

Modelo predictivo para la contaminación del aire en comunas de la Región Metropolitana

TRABAJO DE TITULACIÓN PARA OPTAR AL TÍTULO DE INGENIERO CIVIL
EN COMPUTACIÓN MENCIÓN INFORMÁTICA

PROFESOR GUÍA	ESTUDIANTE
Víctor Hughes Escobar Jeria	Gonzalo Gabriel Salinas Campos
Ingeniero Civil en Computación mención Informática Doctor en Informática – Sistemas Inteligentes Licenciado en Ciencias de la Ingeniería	Estudiante de Ingeniería Civil en Computación mención Informática
vescobar@utem.cl	gonzalo.salinasc@utem.cl

AUTORIZACIÓN PARA LA REPRODUCCIÓN DEL TRABAJO DE TITULACIÓN

Identificación del trabajo de titulación

- **Nombre del estudiante:** Gonzalo Gabriel Salinas Campos.
- **Rut:** 19.291.586-4.
- **Dirección:** psj. Vishnu #2901, Maipú, Santiago, Chile.
- **E-mail:** gonzalo.salinasc@udem.cl.
- **Teléfono:** +56 9 4081 6589.
- **Título del trabajo de título:** Modelo predictivo para la contaminación del aire en comunas de la Región Metropolitana.
- **Escuela:** Informática.
- **Carrera o programa:** Ingeniería Civil en Computación mención Informática.
- **Título al que opta:** Ingeniero Civil en Computación mención Informática.

Autorización de Reproducción (seleccione una opción)

- a) Este trabajo de titulación no puede reproducirse o transmitirse bajo ninguna forma o por ningún medio o procedimiento, sin permiso escrito del(os) autor(es), exceptuando la cita bibliográfica, resumen y metadatos que acreditan al trabajo y a su(s) autor(es).

Fecha: _____ Firma: _____

- b) Se autoriza la reproducción total o parcial de este trabajo de titulación, con fines académicos, por cualquier medio o procedimiento, incluyendo la cita bibliográfica que acredita al trabajo y a su autor. En consideración a lo anterior, se autoriza su reproducción de forma (marque con una X):

a. Inmediata

b. _____ A partir de la siguiente fecha: _____ (mes/año)

Fecha: _____

Firma: _____



Esta autorización se otorga en el marco de la ley N°17.336 sobre Propiedad Intelectual, con carácter gratuito y no exclusivo para la Institución.

NOTA OBTENIDA:

Firma y timbre de la autoridad responsable

*A ti Chemo, viejo amigo de la familia, y a ti,
abuelo Carlos, que en paz descansen,*

...a ti abuela Silvia, lo logré, me pude portar bien,

*...a ti abuela María, agradezco que puedas ver
hasta donde eh llegado,*

*...a mis padres y hermano, por aguantarme
en este proceso y apoyarme,*

...a ti Ignacia, eres fantástica, cambiaste mi mundo

*...a los amigos del Exodus, por todos
los buenos momentos.*

AGRADECIMIENTOS

Quiero agradecer a mi profesor guía, Victor, por ofrecerme sus consejos y guía en este trabajo; a mis compañeros de universidad y amigos Daniel, Roberto y Kevin, por escuchar mis problemas y ofrecerme sus opiniones; a mi padre, Carlos, por todas esas conversaciones sobre mi proyecto y sus ideas tan ingeniosas; a mi hermano, Adolfo, por apoyar mis proyectos y mirar al futuro con ilusión; a mi madre, Claudia, por ser la mujer más cariñosa del mundo y apoyarme en los buenos y malos momentos. Finalmente quiero agradecer a todas esas personas que me ofrecieron su consejo y guía desinteresadamente durante el desarrollo de este trabajo, gracias por su tiempo.

RESUMEN

La ciudad de Santiago, es una de las ciudades con una contaminación del aire más elevada del mundo según el ranking del sitio IQair. IQAir es una empresa suiza de tecnología de la calidad del aire, y en base a su información Santiago de Chile se encuentra actualmente (día 29/04/2021) en el puesto N.^º 8 de las ciudades más contaminadas del mundo con un Índice de calidad del aire (AQI) que alcanza un 121 en su medición (ver Anexo 1), valor que la sitúa en una categoría de “Unhealthy For Sensitive Groups” (Malsano para grupos sensibles) [1]. Es importante regular los contaminantes ambientales que las industrias y las personas producen para evitar daños a la salud y al medioambiente que en su mayoría son irreparables.

El presente documento entrega la propuesta para un proyecto de investigación aplicada, el cual consiste en el modelamiento y posterior desarrollo de un prototipo funcional para la predicción de la contaminación del aire para comunas específicas de Santiago. Se pretende lograr la realización de dicho modelo utilizando la información del aire entregada por la estación móvil UTEM en conjunto con los datos del SINCA (Sistema de Información Nacional de Calidad del Aire), estas entidades presentan diversas variables que se pueden considerar para el desarrollo del proyecto, tales como calidad del aire PM10, PM-Coarse, PM2.5, niveles de SO₂, NO₂, NOx, NO, CO, O₃, CH₄, entre otros datos similares.

PALABRAS CLAVE

Contaminación ambiental, Material particulado, Índice de calidad del aire (AQI), Machine Learning, Random Forest, comunas de la Región Metropolitana de Chile.

ABSTRACT

The city of Santiago is one of the cities with the highest air pollution in the world according to the ranking of the IQair site. IQAir is a Swiss air quality technology company, and based on its information, Santiago de Chile is currently (04/29/2021) in position No. 8 of the most polluted cities in the world with an Index of air quality (AQI) that reaches 121 in its measurement (see Annex 1), a value that places it in a category of “Unhealthy For Sensitive Groups” [1]. It is important to regulate the environmental pollutants that industries and people produce to avoid damage to health and the environment, most of which are irreparable.

This document provides the proposal for an applied research project, which consists of the modeling and subsequent development of a functional prototype for the prediction of air pollution for specific communes belonging to Santiago. It is intended to achieve the realization of said model using the air information provided by the UTEM mobile station in conjunction with the data from SINCA (National Air Quality Information System), these entities present various variables that can be considered for the development project, such as air quality PM10, PM-Coarse, PM2.5, levels of SO₂, NO₂, NOx, NO, CO, O₃, CH₄, among other similar data.

KEYWORDS

Environmental pollution, Particulate matter, Air quality index (AQI), Machine Learning, Random Forest, communes of the Metropolitan Region of Chile.

Índice de Contenido

CAPÍTULO 1 - ASPECTOS GENERALES	14
1.1. Descripción del proyecto de título	14
1.2. Objetivos	15
1.2.1. Objetivo general.....	15
1.2.2. Objetivos específicos.....	15
1.3. Alcances y limitaciones.....	16
1.3.1. Alcances	16
1.3.2. Limitaciones	16
1.4. Metodología de trabajo.....	16
1.5. Recursos.....	22
CAPÍTULO 2 - MARCO REFERENCIAL	23
2.1. Estudio bibliométrico y altmetrics.....	23
2.1.1. Análisis de resultados Búsqueda 1	24
2.1.1.1. Análisis de autores	24
2.1.1.2. Resultado de análisis de autores	27
2.1.1.3. Análisis de publicaciones y resultados	28
2.1.2. Análisis de resultados Búsqueda 2	31
2.1.2.1. Análisis de autores	31
2.1.2.2. Resultado de análisis de autores	34
2.1.2.3. Análisis de publicaciones y resultados	35
2.2. Marco teórico conceptual.....	38
2.2.1. Contaminación ambiental.....	38
2.2.2. Contaminación del agua.....	39
2.2.2.2. Contaminación del suelo.....	42
2.2.3. Contaminación del aire	44
2.2.4. Atmósfera terrestre	46
2.2.5. Presión, densidad y temperatura de la atmósfera	46
2.2.5.1. Presión atmosférica	47
2.2.5.2. Densidad atmosférica.....	48
2.2.5.3. Temperatura atmosférica.....	49
2.2.5.4. Ecuación de estado	49
2.2.6. Composición de la atmósfera.....	52
2.2.7. Elementos y compuestos	53
2.2.7.1. Nitrógeno	53
2.2.7.2. Oxígeno	54
2.2.7.3. Argón.....	55

2.2.7.4. Neón.....	55
2.2.7.5. Helio.....	56
2.2.7.6. Criptón	57
2.2.7.7. Hidrógeno	57
2.2.7.8. Xenón.....	58
2.2.7.9. Vapor de agua	59
2.2.7.10. Dióxido de carbono	59
2.2.7.11. Metano.....	60
2.2.7.12. Óxido Nitroso	60
2.2.7.13. Monóxido de carbono.....	61
2.2.7.14. Ozono.....	62
2.2.7.15. Amoníaco	63
2.2.7.16. Dióxido de nitrógeno.....	64
2.2.7.17. Dióxido de azufre	64
2.2.7.18. Óxido nítrico.....	65
2.2.7.19. Sulfuro de hidrógeno.....	66
2.2.7.20. Material Particulado (PM10/PM2,5).....	67
2.2.8. Índices de contaminación ambiental	68
2.2.9. Área de estudio.....	72
2.2.10. Tecnologías y herramientas	72
2.2.10.1. Machine Learning.....	73
2.2.10.2. Random Forest.....	76
2.2.10.3. TensorFlow.....	80
2.2.10.4. Keras	81
2.2.10.5. NumPy.....	81
2.2.10.6. Pandas.....	81
2.2.10.7. Sklearn	82
2.2.11. Definición de la Solución	82
2.2.11.1. Detalle de solución.....	85
CAPÍTULO 3 - METODOLOGÍA DE INVESTIGACIÓN.....	86
3.1. Enfoque de la investigación	86
3.2. Alcances de la Investigación	87
3.3. Estructura de la investigación	87
3.3.1. Marco de trabajo orientador de investigación	87
3.3.2. Marco de trabajo operacional de investigación	89
3.3.2.1. Marco de trabajo operacional Etapa 1: Búsqueda de Información	89
3.3.2.2. Marco de trabajo operacional Etapa 2: Análisis de información	95
CAPÍTULO 4 - METODOLOGÍA DE DESARROLLO.....	98
4.1. Marco de trabajo orientador de desarrollo	98

4.2. Marco de trabajo operacional de desarrollo	99
4.2.1. Marco de trabajo operacional Etapa 3: Construcción del modelo	99
4.2.2. Marco de trabajo operacional Etapa 4: Entrenamiento y análisis.....	100
4.2.3. Marco de trabajo operacional Etapa 5: Interfaz e integración	101
CAPÍTULO 5 – DESARROLLO DEL MODELO.....	103
5.1. Limpieza del DataSet.....	103
5.1.1. Análisis Outliers columna PM25.....	103
5.1.2. Análisis Outliers columna PM10.....	104
5.1.3. Análisis Outliers columna O3	104
5.1.4. Análisis Outliers columna NO2.....	105
5.1.5. Análisis Outliers columna SO2.....	105
5.1.6. Análisis Outliers columna CO	106
5.2. Diseño de inputs y outputs	106
5.3. Descripción del DataSet	107
5.4. Generación del modelo	110
5.4.1. Algoritmo complementario.....	116
5.5. Diseño de la aplicación web	117
5.5.1. Mockup de la aplicación web.....	118
5.6. Resultados y análisis	120
CAPITULO 6 - CONCLUSIÓN Y TRABAJOS FUTUROS	125
6.1. Conclusiones.....	125
6.2. Trabajos futuros.....	128
BIBLIOGRAFÍA.....	130
ANEXOS	134
Anexo 1: Ranking de ciudades y contaminación.....	134
Anexo 2: Resultados de consultas a BDD POPv7	135
Anexo 3: Depuración y poblamiento de BDD	138
Anexo 4: Cálculo de campo “Total” en relevancia de publicaciones	141
Anexo 5: Publicaciones eliminadas de resultados	143
Anexo 6: Constantes.....	148
Anexo 7: Decisión de la Solución	149
Anexo 8: Algoritmo de complemento para el modelo.....	159

Índice de Ilustraciones

Ilustración 1.1: "Metodología del proyecto y de trabajo general"	21
Ilustración 2.1: "Comparación de tamaño de un cabello humano y PM10, PM2.5"	67
Ilustración 2.2: "Ejemplo de Árbol de decisión"	77
Ilustración 3.1: "Marco de trabajo orientador, investigación"	88
Ilustración 3.2: "Marco de trabajo operacional de investigación Etapa 1, parte 1"	89
Ilustración 3.3: "Marco de trabajo operacional de investigación Etapa 1, parte 2"	90
Ilustración 3.4: "Resumen de utilización de softwares para estudio bibliométrico"	91
Ilustración 3.5: "Modelo estrella de base de datos"	92
Ilustración 3.6: "Modelo BDD de almacenamiento de autores relevantes"	93
Ilustración 3.7: "Marco de trabajo operacional de investigación Etapa 1, parte 3"	94
Ilustración 3.8: "Marco de trabajo operacional de investigación Etapa 2, parte 1"	95
Ilustración 3.9: "Marco de trabajo operacional de investigación Etapa 2, parte 2"	96
Ilustración 4.1: "Marco de trabajo orientador, desarrollo"	98
Ilustración 4.2: "Marco de trabajo operacional de desarrollo Etapa 3"	99
Ilustración 4.3: "Marco de trabajo operacional de desarrollo Etapa 4"	101
Ilustración 4.4: "Marco de trabajo operacional de desarrollo Etapa 5"	102
Ilustración 5.1: "Dataset v1, sin outputs"	108
Ilustración 5.2: "Dataset v2, con output"	108
Ilustración 5.3: "Dataset final, sin output"	110
Ilustración 5.4: "Configuración del modelo"	111
Ilustración 5.5: "Ejecución de código: Generación del modelo y entrenamiento."	112
Ilustración 5.6: "Matriz de Confusión."	113
Ilustración 5.7: "Representación resumida: Modelo Random Forest."	114
Ilustración 5.8: "Resumen, funcionamiento de algoritmo complementario"	116
Ilustración 5.9: "Estructura de la aplicación web"	117
Ilustración 5.10: "Mockup de la aplicación web"	118
Ilustración 5.11: "Captura de la aplicación web"	119
Ilustración 7.1: "Live city ranking, air pollution (AQI)"	134
Ilustración 7.2: "Resultados consultas a bases de datos, búsqueda 1 idioma inglés"	135
Ilustración 7.3: "Resultados consultas a bases de datos, búsqueda 1 idioma español"	135

Ilustración 7.4: "Error 514 POPv7, No matching data found"	136
Ilustración 7.5: "Limpieza de datos de búsqueda 1"	136
Ilustración 7.6: "Resultados consultas a bases de datos, búsqueda 2 idioma inglés"	137
Ilustración 7.7: "Resultados consultas a bases de datos, búsqueda 2 idioma español"	137
Ilustración 7.8: "Limpieza de datos de búsqueda 2"	138
Ilustración 7.9: "Trabajo de datos, carga BDD autores búsqueda 1"	138
Ilustración 7.10: "Trabajo de datos, carga BDD autores búsqueda 2"	139
Ilustración 7.11: "Trabajo de datos, carga BDD resultados búsqueda 1"	139
Ilustración 7.12: "Trabajo de datos, carga BDD resultados búsqueda 2"	140
Ilustración 7.13: "Código de algoritmo LogisticRegression"	149
Ilustración 7.14: "Código de algoritmo DecisionTreeClassifier"	150
Ilustración 7.15: "Código de algoritmo RandomForestClassifier"	151
Ilustración 7.16: "Código de algoritmo AdaBoostClassifier"	152
Ilustración 7.17: "Código de algoritmo CatBoostClassifier"	153
Ilustración 7.18: "Código de algoritmo GaussianNB"	154
Ilustración 7.19: "Código de algoritmo SVC, linear"	155
Ilustración 7.20: "Código de algoritmo SVC, rbf"	156
Ilustración 7.21: "Código de algoritmo SVC, sigmoid"	157
Ilustración 7.22: "Código: Agrupar salidas por índices."	159
Ilustración 7.23: "Código: función DefinirCurva."	160
Ilustración 7.24: "Código: función rectaSubida."	160
Ilustración 7.25: "Código: función rectaBajada."	161
Ilustración 7.26: "Código: función curvaSubida."	161
Ilustración 7.27: "Código: función curvaBajada."	161

Índice de Tablas

Tabla 1.1: "Bases de datos utilizadas y su descripción"	17
Tabla 1.2: "Lista de conceptos Búsqueda 1"	18
Tabla 1.3: "Lista de conceptos Búsqueda 2"	18
Tabla 2.1: "Autores más relevantes de la Búsqueda 1"	28
Tabla 2.2: "Autores más relevantes de la Búsqueda 2"	35
Tabla 2.3: "Componentes permanentes presentes en la atmósfera"	52
Tabla 2.4: "Componentes variables presentes en la atmósfera"	52
Tabla 2.5: "Niveles de preocupación AQI EPA"	71
Tabla 2.6: "Niveles estándar por contaminante e indicador del AQI EPA"	72
Tabla 5.1: "Ingresos al modelo predictivo, comuna de Santiago"	120
Tabla 5.2: "Resultados del modelo predictivo, comuna de Santiago"	120
Tabla 5.3: "Ingresos al modelo predictivo, comuna de Pudahuel"	121
Tabla 5.4: "Resultados del modelo predictivo, comuna de Pudahuel"	121
Tabla 5.5: "Ingresos al modelo predictivo, comuna de El Bosque"	122
Tabla 5.6: "Resultados del modelo predictivo, comuna de El Bosque"	122
Tabla 5.7: "Ingresos al modelo predictivo, comuna de La Florida"	123
Tabla 5.8: "Resultados del modelo predictivo, comuna de La Florida"	123
Tabla 7.1: "Constante de avogadro, valores"	148
Tabla 7.2: "Constante de los gases ideales, valores"	148
Tabla 7.3: "Constante de Boltzmann, valores"	148

Índice de Gráficos

Gráfico 2.1: "Autores vs Productividad Científica últimos 5 años (Búsqueda 1, español)"	24
Gráfico 2.2: "Autores vs Productividad Científica últimos 5 años (Búsqueda 1, inglés)"	25
Gráfico 2.3: "Autores vs Cantidad de Citas últimos 5 años (Búsqueda 1, español)"	25
Gráfico 2.4: "Autores vs Cantidad de Citas últimos 5 años (Búsqueda 1, inglés)"	26
Gráfico 2.5: "Autores vs Relevancia por Obra Últimos 5 años (Búsqueda 1, español e inglés)"	27
Gráfico 2.6: "Publicaciones más relevantes de la búsqueda 1 en inglés"	29
Gráfico 2.7: "Publicaciones más relevantes de la búsqueda 1 en español"	30
Gráfico 2.8: "Autores vs Productividad Científica últimos 5 años (Búsqueda 2, español)"	31
Gráfico 2.9: "Autores vs Productividad Científica últimos 5 años (Búsqueda 2, inglés)"	32
Gráfico 2.10: "Autores vs Cantidad de Citas últimos 5 años (Búsqueda 2, español)"	33
Gráfico 2.11: "Autores vs Cantidad de Citas últimos 5 años (Búsqueda 2, inglés)"	33
Gráfico 2.12: "Autores vs Relevancia por Obra Últimos 5 años (Búsqueda 2, español e inglés)"	34
Gráfico 2.13: "Publicaciones más relevantes de la búsqueda 2 en inglés"	36
Gráfico 2.14: "Publicaciones más relevantes de la búsqueda 2 en español"	37
Gráfico 2.15: "Presión del aire v/s Altitud sobre el nivel del mar"	47
Gráfico 2.16: "Densidad del aire v/s Altitud sobre el nivel del mar"	49
Gráfico 5.1: "Columna pm25, Outliers"	103
Gráfico 5.2: "Columna pm10, Outliers"	104
Gráfico 5.3: "Columna o3, Outliers"	104
Gráfico 5.4: "Columna no2, Outliers"	105
Gráfico 5.5: "Columna so2, datos"	105
Gráfico 5.6: "Columna co, Outliers"	106
Gráfico 5.7: "Importancia de clases en los datos"	111
Gráfico 5.8: "Importancia de variables independientes"	115
Gráfico 7.1: "Publicaciones eliminadas, búsqueda 1 español"	144
Gráfico 7.2: "Publicaciones eliminadas, búsqueda 1 inglés"	145
Gráfico 7.3: "Publicaciones eliminadas, búsqueda 2 español"	146
Gráfico 7.4: "Publicaciones eliminadas, búsqueda 2 inglés"	147

CAPÍTULO 1 - ASPECTOS GENERALES

1.1. Descripción del proyecto de título

El proyecto consiste, en primera instancia, en una investigación exhaustiva acerca de la contaminación ambiental, concretamente la del aire, entender sus causas, evolución, efectos en las personas y en el entorno, puntos críticos, factores que contribuyen a su avance, puntos de no retorno y, sobre todo, los indicadores utilizados para detectar y medir su intensidad, para así determinar de forma correcta la verdadera información a considerar para comenzar a formular el modelo predictivo. Dicho modelo debe predecir de una forma efectiva y precisa los niveles de contaminación del aire en las comunas de Cerrillos, Cerro Navia, El Bosque, Independencia, la Florida, Las Condes, Santiago, Pudahuel, Puente Alto y Talagante. Para lograr lo anterior se utilizará un conjunto base de datos entregados por la unidad móvil UTEM y el SINCA (Sistema de Información Nacional de Calidad del Aire), para esto y después de la investigación anteriormente mencionada, se seleccionará la técnica más adecuada para la elaboración de dicho modelo realizando un análisis del estado de arte sobre proyectos similares e información relevante al tema.

Luego de seleccionar la técnica más adecuada para la realización del modelo, se procederá a su construcción la cual está separada en varias etapas:

- Construcción del algoritmo con la integración de los indicadores de contaminación seleccionados anteriormente.
- Aplicación de técnica seleccionada (se espera, alguna categoría dentro del campo de Machine Learning).
- Realizar tests de funcionamiento.
- Construcción de interfaz visual, web app.
- Comprobar el modelo.

Las etapas anteriormente mencionadas han sido listadas con la intención de ofrecer de forma más clara los puntos clave a lo largo del desarrollo de este proyecto, ya que en general, cada una de ellas tendrá una vasta extensión de subetapas y exploración dentro de diversos tópicos referentes al tema principal del proyecto. Finalmente, luego de todo lo anteriormente mencionado, se realizarán los procesos de término para el proyecto los cuales consisten en la construcción de manuales de usuario, realización de recomendaciones, entrega de documentación adecuada y credenciales, entre otros similares.

1.2. Objetivos

1.2.1. Objetivo general

- Desarrollar un Modelo predictivo para la contaminación del aire en las comunas donde se presente información histórica al respecto, para así, pronosticar situaciones respecto a la contaminación del aire en estas zonas.

1.2.2. Objetivos específicos

- Realizar un estudio bibliométrico y altmetric que permitan relevar información de importancia respecto a la caracterización del modelo predictivo en cuestión y antecedentes de trabajos similares que tengan relación con la contaminación ambiental.
- Analizar los puntos críticos de la contaminación del aire en Santiago o del mundo que puedan tener relevancia para el modelo, para así, establecer las variables de entrada y salida del Dataset a utilizar.
- Realizar una limpieza de datos y crear un Dataset el cual permita entrenar el modelo.
- Entrenar el modelo de Random Forest y perfeccionarlo para lograr los niveles de precisión esperados en sus resultados.
- Validar el modelo en test para realizar su integración a una aplicación web, la cual admita su utilización con datos reales.

1.3. Alcances y limitaciones

1.3.1. Alcances

- El modelo efectuará predicciones referentes a niveles de contaminantes del aire dentro de las comunas de Cerrillos, Cerro Navia, El Bosque, Independencia, la Florida, Las Condes, Santiago, Pudahuel, Puente Alto y Talagante, que son aquellas que disponen de datos para el modelo.
- El modelo solo realizará predicciones a través del uso de la aplicación web, la cual no cuenta con ningún registro de usuario ni de base de datos.

1.3.2. Limitaciones

- La entrega de resultados del modelo estará limitada a la Región Metropolitana y dentro de las comunas en estudio.
- Los resultados que el modelo entregara serán los niveles del índice AQI, no así su predicción exacta, sin embargo, se recreará este valor mediante un algoritmo complementario.
- Los resultados que entregará el modelo no serán precisos al 100%, debido a que se basa en datos estadísticos e históricos, donde no se contemplan hechos fortuitos y aislados, como, por ejemplo, un incendio forestal.
- El rango de predicción estará acotado a un máximo de 30 días futuros.

1.4. Metodología de trabajo

A continuación, se describe la metodología de trabajo que se utilizará para el proyecto, la cual divide el trabajo en etapas y fases.

Etapa 0: Inicio.

Esta etapa corresponde a la planificación del proyecto, donde se establecen los puntos claves para el inicio de este, tales como objetivos, alcances, fechas de hitos, entre otros similares.

Etapa 1: Búsqueda de información.

Esta etapa corresponde a la contextualización del proyecto, con el objetivo final de clarificar las causas de la contaminación del aire y cuáles son los factores o indicadores clave para medirlo. Para realizar esta tarea, se busca información en diversas bases de datos que albergan una basta cantidad de conocimiento científico, técnico y humano. Las cuales se listan a continuación:

Tabla 1.1: "Bases de datos utilizadas y su descripción"

N.º	Nombre BDD	Descripción
1	Scopus	Es una base de datos de resúmenes y citas de Elsevier lanzada en 2004. Scopus cubre casi 36.377 títulos de aproximadamente 11.678 editores, de los cuales 34.346 son revistas revisadas por pares en campos temáticos de alto nivel.
2	Google Scholar	Es un indexador de documentos académicos de acceso gratuito. Los tipos de textos que indexa son las siguientes citas, enlaces a libros, artículos de revistas científicas, comunicaciones y congresos, informes científicos-técnicos, tesis y archivos depositados en repositorios.
3	Crossref	Crossref interconecta millones de artículos en una variedad de tipos de contenido, incluidos diarios, libros, actas de congresos, documentos de trabajo, informes técnicos y conjuntos de datos.
4	Microsoft Academic (Project Academic Knowledge API)	API de consulta de datos que aprovecha la riqueza del contenido académico en Microsoft Academic Graph la cual cuenta con más de 10,000 transacciones por mes.

Fuente: Creación Propia.

Luego de establecer las principales fuentes de información, las cuales se muestran en la Tabla 1.1, se deben establecer las palabras claves para realizar las búsquedas en estas bases de datos, todo esto con el fin de ajustar los resultados lo máximo posible a los esperados, ya que los temas a tratar tanto de machine learning como de contaminación ambiental son muy extensos. Adicionalmente, se establecen dos búsquedas de información por separado pero que tienen cierto grado de relación entre ellas, la primera, está relacionada con el área científica del proyecto y trata sobre la contaminación ambiental y todos los factores que se puedan considerar sobre esta, con el propósito de analizar de raíz y de forma completa el contexto en el cual se debe desarrollar el modelo predictivo. La segunda búsqueda de información está enfocada en las soluciones actuales que se le dan al problema para así, establecer las bases del modelo predictivo que se desarrolla y robustecer su fundamento. Esta etapa del proyecto se divide en 4 fases:

- Fase 1: Búsqueda de Información contextualizadora e indicadores.
- Fase donde se realiza la búsqueda de información sobre la contaminación del ambiente en el mundo, además, se procede a buscar todos los posibles indicadores de contaminación del ambiente. Para la realización de esta búsqueda en bases de datos se utilizarán las siguientes conjugaciones de términos clave:

Tabla 1.2: “Lista de conceptos Búsqueda 1”

N.º	Concepto 1	Concepto 2	Concepto 3
1	Environmental pollution	Air quality	-
2	Environmental pollution	Air pollutants	-
3	Environmental pollution	Air pollutants	indicator
4	Environmental pollution	PM10	indicator
5	Environmental pollution	PM2,5	indicator
6	Environmental pollution	Air compounds	-
7	Environmental pollution	Air compounds	Breathable air

Fuente: Creación Propia.

Los conceptos de búsqueda indicados en la Tabla 1.2 serán ingresados a las bases de datos anteriormente mencionadas en la Tabla 1.1, con una unión de conector “AND” y se realizarán dichas búsquedas en dos idiomas diferentes, el inglés y el español.

- Fase 2: Búsqueda de modelos predictivos sobre la contaminación ambiental y trabajos similares.

Se buscan proyectos similares con el objetivo de obtener ejemplos de técnicas de desarrollo, metodologías, frameworks, y fundamentos para la solución propuesta en este trabajo. Para la realización de esta búsqueda en bases de datos se utilizarán las siguientes conjugaciones de términos clave:

Tabla 1.3: “Lista de conceptos Búsqueda 2”

N.º	Concepto 1	Concepto 2	Concepto 3
1	Predictive model	Environmental pollution	-
2	Predictive model	Air pollutants	-
3	Predictive model	Air pollutants	PM10
4	Predictive model	Air pollutants	PM2,5
5	Predictive model	Air quality	-
6	Predictive model	Breathable air	-

Fuente: Creación Propia.

Los conceptos de búsqueda indicados en la Tabla 1.3 serán ingresados a las bases de datos anteriormente mencionadas en la Tabla 1.1, con una unión de conector “AND” y se realizarán dichas búsquedas en dos idiomas diferentes, el inglés y el español.

- Fase 3: Análisis y Relevancia de Información.

Se selecciona la información obtenida en las fases anteriores a aquella que pueda presentar utilidad real para el proyecto. Para la realización de esta tarea se utiliza el software POPv7 (Publish or Perish Versión 7). POPv7 es un software de recuperación y análisis de citas académicas el cual ofrece la posibilidad de realizar búsquedas en las bases de datos mencionadas en la Tabla 1.1, y con los resultados, permite un análisis de relevancia en base a índices tales como: Índice h de Hirsch, Índice g de Egghe, Índice E de Zhang, entre otros. La información respecto a la metodología exacta de análisis y relevancia de información se encuentra detallada en el Capítulo N.^o 2 (punto 2.3. Estructura de la Investigación) del presente trabajo.

Etapa 2: Análisis de información.

Esta etapa determina cuales son los factores o indicadores que se deben tomar en consideración con mayor o menor relevancia para el modelo predictivo de contaminación. Esta etapa del proyecto se divide en 4 fases:

- Fase 1: Ordenamiento y categorización de indicadores.

Se ordenan los indicadores según su área de estudio con respecto al medio ambiente (aire, agua, entre otros.) y otras categorías que se relevarán en las investigaciones previas.

- Fase 2: Selección de información relevante para el modelo.

En base a la fase 1, se definen los indicadores a utilizar para el modelo y sus relevancias.

- Fase 3: Análisis de técnicas de desarrollo y estructuras del modelo predictivo.

- Se selecciona la técnica a utilizar para el desarrollo del modelo.
- Fase 4: Elaboración de Marco Teórico.

Se elabora un marco teórico para así obtener una imagen general de los conceptos y bases de la investigación del proyecto.

Etapa 3: Construcción del modelo, lógica y algoritmo.

Esta etapa se enfoca en la planificación y construcción del modelo, se divide en 2 fases:

- Fase 1: Diseño lógico.

Corresponde a la fase de definición de lógica de algoritmos y procesamiento de datos del modelo.
- Fase 2: Preparación del modelo.

Se diseña el modelo y se prepara para ser entrenado, sin ningún tipo de interfaz gráfica.

Etapa 4: Entrenamiento del modelo y análisis de resultados.

En esta etapa se utilizarán bancos de datos para el entrenamiento del modelo y luego, en base a parámetros previamente definidos, se analizarán sus resultados para realizar correcciones sobre el proceso. Esta etapa se divide en 3 fases:

- Fase 1: Entrenamiento del modelo.

Se otorga un banco de datos al modelo para su entrenamiento.
- Fase 2: Análisis de resultados.

Se comparan los resultados del modelo con parámetros definidos y se realizan correcciones de ser necesarias.
- Fase 3: Aprobación del funcionamiento del modelo.

Luego de las fases anteriores, y de haberse cumplido los resultados esperados del modelo, se aprueba su funcionamiento.

Etapa 5: Construcción de interfaz de usuario e integración.

Esta etapa se enfoca en el diseño visual del modelo, se divide en 3 fases:

- Fase 1: Diseño.

Corresponde a la fase de definición de niveles de menú, gráficas a mostrar, entre otros apartados visuales.

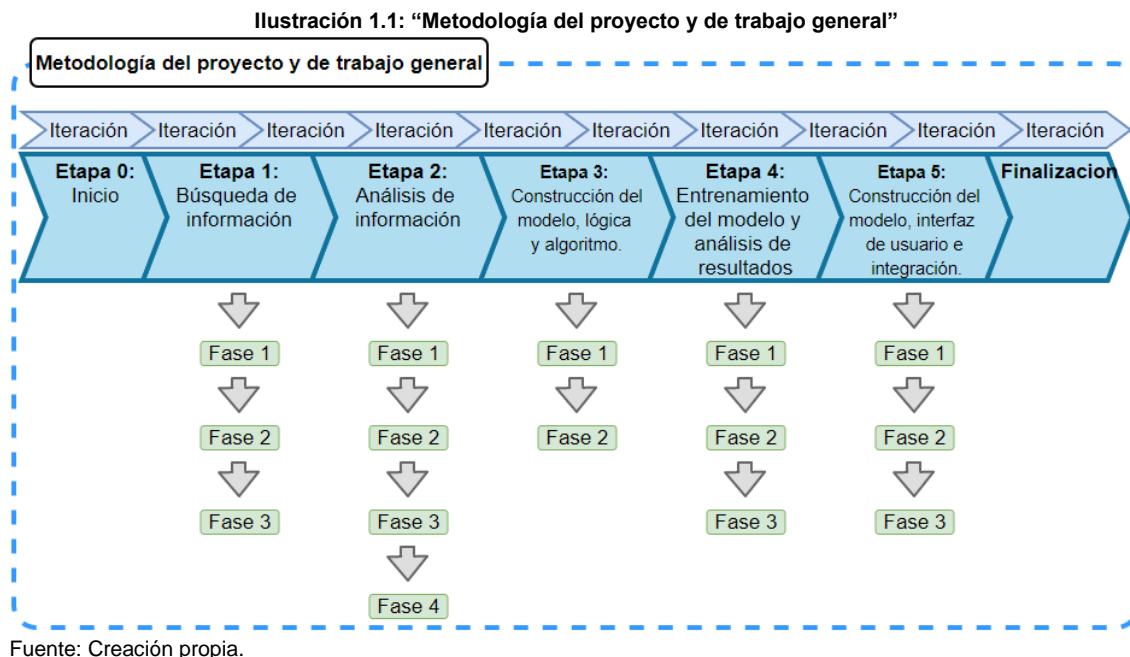
- Fase 2: Construcción visual.

Se construye la maqueta la app web, utilizando los diseños de la fase 1.

- Fase 3: Integración.

Se une la parte lógica con la visual del modelo y de ser requerido, se integra a servidores designados para este.

Adicionalmente, se debe recalcar que esta metodología sólo representa la estructuración del proyecto como tal, no se debe confundir con metodologías de trabajo, para esta última se utilizaron formatos ágiles en conjunto con el académico guía. A continuación, una Ilustración que representa gráficamente lo anteriormente mencionado:



La Ilustración 1.1 representa de una forma simple y resumida la información descrita en este capítulo, además, se debe mencionar que el número

de iteraciones (reuniones) en cada Etapa del proyecto varían según el plazo que se le asignó a cada una de estas.

Luego de entregar la información necesaria para conocer los aspectos generales del proyecto, se da el cierre a este capítulo para dar inicio al Capítulo N.^o 2, donde se explican al detalle las metodologías de estudio que se emplean para realizar los trabajos de investigación y análisis de información.

1.5. Recursos

Para la realización del proyecto, en primera instancia, para efectuar las investigaciones serán necesarios recursos bibliográficos tales como libros, revistas, informes, bases de datos, entre otros. Además, serán necesarios recursos humanos, de software y de hardware los cuales se describen a continuación:

- Recursos humanos: En este apartado se considera al alumno que desarrolla el proyecto, al apoyo del académico que actúa como su guía y al equipo de la estación móvil UTEM que provee información.
- Hardware: Será necesario un ordenador con conexión a internet con la capacidad para desarrollar el proyecto y la investigación previa.
- Software: Sistema operativo (Linux/Windows 10), Publish or Perish v7, Talend Open Studio, MySQL Workbench, Tableau, y softwares ofimáticos como Microsoft Word y Excel. Para el caso de softwares de Desarrollo, se utiliza: framework Django (python), framework Angular (typescript, css, html), Google Colab (python), Visual Studio code, navegador web.

CAPÍTULO 2 - MARCO REFERENCIAL

2.1. Estudio bibliométrico y altmetrics

En base a los resultados de la búsqueda de información (ver Anexo 2) y todo el proceso que se comprende en la limpieza, carga y utilización de dicha información a partir de las bases de datos creadas en MySQL Workbench (ver Anexo 3), se ofrece a continuación, un análisis de los datos obtenidos presentando diversos gráficos que permiten relevar la información y definir un grupo o listado de autores y publicaciones importantes para el proyecto. Las métricas o criterios utilizados para el análisis son los siguientes [3]:

- GSrank (Result Ranking): Este es simplemente el orden en el que la fuente de datos devolvió los resultados (1 = primero, 2 = segundo, etc.). Normalmente, las entradas clasificadas anteriormente indican resultados de búsqueda más relevantes.
- Ecc (Estimated citation Count): Algunas fuentes de datos proporcionan por separado el recuento de citas estimado, por ejemplo, Microsoft Academic. Si no está disponible, este campo se establece en el mismo valor que el campo Cites.
- Cites: Número de citas que tiene la publicación.
- AuthorCount: Número de autores que tiene la publicación, utilizado para el cálculo de citas por autor.
- CitesPerAuthor: Número de citas que se le otorga a cada autor de una publicación, se calcula con la división de Cites por AuthorCount y se presenta como el resultado entero redondeado.
- AuthorProductivity: Representa el número de publicaciones en las cuales participa o realiza un autor en el periodo de tiempo estudiado, define su aporte científico en el ámbito en estudio.
- Search Language: Es el lenguaje en el cual se realiza la búsqueda, no así sus resultados, ya que perfectamente se pueden encontrar documentos

en español si se realizan búsquedas en inglés y viceversa (u otros idiomas). Su utilidad es la de otorgar visualización a documentos en español, debido a que la mayoría de la información presente se encuentra en inglés, en otras palabras, permite igualar las cosas para ambos idiomas en lo que a resultados bibliográficos se refiere.

- Relevancia Por Obra (CitesPerAuthor/AuthorProductivity): Entrega un valor que representa la relevancia que tienen las obras de cada autor dentro del área en estudio, se considera la medida más importante dentro del relevamiento de autores.

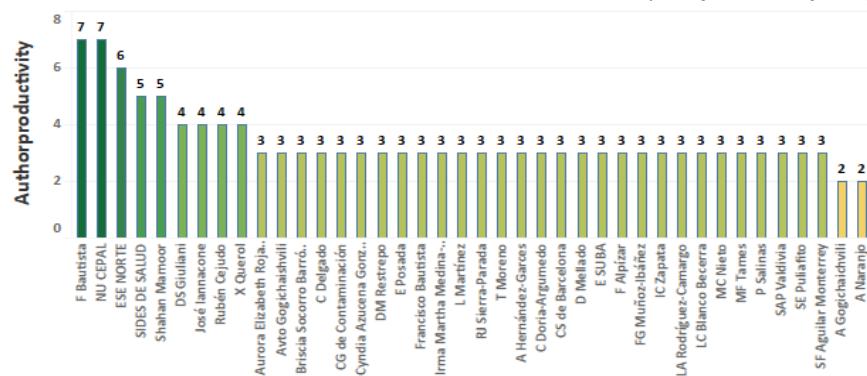
2.1.1. Análisis de resultados Búsqueda 1

La búsqueda 1, como se menciona anteriormente en este informe, pretende establecer la información necesaria para la contextualización del proyecto en el ámbito de la contaminación ambiental. A continuación, los análisis que permiten establecer en qué autores y publicaciones se basará principalmente el proyecto para la obtención de información respecto a la contaminación ambiental.

2.1.1.1. Análisis de autores

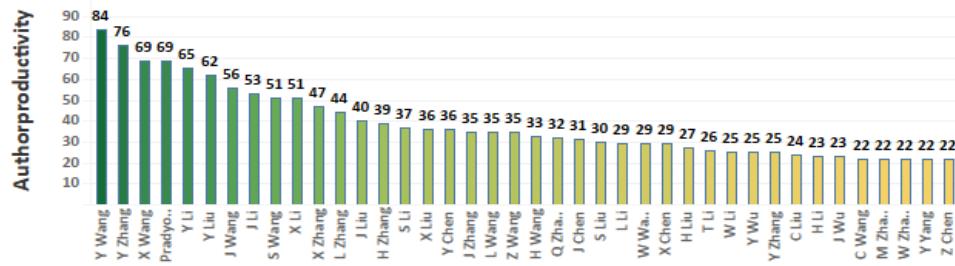
Se presentan a continuación, los gráficos que muestran los 40 autores con la productividad científica más alta (40 por idioma de búsqueda) dentro de los resultados de la búsqueda 1:

Gráfico 2.1: “Autores vs Productividad Científica últimos 5 años (Búsqueda 1, español)”



Fuente: Creación propia, utilizando el software Tableau.

