



**EDUCACIÓN**  
SECRETARÍA DE EDUCACIÓN PÚBLICA



TECNOLÓGICO  
NACIONAL DE MEXICO

**INSTITUTO TECNOLÓGICO DE CIUDAD MADERO**  
**DIVISIÓN DE ESTUDIOS DE POSGRADO E INVESTIGACIÓN**  
**MAESTRIA EN CIENCIAS DE LA INGENIERÍA**



**TESIS**

**VARIABLES DEL CAMBIO CLIMATICO EN MEXICO EN BASE A ANALISIS  
ESTADISTICOS Y MODELOS MATEMATICOS**

Que para obtener el grado de  
**Maestro en Ciencias de la Ingeniería**

Presenta:

**Ing. Lucero Lagunes Cuevas**

**G15070821**

**No. CVU: 1189157**

Director de tesis

**Dr. Juan Javier González Barbosa**

**No. CVU: 202134**

Co- director de Tesis

**Dr. Juan Frausto Solís**

Cd. Madero, Tamaulipas    abril 2024

Ciudad Madero, Tamaulipas, **06/marzo/2024**

**OFICIO No.:** U.028/2024

**ASUNTO:** Autorización de impresión de tesis

**C. LUCERO LAGUNES CUEVAS**  
**No. DE CONTROL G15070821**  
**P R E S E N T E**

Me es grato comunicarle que después de la revisión realizada por el Jurado designado para su Examen de Grado de Maestría en Ciencias de la Ingeniería, se acordó autorizar la impresión de su tesis titulada:

**“VARIABLES DEL CAMBIO CLIMÁTICO EN MÉXICO EN BASE A ANÁLISIS ESTADÍSTICOS Y MODELOS MATEMÁTICOS”**

El Jurado está integrado por los siguientes catedráticos:

PRESIDENTE:	DR.	JUAN JAVIER GONZÁLEZ BARBOSA
SECRETARIO:	DR.	JUAN FRAUSTO SOLÍS
VOCAL:	DR.	ULISES PÁRAMO GARCÍA
SUPLENTE:	DRA.	ANA LIDIA MARTÍNEZ SALAZAR
DIRECTOR DE TESIS:	DR.	JUAN JAVIER GONZÁLEZ BARBOSA
CO-DIRECTOR:	DR.	JUAN FRAUSTO SOLÍS

Es muy satisfactorio para la División de Estudios de Posgrado e Investigación compartir con usted el logro de esta meta. Espero que continúe con éxito su desarrollo profesional y dedique su experiencia e inteligencia en beneficio de México.

**ATENTAMENTE**  
*Excelencia en Educación Tecnológica®*  
*“Por mi patria y por mi bien”®*



**MARCO ANTONIO CORONEL GARCÍA**  
**JEFE DE LA DIVISIÓN DE ESTUDIOS DE**  
**POSGRADO E INVESTIGACIÓN**



c.c.p.- Archivo

MACG/NRV



Av. 1º de Mayo y Sor Juana I. de la Cruz S/N Col. Los Mangos C.P. 89440 Cd. Madero, Tam.  
Tel. 01 (833) 357 48 20, ext. 3110, e-mail: [depi\\_cdmadero@tecnm.mx](mailto:depi_cdmadero@tecnm.mx)  
[tecnm.mx](http://tecnm.mx) | [cdmadero.tecnm.mx](http://cdmadero.tecnm.mx)





## INSTITUTO TECNOLÓGICO DE CIUDAD MADERO

### DIVISIÓN DE ESTUDIOS DE POSGRADO INVESTIGACIÓN

Por medio de la presente se hace constar que la C. LUCERO LAGUNES CUEVAS, con número de control G15070821 Estudiante de la MAESTRÍA EN CIENCIAS DE LA INGENIERÍA no tiene adeudo en el departamento a mi cargo:

#### CENTRO DE CÓMPUTO DE LA DIVISIÓN DE ESTUDIOS DE POSGRADO E INVESTIGACIÓN

#### COORDINACIÓN DE LA MAESTRÍA EN CIENCIAS DE LA INGENIERÍA

FECHA: 06 de marzo de 2024

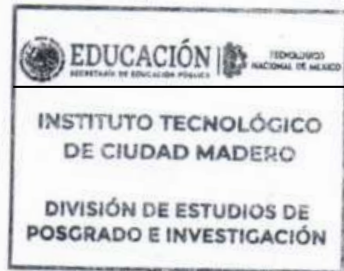
FECHA: 06 de marzo de 2024

FIRMA: \_\_\_\_\_

FIRMA: \_\_\_\_\_

SELLO: \_\_\_\_\_

SELLO: \_\_\_\_\_



#### JEFE DE LA DIVISIÓN DE ESTUDIOS DE POSGRADO E INVESTIGACIÓN

FECHA: 06 de marzo de 2024

FIRMA: \_\_\_\_\_

DR. MARCO ANTONIO CORONEL GARCÍA

SELLO: \_\_\_\_\_



**DECLARACIONES DE ORIGINALIDAD, PROPIEDAD  
INTELLECTUAL, CESION DE DERECHOS Y/O  
CONFIDENCIALIDAD**

Esta tesis consta de la originalidad, propiedad intelectual y cedo los derechos correspondientes, así como también la confidencialidad de los métodos, los resultados y todo lo generado en esta investigación.

## **AGRADECIMIENTOS**

Le quiero agradecer especialmente a mis asesores y maestros que han estado al pendiente del avance, a mi familia y pareja que estuvieron al pendiente de mí en todo el proceso y también a CONACyT por recibir la beca y por supuesto el Instituto Tecnológico de Ciudad Madero.

# **Variables del cambio climático en México en base análisis estadísticos y matemáticos.**

Lucero Lagunes Cuevas

## **Resumen**

El cambio climático es la modificación de la temperatura y del resto de variables del clima, que se está produciendo con una velocidad e intensidad que se ha visto afectado en la historia de la humanidad, como consecuencia de la actividad humana y de otros factores. Es por eso que las variables del cambio climático como temperatura, presión atmosférica, punto de rocío, monóxido de carbono y el óxido nítrico que son fundamentales para avanzar en la comprensión científica del clima aportando información climática necesaria para apoyar la toma de decisiones en muchos sectores. Por lo tanto en cada ciudad son diferentes las variables más relevantes ya que las condiciones climatológicas siempre serán diferentes dada la región de interés, la altura sobre el nivel del mar, las industrias cercanas etc.

El problema que se aborda es el obtener las variables relevantes del conjunto de variables de 4 ciudades que son Monterrey, Guadalajara, Tampico y Tijuana. Además, aparte de buscar las variables relevantes también se busca saber que ciudades les afecta unas más que otras, y cuáles de los métodos utilizados es el mejor para el estudio. Se utilizarán 4 métodos matemáticos y estadísticos: que son PCA, selección de atributos con correlación, selección de atributos por ganancia de información y basada en el aprendizaje. Con los cuales podremos hacer tal comparación utilizando regresión lineal múltiple en donde obtendremos como resultado que los métodos de ganancia de información y basada en aprendizaje son los mejores para este tipo de experimentación.

# **Variables del cambio climático en México en base análisis estadísticos y matemáticos.**

Lucero Lagunes Cuevas

## **Abstract**

Climate change is the modification of temperature and other climate variables, which is occurring with a speed and intensity that has been affected in human history, as a consequence of human activity and other factors. That is why climate change variables such as temperature, atmospheric pressure, dew point, carbon monoxide and nitric oxide are fundamental to advancing scientific understanding of climate, providing the climate information needed to support decision making in many sectors. Therefore, the most relevant variables are different in each city as the climatic conditions will always be different given the region of interest, the height above sea level, nearby industries etc..

The problem addressed is to obtain the relevant variables from the set of variables of 4 cities which are Monterrey, Guadalajara, Tampico and Tijuana. In addition, apart from looking for the relevant variables, the aim is also to find out which cities are affected more than others, and which of the methods used is the best for the study. Four mathematical and statistical methods will be used: PCA, attribute selection with correlation, attribute selection by information gain and based on learning. With which we will be able to make such a comparison using multiple linear regression where we will obtain as a result that the information gain and learning based methods are the best for this type of experimentation.

# Contenido

## Índice general

Resumen .....	V
Abstrac.....	VI
Índice de tablas.....	VIII
Índice de figuras .....	XI
<b>Capítulo 1 .....</b>	<b>1</b>
<b>1 Introducción .....</b>	<b>1</b>
1.1 Planteamiento del problema.....	2
<b>1.2 Objetivos .....</b>	<b>5</b>
<b>1.3 Justificación del estudio.....</b>	<b>6</b>
1.4 Organización de la tesis .....	7
1.5 Organizacion del documento.....	7
<b>Capítulo 2 .....</b>	<b>8</b>
<b>2 Antecedentes/Marco Teórico .....</b>	<b>8</b>
<b>Capítulo 3 .....</b>	<b>24</b>
<b>3 Metodología.....</b>	<b>24</b>
<b>Capítulo 4 .....</b>	<b>29</b>
<b>4 Resultados y conclusiones .....</b>	<b>29</b>
<b>5. Bibliografía .....</b>	<b>62</b>



## Índice Tablas

<b>Tabla 1</b> Nos muestra las contribuciones de las variables, el porcentaje representado de cada uno en los dos ejes representativos de Monterrey .....	33
<b>Tabla 2</b> De las variables más representativas y porcentaje con PCA de monterrey .....	33
<b>Tabla 3</b> Nos muestra las contribuciones de las variables, el porcentaje representado de cada uno en los dos ejes representativos de Guadalajara.....	35
<b>Tabla 4</b> De las variables más representativas y porcentaje con PCA de Guadalajara.....	35
<b>Tabla 5</b> Nos muestra las contribuciones de las variables, el porcentaje representado de cada uno en los dos ejes representativos de Tijuana.....	38
<b>Tabla 6</b> De las variables más representativas y porcentaje con PCA de Tijuana.....	38
<b>Tabla 7</b> Nos muestra las contribuciones de las variables, el porcentaje representado de cada uno en los dos ejes representativos de Tijuana.....	41
<b>Tabla 8</b> De las variables más representativas y porcentaje con PCA de Tampico.....	41
<b>Tabla 9</b> Resultados de correlación de Monterrey .....	42
<b>Tabla 10</b> Resultados de correlación de Tijuana .....	42
<b>Tabla 11</b> Resultados de correlación de Guadalajara.....	42
<b>Tabla 12</b> Resultados de correlación de Tijuana .....	43
<b>Tabla 13</b> Resultados de ganancia de información de Monterrey.....	43
<b>Tabla 14</b> Resultados de ganancia de información de Tijuana.....	44
<b>Tabla 15</b> Resultados de ganancia de información de Guadalajara.....	44
<b>Tabla 16</b> Resultados de ganancia de información de Tampico.....	44
<b>Tabla 17</b> Resultados de basado en el aprendizaje de Monterrey.....	45
<b>Tabla 18</b> Resultados de basado en el aprendizaje de Tijuana.....	45
<b>Tabla 19</b> Resultados de basado en el aprendizaje de Guadalajara.....	46
<b>Tabla 20</b> Resultados de basado en el aprendizaje de Tampico.....	46
<b>Tabla 21</b> Resultados del MAPE de Regresión Lineal múltiple de primera experimentación.....	47

<b>Tabla 22</b> nos muestra las contribuciones de las variables, el porcentaje representado de cada uno en los dos ejes representativos de Monterrey.....	50
<b>Tabla 23</b> de las variables más representativas y porcentaje con PCA de Monterrey.....	50
<b>Tabla 24</b> nos muestra las contribuciones de las variables, el porcentaje representado de cada uno en los dos ejes representativos de Guadalajara.....	52
<b>Tabla 25</b> de las variables más representativas y porcentaje con PCA de Guadalajara.....	53
<b>Tabla 26</b> nos muestra las contribuciones de las variables, el porcentaje representado de cada uno en los dos ejes representativos de Tijuana.....	55
<b>Tabla 27</b> de las variables más representativas y porcentaje con PCA de Tijuana.....	55
<b>Tabla 28</b> resultados de correlación de Monterrey.....	56
<b>Tabla 29</b> resultados de correlación de Guadalajara.....	56
<b>Tabla 30</b> resultados de correlación de Monterrey.....	56
<b>Tabla 31</b> resultados de ganancia de información de Monterrey.....	57
<b>Tabla 32</b> resultados de ganancia de información de Guadalajara.....	57
<b>Tabla 33</b> resultados de ganancia de información de Tijuana.....	58
<b>Tabla 34</b> resultados de basado en el aprendizaje de Tampico.....	58
<b>Tabla 35</b> resultados de basado en el aprendizaje de Guadalajara.....	59
<b>Tabla 36</b> resultados de basado en el aprendizaje de Tijuana.....	59
<b>Tabla 37</b> Resultados del MAPE de Regresión Lineal múltiple de segunda experimentación.....	60

# Índice de Figuras

<b>Figura 1.</b> . Promedio de las temperaturas de los 10 años de las ciudades Monterrey, Guadalajara, Tijuana y Tampico.....	4
<b>Figura 2.</b> Esquema de las interacciones a través de los componentes del sistema climático global..	14
<b>Figura 3</b> varianza.....	19
<b>Figura 4</b> formula de la covarianza.....	20
<b>Figura 5</b> Estructura de un árbol de decisión.....	21
<b>Figura 6</b> Es la metodología detallada de lo que se realizara en el proyecto.....	24
<b>Figura 7</b> Representacion de RLM.....	26
<b>Figura 8</b> De grafico de valores propios y variabilidad acumulada de PCA en Monterrey...	31
<b>Figura 9</b> muestra el mapa del círculo de correlaciones de PCA de Monterrey.....	32
<b>Figura 10</b> de grafico de valores propios y variabilidad acumulada de PCA en Guadalajara.	33
<b>Figura 11</b> muestra el mapa del círculo de correlaciones de PCA de Guadalajara.....	34
<b>Figura 12</b> de grafico de valores propios y variabilidad acumulada de PCA en Tijuana.....	36
<b>Figura 13</b> muestra el mapa del círculo de correlaciones de PCA de Tijuana.....	36
<b>Figura 14</b> de grafico de valores propios y variabilidad acumulada de PCA en Tampico....	39
<b>Figura 15</b> muestra el mapa del círculo de correlaciones de PCA de Tampico.....	40
<b>Figura 16</b> de grafico de valores propios y variabilidad acumulada de PCA en Monterrey...	48
<b>Figura 17</b> muestra el mapa del círculo de correlaciones de PCA de Monterrey.....	49
<b>Figura 18</b> de grafico de valores propios y variabilidad acumulada de PCA en Guadalajara.	51
<b>Figura 19</b> muestra el mapa del círculo de correlaciones de PCA de Guadalajara.....	51
<b>Figura 20</b> de grafico de valores propios y variabilidad acumulada de PCA en Tijuana.....	53
<b>Figura 21</b> muestra el mapa del círculo de correlaciones de PCA de Tijuana.....	54

---

## 1 Introducción

El cambio climático es uno de los problemas más importantes en la actualidad ya que los cambios no solo son notorios, si no que afectan el día a día desde el cambio de las estaciones que se han tenido los últimos años. Es importante tomar en cuenta que cada variable afecta de modo distinto cada ciudad por múltiples motivos como su geografía, desarrollo urbano etc.

Es por la razón anterior que se decide realizar este trabajo, el cual busca encontrar las variables del cambio climático que tienen mayor impacto en Monterrey, Guadalajara, Tijuana y Tampico. Tales ciudades se eligieron de forma aleatoria por la poca información encontrada hasta el momento. Esta investigación desarrollara conocimiento de cambio climático para estas ciudades.

Hoy en día de acuerdo a la información estudiada, no se ha encontrado en la bibliografía leída que exista alguna diferencia entre variables relevantes al comparar una ciudad grande y una ciudad pequeña al menos en revistas de alto impacto en donde se vea marcada dicha comparación.

Es por eso, por lo que el hecho de utilizar ciertas técnicas matemáticas como análisis de componentes principales y estadísticas como correlación, ganancia de información, basadas en el aprendizaje y regresión lineal múltiple entre otras de machine learning nos ayudarán para poder saber cuáles son las variables que más afectaron hasta la actualidad el cambio climático, en 4 ciudades importantes las cuales son monterrey, Guadalajara Tijuana y

Tampoco por lo siguiente saber que método matemático y estadístico es el mejor para realizar dichas evaluaciones.

