Facultad de Estadística
Tesis de Grado
Junio 2023

Aplicación de un modelo multinomial logístico con categorías ordinales y binarias para estudiar los homicidios producidos en el Valle del Cauca durante los años 2016 al 2019

Application of a logistic multinomial model with categories ordinal and binary to study homicides produced in Valle del Cauca during the years 2016 to 2019

Autor: Johnny Novoa Acosta[1]
johnnynovoa@usantotomas.edu.co

Director: Dagoberto Bermúdez Rubio[2]
dagobertobermúdez@usantotomas.edu.co

Resumen

El homicidio es un problema que ha sido latente en el Valle del Cauca durante muchas décadas y que ha permeado en gran medida la historia de este departamento; ubicado en el Pacífico colombiano. Los homicidios (o también llamado muertes intencionales o muertes violentas) son hechos que se realizan por muchas razones; que son asociadas al ajuste de cuentas, riñas, hurtos, etc. Esta coyuntura social, al ser tan importante llama la atención de los medios de comunicación nacional y medios a nivel internacional cada año por sus índices y tasas de homicidio, que en ocasiones, suelen ser preocupantes. Por tal motivo, en esta investigación se implementa un modelo multinomial logístico ordinal con distintas categorías para analizar y caracterizar las muertes violentas ocurridas en el departamento del Valle del Cauca, Colombia. Se contempla el periodo de tiempo comprendido entre los años 2016 y 2019. A partir del nivel de escolaridad (analfabeta, primaria, secundaria, técnico/ tecnólogo y superior) alcanzado por las víctimas; se propone modelar e identificar los factores asociados a estas muertes y comprender su influencia e incidencia en la región. Se recopilan datos de fuentes oficiales y se consideran variables relevantes como características demográficas, socioeconómicas y geográficas de los municipios del Valle del Cauca. Estos datos se utilizan para desarrollar y ajustar el modelo, el cual permite examinar las relaciones entre las variables predictoras. El modelo multinomial logístico ordinal utiliza técnicas de estadística inferencial clásica y modelos lineales generalizados para estimar los parámetros del modelo y proporcionar medidas de incertidumbre para estas estimaciones. El análisis de los resultados contribuye a una mejor comprensión de los factores asociados a las muertes violentas en la región y proporciona información valiosa para el diseño de políticas sociales y estrategias de prevención y control de este delito en el Valle del Cauca.

***Palabras clave*:** . Homicidios, muertes violentas, nivel de escolaridad, modelo multinomial logístico, modelos lineales generalizados.

Abstract

Homicide is a problem that has been latent in Valle del Cauca for many decades and has greatly permeated the history of this department; Located in the Colombian Pacific. Homicides (or also called intentional deaths or violent deaths) are events that are carried out for many reasons; that are associated with the settling of accounts, quarrels, theft, etc. This social situation, being so important, draws the attention of

[1]Estudiante Universidad Santo Tomás
[2]Estudiante Universidad Santo Tomás

the national and international media every year due to its homicide rates and rates, which are sometimes worrying. For this reason, in this research an ordinal logistic multinomial model with different categories is implemented to analyze and characterize the violent deaths that occurred in the department of Valle del Cauca, Colombia, during the short period between 2016 and 2019. of schooling (illiterate, primary, secondary, technical/technologist and higher education) reached by the victims; It is proposed to model and identify the factors associated with these deaths and understand their influence and incidence in the region. Data from official sources are collected and relevant variables such as demographic, socioeconomic and geographical characteristics of the municipalities of Valle del Cauca are considered. These data areused to develop and fit the model, which allows one to examine the relationships between the predictor variables. The ordinal logistic multinomial model uses classical inferential statistical techniques and generalized linear models to estimate model parameters and provide uncertainty measures. for these estimates. The analysis of the results contributes to a better understanding of the factors associated with violent deaths in the region and provides valuable information for the design of social policies and crime prevention and control strategies in Valle del Cauca.

***Keywords*:** . Homicides, violent deaths, education level, logistic multinomial model, generalized lineal model.

1. Introducción

El homicidio en Colombia durante los años 2016 al 2019 representó un desafío significativo en términos de seguridad y violencia. Durante este periodo, el país enfrentó altas tasas de homicidio que demandaron una atención urgente por parte de las autoridades y la sociedad en general; incluso a pesar de que la cantidad y la tasa de homicidios bajaron después del proceso de paz con las guerrilas de las FARC-EP (Reportes de Medicina Legal, 2016-2017). El análisis de este periodo específico permite comprender mejor las tendencias y los factores asociados con el homicidio en Colombia. Según el informe "Global Study on Homicide 2019"de la Oficina de las Naciones Unidas contra la Droga y el Delito (UNODC), Colombia continuo siendo uno de los países con altas tasas de homicidio por muchas razones; durante estos años, evidenciando la persistencia del problema (UNODC, 2019). Estos datos reafirman la importancia de estudiar y comprender en detalle las circunstancias que rodearon los homicidios y las muertes violentas en Colombia y especialmente en el departamento del Valle del Cauca; que se ubica en el occidente del país, ya que durante estos cuatro años, algunos municipios y ciudades de este departamento aparecen en el top de las 50 ciudades más violentas del mundo, por sus altas tasas de homicidios por cada 100.000 habitantes; estas ciudades son Santiago de Cali, Palmira y Buenaventura (Ciudadano para la Seguridad Pública y la Justicia Penal 2016-2021).

La implementación de modelos estadísticos avanzados desempeña un papel fundamental en el análisis de fenómenos complejos, como las muertes violentas. En el contexto del Valle del Cauca, Colombia, durante los años 2016, 2017, 2018 y 2019, las tasas de muertes violentas han sido motivo de preocupación tanto para la comunidad académica como para las autoridades locales de cada municipio del departamento. En este sentido, la aplicación de un modelo logístico multinomial ordinal con diferentes categorías puede ofrecer un enfoque real y efectivo para comprender y clasificar las muertes violentas en función de factores demográficos, socioeconómicos y geográficos.

El modelo logístico multinomial ordinal es una extensión del modelo logístico binomial y permite analizar múltiples categorías de un resultado categórico (Alan Agresti 2012). En el contexto de las muertes violentas, se busca caracterizar el homicidio ejecutado apartir del nivel de escolaridad (analfabeta, primaria, secundaria, técnico/ tecnólogo y superior) que tenia la víctima; en función de variables explicativas como el grupo de edad, el género, zona, clase de empleado, la ubicación geográfica entre otras.

La aplicación de un modelo logístico multinomial en el estudio de las muertes violentas en el Valle del Cauca ha sido respaldada por investigaciones previas. Por ejemplo, en un estudio realizado por Rodríguez en el año 2019, se utilizó un enfoque similar para analizar los factores asociados a los diferentes tipos de muertes violentas en esta región. Los resultados revelaron la importancia de variables como la edad, el

género y la ubicación geográfica en la clasificación de las muertes violentas. En este sentido; es importante señalar que la información de los homicidios fueron obtenidos a través de la fuente del portal web Datos Abiertos del Gobierno Nacional; los cuales son bases de datos consolidadas y corroboradas por el Instituto Nacional de Medicína Legal y Ciencias Forenses en colaboración con las entidades como la Gobernación del departamento del Valle del Cauca, Alcaldía de Santiago de Cali y las estadísticas delictivas de la Policía Nacional de Colombia.

Este documento de investigación se encuentra dividido en las siguientes 11 secciones: en la sección 1, para iniciar, se hace una breve introducción del documento de investigación. En las secciónes 2 y 3 se realiza el plantamiento y la pregunta problema de la investigació. La sección 4 se plantean los objetivos propuestos tanto el general como los específicos. En las secciónes 5 y 6 las cuales están integradas por la justificación e hipótesis que contextualiza acerca de lo que se cree y el por qué se realiza esta investigación. En la sección 7 se exponen los antecedentes que giran alrededor de los estudios estadísticos que se han realizado entorno a esta problemática y que se usaron para orientar y dar objetividad a la investigación. En la sección 8, se encuentra el marco conceptual que nos presenta los conceptos definidos para dar una mayor contextualización. La sección 9 en el que se encuentra el marco teórico, se muestra un método de clasificación de las variables más importantes y la especificación del modelo logístico multinomial ordinal con con categorías binarias y multiclase junto con su componente matemático y estadístico, en la sección 10, se presentan los resultados de la aplicación del método de clasificación y el modelo logístico multinomial ordinal; y por último en la sección 11, se presentan las conclusiones que se obtuvieron en este documento de investigación.

