

Comparativa de diferentes técnicas de Minería de Datos para la predicción del uso de la bicicleta de acuerdo con las condiciones climáticas y estacionales en Washington

Comparison of different Data Mining techniques for predicting bicycle use according to climatic and seasonal conditions in Washington

Ana Rosa Velázquez Cordero ^{a,*}, Francisco-Jacob Ávila-Camacho ^b

^a Maestría en Ingeniería en Sistemas Computacionales, Tecnológico de Estudios Superiores de Ecatepec, 55210, Ecatepec, Estado de México, México.

^b División de Ingeniería en Sistemas Computacionales, Tecnológico Nacional de México / TES Ecatepec, 55210, Ecatepec, Estado de México, México

Resumen

En este artículo se aborda la importancia de los sistemas de bicicletas compartidas como una evolución significativa en el paradigma del alquiler de bicicletas, destacando su impacto positivo en el tráfico, el medio ambiente y la salud, así como la extensa red de programas de bicicletas compartidas a nivel mundial junto con la valiosa cantidad de datos generados por estos sistemas, en especial en las estaciones de Washington. Se realizó un análisis de datos histórico del sistema Capital Bikeshare en Washington D.C. durante los años 2011 y 2012, lo que proporciona una base sólida para capturar variaciones estacionales y tendencias en los patrones de alquiler de bicicletas y con ello realizar la formulación de modelos predictivos precisos. Finalmente se destaca la importancia de los sistemas de bicicletas compartidas, la disponibilidad de datos y la relación entre el proceso de alquiler de bicicletas, así como diversos factores ambientales para comprender y predecir la demanda de bicicletas en entornos urbanos, para el caso de la predicción se compararon 2 técnicas de minería de datos las cuales son “Árboles de regresión y Regresión Simple” con el histórico y elegir la que presentó la menor desviación.

Palabras Clave: Predicción, Árboles de Regresión, redes neuronales, regresión, error cuadrático medio (MSE) y machine learning.

Abstract

This article addresses the importance of bike-sharing systems as a significant evolution in the bike rental paradigm, highlighting their positive impact on traffic, the environment, and health, as well as the extensive network of bike-sharing programs globally along with the valuable amount of data generated by these systems, especially at the Washington stations. A historical data analysis of the Capital Bikeshare system in Washington D.C. was performed during the years 2011 and 2012, which provides a solid basis for capturing seasonal variations and trends in bicycle rental patterns and thereby formulating accurate predictive models. Finally, the importance of bicycle sharing systems, the availability of data and the relationship between the bicycle rental process, as well as various environmental factors to understand and predict the demand for bicycles in urban environments, in the case of prediction, are highlighted. Two data mining techniques were compared, which are “regression trees and Simple Regression” with the historical data and the one that presented the least deviation was chosen.

Keywords: Prediction, Regression Trees, Neural networks, regression, mean square error (MSE) and machine learning.

1. Introducción

*Autor para la correspondencia: 202222@tese.edu.mx

Correo electrónico: 202222@tese.edu.mx (Ana Rosa Velázquez Cordero), fjacobavila@tese.edu.mx (Francisco Jacob Ávila Camacho).

Los sistemas de bicicletas compartidas representan una evolución significativa en el paradigma de alquiler de bicicletas tradicional. Estos sistemas automatizan todo el proceso, desde la membresía hasta el alquiler y la devolución, facilitando a los usuarios la posibilidad de alquilar una bicicleta desde una ubicación específica y devolverla en otra. (Shaheen et al., 2013). La singularidad de estos sistemas radica en que se registra la duración del viaje y las posiciones de salida y llegada. (Rosales-Asensio et al., 2019). A nivel mundial, existen alrededor de 500 programas de bicicletas compartidas, con más de medio millón de bicicletas en circulación (Eren & Uz, 2020). Esta extensa red de alquiler de bicicletas no solo proporciona una alternativa de transporte eficiente, sino que también genera grandes cantidades de datos que resultan valiosos para este artículo.

La importancia de los sistemas de bicicletas compartidas se destaca en su impacto positivo en cuestiones cruciales como el tráfico, el medio ambiente y la salud pública. Al fomentar el uso de la bicicleta como medio de transporte, estos sistemas contribuyen a la reducción de la congestión vehicular en las ciudades, disminuyendo las emisiones de gases contaminantes (Gong et al., 2024). Además, promueven un estilo de vida activo y saludable al integrar el ejercicio físico en las rutinas diarias de los usuarios (Zhou et al., 2022). Desde una perspectiva medioambiental, la reducción de la dependencia de los vehículos motorizados también tiene implicaciones positivas para la calidad del aire y la sostenibilidad urbana (Zhou et al., 2022).

