Universidad de Buenos Aires
Facultad de Ciencias Económicas
Escuela de Estudios de Posgrado

MAESTRÍA EN ADMINISTRACIÓN DE EMPRESAS DE BASE TECNOLÓGICA - MBA BT

TRABAJO FINAL DE MAESTRÍA

Machine Learning como herramienta para el uso eficaz de inversiones en bienes raíces

AUTOR: OSCAR R. GOLDSTEIN C.

DIRECTOR: MARIA EUGENIA DE SIMONI

NOVIEMBRE 2019

Resumen

Esta herramienta para el uso eficaz de inversiones en bienes raíces, está enfocada en un modelado de datos que permita conocer el impacto de un proyecto de construcción residencial basándose en el valor del inmueble (ROI), a partir de la información suministrada con proyección a veinte años, en el mercado inmobiliario estadounidense.

Abordamos este proyecto por la necesidad esencial de proteger cercanamente cualquier inversión, y por ser nuestro hogar un activo valioso, es natural que, apoyados en la tecnología, se busque respuestas posibles a la pregunta de cómo invertir inteligentemente.

La idea principal es poder identificar dentro del mercado inmobiliario residencial cuales son las mejores renovaciones de acuerdo con cada caso individual. Democratizando el acceso a la información y permitiendo que los usuarios conozcan el impacto real de cada renovación en el costo total.

Estamos haciendo uso de las nuevas tecnologías para solucionar un problema actual, permitiendo el acceso a la información y desmitificando la industria de los bienes raíces. La utilidad e implicaciones de este proyecto son palpables, haciendo posible el cálculo a futuro, del impacto de una renovación en el precio real de venta sin el gasto y la parcialidad asociadas al contratar a un especialista.

La información proporcionada por el algoritmo nos permitirá tomar decisiones con una variedad de posibilidades. Que los usuarios tengan en sus manos la facilidad de obtener un valor confiable de cuál es el lugar ideal para renovar, cómo enfocar dicha inversión y cuánto puede esperar percibir de ella.

Se abren de esta manera, múltiples alternativas a futuro. Desde visualizar patrones de comportamiento para generar herramientas de carácter educacional, hasta su adaptación en mercados emergentes y en diversos países.

Palabras clave

Retorno de inversión, Aprendizaje automático, Inversiones residenciales, Analítica de datos.

TABLA DE CONTENIDO

1. Introducción

Iniciaremos sosteniendo la idea de que los bienes inmuebles no son únicamente nuestro lugar de residencia, en muchos casos representan también uno de nuestros recursos más valiosos, por su alto impacto económico. Es por ello que, las remodelaciones tienen una importancia significativa, ya que muchas viviendas comienzan a ser, tanto obsoletas como poco funcionales, y se hace necesario remodelar, no únicamente para revitalizar los espacios, sino también como una forma de asegurar su inversión y por ende la ganancia a futuro. El dueño del inmueble debería ser capaz de evaluar sus opciones con claridad, esto significa, cuánto va a invertir, cuál es el proyecto más idóneo en su caso, en cuánto podría revalorizar su propiedad, etc. Por lo tanto, para poder calcular estos datos se hace necesario recopilar información descriptiva, tanto del mercado como de los detalles de la propiedad.

Atendiendo a lo antes expuesto, el presente proyecto intenta describir de qué se trata el tema, desde el planteamiento del problema, hasta el objetivo general que señala textualmente: "Desarrollar un modelado de datos que permita conocer el impacto de un proyecto de construcción residencial sobre el valor del inmueble (ROI) a partir de datos basados en una proyección temporal de veinte años en el mercado inmobiliario estadounidense", dicho propósito se verá consolidado por medio de los objetivos específicos que nos hemos planteado, los cuales incluyen recopilar datos geográficos, sociales y culturales de fuentes fidedignas, estandarizar la información recopilada para ser consumida por algoritmos, generar una proyección usando algoritmos de *machine learning*, evaluar los datos obtenidos con profesionales en el área y datos históricos.

Vale señalar que, hemos recopilado una lista de definiciones y conceptos que, con una vista desde lo macro a lo micro, esclarecerán dudas para abordar un tema que, aun cuando puede parecer complicado, está continuamente presente a lo largo de nuestras vidas. Más allá del retorno de inversión o de los algoritmos de aprendizaje automático para la resolución de problemas, empezaremos mostrando el impacto de las nuevas tecnologías en la vida cotidiana y como hemos sido partícipes indirectos de una nueva revolución industrial.

También estableceremos la metodología a seguir, partiendo de la recolección de datos provistos por ciudades previamente determinadas, que posteriormente organizadas,

podrán explicar cómo ese sistema organizativo será engullido por el algoritmo. La explicación de la estructura del algoritmo se funde con procesos posteriores de edición y retroalimentación.

Por último, como resultado, se planteará hacer uso de herramientas de *machine learning* para comparar predicciones resultantes del algoritmo con datos históricos de la construcción en los Estados Unidos, con la finalidad de poder identificar cuáles son las modificaciones más rentables que pueden realizarse en edificaciones residenciales de acuerdo con sus características y particularidades estéticas. Factores tanto internos como externos a la residencia entran en consideración y son de vital importancia para tomar una decisión consensuada.

2. Planteamiento del problema

"No hay lugar como el hogar", es una frase que tendemos a escuchar con regularidad y que ha sido plenamente cimentada en la cultura popular. Pero más allá del uso estereotipado que se le pueda otorgar a esta expresión, se esconde la realidad innegable de que nuestras casas son a menudo nuestras mayores inversiones. Colocamos nuestro dinero duramente ganado, nuestro futuro y nuestras emociones en ella, es por eso que se hace indispensable medir el crecimiento económico que estos cambios o remodelaciones de viviendas tienen en pro de revalorizar el inmueble, así como se haría con cualquier otra inversión.

El retorno de la inversión (ROI) para un proyecto de renovación es una métrica simple, que mide la restitución obtenida en relación con una inversión realizada para un proyecto. Ella indica si lo invertido mejorará el valor de la vivienda de manera positiva o negativa. Predecir con exactitud el ROI de diferentes proyectos de renovación tiene muchas ventajas. Los propietarios de viviendas pueden tomar decisiones con suficiente información con respecto a las renovaciones, además de abrir un sinfín de posibilidades en el ámbito comercial. Esto incluye a los agentes de bienes raíces, quienes deberán tomar decisiones certeras y no corazonadas; bancos que otorgarán créditos basados en datos sólidos y con menor chance de pérdida; desarrolladores inmobiliarios que podrán conocer el potencial de las zonas geográficas antes de que los precios suban; entre muchos otros profesionales de diferentes ramos, que podrán beneficiarse con el acceso a la información.

El ROI de una renovación dependerá de múltiples factores, entre ellas, la ubicación, la configuración, los detalles de renovación, además de factores demográficos y socioeconómicos del vecindario. Uno de los principales desafíos que encontramos es precisamente la gran cantidad de variables que podrían influir en el cálculo. Su recopilado procesamiento son procesos complejos y requieren un gran compromiso.

En atención a este señalamiento, hemos considerado los siguientes datos para el modelado:

1. **La propiedad o vivienda:** Configuración básica de la casa y su ubicación. Dónde está localizada y cuáles son las condiciones arquitectónicas que presenta: Tamaño, distribución y acabados.
2. **El vecindario:** El número y la calidad de las escuelas, hospitales, parques, iglesias, índice de criminalidad, costo de vida, nivel socioeconómico, estatus de la seguridad personal, número de empresas productoras que la circundan, centros comerciales y diferentes tipos de comercio.
3. **La ciudad:** Tasa promedio de empleo, la economía regional, el ingreso familiar medio, la tasa de natalidad y mortalidad, el índice de seguridad, clima, nivel de contaminación, índice económico de sus pobladores, entre otros.
4. **La renovación:** proyecto de renovación a realizar, el tamaño del espacio agregado y el costo.

La predicción del ROI se logrará gracias al uso de unos algoritmos que generan una predicción estimada a través del aprendizaje automático, que es una disciplina de la "Inteligencia Artificial" que crea sistemas que aprenden automáticamente. Aprender, en este nuevo contexto tecnológico, quiere decir identificar patrones complejos a partir de diversos datos.

El algoritmo, al revisar todos los datos concernientes a la propiedad, es capaz de predecir comportamientos futuros. Por lo tanto, la propuesta que se ofrece es el modelado de los datos relevantes para una plataforma digital, que, conociendo toda la información descrita anteriormente, pueda identificar el impacto de los proyectos de renovación en su precio futuro. Nos estamos enfocando específicamente en el mercado estadounidense, y utilizando la recopilación de data de dominio público del año 2000 hasta la actualidad, para la realización del estudio que nos interesa.

Justificación

Predecir el retorno de la inversión no es una tarea simple, incluso las personas que trabajan en la industria de la renovación y tienen experiencia o las que están en el negocio de compra y venta de inmuebles, solamente hacen estimaciones que, en su mayoría, dependen de valores subjetivos apoyados en sus años de experiencia en el mercado.

La razón principal de la dificultad para predecir el ROI surge en la precisión, debido a la gran cantidad de factores que influyen, entre ellos, la ubicación, la configuración, el tipo de renovación, además de la cercanía de escuelas, hospitales y parques, que pueden influir indirectamente en la valoración.

Un ejemplo ilustrativo; el proyecto de construcción de una piscina residencial en cualquier ciudad del estado de Florida produce un retorno significativamente mayor si se compara con la misma construcción en Minnesota. En regiones en donde estos proyectos pueden llegar a ser contraproducentes por la relación costo veneficio, podemos ver un impacto negativo en el precio de la vivienda.

Para estimar el Retorno de la Inversión, nos basaremos entonces en datos cuantificables del mercado residencial norteamericano que son de libre acceso, para conseguir un modelo de manejo de datos que permita obtener una visión global de qué esperar y cuál es la mejor manera de tratar las inversiones residenciales.

Estas nuevas propuestas, donde se hace uso consciente de la tecnología y medios digitales para predecir una variable económica determinada, vienen muy a tono con la estructuración de la gestión de la innovación, pues el aprovechamiento de nuevas tecnologías es vital para la subsistencia de cualquier negocio en la actualidad.

Objetivo General:

Desarrollar un modelado de datos que permita conocer el impacto de un proyecto de construcción residencial sobre el valor del inmueble (ROI) a partir de información de veinte años, en el mercado inmobiliario estadounidense.

Objetivos Específicos:

- Recopilar datos demográficos y socioeconómicos de fuentes fidedignas alimentar el algoritmo.
- Estandarizar la información recopilada para ser consumida por algoritmos.
- Generar una proyección estimada usando los algoritmos de aprendizaje automático.
- Evaluar los datos obtenidos con profesionales en el área, así como datos históricos.

Hipótesis:

Un algoritmo de aprendizaje automático es capaz de predecir el retorno de la inversión en edificaciones residenciales, en el mercado norteamericano.

3. Marco teórico

Basaremos el desarrollo en tres referenciales: El primero sobre inteligencia artificial, el segundo sobre algoritmos de aprendizaje automático y por último, todo lo referente al mercado de inmuebles según la perspectiva humana. Estos antecedentes nos ayudarán a comprender mejor las áreas en las cuales el estudio tendrá su enfoque. De tal manera, otorgaremos un contexto que sirva como guía para cumplir lo pactado en los objetivos propuestos para este estudio.

La primera publicación que queremos citar es *Historia y evolución de la inteligencia artificial* (Casella, 2017). Este texto nos presenta un panorama amplio sobre los inicios de esta forma de inteligencia. Comienza informando que, sin duda alguna, el concepto de Inteligencia Artificial (IA) surge en 1956. Lo que posteriormente si cuestiona, es que los filósofos e investigadores no se han puesto de acuerdo en si esta IA es una ciencia o un híbrido entre la informática y la ingeniería. Casella concluye que el objetivo de la IA es "la construcción de artefactos que ayudan y asisten al hombre - y en algunos casos - los sustituyen en la resolución de tares teóricas o prácticas de diferente complejidad".

Según Casella, el antecedente más cercano a la IA es el matemático inglés Alan Turing, a quien se considera como el precursor de la informática moderna. Turing presentó una tesis sobre la naturaleza de los cálculos, cuando en 1936, diseñó una máquina que demuestra la viabilidad de un dispositivo físico que permitió implementar cualquier cómputo formalmente definido. Este adelanto sirvió para que, en 1943, Warren McCulloch y Walter Pitts presentaran su modelo de neuronas artificiales, el cual ha sido considerado como el primer trabajo en este campo, aun cuando todavía en ese entonces, no existía el término de Inteligencia Artificial.

En 1955 Herbert Simón, Allen Newell y J. C. Shaw, desarrollaron el primer lenguaje de programación, cuyo fin apuntaba hacia la resolución de problemas. Posteriormente, tan solo un año después en 1956, desarrollan el *Logic Theorist*, prototipo que permitía demostrar teoremas matemáticos. Fue en ese mismo año de 1956, que el término inteligencia artificial se comenzó a usar por los matemáticos John McCarthy, Marvin Minsky y Claude Shannon.

Más de una década después, en 1967 surge el primer periódico sobre IA, llamado "Inteligencia Artificial". Esta publicación testimonia "los cambios de rumbo en la mitad de los años 70, cuando la insatisfacción por los avances logrados por la investigación en los refinamientos de la resolución resucitó el interés por una demostración de teoremas menos sensibles al requisito de la integridad y más enfocada hacia los procedimientos heurísticos inspirados en los métodos humanos de solución de problemas" (Casella, 2017).

A lo largo de su libro, Casella propone un sinnúmero de tesis e investigaciones que han aportado avances en el campo de la inteligencia artificial - concepto éste que sigue en reconstrucción y es aplicado en todos los campos - tal como el propio autor lo señala. Él sugiere que los distintos estudios y avances sobre la IA se nutren entre sí, mientras que al mismo tiempo son rivales, pues se han seguido los caminos más diversos en pro de ir generando nuevos paradigmas que alimenten lo ya conocido y puedan explicar de forma predictiva lo nuevo por conocer.

Este texto, resuena con el presente proyecto bajo el marco de la IA y particularmente en los cambios de paradigma de la llamada cuarta revolución industrial, debido a los avances radicales en los campos tecnológicos, inteligencia artificial y conectividad, aunado a que el potencial de procesamiento de datos evoluciona radicalmente abriendo posibilidades nunca vistas y permitiendo la predicción de gran cantidad de comportamientos en distintos planos de la actividad humana.

El segundo referencial que tomaremos en cuenta en este trabajo de Machine Learning como herramienta para el uso eficaz de inversiones en bienes raíces, es una tesis presentada ante la Universidad de Uruguay llamada *Estudio de factibilidad del uso de machine Learning con múltiples fuentes de datos en el pronóstico del tiempo*, trabajo realizado por Natalie Gnoza y Marcelo Barberena en el año 2008. De esta disertación interesa específicamente el uso del aprendizaje automático - concepto que históricamente tiene un desarrollo relativamente reciente - y cómo desde sus orígenes ha tenido repercusión en todos los niveles de la vida cotidiana. El concepto de aprendizaje automático en el contexto de la IA, se define como aquel que "aprende identificando patrones complejos en grandes volúmenes de datos, para luego generalizar y realizar asociaciones entre ellos. Son capaces de mejorarse en forma autónoma a partir de la experiencia" (Gnoza & Barberena, 2018).

Con el auge y crecimiento de la Gran data (concepto que definiremos más adelante), se ha podido avanzar en la resolución de muchos problemas que anteriormente eran desafíos para el hombre. El aprendizaje automático se ha transformado en una herramienta exponencial que contribuye en cualquier aspecto. Así es que podemos verlo en nuestros dispositivos móviles como reconocimiento facial, de voz y de objetos. En los buscadores, para mejorar los resultados y sugerencias de la búsqueda y antispam. En el software de seguridad como antivirus y predicción de usos maliciosos. En la genética con la clasificación de secuencias de ADN. En la predicción y pronósticos del clima o tráfico entre muchos otros.

Utilizaremos este texto de los autores anteriormente citados, para comprender conceptos un poco más técnicos y relevantes al campo de la ingeniería, tales como las diferencias específicas entre las técnicas de aprendizaje, supervisadas o no. Gnoza & Barberena, 2018, destacan dos aspectos importantes sobre el aprendizaje autónomo:

El primero trata sobre cómo los datos se clasifican en categorías previamente aprendidas. Este es un aprendizaje basado en la clasificación y en la regresión de los datos, y, en segundo lugar, un arreglo que "encuentra patrones ocultos o estructuras intrínsecas en los datos". Quiere esto decir, que será capaz de inferir resultados a partir de datos de entrada que no han tenido respuestas etiquetadas.

Nuestro tercer referencial es la tesis doctoral de (Hernando, 2016), presentada ante la Universidad de Catalunya. Dicho estudio lleva por título *Análisis e inversión en el mercado inmobiliario desde una perspectiva conductual*. Esta tesis tuvo como objetivo general aplicar las teorías de la escuela conductual en los análisis y modelos de toma de decisiones e inversión para mejorar la rentabilidad de la cartera de un inversor. El autor concluye que en los últimos años la inversión en inmuebles ha crecido, así como también la importancia de este activo como inversión. La investigación utilizó una metodología experimental, pues trata de conciliar dos áreas que a simple vista parecen muy distantes - la psicología y las finanzas - para lo cual se valió del método deductivo-inductivo. Así pudo comprender "que la aplicación de estrategias conductuales a la estrategia corporativa llevaría a la obtención de fondos externos para la realización de nuevas inversiones inmobiliarias" (Hernando, 2016, p. 301).

El aumento del interés inversor en el sector inmobiliario, el cual forma parte, tanto de carteras de pequeños inversores como de grandes fondos de inversión internacionales, ha llevado a la aparición de nuevos vehículos de inversión inmobiliaria, dotando al mercado de mayor complejidad, asemejándolo cada vez más a los mercados financieros.

El presente proyecto utiliza conceptos que están estrechamente racionados con el área de bienes raíces, para ser combinados con el sector de los mercados de inversión y finanzas, con la finalidad de dar una respuesta holística al problema en cuestión y generar una fuente fidedigna de la capacidad de retorno en una remodelación de bienes inmuebles. Por lo cual, el nexo con estos estudios anteriormente citados es el manejo teórico y pragmático asumido para dar luces sobre el mercado.

El entorno

Internet de las cosas [*Internet of things*] (IoT): Término utilizado para referirse al como elementos de uso común han adoptado la conectividad total para automatizar y facilitar procesos por parte del usuario. El Internet de las cosas, consiste entonces en extender el poder de Internet más allá de las computadoras y los teléfonos inteligentes a una amplia gama de otras situaciones, procesos y entornos. Esas cosas "conectadas" se usan para recopilar y enviar información. “El IoT tiene el potencial de extraer y analizar datos en tiempo real de los millones de sensores conectados a él y luego aplicarlos para ayudar a "procesos automatizados y basados en personas”. (Miraz & Ali, 2015)

IoT proporcionará una mejor visión y control sobre el 99% de los objetos y entornos que permanecen fuera del alcance de Internet. Y al hacerlo, permite a las empresas y a las personas, estar más conectadas con el mundo que les rodea para hacer un trabajo más significativo y de mayor nivel.

Según un informe del desarrollo mundial (World Development Report, 2016), la aplicación del internet de las cosas puede ser distribuida en cinco categorías: Dispositivos usables, casas inteligentes, ciudades inteligentes, sensores medioambientales y aplicaciones de negocios.

Esta nueva forma de ver el mundo repercute especialmente en los negocios y por ende ha sido ampliamente asimilada por las empresas de servicios. Las “Apps” se han convertido en un puente de comunicación continua entre el usuario final y las empresas. A través de una aplicación podemos solicitar un servicio completo, asistencia técnica, información y mucho más.

