REVISTA
SINERGIAS

Publicación Semestral de la Secretaría de Posgrado de la Universidad Nacional Guillermo Brown, en colaboración con UNaB Editora.

NÚMERO #01
JULIO DE 2024

Métodos computacionales aplicados a la Política Pública en Argentina

sinergias.unab.edu.ar

REVISTA SINERGIAS

REVISTA SINERGIAS
Publicación Semestral de la Secretaría de Posgrado de la Universidad Nacional Guillermo Brown, en colaboración con UNAB Editora

UNIVERSIDAD NACIONAL
GUILLERMO BROWN

AUTORIDADES

RECTOR

Lic. Pablo Matías Domenichini

VICERRECTOR

Lic. Facundo Nejamkis

SECRETARÍAS

SECRETARÍA ACADÉMICA

Matías Triguboff

SECRETARÍA GENERAL

Stella Salamone

SECRETARÍA ECONÓMICO ADMINISTRATIVA

Diego Otero

SECRETARÍA DE EXTENSIÓN Y BIENESTAR

Ignacio Jawtuschenko

SECRETARÍA DE POSGRADO

Andrés Gilio

La revista electrónica **Sinergias** es una publicación de la Universidad Nacional Guillermo Brown. Ha sido pensada con el objetivo de estimular la reflexión crítica desde las ciencias sociales sobre el desarrollo de políticas públicas. Esta publicación visibiliza el trabajo que se realiza desde la Universidad en articulación/extensión con el Estado y el territorio.
La revista está dirigida a la comunidad universitaria, a los hacedores de políticas públicas y a quienes se interesen por conocer, ampliar y profundizar sobre esta temática.


REVISTA
SINERGIAS

SINERGIAS #1

Métodos computacionales aplicados
a la Política Pública en Argentina

PRESENTACIÓN

Con la publicación de este primer número de la revista Sinergías, la UNAB abre un nuevo espacio de reflexión, producción y debate político y académico. El objetivo es estimular los vínculos de trabajo colaborativo entre la academia, los diferentes niveles de decisión política en el Estado y el territorio. Con esta revista nos proponemos contribuir con la formación de los estudiantes y con el conocimiento sobre el desarrollo de políticas públicas basadas en evidencia, y facilitar el intercambio entre los integrantes de la comunidad académica y los hacedores de políticas públicas.

La revista se suma a las diferentes iniciativas que integran el dispositivo académico de la UNAB dedicado a la investigación, la formación de posgrado y la extensión universitaria en el campo de las políticas públicas a partir del Laboratorio de Datos Políticos y Sociales (LDPS), la Diplomatura en análisis de datos aplicados al desarrollo de políticas pública y la Diplomatura en Ciencias Sociales Computacionales. Estas iniciativas tienen en común el reconocer como objeto de estudio, de formación e intervención los procesos y situaciones que definen y caracterizan las condiciones de vida de la población argentina desde un enfoque interdisciplinario, a través de los cuales se proponen abordar los fenómenos políticos y sociales con políticas públicas basadas en evidencia. Enraizada en este enfoque, esta revista se encuentra disponible para docentes, investigadores, estudiantes de la UNAB y de otras instituciones académicas tanto nacionales como internacionales, así como para funcionarios y expertos en el ámbito político, con el propósito de difundir, debatir eintercambiar conocimientos y experiencias en los ámbitos académicos y políticos.

¿POR QUÉ UNA REVISTA SOBRE SINERGÍAS ENTRE LA UNIVERSIDAD, EL ESTADO Y EL TERRITORIO?

La revista nace con el objetivo de estimular los vínculos de trabajo colaborativo entre la academia y el Estado, contribuyendo a la producción de conocimiento sobre el desarrollo de políticas públicas basadas en evidencia. Su nombre, Sinergías, refleja la colaboración y el fortalecimiento mutuo entre la Universidad, el Estado y el territorio, creando un espacio donde el conocimiento y la acción conjunta generan impactos más significativos en la sociedad. El recorte e identificación de este objetivo apunta a desarrollar un enfoque de la política pública basada en evidencia que atienda las múltiples dimensiones de las problemáticas socioterritoriales con el fin de contribuir a la superación de las desigualdades sociales y de los problemas institucionales del país. En aras de generar políticas públicas que atiendan a estas problemáticas desde la Universidad colaboramos con los distintos niveles del Estado en la realización de estudios diagnósticos, capacitaciones y construcción conjunta de respuestas a necesidades y problemas concretos.

El intercambio de saberes y puntos de vista es un proceso de aprendizaje mutuo entre actores científicos y no académicos; esta propuesta valoriza el saber popular y el intercambio de saberes entre diferentes disciplinas y con los sujetos con quienes trabajamos como condición central en la investigación. El trabajo territorial es entendido desde este enfoque como un conjunto de actividades definidas en función de la agenda de sujetos concretos, es decir, una antropología por demanda a partir de las solicitudes de los sujetos e instituciones con quienes interactuamos. En esta articulación entre investigación y extensión se reconoce los diferentes intereses a los que responden los distintos actores -universidad, territorio y Estado- de este diálogo. El diseño de abordajes metodológicos para el gobierno responde a las necesidades de la gestión, y debe adecuarse a factores contextuales, recursos, tiempos y urgencias. Los procesos de extensión/articulación requieren de adecuaciones, adaptaciones y flexibilidad respecto a la rigurosidad de las técnicas de investigación en las Ciencias Sociales. La incorporación de los saberes situados de los equipos técnicos y decisores ejecutivos de la administración pública permite no solo la co-producción de conocimiento, sino también mayores posibilidades para que se utilicen estos datos para la planificación de políticas públicas. No se trata de una transferencia pasiva de un contenido altamente valioso los agentes estatales, sino de involucrar activamente a los encargados de la toma de decisiones y a sus equipos técnicos en el desarrollo de la investigación para generar evidencia co-construida que incorpore sus saberes situados y la propia idiosincrasia.

Esta revista se propone ser un espacio para la difusión y discusión de formas innovadoras de tomar decisiones en la gestión pública a partir de las herramientas provistas por el análisis de datos. En la actualidad, el aumento de disponibilidad de grandes fuentes de información observacionales que refieren a una gran variedad de fenómenos sociales representa un momento decisivo para las Ciencias Sociales. Las tecnologías móviles (celulares, apps, etc.), la “internet de las cosas” (IoT), la información producida por los usuarios de redes sociales son algunos de los vectores de este incremento en la disponibilidad de información. Estos datos han creado la demanda de nuevos métodos que reducen/simplifican su dimensionalidad, identifican nuevos patrones y relaciones y predicen resultados. En la búsqueda de modos de reducirlos a dimensiones abordables y significativas existen diversas variedades de técnicas computacionales.

¿POR QUÉ UNA REVISTA SOBRE SINERGÍAS ENTRE LA UNIVERSIDAD, EL ESTADO Y EL TERRITORIO?

Esta revista se propone como una nueva herramienta para difundir trabajos académicos e intelectuales recientes que emplean grandes volúmenes de datos y técnicas computacionales y del mundo de la ciencia de datos al estudio de fenómenos políticos y sociales. De esta manera se busca promover programas de políticas públicas basadas en evidencias construidas en plazos de tiempos reducidos y optimizando los límites presupuestarios, y/o que faciliten el seguimiento de las políticas públicas en tiempo real, originando mejoras a partir de la evaluación continua de resultados.

Hecho el planteo en estos términos, los retos derivados de la necesidad de adquirir conocimiento en estas áreas y la relevancia de ampliar los debates mediante la inclusión de diversos agentes sociales y políticos abren un vasto ámbito de actividad y generación de teorías y metodologías con el fin de contribuir a diseñar políticas basadas en evidencia. Estos problemas convocan de manera equitativa a los hacedores de políticas públicas, equipos técnicos de la administración gubernamental y las distintas disciplinas que, desde la esfera de la sociología, la economía, la antropología, la planificación urbana, las ciencias políticas, las ciencias de datos y las ciencias computacionales, pueden aportar desde sus respectivas especialidades y con una mentalidad transformadora al desarrollo de una sociedad más equitativa, participativa y solidaria, donde se garanticen los derechos de todos sus habitantes.

Esperamos que Sinergías contribuya al desarrollo de políticas públicas más justas y efectivas, y que sirva como un puente entre la academia y los distintos niveles de gobierno. Agradecemos profundamente a todos los que han hecho posible este proyecto, y confiamos en que este espacio se consolidará como un referente en el campo de las políticas públicas basadas en evidencia.

FLORENCIA PIÑEYRÚA
COORDINADORA EDITORIAL

EDITORIAL

La revista electrónica *Sinergia* es una publicación de la Universidad Nacional Guillermo Brown. Esta publicación visibiliza el trabajo que se realiza desde la Universidad en articulación con el Estado y el territorio. La revista está dirigida a la comunidad universitaria, a los hacedores de política pública y a quienes se interesen por conocer, ampliar y profundizar sobre esta temática.

El nombre elegido para la revista hace referencia a nuestra perspectiva, que sostiene el carácter colectivo de la construcción de conocimiento, una acción conjunta que no se alcanza con la mera suma de las partes, sino que se produce como efecto mismo de la interacción. Apostamos a ese horizonte, a esa manera de concebir la educación y la investigación.

En este primer número se presentan una serie de trabajos que se han realizado en el marco de la Diplomatura en Análisis de Datos Aplicados al Desarrollo de Políticas Públicas y de la Diplomatura en Ciencias Sociales Computacionales. El eje que estructura a todos ellos es la intención de pensar los problemas sociales problematizados con evidencia empírica, apoyados en la elaboración de datos que fundamentan los temas trabajados. Es un gran desafío contribuir de este modo con la producción de conocimiento basado en experiencias concretas de análisis, utilizando herramientas que nos permiten abordar «*lo real*»; desde nuevas miradas.

Esperamos en un futuro próximo contar con nuevas secciones, ampliar nuestro staff, sumar reflexiones y hacer de esta revista un órgano de difusión de trabajos y debates que se vienen dando en nuestra Universidad y otras casas de estudio que comparten estas inquietudes. Los convocamos a leer, discutir y aportar en esta nueva aventura.

LAUTARO WALLACE

BUENOS AIRES, NOVIEMBRE 2023

DOSSIER

En este primer número de Sinergías incluye trabajos que analizan temas y problemas cruciales que atañen a distintos grupos sociales en la Argentina contemporánea y que son objeto de políticas públicas en distintos niveles del Estado. Tres han sido presentados como trabajos finales en la Diplomatura en análisis de datos aplicados al desarrollo de políticas públicas de la Universidad Nacional Guillermo Brown durante 2023. Los otros dos artículos han sido elaborados en el marco de la Diplomatura en Ciencias Sociales Computacionales. Los cuatro artículos fueron seleccionados con la idea de proveer un compendio de lecturas sobre aspectos teóricos y prácticos que son abordados en estos estudios de posgrado. En su conjunto, estos trabajos están orientados a abordar cuestiones vinculadas a la interacción entre el territorio, la Universidad y el Estado.

Inicia el volumen el artículo de Romina De León y Magali Wettstein que analiza las notificaciones de casos de tuberculosis antes y durante la pandemia de COVID-19 mediante aprendizaje automático. Las fuentes de datos son de carácter secundario, provienen de los reportes obtenidos del Sistema Nacional de Vigilancia de la Salud de la República Argentina. Para el análisis se realizó una combinación de enfoques computacionales en la búsqueda de proporcionar una visión integral del conjunto de datos, permitiendo interpretaciones más robustas y contextualizadas. Las técnicas empleadas fueron herramientas de regresión (que modela las relaciones entre variables), la georreferenciación (que explora la dimensión espacial de los datos) y el análisis de supervivencia, que profundiza en la duración de eventos de interés a lo largo del tiempo. Las autoras evidencian una compleja interacción entre la pandemia de COVID-19 y el diagnóstico y tratamiento de la tuberculosis en Argentina, destacando la importancia de considerar contextos demográficos regionales y específicos en la eficiencia dispar en la atención de la salud de los casos de tuberculosis.

Lucia Julia Gaztañaga examina los datos de violencias por motivos de género dado que presentan importantes desafíos para su análisis, al mismo tiempo, que requieren de metodologías específicas para obtener estrategias de diagnóstico e intervención. El trabajo es una contribución al seguimiento y evaluación de políticas públicas en materia de género a nivel nacional, dado que su objetivo es construir nuevos indicadores que puedan realizar un aporte al análisis de los casos de violencias por motivos de género que aborda periódicamente la Línea. La autora realiza propuesta que abarca la construcción de indicadores a partir del análisis de los datos públicos 2020-2022 de la Línea 144, destinada a la atención y abordaje de consultas de mujeres y personas LGBTI+ que atraviesan diversas situaciones de violencias por motivos de género. Para ello emplea una estrategia metodológica computacional al explorar dos modelos de clasificación mediante distintas técnicas, “árbol de decisión” y el algoritmo del tipo “random forest” (bosque aleatorio).

La segunda sección de Dossier, inicia con el trabajo final de Fernando Ashbey, Julieta Coll, Adrián Ibarra Adrián y Juan Pablo Zumárraga sobre la interoperabilidad y la gobernanza, conceptos cada vez más importantes en el ámbito de la Administración Pública Nacional. Este trabajo se interroga sobre las posibilidades de optimización de las políticas públicas con el fin de estudiar las posibilidades de interoperabilidad dentro del Estado Nacional y así obtener evidencia para el diseño integrado de las mismas. Para avanzar una respuesta, se tomó un caso de estudio (la producción apícola) y se identificó las distintas fuentes de datos que pueden

potencialmente emplear los distintos organismos nacionales. Lo anterior implicó indagar tanto en la información pública disponible, así como también datos obtenidos por medio de pedidos de acceso a la información pública o a la que podían acceder los organismos de la Administración Pública mediante convenios de cooperación interinstitucional.

El trabajo de la autoría de Brián Covaro analiza desde un enfoque computacional las particularidades de la fatalidad vial en la Región del Noreste Argentino durante el período 2019 a 2020. A partir de datos oficiales sobre siniestros viales de este bienio el autor construyó distintos modelos de Machine Learning basados en algoritmos distintos que buscan caracterizar y contribuir a explicar los patrones que presentan los siniestros viales tipificados como fatales. Este análisis computacional tiene el potencial de aportar, no solo a la caracterización sino también a la predicción o prevenir (traduciendo a políticas públicas) la fatalidad vial.

Por último, en el trabajo titulado "Programa Cambio Rural (CR) – SAGyP: identificación del perfil de beneficiarios con IA", Patricia Perrone, Marianela Pi y Constanza Guerrini caracterizan a los beneficiarios/as de una política pública que busca promover y facilitar la intensificación y reconversión productiva. La Secretaría de Agricultura, Ganadería y Pesca de la Nación a través de la asistencia técnica busca mejorar la situación productiva y socioeconómica de los pequeños y medianos productores rurales y propender al desarrollo agroindustrial en todo el territorio nacional, impulsando el aprendizaje grupal. A partir de emplear algoritmos de K-Modes con CAO, las autoras muestran resultados que coinciden con el conocimiento empírico de la base de datos del Programa Cambio Rural.

SINERGIAS
EQUIPO EDITORIAL

DOSSIER

14

Análisis de notificación de casos de tuberculosis antes y durante la pandemia de COVID-19 mediante aprendizaje automático

De León, Romina & Wettstein, Magali

De León, Romina[1] & Wettstein, Magali [2]

Análisis de notificación de casos de tuberculosis antes y durante la pandemia de COVID-19 mediante aprendizaje automático

Resumen

La integración de la Inteligencia Artificial en bases de datos masivas, inabarcables al ojo humano, como aquellas provenientes de las notificaciones de casos de enfermedades infecciosas podría proporcionar nuevos conocimientos sobre la afección que resultó de los aislamientos sociales durante la pandemia de COVID-19, especialmente entre la declaración de la misma y el primer año de desarrollo en pacientes con tuberculosis. Para ello, se empleará aprendizaje automático supervisado en el procesamiento de datos con variables predictoras diversas. En este contexto, también se llevará a cabo un análisis de series temporales para evaluar si las medidas de dichas circunstancias afectaron los tiempos de demora en el diagnóstico y en el tratamiento de la tuberculosis. Este enfoque tendrá como objetivo identificar patrones, tendencias y posibles correlaciones que podrían proporcionar nuevos conocimientos sobre cómo la pandemia de COVID-19 impactó específicamente en la gestión de la tuberculosis, contribuyendo así a una comprensión más completa de las dinámicas de salud pública en situaciones de crisis.

Palabras clave

Tuberculosis, COVID-19, aprendizaje automático, inteligencia artificial, aislamiento social.

1. Introducción[3]

Los avances tecnológicos posteriores al período de las guerras mundiales no tienen precedentes. Incluso desde la mitad del siglo pasado, se han producido actualizaciones digitales a pasos colosales, y en todas las áreas de nuestra vida, por ende no son ajenos todos los ámbitos académicos. Eric Hobsbawm (2018) asegura que ningún otro período de la historia ha sido más impregnado ni ha dependido más de las Ciencias que el siglo XX (443).

La estructura tecnológica satura la vida humana, demostrando su indispensable omnipresencia, incluso en las Ciencias de la Salud. Por ello, en las últimas décadas, la aplicación y análisis con diversas metodologías computacionales ligadas a la Inteligencia Artificial (IA), en grandes cantidades de datos, ha permitido la toma de decisiones proactivas y la predicción de eventos de enfermedades a través de modelos de aprendizaje automáticos. La IA forma parte de la Informática y tiene como objetivo resolver problemas, en torno al aprendizaje y el reconocimiento de patrones.

Una subdisciplina de la IA es el aprendizaje automático o *machine learning* (en inglés, ML). Este tuvo sus primeros avances a mediados de siglo XX por científicos que trabajaban en la empresa IBM (International Business Machines Corporation), con el desarrollo de un programa para jugar damas, presentando el potencial que tendría esta metodología. Asimismo el ML utiliza algoritmos que aprenden de los datos masivos, estructurados o no estructurados, para predecir resultados y extraer nuevos conocimientos. Además, los algoritmos son los que toman decisiones de manera similar al humano después de un proceso de entrenamiento con datos. Es por ello que tanto la IA como el aprendizaje automático, al ser incorporados a las Ciencias de la Salud, en particular para trabajar con datos masivos de enfermedades infecciosas cómo la tuberculosis, permite agilizar y generar nuevos conocimientos de datos médicos, procurando mejorar resultados de salud y vivencias en tres niveles, en lo individual, de cada paciente, en las comunidades de pacientes y, en la sociedad en su conjunto (Martínez Sesmero, 2015).

La tuberculosis (TB), enfermedad infecciosa con antecedentes que datan de hace más de 10.000 años, es causada por la bacteria *Mycobacterium tuberculosis* y suele afectar a los pulmones. Es prevenible y curable,

[1] IIBICRIT-CONICET, rdeleon@conicet.gov.ar, Romina Soledad De León (0000-0003-2495-7213) - ORCID.
[2] INER-ANLIS, mwettstein@anlis.gob.ar
[3] El presente trabajo se desarrolló en el marco de la Diplomatura de Ciencias Sociales Computacionales de la Universidad Nacional Guillermo Brown (UNAB) cohorte 2023.

transmitiéndose de una persona a otra a través de gotitas de aerosol suspendidas en el aire expulsadas por personas con enfermedad pulmonar activa. El diagnóstico se basa en radiografías torácicas, pruebas cutáneas de tuberculina y análisis sanguíneo. Su tratamiento se realiza con antibióticos durante períodos de alrededor de seis meses. En 2022, la TB fue la segunda causa de muerte por un solo agente infeccioso en el mundo, después de la enfermedad por coronavirus (COVID-19), y causó casi el doble de muertes que el VIH/SIDA. Actualmente, la Organización Mundial de la Salud (OMS)[4] estima que existen 2.000 millones de infectados por esta bacteria, y por año se contagian alrededor de 8 millones de personas.

En la actualidad esta infección continúa siendo una de las que posee tasas de morbimortalidad[5] más alta a nivel mundial. Incluso, la Organización Panamericana de la Salud (OPS)[6] considera que las muertes anuales en América han aumentado a más de 3.000 en 2020 debido a la interrupción de los servicios esenciales en el contexto de la pandemia mundial de la COVID-19[7]. Por lo anterior, tanto los Estados Miembros de las Naciones Unidas (ONU) como la Organización Mundial de la Salud (OMS)[8] expresan la necesidad de medidas urgentes para finalizar la epidemia global de TB para 2030, destacando como requerimiento la inversión en recursos para intensificar la respuesta para los enfermos, sobre todo a los más vulnerables.

En conjunto, se puede entender que el ingreso de la IA al ámbito de la salud, con modelos de aprendizaje, algoritmos de ML, permitirán detectar patrones en los datos, facilitando el estudio predictivo y generando mayores conocimientos que impactarán en la calidad de vida de los pacientes. Por lo cual, el presente trabajo tiene por objetivo realizar una evaluación actualizada de la notificación de casos de tuberculosis en Argentina[9], especialmente durante el período 2019-2021, para llevar a cabo una comparación en los sucesos de la pandemia de COVID-19. Se utilizará como base de datos el reporte sobre la comunicación de casos que genera el Sistema Nacional de Vigilancia de la Salud (SNVS 2.0)[10], considerando, sólo, el período mencionado en busca de revelar la modificación o no de la notificación de afectados antes y durante la pandemia, en particular teniendo en cuenta las restricciones sociales que se llevaron a cabo en todo el mundo, durante varios meses[11].

2. Metodología

En este estudio se utilizó el informe del 31 de agosto de 2023 del SNVS[12], que contiene las notificaciones de casos registrados y validados de las 24 jurisdicciones argentinas entre los años 2019 y 2021. El dataset no estaba disponible directamente desde la URL, por lo tanto, después de descargarlo en formato CSV, se subió al repositorio GitHub. Esto se hizo con el objetivo de facilitar el trabajo colaborativo, que permite realizar

---

[4] Para más información, véase Tuberculosis.

[5] Tasa de muertes por enfermedad en una población y en un tiempo determinados.

[6] La Organización Panamericana de la Salud actúa como oficina regional de la Organización Mundial de la Salud. Su primer encuentro fue en diciembre de 1889 en Washington, EE. UU., pero se conformó como tal en 1902 en la misma ciudad.

[7] La pandemia de COVID-19 se derivó de la enfermedad causada por el virus SARS-CoV-2.78. Los primeros casos identificados fueron a finales de noviembre de 2019 en Wuhan, ciudad de China. El 30 de enero de 2020 fue declarada la emergencia de salud pública internacional por la OMS, condición que mantuvo hasta el 5 de mayo de 2023. El 11 de marzo de 2020 fue declarada como pandemia por la alta cantidad de personas infectadas y muertes alrededor del mundo. En nuestro país, el 20 de marzo del mismo año, se declaró el Aislamiento social, preventivo y obligatorio (ASPO) con el decreto N° 297/2020, accesible desde el siguiente link. El ASPO se extendió hasta el hasta el 31 de enero de 2021 reemplazado con el Distanciamiento social, preventivo y obligatorio (DIASPO) hasta junio de 2021. El proceso de vacunación comenzó el 29 de diciembre de 2020, el 76, 34% de la población argentina había recibido al menos una dosis para el 29 de octubre de 2021. Actualmente, según la OMS la enfermedad se encuentra en nivel endémico, debido a que ya no se reportan casos con tanta frecuencia como en sus primeros años.

[8] Para acceder al informe mundial sobre tuberculosis 2023 generado por la OMS véase: Global Tuberculosis Report 2023.

[9] La ley nacional N°15.46 de Argentina promulgada en 1960 estableció la obligatoriedad, dentro del territorio nacional, de comunicar todos los casos de enfermedades infecciosas.

[10] Desde SNVS se genera un reporte individual por parte de los servicios de salud a lo largo y ancho del país, de manera remota e inmediata. Accesible desde SISA.

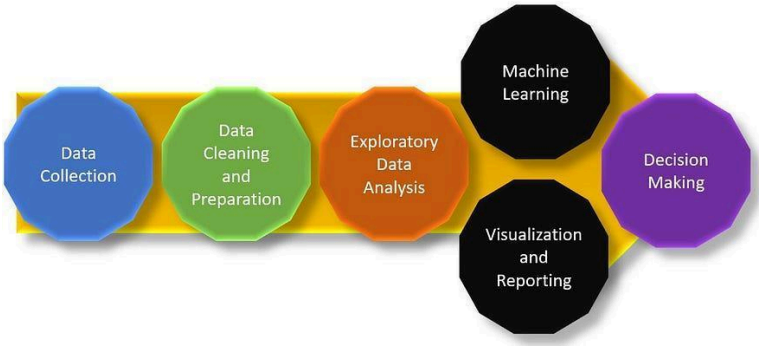
[11] Véase nota 6.

[12] Véase nota 8.

cambios, incluir comentarios y mantener actualizado el seguimiento de los progresos[13].

El desarrollo práctico se inició con un análisis exploratorio de datos (EDA por sus siglas en inglés). Este conjunto de técnicas estadísticas busca acercarse a los datos y explorar las relaciones entre variables. Para ello, mediante métodos sistemáticos, se organiza, limpia y prepara los datos. Además se tratan y evalúan la presencia de valores ausentes (missing) y se detectan valores atípicos (outliers). Además, se generan visualizaciones que permiten comprender y revelar relaciones entre variables, así como seleccionar las herramientas a utilizar en el aprendizaje automático (Monterde i Bort y Perea Lara, 1991). A continuación se presenta un esquema que representa este análisis (Gráfico 1).

Gráfico 1. Flujo de proceso EDA desde la recopilación de datos hasta la toma de decisiones.



Fuente: How to ace Exploratory Data Analysis | by Rahul Pandey | Medium | Analytics Vidhya.

Para continuar la exploración de datos se transformaron en dataframe[14]. Luego, se realizó una inspección estadística de las variables numéricas. Por lo cual, se convirtieron a formatos más adecuados, en el caso de las fechas. Se codificaron variables categóricas, cómo los grupos etarios, sexo, residencia, clasificación del caso. Se reajustaron las variables de *Provincias de residencia*, en particular los valores CABA por Ciudad Autónoma de Buenos Aires[15], información relevante para las gráficas georreferenciadas. Se eliminaron columnas superfluas, como las semanas epidemiológicas[16]. Se determinó la correlación entre variables y se adecuaron valores vacíos, duplicados[17] o faltantes. Se retuvieron únicamente pacientes por Documento Nacional de Identidad (DNI) y fecha de notificación. Además, se organizó de manera ascendente al dataframe. Se destaca respecto a la eliminación de valores faltantes que puede generar pérdida de información por la posible relevancia de los datos en las filas eliminadas. Sin embargo, luego de considerar el porcentaje que los valores vacíos implicaban, o de realizar un resumen del dataframe y de las filas a eliminar, se determinó que la opción más pertinente sería la de reemplazar dichos valores por una estimación basada en los datos existentes o eliminar filas que no aportaban trascendencia a los requerimientos del análisis. Los reemplazos para valores vacíos se hicieron por la media, la mediana o algún otro método de imputación.

La limpieza y análisis del dataframe se limitaron al contexto de la pandemia COVID-19 en Argentina y las

[13] El código e indicaciones de implementación del presente trabajo se encuentran disponibles en el repositorio GitHub y en Google Colab, pudiendo replicar la evaluación y llegando a los mismos resultados.

[14] Es importante señalar la diferencia entre base de datos, dataset y dataframe, que durante todo este trabajo se utilizarán casi como sinónimos. Una base de datos recopila y organiza información, puede almacenar referencias sobre personas, productos, pedidos u otras cosas; por ello, puede entenderse a la misma como dataset. Este, por su parte, es un conjunto de datos, presentando la información en tablas o matrices, que son conformadas por columnas que representan variables particulares, y filas, como categorías de datos. Finalmente, un dataframe es una matriz de datos, que permite datos numéricos y alfanuméricos, además facilita el análisis de los objetos en una muestra de datos, y la información es estructurada en columnas identificadas respectivamente. Por todo lo anterior, se comprende que la consulta, modificación y análisis de información es más sencilla en dataframes.

[15] Como datos de georreferencia se tomaron los datos censales, accesibles desde: https://infra.datos.gob.ar/catalog/modernizacion/dataset/7/distribution/7.31/download/localidades-censales.geojson.

[16] Cabe destacar que la decisión de prescindir de esta información se basa en la premisa de que la fecha proporcionará de manera más efectiva la información relevante para el análisis.

[17] Los datos duplicados refieren a las tomas de muestras, individuales enviadas para confirmar el evento, las de control del tratamiento, así como si el paciente tiene más de una comorbilidad (presencia de otras enfermedades en simultáneo).

fechas de inicio del ASPO[18]. El decreto que anunciaba dicho proceso presentaba en el segundo artículo que "durante la vigencia del 'aislamiento social, preventivo y obligatorio', las personas deberán permanecer en sus residencias habituales o en la residencia en que se encuentren a las 00:00 horas del día 20 de marzo de 2020, momento de inicio de la medida dispuesta". Por ende, se creó una nueva variable clasificatoria de casos que los etiquetaba según **antes** y **después** de la fecha mencionada. El objetivo principal de dicha clasificación fue comparar grupos etarios y regionales respecto a las características de cada caso así como de los tiempos de demora (diagnóstico, tratamiento). Sumado a ello, también se recortó según la clasificación final del resultado de tratamiento de tuberculosis. Es decir, solamente se mantuvieron los pacientes que fueron clasificados como **curados** o **tratamiento completo**, ya que según las *Definiciones y marco de trabajo para la notificación de Tuberculosis* de la OMS (2013) son considerados casos con tratamiento **exitoso**. Entendemos que al filtrar únicamente los casos de éxito, se puede reducir el *ruido* en los datos. Este tipo de eliminaciones puede generar beneficios en la búsqueda de patrones claros o tendencias específicas asociadas a los casos exitosos, pues se elimina la variabilidad introducida por los casos con resultados diferentes.

De manera similar, se agruparon, para realizar visualizaciones más pertinentes, por grupos de edad. Para lo cual se generó una nueva columna con *grupo de edad* que permite catalogarlas de 0 a 18 años, de 19 a 30 años, de 31 a 45 años, de 46 a 60 años, y mayores de 61. Finalmente, se añadieron dos columnas: una que indicaba el tiempo de demora respecto al inicio de síntomas y fecha de diagnóstico, y otra con el tiempo de demora del tratamiento, que cuantifica el comienzo del tratamiento y la finalización del mismo. Los valores negativos de estas nuevas variables, que pueden deberse a errores de cálculo o en la carga de datos, se sustituyen por la media de los valores positivos.

La visualización gráfica y de datos tabulares es propicia para mejorar la interpretación de los datos, así como para limitar los pasos a seguir en el análisis, la toma de decisiones y seleccionar los mejores métodos para el aprendizaje automático supervisado puesto que permiten captar conceptos y patrones que no son relevados sencillamente en las grandes bases de datos (Wickham y Grolemund, 2016). Asimismo, para analizar las relaciones entre las variables, se calculó la matriz de covarianza, que permite apreciar en su diagonal principal la varianza de cada variable, mientras que los elementos restantes arrojan las covarianzas entre las variables. Análogamente, se realizaron diagramas de dispersión evaluando los conjuntos de datos en relación a dos variables, una en cada eje cartesiano. Las representaciones gráficas por medio de histogramas evidencian variables cuantitativas continuas de un conjunto de valores, en forma de barras que proporcionan la frecuencia de valores observados (Murray y Spiegel, 2009). Con ellos, se pueden diferenciar parámetros, como posición, dispersión, asimetría, etc., que son de gran aporte dentro del análisis descriptivo de datos.

En lo que respecta al *machine learning* o aprendizaje automático, campo de estudio derivado de la inteligencia artificial, se ocupa del desarrollo y estudio de algoritmos estadísticos que pueden generalizar eficazmente, es decir, realizar tareas sin instrucciones explícitas. Este enfoque se ha aplicado en diferentes áreas, con desarrollo de diversos modelos como los lingüísticos, de visión por computadora, reconocimiento de voz, filtrado de correo electrónico, agricultura y medicina, etc. Dentro de este subcampo existen subdivisiones, las más destacadas y utilizadas son el aprendizaje supervisado y el no supervisado[19]. El supervisado es utilizado para datos donde las etiquetas son conocidas, ejemplo de ello podría ser la clasificación o la regresión. En este trabajo se utilizaron varios de estos modelos, pero fueron descartados o conservados según el ajuste más aceptable de los resultados. Incluso se consideraron modelos con menor sensibilidad a valores atípicos que otros, sin embargo, resultaron poco concluyentes y no se tuvieron en cuenta[20].

---

[18] Véase nota 6.

[19] Este tipo no fue utilizado para la predicción de los datos de este trabajo debido a que arrojaba segmentaciones poco concluyentes. El aprendizaje automático no supervisado se utiliza para indagar y hallar patrones en datos sin etiquetar. Una de las técnicas de aprendizaje no supervisado es la de clustering K-means. Este algoritmo es un método de cuantificación vectorial, cuyo origen proviene del procesamiento de señales. Tiene como objetivo particionar n observaciones en k clusters o grupos, donde cada observación pertenece al clúster con la media más cercana (es decir, los centros de clúster o clúster centroide), actuando como prototipo del grupo.

[20] Ejemplo de ello, son los bosques aleatorios para árboles de regresión, que si bien representan uno de los algoritmos más importantes y usados en machine learning. Realizan predicciones sobre nuevas observaciones combinando las predicciones de todos los árboles que conforman el modelo. Su potencial radica en que métodos estadísticos y de machine learning basados en estos engloban técnicas supervisadas no paramétricas que consiguen fraccionar el espacio de predictores en regiones simples, donde es más sencillo manejar las interacciones. Estos modelos son más robustos

Dentro de los supervisados, se consideró la regresión lineal multivariada[21]. Los modelos de regresión lineal simple, múltiple o multivariado buscan obtener la regresión entre variables independientes y una variable dependiente. Es decir, teniendo una serie de variables predictoras, obtiene la relación con una variable cuantitativa a predecir; la regresión lineal explica la variable *y* con las variables *x*, generando una función lineal que mejor se ajusta. Pero en la práctica generalmente tenemos *n* variables predictoras, por ello podríamos hacer *n* regresiones lineales simples; sin embargo, cada una ignoraría a las otras *n-1* variables y no se sacaría ventaja de las relaciones entre variables. Por ende, será necesario considerar cómo dos o más variables independientes influyen sobre una variable dependiente (Hastie, Tibshirani y Friedman, 2009). Estos modelos pueden emplearse para predecir el valor de una variable dependiente o para evaluar la influencia que tienen los predictores sobre ella (esto último se debe analizar con cautela para no malinterpretar causa-efecto).

Adicionalmente, se incorporó un análisis estadístico de supervivencia utilizando la técnica de Kaplan-Meier (Kaplan y Meier, 1958). Este enfoque proporciona una comprensión detallada de la duración de eventos en función del tiempo, particularmente valioso para eventos de interés en el estudio. Se exploraron patrones de supervivencia a lo largo del tiempo, permitiendo identificar posibles variaciones en la duración de eventos relevantes para el análisis.

Finalmente, se tuvieron en consideración las localidades de los pacientes. Se puede entender que los sistemas de información geográficos (SIG) son una herramienta importante para el estudio, análisis y control epidemiológico, pues a la información médica, se suman datos geográficos, variables ambientales, socioeconómicas, entre otras, que en el contexto de la pandemia de COVID-19 podrían revelar los retrocesos en la prevención así como el empeoramiento de la calidad de vida de las personas por el aumento de la pobreza y la accesibilidad a los sistemas de salud (OPS, 2002; Cuello-Rüttler y Gudiño, 2017). Por lo anterior, se entiende que el contexto geográfico agrega información cualitativa a los datos, permitiendo revelar si existen o no patrones geográficos nacionales respecto a los enfermos de tuberculosis antes y después de la pandemia. Con este fin, se utilizaron las localidades censales, descargados en GeoJSON desde el Sistema de almacenamiento de archivos y catálogos de la Red de Nodos de Datos Abiertos de la Administración Pública Nacional[22]. Se contabilizaron los casos por cada localidad y se representaron en mapas con nodos proporcionales.

En resumen, el análisis abarcó desde la regresión, que modela las relaciones entre variables, hasta la georreferenciación, que explora la dimensión espacial de los datos, y culminó con un análisis de supervivencia que profundiza en la duración de eventos de interés a lo largo del tiempo. Esta combinación de enfoques proporciona una visión integral del conjunto de datos, permitiendo interpretaciones más robustas y contextualizadas.

3. Análisis de resultados

Los resultados se evaluaron comenzando con el recorte a la base de datos descargada, como se mencionó en la metodología. Se examinaron 36.391 casos de tuberculosis notificados en Argentina durante el período comprendido entre los años 2019 y 2021, con el objetivo de analizar posibles patrones y variaciones en la incidencia de la enfermedad, intermediada por el aislamiento social de la pandemia de COVID-19. Analíticamente se ha visualizado la cronología de los casos, dividiéndolos en aquellos notificados antes del inicio de la pandemia (n= 16.042) y aquellos registrados después de la declaración del ASPO (n=20.349). Estos valores corresponden únicamente a los reportados, sin duplicados, entre 2019 y 2021. De esta manera, se puede apreciar en esta distinción temporal la posible influencia de la pandemia en la notificación y gestión de casos de tuberculosis, brindando una perspectiva más completa sobre la dinámica de la enfermedad en el contexto de eventos sanitarios imperantes. En la tabla siguiente (Tabla 1a) se muestra la cantidad total de casos discriminados por sexo; en la tabla 1b se presentan los datos con filtros mencionados aplicados, incluso con duplicados eliminados, y en la tabla 2 los tratamientos que figuran como curados o con tratamiento

ante valores atípicos, pues asignan menor peso a las observaciones individuales, por lo que los valores atípicos no influyen tanto en la toma de decisiones.

