ESCUELA DE POSGRADO

CARRERA: **ESPECIALIZACIÓN EN CIENCIA DE DATOS (ECD)**

TRABAJO FINAL

“Análisis del impacto del tipo y ubicación de los locales comerciales de un centro o corredor comercial abierto en la probabilidad de vacancia de los mismos, mediante herramientas de visualización, análisis de información geoespacial y algoritmos de aprendizaje supervisado”

Autor: Diego Sanguinetti (Legajo Nro 29206)

Tutor: Lic. Diego Ariel Aizemberg

Lugar y Fecha: CABA, 26 de mayo de 2020

Tabla de Contenidos

1. Introducción

El comercio tradicional en Argentina ha venido sufriendo la combinación de una caída de consumo generalizada y la competencia creciente de la venta on-line, lo que ha puesto en serias dificultades la subsistencia del negocio. Algunos informes puntuales publicados en los principales medios alertan acerca del incremento de la vacancia en los distintos centros comerciales o corredores abiertos de la ciudad de Buenos Aires y Gran Buenos Aires en el último año. Pero la falta de información pública sistemática y granular del sector no permiten entender este fenómeno en detalle. Por otro lado, la creciente disponibilidad pública y de libre acceso tanto de imágenes satelitales como de mapas espaciales detallados de calles, junto con algoritmos de análisis predictivo y sistemas de análisis y visualización de información espacial también al alcance de la mano, complementados con trabajos de campo puntuales, hacen posible ahora el desarrollo de herramientas simples, pero a la vez potentes, que faciliten la toma de decisión de inversores en el sector.

2. Antecedentes

Argentina ya lleva tres años de caída en su actividad económica, como así lo demuestra el "Estimador Mensual de Actividad Económica (EMAE)" elaborado por el INDEC.

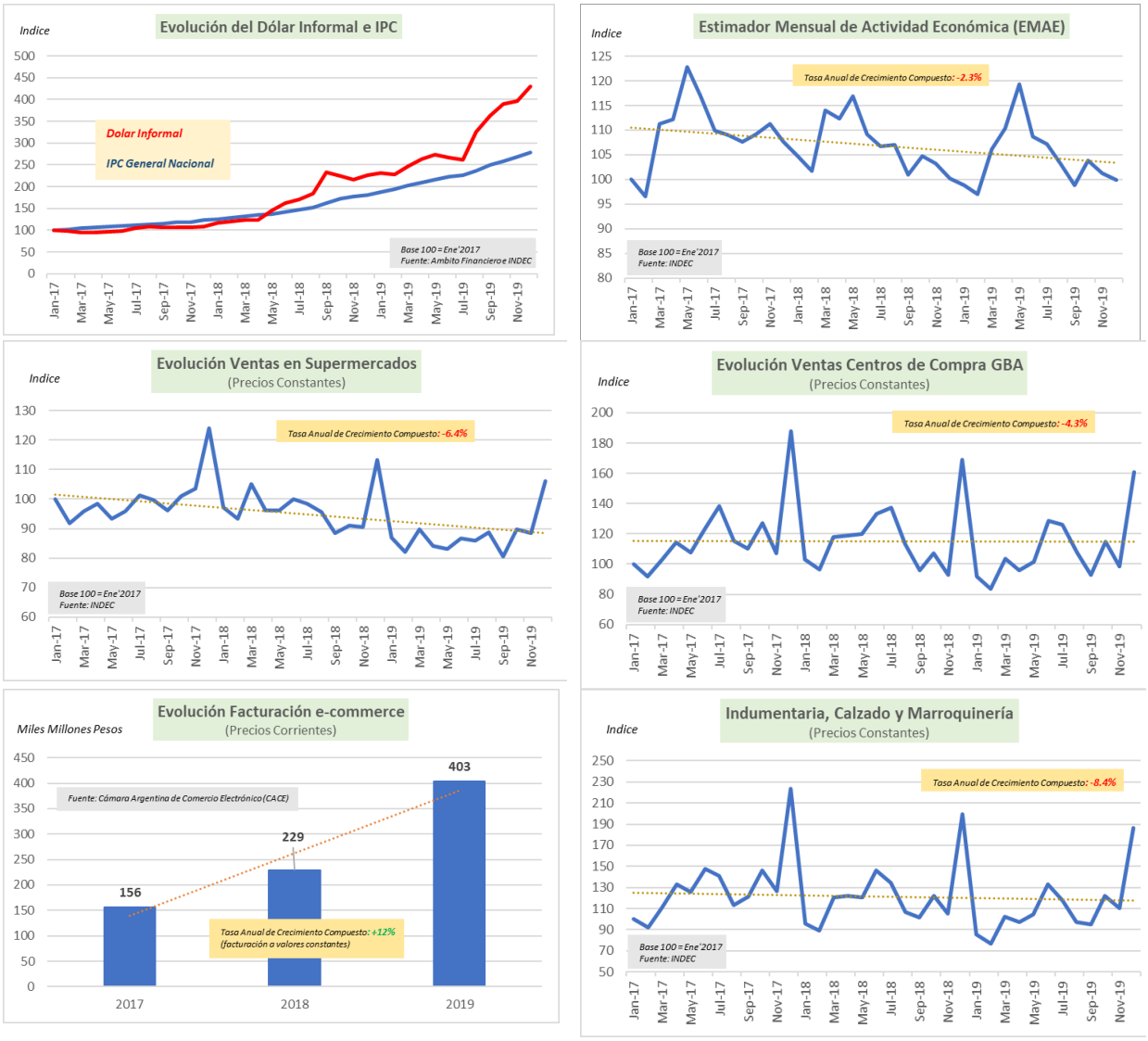
La devaluación y la inflación, que se aceleran a partir de mayo de 2018, implicaron una pérdida significativa en la capacidad de compra de las personas, y por tanto una caída fuerte en el consumo. La incertidumbre producto del año electoral pasado y el cambio de gobierno no hicieron más que agravar la cuestión.

Dentro de los actores del comercio, los centros de compra (shopping centers) han visto reducidas sus ventas a precio constante, aunque en menor medida que los supermercados, que fueron los principales afectados. Dentro de los centros de compra, uno de los rubros más afectados ha sido el de Indumentaria, Calzado y Marroquinería que representa un 40% de sus ventas, según informes del INDEC.

Por otro lado, la facturación del comercio electrónico sigue creciendo (+76% en 2019 versus 2018) y la penetración en los hogares ya llega a niveles considerables, con un 90% de los adultos conectados que ya compraron on-line alguna vez, según consigna la Cámara Argentina de Comercio Electrónico (CAME).

Una evolución de algunos indicadores relacionados puede ser visto en el Gráfico 1.

Gráfico 1 – Evolución principales Indicadores Económicos



Fuente: Ambito Financiero, INDEC, CAC (según se indica en cada gráfico)

Incertidumbre política, caída de la actividad económica, retracción del consumo y baja en las ventas en rubros importantes como Indumentaria, Calzado y Marroquinería (25%-35% del mix de rubros de un centro comercial) por un lado, y el incremento de la venta online por el otro constituyen una doble amenaza para los centros comerciales o corredores abiertos de la Ciudad de Buenos Aires y el Gran Buenos Aires, que observan como la apertura de nuevos locales se enlentece.

