Universidad Popular Autónoma del Estado de Puebla
Centro Interdisciplinario de Posgrados
Investigación y Consultoría
Departamento de Ingeniería
Doctorado en Planeación Estratégica y Dirección
de Tecnología

“Estrategias para la identificación de preferencias en la selección de vivienda”

Tesis en para obtener el Grado de Doctor
en Planeación Estratégica y Dirección de Tecnología

Presenta
Julio César Arreola Frías

UPAEP – Secretaría General
Dirección General de Apoyos Académicos
Dirección del Centro de Recursos para el Aprendizaje y la Investigación.
Biblioteca Central - **Karol Wojtyla**

Tesis Digitales Restricciones de uso:

DERECHOS RESERVADOS ©
PROHIBIDA SU REPRODUCCIÓN TOTAL O PARCIAL

Todo el material contenido en esta tesis está protegido por la Ley Federal del Derecho de Autor (LFDA) de los Estados Unidos Mexicanos (México).

El uso de textos, imágenes, gráficas, fragmentos de videos, y demás material que sea objeto de protección de los derechos de autor, será exclusivamente para fines educativos e informativos y deberá citar la fuente de donde la obtuvo mencionando el autor o autores involucrados en el documento.

Cualquier uso distinto como el lucro, reproducción, edición o modificación, será perseguido y sancionado por el respectivo titular de los Derechos de Autor.

Agradecimientos

Gracias a Dios por darme esperanza y salud y todas esas casualidades que solo el nos da.

Dedico esta Tesis a mi madre, Carolina Frías, por ser un ejemplo de fortaleza y perseverancia, por tu amor y oraciones, siempre seras lo mas importante en mi vida, te amo.

Para la culminación de la presente Tesis se ha contado con el apoyo de las siguientes Instituciones y personas

Institucional:

SEP
PROMEP

Personas:

Dr. Marcelo Sánchez-Oro Sánchez, por su generosidad, hospitalidad y apoyo en mi estancia de doctorado, a Auxi su eposa y a sus amistades que fueron muy gentiles en todo momento.

A mis directores de tesis por su tiempo y dedicación:

Dr. J. Agustín Francon y Dr. Damian Gibaja

A Verónica González por su amistad y apoyo en momentos críticos y felices.

A mi amigo de doctorado Oscar José Luis Crruz Reyes, un abrazo al cielo.

A todos mis amigos por su apoyo en la distancia.

Índice

Índice de Figuras

Índice de Tablas

Capítulo 1

Generalidades

1 Introducción

La toma de decisiones (TD) orientadas a cubrir las necesidades de vivienda representa la base fundamental para el desarrollo ordenado de las comunidades, sin embargo, es una tarea compleja que debe integrar los distintos enfoques que definen los criterios de vivienda adecuada. Según la Oficina del Alto Comisionado para los Derechos Humanos (ACNUDH - ONU Habitat, 2010), la evidencia disponible muestra que la vivienda inadecuada afecta a muchas más personas en las áreas urbanas, a pesar de ser más aguda en las áreas rurales. En algunas ciudades del mundo, hasta el 80% de la población reside en vivienda inadecuada. Este déficit de vivienda adecuada se manifiesta principalmente en los países en desarrollo: en África, al menos 211 millones de personas viven en asentamientos de adecuación deficiente; en Asia poco más de 504 millones y América Latina y el Caribe con cerca de 111 millones (ONU Habitat, 2012). Estos datos proporcionan una clara exposición de la mala planificación y administración del desarrollo urbano, en particular, del mal desempeño de las políticas públicas y de las estrategias de la iniciativa privada para facilitar viviendas inclusivas sostenibles y adecuadas.

De acuerdo con la ACNUDH (2010), una vivienda adecuada debe reunir como mínimo los siguientes criterios:

- Seguridad de la tenencia: garantizar a los usuarios protección jurídica contra el desalojo forzoso, el hostigamiento y otras amenazas.
- Asequibilidad: que su costo no sea una amenaza o comprometa el disfrute de otros derechos humanos de los usuarios.

- Disponibilidad de servicios, materiales, instalaciones e infraestructura: el usuario de la vivienda debe tener acceso al servicio de agua potable, instalaciones sanitarias adecuadas, energía para la cocción, la calefacción y el alumbrado.
- Habitabilidad: garantizar seguridad física, protección contra el frío, la humedad, el calor, la lluvia, el viento u otros riesgos para la salud, así como evitar problemas de hacinamiento proporcionando el espacio adecuado en relación con el número de usuarios por vivienda.
- Accesibilidad: que considere las necesidades específicas de los grupos desfavorecidos y marginados.
- Ubicación: que ofrezca acceso a oportunidades de empleo, servicios de salud, escuelas, guarderías y otros servicios e instalaciones sociales, además, que no esté ubicada en zonas contaminadas o peligrosas.
- Adecuación cultural: implica que las características de diseño permitan que se tome en cuenta y se respete, la expresión de la identidad cultural, en función de su dimensión étnica, regional o urbana.

La ACNUDH (2018) afirma que conforme la asequibilidad de la vivienda se convierte en una crisis global, con un fuerte impacto negativo en el bienestar de las personas y en la exacerbación de la desigualdad urbana, también ha surgido como la medida más apropiada para la adecuación de la vivienda. Aunque la importancia de la asequibilidad es irrefutable, deja de lado la participación de ciudadana para adecuar la vivienda a sus necesidades, de manera que permita incrementar el bienestar y satisfacción en los usuarios por contribuir con la adecuación de su vivienda y al desarrollo de su entorno.

1.1 Planteamiento del problema

La vivienda se ha convertido en un factor de especulación motivado por sectores inmobiliarios privados quienes trabajan solamente bajo la lógica del rendimiento económico (ACNUDH - ONU Habitat, 2010). El sector de la construcción constituye un papel clave en el desarrollo global sostenible. Con respecto al ciclo de vida de una vivienda, se sabe que el sector inmobiliario, en su situación actual, no ha logrado un balance entre factores ambientales, económicos y sociales (Gervio et al., 2014, citado por, Pombo, Rivela, & Neila, 2016). Resaltan

los problemas de asequibilidad y accesibilidad que se viven a nivel mundial en el que la relación entre costos e ingreso no asegura la calidad o el entorno en el que se encuentra la vivienda (ACNUDH - ONU Habitat, 2010). En el desarrollo de nuevos proyectos de vivienda es frecuente que los inversores privados obtengan una tasa de rendimiento superior a la tasa de bonos del gobierno, aunque la mayor parte o la totalidad del riesgo de ingresos asociado con el proyecto es asumido por el sector público (Barlow, Roehrich, & Wright, 2010). En otras palabras, se tiene que armonizar los factores del lado de la oferta y la demanda en el mercado de la vivienda para garantizar que la oferta de vivienda se adapte a las necesidades de los ciudadanos en función de la ubicación, el precio y el grupo objetivo (Samad, Zainon, Rahim, & Lou, 2017).

El derecho a la vivienda digna y decorosa es un derecho humano universal recogido en las declaraciones internacionales y muchas constituciones nacionales (Gledhill, 2010). No obstante, es complicado establecer los parámetros adecuados para llevar a la práctica esta definición debido a la complejidad que tiene la generación de vivienda adecuada en un contexto particular. Especialmente, la implementación de dicha definición se ve impactada por la cultura, los cambios en las necesidades sociales, las tradiciones, la ideología y el hábitat (Kunz-Bolaños & Romero Vadillo, 2008). La definición de vivienda digna y decorosa carece de contenidos útiles para constituir las políticas públicas de vivienda adecuada. En estas circunstancias, se requiere construir la conceptualización desde un contexto sociocultural determinado, y de esta manera identificar los indicadores que describen la calidad de la vivienda adecuada para evaluar el progreso de los países en este rubro (López, 2014).

Debido a su importancia, la vivienda se ubica en el centro de las políticas sociales. Sin embargo, en las políticas de vivienda no se consideran las diferencias en el tamaño o composición del hogar, así mismo se omite la cantidad de unidades de ingreso que contribuyen económicamente a los hogares, en otras palabras, el estado actual de la generación de vivienda social no satisface el criterio de asequibilidad (Baker, Mason, & Bentley, 2015). La evidencia internacional sobre la vivienda asequible se centra en dos aspectos generales: las definiciones globales de asequibilidad y la implementación de

mejores prácticas empresariales y políticas públicas en el desarrollo de vivienda social (Gopalan & Venkataraman, 2015). En los últimos años, la participación ciudadana ha sido cada vez más activa en los diferentes ámbitos que tienen relación con las políticas públicas, eliminando los límites entre lo político y lo social (Mannarini, Legittimo y Taló 2008, citado por Sorribas & Garay Reyna, 2014). No obstante, a pesar de que la sociedad civil ha ganado espacios para ejercer sus derechos, el conocimiento sobre estos procesos participativos no ha sido incluido en la TD para construir y asignar vivienda adecuada.

Las metodologías orientadas a la TD para el desarrollo de comunidades sustentables pueden tener un enfoque ecológico, económico, participativo, multicriterio, público-privado, entre otros. Conforme a las características de la TD comunitarias de vivienda adecuada, en las que se busque alcanzar la satisfacción y adecuación de la vivienda es necesario incluir un enfoque participativo para mejorar la accesibilidad, la asequibilidad y la adecuación cultural de la vivienda. Según Ong y Lenard (2002), citado por (Abdul-Aziz & Jahn Kassim, 2011), el significado esencial de la evaluación multicriterio en un contexto comunitario es tolerancia y democracia, cualidades importantes cuando se trata de temas de sostenibilidad, ya que el conflicto entre valores e intereses diferentes, pero igualmente legítimos es un estado latente en TD comunitarias. En el estudio de las comunidades para el desarrollo de vivienda adecuada es fundamental que los tomadores de decisiones tanto del gobierno como de la iniciativa privada analicen el contexto en el que se eligen los criterios a evaluar, además de considerar el efecto en la TD que tienen las reglas heurísticas (Topolinski & Strack, 2015) y los sesgos cognitivos al ponderar cada criterio (Turiel, 2012).

De acuerdo con la literatura, en muchos países se reconoce cada vez más el desequilibrio entre las necesidades locales y las políticas públicas de planificación urbana que se implementan a nivel nacional. En México, de acuerdo con el Consejo Nacional de Evaluación de la Política de Desarrollo Social (CONEVAL, 2018), se identificaron niveles bajos en la satisfacción de las viviendas mexicanas. Si bien, la adecuación cultural, mediante la cual se mide la satisfacción y adaptabilidad de la vivienda, es aceptada como uno de los principios rectores de la política de vivienda no está incorporada al diseño de las

políticas públicas de vivienda. Al analizar los proyectos de desarrollo de vivienda los gobiernos y la iniciativa privada no poseen la información y los conocimientos necesarios para establecer juicios precisos en relación con la ponderación de los criterios que se deben incluir y evaluar (Rolnik, 2013). En otras palabras, la incertidumbre y el sesgo cognitivo en la ponderación de criterios, así como la correlación entre los criterios, son cuestiones que deben ser estudiadas a través de los instrumentos de participación ciudadana. La opinión autocrática en la TD puede tornarse insuficiente cuando se analizan problemas complejos, sobre todo aquellos problemas en donde la solución puede afectar a muchas otras personas, como la TD de vivienda comunitaria en la que el gobierno y las empresas deciden omitiendo las necesidades del usuario final (Munda, 2004; Mendoza & Martins, 2006; Dou, Zhu & Simon, 2012; citado por Grajales Quintero, Serrano Moya, & Hahn Von-H, 2013).

El proceso de diseño para la TD exige identificar y evaluar las diversas necesidades, requerimientos y deseos de los usuarios (Flach, Stumpf González, & Parisi Kern, 2012), mismos que deben ser traducidos adecuadamente al lenguaje constructivo para ser incorporados al producto final. El problema debe ser abordado involucrando igualmente al sector público, al sector privado y a la comunidad para tomar decisiones (Shan, Hou, Ye, & Wu, 2011), sin embargo, como indico previamente, en los procesos TD de desarrollo de vivienda social se realizan de manera bilateral, por una parte el gobierno, con la creación de políticas estructuradas sin un consenso previo y la falta de mecanismos de control que vigilen su cumplimiento, por ejemplo en México la reforma a la ley del Infonavit que prevé la construcción de vivienda con un mínimo de 40 m^2 y aun se construyen viviendas con una superficie menor (INFONAVIT, 2016); por otra parte los desarrolladores de vivienda, que realizan estudios que no toman en cuenta las necesidades de los usuarios, generando problemas de hacinamiento y abandono de vivienda, cartera vencida entre otros (CONEVAL, 2018). La diferencia más importante entre beneficios económicos inmediatos y vivienda sostenible es que la primera ignora los costos y beneficios futuros, mientras que la segunda busca incluirlos (Xiao, Qiu, & Gao, 2016).

Así, es necesario estructurar un proceso de TD comunitarias concernientes a la generación de vivienda adecuada, mediante el cual se tomen en cuenta las

necesidades de los futuros compradores de vivienda para alcanzar el desarrollo de comunidades sustentables donde la sustentabilidad implique la construcción de viviendas adecuadas que satisfagan las expectativas de los usuarios al elegir dónde y cómo desean vivir.

1.2 Justificación de la investigación

Es relevante fomentar la participación ciudadana en la TD relativa en el desarrollo de vivienda adecuada, es necesario mejorar la calidad de vida, disminuir los problemas de hacinamiento, lograr comunidades mejor comunicadas, viviendas adecuadas a la sociedad donde se construyen, políticas públicas más justas y eficientes (CONEVAL, 2018). El análisis detallado del problema, al considerar las preferencias de la comunidad, permitirá a los gobiernos y a los proveedores de vivienda una mejor planificación para eficientizar los recursos.

Para evitar problemas de hacinamiento y abandono de vivienda es necesario que las políticas de vivienda cuenten con un proceso de TD que se fundamente en la accesibilidad, la asequibilidad y la adecuación cultural, para planear comunidades sustentables que motiven a los usuarios de vivienda valorar su inversión mediante la participación en el diseño de su entorno.

Diseñar una metodología que incorpore un conjunto de indicadores de vivienda de calidad en el contexto de los estudios de calidad de vida puede contribuir al cumplimiento de los objetivos para generar comunidades sustentables. Con la implementación del proceso de TD propuesto es posible disminuir el hacinamiento al conocer las necesidades de espacio de los usuarios de vivienda, así como alcanzar beneficios de a una adecuada planeación urbana y territorial. Además, incluir un el enfoque de sustentabilidad en las características de la vivienda permitirá generar acciones que promuevan ciudades sustentables en lo económico, lo ambiental y lo social.

1.3 Objetivos de investigación

1.3.1 Objetivo general

Diseñar un proceso de toma de decisiones para la generación de comunidades sustentables, a través del uso de instrumentos de juicio de expertos y lógica difusa en la ponderación de criterios, para desarrollar alternativas de vivienda que tengan un impacto en el bienestar social.

1.3.2 Objetivos Específicos

- Analizar los diferentes enfoques para TD comunitarias de vivienda.
- Establecer los criterios de vivienda a evaluar mediante el análisis de encuestas de vivienda.
- Seleccionar las metodologías de toma de decisiones que permitan medir la percepción de los usuarios sobre las preferencias de vivienda .
- Diseñar el proceso de TD comunitarias de vivienda.

1.4 Pregunta de investigación

¿Qué factores deben considerarse en el diseño de un proceso de análisis multicriterio para la toma de decisiones comunitarias de desarrollo de vivienda?

1.5 Alcances y limitaciones

1.5.1 Alcances

El alcance de la investigación se centra en la TD en el ámbito del análisis multicriterio, instrumentos de juicio de expertos, de la lógica difusa y minería de datos, como herramientas para el diseño de un proceso de comparación y selección de alternativas de vivienda social.

El proceso de TD comunitarias de desarrollo de vivienda propuesto incorpora los elementos que definen la vivienda adecuada haciendo énfasis en la accesibilidad, la asequibilidad y la adecuación cultural, lo que permitirá constituir una herramienta mediante la cual será posible identificar las necesidades de los usuarios de vivienda para su inclusión en los proyectos comunitarios de desarrollo de vivienda sostenible.

El proceso de TD propuesto también representa una herramienta para mejorar la planeación de los desarrolladores de vivienda ya que a través de uso es posible conocer los atributos, los prototipos y el número de viviendas que demanda la comunidad que se estudia. De igual forma, con esta información recolectada a través de la implementación del proceso propuesto el gobierno puede fundamentar las políticas públicas de vivienda.

1.5.2 Limitaciones

En el desarrollo de esta investigación se propondrá dar respuesta con fundamento en intereses científicos específicos y en un contexto determinado, esta tesis doctoral se enfrenta a diferentes limitaciones. Entre ellas cabe resaltar:

- Las ponderaciones de los criterios de vivienda adecuada que realizan los usuarios representan una medida general que pueden implicar valoraciones transversales de la experiencia en la vivienda en un sentido muy amplio, que involucre aspectos relacionales, afectivos y simbólicos.
- Esta investigación se orienta al estudio de las necesidades comunitarias de acuerdo con las características de vivienda adecuada, y se fundamenta en las características de accesibilidad, asequibilidad y adecuación cultural.
- El diseño del proceso de TD comunitarias de vivienda propuesto, se estructura con base en los resultados de las encuestas sobre vivienda adecuada que se realizan en esta investigación, mismas que se orientan al estudio de vivienda de interés social exclusivamente.

1.6 Viabilidad de la investigación

La investigación es factible debido a que el juicio de expertos y la lógica difusa son metodologías flexibles que pueden adaptarse a la TD para el desarrollo de vivienda social. Por esta razón, los recursos iniciales que se requieren son solo intelectuales en el corto plazo. El tiempo de aplicación de encuestas, análisis de datos y generación de la propuesta se contempla para un mediano plazo y experimentar y afinar el proceso propuesto se prevé para un largo plazo tal como se establece en el cronograma de trabajo de la investigación.

1.7 Metodología de investigación

En la presente investigación se estudia la relación entre las variables que los usuarios de vivienda consideran importantes para determinar la adecuación de la vivienda; por lo que se requiere el análisis de la información relevante relativa a las comunidades de estudio, la cual será recolectada de fuentes oficiales disponibles de instituciones y organizaciones enfocadas al desarrollo de vivienda de España y México que es el contexto en donde se desarrolla la investigación con el propósito de abarcar diversos tipos de vivienda, encontrar las diferencias y similitudes de los dos entornos y determinar si las metodologías utilizadas cumplen con el propósito de medir la percepción de los usuarios. La información relevante relativa a las comunidades de estudio se recaba mediante encuestas que se analizan las etiquetas lingüísticas de medición mediante la lógica difusa para medir la percepción de la importancia de los criterios de vivienda adecuada y las preferencias comunitarias sobre los tipos de vivienda asequibles para las comunidades. El análisis multicriterio de la información permitirá constituir un proceso de TD comunitarias de vivienda que permita la participación de los usuarios para mejorar su bienestar.

En el primer artículo de la investigación se buscó medir la precepción de los criterios a través de la lógica difuzza, donde los valores de las variables lingüísticas representan conjuntos difusos definidos en términos nominales que se utilizan para aproximar y combinar mediante el método TOPSIS (Technique for Order Performance by Similaritiy to Idea Solution) para determinar las preferencias en los usuarios. La lógica difusa ofrece un modelo de la percepción clasificadora del universo gracias a la posibilidad de permitir la atribución de un objeto a varias clases en el grado en que sea necesario, confiriéndole grados de elasticidad a los grupos (Zimmermann, 1996). En la segunda parte se construye una red neuronal utilizando de la misma forma las etiquetas lingüísticas de las variables nominales como conjuntos difusos definidos en las encuestas oficiales de vivienda de España y México y también se utilizan variables cuantitativas y dicotómicas, con el propósito de determinar cuáles son las características de vivienda de mayor importancia para los usuarios y los tipos de vivienda de mayor aceptación en las comunidades de estudio. En el tercer artículo se fundamenta y se estructurar el diseño del proceso de TD comunitarias que se plantea

analizando las diferentes regiones de México mediante el análisis de dos esquemas de división regional económica del país con el propósito de analizar el sesgo de los tipos de vivienda por su difusión y su relación con los tipos de vivienda de mayor importancia en cada región de acuerdo con los resultados de las redes neuronales.

Además, se utilizarán los recursos disponibles para recolección de información: bases de datos de artículos científicos, libros especializados, tutoriales, etc. También, se plantearán colaboraciones con otros miembros del grupo de investigación y miembros de grupos externos, así como la asistencia a cursos y congresos especializados en la materia.

1.8 Resultados esperados

Las expectativas de la presente investigación es identificar los criterios que definen las necesidades referentes la vivienda en las comunidades. Una vez identificados estos criterios, es posible constituir el diseño de un proceso de TD que permita a las personas intervenir en el proceso de la configuración de su futuro hogar para alcanzar mayor bienestar.

1.9 Contribuciones originales esperadas

Diseñar un proceso de TD comunitarias de vivienda para potencializar la participación de las personas sobre el tipo de vivienda en el que desean vivir conforme a sus ingresos y expectativas. Además, se establecerá un precedente de los criterios que actualmente consideran las personas al tomar la decisión de adquirir una vivienda. Por último, el proceso propuesto en esta investigación servirá como herramienta para realizar proyecciones de demanda de vivienda a los desarrolladores.

1.10 Referencias

1. Abdul-Aziz, A. R., & Jahn Kassim, P. S. (2011). Objectives, success and failure factors of housing public-private partnerships in Malaysia. *Habitat International*. https://doi.org/10.1016/j.habitatint.2010.06.005
2. ACNUDH - ONU Habitat. (2010). El derecho a una vivienda adecuada. Folleto informativo nº21. *Revista de Antropología Social*, *19*, 103–129. https://doi.org/-
3. Baker, E., Mason, K., & Bentley, R. (2015). Measuring Housing Affordability: A Longitudinal Approach. *Urban Policy and Research*. https://doi.org/10.1080/08111146.2015.1034853
4. Barlow, J., Roehrich, J. K., & Wright, S. (2010). De facto privatization or a renewed role for the EU? Paying for Europe's healthcare infrastructure in a recession. *Journal of the Royal Society of Medicine*. https://doi.org/10.1258/jrsm.2009.090296
5. Colegio Mexiquense., I., & Romero Vadillo, G. (2008). *Economía, sociedad y territorio : EST. Economía, sociedad y territorio* (Vol. 8). Retrieved from https://biblat.unam.mx/es/revista/economia-sociedad-y-territorio/articulo/naturaleza-y-dimension-del-rezago-habitacional-en-mexico
6. CONEVAL. (2018). *Estudio Diagnóstico del Derecho a la Vivienda Digna y Decorosa 2018*. Ciudad de México. Retrieved from https://www.coneval.org.mx/Evaluacion/IEPSM/Documents/Derechos_Sociales/Estudio_Diag_Vivienda_2018.pdf
7. Gledhill, J. (2010). El derecho a una vivienda. *Revista de Antropologia Social*. https://doi.org/-
8. Grajales Quintero, A., Serrano Moya, E. D., & Hahn Von-H, C. M. (2013). Los métodos y procesos multicriterio para la evaluación. *Revista Luna Azul*, (36), 285–306. https://doi.org/10.17151/luaz.2015.40.14
9. Kochen, J. J. (2016). El ideal del multifamiliar. *Vivienda Infonavit*, *1*(1).
10. López Ramón, F. (2014). El derecho subjetivo a la vivienda. *Revista Española de Derecho Constitucional*.
11. ONU Habitat. (2012). Viviendas y mejora de asentamientos precarios. Retrieved from http://es.unhabitat.org/temas-urbanos/viviendas/
12. Pombo, O., Rivela, B., & Neila, J. (2016). The challenge of sustainable building renovation: Assessment of current criteria and future outlook. *Journal of Cleaner Production*. https://doi.org/10.1016/j.jclepro.2015.06.137
13. Rodríguez, A. S. (2015). La participación ciudadana en México. *Estudios Políticos*, *34*, 93–116. https://doi.org/10.1016/J.ESPOL.2015.05.001
14. Rolnik, R. (2013). Late Neoliberalism: The Financialization of Homeownership and Housing Rights. *International Journal of Urban and Regional Research*. https://doi.org/10.1111/1468-2427.12062
15. Samad, D., Zainon, N., Rahim, F. A. M., & Lou, E. (2017). Malaysian affordability housing policies revisited. *Open House International*. https://doi.org/10.1051/matecconf/20166600010
16. Shan, X., Hou, W., Ye, X., & Wu, C. (2011). Decision-Making Criteria of PPP Projects : Stakeholder Theoretic Perspective, *5*(5), 627–631.
17. SNIIV2.0. (2018). Demanda Potencial - Chihuahua - INFONAVIT (Número de derechohabientes). Retrieved from http://sniiv.conavi.gob.mx/Reports/Infonavit/Demanda_Pot.aspx
18. Sorribas, P. M., & Garay Reyna, Z. (2014). La participación, entre la

democracia participativa y la democracia directa. Aportes desde un enfoque psicosocial. *Polis*, *10*(2), 39–69. Retrieved from http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S1870-23332014000200003&lng=es&nrm=iso&tlng=es

19. Topolinski, S., & Strack, F. (2015). Heuristics in Social Cognition. In *International Encyclopedia of the Social & Behavioral Sciences: Second Edition*. https://doi.org/10.1016/B978-0-08-097086-8.24018-X
20. Turiel, E. (2012). Social decisions, social interactions, and the coordination of diverse judgments. In *Social Life and Social Knowledge: Toward a Process Account of Development*. https://doi.org/10.4324/9780203809587
21. Xiao, L., Qiu, Q., & Gao, L. (2016). Chinese housing reform and social sustainability: Evidence from post-reform home ownership. *Sustainability (Switzerland)*. https://doi.org/10.3390/su8101053

Capítulo 2

Selección de vivienda a través de redes neuronales: una comparación entre España y México

Resumen

El presente artículo compara las características de vivienda de usuarios en España y México mediante una red neuronal multicapa cual se entrena para la selección del tipo de vivienda adecuada por nuevos usuarios. El propósito de esta investigación de analizar los sesgos y las ponderaciones sinápticas de las variables que se analizan. Los resultados muestran que el sesgo de los datos y la ponderación de las variables no influyen en los índices de precisión de la red neuronal para la clasificación de vivienda; así, la clasificación que se hace de las viviendas es independiente de los sesgos y captura las preferencias de los usuarios de vivienda en cada país. La robustez de los resultados se hace comparando diferentes arquitecturas de retroalimentación para la red neuronal con la intención de mejorar la exactitud de los resultados mediante diferentes entrenamientos.