[21] Puede verse más información sobre el tema en: Regresión lineal múltiple.

[22] Las localidades censales se encuentran disponibles desde: https://infra.datos.gob.ar/catalog/modernizacion/dataset/7/distribution/7.31/download/localidades-censales.geojson.

completo contra la TB.

Tabla *1a*. Casos totales según distinción por sexo.

Clasificación respecto a la pandemia / Sexo	Antes	Después
A	2.979	1.467
F	36.877	46.395
M	62.260	70.942
X	0	12

Fuente: Elaboración propia.

Tabla *1b*. Casos filtrados, sin duplicados, según distinción por sexo.

Clasificación respecto a la pandemia / Sexo	Antes	Después
A	553	374
F	6.504	8.455
M	8.985	11.518
X	0	2

Fuente: Elaboración propia.

Tabla *2*. Casos con tratamiento completo o curados.

Clasificación Pandemia / Resultado Tratamiento	Antes	Después
Curado	1.015	1.285
Tratamiento completo	4.680	6.724

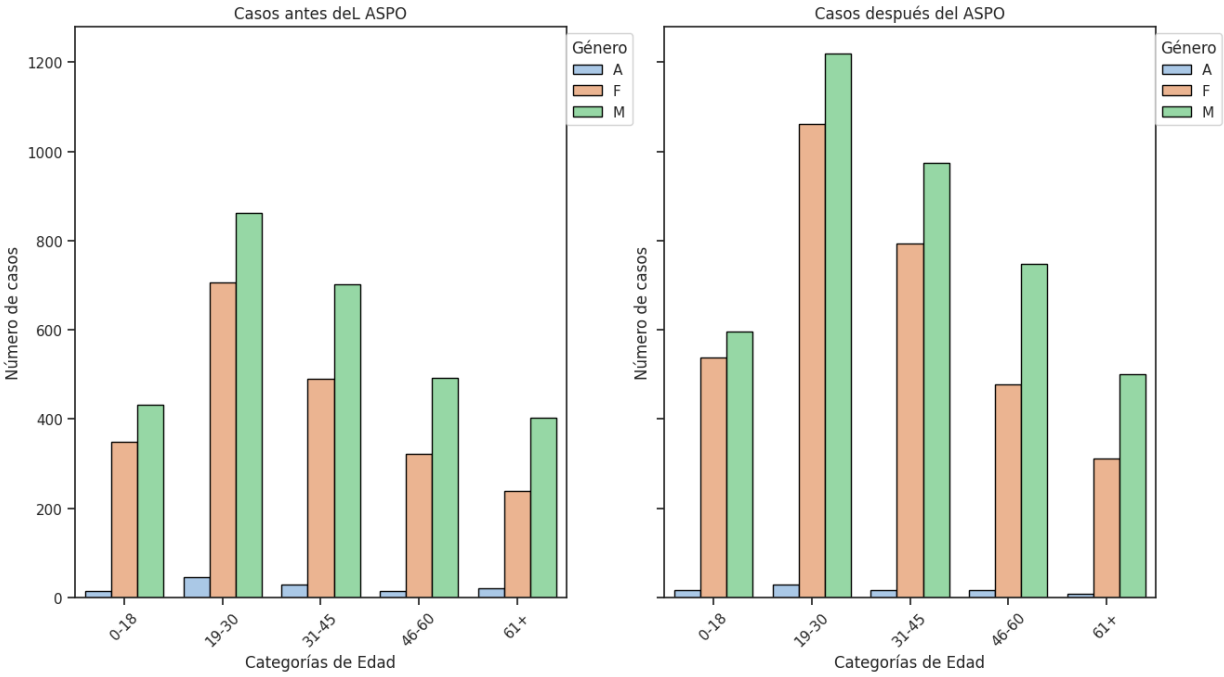
Fuente: Elaboración propia.

Es importante destacar que se optó por analizar los casos con resultado de tratamiento exitoso (curados y tratamiento completo; n=12.505), de los cuales 5.158 casos fueron notificados antes del ASPO y 7.347 casos después (Tabla 2). Al seleccionar casos con resultados exitosos, se buscó la variabilidad inherente en los resultados de tratamiento y reducir la influencia de factores confusos que podrían afectar la comparación, con el objetivo de mejorar la validez de las conclusiones derivadas del análisis.

3.1. Análisis descriptivo de los casos de tuberculosis notificados entre 2019 y 2021

La distribución de casos según grupos de edad se aprecia en el siguiente gráfico, dónde se destaca una mayor incidencia en el rango de 19 a 60 años durante ambos períodos analizados.

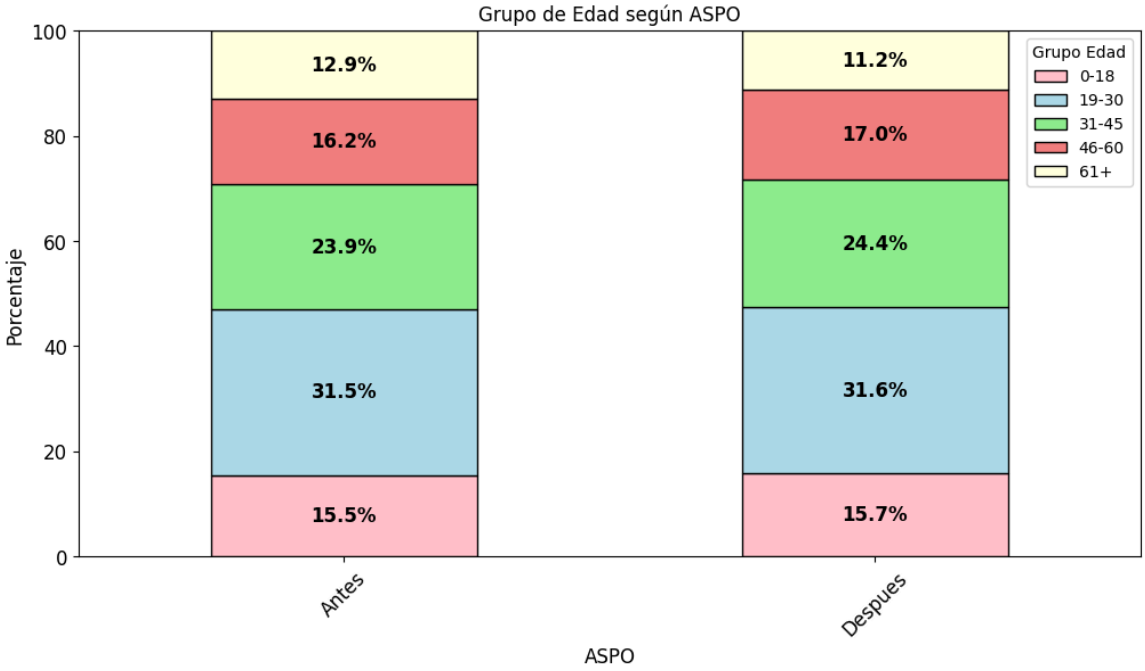
Gráfico 2. Notificación total de casos de tuberculosis notificados antes y después del ASPO según grupo edad, en Argentina.



Fuente: elaboración propia

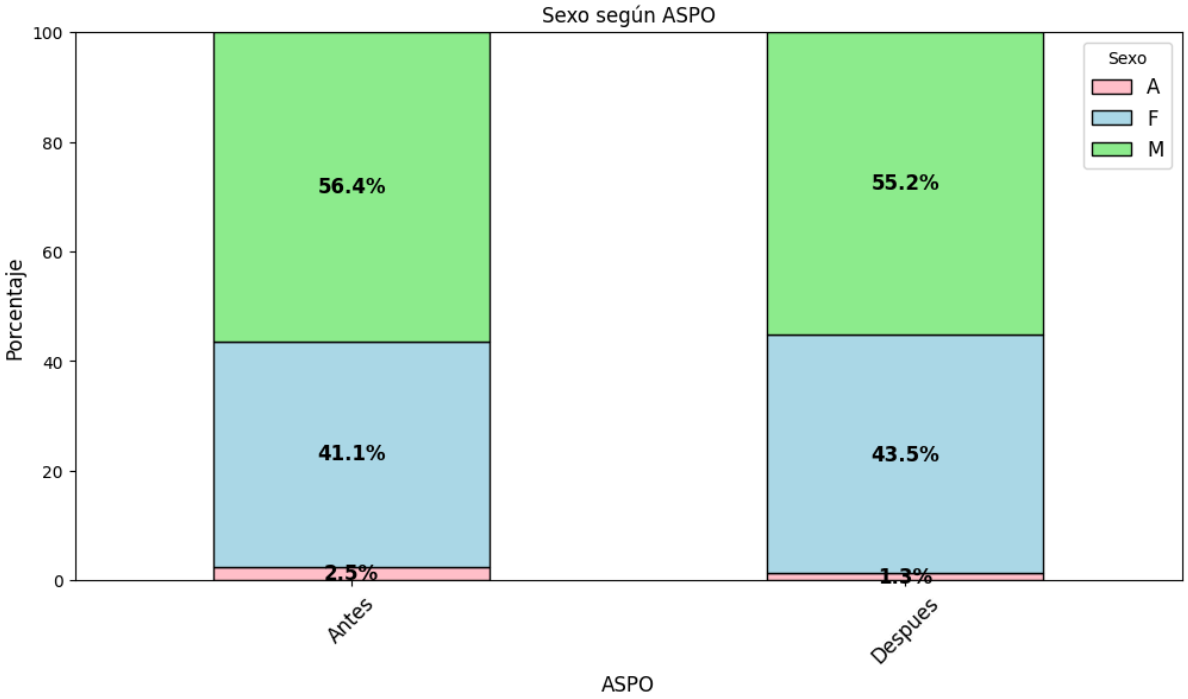
Asimismo, la partición porcentual para los grupos de edad (Gráfico 3) y género (Gráfico 4) evidencia una marcada semejanza.

Gráfico 3. Distribución de la proporción de casos notificados antes y después del ASPO según grupo edad.



Fuente: elaboración propia

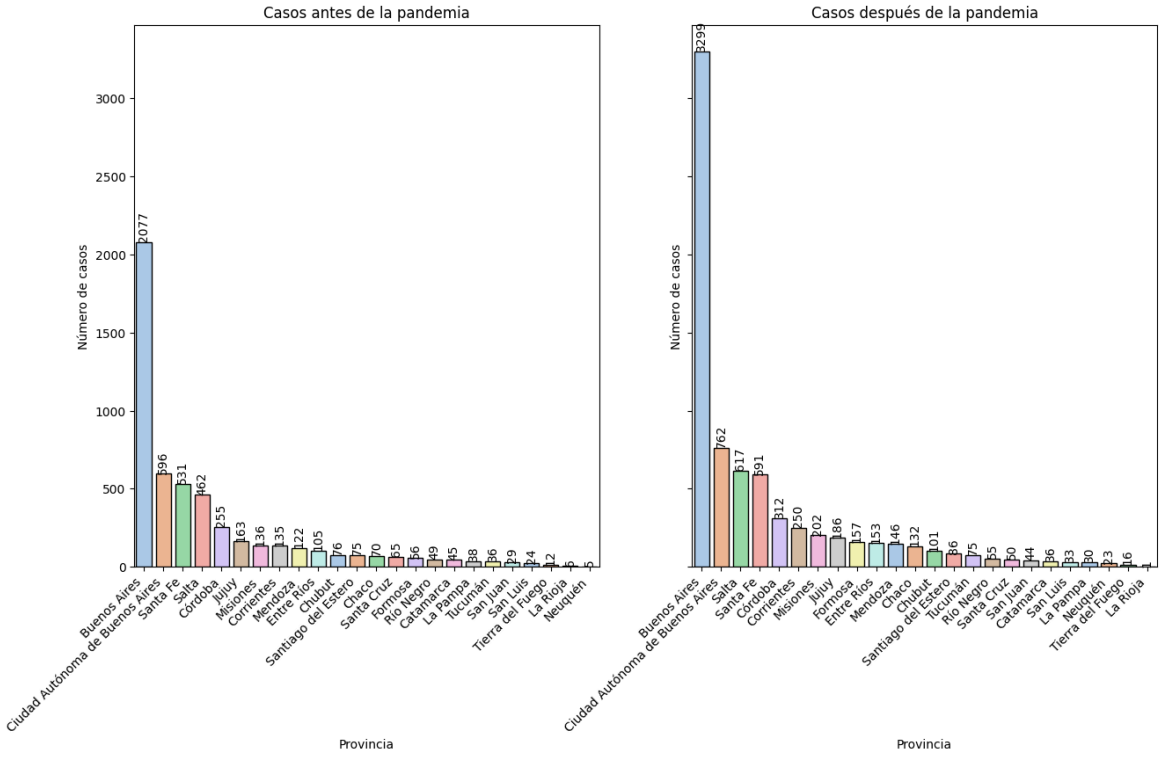
Gráfico 4. Distribución de la proporción de casos notificados antes y después del ASPO según género. (Argentina, 2019 - 2021).



Fuente: elaboración propia

Al considerar pacientes por provincia de residencia (Gráfico 5), la gráfica presenta un aumento en las notificaciones posteriores a la declaración de ASPO durante la pandemia. Las que presentaron mayor cantidad de casos continúan siendo, como antes del 2020, Buenos Aires, Ciudad Autónoma de Buenos Aires (CABA), Santa Fe, Salta y Córdoba.

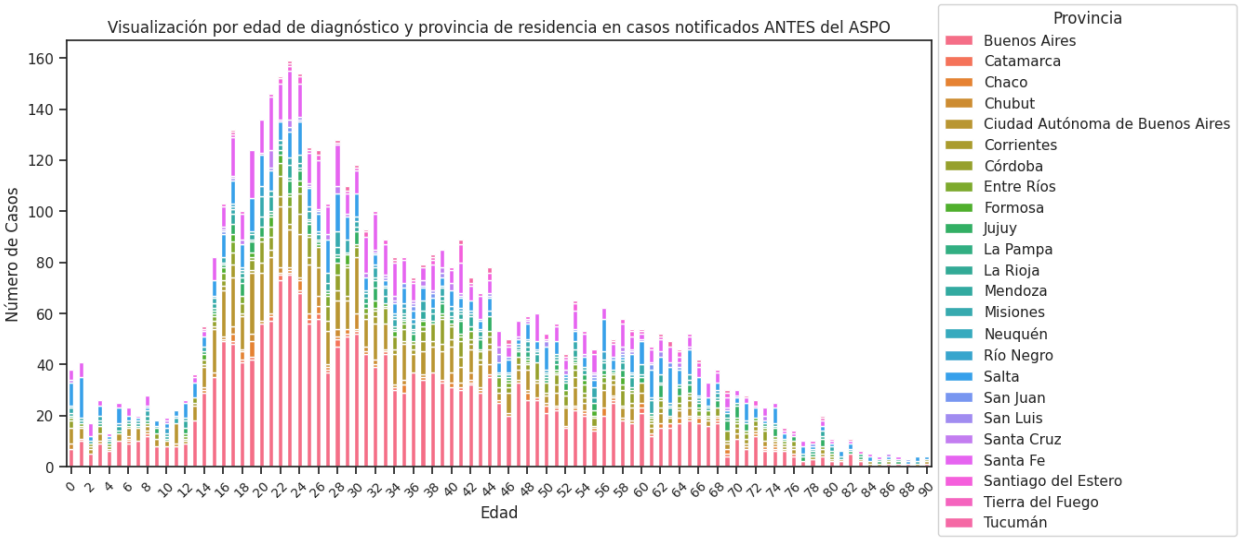
Gráfico 5. Distribución del número de casos notificados antes y después del ASPO según provincia de residencia. (Argentina, 2019 - 2021).



Fuente: elaboración propia

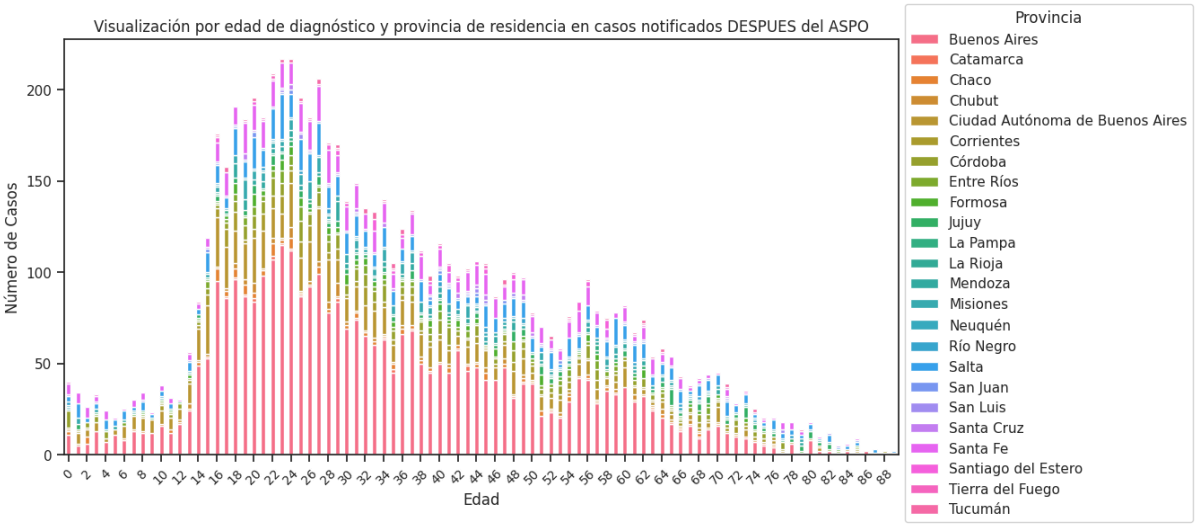
Por otro lado, la variaciones en edades y en provincia de residencia, antes de la pandemia y después del ASPO revela una notable similitud en ambas gráficas (Gráficos 6a y b), con una predominancia de casos notificados en el grupo de pacientes comprendido entre 15 y 40 años.

Gráficos 6a: Distribución por edad de diagnóstico según provincias de residencias en casos de tuberculosis notificados antes del ASPO.



Fuente: Elaboración propia.

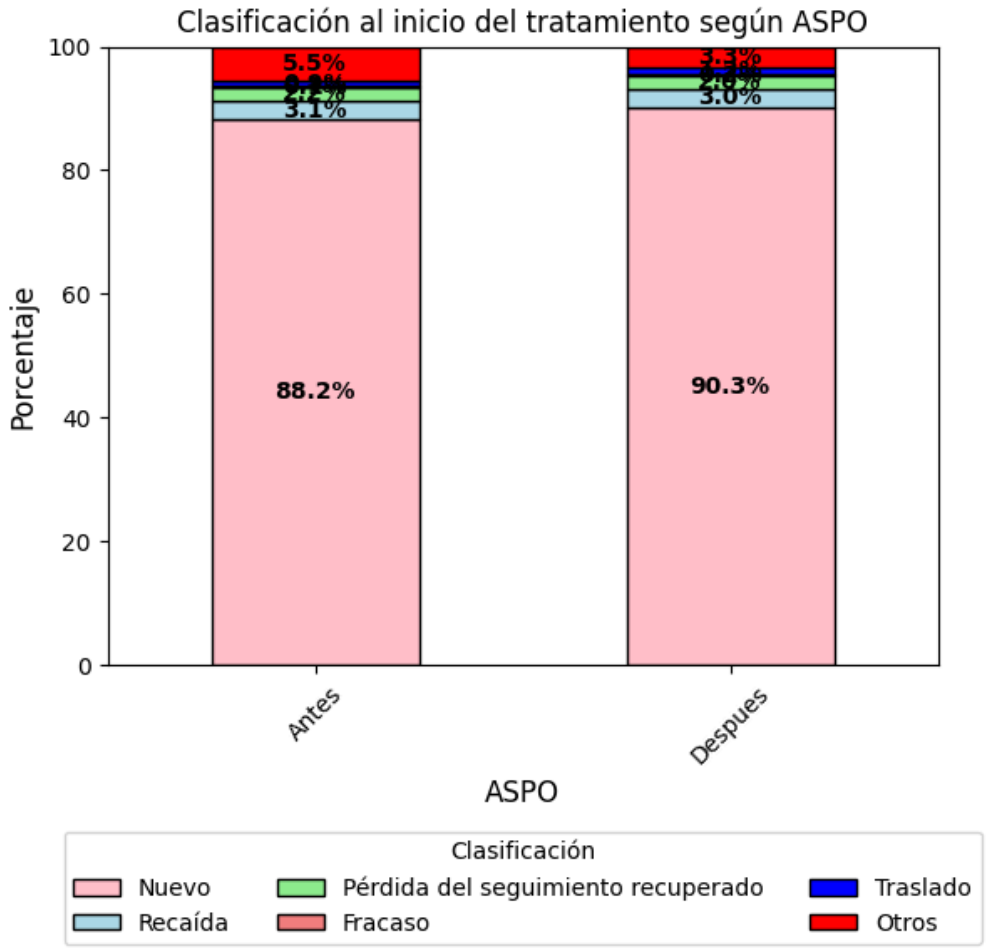
Gráficos 6b: Distribución por edad de diagnóstico según provincias de residencias en casos de tuberculosis notificados después del ASPO.



Fuente: elaboración propia.

En el siguiente gráfico (Gráfico 7) se visualiza que más del 80% de los casos en ambos grupos estudiados se clasificaron como nuevos, es decir que nunca habían sido tratados por tuberculosis o recibieron medicamentos anti-TB al menos por un mes.

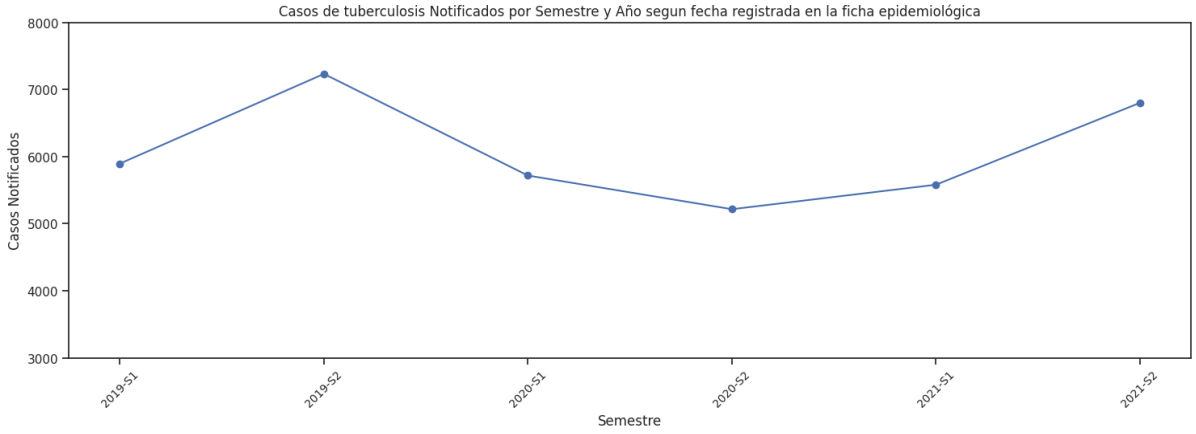
Gráfico 7. Distribución de casos notificados antes y después del ASPO según clasificación del caso al inicio del tratamiento. (Argentina, 2019 - 2021).



Fuente: elaboración propia

A continuación (Gráfico 8), en la evolución semestral de casos notificados, se observa una disminución en los registros a partir del primer trimestre de 2020 en comparación con el segundo semestre del año anterior. Mientras que desde el primer semestre de 2021, se percibe un aumento, sumado a que la tendencia se mantiene hasta el segundo semestre del mismo año.

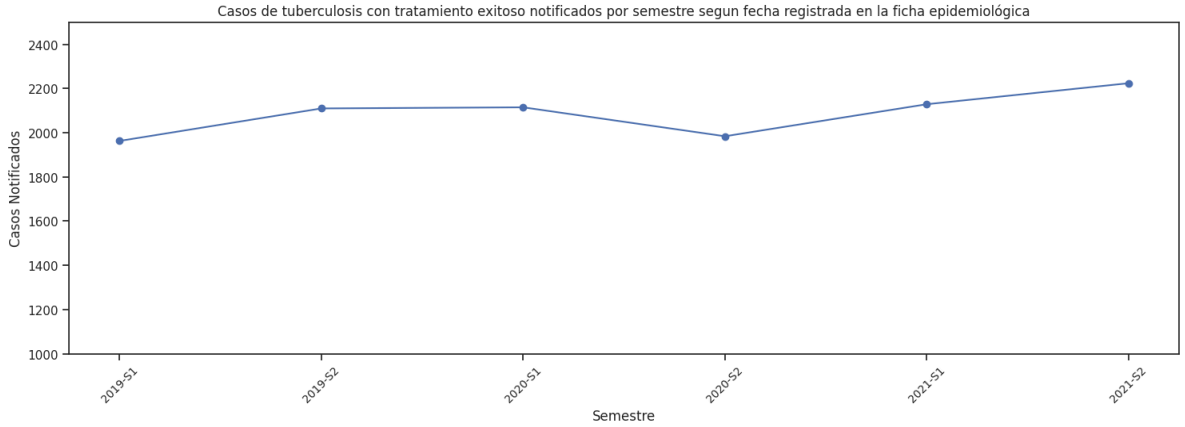
Gráficos 8. Evolución por semestre del total de casos de tuberculosis notificados, según fecha registrada en la ficha de notificación epidemiológica, en Argentina (2019-2021).



Fuente: elaboración propia

Cabe destacar que al analizar la evolución de las notificaciones de tratamientos exitosos, si bien se observa un patrón similar al del total de casos, esta disminución no es tan marcada como en el total de casos (Gráfico 9).

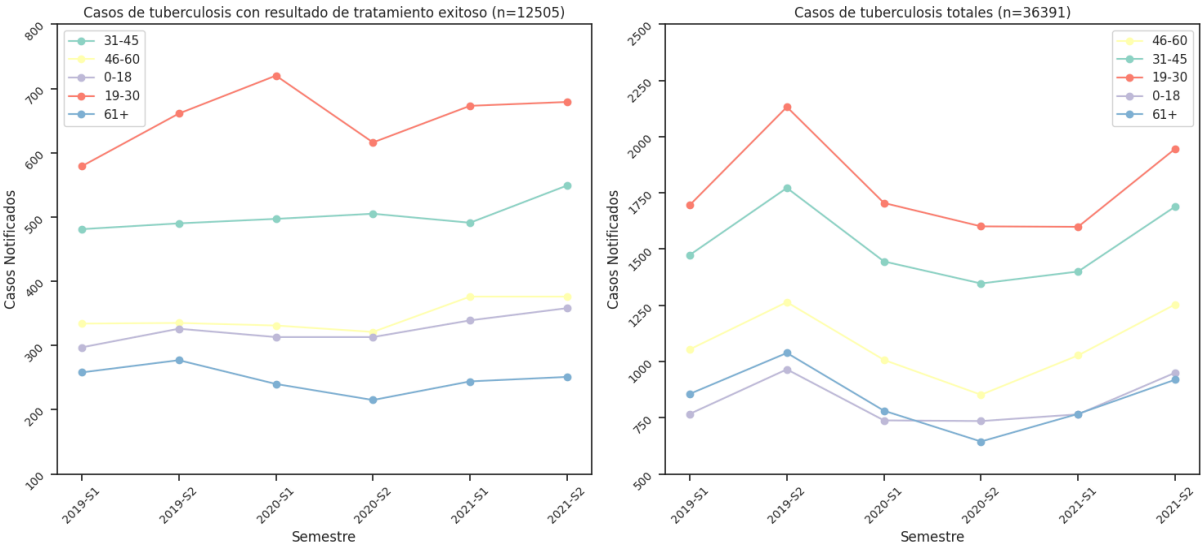
Gráficos 9. Evolución por semestre casos de tuberculosis notificados con resultado de tratamiento exitoso, según fecha registrada en la ficha de notificación epidemiológica, en Argentina (2019-2021).



Fuente: elaboración propia

Sin embargo, al examinar por grupo de edad (Gráfico 10), se destaca una variación consistente en la mayoría de los conjuntos durante el periodo considerado, con una excepción notable en los menores de 18 años. Este último, presenta un descenso durante el primer semestre de 2020, que persiste en el segundo semestre, y experimenta un aumento en el primero del año siguiente. De manera similar, al focalizar solo en casos con tratamiento exitoso existe una mayor variabilidad en cada agrupamiento etario. Sin embargo, es notable que dentro del grupo de 0 a 18 años y mayores de 31 años la variabilidad no es tan marcada.

Gráficos 10. Evolución semestral de la incidencia de casos de tuberculosis totales y casos con resultado de tratamiento exitoso, por grupo edad según fecha registrada en la ficha de notificación epidemiológica, en Argentina (2019-2021).



Fuente: elaboración propia

3.2. Análisis del tiempo de demora en el diagnóstico y tratamiento de la tuberculosis.

Para este análisis se han generado histogramas (Gráfico 11 y 12) que evalúan la distribución de tiempos de demora (calculado en días) en los casos notificados antes y después del ASPO. Estos gráficos revelan asimetría hacia la derecha, donde la media supera tanto a la mediana como a la moda en ambas categorías. Se puede entender que este patrón sugiere una distribución no simétrica. De manera adicional el histograma del

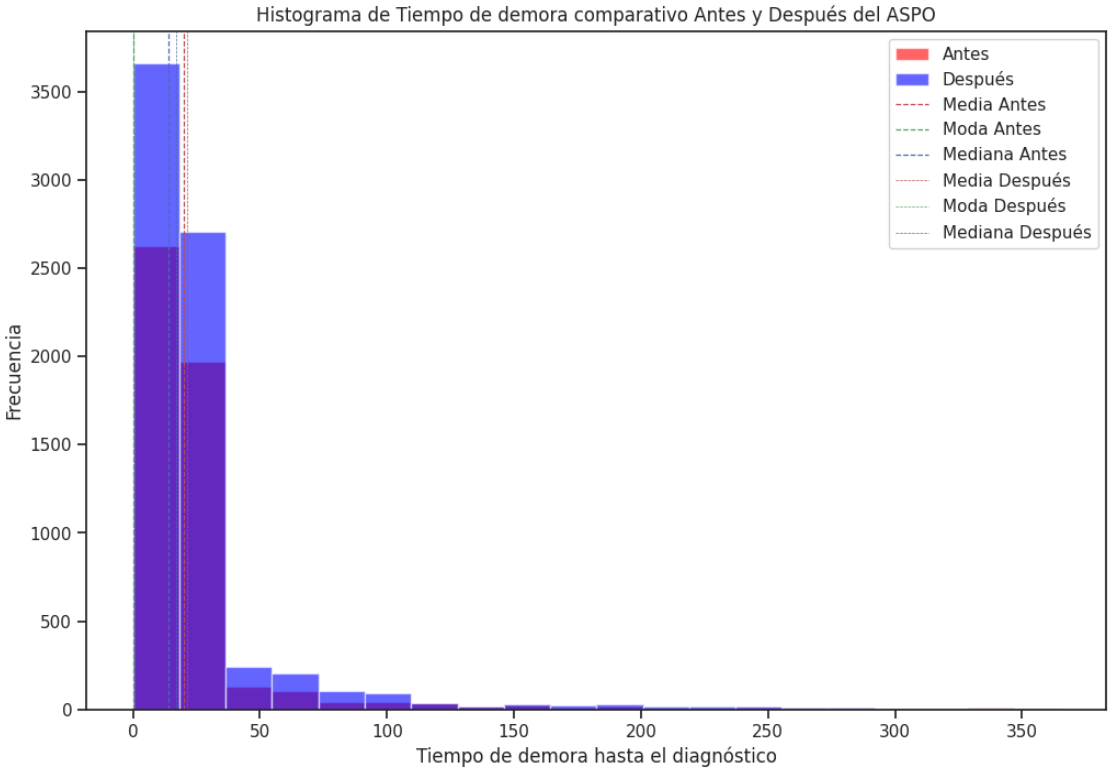
tiempo de demora en el tratamiento (TDT) destaca la presencia de dos modas, indicando una distribución bimodal. Por otro lado, en la Tabla 3, se evidencia una leve variabilidad en tiempos de demora, antes y después de la pandemia, análogamente ocurre en la media como en la mediana. Es decir, que la media aumentó aproximadamente día y medio, mientras que la mediana se incrementó en 3 días, tanto en la demora en el diagnóstico como en el tratamiento.

Tabla 3: Medidas de tendencia central del tiempo de demora hasta el diagnóstico y hasta completar el tratamiento de casos de Tuberculosis con Tratamiento Exitoso.

Clasificación pandemia	Tiempo de Demora (Diagnóstico) - Media	Tiempo de Demora (Diagnóstico) - Mediana	Tiempo de Demora (Tratamiento) - Media	Tiempo de Demora (Tratamiento) - Mediana
Antes	22.55	14.0	227.36	216.0
Después	23.81	17.0	229.86	219.0

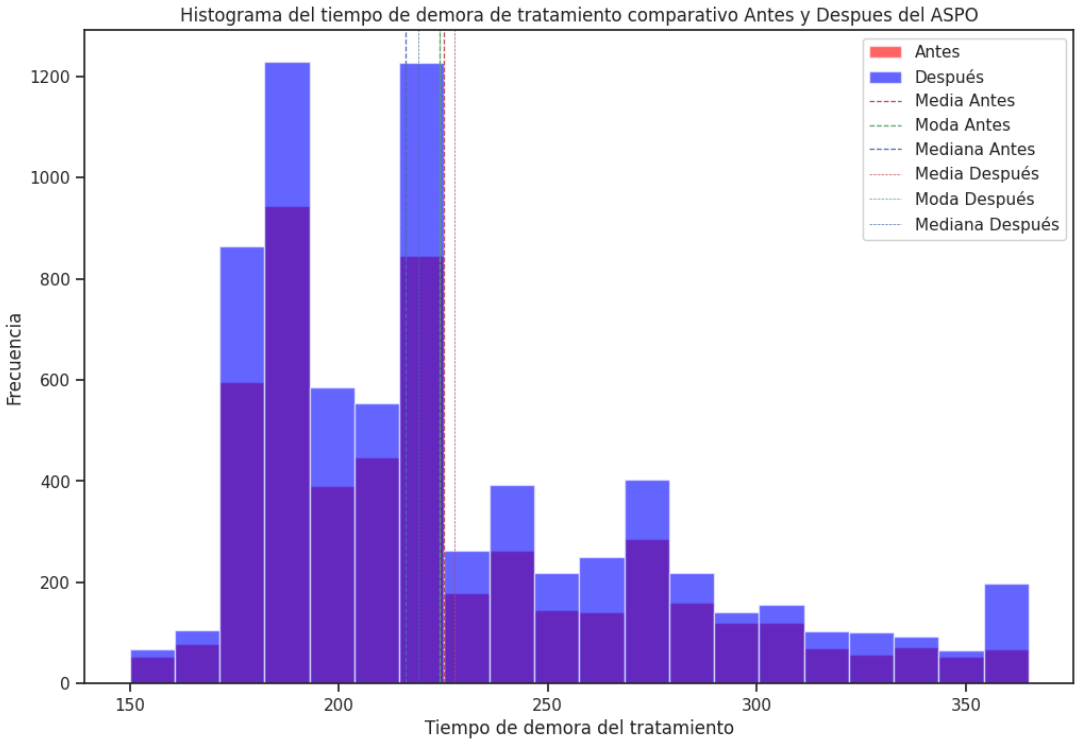
Fuente: elaboración propia.

Gráfico 11. Distribución del Tiempo de Demora en el Diagnóstico de Casos de Tuberculosis con Tratamiento Exitoso.



Fuente: elaboración propia

Gráfico 12. Distribución del Tiempo de Demora hasta completar el tratamiento en Casos de Tuberculosis con Tratamiento Exitoso.

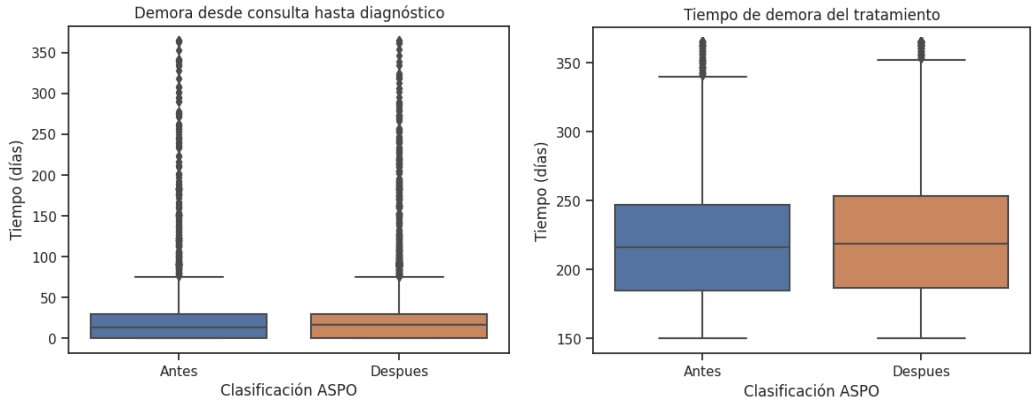


Fuente: elaboración propia

Cabe destacar que si la mediana es mayor a la media en un conjunto de datos, sugiere que la distribución de los datos está sesgada hacia la derecha, es decir, por ende refleja la existencia de valores atípicos o extremadamente altos. Estos son los que aumentan la mediana en comparación con la media. Por ello, en este tipo de situaciones podría ser más pertinente utilizar la mediana como medida de tendencia central en lugar de la media, pues la mediana es menos sensible a los valores atípicos.

