Analysis of the effectiveness of the ID3 method in the process of massive appraisals in the city of Bogota

Edwin Pérez, Ing[1], Pilar Albacando, Msc[2], Miguel Ávila, Msc[3]
[1]Universidad Distrital Francisco José de Caldas, Colombia, edwinperezc@gmail.com, maavila@udistrital.edu.co
[2] Universidad Distrital Francisco José de Caldas, Colombia, arobles.icg@gmail.com

***Abstract* -- This article presents the general theory that corresponds to machine learning, specifically the decision trees, the application of the ID3 method within the process that constitutes the massive appraisals and an analysis of its effectiveness with respect to the traditional method of linear regression and the observed values; In order to carry out this process we worked with data obtained from the Special Administrative Unit of the District Cadastre (UAECD) and the Cadastral Technical Observatory, corresponding to the Zonal Planning Units 65 Arborizadora, 73 Calandaima and 79 Garcés Navas.**

***Keywords*– Machine learning, decision trees, Cross validation, Percentage Split, ID3, Zonal Planning Unit, Linear Regression, WEKA.**

Digital Object Identifier (DOI):
http://dx.doi.org/10.18687/LACCEI2019.1.1.182
ISBN: 978-0-9993443-6-1 ISSN: 2414-6390

17th LACCEI International Multi-Conference for Engineering, Education, and Technology: "Industry, Innovation, And Infrastructure for Sustainable Cities and Communities", 24-26 July 2019, Jamaica.

Análisis de efectividad del método ID3 en el proceso de avalúos masivos en la ciudad de Bogota

Edwin Pérez, Ing[1], Pilar Albacando, Msc[2], Miguel Ávila, Msc[3]
[1]Universidad Distrital Francisco José de Caldas, Colombia, *edwinperezc@gmail.com*, maavila@udistrital.edu.co
[2] Universidad Distrital Francisco José de Caldas, Colombia, arobles.icg@gmail.com

***Resumen*—** Este artículo presenta la teoría general que corresponde al aprendizaje de máquina, específicamente los árboles de decisión, la aplicación del método ID3 dentro del proceso que constituyen los avalúos masivos y un análisis de su efectividad respecto al método tradicional de regresión lineal y los valores observados; con el fin de llevar a cabo este proceso se trabajó con datos obtenidos de la Unidad Administrativa Especial de Catastro Distrital (UAECD) y el Observatorio Técnico Catastral, correspondientes a las Unidades de Planeamiento Zonal 65 Arborizadora, 73 Calandaima y 79 Garcés Navas.

***Palabras clave*—** Aprendizaje de máquina, árboles de decisión, Cross validation, Percentage Split, ID3, Unidad de Planeamiento Zonal, Regresión Lineal, WEKA.

Abstract: This article presents the general theory that corresponds to machine learning, specifically the decision trees, the application of the ID3 method within the process that constitutes the massive appraisals and an analysis of its effectiveness with respect to the traditional method of linear regression and the observed values; In order to carry out this process we worked with data obtained from the Special Administrative Unit of the District Cadastre (UAECD) and the Cadastral Technical Observatory, corresponding to the Zonal Planning Units 65 Arborizadora, 73 Calandaima and 79 Garcés Navas.

Keywords- Machine learning, decision trees, Cross validation, Percentage Split, ID3, Zonal Planning Unit, Linear Regression, WEKA

I. INTRODUCTION (HEADING 1)

Los avalúos masivos de bienes inmuebles constituyen una importante fase en proyectos públicos y privados de gran impacto que implican la adquisición y/o transformación del territorio, por esta razón dentro del desarrollo de los mismos, es necesario emplear métodos que permitan llegar a valores tan cercanos a los reales que funcionen como base para la toma de decisiones, tradicionalmente en Colombia se ha empleado el método de regresión lineal, que aporta resultados cercanos a los reales, no obstante al explorar otros entornos, se consideró explorar el campo de la inteligencia artificial.

La Inteligencia artificial se define como la ciencia de construir máquinas que hagan cosas que, si las hicieran los humanos requerirían inteligencia [1]. Actualmente se está buscando la forma de simplificar procesos y mejorar resultados, como producto de esta búsqueda está el desarrollo en el campo de la inteligencia artificial, cuyo objetivo principal es entrenar a la máquina mediante diferentes elementos para un proceso específico, luego de lo cual se introducen nuevos datos y se espera que la máquina "aprenda" a realizar los mejores procesos o clasificaciones según el caso [2].
Aunque el estado del arte de este tipo algoritmos aplicados a valuación de inmuebles es escaso por no decir nulo, es posible encontrar algunos referentes como: A Machine-Learning Approach to Automated Knowledge-Base Building for Remote Sensing Image Analysis with GIS Data (Una evaluación de aprendizaje mecanizado de construcción basado en conocimiento automatizado para el análisis de imágenes de sensor remoto con datos SIG) por Xueqiao Huang y John R. Jensen. (1997). En este artículo científico, el objetivo de los autores fue presentar la manera en que se podía aplicar el enfoque de aprendizaje de máquina, para la clasificación de imágenes obtenidas mediante sensores remotos.

En segundo lugar, en el año 2008, se publicó para la Revista Colombiana de Estadística, el artículo titulado Aplicación de árboles de decisión en modelos de riesgo crediticio por Paola Andrea Cardona Hernández. Mediante este trabajo, la autora muestra un marco general de la normatividad del sistema de administración de riesgo crediticio y la importancia del papel de la estadística en estos estudios, específicamente el método de árboles de decisión para el cálculo de incumplimiento en crédito presentando sus ventajas y desventajas. Finalmente, en tercer lugar, en el año 2011 fue presentada en la facultad de minas de Ingeniería de sistemas en la Universidad Nacional, sede Medellín, Colombia, la tesis titulada Modelo Basado en Aprendizaje de Máquinas para el Manejo de Riesgo de Falla Durante la Composición de Servicios Web por Byron Enrique Portilla Rosero como requisito para optar al título de magister en ingeniería de sistemas. El objetivo de este trabajo estuvo en proponer un modelo basado en el método de aprendizaje de máquina que permitiera "aprender al sistema" los riesgos que puede presentar en el servicio web a fin de disminuir el riesgo de falla del mismo.

