UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

PROGRAMA DE MAESTRÍA Y DOCTORADO EN INGENIERÍA

INGENIERÍA EN SISTEMAS – INVESTIGACIÓN DE OPERACIONES

LOCALIZACIÓN DE SUCURSALES BANCARIAS UTILIZANDO MODELOS DE APRENDIZAJE SUPERVISADO PARA UNA MICROFINANCIERA MEXICANA

TESIS PARA OPTAR POR EL GRADO DE:

MAESTRA EN INGENIERÍA EN SISTEMAS

PRESENTA:

Act. Karina Edith Velázquez Gómez

TUTOR:

Dra. Mayra Elizondo Cortés
Programa de Maestría y Doctorado en Ingeniería

CIUDAD DE MÉXICO, JUNIO 2019

JURADO:

Presidente:	Dr. Sánchez Guerrero Gabriel D.
Secretario:	Dr. Guillen Burguete Servio Tulio
1er Vocal:	Dra. Elizondo Cortés Mayra
2do Vocal:	Dra. González Videgaray María Del Carmen
3er Vocal:	M. en I. Rodríguez Rubio Jorge

TUTORA DE TESIS:
DRA. MAYRA ELIZONDO CORTÉS

_____

FIRMA

Dedicatorias

A mi padre amado Dios

Por estar conmigo en cada etapa de mi vida y no soltarme de su mano, por permitirme amarlo y vivir junto a él cada día. Por protegerme, amarme y ser la más bendecida a su lado.

A mi mamá Julia y mamá abue

Por enseñarme a ser una gran guerrera como ellas y ver por mi cada día, por estar conmigo en esos días de estudio para alcanzar esta meta, por su paciencia y amor incondicional siempre.

A mis hermanos: Ana, Jovis, Ale y Pao

Por ser parte de mis grandes regalos que Dios me ha dado, porque cada uno de ellos son un ejemplo para mí, por recibir siempre su apoyo, por sus consejos y compañía y sobre todo por el amor que nos tenemos, los amo.

A mi querida tutora Mayra

Por otorgarme su apoyo incondicional, por su paciencia y tiempo que me brindo para poder culminar esta tesis, por todo lo que me ha enseñado, por su confianza y amistad, por escucharme y compartir momentos inolvidables en esta etapa de mi vida.

A mi excelente Jurado y maestros

Mi agradecimiento por su apoyo, paciencia y tolerancia, por formar parte de este gran logro de mi vida y por todo el aprendizaje otorgado.

A mi querida UNAM

Por ser mi alma mater que me ha brindado conocimiento y sabiduría, que me ha permitido ser mejor persona.

A mis amigos y compañeros

Por apoyarme y compartirme sus conocimientos, por escucharme y poder contar con sus consejos.

Índice

Índice figuras

Índice tablas

Capítulo 1. Localización de sucursales bancarias considerando el criterio de ingresos de remesas

1.1 Introducción al capítulo

En este capítulo se describen problemas existentes en la cobertura de servicios financieros en México. También se analiza el impacto del ingreso de remesas para la localización de servicios bancarios, el cual genera sinergias con el negocio bancario que, explotadas adecuadamente, permitirían tanto disminuir los costos operativos como aumentar los ingresos en la industria bancaria y en el servicio financiero de ingreso de remesas recibidas en México.

Se presenta un panorama de las remesas a nivel mundial y por entidad federativa en México. Finalmente, se enfatiza el caso de la Microfinanciera y se plantea el problema concreto que se pretende resolver.

1.2 Planteamiento de problema de localización de sucursales bancarias

En el sector bancario son importantes las sucursales ya que representan el punto de acceso esencial a los servicios bancarios, las sucursales son valoradas por la mayoría de los clientes, de igual manera son importantes para la adquisición de depósitos, esto es, cuando los clientes entregan dinero al banco a cambio de intereses y para ofrecer productos y servicios.

Las sucursales bancarias son un factor fundamental para los clientes a la hora de elegir a su proveedor de servicios, por lo que los bancos aperturan nuevas sucursales, sobre todo en lugares donde haya oportunidades de crecimiento de los depósitos.

Este estudio se enfoca en una Microfinanciera que tiene presencia en la República Mexicana, una de las entidades bancarias importantes en América Latina en microfinanzas[1], especializada en otorgar créditos a los mexicanos, cuyo objetivo es llegar a más personas en el segmento de las microfinanzas en los próximos años. Lo anterior ha generado a la empresa la necesidad de proponer nuevas estrategias de negocio para lograr la adquisición de nuevos clientes.

Como primera estrategia para lograr el objetivo de la empresa financiera, ésta adquirió una empresa transmisora de dinero que ofrece el servicio de pago de

[1] Es la provisión de servicios financieros para personas en situación de pobreza, microempresas o clientes de bajos ingresos.

remesas familiares de forma conveniente, confiable y segura, lo cual ayudó a incrementar en clientes de la Microfinanciera en el ramo de las remesas.

Por otra parte, la Microfinanciera pretende cumplir su objetivo aprovechando sus recursos, entre ellos, su empresa dedicada a ofrecer servicios de pago de remesas, ubicando nuevas sucursales bancarias de manera estratégica donde haya una mayor captación de mercado de clientes potenciales que reciban remesas y puedan utilizar los servicios bancarios. En nuestro estudio nos enfocaremos en los municipios de México.

1.2.1 Problemas en la cobertura de servicios bancarios en México

1.2.1.1 Nivel de bancarización en México

La población Mexicana sufre de un bajo nivel de bancarización[2], cuya principal causa es la falta de acceso a las sucursales bancarias. Prior (2006), en su tesis menciona que uno de los elementos necesarios más importantes para que la población acceda a los servicios bancarios es la cercanía a las sucursales que ofrecen estos servicios, ya que de acuerdo con el estudio de Castellanos, Castellanos y Flores (2009), la población del segmento de menor ingreso, señala la lejanía de las sucursales como el principal factor que les impide el acceso a los servicios bancarios.

El segundo problema fundamental que explica la falta de acceso a los servicios financieros en México son los costos y requerimientos para acceder a ellos, ya que éstos, referidos a los medios de pago, cuentas corrientes (acceso a depósitos) y crédito al consumo; son en México, más altos que en el mercado en general, a pesar de que el PIB *per cápita*[3] es claramente inferior y por tanto los clientes tienen menor poder adquisitivo (Prior, 2006).

Con base en la información del Banco Mundial, a finales del año 2017 México contaba con 14 sucursales por cada 100,000 habitantes adultos (ver Figura1.1). Esta cifra resulta sumamente baja al compararla con otros países de una muestra de 20. También, al compararlo con otros países de la región, el acceso a servicios bancarios en México es menor que en Brasil, España, Colombia y Guatemala.

[2] Es el grado y nivel de utilización que la población dentro de la economía hace de productos y servicios bancarios.

[3] Es un indicador económico que refleja el valor monetario de todos los bienes y servicios finales producidos por un país normalmente anual entre el número de habitantes.

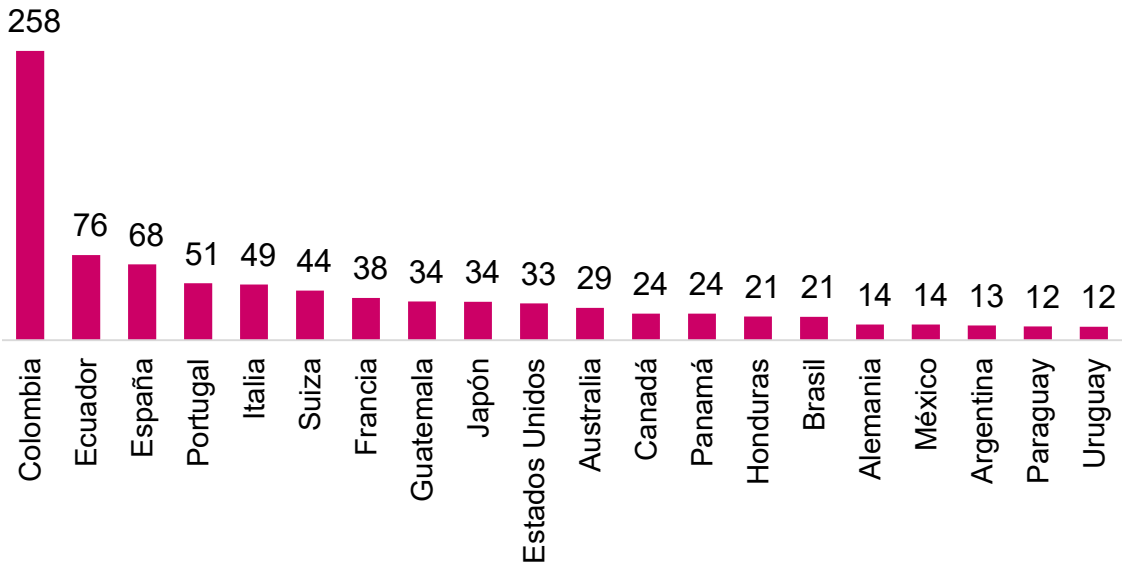

Figura 1.1. Sucursales bancarias por cada 100,000 adultos
Fuente: Elaborado con información del Grupo del Banco Mundial (2017).

A finales de 2014 en México, el 51% de los municipios con menos de 50 mil habitantes no contaba con una sucursal bancaria o infraestructura financiera. Es decir, los habitantes de más de la mitad de los municipios rurales en nuestro país no tenían acceso a recibir un crédito, tampoco la posibilidad de ahorrar o comprar un seguro en sus localidades mediante una sucursal bancaria (Videgaray, 2014).

Oaxaca, Tlaxcala y Puebla son los estados que menos sucursales bancarias presentan a nivel municipio con información de Consejo Nacional de Inclusión Financiera a septiembre 2018, cabe mencionar que Tabasco, Sinaloa y Querétaro son los estados con mayor infraestructura bancaria (ver Figura1.2).

1.2.1.2 Exclusión financiera en estratos de menor ingreso en México

Actualmente, no todos los mexicanos reciben servicios financieros por igual González (2017), por ello se están realizando diferentes iniciativas de inclusión financiera con las que se busca llegar a segmentos de población de menor ingreso en el país, otorgándoles microcréditos a personas que se ubican en la base de la pirámide del ingreso. Al no contar con acceso al sistema financiero tienen opciones limitadas para obtener financiamiento para iniciar o ampliar un negocio, por lo cual recurren a familiares o a otras formas informales de préstamo como tandas, lo que hace que sean excluidos financieramente.

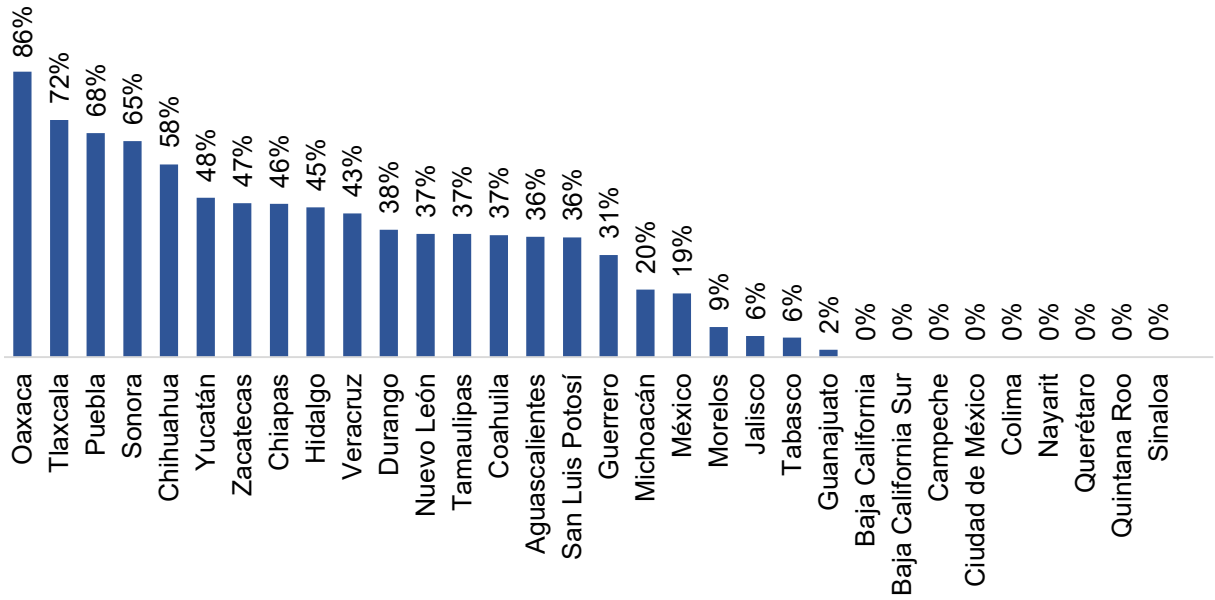

Figura 1.2. Porcentaje de municipios en cada estado, sin sucursales bancarias
Fuente: Elaborado con base en el Consejo Nacional de Inclusión Financiera (2018).

En el país, una tercera parte de las personas adultas no tiene relación con el sistema financiero formal. Específicamente, 56 de cada 100 no cuentan con ningún producto de crédito (González, 2017). Sin embargo, con base en el Instituto Nacional de Estadística y Geografía (INEGI) de un total de 123.5 millones de habitantes en el año 2017, 81 millones de personas en México tienen acceso a un teléfono celular, es decir el 65.6% de los mexicanos dispone de un dispositivo móvil, por lo que mediante esa vía se puede llegar a más personas para que tengan acceso al sistema de pagos, a tener un crédito o a ahorrar y así llegar a millones de personas que hoy en día son excluidas de servicios financieros.

1.3 Impacto de las remesas para la localización de servicios bancarios

El envío de remesas a México, es decir, el envío de dinero de quienes radican en otra nación hacia este país, es muy importante pues es el sustento de muchas familias mexicanas, por lo que también contribuyen al crecimiento económico de sus comunidades. Además, son una fuente importante de divisas para el país.

En 2014, México se ubicó en el quinto lugar de los principales países receptores de remesas en el mundo, con ingresos por 24 mil 231 millones de dólares con base en el Banco de México, tan solo después de India, China, Filipinas y Francia. En 2015, las remesas constituyeron la segunda mayor fuente de ingresos externos del país, después de la inversión extranjera directa, superando a las exportaciones petroleras, con un monto de 24 mil 792 millones de dólares.

Asimismo, las remesas deberían ser un criterio muy importante en la localización de sucursales bancarias porque representan una oportunidad para resolver el problema de la falta de acceso de servicios bancarios en segmentos de bajo ingreso, ya que la transferencia de dicho capital refuerza la capacidad de ofrecer sistemas financieros como es el ahorro y de crédito, además de que los procesos utilizados generan sinergias con el negocio bancario, que explotadas adecuadamente permitirían tanto disminuir los costos operativos como aumentar los ingresos en la industria de ahorro y crédito (Prior, 2006).

Las sinergias entre la industria de envíos de dinero y el negocio bancario se pueden clasificar en dos: sinergias operativas y sinergias de ingresos. Las primeras se basan en procesos o infraestructuras comunes a ambas industrias, que por tanto permiten a entidades interesadas operar en ambas industrias y hacerlo con menores costos operativos. Las sinergias de ingresos, son aquellas que permiten generar ingresos adicionales a operadores de remesas con infraestructura bancaria capaz de ofrecer servicios financieros a los receptores de remesas.

Con base en la tesis de Prior (2006), la infraestructura tecnológica, por su importancia en el negocio bancario como el del negocio de envío de dinero, representa un cuarenta por ciento de los costos operativos. La sinergia más destacada entre ambas industrias, con respecto a la plataforma tecnológica, se refiere a las bases de datos.

Otra sinergia operativa relevante entre la industria de envío de dinero y la industria bancaria es la implantación y gestión de canales telemáticos o alternativos, tales como los servicios de banca telefónica, cajeros automáticos e internet (Prior, 2006).

Por otro lado, las sinergias de ingreso pueden considerarse a partir de la bancarización de las remesas recibidas. Actualmente, las remesas se distribuyen mayoritariamente siguiendo la operativa "*cash to cash*", que consiste en depositar en efectivo el dinero en la red de emisión del operador de envíos de dinero en los Estados Unidos, y recibir el dinero también en efectivo en México sin utilizar ninguna cuenta de transferencia bancaria (Prior, 2006).

Sin embargo, si los flujos recibidos de remesas fueran bancarizados, es decir, recibidos sobre una cuenta corriente en lugar de distribuidos en efectivo a los receptores, las entidades bancarias que ofrecieran este servicio, generarían automáticamente un margen sobre los depósitos recibidos, ya que cada remesa enviada genera un costo cobrado al cliente.

Otra fuente de ingresos potencial para las entidades financieras dispuestas a ofrecer cuentas corrientes a los receptores de remesas, serían las implementadas en función de la contratación de productos básicos como tarjetas de débito, tarjeta de créditos, créditos, cuentas de ahorro, seguros, etc.

1.3.1 Remesas a nivel mundial

En el año 2017, México ocupó el cuarto lugar de los principales países receptores de remesas en el mundo, con ingresos por 30 mil 534 millones de dólares, ver figura 1.3, es decir, el 5.1% del total mundial (Bancomer, 2018).

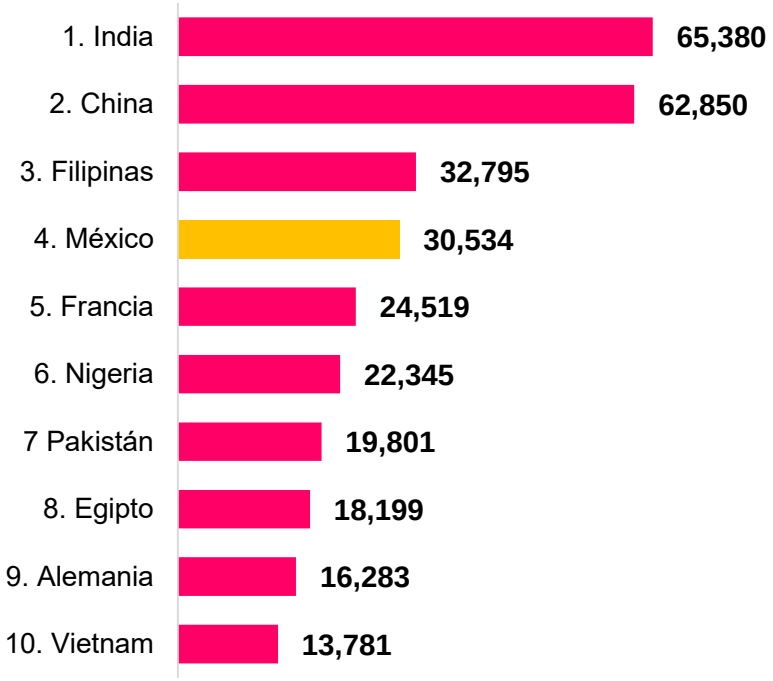

Figura 1.3. Principales países receptores de remesas, 2017(millones de dólares)
Fuente: Elaborado con base en el estudio Bancomer (2018).

Las remesas enviadas a América Latina y el Caribe alcanzaron un monto cercano a 80 mil millones de dólares en 2017, siendo México el principal país receptor en la región, con 38.4% del total. Entre 2000 y 2017, el mayor dinamismo en el crecimiento de las remesas provino de México, que creció 300%, y Centroamérica, que aumentó 500% (Bancomer, 2018).

En 2017, ingresaron a México 28 771 millones de dólares, rompiendo por segundo año consecutivo su máximo histórico, con un monto promedio por envío de 307.8 dólares y más de 93 millones de transacciones. Respecto a 2016, las remesas en 2017 crecieron 6.6%, y entre 2013 y 2017 se elevaron casi 30%, promediando una tasa anual de crecimiento de 6.6%, ver figura 1.4. (Bancomer, 2018).

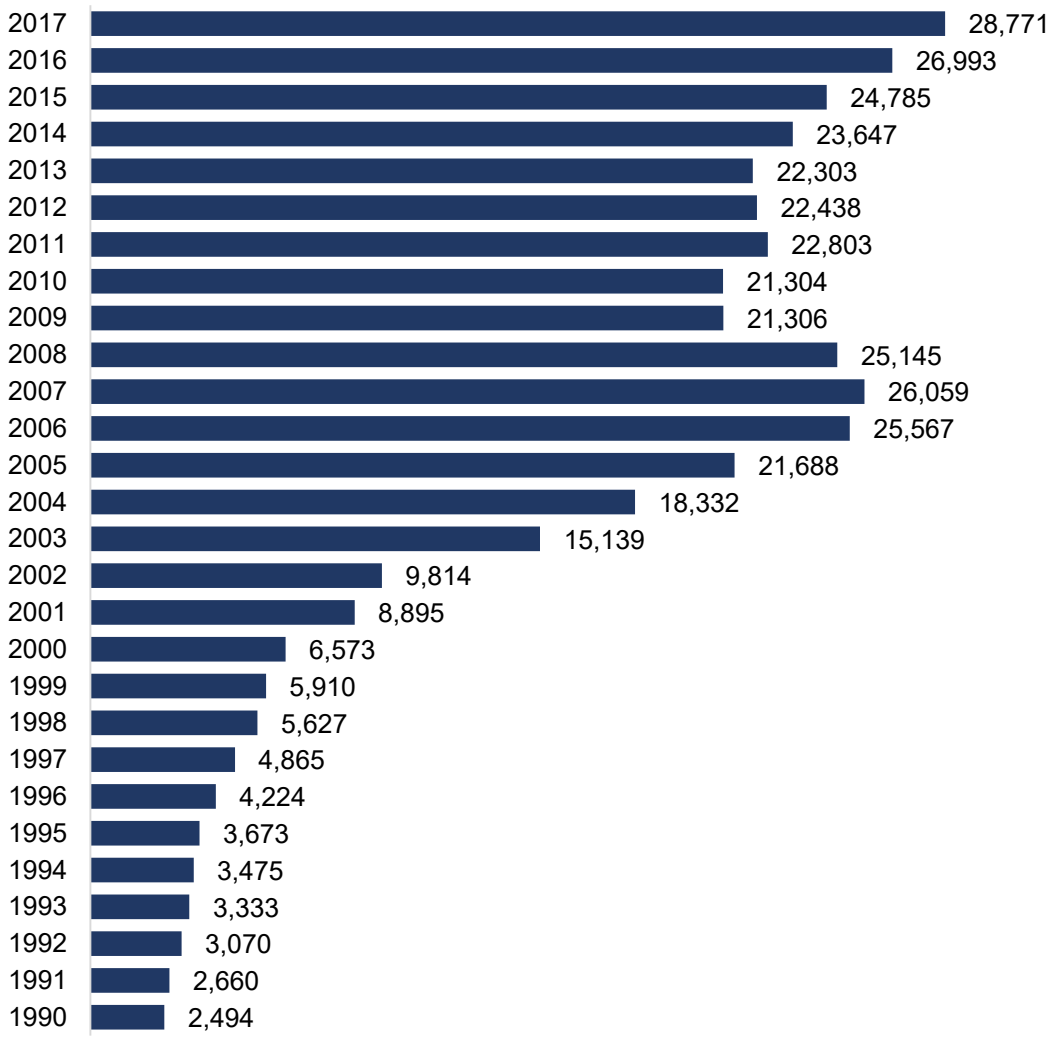

Figura 1.4. Flujo de remesas familiares a México (millones de dólares)
Fuente: Elaborado con base en el estudio Bancomer (2018).

Diez países reciben 50.5% del total de las remesas mundiales para el año 2018. India, China y Filipinas reúnen 26.2% y ocupan el primero, segundo y tercer lugar, respectivamente. México se ubica en el cuarto sitio con 4.9% ver figura 1.5, mientras que Francia en el quinto con 4.0% (Bancomer, 2018).

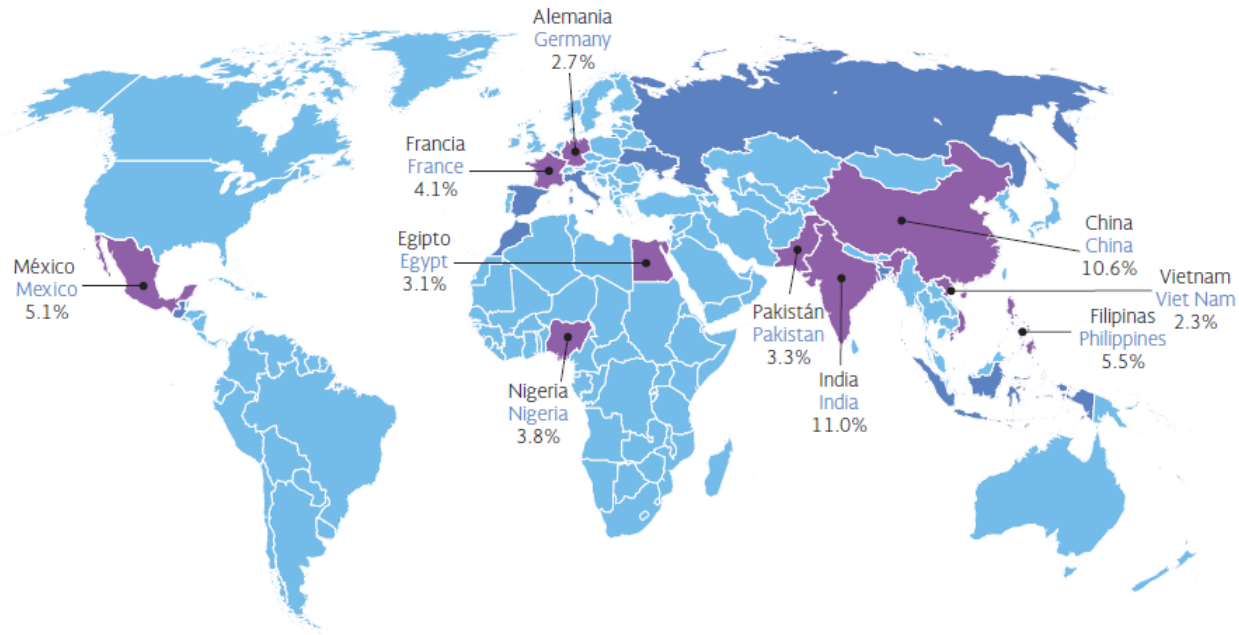

Figura 1.5. Los 20 principales países receptores de remesas, 2018
Fuente: Con base en el estudio Bancomer (2018).

México ocupó el cuarto lugar entre los principales receptores de remesas en el año 2018. Las remesas, efectivamente, son un gran ingreso en la economía de las familias mexicanas que en muchos de los casos viven en condiciones de pobreza extrema, y lo son también para los comercios y la sociedad en general, asimismo, mantienen la tranquilidad y estabilidad en algunas regiones del país. Por ello, tienen un impacto significativo en la economía mexicana, de tal forma, se requiere de servicios financieros para el flujo de remesas.

1.3.2 Remesas por entidad federativa en México

En el ingreso de remesas por entidad federativa, destaca el hecho de que los estados de Michoacán, Jalisco, Guanajuato, Estado de México y Puebla fueron los principales cinco receptores de remesas en 2017, en conjunto, estas cinco entidades concentran el 40% de los ingresos por remesas en el país, ver figura 1.6.

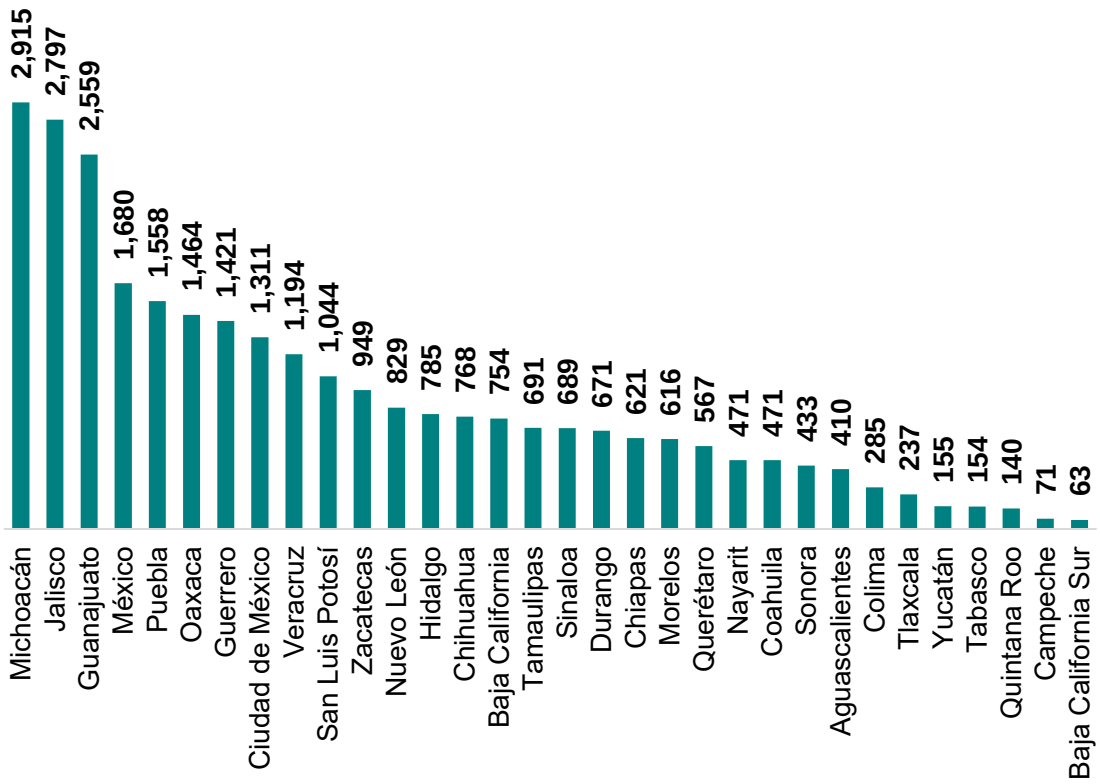

Figura 1.6. Ingreso de remesas por Entidad Federativa (millones de dólares)
Fuente: Elaborado con base en el estudio Bancomer (2018)

Considerando el porcentaje del ingreso de remesas durante el año 2017 por entidad federativa, el porcentaje más alto con 10.1 puntos porcentuales se encuentra el estado de Michoacán por encima de los 31 estados restantes de México (ver figura 1.7).

