Modelo predictivo para la selección de técnica de medición de la opinión pública

Predictive model for the selection of a public opinion measurement technique

Luis Herrero-Corona

De las Heras Demotecnia, México

Recibido: 9 de julio de 2021.
Aprobado: 17 de noviembre de 2021.

Resumen

El objetivo de este artículo es construir un modelo predictivo, a partir de información de estudios realizados en agencias de investigación de mercados y de opinión pública, que indique el método de recolección de datos que se recomienda utilizar, que puede ser encuestas personales, telefónicas o en línea, de acuerdo con los parámetros de cada estudio.

El tipo de modelo predictivo es de clasificación, y se construyen y se analizan modelos a partir de las técnicas de minería de datos de árboles de decisiones, análisis discriminante, análisis de K vecinos más cercanos y análisis de redes neuronales. Adicionalmente, se realiza una segmentación de contactos en clústeres para complementar y enriquecer el conocimiento que aportan las técnicas clasificatorias.

Se concluye que los modelos generados tanto por árboles de decisiones como por redes neuronales son los que mejor predicen la técnica de medición de opinión pública a utilizar.

Palabras clave: *encuestas, opinión pública, modelo predictivo.*
Clasificación JEL: *C38, C45, C83, M30, M31.*

Abstract

The objective of this article is to build a predictive model, based on information from projects carried out in market research and public opinion survey companies, choosing the recommended data collection method, which can be face to face interviews, telephone or online surveys, according to the requirements of each case.

The predictive model type is one of classification, and several are built and analyzed using decision tree data mining techniques, discriminant analysis, K nearest neighbor analysis, and neural network analysis. Additionally, a segmentation of contacts in clusters is carried out to complement and enrich the knowledge provided by the classification techniques.

It is concluded that the models generated by both decision trees and neural networks are the ones that best predict the public opinion measurement technique to be used.

Keywords: *surveys; public opinion; predictive model.*
JEL Classification: *C38, C45, C83, M30, M31.*

1. Introducción

De acuerdo con la European Society for Opinion and Market Research (ESOMAR), la llamada industria de los *insights* —que agrupa los estudios tanto de opinión pública como de investigación de mercados— creció de forma global en 2019 en 5.9 % en términos absolutos (ESOMAR, 2020, 10). La Asociación Mexicana de Agencias de Inteligencia de Mercado y Opinión Pública (AMAI) y el Instituto Tecnológico Autónomo de México (ITAM) estimaron que el valor del mercado nacional de encuestas y estudios de mercado para 2019 en México fue de 7664 millones de pesos (AMAI e ITAM, 2020, 2). La medición de la opinión pública es una fuente de empleo para casi 11 000 personas en el país. Se realizaron 7 millones de entrevistas con informantes, de las cuales 43 % fueron cara a cara (AMAI e ITAM, 2020, 2), 37 % por internet y 20 % vía telefónica (AMAI e ITAM, 2020, 7).

Estas tres metodologías son alternativas para clientes y agencias especializadas en medir la opinión pública y/o hacer investigación de mercados. La decisión de utilizar una en particular conlleva ventajas y limitaciones que pueden tener impacto en los resultados que se obtengan (Zhang *et al*., 2017). Por otro lado, el uso de la minería de datos para abordar soluciones de negocio es una práctica que otorga beneficios en diversos aspectos a las empresas (Escobar Terán *et al*., 2016; Jeffery, 2010). En este estudio, el problema de negocio que se busca resolver es la selección de la técnica de entrevista adecuada para las especificaciones de un estudio de opinión pública o investigación de mercados. Esta decisión puede apoyarse en un modelo predictivo a partir de datos históricos disponibles (Stupakevich *et al*., 2019).

El presente artículo sigue la metodología de construcción de modelos predictivos denominada CRISP-DM (Chapman *et al*., 2007). Se parte de una primera etapa de comprensión del negocio, donde se especifica lo que se pretende lograr, los objetivos y criterios de éxito del trabajo. A continuación, se procede a la comprensión de los datos y se describen las características, su fuente y una primera exploración general de los mismos. La etapa de preparación de datos incluye la selección, limpieza, transformación y otras tareas similares para conseguir una base de conocimiento lista para trabajar con modelos predictivos. En el apartado del modelado, se seleccionan la técnica o técnicas apropiadas para los objetivos y se procede a la construcción de modelos, para posteriormente continuar con su evaluación.

2. Marco teórico

La minería de datos es el proceso que se encarga de descubrir patrones en conjuntos de datos y tiene alguno de los siguientes objetivos: describir, estimar, predecir, clasificar, agrupar o asociar (Larose y Larose, 2015). El objetivo de esta investigación es construir un modelo con base en información de proyectos realizados por empresas especializadas en estudios de mercado y de opinión pública, que permita seleccionar la técnica de medición recomendada —encuestas personales, telefónicas o en línea— a partir de los parámetros que cada estudio requiere.

El concepto de encuesta se refiere al «método que se utiliza para conocer el estado de opinión sobre un determinado tema y que consiste en realizar una serie de preguntas a una muestra representativa de la población, de cuyas respuestas se infieren los valores de la población en su conjunto» (De las Heras, 1999, 184). Como parte de las técnicas para medir la opinión pública, las encuestas se pueden clasificar —según el método de recolección de datos— en encuestas cara a cara, telefónicas o en línea.

Las *encuestas cara a cara* se refieren a la recolección de datos de una encuesta en vivienda o en puntos de afluencia mediante encuestadores que aplican un cuestionario a informantes, cara a cara.

Si la entrevista se realiza por teléfono, se conoce como *encuesta telefónica* y tiene la ventaja de que pueden obtenerse, con relativa facilidad, muestras probabilísticas representativas de la población con acceso a servicios de telefonía fija o móvil (Vehovar *et al*., 2012), siempre y cuando más de 80 % del territorio donde se realicen cuente con este servicio (Steeh, 2008, 221). Las encuestas telefónicas producen resultados que pueden ser comparables a aquellos que se obtienen mediante encuestas personales, pero a menor costo (Lavrakas, 1987).

Las *digitales* o *en línea* son aquellas que corresponden al tipo de encuestas autoadministradas, pero que usan la web en la aplicación del cuestionario (Callegaro *et al*., 2015) y tienen el potencial de describir una población en particular de manera representativa o ser solamente una forma de entretenimiento (Couper, 2000).

La tecnología de computación permite ventajas significativas en comparación con el levantamiento de cuestionarios en papel. Las respuestas se almacenan en bases de datos y se pueden procesar de forma inmediata, reduciendo tiempo, costos y errores como los que se producen cuando se capturan desde el papel (Vehovar

y Lozar Manfreda, 2008). Schober (2018) escribe en su artículo «The future of face-to-face interviewing» que las encuestas cara a cara han sido consideradas el *estándar de oro* en cuanto a las tasas de participación, calidad de datos y satisfacción por los encuestados, pero en el futuro esto no seguirá siendo necesariamente el caso para algunos participantes:

> Los resultados de diversos estudios sugieren que, por lo menos para algunos de los encuestados, las modalidades asincrónicas de entrevistar que reducen la presencia social del entrevistador y permiten que los encuestados participen mientras están en su teléfono móvil o haciendo múltiples tareas [...] bien podrían llevar a la obtención de información de mayor calidad y una mayor satisfacción del encuestado (Schober, 2018, 290).

