
Universidad Autónoma Metropolitana

UNIDAD CUAJIMALPA


# Selección de variables meteorológicas para la clasificación de los índices de contaminación

Idónea comunicación de resultados

QUE PRESENTA:


JUAN ÁNGEL ACOSTA CEJA


PARA OBTENER EL GRADO DE MAESTRÍA EN
Ciencias Naturales e Ingeniería


Comité tutoral:
DIRECTOR: JULIÁN ALBERTO FRESÁN FIGUEROA
CODIRECTOR: DIEGO ANTONIO GONZÁLEZ MORENO
ASESOR: MÁXIMO EDUARDO SÁNCHEZ GUTIÉRREZ
ASESORA: ALMA ROCÍO SAGACETA MEJÍA



Noviembre, 2024

# Índice general

# Resumen

La contaminación atmosférica es uno de los principales problemas en grandes ciudades como la Ciudad de México, donde altos niveles de contaminantes afectan la salud pública y el medio ambiente. Ante esta situación, el análisis de datos meteorológicos y de calidad del aire se vuelve indispensable para desarrollar modelos predictivos precisos que ayuden a los gobiernos de gestionar y reducir los niveles de contaminación. En este contexto, la selección de características de forma eficiente es importante para mejorar el desempeño y eficiencia de los modelos, ya que permite reducir la dimensionalidad de los datos reduciendo la pérdida de la exactitud de las predicciones.

Este estudio se enfoca en la selección de variables meteorológicas relevantes para la clasificación del Índice Aire Salud, empleando una combinación de técnicas de Aprendizaje Automático y Teoría de Gráficas. Se implementaron varios métodos de selección de características, incluyendo análisis de la regresión mediante el Operador de Selección y Contracción Absoluta Mínima (LASSO, por sus siglas en inglés), Análisis de Componentes Principales (PCA, por sus siglas en inglés) y árboles de decisión, como Árboles de Decisión Iterativos (ID3, por sus siglas en inglés) y Clasificación y Regresión con Árboles (CART, por sus siglas en inglés). Además, se utilizaron métricas de variabilidad como la varianza baja, la Diferencia Media Absoluta (MAD, por sus siglas en inglés) y la Relación de Dispersión (DR, por sus siglas en inglés), y se incorporaron métodos basados en Teoría de Gráficas para analizar las relaciones entre variables.

**Palabras clave**: Calidad del aire, selección de características, Aprendizaje Automático, redes neuronales, Teoría de Gráficas, Técnica de Sobremuestreo Sintético para Minorías (SMOTE).

# Introducción

La calidad del aire es un tema de creciente interés e importancia a nivel global debido a sus efectos adversos sobre la salud pública y el medio ambiente. En las grandes ciudades, como la Ciudad de México, el monitoreo y control de los niveles de contaminación es una prioridad para la protección de la salud de sus habitantes. A lo largo de los años, se han implementado diversas medidas para monitorear y reducir los niveles de contaminantes en el aire, y los avances en las tecnologías de sensores y redes de monitoreo han permitido la recopilación continua de datos atmosféricos y meteorológicos en tiempo real. Sin embargo, el problema radica en analizar esta vasta cantidad de información para obtener predicciones precisas y decisiones informadas.

Este trabajo se enfoca en la selección de variables meteorológicas para la clasificación de los índices de contaminación. La tarea de seleccionar variables relevantes es importante, ya que permite construir modelos predictivos más eficientes, al reducir la dimensionalidad de los datos y mejorar el rendimiento de los algoritmos de Aprendizaje Automático. En este contexto, la metodología utilizada se basa en el Proceso Estándar Cruzado para el Aprendizaje Automático con Aseguramiento de Calidad (CRISP-ML(Q), por sus siglas en inglés), un modelo de proceso estándar diseñado para proyectos de Aprendizaje Automático que incluye prácticas de aseguramiento de la calidad. Esta metodología guía el proceso desde la comprensión del problema hasta el despliegue y mantenimiento del modelo.

Para este estudio, se han considerado varios métodos de selección de características, incluyendo análisis de varianza, regresión mediante el Operador de Selección y Contracción Absoluta Mínima (LASSO, por sus siglas en inglés), Análisis de Componentes Principales (PCA, por sus siglas en inglés), y métodos basados en Teoría de Gráficas. Además, se emplean métricas como la va-

rianza baja, la Diferencia Media Absoluta (MAD, por sus siglas en inglés), y la Relación de Dispersión (DR, por sus siglas en inglés), cada una de las cuales aporta una perspectiva diferente sobre la relevancia de cada variable en relación con el índice de calidad del aire. Los métodos de árboles de decisión, específicamente Árboles de Decisión Iterativos (ID3, por sus siglas en inglés) y Clasificación y Regresión con Árboles (CART, por sus siglas en inglés), también son empleados para evaluar la importancia de las características en función de su posición en el árbol de decisión.

En este estudio se incorpora la Técnica de Sobremuestreo Sintético para Minorías (SMOTE, por sus siglas en inglés) para abordar el problema del desequilibrio en las clases de los datos, mejorando la capacidad de los modelos para identificar correctamente las clases minoritarias. El análisis incluye la construcción de redes neuronales y el uso de medidas de centralidad y otros conceptos en Teoría de Gráficas para evaluar la importancia de las variables en una estructura de red.

La finalidad de este trabajo es contribuir a la construcción de modelos robustos y precisos para la clasificación de los índices de contaminación en la Ciudad de México, empleando herramientas de Aprendizaje Automático y técnicas de selección de variables. Este enfoque permitirá mejorar la exactitud de las predicciones y facilitar la toma de decisiones en la gestión de la calidad del aire.

Este trabajo está organizado en los siguientes capítulos:

- El **Capítulo 1** proporciona una revisión de los conocimientos preliminares necesarios para comprender el problema de la calidad del aire, los índices de contaminación y la metodología CRISP-ML(Q) empleada, junto con los conceptos básicos que facilitan la comprensión del documento.

- En el **Capítulo 2**, se describe el proceso de limpieza y preprocesamiento de los datos, incluyendo la imputación de valores faltantes y la eliminación de valores atípicos.

- El **Capítulo 3** aborda el análisis de Aprendizaje Automático y Teoría de Gráficas, detallando los métodos de selección de características y el diseño de las redes neuronales.

- En el **Capítulo 4**, se presentan los resultados del análisis y los experimentos, comparando el desempeño de los modelos con diferentes subconjuntos de características.
- Finalmente, el **Capítulo 5** ofrece las conclusiones y perspectivas para futuros trabajos, destacando las contribuciones de este estudio y posibles mejoras.

Este trabajo busca contribuir a la ciencia de datos aplicada a problemas ambientales, proponiendo nuevas ideas basadas en Teoría de Gráficas para la selección de variables y la predicción de índices de contaminación. Los enfoques desarrollados en este estudio tienen aplicación en otras ciudades y regiones que enfrentan problemas similares de calidad del aire.

# Objetivos

## Objetivo general

Identificar un subconjunto de características significativas que permitan mejorar la exactitud y el *recall* en la clasificación de las clases peligrosas del Índice de Contaminación utilizando modelos de redes neuronales. Este objetivo se logrará minimizando la pérdida de información mediante la aplicación de técnicas de selección de características, Aprendizaje Automático y análisis basado en Teoría de Gráficas, con el propósito de optimizar el número de variables requeridas para el monitoreo ambiental.

## Objetivos específicos

1. **Identificar variables relevantes:** Determinar las variables que contribuyan significativamente al *recall* en la clasificación del Índice Aire Salud, enfocándose en aquellas que permiten identificar correctamente casos de alta contaminación, incluso a costa de una menor exactitud en otras clases.

2. **Identificar variables menos relevantes:** Determinar las variables que no contribuyan al *recall* o a la *exactitud*.

3. **Analizar la interacción de variables dentro de conjuntos:** Evaluar cómo la importancia y el impacto de cada variable en la clasificación del Índice Aire Salud se ven influenciados por su interacción con otras variables dentro del conjunto.

4. **Investigar técnicas de selección de características:** Investigar y aplicar métodos de Teoría de Gráficas para seleccionar variables.

5. **Evaluar el impacto de métodos combinados en el desempeño predictivo:** Comparar y validar enfoques que integren métodos estadísticos, de variabilidad y de Teoría de Gráficas, para desarrollar modelos robustos que mejoren el *recall* y la exactitud.

6. **Evaluar el rendimiento de técnicas para balancear clases minoritarias:** Utilizar métodos como la Técnica de Sobremuestreo Sintético para Minorías (SMOTE, por sus siglas en inglés) para mejorar el *recall* en clases minoritarias, evaluando el impacto de este enfoque en la exactitud.

7. **Desarrollar un modelo de predicción continuo:** Desarrollar un modelo de predicción para el índice de contaminación

# Capítulo 1

# Conocimientos preliminares

En este capítulo se presentan los conocimientos necesarios para la comprensión de este trabajo. En particular introducimos las definiciones necesarias para entender la clasificación y su medición de los índices de contaminación del aire, su impacto en las contingencias. Además, se describe la metodología de Proceso Estándar Cruzado para el Aprendizaje Automático con Aseguramiento de Calidad (CRISP-ML(Q), por sus siglas en inglés) para el uso de algoritmos de Aprendizaje Automático y métodos de selección de características, junto con Teoría de Gráficas, para modelar problemas de clasificación.

## 1.1. Índice aire salud

El Índice Aire Salud es un indicador para informar a la población sobre la condición de la calidad del aire y sus posibles efectos en la salud. Este índice se calcula para cinco contaminantes: dióxido de azufre ($SO_2$), monóxido de carbono ($CO$), dióxido de nitrógeno ($NO_2$), ozono ($O_3$) y partículas suspendidas en el aire menor a 10 y 2,5 micrómetros ($\mu m$) ($PM_{10}$ y $PM_{2,5}$ respectivamente). Cabe mencionar que este índice se mide en una escala de 0 a 500 y establece seis categorías de riesgo en los contaminantes, denotados con colores distintos. Cuanto mayor sea el índice, peor será la calidad del aire, como lo podemos observar en la Cuadro 1.1.

El antecesor del Índice Aire Salud es el Índice Metropolitano de la Calidad del

Aire (IMECA) [11], donde ambos índices comparten las mismas condiciones para calcular los índices de cada contaminante, con la única excepción de que los rangos de concentración correspondientes a cada color pueden ser más amplios o más estrechos. En particular, la diferencia más significativa se encuentra en los intervalos de las categorías "Buena" y "Moderada", que son más amplios en el IMECA y se reducen en el índice Aire y Salud. Desde este contexto, para determinar las clases del índice Aire y Salud, se realizó una reducción en los intervalos de concentración, basada en la Norma Ambiental del Distrito Federal NADF-009-AIRE-2006 [13].

Cuadro 1.1: Intervalos de concentración para la asignación de colores del Índice Aire Salud. En este caso $PM_1(t)$, $PM_8(t)$ y $PM_{24}(t)$ son lo promedios móviles de 1, 8 y 24 horas respectivamente. Por otro lado, $PMP_{12}(t)$ es el promedio móvil ponderado de 12 horas [26].

| | Buena 0-50 | Moderada 51-100 | Dañina 101-150 | Muy dañina 151-200 | Peligro más 200 |
|---|---|---|---|---|---|
| $O_3$ [ppm] $PM_1(t)$ | $[0, 0{,}051]$ | $(0{,}051, 0{,}095]$ | $(0{,}095, 0{,}135]$ | $(0{,}135, 0{,}175]$ | $(0{,}175, \infty)$ |
| $O_3$ [ppm] $PM_8(t)$ | $[0, 0{,}051]$ | $(0{,}051, 0{,}070]$ | $(0{,}070, 0{,}092]$ | $(0{,}092, 0{,}114]$ | $(0{,}114, \infty)$ |
| $NO_2$ [ppm] $PM_1(t)$ | $[0, 0{,}107]$ | $(0{,}107, 0{,}210]$ | $(0{,}210, 0{,}230]$ | $(0{,}230, 0{,}250]$ | $(0{,}250, \infty)$ |
| $SO_2$ [ppm] $PM_{24}(t)$ | $[0, 0{,}008]$ | $(0{,}008, 0{,}110]$ | $(0{,}110, 0{,}165]$ | $(0{,}165, 0{,}220]$ | $[0{,}220, \infty)$ |
| $CO$ [ppm] $PM_8(t)$ | $[0, 8{,}75]$ | $(8{,}75, 11{,}00]$ | $(11{,}00, 13{,}30]$ | $(13{,}30, 15{,}50]$ | $[15{,}51, \infty)$ |
| $PM_{10}$ $[\mu g/\mathrm{m}^3]$ $PMP_{12}(t)$ | $[0, 50]$ | $(50, 75]$ | $(75, 155]$ | $(155, 235]$ | $(235, \infty)$ |
| $PM_{2,5}$ $[\mu g/\mathrm{m}^3]$ $PMP_{12}(t)$ | $[0, 25]$ | $(25, 45]$ | $(45, 79]$ | $(79, 147]$ | $(147, \infty)$ |

Las ecuaciones que describen el índice de cada uno de los contaminantes [17] están dadas por:

$$\text{Índice} = (k(C_{obs} - PC_{inf})) + I_{inf}, \tag{1.1}$$

donde $C_{obs}$ es la concentración observada del contaminante, en partes por millón $(ppm)$ para $O_3$, $NO_2$,$SO_2$ y $CO$, y miligramos entre metros cúbicos $(\frac{\mu g}{m^3})$ para $PM_{10}$ y $PM_{2,5}$

Por otro lado, la constante $k$ se calcula como:

$$k = \frac{I_{sup} - I_{inf}}{PC_{sup} - PC_{inf}}, \tag{1.2}$$

donde:

- $k$ = Constante de proporcionalidad en *ppm* para $O_3$, $NO_2$,$SO_2$ y $CO$, mientras que para $PM_{10}$ y $PM_{2,5}$ en $\frac{m^3}{\mu g}$.
- $PC_{sup}$ = Concentración de punto de corte superior o igual a la concentración a evaluar, en *ppm* para $O_3$, $NO_2$,$SO_2$ y $CO$, mientras que para $PM_{10}$ y $PM_{2,5}$ en $\frac{\mu g}{m^3}$.
- $PC_{inf}$ = Concentración de punto de corte inferior o igual a la concentración a evaluar, en *ppm* para $O_3$, $NO_2$,$SO_2$ y $CO$, mientras que para $PM_{10}$ y $PM_{2,5}$ en $\frac{\mu g}{m^3}$.
- $I_{sup}$ = Índice de la $PC_{sup}$ adimencional.
- $I_{inf}$ = Índice de la $PC_{inf}$ adimencional.

La constante de proporcionalidad $k$ para cada uno de los contaminantes se puede consultar en [17].

### 1.1.1. Contingencia ambiental

La Ciudad de México se destaca a nivel internacional por sus estrictos criterios para declarar contingencias ambientales, con el objetivo de proteger la salud de sus habitantes. A modo de comparación, en Estados Unidos se activa una contingencia ambiental por ozono cuando la concentración promedio en una hora supera las 200 partes por billón (*ppb*). En contraste, en la Zona Metropolitana del Valle de México (ZMCM), la Comisión Ambiental de la Megalópolis (CAMe) establece la contingencia ambiental por ozono cuando se alcanzan 155 *ppb* [18].

En los últimos años, el umbral para activar la contingencia ambiental se ha ajustado en ocho ocasiones, haciéndolo progresivamente más estricto para mejorar la protección de la salud pública [18], por ejemplo:

- Entre 2011 y 2015, la contingencia ambiental se activaba cuando las concentraciones promedio de ozono en una hora superaban los 199 *ppb*.
- Entre 2015 y 2016, la contingencia ambiental se activaba cuando las concentraciones promedio de ozono en una hora superaban los 185 *ppb*.
- De 2016 a la fecha, la contingencia ambiental se activaba cuando las concentraciones promedio de ozono en una hora superaban los 155 *ppb*

La activación de la contingencia ambiental se basa en las concentraciones de tres contaminante: ozono, partículas $PM_{10}$ y $PM_{2,5}$. En el Cuadro 1.2 podemos observar los valores mínimos y máximos para la activación y desactivación de la contingencia ambiental.

Cuadro 1.2: Activación y suspensión de la Fase I, II y combinada.

| **Contingencia** | **Activación** | | | **Suspención** | | |
|---|---|---|---|---|---|---|
| | **Índice (concentraciones)** | | | **Índice** | | |
| | $O_3$ **promedio en una hora** | $PM_{10}$ **Promedio móvil 24 horas** | $PM_{2,5}$ **Promedio móvil 24 horas** | $O_3$ | $PM_{10}$ | $PM_{2,5}$ |
| **Fase I** | Mayor a 150 puntos (mayor a 154*ppb*) | Mayor a 150 puntos (mayor a 214 $\frac{\mu g}{m^3}$) | Mayor a 150 puntos (mayor a 97,4 $\frac{\mu g}{m^3}$) | Menor a 150 puntos con pronóstico meteorológico favorable al siguiente día | | |
| **Fase II** | Mayor a 200 puntos (mayor a 204*ppb*) | Mayor a 200 puntos (mayor a 354 $\frac{\mu g}{m^3}$) | Mayor a 200 puntos (mayor a 150,4 $\frac{\mu g}{m^3}$) | Menor a 150 puntos con pronóstico meteorológico favorable al siguiente día | | |
| **Fase combiada** | Ozono mayor a 150 puntos y $PM_{10}$ o $PM_{2,5}$ mayor a 140 puntos<br><br>Ozono mayor a 140 puntos y $PM_{10}$ o $PM_{2,5}$ mayor a 150 punto | | | Menor a 150 y 140 puntos dependiendo del contaminante, con pronóstico meteorológico favorable al siguiente día | | |

### 1.1.2. Sistema de Monitoreo Atmosférico (SIMAT)

Para determinar las concentraciones de contaminación y de otras variables meteorológicas se creó el Sistema de Monitoreo Atmosférico (SIMAT). Actualmente, el SIMAT está integrado por cuatro subsistemas donde se realiza el monitoreo diversas variables, que interactúan entre sí para integrar un

sistema más amplio, confiable y representativo, que incluye un laboratorio central de análisis de otras especies y mantenimiento de equipos, así como el procesamiento de la información y su difusión y proporciona información importante sobre la calidad del aire en la Zona Metropolitana de Ciudad de México (ZMCM). El SIMAT tiene 44 estaciones en total, las cuales se encuentran distribuidas en los cuatro subsistemas o redes de monitoreo [11, 12].

1. Red Automática de Monitoreo Atmosférico (RAMA): Esta red consta de 34 estaciones, los cuales tienen la tarea de medir e informar de manera oportuna a la población sobre los niveles de contaminación registrados como: $O_3$, $NO_x$,$SO_2$ y $CO$, $PM_{10}$ y $PM_{2,5}$.

2. Red de Meteorología (REDMET): Consta de 9 estaciones para registrar partículas suspendidas totales con métodos de medición manuales y complementar el registro de la RAMA. Los parámetros que mide son: $PST$, $PM_{10}$, $PM_{2,5}$, $Pb$.

3. Red de Depósito Atmosférico (REDDA): Esta red tiene 16 estaciones que utilizan equipos semiautomáticos para la recolección de muestras de depósito seco (polvo sedimentable) y depósito húmedo (lluvia, granizo, nieve, rocío) en los sitios de muestreo. Los parámetros que mide son aniones, cationes, pH, nitratos, sulfatos y conductividad eléctrica.

4. Red Manual (REDMET): Esta red consta de 28 estaciones para medir los parámetros meteorológicos que influyen en la dispersión, transporte y transformación de los contaminantes en la atmósfera. Los parámetros que miden son temperatura, humedad relativa, dirección y velocidad de viento, radiación solar (UV-A y UV-B) y presión barométrica.

### 1.1.3. Modelos de predicción metereológica y de emisiones

Además de las conexiones que hay entre los subsistemas de monitoreo, el Sistema de Predicción de la Calidad del Aire de la Ciudad de México (AQFS-Mex) es una herramienta útil para la predicción de las posibles concentraciones de los contaminantes atmosféricos durante el día actual y el siguiente. Con esto, la población puede tomar medidas preventivas para disminuir el

riesgo para su salud. Esta herramienta fue desarrollada por la Secretaría del Medio Ambiente de la Ciudad de México (SEDEMA) en colaboración con el Centro Nacional de Supercomputación de Barcelona (BSC, por sus siglas en inglés) y su objetivo principal es elaborar un pronóstico meteorológico y de contaminantes para las siguientes 24 y 48 horas, utilizando el conocimiento actual sobre la química y dinámica atmosféricas de la región [28]. Sin embargo, esta herramienta originalmente se utilizaba para predecir el IMECA; actualmente, se emplean promedios móviles para estimar el Índice Aire Salud.

El desempeño del sistema AQFS-Mex de la Ciudad de México es razonable, ya que en términos generales, el rendimiento del modelo, definido como el porcentaje de pronósticos categorizados como buenos y regulares, fue de 80 % para el pronóstico de 24 horas y de 72 % para el de 48 horas [27].

### 1.1.4. Modelos meteorológicos y de emisiones

Actualmente, para predecir el Índice Aire Salud, se utiliza el método de promedios móviles, el cual se define como la media de las concentraciones en un intervalo de $n$ horas consecutivas. En otras palabras, consiste en calcular el promedio entre la hora de interés($t$) y las $n-1$ horas/minutos previas, ya sea del mismo día o del día anterior:

$$PM_n(t) = \frac{1}{n}\sum_{i=0}^{n-1} C_{t-i} \tag{1.3}$$

Donde $PM_n(t)$ es el promedio móvil en el tiempo $t$ con $n$ horas/minutos previos y $C_t$ representa la concentración en la hora $t$.

Los promedios móviles ponderados para los contaminantes $PM_{10}$ y $PM_{2,5}$ se multiplica por un factor de ponderación $W$ a la ecuación 1.3:

$$PMP_n(t) = \frac{1}{\sum_1^n W^i}\sum_{i=0}^{n-1} C_{t-i}W^i \tag{1.4}$$

Donde:

$$W = \begin{cases} w & \text{si } w > 0{,}5 \\ 0{,}5 & \text{si } w \leq 0{,}5 \end{cases} \tag{1.5}$$

Con $w = 1 - \dfrac{C_{max} - C_{min}}{C_{max}}$, donde $C_{max}$ y $C_{min}$ es la concentración máxima y mínima respectivamente en el periodo $n$

## 1.2. CRISP-ML(Q)

Uno de los primeros modelos de proceso estándar para el desarrollo de Aprendizaje Automático (ML, por sus siglas en inglés) es el modelo de **Proceso Estándar entre Industrias para la Minería de Datos** (CRISP-DM, por sus siglas en inglés) que están estrechamente relacionados con los modelos de ML [6, 30]. Sin embargo, se ha notado principalmente dos fallas de CRISP-DM:

- el enfoque de CRISP-DM se implementa en la minería de datos y no abarca la implementación de modelos de ML que tomen decisiones en tiempo real durante un largo periodo. Se debe considerar que estos modelos deben de mantener su rendimiento en un entorno en constante cambio en el tiempo, por lo que es necesario supervisar y mantener constantemente el modelo después de su implementación[3].
- la preocupación principal es que CRISP-DM no proporciona guía sobre la metodología de control de calidad. [3].

Dadas las deficiencias anteriores, se propuso el modelo del Proceso Estándar Cruzado para el Aprendizaje Automático con Aseguramiento de Calidad (CRISP-ML(Q)). Este modelo se describe en seis pasos, como se muestra en la Figura 1.1 y se detalla a continuación:

1. **Conocimiento del problema:** en este primer paso, se busca comprender el problema, la información disponible y los recursos necesarios para abordarlo. Es importante establecer claramente los objetivos de

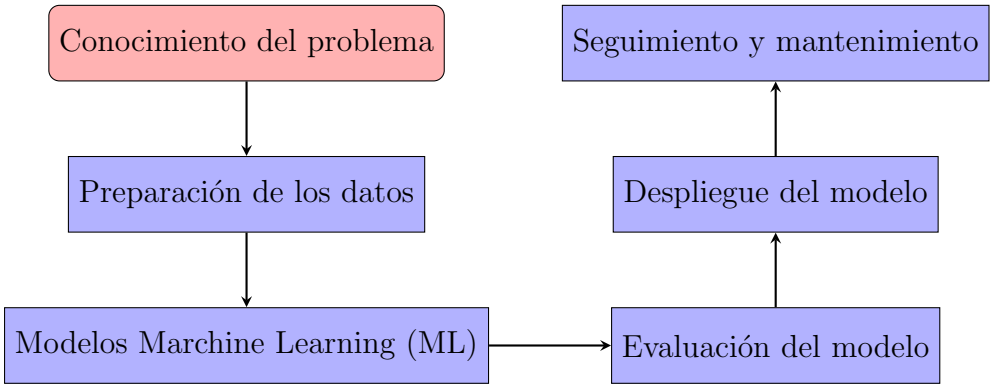



Figura 1.1: Pasos para la implementación de la metodología CRISP-ML(Q).

negocio y las metas de Aprendizaje Automático para poder medir el éxito del proyecto.

2. **Preparación de los datos:** en este paso, se recopilan y analizan los datos necesarios para abordar el problema. Se trata de comprender la estructura y la calidad de los datos, identificar patrones, tendencias y relaciones entre las variables. Es importante evaluar la suficiencia y representatividad de los datos para el problema específico. Además, se limpian, transforman y seleccionan los datos relevantes para el modelo. Se trata de preparar los datos para el proceso de modelado, eliminando valores faltantes, valores atípicos o datos duplicados.

3. **Modelos Aprendizaje Automático:** en este paso, se aplican técnicas de ML y se selecciona el mejor modelo. Es importante evaluar varios modelos y comparar sus desempeños mediante métricas adecuadas. Una vez seleccionado el mejor modelo, se entrena y se ajusta para mejorar su rendimiento.

4. **Evaluación o garantía de modelo:**

   durante esta fase, el rendimiento del modelo entrenado debe validarse en un conjunto de pruebas. Además, la robustez del modelo debe evaluarse utilizando datos de entrada ruidosos o incorrectos. Por otra parte, es una buena práctica desarrollar un modelo de ML explicable para proporcionar confianza, cumplir con los requisitos reglamentarios [5]. Además, Si no se cumplen los criterios de éxito, se puede retroceder

a fases anteriores (modelado o incluso preparación de datos) o detener el proyecto.

5. **Despliegue:** las opciones específicas de ML son, por ejemplo, para optimizar hacia el hardware de destino con respecto a la disponibilidad de CPU y GPU, para optimizar hacia el sistema operativo de destino[5]

6. **Seguimiento y mantenimiento:** en este último paso, se monitorea el rendimiento del modelo en el tiempo, se realiza el mantenimiento y actualización del modelo según sea necesario, y se toman medidas para mejorar el rendimiento y la eficacia del modelo.

Es importante destacar que en el paso 2 de esta metodología se seleccionan las características o variables más relevantes para el nuestro problema. Este trabajo se enfoca principalmente en esta etapa, ya que la selección adecuada de estas características el objetivo principal para el desempeño de los modelos. A continuación, se describen los métodos de selección utilizados.

## 1.3. Técnicas de selección individual

Esta sección se centra en explorar diferentes métodos de selección de características individuales utilizados en Aprendizaje Automático, como la varianza baja [9], Operador de Selección y Contracción Absoluta Mínima (LASSO, por sus siglas en inglés) [4], árboles de decisión [19], Análisis de Componentes Principales (PCA, por sus siglas en inglés) [10], Diferencia Media Absoluta (MAD, por sus siglas en inglés) [31], Relación de Dispersión (DR, por sus siglas en inglés). La idea general de la selección de características individuales es la de asignar un puntuaje a todas las características, con base en un cierto criterio, y posteriormente, elegir las mejores.

