Sistema de clasificación para afiliados en las Cámaras de Comercio

PROYECTO DE GRADO

Camilo José Ibarra Escobar

Asesor
Esteban Piedrahita Uribe

FACULTAD DE INGENIERÍA
MAESTRÍA EN CIENCIA DE DATOS
SANTIAGO DE CALI
2020

Sistema de clasificación para afiliados en las Cámaras de Comercio

Camilo José Ibarra

Trabajo de grado para optar al título de
Máster en Ciencia de Datos

Asesor
Esteban Piedrahita Uribe

FACULTAD DE INGENIERÍA
MAESTRÍA EN CIENCIA DE DATOS
SANTIAGO DE CALI
2020

CONTENIDO

pág.

LISTA DE TABLAS

LISTA DE GRÁFICOS

LISTA DE ANEXOS

RESUMEN

Las Cámaras de Comercio en Colombia tienen una alta dependencia de ingresos por concepto de registro mercantil, el cual es susceptible a cambios por factores externos. Entre las alternativas de ingreso de la organización se encuentran una serie de programas y servicios para empresas que se encuentran en la categoría B2B.

El problema es que en las cámaras de comercio de Colombia y en particular la Cámara de Comercio de Cali, a pesar de tener información suficiente sobre los posibles clientes y los resultados de campañas anteriores, tienen un bajo nivel de captación o de venta de la membresía, que llega sólo al 5,6% del total de empresas, con campañas comerciales que registran un porcentaje de éxito entre 20-30%.

Para abordar este problema se parte de un entendimiento del negocio para obtener la información necesaria para entrenar modelos que permitan realizar la clasificación de clientes. En este trabajo se emplean tres modelos diferentes con diversas técnicas de clasificación (regresión logística, regresión logística paso a paso, árboles de decisión, *Random Forest*, *Robust Random Forest*, árboles de decisión C5.0, *KNN*).

Como resultado se encuentran dos modelos que tras ser validados conservan sus propiedades y podrán ser utilizados para diseñar campañas comerciales más efectivas que las que emplea actualmente la organización.

1. INTRODUCCIÓN

1.1 Contexto y Antecedentes

El origen de las cámaras de comercio se remonta a la edad media en Europa, cuando las empresas crearon entidades que las representaran ante las autoridades locales (Pérez Ramírez, 2014). En Colombia desde 1931[1] las cámaras de comercio son entidades privadas sin ánimo de lucro que recaudan y operan recursos públicos derivados del registro mercantil, que actualmente se encuentra reglamentado por el decreto 650 de 1996 (Presidencia, 1996).

Para las cámaras de comercio, el registro mercantil no es su única fuente de ingreso, pero si la más importante, usualmente superando 70% del total[2]. El flujo de estos ingresos depende de factores externos cómo la legislación existente y la coyuntura económica. Por ejemplo, las leyes 1429 de 2010 y 1780 de 2016 reglamentan una serie de beneficios para ciertas empresas nuevas que incluían exenciones a en el pago de matrículas y disminución en el pago de la renovación de la matricula mercantil. Por su parte, la coyuntura económica influye en las expectativas de los empresarios y por lo tanto en la creación o liquidación de empresas.

En 2019 el Consejo Nacional de Política Económica y Social (CONPES) publicó el documento No. 3956 sobre la política de formalización empresarial, en el cual se comparan las tarifas de registro mercantil en varios países. En este análisis se concluye que en Colombia el registro mercantil resulta mucho más costoso para las empresas pequeñas, además, en otros países no se realiza un pago por su renovación. Al respecto recomienda implementar una reforma en la estructura de pago del registro que mitigue el impacto en la formalización empresarial.

[1] Ley 28 de 1931
[2] 71,4% en Bogotá, 77,9% en Cali y 84,1% en Medellín

Debido a que este tipo de factores afectan el principal rubro de ingreso, para las cámaras de comercio una de las metas de largo plazo consiste en diversificar las fuentes de ingresos mediante la oferta de productos de empresa a empresa (*B2B* por sus siglas en inglés), lo que permitirá continuar con el objetivo misional que es acompañar el desarrollo económico de la región a través del crecimiento empresarial.

Entre los ingresos generados por las Cámaras de Comercio, uno de los rubros más importantes corresponde al pago por afiliación. Las empresas afiliadas adquieren, por un valor adicional, una membresía que les brinda una serie de beneficios (descuentos en programas, espacios de *coworking*, certificados, entre otros) y derechos (voto para elecciones junta directiva). Los afiliados son empresas que cumplen una serie de características mínimas exigidas por la ley[3] y, en algunos casos, se consideran requisitos adicionales definidos por cada Cámara de Comercio.

Las membresías o afiliaciones son muy importantes en las cámaras de comercio del resto del mundo. De hecho, cámaras como las de Nueva York, Londres, Dubái, Tokio, Sídney, entre otras, funcionan a partir de las membresías de sus afiliados y no del pago de una tasa como en Colombia. En Colombia los afiliados de las principales cámaras de comercio no representan más del 6%[4] del total de las empresas, cifra que puede aumentar teniendo en cuenta que se dispone de la información de todas ellas.

[3] Ley 1727 de 2014
[4] Bogotá 2,4%, Medellín 2,9%, Cali 5,6%, Bucaramanga 5,5% y Barranquilla 2,5%, información suministrada por la Cámara de Comercio de Cali

1.2 Planteamiento del Problema

Las Cámaras de Comercio en Colombia tienen una alta dependencia de ingresos por concepto de registro mercantil, el cual es susceptible a cambios por factores externos. Entre las alternativas de ingreso de la organización se encuentran una serie de programas y servicios para empresas que se encuentran en la categoría *B2B*.

Uno de los programas más importantes por las tendencias globales y su participación en los ingresos de la organización es la afiliación, para el cual se diseñan campañas comerciales buscando empresas que cumplan una serie de requisitos mínimos.

El problema es que en las cámaras de comercio de Colombia y en particular la Cámara de Comercio de Cali, a pesar de tener información suficiente sobre los posibles clientes y los resultados de campañas anteriores, tienen un bajo nivel de captación o de venta de la membresía, que llega sólo al 5,6% del total de empresas, con campañas comerciales que registran un porcentaje de éxito entre 20-30%.

1.3 Objetivo General

Implementar un modelo que permita determinar los posibles clientes por su probabilidad de afiliación y que a su vez sea un insumo que aporte en el diseño de campañas comerciales más efectivas para el programa de afiliación en la Cámara de Comercio de Cali.