En los últimos años, han surgido nuevas técnicas de recolección de datos ante la disminución de las tasas de respuesta —tanto de encuestas cara a cara como telefónicas— y debido al incremento en los costos de estas metodologías (Couper, 2017). Sin embargo, existe preocupación por la representatividad de las encuestas en línea, tema que se aborda en estudios que buscan comparar diferentes tipos de entrevistas y diferentes muestras. Grewenig *et al*. (2018) analizan encuestas en línea de entrevistas autoaplicadas, encuestas cara a cara y entrevistas en línea cara a cara. Encuentran diferencias en patrones de respuesta de individuos que contestan en línea en comparación con aquellas que lo hacen en persona, por lo que los investigadores controlan las características demográficas y contextos personales de los participantes y, de esta forma, al asignar los pesos adecuados, obtienen la representatividad deseada. El modo de entrevista afecta los resultados de las encuestas, incluso en el caso de aquellas con diseño de muestras representativas (Zhang *et al*., 2017).

Cuando se permite que el participante elija el modo de entrevista que prefiere, la familiaridad con la tecnología web, la edad y la educación, por sí solas, no explican la elección, pero la afinidad hacia la tecnología está relacionada con elegir el modo en línea como un efecto independiente (Pforr y Dannwolf, 2017).

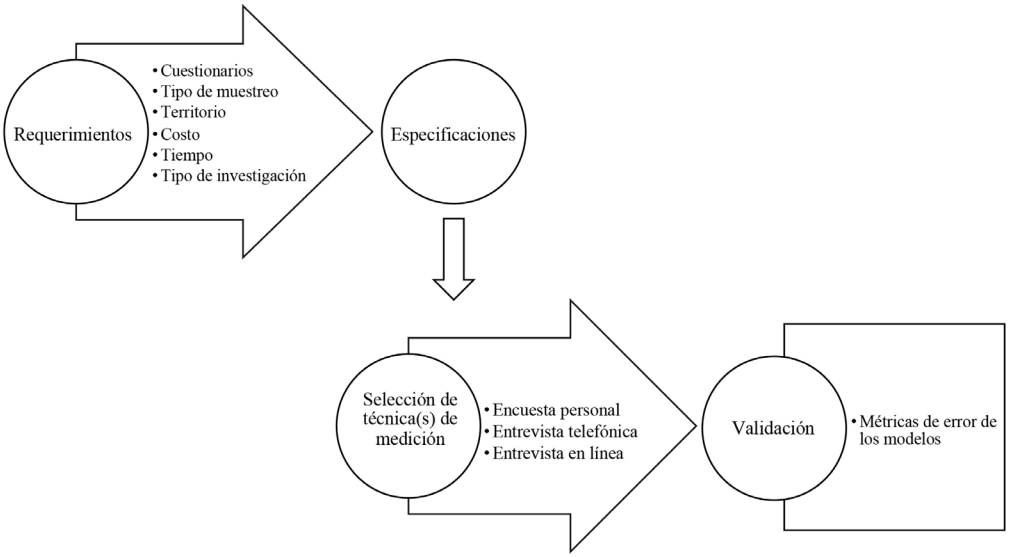
La representatividad en las encuestas es mayor en aquellas que se basan en muestras probabilísticas, como también lo es en encuestas de modos mixtos de recolección de datos que en las de modo único, y tienen mayor representatividad las encuestas distintas a las hechas en línea, de acuerdo con el artículo «Is there an association between survey characteristics and representativeness? A meta-analysis» (Cornesse y Bosnjak, 2018). En este sentido, se han realizado comparaciones del modo de entrevista entre encuestas hechas en computadora de escritorio, en

teléfonos móviles y telefónicas, encontrando que son estas últimas las de mayor precisión, aunque también son las de mayor costo entre las alternativas mencionadas (Lee *et al*., 2019).

Estos estudios muestran que el modo de entrevista puede afectar la precisión de los resultados de una encuesta, por lo que es importante elegirlo adecuadamente. Sin embargo, no se encuentra una metodología que indique al investigador cuándo es apropiado elegir alguna en particular. Para construir el modelo, se pueden redactar especificaciones que servirán como entradas, como se observa en la gráfica 1, con la finalidad de seleccionar la técnica de recolección de datos adecuada al estudio (ver gráfica 1). Para la etapa de minería de datos que permite construir el modelo, se busca como resultado exitoso que la variable que se predice tome alguna de las siguientes opciones:

- Encuesta personal
- Entrevista telefónica
- Entrevista en línea

Gráfica 1. Modelo predictivo para la selección de técnica de medición de la opinión pública



Fuente: elaboración propia con base en experiencia profesional y datos de proyectos realizados en empresas especializadas.

3. Metodología

Para llevar a cabo la recolección de datos iniciales, se contactó a empresas especializadas en estudios de mercado y de opinión pública y se solicitó la información de proyectos, registrando de cada uno tanto las especificaciones como la selección de la técnica de medición empleada.

Se obtuvieron 98 casos en total que corresponden a estudios de opinión pública llevados a cabo en el período comprendido del 27 de junio de 2018 al 7 de febrero de 2020. La información recopilada incluye datos tanto numéricos como de cadenas de texto, con un total de 10 campos por caso. Se estima que con estos se tiene información suficiente y relevante para continuar con la minería de datos.

Los datos incluyen el campo que será la variable para predecir, en este ejemplo, el modo de entrevista. Esta variable puede tomar los siguientes valores: presencial, telefónica o en línea. El dato es de tipo texto, por lo que fue codificado en la siguiente etapa de preparación de datos.

Los campos de la base de conocimiento obtenida son los que se muestran en la tabla 1, a continuación (ver tabla 1).

Tabla 1. Campos de la base de conocimiento

Campo	Dato	Descripción	Tipo
1	Consecutivo	Consecutivo de casos	Numérico
2	Fecha	Fecha del estudio	Fecha
3	Tipo de investigación	Descriptiva o exploratoria	Texto
4	Modo de entrevista	Personal, telefónica o en línea	Texto
5	Proyecto	Nombre del estudio	Texto
6	Cuestionarios	Número de cuestionarios aplicados	Numérico
7	Días	Duración del levantamiento de la información	Numérico
8	Territorio	Territorio de aplicación a nivel nacional, estatal o municipal	Texto
9	Tipo de muestreo	Probabilístico o no	Texto
10	Costo	Costo del estudio en pesos mexicanos	Numérico

Fuente: elaboración propia.

Al verificar la calidad de los datos, se concluye que no hay omitidos ni contienen errores, ya que la información fue recopilada de primera mano y se fue construyendo la base de conocimiento con aquellos casos en los que se contó la información. Tampoco se tienen valores omitidos.

Los datos que se seleccionaron para el análisis son los descritos anteriormente. La única exclusión que se realiza es el nombre del estudio, ya que no contiene información adicional relevante al objetivo que no esté ya considerada en los otros campos, por lo que la base final contempla los nueve campos de información restantes.

