MiBiciudad limpia – AI.rbike

Un modelo de predicción para la calidad del aire basado en el uso de la red de bicicleta pública en ciudades

Autores
Domínguez Balderrama, Jacobo; Garcia Rojas, Raúl;
Tapia Muñoz, Norma Adriana; Uribe López, Kimberly Celeste
Mentor
Jimenez Lepe, Edwin Efrain

---

1. Eje rector del proyecto

Más allá de ser un simple objeto de recreación, las bicicletas están siendo consideradas como un modelo viable de transporte para la vida cotidiana de los habitantes de grandes ciudades. Las calles constantemente se ven pobladas por este medio de transporte y conforme el tiempo avanza y las ciudades crecen, se crean conversaciones para la formación de nuevas políticas entre la población, el gobierno y los sectores privados. Si bien el uso de bicicletas es un tema importante a discutir entorno a sus efectos en la economía local así como la movilidad, el uso de estas también tiene un gran impacto en términos de medio ambiente y salud. La calidad del aire en grandes ciudades es una preocupación que concierne a los habitantes y el gobierno. Muchas ciudades alrededor del mundo han implementado sistemas de bicicleta para promover formas de transporte mucho más sustentables y reemplazar los vehículos de combustibles fósiles. La razón principal es que las contingencias ambientales son la causa del 80% de las enfermedades respiratorias. Por esta razón, el siguiente proyecto plantea explorar el impacto que tiene la infraestructura de bicicleta pública en ciudades urbanas a través de modelos de Machine Learning. Se busca demostrar la importancia y el efecto a largo plazo que tiene el uso de bicicleta en la calidad de aire de acuerdo al uso de éstas por su población. Dicho estudio pretende crear una puerta a la creación de políticas públicas, donde a través de impactos medibles se pueda destinar presupuesto o generar mayor apoyo a las iniciativas de infraestructura para bicicleta pública.

2. Planteamiento del problema

En la última edición del ranking ciclociudades elaborado por el Instituto de Políticas para el Transporte y el Desarrollo México (ITDP, por sus siglas en inglés), Guadalajara logró el mejor puntaje entre 31 ciudades del país. Incluso, le quitó el puesto a la Ciudad de México. Actualmente, se cuenta con un servicio de transporte público llamado MIBICI; una red de estaciones automáticas donde se puede tomar y/o dejar una bicicleta mediante un sistema de membresía adquirido por el usuario. Por otro lado, el Instituto de Planeación y Gestión del Desarrollo del Área Metropolitana de Guadalajara ha puesto sobre la mesa un plan de crecimiento conocido en redes como #CiclovíasEmergentesAMG dentro del cuál se propone el crecimiento de la infraestructura de ciclovías en distintos puntos de la ciudad. Sin embargo, distintos habitantes de la zona siguen en discusión entre aquellos usuarios de bicicletas y los automovilistas. Mientras que los primeros creen firmemente en el cambio que conlleva utilizar este método de transporte además de reclamar sus derechos al espacio público, los segundos se oponen puesto que perciben que esto solo aumentaría los problemas de movilidad ya existentes y preocupantes que tiene la ciudad. Si bien las ciclovías parten desde la esfera política de la ciudad, los servicios de bici pública parten desde la esfera privada y la

pregunta es ¿cómo hacer decisiones de negocio basadas en un contexto tan polarizado; en donde la visión del proyecto no es suficiente para su implementación con éxito? ¿Qué pasaría si se deja de lado el tema de movilidad entorno a las bicicletas y se comienza a visualizar el impacto entorno al medio ambiente?

El sector transporte es uno de los principales emisores de contaminantes a escala nacional en México, siendo responsable de 90.03 por ciento de las emisiones de monóxido de carbono (CO) y de 45.67 por ciento de óxidos de nitrógeno (NOx), en todo el país. Dentro del sector transporte los vehículos de pasajeros —denominados ligeros— emiten 74.41 por ciento de CO, 52.55 de NOx, 73.55 de Compuestos Orgánicos Volátiles (COV) y 94.50 de amoniaco (NH_3). Estos contaminantes contribuyen a la formación de ozono en la atmósfera.

