Modelamiento espacial de la oferta y la demanda en el territorio colombiano a partir de analítica de datos financieros

Trabajo final de grado

Maestría en Ingeniería – Analítica
Profundización

Juan José Ospina Carvajal
jujospinaca@unal.edu.co

Director:
Juan David Ospina Arango
jdospina@unal.edu.co

Universidad Nacional de Colombia – Sede Medellín
Facultad de Minas
Departamento de ciencias de la computación y la decisión
Medellín, Colombia
2021

Resumen

Este trabajo final de maestría propone una herramienta de inteligencia de negocios que permite modelar la oferta y la demanda para una zona geográfica definida a través de la analítica de datos financieros. Se hace una revisión de la literatura sobre sistemas de geolocalización, geomarketing, la información transaccional en el ámbito colombiano y el uso que las entidades financieras le dan a los datos. También plantea una técnica de agrupación geográfica que permite conservar la anonimidad y la privacidad en los datos, además de un modelo de datos y un prototipo funcional de la herramienta. Por último, se exponen tres casos de uso en los que la herramienta aporta información valiosa para el proceso de toma de decisiones dentro de la entidad o el asesoramiento experto a un cliente como estrategia de retención y profundización de la relación cliente – entidad financiera.

Palabras Claves

Geolocalización. Modelamiento de oferta y demanda. Herramienta de datos. Información transaccional. Big data. Inteligencia de negocios.

Title

Spatial modeling of supply and demand in the Colombian territory from financial data analytics.

Abstract

This master's degree final project proposes a business intelligence tool that allows supply and demand modeling within a specific geographic location through financial data analytics. A literature review is made about geolocation systems, geomarketing, transactional data in the Colombian landscape and the use of data in banking sector around the world. This work presents a geographic cluster technique that preserves anonymity and privacy in data, in addition to a data model and a project functional prototype. Finally, three use cases are shown where the prototype provides valuable information in the decision-making process inside the financial entity or in the expert advice to a client as a strategy of customer retention and conservation of the client – bank relationship.

Keywords

Geolocation. Supply and demand modeling. Data tool. Transactional information. Big Data. Business intelligence.

Contenido

Resumen 2

Palabras Claves 2

Abstract 3

Keywords 3

Introducción 9

1. Justificación 10

2. Objetivos 11

2.1. Objetivo general 11

2.2. Objetivos específicos 11

3. Estado del arte 12

3.1. Sistemas de geolocalización 12

3.2. Geomarketing 14

3.3. La información transaccional en Colombia 18

3.4. El uso de datos financieros en el sector bancario 19

3.4.1. Commerce 360, Grupo BBVA 22

3.4.2. Plink, Grupo Bancolombia 23

4. Metodología propuesta 25

4.1. Modelo de datos 25

4.1.1. Clientes 27

4.1.2. Establecimientos 27

4.1.3. Transacciones 28

4.1.4. Diagrama de clases del modelo 28

4.2. Procesamiento de datos 29

4.2.1. Preprocesamiento, protección de datos personales y categorización de variables..... 29

4.2.2. Agrupación geográfica y creación de supertipos 30

4.2.3. Modelado de la oferta en el territorio colombiano 32

4.2.4. Modelado de la demanda en el territorio colombiano..... 33

4.3. Implementación en la entidad financiera 34

4.3.1. Contextualización y actualización sobre *Big Data*..... 34

4.3.2. Entendimiento de los casos de uso y de negocio 35

4.3.3. Implementación de prototipos..... 35

5. Resultados esperados 35

6. Prototipo de herramienta para analizar geográficamente la oferta y la demanda..... 36

6.1. Preprocesamiento de la información 36

6.1.1. Información demográfica de la demanda..... 36

6.1.2. Información general de la demanda 37

6.1.3. Información por categorías de la demanda..... 37

6.1.4. Información demográfica de la oferta 38

6.1.5. Información general de la oferta 39

6.1.6. Información por categorías de la oferta 39

6.2. Desarrollo del prototipo utilizando el lenguaje R 40

6.2.1. Interfaz de usuario..... 40

6.2.2. Servidor y funcionamiento interno 44

7. Casos de uso del prototipo 48

7.1. Identificación de mercados naturales 49

7.2. Asesoría a clientes para la toma de decisiones 51

7.3. Aporte de información para la toma de decisiones frente a grandes proyectos..... 53

8. Consideraciones éticas 57

Referencias..... 59

Lista de Figuras

Figura 1. Mapa de dos rutas de buses con sus cronogramas en el primer y último recorrido. 12

Figura 2. Ubicación estimada del objetivo mediante técnicas basadas en la dirección IP. 13

Figura 3. Ejemplo de comparación entre ubicación real versus ubicación aproximada por el código postal. 15

Figura 4 Clientes potenciales en un radio de 3 Km de una nueva tienda 16

Figura 5 Mapa de curvas isocronas de los centros comerciales. 17

Figura 6. Ejes de trabajo del Proyecto F. 19

Figura 7. Necesidades y técnicas de minería de datos aplicadas en el sector bancario. 20

Figura 8. Transformación de los datos a Inteligencia de Negocios. 21

Figura 9. Ejemplo de indicadores de Commerce 360. 22

Figura 10. Caso de uso de Plink en un escenario real. 24

Figura 11. Métricas que ofrece uno de los módulos de Plink. 25

Figura 12. Actores participantes en el proceso de pago de un establecimiento de comercio. 26

Figura 13 Diagrama de clases UML del modelo de datos. 29

Figura 14. Ejemplo de agrupación geográfica.. 31

Figura 15. Menú del prototipo.. 41

Figura 16. Mapa del prototipo.. 42

Figura 17. Indicadores generales del prototipo.. 43

Figura 18. Indicadores por categorías del prototipo. 43

Figura 19. Interfaz de usuario del prototipo.. 44

Figura 20. Proceso de filtrado de información de la zona de interés. 45

Figura 21. Estructura del prototipo cargada en el servidor web. 47

Figura 22. Carga del prototipo en el servidor GitLab de la entidad financiera.. 48

Figura 23. Modelado de la demanda. Caso de uso 1. 49

Figura 24. Modelado de la oferta. Caso de uso 1. 50

Lista de Tablas

Tabla 1. Ejemplo de información del supertipo Clientes Información Demográfica generado a partir de una muestra de los datos. 31

Tabla 2. Ejemplo de información del supertipo Clientes Información Transaccional generado a partir de una muestra de los datos. 32

Tabla 3. Ejemplo de información del supertipo Clientes Información Transaccional por Categorías generado a partir de una muestra de los datos. 32

Tabla 4. Abreviatura de indicadores.. 47

Tabla 5. Resultados del modelado de oferta y demanda para las zonas preseleccionadas. 52

Tabla 6. Demanda y oferta por categorías en zona Guayabal.. 52

Tabla 7. Demanda y oferta por categorías en zona Buenos Aires. 52

Tabla 8. Detalles del megaproyecto presentado por el cliente constructor... 54

Tabla 9. Análisis de la demanda de la zona de interés a diferentes radios. . 55

Tabla 10. Análisis de la oferta de la zona de interés a diferentes radios.. 56

Introducción

En los últimos años, las transacciones por bienes y servicios a través de medios electrónicos han crecido debido a la adopción de nuevas tecnologías e iniciativas de parte del gobierno y las instituciones financieras en el país. Este crecimiento ha generado un aumento en la cantidad de datos recopilados, dando paso a la analítica para aprovechar esta información y transformarla en beneficios para los prestadores de servicios financieros y para sus consumidores.

La analítica y el aprovechamiento de los datos, conocidos como *Big Data*, ha revolucionado todas las industrias, siendo la financiera una de las más beneficiadas debido a que, por la naturaleza de su negocio, ya cuentan con una gran cantidad de información almacenada y casi lista para su uso. La revolución tecnológica también ha permitido el desarrollo de nuevos conceptos como lo son la geolocalización, situar personas u objetos en el espacio, y el geomarketing, utilizar información geolocalizada para realizar análisis de mercado sobre una zona geográfica específica.

En un contexto económico inestable causado por la pandemia generada por el COVID-19 y las caídas en los mercados por la ruptura de los acuerdos entre Rusia y los miembros de la OPEP, las organizaciones han acelerado la adopción de tendencias como sostenibilidad, computación en la nube y toma de decisiones apalancada en los datos (*Data driven decision making*). En los últimos años, ha resurgido la analítica de negocios o *Business Analytics* con el objetivo de aportar conocimiento desde los datos y así mejorar la efectividad en la toma de decisiones.

En este trabajo se plantea una metodología para aprovechar la información de la transaccionalidad entre los clientes y establecimientos de comercio de bienes o servicios de una entidad financiera para entender la oferta y la demanda sobre una zona geográfica. Para este modelo se crearán supertipos de información demográfica, transaccional y transaccional por categorías que serán geolocalizados y serán analizados si pertenecen a la ubicación de interés.

Se espera poder visualizar cambios entre la oferta y demanda de dos ubicaciones y poder comparar estos resultados con otros modelos o la literatura. También, se espera desarrollar una herramienta en un lenguaje de programación abierto que permita acceder a este modelo para realizar análisis de oferta y demanda.

Este modelo se realiza con información anónima suministrada por una entidad financiera, con el objetivo de preservar la privacidad de los clientes de dicha

entidad y la reserva bancaria. La herramienta desarrollada será propiedad intelectual de la entidad financiera aportante de los datos.

1. Justificación

Gracias al desarrollo de nuevas tecnologías, nuestra cotidianidad se ha digitalizado llevando a dejar un amplio rastro digital a través de: compras con tarjetas débito, crédito o código QR, búsquedas de productos en tiendas online, el monitoreo de ritmo cardiaco y conteo de pasos en el día con el GPS de un *smartphone*, entre muchas otras. Todo este rastro digital genera grandes volúmenes de datos, comúnmente referido como *Big Data*, que se han convertido en activos fundamentales para todas las industrias, dando paso a la analítica, el aprovechamiento de los datos, para modificar la forma en la que las empresas operan y toman decisiones [1].

La industria de los servicios financieros no ha sido ajena al impacto del Big Data y ha adoptado diversas técnicas analíticas buscando el beneficio tanto para sus clientes como para sí mismos. Las técnicas incluyen un amplio espectro que van desde análisis de patrones de gasto y segmentación de sus clientes hasta análisis de sentimientos y gestión de riesgos [2].

Es entonces cuando la analítica de negocios o *Business Analytics* (BA), que se refiere al uso intensivo de los datos, análisis estadísticos y cuantitativos, modelos predictivos y explicativos, y manejo de los procesos para la toma asertiva de decisiones, toma relevancia. Este concepto fue inicialmente desarrollado en la década de los 50s y recientemente se ha convertido en un tema de gran importancia para industrias de todo tipo gracias a las ventajas otorgadas por la tecnología [3].

Como lo exponen los autores en [3], [4], [5] y [6], la información es vital para mejorar la efectividad en la toma de decisiones, teniendo resultados positivos en organizaciones de mediano a gran tamaño e inclusive siendo aún más efectivas en organizaciones con una cultura organizacional madura que impulsa sus decisiones basándose en los datos.

Uno de los aspectos más importantes en BA es la recolección y análisis de información que reduce la incertidumbre al tomar una decisión. Aspectos relacionados con la demanda: hábitos de consumo de su cliente, gustos, lugares frecuentados, nivel de gasto, entre otros; y aspectos relacionados con la oferta: composición del comercio, consumo y gasto promedio en la zona, entre otros; son vital importancia en este proceso [7].

Por otro lado, una de las tendencias que más se ha acogido en la industria de los servicios financieros es el de la sostenibilidad. Los modelos de negocio sostenibles son aquellos que se preocupan por el medio ambiente y la sociedad a la vez que generan un beneficio económico, representando ventajas competitivas para las entidades financieras como incremento en la reputación o reducción de costos [8].

El año 2020 trajo consigo grandes retos, principalmente por la coyuntura generada por la pandemia del COVID-19, la ruptura de los acuerdos entre Rusia y la OPEP, entre otros, formando una desestabilización del panorama económico global [9]. Esto creó en las organizaciones una presión para adaptarse a esta nueva realidad generando aceleraciones en temas como sostenibilidad, computación en la nube, toma de decisiones apalancada en datos y *Open Banking* [10].

Por todo esto se desarrolla un modelo espacial de la oferta y la demanda en el territorio colombiano a partir de analítica de datos financieros con el objetivo de aportar conocimiento en el proceso de toma de decisiones al interior de las organizaciones, contribuyendo en su efectividad y aportando a la sostenibilidad de las instituciones financieras, factores clave ante el complejo contexto actual.

2. Objetivos

2.1. Objetivo general

Desarrollar un producto de datos que permita el análisis de un modelo espacial de oferta y demanda en el territorio colombiano a partir de analítica de datos financieros.

2.2. Objetivos específicos

- Definir las variables relevantes al momento de realizar un análisis de oferta y demanda sobre una zona espacial definida.

- Construir un conjunto de datos, incluyendo transacciones financieras, que permita el modelado espacial de la oferta y la demanda en el territorio colombiano.

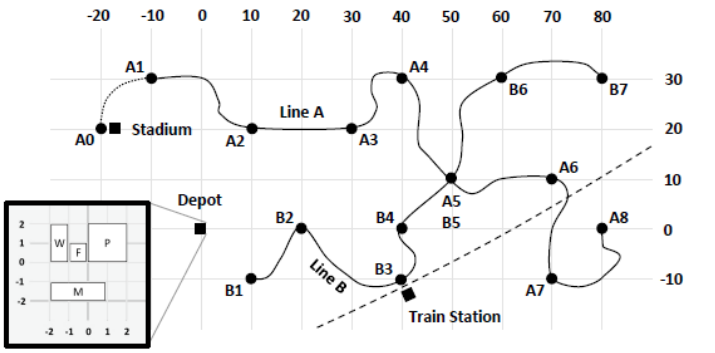
- Desarrollar una herramienta que permita el análisis de los resultados del modelado espacial de oferta y demanda en el territorio colombiano.

3. Estado del arte

3.1. Sistemas de geolocalización

La geolocalización es la habilidad de situar personas u objetos en el espacio mediante un sistema de coordenadas [11]. Google define un sistema de geolocalización como como aquel que “determina la posición de cualquier objeto, persona o vehículo con un margen de error de unos pocos metros” [12].

Con la llegada del internet y los dispositivos móviles, los sistemas de geolocalización toman un nuevo nivel de importancia. Los datos geolocalizados pueden ser recolectados en múltiples contextos y de manera fácil y asequible. Por ejemplo, puede recolectarse la información de la posición en que se encuentra un vehículo de una flota de transportadores mediante los logs generados por dispositivos de rastreo [13].



(a) Map of the bus service lines A and B and the bus depot. The bus depot is identified as four zones: (P) the parking lot, (F) the fuel pumps, (W) the washing machine, and (M) the maintenance facilities.

Stop	First Service		Last Service	
A0	6:50*	8:50*	22:50*	0:50*
A1	7:00	8:40	23:00	0:40
A2	7:10	8:30	23:10	0:30
A3	7:17	8:23	23:17	0:23
A4	7:22	8:18	23:22	0:18
A5/B5	7:30	8:10	23:30	0:10
A6	7:42	7:58	23:42	23:58
A7	7:55	7:45	23:55	23:45
A8	8:10	7:30	0:10	23:30

(b) Line A timetable.

Stop	First Service		Last Service	
B1	6:30	7:50	23:30	0:50
B2	6:40	7:40	23:40	0:40
B3	6:55	7:25	23:55	0:25
B4	7:03	7:17	0:03	0:17
B5/A5	7:09	7:11	0:09	0:11
B6	7:23	6:57	0:23	23:57
B7	7:35	6:45	0:35	23:45

(c) Line B timetable.

Figura 1. Mapa de dos rutas de buses con sus cronogramas en el primer y último recorrido. Tomado de [13]

Gracias a la información geolocalizada que se representa en la Figura 1, es posible identificar el estado en que se encuentra cada vehículo y aplicar técnicas de minería de datos que permitan tener un marco de referencia para el proceso, detectar cuellos de botella, identificar patrones y cambios, cuantificar tiempos de operación y retrasos, validar si se cumplen los cronogramas y comparar rendimientos entre vehículos y conductores, entre otros [13].

Existen otras formas de geolocalizar personas a través de objetos cotidianos como dispositivos móviles o navegadores de internet. Según [14], existen tres

grandes categorías en las que sitios web maliciosos identifican nuestra ubicación:

- Técnicas basadas en la dirección IP: su objetivo es descubrir información geográfica dada una dirección IP, incluyendo el país, ciudad, latitud y longitud. La Figura 2 es un ejemplo de geolocalización a través del cómputo del retraso entre el objetivo y diferentes puntos de referencia para luego calcular sus intersecciones resultando en la ubicación aproximada.

- Técnicas basadas en navegadores: buscan acceder a variables de los servidores web que expongan la ubicación del usuario o mediante la ejecución de scripts dentro del navegador del objetivo.

- Atributos del protocolo de comunicación y transmisión: se enfocan en obtener información del usuario mediante el análisis de los paquetes que se intercambian a través del protocolo HTTP.

Figura 2. Ubicación estimada del objetivo mediante técnicas basadas en la dirección IP. Tomado de [14].

