Universidad ORT Uruguay

Facultad de Ingeniería

Evaluación de algoritmos de aprendizaje automático en la predicción de carnicerías y locales de venta al público infractores en el mercado uruguayo de carnes y derivados

Entregado como requisito para la obtención del título de Maestría en Big Data

Gastón Bernheim – 183884

Tutor: Andrés Ferragut

2021

Declaración de autoría

Yo, Gastón Bernheim, declaro que el trabajo que se presenta en esta obra es de mi propia mano. Puedo asegurar que:

- La obra fue producida en su totalidad mientras realizaba el Trabajo Final de la Maestría en Big Data;
- Cuando he consultado el trabajo publicado por otros, lo he atribuido con claridad;
- Cuando he citado obras de otros, he indicado las fuentes. Con excepción de estas citas, la obra es enteramente mía;
- En la obra, he acusado recibo de las ayudas recibidas;
- Cuando la obra se basa en trabajo realizado conjuntamente con otros, he explicado claramente qué fue contribuido por otros, y qué fue contribuido por mí;
- Ninguna parte de este trabajo ha sido publicada previamente a su entrega, excepto donde se han realizado las aclaraciones correspondientes.

Gastón Bernheim Portela

11 de noviembre de 2021

Abstract

El Instituto Nacional de Carnes tiene el cometido de garantizar la calidad e inocuidad en el mercado de carnes y sus derivados. En el mercado doméstico operan 2.500 carnicerías y 460 vehículos habilitados en todo el país. Dados los recursos inspectivos con los que cuenta el país cada establecimiento es visitado, en promedio, una vez por año. De esta forma, la identificación de potenciales establecimientos infractores es un aspecto crucial para minimizar los impactos negativos en el bienestar de la población y costos económicos asociados.

En este trabajo se desarrollan modelos de clasificación mediante técnicas de aprendizaje automático supervisado para predecir establecimientos infractores. Para ello, se utiliza un conjunto de datos generado que incluye características comerciales de 2.374 puntos de venta del país. Los modelos estimados se basan en regresiones logísticas, redes neuronales artificiales y árboles de decisión.

El análisis realizado permite comprender la relevancia de determinadas características comerciales y su impacto en la probabilidad de que un establecimiento habilitado para la venta de carne y derivados en Uruguay sea considerado infractor. Se destaca la incidencia de variables como la ubicación, el tipo y modalidad de carnicería así como la venta de determinados productos específicos. No obstante, la capacidad predictiva alcanzada por los modelos es insuficiente para clasificar establecimientos infractores de manera certera. De esta forma, se realiza una propuesta de muestreo con el objetivo de maximizar la probabilidad de detectar establecimientos infractores.

Palabras clave

Instituto Nacional de Carnes; carnicerías; locales de venta al consumidor; inocuidad, infractores; aprendizaje automático; aprendizaje supervisado; clasificación; regresión logística; red neuronal artificial; árbol de decisión.

Listado de abreviaturas

ADIFU	Asociación de la Industria Frigorífica del Uruguay
APFMI	Asociación de Plantas de Faena Mercado Interno
ARU	Asociación Rural del Uruguay
CADA	Comisión de Administración del Abasto
CAF	Cooperativas Agrarias Federadas
CFSA	China National Center for Food Safety Risk Assessment
CIF	Cámara de la Industria Frigorífica
CNFR	Comisión Nacional de Fomento Rural
EPC	Equivalente Peso Canal
ETA	Enfermedades Transmitidas por Alimentos
FDMRS	Foodborne Disease Monitoring and Reporting System
FINDER	Foodborne Illness Detector in Real time
FOB	Free on Board
FR	Federación Rural del Uruguay
GAF	Gerencia de Administración y Finanzas
GAL	Gerencia de Asuntos Legales
GCL	Gerencia de Calidad
GCN	Gerencia de Conocimiento
GGN	Gerencia General
GIF	Gerencia de Información
GMI	Gerencia de Mercado Interno
GMK	Gerencia de Marketing
GTI	Gerencia de Tecnologías de la Información
INAC	Instituto Nacional de Carnes
MGAP	Ministerio de Ganadería, Agricultura y Pesca

MIEM	Ministerio de Industria, Energía y Minería
OPS	Organización Panamericana de la Salud
PSD	Presidencia
REGANTEC	Registro de Antecedentes
RUNEC	Registro Único Nacional de Empresas Cárnicas
SRGA	Sistema de Registro y Gestión del Abasto

Índice

Índice de tablas

Índice de ilustraciones

1. Introducción

1.1. Creación del Instituto Nacional de Carnes, funciones y financiamiento

En 1967 existían dos grandes organismos en el ámbito de la carne: el Instituto Nacional de Carnes (INAC) y la Comisión de Administración del Abasto (CADA). El primero dirigido básicamente a la exportación y el segundo al mercado interno. En 1984, el Decreto - Ley 15.605 [1] unifica ambos organismos y crea el INAC como persona pública no estatal con el objetivo de proponer, asesorar y ejecutar la Política Nacional de Carnes, cuya determinación corresponde al Poder Ejecutivo.

Dicho decreto establece que el INAC debe promover, regular, coordinar y vigilar las actividades de producción, transformación, comercialización, almacenamiento y transporte de carnes, menudencias, productos y subproductos cárnicos.

Para dar cumplimiento a dichas actividades, el INAC se financia principalmente a través de gravámenes a la comercialización de carnes en el mercado doméstico y externo.

Se entiende por mercado doméstico a la venta de carnes, menudencias y subproductos provenientes de plantas de faena o de la importación que se comercialicen en el mercado interno. Estas actividades están gravadas a una tasa del 0,7% (cero coma siete por ciento) del precio de venta.

La comercialización en el mercado externo abarca las exportaciones de animales bovinos y ovinos en pie, así como las exportaciones de carne, menudencias, subproductos y productos elaborados en base a carnes y subproductos. Estas actividades están gravadas a una tasa del 0,6% (cero coma seis por ciento) del precio de venta *FOB* neto.

1.2. Dirección del INAC y organización interna

El INAC es dirigido por una Junta de ocho miembros que se reúne semanalmente y está integrada por dos delegados del Poder Ejecutivo, tres representantes de los productores y tres de la industria frigorífica.

Los delegados del Poder Ejecutivo ejercen las funciones de Presidente y Vicepresidente del instituto. El Presidente es propuesto por el Ministerio de Ganadería, Agricultura y Pesca (MGAP) y el Vicepresidente por el Ministerio de Industria, Energía y Minería (MIEM). En cuanto a los representantes de los productores (sector primario), uno lo propone la Asociación Rural del Uruguay (ARU), uno la Comisión Nacional de Fomento Rural (CNFR) y Cooperativas Agrarias Federadas (CAF) y uno la Federación Rural del Uruguay (FR). De los tres representantes de la industria frigorífica (sector industrial), uno es propuesto por la Asociación de la Industria Frigorífica del Uruguay (ADIFU), uno por la Asociación de Plantas de Faena Mercado Interno (APFMI) y uno por la Cámara de la Industria Frigorífica (CIF).

El Presidente del INAC es quien preside y convoca la Junta con el fin de proponer, consultar y validar las estrategias generales del instituto. Una vez validadas, el INAC destina recursos materiales y de personal para el cumplimiento de sus objetivos. En este sentido, vale la pena destacar el doble rol que ejerce el Presidente del INAC: uno político ante la Junta y uno como administrador de los recursos del instituto.

En lo que respecta a la organización interna, el INAC se estructura en cuatro áreas:

1. Presidencia: integrada por el Presidente y el Vicepresidente cuya función principal es dirigir el instituto.
2. Gerencia General: encargada de dialogar e intercambiar con el resto de las áreas con el fin de facilitar el cumplimiento de las metas establecidas.
3. Área de Negocios: integrada por cinco gerencias cuyo trabajo está destinado al cumplimiento de las funciones que la ley establece. A continuación, se mencionan las gerencias que la integran y las actividades que desempeñan.
    - Gerencia de Calidad (GCL): control de calidad de las exportaciones; ingeniería y aprobación de diseño de las plantas frigoríficas.

- Gerencia de Mercado Interno (GMI): promoción de la formalización y control de la cadena de suministro local.
- Gerencia de Información (GIF): generación y análisis de estadísticas; monitoreo del desempeño financiero de las empresas frigoríficas.
- Gerencia de Marketing (GMK): actividades de promoción y marketing; acceso e inteligencia de mercado.
- Gerencia de Conocimiento (GCN): investigación, desarrollo e innovación.

4. Área de Operaciones: integrada por tres gerencias destinadas a proveer infraestructura y soporte a todo el instituto. Éstas son la Gerencia de Tecnologías de la Información (GTI), Gerencia de Asuntos Legales (GAL) y Gerencia de Administración y Finanzas (GAF).

La siguiente ilustración resume la organización interna del INAC.

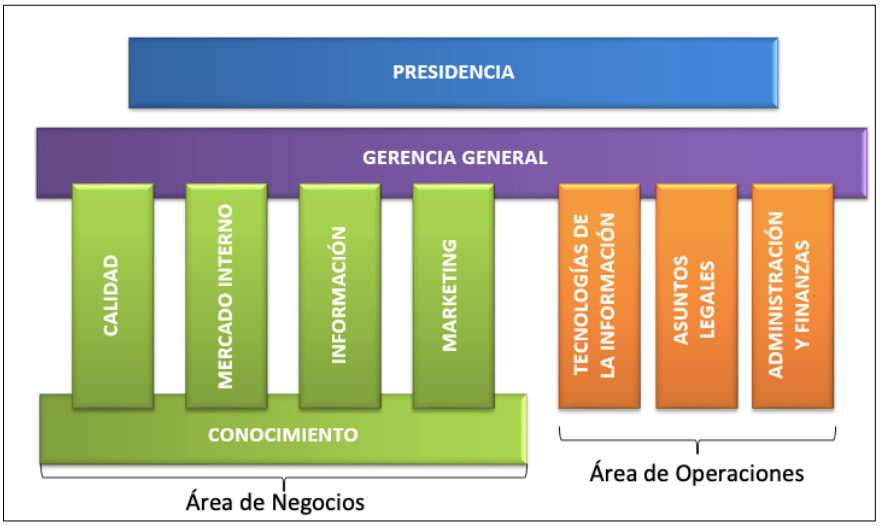


Ilustración 1. Organización interna del INAC

1.3. El rol del INAC en el mercado doméstico y su importancia[1]

El mercado doméstico de proteína animal es importante por dos razones:

1. Desde una perspectiva de salud pública, porque estas proteínas son un componente fundamental de la dieta, con un alto consumo per cápita (entre 2015 y 2020 el consumo promedio fue de 90,6 Kg EPC[2]/persona/año);

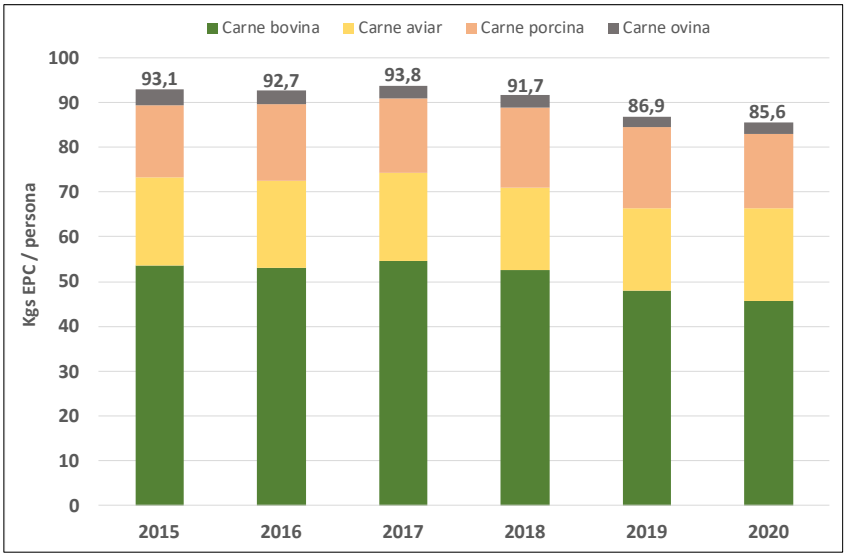


Ilustración 2. Consumo per cápita de carnes

2. Desde una perspectiva económica, porque las cadenas productivas precisan de este mercado para lograr su sostenibilidad en el largo plazo. En términos de volumen, Uruguay exporta el 75% de su producción de carne bovina y el 25% restante se vuelca al mercado interno. Esto convierte al mercado interno en el segundo mercado de destino para la carne bovina uruguaya.

Como ya fue mencionado, el INAC tiene el cometido de apoyar al Poder Ejecutivo en la ejecución de las políticas nacionales. Dichas políticas apuntan a garantizar la calidad e inocuidad del producto, asegurando el abastecimiento a los distintos sectores de la

[1] Este capítulo se basa en [2]

[2] De acuerdo a [3], "El índice equivalente canal es utilizado por todos los países, para la elaboración y publicación de datos referidos a la producción de carne. La utilización de este coeficiente permite convertir el peso de la carne que se comercializa, al peso de las canales que fue necesario procesar para obtenerla (…)".

población y su implementación en un marco de justa competencia entre los agentes de las cadenas de producción y distribución.

El mercado interno consta de más de 4.000 actores que para operar legalmente dentro del circuito comercial de carnes y sus derivados deben estar registrados en el INAC. De esta forma, el instituto se encarga de la habilitación, registro y control de medios de transporte, carnicerías y locales de venta al consumidor así como la determinación, imposición y ejecución de las sanciones por violación a las normas legales y reglamentarias en materia de faena y comercialización interna y externa.

En lo que respecta al control, las tareas del instituto podrían categorizarse en distintos niveles de acción. Un primer nivel de acción refiere a la represión de conductas ilegales como la faena antirreglamentaria. En este aspecto, el INAC realiza tareas de fiscalización conjunta con el Ministerio del Interior y las Intendencias Departamentales en todo el territorio nacional. En general, la carne procedente de esos circuitos ilegales o clandestinos termina comercializándose en puntos minoristas de pueblos y ciudades de menor porte del interior del país.

El segundo nivel de acción viene dado por el control de la justa competencia dentro de las cadenas de producción y distribución. En este nivel, el foco está puesto en el control de establecimientos que carecen de la habilitación correspondiente o donde la misma presenta algún tipo de limitación. Los productos que provienen de estos establecimientos deberán ser transportados para luego ser vendidos en puntos minoristas, eslabones (transporte y carnicerías/locales de venta al consumidor) donde el INAC tiene acceso para fiscalizar.

El tercer nivel de acción representa el control de la evasión, aún dentro de los circuitos comerciales relativamente formalizados.

En el cuarto nivel de acción se ubica el control de la inocuidad en los ámbitos donde el instituto tiene competencia. Entre ellos se destacan carnicerías de Montevideo, carnicerías del interior (con algunas restricciones) y el transporte de carne a nivel nacional. En el último tiempo, dichas restricciones se han ido superando producto del fortalecimiento de las potestades y competencias del INAC detallados más adelante.

1.4. La tarea inspectiva y recursos asociados

Como gerencia encargada de controlar la cadena de suministro local, la GMI realiza diariamente visitas y controles en establecimientos y vehículos habilitados del país. En general, los establecimientos visitados son carnicerías y locales de venta al consumidor.

Dichas visitas pueden coordinarse a solicitud de los titulares de establecimientos y vehículos, como es el caso de la habilitación inicial o renovación de habilitaciones preexistentes, o pueden surgir como parte de las inspecciones de rutina que realiza la gerencia con el fin de garantizar la inocuidad de los productos así como el cumplimiento de las normas y leyes establecidas.

Para ello, la GMI cuenta con un área de Habilitación e Inspección integrada por nueve inspectores. Éstos son asignados en equipos de tres personas que semanalmente planifican y realizan recorridas por todo el territorio nacional. Dos equipos se dedican a la tarea inspectiva – uno en Montevideo y uno en el interior del país – mientras que el tercero se dedica a la habilitación y renovación de habilitaciones.

Con el fin de cubrir todo el territorio nacional a lo largo del año, las tareas inspectivas se planean en base a radios de cobertura. A cada equipo inspectivo se le asignan uno o dos radios de cobertura por semana donde deberán visitar todas las carnicerías y locales de venta al consumidor que se encuentren dentro de ese radio. Los inspectores realizan las recorridas asignadas durante la mañana y en la tarde deben ingresar a los sistemas del INAC los resultados obtenidos. Estos resultados incluyen la actualización de datos asociados a los establecimientos y el ingreso de las actas realizadas por la constatación de incumplimientos a la normativa vigente.

En el caso de los vehículos, el INAC tiene la potestad para fiscalizar en la vía pública. Aquí se controla que la habilitación del vehículo se encuentre vigente y que pueda transportar los productos que lleva en ese momento. Con respecto a la habilitación y su renovación en medios de transporte, el instituto coordina una agenda en distintos puntos del país para que los vehículos se acerquen y realicen el trámite correspondiente.

1.5. Identificación del problema y necesidad de un nuevo modelo de inspección

Previamente se mencionó que en el mercado doméstico operan más de 4.000 actores, de los cuales 2.500 son carnicerías, pollerías y locales de venta al consumidor y 460 son vehículos. Esto implica que cada inspector debe controlar cerca de 330 establecimientos y/o vehículos por año. Además, el mismo cuerpo de inspectores debe atender las denuncias realizadas por la población, limitando en ocasiones la capacidad inspectiva de rutina. Como resultado, cada establecimiento habilitado en el país es inspeccionado en promedio una vez por año, haciendo de la asignación de recursos inspectivos en el territorio un aspecto crucial para el INAC.

Otro aspecto a tener en cuenta son los resultados obtenidos en las inspecciones. Por como se desarrolla la tarea inspectiva, no es posible concluir con relativa certidumbre cuál es el nivel de incumplimiento a las normas vigentes que existe en el mercado dómestico. Si bien cada inspección aporta información adicional, los resultados observados se encuentran sesgados por el hecho de haber decidido inspeccionar en determinado radio de cobertura y no en otro. En otras palabras, la elección del lugar a inspeccionar tiene implicancias en los resultados que serán relevados por los inspectores en el territorio. Tal es el caso de las inspecciones en localidades del interior, donde el primer establecimiento visitado puede ser tomado por sorpresa y en caso de operar con algún grado de incumplimiento éste será relevado por los inspectores. Sin embargo, una vez que los inspectores continúan sus visitas por la localidad las empresas del sector cárnico ya están al tanto de que el INAC está presente y la dinámica habitual del circuito comercial se ve alterada. Este tipo de situaciones son comunes en el interior del país y en menor medida en Montevideo y Canelones.

La conjunción de estos dos aspectos, recursos escasos y potenciales sesgos en los resultados observados, hacen pertinente un cambio en el modelo de inspección desde un modelo presencial - en el que debe cubrirse el territorio a lo largo del año - hacia un modelo de inspección por sistema basado en datos. En este nuevo modelo, la asignación de recursos inspectivos podría realizarse a través de diversas técnicas de muestreo, identificando zonas de riesgo en base al conocimiento de los inspectores y párametros

objetivos como el número de actores, características comerciales de los establecimientos, volúmen comercializado, antecedentes, entre otros.

El camino hacia un nuevo modelo de inspección comenzó hace unos años y fue impulsado por la promulgación de leyes como la Ley de Inocuidad y Transparencia Comercial [4] en el 2019 y la Ley de Urgente Consideración [5] en 2020. Dichos textos fortalecieron las potestades y competencias del INAC, obligando a readecuar el marco normativo del mercado doméstico. En este sentido, se desarrollaron nuevos decretos reglamentarios asociados al registro, distribución y comercialización de carnes y sus derivados que contienen atribuciones puntuales y concretas en el INAC. Como resultado de este proceso se crea el Registro Único Nacional de Empresas Cárnicas (RUNEC) y el Sistema de Registro y Gestión del Abasto (SRGA), pilares fundamentales para el desarrollo de un nuevo modelo de inspección por sistema.

El RUNEC es un registro que incluirá a todos los actores habilitados para operar en el mercado interno, mientras que el SRGA permitirá realizar el seguimiento de forma electrónica de todas las transacciones y movimientos en el circuito comercial de la carne y derivados en el mercado doméstico. Una vez que el SRGA se encuentre en funcionamiento, el instituto comenzará a recibir grandes volúmenes de datos (todas las transacciones y movimientos) que deberá procesar, filtrar y analizar para mejorar su entendimiento del circuito comercial de la carne en el mercado doméstico y desempeñar las tareas que la ley le mandata con mayor eficacia y eficiencia.

1.6. Aspectos a considerar en un nuevo modelo de inspección

El desarrollo de un nuevo modelo de inspección debe tener en cuenta los niveles de acción en los que el instituto debe oficiar como organismo de contralor detallados en el capítulo 1.3. Por un lado, la prevención y represión de conductas ilegales como las detalladas en los niveles de acción uno, dos y tres. Por otro lado, el control de la inocuidad en los ámbitos de competencia detallados en el nivel de acción cuatro. A continuación, se presenta un resumen de la bibliografía estudiada en cada uno de los casos.

1.6.1. Prevención y represión de conductas ilegales

En [6] se analiza el efecto de implementar estrategias de prevención diferenciadas en 56 puntos calientes de actividad criminal. Utilizando el número de denuncias realizadas como métrica, encuentran que en las zonas de tratamiento (en donde se aplicaron técnicas diferenciadas) las llamadas se reducen o crecen en menor medida que en las zonas de control para todos los tipos de delito analizados. Además, destacan la importancia de implementar un enfoque específico y focalizado en ciertos tipos de delitos y lugares.

Por su parte, en [7] se presenta una reseña detallada sobre la evolución de las técnicas estadísticas utilizadas en los últimos años para la predicción de delitos. Se introduce el concepto de *predictive policing* basado en técnicas de aprendizaje automático, su funcionamiento, tipos de delitos en los que se ha implementado y casos de uso como Crime Anticipation System (CAS) en Holanda, PreCobs en Alemania y Suiza y PredPol en Reino Unido y Estados Unidos. Con respecto a los algoritmos utilizados se destacan las redes neuronales artificiales (ANN por sus siglas en inglés), regresiones logísiticas y adaptaciones de modelos lineales como los modelos *time-space* y *near-repeat.* Sin embargo, los autores mencionan que por el momento no existen suficientes evaluaciones sobre la efectividad de las técnicas de *predictive policing* en la reducción de delitos y enfatizan en la necesidad de comunicar adecuadamente lo que se puede esperar de estos modelos.

En [8] se estima la probabilidad de ocurrencia de tres tipos de delitos a nivel espacial mediante la utilización de regresiones logísticas, ANN y un modelo sintético que combina

ambos algoritmos. Para ello, utilizan predictores asociados al tipo de delito, al entorno en donde ocurren los delitos, demográficos, socio económicos y de proximidad a determinados puntos de la ciudad. Según la métrica de performance que se evalúe, tanto la red neuronal artificial como el algoritmo sintético presentan buenos resultados. Sin embargo, se menciona que la regresión logística es el modelo que debe seleccionarse si lo que se busca es simplicidad e interpretabilidad.

Un trabajo similar al anterior es [9], donde se estima la probabilidad de ocurrencia de delitos a la propiedad a nivel espacial mediante *naive Bayes* (NB), ANN, redes neuronales convolucionales (CNN por sus siglas en inglés), *K-Nearest Neighbors* (KNN), *Support Vector Machines* (SVM) y *Random Forest* (RF). En este caso, el entrenamiento y validación de los modelos se realiza en distintos pueblos de una misma ciudad.

1.6.2. Control de la inocuidad

Uno de los aspectos centrales es el control de la incouidad. Su importancia radica en los potenciales efectos negativos que impone a la sociedad en términos de salud y costos económicos asociados [10] – [13]. En este sentido, la inocuidad es un aspecto a monitorear en todos los eslabones de las cadenas productoras de alimentos [14]. Para ello, los países del mundo demandan el cumplimiento de determinados estándares que se ven reflejados en las leyes y normas vigentes.

Uno de los principales problemas en términos de inocuidad viene dado por la aparición de brotes de enfermedades transmitidas por alimentos (ETA). La Organización Panamericana de la Salud (OPS) define un brote de ETA como *"un incidente en el que dos o más personas presentan una enfermedad semejante después de la ingestión de un mismo alimento y los análisis epidemiológicos apuntan al alimento como el origen de la enfermedad"*. No obstante, la determinación del alimento causante de un brote no es tarea sencilla [15].

En [16] y [17] se desarrollan modelos de clasificación basados en aprendizaje automático que permiten la identificación de potenciales brotes y los patógenos causantes de brotes de ETA. Ambos trabajos utilizan datos relevados por el *Foodborne Disease Monitoring*

and Reporting System (FDMRS) del *China National Center for Food Safety Risk Assessment (CFSA).*

Otra línea creciente de trabajos se basan en el análisis de datos no estructurados, específicamente textos, para el monitoreo en *near-real time* de infracciones en restaurantes e identificación de brotes de ETA. [18] - [21] utilizan publicaciones realizadas por los usuarios de *Twitter*, [22] – [24] reseñas del sitio web *Yelp* y [25] reseñas de productos comercializados en *Amazon.com*. En estos trabajos se destaca la ventaja de incorporar análisis de redes sociales y reseñas en los modelos de inspección.

