http://doi.org/10.46480/esj.7.2.188

Aplicación de algoritmos de clasificación, agrupación y predicción en la detección de patrones asociados a la movilidad usando datos de trayectorias vehiculares

Application of classification, clustering and prediction algorithms in the detection of patterns associated with mobility using vehicle trajectory data

Dayana Salvatierra[1], Joshue Laborde[2], y Oscar León-Granizo[3]

RESUMEN

Contexto: Los problemas de transito se consideran un impedimento que repercute directamente en el desarrollo personal de las personas que deben cumplir horarios especıficos, como estudiantes y trabajadores. Con el proposito de comprender mejor esta problematica, se desarrollo este estudio como un antecedente para futuras investigaciones orientadas a proponer soluciones basadas en datos. **Metodo:** Se utilizaron algoritmos de clasificacion, agrupacion y prediccion para detectar patrones asociados a la movilidad. La base de datos empleada contiene trayectorias vehiculares, y se elaboraron tres *datasets* con informacion sobre distancia, duracion, temperatura y hora del dıa. Los algoritmos aplicados fueron *K Nearest Neighbor* (clasificacion), *K-means* (agrupacion) y Regresion Lineal (prediccion). **Resultados:** Se identifico una relacion entre temperaturas elevadas y mayor duracion de trayectos, ası como la existencia de recorridos con igual distancia pero diferentes duraciones. Los periodos con mayor congestion corresponden al mediodıa y la tarde, cuando las temperaturas tambien son mas altas. **Conclusiones:** Los hallazgos evidencian que la temperatura y la hora del dıa son variables relevantes para comprender la congestion vehicular. Esto permite orientar estrategias de movilidad mas eficientes y planificar intervenciones que reduzcan el impacto del transito en horarios crıticos.

Palabras clave: Movilidad Sostenible, Patrones, K-means, Algoritmos, Transito

ABSTRACT

Context: Traffic problems are considered an impediment that directly affects the personal development of people who must meet specific schedules, such as students and workers. To better understand this issue, this study was developed as a precedent for future research aimed at proposing data-driven solutions. **Method:** Classification, clustering, and prediction algorithms were used to detect patterns associated with mobility. The database used contains vehicular trajectories, and three datasets were created with information on distance, duration, temperature, and time of day. The algorithms applied were *K Nearest Neighbor* (classification), *K-means* (clustering), and Linear Regression (prediction). **Results:** A relationship was identified between high temperatures and longer trip durations, as well as the existence of routes with the same distance but different durations. The periods with the greatest congestion correspond to midday and afternoon hours, when temperatures are also higher. **Conclusions:** The findings show that temperature and time of day are relevant variables to understand traffic congestion. This enables the design of more efficient mobility strategies and the planning of interventions to reduce the impact of traffic at critical hours.

Keywords: Sustainable Mobility, Patterns, K-means, Algorithms, Traffic

Fecha de recepción: Septiembre 18, 2022
Fecha de aceptación: Agosto 30, 2023

Introducción

En la actualidad, los miembros de las sociedades urbanas cumplen con un sin número de tareas a lo largo del desarrollo de su día a día, y muchas de estas se deben realizar en diferentes puntos de la ciudad, el gran número de personas que debe movilizarse diariamente va en aumento y el volumen de vehículos también. La importancia de las tareas, como el caso de los estudiantes y trabajadores, confiere preocupación al tema de la necesidad de movilidad sostenible, pues en muchos casos, la impuntualidad genera afectaciones en los roles que desempeñan (Terraza et al., 2021).

Los investigadores que buscan obtener soluciones a problemáticas como las que se acaban de mencionar, necesitan de herramientas y precedentes que permitan cumplir con ese objetivo. Es por esto que el centro principal que caracteriza este trabajo investigativo es la aplicación de algoritmos (predicción, agrupación, clasificación) para el análisis de datos e identificación de patrones asociados a la movilidad. Los datos a analizar serán obtenidos a partir de la base de datos del sistema SIAM desarrollado por la

[1]Universidad de Guayaquil. E-mail: dayana.salvatierrak@ug.edu.ec
[2]Universidad de Guayaquil. E-mail: joshue.labordeg@ug.edu.ec
[3]Universidad de Guayaquil. E-mail: oscar.leong@ug.edu.ec

Cómo citar: Salvatierra, D., Laborde, J. ., & León-Granizo, O. (2024). Aplicación de algoritmos de clasificación, agrupación y predicción en la detección de patrones asociados a la movilidad usando datos de trayectorias vehiculares. *Ecuadorian Science Journal, 7*(2), *10–18, Septiembre–2023.* DOI: https://doi.org/10.46480/esj.7.2.188

Universidad de Guayaquil, que cuenta con el registro de trayectorias circundantes al campus universitario.

Se puede citar la existencia de trabajo relacionados a la movilidad sostenible, como el desarrollado por (Morquecho, 2019) donde indica que, pese al desarrollo de las infraestructuras de seguridad y vialidad, las redes de transporte público urbano, cada vez están más congestionadas y cubren las necesidades de los usuarios con menor calidad, debido, en parte, a la mala planificación urbana. En el caso de la ciudad de Guayaquil, donde se realizó este trabajo, hay muy poca información disponible. La lentitud de la planificación urbanística de los últimos años ha provocado un aumento de la demanda de viajes, pero ninguna mejora de las infraestructuras de transporte. Lo que permite ampliar la visión de los causales que pueden intervenir en el problema planteado y a su vez reforzar la justificación de la importancia del desarrollo de este artículo.