Las nuevas tendencias parecen ir de lo macro de la producción en masa a lo micro del trato personalizado con el cliente. Esto no es casual, distintos factores a lo largo del desarrollo de la conectividad han dado pie a que las empresas tengan hasta cierto punto facilidades para hacerse con la información personal de sus usuarios. El auge de las redes sociales ha sido un revulsivo tremendo porque al darle voz a los usuarios para expresar sus distintos gustos e intereses, les permite a las compañías saber con qué usuario está tratando y hacer lo posible por mejorar su experiencia.

Cuarta revolución industrial [*The fourth Industrial revolution*]: Período de tiempo en el que una gran cantidad de adelantos tecnológicos han convergido y permite presuponer un salto radical en la forma de percibir e interactuar con el mundo.

Este período se caracteriza por la automatización total de procesos de producción, el uso continuo de inteligencia artificial, la capacidad de implementar la conectividad web en todos los aspectos y la capacidad de gestión de gran cantidad de información, para así ofrecer servicios integrados de mayor personalización a los usuarios. Muchos la ubican alrededor del año 1995 y enuncian una proyección exponencial en su implementación en todos los campos del desarrollo humano en los próximos años.

Pero ¿es tan importante este cambio como nos lo han hecho ver? ¿Estamos preparados para adaptarlo a nuestras formas de vida?

Con respecto a la primera pregunta tenemos posturas encontradas. Por un lado, el Foro Económico Mundial hace hincapié en la importancia de la cuarta revolución industrial y en sus múltiples ventajas, así como que el papel de la conectividad ha sido esencial para ayudar a producir una gran cantidad de innovaciones a lo largo de los últimos años.

En contraposición tenemos la postura de (Gordon, 2000) que asegura que muchas de las innovaciones que la cuarta revolución industrial ha traído consigo, pertenecen en realidad a la tercera. En su ensayo titulado "¿Does the new economy measure up to the great inventions of the past?" Gordon argumenta que, hablando proporcionalmente, las innovaciones que la cuarta revolución industrial ha aportado al mundo no pueden compararse con el impacto que significó, por ejemplo, la máquina de vapor en la primera, la electricidad en la segunda, o el transistor en la tercera.

Debemos ver la línea evolutiva de la tecnología en el mundo como una curva creciente con una gran pendiente en un primer momento. Por supuesto, para una sociedad que no conocía el concepto de motor, la máquina de vapor vino a ser una invención con una repercusión tremenda. Para nosotros, que manejamos el concepto del computador desde los años 80, la conectividad total parece ser una adición casi esperada, el siguiente paso lógico del salto tecnológico, y por eso tendemos a darlo por sentado.

Una buena forma de asimilar el impacto tremendo que estos años han generado en la tecnología, es pensar en un día normal de su vida hace veinte años y compararlo con uno en este momento. ¿Podría desarrollar cada una de sus actividades de la misma manera? Ir al banco, conseguir un taxi, comprar un abrigo o hasta pedir una pizza son acciones que en ese entonces obedecían a distintos recursos. La híper-conectividad actualmente es una realidad.

Recopilación de datos en colaboración [*Crowdsourced data*]: Combinación de vocablos "Crowd" (Multitud) y "outsourcing" (dar control a otra entidad). Según (Howe, 2006) el crowdsourcing, es cuando los investigadores recurren a las comunidades de Internet para responder preguntas de investigación, encuestas o comentarios. La recopilación de datos de colaboración múltiple está ganando popularidad porque es conveniente, económica y relativamente rápida. Aprender cómo se compara esta estrategia con enfoques más completos y tradicionales es importante para orientar la investigación futura.

Específicamente, para la creación de este algoritmo, fue necesaria la crowdsourced data para recopilar los datos de percepción de riesgo, para ser comparados con los censos

gubernamentales e introducir una variable de riesgo "percibido" dentro del cálculo especulativo de venta de una vivienda y su repercusión en el análisis del negocio.

Esto nos da un buen ejemplo de cómo la crowdsourced data puede ayudar a preparar una muestra correcta a la hora de resolver un problema. En nuestro caso, son las viviendas las que nos darán información en un primer momento, más adelante, serán las ciudades, pues buscamos abordar todos los factores que condicionan la percepción monetaria de una vivienda.

Retorno de la inversión [*Return of investment*] (RoI): Es una medida de desempeño usada para evaluar la eficiencia de una inversión o compararla frente a un número determinado de inversiones. Según Chen J., 2019 el RoI trata de medir directamente la cantidad de retorno de una inversión en particular, en relación con el costo de ésta. Para calcular el RoI, el beneficio - o rendimiento - de una inversión, se divide por el costo de esta. El resultado se expresa como un porcentaje o una proporción.

Es importante considerar que el término ROI no solo es dominio del mercado de bienes raíces. El acrónimo se usa cada vez con mayor popularidad para referirse a cómo se puede predecir el comportamiento de un determinado rubro en el tiempo al aplicar una cierta inversión. Una estructura de cálculo que es válido para cualquier cálculo de RoI:

$$RoI = \frac{Ganancia - Costo}{Costo} * 100$$

Donde:

-Ganancia: Se refiere a la cantidad de capital generado gracias a la venta o inversión.
-Costo: Se refiere al valor total que el proyecto al que nos abocamos costará.

Esta simple ecuación, que nos da un valor en porcentaje, es el campo base para cualquier cálculo de ROI. Por supuesto un cálculo hecho en base a esta ecuación, presentará una precisión limitada al no contar con múltiples factores que darán más información y ayudarán a que el cálculo converja con mayor criterio.

En este punto, es importante señalar que, en nuestro caso particular, existirán dos tipos de ROI, y es porque un proyecto de renovación puede ser visto desde dos perspectivas: Primeramente, en cómo el proyecto ganará valor en sí mismo (RoI clásico) y en segundo lugar, cómo se relaciona con su entorno (RoI aumentado).

RoI clásico: Se centra en el proceso de la renovación y se desprecia todo dato relevante al entorno o factores externos. El mismo precio de la vivienda no influye directamente en este caso.

Factores para considerar:

- Costo de mano de obra.
- Costo en materiales.
- Tiempo de ejecución de la obra.
- El espacio proyectado a trabajar.

RoI aumentado: Por otra parte, el cálculo de RoI que se apega más a la realidad se da tomando una visión más holística de la propiedad y el entorno. Todo factor externo debe ser considerado ya que los bienes raíces son un mercado altamente especulativo. Algunos de los factores adicionales (sumados a los del caso anterior) que se consideran son:

- El año de construcción.
- Estado físico de la propiedad.
- El costo de la propiedad.
- La ubicación
- Potencial de proyección a futuro (gentrificación)
- Características socioeconómicas del vecindario y sus habitantes
- Riesgo directo o indirecto a desastres naturales
- Disponibilidad y costos de servicios básicos.
- Percepción de riesgos asociados a la propiedad.

Como resultado, el cálculo de RoI se transforma en algo mucho más complejo, que requiere más allá de una tabulación estándar de precios de remodelaciones según el área, un cálculo complejo basado en la recopilación de datos históricos de múltiples fuentes.

La Comisión de Bolsa y Valores de los Estados Unidos [The U.S. Securities and Exchange Commission] (SEC): es una agencia independiente del gobierno federal de los Estados Unidos. La SEC tiene la responsabilidad principal de hacer cumplir las leyes federales de valores, proponer reglas de valores y regular la industria de valores, que es el intercambio de acciones y opciones de la nación, y otras actividades y organizaciones, incluidos los mercados de bienes raíces en los Estados Unidos. (SEC, 2019)

La Ley de Libertad de Información [Freedom of Information Act] (FOIA): es una ley federal de los Estados Unidos que otorga acceso público a la información que poseen las agencias gubernamentales. Previa solicitud por escrito, las agencias del gobierno de EE. UU. deben divulgar la información. Todos los departamentos, agencias y oficinas del Poder Ejecutivo, agencias reguladoras federales y corporaciones federales están sujetos a la Ley de Libertad de Información. Esto incluye a las oficinas de planificación urbana.

La tecnología

Tecnologías de la comunicación y la información [*Information and communications technology*] (ITC): Término referido a cualquier tipo de medio utilizado para comunicar y hacer llegar información a un público de masas. En esta cuarta revolución industrial, el término se aleja de las ITC convencionales de la tercera revolución, como lo son la televisión y la radio, y se avoca al internet, y más en específicamente a las redes sociales. Según Marqués, 2000, la principal característica de estas nuevas ITC radica en la rapidez, tanto del usuario para acceder a la información como del medio para procesarla.

Las ITC tienen fácil acceso a los usuarios, gracias a como la conectividad está presente en muchos de los aspectos de nuestras vidas.

Es necesario entender que en la actualidad cualquier tipo de adelanto debe ir acompañado de capacitación inmediata, que permita una continua producción y retroalimentación. Las tecnologías de la comunicación y la información son responsables directas de los productos informáticos basados en data porque les han permitido a los desarrolladores acceder fácilmente a información que, en otros tiempos, solo se podían conseguir con continuas y extenuantes horas de trabajo en un sitio específico.

No es difícil entrever por qué la hiper conectividad ha sido una participante activa en el desarrollo de cualquier proyecto tecnológico en la actualidad.

Gran data [*Big data*]: Término aplicado a las grandes bases de datos que se crean mediante la interacción de usuarios con un entorno digital. Estas grandes bases de datos aportan información interesante a compañías en cuanto a cómo personalizar la forma en que se acercan a sus clientes.

Ahora bien, la Gran Data no es solo llamada así gracias a su tamaño. A medida que la globalización ha traído distintos procesos y ha permitido poner en las manos de cualquier persona grandes cantidades de datos, gracias a las tecnologías de la comunicación y la información, se hace necesario hacer una distinción en cuanto a qué tan grande debe ser una data para considerarla Gran Data. Este término por lo general se utiliza cuando el volumen de datos que se maneja es tan grande, que los métodos tradicionales de visualización, gestión y análisis no pueden procesar la información.

Según (Elgendy & Elragal, 2014) el Big Data tienes tres factores que denominaron las tres "V": Volumen, Variedad y Velocidad. Podemos decir que el Volumen es la más importante – y la más obvia - El tamaño de las datas significó un gran problema cuando empezaron a pasar de un volumen normal a uno gigantesco. La poca cantidad de espacio de almacenamiento digital hizo que guardar una data se convirtiera en una tarea costosa. Solo cuando la democratización de servidores permitió espacios seguros y continuos de almacenamiento, los costos empezaron a bajar, una vez que la demanda estaba lo suficientemente atendida.

El segundo factor que Elgendy y Elragal tomaron en cuenta es la Variedad. Con esto se refieren a los distintos formatos en los que se puede procesar la Gran Data. Por supuesto, cientos de desarrolladores web han logrado maneras distintas de enfrentarse a tal flujo de información y han generado distintas maneras de afrontarla. Al igual que con el Volumen, cierto proceso de adaptación ha permitido que algunos tipos destaquen sobre otros y tengamos formatos estándar de análisis.

La velocidad - tercer factor manejado por los autores citados - parece ser el término más confuso. Solemos asociarlo concretamente a la manera en que se procesan los datos,

pero nada más lejos de esto. Este término se refiere a la rapidez con que la data debe ser actualizada en función a su contexto. Esto es un punto clave, pues nos muestra que es de crucial importancia que la recolección se planifique con relación a lo que se analiza. Distintos factores pueden exigirle a la data una actualización continua cada semana, mes y año. Imaginemos que realizamos nuestro estudio con data de remodelaciones de los años noventa, ¿Tiene esto sentido? ¿Aún lo tendría para datos del 2010? Estos factores deben ser puestos en tela de juicio, pues una buena data en el momento equivocado es sencillamente una inadecuada información.

También podríamos ver la data como un bosque, ciertamente la información principal está ahí, pero a su vez hay muchos factores escondidos. Una buena data puede permitirnos, al analizarla, encontrar patrones reconocibles que permiten sacar conclusiones sobre su contexto. A esto se le llama análisis de Gran Data y es el concepto básico de cómo el algoritmo se enfrenta a los datos existentes.

Nuestro estudio hará un uso extensivo de grandes volúmenes de data o de Big Data, pues será necesaria la implementación de una gran variedad de detalles concernientes tanto, a proyectos de remodelación, como a los detalles geográficos y demográficos.

Minería de datos, [*Data mining*]: No es raro ver cómo se usan indiferentemente los conceptos minería de datos y aprendizaje automático. Son conceptos primos hermanos. La principal diferencia radica en el objetivo que tiene cada una de estas disciplinas. Mientras que la minería de datos descubre patrones anteriormente desconocidos, el aprendizaje automático se usa para reproducir patrones conocidos y hacer predicciones basadas en los patrones.

La minería de datos es un subcampo interdisciplinario de la informática que implica el proceso computacional del descubrimiento de patrones de grandes conjuntos de datos. El objetivo de este proceso de análisis avanzado es extraer información de un conjunto de datos y transformarla en una estructura comprensible para su uso posterior. (Jain & Srivastava, 2013)

En pocas palabras, se podría decir, que la minería de datos tiene una función exploratoria mientras que, el Machine Learning se focaliza en la predicción. Al contar con

todos los patrones por defecto, gracias al historial de datos, la minería viene a ser un análisis fundamental para el estudio, pues permite descifrar patrones dentro de la data recolectada.

"Un buen sistema de minería de datos debe ser capaz de minar efectivamente varios tipos de data, y ya que la mayor parte de que los puntos de data son hechos en función a data relacional, es muy importante poder extraerla con claridad" (Chen M.-S. , 1996).

El algoritmo utiliza patrones reconocibles que le permiten decir que, bajo un conjunto de condiciones particulares, una misma renovación será más costosa en una zona residencial de alto poder adquisitivo, que en una zona rural. La minería de datos es lo que transforma terabytes datos en información utilizable.

Aprendizaje automático, [*Machine learning*]: Este término se refiere al método de construcción de un algoritmo en el que se le asignan distintos inputs de prueba para que el algoritmo aprenda a crear patrones y sea capaz entonces de predecir resultados parecidos en el futuro. Según la universidad de Stanford, el proceso de machine learning presenta una analogía parecida al conocimiento empírico que el hombre desarrolla al interactuar con su entorno y predecir conductas futuras.

Aprendizaje automático supervisado y no supervisado, [*supervised and unsupervised machine learning*]: El Machine Learning se divide en dos áreas principales: aprendizaje supervisado y aprendizaje no supervisado. Más allá del grado en que el factor humano se involucra en cada proceso, los tipos denotan lo que se hace con la información adquirida:

Uno de los usos más extendidos del aprendizaje supervisado consiste en hacer predicciones a futuro basadas en comportamientos o características que se han visto en los datos ya almacenados. El aprendizaje supervisado permite rastrear patrones en datos históricos relacionando todos campos, uno especialmente llamado campo objetivo. Por ejemplo, los correos electrónicos se etiquetan como "spam" o "legítimo" por parte de los usuarios. El proceso de predicción se inicia con un análisis acerca de qué características o patrones tienen los correos ya marcados con ambas etiquetas. Se puede determinar, por ejemplo, que un correo spam es aquel que viene de determinadas direcciones IP, presenta una determinada relación texto/imágenes, contiene ciertas palabras, no hay nadie en el

campo “para:”, entre otras variables. Este sería tan solo uno de los patrones. Una vez determinados todos los patrones en esta fase que denominamos “de aprendizaje”, los correos nuevos que nunca han sido marcados como spam o legítimos se comparan con los patrones y se clasifican o se predicen, como “spam” o “legítimos” en función de sus características.

Por el contrario, cuando nos referimos al Aprendizaje No Supervisado, encontramos que éste usa datos históricos que no están etiquetados. Su finalidad es explorar dicha data para encontrar alguna estructura o forma de organizarlos. Así, por ejemplo, lo podemos encontramos frecuentemente cuando se trata de agrupar clientes con características o comportamientos similares para los cuales hacer campañas de marketing altamente segmentadas.

Así pues, en un proyecto de ROI que presagie con claridad la ganancia de una renovación, el Aprendizaje Supervisado debe imperar en el código, debido a que partiremos de una base de datos consistente de anteriores permisos de renovaciones. Estos estarán organizados según ciertas variables tales como el tipo de residencia, el proyecto a realizar y el área de trabajo.

A continuación, definiremos algunos conceptos más específicos que abarcan los principales tipos de algoritmo de Machine Learning:

Aprendizaje automático por aprendizaje supervisado [*Supervised Machine Learning*]*:* En este caso se diseña un algoritmo mediante un método predictivo, en el que existen datos de entrada y salida. Los datos de entrada han sido previamente clasificados y estandarizados, a manera de prepararlos previamente para la asimilación y aprendizaje del logaritmo.

Aprendizaje automático por aprendizaje no supervisado [*Unsupervised Machine Learning*]: En esta circunstancia el aprendizaje del algoritmo se realiza sin clasificación de datos previa. Por lo tanto, el algoritmo debe aprender por una estructura de asimilación que le asigna ciertas cualidades parecidas y agrupar este tipo de comportamientos.

Aprendizaje automático de aprendizaje por refuerzo [*Reinforcing Learning*]: Este método se estructura de tal manera que el algoritmo contenga ciertas acciones preestablecidas que, al ser satisfechas, emulan un cierto criterio de satisfacción. De esta manera el algoritmo aprende poco a poco a clasificar los datos de entrada de la manera óptima posible para satisfacer la mayor cantidad de criterios de acción posibles.

Algoritmos clásicos de aprendizaje automático: A la hora de proponer un algoritmo de Machine Learning para la resolución de un determinado problema, por lo general el programador parte de una estructura básica después de identificar la necesidad.

Clasificación y Regresión [*Classification and Regression*]: Son conceptos del Machine Learning supervisado. Un sistema de clasificación predice una categoría, mientras que una regresión pronostica un número.

Un ejemplo de Clasificación es el anteriormente mencionado acerca del spam. Los correos se "categorizan" como "spam" o como "legítimos". Otro ejemplo clásico de Clasificación en el mundo del Machine Learning es la predicción de bajas en un servicio de telefonía, por ejemplo. El objetivo en este caso es detectar los patrones de comportamiento de los clientes para así predecir si optarán por migran hacia la competencia. En este caso los clientes se clasifican como "baja" o "no baja".

La Regresión, por su parte, predice un número. Por ejemplo, cuál podría ser el precio de un artículo, o el número de reservas que se harán en un hotel, en un momento determinado del año.

En nuestro caso, se trabajará con Regresiones, pues lo que buscamos finalmente es un valor de ganancia, tomando siempre en cuenta que el algoritmo también realizará un trabajo de clasificación pues, al estar alimentado con suficientes datos, comenzará un proceso de predicción categórico.

Aprendizaje o Entrenamiento [*Learning, Training*]: Es el proceso en el que se detectan los patrones de un conjunto de datos, es decir, es el corazón del Machine Learning. Una vez identificados los patrones, se podrán hacer predicciones con nuevos datos que se irán incorporando al sistema.

Por ejemplo, los datos históricos de las compras de libros en una web online se pueden utilizar para analizar el comportamiento de los clientes en sus procesos de compra. Esto quiere decir los títulos visitados, las categorías, el historial de compras, etc., agrupándolos en patrones de comportamiento y haciendo recomendaciones de compra a los nuevos clientes que siguen los patrones ya conocidos o aprendidos.

Esto lo haremos una vez que dispongamos de la data. Cada data de remodelación proporcionada por una ciudad, llegará a nuestras manos con una estructura similar que deberá ser correctamente ordenada para alimentar al algoritmo.