En [26] se analiza el efecto de realizar inspecciones estandarizadas en restaurantes sobre el nivel de cumplimiento de la normativa vigente. Para ello, evalúan una serie de factores y generan un puntaje global por restaurante, constatando una mejora de este indicador en inspecciones sucesivas para aquellos restaurantes con menor puntaje inicial. Se concluye que la realización de inspecciones estandarizadas es un mecanismo útil para elevar el nivel general de cumplimiento de la normativa.

En [27] se evalúan modelos basados en regresiones logísiticas, RF y SVM con el objetivo de predecir la probabilidad de constatar una infracción en materia de inocuidad en la ciudad de Toronto, Canadá. Por último, en [28] y [29] se desarrollan modelos de aprendizaje automático con el objetivo de predecir establecimientos con mayor probabilidad de cometer una infracción. En el primer caso, el modelo de inspección identifica a los establecimientos infractores siete días antes en comparación con el modelo inspectivo de rutina. En el segundo caso, los restaurantes identificados como “riesgosos” por *FINDER* (*Foodborne Illness Detector in Real Time*) tuvieron 3,1 veces más probabilidad de ser categorizados como infractores que aquellos inspeccionados de manera tradicional.

2. Objetivo y alcance del trabajo

El presente trabajo tiene por objetivo evaluar la performance de modelos de aprendizaje automático supervisado en la predicción de carnicerías y locales de venta al público infractores en el mercado dómestico uruguayo de carnes y sus derivados.

A partir de una base de datos generada con características comerciales de 2.374 puntos de venta del país, se estiman modelos basados en los siguientes algoritmos de clasificación: regresión logística, redes neuronales artificales y árboles de decisión.

La variable de respuesta es *infractor*, variable categórica que toma valor 1 para los establecimientos que cometieron al menos una infracción en el período enero 2015 - agosto 2021 y 0 en caso contrario. Dentro de las infracciones consideradas se incluyen 49 tipos de infracción, abarcando desde faltas administrativas hasta la comercialización de productos sin origen.

Vale la pena destacar, que si bien la base de datos generada debe ser mejorada, este trabajo constituye una primera aproximación en la identificación de características comerciales que impactan en la probabilidad de que un establecimiento habilitado para la venta de carne y derivados en Uruguay sea considerado infractor.

Por último, en base a las estimaciones arrojadas por los modelos se realiza una propuesta de muestreo de establecimientos a inspeccionar que maximiza la probabilidad de encontrar puntos de venta infractores en el país.

3. Fuentes de información y generación del conjunto de datos

3.1. Fuentes de información

Como fue mencionado previamente, este trabajo se realiza utilizando datos con los que ya cuenta el instituto. En este sentido, se utilizan tres fuentes de información: Padronarios; RUNEC y Registro de Antecedentes (REGANTEC). A continuación se detallan cada una de las fuentes de información así como los datos relevados en cada caso.

3.1.1. Padronarios

Los padronarios son registros a nivel departamental de cáracter no oficial que lleva la GMI. En ellos se relevan datos asociados a la titularidad, ubicación, tipo, modalidad y habilitaciones de cada carnicería y local de venta al consumidor de carnes y derivados.

El siguiente cuadro muestra un resumen de los principales tipos, modalidades y habilitaciones especifícas que pueden tener los distintos puntos de venta.[3]

Tipo de carnicería	Modalidad de carnicería	Habilitaciones específicas (por local)	Especie	Ejemplo
Independiente	De corte	Sector de elaboración	Bovino, Ovino, Porcino, Aves, Conejo y De caza menor	Carnicería tradicional (Media canal + envasado)
		Sector de venta de productos no cárnicos		
		Sector de cocción		
		Sector de chacinados con fraccionamiento		
	De expendio	Sector de chacinados con fraccionamiento		Boutique de carnes
Autoservicio (Supermercado)	De corte	Sector de elaboración	Bovino, Ovino, Porcino, Aves, Conejo y De caza menor	Supermercado con sector carnicería
		Sector de cocción		
		Sector de chacinados con fraccionamiento		
	De expendio	Sector de chacinados con fraccionamiento		Supermercado con góndola de autoservicio
Pollería	De corte	Sector de cocción	Carne aviar	Pollería tradicional

Tabla 1. Tipo, modalidad y habilitaciones específicas por punto de venta.

Los padronarios son actualizados frecuentemente a partir de la información relevada por los equipos inspectivos en el territorio y cuentan con información valiosa de las características comerciales de los establecimientos. Si bien a partir de la implementación del RUNEC como registro oficial del INAC, se ha comenzado un proceso de migración

[3] En [30] se detallan algunas de las definiciones asociadas a las actividades de los distintos tipos y modalidades de las carnicerías.

de datos de la información relevada en los padronarios departamentales a dicho sistema, este proceso aún no ha finalizado. De esta forma, el RUNEC cuenta con información completa y actualizada sobre los establecimientos ubicados en Montevideo pero no así para el resto de los departamentos. Una vez que dicho registro cuente con toda la información a nivel país, los padronarios dejarán de ser utilizados.

A continuación, se listan los datos relevados a nivel de establecimiento en los padronarios departamentales:

- Titularidad: razón social, RUT, teléfono y nro. de habilitación (carpeta). Este último dato lo provee el INAC al momento de habilitar el establecimiento.
- Ubicación: dirección, localidad y departamento donde se encuentra el establecimiento.
- Tipo de carnicería: se identifica si el establecimiento es del tipo independiente, supermercado/autoservicio o pollería.
- Modalidad de carnicería: se identifica si el establecimiento trabaja en modalidad de corte o de expendio.
- Habilitaciones específicas: se identifican las habilitaciones con las que cuenta el establecimiento.
- Actividades realizadas: se identifican actividades desarrolladas por el establecimiento como la elaboración de productos y cocción de productos.
- Productos vendidos: se identifican productos comercializados por el establecimiento como venta de productos no cárnicos y venta de chacinados. En algunos casos se cuenta con el promedio de kgs comprados durante 2018-2019
- Última inspección realizada: se identifica el día, mes y año en que el equipo inspectivo realizo la última visita.

Vale la pena mencionar, que si bien los padronarios cuentan con la información "más actualizada" en lo que hace a datos de locales de venta al público, es un registro con un gran número de datos faltantes e inconsistencias, tema que será abordado más adelante.

3.1.2. RUNEC

El Registro Único Nacional de Empresas Cárnicas es el registro oficial que mantiene el INAC a través de GAL. Dicho registro es donde debe estar inscribirse cualquier persona física o jurídica que participe en actividades de la producción, industrialización, transformación, distribución, transporte, almacenamiento y/o comercialización de carnes y sus derivados.

Previo a la existencia del RUNEC, el INAC mantenía diversos registros incompletos, ineficientes y compartimentados. Esta situación ha cambiado a partir de la migración de datos mencionada previamente, pero aún quedan datos por fuera del nuevo sistema que deberán ser incoporados en el próximo tiempo. En este sentido, el INAC continúa trabajando a la interna y con las Intendencias Departamentales para su incorporación.

Este trabajo utilizó los datos ingresados en el RUNEC con el fin de complementar los datos faltantes en los padronarios departamentales. En algunos casos, el RUNEC permitió incorporar datos asociados a la titularidad, ubicación, tipo, modalidad y habilitaciones de los establecimientos. Aquellos establecimientos para los cuales no se logró obtener datos suficientes fueron eliminados de la base de datos. Este proceso se describe más adelante.

3.1.3. REGANTEC

El Registro de Antecedentes, al igual que RUNEC, es un registro a cargo de GAL. En REGANTEC se ingresan y se da seguimiento a las presuntas infracciones[4] que son constatadas tanto por gerencias del INAC como GMI y GCL, así como por otras instituciones como el Ministerio del Interior.

Este trabajo considera únicamente las presuntas infracciones realizadas por GMI correspondientes al período enero 2015 - agosto 2021. Como resultado se obtiene un listado con los siguientes datos:

- Fecha: día, mes y año en que se constató la presunta infracción.
- Ubicación: localidad y departamento donde se constató la presunta infracción.

[4] Se entiende por presuntas infracciones a aquellas condiciones y/o actividades identificadas por el equipo de inspectores que potencialmente podrían configurar una infracción.

- Infractor: nombre de la persona física o jurídica a la que se le imputa la presunta infracción.
- Tipo de documento y número de documento del infractor: cédula de identidad en el caso de personas físicas y RUT en el caso de personas jurídicas.
- Código y tipo de infracción: código interno y detalle del tipo de infracción constatada.

3.2. Generación del conjunto de datos

Una vez identificadas las fuentes de información y los datos a utilizar se decide generar una única base de datos. Dicha base de datos debe permitir vincular a las carnicerías y locales de venta al consumidor infractores con los datos asociados a las características comerciales de dichos establecimientos.

Idealmente, se deberían vincular las infracciones labradas por los equipos inspectivos con los establecimientos en donde se constataron potenciales actividades ilícitas. De esta forma, sería posible analizar la incidencia - en caso de existir - de las características comerciales de los establecimientos en la probabilidad de cometer una infracción. Sin embargo, la cualidad de infractor no se asocia a un establecimiento sino que se asigna a una persona física o jurídica, lo que complejiza dicha vinculación. Para abordar esta problemática y garantizar la privacidad de los datos se decide realizar la vinculación a través de un id anónimo creado a partir del RUT y departamento, únicas variables en común entre las fuentes de información.

Esta forma de vinculación introduce dos distorsiones que deben ser consideradas al analizar los resultados obtenidos. Por un lado, se excluyen las infracciones asignadas a personas físicas, infracciones que en determinados casos pueden estar asociadas a la actividad comercial de las carnicerías y locales de venta al consumidor. De esta forma, el número total de infracciones que surge del conjunto de datos podría estar subestimado.

Por el otro lado, existen empresas que operan más de un establecimiento con la misma personería jurídica. Las fuentes de información permiten identificar tanto a la empresa a la que le fue labrada un acta de infracción como el departamento en el que ocurrió. Sin embargo, si una empresa cuenta con más de un establecimiento en el mismo

departamento, como es el caso de determinadas cadenas de supermercados, no es posible identificar a cuál de todos los establecimientos le fue asignada la infracción. En este trabajo se decidió considerar como infractores a todos los establecimientos que una empresa infractora opera en los departamentos donde le fue labrada un acta de infracción. Esta decisión podría implicar una sobreestimación del número total de infractores en el conjunto de datos.

A continuación se describe el proceso de extracción, transformación y limpieza de los datos de cada fuente de información para su integración en un único conjunto de datos.

En primer lugar se trabajó sobre los padronarios a nivel departamental relevados por GMI. Se procesaron un total de 60 padronarios, de los cuales 24 corresponden a Montevideo, 19 a Canelones y 17 para el resto de los departamentos. Montevideo y Canelones cuentan con un padronario por radio de cobertura, mientras que el resto de los departamentos cuenta con un único padronario.

A partir de ellos se generó una tabla con 3.058 puntos de venta de carnes y sus derivados en el país. Del total de observaciones, 647 fueron eliminadas por estar catalogadas por GMI como “Cerrada” o “Dada de baja”, lo que implica que en dichos establecimientos no se constató actividad en las últimas visitas o que está operando otro tipo de negocio. Como resultado, se obtiene una tabla con 2.411 puntos de venta, 1.715 con RUT y 696 sin RUT, de las cuales 87 son del interior y 609 de Montevideo.

Cada punto de venta cuenta con un nro. de habilitación único que es registrado en los padronarios. Utilizando este nro. de habilitación se realizó una búsqueda en RUNEC para extraer el RUT de dichos locales, obteniéndolo para todas las carnicerías ubicadas en Montevideo y para 50 de las 87 carnicerías del interior. Las restantes 37 no fueron identificadas y por lo tanto se eliminaron. Al finalizar este proceso, se obtuvo una tabla con 2.374 puntos de venta y características comerciales. La tabla siguiente muestra la distribución de puntos de venta por departamento.

Departamento	Puntos de venta totales	Puntos de venta Cerrados / Dados de baja	Puntos de venta no identificados	Puntos de venta identificados	Porcentaje de puntos de venta identificados
Artigas	54	20	0	34	100%
Canelones	437	81	2	354	99%
Cerro Largo	114	33	5	76	94%
Colonia	164	24	3	137	98%
Durazno	76	15	1	60	98%
Flores	36	9	0	27	100%
Florida	111	24	2	85	98%
Lavalleja	93	29	5	59	92%
Maldonado	241	60	2	179	99%
Montevideo	738	127	0	611	100%
Paysandú	163	32	1	130	99%
Río Negro	59	11	0	48	100%
Rivera	103	35	8	60	88%
Rocha	130	43	0	87	100%
Salto	118	19	0	99	100%
San José	135	23	1	111	99%
Soriano	120	23	1	96	99%
Tacuarembó	95	16	2	77	97%
Treinta y Tres	71	23	4	44	92%
Total	**3.058**	**647**	**37**	**2.374**	**98%**

Tabla 2. Distribución de puntos de venta por departamento.

Para estos 2.374 puntos de venta se generaron variables categóricas que representan el tipo y modalidad de carnicería así como las habilitaciones específicas, actividades comerciales realizadas y productos vendidos. A modo de ejemplo, la variable tipo de carnicería puede tomar tres valores: “independiente”, “supermercado” y “pollería”, por lo que si una carnicería esta registrada en el padronario como independiente la variable “supermercado” y “pollería” toman valor 0 mientras que “independiente” toma valor 1.

Previamente se mencionó que el padronario es un registro no oficial y que su actualización es realizada cuando el equipo inspectivo visita el terreno. Como parte de este proceso se generan algunas inconsistencias en los datos relevados por lo que se decidió asignar el valor NA en aquellas variables para las cuales no se cuenta con datos.

En segundo lugar se identificó a aquellos establecimientos infractores a partir de los datos de REGANTEC. Para ello, se considera infractor a aquel establecimiento que le fue labrada un acta de infracción entre enero de 2015 y agosto de 2021.

La base de datos de infracciones contiene 49 tipos de infracciones que fueron labradas a 1.578 actores, de los cuales 912 son persons físicas y 666 empresas. Al considerar las infracciones a empresas, se observa que dichas infracciones fueron labradas a 406 empresas, lo que arroja un promedio de 1,6 infracciones por empresa. No obstante, vale la pena destacar que dichas infracciones no se distribuyen de manera homogénea entre las empresas. De las 406 empresas infractoras que surgen de REGANTEC, 190 se encuentran en la base de datos generada a partir de los padronarios y RUNEC. En el Anexo 1 se muestran los distintos tipos de infracciones incluidos en el análisis y se observa que más del 80% de las infracciones consideradas se concentran en 12 tipos de infracciones.

Una vez vinculadas ambas fuentes de información a través del id anónimo, se obtiene el conjunto final de datos a utilizar para la estimación de los modelos. Dicho conjunto e datos contiene 2.374 puntos de venta cuya estructura se muestra en la tabla siguiente.

Variable	Descripción
id	Número de identificación anónimo del punto de venta.
infractor	Toma valor 1 si el punto de venta es infractor y 0 en caso contrario.
localidad	Localidad donde se ubica el punto de venta.
departamento	Departamento donde se ubica el punto de venta.
independiente	Toma valor 1 si el punto de venta es de tipo independiente y 0 en caso contrario.
supermercado	Toma valor 1 si el punto de venta es de tipo supermercado y 0 en caso contrario.
polleria	Toma valor 1 si el punto de venta es de tipo pollería y 0 en caso contrario.
corte	Toma valor 1 si el punto de venta es de modalidad corte y 0 en caso contrario.
expendio	Toma valor 1 si el punto de venta es de modalidad expendio y 0 en caso contrario.
vende_no_carnicos	Toma valor 1 si el punto de venta vende productos no cárnicos y 0 en caso contrario.
vende_chacinados	Toma valor 1 si el punto de venta vende chacinados y 0 en caso contrario.
elabora_prod	Toma valor 1 si el punto de venta elabora productos y 0 en caso contrario.
hab_sec_elab	Toma valor 1 si el punto de venta cuenta con un sector de elaboración habilitado y 0 en caso contrario.
realiza_coccion	Toma valor 1 si el punto de venta realiza cocción y 0 en caso contrario.
compras_avg	Cantidad promedio de kgs de carnes y sus derivados comprados por el punto de venta en el período 2018-2019.
ult_visita	Fecha en que se realizó la última visita al punto de venta
dias_desde_30_7	Días transcurridos desde la última visita al punto de venta al 30 de julio de 2021.

Tabla 3. Estructura de la base de datos final.

Por último, en los modelos estimados mediante redes neuronales y árboles de decisión se incorporan variables a nivel departamental como número de habitantes, superficie medida en km^2, ingreso per cápita mensual y número de frigoríficos habilitados.

4. Algoritmos seleccionados

Este trabajo estima modelos de aprendizaje automático basados en tres algoritmos: regresiones logísticas, redes neuronales artificiales y árboles de decisión. A continuación, se describen brevemente los algoritmos seleccionados y su funcionamiento.

4.1. Regresión logística (RL)

La regresión logística es utilizada para predecir el valor de una variable categórica, ya sea de una o más categorías en función de un conjunto de variables independientes. A diferencia del análisis de regresión lineal estándar que modela directamente la variable dependiente o de respuesta, la regresión logísitica modela la probabilidad de que la variable dependiente pertenezca a cierta categoría. En este caso, se modela la probabilidad de pertenecer a la categoría infractor dado el conjunto de características comerciales.

A modo de ejemplo, la probabilidad de ser infractor dado el vector de características comerciales X se puede expresar de la siguiente forma,

$$\Pr(\text{ infractor} = \text{Sí} \mid X)$$

Los valores de $\Pr(\text{ infractor} = \text{Sí} \mid X)$, en adelante abreviada como $p(X)$, se encontrarán entre 0 y 1. De esta forma, para cualquier combinación de X se pueden realizar predicciones sobre la probabilidad de ser infractor. Una posible solución, es asignar la categoría infractor = Sí a aquellos establecimientos cuya $p(X) \geq 0.5$.

Para garantizar que todas las $p(X)$ estimadas se encuentren dentro del rango entre 0 y 1, la regresión logística utiliza la función logística,

$$p(X) = \frac{e^{\beta 0 + \beta 1 X}}{1 + e^{\beta 0 + \beta 1 X}} \quad , \tag{1.1}$$

donde β_0 es un número conocido como intercepto y β_1 es el vector fila que contiene todos los coeficientes asociados a las variables incluidas en X. Para estimar β_0 y β_1 se utiliza el método de máxima verosimilitud. La intuición detrás de este método es simple, consta de estimar valores de β_0 y β_1 tal que para aquellos individuos infractores el valor estimado de $p(X)$ sea cercano a 1 y para los no infractores sea cercano a 0. En [32] se puede

profundizar sobre el proceso de estimación de parámetros mediante el método de máxima verosimilitud, así como el cálculo de probabilidades una vez estimados los parámetros.

Aplicando álgebra sencilla sobre la ecuación anterior se obtiene,

$$\frac{p(X)}{1-p(X)} = e^{\beta 0 + \beta 1 X}, \tag{1.2}$$

donde $\frac{p(X)}{1-p(X)}$ se conoce como *odds ratio* o razón de probabilidades y puede tomar cualquier valor entre 0 e ∞. Valores del *odds ratio* cercanos a 0 indican una probabilidad baja de ser infractor, mientras que valores cercanos a ∞ una probabilidad alta de ser infractor. Al aplicar logaritmos a ambos lados de la ecuación se obtiene la siguiente expresión, donde el término a la izquierda de la igualdad se conoce como *log odds ratio* o *logit.*

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X. \tag{1.3}$$

La interpretación del modelo es la siguiente, por cada unidad que aumenta X_i el *logit* aumenta en β_{1i}. Cabe mencionar que en este modelo, β_{1i} no puede interpretarse como el impacto en $p(X)$ de aumentar X_i en una unidad, ya que como se muestra en (1.1) la relación entre ambas variables no es lineal y por tanto, dicho efecto dependerá del nivel que tenga X_i. Sin embargo, el signo que toma el parámetro β_{1i} indica hacia donde se moverá $p(X)$ cuando se aumente X_i en una unidad independientemente de su nivel. De esta forma, si $\beta_{1i} > 0$, al aumentar X_i en una unidad la probabilidad de ser infractor, $p(X)$, aumentará. Analogamente, si $\beta_{1i} < 0$, al aumentar X_i en una unidad $p(X)$ decrecerá.

4.2. Redes neuronales artificiales (ANN)

Las redes neuronales artificiales son modelos computacionales que buscan replicar la forma en que las neuronas biológicas procesan la información. Es así, que la característica central de estos algoritmos viene dada por su estructura. Dicha estructura se basa en un conjunto de nodos interconectados a través de los cuales viaja la información con el objetivo de resolver tareas o problemas particulares. En [33] se mencionan algunos de los problemas que han sido abordados mediante el uso de redes neuronales artificiales, donde se destacan tareas vinculadas al procesamimento de

imágenes como la detección de caracteres y reconocimiento facial. No obstante, vale la pena mencionar que este tipo de algoritmos son complejos de comunicar ya que por su estructura terminan operando como cajas negras. A continuación, se muestra un esquema sencillo de una red neuronal artifical.

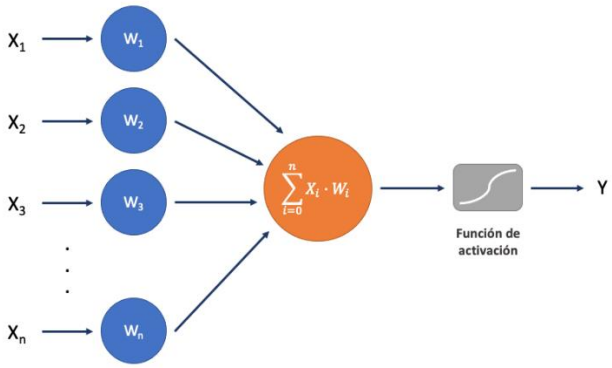


Ilustración 3. Esquema de una red neuronal artificial

En el esquema anterior, X_i representan la información de entrada o inputs del modelo. Cada uno de estos inputs son multiplicados por un peso W_i para posteriormente ser enviados a la neurona. En la neurona de la red es donde se computa el resultado de multiplicar los inputs por sus pesos correspondientes. A dicho resultado se le aplica una función de activación para así obtener el valor de la etiqueta estimado por la red, expresado como Y en la imagen anterior.

Al igual que el cerebro humano, las redes neuronales artificiales cuentan con un gran entramado de neuronas interconectadas y agrupadas en diferentes niveles conocidos como capas. Este tipo de neuronas también se conocen como *Multi Layer Perceptron* (MLP). A modo de ejemplo, la siguiente imagen muestra una red neuronal con cuatro capas.

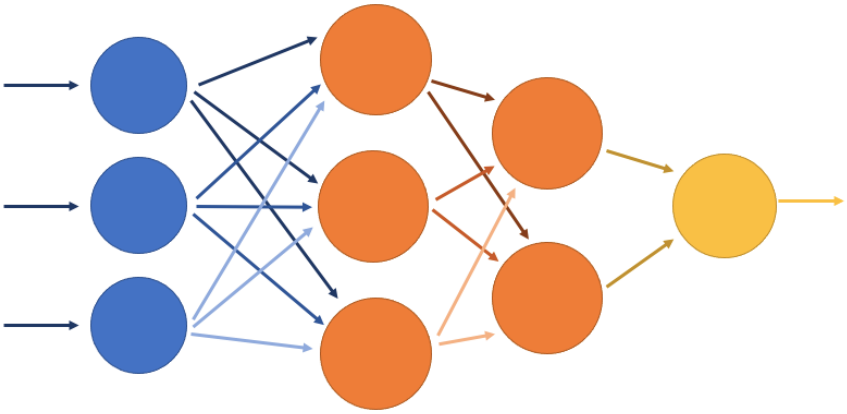


Ilustración 4. Esquema de una red neuronal artificial con cuatro capas

Las tres neuronas de la primera capa reciben los datos reales que alimentan la red neuronal y se conoce como capa de entrada. La salida de la última neurona es el resultado tangible de la red, por lo que recibe el nombre de capa de salida. Las capas intermedias se conocen como capas ocultas, por lo que la red de la Ilustración 4 cuenta con una primera capa oculta de tres neuronas y una segunda capa oculta de dos neuronas. Generalmente, las redes neuronales contienen una capa de entrada, una capa de salida y pueden incluir capas ocultas.

Dado que este trabajo plantea un problema de clasificación, se decide utilizar la función sigmoidea como función de activación. Dicha función genera un valor de salida que es una probabilidad similar a la obtenida en la regresión logística. En este caso, la probabilidad obtenida es de la forma,

$$p(X) = f(X;\, W_i), \tag{2.1}$$

en donde X es el vector de inputs y W_i son los pesos asociados a cada uno de los componentes X_i pertenecientes a X.