3. Descripción de la solución a la problemática detectada

Actualmente existen reportes entorno al efecto que se tiene en términos de movilidad y medio ambiente para regiones como la unión Europea; tales como " Cycling and Urban Air Quality: A study of European Experiences". Sin embargo, para regiones como Latinoamerica, esta línea de investigación aun es un área de investigación. Si bien existen los datos entorno a la calidad del aire de muchas regiones así como aquellos de la infraestructura de bicicleta pública; no existen reportes cuantitativos que exploren la relación entre estos dos aspectos. Por otro lado, los reportes que se pueden generar con estos datos están enfocados en probar y no en mejorar. Es por ello que en este proyecto se buscará plantear un modelo predictivo que permita ver el impacto que se tiene de los sistemas de bicicleta pública en el tiempo futuro. Este estudio podría ser utilizado por las esferas privadas y de gobierno encargadas de temas de movilidad para tomar decisiones entorno a cómo crecer la infraestructura de bicicleta basándose en una proyección del impacto en términos de calidad del aire. Si bien el proyecto estará enfocado en escalarse en Guadalajara, la solución descrita a continuación consta de tres fases. Las primera dos se enfocan en distintas ciudad; Chicago, Boston, Nueva York y Guadalajara. Para estas se realizan las siguientes fases:

1. Creación de modelos específicos de predicción por estación.
2. Generalizar comportamientos, resultados y patrones para la ciudad.

Una vez obtenidas las fases anteriores para cada ciudad se plantea la última fase:

3. Propuesta de un modelo general que pueda ser aplicable para cualquier ciudad

4. Hipótesis

Uno de los principales factores que afectan la calidad de aire en las ciudades es el uso de vehículos de combustibles fósiles. En este modelo se busca comprobar que el número de bicicletas y viajes realizados por estación de bicicleta pública tiene una relación inversamente proporcional al nivel registrado de contaminantes. Con esto se busca comprobar que el uso de la bicicleta tiene un gran impacto en la calidad de aire de una zona.

5. Metodología

a. Descripción de los datos

i. Bases de Datos

Los datos se obtuvieron de bases de datos de acceso público. Para las estaciones de bicicleta de las ciudades correspondientes se accedió a los sitios web de las compañías de bicicleta pública:

Ciudad	**Compañía**
Guadalajara	MiBici
Chicago	Divvy
Boston	Bluebikes
Nueva York	Citibike

Estos *datasets* estaban disponibles para su descarga directamente de su página o en la comunidad de Kaggle.

Los datos de la calidad del aire se obtuvieron de las siguientes fuentes:

Ciudad	**Fuente**
Guadalajara	SEMADET
Nueva York	OpenAQ
Boston & Chicago	Air Quality System (AQS) API

ii. Preprocesamiento

Los siguientes puntos fueron considerados para hacer la limpieza, adecuación y construcción de la base de datos final para cada ciudad:

1. **Distancia de la estación de medición**: Se consideraron solamente los datos de las estaciones de bicicleta más cercanas a los puntos de medición de calidad de aire para cada ciudad. El rango óptimo entre la estación de bicicletas y la estación de medición es de 8km, sin embargo en algunas ciudades no se tenían rangos así de accesibles por lo que solo se consideraron aquellas más cercanas.
2. **Contaminantes**: Aun cuando se tienen medidas atmosféricas de la concentración de contaminantes, no todos los datos son los mismos para todas las ciudades. Por esto se consideraron solo las medidas de ozono y monóxido de carbono (CO) (debido a que el 90% de la presencia de este contaminante se debe al transporte).
3. **Fechas de muestras**: Otra situación a considerar era el rango de fechas a tomar para hacer el modelo. En muchos casos las limitantes encontradas fue el tener fechas de antes y después de la implementación de la infraestructura de bicicleta pública.

