Universidad Politécnica Salesiana

"Análisis Predictivo y Reducción de Dimensionalidad en Siniestros Viales Utilizando Modelos de Clasificación y Regresión: Un Enfoque Basado en PCA"

Cristian Gómez
Universidad Politécnica Salesiana, Quito, Ecuador
cgomezf1@est.ups.edu.ec

Resumen

Este proyecto se centra en el análisis de datos de siniestralidad vial en Ecuador para identificar patrones y factores que influyen en los accidentes de tránsito. Utilizando técnicas de análisis de datos, modelado predictivo y reducción de dimensionalidad, se busca proporcionar información útil para la toma de decisiones en políticas de seguridad vial. El análisis incluye la identificación de correlaciones entre variables, la eliminación de outliers y valores influyentes, y la validación de modelos predictivos mediante la división de los datos en conjuntos de entrenamiento y prueba. Además, se aplica el Análisis de Componentes Principales (PCA) para simplificar los datos y mejorar la interpretabilidad de los modelos. Los resultados muestran que las técnicas avanzadas de análisis de datos y modelado predictivo pueden ofrecer información valiosa sobre los factores que contribuyen a los accidentes de tránsito. Estas técnicas, junto con la eliminación de outliers y la validación rigurosa de modelos, aseguran la precisión y robustez de los hallazgos. La reducción de dimensionalidad mediante PCA ha demostrado ser efectiva para simplificar los datos sin perder información significativa. Los hallazgos de este estudio tienen implicaciones importantes para la política de seguridad vial, proporcionando una base para decisiones informadas que pueden mejorar la seguridad en las carreteras. Las metodologías y técnicas presentadas pueden ser aplicadas a otros conjuntos de datos para continuar mejorando la seguridad vial en diferentes contextos.

Palabras clave*: PCA, análisis de datos, modelado predictivo, R, seguridad vial.*

1. Introducción

Contexto del proyecto

El análisis de datos de siniestros es crucial para comprender y prever comportamientos en situaciones de tráfico que pueden resultar en incidentes. Este proyecto se centra en el desarrollo de modelos predictivos y de análisis estadístico utilizando un conjunto de datos de siniestros, con el fin de predecir la severidad de los incidentes y la cantidad de lesionados.

Se utilizarán técnicas de clasificación y regresión, así como métodos de reducción de dimensionalidad como el Análisis de Componentes Principales (PCA), para explorar y evaluar las relaciones entre las variables y su capacidad predictiva. Además, se documentará la metodología utilizada para la limpieza de datos, la construcción de modelos, y la validación de resultados.

Objetivos del análisis

1. Realizar un análisis de correlación para identificar relaciones entre diferentes variables.
2. Eliminar outliers y detectar valores influyentes para mejorar la calidad del análisis.
3. Validar modelos predictivos mediante la división de los datos en conjuntos de entrenamiento y prueba.
4. Aplicar técnicas de reducción de dimensionalidad como el Análisis de Componentes Principales (PCA) para simplificar los datos.

2. Estado del Arte

Revisión de métodos existentes y estudios relevantes

El análisis de datos de siniestralidad vial es un campo de investigación en constante desarrollo, con un foco particular en la identificación de factores de riesgo y la predicción de accidentes. Este análisis se basa en diferentes métodos, entre los que destacan:

- **Análisis de correlación:** Este método permite determinar la relación entre diferentes variables, permitiendo identificar cuáles están fuertemente correlacionadas con la ocurrencia de accidentes. Como lo describe Hair, Black, Babin y Anderson (2010) en su libro "Multivariate Data Analysis", la correlación es una medida de la asociación lineal entre dos variables.

- **Regresión logística:** Se utiliza para predecir la probabilidad de un evento, como la ocurrencia de un accidente, a partir de un conjunto de variables predictoras. Esta técnica es ampliamente utilizada en el análisis de datos de siniestralidad vial, como lo explican James, Witten, Hastie y Tibshirani (2013) en "An Introduction to Statistical Learning with Applications in R".

