# Modelos de aprendizaje estadístico para la caracterización de los sectores comerciales en la Ciudad de México, usando datos de movilidad alternativos

Rodrigo José Morales Castellanos

UVG UNIVERSIDAD DEL VALLE DE GUATEMALA

UNIVERSIDAD DEL VALLE DE GUATEMALA
Facultad de Ciencias y Humanidades


# Modelos de aprendizaje estadístico para la caracterización de los sectores comerciales en la Ciudad de México, usando datos de movilidad alternativos

Trabajo de graduación en modalidad de tesis presentado por
Rodrigo José Morales Castellanos
para optar al grado académico de Licenciado en Matemática Aplicada

Guatemala,
2022

UNIVERSIDAD DEL VALLE DE GUATEMALA
Facultad de Ciencias y Humanidades


# Modelos de aprendizaje estadístico para la caracterización de los sectores comerciales en la Ciudad de México, usando datos de movilidad alternativos

Trabajo de graduación en modalidad de tesis presentado por
Rodrigo José Morales Castellanos
para optar al grado académico de Licenciado en Matemática Aplicada

Guatemala,
2022

Vo.Bo.:

(f) _____

Lic. Alan Reyes Figueroa

Tribunal Examinador:

(f) _____

Lic. Alan Reyes Figueroa

(f) _____

Lic. Paulo Mejía

(f) _____

Lic. Dorval Carías

Fecha de aprobación: Guatemala, 16 de Junio de 2022.

# Prefacio

La idea principal para realizar esta investigación surgió por proyectos realizados tanto en la universidad como en el trabajo. Ya que junto con un equipo de analistas se desarrolló una solución para medir y predecir el potencial de ventas de casi un millón de abarroterías en México. Este proyecto, iniciado cómo una tarea para la clase de *Machine Learning*, trata de hacer lo mismo pero con datos más baratos. Ya que los usados por el equipo de analistas requerían de una infraestructura computacional mayor y un presupuesto de miles de dólares.

Por lo que la idea era replicar de manera más general y barata los resultados para una pequeña parte de la Ciudad De México por la disponibilidad de los datos gratuitos. En fin es esta investigación es darle herramientas a los emprendedores para saber cómo ubicar su negocio sin tener que tener un presupuesto elevado y con resultados confiables.

Quiero tomar este espacio para agradecer a mi familia por todo el apoyo brindado en estos años de estudio. También quiero agradecer a mis profesores y mi asesor por guiarme.

# Índice

# Lista de figuras

# Lista de cuadros

# Resumen

El objetivo principal de esta investigación es lograr entender los factores que influyen en el potencial de ventas de una tienda de comercio al por menor o un restaurante. Partiendo del supuesto que la afluencia peatonal tiene una relación estrecha con la venta, además de otros factores propios de la ubicación, se busca implementar modelos de aprendizaje estadísticos que logren entender la relación y logre predecir el potencial de venta de una futura tienda.

Para esto se implementaron tres modelos: Redes Neuronales, XGBoost y Regresión Multinivel. Se usaron fuentes de datos alternativas cuyo costo y disponibilidad las hacen accesibles para los usuarios. Estas fuentes de datos miden la afluencia y movilidad de personas y ciclistas, así como también datos socio demográficos del entorno. Usando todo esto se logró obtener un resultado parcial que es capaz de predecir una ubicación mala para una tienda con una precisión del 92 % de acierto. Mientras que las buenas ubicaciones son inciertas y requieren de una mayor investigación. Estos resultados se deben a los múltiples problemas que presentan los datos, tales como el gran desbalance entre tiendas de bajo potencial y alto potencial de ventas.

Por último se concluyó que la caracterización del potencial de venta es posible, pero se deben hacer ajustes en los datos que se utilicen para entrenar el modelo.

# CAPÍTULO 1

## Introducción

La relación entre las volumen de ventas del sector de comercio al por menor y restaurantes y la afluencia de personas, ha sido un problema ampliamente estudiado. Es de principal interés ya que una buena ubicación de una tienda puede traducirse en un nivel de ventas alto, y el opuesto también sucede.

Múltiples estudios, como el de la Universidad de Portland (13), han medido de manera empírica este fenómeno, a través de medir el cambio en las ventas luego de hacer mejoras para los peatones y ciclistas en corredores comerciales a lo largo de Estados Unidos. Por lo que se sabe que esta relación existe y por tanto se busca entender todos los factores que influyen en ella, para luego poder predecirla. Este último punto es de vital importancia para las poder tomar decisiones y estrategias correctas a la hora de poner una tienda. Para esto, investigadores de la Universidad de Seúl(12), propusieron un modelo que fuera capaz de aprender y predecir las ventas de una tienda basados en su ubicación. No tuvieron los resultados que buscaban. Sin embargo, existen empresas que ya desarrollan y comercializan estas soluciones, esto fue posible gracias al avance del *Machine Learning* y datos de afluencia de personas, tales como los datos espaciales registrados por los celulares.

Con esto en mente es posible entender los factores más influyentes en la venta de una tienda del sector *retail* usando su posible ubicación. Sin embargo, este tipo de estudios tiene un precio muy elevado por la cantidad de datos usados y la infraestructura necesaria para procesarlos. Tomando en cuenta, que el sector del comercio al por menor y restaurantes tienen grandes impactos en la economía de un país, es importante que tengan una buena planificación y resultados de ventas. Además, muchos de los involucrados en este sector son pequeños empresarios o emprendedores que no tienen el presupuesto necesario para un estudio del potencial de venta de una ubicación que use datos de celulares.

Esta es la razón de realizar esta investigación. Se busca desarrollar una herramienta que permita entender el potencial de ventas de una tienda usando modelos predictivos y datos alternativos cuyo precio sea menor a los datos celulares y que sean de fácil disponibilidad. Es por ello que se escoge la Ciudad de México como piloto de esta investigación ya que existen datos de afluencia que se pueden tener de manera gratuita o a un precio mucho menor que los normalmente usados. Para esta investigación es importante tomar en cuenta que se realizará con una infraestructura computacional menor y con un equipo de expertos reducido. Además, se usarán las mejores alternativas en modelos predictivos como los son las Redes Neuronales y el XGBoost. Ambos modelos ampliamente usados en el ámbito del análisis de datos y cuya teoría matemática detrás de su funcionamiento es compleja y será explicada en esta investigación.

CAPÍTULO 2

# Objetivos

## 2.1. Objetivo general

Diseñar y proponer modelos matemáticos y estadísticos para la predicción del potencial de ventas de una actividad económica del sector de ventas al por menor y servicio de comida en un sector dado de un área de la Ciudad de México.

## 2.2. Objetivos específicos

- Analizar los factores que influyen en las ventas de los negocios de venta al por menor y servicios de comida.
- Implementar modelos matemáticos y estadísticos para la predicción del potencial de venta.
- Comparar y evaluar varios de estos modelos y determinar aquellos que son más convenientes para su uso a nivel comercial.
- Producir visualizaciones interactivas que le permitan a un potencial emprendedor, ubicar las mejores regiones de oportunidad de negocio en función de las estimaciones del potencial de venta.

# CAPÍTULO 3

## Definición del problema

El sector de las ventas al por menor y los servicios de comida juegan un papel importante en el desarrollo de la economía urbana. Una hipótesis simple consiste en suponer que las ventas de estos sectores sean proporcionales al volumen de peatones del entorno (Liu, J. y Shi, W. 2020)(12), la correlación entre la movilidad de los peatones y las ventas de los sectores de venta al por menor y servicios de comida ha sido objeto de estudio.

En este trabajo se busca analizar la movilidad de los peatones por medio de datos alternativos, y encontrar otros factores del entorno para entender la relación con las ventas de los distintos sectores de la Ciudad de México. A fin de predecir el potencial de venta de una tienda de manera más barata y rápida.

# CAPÍTULO 4

## Justificación

Hoy en día, uno de los factores más importantes para el éxito o fracaso de un negocio de venta al por menor es la correcta elección de la ubicación, para hacer esto es necesario tener predicciones del potencial de venta de un tipo de actividad económica en un sector dado.

Además, por estudios realizados en cinco ciudades de Estados Unidos (Liu, J. y Shi, W. 2020) (13)se sabe que entre más peatones y personas en bicicletas pasen por un lugar más venden los comercios, es por esto por lo que lograr medir con exactitud la afluencia peatonal es de vital importancia. Existen diversas maneras de lograr estas mediciones, una de las que mejores resultados tiene es medirla con datos de celulares. Hay que señalar que estos datos son costosos de analizar, no solo por el precio elevado de las bases de datos sino también porque se requiere de mucha capacidad computacional al tratarse de bases de datos de miles de millones de registros. El estudio se centra en un área limitada de la ciudad de México ya que en dicha zona se cuenta con un servicio de alquiler de bicicletas que es gestionado por el gobierno de la ciudad, por lo que sus datos son de acceso público, esto es de gran ventaja ya que nos da una buena muestra de la movilidad en bicicletas. Para la medición de los datos de movilidad de peatones se usan los datos que provee Facebook, la cual ha demostrado ser de gran valor a la hora de medir el movimiento de las personas (Spryatos S, Vespe, M, Natale F, Weber I, Zagheni E, Rango M. 2019).(16)

Otro de los motivos para la elección del área de estudio fue la accesibilidad de datos que brinda el INEGI (Instituto Nacional de Estadística y Geografía Mexicano) ya que este provee datos tales como una lista de empresas y sus ubicaciones llamada DENUE (Directorio Estadístico Nacional de Unidades Económicas) así como también distintas mediciones de los ingresos de las distintas actividades económicas en los distintos sectores de México, los cuales sirven para estimar el potencial de ventas actual de las empresas existentes.

Por último, la elección del periodo temporal de los datos es del primer trimestre del 2020 ya que aún no se veían efectos de la pandemia del COVID-19 y por la disponibilidad de los datos de UBER, que muestra la conexión entre los distintos sectores de la CDMX.

CAPÍTULO 5

# Marco teórico

## Teoría de los modelos predictivos utilizados

En este capítulo se abordará de manera general, sin descuidar los detalle matemáticos y computacionales, la teoría detrás del funcionamiento de los distintos modelos usados en esta investigación.

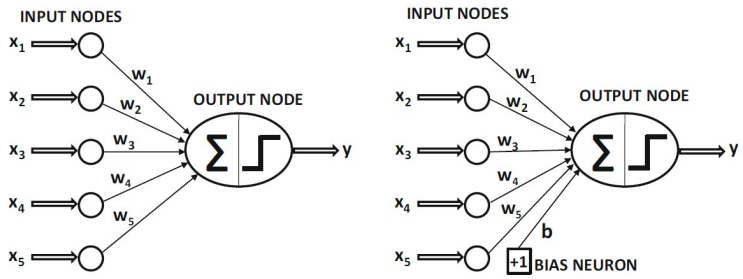
## 5.1. Redes neuronales

Las redes neuronales o redes neuronales artificiales son un subconjunto de los Modelos de *Machine Learning* y el eje central del *Deep Learning.* Estos modelos buscan simular el aprendizaje en organismos biológicos. El en sistema nervioso humano existen células llamadas Neuronas, estas están conectadas con otras a través de axones y dendritas. Las regiones conectadas por estas se les llama sinapsis. La fuerza de estas conexiones cambia en respuesta a estímulos externos. Este cambio eso como el aprendizaje toma lugar en los organismos biológicos.

Este mecanismo es el que buscan simular las redes neuronales artificiales donde a una unidad computacional se le llama neurona. Las unidades computacionales se conectan con otras a través de pesos, los cuales tienen la misma función que la fuerza de las conexiones sinápticas. Cada valor de entrada a una neurona se escala con el peso, el cual afecta a la función computada en esa unidad. Una red neuronal artificial computa una función a los valores provenientes de una neurona de entrada y el resultado lo traslada a una neurona de salida, usando los pesos como parámetros intermedios. El aprendizaje ocurre cuando se cambian los pesos que conectan a las neuronas. Al igual que en los organismos biológicos, para que esto ocurra se necesita un estimulo externo, el cual en este caso son datos que contengan ejemplos de pares de valores de entrada y salida. A estos datos ejemplo se les conoce como datos de entrenamiento. Son estos los que hacen que la red aprenda y sea capaz de reconocer patrones y predecir el resultado en datos futuros.

### 5.1.1. Arquitectura básica

La estructura más simple que se puede tener en un modelo de este tipo es el uso de una solo neurona (Perceptrón). Es decir, que se usa una capa de entrada de los datos y un nodo de salida.




(a) Perceptron without bias (b) Perceptron with bias

Figura 5.1: Arquitectura básica de un perceptrón

(Charu, C. 2018)(4)

Considere un registro de los datos de entrenamiento de la forma $(\bar{X}, y)$, donde $\bar{X} = [x_1, ..., x_d]$, siendo cada $x_i$ un atributo del registro de entrenamiento. Además, sea $y \in \{-1, 1\}$ la variable objetivo o lo que se busca predecir, en este caso es una variable binaria.

Por ejemplo, en esta investigación cada $\bar{X}$ es un punto de venta o tienda, los $x_i$ es su información como el nombre, municipio, et... La variable $y$ sería, en el caso binario, si la tienda tiene buena venta o no.

Por lo tanto, La capa de entrada tendría $d$ nodos, de los cuales cada uno transmite uno de los $d$ atributos de $\bar{X}$ con sus respectivos pesos $\bar{W} = [w_1, ..., w_n]$. Es importante notar que la capa de entrada no realiza ninguna operación. La función $\bar{W} \cdot \bar{X} = \sum_{i=1}^{d} w_i x_i$ se hace en el nodo de salida. Esta es la función que se usa para predecir la variable dependiente $\hat{y} = signo\left\{\bar{W} \cdot \bar{X}\right\} = signo\left\{\sum_{i=1}^{d} w_i x_i\right\}$

La función signo en este caso es lo que se conoce como función de activación, para cada capa que se agregue se puede usar una función diferente de activación, según sea el problema a resolver. Muchos de los modelos clásicos de *Machine Learning* se pueden ver como una red de una sola neurona, esto es importante notarlo ya que ayuda a entender cuanto más profundo se puede llegar con este enfoque.

El objetivo de este modelo es reducir al máximo el número de errores de clasificación. Por lo que en este caso el problema se reduce a:

$$\min_{\bar{W} \in \mathbb{R}^d} L(\bar{W}) = \sum (y - signo\left\{\bar{W} \cdot \bar{X}\right\})^2$$

A la función $L(\bar{W})$ se le conoce como función de pérdida. La mayoría de las redes neuronales están formuladas considerando el uso de una función de pérdida. En este caso se usa una función de mínimos cuadrados. Sin embargo, existen múltiples funciones a utilizar según sea el problema a resolver. El algoritmo de las redes está diseñado para actualizar el vector de pesos $\bar{W}$ en lotes, de tamaño menor al de los datos de entrenamiento. Por lo tanto, el vector de pesos se actualiza recurrentemente de la siguiente manera:

$$\bar{W} \Leftarrow \bar{W} + \alpha(y - \hat{y})\bar{X}$$

El parámetro $\alpha$ controla la tasa de aprendizaje de la red. Este proceso se repite en ciclos escogiendo muestras aleatorias de los datos y va ajustando el vector de pesos hasta que converge. Por lo que un registro dentro del conjunto de datos de entrenamiento puede ser usado múltiples veces durante todo el proceso. Cada ciclo se llama una época y el número de épocas es otro de los parámetros que se definen en el modelo.

Por lo tanto, el algoritmo con una sola neurona puede ser descrito como método de descenso gradiente estocástico, el cual busca minimizar el error al cuadrado de las clasificaciones. Recordando que el descenso gradiente estocástico es un método iterativo que busca optimizar una función objetivo con propiedades adecuadas, por ejemplo ser diferenciable. La diferencia entre el descenso gradiente, es que no calcula el gradiente real (con todos los datos) sino que hace una estimación del mismo usando un subconjunto aleatorio de los datos. En problemas de optimización de grandes dimensiones esto reduce la carga computacional logrando iteraciones más rápidas a cambio de una tasa de convergencia más baja.

Este ejemplo fue el más simple modelo que se puede construir, a continuación se presentan secciones sobre la elección de algunos parámetros, así como la arquitectura de una red más compleja y los problemas más comunes.

### 5.1.2. Elección de la función de activación

La elección de la función de activación es una parte crítica al implementar un modelo de red neuronal. En el caso anterior se usó la función *signo* ya que se buscaba predecir una variable binaria. Sin embargo, según sea el problema a resolver la función de activación deber ser escogida de manera correcta.

Es importante notar que cada neurona realiza dos operaciones, un producto punto entre los registros de entrenamiento y sus pesos (valor pre-activación) y luego ejecuta la función de activación (valor post-activación). Como se ve en la siguiente imagen:

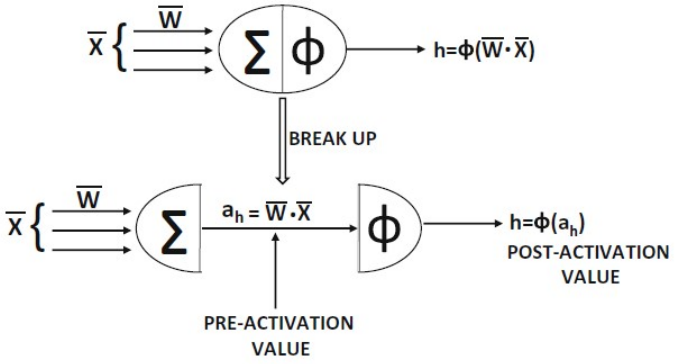



Figura 5.2: Valores pre y post activación

(Charu, C. 2018)(4)

Los valores post-activación es el resultado que devuelve cada neurona y el valor pre-activación es usado en el análisis de propagación hacia atrás, el cual será enunciado más adelante.

Una de la motivaciones de agregar funciones de activación es hacer más dinámica la red y que sea capaz de extraer información compleja de los datos, y principalmente encontrar un mapeo no lineal

entre los datos de entrenamiento y la variable a predecir. Por lo tanto, al agregar la no linealidad con las funciones de activación no lineales a la red, podemos lograr asignaciones no lineales requeridas. Este resultado es probado con la ayuda del siguiente teorema:

**Teorema 5.1.1.** *(4) Una red multicapa que use solo la función identidad como función de activación en todas sus capas se comporta como una red de una capa que realiza una regresión lineal*

Por el resultado anterior, en esta investigación se usaron funciones de activación no lineales con el objetivo de obtener buenos resultados. A continuación, se listan las funciones de activación más comunes y su uso:

- **Función Escalonada Binaria**: es la función de activación más simple. Se define de la siguiente manera:

$$f(x) = \begin{cases} 1 & \text{si } x \geq 0 \\ 0 & \text{si } x < 0 \end{cases}$$

  Es usada en modelos de clasificación binaria, evidentemente no se puede usar en clasificadores de múltiples categorías. Un problema de esta función es que su derivada es 0 por lo que no se puede implementar el método de propagación hacia atrás.

- **Función Lineal**: esta función es directamente proporcional al valor de entrada, ya que está definida de la siguiente manera: $f(x) = ax, a \in \mathbb{R}$. Esta función resuelve el problema de diferenciación de la función anterior pero sigue sin ser útil, ya que su derivada es constante por lo que al usar propagación hacia atrás no se actualizan los pesos. Sin embargo, esta función es usada cuando es necesario interpretar el funcionamiento del modelo y al realizar tareas simples.

- **Función Sigmoide**: es de las funciones más utilizadas ya que no es lineal. Esta función transforma los valores al rango $[0, 1]$, se define de la siguiente manera:

$$f(x) = \frac{1}{1 + e^{-x}}$$

$$f'(x) = f(x)(1 - f(x))$$

  Como se puede observar, la función es derivable y suave.

- **Función Tangente Hiperbólica**: es similar a la sigmoide con la diferencia de ser simétrica con respecto al origen. Se define de la siguiente manera:

$$f(x) = tanh(x)$$

  Es continua y diferenciable, su rango es de -1 a 1. Además, se prefiere a la función sigmoide ya que su gradientes que no está restringido a variar en una sola dirección y también está centrada en cero.

- **Función ReLU**: su nombre se debe a sus a la siglas en ingles (Rectified Linear Unit) es una función no lineal muy utilizada en las redes neuronales. Una de las ventajas de esta función es que no todas las neuronas se activan al mismo tiempo. Ya que por su definición, una neurona se desactiva si su resultado es menor o igual a cero. Se define de la siguiente manera: $f(x) = \text{máx}(0, x)$

  En algunos casos, el valor del gradiente es cero, por lo que los pesos y sesgos no se actualizan

durante el paso de propagación hacia atrás en el entrenamiento de la red neuronal, el cual se explica en una sección posterior. Para resolver este inconveniente existen otras variaciones de la función ReLU:

$$f(x) = \begin{cases} x & \text{si } x \geq 0 \\ ax & \text{si } x < 0 \text{ con a} > 0 \end{cases}$$

Variando el parámetro $a$ se puede lograr una convergencia más rápida.

- **Unidad Lineal Exponencial (ELU)**: es una variante de la función ReLU. En esta función se introduce un nuevo parámetro de la pendiente para los valores negativos. Se define de la siguiente manera:

$$f(x) = \begin{cases} x & \text{si } x \geq 0 \\ a(e^x - 1) & \text{si } x < 0 \end{cases}$$

- **Función Softmax**: es una combinación de de múltiples funciones sigmoides . Esta función devuelve valores entre $[0, 1]$, que pueden ser tratados como probabilidades de pertenencia a una clase o categoría. Se define de la siguiente manera:

$$\sigma(z)_j = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}}, \text{ con } j = 1, ..., K$$

En teoría de probabilidad, la salida de la función Softmax se entiende como la distribución de probabilidad sobre $K$ diferentes categorías. Esta es la gran diferencia con la función sigmoide, ya que la primera solo da la probabilidad de pertenencia categorías binarias, mientras que la Softmax funciona con más de dos categorías.

