UNIVERSIDAD DE CONCEPCIÓN

FACULTAD DE INGENIERÍA

DEPARTAMENTO DE INGENIERÍA INDUSTRIAL

Profesor Patrocinante:

Rodrigo de la Fuente

Informe de Memoria de Título

Para optar al Título de:

Ingeniera Civil Industrial

Desarrollo de modelos predictivos para enfermedades respiratorias y cardiovasculares en zonas pobladas de la región del Biobío

Concepción, Agosto del 2019

María Ignacia Carrasco Huaiquián

UNIVERSIDAD DE CONCEPCIÓN
Facultad de Ingeniería
Departamento de Ingeniería Industrial

Profesor Patrocinante:
Rodrigo de la Fuente

Desarrollo de modelos predictivos para enfermedades respiratorias y cardiovasculares en zonas pobladas de la región del Biobío

María Ignacia Carrasco Huaiquián

Informe de Memoria de Título
Para optar al Título de

Ingeniera Civil Industrial

Agosto 2019

Agradecimientos

Mis mas profundos agradecimientos a mis padres, por quererme, cuidarme y enseñarme casi todo lo que sé. A mi madre, por el excelente modelo a seguir que ha sido toda mi vida; y a mi padre, que jamás ha dudado de mi. Mi querida Lily, que me enseño las cosas más importantes de la vida; y a mi profesor Rodrigo, por su sabiduría y acompañamiento durante esta etapa que en algunos momentos pareció interminable.

Resumen

La contaminación del aire es un problema que afecta no solo a los humanos, sino a todo el ecosistema, deteriorando la salud de cada ser vivo y contribuyendo al cambio climático. En la región del Biobío, al igual que en otras zonas del sur del pais, se registran elevados niveles de material particulado, debido principalmente a la combustión de leña a nivel residencial. Diversas investigaciones han demostrado correlaciones positivas entre efectos adversos en la salud y un aumento en la concentración de material particulado (Tonne et al., 2012; Raaschou-Nielsen et al., 2013; Zhang et al., 2014). El objetivo de este trabajo es construir un modelo para predecir la tasa de ingresos hospitalarios y atenciones de urgencia asociadas a patologías cardiovasculares y respiratorias. En virtud de lo anterior es necesario desarrollar una base de datos, la que se construyó con información sobre las condiciones meteorológicas y de calidad de del aire para la región del Biobío; donde el principal problema fue la cantidad de valores pérdidos, los que se completaron utilizando la imputación múltiple. Para realizar las predicciones fueron desarrollados tres algoritmos. Se elaboró una red neuronal artificial simple (*multilayer perceptron*), un árbol de regresión (*XGBoost*) y un modelo lineal generalizado. Se utilizó la validación cruzada hacia adelante para obtener las predicciones y analizar el desempeño de los algoritmos, y modelos de optimización sequencial (tree parzen estimators) para obtener los hiperparámetros.

Debido al tiempo computacional necesario para ejecutar los algoritmos en una base de datos tan extensa, estos se probaron en una mas pequeña, previamente desarrollada en otra investigación, con información de la comuna de Los Ángeles desde el 2013 al 2017. Los resultados considerados relevantes fueron pocos. El algoritmo XGboost logró explicar el 12,33 % de la varianza para las atenciones de urgencia por enfermedades respiratorias para el grupo etario 3; y el GLM fue capaz de explicar el 48,55 % y el 21,48 % de la varianza en los grupos 2 y 3, respectivamente. El principal problema de esta base de datos fue la gran cantidad de observaciones perdidas, las que fueron completadas con la media, sesgando las estimaciones al homogeneizar la base de datos y reducir su varianza. Es por esto que los resultados no se consideran concluyentes en cuanto a reflejar la capacidad de los algoritmos descritos se refiere.

Índice general

Índice de figuras III

Índice de tablas IV

1. Introducción y objetivos **1**
1.1. Introducción 1
1.2. Objetivos 2
1.3. Descripción del Problema 3

2. Marco conceptual **4**
2.1. Contaminación 4
2.1.1. Contaminantes atmosféricos y sus fuentes de emisión 4
2.1.2. Material particulado (MP) 5
2.1.3. Ozono (O_3) 10
2.1.4. Dióxido de nitrógeno (NO_2) 10
2.1.5. Dióxido de azufre (SO_2) 11
2.1.6. Monóxido de carbono (CO) 11
2.2. Sistema de Salud 12
2.3. Descripción de la zona de estudio 14
2.3.1. Diagnóstico de calidad del aire 16

3. Marco teórico **18**
3.1. Preprocesamiento de la información 18
3.1.1. Imputacion múltiple a través de ecuaciones encadenadas 19
3.2. Redes Neuronales 19
3.2.1. Topología de la red neuronal 21
3.3. Entrenamiento 22

3.3.1. Función de pérdida . . . 24
3.3.2. Optimización . . . 25
3.4. Predicción . . . 26
3.4.1. Hiperparámetros . . . 27
3.5. XGBoost . . . 30
3.6. Modelos Lineales Generalizados . . . 31

4. Base de Datos **32**
4.1. Diagnóstico de variables . . . 32
4.2. Reopilación de información . . . 32
4.3. Creación de Base de Datos . . . 33
4.4. Pre-procesamiento de los datos . . . 33
4.4.1. Material Particulado . . . 33
4.4.2. Análisis de variables de salud . . . 38

5. Resultados y Discusión **39**
5.1. Resultados de los modelos . . . 45
5.1.1. Enfermedades respiratorias . . . 45
5.1.2. Enfermedades cardiovasculares . . . 49
5.1.3. Discusión . . . 50

6. Conclusiones y trabajo futuro **52**

7. Referencias **53**

7. Anexos **58**
7.1. Mapa de flujo servicio de salud público . . . 58
7.2. Lista de contaminantes y estaciones de calidad del aire . . . 59
7.3. Diccionarios . . . 60
7.4. Correlaciones entre estaciones de monitoreo para el material particulado . . . 63
7.5. Cantidad de valores perdidos $MP_{2,5}$ y MP_{10} . . . 65
7.5.1. Registros no validados (2000-2018) . . . 65
7.5.2. Registros validados (2000-2018) . . . 66
7.5.3. Registros validados (2008-2017) . . . 67
7.6. Hiperparámetros . . . 68
7.7. Coeficiente de determinación . . . 69

Índice de figuras

2.1. Ejemplo de la composición y distribución del material particulado 6
2.2. Proporción de emisiones por cada tipo de fuente emisora sobre total de emisiones medidas en toneladas para el año 2017 11
2.3. Organigrama sistema de salud de Chile 13
2.4. Mapa Región del Biobío y Ñuble 15
2.5. Emisiones anuales de $MP_{2,5}$ por región y fuente para el año 2015 17

3.1. Comparación entre una neurona en biológica y una neurona artificial 20
3.2. Validación hacia adelante 23
3.3. A la izquierda una red neuronal estándar y a la derecha después de aplicar dropout . . 23

4.1. Evaluación de las concentraciones diarias ambientales de $MP_{2,5}$ µg/m3 35
4.2. Evaluación de las concentraciones diarias ambientales de MP_{10} µg/m3 35
4.3. Gráfico valores perdidos $MP_{2,5}$ (a la izquierda) y MP_{10} (a la derecha) (2000-2018) . . 37
4.4. Gráfico valores perdidos de $MP_{2,5}$ (a la izquierda) y $MP_{2,5}$ (a la derecha) (2008-2017) 37
4.5. Urgencias por enfermedades Respiratorias para el grupo etario 2 38

5.1. Estructura modelo para ingresos hospitalarios por enfermedades respiratorias G1 . . 41
5.2. Importancia relativa atenciones de urgencia por enfermedades respiratorias G1 43
5.3. Árbol de decisión para atenciones de urgencia por enfermedades respiratorias G1 . . 43
5.4. Predicciones semanales XGboost 44
5.5. Predicciones IG1_R, utilizando MLP con TimeSeriesSplit 47

7.1. Mapa de Flujo 58

Índice de tablas

2.1. Descripción de los principales contaminantes atmosféricos químicos y sus fuentes . . 5
2.2. Efectos adversos de los contaminantes aéreos sobre el sistema respiratorio 8
2.3. Efectos no respiratorios de los contaminantes atmosféricos 9
2.4. Mortalidad y morbilidad asociada a la exposición a $MP_{2,5}$ durante el 2015 16

3.1. Número de imputaciones necesarias 20
3.2. Ventajas y desventajas de funciones de pérdida para problemas de regresión 25
3.3. Softwares de Modelos de optimización sequencial 28

4.1. Lista de variables 34

5.1. Hiperparámetros MLP 40
5.2. Hiperparámetros XGBoost 42
5.3. Error cuadrático medio de los distintos modelos 46

7.1. Lista de compuestos y moléculas por estación 59
7.2. Diccionario de base de datos egresos hospitalarios 60
7.3. Diccionario de base de datos atenciones de urgencia 62
7.4. Correlaciones $MP_{2,5}$ para el promedio diario 63
7.5. Correlaciones MP_{10} para el promedio diario 64
7.6. Cantidad de valores perdidos $MP_{2,5}$ y MP_{10} (2000-2018).
gistros no validados) 65
7.7. Cantidad de valores perdidos $MP_{2,5}$ y MP_{10} (2000-2018).
gistros validados) 66
7.8. Cantidad de valores perdidos $MP_{2,5}$ y MP_{10} (2008-2017).
gistros validados) 67
7.9. Valores finales hiperparámetros 68
7.10. Coeficiente de determinación de los distintos modelos 69

Capítulo 1
Introducción y objetivos

1.1. Introducción

Todos los días y con cada inhalación las personas estan expuestas a una serie de contaminantes, muchos de estos de origen natural, los que no suelen ser perjudiciales para la salud(Palacios et al., 1997). Sin embargo, la expansión de las ciudades y los actuales patrones de consumo han provocado un aumento en los requerimientos energéticos y en la emisión de contaminantes. Los medios de transporte, la calefacción de las viviendas y las actividades industriales se identifican como las tres grandes fuentes de contaminación en Chile. Asimismo, la actividad productiva de algunos sectores ha sido crucial en la generación de problemas de contaminación en varias zonas del país; es por ello que actualmente la calidad del aire constituye una prioridad en la gestión medioambiental y se han adoptado varias medidas como el establecimiento de normas para las principales fuentes industriales emisoras de contaminantes, el aumento de estaciones de monitoreo en el país y restricciones al transporte público y vehículos generales, entre otras, para mejorar la calidad del aire (Ministerio del Medio Ambiente, 2017).

Estas medidas se relacionan directamente con el aumento de la disponibilidad de datos medioambientales, los que junto con la información clínica de libre acceso, han hecho del estudio de la calidad del aire y sus efectos en la salud un tema factible de ser investigado; sobre todo con la expansión y disponibilidad de nuevos métodos computacionales para el análisis de grandes sets de datos. Minería de datos (*data mining*) se le llama al proceso de analizar y descubrir patrones en los datos (Witten et al., 2016); las herramientas utilizadas en estos procesos encuentran su origen en diferentes campos como la estadística, la inteligencia artificial, las matemáticas y el aprendizaje de máquinas (Bellinger et al., 2017). El aprendizaje de máquinas (*machine learning*) es la ciencia y arte de programar computadoras para que puedan aprender de los datos (Géron, 2017), y es de esta forma y con las herramientas descritas que se abordará el tema en este trabajo.

1.2. Objetivos

Objetivo General

Desarrollar modelos predictivos para enfermedades respiratorias y cardiovasculares en las zonas pobladas de la región del Biobío.

Objetivos Específicos

1. Elaboración de una base de datos georreferenciada para la región del Biobío.
2. Desarrollo de un modelo predictivo basado en redes neuronales.
3. Desarrollo de un modelo predictivo basado en árboles de decisión.
4. Análisis comparativo del desempeño de un modelo estadístico tradicional versus nuevas técnicas relativas al aprendizaje de máquinas.

1.3. Descripción del Problema

La contaminación del aire es un problema que afecta no solo a los humanos, sino a todo el ecosistema, deteriorando la salud de cada ser vivo y contribuyendo al cambio climático. En la región del Biobío, al igual que en otras zonas del sur del pais, se registran altos niveles de concentraciones de material particulado, debido principalmente a la combustión de leña a nivel residencial (Ministerio del Medio Ambiente, 2017). Estudios epidemiológicos han demostrado correlaciones positivas entre efectos adversos en la salud y un aumento en la concentración de material particulado (Lighty et al., 2000). Considerando lo anterior resulta interesante estudiar la relación entre la calidad del aire, las variables meteorológicas, y sus efectos en la salud. Fernández (2018) estudió estas relaciones y demostró que existen vínculos significativos entre un aumento en las concentraciones de material particulado y los ingresos hospitalarios en la comuna de Los Ángeles, así como tambien Mardones et al. (2015), él que además logró cuantificar los beneficios económicos de reducir el material particulado en el Gran Concepción .

El desafío de este trabajo es recopilar la información disponible sobre la calidad del aire y variables meteorológicas, para crear una base de datos que permita construir un modelo que describa los ingresos hospitalarios y atenciones de urgencia en las principales comunas de la Región del Biobío. Utilizando Redes Neuronales se espera poder captar las relaciones subyacentes entre las variables, que expliquen de la mejor manera las fluctuaciones de las atenciones en el servicio público debido a la contaminación atmosférica; y con esta información, eventualmente, poder prever situaciones críticas y ofrecer una atención mas eficiente a la población. El alcance de este trabajo considera los efectos en el sistema respiratorio y cardiovascular, ya que en ambos, se han demostrado efectos adversos por la contaminación atmosférica.

Capítulo 2
Marco conceptual

2.1. Contaminación

La contaminación se define como la presencia en el ambiente de sustancias, elementos, energía o combinación de ellos, en concentraciones y/o permanencia superiores o inferiores, según corresponda, a las establecidas en la legislación vigente (Chile, 1994) y que provoca inestabilidad y daña el funcionamiento de un ecosistema. En relación a lo anterior se considera como contaminación atmosférica la presencia de elementos contaminantes en la atmósfera que alteran su composición y afectan a cualquier componente del ecosistema. Desde un punto de vista antropocéntrico se considera solo a aquellos contaminantes que afectan a la salud o bienestar humano (Oyarzún, 2010). La Organización Mundial de la Salud (OMS) la considera como una de las prioridades mundiales más importantes en salud, ya que mediante la disminución de los niveles de contaminación del aire los países pueden reducir la carga de morbilidad derivada de accidentes cerebrovasculares, cánceres de pulmón y neumopatías crónicas y agudas, entre ellas el asma. Además se estima que la contaminación ambiental del aire, tanto en las ciudades como en las zonas rurales, causa mas de 4,2 millones de muertes prematuras cada año (OMS, 2018). Los componentes mas riesgosos para la salud humana son el material particulado inhalable y compuestos químicos gaseosos tales como dióxido de nitrógeno, ozono, dióxido de azufre y monóxido de carbono (Oyarzún, 2010).

2.1.1. Contaminantes atmosféricos y sus fuentes de emisión

Los principales contaminantes del aire son el material particulado (MP), el ozono (O_3), los óxidos de nitrógeno (NO_x) (especialmente el dióxido de nitrógeno (NO_2)), el dióxido de azufre (SO_2), el monóxido de carbono (CO) y los compuestos orgánicos volátiles (COVs). Además, los contaminantes pueden clasificarse en primarios y secundarios, siendo primarios aquellos que proceden directamente de la fuente de emisión y secundarios aquellos que se forman como consecuencia de las transformaciones y reacciones químicas y físicas que sufren los contaminantes primarios y algunas especies no

contaminantes en la atmósfera. Las características de los principales contaminantes químicos y sus fuentes se resumen en la Tabla 2.1.

Tabla 2.1: Descripción de los principales contaminantes atmosféricos químicos y sus fuentes

Contaminante	**Formación**	**Estado físico**	**Fuentes**
Partículas en suspensión (MP): MP_{10}, Humos negros	Primaria y secundaria	Sólido, líquido	Vehículos, Procesos industriales, Humo del tabaco
Dióxido de azufre (SO_2)	Primaria	Gas	Procesos industriales, Vehículos
Dióxido de nitrógeno (NO_2)	Primaria y secundaria	Gas	Vehículos, Estufas y cocinas de gas
Monóxido de carbono (CO)	Primaria	Gas	Vehículos, Combustiones en interiores, Humo del tabaco
Compuestos orgánicos volátiles (COVs)	Primaria y secundaria	Gas	Vehículos, Industria, Humo del tabaco, Combustiones en interiores
Plomo (Pb)	Primaria	Sólido (partículas finas)	Vehículos, Industria
Ozono(O_3)	Secundaria	Gas	Vehículos (secundario a foto-exidación de NO_x y COVs)

Fuente: Oyarzún (2010)

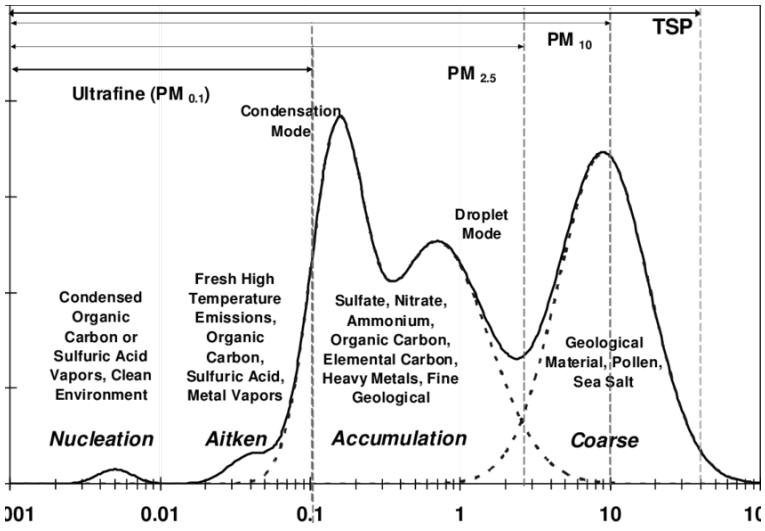
Los contaminantes provienen de diferentes actividades, tanto de fuentes naturales como antropogénicas . Entre las actividades naturales se encuentran las emisiones volcánicas, los incendios forestales, la erosión del suelo, los procesos de pudrición de materia orgánica, entre otros (Semmartin, 2013). Sin embargo, las consecuencias negativas provenientes de estas fuentes no se comparan a las de origen antropogénico debido sobretodo a que sus efectos a largo plazo son menores (Palacios et al., 1997). En las Tablas 2.2 y 2.3 se ilustran algunos efectos adversos de los principales contaminantes atmosféricos.

2.1.2. Material particulado (MP)

El material particulado, también llamado material inhalable, es un indicador representativo común de la contaminación del aire, ya que afecta a más personas que cualquier otro contaminante (OMS, 2018). Se define como "el conjunto de partículas sólidas y/o líquidas (a excepción del agua pura) presentes en suspensión en la atmósfera[1], que se originan a partir de una gran variedad de fuentes naturales o antropogénicas y poseen un amplio rango de propiedades morfológicas, físicas, químicas

[1]Generalmente las mediciones de calidad del aire se notifican como concentraciones medias diarias o anuales de microgramos de partículas por metro cúbico de aire ($\mu g/m^3$)

y termodinámicas" (Suárez, 2012). Los principales componentes del MP son los sulfatos, los nitratos, el amoniaco, el cloruro de sodio, el hollín y los polvos minerales (OMS, 2018). Estas partículas en suspensión son una compleja mezcla de productos químicos y/o elementos biológicos. Debido a que son de tamaño, forma y composición variada se han clasificado en términos de su diámetro aerodinámico, que corresponde al diámetro de una esfera uniforme en unidad de densidad que alcanza la misma velocidad terminal de asentamiento que la partícula de interés y que está determinado por la forma y densidad de la partícula. De acuerdo a lo anterior éstas pueden ser clasificadas como partículas ultrafinas, finas o gruesas, si es que su diámetro aerodinámico es menor a 0.1 μm ($MP_{0,1}$), 2.5 μm ($MP_{2,5}$) y 10 μm (MP_{10}), respectivamente (Suárez, 2012). En la Figura 2.1 se observa un ejemplo de la distribución típica de las diferentes partículas en la atmósfera[2], siendo las principales fuentes de emisión de MP_{10} y $MP_{2,5}$, los procesos mecánicos y de combustión, respectivamente (Lighty et al., 2000).



Fuente: Watson et al. (2010)

Figura 2.1: Ejemplo de la composición y distribución del material particulado

Efectos sobre la salud

Las partículas con un diámetro de 10 micrones o menos (MP_{10}) son peligrosas porque pueden penetrar y alojarse profundamente dentro de los pulmones, pero aquellas con un diámetro menor o igual 2,5 ($MP_{2,5}$) lo son aún más, ya que pueden atravesar la barrera pulmonar y entrar en el sistema

[2]El eje vertical representa la concentración de partículas (%) y el eje horizontal el diámetro de estas. Las partículas totales suspendidas (TSP por sus siglas en inglés) por alto volumen hace referencia a todos los tamaños de partículas del rango de 0 a $\sim$ 30–50 μm.

sanguíneo. Es por ello que una exposición crónica a estas partículas contribuye al riesgo de desarrollar enfermedades cardiovasculares y respiratorias (OMS, 2018; Dominici et al., 2006). Se ha comprobado que existe una correlación significativa entre un aumento en la concentración de material particulado y el espesor medio de la pared de la carótida (Tonne et al., 2012); el cáncer de pulmón (Raaschou-Nielsen et al., 2013); la mortalidad por enfermedades cardiovasculares, especialmente por cardiopatías isquémicas a largo plazo (Zhang et al., 2014) y el riesgo de defectos congénitos cardiacos, debido a la exposición durante la gestación (Agay-Shay et al., 2013), entre otros. El abanico de efectos en la salud es amplio y se ve afectada toda la población, sin embargo la susceptibilidad a la contaminación puede variar con la salud o la edad (OMS, 2006).

La OMS (2018) estima que en 2016, aproximadamente el 58 % de las muertes prematuras relacionadas con la contaminación atmosférica se debieron a cardiopatías isquémicas y accidentes cerebrovasculares, mientras que el 18 % de las muertes se debieron a enfermedad pulmonar obstructiva crónica e infecciones respiratorias agudas, y el 6 % de las muertes se debieron a cáncer de pulmón. Se ha demostrado que el riesgo de diversos efectos aumenta con la exposición, y hay pocas pruebas que indiquen un umbral por debajo del cual no quepa prever efectos adversos en la salud. Es por ello que la OMS recomienda que el proceso de fijación de normas se oriente a alcanzar las concentraciones más bajas posibles teniendo en cuenta las limitaciones, la capacidad y las prioridades en materia de salud pública en el ámbito local (OMS, 2006).

