ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL

Facultad de Ciencias Naturales y Matemáticas

"Creación de un scoring de seguros usando un modelo de aprendizaje automático para la zona 8 del Guayas (Guayaquil, Durán y Samborondón)"

PROYECTO INTEGRADOR

Previo la obtención del Título de:

INGENIERO EN ESTADÍSTICA INFORMÁTICA

Presentado por:

Ricardo Murillo

GUAYAQUIL - ECUADOR

Año: 2022 PAO I

DEDICATORIA

A Jesús, quien es mi Señor y salvador, que me he capacitado para la terminar esta etapa de mi vida.

A mi familia, que siempre me apoyó y empujó a seguir adelante.

AGRADECIMIENTOS

A Dios, quien ha sido mi motor y mi fuerza para seguir adelante en lo que me he propuesto.

A mis padres Kleber y Flor, a mis hermanos Fabricio, Maicol y Mia a mi cuñada Ana, a mi abuela Hipólita y a mis tíos los cuales siempre me dieron su apoyo incondicional y motivación.

A todos esos profesores que con su dedicación y capacidad de enseñanza fueron mi base y mi inspiración a seguir.

A mis amigos con los cuales empecé la carrera y a los que conocí en el camino, que siempre me brindaron su mano amiga sin condición alguna.

DECLARACIÓN EXPRESA

“Los derechos de titularidad y explotación, me corresponde conforme al reglamento de propiedad intelectual de la institución; Yo, Ricardo Raúl Murillo Portillo doy mi consentimiento para que la ESPOL realice la comunicación pública de la obra por cualquier medio con el fin de promover la consulta, difusión y uso público de la producción intelectual"

Ricardo Raúl Murillo
Portillo

EVALUADORES

Firmado electrónicamente por:
SANDRA LORENA
GARCIA BUSTOS

Ph.D. Sandra García Bustos

PROFESOR DE LA MATERIA

SERGIO ALEX
BAUZ OLVERA

Ph.D. Sergio Bauz

PROFESOR TUTOR

Resumen

La creación de nuevos productos de seguros de vida que se ajusten a todas las distantes clases económicas del país, resulta ser un mercado nuevo que necesita la aplicación de modelos estadísticos que minimicen la cantidad de siniestros, esto permite que las compañías aseguradoras puedan ofrecer sus productos sin llegar a afectar la rentabilidad de esta. En el presente documento, se describe la aplicación de modelos de aprendizaje automático para estimar un scoring de seguros de vida, el cual consiste en clasificar por niveles de riesgo a los individuos de acuerdo con su probabilidad de fallecer, modelando datos con variables de tipo clínico que revelen el historial clínico del posible asegurado.

Se utilizaron técnicas de balanceo de la muestra para evitar el sesgo de prevalencia de alguna clase, tales como: resampling y validación cruzada. Se evaluará y se seleccionará los modelos con métricas como sensibilidad, especificidad, precisión y exactitud del modelo. Además, para la identificación de perfiles de riesgos y segmentos de clientes, se utilizó el análisis factorial mixto y clustering para variables cualitativas y cuantitativas.

Finalmente, se seleccionó al modelo de regresión logística como el modelo que mejores métricas de evaluación y selección obtuvo, y se estableció la tabla estimada de descuento del valor de la prima de acuerdo con el punto de corte óptimo calculado del modelo seleccionado.

Palabras claves: Modelos de aprendizaje automático, validación cruzada, prima, seguros de vida, análisis factorial mixto

Abstract

The creation of new life insurance products that are adjusted to all distant economic classes in the country, turns out to be a new market that needs the application of statistical models that minimize the number of claims, this allows insurers to offer their products without affecting its profitability. This document describes the application of machine learning models to estimate a life insurance score, which consists of classifying individuals by risk levels according to their probability of dying, modeling data with clinical variables. that reveal the clinical history of the potential insured.

Sample balancing techniques were used to avoid prevalence bias of any kind, such as: resampling and cross-validation. Models will be evaluated and selected using metrics such as model sensitivity, specificity, precision, and accuracy. In addition, for the identification of risk profiles and customer segments, mixed factorial analysis and clustering were used for qualitative and quantitative variables.

Finally, the logistic regression model was selected as the model that obtained the best evaluation and selection metrics, and the discount table of the estimated value of the premium was established according to the optimal cut-off point calculated from the selected model.

Keywords: machine learning models, cross-validation, premium, life insurance, mixed factor analysis

CAPÍTULO 1

1. Planteamiento del problema

1.1. Introducción

La situación del país después de la segunda ola de la pandemia por COVID -19, empezó a mostrar las debilidades en los sectores sociales, de salud, económicos y productivos; la tendencia al alza en la delincuencia y de muertes violentas tuvo un repunte siendo la más alta vista en el país desde el 2012 (Gonzáles, 2021).

La crisis carcelaria representa un tema importante a resolver de manera urgente por el actual Gobierno, siendo una problemática que resuena fuertemente desde hace 3 años, que ha dejado un total de 276 personas fallecidas, esto por la punga entre bandas narcodelictivas dentro de los centros carcelarios (Gobierno, 2022). Los conflictos ya no solo permanecen dentro de las cárceles, sino que se han sitiado en varias ciudades dentro del país, y como evidencia las noticias nacionales presentan al menos una muerte por sicariato cada día, actualmente, la cifra de muerte por cada 100 mil habitantes es de 12,2.

De acuerdo con datos de la Dirección Nacional de Investigación de Delitos Contra la Vida, Muertes Violentas, Desapariciones, Secuestros y Extorsión (DINASED), en el primer cuatrimestre del 2022 se registraron de 1241 muertes violentas, frente a las 1379 registradas al cierre del 2020, es decir, tan solo en el primer cuatrimestre del 2022 ya se obtuvo el 90% de los casos suscitados en 2020. En el cierre del año 2021, se registraron 2532 muertes violentas, para lo cual, en lo que va del primer cuatrimestre en el 2022, ya representa el 49% de los casos totales del 2021.

1.2. Definición del Problema

La situación de seguridad en el país es un tema que genera diferentes tipos de emociones dentro de la población, y la gran cantidad de hechos atroces nunca vistos en Ecuador han dejado en la incertidumbre a muchos ciudadanos, quienes se mantienen aún optimistas sobre las respuestas que el Gobierno actual implementará para reducir los índices de criminalidad. De acuerdo con datos del Ministerio de Gobierno del 1 de enero al 28 de mayo de 2022, existen 568 casos de muertes violentas que incluye a los cantones de Guayaquil, Durán y Samborondón correspondientes a la zona 8.

Por lo tanto, muchas personas cabezas de hogar han comenzado a considerar la adquisición de un seguro de vida que les permita no dejar desamparado a sus hogares y de acuerdo con datos de la Federación Ecuatoriana de Empresas de Seguros (FEDESEG), la venta de seguros de vida creció un 6,1% entre 2020 y 2021. Sin embargo, la mayoría cree que estos tienen un precio muy alto y que no pueden ser costeados dentro de los gastos familiares, Andrés Cordovez, Gerente General Seguros Equinoccial (Meléndez, 2022), menciona que debe ser considerado como adquisición de un respaldo y como una inversión para la familia.

Ante esta situación de inseguridad ciudadana, surge la necesidad de nuevos productos de seguros que se ajusten de mejor manera a la población de clase media y media-baja. Esta sería una propuesta que se está planteando Seguros Equinoccial, la aseguradora más grande del país, que lo resalta como uno de sus objetivos a mediano y largo plazo.

1.3. Justificación del problema

Desde hace mucho tiempo, los modelos actuariales han sido aplicados para medir el scoring de riesgos de seguros y para otorgar créditos. Un estudio de la Universidad de Antioquia creó un indicador para créditos de una entidad financiera, que le permitía otorgar un crédito con base a variables financieras volátiles, asegurando la rentabilidad y eficiencia de la institución bancaria (Juan Ochoa, 2010). Partiendo de

este antecedente, este modelo también puede ser aplicable al proyecto propuesto por la similitud en variables tales como: el perfil del posible cliente del producto del seguro, condiciones de salud y siniestros de tránsito.

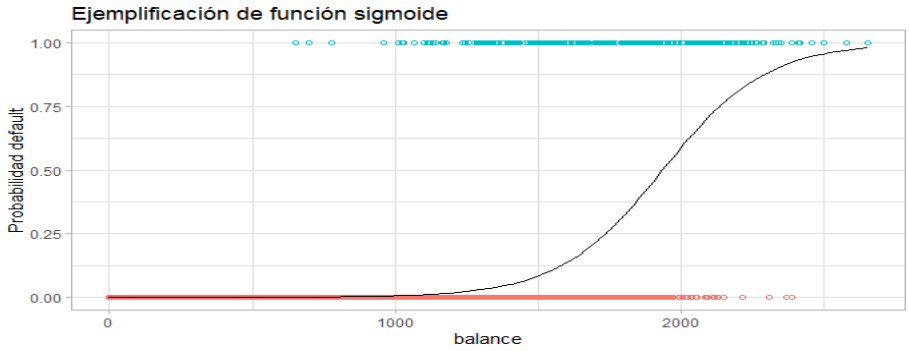
En la actualidad el aprendizaje automático juega un papel importante en la ejecución de este proyecto, mediante la estimación de la probabilidad de riesgo que permitirá definir un costo estimado de la prima a pagar por el cliente. Fue definida por el matemático Alan Turing, el padre de la computación, dentro del ámbito de la ingeniería (Serrano, 2012). Mediante su máquina, Turing resolvió problemas de operaciones básicas como sumas, restas, multiplicaciones, usando algoritmos o procedimiento de pasos hasta lograr completar la tarea, los cuales, aunque sencillos siguen siendo útiles hoy en día, sirviendo de base para procesos más complejos.