Instancia, Ejemplo o Registro [*Instance, Sample, Record*]: Denominamos Instancia a cada uno de los datos disponibles para hacer un análisis. Si se quiere predecir por ejemplo el comportamiento de los clientes de un servicio de telefonía, cada Instancia corresponderá a un abonado. Cada una de ellas a su vez, está plagada con características que la describen: La antigüedad del cliente en la compañía, el gasto diario en llamadas, etc. En una hoja de cálculo, las Instancias serían las filas, mientras que las Características, ocuparían las columnas.

Característica, Atributo, Factor, Propiedad o Campo [*Feature, Attribute, Property, Field*]: Son las cualidades que describen cada una de las Instancias del conjunto de datos. Las denominaciones se usan indistintamente en función del autor y del contexto. En el caso de una cartera de clientes, estaríamos hablando del número de compras de cada cliente, su antigüedad, si es seguidor en redes sociales, si se ha suscrito en el reporte semanal, qué productos comprados, entre otros.

Ingeniería de Factores [*feature engineering*]: Este concepto atañe al proceso previo de la creación del modelo de predicción en el que se hace el análisis y la estructuración de los campos de los datos. Este proceso es uno de los más importantes y costosos del proceso de predicción. El objetivo es eliminar los campos que no son útiles para hacer la predicción y así organizarlos adecuadamente, con la finalidad de que el modelo no reciba información inútil que podría provocar predicciones de pobre calidad o poca confiabilidad.

Este proceso de filtrado es esencial. Muchas veces podríamos toparnos con información que pudiese confundir al algoritmo. En nuestro caso, ejemplos claves pueden ser que el algoritmo reciba algún tipo de data de proyectos comerciales. Si el algoritmo no sabe que la combinación de palabras "No + Residencial" se refiere a un proyecto comercial, lo filtra como residencial solo por el hecho de reconocer la palabra *residential.* Según Scott Locklin (2014) la ingeniería de factores es aún tratada con un carácter bastante informal, a pesar de que la literatura sobre Aprendizaje Automático es ya bastante basta. Sin embargo, el papel que juega en el entorno de Machine Learning es crucial y debemos verla como prioritaria.

Modelo [*Model*]: Tras entrenar al sistema, es decir, después de detectar los diferentes patrones en los datos, se creará un modelo que permitirá para hacer las predicciones. Podemos asimilar un modelo a un filtro en el que entran datos nuevos y cuya salida es la clasificación de esos datos según los patrones que se han detectado en el entrenamiento. En nuestro caso, esto es ya referido al proceso como tal, donde se tomará los datos de entrada para predecir el ROI en la salida.

Red neuronal: Una Red Neuronal, en el caso específico del lenguaje de programación, se refiere a una forma determinada de ordenación y procesamiento de parámetros, en la que, disponiendo de un grupo de datos de entrada, distintos procesos se entrelazan entre sí para clasificar y dar nuevas características - totalmente de carácter matemático - con el fin de conseguir un input que pueda resolver un tipo determinado de problema. Su nombre viene de la gran similitud estructural que tiene con las redes neuronales que regulan complejos de sinapsis para la generación del pensamiento. Existen varios tipos de redes neuronales, entre ellas tenemos:

Red de Alimentación Neuronal hacia Adelante: [*Feedforward Neural Network*] Esta es una de las más simples formas de estructura neuronal que existen. Los datos de entrada siguen un flujo lineal a través de cada capa para finalmente desembocar en la salida, cumpliendo con el objetivo planteado. En este caso pueden o no haber capas ocultas dentro de la red. Según Zhang Jian y Wang XueWu (2011), una red de alimentación neuronal de tres capas ajusta automáticamente los parámetros de estudio para reducir así los errores entre valores reales y simulados.

Red Neuronal de Base Radial [*Radial basisfunction Neural Network*]*:* Las redes Neuronales de Base Radial consideran la distancia que existe entre un punto de estudio con respecto a un centro previamente establecido. Este tipo de redes presenta dos capas, una primaria interna, en la que los datos de entrada son comparados con el centro, y otra en la que estos nuevos valores son considerados y, mediante una fase de cómputos matemáticos equivalente a capas ocultas, arroja el resultado final. Según Iman Sadeghkhani, Abbas Ketabi y Rene Feuillet (2012), la función matemática más utilizada en el procesamiento de datos para este tipo de red es la función Gaussiana, pues la red en sí presenta un tiempo reducido de cómputos y una topología más compacta.

Red de Auto Organización de Kohonen [*Kohonen Self Organizing Neural Network*]: La red de Kohonen se estructura de forma vectorial. Esto es, que cada neurona hace la suerte de un vértice de un cuadrado, y cada cuadrado se organiza en conjunto para crear una red al lado del otro. Esta particular manera de ajustarse, permite reconocer patrones de comportamiento complejos y por lo tanto es bastante utilizada en el campo médico, por ejemplo, para clasificar pacientes con distintas tasas de filtración glomerular. Según Van Biesen W, Sieben G, Lameire N, Vanholder R. (1998), la red de Kohonen funciona creando un ordenamiento neuronal en el que cada neurona se ubica utilizando el algoritmo euclidiano más cercano a la próxima.

Red Neuronal de Memoria Larga de Corto Plazo [*Long Short Term Memory*]: Este tipo de red neuronal se basa en la retroalimentación. En este caso cada capa cuenta con un proceso de redirección en el cual los resultados puntales se toman en consideración para la época siguiente. Esto quiere decir que cada neurona recuerda los valores previos inmediatos e influencia la siguiente predicción. Las aplicaciones de este tipo de red neuronal son bastante amplias y prácticas.

En un ensayo elaborado en el año 2017 por un grupo de científicos del laboratorio Baidu de inteligencia artificial de Silicon Valley, podemos ver como el uso de este tipo de red neuronal se usa para alimentar un software de traducción de fonemas a un lenguaje escrito. Esta red se centra en cómo, al tener un input (una palabra hablada) la capa previa a la salida del sistema realimenta el algoritmo con el fin de que se registre la nueva forma de pronunciación que recibe y pueda ser incorporada para futuras predicciones a la hora de plasmar la palabra escrita en el output.

Red Neuronal Convolucional [*Convolutional Neural Network*]: Este tipo de red neuronal se centra en el aprendizaje y predicción de recursos visuales. Se alimentan de manera similar a las Redes Neuronales Básicas hacia Adelante. Al tener una imagen determinada como input, la Red Neuronal Convolucional divide la imagen en un número de píxeles determinados y procede a aprender una combinación lineal de estructuración que le permite identificar entradas parecidas futuras. Entre el entramado de capas, se sucede una conversión de un código de colores RGB a una escala de grises, donde cada tonalidad colabora a la hora de efectuar una futura predicción.

Red Neuronal Modular [*Modular Neural Network*]: Como su nombre lo indica, las Redes Neuronales Modulares desglosan el trabajo a realizar en distintas partes Esto quiere decir que, partiendo de un grupo de inputs definidos, las neuronas de las capas ocultas no están necesariamente interrelacionadas entre sí. Esto es bastante útil ya que permite establecer delimitaciones claras entre tareas y, en caso de ser necesario, corregir y retroalimentar o conseguir fácilmente el eslabón a reformular. También permite al algoritmo administrar su memoria virtual de manera inteligente. Según Kishan Maladkar (2018) sólo estamos empezando a ver la implementación de las redes neuronales modulares. Su proyección para los próximos años es prometedora.

El mercado inmobiliario

Tamaño: Según el "Urban Institute", El valor total del mercado inmobiliario residencial de EE. UU. Es de $ 27.2 billones de dólares.

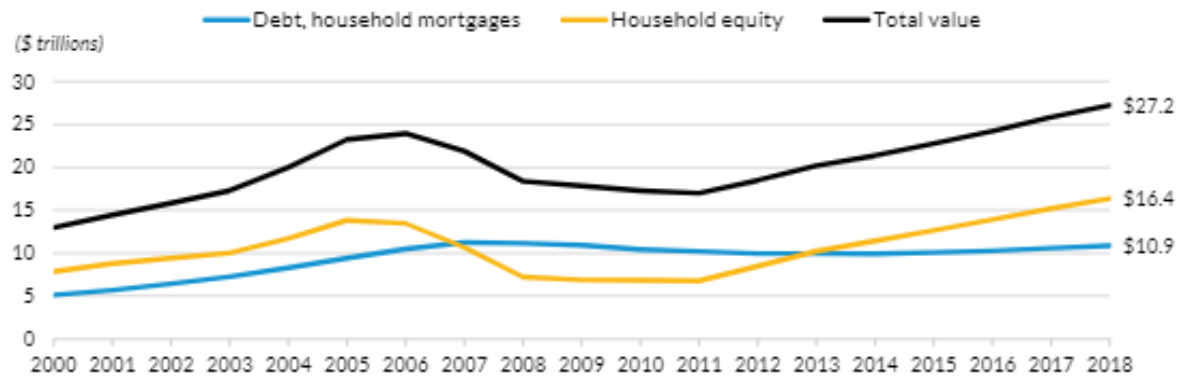

Imagen 1. Valor del mercado inmobiliario de EE. UU. Fuente: (Laurie Goodman, 2019)

Para poner este número en perspectiva; el PIB de EE. UU., el valor total de producción económica de la primera economía del mundo para 2018, fue de $ 21.5 billones. La vivienda es $ 5,7 billones más grande que eso.

Gentrificación: derivación del inglés [*Gentry*] (burguesía). “El proceso de gentrificación consiste en una serie de mejoras materiales en un área histórica de viviendas que hace que todas las condiciones del sector crezcan económicamente.” (Sargatal, 2000)

La gentrificación es un tema del cual se habla cada vez más, aun cuando el término se acuñó en los años sesenta. Factores como la globalización, los tratados económicos y la movilización de personas hacia áreas de trabajo en desarrollo, han hecho que el término cobre una validez impresionante. Y es que una sola condición, puede lanzar a rodar una bola que derive en la gentrificación de una localidad.

Nos viene a la mente una localidad en la que descubren algún yacimiento mineral, esto podría dinamitar el flujo y desarrollo de personas hacia ese sitio. Pero ¿qué pasaría si la extracción de dicho mineral trajera consigo contaminación? ¿sería igual el potencial de gentrificación?

Esta sencilla situación nos permite perfectamente ilustrar las tres etapas de la gentrificación, que funcionan como una suerte de subidas y bajadas. Primeramente, el descubrimiento de dicho mineral traería una fuerte presencia de inversión y la población se desplazaría a la zona, esta etapa se llama nacimiento. En segundo lugar, el alto porcentaje de personas, junto el desarrollo sin medidas de control, acarrearía que muchos ya no vieran al lugar como rentable y decidieran partir. Esta etapa se conoce como abandono. Finalmente, el abandono produciría una caída en precios de renta y producción, por lo que eventualmente, una segunda ola de inversión llegaría al sitio para reformular políticas anteriores y promover un crecimiento más ordenado. Esta etapa final se llama revalorización.

Teniendo la data necesaria, el algoritmo usando los distintos índices y factores, deberá ser capaz de identificar qué parte y de qué ciudades se encuentra en alguna de las tres etapas identificadas de Gentrificación. Este factor condiciona como ningún otro el alza de los precios del mercado inmobiliario.

CBP data: [*County Business Patterns*]: El CBP se refiere a una serie de data anual que recoge, analiza y brinda a los usuarios una cantidad de factores económicos según su división en la industria. Según la oficina de censos de Estados Unidos, la data es bastante útil para estudiar factores de crecimiento económico en áreas determinadas de la geografía estadounidense.

CPI: [*Construction Price Index*]: Este es el índice que determina cuánto del costo total de una remodelación es debido a factores físicos como el tamaño, la ubicación y el proyecto en sí, y cuánto por otra parte, se debe a la tasa de inflación actual. Según la oficina de censos de Estados Unidos, este índice se puede ver de tres diferentes maneras: Primero, cómo el precio se mantiene hoy en función al año uno de la vivienda, en segundo lugar, cuál es la diferencia entre el precio en dólares hoy y en el año uno de la vivienda, y por último, el precio de la vivienda construida hoy con respecto al precio en dólares constante.

Código ZIP: [*Zone Improvement Plan*]: Esto es el código utilizado por el servicio postal estadounidense que designa un cierto sector de la geografía del país. El código ZIP fue inventado en 1963 con el fin de organizar el creciente número de correo que era enviado a lo largo de todos los Estados Unidos. (Scheele, 1970)

Glosario de términos

Estudio Longitudinal: Es un tipo de estudio correlacional de datos donde las variables a estudiar son puestas bajo monitoreo durante un periodo determinado de tiempo, mientras se observa su desarrollo bajo ciertas características. (Cherry, 2019) Este tipo de estudios tiene grandes ventajas al permitir hacer mediciones durante largos períodos de tiempo, en el cual distintos factores externos pueden influir en las relaciones entre variables.

Comunicación Digital Síncrona: Esto significa que la interacción se sucede en tiempo real, y el entorno web permite esto gracias a plataformas de charlas (IRC) y audioconferencias y videoconferencias.

Obtención y Utilización de Recursos: De forma general la obtención y utilización de recursos se da mediante cualquier plataforma web de contenido o transferencia de

ficheros. Como podemos ver, en este caso la comunicación se da en una sola vía, o al existir algún tipo de intercambio, es de manera indirecta.

Conjunto de Datos [*Dataset*]: Esta es la materia prima del sistema de predicción. Es el historial de datos que se usa para entrenar al sistema que detecta los patrones. El conjunto de datos se compone de instancias, de instancias de factores, características o propiedades. Contamos con factores como: la fecha de la remodelación, el costo, el tipo de proyecto, la ubicación, entre muchos otros.

Objetivo [*Objective*]: Esto es el atributo o factor que queremos predecir. El objetivo de la predicción, en nuestro caso, es la ganancia a futuro de un proyecto de renovación de una vivienda.

Época: En el lenguaje de programación relacionado al aprendizaje automático, una época se puede definir como cada corrida de datos que se hace a través de una red neuronal.

Confianza [*Confidence*]: Es la probabilidad de acierto que calcula el sistema para cada una de las predicciones. Este factor de confianza se expresa en porcentajes que pueden estar formados por dos o tres categorías. Por ejemplo, puede estar dado por cómo se relacionan los factores de tipo de proyecto y costo; o ubicación y tipo de residencia. Por lo general los índices de confianza son diversos y ayudan a detectar vulnerabilidades en el código para poder ser reforzadas.

Caracteres [*Features*]: Referido a un conjunto determinado de parámetros con condiciones establecidas, que forman parte del entorno de un problema a resolver.

Datos de entrada: Conjunto de caracteres que se tienen al comienzo de la resolución del problema y se entrelazan con las capas ocultas para generar un resultado.

Datos de salida: Es el conjunto de resultados o salidas de una red neuronal. Se aprecian como la resolución esperada del problema que se quiere resolver.

Capas ocultas: Desarrollo matemático de la red neuronal. Está referido al meollo del algoritmo y puede tener una cantidad diversa de capas internas, cada una genera una operación distinta que deviene en la resolución del problema.

Grupo Entrenamiento: En la resolución de un problema de aprendizaje automático, se generan diversos grupos de datos que sirven distintos propósitos El grupo de entrenamiento se refiere a la primera fase del estudio, donde un grupo determinado de datos se encargan de proporcionar un patrón de pensamiento al algoritmo sin que éste desarrolle una capacidad centralizada o de acción limitada (Grupo de validación). El grupo "entrena" al algoritmo para su posterior implementación.

Grupo de Validación: Aproximadamente un 10% del grupo Entrenamiento puede ser posicionado en un grupo de Validación. Dicho grupo funciona como un grupo de comparación para garantizar que el algoritmo no se vuelva muy especializado.

Grupo Prueba: El Grupo de Prueba abarca un área posterior del estudio. Al ser entrenado el algoritmo, él mismo debe ser probado para garantizar que el proceso fue exitoso. Para esto, se crea un Grupo de Prueba, que exige al algoritmo creado con la finalidad de encontrar posibles fallas, que puedan ser corregidas en un proceso de retroalimentación.

Regresión Ordinaria por Mínimos Cuadrados: Este método trata de predecir nuevos parámetros desconocidos. Recurriendo a la matemática se usa un método de estimación llamado regresión lineal por mínimos cuadrados, que busca trazar una línea de predicción conociendo un grupo de elementos y a partir de ahí formular una aproximación.

Algoritmo de Análisis de Componentes Principales [*Principal Component Analysis*] (PCA): Este algoritmo busca, a partir de datos de entrada, reducir posibilidades de elección en función a parámetros establecidos, de manera que el resultado esperado suceda en función a una salida de datos.

Árbol de Decisión: Extrapolando el concepto general que enuncia la capacidad de generar un modelo que permita, partiendo de una primera hipótesis, plantear distintas alternativas de resolución, el Árbol de Decisión posibilita clasificar una serie de elementos

en el código a partir de una serie de variables. Por lo tanto, cabría esperar en este algoritmo, estructuras del tipo: “en caso de realizar tal o cual acción”, hasta que arroje un resultado entendible y comprobable.

Funciones de Activación: Son funciones matemáticas que hacen la suerte de interruptores entre una interacción (capa) y otra. La función matemática registra el valor que llega de la capa anterior y lo compara con una escala preestablecida. Si el valor supera el valor máximo de la escala, la función da paso a la data para seguir a otra capa que lleva a otra interacción.

4. Metodología y técnicas por utilizar

Con vistas a tener un mejor entendimiento del proyecto, explicaremos a continuación una serie de conceptos importantes a tener en consideración, y los abordaremos desde lo macro hasta lo micro.

Tipo de investigación

Este estudio se inscribe dentro del protocolo de proyecto especial con énfasis en lo tecnológico y la utilidad real, se encuentra enmarcado en una investigación-acción de campo con carácter descriptivo. (Arias, 1999, pág. 21) define la investigación de campo como "aquel estudio que se basa en la obtención y análisis de datos provenientes de materiales impresos u otros tipos de documentos, ya que se utilizan fuentes referenciales". Estas fuentes apoyan y dan sistematicidad y rigor al desarrollo del proyecto. De igual forma cabe señalar que es de investigación- acción, porque propone una descripción, análisis y evaluación de procesos de investigación, donde se pretende desarrollar y resolver un problema que va a ser útil para un conglomerado específico.

Además, se hace necesario aclarar que siendo un proyecto especial no cuenta con un tipo específico y riguroso a nivel metodológico, más bien se presentan procesos intuitivos de corte analítico que dan como resultado objetos de valor tecnológico y de utilidad en el marco de la sociedad actual.

A este tipo de proyecto se le vincula con diseños novedosos y de impacto social, económico y humano, además promueven el desarrollo científico a través de soluciones de alcance en todos los campos de la vida cotidiana. Esta apreciación coincide a cabalidad con el proyecto que se aspira consolidar bajo la aplicación del RoI. Por tanto, el proyecto especial será el protocolo metodológico que más se ajusta al presente estudio, puesto que se tratará de dar solución a una problemática real.

Enfoque de la investigación

El enfoque que tendrá este proyecto se fundamenta en el paradigma cuantitativo para lo cual se utiliza variedad de instrumentos para recoger información y dar sentido al análisis de los datos.

Para este caso en particular la observación y la encuesta a los dueños de varias empresas del ramo de la construcción es muy útil para darle respaldo a todo lo que se pretende desarrollar.