Normativa calidad del aire

La Organización Mundial de la Salud ha fijado un limite superior de $10\mu g/m^3$ para la media anual y $25\mu g/m^3$ para la media diaria (24 horas) para partículas finas ($MP_{2,5}$) y un límite de $25\mu g/m^3$ de media anual y $50\mu g/m^3$ de media diaria para las partículas gruesas (MP_{10}). Además de estos valores, las directrices sobre la calidad del aire establecen metas intermedias para las concentraciones de $MP_{2,5}$ y MP_{10} destinadas a promover una reducción gradual, de concentraciones altas a otras más bajas. Se estima que una reducción media de las concentraciones de MP_{10} de 35 $\mu g/m^3$ a 10 $\mu g/m^3$, permitiría reducir el número de defunciones relacionadas con la contaminación aproximadamente en un 15 %. Sin embargo, incluso en la Unión Europea, donde las concentraciones de MP de muchas ciudades cumplen con los niveles fijados en las directrices, se estima que la exposición a partículas de origen antropogénico reduce la esperanza media de vida en 8,6 meses (OMS, 2018). En Chile la norma exige niveles de ($MP_{2,5}$) debajo de $20\mu g/m^3$ para la media anual y $50\mu g/m^3$ para la media diaria; y $50\mu g/m^3$ para la media anual y $150\mu g/m^3$ para la media diaria de MP_{10}.

Tabla 2.2: Efectos adversos de los contaminantes aéreos sobre el sistema respiratorio

Contaminante	**Efecto a corto plazo**	**Efecto a largo plazo**
Material particulado "respirable" (MP_{10}) y fino ($MP_{2,5}$)	Aumento de morbimortalidad respiratoria Disminución en la función pulmonar Interferencia en mecanismos de defensa pulmonar: fagocitosis y depuración mucociliar Síndrome bronquial obstructivo	Menor desarrollo de la estructura y función del sistema respiratorio Mayor riesgo de cáncer en la edad adulta (HAPs [3])
Particulado ultrafino ($MP_{0,1}$)	Mayor respuesta inflamatoria. (comparado con MP_{10} y $MP_{2,5}$) Pasaje rápido a la circulación y a otros órganos	
Ozono (O_3)	Disminución de frecuencia respiratoria y disminución de CVF [4] y VEF [5] Alveolitis neutrofílica, aumento de permeabilidad e hiperreactividad bronquial Alteración del epitelio alveolar (células tipo II)	Daño de células epiteliales, "bronquiolización" alveolar Disminución del desarrollo de CVF y VEFt
Dióxido de azufre (SO_2)	Obstrucción bronquial Hipersecreción bronquial	Bronquitis crónica
Dióxido de nitrógeno (NO_2)	Hiperreactividad bronquial Aumento de síntomas respi ratorios y exacerbaciones de asma Aumenta la respuesta a la provocación con alergenos Disminución de la actividad mucociliar	Posible decremento del desarrollo pulmonar
Monóxido de carbono (CO)	Disminución en la capacidad de ejercicio	
Plomo (Pb)	Alteración del epitelio bronquiolar (células de Clara)	

Fuente: Oyarzún (2010)

Tabla 2.3: Efectos no respiratorios de los contaminantes atmosféricos

Órganos / Sistemas	**Contaminantes**	**Efectos**
Cardiovascular	Material particulado	Disminución de la variabilidad en la frecuencia cardíaca ante el estrés
	Monóxido de carbono	Interfiere el transporte de O_2 por la hemoglobina
	Plomo / Vanadio	Mayor frecuencia de hipertensión arterial en población adulta
	Ozono (O_3)	Comunicación interventricular (administración prenatal en ratas)
Unidad materno-fetal	Monóxido de carbono y $MP_{2,5}$ (hidrocarburos aromáticos policíclicos: HAP)	Bajo peso de nacimiento
		Baja talla al nacer
Sistema nervioso central y autonómico	Monóxido de carbono	Cefalea, irritabilidad, disminución de percepción auditiva y visual. Compromiso progresivo y letal de conciencia en concentraciones altas
	Plomo	Hiperquinesia, trastornos del aprendizaje; encefalopatía; cólicos intestina les
	Ozono (O_3)	Daño cerebeloso en células de Purkinje (administrado prenatalmente en ratas)
Renal	Cadmio y Vanadio Plomo	Toxicidad renal Tubulopatía
Hematopoyético	Plomo	Anemia
Óseo	Plomo	Reemplazo del Ca+2 en los huesos produciendo descalcificación

Fuente: Oyarzún (2010)

2.1.3. Ozono (O_3)

El ozono a nivel del suelo es uno de los principales componentes de la niebla tóxica. Este se forma por la reacción con la luz solar (fotoquímica) de contaminantes como los óxidos de nitrógeno (NO_x) y los compuestos orgánicos volátiles (COV), emitidos por los vehículos, los disolventes y la industria. Los niveles de ozono más elevados se registran durante los periodos de tiempo soleado. Su exceso en el aire puede producir efectos adversos considerables, es un importante factor de mortalidad y morbilidad. Puede causar problemas respiratorios, provocar asma, reducir la función pulmonar y originar enfermedades pulmonares. La OMS ha fijado un valor de $100\mu g/m^3$ de media en 8 horas para el ozono. Este nivel se establece ya que se ha demostrado una relación entre concentraciones superiores de ozono y el aumento de la mortalidad diaria (OMS, 2018).

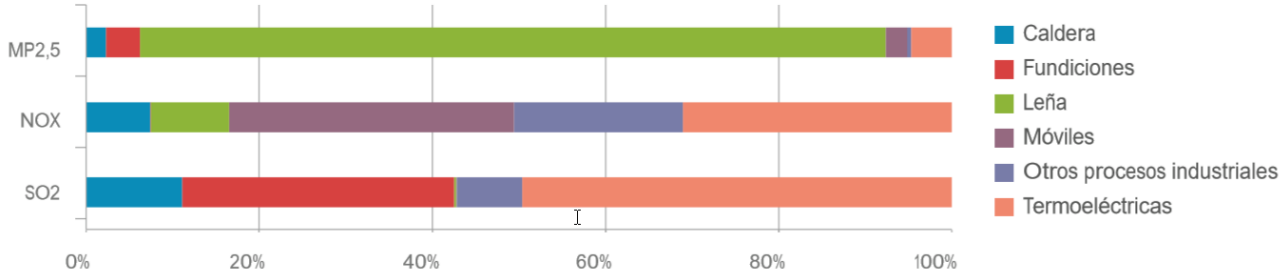
2.1.4. Dióxido de nitrógeno (NO_2)

Es un producto derivado de los procesos de combustión (calefacción, generación de electricidad y motores de vehículos y barcos) y se suele encontrar en la atmósfera íntimamente asociado con otros contaminantes primarios, como las partículas ultrafinas. Es de por sí tóxico y también es precursor del ozono, con el que coexiste junto con varios otros oxidantes generados en procesos fotoquímicos. Las concentraciones de NO_2 muestran con frecuencia una fuerte correlación con la de otros contaminantes tóxicos y, dado que es más fácil de medir, a menudo se utiliza en lugar de la mezcla completa. Por lo tanto, la obtención de concentraciones guía para un solo contaminantes, como el NO_2 puede aportar beneficios para la salud pública superiores a los previstos sobre la base de las estimaciones de la toxicidad de un solo contaminante.

Investigaciones como la Khreis et al. (2018) han demostrado que los síntomas de bronquitis en niños asmáticos se acentuan con la exposición prolongada al NO_2, así como también disminuye el desarrollo de la función pulmonar (OMS, 2018). Sin embargo, no esta claro hasta que punto estos efectos se le pueden atribuir al dióxido de nitrógeno o a otros contaminantes primarios o secundarios con lo que tiene una correlación directa. La OMS recomienda cocentraciones inferiores o iguales $40\mu g/m^3$ de media en anual y $200\mu g/m^3$ de media en una hora. La elección del valor medio en una hora se sustenta en los estudios sobre la capacidad de respuesta bronquial, la que se ve afectada con concentraciones superiores (OMS, 2006).

2.1.5. Dióxido de azufre (SO_2)

El dióxido de azufre es un gas que se forma de la oxidación del azufre, mineral que esta presente en materias primas como el petróleo, carbón, aluminio, cobre y hierro. Esta oxidación generalmente sucede durante la quema de estos combustibles o la extracción de estos metales del mineral (Mihelcic and Zimmerman, 2012). El SO_2 forma pequeñas partículas de sulfato, las que pueden afectar al sistema respiratorio y las funciones pulmonares (Mihelcic and Zimmerman, 2012). La inflamación del sistema respiratorio provoca tos, secreción mucosa, además de agravar el asma y la bronquitis crónica; asimismo, aumenta la propensión de las personas a contraer infecciones del sistema respiratorio. Los ingresos hospitalarios por cardiopatías y la mortalidad aumentan en los días en que los niveles de SO_2 son más elevados (OMS, 2006). La OMS ha fijado como recomendación un límite de 20$\mu g/m^3$ de media diaria y de 500$\mu g/m^3$ de media en 10 minutos.



Fuente: Ministerio del Medio Ambiente (2017)

Figura 2.2: Proporción de emisiones por cada tipo de fuente emisora sobre total de emisiones medidas en toneladas para el año 2017

2.1.6. Monóxido de carbono (CO)

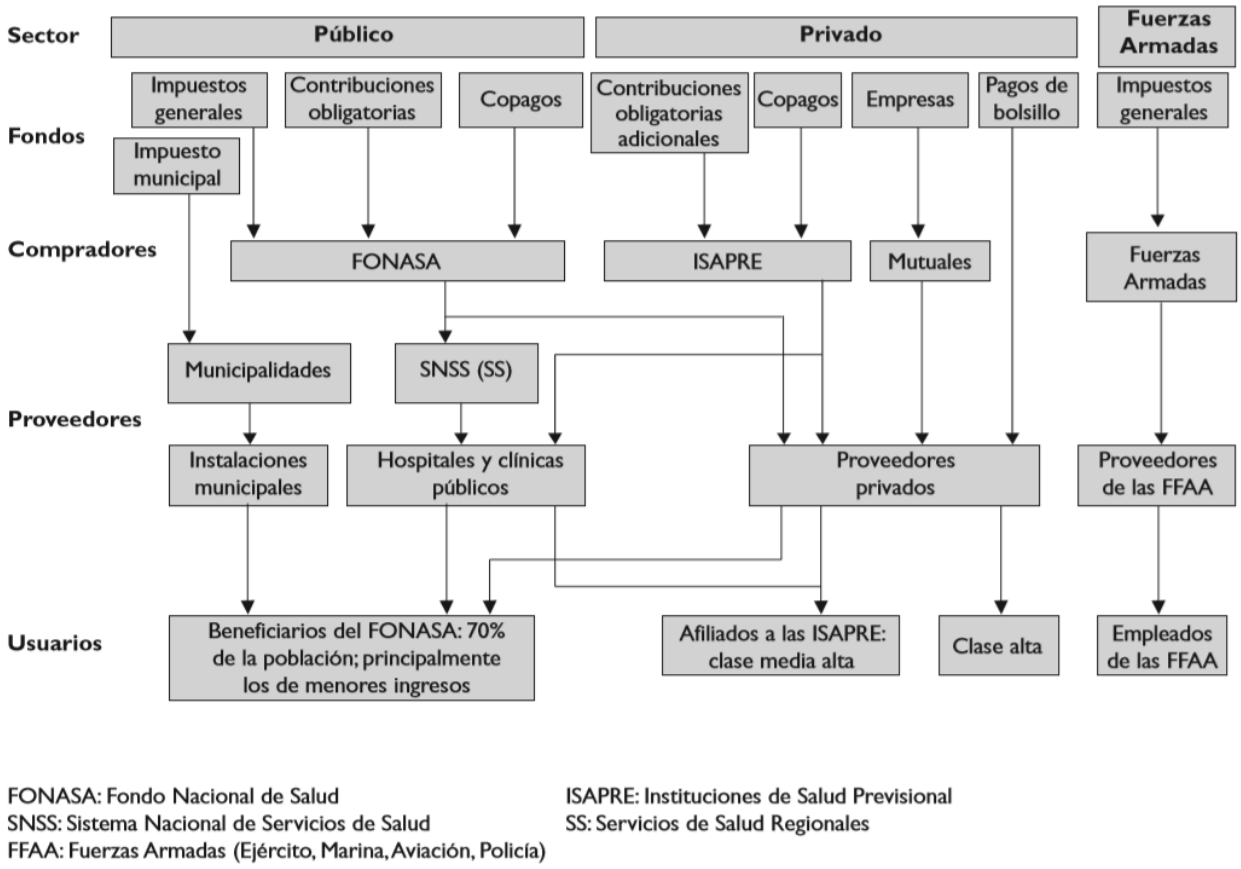
El monóxido de carbono se forma producto de una combustión incompleta, lo que sucede generalmente debido a una cantidad insuficiente de oxígeno para una determinada cantidad de combustible. Los vehículos, las plantas de energía y las estufas a leña constituyen las principales fuentes de emisión. El efecto del monóxido de carbono en el organismo puede ser letal, esto es debido a la afinidad existente entre la hemoglobina y el CO que impide el transporte de óxigeno y puede causar asfixia. Esto es especialmente peligroso en ambientes cerrados, ya que en espacios abiertos este compuesto se diluye fácilmente (Semmartin, 2013). La Norma de Calidad Primaria de Aire para Monóxido de Carbono (CO) en chile fija un límite de 30 mg/m^3 de media en una hora y de 10 mg/m^3 de media en 8 horas.

2.2. Sistema de Salud

"El sistema de salud, se refiere al conjunto formal de personas y entidades públicas y privadas que se relacionan con la organización, financiamiento, aseguramiento, recursos o provisión de bienes y servicios en materias de promoción, prevención, cuidado o recuperación de la salud" (Observatorio Chileno de Salud Pública, ndb). En el periodo entre 1952 y 2014 la salud pública logró grandes avances para la población chilena, los que se retratan en indicadores de salud como la mortalidad infantil, que disminuyo de 117,8 a 7,2 por cada 1.000 nacidos vivos; la mortalidad maternal, que pasó de 276 a 18,5 por cada 100.000 madres; la desnutrición infantil, que bajo desde un 63 % a un 0,5 % para niños menores a 5 años; y la esperanza de vida, que aumento de 50 a 79,8 años. Durante este periodo también mejoró el acceso a agua potable y alcantarillados, todo esto de la mano con el crecimiento económico del país (Becerril-Montekio et al., 2011).

Otros índices como el de analfabetismo y los años de escolaridad también mejoraron. Sin embargo, la inversión en salud pública continúa siendo una de las más bajas de los paises de la Organización para la Cooperación y el Desarrollo Económicos (OCDE), lo que se refleja en un déficit en el número de médicos (1,7 por cada 1.000 habitantes), de enfermeras (4,8 por cada 1.000 habitantes), en la cantidad de camas en los hospitales (2,1 por cada 1.000), y en la disponibilidad de medicamentos genéricos en el mercado (Goic, 2015). Actualmente, la población registrada asciende a 17.574.003 personas a nivel nacional, de las cuales 8.972.014 son mujeres y 8.601.989 son hombres; el 87,8 % de la población vive en áreas urbanas y el 12,2 % en áreas rurales (Instituto Nacional de Estadísticas, 2017). A todos ellos el estado les debe garantizar acceso a la salud, ya que este es un derecho constitucional (Honorable Junta de Gobierno, 1980). Es por ello que este tiene un rol regulador, a través del Ministerio de Salud (MINSAL) y proveedor, como se describirá a continuación.

El Sistema de Salud en Chile es de naturaleza mixta, es decir es una mezcla público/privada tanto en la previsión como en la provisión de servicios. "El sistema de salud, en su aspecto formal, incluye un doble nivel organizacional sobrepuesto: (a) El sistema nacional de salud, que incluye a todas las personas y entidades, ya sean estatales, públicas o privadas, que cumplen funciones relativas a la estructura, financiamiento, aseguramiento y funcionamiento del sistema en su conjunto; (b) El Sistema Nacional de Servicios de Salud (SNSS), que es una entidad pública con una claro marco normativo, que está centrada en la provisión de servicios asistenciales a la población, para lo cual cuenta en forma descentralizada con Servicios de Salud de ámbito regional o subregional; también participan del SNSS, aquellas instituciones que se adscriben a través de convenios, destacando los municipios y los servicios delegados"(Observatorio Chileno de Salud Pública, ndb) En la Figura 2.3 se muestra la estructura del sistema de salud chileno, ilustrando su financiamiento, y los principales actores dentro del proceso (Becerril-Montekio et al., 2011).



Fuente: Becerril-Montekio et al. (2011)

Figura 2.3: Organigrama sistema de salud de Chile

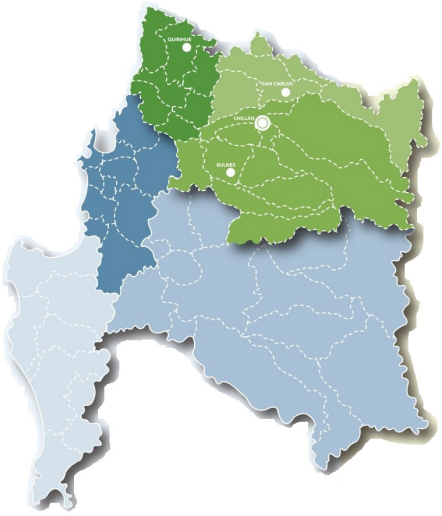
El sector público cubre a aproximadamente el 70 % de la población, mientras que el sector privado provee servicios al 17,5 % de la población (aquellos con mayores ingresos). Aproximadamente el 10 % de la población se atiende en los Servicios de Salud de las Fuerzas Armadas y un porcentaje muy reducido, perteneciente a la clase alta, realiza pagos directamente de su bolsillo a proveedores de salud privados (Becerril-Montekio et al., 2011). El SNSS cuenta con 29 Servicios de Salud territoriales. Estos estan a cargo de la articulación, gestión y desarrollo de la red asistencial para la ejecución de las acciones de fomento, protección y recuperación de la salud y rehabilitación de personas enfermas (Observatorio Chileno de Salud Pública, nda). La red asistencial se compone por un conjunto de establecimientos públicos (a cargo de las municipalidades o los servicios de salud) y privados (aquellos con convenios para prestar servicios delegados). El objetivo de esta red es actuar de forma colaborativa y oportuna para resolver de manera efectiva las necesidades de salud de la población, tanto en el ámbito individual como colectivo (Observatorio Chileno de Salud Pública, ndb).

2.3. Descripción de la zona de estudio

La Región del Biobío y la nueva Región de Ñuble constituyen la zona de estudio, en la Figura 2.4 se ilustra en tonos celestes la Región del Biobío y en tonos verdes la Región de Ñuble. De aquí en adelante se les considerará a ambas como una región (Región del Biobío) para simplificar tanto el lenguaje como el análisis debido a que los datos se encuentran agrupados teniendo las consideraciones del mapa politico previo al cambio realizado en septiembre del 2018. El territorio descrito abarca 37.068,7 km^2 y el 2017 albergaba a 2.037.414 personas, 70 % en zonas urbanas y 30 % en zonas rurales. La provincia de Concepción compuesta por las comunas de Concepción, Coronel, Chiguayante, Hualpén, Hualqui, Lota, Penco, San Pedro de la Paz, Talcahuano y Tomé, es el segundo conglomerado urbano del país. La región aporta un promedio de 7,6 % al PIB nacional ubicándose en el cuarto lugar tras la Región Metropolitana, Antofagasta y Valparaiso. Su economía es de base exportadora respaldándose principalmente en la producción forestal, pesquera e industrial. El territorio se caracteriza por un clima templado mediterraneo cálido y un clima templado húmedo o lluvioso, llegando a más de 2.400 mm de precipitaciones concentradas en invierno (Departamento de Estudios, 2015).

Servicios de Salud

En la Región del Biobío existen 507 establecimientos de salud, públicos y privados, destinados a prevenir, promover y otorgar cuidados de salud. En esta región el 84,8 % de la población se atiende en el sistema público a través de FONASA, el 2,3 % ocupa los servicios de las Fuerzas Armadas, el 9 % se atiende a través de Isapres, el 1,8 % de forma particular y un 2,1 % en otros sistemas (Biblioteca del Congreso Nacional de Chile, 2017). Existen diferentes tipos de establecimientos, los que



Fuente: Oreña (2018)

Figura 2.4: Mapa Región del Biobío y Ñuble

generalmente se clasifican como de atención primaria, secundaria o terciaria. Los primeros se dedican principalmente a la promoción y prevencion, mientras que los últimos se enfocan en el tratamiento mas especializado de enfermedades. En la Red de Urgencias se encuentran los establecimientos de Atención Primaria y las Urgencias Hospitalarias. La Red de Atención Primaria está fromada por los Servicios de Atención Primaria de Urgencia (SAPU), que trabajan en un horario complementario al de los Centros de Salud Familiar, y estan capacitados para resolver problemas de mediana gravedad; las Postas de Salud Rural (PSR), que son establecimientos de menor complejidad lozalizados en zonas de baja densidad poblacional donde se dedican a prevenir, promover, fomentar, proteger y recuperar la situación de salud de las comunidades rurales; los Servicios de Urgencia Rural (SUR), que otorgan atención de urgencia en las zonas mas apartadas en horarios extendidos; los Servicios de Urgencia de Alta Resolución (SAR), que atienden urgencias de menor y mediana complejidad; y los Servicios Médicos de Atención de Urgencia (SAMU), que ubicados en los hospitales entregan una atención de urgencia para casos de mayor complejidad (Servicio de Salud Metropolitano Sur, Ministerio de Salud, 2018b). También existen otros establecimientos de Atención Primaria como son los Centros de Salud Familiar (CESFAM), Centros Comunitarios de Salud Familiar (CECOSF) y Hospitales, que entregan atención de morbilidad y se dedican a prevenir y promover la salud en las comunidades y los Consultorios Adosados de Especialidades (CAE) (Servicio de Salud Metropolitano Sur, Ministerio de Salud, 2018a). En la Sección 7.1 de los anexos se ilustran los niveles de atención de acuerdo al nivel de complejidad necesario para cada tratamiento (Servicio de Salud Biobío, Ministerio de Salud, nd).