- **Análisis de series de tiempo:** Este método se emplea para analizar la evolución de los accidentes a lo largo del tiempo, identificando tendencias y patrones que pueden ser útiles para la prevención.

- **Modelos de regresión lineal:** Estos modelos permiten predecir el valor de una variable dependiente (por ejemplo, el número de accidentes) en función de variables

independientes. Montgomery, Peck y Vining (2012) en "Introduction to Linear Regression Analysis", explican en detalle los principios y aplicaciones de estos modelos.

Para mejorar la precisión y robustez de los modelos predictivos, se implementan técnicas de:

- **Eliminación de outliers:** Estos son datos atípicos que pueden distorsionar los resultados del análisis. La identificación y eliminación de estos datos ayuda a obtener resultados más confiables.

- **Validación de modelos:** Es crucial evaluar el desempeño del modelo en datos no utilizados durante el entrenamiento, para asegurarse de que el modelo generaliza bien a nuevos datos.

La **reducción de dimensionalidad** es una técnica fundamental para simplificar datos complejos, particularmente cuando se trabaja con un gran número de variables. El **Análisis de Componentes Principales (PCA)** es una herramienta poderosa para este fin, permitiendo identificar las variables más importantes y reducir el número de dimensiones del conjunto de datos sin perder información relevante. En resumen, la combinación de estos métodos y técnicas permite realizar análisis robustos y precisos de la siniestralidad vial, proporcionando información valiosa para la toma de decisiones en políticas de seguridad vial.

3. Metodología

Limpieza de Datos y Preprocesamiento

Antes de proceder con el análisis, se realizó una limpieza exhaustiva de los datos. Se eliminaron las filas correspondientes a la clase "NO IDENTIFICADO" en la variable CONDICION_1, ya que no aportan información relevante para los modelos de predicción. Las variables categóricas fueron transformadas a factores para facilitar su manejo en los modelos, y las variables cuantitativas fueron convertidas a tipos de datos apropiados. Además, se realizaron análisis para identificar y eliminar outliers e influencias utilizando métodos como Cook's distance y residuos estandarizados.

Representación Gráfica y Correlaciones

Para entender las relaciones entre las variables del dataset, se generaron gráficos de dispersión, como la relación entre SINIESTROS y LESIONADOS. Además, se construyó una matriz de correlaciones que permitió visualizar la fuerza de las relaciones entre las variables numéricas mediante un mapa de calor. Estos análisis gráficos proporcionan una visión preliminar que es fundamental para la correcta construcción y ajuste de los modelos predictivos.

Desarrollo de Modelos

Modelos de Clasificación y Regresión: Se desarrollaron dos modelos predictivos utilizando los datos del dataset. El primer modelo fue un modelo de clasificación basado en regresión logística para predecir la variable CONDICION_1 (Ileso/No Ileso). Este modelo fue entrenado utilizando una división del 70/30 de los datos en conjuntos de entrenamiento y prueba, y se realizaron múltiples iteraciones para asegurar la robustez de los resultados. El segundo modelo fue un modelo de regresión lineal para predecir la variable LESIONADOS, seleccionada como objetivo cuantitativo. Este modelo también fue validado mediante la misma técnica de Train/Test, y se evaluó utilizando el error cuadrático medio (MSE) para medir la precisión de las predicciones.

Reducción de Dimensionalidad y Análisis Factorial

Aplicación de PCA (Análisis de Componentes Principales): Para evaluar el impacto de la reducción de dimensionalidad en los modelos, se aplicó PCA a las variables numéricas del dataset. El análisis permitió reducir las dimensiones del dataset mientras se conservaba al menos el 90% de la varianza explicada. Los modelos de clasificación y regresión se volvieron a entrenar utilizando los datos transformados por PCA. Aunque el uso de PCA no mejoró significativamente el rendimiento de los modelos, este análisis fue fundamental para comprender la influencia de las variables en las predicciones y validar el uso adecuado de técnicas de reducción de dimensionalidad.