A este respecto, los proyectos especiales son interpretados como "intervenciones que independientemente de su grado de complejidad tienen como propósito específico o especial resolver aquellos problemas que surgen en cualquier ámbito del desempeño humano, con el uso de los conocimientos existentes" (Pérez, 2009)

El conocimiento de la industria inmobiliaria y su incorporación a las nuevas tecnologías son los factores clave en el desarrollo de un modelado de datos que permita conocer el impacto de un proyecto de construcción residencial sobre el valor del inmueble (ROI), a partir de datos proyectados a veinte años en el mercado inmobiliario estadounidense.

Universo, Muestra y Unidad de Análisis

La búsqueda y clasificación de la información histórica de los proyectos de renovación fungen como pieza central en el análisis. Apoyados en la Ley de Libertad de Información, a cada ente gubernamental se le serán solicitados todos los permisos de renovación residenciales desde el año 2000 hasta la actualidad.

Universo: La totalidad de elementos en los cuales se pretende estudiar el comportamiento de Retorno de la Inversión se encuentra localizados geográficamente en los 48 estados continentales de los Estados Unidos. Aun cuando es finito, el universo es tan grande que no se podrá estudiar en su totalidad. De manera que se escogerá un subconjunto para poder llevar a cabo el estudio.

Población: El grupo del cual se procederá a obtener información está definido por todas las construcciones residenciales unifamiliares que tengan al menos una renovación estructural y dos ventas. De manera de poder comparar los precios reales de venta y extraer el aproximado de retorno.

Muestra: El subconjunto de la población que identificaremos como población muestral, está definido por 65 ciudades seleccionadas por sus diferencias poblacionales, étnicas y distribución geográfica dentro de los Estados Unidos.

Alpharetta	Chicago	Fayetteville	Los Angeles	Oklahoma City	San Francisco
Austin	Cleveland OH	Gaithersburg	Las Vegas	Philadelphia	San Jose
Atlanta	Cleveland TN	Gastonia	Memphis	Phoenix	Somerville
Augusta	College Park	Greely	Miami	Portland	St Charles City
Baltimore	Columbus	Hillsboro	Minneapolis	Reading	Stillwater
Bossier City	Cottonwood	Houston	Murfreesboro	Redmond	Stockton
Boston	Dallas	Ithaca	New Bedford	Reno	Surprise
Boulder	Denver	Jacksonville	New Orleans	Richland	Tacoma Lakewood
Cape Coral	Des Plaines City	Kenner	Newark	Saint Paul	Walnut Creek
Carson	Eden Prairie	Kissimmee	New York	South Dakota	Washington
Charlotte	Edmond	Lancaster	North Tonawanda	Seattle	

Tabla 1. Lista de las 65 ciudades seleccionadas.

Parte importante del criterio de selección está relacionado a la calidad de información proporcionada por los entes de la ciudad. Ninguna de ellas está obligada a seguir estándares en los procesos, así que un proceso de reconocimiento y verificación es necesario.

5. Hallazgos / desarrollo

La primera etapa del proyecto conlleva la recolección de la información referente a las características socioeconómicas de las ciudades, los permisos de construcción y las posibles características a evaluar, esta data puede ser encontrada mediante diversas formas, pero siempre enfocándonos en buscar páginas del gobierno. Esto nos garantiza dos puntos principales, uno, que la información que buscamos es oficial y actualizada, y dos, que es completamente gratuita y por ende su uso es legal para nuestro proyecto.

Identificación de los vectores:

Partiremos de la generación de cuatro vectores principales de información, los cuales son necesarios para el cálculo del ROI y que se acoplan al algoritmo de machine learning a fin de poder cubrir el espectro más amplio posible y garantizar una estimación confiable.

1. Características de la ciudad **[*City feature vector*]:**

Este vector almacena un compendio de información relacionada a la ciudad, vista como un todo y sin distinciones internas.

Es el vector más general y nos proporciona una visión holística del entorno, toda la data obtenida para dicho vector es extraída de páginas gubernamentales de los Estados Unidos o requerida a través de solicitudes FOIA en los organismos correspondientes. Ninguna fuente de terceros es utilizada en la mezcla. Se detallan las fuentes utilizadas para la obtención de cada uno de los factores considerados:

- Población:

  La oficina de censos de Estados Unidos (*Census Bureau)* provee, hasta el año 2018, cifras sobre la cantidad de habitantes en el país. Dividida geográficamente por estados, áreas metropolitanas, ciudades e incluso códigos postales. Se utilizan principalmente 3 subconjuntos de datos: La lista de cantidad de habitantes por Estado, la tabla de poblaciones por código postal y la lista de densidades de población.

- Tasa de desempleo:

  Para la tasa de desempleo, nos valemos de la data aportada por la Oficina de Estadísticas Laborales (BLS). A principios de cada mes, del Departamento de Trabajo de EE. UU. a través de la BLS, anuncia el número total de personas empleadas y desempleadas en los Estados Unidos, junto con sus respectivas características socioeconómicas.

- Distribución demográfica:

  Para la distribución demográfica nos apoyamos en el World Urbanization Prospects 2018, de Naciones Unidas, pues proporciona una data completamente actualizada sobre la distribución demográfica en el país. La importancia de este factor es crucial si tomamos en cuenta que los estados unidos es un país altamente urbanizado, con el 81 % de la población residiendo en suburbios y ciudades a la fecha de 2014, mientras la tasa de urbanización mundial es del 54 % (Department of Economic and Social Affair, 2019)

- Margen de criminalidad:

  En el caso del margen de criminalidad, la oficina de investigaciones federales (*Federal Bureau of Investigation - FBI)* proporciona un índice de criminalidad según estados, por delito, por región y por agencia local, así como también el tipo de crimen según su intensidad, incluyendo crímenes a la propiedad privada, de manera anual.

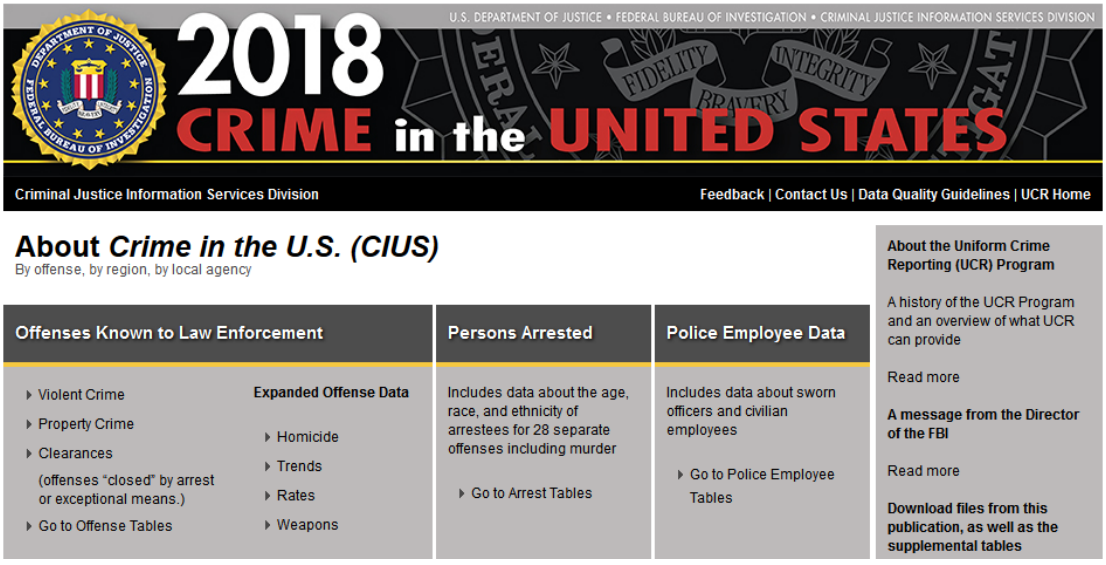

Imagen 2. Índice de criminalidad. Fuente: FBI, UCR

- Costo básico de vida:

Para el costo básico de vida, la Oficina de Estadísticas Laborales publica regularmente el Índice de Precios al Consumidor (CPI). Esta es una medida del cambio promedio en el tiempo, en los precios pagados por los consumidores urbanos por una canasta de mercado de bienes y servicios de consumo. Los índices están disponibles para los EE. UU. y se encuentran además datos de precios promedio para servicios públicos, combustible, salud y recreación.

Imagen 3. Valores de CPI. Fuente: U.S. Bureau of labor.

- Fluctuación del mercado de bienes raíces:

Este es uno de los numerales más importantes por lo cual no se usará solo una fuente pública, sino que se enfrentan valores otorgados por el departamento de desarrollo urbano *(Department of Housing and Urban Development - HUD)* del gobierno con data privada de Zillow, que mantiene un conjunto de datos de costos de viviendas y alquileres para uso público (ZHVI). Para nuestro análisis utilizaremos los datos del precio medio de la vivienda, disponible en 13.105 códigos postales por mes, donde podremos observar la distribución y el precio de todas las construcciones residenciales vendidas hasta el año 2018.

- Producto interno bruto. PIB.

  Como sabemos, el PIB o producto interno bruto, es un valor estimado de la producción interna de una región en un período determinado. A través del Banco Mundial, organismo internacional encargado del manejo y la gestión de datos relacionados con la producción y manejo de recursos de un país en términos de unidades monetarias, podemos encontrar los datos relacionados al PIB de cada estado, en función a su producción anual.

  En el siguiente mapa podemos ver la distribución de PIB dentro de Estados Unidos. Caben destacar estados como Dakota del Norte y Alaska, donde existen yacimientos de petróleo.

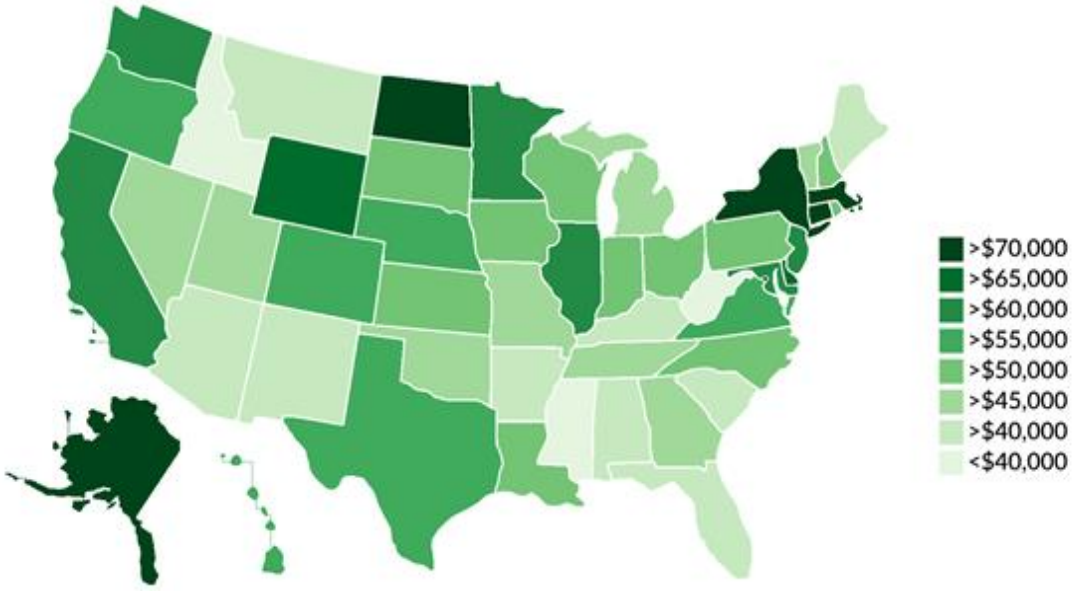

Imagen 4. Comparación de PIB según Estado para el 2018. Fuente: Bureau of Economic Analysis

- Ingreso per cápita.

  A diferencia del PIB, el ingreso per cápita mide la cantidad de ingresos promedio que una persona recibe en una cantidad determinada de tiempo.

  Para calcular el valor estimado del ingreso per cápita, tomamos los valores de PIB obtenidos del Banco Mundial, y los valores de densidad poblacional para cada estado obtenidos de la oficina de censos de Estados Unidos (*Census Bureau)*.

$$\text{Ingreso per cápita} = \frac{pib}{Cantidad\ de\ población}$$

2. Características del vecindario **[*Neighborhood feature vector*]:**

Esta información es específica al lugar de estudio y contiene características sobre los servicios existentes en el área.

Estimar el impacto de los servicios no es sencillo y es, además, un arte algo subjetivo. Según Rivas, "La calificación de la escuela de un vecindario es un reflejo del vecindario y sus características, demografía y económica. Por ende, afecta los precios de sus viviendas, principalmente a través del valor del terreno." (Rivas, 2019)

El análisis del vecindario se centrará entonces en el impacto de los servicios en el precio de la vivienda. Este impacto se mide entonces a partir de la correlación del precio de la vivienda con la distancia del servicio en cuestión. Nótese que el comportamiento no será lineal, por lo tanto, más servicios o menos distancia no siempre significa mayor precio. El secreto consiste en encontrar el "sweet spot" en donde los servicios estén disponibles, pero que no sean una carga. Nadie quiere vivir justo al lado de un hospital, sin embargo, si se encuentra a una distancia segura, aumentará el valor de la propiedad.

Detallaremos las fuentes utilizadas para la obtención de cada uno de los factores considerados:

- Sector educativo: Escuelas y universidades.

Para este particular no existe una fuente centralizada que provea la información completa sobre tan amplio sector por lo que es necesario concatenar las siguientes:

El Centro Nacional de Estadísticas de Educación (NCES) que es la principal entidad federal para recopilar y analizar datos relacionados con la educación. De esta fuente se tomarán la mayoría de los datos necesarios.

La Oficina de Ayuda Federal para Estudiantes elabora un mapa de universidades, la primera herramienta de búsqueda geográfica publicada por IPEDS (Sistema Integrado de Datos de Educación Postsecundaria) que proporciona acceso a más de 7,000 escuelas certificadas, de pregrado y posgrado.

El Departamento de Educación de los Estados Unidos (*US Department of Education*) pues pone a la disposición estadísticas con respecto a la calidad de las instituciones educativas. Esto es válido para cualquier nivel educativo, sea básica, media o diversificada.

El Ranking de US News and World Report también fue revisado, pero está limitado a las primeras 200 universidades clasificadas, mientras que las demás no están clasificadas.

Para nuestros análisis entonces, tres factores del área de las instituciones educativas serán importantes. Primero, su presencia en las cercanías de la edificación en cuestión, luego el tamaño de dicha institución educativa y por último, pero lo más importante, la calidad que la identifique.

- Sector salud: Clínicas y hospitales.

La Asociación Americana de Hospitales (*American Hospital Association - AHA)* provee data con relación a la distribución de clínicas y centros hospitalarios afines en Estados Unidos. Esta data no solo abarca la ubicación del centro, sino también el tipo de ayuda que brindan. De esta manera se puede tener un amplio espectro de cuánta importancia representa cada centro de salud.

El Grupo Leapfrog es un organismo de salud sin fines de lucro encargado de proveer información concisa con respecto a la distribución de centros de salud en Estados Unidos. Provee data actualizada con respecto a los hospitales más grandes en el país. Esto es válido tanto para alternativas públicas como privadas y considera también otras alternativas de ayudas humanitarias.

En resumen, los factores a considerar en orden de importancia para un centro de salud son:

- Ubicación.
- Volumen de personas tratadas en un período determinado.
- Espectro de salud y servicios que abarcan.

- Sector Comercio: Tiendas y Centros Comerciales.

  Debido a que el sector comercio es bastante amplio, los datos necesarios se obtienen de las siguientes fuentes:

  Por un lado, los pequeños comercios y establecimientos, que podrían ser difíciles de conseguir. Para este apartado podemos encontrar información gracias al departamento de datas de Estados Unidos, el cual provee información con respecto a rubros como:

  - Mercados de frutas y vegetales.
  - Tiendas de comestibles.
  - Centros comerciales.
  - Ferreterías.

  Con respecto a las grandes cadenas comerciales, la cadena de distribución "*MWPVL International"* provee información concerniente a las principales tiendas de comestibles y productos de Estados Unidos.

  El sector de grandes tiendas en Estados Unidos se ha diversificado en los últimos años, con tiendas como Wallmart donde se puede encontrar una gran variedad de productos comestibles, aparatos eléctricos y hasta mobiliario. Esta tendencia se ha incrementado en los últimos años, por lo tanto, teniendo estos dos frentes a la hora de gestionar la data referida a comercios, tenemos un espectro de acción bastante amplio.

- Sector Recreación: Parques, Cines, Teatros, Centros Culturales.

  El departamento nacional de parques y recreación (*National Recreation and Park Association - NRPA)* proporciona data relacionada con los parques existentes en zonas determinadas de la geografía estadounidense.

Por lo general, cada ciudad tiene su propio departamento de parques y recreación por lo que de nuestra parte es necesario entrar en contacto de forma individual. Entre los datos que la data proporciona se encuentran:

- Centros culturales.
- Teatros.
- Museos.
- Centros y áreas deportivas.
- Ferias y festivales.
- Terrenos destinados a actividades recreativas en general.

Como factores fundamentales para nuestro estudio, tenemos la distancia y la cantidad de áreas de recreación cerca de la propiedad en cuestión, teniendo como prioridad las actividades recreativas para toda la familia.

3. Características de la Propiedad **[*Property feature vector*]:**

Este vector contiene información en relación con la residencia donde se realizará el proyecto de renovación, sin considerar aún el proyecto en sí. Se toman en consideración tres dimensiones principales:

- El tipo de propiedad:
    - Unifamiliar: Edificaciones destinadas a alojar una sola familia. Casas con jardín y espacio entre unidades.
    - Multifamiliar: Departamentos y condominios donde las unidades están adyacentes unas de otras.

- La distribución:
    - 1 habitación.
    - 2 habitaciones
    - 3 habitaciones
    - 4 habitaciones.
    - + de 4 habitaciones.

- Nivel de la vivienda.

  La comparación entre la propiedad en cuestión y los vecinos entra en juego en este numeral en donde se podrá ver el impacto de la diferencia del tipo de construcción en los retornos de inversión al remodelarla.

  - Tier 1: Nivel de precio **superior** entre viviendas dentro del misma área.
  - Tier 2: Nivel de precio **promedio** entre viviendas dentro de la misma área.
  - Tier 3: Nivel de precio **inferior** entre viviendas dentro de la misma área.

- Tamaño de la propiedad, tamaño del lote.

- Fecha de construcción.

  Si es un apartamento la fecha de refacción de la unidad y la fecha de construcción del edificio completo.

  Como podemos ver, nuestro criterio de selección para cada uno de los vectores busca ir desde lo macro a lo micro del problema. De esta manera se asegura contar con la mayor información posible, que será alimentada al algoritmo para predecir el mejor resultado.

  Un factor intrínseco del mercado americano son las llamadas “Burbujas” en donde los precios son inflados artificialmente por múltiples razones y pueden ocasionar perdidas sustanciales a todos los involucrados. Para este proyecto en particular no se incluyen valores directos para predecir este fenómeno debido a su impredecibilidad, pero al considerar factores como las condiciones socioeconómicas y en particular el histórico de permisos de renovación actualizado podemos tomarlo en consideración indirectamente.

4. Características del Proyecto **[*Permit feature vector*]**:

Este punto se centra en la actividad de remodelación que se quiere realizar. Un proyecto de renovación se puede definir usando tres características:

- El espacio:

Esta variable toma en consideración cada posible área de una residencia promedio estadounidense. Áreas como casa de visitas, medio baño y cuarto de usos múltiples pueden parecer extrañas para nosotros, pero son bastante comunes en las residencias americanas.