Los valores atípicos se presentan en el gráfico 13, sin reflejar variaciones considerables en el tiempo de demora hasta el diagnóstico. Similarmente, la mediana como los cuartiles son similares en ambos periodos. No obstante, luego del ASPO se aprecia un ligero aumento en el tiempo de demora hasta finalizar el tratamiento, y un mayor intervalo de confianza, lo que indica una mayor variabilidad en los tiempos de demora.

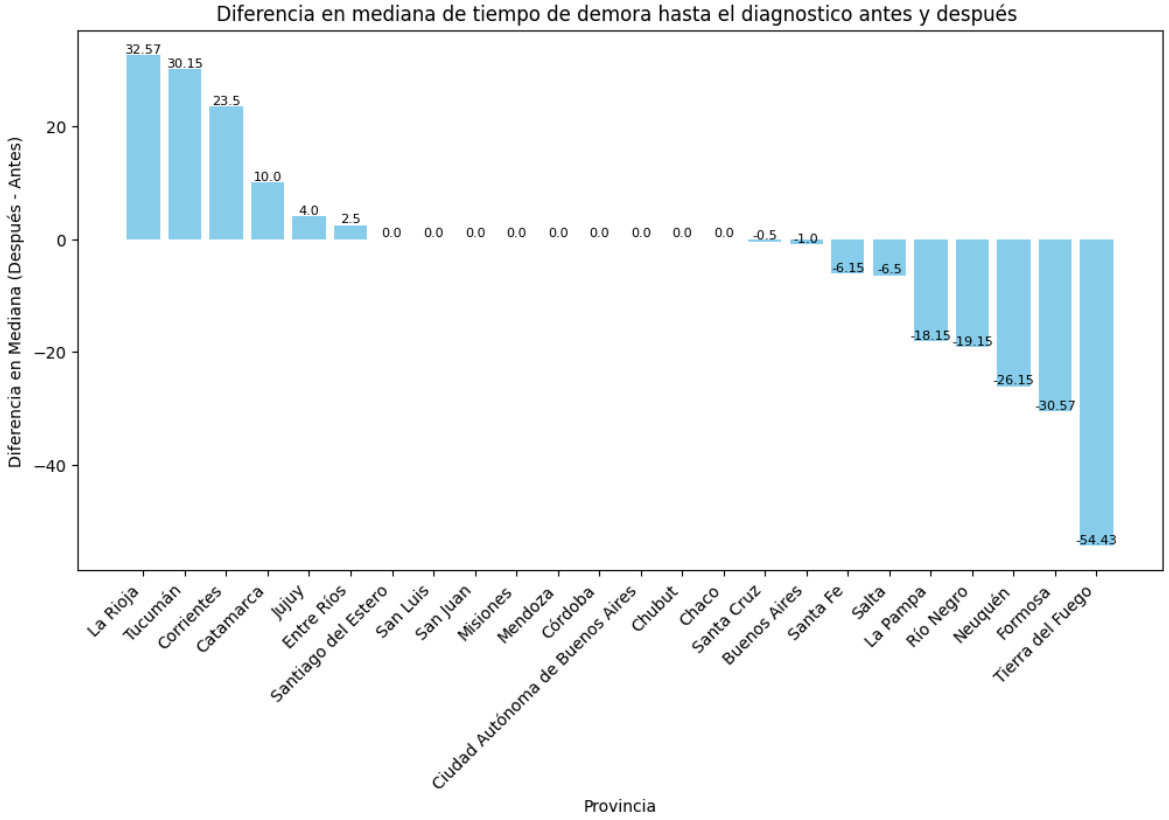
Gráfico 13: Tiempos de Demora en el Diagnóstico y Tratamiento de casos de Tuberculosis en Argentina según momento de notificación (2019-2021).



Fuente: elaboración propia.

En cambio, al analizar cada provincia (Gráfico 14), se visualiza que la diferencia en la mediana del tiempo hasta el diagnóstico no se comporta de la misma manera que en el total de los datos. Se aprecian importantes disimilitudes, por ejemplo en algunas provincias no presentan distinciones entre los periodos analizados, en tanto que en otras como La Rioja, Tucumán, Corrientes, Catamarca, Jujuy y Entre Ríos el tiempo de demora fue mayor después de declararse el ASPO. Por el contrario en Santa Fe, Salta, La Pampa, Río Negro, Neuquén, Formosa y Tierra del Fuego fue menor.

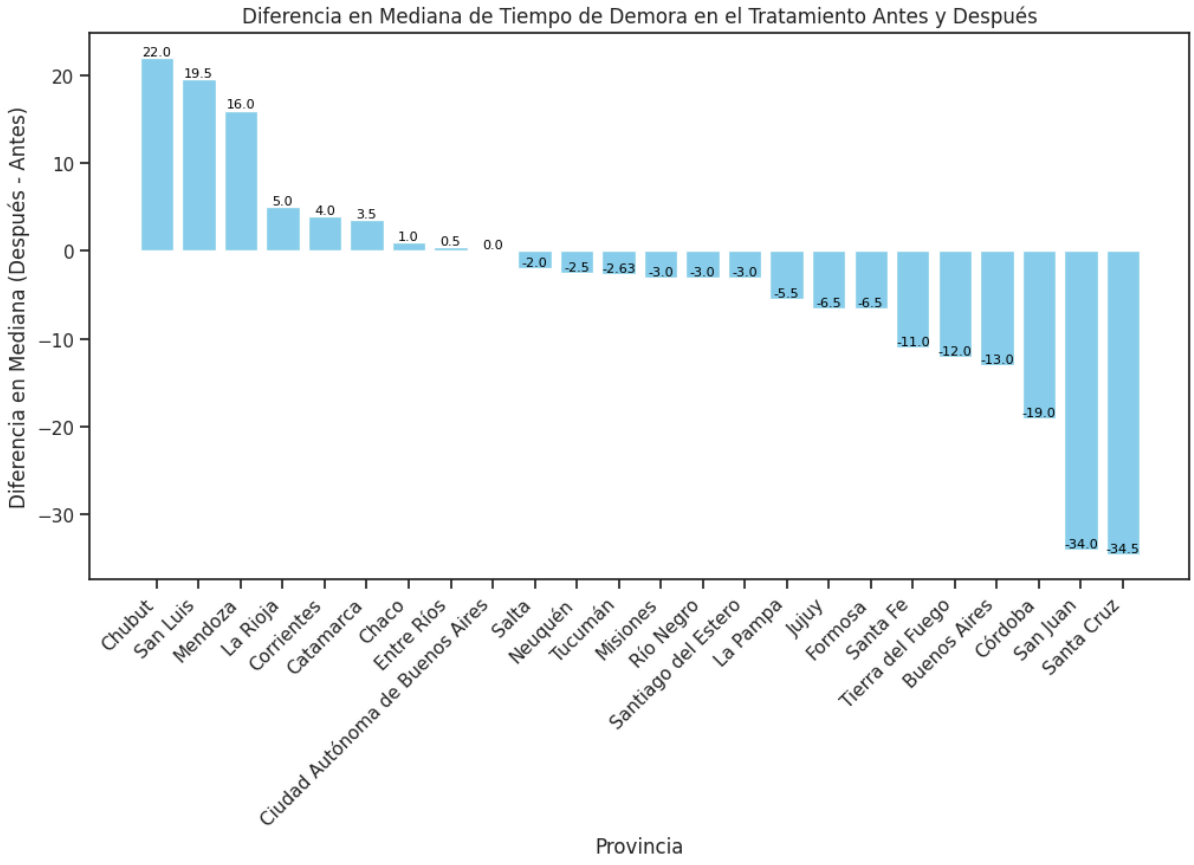
Gráfico 14: Diferencia de la mediana en el tiempo de demora hasta el diagnóstico de casos de Tuberculosis con Tratamiento Exitoso en según provincia en Argentina (2019-2021).



Fuente: elaboración propia.

La diferencia en los tiempos de demora hasta completar el tratamiento (Gráfico 15) se destaca una menor variación en las provincias de Entre Ríos y CABA, inferior a un día, tanto antes como después del ASPO. En contraste, Chubut, San Luis y Mendoza experimentaron un aumento, que superaron los 15 días de diferencia en el tiempo de tratamiento después del ASPO. No obstante, en San Juan y Santa Cruz, a pesar de presentar una mayor dispersión, reducen sus tiempos de tratamiento posterior al aislamiento, con discrepancias superiores a 30 días. Estos datos resaltan los antagonismos del impacto de la pandemia en los tiempos de tratamiento entre distintas provincias, con incrementos y reducciones en la demora. Lo anterior, acentúa lo fundamental de mejorar la eficiencia en la prestación de servicios de salud en cada región.

Gráfico 15: Diferencia de la mediana en el tiempo de demora hasta completar el tratamiento de casos de Tuberculosis con Tratamiento Exitoso en según provincia en Argentina (2019-2021).



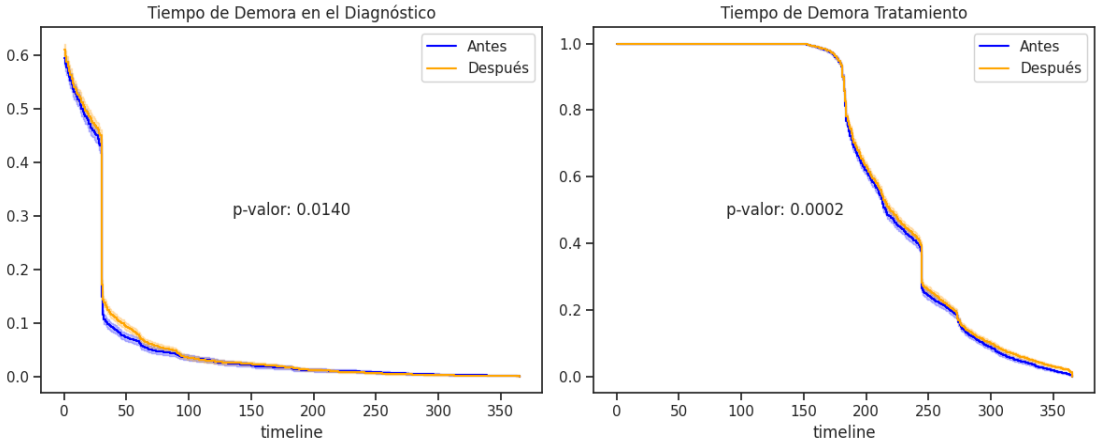
Fuente: elaboración propia.

3.3 Análisis estadístico de supervivencia

Los gráficos de supervivencia grafican la probabilidad de que los eventos, en este caso los tiempos de demora en el diagnóstico y de demora en el tratamiento, no hayan ocurrido hasta un tiempo dado. En el contexto del presente artículo se tienen en cuenta dos periodos distintos *antes* y *después* de la declaración del ASPO, en el gráfico 16 las líneas azul y naranja, respectivamente, representan la supervivencia estimada. En ambos conjuntos, las curvas difieren entre dichos lapsos, y sugieren cambios en la distribución de los tiempos de demora. Es por ello, que para respaldar estas observaciones, se utilizó la prueba de *log-rank*, herramienta estadística que evalúa la existencia de diferencias significativas en las funciones de supervivencia entre dos grupos. Por medio de este test se obtuvo que los *p-valores* son menores a 0.05[23] para ambos conjuntos de datos, e indicó que existen diferencias estadísticamente significativas entre los tiempos de demora en el diagnóstico y en la demora en el tratamiento respecto al antes y después del aislamiento. En consecuencia, se concluye que la clasificación mencionada está asociada a los cambios significativos en estos tiempos.

[23] Se señala que el p-valor = 0.014 en la demora en el diagnóstico y el p-valor = 0.0002 en la demora en el tratamiento.

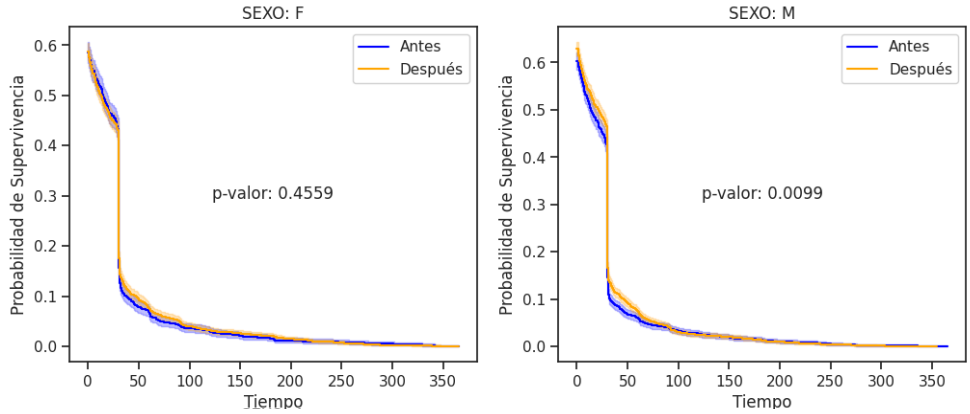
Gráfico 16. Análisis de Supervivencia: Tiempos de Demora en Casos Notificados antes y después del ASPO en Argentina (2019-2021)



Fuente: elaboración propia.

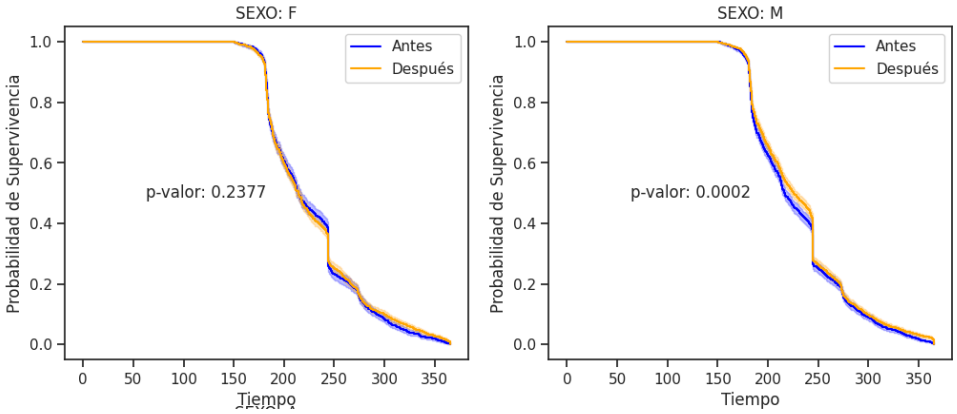
Al considerar los sexos en gráficos de supervivencia (Gráficos 17 y 18), en el femenino, no se observan diferencias significativas en ninguno de los tiempos evaluados. Esta consistencia sugiere que las dinámicas en torno al diagnóstico y tratamiento de la tuberculosis en dicho grupo permanecieron relativamente estables. En contraste, en el grupo masculino, la diferencia significativa se refleja antes y después del ASPO en la demora en el diagnóstico y en la duración del tratamiento. Estas evidencias indican que las medidas de aislamiento reflejaron un impacto específico en los patrones de la TB en hombres, introduciendo variaciones notables en la probabilidad de ser diagnosticado y en la progresión durante el tratamiento.

Gráfico 17. Análisis de Supervivencia: Tiempos de Demora en el Diagnóstico de Casos Notificados antes y después del ASPO en Argentina (2019-2021)



Fuente: Elaboración propia.

Gráfico 18. Análisis de Supervivencia: Tiempos de Demora en el Tratamiento de Casos Notificados antes y después del ASPO en Argentina segun sexo (2019-2021)

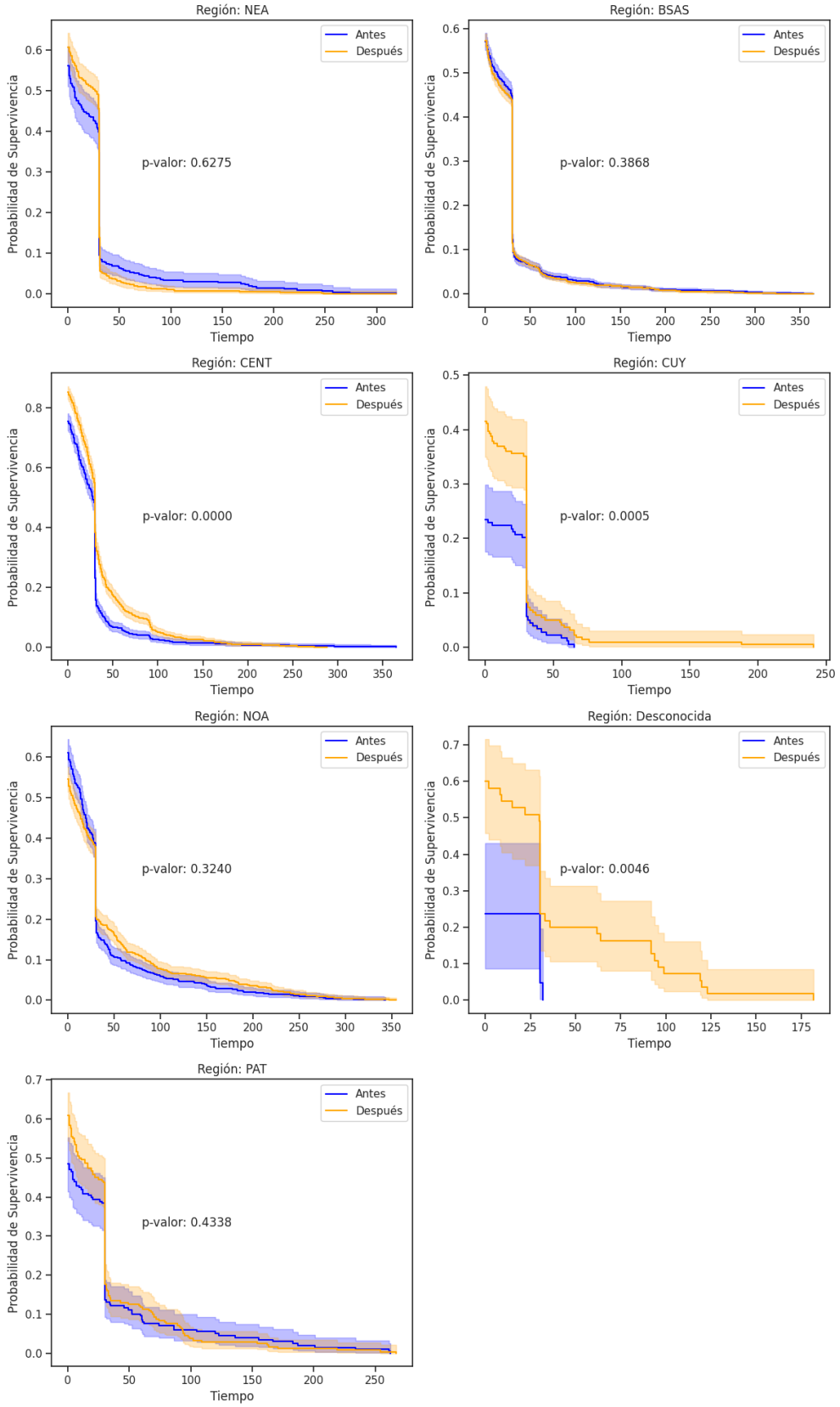


Fuente: elaboración propia.

En lo que respecta a tiempos de demora en el diagnóstico según las diferentes regiones de Argentina (Gráfico 19)[24], se contempla que no hay uniformidad en la probabilidad de supervivencia. En particular, las regiones Centro y Cuyo exhiben diferencias estadísticamente significativas en ambos periodos (*p-valor* < 0.05), y sugiere que las variaciones en los tiempos de demora antes y después del ASPO son significativas y no pueden atribuirse simplemente al azar. En cambio, en las regiones restantes, los *p-valores* no alcanzaron el umbral crítico de 0.05, es decir no hay diferencias significativas. Esto implica que las disparidades en tiempos de demora podrían deberse a variaciones aleatorias o no sistemáticas, en lugar de patrones discernibles.

[24] Se hizo una separación en las siguientes regiones. NOA: Catamarca, Jujuy, Salta, Santiago del Estero y Tucumán; NEA: Chaco, COrrientes, Formosa, Misiones; Patagonia: Chubut, Neuquén, Río Negro, Santa Cruz y Tierra del Fuego; Centro: Córdoba, Entre Ríos, Santa Fe, La Pampa; Cuyo: Mendoza, La Rioja, San Juan, San Luis; BSAS: Buenos Aires y CABA.

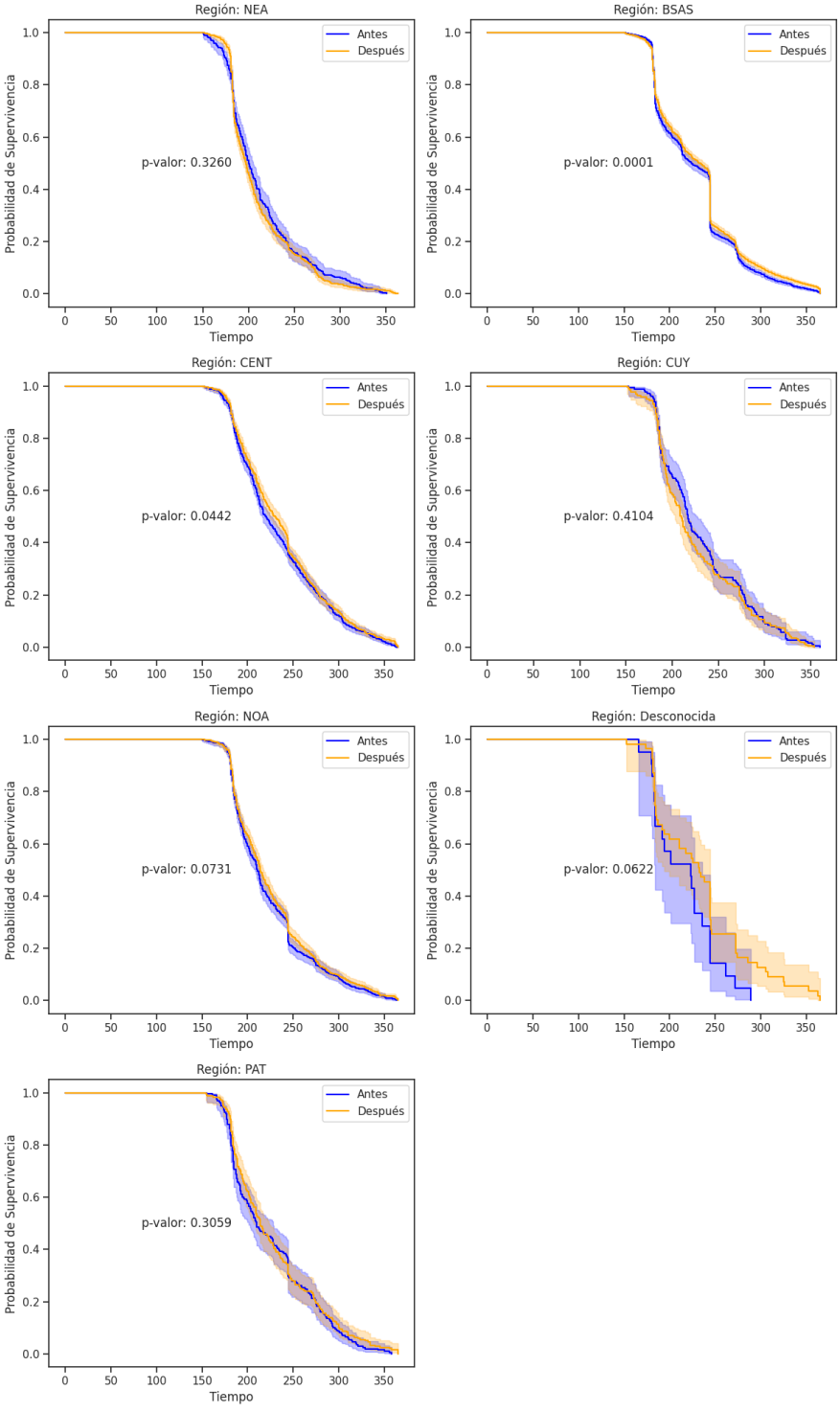
Gráfico 19. Análisis de Supervivencia: Tiempos de Demora en el Diagnóstico de Casos Notificados antes y después del ASPO en Argentina (2019-2021)



Fuente: elaboración propia.

En lo que respecta a los tiempos de demora en el tratamiento (Gráfico 20), se observa que las regiones de Buenos Aires y Centro, fueron las únicas que mostraron un p-valor significativamente inferior a 0.05. Estas diferencias podrían estar relacionadas con diversos factores, como la infraestructura de salud, los recursos disponibles o políticas regionales específicas que fueron tomadas durante el ASPO e impactaron en la eficiencia de los procesos de tratamiento.

Gráfico 20. Análisis de Supervivencia: Tiempos de Demora en el Tratamiento de Casos Notificados antes y después del ASPO en Argentina (2019-2021)



Fuente: elaboración propia.

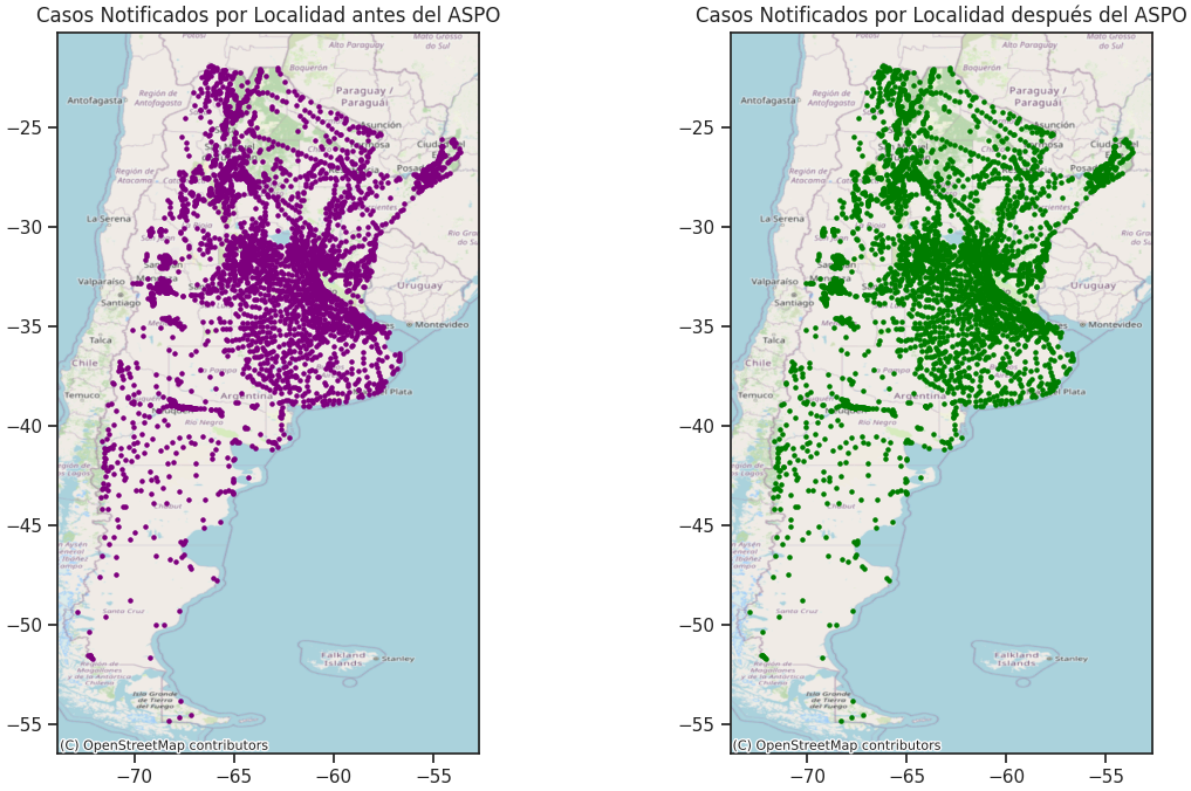
3.4 Georreferenciación de datos

En metodología se refirió a las herramientas y bases de datos utilizadas para realizar el solapamiento de datos georreferenciados y presentar las proporciones entre cantidad de casos y regiones, provincias o localidades.

3.4.1 Visualización de casos en localidades antes de la pandemia

Las georreferencias de localidades (Gráfico 23) reportaron los casos antes como después de la pandemia buscando evaluar posibles patrones geográficos. Asimismo, no se presentan diferencias sustanciales, presentando distribuciones uniformes en ambos periodos así como en regiones.

Gráfico 23: Georreferenciación de casos de tuberculosis según localidad de residencia de casos notificados antes y después del ASPO.

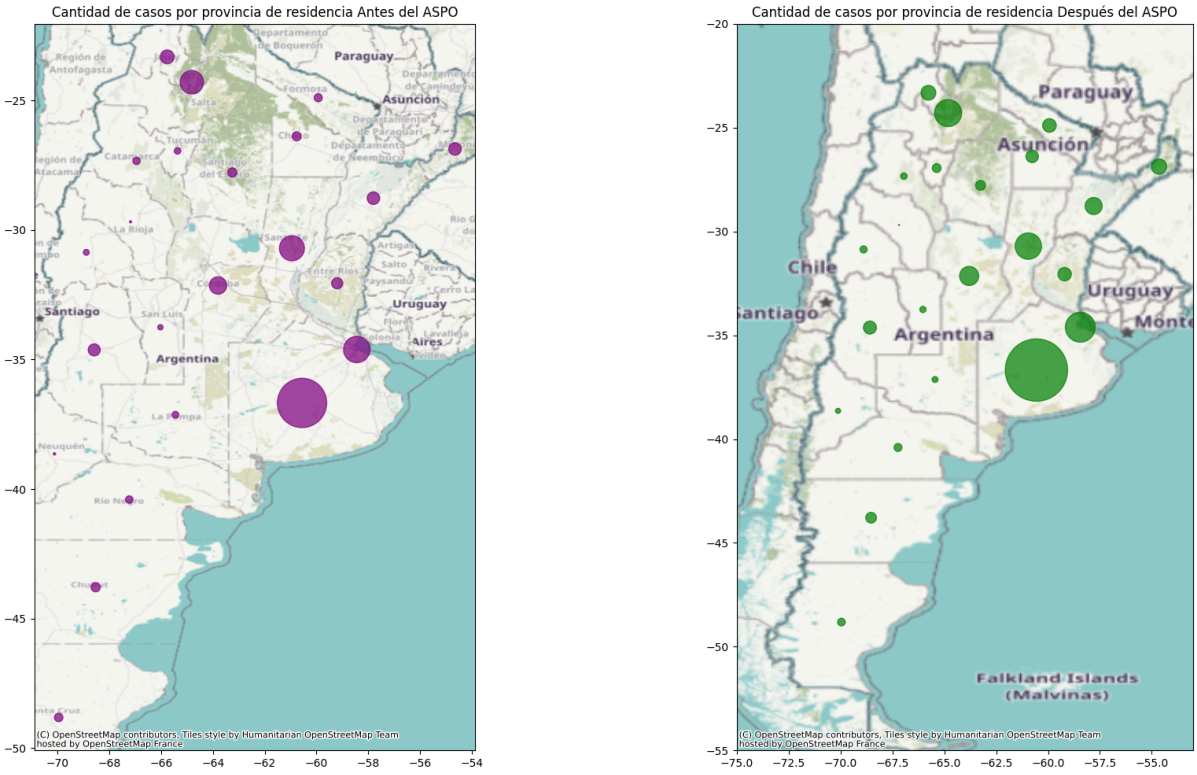


Fuente: elaboración propia.

3.4.2. Mapa con conteo de casos por provincias

Al momento de considerar la relación entre provincias y densidad de casos, antes y después de la pandemia, el patrón espacial refleja similitudes, aunque solo evidencia mayor cantidad de casos en algunas provincias, como la de Buenos Aires (Gráfico 24).

Gráfico 24. Georreferenciación según densidad de conteo de casos notificados por provincia antes y después del ASPO.



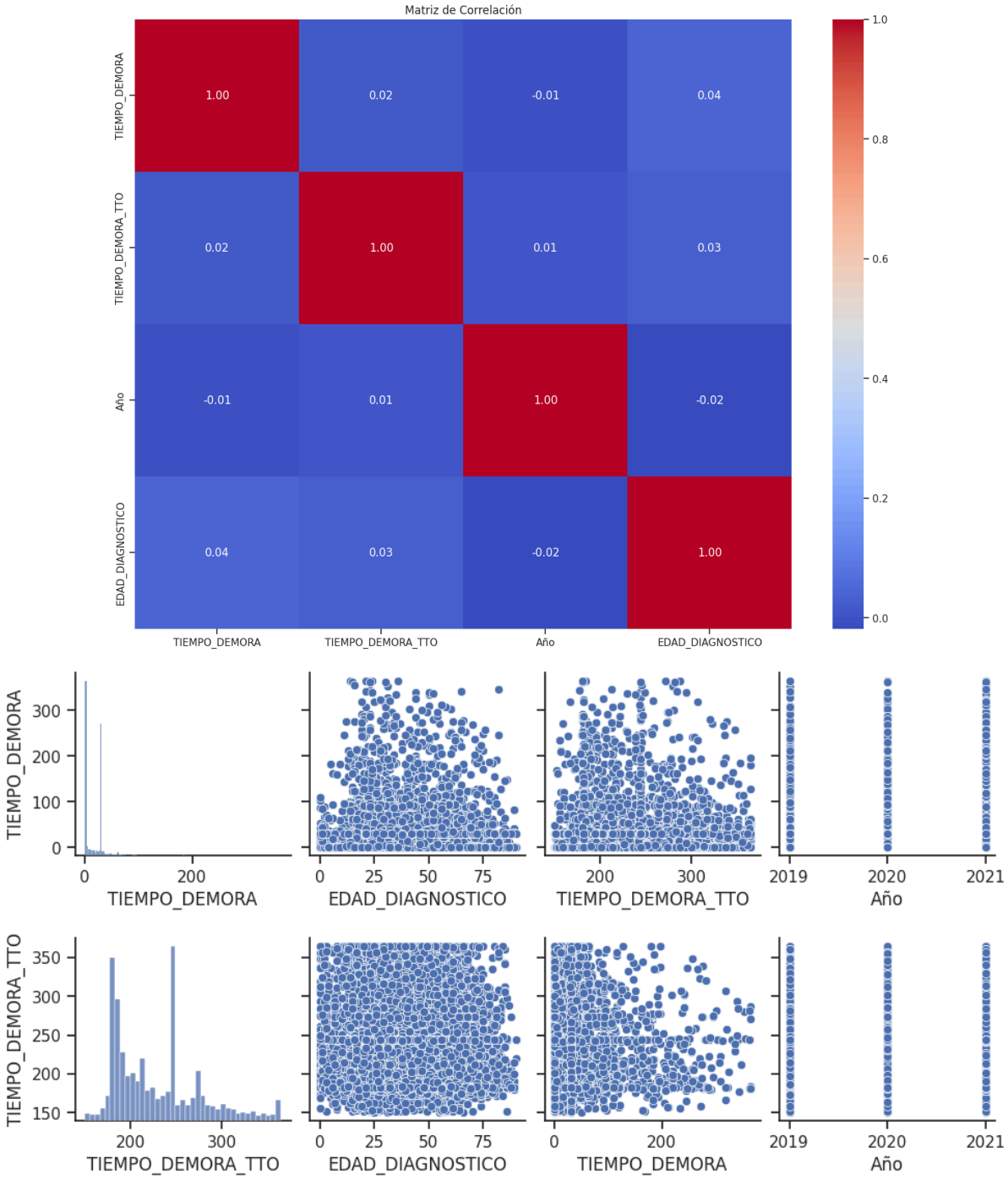
Fuente: elaboración propia.

3.4. Análisis de correlación

Antes de continuar con la regresión multivariada, debe presentarse cómo se relacionan entre sí las variables predictoras y la de respuesta. Si existen fuertes correlaciones entre las predictoras, podría hallarse multicolinealidad, y ello podría afectar la precisión e interpretación del modelo de regresión.

Por lo cual, en los tres gráficos siguientes (Gráfico 21a, b y c) se puede observar que, las cuatro variables analizadas, tiempo de demora en el diagnóstico (TDD), tratamiento, edad y año de notificación presentan un valor cercano a cero. Este valor indica una correlación muy débil entre las variables, es decir que no hay una relación lineal fuerte entre las mismas.

Gráfico 21a, b y c. Análisis de correlación entre variables cuantitativas de casos notificados antes y después del ASPO.

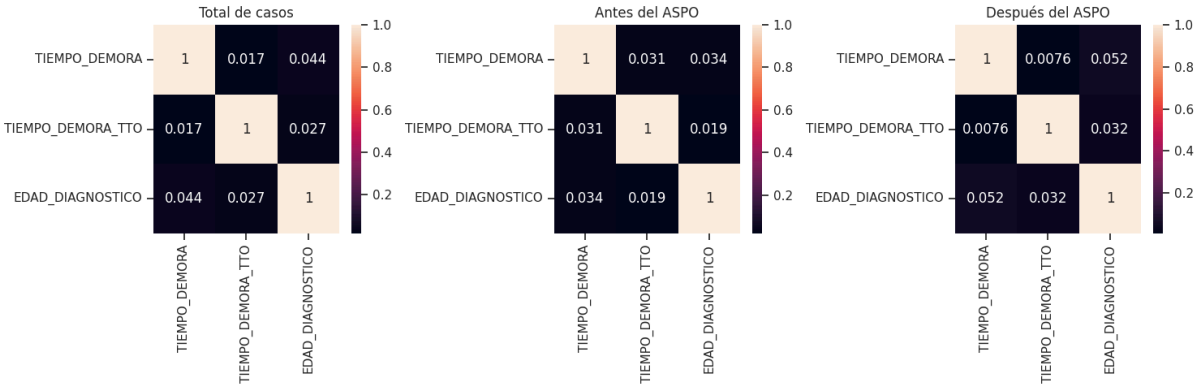


Fuente: elaboración propia.