Se revisó y corrigió cada uno de los registros a partir de la congruencia que se buscaba para cada tipo de campo. Por ejemplo, se unificó el formato de fecha de cada proyecto, para que contuviera adecuadamente la información de día, mes y año en el orden apropiado. Se renombraron los campos a una sola palabra en minúsculas, para facilitar su importación posterior en programas estadísticos. De igual forma, se verificó que todos los registros estuvieran completos y que los criterios usados en los campos de texto fueran uniformes, como es el uso de mayúsculas y minúsculas, abreviaturas, acentos, etcétera.

Se realizó la codificación de las variables que se encontraban como texto para poder trabajar con ellas como variables nominales en los modelos a construir, y que son las siguientes:

- Territorio
  - Codificado en 3 valores diferentes: nacional, estatal o municipal
- Tipo de muestreo
  - Codificado en probabilístico o no probabilístico
- Tipo de investigación
  - Exploratoria o descriptiva
- Modo de entrevista (variable a predecir)
  - Personal, telefónica o en línea

La integración de datos se realizó en la construcción de la base original, reuniendo fuentes que permitieron hacer la base a partir de información comercial (proyecto, cuestionarios), contable (costo) y operativa (días de levantamiento, modo de entrevista). Una vez construida la base, no se integraron fuentes de información adicionales.

Se ordenó la base para que se mostraran los campos de la manera que se presenta a continuación (ver tabla 2) y facilitar así su comprensión y análisis, comenzando por una primera columna con la numeración consecutiva y continuando con una forma lógica que obedece a la fecha del estudio, número de cuestionarios, el territorio, tipo de muestreo, costo, días, tipo de investigación y, finalmente, la variable predictiva, modo de entrevista. El orden final de la base se muestra en la tabla 2.

Tabla 2. Campos en la base final de conocimiento

Campo	Campo	Descripción
1	Consecutivo	Consecutivo de casos
2	Fecha	Fecha del estudio
3	Cuestionarios	Número de cuestionarios aplicados
4	Territorio	Territorio de aplicación a nivel nacional, estatal o municipal (codificado)
5	Tipo de muestreo	Probabilístico o no (codificado)
6	Costo	Costo del estudio en pesos
7	Días	Duración del levantamiento de información
8	Tipo de investigación	Descriptiva o exploratoria (codificado)
9	Modo de entrevista	Personal, telefónica o en línea (codificado)

Fuente: elaboración propia.

Los datos originales se construyeron en una hoja de cálculo, por lo que adicionalmente se preparó un archivo en SPSS (IBM Corp., 2020) para poder trabajar con los modelos predictivos de ese programa estadístico. Las características de la base final, así como el catálogo de codificación, se agregan al final del artículo como anexo (ver tablas A1 y A2).

De acuerdo con los objetivos de la minería de datos, se busca construir un modelo predictivo que permita seleccionar la técnica de medición de la opinión pública adecuada a partir de los requerimientos de cada estudio de mercado, y se sabe que la variable predictiva es categórica, ya que predice la técnica de recolección de datos: encuestas personales, telefónicas o en línea. Por lo anterior, se construye un *modelo de clasificación* (Gera y Goel, 2015; Han *et al*., 2011; Larose y Larose, 2015) utilizando

el conjunto de datos que se recopilaron previamente para predecir la técnica de recolección, que es la clase o categoría que el modelo asignará a nuevos estudios futuros (Joseph *et al*., 2016; Stupakevich *et al*., 2019).

En consecuencia, se prueban las técnicas de clasificación a partir de árboles de decisiones, análisis discriminante, K vecinos más cercanos y redes neuronales (Berry y Linoff, 2004; Larose y Larose, 2015). Adicionalmente, se utiliza la herramienta de Marketing Directo del SPSS (IBM Corp., 2020) para segmentación de contactos en clústeres, porque puede proveer otro enfoque de análisis que aporte un modelo de clasificación válido.

Para cada técnica que se utiliza, se examina el error que se obtiene al comparar los datos observados con los que el modelo predice. Además, se hace una separación del conjunto de datos en aquellos de entrenamiento y de prueba, para las técnicas que permite el programa estadístico, y se estiman las tasas de error en cada uno (Berry y Linoff, 2004; Joseph *et al*., 2016).

4. Resultados

Los modelos fueron construidos en el programa SPSS (IBM Corp., 2020). Una de las fortalezas que tiene el hacerlo de esta forma es que los modelos que se obtienen pueden ser exportados en formato XML para aplicarse posteriormente a un conjunto de datos con una estructura como la de la base en la que se originaron, y obtener el valor de la variable predictiva para nuevos casos.

Técnica: árboles de decisiones

Se utilizó la técnica de árboles de decisiones del SPSS (IBM Corp., 2020), eligiendo como variable dependiente el modo de entrevista y como independientes el número de cuestionarios, el tipo de territorio, el tipo de muestreo, el costo, los días de levantamiento y el tipo de investigación. Se utilizó el método CHAID y, posteriormente, se recurrió a manera de comparación a los métodos CRT y QUEST del árbol de decisiones. Se hizo una corrida con el 100 % de los datos, así como otras, solicitando, para validación, dividir la base en 70 % de entrenamiento y 30 % de prueba, o también 60 % - 40 % (Han *et al*., 2011), manteniendo el resto de los parámetros en sus valores por defecto. Por el tamaño de la base de datos (98 casos), se consideró como nodo paterno al menos 10 casos y, como nodo hijo, 5 casos.

Al cambiar tanto el método de crecimiento del árbol como la proporción de entrenamiento y prueba de los datos, se obtienen diferentes porcentajes de error. La tabla 3 compara estas combinaciones (ver tabla 3).

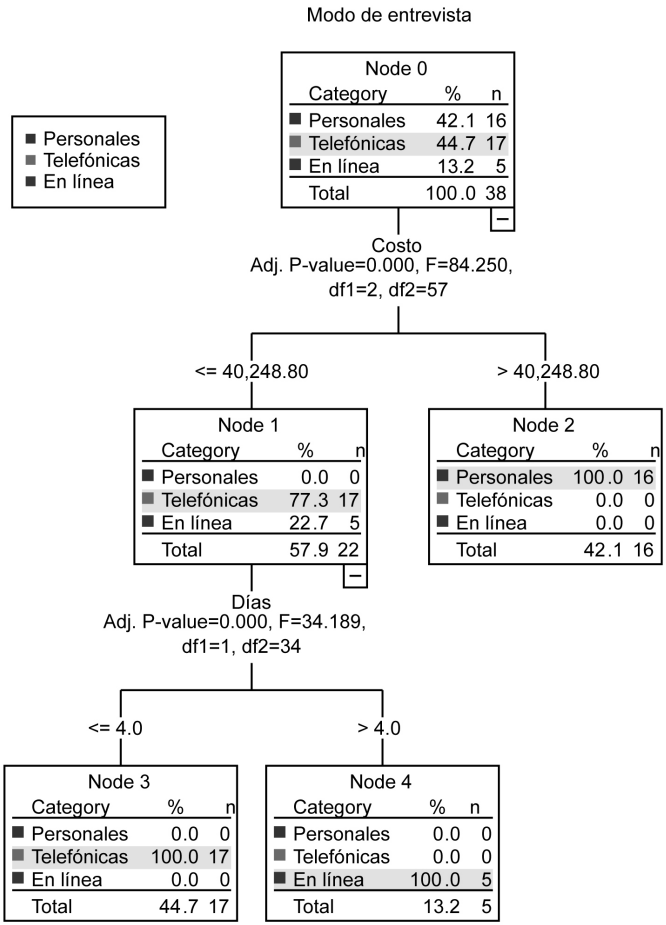
Tabla 3. Comparación entre árboles de decisiones

