UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA
DEPARTAMENTO DE INDUSTRIAS
VALPARAÍSO - CHILE

ANÁLISIS DE FACTORES DE RIESGOS EN LOS EGRESOS HOSPITALARIOS DE CHILE ENTRE LOS AÑOS 2011-2018

VALENTINA CAMILA CONSTANZA AGUILAR CONCHA

MEMORIA PARA OPTAR AL TÍTULO DE
INGENIERA CIVIL INDUSTRIAL

PROFESOR GUÍA	:	RAFAEL FAVEREAU URQUIZA
PROFESOR CORREFERENTE	:	MÓNICA LÓPEZ CAMPOS.

OCTUBRE 2020

AGRADECIMIENTOS

Con este proyecto de tesis culmina una larga etapa de aprendizaje y crecimiento personal, que estuvo colmado de buenos momentos, alegrías y también de sacrificios. En primer lugar, quiero agradecer al profesor Rafael Favereau por el entusiasmo con que imparte sus clases y la motivación que logra transmitir, fue una de las razones por la cual decidí que fuera mi guía en este proceso. Quiero subrayar que es una gran inspiración como docente y como persona, me llevo grandes enseñanzas y siento un gran privilegio de haber trabajado con él como una de sus ayudantes y ser una de sus memoristas.

Esta investigación nace en conjunto con las ideas de René Miranda, quien fue un gran orientador y motivador para adentrarme en el área de la salud, me convenció de la importancia de lo que haría, y me ayudó a trazar el rumbo de la investigación.

Agradecer también a don German Goñi, a su tremenda voluntad y disposición de responder dudas cada instancia que lo necesitaba, le agradezco la ayuda desinteresada, su tiempo y conocimiento transmitido. Asimismo, quiero destacar la voluntad de don Juan Godoy, quién me brindó su ayuda al instante que la solicité, me otorgó acceso de manera remota a los equipos de la universidad, y siempre estuvo pendiente si tenía algún nuevo requerimiento para resolverlo a la brevedad.

Reconocer la labor de mi padre, que gracias a su esfuerzo pude estudiar sin distracciones y siempre tuve todo lo que necesitaba, también quiero reconocer la importancia de mis hermanos Marcelo y Leonardo, y a mis enanos Martín y Bruno que siempre me empujaron a seguir con su amor, alegría y palabras de aliento. Un abrazo gigante al cielo para mi abuelita que mientras pudo me regaloneó y siempre estuvo a mi lado , no me quedan dudas que el día que reciba mi título estará sentada en primera fila aplaudiendo y compartiendo la alegría del momento. Mi pequeña familia, mi principal motor de cada día.

A mis amigos de la vida y los que hice durante la universidad, que siempre me envíaban sus buenas vibras y me acompañaron y alegraron todo el tiempo, quienes me perdonaban las veces que no podía juntarme con ellos.

A todos, de corazón, muchas gracias.

RESUMEN EJECUTIVO

El uso adecuado de los medicamentos implica que el paciente reciba cada medicamento en la dosis correcta durante el tiempo adecuado de tratamiento y al menor costo posible, no obstante, cumpliendo estas características el paciente no esta exento de sufrir efectos no deseados, ya que no se puede predecir el efecto de un medicamento en los distintos organismos. Es por ello, que el presente estudio tiene como finalidad conocer si existen factores relacionados a los pacientes que signifiquen un mayor riesgo de sufrir Eventos Adversos a Medicamentos (EAMs), ya sean por factores biológicos o del entorno donde se desarrollan.

En primer lugar, para conocer el universo de pacientes que han sufrido daño por el consumo o abuso de medicamentos, se utilizan los registros electrónicos de egresos hospitalarios publicados por el Departamento de Estadística e Información de Salud del Ministerio de Salud, donde a través del diagnóstico principal y/o secundario se determinan los registros que están asociados a EAMs, esto se logra gracias a que los diagnósticos responden al estándar de Clasificación Internacional de Enfermedades (CIE), y a que Jürgen Stausberg, bajo cuidadosos criterios de selección, propone un grupo de códigos de la CIE que indican relación con los EAM . Con esta metodología se determina la variable de respuesta de esta investigación, correspondiente a una variable cualitativa de clases binarias que responde a 1 si un egreso es atribuible al daño por medicamentos y, por otro lado, adquiere el valor 0 si el registro no es atribuible a EAM. Para reunir el resto de las variables, se utiliza la información de los registros electrónicos y el riesgo del entorno se aproxima a través de la comuna de residencia del paciente, caracterizando cada egreso a través del índice de pobreza, superficie y población comunal, dado que son las bases de datos disponibles con la mejor calidad de registro. También, se incluyen al análisis el número de recintos de salud de cada nivel de atención, entendiéndolos como el acceso a la salud que posee una zona, y se agrega el número de farmacias y almacenes farmacéuticos por comuna, asociándolos a la disponibilidad y acceso a los medicamentos.

En segundo lugar, para encontrar las asociaciones entre el daño por medicamentos y las variables de interés, se utilizan algoritmos de minería de datos correspondientes a la regresión logística y bosques aleatorios, ambos con gran trayectoría en la clasificación. Se ajustan ambos modelos “a secas” con todas las variables recopiladas y se utiliza la eliminación de las variables menos influyentes en cada iteración, midiendo el rendimiento a través de la capacidad discriminativa, es decir, la capacidad de diferenciar sujetos de la clase 0 y sujetos de la clase 1, donde se obtiene una mala capacidad discriminativa en cada una de las iteraciones evidenciando un claro sesgo a clasificar solo la clase mayoritaria dada la distribución de las clases: 0.68 % para la clase 1 y 99.32 % para la clase 0. Para solucionar este problema, se utiliza la metodología de ajustar parámetros propios de los algoritmos para balancear las clases durante el entrenamiento de cada modelo, sin embargo esta técnica no es capaz de mejorar los resultados, razón por la cual se agrupan algunas variables con el objetivo de reducir la dimensionalidad ya que podría ser un factor causante de los malos ajustes. En el caso de la regresión logística se obtienen leves mejorías en el rendimiento alcanzando un desempeño regular,

pero en bosques aleatorios continúa el mal desempeño discriminativo.

En consecuencia, se aplica una segunda técnica para mejorar el desbalance de las clases ya que las mejoras implementadas no resultan significativas, se elije la técnica de sobremuestreo de minorías sintéticas probando el rendimiento de los modelos con distintos niveles de la clase minoritaria. Cuando se cuenta con un 10 % de dicha clase se obtienen grandes cambios en los rendimientos de ambos modelos, validando que el principal problema del conjunto de datos es el desbalance entre las clases. Adicionalmente, dada la gran magnitud de algunos coeficientes obtenidos en la regresión logística, se sospecha que responden a la elección de la categoría de referencia dado que representan una porción muy pequeña de los datos, por lo que se cambian las categorías de comparación para la variable *sexo* y *previsión*. Se ajustan nuevamente los algoritmos con las muestras sintéticas, nuevas variables y nuevas categorías, obteniendo un 75 % de rendimiento para la regresión logística, y un 93 % en el caso de bosques aleatorios, indicando una buena capacidad discriminativa en ambos casos.

En el caso de la regresión logística, se descubre que existe 2.61 veces más riesgo de sufrir EAM, que no sufrirlos, siendo mujer. Asimismo, se desprende que todas las categorías de la previsión de salud resultan ser un factor protector, sin embargo, de acuerdo a la magnitud de los coeficientes, se tiene que el tramo B de FONASA es el que presenta más riesgo dentro de este grupo, y por el otro extremo, el tramo D de FONASA (grupo con mayores recursos económicos) es el que presenta menos riesgo en comparación al resto de los tramos y otras previsiones de salud, cabe señalar que el resto de las variables no resultan relevantes. Con respecto al algoritmo de bosques aleatorios se determina que la edad es el factor más relevante a la hora de detectar el daño por medicamentos, concluyendo que el peak de casos asociados a EAM ocurren durante la adolescencia, y en un menor grado, durante los cuatro primeros años de vida de una persona, asimismo se desprende que mientras más avanzada la edad de un paciente, menor es la frecuencia de casos EAM. Por otra parte, se determina que el mes de ingreso también es un factor relevante, siendo febrero donde se presenta una menor cantidad de ingresos debido al daño por medicamentos, mientras que, la mayor cantidad de ingresos ocurren durante los meses de agosto, octubre y noviembre. En resumen, se tiene que las variables sexo, edad y mes de ingreso son las variables de mayor riesgo que se deben considerar para generar las acciones sanitarias a fin de reducir los ingresos hospitalarios por el consumo o abuso de medicamentos.

Finalmente, para estudios posteriores con este conjunto de datos, se recomienda reducir la clase mayoritaria bajo el estudio de la composición de los grupos de diagnósticos, a fin de extraer del análisis los egresos que no tienen relación con el consumo de medicamentos, generando al menos que un 10 % de los datos represente los EAM. También, es necesario mencionar que no existe justificación alguna para buscar el equilibrio perfecto entre las clases, ya que con un 30 % de los datos representando al grupo minoritario, los modelos expuestos convergen a su máximo rendimiento. Si bien es difícil lograr este valor, estudios han demostrado que utilizar la reducción de la clase mayoritaria y el sobremuestreo de las minorías en conjunto generan grandes resultados.

.

Índice de Contenidos

1. INTRODUCCIÓN **1**

2. PROBLEMA DE INVESTIGACIÓN **3**

3. OBJETIVOS **6**
3.1. Objetivo General . . . 6
3.2. Objetivos Específicos . . . 6

4. MARCO TEÓRICO **7**
4.1. Sistema de salud chileno . . . 7
4.1.1. Estructura . . . 7
4.1.2. Situación nacional actual . . . 7
4.1.3. Caracterización Comunal . . . 9
4.1.4. Marco Institucional . . . 10
4.1.4.1. Sector Público . . . 10
4.1.4.2. Sector Privado . . . 16
4.2. Los medicamentos en Chile . . . 17
4.2.1. Antecedentes del mercado de medicamentos . . . 17
4.2.2. ¿Qué son los establecimientos farmacéuticos? . . . 18
4.2.2.1. Comercialización . . . 19
4.2.3. Ética del mercado farmacéutico . . . 20
4.3. Minería de datos . . . 22
4.3.1. Métodos y tipos de tareas . . . 23
4.3.2. Algoritmos de clasificación . . . 23
4.3.2.1. Regresión logística . . . 24
4.3.2.2. Bosques Aleatorios . . . 25
4.3.3. Herramientas utilizadas . . . 27
4.3.4. Métricas de evaluación . . . 27
4.3.5. Problemas en la clasificación . . . 30
4.4. Fuentes de información . . . 31
4.4.1. ¿Cómo medir la calidad de los datos? . . . 32

5. METODOLOGÍA **34**
5.1. Comprensión del negocio . . . 34
5.2. Comprensión de los datos . . . 34
5.3. Preparación de los datos . . . 37
5.4. Modelado . . . 38
5.5. Evaluación y Despliegue . . . 38

6. RESULTADOS **39**
6.1. Comprensión del negocio . . . 39
6.1.1. Archivos con origen en SINIM . . . 39

6.1.2. Distribución de farmacias . . . 40
6.1.3. Listado Establecimientos . . . 40
6.1.4. Códigos Únicos Territoriales . . . 41
6.1.5. Listado de códigos EAM . . . 41
6.1.5.1. Listado de códigos según Wu et al. . . . 42
6.1.5.2. Listado de códigos según Stausberg . . . 42
6.1.6. Informes Estadísticos de Egresos Hospitalarios . . . 43
6.2. Comprensión de los datos . . . 45
6.3. Preparación de los datos . . . 48
6.3.1. Relación entre las bases de datos . . . 49
6.4. Modelado . . . 52
6.4.1. Regresión Logística . . . 53
6.4.2. Bosques Aleatorios . . . 61
6.5. Análisis del modelado . . . 64
6.6. Modelado: Segunda parte . . . 64
6.6.1. Regresión Logística . . . 64
6.6.2. Bosques Aleatorios . . . 66
6.7. Análisis del modelado . . . 68
6.8. Modelado: Tercera Parte . . . 69
6.8.1. Regresión Logística . . . 71
6.8.2. Bosques Aleatorios . . . 73
6.9. Análisis del modelado . . . 76
6.10. Modelado: Cuarta Parte . . . 76
6.10.1. Regresión Logística . . . 77
6.10.2. Bosques Aleatorios . . . 78
6.11. Evaluación y Despliegue . . . 81

7. CONCLUSIONES **82**

8. DISCUSIÓN **86**

Bibliografía **88**

A. ANEXOS **93**
A.1. Marco Teórico . . . 93
A.1.1. Distribución de población según situación de afiliación a sistema previsional de salud 93
A.1.2. La farmacia en Chile . . . 94
A.1.3. Metodología CRISP-DM . . . 94
A.2. Metodología . . . 95
A.2.1. Diccionario de datos SINIM . . . 95
A.2.2. Listados de códigos *diag1* y *diag2* según DEIS . . . 95
A.3. Resultados . . . 95
A.3.1. Comprensión de los datos . . . 95
A.3.2. Modelado . . . 99
A.3.2.1. Regresión logística: Variables antiguas . . . 99
A.3.2.2. Regresión logística: Variables nuevas . . . 102
A.3.2.3. Bosques Aleatorios: Variables antiguas . . . 109
A.3.2.4. Bosques Aleatorios: Variables nuevas . . . 111
A.3.2.5. Regresión Logística: Categorías nuevas . . . 113
A.3.2.6. Bosques Aleatorios: Categorías nuevas . . . 113
A.4. Códigos programación en Python . . . 114
A.4.1. Medición calidad de datos Egresos Hospitalarios . . . 114
A.4.2. SMOTE Regresión logística . . . 124
A.4.3. SMOTE Bosques Aleatorios . . . 127

Índice de Tablas

4.1. Clasificación de Establcimientos según complejidad y cobertura 11
4.2. Matriz de confusión 28
4.3. Matriz de confusión clases binarias 28

6.1. Esquema de registro Egreso Hospitalario 44
6.2. Datos válidos en archivos SINIM 45
6.3. Datos inválidos IEEH (%) 46
6.4. Resultados inconsistencia IEEH (%) 47
6.5. Registros útiles Egresos Hospitalarios 48
6.6. Métricas Regresión logística sin balancear 53
6.7. Métricas Regresión logística balanceada 54
6.8. Métricas tras primera reducción de dimensionalidad 55
6.9. Métricas Regresión logística tras reducción 57
6.10. Puntuación variables influyentes 57
6.11. Variables Statsmodels Final (primera parte) 60
6.12. Variables Statsmodels Final (segunda parte) 60
6.13. Métricas tras reducción de dimensionalidad, variables nuevas 65
6.14. Puntuación variables influyentes (segunda parte) 66
6.15. Rendimiento con muestras sintéticas 69
6.16. Métricas Regresión logística con muestras sintéticas 71
6.17. Variables Regresión logística con muestras sintéticas 73
6.18. Métricas Bosques aleatorios con muestras sintéticas 73
6.19. Variables Bosques Aleatorios con muestras sintéticas 75
6.20. Métricas Regresión logística nuevas categorías 77
6.21. Variables Regresión logística con muestras sintéticas y nuevas categorías 77
6.22. Métricas Bosques aleatorios nuevas categorías 78
6.23. Variables Bosques Aleatorios con nuevas categorías 79

A.1. Listado de códigos por año informados por DEIS 95
A.2. Contabilización casos EAM 98

Índice de Figuras

4.1. Obesidad y Sobrepeso 8
4.2. Estructura funcional del sistema de salud chileno 10
4.3. Estructura detallada de los Establecimientos de Atención Primaria 12
4.4. Ranking productos por demanda 16
4.5. Polifarmacia, según sexo y edad 18
4.6. ¿Donde obtuvo este medicamento? 18
4.7. Proceso de creación y validación de un modelo basado en aprendizaje supervisado 22
4.8. Gráfica función sigmoide 25
4.9. Ejemplo de árbol de decisión y partición del espacio que genera 26
4.10. Funcionamiento algortimo Bosques Aleatorios 26
4.11. Ejemplos de curvas ROC 30

6.1. Universos de códigos CIE-10 utilizados 42
6.2. Modelo Conceptual de Datos 50
6.3. Histograma de clases 52
6.4. Resultados regresión logística sin balancear 53
6.5. Resultados regresión logística balanceada 54
6.6. Importancia de las variables Regresión Logística Balanceada 55
6.7. Importancia de las variables tras primera reducción de dimensionalidad 56
6.8. Importancia de las variables, reducción 3 56
6.9. Logit completo (Statsmodels) 58
6.10. Logit reducción 1 (Statsmodels) 59
6.11. Logit sin intercepto reducción 2 (Statsmodels) 60
6.12. Random Forest: Primeras ejecuciones 61
6.13. Bosques Aleatorios, todas las variables 62
6.14. Bosques Aleatorios, reducción 1 63
6.15. Bosques Aleatorios, reducción 2 63
6.16. Curva ROC tras reducción con variables nuevas 65
6.17. Bosques Aleatorios, todas las variables, variables nuevas 67
6.18. Bosques Aleatorios, primera reducción, variables nuevas 67
6.19. Bosques Aleatorios, segunda reducción, variables nuevas 68
6.20. Comportamiento AUC a distintos niveles de muestras sintéticas 70
6.21. Gráficas Regresión logística con muestas sintéticas 71
6.22. Importancia de las variables regresión logística con muestras sintéticas 72
6.23. Gráficas Bosques Aleatorios con muestras sintéticas 74
6.24. Importancia de las variables Bosques Aleatorios con muestras sintéticas 74
6.25. Importancia de las variables con categorías nuevas en regresión logística 78
6.26. Importancia de las variables con categorías nuevas en regresión logística 80
6.27. Frecuencia EAM según edad del paciente 80
6.28. Frecuencia EAM según mes de ingreso 81

A.1. Distribución de la población según situación de afiliación a sistema previsional de salud (1990-2017) . . . 93
A.2. Comparación de la participación de farmacias por cadena . . . 94
A.3. Fases de la metodología CRISP-DM . . . 94
A.4. Formato de registro Población Inscrita Validada en Servicios de Salud Municipal (FONASA) 95
A.5. Conteo de servicio de egreso nulo por establecimiento (2018) . . . 96
A.6. Inconsistencia FONASA 2018 . . . 96
A.7. Inconsistencia FONASA 2017 . . . 96
A.8. Inconsistencia FONASA 2016 . . . 97
A.9. Inconsistencia FONASA 2015 . . . 97
A.10. Inconsistencia FONASA 2014 . . . 97
A.11. Inconsistencia FONASA 2013 . . . 98
A.12. Inconsistencia FONASA 2012 . . . 98
A.13. Regresión logística, sklearn, reducción 1 . . . 99
A.14. Regresión logística, sklearn, reducción 2 . . . 99
A.15. Regresión logística, sklearn, reducción 3 . . . 100
A.16. Regresión logística, sklearn, reducción 4 . . . 100
A.17. Logit sin intercepto: Completo. Statsmodels . . . 101
A.18. Logit sin intercepto: Reducción 1. Statsmodels . . . 102
A.19. Regresión logística, Sklearn, todas las variables, variables nuevas . . . 102
A.20. Importancia de las variables, Sklearn, todas las variables, variables nuevas . . . 103
A.21. Regresión logística, Sklearn, reducción 1, variables nuevas . . . 103
A.22. Importancia de las variables, Sklearn, reducción 1, variables nuevas . . . 104
A.23. Regresión logística, Sklearn, reducción 2, variables nuevas . . . 104
A.24. Importancia de las variables, Sklearn, reducción 2, variables nuevas . . . 105
A.25. Regresión logística, Sklearn, reducción 3, variables nuevas . . . 105
A.26. Importancia de las variables, Sklearn, reducción 3, variables nuevas . . . 106
A.27. Regresión logística, Sklearn, reducción 4, variables nuevas . . . 106
A.28. Importancia de las variables, Sklearn, reducción 4, variables nuevas . . . 107
A.29. Logit con intercepto: Completo - Variables nuevas . . . 107
A.30. Logit con intercepto: Reducción 1 - Variables nuevas . . . 108
A.31. Logit sin intercepto: Completo - Variables nuevas . . . 108
A.32. Logit sin intercepto: Reducción 1 - Variables nuevas . . . 109
A.33. Bosques Aleatorios, todas las variables, variables antiguas . . . 109
A.34. Bosques Aleatorios, reducción 1, variables antiguas . . . 110
A.35. Bosques Aleatorios, reducción 2, variables antiguas . . . 110
A.36. Bosques Aleatorios, reducción 3, variables antiguas . . . 111
A.37. Bosques Aleatorios, todas las variables, variables nuevas . . . 111
A.38. Bosques Aleatorios, reducción 1, variables nuevas . . . 112
A.39. Bosques Aleatorios, reducción 2, variables nuevas . . . 112
A.40. Bosques Aleatorios: Categorías nuevas . . . 113
A.41. Regresión Logística: Categorías nuevas . . . 113

1 | INTRODUCCIÓN

La mayoría de las atenciones en establecimientos de salud culmina en la prescripción de un medicamento (Marovac, 2001), esto se debe a que los medicamentos modernos han cambiado la forma de tratar las enfermedades o las diversas alteraciones del estado de salud. A pesar que se conocen los beneficios de su administración, cada vez existe mayor evidencia que las reacciones adversas a medicamentos son más frecuentes aún cuando se pueden prevenir, evitando consecuencias como enfermedades, discapacidad o incluso con un desenlace fatal (Red PARF, 2010).

Los países que lideran la ingesta de medicamentos responden a los mercados farmacéuticos emergentes consumiendo dos tercios de los volúmenes mundiales de fármacos, que en su mayoría corresponden a medicamentos genéricos, dentro de este grupo se encuentran países como Chile, Argentina, Colombia, México, Egipto y Brasil, por mencionar algunos. De acuerdo a las estimaciones del mercado de medicamentos provenientes del reporte "*Global medicines use in 2020*" (IMS, 2015) del Institute for Healthcare Informatics, indican que el gasto global en medicamentos para el año 2020 va a alcanzar los 1.4 billones de dólares correspondientes a 4.5 billones de dosis.

Particularmente en Chile, uno de cada diez adultos usa cinco o más fármacos y, uno de cada tres adultos mayores emplea cinco o más fármacos (Margozzini y Passi, 2018), la administración simultánea de varios medicamentos al mismo paciente en el mismo periodo de tiempo es lo que la Organización Mundial de la Salud define como "*polifarmacia*". Si bien, la medicación que es apropiada para una afección en particular, puede poner al paciente en riesgo de interacciones farmacológicas y efectos secundarios nocivos cuando se toma en conjunto con otros fármacos (OMS, 2015). A pesar que se conoce que todos los medicamentos tienen algún grado de riesgo en causar efectos indeseados, algunos fármacos han sido identificados con un potencial significativamente mayor para causar problemas cuando se prescribe a adultos mayores, ya que los efectos de la medicación dependen de la respuesta biológica, vale decir, como los órganos responden a la droga. Las reacciones adversas de los medicamentos son hasta siete veces más común en personas de 70 a 79 años que en aquellas de 20 a 29 años, incluso llegando a ser casos más severos ya que la capacidad de respuesta de los órganos se ve reducida al envejecer debido a la multimorbilidad, es decir, varias enfermedades crónicas al mismo tiempo que generan un efecto acumulado en el paciente que las padece. (Chutka et al., 2004).

La Organización Mundial de la Salud ha impulsado el desarrollo de un programa internacional de vigilancia de los medicamentos, coordinado por el Centro de Vigilancia de Uppsala, Suecia. Este programa esta compuesto por 127 países, incluido Chile desde 1996, el cual desde 2011 ha implementado la normativa que regula la actividad en el territorio nacional a partir de la inclusión de la farmacovigilancia en el DS N°3 y de la aprobación de la Norma Técnica N°140 por parte del Ministerio de Salud, la cual establece el "Reglamento del Sistema Nacional de Control de Productos Farmacéuticos de Uso Humano". Actualmente, Chile cuenta con un Sistema Nacional de Farmacovigilancia a cargo del Instituto de Salud Pública (ISP), cuyo programa utiliza el método de la notificación espontánea que consiste en que un profesional de la salud (o de otra entidad relacionada) comunica al ISP las sospechas de reacciones adversas a medicamentos (RAM) desde que toma conocimiento, ya sea por el abuso, mal uso, falta de eficacia o la dependencia hacia los medicamentos (Roldán, 2016).

Según un informe emitido por el ISP, durante el primer semestre de 2017 el 55.6 % de las notificaciones de RAM recibidas en el Centro Nacional de Farmacovigilancia fue realizada en el servicio público, esto implica que la red asistencial tanto pública como privada tuvo la mayor participación en los casos de sospechas de RAM con un 63.7 % del total de las notificaciones, seguido de la industria farmacéutica quienes notifican un 36.2 % del total durante dicho periodo (Galaz, 2018). Sin embargo, un reciente estudio revela que a través del análisis de las bases de datos electrónicas de registros hospitalarios chilenos, se detecta hasta 10 veces más daño por medicamentos que el reporte espontáneo que recopila el ISP (Collao et al., 2019), para ello se basan en el estándar de registro de la Clasificación Internacional de Enfermedades, la cual consiste en un sistema de códigos alfanúmericos, vale decir, códigos compuestos por una letra en la primera posición seguida de números, que se asignan a diagnósticos debidamente ordenados para su identificación (OPS, sf). Esto se logra, gracias a que diversos autores plantean sets de códigos que, desde su punto de vista, están relacionados a los eventos adversos a medicamentos, particularmente el estudio chileno utiliza los códigos propuestos por Jürgen Stausberg y Tai-Yun Wu *et al*.

2 | PROBLEMA DE INVESTIGACIÓN

Los Eventos Adversos a Medicamentos (EAM) han sido objeto de múltiples estudios, dado que ocurren con frecuencia y aumentan la morbilidad y mortalidad de los pacientes, siendo un problema presente en la salud pública. De acuerdo al daño como se generan los EAM, pueden clasificarse en dos grupos: daño intrínseco que responde directamente a las propiedades de la droga lo que se conoce como Reacción Adversa a Medicamentos (RAM), o daño extrínseco que se relaciona con la forma de utilizar la droga o medicamento, lo que se conoce como error de medicación. Dado que, una cantidad importante de EAM puede ser prevenible (principalmente los errores de medicación) es importante considerar la prevención y no tan sólo la detección, para ello, un paso significativo es determinar los pacientes con mayor riesgo (Rommers et al., 2007).

El principal objetivo de la farmacovigilancia es la detección de RAM severas en forma precoz que no fueron detectadas en los ensayos clínicos, ya que cuando un medicamento es aprobado para su comercialización, ha pasado por estudios clínicos en alrededor de 3000 pacientes generalmente sanos, jóvenes, a lo más con una patología y sin mayor interacción con otros medicamentos. Sin embargo las RAM severas se dan con baja frecuencia, presentándose un caso cada dos mil o cien mil pacientes, lo que significa que se requieren al menos cinco mil individuos en estudio para su detección. Esta cifra se alcanza cuando el medicamento ya es comercializado, además, es necesario considerar que el consumo se genera en otras condiciones, vale decir, se emplea en niños y adultos mayores, en pacientes con otras patologías y en conjunto con otros tratamientos farmacológicos. Para determinar la magnitud de las RAM en Chile, se utiliza el sistema de notificación espontánea, pese que es conocido que existe una subnotificación de los casos ya sea por ignorancia, inseguridad o indiferencia de los profesionales pertinentes (Varallo et al., 2014), pero se utiliza de manera global debido al bajo costo y a que ha demostrado ser eficaz (Morales et al., 2002).

Dada esta subnotificación, es que diversos investigadores han propuesto nuevas metodologías para contabilizar los eventos adversos a medicamentos y han sido utilizadas en otros estudios como guías, como el caso del estudio denominado "*Daño asociado al uso de medicamentos en hospitales chilenos: análisis de prevalencia 2010-2017*" (Collao et al., 2019), donde los autores utilizan la información disponible de los Informes Estadísticos de Egresos Hospitalarios que provee el Ministerio de Salud a través del Departamento de Estadísticas e Información de Salud en su sitio web. Esta información es trabajada como tablas en un sistema de gestión de bases de datos, y a través del diagnóstico principal y secundario registrado en estos

informes, es que se determina si un registro de egreso hospitalario es atribuible a un EAM, pero ¿en qué se basan para decir si estos códigos están relacionados a una EAM?. Primero, cabe mencionar que tanto el diagnóstico principal y secundario responden a la Clasificación Internacional de Enfermedades cuyo acrónimo es "CIE-10", dado que se utiliza la décima revisión.

La metodología empleada por Collao *et al* se basa en la utilización de dos sets de códigos distintos que indican relación con EAM, estos listados de códigos son propuestos por dos investigadores en sus respectivos estudios: Jürgen Stausberg y Tai-Yin Wu *et al*, ambos estudios analizaron los registros de las admisiones de los hospitales en búsqueda de relaciones entre los los códigos de los diagnósticos y los EAM. Cada uno de ellos propone una manera de selección de estos códigos de acuerdo a la descripción contenida en ellos, en el caso de Stausberg selecciona 338 códigos de la CIE-10, de los cuales indican al menos que la causalidad con las drogas era muy probable, mientas que el estudio de Wu *et al*, utiliza 260 códigos que en el descriptor del código se indica claramente la asociación al daño del medicamento, siendo el segundo estudio más conservador al momento de elegir el set de códigos. Estas metodologías empleadas dejan en evidencia que pueden detectar diez veces más daño debido a medicamentos, en comparación al reporte espontáneo. Sin embargo, pese a detectar la magnitud del daño debido a medicamentos, es necesario conocer los factores de riesgos asociados al ingreso de una persona a un establecimiento de salud a causa de EAM para poner mayor atención a estas variables.

En otro orden de ideas, en el estudio "*Inequidad en el acceso a salud en Chile: Estudio multifactorial basado en la Encuesta CASEN del año 2013*" (Jiménez et al., 2018), se busca describir la inequidad en el acceso a salud entre los Servicios de Salud de Chile caracterizándolos a través de indicadores de la encuesta CASEN, donde se obtiene que la proporción más alta de la población carente de previsión se encuentra en el Servicio de Salud de Chiloé, seguido por Arica y Aysén, todas zonas geográficas aisladas dentro del territorio nacional. Asimismo, se logra determinar que existe una profunda desigualdad social ya que los servicios de salud con mayor proporción de población indígena tienen, en general, los peores indicadores de pobreza y una alta participación en el sistema de salud público, pese a ello no se logra establecer la existencia de una estructura de inequidad en el acceso a salud en los servicios de salud país, ya que las variables estudiadas no se relacionan directamente en los servicios de salud. Es por ello que se decide utilizar esta idea en conjunto con la técnica de contabilización de EAM para determinar los factores que estan asociados al daño por medicamentos, ya que se registra la comuna de residencia de los pacientes que ingresan a establecimientos de salud.

Sumado a esto, es necesario considerar que la OMS establece que los factores de riesgo son cualquier rasgo, característica o exposición de una persona que acreciente la probabilidad de sufrir una enfermedad o lesión. Asimismo, establece que los determinantes sociales de la salud son "*las circunstancias en que las personas nacen, crecen, viven, trabajan y envejecen, incluido el sistema de salud"* (OMS, 2008). Cabe mencionar que los determinantes de la salud se clasifican en cuatro categorías, en primer lugar se encuentra el

medio ambiente, responde a factores relacionados externos al cuerpo humano y sobre los cuales la persona tiene escaso o nulo control, como la calidad del agua y aire de la zona donde vive. En segundo lugar, se encuentra el **estilo de vida** que representa el conjunto de decisiones de un individuo con respecto a su propia salud como los hábitos de higiene, sociales y físicos, por lo cual ejerce cierto grado de control sobre este factor. En tercer lugar, se ubica la **biología humana** que se relaciona con la salud física y mental del individuo, como la constitución del organismo. Finalmente, la cuarta categoría corresponde a los **servicios de atención** que responden al acceso a la salud ya sea como práctica de medicina, acceso a hospitales o medicamentos (Galli et al., 2017).

Con esta idea en mente surgen las siguientes interrogantes: ¿cómo afectan las características del entorno a los egresos hospitalarios relacionados al daño por medicamentos?, ¿la disponibilidad y acceso a medicamentos genera mayor impacto en los casos atribuibles a EAM?, ¿los recursos de la red asistencial de salud y su distribución a lo largo del territorio nacional es significativa a la hora de cuantificar el daño debido a medicamentos?, ¿qué relevancia tienen los factores biológicos de un individuo a la hora de sufrir un EAM?, ¿la elección del sistema previsión de salud es concluyente en los casos relacionados a medicamentos?, ¿la fecha de ingreso tiene relación con el diagnóstico del paciente y el daño de medicamentos?.

3 | OBJETIVOS

3.1. Objetivo General

Determinar los factores de riesgo asociados al daño por medicamentos en los egresos hospitalarios ocurridos entre 2011 y 2018 en Chile, mediante la caracterización a través de los determinantes de la salud y algoritmos de minería de datos a fin de generar información de valor en el área de la salud.

3.2. Objetivos Específicos

- Investigar, recopilar y medir la calidad de los antecedentes relacionados a los determinantes de la salud.
- Definir los algoritmos y técnicas de minería de datos correctas para resolver la problemática planteada.
- Detectar si existen asociaciones, y el grado, entre los determinantes de la salud y los casos de efectos adversos a medicamentos.

4 | MARCO TEÓRICO

4.1. Sistema de salud chileno

4.1.1. Estructura

"*El sistema de salud, se refiere al conjunto de personas y entidades públicas y privadas que se relacionan con la organización, financiamiento, aseguramiento, recursos o provisión de bienes y servicios en materias de promoción, prevención, cuidado o recuperación de la salud*".(OCHISAP, sfa)

El sistema de salud chileno se define como un sistema mixto en la previsión y provisión, vale decir, el financiamiento proviene del Estado y de las cotizaciones de los trabajadores y empresas. En cuanto a la previsión, se divide en sector público, privado y otros seguros. El sector público está a cargo del Fondo Nacional de Salud (FONASA), el sector privado a cargo de las Instituciones de Salud Previsional (ISAPRE) y los seguros específicos como los de las Fuerzas Armadas y de Orden. Sin embargo, el sector público es el que define las directrices generales y es el encargado de elaborar las políticas para todo el país (Durán y Narbona, 2009).

Según los resultados de la encuesta CASEN del año 2017, el 78 % de la población está asegurado por el sector público, el 14 % por el sector privado y apenas un 3 % por los seguros de las Fuerzas Armadas y de Orden, el gráfico se encuentra disponible en la Subsección A.1.1.

4.1.2. Situación nacional actual

La Organización Panamericana de la Salud (OPS) genera un informe con los indicadores básicos de las tendencias de salud en las Américas, en el último reporte (2019) entrega datos e indicadores que permiten comprender el panorama de cada nación, dentro de los que cabe destacar es el grado de obesidad, donde Estados Unidos lidera el ranking de obesidad y sobrepeso con un 68 % de la población afectada. En la Figura 4.1 se compara la situación de Chile con el estado americano y el promedio de la región, donde se puede apreciar que Chile no escapa de este problema de salud (OPS, 2019). En los últimos años, en Chile

ha aumentado significativamente el nivel de obesidad tanto en hombres y mujeres, por lo que el estado en el año 2013 decide promulgar la ley que crea el "Sistema Elige Vivir Sano" que tiene por objeto promover hábitos y estilos de vida saludables para mejorar la calidad de vida y bienestar de los chilenos (MINSAL, 2015).

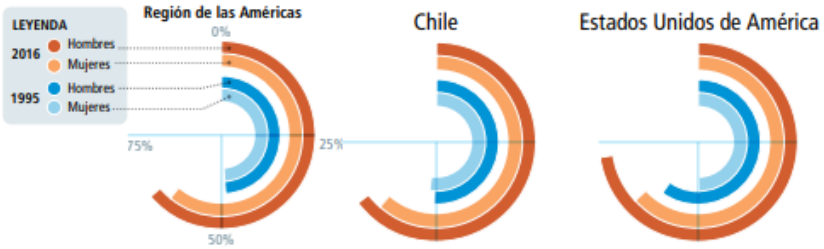


Figura 4.1: Obesidad y Sobrepeso
Fuente: Indicadores Básicos 2019, Tendencias de la salud en las Américas. Organización Panamericana de la Salud.

Dentro de los indicadores demográficos y socioeconómicos, se puede mencionar que Chile posee un 88 % de población urbana, el promedio de años de escolaridad total es es 10.5 años para los hombres y 10.2 años para las mujeres. En cuando al crecimiento del PIB, Chile adquiere un valor de 4 % anual, siendo mayor que el promedio de la región de las Américas (1.9 %). Finalmente, se presente el coeficiente de Gini, el cual mide la desproporcionalidad distributiva del ingreso en los miembros de una población y se sitúa entre 0 % (igualdad perfecta) y 100 % (desigualdad perfecta), para las Américas el coeficiente de Gini es de un 45.1 % y Chile posee 46.6 %, lo que significa que está por sobre la media en desigualdad en los ingresos.

En el año 2011, Chile se incorpora a la Organización para la Cooperación y el Desarrollo Económicos (OCDE) lo que se traduce en nuevos parámetros o puntos de referencia, ahora Chile se compara con los países más desarrollados del mundo, si se revisa la esperanza de vida actual es de 82.1 años para las mujeres, y 77.3 años para los hombres , valor cercano a los 80 años del promedio de la OCDE (INE, sf). Por otro lado, si se considera una de las enfermedades como el cáncer, Chile sigue rezagado respecto de muchos países de la OCDE en términos de control de esta enfermedad, a pesar de tener baja incidencia, 35 % inferior al promedio de la OCDE, la mortalidad por cáncer en Chile es sólo un 5 % más baja a la media de la OCDE (OCDE, 2019). Por otro lado, si se compara el gasto público en salud, queda en evidencia la falencia del Estado chileno, presentando niveles de gasto por debajo del 9.0 % promedio OCDE, ubicándose en el décimo sexto lugar con un 8.5 %, valor lejano al 17.2 % de Estados Unidos, país que lidera el ranking. Las estadísticas revelan que se gastan, en promedio, dos mil dólares per cápita en salud, de los cuales se dividen en partes iguales entre el sistema público y privado, mientras que en países de la OCDE se gastan tres mil dólares en promedio, pero solo mil corresponden a los sistemas privados, lo que genera que no exista asimetría en el gasto privado como en el territorio nacional. Otro indicador relevante, es el número de camas básicas disponibles, en Chile existen menos de dos camas por cada mil habitantes, siendo un tercio de la capacidad que tienen, en promedio, los países de la OCDE. (Cisternas, 2020)

4.1.3. Caracterización Comunal

La Subsecretaría de Desarrollo Regional y Administrativo, a través del Sistema Nacional de Información Municipal (SINIM) entrega datos y estadísticas para las 345 municipalidades del país. SINIM recopila, ordena, procesa y pone a disposición pública información dispersa del ámbito local municipal para distintas áreas como salud, educación, territorial, entre otras. Dentro de la informaciónque se puede obtener en su sitio web a través del portal de Datos Municipales [1], se encuentran indicadores como el índice de pobreza de la última encuesta CASEN, el que evidencia que las comunas más pobres son Cholchol con un 41.6 % de pobreza, seguido de la comuna de Alto Bío Bío (39.73 %) y Galvarino (37.29 %). Por el otro extremo, las comunas con menor índice de pobreza se encuentran Vitacura con un 0.13 %, seguido de Las Condes con 0.19 % y en tercer lugar la comuna de Torres del Paine con un 0.32 %.

En relación a la cobertura de agua potable, para el año 2019, se puede mencionar que las tres comunas con mayor cobertura son: Lo Espejo, El Bosque y Lo Prado todas con más de un 97 % de cobertura, mientras que Río Verde (1.03 %), Torres del Paine (9.09 %) y San Juan de la Costa (15.94 %) son las que poseen menor cobertura de este recurso.

En cuanto al área de salud, se puede destacar el número de personas inscritas validadas en salud municipal (FONASA), las comunas que presentan más inscritos son Puente Alto con 406 601 personas, seguido de La Florida con 311 842 personas, Viña del Mar con 248 602 personas y Antofagasta con 245 691 personas. En la misma línea, las comunas que reciben mayor aporte per cápita del MINSAL respecto del Ingreso Total del Sector Salud son Punta Arenas (94.89 %), Quintero (92.62 %) y Longaví (92.03 %).

Finalmente, cabe mencionar la cobertura de la educación municipal, dentro de las comunas con mayor cobertura se encuentra San Nicolás con un 142.17 % y su comuna posee 2 203 habitantes en edad escolar. A continuación en la lista sigue la comuna de Pica, con una cobertura municipal del 118.5 % de 1 108 personas, y la comuna de O'Higgins que cuenta con una cobertura de 117.39 % de los 92 escolares. Los valores de este indicador representan el total de matrículas iniciales en establecimientos municipales de educación, respecto al total de la población en edad escolar de la comuna considerando a alumnos entre 6 y 19 años. En contraste con estos números se encuentra Lo Barnechea que tan sólo un 5.98 % asiste a educación municipal de las 28 073 personas en edad escolar. Esta realidad también la representa Alto Hospicio, con 30 728 personas en edad escolar, solo el 7.22 % opta por educación municipal, en tercer lugar se encuentra Padre las Casas con un 9.04 % de la población que asiste a educación municipal de un total de 17 183 personas en edad escolar.

[1] Sitio web SINIM: http://datos.sinim.gov.cl/

4.1.4. Marco Institucional

4.1.4.1. Sector Público

El Sistema Nacional de Servicios de Salud (SNSS) es el principal coordinador de las prestaciones del sector público, está compuesto por el Ministerio de Salud (MINSAL) y sus organismos dependientes: Fondo Nacional de Salud, Servicios de Salud, Instituto de Salud Pública (ISP) y la Central de Abastecimiento (CENABAST) (OCHISAP, sfb).

Ministerio de Salud

La misión del MINSAL es construir un modelo de salud sobre la base de una atención primaria fortalecida e integrada, que pone al paciente en el centro, centrándose en el cuidado de las personas durante todo el ciclo de vida. Asimismo, busca promover y la prevención en salud, así como el seguimiento, trazabilidad y cobertura financiera (MINSAL, sfc).

En la Figura 4.2 se puede ver un esquema de la estructura funcional de este ministerio, bajo su dependencia jerárquica se encuentra la Subsecretaría de Salud Pública, la cual está a cargo de diseñar políticas, normas, planes y programas en materias que buscan la promoción de la salud, vigilancia, prevención y control de enfermedades que afectan a la población. Asimismo, debe coordinar las acciones de FONASA y del ISP e impartirles instrucciones.

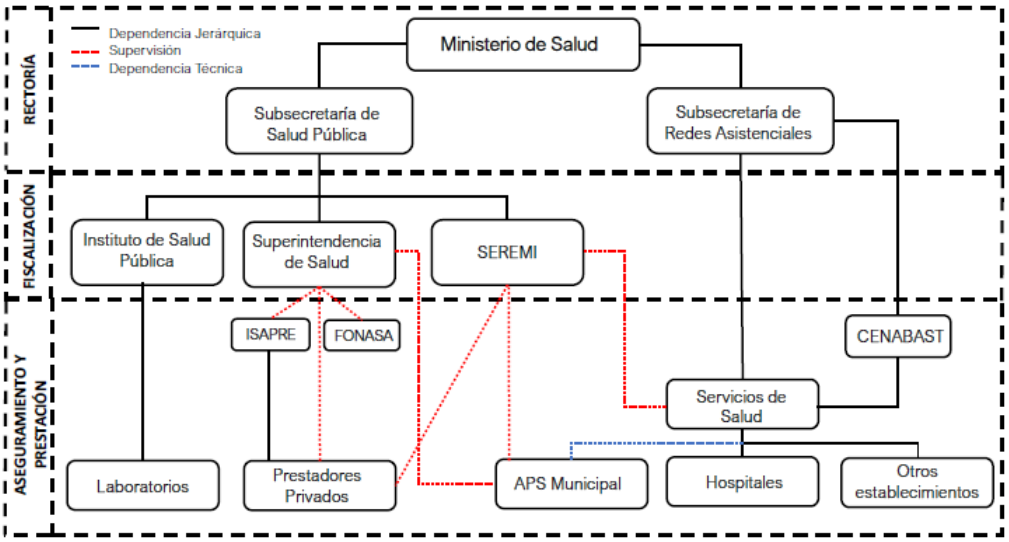


Figura 4.2: Estructura funcional del sistema de salud chileno

Fuente: Estructura y funcionamiento del sistema de salud chileno. Facultad de Medicina Universidad del Desarrollo.