Una vez definida la estructura de la red se procede a realizar el proceso de entrenamiento con el fin de resolver el problema planteado. El entrenamiento de las redes neuronales artificiales consiste en la estimación de todos los pesos W_i de las entradas de las neuronas de forma que los resultados de la capa de salida sean lo más similares a los valores reales. El proceso de entrenamiento de la red se realiza mediante la minimización de una función

de pérdida que evalúa la red en su totalidad, proceso de optimización que se logra a través de la actualización de los pesos W_i de las neuronas. Dentro de los algoritmos más utilizados para minimizar la función de pérdida se encuentran: descenso del gradiente o *gradient descent* por su nombre en inglés, *stochastic gradient descent* y *backpropagation*. En caso de querer profundizar en cómo opera cada uno de estos algoritmos de optimización se recomienda ver [33].

4.3. Árboles de decisión (CART)

El árbol de decisión pertenece a una categoría de métodos conocidos como basados en árboles. Dichos métodos son ampliamente utilizados en problemas de aprendizaje supervisado, ya que generan diagramas de fácil interpretación que permiten representar y categorizar una serie de condiciones de forma sucesiva. Dichos modelos pueden ser utilizados tanto para tareas de regresión como de clasificación.

Al analizar la estructura de un árbol de decisión desde arriba hacia abajo, los tres principales componentes son: nodos internos, ramas y nodos terminales u hojas. A continuación, se muestra un ejemplo de un diagrama de árbol de clasificación.

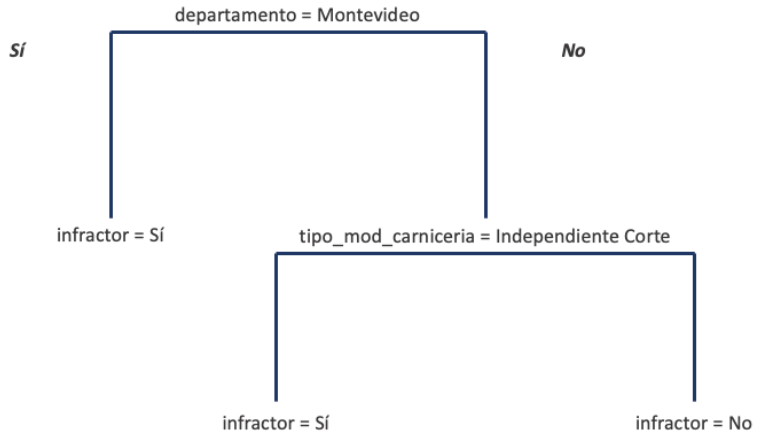


Ilustración 5. Diagrama de árbol de clasificación

Los nodos internos son aquellas preguntas o condiciones, también llamadas *splits*, que separan el espacio de predicción en dos ramas. Generalmente, dichas condiciones son de la forma $X_i < t_k$, donde X_i es la variable sobre la cual se impone el umbral t_k. En la

imagen anterior, los nodos internos vienen dados por las condiciones *departamento = Montevideo* y *tipo_mod_carniceria = Independiente Corte.*

A la izquierda de estos nodos se llega al nodo terminal u hoja en donde se cumple la condición establecida. Siguiendo con el ejemplo de la Ilustración 5, se llega a la región en donde se encuentran las carnicerías que pertenecen al departamento de Montevideo. El valor que se muestra en esta región, infractor = Sí, corresponde a la clase con mayor probabilidad de ocurrencia dentro de las carnicerías de Montevideo. De esta forma, el árbol predice la categoría infractor = Sí para cualquier establecimiento ubicado en Montevideo.

Por el contrario, siguiendo la rama de la derecha se llega a la región en la que se incluyen todas las observaciones que no cumplen con la condición, es decir, carnicerías que no están en el departamento de Montevideo. Para estas observaciones, el árbol de decisión realiza un nuevo split determinado por la condición *tipo_mod_carniceria = Independiente Corte.* A partir de esta nueva segmentación, el árbol predice la clase infractor = Sí para las carnicerías del tipo Independiente Corte e infractor = No para las carnicerías de otro tipo.

A diferencia de la regresión logística y la ANN que estiman la probabilidad de ser infractor, el árbol de decisión asigna el valor de la clase más probable entre las observaciones. Dicho valor dependerá de como se separe el espacio de predicción, por lo que el proceso mediante el cual se realizan los *splits* es relevante. Este proceso se conoce como *binary recursive splitting* y consta de considerar cada una de las variables (X_i) y umbrales (t_k) posibles para realizar los *splits* y seleccionar aquellos que minimizan una función de error global del árbol.

Este es un proceso iterativo, por lo que una vez realizado un *split* se generan dos nuevas regiones en donde se seleccionará una de ellas para realizar una nueva segmentación que minimice la función de error. La optimización finalizará en cuanto se alcance alguna condición prestablecida como ser un número mínimo de observaciones por región o determinado nivel de profundidad del árbol. En [34] se analiza en detalle el *binary recursive splitting* así como técnicas para evitar el sobreajuste de los modelos como el *tree pruning* o podado de árboles.

En el caso de los árboles de clasificación, la función a minimizar es el error de clasificación. Éste se define como la proporción de observaciones dentro de la región que no pertenecen a la clase más probable,

$$E = 1 - \max_k(p_{mk}), \tag{3.1}$$

donde p_{mk} representa la proporción de las observaciones dentro de la región m que pertenecen a la clase k.

El error de clasificación es un indicador relevante al analizar la precisión de las predicciones, sin embargo, no resulta adecuado para evaluar la calidad de los *splits.* A modo de ejemplo, si en determinada región del espacio de predicciones gran parte de las observaciones están categorizadas bajo la misma clase, el error de clasificación resultante sería relativamente bajo. No obstante, esto no implica que el *split* realizado sea de buena calidad, sino que por el contrario representa un caso de desbalance en el conjunto de datos. Es así, que métricas más sensibles a la pureza de los nodos como el índice de Gini y la entropía son preferidas para evaluar la calidad de los *splits*.

El índice de Gini representa una medida de la varianza total entre las K clases y se define como, $$G = \sum_{k=1}^{K} p_{mk}\,(1 - p_{mk}), \tag{3.2}$$

G toma valores pequeños cuando las p_{mk}´s son cercanas a cero o uno, por lo que también se lo conoce como una medida de pureza del nodo. Valores pequeños de G indican que un nodo contiene, en su mayoría, observaciones correspondientes a una única clase.

Por su parte, la entropía es una medida similar al índice Gini, por lo que también tomará valores pequeños en caso de que los nodos sean puros. La entropía se define como,

$$D = -\sum_{k=1}^{K} p_{mk}\, log\, p_{mk}. \tag{3.3}$$

Por el número de observaciones y el desbalance que presentan las clases del conjunto de datos utilizado en este trabajo, es probable que este algoritmo no arroje resultados favorables en la clasificación de infractores.

5. Análisis exploratorio del conjunto de datos

Una vez generada la base de datos final y seleccionados los algoritmos para la estimación de los modelos, se procede a realizar un análisis exploratorio del conjunto de datos. En primer lugar, se analiza el porcentaje de datos faltantes para cada una de las variables que integran el dataset. La imagen siguiente muestra que existe un 10% de datos faltantes en el dataset, donde la variable *compras_avg* es la que presenta mayor porcentaje de faltantes (97,4%). De esta forma, se decide eliminar dicha variable del análisis.

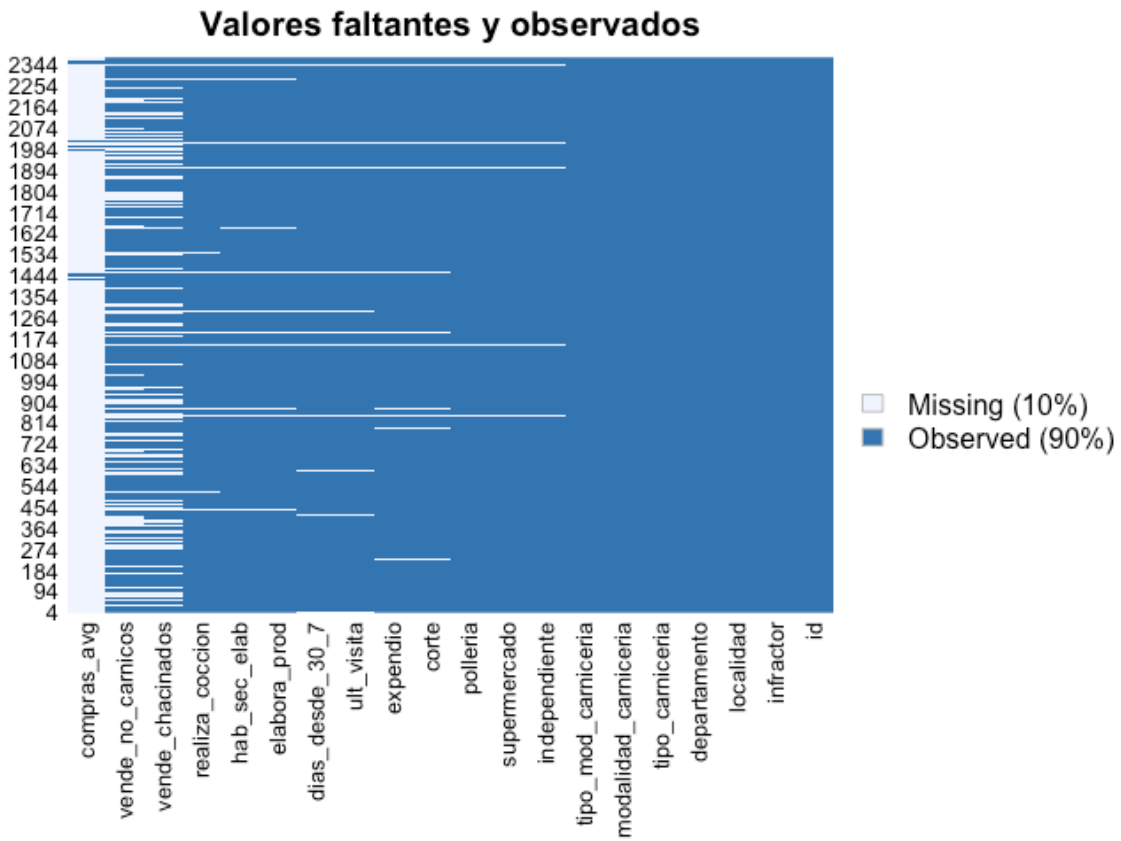


Ilustración 6. Valores faltantes y observados en el dataset

Las variables *vende_no_carnicos* y *vende_chacinados* presentan datos faltantes en el 31,9% y 29,4% de las observaciones respectivamente. No obstante, dado que estas variables podrían ser relevantes para estimar la cualidad de infractor de los establecimientos se decide asignar el valor 0 en aquellas observaciones para las que no se cuenta con datos. Por su parte, las líneas celestes horizontales que van desde la variable

compras_avg hasta *independiente* son establecimientos para los cuales no se cuenta con covariable alguna por lo que se decide eliminarlas del dataset.

La principal característica comercial que será incluida como covariable en los modelos es la variable *tipo_mod_carniceria*. Dicha variable se forma mediante la unión del tipo y modalidad del establecimiento y define gran parte de las actividades que cada punto de venta puede realizar (ver Tabla 1). Al analizar la composición del tipo y modalidad de los establecimientos se observa que existe una observación del tipo Pollería y modalidad Expendio, un caso que según las combinaciones detalladas en la Tabla 1 no existe. De esta forma, se decide eliminar dicha observación del dataset.

Como resultado, se obtiene un conjunto de datos con 2.175 puntos de venta que se distribuyen a nivel departamental según la Ilustración 7. La Ilustración 8 muestra la cantidad de establecimientos según tipo y modalidad.

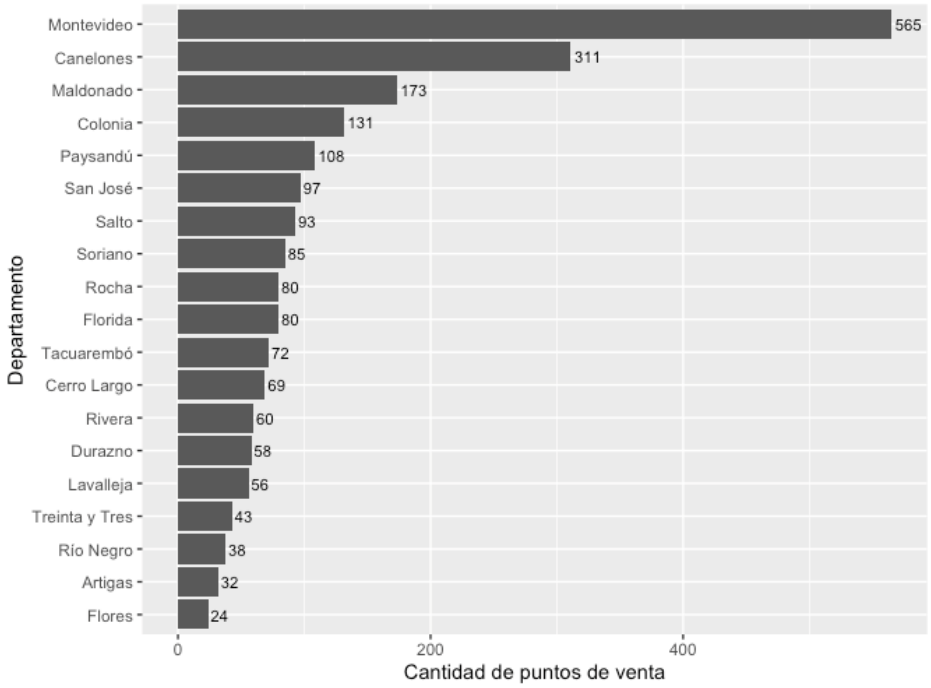


Ilustración 7. Puntos de venta por departamento

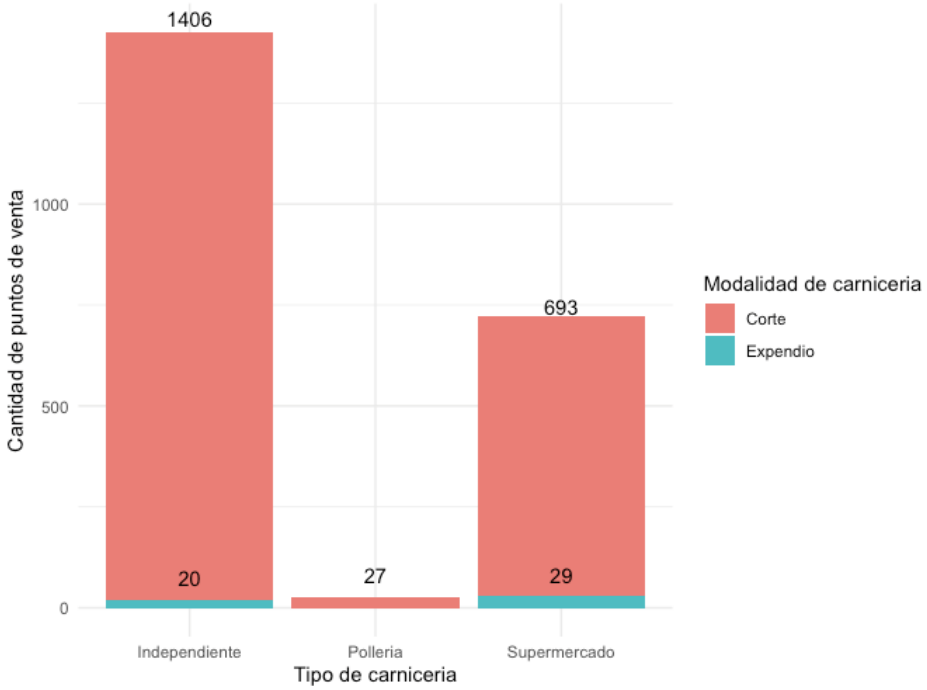


Ilustración 8. Puntos de venta según tipo y modalidad

Como puede observarse Montevideo, Canelones y Maldonado son los departamentos con mayor número de puntos de venta de carnes y sus derivados. Con respecto al tipo y modalidad de los establecimientos, se observa que la mayor parte opera bajo la modalidad de Corte, esto es, reciben la media canal[5] para realizar el fraccionamiento y vender cortes al público. Por el contrario, los que operan bajo la modalidad de Expendio solo pueden vender cortes envasados en origen.

Otra característica comercial relevante es si los puntos de venta elaboran productos. Para ello, se debe contar con un sector de elaboración habilitado. En la Ilustración 9 se puede apreciar como la mayor parte de los establecimientos que operan bajo modalidad de Corte en la práctica elaboran productos, pero a su vez, solo unos pocos se encuentran habilitados para hacerlo. Esta situación se analizará en los modelos mediante la creación de la variable *elaboración* que contemplará la elaboración de productos y la existencia de un sector de elaboración habilitado en conjunto.

[5] Según [3], "Es cada una de las dos partes resultantes de dividir la canal, mediante un corte longitudinal que pasa por la línea media de la columna vertebral."

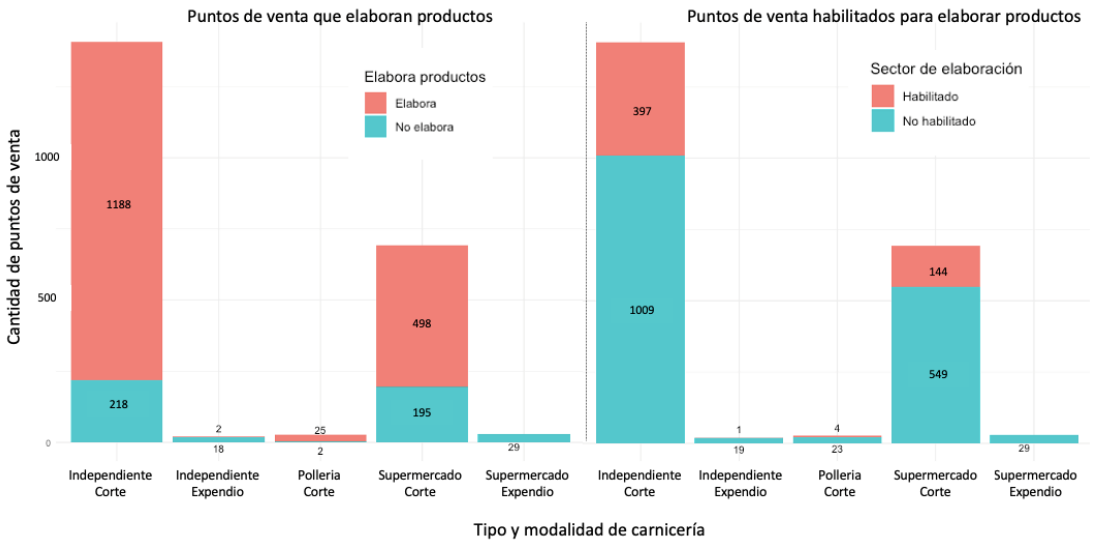


Ilustración 9. Puntos de venta que elaboran productos y habilitados para elaborar productos

Por último, se analiza la frecuencia de las variables *realiza_coccion*, *vende_no_carnicos* y *vende_chacinados* según tipo y modalidad de carnicería. En la siguiente ilustración se puede observar que únicamente el 5% de los puntos de venta analizados realiza cocción en sus locales. Con respecto a la venta de productos no cárnicos[6], poco menos de la mitad de los establecimientos (46%) vende este tipo de productos, mientras que solo el 21% de los locales vende chacinados. Vale la pena destacar, que la oferta de estos productos se concentra en locales que operan bajo la modalidad de Corte, lo que es consistente con lo relevado en la Tabla 1.

[6] Ver productos no cárnicos habilitados por INAC en [31]

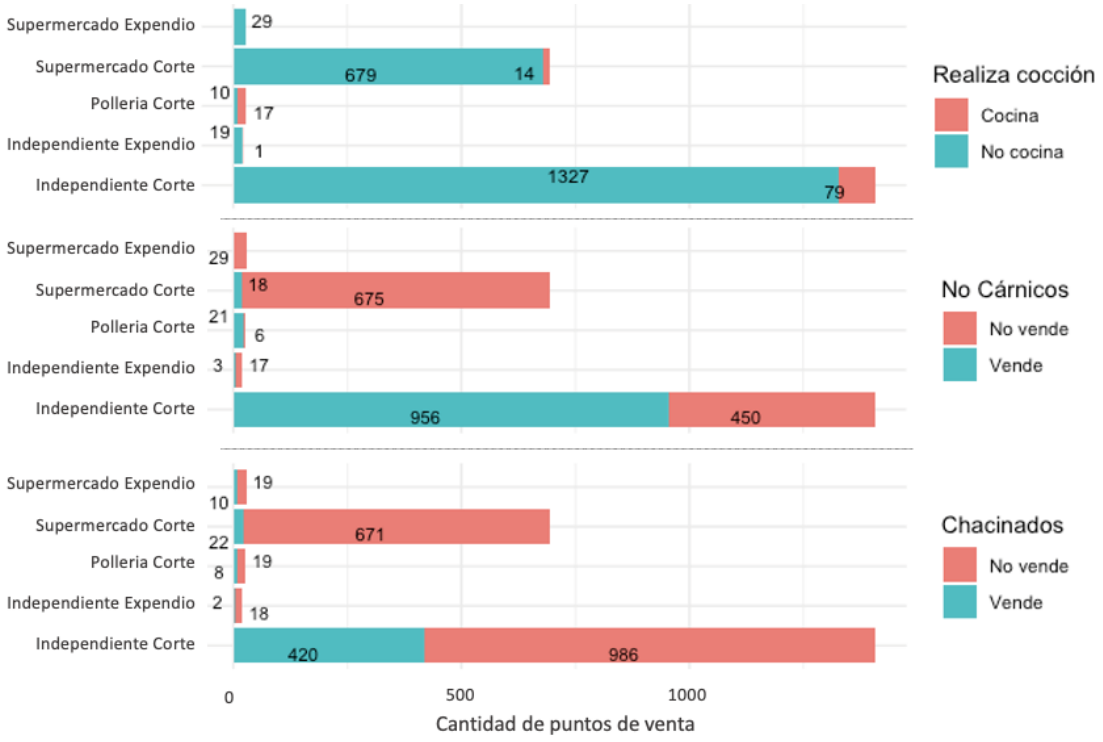


Ilustración 10. Otras características comerciales según tipo y modalidad de carnicería

Un aspecto relevante a conocer es la proporción de establecimientos infractores en el dataset. Dicha proporción puede entenderse como la probabilidad de encontrar un establecimiento infractor a lo largo del país. Como puede observarse en la siguiente imagen, el conjunto de datos cuenta con 340 establecimientos infractores, valor que representa el 15,6% de las observaciones.

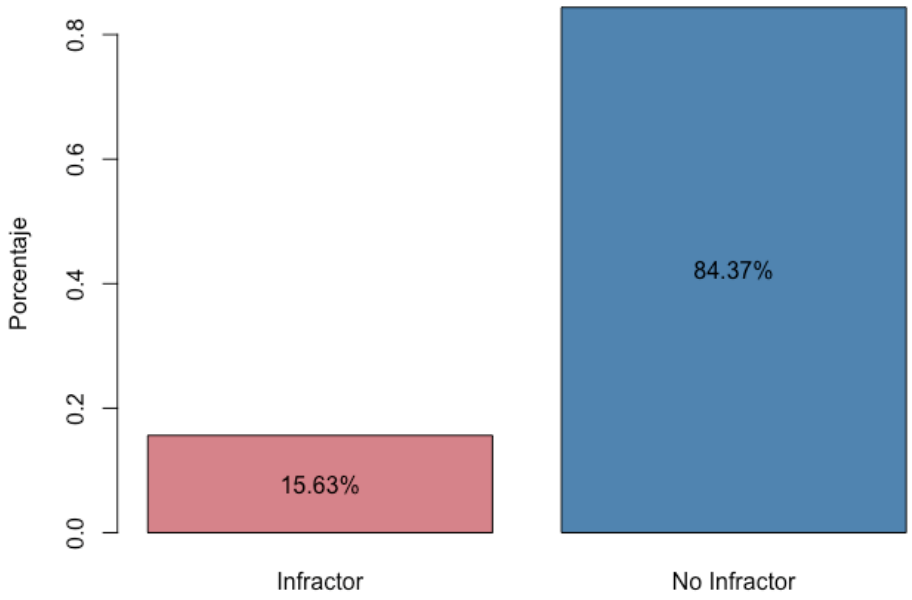


Ilustración 11. Proporción de infractores en el conjunto de datos

Al analizar la proporción de infractores a lo largo del país, se observa una alta heterogeneidad a nivel departamental. Además, se observa que los departamentos con mayor número de establecimientos infractores también son los que presentan un mayor porcentaje. A partir de esta imagen, se decide fijar al departamento de Canelones como la base de comparación de la variable *departamento* en los modelos.