Validación	Método de crecimiento	Porcentaje de error entrenamiento	Porcentaje de error prueba
Sin división de base de datos	CHAID	5.1 %	
	CRT	4.1 %	
	QUEST	6.1 %	
60-40 %	CHAID	10 %	15.8 %
	CRT	1.9 %	9.1 %
	QUEST	8.3 %	0.0 %
70-30 %	CHAID	7.9 %	4.5 %
	CRT	1.5 %	10 %
	QUEST	3 %	9.7 %

Fuente: elaboración propia.

En la gráfica 2 se muestra el árbol de decisiones generado a partir del método QUEST, con división de la base de datos de la siguiente forma: 60 % en datos de entrenamiento y 40 % de prueba, que es el que tiene menor porcentaje de error en la base de prueba (ver gráfica 2).

Gráfica 2. Árbol de decisiones



Fuente: elaboración propia.

En el árbol de la gráfica 2 se muestra que, para predecir si conviene realizar un estudio de mercado mediante encuestas personales, telefónicas o en línea, debe tomarse en cuenta, primero, el costo del estudio de mercado a realizar. Si se tienen más de 40 248 pesos, ha de realizarse de forma personal. Si la cantidad que se tiene es menor, entonces ha de revisarse el número de días disponibles para el levantamiento de información. Si son cuatro días o menos, debe realizarse de forma telefónica. Para más de cuatro días, la encuesta tiene que hacerse en línea.

Técnica: análisis discriminante

Se seleccionó como variable de agrupación el modo de entrevista, en el rango de valores de 1 a 3, ya que son los tres valores que puede tomar esta variable predictiva. El resto de las variables se contemplaron como independientes. Se eligió el método de inclusión por pasos, para comparar en cada uno de ellos si agregar cada variable independiente mejora el modelo o no (Larose y Larose, 2015; Meyers *et al*., 2013). Los demás parámetros se dejaron en sus valores por defecto, solicitando únicamente la gráfica de grupos combinados (ver gráfica 3).

El modelo resultante consta de dos funciones en las que se encuentran asociadas las variables independientes. De acuerdo con la lógica de incorporación de variables de este tipo de análisis, los coeficientes de correlación de cada variable con las funciones son los que se muestran en la tabla 4, eliminando aquellas que no aportan mayor valor predictivo (ver tabla 4).

Tabla 4. Coeficientes de correlación del análisis discriminante

Matriz de estructuras		
	Función	
	1	**2**
Tipo de muestreo	.999*	-.046
Tipo de investigación[b]	.264*	.148
Territorio[b]	-.031*	.005
Días	.053	.999*
Cuestionarios[b]	.008	.254*
Costo[b]	.156	.181*

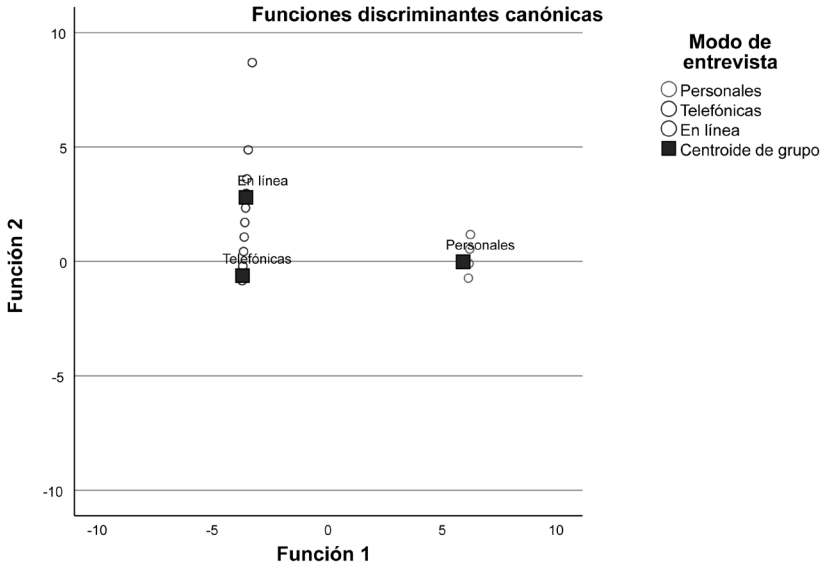
Fuente: elaboración propia.

Notas: variables ordenadas por el tamaño absoluto de la correlación dentro de la función.

*. La mayor correlación absoluta entre cada variable y cualquier función discriminante.

[b]. Esta variable no se utiliza en el análisis.

Gráfica 3. Análisis discriminante



Fuente: elaboración propia.

El modelo logra un error de 6.1 % global, como se detalla en la tabla 5, con el 93.9 % de los casos originales clasificados correctamente (ver tabla 5).

Tabla 5. Resultados de clasificación del análisis discriminante

Resultados de clasificación[a]						
		Modo de entrevista	Pertenencia pronosticada a grupos			Total
			Personales	Telefónicas	En línea	
Original	Recuento	Personales	37	1	0	38
		Telefónicas	0	47	2	49
		En línea	0	3	8	11
	%	Personales	97.4	2.6	0	100
		Telefónicas	0	95.9	4.1	100
		En línea	0	27.3	72.7	100

Notas:
[a]. 93.9 % de casos agrupados originales clasificados correctamente.
Fuente: elaboración propia.

