UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio Ciencias e Ingenierías

Análisis y predicción de las tendencias de venta en el mercado usando árboles de regresión

Proyecto de investigación

Cristopher Andre Palacios Utreras

Ingeniería en Sistemas

Trabajo de titulación presentado como requisito para
la obtención del título de
Ingeniero en Sistemas

Quito, 07 de mayo de 2020

UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio Ciencias e Ingenierías

HOJA DE CALIFICACIÓN
DE TRABAJO DE TITULACIÓN

Análisis y predicción de las tendencias de venta en el mercado usando árboles de regresión

Cristopher Andre Palacios Utreras

Calificación:

Nombre del profesor, Título académico:

Noel Pérez, Ph.D.

Firma del profesor:

_____

Quito, 07 de mayo de 2020

DERECHOS DE AUTOR

Por medio del presente documento certifico que he leído todas las Políticas y Manuales de la Universidad San Francisco de Quito USFQ, incluyendo la Política de Propiedad Intelectual USFQ, y estoy de acuerdo con su contenido, por lo que los derechos de propiedad intelectual del presente trabajo quedan sujetos a lo dispuesto en esas Políticas.

Asimismo, autorizo a la USFQ para que realice la digitalización y publicación de este trabajo en el repositorio virtual, de conformidad a lo dispuesto en el Art. 144 de la Ley Orgánica de Educación Superior.

Firma del estudiante: _____

Nombres y apellidos: Cristopher Andre Palacios Utreras

Código: 00128585

Cédula de Identidad: 1717669822

Lugar y fecha: Quito, 07 de mayo de 2020

ACLARACIÓN PARA PUBLICACIÓN

Nota: El presente trabajo, en su totalidad o cualquiera de sus partes, no debe ser considerado como una publicación, incluso a pesar de estar disponible sin restricciones a través de un repositorio institucional. Esta declaración se alinea con las prácticas y recomendaciones presentadas por el Committee on Publication Ethics COPE descritas por Barbour et al. (2017) Discussion document on best practice for issues around theses publishing, disponible en http://bit.ly/COPETheses.

UNPUBLISHED DOCUMENT

Note: The following capstone project is available through Universidad San Francisco de Quito USFQ institutional repository. Nonetheless, this project – in whole or in part – should not be considered a publication. This statement follows the recommendations presented by the Committee on Publication Ethics COPE described by Barbour et al. (2017) Discussion document on best practice for issues around theses publishing available on http://bit.ly/COPETheses.

RESUMEN

Los recientes avances de las técnicas de inteligencia artificial cada vez se encuentran más relacionadas al aprendizaje profundo, los cuales a su vez son más sofisticados y costosos con respecto al costo computacional y a la infraestructura. Estas técnicas han demostrado ser exitosas en diversas áreas como la médica, industria alimentaria, bancos, entre otras, ayudando en tareas de procesamiento de datos y predicciones relacionadas a estos, lo cual genera una ventaja competitiva. Por ello nace la necesidad de que la inteligencia artificial se encuentre a disponibilidad de todos los sectores empresariales, en especial de las empresas pequeñas o pymes. Para muchas de las pymes su principal fuente de ingreso son las ventas. Por ende, conocer el comportamiento de la cartera de clientes con respecto a la variabilidad de los precios es un asunto primordial para el éxito. Desafortunadamente, todo negocio a sus inicios no cuenta con el capital suficiente para cubrir el apoyo empresarial basado en técnicas de inteligencia artificial. Generalmente, las inversiones son realizadas dependiendo de necesidades básicas del negocio. Por tanto, en el presente trabajo se desarrolla una estrategia de predicción de bajo costo basada en el uso de modelos de árboles de regresión. En particular, se pretende realizar un análisis de desempeño usando la métrica de precisión de los modelos implementados en tres bases de datos experimentales relacionadas a las ventas de bienes raíces y comerciales, respectivamente. Los mejores resultados fueron obtenidos por los modelos CART, los cuales alcanzaron valores de precisión entre el 80 y 94%. Estos resultados brindan evidencia experimental de un buen desempeño de predicción en las distintas bases de datos, mediante el uso de modelos más simples que los basados en el aprendizaje profundo. Por tanto, estos constituyen alternativas factibles y atractivas para el apoyo en la administración de cualquier negocio, especialmente empresas pequeñas.

Palabras claves: fluctuaciones, regresión, árboles de decisiones, tendencias de mercado, predicción

ABSTRACT

The latest advances in artificial intelligence techniques are increasingly related to deep learning, which are more sophisticated and expensive in terms of computational and structural aspects. These techniques have been successful in various areas such as medical, food, industry, banks, and others, helping in data processing and related predictions, which creates a competitive advantage against other industries. Therefore, all business sectors need artificial intelligence to be available, especially the small companies or SMEs. Many of the SMEs have their main source of income in sales. For that reason, is essential for them to predict the behavior of customers with respect to the price variability. Unfortunately, every business in its infancy does not have enough capital to cover business support based on artificial intelligence techniques. Necessarily, investments are made specially for the basic needs of the business. Thus, this research presents a low-cost prediction strategy based on regression trees models. We seek to obtain a performance analysis using the precision metric of the models implemented in three experimental databases related to the field of house sales and commercial sales. The best results were obtained by CART models, which reached an accuracy between 80 to 94%. These results provide experimental evidence of good prediction performance in different databases, using simpler models than those specified in deep learning. Consequently, this feasible and attractive methods can support as administration tool of any business, especially for small companies.

Key words: fluctuations, regression, decision trees, market trends, prediction

TABLA DE CONTENIDO

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

GLOSARIO

ANN: artificial neural network
AI: artificial intelligence
ML: machine learning
SVM: support vector machine
DT: decision tree
RT: regression tree
SVR: support vector regressor
RFR: random forest regressor
CART: classification and regression trees
MSE: mean squared error
MAE: mean absolute error
CBFDT: case base fuzzy decision tree
CRISP-DM: cross industry standard process for data mining
LR: logistic regression
GBDT: gradient boost decision tree
CV: cross validation
MLP: multilayer perceptron
ACC: accuracy o score

CAPÍTULO 1: INTRODUCCIÓN

El día a día en el mundo empresarial gira entorno a factores de incertidumbre y riesgos, para los cuales las empresas deben buscar la mejor estrategia financiera. Por ello, es inherente que una empresa está destinada a generar pérdidas, si no mantiene una política correcta de gestión y mitigación de riesgos. Los principales riesgos a afrontar quedaron evidenciados en la crisis de los años desde 1994 a 2002, donde se determinó la estrecha relación que existe en el desarrollo económico de los mercados con el proceso de globalización, la volatilidad de precios y la incertidumbre de los mercados (San-Martín-Albizuri & Rodríguez-Castellanos, 2011). Ahora bien, aunque el desarrollo de las crisis cuenta con particularidades únicas, guarda también ciertas similitudes que permite elaborar un patrón de acontecimientos similares; es por eso que las organizaciones exitosas están preparadas para saber que podría pasar y minimizar el riesgo en base a predicciones futuras (Nooteboom, 2019).

1.1. Antecedentes

1.1.1. Análisis económico-financiero.

La disponibilidad de la información financiera conjuntamente de un análisis económico financiero define el éxito de la empresa y a su vez tiene la capacidad de dar información para la toma de decisiones relacionadas al futuro de la empresa. Para tal fin, se requiere de una aplicación metodológica que permita la interpretación de los valores, de otra forma sería imposible la formulación de un juicio sobre los datos analizados (Fernández & ernández, 1986).

Una de las formas tradicionales del análisis financiero está relacionada a la interpretación del flujo de valores correspondiente a la actividad desempeñada, el cual se refiere a la comparación de las cifras previstas contra las reales y cuya conclusión determina el estado actual de la empresa.

Esta última tiene como objetivo: analizar si los recursos de la empresa son los más adecuados parar llevar un desarrollo estable y si el alcance de la empresa permite hacer frente a las obligaciones. En general se busca obtener un equilibrio económico en conjunto con los componentes que conforman el sistema de la empresa (Hernández, 1986).

1.1.2. Econometría.

La ciencia económica se refiere a la administración de recursos escasos que están a disposición de las colectividades humanas, debido a esto se evidencia la necesidad de los economistas de orientar sus investigaciones a mejorar el conocimiento de la realidad económica. Como toda ciencia, la economía rige sus primeras bases de diagnóstico en conclusiones descriptivas haciendo que los investigadores tengan un análisis limitado (Garayar P., 1953). Esta tarea comienza a ser más fecunda cuando se alimenta de otros campos como son la estadística y las matemáticas. La importancia del análisis mediante métricas numéricas radica en la generación de relaciones entre magnitudes y la interpretación de su comportamiento mediante símbolos matemáticos. De tal forma, la econometría se define como un neologismo, la conexión de dos palabras griegas "oikonomía" (administración, o economía) y "metron" (medida)" cuyo objetivo es la prueba y desarrollo de un modelo basado en la naturaleza numérica (Tintner, 1953).

1.1.3. Limitaciones de la econometría en el análisis financiero.

El conjunto de herramientas financieras y económicas, aunque irremplazables, resultan un poco ineficientes en cuanto al campo de toma de decisiones. Según la experiencia de (Fernández & ernández, 1986) se ha demostrado que las situaciones del pasado no se repiten, de hecho existe una tendencia actual con respecto a la “globalización temprana” cuya estabilidad y cambios del

mercado se ven afectados por el protagonismo de China en la economía mundial (Bonialian, 2018). Por lo cual, el ritmo de los cambios sobre los productos, mercados o servicios ha aumentado notablemente. Por otro lado, la evaluación de los cambios producidos en el valor de distintos elementos pertenecientes a los estados financieros de una empresa, llegan a ser ineficaces si la cantidad de documentos almacenados son elevados, ya que se dificulta la obtención de las relaciones pertinentes entre distintas variables para la obtención de un modelo.

En resumen, las limitaciones de análisis más comunes se relacionan con la selección de variables más importantes, la determinación de las relaciones de sustitución de los datos, utilización de datos históricos para predicción, entre otras. Como resultado, la información entregada por los estados financieros puede no ser acorde con la realidad económica e inducir un error dentro del desarrollo de los modelos económicos, ya que se puede distorsionar la normalidad de los datos. Por otro lado, la dispersión de datos en relación a solo una tendencia central resulta insuficiente para el diagnóstico financiero y creación de métodos flexibles que permitan adaptación a los objetivos divergentes y complejos que presentan las pequeñas, medianas y grandes empresas (Fernández & ernández, 1986).

1.1.4. Machine learning.

Actualmente las ciencias, se encuentran afirmadas por el desarrollo tecnológico en cuanto al cálculo de procesos, herramientas, almacenamiento y sistemas de información, enfocados al incremento de la productividad en el área de desarrollo. En particular, el campo de la inteligencia artificial (AI) ha demostrado la capacidad de impulsar las ciencias, a tal punto de desplazar y enfocar las ocupaciones humanas simplemente en el control. En el campo económico, AI ha

contribuido con varios algoritmos consistentes para analizar las características del mercado y la predicción de costos a través de datos previamente recopilados (Agrawal et al., 2019).

El machine learning (ML) nace como una disciplina científica y parte de AI cuyo objetivo se enfoca en la creación de sistemas que aprendan automáticamente, por medio de la identificación de patrones complejos en base a datos históricos y que sean capaces de predecir comportamientos futuros. Algunos de los modelos implementados exitosamente en el campo de las tendencias de mercado son los modelos complejos como: redes neuronales artificiales (ANN), support vector machine (SVM), modelos mixtos y los modelos simples como: regresiones simples y árboles de decisiones (Menon et al., 2019).

1.1.5. Árboles de decisiones.