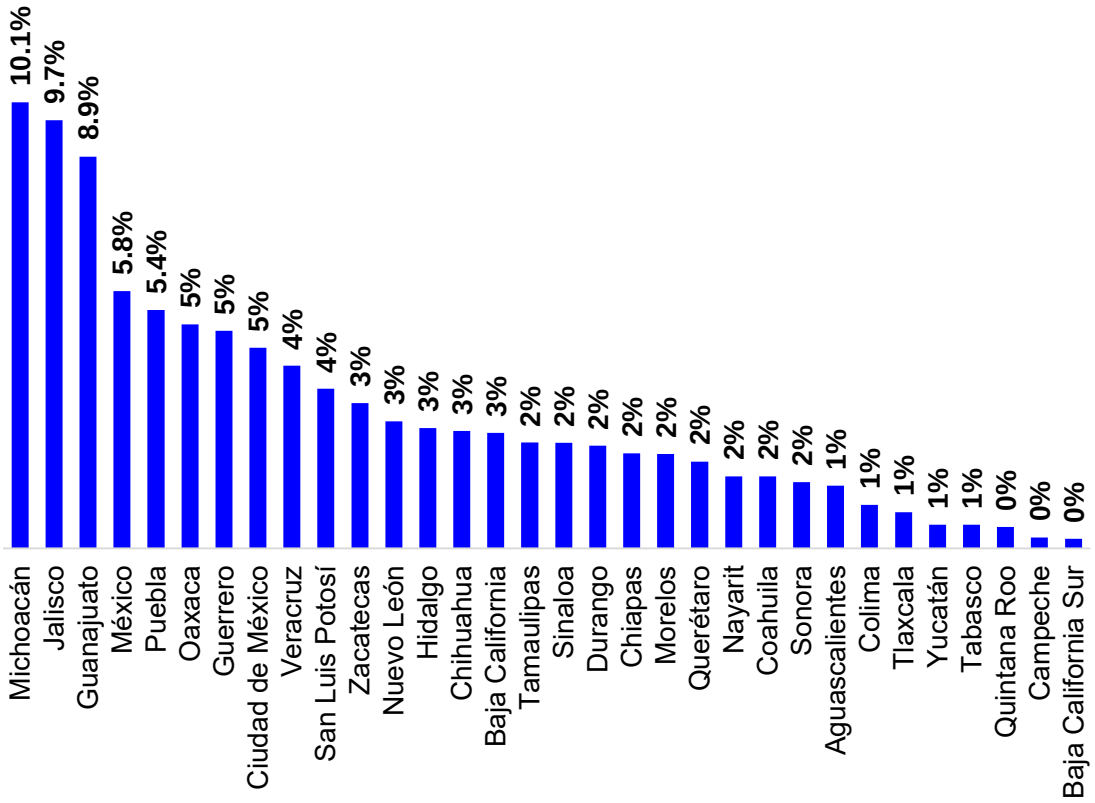

Figura 1.7. Porcentaje de remesas anuales 2017
Fuente: Elaborado con base en el estudio Bancomer (2018).

Asimismo, Michoacán, Oaxaca y Zacatecas son los estados con mayor dependencia de las remesas a nivel nacional, esto es, presentan mayor porcentaje de PIB estatal, siendo de 10.9%, 9.6% y 9.4% su PIB estatal respectivamente en el año 2017.

Entidad State	%	Entidad State	%
1. Michoacán	10.9%	17. Aguascalientes	2.8%
2. Oaxaca	9.6%	18. Veracruz	2.4%
3. Zacatecas	9.4%	19. Querétaro	2.2%
4. Guerrero	9.2%	20. Tamaulipas	2.2%
5. Nayarit	6.1%	21. Chihuahua	2.1%
6. Guanajuato	5.5%	22. Baja california	2.1%
7. Durango	5.2%	23. México	1.7%
8. Morelos	5.1%	24. Sonora	1.2%
9. Hidalgo	4.7%	25. Coahuila	1.2%
10. San Luis Potosí	4.4%	26. Nuevo León	1.0%
11. Colima	4.4%	27. Yucatán	1.0%
12. Puebla	4.0%	28. Quintana Roo	0.8%
13. Tlaxcala	3.9%	29. Ciudad de México	0.7%
14. Jalisco	3.6%	30. Tabasco	0.7%
15. Chiapas	3.5%	31. Baja California Sur	0.6%
16. Sinaloa	2.8%	32. Campeche	0.3%

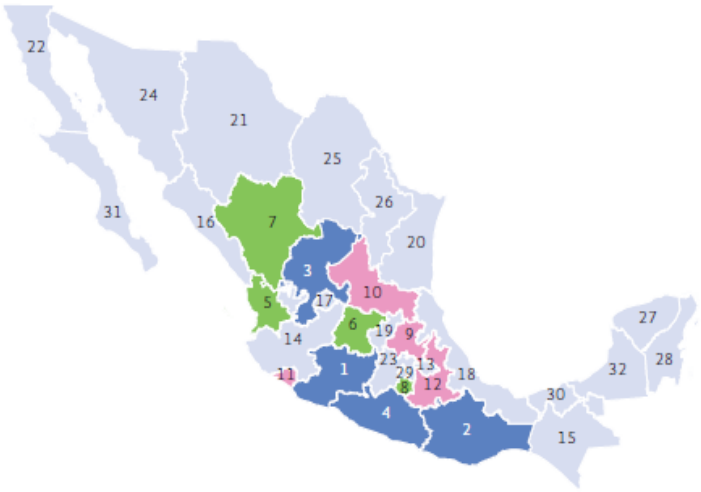

Figura 1.8. Las 10 entidades con mayor dependencia de remesas (remesas como % del PIB estatal)
Fuente: Con base en el estudio Bancomer (2018).

Como se puede observar, el estado de Michoacán es el que recibió el porcentaje más alto de ingreso de remesas 10.1% anual en 2017 y también el de mayor dependencia de las remesas a nivel nacional, siendo de 10.9% su PIB estatal en el año 2017.

1.4 Caso Microfinanciera mexicana

Microfinanciera mexicana tiene como propósito la introducción financiera en la base de la pirámide poblacional[4] en América, es decir, generar oportunidad de desarrollo a través de servicios financieros a personas de segmentos populares. Esto lo hace por medio de un modelo de negocio estratégico, basado en segmentos de clientes potenciales, para reducir el riesgo de incumplir con el pago, con base en estudios de la Microfinanciera (que no han sido publicados). La Microfinanciera mexicana cuyo objetivo es aumentar en clientes en los próximos años, pretende lograrlo aprovechando sus recursos utilizando entre ellos su empresa dedicada a ofrecer servicios de pago de remesas, por este motivo el criterio de ingreso de remesas en la localización de sus nuevas sucursales es muy importante ya que cubre las necesidades de los posibles clientes potenciales.

La Microfinanciera tiene presencia en los 32 estados de la República Mexicana (ver Figura 1.9) considerando sucursales bancarias, cuenta con clientes del sector medio bajo, bajo alto y bajo, con un perfil del cliente predominante en su mayoría

[4] Segmento de la población con menores ingresos, es decir niveles socioeconómicos D, D+ y C-

mujeres. Dispone de productos de créditos para diferentes tipos de clientes y necesidades y seguros.

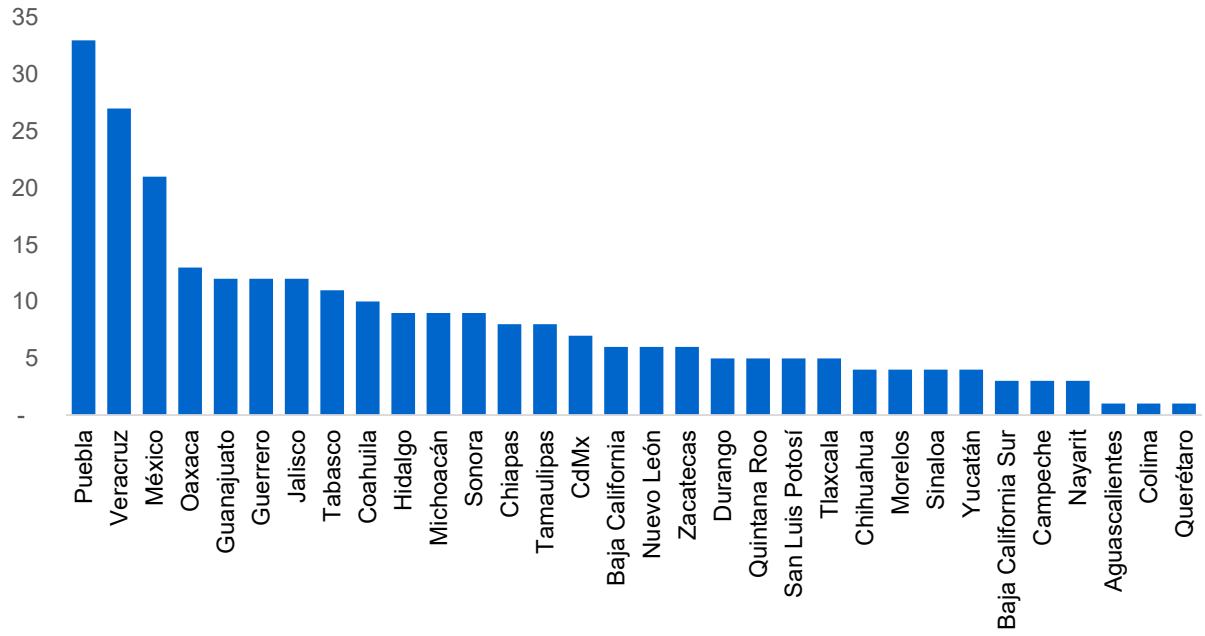

Figura 1.9. Sucursales bancarias de Microfinanciera
Fuente: Elaborado con base en el CNBV[5] septiembre 2018

En la Figura 1.10 se muestra el número de población adulta, es decir personas mayores a 18 años, correspondiente a cada sucursal bancaria de la Microfinanciera por entidad federativa.

[5] CNBV (Comisión Nacional Bancaria y de Valores), se ofrecen información de sucursales, cajeros y otras variables

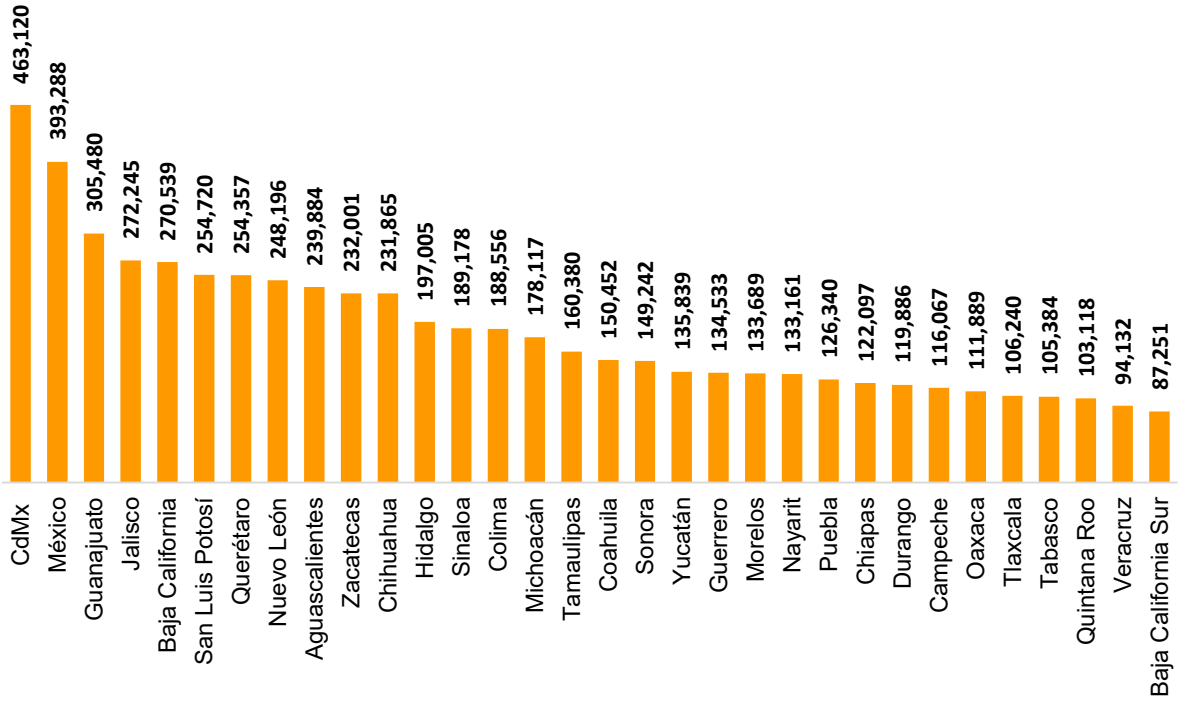

Figura 1.10. Población adulta por sucursal bancaria de la Microfinanciera
Fuente: Elaborado con base en el DENUE[6], 2018

En la Figura 1.11 se muestra el número de sucursales bancarias de la Microfinanciera y de sus cinco principales competidores por entidad federativa. De las 12,792 sucursales bancarias existentes en el país a septiembre de 2018, la presencia de la Microfinanciera era menor al 2.1%, esto es, existe una gran oportunidad de localización de sucursales para la captación de clientes potenciales.

[6] DENUE (Directorio Estadístico Nacional de Unidades Económicas), se ofrecen los datos de identificación, ubicación, actividad económica y tamaño de los negocios activos en el territorio nacional.

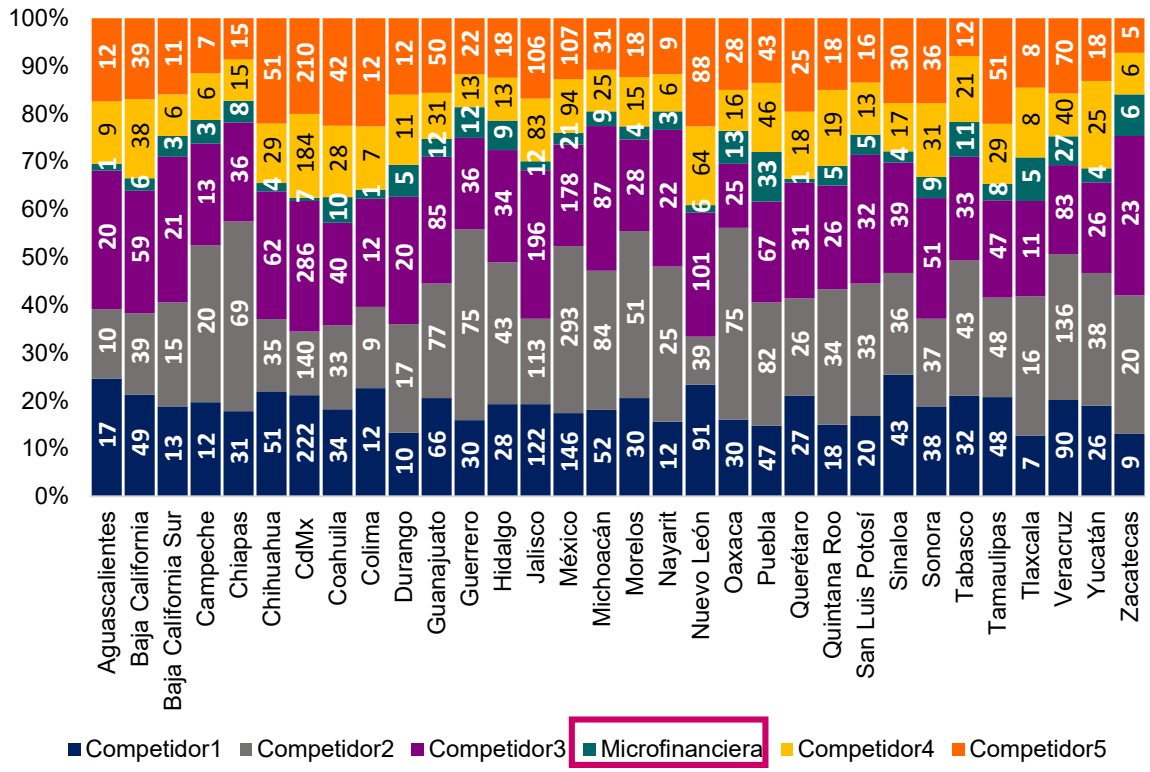

Figura 1.11. Número de sucursales de los principales competidores y Microfinanciera
Fuente: Elaborado con base en CNBV (Comisión Nacional Bancaria y de Valores), septiembre 2018

En la Figura 1.12 se muestra el mercado potencial con base en el objetivo de crecimiento del segmento sub-atendido de la Microfinanciera mexicana, es decir, se considerará como mercado potencial a la población que cumpla con el perfil de cliente sub-atendido por la Microfinanciera, esto es, personas mayores a 20 años, económicamente activas con actividad económica independiente.

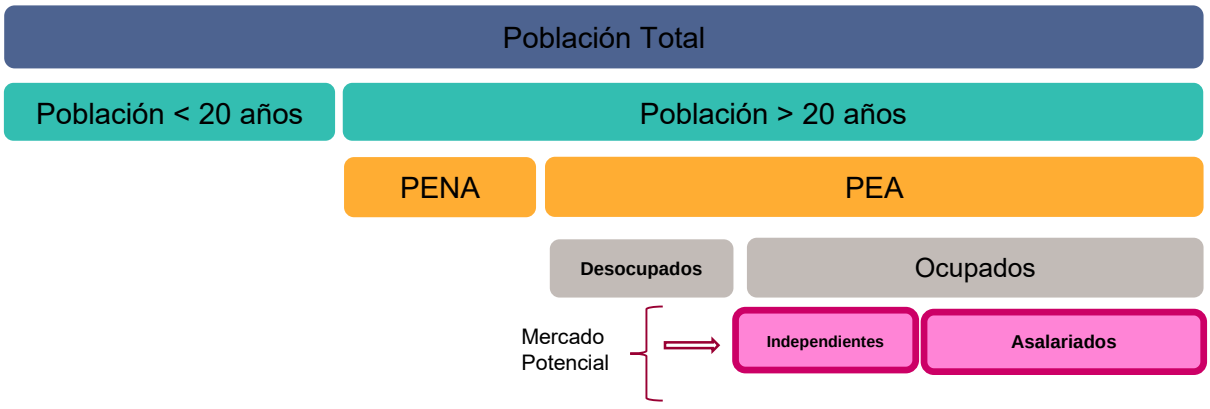

Figura 1.12. Mercado Potencial de Microfinanciera
Fuente: Elaboración propia
Nota: PENA= Población Económicamente No Activa, PEA= Población Económicamente Activa

La localización de sucursales bancarias de la Microfinanciera asocia un mercado meta de niveles socioeconómicos D, D+ y C, que conforman la base de la pirámide, con un nivel educativo promedio de secundaria, con actividad económica predominante en el sector de comercio.

Actualmente, los criterios que utiliza la Microfinanciera mexicana para la localización de sus sucursales bancarias en municipios se pueden ver en la figura 1.13:

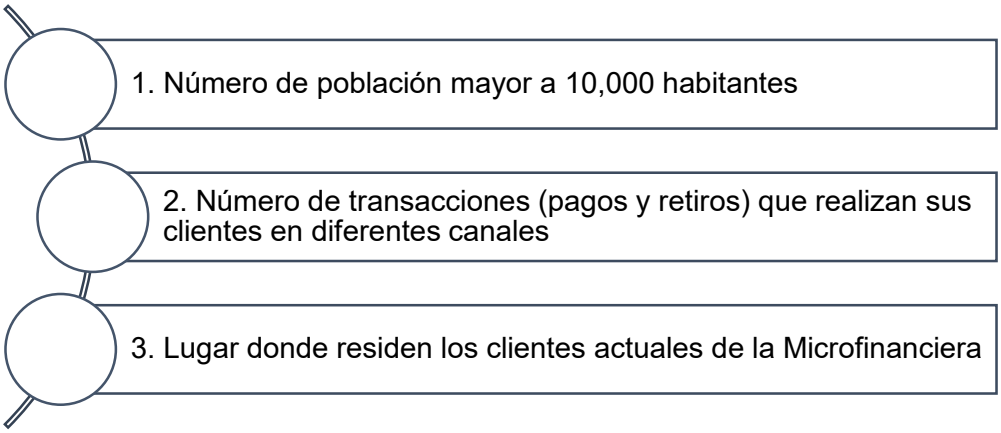

Figura 1.13. Criterios que se utilizan para la localización de sus sucursales bancarias en municipios
Fuente: Elaborado con base en información de la Microfinanciera

Los criterios que utiliza la Microfinanciera se describen a continuación:

1. Número de población mayor a 10,000 habitantes

El criterio más importante que utiliza la Microfinanciera para la localización de sus sucursales bancarias es que el número de habitantes sea mayor a 10,000. En el estudio de Castellanos, *et al. (*2009), se coincide en utilizar cómo principal criterio para la localización de sucursales bancarias el número de habitantes en la región, zona o área geográfica delimitada por ciertas medidas específicas por parte de la Microfinanciera, dónde será ubicada la sucursal bancaria.

2. Número de transacciones (pagos y retiros) que realizan sus clientes en otros canales

El número de transacciones que realizan los clientes de la Microfinanciera en otros canales, es también un criterio importante debido a que los principales medios donde los clientes realizan sus pagos y retiros son competidores de la Microfinanciera por lo que la estrategia es cubrir esta demanda así cómo poder quitarles clientes a sus competidores, así, una manera de lograrlo es poner sucursales en lugares donde los clientes realizan más transacciones, teniendo en cuenta que el objetivo de localizar sucursales bancarias es incrementar la captación de clientes.

3. Lugar donde residen los clientes actuales de la Microfinanciera

La Microfinanciera considera el criterio del lugar donde se ubican los clientes actuales ya que para la ella es importante mantener la cercanía con sus clientes y sus sucursales bancarias.

Cabe mencionar que el parámetro que utiliza la Microfinanciera para localizar sus sucursales bancarias es de un diámetro de 0 a 15 kilómetros a la redonda, partiendo de un punto específico, basado en las sucursales existentes y clientes existentes.

En esta tesis los datos utilizados para el modelo tales como: número de habitantes o población, superficie en kilómetros cuadrados, población adulta, tipo población, total de sucursales de las instituciones financieras, cajeros, Terminales Punto de Venta, transacciones en TPV (Terminal Punto de Venta) y transacciones en cajeros automáticos se obtienen de Banxico (2018) y de la CNBV (2017) que se encuentran publicados a nivel municipio, por tal motivo se manejó a este nivel de detalle. Al realizarlo a nivel municipio fue de gran beneficio ya que delimita a una menor área para la localización de sus sucursales bancarias de la Microfinanciera y hace más precisa el área geográfica para la localización de sucursales y no a nivel estado cómo actualmente lo realiza la institución.

Considerando que la pérdida mínima al cerrar una sucursal es de aproximadamente 2 millones de dólares (mdd) de acuerdo con Rodríguez (2014), y que el capital promedio para constituir un banco es de 33 mdd en México , cifra muy por arriba de los 16 mdd que se necesitan en Suiza , 13 mdd en Costa Rica o Guatemala, los 11 mdd en Nicaragua y los 10 mdd en Brasil y Panamá, números mencionadas por la Comisión Federal de Competencia Económica (CFCE); es de gran importancia establecer un modelo que ayude en la mejor toma de decisiones para la localización de las nuevas sucursales bancarias de la Microfinanciera.

La Microfinanciera entonces, al decidir dónde ubicará sus nuevas sucursales bancarias, debe saber qué criterios son los más convenientes para la localización de sus sucursales que le permitan incrementar el número de clientes existentes y expandirse a un mayor número de clientes potenciales. De igual manera, estos clientes potenciales se concentran en su mayoría en personas que reciben remesas o tienen la oportunidad de solicitar un crédito, ya que su propósito es incrementar las remesas y el crédito del sector popular.

Hasta ahora no se ha considerado el criterio de remesas en la localización de sucursales bancarias, y dado que México ocupa el cuarto lugar como receptor de remesas a nivel mundial es importante considerarlo, además en México dirigirse al lugar para cobrar las remesas tiene un costo promedio de transporte de 50 pesos, lo que representa el 50% de un salario mínimo diario con base en Bancomer (2017). En la presente tesis se muestra una propuesta para localizar sucursales bancarias de la Microfinanciera mexicana considerando el criterio de ingreso de remesas que

le permitan llegar a un mayor número de clientes potenciales en los municipios de México.

1.5 Problema concreto por resolver

Así el problema es el siguiente:

Determinar los sitios donde exista mayor concentración de número de clientes potenciales de la República Mexicana, a nivel de municipios del país para localizar sucursales bancarias de una Microfinanciera, considerando el número de clientes existentes de la Microfinanciera e incluyendo el criterio de ingreso de remesas para reforzar la toma de decisiones e incrementar el número de clientes potenciales.

La Microfinanciera también desea cubrir las necesidades de los clientes de manera rápida, esto significa que debe producir modelos de forma ágil y automática, que puedan analizar volumen de datos grandes y complejos, considerando que el procesamiento computacional debe ser veloz y económico para la cantidad de datos, y producir resultados más rápidos y precisos, así la organización tiene una mejor oportunidad de identificar oportunidades rentables para sus clientes, tomar decisiones en tiempos más cortos y de evitar riesgos desconocidos.

1.6 Conclusión del capítulo

Finalmente, como término de este capítulo tenemos que el problema concreto de esta tesis es un problema real de la Microfinanciera, cuyo objetivo es crecer en clientes y utilizar de manera ventajosa los recursos con los que cuenta, en este caso, la captación de clientes que reciben remesas, una estrategia para lograrlo es localizando nuevas sucursales bancarias para expandir el número de clientes y productos.

En el siguiente capítulo se describen los aportes teórico - metodológicos de la literatura con base en el objeto de estudio de esta tesis.

Capítulo 2. Marco teórico

2.1 Introducción al capítulo

El objeto de estudio de la presente tesis, es la localización de servicios, en particular servicios financieros como son las sucursales bancarias. A partir de él, en este capítulo se detallan los autores que han aportado e investigado en el tema, así como los enfoques con los que lo han abordado.

Se define la línea de investigación en la que encaja esta tesis para definir la aportación que hará la misma: un método para la localización de sucursales bancarias considerando el criterio de ingreso de remesas en los municipios de la República Mexicana.

En la sección de métodos y modelos, se describe el proceso de selección de variables para la localización, se detalla el modelo de aprendizaje de máquina y el algoritmo de Bosque aleatorio (o *Random Forest)*, la evaluación del modelo, y finalmente los conceptos para el desarrollo de la presente tesis.

Después, se describe cada una de las etapas que forman el procedimiento que se seguirá para resolver el problema de investigación.

2.2. Marco teórico

Se presentan los diferentes abordajes teóricos y los métodos utilizados con respecto al objeto de estudio que es la localización de servicios financieros, en particular sucursales bancarias.

Con base en la revisión teórico-metodológica de la localización de servicios financieros y bancarios, se muestra en la figura 2.2 el mapa conceptual basado en las fuentes consultadas acerca de la localización de sucursales bancarias, también se puede apreciar la clasificación de los modelos para la localización de sucursales bancarias con base en la revisión de literatura: modelos de análisis estadístico, de análisis multicriterio, de aprendizaje supervisado y de máxima cobertura.

Se encontró que la línea de investigación que predomina en esta tesis es la localización de sucursales bancarias, la cual es de gran interés en esta tesis ya que contempla la utilización de ellas para dar solución al problema de investigación.