Algunas aplicaciones del aprendizaje automático corresponden con situaciones que se podrían realizar a diario como: el reconocimiento facial de los dispositivos celulares, las acciones antispam de Google en Gmail, regadores automáticos en plantaciones agrícolas, o en el área de la medicina para la prevención de enfermedades y/o diagnósticos tempranos de padecimientos como cánceres y tumores. Por ejemplo, mediante la aplicación de software de machine learning se ha logrado la identificación del virus de papiloma humano, o de células con transformaciones oncogénicas, además, de la aplicación de algoritmos para la identificación temprana de cáncer en el útero (Ávila-Tomás et al., 2021).

En el proyecto se usará un modelo basado en aprendizaje automático que estime una probabilidad de riesgo de seguro, que servirá para la construcción de una tabla referencia de primas netas estimadas a pagar por el posible cliente. El modelo se entrenará en base a variables como el perfil personal del cliente, el nivel de inseguridad y violencia dentro de la zona 8, así mismo, de factores clínicos, usando una función sigmoidal que es un tipo de función de activación, la cual toma los valores de las variables ingresadas y da como respuesta valores entre 0 y 1, como se muestra en la ilustración 1, que permiten establecer los grupos de primas a pagar por el usuario (Ponce Cruz, 2010)

Ilustración 1

Modelo usando una función de enlace sigmoidal



Nota: Los datos corresponde a valores ficticios. Fuente: Autoría propia, elaborado en Rstudio.

Existen técnicas que aplican el aprendizaje automático para encontrar patrones y relaciones dentro de un gran volumen de datos al que no se puede acceder mediante los procedimientos clásicos, a esta técnica se la conoce como minería de datos o data mining (Gallego, 2011).

La minería de datos ayuda a extraer patrones interesantes que antes eran desconocidos como: grupos de conjuntos de datos (análisis de conglomerados), registro de datos de anomalías (detección de anomalías) y dependencias (minería por reglas de asociación). Por lo tanto, estos patrones pueden considerarse como una especie de resumen de los datos de entrada. Además, se puede utilizar para análisis adicionales como aprendizaje automático y análisis predictivo (Vallejo & Tenelanda, 2012).

Los inconvenientes que presenta la minería de texto son examinar grandes y medianos volúmenes de datos, la naturaleza no organizada de los datos y la dificultad de hallar claves para estandarizar el lenguaje con objetivos inferenciales. Por ejemplo, el lenguaje escrito que utilizan las compañías de seguros cambia de coloquial a formal, los reportes policiales o las notas de los ajustadores de siniestros suelen tener terminología repetida, con construcciones léxicas que algunas veces son estáticas y predecibles, sin embargo, esto no se aplica a las redes sociales, que

continuamente cambian y, por consiguiente, son imposibles de examinar con procedimientos estándar (Asencios, 2004).

(María Paula Ávila Rodríguez, 2020) desarrolló a partir de aprendizaje automático supervisado un modelo de clasificación de tweets con redes neuronales artificiales para calcular el descuento sobre la tarifa del seguro de vida, segmentando en cinco categorías de riesgo a los usuarios de Twitter, conforme a las publicaciones relacionadas con la práctica de deportes, niveles de estrés, consumo de tabaco y alimentación

Parte de los datos que se usarán corresponde a datos de entes gubernamentales o de seguridad nacional, pero también será utilizada la técnica de text mining mediante el software estadístico Rstudio anclado con la API de Twitter, lo que nos permitirá conocer ciertas características o variables que puedan suponer un mayor riesgo de asegurabilidad, como pueden ser: número de muertes violentas por distritos, el número de intentos de asesinatos, robos entre otros. Existen actualmente investigadores que usan esta técnica para reconocer patrones de conductas como la depresión, mediante tweets en 2 fases; en la primera levantaron datos de usuarios que especificaron abiertamente en un tweet que sufrían de depresión, recolectaron datos de usuarios que ya habían sido identificados con depresión y finalmente otro grupo de 450 usuarios elegidos al azar.

En la segunda fase realizaron comparaciones de donde encontraron patrones con respecto a franjas horarias que evidenciaban una mayor actividad de este tipo de tweets en la noche, sustantivos usados que se presentaban con mayor frecuencia, formas de escritura como formal o poética y el uso de la gramática entre otros (Leis et al., 2019).

1.4. Objetivos

1.4.1. Objetivo General

Desarrollar un scoring de seguro de vida para la población de clase económica media y media-baja de la zona 8 del Guayas.

1.4.2. Objetivos Específicos

- Estudiar las variables más relevantes para la otorgación de un seguro de vida de acuerdo con las políticas de asegurabilidad.
- Desarrollar un modelo de aprendizaje automático y multivariante que estime el scoring de seguro de vida mediante los factores clínicos.
- Evaluar y seleccionar el modelo que mejor se ajuste a los datos con factores clínicos.

1.5. Alcance

El estudio se va a realizar para los habitantes de la zona 8 del Guayas que comprenden Guayaquil, Durán y Samborondón, considerando personas que sean de clase económica media, media baja dentro de un rango de edad de 18 hasta los 65 años.

CAPÍTULO 2

2. Marco Teórico

2.1.1. Estado del arte

Las compañías aseguradoras continúan en su búsqueda de nuevos métodos que le permitan llegar a más público sin dejar de ser rentables, además de ser más competitivos al cumplir con la demanda de los posibles clientes o asegurados.

En su búsqueda comienzan a considerar algún tipo de incentivo para los clientes como descuentos en las primas, tratando de poder abarcar a las clases económicas medias y media-bajas por lo que la implementación de modelos estadísticos de aprendizaje automático en los últimos años resulta una herramienta muy útil para determinar el riesgo de asegurabilidad de un posible cliente, evitando grandes pérdidas para la compañía aseguradora.

Una gran aseguradora a nivel mundial como John Hancock, estableció un nuevo método para calcular la prima de un seguro de vida, esto mediante un análisis previo a los posibles asegurados donde se les solicitaba el uso de smartwatchs o aplicaciones específicas en sus celulares, con el fin de hacer un monitoreo constante de sus actividades físicas, su ritmo cardiaco, presión, horas de sueño entre otras. Luego del monitoreo y de acuerdo con el resultado obtenido de sus hábitos diarios, otorgaban descuentos a las personas con menor riesgo de asegurabilidad u otros incentivos (María Ávila Rodríguez, 2020).

Otro caso de la aplicación de un modelo de aprendizaje automático fue realizado por la Association for Phychological Science en 2015, el cual implementó un modelo de regresión ridge, resultando ser modelo un más preciso para determinar la mortalidad por enfermedades cardiacas, que el método de verificación tradicional que determinaba esta mortalidad solamente con variable de tipo demográficas, sociales y de consumo de tabaco, hipertensión, diabetes y obesidad. El método consistía en un análisis de tweets

de la población estadounidense, reconociendo factores de riesgos mediante las publicaciones realizadas por los usuarios sobre relaciones sociales negativas, emociones negativas y demás patrones de lenguaje

2.1.2. Prima

El término de prima hace referencia al costo de adquirir una póliza de seguro ya sea esta de vida y salud, contra daños materiales o accidentes y seguros de servicios, valor que es pagado por la persona que está siendo asegurada por la cobertura de riesgo ante la ocurrencia de un siniestro (Díaz Díaz et al., 2008).

2.1.3. Siniestro

Cuando el asegurado sufre la ocurrencia del riesgo, la empresa aseguradora debe retribuir económicamente a las partes convenidas en el contrato de la póliza, esta remuneración es comúnmente proporcional al tiempo que queda por culminar el contrato de la póliza del seguro (Nationale Nederlanden, n.d.).

2.1.4. Riesgo

El riesgo corresponde a la probabilidad de la ocurrencia de un siniestro, en el caso de nuestra investigación corresponde a que el asegurado fallezca. Para la construcción de las tablas de las primas estimadas a pagar, entre más alto sea el riesgo menor será el descuento que se pueda aplicar al valor de la prima establecida por la aseguradora, se puede apreciar esto en la tabla 1.

Tabla 1

Descuentos estimados según el riesgo del individuo

Riesgo	% de descuento
Bajo	20%
Medio	10%
Alto	0%

**Nota.* Esta tabla presenta un valor estimado de descuento en el costo de la prima del seguro de acuerdo con el nivel de riesgo del cliente.

2.1.5. Prevalencia de fallecidos

La prevalencia corresponde a la cantidad o frecuencia de individuos de la población, que presenta una característica de interés (Segura Egea, 2002), para efecto del estudio corresponde a aquellos que han fallecido y de los cuales se conocen variables demográficas y de enfermedades. Definimos P(e) como la probabilidad de que el individuo fallezca y la calculamos de la siguiente manera:

$$P_{(e)} = \frac{No.\ de\ fallecidos}{Total\ población} \quad \textbf{(1)}$$

2.1.6. Matriz de confusión

Es una técnica de visualización del desempeño del modelo de clasificación u algoritmos de aprendizaje supervisado mediante una matriz, cada una de las columnas representa el número de predicciones de las clases, mientras que los renglones representan la instancia o clase real, de esta manera se puede visualizar fácilmente si el modelo se está confundiendo entre las dos clases, es decir, entre lo real u observado frente a lo predicho como se aprecia en la tabla 2.

