UNIVERSIDAD PANAMERICANA
Campus México

UNIVERSIDAD PANAMERICANA

Materia: Gestión de Análisis y Diseño de Comercialización(COM145)

Profesor: Sarahí Aguilar González

Fecha de entrega:

Ciclo: 1222

Nombre del proyecto: Air Pollution

Miembros del Equipo		
ID	**Nombre**	**Carrera**
0194137	Daniela Ahumada Vallejo	ITISI
0211637	Juan Manuel Guerrero Valadez	ITISI

Rúbricas				
ID	**2-social**		**7-knowledge**	
	D	**C**	**A**	**JI**

***Resumen*— El presente documento tiene como objetivo mostrar el desarrollo que se llevó a cabo para el proyecto final de la materia de Gestión de Análisis y Diseño de Comercialización. Nuestra problemática está centrada en la contaminación que se vive en la Ciudad de México, y mediante distintas herramientas, como puede ser el método de agrupamiento K-means y el algoritmo LSTM; se obtuvo un modelo entrenado que predice la contaminación.**
Con este modelo se espera que pueda ser de utilidad para aquellas zonas que se vean más afectadas por la mala calidad de aire.

***Índice de Términos*— Data Science, Contaminación, Machine Learning, K-means, LSTM**

I. Introducción

Uno de los problemas a los que nos enfrentamos hoy en día es a la contaminación del aire. Especialmente en México es un problema que va en aumento y está llegando a niveles nunca antes vistos. Tan solo en la Ciudad de México transitan aproximadamente 5 millones 400 mil coches que emiten gases nocivos para la salud y el planeta, por esto el año pasado superó por seis veces el límite aceptable para poder salir sin tener ninguna repercusión sanitaria.

La contaminación del aire ocurre debido a que la actividad humana modifica la composición del aire natural. Los medios de transporte, la industria y las diferentes formas de generar energía producen sustancias químicas que deterioran la atmósfera y la salud de los seres vivos. De todas las sustancias que se encuentran en el aire las más alarmantes son las partículas suspendidas. Las principales sustancias y partículas que afectan al aire son las siguientes:

- Dióxido de azufre (SO2)
- Monóxido de carbono (CO)
- Ozono (O3)
- Dióxido de nitrógeno (NO2)
- Partículas inhalables (PM10, PM 2.5)

Además de lo mencionado anteriormente también el clima afecta este fenómeno por ejemplo la precipitación, la temperatura, humedad, velocidad y dirección del viento y más son factores que se deben de tomar en cuenta para el estudio de la contaminación del aire.

Debido a las grandes amenazas que presentan se han pensado diversas estrategias para combatir el problema. Hoy en día con todos los avances tecnológicos es importante aprovechar y utilizar estas nuevas herramientas tecnológicas para acercarnos a encontrar una solución que pueda disminuir el problema. En lo que ahora este proyecto se va a enfocar es en contestar la pregunta sobre si es posible saber con precisión el pronóstico de la calidad del aire en una alcaldía de la Ciudad de México utilizando un modelo de Machine Learning.

La inteligencia artificial nos permite desarrollar modelos de machine learning. Machine learning permite que un sistema aprenda a través de los datos, hay dos tipos de aprendizaje: supervisado y no supervisado.

El aprendizaje supervisado es cuando tenemos un conjunto establecido de datos y se tiene una idea de cómo se clasifica, se tiene la intención de encontrar patrones en datos que se pueden aplicar a un proceso de analítica.

El aprendizaje no supervisado se utiliza cuando el problema requiere de una cantidad masiva de datos sin etiquetar, la comprensión del significado detrás de estos datos requiere algoritmos que clasifican los datos con base en los patrones o clústeres que encuentra.

Para un modelo de predicción se necesitan de varios datos y tener las variables independientes y dependientes bien definidas.

II. Visión general del desarrollo

Solución Actual

La solución al problema es desarrollar un modelo de machine learning que nos ayude a predecir con precisión el pronóstico de la calidad del aire específicamente en la Ciudad de México.

