

22340810_tfm_JorgeLuisVasquezAraya-1.pdf

by Jorge Luis VÁSQUEZ ARAYA

Submission date: 07-Nov-2024 03:03AM (UTC+0100)

Submission ID: 2511038833

File name: 22340810_tfm_JorgeLuisVasquezAraya.pdf (1.88M)

Word count: 11915

Character count: 69859



Máster en BioInformática

**"Pronóstico de número incidencias de
enfermedades respiratorias en Chile
utilizando algoritmos de aprendizaje
automático"**

Autor: Jorge Luis Vásquez Araya

Tutor: Ana Monsalve Torra

Curso 2023-24

AGRADECIMIENTOS

Agradezco muy especialmente a mi familia, particularmente a mi núcleo conformado por mi madre, padre, hermana y sobrino quienes me han entregado afecto en cada momento que he necesitado, así como la cordura y disciplina cuando se ha requerido.

A mi pareja Vanessa Saldías y a mi perrita Luna Almendra por haber tenido la paciencia y comprensión de tener un compañero/humano algo ausente a momentos, sobre todo en nuestras actividades cotidianas a lo largo de este proceso, considerando que mi semana transcurría mayormente trabajando o estudiando.

Muy afectuosamente a mis abuelos(as) quienes ya no me acompañan en vida, pero que en los años que compartí con ellos sembraron con enorme cariño la fuerza, dedicación y un buen ejemplo de persona trabajadora y protectores empedernidos de sus sueños y familia.

A cada una de mis amistades que me han sido regaladas a lo largo de mi vida, porque siempre me han acompañado y ofrecido sabios consejos en innumerables momentos de conversación, así como el comprender que durante un año su amigo estaría atendiendo otras prioridades.

Por último, a Ana Monsalve Torra profesora quien tuvo la dedicación de guiarme durante este trabajo de final de maestría.

RESUMEN

Objetivos:

Esta investigación tiene como objetivo entrenar un modelo predictivo avanzado basado en variables climáticas y de calidad del aire para anticipar urgencias respiratorias en Chile. Incluye el análisis de datos, entrenamiento con algoritmos de Regresión Lineal y *CatBoost*, evaluar y comparar sus rendimientos. Además, se evaluará la importancia de las variables predictoras para cada modelo.

Material y métodos:

Desde tres fuentes públicas distintas se obtuvieron los datos de números de urgencias respiratorias, datos de climatología y condición del aire (específicamente material particulado 2.5) registrados entre los años 2014-2023. Se seleccionaron cuatro 4 características existentes y se generó una adicional para calcular la diferencia de temperaturas. Se consolidaron las variables en un solo *dataset* agrupado por año y semana epidemiológica, comuna (Santiago, La Serena y Concepción) resultando un *dataset* con 1572 observaciones y 8 variables predictoras. Se utilizaron dos algoritmos distintos para entrenar dos modelos de predicción de número de urgencias respiratorias, la Regresión Lineal y *CatBoost*. Para evaluar y comparar el rendimiento de cada modelo, fue calculado el coeficiente de determinación (R^2) y el error cuadrático medio (RMSE).

Finalmente utilizando la librería de *Caret* se obtuvo la importancia de las variables para cada modelo de aprendizaje automático.

Resultados:

Entrenados y evaluados ambos modelos, el modelo *CatBoost* destacó por su alta precisión con un $R^2 = 0.87$ en comparación con aquel entrenado con la Regresión Lineal que obtuvo un R^2 de 0.34. También *CatBoost* obtuvo una mejor desviación respecto a los valores reales, con un RMSE de 228.77 contra 521.95 del modelo entrenado con Regresión Lineal. Adicionalmente, se obtuvo la importancia de las variables predictoras destacando la mayor relevancia de la temperatura y el año.

Conclusiones:

Esta investigación demuestra que los modelos de aprendizaje automático, especialmente *CatBoost*, basado en un algoritmo de refuerzo de gradiente, son efectivos para predecir el número de urgencias por enfermedades respiratorias. Así también, estos hallazgos destacan el uso de distintas variables y el potencial de los algoritmos de aprendizaje automático para prever eventos y apoyar en el uso de los recursos en salud pública.

Palabras claves: Aprendizaje automático, Urgencias Respiratorias, Meteorología, Condición del aire, algoritmos, *CatBoost*, Regresión Lineal.

ABSTRACT

Objectives:

This research aims to train an advanced predictive model based on climate and air quality variables to anticipate respiratory emergencies in Chile. It includes data analysis, training with Linear Regression algorithms and CatBoost, evaluating and comparing their performances. In addition, the importance of the predictor variables for each model will be evaluated.

Material and methods:

Data on the number of respiratory emergencies, climatology data and air condition (specifically particulate matter 2.5) were collected from three different public sources, covering the period from 2014 to 2023. Four existing characteristics were selected and an additional one was generated to calculate the temperature difference. The variables were consolidated into a single data set grouped by year and epidemiological week, commune (Santiago, La Serena and Concepción) resulting in a data set with 1572 observations and 8 predictor variables. Two different algorithms were used to train two models for predicting the number of respiratory emergencies, Linear Regression and CatBoost. To evaluate and compare the performance of each model, the coefficient of determination (R^2) and the root mean square error (RMSE) were calculated.

Finally, using the Caret library, the importance of the variables for each machine learning model was obtained.

Results:

Once both models were trained and evaluated, the CatBoost model stood out for its high accuracy with an $R^2 = 0.87$ compared to the one trained with Linear Regression which obtained an R^2 of 0.34. CatBoost also obtained a better deviation from the real values, with an RMSE of 228.77 against 521.95 of the model trained with Linear Regression. Additionally, the importance of the predictor variables was obtained, highlighting the greater relevance of temperature and year.

Conclusion:

This research shows that machine learning models, especially CatBoost, based on a gradient boosting algorithm, are effective in predicting the number of emergencies due to respiratory diseases. Likewise, these findings highlight the use of different variables and the potential of machine learning algorithms to prevent events and support the use of resources in public health.

Keywords: Machine learning, Respiratory emergencies, Meteorology, Air condition, algorithms, CatBoost, Linear regression.

Índice de Contenido

1.	<i>Introducción</i>	10
2.	<i>Hipótesis</i>	12
3.	<i>Objetivos</i>	12
3.1.	<i>Objetivo General</i>	12
3.2.	<i>Objetivos específicos</i>	12
4.	<i>Estado del Arte</i>	13
5.	<i>Marco Conceptual</i>	16
5.1.	<i>Introducción</i>	16
6.	<i>Metodología</i>	22
6.1.	<i>Diseño</i>	22
6.2.	<i>Población</i>	22
6.3.	<i>Ámbito y tiempo</i>	23
6.4.	<i>Criterios de exclusión</i>	24
6.5.	<i>Variables</i>	25
6.6.	<i>Método de recogida de datos</i>	25
6.7.	<i>Manejo del dato</i>	26
6.8.	<i>Análisis estadístico</i>	26
6.9.	<i>Herramientas utilizadas</i>	27
6.10.	<i>Visión General del proceso de investigación</i>	28
7.	<i>Resultados</i>	31
8.	<i>Discusión</i>	47
9.	<i>Conclusiones</i>	50
10.	<i>Referencias</i>	51
	<i>Bibliografía</i>	51
11.	<i>Anexos</i>	53
12.	<i>Declaración obligatoria del uso de herramientas de IA</i>	53

Índice de Figuras

Figura 1 El diagrama de flujo del modelo CatBoost.....	19
Figura 2 Flujo General del proceso de la investigación.....	28
Figura 3 Urgencias respiratorias por año.....	32
Figura 4 Urgencias COVID-19 por año.....	33
Figura 5 <i>Total Anual de Urgencias Respiratorias por Comuna</i>	35
Figura 6 Mediana Temperatura Anual por Comuna	36
Figura 7 Total Precipitación Anual por Comuna	37
Figura 8 <i>Mediana Humedad Relativa Anual por Comuna</i>	38
Figura 9 Mediana Diferencia de Temperatura Anual por Comuna.....	39
Figura 10 Mediana Anual de Material Particulado 2.5 por Comuna	40
Figura 11 <i>Correlación entre variables</i>	41
Figura 12 <i>Valores Predichos vs. Reales regresión lineal</i>	44
Figura 13 <i>Valores Predichos vs. Reales CatBoost</i>	46



Índice de Tablas

Tabla 1 Orden y Causas de Urgencias	31
Tabla 2 Detalle urgencias anuales por comuna	34
Tabla 3 Características Dataset final	42
Tabla 4 Importancia de variables regresión lineal	43
Tabla 5 Importancia de variables CatBoost	45

1. Introducción

La interacción entre factores climáticos, contaminación del aire y enfermedades respiratorias constituye una preocupación creciente en Chile, un país con variada geografía a lo largo de su extenso territorio lo que origina diversidad de clima (Sarricolea, Herrera-Ossandon, & Meseguer-Ruiz, 2016), el que eventualmente se combina con altos niveles de contaminación atmosférica, especialmente en zonas urbanas y áreas con alta concentración industrial. Este fenómeno afecta desproporcionadamente a grupos vulnerables de la población, como los niños, los ancianos y las personas con condiciones de salud preexistentes, quienes son particularmente sensibles a las variaciones en la calidad del aire y las condiciones climáticas. A lo largo de los años, numerosos estudios han subrayado el impacto que tiene la contaminación ambiental en la salud pública, evidenciando una correlación entre la exposición a partículas contaminantes y el aumento en la incidencia de enfermedades respiratorias como el asma, bronquitis crónica y otras afecciones agudas y crónicas del sistema respiratorio (Cifuentes Martínez et al., 2020).

Uno de los factores determinantes en la dispersión de los contaminantes atmosféricos es el clima. Variables como la temperatura, la humedad y precipitaciones juegan un papel clave en la capacidad de las partículas finas y gases contaminantes para permanecer en la atmósfera o ser dispersados (Romero et al., 2010). En ciudades como Santiago, donde la geografía rodeada de montañas genera condiciones de estancamiento del aire, la acumulación de contaminantes durante ciertos meses del año, particularmente en invierno, agrava significativamente los problemas respiratorios en la población. Este fenómeno, conocido como inversión térmica, contribuye a que los niveles de partículas finas en el aire, como el material particulado PM_{2.5}, superen los límites recomendados por las organizaciones internacionales de salud, exacerbando los problemas respiratorios y aumentando las tasas de hospitalización por enfermedades relacionadas (Organización Mundial de la Salud, 2021).

Ante este contexto, surge la necesidad de desarrollar herramientas predictivas que permitan anticipar las consecuencias de la contaminación del aire sobre la salud respiratoria de la población, considerando tanto las variables climáticas como las de calidad del aire. El presente estudio tiene como objetivo generar un modelo fiable que

logre predecir el impacto de estas variables en la salud respiratoria en tres de las comunas más pobladas de Chile: **Santiago, La Serena y Concepción**, seleccionadas estratégicamente por ser las capitales regionales y también representativas del centro, norte y sur del país, respectivamente. Estas comunas presentan características geográficas y climáticas diversas, lo que ofrece una oportunidad única para estudiar cómo interactúan los factores climáticos y la contaminación del aire en diferentes contextos ambientales (Rojas & Martínez, 2022).

A través de un enfoque interdisciplinario que integra la climatología, la salud pública, las tecnologías de la información y las ciencias de datos, este estudio busca utilizar la gran cantidad de información generada por distintas instituciones públicas para entrenar un modelo que prediga el número de urgencias respiratorias reportadas. Este modelo no solo permitirá comprender mejor el impacto de la interacción entre el clima y la contaminación del aire sobre la salud pública (Oyarzún G. & Valdivia C., 2021) sino que también podría ser útil para proyectar los costos, ya que las enfermedades respiratorias también generan importantes pérdidas económicas indirectas relacionadas con el ausentismo laboral, considerando que las personas que padecen afecciones respiratorias crónicas o recurrentes suelen necesitar tomar días libres del trabajo, lo que puede afectar la productividad de las empresas y por ende generar un impacto en la economía nacional como lo señalan algunas investigaciones (Mosegui, Antoñanzas, & de Mello Vianna, 2023).

Finalmente, se espera que, tras entrenar, evaluar y comparar dos modelos, el que tenga mejor rendimiento anticipando el número de urgencias por enfermedades respiratorias, pueda en un futuro ser usado para contribuir significativamente en la formulación de políticas públicas más efectivas, orientadas a reducir la exposición de la población a estas enfermedades. Asimismo, este modelo puede ayudar a establecer bases sólidas para que el estado en colaboración con el sector privado, desarrollen estrategias preventivas que minimicen el impacto que estas enfermedades tienen sobre la sociedad y la economía.

2. Hipótesis

Los factores climáticos y la calidad del aire tienen un impacto significativo en la incidencia de enfermedades respiratorias.

Las variaciones en estos factores tales como cambios en la temperatura, humedad, precipitaciones y niveles de contaminantes atmosféricos, pueden repercutir en la aparición de dichas enfermedades y por ende en el aumento de las urgencias reportadas por las instituciones de salud pública.