Para poder homologar la información obtenida se armo un dataset basado en la información en común así como aquella considerada pertinente para poder utilizar el modelo. El dataset correspondiente para cada ciudad quedaría armado de la siguiente manera:

Cantidad de viajes	Distancia recorrida (sum)	Duracion del viaje (sum)	Hora	Estacion Inicio	Fecha	Métrica Contaminante

b. Descripción del modelo a utilizar

Los modelos a entrenar serían: Regresor Logístico, Árbol de decisión y un Perceptrón Multicapa como red neuronal. Estos tres modelos se compararían entre sí utilizando su coheficiente de determinación; dato estadístico que determina la calidad del modelo para replicar los resultados, y la proporción de variación de los resultados que puede explicarse por el modelo y el root mean square error el cual mide la distancia promedio entre los puntos reales y aquellos dados por la predicción, medidos en una línea vertical.

c. Delimitaciones
   i. De datos
      La principal problemática que se encontró en torno a los datos era entorno a la cantidad de mediciones que se pueden encontrar. Si bien existían bases de datos muy contundentes entorno a las estaciones de bicicleta así como para la medición de calidad del aire, era importante corroborar que se estuvieran utilizando los mismos contaminantes para todos y en ocasiones esto era complicado debido a la disponibilidad. Por otro lado la distancia entre la medición de calidad del aire con las estaciones de bicicleta era importante y en ocasiones esta distancia era muy grande.
   ii. Temporales: tiempo de la limpieza y obtención de base de datos. Tiempo de Durante este proyecto se pudo comprobar que el 80% del tiempo invertido en un proyecto de Ciencias de Datos y Machine Learning se utiliza para la limpieza de datos. Es importante familiarizarse con el lenguaje así como todas las herramientas que este puede dar para hacer esta limpieza de manera óptima. Por otro lado, el tiempo de procesamiento de grandes cantidades de datos puede ser considerablemente tiempo a considerar.
   iii. Teóricas: Investigación de la problemática desde el punto de vista científico y no Una de los mayores aspectos a considerar pata futuros proyectos y no solo de ésta índole o área es la integración de equipos multidisciplinarios entorno al área de ciencias. Si bien el enfoque desde la ingeniería pudiera ser pragmático; hay muchos aspectos relevantes a considerar que se pierden o se concluyen durante el proceso. En el caso específico de éste proyecto, contar con expertise del área medio ambiental durante todas las etapas del proyecto sería importante, mandatorio y de utilidad para poder generar conclusiones más asertadas y poder ir guiando el camino del proyecto entorno a esta línea de investigación.

6. Marco Teórico

Se hizo una búsqueda del estado del arte entorno a proyectos en donde se considerara el uso de la bicicleta como un factor para predecir la calidad del aire. Si bien existen reportes para europa que hacen estudios históricos de esto, no existen estudios que puedan permitir la mejora basada en predicciones. El único estudio que se encontró respecto a este tema es uno llamado *The Impact of Bike-Sharing Ridership on Air Quality: A Scalable Data Science*

Framework. Esta investigación explora la relación entre el indicador de calidad de aire al día (AQI) y la intensidad diaria de uso de bicicleta en Nueva York. Los autores designaron y utilizaron los modelos de Elastic Net, Random Forest Regression, and Gradient Boosted Regression Trees. La información que se encontró respecto a este modelo fue que las características utilizadas fueron: indicador de estación del año, medición AQI del día previo y el número total de bicicletas que se utilizan diariamente del 2016 al 2018.
Para poder hacer una comparativa numérica de los modelos era necesario escalar los valores de RMSE. Sin embargo se pudieron hacer comparaciones desde el enfoque conceptual. Este estudio utiliza como características la calidad del aire del día pasado así como la estación. En comparación con el estudio propuesto en este entorno, el uso de estas características pudiera dar referencia al modelo del valor esperado, al dar una importancia a los valores pasados. Por otro lado, se utilizan valores por día. El estudio propuesto en este informa busca la utilización solamente de las bicicletas para comprobar su relación directa, por lo que se propone un modelo de predicciones más finas.

