

**Limpieza de datos conjuntos**

Javier Lizarraga Beyles - A01253105

Natalia Velasco Garcia - A01638047

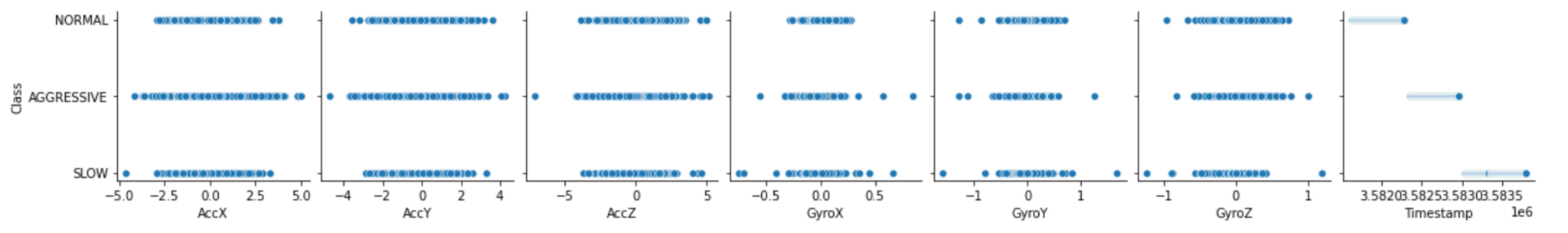
Jose Luis Rosa Cruz - A01638241

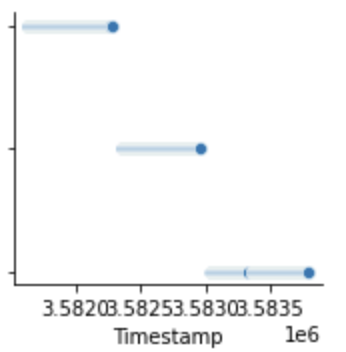
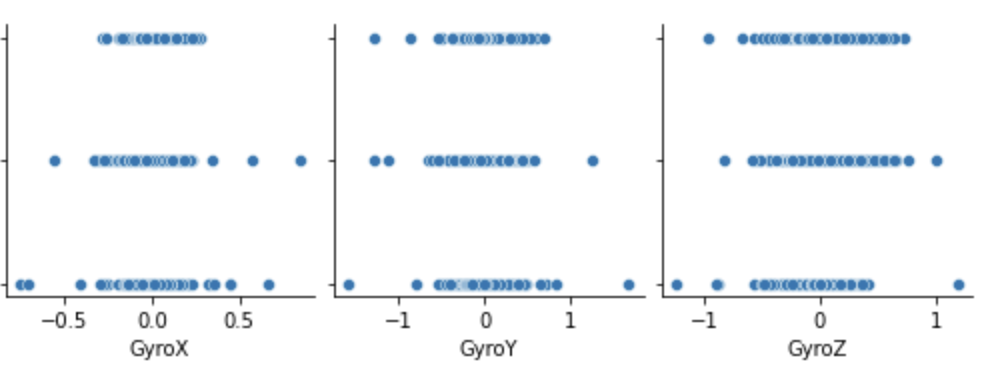
Cesar Ivann Llamas Macias - A01625272

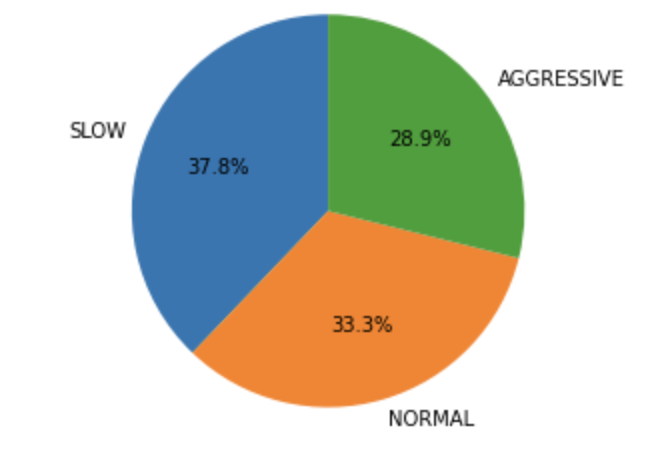
26 de agosto de 2022

Nuestro reto es analizar una base de datos sobre el comportamiento de un conductor con dos muestras de datos por segundo, capturando los datos de sus movimientos en aceleración y dirección en ejes X, Y y Z además de una clasificación sobre el movimiento, con etiquetas de SLOW(lento), NORMAL(normal) y AGGRESSIVE(agresivo) con además una muestra del tiempo con el que se tomó el dato. El objetivo es identificar cuales son las características para clasificar un movimiento con las etiquetas anteriormente mencionadas, con esto podremos identificar cuáles son los comportamientos que pudieran causar algún riesgo en el tráfico.