Dado lo anterior, la vacancia en los principales corredores comerciales se ha visto duplicada en el último año en comparación con niveles históricos (llega estos días hasta 10-20%) y la publicación de un aviso que, típicamente se mantenía por 20-60 días, ahora requiere de no menos de 3 a 4 meses antes del cierre de una operación de alquiler, según consignan informes de distintos medios periodísticos [a].

3. *Definición, Justificación y Alcance del Problema*

En el contexto descripto en el punto anterior, entender si alguna de las características de un local comercial en particular (como ser el tipo de local o su ubicación dentro de un centro comercial) influyen en el nivel de vacancia esperado para el mismo resulta clave para inversores en *real estate* comercial antes de embarcarse en la compra o venta de un activo de este tipo.

Permitiría, por ejemplo, establecer determinados segmentos o sub-mercados de locales dentro de los centros comerciales abiertos cuyo valor de compra/venta no solo debería reflejar la capacidad de generar una renta producto de su alquiler (forma de valoración habitual [b]) sino también el nivel de vacancia esperado y, por ende, de no generación de renta durante el período que dure la misma.

Asimismo, una medida objetiva del nivel de riesgo de vacancia de un local comercial en particular podría ser muy valioso a la hora de decidir si invertir o desinvertir en un local comercial para inversores que, dado su perfil financiero, requieren de una continuidad en la renta que reciben.

La falta de información pública y accesible respecto tanto de las características de los locales comerciales a nivel individual que conforman los centros comerciales abiertos de la ciudad, como de su evolución en el tiempo en términos de rubros de explotación o vacancia no favorecieron el desarrollo de herramientas que en forma objetiva permitieran determinar los factores que impactan sobre el tiempo y la cantidad de dicha vacancia (ni de otras variables clave para la toma de decisiones de negocio).

Por otro lado, la creciente disponibilidad pública y de libre acceso a través de internet de imágenes satelitales, mapas espaciales detallados de calles, sistemas de análisis y visualización de información espacial y algoritmos de análisis predictivo (de clasificación mediante aprendizaje supervisado), complementado con trabajos de campo puntuales deberían facilitar significativamente y hacer viable ahora el estudio de dichos factores para un centro comercial abierto.

Este mismo concepto, el de apalancar la ciencia de datos espaciales buscando mejorar las decisiones de negocios (p.e. para predecir las ventas de los locales o para planificar efectivamente las ubicaciones de nuevas aperturas), es presentado en distintos artículos publicados en los últimos dos años [c] como la única salida posible que tienen las grandes cadenas de "retail" mundiales (como p.e. Gap, Pier 1, Sears, etc) para evitar el colapso en sus negocios.

Marco Conceptual y Teórico

i. ¿Qué determina la **atractividad de una inversión en un local comercial en un corredor abierto**?

- Su precio de compra (más gastos de la compra)
- Su precio de alquiler mensual (renta esperada)
- Su costo de mantenimiento (pocas veces considerado dado su escasa relevancia)
- **Su nivel de riesgo de vacancia (período durante el cual no genera renta)**
- Otras consideraciones como la situación patrimonial, legal, o regulatoria de la propiedad (deuda, embargo, estado del título, zonificación/habilitación, etc)

ii. ¿Qué **factores podrían impactar en la vacancia** de un local comercial?

En general para todos los locales de un centro o corredor comercial:

- El contexto económico y de consumo en general (demanda)
- La superficie comercial ofrecida en la zona (oferta)
- La incidencia de nuevas formas de consumo, p.e. la venta on-line (oferta)

En particular, para un local comercial dentro de un centro comercial abierto:

- Su precio de alquiler y/o gastos mensuales fuera de valores de mercado [d]
- Su estado de conservación (no muy relevante) [e]
- **El tipo y ubicación del local dentro del centro comercial**

iii. ¿Cómo se podría **relevar la vacancia de un local comercial**?

- Se entiende por vacancia al período de tiempo en el cual un local se encuentra cerrado al público sin explotación comercial (en venta/alquiler, en remodelación o cerrado) y, por ende, es razonable asumir no genera una renta para su dueño

- Si bien existen datos de nivel general de vacancia de centros o corredores comerciales en textos periodístico (5-20%) que citan informes de Inmobiliarias o Cámaras Inmobiliarias, no hay una fuente pública, oficial y exhaustiva de vacancia a nivel de locales individuales en Argentina

- Dada la creciente disponibilidad pública y de libre acceso de información y herramientas para su análisis, el relevamiento de la vacancia de los locales individuales que conforman un centro comercial objetivo a lo largo de un período significativo (que incluya meses de alta y baja actividad comercial) y mediante trabajo de campo sería factible de realizar

iv. ¿Qué **variables permitirían caracterizar el tipo y ubicación de un local comercial** y cuya influencia en el nivel de vacancia se podría demostrar estadísticamente?

- Las siguientes variables permitirían caracterizar el **tipo de local de un centro comercial abierto**:

  ✓ ***Disposición***: discrimina locales comerciales cuyo salón tiene acceso directo a la calle de los ubicados en plantas elevadas y que requieren de una escalera o ascensor para acceder a los mismos
  ✓ ***Galería***: discrimina locales que forman parte de una galería comercial con acceso común desde la calle de locales individuales con salida propia a la misma
  ✓ ***Área***: área de la planta del local a nivel de calle medida en metros cuadrados (mts^2), sin considerar espacios en subsuelos, entrepisos u otros pisos
  ✓ ***Perímetro***: perímetro de la planta del local a nivel de la calle medida en metros lineales (mts)

- Las siguientes, caracterizar la **ubicación dentro del centro comercial**:

  ✓ ***Zona***: división natural del centro comercial a partir de avenidas importantes, vías de tren, etc
  ✓ ***Calle***: calle particular donde está ubicado el local
  ✓ ***Numeración***: altura de la calle donde se encuentra la puerta del local
  ✓ ***Par***: si la altura de la calle es par o impar, lo que determina la orientación del local
  ✓ ***Tipo-calle***: calle principal, lateral o paralela del centro comercial
  ✓ ***Esquina***: identifica locales ubicados en final de cuadra o esquinas (exposición a dos calles)

- ✓ ***Distancia-esquina***: distancia medida en metros del local a la esquina más próxima
- ✓ ***Distancia-principal***: distancia ortogonal en metros del local a la calle principal del centro
- ✓ ***Distancia-transporte***: distancia en línea recta en metros entre los locales comerciales y los puntos neurálgicos de transporte cercanos al centro comercial

- A partir de la combinación de imágenes satelitales, mapas detallados de calles y trabajo de campo se podría crear un mapa del centro comercial objetivo identificando todos los locales comerciales que lo componen mediante el uso de una herramienta de información geoespacial como QGIS, que permita/facilite el análisis espacial, la determinación de los valores de dichas variables para cada uno de los locales y su visualización posterior.

v. ¿Cómo se puede determinar el **impacto en el nivel de vacancia del tipo y ubicación** de un local comercial dentro de un centro comercial abierto?

Se puede estimar la probabilidad de vacancia de un local en particular y la relevancia estadística de las distintas variables en la predicción de la misma a partir de la utilización de algoritmos de análisis predictivo y modelos de aprendizaje supervisado basado en la información relevada en el punto anterior. Algunos de estos algoritmos más conocidos son los **árboles de decisión o clasificación, el discriminante "support vector machine" (SVM) o el clasificador de Naive Bayes (NBC),** entre otros.

