Universidad del Desarrollo
Facultad de Ingeniería

DETERMINACIÓN DE LA VARIACIÓN DE LOS TIEMPOS DE VIAJE AL INCORPORAR UNA NUEVA ESTACIÓN DE METRO

POR: FABIAN NOVA Y WILDER PRADO

Proyecto de grado presentado a la Facultad de Ingeniería de la Universidad del Desarrollo para optar al grado académico de Magíster en Data Science

PROFESOR GUÍA:

Dra. Daniel Opitz

Diciembre 2023

SANTIAGO

TABLA DE CONTENIDO

INDICE DE FIGURAS

INDICE DE TABLAS

Resumen

El presente trabajo busca analizar y validar la hipótesis planteada sobre la variación de los tiempos de viaje en el transporte público de las estaciones de metro de la región Metropolitana de Santiago de Chile. La manera como se procedió se dividió en 3 etapas:

La primera consistió en recopilar los datos basados fundamentalmente en las estadísticas de la página del Directorio de Transporte Público Metropolitano (DTPM) y también se complementaron los datos con el ingreso promedio del hogar según comuna, fuente Encuesta CASEN 2023, la densidad Poblacional por comunas, Nivel educacional en las comunas de la Región Metropolitana de Santiago de Chile. También, se seleccionaron los datos del año 2019 filtrando solo los días laborales, se llevó a cabo un análisis grafico descriptivo de las variables, luego se consolidaron los datos definiendo las variables que intervendrían para el posterior análisis del modelo.

En la segunda fase, se utilizaron 11.5 Millones de registros y las variables que participarían en la ejecución de los modelos, las cuales fueron 22: Comuna de Subida, comuna de Bajada, zona de Subida, zona de Bajada, periodo de subida (que se convirtió en variable dummy), cantidad de transbordos realizados, metros por zonas, metros por comuna, promedio de año de estudio, cantidad de personas por zonas y Cantidad de paraderos por zonas y como variable objetivo tiempo promedio de viaje. Estas variables fueron utilizadas para entrenar los cuatro modelos de regresión previamente seleccionados: XGBOOST, Random Forest Regressor, LightGBM y Árbol de Decisión Regressor. De entre ellos, se determinó que el mejor desempeño lo obtuvo el modelo Random Forest Regressor.

Finalmente, se llevaron a cabo dos mediciones. En primer lugar, se determinó la predicción del tiempo de viaje entre una zona de origen y un destino. Posteriormente, se llevó a cabo la predicción para este mismo recorrido, pero considerando la incorporación de una o más

estaciones de metro. La diferencia entre estas dos mediciones nos permitió evaluar el impacto en el tiempo de viaje al introducir nuevas estaciones de metro. La comuna de Maipú fue seleccionada para las pruebas de validación debido a ser la segunda con mayor demanda en el uso del transporte público.[1]

[1] https://www.emol.com/noticias/Nacional/2023/08/10/1103724/alcalde-maipu-extension-metro-l9.html#:~:text=El%20alcalde%20apunta%20a%20que,con%2010%20y%2015%20respectivamente

1. Introducción

Los sistemas de transporte público desempeñan un papel fundamental en la vida de millones de personas en las ciudades modernas. En un mundo cada vez más conectado y con un ritmo de vida vertiginoso, la eficiencia y accesibilidad de este sistema son cruciales para mejorar la calidad de vida de sus ciudadanos.

La ciudad de Santiago de Chile, como muchas metrópolis en crecimiento en todo el mundo, se enfrenta a desafíos significativos en términos de movilidad urbana. La congestión del tráfico, los tiempos de viaje prolongados y la eficiencia del transporte público son preocupaciones comunes que afectan el bienestar de sus habitantes. En este contexto, surge una hipótesis de gran relevancia: la adición de nuevas estaciones al sistema de metro de Santiago reducirá el tiempo promedio de viaje de los usuarios.

Esta investigación se propone explorar y analizar críticamente esta hipótesis, evaluando el impacto potencial de la expansión del sistema de metro en el tiempo de viaje promedio de los usuarios y, en última instancia, en la calidad de vida de los residentes de Santiago. A través de un enfoque multidisciplinario que involucra, análisis exploratorio de datos, definición de un modelo, análisis de resultados y validación de la hipótesis, buscamos validar cómo el crecimiento planificado del metro podría influir en la movilidad urbana y en la transformación del tejido urbano de la ciudad.

Los hallazgos de este estudio podrían tener implicaciones significativas en la planificación del transporte y contribuir a la mejora de la movilidad urbana en una ciudad en constante evolución. La expansión y mejora de la red de metro de una ciudad se considera una estrategia fundamental para abordar los problemas de movilidad urbana. Se espera que la incorporación de nuevas estaciones no solo aumente la accesibilidad

al transporte público, sino que también optimice los tiempos de desplazamiento, contribuyendo así a la eficiencia y comodidad de los desplazamientos diarios de los residentes.

2. Trabajo Relacionado

El transporte juega un papel crucial en las ciudades porque es la forma en que las personas se mueven para hacer lo que necesitan y llegar a donde quieren ir. Según (Miralles-Guasch, 2002), el transporte es esencial para el crecimiento y funcionamiento de las ciudades modernas. En otras palabras, sin el transporte, las ciudades no serían lo que son hoy en día. No se puede considerar una ciudad sin pensar en cómo las personas se mueven en su interior; ambas cosas están estrechamente relacionadas en una interacción compleja entre las personas, sus necesidades y los servicios que utilizan. Se ha visto que, en países desarrollados como Hong Kong, Londres y Paris, el 40% de los viajes motorizados se realizan en transporte público (Ortega, 2018). En la Región Metropolitana el porcentaje de viajes que se realiza en transporte público alcanza casi el 50% (Miguel et al., 2014), sin embargo, durante la pandemia en el 2020 se produjo una disminución en el transporte público del 84% en relación al año anterior (Pezoa et al., 2023) (Artículo publicado por Journal of Transport Geography (2023)).

Acerca de las causas y efectos de los retrasos en el metro de Santiago de Chile publicado en el artículo: "Analysis of Metro Delays in Santiago, Chile: Causes and Effects" señala que las causas podrían incluir problemas técnicos, mantenimiento, congestión en las vías, accidentes, huelgas, entre otros. Analiza los factores que contribuyen a los tiempos de demora en el sistema de transporte subterráneo de la

ciudad tales como la congestión en las estaciones, insatisfacción de los pasajeros y el impacto de en la puntualidad de los viajeros examinaron los efectos de los retrasos en la operación del metro y en la experiencia de los usuarios. Esto podría incluir la pérdida de tiempo, la congestión en las estaciones, la insatisfacción de los pasajeros y el impacto en la puntualidad de los viajeros.

El incremento en el uso de este modo ayuda a disminuir la congestión, contaminación, y al mismo tiempo, trae beneficios a las personas (Gotschi, 2011; Krizec, 2016). Por otro lado, las áreas verdes también se están haciendo cada día más importantes, pues no solo influyen en la configuración y habitabilidad de los espacios públicos, sino también en el incremento de la calidad de vida de los habitantes de una ciudad (Krekel, Kolbe, & Wüstemann, 2015).

Obtener información precisa sobre los horarios de llegada y salida de los autobuses en las paradas es uno de los parámetros clave del transporte público, según un estudio publicado por la Universidad de Maribor en el 2021, “Modeling and Predicting Bus Arrival Delays in Urban Transit Networks", enfoca la predicción de retrasos en la llegada de autobuses en sistemas de transporte urbano. El modelo que proponen en este estudio se basa en la ubicación actual del autobús , la clasificación de los recorridos en periodos de tiempo con respecto a los datos históricos y el modelo de datos de la red de autobuses.

La puntualidad de los trenes en Suecia alcanzó el 94 % para el año 2020 durante la pandemia (Eliasson, 2022), pero volvió a caer al 90 % en 2021 (cerca del período 2013-2017 (Eliasson, 2022)). Un trabajo interesante que se realizó en la ciudad de New York publicado como: “Analysis of Factors Influencing Subway Delays in New York City” En

este estudio se propusieron modelos para modelar y predecir retrasos en los trenes, que básicamente consistieron en dos grandes aspectos.

1. Un modelo de regresión global que son adoptados de los modelos de regresión estadística (Gorman, 2009).
2. Variables explicativas que se basan en el tipo de tren, velocidad, número de trenes, retrasos en salidas, tiempo de permanencia entre otras.

En un estudio sobre investigación de tecnologías emergentes sobre el transporte llamado “Un enfoque de fusión de datos con datos de teléfonos móviles para actualizar estimaciones de división modal basadas en encuestas de viajes” (Graells-Garrido, Opitz, Rowe, & Arriagada, 2023), se propone un método que utiliza datos de teléfonos móviles para capturar en tiempo real los patrones de movilidad humana con alta precisión espacial y temporal. Este enfoque, que integra datos de aplicaciones móviles con información oficial, permite mejorar las estimaciones sobre cómo se distribuye la movilidad en diferentes modos de transporte a nivel de área. El estudio, centrado en Santiago, Chile, identifica cambios significativos en los patrones de transporte entre 2012 y 2020, incluyendo una reducción en el uso del transporte público y un incremento en el transporte motorizado. Además, resalta un cambio hacia modos de viaje menos sostenibles, como el uso de taxi.

Uno de los estudios significativos sobre el impacto de la demanda del transporte público producto de una nueva línea de metro (Alcalá, 2017) utiliza datos masivos provenientes de tarjetas inteligentes con el objeto de evaluar el impacto generado por la implementación de la línea 6 del metro de Santiago de Chile, en este estudio se contemplan dos fases, la primera evaluación estima el impacto que tienen los atributos de uso de suelo, sociodemografía y oferta sobre la demanda del transporte público, la

segunda estima cuánto fue el cambio de demanda provocado por la implementación de la línea 6 del metro, esta variación es agregada a nivel de origen y destino para luego analizar cuáles son los sectores más beneficiados.

De otra manera también se han realizado estudios cuyo objetivo es comparar y validar la información sobre patrones de viaje y cargas de pasajeros en el Metro de Santiago a partir de dos fuentes de datos: encuestas de origen-destino y datos de tarjetas inteligentes. Los resultados muestran una fuerte correlación entre ambas fuentes, lo que podría llevar a una futura complementación de datos y ahorro de recursos en la obtención de información detallada sobre el comportamiento de los viajeros (Pineda, Schwarz, Godoy, 2015).

La variabilidad del tiempo de viaje puede ser un factor importante a considerar en la planificación de rutas y horarios, especialmente para quienes dependen del transporte público o necesitan llegar a destinos con horarios ajustados, se ha demostrado que los usuarios no sólo tienen disposición a pagar por ahorrar tiempo de viaje promedio, sino que además están dispuestos a pagar por disminuir la variabilidad de su tiempo de viaje (Bates, Polak, Jones, & Cook, 2001), ya que de esta manera podrían planificar de un modo más eficiente el horario de salida desde el origen para llegar a una hora deseada a destino.

También se ha realizado caracterizaciones de la variabilidad del tiempo de viaje de la ciudad de Santiago en todas las etapas de un viaje en transporte público por separado: acceso, espera, en vehículo y trasbordo. Para el caso de tiempo en vehículo se comparará la VTV de tres modos, utilizando el tiempo de viaje normalizado por la

distancia (min/km), de manera de tener una medida que permita comparar los resultados obtenidos para bus, metro y automóvil (Durán, 2016).

3. Hipótesis

La adición de nuevas estaciones de metro de Santiago de Chile reducirá el tiempo promedio de viaje de los usuarios.

4. Objetivos Generales

Estimar el tiempo de viaje y la variación que existe de un mismo recorrido al aumentar nuevas estaciones de metro seleccionando el mejor modelo predictivo.

4.1. Objetivos Específicos

- Recopilar y analizar los datos de los tiempos de viaje promedio de los usuarios del transporte público.
- Definir la metodología a utilizar.
- Evaluar diferentes tipos modelos de aprendizaje supervisado.
- Seleccionar el modelo con mejor rendimiento predictivo.
- Predecir los tiempos de viaje de un recorrido determinado y luego utilizar el mismo modelo incorporando nuevas estaciones de metro.