Espacio	
1	Baño
2	Medio baño
3	Cocina
4	Comedor
5	Cuarto de lavado
6	Oficina
7	Habitación
8	Cuarto de usos múltiples
9	Cuarto de invitados
10	Sala
11	Salón
12	Pasillo
13	Sótano
14	Garaje
15	Porche
16	Jardín
17	Piscina
18	Techo
19	Casa
20	Exterior

Categoría	
1	Reemplazo
2	Adición
3	Alteración
4	Construcción
5	Instalación
6	Demolición
7	Mantenimiento
8	Remodelación
9	Mejora
10	Encierro
11	Traslado
12	Expansión
13	Inspección
14	Eliminación
15	Reparación
16	Diseño
17	Reforzamiento

Tabla 2. Opciones para espacio y categoría. Fuente: Elaboración propia.

- La categoría:

  La categoría describe la naturaleza de la modificación a realizar en las renovaciones. Ya sea la construcción de una nueva edificación o la mejora de una habitación particular, el campo de categoría será el encargado de identificar estas diferencias.

- El tipo:

  El tipo es la sección más específica del proyecto. Aquí se muestran una serie de proyectos a realizar en cada área determinada y que van ligados al proyecto general o su categoría.

Tipo			
1	Plomería	23	Cerca
2	Electricidad	24	Sistemas contra incendios
3	Panel eléctrico	25	Chimenea
4	Cableado	26	Techo interno
5	Suich	27	Piso
6	Iluminación	28	Fundación
7	Gabinete	29	Horno
8	Closet	30	Mecánica
9	Piso exterior	31	Cloaca
10	Letrero	32	Ducha
11	Puerta	33	Lavabo
12	Ventana	34	Canalón
13	Ducto	35	Elevador
14	Calefactor	36	Muro de contención
15	Aire acondicionado	37	Estufa
16	Alarma	38	Tanque de agua
17	Antena	39	Barrera
18	Ventilador	40	Nivelación
19	Panel solar	41	Concreto
20	Conexión a gas	42	Pintura
21	Aspersor	43	Revestimiento

22	Escalera	44	Enyesado

Tabla 3. Opciones tipo de renovación. Fuente: Elaboración propia.

Cabe destacar que al algoritmo tiene restricciones en este caso, pues existen tipos y categorías que no pueden coexistir juntos en un proyecto. Por ejemplo: No se podría tener un proyecto de fundación en un techo.

Este será el punto de pivote principal de todo el algoritmo. De acuerdo con todos los otros vectores que consideramos fijos, ¿cuál será la mejor combinación de proyecto para que el retorno de la inversión sea el más alto posible?

Dos ejemplos simples que se pueden intuir de la yuxtaposición de dos vectores serían:

1. Si la distancia a un puerto marítimo o aeropuerto está en el rango de ruidos molestos, un proyecto de renovación de ventanas y puertas tendrá un RoI positivo.
2. Si el año de construcción de la edificación data de unos pocos años, remodelaciones no visuales como la plomería y la electricidad tendrán RoI negativos.

Estos son ejemplos muy básicos para ilustrar como uno de los puntos de data de uno de los vectores elegidos puede hacer un impacto claro en el retorno. Ahora para el procesamiento humano será imposible combinar las múltiples fuentes citadas en los vectores anteriores de manera de comparar el impacto de todas juntas en los proyectos de renovación. De aquí nace la necesidad de usar Inteligencia Artificial para asistirnos en el cálculo.

Un punto importante para recalcar en este numeral es el caso particular de los proyectos de domótica. Debido a que la normativa todavía no fue ajustada para incorporar estos proyectos la gran mayoría son clasificados en las categorías de “Electricidad” o alguna otra clasificación general por lo que no se tomaran en consideración directa para el desarrollo del algoritmo.

Normalización de los datos

Todas las fuentes antes descritas proporcionan información segregada y con multitudinarios formatos. Así, es de vital importancia, que antes de alimentar el algoritmo la información esté lo más heterogénea posible de manera que el algoritmo la pueda digerir satisfactoriamente.

El proceso aplicado tiene varias etapas. En primer lugar, la información pasa por un proceso de geolocalización, donde la información geográfica es revisada y unificada en formato. Posterior a esto, la información pasa por un proceso de limpieza general, y seguidamente, se aplica una gestión de reporte con la cual se podrá retroalimentar el proceso.

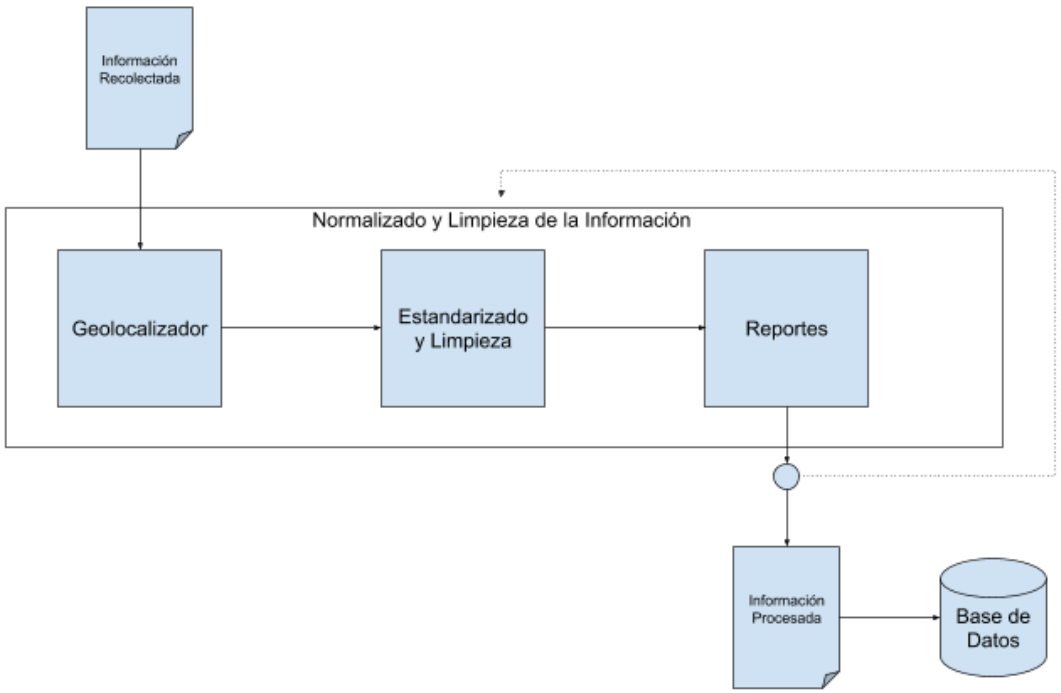

Imagen 5. Ilustración del proceso de normalizado. Fuente: Elaboración propia.

- El Geo-Localizador:

  Es una herramienta creada por el Departamento de Censo de los Estados Unidos. Esta se basa en una base de datos con una extensión que permite almacenar relaciones geográficas. Gracias a esto, el Departamento de Censo ha recopilado la información de la geografía de Estados Unidos y la ha hecho pública de manera que se puede, a partir de una

dirección en un formato preestablecido, consultar la herramienta y extraer más información, como por ejemplo las coordenadas de dicha dirección.

Esta herramienta provee una medida en cuanto a la seguridad que tiene la información que se está devolviendo. Esta medida es un número entre 0 y 100. Cuanto más baja es la medida, más seguro es el resultado de la base de datos. Usando esto, lograremos entender cómo se comporta la herramienta con respecto a su medida de seguridad. Usando diferentes direcciones, tomando como ejemplos direcciones correctas en un buen formato y en malos formatos, direcciones erradas, con algún dato que no sea correcto, y también direcciones correctas, pero con un dato faltante.

Tomemos como ejemplo la dirección 1331 Cavendish Ct, Charlotte NC 28211:

Entrada	Rating	Ciudad	Estado
1331 Cavendish Ct, Charlotte NC 28211	0	Charlotte	North Carolina
1331 Cavendish Ct, Charlotte NC	1	Charlotte	North Carolina
1331 Cavendish Ct, NC 28211	9	Charlotte	North Carolina
1331 Cavendish Ct, Charlotte 28211	0	Charlotte	North Carolina
1331 Cavendish Ct, 28211	32	Avon	Connecticut
1331 Cavendish Ct	27	Avon	Connecticut

Tabla 4. Ejemplos de normalización de las direcciones.

En esta muestra podemos ver como los ratings van variando cuando eliminamos algunos datos. Incluso, podemos notar que, en la medida que falten datos, se sucederán errores. En este caso en particular, la dirección original fue trasladada al estado de Connecticut.

Con esta muestra, revisaremos las medidas y tomaremos 2 umbrales de tolerancia para mejorar y enriquecer la información. El primer umbral es un umbral muy bajo, que

indica casi la perfección de la dirección. La decisión que tomaremos es, si la dirección tiene suficiente seguridad, podremos usar la información retornada por la herramienta de geolocalización, y sustituir la nuestra (en los casos pertinentes como lo son ciudad, estado, código postal, etc.). Además, podremos enriquecerla con las coordenadas provistas por la herramienta. El segundo umbral es menos permisivo, ya que es más alto y permite direcciones que no están totalmente perfectas. En este caso, lo que haremos es solo incluir la información que nos provee la herramienta, si está faltando para nosotros, y usaremos las coordenadas provistas. Todo lo que está por encima del segundo umbral, no se tomará en cuenta.

La data recopilada gracias a las ciudades será capaz de relacionar cada parámetro en función al código ZIP, lo que hace sencillo establecerlo como un punto de control para enlazar toda la data. A fin de facilitar el número de interacciones del usuario con el software, se requiere en este punto, que el usuario solo introduzca el código ZIP perteneciente a su residencia.

Al terminar este paso, tendremos ya la información enriquecida geográficamente y podremos pasar al siguiente paso, que es la limpieza de esta.

- Estandarizado y limpieza:

  La información que es recolectada proviene de muchas fuentes, y además es, casi en su totalidad, generada por un humano detrás de un escritorio, lo que nos genera 2 problemas principales. El primero, la información viene en muchos formatos distintos, un mismo número se puede representar de muchas maneras (172.563,2 o 172,563.2 o 172563.2, etc.). Igual pasa con el nombre de una ciudad que puede escribirse de varias maneras (Port Saint Lucie o Port St. Lucie). Como estos ejemplos hay muchos más, por lo que es necesario llevar la información a un mismo formato.

  El segundo problema que nos encontramos es que la información es generada casi siempre por un humano, lo que la hace más propensa al error. Confusiones como un número de código postal mal escrito, o un problema de ortografía, de tipeo y muchos otros. Por esto, es necesario aplicar un proceso de limpieza a la información, en donde se eliminan caracteres no deseados, los datos incoherentes (números donde solo debería haber texto y viceversa), ubicaciones erróneas, entre otros.

Para resolver este problema de múltiples formatos y errores en la información, usamos un proceso que busca estandarizar el formato en el que se presenta la información, y que además permite corregirla.

Para esto, se toman muestras de múltiples ciudades y se estudia la naturaleza de cada uno de los campos que son recolectados. En este estudio se busca definir aspectos como el tipo y alcance de cada dato (como valores posibles, máximos, mínimos, etc.). Al tener esto definido para cada uno de los campos recolectados, se construye un algoritmo que procesa cada dato, y aplica las reglas de negocio previamente establecidas en el estudio, eliminando la información que no es válida, estandarizando el formato, limpiando los caracteres no deseados, entre otros procesos.

El algoritmo por su parte, busca mejorar la información, y en los casos en donde hay falta de ésta y que además puede ser extraída de otro campo, por ejemplo, del código postal, este proceso es aplicado para mejorar y enriquecer la información recolectada.

Además, el algoritmo también busca modularizar la información. Aspectos como dividir las fechas, de manera que no se tenga una fecha entera sino 3 campos - donde uno es el día, otro el mes y por último el año. También de manera similar sucede con las direcciones - se extraen los componentes de la dirección (numero, tipo de calle, nombre de calle, dirección, etc.), para su posterior procesamiento.

Veamos unos ejemplos del proceso con los costos:

Entrada	Salida
123.456,3	123456.3
1,456,8	1456.8
185.8	185.8
83	83

Tabla 5. Ejemplos de normalización de los costos.

Con los números de teléfono:

Entrada	Salida
(260) 587 4679	2605874679
260-587-4679	2605874679
260-5874679	2605874679
260 587 4679	2605874679

Tabla 6. Ejemplos de normalización de los teléfonos.

Con el código postal:

Entrada	Salida
544	00544
123	-----
33990	33990
33990-6717	33990

Tabla 7. Ejemplos de normalización de los códigos postales.

Podemos ver en los ejemplos anteriores, como la información es limpiada y llevada a un mismo formato. Cabe destacar que en el ejemplo del código postal una de las salidas está vacía, y es debido a que el código postal 00123 no existe en los Estados Unidos, por lo que nuestro algoritmo lo elimina.

Al terminar este proceso, podremos contar con que tenemos información que está en un formato estándar y también está limpia, con lo que podremos alimentar la base de datos que tenemos.

Antes de proceder a la carga de datos a la base, se realiza un último proceso que no modifica la información, pero sí es importante para la retroalimentación del proceso.

Un sistema de reporte pasa por los archivos de entrada y salida, comparándolos e informando de los campos que fueron eliminados y agregados, de manera que el reporte luego sea revisado por los administradores del proceso, los cuales pueden tomar decisiones informadas respecto a cómo mejorar el proceso para disminuir la cantidad de datos eliminados erróneamente, y aumentar la cantidad de datos eliminados correctamente.

Para esto nos valemos del código ZIP. Estados Unidos utiliza el código ZIP como una forma de estandarización y seguimiento geográfico dentro del país. Con una utilidad inicial orientada a la distribución de correo, el código ZIP se ha adaptado a la era digital permitiendo una vía rápida de búsqueda de direcciones en tareas de geolocalización.

El segundo grupo pertenece a los vectores de conocimiento de preferencias de usuario (Vectores de permiso y propiedad). Aquí, la interacción del usuario con el software debe ser mucho más extensa, ya que requiere por su parte la inserción de información precisa concerniente a su caso.

Elección del Algoritmo

Todas estas nuevas formas de hacer las cosas tienen un punto en común, este es el internet. La web se ha adueñado de muchos espacios a lo largo de los años y si la encauzamos de forma correcta, puede ofrecernos grandes beneficios que harán más fácil nuestras vidas. El aprendizaje automático, al ser acondicionado con la información y maleabilidad de la web, nos permitirá crear patrones predictivos al conocer una cantidad predeterminada de datos. Estos patrones se generan alimentando al algoritmo con ejemplos reales según las diferentes posibilidades, enseñándole así la mejor manera de pensar.

Ahora bien, existen distintos tipos de aprendizaje automático que se adaptan a diferentes necesidades según el problema que se quiera resolver. Pero, basándonos en las características de este proyecto, hemos decidido trabajar con Soporte Vectorial. El Soporte Vectorial es un tipo de algoritmo de aprendizaje automático que se basa en técnicas de

regresión y es de carácter supervisado. En líneas generales, se basa en una división de datos generada para crear un medio de categorización.

¿Por qué escoger el Soporte Vectorial? La principal ventaja que nos aporta el Soporte Vectorial es la capacidad de poder gestionar y hacer predicciones correctas aun contando con pocos datos en un determinado conjunto. Esto es esencial en nuestro proyecto ya que, si bien las ciudades son más que amables al aportar la información de la que disponen, en algunas ocasiones no se cuenta con una gran cantidad de data de un determinado rubro. Esto es por razones tales como la incapacidad de clasificar años anteriores, errores de procesamiento al cambiar la data de analógica a digital, entre otras.

Veamos un grupo de datos como puntos en un plano, estos puntos pueden o no tener criterios en común.

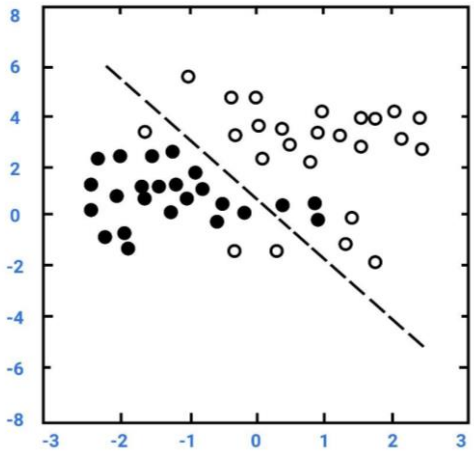

Imagen 6. Grupo de datos clasificados en función a una función de una recta.

En la imagen 5 podemos ver un ejemplo de cómo se clasifican los datos tomando en cuenta sólo una dimensión - los puntos pueden estar llenos o vacíos - Esto se hace mediante la adición de la ecuación de una recta, que separa los datos en dos grupos. Es de hacer notar que no todos los puntos llenos quedaron a la izquierda de la recta, en función al grado de dispersión de los datos, la exactitud y la precisión de la categorización de los datos, varía significativamente.

El algoritmo entonces genera una función principal que delimita cada categoría. La función principal genera dos funciones paralelas a una distancia determinada llamadas hiperplanos.

En el caso de nuestro proyecto, los hiperplanos se generan en función de factores tales como el tipo de renovación, dividiendo los datos en cuanto a baños, cocinas, o a la ciudad en donde se encuentre el inmueble.

Con esto, se generan distintos planos de comportamiento que permiten prever hasta dónde un proyecto de remodelación será rentable en función de la ciudad y a la idea deseada.

Selección de las ciudades

Es importante hacer notar que la elección de las ciudades no solo está enfocada geográficamente sino demográficamente. Con esto podemos proponer una alternativa para el resto de las ciudades.

Imagen 7. Distribución en mapa de las 65 ciudades estudiadas. Fuente: Elaboración propia.

Si un usuario desea buscar la información de cuánto sería el costo de una remodelación y su ciudad no está entre las 65 ciudades estudiadas, el algoritmo simplemente toma las 5 más cercanas y calcula una media. Esto ofrece un cálculo bastante acertado en función a la cercanía e índice demográfico de la región.

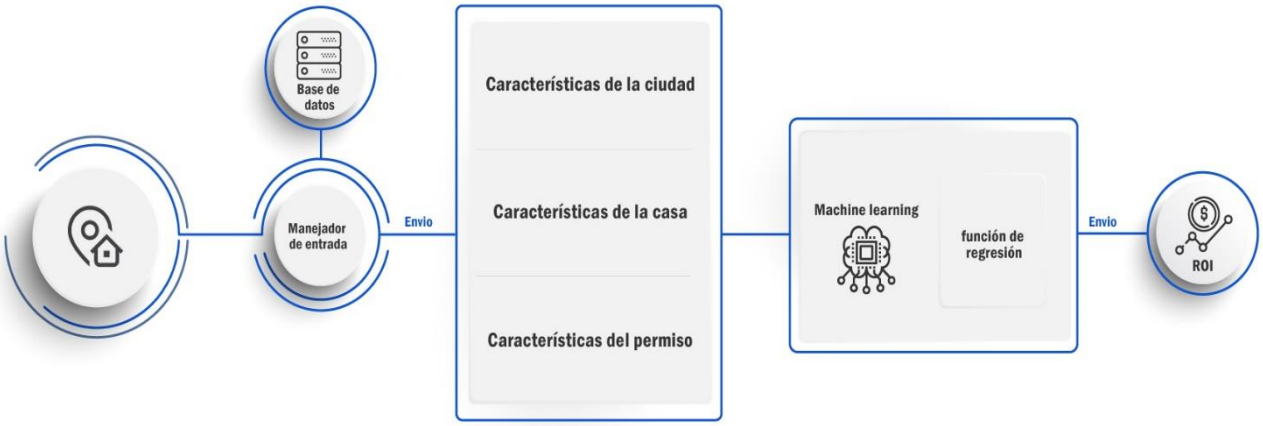

Imagen 8. Diagrama de funcionamiento del Algoritmo. Fuente: Elaboración propia.