Otra forma de geolocalizar es utilizar información almacenada en otras fuentes, como el código postal o la dirección de residencia, y convertirla en información geocodificada, asignándole unas coordenadas que se aproximen a su

ubicación real de residencia. En [15], los autores realizan un proceso de geolocalización iterativo para asignar latitud y longitud a las residencias de los participantes de un estudio médico. Este proceso consistía en la estandarización de su dirección de residencia bajo los estándares del Servicio Postal Estadounidense para luego contrastarlo con varios Sistemas de Información Geográfica hasta obtener un resultado congruente. Dado el caso de que el primer resultado no fuera exitoso, se contactó el participante para rectificar la información y nuevamente ejecutar el proceso. Como resultado, el 97% de los participantes fueron correctamente asignados, mientras que al restante 3% se le asignó las coordenadas de los centroides de su código postal.

Por otro lado, en [16], los autores quisieron validar la efectividad de imputar las coordenadas del código postal como un aproximado de la dirección de residencia, comparando las distancias lineales entre estas. En la Figura 3 se ve un ejemplo en la que se comparan las ubicaciones reales versus las distancias del código postal.

Los resultados mostraron que el 87.9% de las ubicaciones estimadas estaban a un rango de 200 metros de distancia de las ubicaciones reales. Además, el 96.5% estaban a un rango de 500 metros. Con esto, los autores concluyen que las ubicaciones del código postal son un buen aproximado de la ubicación real de residencia. Sin embargo, los autores aclaran que esto depende del tipo de uso que se le vaya a dar a la información ya que, por ejemplo, algunos estudios en salud requieren mayor precisión en la ubicación [16].

3.2. Geomarketing

El geomarketing es un concepto que utiliza información geolocalizada como una herramienta de marketing para la toma de decisiones y promoción de una organización. Con la aparición del internet y los dispositivos móviles se ha generado una mejora tanto en la calidad como en la cantidad de información, convirtiéndolo en una herramienta estratégica de vital importancia [11].

Existen muchas aplicaciones del geomarketing que van desde temas más macro, como medir la influencia de un centro comercial en una población o medir el éxito dentro del plan de expansión del comercio minorista, hasta temas más micro, como geolocalizar a un usuario y guiarlo en el proceso de compras.

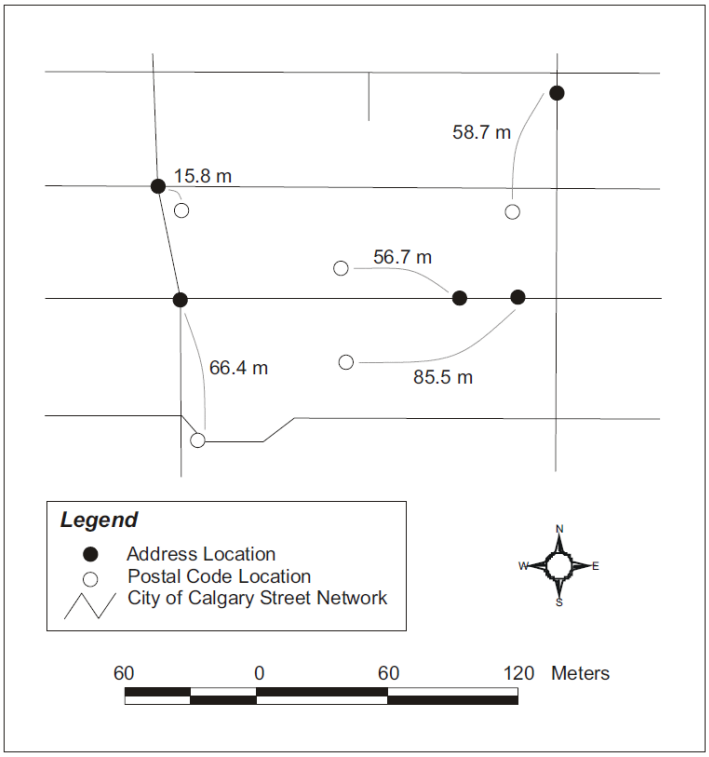


Figura 3. Ejemplo de comparación entre ubicación real versus ubicación aproximada por el código postal. Tomado de [16]

En [17], los autores proponen una solución de geomarketing para un almacén mediante la geolocalización del usuario a través del Bluetooth de su dispositivo móvil donde buscan hacer la experiencia de compras más fácil y rápida, ayudando a los clientes a encontrar productos fácilmente, recibir mensajes de mercadeo personalizados y evitar cualquier promoción poco interesante.

Esta solución de geomarketing no solo ayuda a los clientes, sino que le permite a la administración del almacén conocer información de la relación de los consumidores con los diferentes productos y, por lo tanto, organizar los productos, exposiciones y diseño de la tienda con base en las decisiones y comportamientos de los clientes [17].

Una de las actividades del geomarketing es la segmentación de mercados: *clustering* de la información en varios grupos basados en criterios geográficos. En [18], los autores proponen una mezcla entre segmentación de mercados basado en criterios geográficos y algoritmos de clustering para el manejo de datos geolocalizados tridimensionalmente, considerando el desarrollo mixto de zonas urbanas donde cada vez más se construyen edificios de varios niveles,

esto con el fin de evitar resultados imprecisos. Las aplicaciones de este enfoque pueden ir desde listar la competencia hasta encontrar los clientes potenciales de una nueva tienda en un radio específico, como se puede observar en la Figura 4.

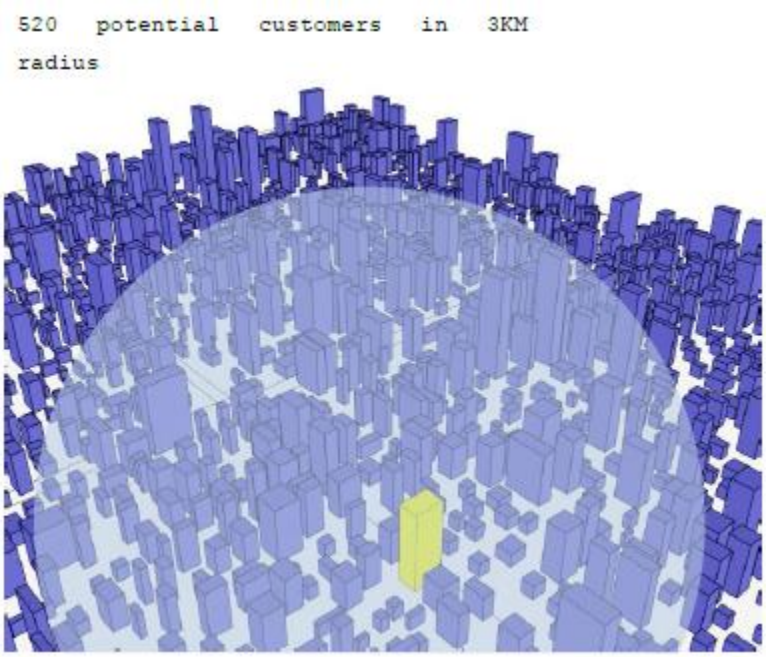


Figura 4 Clientes potenciales en un radio de 3 Km de una nueva tienda (edificio amarillo). Tomado de [18]

Otro uso importante del geomarketing es apoyar la toma de decisiones sobre la apertura de tiendas de comercio minorista. Aplicar un plan de expansión es costoso y, por lo tanto, representa una exposición a riesgos financieros. Además, si la nueva tienda falla debido a una mala ubicación se puede afectar la imagen de la marca y dificultar la experiencia a los consumidores, disminuyendo su lealtad. Eso sumado a un ambiente más competitivo ha resultado en que tomar esta decisión sea aún más difícil [19].

Para atacar este problema, en [19], los autores desarrollan un modelo espacial iterativo que determina la probabilidad de comprar de los clientes combinando teoría de ubicación de supermercados, geomarketing, sistemas de información geográfica y juicio gerencial. Los resultados muestran que las características sociodemográficas del área de influencia del supermercado afectan los resultados del modelo y, por ende, incorporar esta información en la toma de decisiones al respecto.

Similarmente, en [7], el autor desarrolla una metodología para estimar la compatibilidad demográfica entre las empresas de comercio y los centros comerciales permitiendo medir el desempeño de las empresas y sus diferentes establecimientos. Esta metodología tiene el potencial para suplir la necesidad de información de los directivos al momento de localizar nuevos puntos de venta.

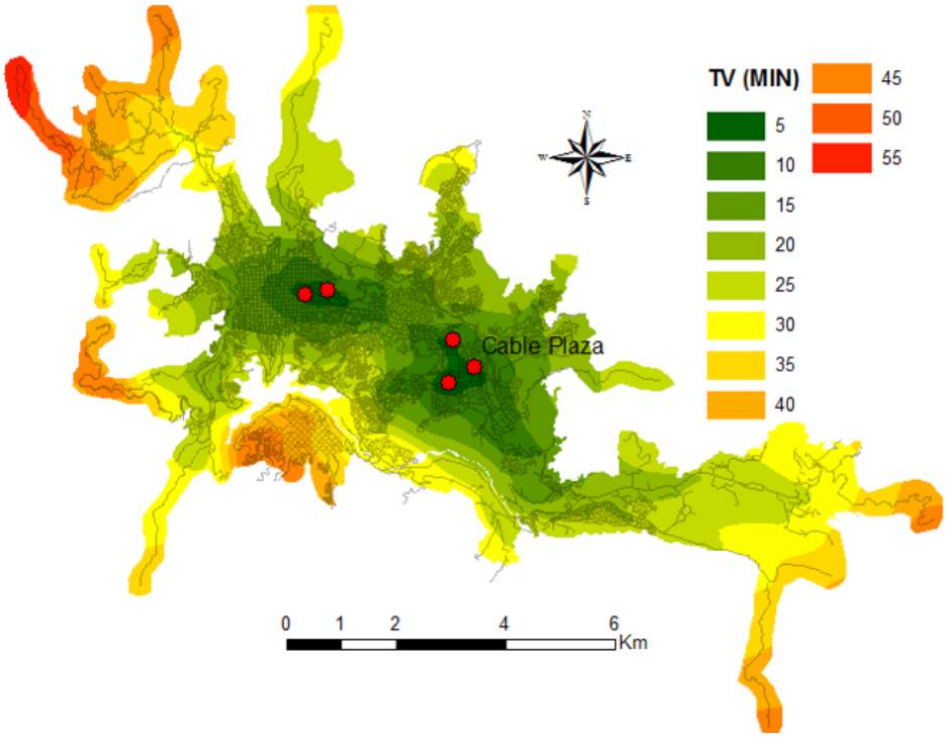


Figura 5 Mapa de curvas isocronas de los centros comerciales (puntos rojos). Tomado de [20].

Medir la influencia también es una aplicación del geomarketing. En [20], los autores analizan la relación entre la ubicación geográfica de los centros comerciales más grandes y su influencia sobre la población de la ciudad de Manizales. Para esto, calculan los tiempos de desplazamiento hacia los centros comerciales a través de la red de transporte público. A esto se le conoce como curvas isocronas y se representan en la Figura 5.

En conclusión, el geomarketing comprende un amplio espectro de técnicas, alcances y enfoques pero que se basan en el uso de información geográfica para apoyar los procesos, la toma de decisiones y generar beneficios tanto para las empresas como para los consumidores.

3.3. La información transaccional en Colombia

La nueva era tecnológica trajo consigo cambios e innovación en las formas de pago y flujos de efectivo, generando impactos positivos especialmente en las instituciones financieras ya que han diversificado sus productos y se han adaptado para fortalecer y apalancar el ámbito transaccional en todo el territorio. La evolución en plataformas y *apps* utilizadas para este fin ha facilitado la interacción entre usuarios y entidades, resultando en un creciente comportamiento transaccional [21].

En Colombia para el año 2015, el 75.4% de los adultos tenía al menos un producto financiero con algún establecimiento de crédito, representando un incremento del 24.3% comparado con el 2006. Adicionalmente en ese mismo periodo, los pagos electrónicos han incrementado a un ritmo del 13.7% anual. Para este mismo año, en promedio una persona realiza 12 pagos electrónicos (incluye transacciones hechas con tarjetas débito, crédito y transferencias electrónicas entre cuentas) al año y la cantidad de datáfonos por cada millón de habitantes es de 6.557,08 [22].

Esta cifra muestra que el crecimiento ha sido sostenido y que continúa con tendencias positivas. Sin embargo, a pesar de que se cuenta con una cifra de datáfonos por millón de habitantes similar a países como Perú y Sudáfrica, con 6.634,19 y 6.686,11, la cantidad de pagos electrónicos per cápita es inferior, con 16,68 y 68,81 respectivamente, y alejada de economías establecidas como Canadá, con 316,60 pagos electrónicos per cápita y 36.299,46 datáfonos por millón de habitantes, o Suecia, con 428,52 pagos electrónicos per cápita y 18.759,30 datáfonos por millón de habitantes [22].

Estos indicadores muestran un panorama en que la inclusión financiera se convierte en un tema clave reducir los riesgos que trae la masificación del uso del efectivo, como la reducción del potencial de crecimiento económico, ineficiencias que promueven economías sumergidas (puesto que las transacciones con este medio no tienen trazabilidad), la reducción de la capacidad de control fiscal facilitando la evasión y elusión de impuestos, la facilitación de actividades ilegales y la promoción de la informalidad. Por lo que se lanzó el Proyecto F, un proyecto interinstitucional liderado por las instituciones bancarias en compañía del gobierno nacional para la masificación de los pagos electrónicos en el país y la reducción del uso del efectivo a través de iniciativas de recaudo electrónico, pagos electrónicos e interoperabilidad, como lo muestra la Figura 6 [23].

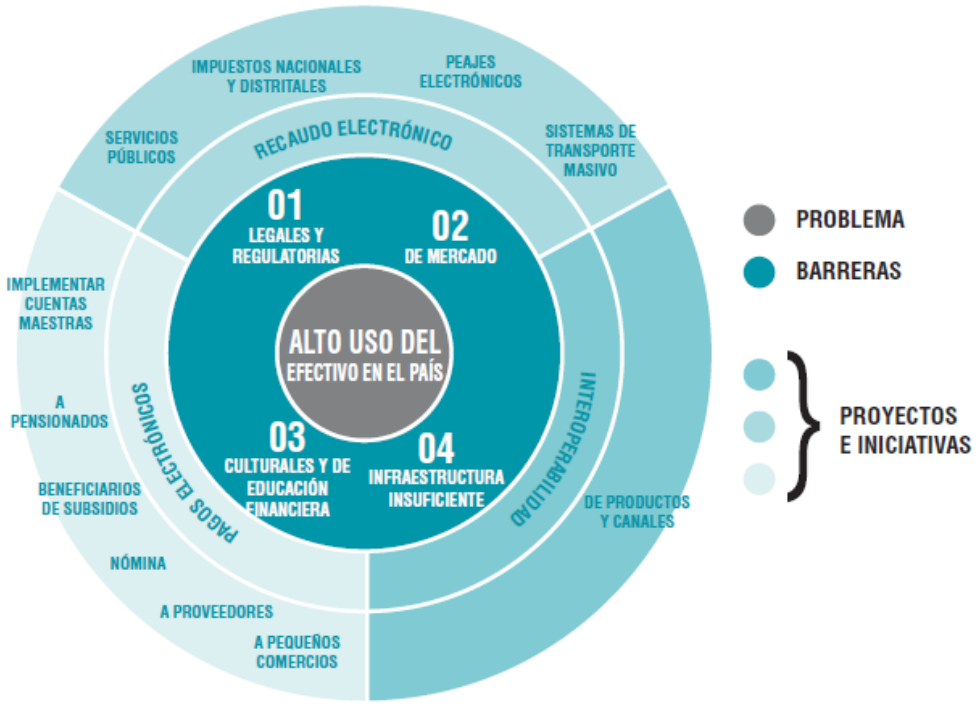


Figura 6. Ejes de trabajo del Proyecto F. Tomado de [23]

El trabajo conjunto de las instituciones financieras del país ha mejorado los indicadores de accesibilidad. Según los indicadores de inclusión financiera, tomados de [24], para el cierre de 2019 en el país había 16,05 millones de tarjetas crédito y 33,09 millones de tarjetas débito vigentes, representando un aumento del 7,50% y 31,46% frente al 2016 respectivamente. También, para el año 2019, el 81,6% de adultos contaba con al menos un producto financiero con algún establecimiento de crédito, representado un incremento del 6,2% respecto a la cifra de 2015.

3.4. El uso de datos financieros en el sector bancario

En los últimos años las entidades financieras han experimentado un incremento en la habilidad de generar, capturar y almacenar datos con el fin de crear conocimiento y generar un beneficio propio que le de ventaja sobre sus competidores. La información contenida en los datos crudos es de vital importancia, por lo cual las entidades financieras alrededor del mundo están implementando técnicas de minería de datos para extraer este conocimiento de la manera más eficiente, rápida y efectiva posible. El concepto del negocio bancario ha evolucionado para ser un sistema técnicamente fuerte y orientado en los clientes. Es entonces cuando los datos pueden ser uno de los recursos

más valiosos con los que cuentas las entidades financieras, siempre y cuando sepan como extraer el conocimiento valioso que se encuentra escondido en los datos crudos [25].

De acuerdo con [26], la minería de datos en las entidades financieras se concentra principalmente en resolver tres necesidades: retención de clientes, prevención del fraude y detección del fraude. Los árboles de decisión que utilizan el algoritmo CART (árboles de regresión y clasificación, por sus siglas en inglés) son de los más utilizados para la retención y profundización de los clientes. Por otro lado, prevenir el fraude es más económico que detectarlo, por esto técnicas como árboles de decisiones, SVM (máquinas de soporte vectorial, por sus siglas en inglés) y regresiones logísticas son aplicadas en el proceso de aprobación de tarjetas de crédito. Para la detección del fraude, la clusterización mediante el algoritmo EM (Esperanza – Maximización) es de los más usados. En la Figura 7 se muestran las principales necesidades seguidas de las técnicas y algoritmos más utilizadas.