2. Plateamiento del Problema

El departamento del Valle del Cauca es el tercer departamento más poblado de Colombia, con una población aproximada de 4.622.450 personas para el 2023 (DANE-Demografía y población); este departamento ubicado en la Región Pacífica colombiana siempre ha tenido, por muchos años, un preocupante problema con los homicidios y muertes violentas; y esto no ha sido diferente durante los años 2016, 2017, 2018 y 2019. Esta situación ha generado una profunda preocupación en la sociedad en general y plantea un desafío para las autoridades y los investigadores en términos de comprender y abordar eficazmente este problema. El nivel de escolaridad es un factor que probablemente tenga importancia y que influye en diferentes aspectos de la vida de las personas, incluyendo su desarrollo socioeconómico, bienestar y calidad de vida.

En el contexto del Valle del Cauca, es necesario analizar cómo el nivel de escolaridad se relaciona con los homicidios ocurridos durante los años 2016 al 2019, con el fin de comprender su impacto en este fenómeno y desarrollar estrategias efectivas de prevención. A pesar de los esfuerzos realizados para reducir la incidencia de los homicidios en esta región colombiana, persisten los altos niveles de violencia. Se ha observado que existen disparidades en la distribución de los homicidios en relación con el nivel de escolaridad de las personas involucradas, lo que sugiere la existencia de posibles vínculos entre la educación y la comisión de actos violentos.

Sin embargo, hasta el momento, la investigación sobre la relación entre el nivel de escolaridad y los homicidios que se cometen en el Valle del Cauca ha sido bastante limitada. Por esto, se requiere un análisis más profundo que permita identificar cómo el nivel de escolaridad influye en la propensión de las personas a ser víctimas del homicidio; así como en las circunstancias y dinámicas asociadas a estos incidentes. El objetivo principal de este estudio es analizar la relación entre el nivel de escolaridad y los homicidios en el Valle del Cauca durante los años 2016 al 2019 a través de un modelo logístico multinomial ordinal con múltiples variables, con el fin de determinar si existe una asociación significativa y comprender los mecanismos subyacentes. Se pretende investigar si el nivel de escolaridad actúa como un factor protector o de riesgo en la comisión de homicidios, y si existen diferencias en las características de los homicidios según el nivel de escolaridad de las personas involucradas. Los hallazgos de este estudio podrían contribuir al diseño e implementación de intervenciones específicas y basadas en evidencia,

dirigidas a prevenir y reducir los homicidios en el Valle del Cauca, centrándose en la mejora del acceso y la calidad de la educación en este departamento.

3. Pregunta Problema

¿Cómo se relaciona el nivel de escolaridad de la víctima del homicidio y sus múltiples características sociales y demográficas con la incidencia de los homicidios en el Valle del Cauca durante el periodo de tiempo comprendido entre los años 2016 al 2019 ?

4. Objetivos

4.1. Objetivo general

Analizar la relación que existe entre el nivel de escolaridad de una persona víctima de un homicidio en el Valle del Cauca durante los años 2016 al 2019 con sus características sociales y demográficas.

4.2. Objetivos específicos

- Identificar la importancia que tiene cada una de las variables explicativas para que el modelo tenga un mayor alcance en la variable respuesta.
- Aplicar los métodos estadísticos prácticos y teóricos del modelo logístico multinomial ordinal con variables ordinales, multiclase y binarias con respuesta a una variable categórica como lo es el nivel de escolaridad.
- Indicar y analizar los resultados conseguidos a través de los métodos estadísticas para encontrar posibles alternativas al pro de la mejora de esta problemática

5. Justificación

La realización de este proyecto tiene como justificación múltiples razones; una de ellas apunta a encontrar una relación objetiva entre las variables de los homicidios y el nivel de escolaridad que hayan alzado las víctimas e homicidio; para tal fin se tiene herramientas y métodos estadístico-matemáticos como lo es el modelo logístico multinomial ordinal para distintas variables categorías; así como también herramientas de programación para así poder demostrar que existen posibles patrones y asociaciones entre las variables en cuestión. Así mismo, no se encuentran muchos artículos relacionados con este tipo de componentes; en caso estrictamente colombiano. Otra razón para justificar este proyecto es que se puede generar un conocimiento científico en el campo de la criminología; y que los resultados que se puedan obtener a través de esta investigación podrán ser utilizados para orientar el diseño y la implementación de políticas públicas y programas de prevención de homicidios en el Valle del Cauca. Este estudio puede arrojar información actualizada y puntual sobre la relación entre el nivel de escolaridad y los homicidios en este departamento, enriqueciendo la literatura y estudios investigativos existentes y fomentando próximas investigaciones venideras en el área.

6. Hipótesis

La metodología estadística del modelo logístico multinomial ordinal es la adecuada para estudiar los homicidios ocurridos en el Valle del Cauca durante el periodo comprendido entre los años 2016 al 2019; y que la variable dependiente; nivel de escolaridad es importante a la hora de abordar un caso de homicidio; ya que a priori se infiere que una persona que tenga un menor nivel de escolaridad y cumplan con ciertas características socieconómicas y demográficas tiene más probabilidades de que sea víctima de una muerte intencional u homicidio.

7. Antecedentes

El análisis de los homicidios mediante el uso de modelos logísticos multinomiales con distintas categórias ha sido objeto de investigación en diversos contextos. Algunos de los tantos antecedentes relevantes que han abordado este enfoque en relación con los homicidios es un estudio hecho en Colombia realizado por Rincón en el 2020, llamado: "Modelos logísticos multinomiales y características socioeconómicas asociadas a los homicidios en Colombia", se implementó un modelo logístico multinomial para analizar los homicidios en Colombia. El objetivo era clasificar los homicidios en diferentes categorías y evaluar los factores asociados a cada una de ellas. Se encontró que las variables como el género, la edad, el nivel socioeconómico y la ubicación geográfica eran determinantes en la clasificación de los homicidios. Otro estudio de aplicación de modelo logístico multinomial en homicidios se hizo en México realizado por Cruz en el 2019, titulado: "Estudio multinomial de los homicidios en México", se utilizó un modelo logístico multinomial para analizar los homicidios en México. El estudio clasificó los homicidios en diferentes categorías y evaluó los factores socioeconómicos, demográficos y geográficos asociados a cada una de ellas. Se encontró que variables como la edad, el género, el nivel de educación y la densidad poblacional tenían una influencia significativa e importante en la clasificación de los homicidios.

Otros estudios más internacionales y con influencia global donde aplicaron esta técnica estadística fue uno que se realizó en China hecho por Li, F., Liu, T., y Huang, L. en el 2017, titulado: "Aplicación de regresión logística multinomial para investigar factores asociados con diferentes tipos de homicidio en China". Este estudio aplicó un modelo logístico multinomial para analizar los factores asociados a diferentes tipos de homicidios en China. Se utilizaron variables socioeconómicas, demográficas y de comportamiento delictivo para predecir la probabilidad de diferentes categorías de homicidio. Existen dos estudios de origen anglosajón que demuestran la aplicabilidad de esta metodología estadística; uno de ellos realizado en California, Estados Unidos y que fue realizado por Wintemute, G. J., Parham, C. A., Beaumont, J. J., Wright, M., y Drake, C. en el año 2019, llamado: "Epidemiología y aspectos clínicos de la violencia homicida con armas de fuego en California". Este estudio examinó la epidemiología y los aspectos clínicos de la violencia homicida y muerte intencional con armas de fuego en estado de California. Se utilizó un enfoque de modelo logístico multinomial para identificar los factores de riesgo asociados con diferentes tipos de homicidios con armas de fuego. Por último, es un estudio realizado en el Reino Unido hecho por Hu, Z., Gray, R. también el año 2019, nombrado: "Un modelo de regresión logística multinomial de homicidios domésticos en Inglaterra y Gales". En aquel estudio se utilizó también un modelo logístico multinomial para investigar los factores asociados con los homicidios domésticos en Inglaterra y Gales. Se exploraron variables relacionadas con la relación entre la víctima y el agresor en cuestión, antecedentes de violencia doméstica y factores socioeconómicos de los mismos.