Palabras calve: Selección de vivienda, Adecuación cultural, RNA.

2.1 Introducción

El sector de la construcción tiene un papel clave para el desarrollo sostenible, en particular de las áreas urbanas. Sin embargo, este no ha logrado un balance equilibrado entre factores ambientales, económicos y sociales con respecto a la vivienda en que se contemple, entre otros factores, la funcionalidad de diseño de la distribución y un proceso de selección-asignación que favorezca el desarrollo y bienestar de la familia El reto consiste en generar “viviendas adecuadas” de manera exitosa (Samad, Zainon, Rahim, & Lou, 2017). En torno a lo cual, existen diferentes enfoques. Por ejemplo, el Consejo Nacional de Evaluación de la

Política de Desarrollo Social de México (CONEVAL, 2018) considera la asequibilidad como un criterio fundamental para determinar si una vivienda es adecuada o no. Aunque la asequibilidad es esencial para el correcto desarrollo urbano, con frecuencia, las regulaciones normativas y/o su aplicación concreta, por parte de los gestores, ignora las preferencias de los individuos, muchas de ellas motivadas desde el punto de vista económico, social, cultural, e incluso por una elemental demanda de adecuación a las necesidades de una movilidad sostenible. Por tanto, la satisfacción de los usuarios no está garantizada. La Oficina del Alto Comisionado para los Derechos Humanos (ACNUDH - ONU Habitat, 2010), establece que una "vivienda adecuada" debe cumplir con los criterios de 1. Seguridad de la tenencia, 2. Disponibilidad de servicios, 3. Asequibilidad, 4. Habitabilidad, 5. Accesibilidad, 6. Ubicación y 7. Adecuación Cultural.

Adquirir una vivienda es una decisión que condiciona la economía doméstica de las familias. Su bienestar queda determinado por años al pago de la hipoteca y/o la posibilidad de no disponer de una vivienda adecuada. En principio, la lógica de asignación de "viviendas adecuadas" supone que los gobiernos diseñan políticas públicas de construcción que garanticen a las empresas la recuperación de su inversión, así mismo se trata de garantizar un nivel de satisfacción alto a los usuarios potenciales de las viviendas (Samad, Zainon, Rahim, & Lou, 2017). En otras palabras, la toma de decisiones (TD) orientada a cubrir las necesidades de vivienda es fundamental para el desarrollo eficiente del sector y de las comunidades.

La mayoría de los estudios sobre "vivienda adecuada" se centran en la asequibilidad de la vivienda. La literatura se enfoca en determinar los factores que influyen en la formación del nivel de precios de la vivienda (Gaspareniene et al., 2014) dejando en segundo plano la satisfacción de los usuarios (ONU-Habitat, 2019). En ésta línea, Rahadi, Wiryono, Koesrindartoto, & Syawmil (2018) utilizan el modelo ANOVA y la regresión lineal para identificar las variables con mayor impacto en la asignación de viviendas, mientras que Hassanudin (2016) utiliza el Proceso Analítico Jerárquico Difuso (FAHP). Similarmente, Choy, Ho, & Mak (2012) analizan la influencia de las prioridades, expresadas por los usuarios, en el precio de la vivienda, mediante una regresión cuantílica. Con respecto a la predicción precios, Choy et al., (2012) y Yao et al., (2018) lo hacen

por medio de redes neuronales; en ambos casos proporcionan evidencia de que la cercanía con estaciones de transporte y/o lugares turísticas incrementa los precios y genera desigualdad entre la población.

En general, ONU-Habitat (2019) señala que tanto investigadores como creadores política vivienda han dejado de lado la adecuación cultural y se han preocupado por analizar este mercado, principalmente, como motor de crecimiento económico y de creación de empleo, eludiendo otras implicaciones que afectan directamente a la sostenibilidad urbana. Por tanto se evidencia la omisión de la adecuación a características de orden cultural del usuario[1], en el desarrollo de los planes de promoción de vivienda (ONU-Habitat, 2019). El resultado es que con frecuencia, se genera un modelo de ciudad poco habitable, caracterizada por la degradación urbana y la segregación especial; donde la nota dominante es la dispersión de las comunidades y la infrautilización de zonas comunes (Fariña y Naredo, 2010).

La presente investigación propone un mecanismo de priorización de las características de las viviendas, basado en redes neuronales artificiales (RNA), el cual contempla la adecuación cultural (en los términos establecidos por ACNUDH - ONU Habitad 2010) y los tipos de vivienda identificadas en los registros de las bases de datos generadas de Encuesta Continua de Hogares del Instituto Nacional de Estadística (INE, 2017) de España y de la Encuesta Nacional de los Hogares del Instituto Nacional de Geografía Estadística e Informática (INEGI, 2017) de México. Como se mostrará, el mecanismo propuesto permite prever el tipo de vivienda que se adapte a las necesidades de los nuevos adquisidores por medio de la identificación de los atributos que caracterizan cada territorio. La RNA permite ponderar la importancia relativa de las variables sobre el tipo de vivienda que los usuarios prefieren, y determinar los sesgos de la última capa oculta para las neuronas de salida de las redes de España y México. Para este propósito, el algoritmo utiliza la función tangente hiperbólica para el entrenamiento de las capas ocultas pues se trata de un problema de clasificación; mientras que la función de activación de la capa de

[1] La adecuación cultural, en este caso implica que las características de diseño permitan que se tome en cuenta y se respete, la expresión de la identidad cultural, en función de su dimensión étnica, regional o urbana (ACNUDH - ONU Habitad 2010).

salida es de tipo softmax, que se usa para la regresión logística multiclase (Llinás et al., 2016), o regresión logística multinomial, adecuado a los siete criterios de "vivienda adecuada" establecidos por ONU-Habitat. Mediante la comparación de las RNA con datos de México y España, se obtuvieron predicciones sobre el tipo de vivienda conveniente para nuevos registros de prueba. Como podrá observarse, los resultados muestran sesgo en la base de datos de España, probablemente atribuibles a que la recolección de datos se realiza de forma telefónica. Sin embargo, estadísticamente, este sesgo no representa un obstáculo para que la RNA identifique el tipo de vivienda adecuado para nuevos registros.

2.2 Metodología

2.2.1 Redes Neuronales Artificiales

Las RNA son una rama de la inteligencia artificial que permite el estudio de problemas multifactoriales por medio de la que ejecución de funciones que analizan un número fijo de entradas y producen un número fijo de salidas (Eichie, Oyedum, Ajewole, & Aibinu, 2017). En otras palabras, son modelos matemáticos programables cuya integración computacional se hace en paralelo, es decir, se constituyen en capas con múltiples conexiones, emulando la conexión de neuronas del cerebro humano(Anand, 2017). La mayoría de los tipos de arquitectura de las RNA incluye una capa de neuronas ocultas, las cuales son el puente entre la entrada y la salida. Las conexiones entre las neuronas tienen un valor de ponderación asociado (peso), además del sesgo (umbral). Este último es un parámetro adicional que ayuda a ajustar al modelo a los datos dados controlando la función de activación (Du & Swamy, 2014). Entonces, los pesos y umbrales fijan los valores de salida del conjunto de valores de entrada.

La red neuronal sigue un proceso de entrenamiento para determinar los pesos y umbrales de mejor ajuste, ver Figura 2.1. La neurona de sesgo se agrega al inicio / final de la entrada y a cada capa oculta, es decir, no está influenciada por los valores de la capa anterior, por lo que estas neuronas no tienen conexiones entrantes. Matemáticamente, el proceso de entrenamiento se expresa de la siguiente manera

$$Salida = \Sigma(\, entradas * ponderaciones) + sesgo \qquad (1)$$

Según el tipo de entrenamiento (Ferreira et al., 2018), las redes se pueden clasificar en dos grupos:

- **RNA con entrenamiento supervisado**. Se entrenan presentando, para cada conjunto de entradas, las salidas que se esperan ellas produzcan. Entonces, los algoritmos de entrenamiento calculan pesos y sesgos nuevos con el objetivo de minimizar el error entre la salida deseada y la resultante.
- **RNA sin entrenamiento supervisado**. Los algoritmos de entrenamiento calculan nuevos pesos de manera aleatoria. Estas redes se utilizan como clasificadores, pues se caracterizan por asociar una combinación de entradas específicas con una sola salida; en otras palabras, identifica similitudes en los datos y reacciona con base en la presencia o ausencia de los elementos comunes de cada registro de datos.

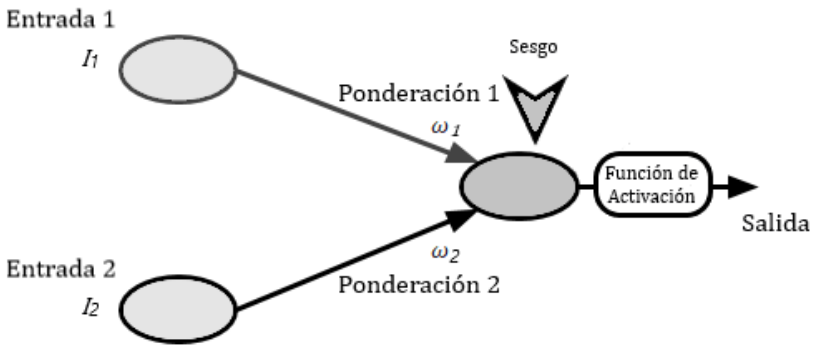

Figura 2.1. Proceso de entrenamiento de una RNA. (Fuente: Elaboración propia con información de Du & Swamy (2014), *Neural networks and statistical learning*. *Neural Networks and Statistical Learning* , pag. 5).

Las funciones de activación más comunes en RNA son las siguientes:

- **Función Heaviside**, también conocida como *función umbral* o *función escalón unitario*, es una función discontinua cuyo valor es uno para valores positivos y cero para valores negativos.

$$\gamma(c) = \begin{cases} 1 \; si \; c \geq 0 \\ 0 \; si \; c < 0 \end{cases}. \tag{2}$$

- **Función sigmoide logístico** es una curva en forma de *S* que se basa en una matriz de ganancia de información. Dicha matriz sugiere una transición de valores bajos hacia valores altas a través de un punto de inflexión.

$$\gamma_{(c)} = \frac{1}{1 + e^{-c}}. \tag{3}$$

- **Función softmax (exponencial normalizada)**, generaliza la distribución de Bernoulli para problemas multiclase. Es de tipo sigmoide y separa la clasificación multiclase debido a que escala las entradas previas, en un rango entre 0 y 1, y normaliza la capa de salida, es decir, la suma de las neuronas es igual a uno.

$$\gamma_{(c)_j} = \frac{e^{c_j}}{\sum_{k=1}^{K} e^{c_k}}. \tag{4}$$

- **Función tangente hiperbólica,** es una curva con propiedades similares a la activación sigmoide, por lo que se le considera su alternativa. Las salidas están acotadas entre -1 y 1.

$$\gamma_{(c)} = tanh(c)\frac{e^{c} - e^{-c}}{e^{c} + e^{-c}}. \tag{5}$$

Es importante mencionar que la selección de la función de activación depende del objetivo que se persiga. Por ejemplo, los modelos de clasificación comúnmente utilizan la función de activación sigmoide, mientras que los modelos predictivos asumen una función de activación lineal (Vanus et al., 2020).

2.2.3 El modelo Perceptrón Multicapa (MLP)

El modelo de perceptrón de alimentación multicapa (MLP) es una RNA que consiste en un número finito de capas sucesivas, a su vez, cada una posee un número finito de neuronas, las cuales se conectan entre sí en la siguiente capa, como si hicieran sinapsis. De este modo, la información fluye en una sola dirección, de una capa a la siguiente (feedforward). La arquitectura de una red MLP (Figura 2.2) se conforma por la primera capa o capa de entrada. Posteriormente se encuentran las capas intermedias, u capas ocultas, y por último se ubica la capa de salida, cuyas neuronas también son llamadas nodos de respuesta (ElKessab, Daoui, Boukhalene, & Salouan, 2014).

Como indica la expresión (1), las neuronas de la primera capa oculta están en función de la suma ponderada de las neuronas de la capa de entrada. Dicha función se denomina *función de activación*, y los valores de las ponderaciones se determinan mediante el algoritmo de estimación, el cual busca minimizar los

errores. Si la red contiene una segunda capa oculta, sus neuronas son una transformación de la suma ponderada de las neuronas de la primera capa oculta. El proceso anterior se repite de manera iterativa para cada capa oculta hasta llegar a la capa de salida.

Los modelos MLP se clasifican de acuerdo con el número de capas ocultas que poseen, y no por su número total de capas, es decir, se clasifican por el número de capas excluyendo la capa de entrada y la de salida. El término MLP se aplica genéricamente a todos los modelos con al menos una capa oculta. Las normas y reglamentos que rigen el modelo MLP (Pinkus, 1999), son las siguientes:

- La capa de entrada tiene como salida de su neurona *j* el valorx_{0j}.
- La neurona *k* de la capa *i* recibe la salida x_{ij} de cada neurona *j* de la capa (*i - 1*). Los valores x_{ij} se multiplican por las ponderaciones sinápticas ω_{ijk}, y los productos se suman para obtener el siguiente valor de salida. Matemáticamente, para cada neurona k, la capa $i+1$ se construye de la siguiente manera:

$$x_{i+1,k} = \gamma\left(\sum_{j} \omega_{ijk} x_{ij} - \theta_{ik}\right), \tag{6}$$

  Donde γ es la función de activación, θ_{ik} son los umbrales o sesgos y ω_{ijk} son las ponderaciones sinápticas.

Los pesos y umbrales se determinan durante el aprendizaje o entrenamiento por medio de algún mecanismo definido exógenamente; por ejemplo, el algoritmo de retropropagación (backpropagation) es de los más utilizados (Du & Swamy, 2014). El termino retropropagación se refiere a la forma en que se propaga la información hacia atrás, iniciando desde la capa de salida; esto permite que las ponderaciones sinápticas de las neuronas ubicadas en las capas ocultas cambien durante el entrenamiento. La variación de las ponderaciones sinápticas se debe a la sensibilidad de la función activación. La misma función de activación se utiliza en cada una de las capas ocultas; sin embargo, la elección de la función de activación en la capa de salida depende del tipo de tarea impuesto (Marín Diazaraque, 2007).

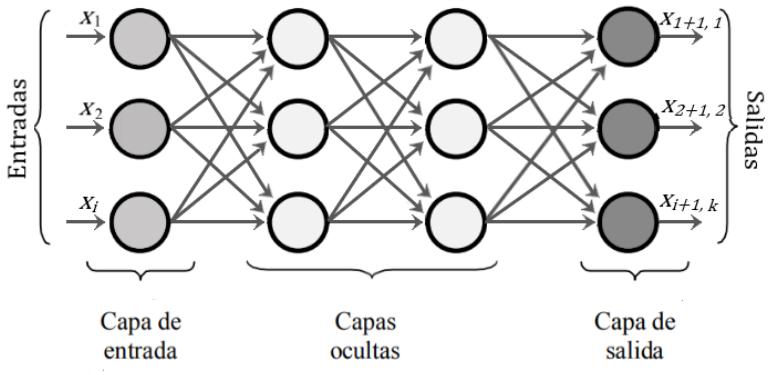

Figura 2.2. Arquitectura de la RNA MLP. (Fuente: Elaboración propia con información de ElKessab et al., (2014), International Journal of Computer Applications, 107 (21), 25–30).

Los resultados en cada vector de entrada se utilizan para la formación y validación mediante múltiples iteraciones en las que el modelo identifica las ponderaciones y sesgos que mejor ajusten la red al comportamiento de los datos. Es importante señalar que el resultado es un *aprendizaje automático,* que se refiere al proceso por el que el algoritmo toma decisiones inteligentes mediante el reconocimiento de patrones, aprendizaje, basados en los datos de la muestra. En otras palabras, se establece un patrón a seguir entre el problema y la respuesta (Han, Kamber, & Pei, 2012).

2.3 Modelo

La construcción del modelo de clasificación y predicción del tipo de vivienda, para México y España, sigue los siguientes pasos:

- Análisis de la información y recopilación de las bases de datos.
- Preprocesamiento de las bases de datos y selección de variables a comparar.
- Diseño del modelo de clasificación (mediante el uso de IBM SPSS Modeler 18).
- Evaluación de los resultados obtenidos.

2.3.1 Análisis de la información y recopilación de las bases de datos

En esta investigación se utilizan los datos de la Encuesta Continua de Hogares del Instituto Nacional de Estadística (INE, 2017) de España y de la Encuesta Nacional de los Hogares del Instituto Nacional de Geografía Estadística e Informática (INEGI, 2017) de México. Ambas encuestas se realizan con el objetivo de conocer las principales características de la población, los hogares y

las viviendas. Se seleccionaron datos de 2017 ya que son los registros más recientes que se tienen para México.

2.3.2 Preprocesamiento de bases de datos y selección de variables

Se ha procedido a depurar los registros para homogeneizar las bases de datos seleccionadas. Posteriormente se obtuvieron 100,542 registros de España y 31,698 registros de México. Por último, se ha determinado el conjunto de variables que se van a someter a prueba, mediante el uso de los siguientes criterios de selección:

- Se consideran solo las variables de relativas a la vivienda presentes en ambas bases de datos de España y México.
- Considerando las variables anteriores, en la base de datos unificada resultante, sólo se incluyen aquellas que sean afines a los 7 criterios de vivienda adecuada de ONU-Habitat (ACNUDH - ONU Habitat, 2010).
- Se descartan las variables que tienen datos idénticos en todos los registros y también aquellas con error de captura o ausencia de información.
- Se concatenan las variables con formatos diferentes en ediciones previas.

La descripción de las variables estudiadas, su tipo, valores de registro y su justificación de selección se presentan en la Tabla 2.1.

Tabla 2.1. Descripción de las variables. (Fuente: elaboración propia a parir de Encuesta Continua de Hogares del Instituto Nacional de Estadística (INE, 2017) y Encuesta Nacional de los Hogares del Instituto Nacional de Geografía Estadística e Informática (INEGI, 2017)).

Variable	Tipo	Valores	Justificación
Tipo de vivienda (TV)	Nominal	1,2,3,4,5,6 España 1,2,3,4,5 México	Variable dependiente
Cocina (C)	Dicotómica	1 y 2	Fundamental para la alimentación de los usuarios de vivienda.
Número de cuartos de dormitorio (CD)	Continua	1,2,3,4,…,n	Permite disminuir el hacinamiento.
Número de cuartos en la vivienda (NC)	Continua	1,2,3,4,….,n	Considera la adecuación cultural mediante la personalización de la vivienda.
Número de baños completos (BC)	Continua	1,2,3,4,….,n	Fundamental para la higiene de los usuarios de vivienda.
Número total de residentes (TR)	Continua	1,2,3,4,….,n	Facilita determinar el tipo de vivienda.
Tipo de pago de tenencia (T)		1,2,3,4	Identifica la seguridad y protección jurídica
Ubicación geográfica (UG)	Nominal	Clave numérica	Conectividad y proximidad a centros de trabajo
Tamaño de la localidad (TL)	Nominal	1,2,3,4,5,6,7,8,9,10,11 España 1,2,3,4 México	Permite identificar la cobertura de servicios.

2.3.3 Diseño de la red neuronal

La arquitectura de la RNA MLP se diseñó en la herramienta de software IBM SPSS Modeler que trabaja mediante la interfaz virtual Watson Studio. La figura 2.3 muestra el flujo del procesamiento de datos para la construcción de los modelos. La modelación de cada una de las redes neuronales de España y México se construye mediante el uso de diferentes herramientas que se enfocan a establecer las arquitecturas de las RNA; cada herramienta establece una actividad en particular para la red neural y se representan mediante íconos en el software.

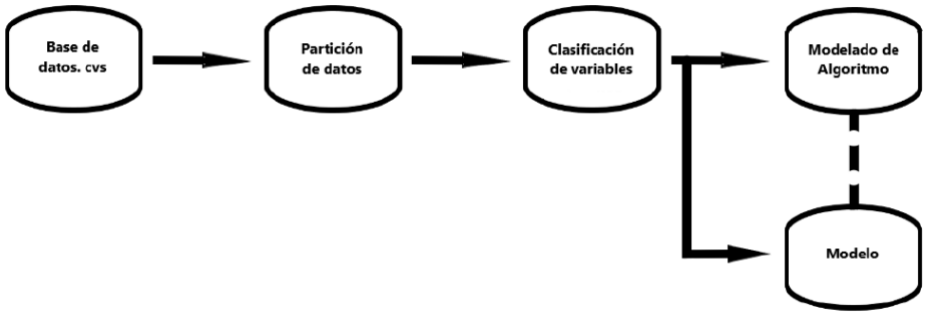

Figura 2.3. Flujo de procesamiento de datos para el desarrollo de los modelos.
(Fuente: elaboración propia IBM SPSS Modeler (2020)).

El flujo de la arquitectura para cada una de las redes consta de los siguientes pasos:

1. Se importan los registros de entrada que integran las bases de datos a través del ícono de Asignación de Datos, el cual permite trabajar con archivos de Excel. CSV.
2. Se determina la partición de los registros de las bases de datos considerando tres conjuntos: uno de entrenamiento, otro de prueba y un tercero de validación; éste último se utiliza para determinar el fin del entrenamiento de la red con el propósito de evitar el sobre entrenamiento. La partición se realiza mediante el icono de Partición, con una proporción de 65% de entrenamiento, 25% de pruebas y 10% de validación.
3. El icono Tipo se utiliza asignar la tipología de las variables en función de sus componentes internos (continua, nominal, categórica, etc.), el rol que desempeñan en el sistema de hipótesis (dependiente, independiente, identificación de registro, etc.) y el valor de medición de cada una de las variables de interés. Para minimizar el efecto del tamaño de los valores de entradas y de salidas, y aumentar la efectividad de aprendizaje, el

algoritmo normaliza los registros de las variables de entrada a valores entre cero y uno. El software permite ajustar el tipo de normalización en caso de que se deseen valores más grandes.

4. Se integra el icono de modelado RNA en el cual se seleccionan las características del algoritmo a ejecutar (elección de variables de entrada, el nivel de confianza de las predicciones o el número de ciclos que se desea ejecutar). En este caso, seleccionamos 1000 ciclos, predeterminado por el software, y establecemos un conjunto de prevención de sobreajuste del 10% para rastrear errores durante el entrenamiento a fin de evitar que el método modele la variación de probabilidad en los datos. Por último, se especifica el número de capas ocultas y el número de neuronas en cada capa oculta.

2.3.4 Características de la red neuronal

En el presente trabajo se emplea una RNA MLP para España y México con tres y cuatro capas, respectivamente. La capa de entrada tiene ocho neuronas (una por cada una de las variables que se especifican en la sección anterior), mientras que las redes difieren por el número de capas ocultas, una para España y dos para México debido a la diferencia de registros. Una capa es lo más recomendado si el error es bajo y con dos capas ocultas representa un clasificador universal (Coloma et al., 2019). Sin embargo, se ha demostrado que para la mayoría de problemas es suficiente una sola capa oculta (Funahashi, 1989 citado por Picón, 2011). La información de la arquitectura de cada red se describe en la Tabla 2.2.

Tabla 2.2. Información de la arquitectura del modelo. (Fuente: elaboración propia).

Información del Modelo	Número de capas ocultas	
	1	2
Variable dependiente	tipo_viv	tipo_viv
Método de contrucción de modelos	Perceptron Multicapa	Perceptron Multicapa
Tipo de modelo	Clasificación	Clasificación
Número de varibles dependientes	8	8
Número de neuronas en la capa oculta #1	18 España, 16 México	15 España y México
Función de activación en la capa oculta	Tangente hiperbólica	Tangente hiperbólica
Número de neuronas en la capa oculta #2		11 España y México
Función de activación en la capa de salida	Softmax	Softmax
Número de neuronas en la capa de salida	6 España, 4 México	6 España, 4 México

La presente investigación busca un error bajo en la eficiencia de clasificación de la RNA ya que la variable dependiente es multiclase. En las capas ocultas se utilizó la tangente hiperbólica como función de activación y en la capa de salida, donde están las neuronas de clasificación del tipo de vivienda, la función de activación softmax ya que la variable dependiente es multiclase, con seis tipos de vivienda para España (Figura 2.4) y cinco tipos de vivienda para México (Figura 2.5) de acuerdo con las bases de datos de cada país.

En las Figuras 2.4 y 2.5 se muestra, en color oscuro, el sesgo de los registros respecto al tipo de vivienda. Para el caso de España, se tiene el tipo de vivienda 2 (vivienda unifamiliar adosada o pareada), mientras que, en el caso de México, el tipo de vivienda 1 (vivienda independiente). El resultado es esperado pues ambos representan las viviendas de mayor uso en cada país.

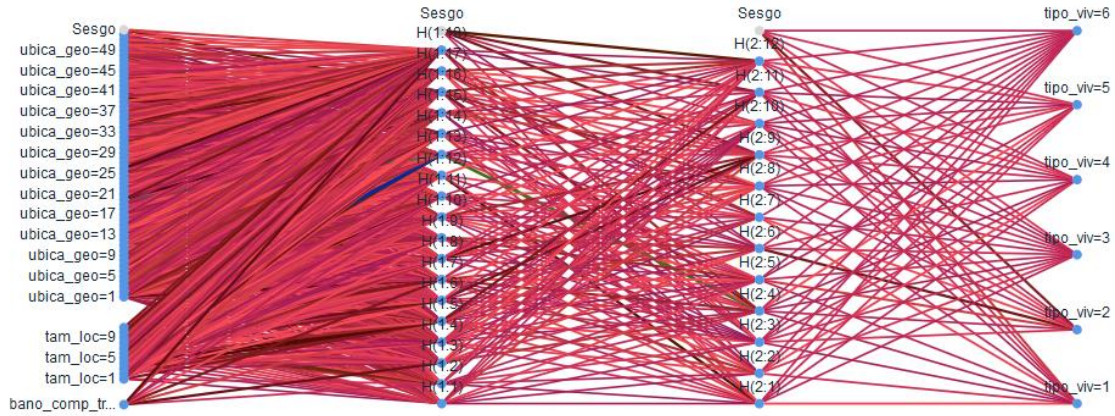

Figura 2.4. Diagrama de RNA con cuatro capas (dos ocultas) de la base de datos de España.
(Fuente: elaboración propia IBM SPSS Modeler (2020)).