Escoger la función de activación es un trabajo que se puede volver largo y tedioso ya que cada caso requiere una investigación exhaustiva. Si bien no existe una regla de que función es la mejor en cada caso, se puede saber según el contexto del problema. Cada función tiene sus ventajas y desventajas, por ejemplo:

- Para los problemas de clasificación una combinación de funciones sigmoide dan buenos resultados. Sin embargo, debido al problema del gradiente llegando a 0, las funciones sigmoide y tanh se evitan.

- La función ReLU es muy usada y funciona de mejor manera en la mayoría de los casos. Sin embargo, si hay neurona muertas en nuestra red se recomienda una variante de la función ReLU y por último, esta función solo puede ser usada en las capas ocultas y no el la capa de salida.

Múltiples experimentos y estudios han sugerido que se empieza usando la función ReLU en las capas ocultas y dependiendo de los resultados ir variando. Además, las funciones sigmoide no son usadas en capas ocultas ya que la pendiente de la función se vuelve muy pequeña a medida que la entrada se vuelve muy grande o muy pequeña, lo que a su vez ralentiza el descenso del gradiente. (15)

### 5.1.3. Elección del número de nodos de salida

La elección del número de nodos de salida se hace según el problema a resolver. Por ejemplo, si se está usando una red para un regresión el número de nodos de salida es igual a la dimensión del

vector que se está estimando. Ahora bien, en este trabajo se está investigando el potencial de ventas de una tienda de manera categórica, por lo que se usan k-nodos de salida. Siendo k el número de categorías posibles y usando la función de activación Softmax en la capa de salida, por cada registro de entrenamiento se obtiene la probabilidad de pertenencia a cada categoría.

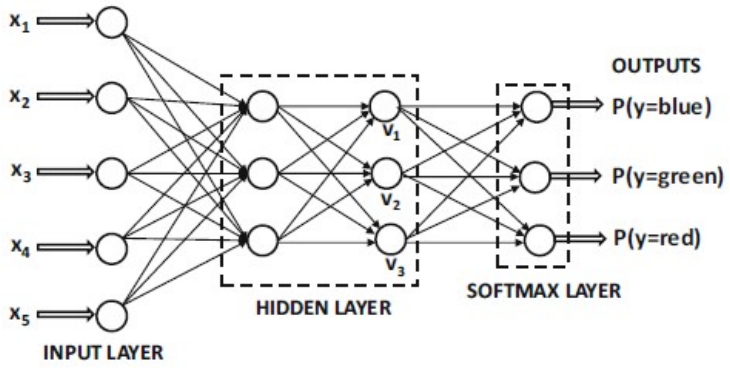



Figura 5.3: Ejemplo de múltiples salidas en un problema de clasificación usando redes neuronales (Charu, C. 2018)(4)

En este caso hay tres neuronas, ya que se está clasificando en tres categorías (Azul, verde y rojo).

### 5.1.4. Elección de la función de pérdida

La elección de la función de pérdida es un paso crítico para definir la salida del modelo y que sea congruente con el problema a resolver. Primero se exploran las funciones de pérdida para clasificación (usadas en esta investigación) y luego se mencionan las funciones $\mathcal{L}_p$ (funciones en las cuales la p-norma es finita)y sus resultados en problemas de clasificación.

**Definición 5.1.1.** Sea $(X, \mu)$ un espacio de medida y sea $p \in [0, \infty)$. Una función $\mathcal{L}_p$ en $X$ es una función medible $f$ en $X$ la cual:

$$\int_X |f|^p d\mu < \infty$$

Para predicciones de clasificación de múltiples clases, la salida softmax es útil. Sin embargo, es probabilística y, por lo tanto, requiere un tipo diferente de función de pérdida. Se utilizan dos tipos diferentes de funciones de pérdida, dependiendo del tipo de predicción:

- **Categorías binarias** (Regresión Logística): suponga que $y \in \{-1, 1\}$ y que $\hat{y}$ es la predicción del modelo. En este caso, la función de pérdida para un valor se define de la siguiente manera:

  $$L = \log(1 + e^{-y \cdot \hat{y}})$$

  Aquí se asume que la función de activación es la identidad. Sin embargo, se puede adaptar para el uso de una sigmoide cuya salida sea la probabilidad de que la predicción sea 1.

- **Múltiples categorías**: suponga que se usa la función de activación Softmax, por lo que sean $\hat{y_1}, ..., \hat{y_k}$ las probabilidades de pertenencia a las $k$ categorías. Además, sea $r$ la clase correcta de un registro de entrenamiento, entonces la función se define:

  $$L = -\log(\hat{y_r})$$

Esta función implementa regresión logística multivariable y se le llama *Pérdida de entropía cruzada*. Nótese que es una extensión del caso binario.

Como parte de una estudio realizado por investigadores de la facultad de Matemáticas y Ciencias de la Computación de la Universidad Jagellónica de Polonia(11), se analizó el uso de las funciones $\mathcal{L}_p$ para problemas de clasificación. En este caso se usan solo la funciones $\mathcal{L}_1 y \mathcal{L}_2$, que se definen de la siguiente manera:

$$L_n = \| y - \hat{y} \|_n^n$$

En el estudio se brinda una prueba para las siguientes dos proposiciones:

**Proposición 5.1.1.** *(11) La pérdida $\mathcal{L}_1$ aplicada a la probabilidad p devuelta por una función de activación (Sigmoide o Softmax) tiende a minimizar la probabilidad esperada de errores de clasificación (lo opuesto a la maximización de registros clasificados correctamente dada por la pérdida logarítmica). De manera similar, $\mathcal{L}_2$ minimiza el mismo factor, pero regularizado con la mitad de la norma L2 cuadrática esperada de las estimaciones de probabilidad de las predicciones.*

**Proposición 5.1.2.** *(11) Las pérdidas $\mathcal{L}_1$ y $\mathcal{L}_2$ aplicadas a las probabilidades devueltas por una función sigmoide o Softmax tienen derivadas parciales no monótonas con respecto a la salida de la última capa (y la pérdida no es convexa ni cóncava respecto a los pesos de la última capa). Además, desaparecen en ambos infinitos, lo que ralentiza el aprendizaje de ejemplos muy mal clasificados.*

Los resultados del estudio concluyeron que el uso de estas funciones en problemas de clasificación es posible pero que hacen lento el entrenamiento por lo cual esta alternativa no ha sido explorada a fondo. Además, concluyeron que dependiendo el problema es preferible buscar alternativas a las funciones de pérdidas logarítmicas.

### 5.1.5. Arquitectura de una red neuronal multicapas

Normalmente, cuando se implementan modelos de redes neuronales suele utilizarse más de una capa a comparación del ejemplo utilizados. A estas capas intermedias entre la capa de entrada y salida, se les conoce como capas ocultas, ya que su funcionamiento no puede ser observado por el usuario. El funcionamiento de la red es hacia adelante, es decir que cada capa tiene conectados todos sus nodos con los nodos de la capa anterior. A cada capa se le asigna una función de activación y con el métodos de propagación hacia atrás, se actualizan los pesos asignados a cada nodo.

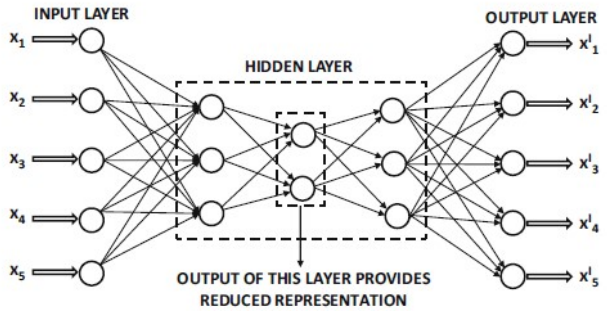



Figura 5.4: Ejemplo de un modelo de red neuronal con capas ocultas

(Charu, C. 2018)(4)

Si bien una arquitectura completamente conectada puede funcionar bien en muchos entornos, a menudo se logra un mejor rendimiento eliminando muchas de las conexiones o compartiéndolas de manera perspicaz. Por lo general, estos conocimientos se obtienen mediante el uso de una comprensión específica del dominio de los datos. En esta investigación se usan funciones para eliminar un porcentaje de los datos entre capas y capa.

### 5.1.6. Propagación hacia atrás

En las redes neuronales la propagación hacia atrás es un paso clave en el entrenamiento. Básicamente es el método por el cual se ajustan los pesos de la red en función de la tasa de error obtenida en la iteración anterior. El correcto ajuste de los pesos permite al modelo reducir su tasa de error y que sea más confiable al aumentar su generalización. El método funciona calculando el gradiente de una función de pérdida con respecto a todos los pesos de la red.

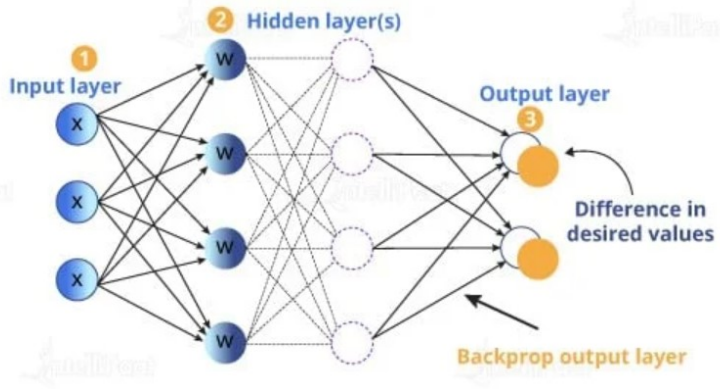



Figura 5.5: Ejemplo del algoritmo de propagación hacia atrás.

En la red neuronal de una sola capa, el proceso de entrenamiento es relativamente sencillo porque el error (o función de pérdida) se puede calcular como una función directa de los pesos, lo que permite calcular fácilmente el gradiente. En el caso de redes multicapa, el problema es que la pérdida es una función de composición complicada de los pesos en capas anteriores. El gradiente de una función de composición se calcula usando el algoritmo de propagación hacia atrás. El algoritmo aprovecha la regla de la cadena para calcular el gradiente como la suma de productos de gradiente locales en las rutas desde un nodo hasta la salida. El algoritmo tiene dos fases:

- **Fase directa**: en esta fase se ingresan los datos de entrada en una red neuronal. El resultado es una serie de cálculos hacia adelante a través de las capas, utilizando el conjunto actual de pesos. Esto llega a la salida final, la cual es comparada al valor que tienen los datos de entrenamiento. La derivada de esta pérdida se debe calcular con respecto a los pesos en todas las capas en la fase hacia atrás.

- **Fase hacia atrás**: El objetivo principal de la fase hacia atrás es aprender el gradiente de la función de pérdida con respecto a los diferentes pesos utilizando la regla de la cadena del cálculo diferencial. Estos gradientes se utilizan para actualizar los pesos. Dado que estos gradientes se aprenden hacia atrás, a partir del nodo de salida, este proceso de aprendizaje se denomina fase hacia atrás.

Al igual que con la red de una sola capa, el proceso de actualización de los nodos se repite hasta la convergencia mediante ciclos repetidos a través de los datos de entrenamiento en épocas. A

veces, una red neuronal puede requerir miles de épocas a través de los datos de entrenamiento para aprender los pesos en los diferentes nodos.

### 5.1.7. Problemas comunes de las redes neuronales

Las redes neuronales, al igual que muchos otros modelos, tienen problemas recurrentes. A continuación se listan algunos de los más comunes.(5)

- **Valores iniciales**: Generalmente los valores iniciales de los pesos se eligen de manera aleatoria, con valores cercanos a cero. Nótese que con pesos cercanos a cero, las funciones sigmoide tienen un comportamiento lineal. Por lo tanto, al iniciar el modelo tiene un comportamiento lineal y va cambiando conforme los pesos cambian.

  El uso de pesos de valor cero conduce a derivadas cero y una simetría perfecta, y el algoritmo nunca se evoluciona. Comenzar en cambio con pesos grandes a menudo lleva a las malas soluciones.

- **Sobre ajuste (*Overfitting)***: Un elemento importante que no ha sido mencionado son los datos de entrenamiento y prueba. Estos datos se usan para encontrar el modelo que mejor rendimiento tenga. Es decir, se tiene un conjunto de datos de los cuales se sabe toda su información de entrada y su valor de la variable a predecir. Este conjunto se separa, usualmente en proporciones 70 % entrenamiento y 30 %, o bien 80 % y 20 %, con los datos de entrenamiento el proceso encuentra el mejor modelo que reconozca los patrones dentro de los datos. Con los datos restantes se prueba y se obtiene una aproximación de que tan bien aprendió y como funcionará con datos de los cuales no se conozca el valor de la variable a predecir. Es muy importante que el modelo nunca se entrene con los datos de prueba ya que los resultados no serán correctos.

  El overfitting es un problema que se da cuando una red tiene un funcionamiento muy bueno en los datos de entrenamiento pero al predecir los datos de prueba funciona de mala manera. Esto ocurre cuando la cantidad de conexiones de peso es mucho mayor que la cantidad de datos de entrenamiento. En esos casos, el modelo memoriza datos específicos de los datos de entrenamiento, pero no reconoce los patrones significativos para clasificar los datos de prueba. Por lo que, aumentar el número de nodos en una red tiende a aumentar la probabilidad de un sobre ajuste.

## 5.2. XGBoost

El nombre *XGBoost* significa *Extreme Gradient Boosting*. Es una librería de código abierto que proporciona un marco de trabajo para la implementación del modelo de *Gradient Tree Boosting*. Esta es una técnica de *machine learning*, publicada por Chen y Guestrin en 2016 (6), es usada para regresión y clasificación, la cual consiste en juntar múltiples modelos predictivos débiles, generalmente árboles de decisión, para crear un solo modelo predictivo robusto. Por otra parte, el XGBoost proporciona una forma más eficaz, en términos de recursos, para lograr el resultado que otros sistemas existentes.

Por lo que en esta sección se expone la teoría de los modelos de Árbol de Gradiente Impulsado (*Gradient Tree Boosting*) y las principales diferencias entre las implementaciones clásicas y el XGBoost.

### 5.2.1. Regularización de la función objetivo de aprendizaje

Suponga que se tiene una fuente de datos $D$ con $n$ elementos o filas, y $m$ atributos o características.

$$D = \{(x_i, y_i)\}, \text{ donde } |D| = n, x_i \in \mathbb{R}^m \text{ y } y_i \in \mathbb{R}$$

Un modelo conjunto de árboles de decisión se puede expresar de la siguiente manera:

$$\hat{y}_i = \phi(x_i) = \sum_{k=1}^{K} f_k(x_i),\ f_k \in \mathcal{F}$$

Donde $K$ es el número de árboles, $f$ es una función en el espacio funcional $\mathcal{F}$ de todos los posibles árboles de clasificación y regresión (CART por sus siglas en inglés). Por lo que al final, la predicción para cada elemento es la suma de las predicciones de cada árbol.

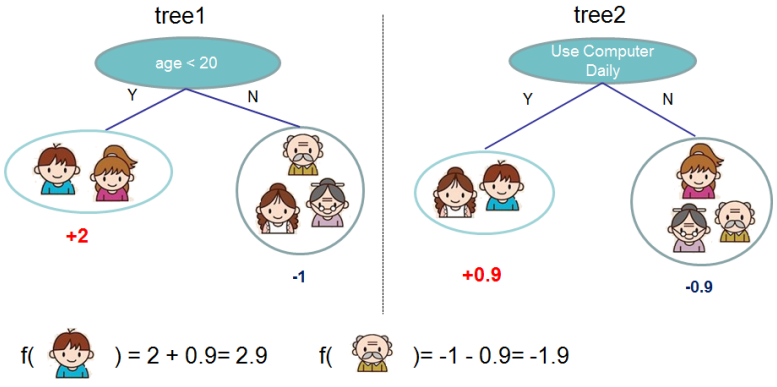



Figura 5.6: Ejemplo del modelo conjunto de árboles

(Chen, T. Guestrin, C. 2016)(6)

Ahora bien, el espacion funcional $\mathcal{F}$ se define formalmente como:

$$\mathcal{F} = \left\{ f(X) = w_q(x) | q : \mathbb{R}^m \to T, w \in \mathbb{R}^T \right\}$$

Donde $q$ representa la estructura de cada árbol que mapea un elemento a al índice de la hoja del árbol que le corresponde. $T$ es el número de hojas de cada árbol y $w$ es el vector de puntuaciones o valores de cada hoja. Entonces, se usan las reglas de cada árbol, dadas por $q$, para clasificar un elemento en su hoja correspondiente y calcular la predicción final al sumar el resultado de cada hoja.

Para que el modelo aprenda se busca minimizar el valor de la función objetivo regularizada:

$$\mathcal{L}(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k)$$

donde

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \parallel w \parallel^2$$

El primer término es la función de pérdida que mide la calidad de la predicción contra el valor real. Esta función debe ser convexa y diferenciable. Una elección común para esta función es el error medio al cuadrado:

$$l(\hat{y}_i, y_i) = (y_i - \hat{y}_i)^2$$

El segundo término $\Omega$ penaliza y controla la complejidad del modelo, esto ayuda a prevenir el sobre ajuste. Por lo tanto, la técnica de regularización de la función objetivo de aprendizaje, tiende a escoger las funciones que sean más simples y predigan mejor.

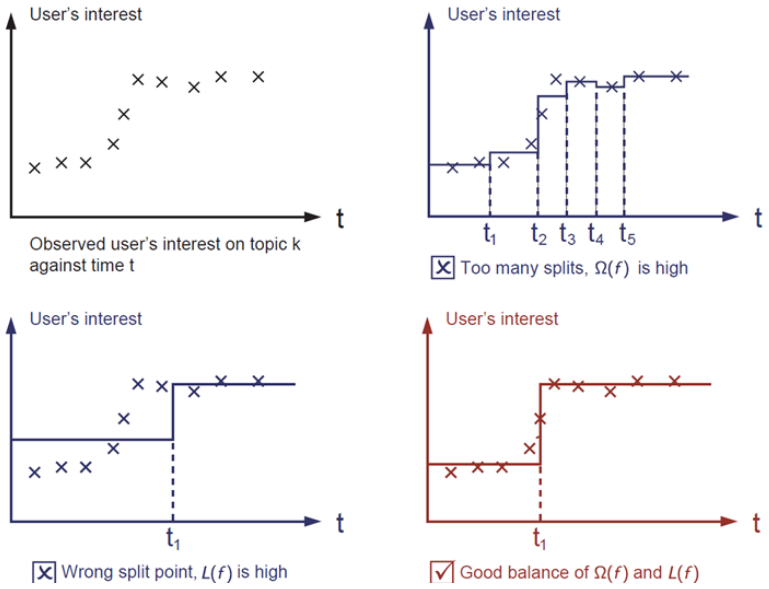



Figura 5.7: Ejemplo del funcionamiento del término de regularización $\Omega$ en la función objetivo de aprendizaje regularizada

(Chen, T. Guestrin, C. 2016)(6)

Cuando el término de regularización es 0, la función objetivo de aprendizaje se vuelve la de un modelo tradicional de Árbol de Gradiente Impulsado (*Gradient Tree Boosting*)

### 5.2.2. Árbol de gradiente impulsado (*Gradient Tree Boosting*)

El modelo conjunto de árboles, que describe la función objetivo $\mathcal{L}(\phi)$ enunciada en la sección anterior, incluye funciones como parámetros y no puede ser optimizada usando métodos tradicionales

en el espacio euclidiano. Por lo que se usa una estrategia aditiva que consiste en fijar lo que el modelo ya aprendió y añadir un nuevo árbol a la vez. Por lo que se escribe el valor de la predicción en el paso $t$ como $\hat{y_i}^{(t)}$. Por lo que se tiene lo siguiente:

$$\hat{y_i}^{(0)} = 0$$

$$\hat{y_i}^{(1)} = f_1(x_i) = \hat{y_i}^{(0)} + f_1(x_i)$$

$$\hat{y_i}^{(2)} = f_1(x_i) + f_2(x_i) = \hat{y_i}^{(1)} + f_2(x_i)$$

$$\vdots$$

$$\hat{y_i}^{(t)} = \sum_{k=1}^{t} f_k(x_i) = \hat{y_i}^{(t-1)} + f_t(x_i)$$

Entonces, usando la expansión de Taylor de segundo orden para optimizarla rápidamente se obtiene:

$$\mathcal{L}^{(t)}(\phi) \simeq \sum_{i=1}^{n} [l(\hat{y_i}^{(t-1)}, y_i) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t)$$

donde

$$g_i = \partial_{\hat{y_i}^{(t-1)}} l(y_i, \hat{y_i}^{(t-1)})$$

$$h_i = \partial^2_{\hat{y_i}^{(t-1)}} l(y_i, \hat{y_i}^{(t-1)})$$

son gradientes estadísticos de primer y segundo orden sobre la función de pérdida. Quitando todas las constantes se obtiene la siguiente expresión simplificada de la función objetivo de aprendizaje en el paso $t$:

$$\tilde{\mathcal{L}}^{(t)} = \sum_{i=1}^{n} [g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t)$$

Una vez reformulada la función objetivo de aprendizaje se define $I_j = \{i | q(x_i = j)\}$ es el conjunto de índices de los elementos $x_i$ de entrenamiento que son mapeados por $q$ a la j-ésima hoja del árbol. Como todos los elementos de entrenamiento que son asignados a la misma hoja tienen la misma puntuación, se puede escribir la función de la siguiente manera:

$$\tilde{\mathcal{L}}^{(t)} = \sum_{j=1}^{T} [(\sum_{i \in I_j} g_i) w_j + \frac{1}{2} (\sum_{i \in I_j} h_i + \lambda) w_j^2] + \gamma T$$

Para una estructura $q(x)$ fija, se puede calcular el peso $w_j^*$ óptimo de la j-ésima hoja como:

$$w_j^* = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda}$$

y se calcula el valor óptimo correspondiente como:

$$\tilde{\mathcal{L}}^{(t)}(q) = -\frac{1}{2} \sum_{j=1}^{T} \frac{(\sum_{i \in I_j} g_i)^2}{\sum_{i \in I_j} h_i + \lambda} + \gamma T$$

La expresión anterior se usa como función de puntuación para calcular la calidad de la estructura del árbol $q$. Esta puntuación es como la puntuación de impurezas para evaluar árboles de decisión, excepto que se deriva de una gama más amplia de funciones objetivo.

