

TRABAJO FINAL CODER HOUSE

Fundamentos de la ciencia de datos 1

Alumno: Jiménez Hernán Matías

ABSTRACT

El objetivo de este estudio es clasificar a las víctimas de accidentes de tráfico prediciendo su tipo de víctima, como "Conductor", "Acompañante", "Peatón", entre otros, utilizando modelos de aprendizaje automático. La variable objetivo en este análisis es "clase víctima", y la clasificación se basa en diversos factores, incluidos el tipo de vehículo, la edad, el sexo, la hora del incidente y el tipo de vía. Al analizar estas variables, el modelo busca identificar patrones y asociaciones que influyan en la probabilidad de pertenecer a una determinada clase de víctima. Comprender estas relaciones puede proporcionar información valiosa para la toma de decisiones y medidas preventivas.

Se llevó a cabo un análisis exploratorio de datos (EDA) para comprender mejor el conjunto de datos. Durante este proceso, se realizó lo siguiente:

Conversión de variables: Se utilizaron dos técnicas de codificación para transformar las variables categóricas: Label Encoder se aplicó a columnas con un número limitado de categorías para convertirlas en valores numéricos. Además, se implementó la técnica de codificación one-hot (variables dummy) para columnas categóricas con más de dos categorías, facilitando así la inclusión de estas variables en los modelos de aprendizaje automático.

Limpieza de datos: Se identificaron y eliminaron los valores NaN del conjunto de datos, asegurando que los modelos pudieran entrenarse con datos completos y coherentes.

Análisis descriptivo: Se realizaron estadísticas descriptivas para explorar la distribución y características de cada variable, identificando patrones y tendencias iniciales en los datos.

Las columnas del conjunto de datos son las siguientes:

fuelle: Indica el origen de los datos sobre el accidente.

numero victima: Un identificador único para cada víctima en el conjunto de datos.

municipio id: Identificador del municipio donde ocurrió el accidente.

municipio nombre: Nombre del municipio relacionado con el accidente.

fecha hecho: Fecha en la que ocurrió el accidente.

hora hecho: Hora específica en que tuvo lugar el accidente.

momento hecho: Momento del día en que ocurrió el accidente (por ejemplo, mañana, tarde, noche).

edad: Edad de la víctima en el momento del accidente.

sexo: Género de la víctima.

clase victima: Clasificación de la víctima (por ejemplo, Conductor, Acompañante, Peatón).

vehículo victima: Tipo de vehículo involucrado en el accidente.

tipo vía: Clasificación del tipo de vía (por ejemplo, carretera, calle urbana).

tipo incidente: Tipo de incidente que ocurrió (por ejemplo, colisión, atropello).

clima: Condiciones climáticas en el momento del accidente.

Se aplicaron distintos modelos de aprendizaje automático, entre ellos Random Forest, XGBoost y Regresión Logística, para evaluar la precisión y el rendimiento en la clasificación de las víctimas. Los modelos fueron evaluados en métricas como precisión, real y f1-score para determinar su efectividad en la predicción de la clase de víctima. Los resultados mostraron que el modelo de XGBoost obtuvo la mejor exactitud, alcanzando un 81.27%, seguido por el Random Forest con un 80.05%. En contraste, la Regresión Logística mostró un rendimiento inferior con una exactitud del 59.37%. Este análisis proporciona una comprensión integral del poder predictivo de las diferentes características y la capacidad de las técnicas de aprendizaje automático para clasificar a las víctimas de accidentes con un alto grado de precisión.

Conclusión

Los resultados obtenidos destacan la eficacia de los modelos de aprendizaje automático en la clasificación de víctimas de accidentes de tráfico. El modelo XGBoost demostró ser el más eficiente, logrando la mayor exactitud y ofreciendo un rendimiento sólido en métricas de evaluación. Estos hallazgos sugieren que factores como la edad, el sexo y el tipo de vía son influyentes en la determinación de la clase de víctima. La implementación de estos modelos.

puede contribuir a una mejor comprensión de los patrones en los accidentes de tráfico y facilitar el desarrollo de estrategias preventivas más efectivas.