Reporte Final

David Guzmán Leyva *Matrícula: A01706417*
Enrique Santos Fraire *Matrícula: A01705746*
Leonardo Alvarado Menéndez *Matrícula: A01705998*
Oscar Enrique Delgadillo Ochoa *Matrícula: A01705935*

I. Introducción

Las encuestas de movilidad realizadas previamente para conocer el movimiento de los ciudadanos de la zona metropolitana de Santiago de Chile contaban con algunas complicaciones, como:

- El costo elevado de realización de encuestas.
- El rechazo de las personas a colaborar en estos instrumentos.
- La extensa duración de la etapa de levantamiento de información [2].

Debido a esto se optó por utilizar datos pasivos captados mediante dispositivos móviles con el fin de obtener información de una manera más eficiente. De esta forma se pudo estimar y visualizar los Viajes Origen-Destino de los usuarios de la ciudad de Santiago de Chile en su región metropolitana.

Dichos datos se obtuvieron de la empresa de telefonía móvil Movistar, y se trabajaron en conjunto con el centro de investigación de Data Science de Chile. Los objetivos a los que llegamos con la información disponible son:

- Determinar las comunas con una mayor presencia de home office.
- Identificar factores determinantes que sean indicadores porcentuales de presencia de home office.

Estos objetivos ayudaron a predecir, según las características de cada comuna, cuánto home-office hay y qué factores afectan más a que se dé. Con esto el gobierno se puede anticipar y adecuar las comunas para que se mantenga el home-office sin complicaciones, y se tomen las medidas necesarias.

II. Estado del arte

A lo largo de la realización del proyecto hicimos uso de CRISP-DM, siglas para Cross-Industry Standard Process for Data Mining, una metodología para orientar trabajos de minería de datos [3].

Posteriormente para el preprocesado de los datos se realizó un análisis exploratorio de datos o EDA para revisar las principales características que se presentan en el dataset proporcionado por el socio formador [4]. Una vez teniendo un entendimiento completo de los datos y determinar lo que para nosotros era información relevante, continuamos con un proceso de ETL por sus siglas en inglés Extract Transform Load [5], para limpiar nuestros datos y transformarlos a algo que pudiéramos usar en un modelo de Machine Learning, que es la forma en la que un sistema puede aprender de estos [6].

En nuestro caso, se optó por usar la herramienta de Pyspark [7] por la gama de algoritmos de machine learning que posee en su librería ml para generar el modelo de predicción que se describe en la fase de método con más detalle. Se generaron 2 modelos de machine learning [6] y 1 de deep learning, que es un subconjunto de machine learning donde algoritmos inspirados en el funcionamiento del cerebro humano aprenden de grandes cantidades de datos [8].

Gracias a nuestra socia formadora y a investigaciones realizadas por nuestra cuenta, sabemos que otros proyectos de movilidad urbana se han efectuado en Santiago de Chile. Por ejemplo, el Data Science Institute de Las Condes 7610658, Chile [9], realizó un trabajo de investigación en el que, por medio de información de los dispositivos móviles conectados a las antenas telefónicas en Santiago de Chile, lograron obtener lo siguiente:

- Mejoraron la cobertura en la ciudad de datos CDR al geolocalizar dispositivos en más áreas de la ciudad que usando métodos estándar.
- Encontraron lugares importantes (hogar y trabajo) para un 10 % de la muestra usando sólo información diaria y recrearon la distribución de la población, así como los viajes de ida y vuelta que hacen los usuarios durante el día.
- A partir de los datos generados pueden ayudar a ubicar lugares, rutas, ubicación de tiendas minoristas y estimación de efectos de transporte a partir de alertas de contaminación.

Por nuestra parte, el propósito del proyecto fue conocer las causas principales por las que se genera home-office, esto con el fin de que en un futuro, en caso de ser necesario, un gobierno pueda predecir la cantidad de home-office que hay en la comuna o municipio y tomar las acciones necesarias.