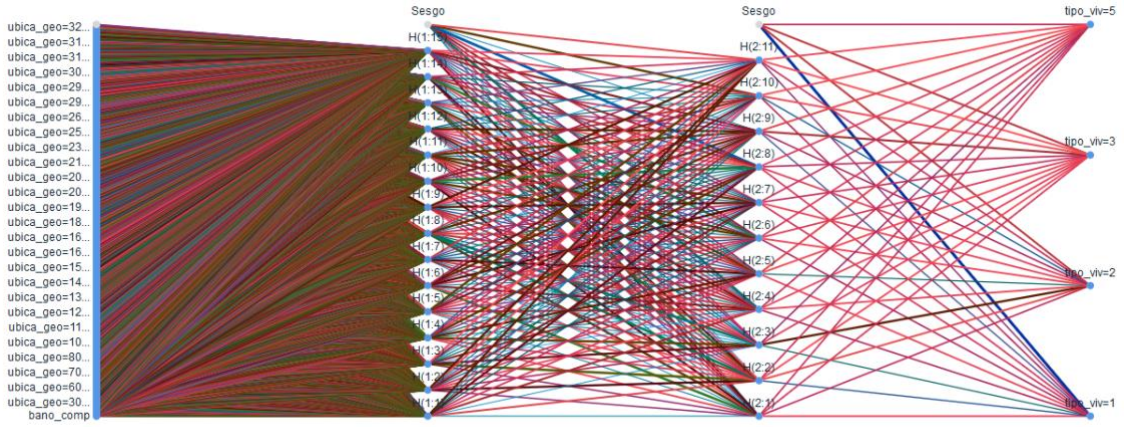

Figura 2.5. Diagrama de RNA con cuatro capas (dos ocultas) de la base de datos de México.
(Fuente: elaboración propia IBM SPSS Modeler 2020).

2.3.5 Evaluación de los resultados obtenidos

Tras el entrenamiento de las RNA's, se procede a evaluar la eficiencia y precisión de los resultados. Además de los resultados, el software proporciona diferentes indicadores de evaluación, los cuales se muestran en la Tabla 2.3; dicho análisis considera tres variantes de la red neuronal:

1. Valores reales de las bases de datos y una sola capa oculta,
2. Valores reales de las variables nominales y valores normalizados de las variables continuas de las bases de datos con una capa oculta y
3. Valores reales de las bases de datos con dos capas ocultas).

Los índices que consideramos para la evaluación de los resultados son los siguientes:

Precisión. Valor predictivo positivo que representa la proporción de predicciones correctas que contiene el modelo. El cálculo de precisión (*P*) está dado por:

$$P = \frac{verdaderos\ positivos + verdaderos\ negativos}{(verdaderos\ positivos + verdaderos\ negativos + falsos\ positivos + falsos\ negativos)} \quad (7)$$

Índice de verdaderos positivos. Presenta la proporción de predicciones correctas en predicciones de clase positiva. Su cálculo esta dado por:

$$Indice\ de\ verdaderos\ positivos = \frac{verdaderos\ positivos}{(verdaderos\ positivos + falsos\ negativos)} \quad (8)$$

Índice de falsos positivos. El índice de falsos positivos representa la proporción de predicciones incorrectas en clase positiva. El cálculo del índice de falsos positivos se calcula de la siguiente forma:

$$Indice\ de\ falsos\ positivos = \frac{falsos\ positivos}{(negativos\ verdaderos + falsos\ positivos)} \quad (9)$$

Exhaustividad ponderada. Es una medida de sensibilidad que determina la proporción de predicciones correctas en clase positiva. La exhaustividad está dada por:

$$Exhaustividad\ ponderada = \frac{verdaderos\ positivos}{(verdaderos\ positivos + falsos\ negativos)} \quad (10)$$

Precisión ponderada. Proporciona el promedio ponderado de precisión con ponderaciones iguales a la probabilidad de clase. La precisión ponderada está dada por:

$$Precisión\ ponderada = \frac{verdaderos\ positivos}{(verdaderos\ positivos + falsos\ negativos)} \quad (11)$$

Medida F1 ponderada. Es una medida de exactitud que proporciona el promedio absoluto de precisión y exhaustividad. La medida F1 ponderada está dada por:

$$Medida\ F1\ ponderada = 2 * \frac{(precisión * exhaustividad)}{(precisión + exhaustividad)} \quad (12)$$

La Tabla 2.3 muestra la evaluación que se hace de las redes neuronales estudiadas. Con respecto a la eficiencia objetiva de los modelos utilizados, se observa que la precisión más baja se obtuvo en los modelos con Valores Reales de la base de datos de México, con una precisión de 97 %; mientras que el modelo de Valores Reales de la base de datos de España tiene el mejor resultado con una precisión del 60.5%. Para esta investigación se ejecutó el algoritmo con valores de normalizados de las variables continuas, sin embargo, como se observa en la tabla 2.3, no hay variación significativa en el conjunto de medidas de evaluación de los modelos en comparación con las otras alternativas ejecutadas. Por lo tanto, varias neuronas no afectan de manera significativa a la precisión de los modelos de clasificación.

En relación con los índices de verdaderos positivos y de falsos negativos, los resultados son congruentes con la precisión de los modelos, ya que el modelo es más confiable, cuando el valor del índice de verdaderos positivos se aproxima a uno, y el índice de falsos negativos es cercano a cero. La puntuación F1, entre cero y uno, representa un promedio ponderado de los valores de precisión y de exhaustividad; en consecuencia, la precisión del modelo es mejor (peor) cuando F1=1 (F1=0).

Tabla 2.3. Medidas de evaluación de modelos. (Fuente: elaboración propia).

Medidas de evaluación de modelos	Formato de datos utilizado					
	España			México		
	Valores Reales	Continuos Normalizados	Valores Reales 2 Capas Ocultas	Valores Reales	Continuos Normalizados	Valores Reales 2 Capas Ocultas
Precisión	0.605	0.599	0.585	0.970	0.970	0.970
Índice de verdaderos positivos ponderados	0.605	0.599	0.585	0.970	0.970	0.970
Índice de falsos positivos ponderados	0.111	0.112	0.092	0.000	0.000	0.000
Precisión ponderada	0.766	0.764	0.797	1.000	1.000	1.000
Exhaustividad ponderada	0.605	0.599	0.585	0.970	0.970	0.970
Medida de F_1 ponderada	0.667	0.663	0.664	0.985	0.985	0.985

En relación con los resultados de los sesgos de la capa de salida, respecto a cada una de las clases de la variable dependiente, se obtienen los valores que

se presentan en la Tabla 2.4. Notemos que existe similitud con el sesgo de la muestra de los datos de entrada, ya que en las dos bases de datos que se estudian predomina un tipo de vivienda, en el caso de España el tipo de vivienda 2 y en el caso de México el tipo de vivienda 1, sin contar con registros del tipo de vivienda 4.

Tabla 2.4. Sesgo de la capa de salida. (Fuente: elaboración propia).

Sesgo del Tipo de Vivienda					
España			México		
Valores Reales 1 Capa Oculta	Continuos Normalizados 1 Capa Oculta	Valores Reales 2 Capas Ocultas	Valores Reales 1 Capa Oculta	Continuos Normalizados 1 Capa Oculta	Valores Reales 2 Capas Ocultas
TV2 0.7690	TV1 0.7300	TV1 0.4761	TV1 1.2093	TV1 1.2093	TV1 0.6727
TV1 0.6734	TV2 0.3153	TV2 0.4749	TV2 0.2218	TV2 0.2218	TV2 0.1618
TV4 0.0433	TV4 -0.0516	TV5 0.1936	TV3 0.0026	TV3 0.0026	TV3 0.1605
TV5 0.0277	TV5 -0.0566	TV6 0.0134	TV5 -0.6392	TV5 -0.6392	TV5 -0.0326
TV3 -0.3375	TV3 -0.5809	TV3 0.0075			
TV6 -0.7563	TV6 -1.1073	TV4 -0.1650			

En lo relativo a los pesos de las variables independientes (Tabla 2.5), los resultados de la red neuronal de España evidencian una menor variación del tipo de vivienda en todo el país, una explicación puede ser que en las áreas urbanas de España predomina la vivienda vertical, incluso en ciudades menores a 100,000 habitantes. Por esta razón se considera que la variable de Ubicación Geográfica (UG) tiene una importancia menor cuando se compara el estado de la vivienda en México, donde es necesario regionalizar la muestra de los datos para que UG permita al algoritmo identificar una ponderación de variables más coherente respecto de las preferencias de los usuarios. En los resultados de España, la variable Número de Cuartos, que corresponde al número total de habitaciones de la vivienda, incluyendo Cocina (C) y Cuartos de Dormitorio (CD), en el análisis de la base de datos se encontró incoherencia en los datos, por ejemplo, en algunos registros, el número de Cuartos Dormitorio es mayor al número total de cuartos de la vivienda representado por la variable NC. No obstante, se depuraron estos registros y el comportamiento de las medidas de evaluación de los modelos presentaban cambios irrelevantes.

Tabla 2.5. Ponderación de la importancia de las variables. (Fuente: elaboración propia).

Impotancia de las Variables					
España			México		
Valores Reales 1 Capa Oculta	Continuos Normalizados 1 Capa Oculta	Valores Reales 2 Capas Ocultas	Valores Reales 1 Capa Oculta	Continuos Normalizados 1 Capa Oculta	Valores Reales 2 Capas Ocultas
NC 21.51	NC 21.72	NC 23.37	UG 26.42	UG 26.42	UG 25.89
TL 18.16	TL 20.11	TL 20.73	BC 19.29	BC 19.29	CD 19.64
BC 15.99	BC 17.37	BC 16.50	TL 15.41	TL 15.41	TR 14.85
UG 14.03	UG 16.31	UG 13.76	CD 11.13	CD 11.13	BC 12.41
CD 12.52	CD 10.16	CD 12.11	NC 10.52	NC 10.52	NC 10.22
TR 9.17	TR 5.21	C 5.48	TR 9.51	TR 9.51	TL 6.84
C 4.51	C 4.57	TR 4.07	T 5.97	T 5.97	C 5.72
T 4.11	T 4.55	T 3.97	C 1.75	C 1.75	T 4.44

2.3.6 Discusión

El modelo presentado permite establecer el proceso de decisión de selección de “vivienda adecuada” a través de la identificación de las necesidades de cada usuario, es decir si se incluye un nuevo registro con las características de vivienda que requiere el usuario el modelo determina el tipo de vivienda adecuada para éste. El modelo de RNA de la base de datos de México resulto más eficiente que la red neuronal de la base de datos de España ya que se identificó incongruencia en algunos registros, por ejemplo, el número cuartos dormitorio excedía al número de cuartos de toda la vivienda. Sin embargo, al probar con nuevos registros la respuesta de la red asigna un tipo de vivienda congruente, por este motivo se considera que el modelo permite seleccionar vivienda adecuada para nuevos usuarios de España y México.

2.4 Conclusiones

Se sabe que las redes neuronales contemplan una clase de funciones capaces de adaptarse incluso a asignaciones aleatorias de entrada-salida con una precisión superior a otros métodos. En este trabajo presentamos propiedades de redes neuronales que complementan este aspecto de la expresividad, ya que la percepción recabada en las bases de datos en los datos de entrada tiene un comportamiento similar en los datos de salida. Al usar herramientas del análisis destacamos un sesgo de aprendizaje de las redes profundas hacia funciones

como la función de activación softmax que permite realizar una clasificación precisa de los tipos de vivienda, aunque los datos no sean coherentes como en el caso de la base de datos de España. Intuitivamente, esta propiedad está en línea con la observación de los resultados de las particiones de prueba y validación que presentan el mismo comportamiento de la partición de datos de entrenamiento. No obstante, en la literatura se recomienda utilizar con cautela la función de activación softmax, ya que predispone que cada registro es miembro de una sola clase, sin embargo, en este caso se comparó su eficiencia al comparar esta función de activación recomendadas para la capa de salida como las funciones de identidad y tangente hiperbólica.

Aunque algunos autores mencionan que los resultados de las variables de salida no se pueden interpretar de la misma manera que en los en métodos estadísticos, con los resultados de esta investigación es posible afirmar que el comportamiento de los sesgos y de las ponderaciones sinápticas de la capa de salida tienen un comportamiento similar al comportamiento estadístico. Para aplicaciones similares, incluso para esta misma investigación, se recomienda indagar sobre el número de clases de la variable dependiente dado que la falta de coherencia de la base de datos de España no se considera la causa de la baja precisión del modelo ya que se omitieron los registros incoherentes y los resultados presentaron cambios irrelevantes.

En relación con los resultados obtenidos esta investigación representa la primera etapa para el diseño de un modelo de selección de vivienda que contemple un mayor número de variables bajo la óptica de selección de vivienda en lugar del enfoque de asignación de vivienda que es el más utilizado. Por otra parte, se plantea como investigación futura probar el modelo en campo para determinar su verdadera eficiencia además de afinar la selección de variables a incluir.

2.5 Referencias

1. ACNUDH - ONU Habitat. (2010). El derecho a una vivienda adecuada. Folleto informativo nº21. *Revista de Antropología Social*, *19*, 103–129. https://doi.org/-
2. Anand, P. (2017). Bias in machine learning. Recuperado de https://iq.opengenus.org/bias-machine-learning/
3. Choy, L. H. T., Ho, W. K. O., & Mak, S. W. K. (2012). Housing attributes and Hong Kong real estate prices: A quantile regression analysis. *Construction Management and Economics*, *30*(5), 359–366. https://doi.org/10.1080/01446193.2012.677542
4. Coloma, J. F., Valverde, L. R., & García, M. (2019). Estimation of rustic housing construction costs through Artificial Neural Networks | Estimación de los costes de construcción de viviendas rústicas mediante Redes Neuronales Artificiales. *Informes de la Construccion*. https://doi.org/10.3989/ic.62206
5. CONEVAL. (2018). *Estudio Diagnóstico del Derecho a la Vivienda Digna y Decorosa 2018*. Ciudad de México. Recuperado de https://www.coneval.org.mx/Evaluacion/IEPSM/Documents/Derechos_Sociales/Estudio_Diag_Vivienda_2018.pdf
6. Du, K. L., & Swamy, M. N. S. (2014). *Neural networks and statistical learning*. *Neural Networks and Statistical Learning* (Vol. 9781447155713). Springer-Verlag London Ltd. https://doi.org/10.1007/978-1-4471-5571-3
7. Eichie, J. O., Oyedum, O. D., Ajewole, M. O., & Aibinu, A. M. (2017). Comparative analysis of basic models and artificial neural network based model for path loss prediction. *Progress In Electromagnetics Research M*, *61*, 133–146. https://doi.org/10.2528/PIERM17060601
8. ElKessab, B., Daoui, C., Boukhalene, B., & Salouan, R. (2014). A Comparative Study between the K-Nearest Neighbors and the Multi-Layer Perceptron for Cursive Handwritten Arabic Numerals Recognition. *International Journal of Computer Applications*, *107*(21), 25–30. https://doi.org/10.5120/19140-0117
9. Ferreira, R. P., Martiniano, A., Napolitano, D., Romero, M., De Oliveira Gatto, D. D., Farias, E. B. P., & Sassi, R. J. (2018). Artificial Neural Network for Websites Classification with Phishing Characteristics. *Social Networking*, *07*(02), 97–109. https://doi.org/10.4236/sn.2018.72008
10. Gaspareniene, L., Venclauskiene, D., & Remeikiene, R. (2014). Critical Review of Selected Housing Market Models Concerning the Factors that Make Influence on Housing Price Level Formation in the Countries with Transition Economy. *Procedia - Social and Behavioral Sciences*. https://doi.org/10.1016/j.sbspro.2013.12.886
11. Han, J., Kamber, M., & Pei, J. (2012). 9.2 Classification by backpropagation. En *Data Mining: Concepts and Techniques* (p. 398). Morgan Kaufmann Publishers. https://doi.org/10.1016/C2009-0-61819-5
12. Hassanudin, M. T. T., & K., C. S. (2016). Prioritisation of key attributes influencing the decision to purchase a residential property in Malaysia: An analytic hierarchy process (AHP) approach. *International Journal of Housing Markets and Analysis*, *9*(4), 446–467.

https://doi.org/10.1108/IJHMA-09-2015-0052
13. INE. (2017). Encuesta Continua de Hogares. Recuperado de https://www.ine.es/CDINEbase/consultar.do?mes=&operacion=Encuesta+continua+de+hogares&id_oper=Ir
14. INEGI. (2017). Encuesta Nacional de los Hogares 2017. Recuperado de https://www.inegi.org.mx/programas/enh/2017/
15. Llinás Solano, H., Arteta Charris, M., & Tilano Hernández, J. (2016). El modelo de regresión logística para el caso en que la variable de respuesta puede asumir uno de tres niveles: Estimaciones, pruebas de hipótesis y selección de modelos. *Revista de Matemática: Teoría y Aplicaciones*. https://doi.org/10.15517/rmta.v23i1.22442
16. Marín Diazaraque, J. M. (2007). Introducción a las redes neuronales aplicadas. *Manual Data Mining*, 1–31.
17. ONU-Habitat. (2019). Elementos de una Vivienda Adecuada. Recuperado el 16 de mayo de 2020, de https://onuhabitat.org.mx/index.php/elementos-de-una-vivienda-adecuada
18. Picón, C. (2011). ¿Son más corruptos los países menos abiertos a los mercados internacionales? Aplicación de un modelo predictivo de clasificación basado en Redes Neuronales. *Economía del Caribe*. https://doi.org/10.14482/rec.v0i8.3262
19. Pinkus, A. (1999). Approximation theory of the MLP model in neural networks. *Acta Numerica*. https://doi.org/10.1017/S0962492900002919
20. Rahadi, R. A., Wiryono, S. K., Koesrindartoto, D. P., & Syawmil, I. B. (2018). External Factors Influencing Housing Product Price in Jakarta Metropolitan Region. *fatores que influencia o preço*, *7*(4), 179–192.
21. Samad, D., Zainon, N., Rahim, F. A. M., & Lou, E. (2017). Malaysian affordability housing policies revisited. *Open House International*. https://doi.org/10.1051/matecconf/20166600010
22. Vanus, J., Fiedorova, K., Kubicek, J., Gorjani, O. M., & Augustynek, M. (2020). Wavelet-based filtration procedure for denoising the predicted CO2 waveforms in smart home within the internet of things. *Sensors (Switzerland)*. https://doi.org/10.3390/s20030620
23. Yao, Y., Zhang, J., Hong, Y., Liang, H., & He, J. (2018). Mapping fine-scale urban housing prices by fusing remotely sensed imagery and social media data. *Transactions in GIS*, *22*(2), 561–581. https://doi.org/10.1111/tgis.12330

Capítulo 3

Estrategia para la selección de vivienda en México a través de redes neuronales

Resumen

Este estudio compara la clasificación de los tipos de vivienda en diferentes regiones de México mediante el uso de redes neuronales artificiales entrenadas para la selección del tipo de vivienda adecuada de nuevos usuarios. Se utilizan dos esquemas de división regional económica del país con el propósito de analizar el sesgo de los tipos de vivienda por su difusión y su relación con los tipos de vivienda de mayor importancia en cada región de acuerdo con los resultados de las redes neuronales. El sesgo de los datos y las ponderaciones de los tipos de vivienda resultantes de la clasificación en las redes neuronales de cada región no influyen en los índices de precisión de la red neuronal para la clasificación de vivienda; así, la clasificación que se hace de las viviendas es independiente de los sesgos y revela las preferencias de los usuarios respecto a las características de vivienda y los tipos de vivienda en cada región. La robustez de los resultados se confirma comparando los dos esquemas de división regional económica para identificar la similitud de las características de vivienda que existen entre las comunidades que integran cada región para mejorar la estrategia de selección en los usuarios.

Palabras clave: Adequate Housing, Cultural Adaptation, Decision Making, ANN.

3.1 Introducción

La toma de decisiones (TD) orientadas a cubrir las necesidades de vivienda representa la base fundamental para el desarrollo ordenado de las comunidades, sin embargo, es una tarea compleja que debe integrar los distintos enfoques que

definen los criterios de vivienda adecuada. La falta de información sobre el comportamiento, el nivel de calidad de vida y la satisfacción lograda por los usuarios que viven en una vivienda, resaltan la necesidad de implementar mecanismos de medición que validen esta situación y sirvan como rayos X de las condiciones de habitabilidad en las que viven. El reto consiste en generar “viviendas adecuadas” de manera exitosa [1]. En torno a lo cual, existen diferentes enfoques. Por ejemplo, el Consejo Nacional de Evaluación de la Política de Desarrollo Social de México (CONEVAL) [2] considera la asequibilidad como un criterio fundamental para determinar si una vivienda es adecuada o no. Aunque la asequibilidad es esencial para el correcto desarrollo urbano, con frecuencia, las regulaciones normativas y/o su aplicación concreta, por parte de los gestores, ignora las preferencias de los individuos, muchas de ellas motivadas desde el punto de vista económico, social, cultural, e incluso por una elemental demanda de adecuación a las necesidades de una movilidad sostenible. Por tanto, la satisfacción de los usuarios no está garantizada. La Oficina del Alto Comisionado para los Derechos Humanos ACNUDH [3], establece que una “vivienda adecuada” debe cumplir con los siguientes criterios de 1. Seguridad de la tenencia, 2. Disponibilidad de servicios, 3. Asequibilidad, 4. Habitabilidad, 5. Accesibilidad, 6. Ubicación y 7. Adecuación Cultural.

En principio, la lógica de asignación de “viviendas adecuadas” supone que los gobiernos diseñan políticas públicas de construcción que garanticen a las empresas la recuperación de su inversión, así mismo se trata de garantizar un nivel de satisfacción alto a los usuarios potenciales de las viviendas [1]. En otras palabras, la toma de decisiones (TD) orientada a cubrir las necesidades de vivienda es fundamental para el desarrollo eficiente del sector y de las comunidades.

En México, de acuerdo con Villavicencio y Duran [4] no se conocen las aspiraciones y los mejoras que desean las familias mexicanas en relación al tipo de vivienda, factores que no parecen estar contemplados en la actual oferta de vivienda de tipo social. Por lo anterior en México la mayor parte de la vivienda social se ha realizado mediante la autoproducción y sin asistencia técnica [5].

La mayoría de los estudios sobre “vivienda adecuada” se centran en la asequibilidad de la vivienda. La literatura se enfoca en determinar los factores que influyen en la formación del nivel de precios de la vivienda [6] dejando en

segundo plano la satisfacción de los usuarios [7]. En ésta línea, Choy et al. [8] analizan la influencia de las prioridades, expresadas por los usuarios, en el precio de la vivienda, mediante una regresión cuantílica. Con respecto a la predicción precios, Choy et al. [8] and Yao et al. [9] lo hacen por medio de redes neuronales; en ambos casos proporcionan evidencia de que la cercanía con estaciones de transporte y/o lugares turísticas incrementa los precios y genera desigualdad entre la población.

En general, ONU-Habitat [7] señala que tanto investigadores como creadores política vivienda han dejado de lado la adecuación cultural y se han preocupado por analizar este mercado, principalmente, como motor de crecimiento económico y de creación de empleo, eludiendo otras implicaciones que afectan directamente a la sostenibilidad urbana. Por tanto se evidencia la omisión de la adecuación a características de orden cultural del usuario, en el desarrollo de los planes de promoción de vivienda [7]. El resultado es que con frecuencia, se genera un modelo de ciudad poco habitable, caracterizada por la degradación urbana y la segregación especial; donde la nota dominante es la dispersión de las comunidades y la infrautilización de zonas comunes [10].

La presente investigación propone un modelo de selección de vivienda basado en la clasificación de las viviendas de cada una de las regiones económicas de México, mediante redes neuronales artificiales (RNA). Este modelo relaciona los criterios de vivienda adecuada con las variables vivienda documentadas en la Encuesta Nacional de los Hogares (ENH) del Instituto Nacional de Geografía Estadística e Informática (INEGI) de México [11]. Como se mostrará, el modelo propuesto permite prever el tipo de vivienda que se adapte a las necesidades de los nuevos adquisidores por medio de la identificación de los atributos que caracterizan cada territorio.

La RNA permite ponderar la importancia relativa de las variables sobre el tipo de vivienda que los usuarios prefieren, y determinar los sesgos de la última capa oculta para las neuronas de salida de las redes de los dos esquemas de división regional económica de México. Para este propósito, el algoritmo utiliza la función tangente hiperbólica para el entrenamiento de las capas ocultas pues se trata de un problema de clasificación; mientras que la función de activación de la capa de salida es de tipo softmax, que se usa para la regresión logística multiclase [12], o regresión logística multinomial, adecuado a los siete criterios de “vivienda

adecuada” establecidos por ONU-Habitat. Mediante la comparación de las RNA con datos de los dos esquemas, se obtuvieron predicciones sobre el tipo de vivienda conveniente para nuevos registros de prueba en cada región.

Como podrá observarse, los resultados muestran sesgo sobre los tipos de vivienda de mayor frecuencia en contraste con los tipos de vivienda de mayor importancia en cada región, lo cual permite identificar cual esquema de división regional económica es más apropiado para determinar la similitud de los tipos de vivienda entre los estados que constituyen cada región.

3.2 Metodología

3.2.1 Redes Neuronales Artificiales

Las RNA son una rama de la inteligencia artificial que permite el estudio de problemas multifactoriales por medio de la que ejecución de funciones que analizan un número fijo de entradas y producen un número fijo de salidas [13]. Son modelos matemáticos programables cuya integración computacional se hace en paralelo, es decir, se constituyen en capas con múltiples conexiones, emulando la conexión de neuronas del cerebro humano [14]. La mayoría de los tipos de arquitectura de las RNA incluye una capa de neuronas ocultas, las cuales son el puente entre la entrada y la salida. Las conexiones entre las neuronas tienen un valor de ponderación asociado (peso), además del sesgo (umbral). Este último es un parámetro adicional que ayuda a ajustar al modelo a los datos dados controlando la función de activación [15]. Entonces, los pesos y umbrales fijan los valores de salida del conjunto de valores de entrada.