5. Datos y Metodología

5.1. Datos

Los datos se obtuvieron de la página del Directorio de Transporte Público Metropolitano (DTPM) que depende del Ministerio de Transporte y Telecomunicaciones, generadas a partir del software ADATRAP desarrollado por la Universidad de Chile. Para el caso particular de estudio se tomaron en cuenta las Tablas Agregadas de MatrizOD, Subidas y Bajadas del año 2019.

Detalle de las tablas utilizadas:

- ✓ Tabla de Viajes (.csv) y Estructura de Tabla de Viajes (.txt): Periodo: Abril (días 8, 9, 10, 11, 12, 13 y 14)
- ✓ Tabla de Etapas (.csv) y Estructura de Tabla de Etapas (.txt), Año 2019
- ✓ Tablas Agregadas de Matriz OD, Subidas y Bajadas: Años 2013 al 2019

También se incluyeron otras fuentes complementarias como:

- ✓ Ingreso promedio del hogar según comuna, fuente Encuesta CASEN 2018, Ministerio del Desarrollo Social.
- ✓ Densidad Poblacional por comunas, fuente: Censo 2017.
- ✓ Nivel educacional en las comunas de la Región Metropolitana de Santiago de Chile, fuente: Censo 2017

5.1.1 Descripción de la base de datos

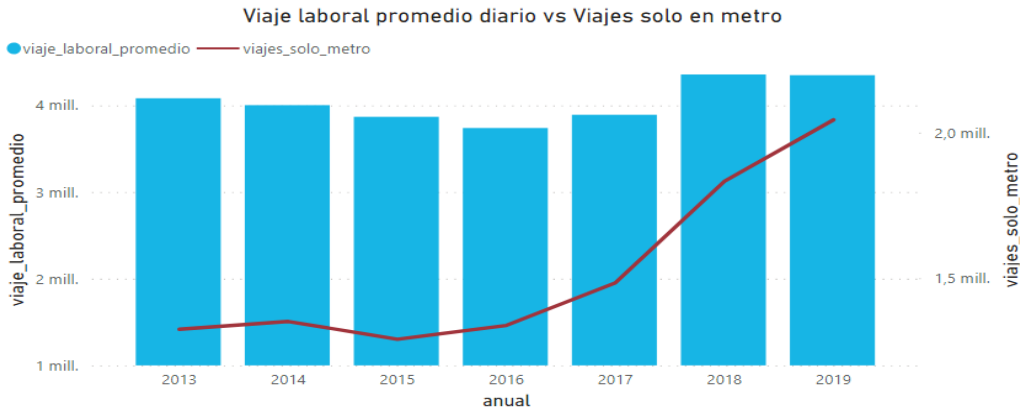
Las bases de datos se encuentran divididas por años desde 2011 a 2019, cada una de ellas tiene diferente estructura, algunas contienen tabla de viajes, etapas y tablas agregadas de matriz OD, subidas y bajadas. En el caso del año 2019 el cual fue seleccionado para el estudio contiene la información completa en cuanto a tablas agregadas.

Para el procesamiento de los datos se cargaron las bases de datos a Big Query debido a su tamaño, esta herramienta permite explorar y visualizar el análisis de datos a través de Looker Studio. A continuación, se presenta el análisis exploratorio de los datos:

Análisis de datos agrupados desde el año 2013 al 2019

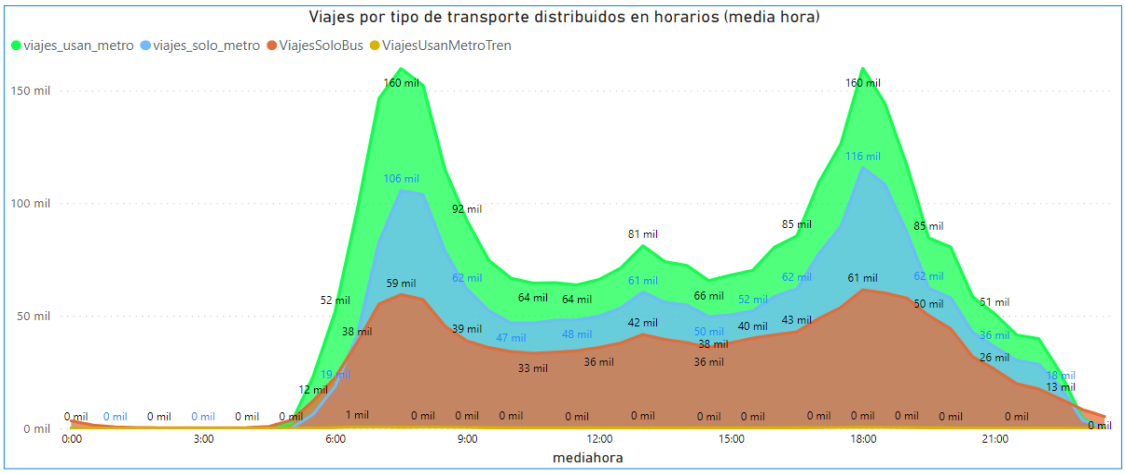
La figura 1 muestra que el año 2019, se evidenció un incremento en comparación con los años previos, especialmente en lo que respecta a los viajes realizados en el metro.

Figura 1. *Viaje promedio diario vs viajes solo en metro*



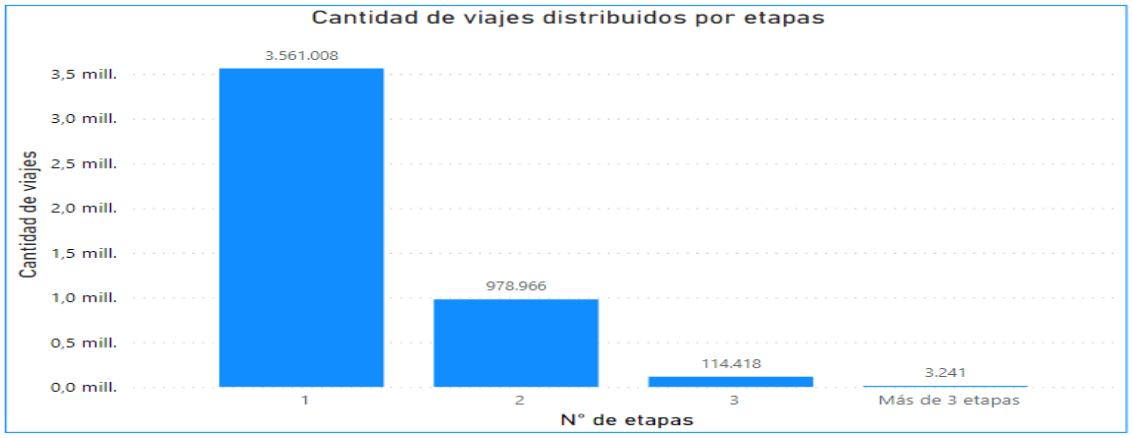
Los viajes en metro presentan los picos de viaje más altos durante los días laborales tal como se puede apreciar en la figura 2.

Figura 2. *Viajes por tipo de transporte distribuidos en horarios (media hora)*



Al analizar la distribución de estos viajes, se pudo constatar un patrón interesante. La mayoría de los viajes se llevaron a cabo en una sola etapa, este hallazgo se encuentra respaldado por los datos representados en la Figura 3, que ofrece una representación visual de esta tendencia.

Figura 3. *Cantidad de viajes por etapas*



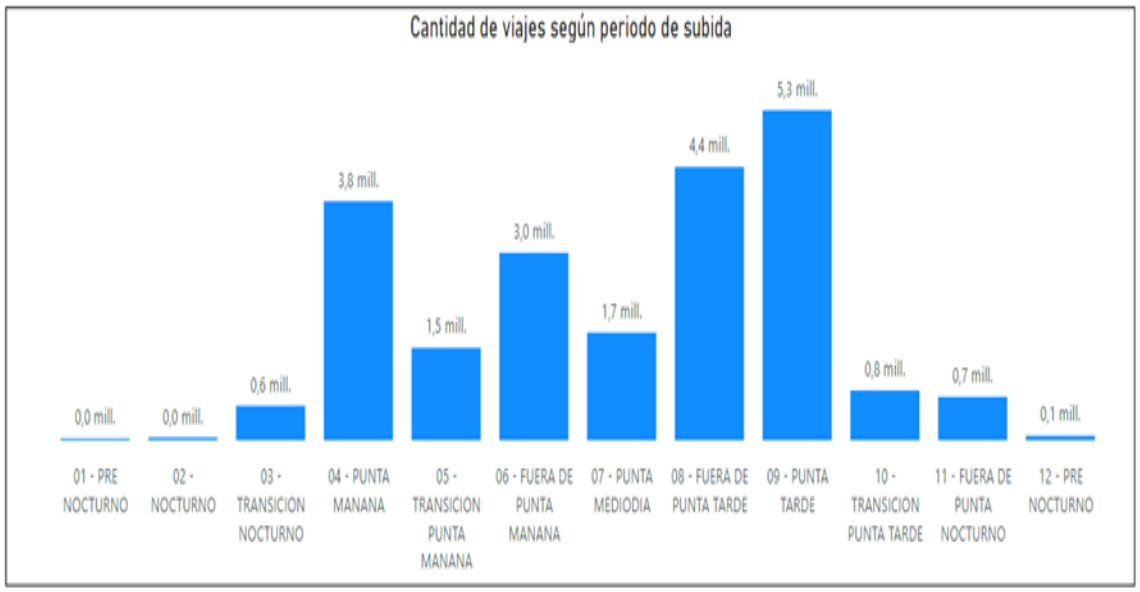
Análisis de la tabla de Viajes

En base al análisis realizado en cuanto a los datos agrupados y tabla de etapas, se elige el año 2019 para realizar el estudio, tomando como base la tabla de viajes del año 2019. Teniendo en cuenta la base de datos tabla de viajes se realizó un análisis de la cantidad de viajes realizados en días laborables, segmentados de acuerdo con los diferentes períodos de subida. Se observa un aumento significativo en los periodos catalogados como "Punta tarde", así como en los segmentos de "Fuera de punta tarde" y "Punta mañana". Para obtener información específica sobre los horarios comprendidos en cada uno de estos periodos, ver anexo B los horarios que incluyen estos periodos.

Base de datos año 2019: Se analiza la cantidad de viajes que se realizan en días laborales clasificados según los periodos de subida donde se observa un importante aumento en los periodos “Punta tarde”, junto a los periodos de “Fuera de punto tarde” y también en “Punta mañana”, ver anexo B los horarios que incluyen estos periodos, tal como se puede apreciar en la figura 4.

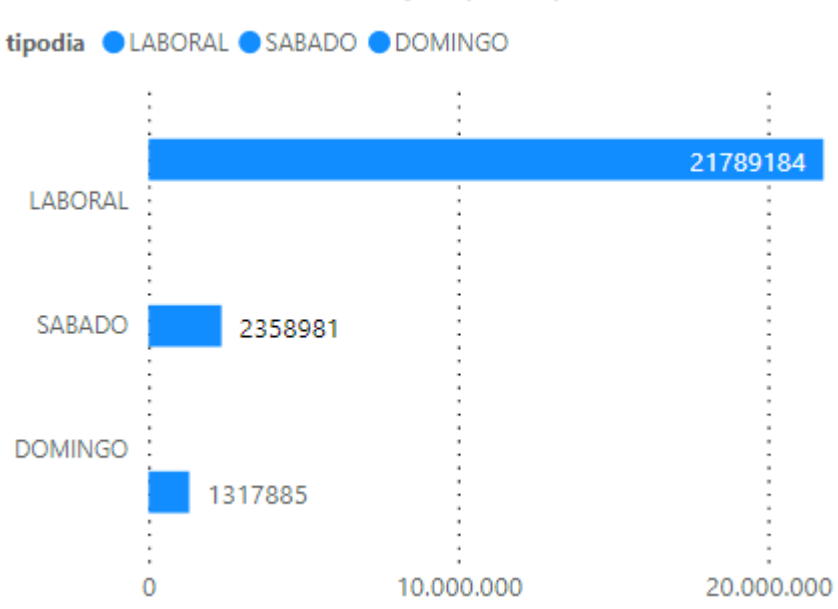
La cantidad de viajes en días laborales según los datos de esa semana fueron de 21.8 millones de viajes.