La Subsecretaría de Salud Pública, se divide en cuatro áreas: División de Prevención y Control de Enfermedades (DIPRECE), División de Políticas Públicas Saludables y Promoción (DIPOL), División de Administración Finanzas (DIFAI) y División de Planificación Sanitaria (DIPLAS), encontrando en esta última el Departamento de Estadísticas e Información de la Salud (DEIS) que dentro de sus funciones destaca producir información estadística oficial, oportuna y de calidad del sector de salud [2].

Por otro lado, la Subsecretaría de Redes Asistenciales está a cargo de la red asistencial, debe regular la prestación de acciones de salud, además de definir normas que determinen niveles de complejidad y de calidad de la atención en salud. También está a cargo de coordinar la ejecución de los Servicios de Salud, la CENABAST y otros organismos que integren el SNSS (Aguilera et al., 2019).

Servicios de Salud

El SNSS cuenta con veintinueve Servicios de Salud a lo largo del territorio nacional, son organismos estatales y funcionalmente descentralizados con personalidad jurídica y patrimonio propio para ejercer responsabilidades de acción sanitaria sobre territorios geográficos definidos. Dentro de las responsabildiades se encuentra la articulación y gestión de la red asistencial correspondiente para la recuperación de la salud y rehabilitación de las personas enfermas (OCHISAP, sfc).

Cada Servicio de Salud posee un director nombrado por el Ministerio de Salud, quién debe supervisar, coordinar y controlar los establecimientos y servicios del sistema comprendido en el territorio a cargo. También, cuenta con una red asistencial de establecimientos y niveles de atención que se organizan de acuerdo a su cobertura poblacional y complejidad asistencial como se ve en la Tabla 4.1.

Establecimientos	Complejidad	Cobertura
Nivel Primario	Baja	Alta
Nivel Secundario	Media	Media
Nivel Terciario	Alta	Baja

Tabla 4.1: Clasificación de Establcimientos según complejidad y cobertura

Fuente: Caracterización del Sistema de Salud Chileno, pág 14. Fundación Sol.

Según el Observatorio Chileno de Salud Pública (OCHISAP) el nivel de atención primario tiene baja complejidad y alta cobertura dado que son la primera instancia a la que deben acudir las personas cuando presentan problemas de salud, son quienes realizan atenciones ambulatorias donde se efectuan los programas básicos de salud. Las principales actividades responden a controles, consultas, educación en grupos, vacunaciones y visitas domiciliarias. Estos establecimientos cuentan con medios simples de diagnóstico y un acotado equipo terapéutico acorde a los servicios que brindan (OCHISAP, sfb).

Los tipos de establecimiento que componen la red asistencial de la atención primaria son los Consul-

[2] Fuente: Organigrama Subsecretaría de Salud Pública. Recuperado de https://www.minsal.cl/organigrama/

torios Generales, Centros de Salud, Servicios de Urgencia, Postas de Salud y Estaciones Médicas como se puede ver en la Figura 4.3. Estos establecimientos cuentan con médicos generales y personal de colaboración como técnicos y auxiliares.

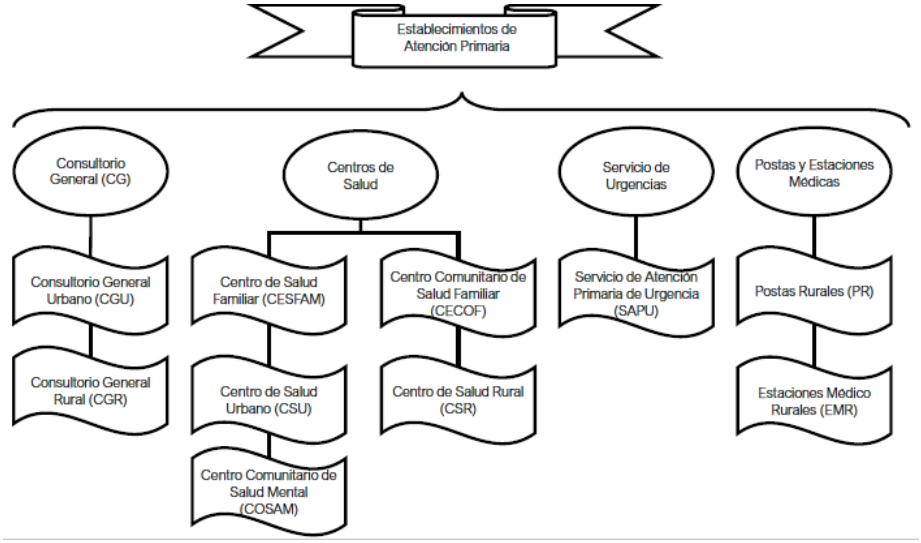


Figura 4.3: Estructura detallada de los Establecimientos de Atención Primaria
Fuente: Caracterización del Sistema de Salud Chileno, pág. 16. Fundación Sol.

Los Consultorios Generales se encargan de la promoción, fomento y protección de la salud en su área poblacional asignada. Se realizan controles a enfermos que no requieren hospitalización, atención de primeros auxilios y actividades integrales para el desarrollo de la población, los Consultorio Generales Urbanos (CGU) pueden a tender hasta 30 000 habitantes, mientras que los Consultorios Generales Rurales (CGR) atienden a localidades que poseen entre 2 000 y 5 000 habitantes, teniendo como límite los 20 000 habitantes. En cuanto a los Centros de Salud, buscan ser un continuador de los consultorios generales, sin embargo con distintos enfoques, el Centro de Salud Familiar (CESFAM), como lo indica su nombre, tiene como foco el modelo de salud para toda la familia, mientras que el Centro Comunitario de Salud Familiar (CECOF) nace con una base comunitaria en su origen y atiende a una población más pequeña (entre 3 500 y 5 000 habitantes).

Por otra parte, los Centros de Salud Urbano y Rural, son establecimientos de atención ambulatoria, que se diferencian en la magnitud de población asignada. Además, el primero puede estar adosado a un hospital de baja complejidad y el segundo depende técnicamente de un consultorio urbano y/o un establecimiento de baja complejidad. Para finalizar este punto, el Centro Comunitario de Salud Mental (COSAM) es un consultorio especializado en salud mental que está orientado a áreas con población de alrededor de 50 000 habitantes.

En cuanto, al servicio de urgencias se encuentra el Servicio de Atención Primaria de Urgencia (SAPU) que se encarga de la demanda de emergencia y urgencia médica de mediana y baja complejidad. Por otro lado, las Postas Rurales son unidades de atención ambulatoria básicas ubicadas en un área geográfica con una población menor a 1 200 habitantes, llevan a cabo atenciones sencillas de recuperación, y las situaciones

que no pueden atender con sus medios, son derivadas a establecimientos de mayor complejidad, dado que están bajo la tutela de un técnico paramédico de salud rural residente que recibe periódicamente el apoyo del equipo profesional. Finalmente, las Estaciones Médicas son centros para la atención ambulatoria básica, cuyo espacio físico es cedido por la comunidad. Estos no poseen auxiliar permanente, sino que son atendidos por un Equipo de Salud Rural que acude en rondas periódicas.

Más adelante, el nivel secundario considera una complejidad y cobertura intermedia, principalmente basado en especialidades básicas, y actúa por referencia de la atención primaria, estos establecimientos cuentan con atención ambulatoria y hospitalización, donde se lleva a cabo principalmente actividades de recuperación y rehabilitación de pacientes más complejos, existe un mayor número de médicos generales que en el nivel primari, además de contar con médicos especialistas y personal de colaboración. Los establecimientos que se catalogan en este nivel son los hospitales y centros de atención ambulatoria con tecnología de especialidad.

Por último, el nivel terciario responde a la alta complejidad y baja cobertura, cuenta con medios complejos de atención directa al paciente, apoyo diagnóstico y terapéutico, y pueden desarrollar funciones del nivel secundario, ayudando a descongestionar dicho nivel. Los establecimientos que responden a este nivel son los institutos, hospitales de especialidad y centros clínicos especializados, cuentan con médicos especialistas y personal de colaboración.(Durán y Narbona, 2009)

El SNSS cuenta con una red de aproximadamente 200 establecimientos hospitalarios de diversa complejidad, y establecimientos de atención ambulatoria secundaria. A lo anterior, se suma la red de consultorios de atención primaria, en su gran mayoría administrados por las municipalidades y que en los últimos años han presentado un giro de modelo de atención orientándose al denominado "*Enfoque de Atención Integral y Comunitario*", donde el Centro de Salud Familiar es un eje articulador. Cabe destacar que el nivel secundario y terciario están bajo la tutela de los Servicios de Salud (MINSAL, sfd).

Fondo Nacional de Salud

FONASA es un organismo público encargado de velar por la protección y cobertura de salud de sus cotizantes y a todas aquellas personas que carecen de recursos, para ello se encarga de recaudar, administrar y distribuir los recursos financieros del sector de salud (FONASA, sf). Debe caracterizar a las personas asociadas a este sistema, para ello cataloga a las personas en cuatro grupos, a continuación se presentan los tramos que rigen para el año 2020[3]:

- **Fondo A:** entran en esta clasificación quienes no cuentan con recursos y los causantes de subsidio único familiar. Este grupo se atiende de manera gratuita en hospitales y consultorios públicos.
- **Fondo B:** personas cuyo ingreso imponible mensual sea inferior (o igual) a $320 500, también entran en este grupo quienes reciben la Pensión Básica Solidaria. Al igual que el grupo anterior, reciben

[3]Fuente: FONASA. Sitio web: `https://www.fonasa.cl/sites/fonasa/tramos`

atención gratuita en hospitales y consultorios públicos.

- **Fondo C:** Pertenecen a esta clase, quienes poseen un ingreso mensual imponible entre $320 500 y $467 930, quienes tengan tres o más cargas familiares pasan al grupo B. El fondo C paga el 10 % del arancel en hospitales públicos.
- **Fondo D:** corresponden a personas con un ingreso imponible mensual mayor a $467 930, quienes tengan tres o más cargas familiares pasan al grupo C. El fondo D paga el 20 % del arancel en hospitales públicos.

En FONASA existen dos modalidades de atención para sus beneficiarios, en primer lugar se encuentra la Modalidad de Atención Institucional (MAI) que corresponde a las prestaciones médicas que se brindan en los establecimientos públicos de la red asistencial como los Centros de Salud Familiar (CESFAM), Servicio de Atención Primaria de Urgencia (SAPU), Centros de Referencia de Salud (CRS), Centros de Diagnóstico Terapéutico (CDT) y hospitales públicos a lo largo de Chile. En segundo lugar, se encuentra la Modalidad de Libre Elección (MLE), bajo esta variante las personas pueden concurrir a establecimientos privados o con profesionales que tengan convenio con FONASA, accediendo a través de la compra de un bono de atención médica (Supersalud, sfb).

Instituto de Salud Pública

El ISP es un servicio público, que posee autonomía de gestión y está dotado de patrimonio y personalidad jurídica propia. Dentro de sus funciones se encuentra la evaluación de calidad de laboratorios, vigilancia de enfermedades, control y fiscalización de medicamentos, cosméticos y dispositivos de uso médico, salud ambiental, salud ocupacional, producción y control de calidad de vacunas, entre otros.

Actualmente, el ISP se estructura en distintos departamentos, sin embargo cabe destacar a la Agencia Nacional de Medicamentos (ANAMED) por ser los encargados del control de los productos farmacéuticos, son quienes deben otorgar las autorizaciones sanitarias y registros que garanticen la calidad, seguridad y eficacia de los medicamentos comercializados en Chile. Este departamento posee, a su vez, siete subdivisiones: 1. Laboratorio Nacional de Control, 2. Cosméticos, 3. Comercio exterior, estupefacientes y psicotrópicos, 4. Biofarmacia y equivalencia terapéutica, 5. Fiscalización, 6. Autorizaciones y Registro Sanitario y 7. Farmacovigilancia (BCN, 2018b).

Para el interés de este estudio es necesario ahondar en el subdepartamentos Farmacovigilancia, división que lleva a cabo actividades relacionadas con la detección, evaluación, compresión y prevención de los efectos adversos asociados al uso de los medicamentos (EAM) o cualquier otro problema relacionado con medicamentos. Dentro de sus objetivos se encuentra conocer la realidad de las Reacciones Adversas a Medicamentos (en adelante, RAM) de la población chilena, las RAM son una respuesta nociva no intencionada que se produce a dosis normalmente utilizadas en las personas para el diagnóstico o tratamiento de una enfermedad, o bien para la modificación de una función fisiológica (OMS, 2012). Para lograr este objetivo, el

ISP debe detectar aumentos en la frecuencia, identificar los factores de riesgo que determinan su aparición y prevenir que los pacientes sean afectados innecesariamente por fármacos potencialmente riesgosos. Para ello, evalúan las sospechas de RAM que les informan, y también evalúan publicaciones obtenidas de agencias de reguladoras de medicamentos de otros países, a fin de conocer si existen nuevos riesgos asociados a medicamentos utilizados en nuestro país. (ISP, sf)

Finalizando, es necesario mencionar que el ISP realiza un proceso de planificación estratégica institucional, el que contempla el Plan Nacional de Salud (2011-2020), elaboración y gestión de la Política Nacional de Medicamentos y lineamientos gubernamentales particulares. Cabe detallar la Política Nacional de Medicamentos, la cual se trata de 31 medidas que buscan mejorar la disponibilidad de medicamentos, disminuir el gasto de bolsillo de las familias y asegurar la calidad de los productos farmacéuticos que se comercializan. Algunas de las medidas son: implementar al menos un almacén farmacéutico para la venta de medicamentos en todas las comunas que cuentan con salud municipalizada, y que no cuenten con farmacias, crear la aplicación móvil "Pastillero Digital"que buscará apoyar al paciente durante su tratamiento mediante el registro de los medicamentos que consume la persona junto con los de horarios de consumo para que operare como recordatorio. También se busca entregar medicamentos a domicilio para personas con dependencia severa y que se controlan en centros de atención primaria, simplificación de los plazos para el registro de medicamentos a un plazo máximo de tres meses, entre otras. (MINSAL, sfa)

Central de Abastecimiento

CENABAST es una institución pública, cuya misión es contribuir al bienestar de la población asegurando la disponibilidad de medicamentos, alimentos, insumos y equipamiento a la red de salud, mediante la gestión de un servicio de abastecimiento de excelencia, eficiente y de calidad, vale decir, gestiona los procesos de compra mandatados por el Ministerio de Salud o de las entidades que se adscriban al SNSS para el ejercicio de acciones de salud.

Su objetivo principal es lograr ahorros significativos para el sistema público a través de la centralización de compras, anualmente esta institución realiza un proceso de consolidación de la demanda de todos los establecimientos de salud que intermedian sus compras a través de CENABAST.(CENABAST, sf)

En la siguiente imagen se presentan los diez productos con mayor demanda en el territorio nacional, destacando en los tres primeros puestos se encuentran medicamentos: *paracetamol*, analgésico y antipirético utilizado para combatir la fiebre o dolor de intensidad leve o moderado, *losartan* utilizado para el tratamiento de la presión arterial alta, y *metformina* que se utiliza para el tratamiento de la diabetes mellitus no dependiente de insulina en pacientes cuya glicemia no puede ser controlada sólo con la dieta, ejercicio o reducción de peso [4].

[4] Fuente: LaboratorioChile. Sitio web: `www.laboratoriochile.cl/producto`

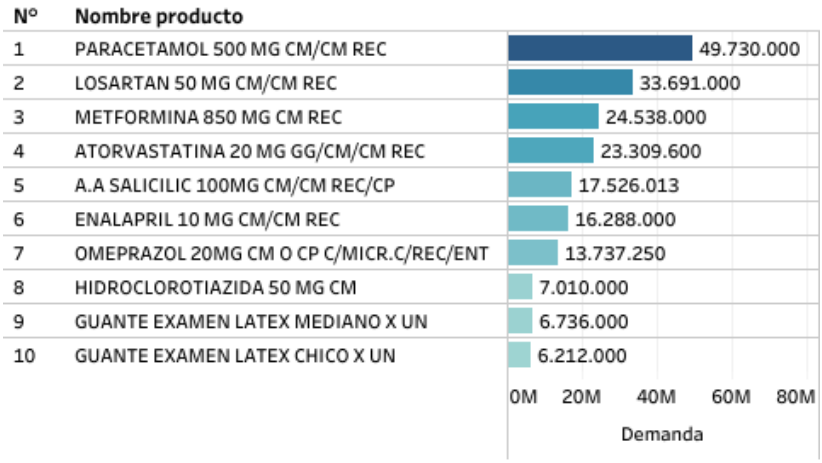


Figura 4.4: Ranking productos por demanda

Fuente: Demanda estimada de productos de Intermediación. CENABAST.

4.1.4.2. Sector Privado

El sector privado se divide en las industrias de seguro, prestaciones asistenciales y productos sanitarios, para el interés de esta investigación se centra en la industria de seguros, ya que aquí pertenecen las ISAPRES junto a otras instituciones aseguradoras.

Las ISAPRES pueden ser de carácter abiertas cuando están orientadas a trabajadores de cualquier empresa o cerradas cuando están orientadas a trabajadores de una empresa determinada, tienen por objetivo la administración y otorgamiento de las prestaciones de salud contratadas por sus beneficiarios. Los beneficiarios son quienes cotizan el 7 % u otro valor mayor convenido de sus ingresos mensuales y, sus respectivas cargas familiares.

Todo esto queda plasmado en un Contrato de Salud, documento de carácter individual entre la ISAPRE y el afiliado, y se expresa en un escrito formal donde se establecen los derechos y obligaciones de las partes (Supersalud, sfa), dentro de los beneficios mínimos obligatorios se encuentran las Garantías Explícitas en Salud (GES) que constituyen beneficios garantizados por ley como el acceso, oportunidad, protección financiera y calidad, estos se traducen en el derecho por ley a la prestación de salud con un tiempo máximo de espera cancelando el porcentaje de afiliación correspondiente ante un prestador acreditado o certificado (Supersalud, sfc). Asimismo, se incluye como beneficio mínino obligatorio los exámenes de medicina preventiva gratuitos, el pago de subsidios por incapacidad laboral, lo que comunmente se conoce como el pago de “licencias médicas”, atención gratuita de la mujer en el periodo de embarazo y hasta el sexto mes de vida del hijo, entre otros. (Supersalud, sfb)

Finalmente, es necesario mencionar que las ISAPRES presentan un bajo compromiso en la salud pública y la prevención según indica la Asociación de ISAPRES, esto se explica por la alta tasa de rotación

en la población asegurada, cada año un 10 % de los afiliados cambia su esquema de seguro lo que reduce el incentivo para invertir en iniciativas de prevención y salud pública. Otro factor que reduce el incentivo a invertir en prevención se debe a que la mayoría de la población adulta mayor pertenece a FONASA, debido a que las ISAPRES tienen la capacidad de cobrar una prima de acuerdo al riesgo del afiliado, por lo que a mayor edad los precios de los planes de aseguramientos se encarecen, por ello es que no inviertan en programas de envejecimiento saludable. (OCDE, 2019)

4.2. Los medicamentos en Chile

4.2.1. Antecedentes del mercado de medicamentos

De acuerdo al Decreto de ley del Código Sanitario, medicamento (o producto farmacéutico) se define como "*cualquier substancia natural, biológica, sintética o las mezclas de ellas, originada mediante procesos químicos, biológicos o biotecnológicos, que se destine a las personas con fines de prevención, diagnóstico, atenuación, tratamiento o curación de las enfermedades o sus síntomas o de regulación de sus sistemas o estados fisiológicos particulares, incluyéndose en este concepto los elementos que acompañan su presentación y que se destinan a su administración*" (BCN, 2020).

La producción de medicamentos puede efectuarse solo en los laboratorios farmacéuticos autorizados por el Instituto de Salud Pública, quien tiene la misión de fiscalizarlos y controlarlos. Todos los productos farmacéuticos, ya sean importados o elaborados en el territorio nacional deben contar con el registro sanitario para poder ser distribuidos en el país. En cuanto a la distribución de medicamentos, es de carácter minorista y se lleva a cabo principalmente por las farmacias, las cuales durante el año 2018 sumaron 1 514 millones de dólares en ventas. (Castro et al., 2020)

En otro orden de ideas, el MINSAL realiza la Encuesta Nacional de Salud (ENS) para conocer las enfermedades y tratamientos que recibe la población mayor de 15 años con el objetivo de esclarecer el panorama de salud para formular las políticas de salud, planes preventivos, además de conocer los factores de riesgo y los determinantes de la salud de las personas (MINSAL, sfb). Esta información fue estudiada por la Facultad de Medicina de la Pontifica Universidad Católica de Chile, quienes realizan un seminario enfocándose en mostrar los resultados del uso de medicamentos de la ENS 2016-2017 [5], donde los aspectos más relevantes que se concluyen de esta encuesta son:

- Uno de cada diez adultos usa cinco o más fármacos, y uno de cada tres adultos mayores usa cinco o más fármacos, como se puede observar en la Figura 4.5

[5] Seminario disponible en: https://www.colegiofarmaceutico.cl/index.php/noticias-nacionales/3125-resultados-encuesta-nacional-de-salud

- La mayor parte de los fármacos consumidos por los chilenos adultos fueron entregados gratuitamente en el sistema público de atención, como se puede evidenciar en Figura 4.6.

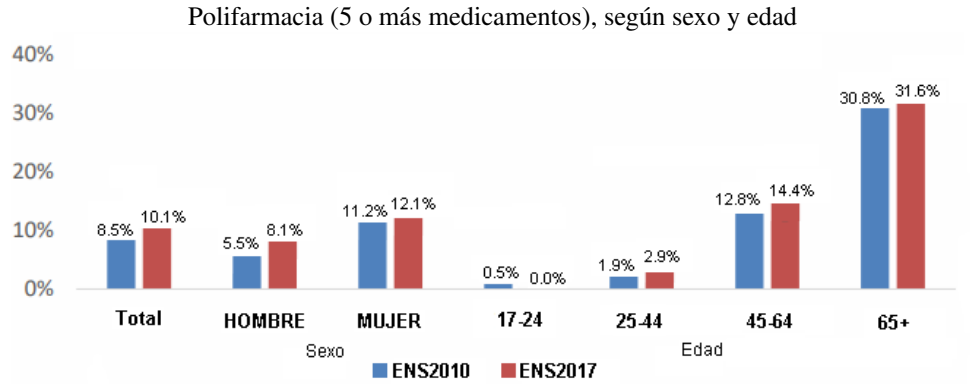


Figura 4.5: Polifarmacia, según sexo y edad

Fuente: Qué nos dice la Encuesta Nacional de Salud, ENS 2016-2017. Departamento de Salud Pública, PUC.

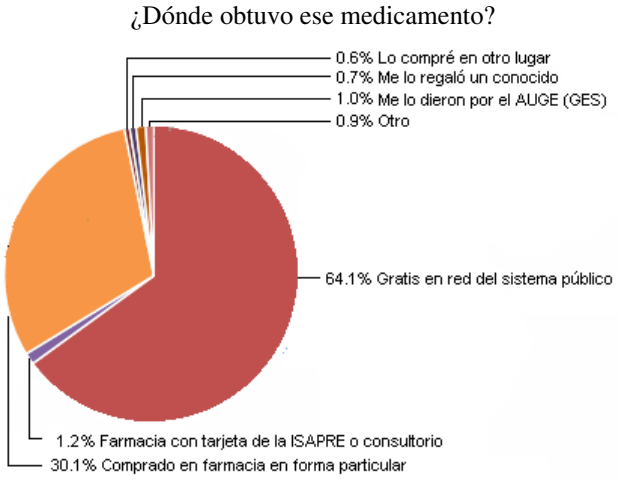


Figura 4.6: ¿Donde obtuvo este medicamento?

Fuente: Qué nos dice la Encuesta Nacional de Salud, ENS 2016-2017. Departamento de Salud Pública, PUC.

4.2.2. ¿Qué son los establecimientos farmacéuticos?

Se considera establecimiento farmacéutico a toda instalación dedicada a la producción, almacenamiento, distribución, comercialización, dispensación, control o aseguramiento de la calidad de los medicamentos, dispositivos médicos o de los elementos para su elaboración (BCN, 2019). En la normativa chilena se distinguen tres tipos de establecimientos farmacéuticos:

- Las **farmacias** son centros de salud que deben contar con un listado mínino de medicamentos que exige la autoridad sanitaria, están dirigidas por un Químico Farmacéutico que debe permanecer presente durante todo el horario de funcionamiento del local.

- Los **botiquines** son recintos donde se mantienen productos farmacéuticos para el uso interno de clínicas, casas de socorro, campamentos mineros, postas médicas, cuarteles, navíos, clínicas veterinarias y otros establecimientos más.
- Los **almacenes farmacéuticos** son los recintos destinados a la venta de los siguientes elementos: medicamentos de venta directa, elementos médico-quirúrgicos de primeros auxilios y de curación, y medicamentos bajo receta médica siempre y cuando se encuentren en la nómina del Decreto de Ley 466 (Título X), tienen por objetivo acercar los fármacos a zonas donde actualmente no existen farmacias.

En 2019, se publica un estudio realizado por el Centro de Profesional Farmacéuticos (CEPROFAR) denominado "*La farmacia en Chile*", en este se puede encontrar un gráfico que muestra la participación de farmacias en el mercado nacional (ver Subsección A.1.2), donde se puede apreciar que las tres grandes cadenas de farmacias que operan en Chile son: Cruz Verde concentrando un 17 % del mercado, Salcobrand con un 11 % de participación y Farmacias Ahumada con 9 %, y las cadenas medianas que se conocen son Dr.Simi, Farmacias Dr.Ahorro, Manriquez, Knop, Belén, La Botika y Galénica concentrando entre todas ellas un 11 % del mercado farmacéutico.

Las grandes cadenas farmacéuticas negocian directamente con los laboratorios productores e importadores debido al gran volumen de compra permitiéndoles acceder a ventajas de precios, sin embargo, aplican precios con altos márgenes de utilidad y no transmiten al público este beneficio. En cambio, las farmacias independientes se abastecen, en su mayoría, a través de distribuidoras, poseen un menor volumen de compra lo que les impide realizar rebajas significativas generando altos precios de venta (Vergara, 2019).

4.2.2.1. Comercialización

La metodología utilizada por las grandes farmacias es proveer gran variedad de productos con el objetivo de satisfacer los requerimientos de los clientes que llegan buscando nombres de fantasías o marcas recetadas por los médicos en lugar de poseer solo una marca (Ricchione, 2020), lo que se corresponde con la encuesta realizada por la Fiscalía Nacional Económica a consumidores de medicamentos, donde el 96 % de las personas dice comprar el medicamento que el doctor le receta y, un porcentaje muy relevante de los consumidores no cambiaría el medicamento prescrito por uno más barato, aunque que se les asegure que es igual de efectivo, ya que confían más en lo recomendado por el médico (Castro et al., 2020).

Según el reporte "*Distribución de farmacias por región*" publicado por el MINSAL, hasta noviembre de 2014 existen 2924 farmacias en Chile, de las cuales un 48.84 % se encuentran en la región Metropolitana, y el resto se distribuye en el resto del país. Otro dato relevente es que el 89.7 % de las ventas de medicamentos se reúnen en las Farmacias Ahumada, Cruz Verde y Salcobrand. Cabe mencionar que el 65 % de las ventas corresponden a medicamentos que requieren receta médica, y el 38 % de las ventas corresponden a medicamentos genéricos.

4.2.3. Ética del mercado farmacéutico

La Organización Mundial de la Salud (OMS) establece criterios éticos sobre la promoción, propaganda y publicidad de medicamentos, con el objetivo de apoyar y promover la protección de la salud de las personas por medio del correcto uso de medicamentos. Estos principios no son obligaciones jurídicas y se pueden adaptar a cada país, comprometen a diversos actores como al gobierno, industria farmacéutica, industria de la publicidad, personal de salud participante en la prescripción, la dispensación, el suministro y la distribución de medicamentos, universidades, entre varios más (Red PARF, 2011).

Algunos de los principios generales son :

- Los medicamentos son un bien social por tanto se establece que sean tratados como bienes de salud y no como simples productos de consumo.
- Los gobiernos deben promover y fomentar la educación en los usuarios y profesionales involucrados en el tema para crear una actitud reflexiva y crítica frente a los diferentes tipos de promoción, propaganda y publicidad de los medicamentos.
- Los medicamentos de venta directa, es decir, los que se venden sin receta, podrán ser objeto de promoción, publicidad y propaganda dirigida a la población en general.
- La promoción, propaganda y publicidad de medicamentos de venta sin receta no debe inducir al uso indiscriminado, innecesario, incorrecto o inadecuado.
- La promoción, propaganda y publicidad de medicamentos no podrá utilizar expresiones capaces de causar miedo o angustia, o sugerir que la salud puede ser afectada por no usar el medicamento.
- Cualquier tipo de promoción, propaganda y publicidad de medicamentos no debe exagerar lo que se espera del producto por encima de lo científicamente comprobado. Asimismo, no deberá atribuir al producto propiedades terapéuticas, nutricionales, cosméticas, preventivas o de cualquier otra naturaleza que no hayan sido expresamente reconocidas o aprobadas por la autoridad sanitaria.

Los profesionales de la salud desempeñan un papel relevante en asegurar el uso racional de los medicamentos, en particular el personal médico y personal de farmacia, son quienes deben evaluar las distintas opciones de tratamientos disponibles y obtener la relación del beneficio y daño potencial de cada opción, es por ello que es necesario mencionar la ética desde ambos puntos de vista. En los últimos años, se ha vuelto preocupante la relación entre los profesionales de la salud y la industria farmacéutica, ya que ésta última tiene la capacidad de influir en las decisiones de prescripción y dispensación de medicamentos afectando la selección racional de los tratamientos, toda persona puede verse influenciada por técnicas sofisticadas de mercadeo, sin embargo, lo relevante de esta industria es que el mercadeo va dirijido a los profesionales de la salud e influye en los tratamientos que otorgaran a los pacientes y no en sus propias decisiones de consumo.

La OMS en conjunto con *Health Action International* [6] elaboran una guía práctica denominada "*Comprender la promoción farmacéutica y responder a ella*" (Goodman et al., 2011), donde se menciona que estudios han revelado que 8/10 de médicos recibían obsequios como comidas gratuitas en su lugar de trabajo, 8/10 recibían muestras gratuitas de medicamentosy 4/10 tenían los gastos pagados por asistir a reuniones y conferencias, generando cambios y claras tendencias en sus prescripciones, aunque pocos reconocen que estas técnicas tengan incidencia sobre ellos. Sin ir más lejos, existe evidencia que la mercadotecnia también afecta a los pacientes, y solicitan remedios en particular para su tratamiento, en esta guía práctica se menciona el caso de la tasa de prescripción de dos tipos de inhaladores para el asma (beclometasona y fluticasona) antes y después de una campaña publicitaria iniciada en abril de 2002, donde se evidencia que muchos pacientes que mantenían un tratamiento con el inhalador ya presente en el mercado (beclometasona) se cambiaron al nuevo, incluso siendo más costoso y se sabe que, a dosis equivalentes la fluticasona no es más efectiva dejando en evidencia el poder de influencia del marketing.

Por otro lado, el comportamiento ético del personal de farmacia se puede evidenciar en el estudio *"Dispensación de medicamentos en las grandes farmacias de Chile: Análisis Ético sobre la profesión del Químico Farmacéutico"* (Marín, 2017), en el cual se efectúan varias entrevistas a Químicos Farmacéuticos (en adelante, QF) de las grandes cadenas de farmacias de la región de Valparaíso y de la región Metropolitana, donde algunos testimonios dan cuenta a la disyuntiva entre la ética profesional y la retribución económica por las comisiones de ventas o cumplimiento de metas. Se menciona que el desempeño de "buen farmacéutico" es otorgado a quién vende más, y no quien restringe la venta de medicamentos para el uso racional, o quien que realiza labores de farmacovigilancia. Algunos de los casos que dan cuenta, son el testimonio de una mujer QF que dice "si yo sé que dos medicamentos que tienen el mismo principio activo y uno me renta el triple, no me voy a quedar tranquila hasta vender el que me renta el triple". Otro caso que se expone es de un QF que menciona que su trabajo no le permite cumplir su rol y enseñar sobre medicamentos, porque al otorgar alternativas más económicas no va a cumplir sus metas.

Para finalizar, otro ejemplo que se menciona es cuando un cliente requiere un medicamento para una enfermedad y existen dos opciones que cumplen la misma función y van a lograr los mismos efectos, uno genérico a bajo costo (entre $300-$400), y uno de marca con un precio superior a los $5000. La QF entrevistada menciona que entra en discordia la ética cuando se ve que el cliente no tiene para pagar , y expone :"¿le vas a negar otro más barato? No puedes, o sea, desde mi punto de vista no puedes negarlo, hay colegas que sí, que lo niegan y prefieren que el cliente se vaya antes de dar la opción del genérico".

[6] La HAI es una organización independiente sin fines de lugro, que lleva a cabo investigaciones y actividades para promover políticas que permitan el acceso a medicamentos seguros, efectivos, asequibles y de calidad garantizada y el uso racional de medicamentos. Recuperado de https://haiweb.org/

4.3. Minería de datos

La minería de datos es el conjunto de metodologías, aplicaciones y tecnologías que permiten reunir, limpiar y transformar datos de los sistemas de información no estructurada en información estructurada, para su posterior explotación o para análizar y crear conocimiento con el objetivo de apoyar la toma de decisiones del negocio (Marcano y Talavera, 2007).

En adelante se utiliza como referencia el libro "Minería de datos modelos y algoritmos" (Gironés et al., 2017), donde se se expone una metodología propuesta por las empresas DaimlerChrysler y SPSS denominada "*Cross Industry Standard Process for Data Mining*" (CRIPS-DM), la cual establece un proyecto de minería de datos como una secuencia de fases que se detalla a continuación, la representación se puede observar en la Figura A.3 presente en anexos.

La primera etapa corresponde a la **comprensión del negocio**, la cual tiene como finalidad establecer los objetivos de la investigación y evaluar la situación actual del problema con el objetivo de plantear las preguntas correctas y no enfocar esfuerzo en un proyecto que no genera información relevante a la toma de decisiones. Más adelante, en la segunda fase correspondiente a la **comprensión de los datos** se debe trabajar con los datos para familiarizarse en profundidad, a fin de saber de dónde se obtienen, cuál es la estructura, que inconvenientes presentan y como poder mitigarlos. Ésta fase se considera crítica, ya que es donde se define la calidad de los datos, que se entienden como la materia prima en la minería de datos.

La tercera etapa es la **preparación de los datos** cuyo objetivo es disponer del juego de datos final sobre el que se aplicarán los modelos, se debe establecer el universo de los datos, realizar tarea de limpieza e integrar datos de diversas fuentes si es necesario. Una vez que los datos están preparados, es necesario dividirlos en dos grupos: *entrenamiento* y *prueba*, donde el conjunto de entrenamiento permite crear el modelo, mientras que el conjunto de prueba se utiliza para medir la precisión alcanzada por el modelo, como se puede ver en la Figura 4.7.

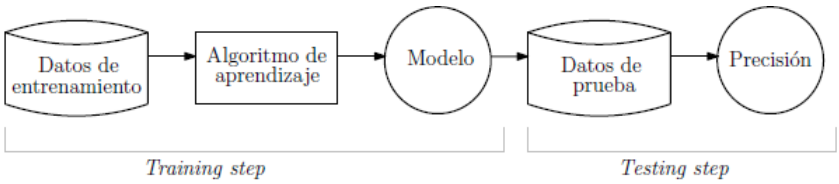


Figura 4.7: Proceso de creación y validación de un modelo basado en aprendizaje supervisado

Fuente: Minería de datos, modelos y algoritmos (pág. 72)

Siguiendo, en la cuarta fase se encuentra el **modelado**, donde se busca construir un modelo que permita alcanzar los objetivos del proyecto, siendo capaces de elegir la técnica de modelado más adecuada para el set de datos, asimismo se debe establecer una estrategia de verificación de la calidad del modelo.

Las dos últimas etapas corresponden a la **evaluación** y **despliegue**, donde la primera responde a evaluar el grado de acercamiento del modelo a los objetivos planteados inicialmente a través de métricas de evaluación, y adicionalmente, se debe revisar todo el proceso de minería de datos que se ha llevado hasta este punto con el fin de establecer los siguientes pasos del proyecto. En la etapa final de despliegue, el conocimiento obtenido se debe organizar y presentar para que pueda ser usado más adelante, asimismo desprender las lecciones aprendidas para futuros proyectos.

4.3.1. Métodos y tipos de tareas

La minería de datos se divide en dos grandes grupos: métodos supervisados, los que requieren de un conjunto de datos previamente etiquetados de acuerdo a conjunto de clases, y los métodos no supervisados, en donde los datos no poseen etiquetas o clasificación previa.

Es importante distinguir los tres problemas que se puede resolver con la minería de datos, los cuales son problemas de clasificación, regresión y agrupamiento, esta última más conocida como *clustering*. La clasificación y regresión responden a métodos supervisados y el agrupamiento a métodos no supervisados.

Los problemas de clasificación radican en asignar las instancias a un conjunto de clases, es decir, el resultado es una categoría arbitraria entre un número limitado de clases, según el problema a tratar. La función de clasificación puede denotarse como:

$$c : X \rightarrow C$$

Donde c representa la función de clasificación, X el conjunto de atributos que representan una instancia y C la etiqueta de clase de la instancia, en esta investigación se utiliza particularmente la clasificación binaria, donde el conjunto de datos pertenece a dos clases $C = \{0, 1\}$.

Sin embargo, es necesario diferenciar los problemas de regresión, ya que estos se definen como un problema de clasificación con clases contínuas, el objetivo de este tipo de problemas consiste en asignar valores numéricos a las instancias y se conocen comunmente como predicción numérica. Por otro lado, el clustering se puede considerar como un problema de clasificación, pero donde las clases no están definidas y a través de las similitudes de las instancias, la minería de datos descubre las clases.

4.3.2. Algoritmos de clasificación

A pesar que, en el libro aludido anteriormente se mencionan algunos algoritmos para los métodos supervisados de clasificación como el algoritmo de k vecinos más cercanos, que si bien es un método simple y robusto, el rendimiento se ve afectado cuando la dimensionalidad del problema es grande. También se exhibe el algortimo de máquinas de soporte vectorial, que es un algortimo potente para clasificar, pero el problema

radica en la comprensión del hiperplano de separación para las clases, lo mismo ocurre en el caso de redes neuronales, ya que explicitar el conocimiento de estos modelos es una tarea compleja. Es por ello, que se selecciona un método de ensamblado correspondiente a Bosques Aleatorios, dado que crean mejores modelos a través de la unión de métodos más sencillos. Por otro lado, se estudia en fuentes externas la regresión logística, ampliamente utilizada para la clasificación.

4.3.2.1. Regresión logística

Como se menciona anteriormente, a pesar de su nombre, la regresión logística es ampliamente usada para resolver problemas de clasificación, para ello se utiliza la función logística (también llamada sigmoide) que determina la probabilidad de la variable dependiente dado un conjunto de variables independientes, se utiliza como umbral de corte la probalidad de 0.5, vale decir, si el valor de Y es mayor a 0.5 se aproxima a 1, y si es menor o igual se aproxima a cero, en consecuencia se obtiene una variable dependiente binaria de clases 0 y 1 (Rodríguez, 2018).

La función sigmoide es una curva en forma de "S'' como se puede observar en la Figura 4.8, puede tomar cualquier valor entre 0 y 1, pero nunca valores fuera de estos límites, la función se describe por (Gujarati, 2004, pág. 574):

$$\sigma(x) = \frac{1}{1 + e^{-(\beta_0+\beta_1 X_1+...+\beta_k X_k)}} = P_i = E(Y = 1|X_i) \tag{4.1}$$

Donde P_i es la probabilidad de tener un caso favorable (Y=1), por lo tanto la probabilidad de no tener un caso favorable es $1 - P_i$, que se define como:

$$1 - P_i = \frac{1}{1 + e^{-(\beta_0+\beta_1 X_1+...+\beta_k X_k)}} \tag{4.2}$$

Sin embargo, se genera un problema en la estimación, ya que P_i no es lineal en los parámetros (β), ni en X, por lo que, convenientemente se trabaja de la siguiente manera:

$$\frac{P_i}{1 - P_i} = \frac{1 + e^{(\beta_0+\beta_1 X_1+...+\beta_k X_k)}}{1 + e^{-(\beta_0+\beta_1 X_1+...+\beta_k X_k)}} = e^{(\beta_0+\beta_1 X_1+...+\beta_k X_k)} \tag{4.3}$$

Se tiene que $P_i/(1 - P_i)$ es la razón de probabilidades en favor del caso exitoso, respecto del caso no exitoso conocido como odds ratio. Aplicando logaritmo natural, se obtiene:

$$ln(\frac{P_i}{1 - P_i}) = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k \tag{4.4}$$

Donde β_0 es la ordenada en el origen de la función de regresión, $\beta_1, \beta_2, ..., \beta_k$ representan los coeficientes de la pendiente de la recta, y las $X_1, X_2, ..., X_k$ representan las variables de riesgo. El criterio de entrada o salida de las variables de riesgo es la significación estadística de los coeficiente de regresión o la influencia de los coeficientes, las dos variantes conocidas son añadir variables una a una (Forward Stepwise) o construir el modelo completo e ir eliminando variables una a una (Backward Stepwise). Si el modelo se ajusta de manera adecuada se puede desprender la fuerza de asociación de las variables (Dolores y Rodríguez, 2000).

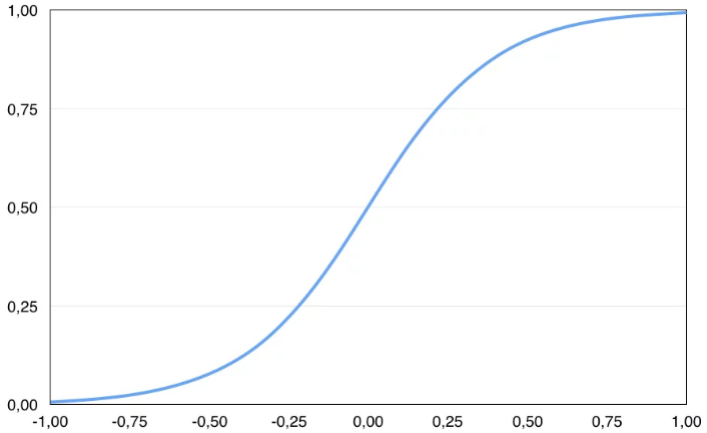


Figura 4.8: Gráfica función sigmoide
La regresión logística, Analytics Lane.

El método de resolución de este algoritmo corresponde a la máxima verosimilitud, donde en el entrenamiento se busca maximizar la probabilidad de que los puntos del conjunto de datos se clasifiquen correctamente (Rodríguez, 2018).

La ventaja de este algortimo es la simplicidad de la programación y que sus resultados son altamente interpretables, mientras que dentro de los aspectos deficientes se encuentra que la variable dependiente debe ser linealmente separable, de lo contrario no clasifica correctamente, sin embargo, aunque no es el algoritmo más potente para clasificar, el costo computacional no es elevado.