Si se analiza por momento de diagnóstico (Gráfico 22) tampoco se aprecia correlación entre edad y tiempos de demora en ninguno de los períodos. Por ende, no se constata una relación lineal entre la edad de los pacientes y el TDD (tiempo de demora en el diagnóstico) y TDT (tiempo de demora en el tratamiento). En otras palabras, la edad de los pacientes no influiría en la cantidad de tiempo que transcurre antes de recibir atención médica.

Es importante destacar que, aunque no se observa una correlación evidente en el gráfico, esto no excluye la posibilidad de que otras variables o factores no representados en este análisis puedan estar influyendo en el TDD y TDT.

Gráfico 22: Análisis de covarianza entre las variables tiempo de demora en el diagnóstico, tratamiento y edad de diagnóstico de casos notificados antes y después del ASPO.



Fuente: elaboración propia.

3.5 Aprendizaje Supervisado

Por medio del aprendizaje automático se busca la predicción del tiempo de demora en el diagnóstico de individuos con diagnóstico de tuberculosis, para ello, se aplica el modelo de regresión lineal multivariado[25].

Tabla 4. Modelo de regresión multivariado para el total de casos notificados de tuberculosis del 2019 a 2021 en Argentina.

```
==============================================================================
Dep. Variable:      TIEMPO_DEMORA_TTO   R-squared:                       0.009
Model:                            OLS   Adj. R-squared:                  0.008
Method:                 Least Squares   F-statistic:                     12.52
Date:                Tue, 12 Dec 2023   Prob (F-statistic):           5.31e-20
Time:                        16:29:19   Log-Likelihood:                -65402.
No. Observations:               12395   AIC:                         1.308e+05
Df Residuals:                   12385   BIC:                         1.309e+05
Df Model:                           9
Covariance Type:            nonrobust
=============================================================================================================
                                                coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------------------------
const                                       151.6785      2.329     65.129      0.000     147.114     156.244
CLASIFICACION_PANDEMIA_Antes                 74.4909      1.232     60.461      0.000      72.076      76.906
CLASIFICACION_PANDEMIA_Despues               77.1876      1.253     61.625      0.000      74.732      79.643
TIEMPO_DEMORA                                 0.0170      0.011      1.538      0.124      -0.005       0.039
CLASIF_INICIO_TRAT_Recaída                    7.4951      2.944      2.546      0.011       1.725      13.266
EDAD_DIAGNOSTICO                              0.0612      0.023      2.605      0.009       0.015       0.107
SEXO_M                                        2.7222      3.141      0.867      0.386      -3.434       8.878
SEXO_F                                       -0.1833      3.157     -0.058      0.954      -6.371       6.004
CLASIF_INICIO_TRAT_Nuevo                     -4.1122      1.629     -2.525      0.012      -7.305      -0.919
CLASIF_INICIO_TRAT_Pérdida del seguimiento recuperado    14.0136      3.359      4.172      0.000       7.430      20.597
CLASIFICACION_EXTRAPULMONAR_Diseminada       22.7289      5.157      4.408      0.000      12.621      32.837
==============================================================================
Omnibus:                     1249.843   Durbin-Watson:                   1.958
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1662.644
Skew:                           0.890   Prob(JB):                         0.00
Kurtosis:                       3.223   Cond. No.                     1.03e+16
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 3.14e-25. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.
```

Fuente: elaboración propia.

En la tabla anterior se aprecia que el coeficiente de determinación (R^2) es de 0.009. Este valor indica que aproximadamente el 0.9% de la variabilidad en el tiempo de demora en el tratamiento puede ser explicada por las variables predictoras incluidas en el modelo. No obstante, este valor es bajo, la prueba F arroja un p-valor significativamente pequeño ($p<0.05$), por lo cual sugiere que al menos una de las variables predictoras es estadísticamente significativa en relación con la variable de respuesta.

Cuando se analizan los coeficientes de las variables predictoras, la que representa la clasificación de la pandemia, como se ha mencionado a lo largo del presente trabajo, posee dos niveles (antes y después). Ambos

[25] Véase metodología.

evidencian diferencias significativas en lo que respecta al tiempo de demora en relación a un nivel de referencia no especificado. Además, el TDT posee un coeficiente positivo de 0.0170, ello implica que manteniendo todas las demás variables constantes, manifiesta un aumento esperado de 0.0170 unidades sobre el TDT por cada unidad de cambio en el tiempo de demora en el diagnóstico. Empero, este coeficiente no es estadísticamente significativo (p=0.124).

En otro orden, la variable de clasificación al inicio del tratamiento goza de varios niveles ('Recaída', 'Nuevo', 'Pérdida del seguimiento recuperado'), cada uno expone diferencias significativas respecto al tiempo de demora en relación al nivel de referencia no especificado. Asimismo, las variables de sexo (masculino y femenino) no son estadísticamente significativas, es decir que no exhibe un efecto significativo en el tiempo de demora. Por último, la variable de clasificación extrapulmonar revela un coeficiente positivo de 22.7289, esto sugiere que preservando las demás variables constantes, expone un aumento esperado de 22.7289 unidades en el TDT cuando la clasificación es 'Diseminada'.

Por todo lo anterior, se considera que si bien el modelo expone algunas relaciones estadísticamente significativas, el pequeño R^2 sugiere que las variables incluidas explican una pequeña proporción de la variabilidad en el tiempo de demora en el tratamiento.

Tabla 5. Modelo de regresión multivariado para el total de casos notificados de tuberculosis del 2019 a 2021 en Argentina.

OLS Regression Results

Dep. Variable:	TIEMPO_DEMORA_TTO	R-squared:	0.001
Model:	OLS	Adj. R-squared:	0.001
Method:	Least Squares	F-statistic:	5.997
Date:	Mon, 11 Dec 2023	Prob (F-statistic):	0.00249
Time:	14:10:15	Log-Likelihood:	-65050.
No. Observations:	12319	AIC:	1.301e+05
Df Residuals:	12316	BIC:	1.301e+05
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	225.9477	0.966	233.841	0.000	224.054	227.842
EDAD_DIAGNOSTICO	0.0679	0.024	2.883	0.004	0.022	0.114
TIEMPO_DEMORA	0.0199	0.011	1.791	0.073	-0.002	0.042

Omnibus:	1254.165	Durbin-Watson:	1.954
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1673.382
Skew:	0.896	Prob(JB):	0.00
Kurtosis:	3.216	Cond. No.	117.

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Fuente: elaboración propia.

De manera análoga, el modelo de regresión lineal múltiple se aplicó para analizar la relación entre la variable dependiente, el tiempo de demora en el tratamiento, y dos variables independientes: la edad en el momento del diagnóstico y el tiempo de demora inicial. Este arrojó un bajo poder explicativo (R^2=0.001), que indica una pequeña variabilidad en el tiempo de demora en el tratamiento respecto a las variables independientes.

Por su parte, los coeficientes de regresión señalan que si se mantiene constante el tiempo de demora inicial, el aumento de una unidad en la edad de diagnóstico se asocia a un incremento de 0.0679 en el TDT. No obstante, un aumento de una unidad en el tiempo de demora inicial se asocia a un aumento de 0.0199 en el TDT. Es decir que ambos coeficientes fueron estadísticamente significativos, con *p-valores* de 0.004 y 0.073, respectivamente.

A pesar de que existe una relación estadísticamente significativa entre las variables independientes y la

demora en el tratamiento, esta es muy débil y explica solo una pequeña parte de la variabilidad observada en la demora del tratamiento. Por lo tanto, es posible que otros factores no incluidos en el modelo también influyan.

Cabe destacar que estos valores no mejoran al analizar las relaciones por grupo según momentos de notificación del caso. Pues, en ambos modelos, los coeficientes de regresión para las variables independientes edad de diagnóstico no son estadísticamente significativos, con *p-valores* de 0.189 y 0.189, respectivamente. Mientras que para la variable independiente tiempo de demora en el diagnóstico sin son estadísticamente significativos, con *p-valores* de 0.032, para los casos antes y después del ASPO. Sin embargo, al igual que para el total de los casos es importante destacar que, aunque los p-valores son estadísticamente significativos, la magnitud de los coeficientes es pequeña, lo que indica que las variables incluidas tienen un impacto limitado en el tiempo de demora en el tratamiento.

Tabla 6. Modelo de regresión multivariado para los casos notificados de tuberculosis antes del ASPO en Argentina.

OLS Regression Results

Dep. Variable:	TIEMPO_DEMORA_TTO	R-squared:	0.001
Model:	OLS	Adj. R-squared:	0.001
Method:	Least Squares	F-statistic:	3.254
Date:	Mon, 11 Dec 2023	Prob (F-statistic):	0.0387
Time:	14:12:07	Log-Likelihood:	-26710.
No. Observations:	5084	AIC:	5.343e+04
Df Residuals:	5081	BIC:	5.345e+04
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	224.9016	1.447	155.436	0.000	222.065	227.738
EDAD_DIAGNOSTICO	0.0457	0.035	1.313	0.189	-0.023	0.114
TIEMPO_DEMORA	0.0361	0.017	2.142	0.032	0.003	0.069

Omnibus:	486.786	Durbin-Watson:	2.006
Prob(Omnibus):	0.000	Jarque-Bera (JB):	638.523
Skew:	0.866	Prob(JB):	2.22e-139
Kurtosis:	3.130	Cond. No.	115.

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Fuente: Elaboración propia.

Tabla 7. Modelo de regresión multivariado para los casos notificados de tuberculosis después del ASPO en Argentina.

OLS Regression Results

Dep. Variable:	TIEMPO_DEMORA_TTO	R-squared:	0.001
Model:	OLS	Adj. R-squared:	0.001
Method:	Least Squares	F-statistic:	3.254
Date:	Mon, 11 Dec 2023	Prob (F-statistic):	0.0387
Time:	14:13:05	Log-Likelihood:	-26710.
No. Observations:	5084	AIC:	5.343e+04
Df Residuals:	5081	BIC:	5.345e+04
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	224.9016	1.447	155.436	0.000	222.065	227.738
EDAD_DIAGNOSTICO	0.0457	0.035	1.313	0.189	-0.023	0.114
TIEMPO_DEMORA	0.0361	0.017	2.142	0.032	0.003	0.069

Omnibus:	486.786	Durbin-Watson:	2.006
Prob(Omnibus):	0.000	Jarque-Bera (JB):	638.523
Skew:	0.866	Prob(JB):	2.22e-139
Kurtosis:	3.130	Cond. No.	115.

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Fuente: elaboración propia.

4. Conclusiones

Teniendo en cuenta la presencia de las Ciencias computacionales en los diversos ámbitos de la vida, en este trabajo se ha buscado relacionar el aprendizaje automático con una enfermedad infecciosa, como es la tuberculosis. Para ello, con los reportes obtenidos del Sistema Nacional de Vigilancia de la Salud de la República Argentina, se evaluaron las notificaciones de casos de TB en el contexto previo y durante la pandemia de COVID-19. El objetivo principal radicó en determinar las relaciones entre tiempos de diagnóstico y de tratamiento, edades de pacientes y ubicación de estos.

La notificación de casos por provincia evidencia un aumento después de la declaración del ASPO, siendo Buenos Aires, la Ciudad Autónoma de Buenos Aires, Santa Fe, Salta y Córdoba las que presentaron mayor incidencia. Asimismo, al considerar la distribución etaria y de sexo, se observó una notable similitud antes y después del ASPO, con un leve predominio de casos en el grupo de entre 15 y 40 años.

En lo que respecta a la temporalidad de las notificaciones se constató una disminución de estas a principios del año 2020, que coincidiría con el principio del aislamiento obligatorio en Argentina, seguida de un notable aumento en el primer semestre de 2021. Esta tendencia se mantiene en los casos tratados con éxito, aunque la disminución no es tan marcada. Al enfocarse en grupos de edad específicos, se destaca una variación en el grupo de 0 a 18 años, con un descenso, seguido de un aumento en el primer semestre de 2021.

El análisis de los tiempos de demora en el diagnóstico y tratamiento presenta distribuciones asimétricas hacia la derecha, sugiriendo una presencia de valores atípicos. Tras el ASPO, se observa un leve aumento en los tiempos de demora, tanto en diagnóstico como en tratamiento. Sin embargo, al desglosar por provincia, se revelan diferencias notables. Algunas provincias mantienen estabilidad, mientras que otras experimentan aumentos o reducciones notables en los tiempos de demora, evidenciando disparidades regionales.

Por otra parte, se pudo observar que la georreferenciación de los casos no mostró diferencias sustanciales en la distribución geográfica de la tuberculosis antes y después de la pandemia. Esto indica que la enfermedad mantuvo una distribución comparativamente uniforme en todo el país durante ambos períodos.

El análisis reveló que el valor de R-cuadrado para la relación entre las variables independientes y la demora en el tratamiento es extremadamente bajo, alrededor del 0.1%. Esto indica que solo una pequeña fracción de la variabilidad en la demora del tratamiento se puede atribuir a las variables independientes incluidas en el modelo.

En particular, no se observó una relación significativa entre la edad de los pacientes y la demora en el tratamiento, lo que sugiere que las variables analizadas tienen una relación muy débil con esta variable.

La prueba de *log-rank* respalda cambios significativos en los tiempos de demora asociados con la clasificación de la pandemia, sugiriendo la necesidad de adaptar estrategias de salud pública para abordar estas variaciones. En resumen, estos hallazgos subrayan la compleja interacción entre la pandemia de COVID-19 y el diagnóstico y tratamiento de la tuberculosis en Argentina, destacando la importancia de considerar contextos demográficos regionales y específicos en la eficiencia dispar en la atención de la salud de los casos de tuberculosis.

Referencias bibliográficas

Banco de recursos de comunicación del Ministerio de Salud de la Nación (2023). Boletín n° 6: Tuberculosis y lepra en la Argentina. https://bancos.salud.gob.ar/recurso/boletin-ndeg-6-tuberculosis-y-lepra-en-la-argentina

Cuello-Rüttler, L., & Gudiño, M. E. (2017). Los sistemas de información geográfica (SIG) como herramienta para el ordenamiento territorial y la salud pública. https://bdigital.uncu.edu.ar/objetos_digitales/10881/01e3cuello-gudio.pdf

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer.

Hobsbawm, E. J. (2018). *Historia del siglo XX: 1914-1991*. Crítica.

Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. Journal of the American Statistical Association, 53(282), 457–481. https://doi.org/10.1080/01621459.1958.10501452

Martínez Sesmero, J. M. (2015). "Big Data"; aplicación y utilidad para el sistema sanitario. *Farmacia hospitalaria : organo oficial de expresion cientifica de la Sociedad Espanola de Farmacia Hospitalaria*, *39*(2), 69-70. https://doi.org/10.7399/fh.2015.39.2.8835

Monterde i Bort, H., & Perea Lara, M. (1991). Capítulo 1: Introducción y principios básicos. *El enfoque del análisis exploratorio de datos: (Y su aplicación al campo de la psicología)* (pp. 9-39).

Organization, P. A. H., & Salud, O. P. de la. (2002). *Sistemas de información geográfica en salud: Conceptos básicos*. https://iris.paho.org/handle/10665.2/40000

Spiegel, M. R. y Stephens, L.J. (2009). *Estadística*. Serie Schaum. Mc Graw Hill.

Wickham, H. y Grolemund, G. (2016). *R for Data Science*. https://r4ds.hadley.nz/

World Health Organization (2013). *Definiciones y marco de trabajo para la notificación de tuberculosis-Revisión 2013*. https://www.who.int/es/publications/i/item/978924150534

World Health Organization (2023). Global Tuberculosis Report 2023. https://www.who.int/publications-detail-redirect/9789240083851

42

Violencia por motivos de género: Medición y predicción de riesgo en los casos abordados por la Línea 144 (2020-2022)

De Lucia Julia Gaztañaga

Lucia Julia Gaztañaga

Violencia por motivos de género: Medición y predicción de riesgo en los casos abordados por la Línea 144 (2020-2022)

Gender-based violence: risk measurements and statistical predictions in communications assisted by 144 National Telephone Line (2020-2022)

Lucia Julia Gaztañaga (UBA/UNAB)
Buenos Aires, Argentina
lugaztanaga@gmail.com

Resumen

El abordaje de los datos de violencias por motivos de género presenta importantes desafíos para su análisis, y requiere de metodologías específicas para obtener estrategias de diagnóstico e intervención. En ese marco, el trabajo presenta una propuesta que abarca la construcción de indicadores a partir del análisis de los datos públicos 2020-2022 de la Línea 144, destinada a la atención y abordaje de consultas de mujeres y personas LGBTI+ que atraviesan diversas situaciones de violencias por motivos de género. El objetivo es construir nuevos indicadores que puedan realizar un aporte al análisis de los casos de violencias por motivos de género que aborda periódicamente la Línea.

Para ello, se utilizan las definiciones de **nivel de riesgo** obtenidas gracias a la funcionalidad de **medición de riesgo** que ofrece el Sistema Integrado de Casos de Violencias por Motivos de Género (SICVG). Así, se pretende, en primer lugar, **calcular y establecer** el indicador correspondiente al **nivel de riesgo** de cada una de las *intervenciones* realizadas por la Línea 144 entre el 2020 y el 2022.

Una vez obtenido dicho indicador, se presenta la propuesta de un modelo de predicción del nivel de riesgo de los casos, con el objeto de **predecir el nivel de riesgo** de un caso de violencia por motivos de género, si se verifica la existencia de determinadas características dentro del caso en cuestión.

Palabras clave

Violencia de género – nivel de riesgo – Línea 144 – SICVG-predictor de riesgo

1. Introducción

Creada en el año 2013 para cumplir los objetivos establecidos por el artículo 9 de la Ley 26.485 (ley de Protección Integral para Prevenir, Sancionar y Erradicar la Violencia Contra las Mujeres en los ámbitos en que desarrollen sus relaciones interpersonales), la Línea 144 brinda atención, contención y asesoramiento a mujeres y LGBTI+ en situación de violencia de género, de manera gratuita, las 24 hs., los 365 días, a través de un llamado al 144, o por los otros canales habilitados tales como WhatsApp, correo electrónico e incluso descargando la app específica. También dispone de servicio de comunicación por videollamadas para personas Sordas e Hipoacúsicas.

La Línea 144 bajo gestión nacional se encontraba bajo la órbita del Ministerio de las Mujeres, Géneros y Diversidad de la Nación[1] desde el 10 diciembre de 2019 hasta el mismo periodo de 2023. De acuerdo al documento publicado por dicho ministerio en septiembre de 2023 a propósito del 10mo aniversario de la implementación de la Línea, *"Línea 144. 10 años. Una década del Dispositivo Federal de Atención de las Violencias*

[1] A la fecha del presente artículo, a partir del 10 de diciembre de 2023, de acuerdo al DNU 8/2023 el Ministerio de Mujeres, Géneros y Diversidad pierde su status como tal y todos sus compromisos, políticas y programas vigentes, incluyendo la Línea 144, quedan bajo la órbita del nuevo Ministerio de Capital Humano. Para más detalle: https://www.boletinoficial.gob.ar/detalleAviso/primera/300727/20231211

de Género" (https://www.argentina.gob.ar/sites/default/files/2022/05/linea144aniversario-web-v4.pdf), esta atiende un promedio de 125.000 comunicaciones al año, provenientes de todo el país.

- **Abordaje y seguimiento de casos de violencias por motivos de género en la Línea 144: el papel de los indicadores de riesgo**

En el mencionado documento se detalla que uno de los pilares fundamentales en el proceso de abordaje de los casos de violencias por motivos de género que ingresan a la Línea es la evaluación de i*ndicadores de riesgo* que, al brindar información sobre la situación de riesgo de vida en el que se encuentra la persona en situación de violencia de género permiten *"establecer el nivel de urgencia en la intervención, acompañamiento y/o derivación a otras instituciones estatales para evitar una situación de gravedad o violencia extrema"*[2] (MMGyD;2023; pp.25).

Entre los aspectos que se indagan para poder caracterizar y contextualizar la situación de violencia a abordar y determinar el correspondiente *nivel de riesgo*, el documento informa que se indaga sobre la situación socioeconómica, *"la salud, el lugar donde vive y datos sobre la persona agresora"* además de aspectos relacionados a las características propias de las violencias que atraviesa la persona asistida.

Algunos de estos datos se encuentran disponibles al público[3]: el 98,0% de las personas asistidas por situaciones de violencias por motivos de género asistidas por la Línea 144 desde el 2013 son mujeres. El 81,2% de las personas agresoras son varones, y en el 39,0% de los casos quien ejerce conductas violentas es la ex pareja. En el 42,8% de los casos, en cambio, se trata de la pareja actual. En relación con las características de la situación de violencia, la gran mayoría (95,3%) de los casos asistidos corresponde a situaciones de violencia doméstica. En una proporción similar, las personas en situación de violencia manifestaron haber atravesado violencia psicológica, mientras que un 70,4% refieren haber sufrido violencia física. Un 10,1% manifestó haber atravesado violencia sexual.

En otros informes publicados, [4]también se hace mención de otros datos como la edad y nacionalidad de las personas en situación de violencia, así como la proporción de personas que se encontraban cursando un embarazo o padecían algún tipo de discapacidad. También se mencionan otros aspectos sobre la situación de violencia, como la presencia, tenencia o uso de armas.

No obstante esta caracterización general de las intervenciones realizadas por la Línea 144, es relevante destacar que otro punto nodal para las evaluaciones de riesgo realizadas por lxs profesionalxs de la Línea son las particularidades y especificidades de cada caso: se parte de la base que *"cada comunicación tiene características únicas"* (MMGyD;2023; pp.25), con lo cual las estrategias de abordaje se corresponden con las particularidades de cada situación de violencia y de la persona que la atraviesa.

- **¿Cómo medir el riesgo? Funcionalidades para la medición de riesgo en el Sistema Integrado de Casos de Violencias por Motivos de Género (SICVG).**

De acuerdo al mencionado documento en el marco del 10mo aniversario del dispositivo de atención correspondiente a la Línea 144, se detalla que actualmente los registros históricos de las comunicaciones ingresadas forman parte del Sistema Integrado de Casos de Violencias por Motivos de Género (SICVG), así como el registro y sistematización de las nuevas comunicaciones ingresadas a partir del segundo semestre de 2023.

Creado por el Ministerio de las Mujeres, Géneros y Diversidad en el marco del Plan Nacional de Acción contra las Violencias por Motivos de Género 2020-2022, el SICVG consiste en una herramienta destinada a la sistematización de la información sobre diferentes tipos de registros administrativos vinculados a situaciones de violencias de género de todo el país, posibilitando no solo el registro, gestión, seguimiento e

---

[2] Las "violencias extremas" refieren a las muertes violentas de mujeres y todas aquellas personas que padecieron violencias letales causa de su identidad de género y/u orientación sexual (MMGYD, MINJUS, MINSEG; 2020).

[3] Ministerio de las Mujeres, Géneros y Diversidad. (2023). "Línea 144. 10 años. Una década del Dispositivo Federal de Atención de las Violencias de Género". Disponible en https://www.argentina.gob.ar/sites/default/files/2022/05/linea144aniversario-web-v4.pdf.

[4] Para más detalles, se encuentran accesibles al público las infografías anuales de datos correspondientes a la Linea 144 en https://www.argentina.gob.ar/generos/linea-144/informacion-estadistica

intervención de los casos asistidos, sino que también posibilita el análisis estadístico y la **medición de riesgo automática** de casa caso.

Esta funcionalidad opera partiendo desde un modelo de ponderación elaborado en conjunto con un set de indicadores diseñados específicamente y que resultan determinantes para la clasificación inmediata del caso en el que se interviene según el **nivel de riesgo** en el que se encuentra una persona para la pronta intervención, Esta funcionalidad permite además sistematizar las mediciones de riesgo, para poder avanzar, en un futuro, hacia modelos de predicción. Es importante además destacar que, en cuanto al funcionamiento de esta medición, *cuanto más detalle se conozca de la situación de violencia, mejor será el desempeño de la medición de riesgo* (MMGyD; 2023; pp.2)

Según la documentación disponible, los niveles de riesgo definidos por sistema son: *altísimo* (riesgo de vida inminente, requiere intervención urgente), *alto* (existe riesgo de vida, se requiere intervención), *mediano* (no se detecta riesgo de vida, pero existe requerimiento de acompañamiento y seguimiento) y *bajo (no se detecta riesgo y solo requiere acciones tales como asesoramientos simples, brindado de información, etc.).*

2. Metodología

- **Objetivos de trabajo:**

- Clasificar las intervenciones realizadas por la Línea 144 en casos de violencias por motivos de género entre 2020-2022 según los niveles de riesgo establecidos por el Sistema Integrado de Casos de Violencias por Motivos de Género (SICVG).
- Elaborar un modelo de predicción de riesgo, teniendo en cuenta los datos disponibles, que permita determinar el nivel de riesgo en los eventuales casos que intervenga la Línea 144.

Para dar cuenta de los objetivos planteados, se trabajarán con los data sets públicos de la Línea 144, correspondientes a los registros de comunicaciones ingresadas a la misma. Es posible acceder a estos registros, ya que los data sets correspondientes se encuentran accesibles al público y alojados tanto en la web institucional https://www.argentina.gob.ar/generos/linea-144/informacion-estadistica como en el Portal Nacional de Datos Abiertos desde https://datos.gob.ar/dataset/generos-base-datos-linea-144.

De acuerdo al detalle publicado en datos.gob.ar, la información publicada [5]de la Línea 144 corresponde a comunicaciones sobre las cuales, según las particularidades del caso reportado, se han realizado acciones de intervención, abordaje y seguimiento personalizados. Este tipo de registros, denominados *intervenciones*, corresponden no solo a la sede a cargo del MMGyD nacional sino también tanto a la sede gestionada por el gobierno de la Provincia de Buenos Aires, como a la perteneciente a la Ciudad Autónoma de Buenos Aires.

Los datos se encuentran organizados en distintos data sets según el año de registro. Por el momento solo se encuentran disponibles en este sitio los datos a partir del año 2020.

El proceso de trabajo fue realizado en tres etapas. Las mismas se realizaron mediante métodos propios de la ciencia de datos, ya que los mismos implican una serie de técnicas necesarias para el tratamiento y manipulación de información masiva desde un enfoque estadístico e informático (Hernandez Leal, Duque Mendezy, Moreno-Cadavid; 2017; pp.2), Se utilizaron diferentes métodos aplicados a través del lenguaje de programación Python.

El desarrollo de diversos métodos de ciencias de datos aplicado a las ciencias sociales aparece como corolario de la expansión del *big data*, entendiendo a éste desde "su dimensión tecnológica, señalando la disponibilidad de grandes volúmenes de datos en diversos formatos, la proliferación de nuevas técnicas para su procesamiento, y el desarrollo de una infraestructura de sistemas capaz de soportar todo esto" (Becerra; 2018, pp.141). La utilización de estos métodos obedece no solo a los propósitos de este trabajo en particular sino también a cuestiones estructurales de las fuentes de datos a utilizar: al analizar programas y políticas públicas, por ejemplo, la información suele encontrarse dispersa y de manera fragmentada, o bien la misma presenta inconsistencias que dificultan su interpretación. La utilización de técnicas novedosas que permitan

[5] De acuerdo a la definición publicada en datos.gob.ar sobre los registros de la Linea 144, estos corresponden a "*a aquellas comunicaciones recibidas por la misma a las que se denomina* intervenciones*, en donde las personas que se comunican acceden a dejar sus datos* para un adecuado abordaje y seguimiento".

acceder y consistir información de manera más ágil y automatizada abriría el paso para maximizar y optimizar los resultados en el uso de los datos disponibles, sin que necesariamente se trate de grandes volúmenes de datos.

Así, la primera etapa consistió en implementar el proceso de *data wrangling* (exploración, transformación y limpieza de datos, entre otras tareas), a través del cual finalmente se obtuvo un único data set con el total de intervenciones realizadas por la Línea 144 durante el periodo analizado.

En paralelo, se realiza una revisión de los documentos de trabajo disponibles acerca de la Línea 144 y el Sistema Integrado de Casos de Violencias por Motivos de Género (SICVG), con el objetivo de indagar acerca de los datos disponibles, estructura y metodología de registros, definición de sus variables y categorías. La información aquí recabada resulta de vital importancia para orientar los procesos de transformación y posterior análisis de datos en pos de los objetivos propuestos.

Finalizada esta etapa, al obtener un data set con el total de intervenciones de la Línea 144 del periodo seleccionado, con todas las transformaciones necesarias, se procede a la selección de aquellas variables y categorías del data set que conforman el set de *indicadores de riesgo* que van a contribuir a la medición del mismo, conforme a las categorías de nivel de riesgo establecidas en el SICVG.

Una vez seleccionados estos, se le asigna una puntuación específica al azar a cada uno, cuya sumatoria conformaría el *índice de valor riesgo* que, de acuerdo al valor obtenido, permite la clasificación en alguno de los cuatro niveles de riesgo.

Al lograr esta clasificación, se procede al análisis de resultados y, a modo exploratorio, se avanza en la elaboración de un *modelo de predicción de riesgo* en base a la información disponible. Previo a este paso se realizan algunos ajustes en el data set para adecuar los datos a la aplicación de los métodos seleccionados.

Así, teniendo en cuenta la estructura de datos con la que se cuenta y el tipo de información que se desea predecir, se exploran dos modelos de clasificación mediante distintas técnicas: el primero mediante la denominada "árbol de decisión", mientras que en la segunda se utiliza el algoritmo del tipo "*random forest*" (bosque aleatorio).

3. Análisis y resultados

- **Etapa 1: Transformación de datos. Proceso de clasificación según nivel de riesgo.**

Una vez obtenidos los datos seleccionados desde su web pública, para el análisis en cuestión, se procede trabajar sobre el contenido de los mismos.

Según las definiciones obtenidas del sitio web, la información con la que se cuenta es la siguiente:

Tabla 1. Línea 144. Listado y definición de variables. Datos públicos 2020-2022.

fecha	Fecha del ingreso de la consulta
prov_residencia_persona_en_situacion_violencia	Provincia en donde la persona en situación de violencia declara estar residiendo al momento de la consulta
genero_persona_en_situacion_de_violencia	Identidad de género autopercibida de la persona en situación de violencia
edad_persona_en_situacion_de_violencia	Edad declarada de la persona en situación de violencia al momento en que se ingresó la consulta
pais_nacimiento_persona_en_situacion_de_violencia	País de nacimiento de la persona en situación de violencia

tipo_de_violencia (física, psicológica, sexual, económica y patrimonial, simbólica)	Tipo de violencia, contemplado por la ley 26.485, que motiva la consulta
modalidad_de_violencia (domestica, institucional, laboral, contra la libertad reproductiva, obstétrica, mediática)	Modalidad de violencia, contempladas por la ley 26.485, que motivan la consulta bajo la cual se desarrollaron
vinculo_con_la_persona_agresora	Vínculo que tiene o tenía la persona en situación de violencia con quien ejerce la agresión en el momento en que sucedieron las situaciones de violencia declaradas
genero_de_la_persona_agresora	Identidad de género de la persona agresora

Fuente: datos.gob.ar

Aquí se ve que existen datos referidos a las *características de la persona en situación de violencia* (provincia de residencia, género, edad, país de nacimiento**),** *características de la situación de violencia* (tipo y modalidad de violencia, vínculo con la persona agresora) y sobre la *persona agresora,* se cuenta con el dato del género. Además, se cuenta con el dato de la fecha de registro.

Tomando en cuenta la estructura y definición de variables del SICVG (disponible en https://www.argentina.gob.ar/sites/default/files/2021/12/definicion_de_variables_y_categorias-modulo_federal_.pdf) las variables de la base de Intervenciones de la Línea 144 coinciden en su definición conceptual y en algunas de sus categorías, si bien esta última contiene muchas menos que dicho sistema. Además, cabe recordar que muchos de los sets de variables que se registran en el SICVG fueron específicamente diseñadas a propósito de la medición de riesgo. No obstante, desde los registros disponibles de la Linea se podrían reconstruir algunos datos tal cual la estructura de dicho sistema.

Como tal, entonces, es pertinente definir, del set de variables y categorías disponibles, cuales resultarían *indicadores de riesgo*. En principio, se consideran como tal a aquellos datos cuya presencia impliquen una mayor *amenaza* a la integridad de la persona y su vida y /o a una mayor *vulnerabilidad* de la persona frente a la situación de violencia que atraviesa, que puede agravar la misma. En definitiva, se entiende que la presencia de estos indicadores incrementaría los niveles de riesgo en cuanto a la integridad física y psíquica de la persona. Estos son:

- La identidad de género de la persona en situación de violencia.
- Características de *interseccionalidad*[6] de las personas en situación de violencia: persona migrante, persona mayor de 60 años, niñxs y adolescentes.
- Tipos de violencia, especialmente la violencia física y la violencia sexual.
- Modalidad de violencia[7]

Se excluyen del set de indicadores a los datos correspondientes a la fecha y provincia de residencia ya que no tienen relación directa con la situación de violencia. También se excluyen los datos correspondientes al vínculo con la persona agresora y al género de la persona agresora, ya que si bien son datos fundamentales para caracterizar y contextualizar las situaciones de violencia por motivos de género (de hecho son datos determinantes para clasificar las situaciones de violencia como tal), teniendo en cuenta este análisis en particular y los datos disponibles, no se consideran factores de peso en cuanto a la medición de amenaza y vulnerabilidad, o bien no inciden en cuestiones de peligro de vida inmediato.

---

[6] *De acuerdo a la definición de Kimberly Creenshaw , el término "interseccionalidad" refiere a la interacción entre "categorías de diferenciación" (como el género o la pertenencia a determinada cultura), en la vida de las personas, en sus prácticas sociales, en las instituciones e ideologías culturales (Creenshaw;1998). En el marco de una situación de violencia por motivos de género, estas "categorías de diferenciación" pueden posicionar a la persona que la atraviesa en un escenario de mayor vulnerabilidad.*

[7] *Refiere a los ámbitos en donde se manifiestan los distintos tipos de violencias por motivos de género. Como se mencionó al inicio del documento, entre quienes se comunicaron a la Línea 144 desde el año 2013 predominan ampliamente las situaciones de violencia sucedidas dentro del ámbito doméstico, en su gran mayoría siendo ejercidas por parejas o ex parejas.*

Como hemos definido que ciertas características de la persona en situación de violencia pueden colocarla en mayor situación de vulnerabilidad frente a la misma, en línea con las definiciones de SICVG se procede a transformar algunas variables y categorías de los data sets de la Línea para equipararlos:

Tabla 2. Línea 144. Transformaciones a realizar sobre variables para la medición de riesgo.

SICVG	Línea 144	Línea 144_categorias
Persona migrante	País de Nacimiento de la persona en situación de violencia	Cualquier país de nacimiento distinto de Argentina
Persona LGBTTI+	Género de la persona en situación de violencia	Identidades de género distintas a "Mujer"
NNyA*	Edad de la persona en situación de violencia	Edad menor a 18 años
Persona Mayor	Edad de la persona en situación de violencia	Edad mayor a 65

Fuente: Elaboración propia en base a datos de la Línea 144-MMGyD

Una vez construidas las nuevas variables, se avanza en los métodos para la clasificación de las intervenciones de la Línea 144 según su nivel de riesgo, mediante la medición del peso de los indicadores de riesgo previamente definidos.

Tal cual se detalló anteriormente, la estructura de medición de riesgo del SICVG se realiza a través del análisis de determinadas variables del sistema, cuyo registro le asigna un puntaje específico que permite el cálculo automático del nivel de riesgo que representa el caso asistido. Entonces, para el cálculo de riesgo sobre los datos de la Línea 144, se decide también asignar un puntaje específico por cada indicador correspondiente, cuya sumatoria, según el rango de puntuación preestablecido, posibilitaría arrojar el nivel de riesgo equivalente. A este "puntaje" asignado, lo denominamos *"valor riesgo"*.

El puntaje asignado por cada categoría se realiza de manera aleatoria, respetando el siguiente criterio:

- 5 puntos en aquellos ítems que impliquen un mayor daño a la integridad física (violencia física, violencia sexual), o signifiquen una situación de extrema vulnerabilidad según las características de las personas en situación de violencia (Niñxs y Adolescentes), independientemente de la presencia de cualquier otro factor.
- 2 puntos si la persona en situación de violencia registra alguna característica *interseccional* que pueda posicionarla con una vulnerabilidad aún mayor (personas mayores, personas migrantes, personas LGBTI+)
- 1 punto si se registra la presencia de alguna otra característica que pueda tener peso en cuanto al nivel de daño producido.

A continuación, se resume en el siguiente cuadro aquellas que se tendrán en cuenta para la medición de riesgo, así como las categorías que representan "peso" en la medición del mismo y por último el *valor riesgo* asignado:

Tabla 3. Línea 144. Valor de riesgo asignado por cada categoría (indicador de riesgo).

