Tecnológico de Monterrey
Escuela de Ingeniería y Ciencias

Ei Expo ingenierías

Poster

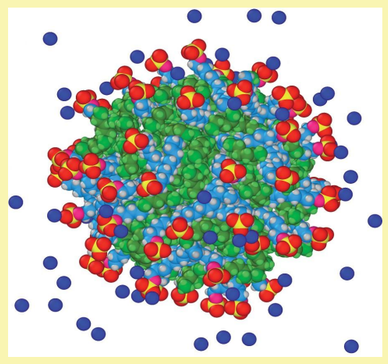
Vista esto: yoingenio, 12 junio del 2023 tec.mx/es/santa-fe

Uso de algoritmos de Machine Learning para la toma de decisiones

V Parraga-Muñoz, A Ochoa-Zuluaga, S Zamarrón-Orrantia, A Cárdenas-Arredondo, **G.J. Camacho González**

Escuela de Ingeniería y Ciencias, Tec de Monterrey en Santa Fe, Av. Carlos Lazo No. 100, Delegación Álvaro Obregón, Ciudad de México, 01389, México. A01029428@tec.mx, A01660137@tec.mx, A01781507@tec.mx, A010219917@tec.mx. **gjcamacho@tec.mx**

ABSTRACTO: Citibike es un sistema y programa de intercambio de bicicletas en la ciudad de Nueva York. Su misión es proporcionar un sistema de intercambio de bicicletas seguro, eficiente y sostenible para los residentes y visitantes de la ciudad de Nueva York, fomentando la movilidad urbana y mejorando la calidad de vida de las personas (Motivate International Inc., 2023). Su visión es convertirse en el principal sistema de bicicletas compartidas en la ciudad de Nueva York, creando una experiencia excepcional para los usuarios. Contribuir a la reducción del tráfico, promover un estilo de vida saludable y sostenible siendo un componente integral de la infraestructura de transporte de la ciudad (Motivate International Inc., 2023).Para acercarse cada vez más a sus metas y visión, Citibike debe crear estrategias inteligentes que le permitan crecer. Esto se logrará a través del análisis de datos. Se utilizarán los datos de "viajes" de años anteriores. A través del análisis descriptivo, se explorarán los datos de Citibike y se encontrarán ideas que respalden una mejor comprensión del negocio, con el objetivo de apoyar la definición de estrategias. En Cycle-Sense, se puede ver cómo se analizó el conjunto de datos proporcionado por Citibike a través de modelos de aprendizaje automático y se generaron predicciones que buscan ayudar a los tomadores de decisiones. Se creó un análisis de los datos dentro de la cual se aplicaron los conocimientos adquiridos del aprendizaje supervisado y no supervisado.



1. INTRODUCCIÓN

Machine Learning es la ciencia y el arte de programar computadoras para que puedan aprender de los datos. Estos modelos pueden ser supervisados y no supervisados. Los primeros se refieren a modelos de predicción donde los datos están “labeled” o etiquetado, entonces el modelo sabe las características del dato que está clasificando y puede predecir su comportamiento, y los segundos se tratan de modelos que clasifican por sí solos los datos y predicen su comportamiento. (1)

Guiada de esta herramienta tenemos a la minería de datos, esta permite encontrar patrones ocultos dentro de grandes volúmenes de datos en ella se busca encontrar relaciones, tendencias y conocimientos significativos que no son evidentes a simple vista, esta técnica permite extraer información significantes para la toma de decisiones y entendimiento del comportamiento de los datos.

Con la ayuda de estas dos herramientas someteremos la base de datos de Citibike con el objetivo de obtener información valiosa sobre la empresa y sus usuarios, para una toma de decisiones objetiva.

2. Exploración y limpieza de datos

Primero se conectaron las bases de datos de Citibike y el Clima, las cuales estaban localizadas en Snowflake. Se conjuntaron con “INNER JOIN” y quedó como resultado el siguiente dataframe

Esta dataframe se limpió; se llenaron los valores nulos con el promedio o la palabra Unknown. Para poder tener todos los datos completos, así como renombrar aquellos valores con nombres extraños o no reconocibles.

3. Algoritmo Predictivo

La empresa de Citibike busca maximizar ventas y minimizar costos. Para esto se plantearon las siguientes preguntas: ¿Cuál es el comportamiento de los usuarios? ¿Qué estaciones vale la pena tener? ¿Cuál es el kilometraje de las bicicletas en el inventario? y ¿Cuántas bicicletas se necesitan para las diferentes estaciones dependiendo del tiempo dado?

Para responder cada una de las preguntas se implementaron modelos de machine learning tanto supervisados como no supervisados, dependiendo de los datos relevantes para la pregunta. Los modelos utilizados fueron: Clustering, Árbol de decisión y Regresión Lineal respectivamente.

Aunque los datos fueron previamente limpiados, cada modelo requiere su propia estructuración y limpieza de datos específicos, en el caso de la Regresión Lineal es necesario tener un nuevo data frame con las variables específicas, agrupadas por las estaciones utilizadas al comienzo y al final de cada viaje. Para el árbol de decisión se agrupó con cada bicicleta diferente y su kilometraje. Y para los modelos de Clusterización se buscó agrupar las variables que responden al comportamiento tanto de usuarios (si son miembros o usuarios casuales, su género y edad) como de las estaciones (zona geográfica, coordenadas y cantidad uso total de la estación).