II. ESPACIALIZACIÓN

Para el desarrollo del proyecto se seleccionaron tres zonas denominadas Unidades de Planeamiento Zonal, estás fueron, la 73, Garcés Navas, 65, Arborizadora y 79, Calandaima, contenidas en la ciudad de Bogotá D.C en la República de Colombia.

A. UPZ 73 Garcés Navas

Ubicada al occidente de la localidad 10 de Engativá cuenta con 557.43 hectáreas comprendidas entre los siguientes límites: Al norte con la Avenida Medellín y la UPZ Bolivia, al sur con la calle 66, el humedal Jaboque y la UPZ Álamos, al Oriente con la Avenida Longitudinal de Occidente y la UPZ Boyacá Real y al Occidente con el río Bogotá, límite del Distrito Capital.

Digital Object Identifier (DOI):
http://dx.doi.org/10.18687/LACCEI2019.1.1.182
ISBN: 978-0-9993443-6-1 ISSN: 2414-6390

La Unidad Garcés Navas cuenta con variedad de usos del suelo, entre los que se encuentran el de vivienda, vivienda con locales comerciales, uso zonal con gran actividad comercial y zonas con usos mixtos (vivienda, comercio, equipamientos) [4] y cuenta con los sectores catastrales presentados en la Figura **1.**

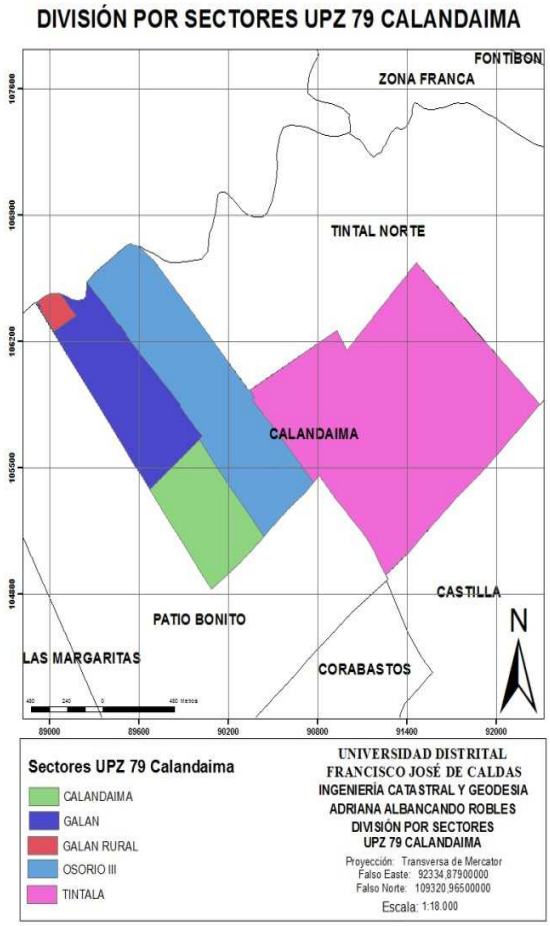

Fig. 1 Sectores UPZ Calandaima.

B. UPZ 65 Arborizadora

Localizada en el nororiente de la localidad 19, Simón Bolívar, contiene un área de 326.97 hectáreas limitadas al norte con la avenida del ferrocarril del sur en límite con la localidad de bosa, al sur con la avenida Villavicencio en límite con la UPZ San Francisco, al oriente con el río Tunjuelo, la localidad de Tunjuelito y la localidad de Kennedy y al occidente con la avenida Villavicencio en límite con las UPZ Ismael Perdomo, Jerusalén y San Francisco, contiene los sectores catastrales que se visualizan en la Figura 2.

Los usos de suelo con los que cuenta la UPZ son los de vivienda, vivienda con zonas de comercio, vivienda con locales comerciales, grandes almacenes y supermercados, Industria y zona para usos mixtos (vivienda, comercio, equipamientos) [5].

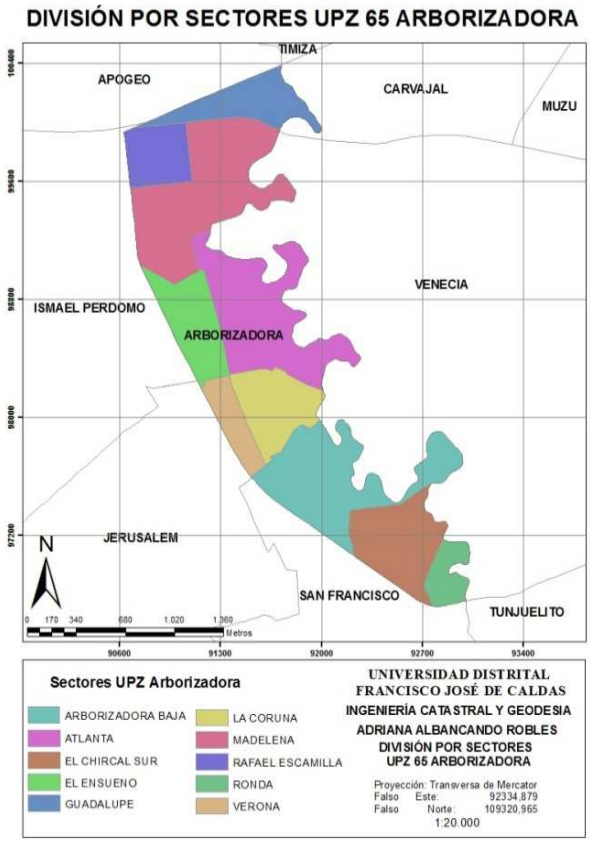

Fig. 2 Sectores UPZ Arborizadora

C. UPZ 79 Calandaima

Localizada en el centro occidente de la localidad 8 de Kennedy tiene una extensión de 319 hectáreas y está limitada al norte por la avenida las Américas en límite con la UPZ Tintal norte, al sur con la avenida las Américas y la avenida de los Muiscas o calle 38 sur en límite con la UPZ Patio Bonito, al oriente con la avenida Ciudad de Cali y la avenida el Tintal en límite con las UPZ Castilla y Patio Bonito y al occidente con el río Bogotá en límite con el municipio de Mosquera [6], contiene los sectores presentados en la Figura 3.