Figura 4 *Cantidad de viajes según periodo de subida*



Como se puede apreciar en la figura 5 el día laboral representa la mayor cantidad de viajes durante el año 2019 con un valor de 21.789.184.

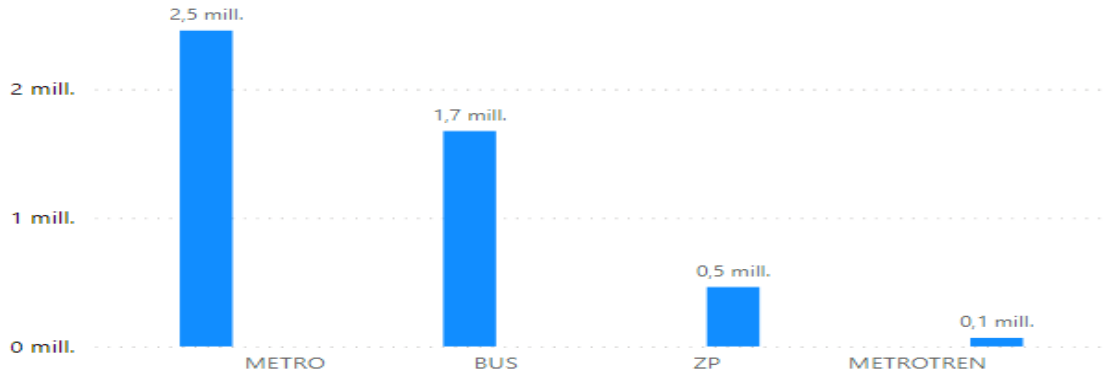
Figura 5. *Cantidad de viajes por tipo de día, año 2019*



Análisis por tipo de transporte

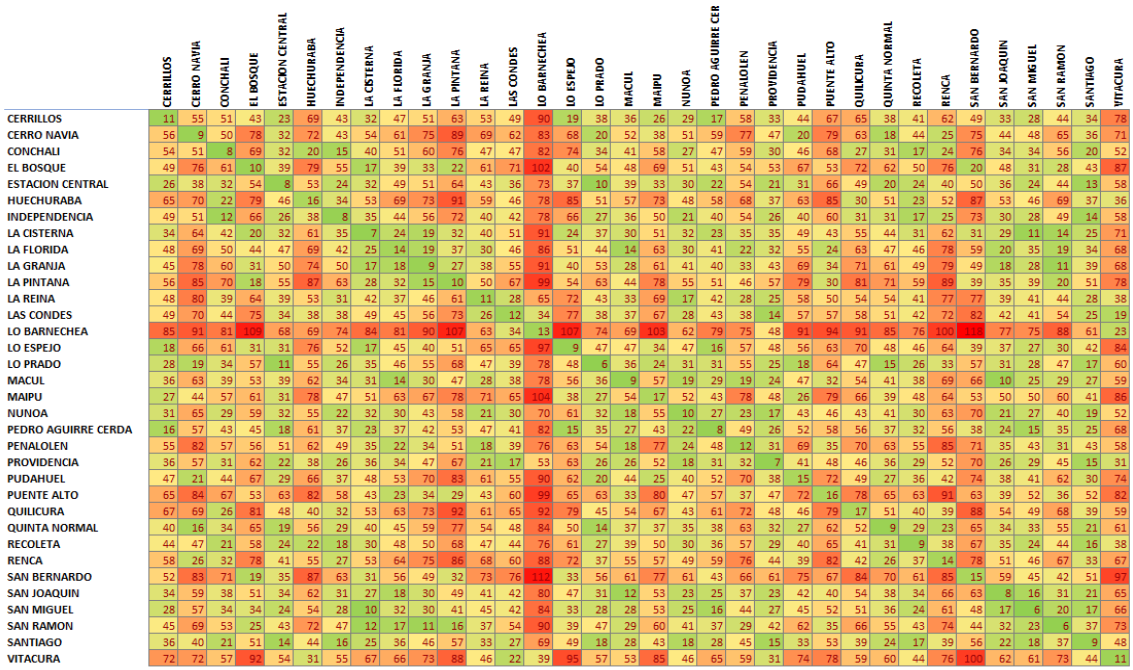
En la figura 6 se puede apreciar que la mayor cantidad de viajes diarios correspondió a la realizada en metro con un cantidad de 2.5 millones de viajes, seguida de en bus con 1,7 millones de viajes, por esta razón se confirma la relevancia que tienen los viajes en metro del cual es objeto el presente estudio.

Figura 6. *Cantidad de viajes diarios según tipo de transporte*



Entre las comunas que presentaron mayores tiempos de viaje se encuentra Barnechea, Vitacura y Maipu entre otras asi como se puede observar en la figura 7.

Figura 7. *Mapa de calor de tiempos de viaje entre comunas*



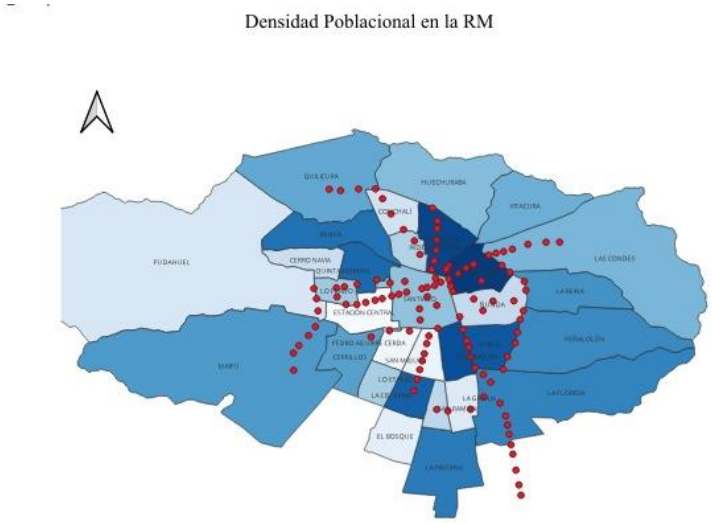
	CERRILLOS	CERRO NAVIA	CONCHALI	EL BOSQUE	ESTACION CENTRAL	HUECHURABA	INDEPENDENCIA	LA CISTERNA	LA FLORIDA	LA GRANJA	LA PINTANA	LA REINA	LAS CONDES	LO BARNECHEA	LO ESPEJO	LO PRADO	MACUL	MAIPU	NUNOA	PEDRO AGUIRRE CER	PENALOLEN	PROVIDENCIA	PUDAHUEL	PUENTE ALTO	QUILICURA	QUINTA NORMAL	RECOLETA	RENCA	SAN BERNARDO	SAN JOAQUIN	SAN MIGUEL	SAN RAMON	SANTIAGO	VITACURA
CERRILLOS	11	55	51	43	23	69	43	32	47	51	63	53	49	90	19	38	36	26	29	17	58	33	44	67	65	38	41	62	49	33	28	44	34	78
CERRO NAVIA	56	9	50	78	32	72	43	54	61	75	89	69	62	83	68	20	52	38	51	59	77	47	20	79	63	18	44	25	75	44	48	65	36	71
CONCHALI	54	51	8	69	32	20	15	40	51	60	76	47	47	82	74	34	41	58	27	47	59	30	46	68	27	31	17	24	76	34	34	56	20	52
EL BOSQUE	49	76	61	10	39	79	55	17	39	33	22	61	71	102	40	54	48	69	51	43	54	53	67	53	72	62	50	76	20	48	31	28	43	87
ESTACION CENTRAL	26	38	32	54	8	53	24	32	49	51	64	43	36	73	37	10	39	33	30	22	54	21	31	66	49	20	24	40	50	36	24	44	13	58
HUECHURABA	65	70	22	79	46	16	34	53	69	73	91	59	46	78	85	51	57	73	48	58	68	37	63	85	30	51	23	52	87	53	46	69	37	36
INDEPENDENCIA	49	51	12	66	26	38	8	35	44	56	72	40	42	78	66	27	36	50	21	40	54	26	40	60	31	31	17	25	73	30	28	49	14	58
LA CISTERNA	34	64	42	20	32	61	35	7	24	19	32	40	51	91	24	37	30	51	32	23	35	35	49	43	55	44	31	62	31	29	11	14	25	71
LA FLORIDA	48	69	50	44	47	69	42	25	14	19	37	30	46	86	51	44	14	63	30	41	22	32	55	24	63	47	46	78	59	20	35	19	34	68
LA GRANJA	45	78	60	31	50	74	50	17	18	9	27	38	55	91	40	53	28	61	41	40	33	43	69	34	71	61	49	79	49	18	28	11	39	68
LA PINTANA	56	85	70	18	55	87	63	28	32	15	10	50	67	99	54	63	44	78	55	51	46	57	79	30	81	71	59	89	39	35	39	20	51	78
LA REINA	48	80	39	64	39	53	31	42	37	46	61	11	28	65	72	43	33	69	17	42	28	25	58	50	54	54	41	77	77	39	41	44	28	38
LAS CONDES	49	70	44	75	34	38	38	49	45	56	73	26	12	34	77	38	37	67	28	43	38	14	57	57	58	51	42	72	82	42	41	54	25	19
LO BARNECHEA	85	91	81	109	68	69	74	84	81	90	107	63	34	13	107	74	69	103	62	79	75	48	91	94	91	85	76	100	118	77	75	88	61	23
LO ESPEJO	18	66	61	31	31	76	52	17	45	40	51	65	65	97	9	47	47	34	47	16	57	48	56	63	70	48	46	64	39	37	27	30	42	84
LO PRADO	28	19	34	57	11	55	26	35	46	55	68	47	39	78	48	6	36	24	31	31	55	25	18	64	47	15	26	33	57	31	28	47	17	60
MACUL	36	63	39	53	39	62	34	31	14	30	47	28	38	78	56	36	9	57	19	29	19	24	47	32	54	41	38	69	66	10	25	29	27	59
MAIPU	27	44	57	61	31	78	47	51	63	67	78	71	65	104	38	27	54	17	52	43	78	48	26	79	66	39	48	64	53	50	50	60	41	86
NUNOA	31	65	29	59	32	55	22	32	30	43	58	21	30	70	61	32	18	55	10	27	23	17	43	46	43	41	30	63	70	21	27	40	19	52
PEDRO AGUIRRE CERDA	16	57	43	45	18	61	37	23	37	42	53	47	41	82	15	35	27	43	22	8	49	26	52	58	56	37	32	56	38	24	15	35	25	68
PENALOLEN	55	82	57	56	51	62	49	35	22	34	51	18	39	76	63	54	18	77	24	48	12	31	69	35	70	63	55	85	71	35	43	31	43	58
PROVIDENCIA	36	57	31	62	22	38	26	36	34	47	67	21	17	53	63	26	26	52	18	31	32	7	41	48	46	36	29	52	70	26	29	45	15	31
PUDAHUEL	47	21	44	67	29	66	37	48	53	70	83	61	55	90	62	20	44	25	40	52	70	38	15	72	49	27	36	42	74	38	41	62	30	74
PUENTE ALTO	65	84	67	53	63	82	58	43	23	34	29	43	60	99	65	63	33	80	47	57	37	47	72	16	78	65	63	91	63	39	52	36	52	82
QUILICURA	67	69	26	81	48	40	32	53	63	73	92	61	65	92	79	45	54	67	43	61	72	48	46	79	17	51	40	39	88	54	49	68	39	59
QUINTA NORMAL	40	16	34	65	19	56	29	40	45	59	77	54	48	84	50	14	37	37	35	38	63	32	27	62	52	9	29	23	65	34	33	55	21	61
RECOLETA	44	47	21	58	24	22	18	30	48	50	68	47	44	76	61	27	39	50	30	36	57	29	40	65	41	31	9	38	67	35	24	44	16	38
RENCA	58	26	32	78	41	55	27	53	64	75	86	68	60	88	72	37	55	57	49	59	76	44	39	82	42	26	37	14	78	51	46	67	33	67
SAN BERNARDO	52	83	71	19	35	87	63	31	56	49	32	73	76	112	33	56	61	77	61	43	66	61	75	67	84	70	61	85	15	59	45	42	51	97
SAN JOAQUIN	34	59	38	51	34	62	31	27	18	30	49	41	42	80	47	31	12	53	23	25	37	23	42	40	54	38	34	66	63	8	16	31	21	65
SAN MIGUEL	28	57	34	34	24	54	28	10	32	30	41	45	42	84	33	28	28	53	25	16	44	27	45	52	51	36	24	61	48	17	6	20	17	66
SAN RAMON	45	69	53	25	43	72	47	12	17	11	16	37	54	90	39	47	29	60	41	37	29	42	62	35	66	55	43	74	44	32	23	6	37	73
SANTIAGO	36	40	21	51	14	44	16	25	36	46	57	33	27	69	49	18	28	43	18	28	45	15	33	53	39	24	17	39	56	22	18	37	9	48
VITACURA	72	72	57	92	54	31	55	67	66	73	88	46	22	39	95	57	53	85	46	65	59	31	74	78	59	60	44	76	100	62	61	73	44	11