Normalmente es imposible enumerar todas las posibles estructuras $q$. En su lugar, se utiliza un algoritmo que parte de una sola hoja y agrega iterativamente ramas al árbol. Suponga que $I_I$ y $I_D$ son los conjuntos de los nodos a la izquierda y derecha después de la separación de los datos en cada hoja. Entonces, $I = I_L \cup I_D$, entonces la reducción de pérdidas después de la división se define:

$$\mathcal{L}_{\text{separación}} = \frac{1}{2}\left[\frac{(\sum_{i\in I_I} g_i)^2}{\sum_{i\in I_I} h_i + \lambda} + \frac{(\sum_{i\in I_D} g_i)^2}{\sum_{i\in I_D} h_i + \lambda} - \frac{(\sum_{i\in I} g_i)^2}{\sum_{i\in I} h_i + \lambda}\right] - \gamma$$

En la práctica esta expresión se usa para evaluar los candidatos para la separación, es decir si es conveniente crear un rama nueva en el árbol.

### 5.2.3. Diferencias principales del XGBoost

En la secciones anteriores se describió el funcionamiento del modelo de Árbol de Gradiente Impulsado (*Gradient Tree Boosting*). Además, se presentó con los cambios que implementa el XGBoost pero estos no fueron propiamente denotados. Por lo que en esta sección se exponen los principales cambios que se implementas y sus ventajas sobre los algoritmos ordinarios.

- **Impulso de Newton**: la principal diferencia entre el XGBoost y otros modelos de árbol de gradiente impulsado es la forma en la que calcula el gradiente. Los modelos tradicionales calculan el gradiente y siguen la técnica de descenso gradiente, mientras que aquí se usa la derivada de segundo orden para encontrar hacer una mejor aproximación de la dirección de máxima disminución de la función de pérdida (en este modelo es error medio al cuadrado). Esto hace que la convergencia se de de una manera más rápida.

- **Complejidad del modelo**: mientras otras implementaciones solo se centran en disminuir la impureza y maximizar la ganancia del modelo, el XGBoost aplica técnicas para evadir el sobre ajuste. El término $\Omega$ en la función objetivo de aprendizaje es el encargado de esto.

- **Puntuación de la estructura**: al reescribir el modelo usando la expansión de Taylor se obtiene una métrica para puntuar la estructura de un árbol. Esto ayuda a buscar árboles que minimicen la pérdida total del modelo.

- **Aprender a construir la estructura**: con la métrica anterior se puede medir la calidad de un árbol. El siguiente paso sería computar todas las posibles estructuras y quedarse con la mejor. Sin embargo, este proceso es computacionalmente muy caro, por lo que el algoritmo del XGBoost utiliza otra técnica. Optimiza un nivel a la vez, entonces partiendo de la primera hoja, usa la función $\mathcal{L}_{\text{separación}}$ (enunciada en la sección anterior) para expresar la ganancia de agregar una nueva división o rama al árbol. Por lo tanto, estamos usando la función de pérdida para construir el árbol controlado por la complejidad del modelo.

## 5.3. Regresión multinivel

El último modelo propuesto en esta investigación es la regresión multinivel. La idea de usar este modelo es dada por una de las investigaciones consultadas previamente. Sin embargo, estos modelos son antiguos que los dos presentados anteriormente. En esta investigación no se obtuvieron resultados con su implementación y esto puede deberse a que tanto las redes neuronales y XGBoost son mucho mejores para reconocer los patrones en los datos utilizados. Por lo tanto, en esta sección se presenta una breve explicación de que son y la teoría detrás sin entrar en el detalle de los anteriores.

### 5.3.1. Modelos multinivel

Muchos tipos de fuentes de datos tienen estructura jerárquica o agrupada. Los modelos multinivel (7) reconocen la existencia de estos tipos de estructura al permitir componentes residuales en cada nivel de la jerarquía. Por ejemplo, para un modelo de dos niveles que estudie las calificaciones de niños de diferentes escuelas, incluiría residuos a nivel de los niños y a nivel de la escuela. Por lo tanto la varianza de los residuos estaría partida en dos partes, una de parte la varianza entre escuelas que sería la varianza de los residuos del nivel de las escuelas. La otra parte sería la varianza dentro de las escuelas que es la varianza de los residuos del nivel de los niños. A los residuos del nivel escuelas se les llama ‹‹Efecto de las escuelas››, el cual representa a las variables de la escuela, que no son observadas, que afectan las calificaciones de cada niño. Son estas variables que llevan a la correlación de las notas de niños de la misma escuela.

Algo a tener en cuenta en estos modelos es el tipo de efectos que causan las variables y sus niveles. Existen efectos fijos, de los cuales ya se sabe que tienen un efecto en la variable dependiente y los efectos aleatorios ya que no se conoce en que medida afectan estos a la variable dependiente. En el caso del ejemplo anterior, el efecto de las escuelas es aleatorio ya que no se sabe como este afecta a las calificaciones de un niño. El decidir el tipo de efecto depende del contexto y objetivo de la investigación correspondiente.

En el caso de los modelos multinivel, las unidades o agrupamientos que definen un nivel son vistas como de efecto aleatorios. Por ejemplo, la escuela o en el caso de esta investigación las zonas de movimiento de UBER. El considerar los efectos como aleatorios hace que el modelo se convierta en un modelo de coeficiente aleatorios que toma la varianza entre los grupos.

### 5.3.2. Modelo de regresión básico de dos niveles

Este modelo de regresión multinivel presupone la existencia del conjunto de datos jerárquicos, con una variable dependiente que se mide en el nivel más bajo. Por ejemplo, en esta investigación son las ventas de cada tienda las cuales son medidas en el nivel de tiendas, el más bajo. También toma en cuenta todas las demás variables explicativas que son medidas a todos los posibles niveles. Teóricamente, este modelo puede ser visto como un sistema jerárquico de ecuaciones de regresión.

Siguiendo con el problema de esta investigación, sea la variable dependiente $y_{ij}$ la venta de una tienda $i$ en la zona de movimiento $j$, que es definida como predictores del nivel 2. Además, es válido esperar que las zonas de movimiento de UBER tengas ventas promedio diferentes, por eso sabiendo en que zona de movimiento de UBER se encuentra una tienda dice mucho de sus posibles ventas. El único predictor del nivel 1 que se tiene es el código de actividad económica. Por lo que un modelo adecuado para esto sería:

$$y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + \epsilon_{ij}$$

La expresión anterior se llama la ecuación de regresión 1. Donde,

- $X_{ij}$: es el predictor del nivel 1, es decir el código de actividad económica.
- $\beta_{0j}$: es el intercepto de la variable dependiente en la zona de movimiento j (Nivel 2)
- $\beta_{1j}$: es la pendiente de la relación en la zona de movimiento j (Nivel 2) entre el código de actividad económica (Nivel 1) y la variable dependiente (venta).
- $\epsilon_{ij}$: se refiere a los errores de la predicción para la ecuación del nivel 1.

Como el modelo es de componentes aleatorios, las intersecciones y pendientes varían aleatoriamente en los diferentes grupos (zonas de movimiento), y cada una tiene su propia media y varianza general. Cuando hay más de un predictor en algunos de los niveles, se pueden sustituir los vectores por matrices. Además, si la relación entre $Y_{ij}$ y $X_{ij}$ no es lineal se puede extender a una regresión logística multinivel. Por otra parte,

$$\beta_{0j} = \gamma_{00} + \gamma_{01} W_j + u_{0j}$$

$$\beta_{1j} = \gamma_{10} + u_{1j}$$

Donde,

- $\gamma_{00}$: es el intercepto general. Es decir, la media de la variable dependiente (venta) de todos los grupos cuando todos los predictores son igual a 0.
- $W_j$: es el predictor del nivel 2, en este caso son todas las variables medidas a nivel de zona de movimiento de UBER.
- $\gamma_{01}$: es la pendiente entre la variable dependiente (venta) y el predictor del nivel 2.
- $u_{0j}$: es el error aleatorio para la desviación de la intersección de un grupo con respecto a la intersección general.
- $\gamma_{10}$: es la pendiente entre la variable dependiente (venta) y el predictor del nivel 1 (código de actividad económica).
- $u_{1j}$: es el error de la pendiente.

### 5.3.3. Correlación intraclase

Suponga que no existen variables explicativas en ninguno de los niveles. Entonces el modelo se convierte en:

$$y_{ij} = \gamma_{00} + u_{0j} + \epsilon_{ij}$$

Para esto se asume también que: $u_{0j} \sim N(0, \sigma^2_{u_{0j}})$ , $\epsilon_{ij} \sim N(0, \sigma^2_{e_{ij}})$ y $COV(\epsilon_{ij}, u_{0j}) = 0$. Entonces, la varianza de cada predicción sería: $VAR(Y_{ij}) = \sigma^2_{e_{ij}} + \sigma^2_{u_{0j}}$. Por lo tanto, la covarianza para dos observaciones $i$ de distintos grupos $j$ y $k$, por propiedades de la covarianza:

$$Cov(Y_{ij}, Y_{ik}) = \sigma^2_{u_{0j}}$$

Entonces,

$$Cor(Y_{ij}, Y_{ik}) = \frac{Cov(Y_{ij}, Y_{ik})}{\sqrt{VAR(Y_{ij}) VAR(Y_{ik})}} = \frac{\sigma^2_{u_{0j}}}{\sigma^2_{e_{ij}} + \sigma^2_{u_{0j}}}$$

Esta expresión es la correlación intraclase, y es un estimador de la proporción de varianza explicada en la población. Esto se usa para determinar si la consideración de un nivel es apropiada para realizar un modelo de regresión multinivel. Generalmente un valor mayor a 0.05 justifica el uso de este tipo de regresiones, sino el valor es menor se recomiendo usar una regresión normal.

CAPÍTULO 6

# Trabajos previos

## Breve explicación de los trabajos realizados con anterioridad del mismo tema

En esta sección se mencionan y resumen los trabajos consultados previo a realizar el estudio. Algunos sirvieron para justificar los supuestos utilizados en el estudio. Otros fueron usados como referencia en cuanto a técnicas y data que se debía considerar y por último otros trabajos fueron consultados con el objetivo de confirmar la calidad de las fuentes de datos alternativas.

### 6.1. Entendiendo el impacto económico y comercial de las mejoras de las calles para bicicletas y movilidad: *Una Exploración de múltiples enfoques y ciudades.* (13)

Este estudio realizado por investigadores de la universidad del estado de Portland, para el instituto nacional de transportes y comunidades, se centra en comprobar de manera empírica que la movilidad y accesibilidad de los peatones y ciclistas tiene un impacto económico significativo.

Para lo cual se analizaron distintas calles, a las cuales se les harían mejoras para hacerlas más accesibles, de cuatro ciudades de Estados Unidos. La ciudades analizadas fueron: Portland, OR. San Fransisco, CA. Minneapolis, MN. y Memphis, TN.

La metodología usada fue la siguiente:

1. **Selección del corredor:** las mejoras de las calles usualmente se realizan en una sección grande de la calle o corredores, por lo tanto todos los datos analizados se usarán a nivel corredor para poder hacer la comparación de la mejor manera. Los corredores seleccionados fueron aquellos que estuvieran ubicados en un área comercial.
2. **Análisis econométrico:** para examinar el impacto económico de los corredores con mejoras, se realizó un análisis econométrico espacial en los corredores con mejoras y los corredores de control correspondientes, para estimar los impactos en las actividades económicas y comerciales

en todos los sectores industriales. Este análisis se hizo usando distintas técnicas descritas a continuación.

*a*) **Análisis de tendencias agregadas:** El enfoque compara las tendencias de los corredores de interés y control además de las tendencias de toda la ciudad durante el período de tiempo completo para el que se tiene datos. Si los corredores de tratamiento muestran mayores tasas de crecimiento en el empleo o los ingresos por impuestos sobre las ventas, o un aumento en el nivel de empleo o ventas, eso representaría un impacto positivo de la mejora de las calles en las actividades comerciales. Este método es fácil de seguir y representa la tendencia agregada de las actividades comerciales.

*b*) **Análisis de diferencias en diferencias**: El segundo método tiene como objetivo estimar la diferencia en la vitalidad comercial de los períodos previos y posteriores a la mejora entre los corredores de tratamiento y control dentro del mismo período de tiempo. Esto se conoce como enfoque de diferencias en diferencias Está diseñado para responder a la pregunta "si no fuera por" de cómo sería la trayectoria económica de un corredor si no se hubieran mejorado las calles. Requiere datos de antes y después de la intervención. El enfoque analiza el cambio en la variable de interés en el corredor de tratamiento antes y después de que se trate. En este caso, esto significa observar algún período de tiempo antes y después de una mejora de la calle y comparar los indicadores económicos con el corredor de control que no ha recibido la mejora de la calle. La diferencia en las trayectorias de crecimiento entre los dos períodos dará una estimación imparcial del efecto del tratamiento.

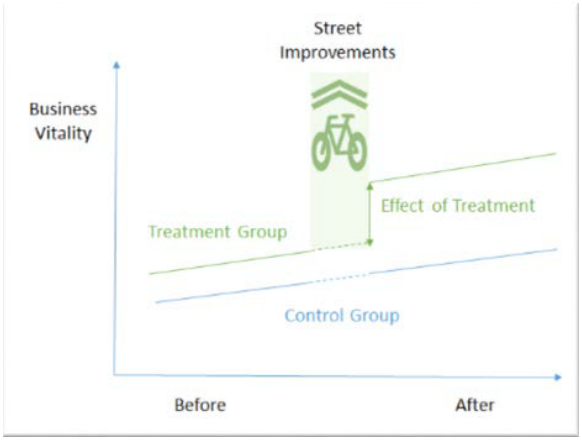



Figura 6.1: Modelo diferencias en diferencias aplicado a corredores.

(Liu, J. y Shi, W. 2020)(13)

*c*) **Series de tiempo interrumpida**: La serie de tiempo interrumpido es una técnica econométrica que estima cómo las mejoras en las calles impactan la vitalidad económica del corredor desde una perspectiva longitudinal. Este enfoque rastrea el corredor de tratamiento a lo largo del tiempo y estima el impacto de la mejora de la calle mediante la identificación de cambios en su tendencia de crecimiento después de las mejoras. Si el tratamiento tiene un impacto causal, los indicadores económicos posteriores a la intervención tendrán un nivel o pendiente diferente a los puntos de datos previos a la intervención. En esta investigación, se usa para distinguir las diferencias en el nivel económico o el crecimiento antes y después de un período de tiempo específico cuando se construya una mejora en la calle, como un nuevo carril para bicicletas protegido o protegido.

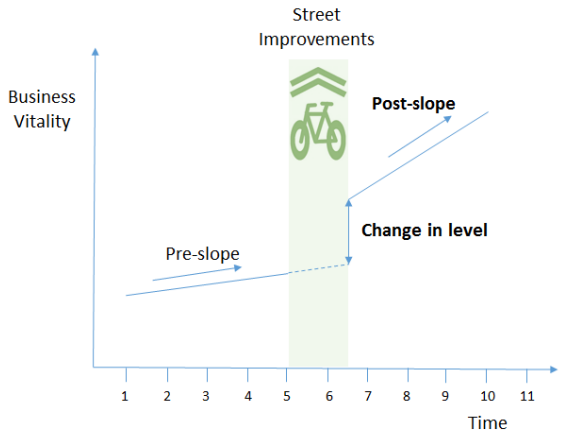



Figura 6.2: Modelo de series de tiempo interrumpida aplicado a corredores.

(Liu, J. y Shi, W. 2020)(13)

3. **Análisis de distribución**: Para comprender los impactos en la equidad y la diversidad de las mejoras en las calles, se llevo a cabo un análisis de distribución para caracterizar la distribución y la tendencia de las oportunidades de empleo y residencia en los corredores de interés/control. Este análisis de distribución sirve para brindar un examen aproximado de si hay cambios demográficos significativos de los residentes a lo largo de los corredores de mejoramiento de calles en comparación con las tendencias a lo largo de los corredores de control correspondientes o dentro de la ciudad en su conjunto.

Los resultados obtenidos por el estudio variaron dependiendo la ciudad. En Portland se encontró en algunos de los tres análisis un impacto positivo en las ventas y contratación minorista. Los resultado en San Francisco mostraron impactos positivos en un corredor, mientras que el otro mostró que no hubo impacto en las ventas minoristas y en el área de servicio de comida hubo un impacto negativo. En Minneapolis se encontró evidencia de un impacto económico positivo. Por último, en Memphis se tuvo resultados de impactos negativos o insignificantes en el área de estudio. Así como algunas contradicciones en la data de un corredor.

## 6.2. Acceso espacial a peatones y ventas minoristas en Seúl, Corea del Sur. (12)

El estudio investiga el efecto de la accesibilidad y la centralidad del volumen peatonal en las ventas minoristas en Seúl, considerando el volumen peatonal y la estructura de la red de calles. Los modelo de regresión multinivel confirman que el acceso espacial de los peatones tiene diferentes efectos de las ventas al por menor, según el tipo de actividad económica. Específicamente, una mayor accesibilidad y visibilidad, de las tiendas minoristas, para los peatones tienden a mejorar las ventas de todas las actividades económicas de venta al por menor. Además, las ventas de los sectores de servicios médicos y educación son notablemente sensibles a los efectos combinados de la configuración de calles y peatones, a diferencia de los otros tres sectores: alimentos, comercio minorista y servicios.

El estudio fue usado como guía para elaborar este proyecto, ya que es un acercamiento similar al que se plantea en este trabajo. Además, se buscó obtener los mismos datos en México de los que se usaron en Corea. Sin embargo, el estudio no tuvo resultados favorables al aplicar los modelos y los investigadores lo atribuyen a la falta de datos, tales como factores espaciales de cada tienda, precios

relativos, calidad del servicio y datos de movilidad con bicicletas. Así como incluir más sectores de la ciudad, o bien incluir más ciudades para que las conclusiones sean extrapolables y aplicables a nivel general.

## 6.3. Cuantificando los patrones de movilidad humana internacional utilizando datos de la red de Facebook. (16)

El estudio realizado por investigadores de la universidad de Tel Aviv, busca usar fuentes de datos alternativas para llenar los vacíos existentes en las estadísticas de migración. Para lo cual se usó datos anónimos proporcionados por Facebook.

Basándose en estadísticas sobre los usuarios de la red Facebook que han vivido en el extranjero y aplicando un método de corrección de sesgo de muestra, se estimó el número de "migrantes"de la red Facebook (FN) en 119 países de residencia y en dos períodos de tiempo por edad, sexo y país de residencia anterior. Además, se estimó la correlación entre las estimaciones de migración derivadas de FN y las estadísticas de migración oficiales de referencia. Al comparar las estimaciones de migración derivadas de FN en dos períodos de tiempo diferentes, enero-febrero y agosto-septiembre de 2018, se capturó con éxito el aumento de migrantes venezolanos en Colombia y España en 2018.

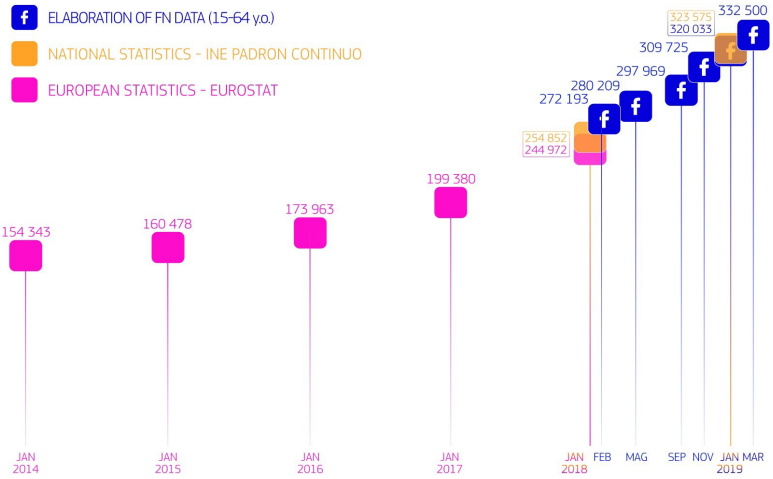



Figura 6.3: Comparativa en el tiempo de los inmigrantes venezolanos en España

(Spyratos S. Vespe M. Natale F. Weber I. Zagheni E. Rango M. 2019)(16)

El estudio concluye que las estimaciones de migración derivadas de FN no pueden reemplazar las estadísticas oficiales de migración, ya que no son representativos, y los métodos exactos que utiliza la FN para clasificar a sus usuarios no se conocen y pueden cambiar con el tiempo. Sin embargo, después de evaluar cuidadosamente la validez de las estimaciones derivadas de FN comparándolas con datos de fuentes confiables, se llegó a la conclusión de que estas estimaciones se pueden usar para el análisis de tendencias y con fines de alerta temprana.