4.3.2.2. Bosques Aleatorios

Como se menciona anteriormente, corresponde a un método de ensamblado, por lo que en primera instancia, es necesario definir qué es un árbol de decisión, ya son la base del entendimiento para este algoritmo. Los árboles de decisión, son una metodología que subdivide el conjunto de datos de entrada con el objetivo de generar regiones que tengan elementos en común, de modo que todos los elementos de una región sean de la misma clase, la cual es utilizada como representante, como se puede observar en la Figura 4.9

Un árbol de decisión consta de "*nodos hoja*" (o terminal) que representan las regiones etiquetadas, y "*nodos internos*"(o splits) que corresponden a las condiciones que permiten decidir a que subregión se va cada elemento que llega al nodo. La raíz del árbol es el elemento superior, y las hojas se sitúan en la parte más

profunda, por lo que la profundidad del árbol es el número máximo de condiciones a resolver para alcanzar una hoja (Gironés et al., 2017).

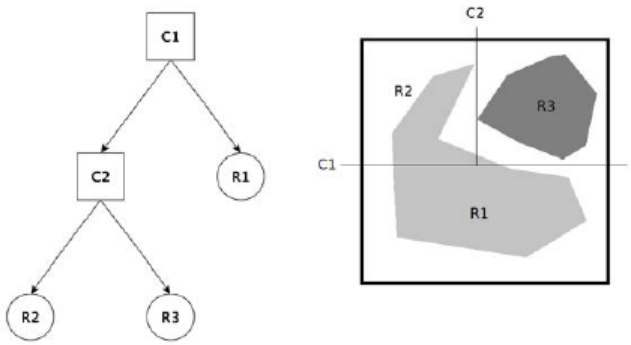


Figura 4.9: Ejemplo de árbol de decisión y partición del espacio que genera

Fuente: Minería de datos, modelos y algoritmos (pág. 213)

Con esta información en mente, los bosques aleatorios son un método de conjunto de árboles de decisión generados en un conjunto de datos dividido aleatoriamente, es por ello el nombre de bosque. Cada árbol depende de una muestra aleatoria independiente con reemplazo, donde clasifica y vota, para posteriormente elegir la clase más popular entre ellos como resultado final, esto se puede comprender gráficamente con la siguiente imagen:

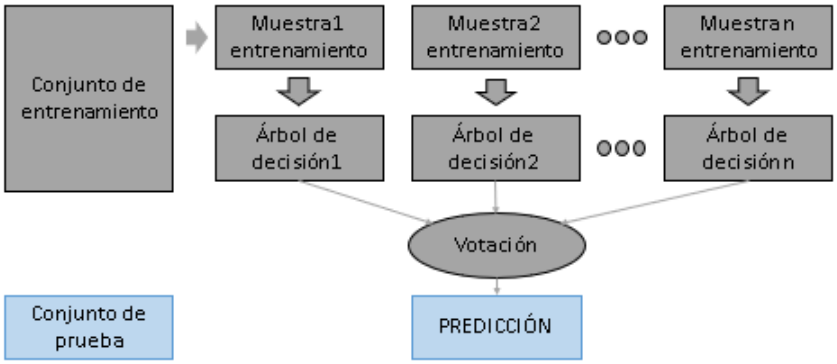


Figura 4.10: Funcionamiento algortimo Bosques Aleatorios

Fuente: Comprender los clasificadores de bosques aleatorios en Python. Datacamp

Cabe mencionar que para los algoritmos de clasificadores combinados, se puede medir el rendimiento de la clasificación de los datos que no fueron utilizados durante el entrenamiento, lo que se conoce como datos fuera de la bolsa (out of bag en inglés), lo que corresponde a utilizar como conjunto de prueba los datos que no fueron seleccionados en el proceso de entrenamiento debido a que la selección de los árboles es un proceso de muestreo con reemplazo, sin embargo, si se dispone de un gran juego de datos se aconseja utilizar un conjunto de prueba adicional para la evaluación del rendimiento.

La ventaja de este método es la precisión y robustez debido a la cantidad de árboles de decisión que participan en el proceso, no sufre el problema de sobreajuste, ya que toma el promedio de todas las predicciones. En cambio, el lado negativo de este algoritmo es el tiempo que tarda en hacer una predicción, porque cada vez que se quiera realizar una predicción, todos los árboles del bosque tienen que hacer un pronóstico para la misma entrada dada y luego realizar una votación sobre ella. También otro problema que presenta es con la interpretación, ya que es complejo en comparación a un árbol de decisión que se obtiene la ruta para conocer el resultado (Navlani, 2018).

4.3.3. Herramientas utilizadas

Para la resolución de los algoritmos se utiliza Python (version 3.8), principalmente se emplea la librería "*Scikit-Learn*" (Sklearn) que incluye paquetes que permiten estructurar y analizar datos, además de graficar y crear arreglos matriciales que ayudan en la resolución de cálculos. Esta librería proporciona una gran variedad de algoritmos supervisados y no supervisados, y el gran beneficio es que existe una gran comunidad que trabaja para mejorarla día a día. Para esta investigación, se utiliza para la resolución del los algoritmos de la regresión logística y bosques aleatorios.

Adicionalmente para la regresión logística, se utiliza el paquete "*Statsmodels*" el cual permite obtener cálculos estadísticos, incluidas estadísticas descriptivas y estimación e inferencia para modelos estadísticos, donde se obtienen los p-valor e intervalos de confianza para las variables que constituyen el modelo, lo que permite estimar las variables con mayor relevancia en el modelo final.

4.3.4. Métricas de evaluación

Las medidas de calidad de los modelos de clasificación se obtienen comparando las predicciones generadas por el modelo con el conjunto de datos de prueba. En primer lugar, se encuentra la **matriz de confusión**, la que muestra el número de instancias correcta e incorrectamente clasificadas, entregando cuatro parámetros: Verdadero Positivo (TP, *True Positive*) que indica el número de clasificaciones correctas de la clase positiva, Verdadero Negativo (TN, *True Negative*) que indica el número de clasificaciones correctas de la clase engativa, Falso Negativo (FN, *False Negative*) que representa el número de clasificaciones incorrectas de la clase positiva que fueron clasificadas como negativa y Falso Positivo (FP, *False Positive*) que representa el número de de clasificaciones incorrectas de la clase negativa que fueron clasificadas como positiva. En la siguiente tabla se puede observar un ejemplo de matriz de confusión:

	T	N
T	TP	FN
N	FP	TN
Clase Verdadera	Clase Predicha	

Tabla 4.2: Matriz de confusión

Fuente: Minería de datos, modelos y algoritmos (pág. 78)

Para una mejor comprensión, se elabora una matriz de clases binaria como es el caso a utilizar en este estudio, explicando cada cuadrante.

	0	1
0	Aciertos clase 0	0 predichos como 1
1	1 predichos como 0	Aciertos clase 1
Clase Verdadera	Clase Predicha	

Tabla 4.3: Matriz de confusión clases binarias

Fuente: Elaboración propia

De la matriz de confusión de la Tabla 4.2, se pueden derivar algunas métricas que permiten cuantificar la bondad de un modelo de clasificación. Para comenzar, se puede mencionar el **error de clasificación** (ER) y la **exactitud** (AC, en inglés *accuracy*) los que indican de manera general el número de instancias incorrecta y correctamente clasificadas respectivamente, se calculan a través de la siguiente ecuación:

$$ERR = \frac{FP + FN}{FP + FN + TP + TN} \tag{4.5}$$

$$ACC = \frac{TP + TN}{FP + FN + TP + TN} \tag{4.6}$$

A través de estas ecuaciones se puede reconocer que la exactitud es el complemento del error de clasificación (AC=1-ERR). Otra métrica que se desprende, es la **precisión** (PRE) la cual mide el rendimiento de las instancias verdaderas positivas con respecto a todas las instancias positivas predichas, como se puede apreciar en la siguiente ecuación:

$$PRE = \frac{TP}{TP + FP} \tag{4.7}$$

Continuando, se pueden desprender dos métricas más: la **sensibilidad** (SEN) (también conocida como *recall*) y la **especificidad** (SPE, *specificity* en inglés), la primera responde a la tasa de de verdaderos positivos,

mientras que la segunda representa la tasa de instancias correctamente clasificadas como negativas respecto a todas las instancias negativas reales, las ecuaciones que las representan son:

$$SEN = \frac{TP}{FN + TP} \tag{4.8}$$

$$SPE = \frac{TN}{TN + FP} \tag{4.9}$$

La precisión de una clase define cuan confiable es un modelo en responder si un punto pertenece a esa clase, mientras que la sensibilidad de una clase expresa cuan bien puede el modelo detectar a esa clase, por lo que se pueden desprender cuatro escenarios:

- **Alta precisión y alta sensibilidad:** el modelo maneja perfectamente la clase.
- **Alta precisión y baja sensibilidad:** el modelo no detecta muy bien la clase, pero cuando lo hace es altamente confiable.
- **Baja precisión y alta sensibilidad:** el modelo detecta bien la clase pero incluye también muestras de otras clases.
- **Baja precisión y baja sensibilidad:** el modelo no logra clasificar la clase correctamente.

Mientras más cercano a 1 sea el valor obtenido en las métricas presentadas, el indicador es mejor. Por otro lado, adicional a las métricas derivadas de la matriz de confusión se tiene la **curva ROC** (*Receiver Operating Characteristic*) la cual mide la sensibilidad de una prueba, en función de los falsos positivos (complementario de la especificidad) como se puede observar en la Figura 4.11, la curva ideal es como muestra el panel (a) donde la tasa de verdaderos positivos es 1, y los falsos positivos son 0. De esta curva se calcula el área bajo la curva (area under the curve: AUC), la cual permite caracterizar el rendimiento del modelo de clasificación y se han establecido los siguientes intervalos para medir la calidad del test:

$$\begin{aligned}
&\text{Si } 0.5 \leq \text{AUC} < 0.6 && \rightarrow && \text{Test malo} \\
&\text{Si } 0.6 \leq \text{AUC} < 0.75 && \rightarrow && \text{Test regular} \\
&\text{Si } 0.75 \leq \text{AUC} < 0.9 && \rightarrow && \text{Test bueno} \\
&\text{Si } 0.9 \leq \text{AUC} < 0.97 && \rightarrow && \text{Test muy bueno} \\
&\text{Si } 0.97 \leq \text{AUC} \leq 1 && \rightarrow && \text{Test excelente}
\end{aligned}$$

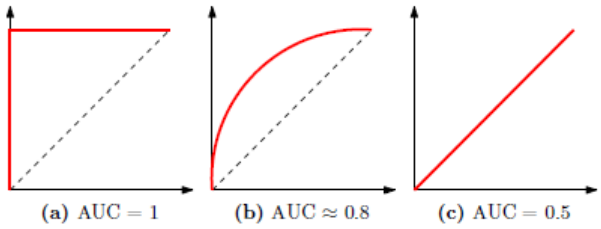


Figura 4.11: Ejemplos de curvas ROC

Fuente: Minería de datos, modelos y algoritmos (pág. 82)

Adicionalmente, en el caso de la regresión logística se utiliza la inferencia estadística, y con ello el p-valor para decidir si la variable continúa en el modelo cuando se ejecuta con el paquete de *Statsmodels*, donde se contrasta la siguiente hipótesis:

$$H_0 : \beta_k = 0$$

$$H_a : \beta_k \neq 0$$

La hipótesis nula (H_0) plantea que el coeficiente de la variable k-ésima es igual a cero, lo que implica que no hay asociación entre el término y la variable de respuesta, mientras que la hipótesis alternativa (H_a) indica que el coeficiente es distinto de cero y si existe asociación entre ellos. Cuando el p-valor es menor o igual al valor de significancia (α), se puede concluir que hay una asociación estadísticamente significativa entre la variable de respuesta y el término que acompaña al coeficiente, por lo tanto se rechaza la hipótesis nula, y se puede concluir que la variable es relevante en el modelo (Minitab, 2019). Por otro lado, el nivel de significancia es el grado de error que se está dispuesto a cometer el investigador al rechazar la hipótesis nula suponiendo que es verdadera, por lo general se utiliza un valor de 0.05 (5 %) (Ventura-León, 2017).

En relación a la bondad de ajuste, en el artículo titulado "*¿Cuál es el mejor R-cuadrado para la regresión logística?*" se exhiben las distintas pruebas para la bondad de ajuste para la regresión logística, donde el autor afirma que "hay muchas formas diferentes de calcular un R^2 para la regresión logística y no hay consenso sobre cuál es la mejor" (Allison, 2013). Asimismo, menciona que los paquetes estadísticos generalmente proporcionan el pseudo R^2 de McFadden o el de Cox-Snell, y que bajo su criterio el de McFadden es la mejor opción. En este caso, *Statsmodels* otorga el R^2 de McFadden.

4.3.5. Problemas en la clasificación

Un problema recurrente en la clasificación es el desbalance entre las clases, esto se evidencia en los problemas del área de la salud donde una fracción pequeña de los pacientes posee la enfermedad a clasificar, y la gran mayoría de los registros son individuos que no presentan la enfermedad.

Esto afecta a los algoritmos en su proceso de generalización de la información lo que perjudica a las clases minoritarias, si el algoritmo en la fase de entrenamiento tiene buen rendimiento y tiende a clasificar como cero (que no padece la enfermedad) para la fase de prueba los resultados serán similares, debido a que la mayoría de los datos responde a esta clase.

Existen estratégias para mejorar este problema, dentro de las cuales se encuentran (Na8, 2019):

- **Ajuste de parámetros del modelo**, consiste en ajustar parametros propios algoritmo para intentar equilibrar a la clase minoritaria y penalizar la clase mayoritaria durante el entrenamiento, sin embargo, no todos los algoritmos tienen esta posibilidad.
- **Modificación de la base de datos**, esta estrategia se puede utilizar desde dos puntos de vista, el primero eliminar muestras aleatorias de la clase mayoritaria corriendo el riesgo que se elimine información relevante y empeorando el modelo, o se puede replicar las muestras minoritarias asumiendo el riesgo de un sobreentrenamiento.
- **Muestras artificiales**, consiste en crear muestras no idénticas utilizando diversos algoritmos que intentan seguir la tendencia del grupo minoritario, como consecuencia se puede alterar la distribución propia de los datos.

4.4. Fuentes de información

Las fuentes de información son un instrumento para el conocimiento, la búsqueda y acceso a la información. Se pueden clasificar como primarias o secundarias, la primera responde a los instrumentos que contienen información original, provienen de una fuente directa sin ser interpretada. En cambio, las fuentes secundarias corresponden a las que la información ha sido procesada, analizada o reorganizada a partir de una fuente primaria con el fin de utilizarla en una investigación distinta a la fuente original, lo que permitirá responder una nueva pregunta de investigación o entregará una perspectiva alternativa a la interrogante original. Es importante mencionar que quienes generan fuentes de información secundaria son especialistas en el área.(Marantao y González, 2015)

Algunas de las ventajas del análisis secundario es la minimización de los efectos que el investigador puede ocasionar sobre los datos, es más económica en tiempo, recursos humanos y materiales, facilita la formulación de distintas hipótesis para los problemas de investigación. Sin embargo, se debe comprobar la validez y fiabilidad de la fuente, vale decir, el investigador debe revisar la consistencia de la información, el diseño muestral, la técnica de obtención, la cantidad de respuestas obtenidas, conceptos utilizados entre otros (Scribano y De Sena, 2009). Cabe mencionar que una buena base de datos debe asignar un apartado o capítulo dedicado a explicar los criterios que han determinado la selección de las fuentes, las definiciones utilizadas, los límites cronológicos o temáticos impuestos y la organización dada a la información.

4.4.1. ¿Cómo medir la calidad de los datos?

La norma ISO 25000 referente a la calidad de software y datos, explica que la calidad del producto de datos es "*el grado en que los datos satisfacen los requisitos definidos por la organización a la que pertenece el producto*" (ISO, 2019), estos requisitos están compuestos por quince características según lo definido por el estándar ISO 25012, y se dividen en grupos: las características inherentes a los datos, y las características que dependen del sistema.

Las características inherentes se refieren al dato en sí mismo y a la correspondencia del mismo con la información del mundo real, dentro de este grupo se encuentran la exactitud, completitud, consistencia, credibilidad y actualidad. Por otro lado, las características de la calidad dependientes del sistema se refieren a la calidad que es preservada a través de la plataforma tecnológica en la cual es empleada la información, dentro de este grupo se encuentra la accesibilidad, conformidad, confidencialidad, eficiencia, entre otros que se detallarán a continuación.

- **Exactitud**: tiene relación con el valor deseado del atributo representando correctamente su valor, ya sea por su adecuada sintáxis o semántica dentro del dominio que corresponda.
- **Completitud**: referente a que todas las instancias y atributos tengan valores para su uso.
- **Consistencia**: es el grado en que no se presentan contradicciones entre los datos.
- **Credibilidad**: se refiere a que los datos sean creíbles en un contexto específico.
- **Actualidad**: relacionado a que los atributos tengan la edad adecuada para el uso.
- **Accesibilidad**: alusivo a la capacidad de configurar los datos de manera especial debido a alguna discapacidad de la persona que los requiere.
- **Conformidad**: grado en que los atributos se adhieren a los estándares o convenciones para su uso.
- **Confidencialidad**: los datos tienen atributos que garantizan que han sido accedidos e interpretados solo por personas autorizadas.
- **Eficiencia**: tiene relación con que los datos cuentan con atributos que pueden ser procesados con recursos adecuados para obtener el rendimiento esperado.
- **Precisión**: se refiere a que los atributos son exactos y entregan un criterio claro.
- **Trazabilidad**: se explica como el ciclo de vida del dato, poder conocer desde que fue extraído hasta que se produjo la transformación en información.
- **Comprensibilidad**: los datos tienen atributos que permiten ser interpretados por el usuario, además de expresarse en símbolos, unidades o lenguaje apropiado para el contexto.
- **Disponibilidad**: capacidad para obtener los datos para su uso.

- **Portabilidad**: grado que tienen los datos para ser instalados, reemplazados o eliminados de un sistema a otro preservando sus propiedades, se puede resumir como la transferibilidad de los datos.
- **Recuperabilidad**: capacidad para acceder y extraer los datos que por daño no pueden ser accesibles de manera usual, preservando el nivel de calidad.

Un concepto que se escucha comunmente cuando se habla de calidad de datos es el paradigma "Garbage in, Garbage out" (Entra basura, sale basura), esta expresión hace alusión a que si los datos ingresados al sistema de información son incorrectos o el modelo es deficiente, los resultados que se generen a partir de ellos serán deficientes también (Saltos, 2018). Hoy en día, existen múltiples herramientas tecnológicas para medir la calidad de los datos, dentro de ellas se pueden mencionar *DataCleaner*, *Data Quality Services*, *Open Studio for Data Quality de Talend*, *Oracle Enterprise Data Quality*, entre otras más.

Thomas Redman, presidente de *Data Quality Solutions*, propone un método sencillo para realizar un análisis de datos, entendiendo que los altos cargos de las empresas deben realizar evaluaciones rápidas para determinar si pueden confiar en un conjunto de datos. El método que propone lo denomina "*Friday Afternoon Measurement*" (Medición del viernes por la tarde), que consta de cuatro pasos (Redman, 2016):

- **Paso 1:** Reunir los últimos cien registros que se generaron en la empresa, y determinar los diez a quince elementos (atributos) críticos dentro de los registros y anotarlos en una hoja de papel.
- **Paso 2:** Solicitar a dos o tres personas con conocimiento de los datos que se junten en una reunión para revisar estos registros. Generalmente los viernes por la tarde cuando la productividad decae, he aquí el origen del nombre del método.
- **Paso 3:** Revisar registro a registro, e ir destacando errores obvios en cada uno de ellos, como nombres mal escritos, valores fuera de rango, etc.
- **Paso 4:** Resumir los resultados, para ello a la lista de registros agregar una columna con el nombre "Registro perfecto", y colocar "sí", en el caso que no existan errores, y o "no" cuando existan algún tipo de error. Finalmente, interpretar los resultados.

5 METODOLOGÍA

Basado en la metodología *CRISP-DM*, se procede a detallar cada una de las fases que conlleva.

5.1. Comprensión del negocio

En principio, se estudia el sistema de salud chileno y como se originan las base de egresos hospitalarios, junto a ello se establece y recopila información referente a los determinantes de la salud para complementar la base de egresos hospitalarios. Es importante mencionar que el objetivo de esta investigación radica en determinar los factores de riesgos asociados al daño ocasionado por medicamentos, más que predecir si un egreso es consecuencia de EAM o no, dado un conjunto de atributos.

5.2. Comprensión de los datos

Se consideran las características inherentes a los datos para medir la calidad de las bases que se encuentren con la información estructurada y que posean un esquema de registro. En cambio, para las bases de datos que no cumplan estas características, se puede medir únicamente los errores aparentes que responden a la completitud de los registros y que no escapen de los rangos definidos, y para ello se utiliza la metodología expuesta por Thomas Redman.

En el primer caso, para medir la calidad de las bases de datos de egresos hospitalarios, se deben definir los dominios válidos para cada atributo, se cargan los datos como tablas en un sistema de gestión de bases de datos (MySQL), donde se trabaja en conjunto con Python, y se utilizan los esquemas de registros presentes en la página web del DEIS como se presentan a continuación:

- serv_salud → listado entregado en el esquema de registro.
- estab→ listado de establecimientos publicado por el DEIS.
- sexo → 1: Hombre; 2: Mujer; 3: Indeterminado; 9: Desconocido.
- edad → menor a 120 años*.

- previ → 1: FONASA 2: ISAPRE; 3: No tiene (Particular); 4: Cajas de Previsión FFAA; 5: CAPREDENA; 6: DIPRECA; 7: Otra
- benef → 1: A; 2: B; 3: C; 4: D
- modal: 1: MAI; 2: MLE
- f_egreso → año egreso correspondiente con el año del reporte.
- serv_egr → Listado entregado en el esquema de registro.
- d_estad → menor a 150 días*.
- diag1 → Listado entregado en el esquema de registro.
- diag2 → Listado entregado en el esquema de registro.
- cond_egr → 1: Vivo; 2: Fallecido.
- interv_q → 1: Sí; 2: No.
- región→ 1-15.
- serv_res → Listado entregado en el esquema de registro.

Este esquema de registro corresponde al año 2011, sin embargo existen variaciones entre los años, por ejemplo el dominio de la previsión se mantiene hasta 2013, ya que al año siguiente se elimina el valor "4" que hasta entonces representa a la Caja Previsional de las Fuerzas Armadas, y también se agrega el campo "9" que denota previsión ignorada. En 2018, ocurre nuevamente un cambio en los valores que puede tomar la previsión, manteniéndose constante solo los valores de FONASA e ISAPRE para todos los años de estudio. Lo mismo ocurre con el atributo "benef", que solo para el año 2011 los valores que puede tomar son entre "1" y "4" para los tramos de los beneficiarios de FONASA, y "0" cuando es una previsión distinta. Para los dos años siguientes, en lugar de ser "0" cuando es distinto a FONASA, las casillas se entregan vacías, y a contar del año 2014, los tramos se registran directamente con la letra que los representa. Por esta razón, se define el dominio válido para las variables que presentan estas variaciones de acuerdo al esquema de datos publicado en el sitio web del DEIS para cada año.

Cabe mencionar que los campos con asterisco (*) responden a dominios definidos para este estudio, dado que en el esquema no se acotan los valores que pueden adquirir estas variables, y se definen para su asegurar su credibilidad. Es necesario señalar que para verificar el dominio de los diag1 y diag2 se utilizan los listados del 2016 y 2012 respectivamente, ya que varían año a año los listados reportados por el DEIS, y se utiliza el que contenga el mayor número de códigos, asumiendo que contiene al resto dentro de ellos, en la Tabla A.1 presente en anexos se puede observar la cantidad de códigos CIE-10 por año. Luego, a a través de un programa en Pyhton se verifica que cada uno de los registros y variables pertenezcan a los dominios

correspondientes, este procedimiento se realiza para las ocho bases de datos de egresos hospitalarios, en anexos, especificamente en la Subsección A.4.1 se encuentra el código utilizado para el año 2011.

Para medir las contradicciones que presentar los datos, primeramente se evalúa la previsión (previ), el tramo de fonasa (benef) y la modalidad de atención(modal), ya que los dos últimos campos responden a características propias de los beneficiarios de FONASA, vale decir, cuando un paciente cuenta con FONASA como seguro de salud, debe indicar el tramo al que pertenece y la modalidad de atención elegida. En cambio, si el paciente tiene cualquier otro tipo de previsión que no sea FONASA, el tramo y la modalidad de atención deben ser cero o vacías dependiendo del año de estudio, por lo tanto se tienen los siguientes escenarios:

1. **Considerando la previsión como dato correctamente ingresado**
   - Si la persona es FONASA, el tramo debe tomar los valores A, B, C o D (o 1,2,3 4) y la modalidad debe tomar los valores 1 o 2. Si no cumple, se clasifica como “inconsistencia 1”.
   - Si la persona no es FONASA, el tramo y la modalidad deben adquirir el valor cero o vacíos dependiendo del año. Si no cumple, se clasifica como “inconsistencia 2”.
2. **Considerando el tramo como dato correctamente ingresado**
   - Si el tramo de beneficio es cero o vacío, el paciente no es FONASA, por lo que la modalidad también debe ser cero o vacía. Si no cumple, se clasifica como “inconsistencia 3”.
   - Si el tramo de beneficio es distinto de cero o vacío, el paciente es FONASA, por lo que la modalidad de atención debe ser distinta de cero y no vacía. Si no cumple, se clasifica como “inconsistencia 4”.
3. **Considerando la modalidad de atención como dato correctamente ingresado**
   - Si la modalidad de atención es cero o vacía, significa que el paciente no es FONASA, por lo que el tramo de fonasa debe ser cero (o vacío) y la previsión debe ser distinta a 1. Si no cumple, se calsifica como “inconsistencia 5”.
   - Si la modalidad de atención es 1 o 2, significa que el paciente es FONASA, por lo que la previsión debe ser 1, y el tramo de beneficio debe tomar valores A, B, C o D (o 1, 2, 3, 4). Si no cumple, se clasifica como “inconsistencia 6”.

Cabe mencionar que se clasifica como “inconsistencia FONASA” si el registro presenta al menos una inconsistencia de las anteriormente descritas. Adicionalmente, se puede medir la inconsistencia de la región respecto a la comuna informada, ya que se cuenta con el listado de códigos único territoriales donde el código de la comuna viene asociado a la región que pertenece. Si presentan errores se clasifica como “inconsistencia comuna-region”. También se puede medir la inconsistencia del atributo “f_egr”, validando que el año reportado en la fecha se corresponda con el año de la base de datos, si no cumple, se clasifica como

"inconsistencia fecha". Finalmente, para medir la completitud, de estos archivos basta con contar los registros campos nulos que existan en cada archivo.

En el segundo caso, se utiliza la metodología expuesta por Thomas Redman en la Subsección 4.4.1, sin embargo se utiliza la totalidad de los registros. En primer lugar, se marcan los errores evidentes como las casillas con información de texto cuando la naturaleza de los datos son numéricos, o registros incompletos, luego se agrega la columna "Dato correcto", y se coloca un "1" si no presenta errores, y "0" si presenta errores. Finalmente, se procede a contabilizar el porcentaje válido de ellos.

5.3. Preparación de los datos

El objetivo de este estudio es determinar los factores de riesgos asociados a los egresos hospitalarios debido a Eventos Adversos a Medicamentos dada la caracterización comunal y otras variables. Es por ello que se pretende unificar en un solo archivo los registros de egresos hospitalarios de los ocho años caracterizando el paciente según su comuna. Para ello se sigue la siguiente lógica de limpieza de datos, reducción de dimensionalidad, normalización de variables e integración de fuentes externas:

1. Dado que la caracterización del paciente es a través de la información de las comunas por año, se dejan fuera del estudio los registros que poseen comunas que no pertenecen al dominio válido, y asimismo se dejan fuera los registros cuya fecha de ingreso es menor a 2011.

2. Determinar las bases de datos externas que se van a utilizar para cruzar con la base de egresos, según la medición de calidad.

3. Determinar, estandarizar y normalizar las variables (de ser necesario) que ingresarán al modelo desde el archivo de egresos hospitalarios. Como se menciona anteriormente, existen variables que toman distintas categorías para los años en estudio, por lo que se decide agrupar y acotar algunas de ellas. Lo mismo se realiza para las fuentes externas recopiladas.

4. Definir $m - 1$ categorías, para las variables cualitativas con m categorías dado que si no se respeta esta regla se cae en la trampa de la variable dicotómica generando problemas de perfecta colinealidad (Gujarati, 2004).

5. Identificar los registros que, según sus *diag1* o *diag2* clasifican como EAM según el set de códigos a utilizar, para ello se establece una nueva variable denominada "EAM" que toma el valor "1" si el código del paciente se encuentra dentro del set de códigos relacionados a los EAM, y "0" si el código no es causa de EAM, siendo esta la variable de respuesta del estudio.

5.4. Modelado

Se definen los algoritmos a utilizar, correspondientes a la regresión logística y bosques aleatorios, debido a sus buenos resultados para la clasificación. Se ejecutan los dos algoritmos en forma independiente, y se evalúa la construcción del modelo a través del método de eliminación hacia atrás, el cual consiste en construir el modelo con todas las variables disponibles, y en función de la relevancia e influencia se van eliminando en cada etapa las menos influyentes, hasta que no se pueda suprimir ninguna más.

Para el algoritmo de bosques aleatorios al entrenar un árbol, se puede calcular cuánto contribuye cada característica a disminuir la impureza de los nodos, a través del parámetro *"importancia"* que proporciona la librería *Sklearn*. La importancia de las características se refiere a las técnicas que asignan una puntuación a las características de entrada en función de su utilidad para predecir una variable objetivo, esta importancia también se obtiene en la regresión logística cuando se ejecuta con dicha librería, sin embargo responde a los coeficientes asociados a la regresión.

Particularmente, se utiliza el valor del área bajo la curva ROC para determinar si el modelo es adecuado, y las métricas derivadas de la matriz de confusión para interpretar los resultados.

Adicionalmente, el caso de la regresión logística cuando se ejecuta a través de *Statsmodels* se evalúa el nivel de significancia estadística de los coeficientes para la eliminación de variables, y luego la relevancia final de los ponderadores en el modelo, es decir, se utiliza de dos maneras distintas este algoritmo.

5.5. Evaluación y Despliegue

En esta etapa se evalúa que tan cerca se está del objetivo de la investigación de acuerdo a los resultados de los algortimos ejecutados, se realizan los ajustes convenientes para tener mejores resultados posibles y si es necesario se vuelve a la fase anterior.

Si bien, se utilizan algoritmos de minería de datos, el fin no es realizar prediccions a futuro, por lo que en la etapa de despliegue no se ahonda en desarrollar un sistema para que sea utilizado por otras personas en la posterioridad, solo se obtienen las interpretaciones y las lecciones aprendidas del proceso, adicionalmente se hacen las propuestas de mejora para futuras investigaciones referentes al tema abordado.

6 | RESULTADOS

6.1. Comprensión del negocio

De acuerdo a los determinantes de la salud, se recopilan múltiples bases de datos con información que ayuda a caracterizar los egresos de los pacientes, de manera general se tienen los archivos con origen en SINIM, listado de establecimientos de salud, distribución de farmacias, códigos únicos territoriales y listados de códigos EAM según Stausberg y Wu *et al*.

6.1.1. Archivos con origen en SINIM

En el sitio web del Sistema Nacional de Información Municipal, se dispone de múltiples datos municipales con registros a contar del año 2001 hasta el año 2019 para todas las regiones. El formato de descarga es un archivo Microsoft Excel con la información que se agrega a los filtros de búsqueda como el año, región, municipio, área, subárea e indicador de interés. Las áreas y subáreas que se ofrecen son:

- Social y Comunitaria
- Desarrollo y Gestión Territorial
- Salud Municipal
- Administración y Finanzas Municipales
- Recursos Humanos Municipal
- Género
- Cementerio

Dado que, el objetivo del estudio es determinar los factores de riesgos asociados a los EAM, las principales áreas de utilidad son las primeras cuatro mencionadas, de las cuales se seleccionan los siguientes indicadores: índice de pobreza, composición de la red asistencial de atención primaria, población inscrita y validada en servicios de salud municipal, población rural y urbana, población comunal total, tasa de

mortalidad y superficie comunal. Se descargan todos estos indicadores para todos los municipios disponibles pertenecientes a los años 2011 al 2018.

6.1.2. Distribución de farmacias

Este reporte es elaborado por la Jefa de Departamento de Políticas Farmacéuticas y Profesiones Médicas perteneciente a la División de Políticas Públicas Saludables y Promoción del MINSAL, fue publicado con datos hasta noviembre de 2014.

El reporte cuenta con antecedentes generales como el número de farmacias por región, los tipos de establecimientos que corresponden, estadísticas de ventas del mercado farmacéutico. Más adelante, en el informe se entrega mediante tablas el detalle del número de farmacias instaladas, las farmacias móviles y almacenes farmacéuticos por comuna.

6.1.3. Listado Establecimientos

Dentro de las estadísticas que ofrece el DEIS en su sitio web, se encuentran los recursos para la salud, donde se puede encontrar el listado de establecimientos de salud con sus estrategias, dotación de camas públicas y privadas, por mencionar algunos.

Los parámetros que se encuentran en esta base de datos son:

- Código antiguo establecimiento, corresponde al código vigente hasta diciembre de 2013.
- Código nuevo establecimiento, código vigente a partir de enero de 2014.
- Código establecimiento madre, corresponde al código antiguo del establecimiento del que depende el establecimiento actual
- Código nuevo establecimiento madre, nuevo código del establecimiento que depende el nuevo establecimiento.
- Código y nombre de la región, corresponde a lo establecido según la división administrativa de las regiones del país.
- Código y nombre del SEREMI/ Servicio de Salud, de acuerdo al que pertenece el establecimiento.
- Perteneciente, indica si el establecimiento pertenece o no al SNSS.
- Tipo de establecimiento, como si es clínica, centro de salud, vacunatorio, etc.
- Tipo de estrategia, si es unidad móvil, centro de rehabilitación, entre otros.
- Certificación, señala si posee alguna certificación.

- Dependencia, indica si el establecimiento es de dependencia pública o privada.
- Nivel de atención, ya sea primario, secundario, terciario o si no aplica.
- Nombre oficial y alias, es el nombre oficial del establecimiento y el correspondiente alias.
- Código y nombre comuna, de acuerdo a la división político administrativa del país.
- Vía, número, dirección y teléfono, campos que permiten identificar donde se encuentra el establecimiento.
- Fecha vigencia, cierre y reapertura, corresponden a la fecha de la resolución de la autorización de funcionamiento del establecimiento o de la resolución de cierre según corresponda.
- Clasificación SAPU, indica clasificación de acuerdo al horario de atención.
- Fecha cambio, es la fecha de emisión de la resolución de algún cambio en el establecimiento o estrategia.
- Observación, corresponde a alguna anotación complementaria.
- Longitud y Latitud, para poder determinar la posición geográfica del establecimiento.

6.1.4. Códigos Únicos Territoriales

Los códigos único territoriales actuales nacen con el Decreto 1115, el cual es promulgado en septiembre de 2018, el que establece las abreviaturas para identificar a las regiones del país, y establece un sistema de codificación única para las regiones, provincias y comunas del territorio nacional.

Al crearse la provincia de Marga Marga en 2010 con las comunas de Quilpué, Villa Alemana, Limache y Olmué, las dos primeras pertenecientes anteriormente a la provincia de Valparaíso, y las otras a la provincia de Quillota, se asigna código a la nueva provincia y se modifica el de sus comunas constitutivas. Lo mismo ocurre el 2017 cuando se crea la región del Ñuble con las provincias de Diguillín, Punilla e Itata, por lo que este decreto deja sin efecto el decreto número 1439 de año 2000 y sus modificaciones, estableciendo los códigos únicos territoriales y nombres para las regiones, provinciar y comunas, y su correspondiente abreviatura legal. (BCN, 2018a)

Si bien este archivo no aporta a la caracterización, es imprescindible para relacionar los distintos archivos a través del código de la comuna.

6.1.5. Listado de códigos EAM

Los listados de códigos son proporcionados por un miembro del equipo que lleva a cabo el estudio "Daño asociado al uso de medicamentos en hospitales chilenos: análisis de prevalencia 2010-2017" (Collao

et al., 2019), donde se utilizan los dos enfoques de clasificación de EAM, cuyo universos de códigos con respecto a la CIE-10 se puede ver a continuación, de igual manera se procede a explicar el origen de cada uno de estos listados.

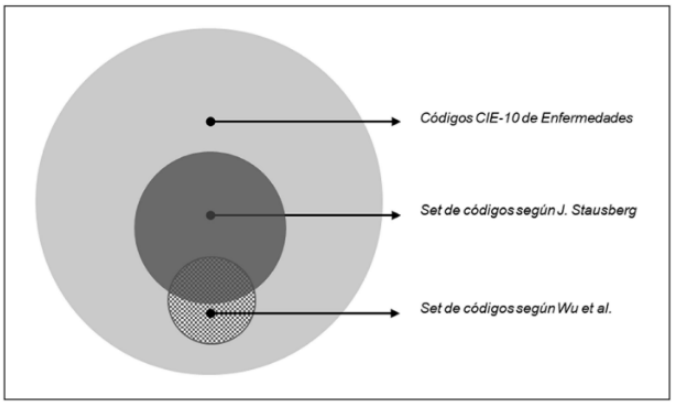


Figura 6.1: Universos de códigos CIE-10 utilizados

Fuente: Daño asociado al uso de medicamentos en hospitales chilenos: Análisis de prevalencia 2010-2017.

6.1.5.1. Listado de códigos según Wu et al.

El año 2010, se publica el estudio "*Ten years trends in hospital admissions for adverse drug reactions in England 1999-2009*" (Wu et al., 2010), el que busca determinar las tendencias de los efectos adversos a medicamentos dada la codificación de los diagnósticos de los pacientes. Para ello, los autores seleccionan los códigos de la CIE-10 que contienen en su descripción las palabras "reacción adversa al medicamento", "inducido por drogas", "debido a drogas", "debido a medicamentos" o "alergia a medicamento". Asimismo, se agregan loa códigos que contengan la palabra "inmunización", dado que las vacunas son otra forma de medicar, alcanzando un total de 85 códigos con dichas descripciones. Además, a este set de códigos se agregan los códigos que se conocen como diagnóstico secundario, incorporando los códigos que pertenecen al rango Y40-Y50 (175 códigos), dejando fuera el envenamiento accidental o intencional debido a drogas. En resumen, se consideran 260 códigos relacionados a efectos adversos a medicamentos, siendo la cota inferior de la magnitud de los EAM.

6.1.5.2. Listado de códigos según Stausberg

En 2014 se publica la investigación "*International prevalence of adverse drug events in hospitals: an analysis of routine data from England, Germany, and the USA*" (Stausberg, 2014), donde se define EAM como una lesión resultante de una intervención médica relacionada con un medicamento, esta definición incluye las reacciones adversas como las consecuencias de los errores de medicación.

Este estudio realiza un análisis comparativo entre Inglaterra, Alemania y Estados Unidos, donde

cada uno cuenta con un sistema de clasificación de enfermedades diferente. Inglaterra utiliza la versión CIE-10, Estados Unidos utiliza la versión CIE-9, mientras que Alemania la CIE-10-GM (CIE-10 Modificación Alemana) cuando se realiza la investigación. Los códigos de cada sistema se clasifican en siete categorías según la validez que tiene cada uno como indicador de EAM, como se puede ver a continuación.

- **A.1**: se observa una causalidad con las drogas en el descriptor del código.
- **A.2**: se observa causalidad con drogas u otra sustancia en la descripción del código.
- **B.1**: en la descripción del código se observa envenenamiento por medicación.
- **B.2**: se evidencia la descripción envenenamietno por uso de una droga u otra sustancia.
- **C**: la causalidad relacionada con las drogas era muy probable.
- **D**: la causalidad relacionada con las drogas era probable.
- **E**: la causalidad relacionada con las drogas era posible.

Para el estudio se consideran válidos sólo los de las categorías A,B,C, ya que son las que probablemente tengan en cuenta la adminsitración de medicamentos y otras sustancias que pueden haber causado el evento adverso. Se identifican 357 códigos en Alemania, 525 códigos en Estados Unidos y 338 códigos en Inglaterra códigos que entran en éstas categorías. Para el análisis que convoca este estudio, se consideran los códigos determinados en el listado inglés ya que se utiliza el mismo sistema de codificación en Chile.

6.1.6. Informes Estadísticos de Egresos Hospitalarios

El MINSAL propone un instrumento para recolectar información de los pacientes, a fin de conocer el perfil y tendencia de morbilidad de las personas que se hospitalizan, para ello elabora el "*Informe Estadístico de Egreso Hospitalario*" (IEEH), para la producción de información estadística sobre causas de egreso hospitalario y variables asociadas.

En noviembre de 2010, se publica el decreto 1617 exento el que aprueba la norma general técnica sobre el uso del formulario IEEH (BCN, 2010), la cual rige para todo establecimiento (perteneciente o no al SNSS) que cuente con camas de hospitalización , y establece que la responsabilidad de velar por el cumpimiento de esta norma recae sobre el Departamento de Estadísticas e Información de Salud del Ministerio de Salud. En este decreto se formaliza la información que debe contener el IEEH, la cual corresponde a:

- Número correlativo anual de egreso, identificación del establecimiento y número de historia clínica.
- Datos de identificación del paciente, como el nombre, RUN, sexo, fecha de nacimiento, edad, nacionalidad, domicilio, comuna de residencia, teléfono, previsión de salud, clase de beneficiario de fonasa, modalidad de atención, entre otros.

- Datos de la hospitalización, como días de estadías, condición de egreso, diagnostico principal, causa externa, otros diagnósticos, intervención quirúrgica principal, hora, fecha y servicio clínico de ingreso y egreso, entre otros.
- Datos de identificación del profesional tratante, como nombre y apellidos, RUN, especialidad y firma.

Los establecimientos de salud, deben enviar las bases electrónicas de datos dos veces al año a la Dirección de Servicio de Salud o Secretarías Regionales Ministeriales de Salud según corresponda, quienes validan y centralizan esta información y la hacen llegar al DEIS, quién se encarga de publicar las estadísticas relativas a la materia y pone a disposición las bases de datos que se generan.

Dentro de estas bases de datos que se generan, se encuentran los "Egresos Hospitalarios" que cuentan con su respectivo diccionario o esquema de datos, ya que no se publica con toda la información que se recolecta. Dentro de estas bases de datos se encuentra la siguiente información:

Nombre del campo	Descripción
Ser_salud	Código del Servicio de Salud de ocurrencia
Estab	Código del establecimiento de ocurrencia
Sexo	Número que identifica el sexo del paciente
Edad	Edad expresada en años
Previ	Previsión en salud
Benef	Clase de beneficiario (tramo de FONASA)
Modal	Modalidad de atención FONASA
Comuna	Código de la comuna de residencia
Fecha_egr	Fecha de alta del paciente
Serc_egr	Código del servicio clínico de egreso
Dias_estad	Días de estadía total
Diag1	Código de la CIE-10 correspondiente al diágnostico principal
Diag2	Código de la CIE-10 correspondiente a causa externa(*)
Cond_egr	Condición de egreso (vivo, fallecido)
Interv_q	Intervención Quirúrgica
Region	Código de la región de residencia
Serv_res	Código del servicio de salud de referencia

Tabla 6.1: Esquema de registro Egreso Hospitalario

Fuente: Elaboración Propia

Cabe mencionar que el d*diag2* existe cuando existe una causa externa a la hospitalización, es decir, cuando el diagnóstico principal (diag1) corresponde a "Traumantismos, envenenamiento y algunas otras consecuencias de causas". Para este estudio , se utilizan las bases de datos de los años 2011 al 2018, contando con más de trece millones de datos inicialmente.

6.2. Comprensión de los datos

Para los archivos con origen en SINIM, se utiliza la metodología denominada "*Friday Afternoon Measurement*" expuesta en la Subsección 4.4.1, en la Tabla 6.2 se muestra el porcentaje (%) de los datos correctos de cada base de datos.