La UPZ Calandaima se encuentra en etapa de desarrollo, razón por la que no cuenta con un decreto que la reglamente, se rige en parte por el Acuerdo 06 de 1990 con Tratamiento Especial de Incorporación al Sector Tintal Central y al Área Suburbana de Expansión mediante el decreto 012 de 1993, considerando el sector de actividad múltiple (Desarrollos urbanísticos residenciales, comerciales, industriales e institucionales) [7].

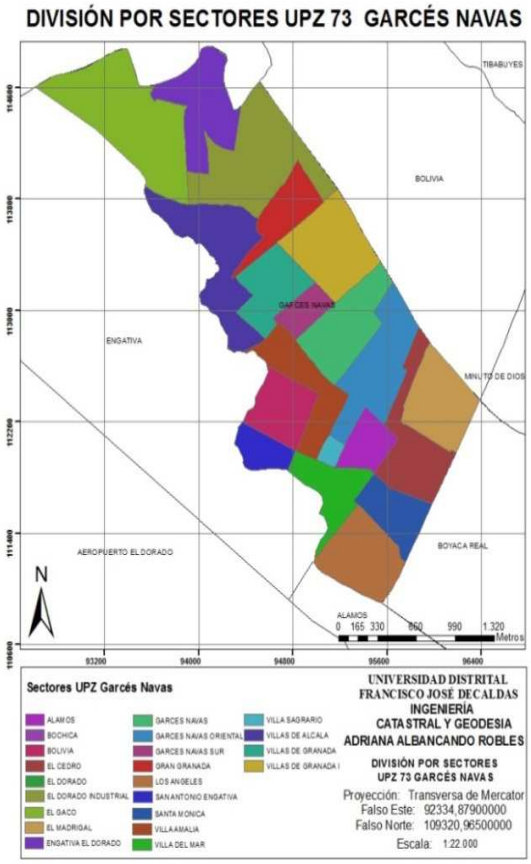

Fig. 3 Sectores UPZ Garcés Navas

III. DEFINICIONES GENERALES

A. Inteligencia Artificial

Según la RAE, Inteligencia es la facultad de conocer, de entender o comprender, por consiguiente, al hablar del término Inteligencia Artificial (IA), se hace referencia a que es un campo de la ciencia y la ingeniería que se ocupa de la creación de artefactos que tengan las facultades mencionadas.

En general la IA se clasifica en las siguientes áreas: Tratamiento de lenguajes naturales, Razonamiento automático – Sistemas expertos, Aprendizaje automático o de máquina, Representación del conocimiento y Visión artificial y robótica [8].

B. Aprendizaje de Máquina

Derivándose de la IA, con el proceso de aprendizaje de máquina, se busca que el objeto sea capaz de deducir, automáticamente y por cuenta propia, una cantidad amplia de consecuencias inmediatas mediante el conocimiento que ya posee [9]. Como se muestra a manera de resumen en la Figura 4, el proceso se completa en dos etapas, en la primera, identificada como Etapa de entrenamiento, el paso inicial es realizado por el componente humano, quienes son los creadores del algoritmo que ejecutará posteriormente la máquina, en segundo lugar se suministran los datos de ejemplo para que sean procesados mediante el algoritmo que generará el aprendizaje y finalmente se obtiene el modelo. La segunda etapa está constituida por cuatro partes, la primera (1) consiste en suministrar, por parte del componente humano un nuevo conjunto de datos, (2) los datos suministrados son procesados por la máquina en el modelo creado y finalmente (3) se genera la información de resultados, es decir, la generada por el ordenador y su validación para ser empleado, posteriormente, como pronóstico [10].

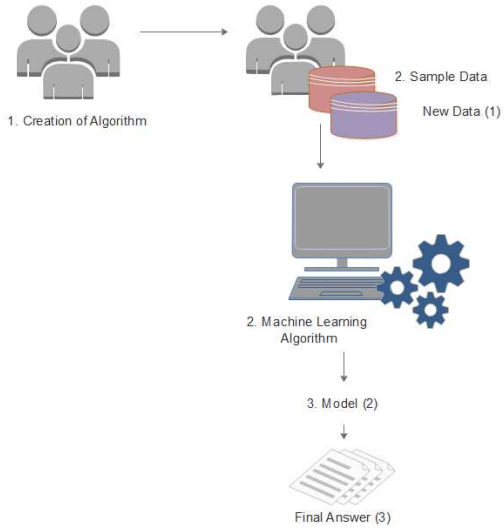

Fig. 4 Proceso del aprendizaje de máquina

C. Árboles de decisión

La técnica de árboles pertenece a los métodos de inferencia inductiva, en donde se deduce información general a partir de información particular [3].

Esta técnica agrupa reglas de forma organizada, en una estructura jerárquica, en donde se llega a la decisión final al seguir las condiciones que establece cada regla desde la raíz hasta las hojas.

Los componentes de los árboles de decisión son: Raíz o nodo inicial (Localizado en la parte superior del árbol, contiene el atributo que da inicio a la clasificación), Ramas (Localizadas al interior del árbol, conectan la raíz con los nodos, los nodos entre sí y finalmente con las hojas), Nodos internos (Localizados al interior del árbol, contienen los atributos que guían la clasificación) y Nodos finales u Hojas (Localizados en los extremos del árbol, contienen las reglas que definen la clasificación final) [11].