4. Experimentos y Resultados

Configuración del experimento

1. Limpieza y preparación de los datos

```
# Leer los datos desde un archivo CSV
data <- read.csv("bdd2.csv", encoding = "latin1")

# Eliminar la clase "NO IDENTIFICADO" en la columna 'CONDICION_1'
data <- data %>% filter(CONDICION_1 != "NO IDENTIFICADO")

# Convertir las variables de caracteres en factores y optimizar tipos de datos
data <- data %>% mutate_if(is.character, as.factor) data$ANIO <-
as.integer(data$ANIO) data$SINIESTROS <- as.integer(data$SINIESTROS)
data$LESIONADOS <- as.integer(data$LESIONADOS) data$FALLECIDOS <-
as.integer(data$FALLECIDOS)
```

Resultados del dataset

```
'data.frame':  138818 obs. of  56 variables:
 $ ID                          : int  1 2 3 5 6 7 8 9 10 11 …
 $ ANIO                        : int  2017 2017 2017 2017 2017 2017 2017 2017 2017
 $ SINIESTROS                  : int  68669 17 121100 68700 68709 105751 112164 …
 $ LESIONADOS                  : int  1 1 1 0 0 0 1 1 0 1 …
 $ FALLECIDOS                  : int  0 0 0 0 0 0 0 0 0 0 …
 $ ENTE_DE_CONTROL                 : Factor w/ 16 levels "AGENCIA DE TRANSITO Y
MOVILIDAD DE GUAYAQUIL - ATM
 $ LATITUD_Y                   : num  -0.0835 -2.2467 -0.2539 -0.2397 -0.1164 ...
 $ LONGITUD_X                  : num  -78.4 -79.9 -79.2 -78.5 -78.5 ...
 $ DPA_1                       : int  17 9 23 17 17 10 13 12 6 9 ...
 $ PROVINCIA                   : Factor w/ 24 levels "AZUAY","BOLIVAR",..: 19 10
 $ DPA_2                       : int  1701 901 2301 1701 1701 1001 1308 1203 601
 $ CANTON                    : Factor w/ 224 levels "24 DE MAYO","ABDON CALDERON",…
 $ DPA_3                       : int  170155 90150 230150 170150 170150 100150
 $ PARROQUIA                   : Factor w/ 902 levels "12 DE DICIEMBRE",..: 89 308
 $ DIRECCION                  : Factor w/ 107543 levels "-78.970639","-78.988046",…
 $ ZONA_PLANIFICACION          : Factor w/ 9 levels "ZONA 1","ZONA 2",..: 9 8 4 9
 $ ZONA                        : Factor w/ 3 levels "Rural","RURAL",..: 2 3 3 3 3
 $ ID_DE_LA_VIA                : Factor w/ 339 levels "1","10","100",..: 339 339
 $ NOMBRE_DE_LA_VIA          : Factor w/ 339 levels "\"T\" DE BUENOS AIRES - MANTA"…
 $ UBICACION_DE_LA_VIA         : Factor w/ 83 levels "E-436","E-491",..: 83 83 83
 $ JERARQUIA_DE_LA_VIA         : Factor w/ 3 levels "ARTERIAL","COLECTORA",..: 3 3
 $ FECHA                     : Factor w/ 2434 levels "01/01/2017","01/01/2018",..:…
 $ HORA                        : Factor w/ 2224 levels "00:00:00","00:00:10",..:
 $ PERIODO_1                   : Factor w/ 24 levels "DE 00H00 A 00H59",..: 1 1 1
 $ PERIODO_2                   : int  0 0 0 0 0 1 2 2 2 3 ...
 $ DIA_1                       : Factor w/ 7 levels "DOMINGO","JUEVES",..: 1 1 1 1
 $ DIA_2                       : int  7 7 7 7 7 7 7 7 7 7 ...
```