Los investigadores de los árboles de decisiones (DTs) trazan los orígenes en los trabajos desarrollados por Hunt y otros alrededor de 1950, Sonquist y Morgan en 1964, los investigadores de Sonquist en el programa “Detección de interacciones Automaticas” en 1971, entre otros (Bouzida & Cuppens, 2006). Con los cuales los investigadores Breiman, Friedaman, Olshen y Stone, introdujeron nuevas métricas al algoritmo de construcción de DTs con el fin de simplificar problemas de regresión y clasificación.

En base a los conceptos desarrollados por los investigadores preliminares, se infiere la determinación de DTs como una técnica enmarcada dentro de los sistemas de razonamiento utilizados en las investigaciones de ML. Una de las principales razones de su aplicación es su estructura, ya que son fáciles de comprender y analizar, su uso más frecuente varía entre diagnósticos médicos, predicciones meteorológicas, controles económicos y demás problemas que requieran de análisis de datos y toma de decisiones.

Los tres principales componentes de un DT son: nodos, ramas y hojas. Cada nodo es etiquetado con el nombre del atributo más significativo, entre los atributos que todavía no han sido seleccionados antes en algún otro nodo raíz. Cada rama es nombrada con un valor de característico para el nodo atributo y cada hoja es etiquetada con una categoría o clase (Bouzida & Cuppens, 2006).

La forma en la que operan los DTs se basa en la aplicación de un conjunto de reglas SI-ENTONCES, haciendo uso de la lógica matemática para determinar distintas disyunciones que nos llevaran a posibles resultados. Su forma de operar es otra de las principales razones de uso, ya que nos permite tener un planteamiento general del problema y llevarlo hacia casos más específicos, de tal forma que todos los casos sean analizados, obteniendo así todas las consecuencias de las posibles decisiones (Zuniga & Abgar, 2011).

La diversificación de valores en categóricos y continuos, no es un impedimento para la predicción cuando se usa DTs, de hecho, el proceso de generación de DTs para la predicción es similar al proceso de generación de DTs para la clasificación variando simplemente las reglas de división de la siguiente manera:

- Para árboles de clasificación, el criterio de división se basa principalmente en el índice de impureza o entropía de cada nodo, los cuales refieren a la probabilidad de no sacar elementos pertenecientes a la misma clase y a la incertidumbre que existe sobre un conjunto de datos describiendo así su grado de desorganización (Timofeev, 2004).
- Para árboles de regresión, el criterio de división se establece en las medidas de error como mínimos cuadrados y mínimas desviaciones absolutas, la primera se refiere a la minimización de la suma del error cuadrado entre las observaciones y el valor de la

media en cada nodo, la segunda se refiere a la minimización de desviación absoluta media de la mediana dentro de un nodo (Moisen, 2008).

Finalmente, debido a su simplicidad y adaptabilidad los DTs dentro de las ciencias económicas se enfocan en fines predictivos, análisis de riesgos, decisiones de inversión o decisiones de gestión financiera, convirtiéndolo en un objeto atractivo de investigación.

1.2. Motivación y objetivos

A pesar de que muchas técnicas de ML han sido aplicadas exitosamente en el campo de la predicción, específicamente de las tendencias de mercado, muchas de ellas vienen a ser técnicas complejas, como por ejemplo ANNs. La dificultad de estos modelos radica en la variabilidad de sus hiperparámetros, ya que la búsqueda de hiperparámetros óptimos se traduce en un problema combinatorio. Además, la cantidad de valores para alimentar al ANN tiende a que mientras más grande sea, más preciso será el modelo y por otro lado, el modelo demanda mayor cantidad de recursos computacionales (Menon et al., 2019).

Con el objetivo de evitar la búsqueda excesiva de hiperparámetros o confiar en un costoso problema de optimización, esta investigación se centra en integrar diferentes técnicas de aprendizaje automático, con hiperparámetros estándar y bajo valor computacional para alcanzar buenas predicciones de valores de venta en el mercado.

Para cumplir con este objetivo, se ha concentrado la atención en el entrenamiento de árboles de regresión (RT) basados en la predicción de valores de venta en el mercado, los cuales son una solución atractiva y fácil de interpretar, ya que son ejecutados desde una hipótesis global hasta la formulación de sus resultados, generalmente en un corto periodo de tiempo (Esmeir & Markovitch, 2007). Durante el camino de obtención de los mejores modelos de RTs y sus mejores

hiperparámetros, este estudio busca explorar a fondo las técnicas de análisis de datos, la comprensión del algoritmo CART y sus métricas de expansión, también demostrar la eficiencia de los modelos de RTs mediante un enfoque comparativo con uno de los modelos complejos definidos en el estado del arte.

1.3. Tareas de investigación

Las tareas de investigación están encaminadas a:

- El estudio del estado del arte de la predicción basadas en modelos de regresión
- El estudio del estado del arte de los árboles de regresión en las tendencias del mercado.
- La determinación de los modelos de *árboles de regresión* para la predicción de valores de venta.
- La aplicación y experimentación de los mejores modelos de *árboles de regresión* en las bases de datos seleccionadas.

CAPÍTULO 2: ESTADO DEL ARTE

2.1. Estado del arte de técnicas de análisis de datos

Hoy en día el manejo de datos almacenados ha rebasado las capacidades de los seres humanos para administrar datos manualmente debido a su gran magnitud. Las herramientas computacionales de análisis de datos nos han permitido obtener conocimiento e interpretar los datos por medio de la identificación de patrones, descripciones estadísticas, técnicas de representación y visualización de datos. Algunas de las técnicas de análisis de datos que están enfocadas en la organización y estructuración de los datos se clasifican según su usabilidad de la siguiente manera (Zuur et al., 2010).

- Descripción estadística: contempla algunas de las técnicas estadísticas, las cuales utilizan a la matemática como su eje fundamental. Estas técnicas buscan detallar promedios, sumas, distribuciones de los datos, tasa de cambio, errores, etc.
- Identificación de valores atípicos: se define como valores atípicos a los valores que son muy grandes o pequeños comparados con la mayoría de las observaciones. Usualmente la técnica de visualización que es usada para la identificación de valores atípicos se denomina diagrama de caja, la cual representa la media como una línea en la mitad de la caja y a los valores atípicos como puntos fuera de la caja.
- Descripción de normalidad: son las técnicas que se usan para describir la distribución de un conjunto de datos. Por ejemplo, los histogramas, los cuales detallan las frecuencias de cada uno de los datos y se comparan con la curva normal.

- Identificación de relaciones: son técnicas que permiten identificar la colinealidad de variables, es decir demostrar la existencia de correlación entre covariantes, y también la independencia de las variables, es decir las variables que no se relacionan. Estos problemas pueden ser identificados mediante el coeficiente de Pearson, el cual mide el grado de relación entre dos variables. Y pueden ser graficados mediante una matriz.

2.2. Estado del arte de modelos de regresión

La definición de regresión fue introducida por Francis Galton en el año de 1886, en un trabajo relacionado a la descripción de los rasgos físicos de los descendientes. El cual expresa el análisis de las técnicas de regresión y define que los modelos de regresión son una técnica estadística usada para estudiar la relación entre variables (Pereira González, 2010). Dicho análisis puede utilizarse para explorar y cuantificar la relación entre variables dependientes e independientes de entrada o salida correspondiente a un sistema lineal o no lineal, con el objetivo de predecir la salida del sistema cuando se proporciona un nuevo dato de entrada. Algunas de las técnicas usadas para la predicción son:

- *Regresión lineal múltiple*: esta regresión nos permite establecer la relación lineal perteneciente a una variable dependiente y a un conjunto de variables independientes, la cual está definida matemáticamente por el modelo determinístico

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + e \qquad (1).$$

Donde Y es la variable para predecir dado un conjunto de variables predictoras o de atributos $X_0 + X_1 + \cdots + X_k$, "e" es el error entre el valor observado y el valor ajustado, $\beta_0 + \beta_1 + \beta_2 + \cdots + \beta_k$ son los pesos estimados que mejor definen la

pendiente lineal del conjunto de datos (Mouhaffel et al., 2017). Este tipo de regresiones son utilizadas para predicciones en varios campos económicos, por ejemplo, la predicción de las ventas totales anuales de un vendedor debido a otras variables como la formación y años de experiencia.

- *Regresión logística:* esta regresión tiene la capacidad de dar a conocer como inciden las variables independientes sobre una dependiente al igual que otras regresiones. Pero, esta cuenta con la particularidad de permitir variables dependientes categóricas o variables independientes categóricas y continuas. Además, no requiere de supuestos como normalidad y homocedasticidad. Esta regresión está determinada por la distribución binomial, ya que ofrece dos estados posibles para la variable de salida. Este proceso esta caracterizado por la probabilidad de ser cierto o falso y se presenta matemáticamente de la siguiente forma:

$$Y = \log\left(\frac{P}{1-P}\right) \qquad (2).$$

Donde P es la probabilidad de observar el resultado a predecir y 1-P es la probabilidad de no observarlo, siendo $P = \frac{n}{T}$ donde *n* es la cantidad de muestras en el atributo y *T* el total de muestras (Alderete, 2006). Estas regresiones son útiles en el campo de la predicción económica donde las etiquetas de salida son binarias. Por ejemplo, en la relación de precio y atributos del producto para determinar si se venderá o no según datos históricos.

- *Suport vector regressor (SVR):* es una técnica de regresión que parte del concepto de estimación de funciones multidimensionales, al igual que el SVM, con la diferencia de que el conjunto de salida es representado por un número continuo,

lo cual permite predecir. Para lograr este fin, la regresión establece un margen de tolerancia, minimiza el error, individualiza el hiperplano que maximiza el margen de predicción tomando en cuenta el error tolerado. Matemáticamente se expresa como:

$$Y = \frac{1}{2}\|w\|^2 + c\sum_{i=1}^{N}\varepsilon_i + \varepsilon_1^* \qquad (3).$$

Donde w representa al hiperplano del vector variable, C denota la compensación entre la regularidad de f y la cuantía hasta la cual toleramos desviaciones mayores que ε, ε_i y ε_1^* representan el error de valores fuera del subespacio superior e inferior respectivamente (Cao et al., 2020). Una de las aplicaciones en el campo financiero, es la determinación de sueldos dependiendo del nivel del puesto.

- *Random forest regressor (RFR)*: es un modelo atractivo por la capacidad de manejar un gran número de variables con relativamente pocas observaciones. El RFR está basado en la unión de varios DTs individuales y no correlacionados que contienen un subconjunto aleatorio de atributos, de los cuales se selecciona el mejor árbol con la mejor predicción. El RFR es construido a partir de la partición recursiva de grupos de datos similares de acuerdo con los criterios de división (Grömping, 2009). Uno de los usos en el campo económico, corresponde a la predicción del precio de una acción de la bolsa de valores tomando en cuenta los atributos y datos históricos.
- *Multilayer perceptron regressor (MLP)*: esta red neuronal consiste en múltiples capas de procesamiento de elementos a través de funciones o neuronas, que interactúan con los datos usando conexiones ponderadas. El modelo se estructura por capas receptoras de datos, capas intermedias u ocultas y capas de salida. En

términos generales, se encuentra inspirado en el modelo simple de las redes neuronales como se muestra en la Figura 1.