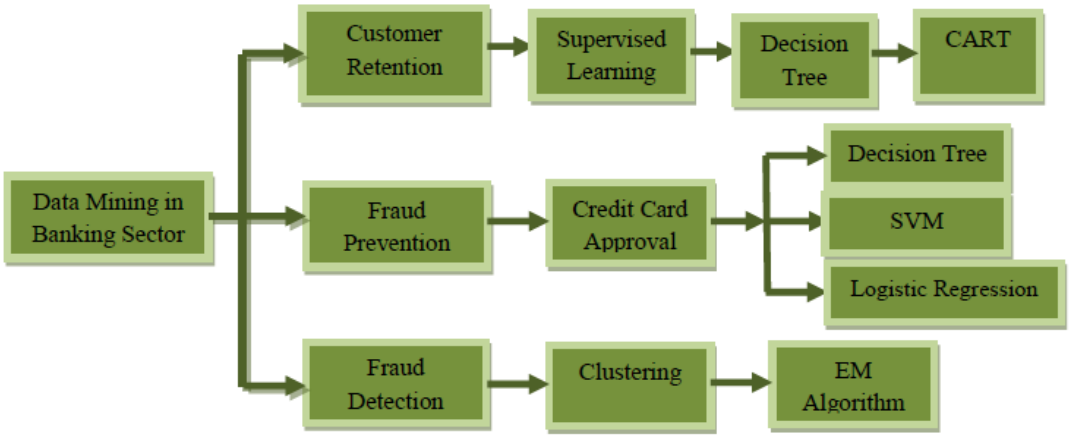


Figura 7. Necesidades y técnicas de minería de datos aplicadas en el sector bancario. Tomado de [26]

A pesar de que las necesidades anteriores son las más comunes dentro del sector bancario, muchas otras necesidades como el manejo de riesgo, marketing, atracción de nuevos clientes, análisis de clientes, búsqueda de patrones transaccionales a lo largo del tiempo, entre otras, son tratadas utilizando diferentes técnicas de la minería de datos buscando siempre crear un ambiente orientado a los datos para mejorar el proceso de toma de decisiones que genere un gran beneficio y derive en una ventaja competitiva considerable que se mantenga en el tiempo frente a la competencia directa [26].

En los últimos años también ha crecido el uso de la inteligencia de negocios al interior de las entidades financieras, la cual se define según [27] como un

conjunto de habilidades, tecnologías y aplicaciones utilizadas para recolectar, almacenar, analizar y crear acceso efectivo a la información con el objetivo de ayudar a las organizaciones a tener un mejor entendimiento del contexto del mercado, a tomar decisiones acertadas y oportunas.

Combinar técnicas de minería de datos con implementaciones de inteligencias de negocios permiten a las organizaciones analizar grandes cantidades de datos en búsqueda de indicadores vitales y conocimiento que sea entregado oportunamente a las personas interesadas en la información. En la Figura 8 se representa, de izquierda a derecha, el proceso de transformación de los datos en indicadores y conocimiento. En un principio se recolecta datos, tanto de fuentes internas como lo son sistemas de bases de datos operacionales, como de fuentes externas como el internet. Luego pasan por un proceso de recolección y ETL (extracción, transformación y carga por sus siglas en inglés). Después son almacenados en almacenes de datos o sistemas de bases de datos agregados para luego ser analizados por diferentes herramientas de la inteligencia de negocios y producir conocimiento e indicadores que ayuden a las organizaciones [27].

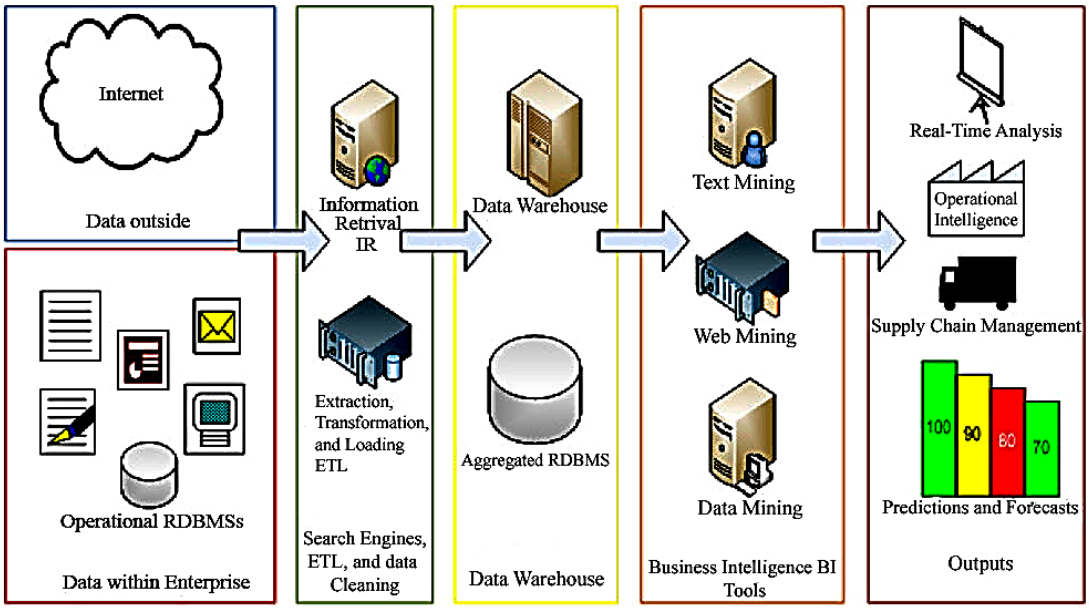


Figura 8. Transformación de los datos a Inteligencia de Negocios. Tomado de [27]

Normalmente en el sector bancario el uso de la minería de datos y la inteligencia de negocios se realizan a un nivel interno en donde los clientes de las entidades financieras no tienen acceso a esta información. Sin embargo, en los últimos años se han venido lanzando aplicaciones que permiten que clientes tengan acceso a cierta información que le ayude a su operación y por la

cual las entidades financieras buscan estrechar los vínculos con sus clientes al convertirse en aliados estratégicos. A continuación se presentan algunas iniciativas lanzadas por entidades financieras alrededor del mundo.

3.4.1. Commerce 360, Grupo BBVA

El Grupo BBVA, a través de BBVA Data & Analytics, puso a disposición de sus clientes en España la herramienta Commerce 360, una herramienta web que convierte en forma anónima y agregada los datos de compras en comercios con tarjetas BBVA realizados en terminal punto de venta (TPV) físicos en información útil para los negocios de sus clientes, siempre y cuando tengan o contraten el servicio de TPV con BBVA. Los clientes tienen acceso a diversos indicadores estadísticos que describen el comportamiento de la actividad económica de su comercio, su cliente, la zona de influencia en la que opera, entre otros [28].

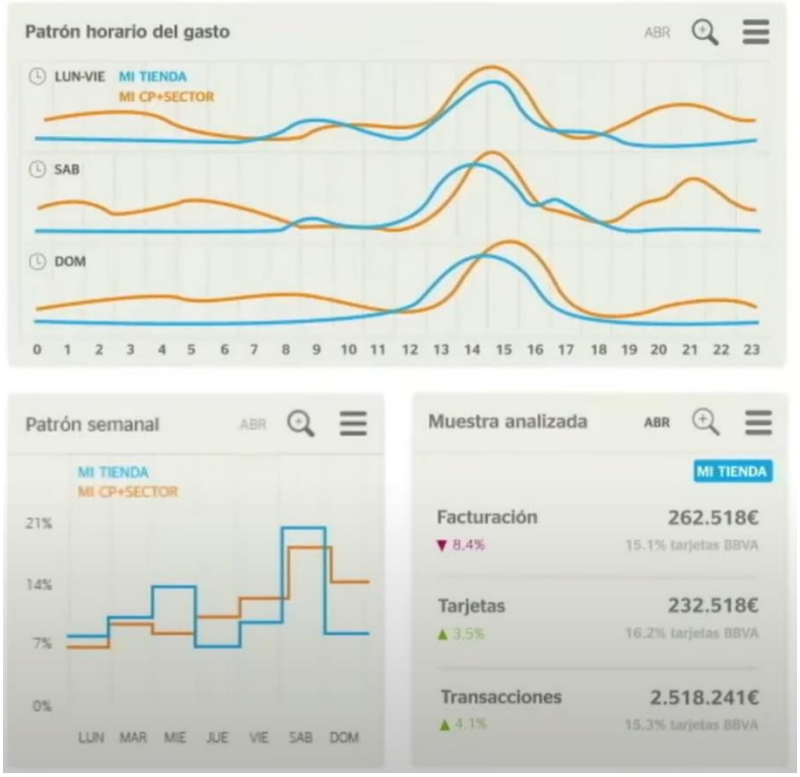


Figura 9. Ejemplo de indicadores de Commerce 360. Tomado de [28]

En la Figura 9 se observan 3 ejemplos de los indicadores que se le presentan al cliente de Commerce 360, los cuales muestran el comportamiento de su establecimiento de comercio versus, por ejemplo, los demás establecimientos

que pertenecen a la misma categoría de comercio y que están ubicados en su mismo sector, es decir su competencia directa. También se muestran indicadores generales y sus variaciones.

El Grupo BBVA tiene una amplia experiencia y es uno de los bancos pioneros en el uso de tecnologías “big data”. Mediante Commerce 360, pone a disposición de sus clientes su gran cantidad de información buscando que estos se vean beneficiados y, a su vez, sumar 100.000 nuevos clientes por año y aumentar la facturación de sus TPV a razón del 25% apoyado en su amplia y cada vez más creciente oferta de productos y servicios [28].

3.4.2. Plink, Grupo Bancolombia

Plink es una herramienta cuyo objetivo es ayudar a sus clientes a aumentar sus ventas mediante la recopilación y procesamiento de las transacciones comerciales realizadas diariamente en toda Colombia para luego presentar gráficos comparativos y proyecciones que permitan buscar soluciones para los clientes, generar mejoras o transformaciones que impacten positivamente las ventas de la empresa [29].

La información que se presenta en la herramienta proviene del servicio de adquirencia contratado con el Bancolombia y, a pesar de que no incluye las ventas en efectivo, la gran participación que este banco representa permite contar con datos estadísticamente significativos para la gran mayoría de los casos [29].

En la Figura 10, se representa un caso de uso real de la aplicación. Para este caso hipotético, un restaurante de comida italiana analiza la procedencia de su clientela. Esta información le permite realizar un mejor proceso de toma de decisiones, ya que ante un plan de expansión, es decir abrir nuevos locales en otras zonas de la ciudad, puede seleccionar o descartar una ubicación según su presencia en los clientes de la zona. También, al tener información de la procedencia de su clientela, se puede realizar una estrategia de marketing más efectiva ya que puede enfocarse en las zonas de procedencia para retener a su clientela o puede enfocarse en otras zonas para atraer nuevos clientes.

Figura 10. Caso de uso de Plink en un escenario real. Tomado de [29]

La herramienta cuenta con varios módulos en los que se puede conocer el desempeño del establecimiento a nivel general para encontrar oportunidades de crecimiento, conocer a profundidad las características principales del perfil de clientes del establecimiento, conocer en detalle los hábitos de consumo de

los clientes y poder comparar el establecimiento con la competencia directa y con la misma categoría de mercado a la que pertenecen [30]. En la Figura 11 se observa unas de las métricas que ofrece el módulo de análisis de mercado.

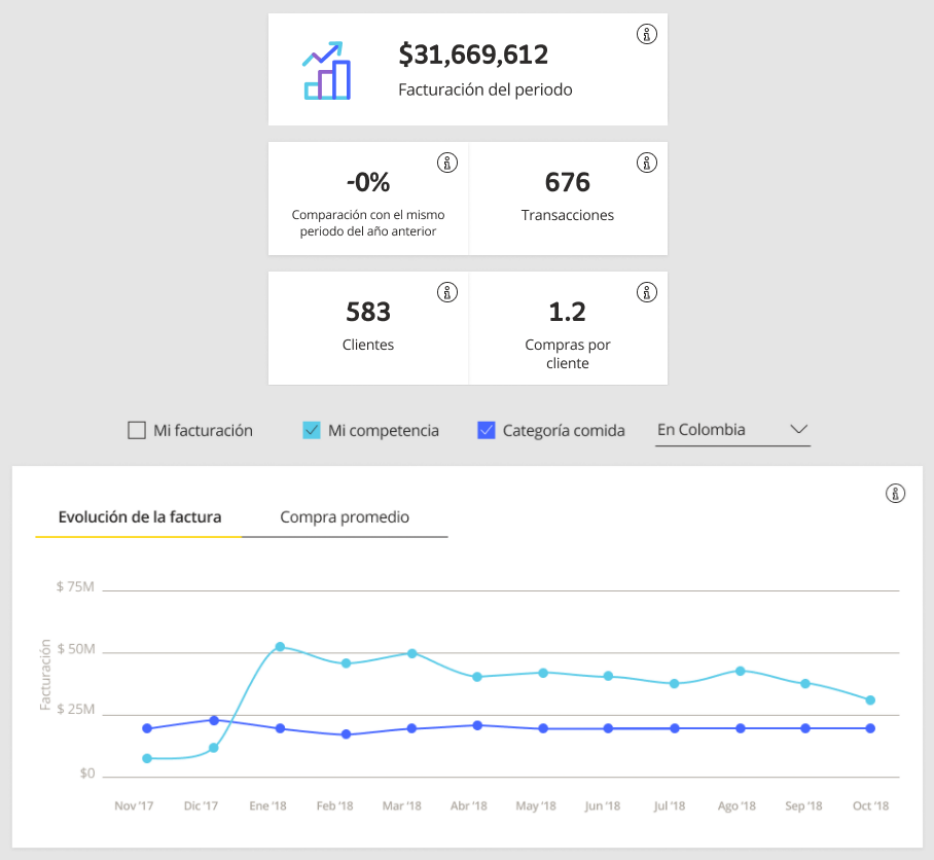


Figura 11. Métricas que ofrece uno de los módulos de Plink. Tomado de [30]

4. Metodología propuesta

4.1. Modelo de datos

El proceso de pagos en establecimientos de comercio tiene varios actores. Por un lado, están los clientes quienes son personas naturales que realizan compras en establecimientos de comercio que ofrecen bienes o servicios. Por otro lado, se encuentran las redes, para el caso colombiano Credibanco y Redeban, que son aquellas empresas encargadas de colocar los datafonos y asegurar el intercambio de fondos entre los clientes y los establecimientos. Por último, están las instituciones financieras, las cuales se encargan de brindar recursos a los clientes mediante sus productos (tarjetas débito y crédito) y de recolectar los recursos que los establecimientos reciben de su actividad,

requisito para poder acceder al servicio. En la Figura 12 se ejemplifica esta relación.

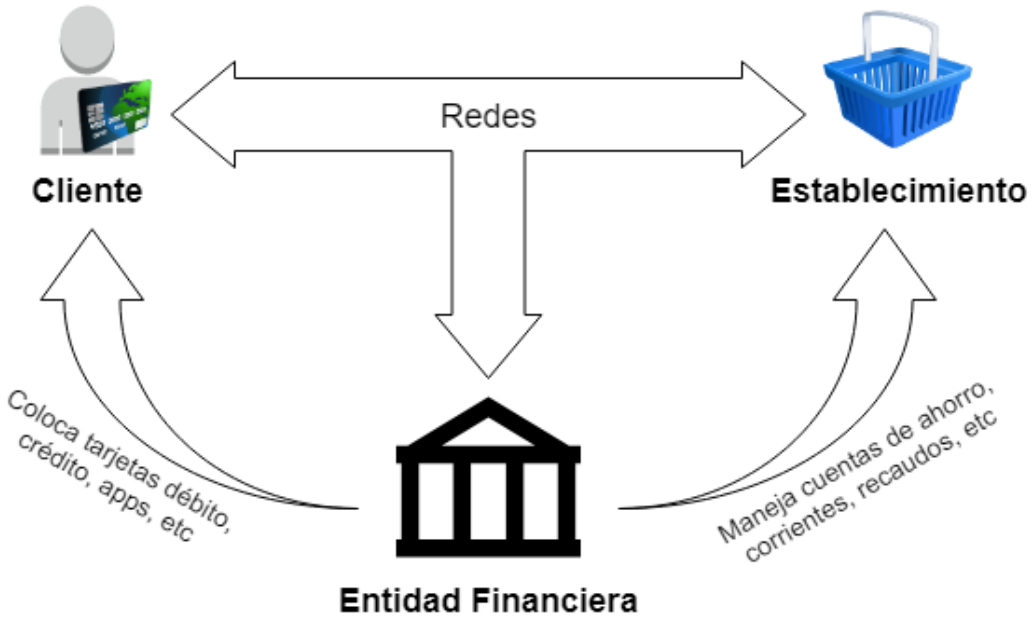


Figura 12. Actores participantes en el proceso de pago de un establecimiento de comercio. Fuente: elaboración propia

Este proceso de pago genera información que puede ser analizada según el actor que intervenga. Las redes pueden medir la información transaccional que pasa por ellas, pero muchas veces no conocen quien es el cliente detrás del medio electrónico con el que hizo la compra, restringiendo su capacidad de análisis demográficos. Los establecimientos pueden almacenar esta información con su sistema de gestión de relaciones (*Customer Relationship Management, CRM*) pero, al igual que las redes, pueden tener información limitada sobre sus clientes y es posible que muchas empresas no cuenten con la capacidad y habilidad para analizar su propia información.