Retroalimentando el algoritmo

Una vez alimentado el algoritmo con toda la información necesaria, el siguiente paso es ponerlo a prueba. Sabemos que nuestro grupo total de datos posee características propias, de las cuales cada una aporta una parte de información extra al algoritmo para ayudar con la predicción. Al empezar a entrenar el algoritmo, formamos dos grupos con los datos, uno llamado entrenamiento y otro llamado prueba. El conjunto de prueba tendrá, aparte de todas las características necesarias para el cálculo final del ROI, un apartado extra con el valor real del ROI. Es importante destacar que este valor no es calculado por el algoritmo, sino que ha sido deducido previamente mediante la ecuación antes mencionada.

La diversidad de cálculos ante distintas condiciones es necesaria que se realice, para garantizar que se calcula un ROI correcto y que los valores no divergen de un rango real. Se perfilan tres tipos de data para poder visualizar mejor los planos en los que se mueven las ganancias:

- ROI min: muestran los valores máximos de pérdidas que podría conllevar una remodelación. Por extraño que parezca, si un proyecto determinado se realiza en un área de baja necesidad y con una gran cantidad de modificaciones, puede generar pérdidas monetarias.
- ROI Max: representa el margen máximo de ganancias que un determinado proyecto podría acarrear.

- ROI promedio: Se obtiene al sumar todos los valores de la muestra y dividirlo por el número de valores que se estudiaron en la misma.
- ROI de la mediana: Se refiere al valor medio calculado entre una cantidad de datos determinada. Si 100 datos son evaluados, la mediana es el valor que atañe al valor 50, después de ser ordenados.

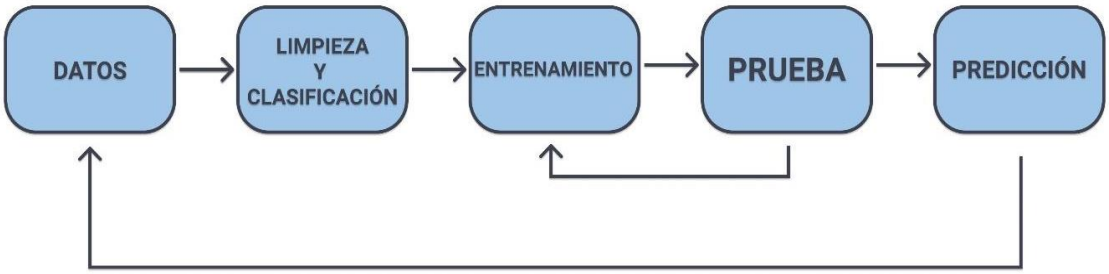

Imagen 9. Pasos del proceso de aprendizaje automático. Fuente: Elaboración propia.

Como podemos ver en la Imagen 9, se crea un bucle cerrado entre los módulos de entrenamiento y prueba que es la columna vertebral de la educación del algoritmo. Por un lado, al módulo de entrenamiento le muestra al algoritmo la información y la interrelación de los datos. Y por el otro lado, el módulo de prueba le dice al algoritmo cuán acertadas son las decisiones que está tomando y si éstas necesitan ser revisadas.

Cuando se empiezan a encontrar errores, se puede encauzar de nuevo al algoritmo alimentándolo con data específica para lograr cambiar la tendencia errónea. Este proceso parte del echo de calcular retornos de inversión de prueba, con los cuales se estima un porcentaje de precisión.

Experimentación y ajuste:

Al momento de empezar a realizar labores de ensayo, el algoritmo tomará al grupo de prueba y se centrará solo en las respuestas, es decir, en los valores de ROI que se le suministraron manualmente. Con esto ajustará el modelo y empezará a predecir por su cuenta, poco a poco, con las características de los datos, siempre comparando el valor del ROI de predicción con el valor del ROI real.

Como el algoritmo buscará minimizar esta diferencia, realizará una cantidad de iteraciones determinadas con el fin de reducir el error y acercarse lo más posible al valor real del ROI. Finalmente hará una última predicción y mostrará qué tan lejos están ambos valores. Esto nos ofrece la posibilidad de fijar un margen de error mínimo. Si la diferencia entre los valores está en el rango aprobado, la prueba se considera exitosa y se pasa al siguiente conjunto de datos. Si por el contrario la diferencia se encuentra fuera del rango, se deberá ajustar la forma en que los datos son suministrados al algoritmo y se procederá de nuevo a realizar otra prueba.

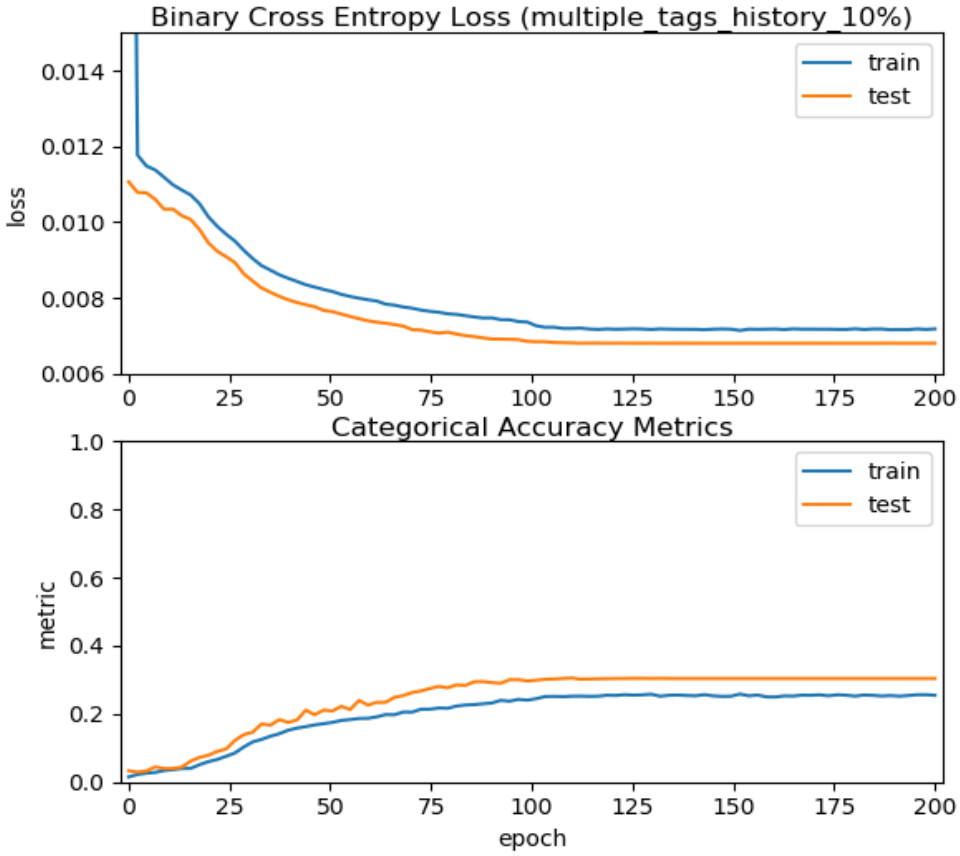

Imagen 9. Comportamiento del algoritmo durante el aprendizaje. Fuente: Elaboración propia.

Las gráficas de la historia para el entrenamiento de los datos originales (para 2.451.980 muestras), arroja características que hay que hacer notar. La idea detrás de estos experimentos es estudiar cómo se comportan las curvas de entrenamiento para distintos valores y determinar cuál es la mejor configuración de la red.

Aun cuando la curva de exactitud (métrica) luce como que converge, en realidad, al ajustar la escala gráfica a valores normalizados, las curvas convergen a valores por debajo de lo esperado (apenas 30%), no pareciera entender todo el conjunto de datos de entrenamiento correctamente.

Al agregar una nueva capa de células recursivas (LSTM):

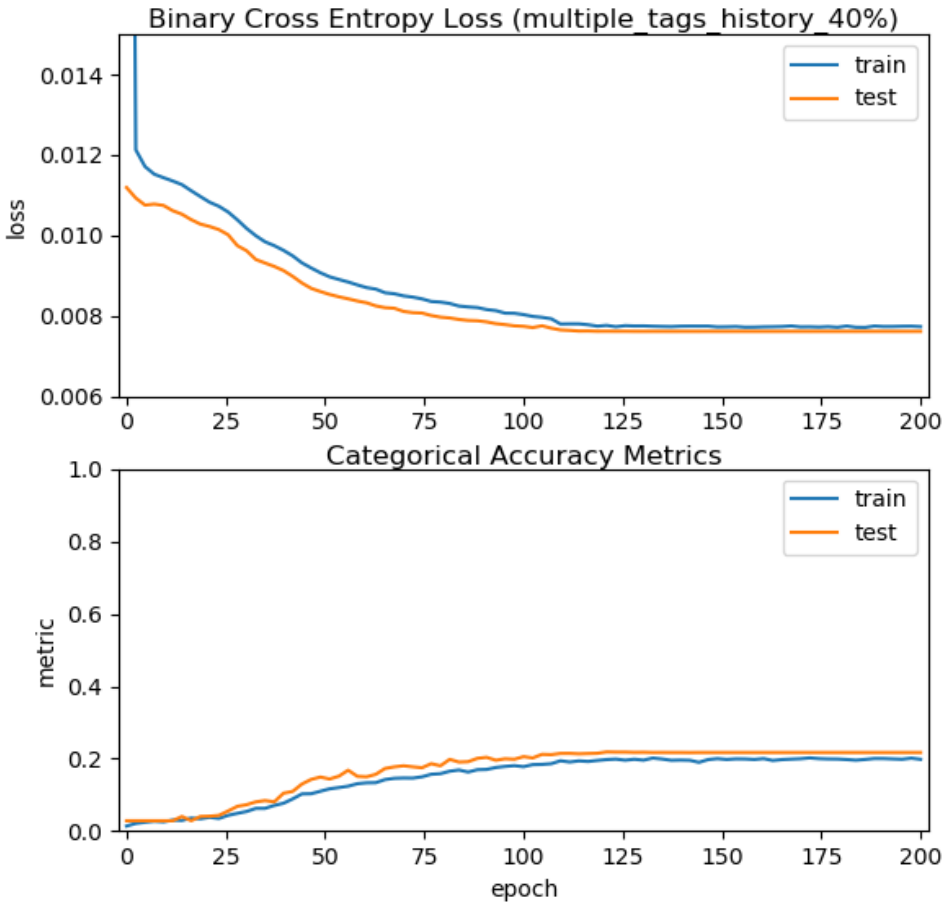

Imagen 10. Comportamiento del algoritmo luego de agregar otra capa LSTM. Fuente: Elaboración propia.

La curva de la función objetivo (pérdida) correspondiente a la muestra de validación, es superior a la de los datos de entrenamiento. Esto puede ser indicativo de que el porcentaje de las muestras de validación (10%) no es significativo. Al aumentar el porcentaje de ejemplos (20%, 30% 40%) las curvas se acercan, pero sigue habiendo una discrepancia, aunque confirma la hipótesis inicial.

Las curvas convergen rápidamente a valores bajos de error. Como se espera dada la complejidad de los datos, la curva tiene un "plateau" o un "lomo".

Ilustrando el procedimiento

La información que tenemos en la Tabla 8 corresponde al historial de una residencia comprada en 2004, con un permiso de remodelación en 2005 y luego vendida en 2008.

Año	Año de Venta (Y después)	CPI	Precio CPI	Permisos	Inversión	Precio Calculado (Y Antes)
2004		CPI_{2004}	$Y_{ref=2004}$	0	0	$Y_{antes=2004}$
2005		CPI_{2005}	$Y_{ref=2005}$	1	I_{2005}	$Y_{antes=2005}$
2006	$Y_{despues=2006}$	CPI_{2006}	$Y_{ref=2006}$	0	0	$Y_{antes=2006}$
2007		CPI_{2007}	$Y_{ref=2007}$	0	0	$Y_{antes=2007}$
2008	$Y_{despues=2008}$	CPI_{2008}	$Y_{ref=2008}$	0	0	$Y_{antes=2008}$

Tabla 8. Historial de ejemplo para una propiedad.

$$Precio\ CPI = \frac{Y_{ref=año} * CPI_{año}}{CPI_{año}} \quad (1)$$

$$Precio\ Calculado = (Y_{ref=año} * (1 + cambio\ porceltual\ del\ CPI\ referencial)) \quad (2)$$

Para calcular el precio de venta de cualquier año, podemos usar las fórmulas (1) y (2) para calcular las variables y llenar las tablas.

Para entender el trabajo de nuestro modelo, veamos la siguiente formula:

$$Y_{despues} = Y_{antes} + W_{permisos} * X_{permisos}$$

$$\left(Y_{despues} - Y_{antes}\right) = W_{permisos} * X_{permisos}$$

La expresión $Y_{despues} - Y_{antes}$ representa el retorno a la inversión. Nuestro modelo aprende mediante variables descritas en este trabajo e información como la de la tabla, sobre la matriz W la cual, al ser multiplicada por la información de los permisos, aproximará de la mejor manera el retorno a la inversión.

Análisis de los resultados

Una vez entrenado el algoritmo, la inserción de nuevos datos para calcular un resultado de ROI es el siguiente paso por seguir. De esta manera, tomaremos entonces nuevas descripciones de los proyectos de renovación a lo largo de las ciudades de prueba, y el algoritmo se encargará de generar las predicciones necesarias. Es de notar que los factores fundamentales que el estudio debe considerar son, tanto el costo de la nueva renovación como el impacto que ésta generará económicamente en la vivienda en cuestión.

Posteriormente, y con miras a comprobar que el error de cálculo del algoritmo se encuentre dentro de valores aceptables, procederemos a compararlo con valores reales de compra y venta en el mercado inmobiliario.

Para tener un punto de control en cuanto a las tendencias de ROI, se escogen 18 ciudades para el estudio, éstas son escogidas por su alto flujo demográfico y tendencias variadas de vida.

Imagen 11. Ciudades metro de estudio para cálculos de ROI de prueba.

Se procederá seguidamente a estudiar cada proyecto de renovación y su posible impacto en los diferentes mercados regionales. A continuación, ofrecemos los resultados obtenidos separados por categoría.

1. Proyectos de comedor:

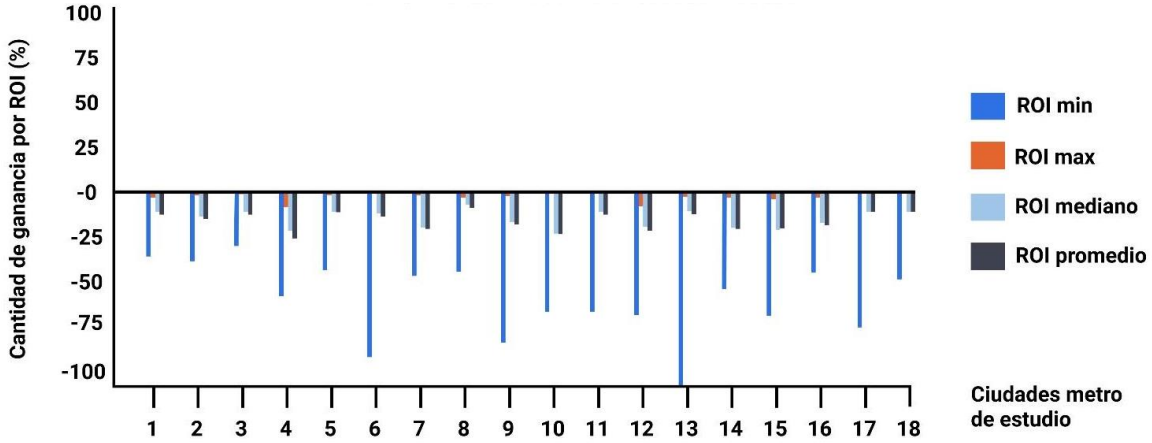

Imagen 12. Tabla de valores de ROI calculado para proyectos de comedor. Fuente: Elaboración propia.

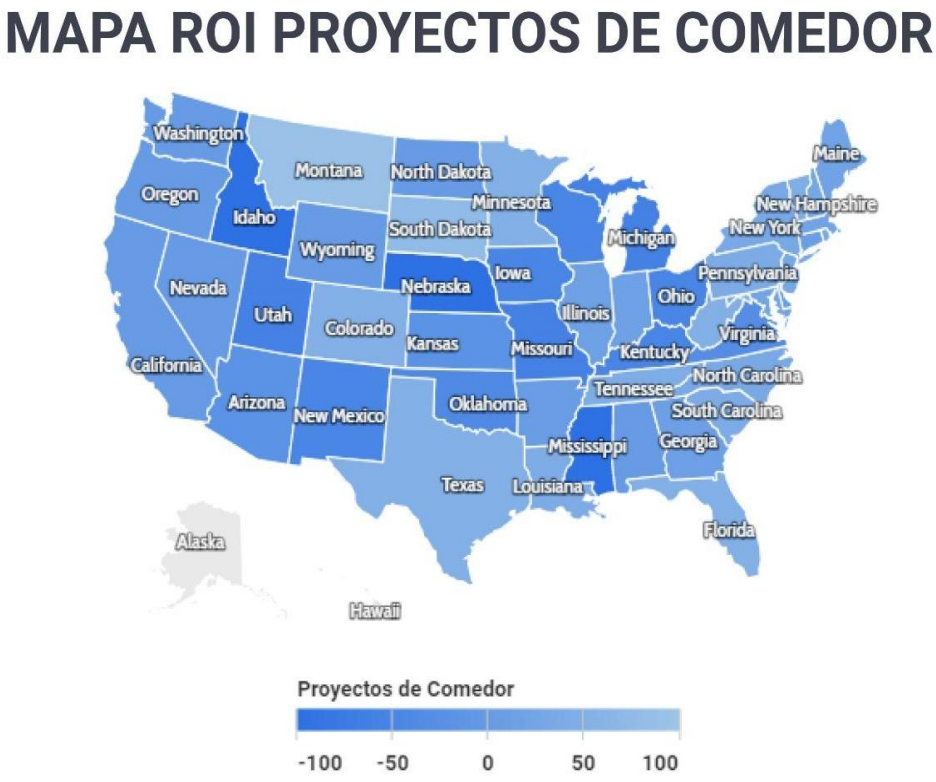

Imagen 13. Mapa de distribución de valores de ROI para proyectos de comedor. Fuente: Elaboración propia.

2. Proyectos de pasillos:

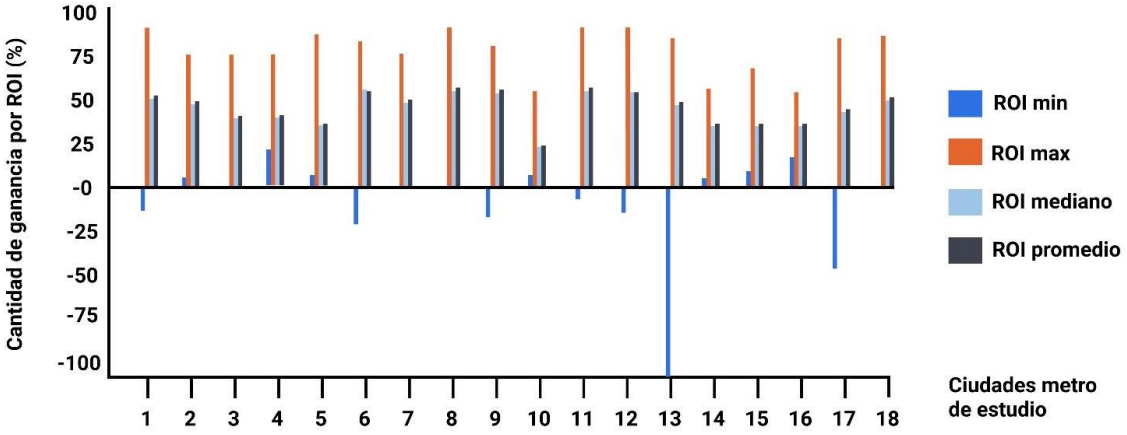

Imagen 14. Tabla de valores de ROI calculado para proyectos de pasillo. Fuente: Elaboración propia.