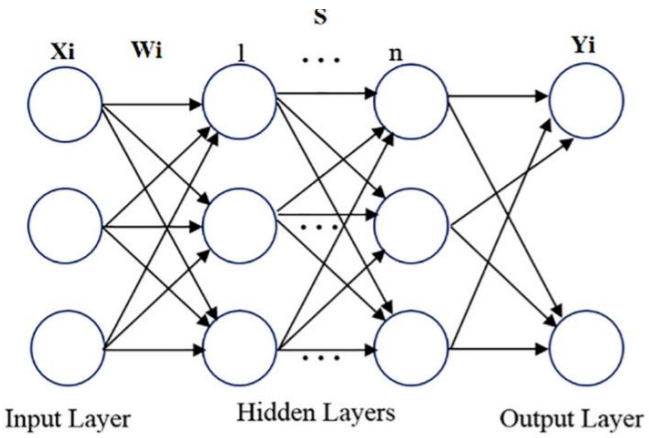

Figura 1. Resumen de una arquitectura MLP de redes neuronales artificiales. Fuente: (ALTobi et al., 2019)

Donde se denota por: $X_i = [X_0, X_1, X_n]$ a las nuevas entradas, $W_i = [w_0, w_1, w_n]$ los pesos y Y_i como la salida. Siendo S las sumas de los productos de las entradas y los pesos (ALTobi et al., 2019). El proceso de obtención de predicciones se logra propagando el error de las primeras predicciones obtenidas desde las capas de salida hasta los pesos, donde se aplican funciones de ajustes para obtener nuevos resultados. Este proceso se conoce como aprendizaje supervisado usando el algoritmo de retro propagación del error hacia atrás (back-propagation error), el cual repetitivamente es aplicado hasta obtener el promedio de la suma de errores cuadrados (MSE) cerca del coeficiente de aprendizaje previamente definido (leraning rate). Uno de los usos de este modelo en el campo económico, corresponde al pronóstico de la varianza mensual de las acciones en el mercado de la bolsa de valores.

2.3. Estado del arte de árboles de regresión

Los árboles de regresión (RT) son un método de aprendizaje supervisado de ML usado para predecir variables continúas basado en un conjunto de datos. Los modelos de RTs son construidos a partir de la estrategia de divide y conquistarás, es decir que el conjunto de datos de entrada del árbol es dividido en múltiples particiones de acuerdo con un criterio de división establecido. Los criterios de división se usan de acuerdo con la mejora de una función de costo y la mejora del criterio de costo-complejidad para podar un árbol. Dicha función está formulada para producir la subsecuencia de divisiones que obtengan predicciones con el error más pequeño posible (Gerber et al., 2013). Un ejemplo de la estructura de un árbol se puede observar en la Figura 2.

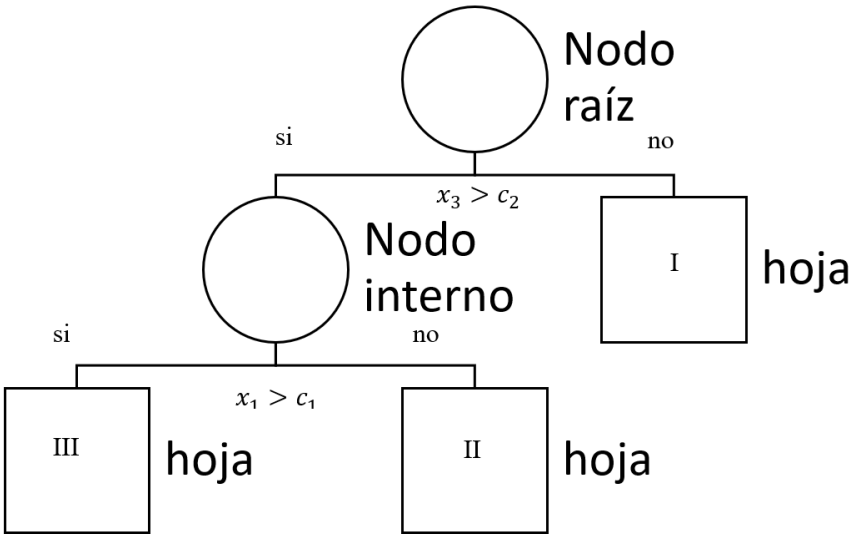

Figura 2. Estructura de un DT. Fuente: (Sepúlveda & Correa, 2013)

Este árbol, en particular contiene tres niveles de nodos y dos de profundidad estructurados de la siguiente manera: el nodo principal, es el nodo con el atributo más general, este nodo es llamado nodo raíz. El nodo interno, en el segundo nivel, contiene una subclasificación del nodo raíz. Los nodos terminales están ubicados en el tercer nivel, los cuales son de carácter homogéneo

y contienen valores específicos que pueden ser usados para la predicción (Sepúlveda & Correa, 2013).

En términos generales, la construcción de DTs es un proceso simple, que no demanda gran cantidad de recursos computacionales y provee una explicación clara de las decisiones tomadas desde el nodo raíz hasta el nodo terminal u hojas. Este procedimiento puede ser llevado a cabo computacionalmente por un sistema recursivo binario denominado classification and regression trees (CART), el cual será analizado posteriormente.

2.3.1. Árboles de regresión CART.

El algoritmo CART fue planteado por Leo Breiman y otros en 1984, rápidamente fue tomada en cuenta por la comunidad científica debido a su fácil implementación en todo tipo de problemas y su clara interpretación de los resultados. Al pasar de los años, este algoritmo comienza a tomar una serie de modificaciones llegando hasta 2006, donde Hothorn, Hornik y Zeileis proponen un marco unificado para el particionamiento recursivo, el cual incorpora modelos de regresión con estructura de árbol juntamente con varios criterios de parada y división que muestran el desempeño predictivo de los árboles resultantes (Sepúlveda & Correa, 2013). Finalmente, la concepción del algoritmo CART se usa para la creación de árboles binarios con fines de clasificación o regresión, es decir que en cada nodo no terminal puede tomar la decisión de crear una nueva división de la muestra continua o categórica. Las particiones para los árboles de regresión se crean conforme al siguiente procedimiento según el algoritmo CART:

- Se forman subgrupos de características similares para cada atributo del conjunto de datos. Los subgrupos más adecuados son determinados por el algoritmo CART como se puede observar en la Figura 3.

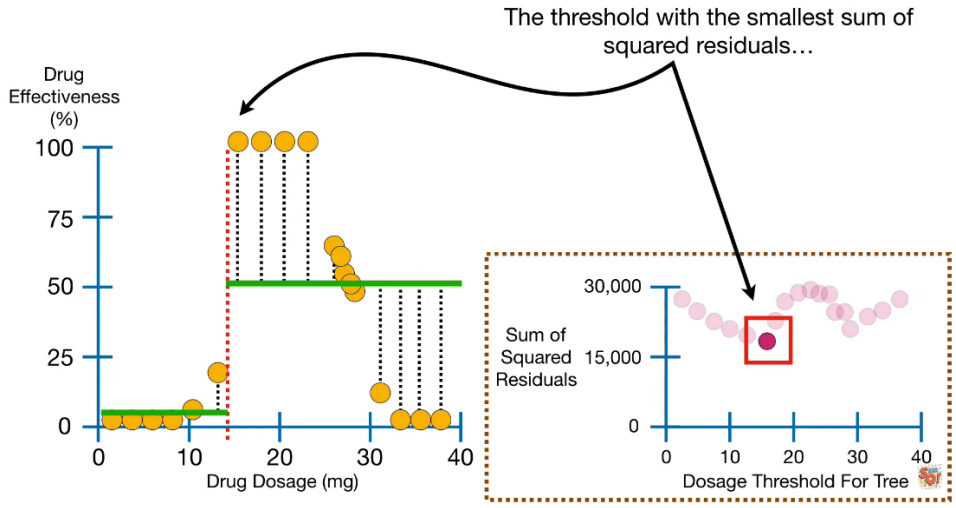

Figura 3. Proceso del algoritmo CART para la división de observaciones. Fuente: (Starmer, 2019)

Donde se evidencia que en cada iteración se van cambiando los rangos de las agrupaciones, se identifica el valor promedio correspondiente al rango, se busca el mínimo error de las diferencias entre las observaciones y el valor promedio del rango; este error se calcula por medio del criterio de división seleccionado en los hiperparámetros, estos criterios pueden ser:

- El MSE, el cual se representa matemáticamente como:

$$MSE = \frac{1}{N}\sum_{i=1}^{n}(y_i - \hat{y}_1)^2 \qquad (4).$$

  Donde y_i representa un valor observado del atributo y $\hat{y}_1$ el valor del límite superior en el rango tomado, es decir que se ira obteniendo el promedio de las sumas de las diferencias al cuadrado de los valores observados y el valor límite del rango seleccionado.

- MSE friedman, el cual es usado para evaluar las particiones potenciales en una región.

$$Friedman\ MSE = \frac{w_i w_r}{w_i + w_r}(y_i - \hat{y}_1)^2 \qquad (5).$$

La Ecuación 5 hace referencia a la Ecuación 4, pero, con la variación del incremento de los pesos, los cuales nos indican los valores seleccionados que están más cerca del resultado promedio.

- El promedio del error absoluto (MAE), el cual es una medida de diferencia promedio entre los valores de un conjunto de datos.

$$MAE = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_1|}{n} \qquad (6).$$

Donde y_i representa un valor observado del atributo y $\hat{y}_1$ el valor representante del nodo.

Finalmente, los rangos con los mínimos valores obtenidos del criterio de selección determinaran los candidatos para ser la raíz de cada nodo y formar los subgrupos de un atributo particular.

- En el caso de tener más de un predictor o atributo, se realiza el procedimiento mencionado en el guion anterior en cada uno de los atributos y se irá tomando los rangos–atributos con los mejores resultados obtenidos por cada uno de los criterios de selección, cuyos resultados serán usados como nodos raíces en el árbol.
- El algoritmo CART reconoce el final de la ejecución de acuerdo con las reglas de parada, las cuales se pueden ejecutar debido a que: se ha alcanzado la máxima profundidad del árbol o no se pueden realizar mas particiones debido a que no existen más variables relevantes, es decir que el número de elementos en el nodo terminal es inferior al número de casos permitidos en la partición.

2.3.2. Poda de los árboles de regresión.

La generación de un modelo de RT demasiado grande y específico para el conjunto de datos de entrenamiento, corre el riesgo de sobre ajustar el modelo a los datos de entrenamiento y producir un pobre rendimiento con los datos de prueba. Por otro lado, la generación de un modelo de RT demasiado pequeño podría no capturar toda la información estructural importante sobre el conjunto de datos. Por estas razones, los algoritmos de creación de árboles de decisiones deben estar asociadas con las técnicas de poda, las cuales se utilizan para la generalización de modelos. Estas técnicas reducen el tamaño de los árboles de decisiones, removiendo las subsecciones de los árboles que proveen menos poder de predicción. Por lo tanto, estas técnicas mejoran la precisión predictiva de un modelo mediante la reducción del sobreajuste. La metodología de poda usada por el algoritmo CART se divide de la siguiente manera:

- Pre-poda: Es un conjunto de parámetros que se establecen para detener el proceso de generación de un árbol antes de que este finalice, este procedimiento se basa en la metodología de mirada hacia adelante, donde se supervisa constantemente las divisiones de los nodos hasta cuando se obtenga un descenso en la tasa de error aceptable en los nodos terminales (Goicoechea, 2002).
- Post poda: Una de las técnicas más usada para la poda de árboles de decisión en el algoritmo CART se define como costo-complejidad, la cual compara las distintas subsecciones de un árbol finalizado, calificando el rendimiento de cada subsección del árbol con respecto al MSE y la penalidad de la complejidad del árbol, esta penalidad se encuentra en función de el número de hojas o de nodos terminales, la cual compensa a las subdivisiones menos complejas por la

diferencia en la cantidad de nodos terminales. Esta técnica se representa matemáticamente por la siguiente ecuación.

$$Tree\ Score = R(T) + \alpha T \qquad (7).$$

Donde el R(T) se define como la tasa total de clasificación errónea de los nodos terminales, α es un parámetro variable que mide el aumento de la taza de error por hoja podada y es inversamente proporcional a la profundidad del árbol, se puede obtener una visión más general usando validación cruzada (CV) para la obtención de alfas y T es el número total de nodos terminales (Esposito et al., 1997).