Bajo este panorama, las entidades financieras se encuentran en una posición privilegiada ya que poseen información sobre los clientes, mediante formularios y procesos de captura de información cuando estos solicitan productos como cuentas de ahorros, tarjetas débito o crédito, y los establecimientos, cuando estos se registran ante las redes, solicitan créditos para apalancar su operación o planes de expansión, entre otros.

Para que una entidad financiera aproveche esta información, se genera un modelo de datos que consta de 3 clases: Clientes, Establecimientos y Transacciones.

4.1.1. Clientes

Esta clase contiene la información de los clientes pertenecientes a la entidad financiera y que son titulares de productos de tarjetas débito y crédito. Los atributos y descripciones de esta clase son:

- `id`: número de documento de identificación del cliente
- `cod_tip_doc`: código del tipo de identificación del cliente
- `tip_doc`: descripción (nombre) del tipo de documento del cliente
- `nombre`: nombre del cliente
- `sexo`: indica el género del cliente
- `edad`: indica la edad actual del cliente
- `ocupación`: indica la ocupación reportada por el cliente
- `ingresos`: indica los ingresos reportados por el cliente
- `estado_civil`: indica el estado civil actual del cliente
- `latitud`: indica la coordenada de latitud de la residencia del cliente
- `longitud`: indica la coordenada de longitud de la residencia del cliente
- `ciudad`: indica la ciudad de residencia del cliente
- `departamento`: indica el departamento de residencia del cliente
- `país`: indica el indica el país de residencia del cliente
- `cod_ciudad`: es el código de la ciudad de residencia del cliente
- `cod_departamento`: es el código del departamento de residencia del cliente
- `cod_pais`: es el código del país de residencia del cliente.

4.1.2. Establecimientos

Esta clase contiene la información de los establecimientos de comercio que reciben pagos con medios electrónicos, específicamente compras con tarjetas débito y crédito. Los atributos y descripciones de esta clase son:

- `cod_estab`: código de identificación del establecimiento
- `nombre`: nombre del establecimiento
- `dirección`: dirección de la ubicación del establecimiento

- `dirección_estandar`: dirección de la ubicación del establecimiento estandarizado mediante la aplicación de reglas de unificación
- `cod_ciudad`: código de la ciudad donde está ubicado el establecimiento
- `ciudad`: ciudad donde está ubicado el establecimiento
- `cod_mcc`: código de categoría mercante (*merchant category code, MCC*) que sirve para clasificar el tipo de bienes o servicios que provee y que está definida por [31]
- `nit`: número de identificación del propietario del establecimiento
- `razón_social`: nombre del propietario del establecimiento
- `latitud`: indica la coordenada de latitud de ubicación del establecimiento
- `longitud`: indica la coordenada de longitud de ubicación del establecimiento

4.1.3. Transacciones

Esta clase contiene la información de las compras hechas por parte de los clientes en los establecimientos mediante pagos con tarjetas débito y crédito. Los atributos y descripciones son:

- `id`: número de documento de identificación del cliente que realiza la transacción
- `tipo_id`: código del tipo de identificación del cliente que realiza la transacción
- `cod_estab`: código del establecimiento donde se realiza la compra
- `num_trj`: número de la tarjeta débito o crédito con la que se hizo el pago en la transacción
- `bin`: código asociado a la tarjeta
- `f_trx`: fecha en la que se hizo la transacción
- `h_trx`: hora en la que se hizo la transacción
- `vlr_trx`: valor en pesos (COP) por la que se hizo la transacción
- `tipo_tarjeta`: indica si la compra se realizó con una tarjeta débito o una tarjeta de crédito

4.1.4. Diagrama de clases del modelo

Como se puede apreciar en el listado de atributos de las clases, en el modelo existe una relación entre ellas, la cual se muestra en la Figura 13 mediante el diagrama de clases de UML. Adicional a estas 3 clases, se creó una cuarta

clase llamada “MCC” que contiene la descripción de los códigos de categoría mercante.

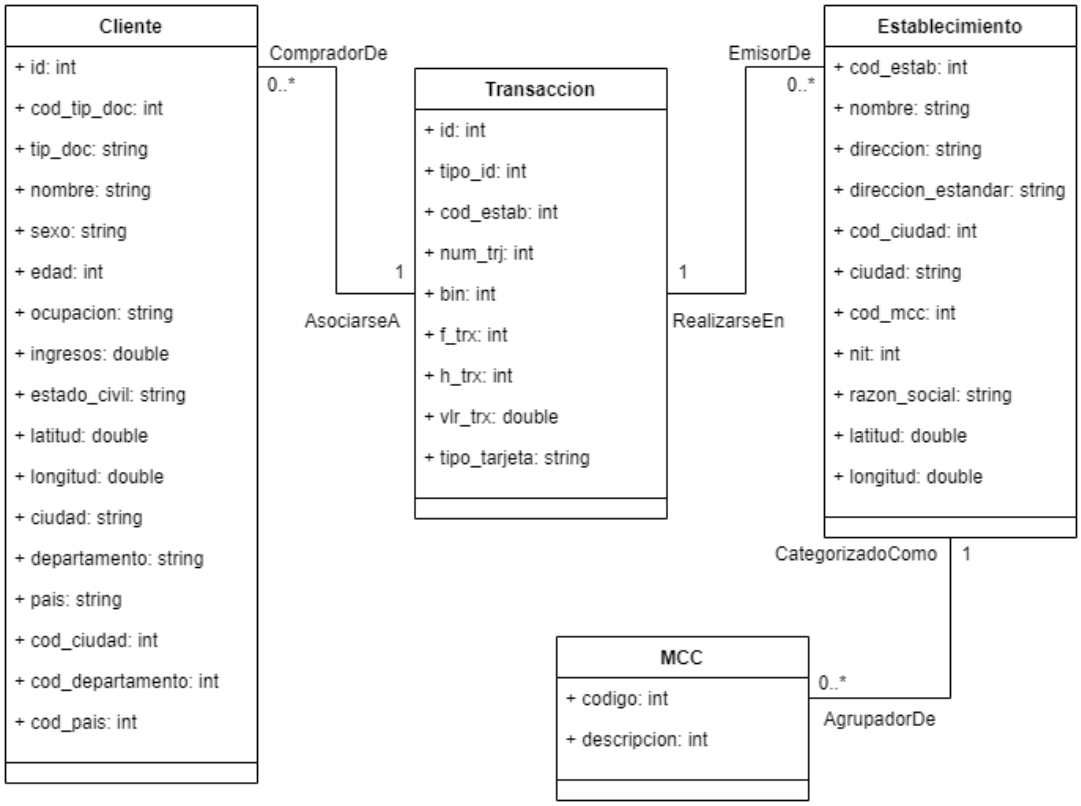


Figura 13 Diagrama de clases UML del modelo de datos. Elaboración propia

4.2. Procesamiento de datos

Para el procesamiento de los datos se tienen en cuenta tres etapas. La primera etapa consta de un proceso de categorización de variables y de protección de datos personales de los clientes. Una vez cumplida esta etapa se da paso a las segunda y tercera etapa que consiste en la extracción de la información para modelar la oferta y la demanda respectivamente.

4.2.1. Preprocesamiento, protección de datos personales y categorización de variables

Lo primero en todo proyecto que utiliza información de personas naturales, como lo son los clientes con compras con tarjetas débito y crédito que hacen parte de este modelado de oferta y demanda, es garantizar la protección de

sus datos personales. Es por esta razón que cualquier dato que ayude a su identificación debe ser excluido o categorizado en grupos demográficos para mantener la privacidad de la información y la reserva bancaria.

Durante la etapa de preprocesamiento y cumpliendo con lo anterior, los atributos `id`, `cod_tip_doc`, `tip_doc`, `nombre`, `ciudad`, `departamento`, `país`, `cod_ciudad`, `cod_pais` y `cod_departamento` deben ser excluidos de la clase Cliente. También se excluyen los atributos `nombre`, `dirección`, `dirección_estandar`, `cod_ciudad`, `nit`, y `razón_social` de la clase Establecimiento. Por último, de la clase Transacción se excluye el atributo `num_trj`.

Los atributos edad e ingresos pertenecientes a la clase Cliente son de importancia para el modelo, sin embargo, se encuentran como atributos cuantitativos. Para facilitar su análisis, se convierten a atributos cualitativos mediante un proceso simple de categorización: para la edad, se crean rangos entre con sus valores y para los ingresos, se dividen entre el salario mínimo mensual legal vigente (SMMLV) y se calculan rangos según este nuevo valor.

Rango_edad:

- Si edad es menor a 17 entonces '< 17 Años'
- Si edad está entre 17 y 24 entonces '17 a 24 Años'
- Si edad está entre 25 y 36 entonces '25 a 36 Años'
- Si edad está entre 37 y 49 entonces '37 a 49 Años'
- Si edad está entre 50 y 64 entonces '50 a 64 Años'
- Si edad es mayor o igual a 65 entonces '> 65 Años'

Rango_ingresos:

- Si ingresos/SMMLV es menor a 1 entonces '< 1 SMMLV
- Si ingresos/SMMLV está entre 1 y 2 entonces '1 a 2 SMMLV'
- Si ingresos/SMMLV está entre 2 y 6 entonces '2 a 6 SMMLV'
- Si ingresos/SMMLV está entre 6 y 10 entonces '6 a 10 SMMLV'
- Si ingresos/SMMLV mayor a 10 entonces '> 10 SMMLV'

4.2.2. Agrupación geográfica y creación de supertipos

Este proceso consiste en realizar una agrupación geográfica a través de las coordenadas y creación de supertipos de clientes mediante agrupación de características e individuos. Primero, redondeamos los atributos de latitud y longitud a tres cifras decimales, que equivalen un mínimo de distancia de 110

metros aproximadamente entre pares de coordenadas. Luego, con estas nuevas coordenadas, agrupamos, por ejemplo, los datos demográficos de los clientes mediante sus atributos demográficos, creando el supertipo Clientes de información demográfica. Para medir el tamaño de este supertipo, se calcula la cantidad de clientes que lo integran.

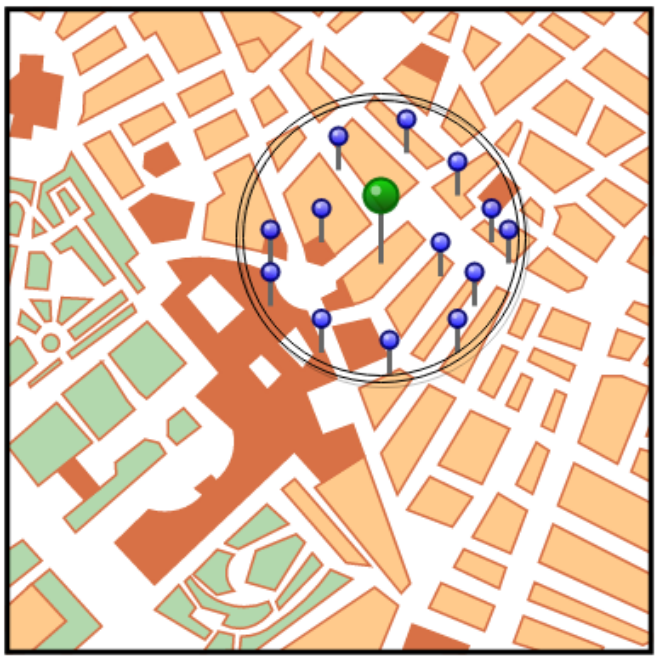


Figura 14. Ejemplo de agrupación geográfica. Fuente: elaboración propia.

En la Figura 14, se puede apreciar cómo los clientes dentro de un rango generado por el redondeo de las coordenadas, representados por los pines azules, se contraen en un supertipo de clientes, representado por el pin verde. En la Tabla 1 también se puede ver un ejemplo del tipo de información que contiene el supertipo de Clientes de información demográfica.

Tabla 1. Ejemplo de información del supertipo Clientes Información Demográfica generado a partir de una muestra de los datos. Fuente: elaboración propia

Latitud	Longitud	Ocupación	Sexo	Rango_edad	Rango_ingresos	Cnt_Clientes
6.241	-75.569	Empleado	M	17 a 24 Años	< 1 SMMLV	6
6.241	-75.569	Empleado	F	25 a 36 Años	2 a 6 SMMLV	10
6.241	-75.569	Empleado	M	17 a 24 Años	1 a 2 SMMLV	11
6.241	-75.569	Independiente	F	25 a 36 Años	> 10 SMMLV	4
6.241	-75.569	Independiente	F	25 a 36 Años	1 a 2 SMMLV	9
6.241	-75.569	Estudiante	M	17 a 24 Años	< 1 SMMLV	15

Para el análisis de este trabajo se crearán, adicional al supertipo de clientes de información geográfica, un supertipo de clientes con información transaccional y otro supertipo de clientes con información transaccional por categorías MCC. En la Tabla 2 y Tabla 3 se puede ver un ejemplo de los supertipos respectivamente.

Tabla 2. Ejemplo de información del supertipo Clientes Información Transaccional generado a partir de una muestra de los datos. Fuente: elaboración propia

Latitud	Longitud	Total_trx	Cant_trx	Ticket_promedio	Cnt_Clientes
6.241	-75.585	1575775944	12270	128425.09	167
6.241	-75.564	505081909	5636	89617.08	175
6.241	-75.546	355547063	3841	92566.27	147
6.241	-75.584	788478352	9237	85360.86	142
6.241	-75.597	629519951	6939	90721.99	106

Tabla 3. Ejemplo de información del supertipo Clientes Información Transaccional por Categorías generado a partir de una muestra de los datos. Fuente: elaboración propia

Latitud	Longitud	Cod_Mcc	Total_trx	Cant_trx	Ticket_promedio	Cnt_Clientes
6.241	-75.6	5411	5782660	83	69670.6	76
6.241	-75.588	7995	85320001	23	3709565.26	10
6.241	-75.553	8011	7446400	43	173172.09	35
6.241	-75.601	7230	16903000	100	169030	53
6.241	-75.571	5211	25096036	80	313700.45	70
6.241	-75.592	5722	92088051	825	111621.88	467

De esta manera se protege la identidad y los datos personales de cada uno de los clientes pertenecientes al análisis y al mismo tiempo se obtiene información útil para desarrollar el modelamiento de la oferta y la demanda, permitiendo responder preguntas sobre la demografía de una zona de interés y sobre la transaccionalidad, a nivel general y por categorías.

4.2.3. Modelado de la oferta en el territorio colombiano

Para modelar la oferta en el territorio colombiano se utilizará la geolocalización de los establecimientos de comercio y las transacciones que los clientes realizan en ellos. El proceso consiste en:

- Seleccionar la ubicación geográfica (latitud, longitud) a la cual se le desea aplicar el modelo de la oferta y que esté ubicado dentro del territorio colombiano. Esta ubicación se define como el centroide de nuestra zona de interés

- Seleccionar el radio en metros para el cual se desea modelar de la oferta, definiendo la zona de interés.
- Realiza el proceso de agrupación geográfica y creación de supertipos de los clientes que realizaron compras en los establecimientos ubicados en la zona de interés.
- En un proceso iterativo, se calcula la distancia lineal entre el centroide de la zona de interés con la ubicación de los supertipos. Aquellos supertipos que tienen una distancia menor al radio seleccionado son seleccionados como miembros de nuestra zona de interés.
- Luego se realizan sumatorias y agrupaciones entre los supertipos seleccionados en donde se calculan, **para los establecimientos dentro de la zona de interés**:
    - Facturación total de compras
    - Ticket promedio por compra
    - Cantidad total de compras
    - Cantidad de clientes
    - Distribución de clientes por género
    - Distribución de clientes por rango de edad
    - Distribución de clientes por ocupación
    - Distribución de clientes por rango de ingresos
    - Información transaccional agrupada por categorías
- Por último, se muestra el resultado del modelo en un tablero de control

4.2.4. Modelado de la demanda en el territorio colombiano

Siguiendo un proceso muy similar al modelado de la oferta expuesto en el numeral 4.2.3., la diferencia radica en que para modelar la demanda analizaremos las compras hechas por los clientes que residen en nuestra zona de interés. Por lo tanto, el proceso consiste en:

- Seleccionar la ubicación geográfica (latitud, longitud) a la cual se le desea realizar el modelo de la demanda y que esté ubicado dentro del territorio colombiano. Esta ubicación se define como el centroide de nuestra zona de interés

- Seleccionar el radio en metros al cual se desea modelar de la demanda, definiendo la zona de interés.

- Realiza el proceso de agrupación geográfica y creación de supertipos de los clientes que residen en la zona de interés y realizaron compras en establecimientos de comercio.

- En un proceso iterativo, se calcula la distancia lineal entre el centroide de la zona de interés con la ubicación de los supertipos de la zona candidata. Aquellos supertipos que tienen una distancia menor al radio seleccionado son seleccionados como miembros de nuestra zona de interés.