A diferencia de otros servicios de transporte, la información detallada registrada en cada viaje, como la duración, la posición de salida y llegada, se convierte en un conjunto de datos robusto que puede considerarse una red de sensores en sí misma. Esta red virtual ofrece una oportunidad única para monitorear la movilidad en la ciudad en tiempo real (Zheng & Li, 2020). El análisis de estos datos puede revelar patrones de comportamiento, preferencias de rutas y cambios en la demanda de bicicletas, lo que tiene implicaciones importantes para la planificación urbana y la optimización de la infraestructura de transporte. En este sentido, la convergencia de la movilidad urbana y la generación de datos en tiempo real abre nuevas posibilidades para abordar desafíos contemporáneos en el diseño de ciudades inteligentes y sostenibles (Guo et al., 2022), (Ricci, 2015).

1.1. Características de los Datos Generados por los Sistemas

Los sistemas de bicicletas compartidas se distinguen notablemente de otros servicios de transporte en la cantidad y tipo de datos recopilados (García-Gutiérrez et al., 2014). A diferencia de servicios como autobuses o metro, donde los datos suelen limitarse a la cantidad de pasajeros o las horas de operación, los sistemas de bicicletas compartidas registran explícitamente detalles específicos de cada viaje. Esta característica única proporciona un conjunto de datos detallado y completo que incluye la duración del viaje, así como la posición exacta de salida y llegada de cada bicicleta alquilada. Mientras que otros servicios de transporte suelen generar datos agregados, los sistemas de bicicletas compartidas ofrecen una granularidad sin precedentes en la información, permitiendo

un análisis más preciso de los patrones de movilidad urbana (Eren & Uz, 2020).

El registro explícito de la duración del viaje, la posición de salida y llegada en los sistemas de bicicletas compartidas no solo representa una ventaja en términos de cantidad de datos, sino que también ofrece una ventana directa a la dinámica de la movilidad urbana. La duración del viaje permite no solo evaluar la eficiencia del sistema en términos de tiempo de desplazamiento, sino también comprender patrones de uso y preferencias de los usuarios. La posición de salida y llegada de cada bicicleta brinda información detallada sobre las rutas más frecuentes, puntos de interés y la distribución geográfica de la demanda (Gómez-Pérez et al., 2020). Estos datos específicos son esenciales para diseñar estrategias efectivas de ubicación de estaciones, mejorar la infraestructura de ciclovías y responder de manera ágil a las fluctuaciones en la demanda de bicicletas en diferentes áreas de la ciudad.

La diferencia en la cantidad y tipo de datos recopilados por los sistemas de bicicletas compartidas resalta su potencial como una valiosa red de sensores virtuales. Esta riqueza de información no solo beneficia a los usuarios en términos de comodidad y accesibilidad, sino que también abre nuevas oportunidades para investigadores y planificadores urbanos (Fin De Máster et al., 2019).

La capacidad de monitorear y analizar detalladamente la movilidad en tiempo real a través de estos datos posiciona a los sistemas de bicicletas compartidas como herramientas fundamentales para comprender la dinámica urbana y mejorar la eficiencia de los servicios de transporte sostenible en nuestras ciudades (Shaheen et al., 2013).

2. Conjunto de Datos del Sistema Capital Bikeshare

El conjunto de datos histórico abarca un periodo significativo de dos años, comprendidos entre 2011 y 2012, del sistema Capital Bikeshare en Washington D.C. son los siguientes:

Fecha	Estación	Clima	Temp	Humedad	Reg Usua	Cnt Bici
2011-01-01	1	6	0.3441	0.8058	654	985
2011-01-02	1	0	0.3634	0.6960	670	801
2011-01-03	1	1	0.1963	0.4372	1229	1349
2011-01-04	1	1	0.2	0.5904	1454	1562

Tabla 1 Datos recolectados del sistema Capital Bikeshare 2011 y 2012

En la tabla anterior se describen las variables que se van a utilizar las cuales son:

- Fecha
- Estación: Corresponde a la temporada del año (1 Primavera, 2 Verano, 3 Otoño y 4 Invierno)
- Clima: 1 Despejado, 2 Niebla, 3 Nieve y 4 Lluvia
- Temperatura
- Humedad
- Req. Usua: Cantidad de usuarios registrados

- Cnt. Bici: Recuento total de bicicletas de alquiler

De acuerdo con las variables descritas en el texto anterior, la variable dependiente que se utilizará para predecir será la columna “Cuento de bicicletas” que corresponde al total de bicicletas requeridas, la cual dependerá del resto de las variables de la tabla anterior.