### 1.3.1. Varianza baja

La varianza es una medida que nos indica el grado de dispersión de los datos de una característica. Una menor varianza implica una menor dispersión de los datos, es decir, cuando los datos de una característica se acercan a cero,

no proporcionan información relevante o suficiente para el análisis. Esto se debe a que mantienen un comportamiento similar o constante a lo largo del tiempo. Para usar la varianza como método de selección de características, la definiremos de la siguiente manera:

Sea $\mathcal{X} = \{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_m\}$ el conjunto de características originales y sea $\mathcal{D}_i = \{d_1^{(i)}, d_2^{(i)} \ldots d_n^{(i)}\}$ el conjunto de datos asociados a la característica $\mathcal{C}_i$ para cada $i \in \{1, 2, \ldots, n\}$. A cada característica $\mathcal{C}_i$ le asociaremos el valor correspondiente a su varianza mediante la siguiente función:

$$f : \mathcal{X} \to \mathbb{R}$$

$$f(\mathcal{C}_i) = \sum_{j=1}^{n} \frac{(d_j^{(i)} - \overline{\mathcal{D}_i})^2}{n},$$

donde $\overline{\mathcal{D}_i}$ es la media del conjunto $\mathcal{D}_i$.

### 1.3.2. Operador de Selección y Contracción Absoluta Mínima

El Operador de Selección y Contracción Absoluta Mínima (LASSO) es una técnica para abordar el problema del sobreajuste, proporcionando nueva información al modelo que generaliza la regresión tradicional eficazmente. El sobreajuste es cuando el modelo aprende de los datos de entrenamiento de tal manera que el rendimiento del modelo empeora cuando se evalúa en los datos de prueba. La regresión LASSO proporciona una forma de seleccionar las características para obtener un modelo que generaliza mejor. La ecuación que define la regresión LASSO es la siguiente:

$$LASSO = \sum_{i}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{j}^{m} |w_j|, \tag{1.6}$$

donde

$$\hat{y}_i = y_0 + \sum_{j}^{m} X_{ij} w_j.$$

Como se puede observar en la ecuación 1.6, LASSO es una versión de la regresión lineal que utiliza una penalización en términos de la norma $L1$, es decir, con base en la suma de los valores absolutos de los términos individuales. Esta penalización es útil para la selección de características, ya que puede reducir la relevancia de una característica en el modelo permitiendo que otras características destaquen.

Cuanto mayor sea el valor elegido para $\lambda$, mayor será la penalización de los pesos de las características $w_j$ y más se eliminarán. Utilizaremos la regresión de LASSO como método de selección de características de la siguiente manera:

Sea $\mathcal{X} = \{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_m\}$ el conjunto de características originales, sea $\mathcal{D}_i = \{d_1^{(i)}, d_2^{(i)} \ldots d_n^{(i)}\}$ el conjunto de datos asociados a la característica $\mathcal{C}_i$ para cada $i \in \{1, 2, \ldots, n\}$. Para un valor de $\lambda_i$ conservamos las $n$ características con mayor relevancia para LASSO, mediante la siguiente función:

$$f : \mathcal{X} \to \mathbb{R}$$
$$f(\mathcal{C}_i; \lambda) = w_i$$

### 1.3.3. Análisis de Componenetes Principales (PCA)

El algoritmo de Análisis de Componentes Principales (PCA) se encarga de transformar un conjunto de variables correlacionadas en un nuevo conjunto de variables no correlacionadas, denominadas componentes principales, con el propósito de reducir la dimensionalidad del conjunto de datos. Es importante destacar que estos componentes principales son combinaciones lineales de las variables originales y se ordenan según la cantidad de información (varianza) que tengan. Sin embargo, estas nuevas variables no necesariamente poseen una interpretación real. Esta técnica se basa principalmente en la matriz de covarianza de todas las características y en sus correspondientes vectores propios.

Se determina la cantidad de componentes principales que se desea conservar, teniendo en cuenta qué parte de la variabilidad de los datos se desea conservar. Finalmente, para usar PCA como una técnica de selección de características, se le asocia un vector a cada característica original, con base

en los vectores propios de la matriz de covarianza, y se utiliza el algoritmo $k-$medias, para clasificar los vectores similares. Se elige el vector más cercano a cada centroide como el representante de la clase, es decir, la característica original asociada a este vector será la seleccionada.

### 1.3.4. Diferencia Media Absoluta (MAD)

La Diferencia Media Absoluta (MAD) determina la disparidad que existe entre los datos de una variable y su media. Mientras una varibale tenga una mayor MAD, los datos se encuentran más dispersos, es decir, esta variable será de mayor relevancia. Para usar MAD como método de selección de características, la definiremos de la siguiente manera:

Sea $\mathcal{X} = \{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_m\}$ el conjunto de características originales y sea $\mathcal{D}_i = \{d_1^{(i)}, d_2^{(i)} \ldots d_n^{(i)}\}$ el conjunto de datos asociados a la característica $\mathcal{C}_i$ para cada $i \in \{1, 2, \ldots, n\}$. A cada característica $\mathcal{C}_i$ le asociaremos el valor correspondiente a su MAD mediante la siguiente función:

$$f : \mathcal{X} \to \mathbb{R}$$

$$f(\mathcal{C}_i) = \frac{\sum |d_j^{(i)} - \overline{\mathcal{D}_i}|}{n},$$

donde $\overline{\mathcal{D}_i}$ es la media del conjunto $\mathcal{D}_i$.

### 1.3.5. Relación de Dispersión (DR)

La Relación de Dispersión (DR) se puede utilizar como una medida de variabilidad de los datos. Cuando todos los datos de una muestra tienen (aproximadamente) el mismo valor, la relación de dispersión estará cerca de 1, lo que se puede interpretar como que la característica asociada a esos datos es de baja en relevancia. Por el contrario, un mayor valor de DR implica una mayor dispersión, por lo tanto, una característica más relevante. A cada característica $\mathcal{C}_i$ le asociaremos el valor correspondiente a su varianza mediante la siguiente función:

$$f : \mathcal{X} \to \mathbb{R}$$
$$f(\mathcal{C}_i) = \frac{\overline{\mathcal{D}_i}}{\mathcal{D}_{gi}},$$

donde $\overline{\mathcal{D}_i}$ es la media y $\mathcal{D}_{gi}$ es la media geométrica del conjunto $\mathcal{D}_i$.

## 1.4. Técnicas de selección por parejas

Los métodos de Chi-Cuadrada [9] y Correlación de Pearson [25] descritos en esta sección, en lugar de asignar un puntaje a cada una de las características, nos permiten establecer un orden de importancia en la relación entre las características y la variable objetivo.

### 1.4.1. Chi-Cuadrada

La prueba de Chi-Cuadrada se utiliza en estadística para probar la independencia de eventos o variables. Esta prueba asigna a una pareja variables un número real. En general, un valor en la prueba de Chi-Cuadrada más grande indica comúnmente una mayor dependencia entre las dos variables, de esta forma, las variables que tengan una mayor dependencia con la variable objetivo tendrán una mayor importancia. El método contiene principalmente dos parámetros, la frecuencia esperada, que se obtiene a partir del cuadro de contingencia asociada a los dos eventos o variables; y la frecuencia observada que se obtiene de los datos originales. Si la frecuencia esperada es igual a la frecuencia observada, el valor de Chi-Cuadrada será cero y no habrá dependencia.

$$\chi^2(\mathcal{C}_i, \mathcal{C}_j) = \sum \frac{(f_0 - f_e)^2}{f_e}, \tag{1.7}$$

donde $f_0$ es la frecuencia observada y $f_e$ es la frecuencia esperada si no existiera relación relación entre las variables. Es importante resaltar que la prueba de Chi-Cuadrada tiene como objetivo principal indagar si existe una relación o no entre las variables.

### 1.4.2. Correlación de Pearson

La Correlación de Pearson mide la relación lineal entre dos variables aleatorias. Sus valores oscilan entre 1 y $-1$. Un valor de 1 indica una correlación perfectamente positiva, mientras que un valor de $-1$ representa una correlación perfectamente negativa. Por otro lado, un valor de 0 indica que no existe una relación lineal entre las variables. Una variable cuya correlación se acerque a una correlación perfectamente positiva o negativa será más relevante que una que se aproxime a cero. Este coeficiente está dado por la siguiente expresión:

$$\rho(\mathcal{C}_i, \mathcal{C}_j) = \frac{\mathrm{Cov}(\mathcal{D}_i, \mathcal{D}_j)}{\sqrt{\mathrm{Var}(\mathcal{D}_i) \cdot \mathrm{Var}(\mathcal{D}_j)}},$$

donde $\mathrm{Cov}(\mathcal{D}_i, \mathcal{D}_j)$ es la covarianza y $\mathrm{Var}(\mathcal{D}_i)$ es la varianza.

## 1.5. Gráficas asociadas a conjuntos de características

En ésta sección presentamos las definiciones y conceptos necesarios de Teoría de Gráficas que utilizamos en este trabajo. Para mayor información sobre definiciones y conceptos que no aparecen aqui se puede consultar el libro de Chartrand, Lesniak y Zhang [7].

Una **gráfica** $G$ es una pareja $(V(G), A(G))$ formada por un conjunto finito y no vacío $V(G)$ de **vértices** y de un conjunto $E(G)$ de parejas no ordenadas de elementos de $V(G)$, a los elementos de $A(G)$ les llamamos **aristas**.

Sea $G$ una gráfica y $u, v \in V(G)$. Si $\{u, v\} \in A(G)$, entonces usamos $uv$ para denotarl a la arista $\{u, v\}$. A las arista de la forma $uu$ se les llama **lazo**.

El **orden** de una gráfica $G$ es la cardinalidad de $V(G)$ y el **tamaño** de $G$ es la cardinalidad de $A(G)$.

Sea $v \in V(G)$. La **vecindad** de $v$ en $G$ es el conjunto:

$$N_G(v) = \{u \in V(G) \mid \ uv \in A(G)\}.$$

El **grado** del vértice $v$ es

$$d_G(v) = |N_G(v)|.$$

Si es claro en que gráfica estamos trabajando, usaremos $d(v)$ y $N(v)$ en lugar de $d_G(v)$ y $N_G(v)$. Si $d_G(v) = 0$, se dice que $v$ es un **vértice aislado** de $G$, y si $d_G(v) = 1$, se dice que $v$ es una **hoja** de $G$.

Una gráfica es **completa** si todo par de vértices esta conectado con una arista. Una subgrafica $H$ de $G$ es un **clan** si $H$ es completa y maximal por contensión.

Una **gráfica ponderada en aristas** es una gráfica $G = (V(G), E(G))$ junto con una función peso $w : E(G) \to \mathbb{R}$ que asigna un valor real a cada arista de $G$. Para cada arista $e \in E$, el valor $w(e)$ se llama el **peso** de la arista $e$. De manera análoga se define una gráfica ponderada en vértices.

Sea $G = (V(G), E(G))$ una gráfica con $V(G) = \{v_1, v_2, \ldots, v_n\}$. La **matriz de adyacencia** de $G$ es una matriz $A = [a_{ij}]$ de $n \times n$ tal que:

$$a_{ij} = \begin{cases} 1 & \text{si } v_i v_j \in E(G), \\ 0 & \text{en caso contrario.} \end{cases}$$

Sea $G$ una gráfica. Un **camino** $W$ es una sucesión de vértices $(v_1, v_2, \ldots, v_k)$ que cumplen que $v_i v_{i+1} \in E(G)$, para todo $i \in \{1, 2, \ldots, k-1\}$. La **longitud** de $W$ es el número de aristas que tiene. Si $W$ comienza en $u$ y finaliza en $v$, diremos que $W$ es un $(u, v)$-camino. Diremos que $W$ es un camino cerrado si comienza y termina en el mismo vértice. Una **trayectoria** es un camino que no repite vértices ni aristas. Un ***ciclo*** es un camino cerrado que no repite aristas ni vértices.

Una **geodésica** entre dos vértices $u$ y $v$ es un $(u, v)$-camino de longitud mínima. La **distancia** entre dos vértices $u$ y $v$, denotada como $d(u, v)$, es la longitud de la $uv-$trayectoria más pequeña, es decir, la longitud de una $uv$-geodésica.

Sea $G$ un gráfica con $V(G) = \{v_1, v_2, \ldots, v_n\}$. La **distancia promedio** de $v_j$, denotada como $l(v_j)$ se define como:

$$l(v_j) = \frac{\sum_{i=1}^{n} d(v_j, v_i)}{n-1}.$$

Sea $G = (V(G), E(G))$ una gráfica. Un **conjunto dominante** en $G$ es un subconjunto de vértices $S \subseteq V(G)$ que cumple que para todo $v \in V(G) \setminus S$, existe un $u \in S$ tal que $uv \in E(G)$.

Un **emparejamiento** es un conjuntos de aristas independientes, es decir, un conjunto de aristas que no tienen un vértice en comun.

Una **digráfica** es una pareja ordenada $D = (V(D), F(D))$, donde $V(D)$ es un conjunto de vértices y $A$ es un conjunto de pares ordenados de vértices llamados **arcos** o **flechas**. Cada arco $uv \in F(D)$ indica una conexión dirigida desde el vértice $u$ al vértice $v$.

En la siguiente sección presentamos algunas medidas de centralidad que estudiaremos en las gráficas obtenida con el objetivo de determinar las características mas importantes en la predicción y clasificación de los índices de contaminación.

### 1.5.1. Medidas de centralidad

En el análisis de gráficas, las medidas de centralidad son herramientas importantes para identificar y evaluar la importancia relativa de los vértices dentro de una red. Estas medidas permiten entender mejor la estructura y la dinámica de la red. En esta sección, se presentan las principales métricas de centralidad que abordamos en este trabajo, tales como intermediación [2], cercanía [2], grado [23], Katz [22] y PageRank [16].

La **Centralidad de Intermediación** de un vértice $v_i$, denotada como $cen_b(v_i)$, se define como:

$$cen_b(v_i) = \sum_{v_k, v_i \in V(G)} \frac{b_{jik}}{b_{jk}},$$

donde $b_{jik}$ es el número de geodésicas desde el vértice $v_j$ hasta el vértice $v_k$ que pasan por el vértice $v_i$ y $b_{jk}$ es el número de todas las $v_j v_l$-geodésicas

La **Centralidad de Cercanía** de un vértice $v$ de una gráfica $G$, denotada como $cen(v)$, se define como el inverso de la distancia promedio $l(v)$, es decir:

$$cen(v) = \frac{1}{l(v)}.$$

En una gráfica la **Centralidad de Grado** de un vértice $v$ es una medida que refleja la importancia de $v$ en función del número de aristas que inciden en él. Formalmente, la centralidad de grado $C_D(v)$ de un vértice $v$ se define como el grado de $v$ dividido entre $|V(G)| - 1$. Si $\deg(v)$ denota el grado de $v$, entonces:

$$cen_{deg}(v) = \frac{\deg(v)}{|V(G)| - 1}.$$

Observa que un vértice una alta centralidad de grado es aquel que tiene muchas conexiones directas con otros vértices, lo que puede indicar una posición más influyente o central en la red.

La **Centralidad de Katz** de un vértice $v$ mide su influencia considerando tanto las conexiones directas como las indirectas. Se define como:

$$C_K(v) = C_{\text{Katz}}(i) = \sum_{k=1}^{\infty} \sum_{j=1}^{n} \alpha^k (A^k)_{ji},$$

donde

- $A$ es la matriz de adyacencia de la gráfica,
- $\alpha$ es un parámetro que controla la influencia de las de los vértices dependiendo de la distancia, con $0 < \alpha < 1$.

Esta medida asigna mayor centralidad a los vértices que están cerca de $v$ y considera la influencia de los caminos indirectos.

El **PageRank** es un algoritmo que tiene el objetivo de medir la importancia o relevancia relativa de un vértice en una gráfica o digráfica en función de sus ex-grados o in-grados. El algoritmo consiste en lo siguiente:

1. Si $|V(D)| = n$, asignamos a cada vértice un PageRank inicial de $1/n$.
2. Cada vértice divide su PageRank entre sus flechas salientes y suma este valor al PageRank de sus ex-vecinos. Si $d^+(\mathcal{C}_i) = 0$ entonces mantiene su PageRank el vértice $\mathcal{C}_i$.
3. Repetir los pasos anteriores hasta que el PageRank de todos los vértices se estabilice.

## 1.6. Técnica de Sobremuestreo Sintético para Minorías (SMOTE)

La Técnica de Sobremuestreo Sintético para Minorías (**SMOTE** por sus siglas en inglés) es una técnica utilizada para abordar el desequilibrio de clases en conjuntos de datos de Aprendizaje Automático. En muchos problemas del mundo real, como la detección de fraudes, el diagnóstico médico y la predicción de fallos, los datos suelen estar desbalanceados, lo que significa que una clase es mucho más frecuente que la otra. En tales casos, los modelos de Aprendizaje Automático pueden sesgarse hacia la clase mayoritaria, lo que resulta en una baja precisión para la clase minoritaria.

SMOTE es una técnica de sobremuestreo que aborda este problema mediante la generación sintética de muestras para la clase minoritaria, lo que equilibra la distribución de clases en el conjunto de datos. El algoritmo funciona de la siguiente manera [8]:

- **Identificación de la clase minoritaria**: Primero, se identifica la clase minoritaria en el conjunto de datos, es decir, la clase con menos ejemplos.
- **Selección de vecinos**: Para cada instancia en la clase minoritaria, se seleccionan k vecinos más cercanos basados en alguna métrica de distancia, como la distancia euclidiana en el espacio de características.
- **Generación de instancias sintéticas**: Para cada instancia de la clase minoritaria, se selecciona aleatoriamente uno de sus $k$ vecinos más cercanos. Luego, se genera una nueva instancia sintética a lo largo de la línea que conecta la instancia original y su vecino seleccionado. La ubicación de la nueva instancia sintética se determina multiplicando la diferencia entre la instancia original y su vecino por un valor aleatorio entre 0 y 1, y luego sumando este resultado a la instancia original.
- **Incorporación de instancias sintéticas**: Se repite el paso anterior hasta que se haya generado un número específico de instancias sintéticas para igualar el tamaño de la clase mayoritaria. Estas instancias sintéticas se agregan al conjunto de datos original.

Un ejemplo lo podemos observar en la Figura 1.2

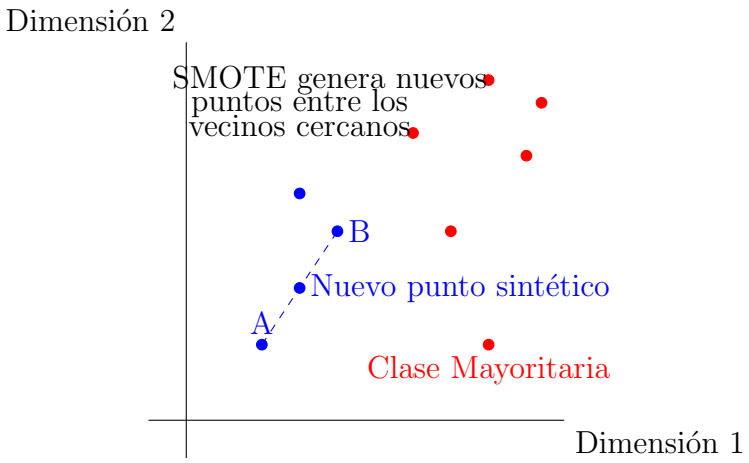



Figura 1.2: Ejemplo de generación de datos con SMOTE

## 1.7. Distancia del Movimiento de Tierra (EMD)

La Distancia del Movimiento de Tierra (EMD, por sus siglas en inglés), también conocida como Distancia Wasserstein. Esta distancia es una medida utilizada para comparar dos distribuciones de probabilidad. Un rasgo particular de esta métrica es que considera la cantidad de *"trabajo"* necesario para transformar una distribución en otra. En nuestro caso, podemos extrapolar estas ideas para el medir el *"esfuerzo"* o similitud para transformar una serie de tiempo en la otra.

La EMD se basa en el problema del transporte óptimo que, intuitivamente, consiste en tener dos montones de tierra que tienen distintas distribuciones de probabilidad, las cuales están asociadas a dos series de tiempos para las que se desea transferir tierra de un montón al otro de la manera más eficiente posible. La EMD calcula la mínima cantidad de trabajo necesaria para completar esta tarea, y cada unidad de tierra tiene un costo asociado con su movimiento desde un montón al otro.

Matemáticamente, la $EMD$ se define como:

$$EMD(P,Q) = min \sum c(i,j) * f(i,j), \tag{1.8}$$

donde:

- $P$ y $Q$ son las dos distribuciones de probabilidad que se comparan.
- $c(i, j)$ es una función de costo que especifica el costo de mover una unidad de masa de la posición $i$ en la distribución $P$ a la posición $j$ en la distribución $Q$.
- $f(i, j)$ es una matriz que indica cuánta masa se mueve de $i$ a $j$.

## 1.8. Análisis de datos

En esta sección se presentan los métodos que se utilizarán para analizar, limpiar y preparar los datos con el objetivo de desarrollar una base de datos confiable, lo que permitirá obtener mejores resultados.

### 1.8.1. Regla empírica

La regla empírica, también conocida como la regla 68-95-99.7, es un método utilizado para la detección de valores atípicos en conjuntos de datos. Esta regla establece que aproximadamente el 68 % de los valores se encuentran dentro de una desviación estándar de la media, el 95 % dentro de dos desviaciones estándar y el 99.7 % dentro de tres desviaciones estándar. En otras palabras establece lo siguiente:

- Aproximadamente el 68 % de los datos están dentro de una desviación estándar de la media.

$$P(\mu - \sigma \leq X \leq \mu + \sigma) \approx 0{,}68$$

- Alrededor del 95 % de los datos están dentro de dos desviaciones estándar de la media.

$$P(\mu - 2\sigma \leq X \leq \mu + 1\sigma) \approx 0{,}95$$

- Casi el 99.7 % de los datos están dentro de tres desviaciones estándar de la media.

$$P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) \approx 0{,}997$$

### 1.8.2. K Vecinos Más Cercanos (KNN)

El algoritmo K Vecinos Más Cercanos (KNN, por sus siglas en inglés) es un método de aprendizaje supervisado que puede ser utilizado tanto para

clasificación como para regresión o completar información de de una base da datos. Para utilizar KNN en problemas de regresión se puede seguir los siguientes pasos [1, 20]

1. **Seleccionar el valor de K:** Elige el número de vecinos más cercanos, K, que se utilizará para hacer la predicción.

2. **Calcular la distancia:** Calcular la distancia de todos los elementos con nuestro punto, tenemos que haber definido la métrica que vamos a utilizar para calcular esta distancia. En particular, si trabajas con valores numéricos es conveniente la distancia euclídeana.

   Dados dos puntos $(p = (p_1, p_2, \ldots, p_n))$ y $(q = (q_1, q_2, \ldots, q_n))$, la distancia euclidiana entre los dos puntos $(p)$ y $(q)$ se define como la longitud del segmento de línea que conecta los dos puntos.

   $$d(p, q) = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}$$

3. **Selección de vecinos:** Coloca todas las distancias de nuestro punto con el resto y ordénalos de menor a mayor. Se elige $K$ como el número de vecinos más cercanos.

4. **Predecir o completar:** Para regresión o completar informacón, la predicción se obtiene promediando los valores de salida (valores numéricos) de los $K$ vecinos seleccionados. Por último se asigna el valor promedio calculado al punto nuevo como la predicción final.

## 1.9. Perceptrones Multicapa (MLP)

Un **Perceptrón Multicapa** (MLP, por sus siglas en inglés) es un tipo de red neuronal artificial que consta de múltiples capas de neuronas organizadas de manera secuencial, es decir, cada capa está completamente conectada a la siguiente. Esta estructura permite al MLP modelar relaciones no lineales en los datos, lo que lo hace particularmente útil en tareas de clasificación y regresión.

1. **Vector de entrada**: Al principio, se define un **vector de entrada**, que es un conjunto de valores numéricos, que representan las características. Denotamos el vector de entrada como:

$$\mathbf{x} = (x_1, x_2, \ldots, x_n)$$

donde $x_i$ representa el valor de la $i$-ésima característica, y $n$ es el número total de características.

2. **Capas ocultas:** Las **capas ocultas** son las encargadas de procesar la información mediante operaciones no lineales. Cada neurona en una capa oculta realiza dos operaciones principales:

   *a*) **Suma ponderada**:

   Cada neurona toma una suma ponderada de las salidas de las neuronas de la capa anterior y agrega un sesgo:

$$z_j^{(l)} = \sum_{i=1}^{n^{(l-1)}} W_{ji}^{(l)} a_i^{(l-1)} + b_j^{(l)}$$

   donde:

   - $z_j^{(l)}$ es la preactivación de la neurona $j$ en la capa $l$.
   - $W_{ji}^{(l)}$ es el peso que conecta la neurona $i$ en la capa $l-1$ con la neurona $j$ en la capa $l$.
   - $a_i^{(l-1)}$ es la activación de la neurona $i$ en la capa $l-1$.
   - $b_j^{(l)}$ es el sesgo de la neurona $j$ en la capa $l$.

   *b*) **Función de activación**

   Después de calcular la suma ponderada $z_j^{(l)}$, la neurona aplica una **función de activación** $g(\cdot)$ para introducir no linealidad:

$$a_j^{(l)} = g(z_j^{(l)})$$

   Las funciones más comunes:

   - **ReLU**: $g(z) = \text{máx}(0, z)$
   - **Tanh**: $g(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$
   - **Sigmoide**: $g(z) = \frac{1}{1+e^{-z}}$

- **Softmax**: $g(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}}$

3. **Capa de salida**: La **capa de salida** transforma la información procesada por las capas ocultas en un vector de salida útil para la tarea específica. En este caso se realizan las mismas tareas de la **suma ponderada** y **gunción de activación**

La **función de costo** mide la diferencia entre las predicciones de la red y los valores reales. Algunas funciones de costo comunes son:

**Error cuadrático medio**

Para problemas de regresión, se utiliza el error cuadrático medio:

$$J(\mathbf{W}, \mathbf{b}) = \frac{1}{2m} \sum_{i=1}^{m} \left(\hat{y}^{(i)} - y^{(i)}\right)^2$$

donde $m$ es el número de ejemplos en el conjunto de entrenamiento.

**Entropía cruzada**

Para problemas de clasificación, se utiliza la entropía cruzada:

$$J(\mathbf{W}, \mathbf{b}) = -\frac{1}{m} \sum_{i=1}^{m} \sum_{k=1}^{K} y_k^{(i)} \log(\hat{y}_k^{(i)})$$

### 1.9.1. Propagación hacia atrás

La **propagación hacia atrás** es un algoritmo de optimización que ajusta los pesos $\mathbf{W}$ y sesgos $\mathbf{b}$ de la red en función del error calculado en la capa de salida. Este proceso se realiza calculando gradientes, que permiten evaluar cómo deben ajustarse los parámetros para reducir el error.

**Definición del error en la capa de salida**

El primer paso es calcular el error en la capa de salida. Para cada neurona de salida $k$, el error se define como la diferencia entre el valor de salida de la

red y el valor real:

$$\delta_k^{(L)} = \frac{\partial J}{\partial z_k^{(L)}} = \frac{\partial J}{\partial \hat{y}_k} \cdot g'(z_k^{(L)})$$

donde:

- $J$ es la función de costo.
- $\hat{y}_k$ es la salida predicha por la neurona $k$.
- $z_k^{(L)}$ es la preactivación de la neurona $k$ en la capa de salida $L$.
- $g'$ es la derivada de la función de activación.

Por ejemplo, en el caso de la función de costo de entropía cruzada combinada con la función de activación softmax, el término $\frac{\partial J}{\partial \hat{y}_k}$ se simplifica a:

$$\delta_k^{(L)} = \hat{y}_k - y_k$$

**Propagación del error a las capas anteriores**

Para cada capa $l = L-1, L-2, \ldots, 1$, calculamos el error de cada neurona utilizando el error de la capa siguiente. La fórmula para el error de la neurona $j$ en la capa $l$ es:

$$\delta_j^{(l)} = \left( \sum_k W_{kj}^{(l+1)} \delta_k^{(l+1)} \right) g'(z_j^{(l)})$$

donde:

- $W_{kj}^{(l+1)}$ es el peso que conecta la neurona $j$ en la capa $l$ con la neurona $k$ en la capa $l+1$.
- $\delta_k^{(l+1)}$ es el error de la neurona $k$ en la capa siguiente.
- $g'$ es la derivada de la función de activación utilizada en la capa $l$.