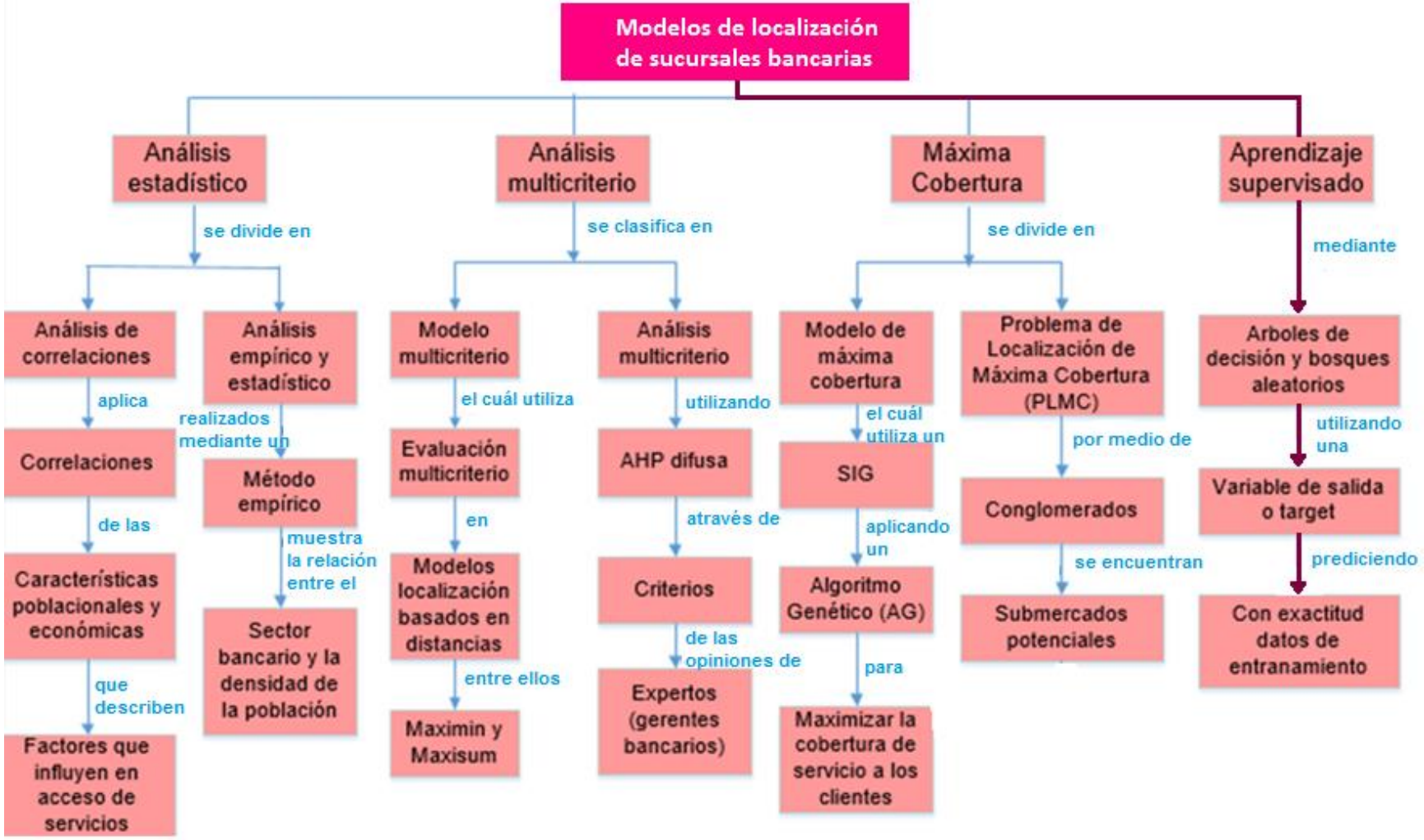

Figura 2.2. Mapa conceptual de modelos de localización encontrados en la revisión bibliográfica

Fuente: Elaboración propia, con base en Steiner y Medina (2002)*;* Peña (2012); Bosque y Franco (1995); Cinar (2009); Tong, Murray, Xiao y Hernández (2009); Schneider, Seifert y Sunyaev (2014); James, Witten, Hastie y Tibshirani (2013).

2.2.1. Localización de servicios financieros y bancarios

El mapa de la figura 2.2 se describen los modelos encontrados en la revisión teórico-metodológica aplicados en la localización de sucursales bancarias.

Con respecto a los modelos basados en análisis estadístico, encontramos los factores que han sido utilizados para la localización de sucursales bancarias:

Beck *et al.* (2006), elaboraron un análisis empírico de los factores que influyen en el alcance geográfico y demográfico de los servicios financieros entre países (donde interpretaron que a mayor cantidad de sucursales bancarias respecto a la extensión geográfica y número de habitantes resulta ser un indicativo de un mayor acceso al uso de los servicios financieros). Sus resultados muestran que el alcance del sector bancario se encuentra directamente ligado al desarrollo financiero, la densidad de población y el nivel económico.

Asimismo, Castellanos *et al.* (2009) con respecto a los factores que se utilizan para la localización de sucursales bancarias, mediante el análisis de correlación describen los factores que influyen en el acceso a servicios financieros a nivel estatal, cuyos resultados fueron: el nivel de ingreso (PIB real) que es una variable relevante en la ubicación de las sucursales bancarias; la variable de población económicamente activa de edad (personas en edad de trabajar es decir entre18 y 64 años) es significativa y positiva. Esto indica que a mayor cantidad de población económicamente activa existirá una mayor importancia en los municipios para determinar la localización de bancos. Lo que indica una relación directa entre la decisión de la instalación de infraestructura y la población en edad laboral. Por otra parte, los autores encontraron que el número de empleados registrados en el sector terciario tiene una gran importancia para la captación de la actividad comercial y de servicios, pero está variable sólo fue significativa en los bancos que presentaron menor cantidad de sus activos totales.

Por su parte Avery (1991) demuestra la existencia de una relación positiva entre el número de sucursales bancarias en los alrededores de Detroit, Cleveland, Filadelfia, Boston y Atlanta en Estados Unidos, y el ingreso promedio de los habitantes, el número de empleados asalariados, el valor de la vivienda y el número de empresas por habitante.

Steiner y Medina (2002), encuentran una correlación positiva entre el número de sucursales bancarias y el nivel de ingreso y bienestar del hogar, la seguridad del vecindario y la calidad de la infraestructura de la zona.

En cuanto a los modelos basados en análisis muticriterio, para conocer los criterios para la localización de una sucursal bancaria, Cinar (2009) define mediante AHP difuso[7], cinco criterios principales: demográficos, socioeconómicos, sector productivo, competencia bancaria y potencial comercial, los cuáles se muestran en la figura 2.1. Cabe mencionar que este estudio estableció un cuestionario para obtener los pesos de los criterios principales y los sub-criterios aplicado a expertos (gerentes bancarios).

Figura 2.1. Criterios y subcriterios para la decisión de ubicación de una sucursal bancaria
Fuente: Elaborado con base en el artículo de Cinar (2009)

Por otro lado, Bosque y Franco (1995) plantean modelos que buscan maximizar las distancias entre la población y las instalaciones no-deseables y pueden ser, a su vez, modelos *maximin* que buscan maximizar la distancia mínima entre población e instalaciones y modelos *maxisum* que pretenden maximizar la distancia total entre

[7] AHP (Proceso Analítico Jerárquico) difuso es una metodología diseñada por Saaty (1990) para resolver problemas de toma de decisión multicriterios. El AHP difuso ayuda a establecer el nivel de importancia entre los criterios y subcriterios definidos para la selección, a partir de una comparación cualitativa entre ellos, que posteriormente se traslada a una escala numérica que permite realizar los cálculos matemáticos. Así mismo, jerarquiza alterativas de una decisión.

población e instalaciones. Estos modelos parten de considerar la distancia entre la población y un conjunto de sitios candidatos para la localización de instalaciones.

Con respecto a los modelos de máxima cobertura:

Love et al. (1988) narran los problemas de localización de un solo servicio, cuyo objetivo es encontrar un sitio óptimo. Las ubicaciones existentes son tratadas como puntos, las demandas y costos son conocidos. Cómo caso particular de este problema de localización de un solo servicio, se considera como métrica la distancia euclidiana[8].

Asimismo*, Love et al.* (1988) describen los problemas de localización de múltiples servicios ya que, en otros casos, a menudo ocurre que se debe ubicar más de una nueva instalación. Se tienen dos casos: si ningún par de nuevas instalaciones interactúa (no existe flujo entre instalaciones existentes), podemos tratar la ubicación de cada nueva instalación como un problema separado y aplicar las características de localización de un sólo servicio, de lo contrario, si las nuevas ubicaciones de las instalaciones interactúan entre sí, se deberán optimizar simultáneamente. En el presente trabajo se pretende dar mejores opciones a la Microfinanciera para la localización de sucursales bancarias.

Schneider *et al.* (2014), identifican regiones con alto potencial de mercado para el desarrollo de la red de sucursales y para identificar sitios potenciales para nuevas sucursales bancarias, utilizan el sistema de apoyo de decisiones BANKMAP, el cual cuenta con tres fases: I: evaluación de mercado potencial, el objetivo de esta fase es determinar el universo para cada región (por ejemplo, código postal o área metropolitana) con base en criterios múltiples de diversas fuentes, como sociodemográficas (por ejemplo, densidad de población), mercado (por ejemplo, ahorro estimado de todos los ciudadanos), competidor (por ejemplo, densidad del competidor), y los datos del banco (por ejemplo, número de clientes del préstamo); II: evaluación de submercado, para la identificación de los estados, ciudades, o municipios con alto potencial de mercado; III: planificación de Red de Sucursales, se calculan las distancias entre las celdas de la cuadrícula hexagonal, utilizando la distancia euclidiana entre los centros de las celdas. Este modelo se basa en el *Maximal Coverage Location Problem (MCLP)*, conocido en español cómo Problema de Localización de Máxima Cobertura, desarrollado por Church y Roberts (1983), donde se busca maximizar la demanda cubierta.

Por otra parte, los algoritmos de aprendizaje supervisado cuentan con varias ventajas entre ellas tenemos, que producen modelos en un periodo corto de tiempo, con mayor precisión y aplicable a gran cantidad de volumen de datos, además de que es capaz de detectar patrones de comportamiento que no se ven a simple vista.

[8] Distancia euclidiana: es la distancia entre dos puntos y se mide como la longitud del segmento que los une en un espacio geométrico.

En esta revisión de literatura también se encontró que se aplican modelos de aprendizaje supervisado en la localización:

Con base en Soco (2017) un equipo de científicos trabajó en la creación de un modelo de aprendizaje supervisado, en este caso un Bosque Aleatorio (*Random Forest*), dicho modelo predice los municipios de México donde hay mayores probabilidades de encontrar fosas clandestinas que a la fecha no hayan sido identificadas ni por las autoridades. Al ser analizados los datos de forma anual, el modelo presenta información relevante sobre dónde se debería empezar a buscar fosas en los años subsecuentes.

La Microfinanciera tiene como meta incrementar el número de clientes, mediante la localización de nuevas sucursales bancarias pretende cubrir una mayor cantidad de clientes potenciales, teniendo en cuenta que la exactitud de llegar a un mayor número de clientes sea la más precisa.

Asimismo, quiere tener rapidez en la ejecución del modelo dando solución en periodos más cortos y considerando que la empresa no cuenta con grandes sistemas operativos para el manejo de altos volúmenes de datos. También se pretende poder manejar criterios nuevos en la localización de sucursales bancarias, sin que esto cause una dificultad para poder ingresar nuevas variables de entrada en el modelo a utilizar.

Con base en lo anterior se utilizó la técnica de aprendizaje supervisado como manera de solución para este problema de investigación en la presente tesis.

2.2.2. Justificación del modelo basado en aprendizaje automático

Considerando la revisión de literatura, Bermejo (2017) nos menciona que el *Machine Learning* (ML), aprendizaje de máquina o aprendizaje automático, es uno de los temas más mencionados e importantes actualmente gracias al impuso que han otorgado gigantes tecnológicos como Microsoft, Google o Facebook. Sin embargo, la creciente cantidad de datos, que producimos y consumimos a diario, así como el abaratamiento de costos en el almacenamiento de estos datos y su procesamiento computacional, han propiciado la aparición del aprendizaje automático como tecnología en auge. Por parte de las empresas se suponen importantes beneficios como la reducción de costos y ahorro de tiempo en el almacenamiento de datos y su procesamiento computacional, en definitiva, una ventaja competitiva a corto y largo plazo.

Asimismo, Bermejo (2017) indica que se debe de usar el aprendizaje de máquina, debido a que el volumen de datos que manejamos a diario es cada vez mayor, además de que estos datos son de diferentes clases, esto es, datos estructurados y datos no estructurados. Los datos estructurados son aquellos que están

conformados por tablas, que permiten relacionar unas tablas con otras y son generalmente conocidos como base de datos SQL; y los datos no estructurados que son aquellos que no tienen algún tipo de orden que permita hacer una categorización, en esta categoría podemos encontrar datos en forma de archivos de texto, pdf, emails, imágenes, archivos de sonido, chats, tweets, páginas web, etc.

Todo lo anterior, ha generado que ahora sea mucho más factible utilizar los algoritmos de aprendizaje de máquina para producir modelos en un periodo corto de tiempo, de forma precisa y aplicarse a una gran cantidad de datos. Y es mediante la construcción de estos modelos que hoy una empresa cuenta con más y mejores posibilidades para identificar oportunidades rentables y evitar riesgos desconocidos en su mercado. El aprendizaje de máquina es capaz de detectar patrones de comportamiento imperceptibles al ojo humano y permite a la empresa una reducción del tiempo de personal dedicado a estos procesos y tener una visión más amplia de sus procesos.

Vallejo (2017) menciona que el Bosque Aleatorio es una técnica de aprendizaje de máquina que mejora la exactitud de las predicciones de valores discretos en los modelos. Alarcón (2017) expone que el modelo Bosque Aleatorio mejora la precisión en la variable de destino, donde dicha variable puede tomar un conjunto finito de valores.

Por tanto, se utilizó el aprendizaje de máquina en esta tesis debido a que produce modelos en un periodo de tiempo más corto y de mayor precisión comparado con el método que utiliza la Microfinanciera que es mediante Sistemas de Información Geográfica y servirá a la Microfinanciera a identificar los sitios viables de la localización de sucursales bancarias.

En las tablas 2.1, 2.2, 2.3 y 2.4 se resumen los principales artículos consultados relacionados con localización de servicios financieros y bancarios, así como los criterios y modelos utilizados para su localización, con base en las diferentes técnicas de los modelos encontrados mostrados en el marco teórico.

Tabla 2.1. Principales fuentes consultadas para la localización de servicios financieros y bancarios basados en análisis estadístico

Técnica	Autor(es) y año	Criterios utilizados	Modelo utilizado	Software utilizado	Aportación a la investigación	País
Análisis estadístico	Steiner, R., & Medina C.,2002	Número de sucursales bancarias, nivel de ingreso y bienestar del hogar, la seguridad del vecindario y la calidad de la infraestructura de la zona	Correlaciones	No se indica	Maximización del bienestar del consumidor y las ganancias de los bancos.	Colombia
	Consoni, E., Taylor P., 2007	Ciudades brasileñas y bancos	Método estadístico multivariado de análisis de componentes principales en una	No se indica	Para localización de bancos.	Brasil

			matriz de ciudades *versus* prestadores de servicios bancarios			
	Castellanos, S., Castellanos, V. & Flores, V., 2009	Características poblacionales, empleo, infraestructura, variables dicotómicas por región.	Análisis de la relación entre infraestructura y varias características poblacionales y económicas de las entidades federativas o municipios	No se indica	Se establecen los factores de influencia en la instalación regional de infraestructura bancaria	México
	De la Fuente, S.,2011	Conglomerados, distancia euclidiana	Metódos de Conglomerados (Clústers)	SPSS estadístico	Permite identificar el método a utilizar para el análisis de conglomerados y definir las etapas.	No se indica
	Peña E.,2012	Pobreza con acceso a servicios financieros	Empirismo	No se indica	Evidencia empírica de la relación entre pobreza y acceso a servicios financieros para población rural.	Colombia
	Espinosa, C.,2013	Ejes viales, parroquias, infraestructura bancaria, comercios y servicios, población, población económicamente activa (PEA), penetración de internet.	Variables de influencia para la localización de áreas de ubicación bancaria usando un Sistema de Información Geográfica (GIS).	SIG	Ponderación de variables para ubicar bancos	Ecuador

Fuente: Elaborado con base en Steiner y Medina (2002); Prior (2006); Consoni y Taylor (2007); Castellanos, Castellanos y Flores (2009); De la Fuente (2011); Peña (2012); Espinosa (2013).

Tabla 2.2. Principales fuentes consultadas para la localización de servicios financieros y bancarios basados en análisis de máxima cobertura

Técnica	Autor(es) y año	Criterios utilizados	Modelo utilizado	Software utilizado	Aportación a la investigación	País
Máxima cobertura	Bosque S. & García, R., 2000	Geotecnia del terreno, que mide lo barato o fácil que resulta construir en cada punto del territorio, distancia a lugares donde existen empleos y calidad visual alrededor de cada punto	Modelo de localización apoyado en Sistemas de Información Geográfica	SIG	Minimización de las distancias que un conjunto de una población tiene que recorrer para utilizar alguna de las instalaciones planteadas.	Venezuela
	Tong D., Murray A., Xiao N., Hernández C., 2009	Número de instalaciones, distancia máxima de servicio aceptable,índice se sitios potenciales , índice de objetos de demanda, etc.	*Modeling Maximal Coverage*	ArcGIS (ArcMap), El GA se codificó en C ++	Modelo matemático que maximice la cobertura de clientes en una región.	Arizona

	Alarcón, Z.,2012	Localización de cada usuario, su demanda y los costos de transporte (tiempo, distancia, utilidad, presupuesto, etc.), número de servicios, ubicación geográfica	Modelo de localización de servicios en redes con 2 objetivos.	MAPLE 13	Para satisfacer la demanda y la mínima inversión.	Estado de México
	Díaz,J. & Pineda, J., 2013	Espacio de localización, competencia, rentabilidad, costos de transporte, distancias con sus competidores, demanda de clientes	Teoría de la interacción espacial	SPSS estadístico	Nos permite identificar los factores para la localización de una sucursal bancaria.	Toluca
	Schneider, S., Seifert, F., & Sunyaev, A., 2014	Sociodemográficos (población y densidad de hogares, tasa de crecimiento de población, población por edad, ingresos del hogar) y características que describen el mercado (número de competidores, sucursales de bancos, sucursales propias).	Problema de localización de máximo cubrimiento (MCLP)	SPSS estadístico y Java	Localización de sucursales bancarias partiendo de la evaluación del mercado y supermercados, posteriormente se aplica el MCLP para localizar las sucursales.	Alemania
	Zamora, D.,2015	Matriz de impacto cruzado o de Haddon	*Model maximal service area problem.*	Google Earth, ArcGIS, ArcMap	Para seleccionar un conjunto de localizaciones para las instalaciones, donde los polígonos de área de servicios combinado de estos sitios abarcan la mayor área posible en la región de la demanda	Toluca

Fuente: Elaborado con base en Bosque y García (2000); Tong, Murray, Xiao y Hernández (2009); De la Fuente (2011); Alarcón (2012); Peña (2012); Díaz y Pineda (2013); Schneider, Seifert y Sunyaev (2014); Zamora (2015).

Tabla 2.3. Principales fuentes consultadas para la localización de servicios financieros y bancarios basados en análisis multicriterio

Técnica	Autor(es) y año	Criterios utilizados	Modelo utilizado	Software utilizado	Aportación a la investigación	País
Análisis multicriterio	Bosque, J. & Franco, S.,1995	Ambientales (hidrografía, litología), criterios que miden la eficiencia espacial de las localizaciones de instalaciones y los que consideran la posible interacción entre las instalaciones no deseables	Evaluación multicriterio	SIG[9] como un sistema de ayuda a la decisión espacial	Modelos de localización basados en la distancia (maximin, maxisum)	No aplica
		Demográficos, socioeconómicos, sector	AHP difuso y TOPSIS	No se indica	El artículo proporciona una referencia de los criterios que se toman en cuenta para elegir la mejor	

[9] Sistema de Información Geográfica funciona como una base de datos con información geográfica que se encuentra asociada por un identificador común a los objetos gráficos de un mapa digital.

	Cinar, N.,2009	productivo, competencia bancaria y potencial comercial			ubicación; sí como una propuesta de solución mediante métodos de Análisis Multicriterio.	Turquía
	Rodríguez, H., Peralta, I. & Delgado, D. 2015	Análisis de potencial de mercado (distribución, concentración, nivel de ingresos, etc.), análisis de competencia (comparación de mercados).	Cobertura de los servicios bancarios con ayuda del análisis espacial y el *geomarketing*	SIG ArcGis Network Analyst	Se determina si una ciudad tiene el potencial para cubrir espacios no proveídos por parte del sector bancario.	Toluca

Fuente: Elaborado con base en Bosque y Franco (1995); Cinar (2009); Rodríguez, Peralta y Delgado (2015).

Tabla 2.4. Principales fuentes consultadas para la localización de servicios financieros y bancarios basados en análisis de aprendizaje supervisado e ingreso de remesas

Técnica	Autor(es) y año	Criterios utilizados	Modelo utilizado	Software utilizado	Aportación a la investigación	País
Aprendizaje Supervisado	Soco M.,2017	35 variables geográficas y socioeconómicas, entre los que se encuentran las tasas de homicidios, el nivel de educación, el decomiso de armas largas.	Bosque Aleatorio (*Random Forest*)	No especifica	Modelo que predice en cuáles municipios en México hay mayores probabilidades de encontrar fosas clandestinas.	México
	James, G., Witten, D., Hastie, T. & Tibshirani R.,2013	No se indica, es un artículo que presenta teoría	Aprendizaje supervisado: problemas de clasificación y regresión	R (software libre)	Dichos modelos nos sirven para predecir sí un municipio es candidato o no lo es, para localizar sucursales bancarias con base en la información de las características de los municipios de México.	E.U.
Ingreso de remesas	Prior F.,2006	Remesas	Modelo propuesto de distribución de servicios financieros de bajo costo para segmentos de bajos ingresos	No se indica	Habla de todos los temas relevantes para la tesis de Microfinanzas.	México

Fuente: Elaborado con base en Soco (2017); Prior (2006); James, Witten, Hastie y Tibshirani (2013), Peralta y Delgado (2015).

Como podemos apreciar en este marco teórico, en los últimos quince años han surgido técnicas para resolver el problema de localización de servicios financieros y bancarios, se desarrollaron nuevas herramientas entre ellas: Sistemas de Información Geográfica como ArcGis, QGis, MapInfo, etc. Desde el área de Investigación de Operaciones, los Problemas de Localización de Máxima Cobertura han sido los modelos de optimización de redes más utilizados para la resolución de

localización de servicios financieros, ya que presentan un proceso que permite maximizar la cobertura de mercado potencial para los problemas de localización.

Por otra parte, el aprendizaje supervisado en los últimos años ha presentado gran importancia gracias a que tiene mayor precisión en sus modelos, así como sus tiempos cortos de ejecución y el volumen de datos que maneja, de tal forma, está técnica ha sido atractiva para dar solución a los problemas de localización.

Cabe mencionar que la búsqueda de literatura ayudó a identificar qué criterios se deben considerar para determinar sitios factibles para la ubicación de sucursales bancarias, con los cuales se han utilizado métodos de análisis multicriterio de correlaciones para ir sumando variables y así poder elegir las variables a utilizar en el modelo de localización.

La línea de investigación en la cual se acota esta tesis es la localización de sucursales bancarias. En esta línea de investigación, la tesis aporta un método para la localización de sucursales bancarias considerando el criterio de ingreso de remesas en los municipios de la República Mexicana, donde los estudios revisados anteriormente ninguno ha considerado este criterio en México.

El problema planteado está enfocado en determinar municipios factibles de México para la localización de sucursales bancarias de una Microfinanciera, dado el ingreso de remesas y mediante técnicas de aprendizaje supervisado para lograr la captación de personas mayores de edad económicamente activas, es decir clientes potenciales, considerando el número de clientes atendidos de la Microfinanciera.

Finalmente, con base en el problema a resolver, tenemos el objetivo general de la presente tesis:

Objetivo general:

Determinar los sitios alternativos para que la Microfinanciera sitúe sucursales bancarias en la República Mexicana considerando los criterios de número de clientes potenciales y clientes actuales, características geográficas (superficie en kilómetros), características poblacionales (población adulta, tipo de población), número de sucursales bancarias operando y monto de ingreso de remesas, para aumentar una mayor cobertura a nivel municipio utilizando técnicas de aprendizaje supervisado.

Capítulo 3. Aprendizaje de máquina o automático (*Machine learning*)

Con base en la literatura y el objetivo planteado, se describen los métodos y modelos que se utilizaron en esta tesis para la localización de sucursales bancarias.

Desde que nacemos vamos interactuando con las personas y medio que nos rodea, de esta manera nos vamos comportando con base en la repetición de acciones de nuestros allegados y a la vez vamos aprendiendo procesos que nos hacen comportarnos de una manera determinada. Todo ese aprendizaje lo realizamos de forma natural y automática, pero las computadoras o máquinas no lo hacen de esta manera, por lo que debemos de "enseñarles" mediante algoritmos de aprendizaje computacionales para que puedan encontrar dichos patrones de comportamiento en los datos. El aprendizaje automático es una rama de la inteligencia artificial o inteligencia computacional, por la cual las máquinas pueden aprender a realizar tareas sin que se les indique explícitamente cómo hacerlo (Bermejo, 2017).

El aprendizaje automático o aprendizaje de máquina (*machine learning*), es el subcampo de la computación y una rama de inteligencia artificial (AI) que proporciona y aporta a las computadoras la capacidad de aprender comportamientos basándose en reglas que presentan los datos para determinar un comportamiento final. Es también utilizado cuando las reglas dependen de demasiados factores y muchas de esas reglas necesitan ser ajustadas muy finamente, cuando hay muchos caminos para llegar a un comportamiento determinado, y es difícil programar una gran cantidad de reglas. Otro caso que se incluye en este campo es la situación en que la cantidad de datos utilizados es muy elevada (Bermejo, 2017).

El aprendizaje de máquina parte de información que es conocida como datos de entrada, que cuentan con una serie de características que sirven como insumo para el modelo, a partir de estos datos se detectan patrones de comportamiento, que no se advierten a simple vista, esté método permite tener una visión más amplia y profunda en poco tiempo acerca de la realidad que esa información contiene.

Para tener éxito al utilizar el aprendizaje de máquina es necesario entenderlo, conocer sus limitaciones. En este caso, para determinar si un municipio es candidato para la localización de sucursales bancarias a partir de datos demográficos, geográficos y estatus de la Microfinanciera, se puede entrenar al modelo a través de aprendizaje supervisado para que establezca una conexión con los datos y pueda llegar a identificar una nueva localización, con base en otros municipios donde anteriormente fue adecuado localizar sucursales bancarias ya que existen un volumen elevado de clientes y transacciones realizados por estos, y así poder predecir comportamientos (Gutiérrez, 2017).

En esta tesis, a partir de datos demográficos tales como:

- superficie de los municipios en kilómetros cuadrados;
- población (número de habitantes);
- población adulta (número de habitantes de edades entre 18 y 65 años);
- tipo de población (rural, urbana, etc.);
- total de sucursales bancarias;
- total cajeros automáticos;
- terminales punto de venta;
- transacciones en Terminales Punto y Venta;
- transacciones en cajeros automáticos;
- ingresos de remesas en millones de pesos;
- información del estatus de la empresa;
- número de sucursales bancarias de la Microfinanciera;
- número de sucursales de la empresa de la Microfinanciera;
- transacciones realizadas en las sucursales bancarias de la empresa;

se utilizará el aprendizaje de máquina para generar patrones, considerando dónde ha sido favorable localizar otras sucursales bancarias, que nos ayuden a identificar los sitios factibles para localizar sucursales bancarias nuevas.

3.1 Tipos de aprendizaje de máquina

Los diferentes tipos de aprendizaje de máquina se muestran en la tabla 3.1 y la figura 3.1 con base en Bermejo (2017) y Granville (2017).

En esta tesis se aplicará el algoritmo de aprendizaje supervisado de clasificación, debido a que se tiene como objetivo predecir a partir de ciertas características otras relacionadas con un elemento, en este caso, la localización de sucursales bancarias. Con estos algoritmos de clasificación se construyen reglas para obtener un modelo y sirve para calcular alguna característica cuyo valor se desconocía.

3.2 Algoritmo de bosque aleatorio (*Random forest*)

Bosque aleatorio es un algoritmo de aprendizaje supervisado de clasificación que a partir de datos de entrada calcula patrones para obtener un modelo, sirve después para calcular alguna característica desconocida. En la presente tesis esta característica que se desea conocer es la variable de localización de sucursales bancarias.

Tabla 3.1. Tipos de aprendizaje de máquina

Tipo de aprendizaje de máquina	Descripción
Aprendizaje supervisado	Es una técnica para deducir una función con base en datos de entrada y salida. El aprendizaje de máquina se encarga de encontrar el patrón.
Aprendizaje no supervisado	Es una técnica en dónde los datos no han sido etiquetados previamente y solo se dispone de datos de entrada, así el aprendizaje de máquina debe ser capaz de encontrar la estructura existente en los datos. Cómo ejemplo son los clusters.

Fuente: Elaborado con base en Bermejo (2017)

.

