Predicción del avalúo catastral de los predios urbanos en la ciudad de Pereira mediante aprendizaje de máquina

Autores:

Walter R. Osorio González

Alexander Rozo Patiño

Trabajo de grado para optar al título de Magíster en Ingeniería de Sistemas y Computación

Director: Mauricio Alexander Álvarez, PhD

Universidad Tecnológica de Pereira

Facultad de Ciencias Básicas

Maestría en Ingeniería de Sistemas y Computación

Pereira

2017

Nota de Aceptación

_____

Director del programa

_____

Director Tesis

_____

Jurado

_____

Jurado

Ciudad: Pereira
Fecha:

A mi padres, hermanos, hermana y a mi esposa, especialmente a mi madre y hermano que me acompañan desde el cielo.

Walter R. Osorio González

A mi madre, a mí amada esposa Tatiana y nuestra amada hija Violeta.

Alexander Rozo Patiño

Agradecimientos

A Dios

A la Universidad Tecnológica de Pereira por ser la institución que ha permitido mi formación profesional

Al Dr. Mauricio Álvarez por su acompañamiento, asesoría y disponibilidad durante todo el desarrollo del proyecto

Walter R. Osorio González

A Dios y a la Universidad Tecnológica de Pereira, mi alma mater y a todo los profesores y amigos que me han acompañado en este proceso de aprendizaje continuo.

Al profesor Mauricio Álvarez por su esmerado esfuerzo y paciencia en hacer posible este trabajo.

Alexander Rozo Patiño

Tabla de contenido

Resumen

El propósito de este estudio es desarrollar e implementar un modelo que estime el avalúo catastral de un predio urbano en la ciudad de Pereira mediante técnicas de aprendizaje de máquina.

Disponer del avalúo catastral de predios es fundamental para la economía de los municipios con el fin de aplicar el impuesto predial y para los propietarios de los predios para tener información del valor de su patrimonio. Se fortalece la comercialización de los predios de una forma más equitativa en negociaciones entre personas y entre personas y estado.

Debe de existir una forma o técnica que le permita dar un valor catastral a los predios, esta se debe poder estandarizar y replicar para ser utilizada donde se necesite. Actualmente este proceso lo realiza el IGAC de una forma manual y poco eficiente que requiere de expertos y toma de muestras del observatorio inmobiliario para definir el valor de las zonas geoeconómicas y el valor de los tipos de edificaciones, el realizar la estimación de forma automática permite que el proceso sea más eficiente y que se pueda realizar en un menor tiempo, lo que reduce costos y permite tener información actualizada más rápidamente. Este medio incluye un mayor número de variables del territorio y sin la necesidad de ajustarse al criterio de las personas haciendo más transparente el proceso de valoración predial.

Para lograr este objetivo se realizó una comparación de las principales técnicas de tipo supervisado en el ámbito de aprendizaje automático: procesos gaussianos, máquinas de soporte vectorial, redes neuronales y regresión lineal múltiple. La comparación se realizó mediante validación cruzada anidada. Se tomaron cinco grupos con 80% de la información para entrenamiento y 20% para validación. Finalmente se determinó el mejor algoritmo mediante la comparación de los resultados del error cuadrático medio (ECM), el error absoluto medio (EAM) y el coeficiente de determinación (r^2).

Para realizar lo anterior se partió de la base de 73.078 registros y 82 variables, registrados en el censo catastral año 2013. Estos corresponden a la totalidad de los predios urbanos de la ciudad de Pereira. Se realizó un análisis exploratorio de datos, depuración de variables y datos atípicos.

En la selección del modelo se realizaron dos experimentos. En el primer experimento se utilizó la técnica de validación cruzada anidada y se optó por el modelo con mejor promedio del error y en el segundo experimento se utilizó el mejor modelo de cada algoritmo del experimento uno y se realizó la estimación con el resto de datos que no fueron utilizaron en la fase de entrenamiento.

En el primer experimento el mejor algoritmo para estimación del avalúo catastral fue el de procesos gaussianos con un ECM de 0,020, un EAM de 0,085, este último equivale a 84.847 pesos colombianos. (Valor por metro cuadrado) y un r^2 del 86%.