### Cálculo del Gradiente de los pesos y sesgos

Una vez que tenemos el error en cada capa, calculamos los gradientes de los pesos y sesgos para actualizar los parámetros.

Para el peso $W_{ji}^{(l)}$, el gradiente es:

$$\frac{\partial J}{\partial W_{ji}^{(l)}} = \delta_j^{(l)} a_i^{(l-1)}$$

donde:

- $\delta_j^{(l)}$ es el error en la neurona $j$ de la capa $l$.
- $a_i^{(l-1)}$ es la activación de la neurona $i$ en la capa anterior.

Para el sesgo $b_j^{(l)}$, el gradiente es:

$$\frac{\partial J}{\partial b_j^{(l)}} = \delta_j^{(l)}$$

### Actualización de pesos y sesgos

Una vez calculados los gradientes, se procede a actualizar los pesos y sesgos mediante el método de descenso de gradiente. La actualización para cada peso y sesgo en la capa $l$ es:

$$W_{ji}^{(l)} := W_{ji}^{(l)} - \eta \frac{\partial J}{\partial W_{ji}^{(l)}}$$

$$b_j^{(l)} := b_j^{(l)} - \eta \frac{\partial J}{\partial b_j^{(l)}}$$

donde:

- $\eta$ es la tasa de aprendizaje, que controla el tamaño de los pasos que tomamos en cada iteración de optimización.

Este proceso se repite en todas las capas y para todos los pesos y sesgos en la red, permitiendo así que la red aprenda a ajustar sus parámetros para minimizar la función de costo y mejorar su precisión en la tarea deseada.

### Validación cruzada

La validación cruzada $k$-fold es una técnica estadística utilizada para evaluar el rendimiento de un modelo de Aprendizaje Automático y su capacidad de generalización a datos no vistos. Consiste en dividir el conjunto de datos disponible en $k$ subconjuntos de tamaño aproximadamente igual, llamados "folds.º pliegues. El proceso sigue los siguientes pasos:

1. **División de los datos**: El conjunto de datos se divide aleatoriamente en $k$ pliegues.
2. **Iteraciones de entrenamiento y validación**: Para cada iteración $i$ de 1 a $k$:
   - Se utiliza el pliegue $i$ como conjunto de validación.
   - Se entrena el modelo usando los $k - 1$ pliegues restantes como conjunto de entrenamiento.
3. **Cálculo de métricas**: Se evalúa el rendimiento del modelo en el pliegue de validación, obteniendo una métrica de evaluación (por ejemplo, exactitud, error cuadrático medio, etc.).
4. **Agregación de resultados**: Al finalizar las $k$ iteraciones, se calcula la media de las métricas obtenidas en cada pliegue, proporcionando una estimación más robusta del rendimiento del modelo.

Esta metodología ayuda a mitigar el sesgo que podría introducir una sola división de entrenamiento y prueba, y proporciona una estimación más confiable del desempeño del modelo en datos no vistos.

# Capítulo 2

# Preparación de datos

Este capítulo comprende la primera parte del segundo paso de la metodología del Proceso Estándar Cruzado para el Aprendizaje Automático con Aseguramiento de Calidad (CRISP-ML(Q), por sus siglas en inglés) (para más detalles, véase la sección 1.2), donde se presentará el procedimiento para la limpieza de datos, comprensión de la naturaleza de los datos y el conocimiento de las variables. Este proceso consiste en analizar la calidad de los datos, lo cual incluye detectar muestras con fechas y horas duplicadas o faltantes, identificar valores atípicos, visualizar posibles relaciones entre las variables mediante gráficas y crear las clases del índice Aire Salud. Es importante señalar que para llevar a cabo esta investigación, se recopilaron los datos proporcionados por la Universidad Iberoamericana.

## 2.1. Descripción de los datos

Los datos proporcionados por la Universidad Iberoamericana se encuentran en archivos de texto (.txt) que están organizados bajo diferentes nombres base: **One Day, Gases, One Min, Partículas y Ten Min**. Cada uno de estos grupos de archivos contiene información relacionada con distintas variables meteorológicas y de contaminación ambiental de la parte Oriente de la Ciudad de México, proporcionando un panorama general sobre las condiciones atmosféricas y los niveles de contaminantes.

Además, las muestras dentro de estos archivos fueron recolectadas empleando diversas metodologías. Por ejemplo, se cuenta con promedios (Avg.), acumulados (Tot.), mediciones puntuales (Smp.), desviaciones estándar (Std.), valores máximos (Max.), valores mínimos (Min.), fecha y hora de registro máximo (TMx.) o fecha y hora de registro mínimo (TMn.), como se detalla en el Cuadro 2.1. Por otro lado, también se observó que hay variables que se tomaron dos veces en distintos grupos de archivo, por lo que conservaremos aquellas que contengan la mayor cantidad de datos. Estas variables están señaladas con el símbolo *, ya que presentan un mayor contidad de datos a lo largo del tiempo.

Para la limpieza de los datos, el proceso se dividió en cinco etapas:

1. Concatenar los archivos según su tipo, ya sea One Day, Gases, One Minute, Partículas o Ten Minutes, e identificar las fechas faltantes y duplicadas en cada uno de ellos.
2. Identificar los valores atípicos y eliminarlos en el análisis de datos, asegurando que estos valores no afecten negativamente los resultados.
3. Completar la información faltante de cada tipo de archivo de manera individual, manteniendo las tendencias en las mediciones de cada archivo.
4. Unir los diferentes tipos de archivos en un único DataFrame, verificando que los datos sean consistentes tanto en fechas como en horarios, y que no haya inconsistencias entre los distintos tipos de archivo.
5. Finalmente, completar la información faltante dentro del DataFrame consolidado, garantizando que todas las entradas tengan datos completos y confiables para la investigación.

## 2.2. Limpieza de datos de datos

En esta sección se describe detalladamente el proceso de limpieza y preparación de los datos utilizados en este proyecto, con el objetivo de obtener una base de datos sin valores atípicos, sin valores nulos o faltantes, sin datos duplicados y unificada en un único archivo csv.

Cuadro 2.1: Información condensada de los datos provenientes de archivos txt. Aquellas variables que se hayan tomado dos veces, únicamente nos quedaremos con aquella contenga mayor cantidad de datos. Aquellas variables que serán eliminadas estarán marcadas con el símbolo *.

| Variables | One Day | Gases | One Min. | Partículas | Ten Min |
|---|---|---|---|---|---|
| Irradiación $\left(\frac{MJ}{m^2}\right)$ | Tot | | | | Tot* |
| Precipitación $(mm)$ | Tot | | | | Tot* |
| Hrs. Sol (hrs.) | Tot | | | | |
| Temp. amb. máx. (°C) | Max | | | | |
| Hr. temp. máx. $(Hrs.)$ | TMx | | | | |
| Temp. mín. (°C) | Min | | | | |
| Hr. temp. mín. $(Hrs.)$ | TMn | | | | |
| $NO_2$ $(ppm)$ | | Smp | | | |
| $SO_2$ $(ppm)$ | | Smp | | | |
| $CO$ $(ppm)$ | | Smp | | | |
| $O_3$ $(ppm)$ | | Smp | | | |
| Temp. del aire (°C) | | Smp | | | |
| Presión atmosférica $(mbar)$ | | Smp* | Avg | | |
| Humedad relativa ( %) | | Smp | Avg* | | |
| Irradiancia $\left(\frac{w}{m^2}\right)$ | | | Avg | | |
| Temp. amb. (°C) | | | Avg | | |
| $PM_{10}$ $\left(\frac{\mu g}{m^3}\right)$ | | | | Smp | |
| $PM_{2,5}$ $\left(\frac{\mu g}{m^3}\right)$ | | | | Smp | |
| Potencial irradiancia $\left(\frac{w}{m^2}\right)$ | | | | | Avg |
| Voltaje de la batería de respaldo Mínima $(Volt.)$ | | | | | Tot |
| Punto de rocío (°C) | | | | | Avg |
| Temp. bulbo húmedo (°C) | | | | | Avg |
| Índice calor (°C) | | | | | Avg |
| Enfr. viento (°C) | | | | | Avg |
| Presión de saturación de vapor $(Pa.)$ | | | | | Avg |
| Vel. de viento $\left(\frac{km}{h}\right)$ | | | | | Avg |
| Vel. de viento $\left(\frac{km}{h}\right)$ | | | | | Std |
| Vel. máx. viento $\left(\frac{km}{h}\right)$ | | | | | Max |
| Dirección de viento (°) | | | | | SMM |
| Dirección de viento (°) | | | | | Avg |
| Dirección de viento (°) | | | | | Std |

### 2.2.1. Completar información de la base de datos

Para llevar a cabo el análisis, se comenzó uniendo los diferentes archivos txt que compartían el mismo nombre base, pero que diferían en las fechas, las cuales indicaban el día en que se recolectaron los datos. Por ejemplo, si contamos con cuatro archivos distintos: Gases_1_4_2023, Gases_1_5_2023, TenMinutes_1_4_2023 y TenMinutes_1_5_2023, en este caso, los archivos con la base de nombre "Gases" se combinarían en uno solo, sin importar la fecha, y lo mismo se haría con los archivos cuyo nombre base es "TenMinutes". Es importante aclarar que cada archivo txt contiene información correspondiente a un día específico, a excepción de los archivos de "One Day", que abarcan datos de un mes entero.

Después de haber concatenado cada uno de los archivos en su respectivo grupo, se realizó una verificación para identificar fechas faltantes y duplicadas. Esto pudo deberse, probablemente, a motivos de mantenimiento, vacaciones o errores de medición, ya que los registros muestran inconsistencias, como se presenta en el Cuadro 2.2. Es importante señalar que estos datos fueron calculados considerando los diferentes inicios de medición; es decir, cada tipo de archivo comenzó a ser medido en fechas distintas, y a partir de esto se calcularon las fechas faltantes y duplicadas. Desde esta perspectiva, se agregaron las fechas que faltaban en cada archivo y se eliminaron las fechas duplicadas, conservando los datos que reflejan las tendencias correspondientes a cada una de las variables.

Cuadro 2.2: Cantidad de fechas faltantes, duplicadas y total de datos de cada uno de los archivos antes de la eliminación de datos en los rangos de fechas no coincidentes.

| Archivo | Fechas faltantes | Fechas duplicadas | Total de datos |
|---|---|---|---|
| One Day | 18 | 30 | 1,571 |
| Gases | 4,614 | 18 | 1,989,314 |
| One Minute | 5,875 | 132 | 2,283,282 |
| Partículas | 476 | 0 | 198,916 |
| Ten Minutes | 976 | 8 | 227,940 |

Por otra parte, para la detección de valores atípicos para cada conjunto de datos se utilizó la regla de la empírica. Este método consiste en establecer un umbral que determina qué valores se consideran atípicos en función de su

distancia con respecto a la media y la desviación estándar de cada una de las variables como se vió en el capítulo 1.8.1. La detección de valores atípicos mediante esta regla es eficiente y fácil de implementar, lo que la convierte en una técnica ampliamente utilizada en diversos campos [29]. En este caso, en el Cuadro 2.3 se muestran la cantidad de datos atípicos que se detectaron en cada una de las variables meteorológicas y de contaminación.

Adicionalmente, para agregar los datos faltantes que se generaron de la detección de valores atípicos y los que teníamos originalmente, se utilizaron técnicas de Aprendizaje Automático. Ejemplos de estos métodos incluyen el uso de Perceptrón Multicapa (MLP, por sus siglas en inglés), mapas de autoorganización (SOM, por sus siglas en inglés) y k-vecinos más cercanos (KNN, por sus siglas en inglés) que son las más apropiadas para abordar la imputación de valores faltantes, generando una mejora significativa en la precisión del pronóstico en comparación con enfoques estadísticos.[20].

Es importante destacar que el algoritmo KNN resulta muy conveniente aplicarlo en este contexto debido a la baja proporción de datos faltantes en comparación con el conjunto total de cada archivo y sus fechas correspondientes. Además, cabe destacar que KNN puede predecir tanto atributos cualitativos (el valor más frecuente entre los k-vecinos más cercanos) como atributos cuantitativos (la media entre los k-vecinos más cercanos). Otra ventaja de este algoritmo es que no hay necesidad de crear un modelo predictivo para cada atributo con datos faltantes [1, 20].

Aunque el algoritmo KNN se explica en 1.8.2, en esta sección proporcionamos una breve explicación sobre su aplicación en la imputación de valores, destacando que este método es particularmente adecuado debido a la baja proporción de datos faltantes. En el Cuadro 2.4 se muestra un ejemplo de la imputación de valores con el algoritmo KNN. En este caso se utiliza un parámetro de vecinos igual a 2 para identificar los dos vecinos más cercanos. Los valores originales e imputados se muestran en la Figura 2.1 para visualizar la efectividad del proceso de imputación.

Observamos que en el Cuadro 2.3 hay dos variables que no presentaron valores atípicos, Hr. temp. máx y mín. Estas dos características se eliminaron del análisis, ya que el algoritmo de KNN no contempla datos temporales.

Cuadro 2.3: Número total de datos atípicos, datos faltantes y la cantidad de datos totales para cada uno de las variables meteorológicas.

| Variables | Datos atípicos | Datos faltantes | Cantidad de datos |
|---|---|---|---|
| Irradiación $\left(\frac{MJ}{m^2}\right)$ | 6.89 % | 1.14 % | 227940 |
| Precipitación (mm) | 0.99 % | 1.14 % | 227940 |
| Hrs. Sol (hrs.) | 4.45 % | 1.14 % | 1571 |
| Temp. amb. máx. (°C) | 4.64 % | 1.14 % | 1571 |
| Hr. temp. máx. (Hrs.) | 0.00 % | 1.14 % | 1571 |
| Temp. mín. (°C) | 5.28 % | 1.14 % | 1571 |
| Hr. temp. mín. (Hrs.) | 0.00 % | 1.14 % | 1571 |
| $NO_2$ (ppm) | 4.85 % | 0.23 % | 1967575 |
| $SO_2$ (ppm) | 2.26 % | 0.23 % | 1967575 |
| $CO$ (ppm) | 4.78 % | 0.23 % | 1967575 |
| $O_3$ (ppm) | 0.00 % | 0.23 % | 1967575 |
| Temp. del aire (°C) | 0.17 % | 0.23 % | 1967575 |
| Presión atmosférica (mbar) | 0.25 % | 0.23 % | 1967575 |
| Humedad relativa ( %) | 0.83 % | 0.25 % | 2283282 |
| Irradiancia $\left(\frac{w}{m^2}\right)$ | 7.32 % | 0.25 % | 2283282 |
| Temp. amb. (°C) | 4.31 % | 0.25 % | 2283282 |
| $PM_{10}$ | 2.40 % | 0.23 % | 196741 |
| $PM_{2,5}$ | 2.58 % | 0.23 % | 196741 |
| Potencial irradiancia $\left(\frac{w}{m^2}\right)$ | 0.00 % | 0.42 % | 227940 |
| Voltaje de la batería de respaldo Mínima (Volt.) | 4.96 % | 0.42 % | 227940 |
| Punto de rocío (°C) | 4.31 % | 0.42 % | 227940 |
| Temp. bulbo húmedo (°C) | 3.63 % | 0.42 % | 227940 |
| Indice calor (°C) | 4.31 % | 0.42 % | 227940 |
| Enfr. viento (°C) | 4.82 % | 0.42 % | 227940 |
| Presión de saturación de vapor (Pa.) | 4.62 % | 0.42 % | 227940 |
| Vel. de viento $\left(\frac{km}{h}\right)$ | 3.56 % | 0.42 % | 227940 |
| Vel. de viento $\left(\frac{km}{h}\right)$ | 3.37 % | 0.42 % | 227940 |
| Vel. máx. viento $\left(\frac{km}{h}\right)$ | 3.78 % | 0.42 % | 227940 |
| Dirección de viento (∘) | 0.00 % | 0.42 % | 227940 |
| Dirección de viento (∘) | 0.18 % | 0.42 % | 227940 |
| Dirección de viento (∘) | 3.23 % | 0.42 % | 227940 |

Cuadro 2.4: Ejemplo para imputar valores con KNN

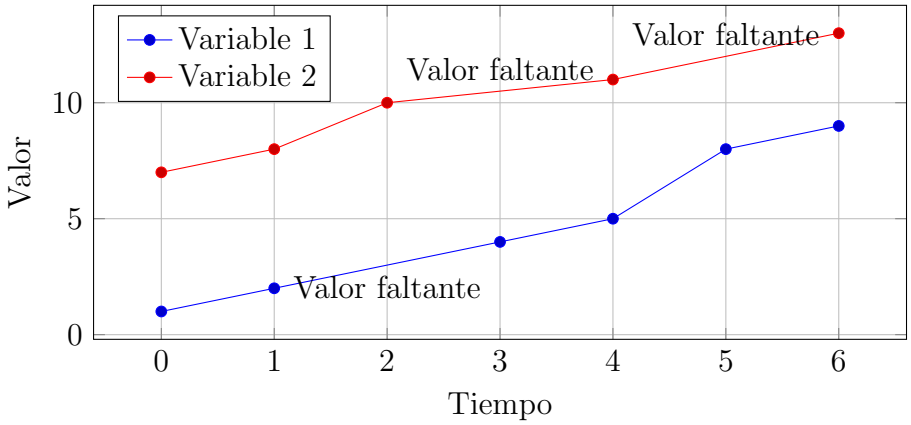
| Datos antes de la imputación | | Datos después de la imputación | |
|---|---|---|---|
| Variable 1 | Variable 2 | Variable 1 | Variable 2 |
| 1.0 | 7.0 | 1.0 | 7.0 |
| 2.0 | 8.0 | 2.0 | 8.0 |
| **NaN** | 10.0 | 3.5 | 10.0 |
| 4.0 | **NaN** | 4.0 | 9.5 |
| 5.0 | 11.0 | 5.0 | 11.0 |
| 8.0 | **NaN** | 8.0 | 12.0 |
| 9.0 | 13.0 | 9.0 | 13.0 |

### 2.2.2. Concatenar y completar de los archivos unidos

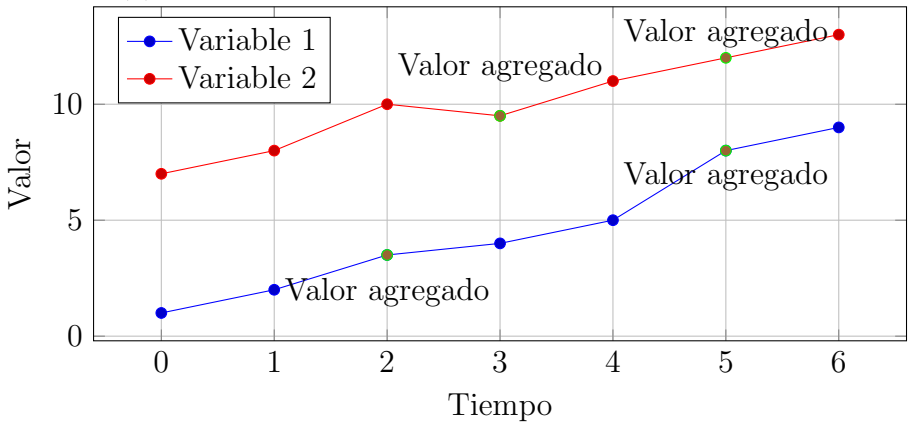
Hasta este punto se han realizado diversas actividades como agregar las fechas y horas faltantes, eliminar las fechas duplicadas, detectar y eliminar los valores atípicos e imputar los datos vacíos para cada archivo. En este apartado se dará a conocer el procedimiento que se realizó para concatenar todos los archivos, a su vez, detectar y completar los intervalos de los datos vacíos de cada uno de las variables.

Para realizar la unión de los archivos, se procedió a combinar todas las fechas y horas comunes, en caso en donde las variables no coincidían en intervalos de fechas y horas, se consideraron esos casos como datos faltantes. Por ejemplo, considerando dos tipos de archivos distintos, $A_1$ y $A_2$, si una fecha está presente en el archivo $A_1$ pero no en el archivo $A_2$, se incorporan esas filas a las columnas del archivo $A_2$ con valores nulos en el conjunto fusionado, y viceversa. Este proceso se implementó con el objetivo de preservar la máxima cantidad de información registrada por los sensores de la Universidad Iberoamericana, como se ilustra en el Cuadro 2.5.

Posteriormente, se procedió a eliminar las fechas iniciales de cada variable hasta que coincidieran sus fechas de inicio tal como se detalla en el Cuadro 2.6. Esta acción se implementó con el objetivo de mitigar sesgos en las variables que presentaban una menor cantidad de datos o cuyo inicio estaba notablemente distante de la fecha más antigua de algún archivo. Antes de esta eliminación, se contaba con 2,289,157 fechas distintas, mientras que, después del proceso, quedaron 1,993,532. Esto implica que se conservó el 87.08 % de




(a) Series antes de la imputación de valores con KNN.




(b) Series después de la imputación de valores con KNN.

Figura 2.1: Gráfricas del ejemplo de imputación de valores del Cuadro 2.4.

las fechas.

Después de concatenar todos los archivos en un único conjunto de datos, se procedió a aplicar tres enfoques distintos para completar los valores para cada característica, distribuidas en tres conjuntos con variables diferentes. El primer enfoque abarca todas las características, con la excepción de tres: “Hrs. Sol”, “Temp. amb. mán.” y “Temp. amb. mín.”. En el segundo, se considera exclusivamente la característica “Hrs. Sol”. Finalmente, el tercer se encuentran los elementos “Temp. amb. máx.” y “Temp. amb. mín.”. Los dividimos de esta manera ya que “Hrs. Sol”, “Temp. amb. máx.” y “Temp.

Cuadro 2.5: Ejemplo de cómo se combinaron los diversos archivos. En casos en los que un archivo no contiene las fechas y horas de todos los demás, se añadieron nuevas filas con valores nulos para las variables correspondientes.

| **Tiempo** | **PM 10** | **PM 2.5** | **NO2** | **SO2** | **CO** | **O3** |
|---|---|---|---|---|---|---|
| 2023-09-21 09:27:00 | NaN | NaN | 0.015 | 0.085 | 0.453 | 0.022 |
| 2023-09-21 09:28:00 | NaN | NaN | 0.015 | 0.084 | 0.451 | 0.025 |
| 2023-09-21 09:29:00 | NaN | NaN | 0.015 | 0.084 | 0.447 | 0.027 |
| 2023-09-21 09:30:00 | 0.7 | 11.7 | 0.015 | 0.083 | 0.442 | 0.027 |

Cuadro 2.6: Se puede observar que los archivos empiezan a coincidir en la fecha y hora de inicio de Partículas y todos tienen el mismo término, con excepción de One Minute.

| Tiempo | Tipo de archivo | Fecha/hora de inicio | Fecha/hora de término |
|---|---|---|---|
| 1 día | One Day | 17/05/19 0:00:00 | 21/09/23 0:00:00 |
| 1 min. | Gases | 06/12/19 17:23:00 | 21/09/23 9:30:00 |
| 1 min. | One Minute | 15/05/19 16:55:00 | 21/09/23 9:31:00 |
| 10 min. | Partículas | 06/12/19 17:40:00 | 21/09/23 9:30:00 |
| 10 min. | Ten Minutes | 15/05/19 17:00:00 | 21/09/23 9:30:00 |

amb. mín." tienen registros por día, lo que ocasiona que sean los que tienen menor cantidad de datos respecto al total, es decir, necesitamos utilizar otros métodos para completar su información.

Para el primer grupo, se aplicó la interpolación spline, dado que presenta un buen rendimiento al completar datos faltantes sobre la calidad del aire en intervalos de entre una a dos horas, pero su desempleño disminuye considerablemente a medida que aumenta la longitud de los intervalos vacíos [21]. Este enfoque divide nuestra curva en segmentos suaves entre los intervalos y cada uno de estos está definido por una función polinomial de grado bajo. Estos segmentos se unen de forma continua, y la suavidad de la curva depende del orden del método, lo que ayuda a mantener las tendencias y patrones presentes en los datos. En este caso, utilizamos una variante lineal, que se basa en conectar los puntos conocidos mediante líneas rectas y usar estas líneas para predecir los valores intermedios. Esta aproximación fue seleccionada porque el intervalo más grande de tiempo con datos faltantes es de 10 minutos, lo

que nos indica que existe poca variabilidad de los datos en estos intervalos. En la Figura 2.2, se muestra un ejemplo de la interpolación aplicada.

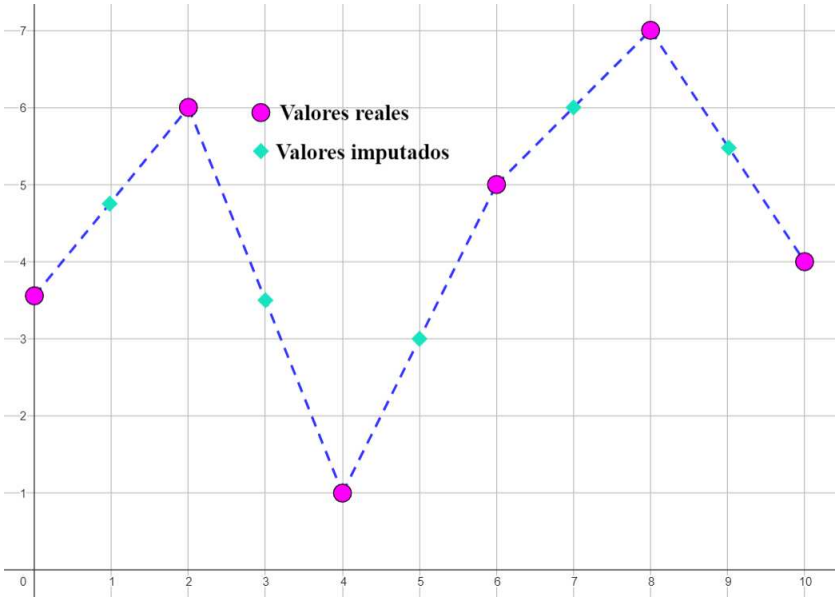



Figura 2.2: Ejemplo de obtener los valores imputados con la interpolación spline.

Para el segundo caso, "Hrs. Sol " se determinó que tiene una estrecha relación con otras características con comportamiento parabólico, entonces este debería tener una tendencia similar. Para determinar estas curvas en todos los días se hicieron dos supociciones:

- La primera consistió en establecer que la variable comenzaría en el origen cada día.
- La segunda asumió que alcanzaría su valor máximo $\mu$ de manera aleatoria entre las 12:00 hrs. y las 14:00 hrs. aleatoriamente.

En el Cuadro 2.1 observamos en la columna de One Day, esta variable es acumulativa, lo que implica que sus valores de un día se suman y posteriormente se reinicia al iniciar el siguiente. Definimos el valor de "Hrs. Sol" como $\alpha$ y se define como la integral de la función cuadrática $f(x) = ax^2 + bx + c$. Por lo tanto, tenemos que

$$\alpha = \int_0^{1440} f(x)\,dx \tag{2.1}$$

$$= \frac{ax^3}{3} + \frac{bx^2}{2} + cx\Big|_0^{1440} \tag{2.2}$$

$$= \frac{a(1440)^3}{3} + \frac{b(1440)^2}{2} + 1440c \tag{2.3}$$

Recordemos que los límites de la integral estan definidos por los minutos que tiene un día, esta unidad de medida es la más pequeña dentro de nuestra base de datos.

Supongamos que nuestro minuto cero comienza exactamente a las 00:00 hrs. entonces se deduce que $c = 0$. Por otro lado, sabemos que en el punto $\mu$ alcanza su valor máximo, entonces $b = -2a\mu$. Es claro que si este último valor se sustituye en la ecuación 2.3 se tiene que

$$a = \frac{3\alpha}{(1441)^2(1440 - 3\mu)} \tag{2.4}$$

Si este resultado lo sustituímos en $b = -2a\mu$ se obtiene

$$b = -\frac{6\alpha}{(1441)^2(1440 - 3\mu)} \tag{2.5}$$

En el tercer grupo se encuentran las variables “Temp. amb. mán.” y “Temp. amb. mín.”, estas se calculan una vez por día. Dado que las temperaturas máxima y mínima de un día son únicas, resulta conveniente asignar el mismo valor a todas las instancias de un mismo día.