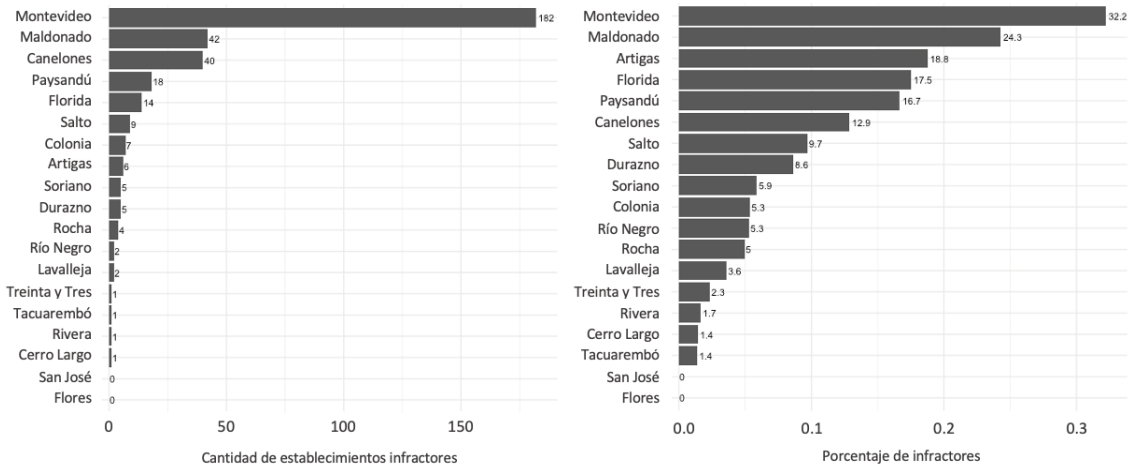


Ilustración 12. Cantidad y porcentaje de infractores por departamento

Por último, se analizan los establecimientos infractores según el tipo y modalidad bajo la que operan. Se puede observar que la mayoría de los infractores operan bajo la modalidad de Corte, dato esperable ya que los establecimientos Independiente Corte y Supermercado Corte representan, en conjunto, el 94% de los establecimientos en el conjunto de datos.

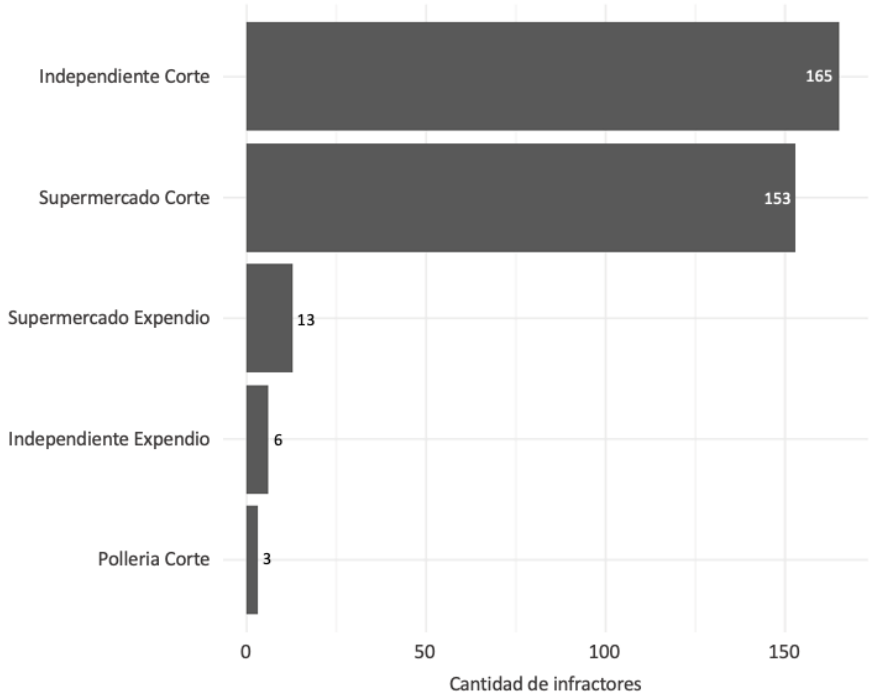


Ilustración 13. Cantidad de infractores según tipo y modalidad de carnicería

Al analizar la proporción de infractores según tipo y modalidad a nivel departamental, se encuentra que los infractores en Cerro Largo, Tacuarembó y Treinta y Tres son en su totalidad Independiente Corte, mientras que en Lavalleja, Río Negro y Rivera la totalidad corresponde a Supermercado Corte. En Montevideo, Canelones y Florida es donde se encuentra otro tipo de establecimientos infractores como son Independiente Expendio o Pollería Corte. Flores y San José no se incluyen en la siguiente imagen ya que no se cuenta con establecimientos infractores para esos departamentos en el conjunto de datos.

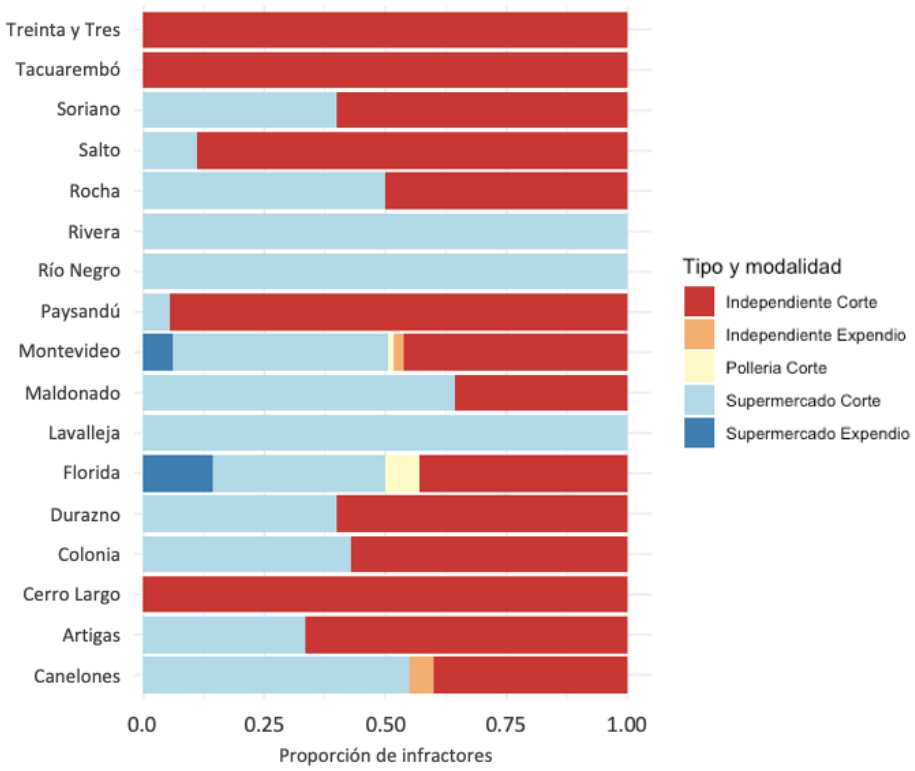


Ilustración 14. Proporción de infractores según tipo y modalidad a nivel departamental

6. Modelos estimados

En este capítulo se detalla el proceso de estimación de modelos y los resultados obtenidos. Para la estimación se utilizo el *validation set approach,* técnica que consiste en dividir el dataset en dos conjuntos de datos: uno de entrenamiento (*train*) al que se le asignan el 70% de las observaciones y uno de prueba (*test*) que contiene el 30% restante. El modelo se ajusta (entrena) en el conjunto de *train* y se evalúa en el conjunto de *test*. De esta forma, el modelo se entrena utilizando determinado conjunto de datos y se evalúa en un nuevo conjunto, desconocido para el modelo, minimizando la posibilidad de sobreajuste.

6.1. Regresión logística (RL)

La regresión logística (*logit)* será el modelo base de este trabajo por dos motivos: (i) por ser una técnica estadística tradicional ampliamente utilizada para abordar este tipo de problemas y (ii) por su simplicidad, ya que permite interpretar y comunicar los resultados de manera clara y sencilla.

Para decidir qué features o covariables del dataset serán incluidos en el análisis, se ajustan diversos modelos utilizando todo el conjunto de datos y se identifican aquellas covariables significativas. Se ajustaron modelos utilizando la ubicación, las características comerciales de las establecimientos y por último, covariables exógenas asociadas al departamento como población, superficie e ingreso. Dado que se quiere prioirzar el impacto de la ubicación y las características comerciales sobre la variable *infractor,* las covariables exógenas no se incluyen en los *logit.*

A partir de este proceso se encuentra que la covariable *departamento* es signifactiva y contribuye al ajuste del modelo con una reducción de la devianza (ver Anexo 2). Al analizar los features vinculados a características comerciales de las carnicerías se encuentra que *tipo_mod_carniceria, elaboración, vende_no_carnicos* y *vende_chacinados* son significativas. Al incluir *vende_no_carnicos* y *vende_chacinados* en la misma ecuación la primera deja de ser significativa, por lo que se decide mantener únicamente *vende_chacinados* (ver Anexo 3).

Al agregar *departamento* junto con las características comerciales, el feature *elaboración* deja de ser significativo. Por su parte, *tipo_mod_carniceria* y *vende_chacinados* mantienen su signo y significancia estadística. De esta forma, se decide trabajar con las variables *departamento, tipo_mod_carniceria* y *vende_chacinados* en los modelos a estimar (ver Anexo 4).

6.1.1. Modelo 1: RL - M1

El primer modelo incluye como único feature dentro del vector X a la variable *departamento*, por lo que se estima un β_i por departamento. Dado que la variable *departamento* tiene como valor base a Canelones, en lugar de estimar 19 parámetros el modelo estimará 18, uno por cada departamento sin contar el valor base. Como resultado, los valores estimados para cada uno de los parámetros β_i se interpretan con respecto a la base de la variable, es decir, el departamento de Canelones.

Como se puede observar en la imagen siguiente, los departamentos de Colonia, Maldonado, Montevideo y Rocha son significativos al 5%. El signo de los parámetros estimados por el modelo se encuentra en línea con lo obtenido en la Ilustración 12, donde a los establecimientos de Montevideo y Maldonado - departamentos con el mayor porcentaje de infractores – se les asigna mayor probabilidad de ser infractores. A modo de ejemplo, un establecimiento de Montevideo tiene un *logit* mayor al de Canelones en 1,1227 unidades. Al exponenciar el coeficiente, se obtiene un *odds ratio* de 3,07 lo que implica que las chances de ser infractor en carnicerías de Montevideo son 3 veces más que en Canelones. Por el contrario, el modelo asigna una menor probabilidad de ser infractor para los establecimientos ubicados en Colonia y Rocha, ya que el *logit* estimado es 1,2219 y 2,1739 unidades menor que el estimado para Canelones respectivamente, lo que significa que las chances de ser infractor en dichos deparamentos son de 0,29 y 0,11 con respecto al departamento base. Dado que Canelones es el valor base de *departamento*, el *intercept* recoge la probabilidad de que una carnicería de Canelones sea infractor. Operando con el coeficiente asociado al *intercept* se llega a que la probabilidad estimada de ser infractor para un establecimiento de Canelones es cercana al 13%, acorde con lo observado en la Ilustración 12.

```
> summary(model_1)

Call:
glm(formula = infractor ~ departamento, family = binomial, data = train.data)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.88094  -0.63999  -0.43660  -0.00013   2.84972

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)                -1.86915    0.19950  -9.369  < 2e-16 ***
departamentoArtigas        -0.02797    0.65049  -0.043  0.96570
departamentoCerro Largo   -16.69692  994.69308  -0.017  0.98661
departamentoColonia        -1.22190    0.54879  -2.227  0.02598 *
departamentoDurazno        -1.04862    0.75288  -1.393  0.16368
departamentoFlores        -16.69692 1809.05447  -0.009  0.99264
departamentoFlorida         0.38754    0.40315   0.961  0.33641
departamentoLavalleja      -1.04862    0.75288  -1.393  0.16368
departamentoMaldonado       0.86015    0.28070   3.064  0.00218 **
departamentoMontevideo      1.12275    0.22654   4.956 7.19e-07 ***
departamentoPaysandú        0.07739    0.38192   0.203  0.83943
departamentoRío Negro      -0.43344    0.76799  -0.564  0.57249
departamentoRivera        -16.69692  994.69308  -0.017  0.98661
departamentoRocha          -2.17391    1.02827  -2.114  0.03450 *
departamentoSalto          -0.24539    0.44710  -0.549  0.58312
departamentoSan José       -16.69692  790.98614  -0.021  0.98316
departamentoSoriano        -1.09268    0.62480  -1.749  0.08032 .
departamentoTacuarembó     -16.69692  941.46181  -0.018  0.98585
departamentoTreinta y Tres  -1.56484    1.03540  -1.511  0.13070
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1320.2  on 1522  degrees of freedom
Residual deviance: 1120.6  on 1504  degrees of freedom
AIC: 1158.6

Number of Fisher Scoring iterations: 17
```

Ilustración 15. Resumen RL - M1

Una vez ajustado el modelo en el conjunto de entrenamiento se realizan predicciones para cada una de las observaciones del conjunto de datos de prueba. Como resultado, se obtiene para cada observación la probabilidad de ser considerado infractor. La siguiente imagen muestra el histograma de probabilidades estimadas.

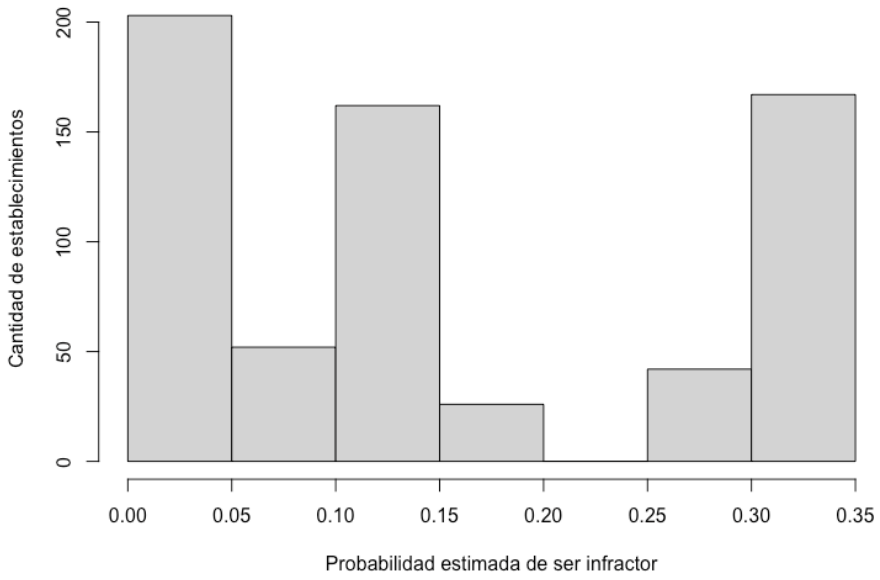


Ilustración 16. Histograma de probabilidades estimadas RL - M1

A modo de evaluar la performance del modelo, las probabilidades estimadas son transformadas en variables discretas con valor 1 para los infractores y 0 para el resto de los establecimientos. Para ello, se debe seleccionar un umbral a partir del cual clasificar las observaciones en infractores (valor 1) y no infractores (valor 0). Al analizar el histograma se observa que el mismo se divide en dos partes, una en donde las probabilidades se encuentran en el rango 0 – 0,20 y otra en donde alrededor de 200 observaciones pertenecen al rango 0,25 – 0,35. De esta forma, se decide clasificar como infractor a los establecimientos cuya probabilidad estimada es igual o mayor a 0,3.

Posteriormente, se computa la matriz de confusión en donde se comparan las categorías estimadas con los valores reales de la variable *infractor* en el conjunto de datos de prueba. Las métricas a utilizar para evaluar la performance son las siguientes:

- Accuracy: porcentaje de verdaderos positivos y verdaderos negativos identificados como porcentaje del total.
- Error de clasificación: porcentaje de falsos positivos y negativos como porcentaje del total. Es equivalente a 1 – Accuracy.
- Recall: también conocido como *Sensitivity*, mide el porcentaje de verdaderos positivos identificados.

- Precisión: también llamado Positive Predictive Value, mide el número de verdaderos positivos identificados como proporción del total de positivos predichos.
- F1 score: métrica que promedia Recall y Precisión. F1 score alcanza su mejor valor en 1 (recall y precisión perfectas) y el peor en 0.
- Balanced accuracy: es una métrica mñas robusta al tratar con clases desbalanceadas. Se calcula como el promedio simple de Recall y *Specificity*.

Vale la pena mencionar, que el accuracy puede ser engañoso al evaluar conjuntos de datos con clases desbalanceadas. En estos casos métricas como recall, precision, F1 score y balanced accuracy son preferibles.

La matriz de confusión y métricas obtenidas se muestran en la Ilustración 17.

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 437  48
         1 113  54

               Accuracy : 0.7531
                 95% CI : (0.7181, 0.7857)
    No Information Rate : 0.8436
    P-Value [Acc > NIR] : 1

                  Kappa : 0.2572

 Mcnemar's Test P-Value : 4.561e-07

              Precision : 0.32335
                 Recall : 0.52941
                     F1 : 0.40149
             Prevalence : 0.15644
         Detection Rate : 0.08282
   Detection Prevalence : 0.25613
      Balanced Accuracy : 0.66198

       'Positive' Class : 1
```

Ilustración 17. Matriz de confusión y métricas de performance RL - M1

Al evaluar la performance del modelo por el accuracy, se puede observar que el modelo predice correctamente el 75,31% de las observaciones, arrojando un error de clasificación de 24,69%. Como se menciono previamente, el accuracy es una métrica engañosa. Dado que el conjunto de datos de prueba cuenta con 652 observaciones, de las cuales 550 son no infractoras, si el modelo predijera *no infractor* para todas las observaciones el accuracy

sería de 84,35%. Teniendo esto en mente y con el objetivo de predecir a los establecimientos infractores es más relevante observar el recall (52,94%). Este valor implica que el modelo identifica correctamente a 5 de cada 10 infractores reales del conjunto de datos de prueba. Además, de los 167 establecimientos que el modelo predice como infractores, únicamente 54 (32,33%) son infractores reales. Por último, el F1 score es de 0,40 mientras que el balanced accuracy es de 0,662. A partir de las métricas obtenidas, se entiende que el modelo no presenta un buen rendimiento al clasificar establecimientos infractores.

6.1.2. Modelo 2: RL – M2

El segundo modelo incluye además de *departamento* el feature *tipo_mod_carniceria* dentro del vector X. El valor base en *tipo_mod_carniceria* es Independiente Corte, por lo que los β_j estimados para cada combinación de tipo y modalidad se interpretan con respecto a dicha base.

Como se puede observar en la Ilustración 18, los departamentos de Maldonado, Montevideo y Rocha mantienen su signo y significancia estadística. Por su parte, los establecimientos Independiente Expendio y Supermercado Corte resultan signifcativos al 5% y con signo positivo, lo que implica que para estos tipos y modalidad de carnicería la probabilidad de ser infractor es mayor que para un establecimiento Independiente Corte. Específicamente, el *logit* aumenta con respecto a la base en 1,7026 unidades para las carnicerías Independiente Expendio y 0,9027 unidades para los Supermercado Corte. Esto arroja que las chances de ser infractor en comparación con un establecimiento Independiente Corte son 5 y 2,5 veces mayores para las carnicerías Independiente Expendio y Supermercado Corte. Cabe destacar que estos resultados podrían ser consecuencia de la asignación de la etiqueta infractor a todos los establecimientos dentro de un mismo departamento operado por una empresa infractora, aspecto que fue detallado en el capítulo 1.8.

El histograma de probabilidades estimadas (ver Anexo 5) presenta algunas diferencias con respecto al estimado para RL – M1. Mientras que en RL – M1 la totalidad de las probabilidades estimadas se encuentra en el rango 0 – 0,35, en este caso se distribuyen en un rango mayor (entre 0 – 0,70). No obstante, dado que gran parte de las probabilidades

se mantiene en el rango 0 – 0,35 se decide continuar con el umbral definido previamente (0,30).

```
> summary(model_2)

Call:
glm(formula = infractor ~ departamento + tipo_mod_carniceria,
    family = binomial, data = train.data)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.42537  -0.63242  -0.41029  -0.00012   3.00898

Coefficients:
                                      Estimate Std. Error z value Pr(>|z|)
(Intercept)                          -2.41065    0.22773 -10.586  < 2e-16 ***
departamentoArtigas                   0.17253    0.65924   0.262  0.79354
departamentoCerro Largo             -16.39012  977.64880  -0.017  0.98662
departamentoColonia                  -0.97380    0.55328  -1.760  0.07840 .
departamentoDurazno                  -0.86840    0.75781  -1.146  0.25182
departamentoFlores                  -16.63655 1776.99541  -0.009  0.99253
departamentoFlorida                   0.62003    0.41455   1.496  0.13474
departamentoLavalleja                -0.82927    0.76366  -1.086  0.27751
departamentoMaldonado                 0.89609    0.28584   3.135  0.00172 **
departamentoMontevideo                1.27423    0.23336   5.460 4.75e-08 ***
departamentoPaysandú                  0.43779    0.39159   1.118  0.26357
departamentoRio Negro                -0.25526    0.77556  -0.329  0.74206
departamentoRivera                  -16.43742  975.94371  -0.017  0.98656
departamentoRocha                    -2.10545    1.03160  -2.041  0.04125 *
departamentoSalto                    -0.02189    0.45365  -0.048  0.96151
departamentoSan José                -16.59877  776.71818  -0.021  0.98295
departamentoSoriano                  -0.77625    0.63043  -1.231  0.21821
departamentoTacuarembó              -16.47897  926.27766  -0.018  0.98581
departamentoTreinta y Tres           -1.23032    1.04112  -1.182  0.23732
tipo_mod_carniceriaIndependiente Expendio   1.70268    0.69353   2.455  0.01408 *
tipo_mod_carniceriaPolleria Corte     1.00633    0.71341   1.411  0.15836
tipo_mod_carniceriaSupermercado Corte   0.90276    0.16135   5.595 2.21e-08 ***
tipo_mod_carniceriaSupermercado Expendio   0.91271    0.47400   1.926  0.05416 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1320.2  on 1522  degrees of freedom
Residual deviance: 1083.7  on 1500  degrees of freedom
AIC: 1129.7

Number of Fisher Scoring iterations: 17
```

Ilustración 18. Resumen RL - M2

Con respecto a la performance del modelo, tanto el accuracy como el error de clasificación mejoran con respecto a RL – M1. Sin embargo, el recall (29,41%), F1 score (0,31) y balanced accuracy (0,592) son menores en comparación con el modelo anterior.

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 490  72
         1  60  30

               Accuracy : 0.7975
                 95% CI : (0.7646, 0.8278)
    No Information Rate : 0.8436
    P-Value [Acc > NIR] : 0.9993

                  Kappa : 0.1943

 Mcnemar's Test P-Value : 0.3384

              Precision : 0.33333
                 Recall : 0.29412
                     F1 : 0.31250
             Prevalence : 0.15644
         Detection Rate : 0.04601
   Detection Prevalence : 0.13804
      Balanced Accuracy : 0.59251

       'Positive' Class : 1
```

Ilustración 19. Matriz de confusión y métricas de performance RL - M2

6.1.3. Modelo 3: RL – M3

El úlitmo modelo agrega *vende_chacinados* a las variables anteriores dentro del vector X. El valor base en *vende_chacinados* es No vende, por lo que los β_k estimados se interpretan con respecto a dicha base.

Como se puede observar en la Ilustración 20, las variables significativas de los modelos anteriores mantienen tanto su signo como la significancia estadística. Con respecto a la variable *vende_chacinados*, en el Anexo 4 se observa que la misma es significativa al 5% cuando se ajusta el modelo al conjunto de datos completo, es decir, sin dividir en conjunto de *train* y *test.* No obstante, al ajustar el modelo a los datos de entrenamiento *vende_chacinados* mantiene el signo positivo pero su significancia es al 10%. El coeficiente estimado implica que aquellos establecimientos que venden productos chacinados tienen mayor probabilidad de ser infractores con respecto a los que no venden ese tipo de productos. El *logit* para vendedores de chacinados aumenta en 0,39 unidades, lo que representa 1,5 veces más chances de ser infractor con respecto a los no vendedores. Además, a partir del coeficiente estimado para el *intercept*, se encuentra que la

probabilidad estimada de ser infractor para una carnicería de Canelones, del tipo Independiente Corte y que no vende chacinados es cercana al 7%.

```
> summary(model_3)

Call:
glm(formula = infractor ~ departamento + tipo_mod_carniceria +
    vende_chacinados, family = binomial, data = train.data)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.41153  -0.62789  -0.38516  -0.00012   3.04979

Coefficients:
                                        Estimate Std. Error z value Pr(>|z|)
(Intercept)                            -2.56400    0.24290 -10.556  < 2e-16 ***
departamentoArtigas                     0.24817    0.66239   0.375  0.70791
departamentoCerro Largo               -16.36529  968.61989  -0.017  0.98652
departamentoColonia                    -0.96242    0.55358  -1.739  0.08212 .
departamentoDurazno                    -0.87467    0.75841  -1.153  0.24879
departamentoFlores                    -16.61352 1774.72527  -0.009  0.99253
departamentoFlorida                     0.64998    0.41436   1.569  0.11674
departamentoLavalleja                  -0.80459    0.76667  -1.049  0.29397
departamentoMaldonado                   0.90135    0.28631   3.148  0.00164 **
departamentoMontevideo                  1.27882    0.23326   5.482 4.20e-08 ***
departamentoPaysandú                    0.46900    0.39274   1.194  0.23241
departamentoRío Negro                  -0.22818    0.77626  -0.294  0.76880
departamentoRivera                    -16.34090  972.23525  -0.017  0.98659
departamentoRocha                      -2.07702    1.03193  -2.013  0.04414 *
departamentoSalto                       0.02257    0.45459   0.050  0.96040
departamentoSan José                  -16.59762  776.46500  -0.021  0.98295
departamentoSoriano                    -0.79470    0.63048  -1.260  0.20750
departamentoTacuarembó                -16.41584  921.17887  -0.018  0.98578
departamentoTreinta y Tres             -1.20873    1.04277  -1.159  0.24639
tipo_mod_carniceriaIndependiente Expendio  1.80768  0.69752   2.592  0.00955 **
tipo_mod_carniceriaPolleria Corte       1.03371    0.70645   1.463  0.14340
tipo_mod_carniceriaSupermercado Corte   1.03801    0.17818   5.826 5.69e-09 ***
tipo_mod_carniceriaSupermercado Expendio  0.91281  0.47870   1.907  0.05654 .
vende_chacinadosVende                   0.39029    0.20254   1.927  0.05398 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1320.2  on 1522  degrees of freedom
Residual deviance: 1080.1  on 1499  degrees of freedom
AIC: 1128.1

Number of Fisher Scoring iterations: 17
```

Ilustración 20. Resumen RL - M3

El histograma de probabilidades estimadas (ver Anexo 6) es muy similar al del modelo RL – M2, por lo que se decide mantener el umbral de 0,3. Al evaluar la performance del modelo, el accuracy (80,21%) y el error de clasificación (19,79%) de este modelo son ligeramente superiores a los dos anteriores. Sin embargo, como puede observarse en la

Ilustración 21, el recall, F1 score y balanced accuracy son similares a los obtenidos en el modelo anterior.