7. Resultados

Boston
Para la ciudad de Boston, cada uno de estos modelos fue aplicado a seis diferentes muestras: cinco estaciones de forma individual y un modelo tomando todas las estaciones de la ciudad. Esto se hizo con la finalidad de comparar modelos entre sí y saber si existía una diferencia entre tomar muestras más pequeñas y un modelo generalizado. Los datos que se utilizaron para los modelos comprenden el año 2019, y son por hora. Una particularidad de los datos de calidad de aire de la ciudad de Boston es que existía un solo lugar donde se hacía le medición de los contaminantes, y muchas estaciones se encontraban bastante alejadas de la toma (más de 10km); el problema que esto genera es que no se puede garantizar calidad de los datos al estar tan alejada la estación de la ubicación de la toma de contaminantes. Otra incoveniente fue que varias de las estaciones que se encontraban cerca de la ubicación de la toma de contaminantes, tenían muy pocos datos de viajes. Estas dos limitantes fueron consideradas para la elección de las estaciones, siendo las seleccionadas aquellas que contaran con la mayor cantidad de viajes que comenzaran en dicha estación, así como su cercanía a la toma de contaminantes (menos de 2km). Al inicio se tomó para el análisis solo la medición de monóxido de carbono (CO); posteriormente se incluyó en ozono (O3). Los resultados obtenidos midiendo el monóxido de carbono no dieron buenos resultados para ninguno de los modelos, ya que el coeficiente de determinación era muy bajo (cercano a cero, o incluso negativo), mientras que por el contrario, el root mean square error era muy grande. Influenciado por los resultados de los modelos de la ciudad de Nueva York donde se medía el ozono y se obtenían mejores resultados en los modelos, se decidió replicar los tres modelos de ML pero analizando el ozono y así comprobar si el ozono era una mejor métrica que el monóxido de carbono. Una vez realizados los modelos y haciendo una comparación se obtuvo como conclusión que para Linear Regression y Decision Tree, tomar el ozono arrojaba mejores resultados que al hacer el análisis sobre el monóxido. El caso donde se obtuvo una mejoría significativa fue para el modelo de linear regression de la Estación 39, donde el MSRE disminuyó de 0.3083 a 0.1058; la mayor diferencia fue en el coeficiente de determinación, que pasó de 0.008 con el monóxido hasta un 0.1899 para el ozono.

	CO				O3	
Linear Regression	Root Mean Square Error	Coefficient of determination			Root Mean Square Error	Coefficient of determination
CIUDAD	0.092	0.0059			0.012	0.0813
Station 12	0.1032	0.0127			0.0122	0.0034
Station 39	**0.3083828789**	**0.008**	**0.55532232**		**0.1058300524**	**0.1899**
Station 51	0.1121	-0.0051			0.0124	0.0233
Station 200	0.119	0.003			0.0124	0.0241
Station 364	0.1033	0.0051			0.0123	-0.0016

Si bien el modelo no es bueno para predecir los datos, es una mejora importante. El modelo de la ciudad lamenteblemente no tuvo una mejora tan importante como el de la estación 39. Una posible razón es que se incluye todas las estaciones del sistema Bluebikes, incluso las que están más alejadas de la toma, lo que podría generar ruido y dificultad para que los modelos aprendan y generen una buena predicción del contaminante (tanto para CO como para O3). Los resultados de todos los modelos para todas las estaciones y para la ciudad pueden encontrarse en los anexos de este documento.

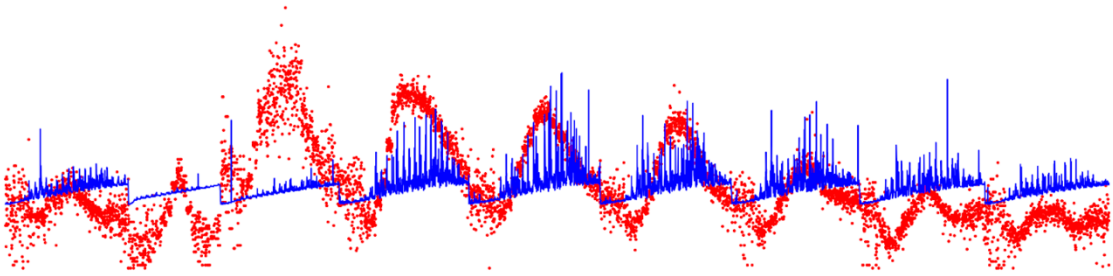
Nueva York

Los resultados de la experimentación para los datos de Nueva York fueron satisfactorios, a pesar de no haber datos suficientes del contaminante que se consideró para la experimentación inicial (CO), el Ozono tuvo un buen desempeño.