- Luego se realizan sumatorias y agrupaciones entre los supertipos seleccionados en donde se calculan, **para los clientes residentes de la zona de interés**:
    - Facturación total de compras
    - Ticket promedio por compra
    - Cantidad total de compras
    - Cantidad de clientes
    - Distribución de clientes por género
    - Distribución de clientes por rango de edad
    - Distribución de clientes por ocupación
    - Distribución de clientes por rango de ingresos
    - Información transaccional agrupada por categorías

- Por último, se muestra el resultado del modelo en un tablero de control

4.3. Implementación en la entidad financiera

Para la implementación de esta metodología inicialmente se contempló ASUM-DM (*Analytics Solutions Unified Method*) de IBM. Esta es una guía paso a paso para realizar y completar la implementación de soluciones con el objetivo de acelerar las ganancias tempranas y minimizar los riesgos [32]. Sin embargo, con el objetivo de darle un enfoque más analítico se eligió la metodología sugerida por [33], la cual combina ASUM-DM junto con otras metodologías y se orienta hacia el desarrollo exitoso de proyectos de *Big Data*.

Esta metodología se compone básicamente de tres etapas:

4.3.1. Contextualización y actualización sobre *Big Data*

Esta es una etapa temprana que busca introducir conceptos generales sobre *Big Data* a la organización interesada. Temas como metodologías, plataformas, casos de éxito, desafíos, etc., son tratados en este espacio. También es una oportunidad para evaluar el nivel de madurez de la organización y algunos factores de riesgo (por ejemplo: soporte de gestión, motivación comercial, entre otros).

4.3.2. Entendimiento de los casos de uso y de negocio

El objetivo de esta etapa es recolectar las necesidades o problemas de negocio que deben ser solucionadas usando *Big Data*. También se debe identificar uno o varios casos de uso que generen más valor desde los datos usados. Al final de esta etapa, se debe tener definido la planeación del proyecto.

4.3.3. Implementación de prototipos

En esta fase aplica conceptos ágiles para el desarrollo de las siguientes actividades:

> "***Comprensión de datos****: Analice conjuntos de datos para detectar subconjuntos interesantes y alcanzar los objetivos comerciales asegurando la calidad de los datos.*
>
> ***Preparación de datos****: Seleccione los datos correctos y limpie / extienda / estructure según sea necesario.*
>
> ***Modelado****: Seleccione técnicas de modelado específicas (por ejemplo, árbol de decisión o redes neuronales). El modelo se construye y se prueba para su precisión y generalidad. Los supuestos de modelado posibles también se verifican. Se pueden revisar los parámetros del modelo o se pueden seleccionar otras técnicas complementarias.*
>
> ***Evaluación****: Evalúe el grado en que el modelo cumple con los objetivos comerciales, utilizando datos simulados o incluso reales.*
>
> ***Implementación****: Transfiera la solución validada a un ambiente de producción, asegurándose que los interesados puedan usarla y comience a monitorear*" [33].

Después de la ejecución parcial de la metodología descrita, se realizará un documento que contenga los aportes significativos y el trabajo futuro para la resolución del problema planteado.

5. Resultados esperados

Luego de ejecutar la metodología planteada, se espera que a través del mismo modelo se visualicen cambios en la oferta y demanda entre distintas zonas de interés. También se espera poder comparar dos ubicaciones determinadas y validar que estas arrojen resultados similares con otros modelos o con la literatura.

Además, se espera desarrollar una herramienta en un lenguaje de programación abierto que permitirá acceder al modelo desarrollado para hacer el análisis de oferta y demanda. Dicha herramienta será propiedad intelectual de la entidad financiera.

6. Prototipo de herramienta para analizar geográficamente la oferta y la demanda

6.1. Preprocesamiento de la información

Teniendo en cuenta el modelo de datos representado por el diagrama de clases de UML en el numeral 4.1.4, se generan las siguientes tablas:

6.1.1. Información demográfica de la demanda

Esta tabla contiene la información demográfica de las personas que realizaron compras durante el periodo de estudio en cualquier establecimiento de comercio. Incluye variables como ocupación, género, rango de edad y rango de ingresos. Los registros están geolocalizados con la dirección de residencia reportada por estos mismos clientes y, para preservar la privacidad, se aplica la técnica de agrupación geográfica y creación de supertipos explicada en el numeral 4.2.2.

Una consulta en Impala SQL con base en el modelo de datos del numeral 4.1.4 para esta tabla sería:

```
select
    round(t1.latitud,3) as lat,
    round(t1.longitud,3) as long,
    t1.ocupacion,
    t1.sexo as genero,
    case
        when t1.edad < 17 then '< 17 Años'
        when t1.edad between 17 and 24 then '17 a 24 Años'
        when t1.edad between 25 and 36 then '25 a 36 Años'
        when t1.edad between 37 and 49 then '37 a 49 Años'
        when t1.edad between 50 and 64 then '50 a 64 Años'
        when t1.edad >= 65 then '> 65 Años'
    end as rango_edad,
    case
        when t1.ingresos/828116 < 1 then '< 1 SMMLV'
        when t1.ingresos/828116 between 1 and 2 then '1 a 2 SMMLV'
        when t1.ingresos/828116 between 2 and 6 then '2 a 6 SMMLV'
        when t1.ingresos/828116 between 6 and 10 then '6 a 10 SMMLV'
        when t1.ingresos/828116 > 10 then '> 10 SMMLV'
    end as rango_ingresos,
```

```
    count(distinct t2.id) as cant_cli
from Cliente t1
inner join Transaccion t2 on t1.id = t2.id and t1.cod_tip_doc = 1
where t2.f_trx between 20190101 and 20191231
group by 1,2,3,4,5,6;
```

Para la consulta anterior, la variable `ingresos` se encuentra en pesos colombianos (COP) por lo cual se divide por el valor del salario mínimo mensual legal vigente de Colombia durante el periodo de análisis. Para la variable `f_trx` (es decir fecha de la transacción) el tipo de dato utilizado es `INT` con formato `YYYYMMDD` (es decir la fecha 01/01/2019 pasa a ser 20190101).

6.1.2. Información general de la demanda

Esta tabla contiene la información de las transacciones realizadas por las personas de una zona geográfica específica en cualquier establecimiento de comercio. Contiene indicadores como la facturación total de los clientes en la zona, la cantidad total de compras realizadas por clientes en la zona, el ticket promedio por compra de los clientes de la zona y la cantidad única de clientes de la zona que realizaron compras. Los registros están geolocalizados con la dirección de residencia reportada de los clientes y, para preservar la privacidad, se aplica la técnica de agrupación geográfica y creación de supertipos explicada en el numeral 4.2.2.

Una consulta en Impala SQL con base en el modelo de datos del numeral 4.1.4 para esta tabla sería:

```
select
    round(t2.latitud, 3) as lat,
    round(t2.longitud,3) as long,
    sum(t1.vlr_trx) as facturacion_total,
    count(*) as cantidad_trx,
    avg(t1.vlr_trx) as ticket_promedio,
    count(distinct t1.id) as cant_unica_cli
from Transaccion t1
inner join Cliente t2 on t1.id = t2.id
group by 1,2;
```

6.1.3. Información por categorías de la demanda

Al igual que la tabla descrita en el numeral anterior, esta tabla contiene la información de las transacciones realizadas por las personas de una zona geográfica específica en cualquier establecimiento de comercio. La diferencia radica en que la información se desglosa en los códigos de categoría mercante (MCC, por sus siglas en ingles).

Los códigos MCC permiten agrupar los establecimientos de comercio según el tipo de mercancía o servicio que ofrecen. A su vez, utilizarlos conserva la privacidad de los establecimientos ya que no es posible identificar ni individualizar un establecimiento con uno de estos códigos.

Una consulta en Impala SQL con base en el modelo de datos del numeral 4.1.4 para esta tabla sería:

```
select
    round(t2.latitud, 3) as lat,
    round(t2.longitud,3) as long,
    t4.cod_mcc,
    sum(t1.vlr_trx) as facturacion_total,
    count(*) as cantidad_trx,
    avg(t1.vlr_trx) as ticket_promedio,
    count(distinct t1.id) as cant_unica_cli
from Transaccion t1
inner join Cliente t2 on t1.id = t2.id
inner join Establecimiento t3 on t1.cod_estab = t3.cod_estab
inner join MCC t4 on t3.cod_mcc = t4.codigo
group by 1,2,3;
```

6.1.4. Información demográfica de la oferta

Esta tabla contiene la información demográfica de las personas que realizaron compras durante el periodo de estudio en los establecimientos de comercio ubicados en la zona de interés. Incluye variables como ocupación, género, rango de edad y rango de ingresos. Los registros están geolocalizados con la dirección proporcionada por los establecimientos de comercio y, para preservar la privacidad, se aplica la técnica de agrupación geográfica y creación de supertipos explicada en el numeral 4.2.2.

Una consulta en Impala SQL con base en el modelo de datos del numeral 4.1.4 para esta tabla sería:

```
select
    round(t3.latitud,3) as lat,
    round(t3.longitud,3) as long,
    t1.ocupacion,
    t1.sexo as genero,
    case
        when t1.edad < 17 then '< 17 Años'
        when t1.edad between 17 and 24 then '17 a 24 Años'
        when t1.edad between 25 and 36 then '25 a 36 Años'
        when t1.edad between 37 and 49 then '37 a 49 Años'
        when t1.edad between 50 and 64 then '50 a 64 Años'
```

```
        when t1.edad >= 65 then '> 65 Años'
    end as rango_edad,
    case
        when t1.ingresos/828116 < 1 then '< 1 SMMLV'
        when t1.ingresos/828116 between 1 and 2 then '1 a 2 SMMLV'
        when t1.ingresos/828116 between 2 and 6 then '2 a 6 SMMLV'
        when t1.ingresos/828116 between 6 and 10 then '6 a 10 SMMLV'
        when t1.ingresos/828116 > 10 then '> 10 SMMLV'
    end as rango_ingresos,
    count(distinct t2.id) as cant_cli
from Cliente t1
inner join Transaccion t2 on t1.id = t2.id and t1.cod_tip_doc = 1
inner join Establecimiento t3 on t2.cod_estab = t3.cod_estab
where t2.f_trx between 20190101 and 20191231
group by 1,2,3,4,5,6;
```

6.1.5. Información general de la oferta

Esta tabla contiene la información de las transacciones realizadas por las personas en establecimientos de una zona geográfica específica. Contiene indicadores como la facturación total de los clientes en la zona, la cantidad total de compras realizadas por clientes en la zona, el ticket promedio por compra de los clientes de la zona y la cantidad única de clientes de la zona que realizaron compras. Los registros están geolocalizados con la dirección proporcionada por los establecimientos de comercio y, para preservar la privacidad, se aplica la técnica de agrupación geográfica y creación de supertipos explicada en el numeral 4.2.2.

Una consulta en Impala SQL con base en el modelo de datos del numeral 4.1.4 para esta tabla sería:

```
select
    round(t2.latitud, 3) as lat,
    round(t2.longitud,3) as long,
    sum(t1.vlr_trx) as facturacion_total,
    count(*) as cantidad_trx,
    avg(t1.vlr_trx) as ticket_promedio,
    count(distinct t1.id) as cant_unica_cli
from Transaccion t1
inner join Establecimiento t2 on t1.cod_estab = t2.cod_estab
group by 1,2;
```

6.1.6. Información por categorías de la oferta

Al igual que la tabla descrita en el numeral anterior, esta tabla contiene la información de las transacciones realizadas por las personas en

establecimientos dentro de una zona geográfica específica. La diferencia radica en que la información se desglosa en los códigos de categoría mercante (MCC, por sus siglas en ingles).

Los códigos MCC permiten agrupar los establecimientos de comercio según el tipo de mercancía o servicio que ofrecen. A su vez, utilizar estos códigos conserva la privacidad de los establecimientos ya que no es posible identificar ni individualizar un establecimiento con uno de estos códigos.

Una consulta en Impala SQL con base en el modelo de datos del numeral 4.1.4 para esta tabla sería:

```
select
    round(t2.latitud, 3) as lat,
    round(t2.longitud,3) as long,
    t3.codigo as cod_mcc,
    sum(t1.vlr_trx) as facturacion_total,
    count(*) as cantidad_trx,
    avg(t1.vlr_trx) as ticket_promedio,
    count(distinct t1.id) as cant_unica_cli
from Transaccion t1
inner join Establecimiento t2 on t1.cod_estab = t2.cod_estab
inner join MCC t3 on t2.cod_mcc = t3.codigo
group by 1,2,3;
```

6.2. Desarrollo del prototipo utilizando el lenguaje R

Para el desarrollo del prototipo de la herramienta que modela espacialmente la oferta y la demanda se utiliza el lenguaje de programación R debido a: primero, R es un lenguaje de programación gratuito que ofrece computación estadística y realización de gráficos; segundo, R compila y se ejecuta en una gran variedad de plataformas UNIX, así como Windows y MacOS; tercero, es R es un lenguaje utilizado en la entidad financiera aportante de los datos para este trabajo; y cuarto, la entidad financiera cuenta con un servidor web local (es decir que solo se puede acceder mientras se encuentra conectado directamente a la red interna de la entidad) que utiliza el paquete Shiny de R, el cual permite crear aplicaciones que se pueden consultar a través de la web. Esta última razón es importante ya que permite que personas que no tengan conocimiento en programación puedan interactuar con la herramienta fácilmente y sin necesidad de instalar software adicional.

6.2.1. Interfaz de usuario

Para la interfaz de usuario se desarrolla un tablero con pestañas de visualización: una para modelar la oferta y otra para modelar la demanda. El tablero se

compone de 4 elementos principales, un menú, un mapa, unos indicadores generales y unos indicadores por categorías.

El menú consta de los botones que indican en que visualización se encuentra (oferta, demanda o demás visualizaciones que se deseen agregar). También cuenta con 2 campos en donde se puede indicar manualmente la latitud y la longitud de la ubicación a analizar. Además cuenta con un control (slider) que indica el radio de la zona de interés que se desea modelar. Por último esta zona es óptima para agregar logos de la entidad financiera, instrucciones, advertencias, marcas de copyright o demás elementos que se deseen. En la Figura 15 se muestra el menú desarrollado en el prototipo.

Figura 15. Menú del prototipo. Fuente: elaboración propia.

El mapa consiste en una visualización interactiva de la geografía del territorio colombiano. También tiene un indicador que muestra el punto central de la zona de interés y una circunferencia con el radio de medición establecido en el control del menú. Adicional a esto se crea un botón “Calcular” que toma la posición actual del indicador y el radio de medición establecido y genera el análisis para esa zona de interés. En la Figura 16 se observa el mapa desarrollado para el prototipo.

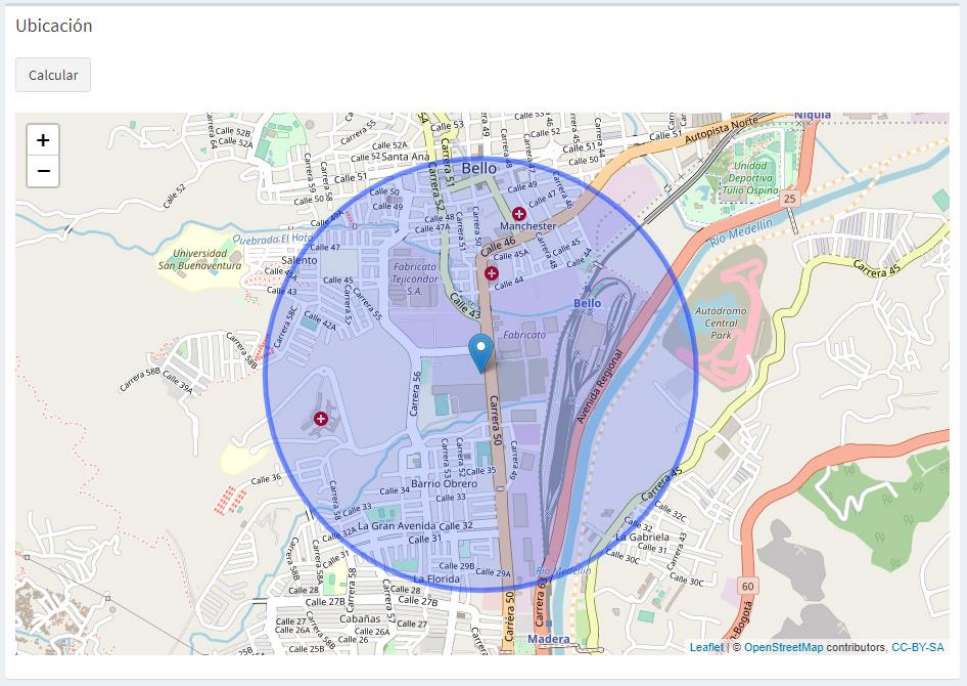


Figura 16. Mapa del prototipo. Fuente: elaboración propia.

Los indicadores generales son objetos de visualización que permiten observar los datos de la zona de interés. Para este prototipo se dividen en 2: valores generales de la zona y unos gráficos de la información demográfica. En la Figura 17 se muestran, en la parte de arriba, cuatro valores generales de la zona de interés: valor de compras totales, ticket promedio, cantidad de compras y cantidad de clientes. En la parte derecha, junto al mapa, se encuentran las gráficas desarrolladas para ver la información demográfica.