La red neuronal sigue un proceso de entrenamiento para determinar los pesos y umbrales de mejor ajuste, ver Figura 3.1. La neurona de sesgo se agrega al inicio / final de la entrada y a cada capa oculta, es decir, no está influenciada por los valores de la capa anterior, por lo que estas neuronas no tienen conexiones entrantes. Matemáticamente, el proceso de entrenamiento se expresa de la siguiente manera:

$$Salida = \Sigma(\, entradas * ponderaciones) + sesgo \qquad (1)$$

Según el tipo de entrenamiento [16], las redes se pueden clasificar en dos grupos:

- **RNA con entrenamiento supervisado.** Se entrenan presentando, para cada conjunto de entradas, las salidas que se esperan ellas produzcan. Entonces, los algoritmos de entrenamiento calculan pesos y sesgos nuevos con el objetivo de minimizar el error entre la salida deseada y la resultante.
- **RNA sin entrenamiento supervisado.** Los algoritmos de entrenamiento calculan nuevos pesos de manera aleatoria. Se utilizan como clasificadores y se caracterizan por asociar una combinación de entradas específicas con una sola salida; en otras palabras, identifica similitudes en los datos y reacciona con base en la presencia o ausencia de los elementos comunes de cada registro de datos.

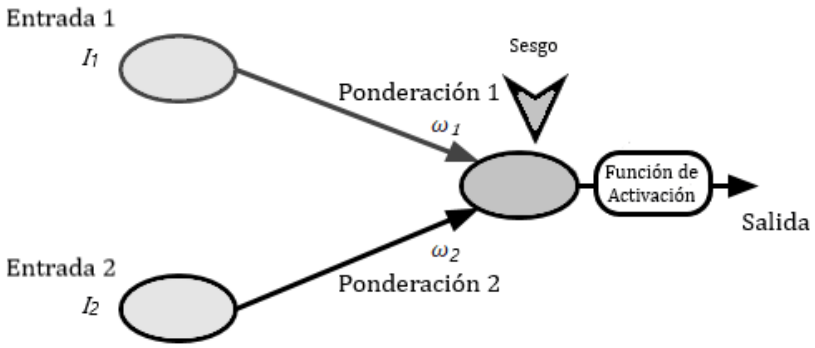

Figura 3.1. Proceso de entrenamiento de una RNA. (Fuente: Elaboración propia con información de Du & Swamy (2014), *Neural networks and statistical learning*. *Neural Networks and Statistical Learning* , pag. 5).

Es importante mencionar que la selección de la función de activación depende del objetivo que se persiga. Por ejemplo, los modelos de clasificación comúnmente utilizan la función de activación sigmoide, mientras que los modelos predictivos asumen una función de activación lineal [17]. Las funciones de activación más comunes en RNA son las siguientes:

- **Función Heaviside.** También conocida como función umbral o función escalón unitario, es una función discontinua cuyo valor es uno para valores positivos y cero para valores negativos.

$$\gamma_{(c)} = \begin{cases} 1 \; si \; c \geq 0 \\ 0 \; si \; c < 0 \end{cases} \tag{2}$$

- **Función sigmoide logístico.** Es una curva en forma de S que se basa en una matriz de ganancia de información. Dicha matriz sugiere una transición de valores bajos hacia valores altas a través de un punto de inflexión.

$$\gamma_{(c)} = \frac{1}{1+e^{-c}} \tag{3}$$

- **Función softmax (exponencial normalizada).** Es de tipo sigmoide y generaliza la distribución de Bernoulli para problemas multiclase, se aplica en la última capa de una red neuronal, para realizar la clasificación de distintos elementos ya que escala las entradas previas, en un rango entre 0 y 1, y normaliza la capa de salida.

$$\gamma_{(c)_j} = \frac{e^{c_j}}{\sum_{k=1}^{K} e^{c_k}} \tag{4}$$

- **Función tangente hiperbólica.** Es una curva con propiedades similares a la activación sigmoide, por lo que se le considera su alternativa. Las salidas están acotadas entre -1 y 1.

$$\gamma_{(c)} = tanh(c)\frac{e^{c}-e^{-c}}{e^{c}+e^{-c}} \tag{5}$$

3.2.2 Perceptrón Multicapa (MLP)

El modelo de perceptrón de alimentación multicapa (MLP) es una RNA que consiste en un número finito de capas sucesivas, a su vez, cada una posee un número finito de neuronas, las cuales se conectan entre sí en la siguiente capa, como si hicieran sinapsis. De este modo, la información fluye en una sola dirección, de una capa a la siguiente (feedforward). La arquitectura de una red MLP (Figura 3.2) se conforma por la primera capa o capa de entrada. Posteriormente se encuentran las capas intermedias, u capas ocultas, y por último se ubica la capa de salida, cuyas neuronas también son llamadas nodos de respuesta [17][18].

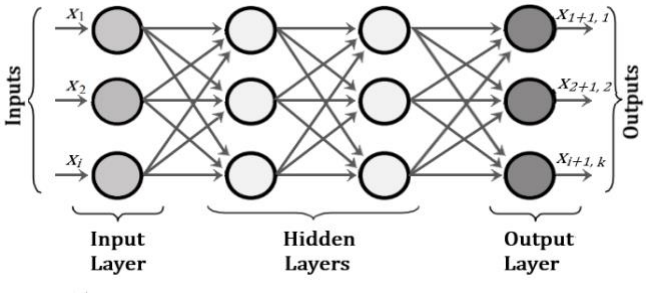

Figura 3.1. Arquitectura de la RNA MLP. (Fuente: Elaboración propia con información de ElKessab et al., (2014), International Journal of Computer Applications, 107 (21), 25–30).

Como indica la expresión (1), las neuronas de la primera capa oculta están en función de la suma ponderada de las neuronas de la capa de entrada. Dicha función se denomina función de activación, y los valores de las ponderaciones se determinan mediante el algoritmo de estimación, el cual busca minimizar los errores. Si la red contiene una segunda capa oculta, sus neuronas son una transformación de la suma ponderada de las neuronas de la primera capa oculta. El proceso anterior se repite de manera iterativa para cada capa oculta hasta llegar a la capa de salida.

Los modelos MLP se clasifican de acuerdo con el número de capas ocultas que poseen, y no por su número total de capas, es decir, se clasifican por el número de capas excluyendo la capa de entrada y la de salida. El término MLP se aplica genéricamente a todos los modelos con al menos una capa oculta. Las normas y reglamentos que rigen el modelo MLP [17] son las siguientes:

- La capa de entrada tiene como salida de su neurona *j* el valor x_{0j}.
- La neurona *k* de la capa *i* recibe la salida x_{ij} de cada neurona *j* de la capa *(i - 1)*. Los valores x_{ij} se multiplican por las ponderaciones sinápticas ω_{ijk}, y los productos se suman para obtener el siguiente valor de salida. Matemáticamente, para cada neurona *k*, la capa *i + 1* se construye de la siguiente manera:

$$x_{i+1,k} = \gamma\left(\sum_j \omega_{ijk} x_{ij} - \theta_{ik}\right) \quad (6)$$

Donde γ es la función de activación, θ_{ik} son los umbrales o sesgos y ω_{ijk} son las ponderaciones sinápticas.

Los pesos y umbrales se determinan durante el aprendizaje o entrenamiento por medio de algún mecanismo definido exógenamente; por ejemplo, el algoritmo de

retropropagación (backpropagation) es de los más utilizados [17]. El termino retropropagación se refiere a la forma en que se propaga la información hacia atrás, iniciando desde la capa de salida; esto permite que las ponderaciones sinápticas de las neuronas ubicadas en las capas ocultas cambien durante el entrenamiento. La variación de las ponderaciones sinápticas se debe a la sensibilidad de la función activación. La misma función de activación se utiliza en cada una de las capas ocultas; sin embargo, la elección de la función de activación en la capa de salida depende del tipo de tarea impuesto [19]. Los resultados en cada vector de entrada se utilizan para la formación y validación mediante múltiples iteraciones en las que el modelo identifica las ponderaciones y sesgos que mejor ajusten la red al comportamiento de los datos. Es importante señalar que el resultado es un aprendizaje automático, que se refiere al proceso por el que el algoritmo toma decisiones inteligentes mediante el reconocimiento de patrones, aprendizaje, basados en los datos de la muestra. En otras palabras, se establece un patrón a seguir entre el problema y la respuesta [20].

3.3 Modelo

La construcción del modelo de clasificación y predicción del tipo de vivienda, para las regiones de México, sigue los siguientes pasos:

- Selección de variables y preprocesamiento de la base de datos.
- Diseño del modelo de clasificación (utilizando IBM SPSS Modeler).
- Evaluación de los resultados obtenidos.

3.3.1 Selección de variables y preprocesamiento de la base de datos

En el proceso de selección de variables se consideran las características con las que cuenta una vivienda nueva en México vivienda así como las características añadidas por los usuarios para adecuar la vivienda a sus necesidades. En México, el INEGI que a partir de 2014 y hasta 2017 realizó anualmente la ENH [11]. La encuesta se divide en tres apartados: Vivienda, Hogar y Persona. Su objetivo es dar a conocer las características de las viviendas seleccionadas; datos sociodemográficos acerca de los integrantes del hogar, su ocupación, educación y su percepción de su estado de salud; así como la disponibilidad de bienes y servicios de tecnologías de la información y las comunicaciones en los

hogares. Para el modelo que se presenta solo se utiliza exclusivamente el apartado correspondiente a vivienda con el propósito de identificar las variables que se relacionan con los criterios de vivienda adecuada.

Para el análisis se utilizan dos esquemas de división regional económica de México y de esta manera conocer el papel que puede desempeñar la geografía en la determinación del sesgo y la importancia de los tipos de vivienda en cada una de las regiones en función de los resultados de las redes neuronales. El esquema 1 (Figura 3.3) es un conjunto de 8 regiones económicas de México que el gobierno implementó en la década de los años 70's con el fin de mejorar las relaciones políticas, sociales y económicas de los estados vecinos entre sí. Esta división se ha utilizado para implementar distintas medidas [21]. El esquema 2 (Figura 3.4) lo organizó Esquivel [22] en 7 regiones económicas basándose en las teorías de Hall y Jones (1996) y Gallup y Sachs (1999) que sugieren que los estados localizados en latitudes altas tienden a tener, en forma consistente, un mayor nivel del ingreso per cápita por efectos que parecen tener las características geográficas en la economía. Aunque los resultados de su investigación son consistentes no se puede omitir la influencia de la cercanía con el mercado estadounidense en los resultados de los estados fronterizos de México. Este esquema considera el clima para agrupar las regiones, factor que influye en el tipo de vivienda de las regiones [23].

Figura 3.3. Esquema de división regional económica 1. (Fuente: Elaboración propia con información de Fouquet, (2008), Sociedad, desarrollo y ciudadanía en México, 229–250).

Figura 3.4. Esquema de división regional económica 2. (Fuente: Elaboración propia con información de Esquivel, (2008), Geografía y Desarrollo Económico en México. Banco Interamericano de. Desarrollo.

Las bases de datos seleccionadas de la ENH corresponden a el periodo 2014 a 2017 que son las únicas 4 ediciones que se han realizado. Posteriormente, se procedió a depurar los registros de vivienda con errores o ausencia de información para homogeneizar y unificar las bases de datos. Los registros de vivienda se desglosan en la Tabla 3.1 de acuerdo con los dos esquemas de división regional económica que se utilizan.

Tabla 3.1. Divisiones regionales económicas de México.
(Fuente: elaboración propia a parir de la ENH, INEGI (2017)) .

Divisiones Regionales Económicas			
Esquema 1		**Esquema 2**	
Región	**Registros**	**Región**	**Registros**
Centro	10595	Capital	7154
Centro Norte	19883	Altiplano	23943
Occidente	14782	Pacífico	19496
Oriente	16035	Este	15898
Sureste	16499	Golfo	20077
Suroccidente	11855	Sur	15959
Noreste	12559	Norte	24931
Noroeste	25250		

En la Tabla 3.2 se muestra el Cuestionario Básico de la ENH 2017 [20], mediante el cual se realiza la consulta sobre las variables asociadas a las características de vivienda, la ubicación y la conformación de los hogares en México. En la ENH se consideran 109 en cada edición. Para la selección de variables se utilizan los siguientes criterios:

- Se incluyen variables afines a los criterios de vivienda adecuada.
- Se descartan las variables que tienen datos idénticos en todos los registros.
- Se descartan las variables mal registradas.

- Se unifican las variables que tengan diferente formato en ediciones previas.

De acuerdo con los criterios de selección de variables, en la Tabla 3.2 se resaltan en color rojo las variables que se descartan y en color verde las variables que se unifican.

Tabla 3.2. Información de las variables de vivienda.
(Fuente: elaboración propia a parir de la ENH, INEGI (2017)).

Cuestionario Básico de la ENH					
#	Variable	Consulta	#	Variable	Consulta
1	folioviv	Identificador de la vivienda	56	const_dorm	Construir dormitorio
2	fipo_viv	Tipo de vivienda	57	const_coci	Construir cocina
3	condominio	Núm. de pisos del condominio	58	const_bano	Construir baño
4	elevador	Disponibilidad de elevador	59	const_neg	Construir negocio
5	mat_pared	Material de paredes	60	comun1	Espacio para sala
6	mat_techos	Material de techos	61	comun2	Espacio para jardín
7	mat_pisos	Material de pisos	62	comun3	Espacio para patio
8	ais_techos	Aislamiento en techo	63	comun4	Espacio para cuarto de lavado
9	ais_pared	Aislamiento en paredes	64	comun5	Espacio para cuarto de televisión
10	ais_ventan	Aislamiento en ventanas	65	comun6	Espacio para cuarto de estudio
11	ais_otro	Otro tipo de aislamiento	66	comun7	Espacio para cuarto de juegos
12	antiguedad	Antigüedad de la vivienda	67	comun8	Espacio para cuarto de ejercicios
13	cocina	Tiene cocina	68	comun9	Espacio para cochera
14	cocina_dor	Utiliza cocina de dormitorio	69	estaciona	Cajones estacionamiento
15	cuart_ dorm	Cuartos dormitorio	70	oomun10	Área común con otras viviendas
16	num_cuarto	Número de cuartos	71	oomun11	Salón de eventos área común
17	disp_agua	Disponibilidad de agua	72	oomun12	Pista para caminar área común
18	dotac_agua	Dotación de agua	73	oomun13	Gimnasio área común
19	excusado	Tiene excusado	74	comun14	Zona de juegos área común
20	uso_compar	Uso compartido del sanitario	75	comun15	Canchas deportiva área común
21	sanit_agua	Sanitario conexión agua	76	comun16	Alberca área común
22	bano_comp	Sanitario excusado regadera	77	comun17	Otra área en común
23	bano_excus	Sanitario excusado	78	tenencia	Tipo de tenencia de la vivienda
24	bano_regad	Sanitario regadera	79	pago_renta	Pago de renta de vivienda
25	drenaje	Destino de drenaje	80	anio_res	Años residiendo en la vivienda
26	disp_elec	Disponibilidad eléctrica	81	mes_res	Meses residiendo en la vivienda
27	anio_panel	Año panel solar	82	familiar	Parentesco con dueño de la vivienda
28	panel_ne	Desconoce año de adquisición	83	tipo_ adqui	Adquisición de la vivienda
29	pot_panel	Conocimiento en potencia inst.	84	financia_1	Recursos propios
30	potencia	Potencia instalada	85	financia_2	Apoyo de FONHAPO[1]
31	focos_inca	Núm. de focos incandescente	86	financia_3	Crédito INFONAVIT[2] o FOVISSSTE[3]
32	focos_ahor	Número de focos ahorradores	87	financia_4	Crédito bancario
33	combustible	Tipo de combustible	88	financia_5	Crédito microfinanciera
34	estufa_chi	Estufa con Chimenea	89	financia_6	Crédito caja de ahorro
35	eli_basura	Eliminación de basura	90	financia_7	Crédito de otra institución
36	lavadero	Dispone de lavadero	91	financia_8	Préstamo familiar
37	fregadero	Dispone de fregadero	92	num_dueno1	Identificador del primer dueño
38	regadera	Dispone de regadera	93	hog_dueno1	Hogar del primer dueño
39	rega_ elect	Dispone de regadera eléctrica	94	num_dueno2	Identificador del segundo dueño
40	tinaco_azo	Dispone de tinaco	95	hog_dueno2	Hogar del segundo dueño
41	cisterna	Dispone de cisterna	96	escrituras	Escrituras de la vivienda
42	pileta	Dispone de pileta o tanque	97	computador	Disponibilidad de computadora
43	calent_sol	Disp. de calentador solar de agua	98	tel_fijo	Disponibilidad de línea telefónica fija
44	calent_gas	Disp. de calentador a gas de agua	99	celular	Disponibilidad de teléfono celular

45	medidor_luz	Dispone de medidor de luz	100	internet	Disponibilidad de internet
46	bomba_agua	Dispone de bomba de agua	101	tv_paga	Servicio de televisión de paga
47	tanque_gas	Disp de tanque gas estacionario	102	tol_resid	Total de residentes
48	aire_acond	Dispone de aire acondicionado	103	tot_hog	Total de hogares en la vivienda
49	calefacc	Dispone de calefacción	104	ubica_geo	Ubicación geográfica
50	chimenea	Dispone de chimenea	105	ageb	Área geoestadística básica
51	repar_pard	Reparar las paredes	106	tam_loc	Tamaño de localidad
52	repar_tech	Reparar el techo	107	est_socio	Estrato socioeconómico
53	repar_agua	Reparar las tuberías del agua	108	esl_dis	Estrato del diseño muestral
54	repar_dren	Reparar las tuberías del drenaje	109	upm	Unidad primaria de muestreo
55	repar_cabl	Reparar el cableado eléctrico	110	factor	Factor de expansión

1 Fideicomiso Fondo Nacional de Habitaciones populares.

2 Instituto de Fondo Nacional de la Vivienda para los Trabajadores.

3 Fondo de la Vivienda del Instituto de Seguridad y Servicios Sociales de los Trabajadores del Estado.

Con el propósito de identificar los criterios de vivienda adecuada que establece la ACNUDH (2010) con las variables que se utilizan en el presente estudio, en la Tabla 3.3 se muestra la clasificación de la afinidad de las variables seleccionadas con las definiciones de cada uno de los criterios.

Tabla 3.3. Clasificación de variables conforme a los criterios ACNUDH.
(Fuente: Elaboración propia con información de la ENH, INEGI (2017), ACNUDH (2010)).

Clasificación de variables		
Criterio	**Definición**	**Variables afines**
Seguridad de la tenencia	Garantizar a los usuarios protección jurídica contra el desalojo forzoso, el hostigamiento y otras amenazas.	Escrituras
Asequibilidad	Que su costo no sea una amenaza o comprometa el disfrute de otros derechos humanos de los usuarios.	Tipo de Financiamiento
Disponibilidad de servicios, materiales, instalaciones e infraestructura	El usuario de la vivienda debe tener acceso al servicio de agua potable, instalaciones sanitarias adecuadas, energía para la cocción, la calefacción y el alumbrado.	Agua Drenaje Electricidad Eliminación de Basura Lavadero Tinaco azotea Calentador de agua Aire acondicionado Calefacción
Habitabilidad	Garantizar seguridad física, protección contra el frío, la humedad, el calor, la lluvia, el viento u otros riesgos para la salud, así como evitar problemas de hacinamiento proporcionando el espacio adecuado en relación con el número de usuarios por vivienda.	Material de paredes Material de techos Material de pisos Cocina Cuartos dormitorio Baño completo Baño excusado Baño regadera Reparación paredes Reparación techos Reparación de tuberías agua Reparación de tuberías drenaje Reparación paredes Reparación cableado
Accesibilidad	Que considere las necesidades específicas de los grupos desfavorecidos y marginados.	Tenencia Estrato socioeconómico
Ubicación	Que ofrezca acceso a oportunidades de empleo,	Ubicación geográfica Tamaño de localidad

	servicios de salud, escuelas, guarderías y otros servicios e instalaciones sociales, además, que no esté ubicada en zonas contaminadas o peligrosas.	
Adecuación cultural	Implica que las características de diseño permitan que se tome en cuenta y se respete, la expresión de la identidad cultural, en función de su dimensión étnica, regional o urbana.	Tipo de vivienda Número de cuartos Construir dormitorio Construir cocina Construir baño Total de residentes Total de hogares

Una vez clasificadas las variables se especifican valores de medición para determinar su tipo (dicotómicas, nominales y continuas) y el rol que cumplirán dentro del modelo, el cual tiene como propósito clasificar el tipo de vivienda mediante el uso de RNA's.

Tabla 3.4. Valores de medición de variables.
(Fuente: Elaboración propia con información de la ENH, INEGI (2017)).

Valores de Medición de Variables				
#	**Variable**	**Valores y Etiquetas**	**Rol**	**Tipo**
1	folioviv	Identificador de la vivienda	ID registro	Nominal
2	tipo_viv	1 Casa independiente 2 Departamento en condominio vertical 3 Vivienda en vecindad 4 Vivienda en cuarto de azotea 5 Local no construido para habitación	Dependiente	Nominal
3	mat_pared	1 Material de desecho 2 Lámina de cartón 3 Lámina de asbesto o metálica 4 Carrizo, bambú o palma 5 Embarro o bajareque 6 Madera 7 Adobe 8 Tabique, ladrillo, block, piedra, cantera, cemento o concreto	Entrada	Nominal
4	mat_techos	1 Material de desecho 2 Lámina de cartón 3 Lámina metálica 4 Lámina de asbesto 5 Palma o paja 6 Madera o tejamanil 7 Terrado con viguería 8 Teja 9 Losa de concreto o viguetas con bovedillas	Entrada	Nominal
5	mat_pisos	1 Tierra 2 Cemento o firme 3 Madera, mosaico u otro recubrimiento	Entrada	Nominal
6	cocina	1 Sí, 2 No	Entrada	Dicotómica
7	cuart_ dorm	Número de cuartos de la vivienda que son usados habitualmente para dormir, aunque también tengan otros usos.	Entrada	Continua
8	num_cuarto	Número total de cuartos que tiene la vivienda, independientemente de su uso.	Entrada	Entrada
9	disp_agua	1 Agua entubada dentro de la vivienda 2 Agua entubada pero dentro del terreno 3 Agua entubada de llave pública (o hidrante) 4 Agua entubada que acarrean de otra vivienda 5 Agua de pipa 6 Agua de un pozo, río, lago, arroyo u otra	Entrada	Nominal
10	bano_comp	Número de baños con excusado y regadera.	Entrada	Continua
11	bano_excus	Número de baños sólo con excusado.	Entrada	Continua
12	bano_regad	Número de baños sólo con regadera.	Entrada	Continua
13	drenaje	1 La red pública 2 Una fosa séptica 3 Una tubería que va a dar a una barranca o grieta 4 Una tubería que va a dar a un río, lago o mar 5 No tiene drenaje	Entrada	Nominal

14	disp_elec	1 Del servicio público 2 De una planta particular 3 De panel solar 4 De otra fuente 5 No tiene luz eléctrica	Entrada	Nominal
15	eli_basura	1 Se la dan a un camión o carrito de basura 2 La llevan al basurero público 3 La dejan en un contenedor o depósito 4 La queman 5 La entierran 6 La tiran en otro lugar (calle, baldío) 7 La tiran en la barranca o grieta 8 La tiran al río, lago o mar	Entrada	Nominal
16	lavadero	1 Sí, 2 No	Entrada	Dicotómica
17	tinaco_azo	1 Sí, 2 No	Entrada	Dicotómica
18	calentador	1 Sí, 2 No	Entrada	Dicotómica
19	aire_acond	1 Sí, 2 No	Entrada	Dicotómica
20	calefacc	1 Sí, 2 No	Entrada	Dicotómica
21	repar_pard	1 Sí, 2 No	Entrada	Dicotómica
22	repar_tech	1 Sí, 2 No	Entrada	Dicotómica
23	repar_agua	1 Sí, 2 No	Entrada	Dicotómica
24	repar_dren	1 Sí, 2 No	Entrada	Dicotómica
25	repar_cabl	1 Sí, 2 No	Entrada	Dicotómica
26	const_dorm	1 Sí, 2 No	Entrada	Dicotómica
27	oonst_coci	1 Sí, 2 No	Entrada	Dicotómica
28	oonst_bano	1 Sí, 2 No	Entrada	Dicotómica
29	tenencia	1 Es rentada 2 Es prestada 3 Es propia, pero la están pagando 4 Es propia 5 Está intestada o en litigio 6 Otra situación	Entrada	Nominal
30	tipo_finan	1 Recursos propios 2 Apoyo de FONHAPO 3 Crédito INFONAVIT o FOVISSSTE 4 Crédito bancario 5 Crédito micro financiero 6 Crédito caja de ahorro 7 Crédito de otra institución 8 Préstamo familiar	Entrada	Nominal
31	escrituras	1 A nombre del dueño 2 A nombre de otra persona 3 No tiene escritura 9 No sabe	Entrada	Nominal
32	tol_resid	Número de residentes de la vivienda	Entrada	Continua
33	tot_hog	Número de hogares en la vivienda	Entrada	Continua
34	ubica_geo	Clave de la entidad, los siguientes tres la clave del municipio y los últimos cuatro la clave de la localidad.	Entrada	Nominal
35	tam_loc	1 Localidades con 100 000 y más habitantes 2 Localidades con 15 000 a 99 999 habitantes 3 Localidades con 2 500 a 14 999 habitantes 4 Localidades con menos de 2 500 habitantes	Entrada	Nominal
36	est_socio	1 Bajo 2 Medio bajo 3 Medio alto 4 Alto	Entrada	Nominal

3.3.2 Diseño de la red neuronal

La arquitectura de la RNA MLP se diseñó en la herramienta de software IBM SPSS Modeler que trabaja mediante la interfaz virtual Watson Studio. La Figura 3.5 muestra el flujo del procesamiento de datos para la construcción de los modelos de cada una de las redes neuronales de las regiones de México. La arquitectura de las RNA's se construye mediante el uso de diferentes

herramientas que desempeñan una actividad en particular para la RNA y se representan mediante íconos en el software.