La visualización del mapa de la región metropolitana de Santiago como se muestra en la figura 8, ofrece una perspectiva de la disposición geográfica de los paraderos de autobuses en relación con las diferentes estaciones del metro. Este análisis geoespacial no solo proporciona información sobre la ubicación física de estos nodos de transporte, sino que también presenta una oportunidad valiosa para examinar aspectos clave relacionados con la movilidad y la eficiencia del sistema de transporte público en la región.

Figura 8. *Mapa de los paraderos de buses con las estaciones de Metro en la Región Metropolitana de Santiago*

En la figura 9 se pudo apreciar la concentración de la población en las diferentes comunas de la región metropolitana, resaltado con color más intenso para establecer una relación con las ubicaciones de las distintas estaciones del metro.

Figura 9. *Densidad poblacional de la zona metropolitana de Santiago*



5.2. Metodología

5.2.1 Recopilación de datos

a) Se recopilaron los datos de la dirección Metropolitana de Transporte Público (DMTP) del año 2019, correspondiente a la semana del 20 de septiembre, solo días laborales.
b) En base a los datos del censo del año 2017, se obtuvo la cantidad de personas por zonas de cada comuna de la Región Metropolitana de Santiago de Chile y los años de estudio por zonas.
c) Se registraron los datos de la ubicación de los paraderos de transporte público y de las estaciones de metro distribuidos por zonas, fuente Open Street Map.

5.2.2 Preparación de los datos

Consolidamos estas tres fuentes de datos para formar una sola base maestra la cual vamos a utilizar para nuestro análisis. Para ello, se realizó el siguiente procesamiento de datos:

Se consideraron solo aquellos registros que tenían información sobre los paraderos de bajada y subida, excluyendo aquellos campos que no lo tuviesen tal como se puede apreciar en el Data_Frame construido en la tabla 1.

Tabla 1. *Carga de datos en un Data_Frame 1*

diseno777subida	diseno777bajada	periodosubida	netapa	Cant_Metros_zonas	CAnt_metros_comuna	Promedio de Promedio_Educacional	Promedio de Cant_Personas	tiempo_promedi
0	149	03 - TRANSICION NOCTURNO	3	NaN	5.0	11,7022	4824	51,3
0	149	04 - PUNTA MANANA	2	NaN	5.0	11,7022	4824	49,
0	150	03 - TRANSICION NOCTURNO	2	NaN	5.0	11,7022	4824	55,86111111111110
0	150	04 - PUNTA MANANA	2	NaN	5.0	11,7022	4824	46,47083333333333
0	150	06 - FUERA DE PUNTA MANANA	1	NaN	5.0	11,7022	4824	79,

5.2.3 Variables seleccionadas para el entrenamiento del modelo

Las variables que se utilizaron fueron:

Variables categóricas:

- ✓ **comunasubida:** Corresponde a la Comuna de subida:
- ✓ **comunabajada:** Comuna de Bajada:
- ✓ **Diseño777 Subida:** Zona de Subid
- ✓ **Diseño777Bajada:** Zona de Bajada
- ✓ **Periodosubida:** Son los horarios agrupados de transporte durante el día, se transformó en variable *dummy* (Pre Nocturno, Nocturno, Transición Nocturno, Punta Mañana, Transición Punta Mañana, Fuera de Punta Mañana, Punta Mediodía, Fuera Punta Tarde, Punta Tarde, Transición Punta Tarde y Fuera de Punta Nocturno).

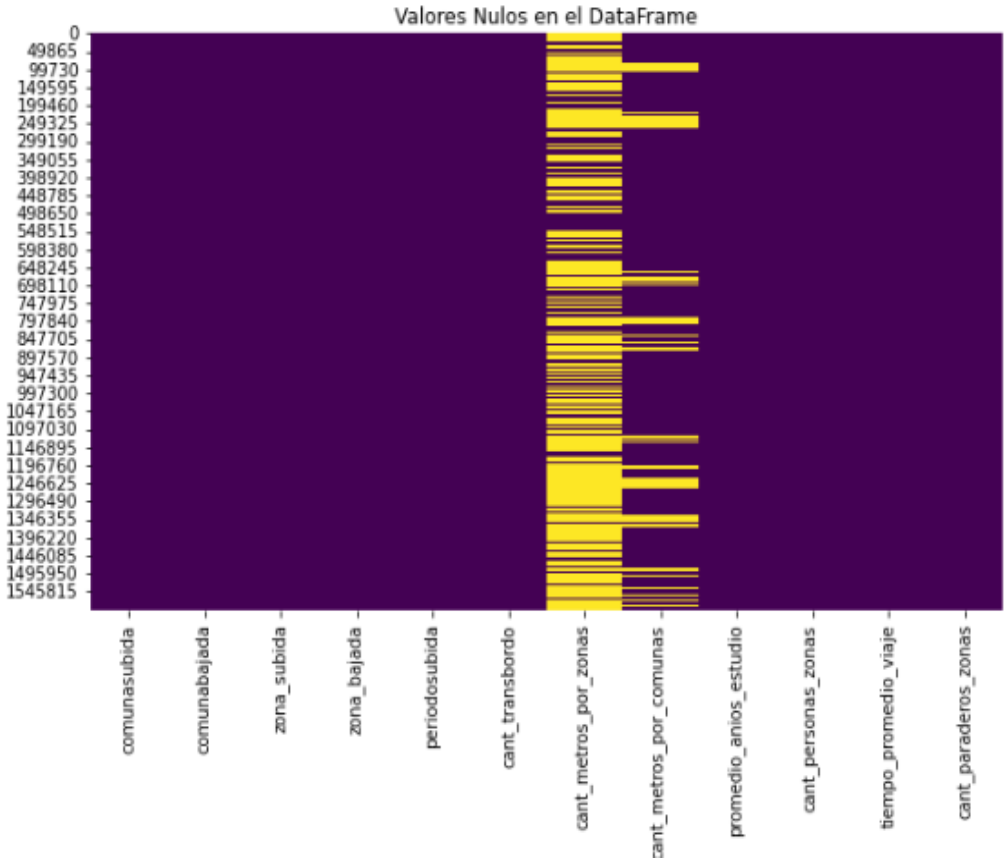
Variables numéricas:

- ✓ **netapas:** Corresponde a la cantidad de transbordo que realizaron los usuarios

- ✓ **Cant_metros_zonas:** Es la cantidad de metros por zonas.
- ✓ **Cant_metros_comuna:** Es la cantidad de metros por comuna.
- ✓ **Promedio de año de estudio:** Es la cantidad promedio de años de estudio.
- ✓ **Cantidad de personas por zonas.**
- ✓ **Cantidad de paraderos por zonas.**

Dentro de la base de datos se identificaron valores nulos, los cuales fueron convertidos en “0”, con el fin asegurar que el conjunto de datos fuera más consistente y fácil de trabajar, especialmente para poder ingresarlos a los algoritmos de predicción, en la figura 13 se aprecia la proporción de datos nulos de la data frame.

Figura 10. *Proporción de datos nulos*



Nota: La gráfica representa los valores nulos (1.315.240) encontrados con la sentencia isnull() que fueron reemplazados por el valor cero, correspondientes a la cantidad de metros por zonas.

Se convirtieron las variables zona subida y zona bajada a categóricas.

5.2.4 Análisis descriptivo

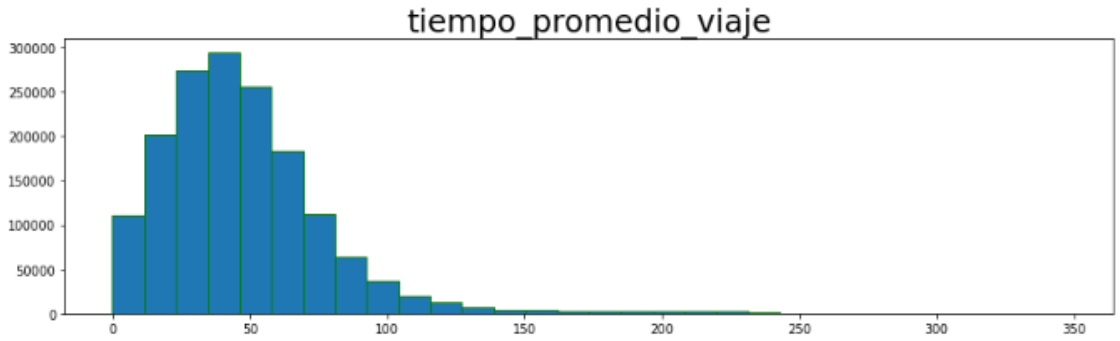
La variable objetivo (target) selecciona fue: "tiempo_promedio_viaje", la cual registra una media de 47 min aprox. con un mínimo de 0 min y máximo de 347 min, esta información se puede observar en la tabla 2.

Tabla 2. *Resumen estadísticas descriptivas*

	count	mean	std	min	25%	50%	75%	max
cant_transbordo	1505070.0	1.013000	0.705395	1.00	1.00	2.00	2.00	9.00
cant_metros_por_zonas	1505070.0	0.402011	0.605412	0.00	0.00	0.00	1.00	2.00
cant_metros_por_comunas	1505070.0	5.854087	5.770255	0.00	2.00	5.00	9.00	20.00
promedio_anios_estudio	1505070.0	13.554877	1.438092	10.09	12.51	13.26	14.52	17.09
cant_personas_zonas	1505070.0	3065.430798	720.834075	955.00	3184.00	3630.00	4146.00	6558.00
tiempo_promedio_viaje	1505070.0	47.110111	20.840091	-0.02	26.98	42.80	61.00	347.13
cant_paraderos_zonas	1505070.0	16.845050	8.210068	1.00	11.00	15.00	21.00	78.00

La variable target u objetivo presenta una distribución asimétrica hacia la derecha tal como se aprecia en la figura 11.

