Date of current version 12, 2022.

ANÁLISIS DE MOVILIDAD EN ZONAS UNIVERSITARIAS

Arista Huerta, Maria de los Ángeles; Castro Payns, Mariana; Ibarra Mora, Marcela; Ramírez Pintor, Manolo; Rodríguez Gil, Eduardo
Instituto Tecnológico de Estudios Superiores de Monterrey, Campus Querétaro
Contacto por orden de autor: A01231973@tec.mx, A01706038@tec.mx, A01369984@tec.mx, A01706155@tec.mx, A01274913@tec.mx

RESUMEN Dentro del desarrollo de este reporte se describe el proceso que se llevó a cabo para la realización de un análisis de minería de datos, con el cual se busca proporcionar conocimiento de valor para la toma de decisiones dentro de la ciudad de Santiago de Chile con respecto a la movilidad, enfocados hacia el efecto que tienen las universidades en la movilidad, para esto se implementaron diferentes modelos de predicción con el propósito de hacer un forecast de la cantidad de viajes que se realizan hacia estas universidades.

PALABRAS CLAVE: inteligencia artificial, matriz de viajes, movilidad, predicción, universidades.

I. INTRODUCCIÓN

Se busca tomar decisiones dentro del Sistema de Transporte Urbano, por medio del uso de información confiable y actualizada. Anteriormente, se realizaban encuestas para conocer la movilidad de los ciudadanos pero, la obtención manual de datos representaba altos costos para ellos, además se obtenían pocas encuestas el tiempo de espera para ingresar manualmente los datos para su análisis era demasiado. Ahora, con los datos obtenidos de Telefónica, ya no presentan estos problemas ya que la obtención de datos, es más rápida y mucho menos costosa.

Gracias a los datos generados por los usuarios, es posible estimar diferentes rutas y comportamientos de movilidad en la población, dichas rutas permitirán realizar un análisis que faciliten una toma de decisiones acertada. [1]

Como principal objetivo, se realizará un análisis de la movilidad enfocada a las sedes de las universidades Top 5 en Santiago de Chile, de acuerdo a rankings como “THE World University Rankings - Times Higher Education” [2] y “QS World University Rankings” [3]. Esto debido a que las universidades pueden tener un efecto significativo dentro de la movilidad de la ciudad .

II. ESTADO DEL ARTE

Para la realización de este proyecto, se hizo uso de la metodología CRISP-DM, Cross Industry Standard Process for Data Mining. Esta es una metodología de ciclo de vida que consiste en orientar trabajos de minería de datos.

CRISP-DM proporciona un ciclo de vida para analizar datos a través de distintas fases y tareas. Dependiendo de los objetivos y el contexto en el que se hará el proyecto, existen distintas formas de aplicarlo. [4]

Aplicando esta metodología, definimos objetivos de negocio y minería de datos con la etapa de Business Understanding. Posteriormente, al tener identificados y confirmados estos objetivos, pasamos a la etapa de Data Understanding, en ella realizamos un análisis exploratorio de datos, EDA, revisamos los datos que nos parecieron importantes, luego en la etapa de Data Preparation pasamos a realizar construcción y limpieza de datos para poder obtener poco a poco un set de datos que pudiéramos utilizar para entrenar y predecir datos que sean de utilidad para el objetivo de negocio. Esta parte de entrenamiento se lleva a cabo en la etapa de Modeling y es realizado a través de modelos de Machine Learning y de Deep Learning. [5]

Durante la etapa de Modeling, se crean distintos modelos y se mejoran para obtener los mejores resultados de predicción modificando hiper parámetros de aprendizaje.[6]

Después de la etapa de Modeling, pasamos a la etapa de Evaluation, donde comparamos el desempeño de los modelos y tomamos el que mejor desempeño tuvo para su aprobación.

Finalmente, en la etapa de despliegue, se toman todos los resultados de la evaluación para entregar los resultados de la minería de datos al negocio.
Con la información obtenida de nuestro socio formador, pudimos tomar bases de investigación que nos permitieron

conocer y tomar en cuenta supuestos que podríamos aplicar para la realización de nuestro proyecto. En este paper que se nos mostró, se realizó una investigación de viajes a los centros comerciales ubicados en Santiago de Chile. [7]

En este trabajo de investigación se obtuvo lo siguiente:

- Análisis de 387,152 teléfonos celulares entre 16 centros comerciales, tomando un mes como rango del análisis.
- Predicciones de la afluencia de personas en los centros comerciales, la distribución en base al perfil del cliente en cada centro comercial, con datos de ubicación, distribución de población y el tamaño del centro comercial.
- La atracción de clases sociales bajas y medias a ciertos centros comerciales.

Tener esta información nos sirvió como guía para generar nuestros datos de viajes a distintas zonas universitarias en Santiago de Chile, con el fin de ver el impacto que generan estas zonas en la movilidad, analizando diversos factores y horarios en los que había mayor actividad.