## 6.4. Índice de tiempo de viaje ponderado basado en los datos de movimiento de UBER. (17)

En este artículo, se combinan datos de Movimiento de Uber y de una encuesta representativa de viajes en hogares para construir un índice de tiempo de viaje ponderado para la Región Metropolitana de Sao Paulo. El índice se calcula en función del tiempo de viaje promedio de los viajes de Uber realizados entre cada par de zonas de tráfico y en cada hora entre el 1 de enero de 2016 y el 31 de diciembre de 2018. El índice se pondera en función de los viajes informados en una encuesta de viajes de hogares que fue diseñado para ser estadísticamente representativo de todos los viajes realizados en la ciudad durante un día laboral típico.

Se muestra que el índice tiene una fuerte correlación con las medidas tradicionales de congestión, sin embargo, con una cobertura más amplia de la red vial. Finalmente, se usa el índice para ejecutar un análisis que estima el efecto de diferentes eventos en la congestión del tráfico en la ciudad, incluidos días festivos, huelgas de transporte público, cierres de carreteras, lluvia y eventos deportivos importantes.

El artículo termina reconociendo las posibles mejoras que se pueden e indicando que los resultado obtenidos son un esbozo inicial de todo lo que se puede desarrollar. Por último, se recomienda la incorporación de otros estudios y la exploración a fondo de los datos de movimiento de UBER que resultaron ser valiosos.

## 6.5. Identificación y caracterización de puntos de venta en canal tradicional. (3)

La empresa Predik Data-Driven desarrolló un producto similar al que se busca en esta investigación. Lo que ellos ofrecen es la capacidad de caracterizar el potencial de venta de puntos de venta en el canal tradicional, se enfocan sobre todo en el sector retail, restaurantes y gasolineras. Esto lo logran mediante la recopilación de información en cada punto como la alfuencia peatonal alrededor de la tienda, la población flotante, capacidad de compra de los residentes y visitantes y un perfil socio demográfico. Esto lo logran mediante la aplicación de modelos de *Machine Learning* tales como Xgboost, Linear Learner, Redes Neuronales, entre otros. Es importante notar que esto lo hacen con datos de entrenamiento otorgados por el cliente de sus ubicaciones para poder estimar el potencial de nuevas tiendas.

El nivel de detalle que el producto maneja es bastante específico ya que llega a múltiples dimensiones con el resultado, como se puede ver a continuación:

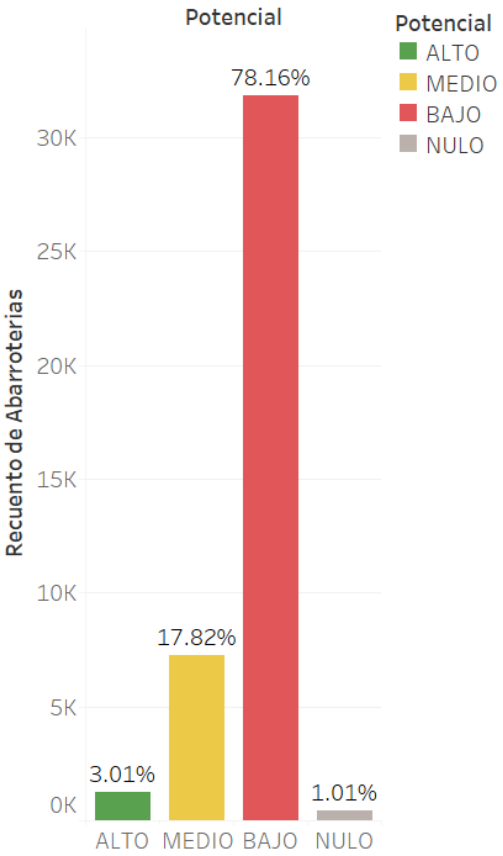



Figura 6.4: Distribución del potencial de venta da las abarroterías en la CDMX realizado por Predik Data-Driven(3)

Así como también se desarrolla el mapa de predicciones para toda la ciudad:

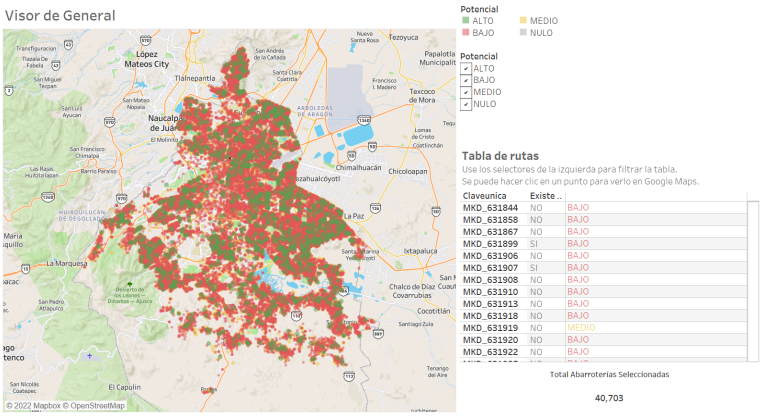



Figura 6.5: Mapa de las abarroterías en CDMX con su potencial de venta realizado por Predik Data-Driven(3)

Esto es un caso se analizaron solo las ventas de las abarroterías y minisuper. Además, aseguran que un gran caso de éxito que han tenido es que una empresa de consumo masivo que ya utiliza esta solución de análisis de puntos de venta para redefinir su estrategia en tiendas retail, estimó que por

cada ruta de distribución que optimizaron, incrementaron su facturación entre 5 % y 10 % en cada punto de venta.

## 6.6. Aporte y usos de cada estudio a la investigación

Los trabajos mencionados anteriormente son los más valiosos de todos los consultados como parte de la investigación previa. El primero (*Entendiendo el impacto económico y comercial de las mejoras de las calles para bicicletas y movilidad: Una Exploración de múltiples enfoques y ciudades*(13)) es el que confirma el supuesto de la relación entre la movilidad de los peatones con las ventas minoristas y de servicio de comidas.

El segundo estudio (*Acceso espacial a Peatones y ventas minoristas en Seúl, Corean del Sur*(12)) sirvió cómo guía para el desarrollo de esta investigación. Las fuentes de datos usadas buscaron parecerse lo más posible a las de ese estudio, y al igual que ellos se tuvo limitantes en las fuentes de datos que se utilizaron. Además, introdujo el uso de los modelos multinivel que al principio no se tenía contemplado.

Tanto el tercer estudio (*Cuantificando los patrones de movilidad humana internacional utilizando datos de la red de Facebook*(16)) y el cuarto artículo (*Índice de Tiempo de Viaje Ponderado basado en los datos de Movimiento de UBER*(17)) sirvieron para validar y justificar el uso de las fuentes de datos alternativas de movilidad como lo son Facebook y UBER.

Por último, el análisis que realiza Predik Data-Driven(3) asegura que es posible el implementar estas soluciones y estimar el potencial de venta mediante modelos de Machine Learning, la única consideración a tomar en cuenta es que ellos tienen acceso a más fuente de datos y a una estructura computacional más robusta.

# CAPÍTULO 7

# Marco metodológico

## 7.1. Definición del área a trabajar y las fuentes de datos

En esta sección se presentan las fuentes de datos a que se usarán, con todos sus atributos. Además, se explica el proceso (donde sea posible) de cómo se construyen las fuentes de datos y de como estas definen el área de estudio.

1. ECOBICI (1):

   En el 2010 inició un servicio de bicicletas públicas en la Ciudad de México. Actualmente cuenta con más de 6800 bicicletas y 480 estaciones en diversos puntos estratégicos. Al ser un servicio público se puede obtener con facilidad los datos de las distintas transacciones que realizan los usuarios. Como se explicó en la justificación, únicamente se toman en cuenta las transacciones del primer trimestre del 2020.
   Además, la ubicaciones de las estaciones de ECOBICI son las que delimitan el área de estudio dentro de la CDMX.

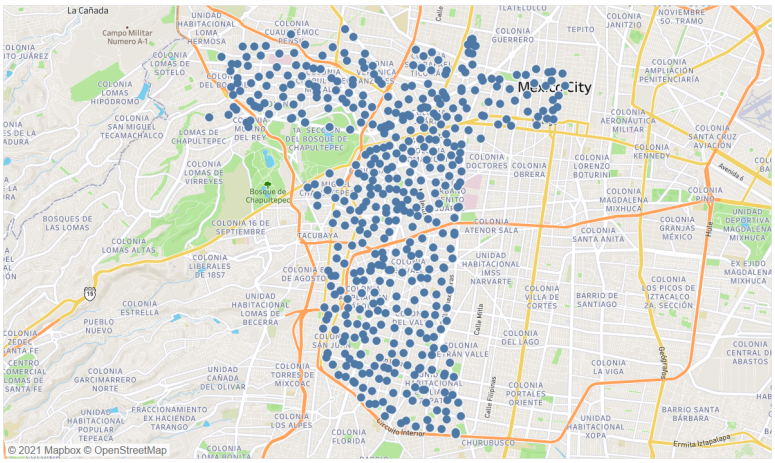



Figura 7.1: Ubicaciones de las estaciones de ECOBICI en la CDMX

2. Metro de la CDMX (14)

   Como parte de la investigación previa se observó el uso de datos del entorno de cada tienda y sector. Por lo tanto se tienen en cuenta las estaciones de metro dentro de cada sector de la ciudad. Estos datos son públicos por ser un servicio del estado, en este caso se usa únicamente las ubicaciones de la estaciones de metro dentro del área de estudio.

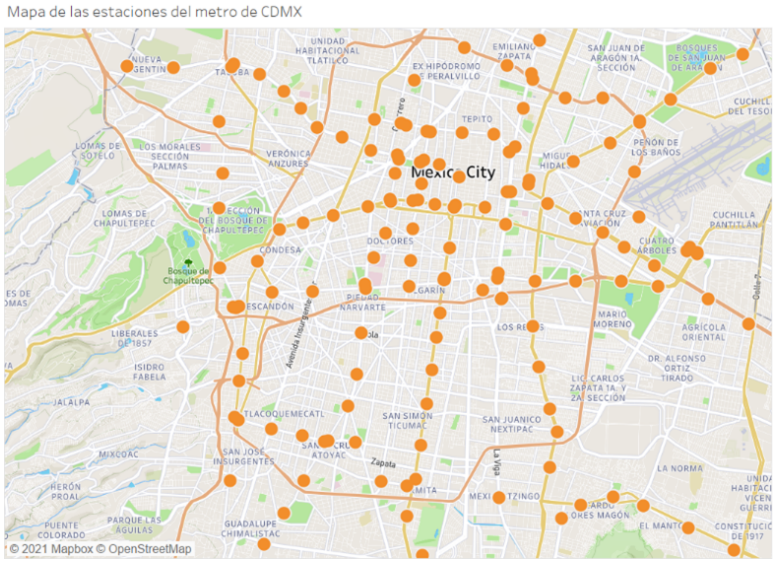



Figura 7.2: Ubicaciones de las estaciones del metro de la CDMX

3. Directorio Estadístico Nacional de Unidades Económicas (DENUE) (9)

   En esta fuente de datos provista por el INEGI, se listan las ubicaciones de todas las empresas registradas en México. Esta fuente de datos tiene muchos atributos por cada empresa, lo más importantes en este caso: Actividad económica y las coordenadas.

   Se utilizan datos provistos por una empresa particular que, con los datos económicos del INEGI, hace una estimación de las ventas anuales de cada empresa. Esta estimación se usa, no como un dato duro, sino que se usa para hacer una variable categórica del potencial de ventas.

   De todas las empresas localizadas en la Ciudad de México se utilizaron únicamente las que estuvieran cerca de alguna estación de ECOBICI y fueran del sector de ventas al por menor y servicio de comida. Esta fuente de datos es con la que se entrenan los modelos predictivos.

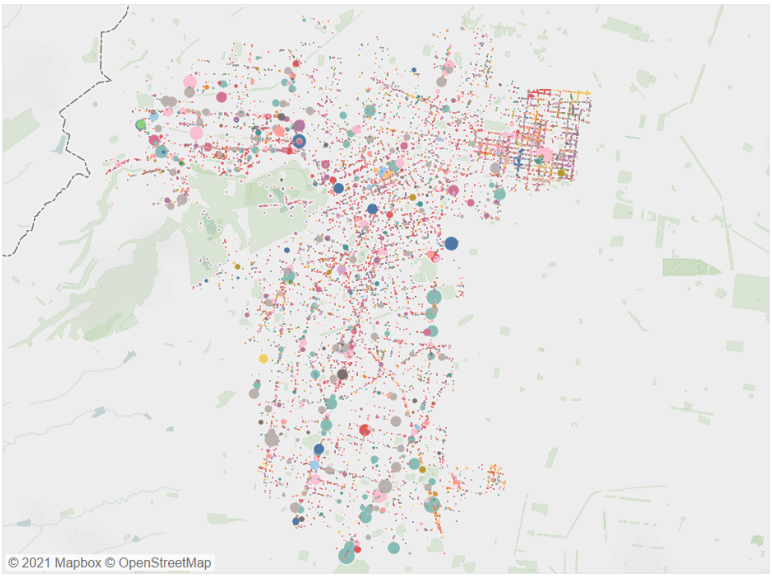



Figura 7.3: Ubicaciones de todos los negocios registrados en el área de influencia de las estaciones de ECOBICI

4. Población flotante (3)

   La fuente de datos de Población Flotante, es provista por la misma empresa particular del apartado anterior. Los datos de población flotante consisten en la población que se mantiene en un lugar, no debe ser necesariamente su hogar. Por ejemplo, si una persona pasa la mitad de su tiempo en su casa y la otra mitad en su trabajo, contará cómo media persona en cada lugar. Estos datos consisten en un cruce de los datos de Facebook y antenas celulares y la información viene dada por cuadrantes de 100 m X 100 m. Los atributos de esta base de datos son los siguientes

   - Cuadrante ID
   - Polígono del cuadrante: Este es un geo JSON con el polígono que encierra cada cuadrante.
   - Porcentaje de la población de 15 a 20 años.
   - Porcentaje de la población de 21 a 30 años.
   - Porcentaje de la población de 31 a 45 años.
   - Porcentaje de la población de 46 a 59 años.
   - Porcentaje de la población de 60 años o más.
   - Porcentaje de la población que es mujer.
   - Porcentaje de la población que es hombre.
   - Edad promedio.
   - Población total.
   - Población con nivel socioeconómico CDE.
   - Población con nivel socioeconómico AB.
   - Población con nivel socioeconómico C+.

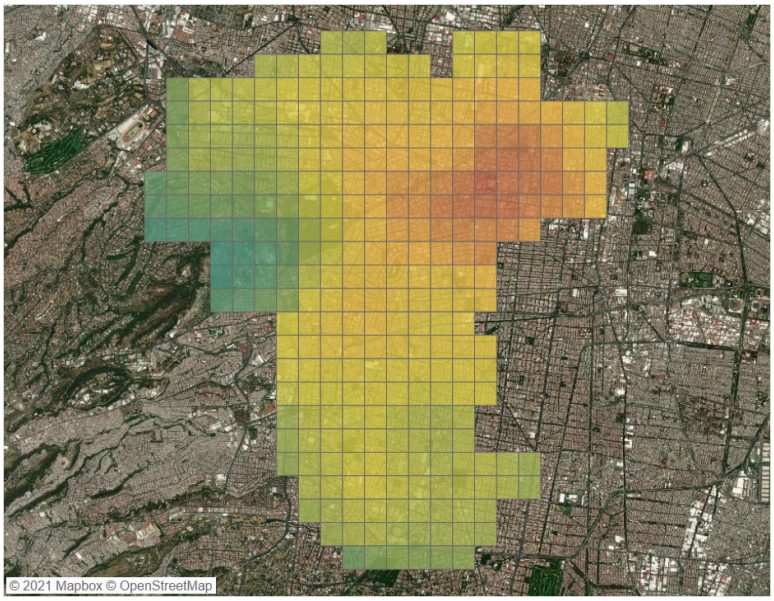



Figura 7.4: Cuadrantes de la población flotante dentro del área de influencia de las estaciones de ECOBICI

5. Características de las localidades y del entorno urbano (8)

   Al igual que los datos del Metro, esta fuente de datos provista por el INEGI brinda un mejor entendimiento de la distribución del entorno y la composición de un lugar en específico. Los datos tienen resolución a nivel de manzanas, por lo que es posible saber cosas como:

   - Disponibilidad de recubrimiento de la calle
   - Disponibilidad de banqueta según disponibilidad de recubrimiento de la calles
   - Disponibilidad de banqueta según disponibilidad de guarnición en sus vialidades
   - Disponibilidad de rampa para silla de ruedas en sus vialidades
   - Disponibilidad de banqueta según disponibilidad de árboles o palmeras en sus vialidades
   - Disponibilidad de teléfono público según disponibilidad de letrero con nombre de la calle en sus vialidades
   - Restricción del paso a automóviles según restricción del paso a peatones en sus vialidades
   - Presencia de puesto ambulante según presencia de puesto semi fijo en sus vialidades
   - Alumbrado público según tipo de recubrimiento de la calle

   Este tipo de datos han demostrado ser de utilidad en un estudio parecido en Seúl, Corea (Chang-Deok, K. 2016)(12)

6. UBER (2)

   Como parte de una iniciativa, UBER liberó ciertos datos de los viajes realizados por sus usuarios en algunas ciudades selectas.

   Los datos consisten en tiempos de viajes entre zonas de movimiento, que son particiones de las ciudades basadas en las definiciones estatales. De esto se obtuvo, el tiempo promedio de los viajes que salen y llegan a cada zona. También se contó el número de zonas con las que se conectan cada una. También proveen las geometrías de cada zona, esto es muy útil para saber cuáles estaciones y empresas están en cada zona de movimiento.

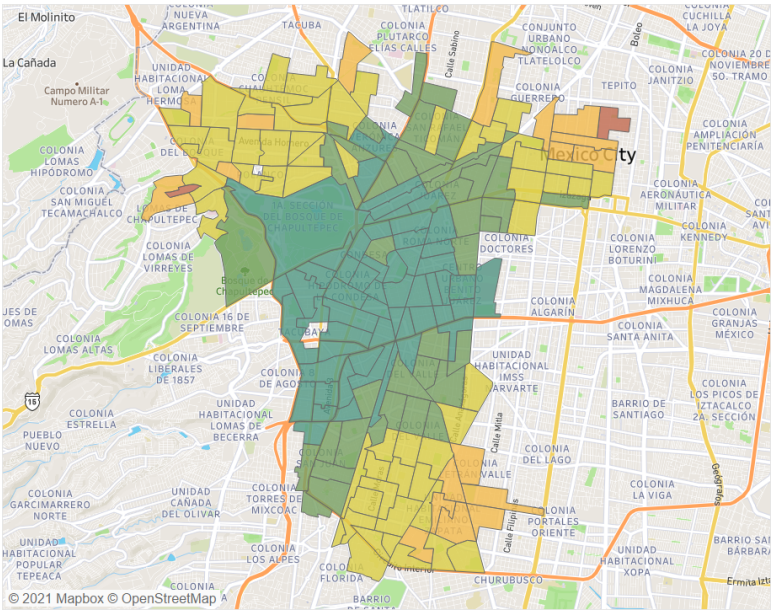



Figura 7.5: Zonas de movimiento de UBER con métricas de tiempos de viaje delimitadas por el área de influencia de las estaciones de ECOBICI.

## 7.2. Recopilación y preparación de los datos

Previo a este proyecto se realizaron estudios sobre los datos de ECOBICI, por lo que ya se sabía con anterioridad que variables utilizar.