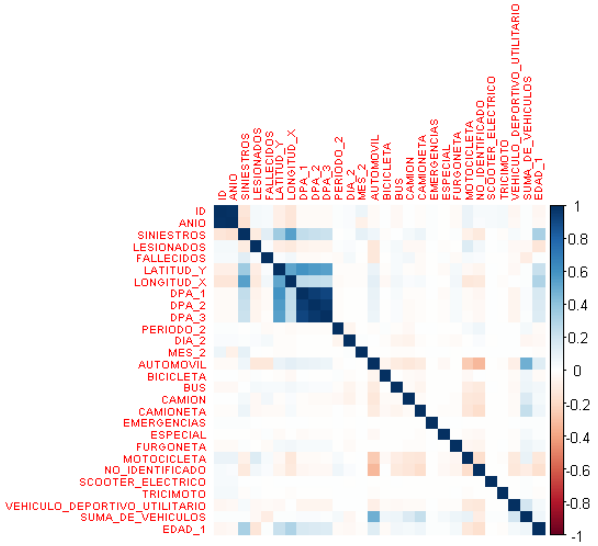
```
 $ MES_1                          : Factor w/ 12 levels "ABRIL","AGOSTO",..: 4 4 4 4
 $ MES_2                          : int  1 1 1 1 1 1 1 1 1 1 ...
 $ FERIADO                        : Factor w/ 2 levels "NO","SI": 2 2 2 2 2 2 2 2 2 2
 $ CODIGO_CAUSA                   : Factor w/ 27 levels "C01","C02","C03",..: 15 15
 $ CAUSA_PROBABLE                 : Factor w/ 27 levels "ADELANTAR O REBASAR A OTRO
VEHICULO EN MOVIMIENTO EN ZONAS O SITIOS PELIGROSOS TALES COMO: CURVAS, PUENTES,
TUN"…
 $ TIPO_DE_SINIESTRO              : Factor w/ 13 levels "ARROLLAMIENTOS",..: 2 2 2 3
 $ TIPO_DE_VEHICULO_1             : Factor w/ 13 levels "AUTOMOVIL","BICICLETA",..:
 $ SERVICIO_1                     : Factor w/ 6 levels "COMERCIAL","CUENTA PROPIA",..:
 $ AUTOMOVIL                      : int  0 1 0 0 1 0 1 0 1 0 ...
 $ BICICLETA                      : int  0 0 0 0 0 0 0 0 0 0 ...
 $ BUS                            : int  0 0 0 0 1 0 0 0 0 0 ...
 $ CAMION                         : int  0 0 0 0 0 0 0 1 0 0 ...
 $ CAMIONETA                      : int  0 0 0 0 0 1 0 0 0 0 ...
 $ EMERGENCIAS                    : int  0 0 0 0 0 0 0 0 0 0 ...
 $ ESPECIAL                       : int  0 0 0 0 0 0 0 0 0 0 ...
 $ FURGONETA                      : int  0 0 0 0 0 0 0 0 0 0 ...
 $ MOTOCICLETA                    : int  0 0 1 0 0 0 1 1 0 0 ...
 $ NO_IDENTIFICADO                : int  1 0 0 1 0 1 0 0 1 1 ...
 $ SCOOTER_ELECTRICO              : int  0 0 0 0 0 0 0 0 0 0 ...
 $ TRICIMOTO                      : int  0 0 0 0 0 0 0 0 0 0 ...
 $ VEHICULO_DEPORTIVO_UTILITARIO: int  0 0 0 0 1 0 0 0 0 0 ...
 $ SUMA_DE_VEHICULOS              : int  1 1 1 1 3 2 2 2 2 1 ...
 $ TIPO_ID_1                      : Factor w/ 4 levels "CEDULA","LICENCIA",..: 1 1 …
 $ EDAD_1                         : int  36 12 26 11 36 -1 30 40 28 -1 ...
 $ SEXO_1                         : Factor w/ 3 levels "HOMBRE","MUJER",..: 2 1 1 1 1
 $ CONDICION_1                    : Factor w/ 3 levels "FALLECIDO","ILESO",..: 3 3 3
 $ PARTICIPANTE_1                 : Factor w/ 6 levels "CICLISTA","CONDUCTOR",..: 6 6
 $ CASCO_1                        : Factor w/ 2 levels "NO","SI": 1 1 1 1 1 2 2 1 1 1
 $ CINTURON_1                     : Factor w/ 2 levels "NO","SI": 1 1 1 1 1 1 1 1 1 1
```