III. MÉTODOS

Inicialmente contamos con un dataset de 49,618,132 datos, de los cuales al analizarlos, se realizan operaciones de construcción y limpieza, para obtener un dataset final para la preparación de los modelos.

Se obtuvo la ubicación de las antenas para saber en qué comuna se ubica, después se obtuvieron las trazas de los viajes mediante cálculos realizados entre los registros de los usuarios.
Se realizó la limpieza de los usuarios sin movimiento y teletransportaciones que daban velocidades inexplicables.
Reemplazo de los IDs de las antenas con una concatenación de las coordenadas para obtener valores únicos. En este punto se redujo la cantidad de datos a 37,732,972 muestras..
Se integró la información de las comunas y posteriormente se realizó la matriz de viajes, donde a partir de varias condiciones generamos un dataset. Ahora sólo quedaban 1,203,010 datos.
Con la información de la matriz, y las sedes universitarias se llevó a cabo la creación del dataset final con datos de zonas universitarias a 1km de radio para la parte de modelado. Así es como nos quedó un registro de la suma de los viajes por hora, universidad y comuna.

Como variables de análisis adicionales se integraron datos como: si la universidad es pública o privada, si la sede tiene residencias, si el viaje registrado era dentro o fuera de la zona, el número de carreras y el tamaño en metros cuadrados.

-Modelos generados

Los modelos serán construidos y entrenados con un conjunto de entrenamiento, su calidad estimada con los conjuntos de validación y finalmente se probará el modelo con un conjunto de datos para pruebas. La división del dataset se determina el 75% de datos para entrenamiento y 25% de datos para pruebas. Del 75% de los datos de entrenamiento se toma un 20% para la validación del modelo.

Para todos los modelos se tuvo que realizar one hot encoding para convertir las variables de string a variables categóricas numéricas. De esta manera los modelos aceptarán y entenderán la entrenar.

- Regresión lineal: describe la relación entre una variable dependiente y una o más variables independientes

  En la implementación del modelo de regresión lineal nos auxiliaremos del modelo de scikit-learn LinearRegression().

- Random Forest: el modelo consiste en la implicación de varios árboles de decisión.

  Para la implementación de este modelo se utilizó el modelo tf.keras.RandomForestModel()de la librería tensorflow decision forests

- Gradient Boosted:Este modelo se basa en crear varias instancias independientes de tipo Random Forest en paralelo, algunos árboles mejores que otros, esto con el propósito de que el siguiente árbol sea capaz de corregir los errores del árbol pasado

  Para este modelo, se utilizó el tf.keras.GradientBoostedTreesModel()de la librería tensorflow decision forests.

- Extra trees Regressor: Extra trees Regressor:El modelo se caracteriza por entrenar numerosos árboles de decisión y agrega los resultados del grupo de árboles de decisión para generar una predicción.

  En este modelo, se usó el from sklearn.ensemble import ExtraTreesRegressor de la librería sklearn ensemble.

- Densely-Connected Neural Network (Sequential): es un tipo de red neuronal que minimiza los errores de predicción, tiene distintas neuronas conectadas entre sí para modificar los pesos aplicados en cada una. Este tipo de red funciona por medio de capas de neuronas que reciben y generan información para las siguientes.

  En la implementación de este modelo se utilizó TensorFlow Keras. Creamos un modelo secuencial para ir añadiendo capas densas y tener neuronas que vayan moviendo sus valores eficientemente a través de la función Adam, funciones de activación y funciones para desactivar aleatoriamente y de forma completa ciertas neuronas para evitar overfitting.

Se realizaron ajustes en los parámetros para mejorar los desempeños de los modelos. Finalmente, se obtuvo un modelo que demostró un mejor desempeño el cual fue el *Extra Trees Regressor.*

IV. RESULTADOS

Las métricas empleadas para la evaluación de los modelos son:

- ➔ Mean Absolute Percentage Error (MAPE)
  Permite conocer el porcentaje de error que presentan las predicciones del modelo, mide la precisión del sistema empleando como función de pérdida en análisis de regresión.
- ➔ R Square (r2)
  Coeficiente de determinación, el cual representa una proporción de la varianza entre una variable dependiente que es posible explicar mediante una o más variables independientes en un modelo de regresión.
- ➔ Root mean square error (RMSE)
  El error cuadrático medio de la raíz describe la diferencia entre los valores previstos y observaciones.