El primer paso fue obtener las transacciones del primer trimestre del 2020, para todas la estaciones de ECOBICI. Las variables de interés son el total de bicicletas depositadas y retiradas, este es un buen insumo de cuánta gente transita alrededor de esa zona, además se generó una métrica de rotación de bicicletas, esta métrica sirve para entender mejor la situación del sector, ya que si la rotación es baja significa que ese sector no es el destino de los usuarios, ya sea porque toman el metro o porque se desplazan a otro lugar mediante otras formas de transporte. Las métricas son calculadas de la siguiente manera:

- Número total de bicicletas por estación:

$$\sum_{m=1}^{\text{meses}} \sum_{d=1}^{\text{días del mes}} depositadas_{d,m} + retiradas_{d,m}$$

- Promedio de rotación de bicicletas por estación:

$$\frac{1}{\text{meses}} \sum_{m=1}^{\text{meses}} \frac{1}{\text{días del mes}} \sum_{d=1}^{\text{días del mes}} 1 - \frac{|depositadas_{d,m} - retiradas_{d,m}|}{depositadas_{d,m} + retiradas_{d,m}}$$

Para ambas métricas se generan las siguientes vistas con el objetivo de obtener una explicación visual:

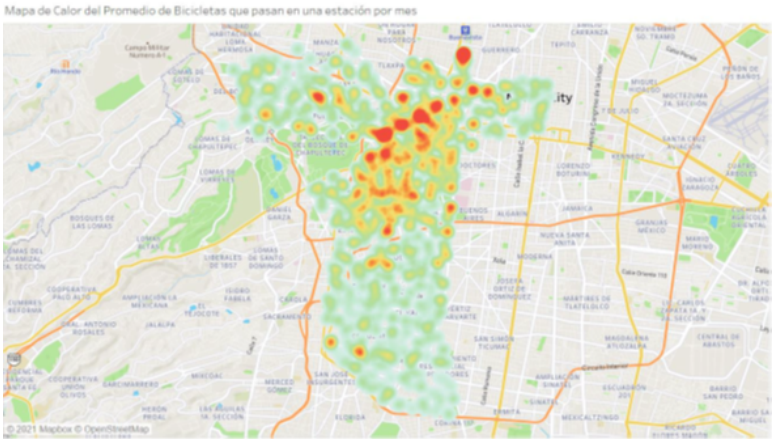



Figura 7.6: Mapa de calor con las transacciones total por estación de ECOBICI en Q1 del 2020


Figura 7.7: Promedio de rotación por estación de ECOBICI en Q1 del 2020

Con estas vistas se explica mejor el movimiento de las estaciones, ya que está sesgado a las zonas ejecutivas de la ciudad. Por otro lado, se observa que el promedio de Rotación es uniforme el la mayoría de las estaciones.

El siguiente paso es usar herramientas geoespaciales para construir el polígono convexo más pequeño que contenga a todas las estaciones de ECOBICI, esto se usa para filtrar las siguientes fuentes de datos.

Por otro lado, la datos que se usaron de UBER, fueron los tiempos de viajes entre zonas de movimiento. Cada registro de la fuentes de datos representa la conexión entre dos zonas, sus tiempo de viajes promedio y la desviación estándar del tiempo de viaje. Las zonas de movimiento que se consideran para toda la Ciudad de México eran más de 5000, las zonas de UBER que tienen al menos una estación de ECOBICI se reduce a 169.

En el resultado del procesamiento de la fuente de datos se obtuvieron cuatro métricas por cada zona de movimiento:

1. Tiempo promedio de un viaje con origen en una zona dada.
2. Tiempo promedio de un viaje con destino en una zona dada.
3. Número de zonas distintas conectadas por viajes con origen en una zona dada.
4. Número de zonas distintas conectadas por viajes con destino en una zona dada.

Adicionalmente, a cada zona de movimiento se le cargaron los datos del entorno urbano. Estos datos buscan medir la accesibilidad para los peatones y bicicletas ya que esto ha demostrado ser de gran impacto para las ventas. (Liu, J. y Shi, W. 2020)(13)

Por último, los datos más importantes y con los cuales se entrenarán los modelos predictivos, los datos del Directorio Estadístico Nacional de Unidades Económicas (DENUE). Primero se tuvo que hacer un refinamiento inicial ya que cómo se está estudiando las ventas basadas en la movilidad de las personas se dejó únicamente aquellas actividades económicas que tuvieran una alta correlación con ella. México usa un sistema numérico para las actividades económicas llamado: *Sistema de Clasificación Industrial de América del Norte 2018.*(10) Los códigos considerados fueron:

- 44: Comercio al por menor
- 72: Servicios de alojamiento temporal y de preparación de alimentos

Así mismo, por cada ubicación se tiene una estimación muy general sobre sus ventas, basada en más indicadores que provee el INEGI sobre el sector y el tipo de actividad. Esta estimación es construida también por la empresa ya mencionada con anterioridad. Debido a que cada tipo de actividad económica genera distintos rangos de ingresos no se puede construir un indicador global. Por lo que se construyó un indicador categórico de potencial de ventas por cada actividad de la siguiente manera:

$$Potencial = \begin{cases} \text{Muy Bajo} & \text{si } v_{i,j} < \mu_i - 2\sigma_i \\ \text{Bajo} & \text{si } \mu_i - 2\sigma_i \leq v_{i,j} < \mu_i - \sigma_i \\ \text{Medio Bajo} & \text{si } \mu_i - 1\sigma_i \leq v_{i,j} < \mu_i \\ \text{Medio Alto} & \text{si } \mu_i \leq v_{i,j} < \mu_i + \sigma_i \\ \text{Alto} & \text{si } \mu_i + \sigma_i \leq v_{i,j} < \mu_i + 2\sigma_i \\ \text{Muy Alto} & \text{si } v_{i,j} \geq \mu_i + 2\sigma_i \end{cases}$$

Donde $v_{ij}$ es el indicador de ventas calculado, $\mu_i$ es la media de ventas por cada actividad económica y $\sigma_i$ es la desviación estándar. Esta es la variable objetivo del modelo, la que intentará predecir. En el caso de la regresión lineal se usó el indicador aproximado.

Por último, se hizo una mapa de las zonas de movimientos de Uber con las ventas estimadas según las empresas del DENUE en cada una:


Figura 7.8: Ventas por zona de movimiento de UBER consideradas en el estudio.

## 7.3. Análisis exploratorio de la base de datos de entrenamiento

En esta sección se tratará el análisis exploratorio a la base de datos de entrenamiento que se el pasará a los diferentes modelos. Este análisis es adicional a los ya presentados anteriormente, que se realizaron en la recopilación y definición de las fuentes de datos.

La primera parte de este análisis consistió en verificar que los números de las tiendas y el potencial de ventas calculado tuviera sentido. Por lo que se realizaron los siguientes dos gráficos:

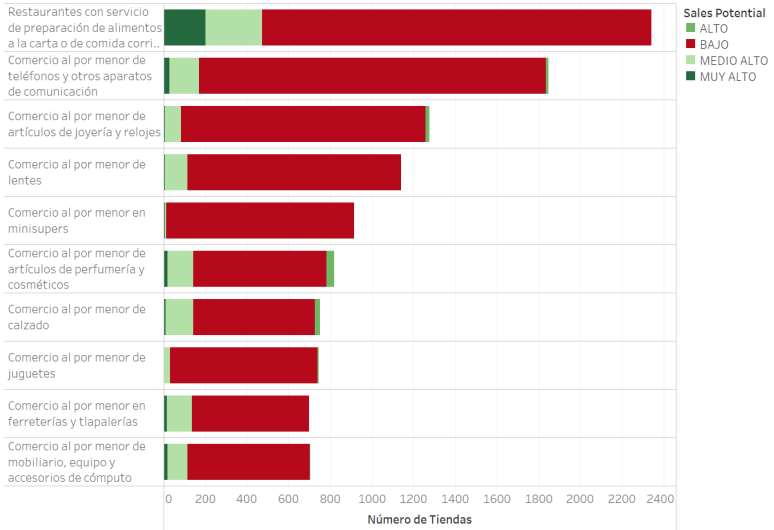



Figura 7.9: Cantidad de tiendas y su potencial de ventas de las 10 actividades económicas con más presencia

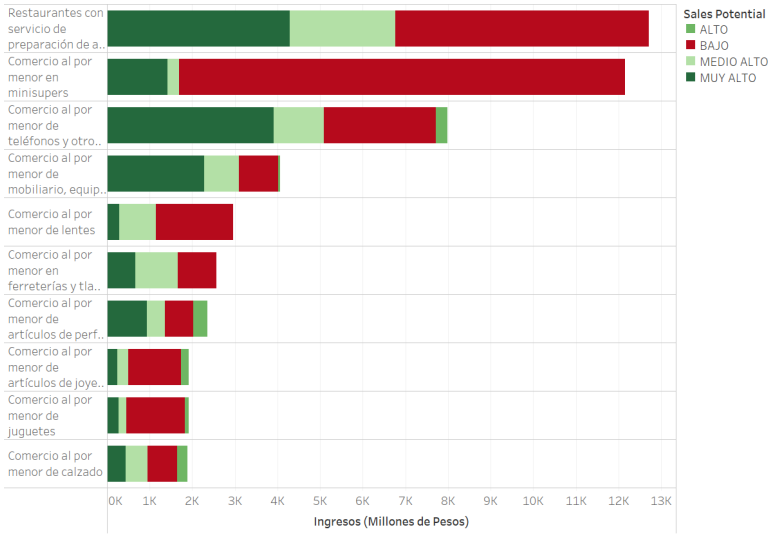



Figura 7.10: Ingresos de las tiendas por su potencial de ventas en las 10 actividades económicas con más presencia

Con estos dos gráficos podemos comprobar que el comportamiento de las tiendas es el esperado. Ya que los muy pocos competidores con potencial alto o mayor se reparten gran parte de los ingresos totales de la actividad económica. Sin embargo, es interesante notar que en las 10 actividades económicas con más presencia de ubicaciones no existen tiendas con potencial *Muy Bajo* o *Medio Bajo*.

Los siguientes análisis se realizaron sobre grupos de variables para averiguar si existía alguna clase de linealidad con los ingresos por tienda o algún fenómeno observable a simple vista que permitiera inferir la importancia de alguna variable sobre el resultado.

## Competidores

Los competidores de cada tienda se calculan en un radio de 500m a la redonda. La siguiente tabla muestra distintas métricas de los competidores según su potencial de ventas:

| **Potencial de venta** | **Máx. competidores** | **Mín. competidores** | **Prom. competidores** |
|---|---|---|---|
| MUY ALTO | 1,212 | 0 | 37 |
| ALTO | 1,087 | 0 | 59 |
| MEDIO ALTO | 2,184 | 0 | 92 |
| MEDIO BAJO | 11 | 0 | 2 |
| BAJO | 2,426 | 0 | 223 |

Tabla 7.1: Métricas de los competidores por potencial de venta

Se puede observar que a excepción del potencial *Medio Bajo*, parece existir ser que el máximo y promedio de competidores es inversamente proporcional al potencial de ventas. Por último, se hace un análisis de linealidad con los ingresos por tienda:

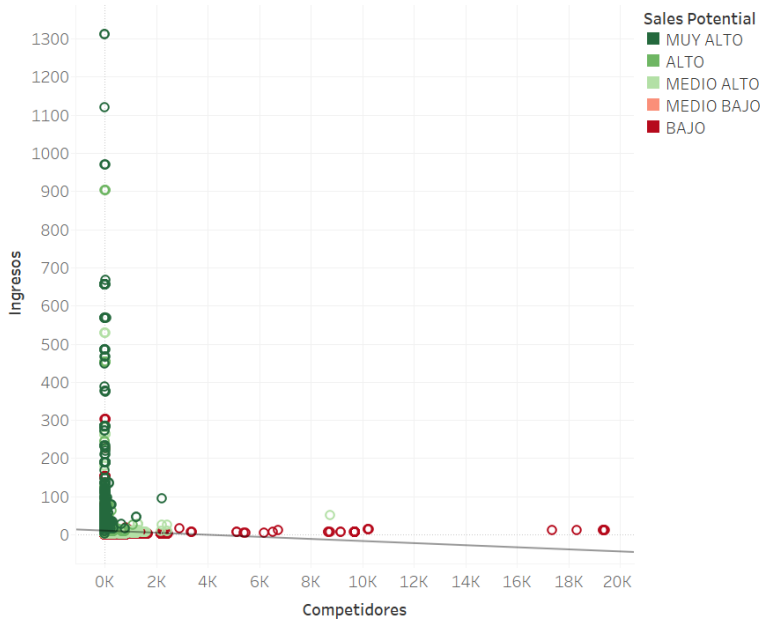



Figura 7.11: Ingresos contra competidores por tienda

Evidentemente no existe una tendencia lineal, ya que el $R^2 = 0.00199$. Esto puede mejorar si se analiza con respecto a cada actividad económica. (Esto se analizará más adelante con el apartado de los modelos de regresiones multinivel).

## Transacciones de Ecobici

Las transacciones de ECOBICI de cada tienda corresponde a la suma de las transacciones de las estaciones en la zona de movimiento de UBER correspondiente. Al igual que en el apartado anterior, se construyen distintas métricas para encontrar alguna relación con el potencial de venta: En este

| **Potencial ventas** | **Máx. transacciones** | **Mín. transacciones** | **Prom. transacciones** | **transacciones** |
|---|---|---|---|---|
| MUY ALTO | 261,815 | 1,724 | 53,509 | 35,797,424 |
| ALTO | 321,771 | 1,724 | 44,318 | 21,183,875 |
| MEDIO ALTO | 321,771 | 1,724 | 48,664 | 127,061,892 |
| MEDIO BAJO | 261,815 | 1,724 | 45,517 | 13,928,053 |
| BAJO | 321,771 | 1,724 | 44,440 | 742,540,858 |

Tabla 7.2: Métricas de las transacciones de ECOBICI por potencial de venta

caso, no se observa que exista ninguna relación entre la cantidad de transacciones y el potencial de ventas.

Además, se hace el análisis de linealidad:

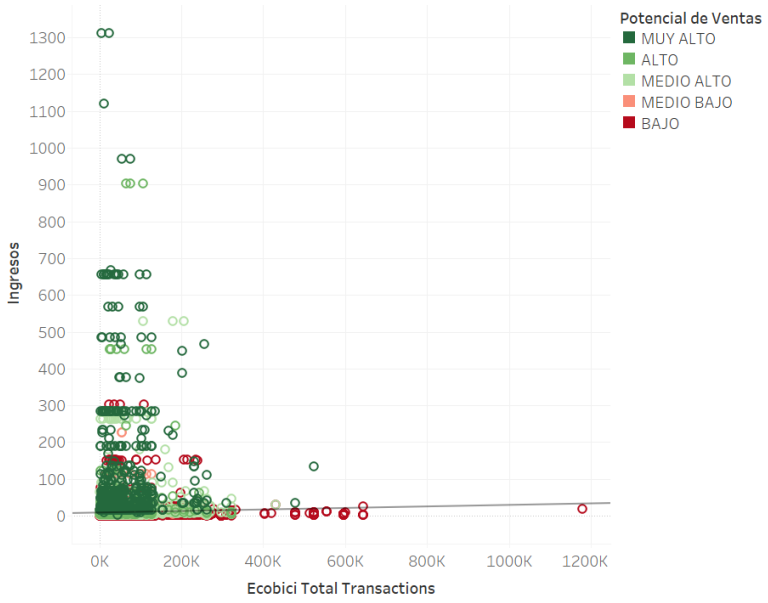



Figura 7.12: Ingresos contra transacciones ECOBICI por tienda

En este caso, el $R^2 = 0.00087$ por lo que no existe ninguna linealidad y por la naturaleza de la variable no tendría sentido un análisis por actividad económica.

## Variables de las zonas de movimiento de UBER

En este grupo de variables no se puede hacer un análisis por tienda ya que todos los indicadores son de un nivel geográfico superior. Para los siguientes análisis, se hace la siguiente aclaración sobre el nombre de las variables:

Dada una zona de movimiento $A$:

- *Connected To*: número de zonas de movimientos conectadas donde la zona $A$ es el destino.
- *Connected With*: número de zonas de movimientos conectadas donde la zona $A$ es el origen.

Las variables de conexiones entre zonas de movimiento no se analizan a mayor profundidad por que la mayoría de las zonas tienen la misma cantidad de conexiones:

| | |
|---|---|
| **Mín. Number Conected To** | 100 |
| **Máx. Number Conected To** | 169 |
| **Prom. Number Conected To** | 165 |
| **Mín. Number Conected With** | 135 |
| **Máx. Number Conected With** | 169 |
| **Prom. Number Conected With** | 167 |

Tabla 7.3: Métricas de conexiones entre zonas de movimiento de UBER

Los siguientes análisis se hacen sobre el tiempo promedio de las conexiones:

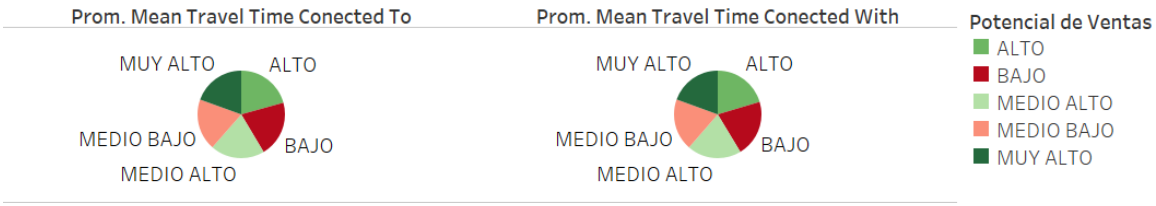



Figura 7.13: Distribución de los tiempos promedios de las conexiones entre zonas de movimientos por potencial de venta

Se puede apreciar que no existe una relación entre los tiempos promedios de conexiones ya que la distribución es casi uniforme. Por último, se hace el análisis de linealidad entre los ingresos por zona de movimiento contra los tiempos promedios de conexiones:


Figura 7.14: Ingresos por zona de movimiento de UBER contra tiempos promedios de conexiones.

Los $R^2_{\text{connected to}} = 0.1113$ y $R^2_{\text{connected with}} = 0.1183$. Son los más grandes por el momento pero sigue sin existir evidencia de linealidad.

## Población flotante

De la misma manera que las variable de UBER, las variables de población flotante son de un nivel geográfico superior, ya que se calculan a nivel zona de movimiento. Por lo que los siguientes análisis son sobre las zonas de movimiento. El primero es de la linealidad de la población total contra los ingresos:

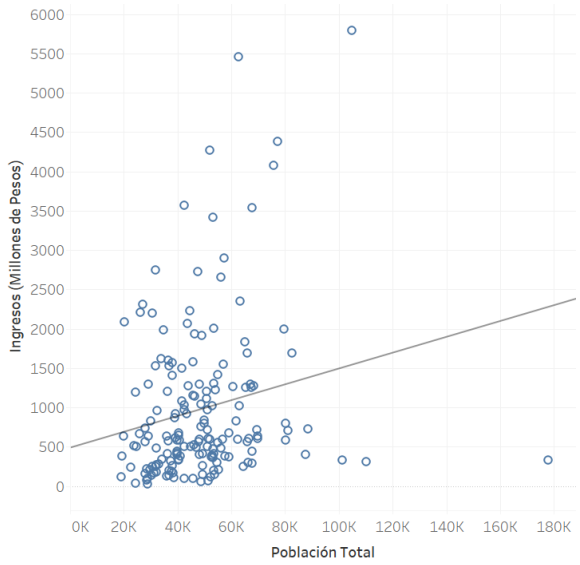



Figura 7.15: Ingresos por zona de movimiento de UBER contra población total

El $R^2 = 0.038$ por lo que no existe evidencia de linealidad.

## Rangos de edad

Se analiza la distribución por edad y potencial de ventas:

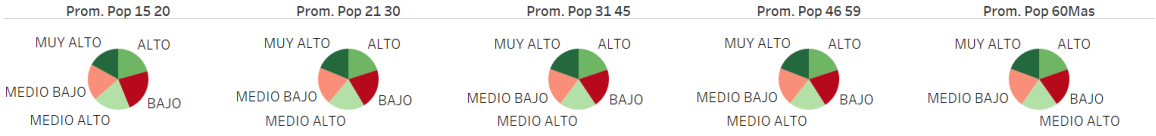



Figura 7.16: Distribución por edad y potencial de ventas

Se puede comprobar que no existe una relación entre las dos variables ya que la distribución se asemeja a una uniforme. Ahora bien, analizamos la linealidad:

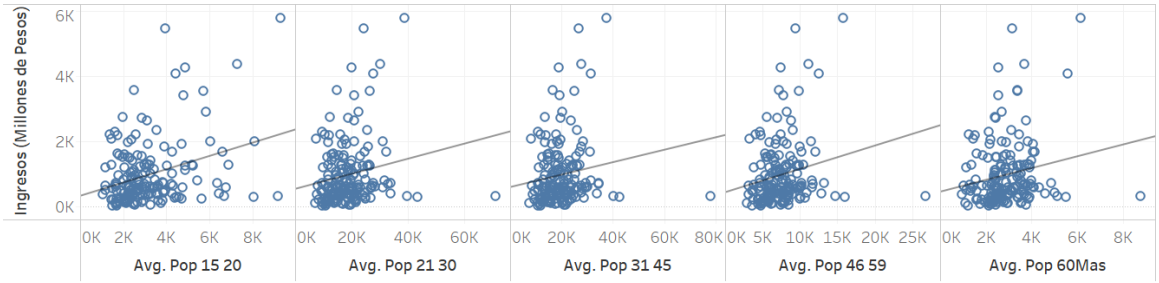



Figura 7.17: Ingresos por zona de movimiento contra rangos de edad

Los $R^2$ son:

- 15-20 años: 0.10138
- 21-30 años: 0.03265
- 31-45 años: 0.02251
- 46-69 años: 0.04108
- 69+ años: 0.03469

No existe evidencia de linealidad.

