Estimación de la cantidad de heridos en accidentes de tránsito dentro de la provincia de Lima utilizando modelos de regresión lineal múltiple y PCA

Alisson Georgeth Avalos Tapia
ag.avalost@alum.up.edu.pe
https://orcid.org/0000-0003-0985-0290

Juan David Cárdenas Zúñiga
jd.cardenasz@alum.up.edu.pe
https://orcid.org/0000-0002-3801-4227

Rodrigo Fernando Caballero Chocano
rf.caballeroc@alum.up.edu.pe
https://orcid.org/0000-0002-3090-1008

Victor Andres Ayma Quirita
va.aymaq@up.edu.pe
https://orcid.org/0000-0003-2987-2761

Universidad del Pacífico, Perú

Recibido: 6 de agosto del 2022 / Aceptado: 23 de septiembre del 2022
https://doi.org/10.26439/ciis2022.6073

RESUMEN. Existe una alta probabilidad de que un accidente de tránsito deje víctimas, muchas de las cuales podrían encontrarse en un estado crítico que requiera de atención médica inmediata para su sobrevivencia. Sin embargo, en el Perú, donde los recursos de atención en salud en estas emergencias son bastante limitados, se tiene la necesidad de priorizar la atención inmediata de ciertos accidentes sobre otros. Por ello, el presente trabajo tiene como objetivo construir un modelo que estime la cantidad de heridos en accidentes de tránsito dentro la provincia de Lima con base en registros de accidentes ocurridos entre los años 2017 y 2016, según el Censo Nacional de Comisarías del 2017. Para este fin, se desarrollará un análisis de

regresión lineal múltiple tomando ciertas variables seleccionadas a partir de los registros de accidentes; asimismo, se realizará un análisis de componentes principales en busca de representar mejor algunas de las variables dentro del modelo y extraer la información más relevante de estos datos. De esta manera, se pretende brindar información valiosa al personal de salud que atiende estos eventos, de forma que les permita priorizar la atención de aquellos accidentes cuyas condiciones permitan maximizar el número de sobrevivientes sobre el total de heridos.

PALABRAS CLAVE: accidentes de tránsito, número de heridos, modelo predictivo, regresión lineal múltiple, PCA

ESTIMATION OF THE NUMBER OF INJURIES IN TRAFFIC ACCIDENTS WITHIN THE PROVINCE OF LIMA USING MULTIPLE LINEAR REGRESSION MODELS AND PCA

ABSTRACT. There is a high probability that traffic accidents will leave several victims in such a critical condition that they will require immediate medical care to survive. In Peru, where healthcare resources for emergencies are quite limited, it is essential to prioritize some cases over others. This work aims to build a model that estimates the number of people injured in traffic accidents in the province of Lima based on 2016 and 2017 accident records from the Police Station National Census. It develops a multiple linear regression analysis based on variables taken from the accident records and principal component analysis to represent better and simplify some of the original model variables and work with the most relevant information. The model seeks to provide valuable information to medical professionals to prioritize attention and maximize the number of survivors.

KEYWORDS: traffic accidents, people injured, predictive model, multiple linear regression, PCA

1. INTRODUCCIÓN

En el Perú, los accidentes de tránsito han sido definidos como prioridad en investigación debido a su alta frecuencia, su elevado nivel de mortalidad y la gran cantidad de afectados que quedan con lesiones irreversibles en estos eventos (Peden et al., 2004). Según la Defensoría del Pueblo (2021), en los últimos cinco años ocurrieron más de 420 000 accidentes de tránsito, los cuales dejaron cerca de 14 000 muertos y 272 000 personas heridas o discapacitadas. Además, los accidentes de tránsito representan entre el 1,5 % al 2 % del PBI nacional, lo que significa 1000 millones de dólares en costos de atenciones, según la Estrategia Sanitaria Nacional de Accidentes de Tránsito (ESNAT, 2009) del Ministerio de Salud. En este contexto, cabe mencionar que, entre todas las provincias peruanas, Lima es la que concentra la mayor cantidad de accidentes registrados, con cerca del 50 % de los casos sobre el total de accidentes a nivel nacional (Instituto Nacional de Estadística e Informática [INEI], 2018).

Frente a un accidente de tránsito, los principales responsables de acudir ante la emergencia son el Sistema de Atención Móvil de Urgencias (SAMU), el Cuerpo General de Bomberos Voluntarios, la red de atención de salud pública y privada, entre otros. Estos entes desempeñan un papel fundamental, puesto que de su respuesta oportuna depende la vida de las personas accidentadas; además, en un accidente de este tipo, solo cuentan con "una hora de oro" para asistir a las personas accidentadas que presenten heridas graves y aumentar su probabilidad de sobrevivir. Según Maldonado (2016), el 75 % de las posibles defunciones en un accidente podrían suceder si no se presta asistencia inmediata a la persona. Sin embargo, en ocasiones, varios accidentes se registran al mismo tiempo y, tomando en cuenta la cantidad limitada de los recursos de atención, es necesario tomar la difícil decisión sobre cuál de los casos debe recibir una atención prioritaria respecto del otro.