Figura 11. *Distribución variable target*

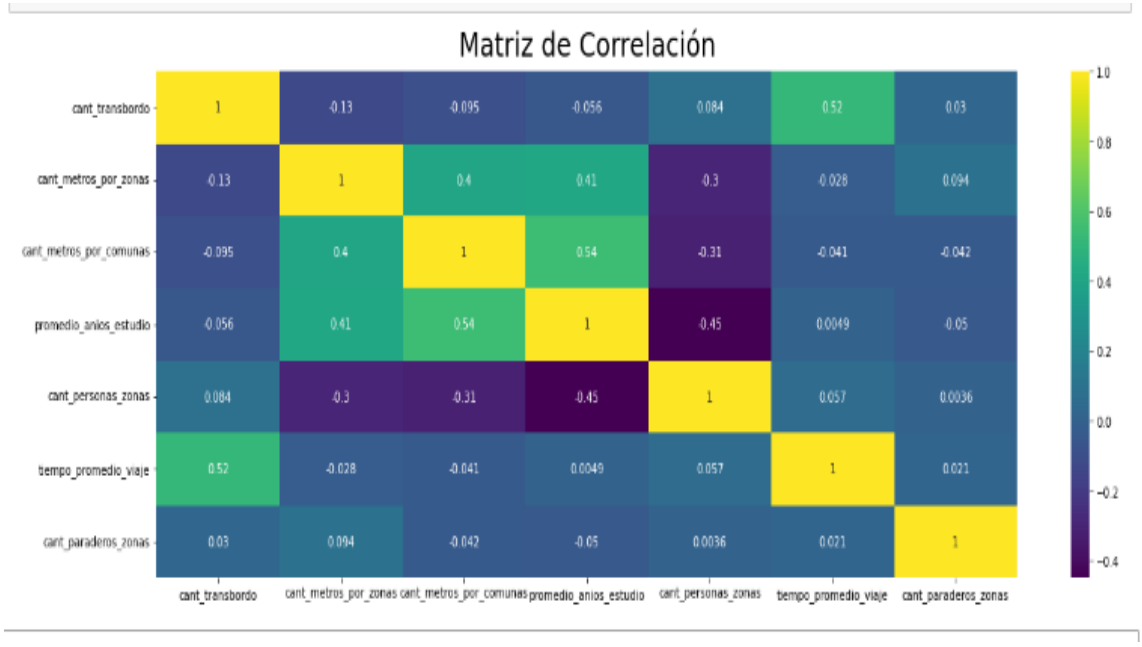


De igual manera, se realizó un análisis de las variables numéricas dentro de las cuales se encuentran:

- ✓ Cantidad promedio de estaciones de metro por zonas es de 1.1, destacando valores máximos de 2 metros por zona.
- ✓ La cantidad promedio de estaciones de metro por comunas es de 6.9 con un máximo de 20 estaciones de metro en alguna comuna.
- ✓ La cantidad promedio de transbordo o número de etapas de viaje es de 1.9, presentando un máximo de 9 etapas.
- ✓ Los años de estudio en promedio por zonas es de: 13.5, y fluctúan entre los 10 y 17 años.
- ✓ La cantidad de personas fue un promedio de 3665 y varían desde 955 hasta 6558.
- ✓ La cantidad de paraderos por zonas en promedio es: 17, con un mínimo de 1 hasta 78 paraderos.

En el proceso de preparación de datos se realizó una caracterización de las variables numéricas para lo cual se llevó a cabo el cálculo de una matriz de correlación con el fin de analizar las relaciones estadísticas entre las variables involucradas:

Figura 12. *Matriz de correlación variables numéricas*



En el análisis de la matriz de correlación ver figura 12, no se observó variables que tenga una correlación demasiado alta, lo cual evita que exista información redundante. Las variables que mejor correlación tuvieron son las de cantidad de transbordo con el tiempo promedio de viajes (0,52). Destacó también la de cantidad de metros por comunas y zonas con el promedio de años de estudio (0.54). Se observó una correlación inversa significativa entre la cantidad de personas por zonas y el promedio de años de estudio, indicando que a medida que aumenta la densidad poblacional en una zona, tiende a disminuir el nivel promedio de educación de las personas en esa área.

También se realizó un análisis descriptivo de las variables categóricas:

- ✓ Las 5 comunas que se destacaron por tener mayor demanda de usuarios de transporte públicos son: Santiago, Maipú, Puente Alto, Las Condes, La Florida.
- ✓ Las comunas de bajada tienen un comportamiento semejante a las comunas de subida, donde igualmente las comunas de Santiago, Maipú, Puente Alto, las Condes y la Florida presentan las frecuencias más altas.

Con el fin de facilitar la compatibilidad con los algoritmos de predicción que se van a utilizar, se realizó una trasformación de las variables categóricas (comuna_subida, comuna_bajada, zona_subida, zona_bajada) en numéricas teniendo en cuenta sus frecuencias.

Para el caso de la última variable categórica “Periodo de subida” se utilizó one hote encode usando la función de get.dummies, tal como se muestra en la tabla 3:

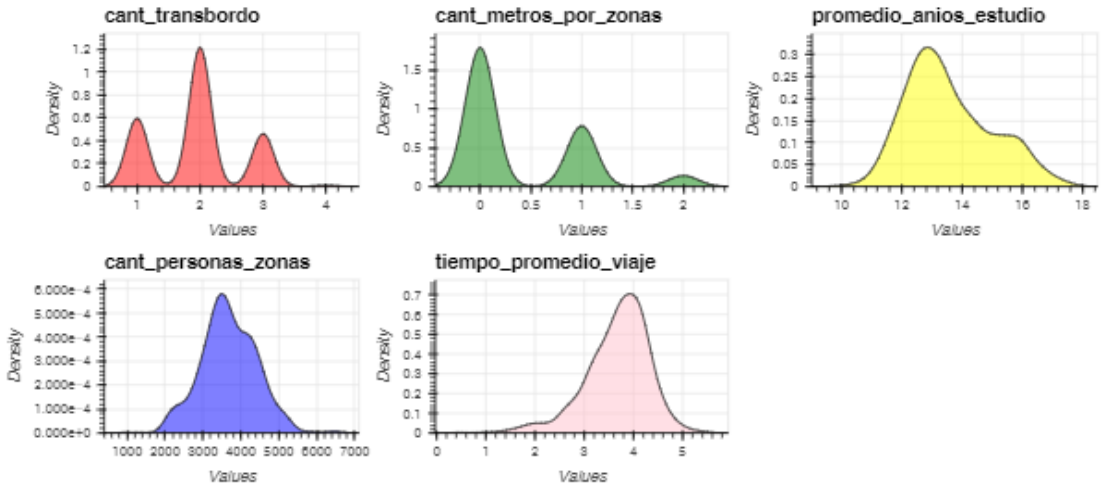
Tabla 3. *Transformación one hote encode*

	PS_02 - NOCTURNO	PS_03 - TRANSICION NOCTURNO	PS_04 - PUNTA MANANA	PS_05 - TRANSICION PUNTA MANANA	PS_06 - FUERA DE PUNTA MANANA	PS_07 - PUNTA MEDIODIA	PS_08 - FUERA DE PUNTA TARDE	PS_09 - PUNTA TARDE	PS_10 - TRANSICION PUNTA TARDE	PS_11 - FUERA DE PUNTA NOCTURNO	PS_12 - PRE NOCTURNO
0	0	1	0	0	0	0	0	0	0	0	0
1	0	0	1	0	0	0	0	0	0	0	0
2	0	1	0	0	0	0	0	0	0	0	0
3	0	0	1	0	0	0	0	0	0	0	0
4	0	0	0	0	1	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...
1595674	0	0	0	0	1	0	0	0	0	0	0
1595675	0	0	0	0	0	0	0	1	0	0	0
1595676	0	1	0	0	0	0	0	0	0	0	0
1595677	0	0	1	0	0	0	0	0	0	0	0
1595678	0	1	0	0	0	0	0	0	0	0	0

5.2.5 Distribución variables

A continuación, en la figura 13, se visualizó la distribución de algunas de las variables con el fin de identificar patrones en los datos, outliers con el fin de respaldar el análisis estadístico de los datos.

Figura 13. *Distribución de variables*

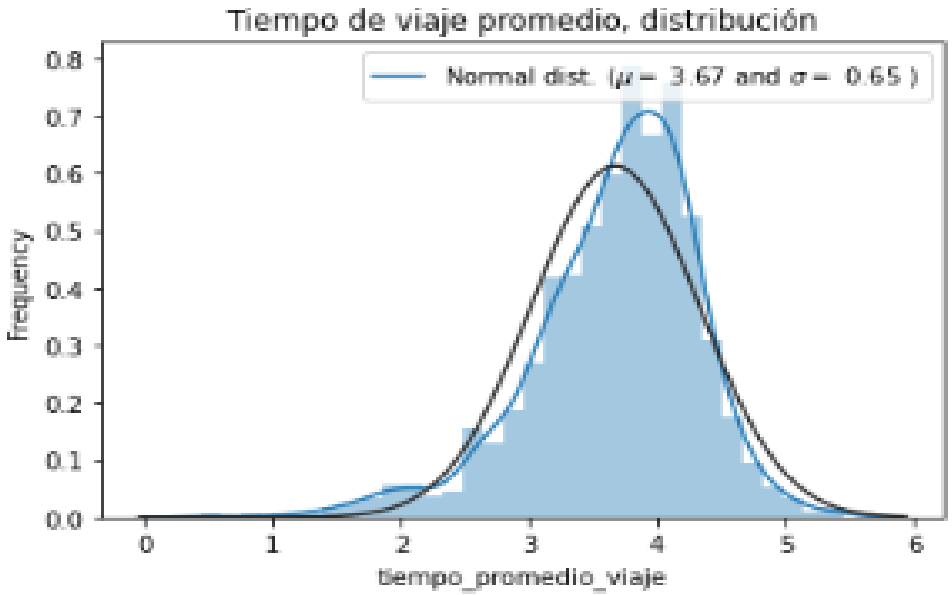


5.2.6 Normalización de variable Target “tiempo_promedio_viaje”

Con el fin de mejorar el rendimiento de la variable target se realizó un proceso de normalización de dicha variable, para esto se utilizó la función numpy log1p. Esta función aplica logaritmo natural a todos los elementos de la variable target, más 1, logrando el efecto de suavizar la distribución de la variable target, lo que ayudó a mejorar el rendimiento de los modelos de aprendizaje utilizados.

Luego de haber tenido una media = 46.26 y una desviación estándar = 27.79, el resultado obtenido con la aplicación de la normalización fue una media de = 3.67 y una sigma = 0.65, tal como se puede apreciar en la figura 14.

Figura 14. *Normalización variable tiempo de viaje*



5.3 Modelos de predicción

Los modelos de aprendizaje supervisado se caracterizan porque enseñan al algoritmo cómo realizar su trabajo (Rojas, 2020). con un conjunto de datos clasificados bajo una cierta apreciación o idea para encontrar patrones que puedan aplicarse en un análisis (Mueller & Massaron, 2016) y producir una salida que ya se conoce.

En la ejecución de este estudio, se exploraron cuidadosamente diversos algoritmos, incluyendo XGBoost, Random Forest Regressor, Random Forest con LightGBM y DecisionTreeRegressor. A continuación, se presenta cada uno de estos algoritmos, así como el proceso de elaboración realizado.

5.3.1 XG BOOST

XG BOOST es un algoritmo de aprendizaje supervisado (Ni, y otros, 2020) basado en un árbol de decisiones y utiliza un marco de potenciación de gradientes. Este algoritmo destaca por su capacidad para lograr predicciones precisas con un esfuerzo computacional relativamente bajo. En muchos casos, sus resultados son comparables o incluso superiores a los obtenidos por modelos más complejos, especialmente en situaciones que involucran datos heterogéneos (Vega, 2020).

A continuación, se presenta el desarrollo del algoritmo en primera medida se utilizaron los siguientes parámetros para la creación del algoritmo XG BOOST:

- n_estimators=200, max_depth=6, eta=0.1, subsample=0.5,min_child_weight= 6

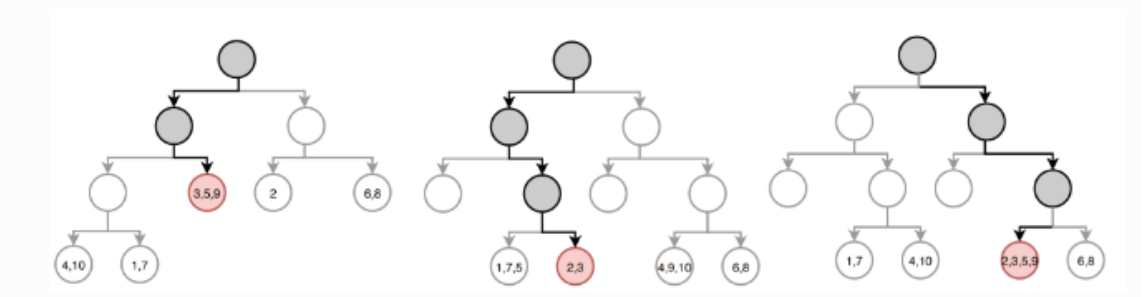
Se utilizó en primera instancia, la siguiente división de la data set (1.500.000 registros): 60% para el conjunto de datos de entrenamiento inicial y el resto para el conjunto de datos de prueba.