Otro presedente, al estudio que se propone, fue plasmado en el artículo científico propuesto por la revista Ingenio titulado "Identificación de Patrones de Movilidad Utilizando Datos en Tiempo Real Generados por Access Points en una Red de Comunicaciones de Campus. Caso de estudio: Universidad Central del Ecuador" donde explica que el objetivo de la investigación es proporcionar información verdadera, objetiva y en tiempo real sobre los patrones de viaje generados por los usuarios en el campus de la Universidad Central de Ecuador, capaz de identificar las características de los usuarios. Para ello se utiliza una plataforma informática compuesta por dos aplicaciones: una app para extraer y almacenar los datos en una base de datos relacional, y una aplicación web para mostrar los resultados a los usuarios. De los resultados finales, se pudieron detectar patrones de movilidad que hubieran sido dificiles y costosos de obtener con métodos tradicionales como el encuestar, permitiendo de esta forma, facilitar la labor de los investigadores de su universidad (Chavez Estrella et al., 2020).

Tambien se analizó, el estudio propuesto por (Alcívar Vargas, 2022), titulado: "Recolección y procesamiento de datos de movilidad para determinar los diferentes modos de transporte utilizados en la Universidad Central del Ecuador", establece que la mayoría de las personas en la actualidad, tienen un dispositivo móvil con varios sensores y aplicaciones para acceder a la información geográfica. El GPS, el acelerómetro y el giroscopio del dispositivo móvil pueden recoger datos sobre la ubicación y la actividad de las personas que se desplazan de un lugar a otro con una precisión aceptable. El estudio se propuso identificar los diferentes modos de transporte utilizados para desplazarse por la Universidad Central de Ecuador, utilizando los datos de movilidad recogidos en un proyecto piloto en el que hubo participación voluntaria por parte de los alumnos de la Facultad de Ingeniería y Ciencias Aplicadas. Utilizando la aplicación Google Maps instalada en la mayoría de los dispositivos móviles, se pudieron recoger los datos del historial de localización de los participantes activando la opción "Your Timeline". Los registros de localización de Google proporcionan información en formato JSON, un formato plano que permite el intercambio de datos entre aplicaciones. Para el análisis computacional, fue necesario transformar este archivo utilizando un paquete desarrollado en Python para obtener un conjunto de datos líquidos.

Los algoritmos de predicción, agrupación y clasificación implementados serán Regresión Lineal, K Nearest Neighbor y K Means respectivamente.

Regresión Lineal

La regresión lineal es un algoritmo de Machine Learning utilizado para el aprendizaje supervisado. El funcionamiento de este algoritmo es predecir una variable dependiente o también llamada objetivo, basada en la o las variables independientes proporcionadas. Por lo tanto, esta técnica de regresión encuentra una relación lineal entre una variable dependiente y las otras variables independientes dadas. Por este motivo, este algoritmo tiene como nombre regresión lineal (Bastien et al., 2005).

K Nearest Neighborn

El K- Nearest Neighbor es un algoritmo de aprendizaje automático supervisado basado en instancias. Puede utilizarse con nuevas muestras para categorizar (valores discretos) o pronosticar (regresión, valores continuos). Es la introducción perfecta al campo del aprendizaje automático, ya que es un enfoque sencillo. Su trabajo fundamental es categorizar valores identificando los puntos de datos "más comparables" (por proximidad) descubiertos durante la fase de entrenamiento, y luego inferir nuevos puntos basados en esta categorización (Khandelwal et al., 2023).

K Means

K-means es un algoritmo de agrupación (clustering) no supervisado que divide los objetos en K grupos en función de sus cualidades. La agrupación se realiza minimizando la suma de las distancias entre cada objeto y los centroides de su grupo o cluster. Se suele utilizar una distancia cuadrática (Ikotun and Ezugwu, 2022).

Metodología

En este trabajo de investigación abarca la búsqueda de patrones de movilidad por medio de tres algoritmos de aprendizaje automático, concretamente K-nearest neighbors, K-means y Regresión Lineal (algoritmo de clasificación, agrupación y predicción respectivamente) haciendo uso de la base de datos proporcionada a partir de una investigación previa con relación a la movilidad sostenible, con el fin de obtener un análisis con resultados más significativos que a su vez permitan plantear soluciones eficaces.

La investigación siguió una metodología cuantitativa con diseño no experimental de tipo trasversal. Esto debido a que los resultados y la forma de obtener los mismos se darán por medio de mediciones universales y operaciones cuantificables implementando los algoritmos propuestos. Además, se considera trasversal porque el procesamiento de los datos para obtener resultados se realizará en un solo momento y usando datos recolectados previamente.

Análisis de fuente de datos

Para llevar a cabo el primer objetivo se realizó un estudio de la información que está alojada en la base de datos Movilidad del sistema SIAMS-UG, sistema desarrollado por la

Universidad de Guayaquil que tiene registros de trayectorias recogidas previamente para otros trabajos investigativos, dentro de ella se buscó las tablas más relevantes que tengan información de trayectorias vehiculares con el fin de poder implementar los algoritmos, a continuación, se especifica los pasos a seguir para la realización del objetivo:

1. **Backup de la base de datos "Movilidad".** En primera instancia se obtuvo el Backup de la Data que fue proporcionada por medio de integrantes del proyecto FCI-010 de partícipe de anteriores proyectos relacionados. El Backup proviene originalmente de la base de datos que se encuentra en los servicios de AWS (Amazon Web Service) cuya información almacenada se origina del sistema SIAM, además de datos relacionados a trayectorias vehiculares que detallan la longitud, latitud, distancia, coordenadas, temperaturas, etc.