3. Objetivos

3.1. Objetivo General

Entrenar un modelo predictivo avanzado utilizando variables climáticas y de calidad del aire, considerando el periodo de datos 2014-2023, con el propósito de anticipar y predecir el número de urgencias por enfermedades respiratorias reportadas en el sistema público.

3.2. Objetivos específicos

- a) Realizar un análisis exploratorio de los *datasets* que se van a usar en el entrenamiento del modelo.
- b) Entrenar un modelo de predicción utilizando el algoritmo de Regresión Lineal.
- c) Entrenar un modelo de predicción utilizando el algoritmo de *CatBoost*.
- d) Comparar el rendimiento de ambos modelos entrenados.
- e) Obtener la importancia de las variables predictoras para ambos modelos.

4. Estado del Arte

En el contexto de las enfermedades respiratorias, el estudio de factores ambientales y su relación con la salud humana ha ganado importancia, especialmente tras la pandemia del COVID-19, la cual resaltó la vulnerabilidad del sistema respiratorio a influencias externas. Si buscamos en la literatura se puede encontrar que los factores climáticos como la temperatura, la humedad, y la calidad del aire juegan un rol fundamental en la aparición de enfermedades respiratorias (Lee & Yoon, 2024) (Behinaein et al., 2023). Sin embargo, a pesar del reconocimiento de esta relación, han sido pocos los estudios que han abordado la predicción de estas enfermedades mediante el uso conjunto de datos climáticos y de contaminación del aire.

Un enfoque emergente en la predicción de la ocurrencia de enfermedades respiratorias (y también de otro tipo) es la aplicación de modelos de aprendizaje automático (*machine learning*). Esta disciplina ha sido utilizada en otros ámbitos de la salud, como la predicción de infecciones virales o accidentes cerebrovasculares, con resultados prometedores (Zhang et al., 2022). Sin embargo, en el caso específico de las enfermedades respiratorias, la mayoría de los estudios existentes se han enfocado en el análisis de un único factor contaminante, como el material particulado fino (PM_{2.5}), sin incluir variables climáticas adicionales como la temperatura, precipitaciones o la humedad relativa, por dar algunos ejemplos (Yang et al., 2018).

Uno de los estudios pioneros en la integración de múltiples factores climáticos y de contaminación en la predicción de enfermedades respiratorias es el realizado por un grupo de investigadores en Korea del Sur (Ku et al., 2022). Este trabajo representa un avance significativo al aplicar técnicas de aprendizaje automático como *gradient boosting* y la regresión de procesos gaussianos (GPR) para predecir la incidencia de enfermedades respiratorias. Al incluir datos sobre temperatura, humedad, presión atmosférica y niveles de contaminantes como el dióxido de azufre (SO₂) y el material particulado (PM_{2.5}), este estudio proporciona una visión integral de los factores que afectan la salud respiratoria. El uso del análisis SHAP (*Shapley Additive exPlanations*) les permitió además interpretar los resultados del modelo, mostrando cómo diferentes variables impactan en la ocurrencia de enfermedades respiratorias.

A lo largo de los últimos años, estudios previos han subrayado la relevancia de factores individuales en la aparición de infecciones respiratorias. Por ejemplo, las bajas temperaturas se han asociado con la propagación del virus respiratorio sincicial en los niños (RSV) y otros patógenos respiratorios (Vandini et al., 2013), mientras que la baja humedad ha sido relacionada con la transmisión aérea de virus a través de partículas de menor tamaño (Yang & Marr, 2012). Asimismo, la contaminación del aire, particularmente el PM2.5, ha sido ampliamente estudiada por su papel en la exacerbación de enfermedades respiratorias, con estudios que indican una correlación directa entre los niveles de este contaminante y el aumento en las hospitalizaciones por infecciones respiratorias (Croft et al., 2019).

A pesar de la importancia de estos hallazgos, la mayoría de los estudios previos se han limitado a un análisis retrospectivo, sin prever la ocurrencia de enfermedades. En este sentido, el estudio de (Ku et al., 2022) marca una innovación al utilizar métodos predictivos en lugar de descriptivos, abriendo una nueva vía para el desarrollo de sistemas de alerta temprana en salud pública. Los modelos propuestos por estos autores, el *gradient boosting* y la regresión de procesos gaussianos (GPR), representa un avance significativo en este campo. Estos modelos permiten no solo predecir la incidencia diaria de enfermedades respiratorias con una precisión competitiva (R^2 de 0.68 para *gradient boosting* y 0.67 para GPR), sino que también ofrecen una interpretación clara de cómo los factores climáticos y de contaminación contribuyen a las variaciones en la incidencia de dichas enfermedades. Sin embargo, también reconocen ciertas limitaciones, como la falta de datos demográficos detallados y la necesidad de expandir el modelo a otras regiones con características climáticas y niveles de contaminación diferentes que les permitiría generalizar su modelo.

En resumen, el estado del arte en la predicción de enfermedades respiratorias mediante factores climáticos y de contaminación del aire ha avanzado significativamente con la introducción de técnicas de aprendizaje automático, proporcionando nuevas herramientas para analizar grandes volúmenes de datos ambientales y de salud, para identificar patrones que anteriormente eran difíciles de detectar utilizando métodos estadísticos tradicionales. Sin embargo, si bien el estudio de (Ku et al., 2022) establece una base sólida para futuras investigaciones, queda pendiente la exploración de su

aplicabilidad en áreas rurales o en regiones con distintos climas, ya que las diferencias en la infraestructura de salud, la distribución geográfica y las condiciones socioeconómicas también pueden influir en la efectividad de estos modelos predictivos en otras ubicaciones.

5. Marco Conceptual

5.1. Introducción

La predicción de urgencias médicas relacionadas con enfermedades respiratorias es un desafío creciente en el ámbito de la salud pública. Los sistemas de atención sanitaria a menudo enfrentan dificultades para planificar la capacidad de atención y la asignación de recursos debido a la fluctuación en la demanda de servicios de urgencia. Las enfermedades respiratorias, que pueden verse exacerbadas por factores estacionales y ambientales, representan una parte significativa de estas urgencias por lo que predecir el número de pacientes que requerirán atención médica permite optimizar los recursos y mejorar la capacidad de respuesta.

En este contexto, el uso de técnicas de aprendizaje automático, como los modelos de Regresión Lineal y *CatBoost*, se presenta como una alternativa poderosa para realizar predicciones precisas.

5.2. Modelos predictivos en el ámbito de la salud

Los modelos predictivos en salud se han utilizado ampliamente para pronosticar diferentes tipos de eventos médicos, como la cantidad de hospitalizaciones por urgencias (Cusidó et al., 2022), la propagación de enfermedades infecciosas o la demanda de servicios de urgencia. Entre las técnicas más utilizada esta la regresión línea, la cual ha demostrado ser una herramienta estadística efectiva para entender las relaciones entre una variable dependiente y una o más variables independientes. En el caso de las urgencias respiratorias, variables como la temperatura, la humedad, los niveles de contaminación y las tendencias históricas de enfermedades respiratorias pueden ser utilizadas para construir el modelo predictivo.

5.3. Algoritmo de Regresión Lineal

La Regresión Lineal es un modelo matemático que asume una relación lineal entre la variable dependiente (en el caso de este trabajo, el número de urgencias respiratorias) y una o más variables independientes (factores como la temperatura, la contaminación del aire y la semana del año). Es un enfoque sencillo y fácil de interpretar, lo que la convierte como una de las primeras opciones al momento de realizar un estudio predictivo.

El modelo de Regresión Lineal puede ser expresado mediante la ecuación:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n + \epsilon \quad (1)$$

Donde:

- Y: variable dependiente (lo que se intenta predecir).
- $X_1, X_2, \dots, X_n$: variables independientes (predictoras).
- β_0 : intercepto (el valor de Y cuando todas las variables independientes son 0).
- $\beta_1, \beta_2, \dots, \beta_n$: coeficientes que representan el cambio en Y por un cambio unitario en cada X.
- ϵ : término de error (diferencia entre el valor real y el predicho de Y).

La Regresión Lineal ha sido ampliamente utilizada en la investigación en salud para modelar y predecir variables continuas (Schober & Vetter, 2020), pero, aunque este enfoque es fácil de interpretar y ha sido efectivo en muchas aplicaciones, también presenta limitaciones, sobre todo cuando la relación entre las variables no es completamente lineal, además, su rendimiento se ve afectado por la multicolinealidad y por la presencia de *outliers*, características que pueden reducir la precisión predictiva de este tipo de modelo.

5.4. Boosting

El *Boosting* está basado en el cuestionamiento planteado por Kearns y Valiant (Freund & Schapire, 1999), acerca si dado un conjunto de clasificadores débiles *¿es posible crear un clasificador más eficiente?*. Esta técnica de aprendizaje automático básicamente combina múltiples modelos débiles, generalmente árboles de decisión, para formar un modelo fuerte que mejora el rendimiento predictivo. La idea central detrás del *boosting* es entrenar de manera secuencial varios modelos, de manera que cada uno se enfoque en corregir los errores cometidos por sus predecesores.

5.5. Algoritmo *CatBoost*

*CatBoost*¹ utiliza un enfoque de *boosting* que combina múltiples árboles de decisión, lo que permite mejorar iterativamente la precisión del modelo. A diferencia de otros algoritmos de *boosting* como *XGBoost*, *CatBoost* maneja directamente las variables categóricas sin necesidad de codificación adicional, lo que reduce el riesgo de perder información relevante durante el preprocesamiento (Dorogush, Ershov, & Gulin, 2018)

En el contexto de predicción de urgencias respiratorias, este modelo puede capturar interacciones no lineales entre múltiples factores, como el clima, los niveles de contaminación, proporcionando una predicción más robusta y precisa.

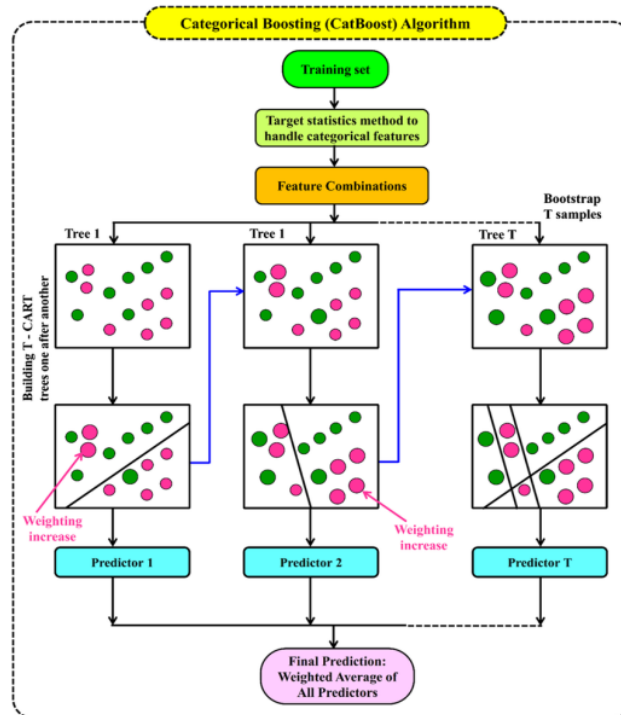
Una de las ventajas de *CatBoost* es su capacidad para proporcionar interpretaciones claras sobre la importancia de las características en la predicción. Se utilizarán técnicas en conjunto con la librería *Caret* (Classification and Regression Training)² para identificar qué variables tienen el mayor impacto en la predicción, lo que facilita obtener conclusiones útiles y comprensibles sobre los factores que más influyen en el número de urgencias respiratorias reportadas.

¹ <https://catboost.ai/>

² <https://cran.r-project.org/web/packages/Caret/index.html>

Figura 1

El diagrama de flujo del modelo CatBoost



Nota. En la figura se representa el flujo que internamente efectúa el algoritmo al entrenar y hacer una predicción. Fuente: (Water Resources Management, 2023)

Breve explicación del flujo de la **Figura 1** paso a paso:

- Conjunto de Entrenamiento (*Training Set*): El proceso comienza con un conjunto de datos de entrenamiento que incluye tanto características numéricas como categóricas.
- Método de Estadísticas de Objetivo para Manejar Características Categóricas (*Target Statistics Method*): Como se mencionó anteriormente, *CatBoost* maneja las características categóricas directamente, sin necesidad de codificarlas en variables numéricas (lo que comúnmente se hace con una codificación *one-hot*).
- Combinaciones de Características (*Feature Combinations*): El algoritmo crea combinaciones de características para generar nuevas variables a partir de las existentes, lo que incrementa la capacidad del modelo de capturar interacciones

complejas entre las variables. Esto permite al modelo descubrir relaciones no triviales en los datos.