La extensión de dos años permite un análisis completo de cómo las condiciones climáticas, la estacionalidad y otros factores pueden afectar los patrones de alquiler de bicicletas. Además, la inclusión de información detallada de estos años específicos permite contextualizar los resultados en un marco temporal específico, considerando posibles eventos externos o cambios en la infraestructura que podrían influir en los datos (Capital Bikeshare DC, 2020.; Fin De Máster et al., 2023.).

La disponibilidad de datos en el portal de <http://capitalbikeshare.com/system-data> agrega un componente crucial para determinar que técnica de minería de datos permite realizar una mejor predicción. La accesibilidad a la información facilita la replicación de estudios, la validación de resultados y el fomento de la transparencia en la investigación de este artículo. La disponibilidad en línea no solo beneficia a los académicos y personas que analizan los datos, sino que también puede ser utilizada por responsables políticos, urbanistas y cualquier persona interesada en comprender la movilidad urbana. Este enfoque abierto contribuye a la democratización del conocimiento y a la colaboración en la búsqueda de soluciones sostenibles y eficientes para los desafíos urbanos (Ricci, 2015b).

La frecuencia de datos, registrada diariamente, aporta gran detalle temporal al conjunto de datos. Esta alta frecuencia de muestreo permite capturar variaciones diarias en los patrones de alquiler de bicicletas. El análisis a nivel de día ofrece una visión más granular de cómo factores como la fecha, temporada del año, mes y condiciones meteorológicas pueden influir en la demanda de bicicletas y la evaluación de tendencias a lo largo de jornadas completas. Esta riqueza temporal es esencial para comprender la dinámica de uso de las bicicletas compartidas y representa una ventaja significativa en la formulación de modelos predictivos precisos (Vogel et al., 2011).

La relación entre el proceso de alquiler de bicicletas y diversos factores ambientales es un aspecto crucial para comprender los patrones de movilidad en los sistemas de bicicletas compartidas. Las condiciones climáticas ejercen una influencia significativa, ya que los usuarios pueden ser más propensos a utilizar bicicletas en días con climas agradables. La presencia de precipitación, por otro lado, puede disminuir la demanda, ya que muchas personas prefieren evitar andar en bicicleta bajo la lluvia. Además, el día de la semana y la temporada también juegan un papel crucial. Por ejemplo, los patrones de alquiler pueden variar en días laborables con respecto a los fines de semana, así como en verano e invierno. La hora del día añade otra capa de complejidad, ya que los comportamientos de alquiler pueden diferir significativamente entre las horas pico y las horas no pico (Fin De Máster et al., 2020).

El impacto de las condiciones climáticas, la precipitación, el día de la semana, la temporada y la hora del día en los comportamientos de alquiler de bicicletas destaca la interconexión entre la movilidad urbana y el entorno. Las preferencias de los usuarios están estrechamente vinculadas a las condiciones ambientales y estacionales. Por ejemplo, la primavera puede ver un aumento en la demanda de bicicletas debido al clima agradable, mientras que el invierno puede influir en la disminución de los alquileres. La variabilidad en la demanda a lo largo de las horas del día también puede relacionarse con patrones de viaje diarios y la conveniencia percibida del uso de bicicletas en momentos específicos. Este análisis detallado de los factores ambientales proporciona una base sólida para el diseño de modelos predictivos precisos que tengan en cuenta la complejidad de las influencias externas en la movilidad urbana (Alvarez-Valdes et al., 2016).

3. Estado del Arte

La revisión de artículos previos sobre la predicción del alquiler de bicicletas proporciona una perspectiva valiosa sobre los enfoques y métodos utilizados en investigaciones similares. En ellos se han abordado esta temática desde diversas disciplinas, incluyendo la ciencia de datos, el machine learning y la estadística. Se han empleado métodos predictivos tradicionales, como regresiones lineales y series temporales, para entender la relación entre las variables ambientales y estacionales y la demanda de bicicletas (Wang et al., 2019). Asimismo, se ha observado un creciente interés en la aplicación de modelos de machine learning, como regresiones polinómicas, árboles de regresión y redes neuronales, para capturar la complejidad de los patrones de alquiler. Estas investigaciones previas han arrojado luz sobre la importancia de considerar múltiples variables y la necesidad de modelos flexibles y adaptables para la predicción precisa.