4. Hipótesis de Investigación

Se puede determinar en forma estadísticamente significativa la influencia relativa de factores que caracterizan a un local comercial en particular dentro de un centro o corredor comercial abierto, tales como su tipo y su ubicación dentro del mismo, en el nivel vacancia esperado para dicho local comercial mediante la combinación de la información pública disponible, observaciones puntuales de campo, el uso de herramientas de análisis geoespacial y el desarrollo de modelos basados en algoritmos de clasificación mediante el aprendizaje supervisado.

Los modelos así desarrollados permitirían segmentar los locales comerciales de un centro comercial abierto en distintos grupos según el nivel de vacancia esperado y la visualización efectiva de dichos segmentos facilitaría la toma de decisión a los inversores en este tipo de activos.

5. Objetivo General y Específicos

Desarrollar una herramienta que permita visualizar distintos segmentos o grupos de locales entre los que conforman un centro o corredor comercial abierto de la ciudad según el nivel de riesgo de vacancia de los mismos, a partir del entendimiento de distintas variables que caractericen a los locales y puedan afectar su vacancia, mejorando/facilitando así las decisiones de compra y venta de locales comerciales por parte de los inversores.

Entre los objetivos específicos del trabajo se encuentran:

- Definir un centro comercial abierto apropiado y que sirva como objeto particular de estudio.

- Definir las variables que alimentarán el algoritmo de clasificación que se utilizará para la segmentar los locales según su nivel de riesgo de vacancia. Esto es la variable target (vacancia) y las variables o atributos independientes que caracterizan a cada local comercial respecto de su tipo y ubicación.
- Definir la metodología para generar los valores de dichas variables respecto de los locales de dicho centro comercial, combinando información pública, trabajo de campo u otras herramientas que se encuentren disponibles.
- Comparar distintos algoritmos de clasificación y aprendizaje supervisado que permitan caracterizar la influencia de las variables anteriores en el nivel de vacancia y seleccionar el de mejor performance.
- Segmentar los locales en base a la predicción de su probabilidad de vacancia y visualizar el resultado.

6. Metodología

Se tomó como objeto de estudio al centro o corredor comercial abierto denominado "Alvear, Martínez" que comprende unas 20 manzanas en las siguientes avenidas/calles: Irigoyen, Alvear, Santa Fe, Tres Sargentos, Albarellos, Arenales, Rawson, Eduardo Costa, General Paunero, V.F. López, L.Martínez, y Gral. Lamarca, ubicadas en la localidad de Martínez, partido de San Isidro, Provincia de Buenos Aires.

Dada la falta de información pública disponible se creó un mapa de los aprox. 600 locales comerciales que componen el centro comercial (ver Gráfico 2) a partir de:

- Imágenes de Google Earth (vistas satelitales 3D y panorámicas a nivel de calle Street View)
- Mapa web detallado de las calles bajo licencia abierta de OpenStreetMap (de tipo "raster tile" y que utiliza el clásico estilo ESRI)
- Visitas puntuales a terreno para recolección de información y depuración de detalles
- Herramienta QGIS, sistema de información geográfico (SIG) libre y de código abierto, que permite crear un formato vectorial en un sistema de referencia de coordenadas (código EPSG:3785 o de Proyección Google Pseudo-Mercator WGS84), su edición, su visualización y el análisis de la información geoespacial relativa a los locales

Gráfico 2 – Mapa del Centro comercial abierto Alvear, Martínez

Se les dio a los locales un número de orden que facilitó la eficiente realización de dos mediciones de vacancia, en noviembre de 2019 y en febrero de 2020. Las mismas fueron realizadas en un día hábil a media mañana y requirieron unas 2 horas de trabajo de una persona cada una. Se definió que un local sería considerado vacante para el estudio si fuera encontrado cerrado (con signos evidentes de no atención al público) en uno (cualquiera de ellas) o en ambos relevamientos realizados. Se identificaron un total de 63 locales o 11.6% de los 594 analizados como vacantes (valor de la clase = 1). (ver Gráfico 3)

Gráfico 3 – Locales vacantes o cerrados al público según los relevamientos realizados

Nota: locales cerrados o vacantes en rojo (clase=1) y locales abiertos en verde (clase=0)

Se definió como punto de partida utilizar las variables independientes que caracterizan el tipo de local de un centro comercial abierto y la ubicación dentro mismo, descriptas en el punto iv. del apartado "Marco Conceptual y Teórico", con las siguientes consideraciones particulares adicionales:

- ***Área y Perímetro***: se basaron en el mapa geoespacial creado con los locales y si bien los polígonos no tienen un grado elevado de exactitud a nivel absoluto e individual respecto de la realidad (fueron dibujados a partir de las imágenes satelitales de Google Earth), sirven para discriminar en forma relativa a los distintos locales dentro del centro comercial

- ***Numeración***: no pudo obtenerse en muchos casos, quedando como valor nulo

- ***Distancia-principal***: se tomó como calle principal al eje de las calles Irigoyen y Alvear

- ***Distancia-transporte***: fueron considerados dos casos: 1) Alvear y Santa Fe, subida y bajada de colectivos, y 2) Estación Martínez del FGBM, subida y bajada del tren urbano

Los valores de las variables independientes surgieron tanto de lo observado en los trabajos de campo como del análisis de la información espacial, mediante la herramienta QGIS.

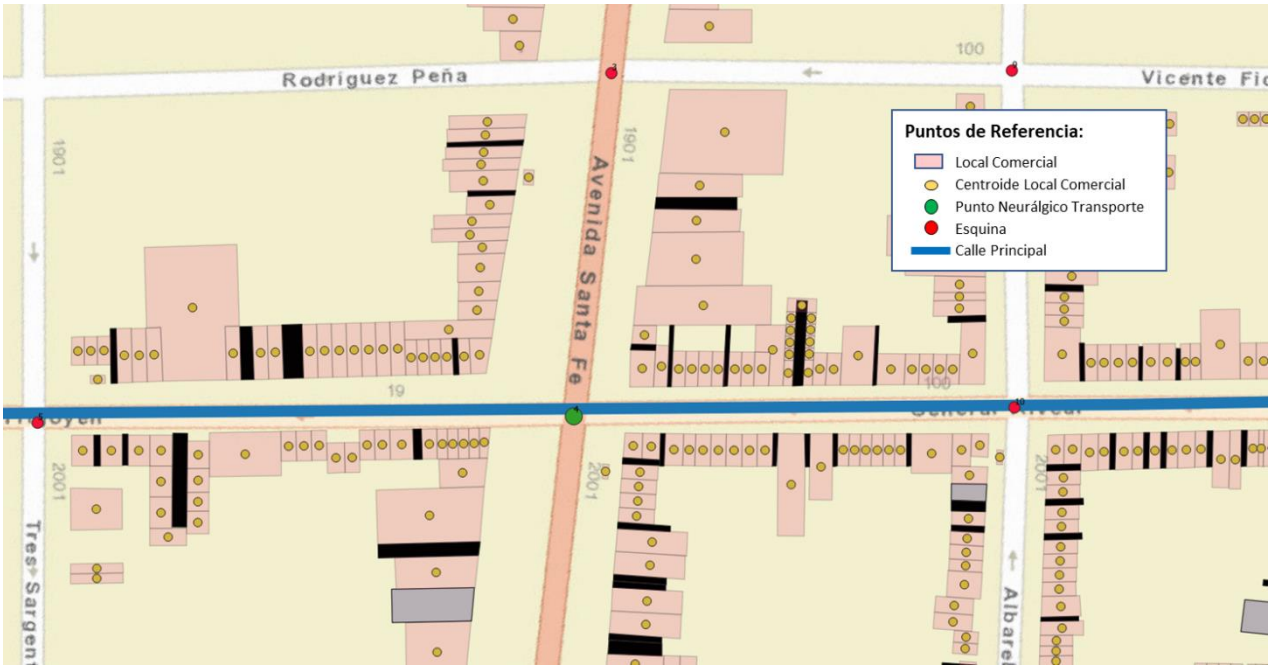
Por ejemplo, si el local forma parte de una galería, la calle o zona donde está emplazado, o si está ubicado en una esquina, esto se determinó a partir de los relevamientos realizados. Luego se visualizó dicha información en un mapa para asegurar que fuera correcta. (ver Gráficos 4)