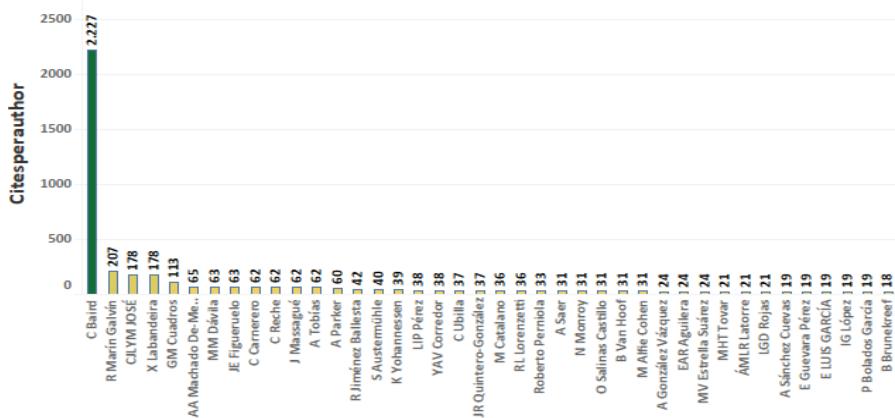
Gráfico 2.2: "Autores vs Productividad Científica últimos 5 años (Búsqueda 1, inglés)"



Fuente: Creación propia, utilizando el software Tableau.

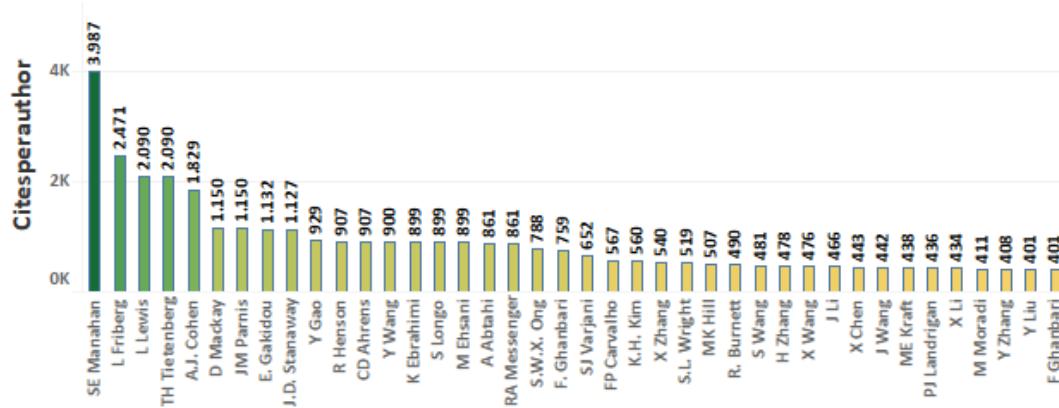
Observando la información presente en los Gráficos 2.1 y 2.2 se distingue una diferencia muy grande, en cuanto a producción científica se refiere, entre los autores obtenidos de la búsqueda en español e inglés, siendo los números de la segunda, los más elevados. A su vez, se puede ver como los autores líderes en producción científica son en su mayoría de origen chino, hecho que refleja la preocupación de la comunidad científica de ese país por la contaminación ambiental presente en sus territorios y ciudades. Los autores presentes en los Gráficos 2.1 y 2.2, serán considerados más adelante para la creación del listado de autores más relevantes de la búsqueda 1. Se presentan a continuación, los gráficos que muestran los 40 autores con la cantidad de citas más altas dentro de los resultados de la búsqueda 1:

Gráfico 2.3: "Autores vs Cantidad de Citas últimos 5 años (Búsqueda 1, español)"



Fuente: Creación propia, utilizando el software Tableau.

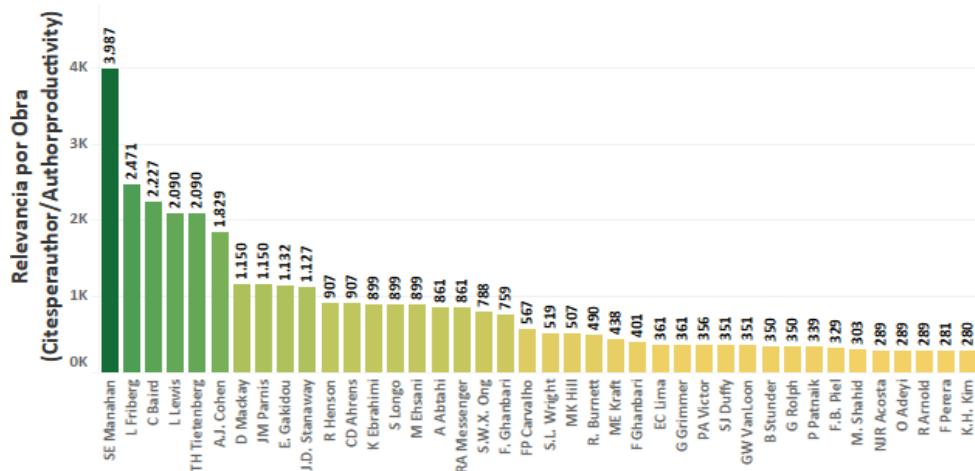
Gráfico 2.4: “Autores vs Cantidad de Citas últimos 5 años (Búsqueda 1, inglés)”



Fuente: Creación propia, utilizando el software Tableau.

En el caso del Gráfico 2.3, se puede observar una situación particularmente curiosa en sus resultados, esta se trata de la diferencia de citaciones entre el primer autor más citado y el segundo. Colin Baird es un autor especializado en el área de la química y la obra donde obtiene esa cantidad de citas se titula “Química Ambiental”, el cual es un trabajo referente en el área que ha obtenido múltiples publicaciones a lo largo de los años, de ahí su gran número de referencias. En el caso del Gráfico 2.4 de los resultados en inglés, se puede ver cómo, nuevamente, el volumen de resultados de las columnas es más alto que el de español y se presentan autores con más grado de relevancia en el área en cuanto a citas se refiere. El listado de autores presente en los Gráficos 2.3 y 2.4 se considerarán para la generación de la lista de autores más relevantes de la búsqueda 1. Finalmente, y como análisis definitivo, se debe considerar la relevancia por obra que tiene cada autor, la cual tiene como objetivo obtener el listado de autores más importantes del ámbito. La relevancia por obra se obtiene dividiendo la cantidad de citas por autor por su cantidad de publicaciones. A continuación, el gráfico que muestra el listado de los 40 autores más relevantes del área:

Gráfico 2.5: “Autores vs Relevancia por Obra Últimos 5 años (Búsqueda 1, español e inglés)”



Fuente: Creación propia, utilizando el software Tableau.

El Gráfico 2.5 tiene la finalidad de establecer una relación entre las citas y la cantidad de obras que se publican por autor, tras esta relación, se pueden eliminar casos como autores que publican demasiado, pero sus obras no han tenido un gran impacto, y a su vez, dar importancia casos como autores que publican poco, pero tienen una gran relevancia.

Juntando la información de todos los gráficos presentes en este punto, se preparará el listado de los autores más importantes de la búsqueda 1.

2.1.1.2. Resultado de análisis de autores

En base a los gráficos del punto 2.1.1.1 “Análisis de Autores”, se presenta a continuación la estructura del listado de autores más relevantes, el cual estará compuesto por 40 autores:

- Se considerarán los primeros 20 autores del Gráfico 2.5.
- Se considerarán los primeros 5 autores de los Gráficos 2.1, 2.2, 2.3 y 2.4.
- En caso de que existan autores que se repitan, se reemplazarán por autores del Gráfico 2.5, a partir del puesto 21.

Aclarado lo anterior, se presenta el listado de los autores más relevantes en el ámbito de la contaminación ambiental que se encontraron con la búsqueda 1:

Tabla 2.1: “Autores más relevantes de la Búsqueda 1”

Búsqueda 1							
N.º	Nombre	N.º	Nombre	N.º	Nombre	N.º	Nombre
1	SE Manahan	11	R Henson	21	R Marín Galván	31	Y Zhang
2	L Friberg	12	CD Ahrens	22	CJLYM JOSÉ	32	X Wang
3	C Baird	13	K Ebrahimi	23	X Labandeira	33	Pradyot Patnaik
4	L Lewis	14	S Longo	24	GM Cuadros	34	Y Li
5	TH Tietenberg	15	M Ehsani	25	F Bautista	35	S.L. Wright
6	A.J. Cohen	16	A Abtahi	26	NU CEPAL	36	MK Hill
7	D Mackay	17	RA Messenger	27	ESE NORTE	37	R. Burnett
8	JM Parnis	18	S.W.X. Ong	28	SIDES DE SALUD	38	ME Kraft
9	E. Gakidou	19	F. Ghanbari	29	Shahan Mamoor	39	F Ghanbari
10	J.D. Stanaway	20	FP Carvalho	30	Y Wang	40	EC Lima

Fuente: Creación propia.

Esta información será de utilidad para la búsqueda de obras alternativas, ya que, en caso de no encontrar información útil en las obras más relevantes, se dará prioridad en buscar apoyo en las publicaciones realizadas por los autores listados en la Tabla 2.1.

2.1.1.3. Análisis de publicaciones y resultados

Para realizar el análisis de las publicaciones más relevantes, se utilizan 3 métricas principales las cuales se operan en conjunto para generar un valor denominado como “Total” (ver Anexo 4), y es en base a este valor que se ordenan los resultados de las búsquedas, además, se separa el análisis según el idioma de estas. Se entregan como resultado final las 20 publicaciones más relevantes por cada idioma, dando un total de 40 publicaciones más relevantes. Adicionalmente, se debe mencionar que ciertas publicaciones debieron ser descartadas ya que su enfoque era otro o tenían una relación muy baja con lo que se quiere abordar en el presente estudio (ver Anexo 5). Aclarado todo lo anterior, se muestran a continuación, los resultados de las publicaciones más relevantes de la búsqueda 1:

Gráfico 2.6: “Publicaciones más relevantes de la búsqueda 1 en inglés”



Fuente: Creación propia, utilizando el software Tableau.

Gráfico 2.7: “Publicaciones más relevantes de la búsqueda 1 en español”



Fuente: Creación propia, utilizando el software Tableau.

Las publicaciones presentes en los Gráficos 2.6 y 2.7 serán revisadas para la construcción de las bases teóricas que conforman al proyecto desde un punto de vista contextualizador en aspectos como la contaminación ambiental y del aire. Si la necesidad de información que se requiere para realizar la tarea anterior no es satisfecha por las publicaciones relevadas, se procederá a buscar otras obras en esta tabla que no estén consideradas dentro de las 20 más relevantes por idioma y, adicionalmente, se puede recurrir a otras publicaciones de los autores más relevantes presentados anteriormente.

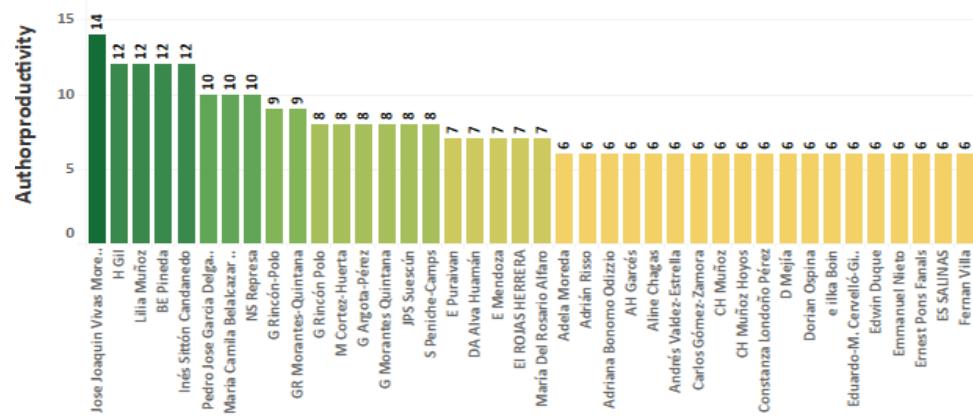
2.1.2. Análisis de resultados Búsqueda 2

La búsqueda 2, pretende otorgar la información necesaria para la creación del modelo predictivo, entrando en temas técnicos de machine learning y de desarrollo, a continuación, los análisis que permiten establecer en qué autores y títulos se basará principalmente el proyecto para la obtención de dicha información.

2.1.2.1. Análisis de autores

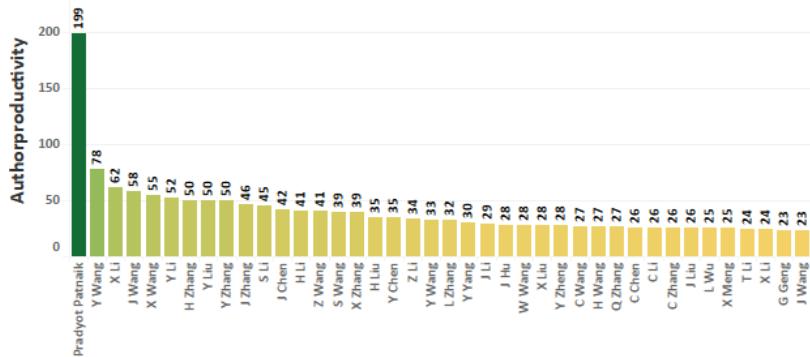
Se presentan a continuación, los gráficos que muestran los 40 autores con la productividad científica más alta dentro de los resultados de la búsqueda 2:

Gráfico 2.8: "Autores vs Productividad Científica últimos 5 años (Búsqueda 2, español)"



Fuente: Creación propia, utilizando el software Tableau.

Gráfico 2.9: “Autores vs Productividad Científica últimos 5 años (Búsqueda 2, inglés)”

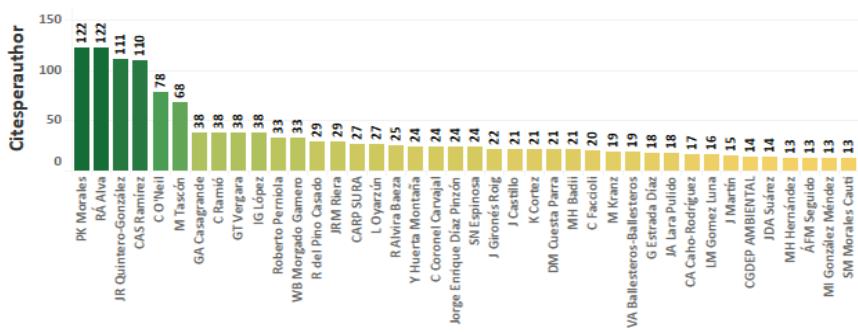


Fuente: Creación propia, utilizando el software Tableau.

Observando la información presente en los Gráficos 2.8 y 2.9, nuevamente se presenta una diferencia importante entre la productividad de los autores de las búsquedas de inglés y español. En el caso de los autores presentes en el Gráfico 2.8 (búsqueda en español), se ve como su nivel de productividad es más homogéneo el cual disminuye gradual y lentamente después del valor más elevado, en cambio, los autores que se encuentran en el Gráfico 2.9 (búsqueda en inglés), presentan un abrupto cambio en el nivel de productividad del primer al segundo puesto. El autor con más productividad científica es Pradyot Patnaik, el cual cuenta con 11 obras en 136 publicaciones en 3 idiomas y 3.860 catálogos de biblioteca, además, el enfoque de sus trabajos está basado principalmente en la creación de manuales sobre la contaminación ambiental y métodos de análisis de ésta, de ahí su elevado número de productividad científica [4].

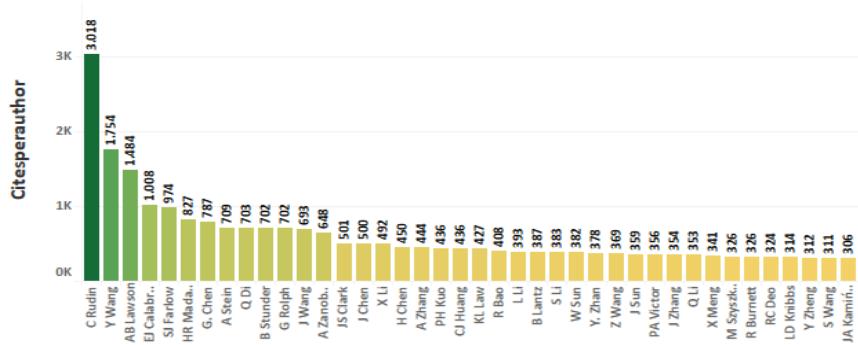
Los autores presentes en los Gráficos 2.8 y 2.9, serán considerados más adelante para la creación del listado de autores más relevantes de la búsqueda 2. Se presentan a continuación, los gráficos que muestran los 40 autores con la cantidad de citas más altas dentro de los resultados de la búsqueda 2:

Gráfico 2.10: “Autores vs Cantidad de Citas últimos 5 años (Búsqueda 2, español)”



Fuente: Creación propia, utilizando el software Tableau.

Gráfico 2.11: “Autores vs Cantidad de Citas últimos 5 años (Búsqueda 2, inglés)”

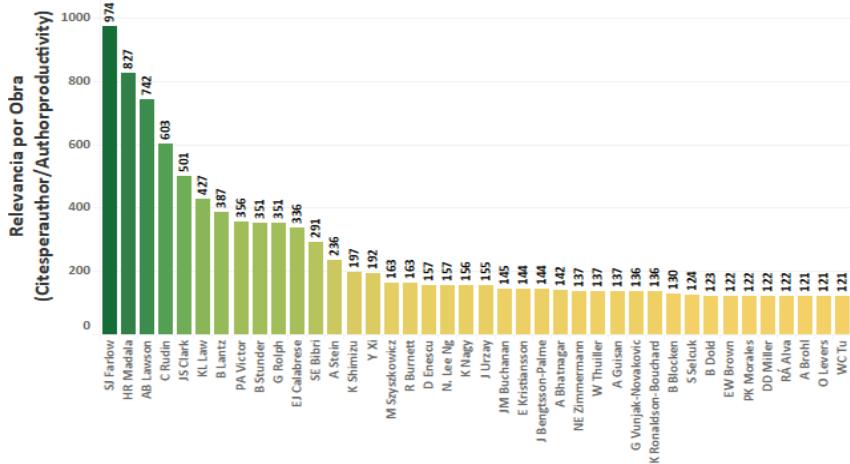


Fuente: Creación propia, utilizando el software Tableau.

En el caso del Gráfico 2.10, se puede observar una situación particularmente conveniente para el presente estudio, la cual tiene que ver con el nivel de citaciones de los 6 primeros autores presentes en la gráfica, dicho nivel de citaciones no solo es el más alto si no que se diferencian considerablemente del resto de autores, siendo el autor en el sexto puesto, citado casi el doble de veces que el del séptimo puesto. En el Gráfico 2.11, existe un nivel más elevado de citaciones en general, sobresaliendo una autora en particular. Cynthia Rudin es directora del Laboratorio de Análisis de Predicciones de la Duke University de Durham, Carolina del Norte Estados Unidos; es una científica informática y se especializa en aprendizaje automático por lo que su trabajo en esta área será un gran referente para el presente proyecto [5]. El listado de autores presente en los Gráficos 2.10 y 2.11 se considerarán para la generación de la lista de autores más relevantes de la búsqueda 2.

Finalmente, y como análisis definitivo, se debe considerar la relevancia por obra que tiene cada autor, la cual tiene como objetivo obtener listado de autores más importantes del ámbito. La relevancia por obra se obtiene dividiendo la cantidad de citas por autor por su cantidad de publicaciones. A continuación, el gráfico que muestra el listado de los 40 autores más relevantes del área:

Gráfico 2.12: “Autores vs Relevancia por Obra Últimos 5 años (Búsqueda 2, español e inglés)”



Fuente: Creación propia, utilizando el software Tableau.

El Gráfico 2.12 tiene la finalidad de establecer una relación entre las citas y la cantidad de obras que se publican por autor, tras esta relación, se pueden eliminar casos como autores que publican demasiado, pero sus obras no han tenido un gran impacto, y a su vez, dar importancia casos como autores que publican poco, pero tienen una gran relevancia.

Juntando la información de todos los gráficos presentes en este punto, se preparará el listado de los autores más importantes de la búsqueda 2.

2.1.2.2. Resultado de análisis de autores

En base a los gráficos del punto 2.1.2.1 “Análisis de Autores”, se presenta a continuación la estructura del listado de autores más relevantes, el cual estará compuesto por 40 autores:

- Se considerarán los primeros 20 autores del Gráfico 2.12.
- Se considerarán los primeros 5 autores de los Gráficos 2.8, 2.9, 2.10 y 2.11.
- En caso de que existan autores que se repitan, se reemplazarán por autores del Gráfico 2.12, a partir del puesto 21.

Aclarado lo anterior, se presenta el listado de los autores más relevantes en el ámbito de modelos predictivos y machine learning que se encontraron con la búsqueda 2:

Tabla 2.2: “Autores más relevantes de la Búsqueda 2”

Búsqueda 2							
N.º	Nombre	N.º	Nombre	N.º	Nombre	N.º	Nombre
1	SJ Farlow	11	EJ Calabrese	21	PK Morales	31	Y Wang
2	HR Madala	12	SE Bibri	22	RÁ Alva	32	Pradyot Patnaik
3	AB Lawson	13	A Stein	23	JR Quintero-González	33	X Li
4	C Rudin	14	K Shimizu	24	CAS Ramírez	34	J Wang
5	JS Clark	15	Y Xi	25	C O'Neil	35	X Wang
6	KL Law	16	M Szyszkowicz	26	Jose Joaquin Vivas Moreno	36	J Urzay
7	B Lantz	17	R Burnett	27	H Gil	37	JM Buchanan
8	PA Victor	18	D Enescu	28	Lilia Muñoz	38	E Kristiansson
9	B Stunder	19	N. Lee Ng	29	BE Pineda	39	J Bengtsson-Palme
10	G Rolph	20	K Nagy	30	Inés Sittón Candanedo	40	A Bhatnagar

Fuente: Creación propia.

Esta información será de utilidad para la búsqueda de obras alternativas, ya que, en caso de no encontrar información útil en las obras más relevantes, se dará prioridad en buscar apoyo en las publicaciones realizadas por los autores listados en la Tabla 2.2.

2.1.2.3. Análisis de publicaciones y resultados

Para realizar el análisis de las publicaciones más relevantes, se utilizan 3 métricas principales las cuales se operan en conjunto para generar un valor denominado como “Total” (ver Anexo 4), y es en base a este valor que se ordenan los resultados de la búsqueda, además, se separa el análisis en español con el del inglés. Se entregan como resultado final las 20 publicaciones más relevantes por cada idioma, dando un total de 40 publicaciones más relevantes.

Adicionalmente, se debe aclarar que ciertas publicaciones debieron ser descartadas ya que su enfoque era otro o tenían una relación muy baja con lo que se quiere abordar en el presente estudio (ver Anexo 5). Aclarado todo lo anterior, se muestran a continuación, los resultados de las publicaciones más relevantes de la búsqueda 2:

Gráfico 2.13: “Publicaciones más relevantes de la búsqueda 2 en inglés”



Fuente: Creación propia, utilizando el software Tableau.

Gráfico 2.14: “Publicaciones más relevantes de la búsqueda 2 en español”



Fuente: Creación propia, utilizando el software Tableau.

Las publicaciones presentes en los Gráficos 2.13 y 2.14 serán revisadas para la construcción de las bases teóricas que conforman al proyecto desde un punto de vista técnico. Si la necesidad de información que se requiere para realizar la tarea anterior no es satisfecha por las publicaciones relevadas, se procederá a buscar otras obras que no estén consideradas dentro de las 20 más relevantes por idioma y, adicionalmente, se puede recurrir a otras publicaciones de los autores más relevantes presentados anteriormente.

2.2. Marco teórico conceptual

2.2.1. Contaminación ambiental

El autor Suarez, afirma que:

La contaminación ambiental es un fenómeno que afecta sobre todo a las áreas urbanas y rurales de nuestro país, y cuyas consecuencias a la salud de la población aún no se encuentran bien identificadas, pero son inmediatas y de afectación a largo plazo. Ante esta situación, resulta alarmante que la sociedad no cuente con información sobre cómo protegerse tanto en su vida diaria, como cuando ocurren fenómenos físicos (pág. 9) [6].

El Instituto de Salud Pública (ISP), afirma que:

Se denomina contaminación ambiental a la presencia en el ambiente de cualquier agente (físico, químico o biológico), o bien de una combinación de varios agentes en lugares, formas y concentraciones tales que sean o puedan ser nocivos para la salud, la seguridad o para el bienestar de la población, o bien, que puedan ser perjudiciales para la vida vegetal o animal, o impidan el uso normal de las propiedades y lugares de recreación y goce de los mismos (pág. 1) [7].

La contaminación también se considera como el ingreso de sustancias químicas nocivas en un entorno al cual no pertenecen, dicho entorno y su equilibrio natural son afectados por el ingreso de estas sustancias y son alterados negativamente. Se entiende por entorno, a cualquier ecosistema, medio físico o ser vivo que esté en el área de observación del fenómeno [8].

A priori y en aspectos generales, la contaminación ambiental se puede dividir en dos categorías principales de causalidad:

- Natural: Causada por la misma naturaleza debido a las emisiones de gas y polvo de incendios forestales, erupciones volcánicas, terremotos y tsunamis.
- Artificial: Causada en su totalidad por el ser humano y se asocia al consumo irresponsable de recursos naturales, mala gestión de residuos, actividad industrial, emisiones de químicos, entre otros.

2.2.2. Contaminación del agua

La autora Zarza, afirma que:

La contaminación hídrica es la presencia de componentes químicos o de otra naturaleza en una densidad superior a la situación natural, de modo que no reúna las condiciones para el uso que se le hubiera destinado en su estado natural [9].

La Organización Mundial de la Salud (OMS), afirma que:

El agua contaminada es aquella que sufre cambios en su composición hasta quedar inservible. Es decir, es agua tóxica que no se puede ni beber ni destinar a actividades esenciales como la agricultura, además de una fuente de insalubridad que provoca más de 500.000 muertes anuales a nivel global por diarrea y transmite enfermedades como el cólera, la disentería, la fiebre tifoidea y la poliomielitis (pág. 1) [10].

Consiste en la alteración de la composición natural del agua, volviéndola inutilizable para el consumo humano y animal. Los orígenes de estas alteraciones pueden ser de [9]:

- Origen doméstico: Agua residual proveniente de entornos urbanos la cual contiene alimentos, deyecciones, basura, productos de limpieza, jabones, shampoo, etc.
- Origen agrícola-ganadero: Resultado de riego y de actividades ganaderas, la cual aporta mucho material orgánico de desechos animales.
- Origen industrial: Aguas residuales que son provocadas como resultante de procesos industriales en maquinaria pesada; puede ser agua que se ha utilizado para limpieza de materiales, enfriamiento, enjuague, etc.
- Origen pluvial: La lluvia, al caer, puede arrastrar diversos gases y partículas nocivas para el medio ambiente y la salud del aire, además, también puede desplazar químicos que ya se encuentren en el suelo.
- Origen Fluvial: Principalmente provocada por navíos que vierten sustancias químicas en el agua, de forma accidental o no, la más conocida de ellas es el vertido de petróleo.

Las consecuencias de la contaminación del agua pueden ser nefastas para el medio ambiente ya que afecta tanto a mares, ríos, lagos, lluvia y aguas subterráneas, lo que compromete por completo el sistema del ciclo del agua. Uno de los contaminantes más presentes en todas las aguas del mundo son los plásticos o fibras de micro plásticos. A continuación, diversas afirmaciones al respecto:

La Organización GreenPeace, afirma que:

Se estima que entre 4,8 y 12,7 millones de toneladas de plástico llegan a los océanos cada año (equivalente al peso de 800 torres Eiffel, suficientes para cubrir 34 veces la isla de Manhattan o el peso de 14.285 aviones

Airbus A380), y el Mediterráneo es uno de los mares más contaminados del mundo (pág. 1) [11].

National Geographic, afirma que:

Cada segundo más de 200 kilogramos de basura van a parar a los océanos, y ya hay cinco islas de basura formadas en su gran mayoría por micro plásticos, algo similar a una «sopa»: dos en el Pacífico, dos en el Atlántico, y una en el Índico. Y casi todo, un 80 por ciento, procede de los continentes terrestres. Se estima que para el año 2020 se superarán los 500 millones de toneladas anuales, un 900 por ciento más que en la década de los 80 [11].

En nuestros océanos hay billones de estos micro plásticos flotando que tienen impactos incluso en las especies más pequeñas que son la base de la red trófica marina. Los micro plásticos (fragmentos de plástico inferiores a 5 mm) pueden ser ingeridos por la fauna marina, la cual a su vez es explotada por el ser humano para consumirla como alimento lo que hace que sea posible la aparición de micro plásticos en comidas de proveniencia marina o de cualquier cuerpo celeste. Las implicaciones en la salud de estas sustancias plásticas se encuentran en estudio, ya que la Organización de las Naciones Unidas (ONU) [12], afirma que: “Se desconocen las implicaciones para la salud humana dado que existen muchas lagunas de conocimiento y por lo tanto se requiere más investigación en este aspecto” (pág. 1). Es necesario recordar que la contaminación del agua también se produce por otras sustancias químicas y no solo por el plástico, estas pueden ser los pesticidas, herbicidas, purines ganaderos, textiles, metales pesados, materiales radioactivos, productos para el hogar en general, entre otros.

2.2.2. Contaminación del suelo

La autora Días, afirma que:

Es la incorporación al suelo de materias extrañas, como basura, desechos tóxicos, productos químicos, y desechos industriales. La contaminación del suelo produce un desequilibrio físico, químico y biológico que afecta negativamente las plantas, animales y humanos (pág. 11) [6].

El autor Gligo, afirma que:

Se entiende por contaminación de un suelo a la presencia de un químico tóxico en el suelo, el cual en altas concentraciones tiene un efecto negativo sobre la salud humana o el ecosistema. La introducción de un contaminante puede estar dada por la actividad humana o ser de origen natural. A grandes rasgos, los contaminantes del suelo se pueden dividir en contaminantes de origen orgánico y aquellos inorgánicos (pág. 286) [13].

La contaminación de los suelos está relacionada con diversos impactos negativos sobre la provisión de servicios ecosistémicos además de problemas ambientales como la pérdida de biodiversidad, el cambio climático y la contaminación. Los servicios ecosistémicos pueden describirse a través de la siguiente cita:

El autor Estévez, afirma que:

Si concretamos que la biodiversidad es la variedad de especies, dentro de cada especie y entre los ecosistemas, entenderemos fácilmente que un servicio ecosistémico se puede definir como los beneficios que obtienen personas y empresas a partir de los ecosistemas. (pág. 1) [14].

Estévez también clasifica los servicios ecosistémicos en 4 categorías principales:

- Servicios de aprovisionamiento: madera, medicamentos, agua dulce, alimentos, fibras.
- Servicios de apoyo: existencia de hábitats, ciclo de nutrientes, dispersión de semillas.
- Servicios de regulación: depuración del agua, polinización, control de plagas.
- Servicios culturales: recreación, turismo, inspiración espiritual.

Si todos estos servicios se ven perjudicados debido a la contaminación de los suelos, se puede afirmar que dicha contaminación afecta a los humanos, su cultura, su desarrollo y sobre todas las cosas, su salud. De igual manera, esto incluye a los animales, los cuales forman parte del ecosistema y necesitan de hábitats para seguir existiendo. Las causas de la contaminación del suelo, se pueden categorizar de la siguiente forma:

- Erosión: Desgaste o destrucción producidos en la superficie por la fricción continua o violenta, la cual puede provenir de una fuerza hídrica y/o eólica. El Sistema de Información Ambiental de Colombia (SIAC), afirma que: “La erosión es un proceso natural; sin embargo, esta se califica como degradación cuando se presentan actividades antrópicas no sostenibles que aceleran, intensifican y magnifican el proceso” (pág. 1) [15]. En otras palabras, existen actividades humanas que aceleran este fenómeno natural, lo que lo convierte, en esos casos, en un daño artificial al medio ambiente.
- Degradación: Puede ser física, química y biológica:

El Sistema Nacional de Información Ambiental y Recursos Naturales (SNIARN), afirma que:

“La degradación química del suelo por polución se debe a la presencia, la concentración y el efecto biológico adverso de algunas sustancias. Éstas pueden provenir de tiraderos a cielo abierto, derrames, residuos

industriales, deposición de compuestos acidificantes y/o metales pesados" (pág. 1) [\[16\]](#).

La degradación física del suelo, no es más que la erosión causada por el ser humano en los terrenos, y finalmente, la degradación biológica del suelo se presenta cuando la materia orgánica se ve drásticamente reducida en la superficie o terreno determinado.

2.2.3. Contaminación del aire

MedlinePlus, afirma que:

La contaminación del aire es una mezcla de partículas sólidas y gases en el aire. Las emisiones de los automóviles, los compuestos químicos de las fábricas, el polvo, el polen y las esporas de moho pueden estar suspendidas como partículas. El ozono, un gas, es un componente fundamental de la contaminación del aire en las ciudades. Cuando el ozono forma la contaminación del aire también se denomina smog (pág. 1) [\[17\]](#).

La Fundación Aquae, afirma que:

La contaminación atmosférica consiste en la presencia de materias o formas de energía en el aire que pueden suponer un riesgo, daño o molestia de diferente gravedad para los seres vivos. Entre las consecuencias directas de la contaminación atmosférica, se podría destacar el desarrollo de enfermedades y afecciones en los seres humanos y la biodiversidad. También la pérdida de visibilidad en zonas de grandes concentraciones o la aparición de olores desagradables(pág. 1) [\[18\]](#).

Se puede describir la procedencia de la contaminación del aire haciendo énfasis en 5 tipos de focos básicos producidos por el ser humano [\[18\]](#):

- Industrias: Existen muchos antecedentes de las últimas décadas tanto en Chile como alrededor del mundo, donde distintas fábricas han emitido contaminantes al aire sin ningún tipo de regulación, las cuales, descargan a la atmósfera sustancias contaminantes sin ningún control de la cantidad, densidad y composición química de estos gases. La causa principal de la contaminación industrial es la quema a gran escala de combustibles fósiles como el petróleo, el carbón y el gas.
- Transporte: Una de las principales fuentes de la contaminación del aire en zonas urbanas altamente pobladas es la quema de combustibles derivados del petróleo en vehículos y medios de transportes. Cerca de 25% de todas las emisiones de CO₂ (dióxido de carbono) relacionadas con la energía provienen del transporte.
- Agricultura: Existen dos formas principales por las cuales se contamina el aire en esta área, la primera de ellas es la quema de los residuos agrícolas, y la segunda, la emisión de metano y amoniaco que genera la ganadería. La actividad agrícola es responsable de la producción del 24% de todos los gases de efecto invernadero en el mundo.
- Quema de residuos: Alrededor del mundo y en diversas actividades, se generan residuos donde aproximadamente el 40% de ellos son quemados al aire libre, lo que genera emisiones a la atmósfera de dioxinas nocivas, furanos, metano y carbono negro.
- Hogares: La contaminación del aire de una procedencia hogareña puede afectar de dos formas, por un lado, es el aire que respiran las personas de forma directa generando enfermedades respiratorias a mediano y largo plazo, y, en segundo lugar, que esta contaminación repercuta en el aire del exterior, aumentando la contaminación del aire general. A día de hoy aún existen muchos hogares donde la quema de combustibles fósiles y madera es usual.

2.2.4. Atmósfera terrestre

La Universidad Católica de Chile (UC), en su sitio web afirma que:

La atmósfera es una capa gaseosa de aproximadamente 10.000 km de espesor que rodea la litosfera e hidrosfera. Está compuesta de gases y de partículas sólidas y líquidas en suspensión atraídas por la gravedad terrestre. En ella se producen todos los fenómenos climáticos y meteorológicos que afectan al planeta, regula la entrada y salida de energía de la tierra y es el principal medio de transferencia del calor (pág. 1) [19].

La atmósfera, a medida que se extiende a más altura de la superficie terrestre, pierde composición de materia hasta alcanzar el espacio interplanetario. La distancia exacta de la atmósfera del planeta no está definida, pero, se puede considerar que tiene un alcance de aproximadamente 10.000 Km de altura. La masa de la atmósfera es de aproximadamente ($5,3 \times 10^{18}$)Kg, mientras que la de la Tierra es de ($5,98 \times 10^{24}$)Kg aproximadamente [20].

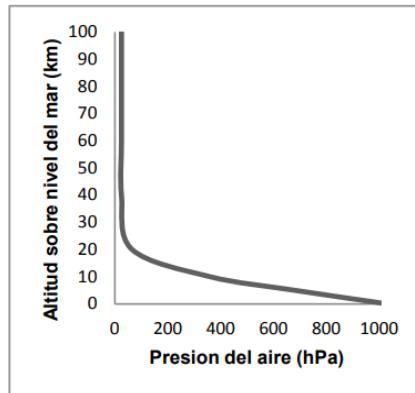
2.2.5. Presión, densidad y temperatura de la atmósfera

Las concentraciones químicas de la atmósfera están controladas por la densidad, presión y temperatura de esta, factores que, dependiendo de la altura, se modifican. La explicación de la importancia de estos factores para el presente estudio estará basada en el texto “Calidad del aire: Monitoreo y modelado de contaminantes atmosféricos. Efectos en la salud pública” el cual fue elaborado por coordinadores de la Facultad de Ciencias Exactas de la Universidad Nacional de la Plata, donde concentran el conocimiento adquirido por más de 10 años de estudios en el área [20].

2.2.5.1. Presión atmosférica

La presión atmosférica puede definirse como la fuerza por unidad de área que ejerce la atmósfera sobre la superficie en medición, esta fuerza en el sistema métrico decimal se mide en HectoPascal (hPa) que corresponde a una fuerza de 100 Newtons sobre un metro cuadrado de superficie. La variación de la presión atmosférica con la altura es mucho mayor que la variación horizontal, de modo que para comparar mediciones en lugares distintos se debe referir a un nivel común, el cual suele ser el nivel del mar, a continuación, un gráfico que representa lo antes mencionado:

Gráfico 2.15: “Presión del aire v/s Altitud sobre el nivel del mar”



Fuente: “Calidad del aire: Monitoreo y modelado de contaminantes atmosféricos. Efectos en la salud pública” [20].

La presión originada por la masa de una columna de aire se le conoce como presión hidrostática del aire, pero, si además de lo anterior, se le añade velocidad vertical al aire que la compone, se habla entonces de presión no-hidrostática. Para el caso de la presión hidrostática, se puede utilizar la siguiente fórmula para estimar su valor a cualquier altura [20]:

$$P_a(Z) = \int_Z^{\infty} \rho_a(Z) * g(Z) * dz$$

donde:

$P_a(Z)$ = Presión del aire en función de la altitud [pascal].

$\rho_a(Z)$ = Densidad del aire en función de la altitud [$\frac{Kg}{m^3}$].

$g(Z)$ = Aceleración gravitacional en función de la altura [$\frac{m}{s^2}$].

Además, la presión media de la atmósfera sobre la superficie terrestre puede ser aproximada con la siguiente fórmula [20]:

$$P_m = \frac{M_A g_0}{4\pi R_E^2}$$

donde:

M_A = Masa total de la atmósfera = ($5,3 \cdot 10^{18}$)[Kg]

g_0 = Aceleración media de la gravedad = 9,8 [$\frac{m}{s^2}$]

R_E^2 = Radio medio de la tierra ($6,37 \cdot 10^6$)[m]

Utilizando los datos anteriores, se puede aproximar el valor de la presión media atmosférica en superficie, el cual equivale aproximadamente a ($1 \cdot 10^5$) Pa.

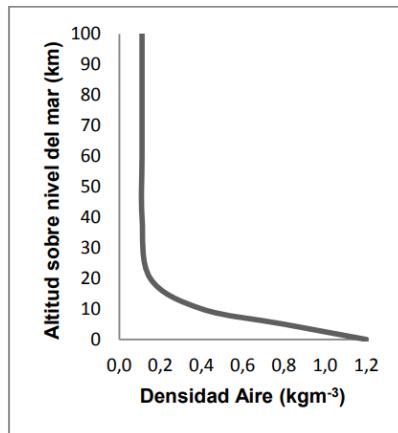
2.2.5.2. Densidad atmosférica

La fuerza de gravedad comprime las moléculas del aire, por lo que, en una unidad de volumen se encontrará un mayor número de dichas moléculas, en otras palabras, aumenta su densidad. Otra forma de explicar lo antes mencionado es que, a mayor cantidad de aire por encima del punto de medición, mayor es el agrupamiento o compresión de las moléculas, de esta forma, se puede expresar la densidad como la cantidad de masa de un cuerpo por unidad de volumen, con la siguiente fórmula [20]:

$$\rho = \frac{\text{masa}}{\text{volumen}}$$

Además, se debe considerar que la relevancia de la variabilidad vertical en la presión es mucho mayor que la horizontal, esto se debe a un crecimiento inverso y exponencial de este valor en base al de la altura, como se demuestra en el siguiente gráfico:

Gráfico 2.16: “Densidad del aire v/s Altitud sobre el nivel del mar”



Fuente: “Calidad del aire: Monitoreo y modelado de contaminantes atmosféricos. Efectos en la salud pública” [20].

2.2.5.3. Temperatura atmosférica

La temperatura atmosférica es el indicador de la cantidad de energía calorífica acumulada en el aire. La temperatura depende de diversos factores, por ejemplo, la inclinación de los rayos solares, del tipo de sustratos (la roca absorbe energía, el hielo la refleja), la dirección y fuerza de los vientos, la latitud, la altura sobre el nivel del mar, la proximidad de masas de agua, entre otros factores. La importancia de este valor es su significancia como la presencia de energía cinética en las moléculas del aire [20].