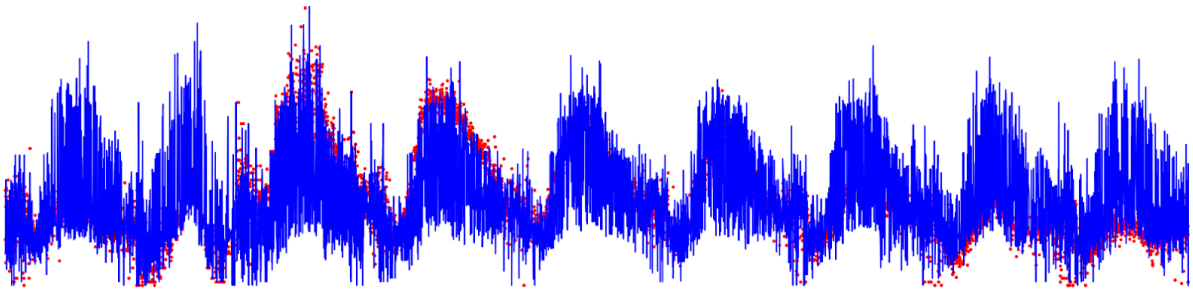
A nivel ciudad de los tres algoritmos usados para entrenar el modelo el que resultó con un mejor desempeño fue MLPRegressor teniendo como valor de RMSE 0.010 y el coeficiente de determinación 0.275, mientras que el DecisionTreeRegressor tuvo un RMSE de 0.013 y un coeficiente de -0.212, a su vez el desempeño del algoritmo Linear Regression fue RMSE con un valor de 0.012 y 0.085 como coeficiente de determinación.

Para entender mejor los resultados se realizaron unas gráficas comparando la predicción (en azul) contra el valor esperado (en rojo), pudiendo así visualizar qué Linear Regresion tiene cierta tendencia a acercarse al valor esperado pero no lográndolo con tanto éxito.

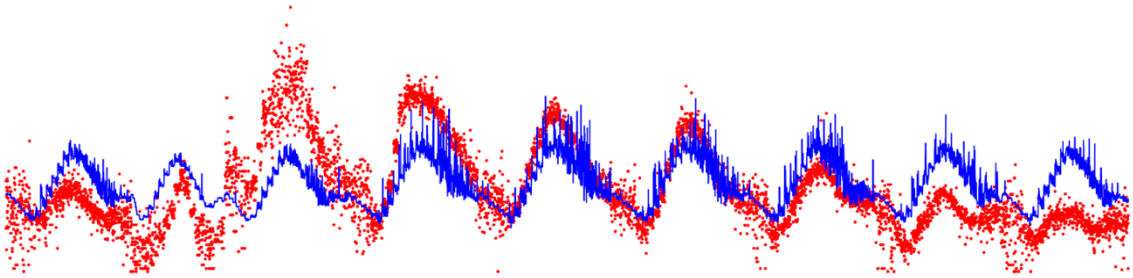
Para el caso del DecisionTreeRegressor podemos ver un poco sobreajustado el modelo casi llegando a parecer con overfiting, mientras que la gráfica del MLPRegressor se pude visualizar como los resultados de la predicción se acercan mucho al valor esperado.



Gráfica 1Linear Regression.



