Laboratorio de Construcción de Software

TP Inicial - Entrega 2

Alumnos:
Perez Giannina - DNI: 43 729 769
Prieto Lucas - DNI: 43 626 494

2023

Introducción

Como ya vimos anteriormente, la tecnología en la actualidad es muy importante ya que influye prácticamente en todos los aspectos de nuestras vidas, desde nuestra forma de comunicarnos hasta la toma de decisiones laborales. Dentro de todos los sistemas y aplicaciones que conforman este mundo tecnológico, se encuentra comúnmente el proceso de preparación de datos.
La preparación de datos es esencial para todas las fases de un proyecto, cuyas necesidades están relacionadas directamente con la necesidad obtener datos de calidad.

En este documento, abordaremos qué implica la preparación de datos, su importancia, su relación con Machine Learning y los pasos que se deben seguir para llevar a cabo este procedimiento. A la par, estaremos desarrollando los modelos seleccionados de Machine Learning para lograr nuestro objetivo propuesto en el documento anterior, determinar qué variables están relacionadas con un mayor riesgo de mortalidad en incidentes viales. Finalmente, documentaremos cómo se llevó a cabo el entrenamiento de nuestros modelos y las conclusiones correspondientes.

Preparación de datos

La preparación de datos puede definirse como las operaciones que se realizan sobre los datos brutos para hacerlos analizables. Es decir, es el proceso de limpiar, transformar y estructurar los datos en bruto de manera que sean adecuados y útiles para el análisis, modelado y aplicación informática. Para poder lograrlo, se deben llevar a cabo una serie de actividades esenciales, las cuales garantizan que los datos sean coherentes, precisos y completos. Esto, a su vez, mejora la calidad de los resultados obtenidos.

Machine Learning y la preparación de datos

Como ya vimos anteriormente, el Machine Learning se utiliza para desarrollar algoritmos y modelos que se entrenan con ejemplos de datos.
Para este proceso de entrenamiento, la preparación de datos es una fase crucial, pues los datos de entrada deben estar limpios, transformados y organizados de manera adecuada para que los modelos de Machine Learning puedan aprender patrones precisos y tomar decisiones informadas basadas en esos datos.

¿Por qué es importante la preparación de los datos?

Calidad de Resultados: Los resultados obtenidos de cualquier análisis o modelo serán tan buenos como los datos que se utilicen. Datos sucios o mal estructurados pueden llevar a conclusiones erróneas.

Eficiencia en el Análisis: Los datos limpios y preparados permiten un análisis más rápido y preciso. Los profesionales no tendrán que gastar tiempo valioso depurando datos en medio del análisis.

Toma de Decisiones Informadas: Las decisiones basadas en datos son más sólidas cuando se trabaja con datos bien preparados. Esto es esencial tanto en el ámbito empresarial como en la investigación académica.

Desarrollo de Modelos Precisos: Los modelos de aprendizaje automático y análisis estadístico dependen de datos de calidad para generar predicciones precisas.

Pasos a seguir

Paso 1: Adquirir datos

El primer paso es adquirir los datos necesarios para el Machine Learning. Estos datos pueden proceder de diferentes lugares y tener diferentes formatos. Se recomienda recolectar datos de una fuente que posea varios formatos, pues esto

puede significar perjudicial a largo plazo. Existe una amplia diversidad de fuentes para recopilar datos, tales como bases de datos, aplicaciones, archivos CSV, APIs, entre otros.

En nuestro caso obtuvimos en primer lugar un conjunto de datos sobre incidentes viales fatales ocurridos en Argentina, pero nos dimos cuenta que necesitábamos datos sobre incidentes no fatales también para que el modelo pueda ver el contraste, aprender y lograr el objetivo de predecir si hay mayor riesgo de sufrir un accidente fatal según las variables independientes. Debido a esto, decidimos buscar otro conjunto de datos complementario para añadir esta información. El conjunto de datos que seleccionamos para lograr esto fue “Road accidents and conditions” del sitio www.kaggle.com. Este dataset contiene información sobre accidentes viales no fatales y fatales ocurridos en Gran Bretaña y lo seleccionamos porque tiene las mismas variables que el primer conjunto de datos seleccionado (momento del hecho, clima, tipo de vehículo, tipo de calle, etc).