Así pues, este trabajo tiene por objetivo desarrollar un modelo que estime la cantidad posible de heridos en un accidente de tránsito en la provincia de Lima. Este modelo pretende convertirse en una primera fuente de consulta e información capaz de ayudar en la priorización de la atención de accidentes de tránsito simultáneos para optimizar la respuesta de atenciones e intentar maximizar la cantidad de sobrevivientes del total de víctimas en un accidente de este tipo. Para tal fin, se utilizó la información del registro de accidentes de tránsito obtenida del Censo Nacional de Comisarías del 2017 (INEI, 2018), a partir de la cual se armaron diferentes modelos de estimación en busca de aquel que se aproxime mejor a los datos reales. Para el desarrollo de tales modelos se emplearon herramientas como la regresión lineal múltiple y el análisis de componentes principales (PCA).

Este artículo se encuentra organizado de la siguiente manera: la segunda sección presenta una revisión general acerca de los conceptos básicos de los métodos estadísticos y algebraicos utilizados; el conjunto de datos empleado y la metodología son descritos en la tercera sección; los resultados se presentan en la cuarta sección; y, finalmente, en la quinta sección, se discuten las conclusiones y recomendaciones a futuro de este trabajo.

2. MARCO TEÓRICO

2.1 Análisis de componentes principales

El análisis de componentes principales (PCA) es un método estadístico que permite simplificar la complejidad de los espacios de muestreo multidimensional, al mismo tiempo que conserva su información (Amat, 2017). En este sentido, mediante PCA se calculan nuevas variables ortogonales, denominadas *componentes principales*, con el fin de extraer la información más importante de una distribución de datos, además de comprimir y simplificar su tamaño, y analizar la estructura de observaciones y variables (Abdi & Williams, 2010). Puesto que los primeros componentes principales son los que llevan consigo la mayor variabilidad del modelo, cuanto mayor sea el número de componentes, mayor es la pérdida de su valor, dada la menor variabilidad con la que representan el sistema. Bajo este concepto, en general, los últimos componentes son los que representan el ruido del proceso y, por consiguiente, no son considerados (García & Fuente, 2011).

2.2 Regresión lineal múltiple

El análisis de regresión lineal múltiple permite establecer la relación entre una variable dependiente (Y) y un conjunto de variables independientes . En este trabajo, la variable independiente es un vector con observaciones, es decir, ($X_1, X_2, \ldots X_i$). Asimismo, cada una de las variables es un vector con observaciones, en otras palabras, $X_i = [X_{i1}, X_{i2}, \ldots X_{ik}]$.

Para la aplicación de esta técnica, se busca que tanto la variable dependiente como las variables independientes sean continuas. Sin embargo, también es posible utilizar esta técnica cuando se relaciona una variable dependiente continua con un conjunto de variables categóricas; o en el caso de que se relacione una variable dependiente nominal con un conjunto de variables continuas (Montero Granados, 2016).

De acuerdo con Montero Granados (2016), en el modelo de regresión lineal múltiple se asume que más de una variable influye o está correlacionada con el valor de una tercera variable; por esta razón, se espera que los sucesos tengan una forma funcional, como se muestra en la ecuación 1:

$$Y_j = b_0 + b_1 x_{ij} + b_2 x_{2j} + \cdots + b_k x_{kj} + u_j \qquad (1)$$

Donde Y es la variable endógena, X las variables exógenas, u los residuos y b los coeficientes estimados del efecto marginal entre cada X y Y. De acuerdo con el método diseñado por Rubio y Toma (2019), se busca minimizar la suma de cuadrados del error del modelo planteado, función señalada en la ecuación 2:

$$Q(b) = \sum^{n} \quad error_i^2 \tag{2}$$

Donde el error es entendido como la diferencia entre el valor real y el estimado, como se muestra en la ecuación 3:

$$error = Y_{estimado} - Y_{real} \tag{3}$$

Adicionalmente, en busca de minimizar el error, se calculan las derivadas parciales respecto a cada variable de la regresión (ecuación 4):

$$\frac{\partial Q(b)}{\partial b} = 0 \tag{4}$$

Cada derivada parcial ofrece una ecuación, las cuales, en conjunto, forman el sistema de ecuaciones normales presentado en la Figura 1. La solución a este sistema consiste en un vector que tiene por elementos los coeficientes del modelo de regresión lineal múltiple [b_0, b_1, b_2 ... b_k]. Usualmente, se utilizan métodos de solución como el cálculo de la pseudoinversa de la matriz A o la factorización QR. Esta última se emplea en el presente trabajo y se explica con más detalle en el ítem 2.3.