1.4 Objetivos Específicos

1. Identificar las variables (internas y externas) relevantes para el modelo de clasificación
2. Aplicar técnicas de clasificación para implementar modelos que permitan identificar posibles clientes

3. Seleccionar los modelos más eficientes para la identificación de clientes

1.5 Organización del Documento

En el capítulo dos se describe un conjunto de técnicas de clasificación ideales para alcanzar el objetivo propuesto y, además, se reseñan trabajos similares en los que se han empleado técnicas de clasificación con productos que se comercializan *B2B* y sus principales resultados.

En el tercer capítulo se describe como se adaptó la metodología CRISP-DM en este caso y la forma en la que se abordaron las distintas etapas del proceso, llegando hasta la evaluación del modelo.

En el capítulo cuatro se describen todos los elementos empleados en el proceso de entrenamiento: la información disponible, los modelos empleados, el protocolo de evaluación y la métrica de clasificación.

En el quinto capítulo se describe el experimento con el que se definirá si los modelos identificados son pertinentes para el cumplimiento del objetivo.

Finalmente, en el capítulo seis se presentan los resultados comparativos del experimento con los mejores modelos, y se analiza la viabilidad de su despliegue.

2. ANTECEDENTES

El problema a resolver consiste en identificar qué empresas tomarán la afiliación, lo cual desde el punto de vista del análisis estadístico se aborda como un problema de clasificación. En estos análisis se busca predecir variables de respuesta cualitativa, estimando la probabilidad de pertenecer a cada una de las posibles categorías.

Para alcanzar el objetivo de este trabajo se implementarán técnicas de aprendizaje supervisado, ya que conocemos de antemano la clasificación que se busca predecir. Las técnicas que se utilizarán son: regresión logística, árboles de decisión, *K-Nearest Neighbor* y algunas variaciones de estos.

2.1 Marco Teórico

2.1.1 Regresión logística

Una de las primeras técnicas empleadas para realizar análisis discriminante es la regresión logística. Esta técnica se basa en la función logística, desarrollada durante el siglo XIX para describir el crecimiento poblacional y algunas reacciones químicas, y que comenzó a aplicarse para análisis discriminante en la década de 1960 (Cramer 2002).

El objetivo de la regresión logística (James, Witten, Hastie & Tibshirani 2013) consiste en estimar la probabilidad de pertenecer a una categoría $\rho(X)$, en función de las variables explicativas o independientes ($X_1, X_2, \dots, X_\rho$) y su respectivo coeficiente $(\beta_1, \beta_2, \dots, \beta_\rho)$.

$$\log\left(\frac{\rho(X)}{1-\rho(X)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_\rho X_\rho$$

$$\rho(X) = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_\rho X_\rho}}{1 + e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_\rho X_\rho}}$$

Los resultados de una regresión logística pueden ser interpretados de manera individual y conjunta. Cada coeficiente indica la magnitud del efecto de la variación de las variables independientes en la clasificación final como puntos porcentuales. Además, es posible determinar la significancia estadística de cada una.

2.1.2 Árboles de decisión

Implementados por Breiman, Friedman, Olshen y Stone (1984) para resolver problemas de clasificación, los árboles de decisión segmentan o estratifican el espacio de las variables independientes con el objetivo de predecir la clase a la que pertenecen. Se caracterizan por su facilidad de interpretación y su manejo de variables categóricas sin necesidad de crear *dummies*.

La pertenencia de una observación a una clase es determinada por la clase más común de las demás observaciones ubicadas en el mismo espacio. Para esto realizan particiones binarias empleando el criterio de reducción de error de clasificación (E), el cual se minimiza cuando aumenta $\hat{p}_{mk}$, que es la proporción de observaciones en la región *m* que pertenecen a la clase *k* (James, Witten, Hastie & Tibshirani 2013).

$$E = 1 - \max_{k}(\hat{p}_{mk})$$

Aunque existen otros índices para evaluar las particiones durante el proceso de poda de los árboles de decisión, se recomienda utilizar la reducción del error de clasificación si el objetivo es realizar predicciones (James, Witten, Hastie & Tibshirani 2013).

Sin embargo, los árboles de decisión no son competitivos en términos de predicción frente a otras técnicas de aprendizaje supervisado. Para mejorar el ajuste de estas técnicas existen variaciones como *Random Forest* y *Ada Boost*, lo cual implica una pérdida en la capacidad de interpretación de los resultados.

2.1.3 Árboles de decisión – *Random Forest*

Los árboles de decisión registran una alta varianza, es decir que pueden arrojar resultados distintos, aunque se utilice el mismo conjunto de datos de entrenamiento. La técnica de *Random Forest* propuesta por Kam (1995) busca reducir la varianza en la predicción, empleando la técnica de *Bootstrap Aggregation* o *Bagging*, que consiste en emplear múltiples muestras del conjunto de datos de entrenamiento, para estimar varios árboles independientes y promediar sus resultados. Además, para cada estimación se emplean de manera aleatoria m predictores tal que $m = \sqrt{p}$, donde p corresponde al número total de predictores.

Este tipo de modelos no es fácilmente interpretable, pero permite conocer la importancia de las variables para realizar las distintas particiones (James, Witten, Hastie & Tibshirani 2013).

2.1.4 Árboles de decisión – ADA

Otra forma de mejorar los resultados de predicción de un árbol de decisión es el *boosting,* propuesto inicialmente por Freud y Schapire (1995). Esta técnica consiste en estimar varios modelos de árboles de decisión con resultados débiles de predicción, pero estos no son independientes sino secuenciales. A diferencia del *Random Forest*, en este caso no se emplea el muestreo de *bootstrap* (James, Witten, Hastie & Tibshirani 2013)

2.1.5 Árboles de decisión – C5.0

Los árboles de decisión realizan la partición de datos basados en la reducción del error de clasificación, en la que se buscan nodos más homogéneos o puros. Como alternativa, se han desarrollado algoritmos para establecer las particiones de datos

basados en la entropía, que consiste en darle más valor a lo altamente improbable, es decir, a las observaciones con valores menos frecuentes.

Quinlan (2014) propone un algoritmo para las particiones de los árboles de decisión basado en la entropía denominado C4.5. Una versión mejorada es el algoritmo C5.0 que agiliza el proceso de entrenamiento y permite ponderar de manera diferenciada los errores de clasificación (Rulequest, 2017)

2.1.6 K-Nearest Neighbours

El método de *K-Nearest Neighbours* desarrollado por Fix y Hodges (1951), determina a que clase pertenece una observación basado en la proximidad con las observaciones más cercanas. Teniendo en cuenta los valores registrados en las variables predictivas se identifican los *k* vecinos más cercanos y de acuerdo a su comportamiento realiza la clasificación.