- d) Construcción de los Árboles de Decisión (*Building Trees*): El algoritmo *CatBoost* construye una serie de árboles de decisión secuenciales, como se muestra en el diagrama, de "Tree 1" a "Tree T". Cada iteración ajusta un árbol basado en los errores cometidos por los anteriores. Este proceso iterativo caracteriza a los algoritmos de *boosting*, cada predictor sucesivo se entrena para corregir los errores del anterior (Prokhorenkova et al., 2019).
- e) Aumento de Pesos (*Weighting Increase*): Después de entrenar cada árbol, se identifican las observaciones que fueron mal clasificadas o predichas de manera incorrecta. A estas observaciones se les asigna un mayor peso en las iteraciones posteriores, de manera que el siguiente árbol enfoque su entrenamiento en corregir estos errores.
- f) Muestras Bootstrap (*Bootstrap Samples*): En cada iteración, el algoritmo toma muestras aleatorias del conjunto de datos para entrenar cada árbol de decisión. Esto introduce aleatoriedad en el proceso y mejora la capacidad generalizadora del modelo, reduciendo el riesgo de sobreajuste (*overfitting*).
- g) Predicción Final (*Final Prediction*): Una vez que se han entrenado todos los árboles, el algoritmo combina los predictores mediante un promedio ponderado para generar la predicción final. El peso de cada predictor se ajusta en función de su precisión, lo que permite que los modelos más precisos tengan una mayor influencia en la predicción final.

5.6. Métricas de evaluación

Para la evaluación y comparación de ambos modelos se va a utilizar el error cuadrático medio (RMSE siglas en inglés) y el R^2 (coeficiente de determinación).

- a) RMSE mide la desviación promedio de las predicciones del modelo respecto a los valores reales, penalizando de forma cuadrática los errores grandes. Cuanto más bajo sea el valor, mejor es el ajuste del modelo, ya que indica que el modelo se desvía menos en promedio. Esto siempre en relación con la unidad de la variable objetivo.

$$RMSE = \sqrt{1/n * \sum (y_i - \hat{y}_i)^2} \quad (2)$$

donde:

y_i son los valores reales,

$\hat{y}_i$ son los valores predichos,

n es el número de observaciones.

- b) R^2 indica la proporción de la variabilidad en la variable dependiente (las urgencias, en tu caso) que es explicada por el modelo de regresión. Va de 0 a 1, donde 1 significa que el modelo explica toda la variabilidad de los datos y 0 significa que no explica nada.

$$R^2 = 1 - (\sum (y_i - \hat{y}_i)^2 / \sum (y_i - \bar{y})^2) \quad (3)$$

donde:

y_i son los valores reales,

$\hat{y}_i$ son los valores predichos,

$\bar{y}$ es la media de los valores reales.

5.7. Conclusión

El uso de modelos predictivos para anticipar el número de urgencias relacionadas con enfermedades respiratorias es un campo emergente en la investigación de salud pública. Comparar la Regresión Lineal y *CatBoost* ofrecerá valiosas perspectivas sobre qué modelos pueden proporcionar mejores predicciones, lo que puede ser relevante para la gestión de recursos hospitalarios. Dado que las enfermedades respiratorias continuarán siendo una preocupación en el futuro, el desarrollo de modelos predictivos efectivos es esencial para mejorar la respuesta de los sistemas de salud.

6. Metodología

6.1. Diseño

La investigación adopta un enfoque cuantitativo, el cual se fundamenta en el uso de técnicas estadísticas y matemáticas para analizar los datos recopilados. Este enfoque es esencial para garantizar la objetividad de los resultados y permite la obtención de conclusiones basadas en evidencias empíricas. En una primera etapa, se procede a la recolección y preprocesamiento de los datos públicos, los cuales son utilizados para entrenar, validar y probar un modelo predictivo del número de casos. El proceso de entrenamiento del modelo involucra la selección de las variables relevantes, así como la **división de los datos en conjuntos de entrenamiento y validación**, lo que permite ajustar los parámetros del modelo de manera óptima.

Una vez que el modelo ha sido entrenado, validado y probado, se procede a la evaluación de su desempeño mediante métricas estadísticas específicas, tales como el coeficiente de determinación (R^2) y el error cuadrático medio (RMSE). Posteriormente, se lleva a cabo un análisis descriptivo de los resultados obtenidos a partir del modelo predictivo, lo que permite identificar patrones y tendencias significativas en los datos de las tres comunas seleccionadas para el estudio (Santiago, La Serena y Concepción). Este análisis descriptivo es fundamental para interpretar los resultados y para comprender mejor la dinámica en la distribución de los datos.

Finalmente, los resultados de este estudio serán comparados con un modelo predictivo secundario, entrenado, validado y probado con los mismos datos que el primario, esto con el objetivo de comparar y evaluar la consistencia y robustez del primer modelo.

Esto no solo permite fortalecer la validez de los resultados, sino también identificar posibles áreas de mejora o refinamientos futuros en la modelización y análisis de los datos.

6.2. Población

Se considera la población de las tres comunas (Santiago, La Serena y Concepción), las cuales, al sumar sus respectivas poblaciones, alcanzan un total de 849,123 personas (Instituto Nacional de Estadística, 2017). Estas comunas fueron seleccionadas debido a su relevancia demográfica, ubicación geográfica y a la disponibilidad de los datos

considerados en este estudio. La elección de estas áreas geográficas también responde a la diversidad de sus características sociodemográficas, lo que proporciona un contexto adecuado para el análisis de las variaciones en la incidencia de enfermedades respiratorias en diferentes grupos de población. Este enfoque podría permitir generalizar los hallazgos a otras regiones con características similares, ampliando la relevancia de la investigación en un futuro.

La presente investigación se centra en los pacientes que fueron atendidos en el sistema de salud público de Chile entre los años 2014 y 2023, un periodo que abarca una década de registros médicos. Durante este tiempo, se recogieron datos de pacientes de todas las edades, desde recién nacidos hasta personas mayores de 65 años, lo que permite un análisis detallado de las tendencias en diferentes grupos etarios. Los pacientes incluidos en el estudio fueron aquellos que presentaron una urgencia médica relacionada con una enfermedad respiratoria, una categoría que abarca una amplia gama de patologías que incluyen desde infecciones respiratorias agudas hasta condiciones crónicas, ejemplo: crisis obstructiva bronquial.

Además, se tiene en cuenta que la población atendida en el sistema público de salud de Chile representa una muestra significativa del total de la población nacional, especialmente en lo que respecta a grupos vulnerables, quienes dependen en gran medida de los servicios de salud pública. El análisis de estos datos es de gran importancia para identificar patrones epidemiológicos y tendencias emergentes en la incidencia de enfermedades respiratorias, lo que a su vez puede contribuir a mejorar las políticas de salud pública y la gestión de los recursos sanitarios en el futuro.

6.3. Ámbito y tiempo

Se hará uso de los datos públicos previamente mencionados, los cuales abarcan el periodo comprendido entre los años 2014 y 2023. Estos datos han sido seleccionados por su relevancia temporal, ya que permiten observar tendencias y patrones en la incidencia de enfermedades respiratorias a lo largo de una década. Sin embargo, es importante señalar que el año 2024 no fue considerado en la fase de entrenamiento del modelo predictivo debido a que los datos correspondientes a dicho año aún no se encuentran completos ni han sido validados por las fuentes que los colectan y publican.

Esto es fundamental para asegurar la calidad de los datos empleados en la modelización, dado que la inclusión de información incompleta o no validada podría introducir sesgos o errores en los resultados obtenidos, afectando así la precisión del modelo.

A pesar de no utilizar el año en curso para el entrenamiento del modelo, es posible emplear los datos preliminares de 2024 para llevar a cabo una comparación entre las predicciones generadas por el modelo y los valores reales que se vayan reportando a medida que avance el proceso de validación de dichos datos. Esta comparación será útil para evaluar la capacidad predictiva del modelo en relación con su objetivo principal, que es anticipar el número de casos relacionados con enfermedades respiratorias en las tres comunas analizadas.

6.4. Criterios de exclusión

Uso de datos anonimizados de casos y/o urgencias reportadas por enfermedades respiratorias, desde el año 2014 al presente.

La anonimización de los datos es un aspecto fundamental para garantizar la privacidad y protección de la información personal de los pacientes, cumpliendo así con las normativas éticas y legales vigentes en Chile, tales como la Ley sobre Protección de la Vida Privada (Biblioteca del Congreso Nacional de Chile, 2023).

En cuanto a las enfermedades respiratorias incluidas en el estudio, se ha limitado la selección a aquellas que se encuentran registradas en los datos publicados oficialmente por el Ministerio de Salud de Chile (MIN-SAL), los cuales son clasificados de acuerdo con la Clasificación Internacional de Enfermedades, 10ª Revisión (CIE-10)³. Esta clasificación es un estándar internacionalmente reconocido que facilita la codificación de enfermedades y otros problemas de salud, permitiendo un análisis sistemático y comparable.

Las enfermedades consideradas son ¹ las siguientes:

IRA alta (J00-J06), Influenza (J09-J11), Neumonía (J12-J18), Bronquitis/Bronquiolitis, aguda (J20-J21), Crisis obstructiva bronquial (J40-J46), Otra causa respiratoria (J22, J30-

³ https://repositoriodeis.minsal.cl/ContenidoSitioWeb2020/uploads/2020/12/CIE-10_2018_VOL1.pdf

J39, J47, J60-J98), COVID-19, virus no identificado (U07.2) y COVID-19, virus identificado (U07.1).

Finalmente, el uso de datos anonimizados y la aplicación de la clasificación CIE-10 garantiza que la investigación esté alineada con los estándares de calidad y rigor científico necesarios para abordar de manera adecuada preguntas, permitiendo una interpretación precisa y confiable de los resultados por parte de los expertos en la materia.

6.5. Variables

- a) Numero de total de casos: Número total de urgencias por semana epidemiológica⁴, esta es la variable dependiente por predecir.
- b) Comuna: Ubicación geográfica para la cual se registró la urgencia; Santiago, La Serena y Concepción.
- c) Anio: Año de registro de la urgencia reportada.
- d) MedianaClimatologica: Mediana climatológica de temperatura para la semana.
- e) DifMaxMin: Mediana de la diferencia de temperaturas máximas y mínimas para la semana.
- f) Mediana_HR_Valor: Mediana de humedad relativa para la semana.
- g) Total_Valor_semPP: Suma de agua caída en mm durante la semana.
- h) vTotal_Valor_semMP2: Suma de cantidad de material particulado menor a 2.5 micrones presentes en el aire para la semana.

6.6. Método de recogida de datos

El estudio hará uso de fuentes de datos abiertos⁵, los cuales son generados y proporcionados por diversas instituciones públicas chilenas. Este tipo de datos, al estar disponibles públicamente y haber sido previamente anonimizados por las entidades que los recopilan, no requieren la obtención de consentimiento informado por parte de los individuos cuyos registros forman parte del conjunto de datos.

⁴ NOTA: Se entiende por semana epidemiológica al lapso de tiempo transcurrido entre las 0:00 horas del día domingo hasta las 24:00 horas del día sábado siguiente. La primera semana epidemiológica del año termina el primer sábado de enero que incluya en los días inmediatamente precedentes cuatro o más días del mes de enero. Fuente: <https://saludresponde.minsal.cl/calendario-epidemiologico/>

⁵ La única condición para hacer uso de estos conjuntos de datos es referenciar a sus autores.

A continuación, las instituciones chilenas desde donde se originan de los datos:

a) Ministerio de Salud (MinSal).

¹Atenciones de urgencias por causas respiratorias por semana epidemiológica, generados por el Ministerio de Salud de Chile y compartidos bajo la licencia Creative Commons Non-Commercial. Datos en formato Parquet.

<https://datos.gob.cl/dataset/atenciones-de-urgencia-causas-respiratorias>

b) Dirección Meteorológica de Chile.

Datos climatológicos históricos, generados y compartidos públicamente por la Dirección Meteorológica de Chile. Datos en formato CSV.

<https://sinca.mma.gob.cl/>

c) Sistema de Información Nacional de Calidad del Aire, Ministerio del Medio Ambiente.

Concentraciones ambientales de los contaminantes atmosféricos actualmente normados en Chile, asociados a los niveles de concentración que originan situaciones de emergencia ambiental, para aquellas zonas que cuentan con estaciones de monitoreo. Datos en formato CSV.

<https://climatologia.meteochile.gob.cl/>

6.7. Manejo del dato

Al tratarse de datos que fueron recolectados por terceros (señalados en 6.6), y que desde esas fuentes fueron anonimizados y disponibilizados públicamente, en este trabajo no se requiere diseñar un proceso de gestión del dato ya que son estas instituciones las que toman esta responsabilidad.

Sin perjuicio de lo anterior, el presente estudio tiene la obligación de mencionar aquellas fuentes como los proveedores de estos datos.

6.8. Análisis estadístico

Se realizará un análisis descriptivo de los datos que se utilizarán para entrenar el modelo predictivo. Este análisis es crucial para entender las características fundamentales del

conjunto de datos, como la distribución de las variables clave, su tendencia central (media, mediana, moda) y la dispersión (desviación estándar, varianza), lo que permitirá identificar patrones generales y posibles sesgos en los datos. La comprensión de estas características básicas proporcionará una base sólida para el desarrollo y entendimiento del modelo predictivo.

Además, se prestará especial atención a la identificación de valores atípicos u *outliers* que podrían influir negativamente en la precisión del modelo, también abordará la presencia de datos faltantes y las estrategias para su tratamiento. Estos pasos son esenciales para asegurar que el modelo se entrene con datos limpios y representativos, lo que incrementará su fiabilidad.