Archivo / Año	2011	2012	2013	2014	2015	2016	2017	2018
Centro Comunitario de Salud Mental	90.1	90.7	87.0	90.4	93.0	91.6	91.9	89.9
Centro de Salud Familiar	90.4	91.0	87.5	91.0	93.3	92.2	92.2	90.4
Centro de Salud Rural	88.4	88.6	85.7	89.4	92.0	88.5	85.5	89.9
Centro de Salud Urbano	88.6	89.2	85.7	89.7	92.4	88.8	85.8	90.1
Servicio Atención Primaria de Urgencia	90.1	90.7	87.2	90.4	93.0	91.6	91.9	90.1
Consultorio General Urbano	90.1	90.7	87.0	90.4	92.8	91.9	91.9	89.9
Consultorio General Rural	88.4	89.2	85.4	89.4	92.2	91.4	91.6	89.9
Postal de salud rural	90.1	90.4	87.2	90.4	93.0	91.9	91.9	90.1
Centro Comunitario de Salud Familiar	90.1	90.7	87.2	90.4	93.0	91.9	91.9	90.1
Índice de pobreza	100	100	100	100	100	100	100	100
Población FONASA	89.6	89.6	89.6	89.9	85.2	85.2	89.0	89.6
Tasa de mortalidad	74.3	74.9	73.4	73.1	70.5	74.0	69.7	72.8
Superficie comunal	97.7	97.7	97.7	97.7	97.7	97.7	97.7	97.7
Población comunal	100	100	100	100	100	100	100	100
Población Rural	89.0	89.0	89.0	0	0	0	0	92.2
Población Urbana	92.5	92.5	92.5	0	0	0	0	92.5

Tabla 6.2: Datos válidos en archivos SINIM

Fuente: Elaboración Propia

El objetivo de estas bases de datos es cruzarlas con la de los egresos hospitalarios para agregar información que caracterice la comuna de residencia del paciente, por lo que se pretende lograr una caracterización limpia, que no perjudique y ensucie los registros de los egresos, razón por la cual se decide incluir en el estudio el índice de pobreza y la población comunal, ya que son las dos bases de datos que cuentan con la totalidad de los registros. Asimismo, se incluye la base de datos de superficie comunal, ya que se

dispone de recursos para poder completarla gracias a la Biblioteca del Congreso Nacional que ofrece reportes estadísticos comunales[7] donde se puede encontrar la superficie del territorio para las comuna que no contaban con información en SINIM, las cuales son Coyhaique, Lago Verde, Aysén, Cisnes, Guaitecas, Cochrane, O'Higgins, Tortel, Chile Chico y Rio Ibañez. Para el resto de los archivos, no se dispone de otras fuentes de información que complemente los registros vacíos o evidentemente incorrectos.

En cuanto a las bases de egresos hospitalarios, lo primero que se mide es la exactitud, completitud y credibilidad, los resultados se representan en porcentaje de registros inválidos Tabla 6.3.

Variable / Año	2011	2012	2013	2014	2015	2016	2017	2018
ser_salud	0.0	0.0	0.0	35.07	35.94	36.97	36.45	36.45
estab	1.66	1.82	1.83	1.54	1.37	1.01	0.53	0.20
sexo	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
edad	0.0	0.0	0.0	0.0	0.0	0.0	0.0	$1.19 \cdot 10^{-6}$
previ	0.0	0.0	$7.75 \cdot 10^{-4}$	0.0	0.0	0.0	0.0	$4.19 \cdot 10^{-4}$
benef	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
modal	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
comuna	0.15	0.27	0.22	0.60	0.30	0.24	0.38	1.17
serc_egr	0.0	0.0	0.0	0.0	0.0	0.0	0.0	44.67
d_estad	0.09	0.09	0.09	0.09	0.10	0.10	0.11	0.12
diag1	0.60	0.54	0.58	0.62	0.01	0	$1.22 \cdot 10^{-3}$	$9.52 \cdot 10^{-3}$
diag2	0.01	0.0	$1.78 \cdot 10^{-4}$	0.0	$1.20 \cdot 10^{-4}$	0	$4.28 \cdot 10^{-4}$	0
cond_egr	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
interv_q	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
region	0.15	0.27	0.22	0.60	0.30	0.24	0.38	1.17

Tabla 6.3: Datos inválidos IEEH (%)
Fuente: Elaboración Propia

Como se puede observar, existen variables con la totalidad de los registros válidos, y otras variables con datos que no pertenecen al dominio esperado o están incompletas. En cuanto al servicio de salud, los datos expuestos representan la cantidad de registros nulos en las bases de datos, sin embargo esto responde a los establecimientos que no pertenecen al SNSS y no deben asignar un valor a este campo.

En el caso de la variable establecimiento de salud, existen varios recintos que reportan sus egresos hospitalarios y no figuran en la base de datos publicada por el DEIS, ya sea porque no funcionan a la fecha del reporte del listado de establecimientos o cambiaron de código al pasar los años. Más adelante, las variables sexo, edad, previsión, modalidad de atención, condición de egreso e intervención quirúrgica no presentan mayores problemas de registro, ya sea por deficiencias de dominio, datos vacíos o nulos.

Con respecto a los datos inválidos de las comunas y las regiones son iguales por año, esto responde al desconocimiento de la comuna de residencia del paciente, en su mayoría no son registros vacíos y responden a la categoría de "ignorada". Si los responsables del registro desconocen la comuna de residencia, desconocerán también la región. Continuando con el listado de las variables, resulta curioso que en el año 2018 existan

[7]Los reportes comunales se pueden encontrar en el siguient enlace https://reportescomunales.bcn.cl/

muchos registros fuera del dominio esperado para el campo servicio de egreso, esto responde exclusivamente a campos nulos registrados principalmente por la Clínica Alemana, Clínica Santa María y Hospital Barros Lucos Trudeau, para mayor detalle ver la Figura A.5 presente en anexos. Más adelante, los *diag1* bajan drásticamente en el año 2016, dado que es la lista que se utiliza para la evaluación, en el caso de los diag2, son muy pocos los que no pertenecen al dominio, y no se cuentan los registros que tengan nulo este campo, dado que por su naturaleza esta varible puede ser nula.

En cuanto a los resultados de la inconsistencia de las bases de egresos hospitalarios se presenta, en porcentaje, en la siguiente tabla:

Variable / Año	2011	2012	2013	2014	2015	2016	2017	2018
Inconsistencia 1	0	0	0.99	0	$5.98{\cdot}10^{-5}$	0	0	0
Inconsistencia 2	5.58	6.06	8.97	0.10	0.05	0.02	0.01	0.02
Inconsistencia 3	5.54	6.06	7.30	0.03	0.02	$7.63{\cdot}10^{-3}$	$8.92{\cdot}10^{-3}$	$7.61{\cdot}10^{-3}$
Inconsistencia 4	0.03	0	2.66	0.06	0.04	$8.18{\cdot}10^{-3}$	$5.01{\cdot}10^{-5}$	0.01
Inconsistencia 5	5.46	6.06	7.16	0.03	$5.21{\cdot}10^{-3}$	$2.26{\cdot}10^{-3}$	$6.72{\cdot}10^{-4}$	$1.08{\cdot}10^{-3}$
Inconsistencia 6	0.11	0	2.80	0.07	0.05	0.01	0.01	0.02
Inconsistencia FONASA	**5.58**	**6.06**	**9.96**	**0.10**	**0.05**	**0.02**	**0.01**	**0.02**
Inconsistencia fecha	0	0	0	$6.02\cdot10^{-5}$	0	0	0	0
Inconsistencia comuna-region	0	0	0	0	0	0	0	0

Tabla 6.4: Resultados inconsistencia IEEH (%)

Fuente: Elaboración Propia

Cabe profundizar los resultados obtenidos de la variable "inconsistencia FONASA", ya que esta identifica los registros con al menos una de las inconsistencias anteriores (1-6), por lo que los registros que indique esta variable son los registros inválidos totales, se procede a evaluar en que establecimientos de salud son los que se presentan mayores inconsistencias en sus registros, donde se obtiene que entre los años 2011-2013 destacan de manera negativa los establecimientos como el Hospital Clínico Universidad de Chile, Hospital Militar de Santiago, Clínica Alemana de Temuco, Hospital Las Higueras de Talcahuano, Clinica Antofagasta, Clínica Santa María y Clínica Las Condes, repitiéndose algunos de ellos en los primeros lugares. Si bien desde al año 2014 en adelante disminuye la magnitud de la inconsistencia, se exponen de igual manera el establecimiento que presentan mayor inconsistencia en los registros asociados a los beneficiarios de FONASA para cada año a contar de esta fecha, dentro de los cuales se encuentra el Hospital Dr. Ernesto Torres Galdamez de Iquique, Hospital del Cobre Salvador Allende, Clinica Ñuñoa, Hospital Clínico San Borja Arriaran y Clínica Iquique.

Con respecto a la información de la distribución de las farmacias, no se puede determinar la calidad de los datos, ya que no se cuenta con información ni descripción de los dominios válidos, solo se puede mencionar que existen múltiples tablas donde se indica la cantidad de farmacias establecidas, farmacias

móviles y almacenes farmacéuticos, donde se asume que los registros que no poseen valor para algunos de los campos se debe a la inexistencia de dichos establecimientos, y no al desconocimiento de la información. Situación similar ocurre con la base de establecimientos de salud, ya que si bien se cuenta con la descripción de cada uno de los campos registrados, no se conoce el dominio válido para todos los atributos, además por el significado de éstos, pueden ser campos vacíos. Cabe señalar, que el ser fuentes de información secundaria, se validan a través del organismo que las publica, ya que son los especialistas en el área y han trabajado en ellas para su formulación.

6.3. Preparación de los datos

Para formar la base de datos global es necesario determinar la cantidad de registros de las bases de egresos a utilizar, dejando fuera los registros que poseen campo inválido en la comuna y/o región, y cuyos registros cuenten con año de ingreso anterior a 2011. Adicionalmente, se retiran los registros cuyos códigos comunales respondan a la Antártica, dado que la información recopilada es a través del sitio web del SINIM que publica la información de las 345 municipalidades del país, siendo este territorio el único que no posee municipalidad dentro de Chile.

A continuación, se puede observar el desglose de los registros, donde se obtiene que finalmente se trabaja con 13 198 363 registros correspondientes al 99.45 % de los registros iniciales.

Año	Registros iniciales	Registros Perdidos	Registros Finales
2011	1 648 687	19 123	1 629 564
2012	1 670 447	4 816	1 665 631
2013	1 676 936	3 859	1 673 077
2014	1 660 151	10 053	1 650 098
2015	1 671 091	5 161	1 665 930
2016	1 637 265	4 042	1 663 223
2017	1 637 150	6 373	1 630 777
2018	1 669 602	19 539	1 650 063
TOTAL	**13 271 329** (100 %)	**72 966** (0.55 %)	**13 198 363** (99.45 %)

Tabla 6.5: Registros útiles Egresos Hospitalarios

Fuente: Elaboración Propia

De estos registro se seleccionan como variables independientes para el análisis el sexo, edad, previsión, tramo de FONASA, año y mes de ingreso, siendo estos dos últimos calculados a partir de la diferencia entre la fecha de egreso y los días de estadía. Se excluyen los atributos de servicio de salud dado que no se

considera relevante, y la modalidad de atención de FONASA además de no ser relevante al estudio, posee la mayor inconsistencia. También se deja fuera el código del establecimiento ya que no se necesita un nivel de detalle menor a lo comunal. También se excluyen el servicio de egreso, condición de egreso e intervención quirúrgica dado que son consecuencias del ingreso hospitalario, y lo que se busca es encontrar variables que estén asociadas al ingreso, y posterior egreso de los pacientes.

A los registros y variables seleccionadas, se les agrega la siguiente caracterización según la comuna de residencia del paciente: población y superficie comunal, índice de pobreza, número de farmacias y almacenes farmacéuticos por comuna, cantidad de establecimientos de salud de atención primaria, secundaria y terciaria por comuna. Por otro lado, a pesar que la totalidad de los diagnósticos principal y secundario no pertenecen al dominio especificado por el DEIS, no significa que sean inválidos, ya que se puedan encontrar dentro de los sets de códigos que proponen los estudios mencionados anteriormente. Finalmente, se agrega la variable de respuesta "eam" de acuerdo si el *diag1* y/o el *diag2* están en los sets de códigos propuestos por los autores.

6.3.1. Relación entre las bases de datos

En la Figura 6.2 se evidencia el modelo conceptual de datos que evidencia las siguientes relaciones entre las bases de datos para construir la base de datos global:

- **Código Territorial - Farmacia**: entidades relacionadas a través del código de la comuna, su relación indica que para cada comuna (código territorial) existe un único registro en la base de datos de farmacias, asimismo ocurre en la otra dirección, cada registro de farmacia está asociado solo a uno código territorial.
- **Código Territorial - SINIM**: entidades relacionadas a través del código de la comuna, su relación indica que existe un y solo un código territorial en los archivos con origen en SINIM, asimismo existe un y solo un código territorial asociado a cada archivo SINIM.
- **Código Territorial - Egreso Hospitalario**: entidades relacionadas a través del código de la comuna, su relación indica que un egreso hospitalario está asociado a una y solo una comuna, mientras que existen cero (dado que la comuna no posea establecimientos de salud que reporten), uno o varios registros de egreso hospitalario que poseen una misma comuna asociada.
- **Código Territorial - Establecimiento**: entidades relacionadas a través del código comunal, la relación indica que existen uno o muchos establecimientos de salud por comuna, pero cada establecimiento está asociado a una comuna en particular.
- **Egreso Hospitalario - Listado EAM**: esta relación existe sólo si el diagnóstico principal y secundario son atribuibles a EAM según cada enfoque de estudio. La relación, en caso de existir, es a través del

diagnóstico principal (diag1) y/o diagnóstico secundario (diag2) la que indica que un código EAM puede estar asociado a ninguno, uno o varios registros de egresos hospitalarios, y un registro de egreso puede corresponderse con ningúno o un código atribuible a EAM.

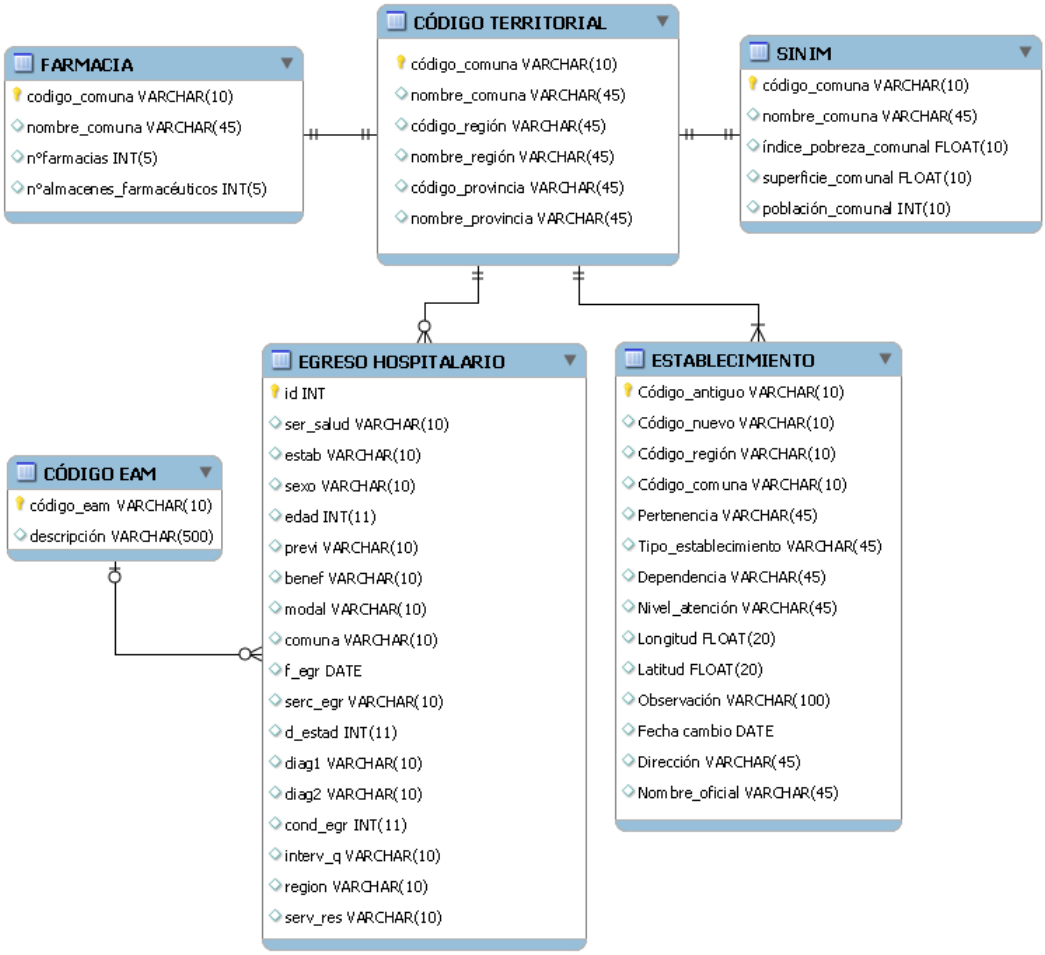


Figura 6.2: Modelo Conceptual de Datos

Fuente: Elaboración Propia

En conclusión, las variables que se utilizan en la siguiente fase son:

- ***año***, responde al año de ingreso del paciente, calculado a partir de la base de egresos hospitalarios.
- ***sexo1***, variable categórica correspondiente a 1 cuando el paciente es hombre, y 0 en otro caso. Generada a partir de la información de la base de datos de egresos hospitalarios.
- ***sexo2***, variable categórica correspondiente a 1 cuando el paciente es mujer, y 0 en otro caso. Generada a partir de la información de la base de datos de egresos hospitalarios.
- ***edad***, responde a la edad del paciente en años, extraída de la base de egresos hospitalarios.

- ***previ1***, variable categórica correspondiente a 1 cuando el paciente consta con previsión FONASA, y 0 en otro caso. Generada a partir de la información de la base de datos de egresos hospitalarios.
- ***previ2***, variable categórica correspondiente a 1 cuando el paciente consta con previsión ISAPRE, y 0 en otro caso. Generada a partir de la información de la base de datos de egresos hospitalarios.
- ***previ3***, variable categórica correspondiente a 1 cuando el paciente no cuenta con previsión, y 0 en otro caso. Generada a partir de la información de la base de datos de egresos hospitalarios.
- ***benef1***, variable categórica que toma el valor 1 cuando el paciente es beneficiario del tramo A de FONASA, y 0 en otro caso. Generada a partir de los datos de la base de egresos hospitalarios.
- ***benef2***, variable categórica que toma el valor 1 cuando el paciente es beneficiario del tramo B de FONASA, y 0 en otro caso. Generada a partir de los datos de la base de egresos hospitalarios.
- ***benef3***, variable categórica que toma el valor 1 cuando el paciente es beneficiario del tramo C de FONASA, y 0 en otro caso. Generada a partir de los datos de la base de egresos hospitalarios.
- ***benef4***, variable categórica que toma el valor 1 cuando el paciente es beneficiario del tramo D de FONASA, y 0 en otro caso. Generada a partir de los datos de la base de egresos hospitalarios.
- ***f_ing***, variable correspondiente al mes de ingreso del paciente, calculada a partir de los datos de la base de egresos hospitalarios.
- ***region***, responde a la región de residencia del paciente, enumeradas de acuerdo al listado de códigos únicos territoriales.
- ***poblacion***, corresponde a la población estimada según diversos tramos de edad, proveniente de las proyecciones de población elaboradas con los datos demográficos del INE, extraído de SINIM.
- ***superficie***, responde a la superficie total que tiene la comuna, medido en km^2, extraído de SINIM.
- ***pobreza***, porcentaje que responde a las estimaciones de la tasa de pobreza por ingresos por comuna de la encuesta CASEN, extraído de SINIM.
- ***farma***, representa el número de farmacias por comuna, extraído del listado de distribución de farmacias.
- ***almacen***, representa el número de almacenes farmacéuticos por comuna, extraído del listado de distribución de farmacias.
- ***primario***, representa el número de establecimientos de atención primaria por comuna, extraído del listado de establecimientos.
- ***secundario***, similar a la anterior, pero contempla los establecimientos del nivel secundario de atención.
- ***terciario***, variable similar a las dos anteriores, pero contempla los establecimientos de atención terciaria.

- ***eam***, variable categórica que considera la clase 1 si el egreseo de un paciente está asociado a un EAM, y la clase 0 si no existe relación.

6.4. Modelado

Para tener una perspectiva general de los registros, se determina la cantidad de egresos hospitalarios que son atribuibles a EAM tras la reducción de instancias debido a la limpieza de datos, se cuenta finalmente con 90 396 casos EAM para el caso de Stausberg y 31 598 para el caso de Wu *et al*, el desglose de cada año se puede observar en la Tabla A.2 presente en anexos.

Primeramente, se grafica para evidenciar magnitud de los efectos adversos a medicamentos, encontrando que sólo un 0.68 % de los registros de egresos responden a daño por medicamentos según Stausberg y un 0.24 % responden a EAM según el enfoque de Wu *et al* como se puede ver en la siguiente imagen.

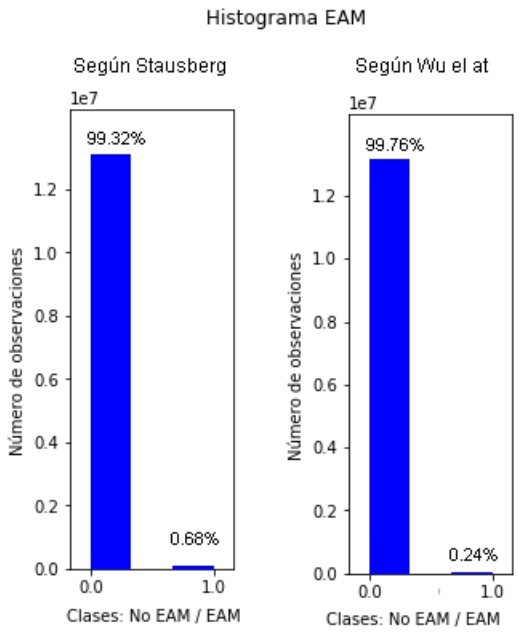


Figura 6.3: Histograma de clases

Fuente: Elaboración propia

Se decide utilizar el enfoque de Stausberg, ya que es capaz de detectar un mayor número de casos relacionados al daño por medicamentos. También es necesario mencionar, que para el desbalance de los datos se utiliza el ajuste de parámetros propios de los algoritmos a través de la estrategia de penalizar para compensar, ya que en un estudio comparativo de diversas estrategias, es el que obtiene el mejor desempeño (Na8, 2019).

6.4.1. Regresión Logística

En primer lugar, se ejecuta el algoritmo de la regresión logística mediante *Sklearn* con todas las variables y sin ningún tipo de ajuste, cabe mencionar que se utiliza un 80 % de los datos para entrenar el modelo y el 20 % restante se utiliza para probar el rendimiento, esto con el objetivo de obtener un modelo robusto (Gironés et al., 2017, pág. 75), obteniendo la siguiente matriz de confusión y curva ROC:

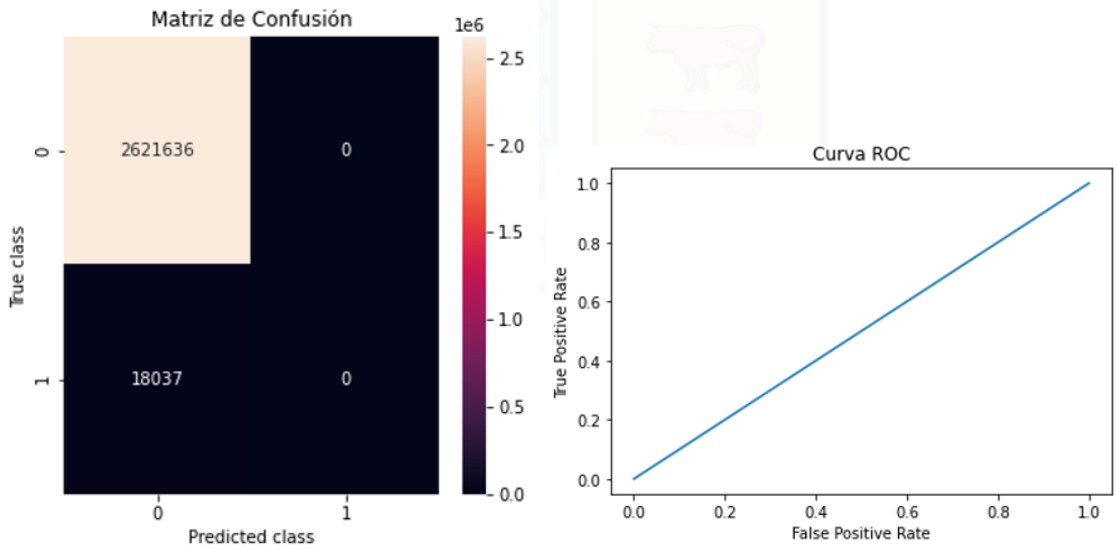


Figura 6.4: Resultados regresión logística sin balancear

Fuente:Elaboración propia

De la matriz se pueden desprender las siguientes métricas de evaluación:

Clase/ Métrica	Precisión	Recall
Clase 0	0.99	1
Clase 1	0	0

Tabla 6.6: Métricas Regresión logística sin balancear

Fuente: Elaboración propia

También se puede obtener la exactitud, cuyo valor es de 0.99, y el área bajo la curva ROC es de 0.5. De las gráficas y métricas, se desprende que el modelo está mal ajustado y que es capaz de predecir muy bien la clase 0, es decir, que el egreso de un paciente no es atribuible a EAM, sin embargo, esto ocurre por el desbalance de los datos como se puede observar en la Figura 6.3 generando un completo sesgo a la clase mayoritaria, si bien la exactitud es un valor muy alto esto se debe a que el modelo no fue capaz de predecir en ningún momento la clase 1, y solo se consideran los aciertos de la clase mayoritaria. Es por ello, que se ajusta el modelo nuevamente, pero se decide utilizar la metodología de penalizar la clase mayoritaria en el proceso

de entrenamiento, con el objetivo que un mayor número de instancias de la clase minoritaria entren en esta etapa y se pueda precedir de mejor manera, donde se obtiene las siguientes gráficas:

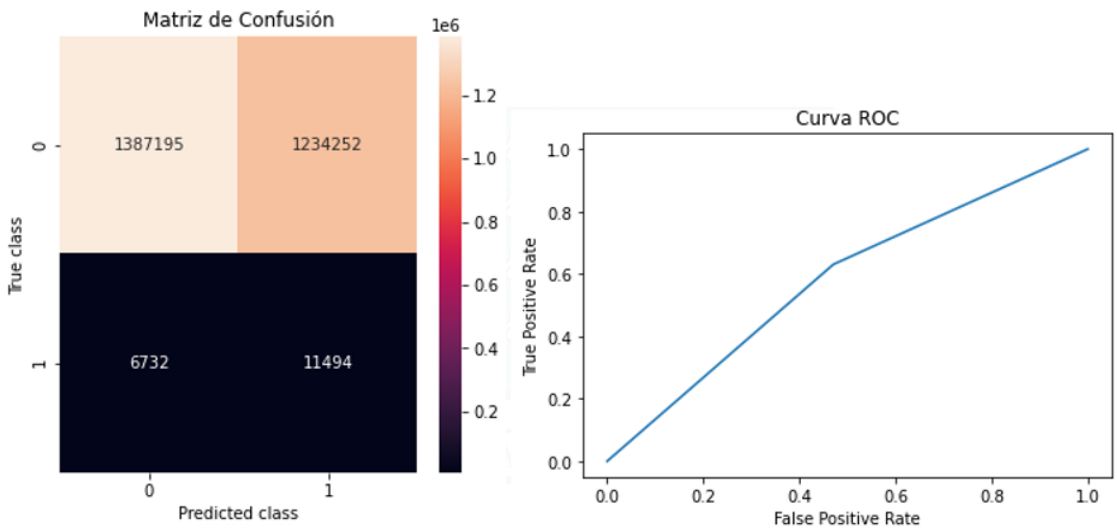


Figura 6.5: Resultados regresión logística balanceada

Fuente:Elaboración propia

De la matriz de confusión se puede observar que los aciertos de la clase 1 (*True Negative*) aumentan considerablemente en relación al caso sin balancear, sin embargo, también existen muchos ceros reales predichos como 1 generando que la sensibilidad para la clase 1 mejore en esta oportunidad, pero disminuyen para la clase 0, los resultados obtenidos son:

Clase/ Métrica	Precisión	Recall
Clase 0	1.00	0.53
Clase 1	0.01	0.63

Tabla 6.7: Métricas Regresión logística balanceada

Fuente: Elaboración propia

El desempeño de la curva ROC mejora levemente, alcanzando un valor de 0.5799, apesar que se perciben leves mejorías en las métricas no son las óptimas. Del modelo se puede obtener la importancia que se le asigna a las variables para la clasificación, las más negativas son las que ayudan a predecir con mayor fuerza la clase cero, y las más positivas son las que ayudan a predecir la clase 1, en el siguiente gráfico se puede observar la importancia en esta etapa de la construcción del modelo.

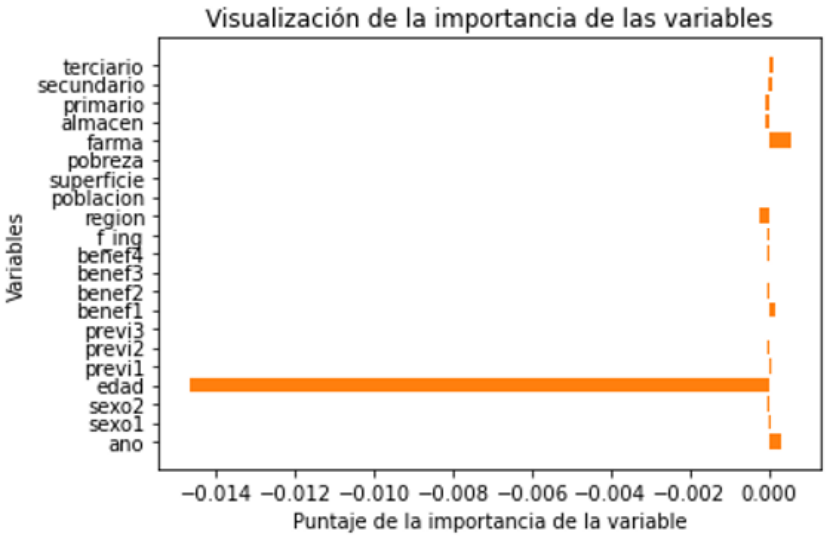


Figura 6.6: Importancia de las variables Regresión Logística Balanceada

Fuente: Elaboración propia

Se puede desprender que la variable correspondiente a la edad del paciente, ayuda notoriamente a que un caso no sea atribuible a EAM. También se puede observar que existen variables con mínimo aporte en la decisión de si es EAM o no, por lo que se deciden remover las que poseen la mínima influencia (en magnitud) en el modelo, en este paso se eliminan las variables: *previ3, benef3, poblacion, pobreza* y *superficie* del gráfico presentado. Se ajusta nuevamente el modelo, donde se obtienen las siguientes métricas de evaluación:

Clase/ Métrica	Precisión	Recall
Clase 0	1.00	0.60
Clase 1	0.01	0.60

Tabla 6.8: Métricas tras primera reducción de dimensionalidad

Fuente: Elaboración propia

Donde la exactitud alcanza un valor de 0.59, y el área bajo la curva ROC es de 0.5995, mejorando levemente en relación a la iteración anterior, se vuelve a graficar la importancia de las variables y se obtiene la Figura 6.7.

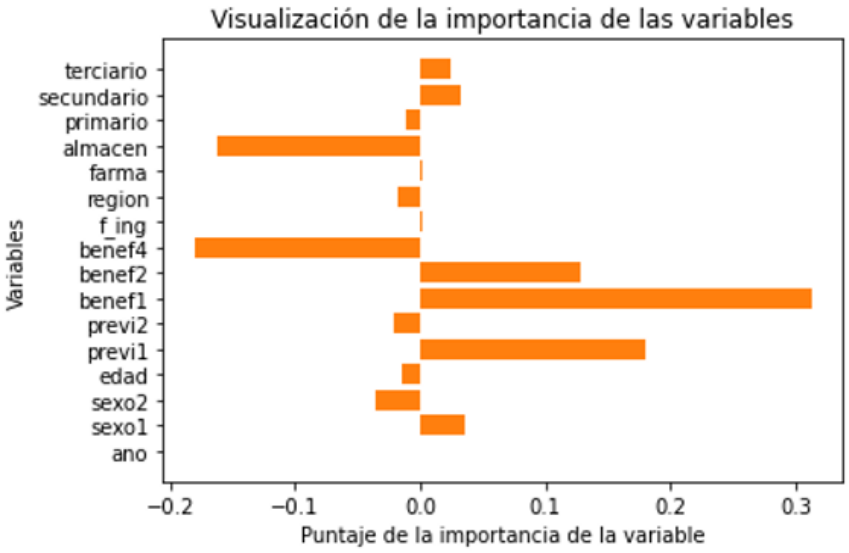


Figura 6.7: Importancia de las variables tras primera reducción de dimensionalidad

Fuente: Elaboración propia

Se puede observar que las variables correspondientes a *año, f_ing* y *farma*, son las que presentan menor magnitud en la importancia. Los valores de la precisión y recall no se ven modificados, y la curva ROC aumenta ínfimamente a 0.6021. En una nueva iteración, las variables candidatas a salir son las variables *edad, region, primario* y *secundario*, tras quitarlas del modelo, el desempeño baja bruscamente a 0.5521, por lo que se decide elimininar de una en una, a fin de determinar cual es la que afecta el rendimiento, encontrando que la variable *edad* es la que provoca el desajuste, por lo que se procede a eliminar *region, primario* y *secundario*, donde se obtiene un rendimiento de 0.6024 y la siguiente gráfica de importancia de variables:

Figura 6.8: Importancia de las variables, reducción 3

Fuente: Elaboración propia

Luego, se continúa eliminando variables para observar si mejora el comportamiento de las métricas derivadas de la matriz de confusión y la curva ROC, en las siguiente iteraciones se eliminan las variables *sexo2* y *terciario*, sin embargo, el modelo no logra mejorar y subir del 60 % de rendimiento, estas iteraciones se encuentran disponibles en anexos en la Subsubsección A.3.2.1. En conclusión, el mejor ajuste lo logra con las variables *sexo1, sexo2, edad, previ1, previ2, benef1, benef2, benef4, almacen* y *terciario*, sin embargo, es clasificado como un test regular según la curva ROC, y según los resultados de la Tabla 6.9 la sensibilidad indica que el algoritmo no detecta bien las clases, y según indica la precisión, cuando se detecta la clase 0 es confiable, pero cuando detecta el daño asociado a medicamentos la confiabilidad es casi nula.

Clase/ Métrica	Precisión	Recall
Clase 0	1.00	0.60
Clase 1	0.01	0.61

Tabla 6.9: Métricas Regresión logística tras reducción

Fuente: Elaboración propia

En la Tabla 6.10 , se presenta el puntaje asignado a cada características de las que influyen en que un egreso sea asociado a daño por medicamentos de acuerdo al mejor desempeño obtenido, donde se desprende que a mayor edad o pertenecer a las comunas con mayor número de almacenes farmacéuticos como Antofagasta, Cunco, El Carmen y Freire existe menor riesgo de sufrir efectos adversos a medicamentos, que sufrirlo. En cambio ser hombre (sexo1) es un factor de riesgo mayor que ser mujer (sexo2). Además, contar con FONASA como previsión indica mayor riesgo de sufrir un egreso hospitalario en comparación a contar con ISAPRES o previsiones de las Fuerzas Armadas y de Orden (u otra), asimismo ser de los tramos A o B, indica un mayor riesgo asociado al propio de ser FONASA, pero si un paciente es del fondo D ayuda a disminuir el riesgo propio de ser FONASA. Finalmente cabe mencionar que residir en comunas donde existe un mayor número de establecimientos de atención terciaria como Santiago y Providencia también supone un mayor riesgo de sufrir daño ocasionado por medicamentos.

N° Variable	Variable	Puntuación
0	sexo1	0.07941
1	sexo2	0.008639
2	edad	-0.01402
3	previ1	0.24004
4	previ2	0.19608
5	benef1	0.38860
6	benef2	0.21653
7	benef4	-0.12338
8	almacen	-0.14595
9	terciario	0.04039

Tabla 6.10: Puntuación variables influyentes
Fuente: Elaboración propia

Desde otra perspectiva, se ejecuta al algoritmo de la regresión logística con el paquete de *Statsmodels*, se ajusta el modelo completo con intercepto con todas las variables obteniendo los siguientes resultados, la metodología de eliminación de variables es a través de la significancia estadística:

```
                           Logit Regression Results
==============================================================================
Dep. Variable:                      y   No. Observations:             10558690
Model:                          Logit   Df Residuals:                 10558668
Method:                           MLE   Df Model:                           21
Date:                Sun, 30 Aug 2020   Pseudo R-squ.:                 0.01542
Time:                        19:36:32   Log-Likelihood:            -4.2511e+05
converged:                       True   LL-Null:                   -4.3177e+05
Covariance Type:            nonrobust   LLR p-value:                    0.000
==============================================================================
                  coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
intercept      -3.8000      3.445     -1.103      0.270     -10.551       2.951
año            -0.0005      0.002     -0.321      0.748      -0.004       0.003
sexo1           0.3316      0.449      0.738      0.460      -0.549       1.212
sexo2           0.2332      0.449      0.519      0.604      -0.647       1.114
edad           -0.0136      0.000    -82.234      0.000      -0.014      -0.013
previ1          0.8579      0.081     10.609      0.000       0.699       1.016
previ2          0.2976      0.021     13.990      0.000       0.256       0.339
previ3          0.3517      0.032     10.951      0.000       0.289       0.415
benef1         -0.0986      0.079     -1.255      0.210      -0.253       0.055
benef2         -0.2895      0.079     -3.659      0.000      -0.445      -0.134
benef3         -0.5125      0.080     -6.417      0.000      -0.669      -0.356
benef4         -0.6448      0.080     -8.099      0.000      -0.801      -0.489
f_ing           0.0019      0.001      1.760      0.078      -0.000       0.004
region         -0.0209      0.001    -18.349      0.000      -0.023      -0.019
poblacion   -1.982e-07   4.13e-08     -4.802      0.000   -2.79e-07   -1.17e-07
superficie     -0.3425      0.047     -7.217      0.000      -0.436      -0.250
pobreza        -0.0080      0.001    -12.832      0.000      -0.009      -0.007
farma          -0.0001      0.000     -0.723      0.469      -0.000       0.000
almacen        -0.1476      0.009    -17.114      0.000      -0.164      -0.131
primario       -0.0041      0.001     -6.113      0.000      -0.005      -0.003
secundario      0.0152      0.004      3.650      0.000       0.007       0.023
terciario       0.0541      0.005     11.131      0.000       0.045       0.064
==============================================================================
```

Figura 6.9: Logit completo (Statsmodels)

Fuente: Elaboración propia

Del resumen anterior, es de particular interés la primera y quinta columna que corresponden a los parámetros (coef β) y p-valor (P>|z|) respectivamente. Como se puede observar, en este ajuste las variables *año, sexo1, sexo2, benef1, f_ing* y *farma*, junto al intercepto, no son significativas, ya que el p-valor es mayor a 0.05 indicando que no existe asociación entre la variable de respuesta y dichas predictoras, por lo que se retiran del modelo. Luego, se ajusta nuevamente el modelo con el resto de las variables, donde se obtiene el siguiente resumen:

```
                           Logit Regression Results
==============================================================================
Dep. Variable:                      y   No. Observations:             10558690
Model:                          Logit   Df Residuals:                 10558675
Method:                           MLE   Df Model:                           14
Date:                Sun, 30 Aug 2020   Pseudo R-squ.:                -0.04157
Time:                        19:37:47   Log-Likelihood:            -4.5082e+05
converged:                       True   LL-Null:                   -4.3282e+05
Covariance Type:            nonrobust   LLR p-value:                    1.000
================================================================================
                  coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------
edad           -0.0253      0.000   -168.113      0.000      -0.026      -0.025
previ1         -1.0727      0.010   -109.601      0.000      -1.092      -1.054
previ2         -1.5676      0.012   -135.155      0.000      -1.590      -1.545
previ3         -1.5649      0.027    -58.636      0.000      -1.617      -1.513
benef2         -0.0522      0.010     -5.171      0.000      -0.072      -0.032
benef3         -0.4359      0.015    -29.108      0.000      -0.465      -0.407
benef4         -0.5174      0.013    -38.577      0.000      -0.544      -0.491
region         -0.1530      0.001   -187.956      0.000      -0.155      -0.151
poblacion   -1.157e-06   4.24e-08    -27.268      0.000   -1.24e-06   -1.07e-06
superficie     -3.4978      0.049    -71.478      0.000      -3.594      -3.402
pobreza        -0.0667      0.001   -120.272      0.000      -0.068      -0.066
almacen        -0.0680      0.007     -9.998      0.000      -0.081      -0.055
primario       -0.0345      0.001    -47.447      0.000      -0.036      -0.033
secundario      0.1028      0.004     25.064      0.000       0.095       0.111
terciario      -0.0543      0.004    -14.328      0.000      -0.062      -0.047
================================================================================
```

Figura 6.10: Logit reducción 1 (Statsmodels)

Fuente: Elaboración propia

Si bien, las variables son significativas, se puede ver que los coeficientes son negativos para todas las variables, excepto para el número de establecimientos de salud de nivel secundario, esto indica que dicha variable es la única que ayuda a predecir la clase 1, y el resto aporta a la clase mayoritaria. Si se observa el valor del pseudo-R^2 se puede notar que baja drásticamente al reducir la dimensión del modelo, sin embargo el primer valor tampoco indica un buen ajuste, ya que según indica la literatura valores entre 0.2 y 0.4 son muy buenos modelos, en equivalencia lo que sería un R^2 en el rango de 0.7 y 0.9 de una regresión lineal (Louviere et al., 2000).

Continuando, se prueba generar un modelo sin intercepto en su origen, donde se obtiene el resumen de la Subsubsección A.3.2.1 presente en anexos, donde las variables no significativas responden a *sexo1, sexo2, benef1* y *farma*, por lo que se sacan del modelo y se ajusta nuevamente. Tras ejecutar el modelo, se obtiene que nuevamente se encuentra una variable no significativa en el modelo, es el caso de la predictora *f_ing* por lo que se retira, y se realiza nuevamente el procedimiento de ajuste, logrando que todas las variables sean significativas como se muestra a continuación.

```
                           Logit Regression Results
==============================================================================
Dep. Variable:                      y   No. Observations:             10558690
Model:                          Logit   Df Residuals:                 10558674
Method:                           MLE   Df Model:                           15
Date:                Sun, 30 Aug 2020   Pseudo R-squ.:                 0.01536
Time:                        20:53:51   Log-Likelihood:            -4.2602e+05
converged:                       True   LL-Null:                   -4.3267e+05
Covariance Type:            nonrobust   LLR p-value:                     0.000
==================================================================================
                   coef    std err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------------------
año           -0.0023   1.27e-05   -179.800      0.000      -0.002      -0.002
edad          -0.0136      0.000    -81.840      0.000      -0.014      -0.013
previ1         0.7644      0.020     37.347      0.000       0.724       0.805
previ2         0.3020      0.021     14.158      0.000       0.260       0.344
previ3         0.3707      0.032     11.579      0.000       0.308       0.433
benef2        -0.1996      0.010    -19.427      0.000      -0.220      -0.179
benef3        -0.4150      0.015    -27.865      0.000      -0.444      -0.386
benef4        -0.5394      0.013    -40.110      0.000      -0.566      -0.513
region        -0.0216      0.001    -19.031      0.000      -0.024      -0.019
poblacion   -2.222e-07   3.69e-08     -6.026      0.000   -2.94e-07    -1.5e-07
superficie    -0.3529      0.047     -7.487      0.000      -0.445      -0.260
pobreza       -0.0082      0.001    -14.228      0.000      -0.009      -0.007
almacen       -0.1547      0.009    -17.840      0.000      -0.172      -0.138
primario      -0.0044      0.001     -6.605      0.000      -0.006      -0.003
secundario     0.0197      0.004      4.771      0.000       0.012       0.028
terciario      0.0533      0.004     14.211      0.000       0.046       0.061
==================================================================================
```

Figura 6.11: Logit sin intercepto reducción 2 (Statsmodels)

Fuente: Elaboración propia

Si bien el resultado del pseudo-R^2 es mejor que en el caso sin intercepto, no logra acercarse a un buen modelo, aunque autores afirman que debe evitarse este coeficiente para resumir y sacar conclusiones del modelo (Gujarati, 2004, pág. 567). A continuación, se presenta una tabla con los coeficientes y odds ratio obtenidos (Exp(β)).