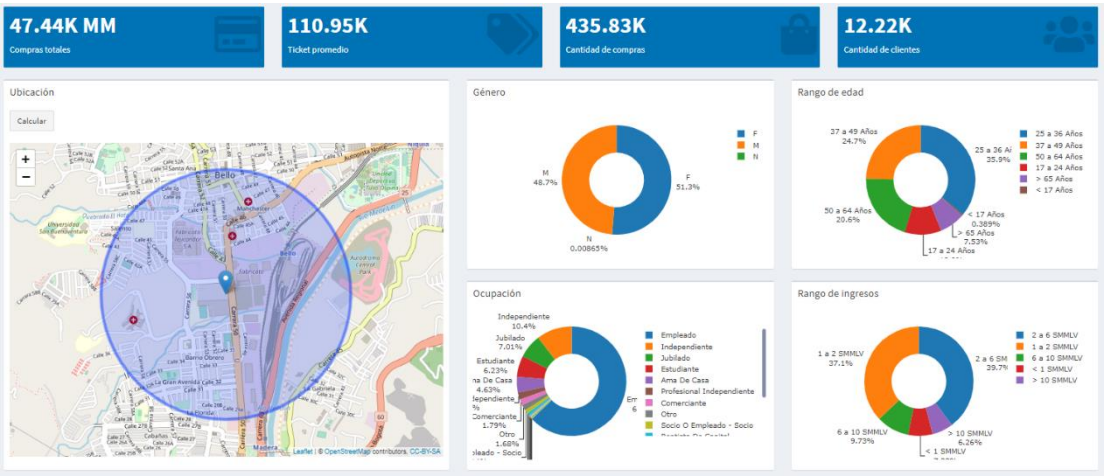


Figura 17. Indicadores generales del prototipo. Fuente: elaboración propia.

Por último, los indicadores por categorías desarrollados en el prototipo constan de una tabla donde se presentan en detalle las categorías de comercio con su participación del total, facturación total, cantidad de compras, cantidad de clientes y ticket promedio, y un diagrama de árbol (treemap) donde visualmente se representan todas las categorías por su facturación total. En la Figura 18 se observa en detalle los indicadores por categorías desarrollados en el prototipo y finalmente en la Figura 19 la interfaz de usuario completa del prototipo.

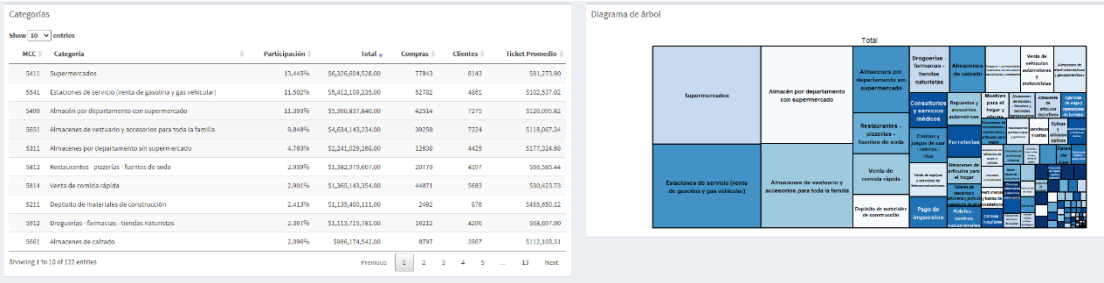


Figura 18. Indicadores por categorías del prototipo. Fuente: elaboración propia.

El uso del prototipo se describe en el siguiente proceso: primero, se selecciona el centro de la zona de interés, ya sea utilizando el mapa y dando clic en el punto o ingresando manualmente los valores de latitud y longitud; segundo, se selecciona el radio de medición de la zona de interés utilizando la barra de control (slider) del menú; por último, se da clic en el botón “Calcular” y se espera a que el tablero recalcule y modele los resultados para la nueva zona.

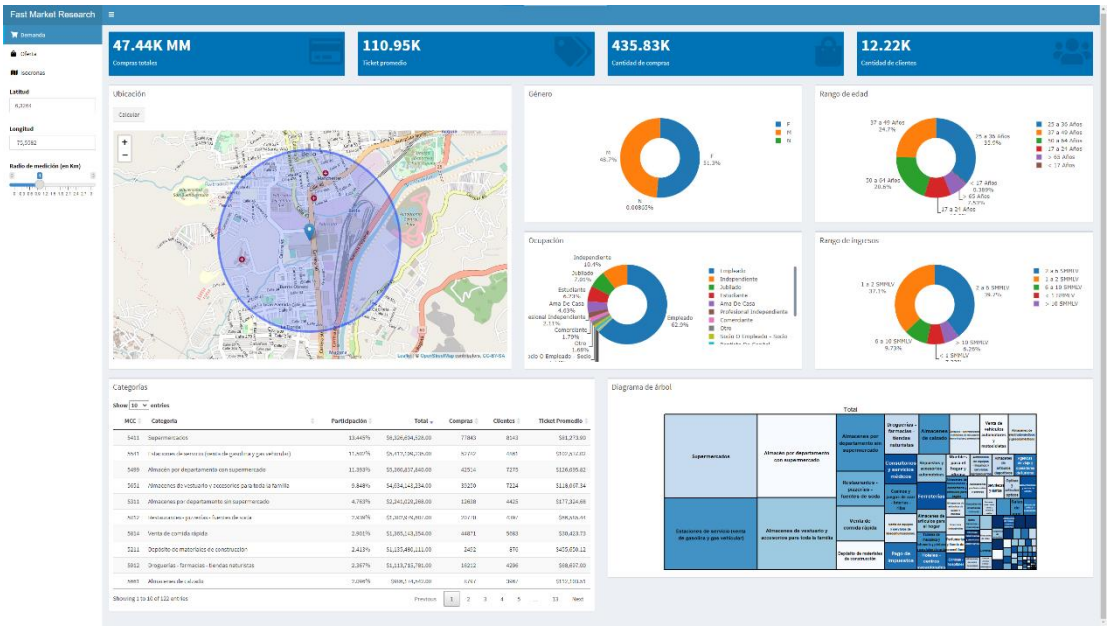


Figura 19. Interfaz de usuario del prototipo. Fuente: elaboración propia.

Este prototipo fue desarrollado en conjunto con personas interesadas pertenecientes a la entidad financiera aportante de los datos, debido a esto se hizo la elección de los indicadores y visualizaciones que lo componen. Sin embargo, el modelo de datos previamente discutido permite hacer muchos más análisis y visualizaciones que las presentadas actualmente. Este prototipo es una guía y puede ser modificado según el objetivo de la entidad financiera que lo implemente.

6.2.2. Servidor y funcionamiento interno

En el paquete Shiny de R utilizado para el prototipo, se le conoce como servidor al script donde se especifica el funcionamiento interno de la aplicación. Es acá donde se especifica el origen de los datos, el comportamiento de la interfaz de usuario, los cálculos y los resultados que arroja el modelado de la oferta y la demanda.

Para el prototipo, utilizamos las consultas explicadas en los numerales 6.1.1 al 6.1.6, generando 6 archivos que contienen los datos de los supertipos en formato CSV. Luego estos datos son cargados en el ambiente de R donde pueden ser procesados.

Cuando el usuario da clic en el botón “Calcular” en la interfaz de usuario, el servidor toma los valores de latitud, longitud y radio de medición para modelar

la oferta o la demanda, según la vista en la que se encuentre el usuario. Luego se procede a filtrar la información que no pertenece a la zona de interés.

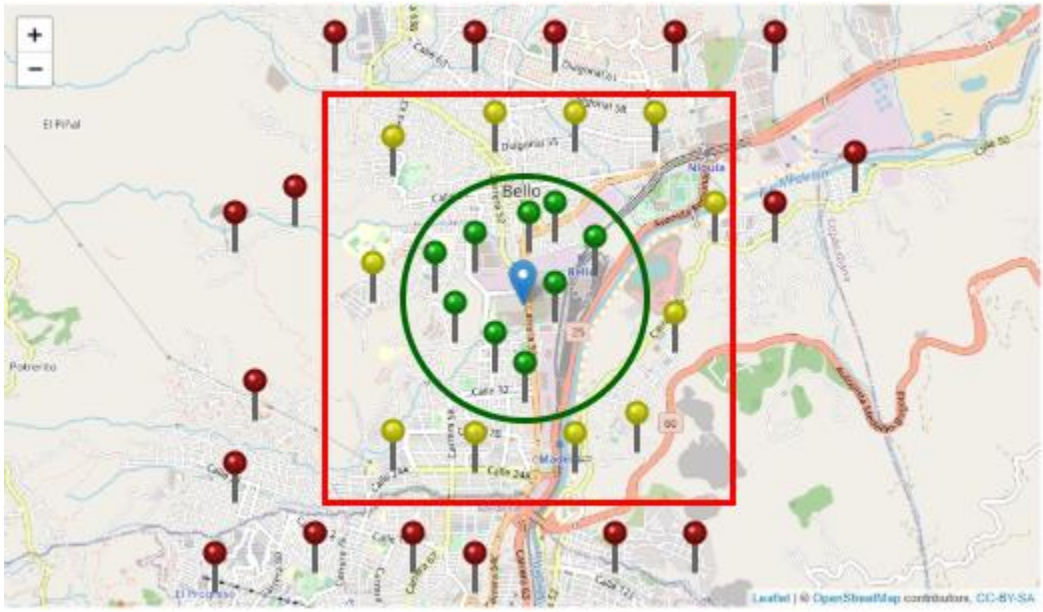


Figura 20. Proceso de filtrado de información de la zona de interés. Fuente: elaboración propia.

Con el ánimo de optimizar el proceso, el filtrado de la información de la zona de interés consiste en 2 pasos:

- Primero se eliminan los puntos que se encuentren por fuera del cuadrado generado por $latitud \pm delta$ y $longitud \pm delta$ del centro de nuestra zona de interés. Este valor $delta$ debe asegurar que la distancia de cualquiera de los supertipos por fuera del cuadrado sea mayor al valor máximo posible del radio de medición. Gráficamente en la Figura 20, este paso se representa con el cuadrado rojo y automáticamente los supertipos representados por pines rojos quedan eliminados. Así se evita calcular la distancia con supertipos que se encuentran muy lejos.

- Segundo, para cada supertipo restante, se calcula la distancia lineal con el centro de nuestra zona de interés para luego descartar los supertipos cuya distancia es mayor al radio de medición. Gráficamente en la Figura 20, se calcula la distancia desde el centro de la zona de interés a cada uno de los supertipos dentro del rectángulo rojo, representados por pines amarillos y verdes, luego se descartan los supertipos representados por pines amarillos debido a que su distancia es mayor al radio de medición y como resultado quedan los supertipos que pertenecen a la zona

de interés representados con pines verdes. En consecuencia nuestra zona de interés queda representada con el círculo verde.

Una vez se tienen los supertipos que pertenecen a la zona de interés, se agrega la información por cada tipo y se generan los indicadores de la interfaz de usuario:

- Para el indicador de compras totales se agrega la información de las consultas 6.1.2. para la demanda o 6.1.5. para la oferta utilizando la función SUMA en la variable *facturación_total*.

- Para el indicador de ticket promedio se agrega la información de las consultas 6.1.2. para la demanda o 6.1.5. para la oferta utilizando la función PROMEDIO en la variable *ticket_promedio*.

- Para el indicador de cantidad de compras se agrega la información de las consultas 6.1.2. para la demanda o 6.1.5. para la oferta utilizando la función SUMA en la variable *cantidad_trx*.

- Para el indicador de cantidad de clientes se agrega la información de las consultas 6.1.2. para la demanda o 6.1.5. para la oferta utilizando la función SUMA en la variable *cantidad_unica_cli*.

- Para las gráficas de género, rango de edad, ocupación y rango de ingresos, se agrega la información de las consultas 6.1.1. para la demanda o 6.1.4. para la oferta utilizando la función SUMA en la variable *cant_cli* y agrupando por las variables *género, ocupación, rango_ingresos y rango_edad*. Cada gráfica es el resultado de comparar la variable de grupo con la sumatoria de la cantidad de clientes.

- Para la tabla de categorías y el diagrama de árbol (treemap) se agrega la información de las consultas 6.1.3. para la demanda o 6.1.6. para la oferta utilizando la función SUMA en la variable *cant_cli* y agrupando por la variable *cod_mcc*

Otra de las ventajas que ofrece el servidor es que permite crear funciones específicas para tratar los datos. Por ejemplo, en el prototipo se quería abreviar los valores de los indicadores para que fueran más fáciles de leer y para esto se creó una función que recibe el número que se desea mostrar y lo abrevia como se muestra en la Tabla 4.

Tabla 4. Abreviatura de indicadores. Fuente: elaboración propia.

Valor original	Valor abreviado
1.000	1 K
1.000.000	1 MM
1.000.000.000	1 K MM
1.000.000.000.000	1 B

Para publicar este prototipo, se crea un paquete que puede ser cargado en un servidor web de Shiny R y que contiene los elementos necesarios para su funcionamiento, mantenimiento y actualización.

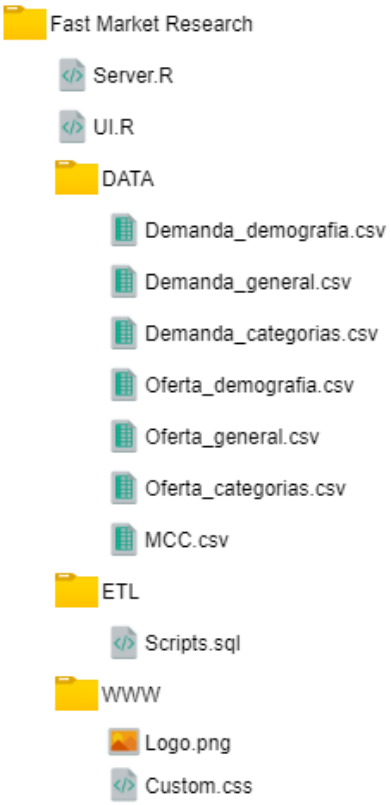


Figura 21. Estructura del prototipo cargada en el servidor web. Fuente: Elaboración propia.

En la Figura 21 se ilustra la estructura del paquete creado. La entidad que proporciona los datos publicó el prototipo dentro de su red privada bajo el nombre de "*Fast Market Research*" y este se puede acceder mediante un link colgado en la intranet de las personas interesadas. El paquete consiste en dos archivos *Server.R* y *UI.R* que contienen toda la programación del prototipo. También tiene una carpeta *DATA* que contiene todas las bases de datos en formato CSV que alimentan la herramienta. Además existe una carpeta *ETL* que contiene las consultas en SQL que general la información de la carpeta

anterior y por último una carpeta *WWW* que contiene todos los elementos gráficos y de estilo que se utilicen en la interfaz de usuario del prototipo.

Adicional a la publicación en un servidor web privado de la entidad financiera, la herramienta se publica en un ambiente colaborativo para que pueda ser compartido y desarrollado por el personal analítico de la entidad. En este caso, se cargó a un servidor de GitLab, también de carácter privado y de dominio de la entidad financiera, como se aprecia en la Figura 22.

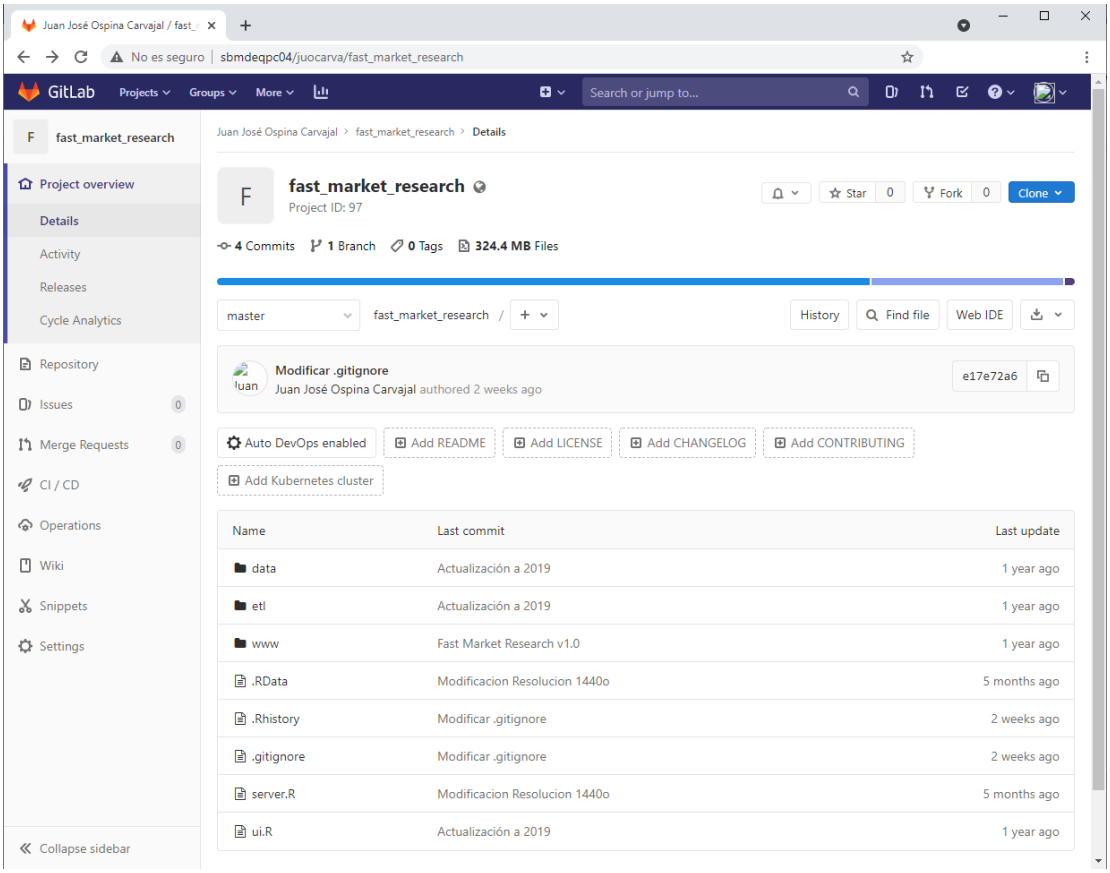


Figura 22. Carga del prototipo en el servidor GitLab de la entidad financiera. Fuente: Elaboración propia.