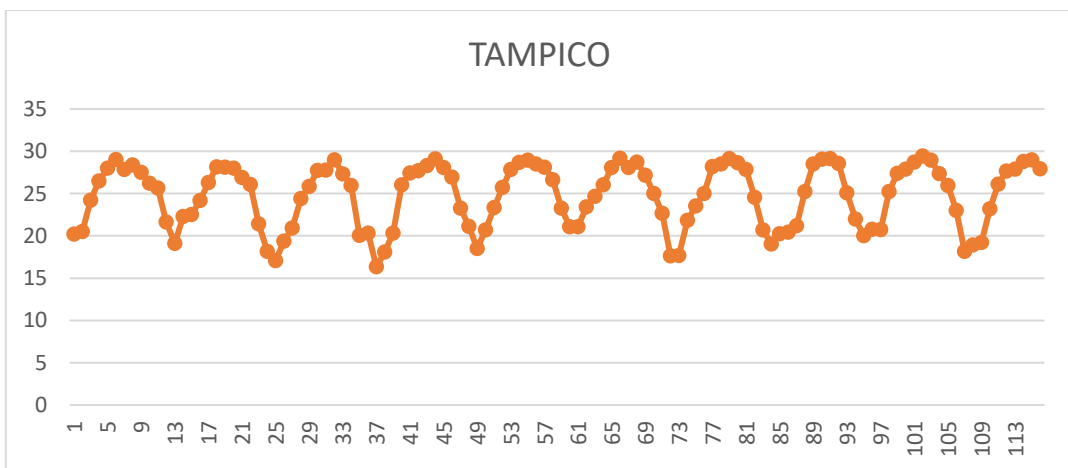
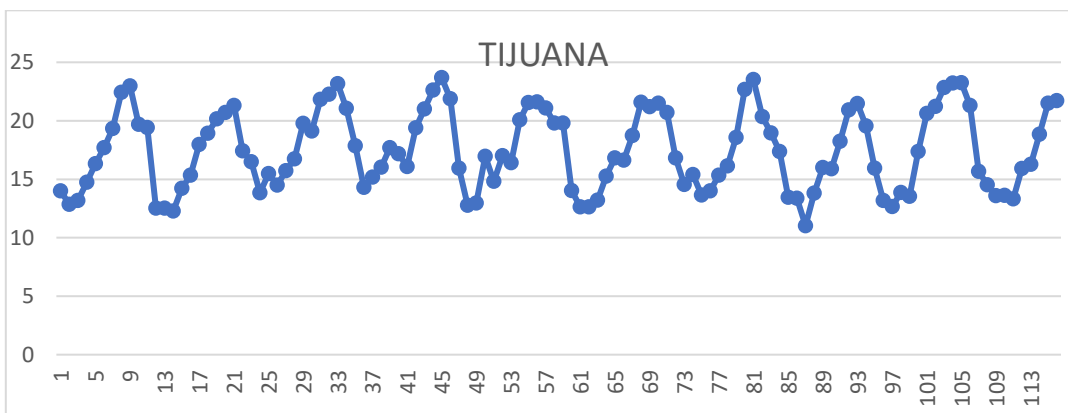
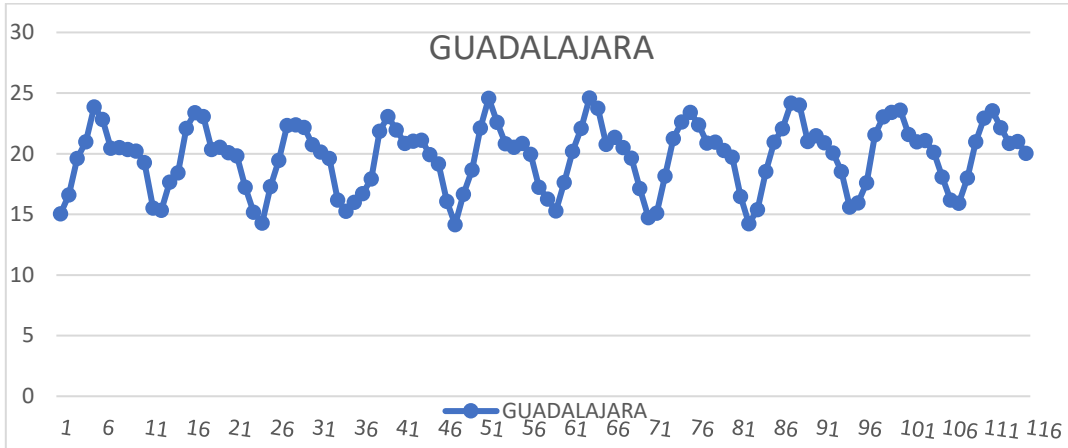
Y por lo tanto se decidió comenzar este proyecto con este fin y conocer cuáles han sido las variables con mayor impacto en las ciudades antes mencionadas y así poder aportar dicha información para programas o estudios futuros que puedan servir de ayuda para lograr en algún momento que se puedan realizar acciones que ayuden a que el cambio climático no avance exponencialmente.

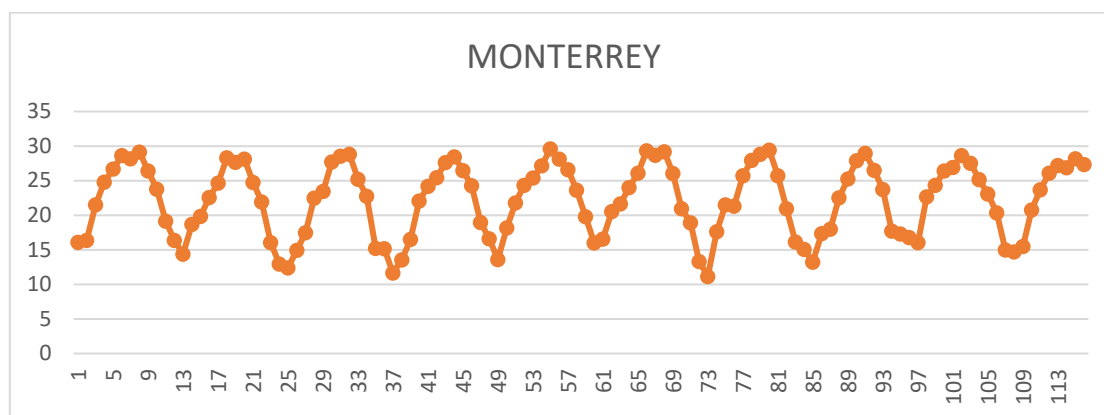
Para esto se analizan las técnicas y variables según la literatura que más estén presentes en algunos estudios previos, para a partir de esto, elegir programas que podamos utilizar de libre y fácil acceso para esta investigación. Después se continuará a analizar cuáles serían las mejores técnicas, para empezar con dicha evaluación y la comparación de estas. Tomando los resultados del estudio completo obtendremos la diferencia si es que la hay, de cada ciudad grande y pequeña, así mediante las métricas utilizadas saber cuáles métodos matemáticos o estadísticos fueron y serán los que nos den mejores resultados.

Y así poder demostrar que este método o métodos puede ser utilizado después para futuras investigaciones y que se sepa que tendrá menor margen de error que otros métodos.

### **1.1 Planteamiento del problema**

El cambio climático se ha considerado como de los problemas ambientales más importantes, y se define como la variación del sistema climático que puede darse por algunas causas naturales o por las actividades humanas. Por lo tanto, el calentamiento global se refiere al incremento promedio de las temperaturas terrestres, marina a nivel global y desde las últimas décadas la temperatura promedio de la tierra se ha vuelto más cálida e incluso supero los registros anteriores a 1850[1]. Lo anterior ha involucrado también el incremento de dióxido de carbono o CO<sub>2</sub> [2]. A pesar que el clima a cambiado mucho por formas naturales por decirlo de alguna manera existen claras evidencias de que el aumento de la temperatura es en gran parte por las actividades humanas [2]. Y este aumento lo podremos observar en las siguientes imágenes del comportamiento de la temperatura del 2012 al 2022 de las 4 ciudades ya mencionadas que se estudiaran a continuación:





**Figura 1.** Promedio de las temperaturas de los 10 años de las ciudades Monterrey, Guadalajara, Tijuana y Tampico

En la figura anterior demuestra el comportamiento de la temperatura promedio año con año de monterrey y Tampico rondando los 30 grados y en el caso de Tijuana y Guadalajara no a aumentado en promedio más de 25 grados. Por lo cual en algunos acuerdos como en el de parís nos da ciertas temperaturas a las cuales no debería aumentar para que nuestro problema sea irremediable. En este acuerdo se menciona que se reducen las emisiones de gases de efecto invernadero para limitar el aumento de la temperatura global en este siglo a 2 °C y tratar de parar el aumento a 1,5. También se revisan estos compromisos cada cinco años; Y ofrecen financiación a los países en desarrollo para que puedan mitigar el cambio climático; Y con esto, el proyecto al sumarse ayudara a estas ciudades en específico. [2]

Siguiendo con las condiciones de vida que el clima determina son las posibilidades de alimentación, la actividad económica en general y la seguridad de las poblaciones. En definitiva, determina cómo son los ecosistemas. Por lo tanto, al verse afectado influirá a todo el demás cambio.

Unas de las variables que afectan el cambio climático son las variaciones de la temperatura, presión atmosférica, la concentración de los gases de efecto invernadero, PM 5, 2.5 (materia particulada), las cuales han sido en gran parte por el incremento de población de los últimos años y entre otras variables, esto se explicarán más a detalle para tener mejor contexto de la situación.

Las variaciones de temperatura atmosférica es uno de los elementos que constituyen el clima. nos da la cantidad de energia calorifica la cual se encunetra en el lugar.

Proviene de los rayos que emite el sol. Los rayos son reflejados por la superficie terrestre y mandados otra vez al espacio. La atmosfera, los rayos solares vuelven a la tierra permitiendo que el calor se

quede más tiempo y así se provoca el efecto invernadero, algo de lo cual se escucha hablar todos los días. La atmósfera está formada por gases gracias a los que nuestro planeta tiene una temperatura media, uno de ellos es el CO<sub>2</sub>. Existen en gran cantidad en la atmósfera se vuelve más gruesa, haciendo que los rayos del sol estén durante más tiempo en la tierra. Al mantenerse ahí y seguir recibiendo rayos de sol, aumenta cada vez mayor de la temperatura de la tierra. Por esto, cuando más gases existan en la atmósfera, aumentará más la temperatura, ya que le impedirá al calor salir hacia el espacio [5]. Por lo tanto, debido a esto el realizar esta investigación de las variables nos dará algunas respuestas de cómo ha sido el comportamiento y afectación en los últimos años de las ciudades a estudiar.

Las variables de interés son llamadas variables dependientes. Y la determinación de estas no es una tarea simple, así como tampoco lo es el lograr cuantificar el nivel de impacto en cada una de las regiones de la República Mexicana. Y por lo tanto la variabilidad que existe en el clima se refiere a las fluctuaciones o variables observadas en el clima, alrededor de una condición promedio durante períodos de tiempo relativamente cortos. La variación observada en el clima durante periodos se llama cambio climático [7].

Para poder entender estas variaciones y obtener las variables más sobresalientes utilizaremos una técnica a utilizar es un análisis de componentes principales en el cual existen altas correlaciones entre las variables (esto indica que existe información redundante y por lo tanto, pocos factores explicaran una gran parte de la variabilidad total) [8].

Todo con el uso del machine learning y métodos estadísticos esto con el conjunto de todas estas técnicas poder realizar la evaluación de las variables que afectan el cambio climático que nos ayudaran a darle respuesta a nuestra situación planteada.

El problema que se busca abordar radica en determinar cuáles son las variables relevantes que afectan al cambio climático específicamente en las ciudades de Tampico, Tijuana y Monterrey, Guadalajara. Tomando en cuenta que en todos los lugares del mundo puede variar el impacto del cambio climático, no se cuenta con los mismos valores las variables dependientes en México y sus regiones. Por otra parte, estudios de este tipo se han realizado en otros lugares, hasta el momento de realizar este

documento, no se han encontrado estudios publicados en revistas de alto impacto para regiones de México. Y tampoco la comparación entre ciudades pequeñas y grandes, además los modelos estadísticos y matemáticos no son suficientemente claros para la evaluación y con los presentados se mejorará eso para poder tener respuestas más claras.



## 1.2 Objetivos

### Objetivo general

Determinar las variables relevantes que tienen mayor impacto de un conjunto de variables de las ciudades de Monterrey, Guadalajara, Tampico y Tijuana con 4 métodos matemáticos y estadísticos y evaluar estos métodos con Regresión Lineal Múltiple para obtener el que tenga menor error, y con esto saber cuál es mejor método de los 4.

### Objetivos específicos

- 1) Hacer una revisión bibliográfica sobre cambio climático para determinar cuáles son las variables con mayor mención y mayor cantidad de datos, utilizadas para medir el desempeño de dichos métodos, así como software libre utilizado para la implantación de dichos métodos y métricas.
- 2) Obtener los datos de cada variable de años de las regiones más pobladas o representativas de México. (Monterrey, Guadalajara, Tijuana y Tampico).
- 3) Evaluar y seleccionar herramientas computacionales de machine learning más adecuadas para elaborar los métodos matemáticos como PCA y estadísticos como correlación, ganancia de información y basado en aprendizaje.
- 4) Diseñar estrategias de prueba y obtener resultados con los métodos desarrollados y las métricas seleccionadas.
- 5) Evaluar qué métodos matemáticos y estadísticos con los años necesarios para poder obtener un buen resultado.
- 6) Realizar las evaluaciones y con Regresión lineal múltiple obtener cual es el mejor método.

### 1.3 Justificación del estudio

La información o evidencia disponible demuestra que el cambio climático es un fenómeno con incidencia global que afecta por lo tanto a México, donde al igual que en otros países existe un debate de las causas y consecuencias en unos lugares más que en otros se encuentra información, razón por la cual se hace este estudio hoy en día [11]. Por lo cual con esta tesis se espera aportar la mayor cantidad de información que nos ayude a tener respuesta de las variables relevantes de las ciudades y que tomen de margen o referencia lo publicado aquí para poder tomar acciones en conjunto.

Por otra parte, en México no se ha encontrado en la bibliografía leída tal información en alguna revista de alto impacto que nos mencione cuales han sido las variables relevantes que más han afectado al cambio climático en comparación de ciudades pequeñas y grandes, los métodos utilizados y la comparación entre ellos. Con esta información que obtendremos en el estudio nos ayudará a tener más información exacta de cuáles son las variables que más han afectado el cambio climático en la última década. En ciertas regiones del país donde a pesar de la gran influencia que hay en cuanto a gente y dimensión del lugar no hay estudio como tal en el que se vea reflejado como estas variables han ido cambiando, y que nos demuestren los cambios que hubo en últimos años.

Social y científicamente hablando, en este estudio se considera que una aportación así es importante siempre para las mediciones, predicciones de los científicos u organizaciones que se dediquen a plantear estrategias para el mejoramiento del planeta, por que como es sabido se está en una etapa muy crítica en la que pueden ocurrir las consecuencias del mal manejo de recursos que han afectado la flora, fauna, y todos nuestros ecosistemas.

Al momento de tener la información del comportamiento de las variables, se utilizarán estos datos estadísticos y matemáticos que ayudarán en un futuro a programas nacionales que puedan aportar económicamente al país para mejorar con distintas acciones, y programas públicos. Y tomar en cuenta que nuestro país no es el único afectado con estas situaciones ya mencionadas, pero de alguna manera al demostrar primero por nuestro país es un gran comienzo para que otras personas puedan o quieran tomar en cuenta esta información, para poder aplicarlas a sus países o regiones en las cuales se encuentren y vean que el conocer lo que ha pasado en estos lugares estudiados, nos puede ayudar para hacer programas en conjunto para el mejoramiento de estas ciudades.

#### **1.4 Organización del documento**

Esta tesis se está realizando en las instalaciones del Instituto Tecnológico de Ciudad madero desde el año 2022 hasta el 2023 que es donde se culminara con dicho proyecto.

En este documento tenemos una pequeña introducción en la cual damos una breve explicación de lo que tratará este trabajo, planteamos el problema por el cual decidimos empezar esta investigación, objetivo general y específicos partiendo de lo que se hará para ser concisos con la información dada, justificación, antecedentes de las técnicas utilizadas aquí o en otros países, la metodología que se utilizó ya que está dividida en dos partes, la primer que consta de 4 ciudades con 14 variables del cambio climático y la segunda parte que consta de 3 ciudades con 17 variables en esta parte las variables agregadas son las de la calidad del aire, después de esto se aplican las técnicas matemáticas y estadísticas y por ultimo obtenemos resultados.

---

## 2 Antecedentes y Marco teórico

### 2.1 Antecedentes

Con el paso de los años el cambio climático se ha visto un evidente cambio; en cambio climático de Mario molina aportaron unas variables, en un estudio de la tendencia de cambio climático en un área de las amazonas, donde se consideran como variables para este fenómeno las siguientes temperaturas máxima, temperatura mínima, evaporación y precipitación por mencionar algunas. Se recolectaron los valores de dichas variables en la estación meteorológica de Puyo. En dicha investigación, los autores cubrieron un conjunto de datos de más de treinta años, en particular del año 1974 hasta 2005 y cubriéndose 380 meses. Se pudo observar que todas las variables consideradas fueren influyentes en el cambio climático, tuvieron un incremento importante desde 1974 hasta 1985. Posteriormente se observó un incremento más suave [12]. Tomando en cuenta en esta ocasión que las variables a considerar son las más aptas para este tipo de estudio.

Otro de los estudios internacionales fue realizado en un área ecológicamente vulnerable en China. En Huang Guanwen 2021 se informa que es muy sensible al aumento de la temperatura global y cambios ambientales. Toma en consideración para el estudio de cambio climático tres factores; temperatura, precipitación y vapor de agua. Dichas variables son estudiadas con modelos de series temporales y con características de la distribución espacial de esos tres factores. Nótese que a pesar de que solo son tres factores en este caso, el número de variables involucradas en la modelación puede ser bastante más grande ya que se pueden tener variables asociadas tamaño de la ventana de muestreo de cada serie temporal y medidas estadísticas asociadas. Los resultados del estudio muestran que en los últimos años la tendencia de aumento de la temperatura en campo de dunas de china ha sido superior al nivel medio en China. Pero esta tendencia ha disminuido gradualmente desde el año 2000.

En el que se utilizó un análisis de componentes principales por vectores propios, función ortogonal empírica o simplemente PCA (de sus siglas en inglés, Principal Component Analysis). Este método es muy eficaz para analizar las características de las variables relevantes, mediante la modelación en una estructura matricial y convertir múltiples variables en pocas componentes principales; esto es posible por medio de la búsqueda de dimensionalidad más pequeña que modela completamente el sistema a estudiar. Lorenz aplicó por primera vez este método a la Meteorología y el clima en 1950 y ahora se ha convertido en un importante método de análisis en el tiempo de la Meteorología y clima esto es aplicable en el campo de los puntos espaciales que cambiaron en el tiempo. [13]. En este método se realiza un procesamiento de datos y se requiere una prueba estadística o de significancia para determinar los componentes principales de mayor impacto.