2.2.5.4. Ecuación de estado

La ecuación de estado describe la relación entre la presión, el volumen y la temperatura absoluta, las variables descriptivas más importantes de la atmósfera. En la ley de los gases ideales esta ecuación de estado se simplifica

para describir el comportamiento de un gas ideal, pero, al utilizarla sobre un gas real, se presentan ciertas diferencias en los cálculos las cuales pueden llevar a errores en los resultados, aun así, según los autores del libro “Calidad del aire”, dicho error es menor al 0,2% por lo que se le considera despreciable, dicho esto, se concluye que es viable aplicar la ecuación de estado para el caso actual, la cual está dada por [20]:

$$PV = nRT$$

donde:

P = Presión absoluta.

V = Volumen.

n = Número de moles del gas.

R = Constante universal de los gases (ver posibles valores en Anexo 8).

T = Temperatura absoluta.

Luego, a partir de la Ley de Dalton, que establece que la presión de una mezcla de gases que no reacciona químicamente, es igual a la suma de las presiones parciales que ejercería cada uno de ellos si sólo ocupase todo el volumen de la mezcla, sin variar la temperatura. Matemáticamente, la presión parcial de un gas se expresa como [20]:

$$P_q = N_q K_B T$$

donde:

N_q = Densidad del gas (número de moléculas por volumen).

K_B = Constante de Boltzmann (ver posibles valores en Anexo 8).

T = Temperatura absoluta.

De este modo, la ecuación de la presión atmosférica total se puede expresar de la siguiente manera:

$$P_a = \sum_q P_q = K_B T \sum_q N_q = N_a K_B T$$

donde:

N_a = Densidad o concentración del aire, determinado por la suma del mismo parámetro de cada gas considerado en la ecuación.

Además, la presión atmosférica total se puede escribir como:

$$P_a = P_d + P_v$$

donde:

$P_d = N_d K_B T \Rightarrow$ presión parcial ejercida por el aire seco (N_d densidad del aire seco).

$P_v = N_v K_B T \Rightarrow$ presión parcial ejercida por el vapor de agua (N_v densidad del vapor de agua).

El aire seco se compone de todos los gases atmosféricos excepto el del vapor de agua.

Finalmente, la densidad del aire se puede representar como:

$$N_a = \frac{A_v PR}{T}$$

y la densidad de un gas específico “g”:

$$N_g = \frac{A_v PR}{T} \chi_g$$

donde:

A_v = Constante de Avogadro (ver posibles valores en Anexo 8).

χ_g = Relación de mezcla de gas “g”.

2.2.6. Composición de la atmósfera

La composición gaseosa de la atmósfera se puede describir en términos de los componentes atmosféricos que han permanecido prácticamente constantes durante milenios y aquellos que han variado durante el curso de tiempo de la experiencia humana, de esta forma, se pueden clasificar estos gases como permanentes y variables respectivamente, además, es necesario agregar que la atmósfera está compuesta por partículas y una mezcla de gases, algunos altamente concentrados y otros más diluidos [21].

Hasta una altura aproximada de 80 km (ver Tabla 2.3), los gases de oxígeno y nitrógeno componen el 99% de la atmósfera, siendo el más presente de estos el nitrógeno con un 78% (oxígeno 21%) de la concentración. Los demás gases presentes en la atmósfera no superan el 1% de la concentración total, pero, son igual de importantes e influyentes en la atmósfera [21].

Tabla 2.3: "Componentes permanentes presentes en la atmósfera"

Componente	Siglas	Concentración (ppm)	Concentración (%)
Nitrógeno	N_2	780.840,00	78,082636942
Oxígeno	O_2	209.460,00	20,945634360
Argón	Ar	9.340,00	0,933983696
Neón	Ne	18,18	0,001817968
Helio	He	5,24	0,000523991
Criptón	Kr	1,14	0,000113998
Hidrógeno	H_2	0,50	0,000049999
Xenón	Xe	0,09	0,000009000

Fuente: Creación propia en base a la información del título "La Atmósfera: Un Sistema del Planeta Tierra" [21].

Tabla 2.4: "Componentes variables presentes en la atmósfera"

Componente	Siglas	Concentración (ppm)	Concentración (%)
Vapor de agua	H_2O	0,1 - 30.000,000000	0,000010000 - 2,999947631
Dióxido de carbono	CO_2	350,000000	0,0349999389
Metano	CH_4	1,670000	0,0001669997
Óxido nitroso	N_2O	0,300000	0,0000299999
Monóxido de carbono	CO	0,190000	0,0000190000
Ozono	O_3	0,040000	0,0000040000
Amoniaco	NH_3	0,004000	0,0000004000
Dióxido de nitrógeno	NO_2	0,001000	0,000000100
Dióxido de azufre	SO_2	0,001000	0,000000100
Oxido nítrico	NO	0,000500	0,000000050
Sulfuro de hidrógeno	H_2S	0,000050	0,000000005
Partículas	PM2,5 - PM10	0,00006	0,000000006

Fuente: Creación propia en base a la información del título "La Atmósfera: Un Sistema del Planeta Tierra" [21].

En la Tabla 2.4, se muestran los componentes variables más relevantes presentes en la atmósfera y, a diferencia de los componentes permanentes, aquí se incluyen partículas las cuales pueden ser de polvo, metales pesados, materia orgánica, etc. A excepción del vapor de agua, todos los componentes variables tienen una concentración menor al 1% por lo que su medición porcentual puede ser inapropiada y dado este caso, se suelen registrar en unidades de ppm (Partes por millón). Siguiendo con el análisis de la Tabla 2.4, se desprende que los gases variables más abundantes en la atmósfera son el vapor de agua y el dióxido de carbono. El rango de porcentaje en volumen del vapor de agua se debe a la variabilidad de su medida de acuerdo a la zona geográfica; en este contexto, en regiones tropicales, el vapor de agua puede constituirse hasta el 3% de los gases atmosféricos, mientras que en regiones polares representa bastante menos, al 1%. El dióxido de carbono por su parte, si bien está presente en la atmósfera en cantidades muy pequeñas, su rol es muy importante para la regulación de la temperatura en el planeta.

2.2.7. Elementos y compuestos

La información plasmada en este punto está basada en el sitio web de LennTech, una empresa dedicada a ofrecer soluciones sostenibles para el tratamiento de agua y la separación de líquidos con aplicaciones industriales [22]. Además, esta información fue contrastada con la expuesta en el texto “Química Ambiental” escrito por Colin Baird, el cual se encontró gracias al estudio bibliométrico [23].

2.2.7.1. Nitrógeno

El nitrógeno molecular es el principal constituyente de la atmósfera (78% por volumen de aire seco). Esta concentración es resultado del balance entre la fijación del nitrógeno atmosférico por acción bacteriana, eléctrica (relámpagos) y química (industrial) y su liberación a través de la descomposición de materias

orgánicas por bacterias o por combustión. En agua y suelos el Nitrógeno puede ser encontrado en forma de nitratos y nitritos [22].

Efectos en la salud del nitrógeno:

- Reacciones con la hemoglobina en la sangre, causando una disminución en la capacidad de transporte de oxígeno por la sangre (nitrito).
- Disminución del funcionamiento de la glándula tiroidea (nitrato).
- Bajo almacenamiento de la vitamina A (nitrato).
- Producción de nitrosaminas, las cuales son conocidas como una de la más común causa de cáncer (nitratos y nitritos).

Propiedades físicas del nitrógeno:

- Masa atómica (g/mol): 14,0067
- Densidad (kg/m³): 0,81
- Punto de ebullición (°C): -195,79
- Punto de fusión (°C): -218,8

2.2.7.2. Oxígeno

Es un elemento esencial en los procesos de respiración de la mayor parte de las células vivas y en los procesos de combustión. Es el elemento más abundante en la corteza terrestre. Cerca de una sexta parte (en volumen) del aire es oxígeno [22].

Efectos en la salud del oxígeno:

- Si bien, el oxígeno es un elemento que el humano respira naturalmente, si se expone a grandes cantidades de oxígeno durante mucho tiempo, se pueden producir daños en los pulmones, así, respirar un 50-100% de oxígeno a presión normal durante un periodo prolongado puede provocar daños en los pulmones.

Propiedades físicas del oxígeno:

- Masa atómica (g/mol): 15,9994
- Densidad (kg/m³): 1.429
- Punto de ebullición (°C): -183
- Punto de fusión (°C): -218,8

2.2.7.3. Argón

Es un gas noble que solo se puede encontrar en la atmósfera de la Tierra; sin embargo, se encuentran trazas de este gas en minerales y meteoritos. El argón es incoloro, inodoro e insípido. En condiciones normales es un gas, pero puede licuarse y solidificarse con facilidad [\[22\]](#).

Efectos en la salud del argón:

- Este gas es inerte y está clasificado como un asfixiante simple. La inhalación de éste en concentraciones excesivas puede resultar en mareos, náuseas, vómitos, pérdida de conciencia y muerte.

Propiedades físicas del argón:

- Masa atómica (g/mol): 39,948
- Densidad (kg/m³): 1,40
- Punto de ebullición (°C): -185,8
- Punto de fusión (°C): -189,4

2.2.7.4. Neón

Es un gas noble que solo se puede encontrar en la atmósfera de la Tierra, aunque se encuentran pequeñas cantidades de neón en el gas natural, en los minerales y en los meteoritos [\[22\]](#).

Efectos en la salud del neón:

- Este gas es inerte y está clasificado como un asfixiante simple. La inhalación de éste en concentraciones excesivas puede resultar en mareos, náuseas, vómitos, pérdida de conciencia y muerte.

Propiedades físicas del neón:

- Masa atómica (g/mol): 20,179
- Densidad (kg/m³): 1,20
- Punto de ebullición (°C): -246
- Punto de fusión (°C): -248,6

2.2.7.5. Helio

Es un gas noble incoloro, inodoro e insípido. El helio terrestre se forma por decaimiento radiactivo natural de elementos más pesados. La mayor parte de este helio migra a la superficie y entra en la atmósfera. Cabría suponer que la concentración atmosférica del helio fuese superior. Sin embargo, su bajo peso molecular le permite escapar al espacio a una velocidad equivalente a la de su formación [22].

Efectos en la salud del helio:

- La sustancia puede ser absorbida por el cuerpo por inhalación provocando elevación de la voz, mareos, pesadez del cuerpo y dolor de cabeza.

Propiedades físicas del helio:

- Masa atómica (g/mol): 4,0026
- Densidad (kg/m³): 0,126
- Punto de ebullición (°C): -268,9
- Punto de fusión (°C): -269,7

2.2.7.6. Criptón

El kriptón es uno de los gases nobles. Es un gas incoloro, inodoro e insípido. Su principal aplicación es el llenado de lámparas eléctricas y aparatos electrónicos de varios tipos. Se utilizan ampliamente mezclas de kriptón-argón para llenar lámparas fluorescentes. El kriptón es un gas raro atmosférico y como tal no es tóxico y es químicamente inerte [22].

Efectos en la salud del criptón:

- Este gas es inerte y está clasificado como un asfixiante simple. La inhalación de éste en concentraciones excesivas puede resultar en mareos, náuseas, vómitos, pérdida de conciencia y muerte.

Propiedades físicas del criptón:

- Masa atómica (g/mol): 83,80
- Densidad (kg/m³): 2,6
- Punto de ebullición (°C): -152
- Punto de fusión (°C): -157,3

2.2.7.7. Hidrógeno

El hidrógeno es la sustancia más inflamable de todas las que se conocen. Aunque por lo general es diatómico (que está formado por dos átomos), el hidrógeno molecular se disocia a temperaturas elevadas en átomos libres. El hidrógeno atómico es un agente reductor poderoso, aun a temperatura ambiente [22].

Efectos en la salud del hidrógeno:

- Este gas es inerte y está clasificado como un asfixiante simple. La inhalación de éste en concentraciones excesivas puede resultar en mareos, náuseas, vómitos, pérdida de conciencia y muerte.

Propiedades físicas del hidrógeno:

- Masa atómica (g/mol): 1,00797
- Densidad (kg/m³): 0,071
- Punto de ebullición (°C): -252,7
- Punto de fusión (°C): -259,2

2.2.7.8. Xenón

El xenón es incoloro, inodoro e insípido; es un gas en condiciones normales. El xenón es el único de los gases nobles no radiactivos que forma compuestos químicos estables a la temperatura ambiente y que, además, es un gas atmosférico no tóxico y es químicamente inerte [22].

Efectos en la salud del xenón:

- Este gas es inerte y está clasificado como un asfixiante simple. La inhalación de éste en concentraciones excesivas puede resultar en mareos, náuseas, vómitos, pérdida de conciencia y muerte.

Propiedades físicas del xenón:

- Masa atómica (g/mol): 131,30
- Densidad (kg/m³): 3,06
- Punto de ebullición (°C): -108,0
- Punto de fusión (°C): -111,9

2.2.7.9. Vapor de agua

El vapor de agua es el componente atmosférico cuya concentración varía más, del 0.1 a 30,000ppm. Al igual que el dióxido de carbono, es también un importante gas invernadero, absorbiendo energía infrarroja e irradiando de regreso al espacio. El vapor del agua es también significativo en la atmósfera porque cambia fácilmente sus fases.

2.2.7.10. Dióxido de carbono

La cantidad del dióxido de carbono en la atmósfera es relativamente baja, comprendiendo sólo cerca de 0.035% o 350 ppm. A pesar de este hecho, este componente es de enorme importancia dado que comprende una de las dos formas principales de materia prima esencial para el proceso de fotosíntesis. El CO₂ es la fuente del carbón, elemento indispensable para la vida. El dióxido de carbono es también un importante gas invernadero y debido a su absorbancia térmica juega un papel significativo en mantener un balance favorable del calor global [22].

Efectos en la salud del dióxido de carbono:

- Si bien el dióxido de carbono no es tóxico ni tan siquiera nocivo para la salud humana, tampoco es útil para la respiración, de manera que altas concentraciones en el aire interior de este gas producen una sensación poco confortable debido a que desplaza el oxígeno del aire y hace que la respiración se vuelva más fatigosa.

Propiedades físicas del dióxido de carbono:

- Masa molecular (g/mol): 44,01
- Densidad (kg/m³): 1,5
- Punto de ebullición (°C): -78

- Punto de fusión (°C): -57

2.2.7.11. Metano

Es una sustancia incolora y no polar, que se presenta en forma de gas a temperaturas y presiones ordinarias, y se caracteriza por su baja solubilidad en fase líquida y elevada persistencia en la atmósfera. Respecto a su incidencia sobre el medio ambiente, se trata del segundo compuesto que más contribuye al calentamiento global de la tierra (efecto invernadero) con un 15 %, sólo superado por el dióxido de carbono con un 76%. Es importante señalar que se trata de una sustancia extremadamente inflamable y el contacto con el aire resulta explosivo, llegando a producir incendios si existen focos de calentamiento [22].

Efectos en la salud del metano:

- Se trata de una sustancia que se puede absorber por inhalación, y al hacerlo, puede originar asfixia por la disminución del contenido de oxígeno en el aire, conllevando una pérdida de conocimiento del individuo e incluso de su muerte.

Propiedades físicas del metano:

- Masa molecular (g/mol): 16,04
- Densidad (kg/m³): 0,6
- Punto de ebullición (°C): -161
- Punto de fusión (°C): -183

2.2.7.12. Óxido Nitroso

El óxido nitroso es un gas volátil, incoloro, con un olor dulce y ligeramente tóxico. Como fuentes principales de emisión de óxido nitroso cabe destacar: los procesos llevados a cabo en agricultura intensiva, quema de biomasa y

combustibles fósiles, uso de fertilizantes nitrogenados ,deforestación. Otras fuentes de emisión se encuentran en procesos biológicos de suelos y océanos (ciclo del nitrógeno), en la desnitrificación del estiércol en los suelos, y en fenómenos tormentosos y emisiones volcánicas. Con respecto a su incidencia sobre el medio ambiente, es un importante gas de efecto invernadero con una permanencia media de 100 años en la atmósfera. Actualmente se le atribuye el 5% del efecto invernadero artificial, además de atacar la capa de ozono, reduciéndolo a oxígeno molecular y liberando dos moléculas de monóxido de nitrógeno (NO) [22].

Efectos en la salud del óxido nitroso:

- Su mecanismo de acción consiste en llegar al cerebro a través de las vías respiratorias y disminuir la actividad normal de las neuronas. Dependiendo de su concentración y exposición, puede generar analgesia, excitación, anestesia quirúrgica (que se manifiesta por pérdida de la conciencia y amnesia) o depresión total del sistema respiratorio (que, sin apoyo artificial, provoca un estado de coma y la muerte).

Propiedades físicas del óxido nitroso:

- Masa molecular (g/mol): 44,013
- Densidad (kg/m³): 1,27
- Punto de ebullición (°C): -89,5
- Punto de fusión (°C): -90,86

2.2.7.13. Monóxido de carbono

El monóxido de carbono es un gas inodoro, incoloro, insípido, tóxico y muy inflamable, aunque no es irritante, por lo que su exposición puede pasar completamente desapercibida. Es menos pesado que el aire, por lo que se acumula en las zonas altas de la atmósfera. La principal fuente de emisión del monóxido de carbono se produce en el sector transporte debido a la combustión

incompleta de gas, petróleo, gasolina, carbón y aceites. Los aparatos domésticos que queman combustibles fósiles como las estufas, hornillos o calentadores, también son una fuente de emisión común [22].

Efectos en la salud del monóxido de carbono:

- La inhalación de este elemento, en pequeñas concentraciones, puede dar lugar a confusión mental, vértigo, dolor de cabeza, náuseas, debilidad y pérdida del conocimiento. Si se produce una exposición prolongada o continua, pueden verse afectados el sistema nervioso y el sistema cardiovascular, dando lugar a alteraciones neurológicas y cardíacas.

Propiedades físicas del monóxido de carbono:

- Masa molecular (g/mol): 28,0
- Densidad (kg/m³): 1,14
- Punto de ebullición (°C): -191
- Punto de fusión (°C): -205

2.2.7.14. Ozono

El ozono es un gas incoloro e inestable que se encuentra en la atmósfera. Puede ser bueno o malo, dependiendo de donde se encuentre. El ozono "bueno" se encuentra en la naturaleza aproximadamente de 16 a 48 kilómetros sobre la superficie terrestre y nos protege de los rayos ultravioleta del sol. En cambio, el ozono perjudicial, se encuentra al nivel del suelo y se forma cuando los contaminantes de los automóviles, las fábricas y otras fuentes reaccionan químicamente con la luz del sol. Es el componente principal del smog [22].

Efectos en la salud del ozono:

- El ozono es un potente oxidante que puede irritar las vías respiratorias causando tos, ardor, resuello, falta de aire; puede agravar el asma y otras dolencias pulmonares.

Propiedades físicas del ozono:

- Masa molecular (g/mol): 48,0
- Densidad (kg/m³): 2,14
- Punto de ebullición (°C): -111,9
- Punto de fusión (°C): -192,7

2.2.7.15. Amoníaco

Se trata de un gas incoloro, de olor muy penetrante, bastante soluble en agua, y en estado líquido es fácilmente evaporable. Debido a que el amoníaco se genera naturalmente en el ambiente, todos estamos expuestos rutinariamente a bajos niveles de amoníaco en el aire, el suelo y el agua. El amoníaco existe naturalmente en el aire en niveles entre 1 y 5 partes en un billón de partes de aire (ppb). Se encuentra comúnmente en el agua de lluvia [22].

Efectos en la salud del amoníaco:

- La exposición a altas concentraciones de amoniaco en el aire, puede producir quemaduras graves en la piel, ojos, garganta y pulmones, y en casos extremos puede provocar ceguera, daño en el pulmón (edema pulmonar) e incluso la muerte. A bajas concentraciones puede causar tos e irritación de nariz y garganta.

Propiedades físicas del amoníaco:

- Masa molecular (g/mol): 17,03
- Densidad (kg/m³): 0,73
- Punto de ebullición (°C): -33
- Punto de fusión (°C): -78

2.2.7.16. Dióxido de nitrógeno

El dióxido de nitrógeno es el principal contaminante de los óxidos de nitrógeno, y se forma como subproducto en todas las combustiones llevadas a cabo a altas temperaturas. Se trata de una sustancia de color amarillento, que se forma en los procesos de combustión en los vehículos motorizados y las plantas eléctricas. Es un gas tóxico, irritante y precursor de la formación de partículas de nitrato, que conllevan la producción de ácidos y elevados niveles de PM2,5 en el ambiente [22].

Efectos en la salud del dióxido de nitrógeno:

- La inhalación en elevadas concentraciones y durante un corto periodo de tiempo, puede originar un edema pulmonar cuyos efectos no se observan hasta pasadas unas horas, agravándose con el esfuerzo físico. Una exposición prolongada puede afectar al sistema inmune y al pulmón, dando lugar a una menor resistencia frente a infecciones y causar cambios irreversibles en el tejido pulmonar.

Propiedades físicas del dióxido de nitrógeno:

- Masa molecular (g/mol): 46,1
- Densidad (kg/m³): 30,01
- Punto de ebullición (°C): 21,2
- Punto de fusión (°C): -11,2

2.2.7.17. Dióxido de azufre

El dióxido de azufre es un gas incoloro y no inflamable, de olor fuerte e irritante. Su vida media en la atmósfera es corta de unos 2 a 4 días, y casi la mitad de las emisiones vuelven a depositarse en la superficie. En la naturaleza, el dióxido de azufre se encuentra sobre todo en las proximidades de los volcanes y las erupciones pueden liberar cantidades importantes de este gas. Los efectos

en el medio ambiente del dióxido de azufre empeoran cuando se combina con partículas o con la humedad del aire ya que se forma ácido sulfúrico, y produce lo que se conoce como lluvia ácida, provocando la destrucción de bosques, vida salvaje y la acidificación de las aguas superficiales [22].

Efectos en la salud del dióxido de azufre:

- Afecta sobre todo las mucosidades y los pulmones provocando ataques de tos, si bien éste es absorbido por el sistema nasal. La exposición de altas concentraciones durante cortos períodos de tiempo puede irritar el tracto respiratorio, causar bronquitis, reacciones asmáticas, espasmos reflejos, parada respiratoria y congestionar los conductos bronquiales de los asmáticos.

Propiedades físicas del dióxido de azufre:

- Masa molecular (g/mol): 64,06
- Densidad (kg/m³): 2,92
- Punto de ebullición (°C): -10
- Punto de fusión (°C): -75,5

2.2.7.18. Óxido nítrico

Es un gas incoloro, tóxico y soluble en agua que es principalmente producido por la combustión en vehículos e industrias, además, es un gas altamente inestable en el aire ya que se oxida rápidamente en presencia de oxígeno convirtiéndose en dióxido de nitrógeno. Su efecto con la radiación solar es doble y mientras existe en la baja atmósfera contribuye al calentamiento global y en el alta lo hace con el oscurecimiento global [22].

Efectos en la salud del óxido nítrico:

- Este gas puede irritar los ojos, la nariz y la garganta, causando tos o falta de aire, además, si se tiene una exposición prolongada puede causar sensación de desvanecimiento, mareo y somnolencia. A altos niveles puede causar desmayo y a niveles muy altos puede causar la muerte.

Propiedades físicas del óxido nítrico:

- Masa molecular (g/mol): 30,01
- Densidad (kg/m³): 1,27
- Punto de ebullición (°C): -152
- Punto de fusión (°C): -164

2.2.7.19. Sulfuro de hidrógeno

El sulfuro de hidrógeno es un gas incoloro, inflamable, muy peligroso y con un olor fuerte. Se puede oler incluso a niveles bajos, pero con la exposición prolongada el olfato pierde el rastro fácilmente, de manera que las personas pueden tener poca alerta de la presencia del gas en concentraciones dañinas. Es un gas que al contacto con la atmósfera se oxida rápidamente para luego transformarse en dióxido de azufre [22].

Efectos en la salud del sulfuro de hidrógeno:

- Este gas puede provocar irritación de la nariz, la garganta y el pulmón, con tos y asfixia (edema pulmonar), dolor de cabeza, náusea y vómitos. Además, las nieblas de ácidos inorgánicos fuertes que contienen ácido sulfúrico causan cáncer de laringe en seres humanos.

Propiedades físicas del sulfuro de hidrógeno:

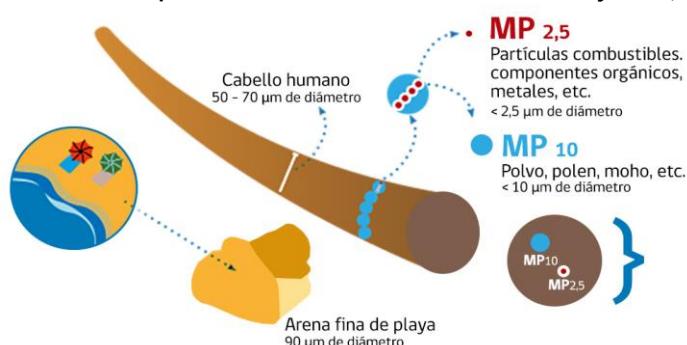
- Masa molecular (g/mol): 34,1

- Densidad (kg/m^3): 1,36
- Punto de ebullición ($^{\circ}\text{C}$): -60
- Punto de fusión ($^{\circ}\text{C}$): -86

2.2.7.20. Material Particulado (PM10/PM2,5)

El material particulado, también llamado contaminación por partículas, consiste en una mezcla de partículas sólidas y gotas líquidas que se encuentran en el aire. Algunas partículas, como el polvo, la suciedad, el hollín o el humo, son lo suficientemente grandes u oscuras para ser vistas a simple vista. Otros son tan pequeños que solo se pueden detectar con un microscopio electrónico. A continuación, una muestra visual del tamaño de estas partículas [23]:

Ilustración 2.1: "Comparación de tamaño de un cabello humano y PM10, PM2.5"



Fuente: Ministerio del medio ambiente, recuperada del sitio <http://airechile.mma.gob.cl/faq>

Estas partículas vienen en muchos tamaños y formas y pueden estar compuestas por cientos de sustancias químicas diferentes, algunas se emiten directamente desde una fuente, como sitios de construcción, caminos sin pavimentar, campos, chimeneas o incendios. La mayoría de las partículas se forman en la atmósfera como resultado de reacciones complejas de sustancias químicas como el dióxido de azufre y los óxidos de nitrógeno, que son contaminantes emitidos por centrales eléctricas, industrias y automóviles. Como se puede ver en la Ilustración 2.1, el material particulado se clasifica según su tamaño, estas clasificaciones pueden ser:

- PM10: Partículas que tienen un diámetro menor a 10 micrones (o micrómetros), y a pesar de ser aquellas con el mayor tamaño dentro de esta clasificación, pueden ingresar al sistema respiratorio de igual forma.
- PM2,5: Partículas que tienen un diámetro menor a 2,5 micrones. Por lo mismo, el MP2,5 se encuentra contenido en el MP10. Estas partículas, dependiendo de su composición, son especialmente peligrosas ya que pueden ingresar al sistema respiratorio fácilmente.

Efectos en la salud del material particulado:

- El material particulado, como se menciona anteriormente, contiene sólidos microscópicos o gotitas de líquido que son tan pequeñas que se pueden inhalar y causar graves problemas de salud. Algunas partículas de menos de 10 micrómetros de diámetro pueden penetrar profundamente en los pulmones y algunas incluso pueden llegar al torrente sanguíneo. De estas, las partículas de menos de 2,5 micrómetros de diámetro, también conocidas como partículas finas o PM2.5, representan el mayor riesgo para la salud. Los efectos de estas partículas en la salud pueden variar según su composición, pero en general pueden provocar: muerte prematura en personas con enfermedades cardíacas o pulmonares, infartos de miocardio no mortales, latidos irregulares, asma agravada, función pulmonar reducida, síntomas respiratorios aumentados, como irritación en las vías respiratorias, tos o dificultad para respirar [24].

2.2.8. Índices de contaminación ambiental

En el mundo, diversas agencias gubernamentales emplean índices de contaminación ambiental para establecer la calidad del aire. Un índice de contaminación ambiental se representa con un valor numérico el cual se calcula (generalmente) en base al nivel de presencia de uno o varios contaminantes en una unidad volumétrica, además, a esta medición, se le suele agregar el factor

del tiempo con el fin de establecer niveles de concentraciones constantes y no únicos (se evita la recolección de información de casos aislados).

Los índices de contaminación ambiental más usados a nivel mundial son los siguientes [20]:

- Air Quality Health Index (AQHI): Índice de salud de la calidad del aire; utilizado por Canadá, consiste en una escala del 1 al 10, en donde se clasifica la calidad del aire (10=la peor calidad).
- Air Pollution Index (AQHI): Índice de contaminación del aire; utilizado en Hong Kong, está basado en medidas de los contaminantes ozono, dióxido de nitrógeno, dióxido de azufre y material particulado (PM_{2,5}-PM₁₀). Para una hora determinada, el AQHI se calcula a partir de la suma del porcentaje de exceso de riesgo de ingresos hospitalarios diarios atribuibles a las concentraciones medias móviles de 3 horas de estos cuatro contaminantes (escala del 1 al 10, donde 10 es la más peligrosa).
- Air Quality Index and Individual Air Quality Index (AQI/IAQI): Índice de contaminación del aire e Índice individual de contaminación del aire; utilizado por China, consiste en la asignación de un puntaje individual (IAQI) a cada contaminante y el AQI final es el más alto de estos seis puntajes. Los seis puntajes corresponden a los niveles de SO₂, NO₂, CO, O₃, PM_{2.5} y PM₁₀. El valor final del AQI se puede calcular por hora o por 24 horas.
- National Air Quality Index (AQI): Índice nacional de calidad del aire; utilizado en India, consiste seis categorías de AQI, a saber, Bueno, Satisfactorio, Moderadamente Contaminado, Deficiente, Muy malo y Severo. El AQI propuesto considerará ocho contaminantes (PM₁₀, PM_{2.5}, NO₂, SO₂, CO, O₃, NH₃ y Pb) para los cuales se prescriben los Estándares Nacionales de Calidad del Aire Ambiental a corto plazo (hasta un período promedio de 24 horas).

- Daily Air Quality Index (DAQI): Índice diario de calidad del aire; utilizado ampliamente por el Reino Unido, consiste en diez puntos, que se agrupan a su vez en cuatro bandas: bajo, moderado, alto y muy alto. El índice se calcula a partir de las concentraciones de los siguientes contaminantes: ozono, dióxido de nitrógeno, dióxido de azufre, PM2,5 y PM10.
- Common Air Quality Index (CAQI): Índice común de calidad del aire; ampliamente utilizado en Europa, consiste en un número en una escala del 1 al 100, donde un valor bajo significa buena calidad del aire y un valor alto significa mala calidad del aire. El índice se define en versiones tanto por hora como por día, y por separado cerca de las carreteras (un índice de "carretera" o "tráfico") o lejos de las carreteras (un índice de "fondo"). A partir de 2012, el CAQI tenía dos componentes obligatorios para el índice de carretera, NO₂ y PM10, y tres componentes obligatorios para el índice de fondo, NO₂, PM10 y O₃. En ambos casos, los componentes opcionales son el PM2.5, CO y SO₂.

En Chile se utiliza (además de otros índices) el Índice de Calidad de Aire por Partículas Respirables (ICAP), el cual fue establecido en el Decreto Supremo 59, el 16 de marzo de 1998, por el Ministerio de Secretaría General de la Presidencia de la República, establece 3 niveles para clasificar las concentraciones de PM10 en el aire, los cuales se distribuyen en una puntuación numérica del 195 al 330 (ug/m³N) en tiempos de medición de 24 horas [25]. Hoy en día, el ICAP se utiliza como complemento del índice Air Quality Index (AQI) de la United States Environmental Protection Agency (EPA), que es el índice de contaminación ambiental más utilizado en América. El AQI de la EPA se basa en los cinco contaminantes de "criterio" regulados por la Ley de Aire Limpio: ozono a nivel del suelo, material particulado, monóxido de carbono, dióxido de azufre y dióxido de nitrógeno. También, se divide en seis categorías que indican niveles crecientes de preocupación por la salud. Un valor de AQI superior a 300

representa una calidad del aire peligrosa y por debajo de 50 la calidad del aire es buena. Ver la siguiente tabla:

Tabla 2.5: “Niveles de preocupación AQI EPA”

Color diario de AQI	Niveles de preocupación	Valores del índice	Descripción de la calidad del aire
Verde	Bueno	0 a 50	La calidad del aire es satisfactoria y la contaminación del aire presenta poco o ningún riesgo.
Amarillo	Moderado	51 a 100	La calidad del aire es aceptable. Sin embargo, puede haber un riesgo para algunas personas, particularmente aquellas que son inusualmente sensibles a la contaminación del aire.
Naranja	Insano para grupos sensibles	101 a 150	Los miembros de grupos sensibles pueden experimentar efectos sobre la salud. Es menos probable que el público en general se vea afectado.
Rojo	Insalubre	151 a 200	Algunos miembros del público en general pueden experimentar efectos sobre la salud; los miembros de grupos sensibles pueden experimentar efectos de salud más graves.
Púrpura	Muy insalubre	201 a 300	Alerta de salud: el riesgo de efectos en la salud aumenta para todos.
Granate	Peligroso	301 <	Advertencia de salud de condiciones de emergencia: todos tienen más probabilidades de verse afectados.

Fuente: Creación propia en base a la información del sitio: <https://www.airnow.gov/aqi/aqi-basics/>

Este índice de calidad del aire se calcula como una función lineal por partes de la concentración de contaminantes. En el límite entre las categorías de AQI, hay un salto discontinuo de una unidad de AQI. Para convertir de concentración a un índice dentro del AQI se utiliza la siguiente ecuación:

$$I = \frac{I_{high} - I_{low}}{C_{high} - C_{low}}(C - C_{low}) + I_{low}$$

donde:

I = Índice de calidad del aire (AQI).

C = Concentración de contaminantes.

C_{low} = Punto de corte de concentración inferior que es $\leq C$

C_{high} = Punto de corte de concentración superior que es $\geq C$

I_{low} = Punto de ruptura del índice correspondiente a C_{low}

I_{high} = Punto de ruptura del índice correspondiente a C_{high}

Para entender mejor qué son estos valores presentes en la fórmula anterior, se presenta a continuación la tabla de valores del índice:

Tabla 2.6: “Niveles estándar por contaminante e indicador del AQI EPA”

O3 (ppb)	O3 (ppb)	PM2.5 ($\mu\text{g}/\text{m}^3$)	PM10 ($\mu\text{g}/\text{m}^3$)	CO (ppm)	SO2 (ppb)	NO2 (ppb)	AQI	AQI
C_{low} - C_{high} (avg)	C_{low} - C_{high} (avg)	C_{low} - C_{high} (avg)	C_{low} - C_{high} (avg)	C_{low} - C_{high} (avg)	C_{low} - C_{high} (avg)	C_{low} - C_{high} (avg)	Category	I_{low} - I_{high}
0-54 (8-hr)	-	0.0-12.0 (24-hr)	0-54 (24-hr)	0.0-4.4 (8-hr)	0-35 (1-hr)	0-53 (1-hr)	Good	0-50
55-70 (8-hr)	-	12.1-35.4 (24-hr)	55-154 (24-hr)	4.5-9.4 (8-hr)	36-75 (1-hr)	54-100 (1-hr)	Moderate	51-100
71-85 (8-hr)	125-164 (1-hr)	35.5-55.4 (24-hr)	155-254 (24-hr)	9.5-12.4 (8-hr)	76-185 (1-hr)	101-360 (1-hr)	Unhealthy for Sensitive Groups	101-150
86-105 (8-hr)	165-204 (1-hr)	55.5-150.4 (24-hr)	255-354 (24-hr)	12.5-15.4 (8-hr)	186-304 (1-hr)	361-649 (1-hr)	Unhealthy	151-200
106-200 (8-hr)	205-404 (1-hr)	150.5-250.4 (24-hr)	355-424 (24-hr)	15.5-30.4 (8-hr)	305-604 (24-hr)	650-1249 (1-hr)	Very Unhealthy	201-300
-	405-504 (1-hr)	250.5-350.4 (24-hr)	425-504 (24-hr)	30.5-40.4 (8-hr)	605-804 (24-hr)	1250-1649 (1-hr)	Hazardous	301-400
-	505-604 (1-hr)	350.5-500.4 (24-hr)	505-604 (24-hr)	40.5-50.4 (8-hr)	805-1004 (24-hr)	1650-2049 (1-hr)		401-500

Fuente: Creación propia en base a la información del sitio:

<https://www.airnow.gov/publications/air-quality-index/technical-assistance-document-for-reporting-the-daily-aqi/>

En la Tabla 2.6 se muestran en cada columna y separados por un “-” los valores de C_{low} y C_{high} que representan los márgenes de tolerancia del valor C , así mismo, el valor de C es el promedio de las mediciones en el rango de horas estipulado en cada columna (según sea el elemento medido). Finalmente, los valores de I_{low} e I_{high} se utilizan en la fórmula según el valor de C .

2.2.9. Área de estudio

El sector o área donde se espera aplicar el modelo predictivo es la comprendida por las comunas que tienen información útil tanto para su análisis, como para su utilización en el entrenamiento del modelo predictivo. A continuación, un listado de las comunas:

Cerrillos, Cerro Navia, El Bosque, Independencia, la Florida, Las Condes, Santiago, Pudahuel, Puente Alto y Talagante.

2.2.10. Tecnologías y herramientas

A continuación, se definen las principales tecnologías o herramientas que se deben conocer para analizar su utilización en la construcción del modelo

predictivo de contaminación del aire, y, al término de este subcapítulo, se definirán las tecnologías y herramientas que se utilizarán.

2.2.10.1. Machine Learning

El aprendizaje automático o machine learning se refiere al proceso de descifrar el patrón subyacente de datos por computadoras automáticamente en lugar de diseñar cualquier regla hecha por el hombre, para así, entregar una respuesta acorde a dichos datos de forma autónoma. El Machine learning se puede clasificar en dos categorías: aprendizaje supervisado y aprendizaje no supervisado [26].

Aprendizaje supervisado:

Se define por el uso de conjuntos de datos etiquetados para entrenar algoritmos que clasifiquen datos o predigan resultados con precisión y a medida que los datos de entrada se introducen en el modelo, esté ajusta sus ponderaciones hasta lograr un resultado esperado. Para la realización del entrenamiento se utiliza un conjunto de datos que contiene entradas y salidas correctas para enseñar a los modelos a producir el resultado deseado. El aprendizaje supervisado se puede dividir en dos tipos de problemas [28]:

- Clasificación: Reconoce entidades específicas dentro del conjunto de datos e intenta sacar algunas conclusiones sobre cómo esas entidades deben etiquetarse o definirse. Los algoritmos de clasificación comunes son clasificadores lineales, máquinas de vectores de soporte (SVM), árboles de decisión, vecino k más cercano y bosque aleatorio.
- Regresión: Se utiliza para comprender la relación entre variables dependientes e independientes, para comúnmente, hacer proyecciones. La regresión lineal, la regresión logística y la regresión polinomial son algoritmos de regresión populares.

Algoritmos pertenecientes a la categoría del aprendizaje supervisado [29]:

- Redes neuronales: Procesan los datos de entrenamiento imitando la interconectividad del cerebro humano a través de capas de nodos. Cada nodo se compone de entradas, pesos, un sesgo (o umbral) y una salida. Si ese valor de salida excede un umbral dado, "dispara" o activa el nodo, pasando datos a la siguiente capa en la red.
- Naive Bayes: Se enfoca en la clasificación que adopta el principio de independencia condicional de clase del teorema de Bayes. Esto significa que la presencia de una característica no afecta la presencia de otra en la probabilidad de un resultado dado, y cada predictor tiene el mismo efecto en ese resultado. Esta técnica se utiliza principalmente en la clasificación de texto, la identificación de spam y los sistemas de recomendación.
- Regresión lineal: La regresión lineal se usa para identificar la relación entre una variable dependiente y una o más variables independientes y generalmente se aprovecha para hacer predicciones sobre resultados futuros. Cuando solo hay una variable independiente y una variable dependiente, se conoce como regresión lineal simple. A medida que aumenta el número de variables independientes, se denomina regresión lineal múltiple.
- Regresión logística: La regresión logística se selecciona cuando la variable dependiente es categórica, lo que significa que tienen salidas binarias, como "verdadero" y "falso" o "sí" y "no". La regresión logística se utiliza principalmente para resolver problemas de clasificación binaria, como la identificación de spam.
- Support vector machine (SVM): Una máquina de vectores de soporte es un modelo que se utiliza tanto para la clasificación como para la regresión de datos, donde generalmente se aprovecha para problemas de clasificación, construyendo un hiperplano donde la distancia entre dos clases de puntos de datos es máxima. Este hiperplano se conoce como el límite de decisión.

- K-nearest neighbor: El vecino K-más cercano, también conocido como algoritmo KNN, es un algoritmo no paramétrico que clasifica los puntos de datos en función de su proximidad y asociación con otros datos disponibles. KNN se utiliza normalmente para motores de recomendación y reconocimiento de imágenes.
- Random forest: El bosque aleatorio es otro algoritmo de aprendizaje automático supervisado flexible que se utiliza tanto para fines de clasificación como de regresión. El "bosque" hace referencia a una colección de árboles de decisión no correlacionados, que luego se fusionan para reducir la varianza y crear predicciones de datos más precisas.