## Nivel socioeconómico

De la misma manera se analiza la distribución por nivel socioeconómico y potencial de venta:

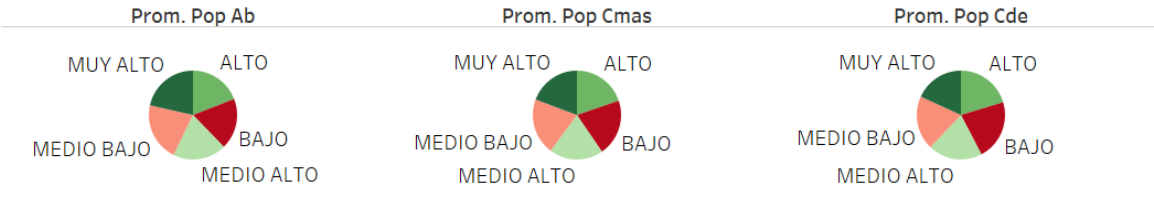



Figura 7.18: Distribución por nivel socioeconómico y potencial de ventas

La distribución es muy semejante a una uniforme por lo que no se encuentra evidencia de una relación entre variables. Ahora, se analiza la linealidad:

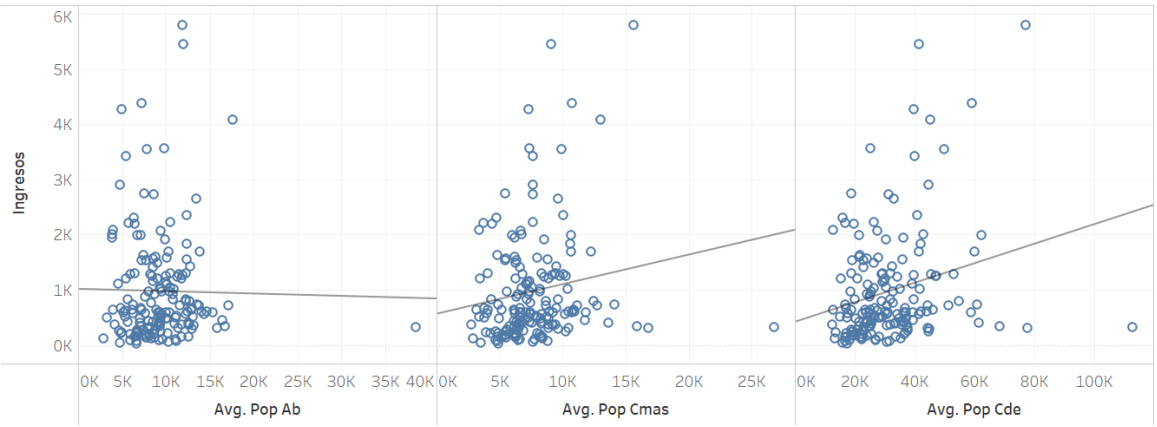



Figura 7.19: Ingresos por zona de movimiento contra el nivel socioeconómico

Los $R^2_{AB} = 0.0002$, $R^2_{C+} = 0.025$ y $R^2_{CDE} = 0.06$. Por lo que no existe linealidad.

## Entorno urbano

Las variables del entorno urbano son calculadas a nivel municipio y se estiman a nivel de zona de movimiento de UBER. Todas las variables son en número de calles, se realizó el análisis de linealidad:

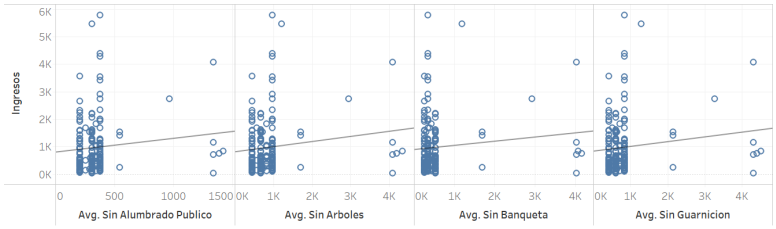



Figura 7.20: Ingresos por zona de movimiento contra entorno urbano (Parte 1)

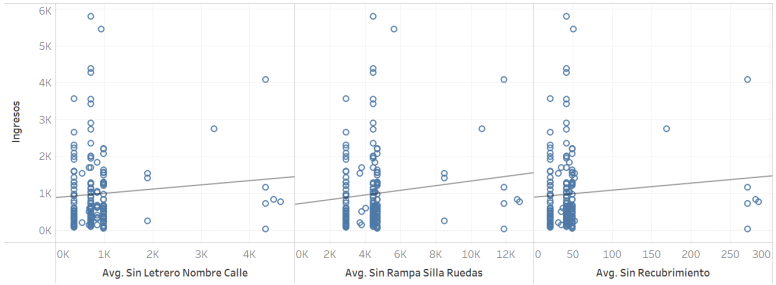



Figura 7.21: Ingresos por zona de movimiento contra entorno urbano (Parte 2)

Los $R^2$ de cada análisis son:

- Sin alumbrado público: 0.0115
- Sin árboles: 0.0168
- Sin banqueta: 0.0123
- Sin guarnición: 0.0171
- Sin letrero con nombre de la calle: 0.0076
- Sin rampa para silla de ruedas: 0.0132
- Sin recubrimiento: 0.0078

No existe evidencia de linealidad.

Ya que en ningún caso se obtuvo evidencia de linealidad, no se usaron métodos de regresión lineal clásicos. Por último, se realiza un análisis de correlaciones entre las variables:

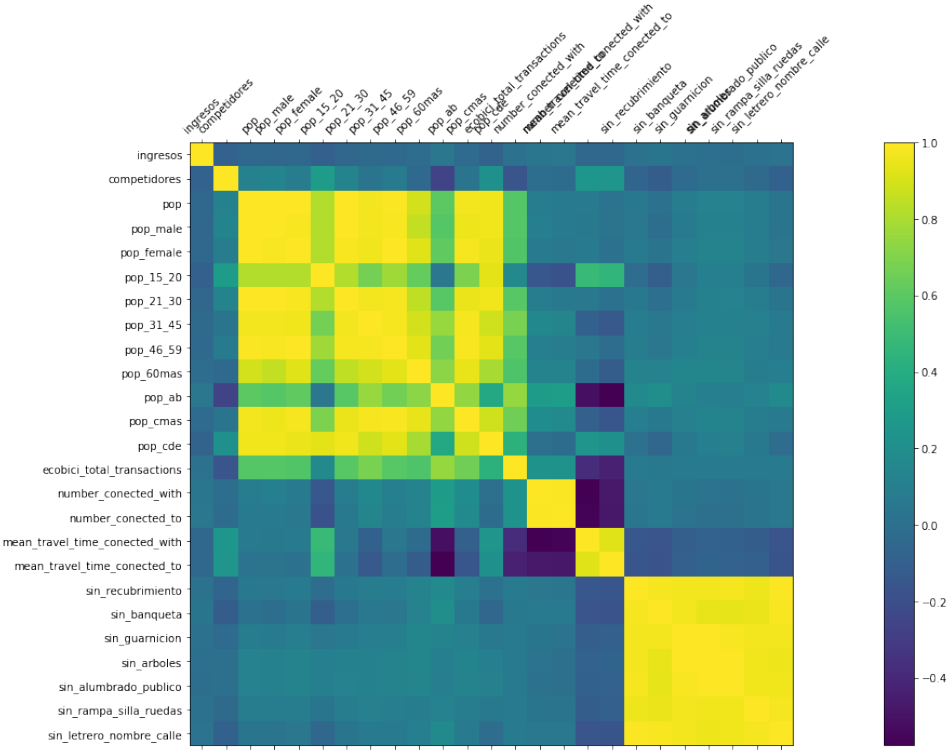



Figura 7.22: Matriz de correlación entre las variables

Se puede notar como los grupos de variables de las mismas fuentes tienen buena correlación. Por otra parte, y a manera de confirmación del análisis previo, la única variable con una correlación fuerte es la de competidores.

## 7.4. Aplicación de los modelos predictivos

Los modelos fueron aplicados a la base de datos de entrenamiento previamente analizada y que tiene la siguiente estructura:

| Fuente de datos | Variable | Tipo | Descripción |
|---|---|---|---|
| Denue | id | string | Id Establecimiento |
| | codigo_act | string | Código SCIAN |
| | nombre_act | string | Nombre SCIAN |
| | nom_estab | string | Nombre Establecimiento |
| | competidores | bigint | # competidores |
| | sales_potential | string | Variable Objetivo |
| | ingresos | double | Variable Objetivo (Regresión) |
| Uber | movement_id | string | Id Zona de Movimiento |
| | number_conected_with | double | # Conexiones |
| | number_conected_to | double | # Conexiones |
| | mean_travel_time_conected_with | double | Tiempo Prom. Conexiones |
| | mean_travel_time_conected_to | double | Tiempo Prom. Conexiones |
| Población flotante | pop | double | Población Total |
| | pop_male | double | Por Género |
| | pop_female | double | Por Género |
| | pop_15_20 | double | Por Rango de Edad |
| | pop_21_30 | double | Por Rango de Edad |
| | pop_31_45 | double | Por Rango de Edad |
| | pop_46_59 | double | Por Rango de Edad |
| | pop_60mas | double | Por Rango de Edad |
| | pop_ab | double | Por NSE |
| | pop_cmas | double | Por NSE |
| | pop_cde | double | Por NSE |
| Ecobici | ecobici_total_transactions | bigint | # Transacciones |
| Entorno | sin_recubrimiento | double | # Calle con carencia |
| | sin_banqueta | double | # Calle con carencia |
| | sin_guarnicion | double | # Calle con carencia |
| | sin_arboles | double | # Calle con carencia |
| | sin_alumbrado_publico | double | # Calle con carencia |
| | sin_rampa_silla_ruedas | double | # Calle con carencia |
| | sin_letrero_nombre_calle | double | # Calle con carencia |

Tabla 7.4: Estructura de la base de datos de entrenamiento

A continuación se listan los modelos usados en este trabajo y cual fue la motivación de su usó:

1. **Redes neuronales artificiales**: este modelo se usó ya que, como se mencionó previamente, se realizó un estudio previo con estos datos y las redes neuronales mostraron una buena efectividad.

2. **XGBoost *(Extreme Gradient Boosting)***: este modelo se utilizó ya que es de los más utilizados en la actualidad en problemas de clasificación con data tabular.

3. **Regresión multinivel**: este modelo se utilizó ya que es el usado en una de las investigaciones consultadas. Estos modelos son utilizados cuando se tienen datos en niveles distintos, como los usados en este trabajo en los niveles superiores al nivel de tienda, como lo son los niveles de zonas de movimiento.

# CAPÍTULO 8

# Experimentos y resultados

En esta sección se listan los múltiples modelos utilizados y los experimentos realizados con cada uno. Además, se discuten los resultados obtenidos con cada uno.

## 8.1. Redes neuronales artificiales

Este fue primer modelo que se aplicó ya que, como se mencionó anteriormente, se utilizó en otro estudio y dio resultados favorables.

Lo primero que se hizo fue cargar la base de datos de entrenamiento que cuenta con 16555 líneas de información. Lo siguiente es estandarizar la data, este proceso consiste en aplicar la siguiente transformación a cada variable numérica:

$$\frac{x_{i,j} - min(X_j)}{max(X_j) - min(X_j)}$$

Donde $x_{i,j}$ es una celda de la base de datos, $i$ es el índice de las filas, $j$ el de las columnas y $X_j$ el vector de todas las celdas de la columna $X$. De esta manera todas las variables numéricas se encuentran en el intervalo $[0, 1]$

El siguiente paso es usar la técnica *One Hot Encoding*, la cual consiste en transformar las posibles categorías en columnas con valores binarios. En este caso la única variable categórica usada fue el código de actividad económica. Luego se dividen los datos en datos de entrenamiento y prueba. Se usó un 80 % en entrenamiento y 20 % en prueba.

El último paso previo a la aplicación de la red es la definición de la estructura. Esta cambió múltiples veces y con la que mejores resultados se obtuvieron fue la siguiente:

| Orden | Tipo | # Nodos | Función de activación |
|---|---|---|---|
| 1 | Capa de entrada | 120 | - |
| 2 | Capa oculta | 2048 | relu |
| 3 | Eliminación de nodos | 20 % | - |
| 4 | Capa oculta | 512 | relu |
| 5 | Eliminación de nodos | 10 % | - |
| 6 | Capa oculta | 256 | relu |
| 7 | Eliminación de nodos | 10 % | - |
| 8 | Capa oculta | 128 | relu |
| 9 | Eliminación de nodos | 10 % | - |
| 10 | Capa de salida | 5 | softmax |

Tabla 8.1: Estructura de la red neuronal con mejores resultados obtenidos.

El principal problema con los datos es que están desbalanceados:

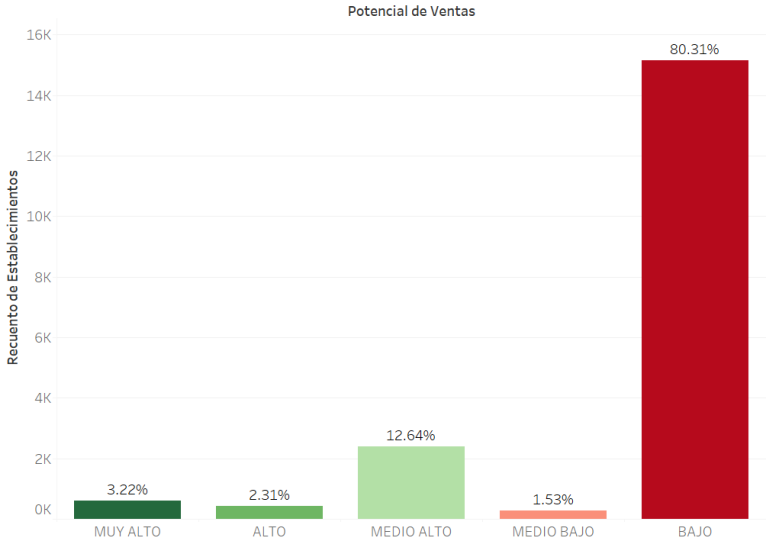



Figura 8.1: Distribución de los datos de entrenamiento

Cómo se puede observar en el gráfico, los datos están muy desbalanceados ya que más de 80 % son establecimientos con potencial *Bajo*, esto causa problemas en el entrenamiento del modelo ya que sesga la manera en la que este aprende.

Todos los experimentos realizados usaron combinaciones de los siguientes parámetros:

| Parámetro | Valores |
|---|---|
| Learning Rate | $1 \times 10^{-3}$, $1 \times 10^{-4}$ |
| Optimizadores | Adam, RMSprop |
| Función de Pérdida | categorical_crossentropy |
| Métricas | CategoricalAccuracy, Accuracy |
| Épocas | 50,100,150 |
| Tamaño del lote | 100,200 |
| Separación para validación | 16 % |

Tabla 8.2: Parámetros usados en la red neuronal

A continuación se listan todas las variaciones del modelo realizadas:

## Aplicación con los datos originales

En este caso, el experimento realizado fue entrenar el modelo con la distribución original de los datos que es desbalanceada.

El tiempo promedio de entrenamiento fue 315 segundos. Obteniendo así una exactitud de 83 % en la fase de entrenamiento y 82 % en la fase de prueba. La matriz de confusión resultante:

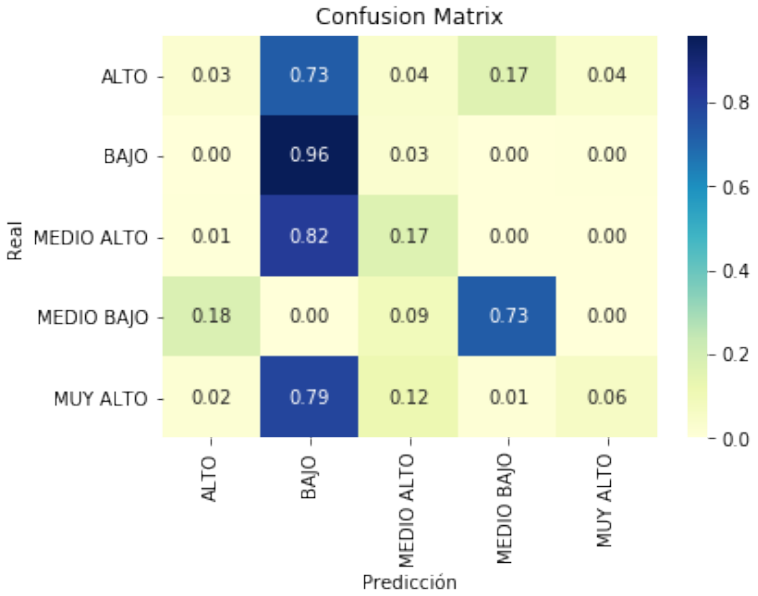



Figura 8.2: Matriz de confusión de la red neuronal con la distribución original

La alta exactitud se debe a la forma en la que se calcula:

$$\frac{\text{Casos correctos}}{\text{Casos totales}}$$

Entonces, debido a que la mayoría de los datos son *BAJOS* y el modelo busca maximizar la exactitud, se sesga a predecir casi todos como *BAJOS* ya que son la mayoría de los casos.

## Aplicación con pesos por clase

Una de las estrategias para mitigar el efecto del desbalanceamiento es asignar pesos cada categoría que compensen en la función de perdida. En este caso los pesos se asignaron con respecto a la categoría: *MUY ALTO* ya que es la de mayor interés. Los pesos usados en este experimento fueron: El tiempo promedio de entrenamiento es de 370 segundos. Mientras que la exactitud de la fase de entrenamiento fue 55 % y en la fase de prueba 53 %. La matriz de confusión resultante:

| Clase | Peso |
|---|---|
| Bajo | 0.040606967 |
| Medio Bajo | 3.247863248 |
| Medio Alto | 0.258152174 |
| Alto | 1.455938697 |
| Muy Alto | 1 |

Tabla 8.3: Pesos por categoría de la red neuronal de la data de entrenamiento

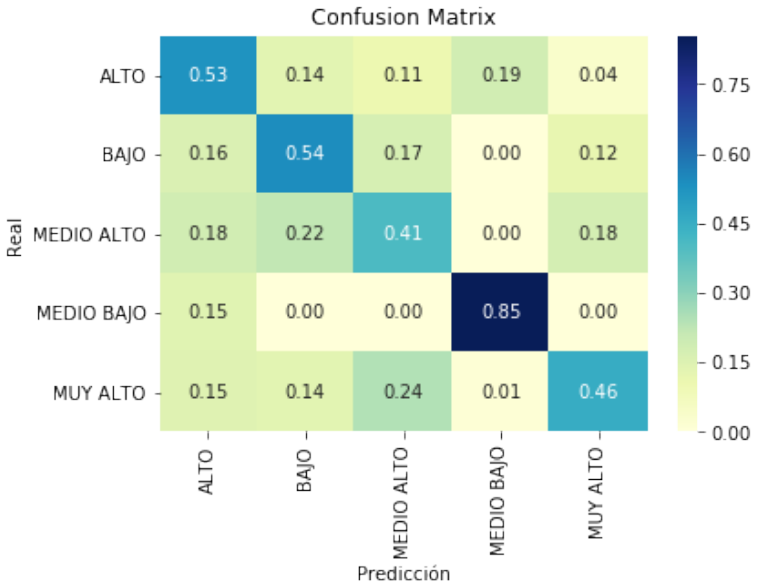



Figura 8.3: Matriz de confusión de la red neuronal con pesos por categoría

En este caso, el modelo tiene una exactitud muy baja porque tiende a estimar un potencial más bajo que el real.

## Aplicación con categorías binarias

Debido a los resultados del experimento anterior, se nota que el modelo no logra diferenciar bien cada categoría. Por lo que se decidió volver el problema en un problema binario. Convirtiendo las categorías a *DEBAJO DE LA MEDIA* y *ENCIMA DE LA MEDIA*.

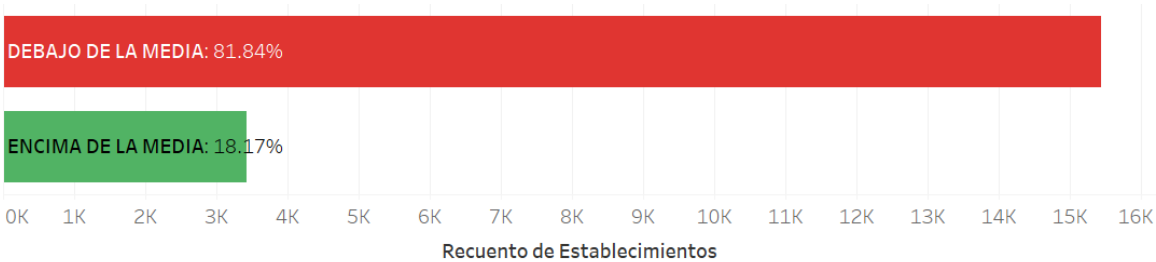



Figura 8.4: Distribución de las categorías binarias en la base de datos de entrenamiento.

El primer experimento se hace con la distribución sin compensar con pesos por clase. El tiempo promedio de entrenamiento fue de 380 segundos. La exactitud en la fase de entrenamiento fue 85 % y en la fase de prueba 81 %. La matriz de confusión resultante fue:

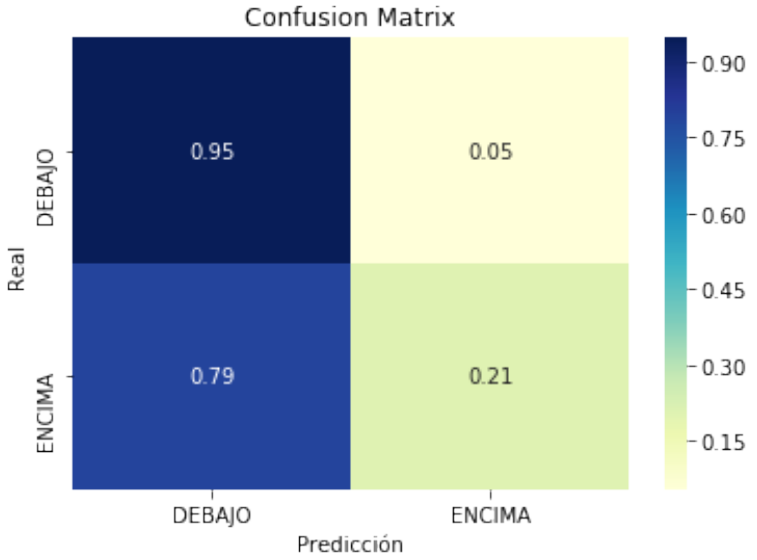



Figura 8.5: Matriz de confusión de la red neuronal con categoría binarias

De la misma manera que en el primer experimento, el modelo se sesga a maximizar la exactitud y predice la mayoría como *DEBAJO DE LA MEDIA*.