Los métodos y enfoques utilizados en investigaciones similares han evolucionado con la expansión de la disponibilidad de datos y el desarrollo de técnicas avanzadas de machine learning. Algunos estudios han incorporado técnicas de agrupamiento para identificar patrones de comportamiento de usuarios similares, mientras que otros han explorado la aplicación de modelos de aprendizaje profundo para capturar relaciones no lineales en los datos (Zheng & Li, 2020). La inclusión de variables interactivas y la consideración de la multicolinealidad entre los factores ambientales han sido áreas de mejora identificadas en la literatura (García-Gutiérrez et al., 2014). Además, la selección adecuada de características y la evaluación robusta del rendimiento del modelo son aspectos críticos abordados por investigadores previos. La revisión de estos enfoques proporciona un marco contextual esencial para la formulación de un modelo efectivo en el contexto específico de predicción del alquiler de bicicletas según condiciones ambientales y estacionales.

4. Metodologías de Predicción

En la selección de algoritmos de inteligencia artificial y minería de datos para la predicción del alquiler de bicicletas, se deben considerar cuidadosamente las características del conjunto de datos y los objetivos específicos de la investigación (Implementación de Modelo Supervisado de Aprendizaje de Máquinas Para La Predicción de Alquiler de Bicicletas. | by Javier Ochoa | Medium, 2020.). Algunos de los algoritmos comúnmente utilizados son:

- Regresión lineal
- Regresión polinómica
- Árboles de Regresión
- Redes neuronales

La elección del algoritmo adecuado dependerá de la naturaleza de las relaciones entre las variables y la complejidad del problema. Los modelos de machine learning (Thirumalai & Koppuravuri, 2018.), al permitir la identificación de patrones no lineales, pueden ser particularmente efectivos para capturar la variabilidad en los comportamientos de alquiler asociados con las condiciones ambientales y estacionales.

En este caso se van a elegir 2 técnicas de predicción de minería de datos las cuales son: Árboles de Regresión y la técnica de Regresión Simple.

Los cálculos y modelos matemáticos que se dan entre las técnicas de Árboles de Regresión y Regresión Simple (Series de Tiempo) consideran las características de los datos y la naturaleza del problema de predicción de la demanda de alquiler de bicicletas, debido a que los Árboles de Regresión (Villafañez et al., 2022.) son modelos de aprendizaje supervisado que dividen repetidamente el conjunto de datos en subconjuntos más pequeños, basándose en las características de las variables predictoras, y luego aplican regresión en cada subconjunto para realizar predicciones. Matemáticamente, estos modelos se basan en algoritmos de partición recursiva como es el algoritmo CART (Árboles de Clasificación y Regresión). Por otro lado, la Regresión Simple en Series de Tiempo (Villafañez et al., 2022.) se basa en la relación entre una variable dependiente y una o más variables independientes en una serie temporal. En este enfoque, se utiliza el histórico de la variable dependiente (en este caso, el alquiler de bicicletas) y otras variables independientes (como las condiciones climáticas) para predecir valores futuros. Matemáticamente, la regresión simple se basa en el método de mínimos cuadrados para encontrar la línea de mejor ajuste a los datos observados.

Técnica Minería Datos	Características
Árboles de Regresión	Modelos predictivos formados por reglas binarias que segmentan el espacio de predictores. Fácil interpretación y manejo de interacciones entre variables. No requieren cumplir distribuciones específicas.
Regresión Simple	Basada en la relación entre una variable dependiente y

Técnica Minería Datos	Características
	variables independientes en una serie temporal. Utiliza histórico de la variable dependiente para predecir valores futuros.

Tabla 2 Características de la técnica árboles de regresión y regresión simple

Para realizar el comparativo de estas 2 técnicas se va a utilizar Python 3.11, junto con las librerías numpy, pandas, matplotlib y seaborn.

Los árboles de regresión son una herramienta fundamental en el campo de la minería de datos para predecir valores numéricos basados en un conjunto de variables predictoras. Funcionan mediante la segmentación progresiva de los datos en subconjuntos más pequeños, donde se ajustan modelos de regresión en cada subdivisión. Este proceso se lleva a cabo de manera iterativa, dividiendo el conjunto de datos en función de la variable que mejor separa los datos en términos de la variable de respuesta. Cada división se realiza buscando minimizar la varianza del valor de respuesta dentro de cada subconjunto, lo que conduce a un árbol que representa un conjunto de reglas de decisión basadas en las variables predictoras (López Pedraza et al., 2019).

