Predicción del precio de viviendas en la RM

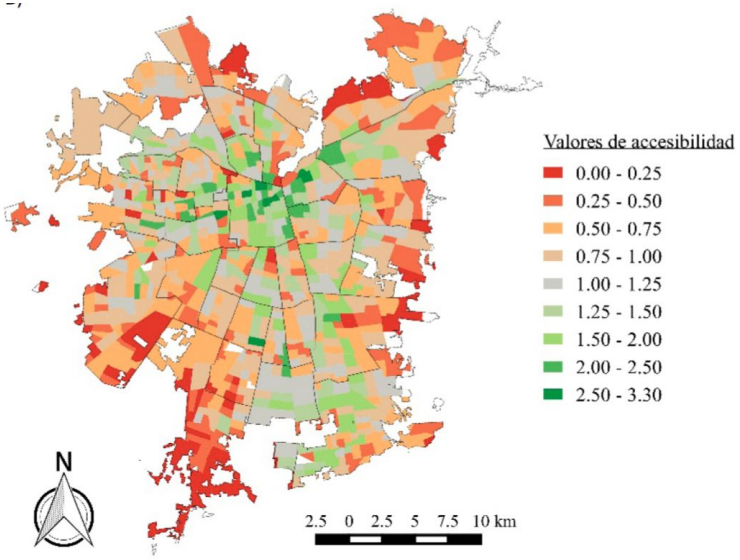
Cristobal Ardiles
Simon Lourido
Cristian Carrión
Roberto Maldonado
Asunción Gómez

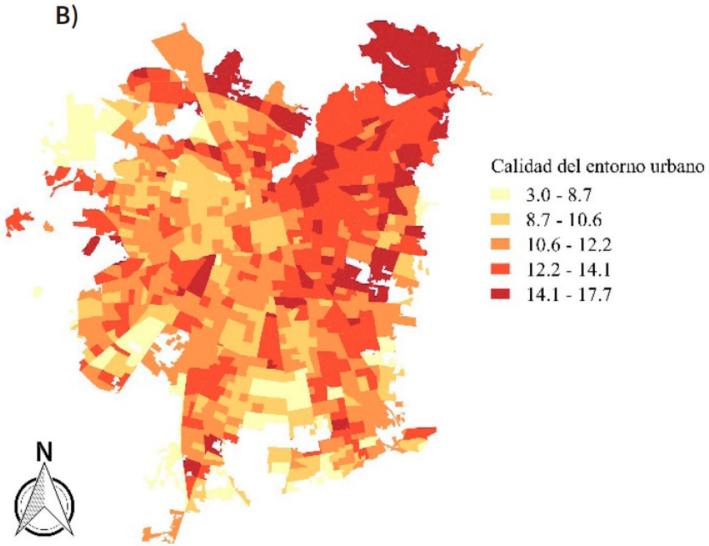
Primer análisis exploratorio

- 2 datasets: viviendas en la RM y delincuencia en el año 2020 y 2022.
- Variable delincuencia a partir de artículo de James, 2021.
- Poca correlación de variables numéricas con el precio.
- ¿Qué variables, o combinación de estas, tienen una correlación más importante con el precio de las viviendas?
- ¿Será la pandemia y la crisis económica un factor a influir en un modelo de predicción?
- ¿Qué otras variables pueden influir en el precio de la vivienda?

Nueva exploración de datos

- ¿Qué otras variables pueden influir en el precio de la vivienda?





A la izquierda las zonas de la ciudad que tienen buena accesibilidad a establecimientos de educación pública gratuita de calidad mediante transporte público. (Larraín y Tiznado-Aitken, 2021).

A la derecha, la calidad del entorno urbano. (Larraín y Tiznado-Aitken, 2021).

Nuevas datasets, nuevas variables

- Estaciones de metros en la RM
- Equipamiento cultural en la RM (teatro, museo, cine, centro cultural, etc)
- Establecimientos educacionales pre básicos
- Establecimientos educacionales
- Clínicas
- Establecimientos de salud (consultorio general, servicio de atención primaria de urgencias, etc).
- Red seguridad (carabineros, investigaciones y bomberos).
- Localización riesgos geofísicos
- Ferias libres y persas
- Supermercados
- Áreas verdes

	tipo	direccion	comuna
0	TEATRO Y SALA DE CONCIERTO	650 ESMERALDA	SANTIAGO
1	TEATRO Y SALA DE CONCIERTO	257 CUETO	SANTIAGO
2	TEATRO Y SALA DE CONCIERTO	1301 BALMACEDA	SANTIAGO
3	TEATRO Y SALA DE CONCIERTO	509 JOSE MIGUEL CLARO	PROVIDENCIA
4	TEATRO Y SALA DE CONCIERTO	1531 MARIO KREUTZBERGER	SANTIAGO

Equipamiento cultural

	tipo	direccion	coord_x	coord_y	localidad
0	ED PREBASICA	1900 ANTONIO MACHADO	347442.72845	6.285191e+06	SANTIAGO
1	ED BASICA	1949 VENANCIA LEIVA	347705.76039	6.285790e+06	SANTIAGO
2	ED BASICA	2769 EL OLIVAR	348779.19141	6.279349e+06	SANTIAGO
3	ED MEDIA	02669 EL OMBU	349217.67379	6.281664e+06	SANTIAGO
4	ED BASICA	02561 MIGUEL ANGEL	349130.79465	6.282411e+06	SANTIAGO

Establecimientos de educación públicos

Nuevas datasets, nuevas variables

De las nuevas variables propuestas, se lograron añadir 2 nuevas columnas a nuestro dataset de propiedades 2022: distancia a estación de metro más cercana y distancia a supermercado más cercano.