Después de la limpieza de los datos, se explicará la segunda parte del paso dos de la metodología CRISP-ML(Q), que incluye la creación de clases y la implementación de métodos para la selección de características, junto con sus respectivos resultados.

Los archivos de la base de datos en formato `.txt` utilizados en este proyecto están disponibles en el repositorio Zenodo bajo el título Muestra_Ibero.

# Capítulo 3

# Análisis de Aprendizaje Automático y teoría de gráficas

En este capítulo presentamos las metodologías empleadas para la creación de clases a partir de las concentraciones de contaminantes, las cuales son útiles para lograr una mayor exactitud en los algoritmos de aprendizaje supervisado, particularmente en redes neuronales. A continuación, exploraremos la aplicación de técnicas de selección de características en el análisis del Índice Aire Salud, integrando enfoques de Aprendizaje Automático y Teoría de Gráficas. Desde esta perspectiva, el Aprendizaje Automático no solo se utiliza para realizar predicciones y clasificaciones, sino también para la selección de variables (características), lo que mejora la exactitud y eficiencia de los modelos de aprendizaje supervisado, complementándose con Teoría de Gráficas, que permite modelar las interacciones entre variables y facilita el análisis de las relaciones presentes en los datos.

## 3.1. Creación de clases

Como se mencionó en el capítulo anterior, contamos con la información de las concentraciones para cada uno de los contaminantes, lo que nos permite generar las clases correspondientes al Índice Aire Salud. Este paso facilita la introducción de algoritmos de aprendizaje supervisado que se caracterizan por la necesidad de una variable dependiente, $Y$, la cuál es conocida por

el algoritmo, y que debe ser explicada a partir de un conjunto de variables independientes $\mathcal{X} = \{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_m\}$.

Para aplicar el aprendizaje supervisado, es útil definir una función que relacione las variables, la cual puede expresarse como:

$$Y = f(\mathcal{C}_i) + \epsilon.$$

Donde $\epsilon$ representa el ruido o error aleatorio, que puede estar asociado a errores en las mediciones o a la influencia de variables no consideradas en el modelo.

Dado que los datos se recopilan en intervalos de un minuto, asignamos una clase correspondiente a cada minuto para cada contaminante, lo cual consideramos como parte del Índice Aire Salud. Este enfoque nos permitirá clasificar con mayor exactitud los niveles de contaminación en cada minuto, proporcionando un análisis más detallado en comparación con otros métodos que utilizan mediciones por hora o en intervalos más amplios.

Al observar el Cuadro 1.2 de contingencia ambiental y los intervalos de concentración para cada contaminante (ver Cuadro 1.1), observamos que la contingencia se activa cuando se alcanzan los 150 puntos en el Índice Aire Salud, lo que corresponde a la clase roja (muy dañina). En este estudio, utilizaremos los intervalos de $PM_1(t)$ para ozono, así como los intervalos de concentración para los demás contaminantes presentes en el Cuadro 1.1. El conteo de la frecuencia de aparición de cada clase para cada contaminante se detalla en el Cuadro 3.1.

El total de clases obtenidas del Índice Aire Salud por minuto es de $1,993,532$ de los cuales el 90,74 % pertenecen a la clase 2, esto indica una clara sobrerepresentación de la clase 2 con respecto a las demás clases. Por otro lado, notamos que 0,49 % y 0,00125 % de los datos pertenecen a la clase 4 y 5, respectivamente, es decir, existe sub-representación de los datos.

Dado que las clases 4 y 5 son de gran importancia para su predicción y clasificación, ya que representan los niveles más altos de peligro, y considerando la insuficiencia de datos en la clase 5, decidimos combinarla con la clase 4 para mejorar los resultados de los modelos, ya que ambas clases implican la declaración de contingencia ambiental.

Cuadro 3.1: Total de clases para cada uno de los contaminantes. El índice total por minuto se toma como el valor máximo de la clase obtenida en ese minuto. Es decir, si en un tiempo determinado el contaminante $O_3$ alcanzó la clase 4 y todos los demás contaminantes se mantuvieron en clase 1, entonces el valor total del Índice Aire y Salud será de clase 4 para ese tiempo.

| **Clase** | $O_3$ | $NO_2$ | $SO_2$ | $CO$ | $PM_{10}$ | $PM_{2,5}$ | Índice Aire Salud por minuto |
|---|---|---|---|---|---|---|---|
| **1** | 1 714 444 | 1 993 532 | 134 973 | 1 993 532 | 1 991 066 | 1 993 532 | 86 904 |
| **2** | 238 189 | 0 | 1 789 040 | 0 | 2 466 | 0 | 1 802 919 |
| **3** | 31 056 | 0 | 69 519 | 0 | 0 | 0 | 93 866 |
| **4** | 9 818 | 0 | 0 | 0 | 0 | 0 | 9 818 |
| **5** | 25 | 0 | 0 | 0 | 0 | 0 | 25 |

Desde esta perspectiva, utilizamos la técnica SMOTE para generar datos sintéticos y equilibrar las clases. Específicamente, se crearon 5 veces más datos para las clases 1 y 3, mientras que para las clases 4 y 5, se generaron 45 veces más datos, como se muestra en el Cuadro 3.2. Es importante mencionar que no se buscó equilibrar todas las clases por completo, ya que generar una cantidad igual de datos para cada clase habría incrementado significativamente el costo computacional y el tiempo necesario para entrenar y procesar las redes neuronales.

La generación de datos, SMOTE permite mantener, en general, las distribuciones originales, como se muestra en [15]. Esto nos asegura que los datos sintéticos no distorsionan la información, es decir, conservan las propiedades del conjunto de datos real, mejorando el equilibrio entre las clases y, por ende, la exactitud de los modelos en su tarea de clasificación. La comparación de distribuciones entre los datos originales y los generados por SMOTE se muestran en el Anexo 5

Los programas desarrollados para la creación de clases y la generación de datos sintéticos con SMOTE están disponibles en el repositorio Zenodo bajo el título Limpieza de datos Maestría. El conjunto de datos completo con la limpieza de datos y las clases generadas se encuentra en el archivo `Data Original.csv`, y los datos sintéticos se encuentran en `SMOTE DATA.csv`.

Cuadro 3.2: Distribución de las clases antes y después de aplicar la técnica SMOTE. Se observa el incremento significativo en el número de muestras para las clases 1, 3, 4, que originalmente estaban subrepresentadas, con el fin de mejorar el equilibrio del conjunto de datos.

| **Clase** | **Total de datos originales** | **Total de datos con SMOTE** |
|---|---|---|
| **1** | 86 904 | 434 520 |
| **2** | 1 802 919 | 1 802 919 |
| **3** | 93 866 | 469 330 |
| **4** | 9 843 | 442 935 |
| **Cantidad de datos** | 1 993 532 | 3 149 704 |

## 3.2. Algoritmos Aprendizaje Automático

Para optimizar el rendimiento de nuestros modelos de clasificación, es importante seleccionar las características más relevantes, pues esto ayuda a reducir significativamente los tiempos de ejecución, minimizando la pérdida de información o evitando una reducción considerable en la exactitud (*accuracy*) de nuestros modelos. En las secciones 1.3 y 1.4, se presentan varios enfoques para la selección de características, cuyo objetivo principal es priorizar las características de mayor a menor relevancia, dependiendo de la propiedad que quieran priorizar. Estas listas se describen a continuación:

1. Varianza baja, MAD y DR: Nos interesa identificar las características con mayor dispersión. Por lo tanto, las características se ordenarán de mayor a menor en función de su varianza, (MAD), (DR), es decir, creamos tres listas de importancia, una por cada método.

2. LASSO: Aplicaremos el método LASSO para cada valor de $\alpha$ que se muestra en el Cuadro 3.5. Este valor determinará el número y las características que serán seleccionadas.

3. PCA: Dado que el número de características seleccionadas dependerá del número de componentes principales, utilizaremos el método PCA

con diferentes números de centroides para evaluar la importancia de cada característica.

4. Árboles de decisión: Con los métodos ID3 y CART, la importancia de una característica se determina por el nivel en el que se encuentra dentro del árbol de decisión. Las características más importantes estarán más cerca de la raíz y las menos importantes estarán cerca de las hojas. Se crearán dos listas de importancia, una para cada método

5. Chi-Cuadrada: Las características se ordenarán según su grado de dependencia con el Índice Aire Salud, desde la característica con mayor dependencia hasta la de menor.

6. Correlación de Pearson: Las características se ordenarán de acuerdo con su correlación con el índice Aire Salud, clasificándolas de mayor a menor correlación.

La lista de importancia para LASSO se puede visualizar en los cuadros 3.3 y 3.4, para los demás métodos, lo podemos visualizar en el Cuadro 3.6

La implementación de los algoritmos de selección automático los podemos encontrar en el archivo Subir_Seleccion_Maquinal en donde se detalla cada uno de los métodos

## 3.3. Selección de características con Teoría de Gráficas

En esta sección explicamos como utilizamos herramientas de la teoría de gráficas para clasificar y seleccionar los mejores conjuntos de características con el objetivo de mejorar el rendimiento de las redes neuronales.

Para construir nuestro modelo consideramos la gráfica completa $K_n$, donde $n$ es el número de variables (características). A cada variable le asociamos un vértice, y definimos su peso de las siguientes maneras:

*a*) El peso de un vértice $v$ se determina por el valor de la varianza de la variable correspondiente.

Cuadro 3.3: Conjunto de variables correspondiente a cada valor de $\alpha$, de 12 a 23 variables.

| | Valores de $\alpha$ | | | |
|---|---|---|---|---|
| Número características | $1 \times 10^{-5}$ | $5 \times 10^{-5}$ | $7 \times 10^{-5}$ | $8 \times 10^{-5}$ |
| 1 | Volt. batería mín. | Volt. batería mín. | Volt. batería mín. | Volt. batería mín. |
| 2 | Enfriamiento aire | Enfriamiento aire | Enfriamiento aire | Enfriamiento aire |
| 3 | Precipitación | Precipitación | Precipitación | Precipitación |
| 4 | Temp. bulbo húmedo | Temp. bulbo húmedo | Temp. bulbo húmedo | Temp. bulbo húmedo |
| 5 | Temp. de aire | Temp. de aire | Temp. de aire | Temp. de aire |
| 6 | Temp. amb. avg | Temp. amb. avg | Temp. amb. avg | Temp. amb. avg |
| 7 | Pot. de irradiancia | Pot. de irradiancia | Pot. de irradiancia | Pot. de irradiancia |
| 8 | Irradiation_Tot | Irradiation_Tot | Irradiation_Tot | Irradiation_Tot |
| 9 | Vel. viento máx | Humedad relativa | Humedad relativa | Humedad relativa |
| 10 | Humedad relativa | Temp. amb. máx | Temp. amb. máx | Temp. amb. máx |
| 11 | Temp. amb. máx | Temp. amb. mín. | Temp. amb. mín. | Temp. amb. mín. |
| 12 | Temp. amb. mín. | Vel. viento máx | Presión admosférica | Presión admosférica |
| 13 | Presión admosférica | Presión admosférica | Hrs. Sol | Hrs. Sol |
| 14 | Punto de rocío | Hrs. Sol | Vel. viento máx | Punto de rocío |
| 15 | Hrs. Sol | Punto de rocío | Punto de rocío | Vel. viento máx |
| 16 | Vel. viento std. | Irradiancia | Irradiancia | Irradiancia |
| 17 | Irradiancia | Vel. viento std. | P. saturación vapor | P. saturación vapor |
| 18 | Índice de calor | Dirección viento std. | Dirección viento std. | Dirección viento std. |
| 19 | Dirección viento std. | P. saturación vapor | Vel. viento std. | Vel. viento avg. |
| 20 | P. saturación vapor | Vel. viento avg. | Vel. viento avg. | Vel. viento std. |
| 21 | Dirección viento avg. | Dirección viento avg. | Dirección viento avg. | |
| 22 | Dirección viento SMM | Dirección viento SMM | | |
| 23 | Vel. viento avg. | | | |
| | Valores de $\alpha$ | | | |
| Número características | $9 \times 10^{-5}$ | $2{,}5 \times 10^{-4}$ | $2{,}7 \times 10^{-4}$ | $4{,}3 \times 10^{-4}$ |
| 1 | Volt. batería mín. | Volt. batería mín. | Volt. batería mín. | Volt. batería mín. |
| 2 | Enfriamiento aire | Enfriamiento aire | Enfriamiento aire | Enfriamiento aire |
| 3 | Precipitación | Temp. bulbo húmedo | Temp. bulbo húmedo | Temp. bulbo húmedo |
| 4 | Temp. bulbo húmedo | Precipitación | Temp. de aire | Temp. de aire |
| 5 | Temp. de aire | Temp. de aire | Precipitación | Temp. amb. avg |
| 6 | Temp. amb. avg | Temp. amb. avg | Temp. amb. avg | Pot. de irradiancia |
| 7 | Pot. de irradiancia | Pot. de irradiancia | Pot. de irradiancia | Temp. amb. máx |
| 8 | Irradiation_Tot | Temp. amb. máx | Temp. amb. máx | Irradiation_Tot |
| 9 | Temp. amb. máx | Irradiation_Tot | Irradiation_Tot | Precipitación |
| 10 | Humedad relativa | Temp. amb. mín. | Temp. amb. mín. | Presión admosférica |
| 11 | Temp. amb. mín. | Presión admosférica | Presión admosférica | Vel. viento máx |
| 12 | Presión admosférica | Vel. viento máx | Vel. viento máx | Temp. amb. mín. |
| 13 | Hrs. Sol | Hrs. Sol | P. saturación vapor | P. saturación vapor |
| 14 | Punto de rocío | P. saturación vapor | Hrs. Sol | Hrs. Sol |
| 15 | Vel. viento máx | Humedad relativa | Humedad relativa | Dirección viento std. |
| 16 | Irradiancia | Dirección viento std. | Dirección viento std. | Vel. viento avg. |
| 17 | P. saturación vapor | Vel. viento avg. | Vel. viento avg. | |
| 18 | Dirección viento std. | Punto de rocío | | |
| 19 | Vel. viento avg. | | | |
| | Valores de $\alpha$ | | | |
| Número características | $4{,}7 \times 10^{-4}$ | $6{,}5 \times 10^{-4}$ | $1 \times 10^{-3}$ | $1{,}5 \times 10^{-3}$ |
| 1 | Volt. batería mín. | Volt. batería mín. | Volt. batería mín. | Volt. batería mín. |
| 2 | Enfriamiento aire | Enfriamiento aire | Enfriamiento aire | Enfriamiento aire |
| 3 | Temp. bulbo húmedo | Temp. bulbo húmedo | Temp. bulbo húmedo | Temp. amb. avg |
| 4 | Temp. amb. avg | Temp. amb. avg | Temp. amb. avg | Temp. amb. máx |
| 5 | Temp. de aire | Temp. de aire | Temp. amb. máx | Irradiation_Tot |
| 6 | Pot. de irradiancia | Pot. de irradiancia | Temp. de aire | Temp. bulbo húmedo |
| 7 | Temp. amb. máx | Temp. amb. máx | Irradiation_Tot | Pot. de irradiancia |
| 8 | Irradiation_Tot | Irradiation_Tot | Pot. de irradiancia | Temp. de aire |
| 9 | Precipitación | Vel. viento máx | P. saturación vapor | P. saturación vapor |
| 10 | Presión admosférica | P. saturación vapor | Vel. viento máx | Vel. viento máx |
| 11 | Vel. viento máx | Presión admosférica | Presión admosférica | Dirección viento std. |
| 12 | P. saturación vapor | Temp. amb. mín. | Temp. amb. mín. | Presión admosférica |
| 13 | Temp. amb. mín. | Dirección viento std. | Dirección viento std. | |
| 14 | Hrs. Sol | Precipitación | | |
| 15 | Dirección viento std. | | | |

Cuadro 3.4: Conjunto de variables correspondiente a cada valor de $\alpha$, de 1 a 11 variables.

| | Valores de $\alpha$ | | | |
|---|---|---|---|---|
| Número características | $3{,}15 \times 10^{-3}$ | $3{,}2 \times 10^{-3}$ | $3{,}25 \times 10^{-3}$ | $4 \times 10^{-3}$ |
| 1 | Irradiation_Tot | Irradiation_Tot | Irradiation_Tot | Irradiation_Tot |
| 2 | Temp. amb. máx | Temp. amb. máx | Temp. amb. máx | Temp. amb. máx |
| 3 | Temp. amb. avg | Temp. amb. avg | Temp. amb. avg | Temp. amb. avg |
| 4 | Enfriamiento aire | Enfriamiento aire | Enfriamiento aire | Enfriamiento aire |
| 5 | Volt. batería mín. | Volt. batería mín. | Volt. batería mín. | Pot. de irradiancia |
| 6 | Pot. de irradiancia | Pot. de irradiancia | Pot. de irradiancia | Dirección viento std. |
| 7 | Índice de calor | Índice de calor | Dirección viento std. | Índice de calor |
| 8 | Dirección viento std. | Dirección viento std. | Índice de calor | Vel. viento máx |
| 9 | Vel. viento máx | Vel. viento máx | Vel. viento máx | |
| 10 | Temp. bulbo húmedo | Temp. bulbo húmedo | | |
| 11 | Temp. de aire | | | |
| | Valores de $\alpha$ | | | |
| Número características | $4{,}5 \times 10^{-3}$ | $5{,}35 \times 10^{-3}$ | $5{,}45 \times 10^{-3}$ | $5{,}5 \times 10^{-3}$ |
| 1 | Irradiation_Tot | Irradiation_Tot | Irradiation_Tot | Irradiation_Tot |
| 2 | Temp. amb. avg | Temp. amb. avg | Temp. amb. avg | Temp. amb. avg |
| 3 | Temp. amb. máx | Dirección viento std. | Dirección viento std. | Dirección viento std. |
| 4 | Enfriamiento aire | Índice de calor | Índice de calor | Índice de calor |
| 5 | Dirección viento std. | Temp. amb. máx | Pot. de irradiancia | |
| 6 | Pot. de irradiancia | Pot. de irradiancia | | |
| 7 | Índice de calor | | | |
| | Valores de $\alpha$ | | | |
| Número características | $6 \times 10^{-3}$ | $7{,}5 \times 10^{-3}$ | $8 \times 10^{-3}$ | |
| 1 | Irradiation_Tot | Irradiation_Tot | Irradiation_Tot | |
| 2 | Temp. amb. avg | Dirección viento std. | | |
| 3 | Dirección viento std. | | | |

Cuadro 3.5: Valores de $\alpha$ y número de características

| **Valores de $\alpha$** | **Número de características** |
|---|---|
| $1 \times 10^{-5}$ | 23 |
| $5 \times 10^{-5}$ | 22 |
| $7 \times 10^{-5}$ | 21 |
| $8 \times 10^{-5}$ | 20 |
| $9 \times 10^{-5}$ | 19 |
| $2{,}5 \times 10^{-4}$ | 18 |
| $2{,}7 \times 10^{-4}$ | 17 |
| $4{,}3 \times 10^{-4}$ | 16 |
| $4{,}7 \times 10^{-4}$ | 15 |
| $6{,}5 \times 10^{-4}$ | 14 |
| $1 \times 10^{-3}$ | 13 |
| $1{,}5 \times 10^{-3}$ | 12 |
| $3{,}15 \times 10^{-3}$ | 11 |
| $3{,}2 \times 10^{-3}$ | 10 |
| $3{,}25 \times 10^{-3}$ | 9 |
| $4 \times 10^{-3}$ | 8 |
| $4{,}5 \times 10^{-3}$ | 7 |
| $5{,}35 \times 10^{-3}$ | 6 |
| $5{,}45 \times 10^{-3}$ | 5 |
| $5{,}5 \times 10^{-3}$ | 4 |
| $6 \times 10^{-3}$ | 3 |
| $7{,}5 \times 10^{-3}$ | 2 |
| $8 \times 10^{-3}$ | 1 |

Cuadro 3.6: Listas de importancia para cada uno de los métodos de Aprendizaje Automático. El número indica la relevancia de una variable, mientras menor sea el número, mayor será su importancia.

| **Importancia** | **Varianza** | **MAD** | **DR** | **PCA** |
|---|---|---|---|---|
| **1** | Dirección viento std. | Dirección viento std. | Dirección viento SMM | Hrs. Sol |
| **2** | Irradiación | Irradiación | Punto de rocío | Temp. amb. máx. |
| **3** | Irradiancia. | Dirección viento SMM | Temp. amb. máx. | Temp. amb. mín. |
| **4** | Dirección viento SMM | Potencial irradiancia. | Temp. amb. mín. | Temp. de aire |
| **5** | Potencial irradiancia. | Irradiancia. | Temp. de aire | Presión atmosférica |
| **6** | Humedad relativa | Humedad relativa | Presión atmosférica | Irradiancia. |
| **7** | Vel. de viento avg. | Hrs. Sol | Irradiancia. | Temp. amb. avg. |
| **8** | Vel. de viento máx. | Dirección viento avg. | Temp. amb. avg. | Humedad relativa |
| **9** | Hrs. Sol | Vel. de viento máx. | Humedad relativa | Volt. batería mín. |
| **10** | Dirección viento avg. | Vel. de viento avg. | Volt. batería mín. | Precipitación |
| **11** | Vel. viento std. | Vel. viento std. | Precipitación | Punto de rocío |
| **12** | Temp. amb. avg. | Temp. bulbo húmedo | Temp. bulbo húmedo | Temp. bulbo húmedo |
| **13** | Índice de calor | Temp. amb. avg. | Vel. de viento máx. | Índice de calor |
| **14** | Enfriamiento aire | Presión atmosférica | Índice de calor | Enfriamiento aire |
| **15** | Volt. batería mín. | Índice de calor | Enfriamiento aire | Potencial irradiancia. |
| **16** | Temp. bulbo húmedo | Enfriamiento aire | Potencial irradiancia. | Presión saturación vapor |
| **17** | Temp. amb. máx. | Volt. batería mín. | Presión saturación | Irradiación |
| **18** | Punto de rocío | Punto de rocío | Irradiación | Vel. de viento avg. |
| **19** | Presión saturación vapor | Temp. amb. máx. | Vel. de viento avg. | Vel. viento std. |
| **20** | Presión atmosférica | Presión saturación vapor | Vel. viento std. | Dirección viento avg. |
| **21** | Temp. amb. mín. | Temp. de aire | Dirección viento avg. | Dirección viento std. |
| **22** | Temp. de aire | Temp. amb. mín. | Dirección viento std. | Vel. de viento máx. |
| **23** | Precipitación | Precipitación | Hrs. Sol | Dirección viento SMM |

| **Importancia** | **CART** | **ID3** | **Chi-Cuadradada** | **Pearson** |
|---|---|---|---|---|
| **1** | Temp. de aire | Temp. de aire | Irradiación | Irradiación |
| **2** | Temp. amb. máx. | Temp. amb. máx. | Irradiancia. | Temp. de aire |
| **3** | Irradiación | Presión saturación vapor | Temp. amb. avg. | Temp. amb. avg. |
| **4** | Volt. batería mín. | Presión atmosférica | Temp. de aire | Presión saturación vapor |
| **5** | Enfriamiento aire | Temp. amb. mín. | Índice de calor | Enfriamiento aire |
| **6** | Humedad relativa | Punto de rocío | Dirección viento std. | Índice de calor |
| **7** | Dirección viento avg. | Índice de calor | Enfriamiento aire | Irradiancia. |
| **8** | Temp. amb. mín. | Potencial irradiancia. | Presión saturación vapor | Vel. viento std. |
| **9** | Temp. bulbo húmedo | Humedad relativa | Humedad relativa | Vel. de viento máx. |
| **10** | Presión atmosférica | Dirección viento avg. | Vel. viento std. | Dirección viento std. |
| **11** | Hrs. Sol | Volt. batería mín. | Vel. de viento máx. | Vel. de viento avg. |
| **12** | Potencial irradiancia. | Vel. viento std. | Dirección viento SMM | Potencial irradiancia. |
| **13** | Irradiancia. | Dirección viento SMM | Potencial irradiancia. | Humedad relativa |
| **14** | Punto de rocío | Temp. bulbo húmedo | Vel. de viento avg. | Dirección viento SMM |
| **15** | Vel. de viento avg. | Hrs. Sol | Temp. amb. máx. | Dirección viento avg. |
| **16** | Vel. de viento máx. | Dirección viento std. | Punto de rocío | Temp. bulbo húmedo |
| **17** | Temp. amb. avg. | Irradiación | Dirección viento avg. | Temp. amb. máx. |
| **18** | Dirección viento SMM | Vel. de viento avg. | Volt. batería mín. | Hrs. Sol |
| **19** | Vel. viento std. | Enfriamiento aire | Hrs. Sol | Temp. amb. mín. |
| **20** | Dirección viento std. | Vel. de viento máx. | Temp. bulbo húmedo | Presión atmosférica |
| **21** | Índice de calor | Temp. amb. avg. | Precipitación | Precipitación |
| **22** | Presión saturación vapor | Irradiancia. | Temp. amb. mín. | Volt. batería mín. |
| **23** | Precipitación | Precipitación | Presión atmosférica | Punto de rocío |

*b*) El peso de un vértice $v$ se establece según la posición que ocupa la variable dentro de los árboles de decisión ID3 y CART.

*c*) El peso de un vértice $v$ también puede definirse mediante una combinación de las importancias obtenidas a partir de varias centralidades, como la centralidad de grado, intermediación, cercanía, Katz y PageRank. Para cada método, generamos una lista de importancia y el peso del vértice se calcula sumando sus posiciones en todas las listas, dividiendo entre el número total de métodos utilizados. Cabe destacar que, como se muestra en el Cuadro 3.7, no se incluye la columna del método Katz en este caso, ya que la gráfica es no conexa y no se puede calcular este método.

Con estos elementos es posible construir cuatro gráficas diferentes, cada una correspondiente a una de las formas en que se puede definir el peso de los vértices.

Por otro lado, podemos extrapolar las ideas de la Distancia del Movimiento de Tierra (EMD, por sus siglas en inglés) para medir la distancia entre dos variables, imaginando una como un "montón de tierra" y la otra como un "agujero" donde esa tierra debe ser trasladada. De esta manera, podemos calcular la similitud entre variables y sus respectivas series de tiempo al medir el trabajo necesario para transformar una serie en otra, como se ilustra en la Figura 3.1. Este enfoque nos permite utilizar la distancia EMD para determinar el peso de las aristas entre los vértices correspondientes.

Es importante destacar que las aristas de mayor peso indican una menor similitud entre las variables. Por lo tanto, para preservar las relaciones más significativas dentro de nuestra gráfica, es deseable conservar las aristas con menor peso. Con base en esto, hemos definido nueve subgráficas $H_i$ para $i \in \{10, 20, \ldots, 90\}$ de la gráfica $K_n$, donde la subgráfica $H_i$ contiene el $i\,\%$ de las aristas con menor peso de $K_n$. Es decir, $H_1$ incluye el $10\,\%$ de las aristas con menor peso, $H_2$ contiene el $20\,\%$ de las aristas con menor peso, y así sucesivamente. Este proceso permite estudiar los cambios en la importancia de las variables a medida que se incluyen más aristas en las subgráficas.