Los vecinos más cercanos son identificados mediante distintos criterios de distancia, y esto dificulta la utilización de variables categóricas como predictores (James, Witten, Hastie & Tibshirani 2013).

2.2 Estado del arte

Las técnicas de analítica de datos tienen una gran variedad de aplicaciones en la gestión de relación con clientes (CRM por sus siglas en inglés). Una de ellas es en la identificación de clientes (Ngai, Xiu, & Chau, 2009). En este tipo de análisis el objetivo es identificar cuáles son los posibles clientes basados en sus características, para lo que se pueden emplear métodos de clasificación, reglas de asociación, *clustering* o regresión.

Es de interés en este caso centrarnos en los problemas de clasificación de clientes para modelos de negocio *B2B*. Sin embargo, estos suelen ser abordados por empresas que realizan la investigación de manera interna, y, por ende, los resultados no son de acceso público (Mortensen 2019).

La predicción del éxito de los posibles acercamientos con clientes empresariales usualmente se basa en la intuición de los vendedores o gerentes encargados. Según Monat (2011), el desperdicio de recursos debido a los pronósticos imprecisos puede reducirse empleando modelos cuantitativos basados en las características de los clientes. Con base en la literatura académica acerca de las ventas, el autor seleccionó 16 variables, entre las que se encuentran información financiera de la empresa, cargo de la persona de contacto y relación previa con esa empresa, entre otras. Con estas elaboró un modelo de análisis discriminante que obtuvo un *accuracy* de 65%.

Yan, Zhag, Zha, Gong y otros (2015) identificaron que los vendedores de servicios B2B disponen de poco tiempo para realizar sus actividades y se propusieron a evaluar si sus resultados podrían estar determinados por las características de sus clientes. Para esto emplearon técnicas de regresión logística, Cox, *Triggering Kernel Learning* y Hawks con un conjunto de variables demográficas de las empresas clientes: ubicación geográfica, valor del negocio, sector y producto. En este caso compararon los modelos a través de AUC, y se sugiere emplearlos para focalizar los sus recursos de acuerdo a las clasificaciones predichas. Los autores resaltan que el modelo empleado es generalizable a otros sectores e industrias debido a su flexibilidad, ya que las variables empleadas se pueden obtener fácilmente.

Bohanec, Borštnar y Robnik-Šikonja (2016) emplearon una metodología para desarrollar un modelo de clasificación (árbol de decisión, random forest y Naive Bayes) que permitiera disminuir el error de pronósticos de los tomadores de decisiones sobre posibles ventas B2B basado en la información histórica. El objetivo

fue encontrar modelos compactos y comprensibles para los agentes de ventas, lo cual justifica la utilización de modelos supervisados y estrategias de análisis exploratorio y validación con los tomadores de decisiones, cuya experticia fue empleada para utilizar solo las variables más relevantes. Los modelos estimados alcanzaron *accuracy* superior al 90%, pero los autores consideran que requieren validar con otro tipo de empresas e industrias la manera en que se realiza la depuración de variables, ya que al parecer cada caso tiene particularidades.

De manera similar Mortensen (2019) estudió el caso de selección de clientes empresariales para una empresa de papeles y empaques. En el proceso identificaron la importancia de la selección de variables en el proceso de limpieza de datos y su efecto en el algoritmo de clasificación. Empleando cuatro modelos de clasificación (regresión logística, arboles de decisión, *Random Forest* y *xgboost)* seleccionaron el *Random Forest* por registrar un accuracy de 82,4% y posteriormente adicionaron meta-variables que no se relacionaban directamente con el proceso de venta, si no con la calidad de los datos, lo cual permitió mejorar las métricas empleadas.

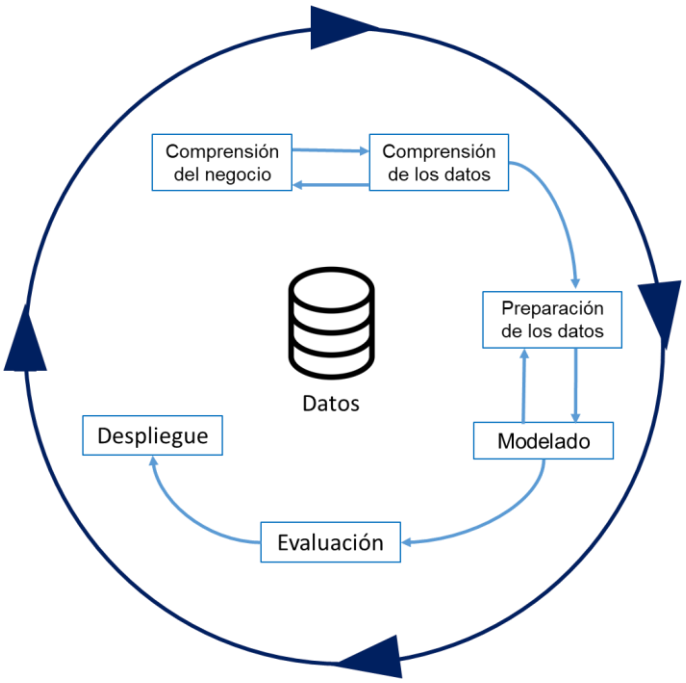
Tabla 1. Comparación de estudios relacionados

Autores	Sector	Variables	Técnicas	Resultado
Monat (2011) Industrial sales lead conversion modeling	Hardware y software (anónima)	Financieras, Cargo persona de contacto, relación previa con la empresa	Análisis Discriminante	Accuracy: 65%
Yan, J., Zhang, C., Zha, H., Gong, M., Sun, C., Huang, J., ... & Yang, X. (2015). On machine learning towards predictive sales pipeline analytics.	Tecnología (anónima)	Ubicación, valor de venta, sector y producto	Regresión logística, Cox, Triggering Kernel Learning, Hawks	AUC: 67-74
Bohanec, Borštnar y Robnik-Šikonja (2016) Integration of machine learning insights into organizational learning: A case of B2B sales forecasting	Software (Salvirt)	Se consideran 16 variables, las más importantes: negociación, reacción, autoridad del contacto, necesidad definida	Árbol de decisión, *Random Forest* y *Naive Bayes*	Accuracy: 93% - 96%
Mortensen, S., Christison, M., Li, B., Zhu, A., & Venkatesan, R (2019) Predicting and Defining B2B Sales Success with Machine Learning	Papel y empaques (anónima)	Se consideran 15 variables, las más importantes: tiempo de negociación, campos completos, ultima interacción, sector y tipo (cliente nuevo)	Regresión logística, árboles de decisión, *Random Forest* y *xgboost*	Accuracy 82,4%
Ibarra (2020) Sistema de clasificación para afiliados en las Cámaras de Comercio	Servicios (Cámara de Comercio de Cali)	Variables financieras (disponibles y creadas), edad, sector, número de establecimientos y otros productos	Regresión logística, árboles de decisión, Random Forest, Robust Random Forest,	F1: 0,51 - 0,60

3. METODOLOGÍA

Para abordar el problema objetivo de este trabajo se empleará la metodología CRISP-DM. Esta metodología es la más usada en los proyectos de ciencia de datos (Piatetsky-Shapiro, 2014), y su principal diferencia frente a otras metodologías (KDD y SEMMA) radica en que parte de la comprensión del negocio y no del análisis de datos. CRISP-DM consta de seis etapas, cuya secuencia no es estricta, y se caracteriza por ser versátil, ya que es posible regresar sobre ciertas fases con el fin de garantizar el éxito del proyecto (IBM 2017).

