# How Severity the Accidents is?

CRISTIAN ALEXANDER CASTAÑO MONTOYA LUIS FELIPE CADAVID CHICA

**Introducción a la Inteligencia Artificial / 2023-1**

En el presente informe se describen los resultados de lo elaborado para nuestro problema de IA, que se basa en un dataset de kaggle.com, que busca calcular qué tan severos son los accidentes en estados unidos.

Primero que todo, se eliminaron los datos duplicados para evitar la introducción de sesgos en el modelo. Luego, se trataron los valores faltantes, que en este caso se eliminó las filas con datos faltantes, pues eran casi un 70% de las filas para este problema, reemplazarlo por un valor aleatorio nos afectaba enormemente.

La normalización también se realizó para garantizar que todas las variables estuvieran en la misma escala y no se diera más importancia a alguna variable sobre otra; En el proceso de normalización, se encontraron variables categóricas que no podían ser directamente normalizadas porque no eran numéricas. Para solucionar esto, se aplicó una técnica de encoding para convertir las variables categóricas en variables numéricas. En este caso, se utilizó la técnica de "one-hot encoding", que crea nuevas columnas binarias para cada posible valor de la variable categórica. Esto permite que el modelo pueda interpretar cada valor como una variable numérica independiente y pueda realizar cálculos con ellas en el proceso de normalización.

El dataset inicialmente contenía 100.000 rows, después de eliminar las duplicadas se redujo a 33.304, pero la Ram del entorno de ejecución del colab por lo que se optó por reducir a 7.000rows los datos usando la librería *df.sample.* Estas técnicas de preprocesado permitieron que el modelo pudiera trabajar con datos limpios y homogéneos, lo que a su vez contribuyó a mejorar la precisión de las predicciones. Este proceso se puede encontrar en el archivo *03- PRE-PROCESADO.ipynb*.

En este caso, se implementó el modelo Naive Bayes para clasificar datos en cuatro clases diferentes. El modelo fue entrenado utilizando los datos preprocesados y normalizados, y se utilizó la función "fit" de la librería scikit-learn para ajustar el modelo a los datos de entrenamiento.

Luego, se utilizó la función "predict" para predecir las clases correspondientes a los datos de prueba y se evaluó el desempeño del modelo utilizando varias métricas de evaluación, incluyendo la precisión, el recall y la matriz de confusión. Este proceso se puede encontrar en *04 - MODELO NAIVE BAYES.ipynb.*

El accuracy del modelo fue de 0.94, lo que indica que el modelo clasificó correctamente el 94% de los casos. Es decir, de cada 100 casos, el modelo clasificó correctamente 94 de ellos y erróneamente 6. El accuracy es una métrica importante para evaluar el desempeño general del modelo, pero es importante complementar con otras métricas, especialmente si hay clases desbalanceadas o si el costo de cometer errores en algunas clases es mayor que en otras.

Para evaluar la calidad del modelo, se han utilizado diversas métricas, entre ellas la precisión (accuracy), la precisión (precisión), el recall (sensibilidad) y la matriz de confusión. En particular, se han analizado los valores de precisión y recall para cada clase, que muestran el desempeño del modelo en cada una de las clases.

Los resultados de la evaluación muestran que el modelo tiene una precisión baja en la clase 0, con un valor de 0.50, lo cual indica que el modelo clasifica incorrectamente aproximadamente la mitad de las instancias de esa clase. Por otro lado, la clase 1 tiene una precisión alta de 0.96, lo que sugiere que el modelo está clasificando la gran mayoría de los casos positivos correctamente. En la clase 2, la precisión es muy baja, con un valor de 0.22, lo que indica que el modelo está cometiendo muchos falsos positivos en esta clase. En la clase 3, la precisión es moderada, con un valor de 0.42, lo que sugiere que el modelo está clasificando algunos casos positivos correctamente en esta clase.

El recall para las clases 0 y 2 es bajo con valores de 0.19 y 0.11 respectivamente, lo que indica que el modelo está identificando una proporción menor de casos positivos en estas clases. Por otro lado, las clases 1 y 3 tienen un recall alto, con valores de 0.99 y 0.25 respectivamente. En la clase 1, esto sugiere que el modelo está identificando correctamente la mayoría de los casos positivos. Sin embargo, en la clase 3, el valor de recall indica que el modelo está identificando sólo una proporción menor de los casos positivos.

Además, se ha utilizado la matriz de confusión para analizar el desempeño del modelo en cada una de las clases. Los resultados muestran que el modelo ha clasificado correctamente la mayoría de los casos en las clases 1, 2 y 3, pero ha cometido errores en la clase 0.

En conclusión, los resultados de la evaluación del modelo indican que el modelo tiene un buen desempeño en la mayoría de las clases, pero necesita mejorar su precisión en la clase 0. Por lo tanto, se buscará en futuros procesos realizar ajustes en el modelo para mejorar su desempeño en esta clase específica. Así mismo, es importante destacar la importancia de evaluar el desempeño del modelo en cada una de las clases, para asegurarse de que el modelo está resolviendo correctamente el problema completo.