En el segundo experimento el algoritmo con mejores resultados fue el de redes neuronales con un ECM 0,041, un EAM 0,140 que equivale 140.155 pesos colombianos y un r^2 de 72 %.

En comparación con otros estudios se ratificó que la técnicas de aprendizaje de máquina son mejores que los modelos de estimación tradicional y finalmente, al realizar la selección de características con la técnica de Sequential Forward Selection (SFS), se mantuvo la coherencia de los errores cuadráticos medios aunque al obtener los resultados no se percibió una mejora sustancial en el error de estimación. Por este motivo se determinó utilizar todas las variables y aplicar este en el desarrollo del software para estimación del avalúo catastral.

Palabras clave: Avalúo catastral, aprendizaje de máquina supervisado, procesos gaussianos, máquinas de soporte vectorial, redes neuronales, regresión lineal múltiple, validación cruzada anidada.

Abstract

The purpose of this study is to develop and implement a model that estimates the housing price of an urban property in the city of Pereira using machine learning techniques.

Having the cadastral value is fundamental for economic of the municipalities in order to apply the property tax and for the landowners to have information on the value of their property. It strengthens the commercialization of land in a more equitable way in negotiations between people and between people and state.

There must be a form or technique that allows it to give a cadastral value, it must be able to standardize and replicate to be used where it is needed, currently this process is performed by the IGAC in a manual and inefficient way that requires experts and sampling of the real estate observatory to define the value of the geo-economic zones and the value of the types of buildings, making the estimate automatically allows the process to be more efficient and can be done in a shorter time, which reduces costs and allows to have updated information more quickly, including a greater number of variables of the territory and without the need to adjust to the criterion of the people making the property valuation process more transparent.

To this goal, was used a comparison of the main supervised type techniques in the field of machine learning: Gaussian processes, vector support machines, neural networks and multiple linear regression. The comparison was performed using nested cross validation. Five groups were taken with 80% of the information for training and 20% for validation. Finally, the best algorithm was determined by comparing the results of mean square error (ECM), mean absolute error (EAM) and coefficient of determination (r^2).

To do this, we started with the base of 73,078 records and 82 variables, registered in the cadastral updating census of the year 2013. These correspond to all of the urban properties in the city of Pereira. An exploratory data analysis, variable debugging and atypical data were performed.

In the selection of the model two experiments were performed. In the first experiment we used the technique of nested cross validation and we opted for the model with better average of the error and in the second experiment using the best model of each algorithm of the experiment one was made the estimation with the rest of data that were not used in the training phase.

In the first experiment, the best algorithm for estimation of the cadastral valuation was that of Gaussian processes with an ECM of 0,020, an AMS of 0,085, equivalent to 84,847 Colombian pesos. (value per square meter) and a r^2 of 86%.

In the second experiment, the algorithm with the best results was the neural networks with a 0,041 ECM, a 0,140 EAM that equals 140,155 Colombian pesos and a r^2 of 72%.

Compared with other studies it was confirmed that machine learning techniques are better than the traditional estimation models and finally, when performing the selection of characteristics with the

technique of Sequential Forward Selection (SFS), the consistency of the quadratic errors means although obtaining the results did not perceive a substantial improvement in the estimation error, so it was determined to use all variables and apply this in the cadastral software.

Key words: Cadastral evaluation, supervised machine learning, gaussian processes, vector support machines, neural networks, multiple linear regression, nested cross validation

1. Introducción

La presente investigación busca desarrollar un modelo que permita solucionar el problema de estimar de manera automática el avalúo catastral de los predios en la ciudad de Pereira, a partir de una información base de los avalúos con sus respectivas características.

En Colombia el ente encargado de la valoración catastral es el Instituto Geográfico Agustín Codazzi (IGAC, 2017).

El IGAC es el encargado de elaborar el catastro nacional de la propiedad inmueble y de realizar el inventario de las características de los suelos. Con la información se realiza el proceso de valoración del predio que incluye las siguientes etapas: identificación predial, determinación de las zonas homogéneas y geoeconómicas, determinación de valores unitarios para los tipos de las edificaciones y por último la liquidación de avalúos.