En la misma línea de las categorías binarias se hizo el experimento con pesos por cada clase. En este caso se asignó un peso de 0.2233952702702702 a la categoría de *DEBAJO DE LA MEDIA* y peso de 1 a *ENCIMA DE LA MEDIA*.

La exactitud en la fase de entrenamiento fue de 79 % y en la fase de prueba 74 %. La matriz resultante fue:

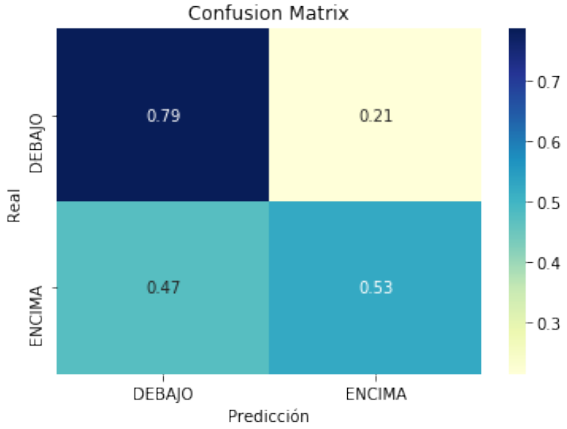



Figura 8.6: Matriz de confusión de la red neuronal con categoría binarias y pesos por clase

### Aplicación con categorías binarias y estructura de la red simplificada.

La última aplicación que se hizo del modelo fue propuesta como una solución al *overfitting* que puede ser causado por muchos nodos o neuronas. En esta aplicación se usa una estructura de la red más simple y con categorías binarias.

| Orden | Tipo | # Nodos | Función de activación |
|---|---|---|---|
| 1 | Capa de entrada | 120 | - |
| 2 | Capa oculta | 50 | relu |
| 3 | Eliminación de nodos | 20 % | - |
| 4 | Capa oculta | 50 | relu |
| 5 | Eliminación de nodos | 10 % | - |
| 6 | Capa oculta | 50 | relu |
| 7 | Eliminación de nodos | 10 % | - |
| 8 | Capa de salida | 2 | softmax |

Tabla 8.4: Estructura simplificada de la red neuronal con buenos resultados.

A este modelo se le aplicaron pesos distintos a cada clase para compensar el desbalanceamiento y con esta estructura se tuvo una exactitud del 67 % en la fase de entrenamiento y 65 % en la de prueba. Además, se obtuvo la siguiente matriz de confusión:

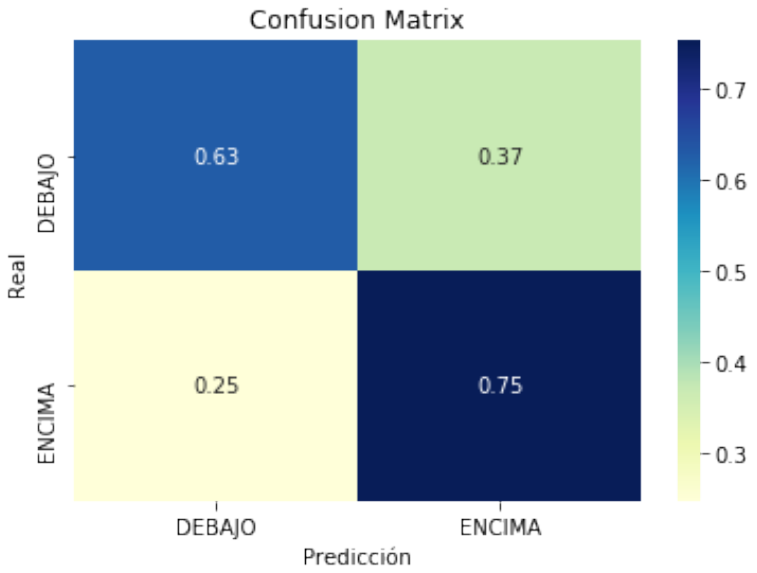



Figura 8.7: Matriz de confusión de la red neuronal con categoría binarias, pesos por clase y estructura simplificada.

Este modelo tiene mejores resultados que los anteriores, por lo que se hace un último análisis para determinar si es conveniente su uso. La matriz de confusión anterior se obtiene de la siguiente tabla:

| | **Predicción: Debajo de la Media** | **Predicción: Encima de la Media** |
|---|---|---|
| **Valor Real: Debajo de la Media** | 2532 | 1499 |
| **Valor Real: Encima de la Media** | 232 | 704 |

Tabla 8.5: Predicciones vs valores reales de la red neuronal con categoría binarias, pesos por clase y estructura simplificada.

Los valores de la matriz de confusión se entienden como: La probabilidad de que un registro sea predicho como $A$ dado que es $B$. Cuando $A = B$ se tienen los valores de la diagonal. Esto es útil para conocer el desempeño en el aprendizaje del modelo, pero se debe calcular de otra manera la probabilidad a la hora de usar las predicciones en la vida real.

Por lo tanto, sea $A_i$ =Una tienda i tiene potencial de venta encima de la media y $B_i$ =El potencial de venta predicho de la tienda i, es Encima de la Media. Entonces, $P(A_i|B_i)$ es la probabilidad de que una tienda tenga potencial de ventas encima de la media, dado que el potencial predicho es encima de la media, y se calcula de la siguiente manera:

$$P(A_i|B_i) = \frac{P(A_i \cap B_i)}{P(B)} = \frac{\frac{704}{4667}}{\frac{2203}{4667}} = \frac{704}{2203} = 0.3195$$

Por lo que si el modelo predice que una nueva tienda tiene potencial encima de la media, solo en el 32 % de los casos es cierto. Ahora bien, el otro caso es cuando $C_i$ =Una tienda i tiene potencial de venta debajo de la media y $D_i$ =El potencial de venta predicho de la tienda i, es debajo de la Media. La probabilidad entonces de que una tienda tenga potencial de ventas debajo de la media, dado que el potencial predicho es debajo de la media es:

$$P(C_i|D_i) = \frac{P(C_i \cap D_i)}{P(D)} = \frac{\frac{2523}{4667}}{\frac{2764}{4667}} = \frac{2523}{2764} = 0.9161$$

En este caso si el modelo dice que una nueva tienda tiene un potencial de ventas debajo de la media, en el 92 % la tienda tendrá un potencial de venta debajo de la media. Si bien el modelo no es perfecto da un resultado muy útil, ya que ayuda a descartar nuevas ubicaciones de tiendas por tener un potencial debajo de la media. Sin embargo, no es concluyente cuando se trata de encontrar una ubicación con potencial encima de la media, para esos casos se necesita una investigación adicional.

Existe una última técnica llamada *Undersampling*, que consiste en tomar una muestra aleatoria de las categoría más grandes con el objetivo de balancear los datos. En este caso no se usó ya que reduciendo la base de datos de entrenamiento no se llega al mínimo de los 10000 datos necesarios.

## 8.2. XGBoost *(Extreme Gradiente Boosting)*

Debido a trabajos realizados previamente y junto a la investigación realizada se decidió utilizar el XGBoost que consiste un modelos simples como los árboles de decisión a los cuales se les optimiza la función de pérdida mediante el descenso gradiente.

En esta implementación se usó como modelo un bosque de árboles de decisión, con la siguiente estructura: Debido a que este modelo se implemento luego de la Red Neural, las estrategias que se

| Parámetro | Valor |
|---|---|
| booster | gbtree |
| max depth | 6 |
| eta | 0.2 |
| objective | multi softmax |
| nthread | 2 |
| number of classes | 2,5 |
| Evaluation metric | Error de clasificación multiclases merror |
| Rondas de entrenamiento | 1000-1500 |

Tabla 8.6: Parámetros usados en el modelo XGBoost

siguieron fueron guiadas con el aprendizaje obtenido previamente. Las estrategias usadas fueron las siguientes:

### Aplicación con la distribución original

La primera alternativa usada del modelo fue la aplicación con la distribución original de los datos. Es decir, con las cinco clases previamente definidas. Usando variaciones de los parámetros anteriormente mencionados, se obtuvo el siguiente resultado del proceso de entrenamiento:

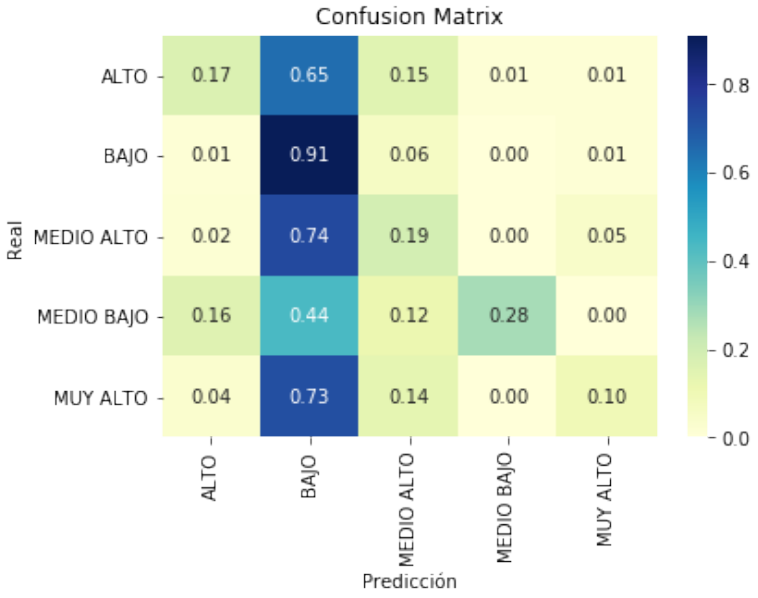



Figura 8.8: Matriz de confusión de la aplicación del Xgboost con la distribución original de clases *undersampling*

Como se puede apreciar en la figura, el modelo tiende a predecir todo como *BAJO* ya que es la clase que maximiza la exactitud. Se probaron distintas métricas de error y se obtuvo resultados iguales.

## Aplicación con clases binarias

Debido a los resultados anteriores, se planteó la alternativa de construir dos clases, cómo en el modelo anterior, con el objetivo de mejorar el rendimiento. Los resultados con las clases binarias, y sin ninguna corrección para el desbalanceamiento de las clases, fue el siguiente:

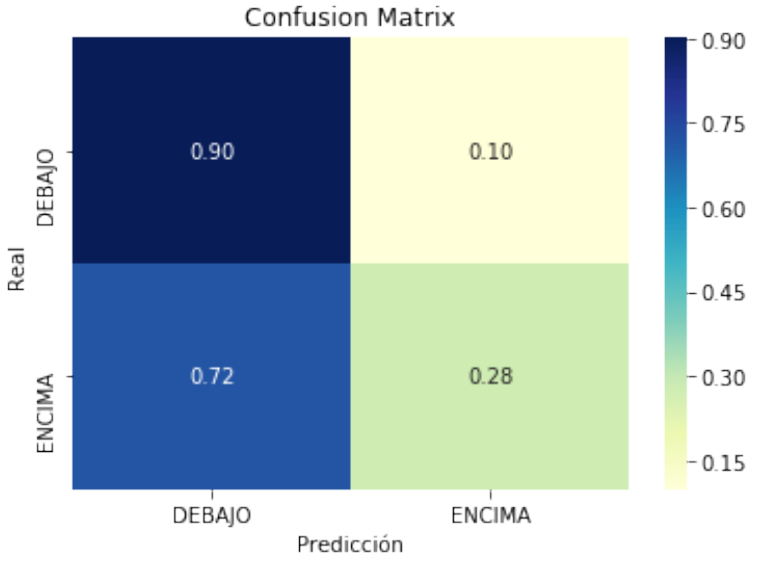



Figura 8.9: Matriz de confusión de la aplicación del Xgboost con clases binarias *undersampling*

Los malos resultados de ambas aplicaciones, se pueden deber en gran parte al desbalance de las clases y también a la poca información de cada ubicación.

## Aplicación con clases binarias y *undersampling*

En este experimento se aplicó la técnica de *Undersampling*, que consistió en tomar una muestra aleatoria de la clase más grande, que en este caso es *DEBAJO DE LA MEDIA*. Esto con el objetivo de balancear las categorías. La estrategia de *undersampling* usada fue la siguiente:

| **Categoría** | **Número de Elementos** | **% del Total** | **Tamaño de la muestra** |
|---|---|---|---|
| DEBAJO DE LA MEDIA | 9530 | 82.24 % | 2058 |
| ENCIMA DE LA MEDIA | 2058 | 17.76 % | 2058 |

Tabla 8.7: Estretegía *undersampling* en el modelo XGBoost

Al aplicar la técnica del *undersampling* se debe ejecutar varias veces el modelo ya que se busca utilizar todo la data de entrenamiento de la clase más grande. La cantidad de veces se calcula:

$$\frac{\text{Número de elementos}}{\text{Tamaño de la muestra}} = \frac{9530}{2058} = 4.63 \approx 5$$

**Aplicación 1:**

Errores:

- Error de evaluación: 2.60 %
- Error de entrenamiento: 41.48 %

Matriz de confusión:

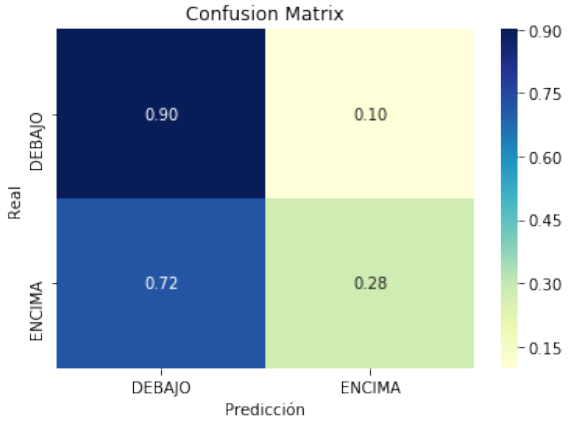



Figura 8.10: Matriz de confusión de la aplicación 1 del Xgboost con clases binarias y *undersampling*

Las aplicaciones subsiguientes tuvieron el mismo resultado, siendo un mal resultado ya que el poder de predicción del modelo es nulo. Por lo tanto se descarta esta alternativa.

Sin embargo, el modelo sirve para entender el funcionamiento de los datos ya que se puede saber cuales variables son las que tienen más peso al dividir los arboles de decisión.

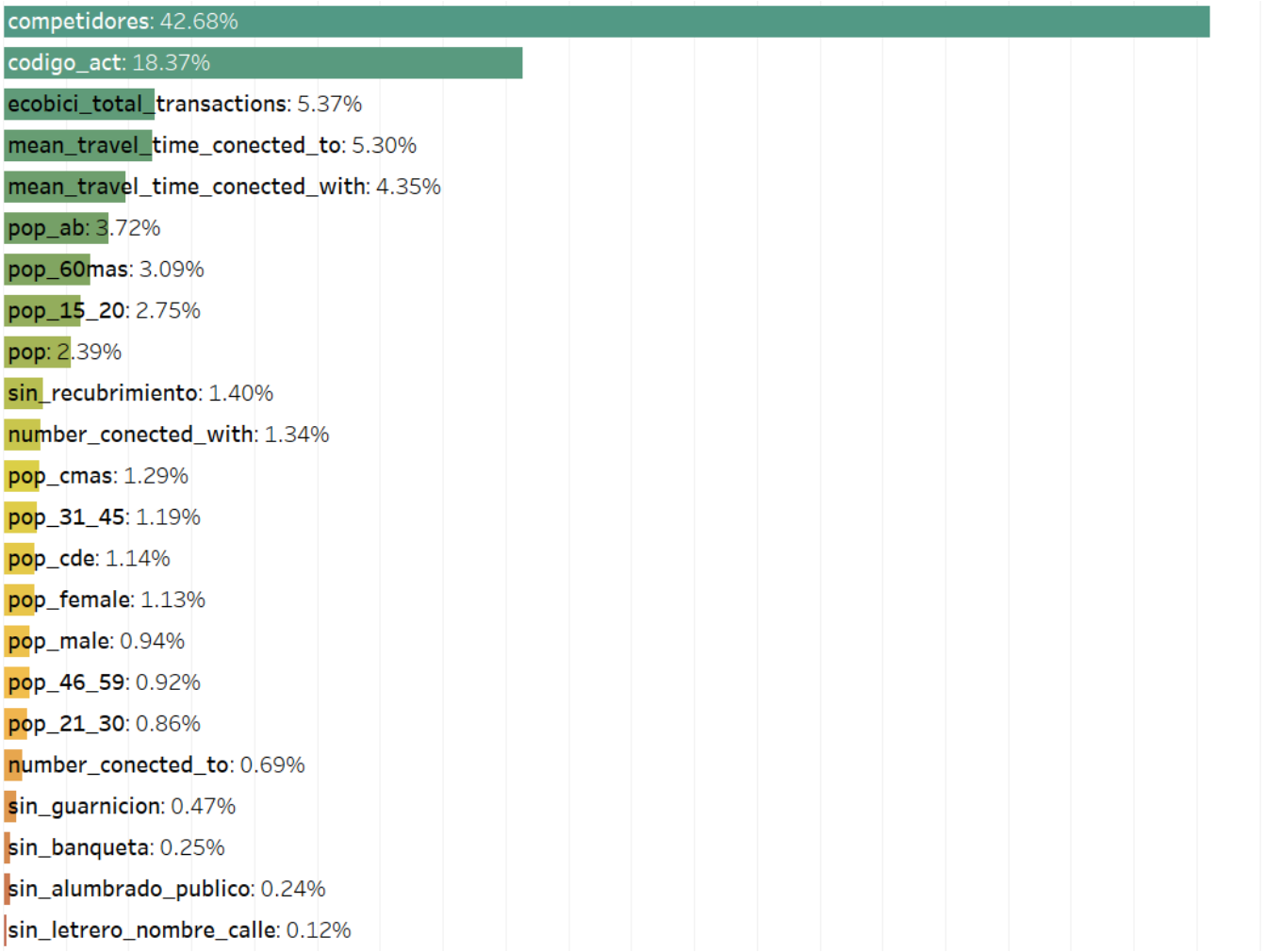



Figura 8.11: Distribución por variables del peso de división de los datos en los árboles de decisión.

Con esto se comprueba el supuesto que se tenía sobre la importancia de las variables de cada tienda, y lleva a la conclusión de que para que el modelo funcione mejor se debe tener más información individual por tienda por sobre la del entorno.

Debido a que el XGBoost es uno de los modelos más populares al usar datos tabulares, los malos resultados aún habiendo aplicado técnicas para corregir el desbalance de las clases, conducen a la hipótesis de que no se tienen los suficientes datos. Ya sea en cantidad de ubicaciones cómo atributos por cada ubicación. Otra posible causa puede ser la variable objetivo, ya que mide el potencial de todas la tiendas en general y por la manera de calcularlo, se pierde mucho conocimiento propio de cada industria al evaluar el potencial de una ubicación.

## 8.3. Regresiones multinivel

Uno de los artículos encontrados en la investigación previa, en el cual buscan implementar un modelo con el mismo objetivo que este trabajo, en Seúl, Corea del Sur (Chang-Deok, K. 2016). El modelo utilizado fueron las regresiones multinivel. Por ello, en este trabajo se usará como alternativa a pesar de los resultados que obtuvo el equipo de Corea.

Las regresiones multinivel funcionan estableciendo niveles jerárquicos, ya que se asume que los elementos varían según los niveles a los que pertenece. Ya que se asume que elementos de un mismo nivel tienden a tener comportamientos parecidos.

En este caso el modelo es de efecto aleatorios, tomando en cuenta que la variable dependiente son los ingresos de cada ubicación (Es importante notar que en este caso estamos usando la variable tal cual viene en la data y no las categorías construidas previamente), y el nivel las zonas de movimiento. Esto surge a partir de la hipótesis de que las ventas se ven influidas en gran parte por el lugar en el que se ubica la tienda.

Antes de ejecutar una regresión multinivel, se debe pasar una prueba para saber si vale la pena un análisis de este tipo, de no pasar la prueba se debe hacer una regresión ordinaria. Se debe calcular el coeficiente de Correlación Intraclases (ICC), este coeficiente puede ser entendido como la proporción de la varianza explicada por los distintos niveles. Es una medida de conocer la semejanza de los individuos de un mismo nivel y de los de otros niveles.

En este caso, se definió como nivel las zonas de movimientos de UBER. El test de ICC para este nivel fue:

$$ICC_{\text{Zona de Movimiento de UBER}} = 6.9\,\%$$

Un valor de ICC que supere 0.05 justifica el uso de la regresión multinivel. Por lo tanto su uso está justificado y el nivel será la zona de movimiento.Se aplicaron dos modelos de regresión multinivel y se hicieron múltiples aplicaciones variando en cada una los predictores más significativos. A continuación se listan las aplicaciones y sus resultados:

### 8.3.1. Regresión multinivel simple

En este caso el modelo aplicado fue una regresión multinivel simple.

