Pontificia Universidad Católica de Chile

Magíster en Ciencia de datos

Problema de clasificación: Análisis de Accidentes Tránsito con resultados de muerte en Estados Unidos

Curso aprendizaje estadístico y computacional

Integrantes Richard Thomas Orellana Taibo, Luciano Lorenzo Davico

PONTIFICIA
UNIVERSIDAD
CATÓLICA
DE CHILE

Santiago, Chile, 04 de octubre del 2023

1.- Introducción

El presente informe se centra en el análisis detallado de una base de datos relacionada con accidentes de tránsito en Estados Unidos, con un enfoque específico en colisiones vehiculares. La base de datos encapsula múltiples dimensiones de información, abarcando desde detalles geográficos y temporales hasta variables específicas como el número de heridos y fallecidos. Dada la riqueza de esta información, el análisis tiene como objetivo primordial modelar la ocurrencia de fallecimientos en estos accidentes. Para ello, se emplean técnicas avanzadas de procesamiento y análisis de datos para extraer, transformar y seleccionar variables relevantes que sirvan como predictores eficaces. Además, se aplican métodos de aprendizaje supervisado y no supervisado para comprender los patrones subyacentes y hacer predicciones basadas en las características del accidente. Este enfoque técnico pretende objetivo de predecir la probabilidad de un deceso bajo ciertas condiciones del siniestro.

2.- Objetivos.

Principal:

- Modelar la ocurrencia de fallecimientos en accidentes de tránsito con el objetivo de predecir la probabilidad de un deceso bajo ciertas condiciones del siniestro. Esta predicción permitirá categorizar el accidente como "tragedia" basándose en la incidencia y relevancia de las muertes ocurridas.

Secundarios:

- Realizar un análisis exploratorio de los datos para comprender la naturaleza y distribución de los accidentes.
- Preprocesar y limpiar la base de datos, eliminando redundancias y generando variables relevantes que faciliten el análisis.
- Implementar técnicas de aprendizaje supervisado adecuadas para el modelado de la ocurrencia de fallecimientos.
- Conducir un análisis de clúster para descubrir patrones subyacentes en los datos y agrupar accidentes con características y consecuencias similares.

3. *Resultados*

a. Análisis exploratorio de datos

- **Preprocesamiento:** Se eliminaron diversas columnas relacionadas con fechas y ubicaciones para centrar el análisis en factores más directamente relacionados con los accidentes. Se introdujo la columna **tragedy** que indica si hubo al menos un fallecimiento

en el accidente.

- **Variables Predictoras:** Se introdujeron variables binarias para determinar si un accidente ocurrió en una avenida. Estas variables podrían ser cruciales para entender si ciertos tipos de carreteras están más asociados con accidentes graves.
- **Consistencia de Datos:** Se validó la relación entre las columnas que indican el número total de heridos y las que desglosan estos números por tipo de persona (peatón, ciclista, motorista). Esta coherencia es fundamental para garantizar la precisión de los análisis posteriores.
- **Tipos de Vehículos:** Se llevó a cabo una reclasificación de los tipos de vehículos involucrados en los accidentes, agrupándolos en categorías más generales. Esto simplifica el análisis y permite identificar tendencias más claramente.

b. Aplicación de técnicas de aprendizaje supervisado

Dado el desafío de predecir la ocurrencia de fallecimientos en accidentes de tránsito, se emplearon varias técnicas de aprendizaje supervisado para construir modelos predictivos eficaces:

- **Random Forest (Bosques Aleatorios):** Se seleccionó este algoritmo por su eficiencia y capacidad para manejar grandes conjuntos de datos con múltiples características.
- **Naive Bayes:** Dado que las covariables predictoras en nuestro conjunto de datos podrían considerarse independientes entre sí, Naive Bayes se consideró una opción adecuada. Además, es eficiente y funciona bien en conjuntos de datos de gran tamaño.
- **Máquinas de Vectores de Soporte (SVM):** Se consideró esta técnica debido a su capacidad para separar espacios de características complejas y manejar datos de alta dimensión.
- **Árbol de Clasificación:** Se utilizó un Árbol de Clasificación como punto de referencia. Los árboles de clasificación son fácilmente interpretables y ofrecen una visualización clara de las decisiones tomadas por el modelo. Si su rendimiento es comparable a modelos más complejos, se podría preferir por su simplicidad y transparencia.

Se utilizó el algoritmo de Bosques Aleatorios (Random Forest) para modelar la ocurrencia de fallecimientos en accidentes de tránsito. Las métricas de rendimiento obtenidas para este modelo son:

- **Accuracy (Exactitud):** 99.78%
- **Precision (Precisión):** 15.66%
- **Recall (Sensibilidad):** 9.89%
- **F1-Score:** 12.12%

Estas métricas sugieren que, aunque el modelo tiene una alta exactitud, tiene dificultades para predecir correctamente los casos de tragedia. La precisión y sensibilidad son particularmente bajas, lo que indica que el modelo tiene un alto

número de falsos positivos y falsos negativos. Dado que el F1-Score (una métrica que combina precisión y sensibilidad) es también relativamente bajo, se puede inferir que el modelo podría no ser el óptimo para esta tarea específica.