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 493  72
         1  57  30

               Accuracy : 0.8021
                 95% CI : (0.7695, 0.8321)
    No Information Rate : 0.8436
    P-Value [Acc > NIR] : 0.9980

                  Kappa : 0.2026

 Mcnemar's Test P-Value : 0.2177

              Precision : 0.34483
                 Recall : 0.29412
                     F1 : 0.31746
             Prevalence : 0.15644
         Detection Rate : 0.04601
   Detection Prevalence : 0.13344
      Balanced Accuracy : 0.59524

       'Positive' Class : 1
```

Ilustración 21. Matriz de confusión y métricas de performance RL - M3

Para este modelo se decide realizar una estimación adicional mediante *k-fold cross validation.* Esta técnica consiste en dividir el conjunto de datos en *k* particiones, donde el modelo se entrenará utilizando *k-1* particiones y será evaluado en la partición restante. Este proceso se repite hasta que cada partición se haya utilizado como conjunto de prueba y el resultado final se computa como el promedio de los resultados obtenidos en cada una de las *k* particiones. Esta técnica presenta ciertas ventajas como son la redución del sesgo y variabilidad de las estimaciones además de prevenir el sobreajuste en los modelos.

Para esta técnica se define $k = 10$, es decir, el conjunto de entrenamiento se divide en 10 particiones. Además, se decide agregar una técnica de oversampling que apunta a resolver el desbalance de clases. Esta técnica utiliza un muestreo aleatorio con reemplazo sobre la clase minoritaria (infractor) hasta igualar el tamaño de la clase mayoritaria (no infractor).

Como se puede observar a continuación, *vende_chacinados* es significativa al 5% y mantiene el signo positivo de su coeficiente. Además, producto del *oversampling*, otros coeficientes asociados a departamentos, tipos y modalidad de carnicerías se convierten

en significativos. El valor NA vinculado a Polleria Expendio se debe a la eliminación de la única observación con esas características, aspecto que fue detallado en el capítulo 5.

```
> summary(model_3_cv)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1702  -0.9824   0.2235   0.8867   2.4817

Coefficients: (1 not defined because of singularities)
                                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                               -0.96608    0.13844  -6.978 2.99e-12 ***
departamentoArtigas                        0.79957    0.33866   2.361 0.018225 *
`departamentoCerro Largo`                -16.98712  580.17902  -0.029 0.976642
departamentoColonia                       -0.56893    0.24351  -2.336 0.019472 *
departamentoDurazno                       -0.87705    0.36961  -2.373 0.017647 *
departamentoFlores                       -17.24239 1058.01487  -0.016 0.986998
departamentoFlorida                        0.55896    0.24906   2.244 0.024816 *
departamentoLavalleja                     -0.90722    0.38450  -2.359 0.018302 *
departamentoMaldonado                      0.88535    0.17294   5.119 3.06e-07 ***
departamentoMontevideo                     1.28976    0.13636   9.458  < 2e-16 ***
departamentoPaysandú                       0.61442    0.22350   2.749 0.005976 **
`departamentoRio Negro`                   -0.03786    0.37028  -0.102 0.918570
departamentoRivera                       -16.95853  581.91966  -0.029 0.976751
departamentoRocha                         -2.06638    0.48649  -4.248 2.16e-05 ***
departamentoSalto                          0.08124    0.25673   0.316 0.751661
`departamentoSan José`                   -17.21586  463.70735  -0.037 0.970384
departamentoSoriano                       -0.88962    0.32448  -2.742 0.006112 **
departamentoTacuarembó                   -17.03257  549.44988  -0.031 0.975270
`departamentoTreinta y Tres`              -0.59865    0.40334  -1.484 0.137748
`tipo_mod_carniceriaIndependiente Expendio`  1.92886    0.53392   3.613 0.000303 ***
`tipo_mod_carniceriaPolleria Corte`        0.95048    0.42267   2.249 0.024527 *
`tipo_mod_carniceriaPolleria Expendio`          NA         NA      NA       NA
`tipo_mod_carniceriaSupermercado Corte`    1.13009    0.11095  10.185  < 2e-16 ***
`tipo_mod_carniceriaSupermercado Expendio`  0.96593    0.33412   2.891 0.003840 **
vende_chacinadosVende                      0.40703    0.12289   3.312 0.000926 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3562.8  on 2569  degrees of freedom
Residual deviance: 2755.5  on 2546  degrees of freedom
AIC: 2803.5

Number of Fisher Scoring iterations: 16
```

Ilustración 22. Resumen RL – M3 utilizando k-fold cross validation (KCV) y oversampling

En cuanto a la performance del modelo, el accuracy decrece a 61,66% y en consecuencia aumenta el error de clasificación a 38,34%. No obstante, el recall es mucho mayor en comparación con los modelos anteriores ya que detecta al 74,51% de los infractores reales. Además, si bien la precisión del modelo decrece alrededor de 10 puntos

porcentuales, tanto el F1 score como balanced accuracy aumentan con respecto a otros modelos.

```
> cm_m3_cv
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 326  26
         1 224  76

               Accuracy : 0.6166
                 95% CI : (0.578, 0.6541)
    No Information Rate : 0.8436
    P-Value [Acc > NIR] : 1

                  Kappa : 0.1887

 Mcnemar's Test P-Value : <2e-16

              Precision : 0.2533
                 Recall : 0.7451
                     F1 : 0.3781
             Prevalence : 0.1564
         Detection Rate : 0.1166
   Detection Prevalence : 0.4601
      Balanced Accuracy : 0.6689

       'Positive' Class : 1
```

Ilustración 23. Matriz de confusión y métricas de performance RL - M3 con KCV y oversampling

6.1.4. Resumen performance modelos regresión logística

A continuación se muestra una tabla resumen con las principales métricas de performance obtenidas.

	RL - M1	RL - M2	RL - M3	RL - M3 KCV
Accuracy	75%	80%	80%	62%
Classification error	25%	20%	20%	38%
Recall	53%	29%	29%	75%
Precision	32%	33%	34%	25%
F1 score	0,40	0,31	0,32	0,38
Balanced accuracy	0,66	0,59	0,60	0,67

Tabla 4. Resumen performance modelos regresión logística

Se puede observar que los modelos en los que se utiliza el *validation set approach* obtienen mayor accuracy - por ende, menor error de clasificación - y mayor precision. No obstante, el recall es sensiblemente mayor en el modelo donde se aplica *k-fold cross validation* y *oversampling.* De esta forma, se concluye que el modelo con mejor performance dentro de los evaluados es RL – M3 KCV con un recall de 75% y balanced accuracy de 0,67.

Si bien la performance de los modelos en la predicción de establecimientos infractores debe ser mejorada, dichos modelos sí pueden ser efectivos para comunicar el impacto que tiene la ubicación y determinadas características comerciales en la probabilidad de ser infractor.

6.2. Redes neuronales artificiales (ANN)

La selección de las variables a incluir en las regresiones logísticas surgió de analizar la significancia estadística de cada covariable en el conjunto de datos. En el caso de las ANN se decide estimar diversas estructuras, variando el número de capas y neuronas, que toman como input todas las variables del dataset.

Previo al ajuste de los modelos se realizaron transformaciones a las variables de del dataset. Dichas variables, que en su mayoría representan factores, fueron transformadas a variables dicotómicas generando una nueva columna por cada nivel. A las nuevas variables *dummy* se le asigna el valor 1 en el nivel correspondiente de la observación y 0 en el resto de los niveles. A modo de ejemplo, la variable *departamento* tiene 19 niveles, uno por cada departamento. De esta forma, para los establecimientos ubicados en Montevideo se asigna el valor 1 a la columna Montevideo y el valor 0 a las 18 columnas restantes. Este proceso se realizó para las variables *departamento, tipo_mod_carniceria* y *elaboracion* generando 19, 5 y 3 nuevas columnas respectivamente.

Con respecto a las variables exógenas, éstas se normalizaron mediante la técnica de estandarización *min-max*. El método de estandarización *min-max* consiste en transformar linealmente los datos originales al aplicar la siguiente fórmula,

$$X_i^{'} = \frac{X_i - \min(X)}{\max(X) - \min(X)}, \quad (4.1)$$

en donde X_i es el i-ésimo valor del vector fila X, $\min(X)$ el valor mínimo y $\max(X)$ el valor máximo del vector X. Como resultado, las variables normalizadas ($X^{'}$) toman valores en el rango 0 – 1. Dicha normalización fue aplicada a las variables *población, superficie_km²* e *ing_p_capita_mes_cte_2005*.

Así, el dataset cuenta con 33 predictores correspondientes a las siguientes variables: *departamento, tipo_mod_carniceria, elaboracion, vende_no_carnicos, vende_chacinados, realiza_coccion, poblacion, superficie_km²* e *ing_p_capita_mes_cte_2005*.

Para el entrenamiento de las redes se utiliza la función de pérdida *binary_crossentropy* recomendada para problemas de clasificación y los optimizadores *sgd (stochastic gradient descent)* y *adam (adaptive moment estimation)*. Además, se realizan 100 *epochs* y se divide el conjunto de datos de *train* dejando un 30% de las observaciones como conjunto de validación.

6.2.1. Modelo 1: ANN 1

La primera ANN estimada consta de una capa de entrada con 33 neuronas, 1 para cada predictor, y una capa de salida con una función de activación sigmoidea. Si bien es una estructura simple, ésta permitirá analizar qué variables tienen mayor incidencia en la probabilidad de ser infractor.

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
flatten_1 (Flatten)          (None, 33)                0
_________________________________________________________________
dense_1 (Dense)              (None, 1)                 34
=================================================================
Total params: 34
Trainable params: 34
Non-trainable params: 0
_________________________________________________________________
```

Ilustración 24. Estructura ANN 1

En la imagen siguiente se muestra la evolución de la función de pérdida y accuracy del modelo en el conjunto de entrenamiento (*loss, accuracy*) y validación (*val_loss, val_accuracy*) respectivamente. Los resultados corresponden al entrenamiento de la red

utilizando el optimizador *sgd*. Como puede observarse, la función de pérdida decrece levemente en el conjunto de *train* y se estabiliza cerca de 0,46. La caída en la función de pérdida es más empinada en el conjunto de datos de validación logrando valores cercanos a 0,28 hacia el final del entrenamiento. De la misma forma, el accuracy del modelo aumenta al inicio y se estabiliza en valores cercanos a 0,80 para el conjunto de datos de *train* y 0,95 para el conjunto de datos de validación.

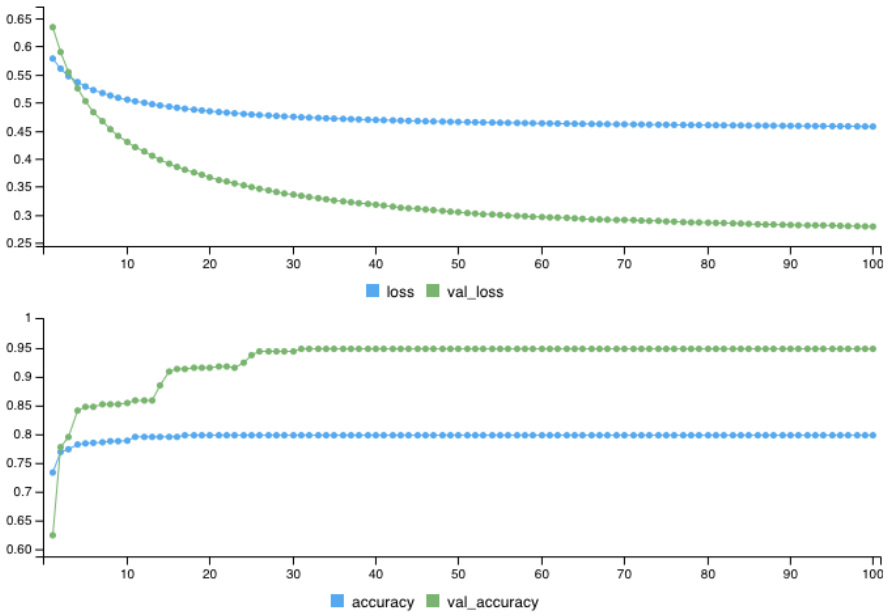


Ilustración 25. Entrenamiento ANN 1 - Optimizador *sgd*

Al evaluar la performance promedio de la red a lo largo de los 100 *epochs* en *train* y *test* se observan resultados similares en ambos conjuntos de datos. Posteriormente, se evalúa la capacidad predictiva de la red en el conjunto de datos de *test.* Para ello se realiza un proceso similar al detallado en los modelos *logit* en donde predice la probabilidad estimada de ser infractor y se asigna dicha categoría a aquellas observaciones con probabilidad estimada mayor o igual a 0,30. Luego se computa la matriz de confusión y métricas asociadas.

	loss	accuracy
train	0,4036	0,8424
test	0,4057	0,8466

Tabla 5. Performance de ANN 1 - Optimizador *sgd* en *train* y *test*

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 504  71
         1  48  29

               Accuracy : 0.8175
                 95% CI : (0.7857, 0.8464)
    No Information Rate : 0.8466
    P-Value [Acc > NIR] : 0.98121

                  Kappa : 0.2242

 Mcnemar's Test P-Value : 0.04372

              Precision : 0.37662
                 Recall : 0.29000
                     F1 : 0.32768
             Prevalence : 0.15337
         Detection Rate : 0.04448
   Detection Prevalence : 0.11810
      Balanced Accuracy : 0.60152

       'Positive' Class : 1
```

Ilustración 26. Matriz de confusión y métricas de performance ANN 1 - Optimizador *sgd*

El modelo presenta un recall de 29%, es decir, detecta a casi 3 de cada 10 infractores reales. Del total de infractores predichos (77) el modelo acierta en el 37,66% de los casos (29). El F1 score y balanced accuracy toman los valores 0,3276 y 0,6015 respectivamente. Los resultados son similares al estimar el modelo utilizando el optimizador *adam* (ver Ilustración 27 y Anexo 7).

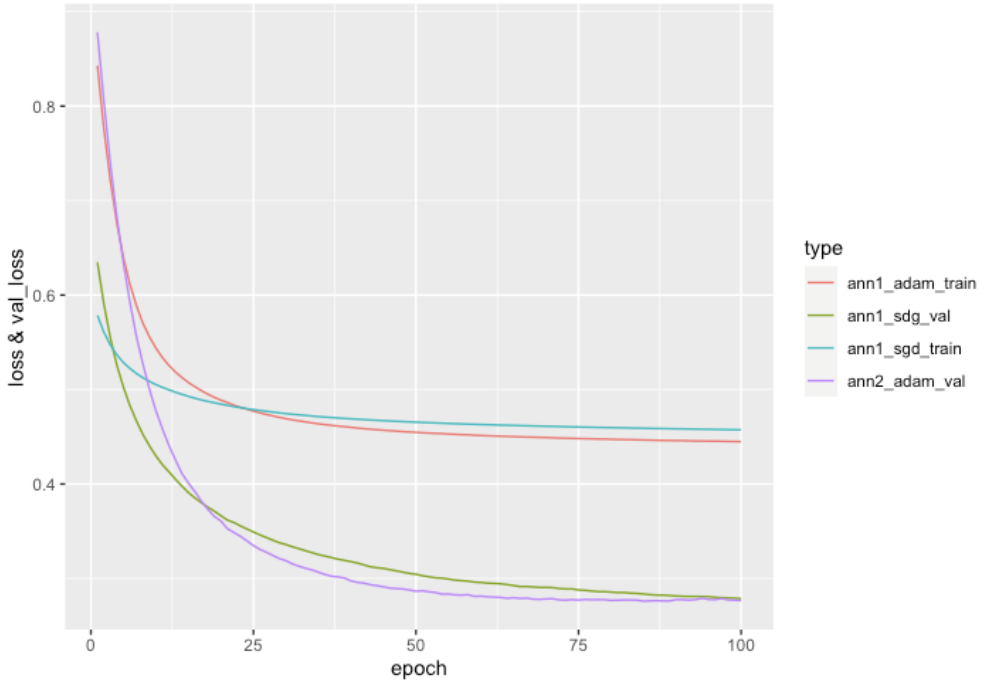


Ilustración 27. *loss* y *val_loss* ANN 1 según optimizador

Por último, en el Anexo 8 se muestran los pesos estimados por cada red según el optimizador utilizado. En ambos casos, los W_i estimados para las carnicerías de tipo Supermercado son positivos, en línea con lo estimado en los modelos *logit.* Esto también sucede en los departamentos con mayor cantidad de infractores como son Montevideo, Artigas y Salto.

6.2.2. Modelo 2: ANN 2

La estructura definida para la segunda ANN es bastante más compleja que la anterior. ANN 2 tiene una capa de entrada con 33 neuronas, 3 capas ocultas – con 100, 50 y 5 neuronas respectivamente y función de activación ReLu – y una capa de salida con función de activación sigmoidea. Producto de esta estructura, la red debe estimar 8.711 parámetros en comparación con los 34 parámetros que se estiman en ANN 1.

Layer (type)	Output Shape	Param #
flatten_3 (Flatten)	(None, 33)	0
dense_6 (Dense)	(None, 100)	3400
dense_5 (Dense)	(None, 50)	5050
dense_4 (Dense)	(None, 5)	255
dense_3 (Dense)	(None, 1)	6

Total params: 8,711
Trainable params: 8,711
Non-trainable params: 0

Ilustración 28. Estructura ANN 2

Al analizar la Ilustración 29, se observa que los resultados son similares a los obtenidos por el modelo anterior a pesar de definir una estructura de red bastante más compleja. Una diferencia con respecto a ANN 1, es la rápidez con la que la red converga a valores estables. Esto sucede en el conjunto de datos de *train* y *validation* en ambas métricas, *loss* y *accuracy.*

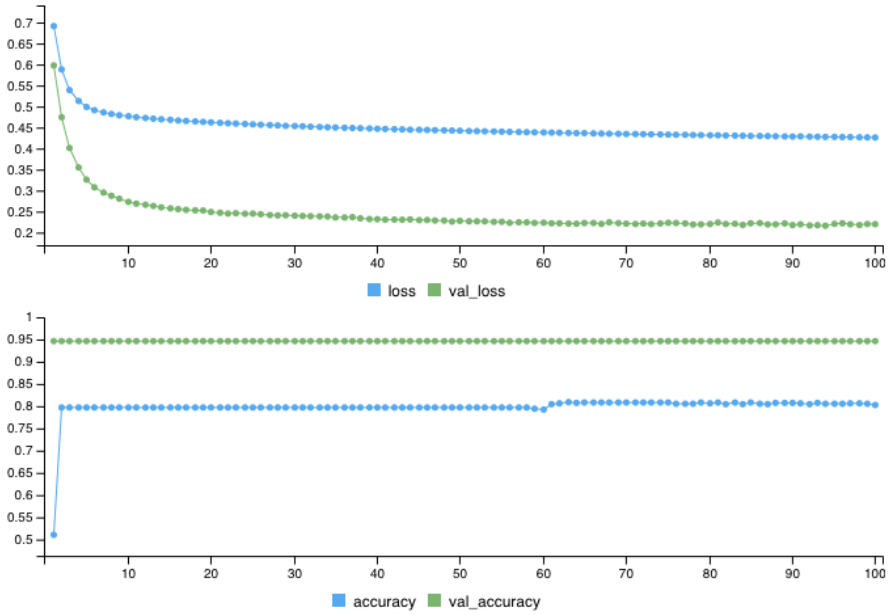


Ilustración 29. Entrenamiento ANN 2 - Optimizador *sgd*

La performance de la red en *train* y *test* se observan en la Tabla 6. Tanto la *loss* como el *accuracy*, es levemente superior en *test* en comparación con *train.* Con respecto al modelo anterior, ANN 2 obtiene una *loss* ligeramente inferior tanto en *train* como en *test.* En

cuanto al accuracy, los resultados obtenidos en ambos modelos son similares. Con respecto a la matriz de confusión, se observa que ANN 2 obtiene mejores resultados en todas las métricas seleccionadas.

	loss	accuracy
train	0,3640	0,8477
test	0,3735	0,8635

Tabla 6. Performance de ANN 2 - Optimizador *sgd* en *train* y *test*

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 489  54
         1  63  46

               Accuracy : 0.8206
                 95% CI : (0.7889, 0.8493)
    No Information Rate : 0.8466
    P-Value [Acc > NIR] : 0.9694

                  Kappa : 0.3336

 Mcnemar's Test P-Value : 0.4595

              Precision : 0.42202
                 Recall : 0.46000
                     F1 : 0.44019
             Prevalence : 0.15337
         Detection Rate : 0.07055
   Detection Prevalence : 0.16718
      Balanced Accuracy : 0.67293

       'Positive' Class : 1
```

Ilustración 30. Matriz de confusión y métricas de performance ANN 2 - Optimizador *sgd*

Los resultados son similares al utilizar el optimizador *adam* (ver Anexo 9). La imagen siguiente muestra que el modelo que utiliza *adam* presenta un proceso más erratico, en donde se constatan picos y valles a lo largo del entrenamiento en el conjunto de validación (*val_loss*).

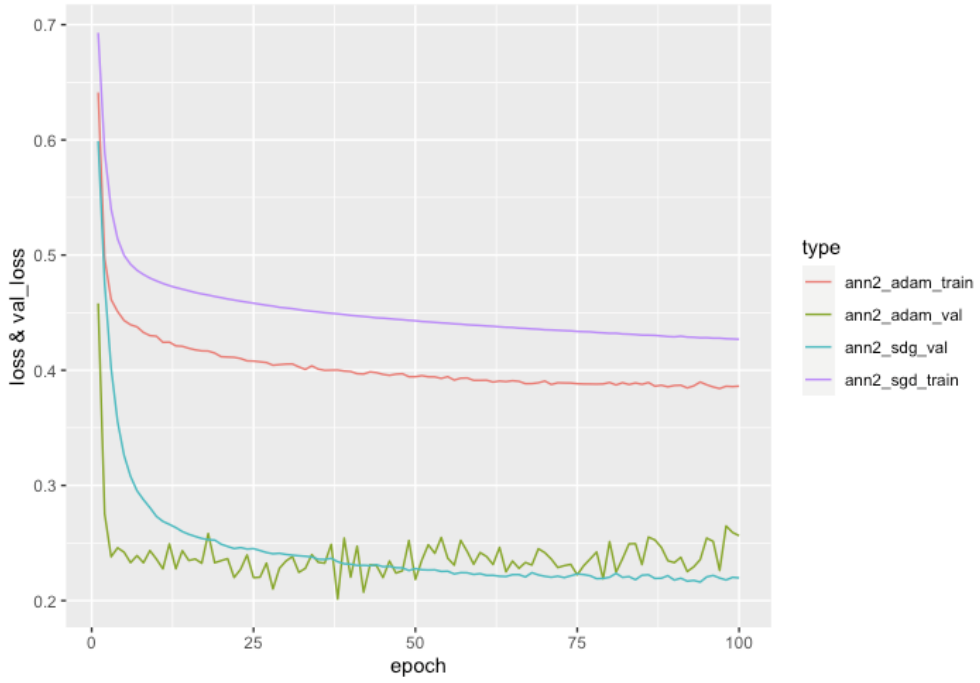


Ilustración 31. loss y val_loss ANN 2 según optimizador

6.2.3. Modelo 3: ANN 3

El último modelo estimado tiene una estructura ubicado en un punto intermedio entre ANN 1 y ANN 2. Se cuenta con la capa de entrada con 33 neuronas, 2 capas ocultas – con 12 y 6 neuronas respectivamente y función de activación ReLu – y la capa de salida con función de activación sigmoidea. Dicha red debe estimar un total de 493 parámetros.