2. **Restore de la base de datos.** Una vez obtenido el Backup, se procedió a realizar el restore de la base de datos a través de la consola de comandos, con el fin de poder de analizar los datos e implementar los algoritmos.

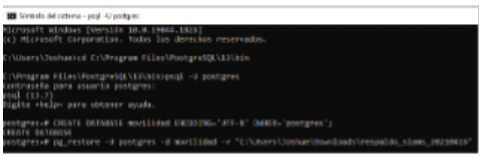

Figura 1. Restore a través de la consola de comandos.

3. **Análisis de la data.** Una vez realizada la restauración, se accede a la base de datos para analizar y comprender las tablas, campos, relaciones y registros cuya finalidad poder decidir qué datos utilizar para aplicar los algoritmos.

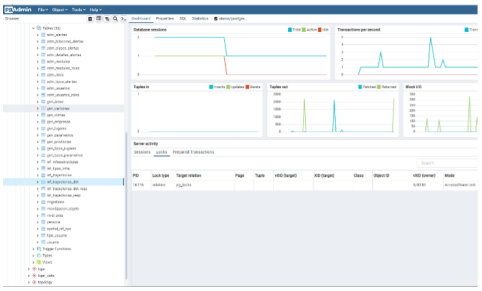

Figura 2. Base de datos restaurada.

A continuación se muestra la estructura que conforma la base de datos restaurada y se procede a la selección de tablas a analizar:

- **gen_climas:** Esta tabla contiene una relación a la tabla inf_trayectorias además de información descriptiva del tiempo climático relacionado a las trayectorias registradas, como datos de: ciudad, temperatura, tipo de clima, condiciones atmosféricas, entre otros.
- **Inf_trayectorias:** Tabla que indica de forma general la trayectoria hacia un punto de destino entre los que destacan los campos que informan

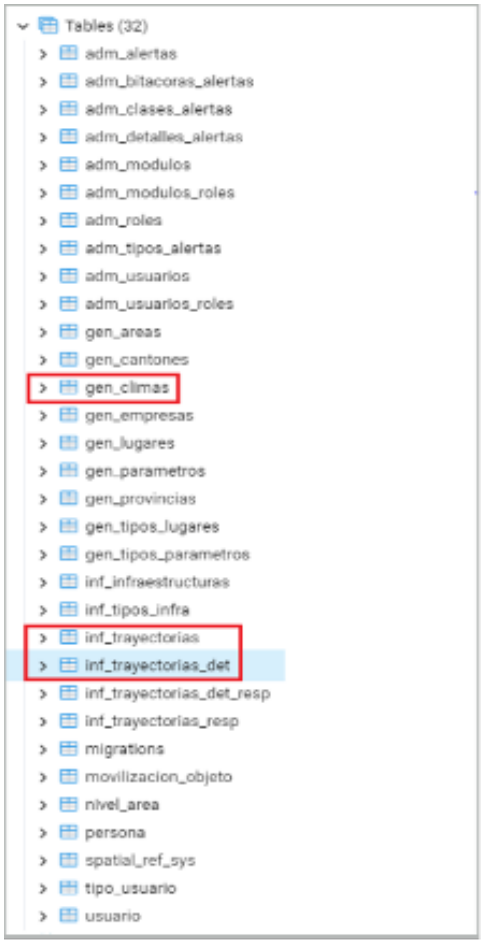

Figura 3. Tablas de la base de datos de movilidad.

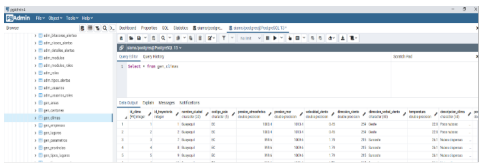

Figura 4. Tablas gen_climas.

de la fecha del registro, área y datos geográficos (punto de origen, última posición y movimiento del trayecto). Además de: tiempo, distancia total y velocidad.

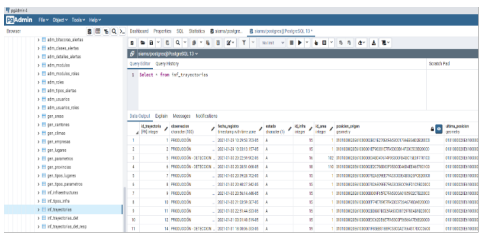

Figura 5. Tablas inf_trayectoria.

- **Inf_trayectorias_det:** Detalles relacionados a un registro de trayectoria principal de la tabla antes mencionada. En este se agrupan las trayectorias que componen a la trayectoria principal, contiene campos como: coordenadas de la trayectoria (latitud y longitud), coordenadas en el mapa, tipos de

coordenadas, duración, velocidad; incluso se puede visualizar un orden de trayectos.