También se ha podido observar en Margarita Caballero en 2021 algunas de las variables relevantes para un estudio de cambio climático para la República Mexicana, contemplaron aspectos asociados valores del clima de la atmosfera y al efecto invernaderos, así como la determinaron los valores de clima de la atmósfera, la radiación solar y el efecto invernadero natural, y también de que, manera estaban distribuidos los climas en México y las causas que lo ocasionaban [14].

Un ejemplo de esto es por Lidija Tadic en su artículo de la aplicación de PCA en la evaluación de cambio climático en sus variables esenciales de temperatura del aire y la precipitación. Utilizando 10 parámetros de dos estaciones de Osijek y Dubraunik de 1985 al 2016 incluyendo precipitación, temperatura media, desviación estándar de la temperatura, temperatura máxima, temperatura máxima, precipitación máxima diaria, número de días con precipitación, días sin precipitación, índice de precipitación estándar del mes y la aridez. Primeramente, se aplicó sobre el conjunto de coeficientes de regresión lineal definidos para los indicadores de cambio climático, continuando con aplicar PCA sobre el estadístico de prueba calculado por Man Kendall para determinar la existencia de tendencias temporales significativas en los valores del indicador. Y demuestra que PCA es una herramienta muy útil para implementar este enfoque, al igual en Europa se estudiaron las estaciones para ver las tendencias respecto al lugar geográfico y donde la precipitación o temperatura estaba en aumento y también se aplicó regresión lineal sobre estos datos. En la estación de Osijek muestra tendencias crecientes en la temperatura del aire, y luego define sus dos componentes esenciales para la descripción de la contribución y el impacto de cada variable. En conclusión, Osijek se dio cuenta que están bajo más influencia de precipitaciones y cambios extremos de temperatura lo cual provoca más sequías e inundaciones [15].

Para el estudio que buscamos abordar hablando de la calidad del aire que es algo que evidentemente afecta, en este caso la materia particulada lo representa el libro “ The Global Health Costo f PM 2.5 Air pollution” que habla acerca del costo del daño a la salud por partículas denominadas PM 2.5, constatándose los riesgos relativos resultantes asociados por la presencia de dichas partículas en el entorno habitable[16].

En otro artículo de que se habla acerca de evaluación de la climatología espacio-temporal se mostraron tendencias del ozono en la región india utilizando aprendizaje automático; este artículo presenta intentos de estudio para detectar anomalías y evaluar tendencias climatológicas estadísticas del ozono columna total registrado sobre los sitios de la India. El estudio considera de 30 a 54 años de datos de grabado por el espectrofotómetro Dobson obtenido del se utiliza el Departamento Meteorológico de la India. [17].

Otra investigación utiliza componentes principales a partir de una base de datos de calidad del aire en el cual utiliza el análisis de matriz de correlaciones en México y evalúa los gases de efecto invernadero, PM 2.5 y PM 10 [19].

Por otra parte otra de las técnicas de los vectores propios con Charles K Stidd 1966, los vectores propios se pueden usar de manera que los polinomios ortogonales sin embargo tienen una ventaja sobre las funciones ortogonales expresables mediante fórmulas simples, ya que se derivan de los datos que estudian y se asemejan mucho a las características importantes de los datos de modo que los primeros vectores propios contienen un porcentaje lo más alto de variancia y estarían con contenido en un número igual de polinomios ortogonales ordinarios. En este artículo describe el proceso mediante el cual se derivó una matriz de vectores propios a partir de un conjunto de 2 valores de precipitación media mensual para 60 estaciones en nueva el efecto de la elevación de la estación en cada vector propio se determina mediante correlación lineal y los multiplicadores de estación corregidos con una elevación media y se trazan y analizan en 3 mapas [20].

Agregando más explicación de las técnicas elaboradas anteriormente En este estudio de Rubaina Kabirse desarrolla un nuevo marco estadístico basado en matrices para la asignación de pesos que pueden considerarse como el método más simple y preciso para asignar pesos a una gran cantidad de indicadores el método MSF se basa en la valoración de matriz de correlación y vectores propios asociados al valor propio basándose en la metodología Inter construcción MSF puede llenar algunas brechas intrigadas entre otros métodos de ponderación también puede ser directamente la decisión de seleccionar los peso relativos que se pueden encontrar en el vector propio correspondiente al valor

propio más grande el nuevo método se aplica asignando pesos a 15 indicadores socioeconómicos y evaluando la vulnerabilidad y el riesgo en la costa de Bangladesh. Comparar con otros métodos de ponderación se encuentra que MSF brinda la explicación física más aceptable sobre los valores relativos de la ponderación de las generaciones de los indicadores [21].

El artículo de Christine Drebeton el cual habla de los componentes principales y el objetivo de proporcionar una gama más amplia de situaciones factibles para ser utilizadas en situación de riesgo. En resumen, se utiliza para disminuir el número de variables y generar variables independientes que puedan generarse por separado [22].

En Stella que es un software en el cual se evaluaron los posibles cambios en precipitación y temperatura para 2010-2099 y también los cambios en la escorrentía y evotranspiración y carga de los mantos acuíferos [23].

Se han realizado diferentes trabajos con respecto a las variables relevantes en el cambio climático con algunos métodos estadísticos o aplicaciones como Weka algunos de los cuales son los siguientes:

Amit Kumar Yadav y Hasmat Mlik 2013 hablan de selección de las variables de entrada más relevantes mediante Weka para modelos de predicción de la radiación basados en redes neuronales artificiales en India. Las variables utilizadas son la latitud, longitud y temperatura se utilizan 3 modelos neuronales artificiales en los cuales la temperatura, la temperatura máxima, la temperatura mínima, la altitud, la latitud, la longitud y el clima son las más relevantes. Y la latitud y la longitud como las variables menos influyentes en la predicción de la radiación solar. [24]

Lidija Tadic en 2019 se utilizó PCA en la evaluación de cambio climático en sus variables esenciales de temperatura del aire y precipitación. Utilizando 10 parámetros de dos estaciones meteorológicas de Osijek y Dubraunik de 1985 al 2016 incluyendo precipitación, temperatura media, desviación estándar de la temperatura, temperatura máxima, temperatura mínima, precipitación máxima diaria, número de días con precipitación, días sin precipitación, índice de precipitación estándar del mes y la aridez.

Y como resultados que la estación meteorológica Osijek muestra tendencias crecientes en la temperatura del aire, y luego define sus dos componentes esenciales para la descripción de la contribución y el impacto de cada variable. Y en conclusión, la ciudad de Osijek se dio cuenta que están bajo más influencia de precipitaciones y cambios extremos de temperatura lo cual provoca más sequías e inundaciones [25].

Huan Guanwen y Zhu Hai hizo en 2021 un estudio internacional que fue realizado en un área ecológicamente vulnerable en China, el cual es muy sensible al aumento de la temperatura global y cambios ambientales. Se toma en consideración para el estudio de cambio climático tres variables; temperatura, precipitación y vapor de agua. Dichas variables son estudiadas con modelos de series temporales y con características de la distribución espacial de esos tres factores.

En dicho estudio se utilizó un análisis de componentes principales por vectores propios, o PCA (de sus siglas en inglés, Principal Component Analysis).

Dando como resultado de esto la tendencia es que el aumento de la temperatura en el campo de dunas de Mu Us ha sido superior al nivel medio de China aunque esta tendencia a disminuido en los últimos años. La distribución de la temperatura tiene una evidente característica de aumento gradual de norte a sur y se ve afectada por la latitud, altitud y la topografía. [26].

P Venkata Ramana Moorthy, B Sarojamma y S Venkatramana Reddy en 2022 En la investigación que realizaron nos dan las siguientes variables atmosféricas, como precipitaciones, temperatura, velocidad del viento, humedad, visibilidad, ráfagas de viento, precipitaciones, etc. Para aprender la covarianza de las máquinas por principios, se realizan enfoques prácticos y probabilísticos utilizando el proceso gaussiano. En este trabajo, tomando la visibilidad, la hora con fecha, la temperatura como variables independientes o de respuesta y la velocidad del viento como variable dependiente o de respuesta, ajustamos un modelo de proceso gaussiano. K star es un clasificador basado en instancias que clasifica los datos. RBF se utiliza para los datos y tiene una estructura similar a la del proceso gaussiano, pero utiliza un método de agrupación con parámetros de peso. La regresión aditiva clasifica las variables utilizando Decision Stump. La regresión de árbol de decisión mejora el modelo eliminando las decisiones del árbol que no son importantes en la clasificación. Se ajustan diferentes modelos en (WEKA) para los datos atmosféricos y determina cuál es el mejor modelo basándonos en los valores de error cuadrático medio. Y en el cual después de utilizar todos los modelos el modelo de estrella K es el mejor modelo para las variables atmosféricas para enfoques prácticos y probabilísticos [27]



Todos los estudios y artículos ya mencionados son una recopilación acerca de los estudios relacionados de este proyecto ya que es importante conocer con anterioridad que se ha estudiado y que no. Por lo cual lo que se estudiará en este proyecto sea gran aporte para los estudios en general de México, ya que los estudios más relacionados a esto se encuentran en otros países y en nuestro país es importante recalcarlo ya que vivimos en él y por lo tanto será de gran ayuda para los futuros estudios o planes de acción para reducir el cambio climático.

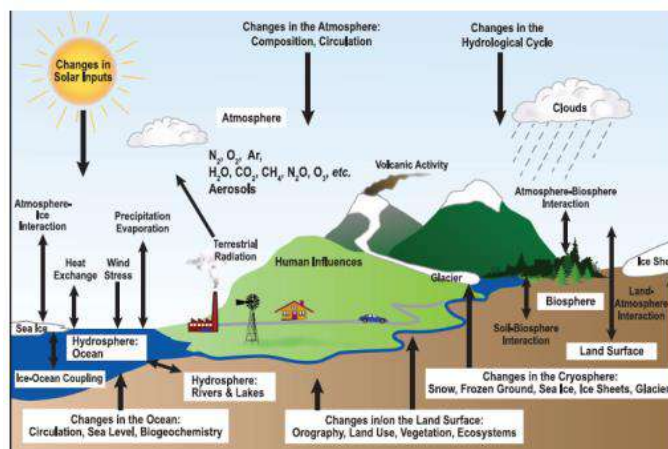
### **2.3 Marco teórico**

México es un país que tiene efectos del cambio climático como cualquier otro país en el mundo, quizás algunos más por diferentes situaciones del ser humano y desarrollos urbanos. Unos de las variables que afectan a esto se mide con el índice de calidad del aire (ICA) es un indicador de las autoridades pertinentes de cada lugar o zona que nos demuestra las variables cantidades de contaminantes que están en el aire. Tomamos en cuenta que la calidad del aire es buena cuando las personas respiran indefinidamente y que su salud nunca se vea afectada por esto. Y cuando es mala suele ser dañina y provocar desde irritación de ojos hasta problemas pulmonares y del corazón. Unas de las situaciones que provoca el cambio climático en México son las inundaciones y la desertificación que afecta un aproximado de 40% de la superficie nacional cultivable. Uno de los estudios que han sido elaborados nos señalan algunas poblaciones que viven cerca del mar, ya que son propensas al aumento de este cuyos efectos hemos tenido evidencia y muestra estos últimos años. [24]. Pero algunas otras sustancias presentes en el aire, hoy en día también son dañinas. Por lo tanto, es importante conocer algunos de los conceptos que involucran y técnicas que podrán ser utilizadas para cuantificar este fenómeno. Comenzamos con los conceptos básicos a continuación

## Clima

Es el resultado de la interacción de la atmósfera, océanos, capas de nieve, continentes. Y se puede explicar como el “tiempo promedio” y las condiciones meteorológicas que lo componen.

Factores claves del clima son el sol que es la fuente de energía, y la Atmósfera es el factor que mantiene el calor recibido por el sol y a su vez a la función de los océanos hace posible la circulación de la humedad y el calor. Los cuales están representados en la figura 1: [28].



**Figura 2.** Esquema de las interacciones a través de los componentes del sistema climático global

El cambio climático se define como un cambio durable en la distribución de los elementos de clima en grandes periodos de tiempo que puede llegar a millones de años. Esto se refiere a que al cambio climático causado por la actividad humana y los causados por procesos naturales de la Tierra y el sistema solar. El clima nos da el estado atmosférico que son desde el viento, precipitación, temperatura media entre otros. En un determinado de tiempo que normalmente se valora en 30 años. La variación estadística promedio en este tiempo se le conoce como cambio climático, tal cambio suele ser influido por los factores antropogénicos y naturales como la quema de fósiles y la atmósfera se ve afectada por esto o incluso en algunos cambios del suelo. [29]. De las variables y factores que influyen a esto es lo siguiente:

### Gases de efecto invernadero

Estos gases están presentes en la atmósfera y entre alguna de sus propiedades es retener la energía momentáneamente que la superficie terrestre, océanos y los hielos devuelven a la atmósfera después de ser calentados por el sol, haciendo con esto posible la vida en el planeta.

Algunos de los gases naturales como el dióxido de carbono ( $\text{CO}_2$ ) que es la quema de combustibles fósiles, algunas reacciones químicas en algunos procesos de acero, por ejemplo. Fuentes del uso de la tierra es en la deforestación y quema de bosque, el metano ( $\text{CH}_4$ ): Fuente industrial es en ellos rellenos sanitarios, minería del carbón, producción de gas natural. Fuente del uso en la tierra es la conversión de humedales arrozales, producción ganadera y óxido Nitroso ( $\text{N}_2\text{O}$ ): En las fuentes industriales es la producción de plásticos y tejidos como el nailon, también es emitido en quema de combustibles fósiles. En el uso de la tierra es el uso de fertilizantes y quema de biomasa.

Gases antropogénicos que son por ejemplo hidrofluorocarbonos (HfCs): Sus usos industriales son en refrigeradores, aerosoles de espuma y solventes, perfluorocarbono (PFCs): En usos industriales es en la producción de aluminio, fabricación de semiconductores, sustituto de las sustancias destructoras del ozono y hexafluorocarbonos (PFCs): Las fuentes industriales son en transmisión eléctrica y sistemas de transmisión [29].

La vida en la tierra tiene influencia o depende de algunos factores de una capa mayormente compuesta de gas: que es la atmosfera. En la cual está compuesta por unos gases como lo es el nitrógeno con 78%, oxígeno y pequeñas cantidades de argón 0.9%. El pequeño porcentaje restante está compuesto por vapor de agua, ozono, bióxido de carbono, hidrogeno, neón, helio y Kriptón. Estos últimos gases se encuentran dispersos en distintas cantidades, en las 5 capas en la que los científicos han dividido .

la atmosfera para poder estudiarla. Y unas de las cosas o funciones de la atmosfera se puede mencionar que su control de filtro de radiaciones ultravioleta puede llegar hasta la superficie terrestre, su acción que protege de una cantidad de meteoritos al destruirlos ya que de una o de otra manera llegarían a la superficie terrestre además ayuda a regular la temperatura.[36].

### **Presión atmosférica**

Al ser un elemento termodinámico del clima, nos establece la presión que está ejerciendo el aire sobre toda la materia dentro de la atmosfera y esta puede cambiar según la latitud o altitud del nivel del mar.

Algunos de los factores que influyen en el constante cambio de la presión atmosférica son por ejemplo la altura sobre el nivel del mar, ya que conforme va aumentando o ganando altura en la atmosfera va bajando la densidad del aire por lo que menor masa ocupa el mismo espacio y siguiente tal cantidad de masa está haciendo menor presión. Si nos encontramos a nivel del mar estamos soportando una mayor cantidad de aire.

Temperatura del aire.- la temperatura tiene la capacidad de que puede cambiar algunas propiedades de algunas sustancias, por ejemplo en el aire atmosférico si se aumenta la temperatura se expande y disminuye la densidad así que si el aire no se encuentra confinado en un contenedor disminuye la presión que ejerce.

Distribución del aire en la Tierra.- en el mundo existe el viento el cual tiene unas características físicas como una velocidad determinada y una temperatura específica y determinada, en tales corrientes al encontrarse en movimiento pueden variar la presión de una zona determinada al pasar por ella.

Zonas de presión Zona de alta presión. – Es llamado anticiclones a los fenómenos atmosféricos donde el viento circula en sentido a las manecillas del reloj en el hemisferio norte.

Zona de baja presión.- se le llama ciclón y se origina en la zona de la atmosfera donde la presión es más baja que la presión de los lados de alrededor, entonces sucede que cuando el aire se calienta y asciende y este lugar es ocupado por el aire frío de la troposfera superior, los vientos giran en sentido contrario las manecillas del reloj en el hemisferio norte y cuando desciende el aire converge en el centro absorbiendo la temperatura de la zona formando nubes hasta saturarse, y así se forman las precipitaciones [37].

Algunas de las variables son el material particulado que es una fracción respirable está dividido en dos grupos principales de acuerdo a su tamaño PM10 y PM2.5. Las partículas menores a 10  $\mu\text{m}$  (PM10) son partículas sólidas o líquidas dispersas en la atmósfera, en su mayoría de pH básico, producto de la combustión no controlada. El segundo tamaño PM2.5, agrupa a partículas generalmente ácidas, que contienen hollín y otros derivados de las emisiones vehiculares e industriales, y corresponde a la fracción más pequeña y agresiva debido a que éstas son respirables en un 100% y tienen mayor efecto en la salud, ya que pueden ingresar al sistema respiratorio depositándose en los alveolos pulmonares y llegar al torrente sanguíneo produciendo efectos negativos en la misma, tales como aumento en las enfermedades respiratorias y cardiacas, asma, bronquitis, enfisema, cáncer pulmonar, y disminución del funcionamiento pulmonar.-Las partículas PM2.5 están compuestas de elementos que son tóxicos como metales pesados y compuestos orgánicos productores de cáncer. [37].