Estos antecedentes demuestran la aplicabilidad, credibilidad y la relevancia del modelo logístico multinomial ordinal con distintas variables categóricas en el análisis de los homicidios. Permiten comprender las características y los factores asociados a diferentes categorías de homicidio, lo que es fundamental para el diseño de estrategias de prevención y respuesta más efectivas en el contexto de la violencia homicida.

8. Marco conceptual

8.1. Homicidio

El homicidio es el acto intencional de causar la muerte a otra persona. Se refiere a la acción voluntaria de una persona de privar de la vida a otro ser humano, con la intención de causarle la muerte. El homicidio puede ser cometido de diversas formas y maneras, como el uso de violencia física desmedida, armas de fuego de cualquier tipo u otros medios.

El homicidio es considerado un delito en la mayoría de los sistemas legales de muchos países en el mundo y está sujeto a muchas sanciones penales. La gravedad que conlleva el homicidio puede variar dependiendo de algunos factores como la intencionalidad, la premeditación, la motivación y las circunstancias específicas en las que ocurrió el hecho.

Es importante resaltar que el homicidio es un fenómeno bastante complejo y a la vez es multidimensional, porque en muchos casos puede estar influenciado por diversos factores sociales, económicos, culturales y políticos. Su estudio y comprensión son fundamentales para el desarrollo de políticas y acciones encaminadas a prevenir y reducir la violencia homicida en las sociedades que sufren a gran escala este flagelo.

El sistema penal y el Código Penal colombiano, define al homicidio como "la muerte causada a una persona por otra, es decir, cuando una persona mata a otra de manera voluntaria"(Congreso de la República de Colombia, 2000, Artículo 103). De igual manera el informe anual de homicidios en Colombia del Instituto Nacional de Medicina Legal y Ciencias Forenses señala que el homicidio es "la acción intencional de causar la muerte a otra persona, ya sea mediante el uso de violencia física, armas de fuego u otros medios como los objetos contundentes"(Instituto Nacional de Medicina Legal y Ciencias Forenses, 2020, p. 6).

8.2. Educación y niveles escolares en Colombia

La educación es aquel proceso que se encarga del aprendizaje mediante el cual todas las personas adquieren conocimientos, habilidades, valores y actitudes que les permiten desarrollarse personal, social y profesionalmente. Es un proceso social para adquirir y compartir conocimiento, fomentar el pensamiento crítico y promover el crecimiento individual y colectivo de cada persona. La educación es un factor determinante y muy importante en la vida de cualquier persona ya que posiblemente influirá en las decisiones y acciones que haga en su vida individual a futuro; además de ser un derecho en la mayoría de las naciones en el mundo; es un proceso que facilita a las personas a su desarrollo personal y ético.

Cada país en el mundo tiene su propio sistema educativo, que se define como una estructura de enseñanza integrada con niveles de conocimientos que se van aprendiendo con la edad del estudiante y las habilidades que tenga, y que a la vez está diseñado para niñas y niños de primera infancia; hasta para las personas que deseen tener conocimientos avanzados en un área específica del conocimiento; que en este caso es la educación superior. El Ministerio de Educación Nacional de Colombia estructura la educación en los siguientes cinco niveles:

8.2.1. Preescolar

Este nivel comprende tres etapas los cuales son prejardín, jardín y preescolar. En este nivel se encuentran niños y niñas en promedio de 2 a 6 años de edad; donde se fortalecen aspectos biológicos, cognitivos, sicomotrices y socioafectivos.

8.2.2. Educación básica primaria

Este nivel está conformado por grados primero, segundo, tercero, cuarto y quinto de primaria; en promedio se encuentran estudiantes entre 6 hasta 11 años de edad. En este nivel principalmente se busca desarrollar habilidades comunicativas, análisis matemático básico, lectura, escritura y valores éticos fundamentales.

8.2.3. Educación básica secundaria

Este nivel comprende los grados sexto, séptimo, octavo y noveno; donde los estudiantes suelen tener entre 11 a 14 años de edad; en este nivel se busca mejorar habilidades de lógica, ciencias sociales y naturales, así como un pensamiento crítico hacia las situaciones y cosas que los rodean.

8.2.4. Educación media

Los grados décimo y once conforman este nivel, normalmente están estudiantes entre las edades 14 a 18 años de edad. Se busca que los estudiantes comprendan ideas y valores universales, que desarrollen habilidades técnicas; y en ocasiones este nivel suele ser una preparación y una orientación vocacional para el trabajo o la educación superior.

8.2.5. Educación superior

Este nivel también comprende varias etapas, como lo es la educación técnica y tecnológica, profesional o pregrado y posgrados que abarcan las maestrías, especializaciones, doctorados y posdoctorados; no existe un rango de edad para este nivel educativo; en este se busca desarrollar una habilidad específica disciplinar e investigativa en un área del conocimiento.

8.3. Nivel de escolaridad agrupada para el Valle del Cauca

El agrupamiento de escolaridad se define de acuerdo a como están categorizados los datos de la variable "*escolaridad*" en la base de datos. Esta variable que es la dependiente y objeto de estudio puede tomar cinco posibles nominaciones y que, a su vez, es la última escolaridad cursada por la víctima. Estas categorías de escolaridad son:

8.3.1. Primaria

Esta categoría comprende tal cual como está en el Sistema Educativo cColombiano, en esta categoría están los grados primero, segundo, tercero, cuarto y quinto.

8.3.2. Secundaria

En esta categoría está la educación básica secundaria y la educación media; esta categoría se encuentra conformado por los grados sexto, séptimo, octavo, noveno, décimo y once.

8.3.3. Técnica/ Tecnológica

Esta categoría se separa de la educación superior para formar una sola denominación independiente, para el caso del Valle del Cauca.

8.3.4. Superior

Este se entiende para víctimas que tenían una escolaridad superior o igual al pregrado profesional (Especialización, maestría, doctorado y posdoctorado).

8.3.5. Analfabeta

Aunque esta categorización educativa no está contemplada en el sistema educativo de Colombia, se infiere que esta persona careció por múltiples razones una educación básica; y analfabeta es el adjetivo que se le da una persona que no sabe leer ni escribir.

9. Marco teórico

9.1. Eliminación recursiva de características (RFE)

El algoritmo de eliminación recursiva o RFE (Recursive Feature Elimination, RFE) es una técnica utilizada en aprendizaje automático y selección de características principales. Su objetivo es seleccionar un subconjunto óptimo y preciso de características de un conjunto de datos para mejorar el rendimiento de un modelo en particular, en este caso, el logístico multinomial ordinal.

El algoritmo de eliminación recursiva RFE se basa en la idea de que en la totalidad de un conjunto de características contiene algunas características mucho más relevantes que otras para algún problema en especícifco. Comienza con un conjunto completo de características y, en cada iteración, entrena un modelo utilizando todas las características, evalúa la importancia de cada una de ellas y elimina las menos relevantes. También se complementa con el algoritmo SVMs (support-vector machines) que son vectores de peso para las dimensiones categóricas; junto con el apoyo de este algoritmo, se repite el proceso con el conjunto reducido de características hasta que se alcanza el número deseado y significativo de características o se cumple algún otro criterio de terminación en específico (Xue-wen Chen, Jong Cheol Jeong, 2007).