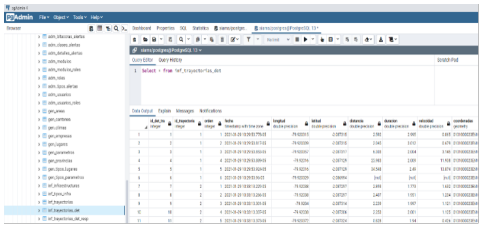

Figura 6. Tablas Inf_trayectorias_det.

4. **Análisis exploratorio de datos y entrenamiento de algoritmos.** Después de revisar y analizar la base de datos, se comenzó a limpiar los datos con el objetivo de evitar cualquier sesgo de valores no deseados. Esto se hizo por medio de un query en PostgreSQL (PgAdmin4) para así, exportarlo a un documento con extensión csv.

Figura 7. Query que filtra datos no nulos y diferentes a cero.

Con el método info() se podrá conocer los tipos de datos de las columnas del Dataset. En este caso, se tiene 2 tipos de datos que son de tipo entero y decimal.

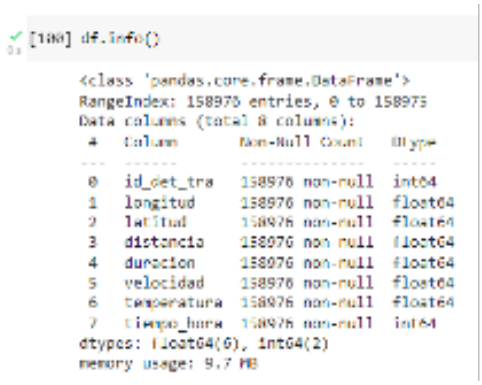

Figura 8. Descripción estadística de los datos.

Por último, se visualizará el gráfico de correlación de los datos del Dataset.

Una vez analizado los datos con las diferentes funciones estadísticas que ofrece Python, se procederá a crear tres datasets que contendrá información relevante para poder seguir con el objeto de estudio. Por consiguiente, se llevará a cabo la implementación de los tres algoritmos seleccionados.

5. **Selección de datos.** Para la implementación de los algoritmos, se crearán tres datasets que contendrán información relevante para el desarrollo del presente trabajo investigativo. Estos datasets se obtendrán del conjunto de datos analizado en el paso anterior, se definió muestras aleatorias de 3000 registros para cada dataset. Por consiguiente, se ejecutará y entrenará los tres algoritmos con el fin de obtener y evaluar los resultados arrojados por cada algoritmo. A

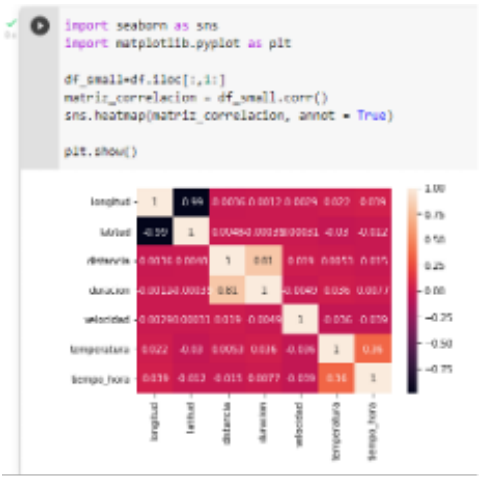

Figura 9. Gráfico de correlación de datos.

continuación, se presentará los datasets o paqueterías de datos a usar:

Dataset #1: Esta paquetería tendrá el nombre de setTemperatura establecida por la relación de las tablas inf_trayectorias, inf_trayectorias_det y gen_climas. El dataset contiene los campos de temperatura y duración relacionadas a los registros de trayectos, se extraerán 3000 datos aleatorios. A partir de estos campos se presenta el campo result que representa el resultado para cada algoritmo.

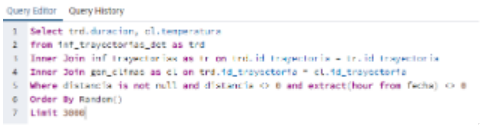

Figura 10. Query para la paquetería setTemperatura.

Tabla 1. Datos de la Paquetería setTemperatura

Tabla	Columna	Tipo de Dato
inf_trayectorias_det	Duración	Double precision
gen_climas	Temperatura	Double precision
setTemperatura	Result	Integer

Dataset #2: Esta paquetería tendrá el nombre de set-Distancia establecida por la tabla inf_trayectorias_det. El dataset contiene los campos de distancia y duración de las trayectorias, se conformará por 3000 datos aleatorios. A partir de estos campos se presenta el campo result que representa el resultado para cada algoritmo.

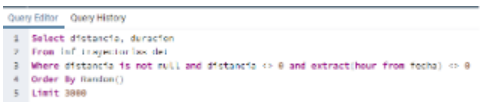

Figura 11. Query para la paquetería setDistancia.

Dataset #3: La (tercera) paquetería que tendrá de nombre setTiempo establecida por la tabla inf_trayectorias_det. Los campos que contienen son hora, que se deriva del campo fecha de la tabla en mención, y duración; este dataset también estará

Tabla 2. Datos de la Paquetería setDistancia

Tabla	Columna	Tipo de Dato
inf_trayectorias_det	Distancia	Double precision
inf_trayectorias_det	Duración	Double precision
setDistancia	Result	Integer

conformado por 3000 registros aleatorios. A partir de estos campos se presenta el campo result que representa el resultado para cada algoritmo.