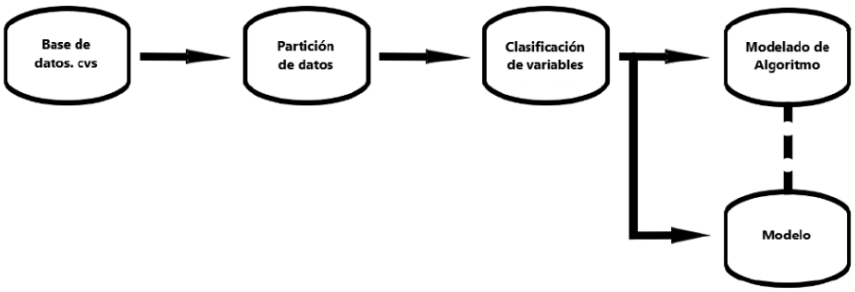

Figura 3.5. Flujo de procesamiento de datos para la construcción de los modelos. (Fuente: elaboración propia en IBM SPSS Modeler (2020)).

El flujo de la arquitectura para cada red consta de los siguientes pasos:

1. Se importan los registros de entrada que integran las bases de datos a través del ícono de Asignación de Datos, el cual permite trabajar con archivos de Excel. CSV.
2. Se determina la partición de los registros de las bases de datos considerando tres conjuntos: uno de entrenamiento, otro de prueba y un tercero de validación; éste último se utiliza para determinar el fin del entrenamiento de la red con el propósito de evitar el sobre entrenamiento. La partición se realiza mediante el icono de Partición, con una proporción de 65% de entrenamiento, 25% de pruebas y 10% de validación.
3. El icono Tipo se utiliza asignar las características de las variables en función de sus datos (continua, nominal, categórica, etc.), el rol que desempeñan en el sistema de hipótesis (dependiente, independiente, identificación de registro, etc.) y el valor de medición de cada una de las variables de interés. Para minimizar el efecto del tamaño de los valores de entradas y de salidas, y aumentar la efectividad de aprendizaje, el algoritmo normaliza los registros de las variables de entrada a valores entre cero y uno. El software permite ajustar el tipo de normalización en caso de que no se deseen valores grandes.
4. Se integra el icono de modelado RNA en el cual se seleccionan las características del algoritmo (elección de variables de entrada, el nivel

de confianza de las predicciones o el número de ciclos que se desea ejecutar). En este caso, seleccionamos 1000 ciclos y establecemos un conjunto de prevención de sobreajuste del 10% para rastrear errores durante el entrenamiento a fin de evitar que el método modele la variación de probabilidad en los datos. Por último se especifica el número de capas ocultas y el número de neuronas en cada capa oculta.

3.3.3 Características de la RNA

En el presente trabajo se emplea una RNA MLP en cada una de las regiones de México de acuerdo con los dos esquemas de división regional económica que se analizan. Se utiliza una arquitectura de una capa oculta. Se recomienda una capa oculta para problemas en los que se busca conseguir un error bajo y dos capas ocultas para problemas de clasificación (Coloma et al., 2019). Sin embargo, se ha demostrado que para la mayoría de los problemas es suficiente una sola capa oculta (Funahashi, 1989 citado por Picón, 2011) ya que un mayor número de capas puede provocar sobre entrenamiento, lo cual favorece la precisión del modelo pero impide que el algoritmo aprenda , es decir, que sea capaz de generalizar la clasificación del tipo de vivienda y que al consultarle con nuevos registros desconocidos proporcione un resultado fiable dada su capacidad de generalización [24]. La capa de entrada tiene 34 neuronas (una por cada variable), mientras que el número de neuronas en la capa oculta es calculado por el algoritmo de manera automática. En la capa de salida de las redes 5 para todas las regiones económicas de México, que corresponden al número de tipos de vivienda, no obstante el tipo de vivienda 4 (vivienda en cuarto de azotea) es el de menor uso en todas las regiones, por esta razón el algoritmo omite predicciones de este tipo de vivienda. La información de la arquitectura de cada red se describe en la Tabla 3.5.

Tabla 3.5. Información de la arquitectura del modelo. (Fuente: elaboración propia).

Información de la Arquitectura del Modelo	
Variable dependiente	tipo_viv
Tipo de modelo	Clasificación
Variables independientes	34
Neuronas en la capa oculta	13 hasta 20
Función de activación en la capa oculta	Tangente hiperbolica
Función de activación en la capa de salida	Softmax
Neuronas en la capa de salida	5

En la capa oculta se utilizó la tangente hiperbólica como función de activación y en la capa de salida, donde están las neuronas de clasificación del tipo de vivienda, la función de activación softmax ya que la variable dependiente es multiclase, con cinco tipos de vivienda para todas las regiones económicas de México, por ejemplo, en la Figura 3.6 se presente la RNA de la Región Sur, en la cual se observa los tipos de vivienda más difundidos en dicha región. En esta RNA también se observa el sesgo de los registros respecto al tipo de vivienda en la capa de salida, donde las líneas en color azul representan un valor mayor y en color rojo un valor menor, en este caso el sesgo favorece el tipo de vivienda 1 (vivienda independiente). Para esta investigación se ejecutó el algoritmo con valores reales ya que el algoritmo realiza automáticamente la normalización de las variables continuas.

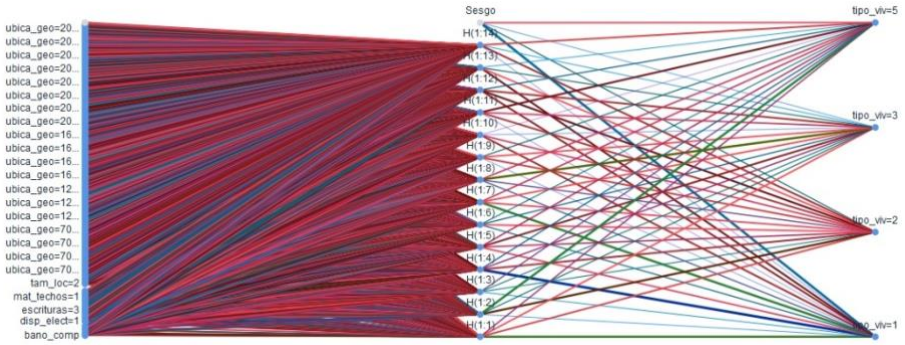

Figura 3.6. Diagrama de la RNA de la Región Sur.
(Fuente: elaboración propia en IBM SPSS Modeler (2020)).

3.4 Evaluación de los resultados obtenidos

Tras el entrenamiento de cada una de las RNA's, se procede a el análisis de la eficiencia los resultados mediante los diferentes indicadores de evaluación de los modelos que proporciona el software, los cuales se definen a continuación:

Precisión. Valor predictivo positivo que representa la proporción de predicciones correctas que contiene el modelo. El cálculo de precisión (*P*) está dado por:

$$P = \frac{verdaderos\ positivos + verdaderos\ negativos}{(verdaderos\ positivos + verdaderos\ negativos + falsos\ positivos + falsos\ negativos)} \quad (7)$$

Índice de verdaderos positivos. Presenta la proporción de predicciones correctas en predicciones de clase positiva. Su cálculo esta dado por:

$$Indice\ de\ verdaderos\ positivos = \frac{verdaderos\ positivos}{(verdaderos\ positivos + falsos\ negativos)} \quad (8)$$

Índice de falsos positivos. El índice de falsos positivos representa la proporción de predicciones incorrectas en clase positiva. El cálculo del índice de falsos positivos se calcula de la siguiente forma:

$$Indice\ de\ falsos\ positivos = \frac{falsos\ positivos}{(negativos\ verdaderos + falsos\ positivos)} \quad (9)$$

Exhaustividad ponderada. Es una medida de sensibilidad que determina la proporción de predicciones correctas en clase positiva. La exhaustividad está dada por:

$$Exhaustividad\ ponderada = \frac{verdaderos\ positivos}{(verdaderos\ positivos + falsos\ negativos)} \quad (10)$$

Precisión ponderada. Proporciona el promedio ponderado de precisión con ponderaciones iguales a la probabilidad de clase. La precisión ponderada está dada por:

$$Precisión\ ponderada = \frac{verdaderos\ positivos}{(verdaderos\ positivos + falsos\ negativos)} \quad (11)$$

Medida F1 ponderada. Es una medida de exactitud que proporciona el promedio absoluto de precisión y exhaustividad. La medida F1 ponderada está dada por:

$$Medida\ F1\ ponderada = 2 * \frac{(precisión * exhaustividad)}{(precisión + exhaustividad)} \quad (12)$$

La Tabla 3.6 muestra las medidas de evaluación de los modelos para comparar los dos esquemas de división regional económica que se analizan. El algoritmo proporciona los resultados de las 3 particiones de los registros (entrenamiento, prueba y validación) en este apartado solo se presentan los resultados correspondientes a la partición de entrenamiento que representa el 60% de los registros. Con respecto a la eficiencia de los modelos que se estudian se observa que la precisión más baja se obtuvo en la Región Capital del Esquema 2, sin embargo el valor del indicador mejora con el entrenamiento con nuevos registros.

En relación con los índices de verdaderos positivos y de falsos negativos, los resultados son congruentes con la precisión de los modelos, ya que el modelo es más confiable, cuando el valor del índice de verdaderos positivos se aproxima

a uno, y el índice de falsos negativos es cercano a cero. Los modelos que presentan menor eficiencia corresponden a las regiones Noroeste del Esquema 1 y Pacífico del Esquema 2, sin embargo de acuerdo con resultados de la partición de validación del índice de falsos positivos en las regiones Noroeste y Pacífico es cero, lo cual indica que con el entrenamiento el aprendizaje mejora la clasificación.

Tabla 3.6. Medidas de evaluación del modelo. (Fuente: elaboración propia).

Medidas de Evaluación del Modelo						
Región Económica Esquema: 2	**Precisión**	**Índice de verdaderos positivos**	**Índice de falsos positivos**	**Precisión ponderada**	**Exhaustividad ponderada**	**Medida F1 ponderada**
Centro	0.926	0.926	0.233	0.934	0.926	0.930
Capital	0.876	0.876	0.251	0.903	0.876	0.887
Centro Norte	0.979	0.979	0.000	1.000	0.979	0.989
Altiplano	0.983	0.983	0.000	1.000	0.983	0.992
Occidente	0.975	0.975	0.143	0.999	0.975	0.987
Pacífico	0.980	0.980	0.449	0.995	0.980	0.987
Oriente	0.973	0.973	0.000	1.000	0.973	0.986
Este	0.968	0.968	0.250	0.999	0.968	0.985
Sureste	0.986	0.986	0.000	1.000	0.986	0.993
Golfo	0.985	0.985	0.222	0.999	0.985	0.992
Suroccidente	0.987	0.987	0.000	1.000	0.987	0.993
Sur	0.986	0.986	0.000	1.000	0.986	0.993
Noroeste	0.990	0.990	0.499	0.997	0.990	0.994
Noreste	0.986	0.986	0.000	0.970	0.970	0.970
Norte	0.986	0.986	0.000	1.000	0. 986	0.993

Los resultados de los sesgos, respecto a cada una de las clases de la variable dependiente, y los pesos de las variables independientes de la capa de salida, se presenta en un análisis que compara las regiones que son similares de los dos esquemas de división regional económica que se utilizan. Este análisis se basa en la capa de salida de las redes neuronales, el software incluye una interfaz que permite observar gráficamente las ponderaciones de mayor o menor importancia del diagrama de la RNA.

3.4.1 Región Centro y Región Capital

La Región Centro agrupa el Estado de México, Morelos y la Ciudad de México, en la Figura 3.7 se muestra el sesgo de esta región (línea roja) para el tipo de vivienda 1 con una importancia ponderada de 0.576, sin embargo el tipo de vivienda 2 es el más importante con una ponderación de 1.516 (línea azul). La Región Capital incluye La Ciudad de México y el Estado de México, en esta región se presenta un sesgo (ver Figura 3.7) dividido para los tipos de vivienda

1 (línea azul) y 2 (línea verde), con pesos de 0.897 y 0.555 respectivamente, sin embargo el tipo de vivienda 2 tiene una importancia ponderada de 0.944 (línea azul). En las dos regiones la preferencia favorece al tipo de vivienda 2 aunque nos sea el tipo de vivienda más usual. En la Figura 3.7 también se observa en la región Capital que los tipos de vivienda 4 (Vivienda en cuarto de azotea) y 5 (Local no construido para habitación) tienen mayor importancia ya que la región Centro que se compone de dos estados, esta característica permite que la RNA alcance mayor precisión.

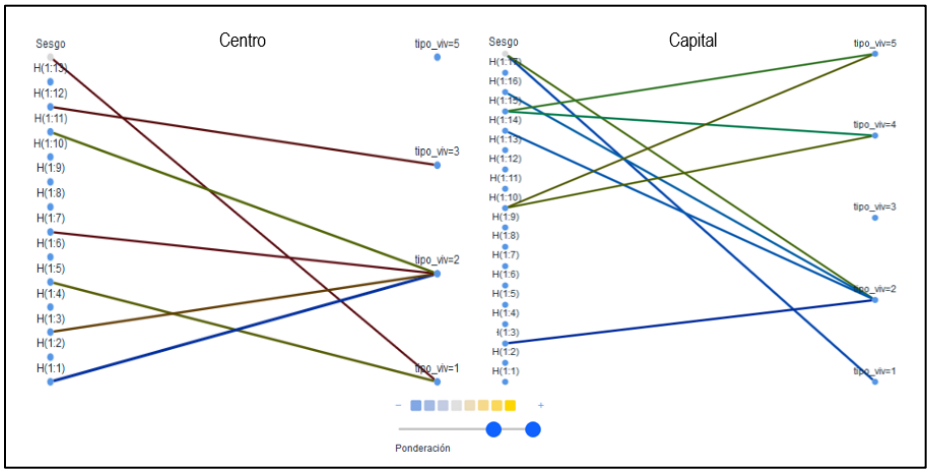

Figura 3.7. Diagramas parciales de las RNA's de las regiones Centro y Capital Sur. (Fuente: elaboración propia en IBM SPSS Modeler (2020)).

En la Tabla 3.7 se presentan los pesos de las variables de mayor importancia para las regiones Centro y Capital. En estas dos regiones es más frecuente el hacinamiento, los porcentajes de hogares que presentan este problema en las dos regiones son 8.83% región Centro y 7.85% región Capital [11], por esta razón se justifica que las variables que tienen relación con el número de cuartos de Baño, Total de Residentes y Total de Hogares sean las más importantes. En estas regiones predomina vivienda vertical, por esta razón se considera que las variables de Ubicación Geográfica y de Número de Cuartos, sean también de mayor peso.

Tabla 3.7. Ponderación de Variables de las regiones Centro y Capital. (Fuente: elaboración propia).

Ponderación de Variables			
Centro	**%**	**Capital**	**%**
Baño excusado	7.99	Baño completo	13.27
Total de hogares	6.96	Baño regadera	10.79
Ubicación geográfica	6.65	Baño excusado	9.79
Número de cuartos	6.28	Total de residentes	6.87
Cuartos dormitorio	5.14	Número de cuartos	6.59
Eliminación de basura	5.02	Total de hogares	5.73

Baño regadera	4.72	Cuartos dormitorio	3.92
Baño completo	4.62	Ubicación geográfica	3.87
Tamaño de la localidad	3.87	Estrato socioeconómico	3.14
Drenaje	3.76	Tipo de financiamiento	2.86
Suma	55.01		66.83

3.4.2 Región Centro Norte y Región Altiplano

La región Centro Norte comprende los estados de Aguascalientes, Guanajuato, Querétaro, San Luis Potosí y Zacatecas, el sesgo en esta región favorece el tipo de vivienda 1 con un peso de 0.472, y es el de mayor importancia con un peso de 1.167 (línea azul Figura 3.8). La Región Altiplano, incluye el estado de Durango al conjunto de estados de la Región Centro Norte, en esta región el de igual forma el sesgo es para el tipo de vivienda 1 con un peso de 0.897 (línea azul Figura 3.8), sin embargo el tipo de vivienda 3 tiene mayor importancia con un peso de 0.629 (línea verde Figura 3.8), seguido de los tipos de vivienda 5 y 1, con un peso de 0.621 y 0.597 respectivamente. En ambas regiones la vivienda en vecindad tiene la misma importancia que el tipo de vivienda 1.

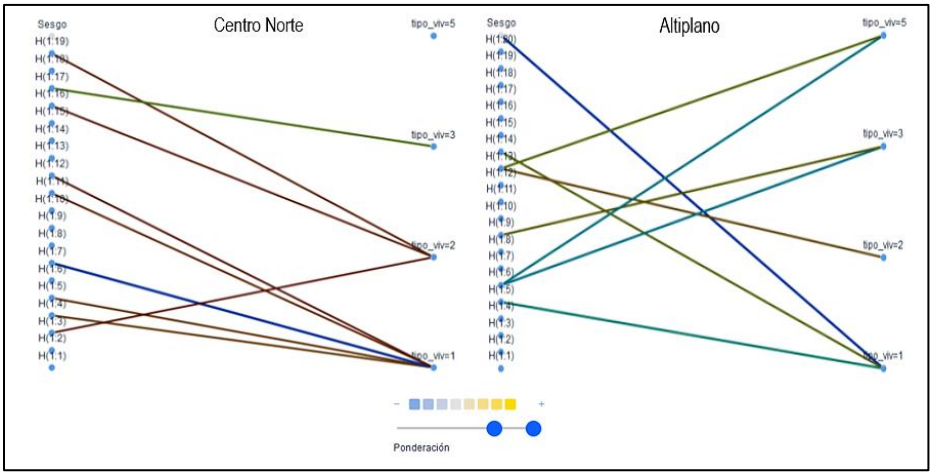

Figura 3.8. Diagramas parciales de las RNAs de las regiones Centro Norte y Altiplano.
(Fuente: elaboración propia en IBM SPSS Modeler (2020)).

En la Tabla 3.8 se presentan los pesos de las variables de mayor importancia para las regiones Centro Norte y Altiplano, En estos dos conjuntos regionales los usuarios consideran importante el tipo de financiamiento para adquirir una vivienda que se puede atribuir al crecimiento económico de la zona. La diferencia entre estas dos regiones no es significativa, los resultados de la eficiencia de ambos modelos solo presentan diferencias en la importancia de las variables.

Tabla 3.8. Ponderación de Variables de las regiones Centro Norte y Altiplano.
(Fuente: elaboración propia).

Ponderación de Variables			
Centro Norte	**%**	**Altiplano**	**%**
Baño excusado	10.15	Cuartos dormitorio	9.61
Total de residentes	9.29	Baño completo	9.59
Total de hogares	7.18	Total de residentes	6.79
Cuartos dormitorio	6.95	Total de hogares	6.67
Número de cuartos	6.86	Baño excusado	5.91
Baño completo	5.89	Número de cuartos	4.48
Baño regadera	4.37	Ubicación geográfica	4.43
Ubicación geográfica	3.87	Baño regadera	4.09
Tamaño de la localidad	2.73	Tipo de financiamiento	3.79
Tipo de financiamiento	2.57	Material de paredes	3.51
Suma	59.86		58.87

3.4.3 Región Occidente y Región Pacífico

La Región Occidente incluye Colima, Jalisco, Michoacán y Nayarit, en esta región se presenta un sesgo (ver Figura 3.9) para el tipo de vivienda 1 (línea azul) con un peso de 0.764, y una importancia ponderada de 0.769 (línea azul), sin embargo los tipos de vivienda 3 y 2 tiene mayor importancia con pesos de 0.893 y 0.785 respectivamente. La Región Pacífico agrupa los estados de Baja California Sur, Colima, Jalisco, Nayarit y Sinaloa. En la Figura 3.9 se muestra el sesgo de esta región (línea verde) para el tipo de vivienda 1 con un peso de 0.552, y una importancia ponderada de 0.758 (línea azul) seguido de los tipos de vivienda 5 y 2 con pesos 0.724 y .683 respectivamente. En las dos regiones presentan una marcada diversificación de las preferencias sobre el tipo de vivienda.

En la Tabla 3.9 se presentan los pesos de las variables independientes de mayor importancia para las regiones Occidente y Pacífico, En estos dos conjuntos regionales se observa que los usuarios vivienda consideran importante evitar el hacinamiento ya que los modelos presentan un peso alto a las variables Cuartos Dormitorio, Número de Cuartos, Total de Residentes y Total de Hogares. Las dos regiones que se comparan son muy diferentes ya que solo 3 estados pertenecen a las dos alternativas regionales, esta circunstancia se refleja en las preferencias de vivienda de los diagramas de las RNA’s así como en los pesos de importancia de las variables.

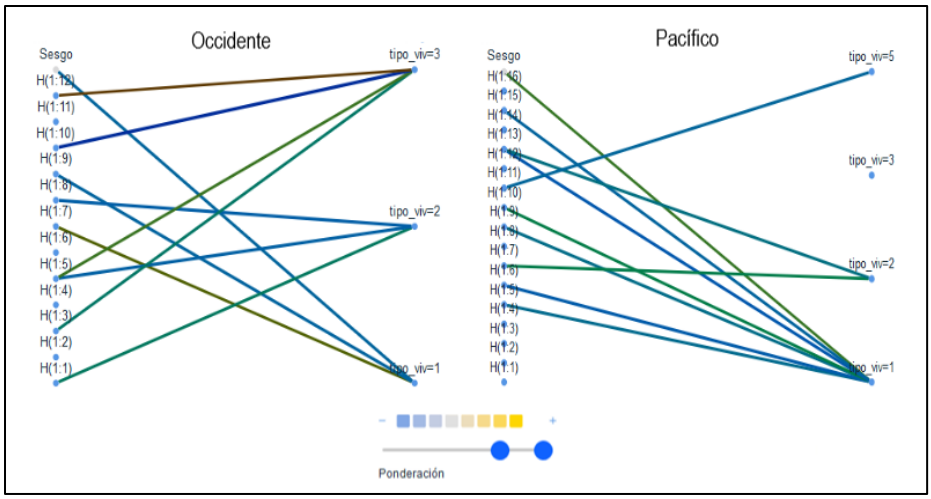

Figura 3.9. Diagramas parciales de las RNAs de las regiones Occidente y Pacífico.
(Fuente: elaboración propia en IBM SPSS Modeler (2020)).

Tabla 3.9. Ponderación de Variables de las regiones Occidente y Pacífico. (Fuente: elaboración propia).

Ponderación de Variables			
Occidente	**%**	**Pacífico**	**%**
Cuartos dormitorio	9.70	Número de cuartos	11.81
Total de residentes	8.79	Total de residentes	8.55
Baño excusado	8.63	Total de hogares	7.43
Baño completo	6.22	Cuartos dormitorio	5.90
Número de cuartos	5.52	Baño completo	4.53
Total de hogares	5.46	Baño excusado	4.43
Ubicación geográfica	4.73	Ubicación geográfica	4.01
Tamaño de la localidad	3.47	Eliminación de basura	3.81
Baño regadera	3.22	Estrato socioeconómico	3.64
Tenencia	2.84	Baño regadera	3.33
Suma	58.58		57.44

3.4.4 Región Oriente y Región Este

La región Oriente está conformada por los estados de Hidalgo, Puebla, Tlaxcala y Veracruz, en la Figura 3.10 se muestra el sesgo para el tipo de vivienda 1 (línea verde) con peso de 0.415 y una importancia ponderada de 0.757 (línea azul). La región Este agrupa los estados de Hidalgo, Morelos, Puebla y Tlaxcala. De igual manera el sesgo en esta región (línea azul Figura 3.10) es para el tipo de vivienda 1 con un peso de 0.810, y una importancia ponderada de 0.374 (línea roja). En las dos regiones existe una diversificación análoga de los tipos de vivienda y los índices de importancia de las variables son también similares.

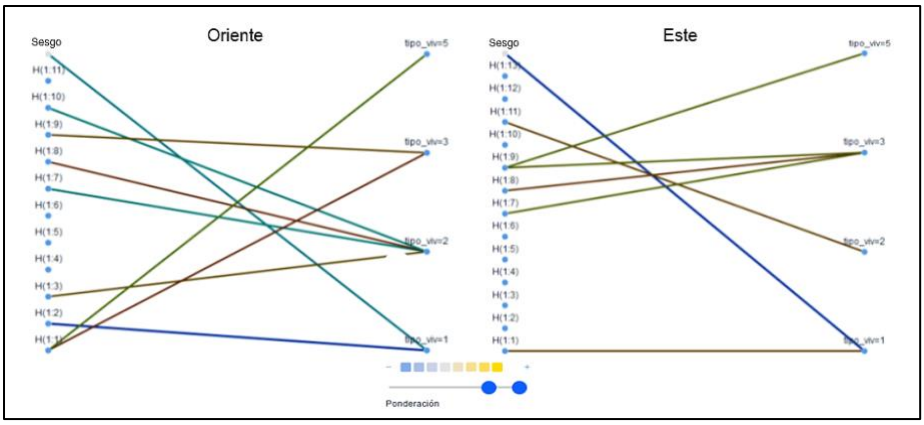

Figura 3.10. Diagramas parciales de las RNAs de las regiones Oriente y Este.
(Fuente: elaboración propia en IBM SPSS Modeler (2020)).

En la Tabla 3.10 se presentan los pesos de las variables independientes de mayor importancia de las regiones Oriente y Este, En esta región de México se presentan los precios de vivienda más altos, por lo que se observan entre las variables de mayor importancia la Tenencia, la Ubicación Geográfica y el tipo de financiamiento. De acuerdo con los resultados el modelo de la región Oriente presenta una mejor eficiencia.

Tabla 3.10. Ponderación de Variables de las regiones Oriente y Este.
(Fuente: elaboración propia).