El método para separar linealmente; empieza con funciń discriminante $g(x_i) = w \cdot xi + b$, donde b es n caracter de sesgo, w es vector de peso y sus datos de prueba es x_i, donde $x_i \in \mathbb{R}, i = 1...m$, para otro caso que no sea separable se implementa una variable holgura ξi, cuyo coeficiente es cero y su propósito es medir la distancia de los puntos en un hiperplano, se tiene la siguiente ecuación

$$\Phi(w, \xi) = \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n} \xi i \tag{1}$$

$$Sujeto : y_i(w \cdot x_i + b) \geq 1 - \xi_i, \xi_i \geq 0 \tag{2}$$

Donde y_i significa la finalidad $y_i = (\pm 1)$, $i = 1, ..., m$.

La optimización que requiere en este problema de elección de características, es un problema dual dado a continuación:

$$W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y_i y_j \alpha_i \alpha_j (x_i \cdot x_i) \tag{3}$$

$$Sujeto : 0 \leq \alpha_i \leq C, i = 1, ...m \tag{4}$$

$$\sum_{i=1}^{m} \alpha_i y_j \tag{5}$$

Donde α_i son los coeficientes de Lagrange. El algoritmo de eliminación recursiva mantiene las características de menos importancia en la base relacionada; adquiriendoles un peso menor w, se calcula el vector de soporte o de peso w, llamados SVMs (support-vector machines) lineales de la siguiente manera:

$$\sum_{i=1}^{m} \alpha_i y_j \tag{6}$$

Para SVMs no lineales se calcula con la siguiente ecuación:

$$w_i = \frac{1}{2}\alpha^T K \alpha - \frac{1}{2}\alpha^T K(-i)\alpha \tag{7}$$

Donde $K(-i)$ es la matriz de Kernel evaluada para omitir la i-écima característica en cuestión en la entrada de la base datos x. El algoritmo RFE omite recursivamente la mayoría de las funciones en cada paso y de nuevo clasifica las demás funciones al volver a entrenar las SVMs en función de estas funciones restantes. Si una característica es débil o no tiene suficiente información que tenga un peso medianamente importante en la base, RFE rápidamenta la omitirá. Aunque puede que exista una característica de x que sea poco importante, esta msima puede adquirir un grado de importancia si y solo si se mezcla con una cantidad de características más importantes.

9.2. Modelos lineales de regresión

En estadística y en análisis de regresión un modelo lineal son tipos de modelos que se usa para predecir, determinar y describir la posibles relaciones que existen entre variables de un estudio en particular. Estos modelos se basan en una especulación previa de que existe algún tipo de ralación lineal entre las variables de entrada y la variable salida (Francesc Carmona 2003).

La variable salida que también es llamada variable dependiente o respuesta, se debe estimar de tal forma de que exista una combinación lineal entre las variables independientes o también llamadas predictoras, que están ponderadas por coeficientes. La forma más general de encontrar un modelo lineal es de la siguiente forma:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon \tag{8}$$

Donde:

- y es la variable dependiente o respuesta que se desea pronosticar o predecir. $x_1, x_2, ..., x_p$ son las variables independientes o predictoras del modelo.
- $\beta_1, \beta_2, ..., \beta_p$ son los coeficientes que contienen la contribución de cada una de las variables de entrada.
- ε es el error, que es el que captura la variabilidad de cada uno y que no se puede explicar en el modelo de regresión.

El finalidad principal de un modelo lineal es estimar los coeficientes $\beta_1, \beta_2, ..., \beta_p$ para que de alguna manera se ajusten mejor a los datos de entrenamiento, minimizando así los posibles errores que hay entre las predicciones del modelo y los valores que son reales de la variable respuesta o de salida. Esto se logra utilizando métodos como la regresión lineal ordinaria, la regresión lineal de mínimos cuadrados o técnicas de optimización.

9.3. Modelos lineales genealizados

Los modelos lineales generlizados o por sus siglas en inglés GLM (generalized linear model) son una generalización de los modelos lineales clásicos, estos modelos lineales se acoplan a la variable dependiente con la implementación de una función de enlace; estos se pueden aplicar a las distribuciones de la familia exponencial, tanto continuas como discretas, como las distribuciones Poisson, gamma, binomial, Bernulli, beta, Weibull, entre otras. Estos modelos se crearon con un proceso iterativo de mínimos cuadrados ponderados para obtener la estimación de máxima verosimilitud de los parámetros del modelo en evaluación (Javier Morales 2001).

A partir de lo anterior, se define Y_i como la función de densidad de una distribución de probabilidad de la familia exponencial y viene dada por:

$$f(y_i;\theta_i,\phi) = exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i,\phi)\right\} \tag{9}$$

Donde $a_i, b(), c()$ son funciones conocidas.

Aparte de la distribución de y, existe el predictor lineal que consiste en la incorporación de la información acerca de las variables predictoras a dicho modelo; y se representa así:

$$\eta = X\beta \tag{10}$$

Para que la media de la función de densidad y el predictor lineal se relacionen matemáticamente debe existir una función de enlace que se presenta a continuación:

$$E(Y|X) = \mu = g^{-1}(\eta) \tag{11}$$

9.4. Modelo logístico multinomial para variables ordinales multicategóricas

El modelo logístico multinomial ordinal o logit multinomial es una extensión del modelo logístico binomial que permite analizar y predecir variables ordinales con más de dos categorías ordenadas. Es útil cuando se tiene una variable de respuesta que no es binaria, sino que tiene múltiples niveles ordenados.

En el modelo logístico multinomial ordinal para distintas variables categóricas, se utiliza la función de enlace logit para modelar las probabilidades de pertenecer a cada categoría ordinal o no; en relación con las variables predictoras. El modelo asume que las categorías ordinales se encuentran en una escala continua y que la relación de odds (probabilidades de que un evento ocurra o no ocurra un evento) entre las categorías es constante a lo largo de la escala y a la vez son acumulativos se así se desea. (Allan Agresti 2010).

La regresión logística se encuentra en una ecucación de la siguiente forma:

$$P(Y) = \frac{1}{1+\exp^{-(\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_nx_n)}} \tag{12}$$

Donde $P(Y)$ es la probabilidad de que ocurra un evento en particula y exp hace referencia a la función exponencial, y que además; destaca que dicha distribución de probabilidad hace parte de la familia exponencial.

Ahora la regresión multinomial se denota la probabilidad de Y:

$$P(Y \leq j|x) = \pi_1(x) + \cdots + \pi_j(x), j = 1,2,3,...,J \tag{13}$$

Para que el modelo logístico sea acumulativo; se presenta la siguiente ecuación:

$$P(Y \leq j|x) = log\frac{P(Y \leq j|x)}{1 - P(Y \leq j|x)} \tag{14}$$

La aplicación del logaritmo, en realidad son para los π para cada x de la siguiente manera:

$$P(Y \leq j|x) = log\frac{\pi_1(x) + \cdots + \pi_j(x)}{\pi_{j+1}(x) + \cdots + \pi_J(x)}, j = 1 + \cdots + J - 1 \tag{15}$$

El modelo consta de logits que se acumulan y se proporcionan con sus probabilidades; cada modelo tiene su propio intercepto α_j:

$$logit[P(Y \leq j|X] = \alpha_j + \beta_1 x_1 + \cdots + \beta_i x_i, donde : j : 1, ..., J - 1, i : 1, ..., M \tag{16}$$

Al ser acumulativo-ordinal en cada y, se verán los efectos en cada β.

10. Resultados

10.1. Variables

La base de datos utilizada en este proyecto es de libre uso y es información pública, y se encuentra en la página web de Datos Abiertos que pertenece al Estado colombiano; se presentan las variables más importantes a nivel socidemográfico que componen la base de datos y otras que a criterio estadístico no tienen mayor información.

1. **Año:** Es una variable de tipo numérica, que demarca el periodo de tiempo en años en el ocurrió el homicidio o muerte violenta, existen cuatro registros, los cuales son los años 2016, 2017, 2018 y 2019.

2. **Zona**: El lugar donde ocurrió el homicidio, es una variable binaria o dummy (0,1), ya que las opciones donde ocurrieron los hechos fueron en una zona urbana o rural.