	Variables	Categorías	Valor riesgo
0	Género	Mujer	1
1	Género	Persona LGBTI+ (Mujer Trans, Varón Trans, Trav...	2
2	País de Nacimiento	Otros paises	2
3	Grupo Edad	NNyA (menores de 18 años)	5
4	Grupo Edad	Persona Mayor (mayores de 60 años)	2
5	Tipo de violencia	Violencia física	5
6	Tipo de violencia	Violencia psicológica	1
7	Tipo de violencia	Violencia económica y patrimonial	1
8	Tipo de Violencia	Violencia sexual	5
9	Modalidad	Violencia doméstica	1
10	Modalidad	Violencia laboral	1
11	Modalidad	Violencia Institucional	1
12	Modalidad	Violencia contra la libertad reproductiva	1
13	Modalidad	Violencia obstétrica	1

Fuente: Elaboración propia en base a datos de la Línea 144-MMGyD

La sumatoria de cada "valor riesgo" dentro de cada caso arrojará un "valor riesgo total" el cual, de acuerdo a su ubicación en un rango pre establecido, permite clasificar cada caso según su nivel de riesgo. Los rangos se establecen a partir del supuesto del máximo índice de valor riesgo posible, de acuerdo a los datos con los que se cuenta, evaluando las combinatorias posibles de los indicadores de riesgo seleccionados, de acuerdo a los valores previamente asignados.

Tabla 4. Línea 144. Clasificación del nivel de riesgo según el rango obtenido en el índice de valor riesgo.

	Niveles	Rango de valores
0	Altísimo	>10
1	Alto	7 a 9
2	Medio	4 a 6
3	Bajo	0 a 3

Fuente: Elaboración propia en base a datos de la Línea 144-MMGyD

A través de los procedimientos detallados se contabilizan un total de 79.565 intervenciones en casos de violencias por motivos de género realizadas por la Línea 144 entre los años 2020 y 2022 inclusive. Este número es consistente con lo reportado en otros informes disponibles. [8]

[8] *Datos de intervenciones realizadas por la Linea 144, por año, disponibles en* https://www.argentina.gob.ar/generos/linea-144/informacion-estadistica

Cada uno de estos registros, de acuerdo a la medición de riesgo detallada en el apartado anterior, se clasifican dentro de los cuatro niveles de riesgo de acuerdo a las definiciones del SICVG.

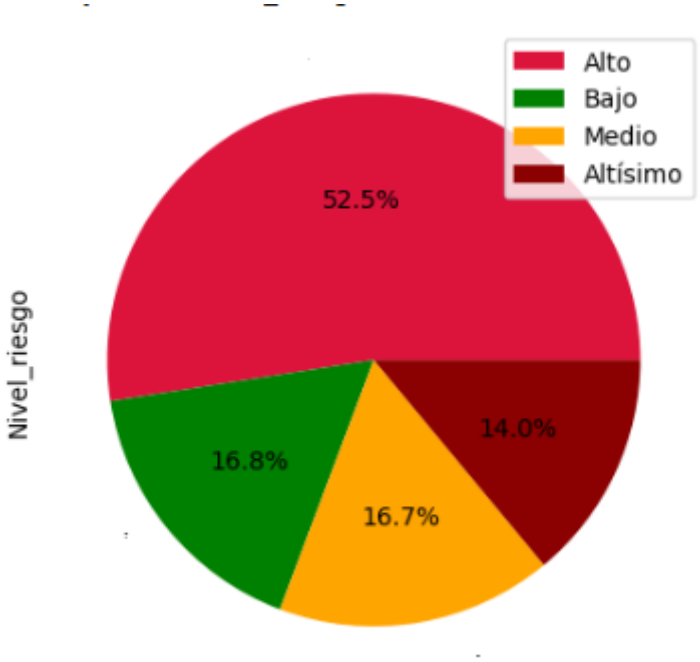
Tabla 4. Línea 144. Intervenciones realizadas por nivel de riesgo. Total país. 2020-2022.

Nivel de riesgo	N
Altísimo	11125
Alto	41751
Medio	13321
Bajo	13368
Total	**79565**

Fuente: elaboración propia en base a datos de la Línea 144.

Así, conforme a los criterios establecidos para la medición del nivel de riesgo de los casos de violencia por motivos de género abordados por la Línea 144, más de la mitad de los casos son de alto nivel de riesgo, mientras que un 14% correspondieron a casos con el máximo nivel de riesgo, con lo cual se supone debieron ser abordajes de casos que requerían actuación inmediata.

Gráfico 1. Línea 144. Intervenciones realizadas por nivel de riesgo. Total país. 2020-2022. En porcentaje.



Fuente: elaboración propia en base a datos de la Línea 144.

Etapa 2: Modelo de predicción de nivel de riesgo para los casos asistidos por la Línea 144

La clasificación de cada una de las intervenciones en casos de violencias por motivos de género realizadas por la Línea 144 durante el periodo de análisis resultaba fundamental para avanzar en la siguiente etapa: la elaboración de un *modelo predictor de riesgo*.

Primero, se procede a realizar otros ajustes y transformaciones adicionales en los datos para evitar inconvenientes al momento de confeccionar los modelos predictores. Luego, se avanza en el diseño de un

modelo estadístico de clasificación del tipo *"árbol de decisión", por un lado,* y otro modelo *clasificatorio mediante la técnica de* "bosque aleatorio"(random forest).

En ambos casos se utiliza el nivel de riesgo como variable objetivo. En cuanto a las variables predictoras, las mismas se seleccionaron en base a aquellos datos disponibles que fueron excluidos del set de indicadores *de riesgo*, al *no* significar un factor de peso respecto al agravamiento o no de una situación de violencias por motivos de género (al menos en el marco del presente análisis). No obstante, estas variables en cuestión sí representan datos fundamentales para caracterizar y contextualizar estas situaciones, y su influencia en cada situación particular es diferencial respecto a cómo se combina con otros factores.[9]

El objetivo es explorar que tan factible es predecir el nivel de riesgo tomando en cuenta aquellos factores que no se consideraron para la medición de riesgo (la edad de la persona en situación de violencia, el vínculo con la persona agresora, el género de la misma, el tipo de violencia simbólica[10] y la modalidad mediática). Luego de elaborar los modelos y verificar sus resultados, se analiza la performance entre ambos modelos, en clave comparativa.

El primer modelo de predicción del nivel de riesgo parece "sesgado" por la amplia proporción de casos de nivel alto que caracteriza al total de los casos analizados: vemos que en la muestra de prueba hay una proporción importante de casos con predicciones correctas, respecto a este nivel de riesgo, pero la performance disminuye significativamente en cuanto a la relación entre los valores observados y esperados para el resto de los niveles.

Tabla 5. Línea 144. Modelo de predicción del nivel de riesgo. Niveles de riesgo observados por nivel de riesgo esperado. Total país. 2020-2022.

Predicciones / Actual	Alto	Altísimo	Bajo	Medio
Alto	9716	133	326	197
Altísimo	1928	526	111	146
Bajo	2646	13	591	117
Medio	2587	128	299	362

Nota: los datos corresponden a la muestra de prueba extraída para aplicar el modelo de clasificación elaborado.

Fuente: elaboración propia en base a datos de la Línea 144.

A continuación, se implementa el segundo modelo de predicción. Se explora esta opción con la expectativa en que supere los resultados arrojados por el modelo anterior.

Tabla 5. Línea 144. Segundo modelo de predicción del nivel de riesgo. Niveles de riesgo observados por nivel de riesgo esperado. Total país. 2020-2022.

[9] *Tanto el corpus teórico vigente en materia de violencia por motivos de género como los datos disponibles a la fecha dan cuenta de la multicausalidad y la multiplicidad de factores que pueden dar cuenta de una situación de violencias por motivos de género y como la combinación de dichos factores tiene efectos diferenciales en cada caso particular.*

[10] *De acuerdo a datos publicados sobre la Línea 144, en más del 30% de las intervenciones realizadas se ha registrado la presencia de violencia simbólica.*

Predicciones1 Actual	Alto	Altísimo	Bajo	Medio
Alto	9833	113	253	173
Altísimo	1915	562	95	139
Bajo	2643	10	645	69
Medio	2574	101	259	442

Nota: los datos corresponden a la muestra de prueba extraída para aplicar el modelo de clasificación elaborado.

Fuente: elaboración propia en base a datos de la Línea 144.

Como se ve, este segundo modelo de predicción de riesgo no arrojo diferencias significativas respecto al aplicado anteriormente.

Conclusiones

Los procesos de producción, recolección y almacenamiento de datos representativos, junto con la garantía de acceso a los mismos y su uso resultan fundamentales para abordar diversas problemáticas sociales (Bercovich, Guaymás, Penna, Yankelevich;2021), entre las que se inscriben las violencias por motivos de género que atraviesan mujeres y personas LGBTI+.

En este sentido, el acceso público a los registros de las intervenciones realizadas por la Línea 144 en casos de violencia por motivos de género resulta absolutamente provechoso para, a través de dichos registros, obtener datos de gran utilidad para conocer algunos aspectos de las características de la violencia por motivos de género en nuestro país. Dicha temática requiere de abordajes específicos y pormenorizados, teniendo en cuenta la complejidad de la misma basada en su multicausalidad y la diversidad de factores que la traviesan.

El tratamiento adecuado de los datos que puedan obtenerse abre un amplio abanico de posibilidades: abren el camino para diseñar estrategias para un adecuado abordaje de los casos propios de dicho dispositivo, permite elaborar estrategias de seguimiento y monitoreo de la información, la obtención de información para insumos de diseño de programas y políticas públicas, etc.

Entre las posibilidades que ofrece la información disponible de la Línea 144, se logró establecer indicadores para la medición de riesgo, que permitieron clasificar cada registro de las intervenciones realizadas de acuerdo al *nivel de riesgo*. El análisis posterior arrojó que más de la mitad de los casos en los que la Línea 144 ha intervenido entre 2020 y 2022 son de riesgo “alto”, lo cual es consistente con las características de las situaciones plasmadas en los registros que componen los datos analizados: es decir, se trata de casos en donde la Línea ha realizado acciones de seguimiento e intervención directa.

A partir de dicho indicador, se exploraron dos modelos de clasificación para verificar cuan factible resulta predecir el nivel de riesgo ante la presencia de ciertos parámetros. Los resultados arrojaron que estos predictores, al menos con la información utilizada, no resultan adecuados para predecir los casos que tengan los máximos niveles de riesgo (altísimo) ni aquellos de nivel medio o bajo. Esto abre la puerta a, por ejemplo, explorar otras alternativas para la construcción de modelos de predicción de riesgo, o bien revisar los criterios elaborados para la clasificación de los registros según el nivel de riesgo.

La construcción tanto del dato del nivel de riesgo y los posteriores ensayos con el modelo predictor fueron posibles gracias a los lineamientos del Sistema Integrado de Casos de Violencias por Motivos de Género (SICVG). El diseño conceptual del mismo permitió transformaciones y adecuaciones relativamente sencillas desde una fuente de datos externa al mismo, originada desde un dispositivo de registro ajeno a dicho sistema.

Sin embargo, el SICVG comenzó a ser utilizado por la Línea 144 con posterioridad al segundo semestre del 2023, con lo cual es válido inferir que ya se encuentra utilizando la medición de riesgo automática del sistema

para la gestión de sus casos abordados. A la fecha no se han publicado datos referidos ni a estas mediciones ni a ese periodo, correspondientes a este dispositivo de atención, pero es de sumo interés acceder a estos en un futuro para poder contrastar con los resultados obtenidos en el presente documento y avanzar en futuras indagaciones.

Bibliografía

- Bercovich, S.; Guaymás, A.; Penna, F. y Yankelevich, D. (2021). Datos y algoritmos para el desarrollo. Buenos Aires. Fundar. Disponible en https://www.fund.ar.
- Crenshaw, K.. (1998) Demarginalising the intersection of race and sex. A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. University of Chicago legal forum, 14, pp. 538-554.
- Hernández-Leal, Emilcy J.; Duque-Méndez, Néstor D.; Moreno-Cadavid, Julián (2017) Big Data: una exploración de investigaciones, tecnologías y casos de aplicación Tecno Lógicas, vol. 20, núm. 39, mayo-agosto, 2017 Instituto Tecnológico Metropolitano Medellín, Colombia. Disponible en https://www.redalyc.org/articulo.oa?id=344251476001
- Ministerio de las Mujeres, Géneros y Diversidad. (2023). Línea 144. 10 años. Una década del Dispositivo Federal de Atención de las Violencias de Género. Disponible en https://www.argentina.gob.ar/sites/default/files/2022/05/linea144aniversario-web-v4.pdf
- Ministerio de las Mujeres, Géneros y Diversidad (2020). Plan Nacional de Acción Contra las Violencias (2020-2022). Disponible en: https://www.argentina.gob.ar/sites/default/files/plan_nacional_de_accion_2020_2022.pdf
- Ministerio de las Mujeres, Géneros y Diversidad (2022). Plan Nacional de Acción Contra las Violencias (2022-2024). Disponible en: https://www.argentina.gob.ar/sites/default/files/2022/08/pna_2022_2024.pdf
- Ministerio de las Mujeres, Géneros y Diversidad (2022) SICVG. Primer informe estadístico. 2022. Disponible en: https://www.argentina.gob.ar/sites/default/files/2023/02/sicvg_-_informe_estadistico.pdf
- Ministerio de las Mujeres, Géneros y Diversidad (2023). SICVG. Segundo informe estadístico. Disponible en: https://www.argentina.gob.ar/sites/default/files/2023/02/2023-segundo-informe-estadistico-sicvg.pdf
- Ministerio de las Mujeres, Géneros y Diversidad (2022). Sistema Integrado de Casos de Violencias por Motivos de Género. Características principales. Disponible en: https://www.argentina.gob.ar/sites/default/files/2022/07/presentacion-sicvg-mmgyd.pdf
- Ministerio de las Mujeres, Géneros y Diversidad (2023). Sistema Integrado de Casos de Violencias por Motivos de Género. Medición de riesgo en el sistema. Disponible en: https://www.argentina.gob.ar/sites/default/files/2022/07/20230509-medicion-de-riesgo-sicvg.pdf
- Sosa Escudero, W. (2019). BIG DATA Big Data: Breve manual para conocer la ciencia de datos que ya invadió nuestras vidas. Buenos Aires: Editorial Siglo XXI

Interoperabilidad y grandes volúmenes de datos

De Ashbey Fernando, Coll Julieta, Ibarra Adrián
& Zumárraga Juan Pablo

Ashbey Fernando; Coll Julieta; Ibarra Adrián; Zumárraga Juan Pablo

Interoperabilidad y grandes volúmenes de datos

Cómo potenciar el diseño de políticas públicas basada en evidencia

La interoperabilidad y la gobernanza son conceptos que se han vuelto cada vez más importantes en el ámbito de la Administración Pública Nacional. En el informe "Adopción de la Interoperabilidad como criterio" (2004), se destaca el Marco de Interoperabilidad Europeo como un nuevo paradigma para entender la gobernanza en Europa. En Argentina, la Oficina Nacional de Tecnología Informática (ONTI) y el Foro Permanente de Responsables Informáticos impulsaron la Resolución 99/2008, que dio origen al componente de Interoperabilidad para el Gobierno Electrónico. Este fue el primer uso del término interoperabilidad para referirse a la interacción entre organismos de la Administración Pública Nacional. Desde entonces, se han creado varias normas y trabajos que complementan la temática de interoperabilidad en el Estado. El concepto de Gobierno Electrónico ha evolucionado para dar lugar al de Gobierno Abierto, donde la ciudadanía ocupa el centro de las acciones de gobierno. La Administración Pública ha puesto a disposición de la ciudadanía numerosas fuentes de datos, en el marco del Decreto 117/2016 y la Ley N° 27.275.

El trabajo investigó acerca de las posibilidades de optimización de las políticas públicas con el fin de estudiar las posibilidades de interoperabilidad dentro del Estado Nacional y de obtener evidencia para el diseño inteligente e integrado de estas. Para ello, se tomó una actividad específica, en este caso la producción apícola y se relevó e identificó las distintas fuentes de datos en poder de los distintos organismos nacionales, indagando en la información disponible y que se encontraban publicadas en los distintos sitios de datos abiertos y solicitando información que, si bien no se encontraba disponible, podía ser solicitada mediante pedidos de acceso a la información pública, de acuerdo a la Ley N°27.275, o a la que podían acceder los organismos de la Administración Pública mediante convenios de cooperación interinstitucional.

Interoperabilidad y grandes volúmenes de datos

How to enhance the design of evidence-based public policies

Interoperability and governance are concepts that have become increasingly important in the field of National Public Administration. In the report "Adoption of Interoperability as a Criterion" (2004), the European Interoperability Framework is highlighted as a new paradigm for understanding governance in Europe. In Argentina, the National Office of Information Technology (ONTI) and the Permanent Forum of IT Managers promoted Resolution 99/2008, which gave rise to the Interoperability Component for Electronic Government. This was the first use of the term interoperability to refer to the interaction between National Public Administration agencies. Since then, several regulations and works have been created that complement the theme of interoperability in the State. The concept of Electronic Government has evolved to give way to Open Government, where citizens are at the center of government actions. The Public Administration has made numerous sources of data available to citizens, under Decree 117/2016 and Law No. 27,275.

The work investigated the possibilities of optimizing public policies in order to study the possibilities of interoperability within the National State and to obtain evidence for the intelligent and integrated design of these. To this end, a specific activity was taken, in this case beekeeping, and the different sources of data in the possession of the different national agencies were surveyed and identified, investigating the available information and that which was published on the different open data sites and requesting information that, although not available, could be requested through requests for public information access, in accordance with Law No. 27,275, or that which could be accessed by Public Administration agencies through inter-institutional cooperation agreements.

El objetivo del presente trabajo fue **pensar una nueva concepción para la gestión de la información y los datos del Estado Nacional**.

Se abordaron y cumplieron los siguientes objetivos:

- Se destacó la importancia de la elección adecuada de fuentes de datos.
- Se realizó un análisis detallado de información en diversas dependencias estatales.
- Se identificaron fuentes de datos interoperables.
- Se generó información significativa para la toma de decisiones y la formulación de propuestas de mejora en políticas públicas.
- Se buscó fomentar el intercambio de información entre diversos actores de la Administración Pública, con el fin de obtener evidencia al momento de diseñar, concebir y evaluar las Políticas Públicas.

Conceptos básicos y procedimiento

Tomando como base los principios de interoperabilidad y apertura de datos, este trabajo pretende, eligiendo el sector apícola, encontrar posibilidades de optimización de políticas públicas y conocimiento del estado de situación de la actividad, entendiendo que este enfoque permitiría orientar las políticas públicas a los objetivos deseados en base a la evidencia disponible.

Es decir, si fomentamos la interacción de la información con la que contamos podremos obtener un mayor conocimiento del ambiente que nos rodea, de la ciudadanía, de los diferentes actores sociales, de su desarrollo económico y de sus necesidades. De esta forma, tanto la ciudadanía como el sector público tendrán herramientas para fortalecer un mejor diseño de Políticas Públicas.

Bajo estos preceptos, el presente trabajo identificó oportunidades de intercambio de información que permitió obtener un mayor conocimiento del sector. Se destaca que este enfoque es aplicable a otras ramas de actividad y objetivos.

En virtud de ello, el primer paso fue encontrar fuentes de información en los organismos a los que pertenecemos con la finalidad de comenzar a investigar las posibilidades de interoperabilidad fácilmente alcanzables, dado la disponibilidad de las fuentes de información. En este sentido, dadas las posibilidades de acceso y confiabilidad, se seleccionó una fuente primaria sobre la que se investigarán las posibilidades de interoperabilidad. A esa fuente la denominamos el “Punto de partida”.

A partir de ello, se exploraron las instancias de información existentes en la plataforma de Datos Abiertos del Sector Público Nacional. A esta fuente le llamamos “Información disponible”.

A los fines de recabar mayor información se hizo uso del Régimen de Acceso a la Información Pública (Ley N° 27.275) con el objetivo de obtener información específica que no se encuentra publicada. A esta fuente le llamamos “Información solicitable”.

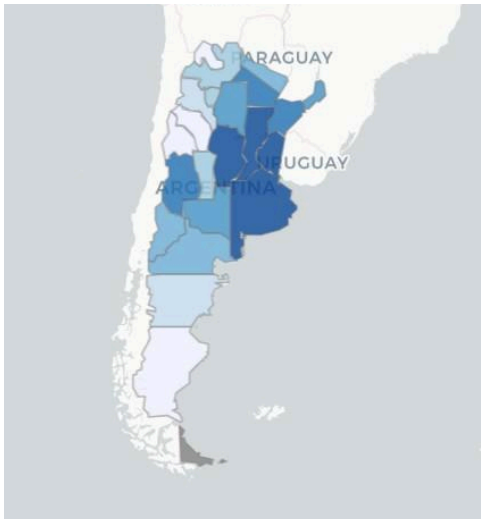
Una vez recabada la información mencionada se comenzaron a explorar las diferentes fuentes de información con la finalidad de identificar posibles interacciones entre las diferentes áreas de gobierno para ensayar herramientas de interoperación que permitieran lograr mejores resultados de políticas públicas.

Punto de partida seleccionado - RENAPA

Información disponible del Registro Nacional de Productores Apícolas

Se selecciona como punto de partida al Registro Nacional de Productores Apícolas (RENAPA) perteneciente al Ministerio de Agricultura, Ganadería y Pesca, por ser una fuente sobre la que tenemos conocimiento y fácil acceso.

El Registro Apícola tiene 15.614 inscriptos y a partir de la exploración de los datos, podemos ver la distribución geográfica de los inscriptos tal como se muestra en el siguiente mapa. Allí vemos según el porcentaje de inscriptos por provincia, la intensidad del color, siendo el color azul oscuro el más alto.

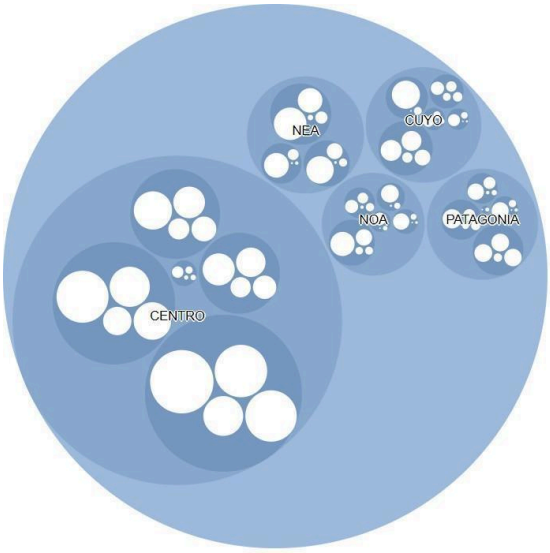


Siguiendo con el análisis categorizamos el registro de acuerdo a una clasificación propia que tiene que ver con la cantidad de colmenas que tiene el productor.

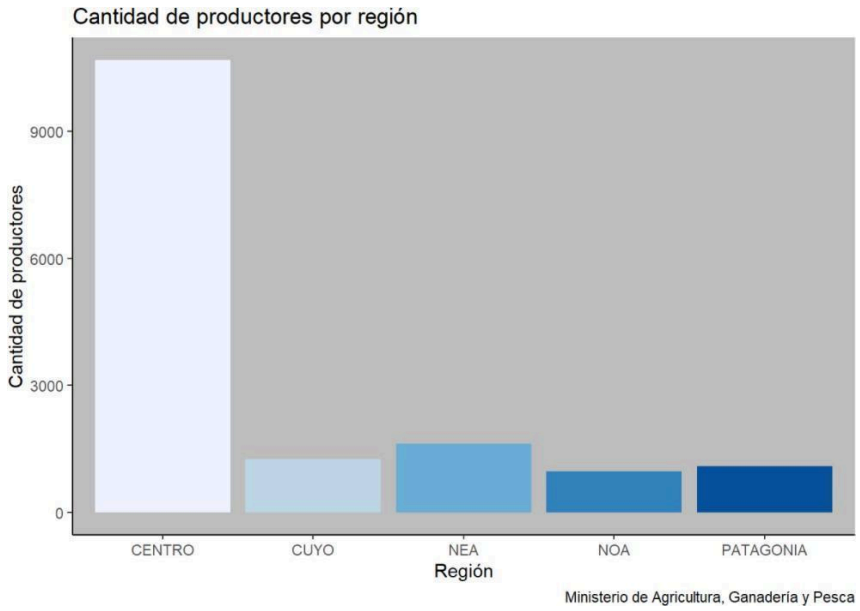
Esta estratificación permite clasificar a los productores en:

- Más de 500 colmenas
- Hasta 500 colmenas
- Más de 200 colmenas
- Hasta 50 colmenas

De esta manera, podemos obtener mayor detalle de la representatividad de la actividad en las distintas regiones, por provincia y por tamaño de los productores, lo que se hace mediante un gráfico:



En la siguiente imagen podemos ver esta información de otra manera:



De esta manera, podemos obtener conocimiento de la distribución y tamaño de la actividad en el país.

A los fines de buscar fuentes de datos relevantes para interoprerar con la base de datos e información disponible, utilizamos:

- Datos propios del organismo de pertenencia (Ministerio de Trabajo, Empleo y Seguridad Social) – Información solicitable
- Exploración de datos abiertos (Ministerio de Desarrollo Productivo / Ministerio de Economía / Ministerio de Agricultura, Ganadería y Pesca) - Información disponible
- Régimen de Acceso a la información Pública (Superintendencia de Riesgos del Trabajo / Ministerio de Desarrollo Productivo) - Información solicitable.

Exploración de otras fuentes de datos

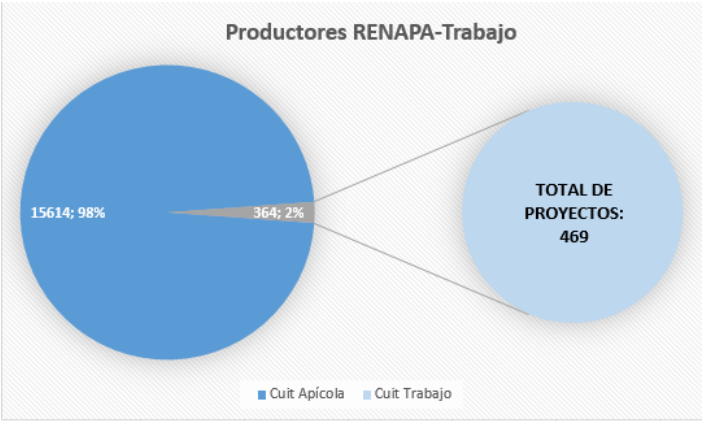
Uso de datos propios de otros organismos

Proyectos y programas de Empleo, Capacitación y Formación Profesional del Ministerio de Trabajo, Empleo y Seguridad Social

Dentro de la información solicitable que no se encuentra publicada a la que tenemos acceso se encuentran los datos sobre proyectos y programas de Empleo, Capacitación y Formación Profesional de la Secretaría de Empleo.

Esta información está disponible desde 2001 hasta la fecha y es posible cruzarla con el RENAPA a través de la información disponible en ambas bases de Código Único de Identificación Laboral (CUIL) y Código Único de Identificación Tributaria (CUIT) de las personas humanas y jurídicas involucradas en la actividad.

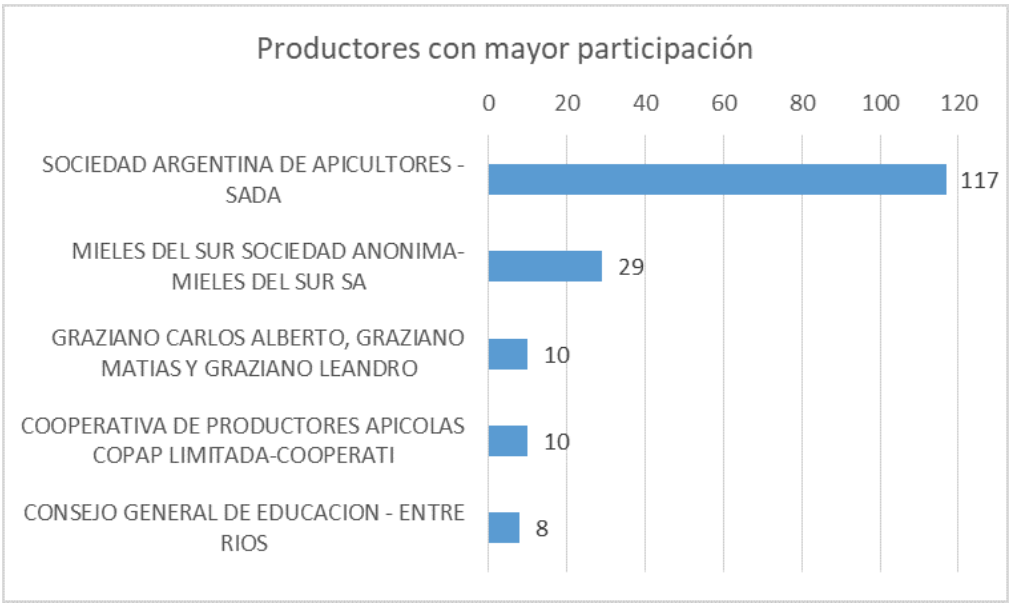
Del cruce de información de RENAPA con las BD del MT, podemos ver que de los 15.614 CuitApicolas, coincidieron 364 con los CuitTrabajo. Un 2% sobre el total de la muestra.



A su vez podemos ver que este 2%, o los 364 productores, participaron en 469 proyectos relacionados con el sector apícola, dentro de las prestaciones de Empleo o Capacitación del MT. Cabe destacar que la participación en ésta cantidad de proyectos se da desde el 2001 a la actualidad.

Programas de Empleo/Capacitación	2001-2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	Total general
APOYO A LAS MIPYME 2001	**4**											**4**
CAPACITACION SECTORIAL	28											28
CREAR TRABAJO	**4**											**4**
Empleo Independiente	0		1									1
ENTRAMADOS PRODUCTIVOS	**0**	**2**			**1**							**3**
Entrenamiento para el trabajo	1	6	10	4	9	10	49	63	47	49	35	283
FORMACION PROFESIONAL	**6**	**2**	**31**	**42**	**29**	**14**						**124**
Fortalecimiento Institucional	0		2			1	1					4
JEFES DE HOGAR	**5**											**5**
PIL	0		1	1		1	4		4			11
PROGRAMA FORMACION DOCENTE	**0**			**1**	**1**							**2**
Total general	48	10	45	48	40	26	54	63	51	49	35	469

Adicionalmente, podemos observar la distribución geográfica de los proyectos:

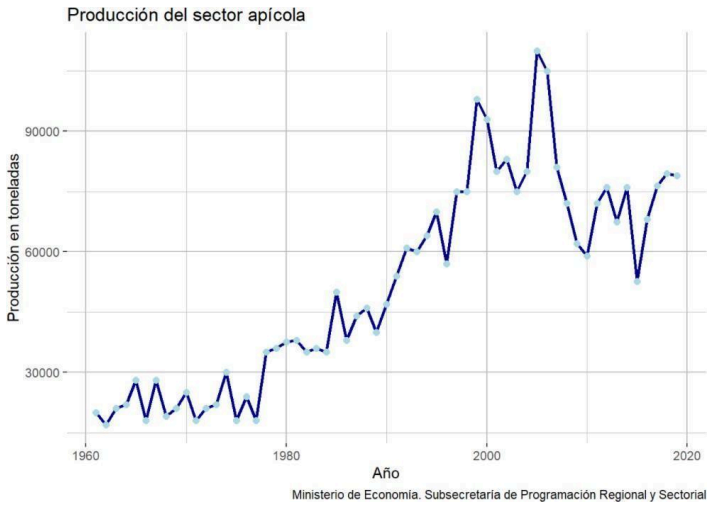


Exploración de datos abiertos

Indicadores Sectoriales y Provinciales del Ministerio de Economía

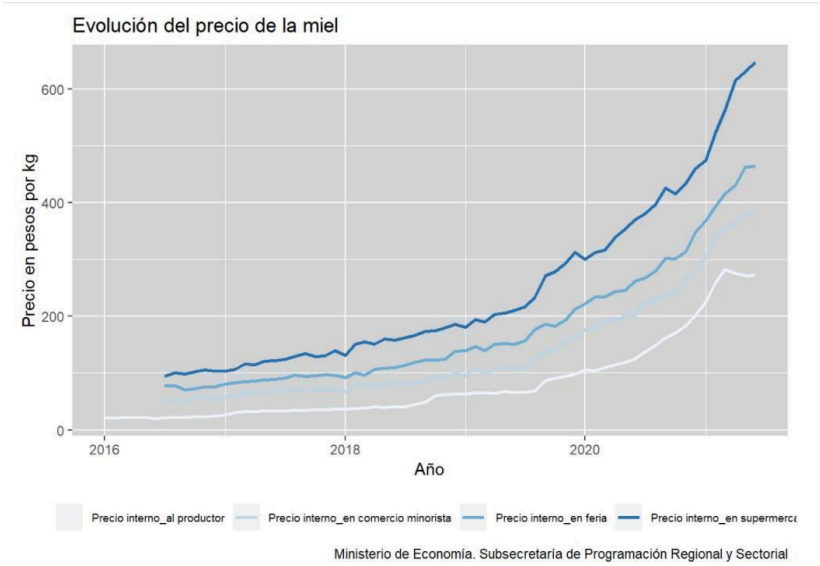
La Subsecretaría de Programación Regional y Sectorial del Ministerio de Economía pública en el portal de Datos Abiertos información sobre la evolución de la actividad apícola.

En el siguiente gráfico vemos la evolución de la producción de miel desde 1961 hasta 2019, y podemos observar como el pico de toneladas de miel producidas se dio en 2005 y luego comenzó a descender la producción hasta llegar en 2015 a los mismos valores que se producían en 1991. En los últimos años, se ve un leve aumento de la producción.



Esta información es útil para conocer las capacidades de producción del sector y puede mostrarnos la necesidad de dirigir políticas a aumentar la producción de miel, máxime habiéndo visto al explorar el RENAPA que algunas regiones como el NEA tienen muchos pequeños productores y pocos grandes productores. En este sentido, de acuerdo a la distribución de los programas de Empleo, Capacitación y Formación Profesional de la Secretaría de Empleo, vemos que de las provincias del NOA únicamente Chaco ha recibido programas de capacitación.

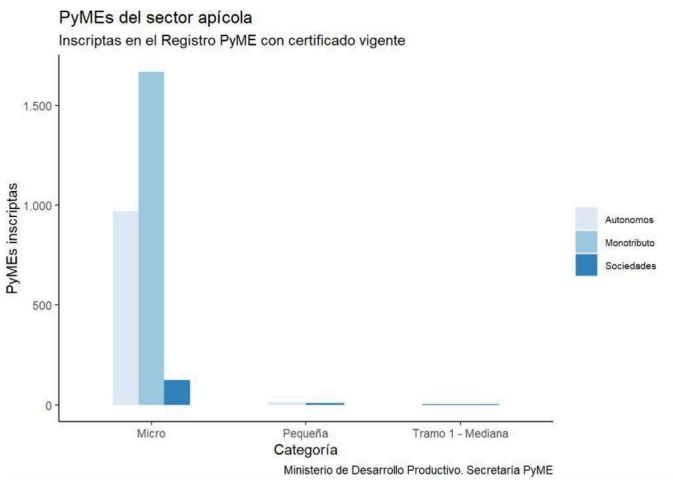
Otra información que podemos obtener del dataset sobre la evolución de la actividad apícola es el precio de la miel según su comercialización.



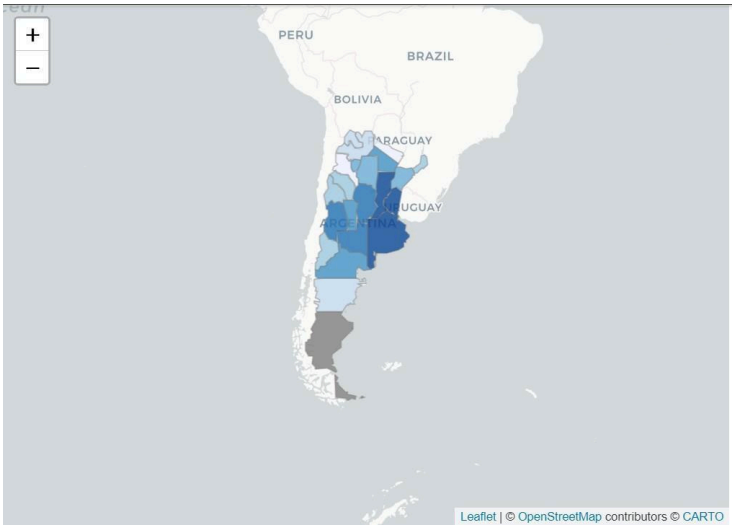
Esta fuente de datos puede ser útil a la hora de desarrollar políticas públicas de informar a los consumidores como así también para dirigir esfuerzos a identificar posibles distorsiones en el mercado.

Registro PyME del Ministerio de Desarrollo Productivo

Mediante la Ley N° 24.467 se estableció un régimen de promoción al crecimiento y desarrollo de las pequeñas y medianas empresas (PyMEs). Se destaca que la información se encuentra publicada anonimizada por lo que no es posible hacer un cruce a nivel CUIT pero si podemos obtener datos de las PyMEs del sector apícola.



Podemos observar que la mayoría de las PyMEs inscriptas con actividad principal apícola son Microempresas bajo el régimen tributario del monotributo.