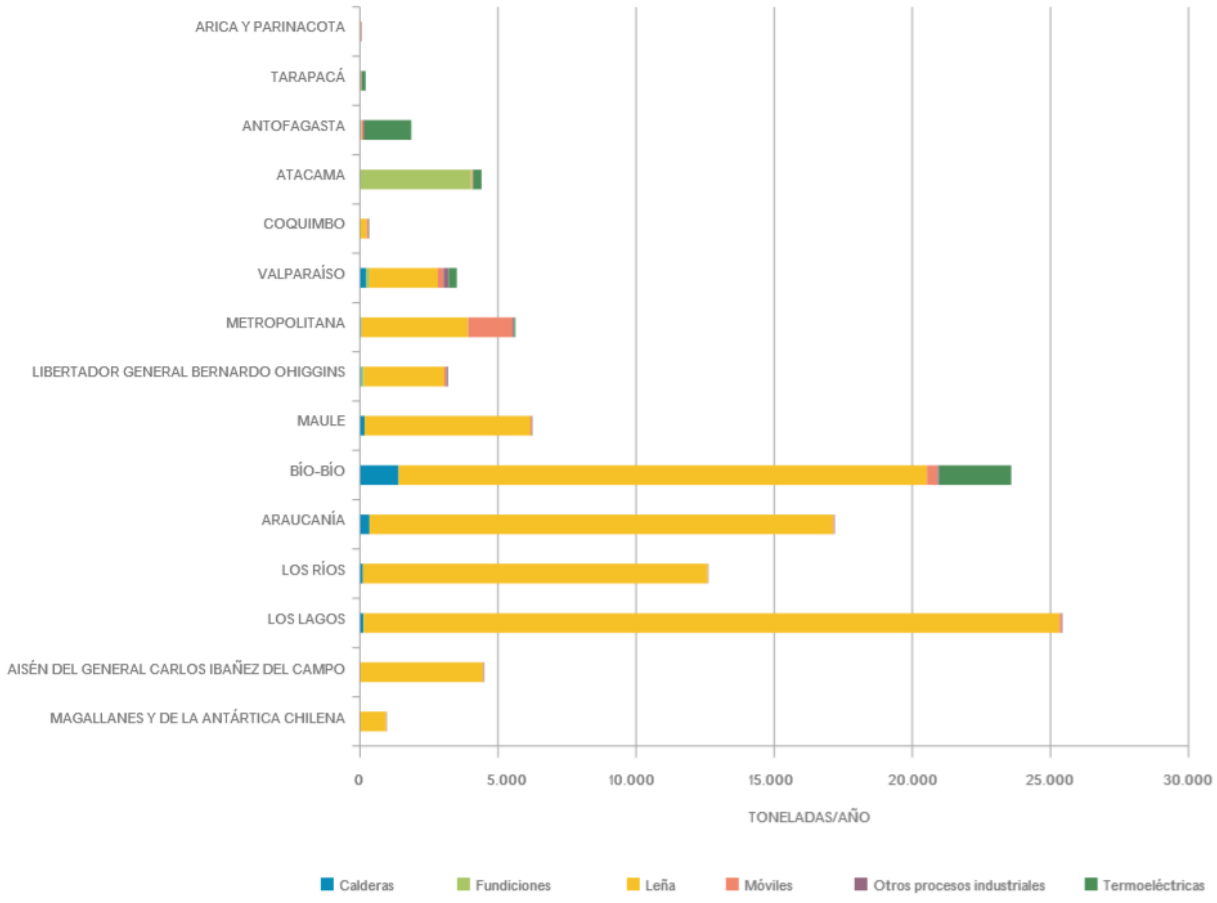
2.3.1. Diagnóstico de calidad del aire

Durante el año 2016 las estaciones 21 de mayo (Los Ángeles) y Purén (Chillán) superaron en más de un 50 % los valores anuales de la norma y más de un 200 % los valores diarios para el $MP_{2,5}$. Para el MP_{10} las estaciones de la región del Biobío se encontraron cercanas al límite de la norma anual y solo 21 de mayo y Consultorio San Vicente (Talcahuano) la superaron. Los valores diarios en las ciudades de Los Ángeles y Chillán tambien fueron sobrepasados. De lo anterior se concluye que estas dos ciudades constituyen puntos críticos para la contaminación en la región. En la Figura 2.5 se muestra que las mayores emisiones de $MP_{2,5}$ se encuentran en las regiones del Biobío y Los Lagos, superando las 23.000 toneladas, provenientes principalmentes de la combustión de leña residencial; en la Tabla 2.4 se reflejan algunas de la consecuencias mas preocupantes debido a una exposición crónica a la contaminación por $MP_{2,5}$, como lo son las admisiones hospitalarias por causas cardiovasculares y respiratorias, el ausentismo laboral y escolar, que se estiman tienen una repercusión en la productividad nacional. A nivel país la región del Biobío es la tercera con mas emisiones de óxidos de nitrógeno, las que constituyen mas de 25.000 toneladas al año producidas por centrales termoeléctricas, calderas y otras actividades industriales principalmente (Ministerio del Medio Ambiente, 2017).

Tabla 2.4: Mortalidad y morbilidad asociada a la exposición a $MP_{2,5}$ durante el 2015

Tipo de Evento	**Evento**	**Grupo de edad**	**Casos**
Mortalidad Prematura	Cardiopulmonar	Mayores de 30 años	3.723
Admisiones Hospitalarias	Cardiovasculares	Mayores de 18 años	1.709
Admisiones Hospitalarias	Pulmonar	Mayores de 18 años	231
Admisiones Hospitalarias	Neumonía	Mayores de 65 años	1.049
Admisiones Hospitalarias	Ataques de asma	0-64 años	152
Visita a Sala de Emergencias	Bronquitis aguda	0-17 años	108.100
Restricción de Actividad	Días de pérdida de trabajo	18-64 años	870.756
Restricción de Actividad	Días de actividad restringida	18-64 años	3.861.706
Restricción de Actividad	Días de actividad restringida menor	18-64 años	7.273.360

Fuente: Ministerio del Medio Ambiente (2017)



Fuente: Ministerio del Medio Ambiente (2017)

Figura 2.5: Emisiones anuales de $MP_{2,5}$ por región y fuente para el año 2015

Capítulo 3
Marco teórico

En este capítulo se describen los métodos utilizados en el desarrollo de este trabajo. Primero se presentan las técnicas relativas al preprocesamiento de la base de datos, luego se describen las redes neuronales, la teoría respecto a las parametrizaciones utilizadas y el algoritmo XGBoost.

3.1. Preprocesamiento de la información

El entrenamiento de un modelo demanda la entrega de información completa para poder estudiar y reflejar las relaciones entre los datos de entrada (variables independientes) y los datos de salida (variables dependientes). Sin embargo, en sets de datos reales es usual que existan problemas de pérdida de información, ya sea por errores humanos, lecturas incorrectas de los sensores, o errores en el proceso de almacenamiento de los datos; lo que suele provocar errores en los códigos. Soluciones sencillas van desde ignorar los datos perdidos hasta imputarlos con criterios como la media o el vecino más cercano. Galván (2007) habla acerca de como la mayoría de investigadores suelen ignorar este problema y dejan que los paquetes estadísticos lo resuelvan, haciendo supuestos a veces inadecuados sobre la información, lo que introduce sesgos y reduce el poder explicativo de los métodos estadísticos, asi como también le resta eficiencia a la fase de inferencia y puede incluso invalidar las conclusiones del estudio. Ignorar o eliminar las muestras donde hay valores pérdidos puede reducir dramáticamente la cantidad de datos disponibles, lo que no es recomendable considerando que las técnicas de machine learning se basan en la utilización de grandes cantidades de datos. Por este motivo, se deciden imputar los datos faltantes. Existen diferentes métodos como de imputación basados en diferentes criterios como la media (*mean imputation*), el vecino mas cercano (*nearest neighbour imputation*), miss forest y la imputación multiple a través de ecuaciones encadenadas, entre otros. Se decidió ocupar la imputación múltiple a través de ecuaciones encadenadas ya que es uno de los mejores métodos según Waljee et al. (2013).

3.1.1. Imputacion múltiple a través de ecuaciones encadenadas

La Imputacion múltiple a través de ecuaciones encadenadas (Multiple imputation by chained equations, MICE) es una técnica que se basa en el supuesto de los datos de las variables usadas en la imputación han sido perdidos al azar (Missing At Random, MAR) y por lo tanto la probabilidad de que un determinado valor este perdido depende de los valores observados de las otras variables (Galván, 2007). En caso de que las obsevaciones perdidas no sean MAR, MICE también puede realizar estimaciones considerando el supuesto de que la ausencia de datos depende de los valores no observados (Missing Not At Random, MNAR) (Orjuela et al., 2018).

Los tres principales pasos de la imputación multiple son la imputación, el análisis y la agrupación. En la imputación se crean *m* sets de datos remplazando los valores perdidos por valores plausibles acorde a una distribucion específicamente modelada para cada entrada perdida teniendo en cuenta su dependencia con los datos observados (Buuren and Groothuis-Oudshoorn, 2010). Los *m* sets solo difieren en los valores perdidos que han sido imputados y la magnitud de las diferencias refleja la incertudumbre respecto a que valor imputar (White et al., 2011). Luego de que los datos han sido imputados, cada set de datos se analiza individualmente y se combinan los *m* conjuntos generados de la imputación múltiple en uno. Se pueden ocupar diferentes métodos con esta técnica de imputación, como el método de ajuste de la media predictiva, la imputación de media incondicional o el bosque aleatorio, entre otros.

El principal atributo del método del ajuste de la media predictiva o comparación predictiva de medias es que las imputaciones estan restringidas a los valores observados y que puede preservar relaciones no lineales (Buuren and Groothuis-Oudshoorn, 2010); el bosque aleatorio al manejar tipos mixtos de datos faltantes hace de estos adaptables a las interacciones y a la no linealidad; y imputación de media incondicional mantiene la media de los datos aunque disminuye su varianza y covarianza con otras variables (Orjuela et al., 2018). El número de imputaciones (*m*) es sumamente relevante para lograr buenas inferencias estadísticas, la Tabla 3.1 resume las recomendaciones segun el nivel de significancia y la fracción de información perdida (γ).

3.2. Redes Neuronales

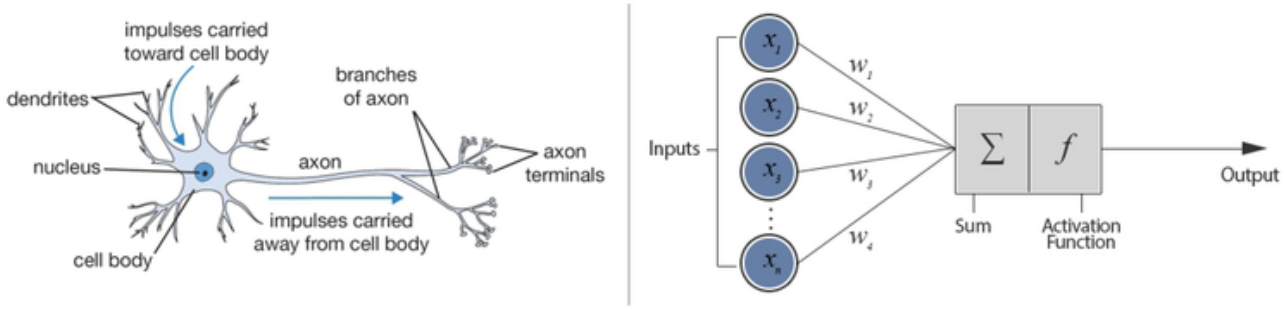
El cerebro humano tiene habilidades sorprendentes, como lo son la habilidad o destreza para reconocer caras o discursos aún estando rodeados de otros estímulos. También es increiblemente robusto, no deja de funcionar a pesar de que algunas células mueran, mientras que dificilmente una CPU podría seguir funcionando si alguno de sus componentes falla, aunque este quizás no es uno de sus aspectos mas fascinastes sino el hecho de que puede aprender, sin la necesidad de una actualización.

Tabla 3.1: Número de imputaciones necesarias

	Nivel de significancia aceptable		
γ	$< 5\,\%$	$< 3\,\%$	$< 1\,\%$
0.1	3	5	20
0.3	10	20	20
0.5	10	20	40
0.7	20	40	40
0.9	40	40	100

Fuente: Graham et al. (2007)

Las valoraciones y cálculos que tienen lugar en el cerebro son realizados por una compleja red de neuronas interconectadas que se comunican a través de la sinapsis, los axones y las dendritas causando la activación o no de ciertas neuronas (Krogh, 2008). El término red neuronal artificial (ANN por sus siglas en inglés) se utiliza para describir un modelo matemático que intenta emular computacionalmente las características esenciales encontradas en las redes neuronales del cerebro. McCulloch and Pitts (1943) fueron los que introdujeron este concepto en su trabajo *Un cálculo lógico de las ideas inmanentes en la actividad nerviosa*, donde presentaron un modelo computacional simplificado que pretendía imitar a las redes neuronales del cerebro animal realizando complejos calculos a través de lógica preposicional (Géron, 2017).



Fuente: Data Camp

Figura 3.1: Comparación entre una neurona en biológica y una neurona artificial

Las neuronas funcionan como un interruptor que recibe una entrada de otras neuronas y dependiendo del peso total de la entrada, la neurona es activada o no. El factor o peso por el que el valor de entrada es multiplicado corresponde a la fortaleza de la sinapsis[1], estos pesos pueden ser positivos o negativos, y excitar o inhibir a la neurona siguiente (Krogh, 2008). En la Figura 3.1 se muestra la comparación entre una neurona animal y una artificial, donde $x_1, x_2, ..., x_n$ corresponde a los valores de entrada y $w_1, w_2, ..., w_n$ a sus pesos, los que son multiplicados, sumados y luego se les aplica una función de activación para obtener el resultado final. En la década de los sensenta se demostró que sistemas modelados como redes neuronales tienen propiedades similares a las del cerebro como la habilidad de reconocer patrones sofisticados, y pueden funcionar incluso si algunas neuronas son destruidas (Krogh, 2008). Sin embargo, el objetivo de estos sistemas no es simplemente crear modelos realistas del cerebro, sino utilizar sus propiedades para desarrollar algoritmos robustos que permitan modelar problemas difíciles (Brownlee, 2017).

Los modelos se pueden clasificar según el tipo de información disponible y la interrogante planteada. Un modelo de aprendizaje supervisado es aquel que aprende en base a información previamente clasificada y de la que conoce la respuesta correcta. Sin embargo, a veces la señal que desea predecir es desconocida, es entonces cuando se recurre a un modelo de aprendizaje no supervisado, aquí el set de datos es un conjunto de atributos para los que no se especifica una salida o respuesta correcta, y para los que la red neuronal intenta encontrar una estructura. Finalmente, está el aprendizaje por refuerzo que se basa en el logro de un objetivo, y si las acciones conducen hacia ese objetivo, entonces se obtiene una recompensa, este tipo de modelos fundamenta sus decisiones en base al movimiento que entregue la mayor recompensa (Salian, 2018).

3.2.1. Topología de la red neuronal

Las redes neuronales estan formadas por arreglos de neuronas, donde a cada fila se le llama capa. La primera capa, también llamada capa visible o de entrada, es la que recoge los datos. Generalmente se utiliza una neurona por cada valor de entrada, los que son transmitidos directamente a la siguiente capa. Las capas consecutivas son llamadas capas ocultas y pueden ser numerosas dependiendo del problema [2]. A la última capa se le llama capa de salida y es la responsable de entregar el vector de valores correspondientes a la salida esperada del problema [3] (Géron, 2017). Existen diferentes tipos de redes según su arquitectura. El Perceptrón es el modelo mas simple (compuesto por una neurona) y es el precursor otros sistemas mas complejos, como el perceptrón multicapa (Multilayer Perceptron,

[1] Sinapsis: espacio donde se produce el contacto entre células nerviosas

[2] Deep learning o aprendizaje profundo se le llama a los problemas con mas de una capa oculta

[3] Es especialmente importante la elección de la función de activación en la última capa, ya que entrega el formato de salida y por lo tanto debe ser coherente con el tipo de problema

MLP), modelo formado por dos o mas capas de neuronas (Brownlee, 2017). El poder predictivo de las redes neuronales y de este modelo en específico viene de su habilidad para aprender de los datos con los que es entrenado. Las salidas se pueden modelar de la siguiente manera:

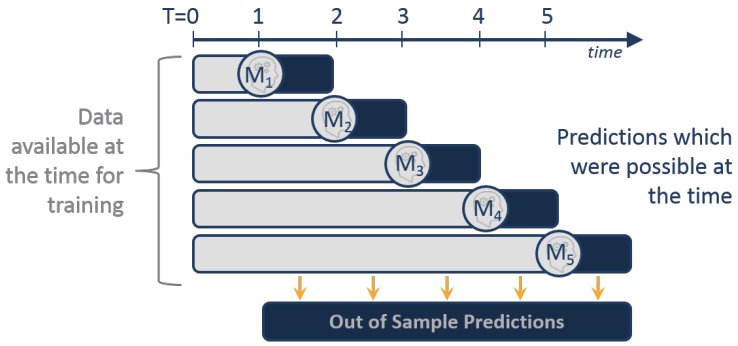
$$t_k = \phi_0(\alpha + \sum_{j \to k} w_{jk}\phi_h(\alpha_j + \sum_{i \to j} w_{ij}x_i)) \tag{3.1}$$

Donde t_v denota a la salida *v* del modelo, α es una constante, ϕ_0 es la función de activacion de la capa de salida y ϕ_h la función de activación de las capas ocultas (Fallah et al., 2009). Otros modelos comunes son las redes convolucionales (Convolutional Neural Network, CNN) y las redes recurrentes (Recurrent Neural Network, RNN) que se utilizan con datos espaciales y para el análisis de sequencias, respectivamente. Las CNN se caracterizan por la función de activación que utilizan y son particularmente buenas para el reconocimiento de imagenes, mientras que las RNN se caracterizan por poseer una unidad de memoria que les permite almacenar información y reflejar las dependencias temporales entre los datos. El problema que enfrentan estas últimas es que el gradiente se desvanece con el tiempo (Ver Sección 3.3.2 para más información sobre el gradiente) y las relaciones de largo plazo se pierden. Para solucionar esta cuestión existen las redes recurrentes con puertas (Gated Recurrent Unit, GRU) que a través de dos puertas, una de actualizacion y otra de reajuste, controlan el modo en el que la información fluye dentro de la unidad; y las de memoria larga de corto plazo (Long Short Term Memory, LSTM) que utilizan tres puertas, una para discriminar la información que se debe recordar, otra para la que se debe olvidar y otra diferenciar los datos en los que debe enfocarse (Gray, 2018).

3.3. Entrenamiento

Una vez configurada la red neuronal es necesario entrenarla en una base datos. Para conocer el desempeño general de la red es necesario realizar pruebas en datos independientes, que no hayan sido vistos por el algoritmo durante su entrenamiento, así se divide el set de datos en dos partes, una llamada de entrenamiento y otra de prueba (Krogh, 2008). Esto puede ocasionar problemas de sobreestimación si es que los datos esconden una estructura subyacente, y para evitarlo existe la validación cruzada o *cross-validation* (Bronshtein, 2017). Uno de los métodos más utilizados es la validación cruzada sobre k carpetas. Sin embargo, cuando se trata con datos temporales la validación cruzada tradicional puede causar pérdidas de información, como por ejemplo las relaciónes temporales, por lo que es importante preservar el orden cronológico de los eventos y para esto se utiliza la validación cruzada anidada o con retención (Cochrane, 2018). El problema con lo anterior es que solo se utiliza una fracción de los datos para el entrenamiento del modelo y no aquellos que se van generando en el tiempo, para solucionarlo

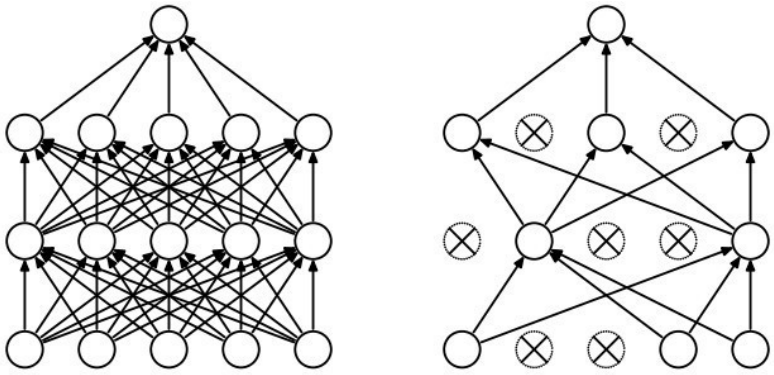
existe la validación hacia adelante o *walk forward validation*. Esta técnica permite realizar pronósticos e incorporar la nueva información en cada paso, de esta manera se comienza con una muestra incial (M_1) donde se entrena el algoritmo y se predice un periodo, luego éste es agregado al conjunto de entrenamiento formando una nueva muestra (M_2) con la que se pronóstica el siguiente periodo, tal como se muestra en la Figura 3.2. Es importante destacar que esta técnica es válida solo para series de tiempo (Gray, 2018).



Fuente: Gray (2018)

Figura 3.2: Validación hacia adelante

Existen otros métodos, además de la validación cruzada, para evitar el sobreajuste. Dropout es otra técnica de regularización utilizada en las redes neuronales profundas y consiste en eliminar aleatoriamente neuronas, junto con sus conexiones durante el entrenamiento para evitar la co-adaptación de los pesos (Srivastava et al., 2014). La Figura 3.3 ilustra este proceso.



Fuente: Srivastava et al. (2014)

Figura 3.3: A la izquierda una red neuronal estándar y a la derecha después de aplicar dropout

3.3.1. Función de pérdida

La función de pérdida es un método que permite conocer que tan bueno es el algoritmo al modelar un set de datos. Mientras más grande sea la pérdida peor sera la estimación, por lo que se utiliza como indicador para reestructurar los cambios en la red y mejorar el modelo (Algorithmia, 2018). Existen varias funciones de pérdida y cada una esta hecha para manejar diferentes tipos de tareas, ya sea un problema de regresión o clasificación (Agrawal, 2017). Las mas comunes son el error cuadrático medio para problemas de regresión, que calcula la diferencia entre el valor estimado y el real, lo eleva al cuadrado y luego lo suma a lo largo del set de datos; la pérdida de entropia cruzada, utilizada en problemas de clasificación y que penaliza fuertemente los errores; o la pérdida absoluta promedio, que captura la diferencia absoluta entre el valor real y el estimado. En general la función de pérdida se puede plantear de la siguiente manera:

$$\mathcal{L}(w) = \sum_{v=1}^{m} L(y^{(v)}, \phi(x^{(v)}, w)) \tag{3.2}$$

Donde $x^{(v)}$ denota a los valores de entrada para la muestra v, w a los parámetros que deben ser aprendidos por el modelo y ϕ a la función de activación (Changhau, 2017). Para evitar problemas de sobreestimación se suele añadir un término de regularización que reduce la magnitud de los parámetros y mantiene al modelo lo mas sencillo posible (Ng, 2018). En la Tabla 3.2 se muestran las ventajas y desventajas de algunas de las funciones de pérdida mas comunes para problemas de regresión.

