultad de encontrar una mejor estrategia de llenado que tuviera en cuenta múltiples factores del conjunto de datos.

Una vez completada esta etapa, nos sumergimos en la investigación sobre cómo abordar este tipo de problemas y cuál sería el siguiente paso para proporcionar una solución más efectiva a los modelos predictivos. Durante nuestra búsqueda, encontramos una imagen que ilustraba el proceso de cómo un científico de datos aborda un problema.

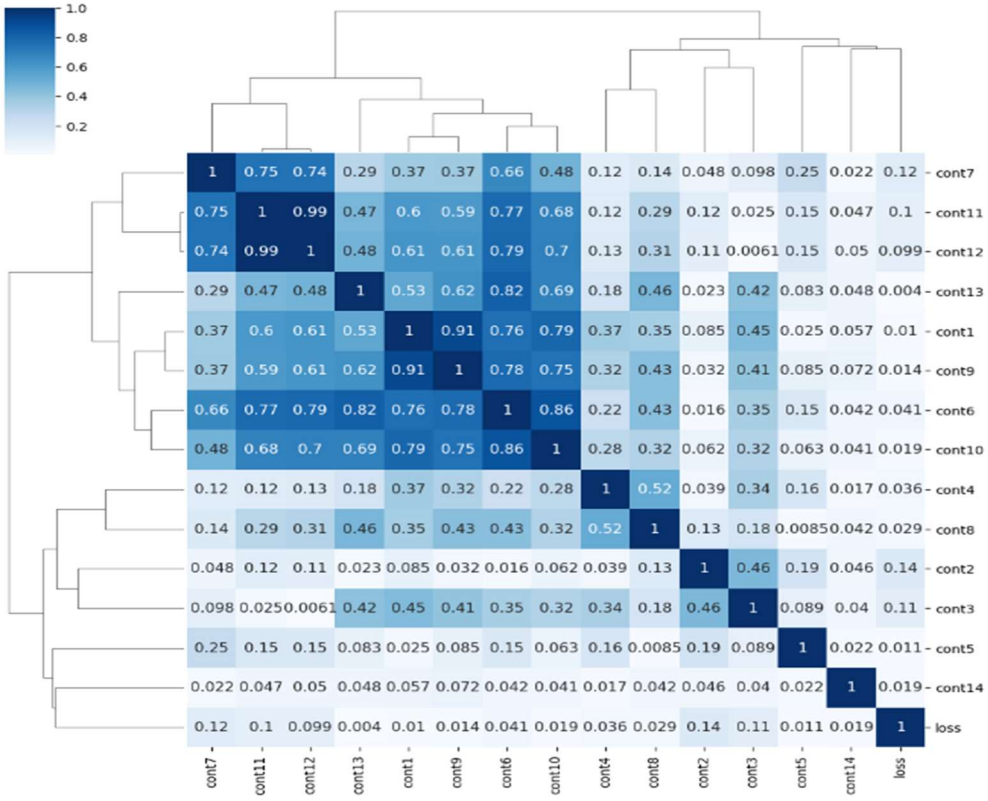


Fuente: https://www.xeridia.com/blog/la-importancia-del-preprocesamiento-de-datos-en-inteligencia-artificial-limpieza-de-datos

En dicha imagen, se nos indica que la mayor parte del tiempo empleado para encontrar una solución efectiva comienza con el análisis de la información que tenemos disponible. Esto implica evaluar cuidadosamente los datos que resultan relevantes para nuestro objetivo y aquellos que podrían ser problemáticos para el modelo. Este proceso de selección es crucial para construir una base sólida.

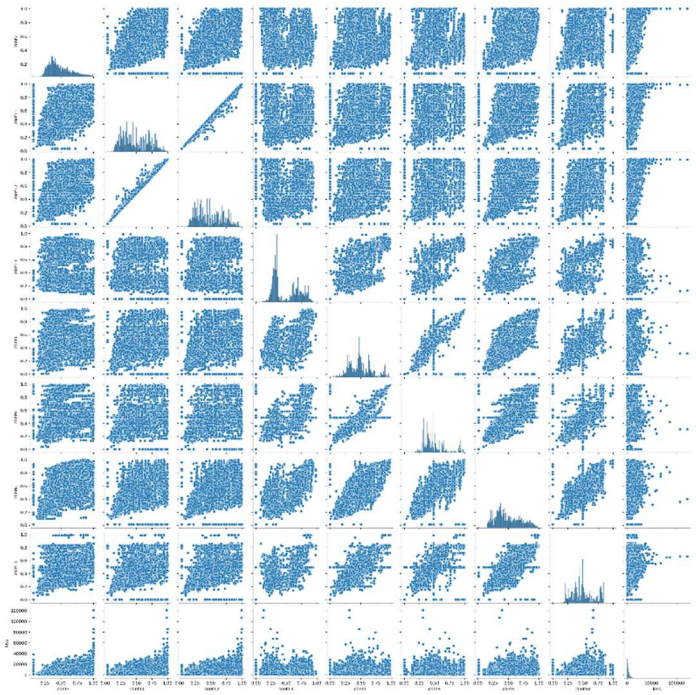
Además, debemos explorar cómo extraer información valiosa de los datos disponibles, de manera que contribuya a mejorar las predicciones. Esto puede implicar la observación de patrones en el comportamiento de los datos, como la agrupación de categorías en ciertos puntos o tendencias específicas.

Cuando se inicia el análisis de los datos, se construye una matriz de correlaciones de las variables continuas. Esto se hace para identificar las relaciones existentes entre ellas, estableciendo si estas correlaciones son fuertes, moderadas o débiles. Esta información es esencial para comprender cómo las diferentes variables interactúan entre sí y para guiar la toma de decisiones en la construcción del modelo.



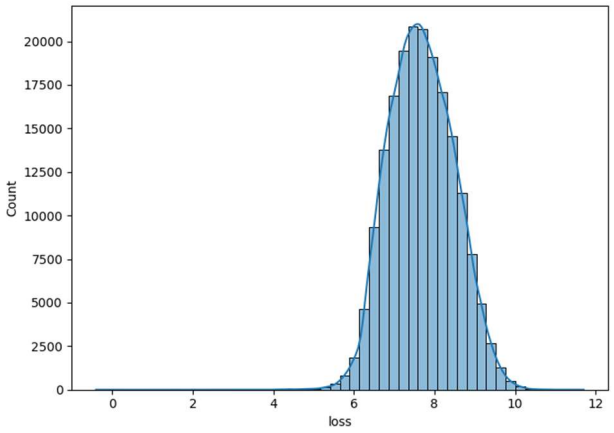
Se observa gráficamente una correlación fuerte entre las variables: cont7, cont11, cont12, cont13, cont1, cont9, cont6 y cont10.

Por lo anterior, se imprime gráficamente la matriz de correlación solo de las 8 variables significativas.



En la grafica se puede observar relaciones lineales entre dos pares de variables cont11-cont12 y cont1-cont9.

Por otro lado, también analizamos la variable respuesta “loss” para detectar el comportamiento de sus datos por lo que nos apoyamos de la librería seaborn, matplotlib para imprimir el histograma de la columna menciona, como se muestra a continuación.



Video: https://youtu.be/i2iSjzopd0k