Podemos observar como, del total de PyMEs con certificado vigente cuya actividad principal pertenece al sector apícola, la mayoría se encuentra en la Provincia de Buenos Aires (48.52%), seguida por Entre Ríos (19.52%) y Santa Fe (11.30%). Esta distribución se corresponde con la propia distribución del RENAPA y, para fortalecer algunas observaciones que se fueron haciendo a la largo del trabajo, notamos lo siguiente:

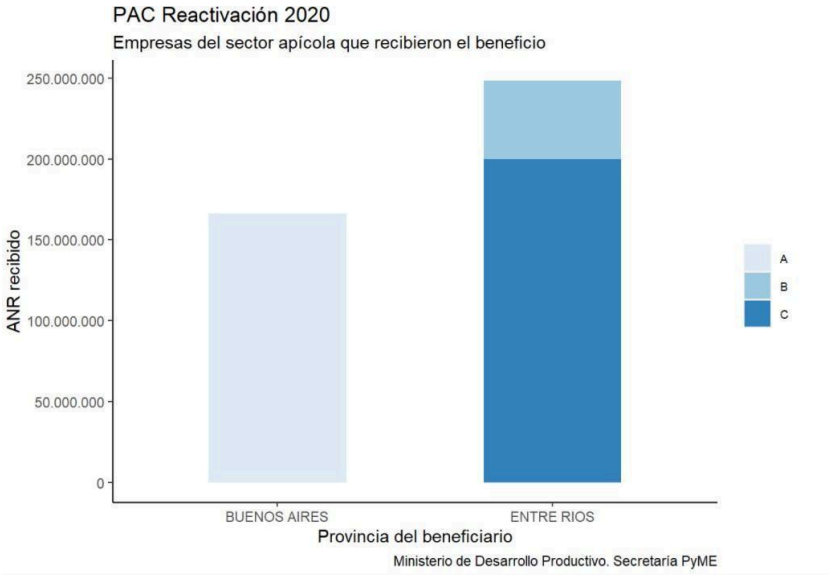
- La Patagonia cuenta con pocas PyMEs en el sector apícola, mayormente concentradas en La Pampa.
- En el NEA hay muy pocas PyMEs del sector apícola inscriptas, lo que es llamativo ya que es la segunda región con más productores y la mayoría de ellos (el 66%) tiene hasta 50 colmenas.

De esta forma vemos como se puede obtener mayor información del sector tomando datos de otros organismos. Por ejemplo, en este caso, se podrían dirigir distintas políticas para que aquellos pequeños productores del sector inscriptos en el RENAPA que no se encuentran inscriptos en el Régimen PyME conozcan los beneficios que el Registro PyME tiene disponible.

Programa de Apoyo a la Competitividad (PAC): Emprendedores Reactivación Productiva 2020 del Ministerio de Desarrollo Productivo

Otro de los Programas del Ministerio de Desarrollo Productivo que cuenta con información publicada en Datos Abiertos es el Programa PAC Emprendedores Reactivación Productiva que estuvo vigente durante el año 2020.

Tomando los datos del Registro Apícola y cruzándolos con la base de datos publicada del Programa, podemos observar que tres beneficiarios del PAC Emprendedores Reactivación Productiva se encuentran registrados en el Registro Apícola. Se muestra la información anonimizada:



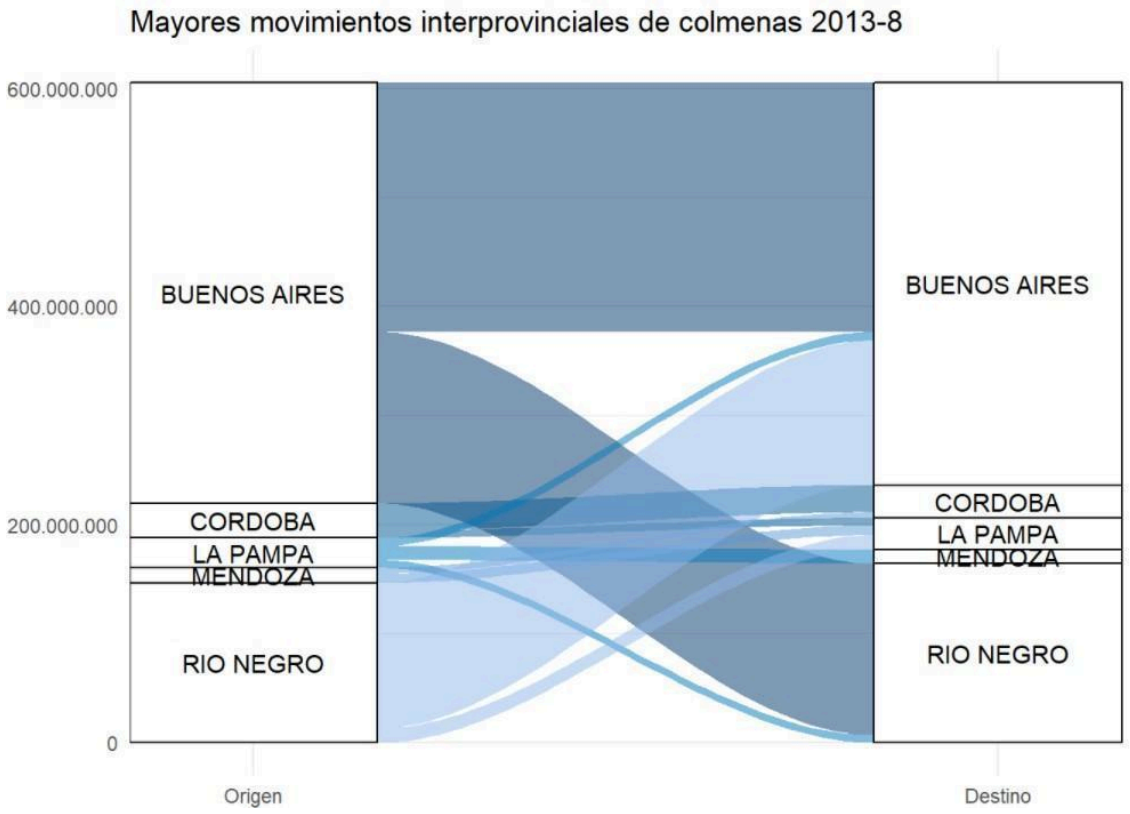
Se puede ver que, como era de esperar, las empresas beneficiadas corresponden a dos de las provincias con mayor actividad apícola.

Este cruce nos permite tener conocimiento adicional del sector ya que vemos que algunas empresas del sector tienen capacidad y están trabajando en mejorar la productividad e introducir innovación en el sector.

Información de los movimientos apícolas del SENASA

El Servicio de Sanidad y Calidad Agroalimentaria (SENASA) es el organismo encargado de velar por la seguridad de los alimentos que se producen y comercializan en el país. Esto comprende autorizar los movimientos, en este caso, de la espcie apícola.

Gracias a la publicación en el portal de Datos Abiertos de los datos de los movimientos de colmenas que el SENASA autoriza, podemos conocer entre qué provincias se da el mayor movimiento de colmenas.



Podemos observar como el mayor movimiento de colmenas se registró dentro de la Provincia de Buenos Aires y también hacia la Provincia de Río Negro, en el mismo sentido, se observan numerosos movimientos desde Río Negro hacia Buenos Aires.

Información de Agricultura Familiar del Ministerio de Agricultura, Ganadería y Pesca

El Ministerio de Agricultura, Ganadería y Pesca publica en el sitio de Datos Abiertos información sobre producciones de la agricultura familiar animal y de la agricultura familiar agroindustrial. Al respecto, se destaca que la información no es clara y en algunos casos está duplicada y mal rotulada.

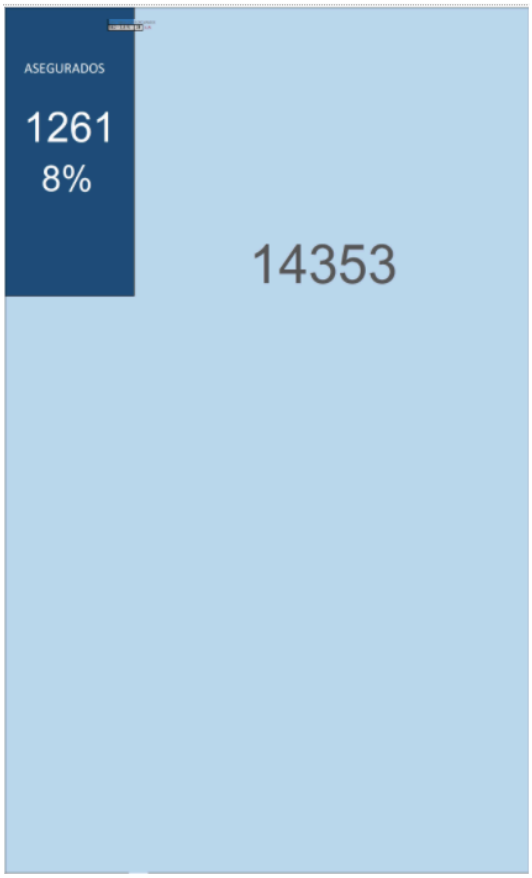
Por ejemplo, en los datos de la agricultura familiar animal aparece un registro para “Miel” y otro para “Polinización”, ambos en el mismo departamento, provincia y con la misma cantidad de producciones, cabezas y vientres, lo que nos lleva a inferir que se trata del mismo núcleo de agricultura familiar que realiza ambas actividades apícolas.

Por estos motivos, no se ha profundizado en el análisis de los datos de la agricultura familiar, pero se destaca que están disponibles en Datos Abiertos y que es una fuente de información que podría utilizarse por parte de los organismos interesados obteniendo las aclaraciones pertinentes.

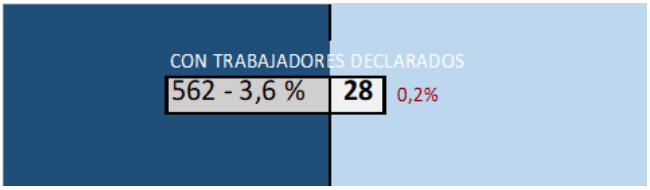
En este caso, se hizo uso de la herramienta provista por la Ley N° 27.275 de Acceso a la Información Pública para solicitar información proveyendo el universo conocido del RENAPA para efectuar la consulta. Al respecto, se recuerda que el régimen de Acceso a la Información Pública tiene como precepto la informalidad, por lo que puede realizarse por mail, por comunicación oficial, mediante nota en las meses de entrada de los organismos, etc. En este caso, se hizo mediante comunicación oficial del Sistema de Gestión Documental Electrónica que es la herramienta de comunicación entre los diversos actores y empleados de la Administración Pública Nacional.

La solicitud se hizo a la Superintendencia de Riesgos de Trabajo y se solicitó información acerca del universo cubierto por Aseguradoras de Riesgos del Trabajo (ART).

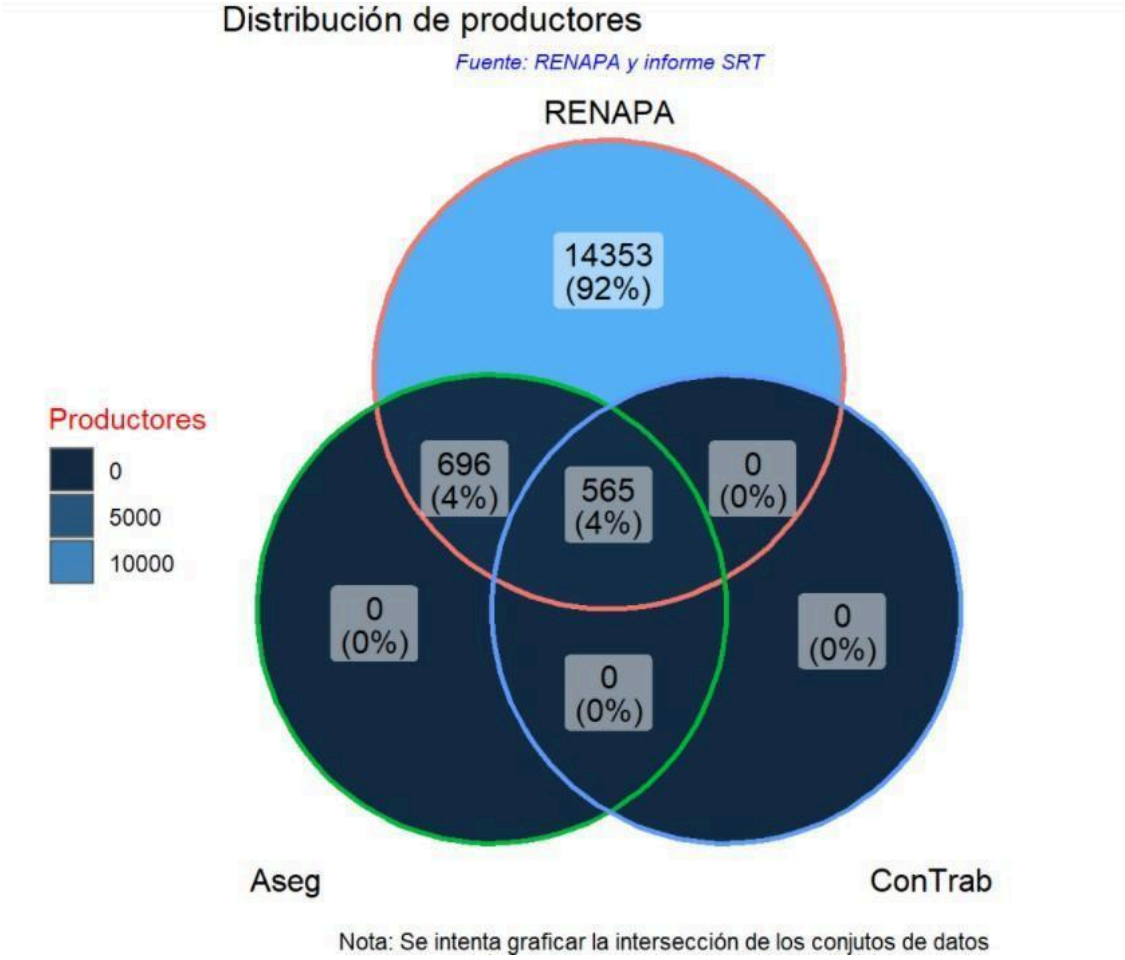
Una vez obtenida, se cruzó la información del RENAPA con la del Sistema de Riesgos del Trabajo y nos encontramos que el 8 % de los integrantes del registro tienen un contrato con una ART, y de ese universo sólo el 45 % declara trabajadores en relación de dependencia. No obstante, es sólo el 3,6% de la totalidad de registro. Encontramos un 0.2 % del Registro que no tiene cobertura y tiene trabajadores en relación de dependencia.



Haciendo Zoom:



Aplicando algunos paquetes de R podemos mostrar esto:



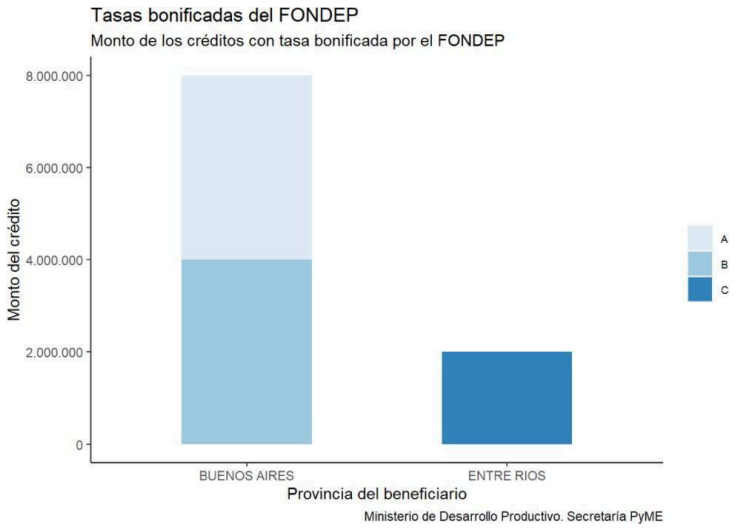
Nota: Se intenta graficar la intersección de los conjutos de datos

Créditos con tasa bonificada del FONDEP del Ministerio de Desarrollo Productivo

El Ministerio de Desarrollo Productivo publica en Datos Abiertos información acerca de la ejecución de líneas de crédito con tasa bonificada por el Fondo Nacional de Desarrollo Productivo (FONDEP) otorgados durante 2020 a través de entidades bancarias.

Sin embargo, esta información no se encuentra desagregada por sector de actividad por lo cual fue solicitada de manera informal. Por este motivo se encuentra dentro de la categoría “información solicitable”.

La información que se obtuvo es la siguiente:

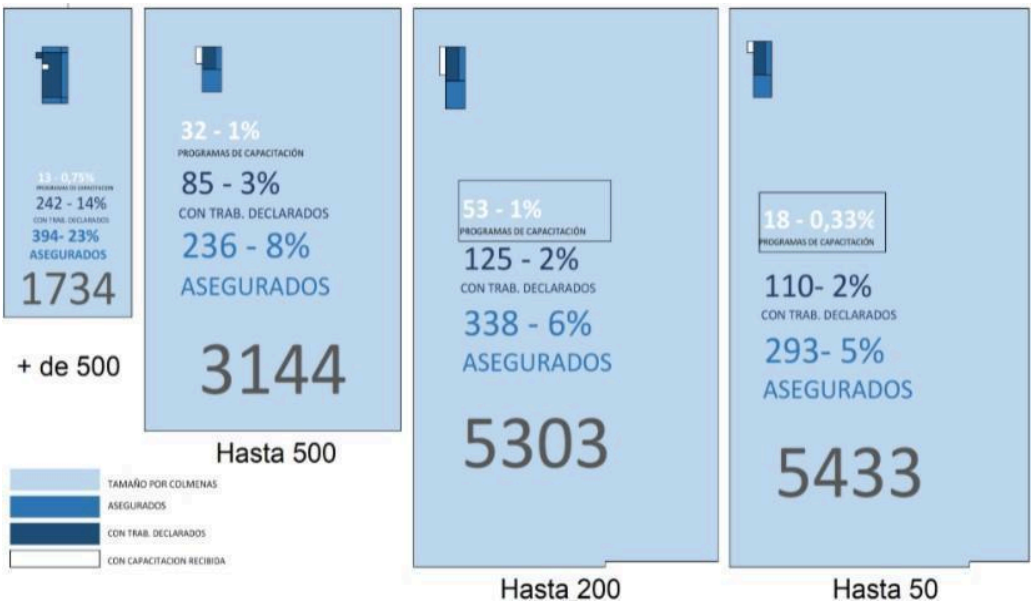


Vemos en este caso que las empresas beneficiarias se encuentran en las provincias que tienen mayor actividad apícola, lo que es un resultado esperable.

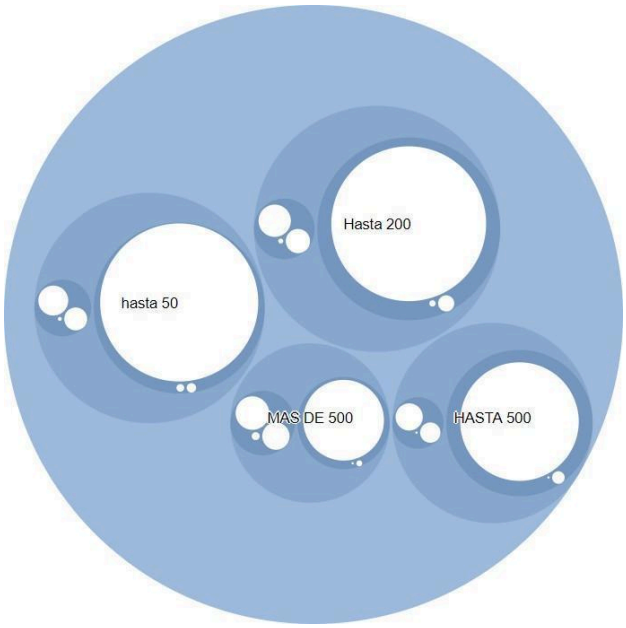
Los datos hablan - oportunidades de mejora

Siguiendo con la categorización de los productores según su tamaño vista al caracterizar el RENAPA, cruzamos además la fuente de datos del Ministerio de Trabajo, Empleo y Seguridad Social y la fuente de datos de la Superintendencia de Riesgos de Trabajo para verificar:

- Universo de productores apícolas asegurados
- Universo de productores apícolas con trabajadores declarados y
- Universo de productores apícolas que fueron beneficiarios de programas de Empleo y Capacitación.



Aplicando algunos paquetes de R podemos mostrar esto:



Con estos cruces podemos observar las siguientes oportunidades de política pública:

- El porcentaje de productores apícolas que recibieron o reciben programas de capacitación y formación del Ministerio de Trabajo, Empleo y Seguridad Social es bajo, lo que podría ser un disparador para articular políticas públicas en conjunto entre el Ministerio de Agricultura, Ganadería y Pesca y el Ministerio de Trabajo, Empleo y Seguridad Social

- A medida que aumenta la cantidad de colmenas por productor aumenta la cantidad de productores con trabajadores declarados y asegurados, sin embargo, hay una diferencia entre estas categorías, por lo que podríamos suponer que el cruce de estas bases de datos podría ser una herramienta de fiscalización.

A partir de esto, con la información recabada y luego de efectuar los cruces correspondientes podemos elegir dos instancias de análisis:

- Corroborar con evidencia los supuestos que se tienen acerca de la conformación de los universos productivos y la distribución por tamaño y/o territorio

- Tomar determinaciones para orientar la política pública en forma conjunta a sectores individualizados.

En este último punto nos animamos a enunciar una política que pueda combinar las acciones de varias áreas de gobierno.

Teniendo en cuenta que las acciones de capacitación aparecen distribuidas en sectores no alcanzados por el resto de las acciones tanto de trabajo registrado como de cobertura de riesgos del trabajo, como se puede ver en los gráficos recuadrados. Se propone desarrollar una política pública que atienda los siguientes aspectos:

- Desarrollar esquemas de incentivo de PYMES con bonificaciones de créditos o exención de tasas para aquellos productores que registren a sus empleados en relación de dependencia (por ejemplo programa de “primer empleo” o “credito fiscal”) y contraten un seguro de riesgos de trabajo (con una alícuota subsidiada) y se les proporcione capacitación en sus proyectos.

- Desarrollar planes de capacitación para trabajadores de las empresas que cuentan con personal temporario no declarado para que cuenten con capacitación sobre los principales accidentes o enfermedades laborales producidas en el ámbito elegido.

Conclusión

Como cierre del trabajo ponemos en valor un decálogo publicado en el sitio MI ARGENTINA denominado Decálogo de la ONTI, que en su punto 4 detalla:

La interoperabilidad y la utilización de estándares abiertos permiten la compatibilidad entre distintas tecnologías y ahorrar en costos de desarrollo o contratación de servicios. Además, facilita la colaboración entre organismos al mismo tiempo que fomenta la transparencia en la Administración Pública y la reducción de la dependencia de oferentes.

Bibliografía

- Clusellas, Pablo; Martelli, Eduardo; Martelo, María José (2019) *Un gobierno inteligente: El cambio de la Administración Pública de la Nación Argentina 2016-2019.* Recuperado el dia 23 de junio de 2021 de https://www.boletinoficial.gob.ar/pdfs/gobierno_inteligente.pdf
- Comisión de Infraestructura Tecnológica y Ciberseguridad del COFEFUP. (24 de septiembre de 2020) *La Interoperabilidad de datos y sistemas y la ciberseguridad en la administración pública desde una perspectiva federal.* [Conclusiones]. 2ª Asamblea del Consejo Federal de Función Pública (CoFeFuP). Recuperado el dia 23 de junio de 2021 de https://www.argentina.gob.ar/noticias/la-interoperabilidad-de-datos-y-sistemas-y-la-ciberseguridad-en-la-administracion-publica
- Decálogo Tecnológico ONTI (8 de noviembre de 2019 Versión 1.1.0) Recuperado el 23 de junio de 2021 de https://www.argentina.gob.ar/jefatura/innovacion-publica/ssetic/onti/decalogo-tecnologico
- Decreto 1273 de 2016 [con fuerza de ley] Por el cual se establece que las entidades y jurisdicciones enumeradas deberán intercambiar información. 19 de diciembre de 2016 http://servicios.infoleg.gob.ar/infolegInternet/verNorma.do?id=269242
- Héctor Poggiese, María Elena Redín, Matías Cerezo, José Manuel Carllini ( ) El Foro de Responsables Informáticos de la Administración Pública Nacional: Una Lectura Interpretativa. Resolución 99 de 2008. [Secretaría de Gabinete y Gestión Pública de la Jefatura de Gabinete de Ministros] Por la cual se crea el Componente de Interoperabilidad para el Gobierno Electrónico en el ámbito de la Oficina Nacional de Tecnologías de Información. 30 de diciembre de 2018 http://servicios.infoleg.gob.ar/infolegInternet/anexos/145000-149999/149270/norma.htm
- Ley 27275 de 2016. Por la cual se garantiza el efectivo ejercicio del acceso a la información pública promoviendo la participación ciudadana y la transparencia en la gestión pública. 29 de septiembre de 2016. http://servicios.infoleg.gob.ar/infolegInternet/verNorma.do?id=265949
- Moreno Escobar, Herán; Sin Triana, Hugo y Silveira Netto, Sérgio (2007). *Conceptualización de arquitectura de Gobierno Electrónico y Plataforma de Interoperabilidad para América Latina y el Caribe.* Publicación de Naciones Unidas. Cáp. II y III www.cepal.org/socinfo/. CEPAL 2007 concep arq ge – 0.pdf
- Naser, Alejandra y Concha, Gastón (2012) *Datos abiertos. Un nuevo desafío de los gobiernos de la región. Instituto Latinoamericano y del Caribe de Planificación Económica y Social.* Publicación de Naciones Unidas. Recuperado el 1 de agosto de 2021 de https://www.cepal.org/es/publicaciones/7331-datos-abiertos-un-nuevo-desafio-gobiernos-la-region
- Naser, Alejandra (2021) *Gobernanza digital e interoperabilidad gubernamental: una guía para su implementación.* Publicación de Naciones Unidas. Recuperado el día 19 de agosto de 2021 de https://www.cepal.org/es/publicaciones/47018-gobernanza-digital-interoperabilidad-gubernamental-guia-su-implementacion
- Poggi, Eduardo (2008) *Modelos de Madurez para la Interoperabilidad*. Monografía presentada en el 2° SIE / 37° JAIIO 2008, Santa Fe, Argentina, Septiembre de 2008. (2° Premio Nacional de Gobierno Electrónico).
- Poggiese Héctor; Redín, María Elena; Cerezo, Matías; Carllini, José Manuel ( ) *El Foro de Responsables*

Informáticos de la Administración Pública Nacional: Una Lectura Interpretativa.

- Resolución 99 de 2008. [Secretaría de Gabinete y Gestión Pública de la Jefatura de Gabinete de Ministros] Por la cual se crea el Componente de Interoperabilidad para el Gobierno Electrónico en el ámbito de la Oficina Nacional de Tecnologías de Información. 30 de diciembre de 2018 http://servicios.infoleg.gob.ar/infolegInternet/anexos/145000-149999/149270/norma.htm
- Resolución 538 de 2013 [Jefatura de Gabinete de MInistros] Por la cual se reglamenta la creación del programa SINDAP, Sistema Nacional de Datos Públicos. 18 de julio de 2013 http://servicios.infoleg.gob.ar/infolegInternet/anexos/215000-219999/218131/norma.htm
- Resolución 19 de 2018 [Secretaría de Modernización Administrativa. Ministerio de Modernización.] Por la cual se implementan los servicios de INTEROPER.AR. 2 de marzo de 2018. http://servicios.infoleg.gob.ar/infolegInternet/anexos/305000-309999/307439/norma.htm
- Sanchez, Claudia A. (2019) *Interoperabilidad en la gestión pública.* [Tesis de Maestría, Universidad de San Andrés]. Recuperado el 23 de junio de 2021 de https://repositorio.udesa.edu.ar/jspui/bitstream/10908/16162/1/%5BP%5D%5BW%5D%20T.%20M.%20Ges.%20S%C3%A1nchez%2C%20Claudia.pdf
- Universidad Privada Dr. Rafael Belloso Chacín. (2009) *Adopción de la Interoperabilidad como criterio.* Venezuela Recuperado el 23 de junio de 2021 de http://www11.urbe.edu/conferencias/images/stories/3_marco_interoperabilidad_europeo_vf.pdf

TRABAJO DE INVESTIGACIÓN PUBLICADO EN LÍNEA: *https://rpubs.com/julietacoll/trabajo_final_UNAB*

72

Caracterización de la fatalidad vial en NEA a partir de modelos de Machine Learning

De Brián Covaro

Brián Covaro

Caracterización de la fatalidad vial en NEA a partir de modelos de Machine Learning

Introducción

El objetivo principal de este trabajo es explorar y analizar las particularidades de la fatalidad vial en NEA[1] para el período 2019 – 2020. A partir de datos oficiales sobre siniestros viales de este bienio, se construyeron distintos modelos de Machine Learning, basados en algoritmos distintos, que buscaron caracterizar y contribuir a explicar los patrones que presentan los siniestros viales tipificados como fatales. Entre 2008 y 2020, en promedio, 5.500 personas pierden la vida en siniestros viales en Argentina. Es la principal causa de muerte en rango etario de 15 a 35 años y tercera causa de muerte por "causas externas".

La elección del período bajo estudio (2019 -2020) responde a la intención de trabajar con los datos a "año completo", lo más actualizados posible que posee el sistema de registro (SIGISVI)[2]. La elección de la región (NEA), se basa en, primero, la disponibilidad de los registros en dicho sistema; y segundo, en que NEA es la región con el mayor nivel de siniestralidad y mortalidad del período bajo estudio.[3] La preparación de datos y la aplicación de modelos se realizó enteramente con el software RStudio y QGIS.

[1] Lamentablemente, este trabajo no puede contar con los datos de siniestralidad vial de la provincia de Misiones por problemas en la carga de los mismos por parte de la jurisdicción.

[2] Sistema Integral de Gestión de la Información de Seguridad Vial. El SIGISVI es el sistema de carga y registro de siniestros viales que posee la Agencia Nacional de Seguridad Vial (ANSV, en adelante). Este sistema es ofrecido a las jurisdicciones para el correcto tratamiento de la información en este ámbito. A la fecha, 16 jurisdicciones cargan sus datos en este sistema.

[3] ANSV. Observatorio Vial. 2019. Anuario 2019. https://www.argentina.gob.ar/sites/default/files/2018/12/ansv_ov_anuario_estadistico_2019_final.pdf.
ANSV. Observatorio Vial. 2020. Informe Anual 2020. Datos preliminares. https://www.argentina.gob.ar/sites/default/files/2018/12/ansv_ov_informe_anual_2020_al_4_de_agosto_2021.pdf

Tratamiento de la información y preparación de los datos

Tal como se expuso, la información y los datos con los que se trabaja son los registros de siniestros viales del último bienio de las provincias de Corrientes, Formosa y Chaco.

La unidad de análisis es el siniestro vial. Un siniestro de tránsito es un suceso que ocurre cuando un vehículo entra en contacto con otro vehículo, peatón, animal u otra obstrucción estacionaria, como un poste, un edificio, un árbol, entre otros, en la vía pública.

El objeto de estudio es la fatalidad en los siniestros. Un siniestro vial es tipificado como fatal cuando a partir de ese evento, una o más personas resultan fallecidas (víctimas fatales). Una víctima fatal por siniestro de tránsito es aquella persona que fallece de inmediato o dentro de los 30 días posteriores al hecho, como consecuencia de un traumatismo causado por el siniestro vial (se exceptúan los suicidios).

Figura 1. Tipificación de Siniestros Viales

Tipo de siniestro	Estado	Detalle
Siniestro Simple	Ileso	Persona sin traumatismo alguno
Siniestro con lesionados	Herido Leve/ Herido Grave	Leve: Persona con al menos un traumatismo que exige atención médica mínima o nula (como esguinces, hematomas, heridas superficiales y rasguños) Grave: Persona con al menos un traumatismo que exige la hospitalización durante al menos 24 horas o una atención especializada, como fracturas, conmoción, shock grave y laceraciones importantes.
Siniestro Fatal	Fallecido	Persona muere de inmediato o en un plazo de 24 hs. después del siniestro debido al traumatismo causado por el mismo

Los datos de los siniestros (datos registrados en SIGISVI) se relevan en el FEU (Formulario Estadístico Único). En este formulario se relevan los datos de todo siniestro acaecido en un territorio determinado y con una fecha específica. Los datos que releva tienen el status de datos censales, dado que cada jurisdicción toma al sistema como el único registro oficial.[4]

El FEU releva indicadores de las 3 poblaciones intervinientes en un siniestro vial: siniestros, vehículos y personas. Los siniestros son eventos únicos que tienen asociados vehículo o vehículos y persona o personas intervinientes. Cada población tiene sus atributos definidos que contribuyen a conformar la tipificación posterior del hecho.

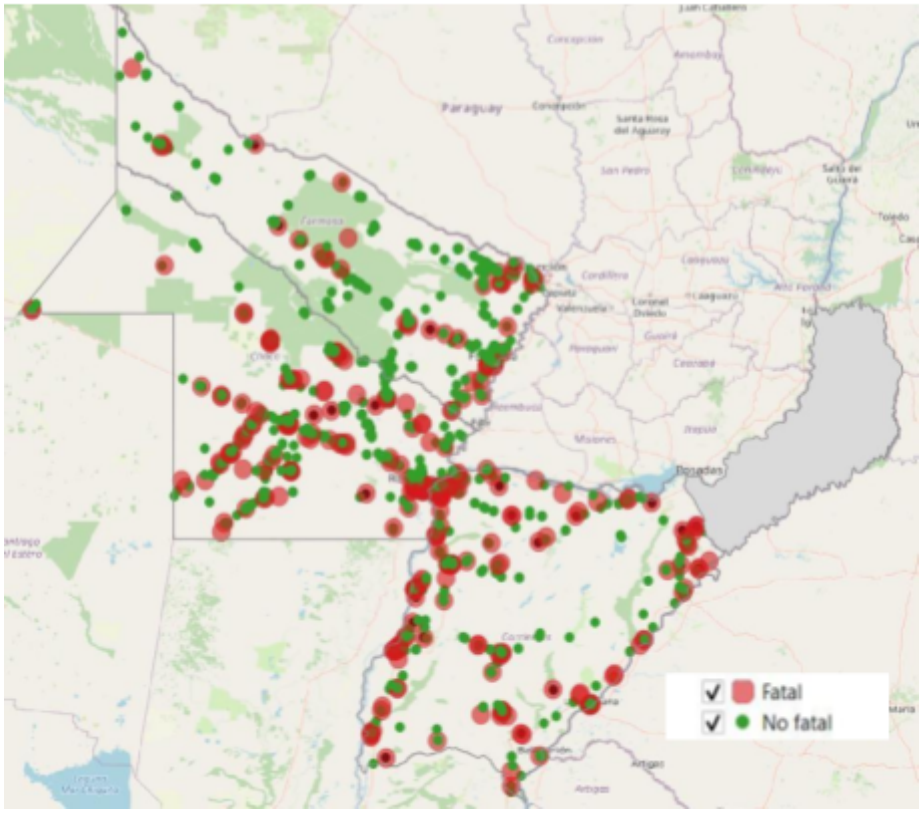
[4] Este atributo de los datos de SIGISVI es importante porque los resultados equivalen a todos los eventos en cada territorio, y no son resultado de ningún diseño muestral. Esto se torna fundamental para diagnósticos espacio temporales como también para la robustez de los estadísticos que se utilicen porque no es necesario poner atención en la significación estadística.

Figura 2. Poblaciones intervinientes en Siniestros Viales

Población	Siniestros	Vehículos	Personas
Indicadores de:	Cuándo y en qué contextos ocurren los siniestros viales. Estados de las vías donde ocurren	Detalle del parque automotor involucrado en siniestros	Perfil de víctimas fatales y heridos Principales grupos de riesgo de siniestros
Variables que contienen:	• Fecha y hora • Ubicación exacta • Categoría de siniestro • Tipo de siniestro • Tipo y estado de la vía • Estado de la calzada • Clima	• Tipos de vehículos involucrados en el siniestro • Datos de cada vehículo involucrado • Tipo de usuario	• Categoría de la víctima • Género • Edad • Condición de la víctima • Uso de elementos de seguridad

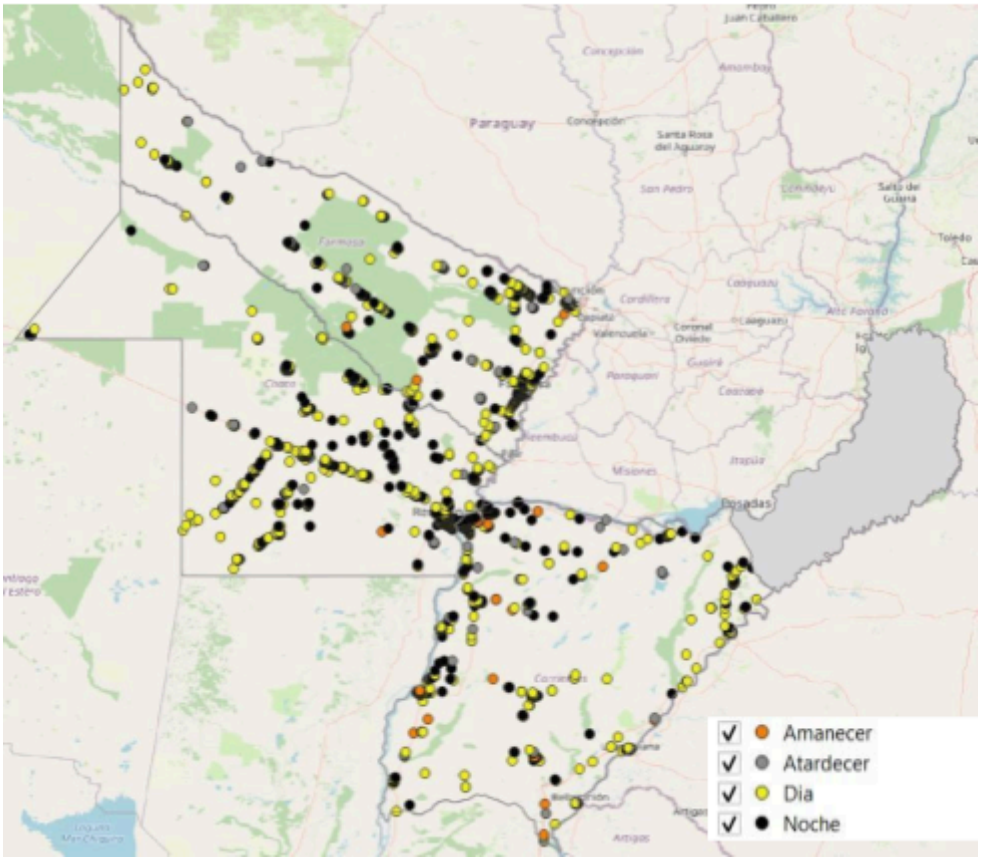
En este sentido, en el FEU se clasifican desde la ubicación geográfica exacta del siniestro, la hora, el tipo, hasta la cantidad y el tipo de vehículos intervinientes, como así también, la o las personas involucradas, con todos sus atributos.