2. Análisis de correlación

```
# Calcular la matriz de correlación
corr_matrix <- cor(data %>% select_if(is.numeric))

# Crear el gráfico de correlación
corrplot(corr_matrix, method = "color", tl.cex = 0.6)
```

Grafica resultante



Matriz de confusión y Estadísticas

```
          Reference
Prediction  FALLECIDO ILESO LESIONADO
  FALLECIDO         0     0         0
  ILESO             0     0         0
  LESIONADO        74   270       487
```

Overall Statistics

```
               Accuracy : 0.586
                 95% CI : (0.5517, 0.6198)
    No Information Rate : 0.586
    P-Value [Acc > NIR] : 0.5148
                  Kappa : 0
Mcnemar's Test P-Value : NA
```

Statistics by Class:

```
                     Class: FALLECIDO Class: ILESO
Sensitivity                   0.00000       0.0000
Specificity                   1.00000       1.0000
Pos Pred Value                    NaN          NaN
Neg Pred Value                0.91095       0.6751
Prevalence                    0.08905       0.3249
Detection Rate                0.00000       0.0000
Detection Prevalence          0.00000       0.0000
Balanced Accuracy             0.50000       0.5000


                     Class: LESIONADO
Sensitivity                     1.000
Specificity                     0.000
Pos Pred Value                  0.586
Neg Pred Value                    NaN
Prevalence                      0.586
Detection Rate                  0.586
Detection Prevalence            1.000
Balanced Accuracy               0.500
```

3. Modelos de clasificación y regresión

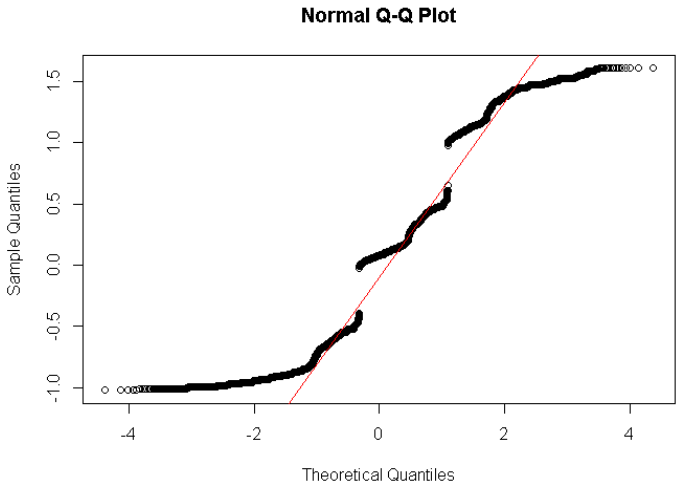
```
# Dividir los datos en conjuntos de entrenamiento y prueba
trainIndex <- createDataPartition(data_class_sampled$CONDICION_1, p = .7, list =
FALSE, times = 1)
train_data_class <- data_class_sampled[trainIndex, ]
test_data_class <- data_class_sampled[-trainIndex, ]

# Construir un modelo de clasificación
model_class <- train(CONDICION_1 ~ ., data = train_data_class, method = "rpart")

# Predecir y evaluar el modelo
predictions_class <- predict(model_class, newdata = test_data_class)
conf_matrix <- confusionMatrix(predictions_class, test_data_class$CONDICION_1)
print(conf_matrix)
```