Para visualizar el funcionamiento de cada uno de los métodos de centralidad en Teoría de Gráficas, consideremos la subgráfica de la gráfica completa $K_5$ mostrada en la Figura 3.2. En este ejemplo, las aristas están ponderadas,

Cuadro 3.7: Ejemplo de las listas de importancia para cada uno de los métodos de centralidad. Cada número en las columnas representa la posición que ocupa la variable en la lista de importancia de cada método. Si tienen el mismo número, significa que tienen la misma importancia dentro del método. La columna "Peso vértice"muestra la suma de las posiciones obtenidas de cada variable en cada método, dividida entre el número total de métodos.

| | **Posición** | | | | |
|---|---|---|---|---|---|
| **Característica** | **Intermediación** | **Cercanía** | **Grado** | **Page Rank** | **Peso vértice** |
| **Temp. amb. máx.** | 1 | 4 | 2 | 1 | 2 |
| **Temp. amb. mín.** | 4 | 3 | 1 | 2 | 5/2 |
| **Índice calor** | 3 | 2 | 4 | 4 | 13/4 |
| **Enfr. viento** | 4 | 1 | 5 | 5 | 15/4 |
| **Volt. batería mín.** | 5 | 6 | 3 | 3 | 17/4 |
| **Temp. bulbo húmedo** | 2 | 7 | 9 | 8 | 13/2 |
| **Temp. del aire** | 5 | 5 | 7 | 9 | 13/2 |
| **Presión saturación vapor** | 5 | 8 | 6 | 7 | 13/2 |
| **Humedad relativa** | 5 | 9 | 8 | 10 | 8 |
| **Vel. de viento std.** | 5 | 10 | 11 | 6 | 8 |
| **Vel. viento máx.** | 5 | 10 | 11 | 6 | 8 |
| **Dirección del viento avg.** | 5 | 11 | 10 | 11 | 37/4 |
| **Hrs. Sol** | 5 | 12 | 12 | 12 | 41/4 |
| **Temp. amb. mín.** | 5 | 12 | 12 | 12 | 41/4 |
| **Presión admosférica** | 5 | 12 | 12 | 12 | 41/4 |
| **Irradiancia** | 5 | 12 | 12 | 12 | 41/4 |
| **Precipitación** | 5 | 12 | 12 | 12 | 41/4 |
| **Punto de rocío** | 5 | 12 | 12 | 12 | 41/4 |
| **Potencial de irradiancia** | 5 | 12 | 12 | 12 | 41/4 |
| **Irradiación** | 5 | 12 | 12 | 12 | 41/4 |
| **Vel. viento avg.** | 5 | 12 | 12 | 12 | 41/4 |
| **Dirección de viento std.** | 5 | 12 | 12 | 12 | 41/4 |
| **Dirección viento SMM** | 5 | 12 | 12 | 12 | 41/4 |

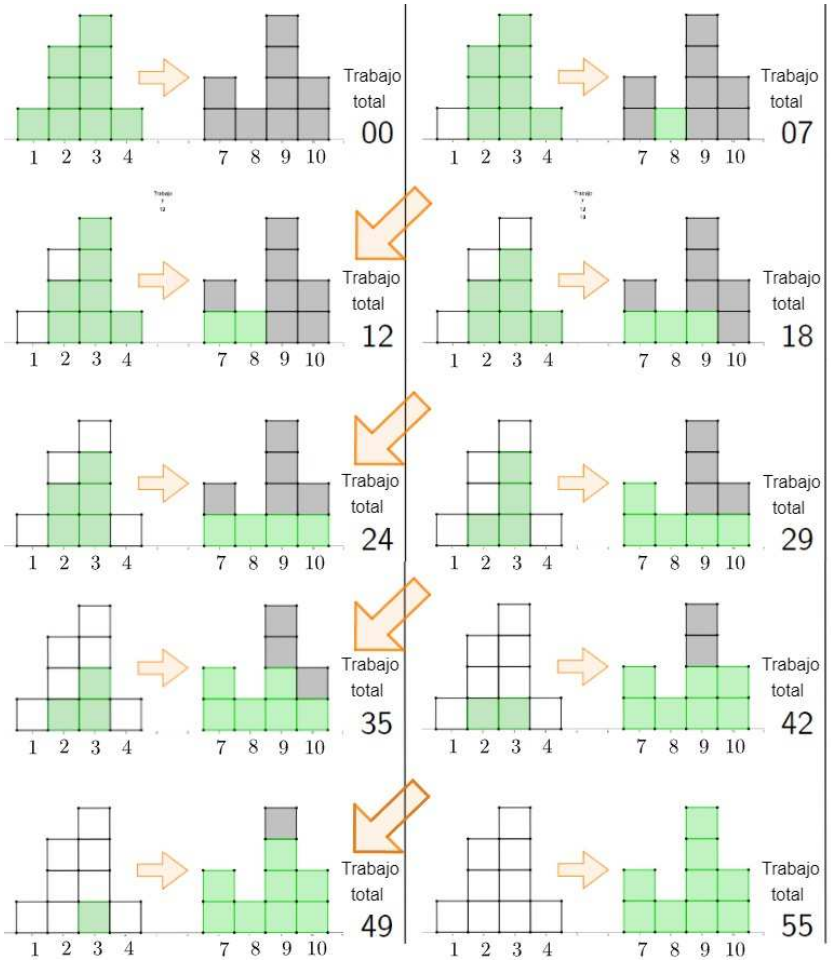



Figura 3.1: Ejemplo del trabajo que se requiere para pasar de la serie de tiempo verde a la gris. Este ejemplo no considera un trabajo óptimo, únicamente propone una posible solución parcial al problema.

representando relaciones entre a partir de la métrice EMD los vértices. A continuación, se describe el cálculo detallado de cada uno de los métodos.

- **Centralidad de Intermediación:**

  En la gráfica de la Figura 3.2, obtenemos todas las geodésicas, como se observa en el Cuadro 3.8. De aquí, observamos que no pasa ninguna geodésica por los vértices verde, café y rojo. Por otro lado, en los vérti-

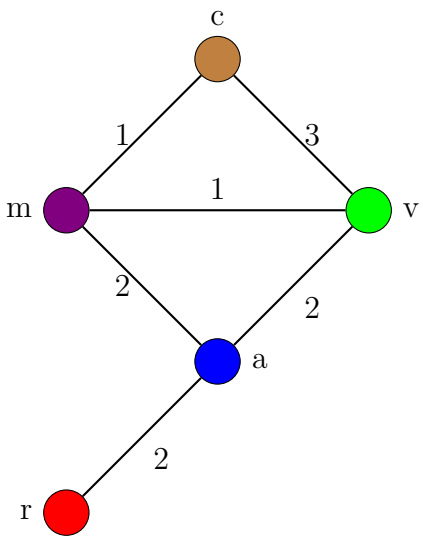



Figura 3.2: Gráfica ponderada que representa relaciones entre vértices con pesos en las aristas.

ces azul y morado hay tres geodésicas que pasan por ellos. Además, solo existe una sola trayectoria para cada par de vértices en este caso.

Por lo tanto, la Centralidad de Intermediación se calcula como:

$$\text{cen}_b(a) = \text{cen}_b(m) = 3, \quad \text{y} \quad \text{cen}_b(v) = \text{cen}_b(c) = \text{cen}_b(r) = 0,$$

donde $a$ es el vértice azul, $m$ es el vértice morado, $v$ es el vértice verde, $c$ es el vértice café y $r$ es el vértice rojo.

- **Centralidad de Cercanía:** del Cuadro 3.8 y la Figura 3.2, podemos

Cuadro 3.8: Descripción de todas las geodésicas de la gráfica de la Figura 3.2

| Geodésica $uv-trayectoria$ | Distancia $d(u,v)$ | Geodésica $uv-trayectoria$ | Distancia $d(u,v)$ |
|---|---|---|---|
| $(c,m,a,r)$ | 5 | $(m,v)$ | 1 |
| $(c,m,a)$ | 3 | $(m,a)$ | 2 |
| $(c,m)$ | 1 | $(v,a,r)$ | 4 |
| $(c,m,v)$ | 2 | $(v,a)$ | 2 |
| $(m,a,r)$ | 4 | $(a,r)$ | 2 |

observar que las distancias promedio de los vértices son:

$$l(a) = l(v) = \frac{9}{5}, \quad l(m) = \frac{8}{5}, \quad l(c) = \frac{11}{5}, \quad l(r) = \frac{15}{5}.$$

Por tanto, la Centralidad de Cercanía en este caso es:

$$\text{cen}(a) = \text{cen}(v) = \frac{5}{9}, \quad \text{cen}(m) = \frac{5}{8}, \quad \text{cen}(c) = \frac{5}{11}, \quad \text{cen}(r) = \frac{5}{15}.$$

- **Centralidad de Grado:** del ejemplo de la Figura 3.2, el grado de cada vértice es:

$$\deg(a) = \deg(v) = 6, \quad \deg(m) = \deg(c) = 4, \quad \deg(r) = 2.$$

Por lo tanto, la Centralidad de Grado es:

$$\text{cen}_{\deg}(a) = \text{cen}_{\deg}(v) = \frac{3}{2}, \quad \text{cen}_{\deg}(m) = \text{cen}_{\deg}(c) = 1, \quad \text{cen}_{\deg}(r) = \frac{1}{2}.$$

- **Centralidad de Katz:** del ejemplo de la Figura 3.2, se utilizaron los siguientes parámetros para calcular la Centralidad de Katz:

$$\alpha = 0{,}3685, \quad \beta = 1.$$

En este caso, se utilizó este valor de $\alpha$ ya que debe ser estrictamente menor que:

$$\frac{1}{\lambda_{\max}} = \frac{1}{2{,}64}.$$

Para profundizar sobre este método, se puede consultar en [22] y la documentación de NetworkX

Los valores obtenidos son:

$$C_K(\text{m}) = C_K(\text{v}) = 42{,}259, \quad C_K(\text{a}) = 37{,}491,$$

$$C_K(\text{c}) = 31{,}892, \quad C_K(\text{r}) = 14{,}188.$$

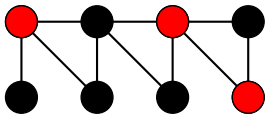
- **PageRank:** dada la gráfica de la Figura 3.2, asignamos un valor inicial de PageRank a cada vértice de $\frac{1}{5}$ con un parámetro de amortiguación $\alpha = 0{,}85$. Para profundizar sobre este método, se puede consultar en la documentación de NetworkX

  Los resultados del PageRank son:

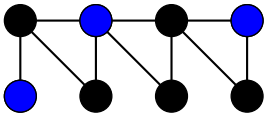
$$PR(\text{a}) = 0{,}2719, \quad PR(\text{v}) = 0{,}2599, \quad PR(\text{m}) = 0{,}1819,$$

$$PR(\text{c}) = 0{,}1791, \quad PR(\text{r}) = 0{,}1071.$$

Para seleccionar las características, consideraremos aquellas que formen parte de un conjunto dominante independiente en la gráfica. Obtener un conjunto dominante independiente implica que cada vértice de la gráfica está en el conjunto o es adyacente a un vértice del conjunto, y que no existen aristas entre los vértices de dicho conjunto. De esta manera, nos aseguramos de obtener un conjunto de variables que mantiene una alta representatividad con respecto al conjunto de todas variables y al mismo tiempo que se evita la redundancia entre las características seleccionadas. Además, es evidente no existe un único conjunto dominante independiente en una gráfica como lo podemos ver en la Figura 3.3. Por lo tanto, entre todos los posibles conjuntos, seleccionaremos aquellos que maximicen la suma de los pesos de los vértices en función de la varianza (aplicado para el primer tipo de gráfica) y minimicen los pesos en relación con su posición dentro de los árboles de decisión (aplicado para el segundo y tercer tipo de gráfica), así como los obtenidos a través de las centralidades (aplicado para el cuarto tipo de gráfica). Los resultados finales obtenidos de esta selección se muestran en los cuadros 3.9, 3.10, 3.11, 3.12.



(a) Conjunto dominante.



(b) Conjunto dominante independiente.

Figura 3.3: Comparación de gráficas con diferentes conjuntos dominantes.

| | **Porcentaje de aristas de la Gráfica $K_n$** | | | | | |
|---|---|---|---|---|---|---|
| **Número de variables** | **10 %** | **20 %** | **30 %-40 %** | **50 %** | **60 %** | **70 %-100 %** |
| **1** | Hrs. Sol | Hrs. Sol | Hrs. Sol | Hrs. Sol | Precipitación | Dirección viento std. |
| **2** | Temp. amb. mín. | Humedad relativa | Precipitación | Precipitación | Dirección viento std. | |
| **3** | Presión admosférica | Precipitación | Potencial Irradiancia | Dirección viento std. | | |
| **4** | Irradiancia | Punto de rocío | Dirección viento std. | Dirección viento SMM | | |
| **5** | Humedad relativa | Potencial Irradiancia | Dirección viento SMM | | | |
| **6** | Precipitación | Irradiánción | | | | |
| **7** | Punto de rocío | Dirección viento std. | | | | |
| **8** | Potencial Irradiancia | Dirección viento SMM | | | | |
| **9** | Irradiánción | | | | | |
| **10** | Vel. viento avg. | | | | | |
| **11** | Dirección viento std. | | | | | |
| **12** | Vel. viento máx. | | | | | |
| **13** | Dirección viento SMM | | | | | |

Cuadro 3.9: Selección de características de las subgráficas con diferentes porcentajes de aristas basadas en el peso de las aristas con baja varianza.

| **Porcentaje de aristas de la Gráfica $K_n$** | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Número de variables** | **10 %** | **20 %** | **30 % - 40 %** | **50 %** | **60 %** | **70 %** | **80 %** | **90 %** | **100 %** |
| **1** | Hrs. Sol | Hrs. Sol | Hrs. Sol | Hrs. Sol | Precipitación | Vel. viento avg. | Vel. viento máx. | Dirección viento std. | Precipitación |
| **2** | Temp. amb. mán. | Temp. amb. mán. | Precipitación | Precipitación | Vel. viento avg. | | | | |
| **3** | Temp. amb. mín. | Irradiancia | Potencial Irradiancia | Dirección viento avg. | | | | | |
| **4** | Presión admosférica | Precipitación | Presión saturación vapor | Dirección viento std. | | | | | |
| **5** | Irradiancia | Punto de rocío | Irradiación | | | | | | |
| **6** | Precipitación | Potencial Irradiancia | | | | | | | |
| **7** | Punto de rocío | Dirección viento std. | | | | | | | |
| **8** | Potencial Irradiancia | Dirección viento SMM | | | | | | | |
| **9** | Irradiación | | | | | | | | |
| **10** | Vel. viento avg. | | | | | | | | |
| **11** | Vel. viento std. | | | | | | | | |
| **12** | Dirección viento std. | | | | | | | | |
| **13** | Dirección viento SMM | | | | | | | | |

Cuadro 3.10: Selección de características de las subgráficas con diferentes porcentajes de aristas basadas en el peso de las aristas con Centralidad.

| | **Porcentaje de aristas de la Gráfica $K_n$** | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Número de variables** | **10 %** | **20 %** | **30 %** | **40 %** | **50 %** | **60 %** | **70 %** | **80 %** | **90 %** | **100 %** |
| **1** | Hrs. Sol | Hrs. Sol | Hrs. Sol | Hrs. Sol | Hrs. Sol | Temp. del aire | Temp. del aire | Temp. del aire | Hrs. Sol | Temp. del aire |
| **2** | Temp. amb. mín. | Temp. amb. mán. | Precipitación | Precipitación | Temp. amb. mán. | Precipitación | Irradiancia | Irradiación | | |
| **3** | Presión admosférica | Precipitación | Potencial Irradiancia | Potencial Irradiancia | Precipitación | Dirección viento std. | | | | |
| **4** | Irradiancia | Punto de rocío | Irradiación | Irradiación | Potencial Irradiancia | | | | | |
| **5** | Precipitación | Potencial Irradiancia | Dirección viento avg. | Vel. viento avg. | Irradiación | | | | | |
| **6** | Punto de rocío | Irradiación | Vel. viento máx. | Dirección viento avg. | | | | | | |
| **7** | Enfr. viento | Vel. viento avg. | | | | | | | | |
| **8** | Potencial Irradiancia | Dirección viento std. | | | | | | | | |
| **9** | Irradiación | Dirección viento SMM | | | | | | | | |
| **10** | Vel. viento avg. | | | | | | | | | |
| **11** | Dirección viento avg. | | | | | | | | | |
| **12** | Dirección viento std. | | | | | | | | | |
| **13** | Vel. viento máx. | | | | | | | | | |
| **14** | Dirección viento SMM | | | | | | | | | |

Cuadro 3.11: Selección de características de las subgráficas con diferentes porcentajes de aristas basadas en el peso de las aristas con CART.

| | **Porcentaje de aristas de la Gráfica $K_n$** | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Número de variables** | **10 %** | **20 %** | **30 %** | **40 %** | **50 %** | **60 %** | **70 %** | **80 %** | **90 %** | **100 %** |
| **1** | Hrs. Sol | Hrs. Sol | Hrs. Sol | Hrs. Sol | Hrs. Sol | Temp. del aire | Temp. del aire | Temp. del aire | Hrs. Sol | Temp. del aire |
| **2** | Temp. amb. mín. | Temp. amb. mán. | Precipitación | Precipitación | Temp. amb. mán. | Precipitación | Irradiancia | Irradiación | | |
| **3** | Presión admosférica | Precipitación | Potencial Irradiancia | Potencial Irradiancia | Precipitación | Dirección viento std. | | | | |
| **4** | Irradiancia | Punto de rocío | Irradiación | Irradiación | Potencial Irradiancia | | | | | |
| **5** | Precipitación | Potencial Irradiancia | Vel. viento std. | Vel. viento avg. | Dirección viento std. | | | | | |
| **6** | Punto de rocío | Irradiación | Dirección viento avg. | Dirección viento avg. | | | | | | |
| **7** | Índice calor | Vel. viento std. | | | | | | | | |
| **8** | Potencial Irradiancia | Dirección viento std. | | | | | | | | |
| **9** | Irradiación | Dirección viento SMM | | | | | | | | |
| **10** | Vel. viento avg. | | | | | | | | | |
| **11** | Vel. viento std. | | | | | | | | | |
| **12** | Dirección viento avg. | | | | | | | | | |
| **13** | Dirección viento std. | | | | | | | | | |
| **14** | Dirección viento SMM | | | | | | | | | |

Cuadro 3.12: Selección de características de las subgráficas con diferentes porcentajes de aristas basadas en el peso de las aristas con ID3.

## 3.4. Diseño de redes neuronales

Para alcanzar el objetivo de determinar un subconjunto de características significativas donde la pérdida de información sea pequeña para la clasificación de los índices de contaminación de dióxido de azufre, monóxido de carbono, dióxido de nitrógeno, ozono y partículas suspendidas, se utilizarán técnicas de selección de características, Aprendizaje Automático y análisis de Teoría de Gráficas, con el fin de reducir el número de sensores (recursos) de un sistema de monitoreo atmosférico en la zona de Santa Fe.

Se ha visto recientemente que las redes neuronales tienen una superioridad para las clasificaciones de índices de contaminación como se puede ver en [14], ya que superan la exactitud a métodos como árboles de decisión [32], Naïve Bayes [24], enfoques difusos [33], entre otros. Por esta razón, utilizamos redes neuronales para la clasificación en donde la variable objetivo es el Índice Aire Salud por minuto.

Es importante decir que se utilizarán todas las variables descritas en el Cuadro 2.3 sin considerar las variables de contaminación $NO_2$, $SO_2$, $CO$, $O_3$, $PM_{10}$, $PM_{2,5}$. Estas variables definen el Índice Aire Salud, y el objetivo es predecirlo sin considerarlas, utilizando únicamente factores meteorológicos. Esto otorgará mayor relevancia a las variables meteorológicas dentro de la red neuronal en comparación con las demás.

Por último, para medir la eficiencia del modelo nos basaremos en el Recall de la clase 4, ya que esta clase es la que tiene un impacto nocivo a la salud y es cuando se activa la contingencia ambiental.

Para llevar a cabo los experimentos de este proyecto, se siguió un proceso detallado que abarca desde la normalización de las variables hasta la evaluación de los modelos resultantes. A continuación, se describe el enfoque implementado:

### 3.4.1. Normalización de las variables

Como primer paso, se normalizaron todas las variables del conjunto de datos utilizando el método **MinMax**. Este método asegura que todas las características estén escaladas en un rango entre 0 y 1, lo que resulta beneficioso

para evitar que algunas variables con mayor magnitud dominen sobre otras durante el entrenamiento de las redes neuronales. Este método es sensible a valores atípicos, sin embargo, como se realizó un preprocesamiento de los datos al eliminar los valores atípicos y completar la información como se discutió en la sección 2, podemos estar seguros que no habrá ningún inconveniente. Esto se hace, ya que los pesos asignados a cada variable pueden verse significativamente afectados por diferencias en las escalas

### 3.4.2. Subconjuntos de selección de características

Para las listas de importancia que se obtuvieron en los métodos basados en Aprendizaje Automático mencionados en la sección 3.2, se generaron subconjuntos de variables de tamaño decreciente. Es decir, comenzando con todas las variables (23 en total), se elimina progresivamente la característica menos importante en cada iteración, conforme a la importancia asignada por cada método. Este enfoque permite observar cómo afecta la eliminación de características al rendimiento del modelo. Esto lo podemos definir de la siguiente manera

Dado el orden de importancia de las variables $\mathcal{X}^i = \{\mathcal{C}_1^i, \mathcal{C}_2^i, \ldots, \mathcal{C}_m^i\}$ para algún método $i$, donde $\mathcal{C}_1^i$ es la variable más importante y $\mathcal{C}_m^i$ es la menos importante según el método $i$.

Definimos $\mathcal{X}_k^i$ como el subconjunto de las primeras $k$ variables más importantes:

$$\mathcal{X}_k^i = \{\mathcal{C}_1^i, \mathcal{C}_2^i, \ldots, \mathcal{C}_k^i\}, \quad \text{para } k = 1, \ldots, m-1$$

De esta forma, podemos verlo de manera iterativa como sigue:

$$\mathcal{X}_k^i = \mathcal{X}_{k+1}^i \setminus \{\mathcal{C}_{k+1}^i\}$$

Este enfoque nos permite observar cómo afecta la eliminación progresiva de características al rendimiento del modelo, al evaluar el modelo con cada subconjunto $\mathcal{X}_k^i$.

En los métodos basados en Teoría de Gráficas, el número y tipo de carac-

terísticas para el conjunto de entrada de las redes neuronales se determinan a partir de las subgráficas $H_i$. Por ejemplo, si para la subgráfica $H_6$, con el peso en los vértices definido con baja varianza es {precipitación, dirección viento std. }, estas dos variables serán las seleccionadas para generar y entrenar una red neuronal.

Además, se integró una lista específica que contiene 12 variables, resultado de un emparejamiento de la gráfica completa presentado en la sección 1.5. Este conjunto se creó minimizando el peso de las aristas en el emparejamiento, y está compuesto por las siguientes variables:

{dirección del viento std, enfriamiento de viento, potencial de irradianza, índice de calor, horas sol, irradiación, velocidad de viento máximo, dirección de viento SMM, temperatura bulbo húmedo, irradiancia, temperatura ambiente, humedad relativa }

### 3.4.3. Diseño de redes neuronales

Para cada conjunto de características generado en el paso anterior, se definió una red neuronal distinta. El número de neuronas en la capa de entrada está determinado por el tamaño del subconjunto de características correspondiente a cada iteración. A continuación, se describen la arquitectura general de las redes neuronales:

- Se emplearon dos capas ocultas. En la primera capa oculta, se aplicó la regla empírica de usar **2/3** de las neuronas de la capa de entrada más las neuronas de la capa de salida. Esta elección se basa en estudios previos que sugieren que esta proporción mejora la capacidad de generalización de la red
- La segunda capa oculta tiene la mitad de las neuronas de la primera capa, lo que ayuda a reducir la complejidad del modelo de manera gradual, evitando el sobreajuste. Ambas capas ocultas utilizaron la función de activación **Tanh**, elegida por su capacidad de manejar relaciones no lineales de manera eficiente
- La capa de salida consta de 4 neuronas, una por cada clase en el Índice Aire Salud, con una función de activación **Softmax** para garantizar que la salida sea una probabilidad distribuida entre las clases

### 3.4.4. Validación cruzada y configuración del entrenamiento

Para evaluar el rendimiento de las redes neuronales, se empleó un esquema de **validación cruzada con *k*-folds** con el 80 % de los datos originales y de los datos generados por Smote. Este enfoque permite evitar el sobreajuste y proporciona una estimación más robusta del rendimiento del modelo al promediar los resultados obtenidos en distintos subconjuntos del conjunto de datos [14].

La configuración del entrenamiento incluyó los siguientes parámetros:

- **Batch size** de 100: Este valor permite actualizar los pesos del modelo después de procesar 100 ejemplos.
- **Épocas (epochs)**: Se fijó en 50 el número máximo de épocas, lo que da suficiente tiempo para que el modelo converja sin riesgo de sobreajuste excesivo. Este valor se definió de forma experimental, ya que al entrenar las redes neuronales se observó que de forma general, la mayoría de estos no había una mejora significativa exactitud del modelo.
- **Shuffle=True**: Se activó el barajado de datos para evitar que el modelo se vea afectado por el orden de los ejemplos.
- **Early Stopping**: Se utilizó un callback de **EarlyStopping** que monitoriza la métrica de `categorical_accuracy` y detiene el entrenamiento si no se observa una mejora en 3 épocas consecutivas. Esta técnica previene el sobreentrenamiento y optimiza el tiempo de entrenamiento.
- **Workers=12**: Se emplearon 12 procesos paralelos para acelerar el entrenamiento, dado que el conjunto de datos es lo suficientemente grande como para beneficiarse de esta optimización.

### 3.4.5. Registro de resultados

Durante el proceso de entrenamiento, se registraron tanto los **tiempos de entrenamiento** como los valores de **exactitud promedio** para cada red

neuronal. Este registro es importante para evaluar la eficiencia de los diferentes conjuntos de características y entender el impacto del número de variables en el tiempo de procesamiento.

El objetivo principal de este proyecto fue lograr la mejor clasificación posible de la **Clase 4**. Por esta razón, es importante identificar el mejor modelo en términos de $k$-fold para la clasificación de la **Clase 4**. Para esto, se evaluó el rendimiento de cada red neuronal considerando el **recall** de la Clase 4, ya que este indicador mide la capacidad del modelo de identificar correctamente los ejemplos de dicha clase. Se registró la información del $k$-fold mejor modelo para cada método y subconjuntos de características.

### 3.4.6. Análisis comparativo

Finalmente, se generaron gráficos de rendimiento para cada método y sus respectivas listas de importancia, mostrando la relación entre el número de variables y la exactitud, el tiempo promedio y el recall de la Clase 4 como se observa en las gráficas 4.1, 4.4, 4.3, 4.6, 4.2 y 4.5. Esto permite comparar de manera clara y concisa el impacto de la selección de variables de cada método en el rendimiento de las redes neuronales.

Además, se crearon dos tipos de cuadros: la primera muestra los tiempos máximos y mínimos, así como el mejor y peor exactitud alcanzado por cada método conforme aumenta el número de variables; la segunda muestra los métodos que lograron los valores más altos de recall para cada cantidad de variables.

El diseño, entrenamiento, validación, registro de resultados y las gráficas de rendimiento de las redes neuronales para los datos originales y los datos generados por SMOTE se encuentran disponibles en los repositorios de Zenodo titulados `Resultados Redes Neuronales Data Originales.zip` y `Resultados` Redes Neuronales Data Smote.zip. En estos repositorios se incluyen todos los archivos `.pkl` correspondientes a las validaciones cruzadas (*k-fold*) de las redes neuronales, así como los archivos que muestran las mejores *k-fold* encontradas para aumentar el *recall* de la clase 4.

# Capítulo 4

# Resultados

En esta sección se presenta el análisis del rendimiento de una red neuronal en función del número y de las características seleccionadas, utilizando los métodos descritos en las secciones 3.2 y 3.3. Se emplean tanto los datos originales como aquellos generados mediante SMOTE para el balanceo de clases. Los indicadores utilizados para medir el rendimiento de los modelos evaluados son el recall, la exactitud promedio, y el tiempo de ejecución.

Desde esta perspectiva, en términos generales, se observa que que el rendimiento de la exactitud promedio y el recall para la clase 4 va disminuyendo conforme va bajando el número de características, y como era de esperarse, también ocurre con el tiempo de entrenamiento de las redes neuronales, como se puede ver en las figuras 4.1, 4.4, 4.3, 4.6, 4.2 y 4.5.