Gráfico 4 – Clasificación de los Locales según zona y tipo de calle donde están emplazados

Notas: Zonas → 1- Irigoyen, 2- Santa Fe, 3- Alvear, 4- Alvear Libertador
Tipo de calles → 1- principal, 2- lateral, 3- paralela

Por otro lado, para determinar los valores de las variables restantes fue utilizado el "toolbox" del sistema QGIS, que permite el procesamiento de la información geoespacial. Como primera medida, fueron identificados los principales puntos espaciales de referencia, que incluyen los centroides de cada local, las esquinas de cada manzana y los dos puntos neurálgicos del transporte, y la línea recta sobre la calle principal. (ver Gráfico 5)

Gráfico 5 – Principales puntos de referencia para mediciones geoespaciales



Luego, a partir de estos puntos espaciales de referencia fueron computadas cada una de las variables. Por ejemplo, para determinar la distancia en metros del local a la calle principal, fueron primero identificados los centroides del polígono que representa cada uno ellos (transformación de "polygon" a "point layer"), luego se dibujó una recta sobre la calle principal convertida más tarde en una serie de puntos a un metro de distancia cada uno (transformación de "line" a "point layer"), y finalmente se calculó la distancia de cada uno de los centroides al punto de la calle principal más cercano ("distance to nearest hub").

7. Datos

Los valores de las variables independientes descriptas en el punto anterior fueron agrupados en un conjunto de datos que incluyó 594 observaciones y 14 variables independientes, totalizando 8169 registros (sin contar valores nulos). Además, se incorporó una clave única identificatoria de cada local u observación ("orden") y la variable dependiente o target ("clase") de tipo binaria con 1 para locales vacantes. (ver Tabla 1 en el Anexo)

El conjunto de variables independientes estuvo compuesto en su origen por:

- 7 **variables numéricas enteras** → "area", "perimetro", "numeracion", "distancia_esquina", "distancia_principal", "distancia_transporte_1", "distancia_trasnporte_2"
- 3 **variables binarias** → "galería", "par", "esquina"
- 4 **variables categóricas nominales** → "disposicion", "zona", "calle", "tipo_calle"

Las variables categóricas fueron transformadas a valores numéricos, mientras que dos de las 14 variables originales fueron finalmente descartadas a la hora de entrenar los modelos:

- "disposicion", no fue necesaria dado que el relevamiento sólo incluyó locales a la calle
- "numeracion", por contar con gran cantidad de valores nulos

Este conjunto de datos fue particionado al azar en un set de entrenamiento (80% o 475 registros) y un set de pruebas (20% o 119 registros) complementario, guardando en ambos la proporcionalidad de la clase (método retención o *holdout method* de evaluación de resultados de los modelos).

8. *Modelo y Resultados obtenidos*

El mismo set de entrenamiento se utilizó para entrenar tres modelos basados en algoritmos de clasificación de aprendizaje supervisado diferentes con parámetros que fueron optimizados, y se utilizó un cuarto modelo que siempre predice la clase mayoritaria como referencia:

1. **Decision Tree**, basado en librería "lightgbm" de Python (script en Anexo), con parámetros [f]:
   - ✓ 'is_unbalance': 'true', → en un dataset binario, busca balancear el peso de ambas clases
   - ✓ 'boosting': 'gbdt',
   - ✓ 'num_leaves': 31,
   - ✓ 'feature_fraction': 0.5,
   - ✓ 'bagging_fraction': 0.5,
   - ✓ 'bagging_freq': 20,
   - ✓ 'Boost_from_average': 'false',
   - ✓ 'learning_rate': 0.01,
   - ✓ 'verbose': 20
2. **Support Vector Machine (SVM)**, basado en la librería "e1071" de R, con parámetros:
   - ✓ Kernel "sigmoid"
   - ✓ Costo = 0.1
3. **NaïveBayes Classifier (NBC)**, basado en la librería "e1071" de R
4. **Modelo que predice siempre la clase mayoritaria (=0),** utilizado como referencia

Dado que se trató de un conjunto de datos desbalanceado (con la probabilidad de ocurrencia positiva de la variable target de menos del 11%) se determinó el mejor modelo comparando las matrices de confusión generadas a partir de la predicción sobre el mismo set de prueba y utilizando la métrica "balanced accuracy" [g], que incorpora los conceptos de sensibilidad y especificidad. Una baja sensibilidad implica una menor capacidad para detectar casos positivos, mientras que una baja especificidad una mayor generación de falsos positivos por parte del modelo.

El modelo que resultó con mejor performance (ver Tabla 2) fue el árbol de clasificación Lightgbm, con un "balanced accuracy" de 58% versus los 54%, 50% y 50% de los modelos NBC y SVM y del modelo tomado como referencia respectivamente. De la comparación de las matrices de confusión de cada uno de ellos se puede observar que el modelo LightGBM no genera falsos positivos (Real 0/Pred 1 = 0) y cuenta con una mayor o similar identificación (aunque baja) de los casos positivos (Real 1/Pred 1 = 2).

En general los árboles de decisión trabajan mejor con datasets desbalanceados y en particular el LightGBM cuenta con un parámetro que busca balancear los pesos de ambas clases, el cual fue utilizado.

A su vez, el valor de la medida "accuracy" simple alcanzado por el algoritmo LightGBM fue superior a la proporción de la clase mayoritaria tomada como referencia (91% versus 89%) y el resultado de su métrica AUC superior a 0.75, ambos sinónimos de un correcto modelo de predicción.

Tabla 2: Evaluación de la Performance de los 3 Modelos y versus la referencia [h]

A- Modelo LightGBM

Matriz de Confusión

	Real 1	Real 0
Pred 1	2	0
Pred 0	11	106

Accuracy	**90.8%**
Sensitivity or Recall	15%
Specificity	100%
Balanced Accuracy	**58%**

B- Modelo SVM

Matriz de Confusión

	Real 1	Real 0
Pred 1	0	0
Pred 0	12	106

Accuracy	**89.8%**
Sensitivity or Recall	0%
Specificity	100%
Balanced Accuracy	**50%**

C- Modelo NBC

Matriz de Confusión

	Real 1	Real 0
Pred 1	2	10
Pred 0	10	96

Accuracy	**83.1%**
Sensitivity or Recall	17%
Specificity	91%
Balanced Accuracy	**54%**

D- Modelo que Predice siempre la clase Mayoritaria (=0)

Matriz de Confusión

	Real 1	Real 0
Pred 1	0	0
Pred 0	13	105

Accuracy	**89.0%**
Sensitivity or Recall	0%
Specificity	100%
Balanced Accuracy	**50%**