2.4. Estudios relacionados al estado del arte de modelos de regresión

En 2006, Bouzida y Cuppens hicieron un estudio donde ponen en evidencia las falencias de los sistemas de autenticación de aquellos tiempos, ya que no eran capaces de detectar ataques de hackers ni realizar técnicas preventivas. Por ello, su necesidad se basa en implementar un sistema de detección de intrusos para la mitigación de los ataques por hackers, por medio de ANNs y DTs. Los investigadores fundamentan que estas técnicas son aplicables para detectar anomalías y sus resultados de experimentación demuestran que las ANNs son altamente exitosas para detectar ataques conocidos, mientras que los DTs son mayormente capaces de detectar nuevos ciber ataques (Bouzida & Cuppens, 2006).

En 2011, Chang y colectivo de autores, proponen una solución al problema de la alta dimensionalidad y variaciones no estacionarias del precio en el mercado de acciones, a través de la construcción de un sistema de comercio de acciones preciso usando árboles de decisión difusa (CBFDT). El cual se somete a pruebas experimentales usando una base de datos de alta numerosidad y dimensionalidad relacionada al mercado de acciones. Finalmente, los

investigadores concluyen que el modelo CBFDT permite generar un conjunto de reglas para aplicarlas en la toma de decisiones sobre el movimiento del precio de las acciones (Chang et al., 2011).

En 2012, Shen y colectivo de autores propusieron el uso del algoritmo de árboles de decisión estimulado (MART), para explotar la correlación temporal entre los mercados bursátiles mundiales y varios productos financieros, con el fin de obtener la tendencia bursátil de días posteriores. Los datos usados como entrenamiento provienen de fuentes como los índices: NASDAQ, S & P500 y DJIA; para los cuales el modelo MART propuesto alcanzó el 74.4%, 76% y 77.6% de valores predichos correctamente. Como conclusión de la investigación, los investigadores llegaron a establecer este modelo como método de predicción comercial de análisis simple y lo contrastaron con un modelo de SVM (Shen et al., 2012).

En 2013, Al-Radaideh y colectivo de autores, propusieron el pronóstico del rendimiento de las acciones del mercado financiero por medio de la generación de DTs en WEKA (software), con el porpósito de ayudar a los investigadores del mercado de valores a decidir cuando es el mejor momento para comprar o vender acciones, en base al conocmiento extraido de los precios históricos. Para la contrucción del modelo propuesto, los investigadores hacen uso de la metodologia CRISP-DM (método para orientar trabajos de data mining) sobre los datos históricos de tres grandes empresas, obteniendo resultados con un promedio de acertividad del 44.68% al 52.59%, los cuales son aceptables para la vizualización de decisiones pero no para predicción. Según los investigadores los resultados obtenidos se deben a la gran cantidad de factores que afectan al mercado (Al-Radaideh et al., 2013).

En 2015, Nasseri y colectivo de autores, en su investigación sugieren un sistema inteligente para el soporte comercial basado en métodos de minería de texto y predicción de desiciones de

compra por medio de DTs, ya que identifican que las publicaciones en foros sociales en linea afectan los precios de los productos y alteran las decisiones de compra. El funcionamiento del sistema desarrollado por estos investigadores se encarga de extraer términos semánticos que expresen un sentimiento particular de compra o venta, posteriomente aplica un filtro de selección de palabras relevates en las publicaciones de los foros, una vez obtenidos los terminos relevantes construye un modelo de DT para determinar si los usuarios de los foros comprarán o no un producto. Los resultados de la investigación fueron positivos y concluyeron que el sistema propuesto permite medir la rentabilidad de un negocio dependiendo de las decisiones de los consumidores (Nasseri et al., 2015).

En 2016, Khaidem y colectivo de autores, proponen una forma de minimizar el riesgo de inversión en acciones del mercado mediante la predicción del rendimiento de las acciones utilizando modelos de RF. Estos modelos reciben como valores de entrada índices estadísticos como: el índice de fuerza relativa (RSI), la media móvil de convergencia-divergencia (MACD), la tasa de cambio del precio (PROC), entre otros. Los cuales son calculados de bases de datos públicas de Samsung y otras marcas populares. Como conclusión, estos investigadores determinaron que para los datos experimentales elegidos, el modelo de RF logró superar a otros modelos de predicción obteniendo un rango del 84 al 96% de valores predichos correctamente (Khaidem et al., 2016).

En 2017, Gupta en su investigación defiende la hipótesis sobre la utilidad y complejidad que existe en la creación de patrones de clasificación en diversos campos como la medicina, valores de mercado, reconocimiento por imágenes y robots autónomos. También, impulsa a que el ML puede ser usado para el reconocimiento de patrones complejos, aunque el rendimiento de ciertos algoritmos puede ser afectado por el tipo de datos a clasificar. Para corroborar dicha afirmación,

Gupta compara el rendimiento de modelos DTs contra ANNs, en bases de datos relacionadas con la salud. Como resultado determinó que con respecto al uso estadístico en las bases de datos, los modelos DTs tienen un mejor rendimiento sobre los basados en ANNs, aunque la afirmación no se puede extender para el reconocimiento de imágenes (Gupta, n.d.).

En 2019, Zhou y colectivo de autores, en su investigación estudiaron los cambios diarios de los índices bursátiles del mercado. La predicción fue realizada a través de un modelo llamado LR2GBDT, que consiste en una mezcla de los modelos de regresión logística (LR) y árboles de decisión impulsados por gradientes (GBDT). El modelo es alimentado por indicadores técnicos de precios tomados del índice compuesto de la Bolsa de Shanghái y del índice compuesto de precios de acciones S&P 500, ambos sin preprocesamientos. El modelo propuesto se evaluó basado en una comparación con otros resultados experimentales provenientes de otros modelos tales como: LR, GBDT, SVM y ANN. Como resultado, el modelo propuesto demostró superar a los restantes en la cantidad de valores predichos correctamente y ser mejor en la explotación de estrategias comerciales (Zhou et al., 2019).

CAPÍTULO 3: MÉTODO PROPUESTO

3.1. Proceso de desarrollo del modelo propuesto

En este trabajo se sigue el ciclo de vida tradicional de cualquier proceso de Data Mining, el cual se resume gráficamente en la Figura 4.

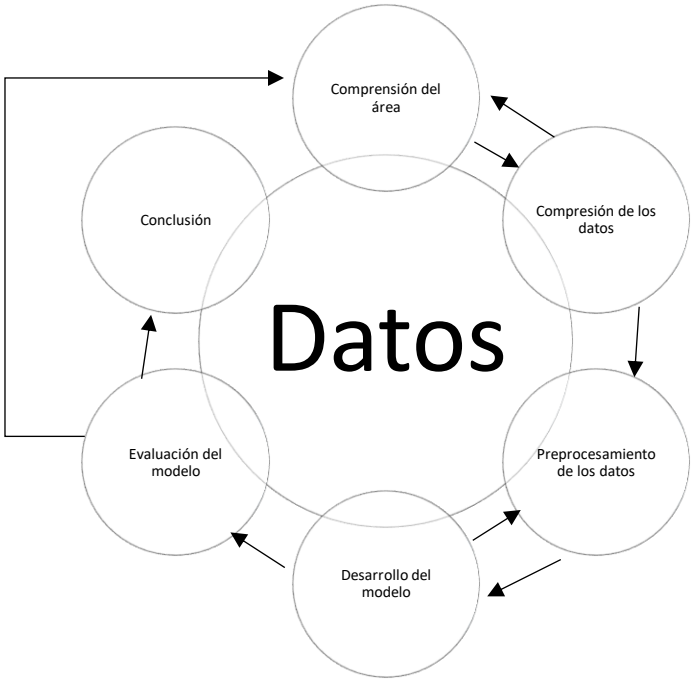

Figura 4. Ciclo de vida de un proceso de minería de datos.

La investigación propone la construcción de tres modelos de RTs para abordar el problema de predicción de valores de venta en el entorno comercial. Para los cuales, se usó bases de datos públicas que estén relacionadas con las tendencias de venta y contengan atributos que generen variabilidad en el valor de salida, por lo que la variable de salida debe ser de carácter continuo. Con el objetivo de analizar y explorar la naturaleza de los datos en cada una de las bases de datos, se usó algunas de las técnicas de exploración propuestas en el estado del arte, las cuales demostraron que existen variables con valores nulos o incompletos, atributos con distribuciones sesgadas, atributos altamente correlacionados, atributos no relacionados con la variable de salida,

atributos con valores fuera de lo común, entre otras observaciones. Lo que sugiere, que es necesario realizar un preprocesamiento de los datos, antes de ser dispuestos a los modelos de RTs. El preprocesamiento incluye un conjunto de pasos generalizados para el tratamiento de los problemas hallados en la exploración de datos de las tres bases de datos experimentales. Una vez obtenidos los datos completos y de calidad, se dispone a que ellos sean distribuidos aleatoriamente en un conjunto de prueba y otro de entrenamiento, los cuales nos permitirán tener una idea de como se ajusta el modelo de RT a diferentes datos proporcionados.

Por otro lado, con el objetivo de crear los mejores modelos de RTs generalizados, se realiza un proceso de tunning o personalización en cada uno de los modelos de RTs. Para lo cual se siguió las siguientes estrategias:

- Pre-poda: estableció los mejores hiperparámetros antes de que el modelo sea construido mediante la prueba y el error.
- Post-poda: se hizo uso de la técnica de costo-complejidad, mencionada en el estado del arte, para obtener el mejor rendimiento de un árbol en base a su profundidad y nodos terminales.
- Búsqueda aleatoria CV: este método permitió el uso de varios modelos de RTs con diversas configuraciones de hiperparámetros en distintos segmentos de las bases de datos, lo cual dejo en evidencia las mejores configuraciones de los modelos de RTs para las bases de datos experimentales.

Como medidas de identificación de los mejores modelos de RTs, se propuso las reglas de oro basadas en la evaluación del tiempo de ejecución, coeficiente de determinación R^2 de la predicción y el MSE de la predicción con respecto a valores reales. Finalmente, con los mejores modelos de RTs logrados, se dispone a mostrar su eficiencia comparándolos con un modelo de

redes neuronales artificiales MLP, para así llegar a una conclusión sobre las motivaciones y objetivos planteados por la investigación.

3.2. Metodología Experimental

Esta sección busca orientar la práctica de la investigación en relación con los conceptos teóricos brindados previamente y adecuar la solución de los objetivos planteados mediante la experimentación. Este diseño experimental contiene los detalles de cada uno de los pasos para la creación de los mejores modelos de RTs para las bases de datos seleccionadas y su objetivo es demostrar mediante resultados la eficiencia de los modelos de RTs. La metodología seleccionada para la experimentación se describe mediante los siguientes puntos:

1. Búsqueda de bases de datos experimentales
2. Análisis exploratorio de datos
3. Preprocesamiento de datos
4. División y entrenamiento de datos
5. Configuración de hiperparámetros del modelo de RT
6. Evaluación y criterios de selección
7. Rendimiento de los modelos de RTs
8. Comparación basada en el estado del arte

3.2.1. Bases de datos experimentales.

Las bases de datos experimentales fueron seleccionadas de acuerdo con el cumplimiento de especificaciones como: un mínimo de 1000 observaciones, inclusión de atributos continuos, una variable de salida de carácter continuo, relación de las bases de datos con el tema de ventas comerciales, mínima cantidad de atributos con valores nulos, origen de fuentes fiables y de libre

uso experimental. En base a las especificaciones, las siguientes bases de datos fueron seleccionadas para la experimentación.