Aprendizaje no supervisado:

El aprendizaje no supervisado utiliza algoritmos de aprendizaje automático para analizar y agrupar conjuntos de datos sin etiquetar. Estos algoritmos descubren patrones ocultos o agrupaciones de datos sin necesidad de intervención humana. Algunas de sus aplicaciones pueden ser [30]:

- Reconocimiento de objetos.
- Detección, clasificación y segmentación de imágenes
- Detección de anomalías
- Crear perfiles de clientes
- Motores de recomendación

Algoritmos pertenecientes a la categoría del aprendizaje no supervisado [31]:

- Association Rules: Método basado en reglas para encontrar relaciones entre variables en un conjunto de datos determinado. Estos métodos se utilizan con frecuencia para el análisis de la canasta de mercado.
- Apriori algorithms: Los algoritmos a priori se utilizan dentro de conjuntos de datos transaccionales para identificar conjuntos de artículos

frecuentes, o colecciones de artículos, para identificar la probabilidad de consumir un producto dado el consumo de otro producto. Los algoritmos a priori utilizan un árbol hash para contar conjuntos de elementos, navegando a través del conjunto de datos de una manera amplia.

- Dimensionality reduction: Reduce la cantidad de entradas de datos a un tamaño manejable y al mismo tiempo preserva la integridad del conjunto de datos tanto como sea posible. Este algoritmo se utiliza comúnmente en la etapa de preprocesamiento de datos, y existen algunos métodos diferentes de reducción de dimensionalidad que se pueden usar, como Principal component analysis y Singular value decomposition.
- Autoencoders: Los codificadores automáticos aprovechan las redes neuronales para comprimir datos y luego recrear una nueva representación de la entrada de datos originales. La etapa de la capa de entrada a la capa oculta se conoce como "codificación", mientras que la etapa de la capa oculta a la capa de salida se conoce como "decodificación".
- Clustering: La agrupación en clústeres es una técnica de minería de datos que agrupa datos sin etiquetar en función de sus similitudes o diferencias. Los algoritmos de agrupación en clústeres se pueden clasificar en unos pocos tipos, específicamente exclusivos, superpuestos, jerárquicos y probabilísticos.

2.2.10.2. Random Forest

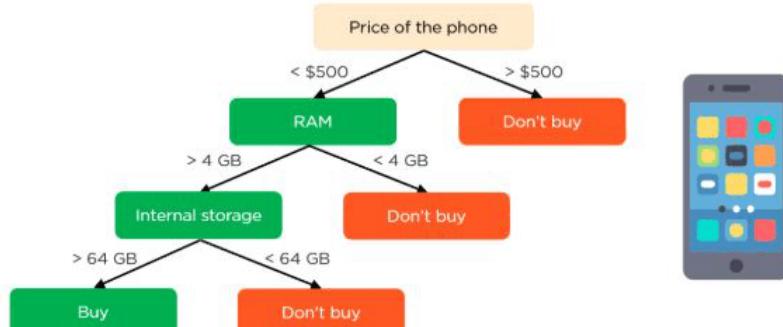
El bosque aleatorio (Random Forest) es un algoritmo de consenso que se utiliza en Machine Learning para resolver problemas de regresión y clasificación. Cada bosque aleatorio se compone de múltiples árboles de decisión que trabajan juntos como un conjunto para producir una predicción. Un árbol de decisión es una construcción lógica que se asemeja a un diagrama de flujo e ilustra una serie de sentencias if-else. Un propósito importante del uso de bosques aleatorios es compensar las limitaciones de los algoritmos de árboles

de decisión mediante el mapeo de varios árboles y el uso de la producción promedio del bosque (media estadística) [32].

Los algoritmos de Random forest pueden producir predicciones aceptables incluso si los árboles individuales del bosque tienen datos incompletos. Estadísticamente, aumentar el número de árboles en el conjunto aumentará correspondientemente la precisión del resultado. A continuación, algunas características claves de los árboles de decisión aleatorios [32]:

- Es más preciso que el algoritmo del árbol de decisiones.
- Proporciona una forma eficaz de gestionar los datos faltantes.
- Puede producir una predicción razonable sin ajuste de hiperparámetros.
- Resuelve el problema del sobreajuste en los árboles de decisión.
- En cada árbol forestal aleatorio, se selecciona aleatoriamente un subconjunto de características en el punto de división del nodo.

Ilustración 2.2: “Ejemplo de Árbol de decisión”



Fuente: IBM, recuperada del sitio <https://n9.cl/70v6x>

Como se muestra en la Ilustración 2.2, los árboles de decisión crean ramificaciones las cuales están conectadas mediante operaciones lógicas condicionales y estas terminan en nodos puros o donde se cumplan las condiciones indicadas por el algoritmo [32].

Hiperparámetros de un modelo Random Forest

Los hiperparámetros de un modelo son los valores de las configuraciones utilizadas durante el proceso de entrenamiento, y para el caso de Random forest utilizando la librería “`sklearn.ensemble.RandomForestClassifier`”, estos son los siguientes [37]:

- **n_estimators (int, default=100)**: El N.^o de árboles de decisión en el bosque.
- **Criterion ({"gini", "entropy"}, default="gini")**: Los criterios con los que dividir en cada nodo (Gini o Entropy para una tarea de clasificación, o el MSE o MAE para regresión).
- **max_depth (int, default=None)**: Cuanto más grande es un árbol individual, más posibilidades tiene de sobre ajustar los datos de entrenamiento, sin embargo, como en Random Forests se tienen muchos árboles individuales, esto no es un gran problema.
- **min_samples_split (int or float, default=2)**: El número mínimo de muestras necesarias para dividir un nodo interno.
- **min_samples_leaf (int or float, default=1)**: Muestras mínimas para dividir en un nodo interno de los árboles. Jugando con este parámetro y el anterior se puede regularizar los árboles individuales si fuera necesario.
- **min_weight_fraction_leaf (float, default=0.0)**: En Random Forest esto no es tan importante, pero en un árbol de decisión individual también puede ayudar en gran medida a reducir el sobreajuste y también ayudar a aumentar la comprensión del árbol al reducir el número posible de rutas a los nodos de hoja.
- **max_features ({"auto", "sqrt", "log2"}, int or float, default = "auto")**: La cantidad de características a considerar al buscar la mejor división.
- **max_leaf_nodes (int, default=None)**: Cultiva árboles desde la mejor manera primero. Los mejores nodos se definen como una reducción

relativa de la impureza. Si es None, entonces tendrá un número ilimitado de nodos hoja.

- **min_impurity_decrease (float, default = 0.0):** Un nodo se dividirá si esta división induce una disminución de la impureza mayor o igual a este valor.
- **Bootstrap (bool, default=True):** Si se utilizan muestras de bootstrap al construir árboles. Si es False, se usa todo el conjunto de datos para construir cada árbol.
- **oob_scorebool (default=False):** Si utilizar muestras fuera de la bolsa para estimar la puntuación de generalización. Solo disponible si bootstrap = True.
- **class_weight ({“balanced”, “balanced_subsample”}, dict or list of dicts, default=None):** Pesos asociados a clases en el formulario . Si no se da, se supone que todas las clases tienen un peso de valor 1. El modo "balanceado" utiliza los valores de "Y" para ajustar automáticamente los pesos de forma inversamente proporcional a las frecuencias de clase en los datos de entrada como n_samples / (n_classes * np.bincount(y)). El modo " balanced_subsample " es lo mismo que " balanced ", excepto que los pesos se calculan en función de la muestra de arranque para cada árbol cultivado.

Ventajas de Random Forest

- Los bosques aleatorios se consideran un método muy preciso y robusto debido a la cantidad de árboles de decisión que participan en el proceso.
- No sufre el problema de sobreajuste. La razón principal es que toma el promedio de todas las predicciones, lo que anula los sesgos.
- El algoritmo se puede utilizar tanto en problemas de clasificación como de regresión.
- Los bosques aleatorios también pueden manejar valores perdidos. Hay dos formas de manejarlos: usando valores medianos para reemplazar las

- variables continuas y calculando el promedio ponderado por proximidad de los valores perdidos.
- Puede obtener la importancia relativa de las características, lo que ayuda a seleccionar las características que más contribuyen al clasificador.

Desventajas de Random Forest

- Los bosques aleatorios generan predicciones con lentitud porque tienen varios árboles de decisión. Siempre que hace una predicción, todos los árboles del bosque tienen que hacer una predicción para la misma entrada dada y luego realizar una votación sobre ella. Todo este proceso requiere mucho tiempo.
- El modelo es difícil de interpretar en comparación con un árbol de decisiones, donde se puede tomar una decisión fácilmente siguiendo la ruta del árbol.

Para ver el detalle de la decisión para la utilización de Random Forest para la creación del modelo predictivo, consultar el Anexo 9.

2.2.10.3. TensorFlow

TensorFlow es una biblioteca de código abierto desarrollada por investigadores de Google para ejecutar Machine Learning, Deep Learning y otras cargas de trabajo de análisis estadístico y predictivo. Está diseñado para agilizar el proceso de desarrollo y ejecución de aplicaciones analíticas avanzadas para usuarios como científicos de datos, estadísticos y modeladores predictivos. Las aplicaciones de TensorFlow pueden ejecutarse en CPU convencionales o en unidades de procesamiento de gráficos (GPU) de mayor rendimiento, así como en las propias unidades de procesamiento de tensor (TPU) de Google, que son dispositivos personalizados diseñados expresamente para acelerar los trabajos de TensorFlow [33].

2.2.10.4. Keras

Keras es una API de aprendizaje profundo de alto nivel desarrollada por Google para implementar redes neuronales. Está escrito en Python y se utiliza para facilitar la implementación de redes neuronales. También es compatible con la computación de múltiples redes neuronales back-end.

Keras es relativamente fácil de aprender y trabajar con él porque proporciona una interfaz de Python con un alto nivel de abstracción y, al mismo tiempo, tiene la opción de múltiples aplicaciones de fondo para fines de cálculo. Esto hace que Keras sea más lento que otros marcos de aprendizaje profundo, pero extremadamente amigable para principiantes [34].

2.2.10.5. NumPy

NumPy, que significa Numerical Python, es una biblioteca que consta de objetos de matriz multidimensionales y una colección de rutinas para procesar esas matrices. Las operaciones que se pueden realizar con NumPy se dividen en tres categorías principales: Transformada de Fourier (transformar señales entre el dominio del tiempo y el dominio de la frecuencia) y manipulación de formas , operaciones matemáticas y lógicas , utilización de álgebra lineal y generación de números aleatorios . Para hacerlo lo más rápido posible, NumPy está escrito en C y Python [35].

2.2.10.6. Pandas

Pandas es una biblioteca de Python de código abierto que se usa más ampliamente para la ciencia de datos / análisis de datos y tareas de aprendizaje automático ya que simplifica la realización de muchas de las tareas repetitivas y que consumen mucho tiempo asociadas con el trabajo con datos. Está construido sobre Numpy , que proporciona soporte para matrices multidimensionales. Como uno de los paquetes de gestión de datos más

populares, Pandas funciona bien con muchos otros módulos de ciencia de datos dentro del ecosistema de Python, y generalmente se incluye en todas las distribuciones de Python, desde las que vienen con su sistema operativo hasta las distribuciones de proveedores comerciales como ActivePython de ActiveState [36].

2.2.10.7. Sklearn

Scikit-learn es una biblioteca en Python que proporciona muchos algoritmos de aprendizaje supervisados y no supervisados. Se basa en algunas de las tecnologías con las que quizás ya esté familiarizado, como NumPy, pandas y Matplotlib. La funcionalidad que proporciona scikit-learn incluye [38]:

- Regresión , incluida la regresión lineal y logística
- Clasificación , incluidos K-vecinos más cercanos
- Agrupación , incluidas K-medias y K-medias ++
- Selección de modelo
- Procesamiento previo , incluida la normalización mínima y máxima

2.2.11. Definición de la Solución

En base a toda la información expuesta en el Capítulo 2, Marco Teórico, se plantea a continuación, la solución a la problemática referente a la necesidad de predecir la presencia de contaminantes en el aire para las comunas pertenecientes a la AMUR.

En primer lugar, se analizan los hechos y la situación actual que presenta la problemática planteada: Originalmente, el proyecto se planteaba como un modelo predictivo de contaminación ambiental general, el cual debía enfocarse en la contaminación a grandes rasgos, pero ¿Qué información se tiene sobre la contaminación del agua, la tierra o el aire?, si nos enfocamos primero en el caso del agua, existen efectivamente formas de medir componentes químicos

presentes en esta y Chile cuenta con información al respecto, pero, ¿Cuál es la utilidad práctica de diseñar un modelo predictivo para tal tarea? ¿Qué acciones cambiarían con tales predicciones de contaminación? ¿Es más útil predecir el estado del agua u obtener esta información en tiempo real? En el año 2016, las empresas Barrick Chile y Pascua-Lama implementaron un sistema de monitoreo del agua para el río Huasco, el cual sigue en funcionamiento hasta el día de hoy y ha probado ser de utilidad para la gente de la localidad de El Valle [27]. Las personas miden el agua (incluso pueden ver sus datos en tiempo real en la web) y saben de inmediato su calidad, de esta forma pueden controlar su nivel de contaminación con seguridad. En el caso específico del agua, es mucho más relevante para las personas conocer su estado en tiempo real, ya que así pueden tomar decisiones rápidamente, como por ejemplo, recolectar agua para regadíos o para dar de beber a animales, en cambio, si se intenta predecir el estado del agua, las complicaciones aumentan demasiado, ya que las variables a considerar son casi en su totalidad relativas y los problemas a considerar para lograr un nivel de predicción aceptable son muy grandes, todas estas dificultades, al menos de momento, se ven superadas por simplemente controlar el estado del agua en tiempo real; además, si bien anteriormente se menciona que existe información para la generación de DataSets de un posible modelo predictivo, ¿Existe dicha información para todos los ríos de Santiago? ¿Qué ríos se deberán estudiar? ¿Qué hay de lagunas, canales y otros cuerpos de agua? Si se selecciona un río para el estudio, ¿A qué altura de este se deberá estimar sus niveles de contaminación? Recordemos que los componentes que pueda llegar a tener el agua no son iguales en todo el caudal del río, etc. Entre otras cuestiones que se deben considerar y que vuelven cada vez más inviable estudiar la contaminación del agua en el presente proyecto. En el caso de la contaminación de los suelos, ocurre algo similar, pero esta vez con la toma de muestras. El mejor ejemplo que se puede dar respecto a esta afirmación es la detección de la erosión del suelo, la cual no puede ser medida con sensores, solo registrada mediante la observación y la detección por medio de redes

neuronales analizando fotografías satelitales del terreno. En el caso de los contaminantes presentes en el suelo, estos generalmente se encuentran concentrados en lugares específicos, donde ciertas actividades humanas generan residuos, como, por ejemplo, la minería, y el hecho de que estos residuos se concentren solo en un lugar específico, al igual que en el caso del agua, vuelven mucho más viable su control en tiempo real en vez de intentar predecir sus cantidades. Finalmente, en el caso del aire, tratar de predecir su nivel de contaminación se vuelve más factible debido a que se cuenta con mucha información de DataSets al respecto, lo que vuelve factible la generación de un modelo, además, se debe considerar la utilidad práctica de un modelo predictivo de contaminación del aire, como por ejemplo, mejor control de la restricción vehicular, regulación de industrias altamente contaminantes con anticipación, prevención ciudadana ante aire altamente contaminado con anterioridad en base a las corrientes del aire, manejo de eventos al aire libre escogiendo una fecha que presente un ambiente más limpio, ayuda para personas con enfermedades respiratorias a la hora de planificar viajes o actividades al aire libre, entre otros. Además de todo lo anterior, Chile cuenta con información actualizada en tiempo real de los factores estudiados en el marco teórico que pueden llegar a influir en el comportamiento de la contaminación del aire, como niveles de presencia de PM_{2,5}-PM₁₀, SO₂, NO₂, NOx, NO, CO, O₃, CH₄, HCNM, información de corrientes de aire, temperatura, presión ambiental, entre otros que puedan ser de utilidad. Por los motivos anteriormente descritos, el modelo predictivo se enfocará solo en la contaminación del aire.

Con respecto a la funcionalidad del modelo, se plantea ofrecer un rango de predicción mínimo de 7 días y máximo de entre 10 a 30 días (depende de la viabilidad del modelo). Además, su implementación se realizará en un servidor a modo de aplicación web. Las características funcionales y visuales específicas del modelo se establecen en los próximos capítulos del presente trabajo.

2.2.11.1. Detalle de solución

En primera instancia se piensa construir el modelo utilizando el algoritmo de Random Forest. El modelo recibirá datos numéricos de entrada tales como fechas (desglosadas), indicadores de comuna (diccionario de clases), niveles de concentración de contaminantes en el aire, temperatura y el día a futuro a predecir. Estos datos de entrada están directamente ligados a la estructura del DataSet, la cual fue escogida en base a la información disponible de La Región Metropolitana y sus comunas, por lo que existen parámetros que no pudieron ser considerados, como elementos contaminantes que no son monitoreados.

Los datos para conformar el DataSet serán extraídos de los sitios web de la Dirección Meteorológica de Chile (Servicios Climáticos), el SINCA y Clima Chile. Los datos están compuestos por niveles de concentración de compuestos por metro cúbico, niveles de concentración de material particulado, presión atmosférica, temperatura, precipitaciones, humedad del aire, fecha de la medición e información de la estación de medición. La comparación de resultados se hará con el cálculo del indicador de contaminación ambiental explicado en el Capítulo N.^o 2, el cual corresponde al indicador AQI EPA.

Finalmente, la interfaz de utilización del modelo predictivo se realizará sobre una aplicación web, la cual consumirá un Web Service que contendrá el modelo predictivo de la contaminación del aire.

CAPÍTULO 3 - METODOLOGÍA DE INVESTIGACIÓN

Este capítulo se enfoca en las Etapas 1 y 2 del proyecto, las cuales corresponden a la búsqueda de información y análisis de esta, estas Etapas se dividen en Fases las cuales se encuentran detalladas en el Capítulo N.^º 1 del presente trabajo y resumidas visualmente en la Ilustración 1.1. Se describe a continuación, el trabajo que se realiza por cada una de las Etapas y Fases anteriormente mencionadas, y metodología de estudio del proyecto.

3.1. Enfoque de la investigación

Si bien el tipo de investigación que se realiza en este trabajo, según su objetivo final, es el de investigación Aplicada, se le debe dar un enfoque para regir claramente cómo se llevará a cabo el proceso, por tanto, se establece que la investigación realizada en el presente trabajo tiene un enfoque descriptivo, transversal, y documental. A continuación, se describen estos enfoques [2]:

- Descriptiva: Debido a que se busca información detallada sobre el fenómeno de estudio, sus características y configuración para tener una visión clara de su naturaleza.
- Transversal: La investigación se llevará a cabo en un grupo acotado de documentos e información, dicho acotamiento se basa en un intervalo de tiempo de 5 años hasta el presente, es decir, a partir del año 2017 al 2021, y debido a que la información de estudio es de un intervalo de tiempo dado, se le considera transversal.
- Documental: La principal fuente de información se obtiene a través de material existente y seleccionado para el estudio, tales como libros, documentos de archivo, hemerografía, registros audiovisuales, entre otros.

3.2. Alcances de la Investigación

En primera instancia, se pretende indagar en distintos aspectos de la contaminación ambiental, sus causas, evolución, efectos en las personas y en el entorno, puntos críticos, factores que contribuyen a su avance, puntos de no retorno y los indicadores utilizados para detectar y medir su intensidad, principalmente los del aire. Es importante mencionar que esta investigación no busca establecer conceptos nuevos ni proponer ajustes o cambios en la información ya existente, su propósito es solo entender las características y naturaleza del tema estudiado para aplicarlos en la elaboración del modelo predictivo. En segunda instancia, la investigación también explora trabajos ya realizados sobre modelos predictivos, principalmente los enfocados en la contaminación del aire, y su finalidad es guiar al autor del proyecto con el desarrollo de un modelo con capacidades similares, pero enfocado en Santiago de Chile.

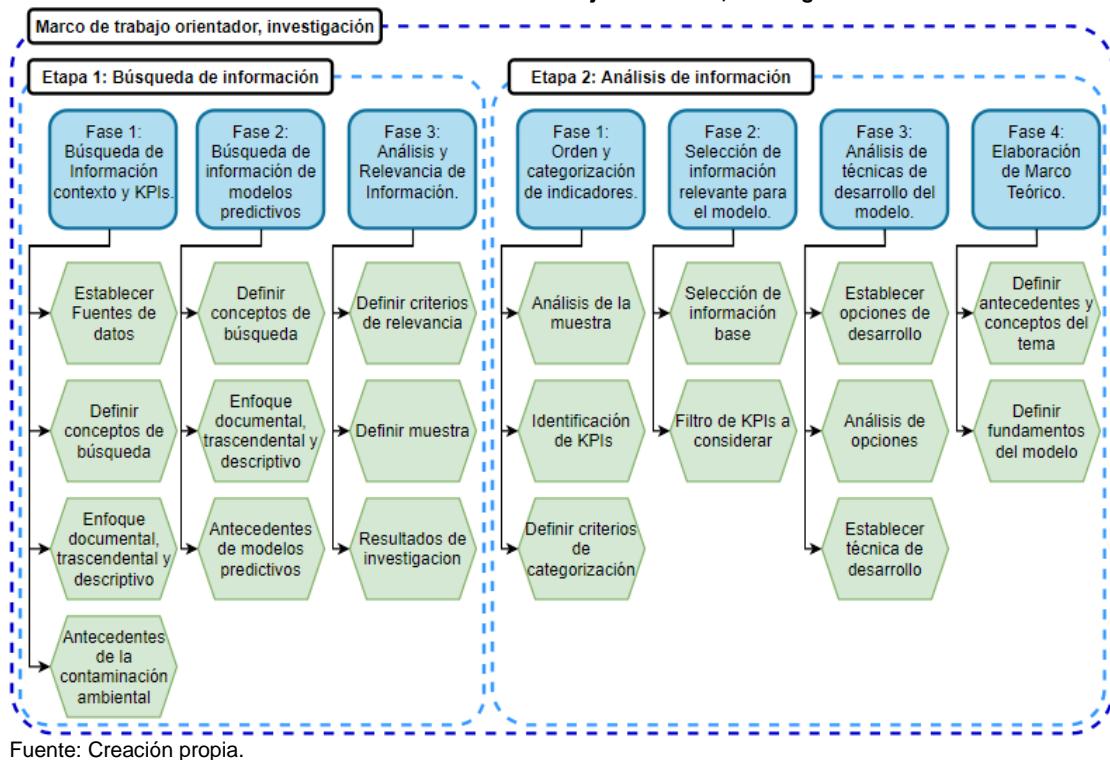
3.3. Estructura de la investigación

La investigación está dividida entre las etapas 1 y 2 del proyecto (ver Capítulo N.^º 1 y la Ilustración 1.1), dichas etapas se componen a su vez de distintas fases las cuales especifican el trabajo a realizar y le dan forma lógica y ordenada. Además, es importante agregar que la investigación cuenta con dos búsquedas de información, ambas con enfoques distintos pero que están relacionadas bajo el tema de la contaminación ambiental. Finalmente, y en base a todo lo anterior, se estructura la investigación dentro de un marco orientador para después, entregar los detalles del proceso en un marco operacional.

3.3.1. Marco de trabajo orientador de investigación

A continuación, se muestra la Ilustración 3.1, en ella, se puede apreciar el marco de trabajo orientador para las Etapas del proyecto correspondientes a la investigación:

Ilustración 3.1: "Marco de trabajo orientador, investigación"



El marco orientador presentado en la Ilustración 3.1 despliega el detalle de actividades que se realizan en cada fase, esto establece un orden en el accionar para dar coherencia a toda la investigación. Como se puede observar, las fases 1 y 2 de la primera etapa corresponden a las dos búsquedas que se realizarán, además, se puede ver como el procedimiento de ambas es prácticamente el mismo, salvo que en la primera búsqueda se define un aspecto esencial para ambas, que corresponde a la decisión de fuentes de información. La fase 3 de la primera etapa si bien corresponde a una fase de análisis, se decidió integrar en esta etapa por las características de sus actividades, dichas actividades tienen como objetivo definir criterios para analizar los títulos y autores encontrados, para así revelar los más importantes o influyentes y entregarlos como resultados, en otras palabras, un estudio bibliométrico. La segunda etapa está centrada en analizar la muestra bibliográfica obtenida en la etapa anterior con el fin de desarrollar un marco teórico lo suficientemente robusto como para sustentar y dar las bases a la propuesta planteada en el presente proyecto, que

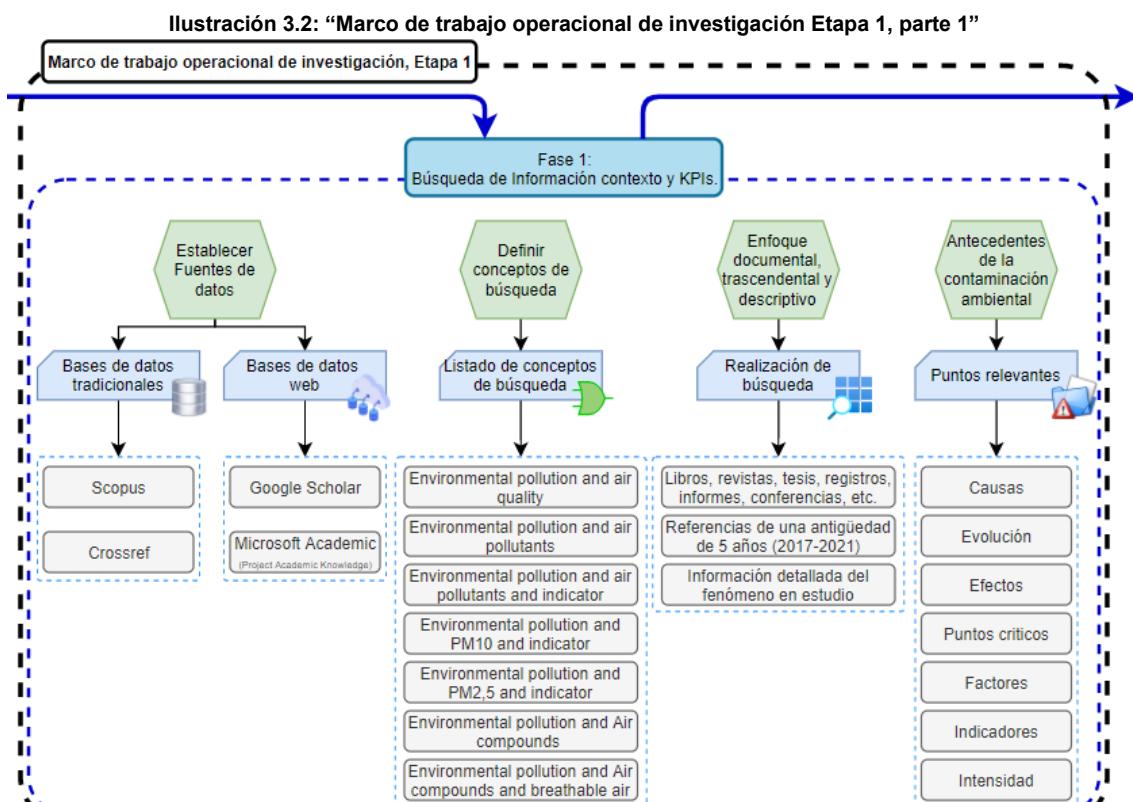
es un modelo predictivo de la contaminación en el aire de las comunas de la Región Metropolitana de Chile.

3.3.2. Marco de trabajo operacional de investigación

Luego de establecer el marco orientador del trabajo, se procede a explicar detalladamente el proceder de las actividades allí descritas (ver Ilustración 3.1). Para esto se utilizarán diversas ilustraciones que permitan hacer más sencilla y resumida la explicación de todo el proceso.

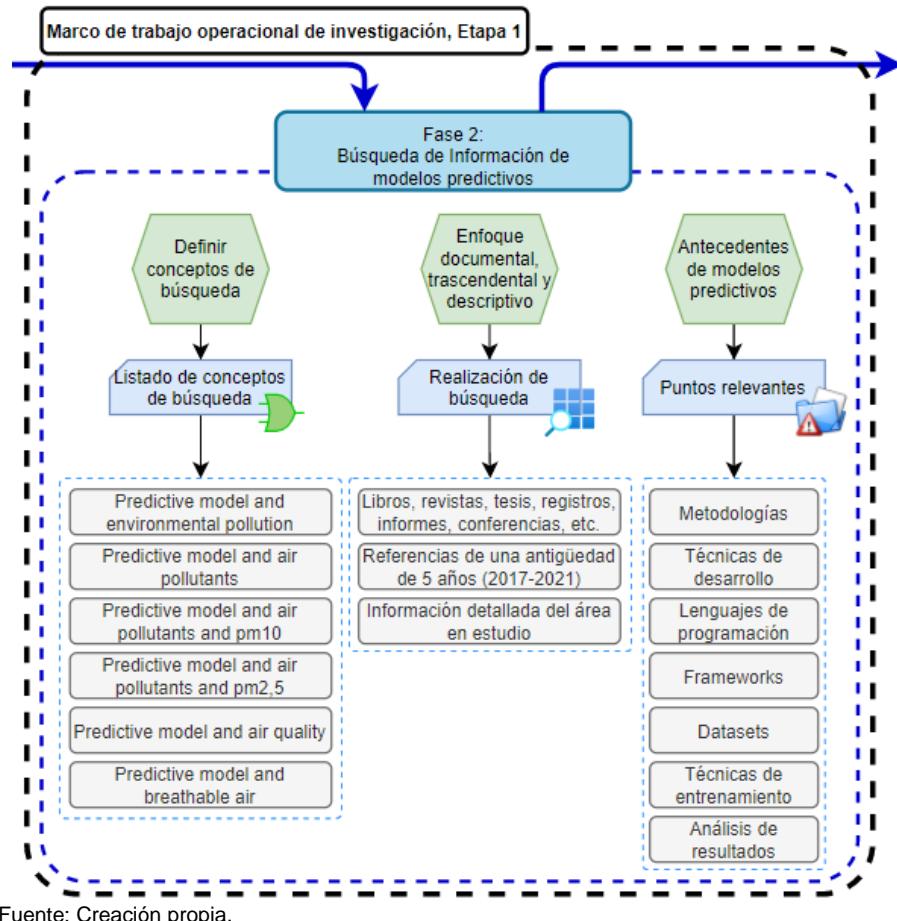
3.3.2.1. Marco de trabajo operacional Etapa 1: Búsqueda de Información

A continuación, se presentan las Ilustraciones de marcos operacionales para las Fases 1 y 2 de la Etapa 1 del proyecto:



Fuente: Creación propia.

Ilustración 3.3: “Marco de trabajo operacional de investigación Etapa 1, parte 2”

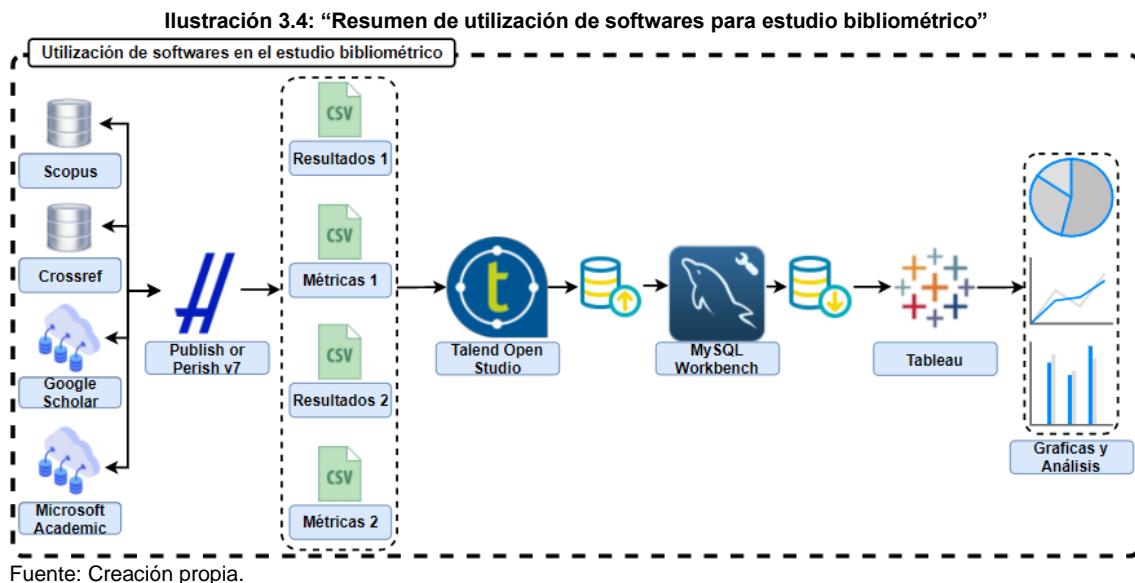


Fuente: Creación propia.

Como se puede ver en la ilustración 3.2, se muestran las principales fuentes de datos para la extracción de información para la construcción del marco teórico del presente trabajo, sin embargo, se debe explicar que para la realización de las búsquedas en estas bases de datos se utilizará el software **Publish or Perish v7** (POPV7) el cual tiene la capacidad de realizar una entrega de datos más ordenada y normalizada en una única estructura a partir de distintas bases de datos, su única desventaja es la incapacidad de realizar una misma consulta a distintas bases de datos al mismo tiempo, por lo que los resultados serán almacenados en múltiples archivos de formato CSV con la conveniente particularidad de poseer el mismo formato y estructura de los datos. Este es el escenario perfecto para la utilización de **Talend Open Studio**, un software ETL (Extract, Transform and Load) que permite la unión, depuración y

carga de datos provenientes de diferentes medios en un único lugar, de esa forma, para la creación y posterior poblamiento de una base de datos se utilizará también **MySQL Workbench**, esta es una herramienta visual de diseño de bases de datos la cual permitirá crear un pequeño Data Warehouse con los resultados de las búsquedas. Los resultados de las búsquedas representadas en las Ilustraciones 3.2 y 3.3 serán almacenados en distintas bases de datos, pero con la misma estructura, todo esto con el fin de manejar los análisis de dichos datos por separado.

Lo explicado en el párrafo anterior se puede resumir gráficamente con la siguiente ilustración:

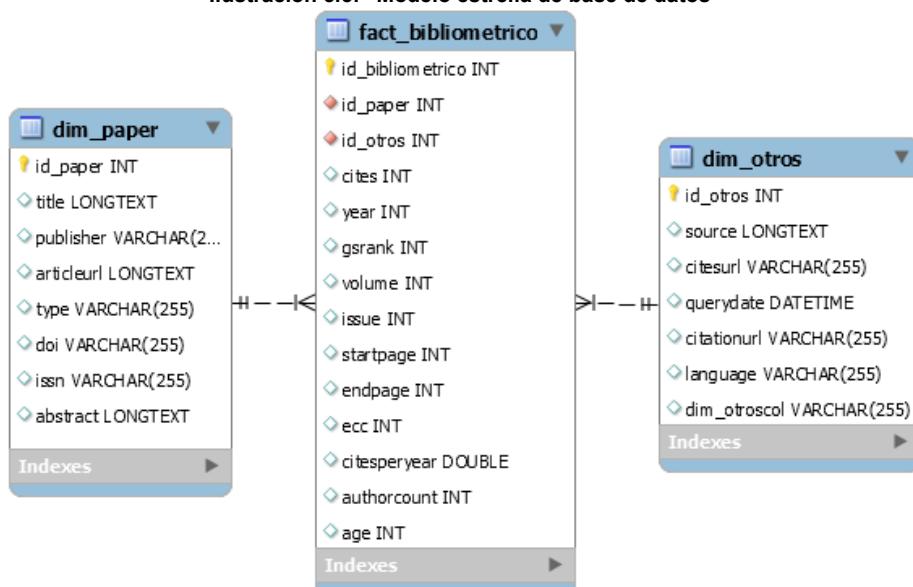


La Ilustración 3.4, como se dijo anteriormente, resume lo explicado referente a la utilización de softwares en el estudio bibliométrico, pero, es necesario agregar que este proceso debe repetirse dos veces, aunque no en su totalidad, ya que, por ejemplo, el modelo utilizado en MySQL Workbench es el mismo para ambas búsquedas. Se puede observar, además, en el diagrama de la Ilustración 3.4, la integración de otro software llamado **Tableau** el cual recibe la información de la base de datos conectándose con el server de MySQL Workbench, para luego, trabajar con ella para la generación de gráficas de

diversos las cuales podrán ser analizadas más adelante en el estudio bibliométrico.

Respecto a las búsquedas que Publish or Perish v7 deberá realizar, si bien representan una tarea simple de efectuar, llevar a cabo todas las búsquedas toma una cantidad de tiempo considerable, debido a que, en primera instancia, el estudio está basado en dos búsquedas generales las cuales a su vez están compuestas de varios conjuntos de conceptos, 7 en el caso de la primera (Ver Ilustración 3.2), y 6 en el caso de la segunda (Ver Ilustración 3.3), por lo que deberán realizarse 13 consultas por base de datos, es decir, que al utilizar 4 de ellas como fuentes de información se deberán realizar 52 consultas en total por cada lenguaje seleccionado, y al ser el inglés y el español los idiomas en los que se realizarán las búsquedas, finalmente se obtiene el número de 104 consultas totales para obtener toda la información necesaria (56 para la primera búsqueda general, 48 para la segunda) para luego, almacenarla en dos bases de datos, una por búsqueda general. Para ver las consultas realizadas en PoPv7, visite el anexo 2. Para la realización del almacenamiento de la información, se construye el siguiente modelo estrella de base de datos:

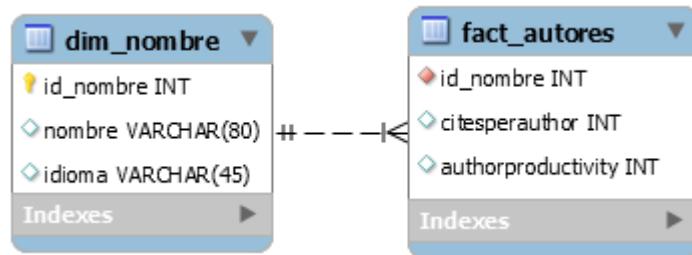
Ilustración 3.5: “Modelo estrella de base de datos”



Fuente: Creación propia, utilizando el software MySQL Workbench.

El modelo que se muestra en la Ilustración 3.5 fue creado con MySQL Workbench y se utilizará para el almacenamiento de toda la información obtenida por las consultas a bases de datos exceptuando la de los autores, además, se recuerda que se utilizaran dos bases de datos por separado con el mismo modelo, una para cada búsqueda general. La información de los autores será almacenada a su vez, en el siguiente modelo:

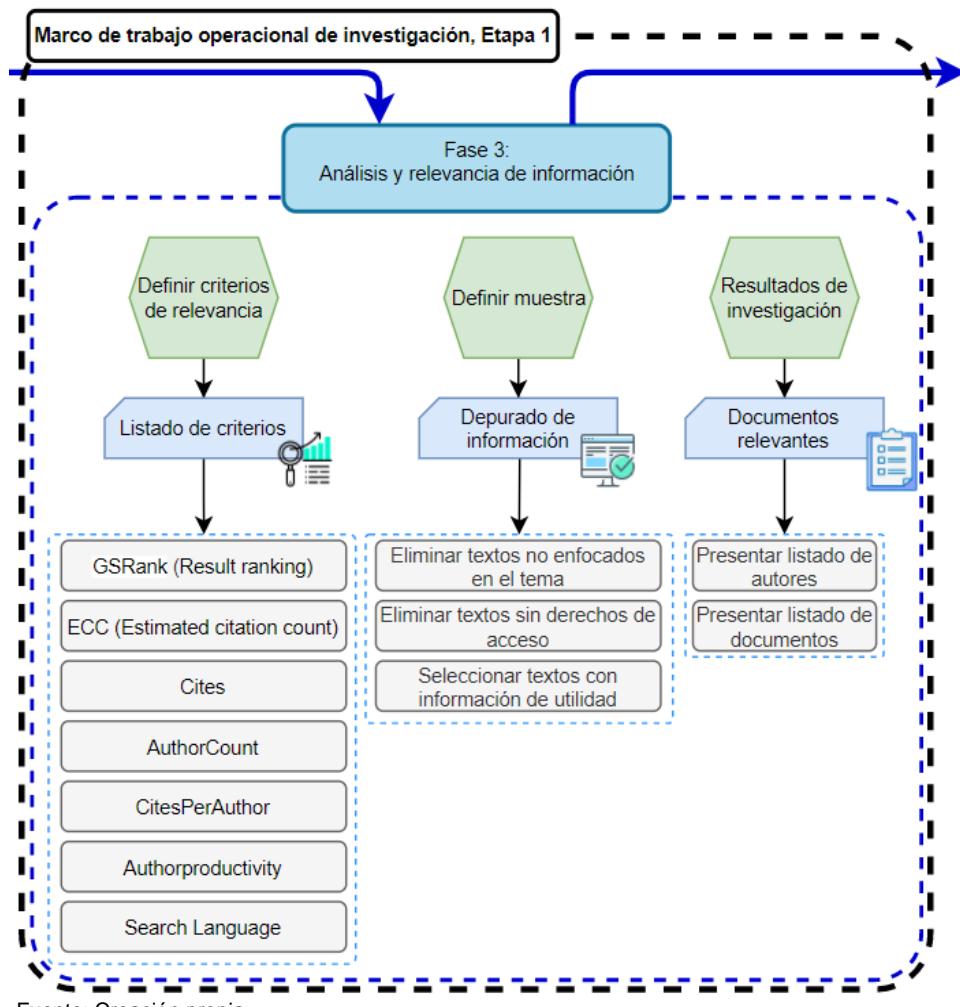
Ilustración 3.6: “Modelo BDD de almacenamiento de autores relevantes”



Fuente: Creación propia, utilizando el software MySQL Workbench.