Referencia clase mayoritaria	*89.1%*

Dado que el dataset analizado era relativamente pequeño, como siguiente paso se utilizó la técnica de validación cruzada de 10 iteraciones o *10-fold cross validation* de forma tal de poder entrenar el Modelo LightGBM, validarlo y realizar predicciones utilizando todo el conjunto de datos. Esta técnica, que permite reducir el sobreajuste u *overfitting* y mejorar la aproximación acerca de la precisión del modelo, consiste en dividir el conjunto de datos al azar en 10 subconjuntos, utilizando uno de ellos como set de pruebas y los otros 9 como set de entrenamiento, repitiendo el procedimiento en 10 iteraciones cuyos resultados son promediados. El "balanced accuracy" del Modelo LightGBM se incrementó en este caso a 61%. (ver Tabla 3)

Tabla 3: Modelo LightGBM con validación cruzada de 10 iteraciones

A'- Modelo LightGBM

(utilizando validación cruzada 10 iteraciones)

	Real 1	Real 0
Pred 1	14	6
Pred 0	49	525

Accuracy	**91%**
Sensitivity or Recall	22%
Specificity	99%
Balanced Accuracy	**61%**

Finalmente, el modelo LightGBM utiliza por defecto en una clasificación binaria como la presente un umbral del 50%. Esto es, asume como positiva toda observación cuya predicción de la probabilidad de ser positiva es superior a dicho 50%. Tomando un umbral de clasificación positiva menor de 28% (es decir, con una mayor agresividad en la detección) fue posible incrementar la capacidad del modelo de identificación de dichos casos o sensitividad a más del triple, sin afectar en gran medida la especificidad o generación de falsos positivos. Este equilibrio permitió alcanzar un "balanced accuracy" máximo para este modelo de casi 80%. (ver Tabla 4)

Tabla 4: Performance del Modelo LightGBM con validación cruzada y umbral optimizado

A''- Modelo LightGBM

(utilizando validación cruzada 10 iteraciones y umbral del 28%)

	Real 1	Real 0
Pred 1	48	99
Pred 0	15	432

Accuracy	**81%**
Sensitivity or Recall	76%
Specificity	81%
Balanced Accuracy	**79%**

A su vez, el modelo LightGBM con validación cruzada permitió asignar a las variables utilizadas distintos niveles de importancia en la predicción de la clase (es decir, en la determinación de la vacancia de los locales). Para ello se basó en la cantidad de veces, entre todas las corridas del modelo, que cada variable fue utilizada en la partición del árbol de decisión (cuanto más fue utilizada, mayor su relevancia). Por ejemplo, en nuestro caso, la variable "area" resultó de gran importancia y fue utilizada en 186 de las 200 iteraciones que realizó el modelo hasta llegar a su nivel óptimo. La variable "galería", por el otro lado, nunca fue utilizado implicando una menor importancia relativa en la predicción. (ver Tabla 4)

Tabla 4: Importancia de las variables en la predicción de la clase

Variables **más** relevantes		Variables **menos** relevantes	
area	186	calle	48
distancia_transporte1	161	par	41
distancia_esquina	156	tipo_calle	31
perimetro	130	zona	22
distancia_principal	125	esquina	4
distancia_transporte2	112	galeria	0

Finalmente, se pudo utilizar el mapa de locales creado en la herramienta QGIS para visualizar los distintos niveles de probabilidad de vacancia predecidos por el modelo LightGBM sobre cada uno de los locales del centro comercial. Para ello se incorporó la información de probabilidad al vector existente de locales mediante una operación de unión (Join attributes by field value) y se graduaron los colores de los polígonos que representan a los locales en el vector resultante según dicha probabilidad, con distintas posibilidades de fijación de niveles (Symbology → Graduated → Natural Breaks (Jenks) → 5 Classes). (ver Gráfico 6)

Gráfico 6: Visualización de los locales del centro comercial según su riesgo de vacancia

De estas dos herramientas (Tabla 4 y Gráfico 6) se desprende que locales de tamaño pequeño a mediano ubicados en la calle principal Alvear, cercanos a la Estación de Tren FGBM y no emplazados en las esquinas son los que cuentan con una menor probabilidad de quedar vacantes (por debajo del 20%). Por otro lado, los locales de mayor superficie y que se encuentran ubicados sobre la Avenida Santa Fe constituyen el otro extremo, con mayor riesgo de vacancia relativa cuanto más alejado a Alvear (entre 30 y 60%).

9. Conclusión

Se puede corroborar el impacto de los distintos factores que caracterizan los locales comerciales en el nivel de vacancia de los mismos a partir de los resultados de algoritmos de clasificación como LightGBM con validación cruzada, posibilitando la visualización de dichos resultados la toma de decisiones de inversión/desinversión más informadas.

Analizar los factores que determinan la vacancia en los centros comerciales abiertos es posible aún sin contar con relevamientos y estadísticas oficiales granulares y sistemáticas. Trabajos puntuales de campo, potenciados con imágenes satelitales y mapas de calle disponibles en internet, procesados utilizando herramientas de creación y análisis de información geoespacial de público acceso como QGIS lo hacen claramente viable.

Extender el presente análisis para abarcar un período de observaciones más extensivo o incluir otros centros comerciales abiertos similares como parte del estudio sin dudas agregarían robustez a los resultados obtenidos y debería considerarse como próximos pasos en la investigación.

10. Notas, Referencias y Bibliografía

(a) *"Por la recesión, aumentó un 57% la cantidad de locales comerciales vacíos en la Ciudad", enero 2019*
https://www.lanacion.com.ar/economia/

"Por la caída del consumo creció 30% la cantidad de locales comerciales vacíos", noviembre 2019
https://www.cronista.com/apertura-negocio/empresas/

"Aumentó un 13% la cantidad de locales vacíos en la Ciudad", marzo 2020
https://www.clarin.com/ciudades/

En base a informes de Colliers International Argentina, "Relevamiento sobre los locales comerciales de las principales arterias de la Ciudad Autónoma de Buenos Aires" (marzo 2020) y relevamientos realizados por la Cámara Argentina de Comercio y Servicio (CAC)

(b) *Típicamente se busca que un local comercial genere una renta neta anual neta de gastos en dólares de 4-6% sobre la inversión realizada para su adquisición, retorno que dadas las circunstancias de mercado actuales se encuentra más cercana al 2-3%*

(c) *Dongjie Fan, "Retail Revenue Prediction Models with Spatial Data Science", September 2019,* https://carto.com/blog/retail-revenue-prediction-data-science/

Jason Sadowski, "Retail trails: Using spatial data to elevate retail", October 2019,
https://towardsdatascience.com/retailtrails-4724d1f12a2f

Alana Podreciks, Nathan Uhlenbrock, and Kelly Ungerman, "Who's shopping where? The power of geospatial analytics in omnichannel retail", July 2018 https://www.mckinsey.com/industries/retail/our-insights/

Ting Choo Yee, "Geospatial Analytics in Retail Site Selection and Sales Prediction", March 2018
https://www.researchgate.net/publication/

(d) *Refiere a gastos periódicos a cargo del inquilino (servicios, impuestos, expensas, mantenimiento, etc).*

Dada la falta de información pública y la imposibilidad de relevar la variable precio de alquiler a nivel de local individual se asume que los mismos son individualmente fijados y siguiendo un comportamiento racional por parte de sus dueños y que no existen variaciones significativas en el nivel de gastos entre los locales o, si los precios de alquiler son fijados siguiendo comportamientos económicamente no racionales o existen variaciones de gastos significativa entre los mismos, estos son independientes de las variables que definen el tipo y la ubicación del local.