Diagrama 1. Ciclo de vida de un proyecto de analítica



Fuente: tomado de IBM (2017)

Las fases de la metodología CRISP-DM implican una serie de tareas y objetivos como se describe a continuación (Chapman, y otros, 2000):

Fase 1. Comprensión del negocio.

En esta fase se busca información relacionada con el estado actual del negocio, que permita entender el problema/objetivo desde el punto de vista comercial. Posteriormente, se identifican los criterios de éxito del proyecto según la perspectiva de los tomadores de decisiones, para finalmente plantear el problema en términos de ciencia de datos.

En este caso, la información del estado actual del negocio se obtuvo de cinco fuentes:

- Información pública disponible de la Cámara de Comercio de Cali (CCC) y sus afiliados.
- Gerente de fidelización de la CCC, persona a cargo del área, que tiene el objetivo de conservar y aumentar el número de afiliados. Proporcionó información sobre la manera en que opera el equipo y cómo han logrado cumplir los objetivos desde 2015, año en el que se implementó la Ley 1727 que modificó la regulación de este producto/servicio.
- Analista de información de la CCC, persona a cargo del procesamiento de información para el diseño de las campañas de afiliación. Compartió el proceso que se utiliza actualmente para identificar los posibles afiliados y diseñar las campañas que realiza el equipo de fidelización.
- Proyecto interno para identificar posibles afiliados. Uno de los equipos de la institución realizó un ejercicio para identificar cuales empresas eran las más propensas a afiliarse, basándose en la información disponible en el registro mercantil y empleando un árbol de decisión. Este modelo dio como resultado una tasa de asertividad de 10%, mientras que las campañas empleadas por el equipo alcanzan aproximadamente 30%.

- Un proyecto externo realizado en conjunto con la Cámara de Comercio de Barranquilla y ejecutado por la firma consultora Quantil, tiene como objetivo predecir cuáles empresas renovarán su registro mercantil.

El área de fidelización ha cumplido su meta de aumentar el número de afiliados cada año desde 2015, diseñando sus campañas basándose en los criterios indicados por la legislación y, los diseñados por el comité de afiliados, y a través de una gestión comercial. El diseño de las campañas se realiza filtrando la información disponible en el registro mercantil. Sin embargo, no emplean técnicas de analítica de datos en el proceso. Actualmente las campañas comerciales tienen una efectividad de 30%; por lo tanto, el modelo que resulte de este proyecto debe superar este valor.

A nivel interno se realizó un proyecto empleando analítica. Sin embargo, sus resultados no fueron satisfactorios. En conjunto con externos se está realizando un proyecto de analítica, pero sus resultados aún no están disponibles. En ambos casos la única información considerada es la disponible en el registro mercantil en un período y no se tiene en cuenta la disponibilidad de información histórica que puede dar luces sobre la condición/salud actual de la empresa y expectativas. Tampoco se emplea información externa o de otras áreas de la entidad.

Las decisiones de los empresarios están determinadas por el valor que perciben en lo que están comprando, así como por sus expectativas, es decir que su entorno y el momento en el que se encuentran influye en la toma de decisiones. Por este motivo se considera relevante incluir información que no se encuentra disponible en el registro mercantil, como si la empresa ha comprado o recibido otros servicios de la Cámara de Comercio de Cali, o si la empresa realiza exportaciones. Para una empresa exportadora, acreditar la afiliación a una cámara de comercio puede permitirle, por ejemplo, generar credibilidad entre posibles clientes extranjeros.

Fase 2. Comprensión de los datos

Es la parte más larga del proyecto, y consiste en el análisis exploratorio de los datos. En esta fase se consideran la recolección de datos y el análisis de las características relevantes de acuerdo con los temas definidos en la fase 1.

- Se solicitó la información correspondiente al registro mercantil.
- Se solicitó la información de las campañas realizadas por el área de fidelización desde 2017 y sus resultados.
- Se solicitó información de la participación de empresas en programas y proyectos de otras áreas, como las Iniciativas Cluster, Internacionalización, Sistemas de Innovación, Alianzas por la Innovación y Valle Impacta.
- Se recopiló información sobre comercio internacional de cada empresa

El análisis exploratorio de los datos se encuentra anexo al presente documento.

Fase 3. Preparación de los datos

Una vez se conocen los datos, es necesario seleccionarlos, acondicionarlos y transformarlos de acuerdo con lo que se considera necesario. Este proceso de limpieza debe documentarse, para tener la trazabilidad de los cambios realizados y garantizar la replicabilidad en el futuro.

En este caso se dispone de un conjunto de bases de datos estructuradas y previamente procesadas por cada una de las áreas responsables al interior de la organización. Estas bases fueron empalmadas y utilizadas para crear nuevas variables que se incluyeron en los modelos a entrenar.

El proceso de limpieza y transformación de los datos da como resultado una base de datos consolidada con la información recolectada en la Fase 2 y que es empleada para el entrenamiento de los modelos.

Fase 4. Modelado

En esta fase se seleccionaron las técnicas adecuadas, los protocolos de evaluación y las métricas de clasificación. Durante el proceso de modelado se realizan varias iteraciones, y se hizo necesario volver sobre las fases 2 y 3. Las técnicas utilizadas en esta fase dependen de los datos disponibles, los objetivos del proyecto y la literatura estudiada.

El detalle de esta etapa se encuentra descrito en capítulos posteriores del presente documento.

Fase 5. Evaluación

Se evaluarán los resultados obtenidos, se revisa el proceso y se identifican los pasos a seguir. Se debe tener en cuenta que los modelos deben ser replicables y los resultados obtenidos deben ser acordes con las métricas de éxito del proyecto que se enfocarán en la asertividad de la predicción de las empresas que se afilian durante las campañas además de la evaluación técnica que requiera cada modelo.