3. **Grupo de Edad:** Esta variable está contemplada en 9 niveles, donde se encuentran agrupadas en edades de 10 años, los 9 niveles de agrupaciones son: 1 a 9, 10 a 19, 20 a 29,30 a 39, 40 a 49, 50 a 59, 60 a 69, 70 a 79 y más de 80 años. Esta variable se codifica de forma ordinal para su debida manipulación.

4. **Sexo:** Variable dummy categórica que tiene 2 posibles opciones; mujer u hombre (0,1).

5. **Estado civil**: Esta variable tiene 6 posibles respuestas categóricas en las cual se encontraba la víctima, estas opciones son casado, soltero, viudo, separado, unión libre y divorciado. Esta variable se codifica a factor.

6. **País de nacimiento**: Nacionalidad de la persona es una variable dummy; 1 si nacio en Colombia y 0 si es extranjero.

7. **Clase de empleo**: Una variable muy importante ya que consta de 13 posibles respuestas que define a que se dedicaba o que tipo de empleo tenia la persona víctima de homicidio, estos son agricultor y ganadero, ama de casa, comerciante, desempleado, empleo particular, empleado público, empleado de salud, estudiante, fuerza pública, grupos ilegales, independiente, pensionado y otro. Esta variable está codificada como factor.

8. **Escolaridad:** Tiene 5 niveles de respuesta que fueron expuestos en el marco teórico y a la vez es una variable ordinal, los niveles de escolaridad son analfabeta, primaria, secundario técnico o tecnólogo y superior.

9. **Otras variables:** Existen variables que no tienen mayor relevancia en la base y tampoco tienen efectos significativos en la víctima de forma directa, esto se demuestra con el algoritmo de eliminación recursiva RFE; esta variables son la hora, el día, la fecha, tipo de arma, tipo de sitio, móvil del agresor, el móvil de la víctima y código del DANE.

10.2. Algoritmo de eliminación recursiva RFE

Este algoritmo ayuda a identificar las características o variables más importantes para nuestra variable dependiente del estudio; que es el nivel de escolaridad de la persona; de las 21 variables originales que tenemos en la base consolidada, el algoritmo divide los datos de forma aleatoria; se asigna 80 % de los datos para entrenamiento y 20 % para la prueba, se designa la variable destino para este algoritmo el cual es el nivel escolaridad. Además de esto se usa validación cruzada que se repite 10 veces con 5 repeticiones cada una, para darle más aleatoriedad al modelo y este proporciona con niveles de importancia, permite saber cuántas variables son óptimas y cuáles son las mejores variables predictivas para ejecutar el modelo.

A continuación se presentan cuántas variables son las más importantes:

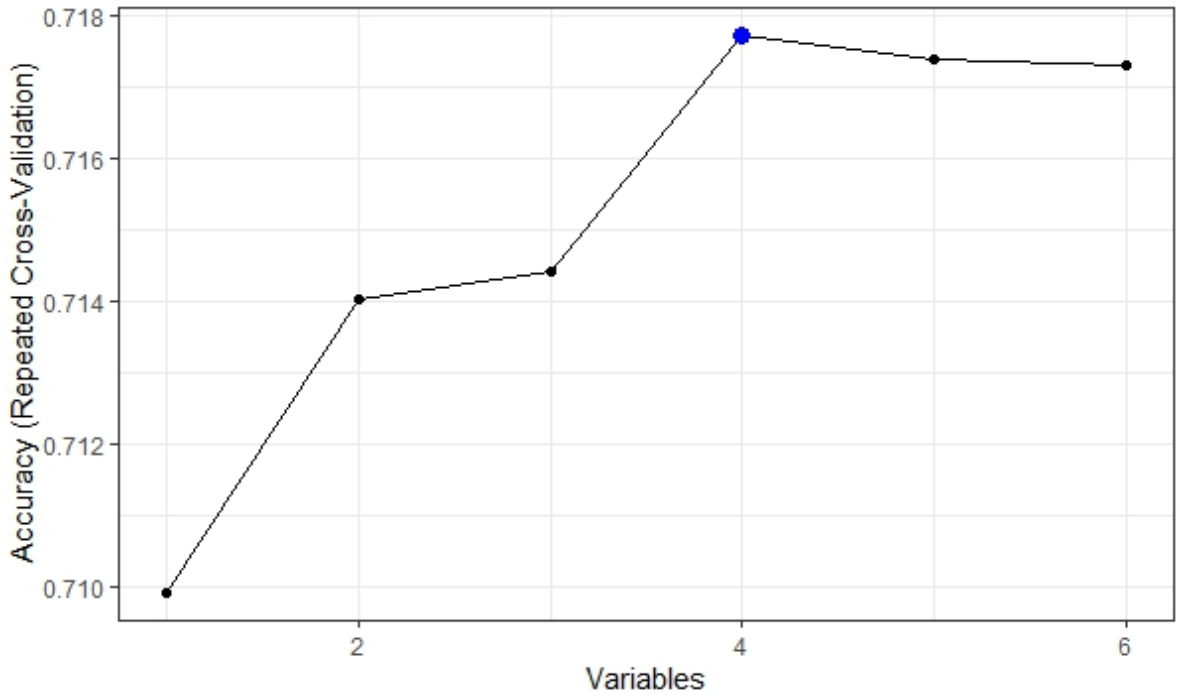

Figura 1: Número de variables que más tienen relación con la variable nivel de escolaridad

En la primera gráfica se evidencia el punto azul, que muestra el número óptimo de variables a usar, que son 4 características significativas. A medida que el algoritmo hace las validaciones cruzadas y con sus respectivas repeticiones sugiere que existen caracteristicas asociadas al nivel de escolaridad, sin embargo existen 2 puntos que están algo alejados del punto óptimo azul pero esto no impide que estas se puedan ser usadas ya que encuentra una relación aceptable con la variable nivel de escolaridad.

Variable	Importancia
Clase de empleado	25.230944
Grupo de edad	19.873611
Estado Civil	4.752146
Zona	3.664341
Sexo	3.523694
País de nacimiento	1.991250

Tabla 1: Variable con su importancia frente a la variable escolaridad

En la tabla se observa la cantidad de importancia que tiene cada variable con respecto a la variable objetivo nivel de escolaridad, donde encabezan la clase de empleo, y la edad agrupada, con un nivel de importancia que ronda el 20 %; seguido por estado civil, zona, sexo y país de nacimiento y que sus niveles de importancia rondan entre el 7 % hasta el 2 % respectivamente.

Estas 6 variables multicategóricas, ordinales y dummy, que el algoritmo de eliminación recursiva estimó con niveles de importancia pueden explicar el posible comportamiendo de la escolaridad de una persona que fue víctima de una muerte intencional.