Esta metodología de valoración masiva por zonas homogéneas fue diseñada por el IGAC en el año 1984, la cual para la valorización de las construcciones se realiza a través de modelos de regresión simples ya sean lineales, exponenciales, logarítmicos o potenciales (Fajardo, 2014).

En la actualidad se han realizado diferentes estudios en varios países con el fin de comparar los métodos tradicionales versus las técnicas de aprendizaje de datos en lo referente a la estimación del avalúo catastral o en el valor comercial de la vivienda, en la mayoría de los estudios expuestos en este trabajo los métodos que han dado mejores resultados son las técnicas de aprendizaje de datos.

El propósito de este trabajo es desarrollar un modelo e implementar un software que permita estimar el avalúo catastral de los predios urbanos de la ciudad de Pereira mediante inteligencia artificial en el campo de aprendizaje de máquina. Para lo anterior se hace necesario determinar cuál es el mejor algoritmo teniendo en cuenta las características de los predios urbanos en la ciudad.

Se partió de una base de datos catastral a la cual se le realizó un análisis exploratorio de datos y limpieza de la información. La base de datos es de la vigencia 2013 en el municipio de Pereira con un total de 73.078 predios urbanos y 82 variables de caracterización. Con el conjunto de datos se realizó la identificación de relaciones entre las variables de análisis, seguido de la comparación de los resultados con las métricas de las siguientes técnicas de aprendizaje de máquina supervisado: máquinas de soporte vectorial (MSV), procesos gaussianos (PG), redes neuronales artificiales (RNA) y con el modelo tradicional de regresión lineal múltiple (RLM). Para realizar las comparaciones se utiliza la técnica de validación cruzada anidada.

En su contenido el documento presenta inicialmente la justificación y varios estudios realizados en la estimación de valoración de vivienda en diferentes países, seguido de una contextualización de las herramientas y métodos utilizados descritos en el marco teórico. Después se describe la metodología aplicada para el estudio comparativo y la selección del modelo, además se presentan los resultados obtenidos del análisis. Finalmente se presentan las conclusiones y recomendaciones para trabajos futuros en este ámbito.

2. Definición de Problema

En la ciudad de Pereira el cálculo del avalúo catastral es realizado por el IGAC (Instituto Geográfico Agustín Codazzi). El proceso se desarrolla en las siguientes etapas: primero la identificación predial, segundo determinación de las zonas homogéneas geoeconómicas, tercero la determinación de los valores unitarios para los tipos de las edificaciones y por último la liquidación de avalúos.

En la identificación predial se recolecta información física y jurídica del predio, son analizadas las áreas del terreno, construcciones que luego son calificadas por el experto y se determina el destino económico del predio. En la calificación se tienen en cuenta el tipo de estructura, los acabados principales y el estado de conservación. A cada una de esas construcciones se le asigna un puntaje.

Para la creación de las zonas homogéneas geoeconómicas se parte de las zonas homogéneas físicas, y para las zonas homogéneas urbanas las variables que se tienen en cuenta son la reglamentación del uso del suelo, uso actual del suelo, vías, topografía, servicios públicos y tipificación de las construcciones. En las zonas rurales las variables utilizadas son: áreas homogéneas de tierra, disponibilidad de aguas superficiales permanentes, influencia de las vías, reglamentación del uso del suelo rural y uso actual del suelo. (Fajardo, 2014).

Desde las zonas homogéneas físicas se crean los puntos para hacer la investigación económica. A cada punto se le hace el avalúo comercial, luego se hace el cálculo del valor unitario por punto. Una vez realizada la investigación directa e indirecta del mercado inmobiliario la información obtenida se analiza estadísticamente y es realizada una depuración para obtener valores comerciales por cada punto. Al finalizar el proceso se obtiene el cálculo del valor unitario de terreno para cada uno de los puntos de investigación.

Una vez adoptados en forma preliminar los valores unitarios de terreno por punto, se ordenan los puntos de investigación por rangos de valores de mayor a menor y se asigna la numeración de zonas partiendo de uno en orden ascendente. El plano de zonas homogéneas geoeconómicas representa la división de la zona del municipio de acuerdo con el valor unitario de terreno definido para calcular el avalúo catastral de los predios del municipio, el cual contiene información del valor por metro cuadrado/hectárea de mayor a menor.