Fase 6. Despliegue

Las consideraciones para un posterior despliegue de los modelos identificados en este trabajo serán presentadas en el capítulo 7.

4. PRESENTACIÓN DE LA PROPUESTA

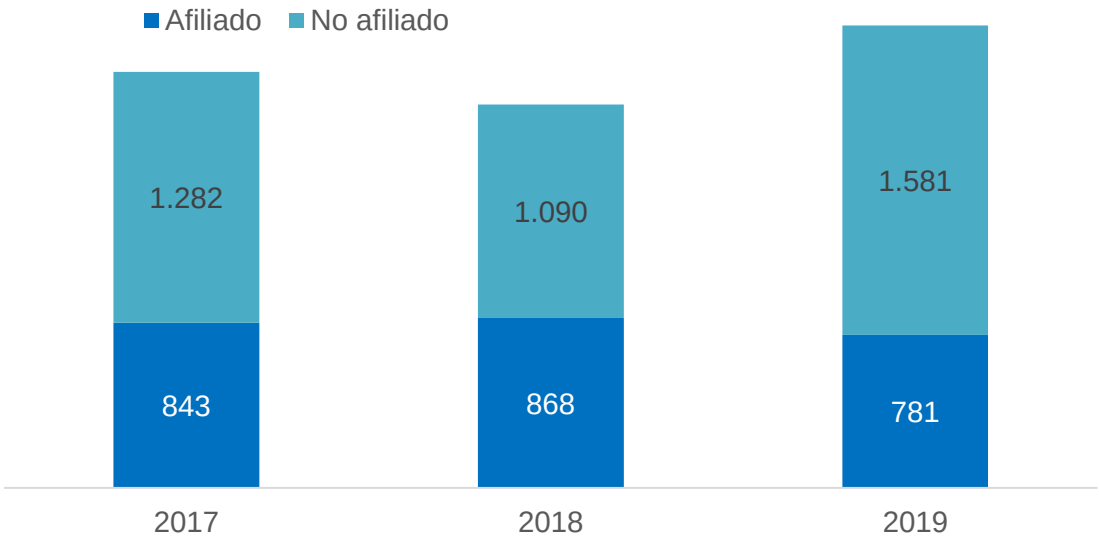
En este capítulo se describen los elementos empleados para el proceso de entrenamiento de los modelos de clasificación. Partiendo de la información disponible se proponen tres modelos con su respectivo protocolo de evaluación y métrica de clasificación.

4.1 Información disponible

Se dispone de información de las campañas de afiliación de empresas para los años 2017, 2018 y 2019. Los modelos serán entrenados con la información correspondiente a las campañas de 2017 y 2018 y validados con la información de 2019.

La campaña realizada en 2017 incluyó 2.125 empresas, de las cuales 39,7% adquirieron la afiliación, en 2018 fueron 1.958 con una tasa de éxito 44,3%, y para 2019 fueron 2.362, con una tasa de 33,1% (Figura 1).

Gráfico 1. Resultados campañas de afiliación 2017-2019



Fuente: Cámara de Comercio de Cali – Elaboración propia

De cada una de las empresas que hacen parte de estas campañas se tiene la información correspondiente al registro mercantil, exportaciones, participación en programas y adquisición de otros productos ofrecidos por la Cámara de Comercio de Cali. Adicionalmente, se calculó el crecimiento de las principales variables financieras de cada empresa durante el último año. Para ver las estadísticas descriptivas ver Anexo 1.

4.1.1 Registro mercantil

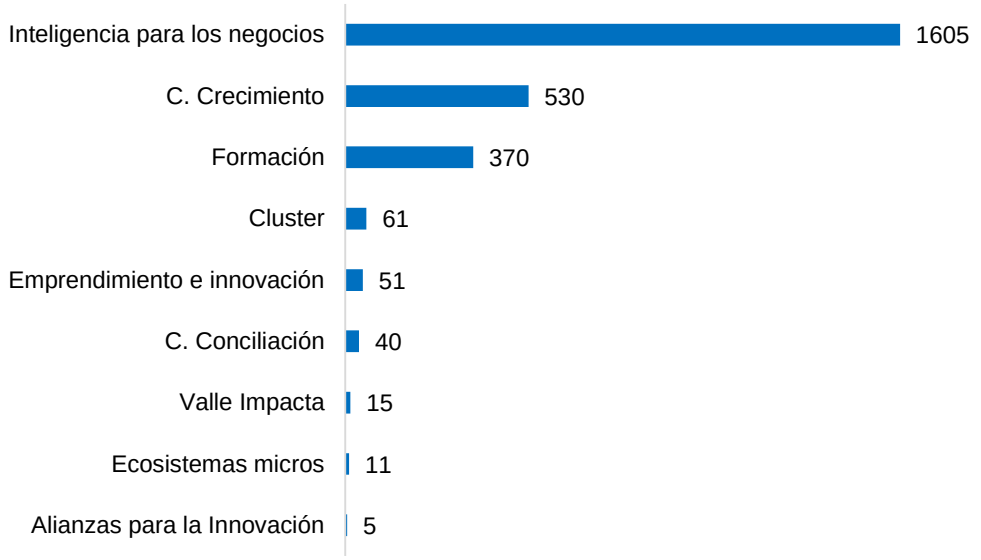
En el registro mercantil se dispone de la siguiente información para cada una de las empresas:

- Balance general: Activos, pasivos y patrimonio
- Estado de resultados: Ingresos operacionales, utilidad bruta y utilidad/pérdida neta
- Sector: variable categórica con 19 clases
- Año de matrícula: año de creación de la empresa, utilizado para determinar la edad de la empresa
- Número de establecimientos

4.1.2 Programas, productos y servicios de la Cámara de Comercio de Cali

Se tuvieron en cuenta nueve programas y servicios ofrecidos por la CCC, de los cuales solo 3 superaron las 300 empresas durante las campañas de 2017-2019: Inteligencia para los negocios, Centro de Crecimiento Empresarial y Formación (Figura 2).

Gráfico 2. Número de empresas por programa incluidas campañas de afiliación 2017-2019



Fuente: Cámara de Comercio de Cali – Elaboración propia

4.1.3 Exportaciones

Como fuente externa se consideró incluir el monto de exportaciones de cada una de las empresas registrado ante el DANE, sin embargo, se descartó debido a que solo 171 empresas en los tres años registraron exportaciones.