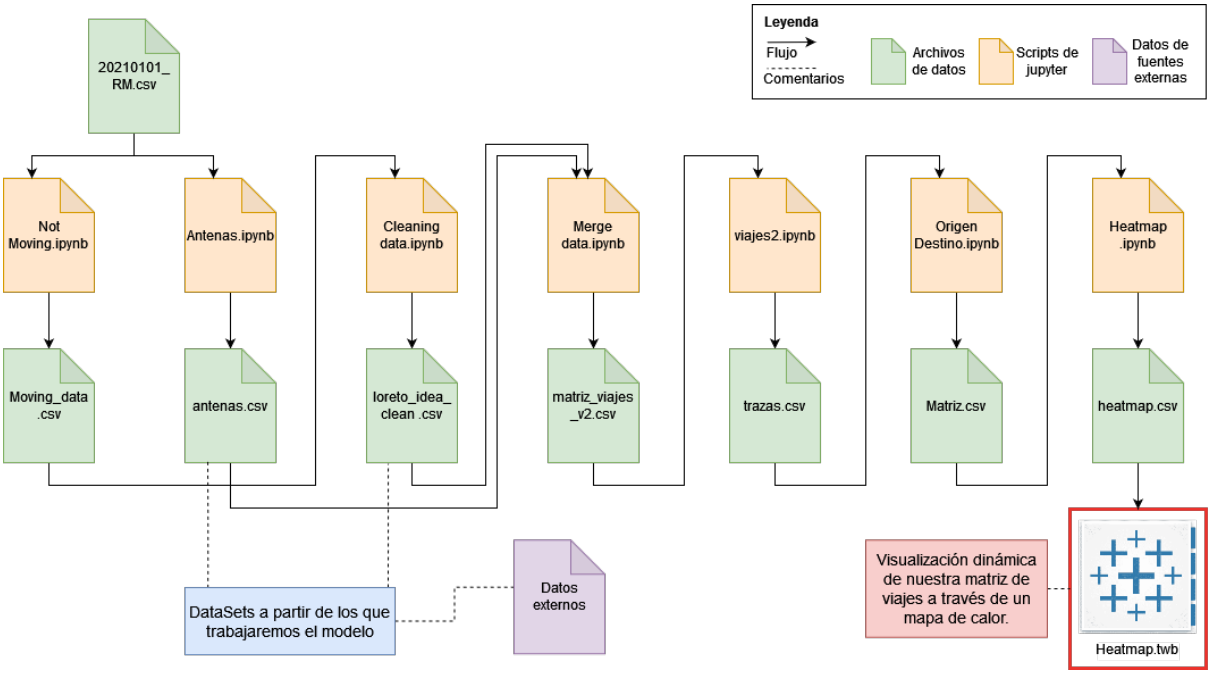
III. Método

En esta sección se presenta la ”Descripción del proceso”, en el que describimos los pasos llevados a cabo para obtener los datasets que tiene la información necesaria para generar nuestra matriz de viajes y para entrenar el modelo que cumpla con lo requerido en los objetivos de minería de datos planteados. Para más detalles, revisar el Anexo 4.

Mientras que, en el apartado de “Modelos generados” desarrollamos lo que queremos predecir y cómo obtener el set de datos para ello.

III-A. Descripción del proceso

En el siguiente diagrama se ve el proceso llevado a cabo:



Comenzamos realizando la limpieza y preparación de los datos proporcionados al inicio del bloque para llegar a un dataset, el cual nos de él número de viajes entre comunas y otro dataset para obtener el porcentaje de home-office por comuna.

Para llegar a esto se realizaron las siguientes actividades:

Primero, un borrado de los registros de las personas que no se movieron durante el día, ya que, al tener el registro de las conexiones de un dispositivo a una antena, este podría ser tanto un teléfono, que nos ayuda a conocer el movimiento de su usuario, como una tableta de alguna empresa o algún dispositivo que cuente con chip y no genera viajes.

Por otro lado, obtuvimos un nuevo dataset, donde cambiamos el ID de las antenas de su nombre dado a las coordenadas de la antena, esto debido a que existían antenas con el mismo nombre en locaciones diferentes y antenas con nombres distintos en las mismas coordenadas.

Finalmente, se agregó como columna la comuna a la que pertenecía cada antena. Además, se borraron los registros donde se hacían saltos de antenas con velocidades inusuales, es decir, de más de 150 km/h. Para considerar un salto como inusual seguimos lo dicho en el documento [2].