Poisson

La función de regresión lineal de poisson es útil al tratar con datos de conteo y es considerada como una extensión de los modelos lineales generalizados para las redes neuronales. La distribución de Poisson expresa la probabilidad de que un determinado numero de eventos suceda en un intervalo de tiempo o espacio, y esta dada por $P[Y = y] = \frac{e^{\lambda}\lambda^{y}}{y!}$, donde y es el número de eventos en el intervalo, y λ el número promedio de eventos en el intervalo. Fallah et al. (2009) propone modelar λ como una función de $x^{(v)}$ a través de una red neuronal simple (perceptrón multicapa) de la siguiente manera:

$$t^{(v)} = \hat{\lambda}^{(v)} = \phi_0(\alpha + \sum_{j} w_j \phi_h(\alpha_j + \sum_{i \to j} w_{ij} x_i^{(v)})) \tag{3.3}$$

Considerando que el error es la probabilidad logaritmica negativa, y las observaciones son independientes, se tiene:

Tabla 3.2: Ventajas y desventajas de funciones de pérdida para problemas de regresión

Funciones de pérdida	Ventajas	Desventajas
Error Cuadrático Medio	El gradiente es grande para pérdidas considerables y decae a medida que se acerca a cero, lo que lo hace mas preciso durante el proceso de entrenamiento	Demasiado sensible a los puntos átipicos
Error Medio Absoluto	Robusto cuando los datos estan corrompidos por outliers o valores átipicos.	El gradiente es el mismo a lo largo del tiempo, lo que significa que el gradiente será grande aún para pérdidas pequeñas.
Huber	Es menos sensible a los puntos atipicos que el error cuadrático medio y es diferenciable en 0.	Es necesario realizar un proceso iterativo para entrenar el parámetro delta [4] .
Log-Cosh	A diferencia del error cuadrático medio no se ve fuertemente afectada por predicciones incorrectas ocasionales. Tiene las mismas ventajas que la función de pérdida de Huber pero es doblemente diferenciable [5].	El gradiente es constante para predicciones que se alejan del valor real.

Fuente: Elaboración propia en base al artículo de Grover (2018)

$$\begin{aligned}\mathcal{L} &= -logP(y|x) = -log\ \prod_{v=1}^{m} p(y^{(v)}|x^{(v)}) = -\sum_{v=1}^{m} log\ \ p(y^{(v)}|x^{(v)}) \\ &= -\sum_{v=1}^{m}(-t^{(v)} + y^{(v)} log\ \ t^{(v)} - ln\ \ y_n!)\end{aligned} \tag{3.4}$$

Eliminando el último término que no esta relacionado con el entrenamiento del modelo:

$$\mathcal{L} = \sum_{v=1}^{m}(t^{(v)} - y^{(v)} log\ \ t^{(v)}) \tag{3.5}$$

3.3.2. Optimización

El objetivo de la optimización es encontrar el vector *w* que minimice la función de pérdida. Una de las estrategias más intuitivas y fundamentales es la del Gradiente descendente. El primer paso es asignar valores aleatorios a *w*, con éstos la red procesa los valores de entrada y produce uno o varios valores de salida. El valor de salida se compara con el valor esperado y se calcula el error correspondiente. Luego, con la ayuda de la derivada de la función de costos, el error es propagado hacia

atras en la red (*Backpropagation*), una capa a la vez, y los parámetros se actualizan en la dirección más prometora, es decir, la opuesta al gradiente (Brownlee, 2017). Lo que se ve reflejado en la siguiente ecuación:

$$w_i := w_i - \eta \frac{\partial}{\partial w_i} \mathcal{L}(w) \tag{3.6}$$

Donde η es el tamaño del paso o ritmo de aprendizaje (learning rate) (Karpathy, 2019). Existen tres variantes del gradiente descendente y dependiendo de la cantidad de datos que se utilicen se tranza entre precisión y rapidez. El gradiente descendente por lotes (*batch gradient descent*) calcula los parámetros *w* para todo el conjunto de entrenamiento y luego los actualiza, por lo que es mas lento e infactible para bases de datos grandes debido al espacio que requiere en la memoria. Garantiza la convergencia al minimo global para funciones convexas y local para funciones no convexas. El gradiente descendente estocástico (*stochastic gradient descent*) actualiza los parámetros para cada muestra del conjunto de entrenamiento, por lo que es más rápido, aunque las actualizaciones constantes causan una gran varianza. Mientras el gradiente por lotes asegura la convergencia al minímo alrededor de donde los parámetros son estimados inicialmente, la fluctiación del gradiente estocástico le permite alcanzar otros minimos locales potencialmente mejores. Para evitar problemas de convergencia la tasa de aprendizaje debe disminuirse paulatinamente. El gradiente descendente por mini lotes (*mini-batch gradient descent*) agrupa lo mejor de los métodos anteriores y realiza actualizaciones de los parámetros por mini lotes de n muestras del conjunto de entrenamiento, asi reduce varianza en la actualización de parametros y entrega resultados diligentemente. Algunos de los algoritmos mas utilizados son Momentum, Nesterov accelerated gradient, Adagrad, Adadelta, RMSprop y Adam, entre otros. Adam supera ligeramete a los otros algoritmos en el tiempo de convergencia por lo que es el mas recomendado sobretodo cuando se trata con redes neuronales complejas o profundas (Ruder, 2016).

3.4. Predicción

Una vez la red neuronal ha sido entrenada puede ser utilizada para realizar predicciones. Las predicciones se realizan sobre datos jamás vistos por el algoritmo con el fin de probar el poder predictivo de este y compararlo con otros modelos (Brownlee, 2017), para ello se utilizan diferentes metricas que miden el grado de acoplamiento entre los datos originales y aquellos estimados. Algunas de éstas métricas son la varianza explicada, el error medio absoluto, el error cuadrático medio y el coeficiente de determinación (R^2), entre otras (Universidad de Valencia, 2007). La varianza mide la dispersión de los datos y puede descomponerse en dos partes, una parte explicada por la regresion y otra no.

$$S_y^2 = S_{y^*}^2 + S_{r(y/x)}^2 \tag{3.7}$$

Donde $S^2_{y^*}$ es la varianza de la regresión y $S^2_{r(y/x)}$ la varianza de los residuos o error cuadrático medio (MSE).

$$S^2_{y^*} = \sum_{i=1}^{n} \frac{(y^*_i - \bar{y})^2}{n} \tag{3.8}$$

$$S^2_{r(y/x)} = \sum_{i=1}^{n} \frac{(e_i - \bar{e})^2}{n} \tag{3.9}$$

Cuanto mayor sea la proporción explicada tanto mejor será el ajuste y tanto más útil la regresión. A la proporción de varianza explicada por la regresión se le llama coeficiente de determinación.

$$R^2 = \frac{S^2_{y^*}}{S^2_y} \tag{3.10}$$

En conclusión, para tener un buen ajuste la varianza explicada debe adoptar valores relativamente grandes, mientras que el MSE debe tener valores pequeños y R^2 acercarse a 1.

3.4.1. Hiperparámetros

Son aquellos parámetros que utiliza el algoritmo para resolver el problema, es decir, no son parámetros propios del problema pero tienen un gran impacto en el entrenamiento y el desempeño final del algoritmo. Los hiperparámetros son especificados antes de que empiece el entrenamiento y no pueden ser optimizados dentro del mismo. Considerando una función $f : \mathcal{X} \longrightarrow \mathbb{R}$ que se quiere minimizar en un dominio $X \subseteq \mathcal{X}$, el problema puede ser representado de la siguiente manera:

$$x^* = argmin_{x \in \mathcal{X}} f(x) \tag{3.11}$$

Donde f representa un valor objetivo a minimizar, como el error cuadrático medio o el coeficiente de determinación, evaluado en el conjunto de validación; y x^* es el conjunto de hiperparámetros que proporcionan el menor valor.

Existen varias técnicas que tienen como objetivo mejorar la elección de estos hiperparámetros, algunas son Grid Search, Random Search y Bayesian Optimization, entre otras. La búsqueda en red o *grid search* requiere la elección de un conjunto de valores para cada parámetro, los que se evalúan para cada combinación de valores posibles, la desventaja de éste método es que exige entrenar el algoritmo en los *K* sets de parámetros, espacio que crece exponencialmente con el número de parámetros (Bergstra and Bengio, 2012). La búsqueda aleatoria o *Random Search*, a diferencia de la anterior, selecciona aleatoriamente subconjuntos de de los parámetros por lo que es mucho mas eficiente en espacios mul-

tidimensionales (Bergstra and Bengio, 2012). Sin embargo, ambos métodos son ineficientes ya que pasan una cantidad considerable de tiempo evaluando malos hiperparámetros y no toman en consideración el conocimiento de los resultados previos en su elección. En contraste a estos métodos esta el ajuste automático de hiperparámetros, que recolecta información del desempeño de las configuraciones previas para hacer una elección inteligente de los parámetros a probar. El objetivo es minimizar la cantidad de pruebas y encontrar a la vez un buen óptimo. Éste puede ser visto como un problema de optimización, donde se deben encontrar los hiperparámetros que maximicen el desempeño del modelo. Pero como este problema no puede ser descrito en una fórmula, no es posible calcular su derivada ni ocupar métodos basados en este cálculo (Pham, 2016). Es por esto que a pesar de que los modelos de optimización sequencial intercambian de manera eficiente la exploración y explotación del espacio de búsqueda, las técnicas no bayesianas son más utilizadas en la práctica debido a la sobrecarga administrativa y la experiencia requerida para la implementación de estos modelos (Dewancker et al., 2015). En este caso, el enfoque utilizado se basa en los modelos de optimización sequencial.

Modelos de optimización sequencial

Los modelos de optimización sequencial, también llamados SMBO por sus siglas en inglés, pueden diferenciarse por sus *modelos de regresión* y por sus *funciones de aquisición* (Dewancker et al., 2015). Existen diversos paquetes que trabajan con estos métodos tal como se resumen en la Tabla 3.3.

Tabla 3.3: Softwares de Modelos de optimización sequencial

Software	**Modelo de regresión**	**Función de adquisición**
Spearmint	Procesos Gaussianos	Mejora Esperada
MOE	Procesos Gaussianos	Mejora Esperada
Hyperopt	Estimadores de Árboles Estructurados	Mejora Esperada
SMAC	Bosque Aleatorio	Mejora Esperada

Fuente: Dewancker et al. (2015)

Modelos de Regresión probabilística

Existen varios modelos de regresión probabilística, éstos deben definir una distribución de predicción $p(y|x, \mathcal{D})$, es decir, crean un modelo probabilístico de la funcion objetivo que mapea los datos de entrada a una probabilidad de pérdida: $p(perdida|datos\ \ de\ \ entrada)$ (Koehrsen, 2018). Esta distribución captura la incertidumbre asociada al modelo sustituto y su reconstrucción de la función. Algunos modelos son los Procesos Gaussianos (Gaussian Processes), donde las predicciones siguen

una distribución normal; los Bosques Aleatorios (Random Forests), donde la función de predicción también se distribuye normal; y los Estimadores de Árboles Estructurados (Tree Parzen Estimators) que, a diferencia de los anteriores, no define una distribución predictiva sobre la función objetivo, sino que crea dos procesos jerárquicos que actuán como modelos generativos de todas las variables del dominio (Dewancker et al., 2015).

Funciones de adquisición

Las funciones de adquisición generalmente evaluan una pérdida esperada relacionada con la evaluación de f en x para luego seleccionar el punto con la menor pérdida. Las funciones de adquisición más populares son la *Probabilidad de Mejora* , que plantea una función de utilidad que entrega una recompensa de una unidad si $f(x)$ es menor o igual a su derivada (la que se asume como el mínimo); y la *Mejora Esperada* que a diferencia de la anterior plantea una función de utilidad que entrega una recomensa igual al tamaño de la mejora, lo que evita el estancamiento en óptimos locales (Garnett, 2019).

Tree Parzen Estimators (TPE)

Los Estimadores de Árboles Estructurados a diferencia de los Procesos Gaussianos que modelan directamente en base a $p(y|x)$, esta estrategia modela $p(x|y)$ y $p(y)$,y crea dos procesos jerárquicos, $l(x)$ y $g(x)$ que actuán como modelos generativos de todas las variables del dominio. Estos procesos modelan los dominios de las variables cuando la función objetivo esta por debajo de un cuantil específico y^*.

$$p(x|y,\mathcal{D}) = \begin{cases} \text{l(x)}, & y < y^* \\ \text{g(x)}, & y \geq y^* \end{cases} \tag{3.12}$$

Donde $l(x)$ es la densidad formada usando las observaciones $x^{(i)}$ que corresponden a la pérdida $f(x^{(i)})$ cuando es menor que y^*, y $g(x)$ que se forma usando las observaciones restantes. Entonces, el algoritmo TPE depende de y^* que es mayor al mejor observado $f(x)$ y se elige y^* tal que sea un cuantil γ de los valores y observados, donde $p(y < y^*) = \gamma$ (Bergstra et al., 2011). Así el algortimo utiliza la mejora esperada como función de adquisicion, obteniendo:

$$EI_{y^*}(x) = \frac{\gamma y^* l(x) - l(x)\int_{-\infty}^{y^*} p(y)dy}{\gamma l(x) + (1-\gamma)g(x)} \propto (\gamma + \frac{g(x)}{l(x)}(1-\gamma))^{-1} \tag{3.13}$$

Esta expresión nos muestra que para maximizar la mejora se desean puntos x tal que la probabilidad $l(x)$ sea alta y la de $g(x)$ sea baja, lo que facilita el proceso de generar candidatos y evaluarlos de acuerdo a $g(x)/l(x)$ (Bergstra et al., 2011).

3.5. XGBoost

XGBoost viene del inglés *eXtreme Gradient Boosting* y es una biblioteca de software a la que se puede acceder desde diferentes interfaces como C++, Python y R, entre otras. Esta biblioteca se basa en la utilizacion de árboles de decisión e implementación del gradiente aumentado (boosting gradient). Boosting es una técnica de ensamblado donde nuevos modelos son añadidos, sequencialmente, para corregir los errores de los modelos existentes hasta que no se pueden hacer mejoras. Se le llama gradiente aumentado porque utiliza el algortimo de gradiente descendente para minimizar la pérdida al añadir nuevos modelos. Soporta tanto problemas de clasificación como de regresión y la velocidad de ejecución y buen desempeño general son sus principales características (Brownlee, 2016).

Los árboles de decisión son un tipo de algoritmo de aprendizaje supervisado. Se divide la población o muestra en conjutos homogéneos basados en la variable de entrada mas significativa y se analiza la mejor variable para la ramificación, a este proceso se le conoce como segmentación recursiva (Orellana, 2018). Considerando Y como la variable de respuesta (aleatoria) y p variables predictivas, $x_1, x_2, ..., x_p$ (fijas), se puede establecer una relación entre las variables x_i e Y con su función de probabilidad o esperanza condicional, segun se trate de un árbol de clasificación o regresión, respectivamente. Los elementos característicos de un árbol son los nodos y ramas, cada nodo representa una prueba sobre algún atributo de la instancia y cada rama proveniente de un nodo corresponde a uno de los posibles valores del atributo. Para dividir el nodo raíz en dos nodos homogéneos, se elige la división tal que la pureza de los dos nodos hijos sea superior a la del nodo madre; ésta suele medirse por la entropía o la información ganada (Gain). La entropía mide la impureza de los datos y se representa matemáticamente como $Entropia(S) = \sum_{i=1}^{k} -p_i log_2(p_i)$ donde S es el conjunto muestral que incluye datos de las k categorías y p_i la proporción de S que pertenece a la clase i. La información ganada es la reducción esperada de la entropía causada por particionar la muestra según un determinado atributo.

$$Gain(n, S_n) = Entropia(n) - \sum_{s \in S_n} \frac{|s|}{|n|} Entropia(s) \tag{3.14}$$

Donde $S_n = \{S_1, ..., S_k\}$ es una partición del nodo n y $|n|$ es la cardinalidad de elementos del nodo n. Se selecciona aquella partición que proporcione la mayor ganancia de información (L'Huillier and Weber, 2010). El proceso de segmentación recursiva continua hasta que no se puedan realizar mas divisiones y cuando esto sucede al nodo final se le llama nodo terminal (Universidad de Valencia, 2011).

3.6. Modelos Lineales Generalizados

Los modelos lineales generalizados (Generalized Lineals Models, GLM) tienen tres componentes básicas, la componente aleatoria que consiste en la variable aleatoria Y y su función de distribución (binomial, poisson, binomial negativa o normal) ; la componente sistemática que especifica las variables explicativas x_i en forma de efectos fijos en un modelo lineal $\eta_i = \sum_i \beta_i x_{ij}$, denominado predictor lineal, donde x_{ij} es el valor del i-ésimo predictor para la muestra j; y la función link, que es una función $g(\cdot)$ del valor esperado de Y ($E(Y) = \mu$) que relaciona las componentes sistemáticas y aleatorias como se muestra en la ecuación 3.15.

$$g(\mu) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k \tag{3.15}$$

Cuando la variable de respuesta es de recuento o conteo, el modelo más simple consiste en asumir una distribución de Poisson para la componente aleatoria (Y). La propiedad más caracteristica de esta distribución es que su media y varianza coinciden. Este modelo asume el logaritmo de los valores esperados (media) como función link, por ello también se le conoce como log-linear model. La funcion de distribución poisson es:

$$f(n; \lambda) = \frac{\lambda^n e^{-\lambda}}{n!} \tag{3.16}$$

Donde n es el número de ocurrencias del evento y λ la media ($\lambda = \mu$). La ecuacion matemática general para el modelo de regresión poisson es:

$$log(\mu) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k \tag{3.17}$$

Despejando se tiene que $\mu = e^{\sum_i \beta_i x_i}$. Con esta información se construye el modelo para predecir la variable aleatoria Y. El principal problema de este modelo es que la media y la varianza no suelen ser iguales en recuentos reales. Cuando la varianza es mayor a la media se le llama sobredispersión y cuando es menor subdispersión. La distribución binomial negativa se utiliza para resolver esta situación ya que cuenta con un parámetro para reflejar la dispersión en los datos. Asi se tiene que la esperanza es $E(Y) = \mu$ y la varianza $Var(Y) = \mu + \frac{\mu^2}{k}$, donde $\frac{1}{k}$ es el parámetro de dispersión. Cuando este se acerca a cero la distribución binomial negativa converge a una distribución poisson. La función de distribución binomial negativa es:

$$f(n; k; \lambda) = \frac{\Gamma(n+k)}{\Gamma(k)\Gamma(n+1)} (\frac{k}{\lambda + k})^k (1 - \frac{k}{\lambda + k})^n \tag{3.18}$$

Donde λ es la media y n es el número de ocurrencias del evento. En general se utiliza la función link de tipo logaritmo en este modelo.

Capítulo 4
Base de Datos

4.1. Diagnóstico de variables

Las variables necesarias para el desarrollo del estudio pueden clasificarse de la siguiente forma:

- Variables de calidad del aire: indica la concentración de las diferentes moléculas y compuestos que están presentes en el aire y pueden ser perjudiciales para la salud y el medio ambiente.
- Variables meteorológicas: ofrecen información relativa a fenómenos atmosféricos como la temperatura, el viento, la humedad relativa y lluvia entre otros.
- Egresos hospitalarios: estas variables indican el diagnóstico, días de estadía y otras características demográficas de los pacientes atendidos en los servicios de salud pública.
- Atenciones de urgencia: al igual que la variable anterior guarda un registro de la causa y algunas de las características demográficas relevantes del paciente pero de aquellos que ingresan al sistema público de urgencias.

4.2. Reopilación de información

Las variables de calidad del aire se obtuvieron del Sistema de Información Nacional de Calidad del Aire (SINCA) de forma desagregada por estación y contaminantes del aire. La temporalidad de los datos recogidos fue diaria debido a que la información de egresos hospitalarios solo se encuentra en esta unidad de tiempo. En el Anexo 7.2 se encuentra la lista de datos de calidad con los que se cuenta para el estudio, los que en total constituyen 99 registros, las fechas de cada registro varían dependiendo de cada estación así como también según el compuesto o molécula. Las variables meteorológicas se recogieron del sistema de la Red Agrometeorológica del Instituto Nacional de Investigación Agraria (INIA) y de la Dirección General de Aguas (DGA). Los datos de egresos hospitalarios se adquirieron

a través de la página web del Departamento de Estadísticas e Información de Salud (DEIS). El registro contiene toda la información relativa a las atenciones en los servicios de salud públicos (exceptuando urgencias) por año, desde el 2001 al 2017. La información de las atenciones de urgencia se descargó de la base de datos del DEIS, desde el año 2008 hasta el año 2017. Ambos set de datos cuentan con un glosario para la interpretación de las variables los que se encuentran en la Sección 7.3 de los Anexos.

4.3. Creación de Base de Datos

Los datos de calidad del aire anteriormente recopilados se agruparon en múltiples archivos con la información referente a cada compuesto en un archivo particular, donde las filas indican la estación y su localización, y las columnas la fecha correspondiente de la medición. Las variables de salud que corresponden a las atenciones en el servicio público se dividen inicialmente en dos grupos, ingresos hospitalarios y atenciones de urgencias, ambas consideran valores diarios. Mientras los ingresos hospitalarios cuentan con un registro detallado de la edad de cada paciente, en las atenciones de urgencia esta característica se encuentra dividida considerando conglomerados de pacientes de 1 a 4 años, de 5 a 14 años, de 15 a 64 años y de 65 años o más. Para facilitar el trabajo con las variables se decide dividir tanto los ingresos hospitalarios como las atenciones de urgencia por grupos etarios, siguiendo lo realizado por Fernández (2018), se decide fusionar los dos primeros grupos obteniendo tres conjuntos. Considerando que ya se ha demostrado que existe una correlación positiva entre la contaminación atmosférica y las enfermedades respiratorias y del sistema circulatorio, como muestran los estudios hechos por Dominici et al. (2006), Tonne et al. (2012), Raaschou-Nielsen et al. (2013) y otros; es que se decide dividir las variables de salud según al sistema que afectan, respiratorio o circulatorio. Así se obtienen las variables mostradas en la Tabla 4.1.

4.4. Pre-procesamiento de los datos

Las variables son estudiadas y modificadas con el objetivo de conocer características clave que permitan la correcta implementación de modelos adecuados a los datos e interrogante planteada.