4.2 Selección de variables:

Tomando como referencia la información disponible y la exploración de datos realizada previamente (Anexo 1), se plantean 3 opciones:

- **Opción 1:** número de establecimientos, exportaciones, ingresos operacionales (ventas), patrimonio, total activos, utilidad bruta, utilidad/pérdida neta, crecimiento en ventas, crecimientos activos, crecimiento en utilidad bruta, crecimiento utilidad/pérdida neta y edad
- **Opción 2:** sector, número de establecimientos, patrimonio, total activos y edad

- **Opción 3:** sector, número de establecimientos, patrimonio, total activos, edad, participación en programas de crecimiento empresarial y participación en programas de formación

La primera opción corresponde a las variables numéricas consideradas relevantes teniendo en cuenta el análisis descriptivo y el conocimiento previo del comportamiento del negocio. La segunda opción solo contiene las variables numéricas que resultaron significativas adicionando la variable categórica de sector. La tercera opción incluye las variables numéricas significativas y adicionalmente considera las variables categóricas consideradas relevantes según el análisis descriptivo.

4.3 Algoritmos de aprendizaje

Se emplearán los algoritmos de aprendizaje: regresión logística, regresión logística paso a paso, *KNN*, árboles de decisión, árbol de decisión *Ada Boost*, *Random Forest*, *Robust Random Forest y C5.0*.

4.4 Protocolo de evaluación

El protocolo de evaluación a utilizar en este ejercicio es *Repeated K-fold Cross-Validation*, que realiza *K* particiones de los datos de entrenamiento e itera igual número de veces utilizando cada una de las particiones para evaluar el desempeño del clasificador. Este protocolo se caracteriza por permitir un balance entre sesgo y varianza, una característica ideal para alcanzar el objetivo del presente trabajo.

En este caso se realizan 10 particiones en el set de entrenamiento y se repite el proceso 10 veces.

4.5 Métricas de clasificación

Para evaluar los modelos existen diferentes tipos de métricas que determinan su asertividad. Es usual enfocarse en una medida completa como el *accuracy* que indica el porcentaje de observaciones que fue clasificada correctamente. Sin embargo, en este caso el objetivo es determinar cuáles empresas serán los clientes que tienen la mayor probabilidad de adquirir el servicio y enfocarse en ellos para realizar campañas con una mayor efectividad. Dicho de otra forma, se busca determinar el mayor porcentaje de verdaderos positivos.

Una vez entrenados y evaluados los modelos descritos anteriormente, se seleccionarán los mejores para ser utilizados en la prueba final. Para esto es necesario definir la métrica de selección ideal para cumplir los objetivos planteados en este trabajo.

Debido al desbalance de clases de la variable objetivo, y que se busca identificar de manera correcta las empresas que tomarán el producto, se empleará como métrica de selección el indicador F1 (Provost y Fawcett 2013), que se construye a partir de la sensibilidad y precisión.

La sensibilidad o *recall* indica el porcentaje de casos positivos correctamente identificados por el clasificador (Siegler, Jain, Raj y Stern 1997):

$$Sensibilidad = \frac{\text{verdaderos positivos}}{\text{verdaderos positivos} + \text{falsos negativos}}$$

Por otra parte, la precisión indica que porcentaje de las predicciones de casos positivos son acertadas (Siegler, Jain, Raj y Stern 1997):

$$Precisión = \frac{\text{verdaderos positivos}}{\text{verdaderos positivos} + \text{falsos positivos}}$$

Tanto la sensibilidad como la precisión son indicadores de la pertinencia de los clasificadores. No obstante, no se puede seleccionar el mejor clasificador empleando dos métricas; por lo tanto, se utiliza el indicador F1,

El indicador F1 es la media armónica de la sensibilidad y la precisión, que, a diferencia de la media aritmética, se acerca al valor más bajo entre los dos indicadores (Siegler, Jain, Raj y Stern 1997):

$$F1 = \frac{2 * \text{Precisión} * \text{Sensibilidad}}{\text{verdaderos positivos} + \text{falsos positivos}}$$

4.6 Selección de modelos

Se entrenaron las 3 opciones de variables, por medio de 8 técnicas para 2 conjuntos de datos (2017 y 2018), dando como total 24 modelos para cada conjunto de datos, de los cuales se seleccionaron los mejores para validar.

5. DISEÑO DE EXPERIMENTO DE VALIDACIÓN

En el proceso de selección de modelos se escogerán los dos que registren el valor más alto del indicador F1, uno entrenado con la información de la campaña realizada en 2017 y otro para 2018.

Estos modelos serán evaluados con la información de la campaña de 2019, y se analizarán los dos escenarios y su pertinencia basada en la efectividad de la predicción. Posteriormente se revisará su viabilidad de aplicación y despliegue al interior de la organización y sus posibles implicaciones.

6. RESULTADOS OBTENIDOS

En total, se entrenaron 48 modelos con la información disponible para los años 2017 y 2018. En términos generales, la tercera opción de variables presentó los mejores resultados ya que registró niveles de F1 superiores a las otras opciones, sin embargo, no presenta el valor más alto para el conjunto de datos de 2018.

Para 2017 el F1 más alto fue registrado por el Modelo 3 entrenado mediante *Robusted Random Forest* (0,511); mientras que en 2018 fue el Modelo 1 entrenado mediante árbol de decisión 5.0 (0,602) (Tabla 2). Con estos resultados, se procederá a evaluar la capacidad predictiva con los datos de 2019.

Tabla 2. **Indicador F1 de modelos entrenados 2017-2018**

	2017			2018		
Técnicas	Modelo 1	Modelo 2	Modelo 3	Modelo 1	Modelo 2	Modelo 3
Regresión logística	0,092	0,315	0,363	0,11	0,347	0,53
Regresión logística Stepwise	0,084	0,271	0,315	0,124	0,347	0,538
Árbol de decisión						
Random Forest	0,402	0,11	0,233	0,488	0,23	0,51
Random Forest robusto	0,467	0,431	**0,511**	0,509	0,455	0,533
C 5.0	0,222	0,381	0,372	**0,602**	0,501	0,52
Ada Boost	0,366	0,313	0,41	0,472	0,377	0,52
KNN	0,337	0,321	0,321	0,494	0,429	0,511

Fuente: Elaboración propia

6.1 Escenario 1

En el primer escenario (mejor modelo de 2017), para 2019 se predice que 768 de las 2.362 empresas de la campaña resultarían afiliadas (Tabla 3). Si se hubiese decidido seguir los resultados de las predicciones, la campaña se habría enfocado en solo el 32,5% de las empresas, con una efectividad del 52,6%.