La operación de un árbol de regresión comienza en la raíz, donde se encuentran todos los datos, y luego se ramifica en nodos más pequeños a medida que se realizan divisiones sucesivas. Estas divisiones se basan en criterios que buscan maximizar la homogeneidad de los valores de respuesta dentro de cada nodo resultante. Los criterios comunes incluyen la reducción de la varianza, el coeficiente de correlación o el error cuadrático medio. Cada nodo del árbol representa una regla de decisión que divide el espacio de características en subconjuntos más simples y homogéneos en términos de la variable de respuesta. Este proceso de crecimiento del árbol continúa hasta que se alcanza algún criterio de parada, como un número máximo de niveles en el árbol o un número mínimo de observaciones en cada nodo (López Pedraza et al., 2019).

La regresión simple es un método estadístico utilizado para analizar la relación entre dos variables, donde una variable, llamada variable dependiente o respuesta, se predice en función de una variable independiente o predictor. En este enfoque, se supone que la relación entre las dos variables puede ser modelada mediante una función lineal. Es decir, se busca una línea recta que mejor se ajuste a los datos observados, de manera que pueda predecir el valor de la variable dependiente para diferentes valores de la variable independiente. Esta línea de mejor ajuste se determina minimizando la suma de los cuadrados de las diferencias entre los valores observados y los valores predichos por el modelo (Carrasquilla-Batista et al., 2016).

La regresión simple utiliza la ecuación de $Y = a + bX$ en donde:

- (Y) representa la variable dependiente.
- (X) representa la variable independiente.

- (a) es la intersección en el eje Y, es decir, el valor de (Y) cuando (X) es 0.
- (b) es la pendiente de la línea, que representa el cambio en (Y) por unidad de cambio en (X).

En el caso del modelo de predicción del uso de la bicicleta de acuerdo a las condiciones climáticas la variable dependiente es el recuento total de las bicicletas de alquiler y las variables dependientes son la temporada del año, el clima, temperatura, humedad y cantidad de usuarios registrados.

Una vez cargado los datos históricos se inicia con el preprocesamiento de datos, es fundamental abordar cuestiones como los valores nulos y la normalización para garantizar la calidad y la eficacia de los modelos. Según López Malca (Implementación de Modelo Supervisado de Aprendizaje de Máquinas Para La Predicción de Alquiler de Bicicletas. | by Javier Ochoa | Medium, 2018.) la gestión de valores nulos puede implicar la eliminación de observaciones incompletas o la imputación de valores faltantes utilizando técnicas como la media o la mediana. Además, la normalización de variables es esencial para llevar a todas las características a una escala comparable, evitando que algunas variables dominen sobre otras. Este proceso facilita la convergencia del modelo durante el entrenamiento y mejora la interpretación de los coeficientes en algoritmos lineales. El preprocesamiento adecuado es un paso crítico para garantizar la robustez y la generalización del modelo.



Figura 1 Proceso del procesamiento de datos

La división del conjunto de datos en conjuntos de entrenamiento y prueba es esencial para evaluar la capacidad predictiva del modelo de manera objetiva. Usualmente, se reserva un porcentaje del conjunto de datos para la evaluación del rendimiento del modelo después de entrenarlo. De acuerdo con Zhang (Ma et al., 2022), la elección del tamaño del conjunto de prueba dependerá de la cantidad total de datos y la complejidad del modelo. Un enfoque común es utilizar un 80% de los datos para entrenamiento y el 20% restante para prueba. Esta división permite verificar la capacidad del modelo para generalizar a datos no vistos y brinda una evaluación realista de su rendimiento en situaciones del mundo real.

5. Modelado y Evaluación

La implementación de modelos predictivos en el alquiler de bicicletas es un paso crucial que requiere la traducción de algoritmos seleccionados a un código ejecutable. La elección

de herramientas y lenguajes de programación dependerá de la complejidad del modelo y la disponibilidad de bibliotecas especializadas. De acuerdo con Zheng (Zheng & Li, 2020) en este proceso, se deben cargar los datos de entrenamiento, ajustar los parámetros del modelo y realizar validaciones cruzadas para evitar el sobreajuste. La implementación eficaz garantiza la capacidad del modelo para aprender patrones a partir de los datos de entrenamiento y realizar predicciones precisas en nuevas instancias.