Paso 2: Exploración de datos

Este paso puede considerarse como una etapa de pre eliminación, el objetivo principal es buscar y encontrar posibles errores. Una visualización rápida es útil para este propósito porque permite reconocer al instante si el formato de los datos es el correcto, identificar valores nulos o atípicos. Es común que en esta fase se utilicen funciones para detectar problemas que se corregirán en la siguiente etapa. Por ejemplo, funciones de clasificación para detectar duplicados.

En nuestro trabajo el proceso de exploración nos sirvió para encontrar filas con valores nulos que nos servirían, y para ver que los dos datasets elegidos tenían similitudes pero también diferencias que tendríamos que homogeneizar para formar un solo archivo con la información estructurada de ambos. Algunas de estas diferencias eran el idioma, la forma de clasificar los tipos de calle, los tipos de vehículos, entre otras.

Paso 3: Limpieza de datos

Esta etapa se destaca por la organización, limpieza de datos y corrección de errores. A la hora examinar datos perdidos o eliminar datos innecesarios, se deben

rellenar estos espacios para garantizar la calidad de los datos. Por otro lado, la limpieza también puede implicar algunas transformaciones. Por ejemplo, se puede detectar una categoría de datos que necesita ser modificada para ser utilizable o reescribir datos que contengan valores atípicos.

En nuestro conjunto de datos de Argentina decidimos prescindir de algunas columnas que pensamos que no serán importante analizar para nuestro objetivo. Las columnas son fuente_datos, numero_victima, municipio_id, municipio_nombre, fecha_hecho, hora_hecho (ya que usaremos la columna momento_hecho que indica si fue de día o de noche).
Mientras que en el conjunto de datos de Gran Bretaña nos quedamos con las columnas que se asemejan a las que utilizaremos del otro conjunto. Estas son:

- Number of Vehicles: la utilizamos como tipo_incidente para determinar si fue entre dos vehículos, un vehículo y una persona, o un vehículo y un objeto / vuelco.
- 1st Road Class: indica el tipo de vía según la clasificación inglesa.
- Lighting Conditions: indica si el incidente fue de día o de noche (columna momento_hecho).
- Weather Conditions: clima.
- Casualty Class: clase_victima.
- Casualty Severity: variable dicotómica, para determinar si el accidente fue letal o no.
- Sex of Casualty: sexo.
- Age of Casualty: edad
- Type of Vehicle: vehiculo_victima

Paso 4: Transformación de datos

Como se comentó anteriormente, los datos adquiridos pueden estar disponibles en muchas formas, tamaños y estructuras. Algunos están listos para el análisis, mientras que otros conjuntos de datos pueden estar contenidos dentro de un formato difícilmente legible. Por ello, es fundamental transformar los datos para garantizar que se encuentren en un formato o una estructura adecuada. Esto variará en función del software o el lenguaje que el analista utilice para realizar su análisis de datos. Dentro de las transformaciones se puede: dinamizar o cambiar la

orientación de los datos, convertir los formatos de fecha, agregar datos de ventas y rendimiento a través del tiempo.

Como se mencionó anteriormente tuvimos que realizar transformaciones de los datos para unir la información de los dos datasets.
En primer lugar agregamos la columna “fatal” que indica si la víctima del incidente vial falleció (1) o no (0).
También se tradujeron los datos del dataset “Road Accidents”. Para esto necesitamos hacer una búsqueda para poder entenderlos bien y traducirlos de manera correcta. Por ejemplo, la clasificación de tipos de vías estaba dividida en las categorías A, B, A(M), Motorway y Unclassified. Estas siglas se determinan por una convención inglesa y significan:

- A: Ruta nacional.
- B: Ruta provincial.
- A(M): Autopista nacional.
- Motorway: Autopista provincial.
- Unclassified: Calle.

Otra transformación fue la de la columna Number of Vehicles. En el caso de tener involucrado un único vehículo y tener como víctima un peatón, se clasificó como tipo_incidente = *Colision vehiculo/Persona*. Y si la víctima era el conductor o un pasajero, el tipo de incidente es *Colision vehiculo/Objeto-Vuelco*. En caso de tener más de un vehículo involucrado en el accidente, el tipo de incidente se clasificó como *Colision vehiculo/Vehiculo*.
También la información de la columna Casualty Severity la pasamos a la nueva columna “Fatal”, donde los valores “Slight” y “Serious” se convirtieron en 0 y los valores “Fatal” en 1.