Uno de los principales desafíos identificados fue el desbalance entre las clases de la variable objetivo: la cantidad de accidentes que resultaron en una tragedia (clase positiva) era significativamente menor en comparación con los que no resultaron en una tragedia (clase negativa). Este desbalance puede llevar a que los modelos tengan un rendimiento excelente en la clase mayoritaria, pero fallen en detectar correctamente la clase minoritaria.

El alto valor de Accuracy (Exactitud) y el bajo F1-Score observados confirmaron este problema. Aunque el modelo pudo clasificar correctamente la mayoría de las instancias, tuvo dificultades para identificar correctamente las tragedias, lo que es esencial para el objetivo del análisis.

Para abordar este desafío, se consideraron varias técnicas, como el undersampling, oversampling y la modificación de la probabilidad de corte. Se optó por ajustar la probabilidad de corte, que implica ajustar el umbral sobre el cual se clasifica una instancia como positiva o negativa. Al modificar este umbral, es posible mejorar la capacidad del modelo para detectar la clase minoritaria sin comprometer demasiado su rendimiento en la clase mayoritaria. Además, se llevó a cabo una optimización de hiperparámetros.

c. Análisis de clúster

1. **Elección del algoritmo:** Se optó por utilizar el algoritmo KMeans para realizar el análisis de clúster debido a su eficiencia en comparación con técnicas como el Agrupamiento Jerárquico, especialmente cuando se trabaja con grandes conjuntos de datos.
2. **Estandarización:** Antes de aplicar el algoritmo, es esencial estandarizar los datos. La estandarización asegura que cada característica contribuye de manera equitativa al proceso de clúster y que ninguna característica domina el proceso debido a su escala.
3. **Determinación del número óptimo de clústeres:** Se utilizó el "Método del codo" para determinar el número óptimo de clústeres. Al graficar las inercias (suma de distancias de cada dato al centroide específico de su clúster) contra diferentes valores de **K** (número de clústeres), se observó que al considerar 11 clústeres, la pendiente empezaba a relajarse, indicando que 11 es un buen número de clústeres para este conjunto de datos.
4. **Aplicación de KMeans:** Se ajustó el algoritmo KMeans utilizando 11 clústeres y se asignaron etiquetas de clúster a cada instancia del conjunto de datos.
5. **Visualización de clústeres usando PCA:** Para visualizar cómo se agrupan los datos en el espacio de características de alta dimensión, se utilizó el Análisis de Componentes

Principales (PCA) para reducir la dimensión de los datos a dos componentes. Estos dos componentes se graficaron para observar cómo se distribuyen los clústeres en el espacio bidimensional.

6. **Interpretación de clústeres:** Se analizaron varios clústeres para entender las características distintivas de cada uno:
    - **Cluster 0:** Principalmente accidentes con motoristas lesionados en las principales avenidas.
    - **Cluster 4:** Similar al cluster 0 pero con mayor gravedad en cuanto a la fatalidad de los accidentes.
    - **Cluster 6:** Accidentes de mayor gravedad ocurridos solo en la avenida principal.
    - **Cluster 8:** Accidentes de gran gravedad que involucran heridos en peatones, ciclistas y motoristas en los tres tipos de avenidas.
    - **Cluster 10:** La tasa de muertes es aún mayor, afectando los tres tipos de avenida y con heridos en peatones, motoristas y ciclistas.

Conclusiones

En el análisis exploratorio de datos, se procesó y preparó el conjunto de datos para modelar, eliminando y reconfigurando características según su relevancia y pertinencia. Esta etapa fue esencial para comprender la naturaleza de los datos y garantizar que las técnicas de modelado subsiguientes fueran efectivas.

En cuanto al aprendizaje supervisado, se exploraron múltiples técnicas de modelado, desde Bosques Aleatorios hasta Máquinas de Vectores de Soporte. A pesar de que algunos modelos, como el Bosque Aleatorio, mostraron alta exactitud, se identificó un problema subyacente de clases desbalanceadas que afectaba la precisión y sensibilidad del modelo. Para abordar este desafío, se consideraron técnicas como la modificación de la probabilidad de corte y la optimización de hiperparámetros.

El análisis de clúster proporcionó una perspectiva adicional, agrupando accidentes en función de sus características subyacentes. Esta agrupación permitió identificar patrones y tendencias comunes entre diferentes tipos de accidentes, ofreciendo una capa adicional de comprensión sobre las condiciones y factores asociados con diversos accidentes.

Bibliografía

Acosta, J. (2023). Análisis de componentes principales (PCA). Semana 8. *Magíster en Ciencias de Datos.* Pontificia Universidad Católica de Chile.

Acosta, J. (2023). Bosques Aleatorios, Naive Bayes y Análisis Discriminante. Semana 5. *Magíster en Ciencias de Datos.* Pontificia Universidad Católica de Chile.

Acosta, J. (2023). Máquinas de vectores de Soporte y Clases desbalanceada. Semana 6. *Magíster en Ciencias de Datos.* Pontificia Universidad Católica de Chile.

Local Government of city of New York. (2023). *Motor Vehicle Collisions – Crashes* [Conjunto de datos]. https://catalog.data.gov/dataset/motor-vehicle-collisions-crashes

Anexo Calificaciones:

NOMBRE	CALIFICACIÓN
Richard Thomas Orellana	100%
Luciano Lorenzo Davico	100%
Mauricio Gallardo	0%