Algunos de los factores que afectan el cambio climático es el aumento de población que ha traído una mayor demanda de recursos naturales, lo que presiona las reservas en la naturaleza. Por ejemplo cubrir el requerimiento de alimentos tanto para uso humano como animal que se utiliza el 11% de la superficie terrestre y el 70% del agua total extraída de los acuíferos, ríos y lagos. [39]. Por lo tanto, el conocer la tendencia de la población es importante.

### **Población mundial**

El crecimiento acelerado de la población mundial es un fenómeno relativamente reciente. Según algunas estimaciones hasta el año 1000, la población no había rebasado los 300 millones de personas, pero sólo 500 años más tarde, el número de habitantes ya había crecido, según diversas estimaciones, a entre 424 y 484 millones. En 1750, esta cantidad ya se había incrementado en poco más de 200 millones llegando a cerca de 700 millones de personas. Para inicios del siglo XX esta población se había incrementado poco más de dos veces llegando a 1 550 millones de habitantes. Las proyecciones del Consejo Nacional de Población (Conapo) señalan que el crecimiento poblacional seguirá hasta el año 2050, cuando alcanzará aproximadamente 150.84 millones de habitantes [40].

Es decir que el cambio climático y sus variables en la actualidad a sido importante reconocerlas para saber a qué nos enfrentamos y como ha sido con el paso de los años su evolución para tener un amplio panorama de lo que se espera.

Técnicas que nos ayudaran con los análisis estadísticos y modelos matemáticos. Como lo son:

### **Análisis de componentes principales**

El análisis de componentes principales (PCA) es una técnica multivariante que analiza una tabla de datos en la que las observaciones se describen mediante varias variables dependientes cuantitativas interrelacionadas. Su objetivo es extraer la información importante de los datos estadísticos para representarla como un conjunto de nuevas variables ortogonales llamadas componentes principales, y mostrar el patrón de similitud entre las observaciones y de las variables como puntos en mapas puntuales. Matemáticamente, el PCA depende de la descomposición propia de las matrices semidefinidas positivas y de la descomposición del valor singular (SVD) de las matrices rectangulares. Está determinado por los vectores propios y los valores propios. Los vectores propios y los valores propios son números y vectores asociados a las matrices cuadradas. Juntos proporcionan la descomposición propia de una matriz, que analiza la estructura de esta matriz, como las matrices de correlación, covarianza o producto cruzado.

Objetivos de la PCA son:

1. Extraer la información más importante de la tabla de datos
2. Comprimir el tamaño del conjunto de datos manteniendo sólo esta información importante
3. Simplificar la descripción del conjunto de datos; y
4. Analizar la estructura de las observaciones y las variables.
5. Comprimir los datos, reduciendo el número de dimensiones, sin mucha pérdida de información.

Para analizar los datos mediante el Análisis de Componentes Principales tenemos que ser minuciosos en estadística y álgebra matricial. Por lo tanto, hablaremos de la estadística, que examina las medidas de distribución, cómo se distribuyen los datos, y también del álgebra matricial, calculando los vectores propios y los valores propios, que es el principio fundamental para determinar el PCA.

### Estadística

La estadística ha sido definida de forma diferente por diferentes autores de vez en cuando. Algunos autores la definen como "datos estadísticos", es decir, enunciados numéricos de hechos, mientras que otros la definen como "métodos estadísticos", es decir, cuerpo completo de los principios y técnicas utilizados en la recogida y el análisis de los datos.

### Desviación estándar

La desviación típica de un conjunto de observaciones de una serie es la raíz cuadrada positiva de la media aritmética de los cuadrados de todas las desviaciones de la media aritmética. Así, en el cálculo de la desviación típica, primero se calcula la media aritmética y se eleva al cuadrado la desviación de varios elementos respecto a la media aritmética. Las desviaciones al cuadrado se suman y la suma se divide por el número de elementos. El proceso de omitir los signos algebraicos + y - de las desviaciones en la desviación media se evita en la desviación estándar. Por lo tanto, la desviación estándar es una medida de dispersión de mayor importancia matemática. La desviación estándar se denota generalmente por  $\sigma$ .

### Varianza

La varianza es otra medida de la dispersión de los datos en un conjunto de datos. Se define como la media de la desviación de cada término respecto a su media aritmética de los datos completos. De hecho, es casi idéntica a la desviación estándar. La fórmula es la siguiente

$$Var(X) = \frac{\sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})}{(n - 1)}$$

**Figura 3** varianza

$\sigma^2$  = Varianza,

$\bar{X}$  = Media aritmética

$X - \bar{X}$  = Desviación de la observación individual de la media aritmética,

$n$  = Número de observaciones.

$\sigma^2$  y la fórmula (no hay raíz cuadrada en la fórmula de la varianza).  $\sigma^2$  es el símbolo habitual de la varianza de una muestra. Ambas medidas son medidas de la dispersión de los datos. Desviación estándar es la medida más común, pero también se utiliza la varianza. La razón por la que he introducido la varianza además de la desviación estándar es para proporcionar una plataforma sólida a partir de la cual se pueda definir la covarianza.

### Covarianza

Muchos conjuntos de datos tienen más de una dimensión, y el objetivo del análisis estadístico de estos conjuntos de datos suele ser ver si existe alguna relación entre las dimensiones. La desviación estándar y la varianza sólo operan en una dimensión, por lo que sólo podríamos calcular la desviación estándar de cada dimensión del conjunto de datos independientemente de las demás dimensiones. Sin embargo, es útil disponer de una medida similar para saber cuánto varían las dimensiones con respecto a la media entre sí. La covarianza es una medida de este tipo. La covarianza se mide siempre entre dos dimensiones. Si calculamos la covarianza entre una dimensión y ella misma, obtendremos la varianza. Así, si tuviéramos un conjunto de datos de 3 dimensiones (A, B, C), podríamos medir la covarianza entre las dimensiones A y B, las dimensiones B y C, y las dimensiones A y C. La medición de la covarianza entre A y A, o B y B, o C y C nos daría la varianza de las dimensiones A, B y C respectivamente. La fórmula de la covarianza es muy similar a la fórmula de la varianza. La fórmula de la covarianza con respecto a la varianza también podría escribirse así

Donde,

$\bar{X}$  = Media aritmética de los datos X

$\bar{Y}$  = Media aritmética de los datos Y

$n$  = Número de observaciones

$$Var(X) = \frac{\sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})}{(n - 1)}$$

$$Cov(X, Y) = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{(n - 1)}$$



**Figura 4** formula de la covarianza

La matriz de covarianza

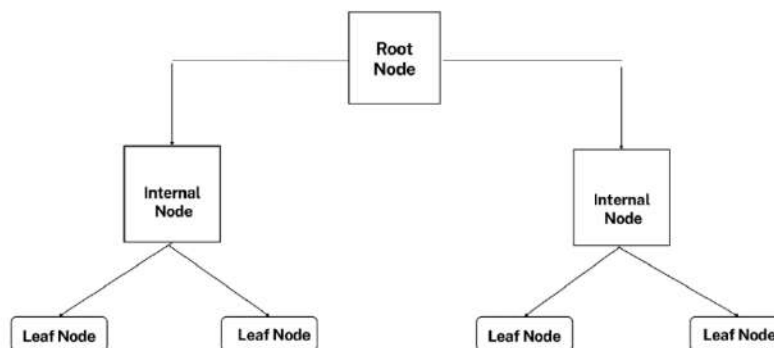
La covarianza se mide siempre entre 2 dimensiones. Si tenemos un conjunto de datos con más de 2 dimensiones, hay más de una medida de covarianza que se puede calcular. Por ejemplo, a partir de un conjunto de datos de 3 dimensiones (dimensiones x, y, z) podríamos calcular la  $cov(x,y)$ ,  $cov(y,z)$  y  $cov(x,z)$ . De hecho, para un conjunto de datos de n dimensiones.

Vectores propios y valores propios

Los vectores propios y los valores propios son números y vectores asociados a matrices cuadradas. Juntos proporcionan la descomposición propia de una matriz, que analiza la estructura de una matriz. Aunque la descomposición propia no existe para todas las matrices cuadradas, tiene una expresión especialmente sencilla para matrices como las de correlación, covarianza o producto cruzado. La descomposición propia de este tipo de matrices es importante porque se utiliza para encontrar el máximo (o el mínimo) de las funciones que implican estas matrices. En concreto, el PCA se obtiene a partir de la descomposición propia de una matriz de covarianza o de correlación [43].

Tomando en cuenta que para poder realizar una selección de variables o selección de características también llamado de esta manera en el cual se aplica el termino para los métodos estadísticos es necesario utilizar algunos de estos métodos uno de ellos es el árbol de decisión el cual es un algoritmo de aprendizaje supervisado que se utiliza para trabajos de clasificación como de regresión. Este árbol de decisión tiene una estructura en particular el cual incluye un nodo raíz, ramas, nodos internos y por ultimo nodos hoja.

Tomando en cuenta que para poder realizar una selección de atributos es necesario utilizar algunos de estos métodos uno de ellos es el árbol de decisión



**Figura 5** Estructura de un árbol de decisión

Como se observa en la figura 1, un árbol de decisión empieza con un nodo raíz. Las ramas salientes alimentan los nodos internos, también conocidos como nodos de decisión. En función de las cosas disponibles, ya que ambos tipos de nodos realizan evaluaciones para formar conjuntos homogéneos.

Algunos de los más populares son el ID3 el cual es un algoritmo que aprovecha la entropía y la ganancia de información como métricas para evaluar las divisiones de candidatos. Y C4.5 el cual se considera un algoritmo como la versión posterior de ID3 que también fue desarrollado por Quinlan. Puede utilizar la ganancia de información o las proporciones de ganancia para evaluar los puntos de división dentro de los árboles de decisión.

Si bien los árboles de decisión se pueden usar en una variedad de casos de uso, los árboles de decisión son particularmente útiles para minería de datos. Algunos de los beneficios son:

Fácil de interpretar: En si la lógica booleana y algunas representaciones visuales de algunos árboles de decisión y así son más fáciles de entender.

Poca o ninguna preparación de datos: los arboles de decisión tienen una serie de características que son más flexibles que otros clasificadores. Y pueden manejar valores discretos y continuos.

Más flexible: Los árboles de decisión se pueden aprovechar para tareas de clasificación y regresión, lo que los hace más flexibles que otros algoritmos.

Así de esta manera se pueden realizar los métodos estadísticos que son;

Correlación que consiste en un procedimiento estadístico para determinar si dos variables están relacionadas o no. El resultado del análisis es un coeficiente de correlación que toma valores entre -1 y +1. El signo nos dice que el tipo de correlación entre las dos variables. Un signo positivo indica que existe una relación positiva entre dos variables. La fuerza de la relación lineal incrementa a medida que el coeficiente de correlación se aproxima a -1 o a +1 [31].

Otra técnica popular de selección de características consiste en calcular la ganancia de información. Puede calcular la ganancia de información (también llamada entropía) de cada atributo para la variable de salida. Los valores de entrada varían de 0 que indica que no contiene tanta información a 1 que es lo máximo de información que puede tener. Aquellos atributos que aporten más información tendrán un valor de ganancia de información más alto y podrán ser seleccionados, mientras que aquellos que no aporten mucha información tendrán una puntuación más baja y podrán ser eliminados.

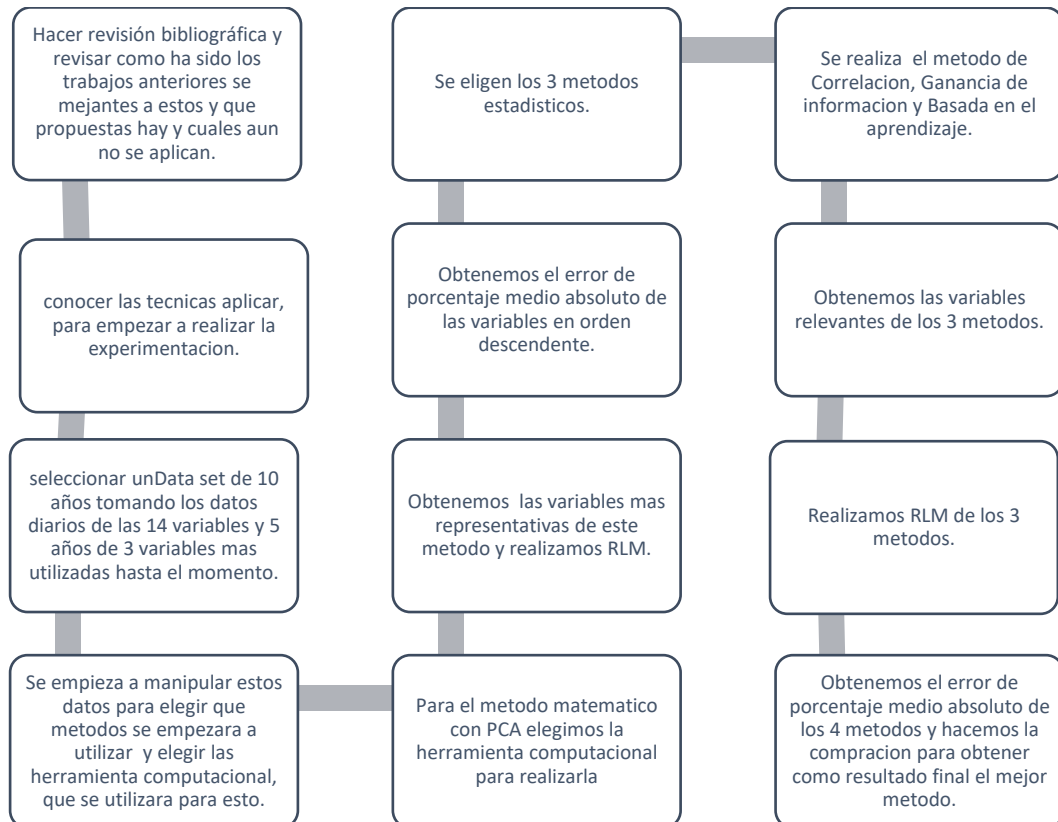
Basado en el aprendizaje esta técnica similar a la anterior, pero siendo una extensión del ID3 en la cual los árboles pueden ser usados para clasificación y por esta razón C4.5 está casi siempre referido como un clasificador estadístico.

En este caso un árbol de decisión comienza con un nodo raíz, que no tiene ramas entrantes. Las ramas salientes del nodo raíz alimentan los nodos internos, también conocidos como nodos de decisión. En función de las características disponibles, ambos tipos de nodos realizan evaluaciones para formar subconjuntos homogéneos, que se indican mediante nodos hoja. Los nodos hoja representan todos los resultados posibles dentro del conjunto de datos. [32]

Para comparar los 4 métodos de selección de variables se utilizará regresión lineal múltiple el cual permite generar un modelo lineal en el que el valor de la variable dependiente o respuesta (Y) se determina a partir de un conjunto de variables independientes llamadas predictores ( $X_1, X_2, X_3 \dots$ ). Es una extensión de la regresión lineal simple, por lo que es fundamental comprender esta última. Los modelos de regresión múltiple pueden emplearse para predecir el valor de la variable dependiente o para evaluar la influencia que tienen los predictores sobre ella. [35]

### 3 Metodología

Para tener un mejor panorama de lo que se realizara en el proyecto se hará una metodología en donde se explique paso a paso lo que se hará.



**Figura 6** Es la metodología detallada de lo que se realizara en el proyecto.

Los procesos de la Figura 5 se describen brevemente a continuación:

1. Hacer revisión bibliográfica y revisar cómo ha sido los trabajos anteriores semejantes a estos y que propuestas hay y cuales aún no se aplican.
2. Conocer las técnicas aplicar, para empezar a realizar la experimentación con valores de prueba de pocos días de 80 a 100 días aproximadamente para tener conocimiento de cómo trabajar con el data set más adelante.
3. Seleccionar el primer data set de 10 años (2012-2022) en wheater ground las 14 variables. Después seleccionar el segundo data set de 5 años (2017-2021) de las 14 variables ya

mencionadas y agregándole las variables de la calidad del aire, sumando así 17 en total. Esto para hacer una evaluación más

4. completa y más certera
5. Se empieza a manipular estos datos para elegir que métodos se empezara a utilizar y elegir las herramientas computacionales, que se utilizara para esto.
6. Se realiza PCA:

Paso 1: obtener los datos del año 2012 al 2022

Paso 2: Para que el PCA funcione correctamente, se resta la media de todos los datos de cada una de las dimensiones de los datos. La media restada es la media de cada dimensión. Así, a todos los valores  $x$  se les resta  $x'$  (la media de los valores  $x$  de todos los puntos de datos), y a todos los valores  $y$  se les resta  $y'$ . Esto produce un conjunto de datos cuya media es cero.

Paso 3: Calcular la matriz de covarianza

Esto se hace exactamente de la misma manera que hemos discutido anteriormente. Como los datos son bidimensionales, la matriz de covarianza será de  $2 \times 2$

Paso 4: Calcular los vectores propios y los valores propios de la matriz de covarianza

Como la matriz de covarianza es cuadrada, se pueden calcular los vectores propios y los valores propios de la matriz. Este paso es el más importante, ya que ayuda a obtener información útil sobre los datos.