Variable	Coeficiente (β)	Exp(β)	1/ Exp(β)
Año	-0.0023	0.998	1.002
Edad	-0.0136	0.986	1.014
Previ1	0.7644	2.148	—
Previ2	0.3020	1.353	—
Previ3	0.3707	1.449	—
Benef2	-0.1996	0.819	1.221
Benef3	-0.4150	0.660	1.514
Benef4	-0.5394	0.583	1.715

Tabla 6.11: Variables Statsmodels Final (primera parte)

Fuente: Elaboración Propia

Variable	Coeficiente (β)	Exp(β)	1/ Exp(β)
Region	-0.0216	0.979	1.022
Poblacion	-2.222e-7	1.000	1.000
Superficie	-0.3529	0.703	1.423
Pobreza	-0.0082	0.992	1.008
Almacen	-0.1547	0.857	1.167
Primario	-0.0044	0.996	1.004
Secundario	0.0197	1.020	—
Terciario	0.0533	1.055	—

Tabla 6.12: Variables Statsmodels Final (segunda parte)

Fuente: Elaboración Propia

Cuando el odds ratio es menor que 1, es conveniente sacar su inversa para la interpretación, y en general, cuando el valor es muy cercano a 1, significa que no tiene mayor relevancia en predecir una clase tras un cambio en la variable asociada. Es relevante el valor obtenido para la previsión de FONASA (previ1), donde nos indica que existe el doble de riesgo de sufrir un EAM, que no sufrirlo, si contamos con este tipo de seguro de salud, adicional a ello los tramos de B-C-D de FONASA indican que son un factor protector, ya que ayudan a reducir el riesgo intrínsico de FONASA, por lo que se podría interpretar que el tramo A es aquel con mayor riesgo. Es relevante mencionar que independiente de la previsión de salud, existe un riesgo asociado en ellas, solo varía la magnitud del riesgo.

Por otro lado, la variable superficie indica que por cada km^2 que aumente en un territorio existe un 1.4 veces menos riesgo de sufrir un EAM, por lo tanto las comunas con menor superficie territorial como Independencia, Lo Prado, Lo Espejo y San Ramón (todas pertenecientes a la Región Metropolitana) son las que presentarían mayor riesgo asociado a medicamentos.

6.4.2. Bosques Aleatorios

En segundo lugar, se ejecuta el algoritmo de Bosques Aleatorios, donde en cada iteración se construyen 100 árboles de decisión, cada uno con $\sqrt{d}$ características que se deben considerar al buscar la mejor división, siendo d la cantidad de variables predictoras (Orellana, 2018). El algoritmo se ejecuta tanto sin ajuste de parámetros (matriz izquierda en la imagen) como balanceado (matriz derecha en la imagen) con todas las variables, obteniendo las siguientes matrices de confusión.

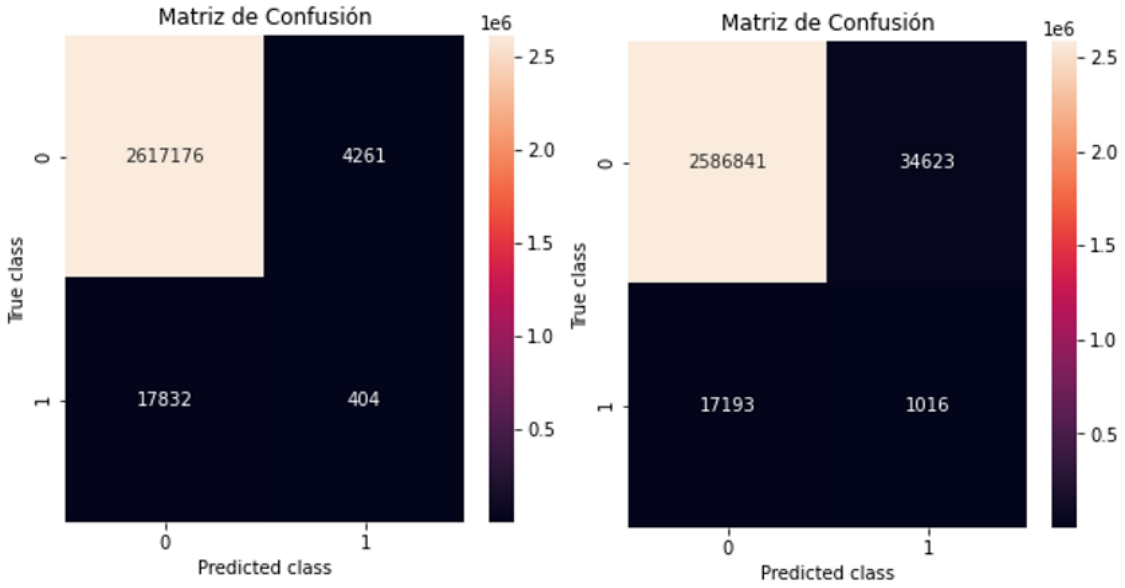


Figura 6.12: Random Forest: Primeras ejecuciones

Fuente: Elaboración propia

En el caso sin ajuste de parámetros se obtiene un área bajo la curva ROC de 0.510, una precisión de 0.09 y una sensibilidad de 0.02 para la clase 0. Por otro lado, en el caso balanceado, el rendimiento de la curva ROC alcanza un 0.521 y el OOB score adquiere un valor de 0.981, lo que indica que al ingresar nuevos valores al algoritmo como lo es el conjunto de prueba, este no es capaz de reconocerlos y predecir de manera adecuada, sin embargo al utilizar un conjunto de datos conocidos para ciertos árboles, si es capaz de predecir correctamente. La precisión y sensibilidad para el caso balanceado alcanzan valores de un 0.03 y 0.06 respectivamente, indicando que no los resultados no son buenos, y tampoco confiables.

Se procede a trabajar con el caso balanceado, eliminando las variables no influyentes según la importancia de las variables, a partir de la siguiente gráfica obtenida al ejecutar el algoritmo con todas las variables:

Visualización de la importancia de las variables

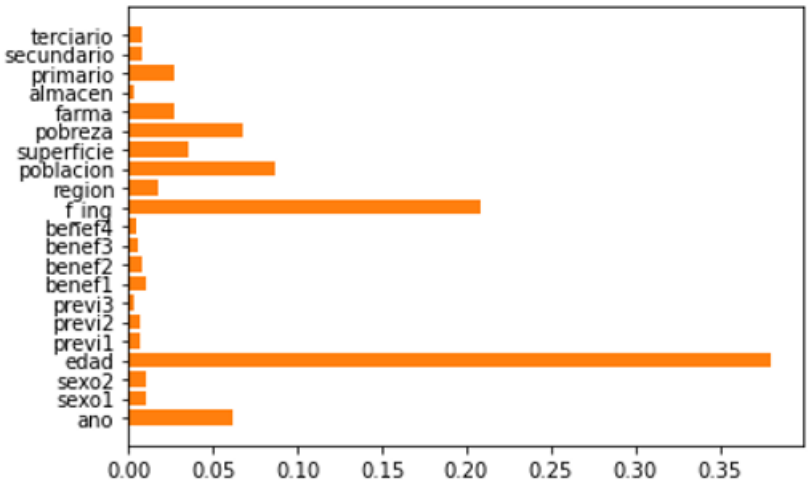


Figura 6.13: Bosques Aleatorios, todas las variables

Fuente: Elaboración propia

Se eliminan las variables *previ3,benef4* y *almacen*, del gráfico presentado ya que son las que poseen menor importancia en la clasificación.

- **Primera reducción** → AUC: 0.522 - OOB score:0.981

  El desempeño del ajuste mejora levemente, por lo que se decide realizar una segunda reducción, de acuerdo al gráfico presentado más abajo, las variables candidatas a salir en esta iteración son *previ1, previ2, secundario* y *terciario*.

Visualización de la importancia de las variables

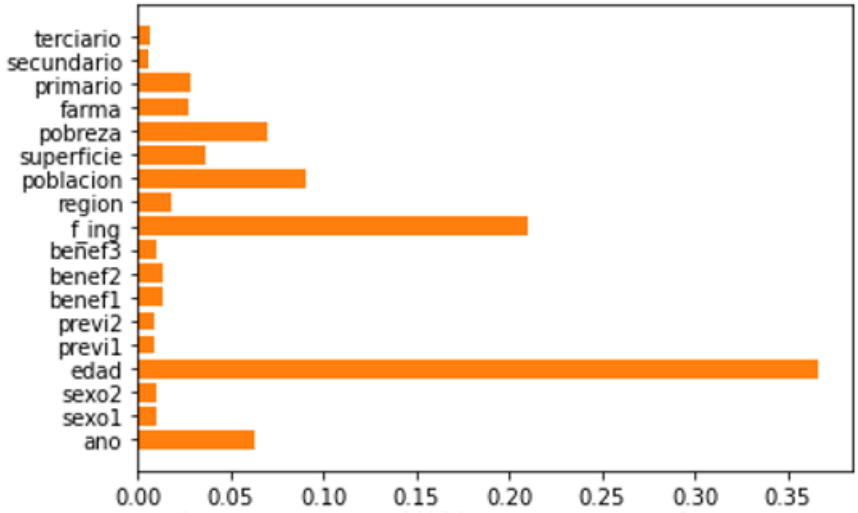


Figura 6.14: Bosques Aleatorios, reducción 1

Fuente: Elaboración propia

- **Segunda reducción** → AUC: 0.521 - OOB score: 0.976

Visualización de la importancia de las variables

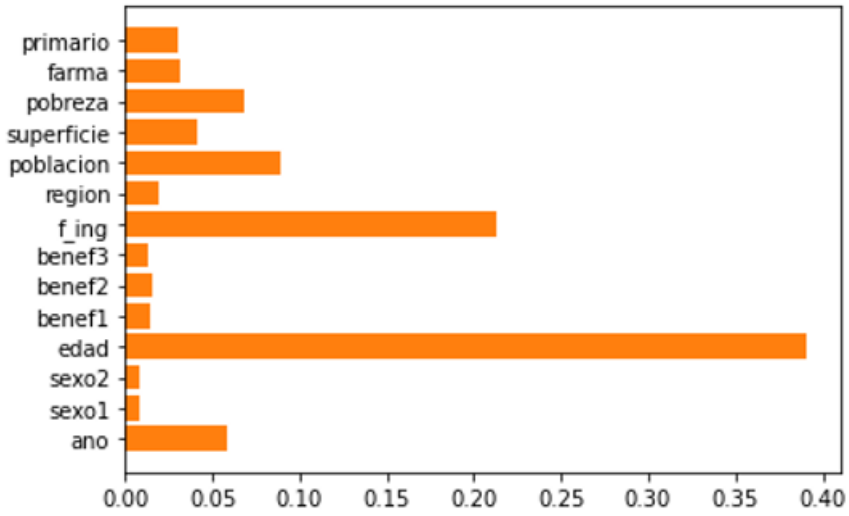


Figura 6.15: Bosques Aleatorios, reducción 2

Fuente: Elaboración propia

El rendimiento no logra grandes cambios, por lo que se continúa eliminando variables, en esta ocasión las candidatas a salir responden a las variables a *sexo1* y *sexo2* del gráfico expuesto.

Tras ejecutar el algoritmo sin las variables *sexo1* y *sexo2*, se obtiene un desempeño de la curva ROC de 0.525 y OOB score de 0.964, por lo que el algoritmo no logra mejorar. Dadas las expectativas en el ajuste y el tiempo que involucra su ejecución no se siguen intentando más reducciones. De acuerdo a los

valores obtenidos de la curva ROC el modelo no logra salir de un mal ajuste, y de acuerdo a las métricas derivadas de la matriz de confusión no logra clasificar bien las clases, y tampoco es confiable cuando lo hace. Con respecto al OOB score, se desprende que el modelo no es capaz de realizar buenas predicciones, los gráficos obtenidos la curva ROC y las matrices de confusión de cada iteración, se encuentran disponibles en Subsubsección A.3.2.3 de anexos.

6.5. Análisis del modelado

Los resultados obtenidos para ambos modelos responden a una deficiente capacidad discriminativa, por lo que en un esfuerzo de mejorar los resultados obtenidos, se decide agrupar algunas variables, para estimar si la dimensionalidad genera confusión al momento de entrenar el algoritmo, es por ello que el cuociente entre la variable población y superficie se convierten en densidad poblacional, también se agrupan las variables de farmacias y almacenes farmacéuticos por comuna, dejándo sólo como establecimientos farmacéuticos (estab_farma), y también se agrupan los establecimientos de atención secundaria y terciaria (*sec_ter*), para distinguir los centros de atención primaria con el resto de los niveles de atención. Finalmente se decide juntar la previsión de FONASA con sus respectivos tramos, quedando como *previ1A, previ1B, previ1C* y *previ1D*, los otros sistemas de seguros de salud previ2 y *previ3* permenecen intactos, en conclusión de 21 variables, se trabaja en una segunda oportunidad con 17 en total.

6.6. Modelado: Segunda parte

6.6.1. Regresión Logística

Se utiliza el mismo procedimiento anterior, buscando el mejor rendimiento de la curva ROC, y los valores más altos de las métricas derivadas de la matriz de confusión en la ejecución con Sklearn. En la primera iteración con todas las variables se obtiene un rendimiento de la curva ROC de 0.6008, superior al valor obtenido con las variables antiguas pero se comienza a iterar para buscar un mejor rendimiento, por lo que las primeras candidatas a eliminar son las variables *año, f_ing, densidad* y *estab_farma*.

Luego, en la segunda iteración, se obtiene un AUC de 0.6017 y que se debe eliminar la variable *pobreza*. Más adelante, en la tercera iteración se obtiene un 0.6033 de rendimiento, y las variables *region* y *primario* son las que salen del modelo. Continuando, en la cuarta iteración se obtiene un rendimiento de 0.6041, para finalmente eliminar la variable *sec_ter* y *edad*, bajando el rendimiento a un 0.5538, se prueba eliminando solo la variable *sec_ter*, ya que *edad* es la que ocasiona la baja en el rendimiento según lo visto en el caso anterior. Sin embargo, el rendimiento no es mejor que en la última iteración. A continuación, se

presentan las métricas antes de realizar la última reducción, ya que es donde se obtiene el mejor desempeño con las variables *sexo1, sexo2, edad, previ1A, previ1B, previ1C, previ1D, previ2, previ3* y *sec_ter*.

Clase/ Métrica	Precisión	Recall
Clase 0	1.00	0.60
Clase 1	0.01	0.61

Tabla 6.13: Métricas tras reducción de dimensionalidad, variables nuevas

Fuente: Elaboración propia

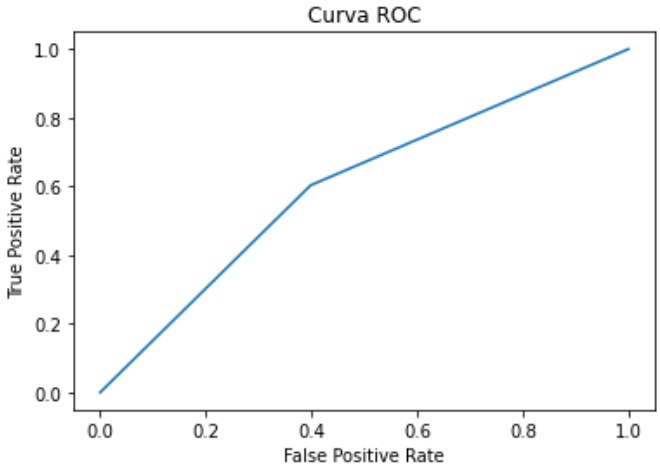


Figura 6.16: Curva ROC tras reducción con variables nuevas

Fuente: Elaboración propia

El valor de la curva ROC sigue reflejando un ajuste regular del test, y se evidencia el sesgo a la clase mayoritaria al estudiar la precisión y sensibilidad de cada clase, sin embargo, cambia el comportamiento de las variables que influyen en el algoritmo como se puede apreciar en la Tabla 6.14. A pesar de ello, la puntuación de la variable que representa el sexo1 se traduce en que existe 1.046 veces más riesgo de sufrir daño por medicamentos, que no sufrirlo al ser hombre, por otro lado, ser mujer sería un factor protector. En cuanto a los seguros de salud, se puede extraer que pertenecer al tramo A de FONASA representa el doble de riesgo de sufrir un EAM, que no sufrirlo en relación a otros seguros de salud. Sin embargo pertenecer a los tramos B y C, no se escapan de esta realidad, ya que ser FONASA y pertenecer al tramo B supone un 1.75 veces mayor riesgo de sufrir daño por medicamentos que no sufrirlo. Asimismo, las personas que cuentan con ISAPRE (previ2) y las que no poseen ningún tipo de cobertura (previ3) no están excentas de riesgo.

Variable	Puntuación	Exp(β)
sexo1	0.045	1.046
sexo2	-0.035	0.966
edad	-0.014	0.986
previ1A	0.741	2.098
previ1B	0.564	1.758
previ1C	0.337	1.401
previ1D	0.217	1.242
previ2	0.303	1.354
previ3	0.349	1.418
sec_ter	0.010	1.010

Tabla 6.14: Puntuación variables influyentes (segunda parte)

Fuente: Elaboración propia

Asimismo, se ejecuta el modelo a través de Statsmodels con el cambio de variables y probando con un modelo con y sin intercepto. De los ajustes, se obtiene que el modelo que inicia con intercepto se comporta de la misma manera descrita con las variables antiguas, por lo que no se considera un buen ajuste, los resultados se pueden observar en la Subsubsección A.3.2.2 presentes en anexos. Con las variables nuevas y sin intercepto, se elimina en cada iteración las variables no significativas, y se concluye que no mejora el pseudo-R^2, en comparación al 0.01536 obtenido en el caso de la segunda reducción sin intercepto con las variables antiguas.

6.6.2. Bosques Aleatorios

Igual que en el caso de la regresión logística, se trabaja con las nuevas variables para probar si mejora el rendimiento del algoritmo, donde se obtienen los siguientes resultados:

- **Ejecución con todas las variables** → AUC: 0.522 - OOB score: 0.9811

  El rendimiento inicial no varía en comparación al caso antiguo, pero se continúa iterando para verificar si sigue el mismo comportamiento o mejora, por lo que se retiran del modelo las variables *sexo1, previ1C, previ1D* y *previ3* de acuerdo a los resultados del siguiente gráfico:

Visualización de la importancia de las variables

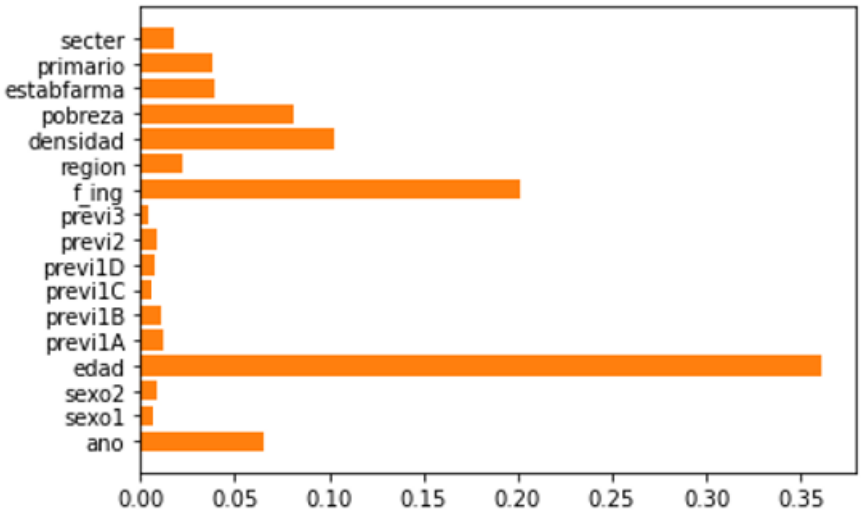


Figura 6.17: Bosques Aleatorios, todas las variables, variables nuevas

Fuente: Elaboración propia

- **Primera reducción** → AUC: 0.521 - OBB score: 0.9787

Visualización de la importancia de las variables

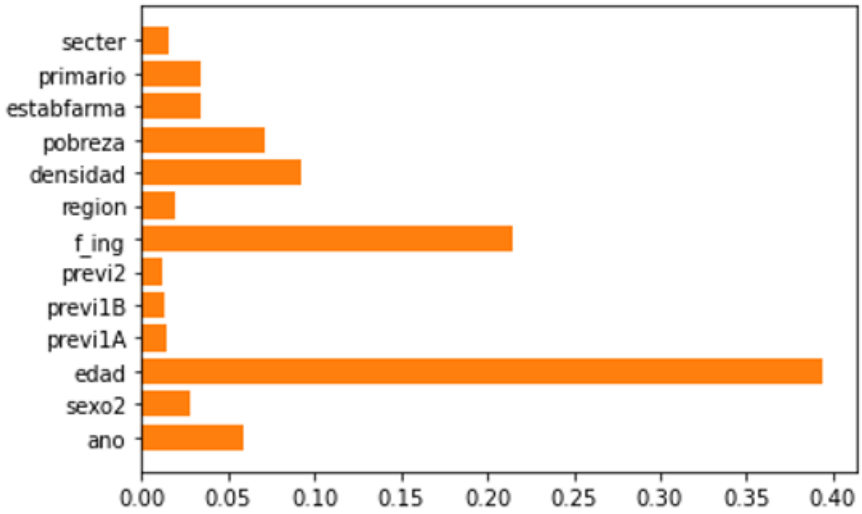


Figura 6.18: Bosques Aleatorios, primera reducción, variables nuevas

Fuente: Elaboración propia

No se perciben diferencias aún en el rendimiento, por lo que se eliminan las variables *previ1A, previ1B* y *previ2* en esta ocasión.

- **Segunda reducción** → AUC: 0.530 - OOB score: 0.9550

Visualización de la importancia de las variables

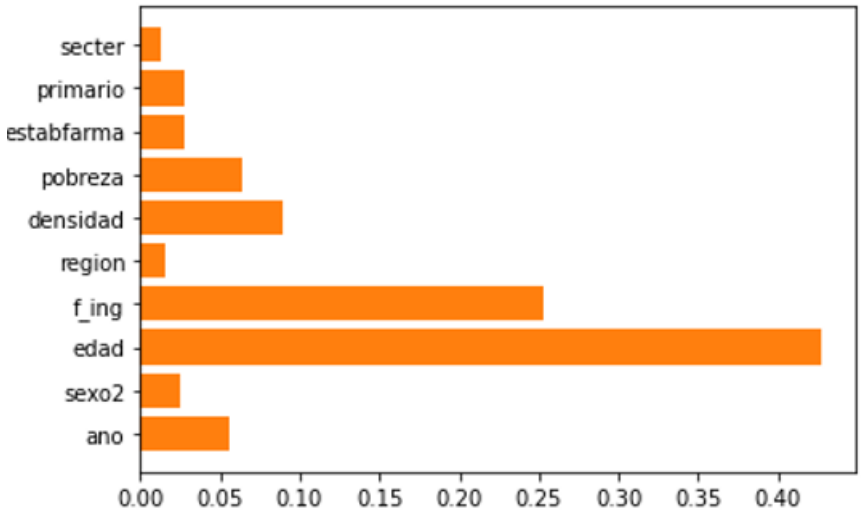


Figura 6.19: Bosques Aleatorios, segunda reducción, variables nuevas

Fuente: Elaboración propia

Sube levemente el valor del área bajo la curva ROC, pero aún en un mal rendimiento, se decide continuar con la eliminación de variables *region* y *sec_ter*.

El rendimiendo de la curva ROC sigue en un desempeño deficiente al eliminarse las últimas variables propuestas, catalogándose como un test malo con un 0.531 y un OOB score de 0.9552, por lo que se decide no seguir realizando iteraciones en vista que no mejoran de manera relevante los indicadores. Si bien el error fuera de la bolsa, indica buenos ajusten durante el proceso de entrenamiento, cuando se mide con datos desconocidos para algunos árboles, el modelo no es capaz de responder de manera óptima.

6.7. Análisis del modelado

Los resultados obtenidos para el algoritmo de la regresión logística se encuentran dentro del rango de un test regular y para bosques aleatorios es un mal desempeño, no se logran mejorar los resultados al utilizar la metodología de ajuste de parámetros para mitigar los datos desbalanceados. Lo mismo ocurre con el cambio de variables, por esta razón, se procede a probar con la metodología de crear muestras sintéticas, ya que los nuevos datos intentan seguir la tendencia del grupo minoritario, aunque esto puede alterar la distribución propia de los datos y confundir al modelo, por lo que no se prueba con la metodología de eliminar según la significancia estadística. En cambio, si se utiliza la metodología de duplicar o triplicar las instancias minoritarias, no se está agregando información nueva el modelo, y este puede caer en sobrentrenamiento, además de la propia alteración de la distribución.

6.8. Modelado: Tercera Parte

Se utiliza la técnica de sobremuestreo de minorías sintéticas , o SMOTE por sus siglas en inglés. La técnica radica en seleccionar una instancia de la clase minoritaria, y luego seleccionar los k vecinos más cercanos de su misma clase, se elije uno de estos puntos al azar y se traza una línea imaginaria en el espacio de las características. Luego, se crea un ejemplo sintético en un punto seleccionado al azar sobre la línea trazada.

En primer lugar, se prueba el rendimiento de cada modelo con distintos porcentajes de muestras sintéticas, se prueba iterando con valores que aumentan en 10 %, y se utilizan los 5 vecinos más cercanos, ya que es un valor que se utiliza típicamente y es el que viene por defecto en la librería *Imblearn* de Python (Brownlee, 2020).

A continuación, se muestran una tabla con los resultados obtenidos que responden al área bajo la curva ROC, es decir, al parámetro AUC, y un gráfico que muestra el comportamiento de este parámetro para los distintos tamaños de muestras sintéticas. Los ajustes de los modelos son a través de la librería de Sklearn con las variables nuevas.

Muestras Sintéticas	Regresión Logística	Bosques Aleatorios
0 %	0.500 ▲	0.521 ▲
10 %	0.731 ▲	0.937 ▲
20 %	0.786 ▲	0.963 ▲
30 %	0.829 ▲	0.971 ▲
40 %	0.834 ▲	0.975 ▲
50 %	0.833 ▲	0.977 ▲
60 %	0.833 ▲	0.978 ▲
70 %	0.835 ▲	0.979 ▲
80 %	0.833 ▲	0.980 ▲
90 %	0.833 ▲	0.980 ▲
100 %	0.835 ▲	0.983 ▲

▲: Test Malo ▲: Test Regular ▲: Test Bueno

▲: Test Muy Bueno ▲: Test Excelente

Tabla 6.15: Rendimiento con muestras sintéticas

Fuente: Elaboración propia

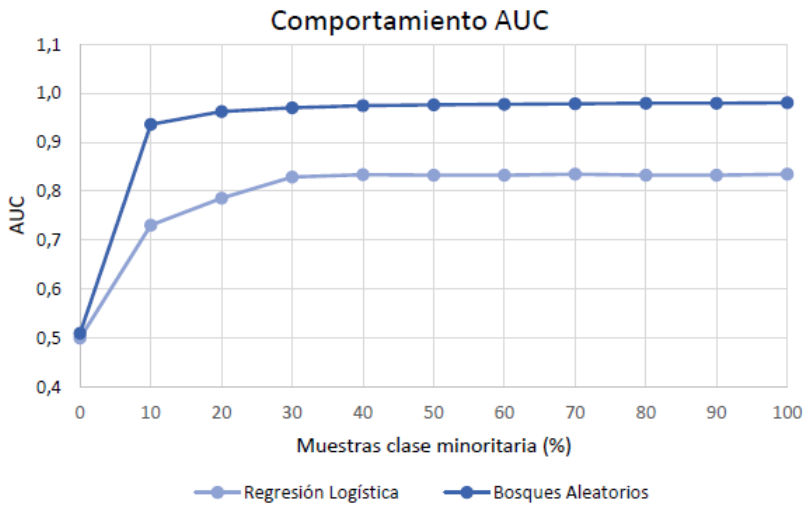


Figura 6.20: Comportamiento AUC a distintos niveles de muestras sintéticas

Fuente: Elaboración propia

De la tabla se desprende, que al ajustar al 10 % el tamaño de la clase minoritaria ambos modelos presentan un gran cambio en el rendimiento, en el caso de la regresión logística pasa de un test malo a un test regular, y tan solo con un 20 % de muestras de la clase 0 logra un buen rendimiento con todas las variables. Por otro lado, el rendimiento con bosques aleatorios aumenta drásticamente, pasando de un test malo a un test muy bueno, incluso llegando a alcanzar rendimientos excelentes desde un 30 % de los datos correspondientes a la clase minoritaria. Por otro lado, del gráfico se desprende que existe un número determinado de muestras de la clase minoritaria donde se converge al máximo rendimiento, por lo que no existe gran diferencia entre crear un 30 % de muestras sintéticas a tener un conjunto totalmente equilibrado para los dos algoritmos presentados.

Para evaluar los factores de riesgo, y la diferencia de ellos en los modelos, se utiliza un 10 % de muestras correspondientes a la clase minoritaria, con el fin de mantenerse en un escenario conservador, ya que esto significa crear más de un millón de muestras. Inicialmente la base de datos con las variables nuevas, se distribuye de la siguiente manera:

$$\underset{\text{(Clase 0)}}{13\,106\,157} \quad + \quad \underset{\text{(Clase 1)}}{90\,394} \quad \rightarrow \quad \underset{\text{(Total)}}{13\,196\,551}$$

Tras el procedimiento de crear muestras sintéticas, la base de datos queda distribuida de la siguiente manera:

$$\underset{\text{(Clase 0)}}{13\,106\,157} \quad + \quad \underset{\text{(Clase 1)}}{1\,310\,615} \quad \rightarrow \quad \underset{\text{(Total)}}{14\,416\,772}$$

6.8.1. Regresión Logística

Al volver a ajustar el algoritmo con todas las variables, para obtener las métricas de evaluación e importancia de variables, se obtiene la siguiente matriz de confusión y curva ROC:

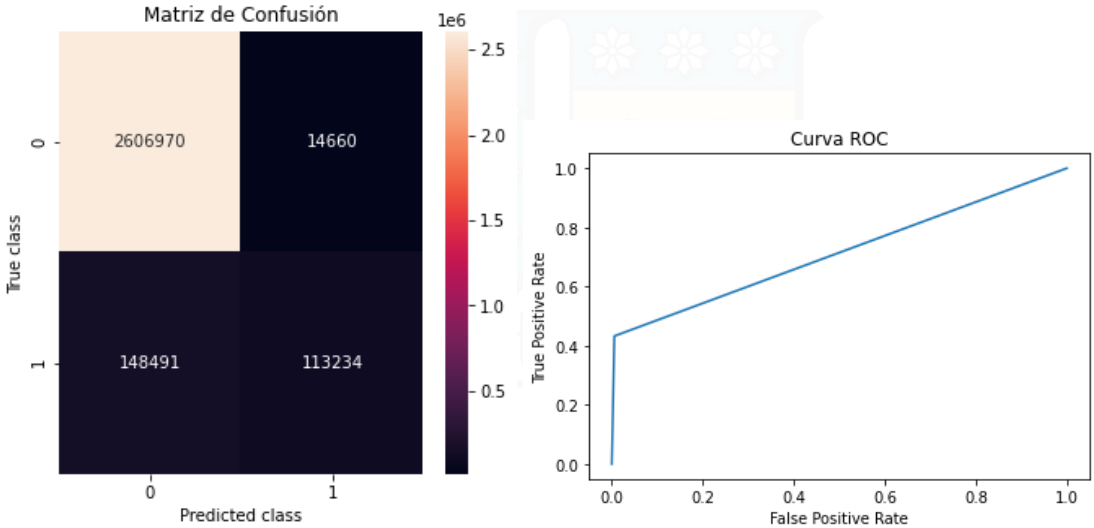


Figura 6.21: Gráficas Regresión logística con muestas sintéticas

Fuente: Elaboración propia

De la matriz se desprenden las siguientes métricas de evaluación:

Clase/ Métrica	Precisión	Recall
Clase 0	0.95	0.99
Clase 1	0.89	0.43

Tabla 6.16: Métricas Regresión logística con muestras sintéticas

Fuente: Elaboración propia

La precisión de ambas clases alcanza altos valores, mientras que la sensibilidad para la clase minoritaria es baja, esto indica que el modelo no detecta muy bien la clase 1, pero es capaz de detectar muy bien la clase 0, sin embargo la precisión indica que al realizar la clasificación, independiente de la clase, el modelo es confiable. La exactitud que alcanza el algortimo es de un 94 %, lo que refleja que, en terminos generales, el modelo no comete tantos errores en la clasificación.

Con respecto a la curva ROC el rendimiento es de un 0.713 que responde a un desempeño regular, similar a lo obtenido cuando se prueba con distintos tamaños de muestras sintéticas. El comportamiento de la influencia de las variables se puede observar en la siguiente gráfica.

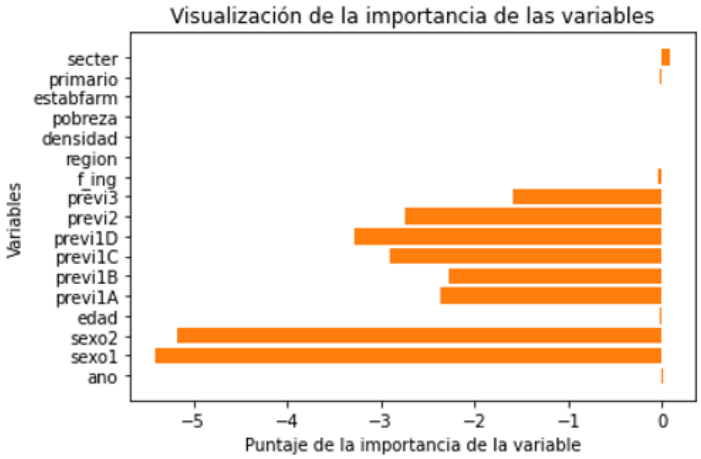


Figura 6.22: Importancia de las variables regresión logística con muestras sintéticas

Fuente: Elaboración propia

Los datos arrojados por el programa responden al valor de los coeficientes de la recta de regresión, por lo que se aplica la función exponencial, y se obtienen los ponderadores de la Tabla 6.21, donde se desprende que las variables que no influyen en el algoritmo son las con un valor cercano a cero como el año, edad, region, densidad, pobreza, estab_farma, primario y sec_ter, si bien estos valores van cambiando al avanzar en las iteraciones solo se interpreta en este etapa del proceso.

Si se observa, el odd ratio de las variables *sexo1* y *sexo2*, se puede notar que son valores que se escapan del resto de los datos, se cree que esto ocurre debido a la elección de la categoría base, ya que se utiliza el sexo "intederminado-desconocido" como categoría de referencia, donde solo representan un 0.01 % de los registros, aunque el valor represente el número de veces de riesgo en relación a no sufrir un evento adverso a medicamentos, también se compara con la categoría base. El valor obtenido para *sexo1* indica que existen 221 veces menos riesgo de no sufrir un EAM, que sufrirlo, al ser hombre, en comparación a una persona de sexo indeterminado. La interpretación es análoga con la variable *sexo2*, que indica que existen 176 veces menos riesgo de no sufrir daño por medicamentos, que sufrirlo, al ser mujer en comparación a una persona de sexo indeterminado, por lo que ser hombre es un factor protector a la hora del daño por medicamentos.

En cuanto a las seguros de salud, los valores no son tan alejados como con las variables que representan el sexo de los pacientes, sin embargo, también su magnitud puede responder a la elección de la categoría base, ya que entre las categorías FONASA - ISAPRES - Particular abarcan el 93.5 % de los datos. Como todos los coeficientes son negativos, indica que el que tenga mayor valor responde a tener menos riesgo en sufrir daño por medicamentos, que sufrirlo en relación a los que cuenta con previsiones de las fuerzas armadas y del orden u otra, por lo que pertenecer a los dos tramos de FONASA con más ingresos implica menos riesgo de sufrir daño por medicamentos, esto mismo ocurre para las personas que cuentan con ISAPRE como seguro

de salud. También se puede desprender que las personas que no cuentan con seguro de salud, es decir, se atienden de manera particular, son las que presentan mayor riesgo de tener daño por medicamentos, aunque se presente como un factor protector.

Finalmente, resulta interesante comprender que indica la variable f_ing, dado que por cada mes que transcurre en el año, existe 1.5 veces menos riesgo de sufrir un EAM, lo que indica que en los primeros meses del año existe mayor riesgo de sufrir daño ocasionado por medicamentos.

Variable	Coeficiente (β)	Exp(β)	1/ Exp(β)
año	0.00282	1.003	—
sexo1	-5.40019	0.005	221.45
sexo2	-5.17260	0.006	176.37
edad	-0.01823	0.982	1.018
previ1A	-2.36318	0.094	10.625
previ1B	-2.28275	0.102	9.804
previ1C	-2.89924	0.055	18.160
previ1D	-3.29085	0.037	26.866
previ2	-2.73686	0.065	15.438
previ3	-1.58232	0.205	4.866
f_ing	-0.04598	0.631	1.584
region	-0.00617	0.994	1.006
densidad	-0.00002	1.000	—
pobreza	-0.00299	0.997	1.003
estab_farma	-0.00283	0.997	1.003
primario	-0.01677	0.983	1.017
sec_ter	0.07918	1.082	—

Tabla 6.17: Variables Regresión logística con muestras sintéticas
Fuente: Elaboración Propia

6.8.2. Bosques Aleatorios

Del ajuste con un 10 % de los datos con la clase minoritaria se obtienen las siguientes métricas de evaluación:

Clase/ Métrica	Precisión	Recall
Clase 0	0.99	0.98
Clase 1	0.79	0.90

Tabla 6.18: Métricas Bosques aleatorios con muestras sintéticas

Fuente: Elaboración propia

Estas son extraídas de la siguiente matriz de confusión, adicionalmente se presenta la curva ROC obtenidaas al ejecutar el algoritmo:

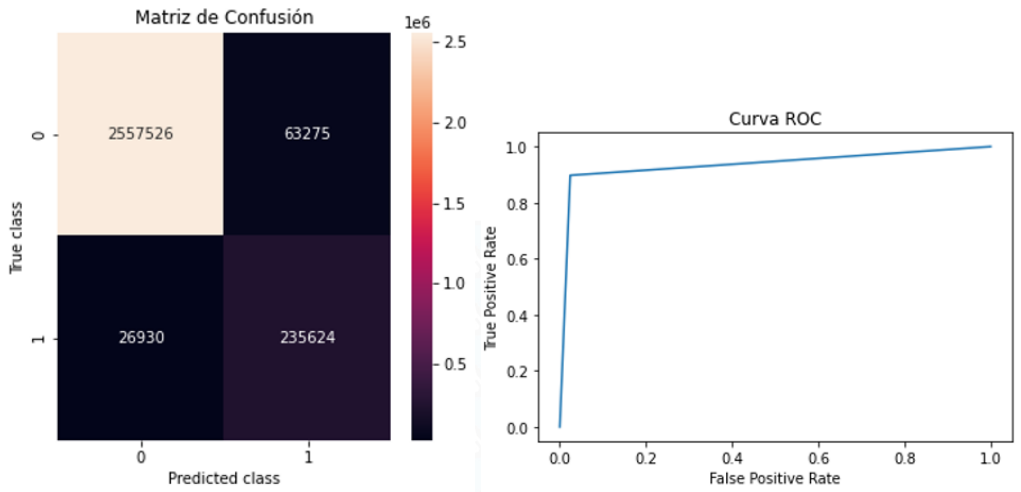


Figura 6.23: Gráficas Bosques Aleatorios con muestras sintéticas

Fuente: Elaboración propia

En cuanto a los valores obtenidos para la clase 0, se obtiene que el modelo clasifica muy bien, y es altamente confiable, lo mismo ocurre para la clase 1, pero en un menor grado de confiabilidad que la clase mayoritaria. El rendimiento del modelo alcanza un 93.66 %, siendo un test muy bueno, lo mismo que indica el OOB score con un valor de 0.97 para los datos que no fueron utilizadon durante el entrenamiento. El comportamiento de la influencia de las variables se puede observar en la siguiente gráfica:

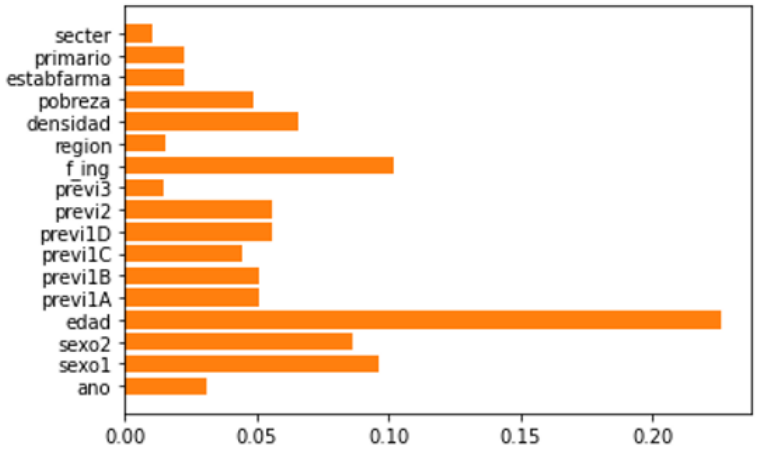


Figura 6.24: Importancia de las variables Bosques Aleatorios con muestras sintéticas

Fuente: Elaboración propia

Al ser un método no paramétrico, no se obtienen coeficientes de una ecuación, pero se puede determinar las variables que ayudan a clasificar un EAM o no, como se muestra en la Tabla 6.19. Para una mejor comprensión al lado derecho se ordena de manera descendente las variables de acuerdo a la puntuación obtenida:

Variable	Puntuación
año	0.03079
sexo1	0.09619
sexo2	0.08645
edad	0.22622
previ1A	0.05116
previ1B	0.05115
previ1C	0.04463
previ1D	0.05576
previ2	0.05547
previ3	0.01470
f_ing	0.10192
region	0.01553
densidad	0.06579
pobreza	0.04887
estab_farma	0.02272
primario	0.02226
sec_ter	0.01040

Tabla 6.19: Variables Bosques Aleatorios con muestras sintéticas

Fuente: Elaboración Propia

1) edad
2) f_ing
3) sexo1
4) sexo2
5) densidad
6) previ1D
7) previ1A
8) previ1B
9) pobreza
10) previ1C
11) año
12) estab_farma
13) primario
14) region
15) previ2
16) sec_ter
17) previ3

De acuerdo a lo obtenido, se puede desprender que variables que responden a factores biológicos como la edad y el sexo del paciente están posicionadas dentro de las más relevantes a la hora de predecir un evento adverso a medicamentos, sin embargo no se puede determinar si el impacto es positivo o negativo, es decir, si es un factor protector o un factor de riesgo. Con respecto a la previsión de salud, se puede notar que la categoría relevante es FONASA con todos sus tramos, ya que las otras dos variables asociadas a las categorías de los seguros de salud se encuentran en los últimos lugares de importancia.

El factor temporal como el mes de ingreso, se encuentra en segundo lugar de las variables relevantes, sin embargo el año de ingreso del paciente al establecimiento de salud, no tienen relevancia a la hora de predecir si está relacionado con el daño por medicamentos, lo mismo ocurre con los niveles de atención, ya que *primario* se encuentran en la doceava posición y los otros niveles de atención se encuentran en el diesiesavo lugar.