Para determinar el valor unitario, los factores que inciden en los precios de las construcciones y edificaciones son los materiales y su calidad, las condiciones urbanísticas y arquitectónicas, el uso de la construcción y/o edificación, la edad o vetustez y la ubicación según clasificación catastral.

Para determinar el avalúo catastral de las construcciones, se ha implementado una metodología basada en tablas de valores unitarios relacionados con la calificación de cada unidad de construcción y determinados mediante el cálculo de regresiones con base en datos obtenidos de la práctica de avalúos individuales a una muestra representativa de las construcciones del municipio. La liquidación de avalúos se realiza calculando el área del terreno en cada una de las zonas homogéneas geoeconómicas que le corresponden por su ubicación geográfica al intersecar con la capa de zonas homogéneas geoeconómicas y aplicando los valores por metro cuadrado asignados a cada zona. A este valor se le suma el valor de las construcciones al cual se le aplican a las tablas

de construcciones según corresponda al puntaje que se le asignó a la construcción. (Fajardo, 2014, pág. 32-36).

Carecer de una correcta valoración de un inmueble, acorde con las características que influyen en mayor medida en el avalúo, puede producir inconformidad en los ciudadanos dado el impuesto generado a partir del valor. Como ejemplo en el año 2016 en Colombia no se realizó un método asertivo en la valoración de los vehículos para efectos de liquidación y el pago de impuestos del año 2016. El problema tuvo que ver con que el valor estimado superaba el valor comercial. Esta situación generó una fuerte polémica e inconformidad de los ciudadanos (EL TIEMPO, 2016).

El patrimonio inmobiliario tiene una importancia fundamental en la vida económica de cualquier país. Como expresión de riqueza presupone una determinada capacidad económica fácil de cuantificar y difícil de ocultar, por lo tanto avaluar una propiedad se convierte en una necesidad.

La metodología de valoración masiva de predios fue diseñada por el IGAC en el año 1984, actualmente tiene 33 años de estar siendo aplicada en el país. Cuando se elaboró la metodología predominaban las casas individuales y cada una tenía terrenos propios y no estaban sometidos al régimen de propiedad horizontal. Actualmente en las zonas urbanas predominan las propiedades horizontales, dada la escasez del suelo.

Ante la dificultad que se tiene para hacer una valoración masiva de los inmuebles más ajustada a la realidad inmobiliaria, de una forma rápida y eficiente implicando menores recursos humanos y económicos, se hace necesario que el IGAC ajuste o complemente las metodologías que viene aplicando de tiempo atrás, explorando nuevos modelos de valoración e incorporando nuevas variables que permitan obtener avalúos de mayor exactitud y congruentes con la realidad del mercado inmobiliario (Fajardo, 2014)

3. Justificación

En el sector inmobiliario el cambio de los precios ha sido de gran interés de los gobiernos, gerentes e individuos debido a sus influencias sobre las condiciones socioeconómicas en la dimensión nacional.

El avalúo está determinado por las propiedades del predio, que son afectadas por variables externas como su ubicación geográfica, variables macroeconómicas, distancias a sitios de interés, características de la estructura de la comunidad y de los servicios ambientales (Kim & Park, 2005) (Kusan, Aytekin, & Özdemir, 2010).

También los índices de precios de vivienda son indicadores importantes para las partes interesadas en el mercado de bienes raíces. El valor de mercado se estima a través de la aplicación de métodos de valoración y procedimientos que reflejan la naturaleza de la propiedad y las circunstancias externas que podrían afectar su valor en el mercado (Pagourtzi, Assimakopoulos, Hatzichristos, & French, 2003, págs. 21-25, 383-401).

Con base en lo anterior tener un mecanismo objetivo de valoración resulta de interés para numerosos colectivos como son propietarios, constructores, agentes de ventas, inversionistas, entidades tasadoras, financieras, aseguradoras y la administración (Villamandos, Caridad, & Núñez, 2008).