Ponderación de Variables			
Oriente	**%**	**Este**	**%**
Cuartos dormitorio	12.03	Cuartos dormitorio	9.60
Total de hogares	11.43	Baño completo	8.45
Baño completo	8.05	Total de hogares	6.84
Baño excusado	7.28	Baño excusado	6.63
Total de residentes	5.61	Número de cuartos	6.36
Número de cuartos	4.66	Total de residentes	4.76
Baño regedera	4.65	Ubicación geográfica	4.64
Ubicación geográfica	4.43	Baño regedera	3.83
Tipo de financiamiento	2.94	Dispone de Electricidad	3.57
Tamaño de la localidad	2.72	Tenencia	3.56
Suma	63.80		58.24

3.4.5 **Región Sureste y Región Golfo**

La región Sureste comprende los estados de Campeche, Quintana Roo, Tabasco y Yucatán. El sesgo de esta región es para el tipo de vivienda 1 (línea verde Figura 3.11) tiene un peso de 0.555, y una importancia ponderada de 0.771 (línea azul), seguido por la preferencia de los tipos de vivienda 3 y 5 con una ponderación de 0.533 y 0.465 respectivamente. La región Golfo se compone de igual manera que la región Sureste y añade al estado de Veracruz. El sesgo en

esta región es para del tipo de vivienda 1 tiene un peso 0.660 (línea roja Figura 3.11), y una importancia ponderada de 1.133 (línea azul), seguido de la preferencia de los tipos de vivienda 2 y 3 con una ponderación de 0.775 y 0.704 respectivamente. En las dos regiones existe una marcada difusión de los tipos de vivienda.

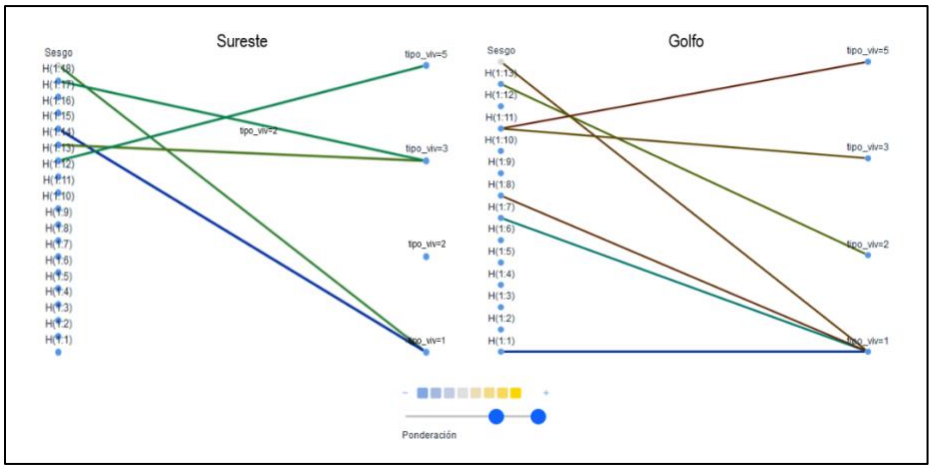

Figura 3.11. Diagramas parciales de las RNAs de las regiones Sureste y Golfo.
(Fuente: elaboración propia en IBM SPSS Modeler (2020)).

En la Tabla 3.11 se presentan los pesos de las variables de mayor importancia para las regiones Sureste y Golfo. Las dos regiones que se comparan son casi idénticas ya que solo se diferencian por el estado de Veracruz que de acuerdo con los resultados se deduce que los usuarios de vivienda de dicho estado tienen las mismas preferencias que sus vecinos.

Tabla 3.11. Ponderación de Variables de las regiones Sureste y Golfo. (Fuente: elaboración propia).

Ponderación de Variables			
Sureste	**%**	**Golfo**	**%**
Baño regadera	9.63	Total de residentes	9.11
Total de hogares	8.19	Número de cuartos	7.56
Total de residentes	6.29	Total de hogares	6.65
Baño completo	5.62	Baño excusado	6.37
Cuartos dormitorio	5.30	Baño completo	6.06
Número de cuartos	4.68	Cuartos dormitorio	5.80
Ubicación geográfica	4.40	Ubicación geográfica	4.92
Tenencia	4.38	Drenaje	4.68
Baño excusado	3.81	Material de paredes	3.92
Eliminación de basura	3.71	Eliminación de basura	3.79
Suma	56.01		58.86

3.4.6 Región Suroeste y Región Sur

La región Suroeste se conforma por los estados de Chiapas, Guerrero y Oaxaca. En la Figura 3.12 se muestra el sesgo de esta región para el tipo de vivienda 1 (línea azul) con peso de 0.928 y con una importancia ponderada de 0.909 (línea azul intermedia), seguido del tipo de vivienda 5 con un peso de 0.753 (línea verde superior). La región Sur integra los estados de Chiapas, Guerrero, Michoacan y Oaxaca. El sesgo en esta región (línea azul Figura 3.12) es para el tipo de vivienda 1 con un peso de 0.813, y una importancia ponderada de 0.924 (línea azul inferior). En las dos regiones se presenta una diversificación de los tipos de vivienda similar, no obstante predomina el tipo de vivienda 5.

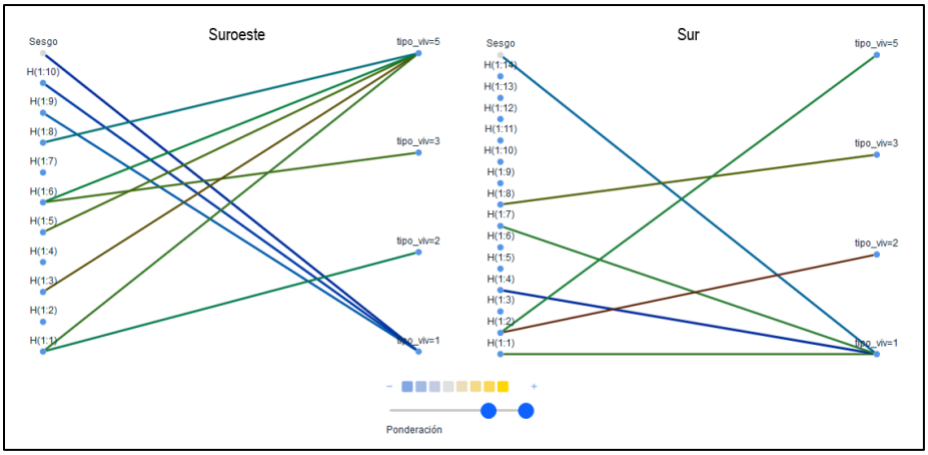

Figura 3.12. Diagramas parciales de las RNAs de las regiones Suroeste y Sur.
(Fuente: elaboración propia en IBM SPSS Modeler (2020)).

En la Tabla 3.12 se presentan los pesos de las variables de mayor importancia de las regiones Suroeste y Sur. La diferencia de la importancia de las variables en las dos regiones es poco significativa.

Tabla 3.12. Ponderación de Variables de las regiones Suroeste y Sur.(Fuente: elaboración propia).

Ponderación de Variables			
Suroeste	**%**	**Sur**	**%**
Total de hogares	13.34	Total de hogares	13.49
Baño excusado	11.62	Baño completo	8.12
Total de residentes	6.82	Número de cuartos	6.18
Cuartos dormitorio	5.02	Cuartos dormitorio	5.67
Número de cuartos	4.53	Ubicación geográfica	4.93
Tipo de financiamiento	4.44	Tenencia	4.51
Baño completo	4.22	Baño excusado	4.45
Ubicación geográfica	4.14	Total de residentes	4.22
Tamaño de la localidad	2.95	Eliminación de basura	3.05
Disponibilidad de Agua	2.84	Tipo de financiamiento	2.88
Suma	59.92		57.50

3.4.7 Región Noroeste, Región Noreste y Región Norte

La región Noroeste está conformada por los estados de Baja California, Baja California Sur, Chihuahua, Durango, Sinaloa y Sonora, en la Figura 3.13 se muestra el sesgo para el tipo de vivienda 1 (línea verde) con peso de 0.881 y una importancia ponderada de 1.192 (línea azul). En esta región se tiene una difusión media de los tipos de vivienda 2 y 5. La región Noreste agrupa los estados de Coahuila, Nuevo León y Tamaulipas. De igual manera el sesgo en esta región (línea roja Figura 3.13) es para el tipo de vivienda 1 con un peso de 0.449, y una importancia ponderada de 0.836 (línea azul) seguido de la vivienda 5 con un peso de 0.585. La región Norte que se compara con las dos anteriores, no se tiene un sesgo fuerte para alguno de los tipos de vivienda (Figura 3.13) solo se presenta una importancia ponderada de 0.472 para el tipo de vivienda 5 (línea amarilla), de 0.689 para el tipo de vivienda 2 (línea verde) y de 0.958 para el tipo de vivienda 1(línea azul). En las tres regiones se presenta una diversificación de los tipos de vivienda similar, no obstante el tipo de vivienda 5 tiene mayor importancia en la región Noroeste.

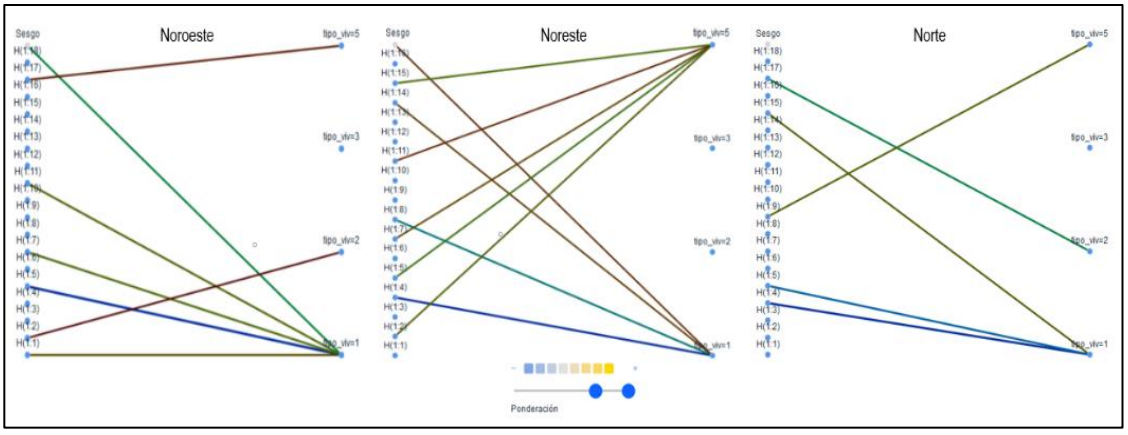

Figura 3.13. Diagramas parciales de las RNAs de las regiones Noroeste, Noreste y Norte.
(Fuente: elaboración propia en IBM SPSS Modeler (2020)).

En la Tabla 3.13 se presentan los pesos de las variables independientes de mayor importancia de las regiones Noroeste, Noreste y Norte. La diferencia de la importancia de las variables entre estas dos regiones es poco significativa. De acuerdo con los resultados el modelo de la región Norte presenta una mejor eficiencia ya que los estados del norte presentan menor variación en las preferencias de los usuarios de vivienda.

Tabla 3.13. Ponderación de Variables de las regiones Noroeste, Noreste y Norte.
(Fuente: elaboración propia).

Ponderación de Variables					
Noroeste	**%**	**Noreste**	**%**	**Norte**	**%**
Total de hogares	12.87	Total de hogares	11.1	Total de hogares	12.57
Número de cuartos	10.08	Total de residentes	7.58	Baño excusado	12.41
Cuartos dormitorio	7.55	Baño excusado	5.51	Baño regadera	9.34
Baño regadera	6.71	Baño completo	5.22	Total de residentes	6.95
Baño completo	4.47	Baño regadera	4.25	Cuartos dormitorio	6.61
Total de residentes	4.31	Ubicación geográfica	4.17	Número de cuartos	5.31
Baño excusado	4.08	Número de cuartos	3.87	Baño completo	4.41
Ubicación geográfica	4.02	Material de techos	3.61	Ubicación geográfica	3.987
Tipo de financiamiento	3.42	Cuartos dormitorio	3.51	Disponibilidad de agua	2.68
Tamaño de la localidad	3.12	Material de paredes	3.16	Tipo de financiamiento	2.60
Suma	60.63		51.98		66.87

3.5 Discusión

Después de analizar las comparaciones entre los dos esquemas de división regional económica se observa una marcada similitud de los tipos de vivienda de cada una de las regiones, aunque el tipo de vivienda1sea el más difundido en todo el país, las diferencias que existen entre las regiones se acentúan por la difusión que se tiene para los otros cuatro tipos de vivienda. Las mismas comunidades que integran una región tienen características propias, sin embargo son más las características en común que aquellas que las individualizan. Es posible generar una RNA para cada comunidad para proporcionar mayor precisión en la selección de vivienda, por ejemplo, si se buscara redensificar ciertas zonas de una ciudad mediante la rehabilitación de vivienda, sería necesario contar con el registro de las viviendas disponibles para nuevos usuarios lo que permitiría que la RNA sea más precisa. En cambio, si se desea analizar las preferencias de los usuarios de vivienda en una región es posible hacerlo mediante una RNA que identifique las características regionales, de esta manera la RNA obtiene un aprendizaje más eficiente ya que identifica diferentes patrones para clasificar los tipos de vivienda.

Conforme a los resultados de la comparación que se presentó en la sección anterior se plantea una división regional de las características de vivienda en México, la cual es una combinación entre el esquema 1 y 2 (Figura 3.14) con una única diferencia respecto a la región norte la cual es una mezcla de las regiones

Noreste y Noroeste del esquema 1 y la región Norte del esquema 2, esta integración se debe a similitud de las preferencias vivienda en los estados que se agrupan por lo que se tiene los resultados que se presentan en la Tabla 3.14.

Figura 3.14. División regional de las características de vivienda en México. (Fuente: elaboración propia).

Tabla 3.14. Medidas de evaluación del modelo. (Fuente: elaboración propia).

Medidas de Evaluación del Modelo						
Región	**Precisión**	**Índice de verdaderos positivos**	**Índice de falsos positivos**	**Precisión ponderada**	**Exhaustividad ponderada**	**Medida F1 ponderada**
Norte	0.986	0.986	0.000	1.000	0.986	0.993

3.6 Conclusiones

En este trabajo presentamos la aplicación de las redes neuronales aplicadas a problema de clasificación para la selección de vivienda, generando una clasificación precisa y con índices de evaluación congruentes. La evaluación de la vivienda representa un reto en cualquier país sin embargo se deben buscar los instrumentos que permitan indagar sobre la satisfacción de los usuarios de vivienda y sus aspiraciones y de esta forma generar las alternativas para alcanzar el desarrollo de comunidades sustentables y el bienestar de las personas.

El aprendizaje de las redes neuronales permite crear modelos que se diseñar de acuerdo con los requerimientos que se tengan. Aunque algunos autores mencionan que los resultados de las variables de salida no se pueden interpretar de la misma manera que en los en métodos estadísticos, con los resultados de esta investigación es posible afirmar que el comportamiento de los sesgos y de las ponderaciones sinápticas de la capa de salida tienen un comportamiento similar al comportamiento estadístico y se pueden interpretar si se conoce el

contexto de los datos que se analizan. Sin embargo en problema de clasificación de un número reducido de clases es probable que se presente el sobre entrenamiento de la red neuronal ya que el algoritmo aprende con una alta precisión.

En lo referente a las bases de datos se considera conveniente que se desarrolle una clasificación más extensa de los tipos de vivienda ya que la clasificación de los tipos de vivienda no abarca los tipos de vivienda que existen en las ciudades mexicanas ni se obedece al menos a los tipos de vivienda de interés social. También se observa la necesidad de un estudio más elaborado de las variables que definen y tienen una relación directa con las características de las viviendas actuales así como las variables indirectas como los tipos de financiamiento, la conectividad de transporte, los servicios de comunicación, entre otros, de manera que la Encuesta Nacional de Hogares ofrezca información actual y más precisa.

En relación con los resultados obtenidos se plantea como investigación futura probar el modelo en campo para determinar su verdadera eficiencia además de afinar la selección de variables a incluir.

3.7 Referencias

1. Samad, D., Zainon, N., Rahim, F.A.M., Lou, E.: Malaysian affordability housing policies revisited. Open House Int. (2017). https://doi.org/10.1051/matecconf/20166600010
2. CONEVAL: Estudio Diagnóstico del Derecho a la Vivienda Digna y Decorosa 2018. , Ciudad de México (2018).
3. ACNUDH - ONU Habitat. (2010). El derecho a una vivienda adecuada. Folleto informativo nº21. *Revista de Antropología Social*, *19*, 103–129. https://www.ohchr.org/Documents/Publications/FS21_rev_1_Housing_sp.pdf.
4. Villavicencio, J., Duran, A.M.: Treinta años de vivienda social en la ciudad de MexicoÑ nuevas necesidades y demandas. Scr. Nov. 146, 1–13 (2014).
5. Kunz Bolaños, I.C., Espinosa Flores, A.S.: Elementos de éxito en la producción social de la vivienda en México. Econ. Soc. y Territ. 683 (2017), https://doi.org/10.22136/est2017875.
6. Gaspareniene, L., Venclauskiene, D., Remeikiene, R.: Critical Review of Selected Housing Market Models Concerning the Factors that Make Influence on Housing Price Level Formation in the Countries with Transition Economy. Procedia - Soc. Behav. Sci. (2014). https://doi.org/10.1016/j.sbspro.2013.12.886.
7. ONU-Habitat. (2019). Elementos de una Vivienda Adecuada. Recuperado el 16 de mayo de 2020, de https://onuhabitat.org.mx/index.php/elementos-de-una-vivienda-adecuada.
8. Choy, L.H.T., Ho, W.K.O., Mak, S.W.K.: Housing attributes and Hong Kong real estate prices: A quantile regression analysis. Constr. Manag. Econ. 30, 359–366 (2012). https://doi.org/10.1080/01446193.2012.677542.
9. Yao, Y., Zhang, J., Hong, Y., Liang, H., He, J.: Mapping fine-scale urban housing prices by fusing remotely sensed imagery and social media data. Trans. GIS. 22, 561–581 (2018). https://doi.org/10.1111/tgis.12330.
10. Ezquiaga Domínguez, J.M., Salat, S., Tojo, J. F., & Naredo, J.M., Ainz Ibarrondo, M.J., Bilbao Uribarri, A., Torres Elizburu, R.: Libro blanco de la sostenibilidad en el planeamiento urbanístico Español. 2nd Natl. Congr. Energy Sp. (2010). https://doi.org/10.3989/estgeogr.201126.
11. INEGI: Encuesta Nacional de los Hogares 2017, https://www.inegi.org.mx/programas/enh/2017/.
12. Llinás Solano, H., Arteta Charris, M., Tilano Hernández, J.: El modelo de regresión logística para el caso en que la variable de respuesta puede asumir uno de tres niveles: Estimaciones, pruebas de hipótesis y selección de modelos. Rev. Matemática Teoría y Apl. (2016). https://doi.org/10.15517/rmta.v23i1.22442.
13. Eichie, J.O., Oyedum, O.D., Ajewole, M.O., Aibinu, A.M.: Comparative analysis of basic models and artificial neural network based model for path loss prediction. Prog. Electromagn. Res. M. 61, 133–146 (2017). https://doi.org/10.2528/PIERM17060601.

14. Anand, P.: Bias in machine learning (2017), https://iq.opengenus.org/bias-machine-learning/.
15. Du, K.L., Swamy, M.N.S.: Neural networks and statistical learning. Springer-Verlag London Ltd (2014).
16. Ferreira, R.P., Martiniano, A., Napolitano, D., Romero, M., De Oliveira Gatto, D.D., Farias, E.B.P., Sassi, R.J.: Artificial Neural Network for Websites Classification with Phishing Characteristics. Soc. Netw. 07, 97–109 (2018), https://doi.org/10.4236/sn.2018.72008.
17. Vanus, J., Fiedorova, K., Kubicek, J., Gorjani, O.M., Augustynek, M.: Wavelet-based filtration procedure for denoising the predicted CO2 waveforms in smart home within the internet of things. Sensors (Switzerland). (2020). https://doi.org/10.3390/s20030620.
18. ElKessab, B., Daoui, C., Boukhalene, B., Salouan, R.: A Comparative Study between the K-Nearest Neighbors and the Multi-Layer Perceptron for Cursive Handwritten Arabic Numerals Recognition. Int. J. Comput. Appl. 107, 25–30 (2014). https://doi.org/10.5120/19140-0117.
19. Marín Diazaraque, J.M.: Introducción a las redes neuronales aplicadas. Man. Data Min. (2007).
20. Han, J., Kamber, M., Pei, J.: 9.2 Classification by backpropagation. En: Data Mining: Concepts and Techniques. p. 398. Morgan Kaufmann Publishers (2012).
21. Fouquet, A.: Disparidades regionales en México: ¿Cuestión de herencia o de geografía? En: Sociedad, desarrollo y ciudadanía en México. pp. 229–250. Limusa (2008).
22. Esquivel, G.: Geografía y Desarrollo Económico en México. Banco Interam. Desarro. (2000).
23. Garza, G., Schteingart, M., Vilalta, C., Sobrino, J., Negrete Salas, M.E., Damián, A., Alegría, T., Chias, L., Reséndiz, H., García Palomares, J.C., Duhau, E., Giglia, Á., Ibarra, V., Salazar Cruz, C.E., Coulomb, R., Azuela, A., Sánchez Mejorada Fernández, C.: II Desarrollo urbano y regional. El Colegio de México, México (2010).
24. IBM: IBM SPSS Neural Networks 25 (2020).

Capítulo 4

Métodos de clasificación multiclase, RNA, XGBoost y BA: **una comparación para la selección de vivienda**

Resumen

Los métodos de clasificación juegan un papel importante en las tareas de toma de decisiones al clasificar registros de preferencias en función de algunos criterios. El objetivo de esta investigación es evaluar el desempeño de algunos métodos de clasificación de Inteligencia Artificial en relación con las preferencias de las características de vivienda de usuarios en México. Se comparan los resultados de los algoritmos de Redes Neuronales Artificiales, Aumento de Gradiente Extremo y Bosque Aleatorio mediante el uso de software SPSS Modeler y Machine Learning en IBM Watson Studio. Nuestra investigación sugiere que en el análisis de decisiones multicriterio es conveniente el estudio de diferentes métodos y de esta manera determinar cuál se adapta mejor a la estructura y rasgos de los datos. Los resultados del estudio pueden ayudar en el diseño de modelos de clasificación con variables de respuesta categórica mediante la comparación de diferentes métodos de clasificación para alcanzar consistencia y mayor precisión en los resultados de la clasificación.

Keywords: Decision Making, RNA, BA, XGBoost.

4.1 Introducción

El estudio de las directrices actuales del mercado de la vivienda es de es fundamental para los sectores público y privado, además de ayudar a los nuevos usuarios a tomar decisiones informadas sobre las opciones vivienda que mejor

se adapten a sus necesidades. La investigación actual sobre la clasificación de vivienda se centra en análisis comparativos de diferentes algoritmos de Inteligencia Artificial (IA). Sin embargo, pocos estudios han examinado en conjunto las variables cualitativas y cuantitativas en la clasificación vivienda desde la perspectiva del usuario ya que principalmente se basan en análisis de riesgo para inversores de vivienda [1][2], clasificación del preció de vivienda [3][4][5], satisfacción [6][7], densidad urbana [8], entre otros. Para abordar este problema, en este documento se expone una comparación de diferentes metodologías de aprendizaje supervisado orientadas a problemas de clasificación multiclase, basada en los datos de la Encuesta Nacional de Hogares (ENH) [9], con el objetivo de construir un modelo que permita identificar el tipo de vivienda en función de los criterios relacionados a las características de vivienda de usuarios de vivienda en México.

Una de las tareas primordiales para de toma de decisiones es la clasificación a través del análisis de problemáticas de predicción o reconocimiento de patrones. El diseño de modelos de clasificación ha tenido mayor difusión a partir del desarrollo de aplicaciones en línea ya que genera un intercambio de información que permite implementar nuevas tecnologías de procesamiento y análisis de datos, como las plataformas de código abierto para la adaptación de software y los sistemas para crear modelos de analítica predictiva y clasificación de datos. Los sistemas de clasificación y regresión permiten a los desarrolladores de aplicaciones el diseño de modelos que permitan, a los clientes de un producto o servicio, sugerencias en línea así como identificar las ventajas para la toma de decisiones [10].

Dado que cada método de clasificación tiene sus fortalezas y limitaciones y que los problemas del mundo real no siempre satisfacen los supuestos de un método en particular, un enfoque es aplicar todos los métodos apropiados y seleccionar el que proporcione la mejor solución [11]. Para esta investigación compara tres algoritmos de IA: Redes Neuronales Artificiales (RNA), Gradiente de Impulso Extremo (XGBoost) y Bosque Aleatorio (BA), los cuales permiten crear modelos de clasificación multiclase además de ser algoritmos que se adaptaron a la

estructura y rasgos de los datos, ya que en este caso se aborda un problema multivariable con un objetivo multiclase. Si bien los orígenes de estos enfoques son distintos y los algoritmos subyacentes difieren sustancialmente, el proceso fundamental es el mismo; todos son métodos de clasificación inductivos. El propósito es comparar la precisión en los diferentes métodos de clasificación en base a los resultados para obtener una mejor comprensión de los métodos de clasificación y construir un precedente de modelos de clasificación multiclase orientados a vivienda.

4.2 Metodología

4.2.1 Redes Neuronales Artificiales y Perceptrón Multicapa

Las Redes Neuronales Artificiales (RNA´s) son una rama específica de la Inteligencia Artificial. Se pueden definir como modelo informático de sinapsis artificial que simula el funcionamiento de las neuronas y sinapsis del cerebro humano para aprender. Su arquitectura está compuesta por un conjunto de neuronas interconectadas que emiten señales bidireccionalmente. La información de entrada fluye en la red a través de distintas operaciones dando como resultado una salida. Se construyen de forma autónoma adaptándose a los datos de entrada. Su difusión se ha dado por su precisión para resolver problemas complejos como el reconocimiento de patrones o el procesamiento del lenguaje natural, mediante el ajuste de los coeficientes de ponderación en una fase de aprendizaje.

MLP por sus siglas en inglés (Multilayer Perceptron), es un algoritmo de clasificación que utiliza una técnica de entrenamiento supervisado llamada Retropropagación (Backpropagation), el cual permiten clasificar datos que no son linealmente separables. Una RNA MLP se distingue por su arquitectura, compuesta por una capa de entrada, una o más capas ocultas, la capa de salida y por los tipos de funciones de activación con que trabaja. La información en la RNA MLP fluye entre las neuronas de una capa a las neuronas de la capa siguiente (feedforward) mediante conexiones que reproducen una sinapsis.