Unimos ambos datasets, tanto el de las antenas, como el de las conexiones sin saltos inusuales.

Después, cambiamos las conexiones a las antenas para tener la hora de entrada y salida a la que un usuario se conectaba a una antena, así como el tiempo de conexión en esa misma y la distancia entre una conexión y la siguiente en lugar de solo la hora de conexión a ésta, de manera que no se repitan los datos.

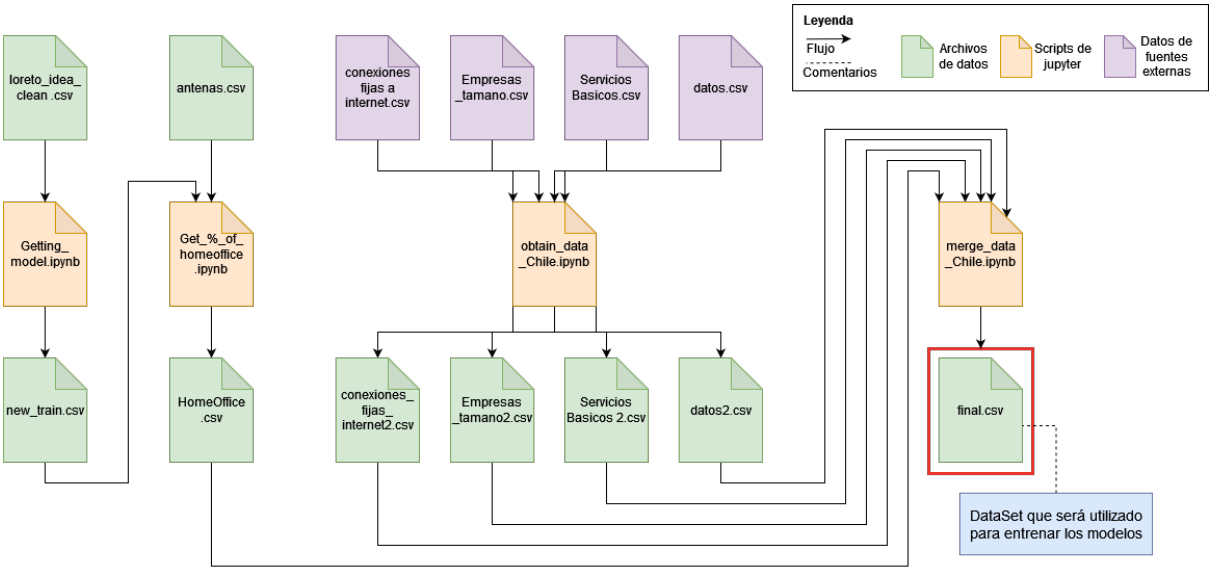
Posterior a ello, generamos el dataset de los viajes que realiza cada usuario, siguiendo las especificaciones del documento [2].

Finalmente, se cuenta el número de viajes realizados entre comunas y se genera un heatmap donde podemos observar los viajes entre comunas. Esto se puede ver en el diagrama dentro del repositorio Anexo 1.

Para más detalles del procesamiento de los datos, revisar el Anexo 2.

III-B. *Modelos generados*

En el siguiente diagrama se ve el proceso llevado a cabo:



Una vez obtenido el dataset para la matriz de viajes, ahora debemos crear los modelos, que en nuestro caso, predicen el porcentaje de home-office de cada comuna según algunas características de esta.

Para llegar al dataset que queremos se siguieron los siguientes pasos:

- A partir de nuestro dataset sin saltos inusuales, obtenemos la antena con mayor número de conexiones entre las 00:00 y las 7:00, la cual llamaremos antena de casa, y la antena con mayor número de conexiones entre las 9:00 y las 18:00, la cual llamaremos antena de trabajo, si una persona realizaba home-office o no en caso de que la antena de casa y la de trabajo sea la misma.
- Calculamos el porcentaje de home-office por comuna en base a los datos anteriores.
- Unimos este porcentaje de home-office por comuna con los datos del Sistema Integrado de Información Territorial (SIIT), los cuales consideramos relevantes para la generación de home-office de cada comuna.