6.9. Análisis del modelado

Al momento de crear las muestras sintéticas, se generan de manera independiente para cada algoritmo, dado que lo que se busca es explicar el comportamiento de los datos por el desbalance, si bien se puede realizar un proceso de eliminación de variables, esto responde a distintos conjuntos de datos para cada caso, por lo que no es correcto realizar la eliminación de variables a partir de distintos inicios.

Se decide realizar un último ajuste: cambiar las categorías base para la variable sexo y previsión, ya que se sospecha que las magnitudes presentadas responden a la comparación que se hace con la categoría de referencia.

6.10. Modelado: Cuarta Parte

En relación a la variable sexo, se decide eliminar la variable que representa a "mujer" quedando en el modelo solo *sexo1*, por lo que se pierden los registros donde ambas categorías tomaban el valor cero, que era la antigua categoría base. Adicionalmente, se tiene la siguiente distribución de registros para las categorías de la variable previsión:

- *previ1A*: 21.69 %
- *previ1B*: 28.71 %
- *previ1C*: 8.54 %
- *previ1D*: 13.46 %
- *previ2* : 19.02 %
- *previ3* : 02.21 %

El 6.37 % restante corresponde a los registros que adquieren valor cero para todas estas variables, respondiendo a la categoría base de las fuerzas armadas y del orden y otras. Se decide definir *previ3* como la nueva categoría base, aunque por los porcentajes presentados es tentador dejar en el estudio solo la comparación entre ISAPRE y FONASA, pero significa arreglar muy convenientemente los datos.

Es necesario mencionar que las técnicas utilizadas para ir mejorando los resultados son acumulativas, por lo que se utiliza el balance en el proceso de entrenamiento, las variables nuevas, muestras sintéticas en conjunto, y ahora se agrega el cambio de categoría base.

6.10.1. Regresión Logística

Se obtienen las siguientes métricas de evaluación:

Clase/ Métrica	Precisión	Recall
Clase 0	0.95	0.98
Clase 1	0.73	0.52

Tabla 6.20: Métricas Regresión logística nuevas categorías
Fuente: Elaboración propia

El rendimiento de la curva ROC alcanza los 0.7513, por lo que se cataloga como un test bueno, aunque en el límite inferior. Si se comparan las métricas derivadas de la matriz de confusión con el caso anterior, se puede notar que el modelo es capaz de detectar un poco mejor la clase 1, sin embargo reduce la confiabilidad cuando lo hace, la matriz de confusión y la gráfica de la curva ROC se encuentran en la Subsubsección A.3.2.5. A continuación se presentan la puntuación obtenida y el respectivo odd ratio:

Variable	Coeficiente (β)	Exp(β)	1/ Exp(β)
año	0.00092	1.001	—
sexo1	-0.96063	0.383	2.613
edad	-0.01380	0.986	1.014
previ1A	-3.88944	0.020	48.884
previ1B	-3.68644	0.025	39.903
previ1C	-3.83455	0.022	46.272
previ1D	-4.34160	0.013	76.830
previ2	-3.85643	0.021	47.296
f_ing	-0.03129	0.969	1.032
region	-0.01345	0.987	1.014
densidad	-0.00000	1.000	—
pobreza	0.01295	1.013	—
estab_farma	-0.00106	0.999	1.001
primario	-0.01519	0.985	1.015
sec_ter	0.04065	1.041	—

Tabla 6.21: Variables Regresión logística con muestras sintéticas y nuevas categorías
Fuente: Elaboración Propia

Se puede evidenciar que la magnitud del odd ratio aterriza para la variable *sexo1*, evidenciando que existe 2.61 veces menos riesgo de sufrir un evento adverso a medicamentos al ser hombre, que sufrirlo, en comparación a una mujer, por lo tanto ser mujer es un factor de riesgo. A pesar de obtener mejor respuesta con esta variable, el cambio de categoría base para la previsión no entrega mejores resultados, se siguen evidenciando grandes magnitudes, incluso mayores que las presentadas anteriormente. Al ser negativo estos valores, indican que son un factor protector, el tramo más pudiente de FONASA es el que presenta menos riesgo de sufrir EAM, y el tramo B es el que podría tener más posibilidades de sufrir un EAM, dentro de su

bajo riesgo. También, se puede observar que el resto de las variables no tienen mayor impacto en la respuesta de la clasificación, ya que el odds ratio (o la inversa de este) son valores cercanos a cero.

La puntuación de los coeficientes se puede apreciar gráficamente en la siguiente imagen:

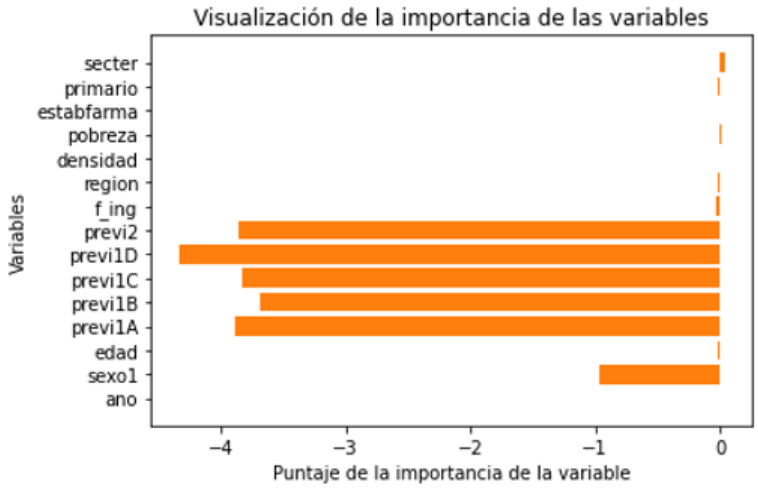


Figura 6.25: Importancia de las variables con categorías nuevas en regresión logística

Fuente: Elaboración propia

6.10.2. Bosques Aleatorios

También se ejecuta con las variables y categorías nuevas, y las muestras sintéticas, donde se obtienen las siguientes métricas de evaluación:

Clase/ Métrica	Precisión	Recall
Clase 0	0.99	0.97
Clase 1	0.73	0.89

Tabla 6.22: Métricas Bosques aleatorios nuevas categorías

Fuente: Elaboración propia

Se obtienen excelentes métricas para la clase mayoritaria, sin embargo los resultados son buenos para la clase 1 también, aunque en un leve mejor grado que el caso sin cambio de categoría base, esto se puede deber a que las muestras sintéticas creadas para los conjuntos responden de manera distinta, y en el caso anterior, se adaptaron mejor al conjunto de datos. Por otro lado, la curva ROC alcanza un rendimiento de 0.9266, siendo un test muy bueno, igual que en el caso anterior.

La importancia de las variables se exponen en la siguiente tabla:

Variable	Puntuación
año	0.03408
sexo1	0.03200
edad	0.25887
previ1A	0.06933
previ1B	0.07415
previ1C	0.05322
previ1D	0.07626
previ2	0.07627
f_ing	0.11577
region	0.01701
densidad	0.07543
pobreza	0.05568
estab_farma	0.02526
primario	0.02531
sec_ter	0.01137

Tabla 6.23: Variables Bosques Aleatorios con nuevas categorías

Fuente: Elaboración Propia

1) edad
2) f_ing
3) previ2
4) previ1D
5) densidad
6) previ1B
7) previ1A
8) pobreza
9) previ1C
10) año
11) sexo1
12) primario
13) estab_farma
14) region
15) sec_ter

La variable *edad* y *f_ing* siguen liderando la importancia para la clasificación, se puede apreciar que la variable *sexo1* pasa del tercer lugar al onceavo lugar tras el cambio de categoría base bajando su relevancia, otro cambio importante es la previsión de ISAPRE, que del lugar número 15 sube al número 3, siendo relevante a la hora de determinar si existe daño por medicamentos. En relación a FONASA con sus tramos de beneficio, el tramo D sigue siendo el más influyente y el tramo C el menos influyente, los tramos A y B enrocan de posición.

Continuando el análisis, las variables *densidad* y *pobreza* no sufren mayores variaciones en sus posiciones de influencia, y los atributos *año, estab_farma, primario,region* y *sec_* se mantienen en los últimos lugares como en el caso anterior.

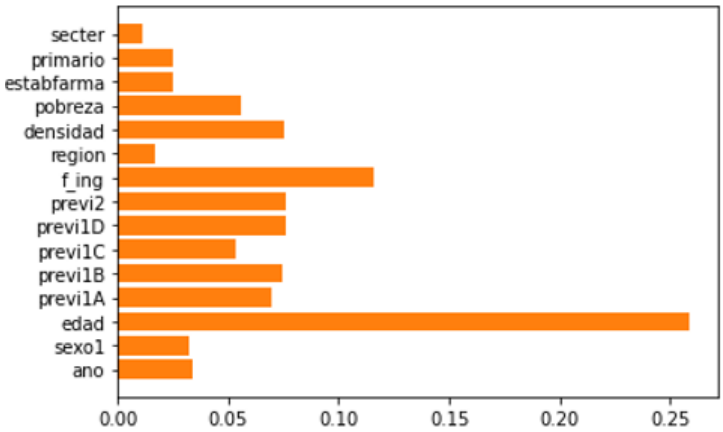


Figura 6.26: Importancia de las variables con categorías nuevas en regresión logística

Fuente: Elaboración propia

Como no se puede determinar el impacto o comportamiento de la variable, se realiza un análisis particular a las dos variables más influyentes y se determina la frecuencia de los casos EAM respecto a ellas, en el primer gráfico se presenta el comportamiento de los casos EAM respecto a la edad, y en segundo lugar los casos EAM respecto al mes de ingreso.

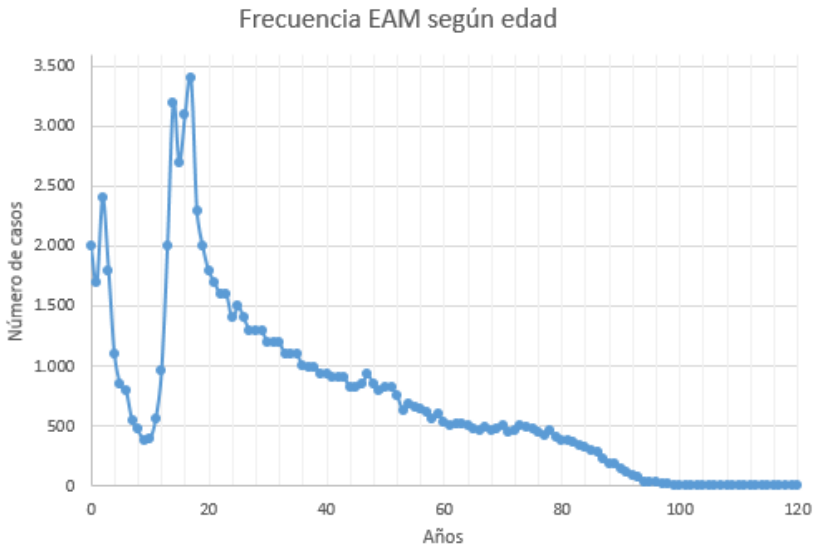


Figura 6.27: Frecuencia EAM según edad del paciente

Fuente: Elaboración propia

Según el gráfico expuesto, la tasa de incidencia de la clase minoritaria durante los tres primeros años de vida es bastante superior al resto de los años de la infancia, cuando se analiza la etapa de la adolescencia y adulto joven se logra la mayor frecuencia de casos atribuibles al daño por medicamentos, para luego comenzar a disminuir constantemente.

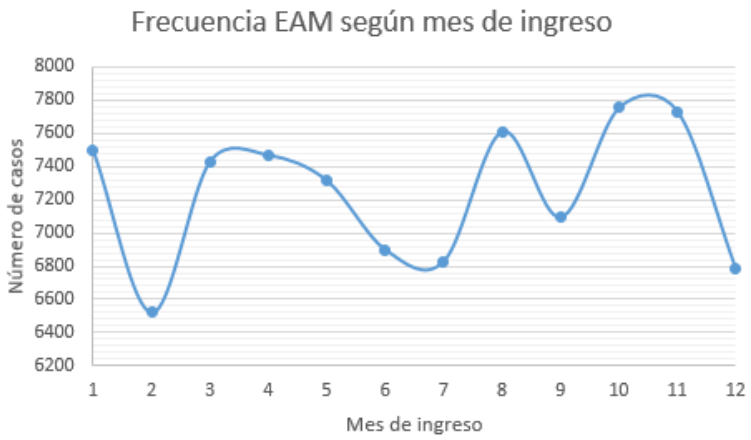


Figura 6.28: Frecuencia EAM según mes de ingreso

Fuente: Elaboración propia

Con respecto a la frecuencia de casos según los meses del año, entre los años 2011 y 2018 se presentaron más casos entre los meses de agosto, octubre, noviembre, y considerablemente menos casos en el mes de febrero.

6.11. Evaluación y Despliegue

Se puede seguir intentando alternativas para mejorar los resultados, como dejar solo las previsiones FONASA e ISAPRE, pese a ello, se decide culminar el estudio, ya que en cada nueva iteración van a surgir nuevas ideas y se convierte en un proceso sin fin. Sin embargo. lo que se determina con certeza es el problema del conjunto de datos: el desbalance entre las clases.

La generación de un modelo no siempre es la fase final de un proyecto, si no como estos van a ser utilizados en la posterioridad, por lo que es importante desprender las lecciones aprendidas de este proceso, como se menciona en el párrafo anterior, el principal problema de este estudio radica en la estrategia para corregir el balance de clases, si bien se aplicaron dos técnicas diferentes (que en casos anteriores han mostrado buenos resultados), sólo la técnica de sobremuestreo de minorias sintéticas es la que logra mejorar el desempeño con este conjunto de datos, y el algoritmo de bosques aleatorios es el que obtiene mejores resultados clasificando, sin embargo, resulta dificil explicar el comportamiento detrás de este tipo de ajuste ya que se comporta como un modelo de caja negra, siendo la mejor alternativa para clasificar pero no para entender como afectan las variables, a menos que se realice un análisis particular a las más influyentes.

7 | CONCLUSIONES

El gran volumen de información producido tanto por instituciones públicas como privadas, hace necesario disponer de instrumentos tecnológicos que faciliten el procesamiento y extracción de información útil para apoyar la toma de decisiones. El objetivo de este estudio apuntó a determinar los factores de riesgos asociados al daño por medicamentos, de acuerdo a la caracterización de la comuna de residencia del paciente aplicando algoritmos de minería de datos, por lo que fue necesario indagar diversas fuentes de información para complementar con información de los determinantes de salud la base de datos de egresos hospitalarios.

De los resultados obtenidos en la medición de la calidad de los datos, se puede desprender que las bases de egresos hospitalarios responden correctamente a los factores inherentes a los datos, ya que la gran mayoría de los registros responden a datos exactos y completos. Cuando se evalúa la consistencia de todo el periodo de estudio existe un 2.7 % de inconsistencia en promedio por año, pero esto esconde que en la realidad los cuatro primeros años son los que presentan un mayor grado de contradicciones en los datos, si bien es un número bajo, responde a que se asigna una modalidad de atención a los pacientes que no son FONASA, cuando es una característica propia de los beneficiarios de dicha previsión. Con respecto a las características de los datos que dependen del sistema, se puede decir que respetan la confidencialidad del paciente identificándolos a través de un código, también existe una buena descripción de las variables y los valores que pueden tomar a través de la publicación de un diccionario de datos, y cada archivo es publicado oportunamente cada año.

Referente a las bases de datos extraídas del Sistema Nacional de Información Municipal se pudo medir los errores aparentes, aunque se obtienen buenos indicadores, existen años completos sin información o comunas que para todos los años en estudio no presentaban información. Con respecto a las otras fuentes de información que complementan la base de egresos, como el listado de establecimientos no se puede medir las características inherentes a los datos, ya que solo se posee la definición de la variable y no los valores que puede tomar para cada campo. En relación a la distribución de las farmacias no es una base de datos, se toma la información del archivo y se estructura como tal, lo mismo ocurre para los códigos que identifican la región y comunas que son extraídos del Decreto que sistematiza la codificación para las divisiones político-administrativa del país.

Más adelante, en la fase de modelado, se determina que la regresión logística y el algoritmo de bosques aleatorios son los adecuados para clasificar y obtener dos puntos de vistas distintos, por un lado se tiene un método paramétrico donde se privilegia la interpretabilidad, y por el otro lado, un método no paramétrico que privilegia la exactitud por sobre la interpretabilidad. Si bien, los resultados de los ajustes no son contundentes y no permiten sacar buenas conclusiones de los factores de riesgo sin ningún tipo de mejoramiento de los datos, permite intuir el problema que origina el mal rendimiento de ambos modelos: el desbalance de las clases del conjunto de datos.

Se prueban distintas alternativas para mejorar los resultados, siendo la creación de muestras sintéticas donde se rescatan buenos resultados, en el caso de la regresión logística se obtiene un buen ajuste con las variables *año, sexo1, edad, previ1A, previ1B, previ1C, previ1D, previ2, f_ing, region, densidad, pobreza, estab_farma, primario* y *sec_ter*, aunque al evaluar la influencia de cada variable por separado se puede desprender que las variables que se agregan a la base de egresos hospitalarios en ningún caso resultan relevantes, y solo factores presentes en las bases de egresos hospitalarios como la edad, sexo y previsión son las que contínuamente aparecen como factores relevantes.

Del último ajuste se concluye que existen 2.61 veces menos riesgo de sufrir daño por medicamentos al ser hombre, que no sufrirlo, en relación a una mujer. Esto se corresponde con el estudio realizado por Collao *et al*, donde indican que los casos EAM ocurren con mayor frecuencia en mujeres al utilizar los códigos propuestos por Stausberg, pero ahora se puede concluir que la relación es de 2.6 veces a 1. Adicionalmente, al estudiar los diagnósticos más recurrentes asociados al daño por medicamentos en las mujeres se encuentra que el *envenenamiento por otras drogas y sustancias biológicas, y por las no especificadas*, *envenenamiento por benzodiazepinas y enterocolitis debida a Clostridium difficile (infección causada por diarrea después del uso de antibióticos)* y *síndrome de dependencia de múltiples drogas y de otras sustancias psicoactivas.*

Por otro lado, del algoritmo de bosques aleatorios se concluye que es una excelente alternativa para la clasificación, pero no entrega información clara de como influyen las variables, solo se puede determinar cual tiene un mayor impacto en la decisión, no obstante, al realizar un análisis particular de las variables como la edad y el mes de ingreso que son las que se presentan en las primera posiciones de relevancia, se puede determinar que el peak de casos de EAM ocurren durante la adolescencia, y luego con mayor frecuencia durante los cuatro primeros años de vida de una persona, y mientras más avanzada la edad, menores son los casos registrados. Con respecto al mes de ingreso se desprende que el valle se encuentre en el mes de febrero, y las mayores frecuencia ocurren durante los meses de octubre y noviembre. Basándose en estos resultados, se analiza los grupos de códigos que se dan con mayor frecuencia en los grupos que el algoritmo detecta como relevantes, siendo el código que describe el *envenenamiento por otras drogas y sustancias biológicas, y por las no especificadas* quien lidera el ranking, también se suma a la lista el *síndrome de dependencia de múltiples drogas y de otras sustancias psicoactivas, envenenamiento por benzodiazepinas* y *envenenamiento por derivados del para-aminofenol* en la adolescencia y, en el caso de los infantes los diagnósticos más

comunes durante los meses de octubre y noviembre se encuentra el *envenenamiento por otras drogas y sustancias biológicas, y por las no especificadas, envenenamiento por benzodiazepinas, púrpura alérgica* y *envenenamiento por drogas antialérgicas y antieméticas.*

Teniendo en cuenta los diagnósticos que generan los mayores casos de EAM en los grupos de riesgo, en el caso de los adolescentes es importante limitar el acceso a los medicamentos que producen estos daños como la benzodiazepina y el paracetamol (principal derivado del para-aminofenol), además de una oportuna y correcta intervención escolar relacionada al consumo de las sustancias psicoactivas. Por otro lado, comunicar a los profesionales que se debe tener una preocupación especial a la hora de otorgar una prescripción farmacológica en los infantes, ya que aunque se conoce que la púrpura alérgica se desarrolla después de una infección respiratoria superior, como un resfriado, existen otros desencadenantes como ciertos medicamentos, alimentos, picaduras de insectos, entre otros [8].

Si se comparan los resultados de las variables obtenidas por ambos algoritmos, resulta curioso por qué la edad es tan influyente para el algoritmo de bosques aleatorios, y no para la regresión logística. Esto responde al supuesto detrás de la regresión logística, ya que parte de la idea que los datos son linealmente separables, y si se observa el comportamiento obtenido de la frecuencia de ocurrencia según la edad y se revisa el estudio de Collao *et al*, se determinan curvas similares del comportamiento según el ciclo vital, donde las frecuencias de incidencias más altas de daño por uso de medicamentos están en los lactantes, preescolares, adolescentes y ancianos, por lo que el algoritmo de bosques aleatorios logra desprender este comportamiento y decir que la edad si es relevante para la clasificación y existen consecuencias diferentes debido al daño por medicamentos de acuerdo a la etapa de la vida que se encuentre una persona, objeto que no logra la regresión logística. En el caso del mes de ingreso, la regresión logística logra captar levemente el comportamiento, sin embargo al asemejarlo a una recta coincide que en los últimos meses del año se presentan mayores casos de EAM.

Finalmente, resulta interesante investigar como utilizar ambos modelos en conjunto para obtener la potencialidad de la interpretabilidad y la exactitud. Si bien se puede desprender información con registros que intentan seguir el comportamiento de los datos reales, lo recomendable es reducir la clase mayoritaria de las bases de egresos hospitalarios, pero para esto es necesario realizar un estudio de los grupos de diagnósticos que componen los egresos, con el objetivo de extraer los registros que no tienen relación con medicamentos, como por ejemplo egresos hospitalarios como consecuencia de accidentes de tránsito ocasionado por terceros, quemaduras, caídas, ataques de animales, agresiones por terceros, etc. Esto para reducir el desbalance bajo un criterio estudiado y fundado, y no producto de la aleatoriedad. Asimismo, vale agregar que con los resultados expuestos se conoce que para ambos algoritmos con un 10 % de la clase minoritaria ya se obtendrían buenos resultados y no es necesario forzar a reducir más en mayor medida la clase mayoritaria.

[8] Fuente: Kids Health, recuperado de https://kidshealth.org/es/parents/hsp-esp.html

Una propuesta complementaria a la reducción de la clase mayoritaria, corresponde a sacrificar una porción de los registros de egresos hospitalarios, dejando fuera los grupos minoritarios de las variables categóricas y centrándose en los que abarcan una mayor proporción de los datos, como se hizo en la última parte de este estudio. Particularmente, enfocarse en hombres y mujeres cuando se refiere al sexo del paciente, y en la variable previsión, enfocarse solo en los dos grandes grupos: FONASA con sus respectivos tramos e ISAPRES, incluso también se puede considerar la opción de reunir los tramos C y D para lograr un balance entre las categorías, o agrupar la previsión de atención particular con la de las Fuerzas Armadas y del Orden, en lugar de eliminar estos registros.

8 | DISCUSIÓN

El Departamento de Estadística e Información de Salud lidera el proceso de gobernanza de datos de la salud, sin embargo, resulta curioso si la información de los egresos hospitalarios solo se recoge para generar estadísticas que se reportan año a año o para realizar estudios transversales, ya que existe un constante cambio en los esquemas de registro. Si bien existen atributos que responden a factores externos como la codificación de los establecimientos que se ven afectados por las modificaciones en el funcionamiento, o la codificación de las comunas y regiones que responden a la división político-administrativa del país, existen atributos que se pueden mantener invariantes en el tiempo como la codificación del sexo, previsión, tramo de fonasa, servicio de salud, por mencionar algunos. Si bien es trabajo del investigador familiarizarse con las fuentes de información secundaria, este constante cambio de codificación genera un gran obstáculo de entrada para el procesamiento de los datos, ya que es necesario invertir una gran cantidad de tiempo para estandarizar todos los atributos para recién comenzar a trabajar con la información. Individualmente, cada base de egresos hospitalarios responde de manera adecuada según su esquema de registro, sin embargo, si se consideran los ocho años de estudio como un conjunto, la comunicación entre estos falla.

En el mismo orden de ideas, es indiscutible la necesidad de capacitar en el correcto registro de datos y transmitir la importancia de esta labor constantemente a quienes corresponda para que se realicen supervisión. Además, se debe comunicar a los digitadores que la información que ingresan al sistema es relevante para realizar estudios de prevalencia, evoluciones y tendencias, permitiendo a los investigadores determinar factores de riesgo a nivel nacional que pueden facilitar la gestión de la sanidad. Si bien existe un instructivo para el registro de los egresos hospitalarios no existe evidencia que se fiscalice y se cumplan con los estándares que define el MINSAL, aunque los resultados de la calidad son buenos a nivel nacional, sería interesante realizar un estudio de la manera de registrar de cada establecimiento de salud, ya que al obtener los promedios, se enconden los recintos que cometen mayores errores con los que realizan la labor de manera ejemplar.

Por otra parte, es importante mencionar que los factores biológicos que se publican son acotados al sexo y edad del paciente, y para realizar un estudio más profundo en el área de los factores de riesgo se hace forzoso conocer la historia clínica del paciente y el estilo de vida, por ejemplo, conocer si el individuo está en un tratamiento farmacológico previo, si posee enfermedades crónicas, si el paciente fuma o posee algún grado

de sobrepeso, siempre manteniendo la confidencialidad bajo la identificación de un código descriptor.

En otro orden de ideas, es necesario ampliar la visión en el análisis de conjuntos de datos, particularmente en salud, ya que muchos estudios dan por sentado las interpretaciones basándose en que los coeficientes son significativos al 5 %. Esto lo menciona Leo Breiman -estadístico, creador del algoritmo de bosques aleatorios- en un artículo donde compara las dos culturas del modelado estadístico: la cultura del modelado de datos y del modelado algorítmico (Breiman, 2001), esta idea viene a cuestionar las decisiones que se han tomado fundadas en modelos lineales aún cuando en la realidad es difícil que esto ocurra. Muchas veces las hipótesis plantean que la relación de una variable de respuesta es directa o indirectamente proporcional a una predictora basándose en la intuición, o lo que dicta la lógica, pero es difícil encontrar casos dond se plantee una hipótesis que no responda a un modelo lineal, como que la variable de respuesta cambie de acuerdo a distintos tramos de la predictora. Si en este estudio solo se utiliza la regresión logística para determinar los factores relevantes, no sería incluida la edad, el mes de ingreso ni la densidad, ya que no son relevantes de manera lineal, sin embargo, se puede notar que son un factor relevante a considerar.

Bibliografía

Aguilera, X. et al. (2019). Estructura y funcionamiento del sistema de salud chileno. *Universidad del Desarrollo, Facultad de Medicina*. Recuperado el 18 de junio de 2020 de https://medicina.udd.cl/centro-epidemiologia-politicas-salud/files/2019/12/ESTRUCTURA-Y-FUNCIONAMIENTO-DE-SALUD-2019.pdf. 4.1.4.1

Allison, Paul (2013). ¿cuál es el mejor r-cuadrado para la regresión logística? Recuperado el 10 de septiembre de https://statisticalhorizons.com/r2logistic. 4.3.4

BCN, Biblioteca del Congreso Nacional de Chile (2010). Decreto 1671 exento: Aprueba normal general técnica sobre uso de formulario "inform estadístico de egreso hospitalario" para la producción de información estadística sobre causas de egreso hospitalario y variables asociadas. Recuperado el 30 de junio de 2020 de https://www.leychile.cl/Navegar?idNorma=1019779#0. 6.1.6

BCN, Biblioteca del Congreso Nacional de Chile (2018a). Decreto 1115: Establece abreviaturas para identificar las regiones del país y sistematiza codificación única para regiones, provinciar y comunas del país dejando sin efecto el decreto 1.39, del año 2000, del ministerio del interior y sus modificaciones. Recuperado el 28 de junio de 2020 de https://www.leychile.cl/Navegar?idNorma=1123248. 6.1.4

BCN, Biblioteca del Congreso Nacional de Chile (2018b). Resolución 1616 exenta: Determina la estructura orgánica del instituto de salud pública de chile. Recuperado el 20 de junio de 2020 de https://www.leychile.cl/Navegar?idNorma=1120468&buscar=resoluci%C3%B3n+exenta+1616. 4.1.4.1

BCN, Biblioteca del Congreso Nacional de Chile (2019). Decreto 3: Aprueba reglamento del sistema nacional de control de los productos farmacéuticos de uso humano. Recuperado el 20 de junio de 2020 de https://www.leychile.cl/Navegar?idNorma=1026879. 4.2.2

BCN, Biblioteca del Congreso Nacional de Chile (2020). Decreto 725: Código sanitario. Recuperado el 22 de junio de 2020 de https://www.leychile.cl/Navegar?idNorma=5595. 4.2.1

Breiman, Leo (2001). Statistical modeling: The two cultures. *Statistical Science*, 16, 199–215. 8

Brownlee, Jason (2020). Smote for imbalanced classification with python. Recuperado el 08 de septiembre de 2020 de https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/. 6.8

Castro, S. et al. (2020). Estudio de mercado sobre medicamentos. Recuperado el 22 de junio de 2020 de https://www.fne.gob.cl/wp-content/uploads/2020/01/Informe-Final.pdf. 4.2.1, 4.2.2.1

CENABAST, Central de Abastecimiento del Sistema Nacional de Servicios de Salud (s.f). Definiciones estratégicas. Recuperado el 21 de junio de 2020 de https://www.cenabast.cl/institucion/definiciones-estrategicas/. 4.1.4.1

Chutka, D. et al. (2004). Inappropriate medications for elderly patients. *Mayo Clinic Proceedings*, 79, 122–129. Recuperado el 04 de julio de 2020 de https://www.sciencedirect.com/science/article/abs/pii/S0025619611632655. 1

Cisternas, M. (2020). Gasto público en salud: la falencia que pone a chile en riesgo frente a la pandemia. *DiarioUchile*. Recuperado el 22 de Junio de 2020 de https://radio.uchile.cl/2020/04/19/gasto-publico-en-salud-la-falencia-que-pone-a-chile-en-riesgo-frente-a-la-pandemia/. 4.1.2

Collao, Juan; Favereau, Rafael; Miranda, René; y Aceitón, Carolina (2019). Daño asociado al uso de medicamentos en hospitales chilenos: análisis de prevalencia 2010-2017. *Revista médica de Chile*, 147, 416 – 425. 1, 2, 6.1.5

Dolores, M. y Rodríguez, C. (2000). La regresión logística: una herramienta versátil. *Nefrología*, 6, 477–565. Recuperado de https://www.revistanefrologia.com/es-la-regresion-logistica-una-herramienta-versatil-articulo-X0211699500035664. 4.3.2.1

Durán, G. y Narbona, K. (2009). Caracterización del sistema de salud chileno: Enfoque laboral, sindical e institucional. *Fundación Sol*. Recuperado el 20 de junio de 2020 de http://www.fundacionsol.cl/wp-content/uploads/2010/09/Cuaderno-11-Salud-y-enfoque-laboral.pdf. 4.1.1, 4.1.4.1

FONASA, Fondo Nacional de Salud (s.f.). Descripción. Recuperado el 20 de junio de 2020 de https://www.fonasa.cl/sites/fonasa/conoce-fonasa. 4.1.4.1

Galaz, Óscar (2018). Sistema público notificó el 56 % de las reacciones adversas a medicamentos. Recuperado el 15 de agosto de 2020 de http://www.ipsuss.cl/ipsuss/analisis-y-estudios/medicamentos/sistema-publico-notifico-el-56-de-las-reacciones-adversas-a/2018-01-22/173409.html. 1

Galli, Amada; Pagés, Marisa; y Swieszkowski, Sandra (2017). Factores determinantes de la salud. *Sociedad Argentina de Cardiología*, (pp. 1–5). Recuperado el 26 de julio de 2020 de https://www.sac.org.ar/wp-content/uploads/2018/04/factores-determinantes-de-la-salud.pdf. 2

Gironés, Jordi; Casas, Jordi; Minguillón, Julià; y Caihuelas, Ramon (2017). *Minería de datos modelos y algoritmos*. Editorial Universitat Oberta de Catalunya (UOC). 4.3, 4.3.2.2, 6.4.1

Goodman, Roberto; Gray, Andy; Hoffman, Jerome; Lexchin, J.; et al. (2011). Comprender la promoción farmacéutica y responder a ella. 4.2.3

Gujarati, Damodar (2004). *Econometría*. McGraw-Hill, cuarta edición edición. 4.3.2.1, 4, 6.4.1

IMS, Institute for Healthcare Informatics (2015). Global medicines use in 2020. Recuperado el 04 de julio de 2020 de https://www.iqvia.com/-/media/iqvia/pdfs/institute-reports/global-medicines-use-in-2020. 1

INE, Instituto Nacional de Estadística (s.f). Esperanza de vida. Recuperado el 22 de junio de 2020 de https://www.ine.cl/ine-ciudadano/definiciones-estadisticas/poblacion/esperanza-de-vida#:~:text=En%20Chile%20la%20Esperanza%20de,77%2C3%20para%20los%20hombres. 4.1.2

ISO, International Organization for Standarization (2019). Iso 25012. Recuperado el 25 de junio de 2020 de https://iso25000.com/index.php/normas-iso-25000/iso-25012?limit=5&start=5. 4.4.1

ISP, Instituto de Salud Pública (s.f.). Farmacovigilancia. Recuperado el 22 de junio de 2020 de http://www.ispch.cl/anamed_/farmacovigilancia_1. 4.1.4.1

Jiménez, Simi; Barriga, Omar; y Salazar, Alide (2018). Inequidad en el acceso a salud en chile: estudio multifactorial basado en la encuesta casen del año 2013. *Revista Chilena de Salud Pública*, 22(1), 31–40. 2

Louviere, Jordan; Hensher, David; y Swait, Joffre (2000). *Stated Choice Methods: Analysis and Applications*. Cambridge University. 6.4.1

Marantao, M. y González, E. (2015). Fuentes de información. *Universidad Autónoma del Estado de Hidalgo*. Recuperado el 20 de junio de 2020 de https://repository.uaeh.edu.mx/bitstream/bitstream/handle/123456789/16700/LECT132.pdf. 4.4

Marcano, Yelitza Josefina y Talavera, Rosalba (2007). Minería de Datos como soporte a la toma de decisiones empresariales. *Opción*, 23, 104 – 118. 4.3

Margozzini, P. y Passi, A. (2018). Seminario: Qué nos dice la encuesta nacional de salud, ens 2016 – 2017. Recuperado el 04 de julio de 2020 de https://www.colegiofarmaceutico.cl/images/2019/Archivos_2019/ENS_2017_Medicamentos_en%20Chile_11Oct2018.pdf. 1

Marovac, J. (2001). Investigación y desarrollo de nuevos medicamentos: de la molécula al fármaco. *Revista médica de Chile*, 129, 99 – 106. 1

Marín, A. (2017). Dispensación de medicamentos en las grandes farmacias de chile: Análisis ético sobre la profesión del químico farmacéutico. *Pontificia Universidad Católica de Valparaíso*, (pp. 345). Recuperado el 21 de junio de 2020 de https://scielo.conicyt.cl/pdf/abioeth/v23n2/1726-569X-abioeth-23-02-00341.pdf. 4.2.3

Minitab (2019). Interpretar todos los estadísticos para regresión logística nominal. Recuperado el 15 de agosto de 2020 de https://support.minitab.com/es-mx/minitab/18/help-and-how-to/modeling-statistics/regression/how-to/nominal-logistic-regression/interpret-the-results/all-statistics/. 4.3.4

MINSAL, Ministerio de Salud (2015). Eligevivirsano. Recuperado el 25 de junio de 2020 de https://www.minsal.cl/promocion_participacion_evs/. 4.1.2

MINSAL, Ministerio de Salud (s.fa). 31 medidas. Recuperado el 21 de junio de 2020 de https://www.minsal.cl/politica-nacional-de-medicamentos/30-medidas/. 4.1.4.1

MINSAL, Ministerio de Salud (s.f.b). Encuesta nacional de salud 2016-2017. Recuperado el 22 de junio de 2020 de http://www.encuestas.uc.cl/ens/presentacion.html. 4.2.1

MINSAL, Ministerio de Salud (s.f.c). Misión y visión. Recuperado el 22 de Junio de 2020 de https://www.minsal.cl/mision-y-vision/. 4.1.4.1

MINSAL, Ministerio de Salud de Chile (sf.d). Diseño e implementación de una metodología de implementación, seguimiento y acompañamiento de la reforma de la salud de chile (resumen ejecutivo). Recuperado el 10 de julio de 2020 de https://url2.cl/2W6rT. 4.1.4.1

Morales, Marte; Ruiz, Inés; Morgado, Cecilia; y González, Ximena (2002). Farmacovigilancia en Chile y el mundo. *Revista chilena de infectología*, 19, S42 – S45. 2

Na8 (2019). Clasificación con datos desbalanceados. Recuperado el 05 de agosto de https://www.aprendemachinelearning.com/clasificacion-con-datos-desbalanceados/. 4.3.5, 6.4

Navlani, Avinash (2018). Comprender los clasificadores de bosques aleatorios en python. Recuperado el 20 de agosto de https://www.datacamp.com/community/tutorials/random-forests-classifier-python. 4.3.2.2

OCDE, Organización para la Cooperación y el Desarrollo Económicos (2019). Estudios de la ocde sobre salud pública: Chile. Recuperado el 23 de Junio de 2020 de https://www.oecd.org/health/health-systems/Revisi%C3%B3n-OCDE-de-Salud-P%C3%BAblica-Chile-Evaluaci%C3%B3n-y-recomendaciones.pdf. 4.1.2, 4.1.4.2

OCHISAP, Observatorio Chileno de Salud Pública (s.f.a). Estructura organizacional. Recuperado el 22 de Junio de 2020 de http://www.ochisap.cl/index.php/organizacion-y-estructura-del-sistema-de-salud/estructura-organizacional. 4.1.1

OCHISAP, Observatorio Chileno de Salud Pública (s.f.b). Estructura organizacional del snss. Recuperado el 22 de Junio de 2020 de http://www.ochisap.cl/index.php/organizacion-y-estructura-del-sistema-de-salud/estructura-organizacional-del-snss. 4.1.4.1, 4.1.4.1

OCHISAP, Observatorio Chileno de Salud Pública (s.f.c). Los servicios de salud del s.n.s.s. Recuperado el 22 de Junio de 2020 de http://www.ochisap.cl/index.php/los-servicios-de-salud-del-s-n-s-s. 4.1.4.1

OMS, Organización Mundial de la Salud (2008). Determinantes sociales de la salud. Recuperado el 26 de julio de 2020 de https://www.who.int/social_determinants/es/. 2

OMS, Organización Mundial de la Salud (2012). Vigilancia de la seguridad de los medicamentos. Recuperado el 22 de junio 2020 de https://www.who.int/medicines/areas/quality_safety/safety_efficacy/WHO-UMC_ReportingGeneralPublic-ESP-GRA3Final.pdf?ua=1. 4.1.4.1

OMS, Organización Mundial de la Salud (2015). Informe mundial sobre el envejecimiento y la salud. Recuperado el 04 de julio de 2020 de https://apps.who.int/iris/bitstream/handle/10665/186466/9789240694873_spa.pdf?sequence=1. 1

OPS, Organización Panamericana de la Salud (sf.). Proósito y aplicabilidad de la cie. Recuperado el 10 de julio de 2020 de http://ais.paho.org/cie/index.asp?xml=purpose.htm. 1

OPS, Organización Panamericana de Salud (2019). Indicadores básicos 2019: Tendencias de la salud en las américas. Recuperado el 23 de junio de 2020 de https://iris.paho.org/bitstream/handle/10665.2/51543/9789275321287_spa.pdf?sequence=7&isAllowed=y. 4.1.2

Orellana, Johanna (2018). Árboles de decisión y random forest. Recuperado el 05 de agosto de 2020 de https://bookdown.org/content/2031/. 6.4.2

Red PARF, Red Panamericana de Armonización de la Reglamentación Farmacéutica (2010). Buenas prácticas de farmacovigilancia para las américas. *Organización Mundial de la Salud*. Recuperado el 21 de junio de 2020 de https://www.paho.org/hq/dmdocuments/2011/Technical-Doc-5-web.pdf. 1

Red PARF, Red Panamericana de Armonización de la Reglamentación Farmacéutica (2011). Criterios Éticos para la promociÓn, propaganda y publicidad de medicamentos. *Organización Mundial de la Salud*. Recuperado el 21 de junio de 2020 de https://www.paho.org/hq/dmdocuments/2011/Criterios-Eticos-Promocion-Propaganda-y-Publicidad-05-2011.pdf. 4.2.3

Redman, T. (2016). Assess whether you have a data quality problem. *Harvard Business Review*. Recuperado el 25 de junio de 2020 de https://hbr.org/2016/07/assess-whether-you-have-a-data-quality-problem. 4.4.1

Ricchione, D. (2020). La farmacia en chile. *Centro de Profesionales Farmaceuticos (CEPROFAR)*. Recuperado el 20 de junio de 2020 de http://www.ceprofar.com.ar/2020/01/21/la-farmacia-en-chile/. 4.2.2.1

Rodríguez, Daniel (2018). La regresión logística. Recuperado el 08 de agosto de 2020 de https://www.analyticslane.com/2018/07/23/la-regresion-logistica/. 4.3.2.1, 4.3.2.1

Roldán, J. (2016). Farmacovigilancia: Datos sobre el estado actual de esta disciplina en chile. *Revista Médica Clínica Las Condes*, 585-593, 585–593. Recuperado el 05 de julio de 2020 de https://www.sciencedirect.com/science/article/pii/S0716864016300839. 1

Rommers, Mirjam; Teepe-Twiss, Irene; y Guchelaar, Henk-Jan (2007). Preventing adverse drug events in hospital practice: An overview. *Pharmacoepidemiology and drug safety*, 16, 1129–35. 2

Saltos, J. (2018). Exponiendo y cuantificando la mala calidad de datos. *Freelance*. Recuperado el 25 de junio de 2020 de https://www.freelancermap.com/blog/es/mala-calidad-datos/. 4.4.1

Scribano, A. y De Sena, A. (2009). Las segundas partes sí pueden ser mejores: Algunas reflexiones sobre el uso de datos secundarios en la investigación cualitativa. Recuperado el 20 de junio de 2020 de https://www.scielo.br/pdf/soc/n22/n22a06.pdf. 4.4

Stausberg, Jürgen (2014). International prevalence of adverse drug events in hospitals: An analysis of routine data from england, germany, and the usa. *BMC health services research*, 14, 125. 6.1.5.2

Supersalud, Superintendencia de Salud (s.f.a). Aprende lo esencial antes de firmar un contrato de salud ante una isapre. Recuperado el 21 de junio de 2020 de http://www.supersalud.gob.cl/difusion/665/w3-article-6329.html. 4.1.4.2

Supersalud, Superintendencia de Salud (s.f.b). Cómo funciona el sistema de salud en chile. Recuperado el 20 de junio de 2020 de http://www.supersalud.gob.cl/difusion/665/w3-article-17328.html#accordion_1. 4.1.4.1, 4.1.4.2

Supersalud, Superintendencia de Salud (s.f.c). ¿qué es auge o ges? Recuperado el 21 de junio de 2020 de http://www.supersalud.gob.cl/consultas/667/w3-article-4605.html. 4.1.4.2

Varallo, Fabiana; Guimarães, Synara; Abjaude, Samir; y Mastroianni, Patricia (2014). Causas del subregistro de los eventos adversos de medicamentos por los profesionales de la salud: revisión sistemática. *Revista Da Escola de Enfermagem Da USP*, (pp. 739–747). 2

Ventura-León, José Luis (2017). El significado de la significancia estadÃstica: comentarios a MartÃnez-Ferrer y colaboradores. *Salud Pública de México*, 59, 499 – 500. 4.3.4

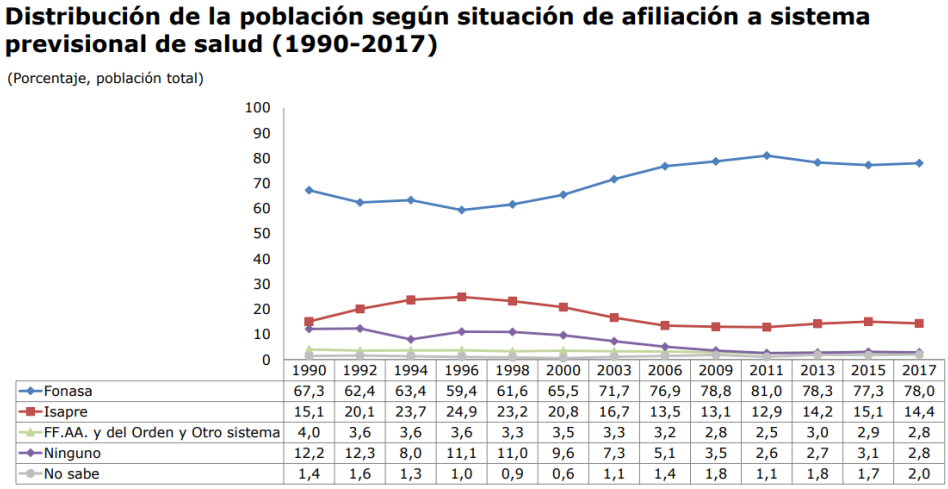
Vergara, H. (2019). El escándalo de los medicamentos en chile. Recuperado el 20 de junio de 2020 de http://www.farmaciadaniela.cl/noticias/el-escandalo-de-los-medicamentos-en-chile. 4.2.2

Wu, Tai-Yin; Jen, Min-Hua; Bottle, Alex; Molokhia, Mariam; Aylin, Paul; Bell, Derek; y Majeed, Azeem (2010). Ten-year trends in hospital admissions for adverse drug reactions in england 1999-2009. *Journal of the Royal Society of Medicine*, 103, 239–50. 6.1.5.1

A | ANEXOS

A.1. Marco Teórico

A.1.1. Distribución de población según situación de afiliación a sistema previsional de salud



	1990	1992	1994	1996	1998	2000	2003	2006	2009	2011	2013	2015	2017
Fonasa	67,3	62,4	63,4	59,4	61,6	65,5	71,7	76,9	78,8	81,0	78,3	77,3	78,0
Isapre	15,1	20,1	23,7	24,9	23,2	20,8	16,7	13,5	13,1	12,9	14,2	15,1	14,4
FF.AA. y del Orden y Otro sistema	4,0	3,6	3,6	3,6	3,3	3,5	3,3	3,2	2,8	2,5	3,0	2,9	2,8
Ninguno	12,2	12,3	8,0	11,1	11,0	9,6	7,3	5,1	3,5	2,6	2,7	3,1	2,8
No sabe	1,4	1,6	1,3	1,0	0,9	0,6	1,1	1,4	1,8	1,1	1,8	1,7	2,0

Figura A.1: Distribución de la población según situación de afiliación a sistema previsional de salud (1990-2017)

Fuente: Síntesis de resultados CASEN 2017. Ministerio de Desarrollo Social

A.1.2. La farmacia en Chile

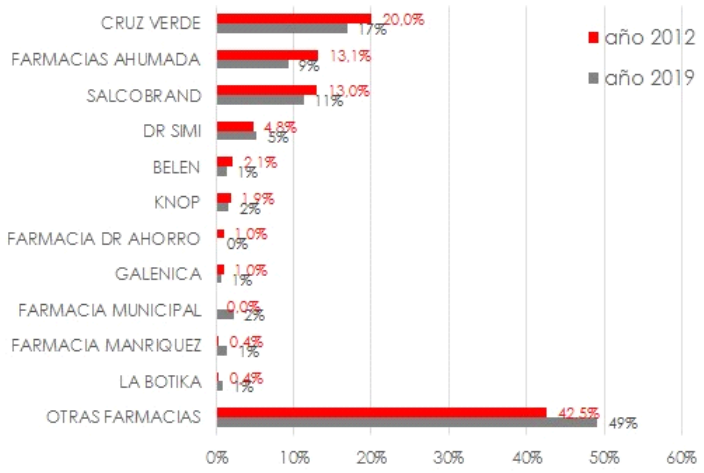


Figura A.2: Comparación de la participación de farmacias por cadena

Fuente: La farmacia en Chile. Centro de Profesionales Farmacéuticos (CEPROFAR, Argentina).