La información de los autores se almacenará por separado debido a que existen casos donde múltiples autores participan de un paper, y ya que, la información que se ofrece de citas por autor es de un solo campo, se debe reestructurar dicha información en dos columnas, una con nombres únicos de autor y otra con la suma total de citas por cada uno de ellos, es una tarea que, a pesar de su aparente sencillez, puede tener cierta complejidad debido a que existen registros en los resultados donde participan hasta más de 40 autores por un solo trabajo. El modelo presentado en la Ilustración 3.6 será utilizado 2 veces, una para cada búsqueda general de la investigación.

Ilustración 3.7: “Marco de trabajo operacional de investigación Etapa 1, parte 3”



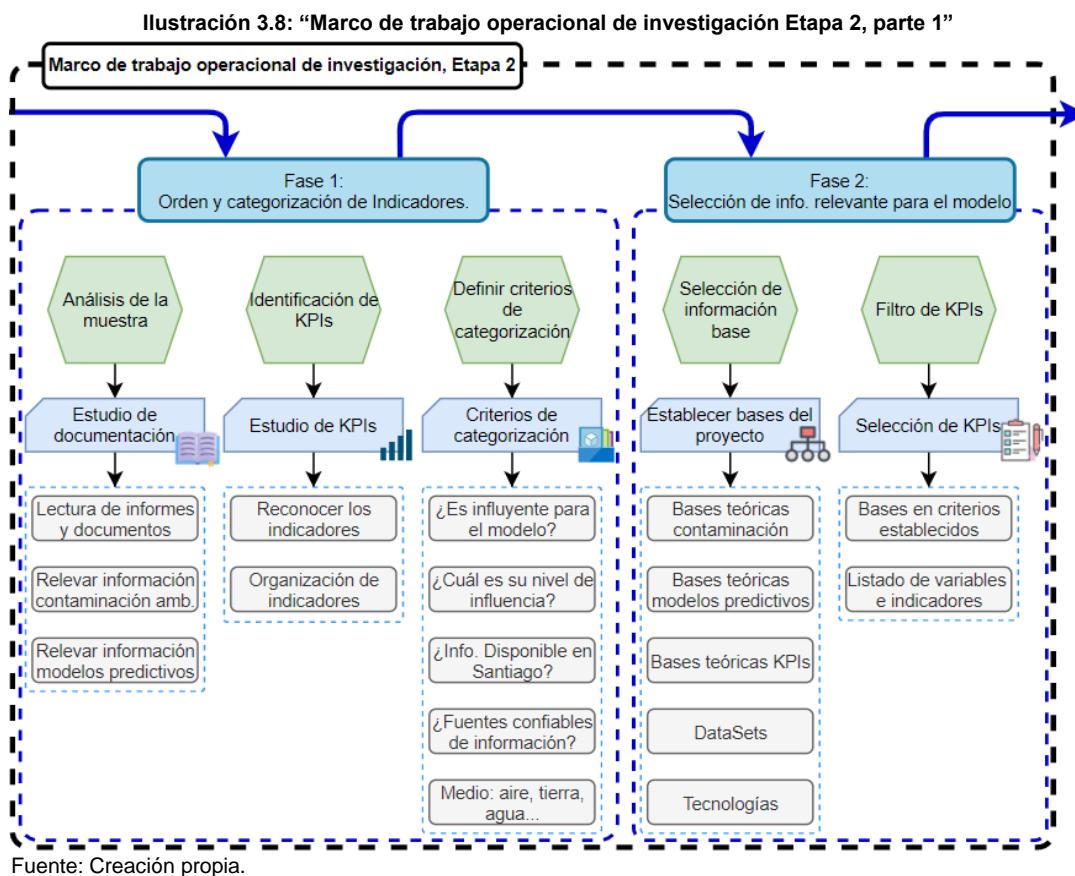
Fuente: Creación propia.

La Ilustración 3.7 muestra la última fase de la primera etapa del proyecto, en ella se presentan las actividades que se realizarán para el relevamiento de información de los resultados de las búsquedas. La primera tarea es definir los criterios de evaluación o relevancia para los resultados encontrados, estos se definen apoyándose principalmente en los ya ofrecidos por el software Publish or Perish v7, salvo por el caso de “AuthorProductivity” y “Search Language”, estos se agregaron con el propósito de ofrecer un análisis más completo. Luego de establecer las mejores publicaciones en base a los criterios anteriormente nombrados, se deben eliminar aquellas que, por motivos de fuerza mayor, no es

posible acceder a su contenido para luego obtener la lista definitiva de resultados.

3.3.2.2. Marco de trabajo operacional Etapa 2: Análisis de información

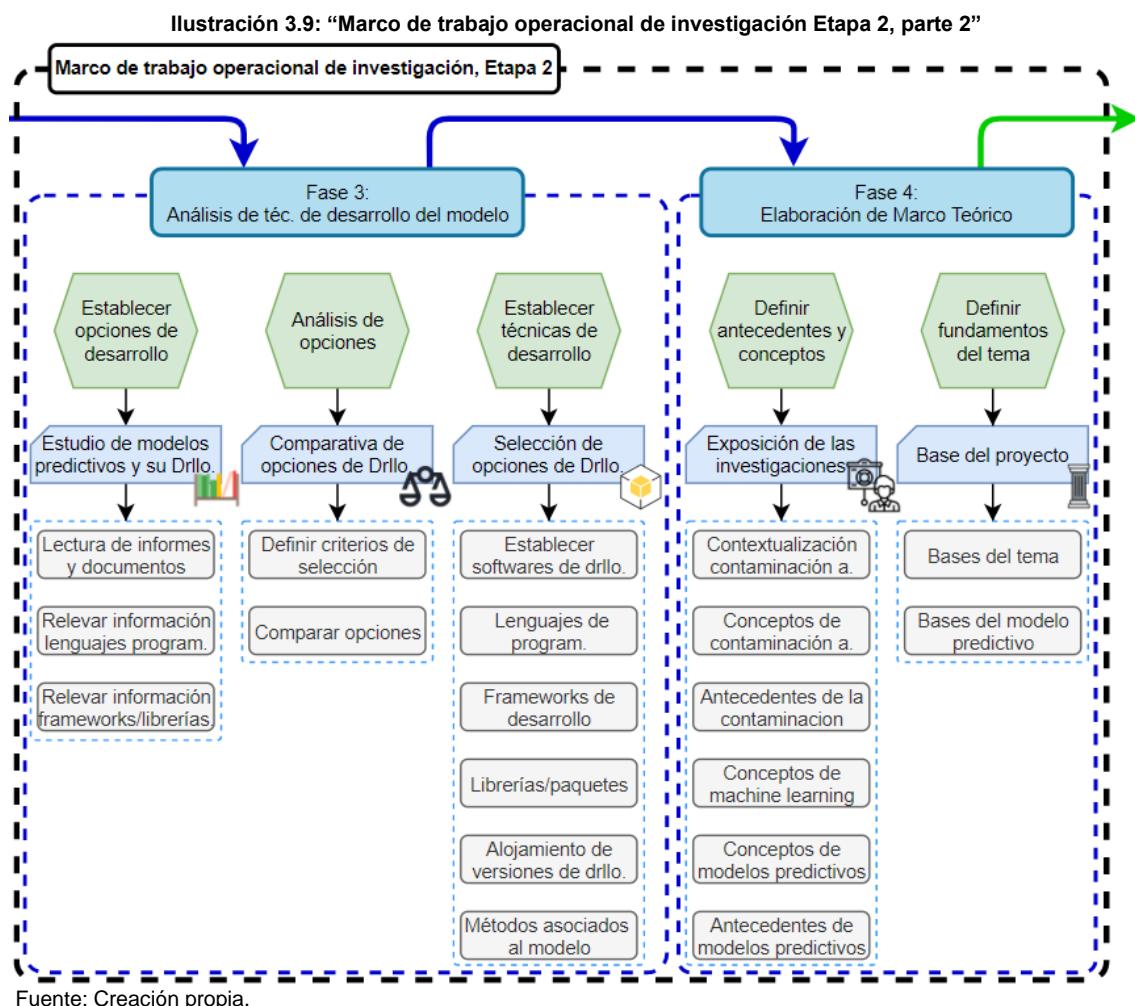
A continuación, se presenta la Ilustración del marco operacional para las Fases 1 y 2 de la Etapa 2 del proyecto:



En la Ilustración 3.8, se representan las actividades a desarrollar para las Fases 1 y 2 de la segunda etapa de investigación del proyecto. En la fase 1, se debe estudiar la información obtenida por el estudio bibliométrico previamente realizado, con el fin de relevar la información más importante y necesaria para el proyecto, luego, se identifican los KPIs de contaminación ambiental que se puedan encontrar en dicha información y finalmente, se crean criterios de

categorización de estos. En la fase 2, se utilizan los criterios de categorización de los KPIs para seleccionar aquellos más relevantes para el modelo predictivo y también aquellos de los cuales se disponga de datos en Santiago, en otras palabras, que también sea posible su utilización. Adicionalmente, la fase 2 del proyecto se asegura de establecer las bases del proyecto, seleccionando información clave y vital para el entendimiento de este y de todo su desarrollo.

A continuación, se presenta la Ilustración del marco operacional para las Fases 3 y 4 de la Etapa 2 del proyecto:



La Ilustración 3.9 muestra las fases 3 y 4 de la investigación del proyecto. En el caso de la fase 3, se puede ver que su propósito es establecer las técnicas de desarrollo que se utilizaran para realizar el modelo predictivo, para esto,

primero se deben establecer las opciones de desarrollo de este, luego, analizar dichas opciones y finalmente compararlas. En la fase 4, y como resultado de toda la investigación, se construye el marco teórico, el cual, debe incluir descripciones de toda la teoría y conceptos necesarios para el desarrollo completo del proyecto. La información necesaria para la construcción del marco teórico está compuesta por la contextualización en el ámbito de la contaminación ambiental, sus conceptos y antecedentes; la contextualización en el ámbito del machine learning y modelos predictivos, sus conceptos y antecedentes, y, toda aquella información que presente una utilidad clave para el proyecto. La Ilustración 3.9, a partir de la fase 4, presenta una flecha de salida de un color distinto al habitual en estos modelos, esto representa, el final del marco operacional de investigación y el comienzo del marco de desarrollo del proyecto.

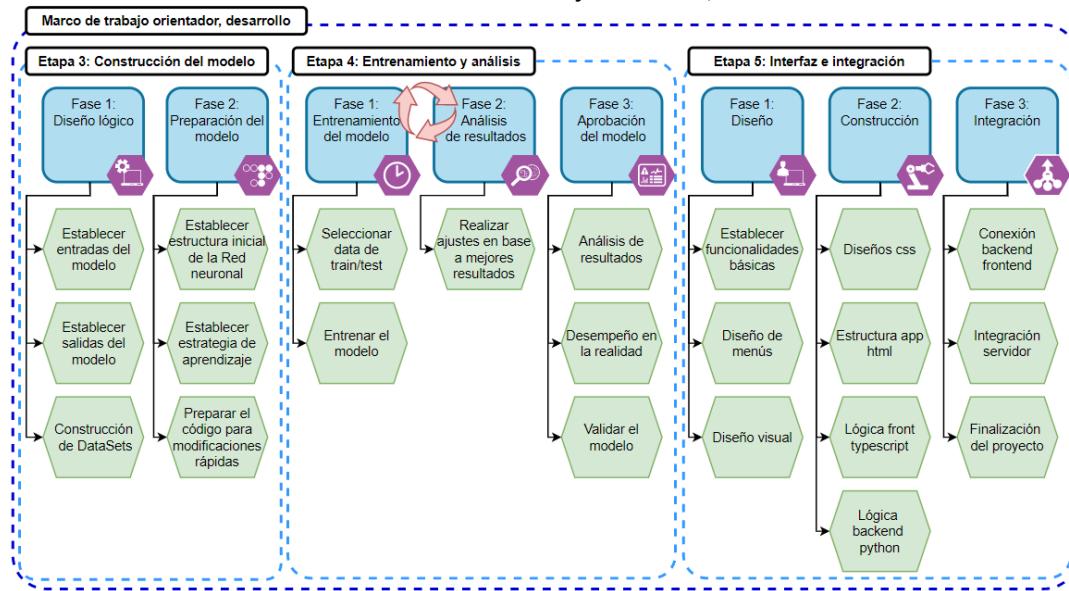
CAPÍTULO 4 - METODOLOGÍA DE DESARROLLO

Este capítulo se enfoca en las Etapas 3, 4 y 5 del proyecto, las cuales corresponden a la construcción del modelo predictivo, su aplicación, y su diseño. Al igual que en las Etapas de investigación (ver Capítulo N.º 3) estas Etapas se dividen en Fases las cuales se encuentran detalladas en el Capítulo N.º 1 del presente trabajo y resumidas visualmente en la Ilustración 1.1. Se describe a continuación, el trabajo que se realiza por cada una de las Etapas y Fases anteriormente mencionadas, y metodología de desarrollo del proyecto.

4.1. Marco de trabajo orientador de desarrollo

A continuación, se muestra la Ilustración 5.1, en ella, se puede apreciar el marco de trabajo orientador para las Etapas del proyecto correspondientes al desarrollo:

Ilustración 4.1: “Marco de trabajo orientador, desarrollo”



Fuente: Creación propia.

Como se puede observar en la Ilustración 5.1, la estructuración del trabajo a realizar en esta parte del proyecto queda bastante clara, iniciando por la Etapa

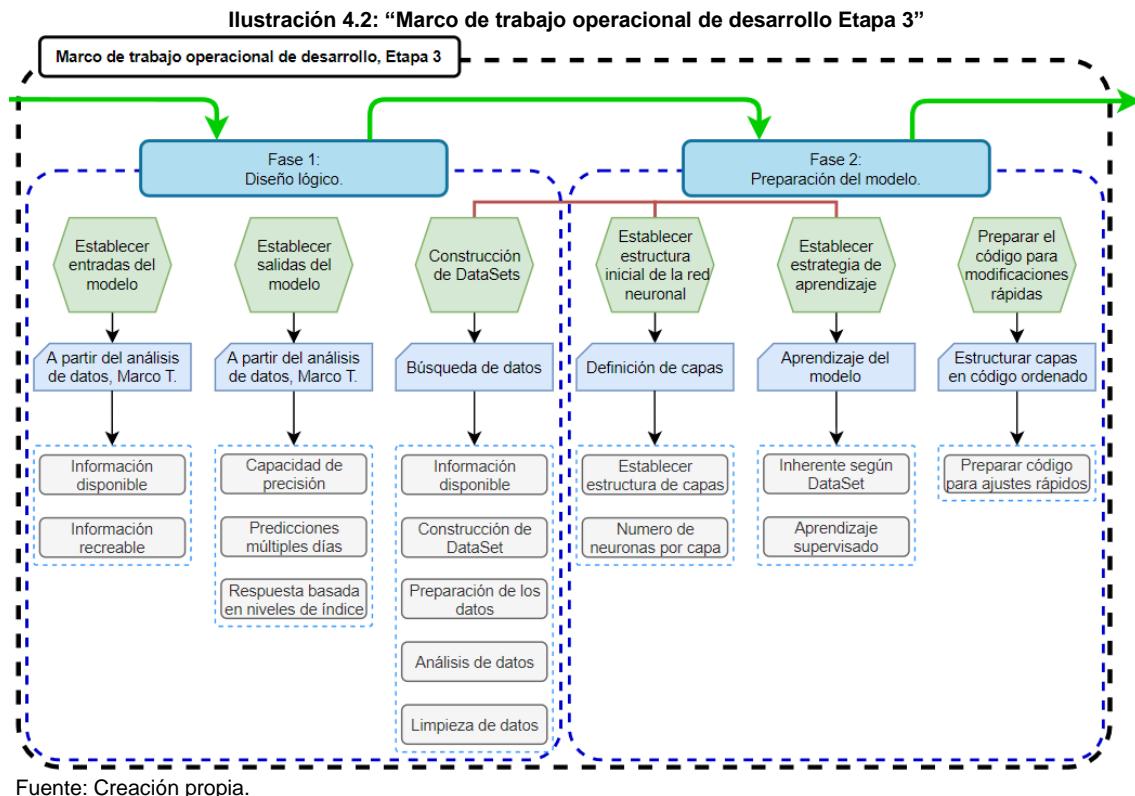
3 donde se plantearán las bases para diseñar la Red Neuronal a utilizar en el modelo, acto seguido, la Etapa 4 donde se realizarán trabajos de ajuste y entrenamiento del modelo, y finalmente, la Etapa 5, que consiste en la construcción de una aplicación web para consumir el modelo predictivo, el cual estará alojado en un servicio web.

4.2. Marco de trabajo operacional de desarrollo

A continuación, se presenta el detalle operacional de cada Etapa (etapas 3, 4 y 5) considerada dentro del desarrollo del proyecto, en donde se describe el detalle de acciones o pasos a seguir para lograr un producto aceptable y un modelo predictivo de contaminación ambiental funcional.

4.2.1. Marco de trabajo operacional Etapa 3: Construcción del modelo

La siguiente ilustración muestra las actividades a desarrollar para la Etapa 3 del proyecto, la cual consiste en la construcción del modelo predictivo:

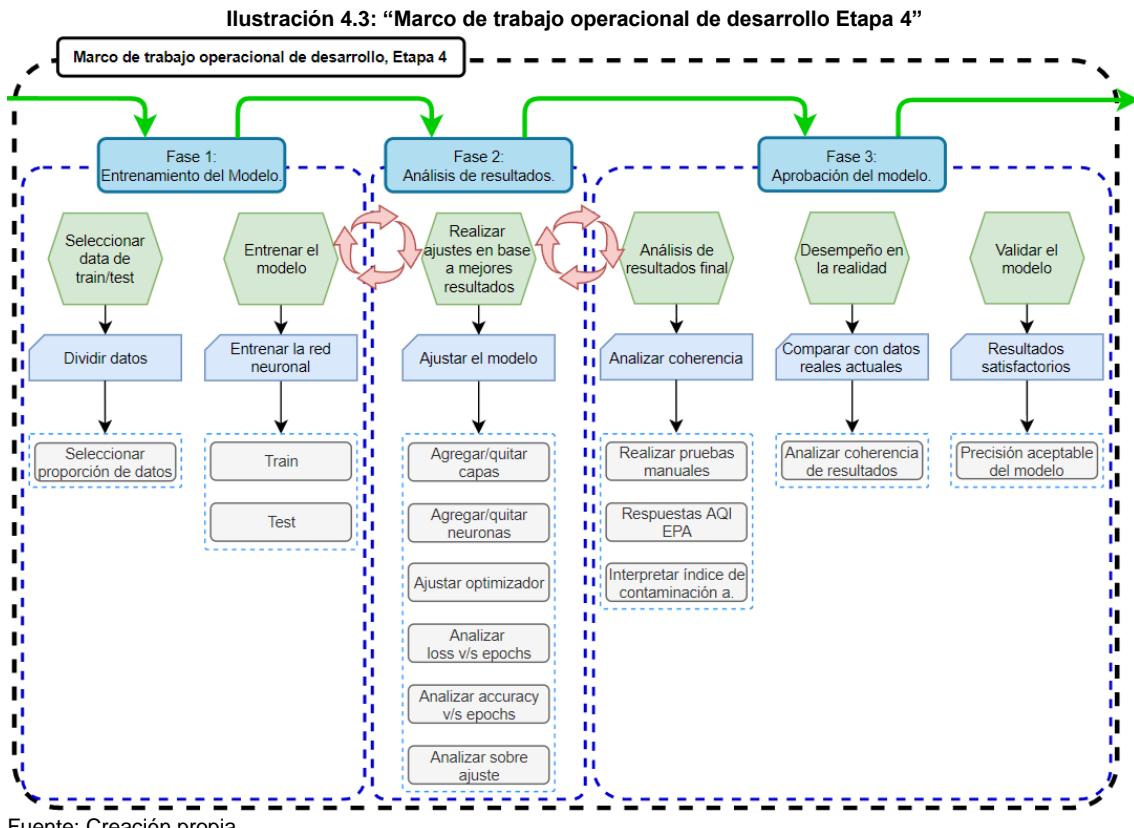


Como se muestra en la Ilustración 4.2, la Fase 1 de esta etapa del proyecto consiste en el diseño lógico de la red neuronal, donde a partir de los datos analizados en el marco teórico y aquellos que se encuentren disponibles se crearán las entradas y salidas del modelo, donde las entradas serán los parámetros de contaminación del aire, presión, temperatura, fecha de medición, entre otros, y las salidas los días a predecir. Recordemos que se busca un numero de 30 días de predicción por lo que se probara la creación de una red neuronal con 30 salidas. Finalmente, dentro de la Etapa 3 se construirá el DataSet para el entrenamiento del modelo, el cual estará conformado con datos de diversas fuentes ya que lamentablemente no se cuenta con un conjunto de datos único, esto hace que la conformación de los datos de entrenamiento sea más laboriosa.

La Fase 2 de la Etapa 3, como se muestra en la Ilustración 4.2, está fuertemente relacionada con la Fase 1, ya que, para establecer la estructura inicial del modelo y la estrategia de entrenamiento de este, se necesita conocer de que información o datos se disponen. Al término de la Fase 2, se debe estructurar un código del modelo de fácil actualización o modificación, ya que es un caso típico en estos desarrollos encontrarse con situaciones de ajuste en base a ensayo y error probando el modelo.

4.2.2. Marco de trabajo operacional Etapa 4: Entrenamiento y análisis

La siguiente ilustración muestra las actividades a desarrollar para la Etapa 4 del proyecto, la cual consiste en el entrenamiento del modelo predictivo del modelo predictivo:



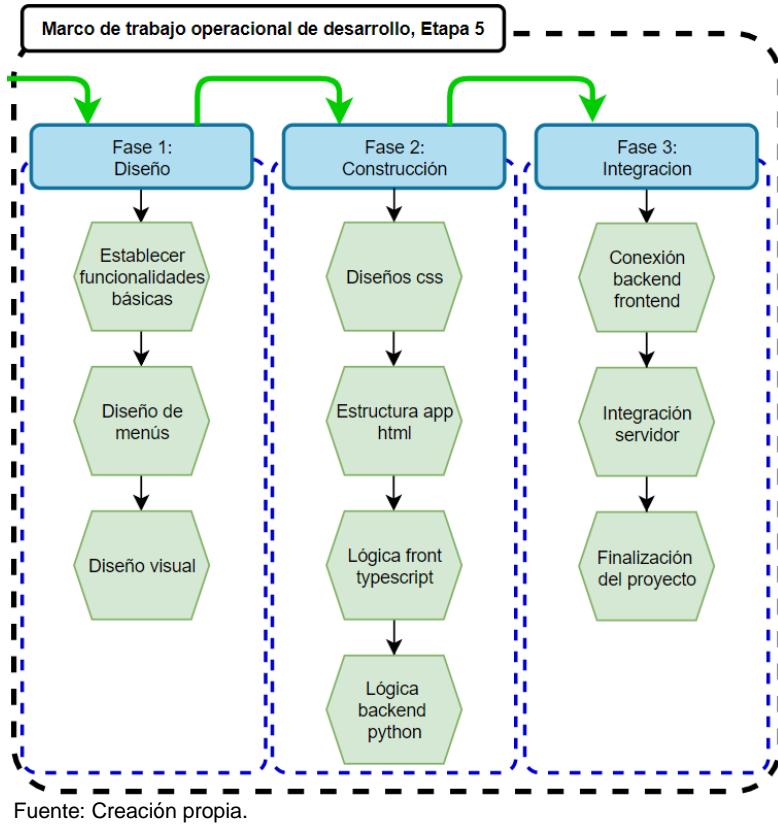
Fuente: Creación propia.

En la Ilustración 4.3 se pueden ver las Fases 1, 2 y 3 de La Etapa 4 del proyecto, las cuales tienen como característica principal una Iteración entre la última actividad de la Fase 1 (Entrenar el modelo), la Fase 2 (Análisis de resultados) y la primera actividad de la Fase 3 (Análisis de resultados). En conjunto, las 3 actividades anteriormente nombradas representan el proceso fundamental para la realización de un modelo predictivo o red neuronal que cumpla con altos parámetros de precisión, luego, una vez obtenido un modelo aceptable, se deben realizar pruebas de desempeño en la realidad para finalmente validar el modelo y exportarlo para su utilización dentro de un Web service.

4.2.3. Marco de trabajo operacional Etapa 5: Interfaz e integración

La siguiente ilustración muestra las actividades a desarrollar para la Etapa 5 del proyecto, la cual consiste en la construcción del web service e interfaz web del modelo predictivo:

Ilustración 4.4: “Marco de trabajo operacional de desarrollo Etapa 5”



El marco de trabajo operacional presentado en la Ilustración 4.4 muestra una forma más simplificada respecto a los anteriores, esto se debe a que las actividades aquí escritas se encuentran dentro de un ámbito no principal para este proyecto, el cual corresponde a desarrollo web. La intención de esta etapa del proyecto es construir un sistema que sea capaz de utilizar el modelo predictivo anteriormente desarrollado y mostrar sus resultados de forma gráfica y atractiva. Para su construcción se utilizarán herramientas como Django y Angular.

Luego del término de la Etapa 5, el proyecto se da por finalizado y se daría pie a la generación de recomendaciones para futuras actualizaciones del modelo.

CAPÍTULO 5 – DESARROLLO DEL MODELO

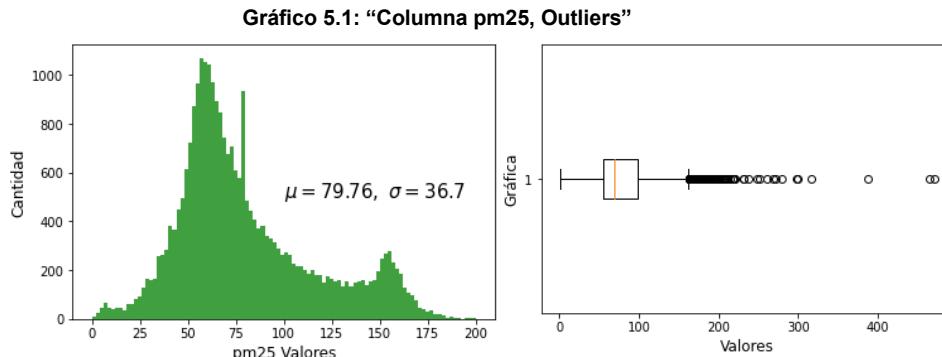
En el presente capítulo se procede a desarrollar el modelo predictivo utilizando la información recolectada para la creación del DataSet. En primera instancia se trabaja sobre los datos limpiándolos y adaptándolos para el entrenamiento del modelo, luego este se construye en base al algoritmo de Random Forest. Finalmente, en este capítulo, se diseñará la interfaz de utilización del modelo predictivo en conjunto de las gráficas de datos que este mostrará al usuario.

5.1. Limpieza del DataSet

En primera instancia se realiza el reemplazo de aquellos datos nulos o de valor NaN (vacío) por el promedio de la columna respectiva, luego de realizar aquella acción, se procede al análisis de Outliers.

5.1.1. Análisis Outliers columna PM25

A continuación, las gráficas de datos y Outliers de la columna PM25:

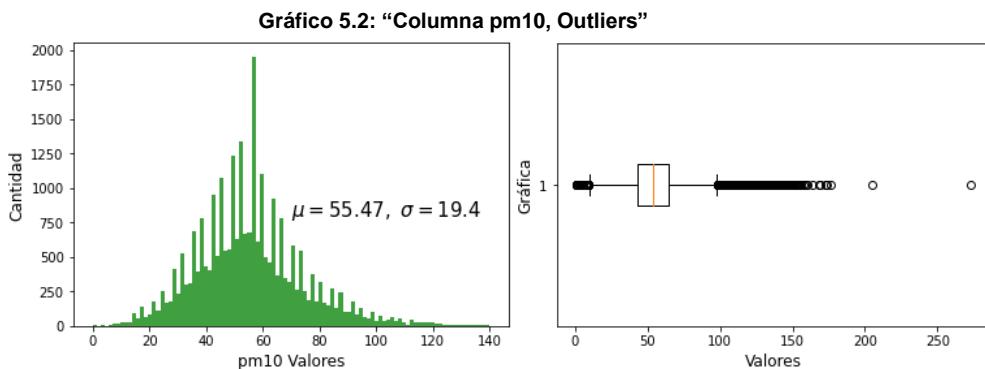


Fuente: Creación propia utilizando Google Colab.

Los Outliers detectados en esta columna corresponden a 612 datos, lo que representa un 2,24% de los datos, por lo que su eliminación es viable.

5.1.2. Análisis Outliers columna PM10

A continuación, las gráficas de datos y Outliers de la columna PM10:

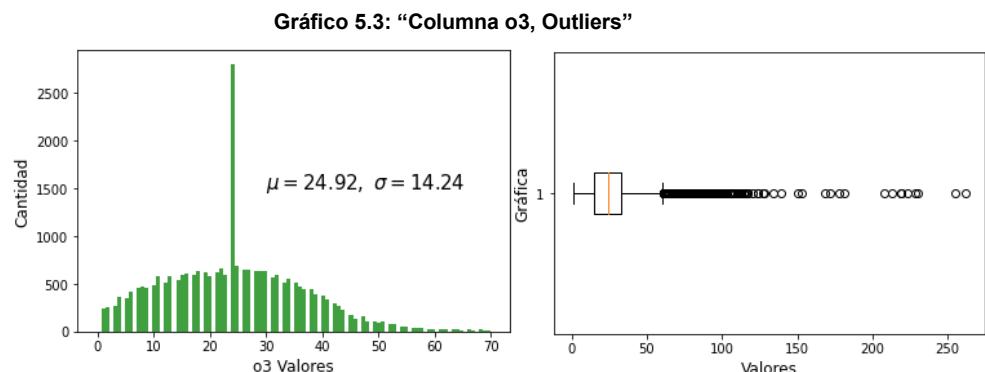


Fuente: Creación propia utilizando Google Colab.

Los Outliers detectados en esta columna corresponden a 778 datos, lo que representa un 2,91% de los datos actuales, por lo que su eliminación es viable.

5.1.3. Análisis Outliers columna O3

A continuación, las gráficas de datos y Outliers de la columna O3:

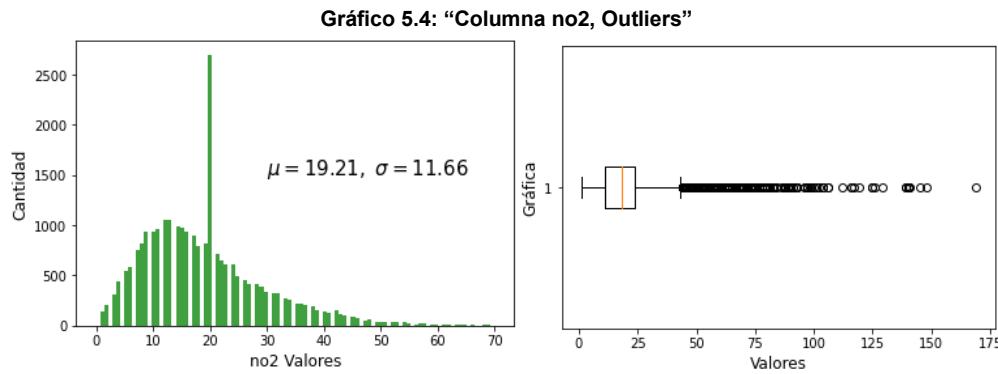


Fuente: Creación propia utilizando Google Colab.

Los Outliers detectados en esta columna corresponden a 367 datos, lo que representa un 1,41% de los datos actuales, por lo que su eliminación es viable.

5.1.4. Análisis Outliers columna NO2

A continuación, las gráficas de datos y Outliers de la columna NO2:

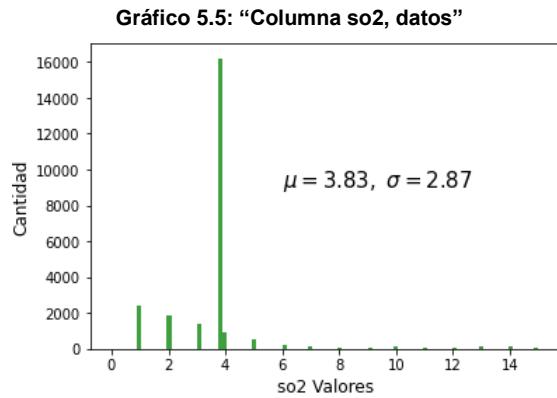


Fuente: Creación propia utilizando Google Colab.

Los Outliers detectados en esta columna corresponden a 868 datos, lo que representa un 3,4% de los datos actuales, por lo que su eliminación es viable.

5.1.5. Análisis Outliers columna SO2

A continuación, las gráficas de datos y Outliers de la columna SO2:



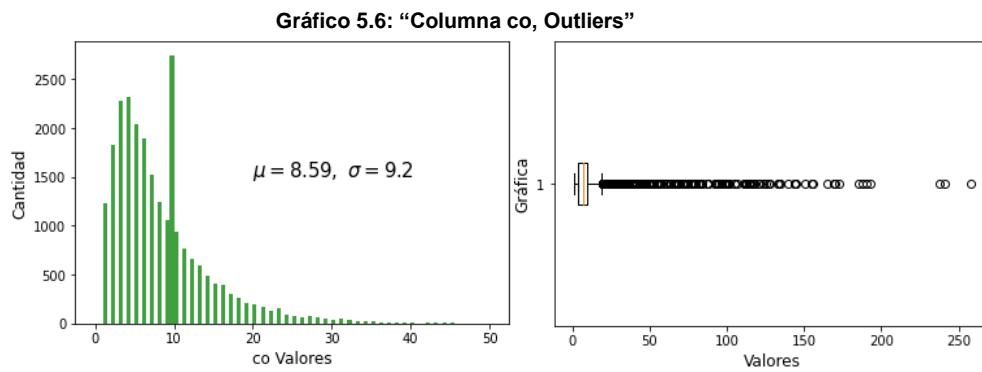
Fuente: Creación propia utilizando Google Colab.

Esta columna presenta un caso especial, y se debe a la gran cantidad de nulos que poseía en un inicio, por lo que, al intercambiar dichos datos nulos al

promedio, perjudican al modelo con un nivel excesivo de datos cercanos a un mismo valor (Como se puede ver en el Gráfico 5.5), por lo que se opta por eliminar la columna del Dataset.

5.1.6. Análisis Outliers columna CO

A continuación, las gráficas de datos y Outliers de la columna CO:



Fuente: Creación propia utilizando Google Colab.

Los Outliers detectados en esta columna corresponden a 1.464 datos, lo que representa un 5,94% de los datos actuales, por lo que su eliminación es viable.

Finalmente, luego de la limpieza de Outliers, el recuento de datos llega a 23.175 filas, este valor seguirá disminuyendo dependiendo del Dataset que se genere para cada uno de los modelos (recordar que cada modelo será entrenado para predecir un día en concreto), a continuación, la explicación de porque esto ocurrirá.

5.2. Diseño de inputs y outputs

Para la construcción del Dataset, se pretende utilizar las entradas:

- PM25, PM10, O3, NO2, CO, Código_comuna, temperatura, Fecha_d, Fecha_m, Fecha_y, Aqi_max_actual, Aqi_mean_ctual.

Todas las entradas del Dataset son de tipo float, y serán de un carácter fijo, a diferencia de su única salida (nivel_o), la cual cambiara en base al día que se deseé predecir, por ejemplo, si se desea predecir el nivel de contaminantes del aire del segundo día a partir de la medición, los datos de entrada X, se mantendrán fijos, mientras que los datos de salida Y se moverán dos “filas” hacia atrás, enlazando así las variables de entrada del día N con las variables de salida del día N+2. Realizar este ajuste en los datos de salida generará naturalmente filas con variables dependientes nulas (básicamente se desplazan los datos), de ahí, la variación en el número de filas de cada Dataset. Es importante considerar que este desplazamiento en los datos se hace en base a las fechas y a los códigos de las comunas, para evitar vinculación entre variables dependientes e independientes que no tengan que ver con el ajuste.

5.3. Descripción del DataSet

El Dataset que se utilizará para el entrenamiento del modelo está compuesto por 30 Datasets, los cuales disponen de toda la información necesaria para poder construir un modelo predictivo, sin embargo, unir toda esta información no representó una tarea sencilla debido a que los datos solo se relacionaban entre si mediante fechas, en algunos casos en distintos formatos, por lo que en primera instancia se debió normalizar la información para poder unirla, luego, se debió crear columnas de códigos de comunas, para que en conjunto con la fecha, se pudieran unir los datos donde correspondían realmente. La creación de múltiples Datasets responde a la necesidad de crear salidas para 30 días a futuro con los mismos datos de entrada, y, tras unirlos, se logró crear un Dataset único el cual dispone de 493.917 filas de datos. A continuación, El primer Dataset creado, con enfoque de salidas para la predicción de un día a futuro:

Ilustración 5.1: “Dataset v1, sin outputs”

	fecha	pm25	pm10	o3	no2	co	codigo_comuna	temperatura	dia_futuro	fecha_d	fecha_m	fecha_y	aqi_max_actual	aqi_mean_actual	
0	03/08/2021	148.000000	68.000000	24.354625	20.126054	9.617034		0	9.454167	1.0	3.0	8.0	2021.0	198.760801	80.142009
1	04/08/2021	124.000000	55.000000	24.354625	20.126054	9.617034		0	7.850000	1.0	4.0	8.0	2021.0	186.368809	76.376742
2	05/08/2021	109.000000	52.000000	24.354625	20.126054	9.617034		0	9.262500	1.0	5.0	8.0	2021.0	178.623815	74.257372
3	06/08/2021	111.000000	69.000000	24.354625	20.126054	9.617034		0	8.387500	1.0	6.0	8.0	2021.0	179.656481	76.420135
4	07/08/2021	151.000000	49.000000	24.354625	20.126054	9.617034		0	9.691667	1.0	7.0	8.0	2021.0	200.500501	78.077154
...	
21605	01/01/2014	79.755928	52.000000	27.000000	6.000000	3.000000		9	20.950000	1.0	1.0	1.0	2014.0	163.524136	55.284714
21606	31/03/2014	79.755928	42.000000	20.000000	16.000000	8.000000		9	14.412500	1.0	31.0	3.0	2014.0	163.524136	64.405177
21607	27/02/2018	79.755928	56.156214	28.000000	11.000000	7.000000		9	18.000000	1.0	27.0	2.0	2018.0	163.524136	65.479937
21608	28/02/2018	79.755928	56.156214	31.000000	14.000000	18.000000		9	18.333333	1.0	28.0	2.0	2018.0	216.778523	94.757235
21609	13/10/2015	79.755928	56.156214	11.000000	15.000000	4.000000		9	15.541667	1.0	13.0	10.0	2015.0	163.524136	56.977415

21610 rows x 14 columns

Fuente: Creación propia utilizando Google Colab.

Luego de todos los ajustes necesarios al Dataset y antes del escalonado de datos, se agregó el Output (columna nivel_o):

Ilustración 5.2: “Dataset v2, con output”

	fecha	pm25	pm10	o3	no2	co	codigo_comuna	temperatura	dia_futuro	fecha_d	fecha_m	fecha_y	aqi_max_actual	aqi_mean_actual	nivel_o	
0	03/08/2021	148.000000	68.000000	24.354625	20.126054	9.617034		0	9.454167	1.0	3.0	8.0	2021.0	198.760801	80.142009	4
1	04/08/2021	124.000000	55.000000	24.354625	20.126054	9.617034		0	7.850000	1.0	4.0	8.0	2021.0	186.368809	76.376742	4
2	05/08/2021	109.000000	52.000000	24.354625	20.126054	9.617034		0	9.262500	1.0	5.0	8.0	2021.0	178.623815	74.257372	4
3	06/08/2021	111.000000	69.000000	24.354625	20.126054	9.617034		0	8.387500	1.0	6.0	8.0	2021.0	179.656481	76.420135	5
4	07/08/2021	151.000000	49.000000	24.354625	20.126054	9.617034		0	9.691667	1.0	7.0	8.0	2021.0	200.500501	78.077154	4
...	
21605	01/01/2014	79.755928	52.000000	27.000000	6.000000	3.000000		9	20.950000	1.0	1.0	1.0	2014.0	163.524136	55.284714	4
21606	31/03/2014	79.755928	42.000000	20.000000	16.000000	8.000000		9	14.412500	1.0	31.0	3.0	2014.0	163.524136	64.405177	3
21607	27/02/2018	79.755928	56.156214	28.000000	11.000000	7.000000		9	18.000000	1.0	27.0	2.0	2018.0	163.524136	65.479937	5
21608	28/02/2018	79.755928	56.156214	31.000000	14.000000	18.000000		9	18.333333	1.0	28.0	2.0	2018.0	216.778523	94.757235	5
21609	13/10/2015	79.755928	56.156214	11.000000	15.000000	4.000000		9	15.541667	1.0	13.0	10.0	2015.0	163.524136	56.977415	4

21610 rows x 15 columns

Fuente: Creación propia utilizando Google Colab.