III. Revisión y uso de datos

Orígenes de datos y control de datos

Los datos fueron recolectados del gobierno de la Secretaría del Medio Ambiente (SEDEMA) del gobierno de la Ciudad de México. Nuestros datos se dividen en dos partes que son: Meteorología y Contaminantes y van desde el 2018 hasta 2022

IV. Proceso de desarrollo

Solución

Como se mencionó anteriormente nuestra pregunta de investigación es saber si podemos saber la precisión en el pronóstico de la calidad del aire en una alcaldía de la Ciudad de México utilizando un modelo de Machine Learning.
Para lograr responder esta pregunta primero tuvimos que seleccionar cuáles serían las variables utilizadas en el modelo. Las variables se dividen en independientes y dependientes.

- ☐ Variables independiente:
  - ❖ Contaminantes: SO2, NO2, CO, O3, PM10, PM2.5
  - ❖ Clima: Temperatura, Humedad, Precipitación, Velocidad del Viento y Dirección del Viento.
  - ❖ Otras: Alcaldía, Año, Mes, Día de la semana y Día
- ☐ Variable dependiente:
  - ❖ PM10

Es importante tratar de encontrar y trabajar en una solución al problema pues afecta demasiado nuestras vidas e incluso nuestro futuro. Al permitir niveles de contaminación tan altos podemos desarrollar diversos tipos de enfermedades como lo son las cerebrovasculares, cánceres de pulmón y neumopatías crónicas y agudas. Todas estas enfermedades tienen una gran probabilidad de que sean fatales por lo que es crucial tratar de reducir este tipo de riesgos.
No solamente afecta la salud también afecta a la atmósfera y cada día nos vemos más cercano a un cambio climático que convierte en inhabitable a la Tierra.

Manejo de Datos

Una vez que encontramos datos de fuentes confiables y que tuvieran las variables necesarias para el modelo se descargaron los CSV y los juntamos pues originalmente se encuentran separados por año y por variable. Una vez que se juntaron creamos dos datasets diferentes, el primero para los datos del clima que cuenta con temperatura, viento, lluvia y más. El segundo contiene todos los contaminantes como PM10, SO2 entre otros. Posteriormente descargamos otro CSV con el diccionario de todas las estaciones registradas para obtener sus coordenadas. Por último juntamos todo en un solo dataframe para que el análisis de los datos sea más sencillo

Limpieza de datos

Al empezar a analizar los datos pudimos observar que se encuentra una gran cantidad de valores nulos como podemos notar en las siguientes gráficas.

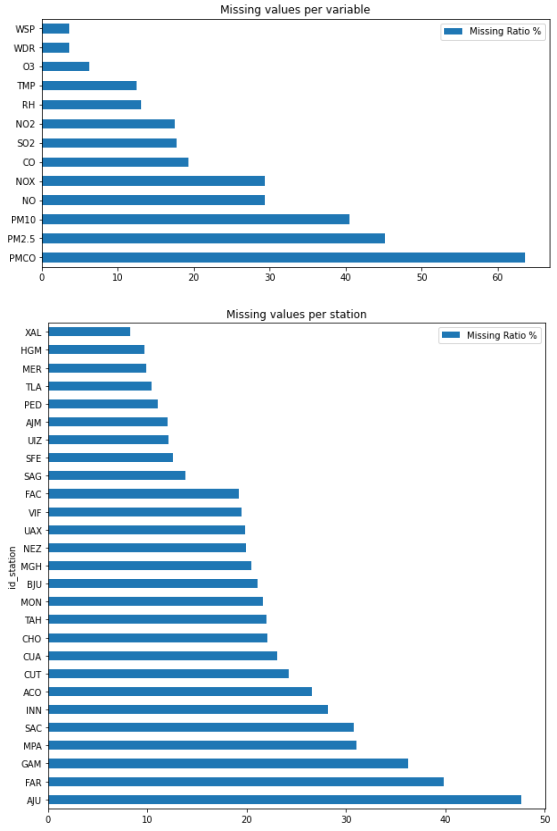


Fig 1: Valores nulos

Sin embargo no nos podemos deshacer de esos valores por lo que aplicamos 3 diferentes pasos para solucionar el problema.