La evaluación de la precisión y rendimiento de los modelos es un paso crítico para validar la eficacia de las predicciones. Se utilizan diversas métricas, como el error cuadrático medio (MSE) o el coeficiente de determinación (R^2), para medir la calidad de las predicciones en relación con los valores reales. Además, se realiza una evaluación del modelo en el conjunto de prueba, que representa datos no vistos durante el entrenamiento. Este proceso ayuda a identificar posibles problemas de sobreajuste o subajuste y proporciona una estimación realista del rendimiento del modelo en situaciones del mundo real. La interpretación de estas métricas no solo permite comparar diferentes modelos, sino que también guía ajustes adicionales para mejorar la precisión predictiva. De acuerdo con Fanaee (Fanaee-T & Gama, 2014), el error cuadrático medio (MSE) es una medida comúnmente utilizada para evaluar la precisión de los modelos de predicción de bicicletas. Por otro lado, Xu (Xu et al., 2020) sugiere que el coeficiente de determinación (R^2) es una métrica útil para evaluar la calidad de las predicciones en relación con los valores reales. Además, Jelic (Jelic & Roncaglia, 2021) menciona que la evaluación del modelo en el conjunto de prueba es una técnica comúnmente utilizada para identificar posibles problemas de sobreajuste o subajuste y proporcionar una estimación realista del rendimiento del modelo en situaciones del mundo real.

La iteración entre la implementación de modelos predictivos y la evaluación de su rendimiento es una parte fundamental del desarrollo de un sistema predictivo robusto. A medida que se analizan los resultados, se pueden realizar ajustes en los parámetros del modelo o considerar la incorporación de nuevas variables para mejorar la precisión (Thirumalai & Koppuravuri, 2017.). Este ciclo iterativo es esencial para afinar el modelo y garantizar su capacidad de generalización en condiciones diversas. De acuerdo con Jelic (Jelic & Roncaglia, 2021), la implementación y evaluación efectivas se complementan mutuamente en el proceso de construcción de modelos predictivos robustos y aplicables al contexto específico de la predicción del alquiler de bicicletas según condiciones ambientales y estacionales.

La proyección de la información respecto a la renta de bicicletas implica evaluar cómo cada técnica modela y predice la demanda de alquiler de bicicletas en función de las condiciones ambientales y estacionales. Con los Árboles de Regresión, se espera que el modelo capture relaciones no lineales y complejas entre las variables predictoras y la variable objetivo, lo que podría ser beneficioso cuando hay interacciones no lineales o efectos de umbral en los datos. Por otro lado, la Regresión Simple en Series de Tiempo se centra en modelar la tendencia y la estacionalidad en los datos de alquiler de bicicletas a lo largo del tiempo, lo que puede ser útil

para capturar patrones temporales y estacionales en la demanda. La elección entre estas técnicas dependerá de la naturaleza de los datos y de la complejidad del problema de predicción.

El resultado de elegir una de las dos técnicas dependerá de varios factores, incluida la calidad de los datos, la interpretabilidad del modelo y el rendimiento predictivo. Los Árboles de Regresión pueden proporcionar modelos más flexibles y capaces de manejar relaciones complejas entre variables, pero pueden ser propensos a sobreajuste si no se controlan adecuadamente. Por otro lado, la Regresión Simple en Series de Tiempo puede proporcionar modelos más interpretables y fácilmente interpretables, pero pueden ser menos flexibles para capturar relaciones no lineales en los datos.

6. Resultados

La presentación de los resultados obtenidos en la predicción del alquiler de bicicletas basada en condiciones ambientales y estacionales es un momento crucial en el proceso de este artículo. Los resultados se expresarán en términos de métricas de evaluación, como el error cuadrático medio (MSE) o el coeficiente de determinación (R^2) (Thirumalai & Koppuravuri, 2017.), que proporcionarán una medida cuantitativa de la precisión del modelo. Se presentarán visualizaciones gráficas, como gráficos de dispersión entre las predicciones y los valores reales, para una comprensión intuitiva de la calidad del modelo. Además, se examinarán las características más influyentes identificadas (Xu et al., 2020) por el modelo para entender cómo las condiciones ambientales y estacionales impactan la demanda de alquiler de bicicletas.

Para la técnica de árboles de regresión se utilizó la librería de `from sklearn.tree import DecisionTreeRegressor` en Spyder y se utilizó el 25% de los 731 registros, donde nos mostró un score de predicción de 0.8362 para la técnica de regresión se utilizó el 25% de los 731 registros, donde nos mostró un score de predicción de 0.8804 quedando de la siguiente forma:

	Árboles Regresión	Regresión Simple
Total Registros	731	731
Datos Prueba	183	183
Datos Entrenamiento	548	548
Score Predicción	0.8362	0.8804

Figura 2 Comparación de los datos con las 2 técnicas de predicción

La interpretación de la relación entre condiciones ambientales, estacionales y las predicciones de alquiler de bicicletas implica analizar la importancia relativa de cada variable en el modelo. Se buscarán patrones y tendencias en las predicciones para entender cómo factores como la temperatura, la temporada o la hora del día afectan la demanda de bicicletas (Jelic & Roncaglia, 2021). Este análisis permitirá una comprensión más profunda de los comportamientos de los usuarios y cómo responden a diferentes condiciones. Además, se explorarán posibles interacciones entre variables para identificar relaciones complejas que podrían no ser evidentes

de manera aislada. La interpretación de estos resultados contribuirá al conocimiento general sobre la influencia de factores ambientales y estacionales en la movilidad urbana a través de sistemas de bicicletas compartidas.