Luego de efectuar una limpieza de datos más profunda, el resultado es aquel que se muestra en la Ilustración 5.2, donde se procede a definir cada una de las columnas:

- PM2,5: Representa los niveles de concentración de material particulado 2,5 en unidades de microgramos.metro cúbico (µg/m3) durante un periodo de medición de 24 horas (promedio).
- PM10: Representa los niveles de concentración de material particulado 10 en unidades de microgramos.metro cúbico (µg/m3) durante un periodo de medición de 24 horas (promedio).
- O3: Representa los niveles de Ozono en partes por billón (ppb) medidos durante un periodo de 8 horas (promedio).

- NO2: Representa los niveles de Nitrato en el aire en partes por billón (ppb) medido durante un periodo de 1 hora (promedio).
- CO: Representa los niveles de monóxido de carbono en el aire en partes por millón (ppm) en un periodo de 8 horas (promedio).
- Código_comuna: Código de diccionario asignado a la comuna de la cual se tratan los datos, a continuación, su asignación:
 - 0: Talagante.
 - 1: Pudahuel.
 - 2: Santiago.
 - 3: Puente Alto.
 - 4: La Florida.
 - 5: Las Condes.
 - 6: Independencia.
 - 7: El Bosque.
 - 8: Cerro Navia.
 - 9: Cerrillos.
- Temperatura: Temperatura promedio del día en grados Celsius (°C).
- Fecha_d: Dia de la fecha de la medición en los datos de entrada.
- Fecha_m: Mes de la fecha de la medición en los datos de entrada.
- Fecha_y: Año de la fecha de la medición en los datos de entrada.
- Dia_futuro: Dia a futuro para predecir, toma valores del 1 al 30.
- Aquí_max_o: Índice de contaminación ambiental AQI EPA calculado en base a las entradas de contaminantes, su relevancia es el cálculo de la clase de salida de nivel de contaminación (nivel_o)
- Nivel_o: Nivel del índice de contaminación del aire representado como un índice de diccionario, el cual representa los siguientes valores:
 - 1: AQI EPA entre 0 y 50, calidad del aire Buena.
 - 2: AQI EPA entre 50 y 100, calidad del aire Moderada.
 - 3: AQI EPA entre 100 y 150, calidad del aire Insano para grupos sensibles.

- 4: AQI EPA entre 150 y 200, calidad del aire Insalubre.
- 5: AQI EPA entre 200 y 300, calidad del aire Muy Insalubre.
- 6: AQI EPA entre 300 y 400, calidad del aire Peligroso.
- 7: AQI EPA superior a 400, calidad del aire Peligroso.
- Aqi_max_actual: Representa el cálculo del índice AQI, y toma el valor más alto de todos los elementos considerados.
- Aqi_mean_actual: Representa el cálculo del índice AQI, y toma el valor promedio de todos los elementos considerados.

Dataset final luego de unir todos los datos:

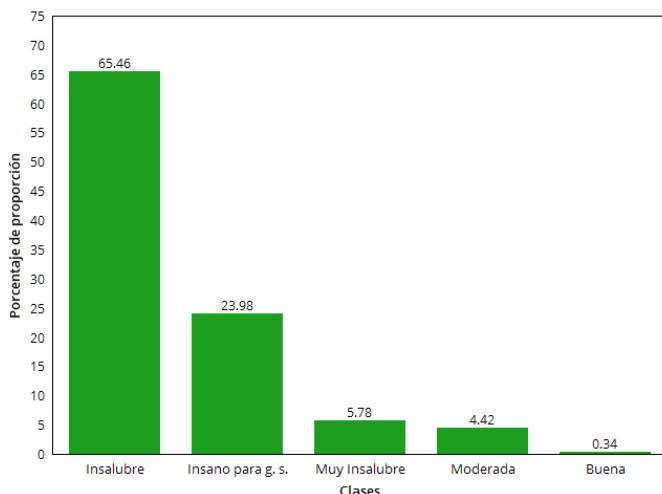
Ilustración 5.3: "Dataset final, sin output"													
	pm25	pm10	o3	no2	co	codigo_comuna	fecha_d	fecha_m	fecha_y	dia_futuro	aqi_max_actual	aqi_mean_actual	temperatura
0	0.366460	0.454545	0.440678	0.119048	0.111111	0.888889	0.033333	0.272727	0.000000	0.0	0.652974	0.381217	0.538812
1	0.509317	0.761364	0.305085	0.404762	0.444444	0.444444	0.900000	0.727273	0.285714	0.0	0.711708	0.571234	0.581531
2	0.559006	0.511364	0.728814	0.261905	0.222222	0.444444	0.366667	0.000000	0.000000	0.0	0.732137	0.511615	0.879637
3	0.211180	0.238636	0.101695	0.547619	0.166667	0.555556	0.433333	0.545455	0.142857	0.0	0.385089	0.255075	0.240204
4	0.490683	0.477273	0.610169	0.166667	0.444444	1.000000	0.800000	1.000000	0.000000	0.0	0.704047	0.554823	0.709945
...
493912	0.335404	0.488636	0.440678	0.476190	0.055556	0.444444	0.766667	0.818182	0.571429	1.0	0.631666	0.383777	0.572534
493913	0.465839	0.454545	0.305085	0.571429	0.478724	0.666667	0.966667	0.727273	0.142857	1.0	0.693833	0.559775	0.563950
493914	0.776398	0.636364	0.254237	0.309524	0.444444	0.000000	1.000000	0.545455	0.428571	1.0	0.821515	0.592022	0.279834
493915	0.354037	0.454545	0.576271	0.190476	0.333333	0.666667	0.933333	1.000000	0.714286	1.0	0.647867	0.485327	0.821961
493916	0.925466	0.522727	0.203390	0.571429	0.666667	0.000000	0.500000	0.454545	0.000000	1.0	0.882803	0.751893	0.387155

Fuente: Creación propia utilizando Google Colab.

5.4. Generación del modelo

Antes de entrenar el modelo, primero se debe ajustar cada hiperparametro de este con el objetivo de encontrar la mejor combinación posible, la cual, permitirá optimizar el modelo a la vez que maximizará la precisión de este. El primer hiperparametro a ajustar, se trata de la importancia de las clases presentes en los datos de entrenamiento, ya que estas no están equilibradas en cantidad y esto puede generar una mala interpretación de los datos por parte del modelo, a continuación, un gráfico que muestra lo anteriormente mencionado:

Gráfico 5.7: “Importancia de clases en los datos”



Fuente: Creación propia.

La situación que se muestra en el Grafico 5.7, enseña claramente como los datos presentan un número elevado de resultados en las clases de “Insalubre” e “Insano para grupos sensibles”, esto puede perjudicar las predicciones del modelo ya que se obtendrá una alta precisión en las clases más importantes, y una mala precisión en aquellas con una importancia menor. Para solucionar esta situación, se ajustó manualmente el hiperparametro de “class_weight”, dejando los pesos de las clases de la siguiente manera:

Ilustración 5.4: “Configuración del modelo”

```
modelos.append(RandomForestClassifier(  
    n_estimators=100,  
    min_samples_leaf=2,  
    min_samples_split=2,  
    max_depth=None,  
    class_weight ={  
        1: 600000000,  
        2: 10,  
        3: 0.002,  
        4: 0.001,  
        5: 35000  
    },  
    n_jobs = -1))
```

Fuente: Creación propia utilizando Google Colab.

Para llegar a la configuración que se muestra en la Ilustración 5.4, hicieron falta más de 30 intentos distintos, con el fin de encontrar la mejor distribución de pesos posibles por clases. Los números representan las clases de la siguiente manera: 1=“Buena”, 2=“Moderada”, 3=“Insano para g. s.”, 4=“insalubre” y

5="Muy Insalubre". Este ajuste en los pesos, en conjunto con los demás hiperparámetros que se muestran en la Ilustración 5.4, permitieron obtener los resultados más altos de todas las pruebas realizadas.

Una vez configurado el modelo, se procede a entrenar, a continuación, las métricas obtenidas de la ejecución del código:

Ilustración 5.5: "Ejecución de código: Generación del modelo y entrenamiento."

```
modelos = []
esti = 100
modelos.append(RandomForestClassifier(n_estimators=esti, min_samples_leaf=2, min_samples_split=2))
#Se entrena el modelo con los datos de entrenamiento
modelos[0].fit(dfs_X_train_final, dfs_Y_train_final)
estimator = modelos[0].estimators_[esti-1]
#Metricas del Modelo
predecido = modelos[0].predict(dfs_X_test_final)
confu = confusion_matrix(dfs_Y_test_final, predecido)
precisionRf_modelo = precision_score(dfs_Y_test_final, predecido, average='micro')
print("\n-----Resultados Modelo 1 ----- \n")
print("Precisión del Modelo: ", precisionRf_modelo)
print("Exactitud del Modelo: ", accuracy_score(dfs_Y_test_final, predecido))
print("Sensibilidad del Modelo: ", recall_score(dfs_Y_test_final, predecido, average='micro'))
print("Mean Absolute Error: ", mean_absolute_error(dfs_Y_test_final, predecido))
print("Mean Squared Error: ", mean_squared_error(dfs_Y_test_final, predecido))
print("Root Mean Squared Error: ", np.sqrt(mean_squared_error(dfs_Y_test_final, predecido)))
print("Puntaje F1 del Modelo: ", f1_score(dfs_Y_test_final, predecido, average='micro'))
print("Matriz de confusión:\n", confu)

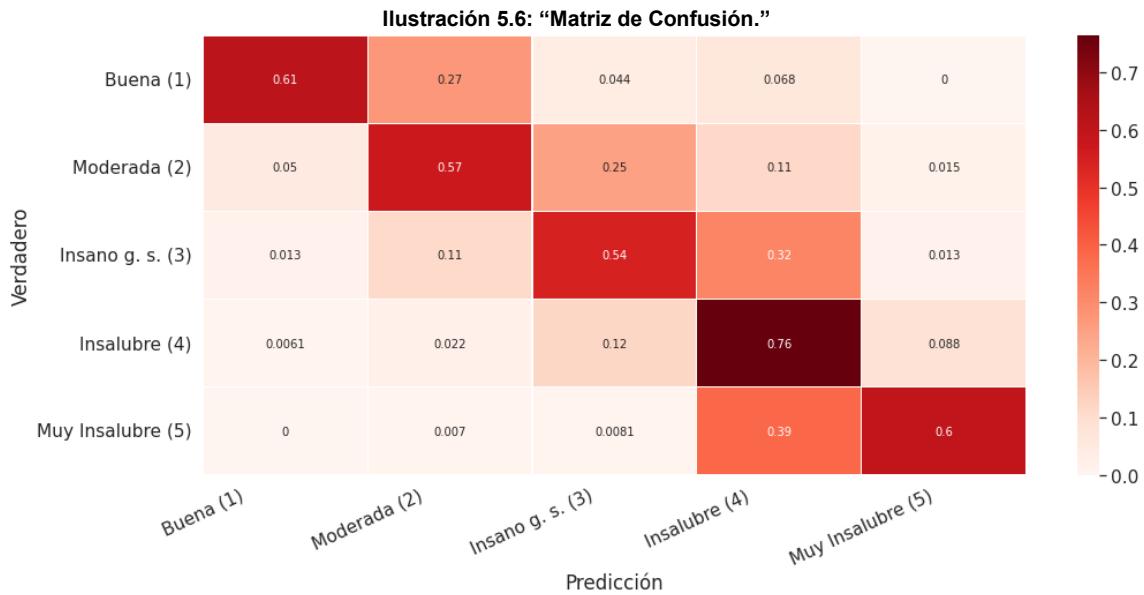
-----Resultados Modelo 1 -----
Precisión del Modelo: 0.6996602532957601
Exactitud del Modelo: 0.6996602532957601
Sensibilidad del Modelo: 0.6996602532957601
Mean Absolute Error: 0.34477537006445763
Mean Squared Error: 0.43035176367699934
Root Mean Squared Error: 0.6560120148876843
Puntaje F1 del Modelo: 0.6996602532957601
Matriz de confusión:
[[ 263  116   21   27    0]
 [ 300  3149 1395   609   86]
 [ 479  3301 15886  9210  417]
 [ 684  1776  9674 61663  7305]
 [   0    56    58  2682  4335]]
```

Fuente: Creación propia utilizando Google Colab.

Tras la ejecución del código que se muestra en la Ilustración 5.5, se obtuvo una precisión general del modelo de 70% aproximadamente, lo que es bastante bueno si se considera que se está intentando predecir la calidad del aire de los próximos 30 días, y que, además, la data disponible en Chile sobre la presencia de contaminantes en el medio ambiente (concretamente la del aire) se encuentra muy incompleta y dañada. Los Dataset de información medioambiental que maneja el SINCA o el gobierno de Chile son completados por diversas estaciones meteorológicas y estaciones de monitoreo del aire, las cuales, en ocasiones, se encuentran inactivas, por lo que hay periodos de tiempo en los que hay información que no queda registrada y se pierde. En el estudio

de los elementos que pueden influir en la contaminación del aire se mencionaron diversas variables a considerar, pero a la hora de incluirlas en el modelo final, se debió sopesar el estado de la información disponible, por lo que algunos elementos importantes debieron descartarse, por falta de información.

En la Ilustración 5.5, se puede ver una estructura básica de la matriz de confusión del modelo, a continuación, se presenta la matriz con más detalle:

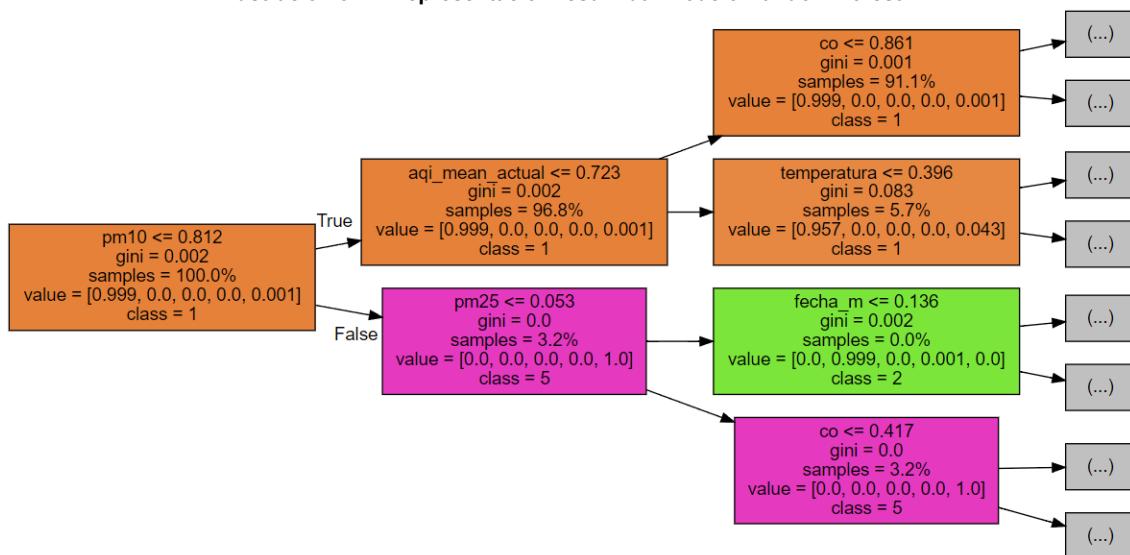


Fuente: Creación propia utilizando Google Colab.

En la matriz mostrada en la Ilustración 5.6, se puede ver como el modelo puede predecir todas las clases con una precisión superior al 50%, siendo la más elevada de ellas, la predicción de la clase “Insalubre”, llegando a un 76%. Además, se puede observar cómo los ajustes a los pesos realizados anteriormente (ver Ilustración 5.3 y 5.4) lograron nivelar la clasificación a aquellas clases correspondientes o cercanas al resultado, reduciendo así, cualquier fallo grave en la predicción del modelo.

A continuación, se muestra una representación gráfica resumida del modelo predictivo basado en un algoritmo de Random Forest:

Ilustración 5.7: “Representación resumida: Modelo Random Forest.”

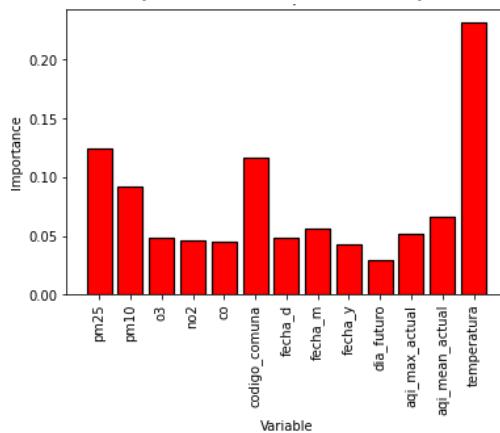


Fuente: Creación propia utilizando Google Colab.

En la Ilustración 5.7, se pueden ver los primeros 2 niveles de uno de los “estimator” o árboles de decisión que componen el Random Forest (está compuesto por 100 árboles), en él, se observa de manera muy resumida cómo funcionan los criterios. Inicialmente, el árbol es “activado” si los parámetros de entrada cumplen con la condición, la cual en este caso es que “pm10<=0,812 and samples <= 2”, luego, y en base a este mismo criterio y si se cumple o no, se sigue dividiendo el árbol y vuelve a repetir el mismo principio, pero con distintos parámetros. Un nodo se seguirá dividiendo a medida que no se cumpla la condición mínima de “samples”.

En la Ilustración 5.7 se muestra un arreglo bajo el nombre “value” en cada una de las casillas, esta variable corresponde a los pesos de cada clase a la hora de evaluarse en cada nodo, y en base a la relevancia de estos pesos, es donde se inclina el modelo para entregar su resultado. De igual forma, y en conjunto a lo anterior, otro parámetro que se debe considerar en la formación del modelo, es la relevancia de las variables de entrada o variables independientes. A continuación, un gráfico que muestra cuales son las variables más relevantes para el modelo:

Gráfico 5.8: “Importancia de variables independientes”



Fuente: Creación propia.

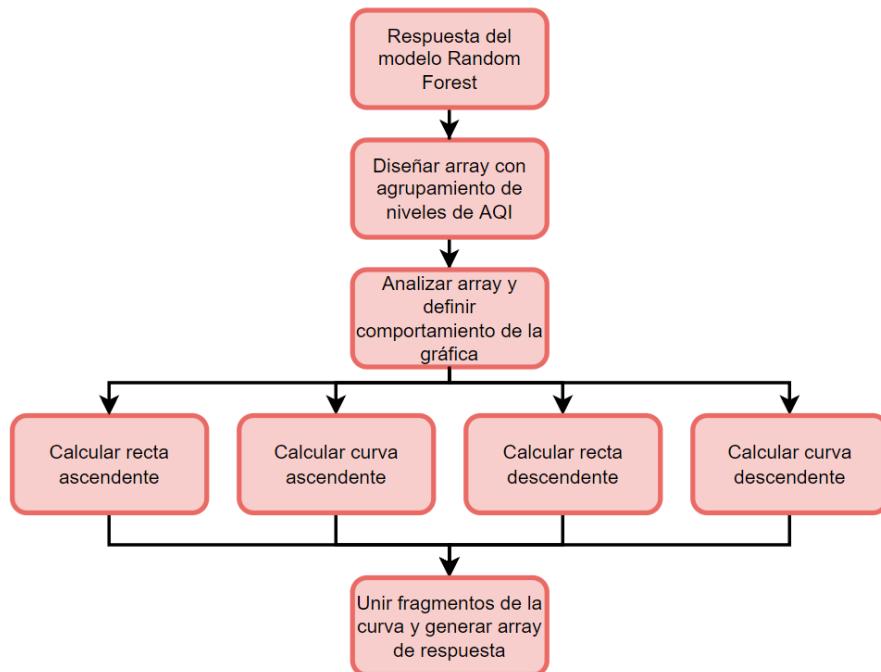
Como se puede observar en el Gráfico 5.8, la variable más relevante para el modelo es la temperatura promedio del día, seguido por el valor del pm25, el código de la comuna y pm10 respectivamente. Esto le otorga un sentido al objetivo del modelo, que era cumplir con distintas predicciones para las comunas de Santiago que se encuentren bajo estudio, además, los elementos más presentes en el aire son los materiales particulados y se demuestra su gran influencia en la calidad del aire. Sin embargo, la variable más importante, que es la temperatura, responde a los parámetros estudiados en el Capítulo 2, puntos 2.2.5.3. Temperatura atmosférica y 2.2.5.4. Ecuación de estado, donde su relevancia aplica para los parámetros de presión, por tanto, la concentración de partículas en el aire.

Inicialmente, el problema al que debe responder el modelo es al de predecir el nivel de contaminación del aire, lo cual, técnicamente, ya cumple, pero hace falta añadir un elemento de precisión para obtener una simulación gráfica del cambio exacto del índice AQI y cómo se comporta en los 30 días de predicción.

5.4.1. Algoritmo complementario

Para complementar y visualizar de mejor manera los resultados entregados por el algoritmo de Random Forest, se añade un tratamiento adicional a las respuestas de este, el cual consiste en un algoritmo estimador de la curva del AQI a lo largo de los días a predecir. A continuación, una representación gráfica resumida de su funcionamiento:

Ilustración 5.8: “Resumen, funcionamiento de algoritmo complementario”

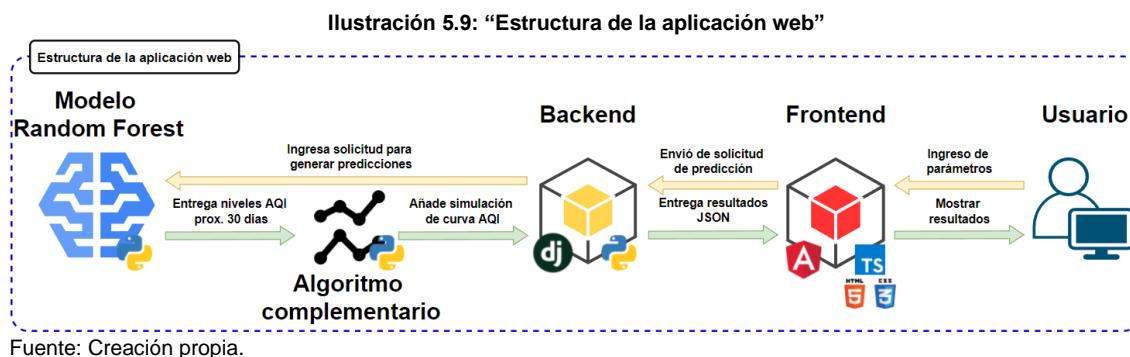


Fuente: Creación propia.

Es gracias a este algoritmo que el modelo es capaz de generar datos para realizar una gráfica más exacta del comportamiento de la curva del índice AQI a lo largo de los días predichos. Para saber más sobre el algoritmo, vea el Anexo 8 del presente informe.

5.5. Diseño de la aplicación web

Inicialmente, para construir la aplicación web donde se mostrarán los resultados del modelo predictivo, es necesario especificar que estructura tendrá y que herramientas se utilizarán para su elaboración. A continuación, el diseño de la de su estructura:



Fuente: Creación propia.

En base a lo mostrado en la Ilustración 5.9, el usuario ingresa los parámetros del día inicial para generar predicciones, luego de hacerlo, la aplicación enviará dichos ingresos a un backend, el cual, luego de realizar las validaciones correspondientes, cargará la información en el Modelo Random Forest. El modelo ejecutará 30 iteraciones para la predicción de los próximos 30 días del nivel categórico del AQI en la comuna correspondiente, luego, estos datos son cargados en el Algoritmo complementario, el cual genera la simulación de la curva del AQI. Los resultados tanto del modelo como del algoritmo son añadidos a un diccionario JSON el cuales son entregados como respuesta al Frontend para que este finalmente los muestre al usuario. Las herramientas utilizadas para la elaboración de la estructura son las siguientes:

- Modelo Random Forest: Librería Sklearn, utilizando el algoritmo RandomForestClassifier; lenguajes de programación: Python.
- Backend: Framework Django; lenguajes de programación: Python.
- Frontend: Framework Angular; lenguajes de programación: Typescript; otros: html, css.

5.5.1. Mockup de la aplicación web

Para la elaboración de la interfaz de la aplicación web, se diseñó de forma inicial en un mockup, el cual se muestra a continuación:

Ilustración 5.10: “Mockup de la aplicación web”

Modelo Predictivo de la contaminación del aire

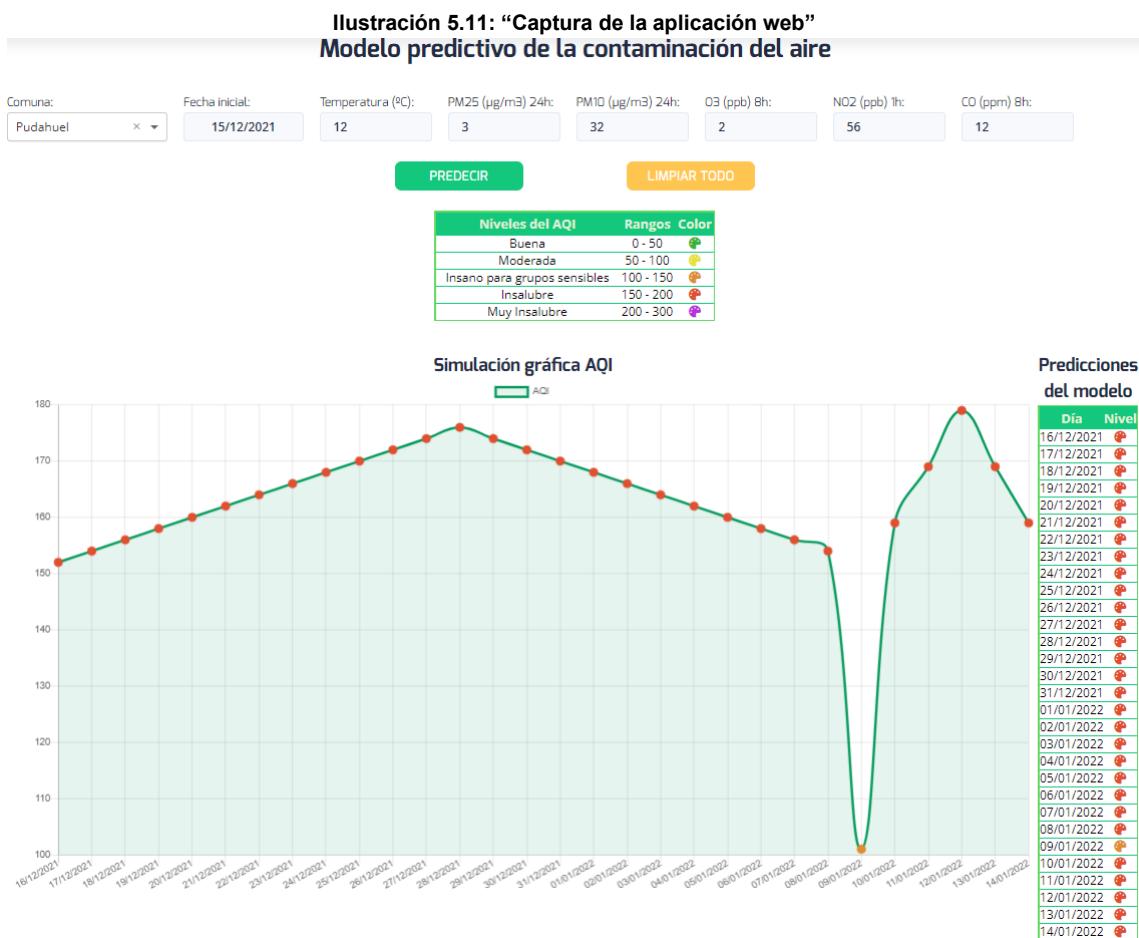
The mockup shows a form titled "Modelo Predictivo de la contaminación del aire". At the top, there are input fields for "Comuna" (dropdown), "Fecha Inicial" (date input), "Temperatura" (text input), and five pollutant inputs ("PM25", "PM10", "O3", "NO2", "CO") each with an "Ingresar" button. Below these are two buttons: "PREDECIR" (green) and "Limpiar Todo" (yellow). A large central area is labeled "ESPACIO PARA TABLA LEYENDA". To the left, a box is labeled "Simulación gráfica AQI" and to the right, "Predicciones del modelo". At the bottom left, a large box is labeled "ESPACIO PARA GRÁFICA". The entire interface is enclosed in a dashed border.

Fuente: Creación propia.

En la Ilustración 5.10 se muestra de forma intuitiva lo que debería ser el resultado final del proyecto, al menos en lo que al apartado visual se refiere. Se cuenta con un selector de comunas el cual contempla todas las opciones disponibles, luego se debe seleccionar la fecha adecuada para el día de inicio de las predicciones y finalmente, se ingresan todos los datos solicitados de contaminantes o elementos en el aire.

Existen dos opciones, “Predecir” y “Limpiar todo”, la primera de ellas efectúa todo el proceso mostrado en la Ilustración 5.9, y el segundo, limpia todos

los datos ingresados y la gráfica. A continuación, se muestra el resultado final luego de toda la planificación del diseño:



Fuente: Creación propia, captura de la aplicación web.

Como se puede ver en la Ilustración 5.11 y 5.10, el diseño de la aplicación web se completó a la perfección, y permite visualizar los resultados del modelo de forma clara y simple. El proyecto completo se encuentra en el siguiente repositorio: https://github.com/GonzaloGSC/modelo_cont_aire

Luego de finalizar la construcción del modelo predictivo y su aplicación web, se comienza a evaluar su comportamiento en casos reales que no se incluyeron en las etapas de entrenamiento y test.

5.6. Resultados y análisis

El modelo predictivo, una vez terminado y montado dentro de un proyecto de aplicación web, está preparado para realizar múltiples pruebas de rendimiento para las cuales se utilizarán casos reales de parámetros presentados en un día escogido al azar. A continuación, la primera predicción realizada, sobre la comuna de Santiago:

Tabla 5.1: “Ingresos al modelo predictivo, comuna de Santiago”

Fecha	Temperatura (°C)	PM25 (µg/m3) 24h	PM10 (µg/m3) 24h	O3 (ppb) 8h	NO2 (ppb) 1h	CO (ppm) 8h
16/11/2021	29	32	44	39	13	4

Fuente: Creación propia.

Tabla 5.2: “Resultados del modelo predictivo, comuna de Santiago”

Fecha	Nivel AQI	Nivel AQI Predicho	Correcto	AQI	AQI Simulado	AQI Diferencia
17/11/2021	3	3	SI	117	146	29
18/11/2021	3	3	SI	124	143	19
19/11/2021	3	3	SI	149	140	9
20/11/2021	3	3	SI	132	136	4
21/11/2021	3	3	SI	144	133	11
22/11/2021	3	3	SI	137	130	7
23/11/2021	3	3	SI	137	127	10
24/11/2021	3	3	SI	142	123	19
25/11/2021	5	3	NO	291	127	164
26/11/2021	3	3	SI	129	130	1
27/11/2021	3	3	SI	144	133	11
28/11/2021	3	3	SI	105	136	31
29/11/2021	2	3	NO	97	140	43
30/11/2021	3	3	SI	110	143	33
01/12/2021	3	3	SI	102	146	44
02/12/2021	3	4	NO	110	166	56
03/12/2021	3	4	NO	129	182	53
04/12/2021	3	4	NO	137	166	29
05/12/2021	3	3	SI	134	141	7
06/12/2021	4	3	NO	152	133	19
07/12/2021	3	3	SI	147	125	22
08/12/2021	3	3	SI	137	117	20
09/12/2021	3	3	SI	132	125	7
10/12/2021	4	3	NO	151	133	18
11/12/2021	3	4	NO	144	166	22
12/12/2021	3	4	NO	112	182	70
13/12/2021	2	4	NO	68	166	98
14/12/2021	3	3	SI	105	133	28
15/12/2021	3	3	SI	122	117	5
16/12/2021	3	3	SI	124	133	9
Precisión			67%	Promedio Dif.		

Fuente: Creación propia.

Como resultado del entrenamiento del modelo, este logró una presión del 70% aproximadamente, por lo que las pruebas realizadas en la comuna de Santiago, las cuales se muestran en las Tablas 5.1 y 5.2, están solo un pequeño

porcentaje bajo esa medida, pero aun presentan un nivel de precisión aceptable, recordemos que se está intentando predecir 30 días a futuro. El modelo, en general, y como se puede ver en la Tabla 5.2, puede predecir de manera muy efectiva el nivel de contaminación más presente en los próximos 30 días, que en este caso es de código 3: “Insano para grupos sensibles”, sin embargo, a la hora de intentar predecir casos muy atípicos como lo es el del día 25/11/2021, el modelo presenta dificultades, ya que a partir de un nivel “3” en la calidad del aire en el día anterior, la calidad del aire empeora a nivel “5” (Muy Insalubre) de golpe. A continuación, la predicción realizada sobre la comuna de Pudahuel:

Tabla 5.3: “Ingresos al modelo predictivo, comuna de Pudahuel”

Fecha	Temperatura (°C)	PM25 ($\mu\text{g}/\text{m}^3$) 24h	PM10 ($\mu\text{g}/\text{m}^3$) 24h	O3 (ppb) 8h	NO2 (ppb) 1h	CO (ppm) 8h
16/11/2021	29	30	40	44	11	2

Fuente: Creación propia.

Tabla 5.4: “Resultados del modelo predictivo, comuna de Pudahuel”

Fecha	Nivel AQI	Nivel AQI Predicho	Correcto	AQI	AQI Simulado	AQI Diferencia
17/11/2021	3	3	SI	107	101	6
18/11/2021	3	3	SI	119	103	16
19/11/2021	3	3	SI	134	104	30
20/11/2021	3	3	SI	134	106	28
21/11/2021	3	3	SI	139	108	31
22/11/2021	3	3	SI	144	109	35
23/11/2021	3	3	SI	134	111	23
24/11/2021	3	3	SI	137	113	24
25/11/2021	3	3	SI	117	114	3
26/11/2021	3	3	SI	122	116	6
27/11/2021	3	3	SI	137	117	20
28/11/2021	2	3	NO	93	119	26
29/11/2021	3	3	SI	117	121	4
30/11/2021	3	3	SI	134	122	12
01/12/2021	2	3	NO	95	124	29
02/12/2021	2	3	NO	99	126	27
03/12/2021	3	3	SI	117	127	10
04/12/2021	3	3	SI	137	129	8
05/12/2021	3	3	SI	142	131	11
06/12/2021	4	3	NO	154	132	22
07/12/2021	3	3	SI	139	134	5
08/12/2021	3	3	SI	119	135	16
09/12/2021	3	3	SI	127	137	10
10/12/2021	4	3	NO	153	139	14
11/12/2021	3	3	SI	129	140	11
12/12/2021	2	3	NO	99	142	43
13/12/2021	2	3	NO	66	144	78
14/12/2021	3	3	SI	110	145	35
15/12/2021	3	3	SI	115	147	32
16/12/2021	3	3	SI	110	149	39

Precisión

77%

Promedio Dif.

22

Fuente: Creación propia.

En las Tablas 5.3 y 5.4, se muestran los parámetros de ingreso al modelo y sus resultados respectivamente, y es en este caso, donde se recalca el comportamiento del modelo predictivo en detectar el nivel de índice de contaminación más relevante en los próximos 30 días, dejando de lado totalmente cualquier dato atípico en la predicción, esto puede ser positivo para aumentar la precisión en los resultados ya que se logra llegar a un 77%, lo que es bastante elevado, pero, existen casos donde puede ser necesario conocer un comportamiento focalizado en ciertos días. A continuación, la predicción realizada sobre la comuna de El Bosque:

Tabla 5.5: “Ingresos al modelo predictivo, comuna de El Bosque”

Fecha	Temperatura (°C)	PM25 (µg/m3) 24h	PM10 (µg/m3) 24h	O3 (ppb) 8h	NO2 (ppb) 1h	CO (ppm) 8h
16/11/2021	29	40	46	33	18	5

Fuente: Creación propia.

Tabla 5.6: “Resultados del modelo predictivo, comuna de El Bosque”

Fecha	Nivel AQI	Nivel AQI Predicho	Correcto	AQI	AQI Simulado	AQI Diferencia
17/11/2021	3	3	SI	122	137	15
18/11/2021	3	3	SI	149	125	24
19/11/2021	4	3	NO	158	113	45
20/11/2021	4	3	NO	152	125	27
21/11/2021	4	4	SI	153	162	9
22/11/2021	4	4	SI	153	174	21
23/11/2021	4	4	SI	152	186	34
24/11/2021	4	4	SI	156	174	18
25/11/2021	3	3	SI	144	141	3
26/11/2021	4	3	NO	154	133	21
27/11/2021	4	3	NO	157	125	32
28/11/2021	3	3	SI	129	117	12
29/11/2021	3	3	SI	122	125	3
30/11/2021	4	3	NO	151	133	18
01/12/2021	3	4	NO	124	154	30
02/12/2021	3	4	NO	137	159	22
03/12/2021	4	4	SI	154	164	10
04/12/2021	4	4	SI	154	169	15
05/12/2021	4	4	SI	154	174	20
06/12/2021	4	4	SI	160	179	19
07/12/2021	4	4	SI	156	174	18
08/12/2021	3	4	NO	149	169	20
09/12/2021	3	4	NO	147	164	17
10/12/2021	4	4	SI	155	159	4
11/12/2021	4	3	NO	154	125	29
12/12/2021	3	3	SI	124	101	23
13/12/2021	2	4	NO	74	162	88
14/12/2021	3	4	NO	144	174	30
15/12/2021	3	4	NO	139	186	47
16/12/2021	4	4	SI	151	174	23

Precisión

57%

Promedio Dif.

23

Fuente: Creación propia.

Los datos reales expuestos en la Tabla 5.6 contrastan totalmente con los dos ejemplos anteriores, ya aquí se logra ver una mayor oscilación en la calidad del aire a lo largo de los días lo que también se puede traducir como inestabilidad en estos datos. El modelo logra una precisión a penas aceptable (superá el 50% de los casos), sin embargo, se observa como este intenta recrear una oscilación un tanto similar a la presentada en los datos reales, lo que, entre otras cosas, puede significar una falta de variables independientes en el entrenamiento del modelo para lograr esa precisión faltante para estos casos. A continuación, la predicción realizada sobre la comuna de La Florida:

Tabla 5.7: “Ingresos al modelo predictivo, comuna de La Florida”

Fecha	Temperatura (°C)	PM25 ($\mu\text{g}/\text{m}^3$) 24h	PM10 ($\mu\text{g}/\text{m}^3$) 24h	O3 (ppb) 8h	NO2 (ppb) 1h	CO (ppm) 8h
16/11/2021	29	32	49	40	17	3

Fuente: Creación propia.

Tabla 5.8: “Resultados del modelo predictivo, comuna de La Florida”

Fecha	Nivel AQI	Nivel AQI Predicho	Correcto	AQI	AQI Simulado	AQI Diferencia
17/11/2021	3	3	SI	127	101	26
18/11/2021	3	3	SI	132	103	29
19/11/2021	3	3	SI	139	104	35
20/11/2021	3	3	SI	137	106	31
21/11/2021	3	3	SI	134	108	26
22/11/2021	3	3	SI	144	109	35
23/11/2021	3	3	SI	144	111	33
24/11/2021	4	3	NO	154	113	41
25/11/2021	3	3	SI	139	114	25
26/11/2021	3	3	SI	134	116	18
27/11/2021	4	3	NO	151	117	34
28/11/2021	3	3	SI	105	119	14
29/11/2021	3	3	SI	117	121	4
30/11/2021	3	3	SI	122	122	0
01/12/2021	2	3	NO	87	124	37
02/12/2021	3	3	SI	117	126	9
03/12/2021	3	3	SI	134	127	7
04/12/2021	3	3	SI	149	129	20
05/12/2021	3	3	SI	149	131	18
06/12/2021	4	3	NO	153	132	21
07/12/2021	4	3	NO	154	134	20
08/12/2021	3	3	SI	147	135	12
09/12/2021	3	3	SI	142	137	5
10/12/2021	3	3	SI	147	139	8
11/12/2021	4	3	NO	153	140	13
12/12/2021	3	3	SI	115	142	27
13/12/2021	2	3	NO	76	144	68
14/12/2021	3	3	SI	105	145	40
15/12/2021	3	3	SI	119	147	28
16/12/2021	3	3	SI	122	149	27
Precisión			77%	Promedio Dif.		

Fuente: Creación propia.

Los resultados obtenidos en esta prueba que se muestran en la Tabla 5.8, son similares a los presentados anteriormente en la Tabla 5.4, donde el modelo detecta eficazmente cual es el nivel de contaminación más presente de los próximos 30 días, logrando nuevamente una precisión del 77% dejando de lado cualquier resultado atípico.

El modelo predictivo, como resultado general de las pruebas logró una precisión promedio del 69,5%, lo que es bastante coherente con los resultados del entrenamiento de este y se considera como una marca en la precisión aceptable. Por otro lado, afocándose en la simulación de la curva AQI, en promedio se obtuvo una diferencia con los datos reales de 24,75 puntos, esto representa una diferencia aproximadamente de medio nivel AQI, lo que es aceptable si se considera que el modelo responde a un problema de clasificación y que el encargado de la recreación de este parámetro es el “algoritmo complementario”.

En termino definitivos, el modelo predictivo funciona correctamente y es coherente con los resultados obtenidos al momento de su construcción, logrando resultados positivos con casos no contemplados anteriormente.