1. Llenar el valor nulo con su valor anterior (hasta 12 horas antes)
2. Juntar por día, al hacer esto podemos conservar la varianza de nuestra variable dependiente que es PM10 y también mantener el comportamiento de contaminación durante el día.
3. Juntar por zonas geográficas, para lograr esto utilizamos k-means. Este algoritmo se utilizó para agrupar en 16 clusters las estaciones según sus coordenadas geográficas.

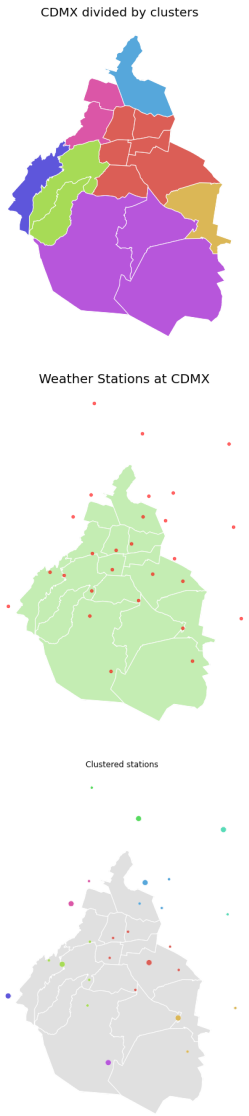
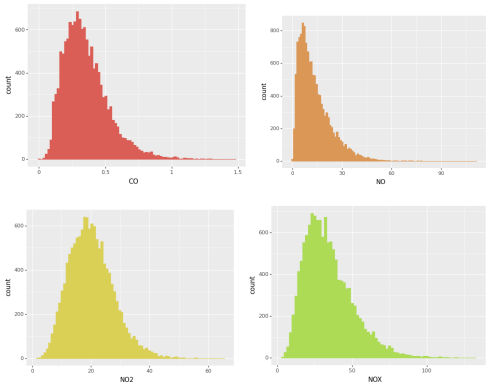


Fig 2: CDMX dividida por clusters

Al hacer un modelo de machine learning tenemos que asegurarnos que los datos estén correctos y limpios para no tener un entrenamiento incorrecto. Por esto generamos diferentes histogramas para ver la distribución de cada variable y asegurarnos que no existan valores anormales que afortunadamente no tenían los datos.



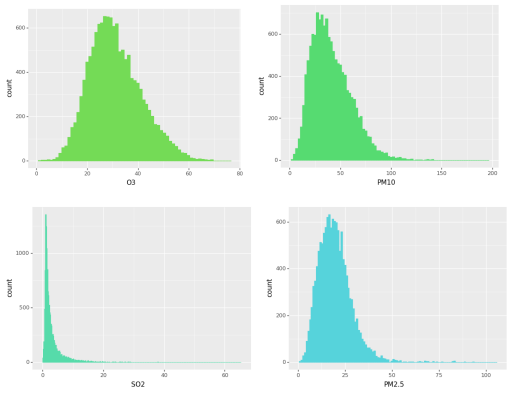


Fig 3: Histogramas

Modelo machine learning

Antes de comenzar el modelo de machine learning realizamos un correlograma para observar qué variables de entrada son las correctas para ingresar en el modelo. Una vez que estas estuvieron definidas empezamos a entrenar nuestro modelo.