7. Casos de uso del prototipo

Para mostrar el uso del prototipo, a continuación se presentan 3 casos de uso de situaciones cotidianas para la entidad financiera.

7.1. Identificación de mercados naturales

Una de las tareas dentro de los equipos comerciales consiste en la búsqueda de oportunidades de negocio para fortalecer las relaciones con los clientes. Es entonces cuando surgen los mercados naturales, definidos como los sitios específicos que cuentan con un nivel alto de demanda pero que la oferta de la zona no la satisface.

Como ejemplo tomaremos la ubicación definida con las coordenadas *latitud = 6,3264*, *longitud = -75,5582* y *radio medición = 1 Km*. Esta zona se ubica en el municipio de Bello, Antioquia y comprende los barrios Obrero, Gran avenida, Manchester, Madera, La Florida y Santa Ana.

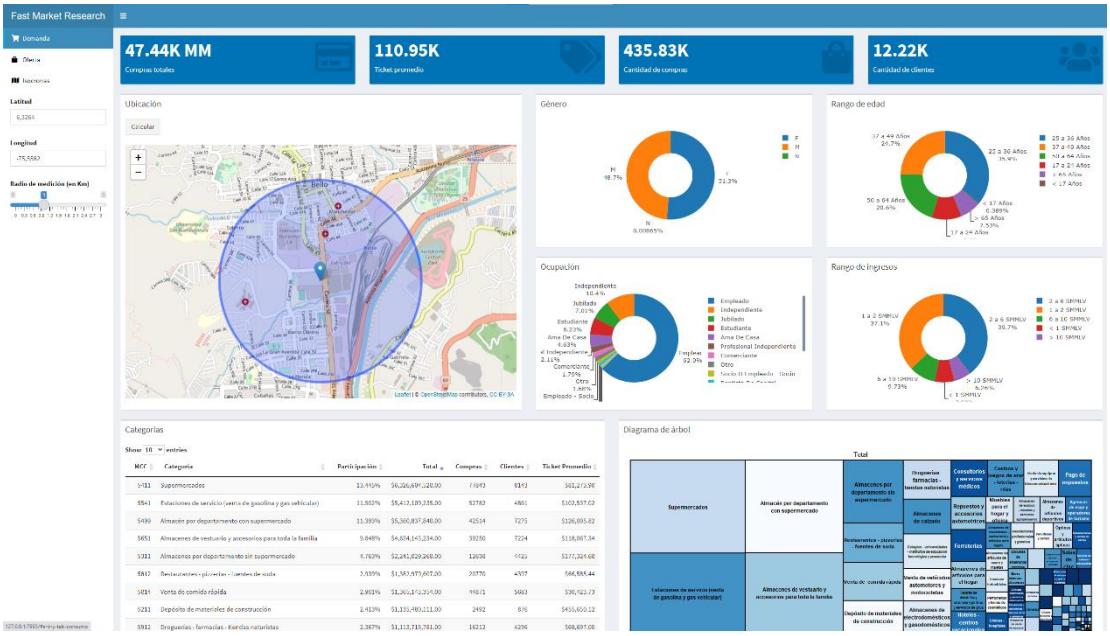


Figura 23. Modelado de la demanda. Caso de uso 1. Fuente: Elaboración propia

Analizando el modelado de la demanda, mostrado en la Figura 23, se observa que la población que vive en la zona realizó un total de compras por valor 47,4 mil millones de COP (pesos colombianos) en una cantidad de 435.830 mil compras con un ticket promedio de 110.950 COP. También se puede observar que la gran mayoría de la población de la zona se encuentra entre los rangos de ingresos de 1 a 6 salarios mínimos mensuales legales vigentes (SMMLV) donde el 37,1% se encuentra en el rango de 1 a 2 SMMLV y el 39,7% se encuentra en el rango de 2 a 6 SMMLV. Además se observa que el 62,9% de la población son empleados. En la información por categorías, las principales categorías son *Supermercados* con un total de compras por valor de 6,32 mil

millones de COP, seguido por *Estaciones de servicio* con 5,41 mil millones de COP y *Almacén por departamento con supermercado* con 5,36 mil millones de COP.

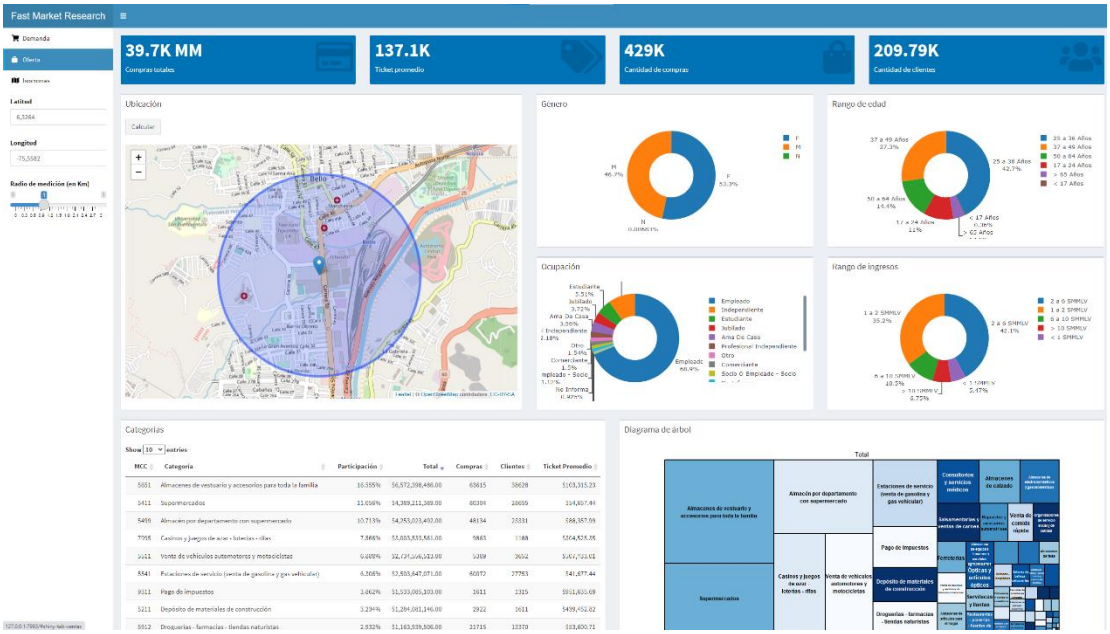


Figura 24. Modelado de la oferta. Caso de uso 1. Fuente: Elaboración propia.

Luego, analizando el modelado de la oferta, mostrado en la Figura 24, se observa que los comercios de la zona tuvieron compras por valor de 37,7 mil millones de COP en un total de 429.000 compras recibidas y un ticket promedio de 137.100 COP. También se puede observar que la gran mayoría de la población que realizó compras en la zona se encuentra entre los rangos de ingresos de 1 a 6 salarios mínimos mensuales legales vigentes (SMMLV) donde el 35,2% se encuentra en el rango de 1 a 2 SMMLV y el 42,1% se encuentra en el rango de 2 a 6 SMMLV. Además se observa que el 68,9% de la población que realizó compras en la zona fueron empleados. En la información por categorías, las principales categorías que recibieron compras fueron *Almacenes de vestuario y accesorios para toda la familia* con un total de compras por valor de 6,57 mil millones de COP, seguido por *Supermercados* con 4,38 mil millones de COP y *Almacén por departamento con supermercado* con 4,25 mil millones de COP.

Comparando los resultados de la demanda y la oferta para la zona se evidencia que hay un déficit de 7,74 mil millones de COP. Es decir, las personas que viven dentro del área analizada realizan desplazamientos fuera de esta para realizar compras. Si analizamos el panorama por categorías, se observa que

la demanda de las personas en la zona para establecimientos en *Supermercados* y *Almacén por departamento con supermercado* es de 6.326.604.528,00 COP y 5.360.837.840,00 COP respectivamente, mientras que la oferta de la zona para las mismas categorías es de 4.389.211.389,00 COP y 4.253.023.492,00 respectivamente. Esto quiere decir que, en conjunto, los establecimientos de la zona pertenecientes a estas categorías están dejando de recibir 3.045.207.487,00 COP de personas que viven en la zona.

La identificación de estos mercados naturales ayuda a los equipos comerciales de la entidad financiera a tener mejores conversaciones con los clientes, presentándole oportunidades de crecimiento y profundización de la relación cliente – entidad financiera con el objetivo de tener un beneficio mutuo y sostenible en el tiempo.

7.2. Asesoría a clientes para la toma de decisiones

Las entidades financieras quieren ser más que un banco para sus clientes y buscan ser aliados estratégicos para sus clientes. Una de las maneras en que las entidades se vuelven aliados estratégicos es en el asesoramiento para la toma de decisiones. Para ejemplificar este caso de uso se utilizará una situación hipotética: un cliente dedicado al comercio minorista y tiene establecimientos en las categorías de *Supermercados* y *Almacén por departamento con supermercado* en diferentes ciudades principales del territorio colombiano. Como parte de su proceso de expansión, desea abrir un nuevo establecimiento en la ciudad de Medellín. La entidad financiera al ser uno de los aliados financieros principales del cliente ofrece su conocimiento de la oferta y la demanda para apalancar la decisión de dónde abrir su nuevo establecimiento y ser el financiador de este proyecto.

El cliente de comercio minorista tiene preseleccionadas 3 ubicaciones para su nuevo establecimiento. La primera ubicación se encuentra en el barrio Laureles, con *latitud = 6,241020* y *longitud = -75,597705*. La segunda ubicación se encuentra en el barrio Guayabal, con *latitud = 6,199925* y *longitud = -75,586597*. Por último, la tercera ubicación se encuentra en el barrio Buenos Aires con *latitud = 6,242118* y *longitud = -75,557439*. Para analizar la oferta y la demanda de las zonas de interés se selecciona un radio de medición igual para todas. Además, debido a que el cliente tiene locales de dos categorías de comercio (*Supermercados* y *Almacén por departamento con supermercado*) se sumarán estos valores para buscar mercados naturales en estas ubicaciones. En la Tabla 5 se observan los valores obtenidos luego de modelar la oferta y la demanda luego de ser escalados utilizando como valor de referencia la demanda e la ubicación Guayabal.

Tabla 5. Resultados del modelado de oferta y demanda para las zonas preseleccionadas. Fuente: Elaboración propia

Ubicación	**Demanda**	**Oferta**	**Diferencia (Demanda – Oferta)**
Laureles	$3{,}6843\ X$	$4{,}3176\ X$	$-0{,}6342\ X$
Guayabal	X	$0{,}3614\ X$	$0{,}6386\ X$
Buenos Aires	$2{,}0433\ X$	$1{,}7502\ X$	$0{,}2931\ X$

Luego de un primer análisis se puede concluir que la ubicación de Laureles, a pesar la de mayor demanda con $3{,}6843\ X$, cuenta también con mucha competencia ya que la oferta la sobrepasa dando como resultado una diferencia de $-0{,}6342\ X$. Este valor negativo se interpreta como que esta zona atrae a clientes que residen por fuera de ella a realizar sus compras.

Teniendo en cuenta los resultados presentados en la Tabla 5 y que el objetivo del cliente es expandir su negocio con la mayor probabilidad de éxito, la recomendación que daría la entidad financiera sería optar por las ubicaciones de Guayabal o Buenos Aires, las cuales cuentan con una diferencia positiva de $0{,}6386\ X$ y $0{,}2931\ X$ respectivamente. Sin embargo, como el cliente cuenta con 2 tipos de establecimientos de comercio, se puede analizar a mayor profundidad la oferta y la demanda para cada una de estos tipos y ubicaciones.

Tabla 6. Demanda y oferta por categorías en zona Guayabal. Fuente: Elaboración propia.

	Guayabal		
	Demanda	**Oferta**	**Diferencia**
Supermercados	$0{,}5613\ X$	$0{,}3612\ X$	$0{,}2001\ X$
Almacén por departamento con supermercado	$0{,}4386\ X$	$0{,}0002\ X$	$0{,}4384\ X$

Tabla 7. Demanda y oferta por categorías en zona Buenos Aires. Fuente: Elaboración propia.

	Buenos Aires		
	Demanda	**Oferta**	**Diferencia**
Supermercados	$1{,}2260\ X$	$1{,}7502\ X$	$-0{,}5242\ X$
Almacén por departamento con supermercado	$0{,}8163\ X$	$Sin\ datos$	$0{,}8163\ X$

Observando los resultados de la Tabla 6 y Tabla 7, los cuales fueron escalados utilizando el mismo valor de la Tabla 5, se observa un panorama más claro según la categoría de establecimiento con la que el cliente de comercio minorista desee expandirse. Si la categoría es *Supermercados* la recomendación sería elegir la ubicación de Guayabal debido a que cuenta con una diferencia positiva de $0{,}2001\ X$, sin embargo si la categoría es *Almacén por departamento*

con supermercado la recomendación se inclinaría por la ubicación de Buenos Aires, debido a que la diferencia positiva es de $0{,}8163\ X$.

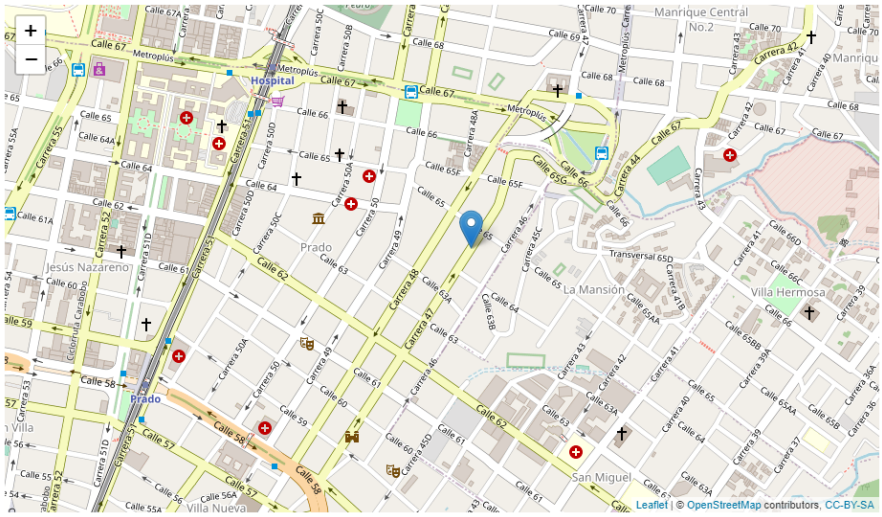
Analizar la demanda y la oferta de las ubicaciones preseleccionadas para los planes de expansión de un cliente permite recomendar no solo la ubicación con mayores posibilidades de éxito sino también que tipo de establecimientos son mejores para la expansión y de esta manera fortalecer la relación cliente – entidad financiera.

7.3. Aporte de información para la toma de decisiones frente a grandes proyectos

Es común en las entidades financieras que a medida que se requiera una participación mayor en un proyecto, las condiciones de aprobación aumentan y pueden pasar a ser revisadas por comités de crédito especializados que incluyen gerentes de zona, vicepresidentes, presidentes e incluso la misma junta directiva de la entidad. Es entonces crítico para las entidades financieras contar con la mayor cantidad de información para ayudar a estos comités a tomar las mejores decisiones.

Para ejemplificar este caso de uso, nuevamente se utilizará una situación hipotética: un cliente constructor presenta un megaproyecto donde busca una financiación de 600 mil millones de COP para la construcción de un centro comercial en la ciudad de Medellín y su posterior operación, es decir que el cliente constructor plantea mantener todo el espacio construido para su posterior arrendamiento. Además, el cliente indica que el centro comercial tendrá un área total de 200.000 m^2, distribuidos en su mayoría en áreas comerciales. También el área a construir incluye consultorios médicos, oficinas y parqueaderos. El público objetivo del centro comercial consta de personas con alto nivel adquisitivo, es decir personas con ingresos superiores a los 10 SMMLV, ya que el cliente planea poblar las áreas comerciales con tiendas exclusivas que comparten este mismo nicho de mercado. La ubicación elegida por el cliente constructor para el desarrollo de este proyecto es el barrio Prado, exactamente en las coordenadas latitud = 6,259929 y longitud = -75,558832. En la Tabla 8 se muestran los principales datos de la propuesta realizada por el cliente constructor.