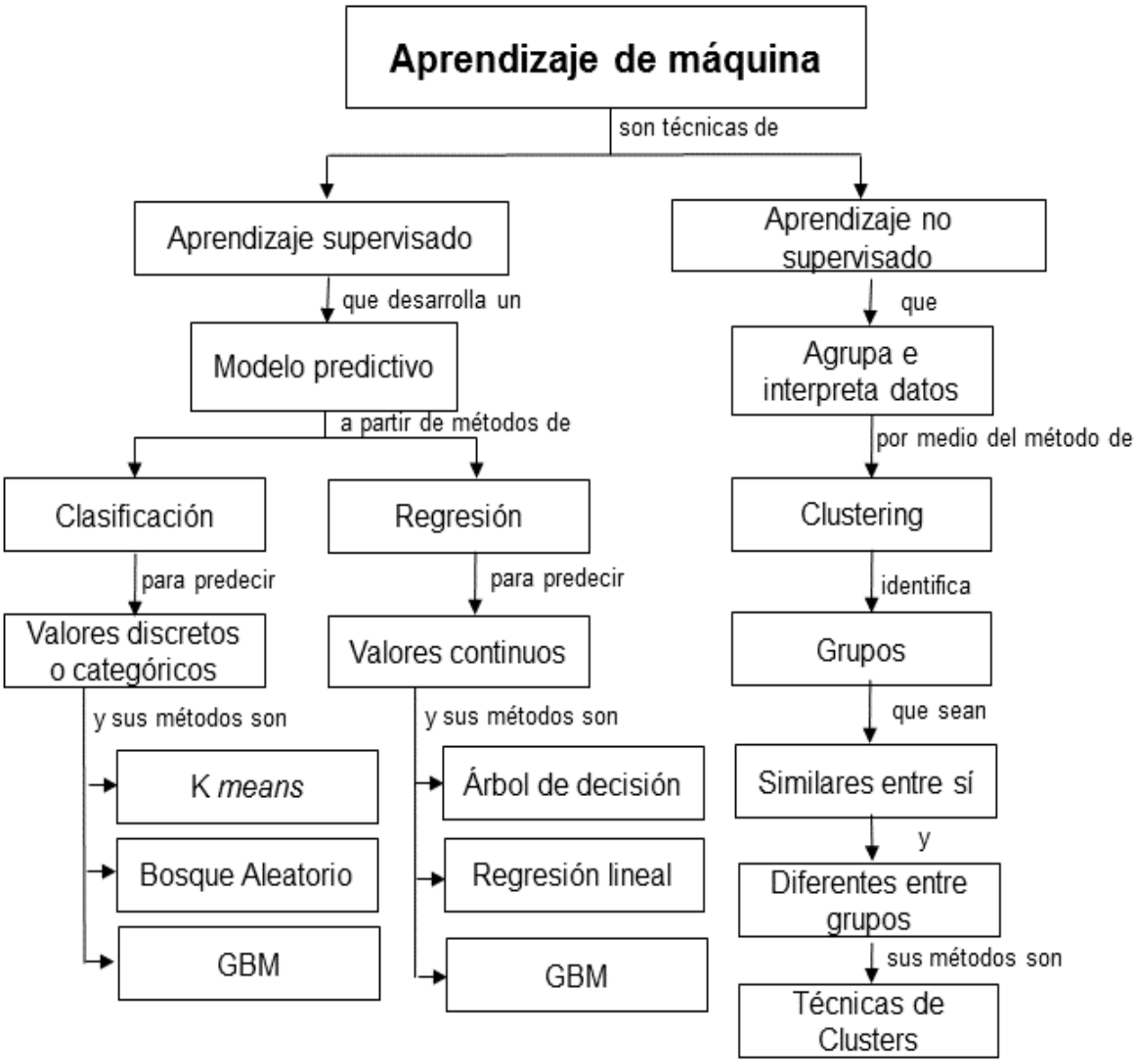

Figura 3.1. Mapa conceptual de tipos de aprendizaje supervisado y no supervisado
Fuente: Elaborado con base en Granville (2017).
Nota: GBM significa Gradient Boosting Machine o Máquina de aumento de gradiente.

El bosque aleatorio no está totalmente comprendido y aún sigue en investigación, se basa en la intuición y la heurística matemática. Fue sugerido por Breiman (2004), entre los años 1996 a 2004, quien demostró que se puede obtener mayor precisión en la clasificación y regresión usando ensambles de árboles, donde cada árbol crece de acuerdo con un parámetro aleatorio y se obtienen predicciones finales mediante agregación de árboles sobre el ensamble inicial. La variable objetivo puede ser una variable continua[10] o categórica[11] (Gutiérrez, 2017).

La popularidad del bosque aleatorio se debe a que es uno de los algoritmos de aprendizaje de máquina con fácil aplicación de árboles de decisión utilizado para la clasificación, también puede usarse para regresión (es decir, variable objetivo continua), pero se comporta especialmente bien en el modelo de clasificación (es decir, en la variable objetivo categórica). El método de bosque aleatorio se ha utilizado muy a menudo para refinar el modelo predictivo de clasificación.

El bosque aleatorio puede utilizarse para clasificar la importancia de las variables en un problema de regresión o clasificación que opera construyendo una gran cantidad de árboles de decisión individuales en la etapa de entrenamiento [12] (Bhalla, 2014). En cada árbol se forma primero, seleccionado aleatoriamente en cada nodo, un pequeño grupo de variables, también llamadas características, del conjunto de datos de entrenamiento.

En el bosque aleatorio, los procesos aleatorios son principalmente: las observaciones para “cultivar” cada árbol y las variables seleccionadas para la división en cada nodo.

El método bosque aleatorio produce resultados muy precisos en tiempos cortos y puede manejar un gran número de variables de entrada sin sobrecargarse, de hecho, aunque se genera una cierta pérdida de interpretabilidad, porque el modelo produce la variable respuesta sin saber la combinación que realizó para llegar a dicha respuesta, se considera una de las técnicas de aprendizaje más precisas que aumenta el rendimiento del modelo final de acuerdo con Bermejo, Vallejo y Alarcón (2017).

[10] Las variables continuas son variables numéricas que tienen un número infinito de valores entre dos valores cualesquiera.
[11] Las variables categóricas contienen un número finito de categorías o grupos distintos. Los datos categóricos pueden no tener un orden lógico. Por ejemplo: sexo, tipo de material, etc.
[12] Es el proceso en el que se detectan los patrones de un conjunto de datos. Una vez identificados los patrones, se pueden hacer predicciones con nuevos datos que se incorporen al sistema.

3.2.1 Operación del método bosque aleatorio

Se describe el proceso del bosque aleatorio dónde cada árbol se crea de la siguiente manera (Bhalla, 2014):

1. Selección aleatoria de registros: Se entrena a cada árbol con aproximadamente con el 63.2% de los datos totales. Esta elección de datos es aleatoria. Esta muestra de datos será el conjunto de entrenamiento para el desarrollo del árbol.

2. Selección aleatoria de variables: se seleccionan algunas combinaciones de variables al azar del total de variables predictoras, por ejemplo, si la cantidad total de variables es m, por defecto se tomará la raíz cuadrada de m. En particular, para un modelo de regresión, m es el número total de todos los predictores dividido por 3. El valor de m se mantiene constante durante el crecimiento del bosque.

3. Para cada árbol, usando los datos sobrantes (36.8%), se calcula la tasa de error de clasificación llamada *Out Of Bag* (OOB) o en español: fuera de la bolsa. Cabe mencionar que el aumento de la tasa de error del bosque aleatorio (OOB) es producto de la correlación entre dos árboles en el bosque. La reducción del número de variables aleatorias utilizadas en cada árbol disminuye la correlación y aumenta el número de variables hace que incrementa la correlación.

En la figura 3.2 se muestra un ejemplo donde se tienen 9 municipios y 4 variables predictores totales, entonces para los datos de entrenamiento se utilizarán 6 municipios que son el 63.2% de los municipios totales y 3 municipios se utilizarán para calcular la tasa de error. Dado que $\sqrt{4} = 2$, se utilizarán 2 variables para crear el modelo de bosque aleatorio.

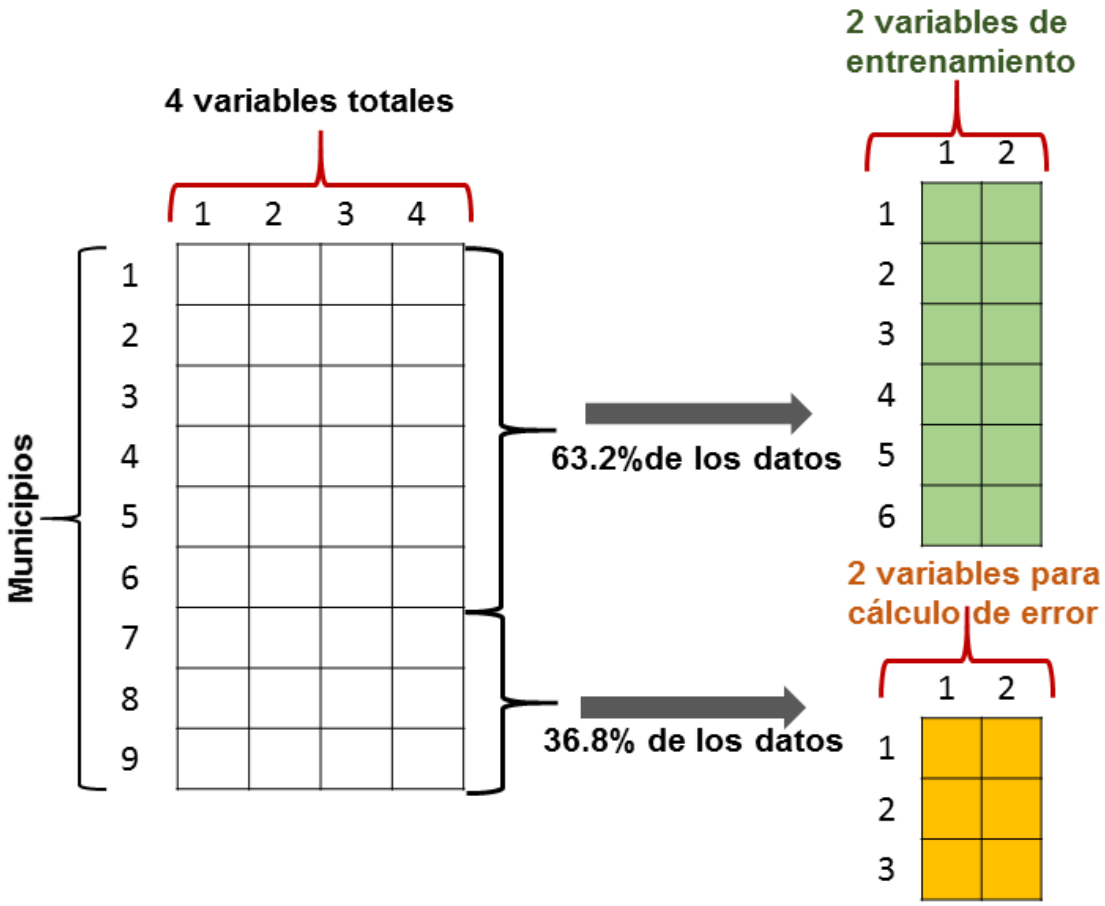

Figura 3.2. Ejemplo de clasificación de variables para el algoritmo de Bosque aleatorio
Fuente: Elaborado con base en Bhalla (2014).

En el bosque aleatorio no hay necesidad de separar el conjunto de registros de prueba del conjunto, para validar el resultado. Se separa internamente durante la ejecución, de la siguiente manera: a medida que el bosque se construye solo con datos de entrenamiento, cada árbol se prueba con la tercera parte de los datos de entrenamiento (36.8%) que no se usan en la construcción del árbol**.**

En el proceso de preparación de los datos para el modelo hay que tener cuidado con los conjuntos de datos desbalanceados de la variable de estudio, esto es, si una clase contiene significativamente más registros que la otra, por ejemplo, si la clase B tienen un número mucho mayor de registros que la clase A, es difícil crear conjuntos de datos de entrenamiento y prueba adecuados y se producirán resultados indeseables, por lo que se sugiere realizar con los datos una proporción en equilibrio, esto es 50% de los datos elegirlos de la clase A y el otro 50% de los

datos de la clase B, o bien la literatura recomienda (Espinar,2018) el 30% y el 70% o 25% y 75% de las clase A y B, respectivamente.

Para dar un ejemplo de un conjunto de datos desbalanceados tenemos el siguiente caso: se estudiará si un municipio es bueno o malo para la localización de sucursales bancarias, donde una clase son los municipios candidatos para la localización de sucursales y la otra clase son municipios que no son buenos candidatos para la localización de sucursales. De un total de 10,000 municipios el 90% del total de municipios no son buenos candidatos para la localización de sucursales y el 10% si lo son. Entonces se recomienda tomar el total de municipios buenos para la localización de sucursales bancarias, esto es, 1,000 municipios, en este ejemplo para aplicar un balanceo 50%-50% se tomarían de manera aleatoria de 1,000 municipios que no son buenos candidatos y finalmente se utilizaría una base de datos de 2,000 municipios.

3.2.2 Ajuste de bosque aleatorio

Para ajustar el algoritmo de bosque aleatorio hay dos parámetros importantes (Bhalla, 2014):

i. Número de variables aleatorias utilizadas en cada árbol, a este párametro se le llama: *mtry*; para encontrar la mejor calidad del número de variables se experimenta con la raíz cuadrada del número total de todos los predictores, la mitad del valor de la raíz cuadrada, y dos veces el valor de la raíz cuadrada. Se comprueba que el número de variables aleatorias utilizadas en cada árbol devuelve el área máxima bajo curva. Por lo tanto, si el número de predictores fueran 1,000 para cada nodo sería 16, 32 y 64 predictores, a seleccionar.

ii. Número de árboles utilizados en el bosque, a este parámetro se le llama: *ntree*: para encontrar el número de árboles que corresponden a un clasificador estable, construimos diferentes bosques aleatorios con diferente número de árboles (por ejemplo: 100, 200, 300, ..., 1,000).

3.3 Machine Learning con H2O

Se pretende mostrar cómo crear modelos de *machine learning* combinando H2O[13] y el lenguaje de programación R[14].

El manejo de H2O puede hacerse íntegramente desde R:, carga de datos, entrenamiento de modelos, predicción de nuevas observaciones, etc.

El H2O es un producto creado por la compañía H2O.ai con el objetivo de combinar los principales algoritmos de *machine learning* y aprendizaje estadístico con el *Big Data*[15]. Gracias a su forma de comprimir y almacenar los datos, H2O es capaz de trabajar con millones de registros en un único equipo de cómputo o en un clúster con procesadores de varios equipos. Los algoritmos de H2O están escritos en Java y *MapReduce*[16].

Aunque la principal ventaja de H2O frente a otras herramientas es su escalabilidad, sus algoritmos son igualmente útiles cuando se trabaja con un volumen de datos reducido.

3.3.1 Máquina de aumento de gradiente (*Gradient Boosting Machine*, GBM)

H2O incorpora los algoritmos de *Machine Learning*, entre sus modelos supervisados se encuentra el *Gradient Boosting Machine* (GBM).

GBM (para regresión y clasificación) es un método de aprendizaje avanzado. El GBM de H2O construye secuencialmente árboles de regresión con todas las

[13] Es un paquete para ejecutar H2O a través de su API (Interfaz de Programación de Aplicaciones que representa la capacidad de comunicación entre componentes de software) desde R. Para comunicarse con una instancia de H2O, la versión del paquete R debe coincidir con la versión de H2O.

[14] Es un lenguaje de programación con un enfoque al análisis estadístico y gráfico. Es un ambiente de programación formado por un conjunto de herramientas flexibles que pueden ampliarse mediante paquetes, librerías o definiendo propias funciones de usuario. También es gratuito y de código abierto.

[15] Big data es la gestión y análisis de volúmenes crecientes de datos, los cuales contienen una mayor variedad y que se presentan a una velocidad superior. Esto se conoce como "las tres V". Hace referencia a una inmensa y compleja colección de todo tipo de datos (estructurados, no estructurados, semi-estructurados, archivos de registro, imágenes, video, audio, etc.).

[16] Es un modelo de programación que da soporte a la computación paralela sobre grandes colecciones de datos en grupos de computadoras. *MapReduce* ha sido adoptado mundialmente, ya que existe una implementación *OpenSource* denominada *Hadoop*.

características del conjunto de datos de una manera completamente distribuida: cada árbol se construye en paralelo (GBM, 2018).

Boosting Machine (BM) es un tipo de modelo obtenido al combinar múltiples modelos sencillos (de regresión o clasificación), también conocidos como *weak learners*. Está combinación se realiza de forma secuencial, de manera que cada nuevo modelo que se incorpora al conjunto intenta corregir los errores de los anteriores. Como resultado de la combinación de múltiples modelos, BM consigue aprender relaciones no lineales entre la variable respuesta y los predictores. Si bien los modelos combinados pueden ser muy variados, H2O, al igual que la mayoría de librerías, utiliza como *weak learners* modelos basados en árboles (Amat, 2018).

GBM es una generalización del modelo de BM que permite aplicar el método de descenso de gradiente para optimizar cualquier función de costo durante el ajuste del modelo.

El valor predicho por un modelo GBM es la agregación (normalmente la moda en problemas de clasificación y la media en problemas de regresión) de las predicciones de todos los modelos individuales que lo forman.

3.3.2 Máquina de aumento de gradiente frente al bosque aleatorio

El bosque aleatorio y la Máquina de aumento de gradiente son métodos de aprendizaje en conjunto y predicen (realizando regresión o clasificación) mediante la combinación de las salidas de árboles individuales. Difieren en la forma en que se construyen los árboles: el orden y la forma en que se combinan los resultados (Ravanshad, 2018).

Bosque aleatorio entrena cada árbol de forma independiente, utilizando una muestra aleatoria de los datos. Esta aleatoriedad ayuda a que el modelo sea más robusto que un solo árbol de decisión, y es menos probable que se sobreajuste en los datos de entrenamiento. Normalmente hay dos parámetros en el Bosque Aleatorio: número de árboles y número de características que se seleccionarán en cada nodo.

La Máquina de aumento de gradiente crea árboles de uno en uno, donde cada árbol nuevo ayuda a corregir los errores cometidos por árboles entrenados previamente. Con cada árbol agregado, el modelo se vuelve aún más expresivo. Normalmente hay tres parámetros: número de árboles, profundidad de árboles y velocidad de aprendizaje.

El entrenamiento de la Máquina de aumento de gradiente generalmente lleva más tiempo debido al hecho de que los árboles se construyen secuencialmente. Sin embargo, los resultados de referencia han demostrado que esta técnica es mejor

aprendiz que el Bosque Aleatorio. Aunque puede parecer que la Máquina de aumento de gradiente es mejor que los bosques aleatorios, también son propensos a sobreajuste, esto es, cuando el modelo se adapta bien al conjunto de datos de entrenamiento, pero falla al conjunto de datos de validación.

En términos de objetivo de entrenamiento, la Máquina de aumento de gradiente intenta agregar árboles nuevos que complementan los ya construidos. Esto normalmente le da una mejor precisión con menos árboles.

3.4 Selección de variables importantes

Antes de comenzar el proceso de construir el modelo, se realiza la selección de variables que es un paso esencial en un proyecto de modelado predictivo. También se llama selección de características. No todas las variables que se tienen en la base de datos para realizar el modelo son importantes para la predicción, por lo tanto, es básico identificar variables importantes y eliminar las variables redundantes.

Con base en Bhalla (2017) se describen algunos puntos acerca de la importancia de la selección de variables:

- eliminar una variable redundante ayuda a mejorar la precisión, del mismo modo, la introducción de una variable relevante tiene un efecto positivo en la precisión del modelo;
- demasiadas variables podrían resultar en sobrecapacidad lo que significa que el modelo no es capaz de generalizar el patrón;
- demasiadas variables conducen a la computación lenta que a su vez requiere más memoria y *hardware*.

Una manera de seleccionar las variables es utilizando el paquete Boruta[17] de R, los motivos para utilizarlo son:

1. funciona bien para el problema de clasificación y regresión;
2. toma en cuenta las relaciones multi-variables;
3. es una mejora en la medida de variables de importancia en el Bosque Aleatorio, y es un método muy popular para la selección de variables;
4. sigue un método de selección de variables relevantes que considera todas las características que son significativa para la variable de resultado.

[17] Es un paquete para seleccionar de características relevantes, capaz de trabajar con cualquier método de clasificación que arroje la medida de importancia variable; por defecto, Boruta usa Bosque Aleatorio.

3.4.1 Algoritmos para la selección de variables

La idea del uso de los algoritmos de selección de variables es lograr el mejor modelo en forma secuencial, incluyendo o excluyendo una sola variable predictora en cada paso de acuerdo con ciertos criterios (González, 2015).

Se describen tres de los algoritmos más usados:

- Métodos *Forward* (Selección hacia adelante): se parte de un modelo sencillo y se van agregando términos con algún criterio, hasta que no procede añadir ningún término más, es decir, en cada etapa se introduce la variable más significativa hasta que se cumple una cierta regla de parada, esto es, la primera variable que se introduce es la de mayor correlación (ya sea negativo o positiva) con la variable dependiente. A continuación, se considera la variable independiente cuya correlación parcial sea la mayor y que no esté en la ecuación. El procedimiento termina cuando ya no quedan variables que cumplan el criterio de entrada.
- Métodos *Backward* (Eliminación hacia atrás): se parte de un modelo complejo, que incorpora todos los efectos que pueden influir en la respuesta, y en cada etapa se elimina la variable menos influyente, hasta que no procede suprimir ningún término más.
- Métodos *Stepwise*: este procedimiento es una combinación de los dos anteriores. Comienza como el de introducción progresiva, pero en cada etapa se plantea si todas las variables introducidas deben de permanecer en el modelo.

Cuando se aplica este tipo de procedimientos tenemos que tener muy claro cuál será la condición para suprimir o incluir un término. Para ello, podemos considerar dos criterios llamados: criterios de significación del término y criterios de ajuste global.

- Criterios de significación: En un método *backward* se suprimirá el término que resulte menos significativo, y en un método *forward* se añadirá aquel término que al ser añadido al modelo resulte más significativo.
- Criterios globales: En lugar de usar la significación de cada coeficiente, podemos basarnos en un criterio global, una medida global de cada modelo, de modo que tenga en cuenta el ajuste y el exceso de parámetros. Se elige el modelo cuya medida global sea mejor, como criterios destacamos el Criterio de Información de *Akaike* (AIC) y el Criterio de información de Bayes (BIC), donde se trata de buscar un modelo cuyo AIC o BIC sea pequeño para elegir entre varios modelos.

3.5 Matriz de Confusión

La matriz de confusión es una herramienta que permite la visualización del desempeño de un algoritmo que se emplea en aprendizaje supervisado. Cada fila de la matriz representa la cantidad de predicciones de cada clase, mientras que cada columna representa los casos en la clase real. Uno de los beneficios de las matrices de confusión es visualizar el número de casos que el modelo está clasificando asertivamente y el número de casos que no clasifica bien. En esta tesis tenemos 4 diferentes escenarios cuando se realiza la comparación de la categoría real del municipio con la clasificación que nos arroja el modelo Bosque Aleatorio. Los casos que clasificaría asertivamente el modelo es cuando un municipio tiene la categoría de "bueno" para la localización de sucursales bancarias y el modelo también lo clasifica como bueno, o cuando el municipio es "malo" para la localización de sucursales bancarias y el modelo lo califica como tal. Caso contario cuando el modelo no clasifica adecuadamente sucede cuando el modelo califica al municipio como bueno pero la categoría real es malo para la localización de sucursales o cuando el modelo nos clasifica al municipio como malo para la localización de sucursales bancarias pero en los datos reales es bueno.

El rendimiento de estos modelos de aprendizaje supervisado es comúnmente evaluado con base en está matriz de confusión, ver tabla 3.2.

Tabla 3.2. Matriz de Confusión

		Valor Real	
		Malo	**Bueno**
Valor de Predicción	**Malo**	Verdaderos Negativos (VN)	Falsos Negativos (FN)
	Bueno	Falsos Positivos (FP)	Verdaderos Positivos (VP)

Fuente: Elaborado con base en James *et al. (2013).*

La tabla 3.3. es la matriz de confusión aplicada en este caso a la localización de sucursales bancarias en los municipios de México, dónde las etiquetas de clasificaciones de los municipios son "malo" o "bueno" para la localización de sucursales bancarias.

Tabla 3.3. Matriz de Confusión para la localización de sucursales bancarias por municipio

Valor de Predicción	Valor Real: Malo	Valor Real: Bueno
Malo	El modelo clasifica un municipio como "malo" para la localización de sucursales bancarias y realmente el municipio es "malo".	El modelo clasifica un municipio como "malo" para la localización de sucursales bancarias pero realmente es un municipio con categoría "bueno".
Bueno	El modelo clasifica un municipio como "bueno" para la localización de sucursales bancarias pero realmente es un municipio con categoría "malo".	El modelo clasifica un municipio como "bueno" para la localización de sucursales bancarias y realmente el municipio es "bueno".

Fuente: Elaborado con base en James *et al. (2013).*

Para saber si el municipio es realmente bueno se consideró con base en la teoría citada factores dónde ha sido favorable la localización de sucursales bancarias y criterios significantes que utiliza la Microfinanciera, entre ellos ingresos de remesas, número de habitantes mayores de edad económicamente activos, número de clientes atendidos por la Microfinanciera.

Las matrices de confusión son muy empleadas en problemas de clasificación y consisten en una tabla que indica cuántas clasificaciones se han hecho para cada tipo, la diagonal de la matriz representa las clasificaciones correctas.

La terminología y sus derivados a partir de una matriz de confusión son:

Exactitud = *Accurary* (ACC): mide la cantidad de éxito, es decir los municipios clasificados para la localización de sucursales bancarias asertivamente por el modelo comparado con el valor real, entre todos los municipios, evalúa la efectividad general de un clasificador.

$$ACC = \frac{VN + VP}{VN + VP + FN + FP} \quad \dots (3.1)$$

Sensibilidad o razón de verdaderos positivos = *Sensitivity* (VPR): es el número de municipios que el modelo predice como buenos candidatos para la localización de

sucursales bancarias y realmente lo son, entre todos los municipios que realmente son buenos para localizar sucursales bancarias.

$$VPR = \frac{VP}{VP + FN} \qquad \dots (3.2)$$

Especificidad o razón de verdaderos negativos = *Specificity* (SPC) : es el número de municipios que el modelo predice como malos candidatos para la localización de sucursales bancarias y realmente lo son, entre todos los municipios que realmente son malos para localizar sucursales bancarias.

$$SPC = 1 - FPR = \frac{VN}{VN + FP} \quad \dots (3.3)$$

El parámetro con el que se medirá la efectividad del modelo depende del criterio que se utilizará para medir los resultados. En general se utiliza la exactitud, pero se recomienda incluir la sensibilidad y/o la especificidad.

3.5.1 Curva ROC

Una curva ROC (*Receiver Operating Characteristic* o Característica Operativa del Receptor) es una representación gráfica de la sensibilidad frente a la especificidad. La sensibilidad y especificidad se correlacionan de forma inversa, y a cada valor de sensibilidad corresponde un valor de especificidad, por lo que se pueden ilustrar formando una curva ROC.

Para dibujar una curva ROC sólo son necesarias las razones de Verdaderos Positivos (VPR) y de Falsos Positivos (FPR). La VPR mide hasta qué punto un clasificador o prueba diagnóstica es capaz de detectar o clasificar los casos positivos correctamente, de entre todos los casos positivos disponibles durante la prueba. La FPR define cuántos resultados positivos son incorrectos de entre todos los casos negativos disponibles durante la prueba.

Dado que VPR es equivalente a sensibilidad y FPR es igual a 1 menos la especificidad o razón de verdaderos negativos, cada resultado de predicción o instancia de la matriz de confusión representa un punto en el espacio ROC.

El mejor método posible de predicción se situaría en un punto en la esquina superior izquierda, o coordenada (0,1) del espacio ROC, representando un 100% de sensibilidad (ningún falso negativo) y un 100% también de especificidad (ningún

falso positivo). A este punto (0,1) también se le llama una clasificación perfecta véase la figura 3.3. Por el contrario, una clasificación totalmente aleatoria (o adivinación aleatoria) daría un punto a lo largo de la línea diagonal, que se llama también línea de no-discriminación, desde el extremo inferior izquierdo hasta la esquina superior derecha. Un ejemplo: lanzar una moneda al aire, a medida que el tamaño de la muestra aumenta, el punto de un clasificador aleatorio de ROC se desplazará hacia la posición (0.5, 0.5).

En la figura 3.3 la diagonal divide el espacio ROC. Los puntos por encima de la diagonal representan los buenos resultados de clasificación (mejor que el azar), puntos por debajo de la línea de los resultados pobres (peor que al azar).