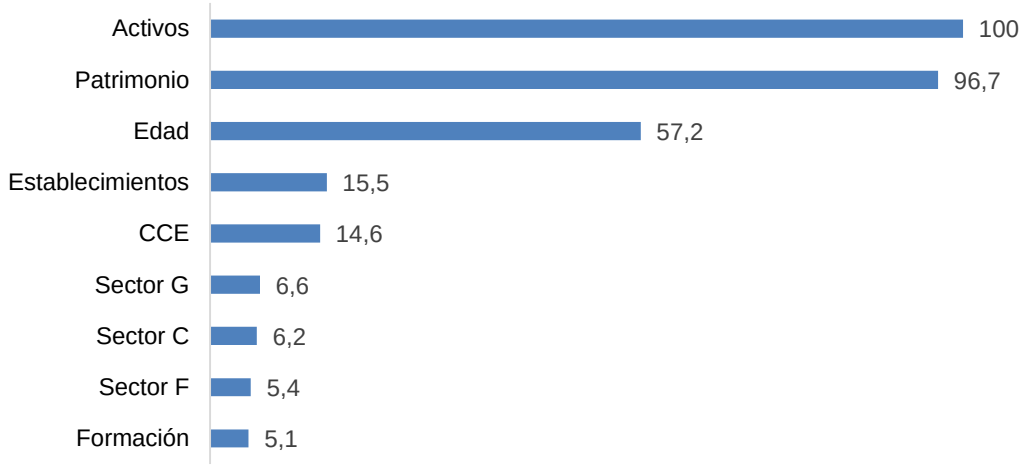
Tabla 3. Matriz de confusión escenario 1

		Observado	
		Afiliado	No Afiliado
Predicción	Afiliado	404	364
	No afiliado	377	1.217

Fuente: Elaboración propia

Aunque la técnica empleada en este caso no permite realizar inferencia con las variables empleadas, si es posible determinar cuáles son los atributos más importantes durante el proceso de entrenamiento. Para este escenario, las variables más importantes son: los activos, el patrimonio y la edad de la empresa (Gráfico 4).

Gráfico 3. Importancia de variables en el escenario 1



Fuente: Elaboración propia

6.2 Escenario 2

En el segundo escenario (mejor modelo 2018), se espera que 1.494 empresas tomarían la afiliación (Tabla 4). Esto corresponde al 63,3% del total de empresas. Si se empleara este modelo para el diseño de la campaña, tendría un éxito del 40,6%.

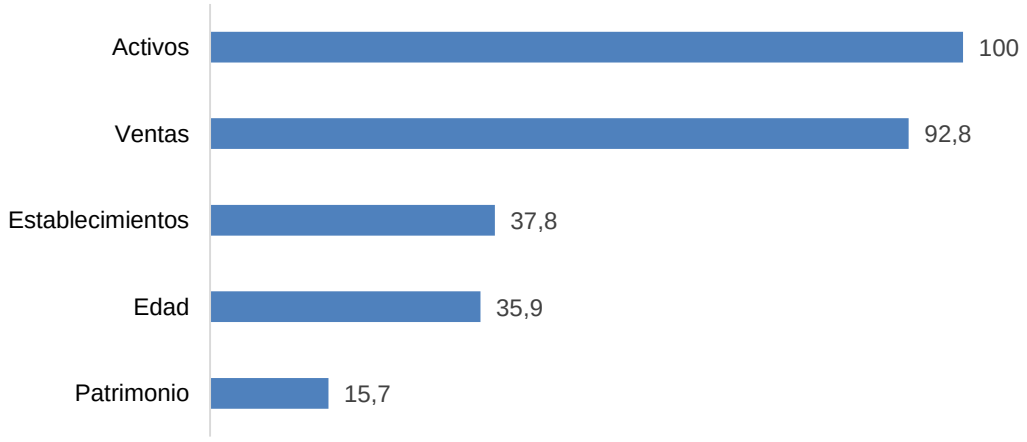
Tabla 4. Matriz de confusión escenario 2

		Observado	
		Afiliado	No Afiliado
Predicción	Afiliado	607	887
	No afiliado	174	694

Fuente: Elaboración propia

En este entrenamiento tampoco es posible realizar inferencia y solo 5 variables resultaron relevantes: Activos, ventas, establecimientos, edad y patrimonio

Gráfico 4. Importancia de variables en el escenario 2



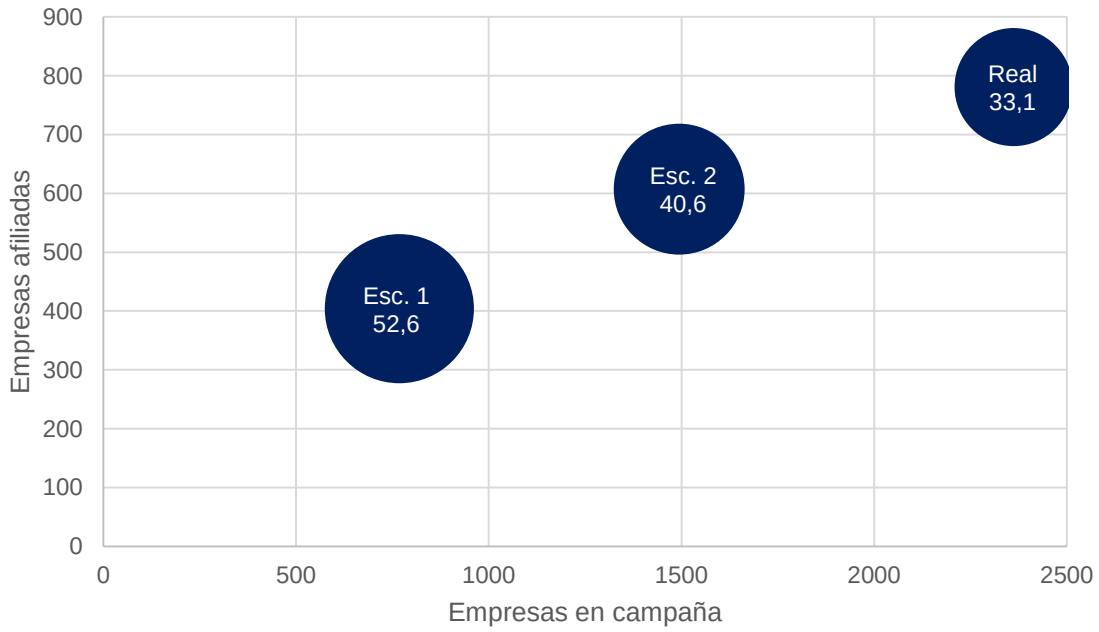
Fuente: Elaboración propia

6.3 Comparación de escenarios

Ambos escenarios presentan resultados validos en el experimento, ya que plantean modelos que conservan niveles de precisión similares a los registrados en el entrenamiento.