La coherencia entre la presentación de resultados y la interpretación de la relación entre condiciones ambientales, estacionales y predicciones de alquiler de bicicletas es esencial para proporcionar una visión integral y significativa del estudio. Este análisis final no solo valida la efectividad del modelo propuesto, sino que también contribuye al entendimiento más profundo de cómo los factores ambientales y estacionales impactan la dinámica de la movilidad en entornos urbanos.

Al comparar las 2 técnicas, la que se acercó a una mayor predicción fue "Predicción" ya que tuvo un mayor grado de eficiencia:

7. Conclusiones

La comparación entre la utilización de la técnica de minería de regresión y árboles de regresión para la predicción del alquiler de bicicletas son fundamentales para entender la eficacia de ambos enfoques en este contexto específico. La minería de regresión, al basarse en la modelización de relaciones lineales entre variables, puede ofrecer resultados precisos cuando la relación entre condiciones ambientales y estacionales es predominantemente lineal. Su capacidad para cuantificar la influencia de cada variable de manera directa puede ser una ventaja en la interpretación de los resultados y la identificación de factores más influyentes. Sin embargo, puede no capturar de manera óptima relaciones no lineales complejas presentes en conjuntos de datos más intrincados.

Por otro lado, la utilización de árboles de regresión ofrece una aproximación más flexible y adaptable. Estos modelos pueden manejar relaciones no lineales de manera más efectiva, permitiendo capturar patrones complejos y no evidentes en un primer análisis. Su capacidad para segmentar el espacio de características de manera jerárquica los hace adecuados para conjuntos de datos con interacciones no lineales. Sin embargo, la interpretación de árboles de regresión puede ser más compleja, ya que la relación entre las variables se expresa en términos de condiciones de bifurcación.

Puede ser beneficioso explorar y comparar ambos enfoques, incluso combinándolos en un modelo conjunto para aprovechar las fortalezas de cada uno. La elección entre minería de regresión y árboles de regresión dependerá de la complejidad subyacente de las relaciones en los datos y los objetivos específicos de la predicción del alquiler de bicicletas según las condiciones ambientales y estacionales.