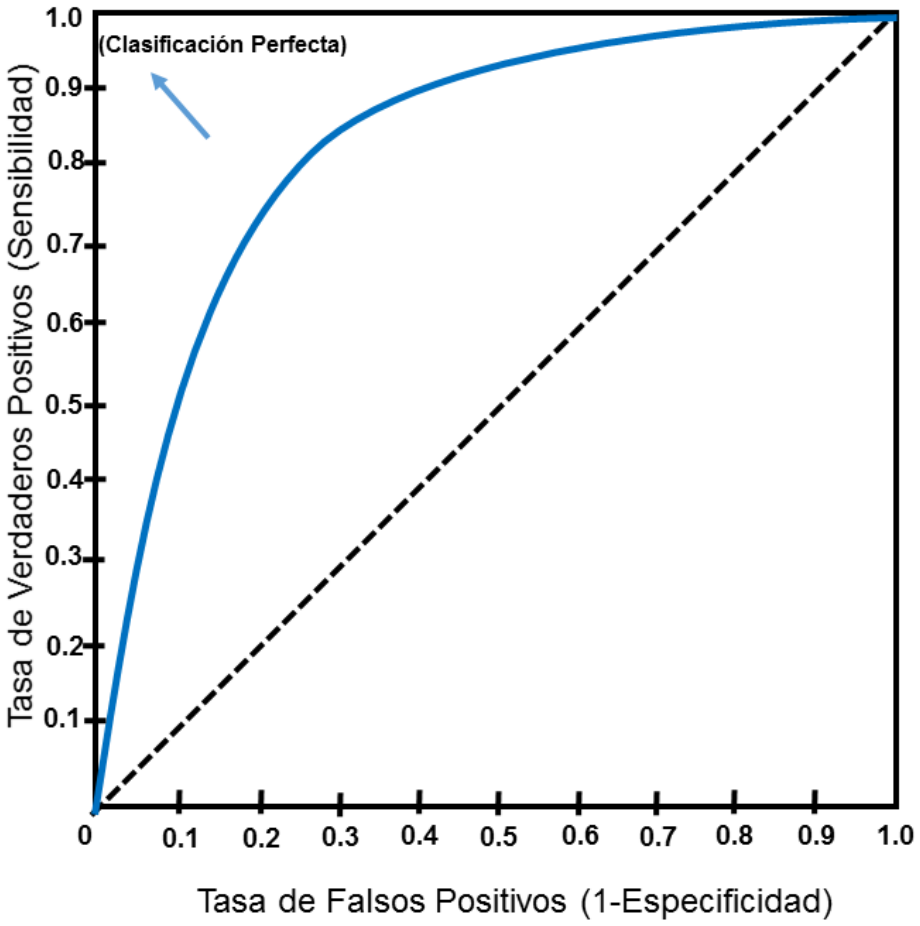

Figura 3.3. Curva ROC
Fuente: Elaborado con base en James (2009).

3.6 Terminología relacionada con el algoritmo Bosque aleatorio

En la tabla 3.4 se describe la terminología básica utilizada para los algoritmos de clasificación:

Tabla 3.4. Terminología relacionada con los algoritmos

Término	Significado
Datos de entrenamiento	Un subconjunto de datos totales, etiquetados con el valor de la variable objetivo y usados como entrada del algoritmo de aprendizaje para producir el modelo.
Datos de prueba	Una porción de los datos de entrenamiento con el valor de la variable objetivo ocultos para que puedan ser usados para evaluar el modelo.
Entrenamiento	El proceso de aprendizaje que usa los datos de entrenamiento para producir un modelo. Este modelo puede estimar la variable objetivo dada por las variables predictoras como entrada.
Variable	En este contexto, es el valor de una característica o una función de varias características.
Registro	Un contenedor donde se almacena un ejemplo. Está compuesto de campos.
Campo	Parte de un registro que contiene el valor de una característica (variable).
Variable predictora	Una característica seleccionada para usarla como entrada al modelo de clasificación. No todas las características necesitan ser usadas. Algunas pueden ser combinaciones algorítmicas de otras.
Variable objetivo	Una característica que el modelo de clasificación intenta estimar, es una variable categórica y determina lo que busca el sistema de clasificación.
Sobreajuste o *Overfitting*	Ocurre cuando el modelo aprende los datos de entrenamiento de memoria en lugar de aprender los patrones que evitan que se pueda generalizar a los datos de prueba. Significa que el modelo se adapta bien al conjunto de datos de entrenamiento, pero falla al conjunto de datos de validación.
Error fuera de la bolsa	Es equivalente a los datos de validación o prueba. En bosques aleatorios no hay necesidad de un conjunto de prueba por separado para validar el resultado. Se estima internamente, durante la ejecución de la siguiente manera: como el bosque se basa en datos de entrenamiento, cada árbol se prueba con la tercera parte de los datos (36.8%) que no se usaron en la construcción de ese árbol (similar al conjunto de datos de validación).

Fuente: Elaborado con base en Laurentio *et al. (2014).*

Enfocándonos en los datos de entrada que utilizará el modelo, en la taba 3.5 se describen los diferentes tipos de datos que pueden tomar las variables de un modelo. Las variables pueden ser: numéricas, texto, etc.

Tabla 3.5. Tipo de variables

Tipo de dato	Descripción
Continua	Es un valor en punto flotante. Podría representar un precio, una talla.
Categórica	Puede tener un valor dentro de un conjunto especificado de valores. Normalmente se trata de un conjunto reducido de valores, como mínimo de dos, aunque a veces puede ser bastante grande. Los booleanos son generalmente tratados como categóricos, otro ejemplo puede ser un ID de vendedor.
Texto	Es una secuencia de valores "palabra", todos del mismo tipo. Un texto es un claro ejemplo de estos valores, pero una lista de correos electrónicos o URLs puede serlo también.

Fuente: Elaborado con base en Laurentio, *et al. (2014).*

Cabe mencionar que se tiene que revisar si los valores son continuos o categóricos, ya que es muy común que se trate números de identificación como continuos cuando son categóricos, debido a que son un conjunto de valores ya especificados cómo por ejemplo la talla de la ropa (talla 28, 30, 32, 34, etc.) que son ya valores determinados (Gutiérrez, 2017).

Asimismo, es necesario el conocimiento del objetivo para la realización del modelo, ya que, si los datos utilizados y el problema a resolver son diferentes en cada caso, implicará obtener resultados erróneos respecto al modelo y al objetivo. También se necesita la recolección de datos, limpieza de datos y pre-procesado de los mismos para que el algoritmo aprenda adecuadamente y después de estos pasos se procede a iniciar la etapa del modelado.

Una vez que se logra encontrar el modelo suficientemente bueno, se evalúa cómo funciona y si los resultados son los esperados, se pone a prueba, se pueden añadir nuevos datos, se limpian y se va perfeccionando el modelo. En cada uno de los modelos se utilizan diferentes parámetros.

3.7 Definición del sistema

Para un mejor entendimiento del procesamiento de la Base de Datos a utilizar en la de generación del modelo, se describen las diferentes etapas que se realizaron para su obtención.

Del lado izquierdo de la figura 3.4 muestra la base total de datos donde cada fila o registro corresponde a cada uno de los municipios de México. Cada columna representa alguna característica demográfica, geográfica, información de la Microfinanciera como es el número de clientes actuales y número de sucursales, así como la variable target para la localización de sucursales bancarias o indicador calculado con base en el ingreso de remesas, clientes de la empresa y población de cada uno de los municipios.

Enseguida se divide aleatoriamente en dos bases: base de entrenamiento la cual contiene el 70% de los datos seleccionados al azar y base de prueba con el 30% de los datos restantes.

Figura 3.4. Base de Datos división de base de entrenamiento y prueba
Fuente: Elaborado con base en Bhalla (2014)

Se utilizará los datos de la base de entrenamiento para generar un modelo manejando técnicas del aprendizaje supervisado, se elegirá el modelo con mayor asertividad en la localización de sucursales bancarias utilizando los datos de la base de prueba, así al tener el modelo final se ingresan al modelo nuevos datos dónde se calcula el indicador que nos dice si es un municipio es bueno o malo para la localización de sucursales bancarias y podamos tomar decisiones con base en los resultados del modelo ver figura 3.5.

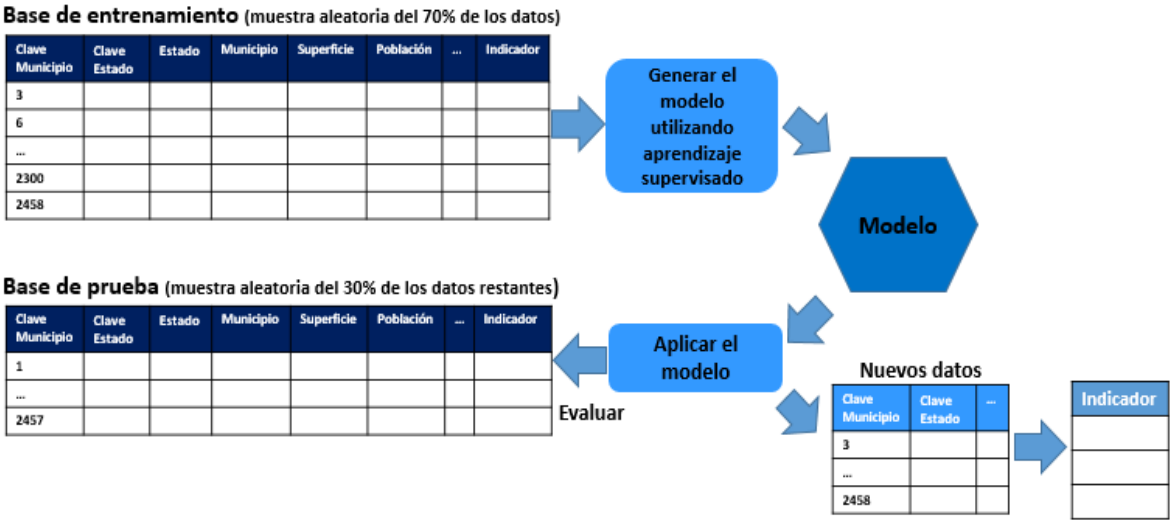

Figura 3.5. Base de Datos división de base de entrenamiento y base de prueba
Fuente: Elaborado con base en Bhalla (2014)

Utilizando la base de datos de entrenamiento (el 70% de datos de la base total), se realizan muestras aleatorias para la detección de patrones del conjunto de datos y una vez identificados los patrones, se pueden hacer predicciones con nuevos datos que se incorporen al modelo. Se utilizan diferentes números de variables (ejemplos: 2, 4, 8) cabe mencionar que el indicador siempre tiene que pertenecer a estas variables. Se elige también el número de árboles a utilizar en el bosque aleatorio (ejemplos: 50, 150, 200, 500, 1000, 2000), finalmente se selecciona el modelo que tenga mayor precisión de asertividad en la localización de sucursales bancarias ver figura 3.6.

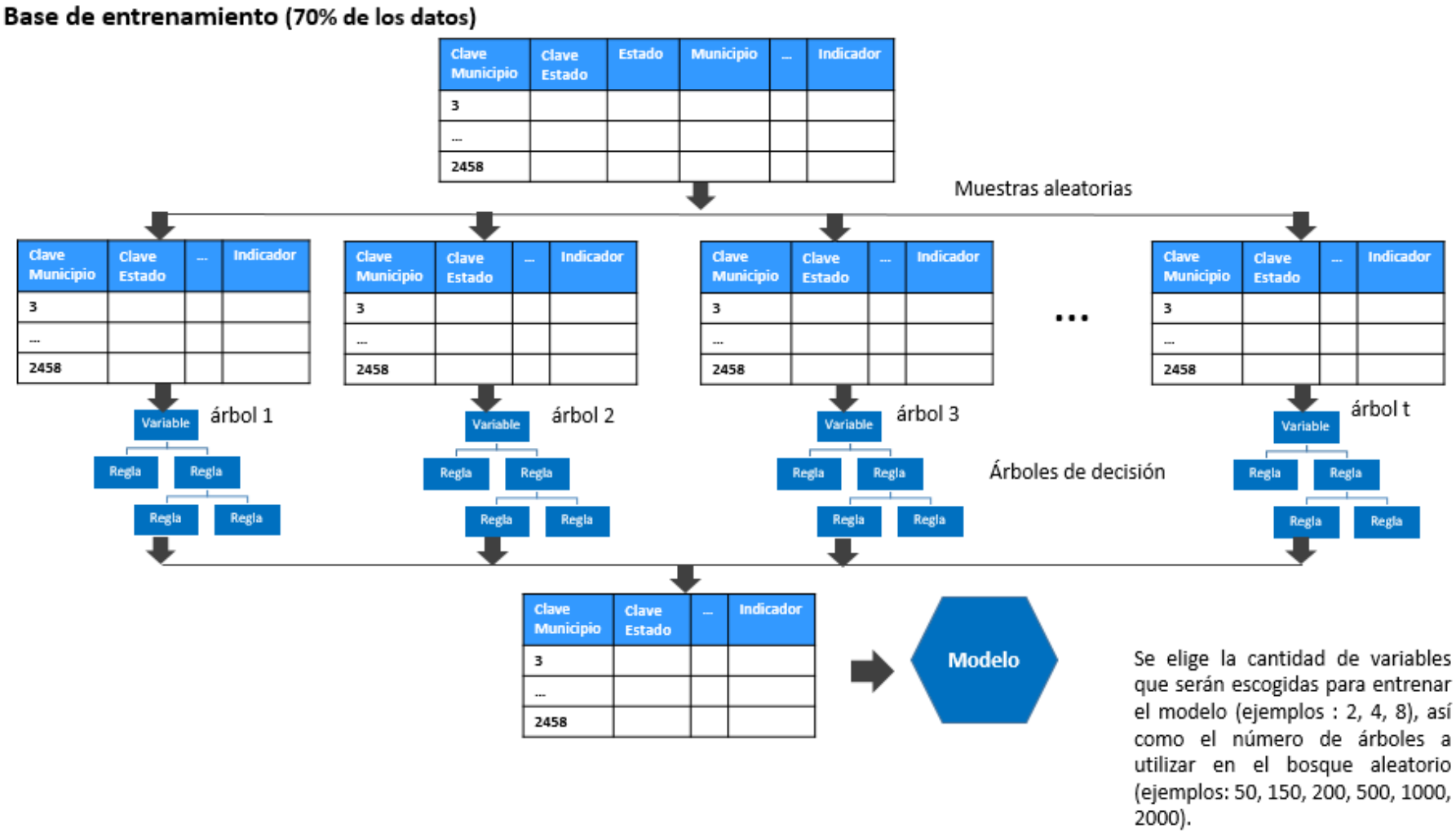

Figura 3.6. Mapa conceptual de modelos de localización encontrados en la revisión bibliográfica
Fuente: Elaborado con base en Bhalla (2014)

Capítulo 4. Metodología propuesta

En esta sección se describe la estrategia para resolver el problema de investigación, esto es, generar un modelo que determine los sitios viables para la localización de sucursales bancarias de la Microfinanciera. Se explica cada una de las etapas de la estrategia, y finalmente las herramientas usadas en cada una de las etapas que nos ayudarán a lograr nuestro objetivo.

4.1 Elección de la metodología utilizada

Es necesario tener una metodología para poner orden y orientación a una serie de herramientas para la solución del problema en la tesis.

Existen tres metodologías dominantes para el proceso de la minería de datos son: la metodología SEMMA propuesta por SAS[18] por sus siglas cuyo significado son *Sample, Explore, Modify, Model, Assess*, en español, Muestra, Explora, Modificación, Modela y Evalúa; CRISP-DM (*Cross-Industry Standard Process for Data Mining*) desarrollado por algunos líderes de la industria: IBM[19], SAS, y la metodología KDD propuesta por Fayyad en 1996, propone 5 fases: Selección, Preprocesamiento, Transformación, Minería de datos y Evaluación e implantación.

Para elegir la metodología utilizada en este problema de investigación se compararon las tres mencionadas anteriormente. Se utilizó la metodología CRISP-DM ya que se apega más que las otras dos en la presente tesis, debido a que es una metodología orientada a objetivos empresariales como es el caso de este proyecto donde se tiene un objetivo concreto que es determinar sitios alternativos para que la Microfinanciera situé sucursales bancarias para aumentar una mayor cobertura a nivel municipio, asimismo se analiza el objetivo o problemática a resolver. Además, cuenta como primera fase el análisis del problema a resolver, que en esta tesis se describe en el Capítulo 1, así como la parte de evaluación e implementación del modelo final cuya etapa de esta metodología se utilizó para evaluar la efectividad del modelo.

El método CRISP-DM, está enfocado en orientar proyectos de minería de datos también conocido *Data Mining* (DM), incluye una metodología estructurada en seis fases. Algunas de estas fases son bidireccionales, lo que significa que permiten revisar parcial o totalmente las fases anteriores. Está metodología se utilizará para

[18] Es una compañía líder de software y servicios de Análisis de Negocio.

[19] Es una compañía estadounidense privada que provee soluciones de hardware, entre los que se incluyen computadoras portátiles y de escritorio, así como software, servicios financieros y una amplia gama de servicios de tecnología de información.

resolver el problema la localización de sucursales bancarias de la Microfinanciera, la cual considera las siguientes 6 fases ver figura 4.1:

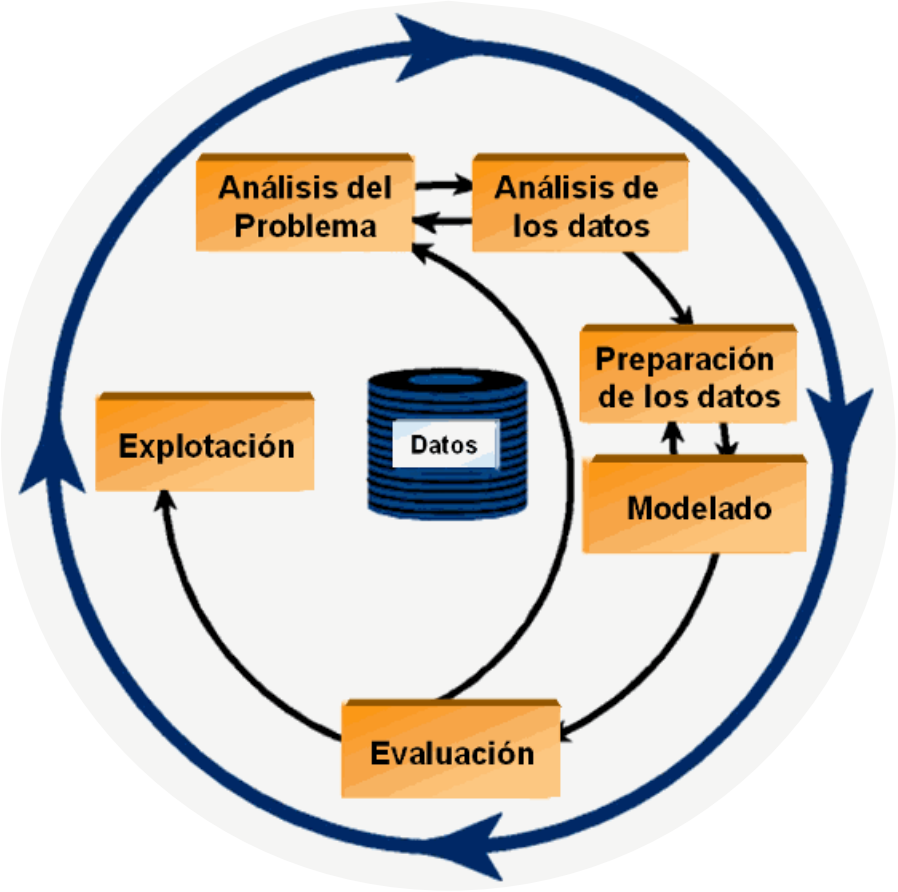

Figura 4.1. Modelo de CRISP DM
Fuente: IBM (International Business Machines) (2012).

4.1.1 Fase 1: Análisis del problema

Se comienza con entender el problema que se desea resolver, lo cual permite recolectar los datos correctos e interpretar adecuadamente los resultados.

En el Capítulo 1 se explica ampliamente cuál es el problema a resolver, seguido del objetivo general en esta presente tesis.

4.1.2 Fase 2: Análisis de los datos

En esta fase se examinan los datos que serán utilizados para la resolución del problema, es decir, explorar los datos con la ayuda de tablas de frecuencia y gráficos para conocerlos, determinar su calidad e identificarlos. Esta fase es elemental para evitar problemas inesperados durante la siguiente que es la preparación de datos, suele ser la fase más larga de un proyecto.

Se recomienda crear una nueva base de datos para el proyecto, debido a que se realizan varias modificaciones en los datos, además de que la información se recolecta de varias fuentes y es más apropiado agruparlos en una sola base.

Se describen las tareas a desarrollar en esta fase:

- Descripción de los datos

En este proceso se muestra la calidad y cantidad de los datos adquiridos, es decir, establecer volúmenes de datos (número de registros y campos), el significado de cada campo y la descripción.

En la presente tesis se utilizará una base de datos que contiene información de 2,458 municipios de la República Mexicana.

Cada registro representa un municipio y está comprendido de los siguientes campos:

- Cajeros;
- Clave Estado;
- Clave Municipio;
- Clientes a junio 2018;
- Clientes a septiembre 2018;
- Depósitos y Pagos realizados en otros canales;
- Estado;
- Grupos de la Microfinancieras junio 2018;
- Indicador conjunto;
- Ingreso Remesas mdd 2do trimestre 2018 (dónde mdd es millones de pesos);
- Ingreso Remesas mdd 3er trimestre 2018;
- Municipio;
- Población;
- Población adulta;
- Región;
- Sucursales de la empresa remesadora;
- Sucursales de la Microfinanciera;
- Superficie en kilómetros cuadrados;

- Terminales Punto de Venta;
- Tipo Población;
- Total de sucursales de las instituciones financieras;
- Transacciones en Cajeros Automáticos;
- Transacciones en TPV (Terminal Punto de Venta).

La variable llamada “indicador conjunto” es la variable objetivo que clasifica a cada municipio del país en dos clases que son: “bueno” y “malo” para la localización de sucursales bancarias, así la Microfinanciera puede enfocarse en los municipios que sean buenos para seleccionar el sitio de localización de sus sucursales bancarias.

Las reglas consideradas para etiquetar a los municipios son:

i. Bueno: cuando el municipio tiene un monto de ingresos de remesas es por arriba del tercer cuartil[20], incluyendo el número de habitantes mayores de edad económicamente activos sea mayor a 10,000 y el número de clientes atendidos de la Microfinanciera sea por mayor del primer cuartil, estos últimos dos factores son criterios que ya utiliza la empresa para la localización de sucursales bancarias asimismo son factores que con base en la teoría citada por Cinar (2009) se utilizan para la localización de sucursales bancarias.

ii. Malo: cuando el municipio presenta monto de ingresos de remesas por debajo del tercer cuartil, incluyendo el número de habitantes mayores de edad económicamente activos sea menores a 10,000 y el número de clientes actuales de la Microfinanciera sea menor al primer cuartil.

Cabe mencionar que el modelo al predecir si un municipio es bueno o malo involucra la importancia de las variables predictoras (variables de entrada) tales como son población, población adulta, sucursales de la Microfinanciera, superficie en kilómetros cuadrados, tipo de población, etc. con respecto al indicador conjunto descrito anteriormente.

- Exploración de los datos

Realizar un análisis descriptivo que muestre propiedades en los datos, cómo son tablas de frecuencia y realizar gráficos de distribución. En esta tesis se graficó la distribución de número de municipios con respecto al nivel del monto de ingreso de remesas, en donde se puede observar el rango de monto de ingreso de remesas donde se concentra la mayoría de los municipios.

- Verificación de la calidad de los datos

[20] Es un número debajo del cual se encuentran las tres cuartas partes de los datos numéricamente ordenados.

Frecuentemente la base de datos contiene valores nulos, es decir, no contienen información, algunos son errores de captura, valores perdidos u otro tipo de contradicciones que hacen que los análisis resulten difíciles y no se obtenga la respuesta final apropiada. Una forma de evitar estos posibles problemas es realizar un análisis de calidad de los datos disponibles antes de comenzar con la etapa de modelado.

Esta tarea se utiliza para encontrar observaciones que son numéricamente apartadas del resto de los datos, los cuales pueden constituirse en ruido para el proceso, valores que no poseen información, vacíos o codificados sin respuesta, cómo por ejemplo: null , ?, NA, etc., su objetivo es asegurar la corrección de los datos y calidad de los mismos.

4.1.3 Fase 3: Preparación de los datos

Se estima que esta fase de preparación de datos suele tomar del 50% a 70% del tiempo y esfuerzo de un proyecto. Se procede a realizar correlaciones entre las variables y distribuciones de las variables para la exploración de los datos.

Esta fase tiene mucha relación con la de modelado ya que, con base en la técnica de modelado elegida, los datos requieren ser procesados de diferentes formas.

La preparación de datos suele implicar las tareas siguientes:

i. Selección de los datos

Se recolectó información de los 2,458 municipios de la República Mexicana y se generó una base de datos de dimensiones 2,458 registros con 18 características, este archivo se dividió en dos, uno para entrenar y otro para probar el entrenamiento con los modelos que presenten mejor desempeño en la localización de sucursales bancarias.

ii. Estructuración de los datos

Esta tarea realiza cálculos con respecto a los datos tales como la generación de nuevos campos a partir de campos ya existentes, es este caso la variable categórica *tipo de población* se transformó en variable numérica, es decir se utilizó la transformación de valores para campos existentes.

En lo referente a la estructuración de los datos, cada registro de la tabla es considerado como un patrón (o ejemplo de entrenamiento), por lo tanto, el modelo de minería tendrá que aprender, para identificar si el municipio es candidato o no para la localización de sucursales bancarias.

iii. Integración de los datos

Para el desarrollo de este proyecto, se dispone de una base de datos que se integró de diferentes fuentes como son: Comisión Nacional Bancaria y de Valores, Consejo Nacional de Inclusión Financiera, Banco de México e información de la Microfinanciera.

iv. Formateo de los datos

Radica principalmente en la realización de transformaciones de los datos sin modificar su significado, por ejemplo, se puede convertir una variable categórica en numérica o viceversa.

En el presente proyecto consideramos utilizar únicamente las variables numéricas para entrenar los modelos y las variables categóricas se transformaron a numéricas, como es el caso de la variable *tipo de población*. Así, no fue necesario realizar algún ajuste de los valores de los campos como eliminar datos, caracteres especiales, máximos y mínimos, etc.

En esta tesis se generó una base de datos, de la cual se creó un archivo que contiene 23 campos y 2,458 registros, donde cada registro representa un municipio y cada campo una característica. La tabla 4.1 muestra las variables predictoras utilizadas para entrenar el modelo, las cuáles se eligieron con base en la revisión teórica-metodológica y los criterios utilizados por la Microfinanciera. Cabe mencionar que el modelo utilizado realiza combinaciones internas entre las variables para seleccionar el de mayor exactitud en la localización de sucursales.

Tabla 4.1 Descripción de variables predictoras

Variable	Descripción
ClaveMunicipio	Clave del municipio
ClaveEstado	Clave del estado
Region	Región
Estado	Estado
Municipio	Municipio
Superficie_km2	Superficie en kilómetros cuadrados
Poblacion	Población
Poblacion_adulta	Población adulta
TipoPoblacion	Tipo de población
TotalSucursales	Total de sucursales de las instituciones financieras
Cajeros	Cajeros
TerminalesPuntodeVenta	Terminales Punto de Venta
TransaccionesEnTPV	Transacciones en TPV (Terminal Punto de Venta)

TransaccionesEnCajerosAutom	Transacciones en Cajeros Automáticos
IngresoRemeasmdd2T18	Ingreso Remesas mdd 4toTrimestre 2017 (dónde mdd es millones de pesos)
IngresoRemeasmdd3T18	Ingreso Remesas mdd 4to Trimestre 2017
SucursalesIntermex	Sucursales de la empresa remesadora
SucursalesMicrofinanciera	Sucursales de la Microfinanciera
GruposGeo2018	Grupos de la Microfinancieras 2017
DepyPagos	Depósitos y Pagos realizados en otros canales
Clientes_jun2018	Clientes a junio 2018
Clientes_sep2018	Clientes a septiembre 2018
Indicador_conjunto	Indicador conjunto

El archivo generado se dividió en 2 archivos: uno para entrenar el modelo y otro para evaluar. Como se mencionó, en el algoritmo Bosque Aleatorio no es necesario realizar esta separación ya que el algoritmo genera internamente los datos de entrenamiento y datos de prueba.

4.1.4 Fase 4: Modelado

- Selección de la técnica de modelado

Para el desarrollo del proyecto y debido a que su objetivo principal era predecir si un municipio era candidato para la localización de sucursales bancarias con base en el criterio de ingreso de remesas, clientes actuales y clientes potenciales de la Microfinanciera, se utilizaron los métodos de aprendizaje supervisado como es el Bosque Aleatorio.

Los métodos fueron entrenados y supervisados para predecir con la mayor exactitud posible, si el municipio es candidato para la localización de sucursales bancarias. El tipo de problema que se maneja es de tipo binario (si o no), ya que solo contempla dos clases o categorías a las cuales puede pertenecer cada registro.

Los métodos que se utilizaron para la construcción del modelo son los siguientes:

- Árboles de decisión;
- Bosques aleatorios;
- Máquina de aumento de gradiente.

- Construcción del modelo

Después de contar con los datos previamente preparados se comienza a generar el modelo. La selección de los mejores parámetros es un proceso iterativo y se basa exclusivamente en los resultados generados.

- Evaluación del modelo

Como se mencionó, el algoritmo Bosque aleatorio es una de las técnicas disponibles más utilizadas actualmente que realiza clasificaciones muy exactas manejando un gran número de campos de entrada. Para el desarrollo de esta tesis cuyo objetivo es la clasificación de los municipios de la República Mexicana como buenos o malos candidatos para la localización de una sucursal bancaria con base en las variables demográficas, geográficas, poblacionales, información de sucursales de la Microfinanciera y bancarias, etc., en este proyecto se realizó el entrenamiento y validación de los modelos con el *software open source R*.

En la sección 3.2.1. Operación del Bosque Aleatorio, se comentó que realiza la partición de datos interna, así el 63.2% de los datos de entrenamiento se usan para generar cada árbol y el complemento para calcular el error OOB. En esta etapa se validará el mejor modelo realizando predicciones con el conjunto de datos de pruebas (datos que se aislaron del conjunto de entrenamiento).