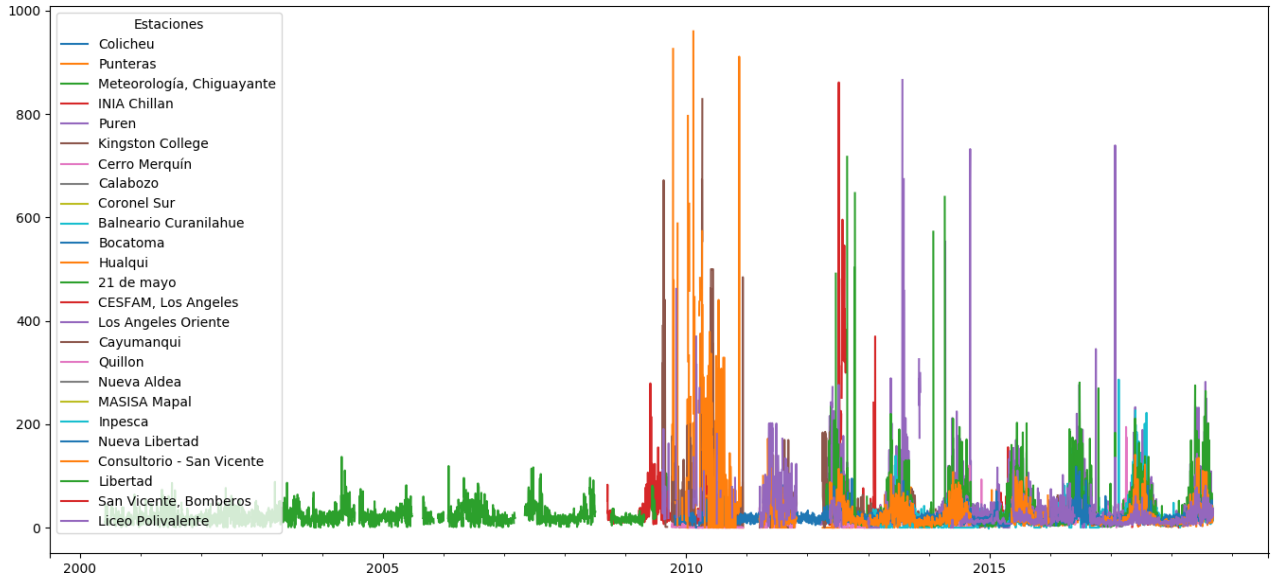
4.4.1. Material Particulado

Para dimensionar el problema y la calidad de los datos estos se grafican, obteniendo una representación clara de los valores perdidos. En las Figuras 4.1 y 4.2 se muestran los valores sin validar de $MP_{2,5}$ y MP_{10}, los principales contaminantes que afectan a la salud, para la región del Biobío y Ñuble
.

Tabla 4.1: Lista de variables

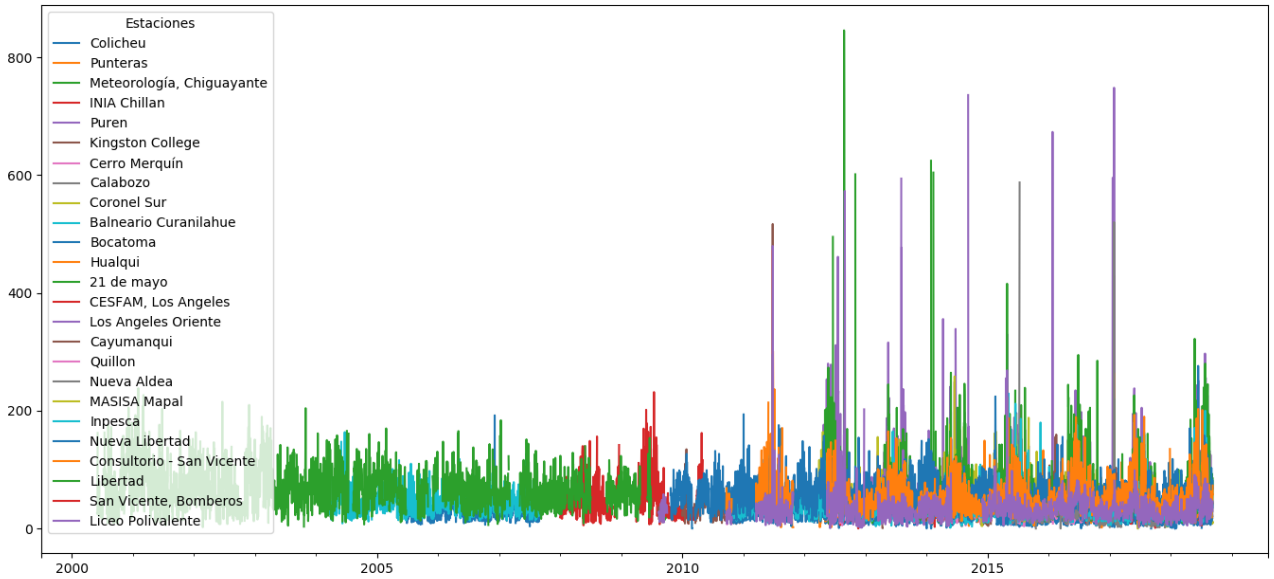
Variables	**Descripción**
PM10	Concentración de Material particulado MP10 (promedio diario).
PM25	Concentración de Material particulado MP2,5 (promedio diario).
SO2	Concentración de Dióxido de azufre (promedio diario).
NO2	Concentración de Dióxido de nitrógeno (promedio diario).
NOX	Concentración de Óxidos de nitrógeno (promedio diario).
NO	Concentración de Monóxido de nitrógeno (promedio diario).
CO	Concentración de Monóxido de carbono (promedio diario).
O3	Concentración de Ozono (promedio diario).
CH4	Concentración de Metano (promedio diario).
HCNM	Concentración de Hidrocarburos no metánicos (promedio diario).
HCT	Concentración de Hidrocarburos totales (promedio diario).
T	Temperatura (promedio diario).
P	Precipitación (promedio diario).
V	Velocidad del viento (promedio diario).
IG1_R	Ingresos hospitalarios por enfermedades respiratorias, G1.
IG2_R	Ingresos hospitalarios por enfermedades respiratorias, G2.
IG3_R	Ingresos hospitalarios por enfermedades respiratorias, G3.
IG1_C	Ingresos hospitalarios por enfermedades circulatorias, G1.
IG2_C	Ingresos hospitalarios por enfermedades circulatorias, G2.
IG3_C	Ingresos hospitalarios por enfermedades circulatorias, G3.
UG1_R	Urgencias por causas respiratorias, G1.
UG2_R	Urgencias por causas respiratorias, G2.
UG3_R	Urgencias por causas respiratorias, G3.
UG1_C	Urgencias por causas circulatorias, G1.
UG2_C	Urgencias por causas circulatorias, G2.
UG3_C	Urgencias por causas circulatorias, G3
IG1_R2	Ingresos hospitalarios por seleccionadas enfermedades respiratorias, G1.
IG2_R2	Ingresos hospitalarios por seleccionadas enfermedades respiratorias, G2.
IG3_R2	Ingresos hospitalarios por seleccionadas enfermedades respiratorias, G3.

Fuente: Elaboración propia



Fuente: Elaboración propia en base a datos del SINCA

Figura 4.1: Evaluación de las concentraciones diarias ambientales de $MP_{2,5}$ μg/m3



Fuente: Elaboración propia en base a datos del SINCA

Figura 4.2: Evaluación de las concentraciones diarias ambientales de MP_{10} μg/m3

Ambas figuras reflejan estacionalidad en sus datos, tal como se esperaba considerando que gran parte del material particulado proviene de procesos de combustión que aumentan en el invierno debido a la calefacción. A grandes rasgos también se observan valores menores de $MP_{2,5}$ en comparación al MP_{10} lo que se condice con lo esperado. Sin embargo, estos datos deben ser validados antes de poder ser utilizados.

Validacion de los datos

Fue necesario comprobar la coherencia de los datos y para esto se utilizaron los siguientes criterios:

1. Siempre debe ser la concentración ambiental de $MP_{2,5} < MP_{10}$. En el caso de detectar una condición anómala. Se debe descartar el dato.

2. En el caso de encontrar valores de concentración de $MP_{2,5}$ o MP_{10} exageradamente elevados se verifica este valor con registros del mismo parámetro en estaciones que se encuentren cercanas.

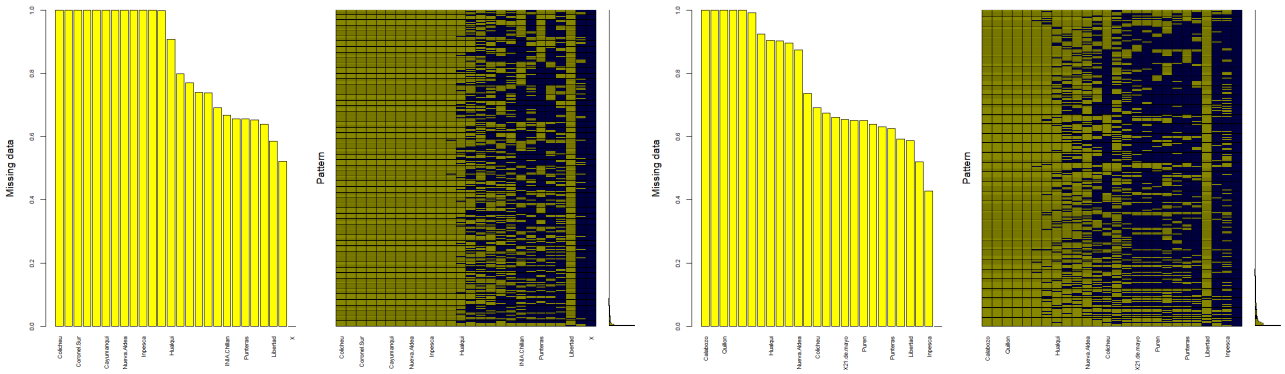
3. En los casos donde se encuentren estaciones de monitoreo en las cercanía, a menos de 30 km (Eum et al., 2015), se identifican las correlaciones entre los valores. Ver Anexo 7.4

4. Si aparecen valores de concentración de MP_{10} y $MP_{2,5}$ que se repiten en forma consecutiva, este puede ser una evidencia que existe un error en el registro. En algunos casos se explica por fallas de transmisión de datos o el logger. Por esto, cuando un valor se repite mas de 7 veces seguidas este se elimina.

Para observar de mejor manera los valores perdidos se utilizó el paquete VIM (Visualization and Imputation of Missing Values, por sus siglas en inglés) en R. En la Figura 4.3 se refleja la cantidad de valores perdidos de $MP_{2,5}$ y MP_{10}, respectivamente, para el periodo que abarca desde febrero del 2000 hasta septiembre del 2018 con los registros ya validados. En ambos gráficos las estaciones han sido ordenadas por la cantidad de valores perdidos. En la Sección 7.5 de los Anexos se encuentra la tabla que muestra el porcentaje de observaciones perdidas para el mismo periodo.

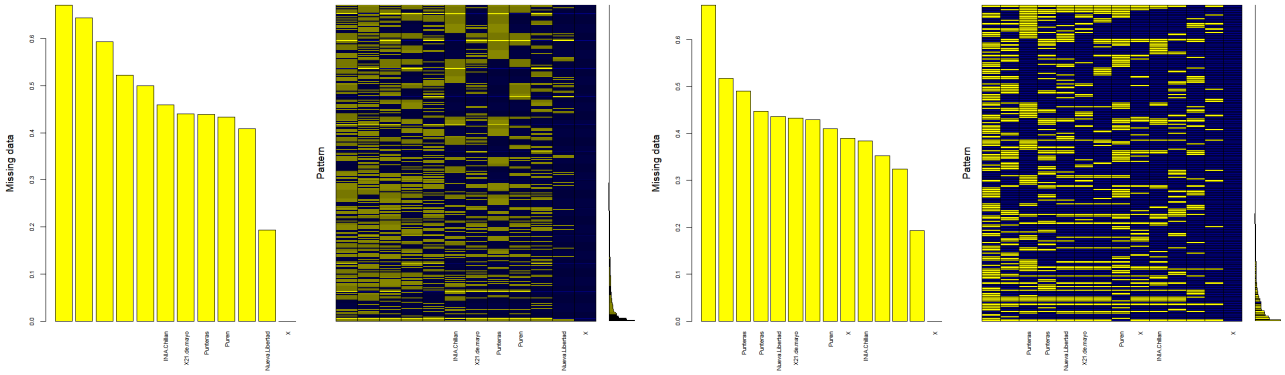
Valores perdidos

La imputación de los valores perdidos se realizó con el paquete `mice` de R. Considerando los resultados anteriores se decide utilizar los datos desde el 2008 hasta el 2017, ya que permitirán una interpolación mas suave y precisa debido a que se dispone de mayor información proporcional al periodo. Además, los datos correspondientes al 2018 no se encontraban disponibles al momento del estudio. Para obtener una mejor resolución se eliminan aquellas estaciones que tienen mas del 70 %



Fuente: Elaboración propia en base a datos del SINCA

Figura 4.3: Gráfico valores perdidos $MP_{2,5}$ (a la izquierda) y MP_{10} (a la derecha) (2000-2018)



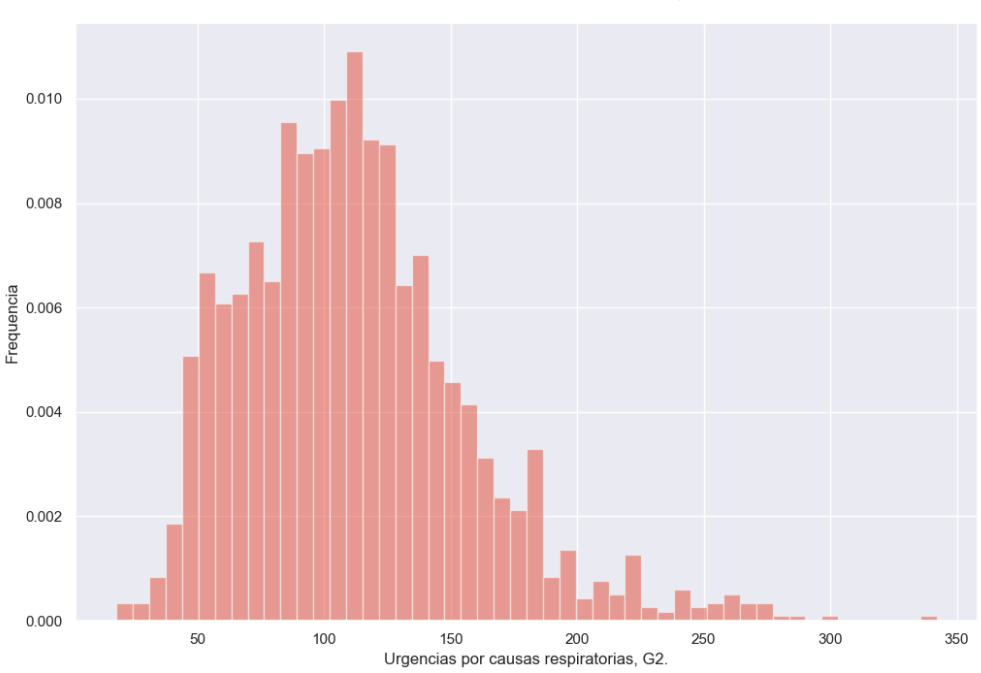
Fuente: Elaboración propia en base a datos del SINCA

Figura 4.4: Gráfico valores perdidos de $MP_{2,5}$ (a la izquierda) y $MP_{2,5}$ (a la derecha) (2008-2017)

de sus valores perdidos o no observados (Ver Tabla 7.8) y considerando un nivel de significancia del 5 % se elige un número de imputaciones igual a veinte ($m = 20$) (Graham et al., 2007). La cantidad de valores perdidos para los registros validados y no validados correspondientes a los diferentes periodos se ilustran en la Sección 7.5 de los anexos. La visualización premilimar de los datos se observa en la Figura 4.4, donde el color azul representa la presencia de datos y el amarillo la ausencia. El resultado final de este proceso se obtuvo combinando las imputaciones generadas utilizando la media para cada estación.

4.4.2. Análisis de variables de salud

La base de datos descargada con las variables de salud debío ser limpiada para poder realizar las agrupaciones descritas en la Sección 4.3. Primero se creó un registro de todos los hospitales, y establecimientos de la red de urgencias en la Región del Biobío. Luego, se cruzo este registro con los datos nacionales de atenciones de urgencia, dando lugar al conjunto de pacientes atendidos en al Región del Biobío. Para los ingresos hospitalarios esto fue inncesario ya que se encontraban agrupados por región. Para estos últimos se revisaron cada una de las enfermedades relacionadas al sistema cardiovascular y respiratorio, correspondientes a aquellas clasificadas con la letra I, J y R en el diagnostico principal (DIAG1), según la décima edición de la clasificación internacional de enfermedades; y las clasificadas con las letras W e Y en el segundo diagnóstico (DIAG2). Luego de agrupar las variables de salud estas se graficaron y se les ajustaron distintas funciones de distribución para conocer las carácteristicas de los datos. En la Figura 4.5 se muestra a modo de ejemplo como se distribuyen los datos de urgencias por enfermedades respiratorias para el grupo etario 2.



Fuente: Elaboración propia

Figura 4.5: Urgencias por enfermedades Respiratorias para el grupo etario 2

Capítulo 5
Resultados y Discusión

En este capítulo se presentan los resultados para comparar la calidad de ajuste de tres formulaciones alternativas al problema presentado, la primera utilizando redes neuronales, la segunda una biblioteca de python basada en árboles de decisión, y la tercera modelos lineales generalizados. Todas ellas consideran las variables presentadas en la Tabla 4.1, donde las primeras 15 variables corresponden a variables independientes o explicativas y las restantes a las variables dependientes. Además, se consideraron los rezagos de las variables meteorológicas y de calidad del aire como nuevas variables; y se agregaron cuatro variables de tendencia, para controlar la estacionalidad lineal, exponencial, sinusoidal y cosenoideal de los datos acorde a los trabajos de Mardones et al. (2015) y Fernández (2018).

Redes Neuronales

El problema en cuestión puede analizarse con un modelo de aprendizaje supervisado, ya que se cuenta con etiquetas para cada muestra en la base de datos, las que corresponden a la cantidad de ingresos hospitalarios y atenciones de urgencia. Una red neuronal simple, normalmente llamada *Multilayer perceptron*, es la elegida; esta decisión se sustenta en que no es necesario predecir ni la calidad del aire ni las variables meteorológicas, ya que estas serán predichas por otro modelo desarrollado de forma paralela a este proyecto. El modelo tiene como entrada los datos predichos de calidad de aire, junto con la información meteorológica, y entregara una estimación de los ingresos hospitalarios y atenciones de urgencia para el periodo.

Inicialmente se consideró llevar a cabo dos estimaciones, la primera con un número limitado de datos para realizar una comparación preliminar de los modelos propuestos y la segunda con una mayor cantidad de datos, considerando mas variables y ampliando el rango temporal, con la esperanza de lograr mayor precisón en las estimaciones a través de un entrenamiento robusto. Sin embargo, debido al tiempo acotado del estudio no fue posible entrenar el modelo en la base de datos ampliados descritos en la sección anterior. Así, la base de datos reducida fue proporcionada por el estudio de

Fernández (2018) y contiene información desde el año 2013 al 2017 para la comuna de Los Ángeles, considera las variables meteorológicas con rezagos de tres días, y el MP_{10} y $MP_{2,5}$ con rezagos de cinco días.

Se realizó un modelo para cada variable dependiente y los datos se estandarizaron utilizando la función `MinMaxScaler` de scikitlearn, escalando los datos entre cero y uno; la elección de esta función fue debido a que los valores atípicos entregan información relevante que se perdería con la estandarización normal, además que de esta manera se mantiene la naturaleza positiva de los datos para su análisis. La estructura general de la red se definió contemplando una capa de entrada con las variables mencionadas anteriormente; seguida de una o más capas ocultas, cada una de ellas ligada a una capa de desecho (*dropout layer*), que elimina aleatoreamente el 20 % de las variables en cada capa oculta con el objetivo de evitar el sobreajuste; y finalmente la capa de salida compuesta por un solo nodo para indicar la cantidad de pacientes esperados. Luego, con esta estructura general se determinaron los hiperparámetros, donde se utilzó el modelo de optimización sequencial basado en estimadores de árboles estructurados, ya que es el único que no define una función de distribución por sobre los datos, permitiendole realizar inferencias que se ajustan mejor a la realidad de estos. Los hiperparámetros estimados fueron la cantidad de unidades o neuronas por capa, las capas ocultas y el número de epochs, los que se ajustaron para que en un principio tomaran valores aleatorios como se ilustra en la Tabla 5.1. En la Sección 7.6 de los anexos se encuentran los valores finales de los hiperparámetros. A modo de ejemplo se muestra la estructura final de los ingresos hospitalarios por enfermedades respiratorias para el grupo etario 1 en la Figura 5.1.