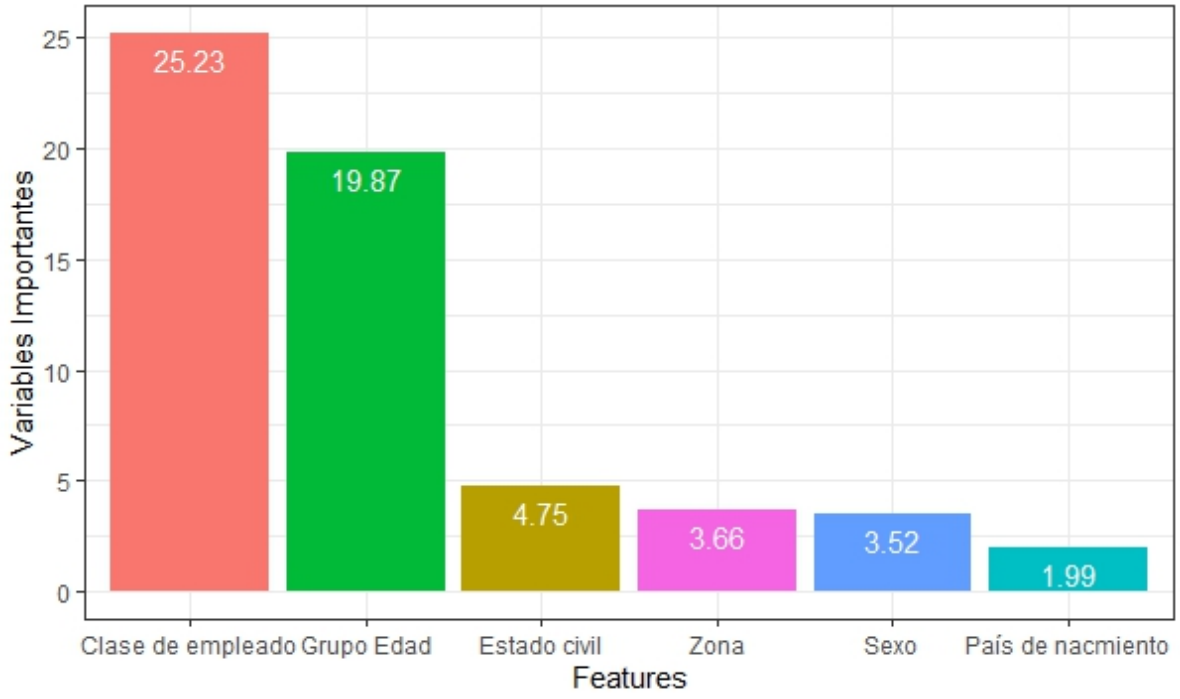

Figura 2: Número de variables que más tienen importancia con la variable nivel de escolaridad

10.3. Mapa de población Valle del Cauca

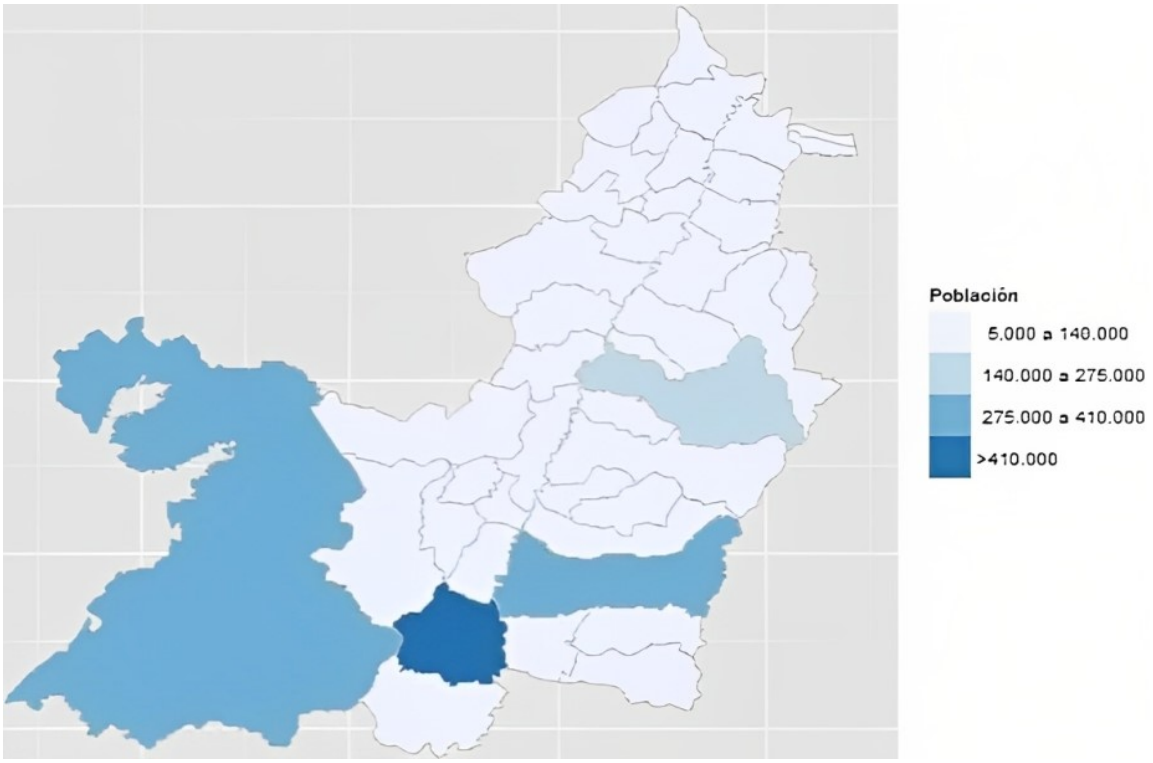

Figura 3: Mapa de distibución de población en el Valle del Cauca

Este mapa del Valle del Cauca representa la cantidad aproximada de personas que habitan en cada uno de los municipios, la ciudad de Santiago de Cali, que se ubica en el centro sur del departamento, es la que tiene tonalidad azul oscura, en esta ciudad habitan aproximadamente más de 2.250.000 personas, es la conurbación más poblada del departamento, y a la vez es el municipio que más homicidios presento en el departamento en los años 2016 al 2019, determinando que en ese tiempo se cometieron 4.789 homicidios comprobables y que tuvo en promedio durante estos cuatro años una tasa de homicidios de 45.97 por cada 100.000 habitantes. Esta ciudad durante muchos años ha tenido un puesto asegurado en las 50 ciudades más violentas del mundo según el Consejo Ciudadano para la Seguridad Pública y la Justicia Penal de México (2020). Las razones principales de los homicidios son las pandillas locales y grupos ilegales, ya que es una ciudad estratégica cerca al pacífico colombiano. Su vecina Palmira, en el oriente sur del departamento, tiene una población tanto en el área rural como en el área urbana de 360.000 personas; también se encuentra entre las ciudades más violentas del mundo, durante los 4 años se cometieron 582 homicidios comprobables y una tasa de homicidios en promedio de 45.15; tiene casi todas las mismas características sociales y demográficas de la ciudad de Cali, pero Palmira tiene una bastante población rural.

Buenaventura con una población de 315.000 habitantes es bien sabido que este municipio es un sitio bastante importante para el país, sobre todo por su puerto, que importa y exporta mercancías desde y al extranjero, es el municipio más grande del departamento y el 77 % de su población se encuentra en la cabecera urbana, por esta misma razón grupos ilegales se han disputa este territorio a sangre y fuego durante muchos años por el control de economías ilegales, por esta razón se presentaron 327 muertes violentas en transcurso de estos años; y estuvo entre las 50 ciudades más violentas del mundo durante este periodo de tiempo. Tuluá y Cartago municipios que cuentan con poblaciones relativamente grandes, tuvieron 314 y 204 homicidios respectivamente, algo que a nivel Colombia es preocupante. El departamento del Valle tuvo durante estos 4 años una tasa de homicidio de

53.89 por cada 100.000, y se presentaron 9.217 homicidios comprobados.

10.4. Modelo logístico multinomial con variables ordinales, multicategóricas y binarias para los años 2016, 2017, 2018 y 2019 con P valor

Categoría	Value	Std Error	Valor T	Valor p
Ama de casa	0.989305	0.235305	4.204348	0.000
Comerciante	2.39676	0.199067	12.03995	0.000
Desempleado	0.616489	0.148222	4.15920	0.000
Empleado Particular	1.743921	0.14289838	12.20393	0.000
Otro Empleado	5.668080	0.569836	9.94685	0.000
Estudiante	2.756483	0.213678	12.90013	0.000
Fuerza pública	4.718861	0.309166	15.26317	0.00
Grupos ilegales	-0.919053	0.676567	-3.58406	0.0174
Independiente	1.274324	0.149246	8.53836	0.000
Pensionado	2.846190	0.406947	6.99400	0.000
Grupo Edad 10-19	0.845326	0.397367	2.12731	0.003
Grupo Edad 20-29	-4.011928	0.397327	-10.09727	0.000
Grupo Edad 30-39	2.418676	0.331094	7.30509	0.000
Grupo Edad 40-49	-2.058019	0.257091	-8.00500	0.000
Grupo Edad 50-59	0.443874	0.187677	2.36508	0.012
Grupo Edad 60-69	-0.594437	0.139505	-4.26102	0.000
Grupo Edad 70-79	0.078232	0.103073	0.75899	0.0448
Masculino	-0.113321	0.109680	5.33192	0.002
Soltero	-0.688158	0.143183	-4.80614	0.000
Unión libre	-0.628247	0.145250	-4.32527	0.000
Viudo	-1.837567	0.459112	-4.00243	0.000
Zona Rural	0.338901	0.064501	5.25417	0.000
Analfabeta—Primaria	-3.068918	0.398742	-7.69649	0.000
Primaria—Secundaria	0.754383	0.391942	3.92473	0.000
Secundaria—Técnico o Tecnólogo	6.02752	0.39938	15.0920	0.000
Técnico o Tecnólogo—Superior	6.673148	0.404728	16.48794	0.000