Paso 5: Elección de los componentes y formación de un vector de características

Aquí es donde entra en juego la compresión de datos y la reducción de la dimensionalidad.

Obtención del nuevo conjunto de datos

Interpretación geométrica

El PCA proyecta los datos a lo largo de las direcciones en las que los datos varían más.

Estas direcciones están determinadas por los vectores propios de la matriz de covarianza correspondientes a los mayores valores propios.

La magnitud de los valores propios corresponde a la varianza de los datos a lo largo de las direcciones de los vectores propios.

(esto fue realizado en Excel para la primera parte de la experimentación y selección de variables)

7. Se realizar la regresión lineal múltiple para poder pronosticar la temperatura que será la variable independiente El modelo de regresión lineal múltiple es idéntico al modelo de regresión lineal simple, con la única diferencia de que aparecen más variables explicativas:  
Modelo de regresión múltiple:

$$Y_i = (\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_n X_{ni}) + e_i$$

**Figura 7** representacion de RLM

Se realiza en excel en analisis de datos en la parte de regresion en donde la temperatura es el valor de Y y las otras variables este procedimiento en X y B.

8. Obtenemos el error de porcentaje medio absoluto. Se calcula de la siguiente manera entre cada par de variables:

Paso 1: Se resta la diferencia entre el valor pronosticado y el valor real en cada punto que se pronostica.

Paso 2: Se saca el valor absoluto del punto anterior.

Paso 3: Se divide cada valor entre el valor real de cada punto en el tiempo que se pronostica.

Paso 4: Se suman todos los valores obtenidos Se divide el resultado de la suma entre la cantidad de puntos pronosticados

9. De métodos estadísticos se eligen 3 los cuales optamos mejor para esta investigación según la literatura.
10. Se realizan los 3 métodos de Correlación, Ganancia de información y basado en aprendizaje con la ayuda de un software de aprendizaje automático llamado weka.

### Selección de atributos por correlación

Calcular el coeficiente de correlación entre dos variables es:  $r = Cov_{xy} / S_{xx} S_{yy}$

El coeficiente de correlación es el resultado de dividir la covarianza entre las variables X y Y entre la raíz cuadrada del producto de la varianza de X y la de Y.

Calcular la covarianza entre la variable X y la variable Y (entre las dos columnas de la matriz) de

acuerdo a la siguiente fórmula:  $Cov_{xy} = \frac{\sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y})}{N}$

Se calcula la media de todos los valores de X y de Y Se realiza la sumatoria del producto de las diferencias entre cada observación de cada variable y su media correspondiente. La sumatoria calculada anteriormente se divide entre el número total de observaciones.

Calcular las varianzas de la variable (X) y la varianza de la variable (Y) y obtener la raíz cuadrada de

cada una:  $\sqrt{\sum_{i=1}^n \frac{(X_i - \bar{X})^2}{N}} \quad \sqrt{\sum_{i=1}^n \frac{(Y_i - \bar{Y})^2}{N}}$

Para cada variable se calcula la desviación estándar y se multiplican.

Se divide la covarianza entre el producto de las desviaciones estándar.

En Weka admite la selección de características basada en la correlación con la técnica `CorrelationAttributeEval` que requiere el uso de un método de búsqueda Ranker.

### **Selección de características basada en la ganancia de información**

Puede calcular la ganancia de información de cada atributo para la variable de salida. Los valores de entrada varían de 0 que indica que no contiene tanta información a 1 que es lo máximo de información que puede tener. Aquellos atributos que aporten más información tendrán un valor de ganancia de información más alto y podrán ser seleccionados, mientras que aquellos que no aporten mucha información tendrán una puntuación más baja y podrán ser eliminados.

En Weka soporta la selección de características mediante la ganancia de información utilizando el evaluador de atributos `InfoGainAttributeEval`. Al igual que la técnica de correlación anterior, se debe utilizar el método de búsqueda Ranker.

### **Selección de características basada en el aprendizaje**

Esta técnica similar a la anterior, pero siendo una extensión del ID3 en la cual los arboles pueden ser usados para clasificación y por esta razón C4.5 está casi siempre referido como un clasificador estadístico.

En este caso un árbol de decisión comienza con un nodo raíz, que no tiene ramas entrantes. Las ramas salientes del nodo raíz alimentan los nodos internos, también conocidos como nodos de decisión. En función de las características disponibles, ambos tipos de nodos realizan evaluaciones para formar subconjuntos homogéneos, que se indican mediante nodos hoja. Los nodos hoja representan todos los resultados posibles dentro del conjunto de datos. [32]

En Weka este tipo de selección de características está soportado por la técnica `WrapperSubsetEval` y debe utilizar un método de búsqueda `GreedyStepwise` o `BestFirst`. Este último, `BestFirst`, es preferible si puede prescindir del tiempo de cálculo.

- a). selecciona la técnica "`WrapperSubsetEval`".
- b). Después "`WrapperSubsetEval`" para abrir la configuración del método.
- c). "Elegir" para el "clasificador" y cámbielo a J48 en "árboles".

11. Se obtienen las variables relevantes de los 4 métodos. (PCA, correlacion, ganancia de información, y basado en el aprendizaje).
12. Se realiza Regresión Lineal Múltiple en los 4 métodos Obtener las variables relevantes que más han afectado al cambio climático contiene la siguiente ecuación:

$$Y_i = (\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_n X_{ni}) + e_i$$

En la cual  $\beta_0$  es la ordenada en el origen, el valor de la variable dependiente y cuando los predictores son cero,  $\beta_i$  es el efecto promedio que tiene como incremento en una unidad de la variable predictor  $X_i$  sobre la variable dependiente y manteniéndose constantes al resto de variables. Se conocen como parciales de regresión. [42]

13. Obtenemos el Error Porcentual Medio Absoluto de los 4 métodos y observar cual es el que contiene menor error y tomar en cuenta que este sería el mejor método.



---

## 4 Resultados y conclusiones

En este capítulo se mostrarán dos experimentaciones, en la primera serán las ciudades Monterrey, Guadalajara, Tampico y Tijuana con 14 variables y 4 métodos matemáticos y estadísticos; en la segunda experimentación serán las ciudades de Guadalajara, Monterrey y Tijuana, con 17 variables y 4 métodos matemáticos y estadísticos.

### 4.1 Experimentación 1

En la primera experimentación se utilizarán las siguientes variables: temperatura máxima ( $T_{max}$ ), temperatura promedio ( $T_{prom}$ ), temperatura mínima ( $T_{min}$ ), punto de rocío máximo ( $PR_{min}$ ), punto de rocío promedio ( $PR_{prom}$ ), punto de rocío mínimo ( $PR_{min}$ ), humedad máxima ( $H_{max}$ ), humedad promedio ( $H_{prom}$ ), velocidad del viento máxima ( $VV_{max}$ ), velocidad del viento promedio ( $VV_{prom}$ ), velocidad del viento mínimo ( $VV_{min}$ ), y presión atmosférica máxima ( $P_{max}$ ), presión atmosférica promedio ( $P_{prom}$ ) y presión atmosférica mínima ( $P_{min}$ ).

Los métodos utilizados en esta experimentación son: PCA, correlación, información de la ganancia y basada en el aprendizaje. Y los datos utilizados fueron obtenidos del sitio weather underground y sinaica y comprendieron del periodo de Enero 2012 a Diciembre 2022.

### 4.2 Experimentación 2

En la segunda parte de la experimentación se utilizarán las siguientes variables: temperatura máxima ( $T_{max}$ ), temperatura promedio ( $T_{prom}$ ), temperatura mínima ( $T_{min}$ ), punto de rocío

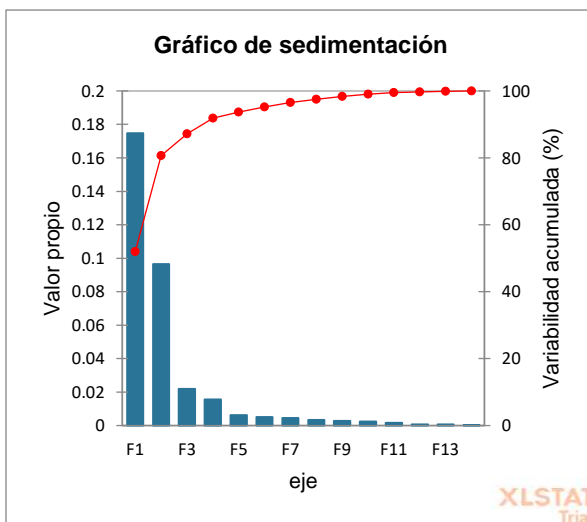
máximo (PR<sub>min</sub>), punto de rocío promedio (PR<sub>prom</sub>), punto de rocío mínimo (PR<sub>min</sub>), humedad máxima (H<sub>max</sub>), humedad promedio (H<sub>prom</sub>), velocidad del viento máxima (VV<sub>max</sub>), velocidad del viento promedio (VV<sub>prom</sub>), velocidad del viento mínimo (VV<sub>min</sub>), y presión atmosférica máxima (P<sub>max</sub>), presión atmosférica promedio (P<sub>prom</sub>), presión atmosférica mínima (P<sub>min</sub>) monóxido de carbono(CO), Óxido nítrico(NO) y PM 2.5

Los métodos utilizados en esta experimentación son: PCA, correlación, información de la ganancia y basada en el aprendizaje. Y los datos utilizados fueron obtenidos del sitio weather underground y sinaica, comprendiendo el periodo de Enero 2017 a Diciembre 2021.

## Resultados de experimentación 1

### PCA Monterrey

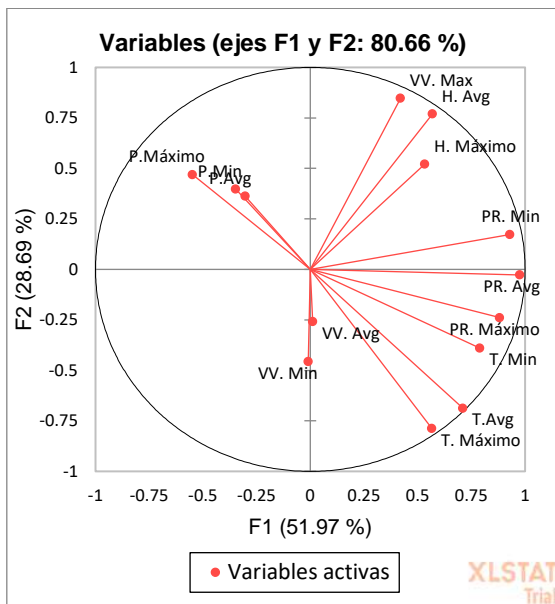
En esta sección mostramos un gráfico de sedimentación del método de PCA de la ciudad de Monterrey donde se muestran los valores propios.



**Figura 8** Grafico de valores propios y variabilidad acumulada de PCA en Monterrey.

En el grafico podemos observar el porcentaje acumulado de la información en la línea punteada color rojo. De la cual utilizaremos el 91% de la información, para poder obtener de este porcentaje de las variables con mayor información, ocuparemos hasta el eje número 4 (F4) .

A continuación se tiene un circulo de correlación utilizado en el método de PCA que muestra una proyección de las variables iniciales en el espacio factorial.



**Figura 9** Mapa del círculo de correlaciones de PCA de Monterrey.

Cuando dos variables están lejos del centro, tenemos varias posibilidades: si están próximas una a la otra, están positivamente correlacionadas (punto de la variable está próximo a 1); si son ortogonales, no están correlacionadas (punto de la variable está próximo a 0); si están en lados opuestos con respecto al centro están negativamente correlacionadas ( punto de la variable está próximo a -1).

La siguiente tabla nos muestra la contribución de las variables en cada uno de los 4 ejes con mayor información de Monterrey.

	F1	F2	F3	F4
<b>T. Máx</b>	5.413	18.974	2.493	0.711
<b>T.Avg</b>	8.515	14.427	0.087	0.370
<b>T. Min</b>	9.372	4.136	9.002	0.340
<b>PR. Máx</b>	9.825	1.301	6.606	0.704
<b>PR. Avg</b>	22.015	0.032	2.208	0.119
<b>PR. Min</b>	30.390	1.900	18.248	1.238
<b>H. Máximo</b>	4.157	7.199	41.209	2.782
<b>H. Avg</b>	5.434	18.008	3.811	1.738
<b>VV. Max</b>	3.806	27.975	9.472	0.542
<b>VV. Avg</b>	0.000	0.397	0.151	16.083
<b>VV. Min</b>	0.001	3.598	3.686	75.183

<b>P.Máx</b>	0.519	0.689	0.553	0.173
<b>P.Avg</b>	0.245	0.635	2.355	0.014
<b>P.Min</b>	0.307	0.729	0.120	0.003

La **tabla 1** Contribuciones de las variables, el porcentaje representado de cada uno en los 4 ejes representativos de Monterrey

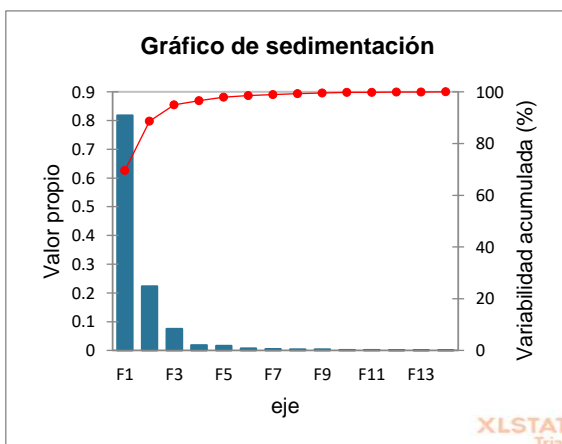
A continuación, están las variables con mayor información acumulada según los 4 ejes más representativos.

<b>Orden</b>	<b>Variables</b>	<b>% de cada eje(1,2,3,4)</b>
1	PR. Min	30
2	PR. avg	22
3	PR. max	9
4	T.min	9
5	T.Avg	8
6	VV.max	27
7	T. max	18
8	VV. Min	75

**Tabla 2** Variables más representativas y porcentaje con PCA de monterrey

### PCA Guadalajara

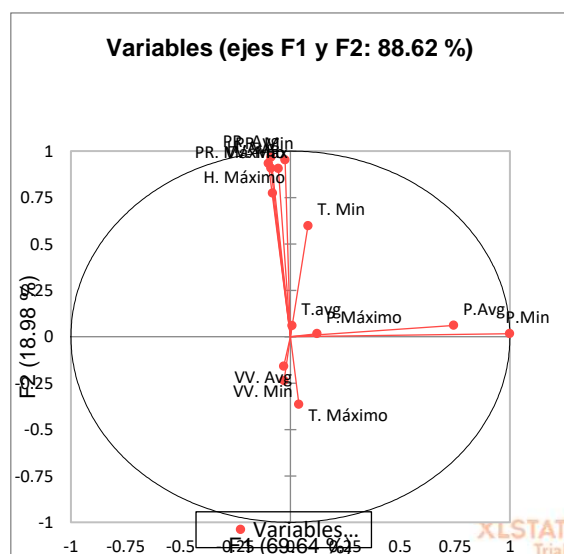
En esta sección mostramos un gráfico de sedimentación del método de PCA de la ciudad de Guadalajara donde se muestran los valores propios.



**Figura 10** Grafico de valores propios y variabilidad acumulada de PCA en Guadalajara.

En el grafico podemos observar el porcentaje acumulado de la información en la línea punteada color rojo. De la cual utilizaremos el 94% de la información, para poder obtener de este porcentaje de las variables con mayor información, ocuparemos hasta el eje número 3 (F3) .

A continuación se tiene un círculo de correlación utilizado en el método de PCA que muestra una proyección de las variables iniciales en el espacio factorial.



**Figura 11** Círculo de correlaciones de PCA de Guadalajara.

Cuando dos variables están lejos del centro, tenemos varias posibilidades: si están próximas una a la otra, están positivamente correlacionadas (punto de la variable está próximo a 1); si son ortogonales, no están correlacionadas (punto de la variable está próximo a 0); si están en lados opuestos con respecto al centro están negativamente correlacionadas (punto de la variable está próximo a -1).

La siguiente tabla nos muestra la contribución de las variables en cada uno de los 3 ejes con mayor información de Guadalajara.

	F1	F2	F3
<b>T. Máximo</b>	0.002	0.741	9.370
<b>T.avg</b>	0.000	0.041	31.724
<b>T. Min</b>	0.030	6.132	26.947
<b>PR. Máximo</b>	0.023	8.912	1.348
<b>PR. Avg</b>	0.041	18.089	1.059
<b>PR. Min</b>	0.005	27.445	1.064
<b>H. Máximo</b>	0.030	10.121	11.119
<b>H. Avg</b>	0.053	16.626	4.919
<b>VV. Max</b>	0.011	11.205	0.451
<b>VV. Avg</b>	0.002	0.151	2.961
<b>VV. Min</b>	0.002	0.436	8.957
<b>P.Máximo</b>	0.006	0.000	0.046
<b>P.Avg</b>	0.279	0.007	0.012
<b>P.Min</b>	99.516	0.094	0.025

La **tabla 3** Contribuciones de las variables, el porcentaje representado de cada uno en los 3 ejes representativos de Guadalajara.