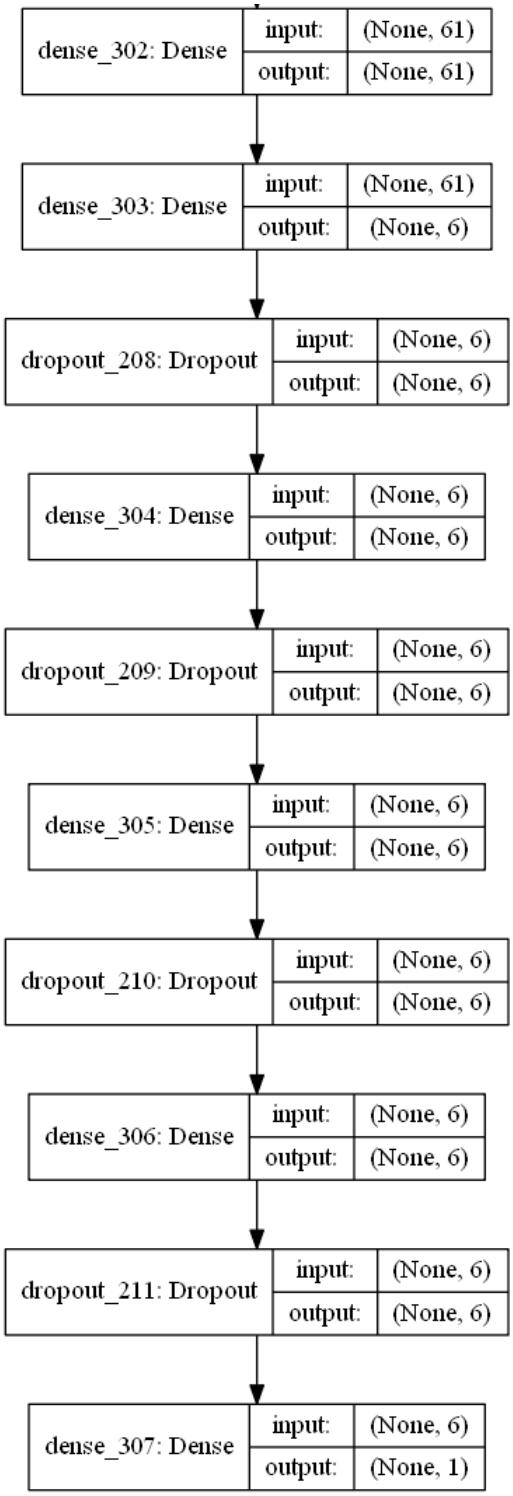
Tabla 5.1: Hiperparámetros MLP

Hiperparámetros	**Función de distribución**	**Rango**
Unidades	hp.randint	[1;18]
Capas	hp.randint	[1;11]
Epochs	hp.randint	[1;200]

Fuente: Elaboración propia

Finalmente, el rendimiento del modelo se probó haciendo uso de la validación hacia adelante (*walk forward validation*), donde se dejaron los datos correspondientes a los últimos dos meses para la validación; realizando pronósticos semanales[1] los que se evaluaron y compararon utilizando el error cuadrático medio. Para las comparaciones se calculó el promedio del desempeño en los ultimos dos meses, asi se resalta en los resultados mostrados mas adelante.

[1]La cantidad de pronósticos puede compararse con la cantidad de carpetas utilizadas en la validación cruzada tradicional ($k = 8$)



Fuente: Elaboración propia

Figura 5.1: Estructura modelo para ingresos hospitalarios por enfermedades respiratorias G1

XGBoost

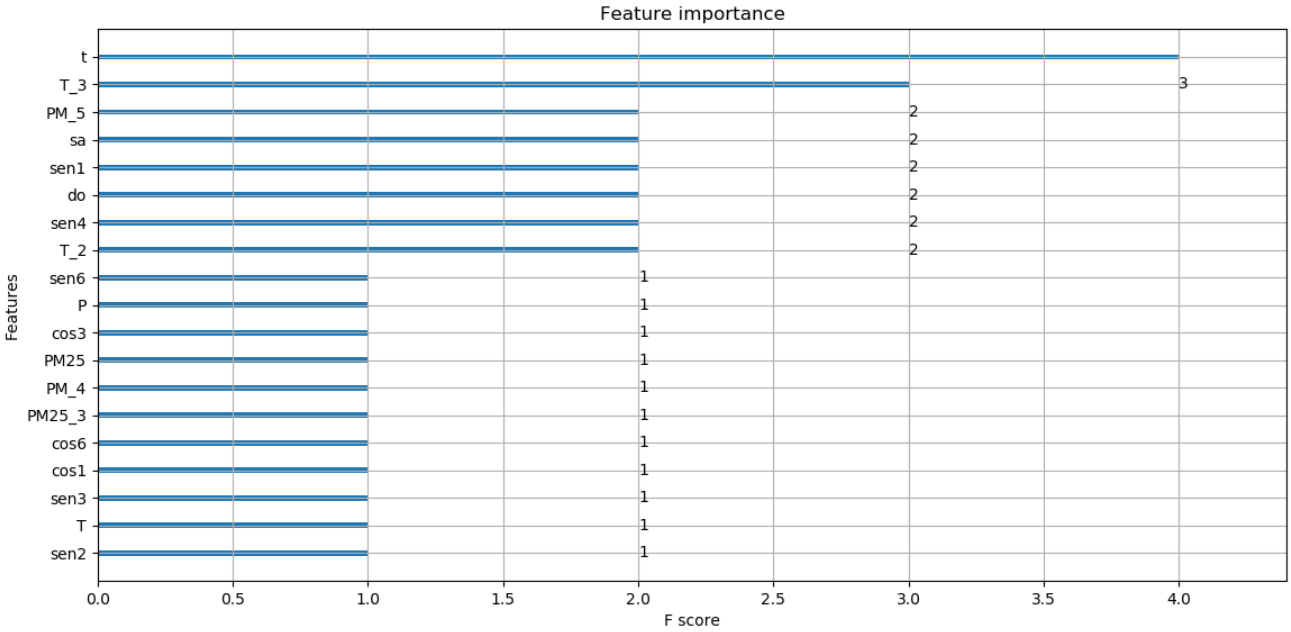
Este modelo, al igual que el anterior, fue probado con la base de datos reducida. Las variables se estandarizaron utilizando la misma función que en el modelo anterior y los hiperparámetros también fueron estimados haciendo uso de los árboles estructurados. Los hipérparámetros relevantes en este caso fueron `eta` (similar a tasa de aprendizaje en el gradiente aumentado); la profundidad máxima del árbol y la suma mínima de los pesos requerida en los nodos hijos, ambas utilizadas para controlar el sobreajuste u *overfitting*, ya que mientras mayor es la profundidad el modelo este es más capaz de aprender las relaciones especificas de los datos particulares, lo mismo sucede cuando la suma de los pesos es pequeña (valores más altos previenen esta situación ya que evita que aprenda en base a una muestra en particular del árbol); la reducción minima de pérdida requerida para una nueva división; y la fracción de columnas o porcentaje de características utilizadas como muestra para cada árbol. En este modelo al igual que en los otros se utilizó la función poisson como función objetivo, ya que se ajusta a los datos de conteo.

Tabla 5.2: Hiperparámetros XGBoost

Hiperparámetros	**Función de distribución**	**Rango**	**Parámetros adicionales (q)**
eta	hp.quniform	[0,001; 0,5]	0,025
profundidad máxima	hp.randint	[1;14]	
suma mínima de los pesos en nodos hijos	hp.quniform	[1;6]	1
reducción mínima de la perdida	hp.quniform	[0;1]	0,05
fracción de columnas	hp.quniform	[0.5;1]	0,05

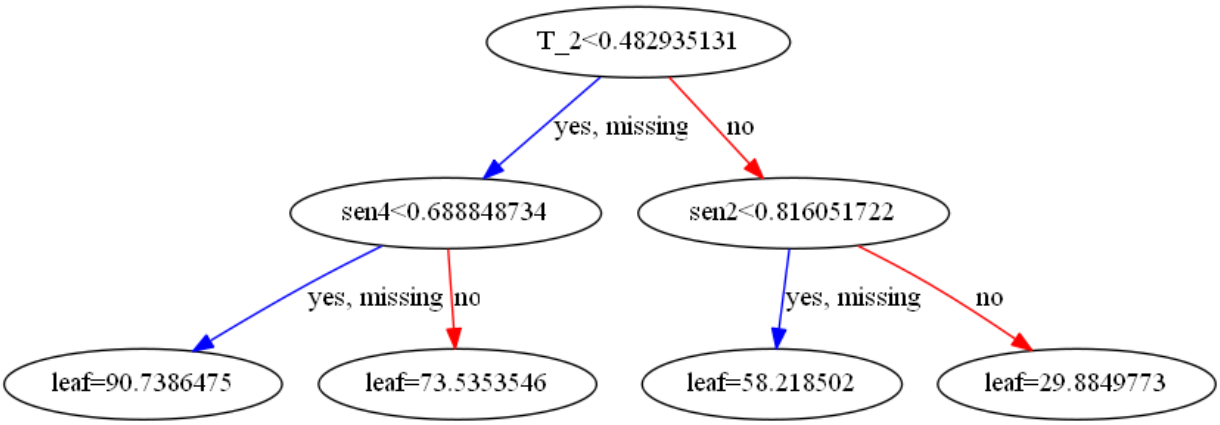
Fuente: Elaboración propia

Adicionalmente, se calculó la importancia relativa de las variables del modelo, como se ilustra a modo de ejemplo en la Figura 5.2. Aquí se observa como la variable mas influyente es aquella para controlar la estacionalidad lineal, seguida por la temperatura de hace tres días y el material particulado de hace cinco. La Figura 5.3 ilustra el proceso de toma de decisión reflejando la condición en cada nodo hasta que se cumple con los criterios especificados anteriormente, como son la reducción minima de la pérdida o la profundidad máxima. La ventaja de los modelos basados en árboles de decisión es que permiten una mayor comprensión de las relaciones entre las variables y ayuda a la interpretación del modelo. Al igual que este árbol se forman otros muchos de forma sucesiva, cada uno con el objetivo de mejorar la predicción hecha por el anterior, enfocandose en aquellas características que le permitan reducir el error en las muestras donde la predicción no fue buena.



Fuente: Elaboración propia

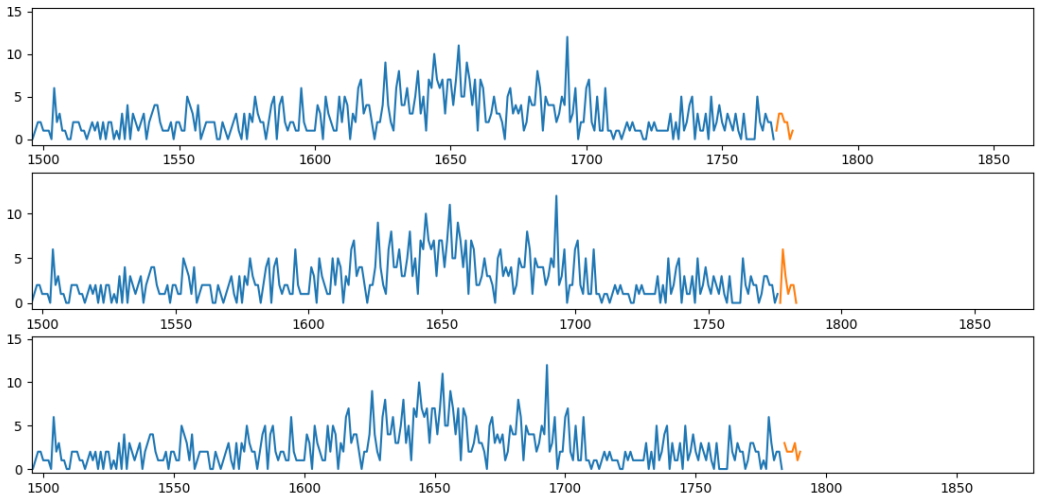
Figura 5.2: Importancia relativa atenciones de urgencia por enfermedades respiratorias G1



Fuente: Elaboración propia

Figura 5.3: Árbol de decisión para atenciones de urgencia por enfermedades respiratorias G1

La Figura 5.4 muestra graficamente las predicciones hechas por el algoritmo XGBoost, como se describió anteriormente, los pronósticos son semanales y se muestran en naranjo. Cada imagen revela el avance de las predicciones cada 30 días, es decir, en el gráfico superior se observa en naranjo las predicciones hechas desde el 6 al 12 de noviembre del 2017, que corresponden a la muestra 1.770 a 1.777, el gráfico medio muestra en naranjo las predicciones desde el 13 al 20 de noviembre; y el inferior las pertinentes al periodo del 20 al 26 de noviembre.



Fuente: Elaboración propia

Figura 5.4: Predicciones semanales XGboost

Modelos Lineales Generalizados

La elección del modelo de distribución para cada variable dependiente se fundamenta en si existe o no sobredispersión en sus parámetros. Debido a que la base de datos es la misma a la utilizada por Fernández (2018), se ocupo su estudio para determinar la distribución a utilizar. La distribución Poisson se utilizó con los Ingresos hospitalarios del grupo etario 1 y 2 por enfermedades cardiovasculares (IG1_C y IG2_C) y con el grupo etario 3 por ciertas enfermedades respiratorias seleccionadas (IG3_R2). El resto de modelos se evaluaron con la distribución Binomial Negativa.

5.1. Resultados de los modelos

La Tabla 5.3 muestra el error cuadrático medio, indicando la media y la desviación estándar, de cada modelo en el conjunto de entrenamiento y validación, resaltando en negrita los mejores valores. Idealmente se esperaría encontrar valores cercanos a cero, ya que esto indicaría que la diferencia entre el valor estimado y real es pequeña, signo de una buena estimación. Sin embargo, se observan grandes diferencias en magnitud, evidenciando un mal ajuste. El análisis en detalle complementado con los resultados para el coeficiente de determinación (Ver Sección 7.7 de los anexos) se encuentra a continuación. Además de estos modelos, a modo de prueba y acorde a lo planteado por Fernández (2018) se crearon dos nuevos, en uno excluyendo el MP_{10} y sus rezagos; y otro excluyendo el $MP_{2,5}$ y sus rezagos. En este caso solo se ejecutó el algoritmo XGBoost junto con el modelo lineal generalizado. Los resultados para el $MP_{2,5}$ y el MP_{10}, junto con otras métricas e información sobre los modelos se pueden encontrar en el repositorio de github del usuario mariaicarrasco: *MT_maria_ignacia_carrasco*.

5.1.1. Enfermedades respiratorias

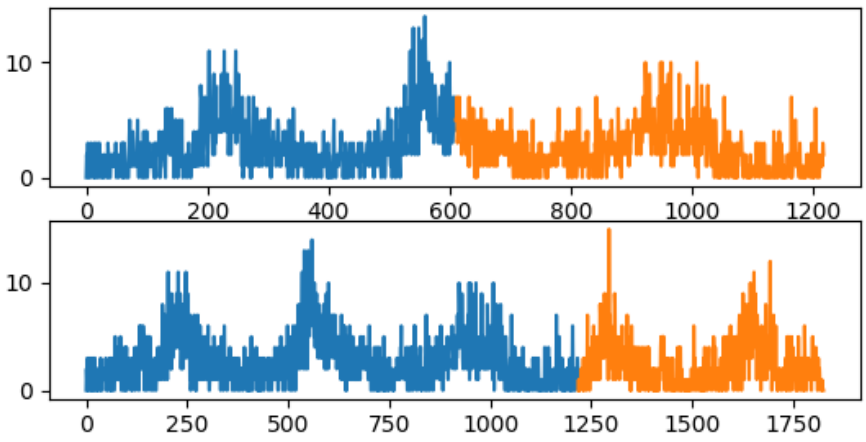
Ingresos hospitalarios por enfermedades respiratorias

En los ingresos hospitalarios por enfermedades respiratorias para el grupo etario 1, utilizando el algoritmo XGBoost, se obtuvo un coeficiente de determinación del 0,6684 en el conjunto de entrenamiento y un valor de -0,9768 en el conjunto de validación. Lo que significa que la media es capaz de explicar mejor los datos que la función ajustada, situación que puede deberse a que se forzó a la función a tomar un intercepto que no era favorable para los datos debido a un sobreajuste en el proceso de entrenamiento. Lo mismo sucede con los demás ingresos por enfermedades respiratorias en todos los modelos. En el conjunto de entrenamiento, considerando el error cuadrático medio, XGBoost superó considerablemente a la red neuronal y esta a su vez al modelo lineal generalizado, mientras que en el conjunto de validación la red neuronal tuvo valores menores para la varianza de los residuos en el grupo 2 y 3; sin embargo, se encuentran valores negativos para el coeficiente de determinación en el proceso de validación para todos los modelos. Cuando se separan las variables $MP_{2,5}$ y MP_{10} se obtiene un leve pero mejor desempeño en el conjunto de validación que en el de entrenamiento para el grupo etario 3, situación que se repite para los grupos 1 y 3 cuando solo se considera el MP_{10} como contaminante.

Tabla 5.3: Error cuadrático medio de los distintos modelos

	Entrenamiento			**Validación**		
Modelos	**XGBoost x(std)**	**GLM x(std)**	**MLP x(std)**	**XGBoost x(std)**	**GLM x(std)**	**MLP x(std)**
IG1_R	**1.648 (1.2519)**	3.327 (0.0099)	5.7208 (3.6549)	**1.8008 (1.4178)**	2.2857 (1.6429)	2.7604 (2.3791)
IG2_R	**1.0524 (0.6252)**	1.7903 (0.0088)	1.3872 (0.1105)	1.209 (0.7148)	1.5357 (0.9442)	**1.0415 (0.6097)**
IG3_R	**1.3526 (0.3855)**	1.842 (0.0055)	1.5127 (0.1115)	1.161 (0.8626)	1.4464 (1.3542)	**1.1477 (0.628)**
IG1_C	**0.0463 (0.0261)**	0.1057 (0.0007)	0.0994 (0.0064)	**0.0298 (0.0366)**	0.0357 (0.0619)	0.0363 (0.0573)
IG2_C	**1.7218 (1.4755)**	2.4194 (0.0104)	2.7627 (1.6484)	2.3918 (2.2668)	2.6786 (2.6086)	**2.3084 (2.0409)**
IG3_C	**1.1997 (1.0927)**	3.1157 (0.0131)	2.8605 (0.2297)	**3.7524 (2.6255)**	4.0357 (3.3895)	3.9329 (2.3759)
UG1_R	**1032.2163 (1142.9027)**	2123.3711 (15.9098)	7073.1235 (8730.133)	899.0283 (539.4074)	**281.7857 (122.1947)**	5253.1284 (6558.7393)
UG2_R	843.0933 (724.85)	**674.3596 (3.4459)**	2240.6191 (1236.6001)	532.9999 (507.6347)	**213.5179 (109.7725)**	838.1298 (642.7481)
UG3_R	**14.8507 (12.125)**	43.5678 (0.2046)	92.3964 (23.3946)	19.7734 (5.4447)	**17.8036 (6.2065)**	24.6973 (9.4312)
UG1_C	**1.2777 (0.1311)**	1.7047 (0.0041)	1.356 (0.2778)	**0.5712 (0.2723)**	0.75 (0.4503)	0.6105 (0.3348)
UG2_C	**11.3634 (5.7604)**	23.4856 (0.0501)	55.6573 (54.8328)	20.8326 (7.4336)	**19.6071 (8.2223)**	50.8281 (59.7299)
UG3_C	**8.0986 (6.9641)**	17.2693 (0.0537)	17.6216 (1.5865)	32.8464 (22.0615)	26.7143 (14.106)	**24.8018 (15.5156)**
IG1_R2	**2.1789 (1.234)**	3.2241 (0.012)	3.2685 (0.2767)	1.6839 (1.1623)	2.3393 (1.6582)	**1.5319 (1.0371)**
IG2_R2	**0.9133 (0.4138)**	1.4926 (0.0101)	1.4591 (0.5508)	**0.7404 (0.3112)**	1.1071 (0.5138)	0.8708 (0.246)
IG3_R2	**1.0496 (0.1464)**	1.5366 (0.0056)	1.2097 (0.0951)	**0.6179 (0.2464)**	1.1429 (0.7143)	0.6326 (0.2059)

Fuente: Elaboración propia



Fuente: Elaboración propia

Figura 5.5: Predicciones IG1_R, utilizando MLP con TimeSeriesSplit

La Figura 5.5 muestra a modo se ejemplo como se distribuyen los datos cuando se particionan considerando periodos más grandes. En este caso se dividieron los datos en dos; primero, dejando desde los datos desde 01/01/2013 al 02/09/2014 para el entrenamiento y realizando un pronóstico para el periodo del 03/09/2014 al 02/05/2016; y luego considerando desde el 01/01/2013 al 02/05/2016 para pronosticar desde el 03/05/2016 al 21/12/2017. Analizando el gráfico, se observa como la red neuronal fue capaz de captar la estacionalidad del modelo, y quizás un analisis con mayor información y pronósticos por periodos mas extensos sería capaz de reducir la varianza de los residuos en comparación a la varianza total y reportar mejores resultados para el coeficiente de determinación

Atenciones de Urgencia por enfermedades respiratorias

Las urgencias por causas respiratorias tienen errores considerablemente altos, lo que en un primer lugar indica un mal ajuste; sin embargo, los errores mayores en magnitud se explican debido a que son justo estas variables las que presentan un mayor número de pacientes promedio. El algoritmo XGBoost posee para los tres grupos etarios excelentes coeficientes de determinación durante el proceso de entrenamiento, todos con valores por sobre el 50 % de varianza explicada por el modelo, lo que podría indicar un buen modelo, sin embargo durante el proceso de validación se observan R^2 negativos lo que indica que el algoritmo aprendió relaciones muy específicas durante el entrenamiento y tiene dificultad para generalizar sobre datos jamás observados. El grupo etario 3 es el único que presenta un ceoficiente positivo, con un valor de 0,1233, lo que significa que el modelo XGBoost es capaz de explicar el 12,33 % de la varianza, lo que si bien no es un valor aceptable para predecir la cantidad de atenciones de urgencia, permite reflejar el impacto del material particulado en la salud.

El modelo lineal generalizado posee valores positivos para el R^2 tanto en el entrenamiento como en la validación. Siendo los de la validación menores que los del entrenamiento. En cuanto a la varianza de los errores, se observan números inferiores en todos los casos. Esta información indica un buen ajuste por parte de este modelo a los datos, especialmente para los grupos 2 y 3, donde el GLM fue capaz de explicar el 48,55 % y el 21,48 % de la varianza, respectivamente.

Por último, en la red neuronal, todos los grupos demostraron valores muy altos en sus errores cuadraticos medios, los que disminuyeron en el conjunto de validación. Contrastando esta información con el coeficiente de determinación, se tiene que solo el grupo 3 presentó un coeficiente positivo durante el entrenamiento, mientras que en los demás casos los valores son negativos por lo que se concluye que este modelo particular necesita mas información para lograr un buen ajuste.

Ingresos hospitalarios por enfermedades respiratorias seleccionadas

Las enfermedades englobadas en esta categoría son las catalogadas como IRA alta, neumonía, bronquitis/bronquiolitis, crisis obstructiva bronquial, asma y otras causas respiratorias, en el sistema de urgencias. En el modelo XGBoost se observan errores pequeños y menores en el conjunto de validación en comparación al de entrenamiento, lo que al igual que en los casos anteriores indica que probablemente los casos en la validación no poseían tantos puntos atípicos como durante el entrenamiento. Además de observan coeficientes de determinación negativos lo que induce a pensar que el modelo en el proceso de entrenamiento forzó a este a pasar por un intercepto desfavorable para los datos en el conjunto de validación. Es decir, el modelo se sobreajustó y mientras es excelente para predecir en casos para los que fue entrenado, falla al generalizar esta información.

El modelo lineal generalizado y la red neuronal, presentan errores levemente inferiores en el conjunto de validación para todos los casos, probablemente por la misma razon descrita anteriormente para el algoritmo XGBoost. En el caso del GLM al observar los coeficientes de determinación, el algortimo no logra superar a la media como predictor en ninguno de los casos. No logrando aprender relaciones significativas para los grupos 2 y 3; y sobreajustando el modelo en el grupo 1. El modelo perceptrón multicapa falla al realizar predicciones para los tres grupos, sin embargo es capaz de encontrar relaciones significativas durante el entrenamiento para el grupo 1.