El modelo MLP extrae aleatoriamente n muestras de tuplas de entrada x_i y salida y_i de los conjuntos de entrenamiento $D = \{(x_i, y_i): i = 1 \ldots N, x_i \in \mathbb{R}^m, y_i \in \mathbb{R}\}$ donde m es el número de características de entrada x_i, y cada muestra n tiene un valorde etiqueta de la variable objetivo y_i . El proceso de entrenamiento del algoritmo Retropropagación es como sigue:

1. A cada característica de entrada se le asigna el mismo peso w_{ij} .
2. Las neuronas i de la capa de entrada propagan los datos de entrenamiento a las neuronas j de la capa oculta en dirección a la capa de salida. Las posibles tuplas de entrada-salida a un vector de características, de valor real de dimensión finita, que se multiplica por la ponderación sináptica w_{ji} . Cada neurona j está dada por:

$$y_j = \gamma\left(\sum_k w_{ji}x_i + \theta_j\right) \quad (1)$$

   Donde γ es la función de activación, θ_j es el sesgo el cual permite desplazar la función de activación añadiendo una constante a la entrada que se ajusta en el entrenamiento. Los pesos sinápticos entre la capa de entrada y la capa oculta se representan por w_{ji} . La variación de los pesos sinápticos se debe a la sensibilidad de la función de activación. El índice i varía entre 1 y el n queson el número de capas de la RNA.

3. Los pesos sinápticos entre la capa oculta y la capa de salida se representan por w_{kj} , y el sesgo en las neuronas de salida se representa por θ_k. La salidadeseada $\hat{y}_k$ esta dada por:

$$\hat{y}_k = \gamma\left(\sum_k w_{kj}x_i + \theta_k\right) \quad (2)$$

4. Se calcula el error de las salidas de todas las neuronas. En las neuronas k, seeleva el error al cuadrado porque interesa su magnitud, no su signo. El proceso de entrenamiento continua hasta que la suma del error cuadrático medio es el mínimo posible [12][13]. El error esta dado por:

$$E = \frac{1}{2}\sum_{k=1}^{m}[(\hat{y}_k - y_k)^2] \quad (3)$$

5. De acuerdo con el algoritmo Retropropagación, el entrenamiento iterativo utiliza la regla delta la cual sustenta que cada peso tiene una tasa de aprendizaje única, y ésta se puede ir modificando durante el entrenamiento [14] [15]. La regla delta es la técnica que actualiza los pesos sinápticos w, y de manera análoga el sesgo, θ, además de añadir una tasa de aprendizaje ε. El gradiente toma la trayectoria opuesta para identificar el decremento más rápido del error, entre cada comparación del objetivo de entrenamiento, y, y el pronóstico, $\hat{y}_k$. Cada uno de los pesos del conjunto de entrenamiento D se ajusta de la siguiente forma:

$$-\sum_{d=1}^{D}\frac{\partial E}{\partial w} \quad (4)$$

6. En cada iteración el entrenamiento actualiza los pesos cuando el resultado de $\hat{y}_k$ no coincide con el objetivo y_i . Mediante la regla de cadena y la propagación hacia atrás, la actualización de los pesos Δw_{kj} (neuronas de salida) y Δw_{ij} (neuronas ocultas) se determinan de la siguiente manera:

$$\Delta w_{kj}(n+1) = -\varepsilon\frac{\partial E}{\partial w_{kj}} = \varepsilon\sum_{d=1}^{D}\delta_k y_j \quad (5)$$

Donde d indica la muestra x_i del conjunto de entrenamiento y $\delta_k = (\hat{y}_k - y_k)$ y $\Delta w_{ij} = \varepsilon y_i x_j$ para $0 < \varepsilon < 1$ que representa la tasa de aprendizaje.

$$\Delta w_{ji}(n+1) = -\varepsilon\frac{\partial E}{\partial w_{ji}} = \varepsilon\sum_{d=1}^{D}\delta_j x_i \quad (6)$$

Para $\delta_j = y_j\sum_{k=1}^{N}\delta_k w_{kj}$

4.2.2 Aumento de gradiente Extremo (XGBoost)

XGBoost es uno de los algoritmos de aprendizaje supervisado de las máquinas de aumento de gradiente (GBM) las cuales se basan en los métodos de Árboles de Clasificación y Regresión (CART por sus siglas en ingles) que se basan el principio de entrenar variables débiles con entrenamiento iterativo para conseguir que el algoritmo identifique los datos incorrectamente clasificados por su predecesor [16]. Por esta razón, el termino Boosting (aumento) en la denominación de método, hace referencia a un tipo de algoritmos cuya finalidad es encontrar una hipótesis fuerte a partir de impulsar hipótesis simples y débiles. Es decir, Boosting genera una regla de predicción muy precisa mediante la combinación de reglas de predicción aproximadas [17]. La categorización multi-clase busca características comunes en las categorías de la variable a predecir por lo que es necesario definir previamente una medida del error conjunto. Durante cada iteración el algoritmo escoge un clasificador de una sola característica con el propósito de identificar las características que son comunes por más categorías, introduciendo una penalización del error de las categorías donde la característica del clasificador está ausente [15]. En cada iteración el árbol crece aprendiendo una nueva función para ajustar el error. XGBoost optimiza la tasa de error ε sobre todas las clases [18].

$$\hat{y}_i = \varepsilon \sum_{j}^{J} f_j\,(x_i), \qquad f_j \in \mathrm{F} \tag{7}$$

El algoritmo itera con la adición constante de árboles transformando las características para hacer crecer un árbol a partir de la predicción previa. A cada nodo hoja corresponde a una puntuación que se suma a la puntuación correspondiente a un árbol, que es el valor predicho de la muestra [8]. La idea del algoritmo Boosting es hacer un muestreo aleatorio con reemplazo, es decir, el algoritmo extrae una muestra aleatoria de las observaciones de los datos de entrenamiento y la retorna a esa población, después de aprender sus características y antes de extraer la siguiente muestra. Algunas observaciones pueden repetirse en cada nuevo conjunto de datos de formación, sin embargo

Boosting pondera las observaciones en cada iteración y por lo que solo las de mayor ponderación se fusionarán más veces en nuevos conjuntos. Boosting basa su entrenamiento en los datos, considerando el éxito de las iteraciones anteriores y de esta forma redistribuir los pesos [19]. Los datos mal clasificados aumentan sus pesos para identificar los casos más difíciles y así en las siguientes iteraciones el algoritmo se centrará en ellos durante su entrenamiento Este proceso se sigue iterativamente hasta que el error se minimiza. Elegimos el subconjunto con el menor error en el conjunto de entrenamiento ponderado para todas las clases. El algoritmo XGBoost sigue las siguientes etapas [15][20]:

1. Cada característica recibe el mismo peso w$_j$.
2. El Algoritmo Boosting selecciona n muestras de los conjuntos de datos de entrenamiento $D = \{(x_i, y_i): i = 1 \dots N, x_i \in \mathbb{R}^m, y_i \in \mathbb{R}\}$. Cada muestra n tiene m características x_i, y un valor de etiqueta de la variable dependiente y_i . Sí se tiene un total de K árboles en el modelo de ensamble, la $\hat{y}_i$ esta dada por:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), \qquad f_k \in F \tag{8}$$

Donde:

$F = \{f(x) = w_q q(x)\}(q : \mathbb{R}^m \to T, w \in \mathbb{R}^T)$; f_k representa una estructura independiente de árbol q con peso w_q; x_i corresponde al vector de características de la muestra. $f_k(x_i)$ denota la puntuación de la i-ésima muestra en k-ésimo árbol y T representa el número de hojas en el árbol. La función objetivo regularizada es la suma de dos partes: la pérdida de formación y la regularización:

$$Obj = \sum_{i=1}^{N} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{9}$$

Donde:

$l(y_i, \hat{y}_i)$ es la función de pérdida (error) que mide la diferencia entre el objetivo y_i y la predicción $\hat{y}_i$ de los conjuntos de entrenamiento.

$\Omega(f_k)$ es la regularización que registra la complejidad del modelo y se determina por:

$$\Omega(f_k) = \gamma T + \frac{1}{2}\,\lambda \sum_{j=1}^{T} {w_j}^2 \tag{10}$$

En la que T es la puntuación de los nodos de hoja, el índice de cada nodo hoja es j, y el peso de la hoja es wj . La variable γ controla la complejidad de la estructura del árbol, y la variable λ es el término de regularización utilizado para controlar la distribución del peso de los nodos de hoja para evitar el sobreajuste [11].

3. El algoritmo XGBoost está diseñado para asociar la función objetivo con la estructura del árbol y luego establecer una relación directa entre la estructura del árbol y el efecto del modelo. El modelo se entrena de forma aditiva para minimiza la función de pérdida $l(y_i, \hat{y}_i)$. Por tanto, a la función de pérdida se le suma la función f_t en la t-ésima iteración para minimizar la función objetivo. En la t-ésima iteración el objetivo viene dado por:

$$L^{(t)} = \sum_{i=1}^{n} l\left[y_i, {\hat{y}_i}^{(t-1)} + f_t(x_i)\right] \sum_{k=1}^{K} \Omega\,(f_k) \tag{11}$$

4. Se realiza una aproximación de segundo orden de la función de pérdida mediante series de Taylor, que puede utilizarse para optimizar el objetivo en el entorno general, el objetivo final viene dado por:

$$L^{(t)} \simeq \sum_{i=1}^{n} l\left[y_i, {\hat{y}_i}^{(t-1)} + g_i\, f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)\right] + \Omega(f_k) \tag{12}$$

Donde, g_i es la primera derivada y h_i es la segunda derivada de la función de perdida.

5. La función objetivo se transforma en la iteración del mínimo de la ecuación cuadrática unidimensional. Si la estructura de árbol q es fija, la iteración sobre el modelo de árbol se transforma en una iteración sobre los nodos hoja del árbol, donde a se deriva para wj , y el peso óptimo de

la hoja j se calcula por:

$$w_j = -\frac{\sum_{i\epsilon I_j} g_i}{\sum_{i\epsilon I_j} h_i + \lambda} \tag{13}$$

6. El valor de pérdida correspondiente a cada nodo hoja puede caracterizarse por la siguiente función de pérdida:

$$\tilde{L}^{(t)}(q) = -\frac{1}{2}\sum_{j=1}^{T} \frac{\left(\sum_{i\epsilon I_j} g_i\right)^2}{\sum_{i\epsilon I_j} h_i + \lambda} + \lambda T \tag{14}$$

7. En general, no es posible definir todos los árboles viables y determinar cuál es el óptimo; el algoritmo comienza con un único nodo hoja y lo divide iterativamente para añadir ramas al árbol. Supongamos que I_L y I_R son conjuntos de nodos izquierdos y derechos después de la división. Dejando que $I = I_L \cup I_L$ entonces la reducción de la pérdida junto a la división se calcula por:

$$\tilde{L}^{(t)} = \frac{1}{2}\left[\frac{\left(\sum_{i\epsilon I_L} g_i\right)^2}{\sum_{i\epsilon I_L} h_i + \lambda} + \frac{\left(\sum_{i\epsilon I_R} g_i\right)^2}{\sum_{i\epsilon I_R} h_i + \lambda} - \frac{(\sum_{i\epsilon I} g_i)^2}{\sum_{i\epsilon I} h_i + \lambda}\right] + \gamma \tag{15}$$

4.2.3 Bosque Aleatorio

BA es una técnica de aprendizaje supervisado que se basa en la generación de múltiples árboles de decisión a partir de un conjunto de datos de entrenamiento donde cada árbol obtiene una predicción de clase y la clase con más votos se convierte en la predicción ganadora. El algoritmo de BA genera cada uno de los árboles de decisión a partir de la selección aleatoria con remplazo de variables de distintos conjuntos de datos de entrenamiento a través del algoritmo de ensamble Banging, también conocido como Booststrap. Banging es una técnica de estadística inferencial que crea nuevos conjuntos de datos simplificados, donde una característica se puede considerar en varios conjuntos. De esta manera se obtienen una variedad de árboles con apariencia una caótica pero producen una salida concreta [21]. Los conjuntos de datos que no se utilizan en la generación de árboles aleatorios, denominados muestras OOB (out of the

bag), se usan para validar el modelo, a partir de la permutación de predictores. La selección de predictores a permutar se basa en a la influencia (importancia) que tienen los predictores en cada árbol sobre la clasificación de salida que es estimada por el error OOB o por validación cruzada. La contribución del predictor permutado se pierde y en consecuencia el error del modelo aumenta [22]. OOB es el error del árbol para las observaciones que no forman parte del conjunto de entrenamiento y está ligado a la importancia de los predictores. Durante el aprendizaje, cada clase obtiene un voto cuando el valor de $\hat{y}_t$ coincide con una de las clases OOB y_i . La estimación del error OOB se determina por la proporción de veces que $\hat{y}_t$ es distinta a la clase verdadera de y_i . Las salidas de todos los árboles se ensamblan en una salida final Y que se obtiene mediante el promedio, cuando las salidas de los árboles del ensamblado son continuas y, por votación, cuando son categóricas. El algoritmo BA funciona de la siguiente manera:

- Considere el vector x ∈ $\mathbb{R}^m$ con N variables x =$[x_1, x_2, x_3, \dots, x_N]$.
- Para problemas de clasificación el objetivo de un modelo predictivo es identificar la clase que genera un caso concreto y asignarlo a una de las M clases y_i de acuerdo con los valores de n variables.
- Mediante el aprendizaje supervisado, el modelo construye conjuntos de entrenamiento $D = \{(x_i, y_i), \dots, (x_D, y_D)\}$, formado por d tuplas, (x_i, y_i); cada tupla contiene un vector x_i con m variables [f$_{i1}$, f$_{i2}$, f$_{i3}$, … , f$_{im}$] y un valor de clase del objetivo y_i ϵ [0,1].
- Las regresiones en BA buscan la clase más dominante entre las predicciones de los árboles individuales. Si hay T árboles, el número de votos de una clase M es:

$$v_o = \sum_{t=1}^{T} \iota\,(\hat{y}_t == o) \tag{16}$$

- Donde, $\hat{y}_t$ es la predicción del t-ésimo árbol en una iteración particular. El indicador de la función $l(\hat{y}_t == o)$ toma el valor de 1 si se cumple el

objetivo, de otra forma es 0.

$$\hat{y}_t = \arg\max\ v_o \qquad o \in \{1, \dots, O\} \tag{17}$$

- El índice de pureza de Gini determina la división optima del nodo raíz y los nodos subsecuentes a partir de la selección aleatoria de una variable. El índice de Gini representa la probabilidad de que una característica determinada sea mal clasificada cuando se elige al azar. Solo se usa para objetivos categóricos. Su valor es cero cuando todas las características del nodo corresponden a una única categoría objetivo. La forma de calcularlo es la siguiente:

$$Gini = \sum_{i=1}^{N} P_{x_i}\left(1 - P_{x_i}\right) \tag{18}$$

4.3 Modelos RNA, XGBoost y BA

4.3.1 Selección de características y preprocesamiento de datos

La construcción de los modelos de clasificación para la predicción del tipo de vivienda que se realizan en los algoritmos RNA, XGBoost y BA, se basa en los datos de usuarios de vivienda de la Ciudad de México y el Estado de México de las ediciones de ENH de 2014 a 2017, se eligió trabajar con los conjuntos de datos de esta región por presentar una mayor difusión de la clasificación de los tipos de vivienda que se documentan en la ENH. La ENH se divide en tres apartados: Vivienda, Hogar y Persona. Para investigación se utiliza exclusivamente el apartado correspondiente a vivienda, en el que Cuestionario Básico de la ENH recaba la información de 109 características asociadas a los atributos de vivienda, la ubicación y la conformación de los hogares en México.

Para limpieza de la base de datos se excluyeron los registros con errores en los datos y los que presentan ausencia de información. Al concluir este proceso se contabilizaron un total de 10,950 registros. El siguiente paso fue la selección de características (ver Apéndice I). En Tabla 4.1 se presentan las características seleccionadas clasificadas de acuerdo con su tipo.

Tabla 4.1. Descripción de variables.
(Fuente: Elaboración propia con información de la ENH, INEGI (2017)).

Valores de Medición de Variables					
#	Nombre	Valores y Etiquetas	#	Nombre	Valores y Etiquetas
Variables Categóricas					
1	tipo_viv	1 Casa independiente 2 Departamento en condominio vertical 3 Vivienda en vecindad 4 Vivienda en cuarto de azotea 5 Local no construido para habitación	6	drenaje	1 La red pública 2 Una fosa séptica 3 Una tubería que va a dar a una barranca 4 Una tubería que va a dar a un río, lago o mar 5 No tiene drenaje
2	tenencia	1 Es rentada 2 Es prestada 3 Es propia, pero la están pagando 4 Es propia 5 Está intestada o en litigio 6 Otra situación	7	disp_agua	1 Agua entubada dentro de la vivienda 2 Agua entubada pero dentro del terreno 3 Agua entubada de llave pública (o hidrante) 4 Agua entubada que acarrean de otra vivienda 5 Agua de pipa 6 Agua de un pozo, río, lago, arroyo u otra
3	tipo_finan	1 Recursos propios 2 Apoyo de FONHAPO 3 Crédito INFONAVIT o FOVISSSTE 4 Crédito bancario 5 Crédito micro financiero 6 Crédito caja de ahorro 7 Crédito de otra institución 8 Préstamo familiar	8	eli_basura	1 Se la dan a un camión o carrito de basura 2 La llevan al basurero público 3 La dejan en un contenedor o depósito 4 La queman 5 La entierran 6 La tiran en otro lugar (calle, baldío) 7 La tiran en la barranca o grieta 8 La tiran al río, lago o mar
4	mat_pisos	1 Tierra 2 Cemento o firme 3 Madera, mosaico u otro recubrimiento	9	escrituras	1 A nombre del dueño 2 A nombre de otra persona 3 No tiene escritura 9 No sabe
5	mat_techos	1 Material de desecho 2 Lámina de cartón 3 Lámina metálica 4 Lámina de asbesto 5 Palma o paja 6 Madera o tejamanil 7 Terrado con viguería 8 Teja 9 Losa de concreto, viguetas y bovedi-llas	10	mat_pared	1 Material de desecho 2 Lámina de cartón 3 Lámina de asbesto o metálica 4 Carrizo, bambú o palma 5 Embarro o bajareque 6 Madera 7 Adobe 8 Tabique, ladrillo, block, piedra u otro
			11	ubica_geo	Clave de la localidad.
Variables Ordinales					
12	tam_loc	1 Localidades 100 000 y más habitantes 2 Localidades 15 000 a 99 999 habitan-tes 3 Localidades 2 500 a 14 999 habitantes 4 Localidades menos de 2 500 habitan-tes	13	est_socio	1 Bajo 2 Medio bajo 3 Medio alto 4 Alto
Variables Continuas					
14	folio_viv	Identificador de la vivienda	18	cuart_ dorm	Número de cuartos de la vivienda para dormir
15	bano_comp	Baños con excusado y regadera.	19	num_cuarto	Número total de cuartos que tiene la vivienda
16	bano_excus	Número de baños sólo con excusado.	20	tot_resid	Número de residentes de la vivienda
17	bano_regad	Número de baños sólo con regadera.	21	tot_hog	Número de hogares en la vivienda
Variables Dicotómicas					
22	const_dorm	1 Sí, 2 No	29	cocina	1 Sí, 2 No
23	oonst_coci	1 Sí, 2 No	30	lavadero	1 Sí, 2 No
24	oonst_bano	1 Sí, 2 No	31	tinaco_azo	1 Sí, 2 No
25	repar_pard	1 Sí, 2 No	32	calentador	1 Sí, 2 No
26	repar_tech	1 Sí, 2 No	33	repar_agua	1 Sí, 2 No
27	repar_dren	1 Sí, 2 No	34	aire_acond	1 Sí, 2 No
28	repar_cabl	1 Sí, 2 No		calefacción	1 Sí, 2 No

4.3.2 Diseño de los Modelos RNA, XGBoost y BA

El diseño del fujo de los modelos RNA , XGBoost y BA en SPSS Modeler sigue

la misma configuración con la diferencia del último nodo que especifica el algoritmo con el que se trabaja en la interfaz virtual Watson Studio. La Figura 4.1 muestra el flujo del procesamiento de datos para la construcción de los modelos mediante el uso de diferentes nodos secuenciales que representan las herramientas del software, los cuales desempeñan una a tarea en específica y se diferencian por un icono particular.

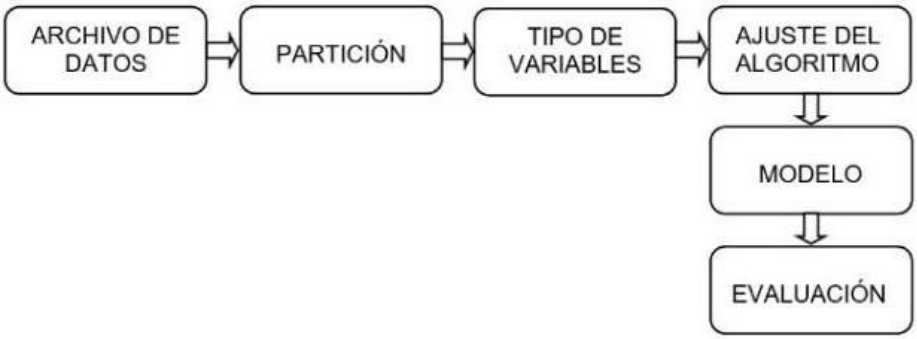

Figura 4.1. Diagrama del flujo de procesamiento de datos para el desarrollo de los modelos.(Fuente: elaboración propia con información de IBM SPSS Modeler (2021).

El diseño del flujo de los modelos RNA, XGBoost y BA se construye en la interfaz de SPSS Modeler Flows mediante el siguiente proceso:

- Se utiliza el nodo *Activo de Datos* para seleccionar los arhivos Excel. CSV quecontienen los conjuntos de datos.
- Se añade el nodo *Partición* para establecer la partición de los conjuntos de datos en tres partes: 60% para entrenamiento, 30% para prueba y 10% validación; éste último se utiliza para determinar el fin del entrenamiento del algoritmo para evitar el sobre entrenamiento.
- Se incorpora el nodo *Tipo* para describir cada una de las variables de acuerdo con su tipo (continua, nominal, categórica, etc.), el rol que desempeñan en el sistema de hipótesis (dependiente, independiente, identificación de registro, etc.) y el valor de medición de cada una de las variables. El algoritmo norma- liza los registros de las variables de entrada a valores entre cero y uno para mejorar su desempeño.
- En cada caso se usan los nodos RNA, XGBoost y BA, que se comparan en esta investigación. En estos nodos se eligen, las variables de entrada,

el nivel de confianza de las predicciones, el número de ciclos, entre otras opciones que se desee considerar. En esta investigación, seleccionamos las 33 variables predictoras, la variable objetivo, 1000 ciclos de iteración y un conjunto de prevención de sobreajuste del 10% para rastrear errores durante el entrenamiento que induzcan al algoritmo a modelar la probabilidad de los datos en vez de clasificarlos.

- Por último se seleccionan los nodos del apartado de resultados de la interfaz que se deseen analizar. Para esta investigación se utiliza el nodo *Análisis* que genera un informe de las métricas del modelo.

4.4 Evaluación del rendimiento

Luego de entrenar cada uno de los algoritmos se generan los resultados con las estadísticas de número y proporción de clasificaciones correctas y erróneas de las etapas de entrenamiento, prueba y validación, como se muestra en la Tabla 4.2.

Tabla 4.2. Clasificaciones correctas y erróneas. (Fuente: elaboración propia).

Modelo	Clasificaciones	Entrenamiento		Prueba		Validación	
RNA	Correctas	1.280	92,96%	383	88,86%	175	91,15%
	Erróneas	97	7,04%	48	11,14%	17	8,85%
	Total	1.377		431		192	
XGBoost	Correctas	1.292	93,83%	381	88,4%	173	90,1%
	Erróneas	85	6,17%	50	11,6%	19	9,9%
	Total	1.377		431		192	
BA	Correctas	1.357	98,55%	395	91,65%	179	93,23%
	Erróneas	20	1,45%	36	8,35%	13	6,77%
	Total	1.377		431		192	

4.4.1 Índices de Evaluación

El número de clasificaciones Correctas y Erróneas forman las matrices de confusión de las cuales se obtiene el desglose de las clasificaciones Correctas y Erróneas en Verdaderos Positivos (VP), Verdaderos Negativos (VN), Falsos Positivos (FP) y Falsos Negativos (FN), de las tres etapas de aprendizaje de los algoritmos. Las matrices de confusión de la etapa de entrenamiento se muestran en la Tabla 4.3, en la que destaca observar la aleatoriedad en elección de las muestras de entrenamiento del modelo BA.

Tabla 4.3. Matrices de Confusión, etapa de entrenamiento. (Fuente: elaboración propia).