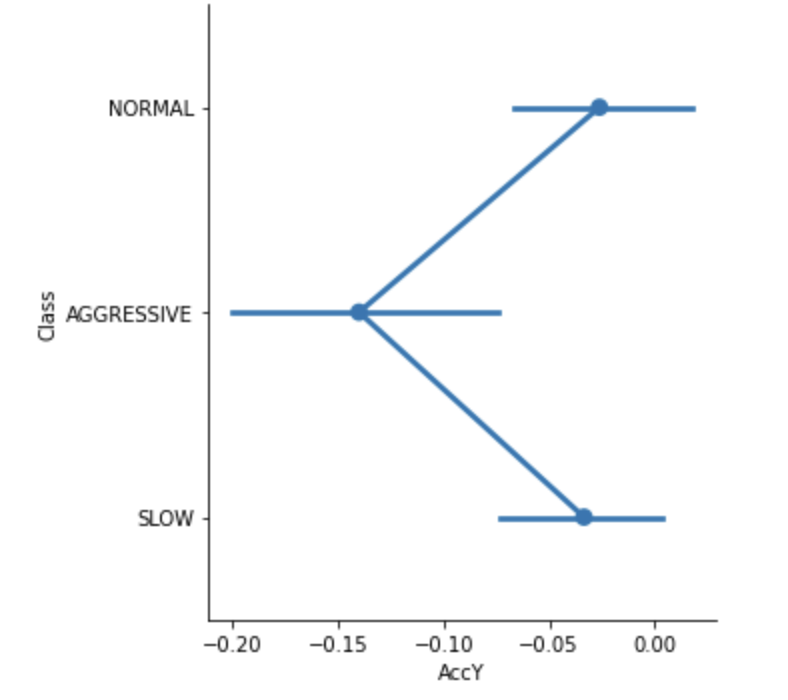
El primer paso es realizar una limpieza de nuestra nueva base de datos, revisamos entre ellos datos faltantes o duplicados, los cuales no encontramos por lo que no hay necesidad de alterar la base de datos para un análisis preciso.

Tras tener nuestros datos completos tratamos de aprender un poco sobre los datos, demostrando gráficamente su distribución.

Lo primero que podemos observar es que las muestras clasificadas como agresivas poseen más muestras con valores extremos o irregulares, aunque no pertenecen únicamente a este grupo, también podemos observar que la variable del tiempo no podrá ser utilizada como una variable independiente o una variable que nos ayude a predecir por sí sola un resultado, pues las muestras fueron tomadas en segmentos ya clasificados por uno tras otro, por lo que no aportan mucho valor al modelo, dando la posibilidad de que nuestro modelo sea erróneo al momento de generar conclusiones (e.g. pudiera concluirse que en la mañana todos conducir normal, en medio día empiezan a conducir agresivamente para terminar el día conduciendo lento) en cambio lo podremos utilizar para comparar cambios bruscos de aceleración y dirección. 

Las variables independientes serán las aceleraciones y direcciones en sus tres dimensiones, mientras que las variables dependientes, es decir las que se intentarán de obtener, o mas bien las clasificaciones en las que caeran nuestros datos son las clasificaciones de SLOW, NORMAL y AGGRESSIVE

Analizando más profundamente la población de datos con la que contamos podemos observar que predominan los datos clasificados como SLOW, seguido de NORMAL y terminando con AGRESSIVE, pero de forma general se puede ver que están relativamente balanceados los tres campos por lo que no tendremos ningún problema con una clasificación menos definida que otra. Cada una de estas categorías claro tienen actitudes caracterizantes como observaremos a continuación.



Podemos observar los rangos de cada una de las muestras, marcando principalmente la diferencia que la categoría AGGRESSIVE contiene valores con un mayor rango, junto con lo entendido anteriormente que es el grupo con menor abundancia de los tres podemos observar que este campo es el que contiene mayor cambio entre muestras, esperando cambios bruscos o extremos por parte de estas muestras.

Con esto podemos empezar a generar hipótesis, por ejemplo, pudiéramos suponer que las clasificaciones son generadas mediante las diferencias de direcciones y aceleraciones entre muestras para calcular la categoría de las nuevas muestras, necesitando un modelo clasificador a comparación de uno predictor, el cual se empieza a pensar que no sería apto para este análisis.