En ambos casos los modelos proponen una focalización en un grupo de empresas con una mayor probabilidad de adquirir la afiliación. En el primer escenario el modelo predice un número más reducido de afiliados con una tasa de éxito mucho más alta. En el segundo escenario el modelo reduce la campaña.

Gráfico 5. Comparación de escenarios



Fuente: Elaboración propia

Bajo el supuesto de que cada una de las empresas incluida en las campañas requiere la misma inversión de recursos para ofrecerle el programa, en el primer escenario alcanzaría 51,7% de los resultados con 32,5% de los recursos, mientras

que en el segundo escenario se llegaría a 77,7% de los resultados con 63,3% de los recursos. En ambas situaciones se obtienen campañas más costo efectivas.

El despliegue de estos modelos debería realizarse con todas las empresas que cumplen los requisitos de ley para obtener los mejores resultados posibles.

7. CONCLUSIONES Y FUTURO TRABAJO

Se identificaron dos modelos que cumplen el objetivo de clasificar de manera acertada las empresas con mayor probabilidad de adquirir la afiliación de la Cámara de Comercio de Cali.

El primer modelo incluye variables numéricas y categóricas y se entrenó con datos de 2017 empleando el método de *Robust Random Forest,* se caracteriza por una buena capacidad predictiva. El segundo modelo solo incluyó variables financieras, se entrenó con datos de 2018 por el método de árboles de decisión C5.0.que se desempeña mejor cuando las observaciones no son las más comunes. Estas técnicas son similares a las identificadas por otros trabajos en el campo.

Los dos clasificadores seleccionados en el presente trabajo mostraron ser consistentes al probarlos con los datos de 2019. Además, podrían emplearse para generar campañas de afiliación más costo efectivas que las que se realizan actualmente.

Aunque no es posible realizar inferencia a partir de los modelos seleccionados, sabemos que ambos tienen entre sus variables relevantes el monto de activos, número de establecimientos y edad. En los trabajos aplicados a negocios *B2B* se han analizado variables cualitativas relacionadas con la forma de contacto con el cliente, e incluso con la percepción de los vendedores. Estas han demostrado ser relevantes en otros casos y como trabajo futuro podrían capturarse para incluirlas en el proceso de entrenamiento y mejorar el desempeño de las predicciones.

Como trabajo futuro, se pueden explorar otras técnicas cómo Triggering Kernel Learning y Hawks que registraron los mejores resultados en estudios similares y no fueron abordadas en este trabajo.

El despliegue requiere la planeación, monitoreo y mantenimiento del modelo seleccionado que permitirá realizar campañas de afiliados con mayor asertividad. para garantizar su replicabilidad en el futuro, se realizará una entrega con el informe final de resultados, el código empleado y la descripción de recolección de información. Es posible que la recolección de información que será utilizada durante el despliegue requerirá implementar un modelo de gobierno de datos o adaptarse a uno existente en la organización.

El despliegue de los modelos evaluados puede afectarse por la coyuntura del COVID-19 y las medidas gubernamentales. Esta situación genera un desequilibro que puede afectar considerablemente las predicciones de modelos entrenados por fuera de este fenómeno. Es posible considerar un despliegue en el futuro si se realiza la validación de los modelos con la información disponible en lo corrido de 2020 y los resultados son consistentes.

La metodología empleada en este trabajo puede ser replicada en otros servicios, en otras Cámaras de Comercio y en otras empresas que ofrezcan servicios B2B. Se debe tener en cuenta que los resultados pueden ser distintos, ya que la literatura indica que en este tipo de aplicaciones son muy particulares para cada caso y no suelen hacerse públicas.

BIBLIOGRAFÍA

Bohanec, M., Borštnar, M. K., & Robnik-Šikonja, M. (2016). Integration of machine learning insights into organizational learning: A case of B2B sales forecasting. In Blurring the boundaries through digital innovation (pp. 71-85). Springer, Cham.

Consejo Nacional de Política Económica y Social. (2019). Documento Conpes 3956.

Cramer, J. S. (2002). The origins of logistic regression.

Fix, E., Hodges,J.L. (1951). Discriminatory analysis: nonparametric discrimination, consistency properties. USAF School of Aviation Medicine.

Freund, Y., & Schapire, R. E. (1995). A decision-theoretic generalization of on-line learning and an application to boosting. In *European conference on computational learning theory* (pp. 23-37). Springer, Berlin, Heidelberg.

IBM. (2017). IBM SPSS Modeler CRISP-DM Guide.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112, pp. 3-7). New York: springer

Kam, H. T. (1995). Random decision forest. In Proceedings of the 3rd International Conference on Document Analysis and Recognition (Vol. 1416, p. 278282). Montreal, Canada, August.

Monat, J. P. (2011). Industrial sales lead conversion modeling. Marketing Intelligence & Planning.

Mortensen, S., Christison, M., Li, B., Zhu, A., & Venkatesan, R. (2019). Predicting and Defining B2B Sales Success with Machine Learning. In 2019 Systems and Information Engineering Design Symposium (SIEDS) (pp. 1-5). IEEE.

Ngai, E. W., Xiu, L., & Chau, D. C. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. Expert systems with applications, 36(2), 2592-2602.

Piatetsky-Shapiro, G. (2014). CRISP-DM, still the top methodology for analytics, data mining, or data science projects. KDnuggetsTM Data Mining, Analytics, Big Data and Data Science.

Presidencia. (1996). Decreto 650.

Presidencia de la República de Colombia. (1971). Decreto 410 de 1971.

Provost, F., & Fawcett, T. (2013). Data Science for Business: What you need to know about data mining and data-analytic thinking. " O'Reilly Media, Inc.".

Quinlan, J. R. (2014). C4. 5: programs for machine learning. Elsevier.

Rulequest. (2017). Rulequest. Obtenido de https://www.rulequest.com/see5-comparison.html

Senado de la República de Colombia. (1931). Ley 28 de 1931.

Senado de la República de Colombia. (2010). Ley 1429 de 210.

Senado de la República de Colombia. (2014). Ley 1727 de 2014.

Senado de la República de Colombia. (2016). Ley 1780 de 2016.

Siegler, M. A., Jain, U., Raj, B., & Stern, R. M. (1997). Automatic segmentation, classification and clustering of broadcast news audio. In Proc. DARPA speech recognition workshop (Vol. 1997).

Vapnik, V. (1995). *The nature of statistical learning theory*. Springer science & business media.

Yan, J., Zhang, C., Zha, H., Gong, M., Sun, C., Huang, J., ... & Yang, X. (2015). On machine learning towards predictive sales pipeline analytics. In Twenty-ninth AAAI conference on artificial intelligence.