Tabla 2: Resumen del modelo para los años 2016 al 2019

Este modelo nos muestra que el valor de p para todas las variables es menos a 0.05, por esto se infiere que son estadísticamente significativas al $95\,\%IC$. Por lo anterior se entiende que J es el número total de categorías dependientes ordenadas j, que es el nivel de escolaridad (Analfabeta, primaria, secundaria, técnio/tecnólogo y superior). M el número de variables independientes i (Clase de empleado, grupo de edad, estado civil, sexo, zona y país de origen). El cojunto de datos presentados son $J = 5$ y $M = 6$.

La variable categórica binaria como sexo se puede afirmar que una persona de sexo masculino, a diferencia de una persona de sexo femenino, está asociado con una menor probabilidad de tener una educación baja (primaria o secundaria) al ser víctima del homicidio. El valor t es mayor que 2 y, por lo tanto, es estadísticamente significativo al nivel del 95 %. Se evidencia que la variable edad pueden interpretarse que con un aumento de categoría ordinal en la edad, el logaritmo de probabilidades va descendiendo, porque la probabilidad de que una persona sea víctima de un homicidio, al ser de edad avanzada y sin importar su nivel de escolaridad es, en ocasiones menos frecuente.

Las intersecciones ordinales del nivel de escolaridad matemáticamente, el intercepto analfabeta—primaria y primaria —secundaria corresponde a logit$[P(Y \leq 2)]$. Puede inferirce que el registro de probabilidades de los niveles de escolaridad analfabeta versus primaria versus secundaria de forma acumulativa que son los que concentran mayor información con el homicidio . Así como logit$[P(Y \leq 4)]$ que son los registros de todas las probabilidades con todos los niveles de escolaridad de forma ordinal.

10.5. Mejor modelo logístico multinomial con variables ordinales, multicategóricas y binarias para los años 2016, 2017 y 2018

Categoría	Estimación	Std Error	Z Valor	$Pr(>\|z\|)$
Intercepto 1	-2.53142	0.53652	-4.718	2.38e-06 ***
Intercepto 2	1.46414	0.52952	2.765	0.05692
Intercepto 3	6.75677	0.53850	12.547	2e-14 ***
Intercepto 4	7.35504	0.54371	13.527	2e-11 ***
Ama de casa	-1.64548	0.53480	-3.077	0.00209 *
Comerciante	-2.63715	0.43538	-6.057	1.39e-09 ***
Desempleado	-0.70525	0.31766	-2.220	7.68e-12 ***
Empleado Particular	-1.98648	0.30423	-6.530	6.60e-4 ***
Otro Empleado	-9.15841	1.81367	-5.050	4.43e-03 ***
Estudiante	-2.64548	0.43220	-6.121	0.0487
Fuerza pública	-4.71963	0.68998	-6.840	7.91e-5 ***
Grupos ilegales	1.29868	1.53439	0.846	0.39734
Independiente	-1.34769	0.31992	-4.213	2.52e-09 ***
Otro	-3.01493	1.69202	-1.782	0.07477 .
Pensionado	-2.64345	0.87365	-3.026	0.01248
Grupo Edad 10-19	3.63836	0.80114	4.541	5.59e-06 ***
Grupo Edad 20-29	-1.49445	0.68184	-2.192	6.72e-07 ***
Grupo Edad 30-39	2.19189	0.52356	4.187	2.83e-05 ***
Grupo Edad 40-49	-0.07246	0.38366	-0.189	0.008502
Grupo Edad 50-59	0.61234	0.28947	2.115	0.03440
Grupo Edad 60-69	0.16404	0.22266	0.737	0.46130
Grupo Edad 70-79	0.09757	0.16705	0.584	0.55917
Sexo	0.01822	0.24474	0.074	0.0094065 **
Soltero	0.60840	0.34722	1.752	0.001974 **
Unión libre	0.39893	0.35320	1.129	0.0035869 *
Viudo	2.82983	0.96164	2.943	0.00929 **
Zona R	0.55459	0.13047	4.251	2.13e-05 ***
País de nacimiento	0.69717	0.43451	1.604	0.10861

Tabla 3: Resumen del mejor modelo para los años 2016 al 2018

Para estos años 2016, 2017 y 2018 el modelo infiere que existen muchas variables categorícas de una persona que fue víctima del homicidio que son significativas con respecto a la variable ordinal categórica; nivel de escolaridad. Las personas que son jóvenes, y que se encuentren entre las edades de 10 a 39 años están muy expuestos a una muerte violenta si tienen un nivel educativo menor o igual a la secundaria ($[P(Y \leq 3)]$).Una de las posibles razones se debe a que las personas jóvenes se ven involucrados en pandillas, en riñas por intolerancia y ajuste de cuentas. La categoría estado civil es relevante cuando la persona víctima principalmente se encuentra soltera o en unión libre y a la vez se encuentra entre las edades más jóvenes. El nivel de escolaridad en una persona se ve muy reflejada también en su ocupación lo es que igual a tipo de empleado, los que tienen un

nivel de escolaridad menor o igual a la secundaria y si se es jóven; normalmente se encuentran desempleados, o en un empleo particular con baja remuneración o de forma informal (Lo cual se encuentra oculto en una menor medida en la categoría independiente y comerciante). El sexo de la persona es importante ya que suele ser el hombre de bajo nivel de escolaridad el que es más propenso al homicidio, pero existe un nivel de escolaridad bajo entre las mujeres víctimas en el área rural.

10.6. Mejor modelo logístico multinomial ordinal con variables ordinales, multicategóricas y binarias con los años 2016, 2017, 2018 y 2019 agrupados

Categooría	Estimación	Std Error	Z Valor	$Pr(>\|z\|)$
Intercepto 1	-3.52389	0.70916	-4.969	0.0073e-07
Intercepto 2	-0.13762	0.69636	-0.198	0.84333
Intercepto 3	5.17954	0.70624	7.334	2.23e-13 ***
Intercepto 4	5.99203	0.72196	8.300	2e-16 ***
Ama de casa	-0.78625	0.26503	-2.967	0.003011 *
Comerciante	-2.33473	0.22513	-10.371	3e-10 ***
Desempleado	-0.58544	0.16744	-3.496	2e-14 ***
Empleado Particular	-1.67596	0.16188	-10.353	3e-11 ***
Otro Empleado	-4.69197	0.69687	-6.733	1.66e-1 **
Estudiante	-2.82374	0.24499	-11.526	0.0096 *
Fuerza pública	-4.75794	0.38369	-12.401	0.002 **
Grupos ilegales	0.83226	0.79654	1.045	0.296096
Independiente	-1.26970	0.16877	-7.523	5.35e-14 ***
Otro	-5.79235	0.66286	-8.738	0.0065 *
Grupo Edad 10-19	4.22559	0.46248	9.137	2e-16 ***
Grupo Edad 20-29	-2.70262	0.38292	-7.058	1.69e-12 ***
Grupo Edad 30-39	2.02785	0.29884	6.786	1.16e-11 ***
Grupo Edad 40-49	-0.59972	0.21849	-2.745	0.00054 **
Grupo Edad 50-59	0.55452	0.16070	3.451	0.0059 *
Grupo Edad 60-69	-0.17625	0.11658	-1.512	0.130585
Grupo Edad 70-79	-0.01195	0.08812	-0.136	0.892152
Sexo	-0.15029	0.12172	1.235	0.216925
Soltero	0.77407	0.15546	4.979	6.38e-07 ***
Unión libre	0.74346	0.15728	4.727	2.28e-06 ***
Zona R	-0.28383	0.07464	-3.803	0.000143 **
País de nacimiento	-0.44439	0.46898	-0.948	0.343344

Tabla 4: Resumen del mejor modelo para el año 2019

Para el modelo del año 2019 existen menos datos, pero son los más recientes y explican la realidad aproximada del comportamiento de los homicidios en el departamento del Valle del Cauca; se infiere no hay diferencias significativas, persiste la tendencia de que hay una relación entre las edades de las personas víctimas de una muerte violenta, que suelen ser jóvenes y que están entre los 10 a 39 años de edad con los niveles de escolaridad menores a secundaria. Así como las categorías asociadas con el estado civil (soltero y unión libre) y ocupación (principalmente desempleado, comerciante e independiente) tienen suma importancia con el nivel de escolaridad de la persona y repercute con el homicidio de este mismo.