(e) *Típicamente un nuevo inquilino remodela por completo y a su cargo el local previo a su inicio de explotación del mismo siguiendo sus propias necesidades y estándares (imagen, arquitectura, seguridad, etc). El nivel de inversión para esta puesta a punto no influye de forma significativa en el precio de compra ni de alquiler del local, aunque si en el plazo de locación pactado en el contrato (el período mínimo legal de un contrato de alquiler es de 2 años, aunque dado lo anterior a menudo este se extiende a 5 o más años).*

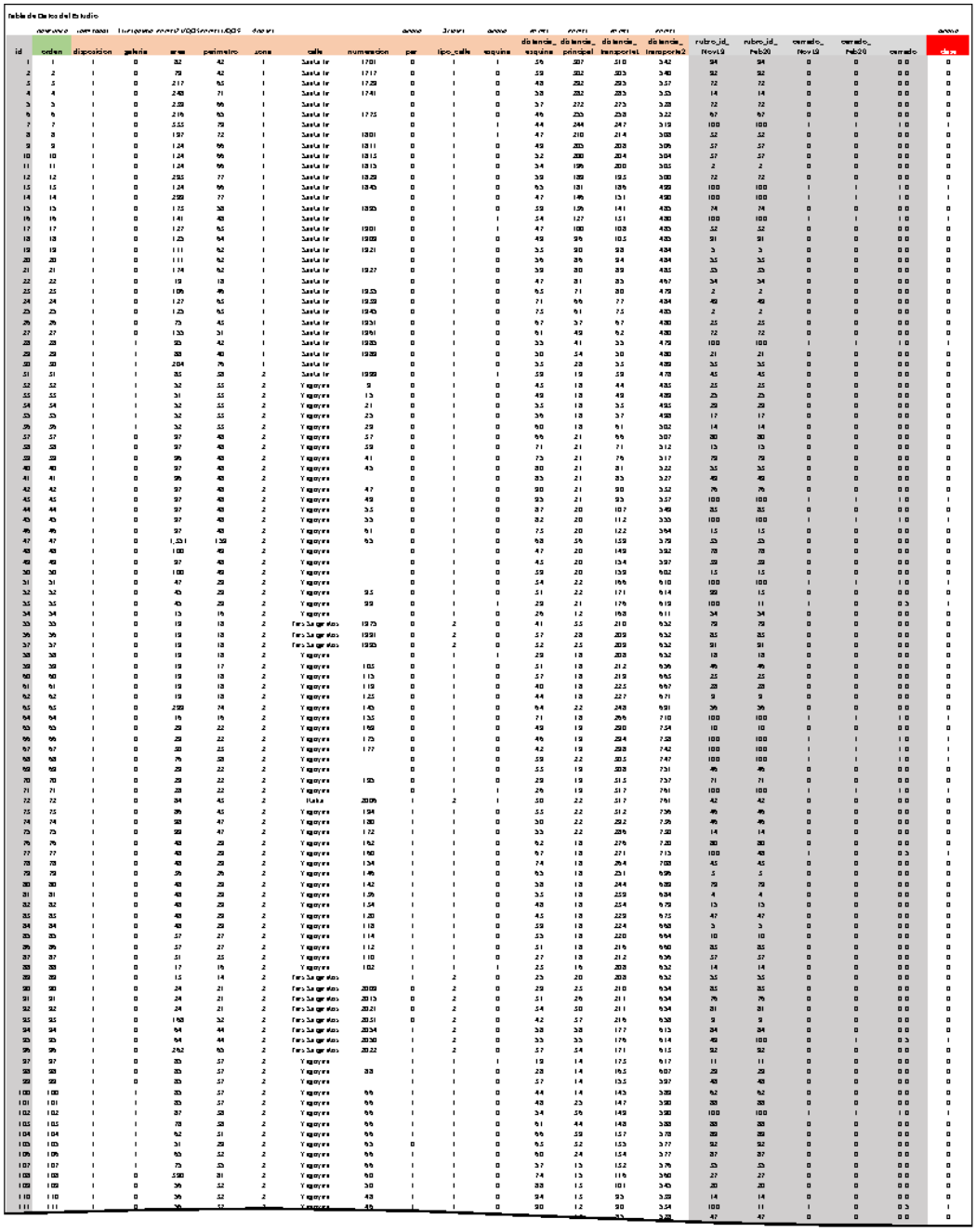
(f) *Una completa información de cómo optimizar los parámetros del algoritmo LightGBM se puede obtener en la sección "Parameters Tuning" del documento "Welcome to LightGBM's documentation" disponible en* https://testlightgbm.readthedocs.io/

(g) *Joset A. Etzel, "balanced accuracy: what and why?", December 2018*
http://mvpa.blogspot.com/2015/12/balanced-accuracy-what-and-why.html

(h) *A continuación, como fueron calculadas cada una de ellas:*
T: true F: false P: positive N: negative
Accuracy = (TP + TN) / (P + N)
Sensitivity = TP / (TP + FN) → habilidad para detectar a los casos positivos
Specificity or Recall = TN / (TN + FP) → habilidad para rechazar los casos negativos
Balanced Accuracy= (Sensitivity + Specificity) / 2

Tabla 3 – Datos del Relevamiento

https://github.com/padisan/infovis/blob/gh-pages/Tabla de Datos Locales TF CSV.csv