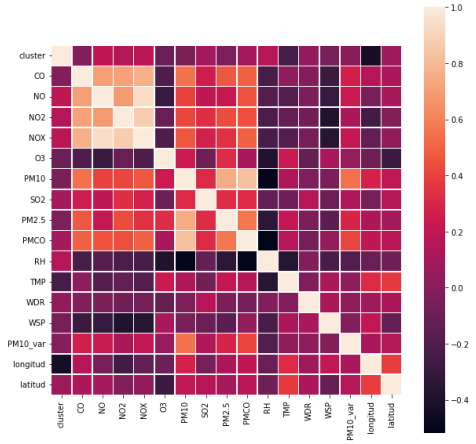


Fig 4: Correlograma

Se decidió utilizar series de tiempo para entrenar el modelo y se eligió la red neuronal de memoria a corto plazo (LSTM) utilizando los datos de los primeros 80% de los días. Como se muestra en la Figura 5, se tomaron los datos meteorológicos de todas las estaciones de los últimos 7 días como variables independientes, y al ser un algoritmo supervisado, se utilizó la velocidad del viento del huracán como variable dependiente.

Para asegurarnos que funcionara correctamente, por último probamos el modelo con el resto de los días que incluían nuestros datos.

V. Resultados

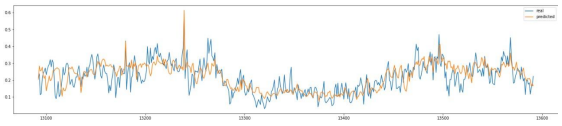


Fig 5: Predicción de Calidad del Aire

En la Figura 5 se muestra en color azul la calidad del aire y en color naranja la predicción hecha por el modelo.
Como se puede observar, el modelo predijo la calidad del aire con una eficacia bastante alta. El modelo cuenta con un MSE de 12.7

Este modelo puede ser utilizado en un ambiente productivo. Si la tendencia de la predicción es positiva, el modelo estaría indicando que puede ocurrir un huracán en los próximos días. Además, al agruparse las estaciones meteorológicas por clusters, con tener al menos un registro climatológico de cada cluster es suficiente para saber si existe una alta probabilidad de contaminación en alguna alcaldía de la Ciudad de México.

VI. Conclusiones

Podemos concluir que el modelo logró predecir el pronóstico de la calidad del aire en la Ciudad de México coincidiendo con las fechas en la que la contaminación suele presentarse más, validando su eficacia.
El modelo puede seguir mejorando, esta fue una primera vista de lo que se puede obtener con una gran cantidad gran cantidad de datos, así como su respectiva preparación, limpieza y procesamiento.

De esta forma se concluye el trabajo desarrollado a lo largo del semestre de la materia de Gestión del Conocimiento. Este tipo de predicciones es sumamente importante para contribuir a la salvación del planeta y a disminuir el riesgo de muerte por enfermedades derivadas por las partículas en el aire. Es nuestra labor como estudiantes y futuros profesionistas aplicar estos conocimientos para buscar soluciones a problemáticas que impactan a nuestra sociedad.

VII. Referencias bibliográficas

1. IBM. (2022). ¿Qué es Machine Learning? 24 de Mayo del 2022, de IBM Sitio Web: https://www.ibm.com/mx-es/analytics/machine-learning
2. National Geographic. (2021). Ciudad de México alcanza niveles históricos de contaminación por partículas suspendidas. 24 de Mayo del 2022, de National Geographic Español Sitio web: https://www.ngenespanol.com/ecologia/ciudad-de-mexico-alcanza-niveles-historicos-de-contaminacion-por-particulas-suspendidas/
3. Aquae fundación. (2021). Contaminación del aire: causas y tipos. 24 de Mayo del 2022, de Aquae fundación Sitio web: https://www.fundacionaquae.org/wiki/causas-y-tipos-de-la-contaminacion-del-aire/

***Abstract* — This document has the objective of showing the development process that took place to realize the final project of Data Science Subject. Our problem is focused on the air pollution that we face in Mexico City, and by applying different tools such as K-means clustering and LSTM algorithm we have the result of a trained model that predicts air pollution. With this model we hope that it can be useful for helping the most affected areas with the biggest levels of air pollution.**

***Index Terms*— Data Science, Contamination, Machine Learning, K-means, LSTM**

I. Introduction

One of the main problems that we face as a society is air pollution. Mexico is one of the countries where this problem is huge, everyday is increasing and reaching new levels never seen before. Just in Mexico City where approximately 5 million 400 thousand cars transit everyday are part of the reason for releasing harmful gasses to the atmosphere and that also affect our health. Because of this the last year surpassed by 6 times more the acceptable limit to go out without having any sanitary repercussions.