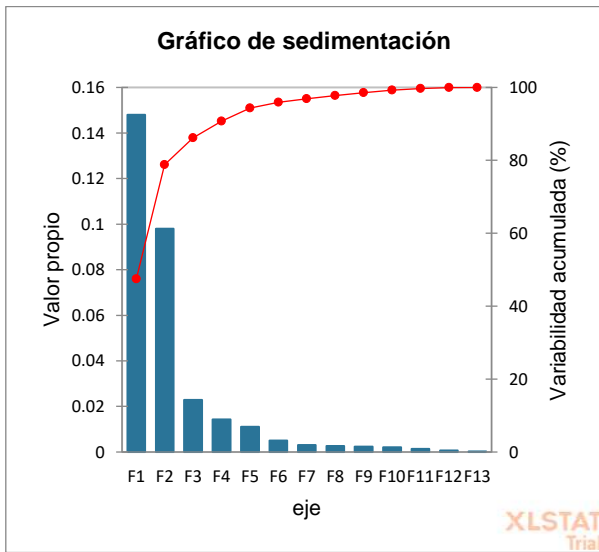
A continuación, están las variables con mayor información acumulada según los 3 ejes más representativos.

Orden	Variables	%
1	P.min	99
2	P.avg	0.2
3	PR.min	27
4	PR.avg	18
5	H. avg	16
6	PR. max	8
7	VV. max	11
8	H. max	10

**Tabla 4** Variables más representativas y porcentaje con PCA de Guadalajara

### PCA Tijuana

En esta sección mostramos un gráfico de sedimentación del método de PCA de la ciudad de Tijuana donde se muestran los valores propios.

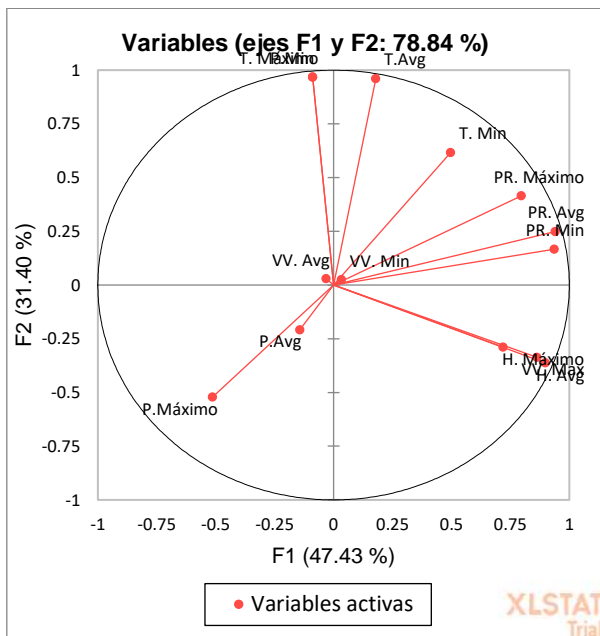


**Figura 12** Grafico de valores propios y variabilidad acumulada de PCA en Tijuana.

En el grafico podemos observar el porcentaje acumulado de la información en la línea punteada color rojo. De la cual utilizaremos el 90% de la información, para poder obtener de este porcentaje de las variables con mayor información, ocuparemos hasta el eje número 4 (F4) .

A continuación se tiene un circulo de correlación utilizado en el método de PCA que muestra una proyección de las variables iniciales en el espacio factorial.





**Figura 13** Mapa del círculo de correlaciones de PCA de Tijuana.

Cuando dos variables están lejos del centro, tenemos varias posibilidades: si están próximas una a la otra, están positivamente correlacionadas ( punto de la variable está próximo a 1); si son ortogonales, no están correlacionadas ( punto de la variable está próximo a 0); si están en lados opuestos con respecto al centro están negativamente correlacionadas (punto de la variable está próximo a -1).

La siguiente tabla nos muestra la contribución de las variables en cada uno de los 4 ejes con mayor información de Tijuana.

	F1	F2	F3	F4
<b>T. Máx</b>	0.135	24.280	3.064	1.558
<b>T.Avg</b>	0.505	21.920	0.606	0.621
<b>T. Min</b>	2.569	5.980	9.549	5.421
<b>PR. Max</b>	5.551	2.276	5.187	0.103
<b>PR. Avg</b>	16.663	1.764	1.750	0.098
<b>PR. Min</b>	22.521	1.077	4.720	6.070
<b>H. Máx</b>	9.054	2.209	34.092	1.224
<b>H. Avg</b>	18.898	4.649	4.947	0.000
<b>VV. Max</b>	19.461	4.464	16.015	1.760
<b>VV. Avg</b>	0.003	0.003	0.750	0.649
<b>VV. Min</b>	0.009	0.007	13.891	6.517
<b>P.Máx</b>	4.440	6.915	2.107	67.787

<b>P.Avg</b>	0.055	0.175	0.258	6.634
<b>P.Min</b>	0.135	24.280	3.064	1.558

La **tabla 5** Contribuciones de las variables, el porcentaje representado de cada uno en los cuatro ejes representativos de Tijuana.

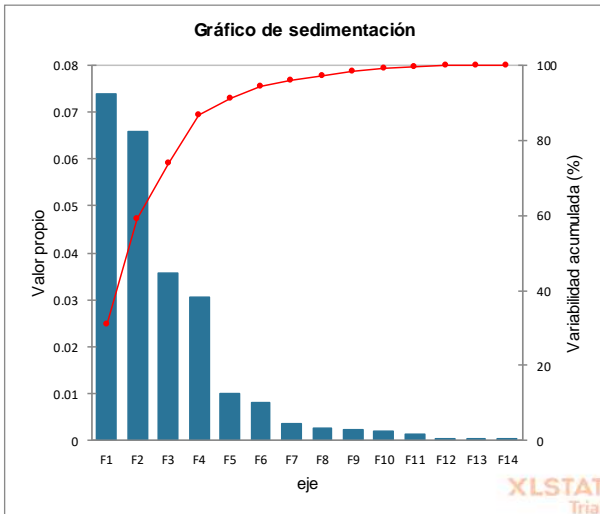
A continuación, están las variables con mayor información acumulada según los 4 ejes más representativos.

<b>Orden</b>	<b>Variables</b>	<b>%</b>
1	PR. avg	16
2	PR. Min	22
3	H. avg	18
4	VV.max	19
5	PR.max	5
6	H.max	9
7	T. Max	24
8	T. avg	21

**Tabla 6** Variables más representativas y porcentaje con PCA de Tijuana.

### **PCA Tampico**

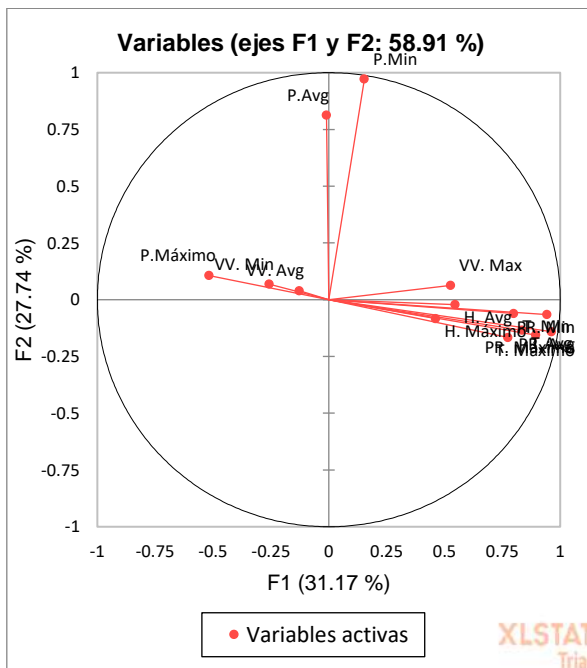
En esta sección mostramos un gráfico de sedimentación del método de PCA de la ciudad de Tampico donde se muestran los valores propios.



**Figura 14** Grafico de valores propios y variabilidad acumulada de PCA en Tampico.

En el grafico podemos observar el porcentaje acumulado de la información en la línea punteada color rojo. De la cual utilizaremos el 91% de la información, para poder obtener de este porcentaje de las variables con mayor información, ocuparemos hasta el eje número 5 (F5) .

A continuación se tiene un circulo de correlación utilizado en el método de PCA que muestra una proyección de las variables iniciales en el espacio factorial.



**Figura 15** Mapa del círculo de correlaciones de PCA de Tampico.

Cuando dos variables están lejos del centro, tenemos varias posibilidades: si están próximas una a la otra, están positivamente correlacionadas (punto de la variable está próximo a 1); si son ortogonales, no están correlacionadas (punto de la variable está próximo a 0); si están en lados opuestos con respecto al centro están negativamente correlacionadas (punto de la variable está próximo a -1).

La siguiente tabla nos muestra la contribución de las variables en cada uno de los 5 ejes con mayor información de Tampico.

	F1	F2	F3	F4	F5
<b>T. Max</b>	11.901	0.617	0.198	14.574	4.267
<b>T.Avg</b>	9.265	0.272	0.190	7.865	0.073
<b>T. Min</b>	8.944	0.058	0.282	4.312	6.066
<b>PR. Max</b>	4.285	0.146	0.109	0.401	1.487
<b>PR. Avg</b>	18.940	0.455	0.208	0.351	1.081
<b>PR. Min</b>	26.396	0.143	0.146	0.022	8.477
<b>H. Max</b>	3.795	0.137	0.006	14.897	47.785
<b>H. Avg</b>	5.059	0.009	0.000	24.821	2.320
<b>VV. Max</b>	6.627	0.107	0.030	31.455	24.236
<b>VV. Avg</b>	0.758	0.083	95.173	0.209	0.091
<b>VV. Min</b>	0.363	0.029	3.529	0.087	0.562

<b>P.Max</b>	1.905	0.090	0.074	0.453	0.821
<b>P.Avg</b>	0.003	19.125	0.044	0.025	1.260
<b>P.Min</b>	1.759	78.729	0.011	0.527	1.475

La **tabla 7** Contribuciones de las variables, el porcentaje representado de cada uno en los cinco ejes representativos de Tampico.

A continuación, están las variables con mayor información acumulada según los 4 ejes más representativos.

<b>Orden</b>	<b>Variables</b>	<b>%</b>
1	PR. avg	18
2	PR. min	26
3	PR.max	4
4	T. avg	9
5	T. min	8
6	T. max	11
7	P. min	78
8	VV. avg	95

**Tabla 8** Variables más representativas y porcentaje con PCA de Tampico.

### **Resultados PCA de todas las ciudades**

Al observar los resultados de PCA de las cuatro ciudades a estudiar, en la que observa que en Monterrey y Guadalajara, el punto de rocío y la humedad son las más relevantes. Y en la ciudad de Tijuana la relevante es el punto de rocío y en Tampico específicamente la velocidad del viento ya que al estar en costa está más expuesto a eso.

Los métodos estadísticos realizados es la correlación, Ganancia de información y basado en aprendizaje.

### 4.3 Resultados de correlación

A continuación, están los resultados del método de correlación los cuales están en orden de mayor a menor en la tabla de cada ciudad.

#### Monterrey

<b>SELECCIÓN DE ATRIBUTOS</b>	
<b>CORRELACION</b>	
<b>ATRIBUTOS</b>	<b>CLASIFICADO</b>
T. avg	0.22
T. Max	0.21
PR. Max	0.19
T. max	0.19
PR. Avg	0.18

**Tabla 9** Resultados de correlación de Monterrey

#### Tijuana

<b>SELECCIÓN DE ATRIBUTOS</b>	
<b>CORRELACION</b>	
<b>ATRIBUTOS</b>	<b>CLASIFICADO</b>
T. avg	0.19
PR. avg	0.18
PR. Max	0.18
P. max	0.17
T. min	0.17

**Tabla 10** Resultados de correlación de Tijuana

#### Guadalajara

<b>SELECCIÓN DE ATRIBUTOS</b>	
<b>CORRELACION</b>	
<b>ATRIBUTOS</b>	<b>CLASIFICADO</b>
PR. avg	0.23
PR. max	0.22
H. avg	0.21
T. avg	0.21
P. min	0.21

**Tabla 11** Resultados de correlación de Guadalajara.

## Tampico

SELECCIÓN DE ATRIBUTOS	
CORRELACION	
ATRIBUTOS	CLASIFICADO
T. avg	0.22
T. max	0.20
PR. max	0.18
PR. avg	0.18
T. min	0.16

**Tabla 12** Resultados de correlación de Tijuana.

En las tablas anteriores podemos ver que el punto de rocío y la temperatura son las variables más relevantes.

### 4.4 Selección de características por ganancia de información

Sabiendo que la ganancia de información es la entropía de un árbol antes de la división menos la entropía ponderada después de la división por un atributo, la cantidad de aleatoriedad disminuye la ganancia de información o la cantidad de certeza aumenta y viceversa.

### Resultados de Ganancia de información

A continuación, se muestra el método de ganancia de información en donde los resultados están ordenados de mayor a menor en cada tabla. Dando como resultado las variables más relevantes de cada ciudad.

## Monterrey

SELECCIÓN DE ATRIBUTOS	
GANANCIA DE INFORMACION	
ATRIBUTOS	CLASIFICADO
T. min	0.85
T. avg	0.62
PR. max	0.68
T. max	0.58
PR. avg	0.66

**Tabla 13** Resultados de ganancia de información de Monterrey

### Tijuana

<b>SELECCIÓN DE ATRIBUTOS</b>	
<b>GANANCIA DE INFORMACION</b>	
<b>ATRIBUTOS</b>	<b>CLASIFICADO</b>
T. min	0.82
PR. avg	0.75
PR. max	0.71
T. avg	0.67
PR. min	0.62

**Tabla 14** Resultados de ganancia de información de Tijuana.

### Guadalajara

<b>SELECCIÓN DE ATRIBUTOS</b>	
<b>GANANCIA DE INFORMACION</b>	
<b>ATRIBUTOS</b>	<b>CLASIFICADO</b>
T. avg	0.99
PR. avg	0.93
T. min	0.91
PR. max	0.86
H. max	0.81

**Tabla 15** Resultados de ganancia de información de Guadalajara.

### Tampico

<b>SELECCIÓN DE ATRIBUTOS</b>	
<b>GANANCIA DE INFORMACION</b>	
<b>ATRIBUTOS</b>	<b>CLASIFICADO</b>
T. avg	0.86
T. min	0.75
T. max	0.68
PR. max	0.60
PR. avg	0.60

**Tabla 16** Resultados de ganancia de información de Tampico.

Los resultados de las cuatro ciudades con el método de ganancia de información muestran que la temperatura y punto de rocío son las variables más relevantes



#### 4.5 Selección de características basada en aprendizaje.

Al realizar esta técnica en la cual se elige un atributo de los datos que más eficazmente dividen el conjunto de muestras en subconjuntos enriquecidos en una clase u otra. Su criterio es el normalizado para ganancia de información que es la diferencia de entropía, que resulta en la elección de un atributo para dividir los datos. El atributo con la mayor ganancia de información normalizada se elige como parámetro de decisión.

#### Resultados de basada en aprendizaje

A continuación, observamos los resultados de las variables más relevantes de las 4 ciudades utilizando el método basado en aprendizaje.

##### Monterrey

SELECCIÓN DE ATRIBUTOS
BASADO EN EL APRENDIZAJE
<b>MONTERREY ATRIBUTOS</b>
T. avg
PR. max
H. max
P. max

**Tabla 17** Resultados de basado en el aprendizaje de Monterrey.

##### Tijuana

SELECCIÓN DE ATRIBUTOS
BASADO EN EL APRENDIZAJE
<b>TIJUANA ATRIBUTOS</b>
T. min
PR. max
PR. min
H. max

**Tabla 18** Resultados de basado en el aprendizaje de Tijuana.

## Guadalajara

SELECCIÓN DE ATRIBUTOS
BASADO EN EL APRENDIZAJE
<b>GUADALAJARA ATRIBUTOS</b>
T. avg
PR. max
H. max
H. avg

**Tabla 19** Resultados de basado en el aprendizaje de Guadalajara.

## Tampico

SELECCIÓN DE ATRIBUTOS
BASADO EN EL APRENDIZAJE
<b>TAMPICO ATRIBUTOS</b>
T. avg
T. min
H. max
H. avg

**Tabla 20** Resultados de basado en el aprendizaje de Tampico.

De las tablas anteriores notamos que las temperaturas y el punto de rocío son las variables más relevantes.

En el caso de monterrey la temperatura y el punto de rocío ha sido lo que más ha sobresalido y se contempla que el clima es algo extremo tanto en verano como en invierno.

Guadalajara punto de rocío, humedad en el cual se sabe que es importante porque es responsable de los procesos de precipitación y por lo tanto un beneficio para la lluvia en nuestro planeta.

Tijuana punto de rocío es el más relevante en el cual el clima se ve reflejado que es templado y seco y con algunas lluvias en invierno.

En Tampico la humedad y temperatura han sido la más relevantes contemplando que es un clima húmedo la mayor parte del tiempo.

#### 4.6 Regresión Lineal Múltiple de los 4 métodos

Como se tenía planteado se realizó la regresión lineal múltiple para poder llegar a una conclusión y ver cuál de los 4 métodos es el mejor contando que se realiza después de esto un error porcentual absoluto medio el cual nos ayudara para saber cuál es el porcentaje de menor error y así tener cual es el mejor método posible y sus comparaciones. Teniendo como resultado que el mejor es el de ganancia de información seguida de la correlación, basada en aprendizaje y por ultimo PCA.