Como se visualiza en la Figura 5, el atributo puntaje ubicado en la raíz del árbol es el seleccionado por el algoritmo para dar inicio a la clasificación, en las ramas que se desprenden está la regla de si es menor o igual a 30.5, si cumple con esta condición, se verifica el atributo edad, en el caso de que sea menor o igual a 16.5, se aplicará la regla 1.

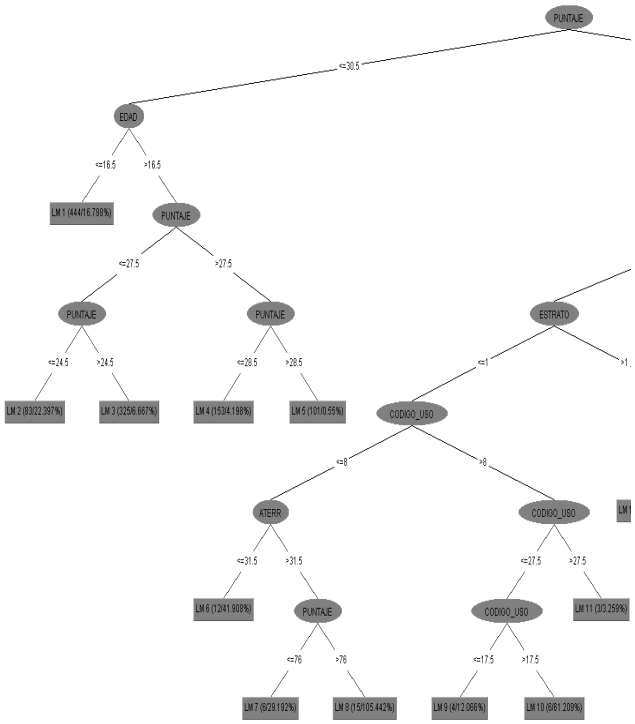

Fig. 5 Ejemplo de árbol de decisión.

D. Entropía de la Información

Conocida también como Entropía de Shannon, mide la incertidumbre de la información suministrada y está dada por la ecuación 1[12].

$$Entropía(S) = -p_+ log_2 p_+ - p_- log_2 p_- \quad \textbf{(1)}$$

En dónde:

$p_+ =$ Promedio de ejemplos positivos en *S*.

$p_- =$ Promedio de ejemplos negativos en *S*.

De forma general se tiene que cuando los ejemplos pertenecen a una misma clase, la entropía es nula, por lo que se muestra una certeza absoluta, en caso contrario, cuando la cantidad de ejemplos negativos es igual a los positivos, la entropía será 1 y en el caso en que los conjuntos sean variables, la entropía estará entre 0 y 1.

Al considerar que el atributo puede tomar diferentes valores, la entropía estará dada por la ecuación 2, en donde p_i es la proporción de S que pertenece a la clase *i*.

$$Entropía(S) = \sum_{i=1}^{c} -p_i log_2 p_i \quad (2)$$

E. Ganancia de la Información

El proceso que implica la creación de árboles de decisión inicia con la selección del atributo que irá en la raíz del árbol por ser el que inicie la clasificación, esta elección se realiza determinando el atributo que más información aporta desarrollando el proceso por medio de la ecuación 3 [9].

$$Ganancia(S,A) = Entropía(S) - \sum_{v \in Valores(A)} \frac{|S_v|}{|S|} Entropía(S_v) \quad (3)$$

En donde:

S= Conjunto de ejemplos

A= Conjunto de los posibles valores para el atributo A

$S_v =$ Subconjunto de *S* para los que el atributo A tiene un valor *v* (Ecuación 4)

$$S_v = \{s \in S \mid A(s) = v\} \quad (4)$$

F. Método ID3

El método ID3 (Iterative Dichotomizer) constituye la versión mejorada de los primeros programas clasificadores por medio de árboles de decisión desarrollados por J. Ross Quinlan [12]. La clasificación realizada mediante el ID3 se realiza al crear de manera automática un árbol de decisión que cumple con las siguientes características: Se crea partiendo de la raíz hacia las hojas, no realiza backtracking, en la fase de entrenamiento emplea únicamente los ejemplos suministrados, para construir el árbol de decisión, el método emplea la ganancia de información, luego de elegir el atributo que ocupará la raíz, elige el que represente más utilidad al modelo en cada paso, dando por terminado el proceso al clasificar el conjunto completo de ejemplos suministrados [3].

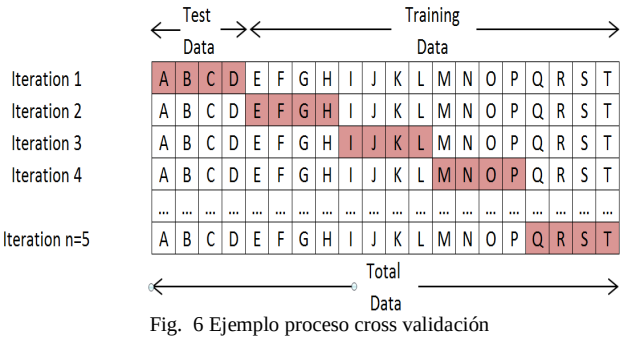

Fig. 6 Ejemplo proceso cross validación

G. Métodos de Validación

1. Cross Validation o Validación cruzada: El usuario elige el número *n* de particiones o *folds*, que será el número en el que se dividirá el conjunto de datos, posteriormente, se construye un clasificador con los *n-1* sub-conjuntos que conformaran el conjunto de datos de entrenamiento y los restantes serán los datos de validación, el proceso se repetirá hasta que el total de los datos hayan sido de entrenamiento y validación en las respectivas iteraciones, tal como se muestra en el ejemplo de la figura 6 para un conjunto de *20* datos y *n=5* [13].
2.

2. Percentage Split: El usuario elije el porcentaje de datos que será empleado para el entrenamiento y el porcentaje restante será

empleado para la validación del modelo [14].