Para más detalles, revisar el Anexo 2.

Finalmente, para la selección de cada modelo, comenzamos generando un benchmark, el cual nos dio una predicción simple para como punto de partida, en nuestro caso realizamos una regresión lineal [11], ya que es un modelo sencillo y útil. Continuamos con un modelo de deep learning, esto para evaluar su desempeño con nuestro dataset, pero al contar con tan pocos datos para entrenar no fue muy eficiente, aunque en caso de contar con más información para entrenar, generaría mejores resultados. Finalmente, se optó por un GBT[10], el cual es un modelo que principalmente utiliza boosting, esto quiere decir que emplea modelos de aprendizaje automático mucho más débiles como los árboles de decisión, donde cada árbol mejora a su antecesor, y debido a que este género los mejores resultados según nuestras métricas, fue el modelo que seleccionamos. Para más detalles, revisar el Anexo 3.

IV. Resultados

IV-A. *Evaluación*

Consideramos un modelo como válido, cuando este cumple con los objetivos de minería de datos y de negocio planteados en el documento de Business Understanding Anexo 4.

La comparación entre modelos se puede ver en la tabla de a continuación:

Modelos	Configuración	RMSE	MSE	MAE	r2
Linear Regression	maxIter = 100	5.2616	27.6845	3.0077	-0.4027
Neural Network (NN) 1.0	Learning rate: 0.01 Epochs: 10 Layers Dense = 3: Neuronas: 180 Activación: relu Neuronas: 512 Activación: relu Neuronas: 256 Activación: relu Neuronas: 1 Activación: linear Dropout = 2: Porcentaje: 20 % Porcentaje: 20 %	52380.125	4647275008.0	42886.35	-1941478980.30
Neural Network (NN) 2.0	Learning rate: 0.001 Epochs: 20 Layers Dense = 3: Neuronas: 160 Activación: relu Neuronas: 480 Activación: relu Neuronas: 256 Activación: relu Neuronas: 1 Activación: linear Dropout = 2: Porcentaje: 20 % Porcentaje: 20 %	30622.20	1582216320.0	25411.75	-660986423.04
Gradient-Boosted Trees (GBTs) 1.0	maxDepth = 5 maxIter = 20 maxBins = 32	3.0273	9.1646	2.2593	-5.7304
Gradient-Boosted Trees (GBTs) 2.0	maxDepth = 21 maxIter = 40 maxBins = 18	2.1 220	4.5032	1.7139	0.6335

Como podemos ver el modelo con mejores resultados es el GBTs v2, por lo que este fue el seleccionado para realizar el deployment.

IV-B. Insights

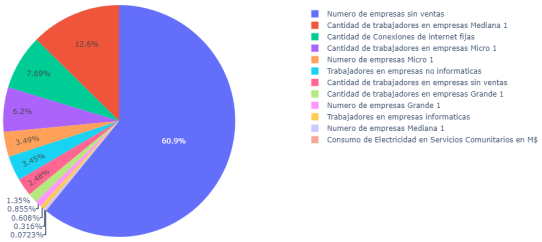
En parte de nuestro proceso, obtuvimos la información del promedio de home-office realizado en cada comuna, así como el número de usuarios que se detectó que realizaban home-office y las que no, conociendo así las comunas con mayor presencia de home-office, Anexo 5.

Según estos datos, podemos intuir que:

- La comuna con mayor presencia de home-office es San Pedro con 73.49 %
- La comuna con menor porcentaje de home-office es Alhué con 46.66 %
- Debido a que el porcentaje de home-office mayor a 46 % en todas las comunas, los datos deben ser en tiempos de pandemia.

Aparte, contamos con un modelo, el cual basándose en datos de cada comuna como la cantidad de empresas micro, chicas, medianas y grandes, número de empleados y otros datos, predice el porcentaje de home-office que se realiza en cada comuna.

En la siguiente imagen, se puede ver la influencia de cada variable con la predicción del modelo:



V. Discussions

V-A. Trabajo futuro

Para proyectos futuros, será bueno contar con una mayor cantidad de datos, ya sea con registros de más días o de otras zonas, esto para poder crear un dataset más extenso y así entrenar de una manera más efectiva el modelo y que con esto se realicen mejores predicciones.