Técnica: análisis de K vecinos más cercanos

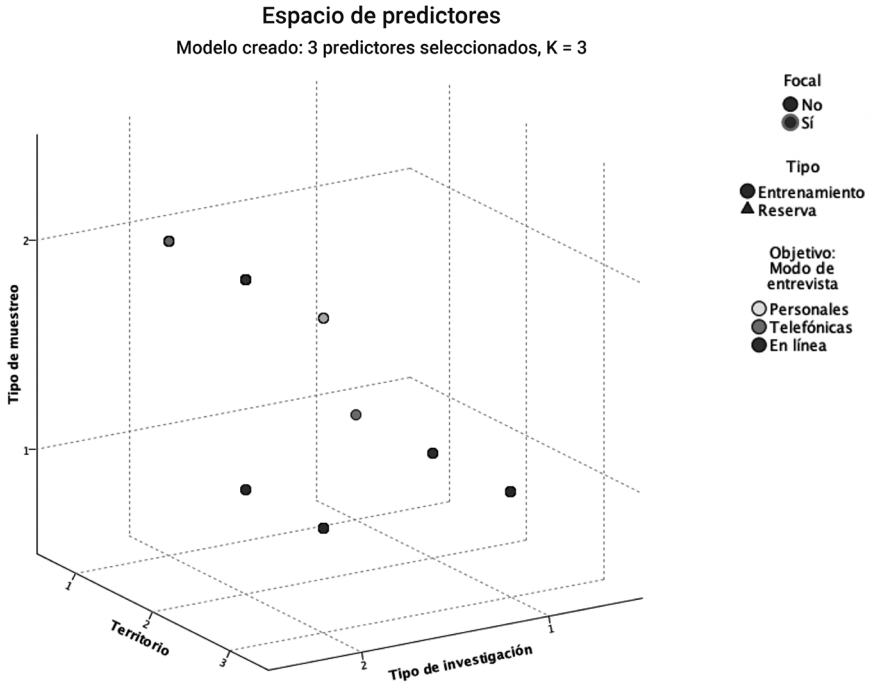
Se considera que el objetivo es la variable de modo de entrevista, sujeto a características en las que se introduce el resto de las variables. Se especifica que el número de vecinos más cercano K se determine de forma automática. En particiones, se utiliza el criterio de 70-30 % para entrenar y probar el modelo, y se mantienen el resto de las opciones en sus valores predeterminados (MacLennan *et al*., 2009).

El modelo resultante contiene un total de seis predictores y logra un nivel de error de 8.7 % para los datos de entrenamiento y de 3.4 % para los de reserva, como se muestra en la tabla 6 y en la gráfica 4 (ver tabla 6 y gráfica 4):

Tabla 6. Clasificación de K vecinos más cercanos

Tabla de clasificación					
Partición	Observado	Previsto			
		Personales	Telefónicas	En línea	Porcentaje correcto
Entrenamiento	Personales	28	1	0	96.6 %
	Telefónicas	0	30	1	96.8 %
	En línea	0	4	5	55.6 %
	Porcentaje global	40.6 %	50.7 %	8.7 %	91.3 %
Reserva	Personales	9	0	0	100.0 %
	Telefónicas	0	17	1	94.4 %
	En línea	0	0	2	100.0 %
	Porcentaje global	31.0 %	58.6 %	10.3 %	96.6 %

Gráfica 4. Análisis de K vecinos más cercanos



Fuente: elaboración propia.

Técnica: análisis de redes neuronales

El tipo de redes neuronales que se construyó fue generado a partir de la técnica de perceptrón multicapa. La variable dependiente es el modo de entrevista. Las demás variables se introducen como covariables. Al igual que con los otros modelos, se hacen pruebas para dividir la base en particiones de entrenamiento y prueba, y se comparan los porcentajes de error en cada caso. También se hace el ejercicio variando el número de capas ocultas en la arquitectura de la red para ver si se logra una mejor precisión en el modelo (Gorunescu, 2011).

La tabla 7 compara los porcentajes de error del modelo al realizar las combinaciones de parámetros descritas (ver tabla 7):

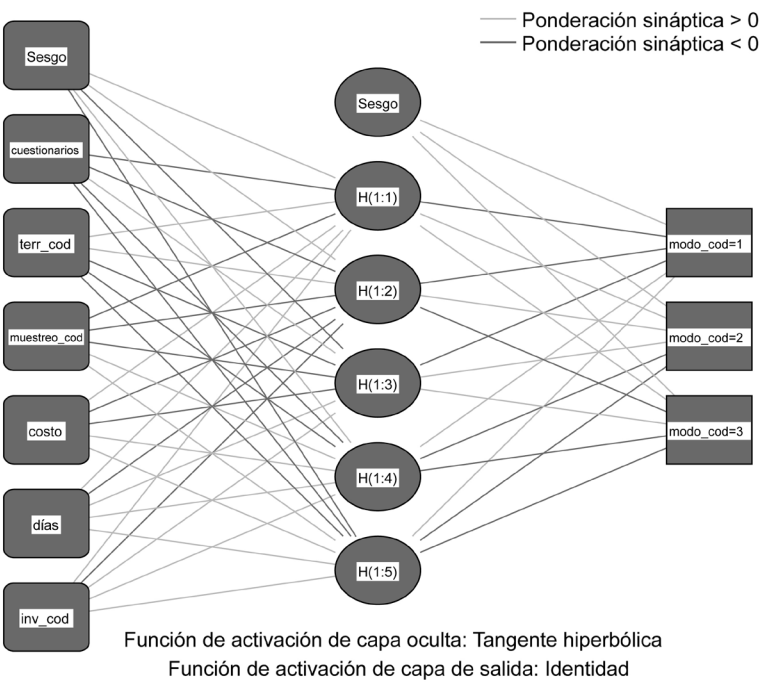
Tabla 7. Comparación de modelos de redes neuronales

Partición de datos	Arquitectura	Porcentaje de error entrenamiento	Porcentaje de error prueba
60-40 %	1 capa oculta	9.4 %	0.0 %
	2 capas ocultas	8.9 %	4.8 %
70-30 %	1 capa oculta	3.1 %	8.8 %
	2 capas ocultas	5.8 %	3.4 %

Fuente: elaboración propia.

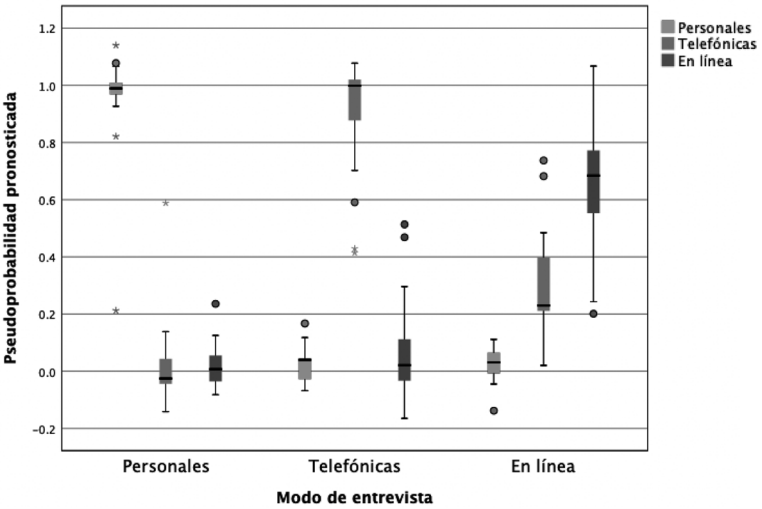
En la gráfica 5 se muestra el ejemplo de red neuronal con partición de datos 60-40 % y arquitectura de una capa oculta, que arrojó un 0.0 % de porcentaje de error de prueba en la tabla 7 (ver gráfica 5 y tabla 7).

Gráfica 5. Diagrama de red neuronal



Fuente: elaboración propia.

Gráfica 6. Pseudoprobabilidad pronosticada



Fuente: elaboración propia.

En la gráfica 6, se aprecia que, para este modelo, la predicción del modo de entrevista personal o telefónica no suele presentar tanta dificultad: ambas tienen valores predictivos cercanos al 1.0 en el modo que les corresponde. Sin embargo, para el modo de entrevista en línea, el modelo predice erróneamente un porcentaje como entrevistas telefónicas e incluso tampoco logra superar el 80 % de predicciones.