Figura 1

Sistema de ecuaciones $A \times b = B$

$$\begin{bmatrix} n & \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i2} & \cdots & \sum_{i=1}^{n} x_{ik} \\ \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i1}^2 & \sum_{i=1}^{n} x_{i1}x_{i2} & \cdots & \sum_{i=1}^{n} x_{i1}x_{ik} \\ \sum_{i=1}^{n} x_{i2} & \sum_{i=1}^{n} x_{i2}x_{i1} & \sum_{i=1}^{n} x_{i2}^2 & \cdots & \sum_{i=1}^{n} x_{i2}x_{ik} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^{n} x_{ik} & \sum_{i=1}^{n} x_{ik}x_{i1} & \sum_{i=1}^{n} x_{ik}x_{i2} & \cdots & \sum_{i=1}^{n} x_{ik}^2 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ \vdots \\ b_k \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} y_{i1} \\ \sum_{i=1}^{n} y_{i1}x_{i1} \\ \sum_{i=1}^{n} y_{i1}x_{i1} \\ \vdots \\ \sum_{i=1}^{n} y_{i1}x_{ik} \end{bmatrix}$$

Variables categóricas

Dado que la regresión lineal que se quiere generar cuenta con variables categóricas entre las que se requiere analizar, se debe realizar un tratamiento adecuado respecto de la naturaleza de este tipo de información. Estas variables permiten darle una interpretación a las *n* categorías que se tienen en una variable a través de la asignación de *n* – 1 coeficientes, los cuales indican ausencia o presencia de cada categoría. Se considera *n* – 1 porque la ausencia de todas las variables igualmente se traduce en la presencia de otra variable, ya que si estuviera representada como otro coeficiente, se trataría de una colinealidad de variables (Montero Granados, 2016).

2.3 Factorización QR

La factorización QR es un método clásico que permite resolver sistemas sobredeterminados (Vadillo, 2018). Es ampliamente utilizado en algoritmos computacionales para diversos cálculos, como la resolución de ecuaciones o la determinación de valores propios (Lay, 2012). Básicamente, este método permite factorizar una matriz con columnas linealmente independientes (A) en función de una matriz ortonormal (Q) y una matriz diagonal superior (R), tal que A = QR. En el presente trabajo, la factorización QR se emplea para resolver un sistema de ecuaciones lineales de forma más simple.

2.4 Coeficiente de regresión ajustado

El coeficiente de determinación *R* cuadrado ajustado mide la proporción de la variación total de la variable explicada por la línea de regresión estimada (Rubio & Toma, 2019). Esta medida estadística tiene una interpretación sencilla, pues un valor de 0 significa que el modelo explica el 0 % de la variación total, mientras que un valor de 1 significa que el 100 % de la variación de la variable es explicada por el modelo. Para el cálculo de este estadístico, se emplea la fórmula presentada en la ecuación 5.

Se plantea el uso del coeficiente *R* cuadrado ajustado, que presenta una interpretación similar a la del estimador *R* cuadrado, debido a que el coeficiente no ajustado suele sobreestimar la proporción de la variación explicada (Rubio & Toma, 2019). En términos sencillos, el coeficiente determina la cercanía de la línea de regresión a los datos.

$$r^2\ ajustado\ =\ 1 - \frac{(n-1)\sum_i^n\ \ (Yreal_i - Yestimada_i)^2}{(n-k-1)\sum_i^n\ \ (Yreal_i - \bar{Y})^2} \qquad (5)$$

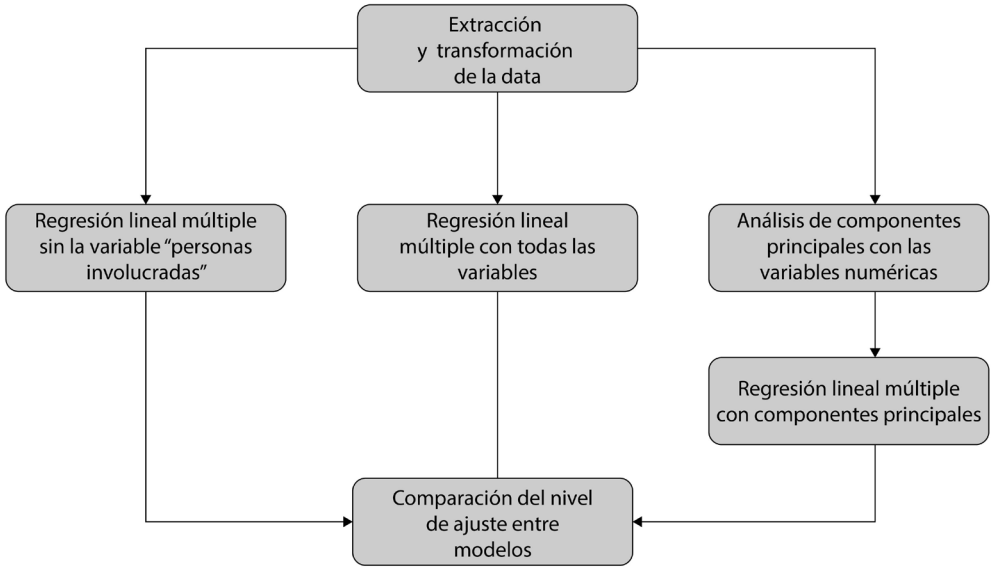
3. METODOLOGÍA

A continuación, se detalla el proceso realizado. Conforme se presenta en la Figura 2, el proceso comienza con la extracción, filtrado y transformación de datos del Censo Nacional de Comisarías del 2017 (INEI, 2018). Posteriormente, se construyeron dos modelos de regresión lineal múltiple con los datos obtenidos. El primero se armó tomando en cuenta todas las variables; para el segundo, dada la gran influencia de la variable "personas involucradas", se propuso un modelo que prescindía de esta última variable. También se generó un análisis de componentes principales entre cuatro variables cuantitativas para luego realizar una nueva regresión lineal múltiple con menos variables.