CAPITULO 6 - CONCLUSIÓN Y TRABAJOS FUTUROS

6.1. Conclusiones

En base a los objetivos específicos del proyecto:

- Realizar un estudio bibliométrico y altmetric que permitan relevar información de importancia respecto a la caracterización del modelo predictivo en cuestión y antecedentes de trabajos similares que tengan relación con la contaminación ambiental.
- Analizar los puntos críticos de la contaminación del aire en Santiago o del mundo que puedan tener relevancia para el modelo, para así, establecer las variables de entrada y salida del Dataset a utilizar.

Se puede afirmar que la estructuración del trabajo es un progreso clave en todo proyecto, herramientas como realización de cronogramas pueden ayudar a ello, pero lo verdaderamente importante es la definición correcta de las metodologías de trabajo, estableciendo un marco orientador y un marco operacional. Un marco orientador permite la estructuración general de un proyecto y de las ideas que lo componen incluso cuando se desconocen algunas etapas o fases de este, luego, estableciendo un marco operacional, se permite la estructuración del trabajo como tal, ordenando las partes de este con niveles de detalle aceptables. Por otra parte, en el desarrollo de investigaciones o proyectos que tengan que ver con esta actividad, que es la generación de un modelo predictivo, es de suma importancia obtener buenas referencias bibliográficas y científicas del tema principal que se esté abordando, ya que más adelante en el desarrollo de este tipo de proyectos, el hecho de haber realizado una investigación deficiente puede repercutir en un prototipo defectuoso. Un buen método para la realización de este tipo de estudios es la clasificación de los documentos en base a diversas métricas que puedan describir su relevancia en el área, en otras palabras, un estudio bibliométrico y altmetrics. En el caso del

presente trabajo, se realizó esta tarea con la utilización de varios softwares los cuales cumplían funciones claves dentro de este proceso (ver capítulo 2, Ilustración 2.4), y permitieron la automatización de tareas como la recolección de información, eliminación de datos incompletos, creación y poblamiento de una base de datos propia y finalmente, la creación de material de análisis como gráficos, tablas y diagramas.

La información que se logró relvar para la confección del marco teórico del presente trabajo fue de gran utilidad para la búsqueda de datos disponibles de las comunas en estudio, sin embargo, si bien se conocen las variables importantes o críticas para la formación del modelo, si no se cuenta con datos históricos de estas, su utilización no se puede llevar a cabo, un ejemplo de esta situación se encuentra documentada en el capítulo 5: Desarrollo del Modelo, punto 5.1.5. Análisis Outliers columna SO₂, donde se debió optar por la eliminación de una variable independiente o de entrada del modelo por falta de registros históricos, lo que da cuenta de una situación bastante problemática a la hora de realizar proyectos de esta índole con la información disponible actualmente en Chile.

En base a los objetivos específicos del proyecto:

- Realizar una limpieza de datos y crear un Dataset el cual permita entrenar el modelo.
- Entrenar el modelo de Random Forest y perfeccionarlo para lograr los niveles de precisión esperados en sus resultados.

La recolección de la información y su unión en un único Dataset representó una tarea ardua y compleja de realizar, principalmente debido a su fragmentación en múltiples Datasets, por cada comuna se debió extraer información respecto a contaminantes del aire y temperatura, es decir que para tener la información de una sola comuna se contaba con dos Datasets y para unirlos, se debió utilizar la fecha como primary key. Luego de la unión de toda la

información de las comunas en un único Dataset, se debieron generar salidas para todos los días a predecir, por lo que ese único Dataset se transformó en 30, los cuales, al unirse, finalmente generaron el Dataset que se utilizó para el entrenamiento del modelo predictivo.

El entrenamiento del modelo consistió en el ajuste de las configuraciones de los hiperparámetros de este de forma iterativa, para ajustarlos en base a los resultados obtenidos en cada entrenamiento, donde el hiperparametro más relevante fue el de los pesos de cada clase. Este hiperparametro permitió la nivelación de la precisión por cada clase, evitando lo máximo posible el sesgo del modelo para con la clase más presente en los datos de entrenamiento.

En base al objetivo específico del proyecto:

- Validar el modelo en test para realizar su integración a una aplicación web, la cual admita su utilización con datos reales.

La terminación del modelo predictivo correspondió a su integración en una aplicación web que permitiera su uso de forma dinámica, llamativa y sencilla. La realización de los componentes necesarios para la confección de la aplicación empezó por la decisión de las herramientas a utilizar para su desarrollo. Para el caso del backend, el componente más importante, se decidió la utilización del framework Django por dos motivos principales, su simplicidad y su lenguaje base, que es python3 al igual que las librerías utilizadas para la generación del modelo predictivo, esto permitió una fácil y rápida integración del modelo con el servicio web. Para el caso del frontend, se decidió la utilización del framework Angular, principalmente por su gran relevancia como uno de los frameworks más utilizado a día de hoy para la generación de proyectos frontend.

Terminar la integración de todos los componentes de la aplicación web, permitió la utilización rápida del modelo para generar predicciones de prueba con datos reales, lo que completa la validación del modelo, y, tal como se muestra en el Capítulo 5: Desarrollo del modelo, punto 5.6. Resultados y análisis, los

resultados de este fueron satisfactorios llegando a un promedio de 69,5% de precisión y se observó un comportamiento de la curva AQI bastante cercano a la realidad, aunque mejorable.

En base al objetivo general del proyecto:

El objetivo general del proyecto se completó correctamente, el cual exigía la creación de un modelo predictivo para las comunas de Cerrillos, Cerro Navia, El Bosque, Independencia, la Florida, Las Condes, Santiago, Pudahuel, Puente Alto y Talagante, obtenido resultados distintos para cada una de ellas en las predicciones, además, la integración del modelo a una aplicación web se llevó a cabo según lo planeado y sin percances, por lo que el resultado final del proyecto es exitoso.

A continuación, se hará énfasis en todas las mejoras que se pueden aplicar al proyecto para elevar aún más la precisión del modelo predictivo.

6.2. Trabajos futuros

Para efectuar mejoras a este proyecto en un futuro, se recomiendan los siguientes puntos:

- Un factor realmente influyente a la hora de la elección del algoritmo para el desarrollo del modelo predictivo fue el volumen y calidad de la información disponible, los cuales son bastante mejorables, por lo que, en un futuro, y cuando se dispongan de los datos necesarios, se recomienda trasladar la solución a un modelo de red neuronal orientado a un problema de regresión y no de clasificación, de esta forma, el componente de algoritmo complementario quedaría totalmente obsoleto y permitiría una generación de un modelo más compacto y simple en usabilidad.
- La utilización de un modelo basado en el algoritmo de Random Forest genera un archivo de salida de un peso superior a los 1.5

GB, por lo que, nuevamente, esto se solucionaría con la utilización de una red neuronal.

- Se puede implementar un sistema de guardado de predicciones en una BDD con facilidad en el Web service actual, de esta forma, se pueden generar vistas interesantes en la aplicación web, como, por ejemplo, comparación entre predicciones. Todas estas mejoras se pueden integrar fácilmente en el frontend, ya que se creó pensado en un sistema de menús.
- Integrar el consumo de un API en tiempo real, para generar predicciones automáticamente.

BIBLIOGRAFÍA

- [1] IQair (2021). *Ranking de ciudades de calidad del aire y contaminación.* Recuperado de <https://www.iqair.com/es/world-air-quality-ranking>
- [2] QuestionPro (2021). *33 tipos de investigación y sus características.* Recuperado de <https://bit.ly/36Wm8V0>
- [3] Adams D. (2021). *Exporting your data.* Harzing, Research in International Management Recuperado de <https://n9.cl/harzing>
- [4] Worldcat (2021). *Patnaik, Pradyot.* Publication Timeline and Most widely held works by Pradyot Patnaik. Recuperado de <http://worldcat.org/identities/lccn-n91118181/>
- [5] Wikipedia (2021). *Cynthia Rudin,* Información de la autora. Recuperado de https://en.wikipedia.org/wiki/Cynthia_Rudin
- [6] Días V. (2017). *La contaminación ambiental.* Educación Básica. UTC. Latacunga. 54 p. Recuperado de <http://repositorio.utc.edu.ec/handle/27000/4101>
- [7] ISP (2020). Contaminación Ambiental. Recuperado de <https://bit.ly/3eUn0O2>
- [8] Fundación Aquae (2021). *¿Qué es la contaminación ambiental?* Recuperado de <https://bit.ly/3iPWjLR>
- [9] Zarza L. (2021). *¿Qué es la contaminación del agua?* IAGUA RESPUESTAS » CALIDAD DEL AGUA. Recuperado de <https://www.iagua.es/respuestas/que-es-contaminacion-agua>
- [10] Iberdrola. (2021). *La contaminación del agua: cómo no poner en peligro nuestra fuente de vida.* Contaminación del agua. Recuperado de <https://www.iberdrola.com/sostenibilidad/contaminacion-del-agua>

- [11] National Geographic (2017). *Del océano al grifo, la contaminación del agua nos afecta a todos: Descubre cómo la contaminación del agua nos afecta a todos: de océanos y mares, al grifo de casa.* Recuperado de <https://www.nationalgeographic.es/medio-ambiente/2017/10/del-oceano-al-grifo-la-contaminacion-del-agua-nos-afecta-todos>
- [12] GreenPeace (2016). *Plásticos en los océanos Datos, comparativas e impactos.* Dossier de prensa. Recuperado de http://archivos.greenpeace.org/espana/Global/espana/2016/report/plasticos/plasticos_en_los_oceanos_LR.pdf
- [13] Gligo V. (2019) *Informe país estado del medioambiente en Chile 2018.* Universidad de Chile, Instituto de Asuntos Públicos (INAP), Centro de Análisis de Políticas Públicas (CAPP). Recuperado de <http://www.cr2.cl/wp-content/uploads/2019/12/Informe-pais-estado-del-medio-ambiente-en-chile-2018.pdf>
- [14] Estévez R. (2015). *Biodiversidad y los servicios ecosistémicos.* Ecoinformación. Recuperado de <https://www.ecoinformacion.com/2015/06/servicios-ecosistemicos/>
- [15] SAIC (2021). *Degradación de suelos.* Recuperado de <http://www.siac.gov.co/erosion>
- [16] SNIARN (2008). *Informe de la situación del medio ambiente en México.* Edición 2008. Recuperado de https://apps1.semarnat.gob.mx:8443/dgeia/informe_2008/03_suelos/cap3_1.html
- [17] MedlinePlus (2020). *Contaminación del aire.* Biblioteca Nacional de Medicina de los EE. UU. Recuperado de <https://medlineplus.gov/spanish/airpollution.html>
- [18] Fundación Aquae (2021). *Contaminación del aire: causas y tipos.* Recuperado de <https://www.fundacionaque.org/causas-y-tipos-de-la-contaminacion-del-aire/>
- [19] Universidad Católica de Chile (2001). *Contaminación atmosférica.* Recuperado de http://www7.uc.cl/sw_educ/contam/fratmosf.htm

- [20] Porta, Yanina y Colman (2018). *Calidad del aire: Monitoreo y modelado de contaminantes atmosféricos. Efectos en la salud pública*. ISBN: 978-950-34-1682-2. Editorial de la Universidad Nacional de La Plata
<http://sedici.unlp.edu.ar/handle/10915/73756>
- [21] Quirós L. (2010). *La Atmósfera: Un Sistema del Planeta Tierra*. Recuperado de <https://www.uv.mx/personal/tcarmona/files/2010/08/Leal.pdf>
- [22] LennTech (2021). *Tabla Periódica: Clasificación periódica de los elementos químicos*. Recuperado de <https://www.lenntech.es/periodica/tabla-periodica.htm>
- [23] Baird C. (2001). *Química Ambiental*. ISBN:9788429179026. Editorial Reverté S.A., España. Recuperado de <https://bit.ly/3i5ZhwF>
- [24] EPA (2021). *Efectos del material particulado (PM) sobre la salud y el medio ambiente*. Recuperado de <https://espanol.epa.gov/espanol/efectos-del-material-particulado-pm-sobre-la-salud-y-el-medioambiente>
- [25] SINCA (1998). *Norma de Calidad Primaria para MP10*. D.S. N.º 59/98 Ministerio Secretaría General de la Presidencia. Recuperado de <https://sinca.mma.gob.cl/uploads/documentos/73881f634e74a87884b626007d5e585f.pdf>
- [26] Hurwitz J. y Kirsch D. (2018). *Machine Learning For Dummies*. Recuperado de <https://www.ibm.com/downloads/cas/GB8ZMQZ3>
- [27] Consejo Minero (2019). *Monitoreo de agua en tiempo real*. Recuperado de <https://consejominero.cl/plataforma-social/monitoreo-de-agua-en-tiempo-real/>
- [28] Zambrano J. (2018). *¿Aprendizaje supervisado o no supervisado? Conoce sus diferencias dentro del machine learning y la automatización inteligente*. Recuperado de <https://medium.com/@juanzambrano/aprendizaje-supervisado-o-no-supervisado-39ccf1fd6e7b>

- [29] Recuerdo de los Santos P. (2017). *Tipos de aprendizaje en Machine Learning: supervisado y no supervisado*. Recuperado de <https://empresas.blogthinkbig.com/que-algoritmo-elegir-en-ml-aprendizaje/>
- [30] Sancho F. (2020). *Aprendizaje Supervisado y No Supervisado*. Recuperado de <http://www.cs.us.es/~fsancho/?e=77>
- [31] Johnson D. (2021). *Unsupervised Machine Learning: What is, Algorithms, Example*. Recuperado de <https://www.guru99.com/unsupervised-machine-learning.html>
- [32] Mbaabu O.(2020). *Introduction to Random Forest in Machine Learning*. Recuperado de <https://www.section.io/engineering-education/introduction-to-random-forest-in-machine-learning/>
- [33] TensorFlow (2021). *Plataforma de extremo a extremo de código abierto para el aprendizaje automático*. Recuperado de <https://www.tensorflow.org/?hl=es-419>
- [34] Keras (2021). *API DOCS*. Recuperado de <https://keras.io/>
- [35] NumPy (2021). *What is NumPy?*. Recuperado de <https://numpy.org/doc/stable/user/whatisnumpy.html>
- [36] Pandas (2021). *Getting started into Pandas*. Recuperado de https://pandas.pydata.org/docs/getting_started/index.html#intro-to-pandas
- [37] Scikit-learn Developers (2021). *sklearn.ensemble.RandomForestClassifier*, librería oficial. Recuperado de <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html#sklearn-ensemble-randomforestclassifier>
- [38] Tutorials Point (2021). *Scikit Learn - Introduction*. Recuperado de https://www.tutorialspoint.com/scikit_learn/scikit_learn_introduction.htm

ANEXOS

Anexo 1: Ranking de ciudades y contaminación

A continuación, se presenta el ranking de ciudades más contaminadas del mundo el 29 de abril del 2021:



12:00, Apr 29

Fuente: Imagen capturada del sitio www.iQAir.com

Anexo 2: Resultados de consultas a BDD POPv7

A continuación, se presentan los resultados obtenidos por POPv7 en la primera búsqueda general:

Ilustración 7.2: “Resultados consultas a bases de datos, búsqueda 1 idioma inglés”

Search terms	Source	Papers	Cites	Cites/year	h	g	hl,norm	hl,annual	hA	acc10	Search date	Cache date	Last result
✓ Environmental pollution Air qu...	Crossref	200	1069	267.25	16	31	10	2.50	12	14	03/05/2021	02/05/2021	0
✓ Environmental pollution Air po...	Crossref	200	586	146.50	12	22	6	1.50	8	5	03/05/2021	03/05/2021	0
✓ Environmental pollution Air po...	Crossref	200	579	144.75	12	22	6	1.50	8	5	03/05/2021	03/05/2021	0
✓ Environmental pollution PM10...	Crossref	200	1287	321.75	22	33	12	3.00	11	13	03/05/2021	03/05/2021	0
✓ Environmental pollution PM2,5...	Crossref	200	863	215.75	16	28	12	3.00	9	9	03/05/2021	03/05/2021	0
✓ Environmental pollution Air co...	Crossref	200	1223	305.75	19	34	12	3.00	12	16	03/05/2021	03/05/2021	0
✓ Environmental pollution Air co...	Crossref	200	357	89.25	8	18	5	1.25	6	3	03/05/2021	03/05/2021	0
✓ Environmental pollution Air qu...	Google Sch...	1000	53698	13424.50	111	177	54	13.50	62	536	03/05/2021	02/05/2021	0
✓ Environmental pollution Air po...	Google Sch...	996	76712	19178.00	123	212	61	15.25	69	757	03/05/2021	03/05/2021	0
✓ Environmental pollution Air po...	Google Sch...	999	46119	11529.75	90	137	43	10.75	52	603	03/05/2021	03/05/2021	0
✓ Environmental pollution PM10...	Google Sch...	995	21652	5413.00	64	95	29	7.25	36	292	03/05/2021	03/05/2021	0
✓ Environmental pollution PM2,5...	Google Sch...	999	31268	7817.00	74	115	31	7.75	39	440	03/05/2021	03/05/2021	0
✓ Environmental pollution Air co...	Google Sch...	996	94828	23707.00	138	233	70	17.50	82	855	03/05/2021	03/05/2021	0
✓ Environmental pollution Air co...	Google Sch...	1000	28393	7098.25	62	135	35	8.75	35	245	03/05/2021	03/05/2021	0
✓ Environmental pollution Air qu...	Microsoft A...	1000	21468	5367.00	67	106	31	7.75	36	202	03/05/2021	02/05/2021	0
✓ Environmental pollution Air po...	Microsoft A...	1000	17534	4383.50	62	93	26	6.50	32	164	03/05/2021	03/05/2021	0
✓ Environmental pollution Air po...	Microsoft A...	54	340	85.00	10	17	5	1.25	7	4	03/05/2021	03/05/2021	0
✓ Environmental pollution PM10...	Microsoft A...	19	173	43.25	5	13	3	0.75	5	1	03/05/2021	03/05/2021	0
✗ Environmental pollution PM2,5...	Microsoft A...	0	0	0.00	0	0	0	0.00	0	0	03/05/2021	03/05/2021	514
✓ Environmental pollution Air co...	Microsoft A...	0	0	0.00	0	0	0	0.00	0	0	03/05/2021	03/05/2021	514
✓ Environmental pollution Air co...	Microsoft A...	498	3684	921.00	27	51	14	3.50	17	34	03/05/2021	03/05/2021	0
✗ Environmental pollution Air co...	Microsoft A...	0	0	0.00	0	0	0	0.00	0	0	03/05/2021	03/05/2021	514
✓ Environmental pollution Air qu...	Scopus	200	30425	7606.25	100	168	100	25.00	57	200	03/05/2021	02/05/2021	0
✓ Environmental pollution Air po...	Scopus	200	30393	7598.25	108	166	108	27.00	55	200	03/05/2021	03/05/2021	0
✓ Environmental pollution Air po...	Scopus	200	5672	1418.00	36	54	36	9.00	19	68	03/05/2021	03/05/2021	0
✓ Environmental pollution PM10...	Scopus	49	428	107.00	11	20	11	2.75	7	4	03/05/2021	03/05/2021	0
✓ Environmental pollution PM2,5...	Scopus	121	1089	272.25	19	28	19	4.75	12	16	03/05/2021	03/05/2021	0
✓ Environmental pollution Air co...	Scopus	200	16231	4057.75	70	109	70	17.50	37	200	03/05/2021	03/05/2021	0
✓ Environmental pollution Air co...	Scopus	5	62	20.67	2	5	2	0.67	2	1	03/05/2021	03/05/2021	0

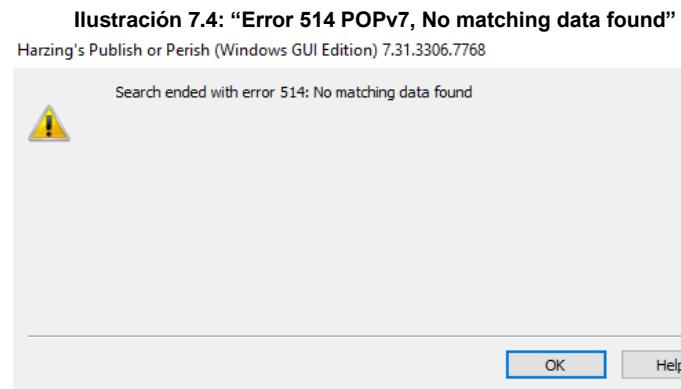
Fuente: Imagen capturada del software POPv7.

Ilustración 7.3: “Resultados consultas a bases de datos, búsqueda 1 idioma español”

Search terms	Source	Papers	Cites	Cites/year	h	g	hl,norm	hl,annual	hA	acc10	Search date	Cache date	Last result
✓ Contaminación Ambiental Cont...	Crossref	200	75	18.75	4	5	2	0.50	2	0	04/05/2021	04/05/2021	0
✓ Contaminación Ambiental Calid...	Crossref	200	38	9.50	3	4	2	0.50	2	0	04/05/2021	04/05/2021	0
✓ Contaminación Ambiental Cont...	Crossref	200	70	17.50	4	5	2	0.50	2	0	04/05/2021	04/05/2021	0
✓ Contaminación Ambiental PM10...	Crossref	200	81	20.25	5	6	3	0.75	2	0	04/05/2021	04/05/2021	0
✓ Contaminación Ambiental PM2,5...	Crossref	200	81	20.25	5	6	3	0.75	2	0	04/05/2021	04/05/2021	0
✓ Contaminación Ambiental Com...	Crossref	200	80	20.00	4	5	2	0.50	2	0	04/05/2021	04/05/2021	0
✓ Contaminación Ambiental Com...	Crossref	200	55	13.75	3	7	2	0.50	1	1	04/05/2021	04/05/2021	0
✓ Contaminación Ambiental Cont...	Google Scholar	974	4783	1195.75	20	62	16	4.00	12	15	04/05/2021	04/05/2021	0
✓ Contaminación Ambiental Calid...	Google Scholar	984	6925	1731.25	29	72	19	4.75	15	24	04/05/2021	04/05/2021	0
✓ Contaminación Ambiental Cont...	Google Scholar	981	2045	511.25	19	34	13	3.25	9	6	04/05/2021	04/05/2021	0
✓ Contaminación Ambiental PM10...	Google Scholar	999	746	82.89	12	20	8	0.89	6	2	04/05/2021	04/05/2021	0
✓ Contaminación Ambiental PM2,5...	Google Scholar	1000	593	65.89	11	18	7	0.78	6	2	04/05/2021	04/05/2021	0
✓ Contaminación Ambiental Com...	Google Scholar	984	5068	1267.00	22	63	15	3.75	11	14	04/05/2021	04/05/2021	0
✓ Contaminación Ambiental Com...	Google Scholar	998	367	91.75	6	13	5	1.25	4	1	04/05/2021	04/05/2021	0
✗ Contaminación Ambiental Com...	Microsoft Academic	0	0	0.00	0	0	0	0.00	0	0	04/05/2021	04/05/2021	514
✗ Contaminación Ambiental Com...	Microsoft Academic	0	0	0.00	0	0	0	0.00	0	0	04/05/2021	04/05/2021	514
✗ Contaminación Ambiental PM2,5...	Microsoft Academic	0	0	0.00	0	0	0	0.00	0	0	04/05/2021	04/05/2021	514
✗ Contaminación Ambiental PM10...	Microsoft Academic	0	0	0.00	0	0	0	0.00	0	0	04/05/2021	04/05/2021	514
✗ Contaminación Ambiental Cont...	Microsoft Academic	0	0	0.00	0	0	0	0.00	0	0	04/05/2021	04/05/2021	514
✗ Contaminación Ambiental Cont...	Microsoft Academic	0	0	0.00	0	0	0	0.00	0	0	04/05/2021	04/05/2021	514
✗ Contaminación Ambiental Calid...	Microsoft Academic	0	0	0.00	0	0	0	0.00	0	0	04/05/2021	04/05/2021	514
✓ Contaminación Ambiental Calid...	Scopus	1	4	2.00	1	1	1	0.50	1	0	04/05/2021	04/05/2021	0
✓ Contaminación Ambiental Cont...	Scopus	2	4	2.00	1	2	1	0.50	1	0	04/05/2021	04/05/2021	0
✗ Contaminación Ambiental Cont...	Scopus	0	0	0.00	0	0	0	0.00	0	0	04/05/2021	04/05/2021	514
✗ Contaminación Ambiental PM10...	Scopus	0	0	0.00	0	0	0	0.00	0	0	04/05/2021	04/05/2021	514
✗ Contaminación Ambiental PM2,5...	Scopus	0	0	0.00	0	0	0	0.00	0	0	04/05/2021	04/05/2021	514
✗ Contaminación Ambiental Com...	Scopus	0	0	0.00	0	0	0	0.00	0	0	04/05/2021	04/05/2021	514
✗ Contaminación Ambiental Com...	Scopus	0	0	0.00	0	0	0	0.00	0	0	04/05/2021	04/05/2021	514

Fuente: Imagen capturada del software POPv7.

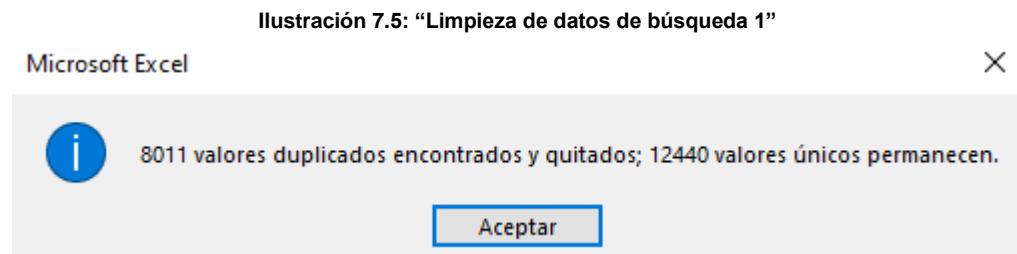
En las Ilustraciones 7.2 y 7.3, se puede ver como la base de datos Crossref siempre devuelve la cantidad máxima de resultados en todos los casos, sin embargo, existen ocurrencias de respuestas con un bajo nivel de resultados, hecho que se puede ver mayormente reflejado en el caso de Microsoft Academic, además, es en esta base de datos que se presenta con ambos idiomas un error de referencias no encontradas, el cual se muestra a continuación:



Fuente: Imagen capturada del software POPv7.

Finalmente, el total de resultados entregados por la búsqueda, sin depurar, es de 20.451 papers.

Debido a la existencia de resultados repetidos (mismos papers en distintas bases de datos), antes de continuar es necesario efectuar una limpieza de los datos para obtener el volumen de resultados final, para esto, existe una característica muy práctica y sencilla de utilizar que ofrece Excel denominada “Quitar duplicados”, a continuación, una imagen del resultado de su ejecución:



Fuente: Imagen capturada del programa Excel.

Luego de lo anterior, se puede afirmar que la primera búsqueda arrojó un volumen de resultados de 12.440 papers distintos.

A continuación, se presentan los resultados obtenidos por POPv7 en la primera búsqueda general:

Ilustración 7.6: “Resultados consultas a bases de datos, búsqueda 2 idioma inglés”

Search terms	Source	Papers	Cites	Cites/year	h	g	hl,norm	hl,annual	hA	acc10	Search date	Cache date	Last result
✓ Predictive model Air pollutants from 2017...	Crossref	200	194	48.50	6	12	5	1.25	4	0	05/05/2021	05/05/2021	0
✓ Predictive model Air pollutants PM10 fro...	Crossref	200	314	78.50	9	15	6	1.50	6	2	05/05/2021	05/05/2021	0
✓ Predictive model Air pollutants PM2,5 fro...	Crossref	200	192	48.00	6	12	5	1.25	4	0	05/05/2021	05/05/2021	0
✓ Predictive model Air quality from 2017 to ...	Crossref	200	391	97.75	9	15	5	1.25	6	2	05/05/2021	05/05/2021	0
✓ Predictive model Breathable air from 201...	Crossref	200	294	73.50	9	14	5	1.25	5	0	05/05/2021	05/05/2021	0
✓ Predictive model Environmental pollutio...	Crossref	200	790	197.50	16	26	11	2.75	8	5	05/05/2021	05/05/2021	0
✓ Predictive model Environmental pollutio...	Crossref	200	790	197.50	16	26	11	2.75	8	5	05/05/2021	05/05/2021	0
✓ Predictive model Air pollutants from 2017...	Google Sc...	1000	52...	13089.25	93	1...	43	10.75	49	655	05/05/2021	05/05/2021	0
✓ Predictive model Air pollutants PM10 fro...	Google Sc...	993	17...	4390.25	62	85	25	6.25	31	246	05/05/2021	05/05/2021	0
✓ Predictive model Air pollutants PM2,5 fro...	Google Sc...	996	20...	5164.25	65	96	26	6.50	32	275	05/05/2021	05/05/2021	0
✓ Predictive model Air quality from 2017 to ...	Google Sc...	987	34...	8596.50	87	1...	40	10.00	44	407	05/05/2021	05/05/2021	0
✓ Predictive model Breathable air from 201...	Google Sc...	998	17...	4402.25	50	88	25	6.25	31	169	05/05/2021	05/05/2021	0
✓ Predictive model Environmental pollutio...	Google Sc...	1000	71...	17908.50	1...	1...	65	16.25	62	824	05/05/2021	05/05/2021	0
✓ Predictive model Air pollutants from 2017...	Microsoft...	86	452	113.00	11	17	4	1.00	6	2	05/05/2021	05/05/2021	0
✓ Predictive model Air pollutants PM10 fro...	Microsoft...	14	101	25.25	6	10	3	0.75	4	0	05/05/2021	05/05/2021	0
✗ Predictive model Air pollutants PM2,5 fro...	Microsoft...	0	0	0.00	0	0	0	0.00	0	0	05/05/2021	05/05/2021	514
✓ Predictive model Air quality from 2017 to ...	Microsoft...	276	1757	439.25	19	33	9	2.25	13	18	05/05/2021	05/05/2021	0
✓ Predictive model Breathable air from 201...	Microsoft...	1	4	1.00	1	1	1	0.25	1	0	05/05/2021	05/05/2021	0
✓ Predictive model Environmental pollutio...	Microsoft...	130	720	180.00	15	22	6	1.50	10	10	05/05/2021	05/05/2021	0
✓ Predictive model Air pollutants from 2017...	Scopus	200	3102	775.50	25	41	25	6.25	15	30	05/05/2021	05/05/2021	0
✓ Predictive model Air pollutants PM10 fro...	Scopus	26	141	35.25	6	11	6	1.50	5	1	05/05/2021	05/05/2021	0
✓ Predictive model Air pollutants PM2,5 fro...	Scopus	56	700	175.00	13	25	13	3.25	7	6	05/05/2021	05/05/2021	0
✓ Predictive model Air quality from 2017 to ...	Scopus	200	3289	822.25	27	41	27	6.75	15	39	05/05/2021	05/05/2021	0
✓ Predictive model Breathable air from 201...	Scopus	2	10	2.50	2	2	2	0.50	1	0	05/05/2021	05/05/2021	0
✓ Predictive model Environmental pollutio...	Scopus	200	4376	1094.00	31	47	31	7.75	18	49	05/05/2021	05/05/2021	0

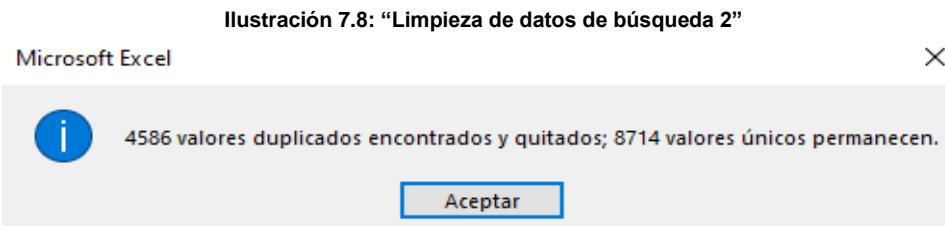
Fuente: Imagen capturada del software POPv7.

Ilustración 7.7: “Resultados consultas a bases de datos, búsqueda 2 idioma español”

Search terms	Source	Papers	Cites	Cites/year	h	g	hl,norm	hl,annual	hA	acc10	Search date	Cache date	Last result
✓ Modelo predictivo Aire respirable from 20...	Crossref	200	106	26.50	5	8	4	1.00	3	1	05/05/2021	05/05/2021	0
✓ Modelo predictivo Calidad del aire from 2...	Crossref	200	40	10.00	3	4	2	0.50	2	0	05/05/2021	05/05/2021	0
✓ Modelo predictivo Contaminación ambi...	Crossref	200	88	22.00	5	6	3	0.75	3	0	05/05/2021	05/05/2021	0
✓ Modelo predictivo Contaminantes del ari...	Crossref	200	53	13.25	4	4	2	0.50	2	0	05/05/2021	05/05/2021	0
✓ Modelo predictivo Contaminantes del ari...	Crossref	200	49	12.25	3	4	2	0.50	2	0	05/05/2021	05/05/2021	0
✓ Modelo predictivo Contaminantes del ari...	Crossref	200	53	13.25	4	4	2	0.50	2	0	05/05/2021	05/05/2021	0
✓ Modelo predictivo Aire respirable from 20...	Google Sc...	148	99	24.75	4	9	2	0.50	2	1	05/05/2021	05/05/2021	0
✓ Modelo predictivo Calidad del aire from 2...	Google Sc...	995	883	220.75	14	22	10	2.50	7	6	05/05/2021	05/05/2021	0
✓ Modelo predictivo Contaminación ambi...	Google Sc...	990	869	217.25	12	22	9	2.25	7	4	05/05/2021	05/05/2021	0
✓ Modelo predictivo Contaminantes del ari...	Google Sc...	999	906	226.50	14	25	11	2.75	9	8	05/05/2021	05/05/2021	0
✓ Modelo predictivo Contaminantes del ari...	Google Sc...	353	62	15.50	4	4	3	0.75	2	0	05/05/2021	05/05/2021	0
✓ Modelo predictivo Contaminantes del ari...	Google Sc...	261	48	12.00	4	4	3	0.75	2	0	05/05/2021	05/05/2021	0
✗ Modelo predictivo Aire respirable from 20...	Microsoft...	0	0	0.00	0	0	0	0.00	0	0	05/05/2021	n/a	514
✗ Modelo predictivo Calidad del aire from 2...	Microsoft...	0	0	0.00	0	0	0	0.00	0	0	05/05/2021	n/a	514
✓ Modelo predictivo Contaminacion ambi...	Microsoft...	1	0	0.00	0	0	0	0.00	0	0	05/05/2021	05/05/2021	0
✗ Modelo predictivo Contaminantes del ari...	Microsoft...	0	0	0.00	0	0	0	0.00	0	0	05/05/2021	n/a	514
✗ Modelo predictivo Contaminantes del ari...	Microsoft...	0	0	0.00	0	0	0	0.00	0	0	05/05/2021	n/a	514
✗ Modelo predictivo Contaminantes del ari...	Microsoft...	0	0	0.00	0	0	0	0.00	0	0	05/05/2021	n/a	514
✗ Modelo predictivo Contaminantes del ari...	Microsoft...	0	0	0.00	0	0	0	0.00	0	0	05/05/2021	n/a	514
✗ Modelo predictivo Contaminacion ambi...	Scopus	0	0	0.00	0	0	0	0.00	0	0	05/05/2021	n/a	514
✗ Modelo predictivo Calidad del aire from 2...	Scopus	0	0	0.00	0	0	0	0.00	0	0	05/05/2021	n/a	514
✗ Modelo predictivo Contaminacion ambi...	Scopus	0	0	0.00	0	0	0	0.00	0	0	05/05/2021	n/a	514
✗ Modelo predictivo Contaminantes del ari...	Scopus	0	0	0.00	0	0	0	0.00	0	0	05/05/2021	n/a	514
✗ Modelo predictivo Contaminantes del ari...	Scopus	0	0	0.00	0	0	0	0.00	0	0	05/05/2021	n/a	514
✗ Modelo predictivo Contaminantes del ari...	Scopus	0	0	0.00	0	0	0	0.00	0	0	05/05/2021	n/a	514
✗ Modelo predictivo Contaminantes del ari...	Scopus	0	0	0.00	0	0	0	0.00	0	0	05/05/2021	n/a	514

Fuente: Imagen capturada del software POPv7.

Los resultados que se muestran en las ilustraciones 7.6 y 7.7 fueron depurados y tratados de la misma manera que los resultados de la primera búsqueda. Los valores de respuesta totales, sin aplicar la limpieza de duplicados, es de 13.300 papers. La siguiente Ilustración muestra cual es el resultado después de la limpieza de estos:

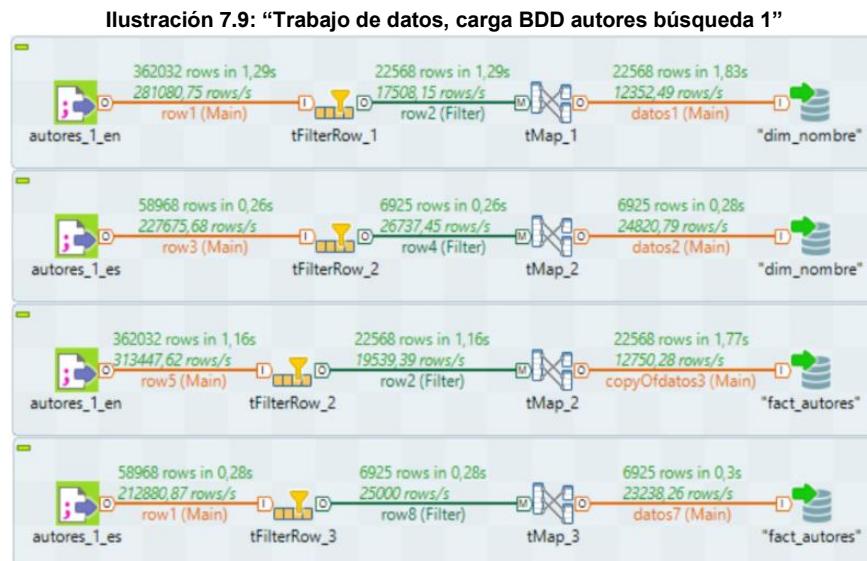


Fuente: Imagen capturada del programa Excel.

En base a la Ilustración 7.8, se puede afirmar que la segunda búsqueda arrojó un volumen de resultados de 8.714 papers distintos.

Anexo 3: Depuración y poblamiento de BDD

A continuación, se presentan las imágenes que muestran el trabajo de transformación, limpieza y carga de datos que se realizó a las BDD, las cuales fueron utilizadas para realizar el estudio bibliométrico:



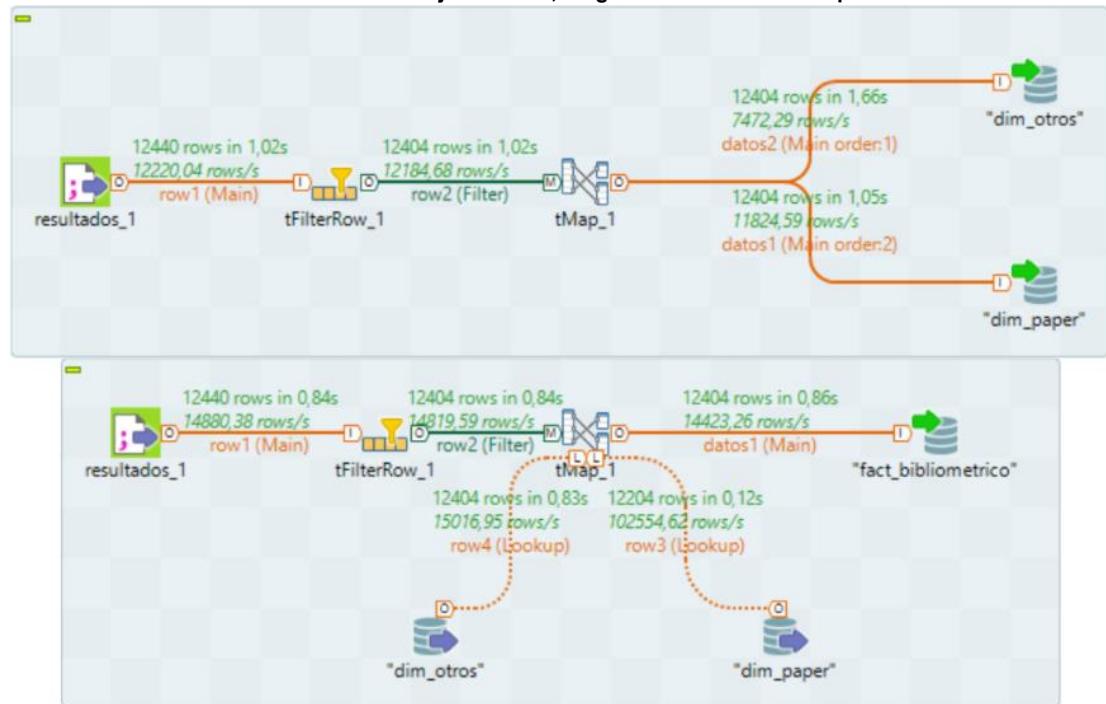
Fuente: Imagen capturada del software Talend Open Studio.