Tabla 2

Representación de una Matriz de confusión

		Predicho	
		1	**0**
Real	**1**	A	B
	0	C	D

**Nota.* La tabla muestra un ejemplo demostrativo en el cual se puede visualizar la construcción de una matriz de confusión, las letras en mayúscula representan cuanto de lo observado es igual a lo predicho.

2.1.7. Sensibilidad

Es una métrica que sirve como herramienta de diagnóstico del ajuste del modelo que especifica la probabilidad de que un individuo presente un diagnóstico positivo, cuando este realmente presente dicho diagnóstico (Segura Egea, 2002), es decir, si se trata de una prueba de una enfermedad diagnostique correctamente al paciente con dicha enfermedad siempre y cuando la tenga. Por lo que entre más sensible sea la prueba más

fiable es el resultado y permite descartar la presencia la enfermedad, se calcula de la siguiente manera:

Definimos (h+) como el resultado positivo de la prueba en la enfermedad (e+), por lo que la probabilidad es condicionada a que el individuo presente la enfermedad, sea (a) el total de enfermos y (a+c) total de verdaderos positivos que se obtendría al aplicar la prueba a todos los enfermos, por lo que formula de la probabilidad condicionada corresponde a $P_{(h^+ \mid e^+)} = \frac{a}{a+c}$ **(2)** y su complemento representa el porcentaje de falsos negativos.

2.1.8. Especificidad

Al igual que la sensibilidad corresponde a una herramienta de diagnóstico del modelo, expresa la probabilidad de que el resultado del diagnóstico sea negativo (h-), cuando la persona no presenta la enfermedad (e-), representa así entonces la tasa de verdaderos negativos y también es una probabilidad condicionada. Sea (d) el total de resultados negativos respecto al total de las personas sin la enfermedad (b+d), obtenemos la especificidad es igual a $P_{(\bar{h} \mid e)^-} = \frac{b}{b+d}$ **(3)** y su complemento representa al porcentaje de falsos positivos.

Cuanto más específica es la prueba, menor probabilidad existe que reportar un caso de obtener un falso positivo.

2.1.9. Precisión del modelo

Después del análisis y como métrica de evaluación del modelo, es necesario establecer la precisión del modelo, y este es la probabilidad de obtener un diagnóstico correcto y se obtiene al calcular el total de verdaderos positivos, sobre el total de verdadero positivos y falsos positivos la fórmula es la siguiente:

$$P_{(p)} = \frac{vp}{vp+fp} \quad \textbf{(4)}$$

Lo esperado es que el modelo tenga una precisión muy cercana al 100%, lo que indica que el modelo capaz de predecir correctamente, no se debe confundir con la exactitud

del modelo la cual no corresponde a una métrica válida de rendimiento cuando la muestra presenta una gran diferencia en el número de casos de alguna clase sobre la otra, un ejemplo corresponde a tener muchos resultados positivos y muy pocos resultados negativos.

2.1.10. Error de Clasificación

Este error corresponde al complemento de la precisión del modelo, por lo que entre más alta sea la precisión del modelo menor será el error de clasificación, se calcula de la siguiente forma:

$$P_{ec} = 1 - {}_{(p)} \quad \textbf{(5)}$$

2.1.11. Validación cruzada con k iteraciones

Es una técnica que se utiliza en muchos modelos de aprendizaje automático, esto con el fin de estimar la precisión del modelo. La técnica consiste en dividir en dos grupos los datos de la muestra, repartiendo un porcentaje como datos de entrenamiento y el restante en datos de testeo, realiza una optimización de los parámetros a estimar del modelo de manera que se ajusta lo mejor posible con la información de los datos de entrenamiento.

Al repetir este proceso en k iteraciones los datos se agrupan en k subconjuntos, donde cada subconjunto es utilizado como datos de testeo y el k-1 restante de los datos, es usado como datos de entrenamiento. Este método se vuelve más preciso puesto que con k combinaciones de datos de entrenamiento y testeo, la media aritmética de los resultados por cada iteración vuelve los parámetros estimados más precisos, sin embargo, tiene un costo computacional más alto. Existen métodos para determinar cuántas iteraciones son necesarias, pero lo usual es usar 10 iteraciones (Luis Enrique et al., n.d.).

2.1.12. Balanceo de la muestra

En ocasiones se tiene una muestra que contiene los resultados de una variable que es tipo dicotómica, y cuando un gran número de registros se agrupa para una sola clase, el sesgo tiende a orientarse hacia esa clase lo que puede llevar a errores en la sensibilidad y la especificad del modelo, la desventaja está en que la muestra se reduce considerablemente, pero existe la mejora en las métricas antes mencionadas lo que permite hacer mejores evaluaciones del modelo a seleccionar (Ledesma, 2008).

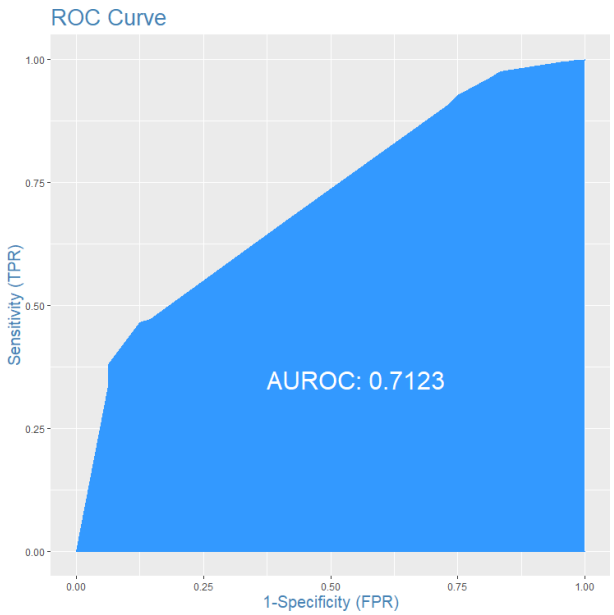
2.1.13. Curva ROC

Es una gráfica que muestra la sensibilidad frente a la especificidad dentro del modelo de clasificación, nos indica la proporción los valores que son verdaderos positivos frente a los que resultan ser falsos positivos. Resulta ser una herramienta para la selección de modelos óptimos y preferirlos sobre modelos óptimos, la ventaja está en que es independiente de cómo se distribuyen las clases dentro de la muestra, en caso de que una clase tenga mayor prevalencia dentro de la muestra.

El área debajo de la curva de ROC se relaciona comúnmente como la probabilidad en la que un modelo clasificador puntúa una instancia positiva aleatoria por encima de una negativa, toma siempre valores entre 0.5 a 1 y entre más cercano se encuentre a 1 más ajustado es el modelo con los datos usados. Si se obtiene un área bajo la curva de 0.5, el modelo es incapaz de diferenciar entre las dos clases, si está entre 0.6 y 0.75 se considera regular, entre 0.75 y 0.9 se considera bueno, y 0.9 a 1 como excelente (Cerda & Cifuentes, 2012).

Ilustración 2

Curva de ROC



2.1.14. Variables

Para la creación del scoring se usaron variables de historial clínico de individuos con fechas del año 2020 y 2021, provenientes de una fuente privada. El total de registros es de 5122 observaciones de las cuales solo 907 registros corresponden a individuos fallecidos, lo que denota un claro desbalance en la muestra. Las variables utilizadas son de tipo dicotómicas y son las siguientes:

- Fallece
- Tiene o ha tenido Covid-19
- Domiciliado en Durán
- Domiciliado en Guayaquil
- Domiciliado en Samborondón
- Sexo
- Sufre de alguna alergia
- Presenta o presentó asma
- Tiene o ha tenido alguna enfermedad autoinmune
- Tiene o ha sufrido alguna enfermedad cardiovascular

- Tiene diagnosticado diabetes

Para la base de datos proporcionada por el ministerio de gobierno acerca de homicidios intencionales que recopila 673 registros de homicidios, asesinatos y femicidios en la zona 8 desde enero hasta junio de 2022 se efectuó un análisis factorial de datos mixtos en el cual se contaron con las variables

- Arma: Variable categórica con 5 niveles
- Sexo: Variable categórica con 2 niveles
- Área del Hecho: Variable categórica con 2 niveles
- Tipo lugar: Variable categórica con 2 niveles
- Edad: Variable cuantitativa
- Franja: Variable categórica con 4 niveles

2.1.15. Modelo de Regresión Logística

Este modelo define Y_1, Y_2, …, Y_n como una muestra proveniente de una población Bernoulli cuyo parámetro a estimar es *p*, y sea el vector de variables explicativas $X_{ik\ x\ 1}$ đe la cual se puede describir la matriz con las *k variables explicativas*, de la siguiente ecuación (MARIO DAVID SOLÓRZANO CARVAJAL, 2016):

$$X_{k\times 1} = \begin{bmatrix} 1 \\ X_1 \\ X_2 \\ \cdot \\ \cdot \\ \cdot \\ [X_{k-1}] \end{bmatrix} \rightarrow X_{i\,k\times 1} = \begin{bmatrix} 1 \\ X_{i1} \\ \text{i2} \\ \cdot \\ \cdot \\ \cdot \\ [X\text{i},_{-1}] \end{bmatrix}; i\text{: } 1,2..,n \qquad \textbf{(6)}$$