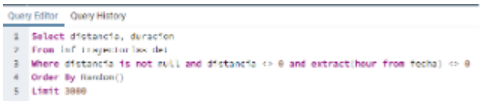

Figura 12. Query para la paquetería setTiempo.

Tabla 3. Datos de la Paquetería setTiempo

Tabla	Columna	Tipo de Dato
inf_trayectorias_det	Hora	Integer
inf_trayectorias_det	Duración	Double precision
setTiempo	Result	Integer

Entrenamiento de los algoritmos

Se tomó en consideración el lenguaje de programación Python para la implementación y análisis de los diferentes tipos de algoritmos. El entorno donde se trabajó fue en Google Colab, debido a que ofrece una gran capacidad de recursos informáticos en su capa gratuita. Para la ejecución de cada algoritmo no se realizó ningún cambio en su estructura, sólo se ajustó los hiperparámetros que ofrece los algoritmos para el presente trabajo de investigación.

Aplicación de algoritmo de agrupación

Parámetros para análisis de algoritmo Kmeans. A continuación, se detallan los parámetros que se considerarán al entrenar el algoritmo kmeans por cada dataset elaborado. Para todos los datasets se utilizaron los siguientes parámetros: Cantidad de registros: 3000; Valor de K: 2; Init: K-means++; Max_iter: 300; N_init: 10; Randon_state: 42.

Tabla 4. Resumen de Análisis K-means

Descripción	Dataset setTemperatura	Dataset setDistancia	Dataset setTiempo
Cantidad de registros	3000	3000	3000
Valor de K	2	2	2
Init	K-means++	K-means++	K-means++
Max_iter	300	300	300
N_init	10	10	10
Random_state	42	42	42

Análisis del dataset setTemperatura para el algoritmo Kmeans. El algoritmo kmeans tiene como finalidad encontrar agrupa-ciones con características similares para identificar patrones que a simple vista no se perciben. En la siguiente tabla de datos se detalla los análisis realizados para la implementación del algoritmo.

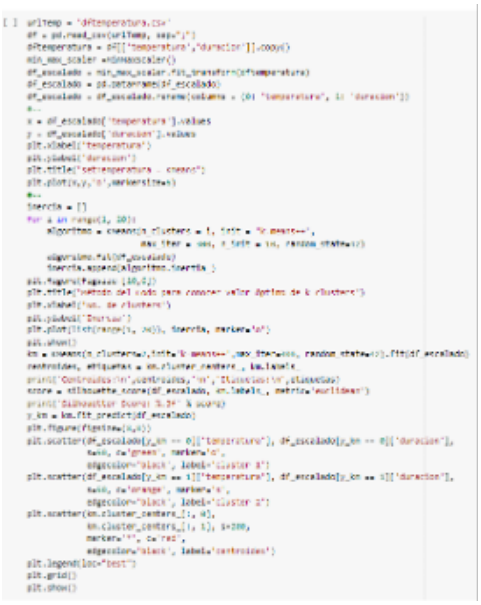

Figura 13. Código de entrenamiento Kmeans dataset setTemperatura.

Análisis del dataset setDistancia para el algoritmo Kmeans. En este dataset se hace uso de los campos distancia y duración. En la siguiente tabla de datos se resume el análisis hecho para su implementación.

Figura 14. Código de entrenamiento Kmeans dataset setDistancia.

Análisis del dataset setTiempo para el algoritmo Kmeans. En este dataset se hace uso de los campos fecha, donde se obtiene la hora del día y la duración. En la siguiente tabla de datos se resume el análisis hecho para su implementación.

Aplicación de algoritmo de predicción

Parámetros para análisis de algoritmo de Regresión Lineal. A continuación, se detallan los parámetros que se

Figura 15. Código de entrenamiento Kmeans dataset setTiempo.

considerarán al entrenar el algoritmo regresión lineal por cada dataset elaborado. Para todos los datasets se utilizó: Datos de entrenamiento: 75% (2250 registros); Datos de prueba: 25% (750 registros).

Tabla 5. Resumen de análisis Regresión Lineal

Descripción	Dataset setTemper-atura	Dataset set-Distancia	Dataset set-Tiempo
Variable in-dependiente	Temperatura	Distancia	Hora
Variable de-pendiente	Duración	Duración	Duración
Datos de en-trenamiento	2250 (75%)	2550 (75%)	2550 (75%)
Datos de prueba	300 (25%)	300 (25%)	300 (25%)

Análisis del dataset setTemperatura para el algoritmo regresión lineal. Para la implementación de este algoritmo se utilizó el campo temperatura como variable independiente y duración como dependiente. Después se prosigue a separar el dataset entre datos de entrenamiento y datos para la prueba, se utilizó el 75% y 25% respectivamente de un total de 3000 registros que vendría ser el 100% de la muestra escogida

Figura 16. Código de entrenamiento regresión lineal dataset setTemperatura.

Análisis del dataset setDistancia para el algoritmo regresión lineal. Para la implementación de este algoritmo se utilizó el campo distancia como variable independiente y du-ración como dependiente. Después se prosigue a separar el dataset entre datos de entrenamiento o aprendizaje y datos para la prueba, se utilizó el 75% y 25% respectivamente de un total de 3000 registros que vendría ser el 100% de la muestra escogida.

Figura 17. Código de entrenamiento regresión lineal dataset setDistancia.