RNA	Pronosticado				
Observado	**1**	**2**	**3**	**5**	**Correctos (%)**
1	1184	0	0	0	100.00%
2	89	96	0	0	51.89%
3	5	0	0	0	0.00%
5	2	1	0	0	0.00%
Correctos (%)	96.30%	69.20%	0.00%	0.00%	93.02%
XGBoost	**Pronosticado**				
Observado	**1**	**2**	**3**	**5**	**Correctos (%)**
1	1169	0	0	0	100.00%
2	77	123	0	0	61.50%
3	5	0	0	0	0.00%
5	3	0	0	0	0.00%
Correctos (%)	93.22%	100.00%	0.00%	0.00%	93.83%
BA	**Pronosticado**				
Observado	**1**	**2**	**3**	**5**	**Correctos (%)**
1	1199	0	0	0	100.00%
2	19	151	0	0	88.82%
3	0	0	5	0	100.00%
5	1	0	0	2	66.67%
Correctos (%)	98.36%	100.00%	100.00%	100.00%	98.55%

Los valores de los resultados de la clasificación de las matrices de confusión se utilizan para determinar los índices de evaluación de los modelos para cada una de las etapas de entrenamiento los cuales permiten conocer la exactitud y precisión del desempeño de los modelos y representan una medida de referencia para poder comparar distintos algoritmos. Los índices de evaluación de los modelos son los siguientes:

Exactitud. Valor predictivo positivo que simboliza la proporción de predicciones correctas del modelo. El cálculo de precisión (*P*) está dado por:

$$Accuracy = \frac{TP + TN}{(TP + TN + FP + FN)} \tag{19}$$

Precisión. indica la tasa de replica de las de las aproximaciones positivas de la clasificación. Su valor esta dado por:

$$Precision = \frac{VP}{(VP + FP)} \tag{20}$$

Exhaustividad. Es una medida de sensibilidad que determina la proporción de predicciones correctas en clase positiva. La exhaustividad está dada por:

$$Recall = \frac{VP}{(VP + FN)} \tag{21}$$

Medida F1. Es un promedio armónico que provee el balance de precisión y exhaustividad. La medida F1 ponderada está dada por:

$$F1 = 2 * \frac{(Precision * Weighted\ recall)}{(Precisión + Weighted\ recall)} \tag{22}$$

Los índices de evaluación de las etapas de entrenamiento de los modelos RNA, XGBoost y BA se presentan en la Tabla 4.4. Los índices de cada etapa de entrenamiento permiten conocer el comportamiento del aprendizaje de los algoritmos, sin embargo los resultados de la etapa de prueba tienen mayor relevancia ya que en esta etapa se utiliza el reconocimiento de patrones, el sesgo y los pesos de las variables predictoras que determinan una salida, los cuales fueron identificados por el algoritmo en la etapa inicial del entrenamiento. Luego, en la etapa de validación, el algoritmo trabaja con los datos que no ha utilizado en las etapas previas para identificar otros patrones, ajustar los pesos y los sesgos, optimizar parámetros de entrenamiento, disminuir el error, entre otras operaciones. De acuerdo con los resultados de la etapa de prueba de la Tabla 4.4, el modelo de RF obtiene las puntuaciones más altas de los índices de evaluación, aunque la diferencia no es significativa comparada con los modelos de la RNA y XGBoost, además de que ambos muestran un mayor incremento en las puntuaciones de validación con respecto a las puntuaciones de la etapa de prueba.

Tabla 4.4. Índices de evaluación de modelos. (Fuente: elaboración propia).

Índices de Evaluación	**RNA**	**XGBoost**	**BA**
Entrenamiento			
Exactitud	0.9325	0.94118	0.9906
Precisión	0.9296	0.93827	0.9855
Exhaustividad	0.9350	0.94375	0.9855
F1	0.9323	0.94101	0.9855
Prueba			
Exactitud	0.8933	0.88863	0.9211
Precisión	0.8886	0.88399	0.9165
Exhaustividad	0.8970	0.89227	0.9251
F1	0.8928	0.88811	0.9207
Validación			
Exactitud	0.9141	0.90365	0.9349
Precisión	0.9115	0.90104	0.9323
Exhaustividad	0.9162	0.90576	0.9372
F1	0.9138	0.90339	0.9347

4.4.2 Curvas ROC

La curva ROC (por sus siglas en Ingles Receiver Operating Characteristic) es un método estadístico que se basa en el umbral de discriminación entre la proporción de VP y la proporción FP. El parámetro AUC (Area Under The Curve), mide el área bajo la curva y mide el rendimiento del algoritmo, de manera que entre más próximo a 1 sea su valor, mejor es el desempeño del modelo. Las Figuras 4.2, 4.3 y 4.4 muestran Las curvas ROC y su AUC de la etapa de validación de los modelos RNA, XGBoost y BA. Se observa que el modelo BA tiene un AUC mayor que los modelos RNA y XGBoost. Las curvas ROC muestran los puntos de corte de la ejecución del algoritmo. Para los algoritmos XGBoost y BA se presentan las curvas ROC de cada clase, ya que fueron realizadas en la herramienta *Auto-AI* de Watson Machine Learning, módulo de IBM Watson Studio.

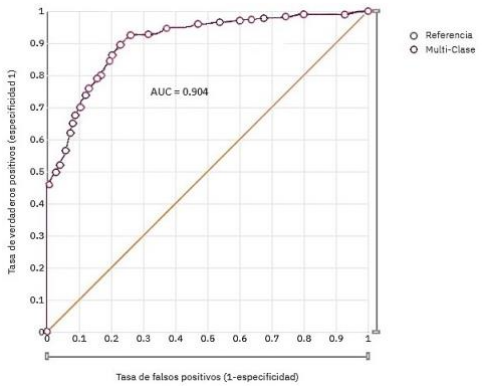

Figura 4.2. Curva ROC del modelo RNA. (Fuente: elaboración propia con información de IBM SPSS Modeler (2021)).

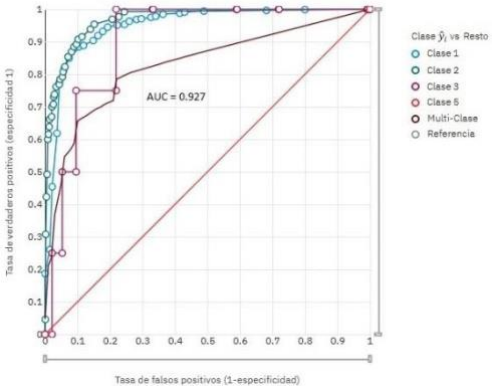

Figura 4.3. Curva ROC del modelo XGBoost. (Fuente: elaboración propia en IBM Watson Machine Learning (2021)).

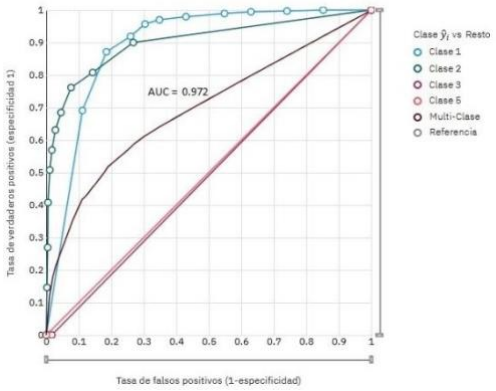

Figura 4.4. Curva ROC del modelo BA. (Fuente: elaboración propia en IBM Watson Machine Learning (2021)).

Otro resultado de salida basado en la curva ROC es el valor de corte óptimo para la prueba. Los puntos de corte dicotomizan los valores de la prueba, por lo que la prueba proporciona el diagnóstico (verdadero o falso). La identificación del valor de corte se determina a través la prueba diagnóstica de exactitud conocida como índice de Youden, que representa una medida simplificada de la curva ROC. El indice de Youden calcula la diferencia máxima de la proporción de VP (exhaustividad) y la proporción de FP (especificidad) −1. Se calcula por:

$$J = \frac{VP}{VP + FN} + -1 \tag{23}$$

El valor máximo del índice de Youden es 1 (prueba perfecta) y el mínimo es 0, cuando la prueba no tiene valor diagnóstico. El mínimo se produce cuando la $exhaustividad$ = 1 − $espcificidad$, representada por la línea diagonal en la gráfica ROC. Para esta investigación en la Figura 4.5 el modelo BA obtuvo el índice de Youden con el valor más alto de 0.854, seguido del modelo XGBoost con 0.768 y por último el modelo RNA con un índice de 0.684.

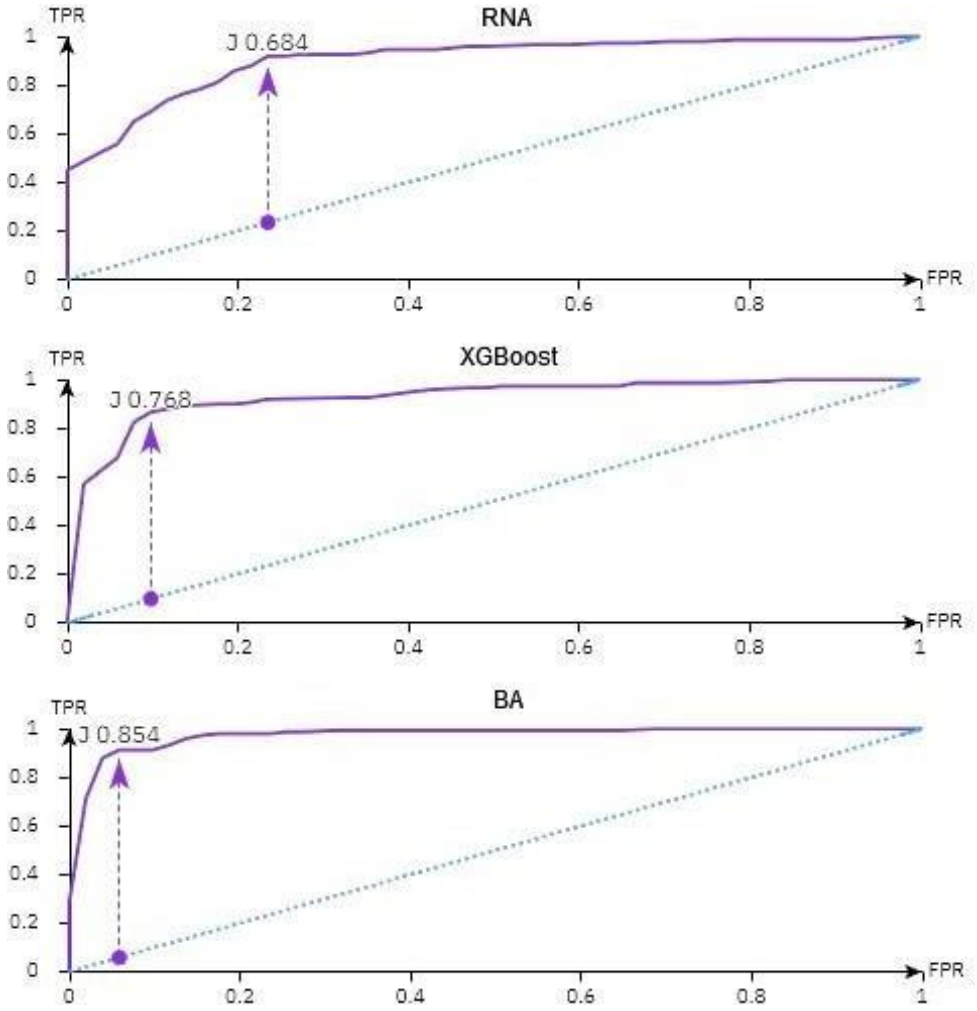

Figura 4.5. Curvas ROC con los Índices de Youden de los modelos RNA, XGBoost y BA. (Fuente: elaboración propia con información de IBM SPSS Modeler (2021)).

4.4.3 Importancia de los predictores

En la interpretación de la importancia de las variables se busca que el modelo muestre la capacidad de identificar las variables que influyen significativamente en el resultado [23], es decir, se analiza si el proceso de aprendizaje puede convertir en reglas las relaciones lógicas de las variables con la salida [24][25]. La importancia de los predictores presenta similitudes entre las características con mayor peso como se muestre en la Tabla 4.5, donde las características *baño excusado*, *ubicación geográfica* y *baño completo* están entre las diez más importantes en los modelos que se comparan.

Tabla 4.5. Importancia de los predictores. (Fuente: elaboración propia con información de IBMSPSS Modeler (2021) e IBM Watson Machine Learning (2021)).

Importancia de los predictores					
RNA	**%**	**XGBoost**	**%**	**BA**	**%**
Baño excusado	7.99	Baño excusado	17.60	Total de residentes	13.10
Total de hogares	6.96	Material de pisos	13.70	Ubicación geográfica	11.90
Ubicación geográfica	6.65	Estrato socioeconómico	11.30	Estrato socioeconómico	10.20
Número de cuartos	6.28	Ubicación geográfica	7.20	Baño completo	8.60
Cuartos dormitorio	5.14	Eliminación de basura	6.00	Tipo de financiamiento	6.90
Eliminación de basura	5.02	Baño completo	5.30	Número de cuartos	6.50
Baño regadera	4.72	Tenencia	4.80	Baño excusado	5.20
Baño completo	4.62	Construcción de baño	3.90	Escrituras	5.00
Tamaño de la localidad	3.87	Tinaco azotea	3.70	Reparación de techos	3.70
Drenaje	3.76	Reparación de paredes	2.60	Tamaño de la localidad	3.30
Suma	55.01		76.10		66.87

En la Figura 4.6 se utiliza un diagrama de Venn para mostrar las variables de mayor importancia que son comunes en las tres modelos comparados. Al comparar los modelos, se puede evidenciar por lógica empírica que la similitud de las variables que convergen como las de mayor importancia en cada modelo fundamenta la confianza en el desempeño de los modelos a pesar de su complejidad para poder interpretarlos.

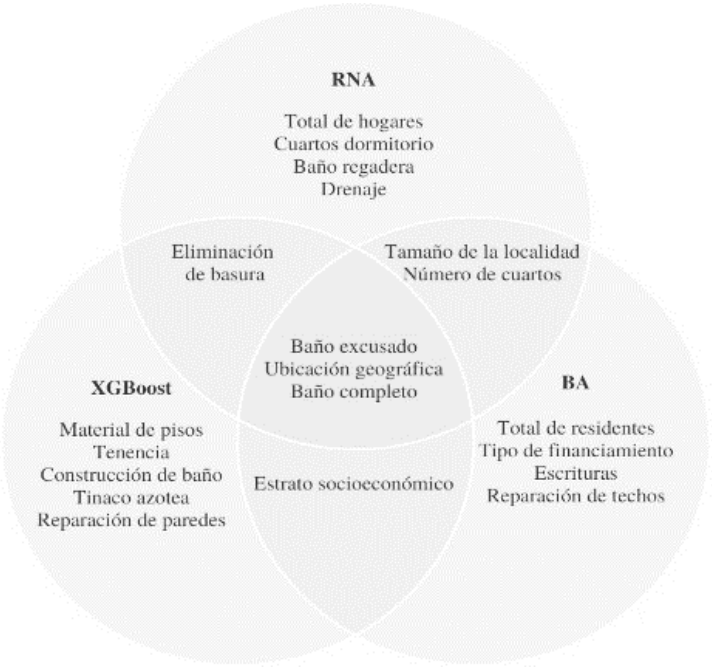

Figura 4.6. Variables importantes comunes en los modelos. (Fuente: elaboración propia).

4.4 Conclusiones

La predicción de la clasificación de los tipos de vivienda que se presenta muestra que desempeño de los modelos RNA, XGBoost y BA cumple con el objetivo de Clasificar los tipos de vivienda utilizando estas metodologías de clasificación multiclase, los conjuntos de entrenamiento de la base de datos de vivienda de la ENH se adaptaron a las tareas de los tres modelos. Al inicio de la investigación se tenía incertidumbre por que los conjuntos de datos contienen variables categóricas, continuas y dicotómicas, sin embargo, la estructura de la base de datos y las herramientas de descripción y registro de variables del software utilizado, permitieron obtener resultados satisfactorios y de fácil comprensión, no obstante se considera conveniente trabajar en la clasificación de los tipos de vivienda con el propósito de ampliar la gama clases, con una descripción más detallada que contemplen atributos como el tamaño del terreno, los años de construcción, acabados de carpintería, celdas solares, entre otros. De esta forma el estudio de la vivienda podría alcanzar mayor influencia en las políticas públicas y permitiría ampliar su estudio de manera que el desarrollo humano y urbano de las ciudades sea beneficiado.

4.5 Referencias

1. Han, S., Ko, Y., Kim, J., Hong, T.: Housing Market Trend Forecasts through Statistical Comparisons based on Big Data Analytic Methods. J. Manag. Eng. 34, (2018). https://doi.org/10.1061/(asce)me.1943-5479.0000583
2. Matthys, C., et al: Prediction of Credit Risks in Lending Bank Loans. Expert Syst. Appl. 83, 1850–1854 (2019)
3. Wu, H., Jiao, H., Yu, Y., Li, Z., Peng, Z., Liu, L., Zeng, Z.: Influence factors and regression model of urban housing prices based on internet open access data. Sustain. 10, (2018). https://doi.org/10.3390/su10051676
4. Rahman, S.N.A., Maimun, N.H.A., Razali, M.N., Ismail, S.: The artificial neural network model (ANN) for Malaysian housing market analysis. Plan. Malaysia. 17, (2019). https://doi.org/10.21837/pmjournal.v17.i9.581
5. Sun, Z., Pedretti, G., Bricalli, A., Ielmini, D.: One-step regression and classification with cross-point resistive memory arrays. Sci. Adv. 6, (2020). https://doi.org/10.1126/sciadv.aay2378
6. Gür, M., Murat, D., Sezer, F.Ş.: The effect of housing and neighborhood satisfaction on perception of happiness in Bursa, Turkey. J. Hous. Built Environ. 35, 679–697 (2020). https://doi.org/10.1007/s10901-019-09708-5
7. Paris, D.E.: The use of artificial intelligence as a decision support system for residential urban environments. En: Proceedings of the Annual Southeastern Symposium on System Theory (2005)
8. Bhooshan, S., Vazquez, A.N.: Homes, Communities and Games: Constructing Social Agency in Our Urban Futures. Archit. Des. 90, (2020). https://doi.org/10.1002/ad.2569
9. INEGI: Encuesta Nacional de los Hogares 2017, https://www.inegi.org.mx/programas/enh/2017/
10. Vishniakou, U.A.: Internet marketing organization with the use of intellegent and block chain technologies. «System Anal. Appl. Inf. Sci. (2020). https://doi.org/10.21122/2309-4923-2020-1-18-23
11. Li, N., Li, B., Gao, L.: Transient Stability Assessment of Power System Based on XGBoost and Factorization Machine. IEEE Access. 8, (2020). https://doi.org/10.1109/ACCESS.2020.2969446
12. Cinar, A.C.: Training Feed-Forward Multi-Layer Perceptron Artificial Neural Networks with a Tree-Seed Algorithm. Arab. J. Sci. Eng. 45, 10915–10938 (2020). https://doi.org/10.1007/s13369-020-04872-1
13. Palmer, A., Montaño, J.J., Jiménez, R.: Tutorial sobre Redes Neuronales Artificiales: El Perceptrón Multicapa. Psicologia.com. 5, (2001)
14. Jacobs, R.A.: Increased rates of convergence through learning rate adaptation. Neural Networks. 1, (1988). https://doi.org/10.1016/0893-6080(88)90003-2
15. Agahian, S., Akan, T.: Battle royale optimizer for training multi-layer perceptron. Evol. Syst. (2021). https://doi.org/10.1007/s12530-021-09401-5
16. Chakraborty, D., Elzarka, H.: Advanced machine learning techniques for building performance simulation: a comparative analysis. J. Build. Perform. Simul. 12, 193–207 (2019). https://doi.org/10.1080/19401493.2018.1498538

17. Freund, Y., Schapire, R.E.: A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. J. Comput. Syst. Sci. 55, (1997). https://doi.org/10.1006/jcss.1997.1504
18. Torralba, A., Murphy, K.P., Freeman, W.T.: Sharing visual features for multiclass and multiview object detection. IEEE Trans. Pattern Anal. Mach. Intell. 29, (2007). https://doi.org/10.1109/TPAMI.2007.1055
19. Janitza, S., Hornung, R.: On the overestimation of random forest's out-of-bag error. PLoS One. 13, (2018). https://doi.org/10.1371/journal.pone.0201904
20. Chen, T., Guestrin, C.: XGBoost: A scalable tree boosting system. En: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2016)
21. Bagnato, J.I.: Algoritmo k-Nearest Neighbor | Aprende Machine Learning
22. Joaquín Amat Rodrigo: Árboles de decisión, random forest, gradient boosting y C5.0
23. Deng, H.: Interpreting tree ensembles with inTrees. Int. J. Data Sci. Anal. 7, (2019). https://doi.org/10.1007/s41060-018-0144-8
24. Doshi-Velez, F., Kim, B.: A Roadmap for a Rigorous Science of Interpretability. arXiv Prepr. arXiv1702.08608v1. (2017)
25. Liang, Y., Li, S., Yan, C., Li, M., Jiang, C.: Explaining the black-box model: A survey of local interpretation methods for deep neural networks. Neurocomputing. 419, (2021). https://doi.org/10.1016/j.neucom.2020.08.011

Apéndice I

En la Tabla 4.6 se muestra el Cuestionario Básico de la ENH 2017 [9], con el que se realiza la consulta sobre las 109 variables asociadas a las características de vivienda, la ubicación y la conformación de los hogares en México. Para la selección de variables se utilizan los siguientes criterios:

- Se incluyen variables afines a los criterios de vivienda adecuada.
- Se descartan las variables que tienen datos idénticos en todos los registros.
- Se descartan las variables mal registradas.
- Se unifican las variables que tengan diferente formato en ediciones previas.

De acuerdo con los criterios de selección de variables, en la Tabla 6 se resaltan en color rojo las variables que se descartan y en color verde las variables que se unifican.

Tabla 4.6. Información de las variables de vivienda.
(Fuente: elaboración propia apartir de la ENH, INEGI (2017)).

Cuestionario Básico de la ENH					
#	Variable	Consulta	#	Variable	Consulta
1	Folio_viv	Identificador de la vivienda	56	const_dorm	Construir dormitorio
2	fipo_viv	Tipo de vivienda	57	const_coci	Construir cocina
3	condominio	Núm. de pisos del condominio	58	const_bano	Construir baño
4	elevador	Disponibilidad de elevador	59	const_neg	Construir negocio
5	mat_pared	Material de paredes	60	comun1	Espacio para sala
6	mat_techos	Material de techos	61	comun2	Espacio para jardín
7	mat_pisos	Material de pisos	62	comun3	Espacio para patio
8	ais_techos	Aislamiento en techo	63	comun4	Espacio para cuarto de lavado
9	ais_pared	Aislamiento en paredes	64	comun5	Espacio para cuarto de televisión
10	ais_ventan	Aislamiento en ventanas	65	comun6	Espacio para cuarto de estudio
11	ais_otro	Otro tipo de aislamiento	66	comun7	Espacio para cuarto de juegos
12	antiguedad	Antigüedad de la vivienda	67	comun8	Espacio para cuarto de ejercicios
13	cocina	Tiene cocina	68	comun9	Espacio para cochera
14	cocina_dor	Utiliza cocina de dormitorio	69	estaciona	Cajones estacionamiento
15	cuart_ dorm	Cuartos dormitorio	70	oomun10	Área común con otras viviendas
16	num_cuarto	Número de cuartos	71	oomun11	Salón de eventos área común
17	disp_agua	Disponibilidad de agua	72	oomun12	Pista para caminar área común
18	dotac_agua	Dotación de agua	73	oomun13	Gimnasio área común
19	excusado	Tiene excusado	74	comun14	Zona de juegos área común
20	uso_compar	Uso compartido del sanitario	75	comun15	Canchas deportiva área común
21	sanit_agua	Sanitario conexión agua	76	comun16	Alberca área común
22	bano_comp	Sanitario excusado regadera	77	comun17	Otra área en común
23	bano_excus	Sanitario excusado	78	tenencia	Tipo de tenencia de la vivienda
24	bano_regad	Sanitario regadera	79	pago_renta	Pago de renta de vivienda
25	drenaje	Destino de drenaje	80	anio_res	Años residiendo en la vivienda
26	disp_elec	Disponibilidad eléctrica	81	mes_res	Meses residiendo en la vivienda
27	anio_panel	Año panel solar	82	familiar	Parentesco con dueño de la vivienda
28	panel_ne	Desconoce año de adquisición	83	tipo_ adqui	Adquisición de la vivienda

29	pot_panel	Conocimiento en potencia inst.	84	financia_1	Recursos propios
30	potencia	Potencia instalada	85	financia_2	Apoyo de FONHAPO[1]
31	focos_inca	Núm. de focos incandescente	86	financia_3	Crédito INFONAVIT[2] o FOVISSSTE[3]
32	focos_ahor	Número de focos ahorradores	87	financia_4	Crédito bancario
33	combustible	Tipo de combustible	88	financia_5	Crédito microfinanciera
34	estufa_chi	Estufa con Chimenea	89	financia_6	Crédito caja de ahorro
35	eli_basura	Eliminación de basura	90	financia_7	Crédito de otra institución
36	lavadero	Dispone de lavadero	91	financia_8	Préstamo familiar
37	fregadero	Dispone de fregadero	92	num_dueno1	Identificador del primer dueño
38	regadera	Dispone de regadera	93	hog_dueno1	Hogar del primer dueño
39	rega_ elect	Dispone de regadera eléctrica	94	num_dueno2	Identificador del segundo dueño
40	tinaco_azo	Dispone de tinaco	95	hog_dueno2	Hogar del segundo dueño
41	cisterna	Dispone de cisterna	96	escrituras	Escrituras de la vivienda
42	pileta	Dispone de pileta o tanque	97	computador	Disponibilidad de computadora
43	calent_sol	Disp. de calentador solar de agua	98	tel_fijo	Disponibilidad de línea telefónica fija
44	calent_gas	Disp. de calentador a gas de agua	99	celular	Disponibilidad de teléfono celular
45	medidor_luz	Dispone de medidor de luz	100	internet	Disponibilidad de internet
46	bomba_agua	Dispone de bomba de agua	101	tv_paga	Servicio de televisión de paga
47	tanque_gas	Disp de tanque gas estacionario	102	tot_resid	Total de residentes
48	aire_acond	Dispone de aire acondicionado	103	tot_hog	Total de hogares en la vivienda
49	calefacc	Dispone de calefacción	104	ubica_geo	Ubicación geográfica
50	chimenea	Dispone de chimenea	105	ageb	Área geoestadística básica
51	repar_pard	Reparar las paredes	106	tam_loc	Tamaño de localidad
52	repar_tech	Reparar el techo	107	est_socio	Estrato socioeconómico
53	repar_agua	Reparar las tuberías del agua	108	esl_dis	Estrato del diseño muestral
54	repar_dren	Reparar las tuberías del drenaje	109	upm	Unidad primaria de muestreo
55	repar_cabl	Reparar el cableado eléctrico	110	factor	Factor de expansión

1 Fideicomiso Fondo Nacional de Habitaciones populares.

2 Instituto de Fondo Nacional de la Vivienda para los Trabajadores.

3 Fondo de la Vivienda del Instituto de Seguridad y Servicios Sociales de los Trabajadores del Estado.