Los parámetros del modelo corresponden a un vector $\beta_{kx1,}$ los cuales se detallan a continuación:

$$\beta_{k\times 1} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \cdot \\ \cdot \\ \cdot \\ [\beta_{k-1}] \end{bmatrix} \qquad \textbf{(7)}$$

Por lo que a partir de la ecuación (1) y (2) y considerando el modelo con una función logística acotando el valor estimado de la variable a predecir entre 0 y 1, como la distribución de probabilidad de la variable dependiente se define la siguiente expresión (Fiuza Pérez & Rodríguez Pérez, 2000):

$$\boldsymbol{X}_i'\mathrm{Q} = \beta_0 + \beta_1 X_{iy} + \cdots + \beta_{k-1} X_{i,-1} \quad \textbf{(8)}$$

2.1.16. Funciones de enlace

Logit

La función de distribución acumulada estándar logit se define como la inversa de la distribución logística, y corresponde a la función enlace por defecto o canónica en modelos que presentan datos provenientes de una distribución binomial. A continuación de se detalla la transformación del modelo logit (Ravelo & Ponsot Balaguer, n.d.):

$$(y)_i = \rho \; i = \frac{\exp(\boldsymbol{X}^F \mathrm{Q})}{1+\exp(\boldsymbol{X}^F_i \mathrm{Q})} \quad \textbf{(9)}$$

$$\log\left(\frac{P}{1-p}\right) = \log\left(\frac{[yi]}{1-[yi]}\right) = \boldsymbol{X}'\mathrm{Q}_i \quad \textbf{(10)}$$

Probit

La función de distribución acumulada normal estándar probit, se define a su vez como la inversa de la distribución normal estándar, su eso está relacionado con pruebas de toxicología o farmacocinética, al usar esta función de enlace el modelo es conocido como modelo probit o regresión probit. A continuación, se detalla dicha transformación:

$$(y = 1|x) = \Phi(\beta_0 + \beta_1 x_i, 1 + \cdots + \beta_{k-1} x_{i,k-1}) \quad \textbf{(11)}$$

$$(z) = \int_{-\infty}^{z} \left(\frac{1}{2\pi}\right)^{1/2} \{-\frac{u^2}{2}\} du \quad \textbf{(12)}$$

2.1.17. Modelo de Random Forest

El método del clasificador consiste en entrenar varios árboles de decisión de manera aleatoria, basado en un modelo de aprendizaje automático que promedia todos los árboles para obtener el mejor de ellos, esto con el fin que el valor predicho sea el más preciso posible. Presenta dos ventajas importantes; la primera corresponde a que es capaz de procesar grandes cantidades de datos y la segunda ventaja es que permite evitar el sobreajuste del modelo gracias a los subconjuntos aleatorios creando ramificaciones más pequeñas (Omar et al., 2021), esto mediante el uso parámetros que pueden ser ajustados directamente dentro del software estadístico de Rstudio.

El bootstrapping garantiza que cada uno de los árboles de decisión sean únicos dentro de todos los árboles de decisiones aleatorios, o también conocidos como el bosque aleatorio. Por lo cual random forest permite tener modelos de clasificación con un menor sesgo y varianza, y en cuanto a precisión resalta sobre el resto de los métodos de clasificación de aprendizaje automático (Misra & Li, 2020).

2.2. Máquina de Vector Soporte

Es una técnica que se aplica principalmente con problemas de clasificación, debido a que este algoritmo se creó para clasificar los grupos de respuesta cuando la variable de explicación era de tipo binaria, y para ciertas funciones Kernel se evidencia que es capaz aprender fácilmente de los datos. Una ventaja es que su error de clasificación puede llegar a ser similar al error de Bayes, esto cuando el conjunto de datos de entrenamiento es suficientemente grande, sin embargo, este algoritmo también puede reducir el conjunto de datos de entrenamiento para simplificar la representación del algoritmo.

El algoritmo realiza la clasificación a través de la búsqueda de un separador que maximice la separación, primero correlaciona los datos usando un espacio de características de alta dimensión representado por un hiperplano, con esto lograr categorizar o segmentar el conjunto de datos de entrenamiento. Cuando el conjunto de datos no es linealmente separable se busca encontrar el hiperplano con menor error empírico posible.

La principal desventaja está en la alta complejidad temporal del algoritmo, por lo que entre más grande sea el conjunto de datos de entrenamiento, más alto es la complejidad. Si se considera m como el total de datos de entrenamiento, la complejidad se relaciona como $O(m^3)$ pero ya se pueden encontrar otros métodos para reducir la complejidad en $O(m^2)$ (García Díaz & Lozano Martínez, 2006).

2.3. Análisis de correspondencia múltiple

Es un procedimiento estadístico dentro de un grupo de procedimientos factoriales destinados a estudiar las relaciones entre cualquier número de características de varias variables categóricas de matrices de datos ordinales y medir personas en términos de variables p.

Se construye una matriz simétrica de bloques llamada tabla de Burt y el siguiente paso es la diagonalización para encontrar la tabla de frecuencias para cada una de las Q variables categóricas de interés y la tabla de contingencia para la intersección de cada variable que se considera que está relacionada. matriz asociada. Especifica los autovalores o autovalores asociados a cada uno de los factores o ejes extraídos. Sea V la matriz a diagonalizar, entonces se puede indicar que:

Después de determinar cada factor, la proyección ortogonal de las diversas modalidades sobre él nos permite determinar qué modalidades hicieron la mayor contribución a la formación del factor.

Desde allí se define la contribución absoluta como la fracción de la varianza "explicada" por un factor. De igual manera son de gran importancia para el análisis las llamadas contribuciones relativas, es decir, las relaciones entre la modalidad y factor.

2.4. Coeficiente de Silhouette

El coeficiente de silueta o puntuación de silueta es una métrica utilizada para calcular la bondad de ajuste una técnica de agrupación (Espinosa Zúñiga, 2020). Su valor oscila entre -1 y 1.

1: Significa que los grupos están bien separados entre sí y claramente diferenciados.

0: Significa que los conglomerados son indiferentes, o podemos decir que la distancia entre conglomerados no es significativa.

-1: significa que los clústeres se asignan de forma incorrecta.

2.5. Análisis factorial de datos mixtos

Es una técnica multivariante que permite estudiar la similaridades entre los individuos de un conjunto de datos tomando en cuenta todas las variables continuas y categóricas.

El término mixto se refiere al uso de variables tanto cuantitativas como cualitativas. A grandes rasgos, podemos decir que FAMD funciona como un análisis de componentes principales (PCA) para variables cuantitativas y como un análisis de correspondencia múltiple (MCA) para variables cualitativas.

2.6. Clusterings jerárquico

Es una herramienta exploratoria diseñada para descubrir agrupaciones naturales o conglomerados en un conjunto de datos que de otro modo serían indetectables. Tiene como objetivo agrupar clústeres en nuevos, de tal manera que cuando este proceso de aglomeración o escisión se realiza sucesivamente, se minimiza alguna distancia o se maximiza alguna similitud a generar (Husson et al., 2010).

Existen dos técnicas de clusterings jerárquico:

Técnicas aglomerativas: A partir del cual se empieza cada caso como un clúster individual para combinar el par de clústeres más cercano hasta quedar solo uno.

Técnicas divisivas: En donde se comienza con un solo clúster hasta que por división cada clúster tenga un único caso.

CAPÍTULO 3

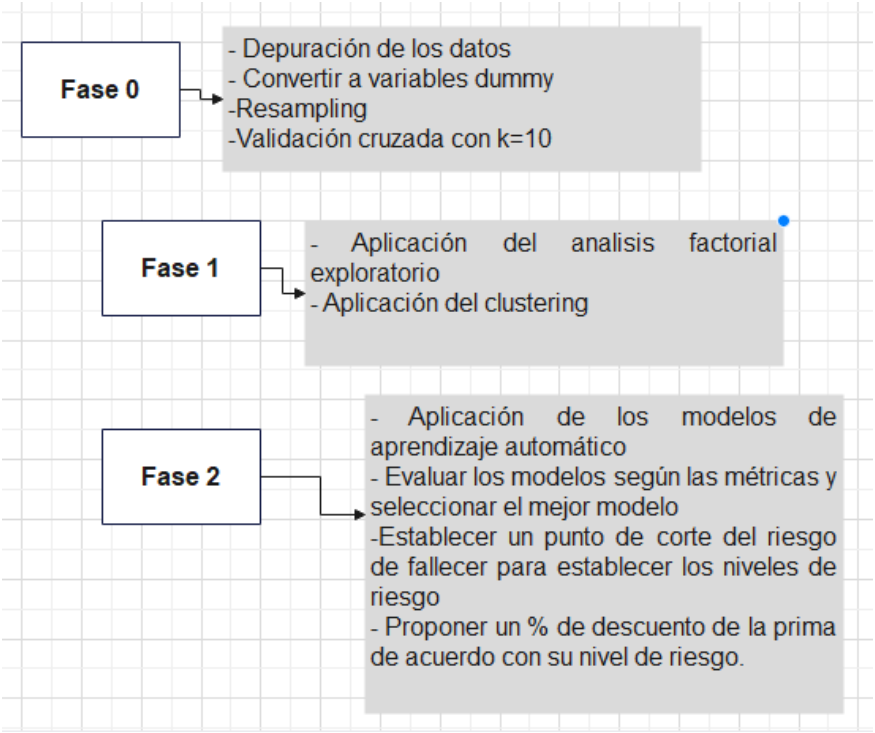
3. Metodología

En el presente capitulo se evidencia el proceso donde se establecen las fases del análisis, en la primera fase se realiza un análisis exploratorio y de clusterings con datos oficiales de homicidios intencionales, con el fin de obtener perfiles de clientes que sean más propensos a fallecer por cusas no naturales como homicidios.