A primera vista, podemos realizar varias observaciones generales para el conjunto de datos originales:

- Hay casos excepcionales como el método de MAD cuando se toma los datos originales; en este caso, este método tuvo los tiempos de entrenamiento más altos entre 15 a 23 características (Figura 4.1), pero con una mayor dispersión en la exactitud promedio (Figura 4.2), y sin una mejora significativa en el recall de la clase 4 (Figura 4.3). Esto lo podemos interpretar como que las redes neuronales se concentraron en clasificar correctamente las clases 1 y 2.
- Cuando hay 5 características, el tiempo de entrenamiento aumentó con-

siderablemente para los métodos de gráficas CART e ID3 con los datos originales. Esto indicaría que se necesitó una mayor cantidad de épocas para el entrenamiento, y como tuvo una mayor exactitud (figuras 4.1 y 4.2), significa que empezó a sesgar los resultados dando mayor importancia en clasificar la clase 2, dada la sobre-representación de esta clase sobre las demás.

- De las tres gráficas asociadas a los rendimientos obtenidos por los datos originales, CART tuvo los mejores rendimientos en recall y exactitud, ya que alcanza en 10 iteraciones el mejor modelo en exactitud (Cuadro 4.1), y no tuvo un gran tiempo de entrenamiento en comparación con los demás métodos (Figura 4.1). Es decir, las características seleccionadas desde el principio tuvieron un buen rendimiento.

- Todos los métodos basados en gráficas presentaron tiempos de ejecución altos con los datos originales, pero obtuvieron buenos rendimientos de 1 a 9 variables en exactitud (Figura 4.2). Sin embargo, fueron deficientes en obtener el recall de la clase 4 (Figura 4.3).

- El método de Chi-Cuadrada mostró un desempeño deficiente, ya que se ubicó en los últimos lugares tanto en exactitud como en recall al considerar entre 15 y 22 variables, aunque presentó tiempos de ejecución bajos. Esto sugiere un mal rendimiento desde el inicio, sin mejoras significativas mediante el descenso de gradiente.

- La exactitud promedio aumenta conforme se añaden más características al modelo, mostrando un crecimiento gradual. Sin embargo, el rendimiento en exactitud varía dependiendo del método de selección de características. Los métodos como CART, PCA, LASSO e ID3 muestran un rendimiento relativamente estable, mientras que otros, como Chi-Cuadrada y Correlación de Pearson, presentan variaciones más significativas en el rendimiento, especialmente con menos de 10 características.

Se pueden hacer varias observaciones sobre el comportamiento de la eficiencia de las redes neuronales. En general, los resultados son caóticos y no se alcanza un rendimiento óptimo. En este caso, el recall es la métrica más importante, ya que ayuda en problemas de clasificación con clases desbalanceadas, reflejando la capacidad del modelo para identificar correctamente los

elementos de una clase. Los resultados obtenidos con los datos originales no son suficientes para concluir sobre las mejores características en nuestro caso de estudio, debido a las bajas tasas de recall. Por tal razón, se implementa el algoritmo de SMOTE (1.6) para generar datos artificiales y mejorar la clasificación de la clase 4. Los resultados se muestran en las figuras 4.5, 4.6 y 4.4.

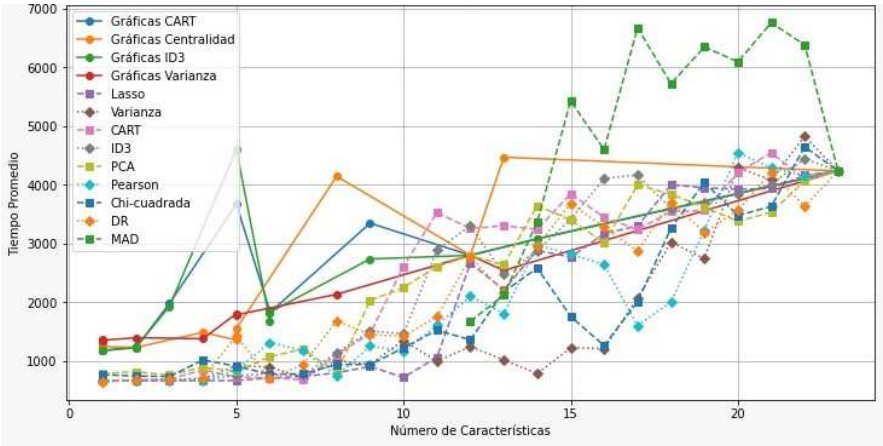



Figura 4.1: Características en función del tiempo (Originales)

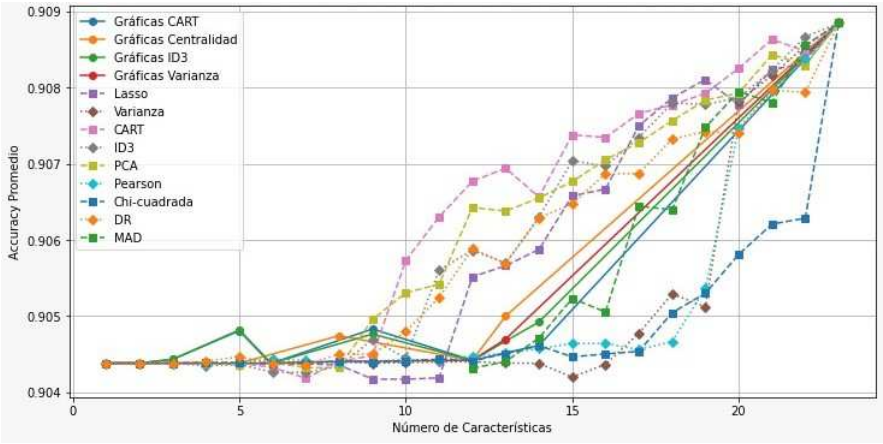



Figura 4.2: Exactitud de características (Originales)

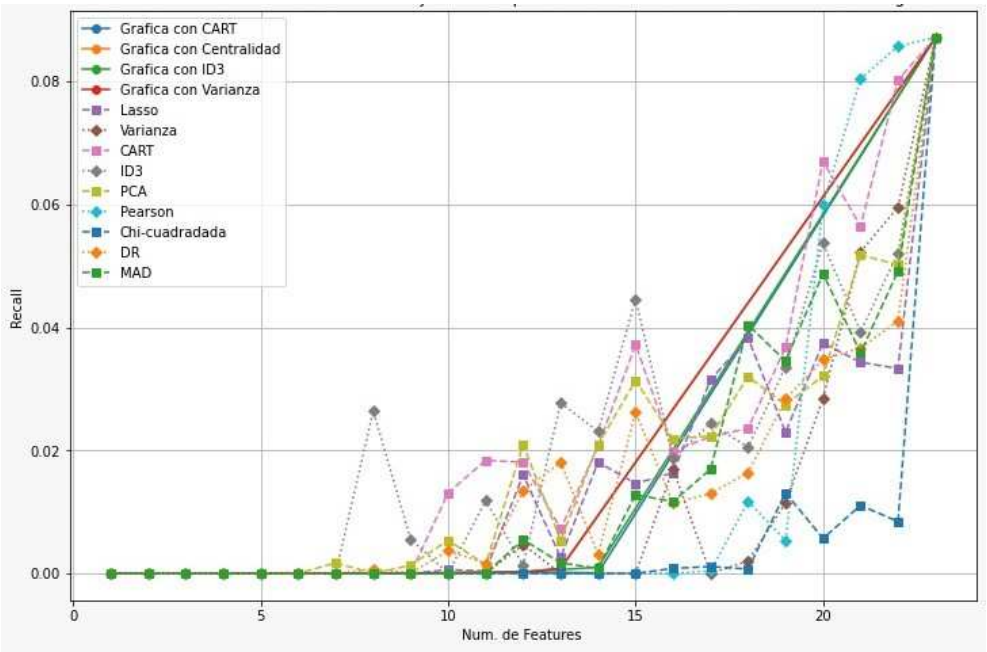



Figura 4.3: Recall de características (Originales)

Cuadro 4.1: Resultados finales de los mejores y peores métodos en cuestión de tiempo y exactitud de los datos originales.

| Datos Originales | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Tiempo | | | | | Exactitud | | | | |
| Núm. Variables | Mejor método | Tiempo mínimo (s.) | Peor método | Tiempo max (s.) | Diferencia | Mejor método | Exactitud max | Peor método | Exactitud min | Diferencia |
| **1** | DR | 640.370019 | Gráficas Varianza | 1346.800027 | 706.430008 | - | 0.904384 | - | 0.904384 | 0.000000 |
| **2** | Pearson | 654.941282 | Gráficas Varianza | 1390.622553 | 735.681271 | - | 0.904384 | - | 0.904384 | 0.000000 |
| **3** | Pearson | 660.122759 | Gráficas CART | 1976.384853 | 1316.262094 | Gráficas CART | 0.904434 | DR | 0.904384 | 0.000050 |
| **4** | LASSO | 659.754435 | Gráficas Centralidad | 1482.173734 | 822.419300 | DR | 0.904409 | ID3 | 0.904348 | 0.000061 |
| **5** | LASSO | 660.940070 | Gráficas ID3 | 4596.261064 | 3935.320993 | Gráficas ID3 | 0.904819 | PCA | 0.904361 | 0.000458 |
| **6** | DR | 692.275152 | Gráficas CART | 1846.628219 | 1154.353067 | Pearson | 0.904435 | ID3 | 0.904265 | 0.000170 |
| **7** | CART | 673.111259 | PCA | 1197.128944 | 524.017686 | Pearson | 0.904418 | CART | 0.904185 | 0.000233 |
| **8** | Pearson | 735.853881 | Gráficas Centralidad | 4139.617955 | 3403.764074 | Gráficas Centralidad | 0.904744 | PCA | 0.904329 | 0.000416 |
| **9** | LASSO | 900.695782 | Gráficas CART | 3341.315331 | 2440.619548 | PCA | 0.904956 | LASSO | 0.904175 | 0.000781 |
| **10** | LASSO | 714.555366 | CART | 2602.957218 | 1888.401852 | CART | 0.905730 | LASSO | 0.904173 | 0.001557 |
| **11** | Varianza | 989.338494 | CART | 3522.684744 | 2533.346250 | CART | 0.906303 | LASSO | 0.904192 | 0.002112 |
| **12** | Varianza | 1242.195288 | ID3 | 3298.000541 | 2055.805253 | CART | 0.906770 | MAD | 0.904316 | 0.002455 |
| **13** | Varianza | 1010.820074 | Gráficas Centralidad | 4463.757315 | 3452.937241 | CART | 0.906941 | Varianza | 0.904387 | 0.002553 |
| **14** | Varianza | 780.327760 | PCA | 3638.011544 | 2857.683784 | CART | 0.906563 | Varianza | 0.904380 | 0.002183 |
| **15** | Varianza | 1220.425912 | MAD | 5419.769402 | 4199.343490 | CART | 0.907378 | Varianza | 0.904201 | 0.003177 |
| **16** | Varianza | 1205.169708 | MAD | 4594.081643 | 3388.911935 | CART | 0.907346 | Varianza | 0.904369 | 0.002978 |
| **17** | Pearson | 1596.849443 | MAD | 6668.714478 | 5071.865034 | CART | 0.907659 | Chi Cuadrada | 0.904538 | 0.003121 |
| **18** | Pearson | 1998.824580 | MAD | 5716.538873 | 3717.714293 | LASSO | 0.907864 | Pearson | 0.904668 | 0.003195 |
| **19** | Varianza | 2738.380961 | MAD | 6342.260796 | 3603.879835 | LASSO | 0.908105 | Varianza | 0.905120 | 0.002985 |
| **20** | PCA | 3374.487402 | MAD | 6087.608769 | 2713.121367 | CART | 0.908253 | Chi Cuadrada | 0.905812 | 0.002441 |
| **21** | PCA | 3526.958179 | MAD | 6754.187491 | 3227.229312 | CART | 0.908632 | Chi Cuadrada | 0.906209 | 0.002423 |
| **22** | DR | 3627.544373 | MAD | 6377.850812 | 2750.306439 | ID3 | 0.908666 | Chi Cuadrada | 0.906286 | 0.002379 |
| **23** | - | 4230.526678 | - | 4230.526678 | 0.000000 | - | 0.908847 | - | 0.908847 | 0.000000 |

Cuadro 4.2: Resultados finales de los mejores y peores métodos en cuestión del número de características y su Recall con los datos originales.

| **Mejores 6 métodos por cantidad de variables con su respectivo Recall con datos originales** | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Num. variables 1-6: | | Num. variables 10: | | Num. variables 14: | | Num. variables 18: | | Num. variables 22: | |
| 0.0000 | Todos | 0.0131 | CART | 0.0231 | ID3 | 0.0404 | MAD | 0.0857 | Pearson |
| | | 0.0054 | PCA | 0.0209 | PCA | 0.0383 | LASSO | 0.0803 | CART |
| | | 0.0037 | DR | 0.0207 | CART | 0.0320 | PCA | 0.0595 | Varianza |
| | | 0.0006 | LASSO | 0.0181 | LASSO | 0.0236 | CART | 0.0521 | ID3 |
| | | 0.0000 | Varianza | 0.0031 | DR | 0.0205 | ID3 | 0.0502 | PCA |
| | | 0.0000 | ID3 | 0.0010 | Grafica ID3 | 0.0164 | DR | 0.0492 | MAD |
| Num. variables 7: | | Num. variables 11: | | Num. variables 15: | | Num. variables 19: | | Num. variables 23: | |
| 0.0017 | PCA | 0.0184 | CART | 0.0446 | ID3 | 0.0368 | CART | 0.0872 | Grafica CART |
| 0.0000 | Métodos restantes | 0.0119 | ID3 | 0.0372 | CART | 0.0345 | MAD | 0.0872 | Grafica Centralidad |
| | | 0.0016 | DR | 0.0314 | PCA | 0.0336 | ID3 | 0.0872 | Grafica ID3 |
| | | 0.0014 | PCA | 0.0262 | DR | 0.0283 | DR | 0.0872 | Grafica Varianza |
| | | 0.0004 | LASSO | 0.0146 | LASSO | 0.0274 | PCA | 0.0872 | LASSO |
| | | 0.0000 | Varianza | 0.0128 | MAD | 0.0230 | LASSO | 0.0872 | Varianza |
| Num. variables 8: | | Num. variables 12: | | Num. variables 16: | | Num. variables 20: | | | |
| 0.0264 | ID3 | 0.0209 | PCA | 0.0219 | PCA | 0.0671 | CART | | |
| 0.0007 | DR | 0.0181 | CART | 0.0199 | CART | 0.0599 | Pearson | | |
| 0.0000 | Métodos restantes | 0.0161 | LASSO | 0.0188 | ID3 | 0.0537 | ID3 | | |
| | | 0.0134 | DR | 0.0170 | Varianza | 0.0488 | MAD | | |
| | | 0.0055 | MAD | 0.0164 | LASSO | 0.0375 | LASSO | | |
| | | 0.0047 | Varianza | 0.0117 | MAD | 0.0347 | DR | | |
| Num. variables 9: | | Num. variables 13: | | Num. variables 17: | | Num. variables 21: | | | |
| 0.0056 | ID3 | 0.0277 | ID3 | 0.0315 | LASSO | 0.0805 | Pearson | | |
| 0.0013 | PCA | 0.0181 | DR | 0.0244 | ID3 | 0.0565 | CART | | |
| 0.0001 | DR | 0.0072 | CART | 0.0224 | PCA | 0.0523 | Varianza | | |
| 0.0000 | Métodos restantes | 0.0054 | PCA | 0.0222 | CART | 0.0518 | PCA | | |
| | | 0.0026 | LASSO | 0.0170 | MAD | 0.0393 | ID3 | | |
| | | 0.0017 | MAD | 0.0130 | DR | 0.0367 | DR | | |

Para los datos generados por SMOTE, se pueden hacer las siguientes observaciones:

- Se nota una disminución considerable en la exactitud en los datos originales, de un 90.8 % (Cuadro 4.1) a un 75.4 % (Cuadro 4.3) con los modelos de redes neuronales. Sin embargo, para la métrica de recall, se mejoró de un 9 % (Cuadro 4.2) a un 89 % (Cuadro 4.4). Esto resalta la importancia de esta técnica cuando el objetivo es identificar elementos de clases minoritarias.

- Es importante destacar que este incremento en el **recall** y disminución en la exactitud global del modelo se debe a que al balancear las clases minoritarias mediante SMOTE, el modelo se vuelve más sensible a estas clases, aumentando la probabilidad de identificarlas correctamente (mayor recall), pero también incrementa la posibilidad de clasificar erróneamente elementos de otras clases (menor exactitud).

- El uso de SMOTE añade un costo computacional adicional, lo que resulta en un mayor tiempo de ejecución en todos los métodos de selección de características (Figura 4.4). Esta técnica requiere generar instancias sintéticas, lo cual incrementa el número de observaciones a procesar y, por ende, un aumento en el tiempo de entrenamiento de cinco a siete veces. Sin embargo, este costo adicional debe considerarse en relación con las mejoras obtenidas en recall y exactitud.

En general, el uso de SMOTE muestra una disminución en la exactitud promedio, aunque el impacto es más notorio en el recall, resaltando la importancia de esta técnica para identificar instancias de clases minoritarias. No obstante, el impacto de esta técnica también depende del método de selección de características empleado. Esto puede observarse claramente en las gráficas 4.5 y 4.6, donde se evidencia cómo diferentes métodos influyen en el rendimiento del modelo.

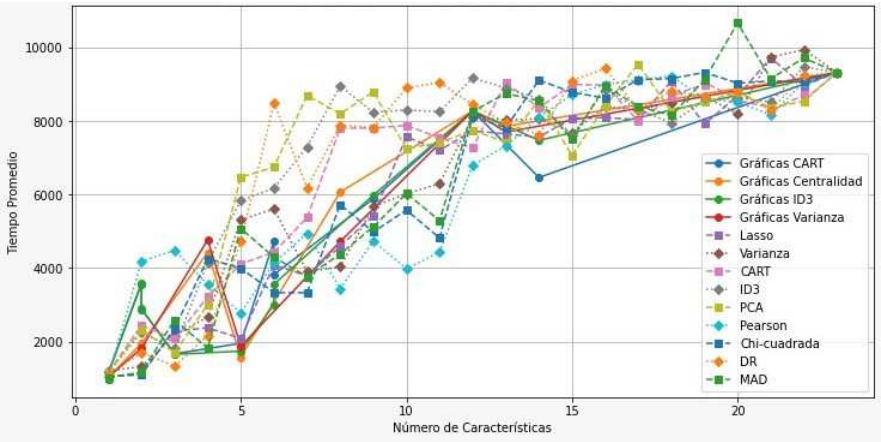



Figura 4.4: Características en función del tiempo (SMOTE)

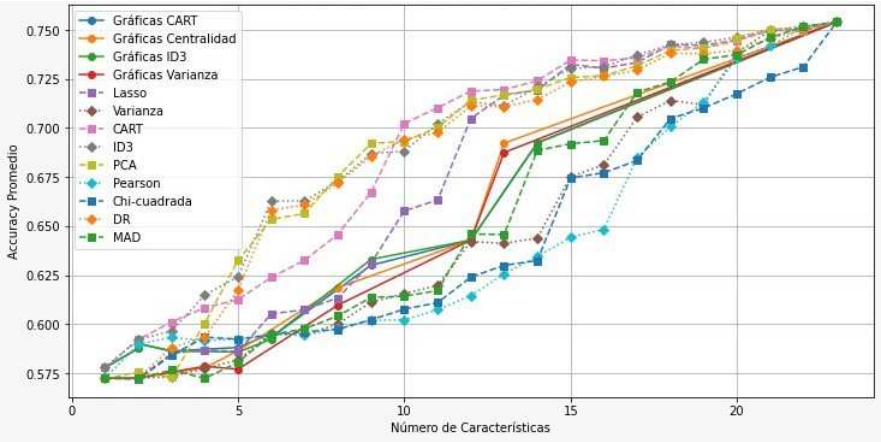



Figura 4.5: Exactitud de características (SMOTE)

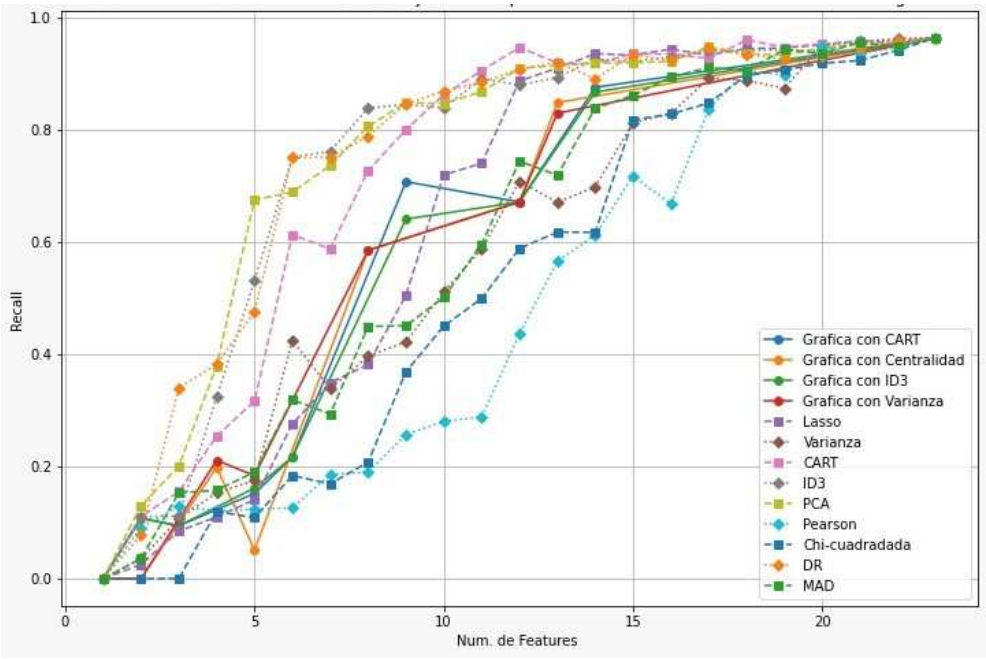



Figura 4.6: Recall de características (SMOTE)

Cuadro 4.3: Resultados finales de los mejores y peores métodos en cuestión de tiempo y exactitud de los datos generados por SMOTE.

| Datos generados por SMOTE | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Tiempo | | | | | Exactitud | | | | |
| **Núm. Variables** | **Mejor método** | **Tiempo mínimo (s.)** | **Peor método** | **Tiempo max (s.)** | **Diferencia** | **Mejor método** | **Exactitud max** | **Peor método** | **Exactitud min** | **Diferencia** |
| **1** | Gráfica ID3 | 977.495286 | Varianza | 1202.650535 | 225.155249 | CART | 0.577845 | Gráfica Varianza | 0.572409 | 0.005436 |
| **2** | Chi Cuadrada | 1097.899384 | Pearson | 4183.583561 | 3085.684178 | CART | 0.592116 | Chi Cuadrada | 0.571986 | 0.020130 |
| **3** | DR | 1333.356089 | Pearson | 4484.394392 | 3151.038303 | CART | 0.601109 | PCA | 0.572949 | 0.028160 |
| **4** | MAD | 1821.291966 | Gráfica Varianza | 4767.977979 | 2946.686013 | ID3 | 0.614843 | MAD | 0.572375 | 0.042468 |
| **5** | Gráfica centralidad | 1580.572958 | PCA | 6471.512323 | 4890.939365 | PCA | 0.632553 | Gráfica Varianza | 0.577045 | 0.055509 |
| **6** | Gráfica ID3 | 3010.597616 | DR | 8489.910544 | 5479.312927 | ID3 | 0.662681 | Gráfica ID3 | 0.592329 | 0.070352 |
| **7** | Chi Cuadrada | 3330.986975 | PCA | 8698.269343 | 5367.282368 | ID3 | 0.662912 | Pearson | 0.594380 | 0.068532 |
| **8** | Pearson | 3441.429967 | ID3 | 8953.220631 | 5511.790664 | PCA | 0.675307 | Chi Cuadrada | 0.597384 | 0.077923 |
| **9** | Pearson | 4735.428231 | PCA | 8792.713803 | 4057.285572 | PCA | 0.692042 | Pearson | 0.601901 | 0.090141 |
| **10** | Pearson | 3982.522343 | DR | 8900.132912 | 4917.610569 | CART | 0.702295 | Pearson | 0.602262 | 0.100033 |
| **11** | Pearson | 4431.343285 | DR | 9047.111199 | 4615.767914 | CART | 0.710330 | Pearson | 0.607275 | 0.103055 |
| **12** | Pearson | 6799.929365 | ID3 | 9172.684928 | 2372.755563 | CART | 0.718777 | Pearson | 0.614538 | 0.104239 |
| **13** | Pearson | 7321.285374 | CART | 9045.498037 | 1724.212663 | CART | 0.719572 | Pearson | 0.625646 | 0.093926 |
| **14** | Gráfica CART | 6464.646221 | Chi Cuadrada | 9112.428895 | 2647.782674 | CART | 0.724028 | Chi Cuadrada | 0.632435 | 0.091593 |
| **15** | PCA | 7063.491824 | DR | 9091.716077 | 2028.224253 | CART | 0.734644 | Pearson | 0.644246 | 0.090398 |
| **16** | LASSO | 8093.693891 | DR | 9425.087368 | 1331.393477 | CART | 0.734239 | Pearson | 0.648390 | 0.085848 |
| **17** | CART | 7995.079044 | PCA | 9547.707560 | 1552.628516 | ID3 | 0.737230 | Chi Cuadrada | 0.683462 | 0.053768 |
| **18** | ID3 | 7951.860226 | Pearson | 9221.920996 | 1270.060770 | ID3 | 0.742491 | Pearson | 0.700639 | 0.041852 |
| **19** | LASSO | 7920.095808 | Chi Cuadrada | 9322.063516 | 1401.967707 | ID3 | 0.743705 | Chi Cuadrada | 0.710167 | 0.033538 |
| **20** | Varianza | 8198.953351 | MAD | 10665.459741 | 2466.506391 | ID3 | 0.746309 | Chi Cuadrada | 0.717437 | 0.028872 |
| **21** | Pearson | 8172.334674 | Varianza | 9733.049307 | 1560.714633 | PCA | 0.750244 | Chi Cuadrada | 0.725834 | 0.024410 |
| **22** | PCA | 8536.766018 | Varianza | 9940.388931 | 1403.622912 | Varianza | 0.751919 | Chi Cuadrada | 0.731170 | 0.020749 |
| **23** | - | 9319.369412 | - | 9319.369412 | 0.000000 | - | 0.754003 | - | 0.754003 | 0.000000 |

Cuadro 4.4: Resultados finales de los mejores y peores métodos en cuestión del número de características y su Recall con los datos generados por SMOTE.

| **Mejores 6 métodos por cantidad de variables con su respectivo Recall con datos generados por SMOTE** | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Num. de variables 1** | | **Num. de variables 5** | | **Num. de variables 9** | | **Num. de variables 13** | | **Num. de variables 17** | | **Num. de variables 21** | |
| 0.0000 | Todos | 0.6756 | PCA | 0.8485 | PCA | 0.9203 | DR | 0.9484 | PCA | 0.9588 | LASSO |
| | | 0.5309 | ID3 | 0.8464 | DR | 0.9193 | CART | 0.9455 | DR | 0.9587 | CART |
| | | 0.4753 | DR | 0.8453 | ID3 | 0.9148 | PCA | 0.9415 | ID3 | 0.9579 | Varianza |
| | | 0.3168 | CART | 0.8005 | CART | 0.9103 | LASSO | 0.9338 | LASSO | 0.9578 | ID3 |
| | | 0.1901 | MAD | 0.7077 | Grafica CART | 0.8934 | ID3 | 0.9274 | CART | 0.9566 | MAD |
| | | 0.1838 | Grafica Varianza | 0.6410 | Grafica ID3 | 0.8486 | Grafica Centralidad | 0.9113 | MAD | 0.9529 | PCA |
| **Num. de variables 2** | | **Num. de variables 6** | | **Num. de variables 10** | | **Num. de variables 14** | | **Num. de variables 18** | | **Num. de variables 22** | |
| 0.1295 | PCA | 0.7515 | ID3 | 0.8690 | DR | 0.9358 | LASSO | 0.9602 | CART | 0.9630 | ID3 |
| 0.1107 | CART | 0.7503 | DR | 0.8627 | CART | 0.9258 | ID3 | 0.9451 | LASSO | 0.9604 | Varianza |
| 0.1107 | ID3 | 0.6892 | PCA | 0.8474 | PCA | 0.9189 | PCA | 0.9420 | ID3 | 0.9592 | DR |
| 0.1080 | Grafica CART | 0.6128 | CART | 0.8395 | ID3 | 0.9184 | CART | 0.9362 | PCA | 0.9545 | CART |
| 0.1080 | Grafica ID3 | 0.4248 | Varianza | 0.7199 | LASSO | 0.8913 | DR | 0.9336 | DR | 0.9541 | LASSO |
| 0.0911 | Pearson | 0.3190 | MAD | 0.5115 | Varianza | 0.8759 | Grafica CART | 0.9061 | Pearson | 0.9496 | PCA |
| **Num. de variables 3** | | **Num. de variables 7** | | **Num. de variables 11** | | **Num. de variables 15** | | **Num. de variables 19** | | **Num. de variables 23** | |
| 0.3399 | DR | 0.7619 | ID3 | 0.9058 | CART | 0.9356 | CART | 0.9476 | CART | 0.9643 | Grafica CART |
| 0.2006 | PCA | 0.7516 | DR | 0.8912 | ID3 | 0.9338 | DR | 0.9460 | LASSO | 0.9643 | Grafica Centralidad |
| 0.1551 | CART | 0.7366 | PCA | 0.8848 | DR | 0.9334 | LASSO | 0.9436 | MAD | 0.9643 | Grafica ID3 |
| 0.1530 | MAD | 0.5881 | CART | 0.8679 | PCA | 0.9226 | ID3 | 0.9420 | ID3 | 0.9643 | Grafica Varianza |
| 0.1288 | Pearson | 0.3492 | LASSO | 0.7406 | LASSO | 0.9205 | PCA | 0.9331 | PCA | 0.9643 | LASSO |
| 0.1288 | ID3 | 0.3387 | Varianza | 0.5963 | MAD | 0.8617 | MAD | 0.9264 | DR | 0.9643 | Varianza |
| **Num. de variables 4** | | **Num. de variables 8** | | **Num. de variables 12** | | **Num. de variables 16** | | **Num. de variables 20** | | | |
| 0.3825 | DR | 0.8391 | ID3 | 0.9470 | CART | 0.9439 | LASSO | 0.9537 | CART | | |
| 0.3793 | PCA | 0.8064 | PCA | 0.9100 | PCA | 0.9352 | CART | 0.9518 | LASSO | | |
| 0.3240 | ID3 | 0.7888 | DR | 0.9081 | DR | 0.9295 | ID3 | 0.9502 | Varianza | | |
| 0.2542 | CART | 0.7278 | CART | 0.8874 | LASSO | 0.9237 | DR | 0.9478 | Pearson | | |
| 0.2111 | Grafica Varianza | 0.5857 | Grafica Varianza | 0.8802 | ID3 | 0.9218 | PCA | 0.9470 | PCA | | |
| 0.1987 | Grafica Centralidad | 0.5851 | Grafica Centralidad | 0.7444 | MAD | 0.8942 | MAD | 0.9408 | ID3 | | |

Para visualizar la importancia de cada variable en la clasificación de los índices en las redes neuronales, se creó un heatmap con los seis mejores subconjuntos de características seleccionados por cada método. Después, se verificó el número de veces que cada característica fue seleccionada dentro de estos seis mejores subconjuntos. Al final, los conteos fueron normalizados dividiendo entre el total de subconjuntos generados por cada método. Para los métodos de selección basados en Teoría de Gráficas con pesos de CART e ID3, se dividió entre 6; para los métodos de varianza baja y centralidad, se dividió entre 5; y para los demás métodos, se dividió entre 21. Es importante mencionar que no se consideraron en el conteo los casos en los que el número de variables es 23 y 1, ya que en estos casos todos los métodos tienen el mismo rendimiento en el Recall, por lo tanto, no son relevantes para el análisis. Tanto las filas (elementos) como las columnas (características) fueron ordenadas en función de la suma de sus valores, de mayor a menor, lo que permite observar primero las características y elementos más relevantes en el proceso de selección, como se muestra en la Figura 4.7.