En la sección 3.2.2. Ajuste de Bosque Aleatorio, se sugiere que para encontrar la mejor calidad del número de variables que en este caso son 18 variables totales para cada árbol, el número de variables sugeridas será de 4 variables aleatorias (esto es, $\sqrt{18} \approx 4$) ,o 2 variables aleatorias (la mitad del valor de la raíz cuadrada $\frac{\sqrt{18}}{2} = 2$) , o de 8 variables aleatorias (dos veces el valor de la raíz cuadrada $2 * \sqrt{18}$ =8) . Con respecto al número de árboles utilizados para que correspondan a un clasificador estable con base en la sección 3.2.2., se construyeron diferentes bosques aleatorios con diferente número de árboles los cuales se propusieron de 50, 150, 200, 500 y 1,000 árboles. Lo que da lugar a 15 experimentos ya que son 5 números diferentes de árboles y 3 diferentes números de variables para elegir el modelo que presente el mejor resultado clasificando a los municipios para la localización de sucursales bancarias, las pruebas fueron las siguientes:

En los primeros 5 experimentos se utilizaron 4 variables aleatorias con un número de árboles que utilizó el algoritmo de 50, 150, 200, 500 y 1000 árboles; con estos primeros resultados se seleccionará el modelo que minimice el error. De igual forma, se manejó este procedimiento para los experimentos con 2 variables aleatorias y 8 variables aleatorias.

4.1.5 Fase 5: Evaluación

En esta fase se evalúa el modelo con las matrices de confusión. Si el modelo generado es válido en función de los criterios de éxito establecidos, se procede a la siguiente etapa que es la explotación del modelo.

4.1.6 Fase 6: Explotación

Esta es la última fase de la metodología CRISP-DM y el objetivo de la misma es el de explicar en este caso a los encargados de la toma de decisiones de la Microfinanciera como poner en funcionamiento el proyecto que se ha construido en las fases anteriores, así como exponer los resultados obtenidos de forma que lo puedan entender fácilmente. Otro objetivo de esta fase es el de crear una estrategia para el mantenimiento del proyecto, producir un informe en el que se incluyan posibles mejoras para el futuro y un listado de las dificultades encontradas a la hora de realizarlo.

Dicho de otra manera, después de que el modelo ha sido construido y validado, en esta fase se transforma el conocimiento obtenido en acciones dentro del negocio, es decir, se recomiendan acciones basadas en la observación del modelo y sus resultados y/o aplicando el modelo a diferentes conjuntos de datos o como parte del proceso.

Con respecto a la supervisión y mantenimiento de la implementación del presente proyecto, la minería de datos debería ser realizada en periodos de tres meses (trimestral) ya que esta es la medida de tiempo utilizada en la Comisión Nacional Bancaria y de Valores (CNBV) como fuente de información demográfica, geográfica, poblacional, etc. y del Banco de México (Banxico) donde son consultados los montos de ingresos por remesas por municipio.

El plan de supervisión y mantenimiento se podría establecer los siguientes procesos:

- Extracción y almacenamiento trimestral de los datos guardando la información obtenida una base de datos.
- Los archivos de la explotación de datos, diferentes escenarios y tipos de modelos deberán ser guardados en soporte de la Microfinanciera, almacenándolos por ejemplo en carpetas ordenadas por procesos trimestrales.
- Los resultados obtenidos en cada explotación de datos y modelo deberán ser llevados a formato para su consulta y generar gráficas de distintos tipos para

una mejor visualización e interpretación de los resultados obtenidos en cada periodo. Así cómo podrían ser presentados los resultados en un mapa de la República Mexicana para una mayor visualización.

En este paso también se debe presentar un informe resumiendo los puntos importantes del proyecto y la experiencia adquirida durante su desarrollo. El público al que va dirigido este informe sería el personal de la Microfinanciera de tal manera que se pueda estudiar la situación actual y tomar medidas correctivas para la mejora del servicio de la empresa.

En el siguiente capítulo, se ejecutan los modelos donde se describe los detalles de su ejecución de cada uno y previamente su evaluación que permita establecer el grado de precisión de cada uno de ellos.

Capítulo 5. Desarrollo de los modelos y resultados

En este capítulo se procederá a realizar el análisis descriptivo de los datos, la selección de las variables con base en la importancia de aportación con el indicador. Se describen los parámetros de los modelos, su ejecución, así como la salida y su descripción de cada uno de ellos.

5.1. Análisis descriptivo de los datos

Con base en la metodología CRISP fase 2, se comienza con el análisis de los datos previo al comienzo de la obtención del modelo, realizando gráficas de distribuciones de la variable objetivo que en este caso es el indicador conjunto y las matrices de correlaciones de las variables con respecto a la variable objetivo.

Primeramente, el tamaño de la base de datos es de 2,458 registros. En la figura 5.1 se muestra la distribución de municipios por ingreso de remesas en millones de dólares a junio de 2018.

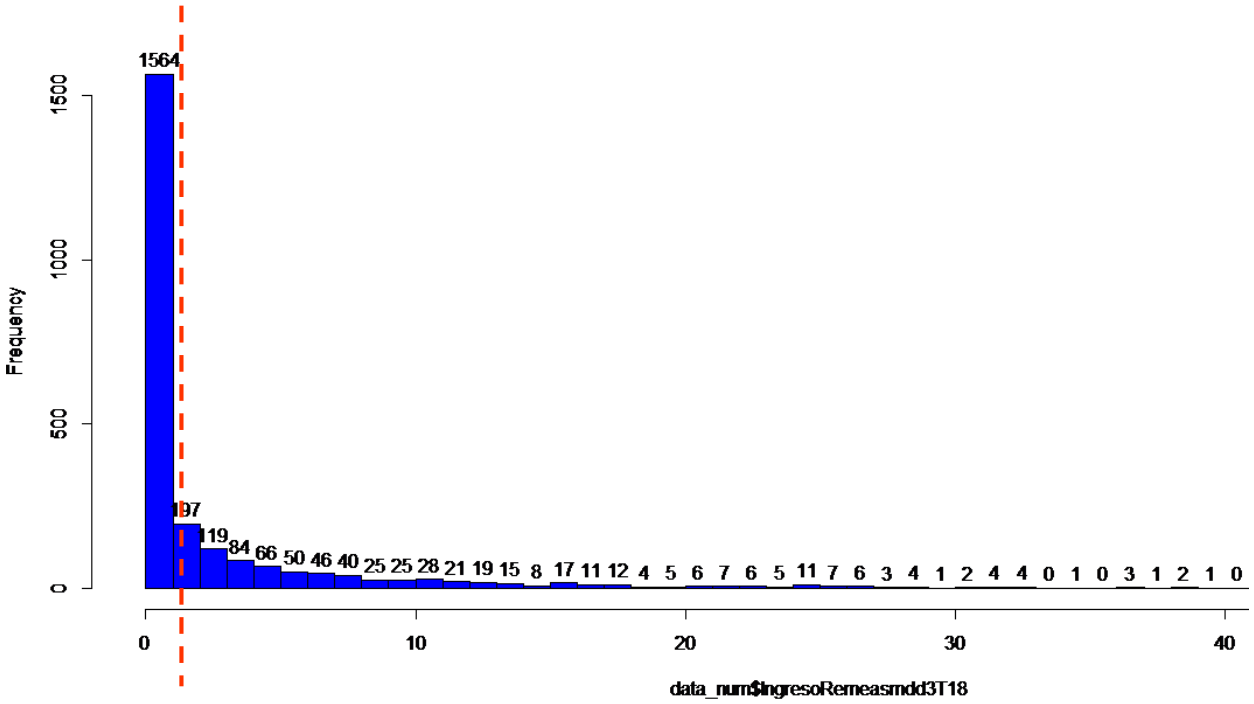

Figura 5.1. Distribución de número de municipios por nivel de ingreso de remesas en millones de dólares.
Fuente: Elaboración propia en R

Como se observa en la figura 5.1 el valor del cuartil 3 o percentil 75 es de 2.873 millones de dólares.

Se realizó la matriz de correlaciones entre las variables dónde entre más cercano a 1 se considera que las variables están correlacionadas, esto es, si tenemos dos variables (A y B) existe correlación entre ellas si al disminuir los valores de A lo hacen también los de B y viceversa (ver Anexo 1 para el código en el *software R*).

En la figura 5.2 se muestran bajas correlaciones positivas de las 18 variables con respecto a la variable indicador conjunto de localización de sucursales, así se observa que no existe una correlación alta entre la variable objetivo y las demás variables, se utilizaron las 18 variables para el modelo. Cabe mencionar que el color del círculo entre más azul fuerte existe una mayor correlación de las variables (se acercan a 1) y entre más claro sea el color azul la correlación es menor entre las variables (se acercan a 0)).

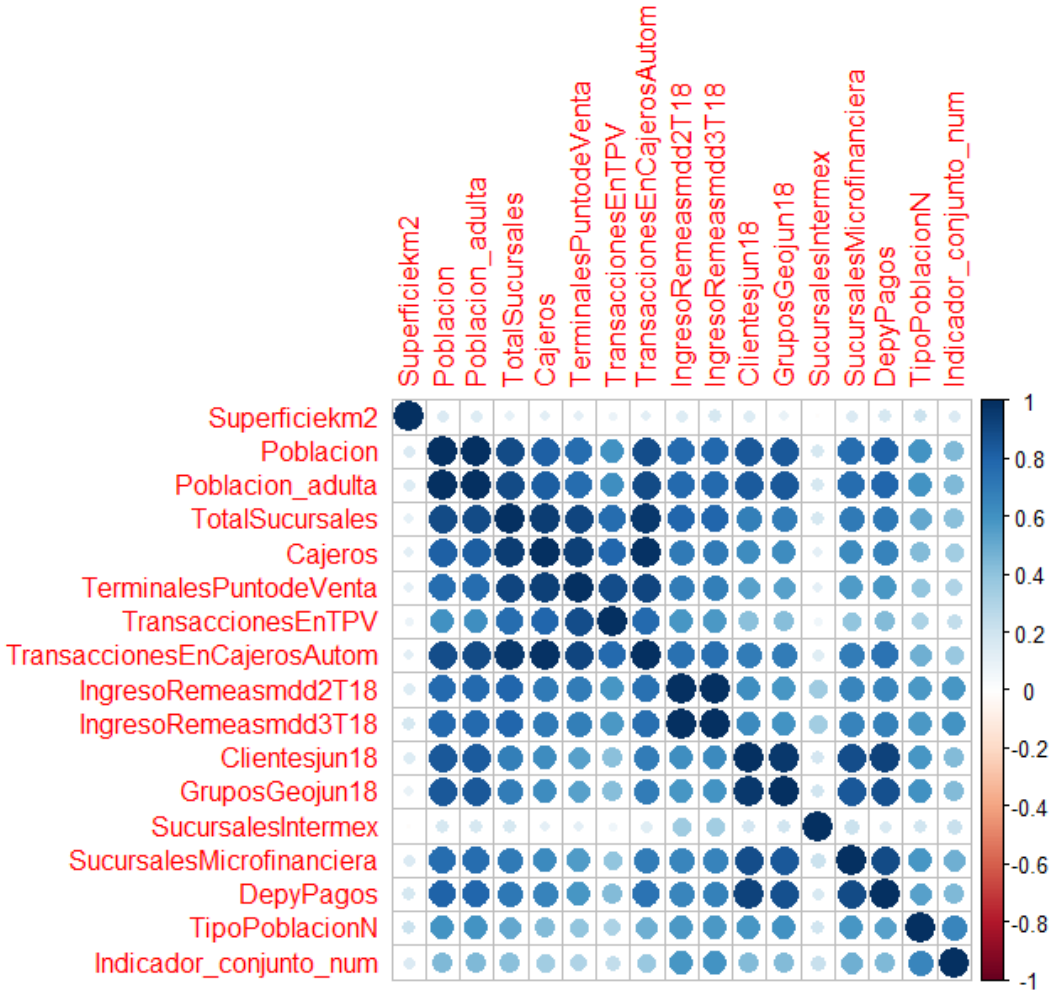

Figura 5.2. Matriz de correlaciones de las variables.
Fuente: Elaboración propia, software R

Como se muestra en la figura 5.2 existe baja correlación de las variables con respecto a la variable indicadora por lo que se utilizaron todas las variables para el modelo sin tener el inconveniente de alguna de ellas tenga dependencia con la variable objetivo. Para mencionar un ejemplo de variables correlacionadas tenemos el ingreso y el gasto familiar que cómo se sabe al aumentar el ingreso y aumenta también los gastos realizados o bien disminuyen juntos.

5.2. Selección de variables importantes

Primeramente se realizará la selección de variables que es un paso importante en un proyecto de modelado predictivo, llamado también selección de características. Es fundamental identificar variables importantes y eliminar variables redundantes antes de construir un modelo predictivo.

La selección de variables es significativa ya que eliminar una variable redundante ayuda a mejorar la precisión. Del mismo modo, la introducción de una variable relevante tiene un efecto positivo en la precisión del modelo.

Demasiadas variables pueden resultar un sobreajuste en el modelo, lo que significa que el modelo no puede generalizar el patrón de los datos y también conducen a un cómputo lento.

Las razones por la cual se eligió el paquete de *software R* llamado Boruta para la selección de variables son:

i. Sigue un método en el cual considera todas las variables que son relevantes para el indicador (*target*).
ii. Funciona bien tanto para la clasificación como para el problema de regresión.
iii. Sigue un método de selección de variables de gran relevancia en el que considera todas las características que son relevantes para la variable indicador. Mientras que, la mayoría de los otros algoritmos de selección de variables siguen un método óptimo mínimo en el que se basan en un pequeño subconjunto de características que produce un error mínimo en un clasificador elegido.

El procedimiento de selección de variables utilizando Boruta se describe en el Anexo 1 sección 2.

A partir del resultado dado por la selección de variables mediante Boruta, la única variable que no es importante es ***Municipio,*** en la figura 5.3 se muestra la impresión que nos arroja Boruta.

```
> print(boruta)
Boruta performed 34 iterations in 42.72542 secs.
 21 attributes confirmed important: Cajeros, ClaveEstado,
ClaveMunicipio, Clientes, DepyPagos and 16 more;
 1 attributes confirmed unimportant: Municipio;
```

Figura 5.3. Resultado de código en la selección de variables mediante Boruta
Fuente: Elaboración propia, software R

La figura 5.4 muestra que las variables utilizadas en el modelo son importantes con excepción de una sola variable que es Municipio, etiquetada como no importante.

La opción plot () muestra el diagrama de caja de todos los atributos de las variables más el puntaje mínimo, promedio y máximo. Las variables que se presentan en el diagrama de caja de color verde en la figura 5.4 muestran que todos los predictores son importantes, caso contario para los diagramas de caja roja, estás variables son rechazadas y el color amarillo de la gráfica de caja indica que son tentativos.

Entonces con base en la figura 5.4 se muestra que todas las variables tienen un diagrama de caja verde por lo que se utilizaran todas para obtener el modelo ya que son importantes para el la variable indicador, con excepción de la variable Municipio que se eliminara de las variables de insumo para el modelo ya que su diagrama es de caja roja lo que significa que es rechazada.

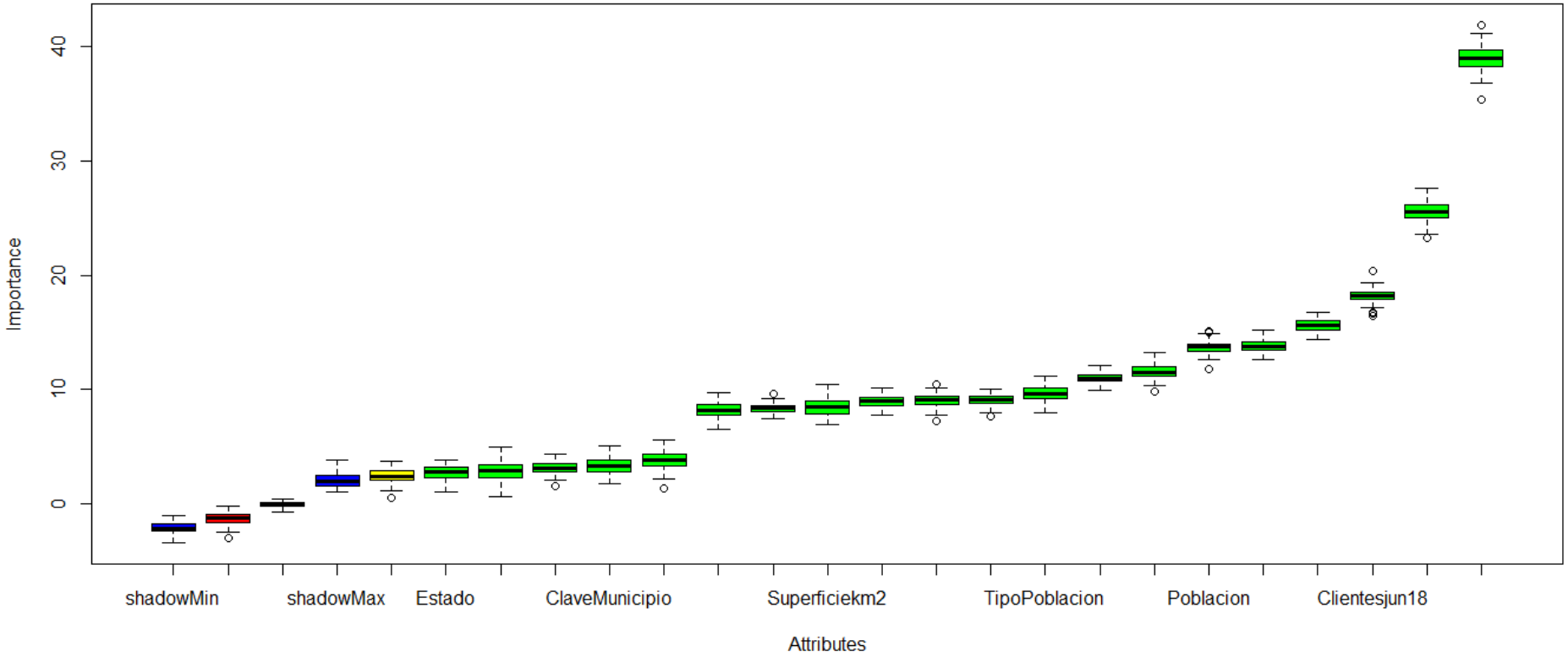

Figura 5.4. Selección de variables o características.
Fuente: Elaboración propia, software R

5.3. Árbol de decisión

El árbol de decisión genera un conjunto de reglas usadas para segmentar la variable predictora. La ventaja es que es un método simple y útil para la interpretación.

Lo que nos muestra el árbol de la figura 5.5 es que cuando tiene un porcentaje alto en color negro en los resultados finales (nodos terminales o cajas rectangulares en la parte inferior del árbol de decisión) nos confirma cuáles son las reglas de las características que tiene que cumplir un municipio candidato para la localización de sucursales bancarias. Se utilizó un árbol de decisión CTree en R, esto es, un árbol de inferencia condicional dónde la partición es recursiva para variables de respuesta.

Se observa en la figura 5.5, que su primer nodo ubicado en la parte superior del árbol de decisión, representa el tipo de población, seguido del total de sucursales por municipio. Los nodos terminales que están representados por rectángulos, entre más sombreados en color negro representa el camino a seguir a la localización de sucursales bancarias. El primer nodo rectangular de lado izquierdo nos dice que hay municipios clasificados cómo buenos candidatos para la localización de sucursales bancarias los cuales cumplen que el tipo de población sean municipios rurales, en transición o semi-urbano (respecto al número es menor o igual a 3), que no existen sucursales bancarias. Otro sitio factible para la localización de sucursales se da cuando el tipo de población de los municipios sea semi-metropoli, metrópoli o urbano, el total de sucursales bancarias existentes en estos municipios sea menor o igual a 6.

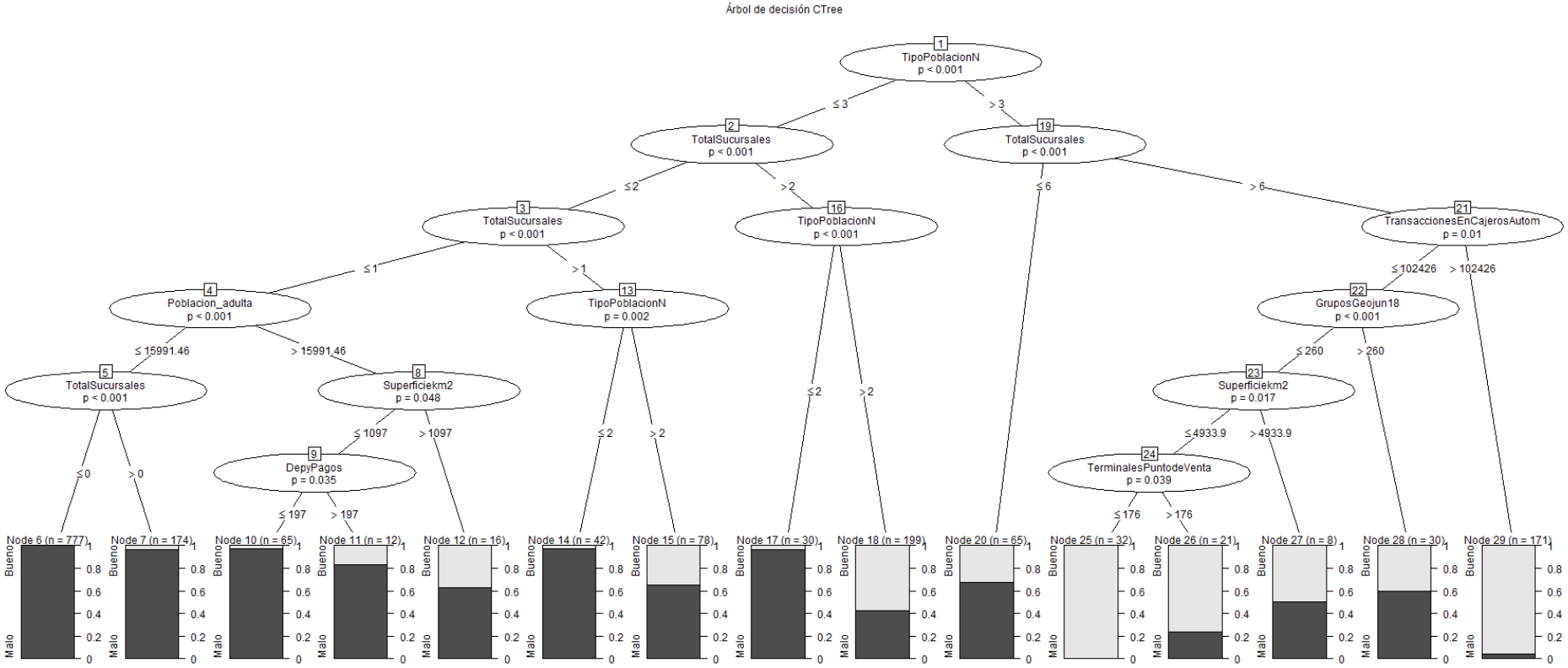

Figura 5.5. Árbol de decisión
Fuente: Elaboración propia, software R

La curva ROC se muestra en la figura 5.6, donde el mejor método posible de predicción se situaría en un punto de la esquina superior izquierda, o coordenada (0,1) del espacio ROC, representando un 100% de sensibilidad (ningún falso negativo) y un 100% también de especificidad (ningún falso positivo). Así tenemos que para este caso una alta predicción de 96%, para el árbol de decisión al predecir la localización de sucursales bancarias por municipio.

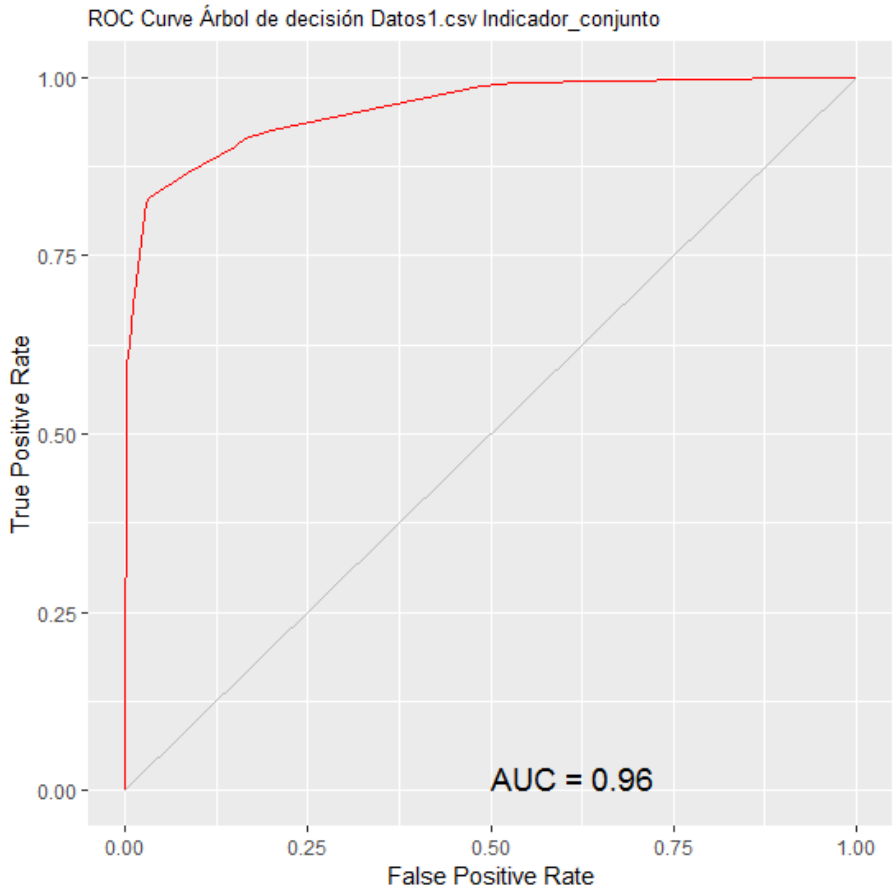

Figura 5.6. Curva ROC (Característica Operativa del Receptor)
Fuente: Elaboración propia, software R

5.4. Evaluación del modelo *Random Forest*

En esta etapa tenemos que la variable *target* llamada “Indicador conjunto” es “bueno” cuando el municipio tiene un monto de ingresos de remesas por arriba del tercer cuartil, incluyendo el número de habitantes mayores de edad

económicamente activos sea mayor a 10,000 y el número de clientes atendidos de la Microfinanciera sea mayor al primer cuartil y es “malo” cuando ocurre lo contario.

Como se mencionó, el algoritmo Bosque Aleatorio es una de las técnicas más utilizadas actualmente, la cual realiza clasificaciones muy exactas manejando un gran número de variables de entrada. Para el desarrollo de este proyecto y debido a que su objetivo principal es una tarea de clasificación, esto es, indicar si un municipio de la República Mexicana es candidato para la localización de una sucursal bancaria, el entrenamiento y validación del algoritmo se realizó con el *software open source* R.

En la sección 3.2.1, se comentó que el Bosque Aleatorio realiza la partición de datos interna, así el 63.2% de los datos de entrenamiento se usan para generar cada árbol y el complemento para calcular el error fuera de la bolsa u *Out Of Bag* (OOB). En esta etapa se validará el mejor modelo realizando predicciones con el conjunto de datos de pruebas (datos que se aislaron del conjunto de entrenamiento).

En esta fase se establecieron 15 experimentos para elegir el modelo que presentara el mejor resultado clasificando los municipios como candidatos para la localización de sucursales bancarias, las pruebas fueron las siguientes:

Se establece el parámetro del número de árboles que utilizará el algoritmo (50, 150, 200, 500, 1,000 y 2,000 árboles) con base en Bhalla (2014), se seleccionará el modelo que minimice el error. Para estas pruebas se utilizó como número de variables aleatorias 2, 4 y 8.

En los primeros 6 experimentos para elegir el modelo que presenté el mejor resultado clasificando los sitios para la localización de sucursales bancarias se utilizaron 4 variables aleatorias, este valor fue obtenido de calcular la raíz cuadrada del número de variables predictoras que son 17 variables:

Experimento 1: Entrenamiento con 50 árboles

	Predicho	
Real	Bueno	Malo
Bueno	532	28
Malo	34	1864
Sensibilidad	95.0%	
Especificidad	98.2%	
Exactitud	97.5%	

Tabla 5.1. Matriz de confusión
Fuente: Elaborado con datos de respuesta software R

Experimento 2: Entrenamiento con 150 árboles

	Predicho	
Real	Bueno	Malo
Bueno	532	28
Malo	33	1865
Sensibilidad	95.0%	
Especificidad	98.3%	
Exactitud	97.5%	

Tabla 5.2. Matriz de confusión
Fuente: Elaborado con datos de respuesta software R

Experimento 3: Entrenamiento con 200 árboles

	Predicho	
Real	Bueno	Malo
Bueno	533	27
Malo	32	1866
Sensibilidad	95.2%	
Especificidad	98.3%	
Exactitud	97.6%	

Tabla 5.3. Matriz de confusión
Fuente: Elaborado con datos de respuesta software R

Experimento 4: Entrenamiento con 500 árboles

	Predicho	
Real	Bueno	Malo
Bueno	532	28
Malo	32	1866
Sensibilidad	95.0%	
Especificidad	98.3%	
Exactitud	97.6%	

Tabla 5.4. Matriz de confusión
Fuente: Elaborado con datos de respuesta software R

Experimento 5: Entrenamiento con 1000 árboles

	Predicho	
Real	Bueno	Malo
Bueno	532	28
Malo	35	1863
Sensibilidad	95.0%	
Especificidad	98.2%	
Exactitud	97.4%	

Tabla 5.5. Matriz de confusión
Fuente: Elaborado con datos de respuesta software R

Experimento 6: Entrenamiento con 2000 árboles

	Predicho	
Real	Bueno	Malo
Bueno	532	28
Malo	35	1863
Sensibilidad	95.0%	
Especificidad	98.2%	
Exactitud	97.4%	

Tabla 5.6. Matriz de confusión
Fuente: Elaborado con datos de respuesta software R

Los resultados de las pruebas de entrenamiento con los diferentes números de árboles y número de variables se encuentran en el anexo 3.