Técnica: segmentación de contactos en clústeres

En la versión utilizada del programa SPSS (IBM Corp., 2020), existe una herramienta que incluye un conjunto de técnicas enfocadas en la actividad de marketing directo. Entre ellas, se encuentra la segmentación de contactos. Se utiliza para generar clústeres que agrupen características distintivas entre los estudios al emplear cada modo de entrevista (Berry y Linoff, 2004; Meyers *et al*., 2013). Una fortaleza de esta herramienta es que no solicita mucha información para ejecutar su trabajo. Simplemente se eligen las variables con las que se desea generar los segmentos, sin hacer distinción entre variable a predecir o no, ya que este es un modelo de segmentación y no de clasificación, y se solicita que proporcione tres segmentos fijos, con la intención de separar los modos de entrevista en cada uno.

El modelo resultante de tres segmentos se caracteriza por la distribución de datos de la tabla 8 y de la gráfica 7 (ver tabla 8 y gráfica 7):

Tabla 8. Análisis de clústeres

Clústeres

Importancia de entrada (predictor)

1.0	0.8	0.6	0.4	0.2	0.0

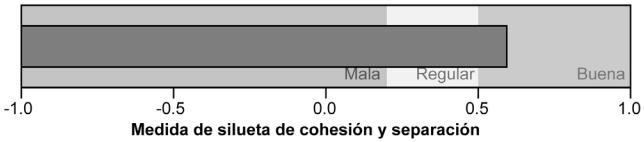
Clúster	Etiqueta	Tamaño	Entradas						
1	Estudios con entrevistas telefónicas	52.0% (51)	Modo de entrevista Telefónicas (94.1%)	Días 1.33	Tipo de muestreo No probabilístico (100.0%)	Costo $12,824.04	Territorio NACIONAL (66.7%)	Tipo de investigación Exploratoria (51.0%)	Cuestionarios 477.35
2	Estudios con entrevistas personales	37.8% (37)	Modo de entrevista Personales (100.0%)	Días 3.11	Tipo de muestreo Probabilístico (100.0%)	Costo $182,882.62	Territorio ESTATAL (62.2%)	Tipo de investigación Descriptiva (100.0%)	Cuestionarios 1,175.41
3	Estudios con entrevistas en línea	10.2% (10)	Modo de entrevista En línea (90.0%)	Días 7.50	Tipo de muestreo No probabilístico (100.0%)	Costo $8,138.97	Territorio NACIONAL (90.0%)	Tipo de investigación Descriptiva (100.0%)	Cuestionarios 1,201.50

Gráfica 7. Análisis de clústeres

Resumen del modelo

Algoritmo	Bietápico
Entradas	7
Clústeres	3

Calidad de clúster



Fuente: elaboración propia.

Se generaron los tres clústeres solicitados y en cada uno de ellos predomina un modo de entrevista sobre los otros. Al examinar las demás variables, pueden distinguirse características asociadas también a cada uno de ellos. Por ejemplo, el costo promedio de cada uno es muy distinto.

El orden en que están las variables de izquierda a derecha representa la importancia que tiene cada una para la clasificación de cada caso en el clúster, por lo que entre más oscuro es el color del fondo, la variable es más importante.

El programa evalúa la calidad de la segmentación realizada y, como se muestra en la gráfica anterior, la considera como buena.

Interpretación de los resultados

Se construyeron varios modelos predictivos de clasificación de datos utilizando diversas técnicas para estimar el valor de la variable predictiva y se obtuvieron diversos porcentajes de error al hacerlo. A continuación, se comparan los modelos obtenidos para determinar la calidad de cada uno.

Con los árboles de decisión, se realizaron ajustes a los parámetros para producir diferentes árboles con distintos porcentajes de error, como se reportó en la tabla 3.

Comparación entre árboles de decisión. Nótese que al variar el método de crecimiento y el porcentaje de la partición de entrenamiento y de prueba se obtienen distintos porcentajes de error, que van desde 0 % de error en la partición de prueba hasta 15.8 %, y desde 1.5 % hasta 10 % en la de entrenamiento. Deben priorizarse aquellos árboles que tienen un mejor desempeño sobre la división de prueba, ya que es la evidencia del modelo puesto en práctica (Larose y Larose, 2015). En consecuencia, se elige el generado con un método QUEST de crecimiento y una división de particiones de 60-40 %, ya que arroja 0 % de error sobre la partición de prueba, como se muestra en la gráfica 2. Árbol de decisiones. No se hizo *pruning* o poda de este árbol, ya que se cuenta con 0 % de error en la predicción.

El análisis discriminante se utilizó para construir un modelo que tiene un porcentaje de error global de 6.1 %, de acuerdo con la gráfica 3. Análisis discriminante, a partir de la construcción de factores representados con diferente magnitud por las variables analizadas. Su debilidad se encuentra en que pronostica erróneamente en 27.3 % de las ocasiones las entrevistas telefónicas cuando en realidad son en línea.

Se utilizó el análisis de K vecinos más cercanos para construir un modelo que se muestra en la gráfica 4. Análisis de K vecinos más cercanos, cuyo error fue de 8.7 % en los datos de entrenamiento y de 3.4 % en los de prueba, por lo que se puede considerar robusto.

La construcción de un modelo a partir de la técnica de perceptrón multicapa de redes neuronales permitió la combinación de parámetros de arquitectura y división de particiones de la base de datos para comparar los porcentajes de error en las estimaciones. Siguiendo con el criterio de aceptar el modelo con menor error de predicción en la partición de datos de prueba, en este caso se trata del modelo que utiliza 60-40 % de porcentaje de partición de la base de datos y una sola capa oculta en la arquitectura de la red que se encuentra en la gráfica 5. Diagrama de red neuronal y en la gráfica 6. Pseudoprobabilidad pronosticada. Este modelo produce 9.4 % de error en la parte de entrenamiento, pero logra 0 % de error en la de prueba.

Finalmente, se utilizó la técnica de segmentación de contactos en clústeres que ofrece el módulo de Marketing Directo del SPSS (IBM Corp., 2020), como se aprecia en la tabla 8 y gráfica 7. Análisis de clústeres, para ganar mayor conocimiento acerca de los atributos que diferencian los tres modos de entrevista que predicen los otros modelos. Esta técnica no sustituye como modelo predictivo a las anteriores —no es un modelo de clasificación, sino de segmentación— pero se puede utilizar

de forma adicional para generar un mejor entendimiento de las características de los estudios que se realizan con cada modo de entrevista.

Se considera que los diferentes modelos construidos cumplen —con diferentes porcentajes de precisión— el objetivo mencionado, ya que permiten seleccionar la técnica de medición en un estudio. Si el criterio de selección del modelo fuera optar por el que arroje el menor error en la partición de datos de prueba (Larose y Larose, 2015), habría un empate entre el modelo recomendado a partir de árboles de decisión y el construido mediante redes neuronales (ambos con 0 % de error en ese tipo de predicción).