3. Supplied test set: El ususrio suministra un conjunto de datos para el entrenamiento y un conjunto distinto para la validación del modelo [15].

H. Errores

Con el fin de verificar la efectividad del empleo del método ID3, se tuvieron en cuenta lo errores relativos que se presentan en las ecuaciones 5 y 6, dado que permiten realizar comparaciones entre modelos con errores medidos en diferentes unidades [16].

Error absoluto relativo (Relative Absolute Error):

$$RAE = \frac{\sum_{i=1}^{N}|\hat{\theta}_i - \theta_i|}{\sum_{i=1}^{N}|\bar{\theta}_i - \theta_i|} \quad (5)$$

Raíz del error cuadrático relativo (Root Relative Squered Error):

$$RRSE = \sqrt{\frac{\sum_{i=1}^{N}(\hat{\theta}_i - \theta_i)^2}{\sum_{i=1}^{N}(\bar{\theta}_i - \theta_i)^2}} \quad (6)$$

J. Avalúos Masivos

Son el conjunto de procesos que permiten extrapolar información de predios particulares a un gran conjunto de diferentes predios, teniendo como base valores de avalúos físicos. Para cumplir con este proceso se deben cumplir cuatro etapas que son: Identificación predial, determinación de zonas homogéneas físicas y geoeconómicas, determinación de valores unitarios para los diferentes tipos de edificaciones y finalmente la liquidación de los avalúos [19].

I. Software WEKA(Waikato Enviroment for Knowledge Analysis)

Software libre desarrollado por la Universidad de Waikato en Nueva Zelanda, permite hacer uso de los paquetes que contiene con datos suministrados por el usuario, el entorno de mayor utilidad es el denominado Explorer, en el que se puede contar con herramientas como Clasiffy, que permite el acceso a algoritmos de regresión y clasificación, Cluster, con la función de mostrar los diferentes métodos de agrupación para los datos, Associate, Select Atributes y Visualize [17]. Como se muestra en la parte inferior de la figura 7, el paso final consistió en crear tres grupos de datos derivados del conjunto de PH y tres del conjunto de No PH. El conjunto de pruebas finales consta de 20 datos extraídos de forma aleatoria con el fin de conservarlo para posteriores experimentos o pruebas necesarias al finalizar todo el proceso, el segundo grupo, contenedor de los datos de entrenamiento y validación, consta del 95% de los datos que quedaron luego de apartar los de pruebas finales y por último, el conjunto de datos para pronóstico consta del 5%. Este procedimiento se siguió con los conjuntos de datos de las 3 Unidades de Planeamiento Zonal obteniendo, al final de esta etapa, grupos conformados como se presenta en la tabla I.

Los archivos que se ingresan al software son tipo *.arff* y la estructura que deben tener para el posterior procesamiento es la siguiente: Encabezado: *@relation* + nombre que identifica el conjunto de datos.

TABLA I
CANTIDADES DE DATOS A PROCESAR POR UPZ

UPZ	Garcés N.	Arborizadora	Calandaima
PH Entrenamiento y Validación	14456	9438	28824
PH Pronóstico	761	497	1517
NO PH Entrenamiento y Validación	17875	6423	2636
NO PH Pronóstico	941	338	139

Luego de tener los conjuntos de datos listos, se renombraron las clases de los atributos que estaban expresados numéricamente, con el fin de que la base de datos quedara presentada nominalmente en su totalidad cumpliendo con los requerimientos del método ID3. Los atributos empleados para el experimento se presentan en la tabla II.

TABLA II
ATRIBUTOS PARA EL EXPERIMENTO POR UPZ

UPZ	Garcés N.	Arborizadora	Calandaima
PH	Sector Edad Puntaje Estrato Actividad Tratamiento Área construida Valor m² de construcción	Sector Uso Edad Puntaje Estrato Actividad Tratamiento Área construida Valor m² de construcción	Sector Uso Pisos Edad Puntaje Estrato Actividad Tratamiento Área construida Valor m² de construcción
NO PH	Sector Pisos Edad Puntaje Estrato Actividad Tratamiento Área de terreno Valor m² de Terreno Área de construcción Valor m² de construcción	Sector Uso Edad Puntaje Estrato Actividad Área construida Valor m² de construcción	Sector Uso Pisos Edad Puntaje Estrato Actividad Tratamiento Área terreno Valor m² terreno Área construida Valor m² de construcción

Al realizar las respectivas comparaciones y verificación de los resultados, valores como las instancias clasificadas correctamente y estadísticos como kappa y la media del error absoluto fueron la base para realizar lo que se menciona como selección 1 y 2, para el caso del método ID3, mientras que en el segundo escenario, al realizar la selección 1 y 2 del método de regresión lineal, los elementos que se tuvieron en cuenta fueron el coeficiente de correlación y los valores dados, para este experimento, en unidades monetarias, la media del error absoluto y la raíz del error medio cuadrático

Declaración de atributos nominales: *@attribute* + {nombre de los atributos separados por comas}

Datos: *@data* + valores de los atributos en el orden en el que fueron declarados y separados por comas a partir de la siguiente fila.

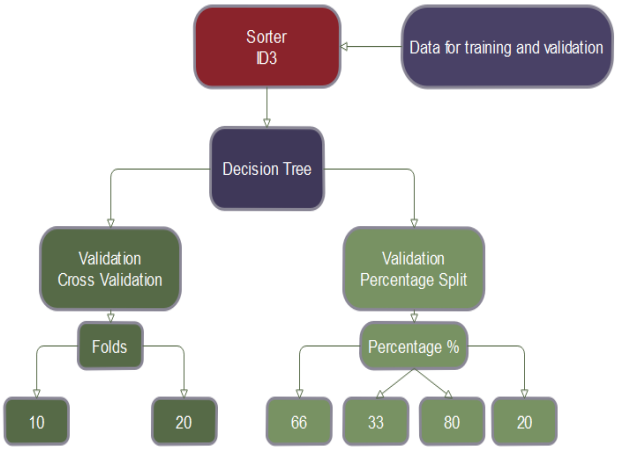

Fig. 8 Etapa de entrenamiento y validación

Al realizar las respectivas comparaciones y verificación de los resultados, valores como las instancias clasificadas correctamente y estadísticos como kappa y la media del error absoluto fueron la base para realizar lo que se menciona como selección 1 y 2, para el caso del método ID3, mientras que en el segundo escenario, al realizar la selección 1 y 2 del método de regresión lineal, los elementos que se tuvieron en cuenta fueron el coeficiente de correlación y los valores dados, para este experimento, en unidades monetarias, la media del error absoluto y la raíz del error medio cuadrático.