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
flatten_5 (Flatten)          (None, 33)                0
_________________________________________________________________
dense_13 (Dense)             (None, 12)                408
_________________________________________________________________
dense_12 (Dense)             (None, 6)                 78
_________________________________________________________________
dense_11 (Dense)             (None, 1)                 7
=================================================================
Total params: 493
Trainable params: 493
Non-trainable params: 0
_________________________________________________________________
```

Ilustración 32. Estructura ANN 3

Los resultados obtenidos, tanto mediante el optimizador *sgd* como *adam*, se ubican en un punto medio entre el modelo ANN 1 y ANN 2 (ver Anexo 10 y 11).

A pesar de definir redes neuronales con estructuras diferentes, no se observan diferencias significativas en los resultados arrojados por los modelos. Esto hace pensar que dichas estructuras estarían afectando de manera superficial, en caso de hacerlo, el proceso de entrenamiento de las redes. Con el fin de optimizar dicho entrenamiento, se decide incorporar nuevas funciones parametrizables definidas más adelante.

Para definir el modelo sobre el cual se aplicarán dichas técnicas de optimización se evalúan los modelos estimados mediante *sgd* y *adam*. La métrica que se evalúa es la función de pérdida en el conjunto de entrenamiento (*loss*) y validación (*val_loss*).

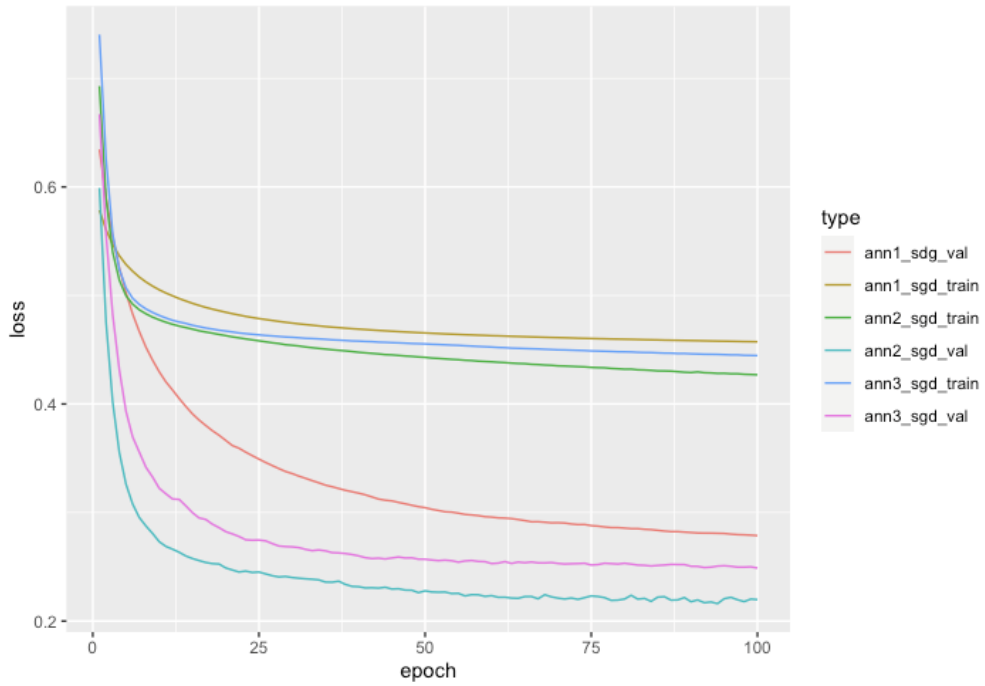


Ilustración 33. Funciones de pérdida estimada mediante ANNs con *sgd optimizer*

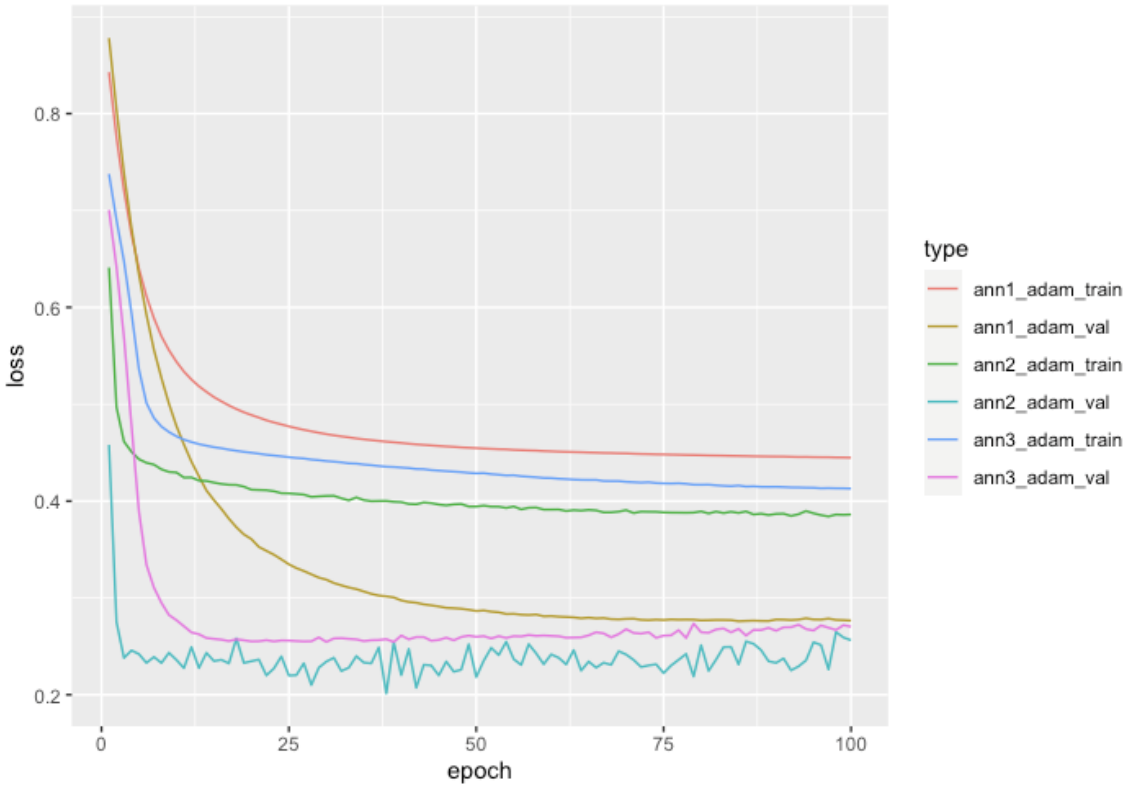


Ilustración 34. Funciones de pérdida estimada mediante ANNs con *adam optimizer*

En la Ilustración 33 y 34 se puede observar que entre las distintas estructuras de redes neuronales estimadas, ANN 2 presenta mejor performance tanto utilizando *sgd* como *adam*. Vale la pena recordar, que dicha red neuronal es la que presenta una estructura más densa y compleja de todos los modelos estimados. A continuación se muestran las funciones de pérdida estimadas para ANN 2 con *sgd* y *adam.* Se observa que la red estimada con *adam* presenta mejores resultados en el conjunto de entrenamiento. Sin embargo, por sus resultados en el conjunto de datos de validación se decide continuar con el modelo ANN 2 con *sgd.*

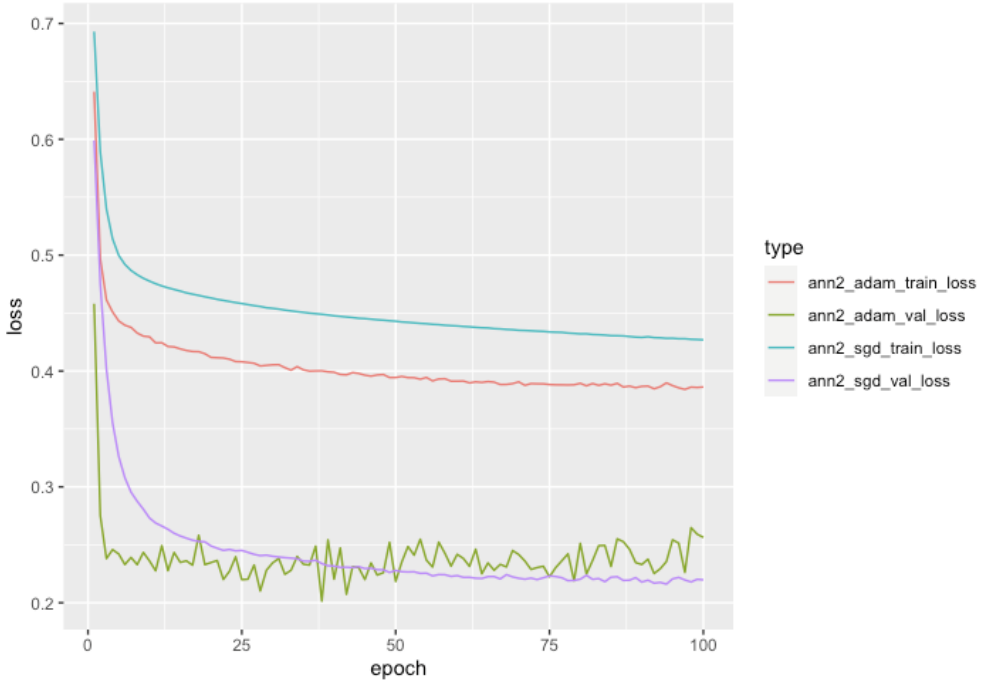


Ilustración 35. Funciones de pérdida estimada ANN 2 según optimizador

A partir de la estructura definida en ANN 2 se decide optimizar el proceso de entrenamiento mediante la inclusión de una reducción de la tasa de aprendizaje (*learning rate*), un *early stopping* y pesos para las clases. El primero, como su nombre lo indica, reduce la tasa de aprendizaje cuando la función de pérdida del conjunto de validación (*val_loss*) no mejora en determinado umbral. En este caso, se multiplica el lr por un factor de 0,75 si *val_loss* no mejora a lo largo de 25 *epochs*. Por su parte, *early stopping* finaliza el proceso de entrenamiento en cuanto *val_loss* no mejora en al menos 0,0001 unidades a lo largo de 60 *epochs*. Además, con el fin de resolver el desbalance de clases se definen pesos entre las categorías que serán utilizados como forma de penalizar errores durante el entrenamiento. De esta forma, se define que la categoría *infractor* tiene un peso 5,4 veces mayor que la categoría *no infractor*, valor que surge de dividir el número total de no infractores (1.835) sobre el total de infractores (340). Por último, dado que se entiende que la estructura definida es compleja para el número de observaciones con el que cuenta el dataset y con el fin de prevenir el sobreajuste del modelo se incorporan dos *dropout layers* con una tasa del 0,10.

Layer (type)	Output Shape	Param #
flatten_10 (Flatten)	(None, 33)	0
dense_31 (Dense)	(None, 100)	3400
dropout_1 (Dropout)	(None, 100)	0
dense_30 (Dense)	(None, 50)	5050
dropout (Dropout)	(None, 50)	0
dense_29 (Dense)	(None, 5)	255
dense_28 (Dense)	(None, 1)	6

Total params: 8,711
Trainable params: 8,711
Non-trainable params: 0

Ilustración 36. Estructura ANN 2 con técnicas de optimización

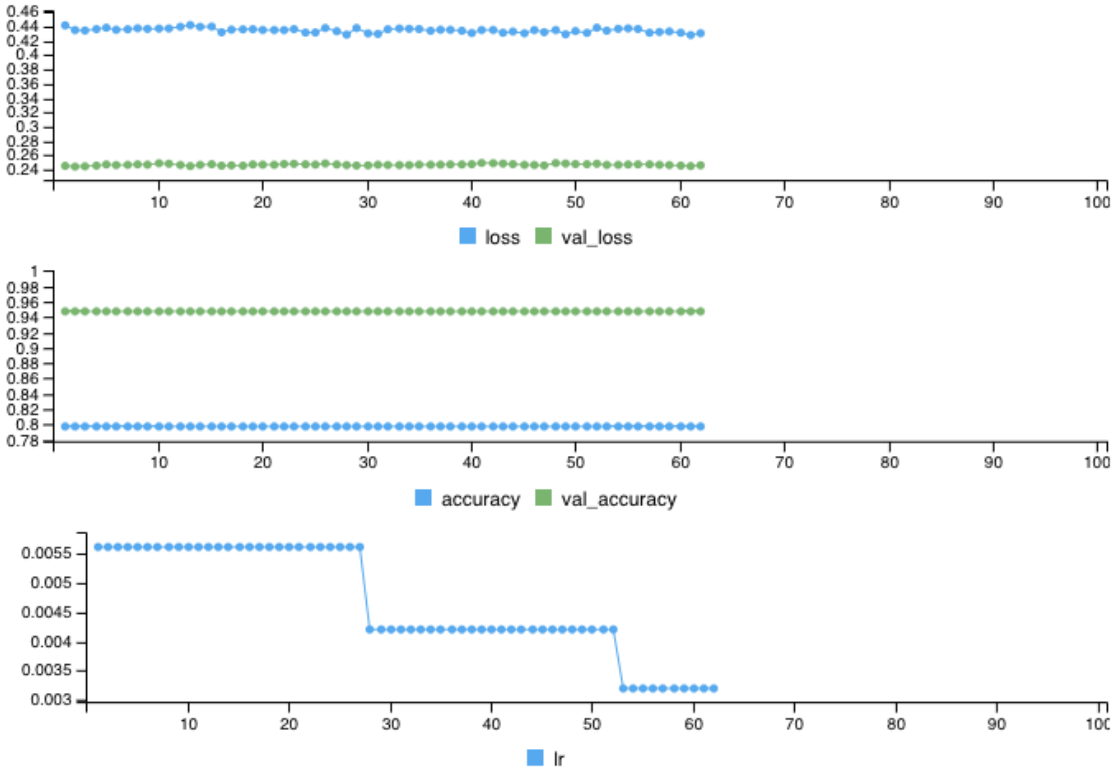


Ilustración 37. Entrenamiento ANN 2 con técnicas de optimización

Como se puede observer en la Ilustración 37, los resultados no son muy distintos a los obtenidos en los modelos anteriores. En este caso se agrega el último gráfico que muestra

la evolución de la tasa de aprendizaje. Se observa que dicha tasa se reduce en dos ocasiones y el proceso de entrenamiento finaliza en el *epoch* 62.

	loss	accuracy
train	0,3733	0,8424
test	0,3813	0,8466

Tabla 7. Performance de ANN 2 con técnicas de optimización

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 489  54
         1  63  46

               Accuracy : 0.8206
                 95% CI : (0.7889, 0.8493)
    No Information Rate : 0.8466
    P-Value [Acc > NIR] : 0.9694

                  Kappa : 0.3336

 Mcnemar's Test P-Value : 0.4595

              Precision : 0.42202
                 Recall : 0.46000
                     F1 : 0.44019
             Prevalence : 0.15337
         Detection Rate : 0.07055
   Detection Prevalence : 0.16718
      Balanced Accuracy : 0.67293

       'Positive' Class : 1
```

Ilustración 38. Matriz de confusión y métricas de performance ANN 2 con técnicas de optimización

6.2.4. Resumen performance modelos redes neuronales

A continuación se muestra una tabla resumen con las principales métricas obtenidas en el conjunto de datos de prueba para las redes neuronales estimadas utilizando *sgd* y *adam*. La estructura de red neuronal que presenta mejores resultados es ANN 2. Al utilizar el optimizador *sgd* se logra la menor función de pérdida mientras que el mayor recall se alcanza al utilizar *adam.*

	sgd				*adam*		
	ANN 1	ANN 2	ANN 2 opt	ANN 3	ANN 1	ANN 2	ANN 3
Accuracy	82%	82%	82%	82%	83%	77%	81%
Classification error	18%	18%	18%	18%	17%	23%	19%
Recall	29%	46%	46%	41%	30%	52%	39%
Precision	38%	42%	42%	41%	44%	33%	39%
F1 score	0,33	0,44	0,44	0,41	0,36	0,41	0,39
Balanced accuracy	0,60	0,67	0,67	0,65	0,62	0,67	0,64
test_loss	0,41	0,37	0,38	0,39	0,40	0,43	0,39

Tabla 8. Resumen performance redes neuronales artificiales

6.3. Árbol de decisión (CART)

Dado el número total de observaciones y el desbalance de clases que presenta el conjunto de datos, se decide estimar un único árbol de decisión que incluya todas las covariables. De esta forma, el modelo estimado incluye: *departamento, tipo_mod_carniceria, vende_no_carnicos, vende_chacinados, elaboracion, realiza_coccion, poblacion, superficie_km2, ing_p_capita_mes_cte_2005* y *frigorificos.*

Al igual que en los algoritmos anteriores, se ajustan los modelos en el conjunto de datos de entrenamiento y se evalúa su performance en el conjunto de datos de prueba. Cabe destacar que la librería utilizada para estimar el árbol, *rpart*, realiza por defecto un procedimiento de *k-fold cross validation* con k=10 en donde obtiene el paramétro de complejidad (*cp*) óptimo del modelo, es decir, aquel que minimiza el error de *cross validation*. A su vez, utiliza el índice de Gini como métrica para evaluar los splits. En la imagen siguiente se muestra el árbol estimado.

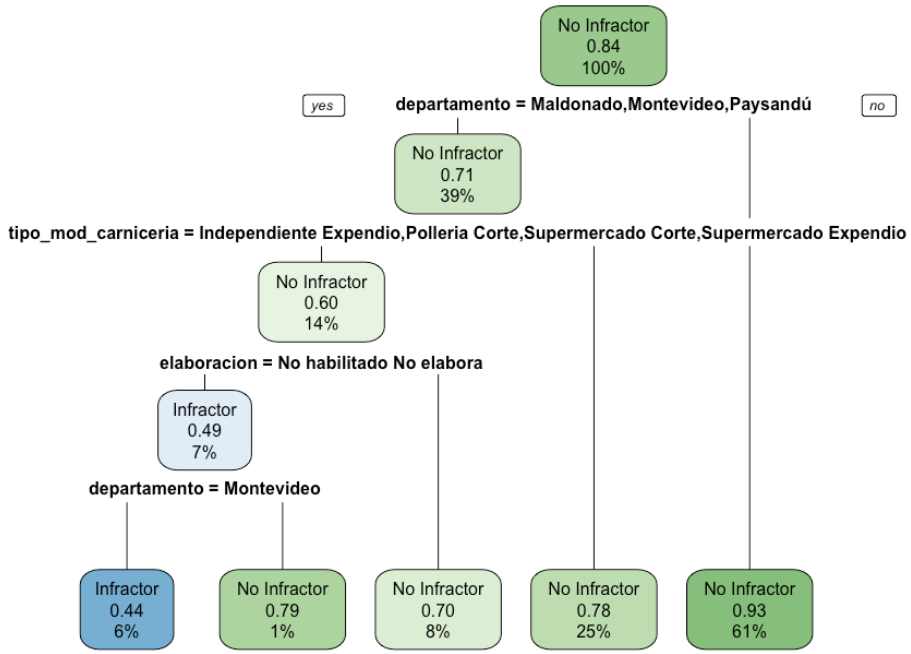


Ilustración 39. Diagrama de árbol de decisión estimado

A modo de entender el diagrama, el primer valor dentro de los cuadros representa la categoría estimada por el modelo, el segundo valor es la probabilidad estimada para las observaciones incluidas en esa región y el último valor es el porcentaje de observaciones que se encuentran en dicha región. De esta forma, el nodo inicial contiene el 100% de las observaciones con una probabilidad estimada de 84% de ser "No Infractor", valores acordes a las proporciones observadas en Ilustración 11.

El primer *split* es realizado por la variable *departamento.* Aquellos establecimientos fuera de Maldonado, Montevideo y Paysandú (la rama de la derecha) representan el 61% de las observaciones del conjunto de entrenamiento y el modelo estima la categoría "No Infractor" con una probabilidad de 93%. A la izquierda, se obtiene la región en donde se encuentran los establecimientos ubicados en Maldonado, Montevideo y Paysandú, para los cuales sin realizar un *split* adicional el modelo estimaría la categoría "No Infractor" con una probabilidad de 71%.

A estas observaciones se aplica el segundo *split* por la variable *tipo_mod_carniceria*, donde el modelo nuevamente estima la categoría "No Infractor" para ambas ramas del árbol, con una probabilidad de 78% para establecimientos del tipo y modalidad

Independiente Corte y 60% para el resto de combinaciones que toma la variable *tipo_mod_carniceria.*

Recién a partir del tercer *split*, el modelo predice la categoría “Infractor” para establecimientos ubicados en Maldonado, Montevideo y Paysandú de algún tipo y modalidad distinto a Independiente Corte que no cuentan con habilitación para elaborar productos y no elaboran productos. En esta rama se encuentran los puntos de venta del tipo Supermercado, por lo que los resultados del árbol son consistentes con lo observado en el análisis exploratorio de datos y modelos anteriores, donde dichos establecimientos tienen una probabilidad mayor de ser infractor. Además, dado que la elaboración de productos se concentra en establecimientos Independiente Corte y estos quedan fuera de dicha rama, los resultados arrojados por el árbol son esperables. Por último, el modelo vuelve a utilizar la variable *departamento* para realizar el *split*, categorizando como infractores a los puntos de venta que cuentan con las características comerciales anteriores y que además se encuentran en Montevideo.

Al analizar la importancia de las distintas variables en el modelo, se observa que *departamento* seguido de las variables exógenas son las más relevantes.

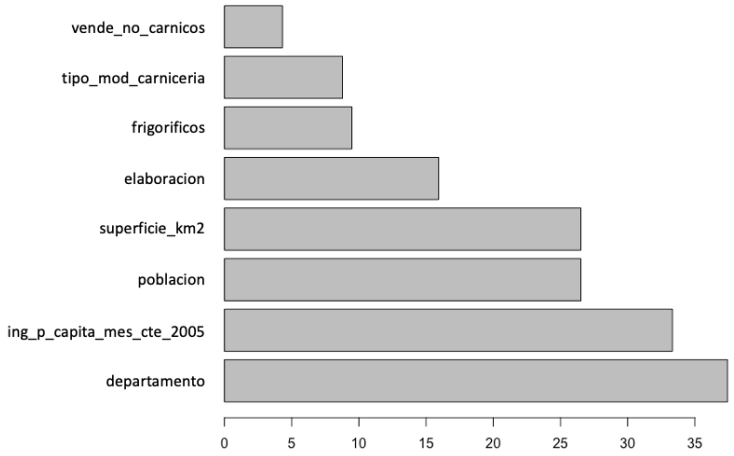


Ilustración 40. Importancia de las variables en el árbol de decisión

	actual	
predicted	Infractor	No Infractor
Infractor	26	13
No Infractor	76	537

Ilustración 41. Matriz de confusión del árbol de decisión en conjunto de *test*

El modelo estimado obtiene una accuracy de 85,09% en *train* y 86,34% en *test.* El recall del modelo es de 25,49%, valor que se encuentra por debajo del obtenido en modelos anteriores.

6.4. Discusión sobre los resultados

Se decidió tomar como modelo base a la regresión logística por sus ventajas a la hora de interpretar y comunicar los resultados. Posteriormente, se estimaron redes neuronales simples (34 parámetros) y complejas (8.711), así como árboles de decisión. En todos los casos se lograron resultados similares, demostrando la consistencia de los mismos independientemente del algoritmo utilizado.

Con respecto a la evaluación de los resultados, se entiende que la capacidad predictiva alcanzada es insuficiente para desarrollar modelos de clasificación eficientes. No obstante, a partir de este análisis se logró comprender la relevancia de determinadas características comerciales y su impacto en la probabilidad de que un establecimiento habilitado para la venta de carne y derivados en Uruguay sea considerado infractor. En este sentido, se destaca la incidencia de variables como la ubicación, el tipo y modalidad de carnicería así como la venta de determinados productos específicos.

7. Propuesta de muestreo

Si bien los modelos de clasificación desarollados en el capítulo anterior ofrecen alguna pista sobre las variables que podrían incidir en la probabilidad de ser infractor, no es posible clasificar con alta certeza los establecimientos infractores.

Es por ello que este capítulo propone utilizar el *score* estimado de cada categoría como indicador de la probabilidad de infracción. En base a ello, se diseña un esquema de muestreo que permite incrementar la capacidad de detección de establecimientos infractores. De esta forma, dado un número de inspecciones finito, éstas se pueden redirigir hacia los puntos de venta con mayor probabilidad de ser infractores.

A continuación se desarrolla la teoría general del procedimiento y posteriormente se aplica en uno de los modelos ajustados previamente.

Probabilidad condicional de infracción

Se supone que se han clasificado los objetos en N categorías disjuntas. A modo de ejemplo, en este trabajo se cuenta con las variables explicativas *departamento* (19 clases) y *tipo_mod_carniceria* (5 clases). Esto implica un total de $N = 19 * 5 = 95$ categorías.