En la segunda fase se usaron modelos de aprendizaje automático, que permitieron estimar la probabilidad de riesgo de fallecer de acuerdo con factores clínicos del posible cliente, una vez aplicados los modelos se evaluaron de acuerdo con las métricas de sensibilidad, especificidad, precisión y exactitud para poder seleccionar el mejor modelo que se ajuste a los datos. Adicionalmente, se describe el proceso completo en la ilustración 2

Ilustración 3

Proceso para la creación del scoring de seguro



3.1. Depuración de los datos

Los datos fueron previamente tratados con la limpieza de inconsistencias, valores faltantes, y de hacer una selección de los datos que correspondan con las áreas de estudios que son Guayaquil, Durán y Samborondón. Para los datos que corresponden a homicidios intencionales, se contabilizaron un total de 673 observaciones los cuales corresponden desde el 1 de enero hasta la última semana de junio del 2022.

Para los datos de registros clínicos se obtuvieron un total de 5122 observaciones.

3.2. Convertir las variables clínicas en variables dummy

En los modelos regresión las variables registran valores continuos comúnmente, es decir, suelen ser de tipo cuantitativa, y en los casos donde la variable es de tipo categórica normalmente con dos o más categorías, es importante utilizar variables indicadoras o dummy que permitan identificar de mejor manera a que clase pertenece cada observación, los valores por defecto que se toman siempre son 0 y 1 por lo que se crean tantas variables dummy como clases tenga la variable categórica.

En el caso de las variables clínicas, todas mostraban valores de si y no y no estaban establecidas como variables dummy, al realizar este tipo de transformación las variables como el domicilio y el sexo del cliente tomaron valores 0 y 1, siendo esto necesario para poder aplicar la regresión logística dentro del software estadístico R.

3.3. Resampling

En el primer análisis de los datos con variables clínicas, se pudo notar que existía un sesgo hacia una clase de la variable de respuesta, lo que hacía que el modelo fuera poco preciso para poder identificar y predecir de manera correcta a las personas que podrían fallecer, de acuerdo con su historial clínico con probabilidades muy cercanos a 0, por lo que se aplicó esta técnica para reducir el sesgo hacia aquella clase, la desventaja presente fue que se pasó de tener 5122 registros a tan solo 1814 registros, de tal manera que había la misma cantidad de fallecidos y no fallecidos, sin embargo, hizo que las métricas del modelo mejoren.

3.4. Validación cruzada

Como otra técnica para obtener parámetros estimados más precisos, se aplicó también a los datos originales de las variables clínicas una validación cruzada con k=10, esto se estableció como una muestra independiente de la obtenida en el proceso de resampling o remuestreo, con el fin de poder comparar las métricas obtenidas en los modelos luego de la aplicación de la validación cruzada y el resampling y en base a ellos obtener el mejor modelo.

3.5. Aplicación del análisis factorial

La aplicación del análisis factorial permitirá descubrir cuales son esos factores subyacentes, que permitan identificar el perfil de riesgo de un cliente, mostrando las características de varios grupos de individuos. Este análisis se realizará con los datos de homicidios intencionales de la zona 8 del Guayas.

3.6. Aplicación del clustering

La aplicación del clustering dentro del conjunto de observaciones de homicidios intencionales de la zona 8 del Guayas, permitirá identificar y segmentar a los posibles clientes de acuerdo con las características de las variables presente en los datos. Se utilizará un algoritmo de clustering que tiene como entrada datos tanto cuantitativos como cualitativos, se calculará el numero óptimo de clúster a obtener y se interpretará aquellos clústeres encontrados.

3.7. Aplicación de modelos de aprendizaje automático

Se aplicarán modelos de aprendizaje automático como: regresión lineal, random forest y máquina de vector soporte, como algoritmos de modelos de clasificación dado su variable de respuesta que es de tipo binaria. Se hará una partición de los datos en dos grupos, uno para el entrenamiento del modelo y el restante para el testeo, de esta manera se entrenará el modelo para obtener las estimaciones de los coeficientes mas adecuados para los datos utilizados, y, además, que sirvan para la entrada de nuevas observaciones y obtener así las probabilidades de riesgo de fallecer por cada individuo.

3.8. Evaluación y selección de los modelos

La evaluación de los modelos se realizará en base a las métricas de sensibilidad, especificidad, precisión y exactitud del modelo, aquel modelo que presente métricas más adecuadas será el seleccionado, sin embargo, las métricas de mayor importancia son la precisión del modelo para predecir verdaderos positivos, la sensibilidad del modelo y que tan complejo sea poder interpretar los parámetros estimados.

3.9. Punto de corte

Una vez seleccionado el modelo que se adecue mejor al conjunto de datos, se establecerá un punto de corte el cual sea el óptimo, tomando en cuenta que no afecte la sensibilidad y especificidad del modelo. Este punto de corte será utilizado para crear las categorías del scoring de seguro. El punto de corte será estimado de acuerdo con el modelo seleccionado usando la función optimalCutoff dentro del software de r.

3.10. Creación del scoring

La creación del scoring de seguro hace referencia a que porcentaje de descuento de la prima puede obtener el cliente, se propone realizarlo en tres niveles: el primer nivel corresponde a un probabilidad o riesgo de fallecer bajo, para el cual se propone que sea un descuento del 20%, para un cliente con riesgo de fallecer medio que se encuentra por debajo del punto de corte, que se le pueda ofrecer un descuento del 10% del costo de la prima, y por último, si el riesgo de fallecer del cliente es igual o superior al punto de corte, se lo considere de riesgo alto y no se aplica a ningún descuento al valor de la prima de la póliza del seguro de vida.

CAPÍTULO 4

4. Resultados

En este capítulo se redactan los resultados obtenidos en las aplicaciones de los modelos sobre los datos, donde se encontraron ciertas novedades que indicaban que el modelo necesitaba un ajuste adicional. Adicionalmente los resultados obtenidos en la aplicación del análisis exploratorio y clusterings.

En la primera aplicación de los modelos de: regresión logística, random forest y máquinas de vector soporte, se encontró un bajo nivel de sensibilidad, es decir, predecía de manera errónea los individuos que iban a fallecer, lo que representa que el modelo no era el adecuado dada la alta probabilidad de no estimar correctamente el riesgo para las aseguradoras, las cuales tendrían mayores registros de siniestros y mayor desembolso de dinero si no se aplicaba un ajuste al modelo.

Lo contrario se pudo observar en la especificidad del modelo, era muy preciso para predecir los casos en los que el individuo no iba a fallecer de acuerdo con su historial clínico, como consecuencia del gran sesgo que existía hacia esta clase. Esto era producido por la gran prevalencia de no fallecidos dentro de la muestra, por lo que se optó realizar una validación cruzada con k iteraciones, con k =10. Una vez realizada mejoró en precisión, la sensibilidad y especificidad en cada uno de los modelos.

Otra técnica también utilizada fue la de balancear la muestra, con un remuestreo de manera que existiría la misma cantidad de fallecidos y no fallecidos. La precisión del modelo se ajustó de mejor manera, e incluso el área debajo de la curva de ROC llegó al 81%, frente a 69% obtenido sin la aplicación del balanceo de la muestra y la validación cruzada.

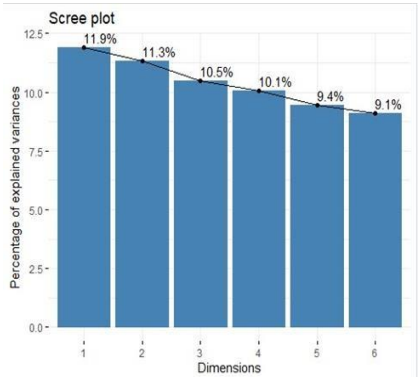
4.1. Aplicación del Análisis Factorial

La variable Franja fue creada a partir de la hora en que el crimen fue cometido, en la cual se estableció para el primer nivel “madrugada” desde las 0:00 horas hasta las 5:59 horas, para el segundo nivel “mañana” desde las 06:00 horas hasta las 11:59 horas, el tercer nivel “tarde” desde las 12:00 horas hasta las 18:59 horas y el cuarto nivel “noche” desde las 19:00 horas hasta las 23:59 horas.