IV. METODOLOGÍA

El proceso por el que se llevó a cabo el experimento constó de cinco etapas, organización de los datos, entrenamiento y validación del modelo, generación de pronósticos, procedimiento con el método de regresión lineal y análisis comparativo.

A. Organización de Datos

Con el fin de realizar el procesamiento, el primer paso fue organizar las bases de datos proporcionadas por la Unidad Especial de Catastro, esta etapa se desarrolló hasta llegar a dejar 6 conjuntos de datos, proceso que se llevó a cabo como se presenta en la figura 7.

En primer lugar, se dejó únicamente la información que se emplearía en el experimento, seguido a esto, se separaron los datos que correspondían a predios bajo el reglamento de Propiedad Horizontal (PH) y los que no correspondían a este (No PH), debido a que sus características son diferentes.

B. Entrenamiento y Validación

El experimento se organizó de la forma que se presenta en la figura 8, procesando los datos en el software WEKA, mediante el clasificador ID3, luego de generar un árbol de decisión, se procedió a realizar la validación del modelo mediante los métodos Cross validation, con 10 y con 20 particiones y Percentage Split, con porcentajes de 66, 33, 80 y 20, obteniendo los resultados presentados en la tabla III por cada validación del experimento, información útil para verificar la efectividad del clasificador.

TABLA III
RESULTADOS OBTENIDOS POR MEDIO DEL CLASIFICADOR ID3

	Resultados Obtenidos
Clasificador ID3	Instancias Cantidad de atributos Instancias clasificadas correctamente Instancias clasificadas incorrectamente Estadístico Kappa Media del error absoluto Raíz del error medio cuadrático Error absoluto relativo Raíz del error cuadrático relativo Instancias sin clasificar

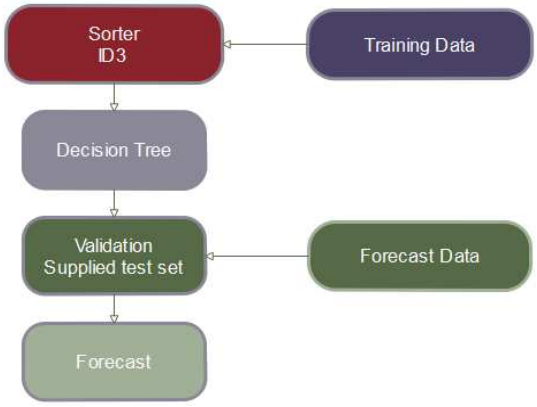

Fig. 9 Etapa de pronóstico

C. Etapa de Pronóstico

Esta etapa consistió en emplear el método de validación Supplied test set, por permitir validar el modelo con datos diferentes a los que se emplearon para el entrenamiento y generación del árbol de decisión, siguiendo el proceso presentado en la figura 9, se generaron, de esta forma, los resultados para verificar la efectividad del método ID3 luego de tener el modelo elaborado.

D. Proceso con método de Regresión Lineal

El proceso que se siguió luego de emplear el método de árboles de decisión ID3, fue emplear los mismos datos y procesarlos por medio del método convencional de regresión lineal, el procedimiento fue semejante, empleando las mismas variaciones de los métodos de validación de Cross Validation, Percentage Split y el método para realizar los respectivos pronósticos Supplied test set, los resultados obtenidos mediante este proceso son los presentados en la tabla IV.

TABLA IV
RESULTADOS OBTENIDOS POR MEDIO DEL MÉTODO DE REGRESIÓN LINEAL

	Resultados Obtenidos
Método de Regresión Lineal	Instancias, Cantidad de atributos RAE, RSSE Coeficiente de correlación Media del error absoluto ($) Raíz del error medio cuadrático ($)

7

Finalmente, se cotejaron los errores absolutos relativos y las raíces de los errores cuadráticos, datos que permiten realizar la comparación entre modelos de ID3 y regresión lineal. Los resultados comparativos se presentan a en las tablas V, VI, VIII.

TABLA V
COMPARACIÓN DE ERRORES BASE DE DATOS UPZ GARCÉS NAVAS NO PH

EXPERIMENT	RAE (%)	RRSE (%)
TRA- VAL (ID3 CROSS 20)	30.16	66.27
TRA- VAL (LR CROSS 20)	51.59	62.31
TRA- VAL (ID3 SPLIT 80)	29.76	65.91
TRA- VAL (LR SPLIT 66)	51.95	62.81
FORECAST (ID3 SUPPLIED)	31.84	67.83
FORECAST (LR SUPPLIED)	52.62	62.64