5.1.2. Enfermedades cardiovasculares

Ingresos hospitalarios por enfermedades cardiovasculares

El modelo XGBoost presenta una varianza de los residuos levemente menor durante el entrenamiento a la obtenida en el conjunto de validación para el grupo etario 2 y 3, lo que se condice con lo esperado y podría indicar un buen ajuste. Sin embargo al analizar el coeficiente de determinación se obtienen valores positivos y relativamente buenos en el conjunto de entrenamiento y valores negativos en el conjunto validación, los que permiten concluir que el modelo se sobreajustó a los datos. En el grupo etario 1 se observa lo contrario, siendo el error en el entrenamiento mayor al de la validación, lo que junto a un R^2 de 0,5236 en el entrenamiento y uno de 0,07105 en el conjunto de validación, lo que también indica un sobreajuste en este caso.

El modelo lineal generalizado en el grupo etario 1 se desempeña considerablemente bien en el conjunto de validación, alcanzando un R^2 de 0,7083 a pesar de su valor de -0,0877 durante el entrenamiento. La explicación de esta diferencia abismal en su desempeño puede deberse a que el conjunto de validación estaba formado por casos relativamente sencillos a diferencia del conjunto de entrenamiento donde encontró valores difíciles o atípicos de los que de aprender. Para los grupos 2 y 3 el modelo se desempeñó mal tanto en el conjunto de entrenamiento como en el de validación mostrando en ambos casos valores negativos para el coeficiente de determinación.

La red neuronal realiza un ajuste mejor que la media para el primer grupo etario, sin embargo no es lo suficientemente bueno como predictor, ya que este buen ajuste bien podria deberse casos particulares de las muestras utilizadas en el conjunto de validación porque en el conjunto de entrenamiento el desempeño medido por el coeficiente de determinación fue malo. Para los grupos 2 y 3 la red presenta errores mayores en el conjunto de validación en comparación al conjunto de entrenamiento y al igual que con el algoritmo XGBoost, el coeficiente de determinación es significativamente menor en el conjunto de validación por lo que es necesaria más información para poder realizar un buen ajuste.

Atenciones de Urgencia por enfermedades cardiovasculares

El algortimo XGBoost, al igual que en el caso anterior presenta errores menores en el conjunto de entranamiento en comparación al de validación, salvo en el grupo 1, donde ocurre lo contrario. Los coeficientes de determinación tambien son positivos y relativamente buenos durante el proceso de aprendizaje y negativos en el conjunto de validación.

El modelo lineal generalizado y la red neuronal presentan errores mayores en el entrenamiento para los grupos 1 y 2; y errores menores para el grupo 3, en relación al conjunto de validación. Además de coeficientes de determinación que indican que ambos modelos realizan un peor ajuste que la media al intentar predecir las atenciones de urgencia, en todos los casos.

5.1.3. Discusión

Los resultados de los tres modelos no arrojaron información relevante ni al considerar los ingresos hospitalarios por enfermedades respiratorias en general ni cuando estos se agrupaban por ciertas enfermedades respiratorias que, de acuerdo a la literatura, han mostrado evidencia de una posible relación significativa con el material particulado (Fernández, 2018); mientras que si hubo relaciones significativas en las atenciones de urgencia, especialmente al utilizar el modelo lineal generalizado. En las enfermedades cardiovasculares se presenta el mismo problema, solo observando un coeficiente de determinación alto al utilizar el modelo lineal generalizado para los ingresos hospitalarios en el primer grupo. La razon detrás del buen desempeño del GLM en estos casos se debe a que la distribución poisson lo inclina a tomar valores cercanos a la media donde la probabilidad de ocurrencia es mayor, por lo que su desempeño es bueno cuando la variable dependiente se acerca a estos valores, sin embargo su desempeño es débil cuando intenta predecir valores cercanos a los limites superior e inferior del dominio, lo que se condice directamente a lo presentado por Kassomenos et al. (2011) en su investigación. Entonces, tiene sentido que el GLM haya sido el que mostró mejores métricas en el conjunto de validación para aquellos casos considerados significativos, porque analizando la muestra, esta se encuentra en un rango cercano a la media, y si además se considera la gran cantidad de valores perdidos rellenados con la media, más del 34 % del $MP_{2,5}$ y 76 % del MP_{10}, la situación ilustrada es la más probable. Se espera que agregando más variables e información, tanto espacial como temporalmente se puedan lograr mejores inferencias, ya que este porcentaje esta limitado a las relaciones aprendidas en el periodo 2013-2017 entre el material particulado y las variables meteorológicas disponibles. Es importante tener presente que existen otros contaminantes y variables meteorológicas relevantes y con gran incidencia que no se han incluido en esta base de datos. Wang et al. (2008) en su estudio concluye que el NO_X es el contaminante mas influyente en la tasa de mortalidad debido a enfermedades respiratorias; y Kassomenos et al. (2011) situa al O_3, CO, NO_2 y el SO_2 como los siguientes en relevancia, tras el material particulado, al determinar la cantidad de ingresos hospitalarios. La circulación viral también consistituye una variable importante y difícil de predecir que causaría gran impacto en el desempeño del modelo.

Por otro lado, existen variables fenomenológicas propias de cada individuo y su situación, que tienen una gran influencia en las admisiones tanto hospitalarias como de urgencia. El grupo etario 1 depende completamente de la disponibilidad del adulto a cargo para acercarse a un centro de salud, así como también el conocimiento y capacidad de este para reconocer un problema de salud. Las condiciones climáticas extremas también pueden afectar al comportamiento de los individuos, por ejemplo es poco probable que un adulto mayor se acerque a un centro de salud si las temperaturas son extremadamente bajas, a pesar de que la contaminación puede ser más crítica en ese momento. Así también, el sexo puede ser un factor relevante al retratar los ingresos hospitalarios, ya que el momento de visitar un hospital o centro de salud puede variar significativamente si se trata de un hombre o una mujer, siendo esta última más propensa a retrasar su visita por estar cuidando a otros como Sancho Cantus et al. (2015) refleja en su estudio, donde investiga las causas fenomenológicas que afectan a la mujer y su consulta tardía al padecer una cardiopatía isquémica. Estos y otros factores no fueron tenidos en cuenta en el estudio y constituyen un aspecto importante a investigar.

En relación a la capacidad de los modelos propuestos, destaca la flexibilidad de la red neuronal al ser capaz de adaptarse a escenarios complejos e incluir los efectos de parámetros no lineales (Kassomenos et al., 2011), por lo que a pesar de su mal desempeño en esta base de datos, se considera una de las mejores opciones si se cuenta con mayor información. La desventaja, sin embargo, es la difícil interpretabilidad de las relaciones, ya que entre la entrada y salida de los datos la información fluye en números direcciones y se vuelve más compleja a medida que atraviesa las diferentes capas. XGBoost es quizás una mejor opción al presentar entrenamientos más eficientes y eficaces, además de proveer una estructura más sencilla de analizar. Cabe destacar la dificultad de aplicar estos modelos en ambientes complejos, con escasa información y comportamientos erráticos [2]. Es así como la región del Biobío, asi como el resto del país, son escenarios complicados para realizar pronósticos, debido principalmente a la falta de información.

[2]Se consideran comportamientos erráticos aquellos que rompen con la usual relación causa efecto, como por ejemplo no encender la estufa cuando hace frío debido al cansancio

Capítulo 6
Conclusiones y trabajo futuro

En este trabajo se abordó la problemática de la contaminación atmosférica y sus efectos en la salud. Se propusieron tres modelos, una red neuronal artificial (MLP), un árbol de regresión (XGBoost) y un modelo estadístico tradicional (GLM). El algoritmo XGboost presentó resultados relevantes para las atenciones de urgencia por enfermedades respiratorias para el grupo etario 3, logrando explicar el 12,33 % de la varianza. El GLM fue capaz de explicar el 48,55 % y el 21,48 % de la varianza en los grupos 2 y 3, respectivamente. El desempeño de estos algortimos fue limitado, y en algunos casos contraituitivo, ya que inicialmente se esperaba que tanto la red neuronal como el XGBoost realizaran mejores predicciones que el GLM. Sin embargo, debido a la calidad de los datos y un conjunto de validación con valores cercanos a la media, el GLM superó a ambos algoritmos en las atenciones de urgencia por enfermedades respiratorias. La recomendación, si se cuenta con una buena base de datos (extensa y con información consistente), es utilizar el algoritmo XGBoost debido a su eficiencia y eficacia. Para su implementación es de suma relevancia la elección de los hiperparámetros, ya que tienen influencia directa en el desempeño del algoritmo, determinando o no la posibilidad de sobreajuste y el tiempo de procesamiento; de forma general se aconseja la utilización de estimadores de árboles estructurados. Si se cuenta con una base de datos reducida se sugiere utilizar métodos mas sencillos como el GLM u otros métodos estadísticos tradicionales.

Las principales limitaciones en el desarrollo de este trabajo fueron las asociadas a la disponibilidad de información, la que era escasa y de difícil acceso. La falta de un sistema de información centralizado, coordinado y consistente retrasaron el trabajo y dificultaron el análisis. La capacidad del hardware y los tiempos asociados al procesamiento de datos, también fueron limitantes importantes. En relación a las dificultades, la especificidad del conocimiento necesario para la aplicación de los métodos se considera una de las más relevantes. Es así, que para estudios futuros, se considera como principal desafío la creación de una base de datos robusta. Junto con esto, se considera el desarrollo de un análisis de sensibilidad con las diferentes imputaciones, considerando distintos escenarios de calidad del aire y sus repercusiones en los ingresos hospitalarios y atenciones de urgencia.

Capítulo 7
Referencias

Agay-Shay, K., Friger, M., Linn, S., Peled, A., Amitai, Y., and Peretz, C. (2013). Air pollution and congenital heart defects. *Environmental research*, 124:28–34.

Agrawal, A. (2017). Loss functions and optimization algorithms. demystified.

Algorithmia (2018). Introduction to loss functions.

Becerril-Montekio, V., Reyes, J. d. D., and Manuel, A. (2011). Sistema de salud de chile. *Salud pública de México*, 53:s132–s142.

Bellinger, C., Jabbar, M. S. M., Zaïane, O., and Osornio-Vargas, A. (2017). A systematic review of data mining and machine learning for air pollution epidemiology. *BMC public health*, 17(1):907.

Bergstra, J. and Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13(Feb):281–305.

Bergstra, J. S., Bardenet, R., Bengio, Y., and Kégl, B. (2011). Algorithms for hyper-parameter optimization. In *Advances in neural information processing systems*, pages 2546–2554.

Biblioteca del Congreso Nacional de Chile (2017). Reporte décima circunscripción electoral 2017. Technical report, Biblioteca del Congreso Nacional de Chile.

Bronshtein, A. (2017). Train/test split and cross validation in python–towards data science. *Towards Data Science*.

Brownlee, J. (2016). A gentle introduction to xgboost for applied machine learning.

Brownlee, J. (2017). *Deep Learning with Python: Develop Deep Learning Models on Theano and TensorFlow Using Keras*. Machine Learning Mastery.

Buuren, S. v. and Groothuis-Oudshoorn, K. (2010). mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, pages 1–68.

Changhau, I. (2017). Loss functions in neural networks.

Chile, M. S. G. d. l. P. (1994). Ley de bases del medio ambiente, ley 19300.

Cochrane, C. (2018). Time series nested cross-validation. *Towards Data Science*.

Departamento de Estudios (2015). Síntesis regional, región del biobío. Technical report, Consejo Nacional de la Cultura y las Artes.

Dewancker, I., McCourt, M., and Clark, S. (2015). Bayesian optimization primer.

Dominici, F., Peng, R. D., Bell, M. L., Pham, L., McDermott, A., Zeger, S. L., and Samet, J. M. (2006). Fine particulate air pollution and hospital admission for cardiovascular and respiratory diseases. *Jama*, 295(10):1127–1134.

Eum, Y., Song, I., Kim, H.-C., Leem, J.-H., and Kim, S.-Y. (2015). Computation of geographic variables for air pollution prediction models in south korea. *Environmental health and toxicology*, 30.

Fallah, N., Gu, H., Mohammad, K., Seyyedsalehi, S. A., Nourijelyani, K., and Eshraghian, M. R. (2009). Nonlinear poisson regression using neural networks: a simulation study. *Neural Computing and Applications*, 18(8):939.

Fernández, A. (2018). Estudio del impacto de las concentraciones de mp2,5 y mp10 en las enfermedades respiratorias de la población de la comuna de los Ángeles, chile. Memoria de Título.

Galván, M. (2007). *Imputación de datos: teoría y práctica*, volume 54. United Nations Publications.

Garnett, R. (2019). Bayesian optimization.

Géron, A. (2017). *Hands-on machine learning with Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems*. .º'Reilly Media, Inc.".

Goic, A. (2015). El sistema de salud de chile: una tarea pendiente. *Revista médica de Chile*, 143(6):774–786.

Graham, J. W., Olchowski, A. E., and Gilreath, T. D. (2007). How many imputations are really needed? some practical clarifications of multiple imputation theory. *Prevention science*, 8(3):206–213.

Gray, C. (2018). Stock prediction with ml: Walk-forward modeling.

Grover, P. (2018). 5 regression loss functions all machine learners should know.

Honorable Junta de Gobierno (1980). Constitución política de la república de chile, s/e. *Santiago*.

Instituto Nacional de Estadísticas (2017). Resultados censo 2017. Technical report, Instituto Nacional de Estadísticas.

Karpathy, A. (2019). Convolutional neural networks for visual recognition. Apuntes de clases.

Kassomenos, P., Petrakis, M., Sarigiannis, D., Gotti, A., and Karakitsios, S. (2011). Identifying the contribution of physical and chemical stressors to the daily number of hospital admissions implementing an artificial neural network model. *Air Quality, Atmosphere & Health*, 4(3-4):263–272.

Khreis, H., de Hoogh, K., and Nieuwenhuijsen, M. J. (2018). Full-chain health impact assessment of traffic-related air pollution and childhood asthma. *Environment international*, 114:365–375.

Koehrsen, W. (2018). An introductory example of bayesian optimization in python with hyperopt.

Krogh, A. (2008). What are artificial neural networks? *Nature biotechnology*, 26(2):195.

L'Huillier, G. and Weber, R. (2010). Introducción a la minería de datos.

Lighty, J. S., Veranth, J. M., and Sarofim, A. F. (2000). Combustion aerosols: factors governing their size and composition and implications to human health. *Journal of the Air & Waste Management Association*, 50(9):1565–1618.

Mardones, C., Saavedra, A., and Jiménez, J. (2015). Cuantificación económica de los beneficios en salud asociados a la reducción de la contaminación por mp10 en concepción metropolitano, chile. *Revista médica de Chile*, 143(4):475–483.

McCulloch, W. S. and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133.

Mihelcic, J. and Zimmerman, J. (2012). Ingeniería ambiental. *Fundamentos, sustentabilidad y diseño. Alfaomega, México DF, México*.

Ministerio del Medio Ambiente (2017). Tercer reporte del estado del medio ambiente. Technical report, Ministerio del Medio Ambiente.

Ng, A. (2018). *Machine Learning Course*. Stanford and Coursera.

Observatorio Chileno de Salud Pública (n.d.a). Los servicios de salud del s.n.s.s.

Observatorio Chileno de Salud Pública (n.d.b). Organización y estructura del sistema de salud.

OMS (2006). Guiías de calidad del aire de la oms relativas al material particulado, el ozono, el dioóxido de nitroógeno y el dióxido de azufre: actualizacioón mundial 2005. Technical report, Ginebra:Organizacioón Mundial de la Salud.

OMS (2018). Calidad del aire y salud.

Oreña, A. (2018). Mapa región biobío.

Orellana, J. (2018). Arboles de decision y random forest.

Orjuela, E. J. R., Rozo, M. E. F., and Gómez, O. A. B. (2018). Evaluación y comparación de métodos de imputación múltiple implementados en el paquete mice de r.

Oyarzún, M. (2010). Contaminación aérea y sus efectos en la salud. *Revista chilena de enfermedades respiratorias*, 26(1):16–25.

Palacios, A., América, L., et al. (1997). Contaminación ambiental. origen, clases, fuentes y efectos. In *Introducción a la toxicología ambiental*, pages 37–52. ECO.

Pham, V. (2016). Bayesian optimization for hyperparameter tuning.

Raaschou-Nielsen, O., Andersen, Z. J., Beelen, R., Samoli, E., Stafoggia, M., Weinmayr, G., Hoffmann, B., Fischer, P., Nieuwenhuijsen, M. J., Brunekreef, B., et al. (2013). Air pollution and lung cancer incidence in 17 european cohorts: prospective analyses from the european study of cohorts for air pollution effects (escape). *The lancet oncology*, 14(9):813–822.

Ruder, S. (2016). An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*.

Salian, I. (2018). Supervize me: What's the difference between supervised, unsupervised, semi-supervised and reinforcement learning. *Nvidia (blog), August*, 2.

Sancho Cantus, D., Solano Ruiz, C., and Solera Gómez, S. (2015). Conocimientos previos de la mujer en la enfermedad coronaria: un estudio fenomenológico. *Index de Enfermería*, 24(3):129–133.

Semmartin, M. (2013). Contaminación atmosférica.

Servicio de Salud Biobío, Ministerio de Salud (n.d.). Nuestros establecimientos.

Servicio de Salud Metropolitano Sur, Ministerio de Salud (2018a). Atención primaria.

Servicio de Salud Metropolitano Sur, Ministerio de Salud (2018b). Red de urgencia.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Suárez, C. A. A. (2012). Diagnóstico y control de material particulado: partículas suspendidas totales y fracción respirable pm10. *Revista Luna Azul*, (34):195–213.

Tonne, C., Yanosky, J. D., Beevers, S., Wilkinson, P., and Kelly, F. J. (2012). Pm mass concentration and pm oxidative potential in relation to carotid intima-media thickness. *Epidemiology*, pages 486–494.

Universidad de Valencia (2007). Regresión. Apuntes de clases.

Universidad de Valencia (2011). Árboles de clasificación y regresión.

Waljee, A. K., Mukherjee, A., Singal, A. G., Zhang, Y., Warren, J., Balis, U., Marrero, J., Zhu, J., and Higgins, P. D. (2013). Comparison of imputation methods for missing laboratory data in medicine. *BMJ open*, 3(8):e002847.

Wang, Q., Liu, Y., and Pan, X. (2008). Atmosphere pollutants and mortality rate of respiratory diseases in beijing. *Science of the Total Environment*, 391(1):143–148.

Watson, J. G., Chow, J. C., Chen, L., Wang, X., and Diamond Bar, C. (2010). Measurement system evaluation for fugitive dust emissions detection and quantification. *Prepared by Desert Research Institute, Reno, NV*.

White, I. R., Royston, P., and Wood, A. M. (2011). Multiple imputation using chained equations: issues and guidance for practice. *Statistics in medicine*, 30(4):377–399.

Witten, I. H., Frank, E., Hall, M. A., and Pal, C. J. (2016). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

Zhang, L.-w., Chen, X., Xue, X.-d., Sun, M., Han, B., Li, C.-p., Ma, J., Yu, H., Sun, Z.-r., Zhao, L.-j., et al. (2014). Long-term exposure to high particulate matter pollution and cardiovascular mortality: a 12-year cohort study in four cities in northern china. *Environment international*, 62:41–47.