1. *House Prices.*

La base de datos "*House Prices"* tiene como objetivo la predicción de precios de viviendas, basado en los atributos físicos de cada propiedad ubicada en Ames, Iowa, Estados Unidos (EE. UU). La base de datos fue recolectada por Dean De Cock y puesto a disposición al público a través de Kaggle (comunidad online de científicos de datos y ML). Los atributos de la base de datos reflejan atributos de interés común para compradores de casas en EE. UU, por ejemplo, el tamaño del lote, el año de construcción, cantidad de habitaciones, con o sin jardín, materiales utilizados en la construcción de la casa, estilo de la casa, entre otros. La base de datos incluye 43 variables categóricas y 36 variables continuas que forman un total de 2920 observaciones. En particular, la variable de salida y de interés es el precio de la casa, el cual es continuo y será el objeto para futuras predicciones.

2. *Video Games Sales.*

La base de datos "*Video Games Sales"* tiene como objetivo la predicción de ventas globales de diversos juegos en varias consolas. La base de datos fue generada por VGChartz y publicada en Kaggle. VGChartz es un sitio web de seguimiento de videojuegos, este sitio proporciona cifras de ventas semanales de software y hardware clasificado por región. Los atributos de la base de datos están compuestos por diversas características de un videojuego como: el ranking, el nombre, la plataforma, el año de lanzamiento, el género del videojuego, las ventas en diferentes regiones, entre otros. Esta base de datos incluye 6 variables categóricas y 5 variables continuas las cuales en total contienen 16600 registros. La variable de salida es el atributo "Global Sales", el cual

representa todas las ventas a nivel mundial (en millones), esta variable es de carácter continuo y será la variable por predecir con nuevos valores ingresados.

3. *Russian Housing Market.*

Esta base de datos está compuesta por un conjunto de características de viviendas en Rusia, la cual plantea una interacción entre las características de una casa y las complicaciones para la predicción de su precio. La base de datos fue planteada por Sberbank, el banco más antiguo y grande de Rusia; el cual ayuda a sus clientes haciendo predicciones sobre los precios de bienes inmuebles para inquilinos o inversores, los cuales buscan tener mayor asertividad y confianza al momento de comprar una vivienda. Los atributos de esta base de datos contienen elementos físicos de una vivienda y su ubicación, por ejemplo, el área de construcción, material de construcción, centros educacionales, centros comerciales, año de construcción, áreas verdes, entre otros. La base de datos cuenta con 276 variables continuas y 16 categóricas, las cuales entre sus observaciones suman un total de 30471, donde su variable de salida se llama “price_doc”, la cual demuestra el precio publicado en el mercado.

3.2.2. Análisis exploratorio de datos.

La exploración de datos se define como el análisis inicial de datos para la construcción de cualquier modelo de aprendizaje automático. Este proceso es esencial para conocer la estructura, propiedades y particularidades de las bases de datos experimentales, con el fin de evitar entradas globales que sean superficiales y engañosas para el modelo de RT. La exploración de datos en esta investigación utiliza algunas de las técnicas de exploración visual y estadística estudiadas en el estado del arte, las cuales son puestas a disposición por las librerías de pyplot y seaborn en Python v3.7.

Con el objetivo de ilustrar este procedimiento generalizado para las tres bases de datos experimentales, se usará la base de datos *House Prices* como objeto de prueba y análisis. De tal manera, la experimentación sigue los siguientes puntos de enfoque:

- El primer punto de enfoque es el análisis de la variable de salida, en este caso "SalePrice", para la cual es necesario tener una descripción del conjunto de observaciones de la variable.

Medidas	**Valor**	**Descripción**
Media	180921,20	Indica el promedio del atributo
Cantidad de observaciones	1460	Informa la cantidad de observaciones que contiene el atributo
Desviación estándar	79442,50	Informa que tan dispersos están los valores con respecto a la media
Valor mínimo	34900	Indica el valor más pequeño del atributo
Primer cuartil	129975	Indica el valor máximo del 1/4 de la muestra 129975
Segundo cuartil	163000	Indica el valor máximo del 1/2 de la muestra
Tercer cuartil	214000	Indica el valor máximo de 3/4 de la muestra
Valor máximo	755000	Informa el valor más grande del atributo

Tabla 1. Descripción estadística de la variable "SalePrice"

Con respecto a la Tabla 1 observamos que la variable "SalePrice" no tiene valores nulos, debido a que la cantidad de observaciones del atributo es la misma cantidad de índices de la base de datos. Además, demuestra que todas las observaciones constituyen valores reales positivos y no inventados, aunque la desviación la desviación estándar deja en evidencia que existe valores atípicos, ya que el valor mínimo y máximo se encuentran fuera del rango. Para corroborar las ideas del análisis estadístico de la variable "SalePrice" es necesario visualizar su distribución.

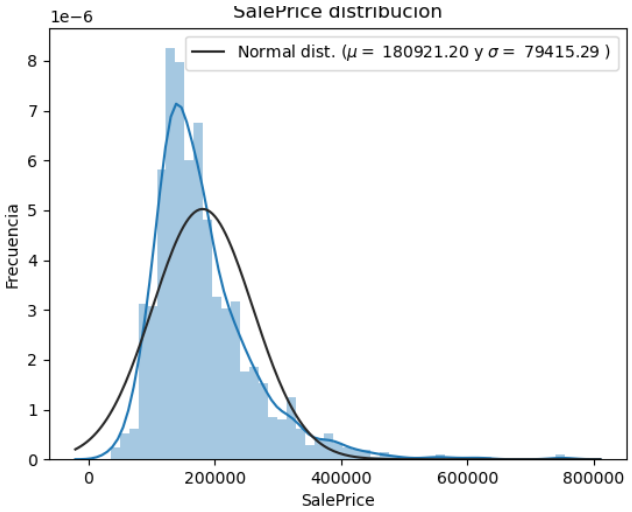

Figura 5. Distribución de datos de la variable “SalePrice”

En base a la Figura 5, se puede determinar que la distribución de esta variable se encuentra desviada de la distribución normal, tiene valores atípicos, tiene asimetría positiva y muestra un pico. Por lo cual, será necesario pre-procesar para llegar a una distribución normal, antes de ingresar la variable al modelo de RT.

- Como segundo punto de enfoque, es necesario determinar la relación de las variables entre sí y con la variable de salida con el fin de descartar aquellas que no se relacionan por medio de la matriz de correlación y el método de Pearson.

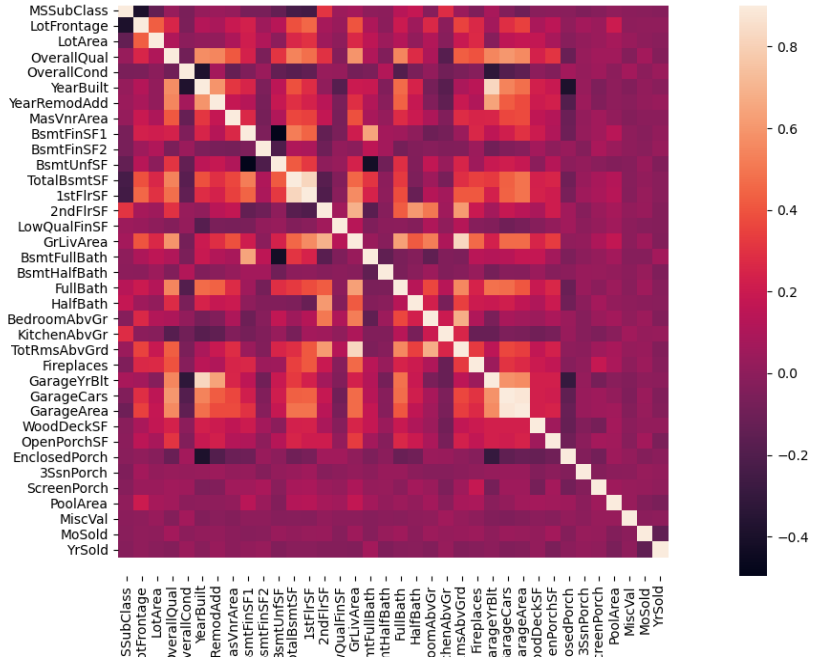

Figura 6. Matriz de correlación de las variables del conjunto de datos “HousePrices”

La matriz de la Figura 6 nos permite inferir que: las variables como “OverallQual”, “GrLivArea” y “TotalBsmtSF” son las variables mayormente relacionadas con la variable de salida “SalePrice”, por lo cual estas variables serán importantes para predecir nuevos valores. También, la matriz permite determinar las variables que actúan como gemelas, por ejemplo: “GarageCars” y “GarageArea”, de las cuales una de las dos puede ser eliminadas, ya que entregan la misma información. Por otro lado, la Figura 6 nos permite tener una idea de las variables que no tienen nada que ver con la variable de salida “SalePrice” y descartarlas.

- En el tercer punto de enfoque, vamos a visualizar cómo reaccionan los datos de las variables predictoras con respecto a la variable de salida para comprender su naturaleza.

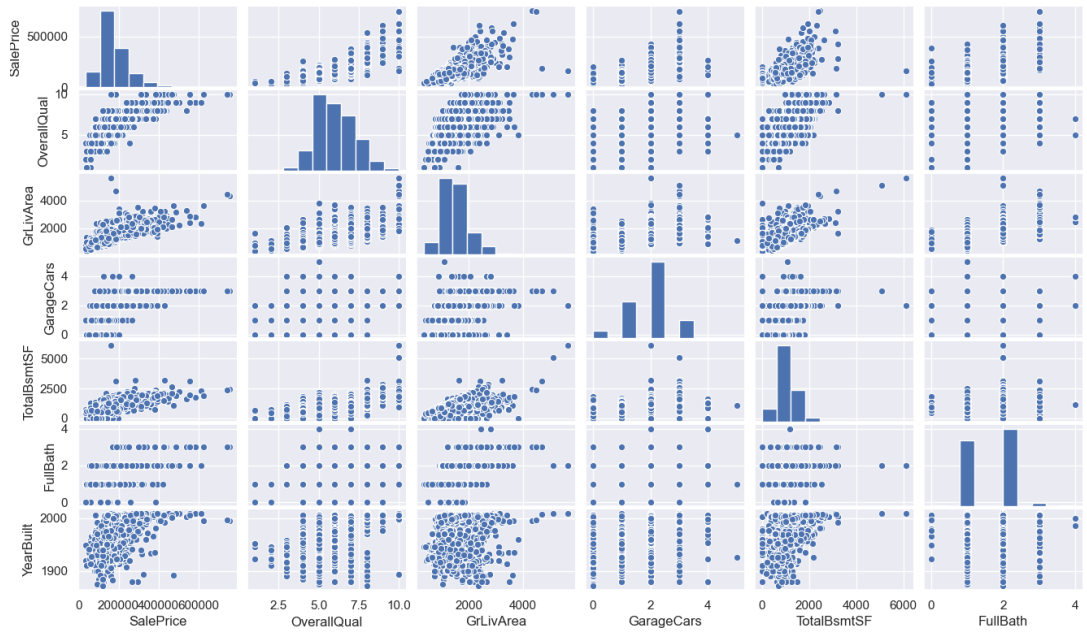

Figura 7. Matriz de la distribución de datos de las variables correlacionadas en la base de datos “HousePrices”

La Figura 7 nos permite comprender de mejor forma como se comportan los datos de las variables relacionadas unas con otras. Brevemente nos informa que la tendencia de los datos es de forma creciente, es decir que mientras una variable crezca las otras también crecerán. Esta premisa nos ayuda a plantear que la relación de las variables predictoras con la variable de salida genera un crecimiento de comportamiento lineal o exponencial, lo cual nos podría dar un indicio con respecto a la conclusión, ya que el crecimiento lineal afecta al rendimiento de un modelo de RT. Adicionalmente, la Figura 7 deja en evidencia que existen valores sesgados en las muestras de los atributos y que muchas de las distribuciones necesitan ser normalizadas en el preprocesamiento.