Dada una carnicería X, sea $Z_i = 1_{\{X \in C_i\}}$, la variable que indica si pertenece a la clase $C_i, i = 1, \dots, N$. Los modelos ajustados (*logit*, ANN, CART) se pueden pensar como un estimador de la probabilidad de ser infractor dado que se pertenece a la cateogría i, es decir, si se define

$$p_i = P(X\ infractor \mid X \in C_i) = E[Z_i \mid X \in C_i],$$

entonces el modelo permite estimar dicha probabilidad mediante una función

$$p = f(Z),$$

Con $0 \leq p \leq 1$ y $f(Z) = f(Z_1, \dots, Z_N)$. Si se toma $Z_i = 1$ y las demás 0, entonces se obtiene p_i. La función f es conocida en el caso del *logit* y es del tipo caja negra en una

red neuronal, sin embargo, en todos los casos se puede obtener p_i para cada clase $i = 1, \dots, N$.

Muestreo dirigido

Utilizando la información anterior se pueden generar muestras con mayor probabilidad de detección. Sea X un establecimiento seleccionado del siguiente modo: se sortea con probabilidad q_i un establecimiento de clase i, donde $\sum_i q_i = 1$. Teniendo en cuenta lo mencionado previamente, la probabilidad de que dicho establecimiento sea infractor es,

$$P(X\ infractor) = \sum_{i=1}^{N} P(X\ infractor \mid X \in C_i)\, P(X \in C_i) = \sum_{i=1}^{n} p_i q_i,$$

siendo p_i los predichos por el modelo.

Esto lleva al planteo del siguiente problema de optimización,

$$\max_{q_i} P(X\ infractor),$$

sujeto a

$$q_i \geq 0, \sum_{i=1}^{n} q_i = 1.$$

Dicho problema se traduce en

$$\max_{q_i} \sum_i p_i q_i,$$

sujeto a

$$q_i \geq 0, \sum_{i=1}^{n} q_i = 1.$$

Como un problema de programación lineal tiene uno de sus óptimos necesariamente en un vértice, es sencillo identificar que la solución al problema anterior es elegir

$$q_i = 1 \text{ si } i = \arg\max_j p_j,$$

y 0 en otro caso. Esto implicaría que para maximizar la probabilidad de detección conviene muestrar únicamente a la clase más probable de ser infractor (clases en caso de empate). Este es un extremo no deseable, ya que se pierde el control sobre las clases menos infractoras.

Penalización mediante la entropía

Se propone considerar dos alternativas al modelo de muestreo dirigido anterior. La primera consiste en penalizar la distribución q entre las clases utilizando la función de entropía. La entropía de un vector q se define como,

$$H(q) = -\sum_i q_i \log q_i.$$

Esta función es cóncava y se maximiza para la distribución uniforme. Resolviendo,

$$\max H(q)$$

sujeto a $q_i \geq 0, \sum_i q_i = 1$. De esta forma, el Lagrangeano queda de la siguiente manera,

$$\mathcal{L}(q, \lambda) = -\sum_i q_i \log q_i + \lambda(\sum_i q_i - 1)$$

Derivando respecto a q_i, se obtiene

$$\frac{\partial \mathcal{L}}{\partial q_i} = -\log q_i - 1 + \lambda$$

donde el punto estacionario es $q_i = e^{\lambda - 1}$ para todo i, lo que implica que todas las probabilidades son iguales, e iguales a $1/N$ por la restricción. El multiplicador queda $\lambda = \log \frac{1}{N} + 1$. En términos de muestreo, esto corresponde a un muestreo estratificado donde cada clase se ve igualmente representada en la muestra.

Si solo se maximiza la entropía de q, corresponde a tomar un establecimiento de cada clase al azar. Por el contrario, si solo se maximiza la probabilidad de detección solo se tomarían establecimientos de la clase más infractora. Es así, que resulta interesante plantear el siguiente problema,

$$\max_{q_i} \sum_i p_i q_i + \beta H(q),$$

sujeto a

$$q_i \geq 0, \sum_{i=1}^{n} q_i = 1,$$

donde β es un término de penalización o *trade-off* entre ambas componentes. El Langrangeano de este problema queda de la siguiente forma,

$$\mathcal{L}(q, \lambda) = \sum_i p_i q_i - \beta \sum_i q_i \log q_i + \lambda(\sum_i q_i - 1)$$

Derivando respecto a q_i, se obtiene

$$\frac{\partial \mathcal{L}}{\partial q_i} = p_i - \beta(\log q_i - 1) + \lambda$$

Donde el punto estacionario es $q_i = e^{\frac{p_i + \lambda}{\beta} - 1}$. Ahora q_i es creciente con p_i por lo que se tenderá a muestrear más a las clases más infractoras. Sin embargo, el efecto de p_i se ve restringido por el valor de β, ya que si $\beta \to \infty$, el multiplicador se ajusta para que converga a la distribución uniforme de clases.

Ejemplo

Tomando como base la regesión logística con la variable *departamento* y *tipo_mod_carniceria,* se estiman las probabilidades estimadas de ser infractor para cada una de las 95 categorías detalladas previamente. La tabla siguiente muestra el top 10 de categorías con mayor probabilidad de ser infractoras.

Categoría	Probabilidad estimada infractor = 1
Supermercado Expendio Montevideo	52,7%
Independiente Expendio Montevideo	43,6%
Supermercado Corte Montevideo	42,2%
Supermercado Expendio Maldonado	40,4%
Supemercado Expendio Paysandú	36,4%
Supermercado Expendio Artigas	35,6%
Supermercado Expendio Florida	33,0%
Independiente Expendio Maldonado	31,9%
Supermercado Corte Maldonado	30,7%
Pollería Corte Montevideo	30,7%

Tabla 9. Top 10 categorías con mayor probabilidad estimada de ser infractor

En línea con lo observado en los modelos previos, los establecimientos del tipo Supermercado son los que obtienen mayor probabilidad estimada de cometer una infracción. Además, la ubicación corresponde con los departamentos con mayor porcentaje de establecimientos infractores.

Una vez estimadas las probabilidades para categoría, se resuelve el problema de optimización planteado. Se parte de una probabilidad de muestreo uniforme de 1,05% (1/95) para cada categoría y se define $\beta = 0{,}5$. La siguiente tabla muestra las probabilidades de muestreo obtenidas para las categorías de la tabla anterior.

Categoría	Probabilidad estimada infractor = 1	Probabilidad muestreo (qi) beta = 0,5
Supermercado Expendio Montevideo	52,7%	2,30%
Independiente Expendio Montevideo	43,6%	1,92%
Supermercado Corte Montevideo	42,2%	1,87%
Supermercado Expendio Maldonado	40,4%	1,80%
Supemercado Expendio Paysandú	36,4%	1,66%
Supermercado Expendio Artigas	35,6%	1,64%
Supermercado Expendio Florida	33,0%	1,55%
Independiente Expendio Maldonado	31,9%	1,52%
Supermercado Corte Maldonado	30,7%	1,48%
Pollería Corte Montevideo	30,7%	1,41%

Tabla 10. Top 10 categorías con mayor probabilidad estimada de ser infractor y probabilidad de muestreo asociada

Se observa que la probabilidad de muestreo estimada es superior a la establecida por defecto (1,05%) para las clases con mayor probabilidad de infracción. Dentro de las 95 categorías disponibles, las que presentan mayor probabilidad de muestreo son Supermercado Expendio Montevideo con 2,30%, Independiente Expendio Montevideo con 1,92%, Supermercado Corte Montevideo con 1,87%, Supermercado Expendio Maldonado con 1,80% y Supermercado Expendio Paysandú con 1,66%.

Por último, esta propuesta presenta una probabilidad de detección de infractores de 15,26%, en donde la probabilidad de muestrear a la categoría menos muestreada es de 0,80%.

Vale la pena mencionar que si bien en este ejemplo se utiliza la regresión logística, este método puede ser aplicado al resto de los algoritmos utilizados en este trabajo.

Muestreo proporcional

Este segundo método propone optimizar la siguiente función para repartir el muestreo entre clases.

$$\max_{q_i} \sum_i p_i \log(q_i),$$

sujeto a

$$q_i \geq 0, \sum_{i=1}^{n} q_i = 1$$

En teoría de asignación de recursos esto se conoce como *proportional fairness* o reparto proporcional y presenta buenas propiedades. En particular, si se resuleve el siguiente Langrangeno,

$$\mathcal{L}(q, \lambda) = \sum_i p_i \log(q_i) + \lambda(\sum_i q_i - 1)$$

Derivando con respecto a q_i, se obtiene

$$\frac{\partial \mathcal{L}}{\partial q_i} = \frac{p_i}{q_i} + \lambda$$

Donde el punto estacionario es $q_i = \left(-\frac{1}{\lambda}\right) p_i$. Vale recordar que λ puede ser negativo y lo será en este caso. Esto es, se elige cada clase con probabilidad proporcional a probabilidad de ser infractor de la clase. Al imponer la restricción,

$$1 = \sum_{i=1}^{N} q_i = -\left(\frac{1}{\lambda}\right) \sum_{i=1}^{N} p_i$$

surge $\lambda = -\sum_{i=1}^{N} p_i$ y la probabilidad de muestrear la clase i óptima queda de la siguiente forma,

$$q_i^* = \frac{p_i}{\sum_{j=1}^{N} p_j}$$

Ejemplo

Partiendo del modelo detallado previamente, se resuelve el problema de optimización planteado para este método.

Al igual que en el caso anterior se parte de una probabilidad de muestreo uniforme (1,05%) para cada categoría. La siguiente tabla muestra las probabilidades de muestreo obtenidas mediante los métodos de penalización por entropía y muestreo proporcional.

Categoría	Probabilidad estimada infractor = 1	Probabilidad muestreo Penalización por entropía	Probabilidad muestreo Proporcional
Supermercado Expendio Montevideo	52,7%	2,30%	4,58%
Independiente Expendio Montevideo	43,6%	1,92%	3,78%
Supermercado Corte Montevideo	42,2%	1,87%	3,68%
Supermercado Expendio Maldonado	40,4%	1,80%	3,51%
Supemercado Expendio Paysandú	36,4%	1,66%	3,17%
Supermercado Expendio Artigas	35,6%	1,64%	3,11%
Supermercado Expendio Florida	33,0%	1,55%	2,87%
Independiente Expendio Maldonado	31,9%	1,52%	2,78%
Supermercado Corte Maldonado	30,7%	1,48%	2,68%
Pollería Corte Montevideo	30,7%	1,48%	2,68%

Tabla 11. Top 10 categorías con mayor probabilidad estimada de ser infractor y probabilidades de muestreo asociadas

Nuevamente la probabilidad de muestreo que surge de resolver el problema de optimización es superior a la establecida por defecto para las clases con mayor probabilidad de infracción. Dado que este método elige la clase a muestrear con probabilidad proporcional a la probabilidad de ser infractor, se observa un aumento en la probabilidad de muestreo para todas las categorías incluidas en el top 10.

Por último, la probabilidad de detección de infractores de este método es de 23,51%, donde la probabilidad de muestrear a la categoría menos muestreada es de $3{,}76 \times 10^{-9}$ %.

8. Conclusiones

El trabajo realizado demuestra que es posible generar modelos de aprendizaje automático supervisado para la predicción de carnicerías y locales de venta al público infractores en el mercado dómestico uruguayo de carnes y sus derivados.

Este análisis representa una primera aproximación a este problema y permite comprender la relevancia de determinadas características comerciales y su impacto en la probabilidad de cometer una infracción. En este sentido, se destaca la incidencia de variables como la ubicación, el tipo y modalidad de carnicería así como la venta de determinados productos específicos.

Sin embargo, la capacidad predictiva de los modelos desarrollados es insuficiente para clasificar establecimientos infractores de manera certera. Con el fin de mejorar este punto, se realizan los siguientes comentarios vinculados a los datos y al proceso de estimación.

Con respecto al conjunto de datos utilizado, se entiende que el mismo debe contar con un mayor número de observaciones y predictores. Dado que la base de datos recoge la mayoría de las carnicerías del país, una forma de aumentar el número de observaciones podría ser dejando de lado los establecimientos como unidad de análisis y monitorear la actividad comercial y flujos asociados. Esto será posible a partir de la puesta en marcha del SRGA, donde se relevarán datos a nivel de transacción.

Otro aspecto a considerar es la vinculación entre las infracciones labradas por los equipos inspectivos y los establecimientos. Este trabajo resolvió de manera parcial este problema mediante un id generado a partir de la ubicación y titularidad de los establecimientos. Sin embargo, dicha vinculación es subóptima ya que deja fuera las infracciones labradas a personas físicas e introduce potenciales sesgos en el conjunto de datos que podrían estar afectando los resultados. Identificar y considerar únicamente los tipos de infracciones graves podría representar mejoras en el ajuste de los modelos. Por último, implementar técnicas de muestreo como las planteadas en este trabajo podría contribuir de manera positiva en la asignación de los recursos inspectivos en el territorio.

9. Referencias bibliográficas

[1] Instituto Nacional de Carnes, “Decreto - Ley N° 15.605 del 27.07.1984”. [Online]. Available: Decreto - Ley N° 15.605 del 27.07.1984 (inac.uy)

[2] Instituto Nacional de Carnes, “Sistema de Registro y Gestión del Abasto (SRGA): Caso de negocio, modelos conceptuales y propuesta de nuevo sistema. Documento Funcional versión 3.1”, unpublished.

[3] Instituto Nacional de Carnes, “Algunas definiciones prácticas”. [Online]. Available: Algunas definiciones prácticas (inac.uy)

[4] Dirección Nacional de Impresiones y Publicaciones Oficiales, “Ley N° 19.783 del 23.08.2019”. [Online]. Available: Ley N° 19.783 del 23.08.2019 (impo.com.uy)

[5] Dirección Nacional de Impresiones y Publicaciones Oficiales, “Ley N° 19.889 del 09.07.2020”. [Online]. Available: Ley N° 19.889 del 09.07.2020 (impo.com.uy)

[6] D. Weisburd and L. Green, “Policing drug hot spots: The Jersey City drug market analysis experiment”, *Justice Quarterly*, vol. 12 (4), 1995.

[7] W. Hardyns and A. Rummens, “Predictive Policing as a New Tool for Law Enforcement? Recent Developments and Challenges”, *Eur J Crim Policy Res*, 2017.

[8] A. Rummens, W. Hardyns and L. Pauwels, “The use of predictive analysis in spatiotemporal crime forecasting: Building and testing a model in an urban context”, *Applied Geography*, vol. 86, 2017.

[9] X. Zhang, L. Liu, L. Xiao and J. Ji, “Comparison of Machine Learning Algorithms for Predicting Crime Hotspots”, *IEEE Access*, vol. 8, 2020.

[10] S. Jaffee, S. Henson, L. Unnevehr, D. Grace and E. Cassou, “The Safe Food Imperative: Accelerating Progress in Low and Middle-Income Countries”, *Agriculture and Food Series*, World Bank, 2019.

[11] News Desk, “2011 Outbreak of Rare E. Coli Strain was Costly for Europe”, *Food Safety News*, Apr., 2015 [Online]. Available:

https://www.foodsafetynews.com/2015/04/2011-outbreak-of-rare-e-coli-strain-was-costly-for-europe/

[12] World Health Organization, "The burden of foodborne diseases in the WHO European region", 2017.

[13] S. Hoffmann, B. Maculloch, and M. Batz, "Economic Burden of Major Foodborne Illnesses Acquired in the United States", U.S. Department of Agriculture, Economic Research Service, EIB-140, May 2015.

[14] H. Marvin, E. Janssen, Y. Bouzembrak, P. Hendriksen and M. Staats, "Big data in food safety: An overview", *Critical Reviews in Food Science and Nutrition*, vol. 57 (11), 2286-2295, 2017.

[15] M. Gertler, I. Czogiel, K. Stark and H. Wilking, "Assessment of recall error in self-reported food consumption histories among adults—Particularly delay of interviews decrease completeness of food histories—Germany, 2013", *PLoS ONE*, vol. 12 (6): e0179121, 2017.

[16] P. Zhang, W. Cui, H. Wang, Y. Du and Y. Zhou, "High-Efficiency Machine Learning Method for Identifying Foodborne Disease Outbreaks and Confounding Factors", *Foodborne Pathogens and Disease*, vol. 18 (8), 590-598, 2021.

[17] H. Wang, W. Cui, Y. Guo, Y. Du and Y. Zhou, "Machine Learning Prediction of Foodborne Disease Pathogens: Algorithm Development and Validation Study", *Journal of Medical Internet Research*, vol. 9 (1), 2021.

[18] J. Harris, R. Mansour, B. Choucair, J. Olson, C. Nissen and J. Bhatt, "Health department use of social media to identify foodborne illness - Chicago, Illinois, 2013-2014.", *Morbidity and Mortality Weekly Report*, vol. 63 (32), 681-685, 2014.

[19] B. Kuehn, "Agencies Use Social Media to Track Foodborne Illness", *Journal of the American Medical Association*, vol. 312 (2), 117–118, 2014.

[20] A. Sadilek, H. Kautz, L. DiPrete, B. Labus, E. Portman, J. Teitel and V. Silenzio, "Deploying NEmesis: Preventing Foodborne Illness by Data Mining Social Media", *AI Magazine*, vol. 38, no. 1, 37-48, 2017.

[21] K. Devinney, A. Bekbay, T. Effland, L. Gravano, D. Howell, D. Hsu, D. O'Hallorhan, V. Reddy, F. Stavinsky, H. Waechter and B. Gutelius, "Evaluating Twitter for Foodborne Illness Outbreak Detection in New York City", *Online Journal of Public Health Informatics*, vol. 10 (1), 2018.

[22] J. Kang, P. Kuznetsova, M. Luca, and Y. Choi, "Where not to eat? Improving public policy by predicting hygiene inspections using online reviews", *Proc. Conf. Empirical Methods Natural Language Processing*, WA, 2013, pp. 1443-1448.

[23] C. Harrison, M. Jorder, H. Stern, F. Stavinsky, V. Reddy, H. Hanson, H. Waechter, L. Lowe, L. Gravano and S. Balter, "Using Online Reviews by Restaurant Patrons to Identify Unreported Cases of Foodborne Illness—New York City, 2012–2013.", *Morbidity and Mortality Weekly Report*, vol. 63 (20), 441-445, 2014.

[24] J. Schomberg, O. Haimson, G. Hayes and H. Anton-Cluver, "Supplementing Public Health Inspection via Social Media", *PLoS ONE*, vol. 11 (3): e0152117, 2016.

[25] A. Maharana, K. Cai, J. Hellerstein, Y. Hswen, M. Munsell, V. Staneva, M. Verma, C. Vint, D. Wijaya and E. O Nsoesie, "Detecting Reports of Unsafe Foods in Consumer Product Reviews", *Journal of the American Medical Informatics Open*, vol. 2, 2019.

[26] T. Jordan, B. Pavlin, B. Lafleur, L. Amanda Ingram and W. Schaffner, "Restaurant Inspection Scores and Foodborne disease", *Emerging Infectious Diseases*, vol. 10, no. 4, April 2004.

[27] K. Morrison and J. Wong, "Predicting severe food safety violations in Toronto, Ontario", *Computer Science 598 Applied Machine Learning*, McGill University, 2014.

[28] City of Chicago, "Food Inspection Forecasting: Optimizing inspections with analytics". [Online]. Available: https://chicago.github.io/food-inspections-evaluation/

[29] A. Sadilek, S. Caty, L. DiPrete, R. Mansour, T. Schenk Jr., M. Bergtholdt, A. Jha, P. Ramaswami and E. Gabrilovich, "Machine-learned epidemiology: real-time detection of foodborne illness at scale", *npj Digital Med*, vol. 1, 36, 2018.

[30] Instituto Nacional de Carnes, "Reglamento Nacional de Carnicerías". [Online] Available: Reglamento Nacional de Carnicerías (inac.uy)

[31] Instituto Nacional de Carnes, "Listado de productos no cárnicos". [Online] Available: Listado de productos no cárnicos (inac.uy)

[32] G. James, D. Witten, T. Hastie and R. Tibshirani, "*Logistic regression*", in *An Introduction to Statistical Learning with applications in R*. New York: Springer, 2017.

[33] T. Mitchell, "*Artificial Neural Network*", in *Machine Learning*. New York: McGraw-Hill, 1997.

[34] G. James, D. Witten, T. Hastie and R. Tibshirani, "*Tree based methods*", in *An Introduction to Statistical Learning with applications in R*. New York: Springer, 2017.

10. Anexos

Anexo 1 – Listado con tipos de infracciones

Código infracción	Detalle de infracción	Frecuencia	Porcentaje	Porcentaje acumulado
1462	Actividades no autorizadas en carnicerías o pollerías.	159	23,9%	23,9%
1505	Tenencia de carne sin documentación y/o sin identificación para su comercialización en carnicerías.	96	14,4%	38,3%
1701	No confección de guías.	60	9,0%	47,3%
1507	Existencia o venta en carnicerías y otros comercios de productos cárnicos sin documentación o chacinados sin justificar procedencia, carentes de etiquetas, etc.	49	7,4%	54,7%
1604	Transporte en vehiculo no habilitado, o con la habilitación vencida o suspendida.	39	5,9%	60,5%
1433	Carne en mal estado en local autorizado.	30	4,5%	65,0%
1428	Venta de otros productos no autorizados o violación de las condiciones de autorización.	22	3,3%	68,3%
1809	Incumplimiento a resolución de INAC, Decreto Ley 15.605.	21	3,2%	71,5%
1402	Carnicería o pollería no habilitada.	18	2,7%	74,2%
1801	Empresa no inscripta en los Registros a cargo de INAC o con inscripción cancelada.	17	2,6%	76,7%
1426	No confecciona Declaración Jurada, atraso y/o irregularidades.	15	2,3%	79,0%
1802	Empresa trabajando con registros de INAC suspendidos.	15	2,3%	81,2%
1106	No cumplimiento de Declaración Jurada	12	1,8%	83,0%
1521	Trasiego de carne.	10	1,5%	84,5%
1601	Depósito no habilitado y/o suspendido.	10	1,5%	86,0%
1401	Venta de carne y/o menudencias en comercio no Carnicería.	9	1,4%	87,4%
1614	Transporte sin guía o con guía en situación irregular.	9	1,4%	88,7%
1411	Carnicería trabajando con habilitacion vencida.	8	1,2%	89,9%
1421	Modificación o incumplimiento de exigencias locativas y de funcionamiento del comercio, sus equipos o instalaciones.	7	1,1%	91,0%
1432	Falta de higiene.	6	0,9%	91,9%
1467	Elaboracion de productos cárnicos sin autorizacion de INAC (milanesas, etc.).	6	0,9%	92,8%
1413	No realizar cambio de titularidad de carnicería.	5	0,8%	93,5%
1624	Transporte de productos cárnicos sin documentación y/o identificacion que acredite su origen.	4	0,6%	94,1%
1451	Venta de productos no cárnicos que no reúnen las condiciones reglamentarias.	3	0,5%	94,6%
1529	Suministro de mercadería a comercio no habilitado a tal efecto (distribuidor a pollería, almacén, etc).	3	0,5%	95,0%
1702	Irregularidad en la confección de gúías.	3	0,5%	95,5%
1813	Incumplimiento de la Resolución de INAC 20/110 del 21.09.2020 referente a la normativa aplicable al Control de Taras de Roldanas en la Industria Frigorífica	3	0,5%	95,9%
1470	Venta de carne cuya fecha de vencimiento expiró.	2	0,3%	96,2%
1514	Elaboración de productos cárnicos para su comercialización en establecimiento no habilitado.	2	0,3%	96,5%
1620	Trasiego de carne en la vía pública.	2	0,3%	96,8%
1805	No presentación de declaración jurada de 0,7%,	2	0,3%	97,1%
1807	Entorpecimiento a las tareas inspectivas o al cumplimiento de las funciones del Instituto.	2	0,3%	97,4%
1301	Incumplimiento en exportación.	1	0,2%	97,6%
1309	Exportación no cubierta por certificado oficial de calidad comercial (total o parcial).	1	0,2%	97,7%
1403	No exhibición al público lista de precios.	1	0,2%	97,9%
1457	Elaboración de chacinados o productos cárnicos embutidos.	1	0,2%	98,0%
1516	Suministro a carnicerías de carne y/o menudencia no tipificada y/o autorizada para abasto.	1	0,2%	98,2%
1526	Existencia o venta en carnicería y otros comercios de subproductos cárnicos sin documentación y/o sin justificar procedencia, etc.	1	0,2%	98,3%
1527	Venta de carne cuya fecha de vencimiento expiró	1	0,2%	98,5%
1531	Elaboración de subproductos cárnicos para su comercialización en establecimiento no habilitado.	1	0,2%	98,6%
1532	Entrega de mercadería por establecimiento de faena a Deposito no habilitado.	1	0,2%	98,8%
1619	Transporte simultáneo de carne y otros productos.	1	0,2%	98,9%
1703	No entrega de guías.	1	0,2%	99,1%
1704	Entrega de guías fuera de plazo.	1	0,2%	99,2%
1707	Utilización de categoría de guías incorrecta.	1	0,2%	99,4%
1709	Utilización de guías ajenas.	1	0,2%	99,5%
1713	No hacer constar el nùmero de la guía en la documentación comercial.	1	0,2%	99,7%
1804	Infracciones reglamentarias prestacion 0,7%,	1	0,2%	99,8%
1812	No cumplir con las obligaciones de depositario.	1	0,2%	100,0%
	Total	**666**	**100%**	