A.1.3. Metodología CRISP-DM

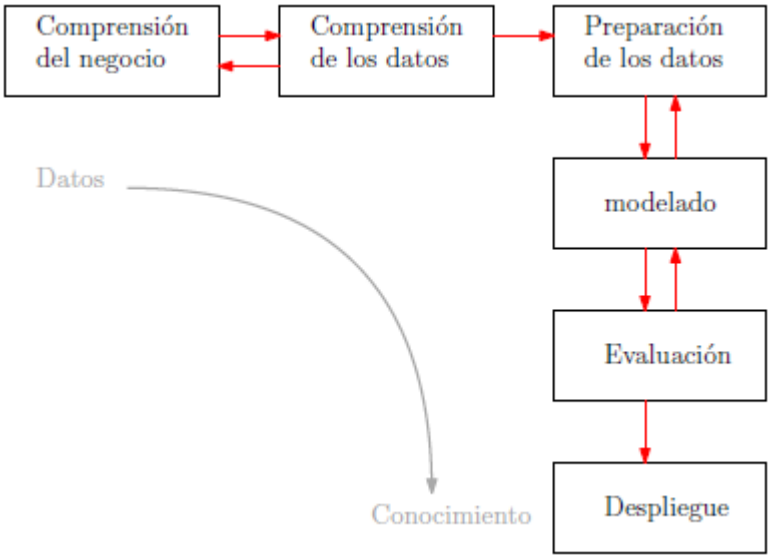


Figura A.3: Fases de la metodología CRISP-DM
Fuente: Minería de datos, modelos y algoritmos (pág. 28)

A.2. Metodología

A.2.1. Diccionario de datos SINIM

Dato	Descripción	Objetivo	Unidad de medida	Rango
HPISM Población Inscrita Validada en Servicios de Salud Municipal (FONASA)	Total de personas inscritas en el servicio de salud y validada por el servicio de salud municipal (FONASA).		Nº NÚMERO ENTERO	0,infinito

Figura A.4: Formato de registro Población Inscrita Validada en Servicios de Salud Municipal (FONASA)

Fuente: Diccionario de Datos. SINIM

A.2.2. Listados de códigos *diag1* y *diag2* según DEIS

Año	diag1	diag2
2011	-	-
2012	8824	3320
2013	8825	3320
2014	8818	3314
2015	8866	3314
2016	8867	3314
2017	8627	3314
2018	8491	3314

*En azul se selecciona el listado con mayor cantidad de códigos

Tabla A.1: Listado de códigos por año informados por DEIS

Fuente: Elaboración propia

A.3. Resultados

A.3.1. Comprensión de los datos

- Reporte del campo servicio de egreso nulo

```
select nombre_estab,estab,count(*) from ieeh18,establecimientos_deis where
estab=ncode and serc_egr="" group by nombre_estab order by 3 DESC limit 10;

+-----------------------------------------------+--------+----------+
| nombre_estab                                  | estab  | count(*) |
+-----------------------------------------------+--------+----------+
| Clinica Alemana                               | 112200 |    37324 |
| Clinica Santa Maria                           | 112249 |    30718 |
| Hospital Barros Luco Trudeau Santiago San Miguel | 113100 |    30233 |
| Clinica Indisa                                | 112211 |    28491 |
| Hospital Dr. Cesar Garavagno Burotto Talca    | 116105 |    26045 |
| Clinica Las Condes                            | 112212 |    25694 |
| Hospital Clinico Universidad de Chile         | 109200 |    22965 |
| Hospital de Puerto Montt                      | 124105 |    22350 |
| Red Salud Santiago ex Clinica Bicentenario    | 111295 |    21615 |
| Hospital Carlos Van Buren Valparaiso          | 106100 |    18949 |
+-----------------------------------------------+--------+----------+
```

Figura A.5: Conteo de servicio de egreso nulo por establecimiento (2018)

Fuente: Elaboración propia (MySQL)

- Inconsistencia FONASA

```
+-------------------------------------------------------------+--------+----------+
| nombre_estab                                                | estab  | count(*) |
+-------------------------------------------------------------+--------+----------+
| Clinica Iquique                                             | 102200 |       80 |
| Pensionado San Jose                                         | 112229 |       33 |
| Hospital de Urgencia Asistencia Publica Dr. Alejandro del Rio | 111195 |       29 |
| Hospital Clinico San Borja Arriaran                         | 111100 |       24 |
| Clinica Psiquiatrica Renacer                                | 112252 |       18 |
| Clinica Psiquiatrica Bretana                                | 112205 |       13 |
| Clinica nunoa                                               | 112222 |       11 |
| Complejo Asistencial Dr. Victor Rios Ruiz Los angeles       | 120101 |        9 |
| Hospital San Vicente de Tagua -Tagua                        | 115105 |        9 |
| Clinica Psicoterapia los Tiempos                            | 112258 |        9 |
+-------------------------------------------------------------+--------+----------+
```

Figura A.6: Inconsistencia FONASA 2018

Elaboración Propia (MySQL)

```
+-------------------------------------------------------------+--------+----------+
| nombre_estab                                                | estab  | count(*) |
+-------------------------------------------------------------+--------+----------+
| Clinica nunoa                                               | 112222 |       34 |
| Hospital Clinico San Borja Arriaran                         | 111100 |       34 |
| Hospital de Urgencia Asistencia Publica Dr. Alejandro del Rio | 111195 |       20 |
| Hospital Regional Dr. Juan Noe Crevanni Arica               | 101100 |       16 |
| Pensionado San Jose                                         | 112229 |       14 |
| Instituto de Seguridad del Trabajo                          | 107211 |       12 |
| Clinica Psiquiatrica Bretana                                | 112205 |       11 |
| Hospital Dr. Gustavo Fricke Vina del Mar                    | 107100 |        6 |
| Hospital Dr. Ricardo Valenzuela Saez Rengo                  | 115104 |        4 |
| Hospital San Juan de Dios Santiago Santiago                 | 110100 |        3 |
+-------------------------------------------------------------+--------+----------+
```

Figura A.7: Inconsistencia FONASA 2017

Elaboración Propia (MySQL)

nombre_estab	estab	count(*)
Clinica nunoa	112222	83
Clinica Cumbres del Norte S.A.	103224	19
Clinica Psiquiatrica Bretana	112205	11
Hospital Dr. Gustavo Fricke Vina del Mar	107100	10
Hospital Clinico Instituto de Seguridad del Trabajo de Santaigo	111205	9
Hospital San Vicente de Tagua -Tagua	115105	7
Hospital de Rio Bueno	122105	7
Hospital San Francisco de Pucon D	121200	7
Hospital San Juan de Dios Santiago Santiago	110100	6
Complejo Asistencial Dr. Victor Rios Ruiz Los angeles	120101	5

Figura A.8: Inconsistencia FONASA 2016

Elaboración Propia (MySQL)

nombre_estab	estab	count(*)
Hospital del Cobre Salvador Allende	103219	336
Clinica nunoa	112222	153
Clinica Juan Pablo II	111221	41
Hospital Clinico Instituto de Seguridad del Trabajo de Santaigo	111205	34
Hospital Dr. Cesar Garavagno Burotto Talca	116105	31
Complejo Hospitalario Dr. Sotero del Rio Santiago Puente Alto	114101	17
Hospital San Juan de Dios Santiago Santiago	110100	15
Hospital Clinico Metropolitano La Florida Dra. Eloisa Diaz Insunza	114105	14
Clinica de Salud Integral	115222	12
Hospital Dr. Abel Fuentealba Lagos de San Javier	116109	11

Figura A.9: Inconsistencia FONASA 2015

Elaboración Propia (MySQL)

nombre_estab	estab	count(*)
Hospital Dr. Ernesto Torres Galdames Iquique	102100	412
Clinica Tarapaca	102201	354
Hospital del Cobre Salvador Allende	103219	208
Clinica El Loa	103218	119
Clinica nunoa	112222	114
Clinica San Jose	101213	99
Hospital Clinico Instituto de Seguridad del Trabajo de Santaigo	111205	81
Hospital San Juan de Dios La Serena	105100	49
Hospital de Quellon	133165	27
Hospital de Rio Bueno	122105	15

Figura A.10: Inconsistencia FONASA 2014

Elaboración Propia (MySQL)

nombre_estab	estab	count(*)
Clinica Santa Maria	12-249	29127
Clinica Las Condes	12-212	20112
Hospital Clinico Universidad de Chile	09-200	14795
Hospital Militar de Santiago	12-530	11099
Clinica Antofagasta	03-203	7566
Clinica Alemana de Temuco	21-202	6937
Clinica Davila	09-201	6092
Clinica Vespucio	14-223	5425
Hospital Clinico de Magallanes Dr. Lautaro Navarro Avaria	26-100	5074
Clinica de la Mujer Sanatorio Aleman	18-202	4349

Figura A.11: Inconsistencia FONASA 2013

Elaboración Propia (MySQL)

nombre_estab	estab	count(*)
Hospital Clinico Universidad de Chile	09-200	14573
Hospital Militar de Santiago	12-530	11286
Clinica Alemana de Temuco	21-202	8456
Hospital Las Higueras Talcahuano	19-100	7810
Clinica Antofagasta	03-203	6077
Clinica de la Mujer Sanatorio Aleman	18-202	5105
Clinica Regional Elqui	05-208	3510
Hospital Clinico del Sur S.A.	18-200	3421
Hospital FFAA Cirujano Guzman	26-200	3327
Clinica Puerto Montt	24-250	2806

Figura A.12: Inconsistencia FONASA 2012

Elaboración Propia (MySQL)

- Contabilización EAM, antes y después de la limpieza de datos

	Stausberg		Wu et al	
	Inicial	Final	Inicial	Final
2011	11 673	11 529	3545	3 476
2012	11 347	11 320	3 388	3 374
2013	11 726	11 700	3 592	3 584
2014	11469	11 426	3 616	3 604
2015	11460	11 418	4 137	4 124
2016	10 327	10 299	4 157	4 143
2017	10 987	10 964	4379	4 373
2018	11 850	11 740	4 969	4 920
Total	**90 839**	**90 396**	**31 783**	**31 598**

Tabla A.2: Contabilización casos EAM

Fuente: Elaboración propia

A.3.2. Modelado

A.3.2.1. Regresión logística: Variables antiguas

- Sklearn

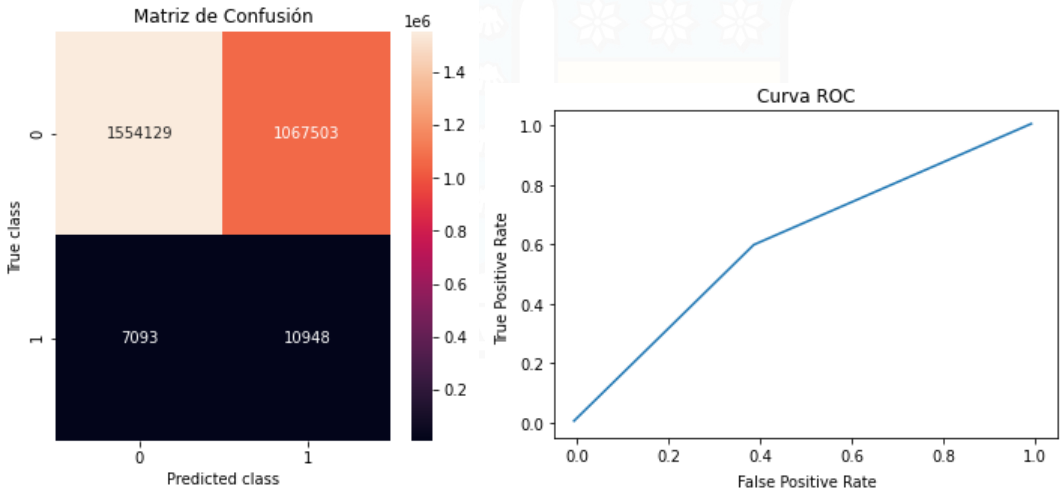


Figura A.13: Regresión logística, sklearn, reducción 1

Fuente: Elaboración propia

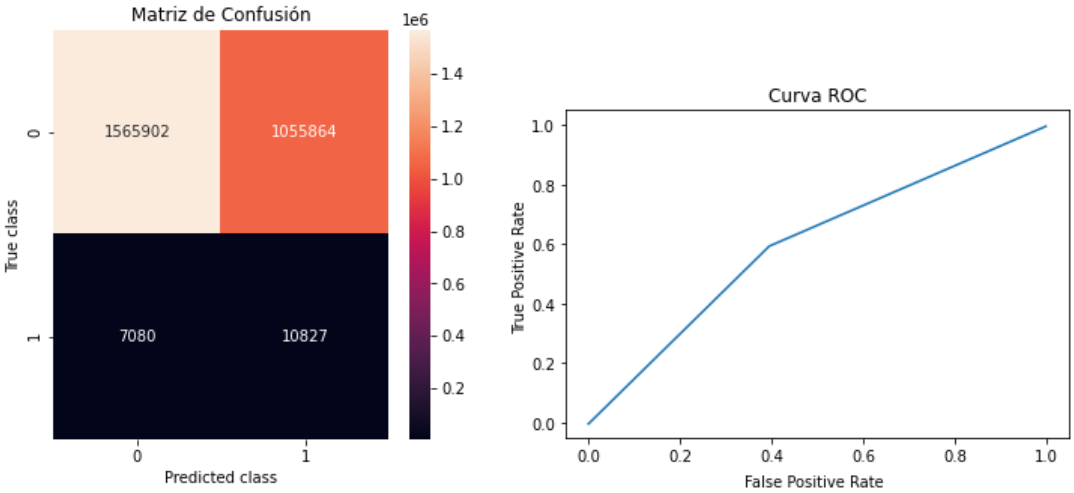


Figura A.14: Regresión logística, sklearn, reducción 2

Fuente: Elaboración propia

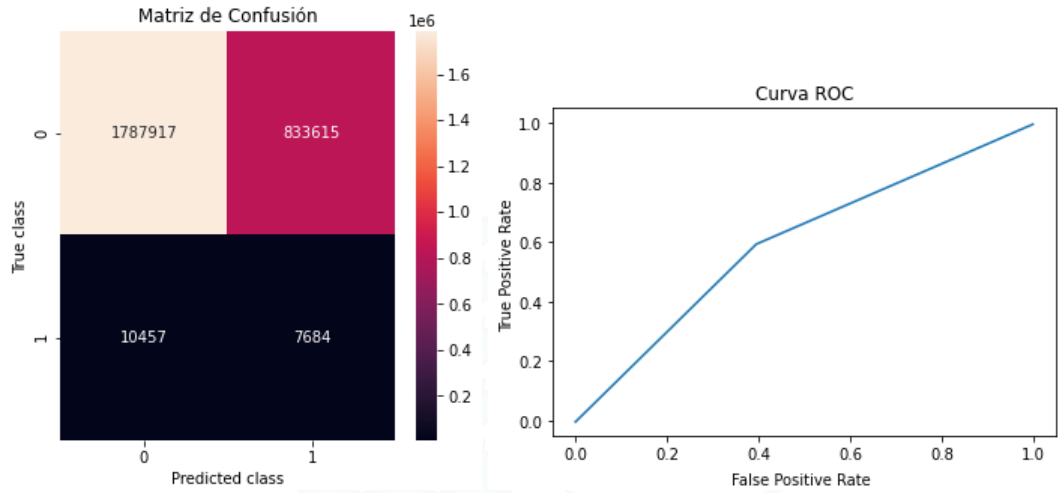


Figura A.15: Regresión logística, sklearn, reducción 3

Fuente: Elaboración propia

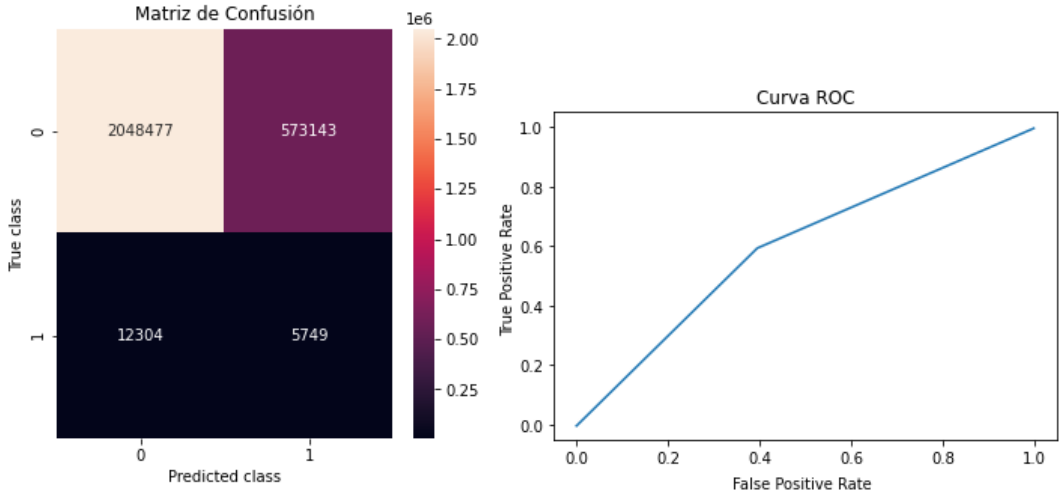


Figura A.16: Regresión logística, sklearn, reducción 4

Fuente: Elaboración propia

- Statsmodels

```
                           Logit Regression Results
==============================================================================
Dep. Variable:                      y   No. Observations:             10558690
Model:                          Logit   Df Residuals:                 10558669
Method:                           MLE   Df Model:                           20
Date:                Sun, 30 Aug 2020   Pseudo R-squ.:                 0.01542
Time:                        20:44:15   Log-Likelihood:            -4.2534e+05
converged:                       True   LL-Null:                   -4.3200e+05
Covariance Type:            nonrobust   LLR p-value:                    0.000
==================================================================================
                  coef    std err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------------------
año            -0.0023      0.000    -11.753      0.000      -0.003      -0.002
sexo1           0.0870      0.395      0.220      0.826      -0.688       0.862
sexo2          -0.0016      0.395     -0.004      0.997      -0.777       0.773
edad           -0.0136      0.000    -82.311      0.000      -0.014      -0.013
previ1          0.8520      0.081     10.565      0.000       0.694       1.010
previ2          0.2895      0.021     13.574      0.000       0.248       0.331
previ3          0.3498      0.032     10.866      0.000       0.287       0.413
benef1         -0.0875      0.078     -1.117      0.264      -0.241       0.066
benef2         -0.2732      0.079     -3.463      0.001      -0.428      -0.119
benef3         -0.5088      0.080     -6.388      0.000      -0.665      -0.353
benef4         -0.6212      0.079     -7.824      0.000      -0.777      -0.466
f_ing           0.0028      0.001      2.547      0.011       0.001       0.005
region         -0.0213      0.001    -18.685      0.000      -0.024      -0.019
poblacion   -1.811e-07   4.11e-08     -4.402      0.000   -2.62e-07      -1e-07
superficie     -0.2991      0.047     -6.377      0.000      -0.391      -0.207
pobreza        -0.0087      0.001    -14.316      0.000      -0.010      -0.007
farma          -0.0002      0.000     -1.135      0.257      -0.001       0.000
almacen        -0.1486      0.009    -17.338      0.000      -0.165      -0.132
primario       -0.0048      0.001     -7.114      0.000      -0.006      -0.003
secundario      0.0180      0.004      4.327      0.000       0.010       0.026
terciario       0.0558      0.005     11.492      0.000       0.046       0.065
==================================================================================
```

Figura A.17: Logit sin intercepto: Completo. Statsmodels

Fuente: Elaboración propia

```
                           Logit Regression Results
==============================================================================
Dep. Variable:                      y   No. Observations:             10558690
Model:                          Logit   Df Residuals:                 10558673
Method:                           MLE   Df Model:                           16
Date:                Sun, 30 Aug 2020   Pseudo R-squ.:                 0.01525
Time:                        20:47:38   Log-Likelihood:            -4.2466e+05
converged:                       True   LL-Null:                   -4.3124e+05
Covariance Type:            nonrobust   LLR p-value:                    0.000
==============================================================================
                  coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
año            -0.0023   1.32e-05   -173.516      0.000      -0.002      -0.002
edad           -0.0135      0.000    -81.497      0.000      -0.014      -0.013
previ1          0.7559      0.020     36.996      0.000       0.716       0.796
previ2          0.2848      0.021     13.361      0.000       0.243       0.327
previ3          0.3584      0.032     11.170      0.000       0.296       0.421
benef2         -0.1936      0.010    -18.841      0.000      -0.214      -0.173
benef3         -0.4300      0.015    -28.640      0.000      -0.459      -0.401
benef4         -0.5331      0.013    -39.650      0.000      -0.559      -0.507
f_ing           0.0019      0.001      1.730      0.084      -0.000       0.004
region         -0.0217      0.001    -19.069      0.000      -0.024      -0.019
poblacion   -2.191e-07    3.7e-08     -5.924      0.000   -2.92e-07   -1.47e-07
superficie     -0.3565      0.047     -7.542      0.000      -0.449      -0.264
pobreza        -0.0080      0.001    -13.882      0.000      -0.009      -0.007
almacen        -0.1518      0.009    -17.527      0.000      -0.169      -0.135
primario       -0.0044      0.001     -6.528      0.000      -0.006      -0.003
secundario      0.0189      0.004      4.569      0.000       0.011       0.027
terciario       0.0513      0.004     13.601      0.000       0.044       0.059
==============================================================================
```

Figura A.18: Logit sin intercepto: Reducción 1. Statsmodels

Fuente: Elaboración propia

A.3.2.2. Regresión logística: Variables nuevas

- Sklearn

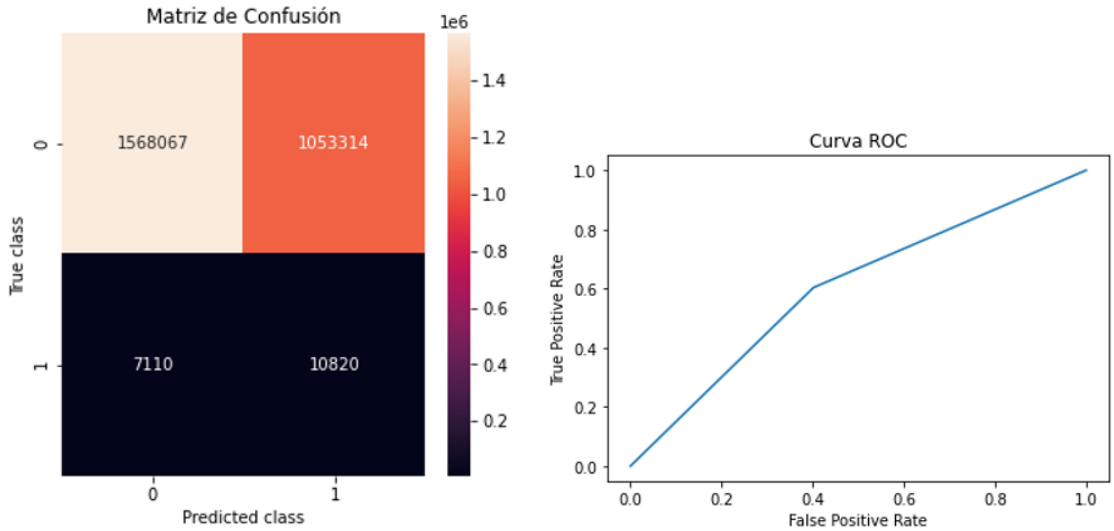


Figura A.19: Regresión logística, Sklearn, todas las variables, variables nuevas

Fuente: Elaboración propia

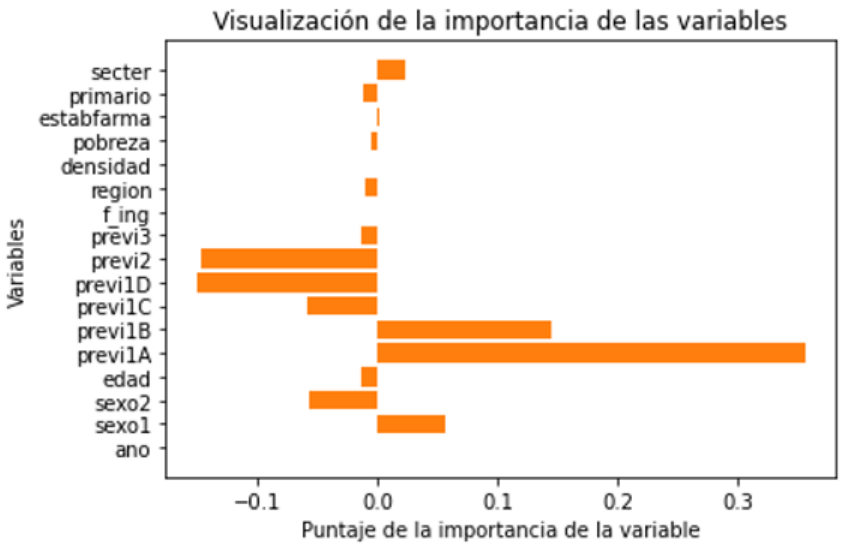


Figura A.20: Importancia de las variables, Sklearn, todas las variables, variables nuevas

Fuente: Elaboración propia

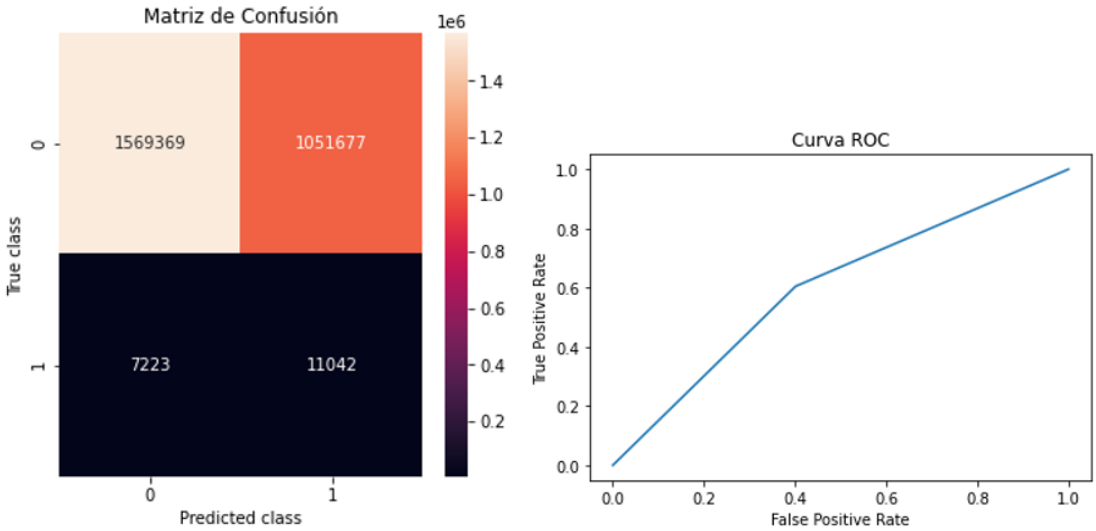


Figura A.21: Regresión logística, Sklearn, reducción 1, variables nuevas

Fuente: Elaboración propia

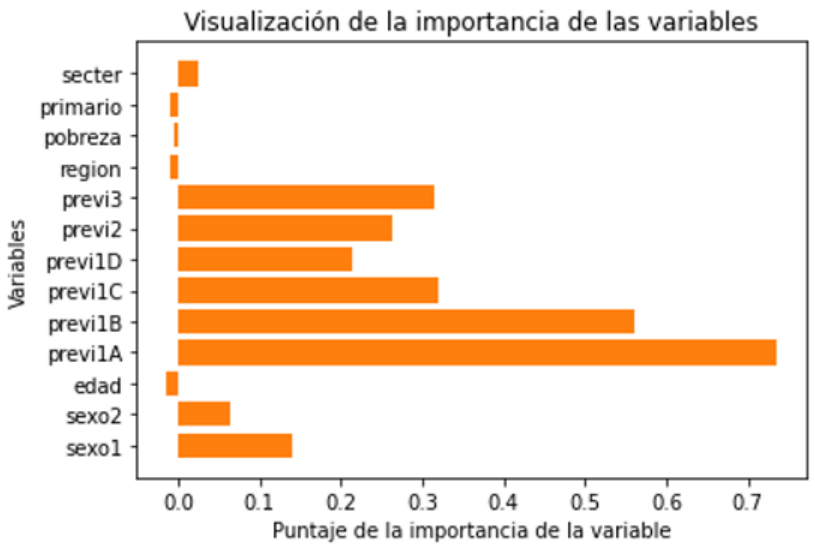


Figura A.22: Importancia de las variables, Sklearn, reducción 1, variables nuevas

Fuente: Elaboración propia

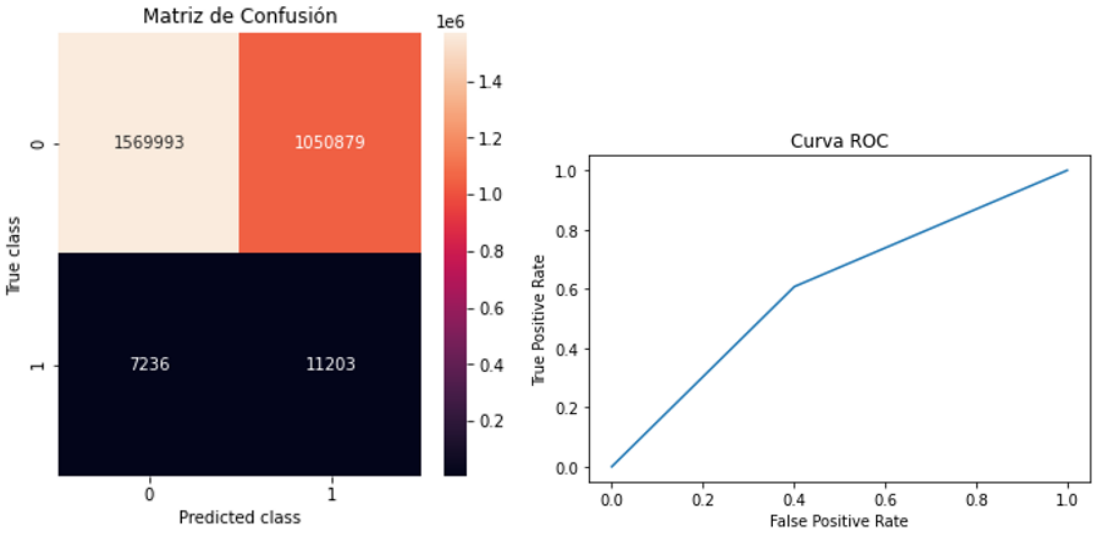


Figura A.23: Regresión logística, Sklearn, reducción 2, variables nuevas

Fuente: Elaboración propia

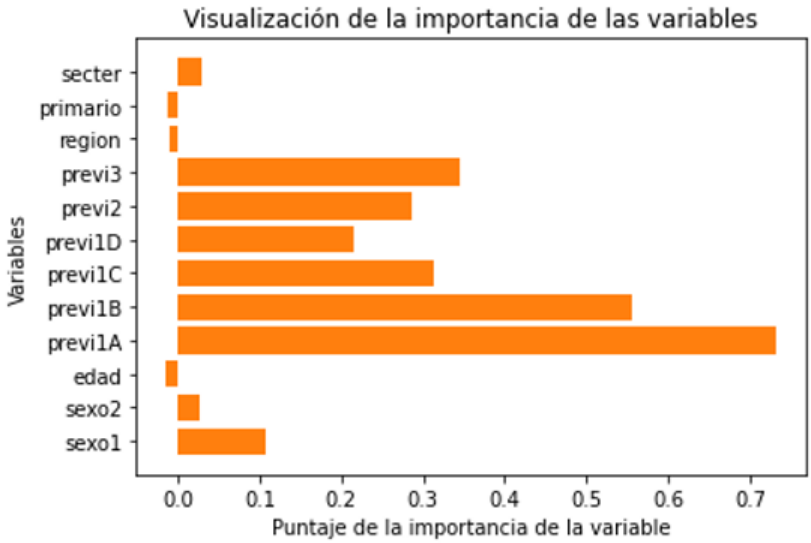


Figura A.24: Importancia de las variables, Sklearn, reducción 2, variables nuevas

Fuente: Elaboración propia

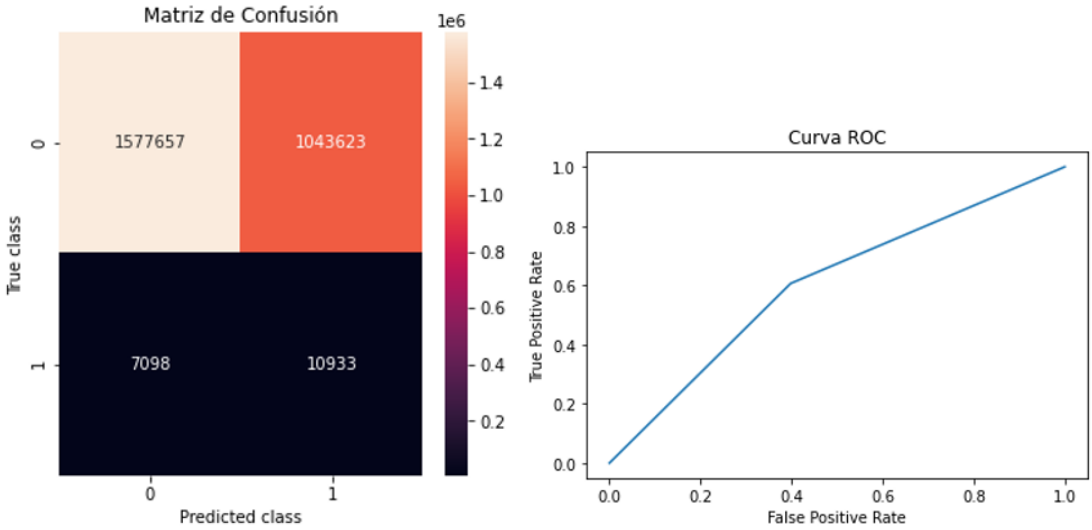


Figura A.25: Regresión logística, Sklearn, reducción 3, variables nuevas

Fuente: Elaboración propia

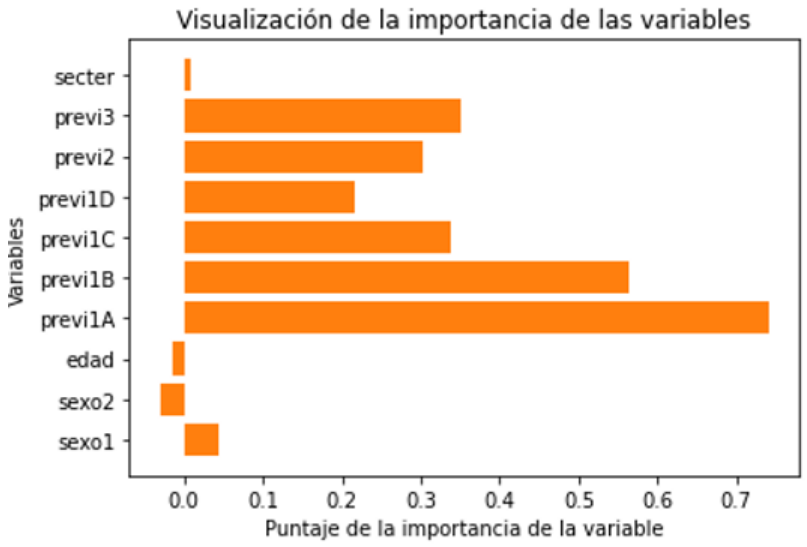


Figura A.26: Importancia de las variables, Sklearn, reducción 3, variables nuevas

Fuente: Elaboración propia

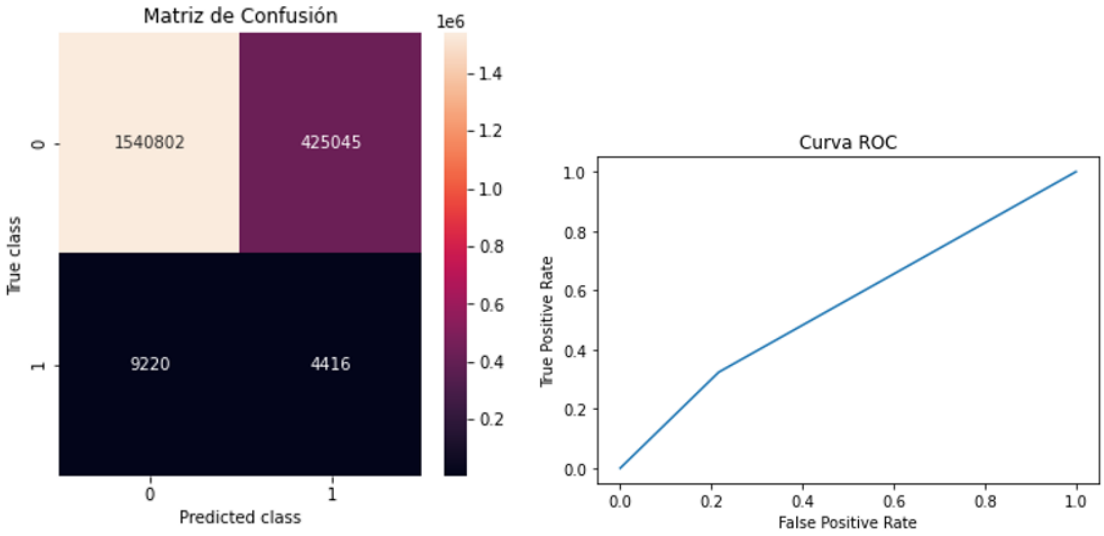


Figura A.27: Regresión logística, Sklearn, reducción 4, variables nuevas

Fuente: Elaboración propia

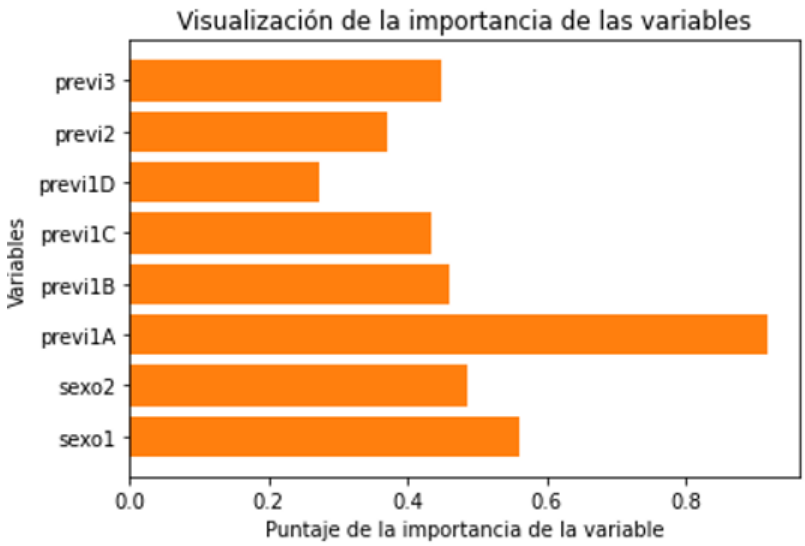


Figura A.28: Importancia de las variables, Sklearn, reducción 4, variables nuevas