11. Conclusiones

El algoritmo de eliminación recursiva RFE si cumplio su objetivo de determinar las variables categóricas tanto multiclase como binarias que más importancia tienen y que a la vez este método sugiere un grado de importancia de mayor a menor para cada una de estas 6 variables. Este método clasificó por orden de importancia a la clase de empleado, grupo de edad, estado civil, zona, sexo y país de origen o nacionalidad como las mejores variables predictivas, para estimar los posibles modelos de regresión, para evidenciar un tipo de relación entre estas mismas y la variable dependiente nivel de escolaridad de la víctima de homicidio.

La variable edad ordinal, el tipo de empleo, el estado civil y zona y sexo son las que más explican los modelos expuestos, estas variables predictivas afirman que si existe una persona jóven, soltero, que se encuentra desempleado y con niveles escolares que oscilan entre primaria y secundaria; tendrá una mayor probabilidad de estar involucrado en una muerte violenta y tener una bajo nivel de escolaridad. Los homicidios ocurridos en el Valle del Cauca durante los años, 2016, 2017, 2018 y 2019 guardan relación objetiva con la hipótesis de que existe una relación entre los homicidios y el nivel educativo de la víctima y que es explicado con variables predictivas con enfoque demográfico y social. A pesar de que el país de origen resulta una variable con una importancia baja, porque se demuestra en los modelos que no tienen mucha significancia, existieron datos de personas de origen venezolano que fueron víctimas de homicidio, y tienen las mismas características sociales, demográficas y educativas que un colombiano, por lo tanto los migrantes no son ajenos a este problema. Los hombres representaron el 92 % de los homicidios ocurridos en el departamento durante estos 4 años, y las mujeres represetaron un 8 %, pero en el área rural ocurrieron el 17 % de los homicidios, y la población rural total del departamento está en un 14 %, con esto se infiere que lo homicidios ocurrieron proporcionamente más en esta área, y las más afectadas son las mujeres que tienen niveles escolares (analfabeta, primaria y secundaria) relativamente inferiores a los hombres. Las personas asesinadas que tuvieron un nivel de escolaridad mayor a la secundaria (técnico o tecnólogo y superior) representaron un 2.4 % de los homicidios durante estos 4 años, pero en el Valle del Cauca el porcentaje de personas que tienen este tipo nivel de escolaridad en el departamento, está entre 14 % a 15 %, por lo tanto los homicidios, con gran diferencia se asocian a personas que no tuvieron niveles escolares superiores a la secundaria.

Los modelos al inicio estaban desbalanceados hacia las categorías de primaria y secundaria ya que componian más del 95 % de la información dependiente, pero no se realizo un método para equilibrar las categorías; ya que la información de estas pudieron sesgarse, por lo tanto se dividieron los mejores modelos y se explican que las variables predictivas se asocian mucho a la variable dependiente nivel de escolaridad.

Este trabajo abre una puerta de investigación para que se hagan estudios e investigaciones más profundos que giren alrededor de este tema con el objetivo de mitigar este flajelo que ha permeado de forma significativa a la sociedad del departamento del Valle del Cauca. Se pueden implementar políticas públicas de prevensión acordes para bajar los índices y cantidad de homicidios sobretodo en personas jóvenes que tienen un nivel de escolaridad entre primaria y secundaria, que se encuentran desempleados o con una ocupaciones laborales poco calificables e inestables.

Referencias

Acosta, C. (2012), 'Anatomía del conflicto armado en el Valle del Cauca durante la primera década del siglo XXI, Universidad San Buenaventuras, Cali'.

Agresti, A. (2002), *categorical data analysis*, tercera edn, Serie de Wiley en Probabilidad y Estadística, Universidad de Florida.

Annette, J. D. (2001), *An introdution to generalize linear models*, primera edn, Chapman Hall/CRC, University of British Columbia, Canada.

Bargent, J. (2014), Los grupos armados hacen del valle del cauca la cápital de violencia de colombia, Technical report, InSight Crime.

Bérmudez, A. J. (2022), 'Interpretabilidad categórica de clasificadores automáticos sobre contenido relacionado a la percepción de la seguridad'.

Christensen, R. H. B. (2019-12-10), *Ordinal - Regression Models for Ordinal Data*, R package version.
*https://CRAN.R-project.org/package=ordinal

Christopher, T. (2014), *Analysis of Categorical Data with R*, primera edn, Chapman Hall/CRC Texts in Statistical Science, University of Nebraska, Simon Fraser University.

Cohen, D. A. & Piquero, A. R. (2009), 'New evidence on the monetary value of saving a high-risk youth. journal of quantitative criminology', **25**, 25–49.

Cordeiro, G. (2010), *Modelos Lineares Generalizados e Extensoes*, primera edn, Departamento de Estatística e Informática, UFRPE.

Datos.Gov.co (2021), Datos abierto, Technical report, https://www.datos.gov.co/Justicia-y-Derecho/Homicidios-Colombia-a-os-2016-a-2019/vtub-3de2.

De Castro, R. (2003), *El universo LaTeX*, segunda edn, Unibiblos, Universidad Nacional de Colombia, Bogotá'.

Di Masso, M. & Granitto, P. (2014), 'Selección estable de variables independientes con rfe', *Centro Internacional Franco Argentino de Ciencias de la Información y Sistemas* pp. 1–9.

Hilbe.J.M (2017), *Logistic Regression Models*, Chapman Hall.

INEGI (2018), 'En números. Documentos de análisis y estadísticas, patrones y tendencias de los homicidios en México' ', pp. 15–45.

Klein, B. (2021), *Machine Learning with python tutorial*, 1st edition edn, Bodenseo.

Li, F., Liu, T. & Huang, L. (2017), 'Aplicación de regresión logística multinomial para investigar factores asociados con diferentes tipos de homicidio en China'.

Menard, S. (2009), *Logistic regression -from introductory to Advanced Concepts and Applications*, Sage, Sam Houston State University.

Milliken, G. & Johnson, D. (1984), *Analysis of Messy Data*, Vol. I of *of Designed Experiments*, Van Nostrand Reinhold, New York.

Polania, M. A. (2020), '*Modelo logístico multinomial ordinal para la caracterización del dengue en el departamento del Caquetá entre los años 2012 hasta el 2019*'.

Reyes, J. C. & Zarama, R. (2017), 'Análisis multinomial de los homicidios en Colombia para los años 2000 a 2004', *Revista de Economía del Rosario* pp. 151–175.

Rincón', C. (2020), 'Modelos logísticos multinomiales y características socioeconómicas asociadas a los homicidios en Colombia'.

Rudas, T. (2018), *Lectures on Categorical Data Analysis*, primera edn, Springer, Hungarian Academy of Sciences.

Velandia, A. (2019), '*Modelamiento de la tasa de homicidios en Norte de Santander por municipio para el año 2016 a través de modelos espaciales bayesianos*'.

Wintemute, G. J.Parham, C. A., Beaumont, J., J. Wright, M. & Drake, C. (2019), 'Epidemiología y aspectos clínicos de la violencia homicida con armas de fuego en California'.