Air pollution occurs due to human activity, human activity modifies the composition of natural air. Transportation, the industry and the different ways to generate energy produce chemical substances that deteriorate the atmosphere and the health of living beings. Of all the substances that are found in the air the most alarming ibes are the suspended particles. The principal substances and particles that affect the air are the following ones:

- Sulfur dioxide (SO2)
- Carbon monoxide (CO)
- Ozone (O3)
- Nitrogen dioxide (NO2)
- Inhalable particles (PM10, PM 2.5)

Moreover there are other factors that affect this, one of them is the weather for instance temperature, precipitation, humidity, wind velocity and more are factors that must be taken into account for the study of air pollution and contamination.

Due to the big threat that it is for our society, mankind has thought of different strategies to solve this problem. Nowadays with all the technological advances it is important to take advantage of them and used these new tools to get closer to a solution to reduce the problem. Now the project will focus on answering the next question. We want to know if it is possible to know with accuracy the pronostic of air pollution in Mexico City using a Machine Learning model.

Artificial Intelligence allow us to develop machine learning models. Machine Learning allows a system to learn through data, there are two types of learning: supervised and unsupervised.
Supervised learning is when we have an established dataset and you have an idea of how to classify them. It has the intention of finding patterns in the data that could be applied to an analytical process.

Unsupervised learning is when the problem requires a massive quantity of data without any labeling, the comprehension of the meaning behind the data requires algorithms that classify the data based on the patterns or clusters it finds.

For a predictive model we need a lot of data and have variables: independent and dependent.

II. Development overview

Current Solution

The current solution to the problem is to develop a machine learning model that helps us predict with accuracy the air quality in Mexico City.

III. Data

Data Source

The data was collected from the government of Secretaría del Medio Ambiente (SEDEMA) of Mexico City. Our data is divided in two sets which are: Meteorology and pollutants. The data starts in 2018 and goes all the way to 2022.

IV. Development

Solution

As mentioned before our question is if we can know the accuracy of the air quality in Mexico City by using Machine Learning.
For us to be able to answer this question we had to select our variables which would be used in our model. The variables are divided into independent and dependent.

- ☐ Independent variables:
  - Pollutants: SO2, NO2, CO, O3, PM10, PM2.5

- ❖ Weather: Temperature, Humidity, Precipitation, Wind Velocity and Wind Direction.
- ❖ Others: Town Hall, Year, Month, Day of the week and Day

☐ Dependent variable:

- ❖ PM10

It is important to find a new way to work on the solution to our problem, because it affects our daily life and even our future. By allowing contamination levels that high we can get different kinds of diseases such as cancer, chronic lung disease and more. All of these diseases have a great chance to end in a fatal way, because of this it is crucial to try and reduce these types of risks.

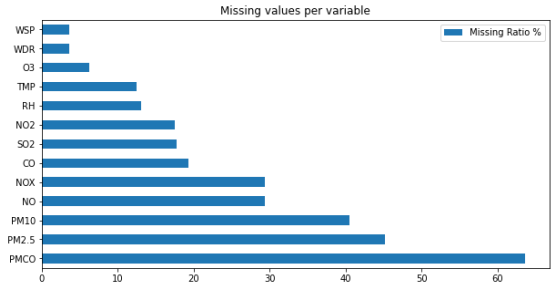
Not only affects our health but also the atmosphere and each day we see closer the effects of climate change making the Earth no longer a viable place to live.