Gráfica 2DecisionTreeRegressor



Gráfica 3 MLPRegressor

La experimentación a nivel estaciones tuvo resultados variados. En algunos casos ciertos algoritmos resultaron mejor que otros viendose afectado MLPRegressor al tener predicciones un poco erróneas la mayoría de las veces, tal vez afectado por la insuficiencia de datos para operar adecuadamente. Mientras los otros dos algoritmos tuvieron resultados similares no pudiendo etiquetar como uno mejor que otro. Los resultados de todos los modelos para todas las estaciones y para la ciudad pueden encontrarse en los anexos de este documento.

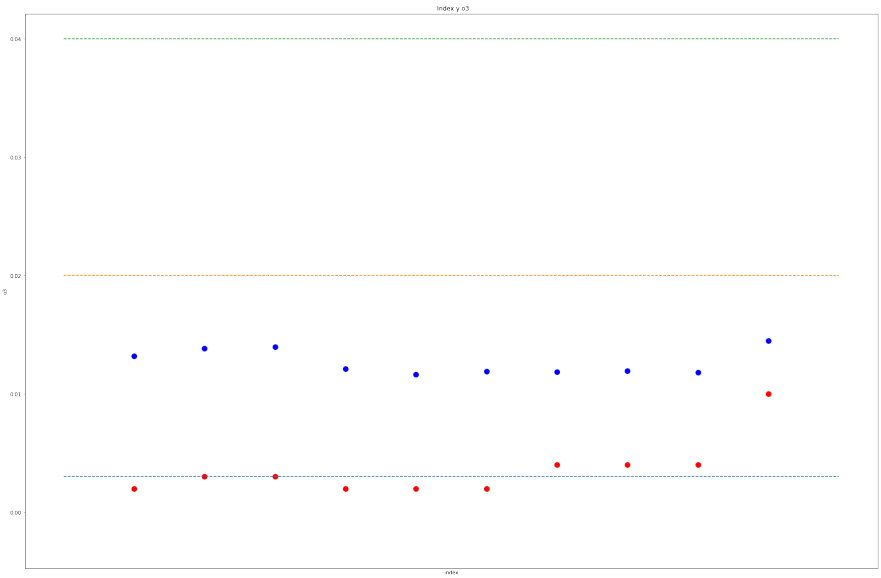
Guadalajara

El dataset está compuesto por todos los viajes del sistema mibici durante 2019 y las mediciones del sistema de calidad del aire de la SEMADET, estas mediciones se realizan cada hora los 365 dias de la año, el sistema consta de 10 estaciones de mediciones

los datos de viajes de mibici se agregaron por hora considerando el número de viajes, la distancia recorrida y el tiempo de los viajes, cada viaje se le relaciono la estación de inicio y fin con la estación de medición de aire más cercana los datos de calidad del aire incluían las siguientes medidas : CO, NO, NO2, NOX, O3, SO2, PM 2.5 y PM10

Los modelos se aplicaron sobre el conjunto de datos de todo el sistema mibici, también se ejecutaron con las 5 estaciones que contaban con más cantidad de viajes y que estuvieran dentro de un radio de 2.5 km de una estación de medición del aire. Los resultados obtenidos con el monóxido de carbono tuvieron una Mean Square Error mayor a lo esperado. Los mismos modelos aplicados a los datos de la ciudad de NY en donde solo se tenían datos de Ozono arrojaron mejores resultados, esto nos llevó a intentar de nuevo los mismos 3 modelos ahora con esta métrica. Los resultados mejoraron notoriamente, siendo el modelo MLPRegressor con el dataset de toda la ciudad el que obtuvo mejores resultados, este mismo modelo con datasets por estación resulta en resultados muy dispares, se presume que la cantidad de datos por estación no es suficiente para entrenar la red

los otros dos modelos regresión lineal, árbol de decisión también mostraron mejoría aunque no al grado que MLPRegressor



Gráfica 4 Mediciones de ozono real y predicción MLP Regressor

8. Conclusiones finales y siguientes pasos

De acuerdo a representantes del Centro Mexicano de Derecho Ambiental AC (Cemda)