Análisis del dataset setTiempo para el algoritmo regresión lineal. PPara la implementación de este algoritmo se utilizó el campo fecha extrayendo la hora como variable independiente y duración como dependiente. Después se prosigue a separar el dataset entre datos de entre-namiento o aprendizaje y datos para la prueba, se utilizó el 75% y 25% respectivamente de un total de 3000 registros que vendría ser el 100% de la mues-tra escogida.

Figura 18. Código de entrenamiento regresión lineal dataset setTiempo.

Aplicación de algoritmo de clasificación

Para este tipo de algoritmo de clasificación es necesario tener una columna que clasifique los otros campos, en consecuencia, se procedió a agregar un campo llamado clasificación que no es más que una columna que mide la duración de la trayectoria en tres niveles determinado por diferentes rangos: nivel 1 (Baja duración, valores entre 0 a 1), nivel 2 (Media duración, valores entre 1 a 2), y nivel 3 (Alta duración, valores mayores a 2).

Parámetros para análisis de algoritmo de vecinos más cercanos (KNN). Se utilizó el mismo esquema de división de datos: 75% para entrenamiento y 25% para prueba.

Análisis del dataset setTemperatura para el algoritmo KNN. Para la implementación de este algoritmo se utilizó el campo temperatura y duración como variables independientes y clasificación como dependiente. Después se prosigue a separar el dataset entre datos de entrenamiento y datos para la prueba, se utilizó el 75% y 25%

Tabla 6. Datos para columna clasificación

Clasificación	Descripción
1	Baja duración. Valores entre 0 a 1
2	Media duración. Valores entre 1 a 2
3	Alta duración. Valores mayores a 2

Tabla 7. Resumen de análisis vecinos más cercanos

Descripción	Dataset setTemperatura	Dataset set-Distancia	Dataset set-Tiempo
Variable independiente	Temperatura	Distancia	Hora
Variable dependiente	Clasificación	Clasificación	Clasificación
Datos de entrenamiento	2250 (75%)	2550 (75%)	2550 (75%)
Datos de prueba	300 (25%)	300 (25%)	300 (25%)

respectivamente de un total de 3000 registros que correspondería al 100% de la muestra escogida.

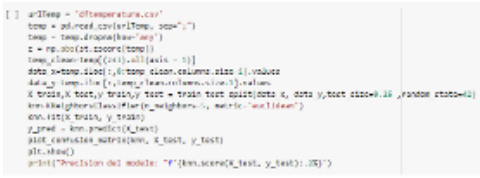

Figura 19. Código de entrenamiento regresión lineal dataset setTemperatura.

Análisis del dataset setDistancia para el algoritmo KNN. Para la implementación de este algoritmo se utilizó el campo distancia y duración como variable independiente y el nuevo campo llamado clasificación como dependiente. Después se prosigue a separar el dataset entre datos de entrenamiento o aprendizaje y datos para la prueba, se utilizó el 75% y 25% respectivamente de un total de 3000 registros que vendría ser el 100% de la muestra escogida.

Figura 20. Código de entrenamiento regresión lineal dataset setDistancia.

Análisis del dataset setTiempo para el algoritmo KNN. Para la implementación de este algoritmo se utilizó el campo fecha extrayendo la hora y duración como variable independiente y clasificación como dependiente. Después se separa el dataset entre datos de entrenamiento o aprendizaje y datos para la prueba, se utilizó el 75% y 25

Figura 21. Código de entrenamiento regresión lineal dataset setTiempo.

Gráficas Resultantes

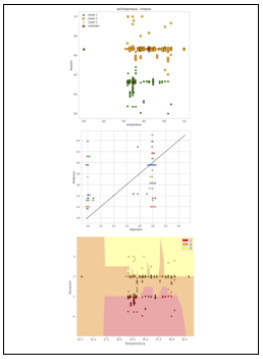

Figura 22. Gráfica de kmeans, regresión lineal y vecinos más cercanos de dataset setTemperatura.

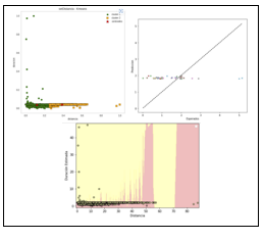

Figura 23. Gráfica de kmeans, regresión lineal y vecinos más cercanos de dataset setDistancia.

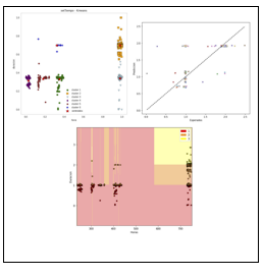

Figura 24. Gráfica de kmeans, regresión lineal y vecinos más cercanos de dataset setTiempo.

Resultados

Como resultado a la problemática planteada en este trabajo investigativo, se obtuvo como respuesta los siguientes puntos analizados:

Los resultados detectados para el dataset de temperatura, con el algoritmo de agrupación Kmeans, nos permite identificar que parece existir una relación entre la duración de los trayectos y las temperaturas, puesto que se identificó que los datos de duración registrados se daban en periodos de tiempo con altas temperatura, lo que nos podría indicar

Tabla 8. Resultados de Análisis por Algoritmo y Dataset

	Algoritmo K-means	**Algoritmo Regresión Lineal**	**Algoritmo K-Nearest-Neighbor**
Métrica	**Coeficiente Silhouette**	**Coeficiente de determinación R^2**	**Precisión de la clasificación**
Dataset Temperatura	0.862	0.539	0.988
Dataset Distancia	0.606	0.006	0.870
Dataset Tiempo	0.914	0.053	0.986

que es probable que los momentos en los que más se prolongaban los trayectos tienen relación al gran volumen de autos que se movilizan al medio día, considerado un momento de hora pico en el tránsito, además del momento en que se manifiestan las más altas temperaturas del día. Pese a tener una gran cantidad de registros para aplicar el análisis, existen varios valores dispersos en las agrupaciones que pueden estar relacionadas a la distribución de los datos o a la selección de las características a evaluar.