3.2.3. Preprocesamiento de los datos.

El preprocesamiento de los datos es una etapa fundamental para el traspaso de simples datos sin significancia a datos consistentes y de calidad para la aplicación en un modelo de ML, tomando en cuenta que el resultado de las predicciones o clasificaciones depende en gran medida de la calidad de datos proporcionados al modelo. El preprocesamiento de datos propuesto en esta investigación esta formado por una serie de técnicas de limpieza de datos, transformación de datos, integración de datos, normalización de datos, imputación de valores faltantes e identificación de ruido; los cuales servirán al algoritmo CART para formular modelos de RTs óptimos. Estas técnicas están enfocadas en solucionar los problemas evidenciados en la exploración de datos de cada una de las bases de datos, para los cuales se generalizó los siguientes pasos:

- La eliminación de observaciones sesgadas, principalmente de los atributos más importantes
- La eliminación de atributos poco correlacionados con la variable de salida, así como la eliminación atributos gemelos.
- La eliminación de atributos que contienen un porcentaje mayor al 50% de observaciones faltantes.
- La imputación de observaciones faltantes a cada uno de los atributos, los valores imputados serán provenientes de la moda o media de los atributos, dependiendo de su característica categórica o continua.
- Codificación de las variables categóricas en variables continuas.
- Integración de variables que reemplacen y entreguen mayor información que otras.
- Estandarización de los datos y obtención de una distribución normal en la variable de salida.

3.2.4. División de datos y entrenamiento del modelo.

La división de datos en un conjunto de entrenamiento y un conjunto de pruebas es necesaria para asegurar la generalización en la creación de un modelo de RT, ya que el set de entrenamiento permite diseñar la regresión del modelo y el set de prueba permite evaluar que tan bien se ajusta el modelo a los datos no vistos antes.

Para esta investigación se asignó aleatoriamente, el 20% de la totalidad de los datos de cada base de datos experimental al conjunto de prueba y el 80% de datos restantes al conjunto de entrenamiento. Estos conjuntos de datos de entrenamiento fueron dispuestos para su cometido a los modelos de RTs, los cuales fueron creados en base al siguiente procedimiento:

- Se determinó el rango funcional de los hiperparámetros mostrados en la Tabla 2 correspondientes a un modelo de RT, por medio de los métodos de pre-poda.

Hiperparámetros	**Rango**	**Observación**
Max depth	0 al 21	Los árboles dentro de este rango de profundidad muestran variabilidad en su rendimiento fuera de él, solo disminución del rendimiento.
Leaf sample	1 al 20	Los árboles con este rango de ejemplos mínimos por hoja muestran buen rendimiento, fuera de él muestran decremento en el rendimiento.

Tabla 2. Rango de hiperparámetros determinados en la pre-poda

  Estos rangos permitirán predefinir algunas de las configuraciones de la construcción de un modelo de RT.

- Se creó diferentes modelos de RTs con diferentes configuraciones de hiperparámetros haciendo uso del método de Randomized Search 10-Folds Cross Validation, con lo cual se dispone 10 modelos de RTs con diferentes

hiperparámetros para cada una de las bases de datos experimentales, los resultados se muestran en el sub-epígrafe 3.2.7.

- Una vez definidos los mejores modelos de RTs en base a los criterios de evaluación, se realiza el proceso de post-poda, en el cual se determina el Alpha ideal en base a la relación con la profundidad de los mejores modelos de RTs. Por ejemplo,

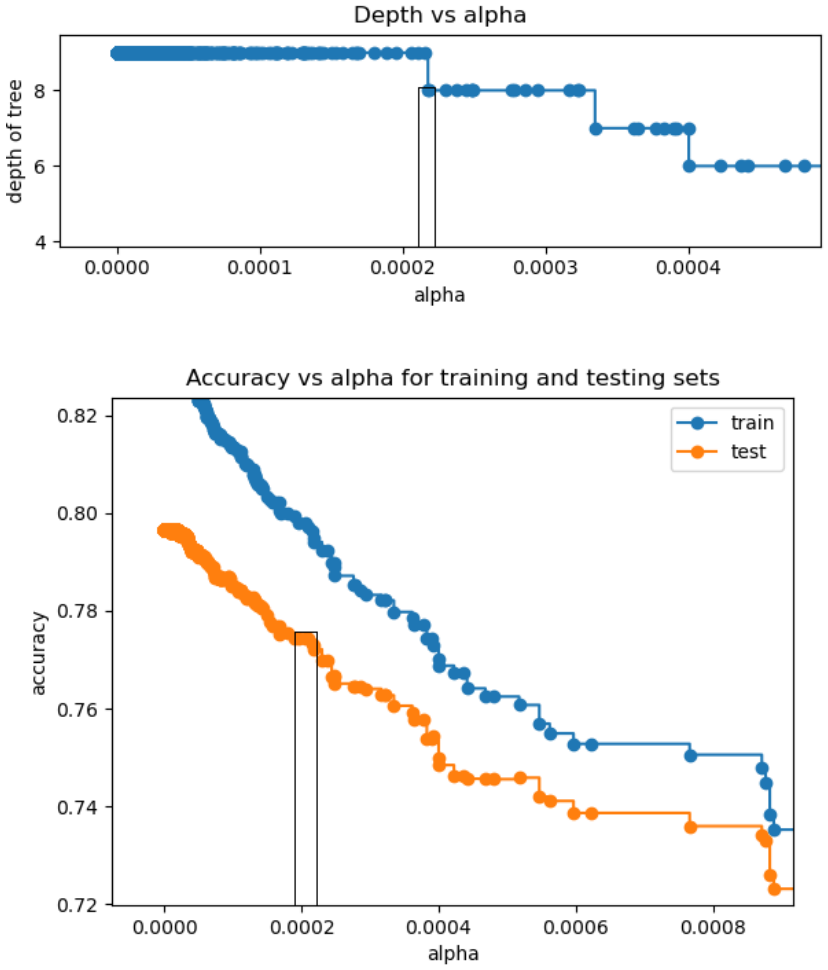

Figura 8. Relación de Alpha con la profundidad y rendimiento de un árbol de regresión.

El modelo de RT de la Figura 8 tiene una profundidad máxima de 9 niveles el cual genera el máximo rendimiento, pero con el fin de disminuir la probabilidad de

sobreajuste del modelo de RT con los datos, se selecciona el Alpha más equilibrado que mantenga una profundidad de buen rendimiento y disminuya la máxima profundidad de 9 niveles. En el caso de la Figura 8, se seleccionó un Alpha que reduce la profundidad del modelo de RT a 7 niveles y mantiene un rendimiento de alrededor de 78% de eficiencia.

3.2.5. Configuración de hiperparámetros del modelo.

Las distintas configuraciones de un modelo de ML permiten variar su comportamiento con respecto a los datos. Dependiendo de las variaciones de sus hiperparámetros el modelo de ML se acoplará de mejor manera a sus datos y obtendrá un mejor rendimiento, para ello es esencial la compresión de los distintos criterios e hiperparámetros.

Como parte del proceso experimental, la metodología de pre-poda y post-poda nos ayudan a comprender algunos de los mejores resultados con respecto a la configuración de un modelo. Se han puesto a prueba varias configuraciones de parámetros en distintos modelos de RTs por medio del método de Randomized Search 10-Folds Cross Validation, con el objetivo de identificar cuáles son los mejores parámetros, los cuales provienen de los modelos de RT con el porcentaje más alto de estimaciones cercanas a la variable real R^2.

El método Randomized Search se encarga de la selección de configuraciones aleatorias para cada modelo de RT con los siguientes parámetros:

- **Criterion:** se refiere a las medidas de selección de atributos, estos pueden ser mse, Friedman mse o mae, los cuales han sido detallados previamente en el estado del arte.

- **Max depth:** indica la máxima expansión del árbol hasta el último nodo terminal, se ha determinado mediante prueba y error que desde 0 hasta 21 se puede obtener un buen resultado
- **Leaf Sample:** se refiere al rango mínimo de observaciones que tendría un nodo hoja, los mejores resultados se han obtenido con un rango del 1 al 20.
- **Divisor:** es la forma en cómo el algoritmo tomara un atributo para dividir el conjunto de datos. Estos pueden ser: al azar, el cual toma el mejor atributo aleatorio para realizar la división; o "best", el cual realiza una búsqueda exhaustiva hasta encontrar la mejor división.
- **Max Features:** refiere al número de atributos para tener en cuenta al buscar el criterio de divisor, pueden ser "sqrt=sqrt(n_atributos)", "log2=log2(n_atributos)" o "None=max_atributos".

3.2.6. Criterio de evaluación del modelo y herramientas de implementación.

Los criterios de evaluación del rendimiento de un modelo indican el nivel de acierto de las predicciones obtenidas de un conjunto de datos mediante un modelo entrenado. El proceso experimental de esta investigación busca evaluar el rendimiento de los valores de venta predichos por medio de:

- Coeficiente de determinación (R^2), el cual representa la capacidad de la predicción del modelo por medio de la proporción de la variación total de la estimación y el valor real, es decir cuan ajustado está el modelo a los datos.

$$R^2 = 1 - \frac{\sigma_r^2}{\sigma^2} \qquad (8).$$

Donde σ_r^2 representa la varianza residual de la variable dependiente Y y σ^2 la varianza de las estimaciones de Y con respecto de los valores reales. El coeficiente R^2 puede variar entre 0 y 1, correspondiendo a que en 0 el modelo no explica nada y en 1 a que la regresión del modelo cruza exactamente por cada uno de los valores reales.

- MSE, este criterio busca mostrarnos el promedio de error de la suma de cuadrados de los residuos entre los valores estimados y los valores reales. Este criterio será usado en la comparación basada en el estado del arte del modelo de RT con el modelo MLP. El resultado de la evaluación define que modelo genera una regresión más precisa.
- Tiempo de ejecución, esta regla permitirá formular las conclusiones de cuan bajo es el costo computacional de la construcción del modelo de RT estandarizado y la ventaja comparativa de un modelo de RT versus un modelo complejo de predicción MLP.

La implementación del proceso experimental y de evaluación de esta investigación se realizó basado en el lenguaje de programación Python 3.7 y algunas de sus bibliotecas, entre las cuales podemos destacar:

- Sklearn, la cual contribuyo con las herramientas de aprendizaje automático y generación de RTs.
- Numpy, la cual permitió el manejo de vectores multidimensionales.
- Matplotlib, permitió la visualización de los datos.
- Pandas, el cual facilito la manipulación y análisis de los datos.

3.2.7. Rendimiento de los modelos de árboles de regresión.

De acuerdo con el objetivo de la investigación, para cada base de datos se aplicó la metodología experimental detallada en esta investigación. Los resultados logrados demuestran la obtención de los modelos de RTs construidos mediante el procedimiento de selección aleatoria y los resultados de la aplicación de la post-poda, con lo cual se asegura la globalidad de los mejores modelos de RTs.

La investigación registró un total de 30 modelos de RTs, como se muestra en la Tabla 2, perteneciendo 10 modelos de RTs a cada una de las tres bases de datos experimentales.