Debido a que los datos de homicidios intencionales no solamente presentaron datos cuantitativos, sino también cualitativos se aplicó un análisis factorial mixto. Se utilizó la función FAMD () del paquete FactoMineR del software estadístico R, donde la variable continua edad se escaló a la varianza unitaria, y las variables categóricas se transformaron en una tabla de datos disyuntiva y luego se escalaron utilizando la escala específica del análisis de correspondencia múltiple

A partir de la ilustración 4 notamos que el porcentaje de varianza explicado entre las 2 primeras componentes es del 23.25%, mientras que las 6 componentes explican alrededor del 62% de la varianza.

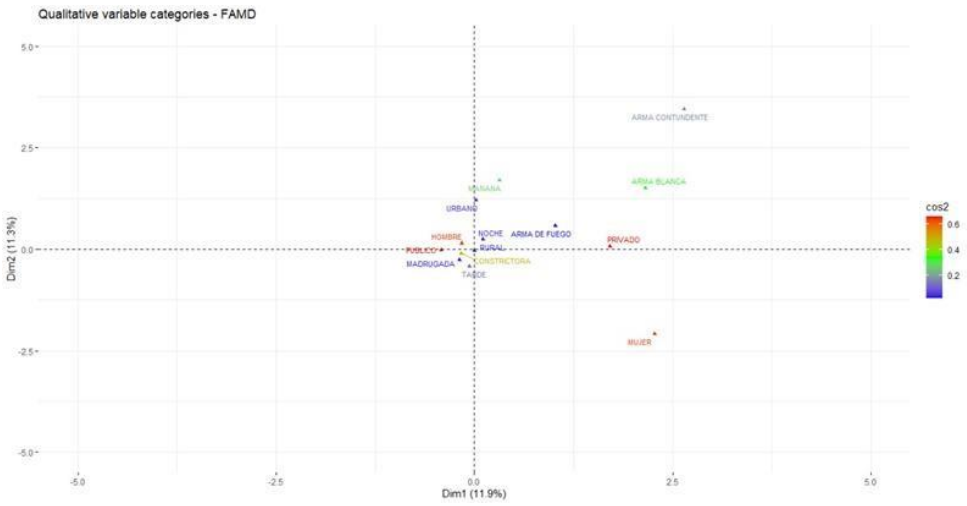
Ilustración 4



Nota. Porcentaje de Varianza explicado Análisis factorial mixto Fuente: Elaboración propia

La ilustración 5 nos muestra tanto el tipo de lugar urbano como rural están más correlacionados con la dimensión 2 y que el área del hecho: público y privado están mayormente correlacionados con la dimensión 1.

Ilustración 5



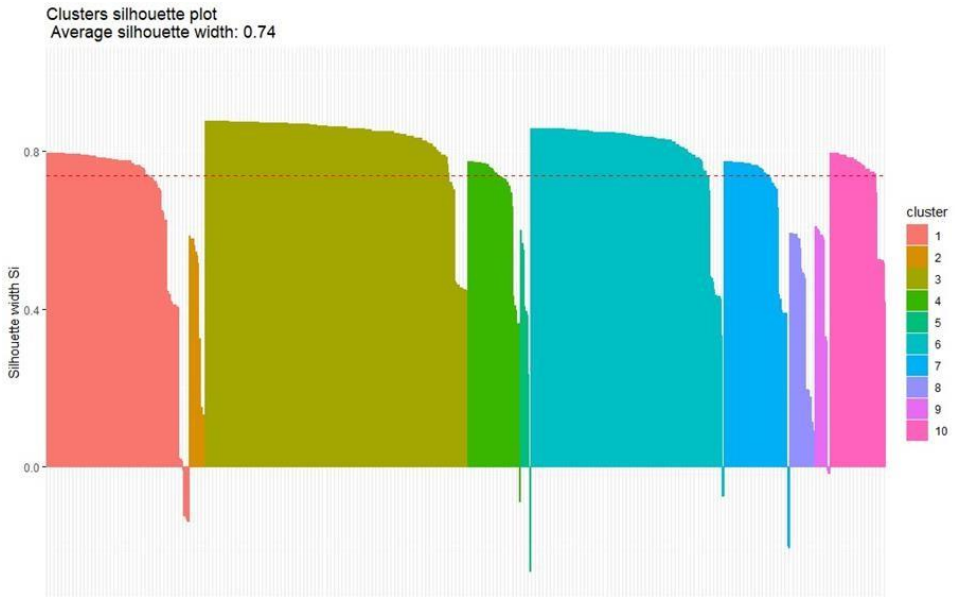
Nota. Mapa de calidad de representación FAMD Fuente: Elaboración propia

4.2. Aplicación del Clusterings jerárquico

Posteriormente se realizó un análisis de clúster donde se calculó la distancia de Gower mediante la función daisy () del paquete clúster, en donde se computó todas las disimilaridades por pares (distancias) entre las observaciones en el conjunto de datos, para luego utilizar la función pam () dándonos como resultado el número de clúster a escoger.

La ilustración 6 nos indica que para cada clúster el análisis Silhouette (Si), mide qué tan bien se agrupa una observación y estima la distancia promedio entre los conglomerados, dando un valor promedio de coeficiente de 0.74 lo cual está dentro del rango entre 0 y 1 concluyendo que están separados entre sí y claramente diferenciados.

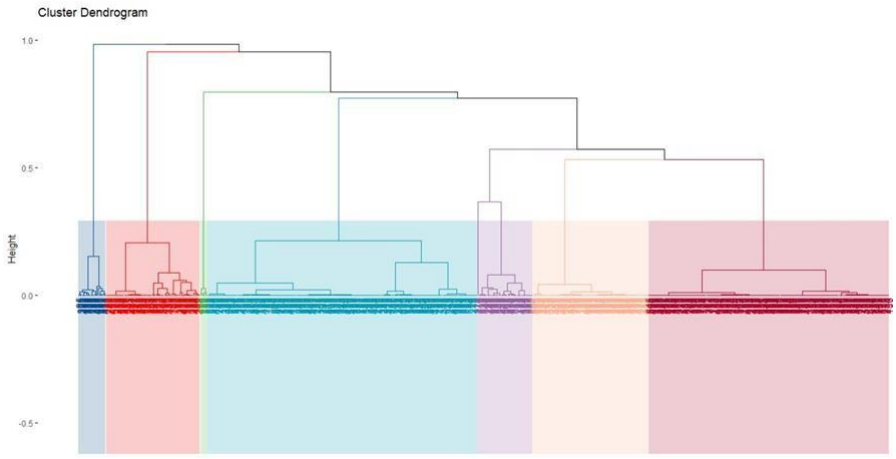
Ilustración 6



Nota. Gráfico de silueta. Particionamiento alrededor de Medoides Fuente: Elaboración propia

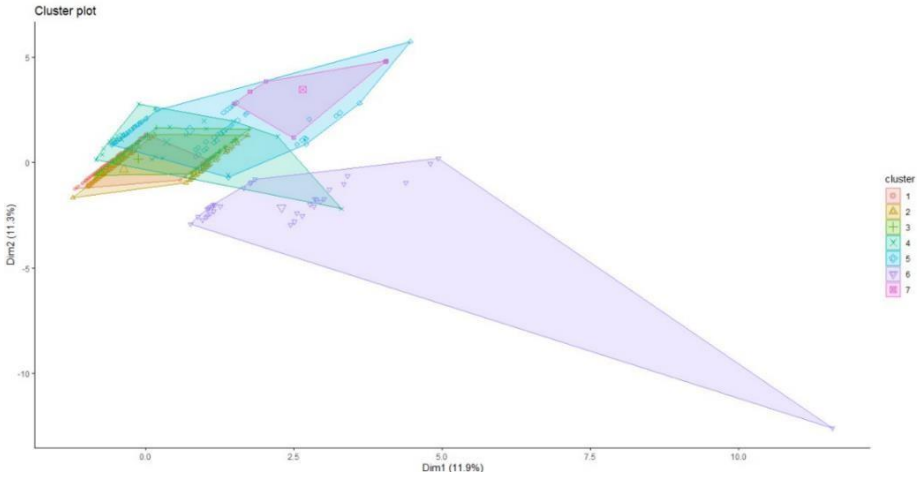
Finalmente se realizó el dendograma al análisis factorial mixto a través de la función HCPC mostrando como resultado 7 conglomerados, lo cual se puede visualizar en la ilustración 7 y la ilustración 8.

Ilustración 7



Nota. Dendograma de conglomerados Fuente: Elaboración propia

Ilustración 8



Nota. Gráfico de conglomerados Fuente: Elaboración propia

4.3. Aplicación de los modelos de aprendizaje automático

Los modelos de regresión logística, random forest y máquina de vector soporte, fueron aplicados tanto en la muestra de resampling, como en la de validación cruzada en el conjunto de datos de variables clínicas, por lo que se utilizaron un total de 6 modelos. Para cada uno de los modelos se aplicó una partición de los datos en un 80% datos de entrenamiento y 20% datos de testeo, con una misma semilla para uno de los modelos.

Una vez aplicado el modelo en el conjunto de datos de entrenamiento, se procedió a obtener las predicciones en base al conjunto de datos de testeo y así poder obtener la matriz de confusión la que permite calcular las métricas que servirán para seleccionar el mejor modelo.

Se obtuvieron 6 matrices de confusión las cuales fueron dividas en dos grupos: el primer grupo corresponde a las 3 matrices obtenidas con los modelos de aprendizaje automático en el conjunto de datos de muestra de resampling, el otro grupo contiene las 3 matrices restantes con el conjunto de datos de muestra de validación cruzada. Para estas matrices de confusión se define 0 como “No fallece” y 1 como “Fallece”.