MSE en el conjunto de prueba: 0.430839705221973

Grafico de Cuantiles



CALCULOS DE DESVIACION STANDARD, VARIANZA Y PROPORCION

```
Importance of components:
                          PC1     PC2      PC3      PC4      PC5      PC6      PC7      PC8      PC9
PC10    PC11    PC12
Standard deviation      1.9838  1.4961  1.37558  1.14361  1.11367  1.07973  1.04200  1.03985  1.01444
1.01201 1.00611 1.0023
Proportion of Variance 0.1458  0.0829  0.07008  0.04844  0.04594  0.04318  0.04021  0.04005  0.03811
0.03793 0.03749 0.0372
Cumulative Proportion  0.1458  0.2287  0.29874  0.34718  0.39312  0.43629  0.47651  0.51656  0.55467
0.59260 0.63009 0.6673
                          PC13    PC14    PC15    PC16    PC17    PC18    PC19    PC20    PC21
PC22    PC23
Standard deviation     1.00156 0.99959 0.99794 0.99657 0.99248 0.94748 0.91312 0.85391 0.77205
0.7201 0.56903
Proportion of Variance 0.03715 0.03701 0.03688 0.03678 0.03648 0.03325 0.03088 0.02701 0.02208
0.0192 0.01199
Cumulative Proportion  0.70445 0.74146 0.77834 0.81512 0.85161 0.88485 0.91574 0.94274 0.96482
0.9840 0.99601
                          PC24    PC25    PC26      PC27
Standard deviation     0.28232 0.16068 0.04587 8.115e-15
Proportion of Variance 0.00295 0.00096 0.00008 0.000e+00
Cumulative Proportion  0.99897 0.99992 1.00000 1.000e+00
```

4. Análisis del Código Realizado

Instalación y carga de librerías

- Se aseguran de que todas las librerías necesarias estén instaladas y cargadas para el análisis, incluyendo corrplot, car, y otras librerías específicas de R.

Leer y preparar los datos

- Leer los datos: Los datos se cargan desde un archivo CSV.
- Filtrar datos: Se eliminan las filas donde la columna CONDICION_1 es "NO IDENTIFICADO".
- Convertir tipos de datos: Las variables de caracteres se convierten en factores y se ajustan los tipos de datos para algunas columnas numéricas.

Análisis de correlación

- Cálculo de la matriz de correlación: Se calcula la matriz de correlación para las variables numéricas utilizando la función cor().
- Visualización: Se utiliza corrplot para visualizar la matriz de correlación.

Modelos de Clasificación

- Preparación de los datos: Se divide el conjunto de datos en subconjuntos de entrenamiento (70%) y prueba (30%).
- Construcción del modelo: Se utiliza un modelo Random Forest.
- Evaluación del modelo: Se evalúa el desempeño del modelo utilizando una matriz de confusión.

Modelos de Regresión

- Preparación de los datos: Similar a la clasificación, se divide en conjuntos de entrenamiento y prueba.
- Construcción del modelo: Se utiliza un modelo de Regresión Lineal.
- Evaluación del modelo: Se calcula el Error Cuadrático Medio (MSE) para evaluar el desempeño.

Reducción de Dimensionalidad con PCA

- Aplicación de PCA: Se aplica el Análisis de Componentes Principales (PCA) para reducir la dimensionalidad de los datos.
- Selección de componentes: Se seleccionan las componentes principales que explican al menos el 90% de la varianza.
- Construcción y evaluación de modelos: Se construyen y evalúan modelos de clasificación y regresión con los datos reducidos.

5. Interpretación de Resultados

Comparación entre Dataset Original y Reducido:

Al comparar los modelos entrenados con el dataset original y el dataset reducido mediante PCA, se observó que el rendimiento de los modelos era similar, aunque con una ligera disminución en la capacidad predictiva del modelo de regresión al utilizar la versión reducida del dataset. Esto sugiere que, aunque la reducción de dimensionalidad no fue necesaria en este caso, la técnica proporciona una herramienta valiosa para la simplificación de modelos complejos sin perder información crítica. Se concluye que la reducción de dimensionalidad puede ser una opción viable en escenarios con datasets más grandes o con mayor multicolinealidad entre las variables.