Finalmente, se compararon ambos modelos para identificar cuál de ellos se aproximaba mejor a la variable dependiente "cantidad de heridos".

Figura 2

Flujograma del procedimiento metodológico



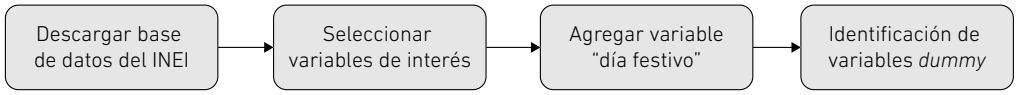
En adelante, la presente sección describirá en detalle cómo se procesaron los datos, así como la aplicación del PCA sobre las variables cuantitativas y también el análisis de regresión múltiple.

3.1 Datasets y preprocesamiento

Para la realización del trabajo, la data fue extraída del Censo Nacional de Comisarías del 2017 (INEI, 2018), de la sección de accidentes de tránsito. Luego se procedió a filtrar todos aquellos registros de accidentes de tránsito en la provincia de Lima. De estos datos, se seleccionaron las variables de interés para trabajar en el modelo, y se agregó la variable "día festivo", que permite identificar si la fecha del accidente corresponde o se acerca a una fecha de ese tipo y analizar su relevancia en el modelo. La inclusión de esta variable se debe a que, por lo general, hay más muertes por accidentes de tránsito en días festivos (NHTSA, 2019). Asimismo, para la inclusión de las variables categóricas, se crearon variables "categóricas". Este proceso se presenta en la Figura 3.

Figura 3

Flujograma de la transformación del dataset



De este modo, se obtuvieron las variables que se incluyen en el modelo de regresión lineal múltiple, como se muestra en la Tabla 1.

Tabla 1

Clasificación de las variables seleccionadas

Variables independientes				
Categoría	Variable	Nombre de la variable	Descripción	Tipo de variable
Fecha	X_1	Hora	Hora del día en que se registró la ocurrencia del accidente.	Cuantitativa
	X_2	Día festivo	Cuantitativa, se asignan probabilidades que indican cuando un día es festivo.	Cuantitativa
Días de la semana (lunes es la referencia)	X_3 - X_8	Martes a domingo	Nombre del día de la semana en que se registró la ocurrencia del accidente.	Categórica
Trimestre (trimestre 1 es la referencia)	X_9 - X_{11}	Trimestre 2 a trimestre 4	Número del trimestre del año en que se registró la ocurrencia del accidente.	Categórica

(*continúa*)

(*continuación*)

Tipo de vía (interprovincial es la referencia)	X_{12} - X_{14}	Urbana, rural, otras	Clasificación del tipo de vía en que se registró la ocurrencia del accidente.	Categórica
Tramo de la vía (intersección es la referencia)	X_{15} - X_{18}	Recta, curva, rotonda, bifurcada	Clasificación del tramo de la vía en que se registró la ocurrencia del accidente.	Categórica
Tipo de accidente (atropello es la referencia)	X_{19} - X_{21}	Otros, colisión, despiste o volcadura	Clasificación del tipo de accidente registrado.	Categórica
Vehículos	X_{22} - X_{24}	Vehículos mayores, vehículos mayores masivos, vehículos menores	Cantidad de vehículos involucrados en el accidente registrado.	Cuantitativa
Personas	X_{25}	Personas involucradas	Cantidad de personas involucradas registradas en el accidente (heridos y muertos).	Cuantitativa
Variable dependiente				
Categoría	Variable	Nombre de la variable	Descripción	Tipo de variable
Personas	Y	Heridos	Cantidad de personas heridas registradas en el accidente.	Cuantitativa

Cabe aclarar que la variable X_2 (día festivo) asigna valores entre 1 y 0 a cada registro de acuerdo con la distancia de la fecha de accidente con la fecha festiva más próxima. Donde 1 corresponde a una fecha festiva, 0,7 a registros con un día de distancia, 0,3 a accidentes con dos días de distancia y 0 para una distancia mayor a dos días.

3.2 Regresión lineal múltiple con todas las variables

Para el desarrollo, implementación y aplicación de la técnica de regresión lineal múltiple, se usó el *software* MATLAB. Primero, se cargaron los datos transformados. Luego, se calcularon la matriz A y el vector B como parte del sistema de ecuaciones normales de la regresión lineal múltiple. Con este fin, se llenaron los datos de forma iterativa, como se indica en la Figura 4.