Por otro lado, al aplicar Regresión lineal para el análisis de algoritmo de predicción, se identificó que la correlación entre datos predichos con datos reales no era del todo congruente, hecho que pudo identificarse y demostrarse a partir del coeficiente de determinación generado que nos indican que las predicciones acertadas correspondían únicamente a un 56,44%, lo que se puede interpretar casi como una probabilidad de azar.

Al aplicar el algoritmo de clasificación, k-nearest neighbors, al mismo dataset, se pudieron identificar tres clasificaciones en el conjunto de datos, para obtener esta cantidad de clasificaciones fue necesario evaluar la precisión en función de todos los posibles valores de k-puntos a considerar; donde se determinó que 2 y 3 son los óptimos. Los resultados obtenidos de las clasificaciones fueron tres rangos de duración que podrían cualificarse como duración: alta, media o baja; los datos de temperatura tenían una misma distribución dentro de cada clasificación. En base a la gráfica de las clasificaciones se halló una mayor concentración de registros para la clasificación de duración media, pues dentro de esta clasificación también se encontraban las temperaturas en su punto más bajo y más alto. También se puede comprobar la concentración de datos a través de la gráfica de matriz de confusión que se generó, donde se indica que la clasificación dos (duración media) tiene 510 datos, mientras que la clasificación uno y dos tienen 124 y 116 datos respectivamente.

Posteriormente, los resultados que se hallaron para el dataset de distancia, al aplicar el algoritmo de agrupación K-means, fueron dos clusters u agrupaciones. La agrupación uno y dos están constituidas por trayectorias de duración baja; pero se puede identificar que la mayor parte de las duraciones más bajas se encuentran en la agrupación uno, al igual que las más altas. Incluso se pueden identificar outliers o datos aislados de duración muy alta que son parte del cluster uno.

Otra particularidad que diferencian al clúster uno y dos son las distancias, mientras que en el clúster uno se encuentran todas las distancias bajas, en el clúster dos se hayan todas las distancias medias a largas. Se puede considerar que las trayectorias de la agrupación uno, recorren distancias cortas con una duración similar a la agrupación dos, pese a que esta última tiene distancias más largas. Por ende, se puede analizar la posibilidad de que las personas partícipes de las trayectorias del cluster uno tenga mayores problemas de movilidad que las personas del clúster dos.

No obstante, al analizar el mismo dataset con el algoritmo Regresión Lineal para hallar predicciones, los resultados de las predicciones no coincidían, incluso el coeficiente de determinación generado indicaba un 0,68% de acierto en las predicciones, en vista de que se consideraron los mismos hiperparámetros de entrenamiento que se aplicaron a los otros datasets estudiados; se decidió alterar los hiperparámetros para mejorar los resultados de predicción, la modificación realizada aumentó el coeficiente de determinación a un 2,10%, lo que se sigue considerando un nivel muy bajo por lo que se considera poco probable acertar las predicciones del algoritmo.

Adicionalmente al estudio del dataset de distancia, se analizaron los datos de duración y distancia de trayectorias usando el algoritmo de clasificación, k-nearest neighbors, desde el que se detectaron dos clasificaciones para los datos. La clasificación dos se distingue por ser la más grande de las dos mencionadas, además de contar con la mayor parte de trayectorias con duración media a alta; a diferencia de la clasificación uno que contiene la mayor parte de duraciones medias a bajas. Ambas clasificaciones comparten datos distribuidos en un mismo rango de distancia; además ambas tienen datos aislados.

De igual forma que los dataset anteriores mencionados, se procedió a analizar el dataset de tiempo, implementando el algoritmo de agrupación Kmeans se obtuvieron como resultado seis agrupaciones con un valor de coeficiente silhouette 0.904 (cercano a 1) lo que indica que la calidad de las agrupaciones es óptima puesto a que los datos dentro de cada grupo tienen coherencia entre sí. Los cluster uno, tres, cuatro y cinco tienen duraciones similares, teniendo la mayor parte de sus datos en la franja de duración entre 0.2 y 0.4; se puede considerar que el promedio de los trayectos tiene ese intervalo de duración. El cluster uno y tres contienen datos de duraciones aisladas. Los clusters dos y seis se caracterizan por tener una mayor duración en sus datos de trayecto; inclusive el cluster dos contiene una mayor concentración de datos que abarca duraciones medias a prolongadas. Los clusters uno, cuatro, cinco y seis se ubican en la primera mitad del eje de hora, lo que indica que estas agrupaciones se dan en horas de la mañana. Los clústeres dos y seis, que cuentan con las duraciones más largas de trayectos, se ubican en franjas cercanas al medio día y en horas de la tarde respectivamente, lo que nos podría indicar que estas agrupaciones pudieron haber transitado en áreas con problemas de tránsito vehicular ocasionado por horas de alta demanda de movilidad.