Finalmente, este análisis descriptivo servirá como una herramienta para la selección de variables relevantes, permitiendo optimizar el rendimiento del modelo predictivo sobre una base estadística sólida y con un alto potencial de precisión para predecir la incidencia de enfermedades respiratorias.

6.9. Herramientas utilizadas

- Lenguaje R⁶: Lenguaje de programación y entorno de software utilizado para análisis estadístico y gráficos, ampliamente utilizado en ciencia de datos y bioestadística.
- Google *Colaboratory*⁷: es una plataforma en la nube que permite ejecutar Python en el navegador, proporcionando acceso gratuito a GPUs y facilitando la colaboración en proyectos de ciencia de datos (Google, 2023).
- GitHub⁸: Es una plataforma de alojamiento para control de versiones con Git, que facilita la colaboración entre desarrolladores de software mediante la gestión de repositorios.
- Jamovi⁹: es un software de análisis estadístico diseñado para ser fácil de usar y accesible, con una interfaz gráfica intuitiva y soporte para análisis avanzado (The jamovi project, 2022). Se basa en librerías del lenguaje R.

⁶ <https://www.r-project.org/>

⁷ <https://colab.research.google.com/>

⁸ <https://github.com/>

⁹ <https://www.jamovi.org>

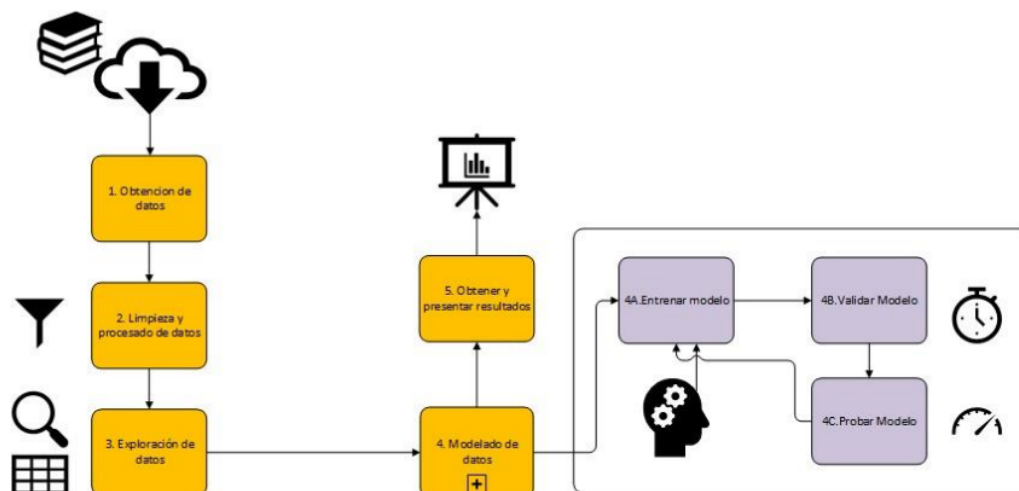
- Microsoft Word¹⁰: Es un procesador de texto popular que forma parte de la suite Microsoft Office, utilizado para crear, editar y compartir documentos de texto (Microsoft, 2023).

6.10. Visión General del proceso de investigación

Para llevar a cabo el presente estudio, se diseñó y se implementó un flujo de trabajo estructurado que abarca una serie de actividades cuidadosamente planificadas, representando cada uno de los pasos necesarios para alcanzar los objetivos propuestos. Este flujo de actividades es fundamental para garantizar un enfoque sistemático y riguroso en todas las etapas del proceso de investigación, desde la obtención inicial de los datos hasta la presentación final de los resultados.

Figura 2

Flujo General del proceso de la investigación.



Nota: La figura muestra el flujo general de trabajo para esta investigación. Fuente: figura de elaboración propia.

A continuación, se describen los pasos:

- Obtención de datos:** Aquí se identifican y adquieren los datos necesarios para cumplir con los objetivos del proyecto, lo que puede implicar consultar bases de

¹⁰ <https://www.microsoft.com>

datos, extraer información de sitios web o acceder a archivos disponibles. En el caso de este estudio se descargaron en 2 formatos distintos (parquet y csv) desde los sitios web de las instituciones que los publicaron.

- b) **Limpieza y procesamiento de datos:** Este paso es el que requiere más tiempo dentro en el proceso, ya que se deben unificar todos los datos en un formato común, organizarlos adecuadamente, eliminar aquellos que no son relevantes para el análisis y reemplazar los datos faltantes para asegurar la integridad del conjunto de datos destinado al modelado. En el caso en particular se tuvieron que seleccionar rangos de fechas similares, para que los datos estuvieran disponibles en los tres *datasets* distintos, luego normalizar las agrupaciones de datos a semanas epidemiológicas, calcular medianas entre otras acciones que se tomaron para llegar a un *dataset* final (consolidado).
- c) **Exploración de datos:** Después de limpiar los datos, se analiza aplicando técnicas analíticas y estadísticas para identificar relaciones entre sus características, así como correlaciones estadísticas relevantes. Estas relaciones permiten predecir una variable de interés en el ámbito de la salud, conocida como etiqueta u objetivo, que puede ser un valor cuantitativo, como el riesgo de desarrollar una enfermedad o en el caso del presente estudio la cantidad de urgencias reportadas para un periodo dado bajo ciertas condiciones.
- d) **Modelado de datos:**
 - i. Entrenar: Se alimenta el modelo de aprendizaje automático con un conjunto de datos etiquetados o no etiquetados, resultado de los procesos anteriores para que de esta forma el modelo pueda aprender patrones y sus relaciones. Durante esta fase, el modelo va ajustando sus parámetros internos en función de los datos de entrenamiento para minimizar el error en la predicción. El objetivo es que el modelo sea capaz de generar predicciones precisas sobre datos nuevos y no vistos.
 - ii. Validar: La validación es un paso intermedio clave que se emplea para ajustar el modelo y evitar problemas como el sobreajuste. Este proceso se puede realizar a través de técnicas como por ejemplo la validación cruzada (*cross-*

validation) (Manorathna, 2020), donde el conjunto de datos se divide en varios subconjuntos, y el modelo se entrena y valida varias veces en diferentes particiones de los datos. La validación proporciona un ajuste más fino de los hiperparámetros del modelo, mejorando así su rendimiento y resultado.

- iii. Probar: La prueba de un modelo de ciencia de datos se realiza después del entrenamiento y validación, se utiliza una parte del conjunto previamente procesado y que no se ha utilizado en el entrenamiento inicial. Esta fase permite evaluar qué tan bien funciona el modelo ya entrenado y validado, con datos que no ha visto previamente. Los resultados de esta prueba ofrecen una medida más realista de su desempeño en el mundo real, lo que ayuda a evitar el sobreajuste (*overfitting*) y garantizar que el modelo es capaz de generalizar.
- iv. Obtener y presentar resultados: La obtención y presentación de resultados es la fase final en la que los datos obtenidos del modelo se analizan, se obtienen conclusiones y estas se comunican de manera clara y comprensible. Esta etapa incluye la interpretación de las métricas de evaluación del modelo, como el coeficiente de determinación (R^2), el error cuadrático medio o el área bajo la curva ROC (*Receiver Operating Characteristic*) (Fawcett, 2006). Los resultados se pueden visualizar con apoyo de gráficos o tablas para facilitar la comprensión de los hallazgos considerando audiencias técnicas y no técnicas.

7. Resultados

7.1. Exploración, procesamiento y consolidación de datos

Como parte del procesamiento de datos, para el *dataset* de urgencia respiratorias reportadas, se filtraron aquellos datos con Orden de Causa 3 (Total Causas Sistema Respiratorio), lo que corresponden al total de urgencias por enfermedad respiratoria reportadas en una semana epidemiológica para un año dado. Este variable va a ser la que se usa como variable objetivo en los modelos predictores.

Todos los tipos de causas y su orden, incluyendo las hospitalizaciones generadas por estas urgencias se pueden observar en la **Tabla 1**.

Tabla 1

Orden y Causas de Urgencias

OrdenCausa	Causa
3	Total Causas Sistema Respiratorio
4	Ira Alta (J00-J06)
5	Influenza (J09-J11)
6	Neumonía (J12-J18)
7	Bronquitis/Bronquiolitis Aguda (J20-J21)
8	Crisis Obstructiva Bronquial (J40-J46)
9	Otra Causa Respiratoria (J22, J30-J39, J47, J60-J98)
10	Covid-19, Virus No Identificado U07.2
11	Covid-19, Virus Identificado U07.1
32	Hospitalizaciones por Causas Sistema Respiratorio
33	Hospitalizaciones Covid-19, Virus No Identificado U07.2
34	Hospitalizaciones Covid-19, Virus Identificado U07.1

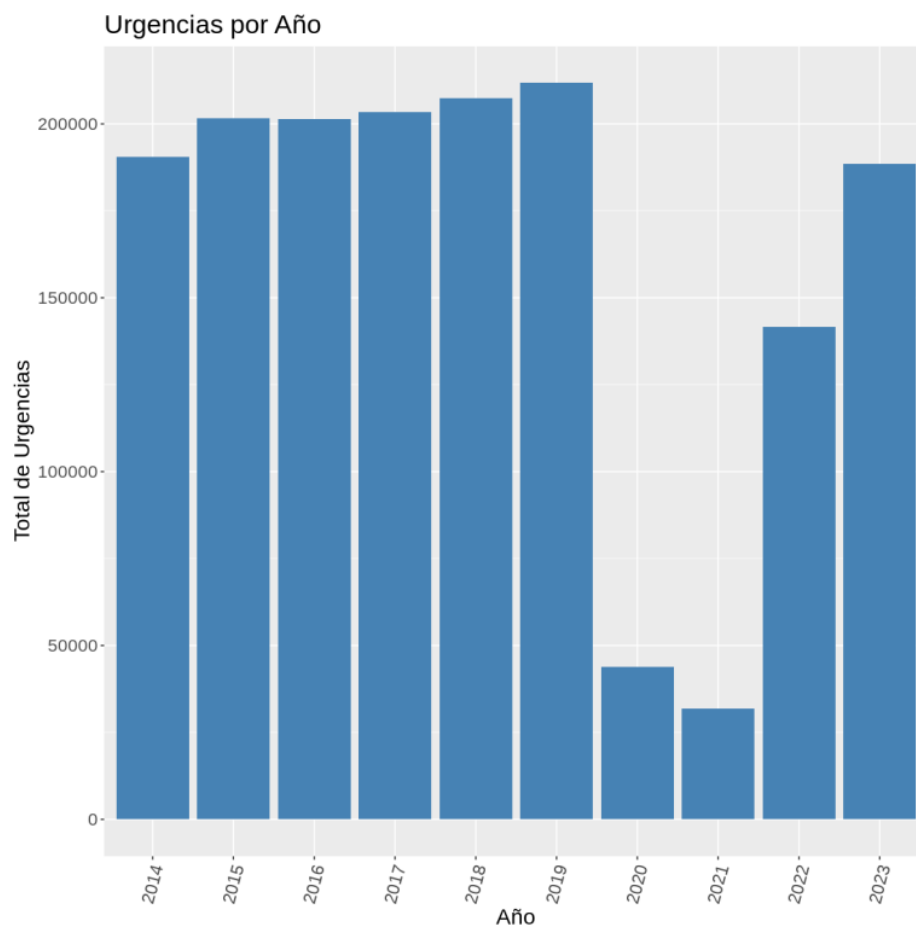
Nota: la tabla muestra todos los tipos de causas registrados en el *dataset* de urgencias respiratorias. Fuente: elaboración propia.

Como parte del análisis exploratorio de datos, luego de filtrar por el tipo de causa que, considerando el periodo 2014-2023 solo para las 3 comunas seleccionadas en esta investigación (Santiago, La Serena y Concepción), se observa que el comportamiento **Total Causas Sistema Respiratorio** por año es considerablemente bajo en los años de pandemia. Luego en los años posteriores se vuelve a incrementar el número reportado de urgencias, ver **Figura 3**. Ante esto algunos estudios indican que las medidas tomadas para evitar el contagio de COVID-19 contribuyeron a frenar la propagación de otros virus

como la gripe o el virus sincial, algo que podría explicar este comportamiento (Reina et al. , 2021).