Figura 4

Pseudocódigo del cálculo de la matriz A y el vector B

```
Crear matriz A y  vector B llenos de ceros

A(1,1) = número de registros
B(1,1) = sumatoria de columna heridos

Para i desde 1 hasta la última columna de los datos:
        A(1,i+1) = suma de elementos de la columna i de los datos

Para j desde 1 hasta la  última columna de los datos:

        A(j+1,1) = suma de elementos de la columna j de los datos
        B(j+1,1) = producto punto de la columna j de los datos y la
                   columna de heridos

        Para i desde 1 hasta la última columna de los datos:

                A(j+1,1) = producto punto de la columna j de los
                   datos y la columna i de los datos
```

Después, se utilizó el método de factorización QR para resolver este sistema de ecuaciones. La factorización se realizó mediante la función predefinida de MATLAB y, posteriormente, se empleó la sustitución hacia atrás del método de eliminación Gauss-Jordan para hallar la solución del sistema (vector con los coeficientes del modelo). La resolución del sistema se desarrolló de esta manera dadas las dimensiones de la matriz R (diagonal superior), las cuales no permitían hallar su inversa debido a limitaciones de la precisión de cálculo del *software*.

Para medir el grado de ajuste del modelo a los datos, se implementó el cálculo del coeficiente de regresión R cuadrado ajustado. Esta medida fue calculada para cada modelo y, luego, se empleó para interpretar los resultados del trabajo.

El modelo de regresión lineal múltiple sin la variable "número de personas" siguió el mismo procedimiento que lo expuesto anteriormente, solo que para este modelo esta variable no fue tomada en consideración.

3.3 Regresión lineal múltiple con PCA

Por otro lado, se realizó un análisis de componentes principales (PCA) mediante la función predefinida en MATLAB sobre las variables cuantitativas X_{22}, X_{23}, X_{24} y X_{25} (relacionadas con la cantidad de vehículos y la cantidad de personas involucradas), por la posibilidad de darse una correlación entre el número de vehículos y el número de personas. El método de PCA

permite, a través de los autovectores (después de hallar los autovalores de las componentes que representan sus respectivas varianzas), reflejar las proyecciones de los datos sobre los respectivos componentes.

Luego de realizar esta transformación, se implementó un modelo de regresión que incluía el resto de las variables que no fueron analizadas mediante PCA, y los componentes principales más representativos de la transformación, los cuales correspondían a los dos primeros componentes principales. Esta regresión se hizo siguiendo los pasos descritos anteriormente: tanto el cálculo de coeficientes como la medida del grado de ajuste.

4. RESULTADOS

En esta sección, se presentan, primero, los coeficientes calculados para cada modelo de estimación desarrollado y su *R* cuadrado ajustado respectivo. Luego, se muestran los resultados obtenidos a partir de la evaluación de tres registros de accidentes para los tres modelos con mejor ajuste.

La Tabla 2 presenta el valor de los coeficientes calculados mediante el programa de MATLAB para cada variable de acuerdo con cada uno de los modelos de estimación desarrollados. Además, muestra el valor del *R* cuadrado ajustado, el cual nos da una idea del nivel de precisión de los modelos.

En la Tabla 2, se puede observar que el modelo que explica mejor esta regresión lineal múltiple es el que se basa en todas las variables, ya que tiene un *R* cuadrado ajustado superior al de los demás modelos. Igualmente, se aprecia que el modelo sin la variable "personas involucradas" es el que realiza de manera menos adecuada las predicciones del número de heridos en los accidentes. Mientras que el *R* cuadrado del modelo de componentes 1 y 2 se asemeja al que se basa en todas las variables, con la consideración de que este modelo optimiza el proceso de alguna manera, ya que usa dos variables menos respecto al modelo inicial, que incluye todas las variables.

Tabla 2

Coeficientes de los distintos modelos elaborados para la estimación de heridos

Variable	Nombre de la variable	Coeficientes	Modelo basado en todas las variables	Modelo sin la variable "personas involucradas"	Modelo con componentes 1 y 2
X_0	Constante	b0	–0,3085	1,1326	1,0953
X_1	Hora	b1	–0,0004	0,0009	–0,0004
X_2	Día festivo	b2	–0,0253	0,0175	–0,0231
X_3	Martes	b3	–0,0049	–0,0424	–0,0145

(*continúa*)

(*continuación*)