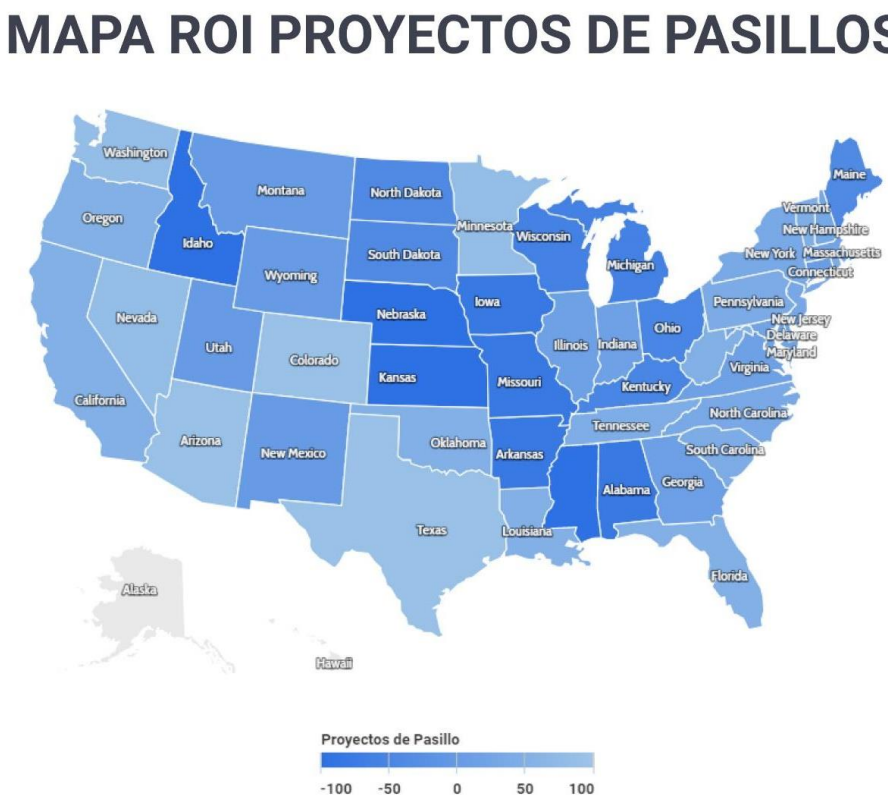

Imagen 15. Mapa de distribución de valores de ROI para proyectos de pasillo. Fuente: Elaboración propia.

3. Proyectos de garaje:

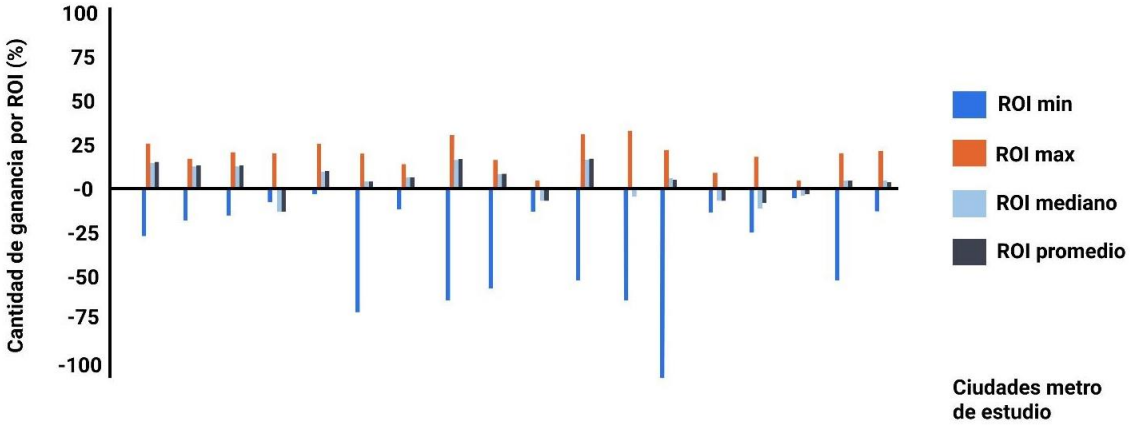

Imagen 16. Tabla de valores de ROI calculado para proyectos de garaje. Fuente: Elaboración propia.

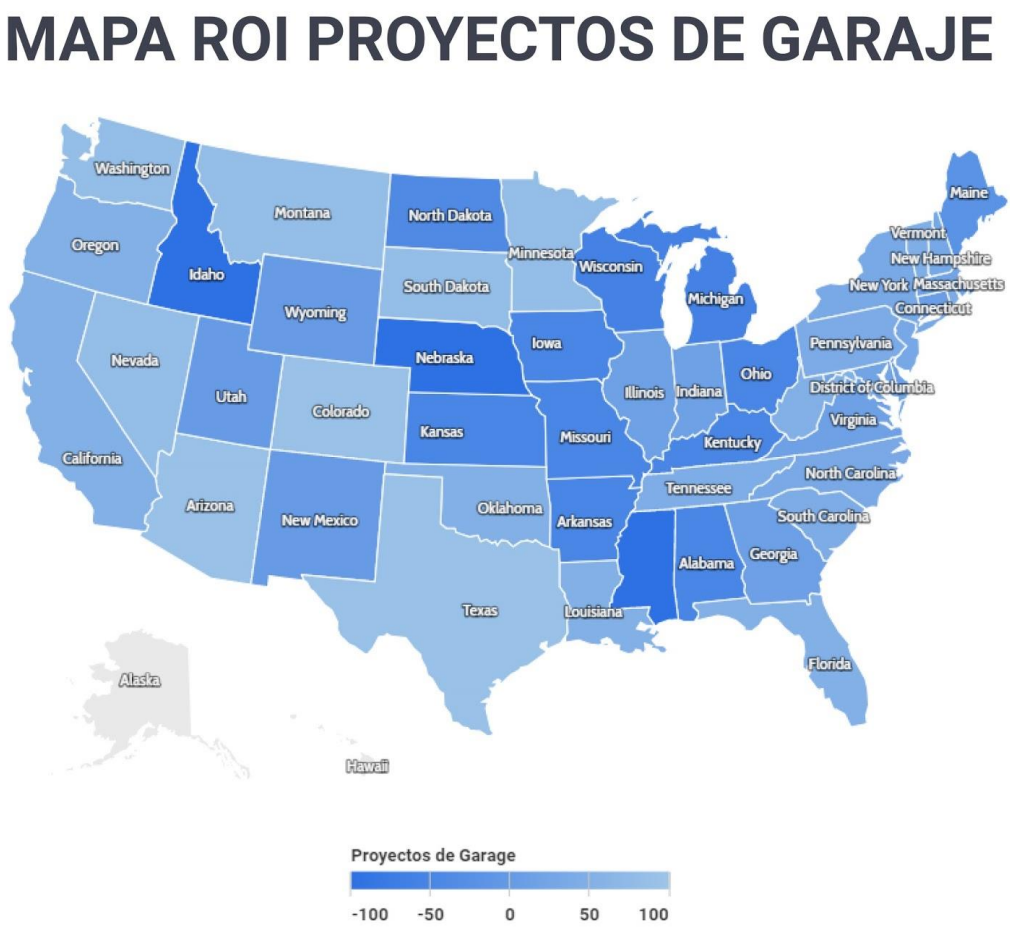

Imagen 17. Mapa de distribución de valores de ROI para proyectos de garaje. Fuente: Elaboración propia.

4. Proyectos de cocina:

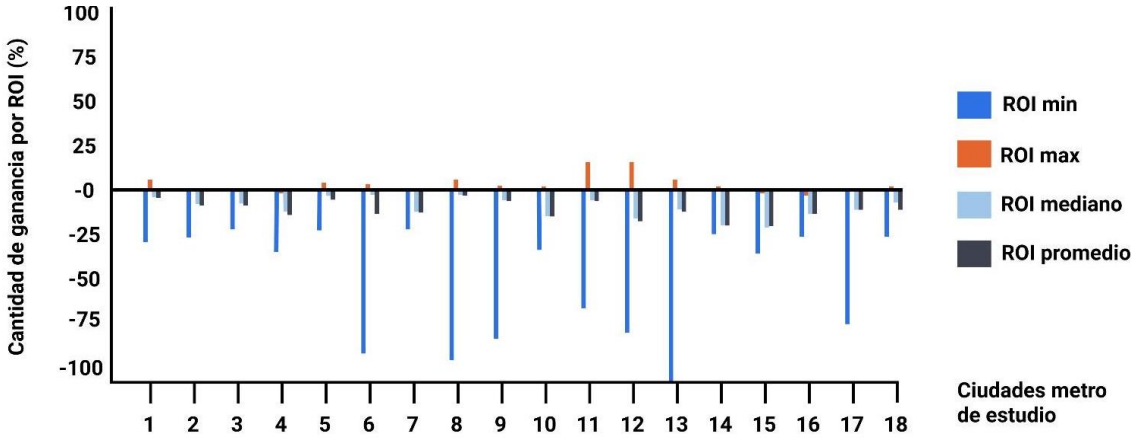

Imagen 18. Tabla de valores de ROI calculado para proyectos de cocina. Fuente: Elaboración propia.

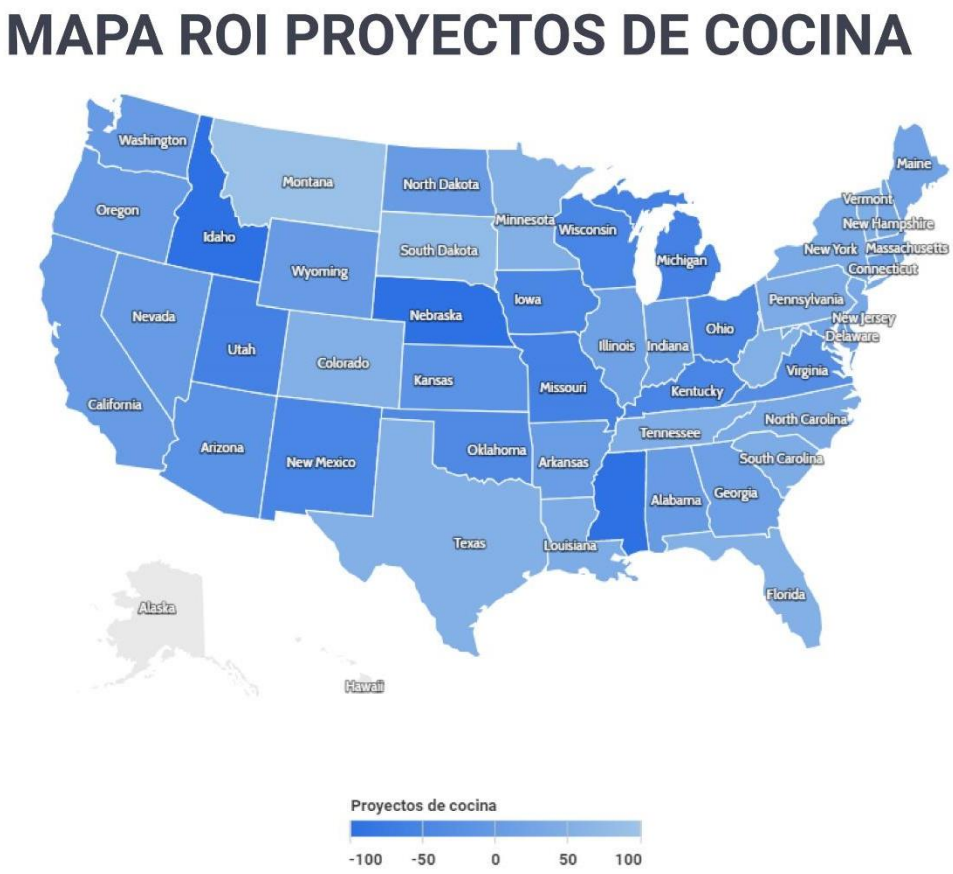

Imagen 19. Mapa de distribución de valores de ROI para proyectos de cocina. Fuente: Elaboración propia.

5. Proyectos de baño:

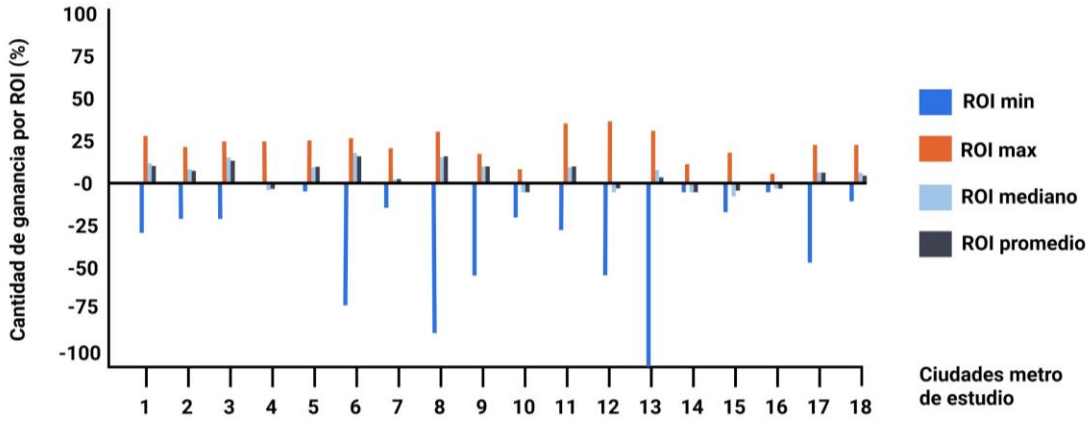

Imagen 20. Tabla de valores de ROI calculado para proyectos de baño. Fuente: Elaboración propia.

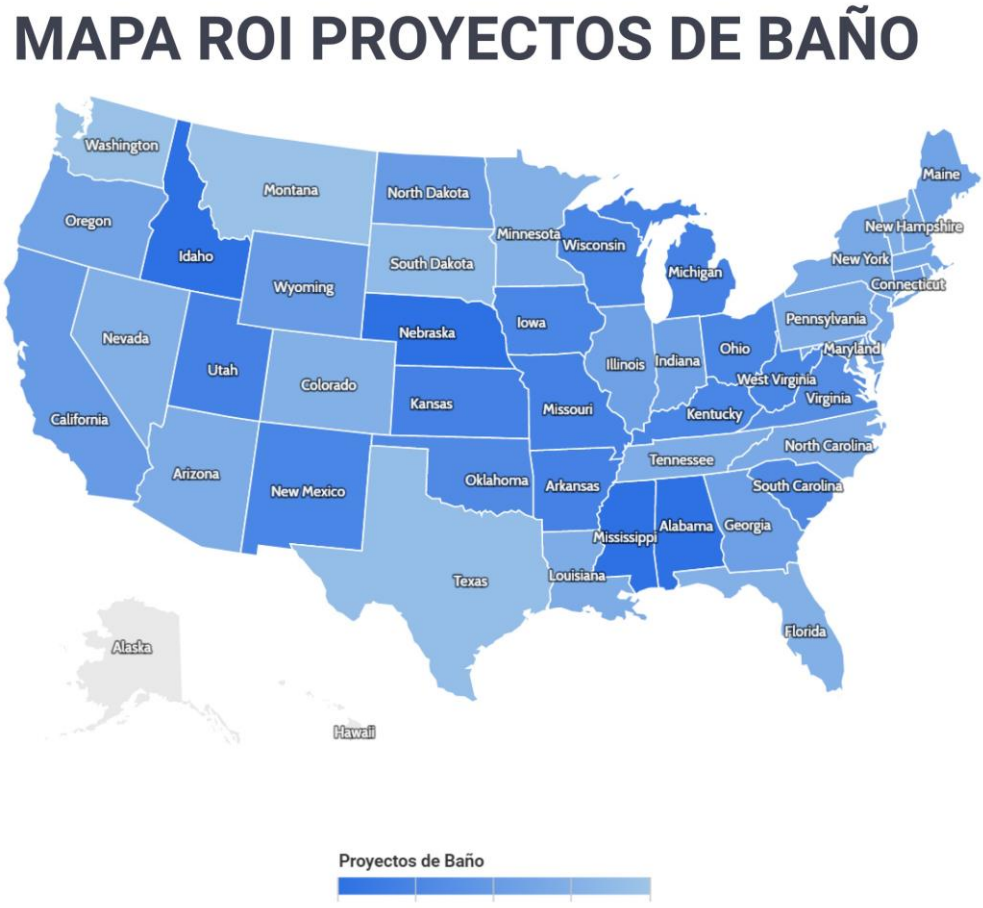

Imagen 21. Mapa de distribución de valores de ROI para proyectos de baño. Fuente: Elaboración propia.

6. Proyectos de terraza:

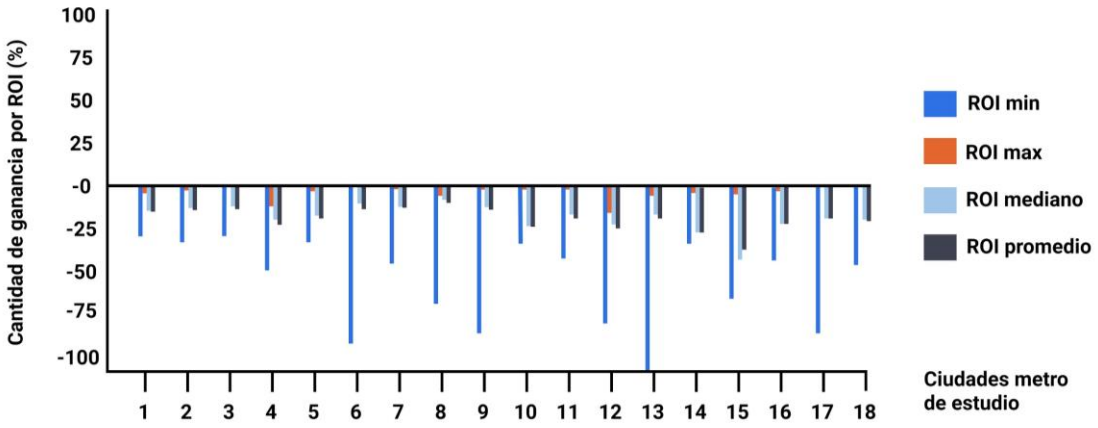

Imagen 22. Tabla de valores de ROI calculado para proyectos de terraza. Fuente: Elaboración propia.