> *El sector transporte es uno de los principales emisores de contaminantes a escala nacional en México, siendo responsable de 90.03 por ciento de las emisiones de monóxido de carbono (CO) y de 45.67 por ciento de óxidos de nitrógeno (NOx), en todo el país. Dentro del sector transporte los vehículos de pasajeros —denominados ligeros— emiten 74.41 por ciento de CO, 52.55 de NOx, 73.55 de Compuestos Orgánicos Volátiles (COV) y 94.50 de amoniaco (NH3). Estos contaminantes contribuyen a la formación de ozono en la atmósfera.*

Esta información fue de gran utilidad para poder interpretar los resultados. Si bien los modelos para predicción no eran tan buenos, aquellos de ozono sí lo eran. Esto se debe a que este contaminante es una reacción de distintos factores y está compuesto por otro contaminantes. Si bien a pequeña escala no era tan relevante el uso de la bicicleta; cuando se compara con este contaminante sí prueba una relación entre estos dos factores. Por otro lado, el uso del ozono como métrica para este estudio es pertinente debido a que en los últimos 5 años de 20 contingencias ambientales encontradas en la Ciudad de México, 17 de ellas fueron causadas por el ozono.

Para Guadalajara, los modelos por estación tenía diferencias mínimas entre sí de valor de Mean Square Error y el Coeficiente de determinación. La máxima distancia entre la medición del contaminante era de 4km. Para este modelo era conveniente utilizar los datos de toda la ciudad. Para Boston, había estaciones con muy pocos datos, por lo que esto afectaba el resultado del modelo por ciudad. Por otro lado, había estaciones incluso a 29 km de distancia

de la locación de medición del contaminante. Esto da pie a una de las conclusiones más importantes para este entorno de trabajo. Es importante contar con datos cuyas mediciones sean cercanas a las estaciones de bicicleta, ya que de lo contrario, el modelo de predicción no tendrá tanta relevancia. Así bien, la cantidad de mediciones es importante.
En este trabajo solo se exploró la relación directa de la bicicleta con el contaminante. Este enfoque fue viable para este estudio ya que se quería conocer su relación directa, sin embargo para trabajos futuros es necesario incoporar otra información relevante tales como temperatura o incluso un estudio más a fondo de los contaminantes y qué factores los afectan. Esto puede ser añadido al modelo de manera de etiqueta, no de característica para evitar el sesgo en el estudio.

Este estudio puede tomarse a consideración para comenzar a ver la bicicleta no solo como una herramienta de movilidad; si no como una puerta a una mejor calidad del aire para zonas urbanas. Estudios como este pueden ser utilizados por:

- La Esfera Privada: para poder crear planes para la infraestructura de sistemas de bicicleta pública y hacer planes de crecimiento urbanos de acuerdo a las necesidades ambientales de la zona.
- Gobierno: utilizar modelos de proyección en el futuro para poder crear planes de crecimiento y destinar fondos a la infraestructura de bici basándose en resultados proyectados en el futuro en planes a corto, mediano y largo plazo.
- Ciudadanos: Comenzar a visualizar el impacto del uso de bici no solo como un medio de transporte para solo algunas personas, si no como una mejora para la zona en donde se habita visualizando las proyecciones a futuro de la calidad del aire. Por otro lado para los usuarios de bicicleta, tener una manera de medir y visualizar su impacto.

9. **Bibliografía**

Woodbridge, N. Hua, V. Suarez, R. Reilly, P. Trinh and P. Intrevado, "The Impact of Bike-Sharing Ridership on Air Quality: A Scalable Data Science Framework," 2019 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI), Leicester, United Kingdom, 2019, pp. 1950-1957, doi: 10.1109/SmartWorld-UIC-ATC-SCALCOM-IOP-SCI.2019.00341.D

Navarro, Israel. “Autos, Primera Fuente de Contaminación En El País.” Milenio, www.milenio.com/estados/autos-primera-fuente-de-contaminacion-en-el-pais. Accessed 4 July 2020.

Repository.usfca.edu. 2020. [online] Available at: <https://repository.usfca.edu/cgi/viewcontent.cgi?article=1132&context=artsci_stu> [Accessed 4 July 2020].