4. RESULTADOS

Los resultados encontrados fueron los siguientes: Algunas estaciones que están muy cerca y tienen poco uso, pueden ser consolidadas con otras que se usan más, esto gracias a una clusterización jerárquica de dichas estaciones

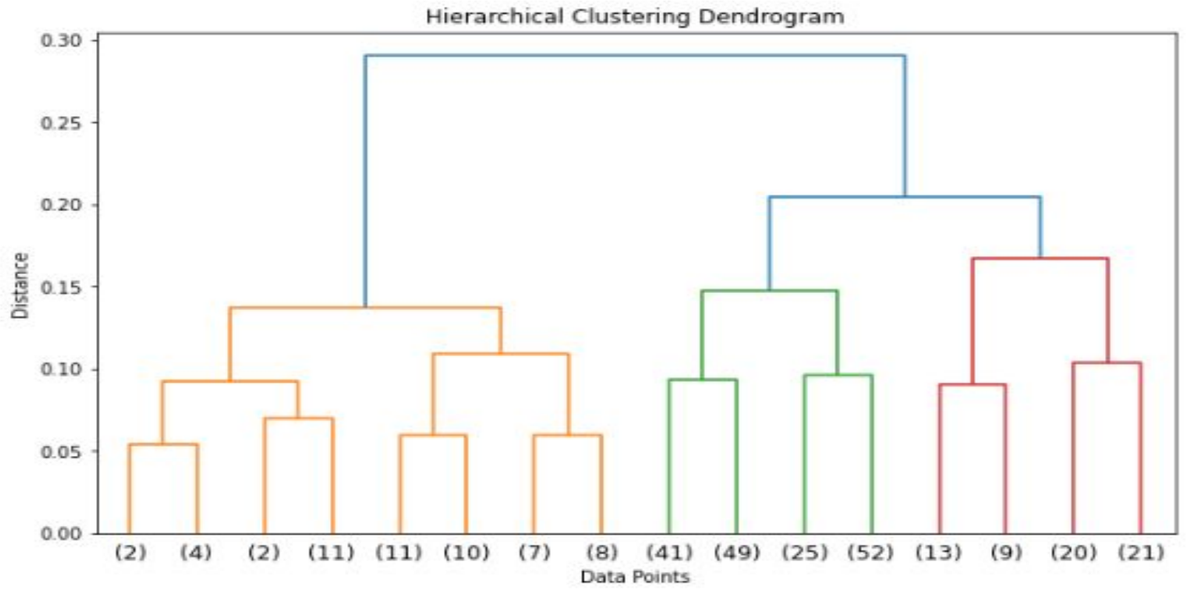


Figure 2. Dendograma de sub-clusters de las estaciones por distancia

También se pudo calcular y predecir cuando una bicicleta en inventario se debe llevar a mantenimiento, esto se hizo calculando su kilometraje y con una investigación básica sobre el debido cuidado de bicicletas normales.

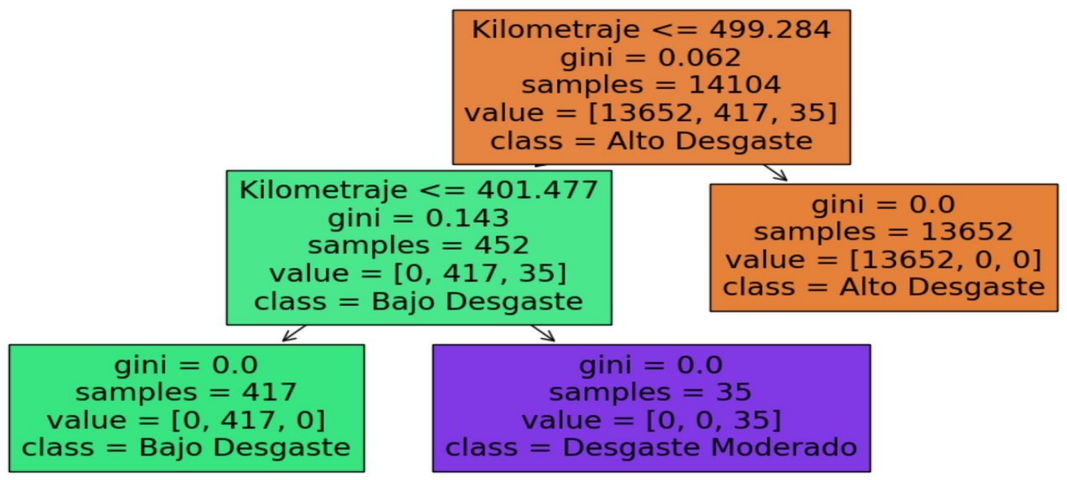


Figure 3. Árbol de decisión para mantenimiento de bicicletas.

5. CONCLUSIÓN

Podemos observar que la minería de datos y el aprendizaje automático son herramientas muy valiosas para la toma de decisiones en cualquier negocio, haciendo énfasis en que solo si nosotros, como usuarios y programadores hacemos las preguntas correctas. Siendo ese el caso, podemos concluir que el modelo brindará a los usuarios ideas valiosas y perspicaces.

References

1. Géron, A. (2019). *Hands-on machine learning with scikit-learn, keras and tensorflow: Concepts, tools, and techniques to build Intelligent Systems*. O’Reilly.