Figura 3

Urgencias respiratorias por año

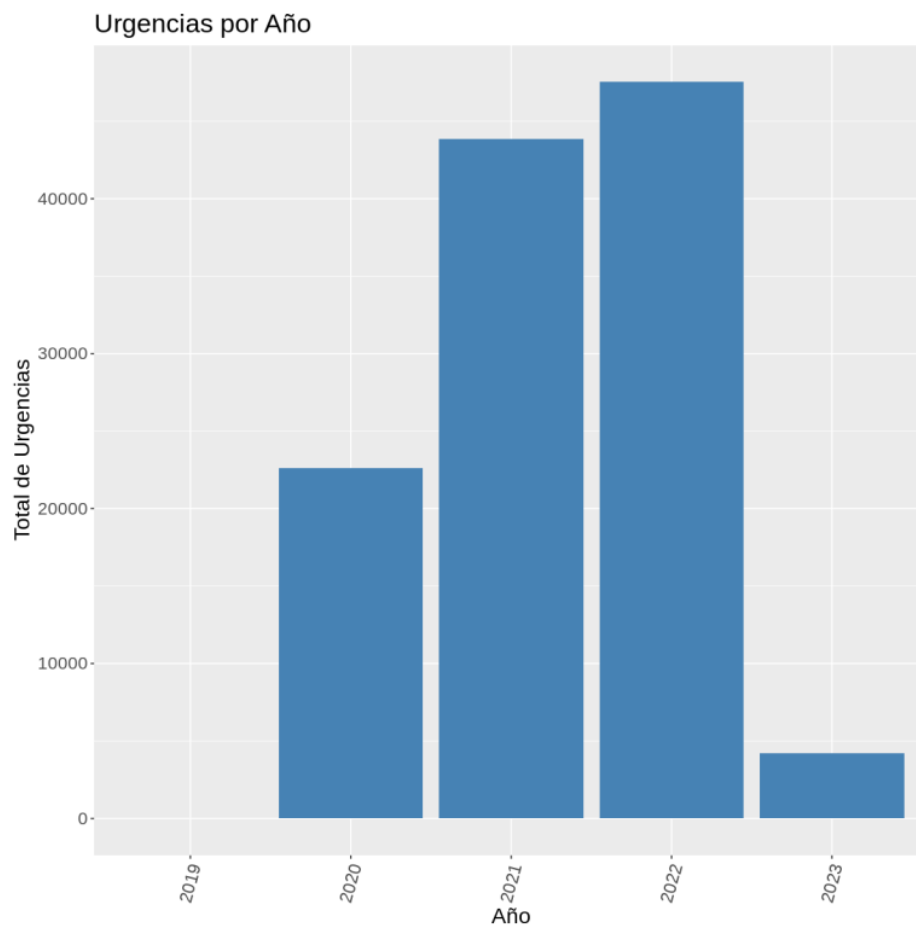


Nota: Cantidad de urgencias respiratorias reportadas para las comunas de Santiago, la Serena y Concepción. Fuente: elaboración propia.

De forma distinta, el comportamiento del virus COVID-19 durante los años de pandemia y postpandemia, si tuvo un incremento en pleno auge que terminó con una reducción a partir del 2023 ya estabilizada la propagación del virus. Ver **Figura 4**.

Figura 4

Urgencias COVID-19 por año



Nota: Cantidad de urgencias respiratorias por COVID 19 reportadas para las comunas de Santiago, la Serena y Concepción. Fuente: elaboración propia.

Ahora como parte de la exploración se puede observar en detalle las urgencias reportadas por cada una de las 3 comunas (Santiago, La Serena y Concepción) durante los años 2014-2023.

Tabla 2

Detalle urgencias anuales por comuna

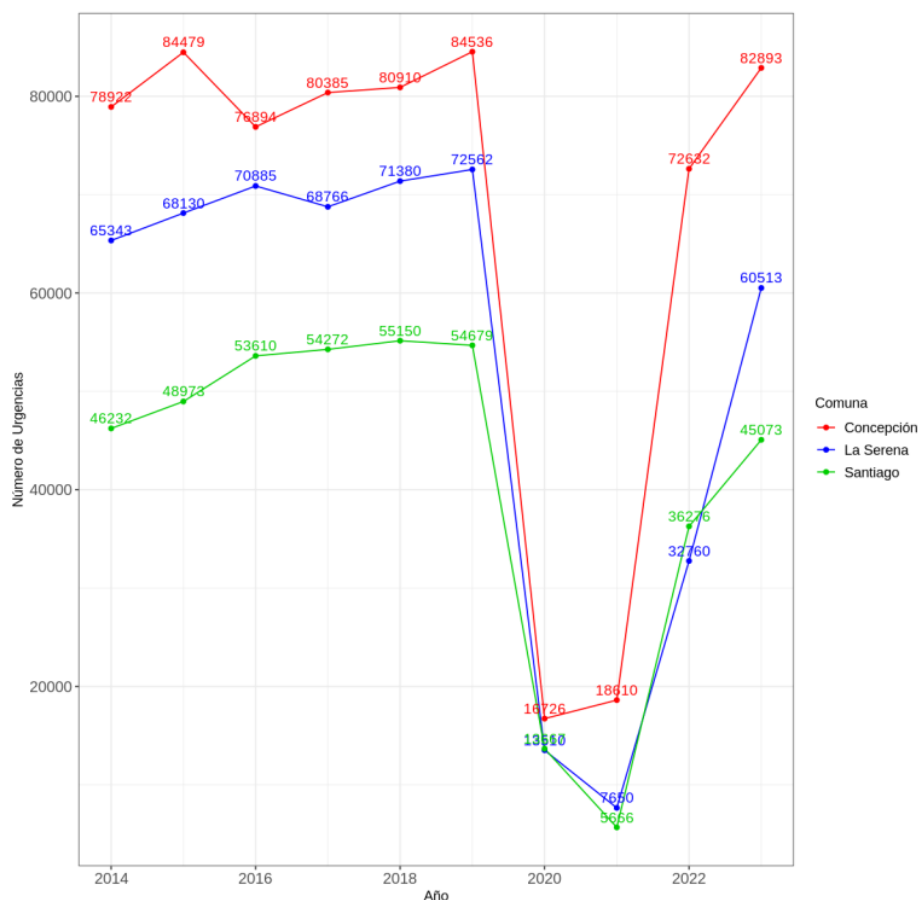
Comuna	Año	Media Semanal	Total Urgencias Anual	Mínimo Anual	Máximo Anual
Concepción	2014	1489	78922	574	2667
	2015	1594	84479	383	2659
	2016	1479	76894	607	2475
	2017	1546	80385	594	2887
	2018	1556	80910	685	2595
	2019	1626	84536	597	2794
	2020	316	16726	102	1508
	2021	351	18610	48	806
	2022	1397	72632	550	2762
	2023	1594	82893	514	3022
La Serena	2014	1233	65343	351	2254
	2015	1285	68130	318	2010
	2016	1363	70885	555	2363
	2017	1322	68766	507	2649
	2018	1373	71380	528	2511
	2019	1395	72562	437	2736
	2020	255	13510	50	1381
	2021	144	7650	37	341
	2022	630	32760	145	1391
	2023	1164	60513	376	2356
Santiago	2014	872	46232	284	1819
	2015	924	48973	203	1749
	2016	1031	53610	401	2000
	2017	1044	54272	383	2155
	2018	1061	55150	481	1763
	2019	1052	54679	411	2020
	2020	258	13667	65	1287
	2021	107	5666	34	205
	2022	698	36276	97	1207
	2023	867	45073	388	1527

Nota: La tabla muestra la cantidad de urgencias anuales por cada comuna, incluyendo la media de semana epidemiológica. Fuente: elaboración propia.

En resumidas cuentas, el comportamiento de las urgencias por enfermedades respiratorias por cada año, tiene el siguiente comportamiento observable en la **Figura 5**.

Figura 5

Total Anual de Urgencias Respiratorias por Comuna



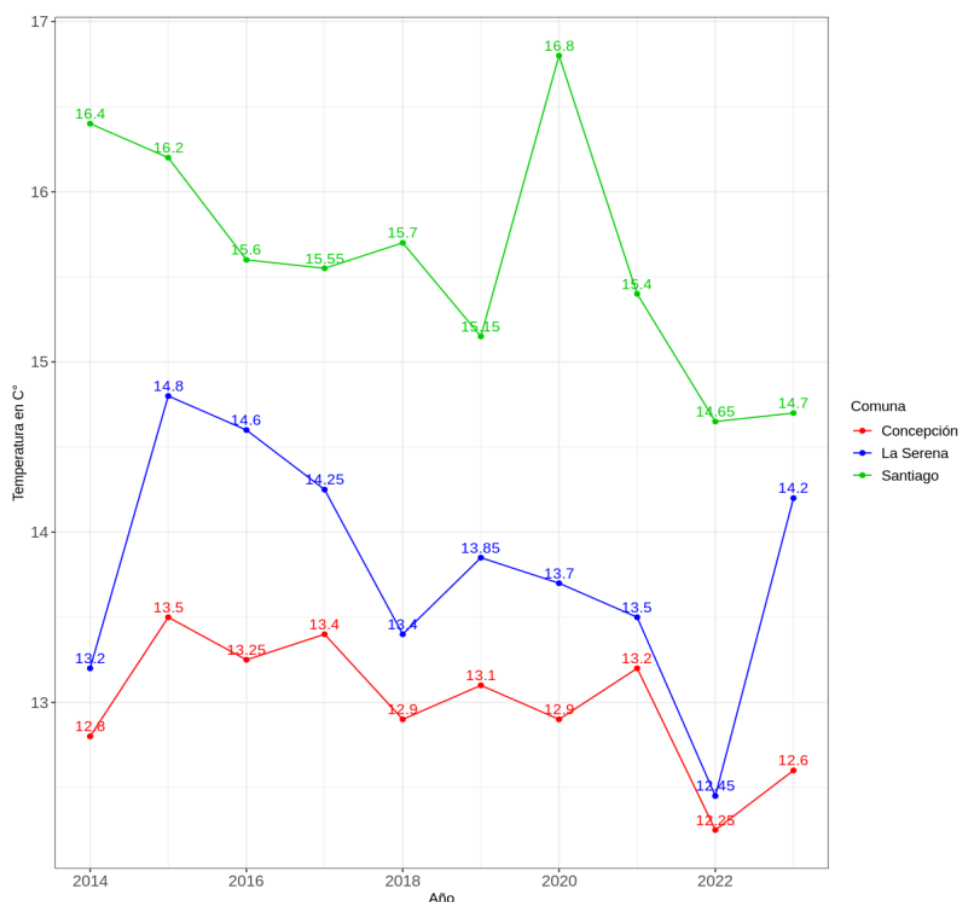
Nota: La figura muestra la evolución de números de urgencias respiratorias reportadas por cada comuna entre los años 2014-2023. Fuente: elaboración propia.

Luego se exploraron los datos de las variables climatológicas encontrando diferentes comportamientos dependiendo de la comuna, lo que se menciona como parte de la diversidad de climas en Chile (Sarricolea, Herrera-Ossandon, & Meseguer-Ruiz, 2016).

Primero está la mediana de temperatura anuales, en la **Figura 6** se puede observar claramente que Santiago es la que tiene la más alta temperatura.

Figura 6

Mediana Temperatura Anual por Comuna



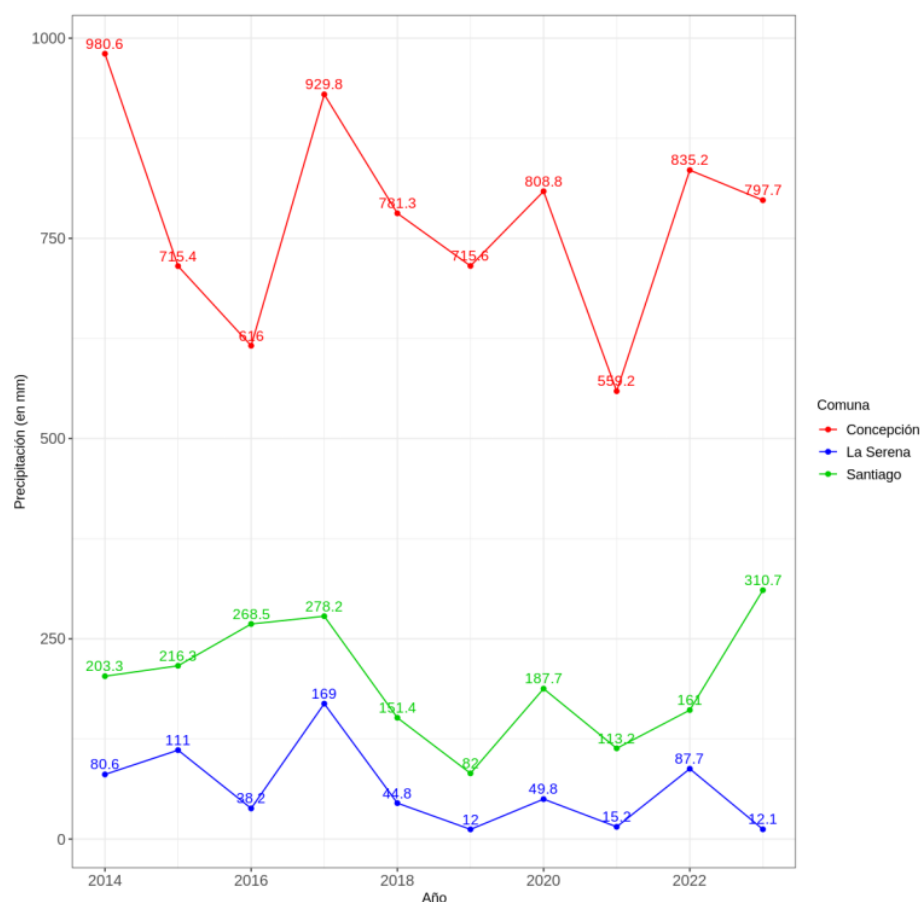
Nota: Mediana anual de las temperaturas por cada una de las comunas. Fuente: elaboración propia.