Capítulo 7
Anexos

7.1. Mapa de flujo servicio de salud público

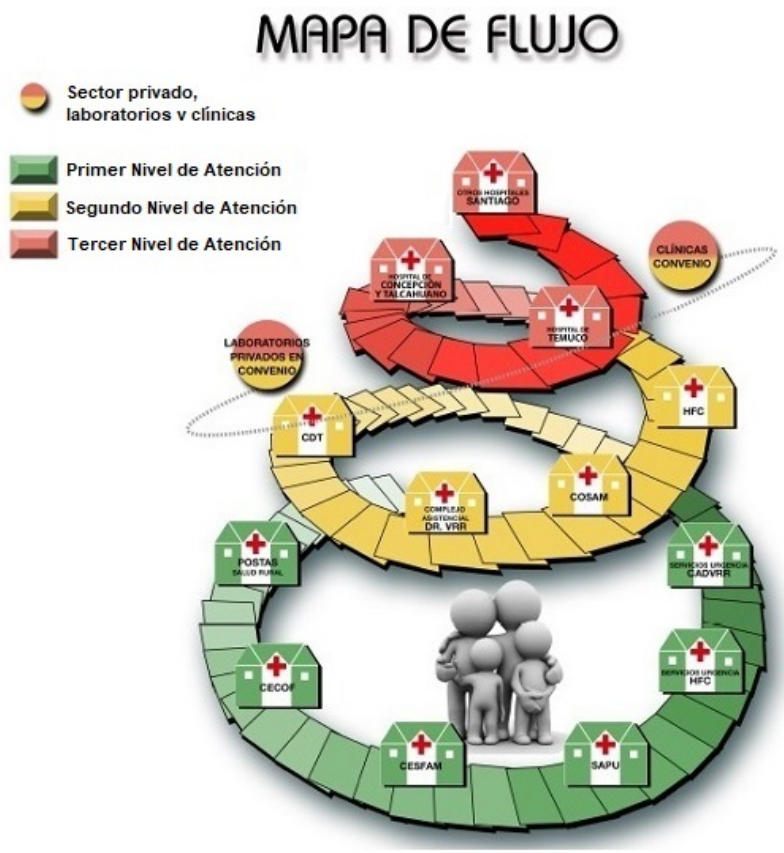


Figura 7.1: Mapa de Flujo

Fuente: Servicio de Salud Biobío, Ministerio de Salud (nd)

7.2. Lista de contaminantes y estaciones de calidad del aire

Tabla 7.1: Lista de compuestos y moléculas por estación

Calidad del aire		$MP_{2,5}$	MP_{10}	SO_2	NO_2	NO_X	NO	CO	O_3	CH_4	HCNM	HCT
X	Colicheu		X	X	X	X	X	X	X	X	X	X
2	Punteras	X	X	X	X	X	X					
3	Meteorología, Chiguayante	X	X									
4	INIA Chillan	X	X					X				
5	Puren	X	X									
6	Kingston College	X	X	X	X	X	X	X	X			
7	Cerro Merquín	X	X	X								
8	Calabozo			X								
9	Coronel Sur		X	X	X	X	X	X		X		X
10	Balneario Curanilahue	X	X	X								
11	Bocatoma		X									
12	Hualqui	X	X	X	X	X	X		X			
13	21 de mayo	X	X									
14	CESFAM, Los Angeles		X									
15	Los Angeles Oriente	X	X									
16	Cayumanqui			X								
17	Quillon			X	X			X	X			
18	Nueva Aldea		X	X	X			X	X			
19	MASISA Mapal		X	X								
20	Inpesca		X	X								
21	Nueva Libertad	X	X	X	X	X	X	X	X	X		X
22	Consultorio - San Vicente	X	X	X	X	X	X					
23	Libertad	X	X									
24	San Vicente, Bomberos			X								
25	Liceo Polivalente	X	X	X	X	X	X		X			

Fuente: Elaboración propia

7.3. Diccionarios

Tabla 7.2: Diccionario de base de datos egresos hospitalarios

Nombre del campo	**Descripción**
ServicioSalud o SER_SALUD	Código Servicio de Salud, corresponde al Servicio de Salud de ocurrencia. (Código de Servicio de Salud)
Seremi	Código SEREMI, corresponde a la SEREMI de ocurrencia
ESTAB	Código de Establecimiento, corresponde al Establecimiento de ocurrencia
SEXO	Sexo, los valores aceptados son: 1 = Hombre, 2 = Mujer, 3 = Indeterminado, 9 = Desconocido
EDAD	Edad en cantidad, expresada en años
PREVI	Previsión en salud
BENEF	Clase de Beneficiario, corresponde a los tramos de FONASA
MOD	Modalidad de Atención de los beneficiarios FONASA
PROCEDENCI	Procedencia de Paciente. 1=Unidad. Emergencia , 2=APS (Atención Primaria de Salud), 3=At. Especialidades, 4=Otro Hospital, 5=Otra Procedencia
COMUNA	Código Comuna
FECHA_EGR	Fecha de alta del paciente, formato fecha DD/MM/AAAA.
SERC_EGR	Servicio Clínico de Egreso o nivel de cuidado,
DIAS_ESTAD	Días estada total.
DIAG1	Diagnóstico principal, corresponde al código de la CIE-10

(Continúa en la página siguiente)

(Continuación Tabla 7.2)

Nombre del campo	**Descripción**
DIAG2	Causa externa, corresponde al código de la CIE-10 cuando el diagnostico principal corresponde a: ”Traumatismos, envenenamientos y algunas otras consecuencias de causas externas”
COND_EGR	Condición al egreso, los valores aceptados son: 1 = Vivo 2 = Fallecido
INTERV_Q	Intervención Quirúrgica, los valores aceptados son: 1 = Sí 2 = No
REGION	Región de residencia, campo creado a partir de la comuna de residencia del paciente, de acuerdo a la División Político Administrativo vigente desde el año 2008
SERV_RES	Servicio de Salud de referencia, campo creado a partir de la comuna de residencia del paciente, de acuerdo a la División Político Administrativo vigente desde el año 2008
NACIONALID	Nacionalidad (Código País)
ETNIA	Pueblo Originario Declarado

Fuente: Elaboración propia en base a los diccionarios del DEIS

Tabla 7.3: Diccionario de base de datos atenciones de urgencia

Campo	**Descripción**
IdEstablecimiento	Código de Establecimiento.
NEstablecimiento	Nombre de Establecimiento
IdCausa	Código de Causa
GlosaCausa	Descripción de Causa (Tabla 1)
Col01	Total de personas atendidas
Col02	Menores de 1 año
Col03	1 - 4 años
Col04	5 - 14 años
Col05	15 - 64 años
Col06	65 y más años
fecha	Fecha
semana	Semana Estadística del año correspondiente
GLOSATIPOESTABLECIMIENTO	Tipo de Establecimiento (Tabla 2)
GLOSATIPOATENCION	Tipo de Atención (Tabla 3)
GlosaTipoCampana	Tipo de Campaña (Tabla 4)

Fuente: DEIS

7.4. Correlaciones entre estaciones de monitoreo para el material particulado

Tabla 7.4: Correlaciones $MP_{2,5}$ para el promedio diario

	Colicheu	Punteras	Meteorología, Chiguayante	INIA Chillan	Puren	Kingston College	Cerro Merquín	Calabozo	Coronel Sur	Balneario Curanilahue	Bocatoma	Hualqui	21 de mayo	CESFAM, Los Angeles	Los Angeles Oriente	Cayumanqui	Quillon	Nueva Aldea	MASISA Mapal	Inpesca	Nueva Libertad	Consultorio - San Vicente	Libertad	San Vicente, Bomberos	Liceo Polivalente
Colicheu	1																								
Punteras		1				0.4	-0	1	1			0.8							1	1		0.3	1	1	
Meteorología, Chiguayante			1								1										1				
INIA Chillan				1	0.7																				
Puren				0.6	1																				
Kingston College		0.3				1	0	1	1			0.8							1	1		0.2	1	1	0.1
Cerro Merquín		-0				0	1	1	1			0.3							1						
Calabozo		1				1	1	1	1			1							1	1					
Coronel Sur		1				1	1	1	1			1							1						
Balneario Curanilahue										1															
Bocatoma			1								1										1				
Hualqui		0.8				0.8	0.3	1	1			1							1						
21 de mayo													1		0.5										
CESFAM, Los Angeles														1											
Los Angeles Oriente													0.2		1										
Cayumanqui																1	1	1							
Quillon																1	1	1							
Nueva Aldea																1	1	1							
MASISA Mapal		1				1	1	1	1			1							1	1		1	1	1	
Inpesca		1				1		1											1	1		1	1	1	1
Nueva Libertad			1								1										1				
Consultorio - San Vicente		0.3				0.2													1	1		1	1	1	0.4
Libertad		1				1													1	1		1	1	1	1
San Vicente, Bomberos		1				1													1	1		1	1	1	1
Liceo Polivalente						0.1														1		0.3	1	1	1

Fuente: Elaboración propia

Tabla 7.5: Correlaciones MP_{10} para el promedio diario

	Colicheu	**Punteras**	**Meteorología, Chiguayante**	**INIA Chillan**	**Puren**	**Kingston College**	**Cerro Merquín**	**Calabozo**	**Coronel Sur**	**Balneario Curanilahue**	**Bocatoma**	**Hualqui**	**21 de mayo**	**CESFAM, Los Angeles**	**Los Angeles Oriente**	**Cayumanqui**	**Quillon**	**Nueva Aldea**	**MASISA Mapal**	**Inpesca**	**Nueva Libertad**	**Consultorio - San Vicente**	**Libertad**	**San Vicente, Bomberos**	**Liceo Polivalente**
Colicheu	1																								
Punteras		1				0.4	-0	1	1			0.8							1	1		0.3	1	1	
Meteorología, Chiguayante			1								1										1				
INIA Chillan				1	0.7																				
Puren				0.6	1																				
Kingston College		0.3				1	0	1	1			0.8							1	1		0.2	1	1	0.1
Cerro Merquín		-0				0	1	1	1			0.3							1						
Calabozo		1				1	1	1	1			1							1	1					
Coronel Sur		1				1	1	1	1			1							1						
Balneario Curanilahue										1															
Bocatoma			1								1										1				
Hualqui		0.8				0.8	0.3	1	1			1							1						
21 de mayo													1		0.5										
CESFAM, Los Angeles														1											
Los Angeles Oriente													0.2		1										
Cayumanqui																1	1	1							
Quillon																1	1	1							
Nueva Aldea																1	1	1							
MASISA Mapal		1				1	1	1	1			1							1	1		1	1	1	
Inpesca		1				1		1											1	1		1	1	1	1
Nueva Libertad			1								1										1				
Consultorio - San Vicente		0.3				0.2													1	1		1	1	1	0.4
Libertad		1				1													1	1		1	1	1	1
San Vicente, Bomberos		1				1													1	1		1	1	1	1
Liceo Polivalente						0.1														1		0.3	1	1	1

Fuente: Elaboración propia

7.5. Cantidad de valores perdidos $MP_{2,5}$ y MP_{10}

7.5.1. Registros no validados (2000-2018)

Tabla 7.6: Cantidad de valores perdidos $MP_{2,5}$ y MP_{10} (2000-2018).

(Registros no validados)

Calidad del aire		Valores perdidos $MP_{2,5}$	Porcentaje $MP_{2,5}$	Valores perdidos MP_{10}	Porcentaje MP_{10}
1	Colicheu	1	100.00 %	0.6917890	69.18 %
2	Punteras	0.5949955	59.50 %	0.6263111	62.63 %
3	Meteorología, Chiguayante	0.9953551	99.54 %	0.9920587	99.21 %
4	INIA Chillan	0.6513335	65.13 %	0.6397962	63.98 %
5	Puren	0.6511837	65.12 %	0.6504345	65.04 %
6	Kingston College	0.6702128	67.02 %	0.5922985	59.23 %
7	Cerro Merquín	0.6050345	60.50 %	0.7358406	73.58 %
8	Calabozo	1	100.00 %	1	100.00 %
9	Coronel Sur	1.0000000	100.00 %	0.6514834	65.15 %
10	Balneario Curanilahue	0.6961343	69.61 %	0.8957147	89.57 %
11	Bocatoma	1	100.00 %	0.6750075	67.50 %
12	Hualqui	0.9072520	90.73 %	0.9039556	90.40 %
13	21 de mayo	0.6547797	65.48 %	0.6537309	65.37 %
14	CESFAM, Los Angeles	1	100.00 %	0.9249326	92.49 %
15	Los Angeles Oriente	0.7947258	79.47 %	0.9036560	90.37 %
16	Cayumanqui	1	100.00 %	1	100.00 %
17	Quillon	1	100.00 %	1	100.00 %
18	Nueva Aldea	1	100.00 %	0.8744381	87.44 %
19	MASISA Mapal	1	100.00 %	1	100.00 %
20	Inpesca	1	100.00 %	0.4270303	42.70 %
21	Nueva Libertad	0.5205274	52.05 %	0.5205274	52.05 %
22	Consultorio - San Vicente	0.6766557	67.67 %	0.6308061	63.08 %
23	Libertad	0.5858556	58.59 %	0.5867546	58.68 %
24	San Vicente, Bomberos	1	100.00 %	1	100.00 %
25	Liceo Polivalente	0.7289482	72.89 %	0.6603236	66.03 %

Fuente: Elaboración propia en base a datos del SINCA

7.5.2. Registros validados (2000-2018)

Tabla 7.7: Cantidad de valores perdidos $MP_{2,5}$ y MP_{10} (2000-2018).

(Registros validados)

Calidad del aire		Valores perdidos $MP_{2,5}$	Porcentaje $MP_{2,5}$	Valores perdidos MP_{10}	Porcentaje MP_{10}
1	Colicheu	1	100.00 %	0.691789	69.18 %
2	Punteras	0.6559784	65.60 %	0.6263111	62.63 %
3	Meteorología, Chiguayante	0.9992508	99.93 %	0.9920587	99.21 %
4	INIA Chillan	0.6684147	66.84 %	0.6397962	63.98 %
5	Puren	0.6528319	65.28 %	0.6504345	65.04 %
6	Kingston College	0.7383878	73.84 %	0.5922985	59.23 %
7	Cerro Merquín	0.6397962	63.98 %	0.7358406	73.58 %
8	Calabozo	1	100.00 %	1	100.00 %
9	Coronel Sur	1	100.00 %	0.6514834	65.15 %
10	Balneario Curanilahue	0.7401858	74.02 %	0.8957147	89.57 %
11	Bocatoma	1	100.00 %	0.6750075	67.50 %
12	Hualqui	0.9078514	90.79 %	0.9039556	90.40 %
13	21 de mayo	0.6561283	65.61 %	0.6537309	65.37 %
14	CESFAM, Los Angeles	1	100.00 %	0.9249326	92.49 %
15	Los Angeles Oriente	0.799071	79.91 %	0.903656	90.37 %
16	Cayumanqui	1	100.00 %	1	100.00 %
17	Quillon	1	100.00 %	1	100.00 %
18	Nueva Aldea	1	100.00 %	0.8744381	87.44 %
19	MASISA Mapal	1	100.00 %	1	100.00 %
20	Inpesca	1	100.00 %	0.4270303	42.70 %
21	Nueva Libertad	0.5211268	52.11 %	0.5205274	52.05 %
22	Consultorio - San Vicente	0.6905904	69.06 %	0.6308061	63.08 %
23	Libertad	0.5861552	58.62 %	0.5867546	58.68 %
24	San Vicente, Bomberos	1	100.00 %	1	100.00 %
25	Liceo Polivalente	0.7695535	76.96 %	0.6603236	66.03 %

Fuente: Elaboración propia en base a datos del SINCA

7.5.3. Registros validados (2008-2017)

Tabla 7.8: Cantidad de valores perdidos $MP_{2,5}$ y MP_{10} (2008-2017).

(Registros validados)

Calidad del aire		Valores perdidos $MP_{2,5}$	Porcentaje $MP_{2,5}$	Valores perdidos MP_{10}	Porcentaje MP_{10}
1	Colicheu	0.9994528	99.95 %	0.4897401	48.97 %
2	Punteras	0.4388509	43.89 %	0.3838577	38.39 %
3	Meteorología, Chiguayante	0.9980848	99.81 %	0.9849521	98.50 %
4	INIA Chillan	0.4596443	45.96 %	0.4098495	40.98 %
5	Puren	0.4328317	43.28 %	0.429275	42.93 %
6	Kingston College	0.5222982	52.23 %	0.3236662	32.37 %
7	Cerro Merquín	0.4090287	40.90 %	0.5170999	51.71 %
8	Calabozo	0.9994528	99.95 %	0.9994528	99.95 %
9	Coronel Sur	0.9994528	99.95 %	0.4320109	43.20 %
10	Balneario Curanilahue	0.5937073	59.37 %	0.8090287	80.90 %
11	Bocatoma	0.9994528	99.95 %	0.672777	67.28 %
12	Hualqui	0.8949384	89.49 %	0.8916553	89.17 %
13	21 de mayo	0.4399453	43.99 %	0.4361149	43.61 %
14	CESFAM, Los Angeles	0.9994528	99.95 %	0.8623803	86.24 %
15	Los Angeles Oriente	0.6708618	67.09 %	0.8235294	82.35 %
16	Cayumanqui	0.9994528	99.95 %	0.9994528	99.95 %
17	Quillon	0.9994528	99.95 %	0.9994528	99.95 %
18	Nueva Aldea	0.9994528	99.95 %	0.7701778	77.02 %
19	MASISA Mapal	0.9994528	99.95 %	0.9994528	99.95 %
20	Inpesca	0.9994528	99.95 %	0.3521204	35.21 %
21	Nueva Libertad	0.1931601	19.32 %	0.1926129	19.26 %
22	Consultorio - San Vicente	0.499316	49.93 %	0.3896033	38.96 %
23	Libertad	0.8968536	89.69 %	0.89658	89.66 %
24	San Vicente, Bomberos	0.9994528	99.95 %	0.9994528	99.95 %
25	Liceo Polivalente	0.6445964	64.46 %	0.4470588	44.71 %

Fuente: Elaboración propia en base a datos del SINCA

7.6. Hiperparámetros

Tabla 7.9: Valores finales hiperparámetros

	MLP		**XGBoost**				
Modelos	**Unidades**	**Capas**	**eta**	**Profundidad máxima**	**Suma minima de los pesos en nodos hijos**	**Reducción mínima de la perdida**	**Fracción de columnas**
IG1_R	6	10	0.45	12	2	0.9	0.85
IG2_R	17	14	0.325	9	3	0.6	0.5
IG3_R	11	20	0.075	12	3	0.65	0.65
IG1_C	12	16	0.375	12	5	0.1	0.55
IG2_C	11	12	0.075	0	6	0.75	0.5
IG3_C	5	8	0	9	6	0.4	0.9
UG1_R	11	4	0.4	9	4	0.65	0.85
UG2_R	14	20	0.475	0	6	0.7	0.85
UG3_R	12	16	0.4	5	2	0.55	0.9
UG1_C	6	12	0.3	6	1	0.2	0.85
UG2_C	2	20	0.34	1	6	0.45	0.95
UG3_C	15	12	0.45	6	3	0.1	0.75
IG1_R2	9	6	0.175	9	4	0.7	0.75
IG2_R2	6	4	0.35	3	5	0.4	0.9
IG3_R2	9	16	0.125	1	3	0.6	0.55

Fuente: Elaboración propia

7.7. Coeficiente de determinación

Tabla 7.10: Coeficiente de determinación de los distintos modelos

	Entrenamiento			**Validación**		
Modelos	**XGBoost x(std)**	**GLM x(std)**	**MLP x(std)**	**XGBoost x(std)**	**GLM x(std)**	**MLP x(std)**
IG1_R	**0.6684 (0.251)**	0.3303 (0.0023)	-0.1519 (0.7363)	**-0.9768 (1.6065)**	-1.1367 (0.8916)	-4.0687 (8.3203)
IG2_R	**0.335 (0.394)**	-0.1317 (0.0055)	0.1231 (0.0695)	-0.1633 (0.3922)	-0.5189 (0.3263)	**-0.1155 (0.5552)**
IG3_R	**0.215 (0.2237)**	-0.0691 (0.0025)	0.122 (0.0654)	-0.6796 (0.5533)	-0.9248 (0.4692)	**-1.0311 (1.0998)**
IG1_C	**0.5236 (0.2674)**	-0.0877 (0.0006)	-0.0229 (0.0701)	**0.0711 (0.1653)**	0.7083 (0.5052)	0.349 (0.5069)
IG2_C	**0.2843 (0.6139)**	-0.0051 (0.0016)	-0.1474 (0.6835)	-0.4114 (0.5346)	-0.58 (0.7806)	**-1.0022 (1.7646)**
IG3_C	**0.6087 (0.3566)**	-0.0153 (0.0015)	0.0679 (0.0753)	**-0.7344 (1.7099)**	-0.4105 (0.8203)	-1.1105 (2.185)
UG1_R	**0.8169 (0.2029)**	0.6234 (0.0009)	-0.2526 (1.5425)	-1.6888 (2.6376)	**0.0296 (1.344)**	-14.1029 (19.7581)
UG2_R	0.5806 (0.3608)	**0.6648 (0.0005)**	-0.1113 (0.6099)	-0.7398 (2.6923)	**0.4855 (0.2581)**	-0.9963 (1.188)
UG3_R	**0.8685 (0.1082)**	0.6151 (0.0009)	0.1844 (0.2038)	0.1233 (0.2559)	**0.2148 (0.2685)**	-0.053 (0.2958)
UG1_C	**0.1609 (0.084)**	-0.1198 (0.0037)	0.1094 (0.1826)	**-0.2298 (0.5054)**	-0.4277 (0.6442)	-0.2412 (0.4474)
UG2_C	**0.5537 (0.2257)**	0.0771 (0.0043)	-1.1899 (2.1605)	-0.6606 (0.8108)	**-0.4799 (0.5479)**	-2.6443 (4.3184)
UG3_C	**0.6041 (0.3403)**	0.1545 (0.0019)	0.1372 (0.0786)	-0.8758 (1.1691)	-0.6893 (1.0313)	**-0.4712 (0.8524)**
IG1_R2	**0.5427 (0.2597)**	0.324 (0.0019)	0.3149 (0.0555)	-1.2813 (2.6627)	-1.1619 (0.8675)	**-1.2481 (3.0864)**
IG2_R2	**0.2984 (0.3177)**	-0.1474 (0.0046)	-0.1217 (0.4236)	**-0.0558 (0.1771)**	-0.6459 (0.6039)	-0.429 (0.6706)
IG3_R2	**0.2322 (0.1067)**	-0.1241 (0.002)	0.115 (0.0703)	**-0.3051 (0.4511)**	-1.1482 (0.9542)	-0.3183 (0.2247)

Fuente: Elaboración propia

UNIVERSIDAD DE CONCEPCIÓN - FACULTAD DE INGENIERÍA
RESUMEN DE MEMORIA DE TÍTULO

Departamento de Ingeniería Industrial			
Título		Desarrollo de modelos predictivos para enfermedades respiratorias y cardiovasculares en zonas pobladas de la región del Biobío	
Nombre Memorista		María Ignacia Carrasco Huaiquián	
Modalidad	Investigación	Profesor(es) Patrocinante(s)	
Concepto		Rodrigo de la Fuente	
Calificación			
Fecha	Agosto 2019	Ingeniero Supervisor	Institución
Comisión (Nombre y Firma)			
Jorge Jiménez			
Resumen			

La contaminación del aire es un problema que afecta no solo a los humanos, sino a todo el ecosistema, deteriorando la salud de cada ser vivo y contribuyendo al cambio climático. En la región del Biobío, al igual que en otras zonas del sur del pais, se registran elevados niveles de material particulado, debido principalmente a la combustión de leña a nivel residencial. Diversas investigaciones han demostrado correlaciones positivas entre efectos adversos en la salud y un aumento en la concentración de material particulado (Tonne et al., 2012; Raaschou-Nielsen et al., 2013; Zhang et al., 2014). El objetivo de este trabajo es construir un modelo para predecir la tasa de ingresos hospitalarios y atenciones de urgencia asociadas a patologías cardiovasculares y respiratorias. En virtud de lo anterior es necesario desarrollar una base de datos, la que se construyó con información sobre las condiciones meteorológicas y de calidad de del aire para la región del Biobío; donde el principal problema fue la cantidad de valores pérdidos, los que se completaron utilizando la imputación múltiple. Para realizar las predicciones fueron desarrollados tres algoritmos. Se elaboró una red neuronal artificial simple (*multilayer perceptron*), un árbol de regresión (*XGBoost*) y un modelo lineal generalizado. Se utilizó la validación cruzada hacia adelante para obtener las predicciones y analizar el desempeño de los algoritmos, y modelos de optimización sequencial (tree parzen estimators) para obtener los hiperparámetros.

Debido al tiempo computacional necesario para ejecutar los algoritmos en una base de datos tan extensa, estos se probaron en una mas pequeña, previamente desarrollada en otra investigación, con información de la comuna de Los Ángeles desde el 2013 al 2017. Los resultados considerados relevantes fueron pocos. El algoritmo XGboost logró explicar el 12,33 % de la varianza para las atenciones de urgencia por enfermedades respiratorias para el grupo etario 3; y el GLM fue capaz de explicar el 48,55 % y el 21,48 % de la varianza en los grupos 2 y 3, respectivamente. El principal problema de esta base de datos fue la gran cantidad de observaciones perdidas, las que fueron completadas con la media, sesgando las estimaciones al homogeneizar la base de datos y reducir su varianza. Es por esto que los resultados no se consideran concluyentes en cuanto a reflejar la capacidad de los algoritmos descritos se refiere.