Para la aplicación de la regresión logística se utilizó la función glm () en el software estadístico r, tomando como función de enlace Logit, la variable dependiente corresponde si la persona fallece o no, y como variables independientes las obtenidas de factor clínico, todas estas variables son categóricas de tipo binaria o dicotómicas.

En la aplicación del modelo de Random Forest se utilizó la librería de randomForest usando la función con el mismo nombre, con un total de 500 árboles de crecimiento para garantizar que cada fila de entrada se prediga al menos un par de veces, además de pedir al modelo que estime la importancia de las variables predictoras para un análisis posterior.

En la Máquina de Vector Soporte se utilizó la librería e1071, usando la función svm () además de establecer la función Kernel sigmoide, debido a que la variable de respuesta es de tipo binaria. También se le especifico que el modelo era de clasificación y con un costo de 20 como término de regularización en la formulación de Lagrange.

4.4. Matriz de confusión obtenida con el balanceo de la muestra

4.4.1. Regresión logística

En la tabla 3 se puede visualizar la matriz de confusión obtenida con resampling, en el cual se evidencia que este modelo no discrimina correctamente las personas que fallecen, ya que presenta casi la misma cantidad de predicciones erróneas como correctas en referencia si la persona fallece, lo contrario ocurre con las predicciones sobre las personas que no fallecerán.

Tabla 3

Matriz de confusión obtenido para regresión logística

		Predicho	
		1	**0**
Real	**1**	99	86
	0	24	153

4.4.2. Random Forest

En la tabla 4, para la clase de fallecido el modelo predice correctamente 135 observaciones de verdaderos positivos del total de 188 observaciones, un resultado similar se obtiene con la clase de no fallecidos, 120 observaciones de verdaderos negativos correctamente predichas del total de 173 observaciones.

Tabla 4

Matriz de confusión obtenida para Random Forest

		Predicho	
		1	**0**
Real	**1**	135	54
	0	53	120

4.4.3. Máquina de Vector Soporte

En la tabla 5, se puede evidenciar que este modelo predice correctamente 102 observaciones de verdaderos positivos, del total de 194 observaciones, en el caso de verdaderos negativos el modelo estimó correctamente 177 observaciones del total de 207 observaciones

Tabla 5

Matriz de confusión para Máquina de Vector Soporte

		Predicho	
		1	**0**
Real	**1**	102	30
	0	92	177

Una vez obtenidas las matrices de confusión de la muestra con resampling o remuestreo, se pudieron obtener las métricas para evaluar y seleccionar los modelos. En cuanto a especificidad, el modelo de regresión logística predice de mejor manera la especificidad que los otros modelos, sin embargo, en cuanto a la tasa de verdaderos positivos es el

peor modelo. Las demás métricas obtenidas con remuestreo de la muestra se pueden visualizar en la tabla 6.

Tabla 6

Métricas para la selección del modelo con balanceo de la muestra

Modelo	Sensibilidad	Exactitud	Especificidad	Precisión
GLM	80%	70%	64%	54%
Random Forest	72%	70%	69%	71%
Máquina Vector Soporte	53%	69%	86%	77%

4.4.4. Matriz de confusión obtenida con Validación cruzada k =10

4.4.5. Regresión Logística

En la tabla 7, se puede visualizar la matriz de confusión obtenida con validación cruzada, en el cual se evidencia que este modelo obtiene una gran mejora para predecir la tasa de verdaderos positivos, pero en base de la tasa de verdaderos negativos también el modelo prediciendo correctamente 10 de 14 observaciones.

Tabla 7

Matriz de confusión obtenido para regresión logística

		Predicho	
		1	**0**
Real	**1**	333	4
	0	63	10

4.4.6. Random Forest

En la tabla 8, se puede visualizar la matriz de confusión obtenida con validación cruzada, parece converger a las mis predicciones que el modelo de regresión logística predice correctamente 334 observaciones de un total de 397 observaciones la tasa de verdaderos positivos, y en cuanto a la tasa de verdaderos negativos predice correctamente 10 observaciones de 11.

Tabla 8

Matriz de confusión obtenido para Random Forest

		Predicho	
		1	**0**
Real	**1**	334	1
	0	63	10

4.4.7. Máquina de Vector Soporte

En la tabla 9, se puede visualizar la matriz de confusión obtenida con validación cruzada, parece converger también a las mis predicciones que el modelo de regresión logística predice correctamente 336 observaciones de un total de 403 observaciones la tasa de verdaderos positivos, y en cuanto a la tasa de verdaderos negativos predice correctamente 5 observaciones de 6.

Tabla 9

Matriz de confusión para Máquina de Vector Soporte

		Predicho	
		1	**0**
Real	**1**	336	1
	0	67	5

Los resultados obtenidos de aplicar la validación cruzada a los modelos, registra las siguientes métricas calculas en la tabla 10.

Tabla 10

Métricas para la selección del modelo con validación cruzada

Modelo	Sensibilidad	Exactitud	Especificidad	Precisión
GLM	84%	84%	71%	99%
Random Forest	84%	83%	91%	100%
Máquina Vector Soporte	83%	84%	83%	100%

4.4.8. Selección del modelo

Luego de la aplicación de la técnica de balanceo de la muestra, se puede notar que el modelo de regresión logística presentó una tasa de especificidad del 64%, el menor de los tres modelos, lo contrario ocurre con la sensibilidad al balancear la muestra mostrando una tasa del 80%. En cuanto a la tasa de verdaderos positivos predichos correctamente obtuvo un 54%, siendo el menos preciso de los tres modelos.

La exactitud presente en los 3 modelos es relativamente igual, sin embargo, la precisión resultante del modelo de Máquina de Vector Soporte es la más alta siendo del 77%, y de igual manera la especificidad de este modelo, sin embargo, la sensibilidad del modelo es la más baja comolo podemos verificar en la tabla 9.

En comparación con la técnica de la validación cruzada con k =10, se obtuvo una mejora sustancial en cuanto a las métricas, sin embargo, la exactitud de la modelo obtenida por la validación cruzada no puede ser comparada con la obtenida en el balanceo de la muestra, debido a que esta no es realmente significativa cuando existe una prevalencia alta de una clase sobre la otra, como se puede observar en la tabla 10.

4.4.9. Modelo Seleccionado

En la selección del modelo adecuado de acuerdo con las métricas de sensibilidad y precisión, se escoge al modelo de regresión logística por su ajuste adecuado de los datos, además de poseer una menor complejidad para interpretar los parámetros estimados. Los coeficientes del modelo se presentan a continuación:

$$Y = -1.76 - 12.44 * Samborondon + 1.64 * Guayaquil - 1.38 * Duran + 1.30 * alergia + 0.16 * Enfermedades\ autoinmunes + 0.82 * Covid - 0.09 * Sexo + 1.35 * amas + 1.07 * Enfermedades\ cardiovasculares + 0.98 * Diabetes$$

De acuerdo con el resumen del modelo las variables que no resultaron ser significativas son: si el individuo vive en Samborondón, adicional si el individuo tiene o tuvo enfermedades autoinmunes y si presenta diabetes. El intervalo de confianza de los odds ratios se puede apreciar en la tabla 11.

Tabla 11

Intervalo de confianza de los odds ratios

	2.5%	**97.5%**
Intercepto	0.1121308	2.555277e-01
Covid1	1.7085092	3.055911e+00
Samborondon1	NA	2.180488e+23
Duran1	0.1667390	3.674857e-01
Guayaquil1	3.5442449	7.719397e+00
Sexo1	0.7166771	1.159350e+00
Alergia1	2.6076399	5.282509e+00
Asma1	1.6248796	1.028005e+01
enfermedades_autoinmunes1	0.4983080	2.851654e+00
enfermadades_cardiovasculares1	1.7077016	5.097725e+00
diabetes1	0.8774319	1.024523e+01

4.4.10. Resultados obtenidos del análisis factorial mixto y del clustering

De los 7 grupos encontrados, el primer grupo está representado por los hombres de 30 años asesinados con arma constrictora durante la madrugada en lugares públicos y en zona rurales, el segundo caracterizado por hombres de 32 años asesinados con arma constrictora durante la tarde, el tercero representado por hombres de 31 años asesinados con arma constrictora durante la noche.

El cuarto grupo representado por hombres con edades de 29,31,35,37 y 41 años asesinados en la madrugada en zonas urbanas con arma constrictora y arma de fuego, el quinto grupo con hombres de 31,33 y 35 años asesinados durante la mañana con arma blanca y arma constrictora mientras que el sexto grupo se vio representado por mujeres de 12,15,21 y 30 años asesinadas con arma constrictora durante la tarde y finalmente el séptimo conglomerado representado por hombres mayores de 50 años asesinados en la noche con arma contundente.

4.4.11. Creación del scoring

El scoring de seguro de vida se establecerá de acuerdo con la probabilidad de fallecer, que corresponde al riesgo obtenida del modelo de regresión logística, este scoring es de tipo cualitativo y contará con 3 niveles que sugieren un porcentaje estimado del seguro de vida, de acuerdo con los factores clínicos del posible cliente como se puede ver en la tabla 12.