Fuente: Elaboración propia

- Statsmodels

```
                           Logit Regression Results
==============================================================================
Dep. Variable:                      y   No. Observations:             10557240
Model:                          Logit   Df Residuals:                 10557222
Method:                           MLE   Df Model:                           17
Date:                Sun, 30 Aug 2020   Pseudo R-squ.:                 0.01433
Time:                        21:01:05   Log-Likelihood:            -4.2678e+05
converged:                       True   LL-Null:                   -4.3298e+05
Covariance Type:            nonrobust   LLR p-value:                    0.000
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
intercept    -11.0036      3.421     -3.217      0.001     -17.708      -4.299
ano            0.0029      0.002      1.743      0.081      -0.000       0.006
sexo1          0.3770      0.449      0.839      0.401      -0.503       1.257
sexo2          0.2852      0.449      0.635      0.525      -0.595       1.166
edad          -0.0134      0.000    -81.381      0.000      -0.014      -0.013
previ1A        0.7668      0.020     37.472      0.000       0.727       0.807
previ1B        0.5802      0.021     28.259      0.000       0.540       0.620
previ1C        0.3478      0.023     14.878      0.000       0.302       0.394
previ1D        0.2224      0.022      9.953      0.000       0.179       0.266
previ2         0.2842      0.021     13.342      0.000       0.242       0.326
previ3         0.3586      0.032     11.211      0.000       0.296       0.421
f_ing          0.0024      0.001      2.152      0.031       0.000       0.005
region        -0.0086      0.001     -7.373      0.000      -0.011      -0.006
densidad   -1.016e-05   1.22e-06     -8.346      0.000   -1.25e-05   -7.77e-06
pobreza       -0.0044      0.001     -7.702      0.000      -0.006      -0.003
estab_far      0.0007      0.000      4.379      0.000       0.000       0.001
primario      -0.0103      0.001    -16.695      0.000      -0.012      -0.009
sec_ter        0.0251      0.003      9.332      0.000       0.020       0.030
==============================================================================
```

Figura A.29: Logit con intercepto: Completo - Variables nuevas

Fuente: Elaboración propia

```
                           Logit Regression Results
==============================================================================
Dep. Variable:                      y   No. Observations:             10557240
Model:                          Logit   Df Residuals:                 10557226
Method:                           MLE   Df Model:                           13
Date:                Sun, 30 Aug 2020   Pseudo R-squ.:                -0.04507
Time:                        21:07:20   Log-Likelihood:            -4.5244e+05
converged:                       True   LL-Null:                   -4.3293e+05
Covariance Type:            nonrobust   LLR p-value:                     1.000
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
edad          -0.0248      0.000   -163.554      0.000      -0.025      -0.025
previ1A       -1.1172      0.010   -114.434      0.000      -1.136      -1.098
previ1B       -1.1442      0.010   -111.361      0.000      -1.164      -1.124
previ1C       -1.5392      0.015   -102.600      0.000      -1.569      -1.510
previ1D       -1.6485      0.013   -122.976      0.000      -1.675      -1.622
previ2        -1.5768      0.012   -136.012      0.000      -1.600      -1.554
previ3        -1.5893      0.027    -59.446      0.000      -1.642      -1.537
f_ing         -0.0848      0.001    -87.672      0.000      -0.087      -0.083
region        -0.1184      0.001   -124.695      0.000      -0.120      -0.117
densidad    1.156e-05   1.19e-06      9.743      0.000    9.23e-06    1.39e-05
pobreza       -0.0522      0.001    -96.897      0.000      -0.053      -0.051
estab_far     -0.0041      0.000    -26.270      0.000      -0.004      -0.004
primario      -0.0489      0.001    -78.198      0.000      -0.050      -0.048
sec_ter        0.0200      0.003      7.140      0.000       0.015       0.026
==============================================================================
```

Figura A.30: Logit con intercepto: Reducción 1 - Variables nuevas

Fuente: Elaboración propia

```
                           Logit Regression Results
==============================================================================
Dep. Variable:                      y   No. Observations:             10557240
Model:                          Logit   Df Residuals:                 10557223
Method:                           MLE   Df Model:                           16
Date:                Sun, 30 Aug 2020   Pseudo R-squ.:                 0.01427
Time:                        21:12:15   Log-Likelihood:            -4.2639e+05
converged:                       True   LL-Null:                   -4.3256e+05
Covariance Type:            nonrobust   LLR p-value:                     0.000
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
año           -0.0025      0.000    -12.046      0.000      -0.003      -0.002
sexo1          0.2344      0.409      0.572      0.567      -0.568       1.037
sexo2          0.1402      0.409      0.342      0.732      -0.662       0.943
edad          -0.0134      0.000    -81.211      0.000      -0.014      -0.013
previ1A        0.7727      0.021     37.607      0.000       0.732       0.813
previ1B        0.5853      0.021     28.397      0.000       0.545       0.626
previ1C        0.3711      0.023     15.860      0.000       0.325       0.417
previ1D        0.2345      0.022     10.462      0.000       0.191       0.278
previ2         0.2951      0.021     13.800      0.000       0.253       0.337
previ3         0.3812      0.032     11.944      0.000       0.319       0.444
f_ing          0.0020      0.001      1.785      0.074      -0.000       0.004
region        -0.0091      0.001     -7.820      0.000      -0.011      -0.007
densidad   -1.003e-05   1.22e-06     -8.239      0.000   -1.24e-05   -7.64e-06
pobreza       -0.0050      0.001     -8.955      0.000      -0.006      -0.004
estab_far      0.0005      0.000      3.546      0.000       0.000       0.001
primario      -0.0103      0.001    -16.687      0.000      -0.012      -0.009
sec_ter        0.0237      0.003      8.764      0.000       0.018       0.029
==============================================================================
```

Figura A.31: Logit sin intercepto: Completo - Variables nuevas

Fuente: Elaboración propia

```
                           Logit Regression Results
==============================================================================
Dep. Variable:                      y   No. Observations:             10557240
Model:                          Logit   Df Residuals:                 10557226
Method:                           MLE   Df Model:                           13
Date:                Sun, 30 Aug 2020   Pseudo R-squ.:                 0.01420
Time:                        21:18:06   Log-Likelihood:            -4.2604e+05
converged:                       True   LL-Null:                   -4.3218e+05
Covariance Type:            nonrobust   LLR p-value:                    0.000
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
año           -0.0023   1.23e-05   -190.895      0.000      -0.002      -0.002
edad          -0.0135      0.000    -81.235      0.000      -0.014      -0.013
previ1A        0.7405      0.020     36.406      0.000       0.701       0.780
previ1B        0.5526      0.020     27.059      0.000       0.513       0.593
previ1C        0.3225      0.023     13.854      0.000       0.277       0.368
previ1D        0.2008      0.022      9.025      0.000       0.157       0.244
previ2         0.2663      0.021     12.547      0.000       0.225       0.308
previ3         0.3503      0.032     11.003      0.000       0.288       0.413
region        -0.0094      0.001     -8.070      0.000      -0.012      -0.007
densidad   -1.019e-05   1.22e-06     -8.353      0.000   -1.26e-05    -7.8e-06
pobreza       -0.0046      0.001     -8.295      0.000      -0.006      -0.004
estab_far      0.0005      0.000      3.384      0.001       0.000       0.001
primario      -0.0105      0.001    -16.983      0.000      -0.012      -0.009
sec_ter        0.0267      0.003      9.906      0.000       0.021       0.032
==============================================================================
```

Figura A.32: Logit sin intercepto: Reducción 1 - Variables nuevas

Fuente: Elaboración propia

A.3.2.3. Bosques Aleatorios: Variables antiguas

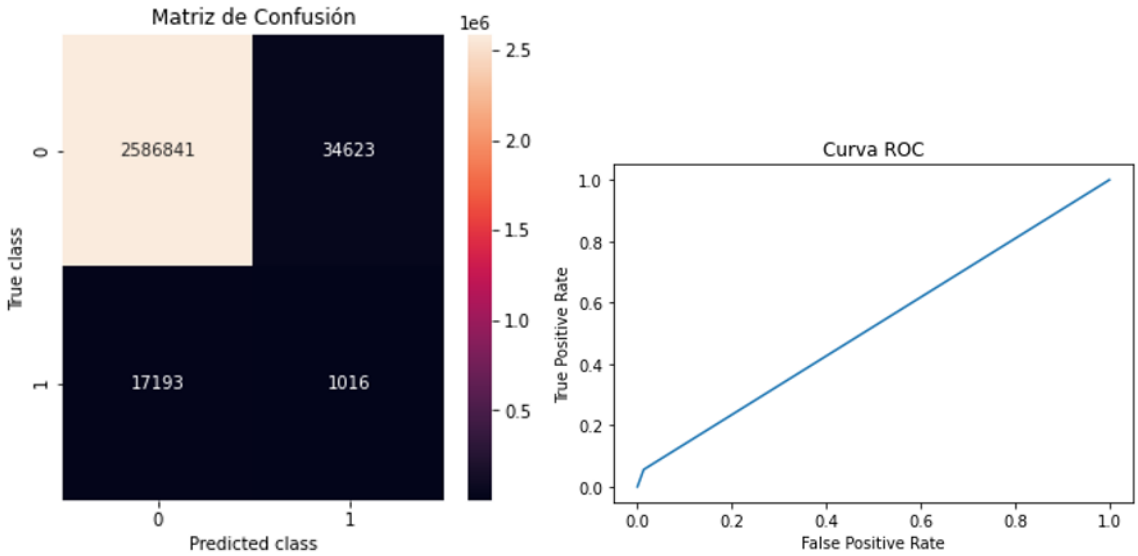


Figura A.33: Bosques Aleatorios, todas las variables, variables antiguas

Fuente: Elaboración propia

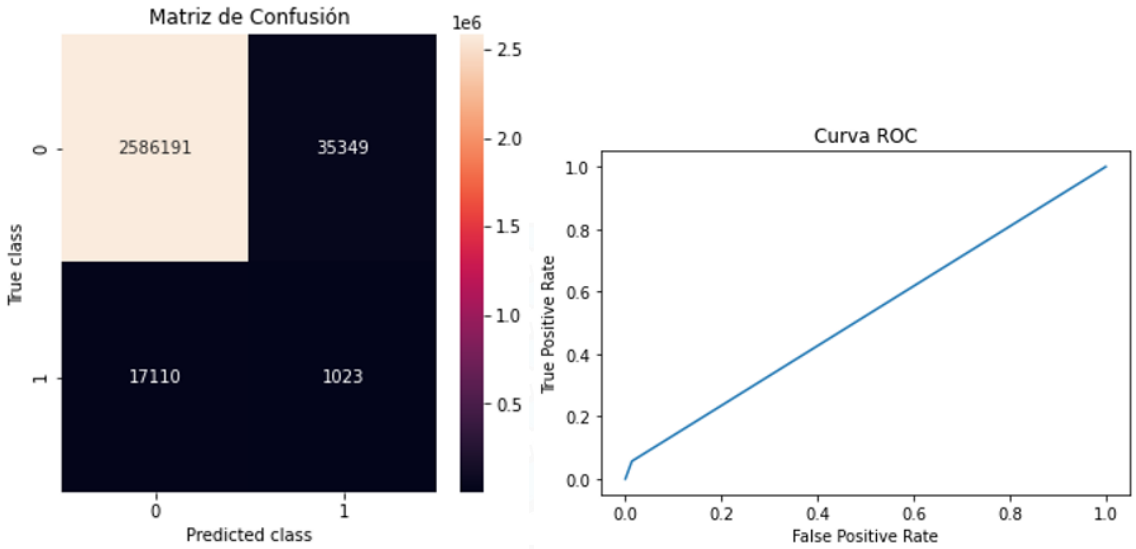


Figura A.34: Bosques Aleatorios, reducción 1, variables antiguas

Fuente: Elaboración propia

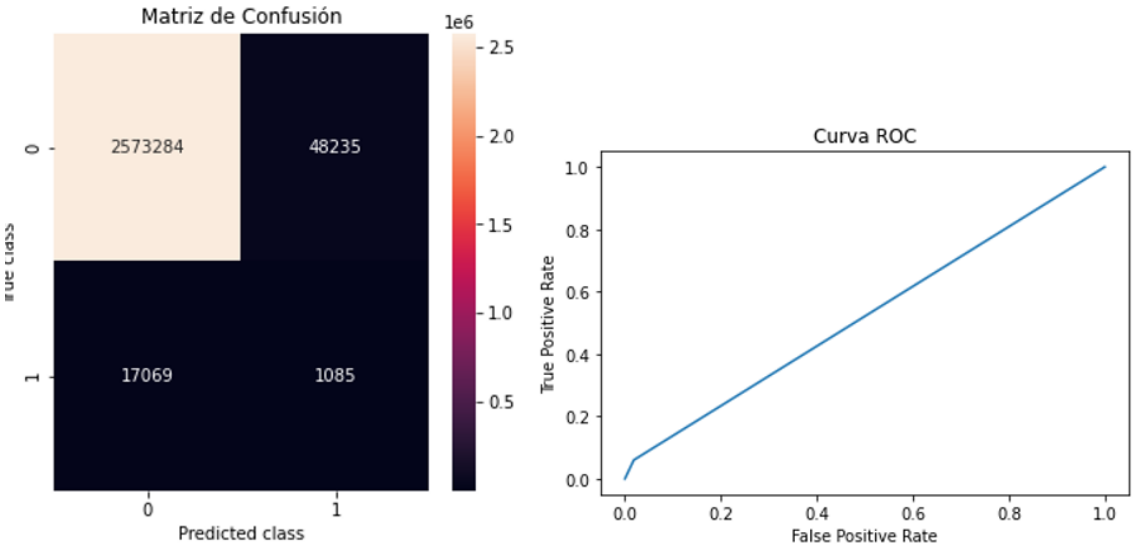


Figura A.35: Bosques Aleatorios, reducción 2, variables antiguas

Fuente: Elaboración propia

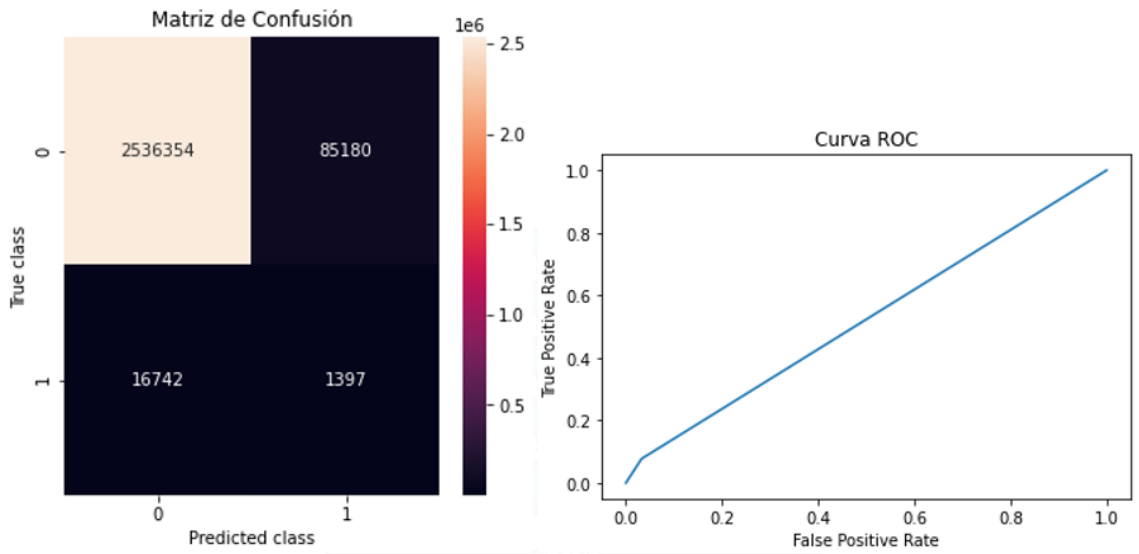


Figura A.36: Bosques Aleatorios, reducción 3, variables antiguas

Fuente: Elaboración propia

A.3.2.4. Bosques Aleatorios: Variables nuevas

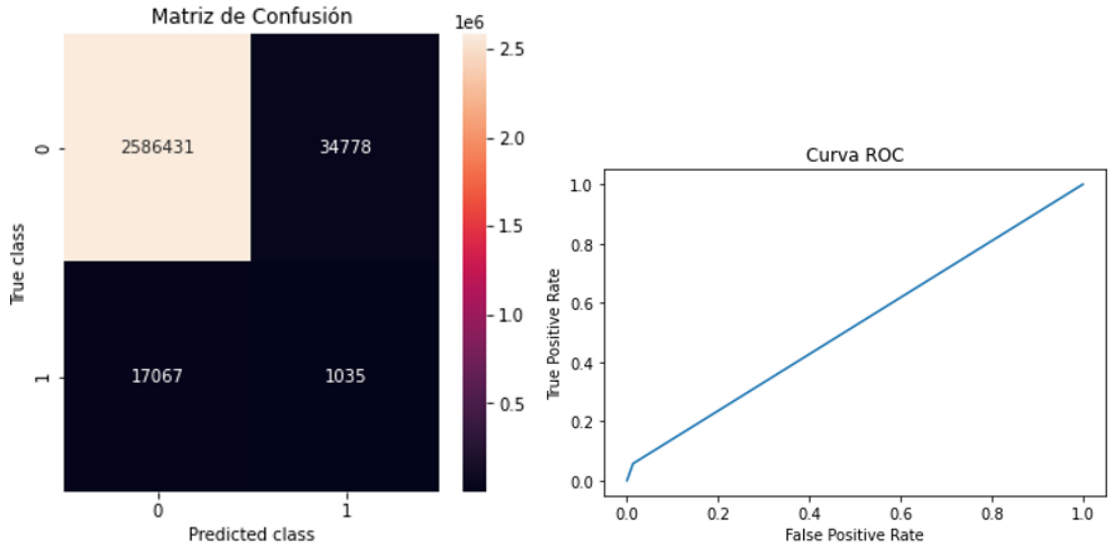


Figura A.37: Bosques Aleatorios, todas las variables, variables nuevas

Fuente: Elaboración propia

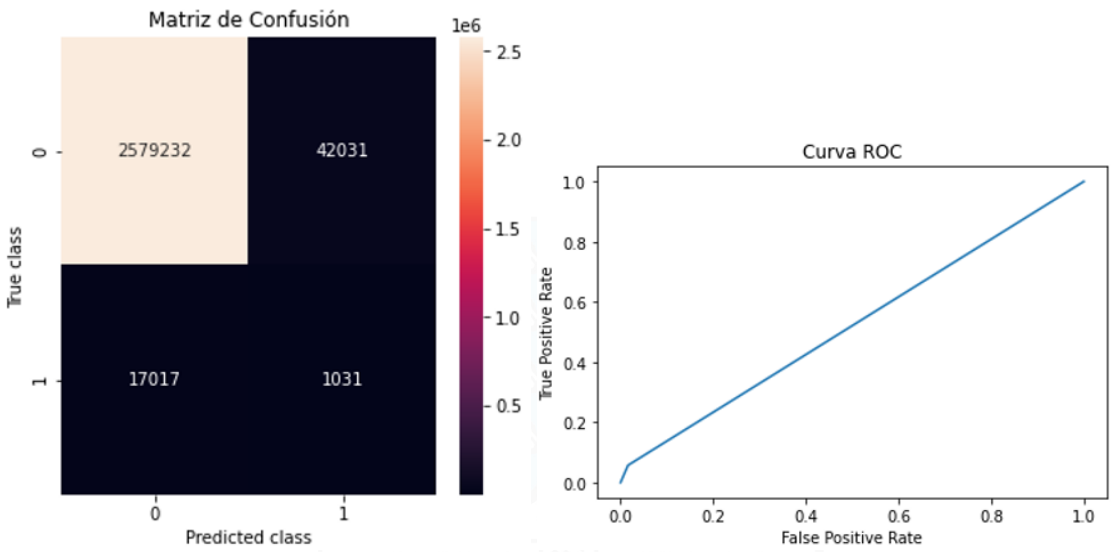


Figura A.38: Bosques Aleatorios, reducción 1, variables nuevas

Fuente: Elaboración propia

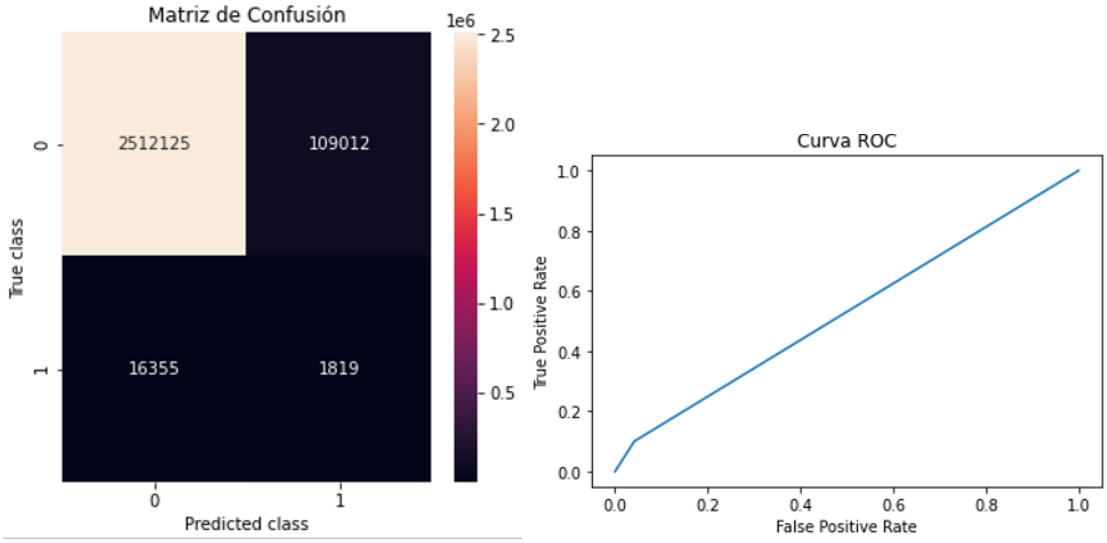


Figura A.39: Bosques Aleatorios, reducción 2, variables nuevas

Fuente: Elaboración propia

A.3.2.5. Regresión Logística: Categorías nuevas

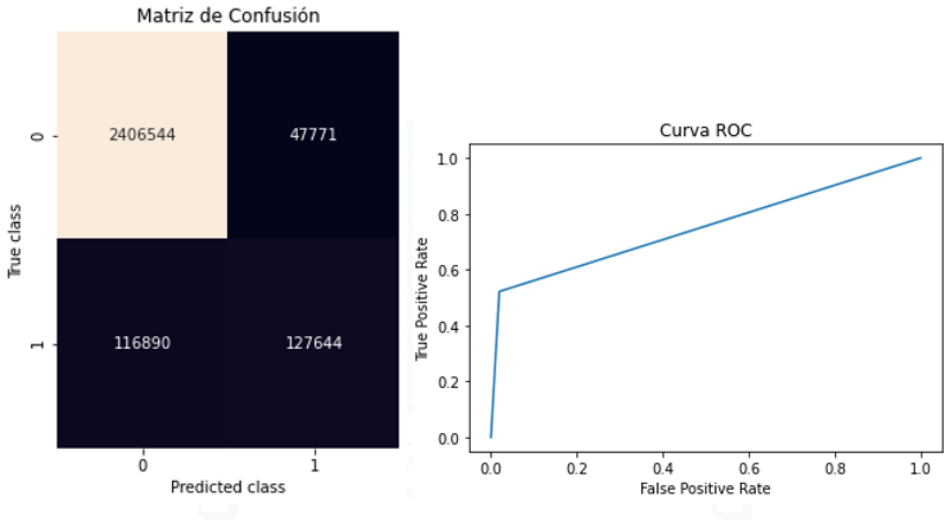


Figura A.40: Bosques Aleatorios: Categorías nuevas

Fuente: Elaboración propia

A.3.2.6. Bosques Aleatorios: Categorías nuevas

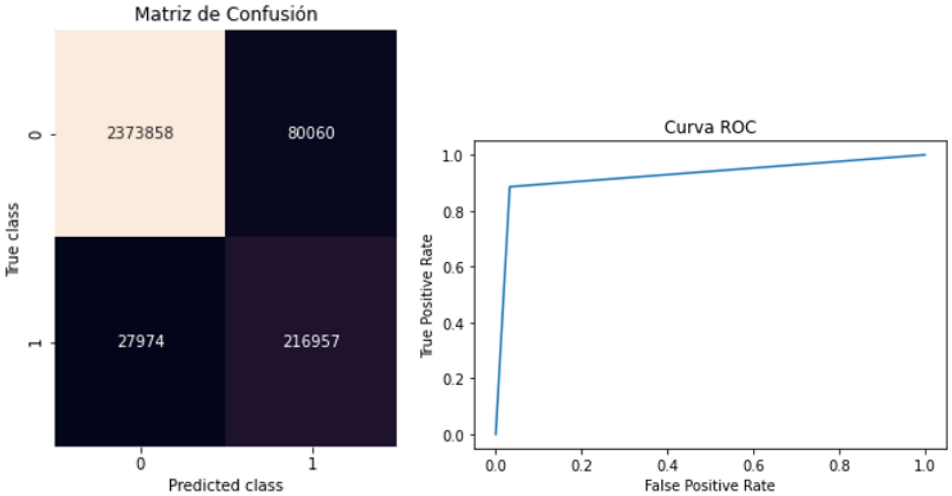


Figura A.41: Regresión Logística: Categorías nuevas

Fuente: Elaboración propia

A.4. Códigos programación en Python

A.4.1. Medición calidad de datos Egresos Hospitalarios

```
import mysql.connector
from datetime import datetime, timedelta

ahora1=datetime.now()
print("Hora inicio: ", ahora1)
con1=mysql.connector.connect(user="root",password="*****",
                             host="localhost",database="egresos")
query1=con1.cursor()
query1.execute("select trim(code) from diag1_2016")
r_query1=query1.fetchall()

r_diag1=[]
for diag_1 in r_query1:
    r_diag1.append(diag_1[0])
print("Códigos diag1: ", len(r_diag1))

con2=mysql.connector.connect(user="root",password="*****",
                             host="localhost",database="egresos")
query2=con1.cursor()
query2.execute("select trim(code) from diag2_2012")
r_query2=query2.fetchall()

r_diag2=[]
for diag_2 in r_query2:
    r_diag2.append(diag_2[0])
print("Códigos diag2: ", len(r_diag2))

con3=mysql.connector.connect(user="root",password="*****",
                             host="localhost",database="egresos")
query3=con1.cursor()
query3.execute("select trim(ocode),trim(ncode) from establecimientos_deis")
r_query3=query3.fetchall()
```

```
ocode_estab=[]
ncode_estab=[]
for estab in r_query3:
    ocode_estab.append(estab[0].zfill(6))
    ncode_estab.append(estab[1].zfill(6))

con4=mysql.connector.connect(user="root",password="*****",
                             host="localhost",database="egresos")
query4=con1.cursor()
query4.execute("select trim(codigo) from comunas")
r_query4=query4.fetchall()

r_com=[]
for com in r_query4:
    r_com.append(com[0].zfill(5))
print("Códigos comunas: ", len(r_com))

con5=mysql.connector.connect(user="root",password="*****",
                             host="localhost",database="egresos")
query5=con1.cursor()
query5.execute("select trim(code) from servicio_egreso")
r_query5=query5.fetchall()

r_ser_egr=[]
for segr in r_query5:
    r_ser_egr.append(segr[0].zfill(3))
print("Códigos servicio de egreso: ", len(r_ser_egr))


con6=mysql.connector.connect(user="root",password="*****",
                             host="localhost",database="egresos")
query6=con1.cursor()
query6.execute("select trim(code),trim(nombre_comuna),
region from comunas_por_region")
r_query6=query6.fetchall()
```

```
comunas_region=[]
for com_reg in r_query6:
    comunas_region.append(com_reg)
print("Códigos comunas_region: ", len(comunas_region))

######### LECTURA BASE EGRESOS HOSPITALARIO

egresos=open("egresos2011.txt")
base_datos=[]
for egreso in egresos:
    base_datos.append(egreso)

print("Egresos 2011: " , len(base_datos))
base_datos2 = [x.strip().split(',') for x in base_datos]

############ DOMINIOS CORRECTOS Y CONTADORES
inconsistente1=0
inconsistente2=0
inconsistente3=0
inconsistente4=0
inconsistente5=0
inconsistente6=0
inconsistente_fecha=0
inconsistente_fonasa=[]
l_inconsistente1=[]
l_inconsistente2=[]
l_inconsistente3=[]
l_inconsistente4=[]
l_inconsistente5=[]
l_inconsistente6=[]
l_fecha=[]
l_region=[]
diag2nulos=0
var_sersalud=["1","2","3","4","5","6","7","8","9","10","11","12","13",
               "14","15","16","17","18","19","20","21","22","23","24",
```

```
              "25","26","28","29","33"]
#var_estab=[] SQL tabla establecimientos_deis
var_sexo=["1","2","3","9"]
#var_edad edad < 120
var_previ1=["1","2","3","4","5","6","7"]   #2011-2013
var_previ2=["1","2","3","5","6","7","9"]    #2014-2017
var_previ3=["1","2","3","4","5","96","99"]  #2018
var_benef1=["0","1","2","3","4"]    #2011
var_benef2=["A","B","C","D", " "]    #2012-2018
var_modal=["0","1","2"]
#var_comuna=[]    SQL tabla comunas
#var_serc_egr=[] SQL tabla servicio egreso
#var_d_estad=[]    d_estad < X
#var_diag1=[] SQL tabla diag1_2016
#var_diag2=[] SQL tabla diag2_2012
var_cond_egr=["1","2"]
var_intervq=["1","2"]
var_region=["1","2","3","4","5","6","7","8","9","10","11","12","13","14","15"]
#var_servres=[] = servicio salud
id_malos=[]

########## EVALUANDO DOMINIO CORRECTO
c_sersalud=0
c_estab=0
c_sexo=0
c_edad=0
c_previ=0
c_benef=0
c_modal=0
c_comuna=0
c_seregreso=0
c_destad=0
c_diag1=0
c_diag2=0
c_intvq=0
c_condegr=0
```

```
c_region=0
for linea in base_datos2:
    num=linea[0]
    ser_salud=linea[1]
    estab=linea[2]
    sexo=linea[3]
    edad=linea[4]
    previ=linea[5]
    benef=linea[6]
    modal=linea[7]
    comuna=linea[8].zfill(5)
    f_egr=datetime.strptime(linea[9], '%Y-%m-%d')
    serc_egr=linea[10].zfill(3)
    d_estad=linea[11]
    diag1=linea[12]
    diag2=linea[13]
    cond_egr=linea[14]
    interv_q=linea[15]
    region=linea[16]
    serv_res=linea[17]
    stat=linea[18]
    cr=linea[19]
    f_ing=f_egr-timedelta(float(d_estad))

    if ser_salud not in var_sersalud:
        c_sersalud+=1
##        if num not in id_malos:
##            id_malos.append([num])
##
    if estab not in ocode_estab:
        c_estab+=1
        if estab not in id_malos:
            id_malos.append(estab)

    if sexo not in var_sexo:
        c_sexo+=1
```

```
##        if num not in id_malos:
##            id_malos.append([num])
##
    if previ not in var_previ1:
        c_previ+=1

##        if num not in id_malos:
##            id_malos.append([num])

    if benef not in var_benef1:
        c_benef+=1
##        if num not in id_malos:
##            id_malos.append(num)
##            print("ID:", num, "BENEF:",benef)

    if modal not in var_modal:
        c_modal+=1
##        if num not in id_malos:
##            id_malos.append(num)


    if comuna not in r_com:
        c_comuna+=1
##        if num not in id_malos:
##            id_malos.append([num])

    if serc_egr not in r_ser_egr:
        c_seregreso+=1
##        if num not in id_malos:
##            id_malos.append([num])

    if int(d_estad) > 150 :
        c_destad+=1
##        if num not in id_malos:
##            id_malos.append(num)
```

```
    if diag1 not in r_diag1:
        c_diag1+=1
##        if num not in id_malos:
##            id_malos.append([num])

    if diag2 not in r_diag2:
        c_diag2+=1
##        if num not in id_malos:
##            id_malos.append(num)

    if diag2 == "NULL":
        diag2nulos+=1

    if interv_q not in var_intervq:
        c_intvq+=1
##        if num not in id_malos:
##            id_malos.append([num])

    if cond_egr not in var_cond_egr:
        c_condegr+=1
##        if num not in id_malos:
##            id_malos.append([num])

    if region not in var_region:
        c_region+=1
##        if num not in id_malos:
##            id_malos.append([num])
    if int(edad) > 120:
        c_edad+=1

###### EVALUANDO INCONSISTENCIA FONASA

###### LOS QUE SON FONASA: PREVI=1, MODAL=1-2, BENEF=1-2-3-4 ó A,B,C,D
####si es fonasa, el benef ni el modal deben ser cero
    if previ =="1" and (benef=="0" or modal=="0"):
        inconsistente1+=1
```

```
        l_inconsistente1.append([previ,benef,modal])
        if num not in inconsistente_fonasa:
            inconsistente_fonasa.append(num)
##        print("1/ PREVI: ",previ,"BENEF: ",benef,"MODAL: ",modal, "ID: ", num)

#### si no es fonasa, el benef y modal deben ser cero
    if previ !="1" and (benef!="0" or modal !="0"):
        inconsistente2+=1
        l_inconsistente2.append([previ,benef,modal])
        if num not in inconsistente_fonasa:
            inconsistente_fonasa.append(num)
##        print("2/ PREVI: ",previ,"BENEF: ",benef,"MODAL: ",modal, "ID: ", num)

##### si la modalidad es 1 o 2, soy fonasa, debería tener benef 1-2-3-4
    if modal !="0" and (previ!="1" or benef =="0"):
        inconsistente3+=1
        l_inconsistente3.append([previ,benef,modal])
        if num not in inconsistente_fonasa:
            inconsistente_fonasa.append(num)
##        print("3/ PREVI: ",previ,"BENEF: ",benef,"MODAL: ",modal, "ID: ", num)

##### si la modalidad es 0, no soy fonasa, debería tener benef 0
    if modal =="0" and (previ=="1" or benef !="0"):
        inconsistente4+=1
        l_inconsistente4.append([previ,benef,modal])
        if num not in inconsistente_fonasa:
            inconsistente_fonasa.append(num)
##        print("4/ PREVI: ",previ,"BENEF: ",benef,"MODAL: ",modal, "ID: ", num)

#### si el benef es 0, no soy fonasa, la modalidad debe ser 0
    if benef =="0" and (previ=="1" or modal!="0"):
        inconsistente5+=1
        l_inconsistente5.append([previ,benef,modal])
        if num not in inconsistente_fonasa:
            inconsistente_fonasa.append(num)
##        print("5/ PREVI: ",previ,"BENEF: ",benef,"MODAL: ",modal, "ID: ", num)
```

```
#### si el benef es 1-2-3-4, soy fonasa, la modalidad debe ser 1-2
    if benef !="0" and (previ!="1" or modal=="0"):
        inconsistente6+=1
        l_inconsistente6.append([previ,benef,modal])
        if num not in inconsistente_fonasa:
            inconsistente_fonasa.append(num)
##        print("6/ PREVI: ",previ,"BENEF: ",benef,"MODAL: ",modal, "ID: ", num)

    if f_egr.year != 2011:
            inconsistente_fecha+=1
            l_fecha.append([num,f_egr])

print("Inconsistencia")
ahora1=datetime.now()
print("Hora: ", ahora1)

################################## INCONSISTENTE REGION RESPECTO A LA COMUNA
inconsistente_region=0
inconsistente_comuna=0

for num,ser_salud,estab,sexo,edad,previ,benef,modal,comuna,f_egr,serc_egr,d_estad,
diag1,diag2,cond_egr,interv_q,region,serv_res,stat,cr in base_datos2:
    #print("ID revisado",num)
    for code,name,region_com in comunas_region:
       # print("Ieeh:", comuna.zfill(5), region, code.zfill(5), region_com)
        if comuna.zfill(5) == code.zfill(5):
            #print("Comunas iguales check", region, region_com)
            if region != region_com:
                inconsistente_region+=1
                l_region.append([region, region_com])
                #print(comuna.zfill, region, region_com)
                #print("inconsitente",inconsistente_region)
            break

fin_diag2=c_diag2-diag2nulos
```

```
ahora1=datetime.now()
print("Hora: ", ahora1)
print("--------------- CONTADORES DE REGISTROS MALOS ---------------")
print("AÑO 2011")
print("Contador servicio salud: ",str(c_sersalud))
print("Contador establecimientos: ", str(c_estab))
print("Contador sexo :", str(c_sexo))
print("Contador edad :", str(c_edad))
print("Contador previ:", str(c_previ))
print("Contador benef :",str(c_benef))
print("Contador modal:", str(c_modal))
print("Contador comunas:", str(c_comuna))
print("Contador servicio egreso:", str(c_seregreso))
print("Contador dias estadia:", str(c_destad))
print("Contador diag1 :", str(c_diag1))
print("Contador c_diag2:", str(c_diag2))
print("Contador diag2 nulos:", str(diag2nulos))
print("Contador diag2 malos:", str(fin_diag2))
print("Contador intervencion quirurgica:", str(c_intvq))
print("Contador condicion egreso:", str(c_condegr))
print("Contador region:",str(c_region))

print("--------- CONTADORES DE REGISTROS INCONSISTENTES ----------")
print("Inconsistentes1: ", str(inconsistente1))
print("Inconsistentes2: ", str(inconsistente2))
print("Inconsistentes3: ", str(inconsistente3))
print("Inconsistentes4: ", str(inconsistente4))
print("Inconsistentes5: ", str(inconsistente5))
print("Inconsistentes6: ", str(inconsistente6))
print("Inconsistente fonasa: ", len(inconsistente_fonasa))
print("Inconsistente region: ", str(inconsistente_region))
print("Inconsistente fecha:", str(inconsistente_fecha))

idmalos=open("idmalos.txt",'w')
idmalos.write("Establecimientos malos: "+ '\n')
for i in id_malos:
```

```
    idmalos.write(str(i)+'\n')

idmalos.write("Inconsiste fecha: "+ '\n')
for i in l_fecha:
    idmalos.write(str(i)+'\n')

idmalos.write("Inconsiste region: "+ '\n')
for i in l_region:
    idmalos.write(str(i)+'\n')

idmalos.write("Inconsistente FONASA: "+ '\n')
for i in inconsistente_fonasa:
    idmalos.write(str(i)+'\n')

egresos.close()
idmalos.close()
ahora1=datetime.now()
print("Hora: ", ahora1)
```

A.4.2. SMOTE Regresión logística

```
import pandas as pd
from sklearn.model_selection import train_test_split
import matplotlib.pyplot as plt
import seaborn as sns
from datetime import datetime
from sklearn import metrics
from sklearn.metrics import classification_report,confusion_matrix,roc_auc_score
from sklearn.linear_model import LogisticRegression
from collections import Counter
from imblearn.over_sampling import SMOTE

ahora=datetime.now()
print("Hora inicio: ", ahora)
data=pd.read_csv('StausbergFINAL.csv',sep=',')
```

```
                ,dtype={'ano':'int32','sexo1':'int32','sexo2':'int32',
                'edad':'int32','previ1A':'int32','previ1B':'int32',
                'previ1C':'int32','previ1D':'int32','previ2':'int32',
                'previ3':'int32','f_ing':'int32','region':'int32',
                'poblacion':'float64','superficie':'float64','densidad':'float64',
                'pobreza':'float64','farma':'int32','almacen':'int32',
                'estabfarma':'int32','primario':'int32','secundario':'int32',
                'terciario':'int32','secter':'int32','cie':'int32'})

dataframe=pd.DataFrame(data)

print('Transform the dataset al 0.1')
# transform the dataset
X=dataframe.drop(['cie','poblacion','superficie','farma',
                  'almacen','terciario','secundario'],axis=1)
y=dataframe['cie']
oversample = SMOTE(sampling_strategy=0.1)
Xt, yt = oversample.fit_resample(X, y)

# summarize the new class distribution
counter = Counter(yt)
print(counter)

# Model
Xt_train,Xt_test,yt_train,yt_test=train_test_split(Xt,yt,test_size=0.2)
model=LogisticRegression(max_iter=2000)
result=model.fit(Xt_train,yt_train)
yt_pred=result.predict(Xt_test)

yt_pred2=[]
for i in yt_pred:
    if i > 0.5:
        yt_pred2.append(1)
    else:
        yt_pred2.append(0)
```

```
# Metrics
print (classification_report(yt_test, yt_pred2))

# Plot Confusion matrix
conf_matrix = confusion_matrix(yt_test, yt_pred2)
plt.figure(figsize=(5, 5))
sns.heatmap(conf_matrix, annot=True, fmt="d");
plt.title("Matriz de Confusión")
plt.ylabel('True class')
plt.xlabel('Predicted class')
plt.show()

#  Plot ROC Curve
roc_auc=roc_auc_score(yt_test,yt_pred2)
print('ROC-AUC: ', roc_auc)

fpr, tpr, thresholds = metrics.roc_curve(yt_test, yt_pred2, pos_label=0)
plt.plot(tpr,fpr)
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title("Curva ROC")
plt.show()

# Get importance
importance = model.coef_[0]

# Summarize feature importance
for i,v in enumerate(importance):
  print('Feature: %0d, Score: %.5f' % (i,v))

# Plot feature importance
plt.barh([x for x in range(len(importance))], importance)
plt.barh(X.columns, importance)
plt.xlabel('Puntaje de la importancia de la variable')
plt.ylabel('Variables')
plt.title("Visualización de la importancia de las variables")
```

```
plt.show()

ahora=datetime.now()
print("Hora: ", ahora)
```

A.4.3. SMOTE Bosques Aleatorios

```
import pandas as pd
from datetime import datetime
from sklearn.model_selection import train_test_split
import matplotlib.pyplot as plt
import seaborn as sns
#from rfpimp import permutation_importances
from sklearn.ensemble import RandomForestClassifier
from sklearn import metrics
from sklearn.metrics import classification_report,confusion_matrix,roc_auc_score
from collections import Counter
from imblearn.over_sampling import SMOTE

ahora=datetime.now()
print("Hora inicio: ", ahora)
data=pd.read_csv('StausbergFINAL.csv',sep=',
                ,dtype={'ano':'int32','sexo1':'int32','sexo2':'int32',
                'edad':'int32','previ1A':'int32','previ1B':'int32',
                'previ1C':'int32','previ1D':'int32','previ2':'int32',
                'previ3':'int32','f_ing':'int32','region':'int32',
                'poblacion':'float64','superficie':'float64','densidad':'float64',
                'pobreza':'float64','farma':'int32','almacen':'int32',
                'estabfarma':'int32','primario':'int32','secundario':'int32',
                'terciario':'int32','secter':'int32','cie':'int32'})

dataframe=pd.DataFrame(data)

print('Transform the dataset al 0.1')
# transform the dataset
oversample = SMOTE(sampling_strategy=0.1)
```

```
Xt, yt = oversample.fit_resample(X, y)

# summarize the new class distribution
counter = Counter(yt)
print(counter)

# Model
Xt_train,Xt_test,yt_train,yt_test=train_test_split(Xt,yt,test_size=0.2)
model=RandomForestClassifier(n_estimators=100,class_weight='balanced',
oob_score=True)
result=model.fit(Xt_train,yt_train)
yt_pred=result.predict(Xt_test)
print (classification_report(yt_test, yt_pred))

# Plot confusion matrix
conf_matrix = confusion_matrix(yt_test, yt_pred)
plt.figure(figsize=(5, 5))
sns.heatmap(conf_matrix, annot=True, fmt="d");
plt.title("Matriz de Confusión")
plt.ylabel('True class')
plt.xlabel('Predicted class')
plt.show()

# Plot ROC Curve
roc_auc=roc_auc_score(yt_test,yt_pred)
print('ROC-AUC: ', roc_auc)

fpr, tpr, thresholds = metrics.roc_curve(yt_test, yt_pred, pos_label=0)
plt.plot(tpr,fpr)
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title("Curva ROC")
plt.show()

# get importance
importance = model.feature_importances_
```

```
# summarize feature importance
for i,v in enumerate(importance):
  print('Feature: %0d, Score: %.5f' % (i,v))

# plot feature importance
plt.barh([x for x in range(len(importance))], importance)
plt.barh(X.columns, importance)
plt.show()

print('OOB score: ',model.oob_score_)
ahora=datetime.now()
print("Hora fin: ", ahora)
```