Anexo 2 – Contribución de la variable *departamento* en la regresión logística al ajuste del modelo

```
> logit <- glm(infractor ~ departamento, data = pais_logit, family = binomial)
> summary(logit)

Call:
glm(formula = infractor ~ departamento, family = binomial, data = pais_logit)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8818  -0.6203  -0.4246  -0.1709   2.9246

Coefficients:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)               -1.9132     0.1694 -11.295  < 2e-16 ***
departamentoArtigas        0.4469     0.4835   0.924  0.35537
departamentoCerro Largo   -2.3063     1.0215  -2.258  0.02396 *
departamentoColonia       -0.9611     0.4238  -2.268  0.02334 *
departamentoDurazno       -0.4476     0.4976  -0.900  0.36832
departamentoFlores        -15.6528   807.5519  -0.019  0.98454
departamentoFlorida        0.3626     0.3395   1.068  0.28547
departamentoLavalleja     -1.3826     0.7397  -1.869  0.06162 .
departamentoMaldonado      0.7757     0.2452   3.163  0.00156 **
departamentoMontevideo     1.1692     0.1918   6.095 1.09e-09 ***
departamentoPaysandú       0.3038     0.3088   0.984  0.32521
departamentoRío Negro     -0.9771     0.7460  -1.310  0.19023
departamentoRivera        -2.1643     1.0226  -2.117  0.03430 *
departamentoRocha         -1.0312     0.5402  -1.909  0.05629 .
departamentoSalto         -0.3204     0.3895  -0.822  0.41080
departamentoSan José      -15.6528   401.6893  -0.039  0.96892
departamentoSoriano       -0.8593     0.4911  -1.750  0.08015 .
departamentoTacuarembó    -2.3494     1.0212  -2.301  0.02141 *
departamentoTreinta y Tres -1.8244    1.0259  -1.778  0.07535 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1885.8  on 2174  degrees of freedom
Residual deviance: 1634.2  on 2156  degrees of freedom
AIC: 1672.2

Number of Fisher Scoring iterations: 16

> anova(logit, test = "Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: infractor

Terms added sequentially (first to last)


             Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                          2174     1885.8
departamento 18   251.57      2156     1634.2 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

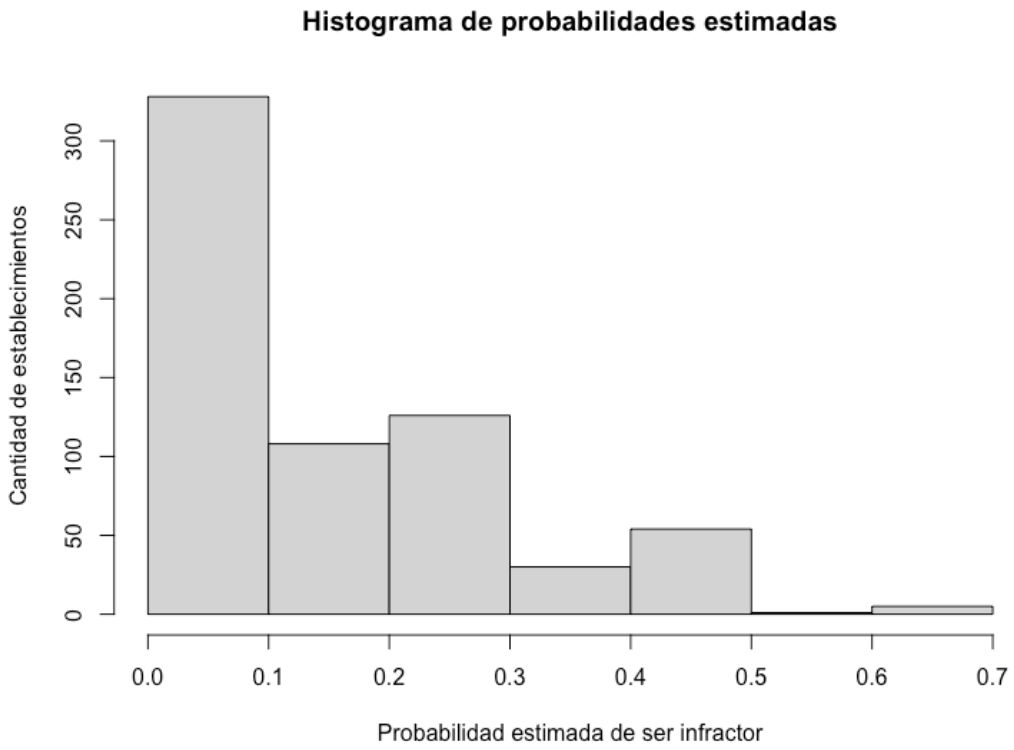
Anexo 3 – Regresión logística con variables asociados a características comerciales

```
> summary(logit_carniceria)

Call:
glm(formula = infractor ~ tipo_mod_carniceria + elaboracion +
    vende_no_carnicos + vende_chacinados, family = binomial,
    data = pais_logit)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2004  -0.6123  -0.4503  -0.3271   2.4309

Coefficients:
                                        Estimate Std. Error z value Pr(>|z|)
(Intercept)                             -1.88924    0.18183 -10.390  < 2e-16 ***
tipo_mod_carniceriaIndependiente Expendio  1.00519    0.51456   1.954 0.050759 .
tipo_mod_carniceriaPolleria Corte          0.15893    0.63015   0.252 0.800876
tipo_mod_carniceriaSupermercado Corte      1.04152    0.17945   5.804 6.47e-09 ***
tipo_mod_carniceriaSupermercado Expendio   1.54159    0.42067   3.665 0.000248 ***
elaboracionHab Elabora                     0.04855    0.16411   0.296 0.767337
elaboracionNo_hab Elabora                 -1.01179    0.15990  -6.328 2.49e-10 ***
vende_no_carnicosVende                     0.26177    0.19680   1.330 0.183479
vende_chacinadosVende                      0.40162    0.17305   2.321 0.020299 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1885.8  on 2174  degrees of freedom
Residual deviance: 1752.7  on 2166  degrees of freedom
AIC: 1770.7

Number of Fisher Scoring iterations: 5
```

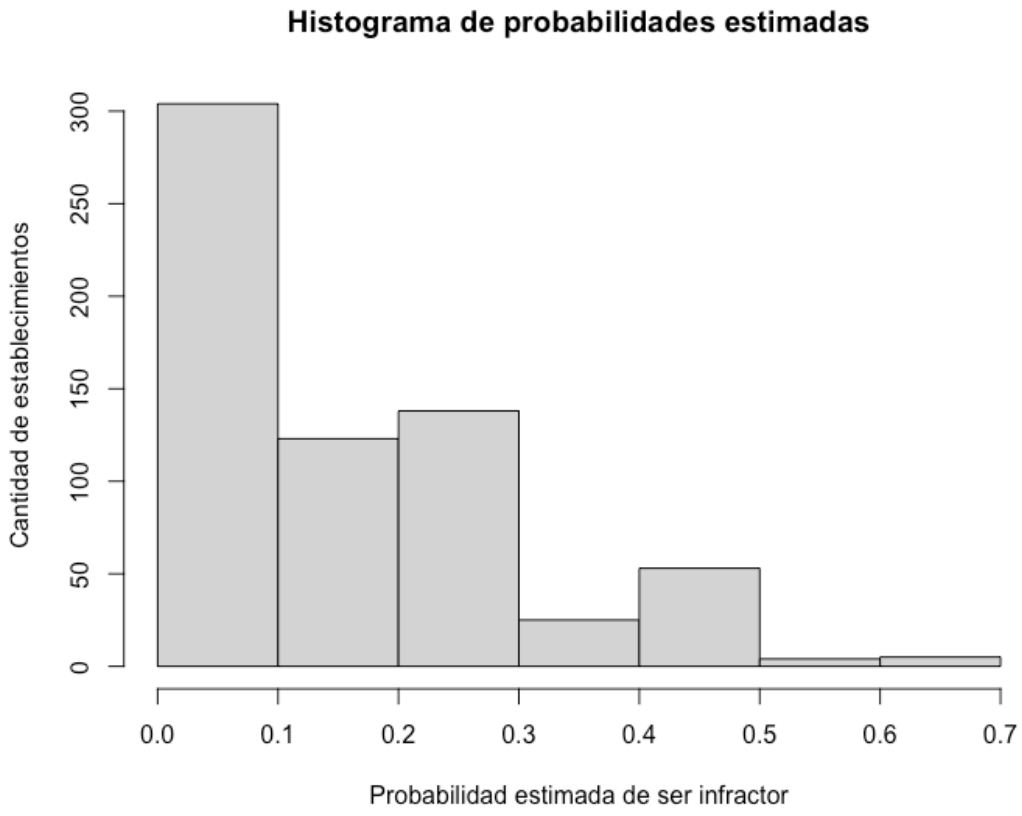
Anexo 4 – Regresión logística con variables de ubicación y asociadas a características comerciales

```
> summary(logit)

Call:
glm(formula = infractor ~ departamento + tipo_mod_carniceria +
    vende_chacinados, family = binomial, data = pais_logit)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3268  -0.6050  -0.3832  -0.1516   3.0593

Coefficients:
                                      Estimate Std. Error z value Pr(>|z|)
(Intercept)                           -2.4795     0.2018 -12.285  < 2e-16 ***
departamentoArtigas                    0.6367     0.4915   1.295  0.19516
departamentoCerro Largo               -1.9809     1.0246  -1.933  0.05320 .
departamentoColonia                   -0.7704     0.4269  -1.804  0.07116 .
departamentoDurazno                   -0.2911     0.5015  -0.580  0.56161
departamentoFlores                   -15.5546   796.8356  -0.020  0.98443
departamentoFlorida                    0.4827     0.3480   1.387  0.16549
departamentoLavalleja                 -1.1567     0.7432  -1.556  0.11963
departamentoMaldonado                  0.7743     0.2482   3.120  0.00181 **
departamentoMontevideo                 1.2767     0.1963   6.503 7.86e-11 ***
departamentoPaysandú                   0.6306     0.3170   1.989  0.04665 *
departamentoRío Negro                 -0.6782     0.7507  -0.903  0.36627
departamentoRivera                    -1.8545     1.0258  -1.808  0.07064 .
departamentoRocha                     -0.9695     0.5429  -1.786  0.07416 .
departamentoSalto                     -0.1210     0.3941  -0.307  0.75874
departamentoSan José                 -15.6007   396.2684  -0.039  0.96860
departamentoSoriano                   -0.6559     0.4955  -1.324  0.18561
departamentoTacuarembó                -2.1910     1.0231  -2.142  0.03223 *
departamentoTreinta y Tres            -1.5538     1.0291  -1.510  0.13109
tipo_mod_carniceriaIndependiente Expendio   0.9163     0.5311   1.725  0.08448 .
tipo_mod_carniceriaPolleria Corte           0.2914     0.6488   0.449  0.65335
tipo_mod_carniceriaSupermercado Corte       0.8743     0.1472   5.941 2.83e-09 ***
tipo_mod_carniceriaSupermercado Expendio    1.1698     0.4017   2.912  0.00359 **
vende_chacinadosVende                       0.3777     0.1658   2.279  0.02268 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1885.8  on 2174  degrees of freedom
Residual deviance: 1592.8  on 2151  degrees of freedom
AIC: 1640.8

Number of Fisher Scoring iterations: 16
```

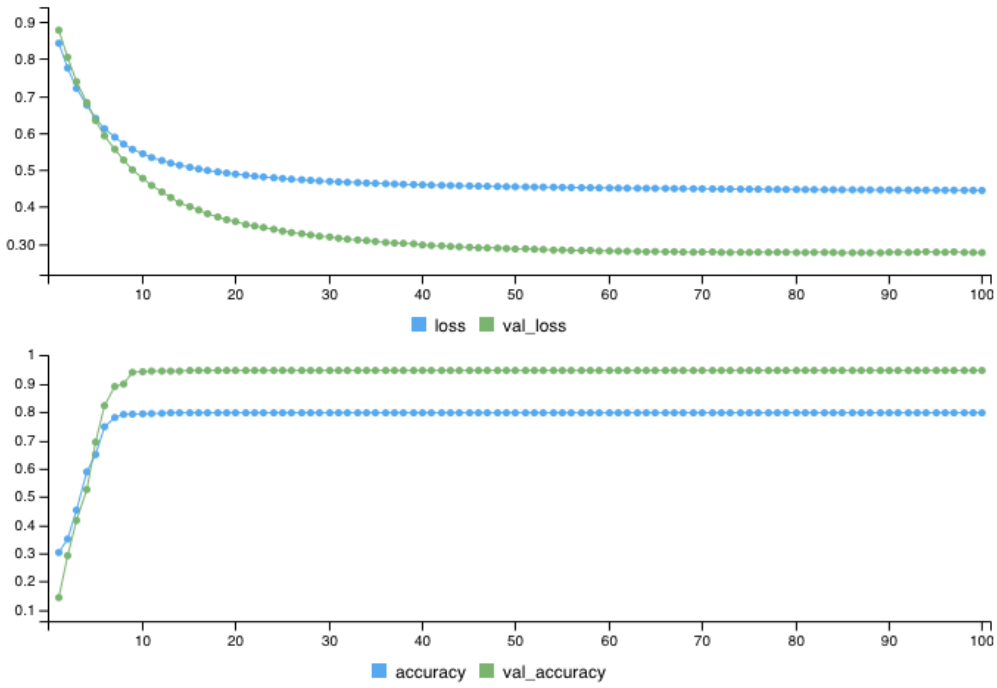
Anexo 5 – Histograma de probabilidades estimadas para el modelo RL – M2



Anexo 6 – Histograma de probabilidades estimadas para el modelo RL – M3



Anexo 7 – Resumen performance ANN 1 – Optimizador *adam*



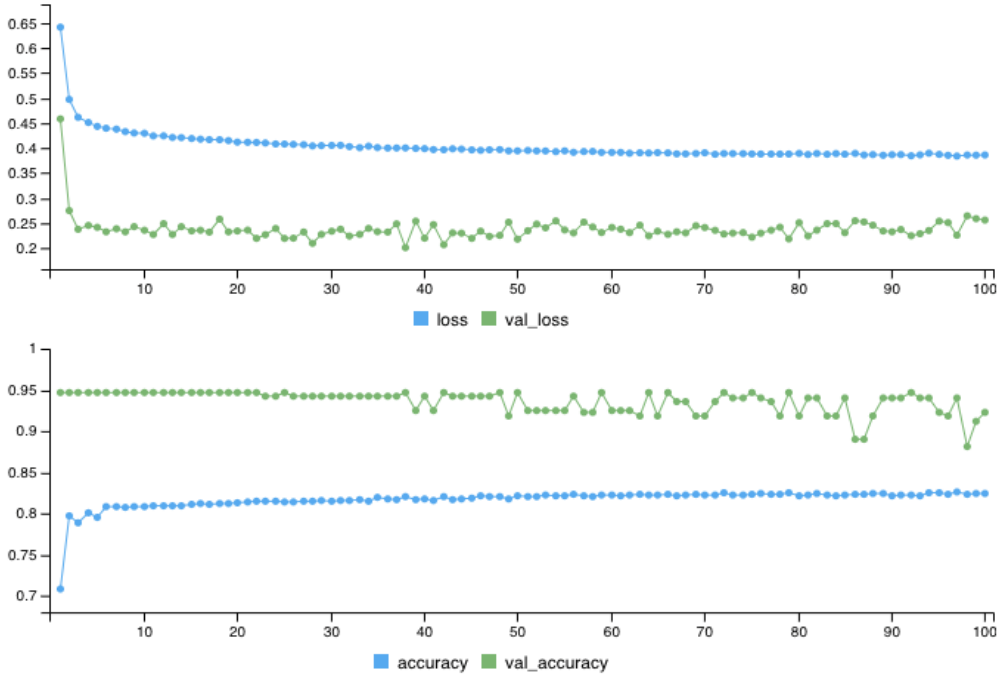
	loss	accuracy
train	0,3942	0,8424
test	0,3963	0,8466

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 514  70
         1  38  30

               Accuracy : 0.8344
                 95% CI : (0.8036, 0.8621)
    No Information Rate : 0.8466
    P-Value [Acc > NIR] : 0.822721

                  Kappa : 0.266

 Mcnemar's Test P-Value : 0.002855

            Precision : 0.44118
               Recall : 0.30000
                   F1 : 0.35714
           Prevalence : 0.15337
       Detection Rate : 0.04601
 Detection Prevalence : 0.10429
    Balanced Accuracy : 0.61558

     'Positive' Class : 1
```

Anexo 8 – W_i estimados por la red ANN 1 utilizando los optimizadores *sgd* y *adam*

id	variable	weights *sgd*	weights *adam*
1	tipo_mod_carniceria_Independiente Corte	-0,7264	-0,5604
2	tipo_mod_carniceria_Independiente Expendio	0,0845	-0,1972
3	tipo_mod_carniceria_Polleria Corte	-0,3205	-0,0134
4	tipo_mod_carniceria_Supermercado Corte	0,1182	0,2266
5	tipo_mod_carniceria_Supermercado Expendio	0,2507	0,1237
6	departamento_Canelones	-0,6525	-0,6523
7	departamento_Artigas	0,1119	0,1041
8	departamento_Cerro Largo	-0,5596	-1,4976
9	departamento_Colonia	-0,5793	-1,1013
10	departamento_Durazno	-0,0839	-0,3754
11	departamento_Flores	-0,4494	-1,616
12	departamento_Florida	-0,2559	0,0136
13	departamento_Lavalleja	0,0247	-0,8936
14	departamento_Maldonado	-0,2255	0,0704
15	departamento_Montevideo	0,4175	0,6332
16	departamento_Paysandú	-0,0907	0,3189
17	departamento_Río Negro	0,0204	-0,2874
18	departamento_Rivera	0,2955	-0,1929
19	departamento_Rocha	-0,1839	0,2463
20	departamento_Salto	0,3245	0,2856
21	departamento_San José	-0,2420	-0,0916
22	departamento_Soriano	-0,4092	-0,104
23	departamento_Tacuarembó	0,2625	0,0378
24	departamento_Treinta y Tres	0,3682	-0,1022
25	elaboracion_No hab No elabora	-0,1174	-0,1785
26	elaboracion_No hab Elabora	-0,4316	-0,3224
27	elaboracion_Hab Elabora	-0,1520	-0,3282
28	vende_no_carnicos	-0,0169	-0,034
29	vende_chacinados	-0,0436	0,101
30	realiza_coccion	-0,2100	-0,3487
31	poblacion	-0,0485	0,084
32	superficie_km2	-0,2377	-0,604
33	ing_p_capita_mes_cte_2005	-0,0539	-0,4280
34	neurona	-0,6385	-0,5069

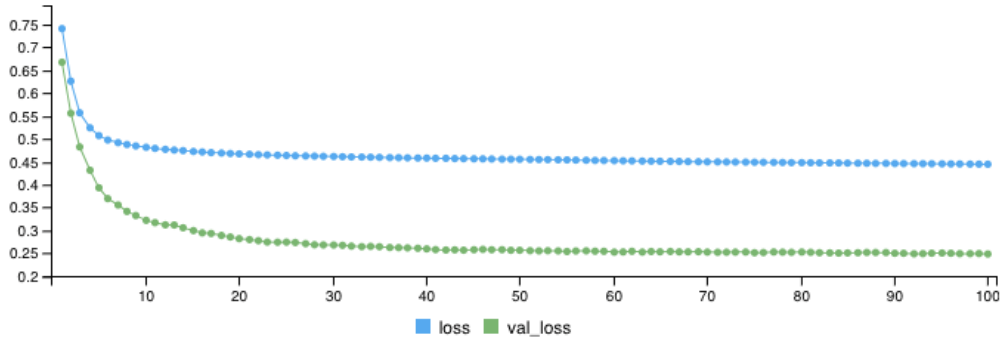
Anexo 9 – Resumen performance ANN 2 – Optimizador *adam*

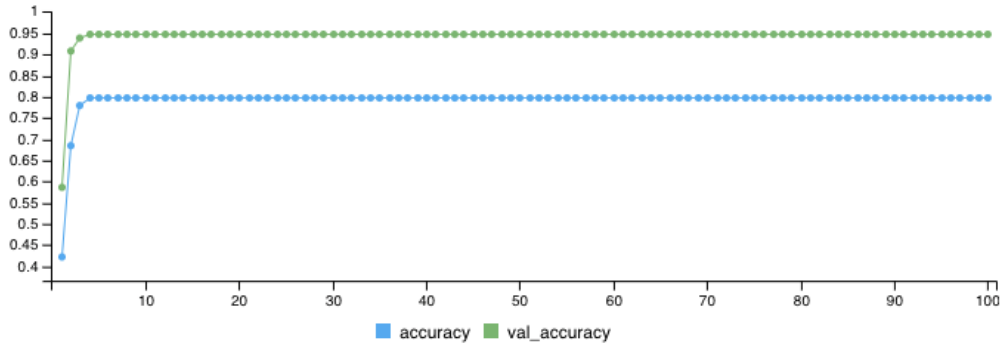


	loss	accuracy
train	0,3423	0,8549
test	0,4270	0,8420

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 448  48
         1 104  52

               Accuracy : 0.7669
                 95% CI : (0.7325, 0.7988)
    No Information Rate : 0.8466
    P-Value [Acc > NIR] : 1

                  Kappa : 0.2697

 Mcnemar's Test P-Value : 8.154e-06

            Precision : 0.33333
               Recall : 0.52000
                   F1 : 0.40625
           Prevalence : 0.15337
       Detection Rate : 0.07975
 Detection Prevalence : 0.23926
    Balanced Accuracy : 0.66580

     'Positive' Class : 1
```

Anexo 10 – Resumen performance ANN 3 – Optimizador *sgd*

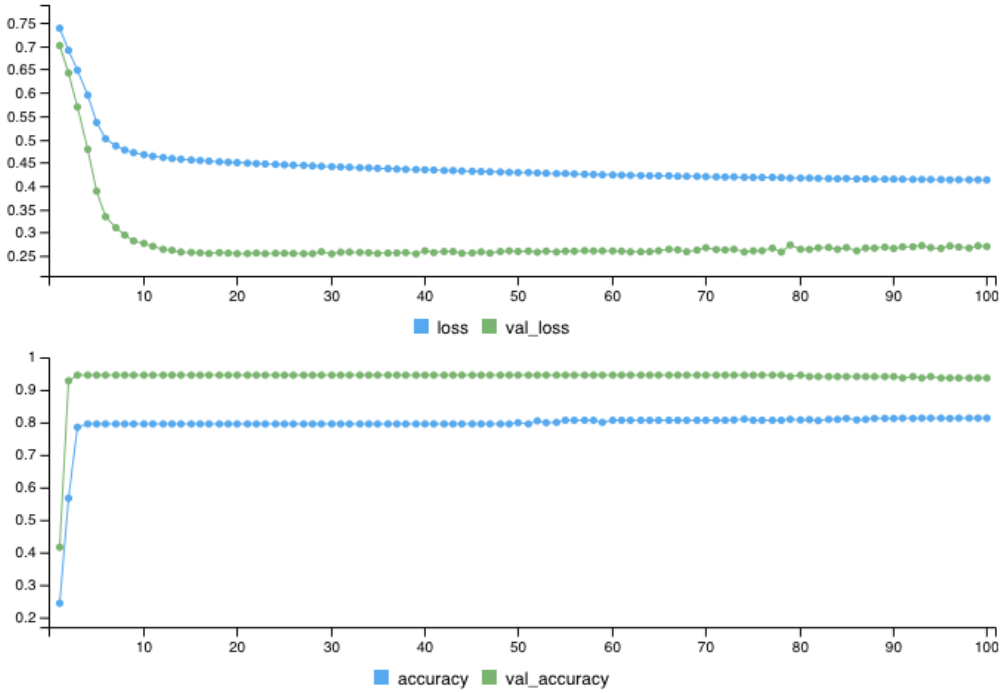




	loss	accuracy
train	0,3855	0,8424
test	0,3894	0,8466

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 493  59
         1  59  41

               Accuracy : 0.819
                 95% CI : (0.7873, 0.8478)
    No Information Rate : 0.8466
    P-Value [Acc > NIR] : 0.9759

                  Kappa : 0.3031

 Mcnemar's Test P-Value : 1.0000

              Precision : 0.41000
                 Recall : 0.41000
                     F1 : 0.41000
             Prevalence : 0.15337
         Detection Rate : 0.06288
   Detection Prevalence : 0.15337
      Balanced Accuracy : 0.65156

       'Positive' Class : 1
```

Anexo 11 – Resumen performance ANN 3 – Optimizador *adam*



	loss	accuracy
train	0,3694	0,8523
test	0,3904	0,8620

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 491  61
         1  61  39

               Accuracy : 0.8129
                 95% CI : (0.7808, 0.8421)
    No Information Rate : 0.8466
    P-Value [Acc > NIR] : 0.9916

                  Kappa : 0.2795

 Mcnemar's Test P-Value : 1.0000

              Precision : 0.39000
                 Recall : 0.39000
                     F1 : 0.39000
             Prevalence : 0.15337
         Detection Rate : 0.05982
   Detection Prevalence : 0.15337
      Balanced Accuracy : 0.63975

       'Positive' Class : 1
```