De igual forma se puede observar la cantidad anuales precipitaciones en las 3 comunas **Figura 7**, mostrándose considerablemente alta en la comuna de Concepción, ubicada al sur de Chile, zona geográfica caracterizada por tener abundantes lluvias en el periodo

invernal. Esto en contraste con Santiago y La Serena, comunas ubicadas en zonas donde las precipitaciones son moderadas a escasas (centro y norte).

Figura 7

Total Precipitación Anual por Comuna



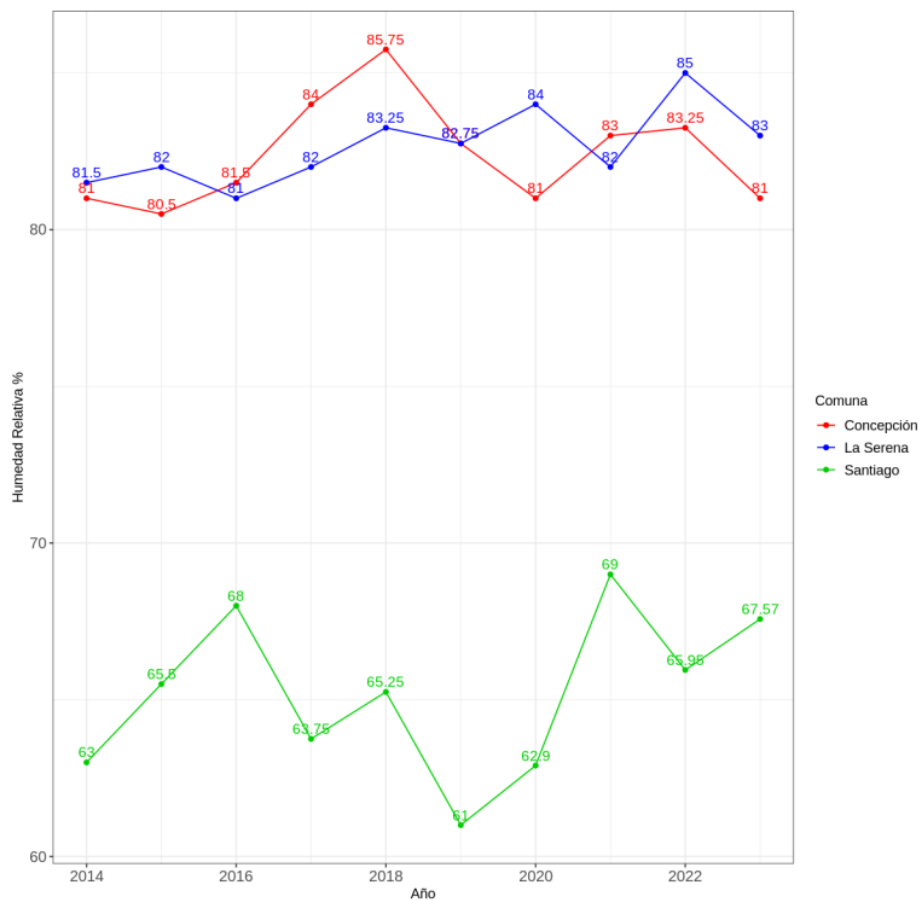
Nota: Total de precipitación anual (medido en milímetros) por cada una de las comunas.

Fuente: elaboración propia.

En el caso de la humedad relativa, se puede observar la diferencia que existe entre las comunas costeras (Concepción y La Serena) en contraposición a Santiago, esta última ubicada geográficamente en un valle rodeado de cadenas montañosas, por lo que se caracteriza por ser moderadamente más “seco” **Figura 8**.

Figura 8

Mediana Humedad Relativa Anual por Comuna

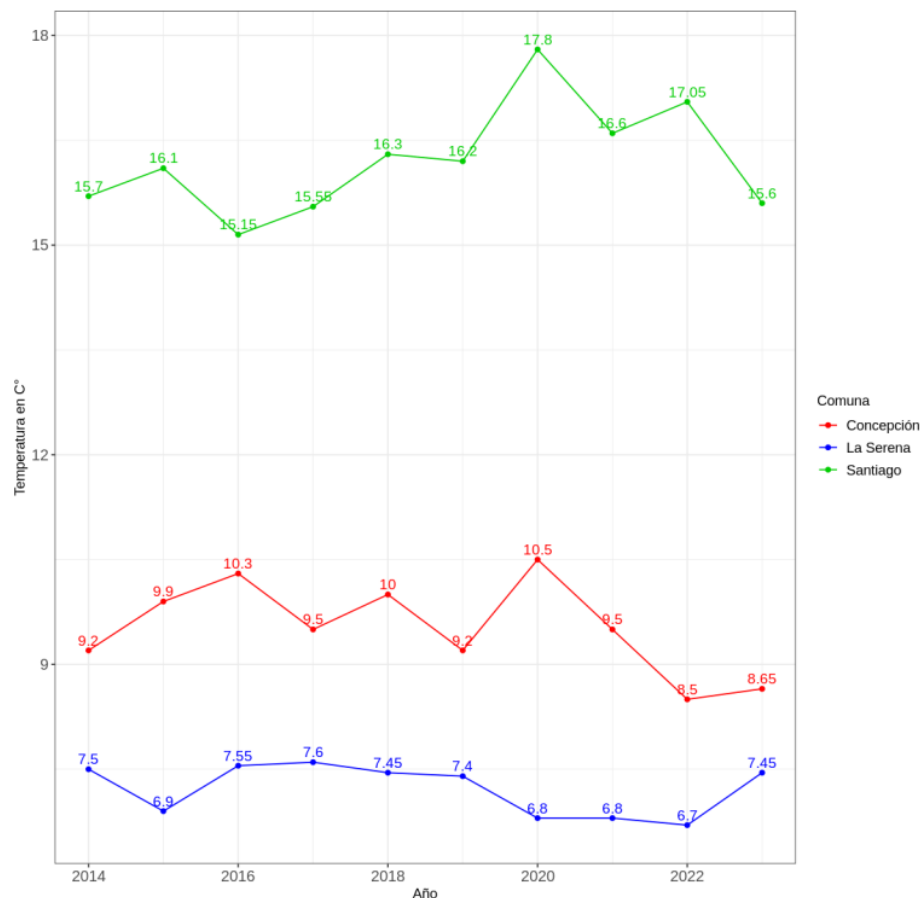


Nota: Mediana de humedad relativa anual para cada comuna. Fuente: elaboración propia.

Por otra parte, se puede visualizar que la diferencia de temperaturas es mucho mayor en comunas como Santiago, en contraste con La Serena y Concepción donde están considerablemente más estables los cambios. Ver **Figura 9**.

Figura 9

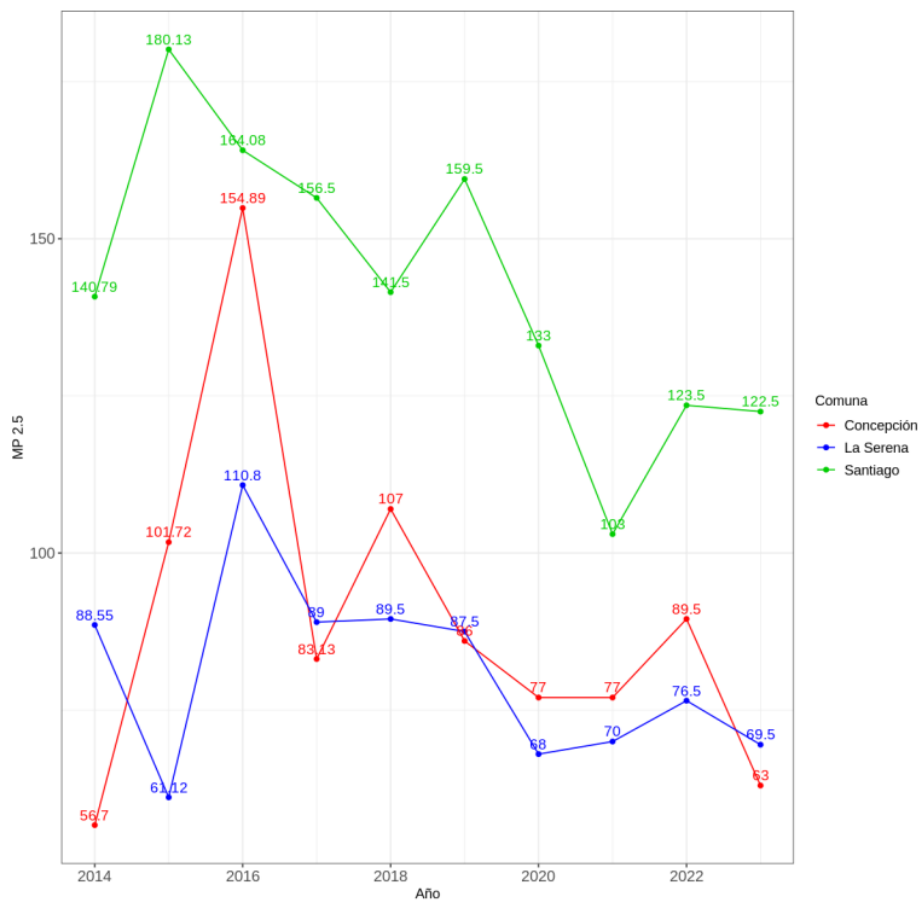
Mediana Diferencia de Temperatura Anual por Comuna



Nota: Mediana de diferencias de temperaturas anual por cada comuna. Fuente: elaboración propia.

Ya exploradas las variables climatológicas, se pone foco en observar los datos de calidad del aire, basándose en la medición del material particulado de 2.5 micrones **Figura 10**, se puede observar a la comuna de Santiago con mayor contaminación, lo que puede deberse a distintos factores como: mayor población, ubicación en un valle rodeado por cadena montañosas por lo que tiene menos ventilación en comparación con comunas costeras, mayor parque vehicular, entre otros factores.

Figura 10

Mediana Anual de Material Particulado 2.5 por Comuna

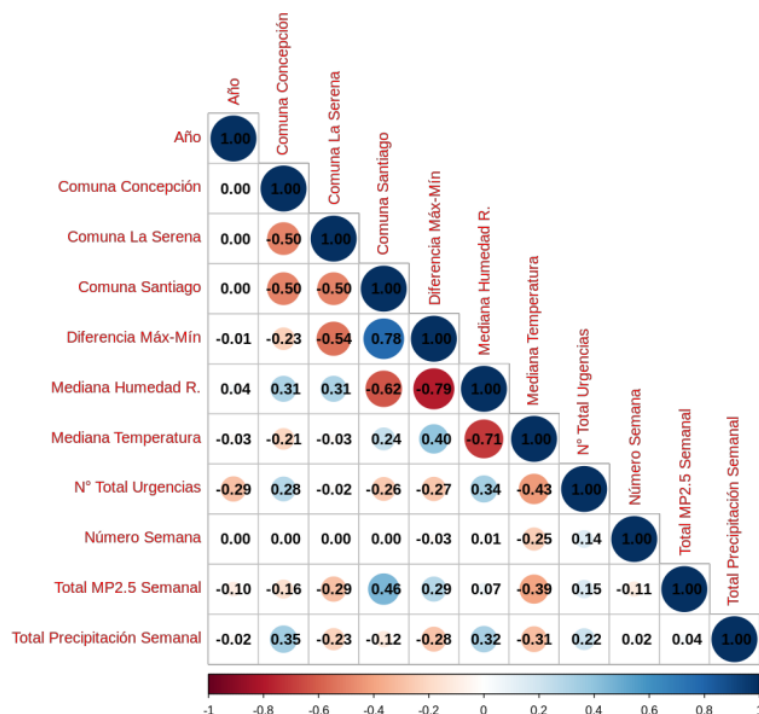
Nota: Mediana de Material Particulado 2.5 anual por cada comuna. Fuente: elaboración propia.

Por último, se realiza una última observación para indagar una probable correlación entre las variables predictoras (climatológicas y de calidad del aire) y la variable objetivo (total de urgencias), para ello se generó un gráfico de correlaciones **Figura 11** en donde en una escala de -1 a 1 se puede observar la correlación entre ellas, mientras más cercana a 0 menor es la correlación.

Para sumar la variable categórica Comuna al gráfico se convirtió en *dummies* (generando Santiago, La Serena y Concepción) por lo que su interpretación tiene que ser en el contexto de lo que realmente interpreta esa variable y no de un número como el resto de las variables.

Figura 11

Correlación entre variables



Nota: Mediana de diferencias de temperaturas anual por cada comuna. Fuente: elaboración propia.

Se puede observar que el **N° Total de Urgencias** tiene una correlación más fuerte con **Temperatura** (-0.43) y la **Humedad Relativa** (0.34), y en menor medida con **Diferencia Máx-Mín** (-0.27). Estos valores pueden sugerir que el número de urgencias respiratorias

podría estar más influenciado por factores meteorológicos que por los factores de contaminación, como lo es la variable **Total MP2.5 Semanal**, pero este análisis previo se debe supeditar a la premisa estadística que dice que la “*correlación no necesariamente implica una casualidad*”, pero es algo a considerar para futuros análisis o selección de características.