Después de realizar los 18 experimentos, se eligió el experimento número 3 este algoritmo dio un mejor resultado para este objetivo, con 200 árboles y 4 variables, ya que presentó una sensibilidad mayor de 95.2% es decir, es el número de municipios que el modelo predice como buenos candidatos para la localización de sucursales bancarias y realmente lo son, entre todos los municipios que realmente son buenos; especificidad de 98.3% (número de municipios que el modelo predice como malos candidatos para la localización de sucursales bancarias y realmente lo son, entre todos los municipios que realmente son malos) la cuál fue la máxima alcanzada para estos 18 experimentos y una exactitud del modelo del 97.6% está precisión también fue el valor más alto alcanzado en estos experimentos, ver figura 5.7.

Número árboles	50	150	200	500	1,000	2,000
Sensibilidad	95.0%	95.0%	95.2%	95.0%	95.0%	95.0%
Especificidad	98.2%	98.3%	98.3%	98.3%	98.2%	98.2%
Exactitud	97.5%	97.5%	97.6%	97.6%	97.4%	97.4%

Tabla 5.7. Métricas de matriz de confusión utilizando 4 variables
Fuente: Elaborado con datos de respuesta software R

Ver el código *R* del modelo en Anexo 2.

Las principales variables con mayor importancia del modelo son: total de sucursales bancarias totales existentes en el municipio, número de clientes de la Microfinanciera, población adulta es decir mayor a 18 años, grupos actuales de la empresa, transacciones en cajeros automáticos por municipio, depósitos y pagos realizados en la competencia, número de cajeros, superficie del municipio en kilómetros cuadrados, terminales de punto y venta y en la figura 5.7 se muestran las 14 variables más importantes del modelo.

	Bueno	Malo	MeanDecreaseAccuracy	MeanDecreaseGini
TotalSucursales	21.82	5.50	23.90	90.82
Clientesjun18	10.28	5.19	13.90	19.96
Poblacion_adulta	11.99	6.85	13.43	33.61
GruposGeojun18	7.31	7.90	13.27	21.35
TransaccionesEnCajerosAutom	12.83	1.65	12.82	50.79
Poblacion	14.25	6.03	11.55	30.01
DepyPagos	10.85	3.44	11.14	28.37
Cajeros	11.32	4.09	10.78	50.17
Superficiekm2	13.03	-5.28	7.90	17.74
TransaccionesEnTPV	2.36	9.08	7.09	18.65
TerminalesPuntodeVenta	4.67	4.18	6.88	22.09
TipoPoblacionN	5.30	2.90	5.20	6.07
SucursalesMicrofinanciera	3.11	0.53	2.83	2.07
SucursalesIntermex	1.57	-0.89	0.40	0.77

Figura 5.7 Importancia de variables de Bosque Aleatorio
Fuente: Elaborado con datos de respuesta software R

Con el modelo de 200 árboles y 4 variables que dio mejor resultado para clasificar los sitios para la localización de sucursales bancarias, se utilizó para predecir la clasificación de buenos o malos municipios para la localización de sucursales bancarias con información al tercer trimestre del año 2019, como resultado el modelo arrojó una buena predicción de 95.7%, y una alta sensibilidad del 89.5% esto es la razón del número de municipios que el modelo predice como buenos candidatos para la localización y realmente lo son, entre todos los municipios que realmente son buenos:

	Predicción	
Real	Bueno	Malo
Bueno	502	59
Malo	46	1851
Sensibilidad	89.5%	
Especificidad	97.6%	
Exactitud	95.7%	

Tabla 5.8. Matriz de confusión
Fuente: Elaborado con datos de respuesta software R

5.5. Evaluación del modelo *Gradient Boosting Machine*

Se realizó un modelo de H2O Máquina de Aumento de Gradiente o *Gradient Boosting Machine* (GBM) mencionado en la sección 3.3.1 para generar un modelo que nos permita comparar los resultados que se obtuvieron con el modelo de Bosque Aleatorio. Cabe mencionar que los algoritmos Máquina de Aumento de Gradiente son igualmente útiles cuando se trabaja con un volumen de datos reducido o con un volumen de datos grande.

Se realizó la partición de datos interna, así el 70% de los datos de entrenamiento se usan para generar el modelo y el complemento (30% de los datos) para calcular el error OOB. Finalmente se evalúa el modelo con toda la base de datos.

Se utilizaron 1000 árboles ya que estudios previos como lo menciona Singh (2018) muestran que este parámetro da mejores resultados.

El modelo da los siguientes resultados:

	Predicción	
Real	Bueno	Malo
Bueno	532	44
Malo	28	1,854
Sensibilidad	92.36%	
Especificidad	98.51%	
Exactitud	97.07%	

Tabla 5.9. Matriz de confusión
Fuente: Elaborado con datos de respuesta software R

En la figura 5.8 se presentan en forma descendente las variables que son más importantes para el modelo GBM, donde la variable total de sucursales por

municipio tiene la mayor relevancia en este modelo seguido de la variable grupos de la Microfinanciera, depósitos y pagos que realizan los clientes en sucursales de la competencia y población por municipio en cuarto lugar.

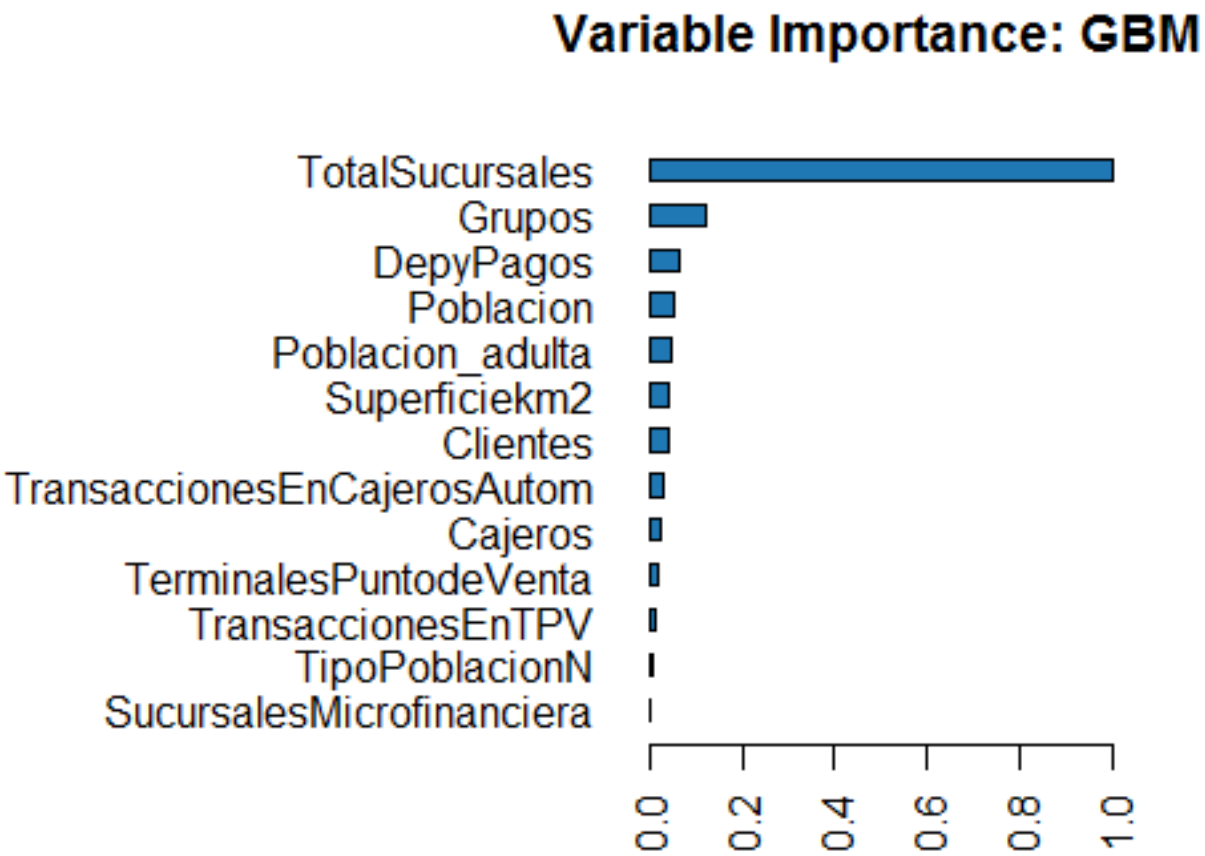

Figura 5.8 Importancia de variables de GBM
Fuente: Elaboración propia, software R

Con el modelo de 1000 árboles, se utilizó para clasificar los sitios para la localización de sucursales bancarias con información al tercer trimestre de 2019, es decir con datos actualizados 3 meses después de realizar el entrenamiento del modelo, el cual tuvo como resultado una predicción de 95.08%, y una sensibilidad de 88.87% que es la proporción del número de municipios buenos predichos y realmente los son entre el total de casos buenos reales:

En general la tabla 5.10 muestra una exactitud alta de 95.08% con porcentaje de sensibilidad también bueno para la localización de sucursales bancarias por municipio. Así se puede utilizar el modelo para la predicción de la clasificación de municipios agregando información reciente y nos dará buenos resultados.

	Predicción	
Real	Bueno	Malo
Bueno	503	63
Malo	58	1,834
Sensibilidad	88.87%	
Especificidad	96.93%	
Exactitud	95.08%	

Tabla 5.10. Matriz de confusión de modelo seleccionado GBM
Fuente: Elaborado con datos de respuesta software R

5.6. Comparación de resultados Bosque Aleatorio y Máquina de Aumento de Gradiente

En la sección 5.4 y 5.5 se entrenaron los modelos Bosque Aleatorio y Máquina de Aumento de Gradiente, los cuales se evaluaron al segundo trimestre del año 2018 y en el tercer trimestre de 2018. En la tabla 5.11 se muestran las métricas utilizadas para evaluar el desempeño de los modelos, donde se observa que el Bosque Aleatorio tiene un mejor desempeño en ambos periodos del año 2018, podemos ver que el modelo de Bosque aleatorio también muestra estabilidad ya que con respecto a la exactitud de la localización de sucursales bancarias disminuye de 97.6% a 95.7%, así decrece solo 2 puntos porcentuales, para la métrica de especificidad o razón de verdaderos negativos decrece 0.7%.

	Sensibilidad		Especificidad		Exactitud	
Modelo	**2T 2018**	**3T 2018**	**2T 2018**	**3T 2018**	**2T 2018**	**3T 2018**
Bosque Aleatorio	**95.2%**	**89.5%**	**98.3%**	**97.6%**	**97.6%**	**95.7%**
Máquina de Aumento de Gradiente	**92.4%**	**88.9%**	**98.5%**	**96.9%**	**97.1%**	**95.1%**
Diferencia	**2.8%**	**0.6%**	**-0.2%**	**0.6%**	**0.5%**	**0.7%**

Tabla 5.11. Resumen comparativo de resultados modelos,
Fuente: Elaborado con datos de respuesta software R

En resumen, se eligió el bosque aleatorio ya que en los dos periodos del año 2018 se obtienen mejores resultados en las métricas utilizados: exactitud, sensibilidad y especificidad, comparando con los datos de respuesta de la Máquina de Aumento de Gradiente.

Se compararon en esta tesis dos modelos para tener punto de comparación, dónde se encontraron hallazgos inesperados que el Bosque aleatorio predijo mejor para el objetivo de esta tesis y los datos utilizados. Se esperaba que la Máquina de Aumento de Gradiente tuviera un mejor desempeño ya que maneja una gran cantidad de datos en menor tiempo.

Conclusiones

Un modelo de clasificación como es el Bosque Aleatorio (*Random Forest*) facilita tener un algoritmo predictivo para la localización de sucursales en el sector bancario, el cual ayudará a la Microfinanciera en la toma de decisiones para localizar sus sucursales con base en la información pública (como la que se puede obtener en las bases de dato públicas del Banco de México, la Comisión Nacional Bancaria y de Valores) y datos de la Microfinanciera en este estudio.

Para esta tesis se investigó el funcionamiento del algoritmo Bosque Aleatorio y se consideraron características importantes del mismo, entre ellas las siguientes: este algoritmo no es afectado por valores *outlier* o atípicos, los cuales son observaciones de los datos que son numéricamente distantes del resto; recomendable contar con una técnica para la ocurrecia de *missing values* o datos perdidos, que son aquellos valores faltantes debido a que no se almacena ningún valor de datos en la observación antes de ser procesados.

Se entrenó el modelo de aprendizaje supervisado Bosque Aleatorio, y el proyecto se enfocó en reducir los falsos negativos (municipios buenos que el modelo predice como malos) más que en reducir los falsos positivos (municipios malos que el modelo predice como buenos). El modelo presentó buenos resultados en su matriz de confusión (sensibilidad de 95.2%, exactitud de 97.6%) para identificar los sitios para la localización de sucursales bancarias, esto quiere decir que con el modelo Bosque Aleatorio se obtuvo una buena exactitud como respuesta del modelo, así como una alta sensibilidad, esto es, clasifica con alta probabilidad asertiva los municipios que el modelo predice como buenos candidatos para la localización de sucursales bancarias y realmente lo son, entre todos los municipios que realmente son buenos para localizar sucursales bancarias.

Utilizando H2O y *Gradient Boosting Machine* el modelo dio una sensibilidad de 92.4% y una exactitud de 97.1%, lo que implica que comparándolo con los resultados del modelo Bosque Aleatorio, queda 2.8 puntos porcentuales y 0.5 puntos porcentuales por debajo del rendimiento, respectivamente. Lo que significa que el modelo GBM no obtuvo los mejores resultados, en cambio con el modelo de Bosque Aleatorio se alcanzaron una mayor precisión y sensibilidad.

Comparando los modelos *Random Forest* y *Gradient Boosting Machine* se obtuvieron mejores resultados de exactitud, sensibilidad del modelo *con el Random Forest* o Bosque Aleatorio, por lo que se eligió como modelo final para la localización de sucursales en esta tesis.

Para mejorar los resultados del modelo podemos concentrarnos en el conjunto de datos de entrenamiento, como se mencionó en la sección 3.2.1. Si los conjuntos de datos se encuentran desbalanceados se pueden presentar resultados inesperados en cuanto a la precisión o exactitud. En este caso, los datos de entrenamiento

contienen más registros de municipios con el indicador “malo”, esto es que no son candidatos para localizar sucursales bancarias. En este caso se utilizó un balance de 70%y 30% con dos diferentes modelos. Es posible probar regulando el tamaño de las clases 50% y 50% utilizar la misma cantidad de municipios malos y buenos seleccionados de forma aleatoria, para ver si disminuye o incrementa la precisión de modelo.

Se recopiló información de los 2,458 municipios del país, entre ellos datos de sucursales bancarias reportadas en Banco de México, así como indicadores geográficos, sociodemográficos y estatus de la Microfinanciera en cada uno de los municipios. Con todos estos datos se crearon dos modelos (Bosque Aleatorio y Máquina de Aumento de Gradiente) haciendo una clasificación de 18 variables.

Mediante una validación interna de los modelos, se midió el área bajo la curva ROC (*Receiver Operating Characteristic*) teniendo en cuenta que un buen indicador de ROC es igual o superior a 0.7, en el caso del modelo elegido fue de 0.9 lo que significa que tiene un buen resultado del área bajo la curva. Asimismo, se construyeron las matrices de confusión que permitieron calcular la sensibilidad que indica la proporción de municipios como buenos candidatos para la localización de sucursales bancarias y que son correctamente identificados, la especificidad que mide la proporción de municipios como malos candidatos para la localización y que realmente lo sean. Con base en estas métricas se eligió el modelo que presentó mayor precisión, sensibilidad y especificidad en los dos periodos evaluados basándonos en el marco teórico descrito anteriormente.

El modelo identificó como primer lugar municipios con probabilidades mayores al 95% para localizar sucursales bancarias. Entre ellos: Toluca, Naucalpan, Nezahualcóyotl en el estado de México, Iztapalapa y Tlalpan en la Cuidad de México, Guadalajara en el estado Jalisco y Morelia en Michoacán. Por otro lado, los municipios con un nivel bajo de probabilidad de ubicar sucursales bancarias pertenecen a los estados de Oaxaca (municipios como: Santo Domingo Tlatayápam, San Miguel del Río) y Sonora (Arivechi y Huépac).

Se muestra que predominan las capitales de los estados como sitios elegidos donde se puede tomar la decisión de la localización de una sucursal bancaria para la Microfinanciera.

A pesar de que en los últimos años ha surgido la era de la digitalización dentro de la banca, con base en estudios e investigaciones realizadas hacia los clientes de los bancos, Ruiz (2019) y Iprofesional (2018) nos muestran que aún existe la tendencia a utilizar las sucursales bancarias buscando el trato uno a uno (banco – cliente) ya que para los usuarios resulta más confiable y seguro tener un contacto físico con el banco, sobre todo para los clientes de edad de 35 a 59 años, lo que conlleva a que los bancos del país tengan y busquen sitios viables para la localización de sucursales bancarias.

La construcción de un modelo de Bosque Aleatorio y la utilización de un algoritmo de aprendizaje supervisado (de los más certeros que hay disponibles), le aporta a la teoría de la localización un modelo que predice en qué municipios de México existe mayor ventaja al localizar sucursales bancarias que no habían sido identificados antes, basándose principalmente en similitudes entre los municipios con mayor penetración de sucursales bancarias, población por municipio, clientes de la Microfinanciera, etc.

Asimismo, gracias a esta clasificación y al ser analizados los datos de forma trimestral, el modelo presenta información relevante que indica dónde debería comenzar a explorar sitios posibles para la localización de sucursales bancarias basándose de datos históricos. Además, las técnicas utilizadas cuentan con la flexibilidad de ir agregando nuevos criterios (nuevas variables) al modelo que pueden ser influyentes para la localización y poder incrementar el nivel de precisión.

Como hallazgo inesperado se encontró que al realizar los dos modelos Bosque Aleatorio y Máquina de Aumento de Gradiente, se mostró que a pesar de que el segundo modelo cuenta con la capacidad de combinar algoritmos de Aprendizaje de máquina con *Big Data*[21] , se obtuvo menor precisión en la identificación de municipios para la localización de sucursales bancarias en los dos trimestres evaluados, al compararlo con el primer modelo es decir, el Bosque aleatorio. Se esperaba que los resultados del modelo GBM fueran mejores comparado con el Bosque Aleatorio ya que es un modelo que maneja gran cantidad de datos con buena precisión según reporta la literatura ver Ravanshad (2018).

Asimismo, el modelo elegido de Bosque Aleatorio presentó una estabilidad en los 2 trimestres evaluados, su precisión no varío en gran proporción en el transcurso del tiempo, lo que muestra que el modelo se puede utilizar en un periodo de 6 meses, sin embargo, los modelos se tienen que estar validando si se utilizarán para periodos más prolongados.

Una implicación teórica de trabajo realizado es que se aportan nuevos criterios para localización de sucursales bancarias: el monto de ingresos de remesas como variable para localizar sucursales bancarias de la empresa.

Random Forest es uno de los algoritmos de clasificación más usados y una de sus ventajas es que aporta una estimación interna de exactitud mediante una forma de validación cruzada de los datos por lo que los resultados encontrados para la localización de sucursales bancarias ayudarán a tomar decisiones a la Microfinanciera de manera más precisa y con mayor rapidez.

Si la Microfinanciera toma la decisión de agregar o cambiar criterios para la localización de sus sucursales bancarias se puede agregar variables de entrada al

[21] Se refiere a los conjuntos de datos o combinaciones de conjuntos de datos cuyo tamaño (volumen), complejidad (variabilidad) y velocidad de crecimiento (velocidad) dificultan su captura, gestión, procesamiento o análisis mediante tecnologías y herramientas convencionales.

modelo, además estos nuevos escenarios pueden ejecutarse rápidamente para obtener una respuesta eficaz y con precisión alta. Este sentido, otra posibilidad ventajosa es poder realizando combinaciones de criterios para ver cuáles son los más importantes para este tipo de problema de localización.

Con respecto a la validación externa de los resultados de los municipios candidatos para la localización de sucursales bancarias, se confirma que mediante los modelos de aprendizaje de máquina como Bosque Aleatorio se puede predecir en qué municipios de México pueden localizarse sucursales bancarias ya que se compararon los resultados con el área encargada de este proyecto cuyos resultados son parte de los estudios de la empresa. Se observó que hubo coincidencias en sitios, cómo el municipio de Morelia, como candidatos para la localización de sucursales de la empresa.

Asimismo, se confirma que el modelo Bosque Aleatorio aprendió el comportamiento de las características de los municipios considerando el ingreso de remesas, dicho modelo se puede utilizar ingresándole nuevos datos y así poder clasificar a los municipios si son candidatos para abrir sucursales.

Recomendaciones

A partir del modelo *Random Forest* entrenado con 4 variables y 200 árboles, se obtuvo un buen clasificador para identificar en qué municipios se podrían localizar sucursales bancarias con base en el criterio de remesas, clientes de la Microfinanciera y población. Sin embargo, se tiene que dar mantenimiento al modelo mínimo cada seis meses, ya que con base en la información de los datos consultados de la Comisión Nacional Bancaria y de Valores, éstos están en constante cambio, lo que hace que el modelo también vaya cambiando y por lo tanto es necesario volver entrenar el modelo.

Otra recomendación para mejorar los resultados del modelo es incluir variables que se encuentren en otras fuentes de datos y que aporten al modelo para mejorar la precisión, por ejemplo, una de las variables sería incluir tasas de delincuencia por tipo de delito por municipio ya que se ve afectado tanto al cliente que visitará la sucursal bancaria, cómo la estabilidad de la empresa en estas zonas.

Finalmente, en lugar de utilizar los cuartiles con los que se realizó esta tesis para la clasificación de municipios en función de su ingreso de remesas, clientes y población adulta, se pueden realizar clústers como etapa para clasificar a los municipios.

Trabajos futuros

La primera línea de continuación de este trabajo de investigación es utilizar datos a mayor detalle, esto es, a nivel código postal, manzana geográfica, etc. El motivo por el cual no se realizó así, fue porque la información pública consultada no se encuentra a este detalle.

Para presentar los resultados se propone utilizar un sistema de visualización donde se utilicen mapas mostrando la República Mexicana como apoyo del cliente final, en este caso los tomadores de decisiones de la Microfinanciera. Esto facilitaría la comprensión y el análisis del modelo por municipio. Una de las herramientas que podría utilizarse es *Qlik Sense*[22] que permite visualizar los sitios en formato de *dashboard*[23] , también es posible usar mapas para gestionar los resultados del modelo de sucursales bancarias mediante visualizaciones con *Power Map*[24] .

Figura 5.9 Mapa a nivel estado
Fuente: Con base en la herramienta Qlik Sense

[22] Una aplicación revolucionaria de descubrimiento y visualización de datos de autoservicio, ayuda a la exploración de datos para ver qué está ocurriendo y conocer el por qué.
[23] Es una representación gráfica de las principales métricas o indicadores que intervienen en la consecución de los objetivos de una estrategia.
[24] Herramienta de visualización de Excel, de datos (en 3D) tridimensionales que le permite ver información de nuevas maneras.

Anexo 1

1. Código para generar la matriz de correlación entre las variables

```
## Carga de base de datos
setwd("C:/Users/Microsoft/Documents/Karina/2019/Tesis/Datos")
data=read.csv("Datos.csv",header=T,na.strings="?")
## Vista de la base de datos
View(data)
ls(data)
## Selección de variables para la matriz de correlaciones
data_num <- data[,c(-1,-2,-3,-4,-5,-9)]
View(data_num)
##Matriz de correlación
library(corrplot)
M <- cor(data_num)
##Gráfica de correlaciones entre las variables
corrplot(M)
corrplot(M, method="number")
```

2. Selección de variables relevantes utilizando Boruta de *R*

```
## Carga de base de datos
setwd("C:/Users/Microsoft/Documents/Karina/2019/Tesis/Datos")
data1=read.csv("Datos1.csv",header=T,na.strings="?")
## Para visualizar Base de datos en R
View(data1)
## Convertir la variable target llamada “indicador conjunto” a tipo categ
órica
data1$Indicador_conjunto= as.factor(data1$Indicador_conjunto)
```

```
## Estadísticos de cada variable de la base de datos cómo: mínimo, máximo
, mediana, promedio, etc.

summary(data1)


## Se instala y carga el paquete Boruta

install.packages ("Boruta")

library(Boruta)

## Se establece una semilla para que el resultado al ejecutar se mantenga
constante

set.seed (456)

## Ejecución de algoritmo

boruta <- Boruta(Indicador_conjunto~., data = data1, doTrace = 2)

print(boruta)

plot(boruta)

summary(data1)
```

```
> summary(data1)
 ClaveMunicipio   ClaveEstado                    Region        Estado              Municipio    Superficiekm2
 Min.   : 1001   Min.   : 1.00   Centro Sur y Oriente:731   Oaxaca  : 570   Benito Juárez  :   7   Min.   :    3.59
 1st Qu.:14084   1st Qu.:14.00   Ciudad de México    : 16   Puebla  : 217   Ocampo         :   6   1st Qu.:   87.24
 Median :20232   Median :20.00   Noreste             :190   Veracruz: 212   Emiliano Zapata:   5   Median :  233.75
 Mean   :19370   Mean   :19.26   Noroeste            :206   Jalisco : 125   Hidalgo        :   5   Mean   :  813.50
 3rd Qu.:24030   3rd Qu.:24.00   Occidente y Bajío   :401   México  : 125   Juárez         :   5   3rd Qu.:  680.01
 Max.   :32058   Max.   :32.00   Sur                 :914   Chiapas : 118   Morelos        :   5   Max.   :51952.30
                                                            (Other) :1091   (Other)        :2425
   Poblacion       Poblacion_adulta          TipoPoblacion TotalSucursales      Cajeros        TerminalesPuntodeVenta
 Min.   :     88   Min.   :     66.9   En Transición :619   Min.   :  0.000   Min.   :   0.00   Min.   :    0.00
 1st Qu.:   4576   1st Qu.:   3311.1   Metrópoli     : 12   1st Qu.:  0.000   1st Qu.:   0.00   1st Qu.:    0.00
 Median :  13739   Median :   9872.4   Rural         :659   Median :  1.000   Median :   1.00   Median :    6.00
 Mean   :  50722   Mean   :  37209.0   Semi-metrópoli: 71   Mean   :  6.987   Mean   :  21.14   Mean   :  381.06
 3rd Qu.:  35695   3rd Qu.:  25552.7   Semi-urbano   :736   3rd Qu.:  4.000   3rd Qu.:   5.00   3rd Qu.:   47.75
 Max.   :1818939   Max.   :1388352.9   Urbano        :361   Max.   :423.000   Max.   :2118.00   Max.   :37466.00
```

```
 TransaccionesEnTPV TransaccionesEnCajerosAutom IngresoRemeasmdd2T18 IngresoRemeasmdd3T18 Clientesjun18     GruposGeojun18
 Min.   :       0   Min.   :      0             Min.   :  0.0000     Min.   :  0.00000    Min.   :    0.0   Min.   :   0.0
 1st Qu.:       0   1st Qu.:      0             1st Qu.:  0.0003     1st Qu.:  0.00064    1st Qu.:   30.0   1st Qu.:   4.0
 Median :     546   Median :   1911             Median :  0.2165     Median :  0.20528    Median :  183.5   Median :  22.0
 Mean   :   74575   Mean   :  75762             Mean   :  3.6780     Mean   :  3.43581    Mean   :  940.5   Mean   : 104.3
 3rd Qu.:    5824   3rd Qu.:  17124             3rd Qu.:  2.8735     3rd Qu.:  2.66430    3rd Qu.:  716.0   3rd Qu.:  81.0
 Max.   :16122547   Max.   :5218755             Max.   :126.6578     Max.   :112.37626    Max.   :37480.0   Max.   :4517.0

 SucursalesIntermex SucursalesMicrofinanciera   DepyPagos       Indicador_conjunto TipoPoblacionN
 Min.   :0.000      Min.   : 0.0000           Min.   :    0.0   Bueno: 560         Min.   :1.000
 1st Qu.:0.000      1st Qu.: 0.0000           1st Qu.:    0.0   Malo :1898         1st Qu.:1.000
 Median :0.000      Median : 0.0000           Median :    0.0                      Median :2.000
 Mean   :0.024      Mean   : 0.2425           Mean   :  646.5                      Mean   :2.612
 3rd Qu.:0.000      3rd Qu.: 0.0000           3rd Qu.:  268.8                      3rd Qu.:3.000
 Max.   :2.000      Max.   :10.0000           Max.   :31880.0                      Max.   :7.000
```