Bases de datos	**N° Config**	**Criteria**	**Max Depth**	**Max Features**	**Min Samples Leaf**	**Split**	R^2
House Prices	1	mae	7	log2	4	best	18,81%
	2		17		8	random	12,96%
	3	**mse**	**21**	**None**	**12**	**best**	**80,37%**
	4		13	None	9		80,23%
	5		21	sqrt	13	random	37,98%
	6		11	None	9		74,13%
	7		5	log2	8		11,65%
	8	friedman_mse	3		6	best	12,47%
	9		17	sqrt	2		51,19%
	10		3	None	10		72,14%
Video games sales	1	friedman_mse	15	None	18	random	66,85%
	2	**mse**	**21**	**log2**	**2**	**best**	**99,00%**
	3	friedman_mse	15	sqrt	7		97,88%
	4	mae	7	None	9		97,09%
	5		3		11		82,68%
	6	friedman_mse	0	log2	11	random	21,19%
	7	mae	13	None	17		62,14%
	8	friedman_mse	11	sqrt	15		11,00%
	9		21	None	9		84,87%
	10	mse	9	sqrt	7		27,78%
Russian Housing Market	1	mse	11	auto	10	best	79,71%
	2	mae	17	sqrt	19		77,74%
	3	mse	5		16	random	51,36%
	4	friedman_mse	13	**None**	1	**best**	75,30%

	5	**mse**	**9**		**15**		**80,00%**
	6	friedman_mse	11	sqrt	16	random	67,82%
	7		15	auto	18	best	79,97%
	8		13	None	9		79,19%
	9		5	sqrt	7	random	44,25%
	10		3	None	7		51,70%

Tabla 3. Hiperparámetros obtenidos de las 10 Fold Cross Validation en los diferentes modelos de RTs para cada base de datos experimental.

La Tabla 3 resume el trabajo experimental realizado por el algoritmo CART y contiene el detalle de los hiperparámetros usados en los modelos de RTs, el número de Fold o partición en el CV y el porcentaje del rendimiento obtenido por la medida de evaluación del modelo, sin haber realizado post-poda. De acuerdo con los resultados obtenidos en la Tabla 3, se seleccionó los mejores modelos de RTs (marcados en negrita) como ejemplares de cada una de las bases de datos:

- *House Prices:* el mejor modelo de esta base de datos usa el criterio de selección MSE, una profundidad máxima de 21 niveles, considera todos los atributos como candidatos para el criterio de división, mantiene un mínimo de 12 ejemplos para las hojas o nodos terminales y realiza búsqueda exhaustiva para encontrar la mejor división, este conjunto logra un R^2 de 80.3%
- *Video games sales:* su mejor modelo usa el criterio de selección MSE, una profundidad máxima de 9 niveles, considera todos los atributos como candidatos para el criterio de división, tiene un mínimo de 1 observaciones para las hojas o nodos terminales y realiza búsqueda exhaustiva para encontrar la mejor división, esta unión de estos parámetros da como resultado un 93.9% de R^2.
- *Russian Housing Market:* el mejor modelo usa el criterio de selección mse, una profundidad máxima de 9 niveles, el criterio de división que considera a todos los

atributos como candidatos para el criterio de división, tiene un mínimo de 15 observaciones para los nodos hojas o terminales y realiza búsqueda exhaustiva para encontrar la mejor división, este conjunto logra un R^2 del 80.0%.

En términos generales, se puede evidenciar que el criterio de selección MSE, el uso de todos los atributos en el criterio de división y la estrategia de búsqueda exhaustiva "best", son los hiperparámetros comunes que brindan el mejor R^2 para el algoritmo CART en estas bases de datos. En realidad, se esperaba que estos hiperparámetros sean parte de los mejores resultados, ya que estas configuraciones fuerzan al modelo de RT para realizar pruebas exhaustivas hasta encontrar las mejores predicciones posibles. Posteriormente, se aplicó proceso de postpoda para los mejores modelos de RTs, con lo cual se obtuvo los siguientes resultados.

Bases de datos	**Criterion**	**Max Depth**	**Max Features**	**Min Samples Leaf**	**Split**	**Alpha**	$\boldsymbol{R^2}$
House prices	mse	21	None	12	best	0,0006	83,02%
Video game sales	mse	9	None	1	best	0,003	93,27%
Russian house market	mse	9	None	15	best	0,0002	77,43%

Tabla 4. Aplicación de la post-poda a los mejores modelos de RTs.

El algoritmo CART permite realizar el método de post-poda por medio de la inclusión del hiperparámetro Alpha, el cual se determinó por medio del procedimiento mencionado en el tercer guion del suepigrafe 3.2.4, con lo cual los resultados de la Tabla 4 demuestran que el método de

post-poda puede disminuir un poco el rendimiento del modelo de RT, pero genera un modelo estandarizado y más sencillo de comprender, como se puede ver en las figuras 9-10-11.

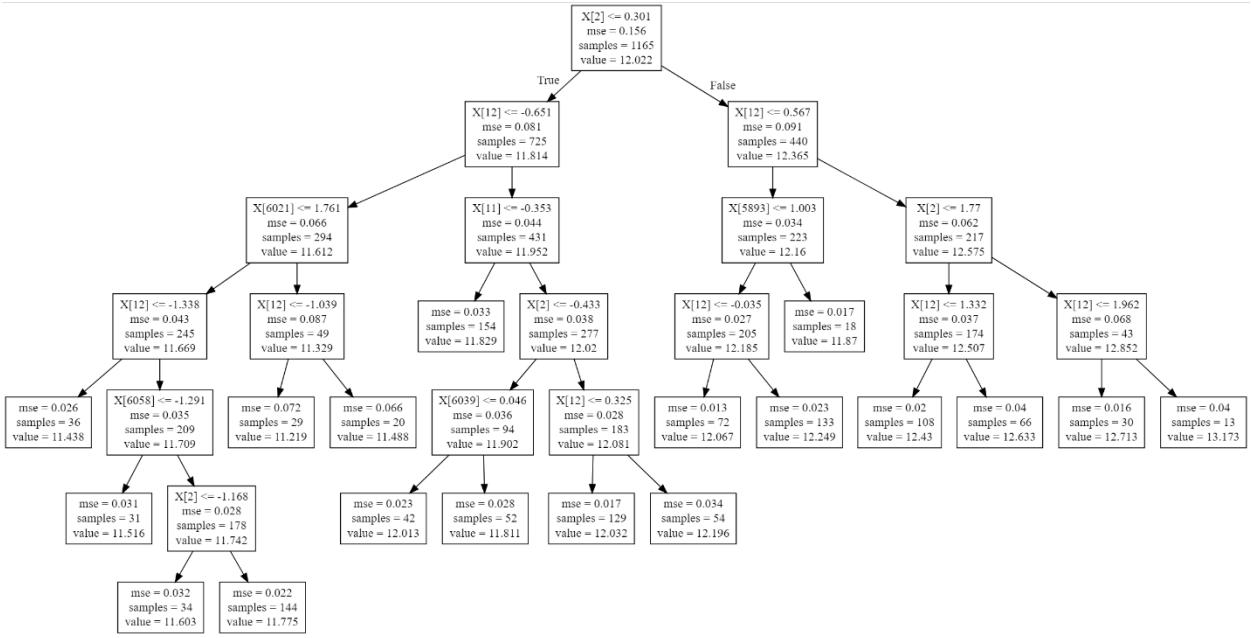

Figura 9. Árbol de decisiones correspondiente a la base de datos *House Prices*.

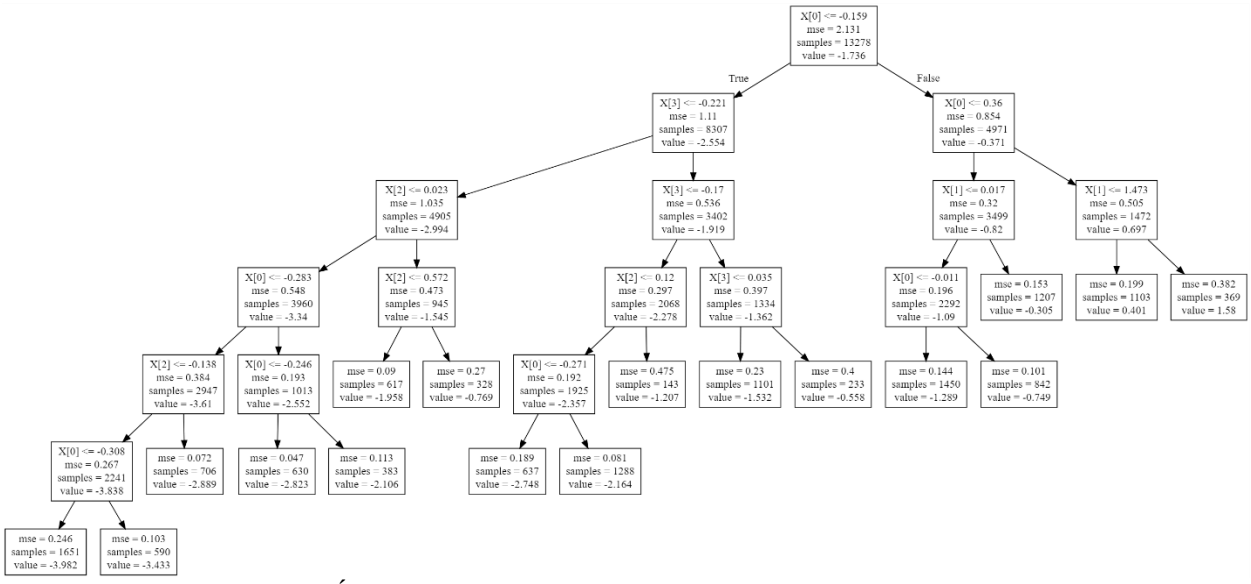

Figura 10. Árbol de decisiones correspondiente a la base de datos *Video Games Sales*.

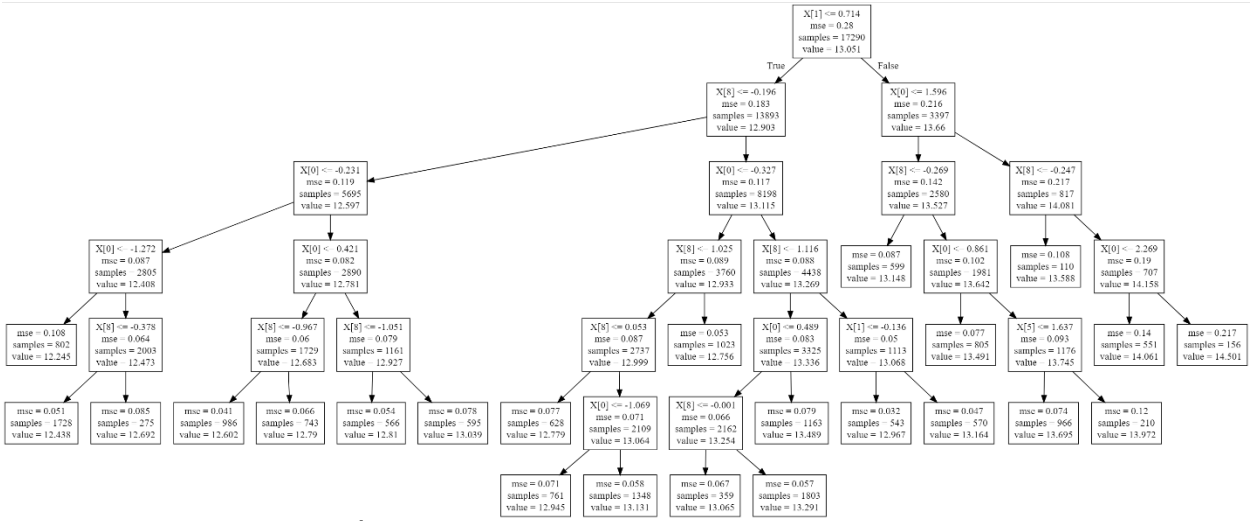

Figura 11. Árbol de decisiones correspondiente a la base de datos *Russian Housing Market*.