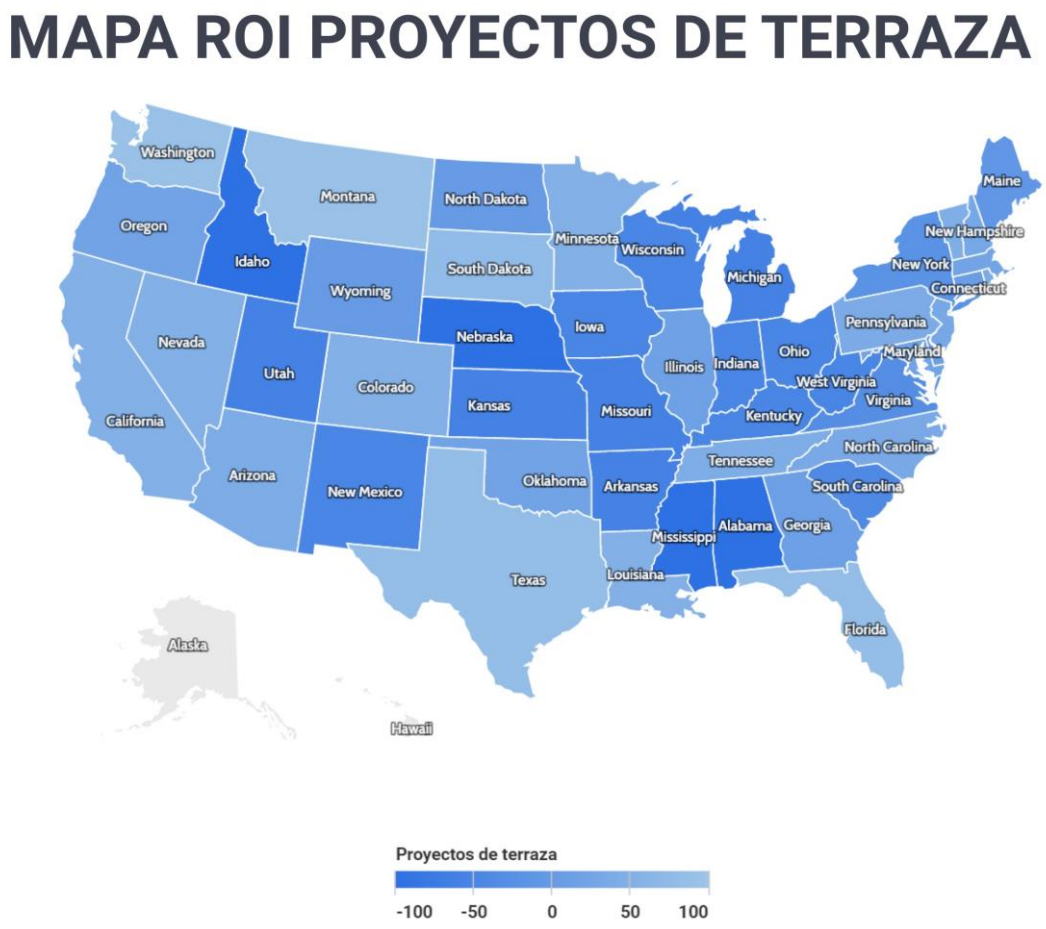

Imagen 23. Mapa de distribución de valores de ROI para proyectos de terraza. Fuente: Elaboración propia.

7. Proyectos de piscina:

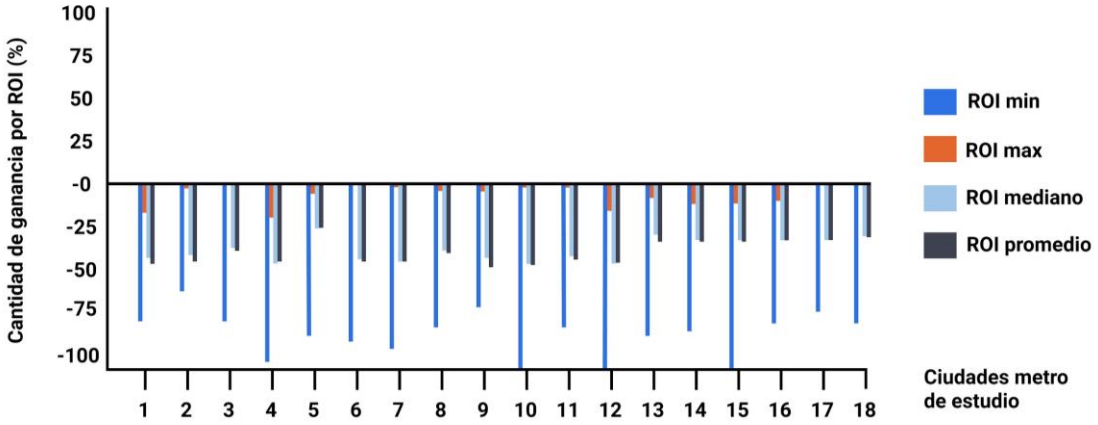

Imagen 24. Tabla de valores de ROI calculado para proyectos de piscina. Fuente: Elaboración propia.

.

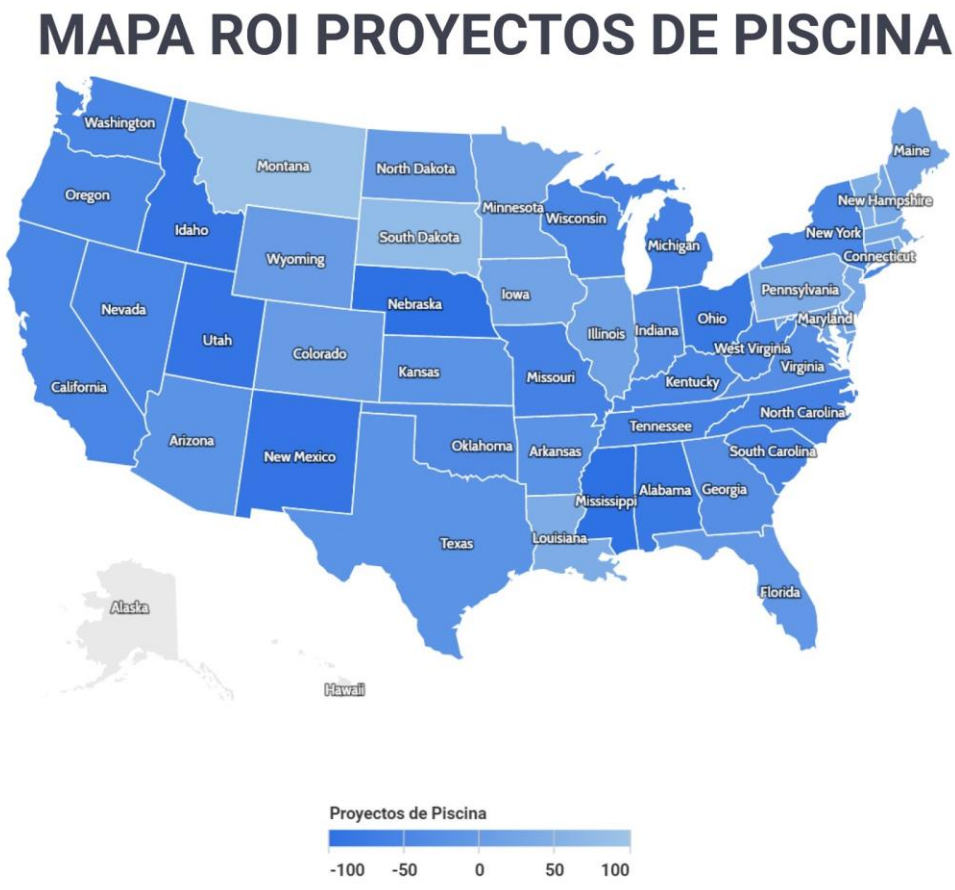

Imagen 25. Mapa de distribución de valores de ROI para proyectos de piscina. Fuente: Elaboración propia.

Se puede apreciar que, en promedio, el proyecto de renovación residencial menos rentable corresponde al de comedor, [*dining-room*], pues se observan valores de pérdidas entre el 100% y 25% del total invertido para todas las ciudades estudiadas.

Por otra parte, en promedio, el proyecto más rentable está ubicado en los pasillos de la residencia, [*Hallway*]. Obteniéndose valores de ganancias entre 40% y 80% de la inversión para todas las ciudades estudiadas. Resultado curioso ya que es una renovación sin gran potencial de cambio significativo.

Estudiando los patrones de distribución, podemos observar que los proyectos de renovación más realizados en Estados Unidos en el año 2017, en una muestra de alrededor de 580.000 permisos gestionados en el año y concentrándose solo en renovaciones internas residenciales, fueron los siguientes:

Proyecto	Costo Promedio	RoI Promedio	Frecuencia
Comedor	6400$	3570$	3.37%
Pasillo	3104$	4900$	0.60%
Garaje	13482$	12900$	5.55%
Cocina	12727$	9600$	12.76%
Baño	9500$	9430$	23.99%
Terraza	29480$	13100$	0.92%
Piscina	11700$	3710$	2.33%

Tabla 9. Distribución de costo y Retorno por proyecto. Fuente: Elaboración propia.

Al contar con una base de resultados para cada una de las localidades, pasamos luego a compararlos con el costo promedio y la frecuencia de aparición para tener una referencia transversal en todo el país.

6. Conclusiones y reflexiones finales

El proyecto que nos ocupa en su fase final debe centrarse en lograr un equilibrio entre una predicción estable del estimado de renovación y el cálculo del impacto de dicha renovación en la edificación, de manera que le permita al usuario acceder y gestionar datos antes desconocidos.

Comparando la cantidad de proyectos realizados con respecto a la posibilidad de que genere ganancias a futuro llegamos a las siguientes conclusiones:

- Si bien el proyecto más realizado en el período de tiempo determinado fue el de renovación de garaje, no se corresponde con la opción que genera más retorno de inversión.

- Aun cuando el proyecto de cocina tiene el segundo lugar, es uno de los que peor retorno genera. Esta correspondencia es interesante, pues hace ver que es necesario educar a la población con respecto a qué proyectos de renovación son más factibles a fin de generar un mayor retorno de inversión.

- Los proyectos de comedor son los únicos que se comportan acorde a la relación costo/beneficio. Es unos de los menos populares, pero se corresponde con que es uno de los menos rentables.

- Los proyectos de renovación de pasillos son los que más retorno de inversión generan. Sin embargo, son los menos realizados.

Análisis FODA

Fortalezas.	**Debilidades.**
-Responde a una necesidad real y tangible de todos los propietarios de bienes raíces.	-Está restringido al mercado estadounidense.
Es una manera novedosa de utilizar tecnologías de vanguardia para resolver un problema. De manera que no cuenta con competencia directa.	-No responde bien a la volatilidad del mercado, sobre todo en un mercado como el americano que crea burbujas financieras que pueden explotar en cualquier momento.
-El algoritmo es sólido y versátil ya que su implementación y retroalimentación son procesos sencillos.	-El cálculo depende de valores extrapolados de data que proviene de muchas fuentes distintas.
-Es de fácil Escalabilidad. No importa la zona geográfica que se esté estudiando el algoritmo se comporta igual. Solo tiene que consumir la data específica de la región.	-Existe la posibilidad de errores humanos en todo el proceso de recolección y acondicionamiento.
-Es una información de fácil digestión y fue validado por las empresas especializadas.	-La rapidez de movilidad residencial y cambios en las ciudades hacen que se requieran actualizaciones constantes.

Tabla 10. Fortalezas y debilidades del proyecto. Fuente: Elaboración propia.

Oportunidades.	**Amenazas.**
-Posibilidades de implementación del algoritmo y software en otros países, considerando factores propios según la situación y el entorno.	-Las nuevas tendencias en programación pueden aumentar la competencia a medida que el uso de estas herramientas se haga más accesible a usuarios no especializados.
-Ampliar el número de ciudades con los que se cuentan datos, a modo de mejorar el cálculo y disminuir los márgenes de errores.	-Cambios en la política del país podrían limitar la publicación de datos importantes ya que todos los datos provienen de entidades gubernamentales.
-Buscar distintas alternativas para conseguir la información necesaria para alimentar el algoritmo.	-El estado cambiante en el mercado de bienes raíces puede generar cambios en el cálculo.

Tabla 11. Oportunidades y amenazas. Fuente: Elaboración propia.

8. Bibliografía

Arias, F. (1999). *El Proyecto de Investigación. Guía para su elaboración.* Caracas, Venezuela: Editorial Episteme.

Casella, M. (2017). *Historia y evolución de la Inteligencia Artificial.* Editorial Marco Casella 2015.

Chen, J. (22 de 2 de 2019). *Investopedia.* Obtenido de https://www.investopedia.com/terms/r/returnoninvestment.asp

Chen, M.-S. (1996). Data mining: an overview from a database perspective. *IEEE Transactions on Knowledge and Data Engineering*, 866 - 883.

Cherry, K. (3 de 4 de 2019). Obtenido de Very Well Mind: https://www.verywellmind.com/what-is-longitudinal-research-2795335

Department of Economic and Social Affair. (2019). *World Urbanization Prospects The 2018 Revision .* New York: United Nations.

Elgendy, N., & Elragal, A. (2014). Big Data Analytics: A Literature Review Paper. *Springer International Publishing*.

Gnoza, N., & Barberena, M. (2018). *Estudio de factibilidad del uso de machine Learnig con múltiples fuentes de datos en el pronóstico el tiempo.* Monte Video: Universidad de Uruguay.

Gordon, R. J. (2000). Does the new economy measure up to the great inventions of the past. *National Bureau of Economic Research*, 49-74.

Hernando, J. (2016). *Análisis e inversión en el mercado inmobiliario desde una perspectiva conductual.* Catalunya: Universidad de Catalunya.

Howe, J. (1 de 6 de 2006). *The Rise of Crowdsourcing.* Obtenido de Wired: https://www.wired.com/2006/06/crowds/

Jain, N., & Srivastava, V. (2013). DATA MINING TECHNIQUES: A SURVEY PAPER. *International Journal of Research in Engineering and Technology*.

Laurie Goodman, A. M. (2019). *Housing Finance At A Glance: A Monthly Chartbook, April 2019.* Urban Institute.

Marquès, P. (2000). Las TIC y sus Aportaciones a la Sociedad. *Researchgate*.

Miraz, D., & Ali, M. (2015). A review on Internet of Things (IoT), Internet of Everything (IoE) and Internet of Nano Things (IoNT). *Researchgate*.

Pérez, A. (2009). Educación y otras ciencias.

Rivas, R. (2019). The impact of colleges and hospitals to local real estate markets. *Journal of Big Data*.

Sargatal, A. (2000). El estudio de la Gentrificación. *Biblio 3W. Revista Bibliográfica de Geografía y Ciencias Sociales*.

Scheele, C. H. (1970). A Short History of the Mail Service. *Smithsonian Institution Press*, 174-175.

SEC. (10 de 2019). *What We Do*. Obtenido de sec.gov: https://www.sec.gov/Article/whatwedo.html

Sharma, S. (6 de Septiembre de 2017). *Activation Functions in Neural Networks*. Obtenido de Towards Data Science: https://towardsdatascience.com/activation-functions-neural-networks-1cbd9f8d91d6

World Development Report. (2016). *Enabling Digital Development. Six Digital Technologies to Watch.*

9. Anexos

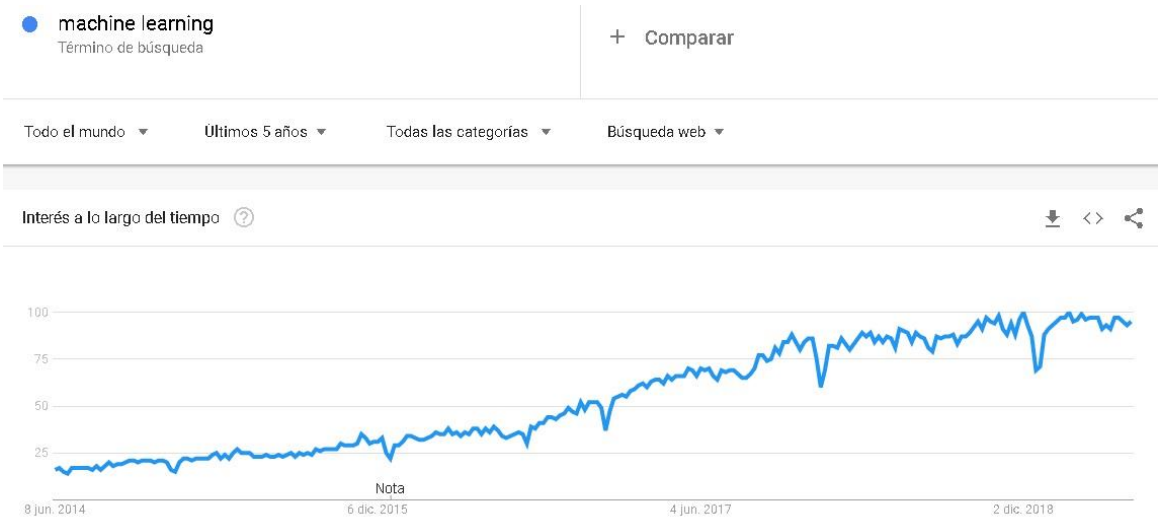

Imagen 26. Gráfico del crecimiento del interés por las tecnologías de aprendizaje automático.

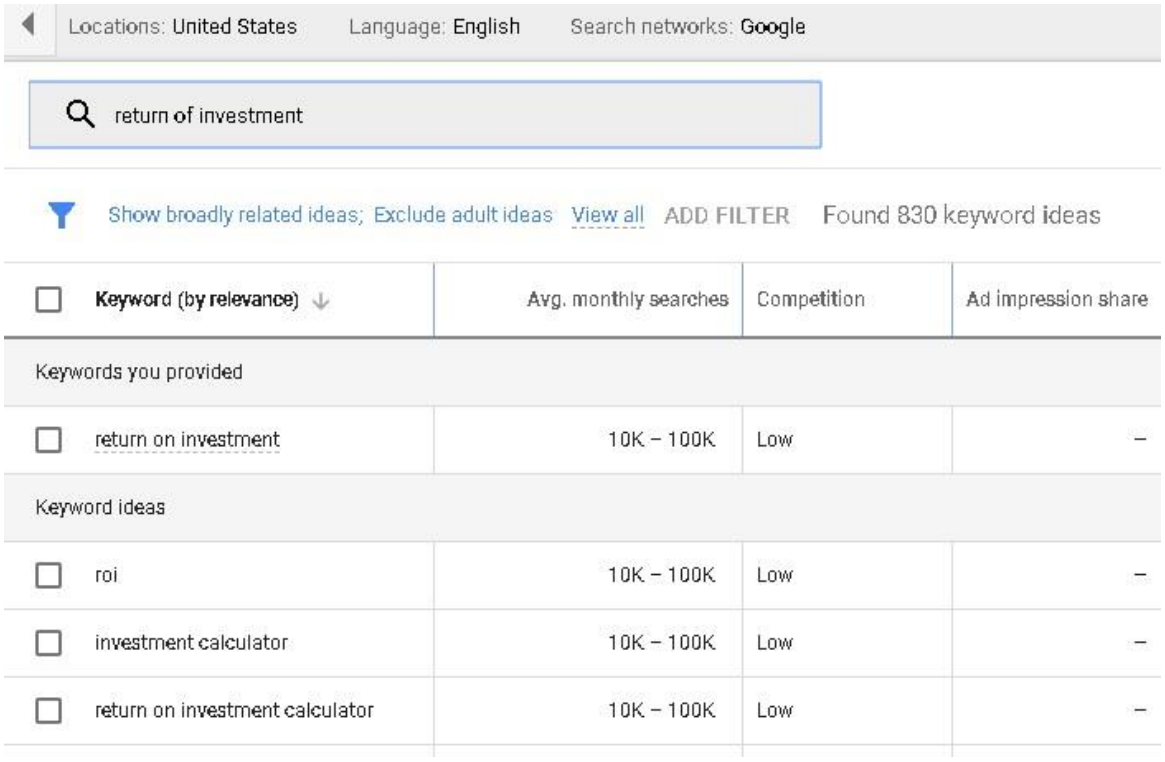

Imagen 27. Cantidad de búsquedas en internet, sobre retorno de inversión.