Luego de realizar las traducciones y unir todos los datos al mismo archivo, nos quedaron las siguientes variables con sus posibles valores:

- momento_hecho: Diurno o Nocturno.
- Clase_victima: Conductor, pasajero o peatón
- Vehiculo_victima: Automóvil, motocicleta, camión, colectivo, camioneta, bicicleta, vehículo, peatón u otro.
- Edad

- Sexo: Masculino o femenino.
- Tipo_via: Autopista nacional, autopista provincial, ruta nacional, ruta provincial o calle.
- Tipo_incidente: Colision vehiculo/Vehiculo, Colision vehiculo/Objeto-Vuelco, Colision vehiculo/Persona u Otro modo.
- Clima: Bueno, lluvia, llovizna, ventoso, niebla, nublado u otro.
- Fatal: 1 (si la víctima murió) o 0 (si la víctima no murió)

Paso 5: Validación y Visualización

En la última etapa los datos están limpios y etiquetados, los equipos de Machine Learning suelen rever los datos para asegurarse de que son correctos y están listos. Las visualizaciones como histogramas, gráficos de dispersión, gráficos de caja, gráficos de línea y gráficos de barra son herramientas útiles para confirmar que los datos son correctos. Además, las visualizaciones también ayudan a los equipos de ciencia de datos a completar análisis exploratorios de datos.

Este proceso utiliza las visualizaciones para detectar patrones, encontrar anomalías, probar una hipótesis o verificar supuestos. Los análisis exploratorios de datos no requieren un modelado formal; en lugar de eso, los equipos de ciencia de datos pueden utilizar visualizaciones para descifrar los datos.

Entrenamiento del modelo

Una vez que conseguimos tener datos variados sobre accidentes fatales y no fatales, nos proponemos continuar con el objetivo de entrenar el modelo con regresión logística clasificar accidentes entre los que tienen más probabilidades de tener víctimas fatales y los que no.

Para realizar el primer entrenamiento a nuestro modelo utilizamos Jupyter Notebooks. El primer ajuste que realizamos fue transformar las variables categóricas. Las variables categóricas son aquellas que representan diferentes categorías o etiquetas, como género, tipo de incidente, clima, etc.

Estas variables deben ser codificadas antes de usarlas en un modelo de regresión logística para que el modelo pueda trabajar con ellas.
Esto lo hicimos utilizando la codificación “one-hot”, la cual convierte las variables categóricas en numéricas. Por ejemplo, la variable clima que contiene diferentes tipos de clima, la codificación one-hot la transforma en variables binarias para cada tipo de clima (clima_Bueno, clima_Lluvia, etc).
También es importante mencionar que manejamos los datos de forma aleatoria para que el algoritmo haga los ajustes de forma automática.

Conclusiones

Nos parece importante destacar la relevancia de la preparación de datos dentro del Machine Learning, pues si los datos ingresados como entrada fueran desorganizados o estuvieran incompletos, los resultados de los modelos podrían llegar a ser erróneos. Debido a esto, creemos que la preparación de datos es fundamental para que los algoritmos funcionen de manera eficiente, pues mejoran la precisión y la capacidad de adaptación. De esta manera, podemos asegurarnos de que nuestros modelos de Machine Learning ofrecerán resultados sólidos. A modo de cierre, podemos decir que la preparación de datos influye considerablemente en la evolución de la Inteligencia Artificial y del Machine Learning.

Bibliografía

Preparación de datos | Into The Minds
¿Qué es la preparación de datos? | AWS
¿Qué es la preparación de datos? | Alteryx

Instalar Python Anaconda | Aprende Machine Learning
Regresión Logística con Python | Aprende Machine Learning

OpenRefine: Carga de CSV | Youtube
OpenRefine: Funciones GREL | Youtube
Guía para limpieza de datos con OpenRefine | GBIF
Road accidents and conditions | Kaggle
Análisis de datos y preparación para la regresión logística [parte 15] | Microsoft Learn