Para concluir, se puede reconocer que la metodología experimental para la creación de modelos de RTs, es capaz de encontrar buenos resultados que proporcionan un alto R^2. Además, demuestra las mejores configuraciones de hiperparámetros que pueden ser replicados a cualquier conjunto de nuevos datos y reproduce con sencillez el conjunto de decisiones tomadas por los modelos de RTs.

3.2.8. Comparación basada en el estado del arte.

La comparación de modelos se realiza con el objetivo de determinar cuál es el rendimiento del modelo experimental frente al rendimiento de otros modelos. Con el objetivo de demostrar la eficacia del rendimiento de los modelos de RTs en esta investigación, se comparó con un modelo complejo de regresión seleccionado del estado del arte en la sección 2.2.2. El modelo seleccionado fue el modelo MLP con 60 capas ocultas en cada fase e hiperparámetros estándar como base de comparativa. Con lo cual se obtuvo los siguientes resultados.

Bases de datos	Modelo	MSE	Time (sec.)	R^2
House Prices	**RT**	**0,027**	**0,14**	**80,37%**
	MLP	0,061	36,87	69,11%
Video games sales	**RT**	**0,051**	**0,011**	**98,82%**
	MLP	0,172	11,12	89,61%
Russian Housing Market	**RT**	**0,054**	**0,047**	**80,00%**
	MLP	0,047	61,42	82,72%

Tabla 5. Comparación basada en la literatura de modelos RT vs MLP

La Tabla 5 muestra: los tres mejores modelos de RTs obtenidos en la Tabla 3, el MSE de los valores predichos contra los valores reales, el tiempo de ejecución y el R^2 correspondiente a cada modelo. Con respecto a los resultados obtenidos, se logra evidenciar que el modelo MLP logra un buen rendimiento en las bases de datos con mayor cantidad de observaciones, aunque cabe mencionar que se podría mejorar el R^2 variando sus hiperparámetros, mientras que los modelos RTs simples muestran buenos resultados en todas las bases de datos. Por otro lado, la cantidad de tiempo empleado en la construcción de modelos tiene resultados resaltables, ya que el tiempo de construcción de un modelo MLP es bastante alto, considerando que el equipo usado para las pruebas cuenta con 16GB de RAM y un procesador Intel i7-8750H de 6 núcleos a 3.5GHz. Se puede concluir que los modelos de RTs son computacionalmente menos costosos y brindan resultados confiables.

CAPÍTULO 4: CONCLUSIONES

Esta investigación ha descrito un acercamiento de los árboles de regresiones y como obtener modelos apropiados en el contexto experimental de la predicción de precios de venta en el mercado. Se ha expuesto una metodología consistente y generalizada para el preprocesamiento de bases de datos, exploración y análisis de datos, una estrategia de selección de hiperparámetros y la comparación del rendimiento contra un modelo MLP.

Por otro lado, los resultados expuestos muestran predicciones aceptables estando entre el 76 y 99% de eficacia; los cuales son valores aceptables en el caso de predicciones de precios, ya que la estimación de un valor correspondiente al precio de venta mayormente depende de cuán importante es para el vendedor que el precio sea preciso o estimado, es decir que para muchos vendedores una estimación efectiva es la solución del problema de la variabilidad de los precios.

Finalmente, en base a los resultados se puede concluir que los árboles de regresión son una herramienta poderosa, simple, poco costosa, sencilla de comprender, adaptable a poca o gran cantidad de datos, comparable contra modelos robustos de predicción como MLP y puede servir de apoyo en diversas áreas de predicción.

4.1. Trabajo futuro

Uno de los puntos que destaca la investigación es el preprocesamiento de los datos como llave estratégica para el uso de los modelos. Se considera que, en futuras investigaciones el desarrollo de diversas estrategias y selección de atributos contribuirá a la mejora de modelos en términos de R^2, tiempo de ejecución y predicción. También, se incita al uso de diversas variaciones de modelos de árboles de regresión como RF, XGBoosted, BayesianTrees o modelos híbridos, los cuales han sido mencionados en investigaciones relacionadas.

CAPÍTULO 5: REFERENCIAS BIBLIOGRÁFICAS

Agrawal, A., Gans, J., Goldfard, A., & Moral, J. V. (2019). *Máquinas predictivas: La sencilla economía de la inteligencia artificial*. Reverte-Management. https://books.google.com.ec/books?id=Pz-7DwAAQBAJ

Al-Radaideh, Q. A., Assaf, A. A., & Alnagi, E. (2013). Predicting stock prices using data mining techniques. *The International Arab Conference on Information Technology (Acit'2013)*.

Alderete, A. M. (2006). Fundamentos del análisis de regresión logística en la investigación psicológica. *Revista Evaluar*, *6*(1).

ALTobi, M. A. S., Bevan, G., Wallace, P., Harrison, D., & Ramachandran, K. P. (2019). Fault diagnosis of a centrifugal pump using MLP-GABP and SVM with CWT. *Engineering Science and Technology, an International Journal*, *22*(3), 854–861. https://doi.org/https://doi.org/10.1016/j.jestch.2019.01.005

Bonialian, M. (2018). LA GLOBALIZACIÓN TEMPRANA. *Historia Mexicana*, *68*(2 (270)), 785–801. http://www.jstor.org/stable/26557182

Bouzida, Y., & Cuppens, F. (2006). Neural networks vs. decision trees for intrusion detection. *IEEE/IST Workshop on Monitoring, Attack Detection and Mitigation (MonAM)*, *28*, 29.

Cao, C., Liao, J., Hou, Z., Wang, G., Feng, W., & Fang, Y. (2020). Parametric uncertainty analysis for CO2 sequestration based on distance correlation and support vector regression. *Journal of Natural Gas Science and Engineering*, *77*, 103237. https://doi.org/https://doi.org/10.1016/j.jngse.2020.103237

Chang, P.-C., Fan, C.-Y., & Lin, J.-L. (2011). Trend discovery in financial time series data using a case based fuzzy decision tree. *Expert Systems with Applications*, *38*(5), 6070–6080. https://doi.org/https://doi.org/10.1016/j.eswa.2010.11.006

Esmeir, S., & Markovitch, S. (2007). Anytime learning of decision trees. *Journal of Machine Learning Research*, *8*(May), 891–933.

Esposito, F., Malerba, D., Semeraro, G., & Kay, J. (1997). A Comparative Analysis of Methods for Pruning Decision Trees. *Pattern Analysis and Machine Intelligence, IEEE Transactions On*, *19*, 476–491. https://doi.org/10.1109/34.589207

Fernández, A. I., & ernández, A. I. (1986). EL DIAGNOSTICO FINANCIERO DE LA EMPRESA. NUEVAS TENDENCIAS EN EL ANALISIS. *Revista Española de Financiación y Contabilidad*, *15*(49), 113–132. http://www.jstor.org/stable/42779772

Garayar P., G. (1953). INTRODUCCIÓN A LA ECONOMETRÍA. *El Trimestre Económico*, *20*(77(1)), 45–74. http://www.jstor.org/stable/20855328

Gerber, S., Rübel, O., Bremer, P.-T., Pascucci, V., & Whitaker, R. T. (2013). Morse-Smale Regression. *Journal of Computational and Graphical Statistics*, *22*(1), 193–214. http://www.jstor.org/stable/43304823

Goicoechea, A. P. (2002). Imputación basada en árboles de clasificación. *Eustat. Available in: Http://Www. Eustat. Es/Documentos/Datos/Ct*, *4*.

Grömping, U. (2009). Variable Importance Assessment in Regression: Linear Regression versus Random Forest. *The American Statistician*, *63*(4), 308–319.

http://www.jstor.org/stable/25652309

Gupta, A. (n.d.). *Decision Trees and Deep Networks for Pattern Classification*.

Hernández, A. (1986). El diagnóstico financiero de la empresa: Nuevas tendencias en el análisis. *Revista Española de Financiación y Contabilidad*, *15*(49), 113–132. http://www.jstor.org/stable/42779772

Khaidem, L., Saha, S., & Dey, S. R. (2016). *Predicting the direction of stock market prices using random forest, CoRR*.

Menon, A., Singh, S., & Parekh, H. (2019). A Review of Stock Market Prediction Using Neural Networks. *2019 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN)*, 1–6. https://doi.org/10.1109/ICSCAN.2019.8878682

Moisen, G. G. (2008). Classification and regression trees. *In: Jørgensen, Sven Erik; Fath, Brian D.(Editor-in-Chief). Encyclopedia of Ecology, Volume 1. Oxford, UK: Elsevier. p. 582-588.*, 582–588.

Mouhaffel, A. G., Domínguez, C. M., Arcones, B., Redonda, F. M., & Martín, R. D. (2017). Using multiple regression analysis lineal to predict occupation market work in occupational hazard prevention services. *International Journal of Applied Engineering Research*, *12*(3), 283–288.

Nasseri, A. Al, Tucker, A., & de Cesare, S. (2015). Quantifying StockTwits semantic terms' trading behavior in financial markets: An effective application of decision tree algorithms. *Expert Systems with Applications*, *42*(23), 9192–9210.

https://doi.org/https://doi.org/10.1016/j.eswa.2015.08.008

Nooteboom, B. (2019). Uncertainty and the Economic Need for Trust. In M. Sasaki (Ed.), *Trust in Contemporary Society* (Vol. 42, pp. 60–74). Brill. https://doi.org/10.1163/j.ctvrxk3cr.9

Pereira González, A. (2010). *Análisis predictivo de datos mediante técnicas de regresión estadística*.

San-Martín-Albizuri, N., & Rodríguez-Castellanos, A. (2011). La imprevisibilidad de las crisis: Un análisis empírico sobre los índices de riesgo país. *Innovar*, *21*(39), 161–178. https://revistas.unal.edu.co/index.php/innovar/article/view/35093/35357

Sepúlveda, J. F. D., & Correa, J. C. (2013). Comparación entre árboles de regresión CART y regresión lineal. *Comunicaciones En Estadística*, *6*(2), 175–195.

Shen, S., Jiang, H., & Zhang, T. (2012). Stock market forecasting using machine learning algorithms. *Department of Electrical Engineering, Stanford University, Stanford, CA*, 1–5.

Starmer, J. (19 de August de 2019). Regression Trees, Clearly Explained. US.

Timofeev, R. (2004). Classification and regression trees (CART) theory and applications. *Humboldt University, Berlin*, 1–40.

Tintner, G. (1953). The Definition of Econometrics. *Econometrica*, *21*(1), 31–40. https://doi.org/10.2307/1906941

Zhou, F., Zhang, Q., Sornette, D., & Jiang, L. (2019). Cascading logistic regression onto gradient boosted decision trees for forecasting and trading stock indices. *Applied Soft Computing*,

84, 105747. https://doi.org/https://doi.org/10.1016/j.asoc.2019.105747

Zuniga, C., & Abgar, N. (2011). Breve aproximación a la técnica de árbol de decisiones. *Recuperado de Https://Niefcz. Files. Wordpress. Com/2011/07/Breve-Aproximacion-a-La-Tecnica-de-Árbol-de-Decisiones. Pd.*

Zuur, A. F., Ieno, E. N., & Elphick, C. S. (2010). A protocol for data exploration to avoid common statistical problems. *Methods in Ecology and Evolution*, *1*(1), 3–14.