Figura A0.2 Resumen de la base de datos.
Fuente: Elaboración propia, software R

3. Selección de variables relevantes utilizando Boruta de *R*

Importancia de las variables

La figura A1.3 muestra que las variables utilizadas en el modelo con tres asteriscos son de gran importancia, es decir entre más cercana a cero sea la probabilidad de la distribución Chi cuadrada, más importante es la variable en relación con la variable target es decir indicador conjunto.

```
                          Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                                      1719    1884.58
Superficiekm2              1    52.48      1718    1832.10 4.350e-13 ***
Poblacion                  1   719.46      1717    1112.64 < 2.2e-16 ***
Poblacion_adulta           1     0.00      1716    1112.64 0.9870406
TotalSucursales            1   164.96      1715     947.68 < 2.2e-16 ***
Cajeros                    1     2.91      1714     944.77 0.0881716 .
TerminalesPuntodeVenta     1    16.12      1713     928.65 5.935e-05 ***
TransaccionesEnTPV         1     0.46      1712     928.19 0.4966862
TransaccionesEnCajerosAutom 1    1.39      1711     926.79 0.2377598
Clientesjun18              1    33.47      1710     893.32 7.231e-09 ***
GruposGeojun18             1    34.75      1709     858.57 3.749e-09 ***
SucursalesIntermex         1     6.36      1708     852.21 0.0116514 *
SucursalesMicrofinanciera  1     4.20      1707     848.00 0.0403261 *
DepyPagos                  1    13.59      1706     834.41 0.0002275 ***
TipoPoblacionN             1    20.68      1705     813.73 5.432e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Tiempo transcurrido: 1.40 segs

Rattle marca de tiempo: 2019-04-09 00:12:43 Microsoft
```

Figura A1.3 Importancia de variables o características.
Fuente: Elaboración propia, software R

Anexo 2. Código para entrenar y validar modelo *Random Forest*

```
# Carga librerias

library(rattle)

rattle()


# Carga base de datos
fname <- "file:///C:/Users/Microsoft/Documents/Karina/2019/Tesis/Datos/Da
tos1.csv"

crs$dataset <- read.csv(fname,

                  na.strings=c(".", "NA", "", "?"),

                  strip.white=TRUE, encoding="UTF-8")

# Crea semilla

crv$seed <- 42

set.seed(crv$seed)


# Selecciona variables para el modelo

crs$input     <- c("Superficiekm2", "Poblacion", "Poblacion_adulta",

                   "TotalSucursales", "Cajeros", "TerminalesPuntodeVenta"
,

                   "TransaccionesEnTPV", "TransaccionesEnCajerosAutom",

                   "Clientes", "Grupos", "SucursalesRemesadora",

                   "SucursalesMicrofinanciera", "DepyPagos", "TipoPoblaci
onN")

crs$target    <- "Indicador_conjunto"


# Preparación de datos, 70 % para entrenar (train) y 30% para probar (tes
t)

crs$nobs     <- nrow(crs$dataset)

crs$train    <- crs$sample <- sample(crs$nobs, 0.7*crs$nobs)
```

```
crs$validate <- sample(setdiff(seq_len(crs$nobs), crs$train), 0.15*crs$no
bs)

crs$test     <- setdiff(setdiff(seq_len(crs$nobs), crs$train), crs$valida
te)



# Entrenamiento del modelo con 150 Árboles

set.seed(crv$seed)

crs$rf <- randomForest::randomForest(Indicador_conjunto ~ .,

  data=crs$dataset[crs$sample, c(crs$input, crs$target)],
  ntree=200,
  mtry=4,
  importance=TRUE,
  na.action=randomForest::na.roughfix,
  replace=FALSE)


# Genera una matriz de error para el modelo Bosque aleatorio.

# Respuesta del modelo Bosque aleatorio.

crs$pr <- predict(crs$rf, newdata=na.omit(crs$dataset[c(crs$input, crs$ta
rget)]))

# Genera matriz de confusión

rattle::errorMatrix(na.omit(crs$dataset[c(crs$input, crs$target)])$Indica
dor_conjunto, crs$pr, count=TRUE)



# Genera matriz de confusión en proporciones

(per <- rattle::errorMatrix(na.omit(crs$dataset[c(crs$input, crs$target)]
)$Indicador_conjunto, crs$pr))

# Matriz de error para el modelo Bosque aleatorio en Datos1.csv (cuentas)
:

 Predicted

Actual  Bueno Malo Error

  Bueno   533   27   4.8

  Malo     32 1866   1.7
```

```
# Evaluación con datos al 3er trimestre 2019

# Carga datos

crs$testset <- read.csv("C:/Users/Microsoft/Documents/Karina/2019/Tesis/Datos/Datos1f.csv", na.strings=c(".", "NA", "", "?"), header=TRUE, sep="," , encoding="UTF-8", strip.white=TRUE)



# Obtenga calificaciones de probabilidad para el modelo Bosque aleatorio en Datos1.csv

crs$pr <- predict(crs$rf, newdata=na.omit(crs$testset[,c(crs$input),drop=FALSE]))



# Extraer las variables relevantes del conjunto de datos.

sdata <- subset(crs$testset[,], select=c())



# Sacar los datos combinados.

write.csv(cbind(sdata, crs$pr), file="C:/Users/Microsoft/Documents/Karina/2019/Tesis/Datos/Datos1_score_idents.csv", row.names=FALSE)
```

Anexo 3. Experimentos del modelo

Utilizando como número de variables aleatorias la raíz cuadrada del número total de variables predictoras, es decir 8 variables, tenemos los siguientes experimentos para elegir el modelo que presente el mejor resultado clasificando los sitios para la

localización de sucursales bancarias:

Experimento 1: Entrenamiento con 50 árboles, 8 variables

	Predicho	
Real	Bueno	Malo
Bueno	533	27
Malo	40	1858
Sensibilidad	95.2%	
Especificidad	97.9%	
Exactitud	97.3%	

Tabla A2.1. Matriz de confusión
Fuente: Elaborado con datos de respuesta software R

Experimento 2: Entrenamiento con 150 árboles, 8 variables

	Predicho	
Real	Bueno	Malo
Bueno	531	29
Malo	37	1861
Sensibilidad	94.8%	
Especificidad	98.1%	
Exactitud	97.3%	

Tabla A2.2. Matriz de confusión
Fuente: Elaborado con datos de respuesta software R

Experimento 3: Entrenamiento con 200 árboles, 8 variables

	Predicho	
Real	Bueno	Malo
Bueno	532	28
Malo	38	1860
Sensibilidad	95.0%	
Especificidad	98.0%	
Exactitud	97.3%	

Tabla A2.3. Matriz de confusión
Fuente: Elaborado con datos de respuesta software R

Experimento 4: Entrenamiento con 500 árboles, 8 variables

	Predicho	
Real	Bueno	Malo
Bueno	532	28
Malo	35	1863
Sensibilidad	95.0%	
Especificidad	98.2%	
Exactitud	97.4%	

Tabla A2.4. Matriz de confusión
Fuente: Elaborado con datos de respuesta software R

Experimento 5: Entrenamiento con 1000 árboles, 8 variables

	Predicho	
Real	Bueno	Malo
Bueno	532	28
Malo	36	1862
Sensibilidad	95.0%	
Especificidad	98.1%	
Exactitud	97.4%	

Tabla A2.5. Matriz de confusión
Fuente: Elaborado con datos de respuesta software R

Experimento 6: Entrenamiento con 2000 árboles, 8 variables

	Predicho	
Real	Bueno	Malo
Bueno	532	28
Malo	36	1862
Sensibilidad	95.0%	

Especificidad	98.1%
Exactitud	97.4%

Tabla A2.6. Matriz de confusión
Fuente: Elaborado con datos de respuesta software R

Cuando el número de variables aleatorias es la mitad de la raíz cuadrada del número total de variables predictoras, es decir 2 variables, tenemos los siguientes experimentos:

Experimento 1: Entrenamiento con 50 árboles, 2 variables

	Predicho	
Real	Bueno	Malo
Bueno	531	29
Malo	34	1864
Sensibilidad	94.8%	
Especificidad	98.2%	
Exactitud	97.4%	

Tabla A2.7. Matriz de confusión
Fuente: Elaborado con datos de respuesta software R

Experimento 2: Entrenamiento con 150 árboles, 2 variables

	Predicho	
Real	Bueno	Malo
Bueno	533	27
Malo	34	1864
Sensibilidad	95.2%	
Especificidad	98.2%	
Exactitud	97.5%	

Tabla A2.8. Matriz de confusión
Fuente: Elaborado con datos de respuesta software R

Experimento 3: Entrenamiento con 200 árboles, 2 variables

	Predicho	
Real	Bueno	Malo
Bueno	533	27
Malo	33	1865
Sensibilidad	95.18%	

Especificidad	98.26%
Exactitud	97.56%

Tabla A2.9. Matriz de confusión
Fuente: Elaborado con datos de respuesta software R

Experimento 4: Entrenamiento con 500 árboles, 2 variables

	Predicho	
Real	Bueno	Malo
Bueno	533	27
Malo	31	1867
Sensibilidad	95.18%	
Especificidad	98.37%	
Exactitud	97.64%	

Tabla A2.10. Matriz de confusión
Fuente: Elaborado con datos de respuesta software R

Experimento 5: Entrenamiento con 1000 árboles, 2 variables

	Predicho	
Real	Bueno	Malo
Bueno	532	28
Malo	31	1867
Sensibilidad	95.0%	
Especificidad	98.4%	
Exactitud	97.6%	

Tabla A2.11. Matriz de confusión
Fuente: Elaborado con datos de respuesta software R

Experimento 6: Entrenamiento con 2000 árboles, 2 variables

	Predicho	
Real	Bueno	Malo
Bueno	532	28
Malo	32	1866
Sensibilidad	95.0%	
Especificidad	98.3%	
Exactitud	97.6%	

Tabla A2.12. Matriz de confusión
Fuente: Elaborado con datos de respuesta software R

Anexo 4. Código para entrenar y validar modelo *Gradient Boosting Machine*

```
# Carga librerías

library(ggplot2)

library(gridExtra)

library(h2o)

library(caret)
```

```
# Carga base de datos
Datos<-read.csv('C:/Users/Microsoft/Documents/Karina/2019/Tesis/Datos/GBM
/Datos1.csv')



# Variables que se utilizarán en el modelo

vars<-c(

  'Superficiekm2',

  'Poblacion',

  'Poblacion_adulta',

  'TotalSucursales',

  'Cajeros',

  'TerminalesPuntodeVenta',

  'TransaccionesEnTPV',

  'TransaccionesEnCajerosAutom',

  'Clientes',

  'Grupos',

  'SucursalesRemesadora',

  'SucursalesMicrofinanciera',

  'DepyPagos',

  'TipoPoblacionN',

  'Indicador')
```

```
# Selecciona las variables que se utilizarán en el modelo y cambia la var
iable target o indicador a factor

Datos_L<-Datos[,c(which(colnames(Datos)%in%(vars)))]

Datos_L$Indicador<-as.factor(Datos_L$Indicador)


# Preparación de datos, 70 % para entrenar (train) y 30% para probar (tes
t)
datos_7_a_3<-Datos_L

set.seed(1234)

train<-sample(1:nrow(datos_7_a_3),round(.7*nrow(datos_7_a_3)))

TrainingSample<- datos_7_a_3[train,]

TestingSample<- datos_7_a_3[-train,]



# Inicializar h2o y guarda bases en h2o datatable

h2o.init(nthreads=-1)

train.h2o <- as.h2o(TrainingSample)

test.h2o <- as.h2o(TestingSample)



#Modelo h2o, con 1,000 árboles

indep<-c(which(colnames(train.h2o)%in%(vars)))

y<-c(which(colnames(train.h2o)%in%("Indicador")))



system.time(

  gbm.model_Localiza <- h2o.gbm(y=y, x=indep, training_frame = train.h2o,
ntrees = 1000, max_depth = 5, learn_rate = 0.01, seed = 1122)

)

Mex_Gbm_Localiza<-h2o.saveModel(object=gbm.model_Localiza, path="C:/Users
/Microsoft/Documents/Karina/2019/Tesis/Datos/GBM", force=TRUE)
```

```
#Guarda Modelo h2o

ModeloLocaliza<-h2o.loadModel(C:/Users/Microsoft/Documents/Karina/2019/Tesis/Datos/GBM/ModeloLocalizacion')


#Matriz de confusión del modelo h2o

ModeloLocaliza


#Evalúa con datos de prueba

predictLocaliza <- as.data.frame(h2o.predict(ModeloLocaliza, test.h2o))

confusionMatrix(TestingSample$Indicador,predictLocaliza$predict)

sdata <- subset(TestingSample[,], select=c("Indicador"))

a<-data.frame(sdata,predictLocaliza$p1)


# Guarda resultados de calificar datos de prueba

write.csv(a, file="C:/Users/Microsoft/Documents/Karina/2019/Tesis/Datos/GBM/ResultadosTest.csv", row.names=FALSE)


#Evalúa utilizando todos los datos

library(data.table)


#Carga todos los datos

DatosTotal <- fread("C:/Users/Microsoft/Documents/Karina/2019/Tesis/Datos/GBM/Datos1.csv",stringsAsFactors = TRUE)

DatosTotal$Indicador<-as.factor(DatosTotal$Indicador)


# Guarda bases en h2o datatable

nvo.h2o <- as.h2o(DatosTotal)

nvo.h2o$Indicador<-as.factor(nvo.h2o$Indicador)
```

```
#Evalúa con todos los datos

predict.gbm <- as.data.frame(h2o.predict(ModeloLocaliza, nvo.h2o))

confusionMatrix(DatosTotal$Indicador,predict.gbm$predict)

sdata <- subset(DatosTotal[,], select=c("ClaveMunicipio","Indicador"))

a<-data.frame(sdata,predict.gbm$p1)


#Guarda resultados

write.csv(a,
file="C:/Users/Microsoft/Documents/Karina/2019/Tesis/Datos/GBM/EvalTodoOK
.csv", row.names=FALSE)


#Evalúa con datos de septiembre 2019

library(data.table)


#Carga todos los datos

Datos2018 <- fread("C:/Users/Microsoft/Documents/Karina/2019/Tesis/Datos/
GBM/Datos3_18.csv",stringsAsFactors = TRUE)

Datos2018$Indicador<-as.factor(Datos2018$Indicador)


# Inicaliza h2o

h2o.init(nthreads=-1)


# Guarda bases en h2o datatable

nvo.h2o <- as.h2o(Datos2018)

nvo.h2o$Indicador<-as.factor(nvo.h2o$Indicador)


#Evalúa con todos los datos

predict.gbm <- as.data.frame(h2o.predict(ModeloLocaliza, nvo.h2o))

confusionMatrix(Datos2018$Indicador,predict.gbm$predict)
```

```
sdata <- subset(Datos2018[,], select=c("IdPersona","Indicador"))

a<-data.frame(sdata,predict.gbm$p1)



#Guarda resultados

write.csv(a,
file="C:/Users/Microsoft/Documents/Karina/2019/Tesis/Datos/GBM/EvalTodoSEP18.csv", row.names=FALSE)
```

Anexo 5. Aplicaciones en R

Parámetros de bosque aleatorio en R

Los principales indicadores utilizados en el bosque aleatorio en R son:

- *Mtry*: es el número de variables seleccionadas en cada división, por defecto es la raíz cuadrada del número de variables independientes para el modelo de clasificación y es el número de variables dividido entre 3 para el modelo de regresión.
- *Ntree*: es el número de árboles a crear, por default se tiene que es igual a 500
- *Nodesize*: es el tamaño mínimo de los nodos terminales, por default es igual a 1
- *Replace*: se utiliza para comprobar si el muestreo es con o sin reemplazo. TRUE implica con reemplazo. FALSO implica sin reemplazo. Cabe mencionar que TRUE se establece de forma predeterminada.
- *Sampsize*: es el tamaño de la muestra a extraer de los datos para el crecimiento de cada árbol de decisión. Por defecto, toma el 63.2% de los datos.

Parámetros de Máquina de aumento de gradiente en R

Los principales indicadores utilizados en el Máquina de aumento de gradiente en R son:

- *model_id*: (opcional) Especifique un nombre personalizado para que el modelo lo use como referencia. Por defecto, H2O genera automáticamente una clave de destino.
- *training_frame*: (obligatorio) Especifique el conjunto de datos utilizado para construir el modelo.
- *validation_frame*: (opcional) Especifique el conjunto de datos utilizado para evaluar la precisión del modelo.
- *y*: (obligatorio) Especifique la columna que se usará como variable dependiente. Los datos pueden ser numéricos o categóricos.
- *x*: Especifique un vector que contenga los nombres o índices de las variables de predicción que se usarán al construir el modelo. Si falta x, se usan todas las columnas excepto *y*.
- *ntrees*: especifique la cantidad de árboles para construir.
- *max_depth*: especifica la profundidad máxima del árbol.
- *seed:* especifique el generador de números aleatorios para los componentes del algoritmo que dependen de la aleatorización. La semilla es consistente

para cada instancia de H2O para que pueda crear modelos con las mismas condiciones de inicio en configuraciones alternativas.

- *learn_rate*: especifica la velocidad de aprendizaje. El rango es de 0.0 a 1.0.

Referencias

1. Alarcón, J. (2017). *Modelos de minería de datos: random forest y adaboost, para identificar los factores asociados al uso de las TIC (internet, telefonía Fija y televisión de paga) en los hogares del Perú. 2014*. Universidad Nacional Mayor de San Marcos. Lima, Perú.

2. Alarcón, Z. (2012). *Desarrollo un modelo de localización de servicios bi - nivel y su algoritmo de solución*. Universidad Nacional Autónoma de México.

3. Amat, J. (2018). *Machine Learning con H2O y R*, sitio web: http://rstudio-pubs-static.s3.amazonaws.com/406480_299f4d339c96450790ecd7982b052f69.html.

4. Arevalillo, J. (2000). *La localización: concepto, nuevas tecnologías y requisitos del nuevo traductor de informática*. Universidad de Alcalá de Henares.

5. Avery, R. (1991). *Deregulation and the Location of Financial Institution Offices*. Federal Reserve Bank of Cleveland Economic Review.

6. Bancomer (2016). *Anuario de migración de remesas México 2016*. BBVA Bancomer.

7. Bancomer (2017). *Anuario de migración de remesas México 2017*. BBVA Bancomer.

8. Bancomer (2018). *Anuario de migración de remesas México 2018*. BBVA Bancomer.

9. Banco Mundial (2017). *Sucursales de bancos comerciales (por cada 100.000 adultos)*, Sitio web: https://datos.bancomundial.org/indicador/FB.CBK.BRCH.P5?view=chart

10. Banxico (2018). *Ingresos por remesas, distribución por municipio*, Sitio web: http://www.banxico.org.mx/SieInternet/consultarDirectorioInternetAction.do?accion=consultarCuadro&idCuadro=CE166&locale=es#

11. Benalcázar, J. (2017). *Análisis comparativo de metodologías de minería de datos y su aplicabilidad a la industria de servicios*. Universidad De Las Américas.
12. Beck, T., Demigurc-Kunt, A. & Martínez, M. (2006*). Reaching Out: Access to and Use of Banking Services Across Countries*. World Bank Policy Research Working Paper

13. Bermejo, E. (2017). *Machine Learning. El SMART Digital Workplace como ventaja competitiva*. Roana. https://es.slideshare.net/raona/machine-learning-whitepaper

14. Bhalla, D. (2014). *Random Forest in R: Step by Step Tutorial*, Sitio web: http://www.listendata.com/2014/11/random-forest-with-r.html

15. Bhalla, D. (2017). *Feature selection: select important variables with boruta package*, Sitio web: http://www.listendata.com/2017/05/feature-selection-boruta-package.html

16. Bosque, J. & Franco, S. (1995). *Modelos de localización - asignación y evaluación multicriterio para la localización de instalaciones no deseables*. Serie Geográfica, n° 5, pp. 97-112.

17. Bosque, S. & García, R. (2000). *El uso de los sistemas de información geográfica en la planificación territorial*. Anales de Geografía de la Universidad Complutense, 20, 49-67.

18. Breiman, L. (2004). *Consistency for a simple model of random forests*. University of California at Berkele, Statistics Department.

19. Carro, R. & González, D. (2013). *Localización de instalaciones*. Facultad de Ciencias económicas y Sociales. Universidad Nacional del Mar de Plata.

20. Castellanos, S., Castellanos, V. & Flores, N. (2009). *Factores de influencia en la localización regional de infraestructura bancaria*. Economía mexicana. Nueva época, vol.18. pp.20.

21. Church, R., & Roberts, K. (1983), *Generalized Coverage. Models and Public Facility Location*, Papers of the Regional Science Association, pp. 117-135.

22. Cinar, N. (2009). *A Decision Support Model for Bank Branch Location Selection*. World Academy of Science, Engineering and Technology.

23. CNBV. (2017). *Base de Datos de Inclusión Financiera Acceso diciembre de 2017*, Sitio web: http://www.cnbv.gob.mx/Inclusi%C3%B3n/Paginas/Bases-de-Datos.aspx
24. CNBV. (2018). *Información de sucursales, tdc, cajeros y otras variables*, Sitio web: http://portafoliodeinformacion.cnbv.gob.mx/bm1/Paginas/infoper.aspx.

25. Consoni, E., Taylor P. (2007). *Gateway cities: círculos bancarios, concentración y dispersión en el ambiente urbano brasileño*. Revista eure (Vol. XXXIII, Nº 100), pp. 115-133. Santiago de Chile.

26. De la Fuente, S. (2011). *Análisis de Conglomerados*. Universidad Autónoma de Madrid, Facultad de Ciencias Económicas y Empresariales.

27. Díaz, J. & Pineda, J. (2013). *Geografía de las sucursales bancarias en el Área Metropolitana de Toluca, 1989-2009*.Universidad Autónoma del Estado de México, México.

28. Espinar, R (2018). *Modelos de Clasificación datos no balanceados*, Universidad de Sevilla, Facultad de Matemáticas, Departamento de Estadística e Investigación Operativa, Sevilla

29. Espinosa, C. (2013). *Localización de zonas potenciales para nueva infraestructura bancaria en la Parroquias Urbanas del Norte de la Cuidad de Quito utilizando análisis espaciales: Línea de investigación: La dimensión territorial. Uso sustentable del espacio*. Pontificia Universidad Católica del Ecuador

30. Esquivel, E. (2015). *La importancia de las remesas y la urgente necesidad del desarrollo regional. SDPnoticias,* http://www.sdpnoticias.com/nacional/2015/09/03/la-importancia-de-las remesas-y-la-urgente-necesidad-del-desarrollo-regional.

31. Frank, I. & Friedman, J. (1993). *A statistical view of some chemometries regression tools*, Technometrics.

32. Fu, W., (1998). *Penalized regressions: The bridge versus the lasso*, Journal of Computational and Graphical Statistics.

33. González, A. (2015). *Selección de variables: Una revisión de métodos existentes*. Universidad de Coruña.

34. González, R. (2017, octubre 17). *El beneficio de la economía no llega a todos: Santander*. La Jornada. Recuperado de: http://www.jornada.unam.mx/2017/10/17/economia/025n1eco

35. *GBM (2018),* http://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science/gbm.html

36. Granville, V. (2017). *Machine Learning Summarized in One Picture*, https://www.datasciencecentral.com/profiles/blogs/machine-learning-summarized-in-one-picture?utm_content=buffer0a1f7&utm_medium=social&utm_source=plus.google.com&utm_campaign=buffer

37. Grupo del Banco Mundial (2016). *Sucursales de bancos comerciales (por cada 100.000 adultos)*. http://datos.bancomundial.org/indicador/FB.CBK.BRCH.P5.

38. Gutiérrez, L. (2017). *Generar un modelo para clasificar las escuelas primarias de México por su rendimiento con tecnología big data que ayude crear a estrategias para mejorar los resultados*. Universidad Cuauhtémoc.

39. Hernández, R., Fernández, C., & Baptista, M. (2010). *Metodología de la Investigación* (5ta. Ed.). McGraw-Hill/Interamericana Editores S.A. de C.V.

40. Hotelling, H. (1929). *Stability in competition.* Economic Journal, 39, Royal Economic Society, St. Andrews, Escocia, pp. 41-57.

41. IBM (2012). *Manual CRISP-DM de IBM SPSS.* Modeler 15.

42. Iprofesional (2018). *Clientes de los bancos todavía prefieren las sucursales sobre los canales digitales*. https://www.iprofesional.com/notas/271197-Por-que-clientes-de-los-bancos-todavia-prefieren-las-sucursales-sobre-los-canales-digitales.

43. James, G., Witten, D., Hastie, T. & Tibshirani, R. (2013). *An Introduction to Statistical Learning, with applications in R*. Springer.

44. Laurentio, A., Sanchez, J. & San Gabino, D. (2014). *Entorno de ejecución de algoritmos inteligentes de Mahout para Hadoop y Big Data*. Universidad Complutense de Madrid,

http://eprints.ucm.es/26534/1/Memoria%20easyMahout%20-%20Sistemas%20Inform%C3%A1ticos.pdf.

45. Love, R., Morris, J. & Wesolowsky, G. (1988). *Facilities Location. Models and Methods*. New York, Elsevier Science Punlishing Co.

46. Maskell, P. & Lorenzen, M. (2004). *The cluster as market organization, Urban Studies*, Vo.41, Número 5/6.

47. *México cuenta con 123.5 millones de habitantes* (2017), https://www.eleconomista.com.mx/politica/Mexico-cuenta-con-123.5-millones-de-habitantes-20170710-0116.html

48. Peña E. (2012). *Acceso a servicios financieros y reducción de la pobreza extrema en el sector rural colombiano (2009-2011).* Universidad Complutense de Madrid.

49. Prior F., (2006). *Desarrollo de servicios microfinancieros en México: propuesta de modelo de distribución de servicios microfinancieros de bajo costo para segmentos de bajos ingresos*. Universidad Autónoma de Madrid.

50. Ramírez, L. & Bosque, J. (2001). *Localización de hospitales: Analogías y diferencias del uso del modelo p-mediano en SIG raster y vectorial*. Anales de Geografía de la Universidad Complutense.

51. Ravanshad, A. (2018). *Gradient Boosting vs Random Forest*. Recuperado de: https://medium.com/@aravanshad/gradient-boosting-versus-random-forest-cfa3fa8f0d80

52. Rodríguez, H., Peralta, I. & Delgado, D. (2015). *Análisis de mercado de las sucursales bancarias en la Ciudad de Toluca con técnicas de geomarketing. En Pasado, presente y futuro de las regiones en México y su estudio*. Asociación Mexicana de Ciencias para el Desarrollo Regional A.C.

53. Rodríguez, K. (2014).*Abrir un banco en México cuesta 33 mdd,* http://www.cnnexpansion.com/negocios/2014/07/10/abrir-un-banco-en-mexico-cuesta-33-mdd.

54. Ruiz B. (2019). *Ante digitalización de la banca, sucursales bancarias aún son importantes*. https://www.mypress.mx/negocios/digitalizacion-banca-sucursales-bancarias-son-importantes-4921.

55. Sánchez, J. & Zamarripa, G. (2015). *Un análisis sobre la infraestructura bancaria en México*. FUNDEF, México, D.F.

56. SAS (2016). *Enterprise Miner Glossary.* SAS Institute Inc.

57. Schneider, S., Seifert, F., & Sunyaev, A. (2014). *Market Potential Analysis and Branch Network Planning: Application in a German Retail Bank*.

58. Singh, H (2018), *Understanding Gradient Boosting Machines,* https://towardsdatascience.com/understanding-gradient-boosting-machines-9be756fe76ab

59. Soco, M. (2017). *Random Forest, ¿se puede predecir en base a estadísticas en qué municipios de México puede haber fosas clandestinas* ?, Sitio web: https://www.xataka.com.mx/otros-1/random-forest-se-puede-predecir-en-base-a-estadisticas-en-que-municipios-de-mexico-puede-haber-fosas-clandestinas.

60. Steiner, R., y C. Medina (2002). *Oferta de servicios financieros a los pobres bancables en Colombia*. Facultad de Economía, Universidad de los Andes.

61. Tibshirani, R. (1996). *Regression shrinkage and selection via the lasso*, Journal of the Royal Society.

62. Tong D., Murray A., Xiao N., Hernández C. (2009). *Heuristics in Spatial Analysis: A Genetic Algorithm for Coverage Maximization. Annals of the Association of American.*

63. Torgo, L. (2010). *Data Mining with R Learning with Case Studies*. Chapman & Hall/CRC.

64. *Tres de cada cuatro usuarios de telefonía celular en México tienen smartphone: INEGI* (2017). https://www.sdpnoticias.com/tecnologia/2017/03/15/3-de-cada-4-usuarios-de-telefonia-celular-en-mexico-tienen-smartphone-inegi

65. Vallejo, E. (2017). *Algoritmos de Machine Learning. Data Science for Business.* Tecnológico de Monterrey.

66. Videgaray, L. (2014, junio 23). *Una banca para todos. El Universal.* Recuperado de: http://www.eluniversalmas.com.mx/editoriales/2014/06/70939.php

67. Yoshibauco (2011). *Diferencia entre CRISP y SEMMA*. Recuperado de: https://yoshibauco.wordpress.com/2011/03/07/diferencia-entre-crisp-y-semma/

68. Zamora, D. (2015). *Metodología para la localización de servicios de emergencia, caso México – Toluca*. Universidad Nacional Autónoma de México.

69. Zhang, Y. (2015). *What are the differences between bagged trees and random forests?,* Recuperado de: https://www.quora.com/What-are-the-differences-between-bagged-trees-and-random-forests