Como se ha mencionado, tras procesar los datos, cada *dataset* (meteorológicos, calidad del aire y urgencias respiratorias) fue agrupado en semanas epidemiológicas, para alinearlo con la variable objetivo (número de urgencias) y finalmente todo fue consolidado en una *dataset* final de 9 variables y 1572 observaciones con las siguientes características.

Tabla 3

Características Dataset final

Variable	Rol	Tipo	Descripción
NumTotalUrg	Objetivo	Númerica	Número total de urgencias en una semana epidemiológica.
Anio	Predictora	Númerica	Año de registro del dato.
NuSemana	Predictora	Númerica	Semana epidemiológica.
COMUNA	Predictora	Categórica	Santiago, La Serena o Concepción.
medianaClim	Predictora	Númerica	Temperatura en C°.
difMaxMin	Predictora	Númerica	Temperatura en C°.
Mediana_HR_Valor	Predictora	Númerica	Medida en %.
Total_Valor_semPP	Predictora	Númerica	Cantidad en milímetros (mm).
Total_Valor_semMP25	Predictora	Númerica	Medida de material particulado.

Nota: Detalle de características del *dataset* con el cual se van a entrenar los modelos.

Fuente: elaboración propia.

7.2. Algoritmo de regresión lineal

Con los datos explorados, procesados y consolidados en un único *dataset*, se comenzó por entrenar el modelo con la regresión línea usando la librería *Caret*¹¹ dentro del entorno de programación R, montado sobre *Google Colaboratory*.

Se separó el *dataset* consolidado en 70% entrenamiento y 30% prueba, no se requirió convertir la variable categórica **Comuna** (tipo factor en R) a variable dummies, ni

²

¹¹ <https://cran.r-project.org/web/packages/Caret/index.html>

tampoco separar un tercer subconjunto de validación debido a que la librería *Caret* maneja internamente este tipo de variables. A su vez con la función `trainControl()` permite realizar la validación cruzada (*cross-validation*) durante el entrenamiento, dividiendo el conjunto de entrenamiento en 5 pliegues que de forma iterada van actuando como validador, hasta obtener el mejor promedio de rendimiento y los parámetros óptimos, dándole robustez y evitando un sobreajuste.

Adicionalmente, haciendo uso de la función `varImp()` de *Caret*, se obtiene la importancia de las variables predictoras para el modelo, en una escala de 0 a 100 en donde 100 es lo más importante, lo que es útil para seleccionar y mejorar aquellas más idóneas para optimizar la capacidad predictora. En este caso los resultados son los siguientes:

Tabla 4

Importancia de variables regresión lineal

Variable	Importancia
Anio	12.495
NuSemana	2.344
medianaClim	7.885
COMUNA.Santiago	6.753
COMUNA.LaSerena	4.731
difMaxMin	3.479
Total_Valor_semMP25	3.752
Mediana_HR_Valor	3.423
Total_Valor_semPP	1.826

Nota1: la tabla muestra el valor de importancia de cada variable para el modelo de regresión lineal. Fuente: elaboración propia.

Nota2: La variable COMUNA, *Caret* la transformó en solo 2 variables y no en 3 *dummies*, esto para evitar el problema de colinealidad perfecta o multicolinealidad.

Luego haciendo uso del comando `print (modelo_lineal)`, se evaluó el rendimiento del modelo entrenado con dos métricas, R^2 , y RMSE obteniendo:

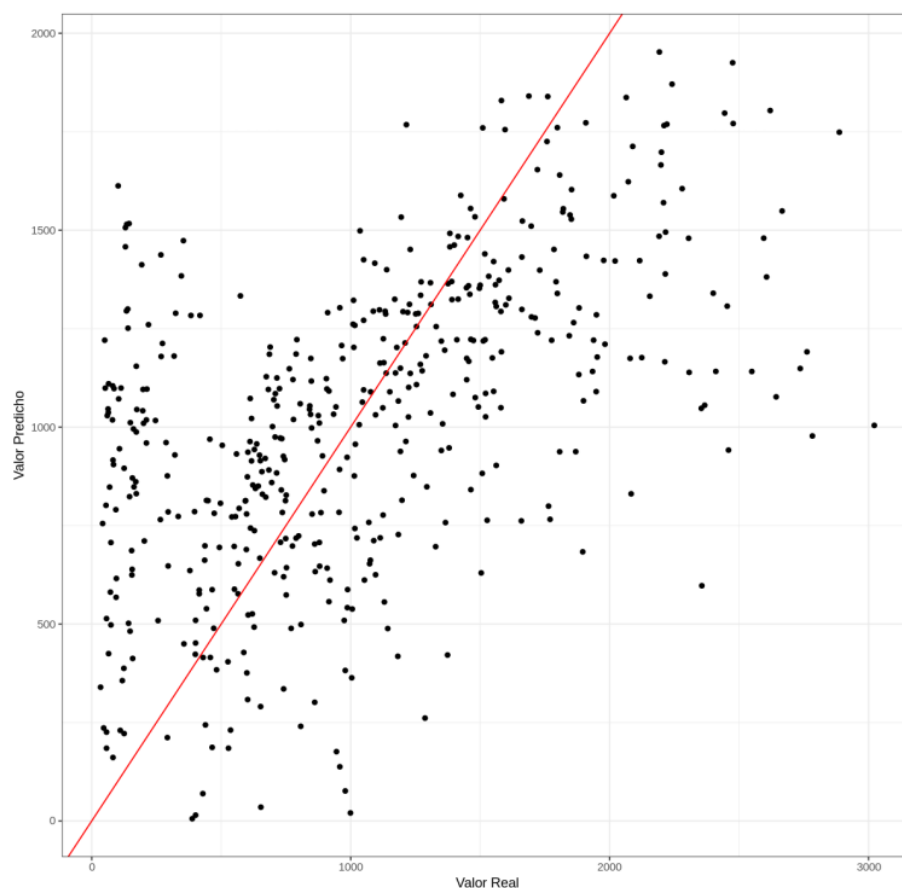
R^2 : **0.34** y RMSE: **521.95**

Lo que, en vista de la descripción de ambas métricas, se puede decir que la capacidad predictora del modelo de Regresión Lineal es pobre, ya que el R^2 es bajo por ende no captura los patrones que afectan la variable objetivo y su vez el RMSE indica una gran desviación respecto a los valores reales, algo que se puede observar en la

Figura 12 que compara valores de pruebas vs los valores predichos por el modelo.

Figura 12

Valores Predichos vs. Reales regresión lineal



Nota: Valores predichos vs reales para Regresión Lineal, mientras más cerca de la línea roja (intercepto) estén los puntos, mejor es la predicción. Fuente: elaboración propia.

7.3. Algoritmo de reforzamiento de gradiente *CatBoost*

Para usar este algoritmo, se utilizó la librería *Caret* en conjunto con la librería *CatBoost*, también en un entorno de lenguaje R montado en *Google Colaboratory*.

Particularidades de este algoritmo, es que no necesita trabajar previamente las variables categóricas, ya que maneja internamente estas de igual forma que las numéricas.

Tal como se hizo para la regresión lineal, se hace uso de la función `varImp()` de *Caret*, para obtener la importancia de las variables predictoras para el modelo, en una escala de 0 a 100. En este caso los resultados son los siguientes:

Tabla 5

Importancia de variables CatBoost

Variable	Importancia
Anio	57.794
NuSemana	13.022
medianaClim	10.253
COMUNA	8.244
difMaxMin	3.057
Total_Valor_semMP25	2.927
Mediana_HR_Valor	2.429
Total_Valor_semPP	2.273

Nota1: la tabla muestra el valor de importancia de cada variable para el modelo *CatBoost*. Elaboración propia.

Al igual que el modelo de regresión lineal, se separó el *dataframe* consolidado en 70% entrenamiento y 30% pruebas, no se requirió separar un tercer subconjunto de validación debido a que la librería *Caret* a través la función `trainControl()` realiza internamente la validación cruzada para obtener de esta forma el mejor promedio de rendimiento y los parámetros óptimos, dándole robustez y evitando un sobreajuste en el modelo.

Luego haciendo uso del comando `print(modelo_catb)` para el R^2 y el RMSE se obtuvo:

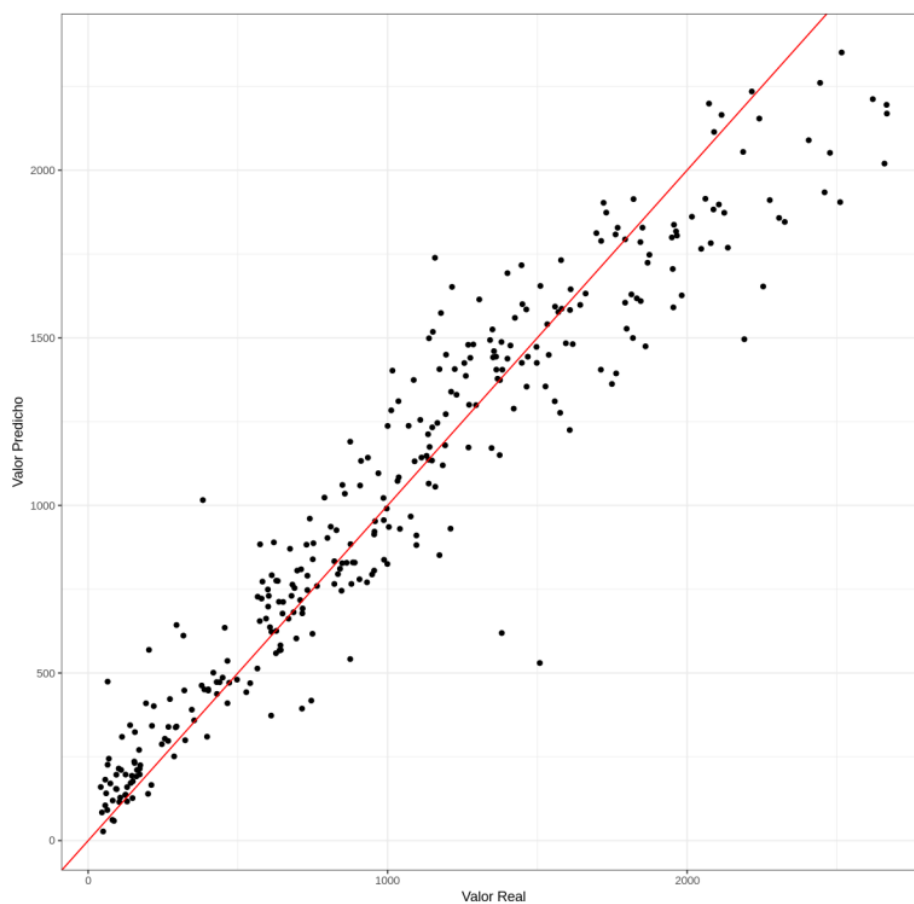
R^2 : **0.87** y RMSE: **228.77**

Lo que, en vista de la descripción de ambas métricas, se puede decir que la capacidad predictora del modelo de *CatBoost* es buena, ya que el R^2 es alto (cercano a 1) por ende captura de buena forma los patrones que afectan la variable objetivo y su vez el RMSE indica una moderada desviación respecto a los valores reales, algo que se puede observar en la

Figura 13 que compara valores de pruebas vs los valores predichos por el modelo.

Figura 13

Valores Predichos vs. Reales CatBoost



Nota: Valores predichos vs reales para *Catboost*, mientras más cerca de la línea roja (intercepto) estén los puntos, mejor es la predicción. Fuente: elaboración propia.

8. Discusión

En este estudio se usaron datos públicos y anonimizados de 3 fuentes distintas, datos con el número de urgencias respiratorias en el sistema público, datos climatológicos y de calidad del aire. Por otra parte, se propuso el uso de 2 algoritmos de aprendizaje automático, para entrenar 2 modelos distintos y comparar su desempeño en cuanto a su capacidad de predecir el número de urgencias respiratorias. Los algoritmos utilizados fueron la Regresión Lineal y *Catboost*, el primero un algoritmo simple que prevé los valores continuos de una variable en relación las variables predictoras, el segundo algoritmo del tipo *boosting*, o que de forma iterada combina una serie de árboles (de débil desempeño) con la finalidad de ir minimizando el error en cada iteración y por ende mejorar la predicción del modelo.

8.1. Interpretación de Resultados

Observando los resultados obtenidos para cada modelo, se observa que *CatBoost* tuvo un rendimiento superior en la predicción del número de urgencias de enfermedades respiratorias en comparación con el modelo de regresión lineal. Mientras que la Regresión Lineal logró explicar un 34% de la variabilidad de los datos ($R^2 = 0.34$), el modelo de *CatBoost* alcanzó un 87% ($R^2 = 0.87$). A su vez el RMSE (error cuadrático medio) para la Regresión Lineal fue **521.95** y para el modelo entrenado con *CatBoost* **228.77**, estos valores se interpretan como la precisión absoluta del modelo en la predicción, en el fondo cuánto en promedio se desvía la predicción del modelo respecto a los valores reales de la variable objetivo, en este caso del número de urgencias respiratorias (comparando el mismo tipo de unidades), lo que muestra una mejor capacidad para capturar patrones complejos en los datos, lo que es clave en contextos donde intervienen múltiples factores y variables que afectan en el aumento de la incidencia de enfermedades respiratorias.