	<i>ERROR</i>	<i>PORCENTUAL</i>	<i>ABSOLUTO MEDIO</i>	<i>(MAPE)</i>
METODOS	PCA	CORRELACION	GANANCIA DE INFORMACION	BASADA EN APRENDIZAJE
CIUDADES				
MONTERREY	2%	0%	0%	0%
GUADAJALARA	2%	0%	0%	1%
TAMPICO	0%	0%	0%	0%
TIJUANA	0%	1%	0%	8%

**Tabla 21** Resultados del MAPE de Regresión Lineal múltiple de primera experimentación.

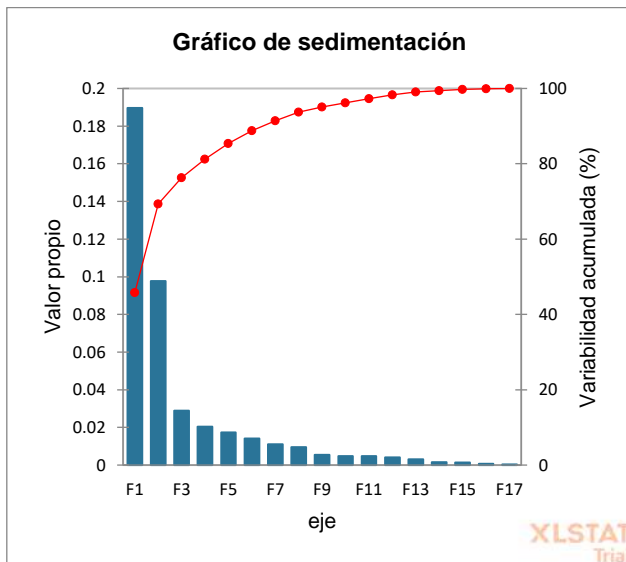
## 4.6 Segunda parte de experimentación

En esta segunda parte de la experimentación agregamos 3 variables más a las ya obtenidas anteriormente con la diferencia que el tiempo a estudiar es de 5 años para las 17 variables de 2017 al 2021. Nuestras variables agregadas son contempladas en la calidad del aire que son monóxido de carbono (CO), Oxido Nítrico(NO) y Partículas menores a 2.5, las cuales se obtuvieron del sistema Nacional información de calidad del Aire. De las ciudades de monterrey, Guadalajara y Tijuana. A diferencia de la primera parte de la experimentación en esta la ciudad de Tampico no se obtuvieron los datos para poder agregarla. Sin embargo, con las 3 ciudades obtenidas se hará esta experimentación.

## 4.7 Resultados PCA

### MONTERREY

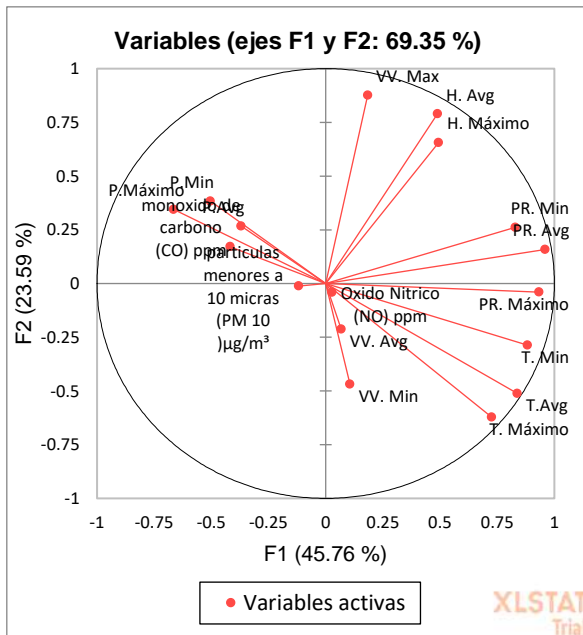
En esta sección mostramos un gráfico de sedimentación del método de PCA de la ciudad de Monterrey donde se muestran los valores propios.



**Figura 16** Grafico de valores propios y variabilidad acumulada de PCA en Monterrey.

En el grafico podemos observar el porcentaje acumulado de la información en la línea punteada color rojo. De la cual utilizaremos el 91% de la información, para poder obtener de este porcentaje de las variables con mayor información, ocuparemos hasta el eje número 5 (F5) .

A continuación se tiene un círculo de correlación utilizado en el método de PCA que muestra una proyección de las variables iniciales en el espacio factorial.



**Figura 17** Mapa del círculo de correlaciones de PCA de Monterrey.

Cuando dos variables están lejos del centro, tenemos varias posibilidades: si están próximas una a la otra, están positivamente correlacionadas ( punto de la variable está próximo a 1); si son ortogonales, no están correlacionadas (punto de la variable está próximo a 0); si están en lados opuestos con respecto al centro están negativamente correlacionadas (punto de la variable está próximo a -1).

La siguiente tabla nos muestra la contribución de las variables en cada uno de los 5 ejes con mayor información de Monterrey.

	F1	F2	F3	F4	F5
<b>T. Máximo</b>	8.514	12.123	2.240	2.146	0.400
<b>T.Avg</b>	10.694	7.726	0.506	0.086	0.001
<b>T. Min</b>	16.298	3.340	0.037	7.861	0.289
<b>PR. Max</b>	7.241	0.026	0.022	0.073	0.571
<b>PR. Avg</b>	27.688	1.461	0.341	1.326	0.136
<b>PR. Min</b>	13.565	2.619	0.413	27.256	0.427
<b>H. Max</b>	3.998	13.684	0.152	24.975	5.688
<b>H. Avg</b>	4.347	22.002	2.241	8.695	0.067
<b>VV. Max</b>	0.585	25.831	2.824	13.002	0.013
<b>VV. Avg</b>	0.015	0.290	5.964	0.179	2.557

VV. Min	0.179	6.744	71.940	0.083	9.851
P.Máx	3.004	1.569	3.219	1.545	1.692
P.Avg	0.916	0.928	2.874	4.549	1.274
P.Min	0.892	1.002	0.696	1.812	2.061
Monóxido de carbono (CO) ppm	1.979	0.644	6.474	5.130	72.071
Partículas menores a 10 micras (PM 10) $\mu\text{g}/\text{m}^3$	0.083	0.002	0.028	1.237	2.690
Óxido Nítrico (NO) ppm	0.002	0.009	0.030	0.046	0.212

La tabla 22 Contribuciones de las variables, el porcentaje representado de cada uno en los cinco ejes representativos de Monterrey.

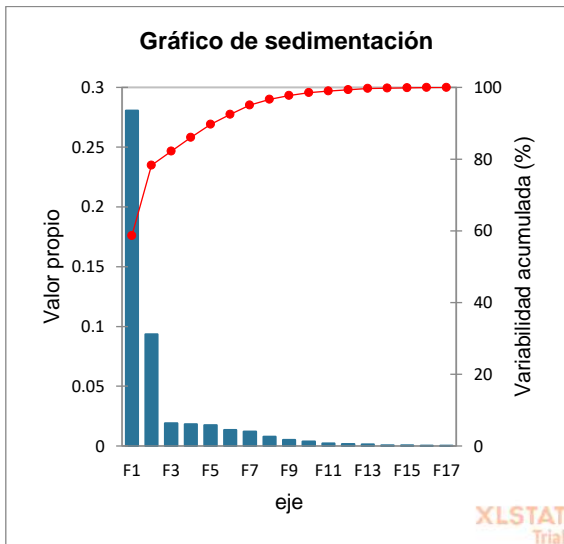
A continuación, están las variables con mayor información acumulada según los 5 ejes

PCA	
Monterrey	%
PR. avg	27
PR. max	7
T. Min	16
T. Avg	13
T. Máximo	10
PR. min	13
VV. max	25
H.avg	22

**Tabla 23** Variables más representativas y porcentaje con PCA de Monterrey.

## PCA GUADALAJARA

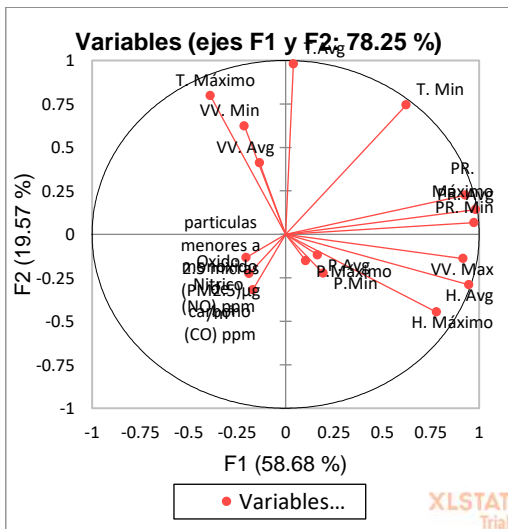
En esta sección mostramos un gráfico de sedimentación del método de PCA de la ciudad de Guadalajara donde se muestran los valores propios.



**Figura 18** Grafico de valores propios y variabilidad acumulada de PCA en Guadalajara.

En el grafico podemos observar el porcentaje acumulado de la información en la línea punteada color rojo. De la cual utilizaremos el 90% de la información, para poder obtener de este porcentaje de las variables con mayor información, ocuparemos hasta el eje número 5 (F5) .

A continuación se tiene un círculo de correlación utilizado en el método de PCA que muestra una proyección de las variables iniciales en el espacio factorial.



**Figura 19** Mapa del círculo de correlaciones de PCA de Guadalajara.

Cuando dos variables están lejos del centro, tenemos varias posibilidades: si están próximas una a la otra, están positivamente correlacionadas ( punto de la variable está próximo a 1); si son ortogonales, no están correlacionadas ( punto de la variable está próximo a 0); si están en lados opuestos con respecto al centro están negativamente correlacionadas (punto de la variable está próximo a -1).

La siguiente tabla nos muestra la contribución de las variables en cada uno de los 5 ejes con mayor información de Guadalajara.

	<b>F1</b>	<b>F2</b>	<b>F3</b>	<b>F4</b>	<b>F5</b>
<b>T. Máximo</b>	1.223	15.413	4.714	7.364	5.251
<b>T.Avg</b>	0.018	32.250	0.395	1.768	1.108
<b>T. Mín</b>	5.482	23.388	0.720	0.014	0.232
<b>PR. Máximo</b>	12.229	2.154	2.524	2.095	0.024
<b>PR. Avg</b>	20.916	1.316	0.956	1.240	0.064
<b>PR. Mín</b>	21.414	0.308	0.156	0.015	0.002
<b>H. Máximo</b>	7.472	7.379	8.537	4.558	0.094
<b>H. Avg</b>	17.752	4.989	0.000	0.000	0.126
<b>VV. Max</b>	12.417	0.863	7.841	6.074	2.997
<b>VV. Avg</b>	0.081	2.262	0.305	1.229	30.625
<b>VV. Mín</b>	0.252	6.466	1.822	3.175	29.977
<b>P. Máx</b>	0.013	0.082	0.000	0.040	0.127
<b>P.Avg</b>	0.030	0.046	0.027	0.038	0.126
<b>P.Mín</b>	0.042	0.169	0.038	0.061	0.044
<b>Monóxido de carbono (CO) ppm</b>	0.145	1.540	0.039	8.966	3.165
<b>Oxido Nítrico (NO) ppm</b>	0.237	1.019	7.723	63.134	10.206
<b>Partículas menores a 2.5 micras (PM2.5)µg/m<sup>3</sup></b>	0.278	0.356	64.204	0.228	15.830

**Tabla 24** Contribuciones de las variables, el porcentaje representado de cada uno en los cinco ejes representativos de Guadalajara.



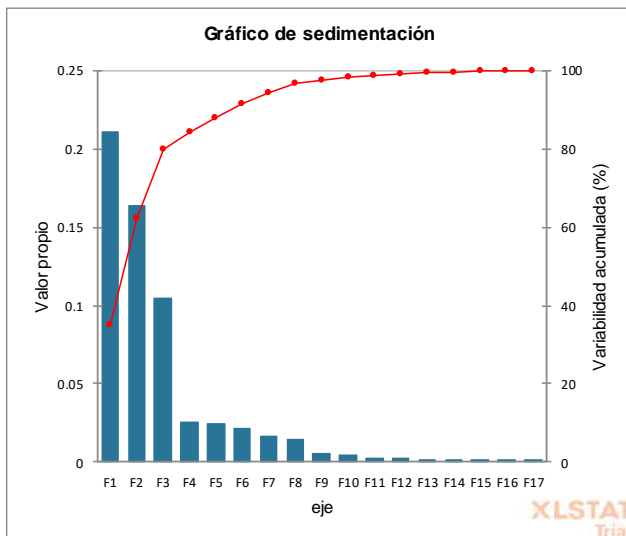
A continuación, están las variables con mayor información acumulada según los 5 ejes

PCA	
Guadalajara	%
PR. Avg	20
PR. Min	21
H. Avg	17
PR. Max	12
VV. Max	17
H. Max	7
T. Avg	32
P.M 2.5	64
NO	63

**Tabla 25** Variables más representativas y porcentaje con PCA de Guadalajara.

## PCA TIJUANA

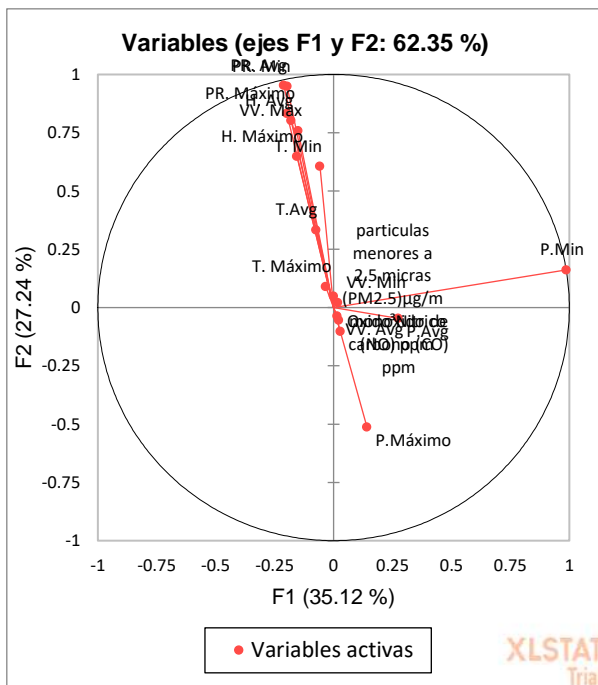
En esta sección mostramos un gráfico de sedimentación del método de PCA de la ciudad de Tijuana donde se muestran los valores propios.



**Figura 20** Grafico de valores propios y variabilidad acumulada de PCA en Tijuana.

En el grafico podemos observar el porcentaje acumulado de la información en la línea punteada color rojo. De la cual utilizaremos el 91% de la información, para poder obtener de este porcentaje de las variables con mayor información, ocuparemos hasta el eje número 3 (F3) .

A continuación se tiene un círculo de correlación utilizado en el método de PCA que muestra una proyección de las variables iniciales en el espacio factorial.



**Figura 21** Mapa del círculo de correlaciones de PCA de Tijuana.

Cuando dos variables están lejos del centro, tenemos varias posibilidades: si están próximas una a la otra, están positivamente correlacionadas ( punto de la variable está próximo a 1); si son ortogonales, no están correlacionadas ( punto de la variable está próximo a 0); si están en lados opuestos con respecto al centro están negativamente correlacionadas (punto de la variable está próximo a -1).

La siguiente tabla nos muestra la contribución de las variables en cada uno de los 3 ejes con mayor información de Tijuana.

<b>Contribuciones de las variables (%):</b>			
	F1	F2	F3
<b>T. Max</b>	0.015	0.141	24.246
<b>T.Avg</b>	0.120	3.080	37.537
<b>T. Min</b>	0.042	5.884	12.700
<b>PR. Max</b>	0.326	7.403	1.485
<b>PR. Avg</b>	0.693	18.174	0.369
<b>PR. Min</b>	0.900	26.844	0.032

<b>H. Max</b>	0.281	6.310	3.857
<b>H. Avg</b>	0.547	13.970	9.397
<b>VV. Max</b>	0.441	14.412	9.737
<b>VV. Avg</b>	0.000	0.003	0.003
<b>VV. Min</b>	0.003	0.005	0.041
<b>P.Max</b>	0.013	0.218	0.172
<b>P.Avg</b>	0.215	0.008	0.138
<b>P.Min</b>	96.392	3.332	0.001
<b>(NO) ppm</b>	0.004	0.032	0.136
<b>(PM2.5)µg/m<sup>3</sup></b>	0.000	0.031	0.146
<b>(CO) ppm</b>	0.009	0.153	0.002

La **tabla 26** Contribuciones de las variables, el porcentaje representado de cada uno en los tres ejes representativos de Tijuana.

A continuación, están las variables con mayor información acumulada según los 3 ejes.

<b>PCA</b>	
<b>Tijuana</b>	<b>%</b>
P.min	96
PR. avg	18
PR. min	26
PR. max	6
H. avg	13

**Tabla 27** Variables más representativas y porcentaje con PCA de Tijuana.

En este caso de PCA podemos observar que los más relevantes en las 3 ciudades es el punto de rocío junto con otras variables, pero el que más influye es ese.

En esta técnica de correlación, la variable más relevante en las 3 ciudades han sido las temperaturas y el punto de rocío es algo similar al anterior solo con la diferencia que en el caso de monterrey una de las relevantes es el monóxido de carbono.

#### 4.8 Resultados de correlación

A continuación, están resultados del método de correlación los cuales están en orden de mayor a menor en la tabla de cada ciudad.