Tabla 12

Creación del scoring de seguros

Riesgo	% de descuento estimado
○ Bajo, si la probabilidad de fallecer está por debajo del 0.40	20%
○ Medio, si la probabilidad de fallecer se encuentra entre 0.41 y 0.52	10%
○ Alto, si la probabilidad de fallecer es mayor a 0.53	0%

CAPÍTULO 5

5. Conclusiones y Recomendaciones

5.1. Conclusiones

- Para la selección del modelo se establece que es de mayor importancia a la hora de asegurar al posible cliente, sin embargo, la exactitud del modelo es realmente importante a considerar, a la par con la sensibilidad y la precisión. Es de ahí que el mejor modelo corresponde a la regresión logística dada las métricas encontradas, además de ser menos complejo poder interpretar los factores estimados, esto lo podemos ver tanto al aplicar el balanceo de la muestra como en la validación cruzada, tal como se pueden observar con los resultados obtenidos en la tabla 9 y tabla 10.
- Del conjunto de variables aplicadas dentro del modelo de regresión logística, se pudo evidenciar que no eran significativas las siguientes: si lugar de domicilio es Samborondón, el sexo de la persona, si presenta diabetes o no, y si presenta o ha presentado enfermedades autoinmunes.
- Luego del análisis exploratorio realizado con los datos de homicidios intencionales, los grupos de edad más propensos a ser asesinados corresponden al rango de 30 a 41 años para hombres, siendo de mayor ocurrencia en horas de la madrugada.
- El grupo de mujeres con más casos de homicidios intencionales corresponde a mujeres de entre 15 y 21 años con una mayor ocurrencia en las tardes.
- La aplicación de modelos estadísticos resultan ser una ayuda importante para la explicación y resolución de problemáticas de cualquier nivel de la sociedad y de diversas áreas.

5.2. Recomendaciones

- Si las aseguradoras disponen de una información más completa y de uso libre, el modelo seguramente el modelo se ajustaría de mejor manera con la realidad.
- Que exista un registro de acceso público que registre mayores datos sobre el fallecimiento de una persona, como si es de causa natural, por enfermedad, por accidente de tránsito, accidente laboral entre otros, el scoring podría ser más preciso y se podría segmentar de otra manera a los clientes.

6. Referencias

Gallego, Ó. M. (2011). *Modelo matemático paramétrico de estimación para proyectos de data mining*. Recuperado el 17 de 7 de 2022, de http://oa.upm.es/282

Gobierno, M. d. (mayo de 2022). *Ministerio de Gobierno de Ecuador*. Obtenido de http://cifras.ministeriodegobierno.gob.ec/comisioncifras/inicio.php#

Gonzáles, M. A. (24 de Septiembre de 2021). *Primicias.Ec*. Obtenido de https://www.primicias.ec/noticias/sociedad/ecuador-tasa-muertes-violentas-alta/

Juan Ochoa, W. G. (25 de Octubre de 2010). Construcciòn de un modelo de scoring para el otorgamiento de crédito en una entidad financiera. *Scielo.org*, 1-32. Obtenido de http://www.scielo.org.co/pdf/pece/n16/n16a10.pdf

Meléndez, Á. (2022). La fusión de dos grandes robustece el mercado de seguros en Ecuador. *Bloomberg Línea*, 1.

Serrano, A. G. (2012). *Intenligencia Artificial Fundamentos, práctica y aplicaciones.* Madrid, España: Grupo RC. doi:978-84-939450-2-2

Vallejo, P. D., & Tenelanda, V. G. (2012). MINERÍA DE DATOS APLICADA EN DETECCIÓN DE INTRUSOS. *Ingenierías USBMed, 3*(1), 50-61. Recuperado el 17 de 7 de 2022, de http://web.usbmed.edu.co/usbmed/fing/v3n1/v3n1a6.pdf

Ávila-Tomás, J. F., Mayer-Pujadas, M. A., & Quesada-Varela, V. J. (2021). La inteligencia artificial y sus aplicaciones en medicina II: importancia actual y aplicaciones prácticas. *Atención Primaria*, *53*(1), 81–88. https://doi.org/10.1016/J.APRIM.2020.04.014

Cerda, J., & Cifuentes, L. (2012). Uso de curvas ROC en investigación clínica: Aspectos teórico-prácticos. *Revista Chilena de Infectología*, *29*(2), 138–141. https://doi.org/10.4067/S0716-10182012000200003

Díaz Díaz, B., Sanfilippo Azofra, S., & López Gutiérrez, C. (2008). Influencia de la prima sobre la creación de valor en las fusiones y adquisiciones bancarias en Europa. *Cuadernos de Economía y Dirección de La Empresa*, *11*(34), 81–106. https://doi.org/10.1016/S1138-5758(08)70054-9

Espinosa Zúñiga, J. J. (2020). Aplicación de metodología CRISP-DM para segmentación geográfica de una base de datos pública. *Ingeniería, Investigación y Tecnología*, *21*(1), 1–13. https://doi.org/10.22201/FI.25940732E.2020.21N1.008

Fiuza Pérez, M. D., & Rodríguez Pérez, J. C. (2000). La regresión logística: una herramienta versátil. *Nefrología*, *20*(6), 495–500. https://www.revistanefrologia.com/es-la-regresion-logistica-una-herramienta-articulo-X0211699500035664

García Díaz, E. E., & Lozano Martínez, F. (2006). Máquinas de vectores de soporte. *Revista de Ingeniería*, *24*, 62–70. http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S0121-49932006000200008&lng=en&nrm=iso&tlng=es

Ledesma, R. (2008). Introduccción al Bootstrap. Desarrollo de un ejemplo acompañado de software de aplicación. *Tutorials in Quantitative Methods for Psychology*, *4*(2), 51–60.

Leis, A., Ronzano, F., Mayer, M. A., Furlong, L. I., & Sanz, F. (2019). Detecting Signs of Depression in Tweets in Spanish: Behavioral and Linguistic Analysis. *J Med Internet Res 2019;21(6):E14199 Https://Www.Jmir.Org/2019/6/E14199*, *21*(6), e14199. https://doi.org/10.2196/14199

Luis Enrique, C.-G., Maya, C.-R., Victor Giovanni, M.-M., & López José Gustavo, L. (n.d.). *Validación de un algoritmo de clasificación para la identificación de interacciones farmacológicas Validation of a classification algorithm for identifying pharmacological interactions*. https://doi.org/10.22201/fi.25940732e.2019.20n2.014

María Ávila Rodríguez. (2020). *Análisis de tweets y su influencia en los seguros de vida en el ámbito colombiano* [Universidad Politécnica de Valencia]. https://riunet.upv.es/bitstream/handle/10251/150168/%C3%81vila%20-%20An%C3%A1lisis%20de%20tweets%20y%20su%20influencia%20en%20los%20seguros%20de%20vida%20en%20el%20%C3%A1mbito%20colombiano..pdf?sequence=1

María Paula Ávila Rodriguez. (2020). *Análisis de Tweets y su influencia en los seguros de vida en el ámbito colombiano* [Universidad Politécnica de Valencia]. https://riunet.upv.es/bitstream/handle/10251/150168/%C3%81vila%20-%20An%C3%A1lisis%20de%20tweets%20y%20su%20influencia%20en%20los%20seguros%20de%20vida%20en%20el%20%C3%A1mbito%20colombiano..pdf?sequence=1&isAllowed=y

MARIO DAVID SOLÓRZANO CARVAJAL. (2016). *CONSTRUCCIÓN DE TABLAS DE MORTALIDAD PARA LA REGIÓN SIERRA DEL ECUADOR 2010. UN ANÁLISIS DE SUPERVIVENCIA DE LA POBLACIÓN* [ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL]. https://www.dspace.espol.edu.ec/bitstream/123456789/37115/1/T-102471.pdf

Misra, S., & Li, H. (2020). Noninvasive fracture characterization based on the classification of sonic wave travel times. *Machine Learning for Subsurface Characterization*, 243–287. https://doi.org/10.1016/B978-0-12-817736-5.00009-0

Nationale Nederlanden. (n.d.). *Entendiendo los seguros: ¿qué es un siniestro?* Retrieved September 8, 2022, from https://www.nnespana.es/blog/seguros/sabias-que/entendiendo-los-seguros-que-es-un-siniestro

Omar, K. S., Islam, M. N., & Khan, N. S. (2021). Exploring tree-based machine learning methods to predict autism spectrum disorder. *Neural Engineering Techniques for Autism Spectrum Disorder: Volume 1: Imaging and Signal Analysis*, 165–183. https://doi.org/10.1016/B978-0-12-822822-7.00009-0

Ponce Cruz, P. (2010). ADALINE (Adaptive Linear Neuron). *Inteligencia Artificial Con Aplicaciones a La Ingeniería*, *1*(Primera Edición), 202–205.

Ravelo, M. C., & Ponsot Balaguer, E. (n.d.). *Funciones de enlace alternativas en modelos de respuesta binomial*.

Segura Egea, J. J. (2002). Sensibilidad y especificidad de los métodos diagnósticos convencionales de la caries oclusal según la evidencia científica disponible. *RCOE*, *7*(5), 491–501. https://scielo.isciii.es/scielo.php?script=sci_arttext&pid=S1138-123X2002000600004&lng=es&nrm=iso&tlng=es