Con un criterio de practicidad, la recomendación es elegir el del árbol de decisiones, ya que cualquier persona puede utilizarlo de forma intuitiva, sin necesidad de una herramienta de cómputo: toma en cuenta únicamente el costo del estudio de mercado a realizar y la cantidad de días que se tienen para el levantamiento de información.

La segmentación de clústeres enriquece nuestra comprensión de los datos, ya que muestra las diferencias entre estudios de mercado y la importancia que tienen las variables para cada modo de entrevista. En cuanto a las variables más importantes, coincide con el árbol de decisiones en cuanto a considerar tiempo y costo como características diferenciadoras, y agrega también el tipo de muestreo como elemento a considerar.

5. Conclusiones

El uso de la minería de datos para abordar problemas de negocio impulsa la generación de ventajas competitivas (White y Rollings, 2021). En este estudio, permitió predecir la forma de actuar de una *agencia tipo*, especializada en medición de opinión pública e investigación de mercados.

Cada técnica de recolección de datos tiene sus ventajas y sus limitaciones, tanto desde aspectos prácticos como desde los de calidad de los resultados que producen (Lavrakas, 1987; Steeh, 2008; Vehovar y Lozar Manfreda, 2008). Y aunque el consenso actual es que las encuestas cara cara son las que mejores resultados obtienen (Schober, 2018), las encuestas telefónicas y en línea surgen como alternativas más económicas y ágiles (Couper, 2017; Grewenig *et al*., 2018), aunque con posible detrimento en la representatividad (Zhang *et al*., 2017). El debate para la elección de la

técnica adecuada es complejo, debido a que intervienen aspectos teóricos, metodológicos, presupuestales y prácticos (De las Heras, 1999; Groves *et al*., 2009).

Al analizar lo que hacen en la práctica las agencias especializadas en el ramo de medición de opinión pública e investigación de mercados, estamos abordando el problema desde la dirección opuesta, es decir, en vez de buscar el fundamento que lleva a una empresa de este tipo a elegir cada metodología, revisamos lo que han realizado de manera histórica en sus propios proyectos para extraer ese conocimiento. Al hacerlo de esta forma, no estamos indagando la razón de elegir cada técnica, sino únicamente la decisión como tal, y estamos extrapolando ese conocimiento a un modelo predictivo basado en técnicas de clasificación.

Algunas de estas técnicas de análisis producen modelos con menos errores que otras, pero todas logran la tarea encomendada. La decisión de quedarse con el más certero obedece a un criterio que puede ser cuantitativo o práctico: el que tenga menor error en sus predicciones o el que sea más sencillo de implementar. Una implicación práctica del modelo es que permite tomar mejores decisiones, que es uno de los valores que se obtienen al realizar analítica predictiva (Goul *et al*., 2018; Halper, 2017). Al poder predecir desde los ojos de una agencia promedio la técnica que se recomendará emplear en un proyecto, los clientes podrán preparar presupuestos adecuados de antemano o, dado el presupuesto asignado, podrán saber la técnica que les será recomendada.

Otras implicaciones para los investigadores, tanto de opinión pública como de mercado, son que, desde antes, pueden tener una buena idea de cuál técnica conviene emplear en un proyecto, con lo que podrán realizar la planeación a detalle y ponerlo en marcha de acuerdo con parámetros ya conocidos. Adicionalmente, las implicaciones para medios de comunicación que buscan difundir resultados de estudios de opinión pública son que podrán revisar la técnica de levantamiento de datos de un estudio y contrastarla con la que el modelo sugiere y, por lo tanto, abrir espacios de diálogo para comprender de mejor forma las limitaciones que enfrentan los resultados obtenidos, dando argumentos para promover una cultura de buenas prácticas de investigación.

Como alcances del presente estudio, no debe perderse de vista que se ha construido un modelo que predice lo que una agencia especializada realizaría. No significa que esa técnica sea siempre la adecuada para cada estudio, además de que hay otras agencias que no se han podido incluir en el alcance de la investigación, por lo que este modelo no refleja las decisiones que esas empresas tomarían.

Una limitación de la investigación es que no fue considerado el tipo de encuesta: analítica o descriptiva. En las descriptivas se desea obtener precisión y representatividad en los resultados, mientras que en las analíticas el propósito es establecer relaciones entre variables (Baker *et al*., 2013; Gill y Johnson, 2010). Este factor puede alterar la decisión de la técnica de recolección de información en una encuesta, pero no se contó con dicha información respecto a los proyectos considerados para la construcción del modelo. Otras limitaciones corresponden a elementos que no fueron incluidos en el modelo y cuya influencia podría afectar el resultado de la clasificación: la rentabilidad de la técnica, las condiciones de seguridad del territorio, la confianza en las estimaciones de cada una y la solicitud expresa del cliente de utilizar alguna en particular.

Como línea de investigación futura, se podrían estudiar las razones o fundamentos que llevan a las agencias a elegir esas técnicas en específico, para comprender su lógica y mejorar el modelo con ese conocimiento. De forma paralela, podría cotejarse si la elección de cada modo de entrevista que efectúa una agencia es la adecuada para la calidad de los resultados, utilizando, por ejemplo, el marco del error total de encuesta o *Total Survey Error* (Biemer, 2010), lo que sería una especie de auditoría que permita mejorar el modelo con una variable de calidad en la elección de la técnica. Asimismo, se puede continuar alimentando el modelo con nuevos datos observados de más proyectos y más agencias, conforme se tenga acceso a ellos, y calibrar nuevamente haciendo los ajustes pertinentes de forma similar a la descrita.

Referencias bibliográficas

Asociación Mexicana de Agencias de Inteligencia de Mercado y Opinión Pública, AMAI e Instituto Tecnológico Autónomo de México, ITAM. (2020). *Estudio Anual de la Industria de Investigación de Mercados y Opinión Pública en México. Edición XXII (2019-2020). Reporte a Participantes*.

Baker, R., Brick, J. M., Bates, N. A., Battaglia, M., Couper, M. P., Dever, J. A., Gile, K. J. y Tourangeau, R. (2013). «Summary report of the AAPOR task force on non-probability sampling». *Journal of Survey Statistics and Methodology*, 1 (2), 90-143. https://doi.org/10.1093/jssam/smt008

Berry, M. J. A. y Linoff, G. S. (2004). *Data Mining Techniques. For Marketing, Sales, and Customer Relationship Management* (2ª ed.). Wiley.

Biemer, P. P. (2010). «Total survey error: Design, implementation, and evaluation». *Public Opinion Quarterly*, 74 (5), 817-848. https://doi.org/10.1093/poq/nfq058

Callegaro, M., Lozar Manfreda, K. y Vehovar, V. (2015). En Metzler, K. (ed.). *Web Survey Methodology*. SAGE Publications.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. y Wirth, R. (2007). *Metodología CRISP-DM para minería de datos*. https://www.dataprix.com/es/book/export/html/107

Cornesse, C. y Bosnjak, M. (2018). «Is there an association between survey characteristics and representativeness? A meta-analysis». *Survey Research Methods*, 12 (1), 1-13. https://doi.org/10.18148/srm/2018.v12i1.7205

Couper, M. P. (2000). «Web Surveys: A Review of Issues and Approaches». *Oxford University Press*, 64 (4), 464-494.