##### Monterrey

<b>SELECCIÓN DE ATRIBUTOS</b>	
<b>CORRELACION</b>	
<b>ATRIBUTOS MONTERREY.</b>	<b>CLASIFICADO</b>
T. avg	0.22
T. min	0.21
PR. Max	0.19
T. max	0.19
PR. Avg	0.18
PR. Min	0.16
CO	0.14

Tabla 28 Resultados de correlación de Monterrey

##### Guadalajara

<b>SELECCIÓN DE ATRIBUTOS</b>	
<b>CORRELACION</b>	
<b>ATRIBUTOS GUADALAJARA</b>	<b>CLASIFICADO</b>
T. min	0.24
PR. Avg	0.24
PR.max	0.23
PR. Min	0.23
H. avg	0.22
T. avg	0.22
VV. max	0.20

Tabla 29 Resultados de correlación de Guadalajara

##### Tijuana

<b>SELECCIÓN DE ATRIBUTOS</b>	
<b>CORRELACION</b>	
<b>ATRIBUTOS TIJUANA</b>	<b>CLASIFICADO</b>
T. min	0.23
T. avg	0.21
PR. Max	0.19
PR. avg	0.19
P. max	0.18
PR. min	0.17
T. max	0.08

**Tabla 30** Resultados de correlación de Monterrey

#### 4.9 Ganancia de información

Sabiendo que la ganancia de información es la entropía de un árbol antes de la división menos la entropía ponderada después de la división por un atributo, la cantidad de aleatoriedad disminuye la ganancia de información o la cantidad de certeza aumenta y viceversa.

A continuación, se muestra el método de ganancia de información en donde los resultados están ordenados de mayor a menor en cada tabla. Dando como resultado las variables más relevantes de cada ciudad.

En esta técnica la temperatura y el punto de rocío han sido las más relevantes, en el caso de la calidad de aire en monterrey el moxido de carbono es relevante.

#### Monterrey

<b>GANANCIA DE INFORMACION</b>	
<b>ATRIBUTOS MONTERREY</b>	<b>CLASIFICADO</b>
T. min	0.84
T. avg	0.77
PR. max	0.66
CO	0.58
PR. avg	0.57
T. max	0.53
PR. min	0.35

**Tabla 31** Resultados de ganancia de información de Monterrey.

#### Guadalajara.

<b>GANANCIA DE INFORMACION</b>	
<b>ATRIBUTOS GUADALAJARA</b>	<b>CLASIFICADO</b>
T. avg	1.00
PR. avg	0.96
T. min	0.94
H. avg	0.90
PR.max	0.86
H. max	0.85
PR. min	0.80

**Tabla 32** Resultados de ganancia de información de Guadalajara.

## Tijuana

<b>GANANCIA DE INFORMACION</b>	
<b>ATRIBUTOS TIJUANA</b>	<b>CLASIFICADO</b>
T. min	1.02
PR. avg	0.86
PR. max	0.84
T. avg	0.84
PR. min	0.70
T. max	0.48
P. max	0.45

**Tabla 33** Resultados de ganancia de información de Tijuana.

### 4.10 Basada en el aprendizaje

Al realizar esta técnica en la cual se elige un atributo de los datos que más eficazmente dividen el conjunto de muestras en subconjuntos enriquecidos en una clase u otra. Su criterio es el normalizado para ganancia de información que es la diferencia de entropía, que resulta en la elección de un atributo para dividir los datos. El atributo con la mayor ganancia de información normalizada se elige como parámetro de decisión.

A continuación, observamos los resultados de las variables más relevantes de las 4 ciudades utilizando el método basado en aprendizaje.

## Monterrey

<b>SELECCIÓN DE ATRIBUTOS</b>
<b>BASADO EN EL APRENDIZAJE</b>
<b>ATRIBUTOS MONTERREY</b>
T. avg
T. min
PR. max
PR. min
H. max
P. avg
CO.

**Tabla 34** Resultados de basado en el aprendizaje de Tampico.

### Guadalajara.

<b>SELECCIÓN DE ATRIBUTOS</b>
<b>BASADO EN EL APRENDIZAJE</b>
<b>ATRIBUTOS GUADALAJARA</b>
T. max
T.avg
T.min
PR. max
PR. avg
PR. min
H. max

**Tabla 35** Resultados de basado en el aprendizaje de Guadalajara.

### Tijuana

<b>SELECCIÓN DE ATRIBUTOS</b>
<b>BASADO EN EL APRENDIZAJE</b>
<b>ATRIBUTOS TIJUANA</b>
T. avg
T. min
PR. max
PR. min
VV. min
CO
NO

**Tabla 36** Resultados de basado en el aprendizaje de Tijuana.

En el caso de Monterrey y Tijuana podemos ver que su variable más relevante es el monóxido de carbono el cual se produce por la combustión incompleta de cualquier material que contenga carbono: gas, petróleo, carbón, kerosén, nafta, madera, plásticos, etc.; en un ambiente con una concentración de oxígeno escasa o insuficiente.

Por otra parte, en Guadalajara es notable que las temperaturas junto con el punto de rocío el óxido nítrico son las más relevantes por lo tanto este último elemento de la calidad del aire en zonas urbanas es la oxidación incompleta del nitrógeno atmosférico en los motores de combustión interna, y por ello, en dichas áreas los automóviles son los principales culpables de las emisiones de este contaminante. Y seguida del monóxido de carbono que es por la combustión incompleta de cualquier material con este elemento.

#### 4.11 Resultados de Regresión Lineal múltiple de los 4 métodos

Los resultados de las técnicas para obtener el menor error (MAPE) en el cual se observa que la técnica basada en aprendizaje es la que tiene menor error y por lo tanto la que puede ser utilizada para poder tener resultados más cercanos a lo ideal.

	<b>ERROR</b>	<b>PORCENTUAL</b>	<b>ABSOLUTO MEDIO</b>	<b>(MAPE)</b>
METODOS	<b>PCA</b>	<b>CORRELACION</b>	<b>GANANCIA DE INFORMACION</b>	<b>BASADA EN APRENDIZAJE</b>
CIUDADES				
<b>MONTERREY</b>	<b>1.6%</b>	<b>0%</b>	<b>0%</b>	<b>0%</b>
<b>GUADAJALARA</b>	<b>0%</b>	<b>0%</b>	<b>0%</b>	<b>0%</b>
<b>TIJUANA</b>	<b>1.9%</b>	<b>2.9%</b>	<b>2.1%</b>	<b>2%</b>

**Tabla 37** Resultados del MAPE de Regresión Lineal múltiple de segunda experimentación



## Conclusiones

Se realizaron los 4 métodos en los cuales nos dieron resultados algo diferentes en cada uno, pero llegando al final con la regresión lineal múltiple la cual nos permitió conocer el error porcentual absoluto medio de cada método y obteniendo que el de ganancia de información y en la segunda parte de la experimentación es la basada en aprendizaje las cuales parten de una misma teoría y fueron los que obtuvieron el menor y por lo tanto el mejor para poder realizarlo en futuras ocasiones en otras ciudades del país o del mundo, teniendo en cuenta que nos daría mejores resultados que los demás. Tomando en cuenta que las dos técnicas tienen como base la entropía que mide la aleatoriedad de las variables así que parten de la misma característica es por eso que ambas obtuvieron los mejores resultados. También se percató que las ciudades tienen predominantes de la calidad del aire en los cuales se puede ver cuáles son, y con los resultados poder tomar alguna acción conjunta para que esto mejore dado el caso de Monterrey y Tijuana que en ambas ciudades de presencia de vehículos e industrias son factores importantes para complementar que los elementos de la calidad del aire como el monóxido de carbono y óxido nítrico son elevados. Por otra parte esperando que los resultados de cada ciudad sean de ayuda para que tomándolos en cuenta en futuras ocasiones se sepa de donde partir, y de que variables tener más atención en específico ya que serían los que más han impactado en el cambio climático, claro que por muchos motivos puede variar dependiendo la ciudad ya que por ejemplo en la ciudad de Monterrey está a 540 metros sobre el nivel del mar y Guadalajara casi 3 veces más con 1566 metros y la distribución geográfica, el incremento de la población pero al todas estar en México comparten ciertas características en común.

## 5. Bibliografía

- [1] Cambio climático, Semarnat ciencia evidencia y acciones, SEMARNAT, Gobierno Federal, México 2009, [http://centro.paot.org.mx/documentos/semarnat/cambio\\_climatico.pdf](http://centro.paot.org.mx/documentos/semarnat/cambio_climatico.pdf), último acceso: 09 Marzo 2022.
- [2] R.K. Pachauri and L.A. Meyer Core Writing Team, Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. IPCC 2014, Geneva, Switzerland, pág.2.
- [3] Huang Guanwen, Zhu Hai, Zhang, Juqing, Liu Bohan, “Analysis of the Characteristics of Climate Change in the Ecologically Vulnerable Area of the Mu Us Dune Field under the Background of Global Warming”. Remote Sensing Journal, 2021, 13, 627.
- [4] IPCC, 2013: Glosario [Planton, S. (ed.)]. En: Cambio Climático 2013. Bases físicas. Contribución del Grupo de trabajo I al Quinto Informe de Evaluación del Grupo Intergubernamental de Expertos sobre el Cambio Climático [Stocker, T.F., D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex y P.M. Midgley (eds.)]. Cambridge University Press, Cambridge, Reino Unido y Nueva York, NY, Estados Unidos de América.

- [5] Jóvenes frente al cambio Climático. <http://jovenesfrentealcambioclimatico.com/glosario/cambio-climatico> Ultimo acceso: 17 de marzo del 2022
- [6] NASA-Global Climate Change, <https://climate.nasa.gov/causas/>. Ultimo acceso: 17 de marzo del 2022
- [7] Henry Oswaldo Benavides, Gloria Esperanza León. Información técnica sobre gases de efecto invernadero y el cambio climático 2007 Nta Técnica de IDEAM Gases de efecto invernadero y el cambio climático ([ideam.gov.co](http://ideam.gov.co))
- [8] Marco A. Cruz Chavez Licenciatura en electrónica y computación: métodos numéricos VALORES Y VECTORES PROPIOS ([uaem.mx](http://uaem.mx)) uaem pag 23-24
- [9] Software del sol Geolit, Parque Científico y Tecnológico Glosario Varianza ¿Qué es? ([sdelsol.com](http://sdelsol.com))
- [10] Luis Miguel Galindo José Luis Samaniego José Eduardo Alatorre Jimmy Ferrer Carbonell Reflexiones metodológicas del análisis del cambio climático Una visión desde América Latina. Comisión Económica para América Latina y el Caribe (CEPAL) 2014 pp 9-32. [https://repositorio.cepal.org/bitstream/handle/11362/37608/1/S1500008\\_es.pdf](https://repositorio.cepal.org/bitstream/handle/11362/37608/1/S1500008_es.pdf)
- [11] H. Millán A. Kalauzi b G. Llerena a, J. Sucoshañay a, D. Piedra a. Climatic trends in the Amazonian area of Ecuador: Classical and multifractal analyses; Atmospheric Research 2007
- [12] Mario Molina, Jose Sarukhan, Julia Carabias “El cambio climático” pág. 25
- [13] Huang Guanwen, Zhu Hai, Zhang, Juqing, Liu Bohan, “Analysis of the Characteristics of Climate Change in the Ecologically Vulnerable Area of the Mu Us Dune Field under the Background of Global Warming”. Remote Sensing Journal, 2021, 13, 627.

- [14] Margarita Caballero, Socorro Lozano-García , Beatriz Ortega-Guerrero a , Alexander Correa-Metrio. “Quantitative estimates of orbital and millennial scale climatic variability in central Mexico during the last ~40,000 years” journal el servier 2021
- [15] Lidija Tadić<sup>1</sup> & Ognjen Bonacci<sup>2</sup> & Tamara Brleković<sup>1</sup> *An example of principal componet analysis application on climate change assessment*. Springer-Verlag GmbH Austria, part of Springer Nature 2019
- [16] International development in focus The Global Health Cost of PM2.5 Air Pollution: A Case for Action Beyond world gran group 2021 pag 6 -7
- [17] Assessment of Spatio-temporal Climatological trends of ozone over the Indian region using Machine Learning
- [18] Braham Sanchez Lopez , Viridiana Cruz-Gutierrez<sup>1</sup> , Mario Alberto Posada-Zamora<sup>1</sup> , M. Teresa Torrijos M.<sup>1</sup> , M. Auxilio Osorio Lama<sup>2</sup> estudio del analisis de componentes principales en bases de datos de calidad del aire Estudio del análisis de componentes principales en bases de datos de calidad del aire (ipn.mx) pag 10 y 11 Estudio del análisis e componentes principales en bases de datos de calidad del aire (ipn.mx)
- [19] Pedro Guido Aldama Cambio climatico selección, clasificación y diseño de medidas de adaptación 2017 pag 13-
- [20] Charles stidd El uso de vectores propios para estimaciones climáticas Instituto de investigación del desierto 1996
- [21] Rubaiya Kabir, Marin Akter , Dewan Sadia Karim, Anisul Haque , Munsur Rahman and Mohiuddin Sakib. Development of a Matrix Based Statistical

Framework to Compute Weight for Composite Hazards, Vulnerability and Risk Assessments journal climate 2019

- [22] Cristine Dreveton Use of a Principal Components Analysis for the Generation of Daily Time Series Escuela de meteorología de Toulouse France Febrero 2004
- [23] Use of Stella Software for the Modelling of Climate Change Impacts on Water Balance for the Rio Yaqui Basin, Sonora, Mexico Article in European Scientific Journal · April 2014
- [24] [1 ] Linear Models with R (Chapman & Hall/CRC Texts in Statistical Science)
- [25] Amit Kumar Yadav, Hasmat Malik SS. Chandel “Selection of most relevant input parameters using Weka for artificial neural network based solar radiation prediction models” Department of Electrical Engineering Indian Institute. El sevier Journal 2013
- [26] Lidija Tadić<sup>1</sup> & Ognjen Bonacci<sup>2</sup> & Tamara Brleković<sup>1</sup> An example of principal component analysis application on climate change assessment. Springer-Verlag GmbH Austria, part of Springer Nature 2019
- [27] Huang Guanwen, Zhu Hai, Zhang, Juqing, Liu Bohan, “Analysis of the Characteristics of Climate Change in the Ecologically Vulnerable Area of the Mu Us Dune Field under the Background of Global Warming”. Remote Sensing Journal, 2021, 13, 627.
- [28] Venkata Ramana Moorthy, B Sarojamma and S Venkatramana Reddy. “Development of Machine Learning Modelos using WEKA for Atmospheric Data” India Jorunal of Physics: Conference Series 2022
- [29] Microsoft PowerPoint - Bloque 1.Conceptos basicos sobre el cambio climatico\_cgl.pptx (conafor.gob.mx) commission federal forestal Diapositiva 8
- [30] Pedro Guido Aldama Cambio climatico selección, clasificación y diseño de medidas de adaptación 2017 pag 13

- [31] Alquirica J. Análisis de correlación. Congonasi, José Alquicira (2017, 25 de Mayo)
- [32] Ingrid Wilford Rivera, Alejandro Rosete Suarez, Alfredo Rodriguez Diaz . (2010) aplicación de la minería de datos para el análisis de información clínica estudio experimental en cardiopatías isquémicas. 10 Junio 2016 de revista cubana de información. [www.rcim.sld.cu/revista\\_18/articulos\\_htm/mineriadatos.htm](http://www.rcim.sld.cu/revista_18/articulos_htm/mineriadatos.htm)
- [33] Venkata Ramana Moorthy, B Sarojamma and S Venkatramana Reddy. “Development of Machine Learning Modelos using WEKA for Atmospheric Data” India Jorunal of Physics: Conference Series 2022
- [34] Microsoft PowerPoint - Bloque 1.Conceptos basicos sobre el cambio climatico\_cgl.pptx (conafor.gob.mx) commission federal forestal Diapositiva 8
- [35] Cambio climático, Semarnat ciencia evidencia y acciones, SEMARNAT, Gobierno Federal, México 2009, [http://centro.paot.org.mx/documentos/semarnat/cambio\\_climatico.pdf](http://centro.paot.org.mx/documentos/semarnat/cambio_climatico.pdf), último acceso: 09 Marzo 2022.
- [36] Clase-Presión.pdf (meteorito.mx) UNAM Pag.4
- [37] Redalyc.Caracterización fisicoquímica del material particuladofracción respirable PM2.5 en Pamplona-Norte de Santander-Colombia
- [38] Cambio climático, Semarnat ciencia evidencia y acciones, SEMARNAT, Gobierno Federal, México 2009, [http://centro.paot.org.mx/documentos/semarnat/cambio\\_climatico.pdf](http://centro.paot.org.mx/documentos/semarnat/cambio_climatico.pdf), último acceso: 09 Marzo 2022.
- [39] Caldwell, J.C. y T. Schindimayr. Historical population estimates: un raveling the concensus. Population and Development Review ervicios ambientales de los ecosistemas y bienestar Pag 10y 28: 183-204. 2002. Cap1\_poblacion.pdf (semarnat.gob.mx)

- [40] Periodistas especializados en periodismo e información ambiental de la Agencia Madrid Journal Efeverde Investigadores del CSIC presentan modelo matemático para explicar clima (efeverde.com) [www.efeverde.com](http://www.efeverde.com)
- [41] Software del sol Geolit, Parque Científico y Tecnológico Glosario Varianza ¿Qué es? (sdelsol.com Periodistas especializados en periodismo e información ambiental de la Agencia Madrid Journal Efeverde Investigadores del CSIC presentan modelo matemático para explicar clima (efeverde.com) [www.efeverde.com](http://www.efeverde.com)
- [42] Linear Models with R (Chapman & Hall/CRC Texts in Statistical Science)  
Julian J. Faraday 1 julio 2014
- [43] Sidharth Prasad Mishra\*, Uttam Sarkar, Subhash Taraphder, Sanjay Datta, Devi Prasanna Swain<sup>1</sup>, Reshma Saikhom, Sasmita Panda<sup>2</sup> and Menalsh Laishram<sup>3</sup>  
Principal Component Analysis Article *in* International Journal of Livestock Research  
· January 2017