Como resultado se obtuvo un modelo con los siguientes coeficientes de R2 = 0.58 y MSE= 0.21.

5.3.2 RANDOM FOREST REGRESOR

Random Forest es un algoritmo de machine learning de uso común registrado por Leo Breiman y Adele Cutler, que une la salida de varios árboles de decisión para alcanzar un solo resultado. Cada uno de los árboles es entrenado con una muestra aleatoria extraída de los datos de entrenamiento, cada una de las observaciones se distribuyen en nodos generando una estructura de árbol. La predicción del modelo constituye un conjunto compuesto por las predicciones de cada uno de los árboles individuales (Rodrigo, Árboles de decisión, random forest, gradient boosting y C5.0, 2020).

Figura 15. *Estructura de árbol Random Forest*



Fuente: https://cienciadedatos.net/documentos/py08_random_forest_python

En cada nodo terminal se detalla el índice de las observaciones de entrenamiento que forman parte de él como se pudo observar en la figura 15.

Los parámetros utilizados para el modelo de Random forest fueron los siguientes:

- n_estimators=200, max_depth=20, random_state=1

Luego de ejecutar el modelo se obtuvieron los siguientes resultados R2=0.75 y MSE= 0.12.

5.3.3 Random Forest con LightGBM

LightGBM utiliza la técnica Gradient Boosting. este método permite que los árboles se construyan de manera secuencial y cada uno que se agrega aporta su granito de arena para refinar la predicción anterior. Es decir, se comienza con un valor constante y cada árbol nuevo se entrena para predecir el error en la suma de todas las predicciones de los árboles anteriores, utilizando el refuerzo de gradiente para construir árboles de manera secuencial, dándole prioridad a las instancias que contribuyen más al error del modelo. Puede sobre ajustar fácilmente datos pequeños debido a su sensibilidad. Se puede utilizar para datos que tengan más de 10.000 filas. No existe un umbral fijo que ayude a decidir el uso de LightGBM. Se puede utilizar para grandes volúmenes de datos, especialmente cuando se necesita lograr una alta precisión (Dwivedi, 2020).

La combinación de Ramdon Forest y LightGBM hace que se construyan árboles basado en gradientes, que prioriza las instancias que más contribuyen al error del modelo. Esta técnica, junto con la capacidad de Random Forest para reducir el sobreajuste manera mucho más eficiente, esto implica que el modelo sea más eficaz y rápido, optimizando la utilización de memoria; lo cual favorece en la manipulación de base de datos grandes, siendo de esta manera atractivo para implementarlo en el presente estudio.

Para el desarrollo del modelo se definieron los siguientes parámetros:

- boosting_type="rf", n_estimators=200, max_depth=10, colsample_bytree=0.8, subsample=0.8, subsample_freq=1, reg_alpha= 1, # L1 regularización, reg_lambda= 1, # L2 regularización

Las métricas obtenidas fueron: R2= 0.40 y MSE= 0.29.

5.3.4 Árboles de decisión (DecisionTreeRegressor.)

Los árboles de decisiones son modelos predictivos donde la variable destino puede tomar un conjunto de valores continuos se denominan árboles de regresión los cuales vamos a utilizar en el presente trabajo. Estos modelos toman decisiones mediante la construcción de un árbol estructurado compuesto por nodos, donde cada nodo representa una decisión basada en características específicas de los datos (Yáñez, 2017).

El árbol de decisiones es muy versátil y puede realizar varias funciones además de solo una predicción de resultados estándar. También sirve para explorar datos, resolver problemas de regresión y clasificación, así como para la segmentación de problemas. (Obando, 2023). Una de las ventajas que tienen los árboles de decisión ea que puede manejar variables categóricas como numéricas, dividiendo los nodos según reglas específicas para cada tipo de característica.

Para desarrollar el modelo de árboles de decisión, se empleó la clase DecisionTreeRegressor, configurando los siguientes parámetros:

- criterion='squared_error', max_depth=10,min_samples_leaf=2

Las métricas obtenidas fueron R2 = 0.66 y MSE= 0.40.

5.4 Comparación y Evaluación de modelos

Los modelos utilizados para realizar la predicción y comprobación de la hipótesis referente a la “adición de nuevas estaciones de metro de Santiago de Chile reducirá el tiempo promedio de viaje de los usuarios” fueron, XGBoost, Random Forest regressor, LightGBM y DecisionTreeRegressor.. Cada modelo fue configurado y entrenado utilizando un conjunto de datos seleccionado, de igual manera se realizaron pruebas para evaluar su capacidad predictiva. La comparación se centró en métricas clave de evaluación, como precisión, coeficiente de determinación (R2) y error cuadrático medio (RMSE). Se analizaron detalladamente las fortalezas y debilidades de cada modelo en términos de su capacidad para generalizar patrones a partir de los datos de entrenamiento y aplicarlos a nuevos datos de prueba.

En la tabla 4 se pueden observar los valores de R2 y RMSE de los cuatro modelos aplicados en la comprobación de la hipótesis. En la Figura 16 lado izquierdo, se observa claramente que el modelo con el R2 más destacado es el de Random Forest Regressor. Y en segundo lugar, se sitúa árbol de DecisionTreeRegressor.

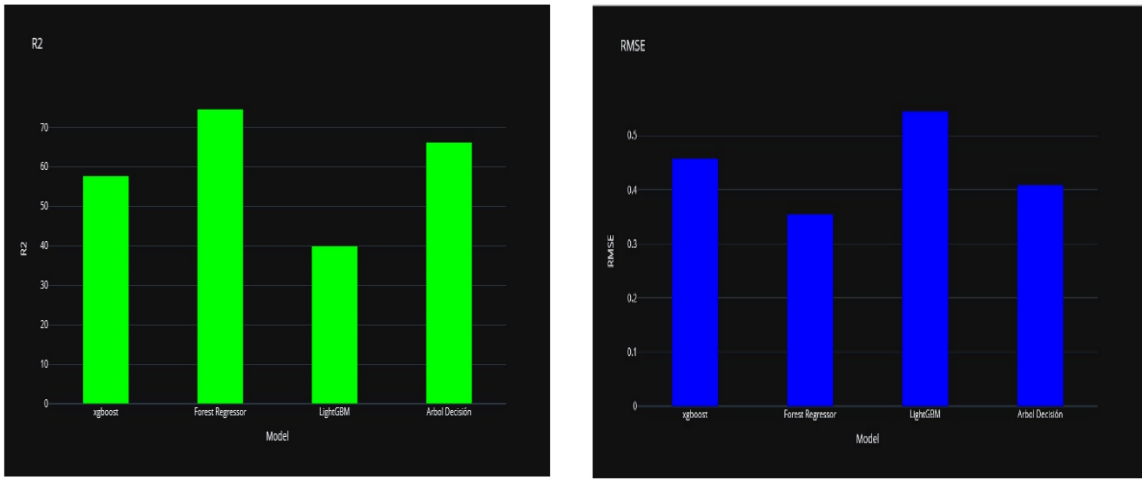
Como la opción más idónea para llevar a cabo la verificación de la hipótesis fue el modelo Random Forest Regressor.

En la figura 16 lado derecho se aprecian los resultados de la comparación del RMSE de cada uno de los modelos, donde el modelo Random Forest Regressor destaca por tener un error medio estándar más bajo sobre los demás modelos.

Tabla 4. *Comparación de R2 y RMSE de los modelos aplicados*

Model	R2	RMSE
xgboost	57.640501	0.457312
Forest Regressor	74.560570	0.354398
LightGBM	39.911580	0.544669
Arbol Decisión	66.168003	0.408697

Figura 16. *Comparación de resultados R2 y RMSE de los modelos aplicados*



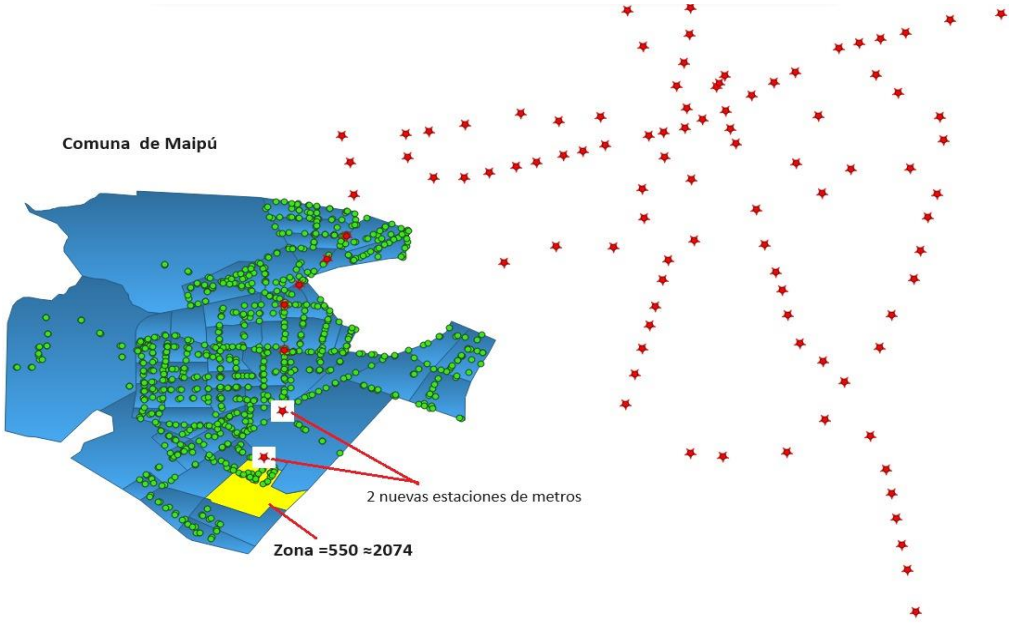
El modelo que decidimos utilizar para este estudio es el Forest Regressor debido a su mejor precisión y un menor RMSE.

5.5 Predicción

Para el proceso de la predicción se tomaron los modelos con mejor desempeño con análisis anterior. La predicción consistió en agregar valores a la variable estaciones de metros para validar si mejoraban los tiempos promedios de viaje. Se seleccionó la comuna de Maipú como ejemplo para esta simulación debido a que se caracteriza por ser una comuna por tener una gran cantidad de habitantes, menor cantidades de estaciones de metro y otra de las razones que motivo a elegir esa comuna fueron las declaraciones de su alcalde respecto de la necesidad de contar con más estaciones de metros para satisfacer la demanda de su población.[2]

En la figura 17 se puede apreciar una simulación visual de los ejercicios prácticos realizados con los modelos seleccionados.

Figura 17. *Visualización de mapa según pruebas de predicción*



[2] https://www.emol.com/noticias/Nacional/2023/08/10/1103724/alcalde-maipu-extension-metro-l9.html#:~:text=El%20alcalde%20apunta%20a%20que,con%2010%20y%2015%20respectivamente

6. Resultados

En seguida, se detallan los tres tipos de pruebas realizadas para evaluar la posible reducción en el tiempo promedio de viaje, en el caso de realizar variaciones en el número de metros. Estos análisis tienen como objetivo explorar cómo las fluctuaciones en la disponibilidad de metros pueden afectar la duración promedio de los viajes, proporcionando así una comprensión más profunda de las dinámicas de transporte en el contexto estudiado.

Seleccionamos tres viajes de los datos pertenecientes a la comuna de Maipú para evaluar la predicción de nuestro modelo y compararlo con los datos originales del tiempo promedio de viaje, esto se observa en la tabla 5.