Referencias

- Alvarez-Valdes, R., Belenguer, J. M., Benavent, E., Bermudez, J. D., Muñoz, F., Vercher, E., & Verdejo, F. (2016). Optimizing the level of service quality of a bike-sharing system. *Omega*, *62*, 163–175. <https://doi.org/10.1016/j.omega.2015.09.007>
- Capital Bikeshare DC. (n.d.). Retrieved January 23, 2024, from <https://capitalbikeshare.com/>
- Carrasquilla-Batista, A., Chacón-Rodríguez, A., Núñez-Montero, K., Gómez-Espinoza, O., Valverde-Cerdas, J., & Guerrero-Barrantes, M. (2016). Regresión lineal simple y múltiple: aplicación en la predicción de variables naturales relacionadas con el crecimiento microalgal. *Revista Tecnología En Marcha*, *29*(8), 33. <https://doi.org/10.18845/tm.v29i8.2983>
- Eren, E., & Uz, V. E. (2020). A review on bike-sharing: The factors affecting bike-sharing demand. *Sustainable Cities and Society*, *54*, 101882. <https://doi.org/10.1016/J.SCS.2019.101882>
- Fanaee-T, H., & Gama, J. (2014). Event labeling combining ensemble detectors and background knowledge. *Progress in Artificial Intelligence*, *2*(2–3), 113–127. <https://doi.org/10.1007/s13748-013-0040-3>
- Fin De Máster, T., Beltrante, A., & Santana, A. E. (n.d.). *Predicción del uso de bicis compartidas dependiendo de las condiciones climáticas del día*.
- Gámez-Pérez, K., López, P. E. A., & Iniestra, J. G. (2020). Supporting the strategic design of public bicycle sharing systems: The experience of a large Mexican city. *Contaduría y Administración*, *65*(3). <https://doi.org/10.22201/FCA.24488410E.2020.2192>
- García-Gutiérrez, J., Romero-Torres, J., & Gaytan-Iniestra, J. (2014). Dimensioning of a Bike Sharing System (BSS): A study case in Nezahualcoyotl, Mexico. *Procedia - Social and Behavioral Sciences*, *162*, 253–262. <https://doi.org/10.1016/j.sbspro.2014.12.206>
- Gong, W., Rui, J., & Li, T. (2024). Deciphering urban bike-sharing patterns: An in-depth analysis of natural environment and visual quality in New York's Citi bike system. *Journal of Transport Geography*, *115*, 103799. <https://doi.org/10.1016/J.JTRANGE.2024.103799>
- Guo, Y., Yang, L., & Chen, Y. (2022). Bike Share Usage and the Built Environment: A Review. In *Frontiers in Public Health* (Vol. 10). Frontiers Media S.A. <https://doi.org/10.3389/fpubh.2022.848169>
- Implementación de modelo supervisado de aprendizaje de máquinas para la predicción de alquiler de bicicletas.* / by Javier Ochoa / Medium. (n.d.). Retrieved January 23, 2024, from <https://medium.com/@javier8amoreno/implementaci%C3%B3n-de-modelo-supervisado-de-aprendizaje-de-m%C3%A1quinas-para-la-predicci%C3%B3n-de-alquiler-de-d504b046e09b>
- Jelic, A., & Roncaglia, P. (2021). *Predicting bike sharing demand with machine learning*.
- López Pedraza, F. J., Macías González, Ma. D. C., & Sandoval García, E. R. (2019). Minería de datos: identificando causas de deserción en las instituciones públicas de Educación Superior de México. *TIES, Revista de Tecnología e Innovación En Educación Superior*, *2*, 1–14. <https://doi.org/10.22201/dgtic.26832968e.2019.2.4>
- Ma, X., Zhang, S., Jin, Y., Zhu, M., & Yuan, Y. (2022). Identification of metro-bikeshare transfer trip chains by matching docked bikeshare and metro smartcards. *Energies*, *15*(1). <https://doi.org/10.3390/en15010203>
- Ricci, M. (2015a). Bike sharing: A review of evidence on impacts and processes of implementation and operation. *Research in Transportation Business & Management*, *15*, 28–38. <https://doi.org/10.1016/J.RTBM.2015.03.003>
- Ricci, M. (2015b). Bike sharing: A review of evidence on the impacts and processes of implementation and operation. *Research in Transportation Business & Management*, *15*, 28–38. <https://doi.org/10.1016/j.rtbm.2015.03.003>
- Rosales-Asensio, E., Borge-Diez, D., Blanes-Peiró, J. J., Pérez-Hoyos, A., & Comenar-Santos, A. (2019). Review of wind energy technology and associated market and economic conditions in Spain. *Renewable and Sustainable Energy Reviews*, *101*, 415–427. <https://doi.org/10.1016/J.RSER.2018.11.029>
- Shaheen, S., Cohen, A., & Martin, E. (2013). Public bikesharing in North America. *Transportation Research Record*, *2387*, 83–92. <https://doi.org/10.3141/2387-10>
- Thirumalai, C., & Koppuravuri, R. (n.d.). *Bike Sharing Prediction using Deep Neural Networks*.
- Villafañez, F., Escudero, G., & Ángel, L. (n.d.). *MÉTODOS DE REGRESIÓN Y CLASIFICACIÓN BASADOS EN ÁRBOLES*.
- Vogel, P., Greiser, T., & Mattfeld, D. C. (2011). Understanding Bike-Sharing Systems using Data Mining: Exploring Activity Patterns. *Procedia Social and Behavioral Sciences*, *20*, 514–523. <https://doi.org/10.1016/j.sbspro.2011.08.058>
- Wang, J., Huang, J., & Dunford, M. (2019). Rethinking the utility of public bicycles: The development and challenges of station-less bike sharing in China. *Sustainability (Switzerland)*, *11*(6). <https://doi.org/10.3390/su11061539>
- Xu, M., Liu, H., & Yang, H. (2020). A Deep Learning Based Multi-Block Hybrid Model for Bike-Sharing Supply-Demand Prediction. *IEEE Access*, *8*, 85826–85838. <https://doi.org/10.1109/ACCESS.2020.2987934>
- Zheng, L., & Li, Y. (2020). The development, characteristics and impact of bike-sharing systems: A literature review. In *International Review for Spatial Planning and Sustainable Development* (Vol. 8, Issue 2, pp. 37–52). SPSD Press. https://doi.org/10.14246/irspsd.8.2_37
- Zhou, J., Guo, Y., Sun, J., Yu, E., & Wang, R. (2022). Review of bike-sharing system studies using bibliometrics method. *Journal of Traffic and Transportation Engineering (English Edition)*, *9*(4), 608–630. <https://doi.org/10.1016/J.JTTE.2021.08.003>