Modelo	HyperParametros	MAPE	R^2	RMSE
Regresión lineal	*fit_intercept=True* *normalize=False* *copy_X=True* *n_jobs=None* *positive = False*	1.98	0.08	6.59
Regresión lineal	positive=True normalize=True n_jobs=100	1.86	0.056	6.65
Extra Trees Regressor	Max_features = 1.0	53.71	0.7687	3.846
Extra Trees Regressor	Max_features = 40	52.97	0.7692	3.891
Random Forest Model	Número de árboles = 73, Profundidad max = 10	92.53	-	13.86
	Número de árboles = 200, profundidad max = 50	75.14	-	13.86
Gradient Boosted	Número de árboles = 73, Profundidad max = 10	88.73	-	4.5
	Número de árboles = 63, profundidad max = 13	73.31	-	4.5
Densely-Connected Neural Network	Número de capas = 4 Número de entradas = 7 Capa 1: • Capa densa con 100 neuronas, activación softmax Dropout de 60% Capa 2 • Capa densa de 80 neuronas, activación relu. Dropout de 20% Capa 3: • Capa densa con 60 neuronas , activación relu	45.76	-	9.83
	Capa 4: • Capa densa de 8 neuronas, activación sigmoid. Número de salidas = 1			

Comparación de los modelos

Consideramos que los resultados obtenidos, son de utilidad e importancia ya que se está dando a conocer cómo es el comportamiento de la movilidad generada por las zonas universitarias de Santiago.

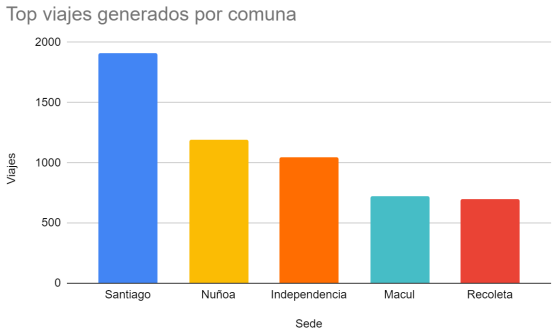
Con esto pudimos dar a conocer las comunas que registran mayor cantidad de viajes a determinada zona universitaria, también conocemos los horarios en los que las personas se desplazan y cómo es el desplazamiento, si dentro de la zona o fuera de ella.

Los hallazgos obtenidos son:

Número de viajes por cada comuna

El top 5 de viajes por cada comuna es:

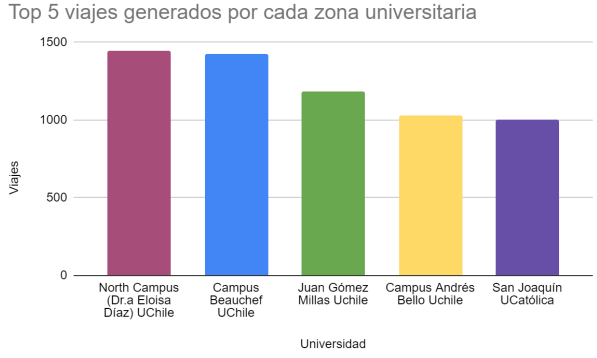
1. Santiago = 1910
2. Ñuñoa = 1189
3. Independencia = 1042
4. Macul = 721
5. Recoleta = 699



Número de viajes por cada sede universitaria

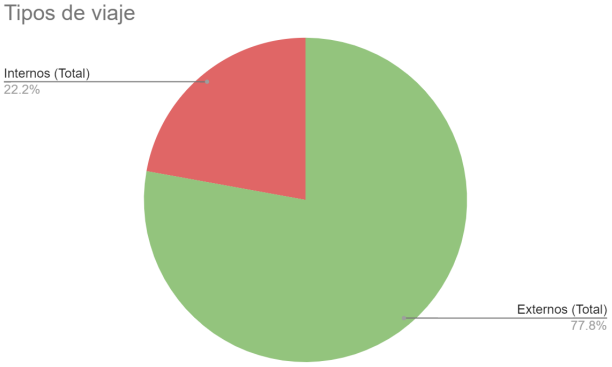
El top 5 de viajes por cada sede es:

1. North Campus (Dra. Eloisa Diaz) Uchile = 1446
2. Campus Beauchef Uchile = 1426
3. Juan Goméz Millas Uchile = 1182
4. Campus Andrés Bello Uchile = 1027
5. San Joaquín UC = 1003



Viajes internos y externos

Los viajes internos se dan más en las sedes de mayor tamaño y que cuentan con varias facultades. Se puede rectificar con la media de los viajes internos siendo de 166 y los viajes externos de 534, viéndose la clara diferencia de viajes entre uno y otro.