X_4	Miércoles	b4	–0,0131	–0,0528	–0,0189
X_5	Jueves	b5	–0,0044	–0,0366	–0,0060
X_6	Viernes	b6	–0,0179	–0,0575	–0,0199
X_7	Sábado	b7	–0,0094	0,0052	–0,0077
X_8	Domingo	b8	–0,0133	0,0543	0,0101
X_9	Trimestre 2	b9	0,0118	0,0186	0,0143
X_{10}	Trimestre 3	b10	0,0484	0,0299	0,0408
X_{11}	Trimestre 4	b11	0,0377	0,0369	0,0371
X_{12}	Vía urbana	b12	0,0007	–0,1326	0,0239
X_{13}	Vía rural	b13	–0,1454	0,0215	–0,0551
X_{14}	Otro tipo de vía	b14	–0,0114	–0,1633	0,0313
X_{15}	Recta	b15	–0,0019	–0,1268	–0,0105
X_{16}	Curva	b16	–0,0074	–0,1506	–0,0407
X_{17}	Rotonda	b17	–0,0556	–0,3323	–0,1051
X_{18}	Bifurcación	b18	0,0261	–0,3268	–0,0245
X_{19}	Otro tipo de accidente	b19	–0,0414	–0,1905	–0,1828
X_{20}	Colisión y choque	b20	–0,2245	–0,3905	–0,4882
X_{21}	Despiste y/o volcadura	b21	0,2533	–0,1351	0,1999
X_{22}	Cantidad de vehículos mayores	b22	–0,5632	–0,0404	
X_{23}	Cantidad de vehículos mayores masivos	b23	–0,4681	0,3366	
X_{24}	Cantidad de vehículos menores	b24	–0,1747	0,4232	
X_{25}	Personas involucradas	b25	0,8592		
X_{26}	Componente principal 1	b26			0,7562
X_{27}	Componente principal 2	b27			–0,4060
R cuadrado ajustado	0,7673	0,0863	0,7375		

Nota. Resultados obtenidos mediante el procesamiento de datos en MATLAB.

La Tabla 3 muestra tres casos de accidentes tomados de la base de datos, junto con los valores de sus variables. En la parte inferior de la tabla, se observa el total de heridos estimados como resultado de la evaluación de cada uno de los tres accidentes en cada uno de los tres modelos de estimación con mejor ajuste seleccionados: modelo basado en todas las variables, modelo sin variable "personas involucradas", y modelo con componentes 1 y 2.

Tabla 3

Resultados de los tres mejores modelos de estimación

Variable	Nombre de la variable	Accidente 1	Accidente 2	Accidente 3
X_0	Constante	1	1	1
X_1	Hora	6	19	22
X_2	Día festivo	0	0	0,7
X_3	Martes	0	0	0
X_4	Miércoles	1	0	0
X_5	Jueves	0	0	0
X_6	Viernes	0	0	0
X_7	Sábado	0	1	0
X_8	Domingo	0	0	0
X_9	Trimestre 2	0	1	0
X_{10}	Trimestre 3	0	0	0
X_{11}	Trimestre 4	0	0	1
X_{12}	Vía urbana	0	1	1
X_{13}	Vía rural	0	0	0
X_{14}	Otro tipo de vía	0	0	0
X_{15}	Recta	1	0	1
X_{16}	Curva	0	0	0
X_{17}	Rotonda	0	0	0
X_{18}	Bifurcación	0	0	0
X_{19}	Otro tipo de accidente	0	0	0
X_{20}	Colisión y choque	0	1	0
X_{21}	Despiste y/o volcadura	1	0	0
X_{22}	Cantidad de vehículos mayores	0	1	2
X_{23}	Cantidad de vehículos mayores masivos	1	1	0

(*continúa*)

(*continuación*)

X_{24}	Cantidad de vehículos menores	0	0	0
X_{25}	Personas involucradas	27	11	6
X26	Componente principal 1	24,21	8,6076	3,8
X27	Componente principal 2	–4,1	–1,1647	0,49
Total de heridos estimado según modelo basado en todas las variables	22,6575	7,8818	3,7296	
Total de heridos estimado según modelo sin variable "personas involucradas"	1,1600	0,9469	0,8618	
Total de heridos estimado según modelo con componentes 1 y 2	21,2346	7,6114	3,7950	
Total de heridos reales	25	10	2	

Nota. Datos de prueba tomados del Censo Nacional de Comisarías del 2017 (INEI, 2018).

5. DISCUSIÓN DE LOS RESULTADOS

5.1 Modelo basado en todas las variables

En la Tabla 2, en el análisis del modelo basado en todas las variables, se puede identificar que la principal variable que explica la cantidad de heridos en un accidente de tránsito es la de "personas involucradas" (X_{25}); así, cuando se aumenta una persona involucrada en el accidente, la cantidad de heridos aumenta en 0,86 personas (manteniendo las demás variables como constantes). Vale notar, además, que la inclusión de esta variable explica el modelo de manera bastante potente (con un *R* cuadrado de 0,77). Esto quiere decir que el modelo explica satisfactoriamente la variación total que presentan los datos. Por ello, se realizó la aplicación del modelo sin esta variable explicativa para ver únicamente la influencia de las demás variables sobre el modelo.

En la Tabla 2, a partir del valor de los coeficientes hallados, se puede apreciar que, contrario a las expectativas, la hora (X_1) parece no influir mucho, al igual que el día festivo (X_2). Por otra parte, el día lunes (variable categórica referencial) es el día que más influye en el modelo (aunque sea mínimo). El trimestre 4 (X_{11}) es el que influye más entre todos los trimestres. La bifurcación (X_{18}) es el tramo de vía que afecta más, seguido por la intersección. Por otro lado, la variable despiste y/o volcadura (X_{21}) presenta el segundo coeficiente más alto en el modelo. Se puede interpretar que es la categoría de tipo de accidentes asociada a un mayor número de heridos.