Tabla 5. *Datos de prueba*

	comunasubida	comunabajada	zona_subida	zona_bajada	cant_transbordo	cant_metros_por_zonas	cant_metros_por_comunas	promedio_anios_estudi
1102257	105803	178544	2074	9921	2	0.0	5.0	11.
1102269	105803	178544	2074	9921	2	0.0	5.0	11.
1102258	105803	178544	2074	9921	2	0.0	5.0	11.

6.1 Resultados datos originales

Se realizaron tres casos cuyo valor del tiempo de viaje fue:

Caso N°1: El tiempo promedio de viaje original fue de: 51.6 minutos

Caso N°2: El tiempo promedio de viaje original fue de: 58.6 minutos

Caso N°3: El tiempo promedio de viaje original fue de: 56.2 minutos

Los datos de las otras variables se encuentran descritos en el anexo C.

6.2 Resultados de los tiempos de viaje aplicando el modelo Random Forest Regressor

Caso N°1: El tiempo promedio de viaje que predijo el modelo fue de 44.4 minutos. Al compararlo con el tiempo real que fue de 51.6 minutos observamos que hubo un 14 % menos que el tiempo real.

Caso N°2: El tiempo promedio de viaje que predijo el modelo fue de 51.8 minutos. Se observa un comportamiento similar al primer caso donde el tiempo real fue de 58.6 minutos prediciendo un 12 % menos del real.

Caso N°3: El tiempo promedio de viaje que predijo el modelo fue de 50.2 minutos. Se evidencio que el modelo predijo un 11% menos del real.

En la tabla 6 se observa un resumen de lo descrito anteriormente:

Tabla 6. *Resumen predicción vs datos reales*

N° de Casos	Datos Reales (minutos)	Predicción I (minutos)
Caso N°1	51,6	44,4
Caso N°2	58,6	51,8
Caso N°3	56,2	50,2

6.3 Resultados incorporando estaciones de metro

Ahora se ejecutó el modelo simulando dos estaciones nuevas de metro en la misma comuna de Maipú, a continuación, se modificaron los datos de las siguientes variables:

- Cantidad de metros por zona: Se agregó una estación de metro por zona.
- Cantidad de metros por comuna: Se agregaron dos estaciones de metro a las existentes sumando en total siete.
- Cantidad de trasbordos: Se disminuyeron la cantidad de trasbordos en uno.

Al ejecutar el modelo con las anteriores modificaciones obtuvimos los siguientes resultados, los cuales se aprecian en la tabla 7:

Tabla 7. *Tiempos de viaje predicción I vs predicción con nuevas estaciones de metro*

N° de Casos	Predicción I (minutos)	Predicción II estaciones más de metro (minutos)
Caso N°1	44,4	29,3
Caso N°2	51,8	32,6
Caso N°3	50,2	29,5

Una vez ejecutado el modelo simulando nuevas estaciones de metro, se puede apreciar que el tiempo promedio de viaje disminuyó considerablemente con respecto a la predicción la cual no incluye nuevas estaciones de metro.

En resumen, describimos en la tabla 8 la comparación de los resultados de los tres casos analizados anteriormente.

Tabla 8. *Tabla resumen que incluye nuevas estaciones de metro*

N° de Casos	Datos Reales (minutos)	Predicción I (minutos)	Predicción II estaciones más de metro (minutos)
Caso N°1	51,6	44,4	29,3
Caso N°2	58,6	51,8	32,6
Caso N°3	56,2	50,2	29,5

A continuación, se muestra en la tabla 9 los resultados de la variación del tiempo entre la predicción II (incorporación de nuevas estaciones de metro) y predicción I (sin nuevas estaciones de metro).

Tabla 9. *Resumen de los resultados al aplicar el modelo Random Forest Regressor*

N° de Casos	Datos Reales (minutos)	Predicción I (minutos)	Predicción II estaciones más de metro (minutos)	Variación del Tiempo de viaje entre ambas predicciones (minutos)
Caso N°1	51,6	44,4	29,3	**15,1**
Caso N°2	58,6	51,8	32,6	**19,2**
Caso N°3	56,2	50,2	29,5	**20,7**

7. Conclusiones

- De los cuatro modelos utilizados (XGBOOST, Random Forest Regressor, LightGBM y Árbol de Decisión Regressor) para la comprobación de la hipótesis, el que mejor desempeño tuvo fue **Random Forest Regressor** por ser más preciso y presentar menor porcentaje de error.
- Del total de las 21 variables de entrada que se utilizaron para el entrenamiento de los modelos, las de mayo relevancia fueron las comunas de subida, comuna de bajada y Zona de subida
- La variable de salida u objetivo fue: Tiempo promedio de viaje. Para ayudar a mejorar la precisión del modelo seleccionado, fue necesario ajustar la distribución de esta variable y así mejorar el rendimiento del modelo.
- Se validó la hipótesis que afirmaba que la incorporación de nuevas estaciones de metro que reduciría el tiempo de viaje de los usuarios. Realizando dos pasos importantes, uno mediante la predicción del tiempo de viaje entre zonas sin la inclusión de nuevas estaciones y luego aplicando el mismo modelo, pero incorporando las nuevas estaciones de metro. Los resultados confirmaron una disminución significativa en los tiempos de viaje.
- Las zonas de la comuna de Maipú nos sirvió de referencia para poner a prueba el modelo de predicción por ser una de las comunas de mayor tiempo de demora en los viajes demanda de transporte público por parte de su población y dada la escasez de estaciones de metro que actualmente presenta.
- Las pruebas realizadas consistieron inicialmente en seleccionar zonas geográficas que no tenían estaciones de metro para luego adicionar una estación de metro y comprobar la variación del tiempo promedio de viaje, pero no hubo

una diferencia considerable por lo que se tomó la decisión de modificar las variables de cantidad de transbordo y también la cantidad de paraderos lo cual evidenció una disminución en los tiempos de viaje.

- Los resultados mostraron que los tiempos de viaje en los tres casos analizados al incorporar dos nuevas estaciones de metro, disminuyeron entre 15 y 20 minutos.

8. Limitaciones y Trabajos futuros

- Se intentó utilizar otros modelos de aprendizaje supervisado, como Super Vector Regression (SVR) SVR. Sin embargo, debido a las limitaciones de capacidad de nuestros notebooks, no fue posible llevar a cabo su ejecución.
- Si bien realizamos validación cruzada con Random Forest, sería interesante ería aplicar esto a otros modelos para optimizar la selección de los hiperparámetros.
- Como trabajo futuro se sugiere la incorporación de otras variables como centros comerciales, financieros y salud para evaluar su posible incidencia en la reducción del tiempo promedio de viaje.
- Asimismo, sería beneficioso contar con datos actualizados sobre estadísticas de evasión del pago en el transporte público.

  Contar con datos acerca del impacto ambiental en la construcción de nuevas estaciones de metro y disponer de tasas de crecimiento de la población actualizadas. Estos elementos complementarios podrían enriquecer y ampliar la comprensión de los factores influyentes en el sistema de transporte estudiado.

- Sería conveniente utilizar algoritmos de redes neuronales para mejora el umbral de predicción, como pueden ser redes neuronales convolucionales (CNN) o redes neuronales recurrente (RNN)

9. Bibliografía

Alcalá, A. (10 de febrero de 2017). *Banco de desarrollo de América Latina el caribe*. Recuperado el 2023, de https://www.caf.com/es/conocimiento/visiones/2017/02/transporte-publico-en-america-latina-es-posible-un-cambio-de-paradigma/

Bates, J., Polak, J., Jones, P., & Cook, A. (2001). The valuation of reliability for personal travel. *Transportation Research Part E: Logistics and Transportation Review*, 191-229. Recuperado el 2023, de https://www.sciencedirect.com/science/article/pii/S1366554500000119

Durán, E. A. (2016). Caractarización de la variabilidad del tiempo de viaje en la ciudad de Santiago. 46-52. Santiago, Chile. Recuperado el 2023, de https://repositorio.uchile.cl/bitstream/handle/2250/139518/Caracterizacion-de-la-variabilidad-del-tiempo-de-viaje-en-la-ciudad-de-Santiago.pdf?sequence=1

Dwivedi, R. (26 de junio de 2020). Obtenido de Analitic steps: https://www.analyticssteps.com/blogs/what-light-gbm-algorithm-how-use-it

Eliasson, J. (2022). Transportation Research Interdisciplinary Perspectives. *Science Direct*, 2-3. Recuperado el 2023, de https://www.sciencedirect.com/science/article/pii/S2590198221002141

Gorman, K. (2009). Hierarchical regression modeling for language research. *Institute for Research in Cognitive Science*, 8-14. Recuperado el 2023, de https://www.academia.edu/452296/Hierarchical_Regression_Modeling_for_Language_Research

Gotschi, T. (2011). Costs and benefits of bicycling investments in Portland. *Journal of Physical Activity*, 49-50. Recuperado el 2023, de http://www.healthyweld2020.com/assets/7311d20DD1b9CCdc74c8.pdf

Graells-Garrido, E., Opitz, D., Rowe, F., & Arriagada, J. (2023). A data fusion approach with mobile phone data for updating travel survey-based mode split estimates. *Transportation Research Part C: Emerging Technologies*, 1-3. Recuperado el 2023, de https://www.sciencedirect.com/science/article/pii/S0968090X23002747

Krekel, C., Kolbe, J., & Wüstemann, H. (2015). The effect of urban land use on residential well-being. *Ecological economics*, 12-20. Recuperado el 2023, de https://api-depositonce.tu-berlin.de/server/api/core/bitstreams/7db58f7c-ed2d-4b42-b0e4-df11d08d2395/content

Krizec, K. (2016). Estimating the economic benefits of bicycling and bicycle facilities: an interpretive review and proposed methods. *Eure*, 133-152. Recuperado el 2023, de https://www.redalyc.org/journal/196/19655175007/html/#redalyc_19655175007_ref18

Miguel, M., Pedro, J., De Blas, S., Jimenez Barandalla, C., & Iciar, C. (2014). Estudio empírico sobre la utilización del transporte público en la Comunidad de Madrid como factor clave de movilidad sostenible. *Cuadernos de Economía*, 112-124. doi:10.1016/j.cesjef.2013.12.001

Miralles-Guasch, C. (2002). *Ciudad y transporte.* España: Ariel S.A. Recuperado el 2023, de https://www.researchgate.net/publication/40700310_Movilidad_cotidiana_y_sostenibilidad_una_interpretacion_desde_la_geografia_humana

Mueller, J. P., & Massaron, L. (2016). *Machine learning for dummies.* New Jersey: ohn Wiley & Sons, Inc. Recuperado el 2023

Ni, L., Wang, D., Wu, J., Wang, Y., Tao, Y., Zhang, J.-y., & Liu, J. (01 de 03 de 2020). Streamflow forecasting using extreme gradient boosting model coupled with Gaussian mixture model. *Journal of Hydrology*. Recuperado el 2023

Obando, R. (21 de Enero de 2023). *Blogspot*. Recuperado el 2023, de https://blog.hubspot.es/sales/arbol-decisiones

Ortega, S. F. (2018). Análisis del comportamiento del transporte público a nivel mundial. *Espacios*, 4-10. Recuperado el 2023, de https://www.researchgate.net/publication/325130617_Analisis_del_transporte_publico_a_nivel_mundial

Pezoa, R., Basso, F., & Frez, J. (2023). Crowding on public transport using smart card data during the COVID-19 pandemic: New methodology and case study in Chile. *ScienceDirect*, 1-11. Recuperado el 2023, de https://www.sciencedirect.com/science/article/pii/S2210670723003232

Rodrigo, J. A. (1 de octubre de 2020). *Árboles de decisión, random forest, gradient boosting y C5.0*. Recuperado el 2023, de https://www.cienciadedatos.net/documentos/33_arboles_decision_random_forest_gradient_boosting_C50.html

Rodrigo, J. A. (septiembre de 2023). *Random Forest con Python*. Obtenido de Ciencia de datos. net: https://cienciadedatos.net/documentos/py08_random_forest_python

Rojas, E. M. (2020). Machine Learning: análisis de lenguajes de programación y herramientas para desarrollo. *Revista Iberica de Sistemas y tecnologias de la*

información, 586-599. Recuperado el 2023, de https://www.proquest.com/openview/c7e24c997199215aa26a39107dd2fe98/1?pq-origsite=gscholar&cbl=1006393

Rushikesh, P. d. (16 de junio de 2018). *Medium*. Obtenido de https://towardsdatascience.com/https-medium-com-pupalerushikesh-svm-f4b42800e989

Vega, J. B. (12 de Agosto de 2020). *XGBoost en Python*. Recuperado el 2023, de https://medium.com/@jboscomendoza/tutorial-xgboost-en-python-53e48fc58f73

Yáñez, R. (2017). *Aprendizaje automatico: Estado de la cuestión y casos de estudio.* Madrid. Recuperado el 2023

10. Anexos

Anexo A

Tablas de Procesamiento de datos

Las Tablas A.1, A.2 y A3 presentan los atributos y una corta descripción de su contenido para las tablas resultantes del pre procesamiento de las transacciones al ser agrupadas por etapas y viajes respectivamente.