Figura 3. Categoría de Siniestro NEA (SIGISVI - 2019 – 2020)



Las 3 poblaciones asociadas a un evento vial se vinculan mediante claves, como es usual en este tipo de diseños de sistemas de carga. Esto posibilita analizar cada población por separado como también cada siniestro con todos los elementos intervinientes.

Figura 4. Horario del Siniestro NEA (SIGISVI - 2019 – 2020)

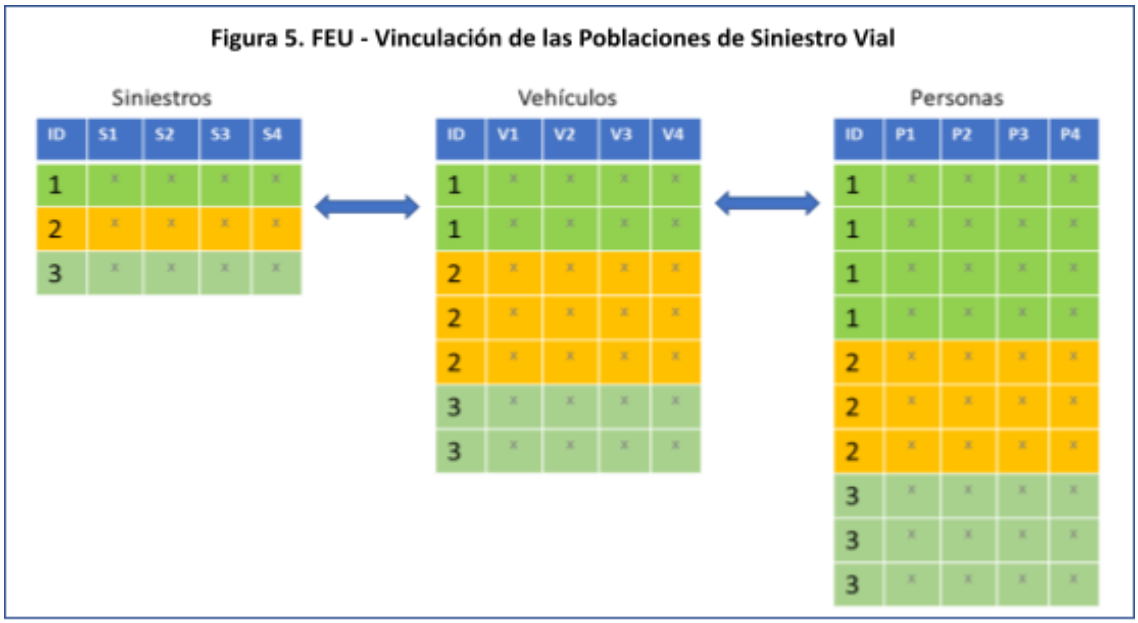


Usualmente, en las estadísticas de siniestralidad vial, los indicadores y variables de las poblaciones que conforman su universo conceptual y empírico, se vinculan en los análisis. De este modo, se puede obtener información sobre qué tipo de siniestros es el preponderante en una zona determinada, o en una época del año u hora del día; como también qué tipo de vehículos se asocian más a siniestros con lesionados, o si algún rango etario se vincula con algunos tipos de siniestros o vehículos.

Estas vinculaciones entre indicadores, además de alinearse a las normativas internacionales de registros de siniestros viales, muestran que los siniestros viales son hechos multicausales, donde intervienen y traccionan multiplicidad de factores. Un siniestro vial puede ser explicado, y parcialmente, por factores ambientales, estructurales, mecánicos, subjetivos, etc.

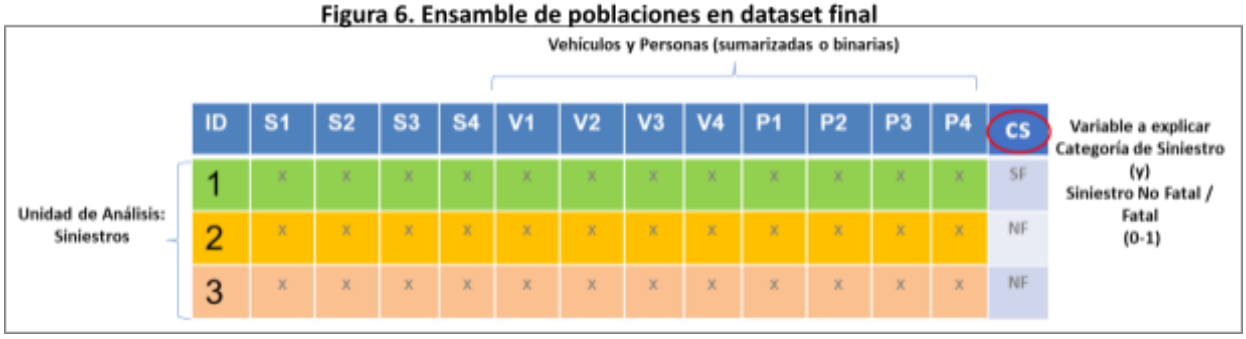
Por esta razón, que da cuenta de la complejidad de estos eventos, al momento de pensar un modelo explicativo de la fatalidad vial, es necesario contar con información asociada de las 3 poblaciones intervinientes.

Figura 5. FEU - Vinculación de las Poblaciones de Siniestro Vial



El proceso de preparación de datos vinculó las 3 poblaciones mediante sus claves, tomando al siniestro como unidad de análisis. Esto último se explica porque lo que buscamos caracterizar es un atributo de los siniestros. La variable objetivo o variable a explicar o *target*, es una variable que es un atributo del siniestro pero que se calcula a partir de un atributo de las personas. El atributo del siniestro (Fatal/ No fatal), deviene del estado de las personas intervinientes en el siniestro[5].

Figura 6. Ensamble de poblaciones en dataset final



Como los datasets tienen jerarquía distinta, al ser el siniestro la unidad de análisis, los indicadores de las personas y vehículos, se sumarizan, en caso de ser variables de intervalos, o se binarizan, en caso de ser variables categóricas. La construcción de la variable a explicar (Siniestro Fatal/ No fatal), se genera a partir del estado fallecido en el lugar del hecho o fallecido post hecho[6]. Al haber una persona (o más de una) con este status, el siniestro es clasificado como Fatal. De esta forma, queda construido el dataset con las variables de las 3 poblaciones para comenzar la exploración y el modelado.

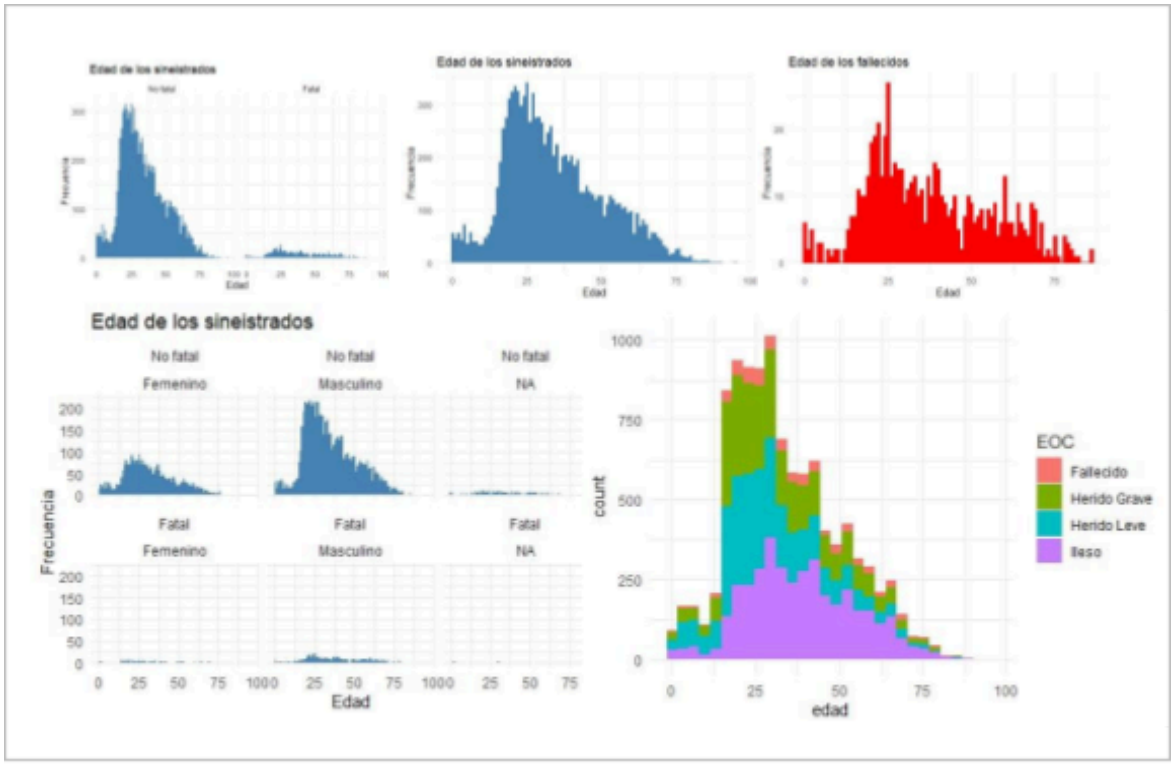
[5] Tal como se expuso en el cuadro N°1, el estado de un interviniente post. Siniestro puede ser ileso, lesionado (leve o grave) o fallecido.

[6] Los fallecidos en seguridad vial son los fallecidos en el lugar del hecho o los fallecidos "a 30 días", es decir, toda persona que haya fallecido por causa de un siniestro vial, desde la fecha del siniestro hasta 30 días. El SIGISVI permite hacer ese seguimiento y calcular este tipo de fallecidos.

Exploración y análisis de los datos

En el bienio 2019 – 2020, en NEA (-M)[7], se registraron 5.514 siniestros viales[8]. De éstos, un 5.9% fueron siniestros fatales. Intervinieron 9.975 vehículos y 12.551 personas, de las cuales fallecieron 714 (5.7%). 6 de cada 10 personas que se vieron involucradas en estos siniestros, tuvieron algún tipo de lesión. [7]

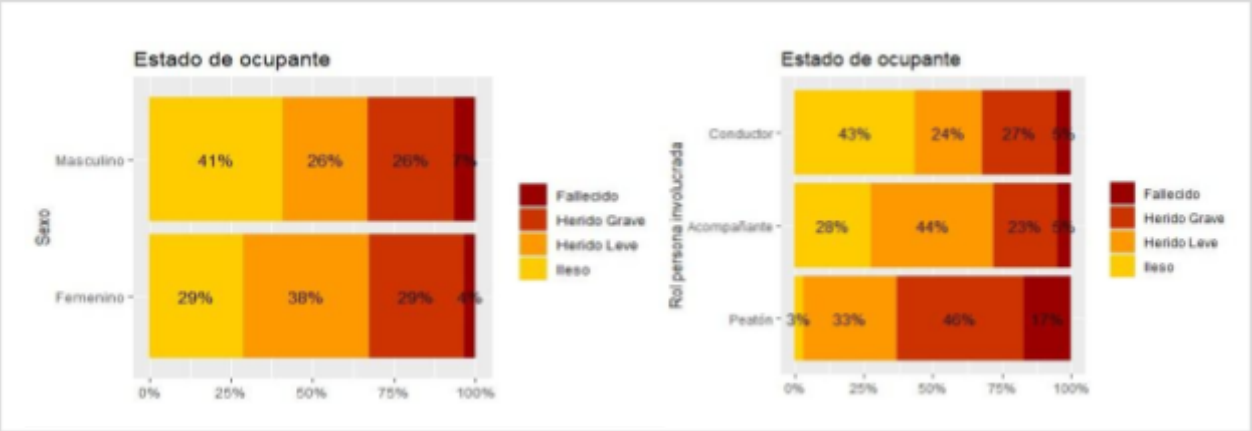
Figura 7. Descripción de las personas involucradas (siniestradas y fallecidas)



La incidencia de género en los siniestros es de 70/30 a favor de los hombres, aumentando a 80/20 en cuento a los fallecimientos. Huelga decir, y ésto se refleja en las estadísticas anuales, que estamos ante un evento predominantemente masculino. El promedio de edad de los siniestrados es de 34 años y de 37 de los fallecidos. El rango de 15 a 35 años, representa el 50% de los fallecidos. Al interior de cada género, los hombres se diferencian en sus niveles sin lesión y en fallecidos. Luego, al interior de los roles, los peatones son los que más intervienen en siniestros donde hay fallecidos.

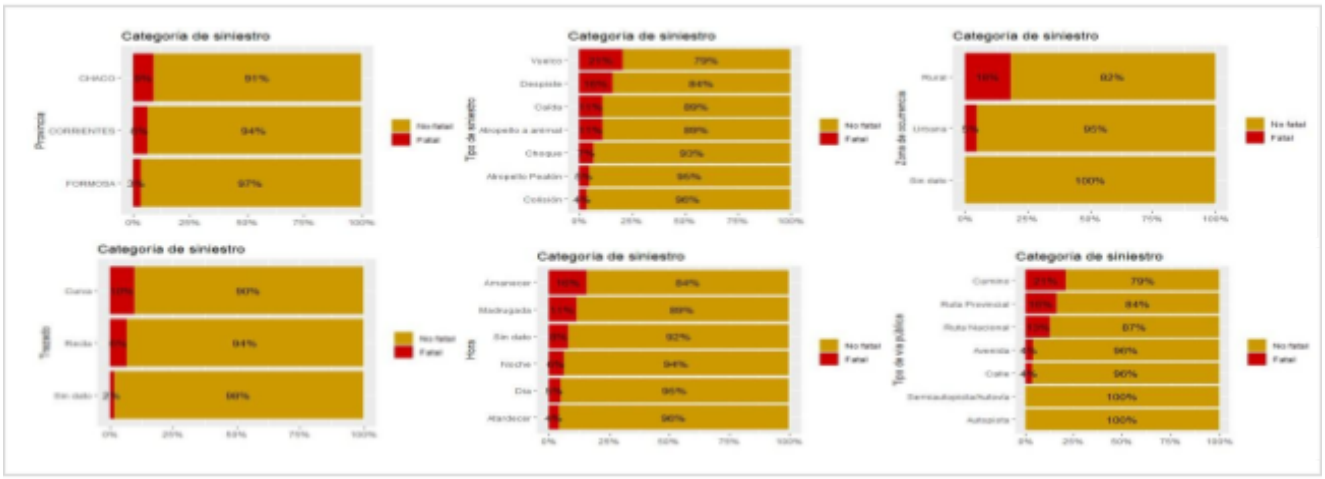
[7] Siglas que explican la ausencia de la provincia de Misiones dentro de lo que se conoce como NEA.
860% ocurrieron en 2020 y 40% en 2019, cosa que es curiosa, dado la baja de circulación que hubo en casi todo 2020 a causa del ASPO

Figura 8. Estado de ocupante según Sexo y Rol



Los siniestros presentan características puntales: Chaco (9%) es la provincia con más incidencia de siniestros fatales, seguido por Corrientes (6%) y Formosa, con un tercio respecto de la primera (3%). La fatalidad es mayor en vuelcos y despistes; en curvas y en ámbitos rurales. Asociado a ésto, se observa que el tipo de vía no es el medio urbano, sino "caminos" o rutas nacionales o provinciales.

Figura 9. Categoría de siniestro según atributos de siniestros



En términos temporales (no se percibe un patrón estacional en los meses del año), la fatalidad se da más en horas del amanecer y madrugada.

El siguiente cuadro muestra el porcentaje de participación de cada tipo de vehículo en siniestros fatales, es decir, qué participación tuvieron cada uno en la fatalidad.[8] Nuevamente hay indicios de ciertos patrones en la fatalidad de los sinestros en esta región. Tres de los cuatro primeros vehículos involucrados, son vehículos no particulares, es decir, de uso privado; son vehículos de uso profesional o laboral (Maquinaria, Transportes de Carga y Transporte de pasajeros). Esto, sumado a la incidencia por el tipo de vía (rutas nacional y provincial, y "caminos"), nos empieza a dar una fotografía de factores propios de la zona (en términos estructurales y productivos) que se asocian con la fatalidad vial.

[8] Es importante comprender que lo que se intenta caracterizar no es la participación en la siniestralidad, sino la incidencia en la fatalidad (participación en siniestros que han resultado fatales). Por lo tanto, lo que buscamos en el análisis de los vehículos involucrados, es la proporción de su participación en siniestros viales fatales, es decir, de todos los siniestros donde han participado, qué proporción fue fatal.

Figura 10. Porcentaje de fatalidad en el tipo de vehículo involucrado	
Maquinaria	16,7%
Transporte de carga	13,4%
Otros	12,5%
Transporte de pasajeros	11,6%
Bicicleta	7,8%
Peatón	6,0%
Motocicleta	5,1%
Camioneta/ Utilitario	4,7%
Automóvil	3,6%

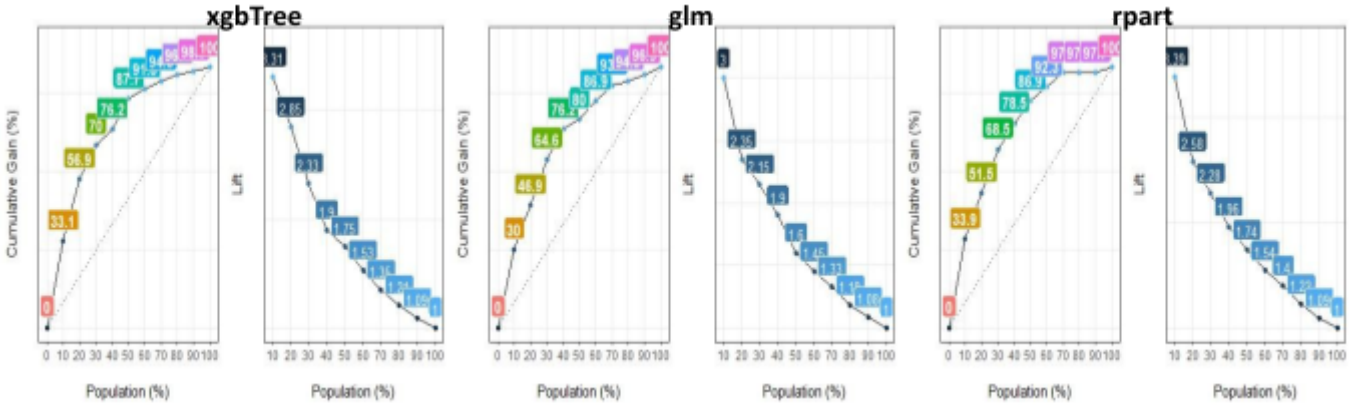
Cuando hablamos de los siniestros viales como eventos multicausales o multifactoriales, estamos hablando de cómo interactúan diversos factores que contribuyen a explicar estos eventos. Esta zona parece tener atributos y características particulares relacionadas no sólo a los ambientes naturales, sino a factores sociales y productivos. Una hipótesis plausible de ser contrastada, podría ser que la siniestralidad en general, y la fatalidad en particular, difiere significativamente (cuantitativa y cualitativamente) a medida que pasamos de una región a otra.

Modelización de la fatalidad vial

Para modelizar la siniestralidad fatal, se pusieron en práctica 3 algoritmos para decidir cuál de ellos desarrollaba la mejor solución ante este problema de investigación y este set de datos. Todos son para procesos basados en técnicas de dependencia o supervisadas. Los algoritmos usados fueron "glm", la función del modelo lineal generaliza, es su opción binomial (logit), "rpart", árbol de regresión y "XGBoost", árboles de decisión basados en el principio de *boosting.*[9]

Para ver el rendimiento de los modelos[10], analizamos las curvas de Ganancia (Gain curve). Como se puede ver, son muy parejas, siendo la curva de xgbTree la que tiene mejor performance (ganancia de 70% en los 30 p./ 87.7% de acumulación de positivos en 50 p. y lift de 1.75).

Figura 10. Curva de Ganancia de los Modelos



[9] La idea principal de esta parte del trabajo no es profundizar en los algoritmos, sino en los resultados. La elección de estos tres tuvo que ver con probar algoritmos distintos o con lógicas de clasificación distintas.

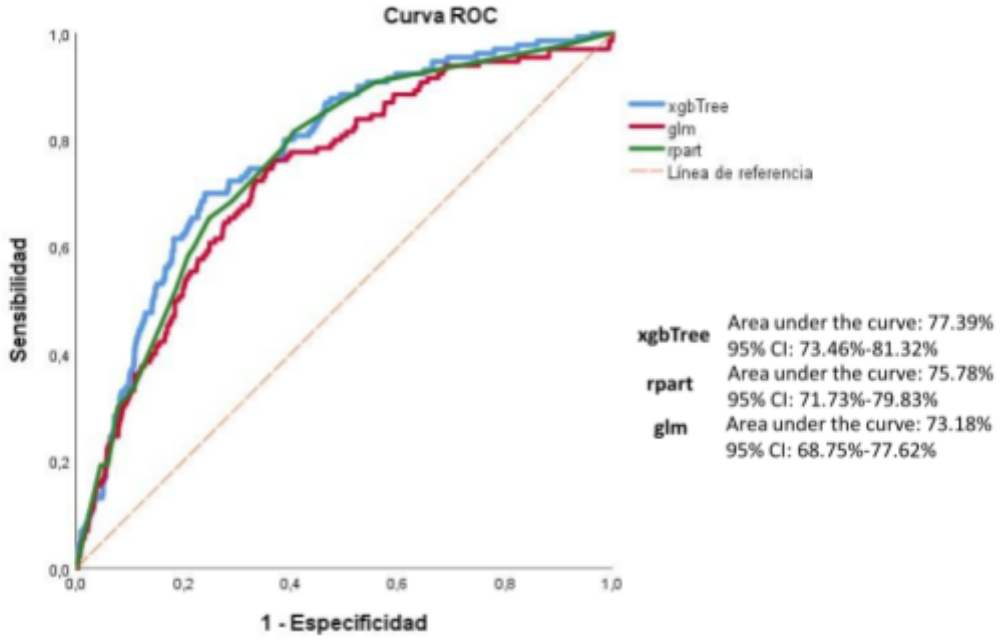
[10] Previamente se dividió el dataset en train/ test (60/40). El análisis posterior de resultados corresponde al dataset de testing.

xgbTree

Population	Gain	Lift	Score.Point
1	10	33.08	3.31	0.76873952
2	20	56.92	2.85	0.63756694
3	30	70.00	2.33	0.51295121
4	40	76.15	1.90	0.42125511
5	50	87.69	1.75	0.33519310
6	60	91.54	1.53	0.27165674
7	70	94.62	1.35	0.21468726
8	80	96.92	1.21	0.16127484
9	90	98.46	1.09	0.10764062
10	100	100.00	1.00	0.03498399

Otra manera de analizar el ajuste de los modelos es comprar su rendimiento mediante la curva ROC, analizando su AUC (Área bajo la curva). La lógica es similar a la curva de Ganancia, en cuanto al análisis de la buena o mala clasificación. La clasificación indica que el modelo es aceptable.

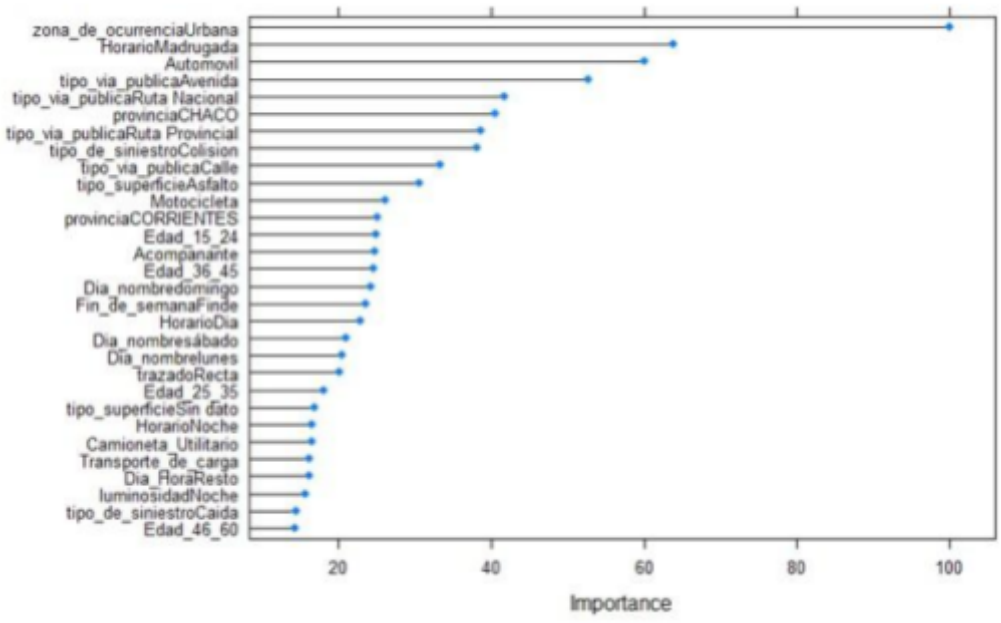
Figura 11. Curva ROC de los Modelos



El algoritmo xgbTree ofrece un gráfico que muestra el aporte (gráfico de importancia) de cada variable al ajuste final del modelo. En este caso, se muestran 30 variables que más aportan a la clasificación de la variable objetivo.[11]

[11] Si bien no deja de ser una síntesis útil, en este tipo de algoritmos basados en el boosteo (*boosting*, o combinación los resultados de varios clasificadores débiles para obtener un clasificador robusto), hay que interpretarlos con prudencia. En este caso, la partición del este tipo de árboles, también trabaja con el peso de cada categoría en cuanto a lo que aporta para la partición de los nodos, es decir, que una categoría con mucha "presencia" (peso), aporte (está más presente) en los nodos, no siendo necesariamente útil para la clasificación de la variable objetivo. En este caso, se puede ver que la categoría "urbana" tiene la mayor "importancia", pero esa importancia es su presencia en los datos y en la partición.

Figura 11. Plot de Importancia de xgbTree



Habiendo resultado aceptable el ajuste final, y como el objetivo principal de este modelo es caracterizar[12] la fatalidad vial, analizamos la distribución de los predictores seleccionados. Como expusimos al principio, las variables seleccionadas (42) para modelizar, corresponden a distintas dimensiones del corpus conceptual de la seguridad vial: estructurales, ambientales, mecánicas, productivas y subjetivas. Para ordenar la lectura de los resultados, dada la cantidad de los mismos, se "ordenaron" según estas dimensiones.

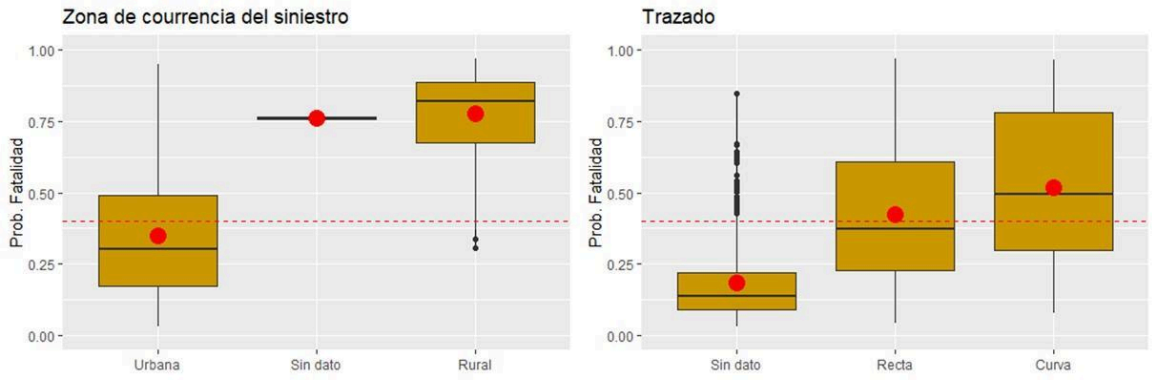
Lo primero que observamos es la preminencia no urbana de la fatalidad. Los gráficos[13] son muy claros en cuanto al promedio de la probabilidad de fatalidad y la agrupación de casos en "rural" como zona de ocurrencia del siniestro[14].

[12] Estos modelos cumplen todos los requisitos para ser modelos predictivos, pero no es el objetivo de este trabajo. Entendemos que lo predictivo en un modelo de estas características, se da al momento de poner en producción al modelo, clasificando datos "nuevos" o datos no supervisados. Esa instancia no es posible en este momento del proceso, por eso optamos por la caracterización de los eventos (siniestros fatales).

[13] Los gráficos tienen el promedio de probabilidad de fatalidad (punto rojo) y los boxplot de la distribución de los casos de las categorías de la variable predictora (la línea negra de cada caja es la media de la distribución). La línea roja punteada, es el promedio de la probabilidad observada del socre.

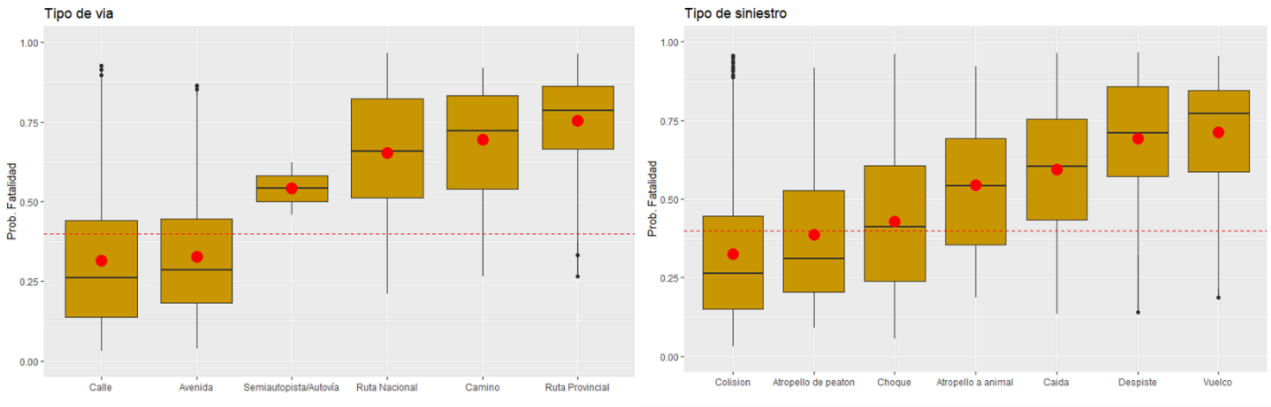
[14] Recordemos que lo que buscamos caracterizar es la fatalidad, no la siniestralidad (91% urbana en NEA). Evidentemente, la siniestralidad es "más fatal" (aunque inmensamente menor en incidencia) en zonas rurales (casi 5 veces más de fatalidad) que urbanas (más siniestralidad). Es una relación entre eventos posibles (fatales) dentro de eventos observables (siniestros).

Figura 12. Probabilidad de fatalidad según zona de ocurrencia y trazado



Relacionado con esto, los tipos de vías con mayor probabilidad de fatalidad son las preeminentes en medios rurales: Rutas nacionales y provinciales y "caminos" rurales. En tipo de siniestro, es muy clara la probabilidad de fatalidad si se dan vuelcos, despistes, caídas o atropellos[15]. La curva es la zona del trazado que se evidencia como más fatal.

Figura 13. Probabilidad de fatalidad según Tipo de vía y siniestro

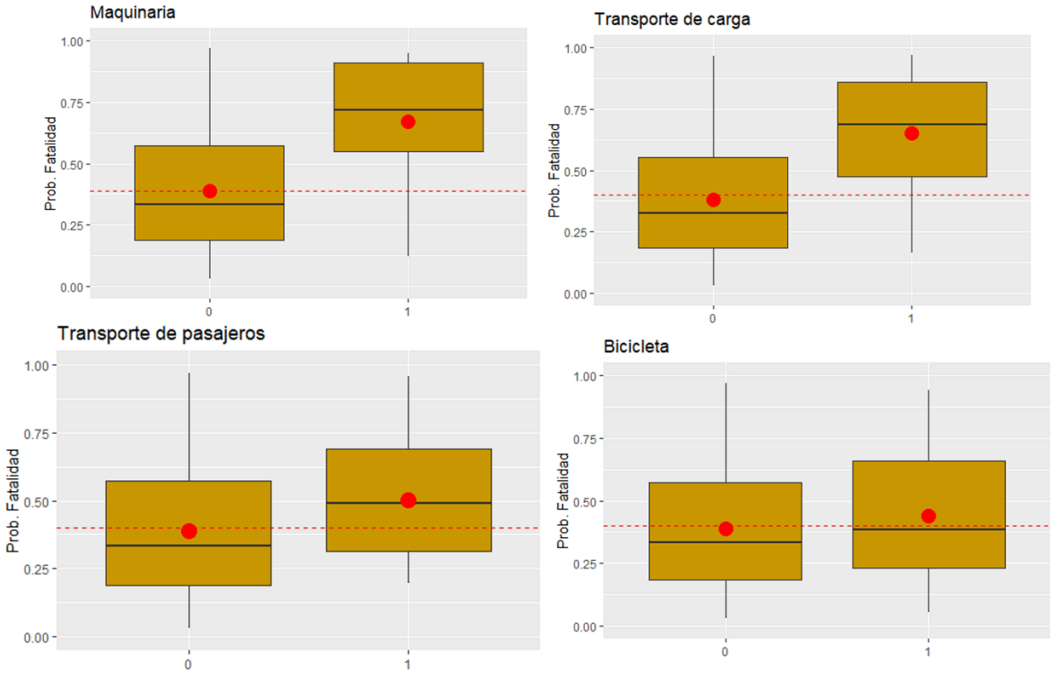


Una parte importante del análisis de la fatalidad vial, es el papel de los vehículos (dimensión mecánica). En este caso, confirmamos la tendencia de la "ruralidad" en la fatalidad en esta región, porque los tipos de vehículos con una probabilidad de fatalidad significativa son: Maquinarias, Transportes de carga, Transporte de pasajeros. Bicicleta es el único tipo de vehículo en que arriesgamos como "urbana" que aparece como significativo en su probabilidad de fatalidad. Los restantes tipos de vehículos (automóvil, motocicletas, camionetas, etc.) no aportaron de forma significativa a la probabilidad de intervenir en un siniestro fatal (similares a la observada o un poco por encima, como motocicletas y automóviles).

También parece haber un patrón temporal en la caracterización de la fatalidad. Principalmente la fatalidad se da los fines de semana, mayormente de noche. De los días, específicamente el domingo. El horario que presenta mayor diferencia a favor de la fatalidad es el amanecer y la madrugada.

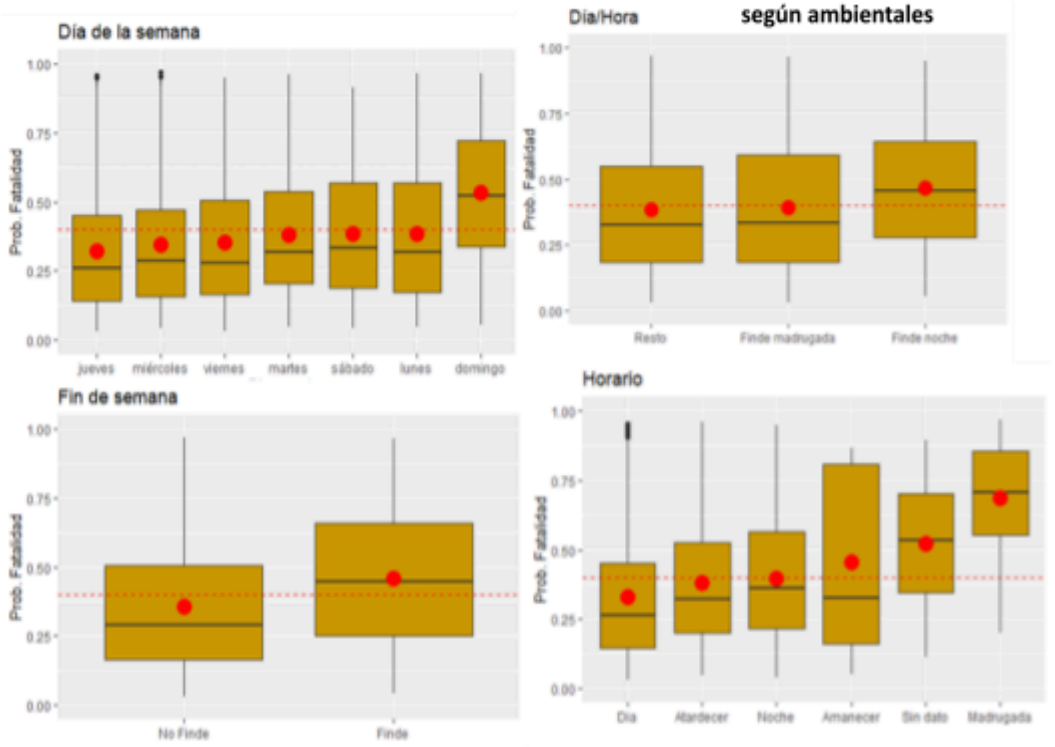
[15] Como se observa, colisión es el tipo de siniestro con mayor incidencia neta, pero el de menor fatalidad relativa.