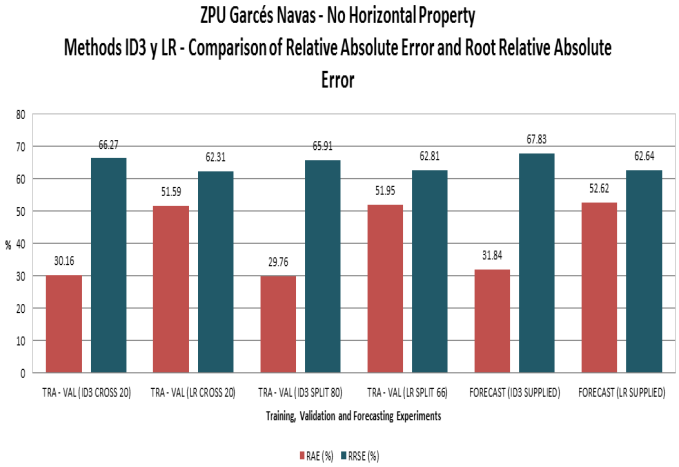

Fig. 10 Comparación de errores base de datos UPZ GARCÉS NAVAS no ph

TABLA VI
COMPARACIÓN DE ERRORES BASE DE DATOS UPZ ARBORIZADORA PH

EXPERIMENT	RAE (%)	RRSE (%)
TRA- VAL (ID3 CROSS 20)	25.83	52.02
TRA- VAL (LR CROSS 20)	79.84	78.3
TRA- VAL (ID3 SPLIT 80)	25.73	51.68
TRA- VAL (LR SPLIT 80)	81.51	79.03
FORECAST (ID3 SUPPLIED)	24.5	49.38
FORECAST (LR SUPPLIED)	86	81.16

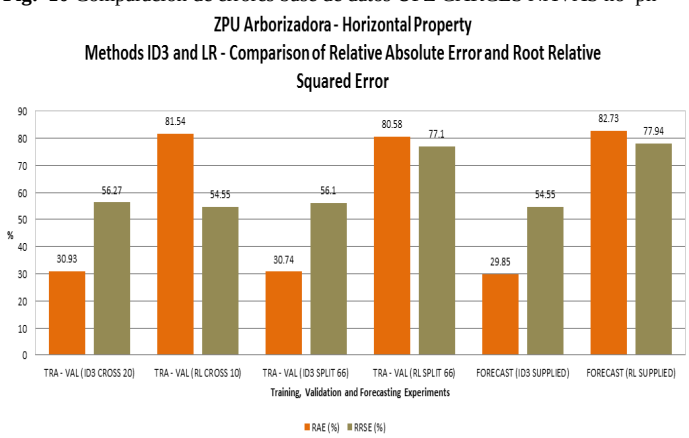

Fig. 11 Comparación de errores base de datos UPZ ARBORIZADORA PH

TABLA VII
COMPARACIÓN DE ERRORES BASE DE DATOS UPZ ARBORIZADORA NO PH

EXPERIMENT	RAE (%)	RRSE (%)
TRA- VAL (ID3 CROSS 20)	38.6	73.82
TRA- VAL (RL CROSS 10)	67.43	72.28
TRA- VAL (ID3 SPLIT 66)	39.42	73.79
TRA- VAL (RL SPLIT 33)	67.01	72.03
FORECAST (ID3 SUPPLIED)	42.05	78.9
FORECAST (RL SUPPLIED)	71.58	97.4

Las gráficas respectivas a cada tabla se relacionan a continuación

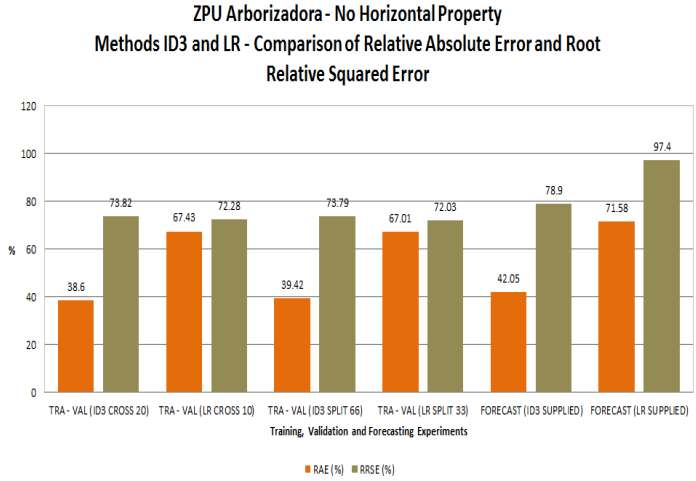

Fig. 12 Comparación de errores base de datos UPZ ARBORIZADORA NO PH

Como se puede observar en las gráficas de los diferentes conjuntos de datos, en general, los valores del error absoluto relativo generados del procesamiento del método ID3 son menores en cantidades numéricas representativas, la mayoría entre 20 y 50 puntos porcentuales a comparación de los valores generados de este error con el método de regresión lineal, en cuanto a la raíz del error cuadrático relativo, las diferencias entre los valores obtenidos por el método ID3 y el de regresión lineal son menores, desde 2 hasta 30 puntos porcentuales, en la mayoría de los casos, siendo menor el obtenido por los árboles de decisión y en algunos otros el método de regresión lineal.

VI. CONCLUSIONES

Por medio del experimento descrito se puede concluir que el método ID3 de árboles de decisión es una herramienta de manejo asequible en los trabajos de ingeniería y procesamiento de datos como los que implican el desarrollo de los avalúos masivos.

La etapa de mayor dedicación, en términos de tiempo, fue la de organización de los datos, el trabajo de clasificar los datos y renombrarlos como etapa anterior al procesamiento, en contraste con la cantidad de tiempo dedicada a procesar los datos y adquirir resultados para las posteriores comparaciones.

En general, por los resultados que se obtuvieron del experimento con las seis bases de datos de las tres UPZ y las divisiones de PH y No PH de cada una de estas, con cantidades significativas de datos, se concluye que el método ID3 es una herramienta útil y efectiva en el proceso que implican los avalúos masivos.

Finalmente, entender el procesamiento y hacer un seguimiento a la clasificación que se realiza por medio de los árboles de decisión, es una facilidad que presenta el método ID3 al emplear el algoritmo para la generación de pronósticos según la necesidad de los usuarios.