Tabla A.1 Tabla de viajes

Variable	Descripción
id	Número identicador de la tarjeta inteligente
nviaje	Correlativo asociado al viaje de la tarjeta
netapa	Correlativo asociada a la etapa dentro del viaje
netapassinbajada	Cantidad de etapas sin una estimación de bajada
ultimaetapaconbajada	1: En caso que la última etapa posea una estimación de
bajada.	0: Si no.
etapas	Resumen de los servicios utilizados en cada etapa del viaje
contrato	Tipo de contrato de la tarjeta
tipodia	Laboral, Sábado o Domingo
mediahora	Periodos de 30 minutos durante un día al que pertenece la hora de la transacción
paraderosubida	Paradero o estación de metro de inicio del viaje
paraderobajada	Paradero o estación de metro del destino nal estimado del viaje
comunasubida	Comuna donde se realizó la transacción de inicio del viaje
comunabajada	Comuna donde se estimo el n del viaje
tiemposubida	Fecha y hora de la transacción de inicio del viaje
tiempobajada	Fecha y hora estimada del n del viaje
periodosubida	Periodo denido por Transantiago al que pertenece la hora de la transacción de inicio del viaje
periodobajada	Periodo denido por Transantiago al que pertenece la hora estimada del n del viaje
dviajeeuclidiana	Distancia euclidiana del total del viaje
dviajeenruta	Distancia sobre la ruta del total del viaje
tviaje_min	Tiempo total del viaje
t_X_etapa	Tiempo total de la etapa X
d_X_etapa	Distancia euclidiana de la etapa X
ttrasbordo_X_etapa	Tiempo de transbordo de la etapa X
tcaminata_X_etapa	Tiempo de caminata de la etapa X
tespera_X_etapa	Tiempo de espera de la etapa X
serv_X_etapa	Servicio y sentido de la etapa X
tipotransporte_X	Tipo de transporte de la etapa X

Tabla A.2: Tabla de etapas

Variable	Descripción
id	Número identicador de la tarjeta
nviaje	Correlativo asociado al viaje de la tarjeta
netapa	Correlativo asociado a la etapa dentro del viaje
tipo_trasporte	Bus, Metro o Zona Paga
serv_un_zp2	Código que especica el servicio y su sentido
tipo_dia	Laboral, Sábado o Domingo
mediahora	Periodos de 30 min al que pertenece la hora de la transacción
tiempo_subida	Fecha y hora de la transacción
tiempo_bajada	Fecha y hora de la bajada estimada
x_subida	Coordenada X de la transacción
y_subida	Coordenada Y de la transacción
par_subida	Paradero o estación de metro de inicio de la etapa
par_bajada	Paradero o estación de metro de termino estimado de la etapa
comuna_subida	Comuna donde se ubica el origen de la etapa
comuna_bajada	Comuna donde se ubica el destino estimado de la etapa
tiempo_etapa	Tiempo total entre la una transacción y la siguiente
tiempo_trasbordo	Tiempo entre la bajada estimada y la siguiente transacción del mismo viaje
tiempo_caminata	Tiempo necesario caminando entre la bajada estimada y el paradero de la siguiente etapa del mismo viaje
tiempo_espera	Tiempo de transbordo menos el tiempo de caminata
disteuclid	Distancia euclidiana entre el paradero de origen y el paradero estimado de destino
distonroute	Distancia sobre la ruta entre el paradero de origen y el paradero estimado de destino

Tabla A.3: Tabla agregadas de Matriz OD, subidas y bajas por comunas

Variable	Descripción
comunasubida	Comuna del paradero de subida de la primera etapa
comunabajada	Comuna del paradero de bajada de la última etapa
mediahora	Periodos de 30 minutos durante un día al que pertenece
viaje_laboral_promedio	Viajes laborales promedio de la suma de todas las etapas
viajes_1_etapa	Viajes promedio 1 etapa
viajes_2_etapas	Viajes promedio 2 etapa
viajes_3_etapas	Viajes promedio 3 etapa
viajes_4_etapas	Viajes promedio 4 etapa
viajes_5omas_etapas	Viajes promedio de 5 o más etapa
viajes_usan_metro	Viajes don usan metro
ViajesUsanMetroTren	Viajes que usan metro Tren
viajes_solo_metro	Viajes solo metro
ViajesSoloBus	Viajes solo bus

Anexo B

Periodos definidos por Transantiago

Las Tablas B.1, presentan los periodos definidos por Transantiago para caracterizar la oferta y la demanda por transporte público en Santiago para los días laborales.

Tabla B.1: Periodos Transantiago en días laborales

Horario	Periodo
00:00 - 01:00	01 - Pre Nocturno
01:00 - 05:30	02 - Nocturno
05:30 - 06:30	03 - Transición Nocturno
06:30 - 08:30	04 - Punta Mañana
08:30 - 09:30	05 - Transición de punta Mañana
09:30 - 12:30	06 - Fuera de Punta Mañana
12:30 - 14:00	07 - Punta Mediodía
14:00 - 17:30	08 - Fuera de Punta Trade
17:30 - 20:30	09 - Punta Trade
20:30 - 21:30	10 - Transición Punta Tarde
21:30 - 23:00	11 - Fuera de Punta Nocturno
23:00 - 00:00	12 - Pre Nocturno

ANEXO C. Casos de prueba

Caso N° 1

comunasubida	105803.000000
comunabajada	178544.000000
zona_subida	2074.000000
zona_bajada	9921.000000
cant_transbordo	2.000000
cant_metros_por_zonas	0.000000
cant_metros_por_comunas	5.000000
promedio_anios_estudio	11.600000
cant_personas_zonas	3814.000000
tiempo_promedio_viaje	3.963096
cant_paraderos_zonas	22.000000
PS_02 - NOCTURNO	1.000000
PS_03 - TRANSICION NOCTURNO	0.000000
PS_04 - PUNTA MANANA	0.000000
PS_05 - TRANSICION PUNTA MANANA	0.000000
PS_06 - FUERA DE PUNTA MANANA	0.000000
PS_07 - PUNTA MEDIODIA	0.000000
PS_08 - FUERA DE PUNTA TARDE	0.000000
PS_09 - PUNTA TARDE	0.000000
PS_10 - TRANSICION PUNTA TARDE	0.000000
PS_11 - FUERA DE PUNTA NOCTURNO	0.000000
PS_12 - PRE NOCTURNO	0.000000

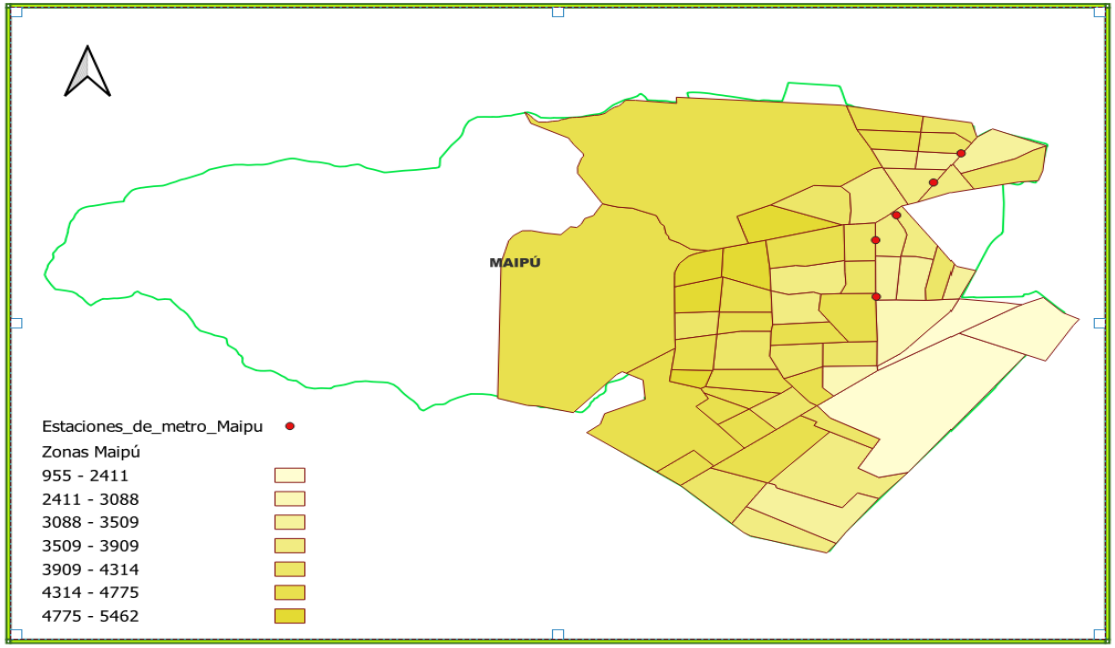
Caso N° 2

comunasubida	105803.000000
comunabajada	178544.000000
zona_subida	2074.000000
zona_bajada	9921.000000
cant_transbordo	2.000000
cant_metros_por_zonas	0.000000
cant_metros_por_comunas	5.000000
promedio_anios_estudio	11.600000
cant_personas_zonas	3814.000000
tiempo_promedio_viaje	4.046903
cant_paraderos_zonas	22.000000
PS_02 - NOCTURNO	0.000000
PS_03 - TRANSICION NOCTURNO	1.000000
PS_04 - PUNTA MANANA	0.000000
PS_05 - TRANSICION PUNTA MANANA	0.000000
PS_06 - FUERA DE PUNTA MANANA	0.000000
PS_07 - PUNTA MEDIODIA	0.000000
PS_08 - FUERA DE PUNTA TARDE	0.000000
PS_09 - PUNTA TARDE	0.000000

Caso N°3

comunasubida	105803.000000
comunabajada	178544.000000
zona_subida	2074.000000
zona_bajada	9921.000000
cant_transbordo	2.000000
cant_metros_por_zonas	0.000000
cant_metros_por_comunas	5.000000
promedio_anios_estudio	11.600000
cant_personas_zonas	3814.000000
tiempo_promedio_viaje	4.088829
cant_paraderos_zonas	22.000000
PS_02 - NOCTURNO	0.000000
PS_03 - TRANSICION NOCTURNO	0.000000
PS_04 - PUNTA MANANA	0.000000
PS_05 - TRANSICION PUNTA MANANA	0.000000
PS_06 - FUERA DE PUNTA MANANA	0.000000
PS_07 - PUNTA MEDIODIA	0.000000
PS_08 - FUERA DE PUNTA TARDE	0.000000
PS_09 - PUNTA TARDE	1.000000
PS_10 - TRANSICION PUNTA TARDE	0.000000
PS_11 - FUERA DE PUNTA NOCTURNO	0.000000
PS_12 - PRE NOCTURNO	0.000000

ANEXO D: Zona de comuna de Maipú



Anexo E

Link del Notebbok:

https://drive.google.com/file/d/1ZMvpGKPaXZ5VlsoctkciOyW9TSN72HT9/view?usp=sharing