Ilustración 7.10: “Trabajo de datos, carga BDD autores búsqueda 2”



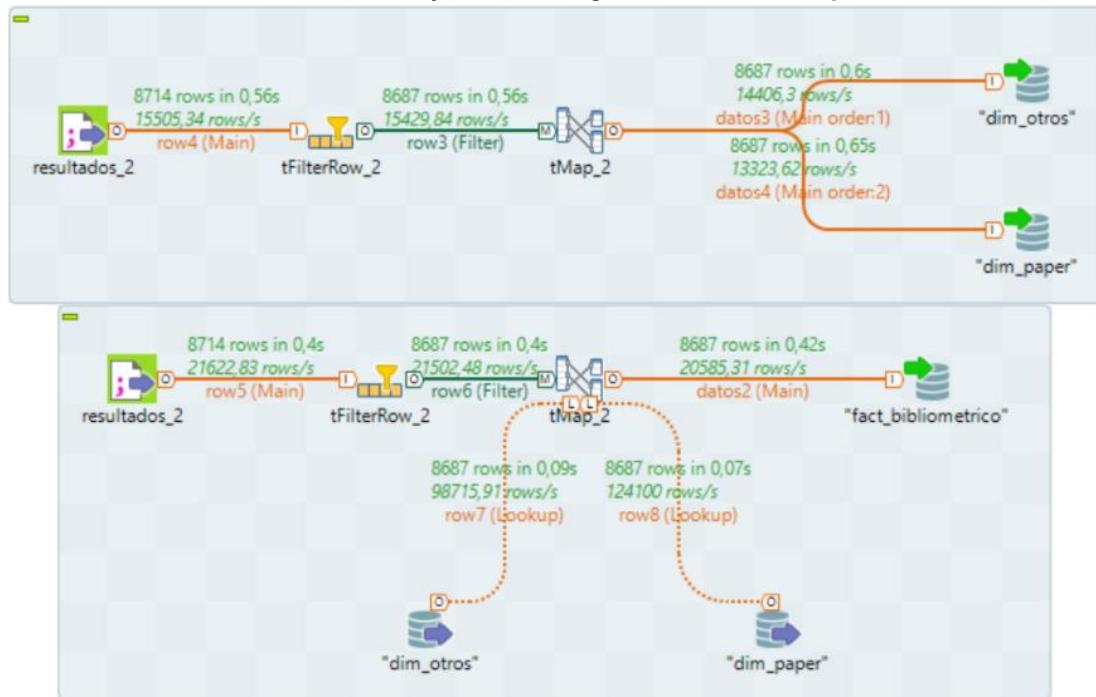
Fuente: Imagen capturada del software Talend Open Studio.

Ilustración 7.11: “Trabajo de datos, carga BDD resultados búsqueda 1”



Fuente: Imagen capturada del software Talend Open Studio.

Ilustración 7.12: “Trabajo de datos, carga BDD resultados búsqueda 2”



Fuente: Imagen capturada del software Talend Open Studio

Anexo 4: Cálculo de campo “Total” en relevancia de publicaciones

Uno de los factores que se debe considerar a la hora de analizar una publicación, es su número de citas, el cual se presenta, en este caso, en dos columnas de datos, una de ellas es “Cites” y la otra “ECC”. Ambas representan lo mismo, pero en algunos casos solo está presente el valor de una de ellas, por tanto, se debe programar la elección de este valor en base a su presencia y esta tarea es sencilla de hacer en Tableau. El primer parámetro entonces, para la creación del valor “Total” es:

$$\text{Citas} = \begin{cases} \text{Si existe Cites} \rightarrow (\text{Cites}) \\ \text{Si no} \rightarrow (\text{ECC}) \end{cases}$$

Luego, se debe considerar el orden de aparición de la publicación en los resultados de la consulta a la base de datos, dicho dato se representa en el parámetro GSrank, pero, existe un problema, ya que mientras menor sea este número, significa que mejor es su valor (relación inversa), por tanto, para poder operarlo con los demás datos (Cites o ECC) se debe invertir:

$$GSrank \text{ ajustado} = \frac{\text{Nº máximo de citas en columna}}{GSrank}$$

Si bien, para invertir un número de posición en base a un ranking, se debe restar ese número al total de datos en dicho ranking y luego sumarle 1, para este caso se decidió utilizar una división por un motivo en particular, la disminución de la influencia del parámetro. Si bien el orden de los resultados es una forma importante de analizar los datos, no debe tener una mayor importancia que las citas, es por esto que al dividir se cumple la función de “invertir” y además al hacerlo por la cantidad más alta de citas nos aseguramos de que este parámetro no tenga más influencia que estas. Finalmente, se tienen todos los parámetros necesarios para calcular el Total de relevancia, el cual está dado por la siguiente fórmula:

$$Total = \frac{2 * \text{Citas} + \text{Gsrkajustado}}{2}$$

Básicamente se trata de un promedio, con la diferencia que es aquí donde se le agrega el verdadero peso de relevancia a las citas, multiplicándose por 2, así, siempre tendrá mayor influencia que la posición del ranking. A continuación, un ejemplo práctico a modo de demostración de la teoría antes explicada:

Publicación 1 → N.º 10 en resultados -> GSrank = 10; citas = 10

Publicación 2 → N.º 125 en resultados -> GSrank = 125; citas = 200

Por tanto:

$$GSrank \text{ ajustado publicación 1} = \frac{200}{10} = 20$$

$$GSrank \text{ ajustado publicación 2} = \frac{200}{125} = 1,6 \approx 1$$

Finalmente:

$$Total \text{ publicacion 1} = \frac{2 * 10 + 20}{2} = 20$$

$$Total \text{ publicacion 2} = \frac{2 * 200 + 2}{1} = 402$$

Por lo que la publicación 2, a pesar de estar en el puesto N.º 125, contará con una mayor relevancia gracias a su gran cantidad de citas. Finalmente, otro ejemplo, pero esta vez con una situación final distinta:

Publicación 1 → N.º 1 en resultados -> GSrank = 1; citas = 105

Publicación 2 → N.º 200 en resultados -> GSrank = 200; citas = 200

Por tanto:

$$GSrank \text{ ajustado publicación 1} = \frac{200}{1} = 200$$

$$GSrank \text{ ajustado publicación 2} = \frac{200}{200} = 1$$

Finalmente:

$$Total\ publicacion\ 1 = \frac{2 * 105 + 200}{2} = 205$$

$$Total\ publicacion\ 2 = \frac{2 * 200 + 1}{2} = 200,5$$

En este caso, la publicación 1 cuenta con menos citas pero que tiene un mejor “GSrank ajustado” (se encontró antes en la búsqueda) lo que hace que tenga un “Total” más alto, en otras palabras, una relevancia más alta. Sin embargo, se puede observar como la influencia del GSrank es muy reducida en comparación con las citas (diferencia en resultados finales muy pequeña) pero lo suficientemente influyente como para modificar los resultados, todo esto es realmente útil para eliminar documentos que puedan estar muy citados pero que no tengan mucho que ver con el tema que se está buscando.

Anexo 5: Publicaciones eliminadas de resultados

Para ofrecer un listado más completo y acertado de publicaciones relevantes, se eliminaron aquellas que estaban enfocadas en otros ámbitos o tenían poca relación con lo que se está investigando. Si bien se utilizaron diversas métricas para filtrar los datos, siempre existirán resultados que no sean acertados dentro de una búsqueda bibliográfica de estas características, por lo que se debe limpiar el listado manualmente. A continuación, las publicaciones eliminadas con medida “Total” más alta (ver Anexo 4 para entender medida “Total”):

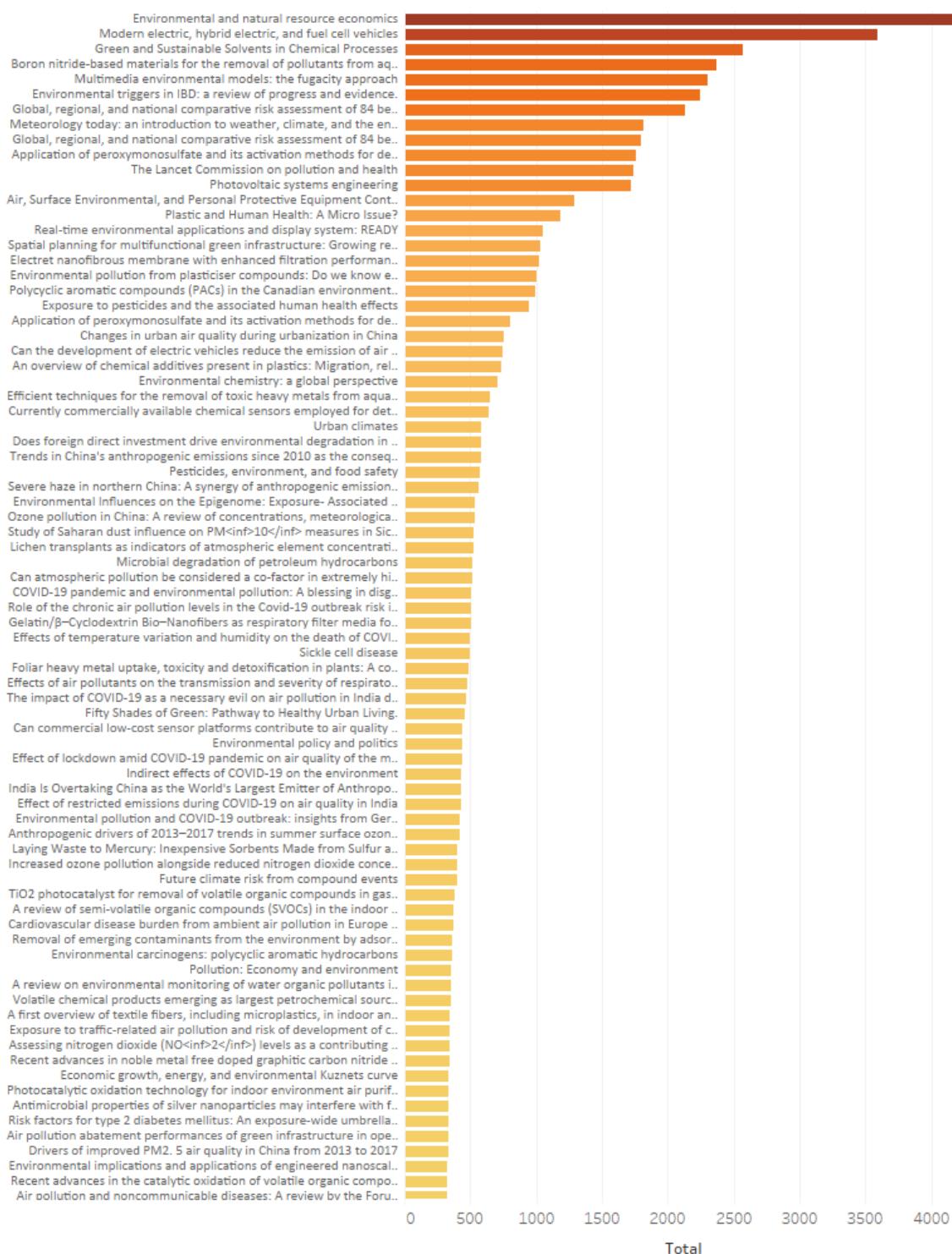
Gráfico 7.1: "Publicaciones eliminadas, búsqueda 1 español"



Fuente: Imagen capturada del software Talend Open Studio

Recuento Gráfico 7.1: Se muestran 73 publicaciones de las 88 eliminadas.

Gráfico 7.2: "Publicaciones eliminadas, búsqueda 1 inglés"



Fuente: Imagen capturada del software Talend Open Studio

Recuento Gráfico 7.2: Se muestran 79 publicaciones de las 113 eliminadas.

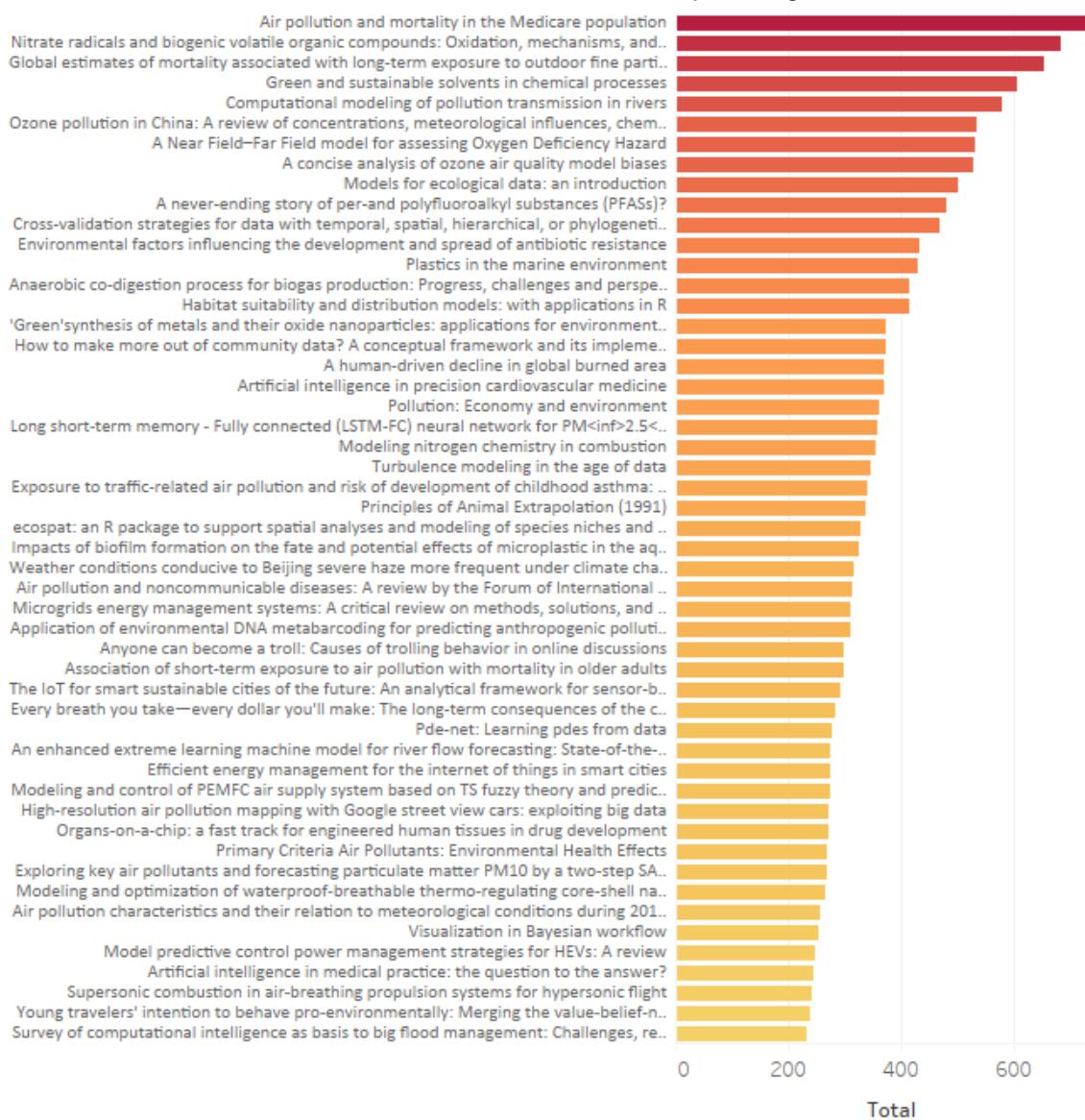
Gráfico 7.3: "Publicaciones eliminadas, búsqueda 2 español"



Fuente: Imagen capturada del software Talend Open Studio

Recuento Gráfico 7.3: Se muestran las 94 publicaciones eliminadas.

Gráfico 7.4: "Publicaciones eliminadas, búsqueda 2 inglés"



Fuente: Imagen capturada del software Talend Open Studio

Recuento Gráfico 7.4: Se muestran las 51 publicaciones eliminadas.

Anexo 6: Constantes

Tabla 7.1: “Constante de avogadro, valores”

Na	Unidad
6,02214129(27)×10 ²³	1/mol
2,73159757(14)×10 ²⁶	lb/mol
1,707248479(85)×10 ²⁵	oz/mol

Fuente: Creación propia.

Tabla 7.2: “Constante de los gases ideales, valores”

R	Unidad	Observación
8,314472 x 10 ⁻³	kJ / (K mol)	
8,314472	J / (K mol)	
0,08205746	L atm / (K mol)	
8,205746 x 10 ⁻⁵	m ³ atm / (K mol)	
8,314472	dm ³ kPa / (K mol)	
8,314472	L kPa / (K mol)	
8,314472	m ³ Pa / (K mol)	
62,36367	L mmHg / (K mol)	
62,36365	L Torr / (K mol)	
83,14472	L mbar / (K mol)	
1,987	cal / (K mol)	
6,13244	lbf ft / (K g-mol)	
10,73159	ft ³ psi / (°R lb-mol)	
0,7302413	ft ³ atm / (°R lb-mol)	
1,986	Btu / (°R lb-mol)	
2,2024	ft ³ mmHg / (K mol)	
8,314472 x 10 ⁷	erg / (K mol)	
1716	ft lb / (°R slug)	Sólo aire, sin vapor de agua
286,9	N m / (kg K)	Sólo aire, sin vapor de agua
286,9	J / (kg K)	Sólo aire, sin vapor de agua
0,08205746	dm ³ atm / (K mol)	
8,314472 x 10 ⁻⁵	m ³ bar / (K mol)	

Fuente: Creación propia.

Tabla 7.3: “Constante de Boltzmann, valores”

Valores de k	Unidades
1,380649 × 10 ⁻²³	J/K
8,617333262 × 10 ⁻⁵	eV/K

Fuente: Creación propia.

Anexo 7: Decisión de la Solución

Para la generación de la solución, se consideraron múltiples algoritmos y modelos de Machine Learning, y en base a sus mejores resultados tras múltiples ajustes se llegó a la conclusión de que la mejor opción a utilizar para el caso actual, es el algoritmo de Random Forest. A continuación, un resumen obtenido de todos los modelos generados:

Regresión logística

Se utilizó la librería sklearn con el modelo LogisticRegression, a continuación, la ilustración del código empleado:

Ilustración 7.13: “Código de algoritmo LogisticRegression”

```
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import confusion_matrix, precision_score, accuracy_score, f1_score, recall_score
data_lr=LogisticRegression(solver='newton-cg')
data_lr.fit(X_train, Y_train)

LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True,
                    intercept_scaling=1, l1_ratio=None, max_iter=100,
                    multi_class='auto', n_jobs=None, penalty='l2',
                    random_state=None, solver='newton-cg', tol=0.0001, verbose=0,
                    warm_start=False)

precision_lr=precision_score(Y_test, data_lr.predict(X_test), average='micro')
print('Precisión del Modelo: ', precision_lr)
print("Exactitud del Modelo: ", accuracy_score(Y_test, data_lr.predict(X_test)))
print("Sensibilidad del Modelo: ", recall_score(Y_test, data_lr.predict(X_test), average='micro'))
print("Puntaje F1 del Modelo: ", f1_score(Y_test, data_lr.predict(X_test), average='micro'))
print("Matriz de confusión: \n", confusion_matrix(Y_test, data_lr.predict(X_test)))

Precisión del Modelo:  0.7760296159185562
Exactitud del Modelo:  0.7760296159185562
Sensibilidad del Modelo:  0.7760296159185562
Puntaje F1 del Modelo:  0.7760296159185563
Matriz de confusión:
[[ 0   3   0   0   0]
 [ 0  14  71  13   0]
 [ 0   2 317 195   0]
 [ 0   0  96 1314  20]
 [ 0   0   2   82  32]]
```

Fuente: Creación propia utilizando Google Colab.

El modelo alcanzó una precisión aproximada del 77,9% para el primer caso, que corresponde a 1 día de predicción.

Árbol de Decisiones

Se utilizó la librería sklearn con el modelo DecisionTreeClassifier, a continuación, la ilustración del código empleado:

```
Ilustración 7.14: "Código de algoritmo DecisionTreeClassifier"
#Importando el arbol de decisión
from sklearn.tree import DecisionTreeClassifier, export_graphviz
from sklearn import tree
import graphviz

#Se define el algoritmo
arbolid4 = DecisionTreeClassifier(criterion='entropy', max_depth=5)

#Se entrena el modelo con los datos de entrenamiento
arbolid4.fit(X_train, Y_train)

#Metricas del Modelo
precision_tree = precision_score(Y_test, arbolid4.predict(X_test), average='micro')
print("Precisión del Modelo: ", precision_tree)
print("Exactitud del Modelo: ", accuracy_score(Y_test, arbolid4.predict(X_test)))
print("Sensibilidad del Modelo: ", recall_score(Y_test, arbolid4.predict(X_test), average='micro'))
print("Puntaje F1 del Modelo: ", f1_score(Y_test, arbolid4.predict(X_test), average='micro'))
print("Matriz de confusión:\n", confusion_matrix(Y_test, arbolid4.predict(X_test)))

Precisión del Modelo:  0.7547431744562703
Exactitud del Modelo:  0.7547431744562703
Sensibilidad del Modelo:  0.7547431744562703
Puntaje F1 del Modelo:  0.7547431744562703
Matriz de confusión:
[[ 0   3   0   0   0]
 [ 0   38  50  10   0]
 [ 0    9 293 212   0]
 [ 0    3 120 1288  19]
 [ 0    0   3 101   12]]
```

Fuente: Creación propia utilizando Google Colab.

El modelo alcanzó una precisión aproximada del 75,47% para el primer caso, que corresponde a 1 día de predicción.

Random Forest

Se utilizó la librería sklearn con el modelo RandomForestClassifier, a continuación, la ilustración del código empleado:

Ilustración 7.15: “Código de algoritmo RandomForestClassifier”

```
from sklearn.ensemble import RandomForestClassifier

#Se define el algoritmo
rForest = RandomForestClassifier(n_estimators=400, random_state = 1)

#Se entrena el modelo con los datos de entrenamiento
rForest.fit(X_train, Y_train)
estimator = rForest.estimators_[399]
#Metricas del Modelo
precisionRf = precision_score(Y_test, rForest.predict(X_test), average='micro')
print('Precisión del Modelo: ', precisionRf)
print("Exactitud del Modelo: ", accuracy_score(Y_test, rForest.predict(X_test)))
print("Sensibilidad del Modelo: ", recall_score(Y_test, rForest.predict(X_test), average='micro'))
print("Puntaje F1 del Modelo: ", f1_score(Y_test, rForest.predict(X_test), average='micro'))
print("Matriz de confusión:\n", confusion_matrix(Y_test, rForest.predict(X_test)))

Precisión del Modelo:  0.8028690421101342
Exactitud del Modelo:  0.8028690421101342
Sensibilidad del Modelo:  0.8028690421101342
Puntaje F1 del Modelo:  0.8028690421101342
Matriz de confusión:
[[ 2   1   0   0   0]
 [ 0  41  45  12   0]
 [ 0  14 354 146   0]
 [ 1   1  97 1305  26]
 [ 0   0   3   80  33]]
```

Fuente: Creación propia utilizando Google Colab.

El modelo alcanzó una precisión aproximada del 80,28% para el primer caso, que corresponde a 1 día de predicción.

Ada Boost

Se utilizó la librería sklearn con el modelo AdaBoostClassifier, a continuación, la ilustración del código empleado:

Ilustración 7.16: “Código de algoritmo AdaBoostClassifier”

```
#Se importa la libreria
from sklearn.ensemble import AdaBoostClassifier

#Se define el algoritmo
adaClassif = AdaBoostClassifier(base_estimator=rForest, n_estimators=450, learning_rate=1.5)

#Se entrena el modelo con los datos de entrenamiento
adaClassif = adaClassif.fit(X_train, Y_train)

#Metricas del Modelo
precision_ada = precision_score(Y_test, adaClassif.predict(X_test), average='micro')
print('Precisión del Modelo: ', precision_ada)
print("Exactitud del Modelo: ", accuracy_score(Y_test, adaClassif.predict(X_test)))
print("Sensibilidad del Modelo: ", recall_score(Y_test, adaClassif.predict(X_test), average='micro'))
print("Puntaje F1 del Modelo: ", f1_score(Y_test, adaClassif.predict(X_test), average='micro'))
print("Matriz de confusión:\n", confusion_matrix(Y_test, adaClassif.predict(X_test)))

Precisión del Modelo:  0.8019435446552522
Exactitud del Modelo:  0.8019435446552522
Sensibilidad del Modelo:  0.8019435446552522
Puntaje F1 del Modelo:  0.8019435446552522
Matriz de confusión:
 [[ 2   1   0   0   0]
 [ 0   41  46  11   0]
 [ 0   16  354 144   0]
 [ 1   1  100 1302  26]
 [ 0   0    3   79  34]]
```

Fuente: Creación propia utilizando Google Colab.

El modelo alcanzó una precisión aproximada del 80,19% para el primer caso, que corresponde a 1 día de predicción.

Cat Boost

Se utilizó la librería sklearn con el modelo CatBoostClassifier, a continuación, la ilustración del código empleado:

Ilustración 7.17: “Código de algoritmo CatBoostClassifier”

```
#Se importa la libreria
from catboost import CatBoostClassifier

#Se define el algoritmo
CatBoostC = CatBoostClassifier(iterations=10, learning_rate=0.9, depth=10)
CatBoostC.fit(X_train, Y_train)

#Metricas del Modelo
precision_ada = precision_score(Y_test, CatBoostC.predict(X_test), average='micro')
print('Precisión del Modelo: ', precision_ada)
print("Exactitud del Modelo: ", accuracy_score(Y_test, CatBoostC.predict(X_test)))
print("Sensibilidad del Modelo: ", recall_score(Y_test, CatBoostC.predict(X_test), average='micro'))
print("Puntaje F1 del Modelo: ", f1_score(Y_test, CatBoostC.predict(X_test), average='micro'))
print("Matriz de confusión:\n", confusion_matrix(Y_test, CatBoostC.predict(X_test)))
```

0: learn: 0.7399772 total: 152ms remaining: 1.37s
1: learn: 0.6180394 total: 244ms remaining: 974ms
2: learn: 0.5623813 total: 326ms remaining: 761ms
3: learn: 0.5278504 total: 427ms remaining: 640ms
4: learn: 0.5034667 total: 516ms remaining: 516ms
5: learn: 0.4855914 total: 593ms remaining: 395ms
6: learn: 0.4699925 total: 682ms remaining: 292ms
7: learn: 0.4576448 total: 757ms remaining: 189ms
8: learn: 0.4466033 total: 840ms remaining: 93.4ms
9: learn: 0.4334138 total: 922ms remaining: 0us
Precisión del Modelo: 0.7857473391948172
Exactitud del Modelo: 0.7857473391948172
Sensibilidad del Modelo: 0.7857473391948172
Puntaje F1 del Modelo: 0.7857473391948172
Matriz de confusión:
[[0 1 2 0 0]
 [0 46 40 12 0]
 [0 29 347 138 0]
 [1 2 130 1265 32]
 [0 0 2 74 40]]

Fuente: Creación propia utilizando Google Colab.

El modelo alcanzó una precisión aproximada del 78,57% para el primer caso, que corresponde a 1 día de predicción.

Clasificador Bayesiano

Se utilizó la librería sklearn con el modelo GaussianNB, a continuación, la ilustración del código empleado:

Ilustración 7.18: “Código de algoritmo GaussianNB”

```
# Se importa la librería
from sklearn.naive_bayes import GaussianNB

bayesClassif = GaussianNB() #Definición del algoritmo
bayesClassif.fit(X_train, Y_train) #Entrenamiento del algoritmo

Y_predict = bayesClassif.predict(X_test) #Predicción del target en base a los datos de prueba

#Calculo de la precisión del modelo
precision_bayes = precision_score(Y_test, Y_predict, average='micro')
print('Precisión del Modelo: ', precision_bayes)
print("Exactitud del Modelo: ", accuracy_score(Y_test, bayesClassif.predict(X_test)))
print("Sensibilidad del Modelo: ", recall_score(Y_test, bayesClassif.predict(X_test), average='micro'))
print("Puntaje F1 del Modelo: ", f1_score(Y_test, bayesClassif.predict(X_test), average='micro'))
print("Matriz de confusión:\n", confusion_matrix(Y_test, bayesClassif.predict(X_test)))

Precisión del Modelo:  0.6779268857010643
Exactitud del Modelo:  0.6779268857010643
Sensibilidad del Modelo:  0.6779268857010643
Puntaje F1 del Modelo:  0.6779268857010643
Matriz de confusión:
[[ 0  3  0  0  0]
 [ 0  26  55  17  0]
 [ 0  15  400  98  1]
 [ 0   2  300  956 172]
 [ 0   0   0  33  83]]
```

Fuente: Creación propia utilizando Google Colab.

El modelo alcanzó una precisión aproximada del 67,79% para el primer caso, que corresponde a 1 día de predicción.

Máquina de soporte Vectorial, Kernel Lineal

Se utilizó la librería sklearn con el modelo SVC, a continuación, la ilustración del código empleado:

Ilustración 7.19: “Código de algoritmo SVC, linear”

```
from sklearn import svm

# Kernel Lineal
SVM1 = svm.SVC(kernel='linear')
SVM1.fit(X_train, Y_train)

#Metricas del Modelo
precision_SVM1 = precision_score(Y_test,SVM1.predict(X_test), average='micro')
print('Precisión para SVM 1: Kernel Lineal: ', precision_SVM1)
matriz_SVM1 = confusion_matrix(Y_test, SVM1.predict(X_test))
print("Exactitud del Modelo: ", accuracy_score(Y_test, SVM1.predict(X_test)))
print("Sensibilidad del Modelo: ", recall_score(Y_test, SVM1.predict(X_test), average='micro'))
print("Puntaje F1 del Modelo: ", f1_score(Y_test, SVM1.predict(X_test), average='micro'))
print('Matriz de Confusión: ')
print(matriz_SVM1)

Precisión para SVM 1: Kernel Lineal:  0.7686256362795002
Exactitud del Modelo:  0.7686256362795002
Sensibilidad del Modelo:  0.7686256362795002
Puntaje F1 del Modelo:  0.7686256362795002
Matriz de Confusión:
[[ 0   0   3   0   0]
 [ 0   0   88  10   0]
 [ 0   0  338  176   0]
 [ 0   0  107 1323   0]
 [ 0   0    4  112   0]]
```

Fuente: Creación propia utilizando Google Colab.

El modelo alcanzó una precisión aproximada del 76,86% para el primer caso, que corresponde a 1 día de predicción.

Máquina de soporte Vectorial, Kernel Gaussiano

Se utilizó la librería sklearn con el modelo SVC, a continuación, la ilustración del código empleado:

Ilustración 7.20: “Código de algoritmo SVC, rbf”

```
# Kernel RBF
SVM_2 = svm.SVC(kernel='rbf')
SVM_2.fit(X_train, Y_train)

#Métricas del Modelo
precision_SVM2 = precision_score(Y_test,SVM_2.predict(X_test), average='micro')
print('Precisión para SVM 2: Kernel Gaussiano: ', precision_SVM2)
matriz_SVM_2 = confusion_matrix(Y_test, SVM_2.predict(X_test))
print("Exactitud del Modelo: ", accuracy_score(Y_test, SVM_2.predict(X_test)))
print("Sensibilidad del Modelo: ", recall_score(Y_test, SVM_2.predict(X_test), average='micro'))
print("Puntaje F1 del Modelo: ", f1_score(Y_test, SVM_2.predict(X_test), average='micro'))
print('Matriz de Confusión SVM 2: Kernel Gaussiano: ')
print(matriz_SVM_2)

Precisión para SVM 2: Kernel Gaussiano:  0.7963905599259602
Exactitud del Modelo:  0.7963905599259602
Sensibilidad del Modelo:  0.7963905599259602
Puntaje F1 del Modelo:  0.7963905599259602
Matriz de Confusión SVM 2: Kernel Gaussiano:
[[ 0   2   1   0   0]
 [ 0  28  61   9   0]
 [ 0   8 348 158   0]
 [ 0   1  90 1330   9]
 [ 0   0   1  100  15]]
```

Fuente: Creación propia utilizando Google Colab.

El modelo alcanzó una precisión aproximada del 79,63% para el primer caso, que corresponde a 1 día de predicción.

Máquina de soporte Vectorial, Kernel Sigmoidal

Se utilizó la librería sklearn con el modelo SVC, a continuación, la ilustración del código empleado:

Ilustración 7.21: “Código de algoritmo SVC, sigmoid”

```
# Kernel Sigmoid
SVM3 = svm.SVC(kernel='sigmoid')
SVM3.fit(X_train, Y_train)

#Metricas del Modelo
precision_SVM3 = precision_score(Y_test,SVM3.predict(X_test), average='micro')
print('Precisión para SVM 3: Kernel Sigmoidal: ', precision_SVM3)
matriz_SVM3 = confusion_matrix(Y_test, SVM3.predict(X_test))
print("Exactitud del Modelo: ", accuracy_score(Y_test, SVM3.predict(X_test)))
print("Sensibilidad del Modelo: ", recall_score(Y_test, SVM3.predict(X_test), average='micro'))
print("Puntaje F1 del Modelo: ", f1_score(Y_test, SVM3.predict(X_test), average='micro'))
print('Matriz de Confusión SVM 3: Kernel Sigmoidal: ')
print(matriz_SVM3)

Precisión para SVM 3: Kernel Sigmoidal: 0.4835724201758445
Exactitud del Modelo: 0.4835724201758445
Sensibilidad del Modelo: 0.4835724201758445
Puntaje F1 del Modelo: 0.4835724201758445
Matriz de Confusión SVM 3: Kernel Sigmoidal:
[[ 0  0  0  3  0]
 [ 0  1  15  53  29]
 [ 0  1  126  292  95]
 [ 0  0  432  912  86]
 [ 0  0  41  69  6]]
```

Fuente: Creación propia utilizando Google Colab.

El modelo alcanzó una precisión aproximada del 48,35% para el primer caso, que corresponde a 1 día de predicción.

Redes Neuronales

Las pruebas con modelos de Redes Neuronales Profundas tras múltiples combinaciones de estructuras, capas, funciones de activación, número de neuronas, etc., Solo alcanzaron un máximo de 5,65% de precisión en sus resultados más prometedores, una cifra sumamente baja.

La mayor debilidad de las DNN (Deep Neural Networks) es la gran cantidad de datos necesarios para capacitarlas. A diferencia de las redes neuronales convencionales, para las que se proporcionan detalles de

características como parte de la entrada, las DNN necesitan datos suficientes para identificar las características por sí mismas. Como resultado, a menudo requieren un número elevado de muestras para funcionar de manera confiable. La afirmación anterior no se cumple para el caso de estudio actual, ya que justamente las clases que componen el DataSet no tienen un numero de muestras aceptable para este tipo de modelos.

Los datos de entrada deben proporcionar una mayor variación para evitar el "sobreajuste", que ocurre cuando una red neuronal desarrolla inferencias que no se basan en relaciones reales de los datos, a menudo como resultado del entrenamiento en un conjunto demasiado limitado de incidentes reales. La salida funciona bien en el conjunto de entrenamiento, pero no en un entorno del mundo real. A menos que se tenga acceso a una cantidad significativa de datos etiquetados, podría estar mejor con las técnicas tradicionales de aprendizaje automático.

Conclusión

En base a todos los resultados de los modelos estudiados, el modelo que presenta una mejor precisión y métricas es el de Random Forest, con una precisión del 80,28% para el primer día de predicción. Además de la precisión, se debe considerar el problema como uno de clasificación, no así de regresión, ya que, al intentar calcular una salida de índice de contaminación exacta, la precisión de todos los modelos estudiados no alcanzaba ni siquiera un 5%, por lo que se decidió utilizar rangos de valores para identificar niveles o clases de contaminación del aire. Además, dada la naturaleza del modelo, este debe ser capaz de realizar distintas predicciones según sea el día a calcular a futuro (rango máximo 30 para este caso), por lo que se optó por utilizar una columna específica en los Datasets que indique este parámetro, por ende, se generarán para la solución 30 Datasets los cuales serán combinados en uno solo, para así entrenar el modelo de Random Forest.

Anexo 8: Algoritmo de complemento para el modelo

El algoritmo, inicialmente debe agrupar las salidas del modelo por niveles de contaminación según el índice AQI, para realizar esto, se creó el siguiente código:

Ilustración 7.22: “Código: Agrupar salidas por índices.”

```
519 filename = BASE_DIR + '\\modelo1.sav'
520 loaded_model = pickle.load(open(filename, 'rb'))
521 result = loaded_model.predict(data)
522 marcaAnterior = 0
523 marca = 0
524 arregloTemporal = []
525 arregloNiveles = []
526 for index in range(0,len(result)):
527     if index == 0:
528         marca = result[index]
529         marcaAnterior = result[index]
530     else:
531         marca = result[index]
532     if marca == marcaAnterior:
533         arregloTemporal.append(marca)
534         if index == len(result)-1:
535             arregloNiveles.append(arregloTemporal)
536     else:
537         arregloNiveles.append(arregloTemporal)
538         arregloTemporal = []
539         arregloTemporal.append(marca)
540         marcaAnterior = copy.copy(marca)
541
542 curva = realizarPrediccion.definirCurva(arregloNiveles)
```

Fuente: Creación propia.

El resultado final del código presente en la Ilustración 7.22, es un array que tiene la siguiente forma, ejemplo:

Básicamente, se separan los niveles de contaminación entregados por el modelo según igualdad y respetando el orden de salida del modelo, ya que, estos resultados de predicción van del día 1 al 30 de izquierda a derecha.

Luego de realizada la tarea anterior, se procede a identificar a que tipo de traza se debe dibujar en la gráfica según que agrupamiento de niveles de AQI. Para realizar esta tarea, se diseñó el siguiente código:

Ilustración 7.23: “Código: función DefinirCurva.”

```
for index in range(0, len(arreglo)):
    anterior = 0
    actual = arreglo[index][0]
    siguiente = 0
    if actual == 1:
        maximo = nivel_1_max
        minimo = nivel_1_min
    if actual == 2:
        maximo = nivel_2_max
        minimo = nivel_2_min
    if actual == 3:
        maximo = nivel_3_max
        minimo = nivel_3_min
    if actual == 4:
        maximo = nivel_4_max
        minimo = nivel_4_min
    if actual == 5:
        maximo = nivel_5_max
        minimo = nivel_5_min
    if index == 0:
        anterior = arreglo[index][0]
    else:
        anterior = arreglo[index-1][0]
    if index == len(arreglo)-1:
        siguiente = arreglo[index][0]
    else:
        siguiente = arreglo[index+1][0]
    if actual > anterior and actual < siguiente:
        respuesta = respuesta + realizarPrediccion.rectaSubida(arreglo[index], maximo, minimo)
    elif actual < anterior and actual > siguiente:
        respuesta = respuesta + realizarPrediccion.rectaBajada(arreglo[index], maximo, minimo)
    elif actual > anterior and actual > siguiente:
        respuesta = respuesta + realizarPrediccion.curvaSubida(arreglo[index], maximo, minimo)
    elif actual < anterior and actual < siguiente:
        respuesta = respuesta + realizarPrediccion.curvaBajada(arreglo[index], maximo, minimo)
    else:
        respuesta = respuesta + realizarPrediccion.rectaSubida(arreglo[index], maximo, minimo)
return respuesta
```

Fuente: Creación propia.

El código presente en la Ilustración anterior busca identificar en base a diversos criterios, que tipo de trazo debe dibujar en la gráfica para recrear el comportamiento del índice AQI. Una vez identificado el trazo, se llama a la función correspondiente para que realice los cálculos adecuados en la salida. Las opciones posibles de dichas funciones son “rectaSubida”, “rectaBajada”, “curvaSubida” y “curvaBajada”, a continuación, sus algoritmos:

Ilustración 7.24: “Código: función rectaSubida.”

```
def rectaSubida(arreglo: list, max: float, min: float):
    diferencia = max - min - 1;
    porcion = diferencia / len(arreglo)
    temporal = []
    valor:float = min
    for index in range(0, len(arreglo)):
        valor = valor + porcion
        temporal.append(int(valor))
    print("rectaSubida: ", temporal)
    return temporal
```

Fuente: Creación propia.

Ilustración 7.25: “Código: función rectaBajada.”

```
def rectaBajada(arreglo: list, max: float, min: float):
    diferencia = max - min - 1;
    porcion = diferencia / len(arreglo)
    temporal = []
    valor:float = max
    for index in range(0, len(arreglo)):
        valor = valor - porcion
        temporal.append(int(valor))
    print("rectaBajada: ", temporal)
    return temporal
```

Fuente: Creación propia.

Ilustración 7.26: “Código: función curvaSubida.”

```
def curvaSubida(arreglo: list, max: float, min: float)
    diferencia = max - min - 1;
    porcion = diferencia / len(arreglo)
    temporal = []
    valor:float = min
    for index in range(0, len(arreglo)):
        if index <= int(len(arreglo)/2):
            valor = valor + porcion
        else:
            valor = valor - porcion
        temporal.append(int(valor))
    print("curvaSubida: ", temporal)
    return temporal
```

Fuente: Creación propia.

Ilustración 7.27: “Código: función curvaBajada.”

```
def curvaBajada(arreglo: list, max: float, min: float):
    diferencia = max - min - 1;
    porcion = diferencia / len(arreglo)
    temporal = []
    valor:float = max
    for index in range(0, len(arreglo)):
        if index <= int(len(arreglo)/2):
            valor = valor - porcion
        else:
            valor = valor + porcion
        temporal.append(int(valor))
    print("curvaBajada: ", temporal)
    return temporal
```

Fuente: Creación propia.

Una vez ejecutadas todas las funciones correspondientes por cada grupo de salidas del modelo, lo que se consigue es la interpretación de la clasificación por niveles de AQI, a una simulación del comportamiento de la curva del mismo índice. Ejemplo:

De esta forma, se obtiene un arreglo de puntos de una curva la cual será, junto a las fechas de los próximos 30 días, los componentes a graficar como respuesta final del modelo predictivo.