Data Management

Once we found our data from a trustworthy source and have the necessary variables for the model, we download the CSV and unite them in one because originally they were separated by year and variable. Once we have our two different dataset, the first one is for the weather 's data where we can find temperature, wind, precipitation and more. The second one contains all of the pollutants such as PM10, SO2 and others. After this we download another CSV with the dictionary of all the registered stations to obtain their coordinates. Last but not least we put everything together in a single dataframe, this helped us to make the analysis easier.

Data Cleaning

When we started analyzing our data we could observe that there is a big quantity of null values as we can see in the next graphic.



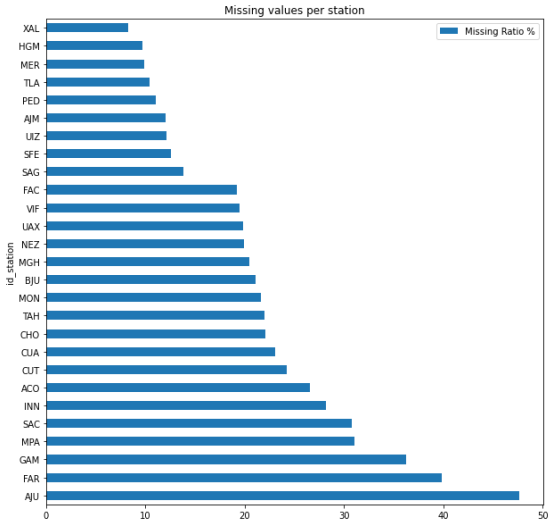
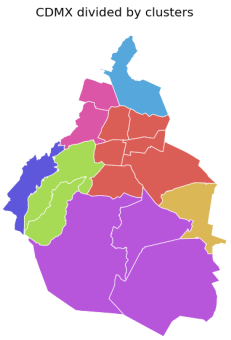


Fig 1: Null Values

However we can not throw away these values, this is why we apply 3 different methods to solve the problem.

1. Fill the null values with the previous value (12 hours before)
2. Put together by day, by doing these we can maintain the variance of our dependent variable that is PM10 and also maintain the behavior of contamination during the day.
3. Put together the geographic areas, we achieved this by using k-means. The algorithm made 16 clusters of the stations according to their geographic coordinates.



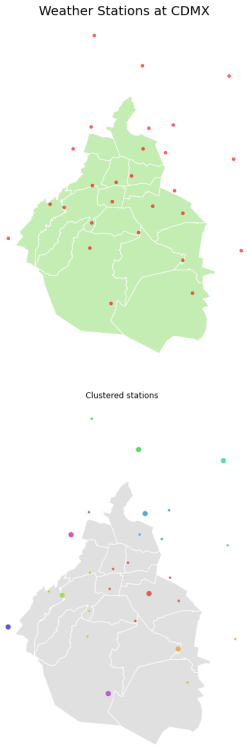
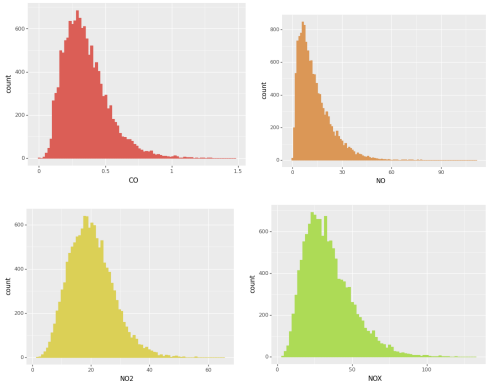


Fig 2: CDMX divided by clusters

When doing a Machine Learning model we need to make sure that all of our data is correct and clean so we do not have incorrect training. Because of this we decided to make different histograms to see the distribution of each variable and make sure there are no abnormal values, luckily the data was clean.