En definitiva, los resultados indican una mejor capacidad para capturar patrones complejos en los datos por parte de *Catboost*, lo que es clave en contextos donde intervienen múltiples factores y variables que afectan en el aumento de la incidencia de enfermedades respiratoria, esto demuestra que modelos de reforzamiento de gradiente

son más efectivos para analizar fenómenos con alta variabilidad y dependencia no lineales entre los factores.

8.2. Comparación con Estudios Previos

Este estudio coincide con lo observado en el de (Ku et al., 2022) en donde los modelos de aprendizaje automático usados demostraron su capacidad para generar acertadas predicciones relacionadas con la salud pública. Sin embargo, a diferencia de aquella investigación anterior, en el presente estudio el uso de *CatBoost* mostró una mejora significativa en la capacidad de predecir ($R^2 = 0.87$ vs. $R^2 = 0.68$). Por otra parte, indicando que la elección del algoritmo de aprendizaje automático es fundamental en la precisión de predicción de este tipo de eventos, al igual que la selección de características con las cuales se entrenan.

Adicionalmente el uso de datos de distintas zonas geográficas, con variadas condiciones climáticas y de calidad del aire, permite a este estudio obtener una mejor generalización del modelo en comparación con el realizado el 2022, que solo contemplaba una única zona geográfica (Seúl).

8.3. Limitaciones del Estudio

Las principales limitaciones del estudio fue la falta de acceso a históricos anteriores al año 2014, datos como la calidad del aire, humedad relativa o precipitaciones, y esto se debe a que la instituciones que sirvieron de fuente no poseen o no median los datos en aquellos años, especialmente en comunas fuera de la región metropolitana que es donde se encuentran ubicadas la comuna de Concepción y La Serena.

La limitación de datos podría tener una influencia en el entrenamiento de los modelos y sus resultados de predicción, por lo que de contar con una mayor amplitud de históricos podría ser algo beneficioso.

8.4. Futuras líneas de investigaciones

Para futuros estudios se podría incluir variables adicionales, como el viento, la presión atmosférica, radiación UV, material particulado de 10 micrones o la tasa de vacunación en las comunas, esto podría mejorar la precisión de los modelos.

Se podría considerar realizar un enfoque más detallado en la predicción, ya teniendo la el *dataset* y la metodología de trabajo, se puede pensar en predecir el número de urgencias respiratorias por semana epidemiológica, pero por cada enfermedad específica y no solo el número total.

También sería útil probar otros modelos de aprendizaje automático, *ADA Boost*, *XG boost* o incluso redes neuronales profundas, las que podrían capturar relaciones aún más complejas en los datos entregados.

Finalmente, sumar más regiones geográficas al estudio, permitiría validar la aplicabilidad del modelo y hacerlo aún más generalizable, incluso pensando en ser entrenado con datos de otro país con condiciones distintas a Chile.

9. Conclusiones

Un gran desafío que tiene el sistema de salud público es anteponerse y estar preparado frente a alta demandas de urgencias, haciendo uso eficiente de sus recursos, sean estos infraestructura, personal o insumos. Las enfermedades respiratorias año tras año provocan colapso del sistema lo que genera una reducción en la calidad de la atención.

Los resultados de esta investigación demuestran la efectividad de los modelos de aprendizaje automático en la predicción de número de urgencias de enfermedades respiratorias, lo que tras su implementación podría beneficiar la planificación de servicios de salud durante los períodos de alta demanda, mejorando el uso de los recursos y otorgando una mejor asistencia a la población.

En un comienzo se expuso las diferencias existentes para las 3 comunas (Santiago, La Serena y Concepción) respecto a las variables consideradas para entrenar el modelo, notando diferencias climatológicas entre zonas geográficas (Norte, Centro, Sur, Costa y Valle), así como se evidenció un comportamiento de baja de urgencias respiratorias durante los años de pandemia por COVID-19 para las comunas estudiadas.

Luego se logró entrenar, evaluar y probar para predecir el número de urgencias de enfermedades respiratorias, dos modelos de aprendizaje automático, la Regresión Lineal y el *CatBoost*. Este último se logró determinar mediante el uso de métricas como R^2 y RMSE, es el modelo que brinda mayor precisión en las predicciones con un $R^2 = 0.87$ y un RMSE: **228.77**.

También durante el entrenamiento de los modelos, se logró obtener la importancia predictora de las variables para cada modelo, siendo año y mediana de temperatura las más relevantes para ambos modelos entrenados. Este ejercicio es útil para en un futuro seleccionar nuevas variables o descartar aquellas intrascendentes al momento de alimentar el modelo.

Estos resultados aportan evidencia valiosa al campo de la predicción en salud pública, sugiriendo que los algoritmos de aprendizaje automático pueden ser herramientas efectivas en la planificación de recursos hospitalarios.

10. Referencias

Bibliografía

- Behinaein, P., Hutchings, H., Knapp, T., & Okereke, I. (2023). The growing impact of air quality on lung-related illness: a narrative review. *15*(9), 5055–5063. Obtenido de <https://doi.org/10.21037/jtd-23-544>
- Biblioteca del Congreso Nacional de Chile. (2023). *Ley 19628 SOBRE PROTECCION DE LA VIDA PRIVADA*. Obtenido de <https://bcn.cl/2eqfn>
- Cifuentes Martínez, P., Rodríguez Fernández, A., Luengo M., C., & Tapia O., L. (2020). Relación entre contaminación atmosférica y consultas por enfermedades respiratorias en atención primaria de urgencia. *Revista Chilena de En-fermedades Respiratorias*. *36*(4), 260-267. Obtenido de <https://doi.org/10.4067/s0717-73482020000400260>
- Croft, D., Zhang, W., Lin, S., & et al. (2019). The Association between Respiratory Infection and Air Pollution in the Setting of Air Quality Policy and Economic Change. *16*(3), 321-330. Obtenido de <https://doi.org/10.1513/AnnalsATS.201810-691OC>
- Cusidó, J., Comalrena, J., Alavi, H., & Lluñas, L. (2022). Predicting Hospital Admissions to Reduce Crowding in the Emergency Departments. *12*(21). Obtenido de <https://doi.org/10.3390/app122110764>
- Dorogush, A., Ershov, V., & Gulin, A. (2018). CatBoost: gradient boosting with categorical features. Obtenido de <https://doi.org/10.48550/arXiv.1810.11363>
- Fawcett, T. (2006). An introduction to ROC analysis. *27*(8), 861-874. Obtenido de <https://doi.org/10.1016/j.patrec.2005.10.010>
- Freund, Y., & Schapire, R. (1999). A Short Introduction to Boosting. *14*(5), 771-780. Obtenido de <https://cseweb.ucsd.edu/~yfreund/papers/IntroToBoosting.pdf>
- Instituto Nacional de Estadística. (2017). *Resultados CENSO 2017*. Obtenido de <http://resultados.censo2017.cl/Home/Download>
- Ku, Y., Kwon, S., Yoon, J.-H., Mun, S.-K., & Chang, M. (2022). Machine Learning Models for Predicting the Occurrence of Respiratory Diseases Using Climatic and Air-Pollution Factors. *15*(2), 168–176. Obtenido de <https://doi.org/10.21053/ceo.2021.01536>
- Lee, H., & Yoon, H.-Y. (2024). Impact of ambient temperature on respiratory disease: a case-crossover study in Seoul. *25*(73). Obtenido de <https://doi.org/10.1186/s12931-024-02699-0>
- Manorathna, R. (19 de 12 de 2020). *k-fold cross-validation explained in plain English (For evaluating a model's performance and hyperparameter tuning)*. Obtenido de https://www.researchgate.net/publication/348237224_k-fold_cross-validation_explained_in_plain_English_For_evaluating_a_model%27s_performance_and_hyperparameter_tuning
- Mosegui, G., Antoñanzas, F., & de Mello Vianna, C. (2023). Cost of lost productivity from acute respiratory infections in South America. *Re-vista Panamericana de Salud Pública*. *47*. Obtenido de <https://doi.org/10.26633/RPSP.2023.65>

- Organización Mundial de la Salud. (2021). *Directrices mundiales de la OMS sobre la calidad del aire: materia particulada (MP2.5 y MP10), ozono, dióxido de nitrógeno, dióxido de azufre y monóxido de carbono: resumen ejecutivo*. Obtenido de <https://iris.who.int/handle/10665/346062>
- Oyarzún G., M., & Valdivia C., G. (2021). Impactos en la salud de la contaminación del aire. *Revista Chilena de Enfermedades Respiratorias*. 37(2), 103-106. Obtenido de <https://doi.org/10.4067/s0717-73482021000200103>
- Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A., & Gulin, A. (2019). CatBoost: unbiased boosting with categorical features. Obtenido de <https://arxiv.org/abs/1706.09516>
- Reina, J., Arcay, R., Busquets, M., & Machado, H. (2021). Impacto de las medidas higiénicas y de distanciamiento social frente al SARS-CoV-2 sobre las infecciones respiratorias causadas por otros virus. 34(4), 365–370. Obtenido de <https://doi.org/10.37201/req/017.2021>
- Romero, H., Irrazábal, F., Opazo, D., Salgado, M., & Smith, P. (2010). Climas urbanos y contaminación atmosférica en Santiago de Chile. 36(109), 35-62. Obtenido de <https://doi.org/10.4067/s0250-71612010000300002>
- Sarricolea, P., Herrera-Ossandon, M., & Meseguer-Ruiz, Ó. (2016). Climatic regionalisation of continental Chile. 13(2), 66-73. Obtenido de <https://doi.org/10.1080/17445647.2016.1259592>
- Schober, P., & Vetter, T. (2020). Linear Regression in Medical Research. 132(1), 108–109. Obtenido de <https://doi.org/10.1213/ANE.00000000000005206>
- Vandini, S., Corvaglia, L., Alessandrini, R., Aquilano, G., Marsico, C., Spinelli, M., . . . Faldella, G. (2013). Respiratory syncytial virus infection in infants and correlation with meteorological factors and air pollutants. 39(1). Obtenido de <https://doi.org/10.1186/1824-7288-39-1>
- Water Resources Management. (2023). A Comprehensive Experimental and Computational Investigation on Estimation of Scour Depth at Bridge Abutment: Emerging Ensemble Intelligent Systems - Scientific Figure on ResearchGate. Obtenido de https://www.researchgate.net/figure/The-flow-diagram-of-the-CatBoost-model_fig3_370695897
- Yang, P.-H., Hsieh, M.-T., Lin, G.-M., Chen, M.-J., Yeh, C.-H., Huang, Z.-X., & Yang, C.-M. (2018). Prediction of Outpatient Visits for Upper Respiratory Tract Infections by Machine Learning of PM2.5 and PM10 Levels in Taiwan. 1-2. Obtenido de <http://dx.doi.org/10.1109/ICCE-China.2018.8448613>
- Yang, W., & Marr, L. (2012). Mechanisms by Which Ambient Humidity May Affect Viruses in Aerosols. 78(19), 6781–6788. Obtenido de <https://doi.org/10.1128/AEM.01658-12>
- Zhang, L., Shi, Z., Chen, M., Chen, Y., Cheng, J., Fan, L., . . . Jin, Z. (2022). Study design of deep learning based automatic detection of cerebrovascular diseases on medical imaging: a position paper from Chinese Association of Radiologists. 2(4), 221–229. Obtenido de <https://doi.org/10.1016/j.imed.2022.07.001>

11. Anexos

El código fuente del presente trabajo se encuentra en la plataforma GitHub, la ejecución del código se realizó sobre un notebook en Google Colaboratory. A continuación, el enlace al proyecto en GitHub:

<https://github.com/cazathoth/tfm>

12. Declaración obligatoria del uso de herramientas de IA

El autor y el tutor de este trabajo de investigación declaran el uso de las herramientas ChatGPT y Gemini durante la fase de codificación y redacción del presente documento. Estas herramientas se emplearon con el propósito de generar ideas y proporcionar apoyo en la redacción, optimizando el proceso de elaboración del trabajo. Sin embargo, es importante destacar que el uso de estas tecnologías fue complementario al trabajo intelectual del autor y del tutor, ya que ambos participaron activamente en la revisión y edición del contenido generado.

Tras la utilización de ambas herramientas, el autor y el tutor revisaron minuciosamente tanto los códigos como los contenidos redactados, haciendo las correcciones y ajustes necesarios. Por lo tanto, la presente memoria es completamente de su autoría, y ambos asumen plena responsabilidad por el contenido expuesto.

ORIGINALITY REPORT

2%

SIMILARITY INDEX

2%

INTERNET SOURCES

0%

PUBLICATIONS

%

STUDENT PAPERS

PRIMARY SOURCES

1

datos.gob.cl

Internet Source

1%

2

repository.udistrital.edu.co

Internet Source

1%

Exclude quotes On

Exclude bibliography On

Exclude matches < 1%