Finalmente, se introdujeron tres registros de accidentes de tránsito tomados de la base de datos y se obtuvieron los resultados mostrados en la Tabla 3. De estos registros, se obtuvo una predicción bastante cercana a la cantidad de heridos reales. Esto concuerda con los coeficientes de determinación, que indican una cercanía considerable entre las predicciones estimadas por el modelo y los datos reales producidos en dichos accidentes.

5.2 Modelo sin la variable "personas involucradas"

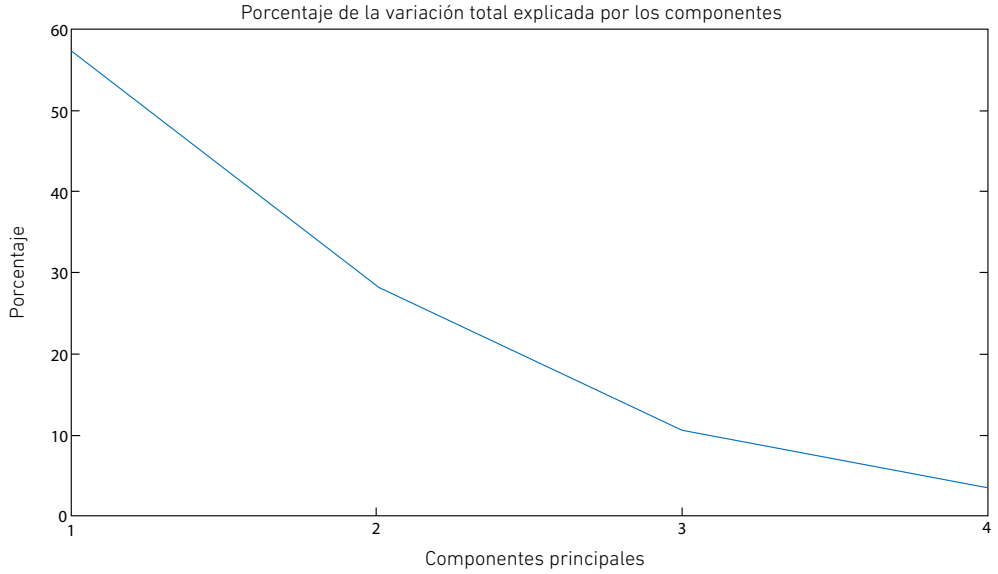
Como se ya se ha mencionado, el primer modelo implementado, que se construye con base en todas las variables, explica bastante bien la cantidad de heridos a partir de la cantidad de personas involucradas (X_{25}). En vista de ello, se plantea una situación en la que no se pueda disponer del número de personas, excluyendo esta variable del modelo. En la Tabla 2 se aprecian los resultados de este nuevo modelo, donde el valor de su métrica *R* cuadrado disminuye a 0,08. Esto nos llevaría a desestimar las predicciones realizadas por este modelo, ya que, como se observa en la Tabla 3, donde se evaluaron registros de accidentes de tránsito extraídos de la base de datos, los resultados obtenidos con este modelo muestran una notoria imprecisión en la estimación de la cantidad de heridos.

5.3 Análisis de componentes principales (PCA)

Para iniciar el análisis de los componentes utilizados respecto de los cuatro atributos originales, en la Figura 5 se presenta la influencia de cada componente. Para ello, se generan cuatro componentes principales que explican la variación de los datos en diferente magnitud. Se puede notar en la Figura 5 que el primer componente principal explica alrededor del 57 % de la variación total del conjunto de datos de las variables tomadas. De manera similar, el segundo componente explica cerca del 30 % de la variación de los datos. Por esta razón, se emplearon ambos componentes para los siguientes modelos de regresión lineal planteados.

Figura 5

Variación explicada por los componentes principales generados



Nota. Resultados obtenidos mediante el procesamiento de datos en MATLAB.

Los modelos planteados en la Tabla 2 logran reducir el número de variables de forma limitada, pues solo aminoran el modelo a 22 variables en el más resumido. De acuerdo con la Tabla 2, el modelo que emplea las dos primeras componentes principales antes mencionadas genera un coeficiente de determinación similar al del modelo más preciso (R cuadrado = 0,74 ≈ 0,77).

Así como en el caso previo, sobre la base de este modelo se evaluaron los mismos tres registros de accidentes de tránsito extraídos de la base de datos y se obtuvieron los resultados presentados en la Tabla 3, que muestran una estimación cercana a la cantidad de heridos reales; sin embargo, es más imprecisa que el modelo con todas las variables. Como en los demás modelos, la precisión de las estimaciones es congruente con los coeficientes de determinación.

6. CONCLUSIONES

Este trabajo presenta un modelo de estimación de la cantidad de heridos que se podrían producir en un accidente de tránsito en la provincia de Lima a través de la implementación de un modelo de regresión lineal múltiple, tomando como base los registros de accidentes ocurridos entre los años 2017 y 2016, de forma que se brinde información importante para el personal de salud que atiende estos eventos, quienes deben priorizar la atención de aquellos accidentes cuyas condiciones permitan maximizar el número de sobrevivientes sobre el total de heridos como producto de la atención oportuna. Así, del análisis desarrollado, se puede concluir que la regresión lineal múltiple permite realizar predicciones adecuadas mediante un modelo con variables explicativas, el cual se aplicó en este trabajo al tratar de predecir el número de heridos mediante las variables propuestas.