Couper, M. P. (2017). «New Developments in Survey Data Collection». *Annual Review of Sociology*, 43 (1), 121-145. https://doi.org/10.1146/annurev-soc-060116-053613

De las Heras, M. (1999). *Uso y abuso de las encuestas: elección 2000, los escenarios*. Océano.

Escobar Terán, H., Alcivar, M. y Puris, A. (2016). «Aplicaciones de Minería de Datos en Marketing». *Revista Publicando*, *3* (8), 503–512.

European Society for Opinion and Market Research, ESOMAR. (2020). *Global Market Research 2020*.

Gera, M. y Goel, S. (2015). «Data Mining - Techniques, Methods and Algorithms: A Review on Tools and their Validity». *International Journal of Computer Applications*, 113 (18), 22-29. https://doi.org/10.5120/19926-2042

Gill, J. y Johnson, P. (2010). *Research Methods for Managers* (4ª ed.). SAGE Publications.

Gorunescu, F. (2011). *Data Mining. Concepts, Models and Techniques* (vol. 12). Springer. https://doi.org/10.1007/978-3-642-19721-5

Goul, M., Raghu, T. S. y Louis, R. D. S. (2018). «APC forum: Governing the wild west of predictive analytics and business intelligence». *MIS Quarterly Executive*, 17 (2), 157-183.

Grewenig, E., Lergetporer, P., Simon, L., Werner, K. y Woesmann, L. (2018). «Can Online Surveys Represent the Entire Population?». *IZA - Institute of Labor Economics*, *11799*, 31.

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E. y Tourangeau, R. (2009). *Survey Methodology* (2nd ed.). John Wiley & Sons.

Halper, F. (2017). *Predictive Analytics*. TDWI. https://tdwi.org/research/2017/06/adv-all-tdwi-navigator-report-predictive-analytics.aspx?tc=page0

Han, J., Pei, J. y Kamber, M. (2011). *Data Mining: Concepts and Techniques.* (3rd ed.) Morgan Kaufmann. https://www.sciencedirect.com/book/9780123814791/data-mining-concepts-and-techniques

IBM Corp. (2020). *IBM SPSS Statistics for Macintosh* (núm. 27). IBM Corp.

Jeffery, M. (2010). *Data-Driven Marketing: The 15 Metrics Everyone in Marketing Should Know*. John Wiley & Sons Inc.

Joseph, S. R., Hlomani, H. y Letsholo, K. (2016). «Data Mining Algorithms: An Overview». *International Journal of Computers & Technology*, 15 (6), 6806-6813. https://doi.org/10.24297/ijct.v15i6.1615

Larose, D. T. y Larose, C. D. (2015). *Data Mining and Predictive Analytics*. John Wiley & Sons Inc. https://www.wiley.com/en-us/Data+Mining+and+Predictive+Analytics%2C+2nd+Edition-p-9781118116197

Lavrakas, P. J. (1987). *Telephone survey methods: Sampling, selection, and supervision.* SAGE Publications.

Lee, H., Kim, S., Couper, M. P. y Woo, Y. (2019). «Experimental Comparison of PC Web, Smartphone Web, and Telephone Surveys in the New Technology Era». *Social Science Computer Review*, 37 (2), 234-247. https://doi.org/10.1177/0894439318756867

MacLennan, J., Tang, Z. y Crivat, B. (2009). *Data Mining with Microsoft SQL Server 2008*. Wiley Publishing, Inc.

Meyers, L. S., Gamst, G. C. y Guarino, A. J. (2013). *Performing Data Analysis Using IBM SPSS*. John Wiley & Sons, Inc.

Pforr, K. y Dannwolf, T. (2017). «What do we Lose with Online-Only Surveys? Estimating the Bias in Selected Political Variables Due to Online Mode Restriction». *Statistics, Politics and Policy*, 8 (1), 105-120. https://doi.org/10.1515/spp-2016-0004

Schober, M. F. (2018). «The future of face-to-face interviewing». *Quality Assurance in Education*, 26 (2), 290-302. https://doi.org/10.1108/QAE-06-2017-0033

Steeh, C. (2008). «Telephone surveys». En Leeuw, Edith D. de, Hox, Joop y Dillman, Don (eds.), *International Handbook of Survey Methodology* (pp. 221-238). Taylor & Francis Group.

Stupakevich, B., Sweenor, D. y Swiderek, S. (2019). *Reporting, Predictive Analytics, and Everything in Between*. O'Reilly Media.

Vehovar, V. y Lozar Manfreda, K. (2008). «Overview: online surveys». En Fielding, N. G, Lee, R. M. y Blank, G. (eds.). *The SAGE Handbook of Online Research Methods* (vol. 1, pp. 177-194). SAGE Publications.

Vehovar, V., Slavec, A. y Berzelak, N. (2012). «Costs and Errors in Fixed and Mobile Phone Surveys», en Gideon, L. (ed.), *Handbook of Survey Methodology for the Social Sciences* (pp. 277-295). Springer New York. https://doi.org/10.1007/978-1-4614-3876-2_16

White, A. y Rollings, M. (2021). *5 Key Actions for IT Leaders for Better Decisions*. Gartner, Inc.

Zhang, X. C., Kuchinke, L., Woud, M. L., Velten, J. y Margraf, J. (2017). «Survey method matters: Online/offline questionnaires and face-to-face or telephone interviews differ». *Computers in Human Behavior*, *71*, 172-180. https://doi.org/10.1016/j.chb.2017.02.006

Anexo

Tabla A1. Características de la base de datos

Variable	Posición	Etiqueta	Nivel de medición
consecutivo	1	Consecutivo	Nominal
Fecha	2	Fecha	Nominal
cuestionarios	3	Cuestionarios	Escala
terr_cod	4	Territorio	Nominal
muestreo_cod	5	Tipo de muestreo	Nominal
Costo	6	Costo	Escala
Días	7	Días	Escala
inv_cod	8	Tipo de investigación	Nominal
modo_cod	9	Modo de entrevista	Nominal

Fuente: elaboración propia.

Tabla A2. Catálogo de codificación

Valor	Etiqueta	
terr_cod	1	Municipal
	2	Estatal
	3	Nacional
muestreo_cod	1	Probabilístico
	2	No probabilístico
inv_cod	1	Exploratoria
	2	Descriptiva
modo_cod	1	Personales
	2	Telefónicas
	3	En línea

Fuente: elaboración propia.

Sobre el autor

El Mtro. Luis Herrero-Corona es alumno del doctorado en Comunicación y Mercadotecnia Estratégica en la Universidad Anáhuac de la Ciudad de México. Obtuvo su MBA por la Universidad de Texas en Austin, graduándose con Mención Honorífica de Excelencia por logro académico sobresaliente. Es licenciado en Mercadotecnia por el Instituto Tecnológico y de Estudios Superiores de Monterrey, Campus Ciudad de México. En su trayectoria profesional ha dirigido durante más de 20 años empresas dedicadas a la investigación de mercados y evaluación de la opinión pública, tanto para el sector privado como para el gubernamental.

luis.herreroco@anahuac.mx
https://orcid.org/0000-0001-8031-3012