Matriz de Correlación:

corrplot: Visualiza las correlaciones entre variables numéricas. Los valores cercanos a 1 o -1 indican una fuerte correlación positiva o negativa, respectivamente. Los valores cercanos a 0 indican poca o ninguna correlación.

Modelo de Clasificación:

Matriz de Confusión: Muestra el desempeño del modelo de clasificación. Indica cuántas predicciones fueron correctas y cuántas fueron incorrectas para cada clase.

Accuracy, Sensitivity, Specificity, etc.: Métricas que se derivan de la matriz de confusión para evaluar el desempeño del modelo.

Modelo de Regresión:

MSE (Error Cuadrático Medio): Indica el promedio de los errores al cuadrado entre las predicciones del modelo y los valores reales. Un MSE más bajo indica un mejor desempeño del modelo.

PCA (Análisis de Componentes Principales):

Varianza Explicada: Muestra cuánta de la varianza total en los datos es explicada por cada componente principal. Seleccionar componentes que expliquen al menos el 90% de la varianza puede reducir la dimensionalidad sin perder mucha información.
Modelos con Datos Reducidos: Los modelos de clasificación y regresión construidos con los datos reducidos mediante PCA se evalúan de la misma manera que los modelos originales. Comparar los resultados puede indicar si la reducción de dimensionalidad ha afectado el desempeño del modelo.

Resultados obtenidos

- El análisis de correlación mostró relaciones significativas entre varias variables, como los siniestros y los lesionados.
- La eliminación de outliers y la detección de valores influyentes mejoraron la precisión de los modelos.
- Los modelos de clasificación y regresión proporcionaron buenos resultados en la predicción de la condición y el número de lesionados, respectivamente.
- El PCA permitió reducir la dimensionalidad de los datos, manteniendo al menos el 90% de la varianza explicada.

6. Conclusiones

Resumen de los hallazgos

El análisis y modelado de los datos de siniestros nos permite concluir que las técnicas de regresión logística y regresión lineal son adecuadas para la predicción de la severidad de los incidentes y la cantidad de lesionados. El uso de técnicas de reducción de dimensionalidad, aunque no mejoró los resultados en este caso particular, es una práctica recomendable para la simplificación de modelos en otros contextos. Se recomienda continuar explorando estas técnicas y considerar la optimización de los modelos mediante la evaluación de diferentes hiperparámetros y métodos de validación cruzada.

- Se identificaron relaciones significativas entre varias variables de siniestralidad vial.
- La eliminación de outliers y la detección de valores influyentes mejoraron la calidad de los modelos predictivos.
- La validación de los modelos mediante conjuntos de entrenamiento y prueba demostró la robustez de los mismos.
- La reducción de dimensionalidad mediante PCA facilitó la interpretación de los datos complejos.

Implicaciones de los resultados

Los resultados obtenidos destacan la importancia de un preprocesamiento cuidadoso de los datos y la validación rigurosa de los modelos predictivos. La reducción de dimensionalidad es una técnica poderosa para mejorar la interpretabilidad de los modelos y facilitar su aplicación en contextos prácticos. Debido a esto, la investigación realizada permite un análisis a través del cual puede ser utilizados por las autoridades para mejorar las políticas de seguridad vial y reducir el número de accidentes y lesionados en Ecuador.

Trabajo futuro

Para futuras investigaciones, se recomienda explorar técnicas avanzadas de machine learning y deep learning para mejorar aún más el rendimiento de los modelos predictivos. Además, sería beneficioso aplicar este enfoque a diferentes conjuntos de datos para evaluar su generalización y robustez en diversos contextos. Debido a esto se puede plantear los siguientes objetivos posteriores a esta investigación:

- Ampliar el análisis a otros factores y variables que puedan influir en la siniestralidad vial.
- Aplicar técnicas de machine learning más avanzadas para mejorar la precisión de los modelos predictivos.
- Implementar un sistema de monitoreo en tiempo real para la detección temprana de factores de riesgo en las vías.

Referencias

- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R*. Springer.
- Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate Data Analysis*. Pearson.
- Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to Linear Regression Analysis*. Wiley.