Para este fin, se crearon tres distintos modelos de estimación: el primero usó todas las variables independientes; el segundo prescindió de la información sobre el número de personas involucradas en el accidente; y el tercero se desarrolló a través de un análisis de componentes principales de sus variables cuantitativas.

El modelo que mejor explica la cantidad de heridos involucrados en un accidente sería el que se basa en todas las variables. Este hecho se refleja en que la variable independiente "personas involucradas" fue determinante para realizar la estimación de la cantidad de personas heridas de manera adecuada; por esta razón, se recomienda priorizar la atención de accidentes en los que se cuente con el registro de este atributo. Sin embargo, se rescata que, si no se contara con la información del número de personas involucradas, se debería priorizar a los accidentes en función del lugar donde ocurren, como las intersecciones, así como también a aquellos eventos donde se tenga una mayor participación de vehículos menores; aunque vale la pena mencionar que las estimaciones de este modelo (sin el número de personas) poseen una precisión muy inferior con respecto al modelo que incluye la cantidad de personas involucradas.

El análisis de componentes principales permitió reducir de 4 a 2 el número de variables mediante la combinación lineal de otras variables, lo que se toma en cuenta para poder volver más simple un modelo sin perder tanta variabilidad.

Para trabajos futuros, se propone emplear distintos métodos para reducir el número de variables sin pérdida de información relevante, especialmente con referencia al tratamiento de variables categóricas como el análisis de correspondencia. De esta forma se busca generar un modelo más simple, pero considerablemente más preciso.

Asimismo, se puede incluir información relacionada con el número de fallecidos que provoca el accidente como indicador de su gravedad, ya que agregar esta información a los modelos de estimación permitiría lograr una mejor toma de decisiones respecto de la priorización de las atenciones.

REFERENCIAS

Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*(4), 433-459. https://doi.org/10.1002/wics.101

Amat, J. (2017). *Análisis de componentes principales (Principal Component Analysis, PCA) y t-SNE*. Cienciadedatos.net. https://www.cienciadedatos.net/documentos/35_principal_component_analysis

Defensoría del Pueblo. (2021, 20 de diciembre). *Defensoría del Pueblo: más de 14 000 personas fallecieron en accidentes de tránsito en los últimos cinco años*. https://www.defensoria.gob.pe/defensoria-del-pueblo-mas-de-14-000-personas-fallecieron-en-accidentes-de-transito-en-ultimos-cinco-anos/

Estrategia Sanitaria Nacional de Accidentes de Tránsito. (2009). *Accidentes de tránsito: problema de salud pública. Informe nacional*. http://bvs.minsa.gob.pe/local/MINSA/829_MINSA1412.pdf

García, D., & Fuente, M. J. (2011). Estudio comparativo de técnicas de detección de fallos basadas en el análisis de componentes principales (PCA). *Revista Iberoamericana de Automática e Informática Industrial*, *8*(3), 182-195. https://doi.org/10.1016/j.riai.2011.06.006

Instituto Nacional de Estadística e Informática. (2018). Análisis de los accidentes de tránsito ocurridos en el año 2016. En *Perú: VI Censo Nacional de Comisarías 2017. Resultados definitivos* (pp. 123-147). https://www.inei.gob.pe/media/MenuRecursivo/publicaciones_digitales/Est/Lib1528/cap03.pdf

Lay, D. C. (2012). *Álgebra lineal y sus aplicaciones* (4.ª ed.). Pearson Educación.

Maldonado, M. (2016, 25 de noviembre). La hora de oro: 60 minutos que pueden salvar vidas. *Revista Digital INESEM*. https://www.inesem.es/revistadigital/biosanitario/hora-de-oro/

Montero Granados, R. (2016). *Modelos de regresión lineal múltiple. Documentos de trabajo en economía aplicada*. Universidad de Granada. https://www.ugr.es/~montero/matematicas/regresion_lineal.pdf

NHTSA. (2019). *Estimate of motor vehicle traffic crash fatalities for the holiday periods of 2019* [Traffic Safety Facts Research Note. Report N.° DOT HS 812 823]. National Highway Traffic Safety Administration. https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812823

Peden, M., Scurfield, R., Sleet, D., Mohan, D., Hyder, A. A., Jarawan, E., & Mathers, C. (Eds.). (2004). *Informe mundial sobre prevención de los traumatismos causados por el tránsito*. Organización Mundial de la Salud.

Rubio, J., & Toma, J. (2019). *Estadística aplicada. Segunda parte* (2.ª ed.). Universidad del Pacífico.

Vadillo, F. (2018). *Problemas de mínimos cuadrados*. Universidad del País Vasco. https://www.ehu.eus/~mepvaarf/ficheros/minimos_cuadrados.pdf