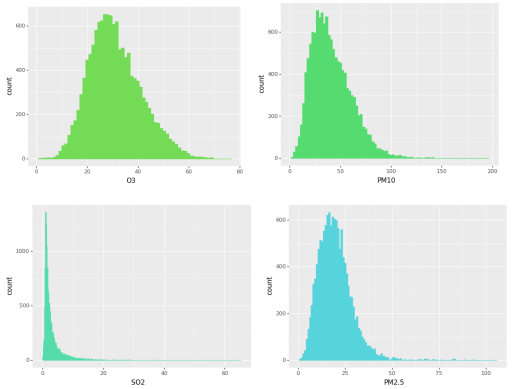


Fig 3: Histograms

Machine Learning Model

Before starting the machine learning model itself, we created a correlogram to observe which input variables are the adequate ones to give the model. Once those variables were defined, we began the training of the model.

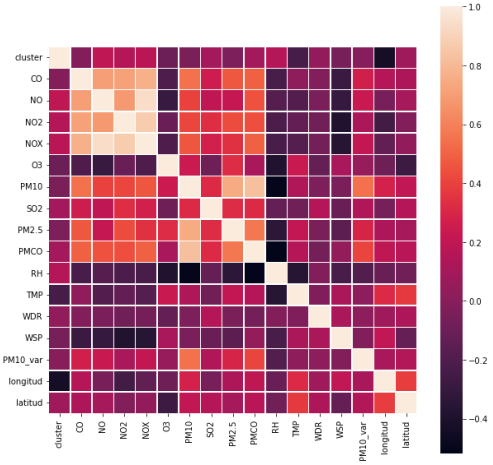


Fig 4: Correlogram

We decided to use time series in order to train the model, as well as a short-term memory neural network (LSTM) with the first 80% of the available days. As shown in Figure 5, the meteorological data of the seasons within the last 7 days were used as independent variables and, as the model was of the supervised kind, the wind speed of the hurricane was used as the dependent variable.

To ensure that the model worked properly, we tested it with the remaining 20% of the available days within the used data.

V. Results

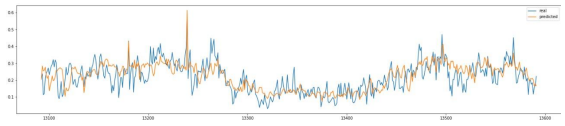


Fig 5: Forecast of Air Quality

As shown in Figure 5, the historical data is plotted with a blue line, whereas the model-predicted values are plotted with an orange line.
The image shows that the model forecasted the air quality with a fairly high accuracy.
Como se puede observar, el modelo predijo la calidad del aire con una eficacia bastante alta. The model has an MSE of 12.75

The designed model can be implemented within a productive environment. By aggregating the meteorological seasons in clusters, it is possible to determine if there is a significantly higher likelihood of poor air quality within Mexico City in a specific period of time.

VI. Conclusions

It is possible to conclude that the model was able to predict the forecast of air quality within Mexico City in those seasons which are, historically, keen to have poor air quality, doing so with a high accuracy.
The model can still be improved due to the fact that the procedure performed throughout this report was only a first view of what it can actually accomplish when fed with a large quantity of data which was previously prepared, cleaned and processed.

The information contained within this report concludes the work developed throughout the semester of the class Data Science. These type of predictions are highly important so as to have a positive impact on the environment and decrease the amount of casualties derived from air particles and poor air quality in general. It is our duty as students and soon-to-be professionals to apply the acquired knowledge in order to find and discover solutions to the problems which have a direct impact on society as a whole.

VII. Bibliographic references

4. IBM. (2022). ¿Qué es Machine Learning? 24 de Mayo del 2022, de IBM Sitio Web: https://www.ibm.com/mx-es/analytics/machine-learning
5. National Geographic. (2021). Ciudad de México alcanza niveles históricos de contaminación por partículas suspendidas. 24 de Mayo del 2022, de National Geographic Español Sitio web: https://www.ngenespanol.com/ecologia/ciudad-de-mexico-alcanza-niveles-historicos-de-contaminacion-por-particulas-suspendidas/
6. Aquae fundación. (2021). Contaminación del aire: causas y tipos. 24 de Mayo del 2022, de Aquae fundación Sitio web: https://www.fundacionaquae.org/wiki/causas-y-tipos-de-la-contaminacion-del-aire/