Al analizar el heatmap generado, se pueden extraer varias observaciones sobre los patrones de selección de características y los métodos utilizados:

- **Preferencia por ciertos elementos**: Las características como temperatura ambiente máxima, temperatura del aire, irradiación, potencial de irradiancia, horas sol, y presión atmosférica tienen una alta frecuencia de selección en la mayoría de los mejores métodos, lo que nos indica que son los más relevantes para los modelos empleados.

- **Elementos de menor relevancia**: Algunas características, como "velocidad del viento (desviación estándar)", "índice de calor promedio" y "velocidad del viento máxima", tienen un conteo bajo en varias columnas. Es decir, las veces que fueron seleccionadas por algún método no fueron significativas para mejorar el rendimiento, salvo la característica "índice de calor " estuvo presente en la mayoría de los subconjuntos generados por ID3.

- **Consistencia en la selección**: Cada método tiende a seleccionar ciertos elementos de manera consistente. Por ejemplo, CART, ID3 y PCA muestran una distribución amplia de mejores elementos seleccionados, lo que indica una mayor variedad de características relevantes.

- Se observa la ausencia de una columna para Chi-Cuadrada, lo que indica que este método tuvo el peor desempeño y en el caso de Gráfica ID3, los menos de la mitad de los subconjuntos generados por estos, no fueron seleccionados como los mejores.

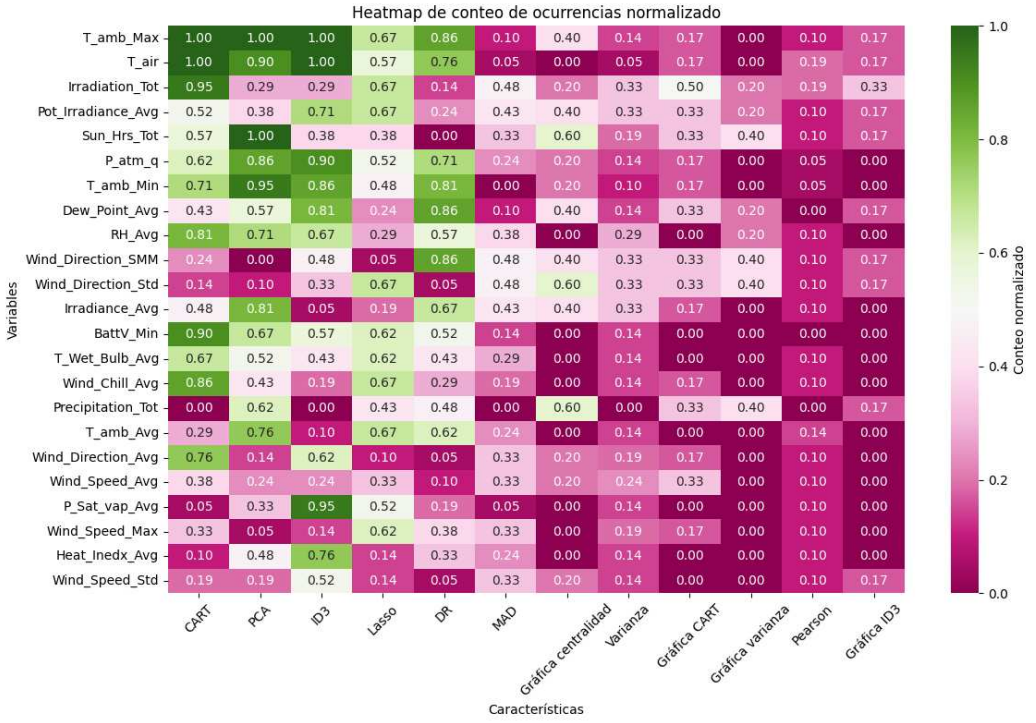



Figura 4.7: "Heatmap de conteo de ocurrencias normalizado para diferentes variables seleccionadas por varios métodos de selección de características. Los valores están normalizados de 0 a 1: un valor de 0 indica baja relevancia para el método o una contribución no significativa al rendimiento del subconjunto de variables. Un valor de 1 indica que la variable estuvo presente en todos los subconjuntos generados por el método y que dichos subconjuntos se encontraron entre los 6 mejores."

Notemos que nuestra metodología basada en Teoría de Gráficas para la selección de conjuntos de variables considera varios aspectos importantes:

- A diferencia de las técnicas de selección individual, nuestros modelos consideran las relaciones entre las variables, lo que permite reducir la

redundancia de información al identificar y eliminar características que contienen información similar.

- Con respecto a las técnicas de selección por pares, nuestro método se diferencia al considerar no solo la relación directa entre variables y el objetivo, sino también las interrelaciones entre las distintas variables. Extraemos información sobre la importancia de los vértices utilizando medidas de centralidad, varianza e importancia derivada de los algoritmos ID3 y CART, combinadas con una reducción basada en conjuntos dominantes independientes, a partir de las relaciones establecidas por la métrica EMD.

- Nuestro modelo se enfoca únicamente en las relaciones entre las variables meteorológicas, sin considerar la relación directa con la variable objetivo. Esto permite que nuestros conjuntos de características sean de alta relevancia no solo para predecir la variable objetivo, sino también entre ellas mismas, lo que abre nuevas posibilidades para la predicción de otras variables.

El análisis detallado de los métodos de selección de características mediante heatmap y métricas de desempeño revela que no existe un método único que considere completamente todas las dimensiones de importancia en un conjunto de datos. Cada método aporta una perspectiva única: algunos capturan la dispersión en los datos, otros evalúan la correlación con la variable objetivo, y otros identifican relaciones entre las variables mismas mediante Teoría de Gráficas. Sin embargo, los métodos **CART**, **PCA**, **ID3** y **LASSO** se destacan por su consistencia y efectividad, maximizando tanto la frecuencia de selección de características en el heatmap como los valores de *recall*, lo que los convierte en los métodos preferidos para la selección de características en este contexto.

# Capítulo 5

# Conclusiones y trabajo futuro

En el contexto de la predicción Índice Aire Salud, priorizar el recall es la tarea más importante, ya que es más importante identificar correctamente los casos en los que se debe disminuir el flujo vehicular por alta contaminación, aunque eso implique una menor exactitud en otras clases.

En este proyecto se ha demostrado que, aunque los subconjuntos de características puedan ser similares en su composición, como ocurre en los casos de ID3 y CART, la importancia de cada característica y su impacto en el *recall* están determinados en gran medida por su interacción con el subconjunto de variables al que pertenece. Esto es evidente en el heatmap, donde observamos cómo el contexto de un conjunto de variables puede influir en la relevancia de una característica particular, resaltando la necesidad de analizar cada variable dentro de su conjunto en lugar de tratarla de forma aislada.

Además, se identificaron ciertas variables que, independientemente del método utilizado, destacan por su importancia debido a su estrecha relación con el Índice Aire Salud. En particular, la **temperatura ambiente máxima** y la **temperatura del aire** resultan ser variables con mayor contribución al modelo de clasificación. Este análisis resalta implicaciones significativas para la predicción del índice de calidad del aire:

- **Incremento en la precisión de los modelos predictivos**: Al centrarse en estas variables esenciales, los modelos pueden reflejar de manera más precisa los procesos que influyen en la formación y dispersión

de contaminantes, generando predicciones más confiables.

- **Uso eficiente de recursos**: Reconocer las variables más influyentes permite asignar los recursos del monitoreo ambiental de manera más efectiva, priorizando las que aportan mayor relevancia.
- **Mejor entendimiento de los factores importantes**: Estas variables están íntimamente relacionadas con procesos atmosféricos que contribuyen al aumento de la concentración de contaminantes, influenciados por altas temperaturas y una mayor irradiación solar.

Por otro lado, algunas características como la **velocidad del viento (desviación estándar)**, el **índice de calor promedio** y la **velocidad del viento máxima** tienen un conteo bajo en varios métodos. Esto sugiere que, en el contexto específico del proyecto, estos factores no son determinantes para la calidad del aire. Su exclusión puede simplificar el modelo sin afectar significativamente su desempeño, facilitando así su interpretación.

Se encontró que el uso de SMOTE mejora significativamente el *recall* (ver figuras 4.3 y 4.6), favoreciendo la identificación de clases minoritarias, aunque con una disminución en la *exactitud promedio* (ver figuras 4.2 y 4.5). Esto ocurre porque al balancear las clases minoritarias, el modelo se vuelve más sensible a estas clases, aumentando la probabilidad de identificarlas correctamente (mayor *recall*), pero también incrementa la posibilidad de clasificar erróneamente instancias de otras clases (menor *exactitud*). En el contexto de la predicción del Índice Aire Salud, priorizar el *recall* es más importante identificar correctamente los casos de alta contaminación para proteger la salud pública.

De acuerdo con los resultados obtenidos, es necesario incluir al menos ocho variables en el conjunto de entrada para alcanzar una tasa de acierto cercana al 80 %, como se observa en la sección 1.1.3. Los cuadros 4.4 y 3.6 muestran que las ocho variables más importantes identificadas por los métodos ID3 y PCA son suficientes para lograr este resultado.

En cuanto a los métodos de selección de características, se encontró que métodos como **CART**, **PCA**, **ID3** y **LASSO** se destacan como los más eficientes, ocupando los primeros lugares en diversos subconjuntos de variables. Por otro lado, métodos como **Correlación de Pearson** y las gráficas generadas con **ID3** muestran subconjuntos de variables con menor relevancia,

lo cual podría hacerlos menos adecuados en aplicaciones prácticas y en el caso de **Chi-Cuadrada** fue el que tuvo peor rendimiento, ya que no nunca aparece entre los mejores seis métodos.

Los métodos enfocados en la variabilidad, tales como **Varianza baja**, **MAD** y **DR**, son útiles para identificar características con gran dispersión en los datos; sin embargo, no siempre garantizan un alto *recall*. Esto sugiere que, aunque la variabilidad es un criterio importante, su uso aislado puede no ser suficiente para maximizar el desempeño predictivo y podría ser más efectivo cuando se combina con métodos que evalúan la correlación directa con la variable objetivo (como Correlación de Pearson o Chi-Cuadrada).

El uso de un enfoque combinado que integre métodos de Teoría de Gráficas, estadísticos y de variabilidad permite desarrollar modelos robustos y precisos, especialmente en aplicaciones donde múltiples factores influyen en los resultados. A partir de este análisis, se identificaron varias áreas de trabajo futuro que podrían profundizar y expandir las soluciones actuales:

1. Crear una nueva lista de relevancia que organice las variables según la importancia observada en el heatmap. Por ejemplo, asignar menor importancia a la "Velocidad del viento estándar" y mayor a la "Temperatura ambiente máxima". Este enfoque podría mejorar los resultados al ordenar las variables en función de la discriminación proporcionada por los métodos de selección y la información que aportan.

2. Explorar nuevos enfoques en Teoría de Gráficas que permitan una discriminación más efectiva en la selección de variables, considerando las relaciones y el peso de las variables en las gráficas.

3. Investigar nuevas técnicas para medir la relación entre variables, con el objetivo de generar representaciones gráficas más robustas para la selección de características. La elección de un conjunto dominante independiente depende de las conexiones generadas en la gráfica, lo que puede dificultar la identificación de buenos subconjuntos de variables.

4. Desarrollar un modelo de predicción del Índice Aire Salud de manera continua en lugar de clasificarlo por categorías, utilizando los subconjuntos de variables obtenidos en este trabajo y evaluando con distintas métricas.

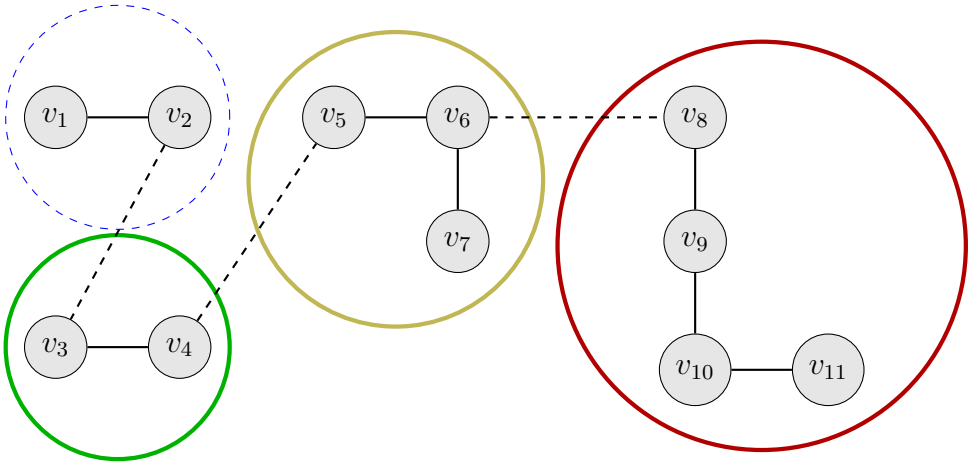



Figura 5.1: Gráfica con cuatro comunidades, cada una con un número diferente de vértices. El peso de cada vértice está determinado por el número de vértices de su comunidad asociada. Es decir, en la comunidad azul, sus vértices tienen peso 2; en la comunidad verde, sus vértices tienen peso 2; en la comunidad amarilla, sus vértices tienen peso 3; y, por último, en la comunidad roja, sus vértices tienen peso 4.

El modelado del problema mediante Teoría de Gráficas proporciona una gran flexibilidad al permitir elegir distintas métricas para medir la relación entre dos variables (vértices), como *Dynamic Time Warping* (DTW), correlación de Spearman, entre otras. Además, ofrece diversas herramientas para evaluar la importancia de cada vértice en relación con los demás, utilizando métodos como agrupación (ver Figura 5.1), propagación de etiquetas (ver Figura 5.2) y detección de comunidades (ver Figura 5.3).

A futuro, se planea profundizar en el uso de estos métodos y explorar nuevas técnicas de Teoría de Gráficas, como la agrupación y la propagación de etiquetas ya mencionadas. El objetivo principal será identificar conjuntos adicionales de variables que tengan una relación significativa con la variable objetivo, el Índice Aire Salud. Esto permitirá integrar de manera más directa la variable objetivo en la representación gráfica, como se ilustra en la Figura 5.3, fortaleciendo la comprensión y el modelado de los factores importantes que afectan su predicción.

En conclusión, el análisis detallado de los métodos de selección de características revela que no existe un método único capaz de abarcar por comple-

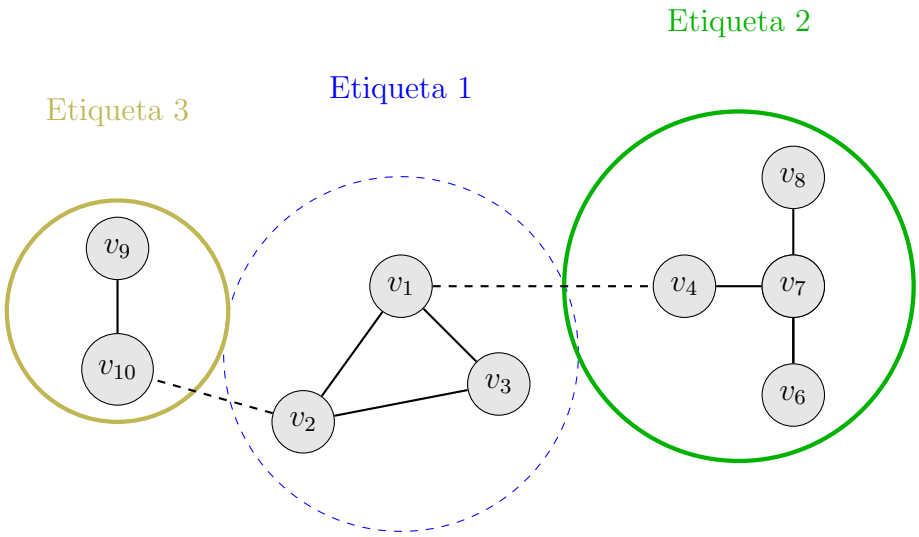



Figura 5.2: El peso de los vértices está determinado por sus respectivas etiquetas. Es decir, los vértices de la subgráfica dentro de la etiqueta 1 tendrán peso uno, los vértices de la subgráfica dentro de la etiqueta 2 tendrán peso dos, y de igual manera, los vértices de la subgráfica dentro de la etiqueta 3 tendrán peso tres.

to todas las dimensiones de importancia en un conjunto de datos. Por ello, una combinación de diferentes metodologías ayuda a desarrollar modelos más robustos y precisos. Esto es particularmente relevante en aplicaciones donde múltiples factores interactúan de manera compleja, como en los problemas ambientales.

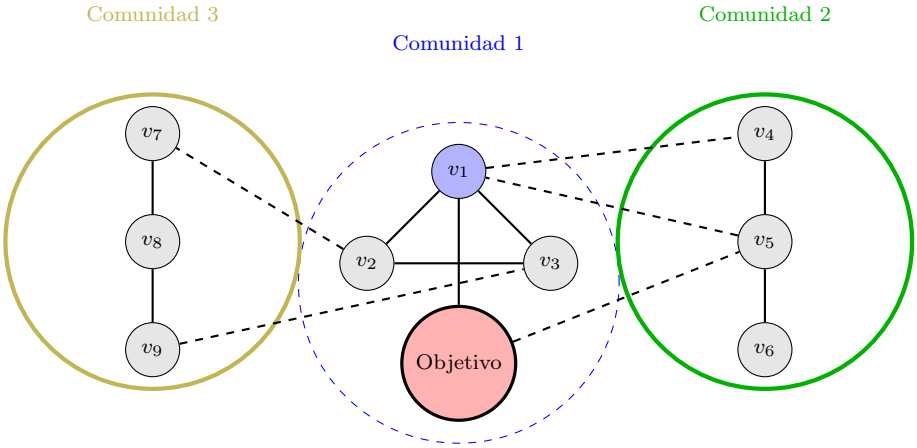



Figura 5.3: Gráfica con tres comunidades, resaltando un vértice objetivo color rojo y un vértice de color azul que representa un conjunto dominante independiente en su comunidad.