id	codem	disposicion	galeria	area	perimetro	zona	calle	numeracion	per	tipo_calle	esquina	distancia_esquina	distancia_principal	distancia_transporte1	distancia_transporte2	rubro_id_Nov19	rubro_id_Feb20	cerrado_Nov19	cerrado_Feb20	cerrado	clase
1	1	1	0	82	42	1	Juana Ir	1701	0	1	1	56	307	510	342	94	94	0	0	0 0	0
2	2	1	0	78	42	1	Juana Ir	1717	0	1	0	59	302	505	340	92	92	0	0	0 0	0
3	5	1	0	217	65	1	Juana Ir	1729	0	1	0	48	292	495	337	72	72	0	0	0 0	0
4	4	1	0	248	71	1	Juana Ir	1741	0	1	0	38	282	485	335	14	14	0	0	0 0	0
5	5	1	0	238	66	1	Juana Ir		0	1	0	37	272	475	328	72	72	0	0	0 0	0
6	6	1	0	216	65	1	Juana Ir	1775	0	1	0	46	255	458	322	67	67	0	0	0 0	0
7	7	1	0	535	79	1	Juana Ir		0	1	1	44	244	247	319	100	100	1	1	1 0	1
8	8	1	0	197	72	1	Juana Ir	1801	0	1	1	47	210	214	308	32	32	0	0	0 0	0
9	9	1	0	124	66	1	Juana Ir	1811	0	1	0	49	205	208	305	27	27	0	0	0 0	0
10	10	1	0	124	66	1	Juana Ir	1815	0	1	0	52	200	204	304	27	27	0	0	0 0	0
11	11	1	0	124	66	1	Juana Ir	1815	0	1	0	54	195	200	305	2	2	0	0	0 0	0
12	12	1	0	295	77	1	Juana Ir	1829	0	1	0	59	189	195	308	72	72	0	0	0 0	0
13	13	1	0	124	66	1	Juana Ir	1845	0	1	0	65	181	186	498	100	100	1	1	1 0	1
14	14	1	0	299	77	1	Juana Ir		0	1	0	67	146	151	498	100	100	1	1	1 0	1
15	15	1	0	175	58	1	Juana Ir	1895	0	1	0	59	136	141	485	74	74	0	0	0 0	0
16	16	1	0	141	48	1	Juana Ir		0	1	1	54	127	151	480	100	100	1	1	1 0	1
17	17	1	0	127	65	1	Juana Ir	1901	0	1	0	47	100	108	485	52	52	0	0	0 0	0
18	18	1	0	125	64	1	Juana Ir	1909	0	1	0	49	96	105	485	91	91	0	0	0 0	0
19	19	1	0	111	62	1	Juana Ir	1921	0	1	0	52	90	98	494	3	3	0	0	0 0	0
20	20	1	0	111	62	1	Juana Ir		0	1	0	56	86	94	494	55	55	0	0	0 0	0
21	21	1	0	124	62	1	Juana Ir	1927	0	1	0	59	80	89	485	55	55	0	0	0 0	0
22	22	1	0	19	18	1	Juana Ir		0	1	0	47	81	85	497	34	34	0	0	0 0	0
23	23	1	0	109	46	1	Juana Ir	1935	0	1	0	65	71	80	479	2	2	0	0	0 0	0
24	24	1	0	127	65	1	Juana Ir	1939	0	1	0	71	66	77	494	49	49	0	0	0 0	0
25	25	1	0	125	65	1	Juana Ir	1945	0	1	0	75	61	75	485	2	2	0	0	0 0	0
26	26	1	0	75	45	1	Juana Ir	1951	0	1	0	67	57	67	480	25	25	0	0	0 0	0
27	27	1	0	133	51	1	Juana Ir	1961	0	1	0	61	49	62	480	72	72	0	0	0 0	0
28	28	1	1	95	42	1	Juana Ir	1965	0	1	0	55	41	55	479	100	100	1	1	1 0	1
29	29	1	1	88	40	1	Juana Ir	1989	0	1	0	50	54	50	480	21	21	0	0	0 0	0
30	30	1	1	284	76	1	Juana Ir		0	1	0	55	28	55	489	55	55	0	0	0 0	0
31	31	1	1	85	38	2	Juana Ir	1999	0	1	1	59	19	59	478	45	45	0	0	0 0	0
32	32	1	1	32	55	2	Yaguarete	9	0	1	0	45	18	44	485	25	25	0	0	0 0	0
33	33	1	1	31	55	2	Yaguarete	13	0	1	0	49	18	49	488	25	25	0	0	0 0	0
34	34	1	1	32	55	2	Yaguarete	21	0	1	0	52	18	55	492	29	29	0	0	0 0	0
35	35	1	1	32	55	2	Yaguarete	25	0	1	0	56	18	57	498	17	17	0	0	0 0	0
36	36	1	1	32	55	2	Yaguarete	29	0	1	0	60	18	61	502	14	14	0	0	0 0	0
37	37	1	0	97	48	2	Yaguarete	37	0	1	0	66	21	66	507	80	80	0	0	0 0	0
38	38	1	0	97	48	2	Yaguarete	39	0	1	0	71	21	71	512	15	15	0	0	0 0	0
39	39	1	0	96	48	2	Yaguarete	41	0	1	0	75	21	76	517	79	79	0	0	0 0	0
40	40	1	0	97	48	2	Yaguarete	45	0	1	0	80	21	81	522	55	55	0	0	0 0	0
41	41	1	0	96	48	2	Yaguarete		0	1	0	85	21	85	527	48	48	0	0	0 0	0
42	42	1	0	97	48	2	Yaguarete	47	0	1	0	90	21	90	532	76	76	0	0	0 0	0
43	43	1	0	97	48	2	Yaguarete	49	0	1	0	95	21	95	537	100	100	1	1	1 0	1
44	44	1	0	97	48	2	Yaguarete	53	0	1	0	97	20	107	548	55	55	0	0	0 0	0
45	45	1	0	97	48	2	Yaguarete	55	0	1	0	82	20	112	555	100	100	1	1	1 0	1
46	46	1	0	97	48	2	Yaguarete	61	0	1	0	75	20	122	564	15	15	0	0	0 0	0
47	47	1	0	1,531	158	2	Yaguarete	63	0	1	0	68	36	158	579	55	55	0	0	0 0	0
48	48	1	0	100	48	2	Yaguarete		0	1	0	47	20	149	592	78	78	0	0	0 0	0
49	49	1	0	97	48	2	Yaguarete		0	1	0	45	20	154	597	28	28	0	0	0 0	0
50	50	1	0	100	48	2	Yaguarete		0	1	0	59	20	159	602	15	15	0	0	0 0	0
51	51	1	0	47	29	2	Yaguarete		0	1	0	54	22	166	610	100	100	1	1	1 0	1
52	52	1	0	45	29	2	Yaguarete	95	0	1	0	51	22	171	614	98	15	0	0	0 0	0
53	53	1	0	45	29	2	Yaguarete	99	0	1	1	29	21	176	619	100	11	1	0	0 3	1
54	54	1	0	15	15	2	Yaguarete		0	1	0	25	12	158	611	54	54	0	0	0 0	0
55	55	1	0	19	18	2	Tiro Suegretas	1975	0	2	0	41	55	210	632	79	79	0	0	0 0	0
56	56	1	0	19	18	2	Tiro Suegretas	1991	0	2	0	57	28	209	632	85	85	0	0	0 0	0
57	57	1	0	19	18	2	Tiro Suegretas	1995	0	2	0	52	25	209	632	91	91	0	0	0 0	0
58	58	1	0	19	18	2	Yaguarete		0	1	1	29	18	208	632	18	18	0	0	0 0	0
59	59	1	0	19	17	2	Yaguarete	105	0	1	0	51	18	212	636	49	49	0	0	0 0	0
60	60	1	0	19	18	2	Yaguarete	115	0	1	0	57	18	219	643	25	25	0	0	0 0	0
61	61	1	0	19	18	2	Yaguarete	119	0	1	0	60	18	225	647	28	28	0	0	0 0	0
62	62	1	0	19	18	2	Yaguarete	125	0	1	0	64	18	227	671	9	9	0	0	0 0	0
63	63	1	0	299	74	2	Yaguarete	145	0	1	0	64	22	248	691	39	39	0	0	0 0	0
64	64	1	0	19	19	2	Yaguarete	152	0	1	0	71	18	255	710	100	100	1	1	1 0	1
65	65	1	0	29	22	2	Yaguarete	159	0	1	0	49	19	290	754	10	10	0	0	0 0	0
66	66	1	0	29	22	2	Yaguarete	175	0	1	0	46	19	294	758	100	100	1	1	1 0	1
67	67	1	0	30	25	2	Yaguarete	177	0	1	0	42	19	298	742	100	100	1	1	1 0	1
68	68	1	0	76	38	2	Yaguarete		0	1	0	59	22	305	747	100	100	1	1	1 0	1
69	69	1	0	29	22	2	Yaguarete		0	1	0	55	19	309	731	49	49	0	0	0 0	0
70	70	1	0	29	22	2	Yaguarete	195	0	1	0	29	19	315	757	71	71	0	0	0 0	0
71	71	1	0	28	22	2	Yaguarete		0	1	1	26	19	317	761	100	100	1	1	1 0	1
72	72	1	0	84	45	2	Italia	2009	1	2	1	50	22	317	761	42	42	0	0	0 0	0
73	73	1	0	86	45	2	Yaguarete	194	1	1	0	55	22	312	736	49	49	0	0	0 0	0
74	74	1	0	88	47	2	Yaguarete	180	1	1	0	50	22	252	726	49	49	0	0	0 0	0
75	75	1	0	88	47	2	Yaguarete	172	1	1	0	55	22	286	720	14	14	0	0	0 0	0
76	76	1	0	48	29	2	Yaguarete	162	1	1	0	62	18	276	720	80	80	0	0	0 0	0
77	77	1	0	48	29	2	Yaguarete	160	1	1	0	67	18	271	715	100	48	1	0	0 3	1
78	78	1	0	48	29	2	Yaguarete	134	1	1	0	74	18	264	708	45	45	0	0	0 0	0
79	79	1	0	26	26	2	Yaguarete	146	1	1	0	65	18	251	695	5	5	0	0	0 0	0
80	80	1	0	48	29	2	Yaguarete	142	1	1	0	58	18	244	688	79	79	0	0	0 0	0
81	81	1	0	48	29	2	Yaguarete	136	1	1	0	55	18	259	684	4	4	0	0	0 0	0
82	82	1	0	48	29	2	Yaguarete	134	1	1	0	48	18	254	679	15	15	0	0	0 0	0
83	83	1	0	48	29	2	Yaguarete	120	1	1	0	45	18	229	675	47	47	0	0	0 0	0
84	84	1	0	48	29	2	Yaguarete	118	1	1	0	59	18	224	592	5	5	0	0	0 0	0
85	85	1	0	57	27	2	Yaguarete	114	1	1	0	55	18	220	664	10	10	0	0	0 0	0
86	86	1	0	57	27	2	Yaguarete	112	1	1	0	51	18	216	660	85	85	0	0	0 0	0
87	87	1	0	51	25	2	Yaguarete	110	1	1	0	27	18	212	636	57	57	0	0	0 0	0
88	88	1	0	17	16	2	Yaguarete	102	1	1	1	25	16	208	632	14	14	0	0	0 0	0
89	89	1	0	15	14	2	Tiro Suegretas		1	2	0	25	20	208	632	35	35	0	0	0 0	0
90	90	1	0	24	21	2	Tiro Suegretas	2009	0	2	0	29	25	210	634	85	85	0	0	0 0	0
91	91	1	0	24	21	2	Tiro Suegretas	2015	0	2	0	51	26	211	634	76	76	0	0	0 0	0
92	92	1	0	24	21	2	Tiro Suegretas	2021	0	2	0	54	30	211	634	81	81	0	0	0 0	0
93	93	1	0	168	52	2	Tiro Suegretas	2051	0	2	0	42	57	216	638	9	9	0	0	0 0	0
94	94	1	0	94	44	2	Tiro Suegretas	2034	1	2	0	30	30	177	615	84	84	0	0	0 0	0
95	95	1	0	94	44	2	Tiro Suegretas	2030	1	2	0	55	55	176	614	48	100	0	1	0 3	1
96	96	1	0	262	66	2	Tiro Suegretas	2022	1	2	0	57	54	171	615	92	92	0	0	0 0	0
97	97	1	0	85	37	2	Yaguarete		1	1	1	19	14	175	617	11	11	0	0	0 0	0
98	98	1	0	85	37	2	Yaguarete	88	1	1	0	28	14	165	607	28	28	0	0	0 0	0
99	99	1	0	85	37	2	Yaguarete		1	1	0	37	14	155	597	45	45	0	0	0 0	0
100	100	1	1	85	37	2	Yaguarete	96	1	1	0	44	14	145	588	62	62	0	0	0 0	0
101	101	1	1	85	37	2	Yaguarete	96	1	1	0	48	23	147	588	88	88	0	0	0 0	0
102	102	1	1	87	38	2	Yaguarete	96	1	1	0	34	36	149	588	100	100	1	1	1 0	1
103	103	1	1	78	38	2	Yaguarete	96	1	1	0	61	44	148	588	88	88	0	0	0 0	0
104	104	1	1	62	31	2	Yaguarete	96	1	1	0	66	50	127	578	88	88	0	0	0 0	0
105	105	1	1	31	29	2	Yaguarete	93	0	1	0	65	52	155	577	92	92	0	0	0 0	0
106	106	1	1	65	32	2	Yaguarete	96	1	1	0	60	24	154	577	87	87	0	0	0 0	0
107	107	1	1	75	35	2	Yaguarete	96	1	1	0	57	13	152	576	35	35	0	0	0 0	0
108	108	1	0	198	81	2	Yaguarete	90	1	1	0	74	13	116	540	27	27	0	0	0 0	0
109	109	1	0	96	52	2	Yaguarete	30	1	1	0	83	13	101	545	20	20	0	0	0 0	0
110	110	1	0	96	52	2	Yaguarete	48	1	1	0	94	15	93	528	14	14	0	0	0 0	0
111	111	1	0	96	52	2	Yaguarete	49	1	1	0	90	12	90	554	100	11	1	0	0 3	1
															528	47	47	0	0	0 0	0