Esto se debe a que los otros datasets consideraban alrededor de 30 comunas de las 52 que contiene nuestro dataset y limitarlo a estas reduciría el número de filas (eliminaría ~1300 filas).

Propuesta experimental

Siguiendo la línea del objetivo del proyecto, y para responder las preguntas principales planteadas en el hito 1, se proponen 3 metodologías para obtener respuestas.

- **PCA**: para reducir la dimensionalidad y así ver las variables con mayor impacto en el precio.
- **Regresión Lineal**: un algoritmo usando como variable dependiente el precio y así lograr ver la relación con las otras variables, e incluso predecir el valor de la variable dependiente.
- **Random Forest:** un algoritmo clasificador que mediante árboles predictores, logre ir clasificando y así crear un modelo predictivo. Esta metodología se ajusta particularmente bien al dataset y al objetivo, dado que maneja bien varias variables y da estimaciones de qué variables son importantes en la clasificación.

Regresión lineal

- Alto nivel de significancia por parte de todas las variables agregadas al modelo.
- R2 ajustado de 0.48, por lo que se podría pensar que existen variables que no se incluyeron en el modelo o que no son observables y que podrían estar explicando parte de la variación del precio en UF. Por ejemplo, la situación económica a nivel mundial y a nivel país.

```
Call:
lm(formula = Valor_UF ~ N_Habitaciones + N_Baños + N_Estacionamientos +
    Total_Superficie_M2 + Tasa.cada.100Mil + dist_closest_metro +
    closest_super_dist, data = casa_usadas)

Residuals:
   Min     1Q Median     3Q    Max
-35799  -3122   -778   1874  35617

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        -4431.8045   493.6418  -8.978  < 2e-16 ***
N_Habitaciones       258.8220   101.7495   2.544   0.0110 *
N_Baños             3372.6308   115.7014  29.149  < 2e-16 ***
N_Estacionamientos  1396.8409   108.7363  12.846  < 2e-16 ***
Total_Superficie_M2    2.1052     0.1842  11.430  < 2e-16 ***
Tasa.cada.100Mil       1.1830     0.4675   2.531   0.0114 *
dist_closest_metro   345.1947    56.0772   6.156 8.56e-10 ***
closest_super_dist  -757.3992   136.2832  -5.558 3.00e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5692 on 2764 degrees of freedom
Multiple R-squared:  0.4875,    Adjusted R-squared:  0.4862
F-statistic: 375.6 on 7 and 2764 DF,  p-value: < 2.2e-16
```

Random Forest

Se aplica un regresor Random Forest para predecir el precio de las viviendas. Se utilizan sólo las variables numéricas del dataset.

Los resultados obtenidos muestran de forma preliminar una relativa coincidencia en los valores predichos. Sin embargo, es necesario validar con técnicas ad-hoc, lo que queda pendiente para el hito 3.

```
[ ] from sklearn.model_selection import train_test_split
```

```
[ ] #Definimos los predictores y la variable objetivo
    df_numeric = df_prop_ext_filter_2022.copy().drop(["id", "Comuna", "Valor_UF"], axis=1)
    df_numeric.head(3)

    target = list(df_prop_ext_filter_2022["Valor_UF"])

    X_train, X_test, y_train, y_test = train_test_split(df_numeric, target, test_size=.33, random_state=37)
```

```
[ ] # Fitting Random Forest Regression to the dataset
    # import the regressor
    from sklearn.ensemble import RandomForestRegressor

    # create regressor object
    regressor = RandomForestRegressor(n_estimators = 100, random_state = 0)

    # fit the regressor with x and y data
    regressor.fit(X_train, y_train)
    y_pred = regressor.predict(X_test)
```

```
    print(y_pred[0:10])
    print(y_test[0:10])
```

Decision Tree

Para complementar, se utiliza un Decision Tree Regressor para poder interpretar las decisiones tomadas por el algoritmo.

De dicho algoritmo, se pueden observar que las variables más importantes para predecir el precio son la superficie total, el número de baños y la cercanía al metro.

Sin embargo, dichas conclusiones deben ser respaldadas por las técnicas correctas, lo que queda pendiente para el hito 3.

Conclusiones y nuevas problemáticas

- Como primer punto, todas las metodologías propuestas funcionan a base de variables numéricas, y dado que el dataset cuenta con variables de tipo categóricas como ‘Comuna’, surge la problemática de cómo incorporar este tipo de variables al futuro modelo.
- Por otro lado, tenemos que, luego de aplicar Random Forest, obtener los resultados y utilizar Decision Tree para interpretar. Las nuevas variables agregadas al dataset tienen un fuerte impacto en el precio (en particular, la distancia al metro más cercano), por lo que concluímos que incorporar estas tienen un efecto positivo en el modelo.

Bibliografía

- Larraín y Tiznado-Aitken (2021). *Análisis de los criterios para definir áreas de integración urbana en Chile*. https://www.scielo.cl/scielo.php?pid=S0717-50512021000200142&script=sci_arttext
- Fuentes y Pezoa (2018). *Nuevas geografías urbanas en Santiago de Chile 1992 - 2012. Entre la explosión y la implosión de lo metropolitano.* https://scielo.conicyt.cl/scielo.php?pid=S0718-34022018000200131&script=sci_arttext&tlng=p
- James et al. (2021). *An Introduction to Statistical Learning*. https://hastie.su.domains/ISLR2/ISLRv2_website.pdf

Predicción del precio de viviendas en la RM

Cristobal Ardiles
Simon Lourido
Cristian Carrión
Roberto Maldonado
Asunción Gómez