#### Aplicación 1:

La primera aplicación se hizo considerando todos los predictores, la estructura fue la siguiente:

| Predictor | Nivel | Descripción del nivel |
|---|---|---|
| codigo_act | 0 | Información propia de la tienda |
| movement_id | 1 | Información de la zona de movimiento |
| competidores | 1 | Información de la zona de movimiento |
| pop_ab | 1 | Información de la zona de movimiento |
| pop_cmas | 1 | Información de la zona de movimiento |
| pop_cde | 1 | Información de la zona de movimiento |
| ecobici_total_transactions | 1 | Información de la zona de movimiento |
| pop_male | 1 | Información de la zona de movimiento |
| pop_female | 1 | Información de la zona de movimiento |
| number_conected_with | 1 | Información de la zona de movimiento |
| number_conected_to | 1 | Información de la zona de movimiento |
| mean_travel_time_conected_with | 1 | Información de la zona de movimiento |
| mean_travel_time_conected_to | 1 | Información de la zona de movimiento |
| sin_recubrimiento | 1 | Información de la zona de movimiento |
| sin_banqueta | 1 | Información de la zona de movimiento |
| sin_guarnicion | 1 | Información de la zona de movimiento |
| sin_arboles | 1 | Información de la zona de movimiento |
| sin_alumbrado_publico | 1 | Información de la zona de movimiento |
| sin_rampa_silla_ruedas | 1 | Información de la zona de movimiento |
| sin_letrero_nombre_calle | 1 | Información de la zona de movimiento |

Tabla 8.8: Estructura de la primera aplicación de la regresión multinivel simple

Con esta aplicación el $R^2$ que se obtuvo fue 0.4559 siendo bastante bajo. Por lo que este resultado no es utilizable, y se confirma viendo la siguiente gráfica:

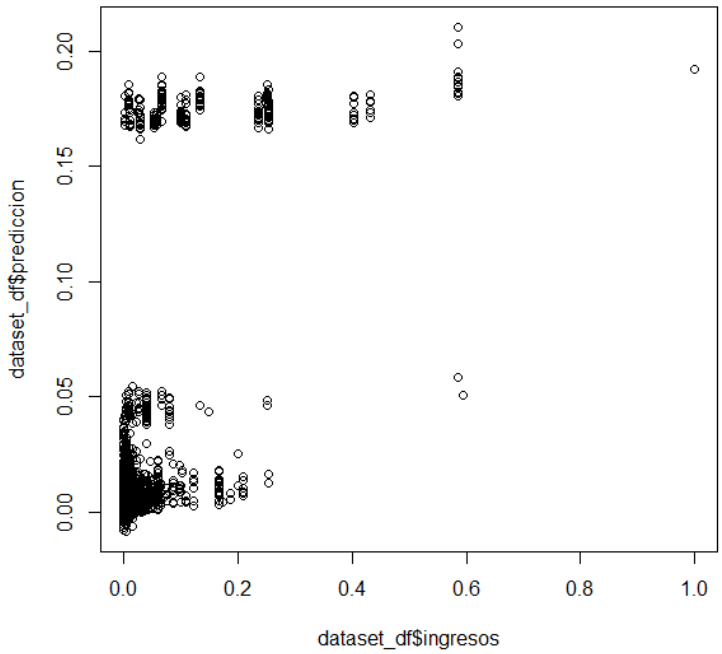



Figura 8.12: Resultados reales vs predicciones en aplicación 1 de la regresión multinivel simple.

Si la regresión fuera perfecta, se debería la función identidad: $f(x) = y$. Los resultados no siguen ese patrón por lo que no es usable esta aplicación.

**Aplicación 2:**

Para esta aplicación se usaron solo los predictores que la aplicación anterior marcó cómo significativos, dejando así la siguiente estructura:

| Predictores | Nivel | Descripción del nivel |
|---|---|---|
| codigo_act | 0 | Información propia de la tienda |
| movement_id | 1 | Información de la zona de movimiento |
| competidores | 1 | Información de la zona de movimiento |
| number_conected_with | 1 | Información de la zona de movimiento |
| number_conected_to | 1 | Información de la zona de movimiento |
| sin_recubrimiento | 1 | Información de la zona de movimiento |
| sin_banqueta | 1 | Información de la zona de movimiento |
| sin_guarnicion | 1 | Información de la zona de movimiento |
| sin_arboles | 1 | Información de la zona de movimiento |
| sin_alumbrado_publico | 1 | Información de la zona de movimiento |
| sin_rampa_silla_ruedas | 1 | Información de la zona de movimiento |
| sin_letrero_nombre_calle | 1 | Información de la zona de movimiento |

Tabla 8.9: Estructura de la segunda aplicación de la regresión multinivel simple

Se puede observar una mejor mínima con respecto a la aplicación anterior ya que el $R^2$ fue de 0.4560825. Sin embargo, sigue siendo un mal modelo ya que al comparar las predicciones con los datos reales se obtiene la siguiente gráfica:

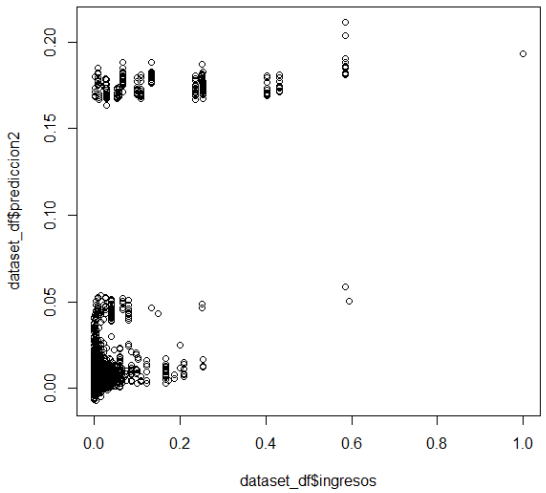



Figura 8.13: Resultados reales vs predicciones en aplicación 2 de la regresión multinivel simple.

**Aplicación 3:**

Al consultar el resumen del modelo de la aplicación anterior, se obtiene que algunos de los predictores dejan de ser significativos. Por lo que se conservan, en esta nueva aplicación, los predictores significativos, dejando así la siguiente estructura:

| Predictores | Nivel | Descripción del nivel |
|---|---|---|
| codigo_act | 0 | Información propia de la tienda |
| movement_id | 1 | Información de la zona de movimiento |
| competidores | 1 | Información de la zona de movimiento |
| number_conected_with | 1 | Información de la zona de movimiento |
| number_conected_to | 1 | Información de la zona de movimiento |
| sin_banqueta | 1 | Información de la zona de movimiento |

Tabla 8.10: Estructura de la tercera aplicación de la regresión multinivel simple

En este caso el $R^2 = 0.4562$ mejoró ligeramente. Sin embargo, sigue sin ser suficiente y la gráfica siguiente lo muestra:

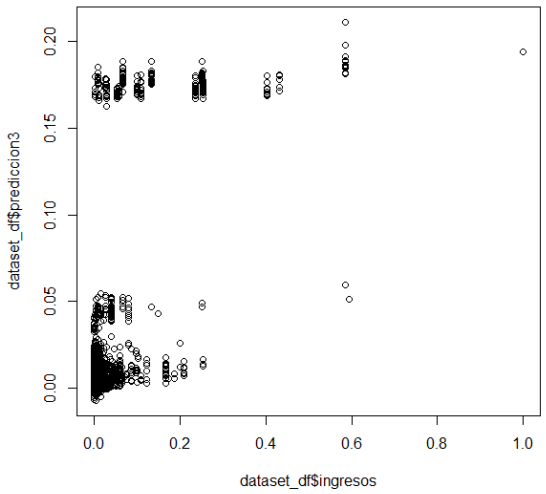



Figura 8.14: Resultados reales vs predicciones en aplicación 3 de la regresión multinivel simple.

### 8.3.2. Regresión multinivel logística

Por último, se usó una regresión multinivel logística. Este modelo se usó para ver si mejoraba los resultados del anterior, por lo que se usó la misma estructura del original.

El modelo tuvo un $R^2 = 0.00618$, por lo que se observa que tiene un resultado mucho peor que el de la regresión multinivel simple. Además, se observa en la siguiente gráfica que son predicciones muy desacertadas:

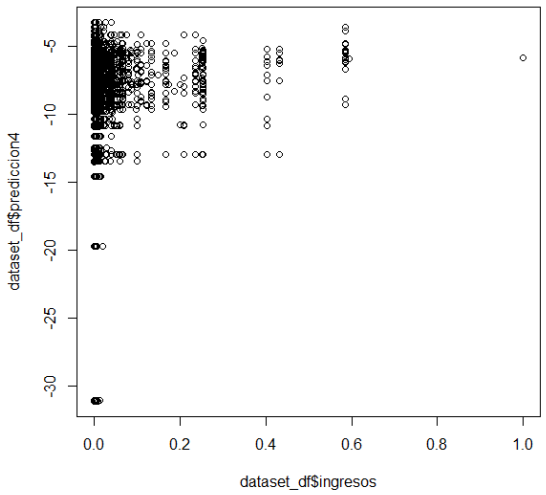



Figura 8.15: Resultados reales vs predicciones en aplicación 1 de la regresión multinivel logística.

Debido a los resultados obtenidos con este modelo se tomó la decisión de descartar su uso. Uno de los posibles fallos fue la incorporación de puntos con ventas demasiado altas. Se recomienda intentar este tipo de análisis con solo un tipo de actividad económica. Otro punto de interés, es la incorporación de más datos del nivel 0 (Información propia de la tienda). Por ejemplo, el área, la movilidad de la tienda, si tiene parqueo, si es parte de una cadena, etc...

Por último, se hizo un descubrimiento interesante y es que considerando el nivel 1 como la actividad económica se tiene un ICC mucho más cercano a 1. Desgraciadamente no se pudo usar este nivel ya que al ser un modelo jerárquico requería información exclusiva de ese nivel,con la cual no se cuenta. Además, el mejor ICC se logró al considerar la actividad económica como el nivel 2. De nuevo, este enfoque no es viable ya que el nivel 2 debe contener (en el sentido conjuntista) al nivel 1.

# CAPÍTULO 9

# Conclusiones y recomendaciones

## 9.1. Conclusiones

El objetivo de esta investigación era encontrar un modelo que fuera capaz de predecir y entender la características que influyen en las ventas de una tienda del sector minorista o un restaurante. Ya que como se vio en el estudio realizado por la Universidad de Portland (Liu, J. y Shi, W. 2020) existe una correlación entre el nivel de afluencia peatonal y las ventas minoristas. Además, de los resultados obtenidos por *Predik Data-Driven* los cuales logran caracterizar el potencial de venta de una tienda usando información de afluencia y entorno. Con estos dos resultados previos, la propuesta de esta investigación era usar fuentes de datos alternativos que permitieran la caracterización de los sectores comerciales basados en sus ventas. Se usaron datos alternativos como posible solución ya que los datos usados en ambos estudios son difícil acceso en México y muy caros en otros casos.

En esta investigación no se logró la caracterización completa, la cual sería poder indicar el volumen de ventas esperado de cada ubicación. Por las investigaciones consultadas previamente se sabe que es posible obtenerla, y las principales variables que influyen son aquellas de población flotante y movilidad específicas del punto. Así como todas las características propias de cada tienda, que en este trabajo no se tenían. Por otro lado, se obtuvieron resultados con las redes neuronales capaces de descartar una ubicación con una predicción del potencial de venta menor a la media. En el caso de las ubicaciones con una predicción del potencial de venta mayor a la media los resultados no son concluyentes, ya que por la probabilidad condicional calculada de las redes, solo el 32 % donde el modelo predice que una tienda tiene potencial por encima de la media es cierto. Debido a esto no se puede usar el modelo para encontrar un buen punto, solo para descartar un mal punto.

Además, se probaron múltiples variantes de cada modelo, alternando parámetros y llegando a los mismos resultados. Por lo que se concluye que son los datos los que no permiten la completa caracterización de los sectores comerciales.

En el análisis exploratorio de los datos se encontró que ninguna variable por si sola tiene una correlación directa con las ventas y por lo tanto el poder de predicción se encuentra en la mezcla de variables de la tienda y el entorno. Además, cómo se pudo ver en la aplicación de las redes neuronales, uno de los principales problemas fue el desbalanceamiento de los datos y lo escasos casos de tiendas con potencial mayor a la media, haciendo difícil la tarea de los modelos para aprender que diferenciaba esos casos, que eran los más interesantes.

Por otro lado, la aplicación del *Xgboost* reveló que las variables de más peso en la decisión son aquellas propias de cada tienda, que en este estudio son los competidores a 500 metros y el código de la actividad económica las que abarcan el 60 % del peso de la decisión, por lo que se llegó a la conclusión de que es necesario incorporar más datos a nivel tienda. Esto se ve reforzado con el hecho de que en las investigaciones consultadas las que más peso en la decisión tenían eran las de movilidad y población en un entorno cercano a la tienda.

Por último, al aplicar las Regresiones Multinivel, aún sabiendo que no funcionaron en estudios anteriores, se obtuvo resultados interesantes. El más significativo fue que la actividad económica dividía de mejor manera los resultados que las zonas de movimiento de UBER. Es decir en el análisis de correlación intraclase se vio que la actividad económica causa un mayor efecto en la variable dependiente de la venta. Esta no se pudo usar por la naturaleza del modelo, usando las zonas de movimiento los valores de las variables que pertenecen al nivel son un subconjunto de la misma. Por ejemplo, si una tienda tiene el nivel de la zona de movimiento A las variables pertenecientes a ese nivel están todas calculadas sobre la zona de movimiento A y no sobre ninguna otra. En otras palabras las zonas de movimiento son una partición del conjunto de los datos, cosa que no pasa con la actividad económica, ya que si una tienda tiene actividad n, las variables de población y movilidad no se calculan exclusivamente para esa actividad económica. Además, no se puede definir los dos niveles para la regresión porque la naturaleza de los niveles debe ser jerárquica. Por ejemplo, si se usa zona de Movimiento, el otro nivel sería municipio, y el siguiente sería el estado.

Esto conduce a la conclusión de que estos análisis se deben hacer de manera independiente para cada actividad económica, esto se ve justificado con la dificultad de la regresión para funcionar ya que había actividades económicas con ventas mucho mayores que otras.

La última conclusión es que este tipo de análisis es posible y funciona de la manera correcta con la incorporación de más datos. Esto se ve justificado tanto por los estudios de la investigación previa, así cómo por los resultados y posibles mejoras de los modelos.

## 9.2. Recomendaciones

Según lo establecido en la sección anterior y los trabajos consultados en la investigación previa, la comprensión y estimación del potencial de venta es posible mediante modelos de *Machine Learning*. Sin embargo, para que la propuesta de esta investigación funcione se tienen una serie de recomendaciones si se quiere replicar el estudio.

1. **Más información por cada punto de venta**: como se reflejó en los resultados los modelos necesitan más información propia de cada tienda para poder discernir de mejor manera. Idealmente, se recomienda el uso de datos de celulares ya que son un insumo muy bueno al medir la afluencia peatonal real.

2. **Más información del entorno**: como se vio en la investigación previa, el uso de más datos de entorno, tal como perfiles socio demográfico de la audiencia de cada tienda, y estadísticas a nivel manzana o municipio, son de gran ayuda para el correcto funcionamiento del modelo.

3. **Datos reales de venta**: en esta investigación se observó el impacto de las estimaciones de venta de cada tienda y la definición del potencial basado en la media y varianza de la actividad económica. En los estudios que han tenido éxito con la caracterización se usan datos reales de venta y la definición de exitoso o no de una tienda, ya que cada actividad económica tiene aspectos distintos a considerar al medir el éxito de una ubicación.

4. **Estudios independientes por actividad económica**: en uno de los experimentos se comprobó que los modelos funcionarían mejor si se pudieran separar por actividad económica ya que se estaba intentando aprender de un conjunto de elementos que no son del todo comparables.

5. **Más ubicaciones para analizar**: en esta investigación el número de tiendas que se analizó estuvo limitado por la fuentes de datos disponibles. Sin embargo, la muestra se quedó corta para los modelos, los cuales aprenden mejor con más datos para analizar.

CAPÍTULO 10

# Bibliografía

A continuación, se encuentra la lista completa de referencias usadas en este trabajo.

## Bibliografía


[1] *Ecobici*. https://www.ecobici.cdmx.gob.mx/.

[2] *Uber Movement*. https://movement.uber.com/?lang=en-US.

[3] *Identificación y caracterización de puntos de venta en Canal Tradicional*, Dec 2021. https://predikdata.com/analisis-de-movilidad/.

[4] Aggarwal, Charu C.: *Neural Networks and Deep Learning*. Springer, Cham, 2018.

[5] Bishop, Christopher M.: *Pattern recognition and machine learning*. Springer New York, 2016.

[6] Chen, Tianqi y Carlos Guestrin: *XGBoost: A Scalable Tree Boosting System*. KDD '16, página 785–794, New York, NY, USA, 2016. Association for Computing Machinery, ISBN 9781450342322. https://doi.org/10.1145/2939672.2939785.

[7] De La Cruz, Francisco: *Modelos multinivel*. Revista Peruana de Epidemiología, 12, Dec 0AD.

[8] Geografía (INEGI), Instituto Nacional de Estadística y: *Características de las localidades y del entorno urbano 2014*. https://www.inegi.org.mx/programas/cleu/2014/.

[9] Geografía (INEGI), Instituto Nacional de Estadística y: *Directorio Estadístico nacional de unidades económicas. DENUE*. https://www.inegi.org.mx/app/mapa/denue/default.aspx.

[10] Geografía (INEGI), Instituto Nacional de Estadística y: *Sistema de Clasificación Industrial de América del Norte 2018 (SCIAN 2018)*. https://www.inegi.org.mx/app/scian/.

[11] Janocha, Katarzyna y Wojciech Marian Czarnecki: *On loss functions for deep neural networks in classification*. Schedae Informaticae, 25:49–59, 2017.

[12] Kang, Chang Deok: *Spatial access to pedestrians and retail sales in Seoul, Korea*. Habitat International, 57:110–120, 2016, ISSN 0197-3975. https://www.sciencedirect.com/science/article/pii/S0197397516303113.

[13] Liu, Jenny H. y Wei Shi: *Understanding Economic and Business Impacts of Street Improvements for Bicycle and Pedestrian Mobility: a Multi-city Multi-approach Exploration.* Portland, OR: Transportation Research and Education Center (TREC), 2020.

[14] Pública, Agencia Digital de Innovación: *Ubicación de líneas y estaciones del Sistema de Transporte Colectivo Metro.* `https://datos.cdmx.gob.mx/dataset/lineas-y-estaciones-del-metro/resource/0869e0dd-6876-4446-a199-8f670a359c00`.

[15] Siddharth Sharma, Simone Sharma, Anidhya Athaiya: *Activation functions in neural networks.* International Journal of Engineering Applied Sciences and Technology, 4:310–316, 2020.

[16] Spyratos, Spyridon, Michele Vespe, Fabrizio Natale, Ingmar Weber, Emilio Zagheni y Marzia Rango: *Quantifying international human mobility patterns using Facebook Network Data.* PLOS ONE, 14(10), 2019.

[17] Vieira, Renato S. y Eduardo A. Haddad: *A weighted travel time index based on data from Uber Movement.* EPJ Data Science, 9(1), 2020.

CAPÍTULO 11

# Anexos

## 11.1. Visor de potencial por actividad económica

Como parte de los objetivos de la investigación se buscaba crear una herramienta que permitiera consultar el potencial de venta de una tienda, según su actividad económica y su zona de movimiento. Esto se realizó usando los resultados de la red neuronal, sabiendo que solo es capaz de descartar lugares por mal potencial y los que son marcados con potencial: *ENCIMA DE LA MEDIA* requieren una mayor investigación.

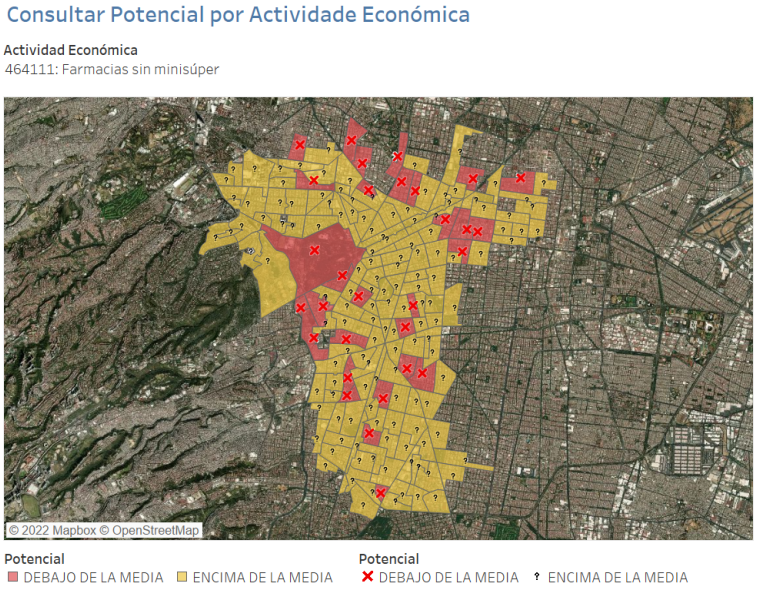



Figura 11.1: Ejemplo del visor de potencial de venta desarrollado con los resultados de la red neuronal