Cabe resaltar los siguientes resultados cuantitativos para dos de las variaciones del experimento que sobresalen por la bondad de sus resultados a saber:

- ID3 Cross 10

El experimento ID3 Cross 10 de entrenamiento y validación se realizó con el método de clasificación ID3 y validó mediante el método Cross Validation con 10 particiones (folds), seleccionadas aleatoriamente del conjunto de datos de PH con valores nominales.
Este experimento dio como resultado 2197 instancias clasificadas correctamente, un 83,346% de las instancias, siendo un porcentaje significativo en comparación con las 371 que se clasificaron incorrectamente y corresponden a un 14,0744% del total de las instancias.
El estadístico Kappa obtuvo un valor de 0,6902, que permite considerar que el experimento obtuvo un grado de concordancia aceptable.
El error absoluto medio tuvo un valor de 0,0425 que permite deducir que los resultados obtenidos son buenos y la raíz del error medio cuadrático con un valor de 0,1741 mayor que el anterior, deja ver como este error castiga fuertemente el hecho de que el 14,0744% de las instancias no hayan sido clasificadas correctamente.
Finalmente, el error absoluto relativo, con un 35,8212% es aceptable, por dejar un 64,1788% confiable, y por su lado, la raíz del error cuadrático relativo, aunque fue grande, con un valor de 72,0379%, fue menor que los métodos con los que se le comparó (Anexo 70).
Es de considerar también el hecho de que en este método quedaron 68 instancias sin clasificar, las cuales equivales a un 2,5797% del conjunto total de datos.

- ID3 Supplied

El experimento ID3 Supplied de pronóstico se realizó por medio del método ID3 para el entrenamiento con un conjunto de 2636 y el método Supplied test set para el pronóstico con un conjunto de 139 datos, ambos conjuntos compuestos de valores nominales.
Este experimento de pronóstico dio como resultado 115 instancias clasificadas correctamente, es decir un 82,7338% de las instancias, siendo un porcentaje significativo en comparación con las 21 que se clasificaron incorrectamente y corresponden a un 15,1079% del total de las instancias.

El estadístico Kappa por estar relativamente cercano a 1 con un valor de 0.6753, permite considerar que el experimento obtuvo un grado aceptable de concordancia al ser evaluado con 139 datos.
El error absoluto medio con un valor de 0,044, permite deducir que los resultados obtenidos son buenos y la raíz del error medio cuadrático con un valor de 0,1734 mayor que el anterior, deja ver como este error castiga fuertemente el hecho de que el 15,1079% de las instancias no hayan sido clasificadas correctamente. Finalmente, el error absoluto relativo, con un 36,4521% es aceptable, por dejar un 63,5479% confiable, y la raíz del error cuadrático relativo, con un valor de 70,5714%. Muestra un valor elevado de error.

REFERENCIAS

[1] Cazorla, M, Alfonso, M, Escolano, F, Colomina, O y Lozano, M, Inteligencia Artificial, Modelos, Técnicas y Áreas de aplicación, Alicante: Paraninfo, S.A, 2003.

[2] Murphy, K, Machine Learning A Probabilistic Perspective, Cambridge, Massachusetts, 2012.

[3] Mitchell, T, Machine Learning, Portland: Mc Hill, 1997.

[4] Secretaría Distrital de Planeación, «21 Monog de las localidades: Diagnóstico de los asp físicos, demográficos y socioeconómicos d localidades – 2011. # 19 Ciudad Bolívar,» 201 Bogotá.

[5] Alcaldía Mayor de Bogotá D.C , «UPZ 65 Acu para construir ciudad,» Bogotá, Oficina aseso prensa y comunicaciones - Secretaría Distrit Planeación, 2008.

[6] Secretaría Distrital de planeación, «Conocien

localidad de Kennedy. diagnóstico de los aspectos físicos, demográficos y socioeconómicos,» Bogotá, 2009.

[7] Alcaldía Mayor de Bogotá, «Decreto 12 de 1993,» Bogotá D.C.

[8] Pino, R, Gómez, A y de Abajo, N, «Introducción a la Inteligencia Artificial: Sistemas Expertos, Redes Neuronales Artificiales, y computación evolutiva,» Asturias, Servicios y publicaciones de la Universidad de Oviedo, 2001.

[9] McCarthy, J, Mechanisation of Thought Processes., Simposio N.10 Volumen I ed., Londres, 24 - 27 de Noviembre de 1958.

[10] García, A, Inteligencia Artificial. Fundamentos, práctica y aplicaciones, Madrid: RC Libros, 2012.

[11] Vizcaino, P. A, Aplicación de técnicas de inducción de árboles de decisión a problemas de clasificación mediante el uso de WEKA, Bogotá.

[12] Sancho, F, «Departamento de Ciencias de la Computación e Inteligencia Artificial, Universidad de Sevilla,» 10 Diciembre 2016. [En línea]. Available: http://www.cs.us.es/~fsancho/?e=104. [Último acceso: 14 Abril 2017].

[13] Corso, C, Aplicación de algorítmos de clasificación supervisada, Buenos Aires: Universidad Tecnológica Nacional, 2009.

[14] García, F, Aplicación de técnicas de minería de datos a datos obtenidos por el Centro Andaluz de Medio Ambiente (CEAMA), Granada: Universidad de Granada, 2013.

[15] Hernández, J, Práctica de minería de datos, Introducción a WEKA, Valencia: Universidad Politécnica de Valencia, 2006.

[16] Mood, A, Graybill, F y Boes, D, Introduction to the Theory of Statistics, Auckland: McGraw Hill, 1974.

[17] Morate, D, mabual de WEKA, Granada, 2000.

[18] Cuevas, A, Teoría de la Información, Codificación y Lenguajes, Madrid: Servicio del Ministerio de Educación y Ciencia, 1975.

[19] Instituto Geográfico Agustín Codazzi (IGAC), Resolución 620 de 2008, Bogotá, 2008.

[20] Secretaría Distrital de Planeación , «21 Monografías de las Localidades: Diagnóstico de los aspectos físicos, demográficos y socioeconómicos,» Bogotá, 2011.