En tanto a que, al realizar el análisis del algoritmo de predicción, Regresión Lineal, se obtuvo una gráfica con pocas coincidencias en las predicciones, pues los valores

esperados no concordaban con los resultados generados por el algoritmo, en reducidos puntos se visualiza una alineación con la recta de tendencia lineal, adicional a ello el coeficiente de determinación resultante corresponde a un 64,19% de acierto para las predicciones halladas.

Al momento de aplicar el algoritmo de clasificación k-nearest neighbors sobre el mismo dataset de tiempo, se obtuvo un indicador de precisión de 98,67%. A partir de la gráfica se identifican gran cantidad de valores superpuesto entre clasificaciones, por lo que a simple vista se visualiza la clasificación uno como la más significativa por ocupar gran parte de la gráfica. A partir del tercer tercio del eje horas, se visualiza de forma más definida la delimitación entre clasificaciones, incluso se puede inferir que se trata de intervalos de duración prolongada, media y corta; donde la clasificación uno corresponde a corta, clasificación dos a medio y clasificación tres a prolongada. Por lo que se considera que la clasificación está directamente relacionada a la duración de los trayectos a lo largo del día, que pueden parecer difusos en un principio, pero se ve claramente la clasificación en el segmento de la tarde (horas).

Conclusiones

A partir del riguroso proceso investigativo aplicado sobre los de movilidad proporcionados, se puede concluir lo siguiente:

- Se logró conformar tres dataset de interés para el estudio, posterior a una limpieza de datos realizada sobre el respaldo proporcionado de la base de datos del sistema SIAM, donde se aplicaron los algoritmos de agrupación, predicción y clasificación (K-means, Regresión Lineal y K Nearest Neighbor respectivamente) a cada dataset conformado. El análisis y estudio de los datos fue realizado en el entorno de desarrollo Google Colab, permitiendo mayor velocidad de ejecución que la que se podría llegar a tener con recursos de hardware limitados.
- Durante la experimentación, las variables dependientes a implementar en la investigación (Coeficiente Silhouette, Coeficiente de determinación R^2 y Precisión de la clasificación) tuvieron resultados admisibles; a excepción del Coeficiente de determinación R^2 que se halla en la Regresión Lineal, este último arrojó valores ínfimos para todos los datasets, lo que no permitió considerarlo un modelo de predicción confiable. Se contempló la posibilidad de que esta desviación puede darse por la distribución de los datos.
- Los patrones detectados se dieron por relaciones entre los datos que conformaban los datasets modelados:
  - En el caso del dataset de temperatura, donde se estudió la duración de trayectoria con relación a la temperatura, se evidenció una relación entre las temperaturas altas y mayores duraciones de trayectos, y terminó vinculándose con el hecho de que se considera que las mayores temperaturas se dan al medio día, y a su vez, estas coinciden con un aumento de tránsito vehicular por considerarse una hora de movimiento masivo.
  - En el dataset de distancia, donde se consideró la duración y la distancia, se obtuvo que, a pesar de que existen agrupaciones de trayectos con distancias similares, se diferencian por tener una duración mayor; incluso el grupo de trayectorias con mayor duración tiene la principal concentración de registros, lo que nos permite percibir que los problemas de tránsitos tienen una afectación consistente.
  - En tanto que, el dataset de tiempo, donde se considera la duración y los intervalos de tiempo que conforman el día, arrojó como resultado que las duraciones más largas se evidenciaban cercanas al medio día y horas de la tarde previo al anochecer, momentos del día considerados horas pico de tránsito donde se forma congestión vehicular.

Agradecimientos

Los autores agradecen a la Universidad de Guayaquil por proporcionar acceso a la base de datos del sistema SIAM utilizada en esta investigación.

Referencias

Alcívar Vargas, M. (2022). Recolección y procesamiento de datos de movilidad para determinar los diferentes modos de transporte utilizados en la universidad central del ecuador. Master's thesis, Universidad Central del Ecuador, Facultad de Ingeniería y Ciencias Aplicadas.

Bastien, P., Vinzi, V. E., and Tenenhaus, M. (2005). Pls generalised linear regression. *Computational Statistics and Data Analysis*, 48(1):17–46.

Chavez Estrella, M., Enríquez-Reyes, R., Cadena Flores, G., and Mocayo Unda, M. (2020). Identificación de patrones de movilidad utilizando datos en tiempo real generados por access points en una red de comunicaciones de campus. caso de estudio: Universidad central del ecuador. *Revista el Ingenio*.

Ikotun, A. M. and Ezugwu, A. E. (2022). Boosting k-means clustering with symbiotic organisms search for automatic clustering problems. *PLoS ONE*, 17(8).

Khandelwal, M., Rout, R. K., Umer, S., Sahoo, K. S., Jhanjhi, N. Z., Shorfuzzaman, M., and Masud, M. (2023). A pattern classification model for vowel data using fuzzy nearest neighbor. *Intelligent Automation and Soft Computing*, 35(3):3587–3598.

Morquecho, R. I. (2019). Análisis de movilidad en entornos urbanos. Master's thesis, Instituto Politécnico Nacional.

Terraza, M., Zhang, J., and Li, Z. (2021). Intersection signal timing optimisation for an urban street network to minimise traffic delays. *Promet - Traffic - Traffico*, 33(4):579–592.