# Anexo:Distribuciones con datos originales y con SMOTE.

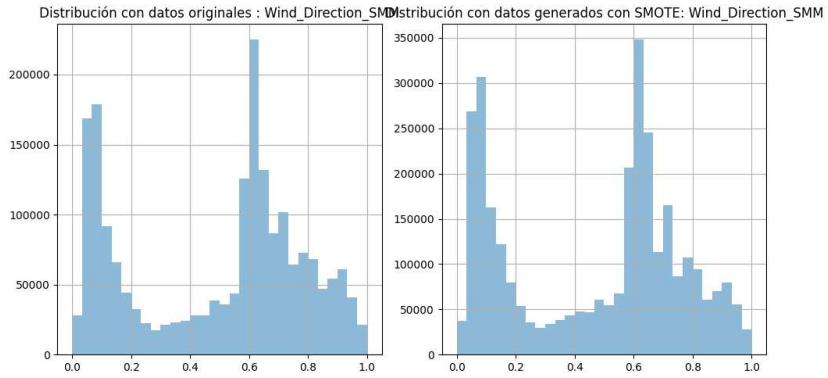



Figura 5.4: Comparación de Wind Direction Std

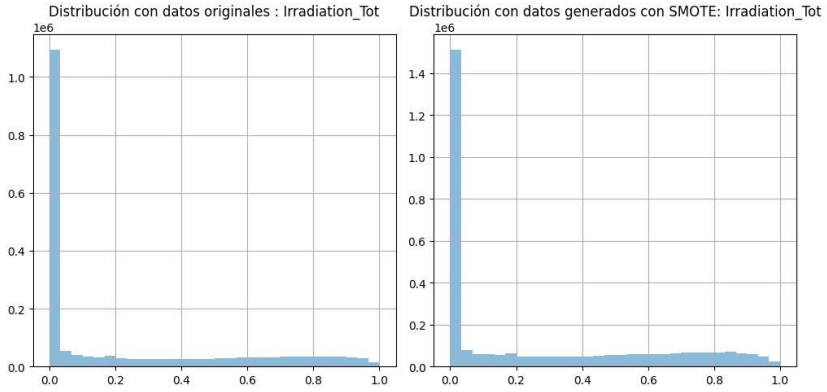



Figura 5.5: Comparación de Irradiation Tot

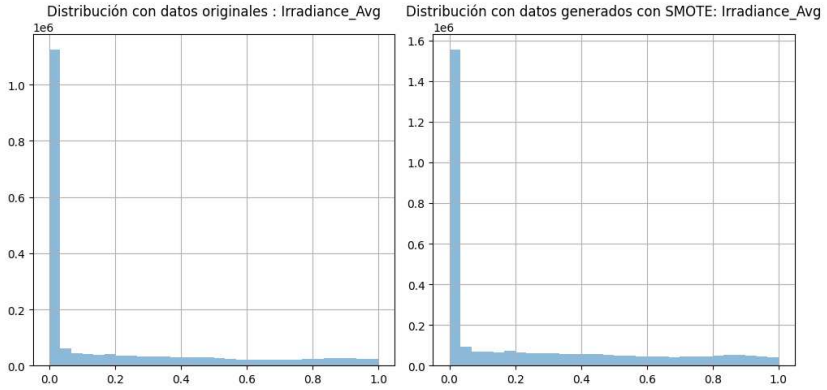



Figura 5.6: Comparación de Irradiance Avg

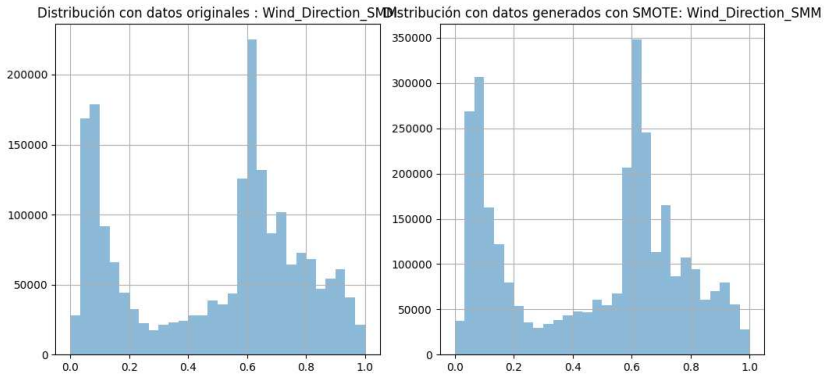



Figura 5.7: Comparación de Wind Direction SMM

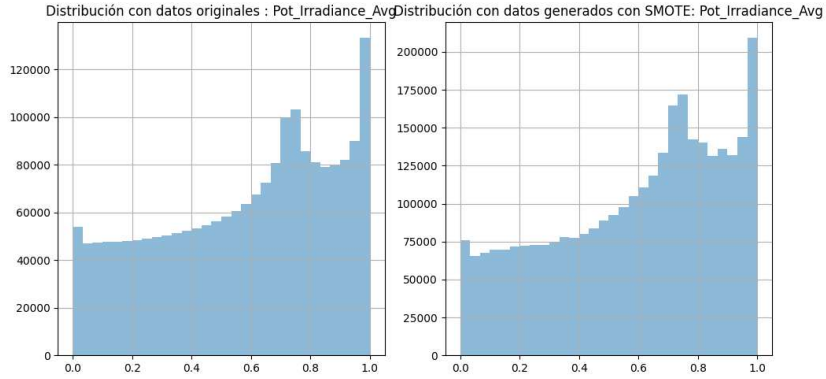



Figura 5.8: Comparación de Pot Irradiance Avg

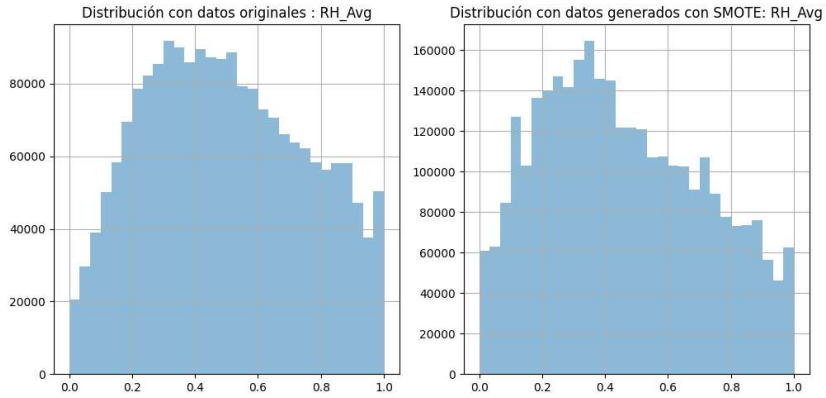



Figura 5.9: Comparación de RH Avg

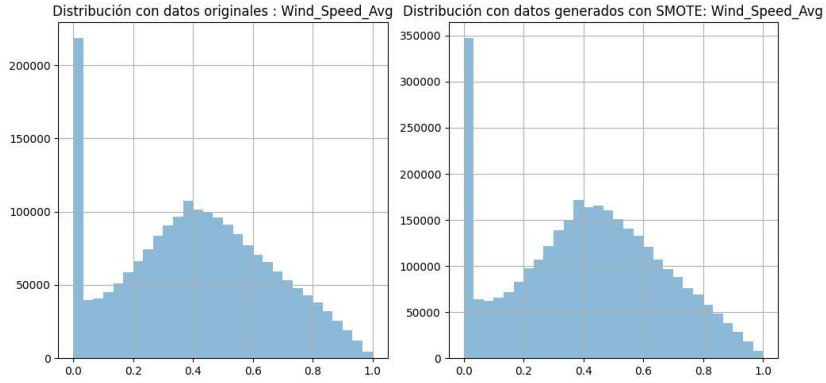



Figura 5.10: Comparación de Wind Speed Avg

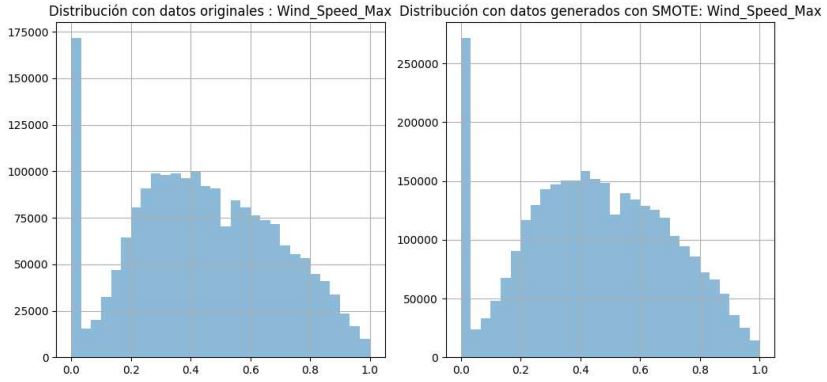



Figura 5.11: Comparación de Wind Speed Max

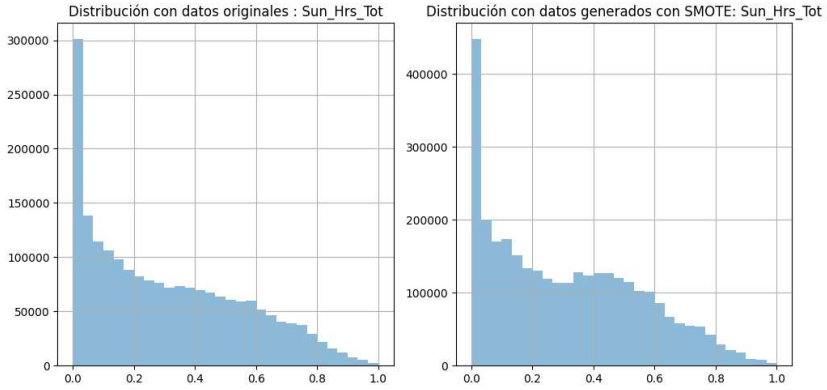



Figura 5.12: Comparación de Sun Hrs Tot

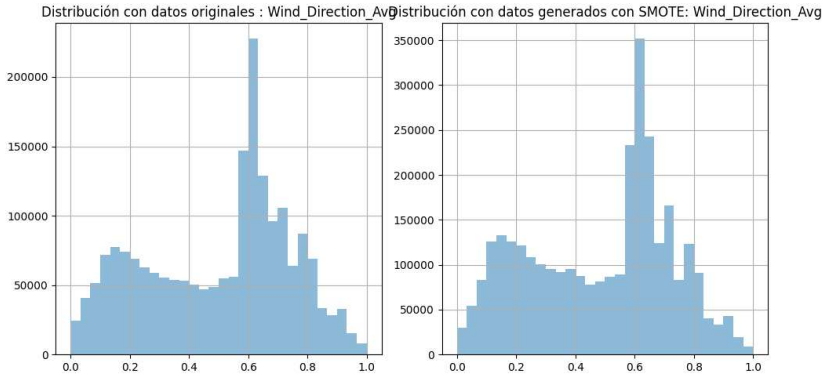



Figura 5.13: Comparación de Wind Direction Avg

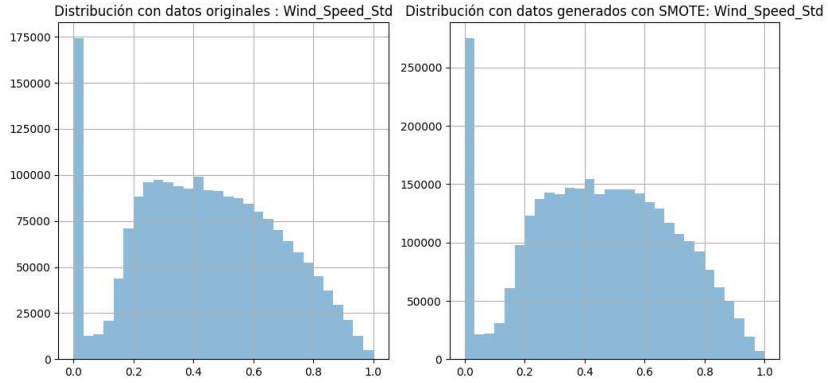



Figura 5.14: Comparación de Wind Speed Std

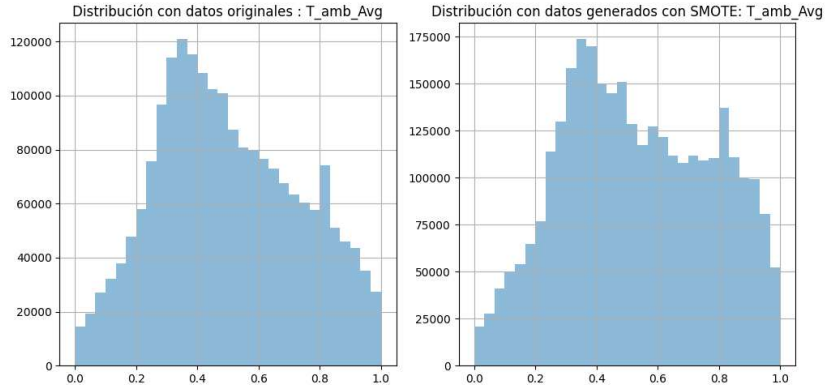



Figura 5.15: Comparación de T amb Avg

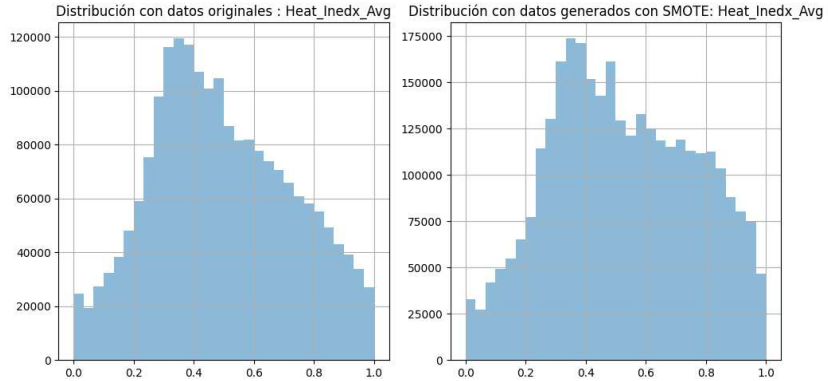



Figura 5.16: Comparación de Heat Index Avg

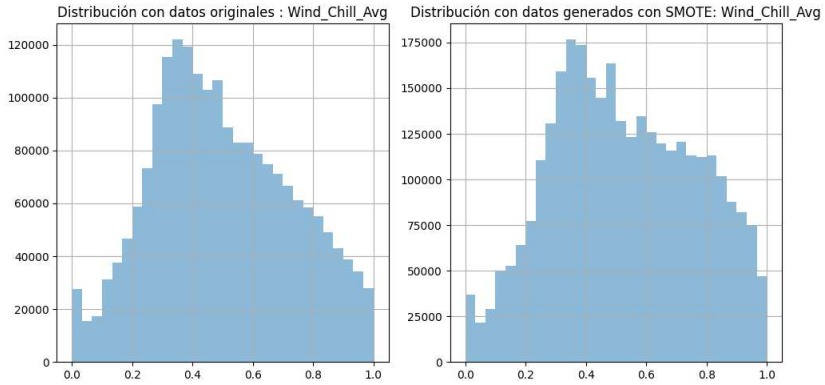



Figura 5.17: Comparación de Wind Chill Avg

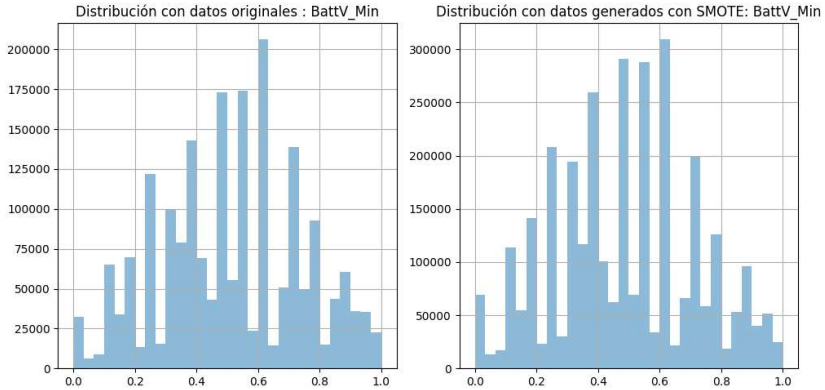



Figura 5.18: Comparación de BattV Min

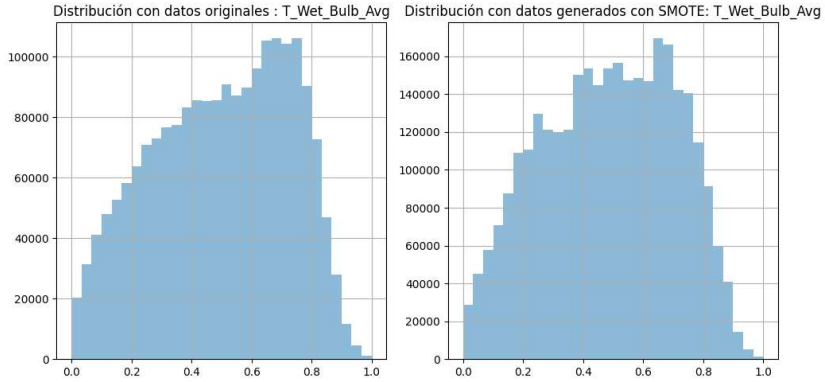



Figura 5.19: Comparación de T Wet Bulb Avg

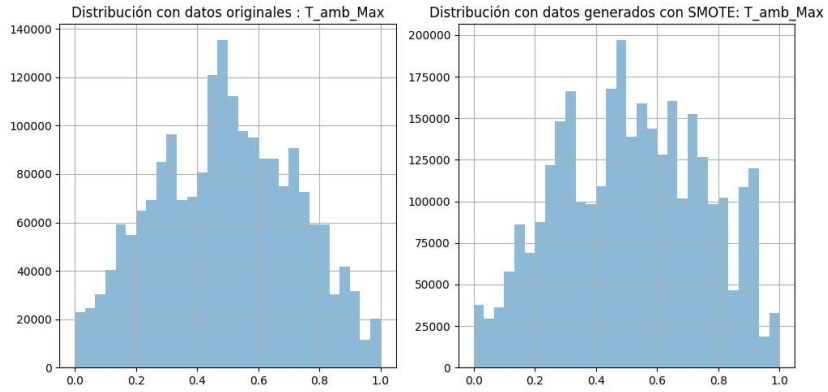



Figura 5.20: Comparación de T amb Max

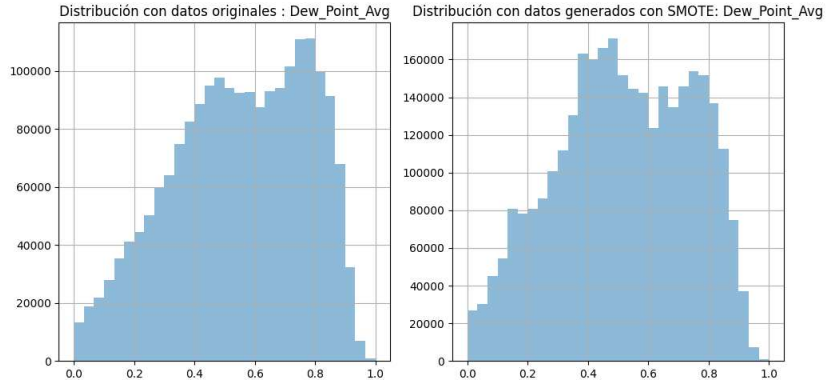



Figura 5.21: Comparación de Dew Point Avg

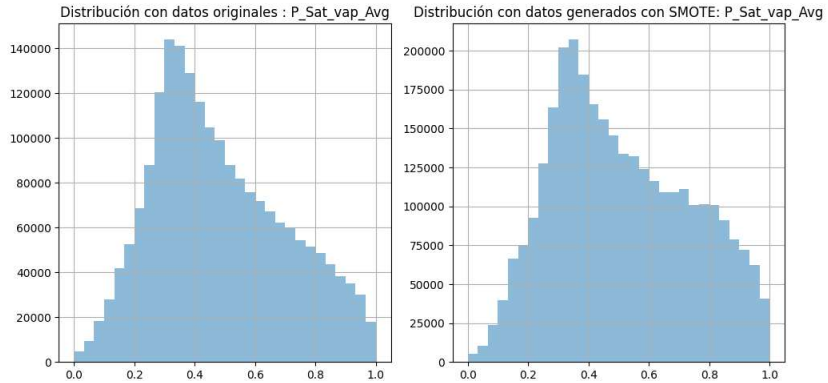



Figura 5.22: Comparación de P Sat vap Avg

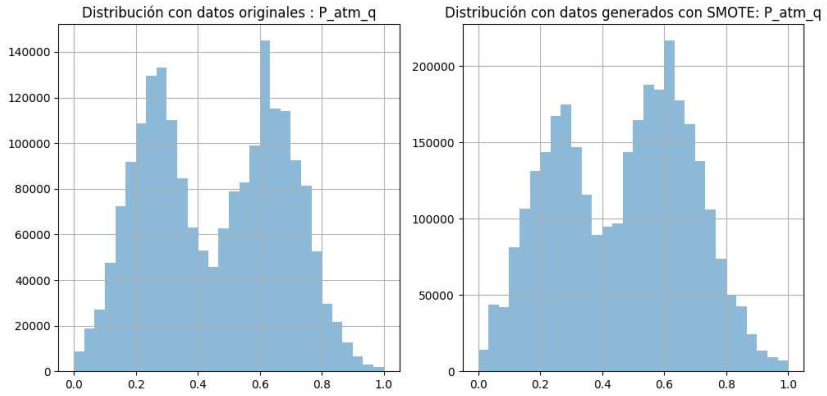



Figura 5.23: Comparación de P atm q

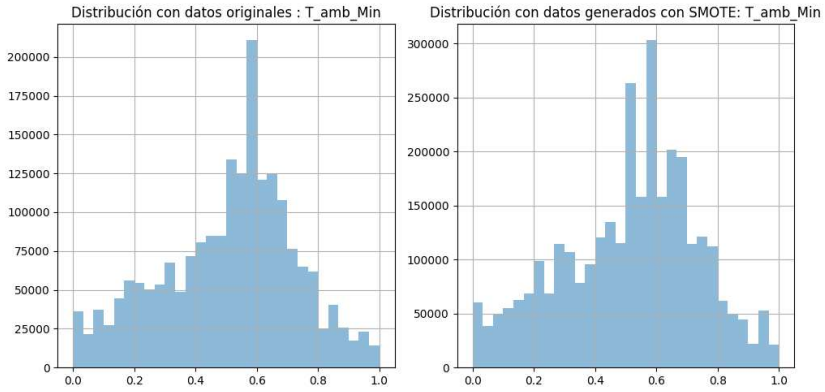



Figura 5.24: Comparación de T amb Min

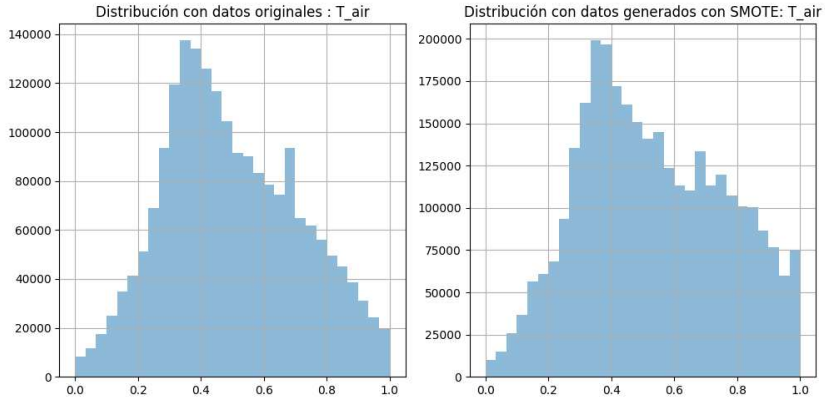



Figura 5.25: Comparación de T air

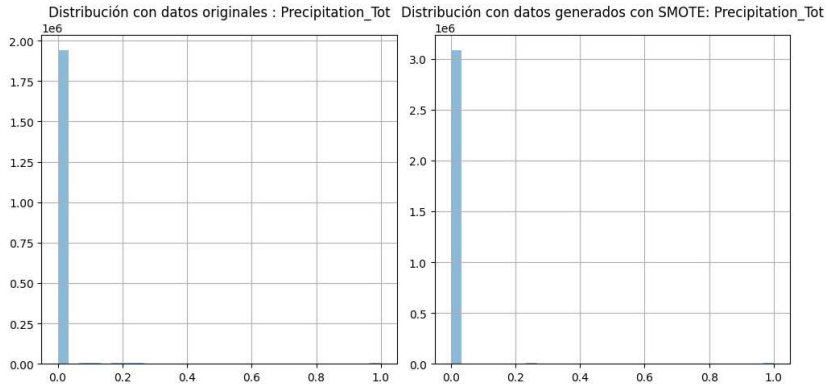



Figura 5.26: Comparación de Precipitation Tot

# Bibliografía


[1] G. Batista and M. Monard. An analysis of four missing data treatment methods for supervised learning. *Applied Artificial Intelligence*, 17(5-6):519–533, 2003.

[2] U. Brandes. A faster algorithm for betweenness centrality. *Journal of Mathematical Sociology*, 25(2):163–177, 2001.

[3] T. B. Bui, C. Drescher, A. Hanuschkin, K. Müller, S. Peters, S. Studer, and L. Winkler. Towards crisp-ml(q): A machine learning process model with quality assurance methodology. *Machine Learning and Knowledge Extraction*, 3(2):392–413, 2021.

[4] M. Bukov, A. G. Day, C. K. Fisher, P. Mehta, C. Richardson, D. J. Schwab, and C. H. Wang. A high-bias, low-variance introduction to machine learning for physicists. *Physics Reports*, 810:1–124, 2019.

[5] I. Bär, A. Kammer, A. Kniesz, L. Visengeriyeva, and M. Plöd. Crisp-ml(q). the ml lifecycle process. s.f.

[6] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer, and R. Wirth. Crisp-dm 1.0 step-by-step data mining guide. Technical report, The CRISP-DM Consortium, 2000.

[7] G. Chartrand, L. Lesniak, and P. Zhang. *Graphs & Digraphs*. CRC Press, Boca Raton, FL, 5th edition, 2015.

[8] N. Chawla, K. Bowyer, L. Hall, and W. Kegelmeyer. Smote: Synthetic minority over-sampling technique. *The Journal of Artificial Intelligence Research*, 16:321–357, 2002.

[9] K. Cheng, J. Li, H. Liu, F. Morstatter, J. Tang, R. P. Trevino, and S. Wang. Feature selection. *ACM Computing Surveys*, 50(6):1–45, 2017.

[10] I. S. Cohen, Y. Lu, Q. Tian, and X. Zhou. Feature selection using principal feature analysis. *Journal not specified*, 2007.

[11] Secretaría del Medio Ambiente de la Ciudad de México. Calidad del aire en la ciudad de méxico, informe 2018, 2020. Dirección General de Calidad del Aire, Dirección de Monitoreo de Calidad del Aire.

[12] Secretaría del Medio Ambiente de la Ciudad de México. Caracterización, evaluación y análisis del entorno físico y de la representatividad en las estaciones del sistema de monitoreo atmosférico de la ciudad de méxico, 2021. Dirección General de Calidad del Aire, Dirección de Monitoreo de Calidad del Aire.

[13] Secretaría del Medio Ambiente del Distrito Federal. Norma ambiental para el distrito federal nadf-009-aire-2006, 2006. Gaceta Oficial del Distrito Federal.

[14] Hamami F. and Fithriyah I. Classification of air pollution levels using artificial neural network, 2020.

[15] A. Fernandez, S. Garcia, F. Herrera, and N. V. Chawla. Smote for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary. *The Journal of Artificial Intelligence Research*, 61:863–905, 2018.

[16] P. Fernández. El secreto de google y el Álgebra lineal. ResearchGate, 2004.

[17] Gaceta Oficial de la Ciudad de México. Índice administración pública de la ciudad de méxico, 2018. Publicado el 14 de noviembre de 2018.

[18] Gaceta Oficial de la Ciudad de México. Índice administración pública de la ciudad de méxico, 2018. Publicado el 28 de mayo de 2019.

[19] R. Gupta, A. Priyam, A. Rathee, and S. Srivastava. Comparative analysis of decision tree classification algorithms. *INPRESSCO*, 2013.

[20] J. M. Jerez, I. Molina, P. J. García-Laencina, E. Alba, N. Ribelles, M. Martín, and L. Franco. Missing data imputation using statistical and machine learning methods in a real breast cancer problem. *Artificial Intelligence in Medicine*, 50(2):105–115, 2010.

[21] H. Junninen, H. Niska, K. Tuppurainen, J. Ruuskanen, and M. Kolehmainen. Methods for imputation of missing values in air quality data sets. *Atmospheric Environment*, 38(18):2895–2907, 2004.

[22] Leo Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43, 1953.

[23] Y. Luo and J. Zhang. Degree centrality, betweenness centrality, and closeness centrality in social network, 2017.

[24] Gore R. and Deshpande D. An approach for classification of health risks based on air quality levels, 2017.

[25] L. Rincón. *Introducción a la probabilidad*. Universidad Nacional Autónoma de México, Facultad de Ciencias, 2 edition, 2016.

[26] SEDEMA Secretaría del Medio Ambiente de la Ciudad de México. Calidad del aire en la ciudad de méxico, informe anual 2020, 2023. Dirección General de Calidad del Aire, Dirección de Monitoreo de Calidad del Aire.

[27] SEDEMA. Calidad del aire en la ciudad de méxico, informe 2017, 2018. Ciudad de México.

[28] SEDEMA. Pronóstico meteorológico para la cdmx, 2018. Recuperado el 30 de enero de 2023.

[29] S. Seo. *A review and comparison of methods for detecting outliers in univariate data sets*. PhD thesis, University of Pittsburgh, 2006.

[30] C. Shearer. The crisp-dm model: The new blueprint for data mining. *Journal of Data Warehousing*, 2000.

[31] Y. Simaan. Estimation risk in portfolio selection: the mean variance model versus the mean absolute deviation model. *Manage. Sci.*, 43:1437–1446, 1997.

[32] B. Sugiarto and R. Sustika. Data classification for air quality on wireless sensor network monitoring system using decision tree algorithm, 2016.

[33] A. Teologo, E. Dadios, R. Baldovino, R. Neyra, and I. Javel. Air quality index (aqi) classification using co and no2 pollutants: A fuzzy-based approach, 2018.

POSGRADO EN CIENCIAS NATURALES E INGENIERÍA
ESPECIALIZACIÓN MAESTRÍA DOCTORADO

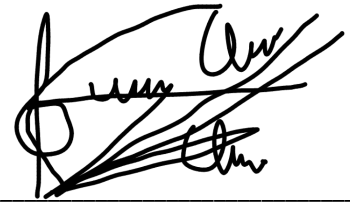
Esta idónea comunicación de resultados fue realizada dentro del Programa de Especialización del **Posgrado en Ciencias Naturales e Ingeniería** de la División de Ciencias Naturales e Ingeniería (DCNI) de la Universidad Autónoma Metropolitana-Unidad Cuajimalpa. El trabajo experimental y programa de cómputo fue realizado del octubre de 2022 al diciembre de 2024 en DMAS de la DCNI

DECLARACIÓN DE CESIÓN DE DERECHOS

En la Ciudad de México, D. F. el día 10 del mes diciembre del año 2024, el (la) que suscribe Juan Angel Acosta Ceja alumno (a) del Programa de Maestría en Ciencias Naturales e Ingeniería de la División de Ciencias Naturales e Ingeniería de la Universidad Autónoma Metropolitana-Unidad Cuajimalpa, manifiesta que es autor (a) intelectual de la presente idónea comunicación de resultados titulada; "Selección de Variables meteorológicas para la clasificación de los índices de contaminación" realizada bajo la dirección de Julián Alberto Fresán Figueroa y cede los derechos de este trabajo a la Universidad Autónoma Metropolitana (UAM) para su difusión con fines académicos y de investigación.

Los usuarios de la información no deben reproducir el contenido textual, gráfico o de datos del trabajo, sin el permiso expreso del (la) director (a) del trabajo como representante de la UAM. Este puede ser obtenido escribiendo a la siguiente dirección: (jfresan@cua.uam.mx)

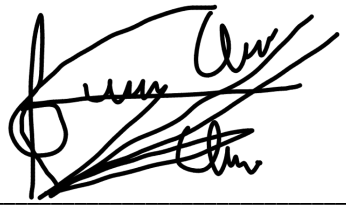
Si el permiso se otorga, el usuario deberá dar el agradecimiento correspondiente y citar la fuente del mismo.



Juan Angel Acosta Ceja

DECLARACIÓN DE ORIGINALIDAD

“El que suscribe, Juan Angel Acosta Ceja, alumno del Programa de Maestría en Ciencias Naturales e Ingeniería declara que los resultados reportados en esta idónea comunicación de resultados, son producto de mi trabajo con el apoyo permitido de terceros en cuanto a su concepción y análisis. Así mismo, declaro que hasta donde es de mi conocimiento no contiene material previamente publicado o escrito por otras (s) persona (s) excepto donde se reconoce como tal a través de citas y que este fue usado con propósitos exclusivos de ilustración o comparación. En este sentido, afirmo que cualquier información sin citar a un tercero es de mi propia autoría. Declaro, finalmente, que la redacción de este trabajo es producto de mi propia labor con la dirección y apoyo de mi director y de mi comité tutorial, en cuanto a la concepción del proyecto, al estilo de la presentación y a la expresión escrita.”



_____

Juan Angel Acosta Ceja

DECLARACIÓN DE NO LUCRO:

El que suscribe, Juan Angel Acosta Ceja, alumno del Programa de Maestría en Ciencias Naturales e Ingeniería, manifiesta su compromiso de no utilizar con fines de difusión, publicación, protección legal por cualquier medio, licenciamiento, venta, cesión de derechos parcial o total o de proporcionar ventajas comerciales o lucrativas a terceros, con respecto a los materiales, datos analíticos o información de toda índole, relacionada con las actividades e intercambios de información derivados de la relación de investigación académica y tecnológica desarrollada entre la Universidad Autónoma Metropolitana (UAM) y Juan Angel Acosta Ceja.

_____

Juan Angel Acosta Ceja