Igualmente se pueden usar datos de un año en el que no haya pandemia, esto para conocer el impacto que tuvo en la vida diaria de las personas.

A futuro, puede usarse este modelo para predecir el porcentaje de home office en otras zonas ya sean de Chile o de cualquier otro país.

V-B. Lecciones aprendidas

Un error cometido durante este proyecto fue el subestimar los tiempos de procesamiento de los datos, al contar con un

dataset tan grande, esto causo retrasos en nuestro plan de trabajo y que tuviéramos que realizar varios cambios.

El desarrollo del proyecto continuó con normalidad en cada una de las fases del CRISP-DM hasta la fase de evaluation. Gracias a una sesión con el socio formador, nos dimos cuenta que al momento de establecer nuestros objetivos de negocio y de minería de datos, no planteamos la solución a un problema en el que fuera necesario una solución que usara machine learning o deep learning, lo que derivó a un modelo poco útil, que no aportaba valor y podía ser calculado por codificación tradicional. Por lo que se decidió regresar a la fase de Business Understanding y plantear correctamente nuestros objetivos a partir de lo que obtuvimos previamente, logrando realizar un deployment adecuado.

Los cambios que se realizaron en el plan durante la elaboración del proyecto fueron:

- Realizar una segunda iteración en CRISP-DM para plantear mejor tanto los objetivos de negocios como los de minería de datos.

VI. Anexos

1. Heatmap. (2022, November). Heatmap. Github. Heatmap
2. Data Preparation V2. (2022, November 23). Data Preparation V2. Github. Data Preparation V2
3. Modeling V2. (2022, November 27). Modeling V2. Google Docs. Modeling V2
4. Business Understanding V2. (2022, November 15). Business Understanding V2. Google Docs. Business Understanding V2
5. HomeOffice (2022, November). HomeOffice. Google Sheets. HomeOffice.csv

Referencias

[1] Graells-Garrido, E., & Saez-Trumper, D. (2016, February 29). A Day of Your Days: Estimating individual daily journeys using mobile data to understand urban flow. arXiv.org. Retrieved November 28, 2022, from https://arxiv.org/abs/1602.09000

[2] Loreto-Bravo (2022, September 19). Estimación de la Matriz Origen Destino a partir de Datos de PropuestaProyectoTECv2

[3] IBM. (2021, August 17). Conceptos básicos de ayuda de CRISP-DM. Conceptos Básicos de Ayuda de Crisp-DM. Retrieved November 29, 2022, from https://www.ibm.com/docs/es/spss-modeler/saas?topic=dm-crisp-help-overview

[4] IBM Cloud Education. (2020, August 20). ¿Qué es el análisis exploratorio de datos? IBM. Retrieved November 29, 2022, from https://www.ibm.com/mx-es/cloud/learn/exploratory-data-analysis

[5] Mahmood, O. (2021, March 2). What's ETL? Medium. Retrieved November 29, 2022, from https://towardsdatascience.com/whats-etl-b4903a57f8ce

[6] IBM. (n.d.). ¿Qué es Machine Learning? IBM. Retrieved November 29, 2022, from https://www.ibm.com/mx-es/analytics/machine-learning

[7] Apache. (n.d.). PySpark documentation. PySpark Documentation - PySpark 3.3.1 documentation. Retrieved November 29, 2022, from https://spark.apache.org/docs/latest/api/python/

[8] IBM. (n.d.). Deep learning - ¿qué es deep learning? IBM. Retrieved November 29, 2022, from https://www.ibm.com/mx-es/cloud/deep-learning

[9] MDPI. (2016). Sensing Urban Patterns with Antenna Mappings: The Case of Santiago, Chile. Retrieved November 29, 2022, from https://www.mdpi.com/1424-8220/16/7/1098

[10] LinearRegression — PySpark 3.3.1 documentation. (2022). Apache. https://spark.apache.org/docs/latest/api/python/reference/api/pyspark.ml.regression.LinearRegression.html

[11] Apache Spark (S.N). GBTRegressor. https://spark.apache.org/docs/latest/api/python/reference/api/pyspark.ml.regression.GBTRegressor.html