Figura 14. Probabilidad de fatalidad según atributos de los vehículos



Tenemos evidencia como para ensayar la hermenéutica de estos resultados. La fatalidad vial en NEA (-M), por lo menos en los datos de siniestros viales del bienio 2019 – 2020, parece estar asociada, en mayor medida, al ámbito rural más que al urbano, a los vehículos de porte y de uso profesional, más que a los vehículos particulares. Esto tiene sentido cuando observamos la traza vial de la región, la misma está atravesada por 12 rutas nacionales con un importante TMDA.[17]

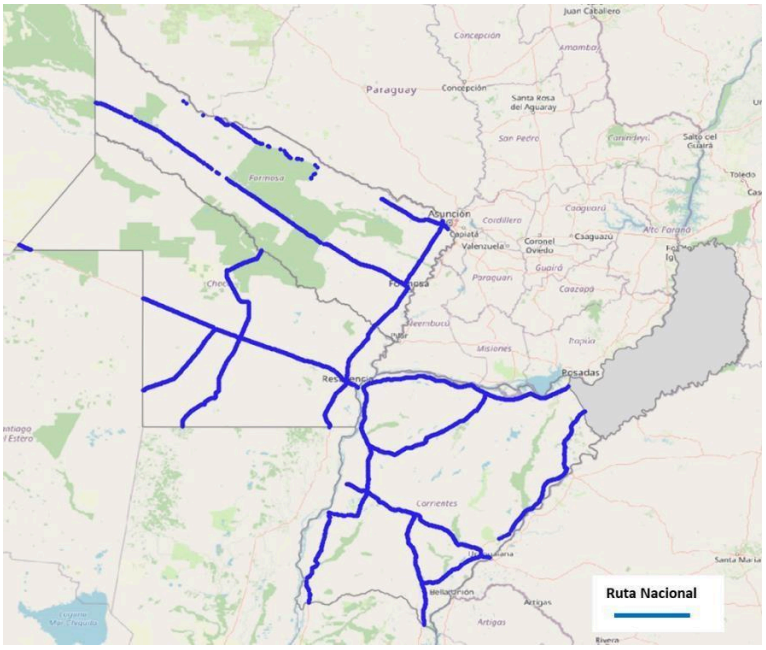
Figura 15. Probabilidad de fatalidad según ambientales



[17] TDMA: Tránsito Diario Medio Anual

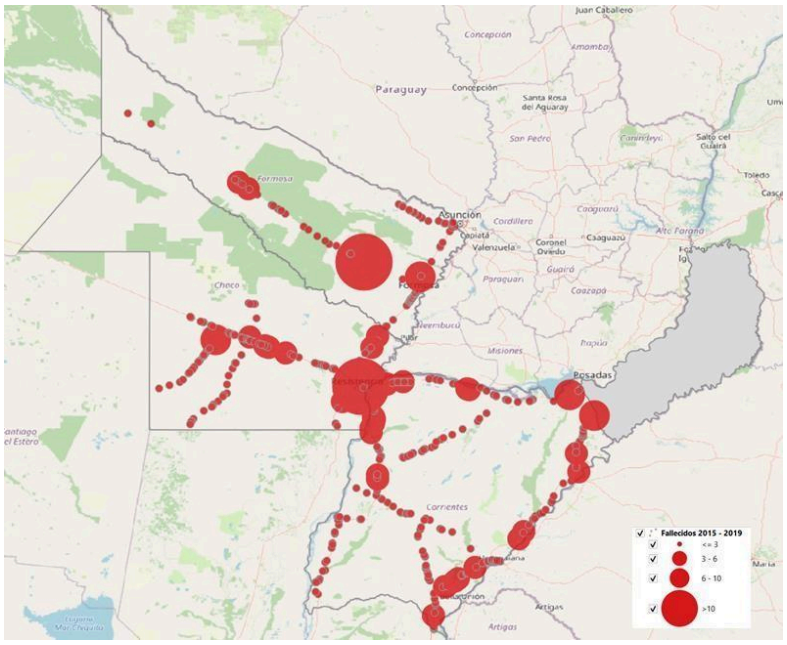
Por lo tanto, estaríamos ante un nuevo patrón de la fatalidad, la asociada al ámbito productivo. La fatalidad vial está asociada a vehículos en actividad en medio de un proceso productivo (trabajo), ya sea maquinaria (suponemos, agrícola) o transporte (de carga o pasajeros).

Figura 16. Rutas nacionales NEA (-M)



Si a esto le sumamos que los siniestros devenidos en fatales se dan en hora de madrugada o amanecer y el tipo de siniestro fatales son vuelcos o despistes, muy probablemente estemos ante siniestros fatales por causa de exceso de tiempo de conducción, es decir por cansancio. Tampoco hay que descartar el factor de visibilidad y el estado de las rutas como causantes de este tipo de siniestros (dimensión estructural).

Figura 17. Fallecidos en Rutas nacionales NEA (-M) (SIAT -2015 -2019)



Consideraciones finales

Hemos explicado y detallado los datos y los resultados del modelo de caracterización de fatalidad en NEA y los asumimos como lógicos y consistentes con los datos y los objetivos planteados. Estimamos que se logró un modelo de caracterización de la fatalidad en esta región que aporta conocimiento sustancial hacia esta problemática. En este cierre, no vamos a repetir los patrones anteriormente detallados y explicado sobre la fatalidad en esta región, la idea es no redundar. Ahora existen indicios por dónde se podría comenzar a trabajar en un área temática (seguridad vial) que no se caracteriza por tener abundancia de estudios y de datos al respecto.

Lo que se podría agregar es lo siguiente:

Primero, lamentamos no haber podido aprovechar al máximo un sistema de gestión de datos como el SIGISVI. Esto se debió no solo por la ausencia de la provincia de Misiones (cuestión que ya explicamos), sino también a que no pudimos contar con indicadores que, estimamos, hubiesen sido de gran utilidad para explicar la fatalidad vial.[16] Esta situación se debe a la cantidad de variables "sin dato" que posee el sistema. No deja de ser un llamado de atención para los encargados de la carga de datos, pero también para reflexionar sobre el diseño del FEU.

Segundo, desde el comienzo de este trabajo, explicamos que el objetivo era la caracterización y no la predicción, por no tener la chance práctica de clasificar datos nuevos. Sin embargo, creemos que este tipo de procesos y de información resultante, pueden aportar a predecir (o prevenir, traduciéndolo a políticas públicas) la fatalidad vial. Este tipo de procesos deberían (aprovechando los datos que se registran) intervenir en algún proceso regular dentro de la ANSV.[17] La ANSV otorga licencias de conducir, intervienen en la reglamentación de vehículos profesionales, es actor interviniente en legislación vial, etc. Este tipo de trabaja con datos, debería complementar y contribuir en alguna de estas tareas, con el fin de mejorar, reforzar la seguridad vial y prevenir siniestros y víctimas.

[16] El sistema cuenta con indicadores como utilización de elementos de protección (caso, cinturón, etc), seguro del vehículo, existencia de señalizaciones, estado de la ruta/calle, presencia de luz artificial, VTV del vehículo, alcoholemia del conductor, color del vehículo, entre otros. Ninguno de éstos se pudo utilizar en el modelo para caracterizar la fatalidad, desaprovechando una fortaleza que tienen estas soluciones de ML.

[17] No olvidar que uno de los objetivos principales de la ANSV es la reducción de la siniestralidad vial en general y las víctimas fatales viales en particular.

Referencias y material de consulta

- OPS. 2016. Sistema de datos Manual de seguridad vial para decisores y profesionales. http://whqlibdoc.who.int/cgibin/repository.pl?url=/publications/2008/9789275316283_spa.pdf
- ANSV. Observatorio Vial. 2019. Glosario de términos y definiciones relativas a la seguridad vial. https://www.argentina.gob.ar/sites/default/files/glosario_de_terminos_seguridad_vial.pdf
- ANSV. Observatorio Vial. 2019. Anuario 2019. https://www.argentina.gob.ar/sites/default/files/2018/12/ansv_ov_anuario_estadistico_2019_final.pdf
- ANSV. Observatorio Vial. 2020. Informe Anual 2020. Datos preliminares. https://www.argentina.gob.ar/sites/default/files/2018/12/ansv_ov_informe_anual_2020_al_4_de_agosto_2021.pdf
- Introduction to Boosted Trees. https://xgboost.readthedocs.io/en/latest/tutorials/model.html
- IRTAD 2019. International Transport Forum. Road Safety Annuel Report – Argentina. https://www.itf-oecd.org/sites/default/files/argentina-road-safety.pdf

Programa Cambio Rural (CR) – SAGyP: Identificación del perfil de beneficiarios con IA

De Patricia Perrone (CR), Marianela Pi, Constanza Guerrini

Patricia Perrone (CR), Marianela Pi, Constanza Guerrini

Programa Cambio Rural (CR) – SAGyP: Identificación del perfil de beneficiarios con IA

Objetivos Generales

El objetivo general de esta investigación es caracterizar a los beneficiarios/as del Programa Cambio Rural perteneciente a la Secretaría de Agricultura, Ganadería y Pesca de la Nación.
Cambio Rural es una política pública que busca, a través de la asistencia técnica, promover y facilitar la intensificación y reconversión productiva, como un medio para mejorar la situación productiva y socioeconómica de los pequeños y medianos productores rurales y propender al desarrollo agroindustrial en todo el territorio nacional, impulsando el aprendizaje grupal.

Objetivos Específicos

Obtener un *dataset* limpio, determinar las variables a utilizar para la caracterización y realizar la clusterización. Realizar análisis estadísticos.

Antecedentes

El Programa actual, es la reestructuración del Programa Federal de Reconversión Productiva que tuvo lugar en septiembre de 2017 y que fue relanzado como Programa Cambio Rural (CR) a través de la Resolución E 249/2017, junto con un nuevo Manual Operativo. En ese mismo documento se crea el Registro de Integrantes de Grupos Cambio Rural y se desarrolla un Sistema de Gestión para sistematizar la información de este Registro. La decisión de reestructuración se tomó debido a que, a pesar de los esfuerzos y recursos invertidos, los resultados obtenidos no fueron suficientes para posicionar a la pequeña y mediana empresa rural en los niveles óptimos y necesarios de eficiencia productiva que les permitieran enfrentar exitosamente las fluctuaciones económicas y climáticas.

Como antecedente, se tomó la Resolución N° 227 de fecha 4 de mayo de 1993 de la Secretaría de Agricultura, Ganadería y Pesca del entonces Ministerio de Economía y Obras y Servicios Públicos, que creó el Programa Federal de Reconversión Productiva con el propósito de promover y facilitar la intensificación y reconversión productiva de la pequeña y mediana empresa rural. La creación del programa se llevó a cabo en el contexto histórico de los años 90 y ante la crisis económica reinante, y se solicitaron propuestas al sector, una de las cuales, elaborada por el INTA, se convirtió en el Programa Federal de Reconversión Productiva. En su origen, el Programa otorgó gran importancia al trabajo coordinado con las Provincias y las entidades del sector, incluidas las intermedias, para posibilitar y potenciar la asistencia técnica, el acceso al crédito y el intercambio tecnológico necesario para una mayor eficiencia y diversificación productiva que, junto al esfuerzo asociativo, generaran economías competitivas.

Políticas similares

https://agriculture.ec.europa.eu/common-agricultural-policy/rural-development_es

Actividades y metodología

a) Análisis de las variables relevadas
b) Selección de las variables a trabajar
c) Limpieza de datos

d) Homogeneización de datos
e) Generación del *dataset*
f) Análisis estadístico de las variables

Factibilidad

Es posible realizar un análisis de *clustering* utilizando variables categóricas, pero es importante utilizar una técnica adecuada para manejarlas correctamente. En este caso, utilizaremos una técnica de codificación de variables categóricas, como la codificación "*One-HotEncoder*", la cual implica la creación de una variable binaria por cada posible valor de la variable categórica original.

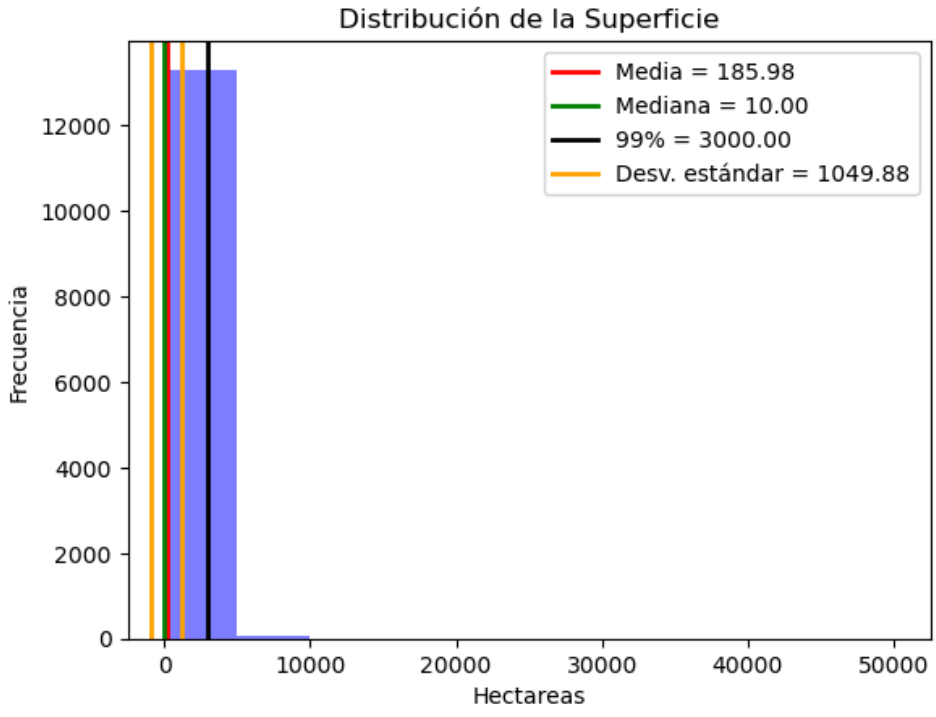
Sin embargo, es importante tener en cuenta que trabajar con demasiadas variables categóricas puede generar un gran número de variables binarias en el análisis, lo que puede afectar la calidad del modelo resultante. En este caso, nos enfocaremos en trabajar con un máximo de 4 variables categóricas, ya que abordar más variables requiere un mayor conocimiento y experiencia en la selección de técnicas de clustering adecuadas y en la interpretación de los resultados obtenidos, lo cual puede exceder el plazo de dos meses disponible para este trabajo.

Proceso

Elección de variables

Se realizó un análisis de la tabla extraída del “Sistema de Registro de Integrantes de Grupos Cambio Rural” y se identificaron varios problemas en las variables de edad y superficie de la explotación, principalmente por errores de carga.

En el caso de la superficie, el 99% de los datos se encuentran por debajo de las 3000 Has. El Resto puede ser un error de carga, tierras comunitarias o casos muy extremos. Podría tomarse sólo el rango de 0 a 3000 Has, siendo que existen actividades que justifican un valor de 0 Has. Podemos valernos de la observación de los percentiles para armar rangos, pero no describirían las distintas actividades que realizan los productores generando un desvío de lo que se busca evidenciar.




```
Desviación estándar: 1049.882016
Varianza: 1102252.247966
Rango intercuartil: 76.0
Coeficiente de variación: 564.508225
```

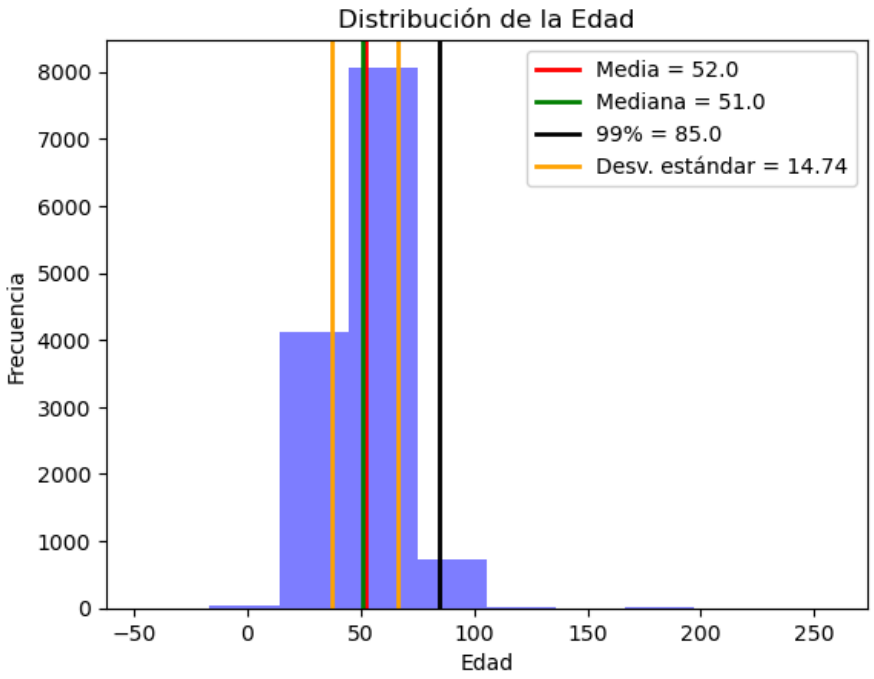
Se intentó resolver el problema de los datos de la superficie utilizando una transformación logarítmica -muy utilizada en estos casos- para reducir el rango de datos, pero no se pudo utilizar por ser una función que no está definida en el cero (dentro de los beneficiarios hay quienes tienen actividades que no requieren de la explotación de tierras por lo que la superficie utilizada tiene un valor genuino de cero). La posibilidad de sustituir por el valor de la media no nos pareció adecuada en estas circunstancias. Otro tema del cual ocuparse son los NaN (valores nulos), por falta de carga.
Se pensó como alternativa la utilización de una función partida para el uso del logaritmo, donde la función es:

$$\begin{cases} \mathbf{0 \text{ para } x=0;} \\ \mathbf{\textit{Null} \text{ para } x=NaN \text{ y}} \\ \mathbf{\textit{Log}(x) \text{ para } x>0.} \end{cases}$$

En este caso tendríamos además una "sobrecarga" de ceros, por los ceros genuinos y los provocados por el *Log*(1), generando datos falsos. Por esta razón se decidió excluir la variable superficie.

Con respecto a la edad, se hizo un estudio estadístico para analizar la importancia de los errores de carga y tomar decisiones con respecto a su utilización.
Por lo analizado hay evidencia de errores de carga en la fecha de nacimiento y por consiguiente en el cálculo de la edad. Por un lado, no puede haber menores de 18 ni productores que aún no nacieron, y ciertamente no resulta factible tener productores de 250 años como surge en algún caso. Del análisis de los percentiles resulta que el 99% de la población CR está por debajo de los 85 años. Se pueden tomar como *outlayers* los valores menores a 18 y mayores a 86.

Otra opción es utilizar rangos etarios, de hecho, ya existe un rango etario diseñado en las encuestas en profundidad que se hacen a los grupos y es el que utilizaremos en lugar de Edad para la clasificación.




```
Desviación estándar: 14.737238
Varianza: 217.186170
Rango intercuartil: 20.0
Coeficiente de variación: 28.332787
```

Esto nos resuelve, también, los datos espurios de edades menores de 18 o mayores de 100 que son de seguro errores de carga.

Por todo esto se decidió trabajar sólo con las siguientes variables categóricas: rango etario, educación, ingresos (% de ingresos mensuales que aporta el emprendimiento acompañado por CR) y situación AFIP.

Resumen estadístico de las variables seleccionadas

	Rango_etario	Educación	Ingresos	SituacionAFIP
count	13391	13391	13391	13391
unique	4	6	3	17
top	Más de 55	Secundario	Más del 50%	Responsable Inscripto
freq	5326	5510	5808	3909

El valor "top" en cada variable muestra la categoría más frecuente en esa variable, lo que indica que hay 5,326 registros con la categoría "Más de 55" en la variable "Rango_etario", 5,510 registros con la categoría "Secundario" en la variable "Educación", 5,808 registros con la categoría "Más del 50%" en la variable "Ingresos" y 3,909 registros con la categoría "Responsable Inscripto" en la variable "SituacionAFIP".

Construcción del Modelo

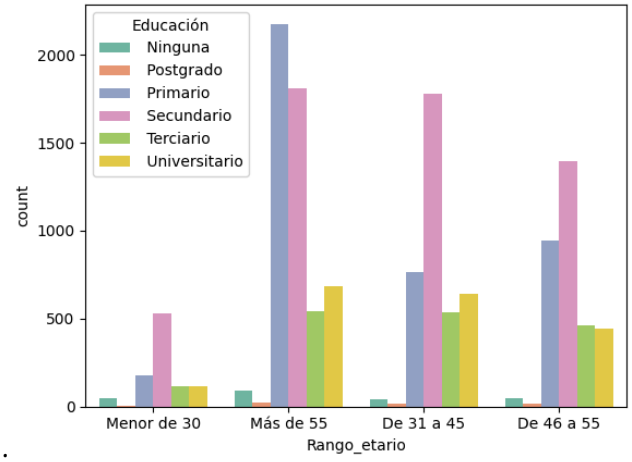
Nuestro objetivo en este cuaderno es sólo mostrar el algoritmo K-Modes, omitiendo en esta ocasión el EDA y pasaremos directamente a la construcción del modelo.

Construcción del modelo - Convertimos las variables categóricas en numéricas y las agregamos a la base original.

	Rango_etario	Educación	Ingresos	SituacionAFIP	Rango_etarioNum	EducacionNum	IngresosNum	SituacionAFIPNum
0	Menor de 30	Ninguna	Menos del 30%	No declarado	0	0	0	0
1	Menor de 30	Ninguna	Más del 50%	Responsable Inscripto	0	0	2	3
2	Menor de 30	Ninguna	Más del 50%	Responsable Inscripto	0	0	2	3
3	Menor de 30	Ninguna	Más del 50%	Monotributo Cat B	0	0	2	6
4	Menor de 30	Ninguna	Más del 50%	Monotributo Cat C	0	0	2	7
...	...	...	...	...	...	...	...	...
13386	Más de 55	Universitario	Menos del 30%	Responsable Inscripto	3	4	0	3
13387	Más de 55	Universitario	Más del 50%	Responsable Inscripto	3	4	2	3
13388	Más de 55	Universitario	Entre 30% y 50%	Responsable Inscripto	3	4	1	3
13389	Más de 55	Universitario	Entre 30% y 50%	Responsable Inscripto	3	4	1	3
13390	Más de 55	Universitario	Entre 30% y 50%	Responsable Inscripto	3	4	1	3

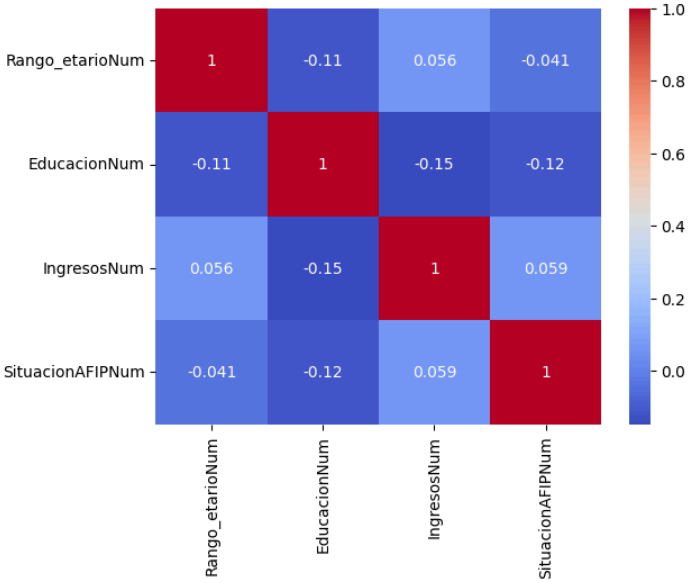
13391 rows × 8 columns

Aprovechamos para analizar la relación de algunas variables como edad y educación.
En este gráfico, se puede observar que en el rango etario 'Más de 55 años', hay una mayor proporción de individuos que sólo han completado la educación primaria, marcando una gran diferencia con el resto de los grupos etarios.



.

Generamos una matriz de correlación para ver cómo están vinculadas las variables elegidas.



Utilización de los métodos para clusterizar

K-Modes con Inicialización "CAO" y K-Modes con Inicialización "Huang"

K-Modes con Inicialización "Huang"

Aplicamos el método y nos tira la siguiente corrida

```
Initialization method and algorithm are deterministic. Setting n_init to 1.
Init: initializing centroids
Init: initializing clusters
Starting iterations...
...
Run 1, iteration: 1/100, moves: 0, cost: 25075.0
```

```
K-Modes con Inicialización "Huang"

Init: initializing centroids
Init: initializing clusters
Starting iterations...
...
Run 1, iteration: 1/100, moves: 0, cost: 25578.0
...
Run 2, iteration: 1/100, moves: 0, cost: 26049.0
...
Run 3, iteration: 1/100, moves: 2198, cost: 25120.0
...
Run 4, iteration: 1/100, moves: 1498, cost: 25010.0
...
Run 5, iteration: 1/100, moves: 0, cost: 26043.0
Best run was number 4
```

En cada corrida del algoritmo se registra el número de iteración actual, el número de movimientos realizados, el costo actual y el número de la ejecución. Se puede observar que en la ejecución número 4 se logró el menor costo, lo que indica que esa es la mejor solución encontrada por el algoritmo.

Se realizan corridas similares con Inicialización "CAO" con resultados similares que compararemos luego.

Resultados

Al comparar las corridas de K-Modes por los dos métodos de inicialización, encontramos los siguientes resultados:

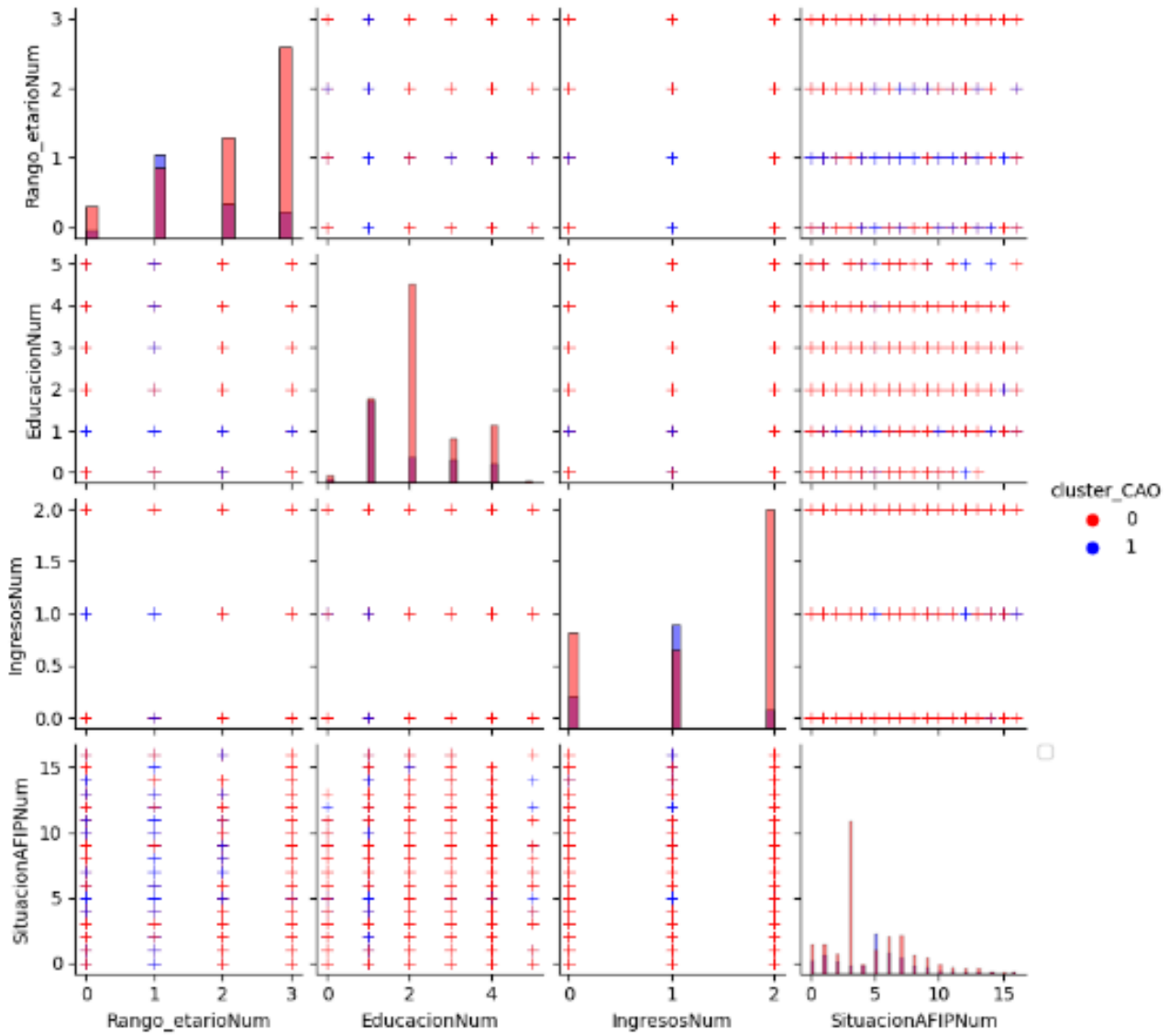
```
Clusers generados por CAO
0    9639
1    3752
Name: cluster_CAO, dtype: int64

Clusers generados por Huang
0    9637
1    3754
Name: cluster_Huang, dtype: int64
```

Para llegar a que generen dos clusters de medidas similares, se utilizó un parámetro que regula la cantidad de corridas y se lo ajustó hasta obtener resultados similares en la cantidad de cada cluster tanto de CAO como de HUANG para poder compararlos.

Mostramos a continuación una tabla comparativa de las descripciones de los dos clusters generados de acuerdo al método utilizado.

Características de los clusters



Comparación de los datos arrojados por ambos métodos (cluster mayor y menor):

Cluster mayor de CAO

```
count          9639
unique            4
top       Más de 55
freq           4669
Name: Rango_etario, dtype: object

count            9639
unique              6
top       Secundario
freq             4871
```

Cluster mayor de HUANG

```
count          9637
unique            4
top       Más de 55
freq           4617
Name: Rango_etario, dtype: object

count            9637
unique              6
top       Secundario
freq             5073
```

```
Name: Educación, dtype: object
```

```
Name: Educación, dtype: object
```

```
count            9639
unique              3
top     Más del 50%
freq             5363
Name: Ingresos, dtype: object
```

```
count                        9639
unique                         17
top     Responsable Inscripto
freq                         3742
Name: SituacionAFIP, dtype: object
```

Cluster menor de CAO

```
count           3752
unique             4
top       De 31 a 45
freq            2059
Name: Rango_etario, dtype: object
```

```
count          3752
unique            6
top        Primario
freq           1993
Name: Educación, dtype: object
```

```
count                3752
unique                  3
top       Entre 30% y 50%
freq                 2529
Name: Ingresos, dtype: object
```

```
count                    3752
unique                     17
top       Monotributo social
freq                      975
Name: SituacionAFIP, dtype: object
```

```
count            9637
unique              3
top       Más del 50%
freq             5347
Name: Ingresos, dtype: object
```

```
count                        9637
unique                         17
top     Responsable Inscripto
freq                         3172
Name: SituacionAFIP, dtype: object
```

Cluster menor de HUANG

```
count           3754
unique             4
top       De 46 a 55
freq            2031
Name: Rango_etario, dtype: object
```

```
count          3754
unique            6
top        Primario
freq           2136
Name: Educación, dtype: object
```

```
count                3754
unique                  3
top       Entre 30% y 50%
freq                 2623
Name: Ingresos, dtype: object
```

```
count                        3754
unique                         17
top       Responsable Inscripto
freq                          737
Name: SituacionAFIP, dtype: object
```

Podemos observar que en los clusters mayores obtenidos por cada método hay coincidencia en las características obtenidas, no así en el caso de los clusters con menor población, donde difieren en la edad y en la categoría de AFIP.

En cuanto a la edad, si bien difieren, son clases contiguas dentro de la variable, lo que no significa mayor distancia, y por lo tanto una menor diferencia entre ambos.

Con respecto a la diferencia en la categoría a la que pertenecen -o están inscriptos- en la AFIP, no encontramos un ordenamiento oficial que nos permita deducir mayores cosas con respecto a la "distancia" entre ambos resultados.

Utilizamos una técnica de reducción de dimensionalidad para visualizar los puntos en el espacio de dos dimensiones, con el Análisis de Componentes Principales (PCA) y el t-SNE.
Estas técnicas permiten proyectar los datos en un espacio de dos dimensiones de tal manera que se preserva la estructura de similitud entre los puntos en el espacio original.
Luego utilizamos un gráfico de dispersión para visualizar los puntos en el espacio de dos dimensiones y colorearlos según su etiqueta de cluster para ver si existe alguna estructura en los datos.

El primero corresponde a un gráfico de dos dimensiones utilizando PCA y el segundo con t-SNE.

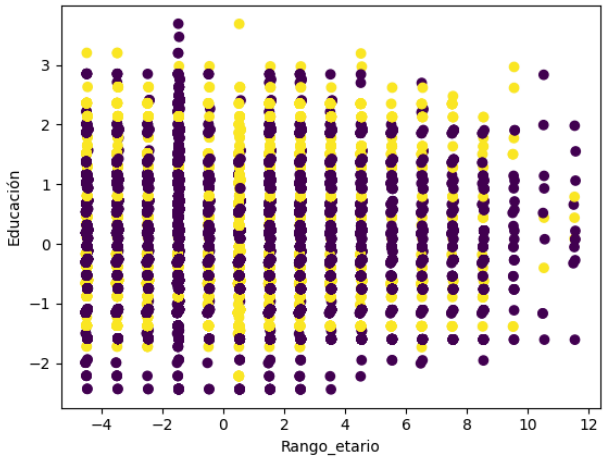


Gráfico PCA

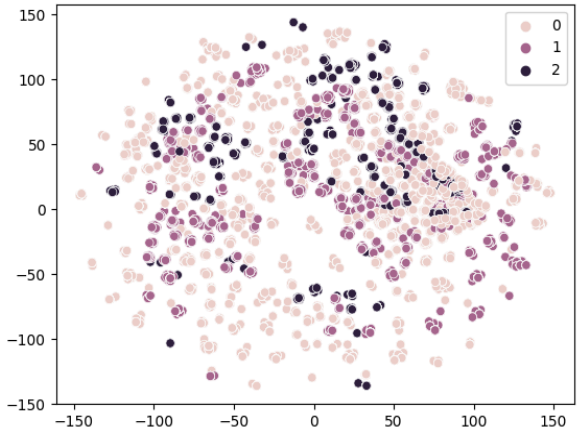


Gráfico t-SNE

En esta instancia, nos dimos cuenta que el ordenamiento de las categorías de AFIP quizás no era el correcto. Al modificarlo y volver a correr los datos, el K-Modes con CAO no sufrió ningún cambio, por el contrario, al usar Huang, los datos variaron bastante y descubrimos que; por cada corrida del algoritmo con los mismos parámetros, arrojaba, no sólo distintos cardinales de los clústers, sino distintas características de los mismos.

Conclusiones

Con los resultados obtenidos llegamos a la conclusión de que no hay, a priori, grupos bien definidos y que hay que seguir investigando sobre el uso del K-Modes (CAO y Huang) y el comportamiento del mismo y de sus parámetros, como así también la incorporación de otros métodos como puede ser DBSCAN ya que también funciona con variables categóricas y ver qué resultados arroja.

El resultado es en parte esperable ya que el programa tiene como población objetivo un estrato específico: las Pequeñas y medianas empresas (PYMES agropecuarias, agroalimentarias y agroindustriales) y las empresas familiares capitalizadas, y en apariencia no es tan sencillo distinguir grupos diferenciables dentro de la población de productores que forman parte del Programa.

Si bien no pudimos sacar conclusiones firmes, en apariencia la utilización de K-Modes con CAO hace más estable el algoritmo, mostrando resultados que coinciden con el conocimiento empírico de la base de datos del Programa Cambio Rural, por lo que nos resulta satisfactorio el resultado general.

Material adicional

- Decisión Administrativa 1441/2020
- DECAD-2020-1441-APN-JGM. Estructura organizativa
- Incorporación CEyCR en la estructura del Ministerio
- Funciones de la CEyCR
- Anexo 4
- Manual operativo
- Notebook del proyecto

Referencias

- https://biblioguias.uma.es/citasybibliografia/ejemplosAPA
- https://www.kaggle.com/code/halflingwizard/clustering-categorical-data-using-gower-distance
- https://www.youtube.com/watch?v=S5cL5MAFon8
- https://www.youtube.com/watch?v=o4bn2ZEGr4g
- http://www.scielo.org.co/pdf/cide/v7n1/v7n1a03.pdf

Esta revista propone un espacio académico propicio para **estimular, ampliar y difundir** investigaciones y debates sobre la problemática de las políticas públicas basadas en evidencia en los distintos niveles del Estado (local, provincial y nacional) así como también regional e internacional. Esperamos que la **primera edición** de Sinergías contribuya al mejoramiento del diseño, seguimiento y evaluación de las políticas públicas como intervenciones específicas y en articulación con los sistemas de protección nacional. La publicación es de carácter periódico semestral.