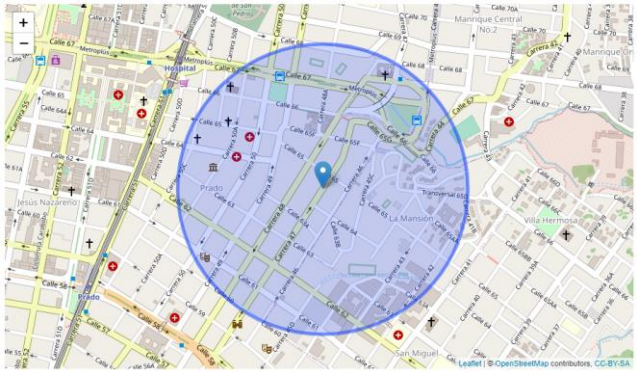
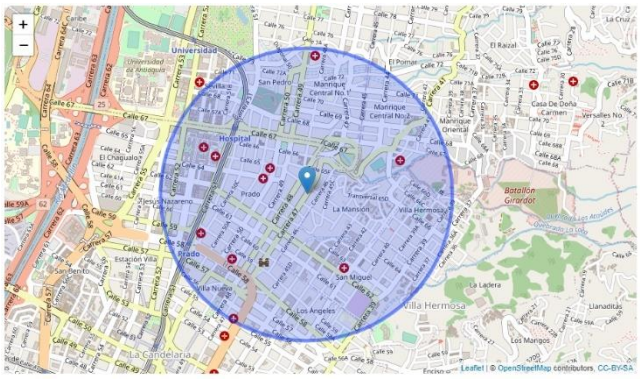
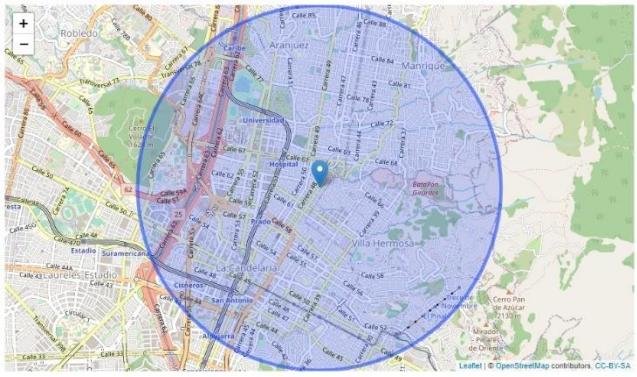
Tabla 8. Detalles del megaproyecto presentado por el cliente constructor. Fuente: Elaboración propia.

Ubicación	Barrio Prado, Medellín, Colombia.
Coordenadas	Latitud = 6,259929 Longitud = -75,558832
Vista aérea	
Financiación solicitada	600.000.000.000 COP
Público objetivo	Personas con alto poder adquisitivo (Ingresos superiores a 10 SMMLV)
Área por construir	200.000 m^2

Debido a que el monto solicitado por el cliente constructor es muy grande, la entidad financiera decide evaluar el megaproyecto ante un comité de crédito especializado para tomar la decisión de aceptar o rechazar la solicitud.

Para tener más información, la entidad financiera utiliza la herramienta de modelado de oferta y demanda y analiza la ubicación proporcionada por el cliente a diferentes radios de medición: 500 m, 1 Km y 2,5 Km en donde se hará enfoque en las características demográficas de la población para saber si la ubicación elegida es la óptima para el público objetivo que tiene el centro comercial.

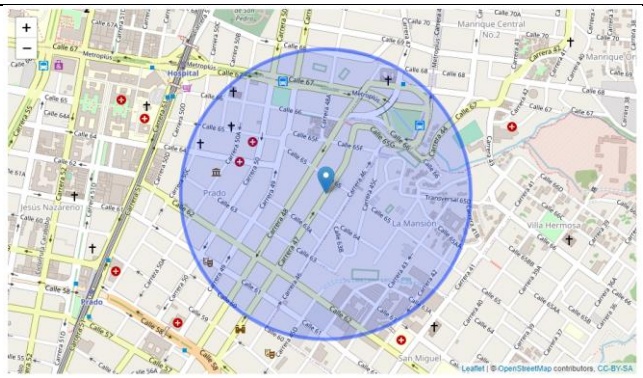
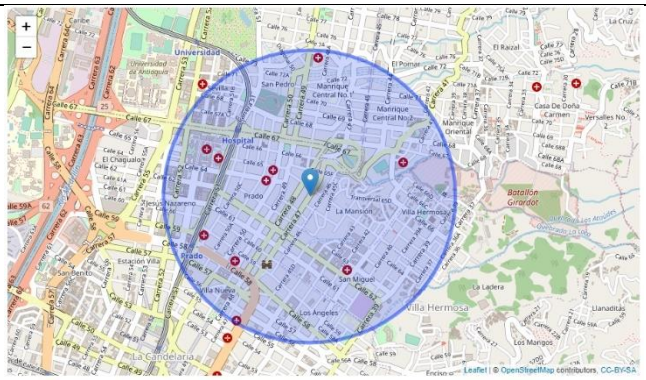
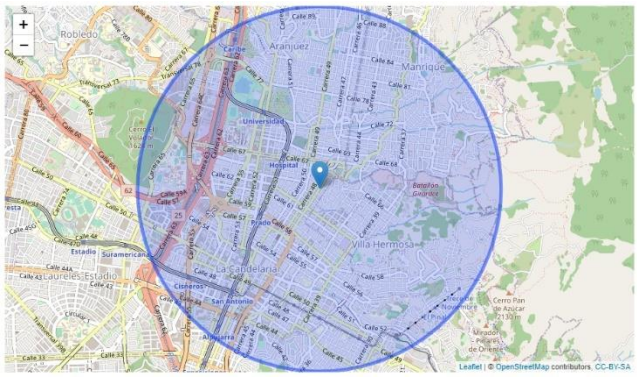
Tabla 9. Análisis de la demanda de la zona de interés a diferentes radios. Fuente: Elaboración propia.

Demanda			
Radio de medición	**Ubicación**	**Población 10+ SMMLV**	**Población 6 a 10 SMMLV**
500 m		5,85%	9,54%
1 km		6,00%	9,62%
2,5 km		6,07%	8,14%
	Promedio	**5,97%**	**9,10%**

Analizando la demanda de la zona, presentados en la Tabla 9, se observa que en promedio la zona solo cuenta con el 5,97% de la población con ingresos superiores a 10 SMMLV. Adicionalmente se observa también que la población del rango salarial anterior, es decir población con ingresos entre 6 y 10 SMMLV, en promedio solo representan un 9,10% de la zona. Estos datos

indican que en la zona no predominan las personas con alto poder adquisitivo en comparación con otras zonas de la ciudad, como por ejemplo el barrio El Poblado donde para las coordenadas *latitud = 6,197472*, *longitud = -75,565408* y con un radio de medición de 1 Km, el 36,1% de la población tiene ingresos superiores a 10 SMMLV.

Tabla 10. Análisis de la oferta de la zona de interés a diferentes radios. Fuente: Elaboración propia.

Oferta			
Radio de medición	**Ubicación**	**Población 10+ SMMLV**	**Población 6 a 10 SMMLV**
500 m		9,63%	11,40%
1 km		9,30%	12,30%
2,5 km		10,50%	13,30%
	Promedio	**9,81%**	**12,3%**

Para realizar un análisis completo, se modela la oferta de la zona a los diferentes radios de medición establecidos y se muestran los resultados en la Tabla 10. En promedio, en los establecimientos de la zona de interés sólo el 9,81% de la población tiene ingresos superiores a 10 SMMLV. La población perteneciente al rango salarial anterior, es decir población con ingresos entre 6 y 10 SMMLV, en promedio solo representan el 12,30%.

Al comparar los resultados de la oferta y demanda, a pesar de que la zona atrae a población con ingresos superiores a 10 SMMLV debido a que la oferta en promedio es mayor a la demanda, los niveles que muestran son muy inferiores comparados con otras zonas de la ciudad. Si miramos la misma ubicación en el barrio El Poblado (latitud = 6,197472, longitud = -75,565408 y radio de medición de 1 Km) el modelado de la oferta indica que el 32,8% de la población que realizó compras en los establecimientos de la zona reciben ingresos de más de 10 SMMLV.

Con esta información, el comité de crédito especializado puede tomar la decisión de rechazar la solicitud de crédito del cliente constructor ya que el público objetivo del megaproyecto no tiene gran representación dentro de la zona de interés. También, con el objetivo de ser un socio estratégico del cliente constructor, la entidad financiera podría asesorar el proyecto, ya sea sugiriendo un público objetivo acorde a la zona de interés seleccionada o buscando una zona de interés donde predomine el público objetivo definido del megaproyecto.

8. Consideraciones éticas

Las empresas e instituciones constantemente recolectan datos para ofrecer o mejorar sus servicios existentes, los cuales muchas de las veces requieren analizar, publicar o compartir información privada y confidencial. Cada vez son más los métodos propuestos que permiten la extracción de información conservando la privacidad (*Privacy Preserving Data Mining, PPDM*) [34].

Este crecimiento trae consigo un choque entre los puntos de vistas de tres interesados en el tema: los consumidores, las empresas y los entes regulatorios. En [35], los autores hablan de que existe una relación positiva entre los interesados cuando los consumidores comparten su información con las empresas, apoyados por entes regulatorios que se encargan de salvaguardar la información y que las empresas se preocupan por mantener confidenciales los datos, entendiendo del potencial que representan.

Esto lo que nos muestra es la fuerte tendencia hacia mantener la información de terceros de manera segura y tratando de mantener la privacidad. Además, en Colombia ya se regula y está vigente la ley de protección de datos

personales. En esta, se le otorgan “facultades al titular de los datos personales para exigir a los administradores de sus datos el acceso, inclusión, exclusión, corrección, adición, actualización y certificación de los datos, así como la limitación de las posibilidades de su divulgación, publicación, o cesión” [36].

Dentro del modelado de la oferta y la demanda toma relevancia mantener la privacidad y cuidar los datos sensibles, tanto de los clientes como de los establecimientos de comercio analizados, volviéndose pertinente la implementación de la agrupación geográfica y creación de supertipos propuesta en la sección 4.2.2.

Este trabajo garantizará en todas sus etapas la privacidad de los usuarios, la reserva bancaria y la propiedad intelectual de las entidades involucradas.

Referencias

[1] Asobancaria, «Big Data: aprovechando los datos, el nuevo gran activo,» *Semana Economica,* nº 1124, 12 Febrero 2018.

[2] U. Srivastava y S. Gopalkrishnan, «Impact of Big Data Analytics on Banking Sector: Learning for indian Banks,» *Procedia Computer Science,* vol. 50, pp. 643-652, 2015.

[3] G. Cao, Y. Duan y G. Li, «Linking business Analytics to decision making effectiveness: a path model analysis,» *IEEE Transactions on Engineering Management,* vol. 62, 2015.

[4] R. Deshpande y G. Zaltman, «Factors affecting the use of market research information: a path analysis,» *Journal of Marketing Research,* vol. 19, pp. 14-31, 1982.

[5] C. Moorman, G. Zaltman y R. Deshpande, «Relationships between providers and users of market research: the dynamics of trust within and between organizations,» *Journal of Marketing Research,* vol. 29, pp. 314-328, 1992.

[6] H. Muñoz Hernández, R. C. Osorio Mass y L. M. Zúñiga Pérez, «Inteligencia de negocios. Clave del éxito en la era de la información,» *Clío América,* vol. 10, pp. 194-211, 2016.

[7] C. A. Gutiérrez Barrera, Planes de expansión del comercio minorista: un acercamiento desde la analítica, grandes conjuntos de datos e información de transacciones financieras, Medellín, Antioquia: Universidad EAFIT, 2018.

[8] S. Nosratabadi, G. Pinter, A. Mosavi y S. Semperger, «Sustainable banking; Evaluation of European business models,» *Sustainability,* vol. 12, pp. 2314-2333, 2020.

[9] Asobancaria, «COVID-19, una primera aproximación de los efectos en las cuentas externas ante un nuevo panorama global y local,» *Semana económica (1226),* vol. 1226, 24 Marzo 2020.

[10] Asobancaria, «Tendencias de la banca para el 2021,» *Semana económica (1263),* 14 Diciembre 2020.

[11] G. Beltrán López, «La geolocalización social,» *Poligonos,* vol. 27, pp. 97-118, 2015.

[12] Google, «Sistemas de Geolocalización,» 2021. [En línea]. Available: https://www.google.com/maps/d/u/0/viewer?msa=0&ie=UTF8&ll=12.758118235683005%2C42.2652991078877&spn=35.231486%2C209.882813&t=m&source=embed&mid=13kxZTfA8HyYkBbNgwPsFtBvQbEU&z=3. [Último acceso: 18 Enero 2021].

[13] J. Ribeiro, F. Tânia, C. Soares y J. L. Borges, «Process discovery on geolocation data,» *Transportation Research Procedia,* vol. 47, pp. 139-146, 2020.

[14] M. Mansoori y I. Welch, «How do the find us? A study of geolocation tracking techniques of malicious web sites,» *Computers & Security,* vol. 97, p. 101948, 2020.

[15] J. McElroy, P. Remington, A. Trentham-Dietz, S. Robert y P. Newcomb, «Geocoding addresses from a large population-based study: lessons learned,» *Epidemiology,* vol. 14, pp. 399-407, 2003.

[16] J. Bow, N. Waters, P. Faris, J. Seidel, D. Galbraith, M. Knudtson y W. Ghali, «Accuracy of city postal code coordinates as a proxy for location

of residence,» *International Journal of Health Geographics,* vol. 3, pp. 5-14, 18 Marzo 2004.

[17] D. Zaim, A. Benomar y M. Bellafkih, «Developing a geomarketing solution,» *Procedia Computer Science,* vol. 148, pp. 353-360, 2019.

[18] A. Suhaibah, U. Uznir, A. A. Rahman, F. Anton y D. Mioc, «3D geomarketing segmentation: a higher spatial dimension planning perspective,» *Proceedings of the International Conference on Geomatic and Geospatial Technology (GGT),* vol. 42, pp. 1-7, 2016.

[19] A. Baviera-Puig, J. Buitrago-Vera y C. Escriba-Perez, «Geomarketing model in supermarket location strategies,» *Journal of Business Economics and Management,* vol. 17, pp. 1205-1221, 2016.

[20] J. D. Zuluaga y D. A. Escobar, «Geomarketing analysis for shopping mall in Manizales (Colombia). Accessibility approach methodology,» *Espacios,* vol. 38, p. 20, 2017.

[21] W. R. Meneses, «Evolución de las tecnologías transaccionales en el sector financiero en Colombia entre los años 2010 y 2015,» 2016.

[22] C. Arango Arango, H. Zárate Solano y N. Suarez Ariza, «Determinantes del acceso, uso y aceptación de pagos electrónicos en Colombia,» de *Ensayos sobre inclusión financiera*, 2017, pp. 287-358.

[23] G. Montoya, N. Rodríguez y W. Borja, «La estrategia interinstitucional para la reducción del uso del efectivo y la masificación de los medios de pago electrónicos,» de *Ensayos sobre inclusión financiera*, 2017, pp. 359-412.

[24] Asobancaria, «Balance y perspectivas crediticias 2020-2021 en medio de la pandemia,» *Semana económica (1265),* 25 Enero 2021.

[25] K. Imran Moin y Q. Baseer Ahmed, «Use of Data Mining in Banking,» *International Journal of Engineering Research and Application (IJERA),* vol. 2, nº 2, pp. 738-742, Mar-Apr 2012.

[26] K. Chitra y B. Subashini, «Data Mining Techniques and its Applications in Banking Sector,» *International Journal of Emerging Technology and Advanced Engineering,* vol. 3, nº 8, pp. 219-226, Agust 2013.

[27] E. Radmehr y M. Bazmara, «A Survey on Business Intelligence Solutions in Banking Industry and Big Data Applications,» *International Journey of Mechatronics, Electrical and Computer Technology (IJMEC),* vol. 7, nº 23, pp. 3280-3298, Jan 2017.

[28] Banco Bilbao Vizcaya Argentaria, S.A., «Llega Commerce360, una herramienta de tecnología 'big data', para ayudar al desarrollo de las pymes,» 27 Jun 2017. [En línea]. Available: https://www.bbva.com/es/bbva-suiza-lanza-new-gen-la-cuenta-digital-que-facilita-las-inversiones-en-los-sectores-con-mayor-impacto-en-el-futuro/. [Último acceso: 12 Agosto 2021].

[29] Grupo Bancolombia S.A., «¿Qué es Plink? Conoce como funciona y puede ayudarte,» 2021. [En línea]. Available: https://plink.grupobancolombia.com/plink/que-es-plink. [Último acceso: 12 Agosto 2021].

[30] Grupo Bancolombia S.A., «¿Cómo usar Plink? Aprende aquí a usar esta poderosa herramienta,» 2021. [En línea]. Available: https://plink.grupobancolombia.com/plink/como-usar-plink. [Último acceso: 12 Agosto 2021].

[31] International Standard Organization, «ISO 18245. Retail financial services — Merchant category codes,» 2003.

[32] IBM Corporation, «Analytics Solutions Unified Method,» 1 Marzo 2016. [En línea]. Available: ftp://ftp.software.ibm.com/software/data/sw-library/services/ASUM.pdf. [Último acceso: 2 Junio 2020].

[33] C. Ponsard, M. Touzani y A. Majchrowski, «Combining process guidance and industrial feedback for successfully deploying big data projects,» *Open Journal of Big Data (OJBD),* vol. 3, nº 1, pp. 26-41, 2017.

[34] R. Mendes, V. Vilela y J. P., «Privacy preserving data mining; Methods, metrics and applications,» *IEEE Access,* vol. 5, 2017.

[35] K. D. Martin, J. J. Kim, R. W. Palmatier, L. Steinhoff, D. W. Stewart, B. A. Walker y Y. W. S. K. Wang, «Data privacy in retail,» *Journal of Retailing,* vol. 30, 2020.

[36] Mineducación, «Política de tratamiento de datos personales,» 2020. [En línea]. Available: https://www.mineducacion.gov.co/1759/articles-353715_recurso_5.pdf. [Último acceso: 20 Enero 2021].