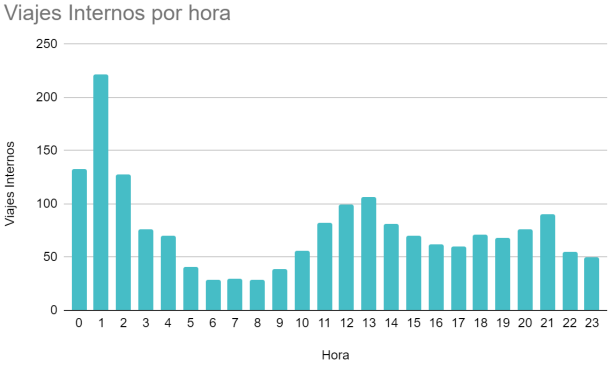
Horario con mayor número de viajes

La hora en la que más viajes se presentaron, en viajes internos y externos fue a la 1 de la mañana. Para el análisis de estos datos se utilizó la media de viajes por hora, para el caso de viajes externos obtenemos una media de 267 y para viajes internos obtenemos una media de 76. A partir de esto buscamos los horarios que cumplan con una cantidad mayor o superior y así podemos concluir que los rangos de hora de viajes con mucho movimiento son:

1. 12:00 am a 3:59 am
2. 01:00 pm a 02:59 pm
3. 09:00 pm a 09:59 pm

En el caso de los viajes internos, también obtenemos mucho movimiento dentro de los siguientes rangos:

1. 12:00 am a 03:59 am
2. 11:00 am a 02:59 pm
3. 08:00 pm a 09:59 pm



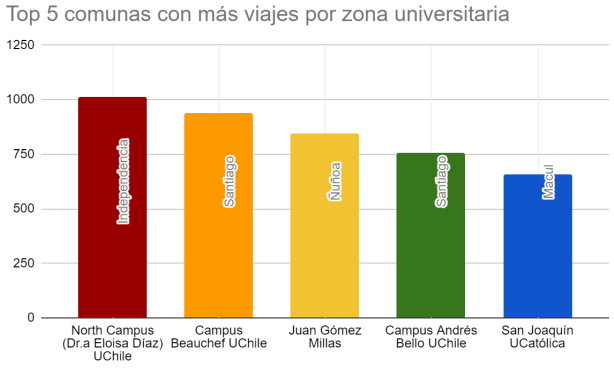
Comuna de origen con la mayor cantidad de viajes por sede

Top 5 universidades con más viajes por comuna de origen,

1. North Campus (Dra. Eloisa Díaz) Uchile
   Independencia - 1013
2. Campus Beauchef Uchile

Santiago - 941

3. Juan Gómez Millas
   Ñuñoa - 844
4. Campus Andrés Bello Uchile
   Santiago - 759
5. San Joaquín UC
   Macul - 659



V. DISCUSIÓN Y ANÁLISIS DE RESULTADOS

En base a los hallazgos obtenidos, es posible decir que hay una gran área de oportunidad de mejora, principalmente dentro de la etapa de modelado, en la cual se pudo haber dedicado más tiempo para ajuste de hiper parámetros, y así obtener mejores resultados.

Los principales problemas con los que nos enfrentamos fueron el reprocesar los datos múltiples veces debido a errores de lógica para limpiar y generar datos, por lo cual presentamos un atraso significativo. Esto fue solucionado al procesar los datos por medio de diferentes recursos computacionales, lo cual optimizó considerablemente el tiempo de procesamiento.

Otro problema que tuvimos fue que el dataset final que generamos durante la fase de para el entrenamiento del modelo, no contaba con datos suficientes para realizar predicciones en las que se pudiera confiar para su uso en el los objetivos de negocio.

VI. REFERENCIAS

[1] A. Azevedo y M. Zantos. "CRISP-DM: La metodología para poner orden en los proyectos - Sngular". Sngular. Disponible en: https://www.sngular.com/es/data-science-crisp-dm-metodologia/ (accedido el 1 de diciembre de 2022).

[2] E. Graells-Garrido y D. Saez-Trumper. "A Day of Your Days: Estimating Individual Daily Journeys Using Mobile Data to Understand Urban Flow". arXiv.org. Disponible en: https://arxiv.org/abs/1602.09000 (accedido el 1 de diciembre de 2022).

[3] QS. World University Rankings. "QS World University Rankings: Top global universities". Top Universities. Disponible en: https://www.topuniversities.com/qs-world-university-rankings (accedido el 1 de diciembre de 2022).

[4] R. Alonso. "IA, Machine Learning y Deep Learning, ¿cuál es la diferencia?" HardZone. Disponible en: https://hardzone.es/tutoriales/rendimiento/diferencias-ia-deep-machine-learning/ (accedido el 1 de diciembre de 2022).

[5] T. Higher Education. "World University Rankings". Times Higher Education (THE). Disponible en https://www.timeshighereducation.com/world-university-rankings/2023/world-ranking (accedido el 1 de diciembre de 2022).

[6] Martínez. M "Optimizando tus hiper-parámetros: una perspectiva teórica". Paradigma. Tecnología con propósito para un mundo mejor. Disponible en https://www.paradigmadigital.com/dev/optimizando-hiper-parametros-una-perspectiva-teorica/ (accedido el 1 de diciembre de 2022).