Script Python Modelo lightGBM con 10-fold cross validation

```
import numpy as np
import pandas as pd
import lightgbm
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import LabelBinarizer

train = pd.read_csv("C:/Users/Adm/Locales.csv")
train
y = train.clase.values
orden = train.orden.values

train.drop(['id', 'orden','disposicion', 'numeracion', 'rubro_id_Nov19', 'rubro_id_Feb20',
'cerrado_Nov19', 'cerrado_Feb20','cerrado','clase'], inplace=True, axis=1)

from sklearn.preprocessing import LabelEncoder

le = LabelEncoder()

train["calle"] = le.fit_transform(train["calle"])

features = train.columns


x = train.values

# para dividir el dataset en train y test
# x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.2, random_state=42,
stratify=y)
# train_data = lightgbm.Dataset(x_train, label=y_train, free_raw_data=False)
# test_data = lightgbm.Dataset(x_test, label=y_test, free_raw_data=False)


train_data = lightgbm.Dataset(x, label=y)
train_data


parameters = {
    'application': 'binary',
    'objective': 'binary',
    'has_header': 'true',
    'metric': ['binary_error', 'auc'],
    'is_unbalance': 'true',
    'boosting': 'gbdt',
    'num_leaves': 31,
    'feature_fraction': 0.5,
    'bagging_fraction': 0.5,
    'bagging_freq': 20,
```

```
    'Boost_from_average': 'false',
    'learning_rate': 0.01,
    'verbose': 20
}

cv_result = lightgbm.cv(parameters,
                    train_data,
                    num_boost_round=5000,
                    nfold=10,
                    stratified=True,
                    shuffle=True,
                    early_stopping_rounds=100,
                    verbose_eval=20,
                    show_stdv=True)

num_boost_rounds_lgb = len(cv_result['binary_error-mean'])

model = lightgbm.train(parameters, train_data, num_boost_round=num_boost_rounds_lgb)

features

# importancia de las variables en la clasificación

sorted(zip(model.feature_importance(), model.feature_name()), reverse = True)


y_pred = model.predict(x)

y_pred


output = pd.DataFrame({'clase': y, 'clase_pred': y_pred})

output.to_csv("C:/Users/Adm/Locales_pred_cv.csv", index=False)
```