La estimación de un avalúo catastral generalmente es un proceso individual realizado por personas expertas en el tema, sin embargo en el análisis de valoración de viviendas la literatura indica dos principales líneas de investigación: el uso de métodos hedónicos (MPH) que determina el precio de la vivienda acorde con diferentes características del predio como son calidad, área, tiempo de construcción, localización y otros, en los que la función del precio hedónico está dada por $P=XB+E$ en el cual P es el vector de precio de venta, X la matriz de características, B el vector de coeficientes de regresión y E el margen de error (Ayan & Erkin, 2013) y técnicas en el área de inteligencia artificial para desarrollar modelos predictivos de precios de vivienda donde el aprendizaje de máquina juega un papel importante para la estimación del precio (Park & Kwon, 2015).

En Colombia la entidad encargada de generar la normatividad y de realizar los avalúos catastrales es el Instituto Geográfico Agustín Codazzi (IGAC). De acuerdo con el marco legislativo de esta institución, el proceso de actualización de la información catastral es donde se actualizan las características del predio para definir el avalúo catastral, "el cual no podrá ser inferior al 60 % del respectivo valor comercial del inmueble, sin llegar a superar este último" (Ley 1450 de 2011), esto está previsto en el Plan de Desarrollo del Gobierno Nacional "Prosperidad para todos, 2.010 - 2.014".

La metodología de valoración masiva con la que actualmente se estiman los avalúos catastrales fue diseñada por el IGAC en el año de 1984. La Ley señala que el avalúo catastral se obtiene mediante investigación y análisis estadístico del mercado inmobiliario, y se determina mediante la adición

de los avalúos parciales practicados independientemente para los terrenos y para las edificaciones en él comprendidas (Ley 14, 1983).

La estimación del valor del terreno se realiza mediante la metodología de zonas homogéneas y para las construcciones se efectúa a través de modelos de regresiones simples ya sean lineales, exponenciales, logarítmicas o potenciales. La división del valor integral de una propiedad entre terreno y construcción con frecuencia genera resultados poco coherentes, pues terrenos con diferente localización o con diferente forma dentro de una área homogénea, deberían tener diferentes precios pero en la metodología de zonas homogéneas se les asigna el mismo valor y en el caso de las propiedades horizontales esta metodología obliga a asignarle a la construcción el efecto de las diferencias de características particulares de cada predio dentro de la propiedad horizontal.

Tener una mejor estimación de los valores catastrales del predio permite que los propietarios de los bienes inmuebles tengan información más aproximada de su patrimonio y así en el momento de transarlo en el mercado lo haga con aproximación a la realidad, con la finalidad de realizar el intercambio a un precio justo.

En nuestra región Pereira (Risaralda) no se ha identificado un sistema de información en aprendizaje de máquina en el campo de la Inteligencia artificial que pueda estimar el valor catastral del predio. Estas técnicas permiten incorporar nuevas variables al proceso de valoración predial para obtener resultados más acordes con el mercado inmobiliario actual.

Dado lo anterior ¿Es posible desarrollar un modelo para estimar el avalúo catastral que esté basado en las principales características del predio empleando técnicas de aprendizaje de máquina?

4. Objetivos

4.1 Objetivo General

Desarrollar un modelo que permita estimar el avalúo catastral de los predios urbanos de la ciudad de Pereira utilizando técnicas de aprendizaje de máquina.

4.2 Objetivos Específicos

- Aplicar técnicas de análisis exploratorio sobre la base de datos de los predios y avalúos catastrales para identificar relaciones entre las variables de análisis

- Realizar una comparación de las principales técnicas de aprendizaje de máquina para estimar el avalúo catastral con las características del predio

- Desarrollar una aplicación Web que permita seleccionar un predio y estimar su avalúo catastral utilizando el algoritmo y los parámetros que mejor se ajustaron a la estimación

5. Estado del Arte

Los Sistemas de Inteligencia Artificial muestran en las pruebas errores medios que se sitúan entre el 5 y el 10%, mientras que los de Regresión Múltiple se sitúan más entre el 10 y el 15%. En alguna de estas pruebas los resultados han sido similares para ambos sistemas, pero muestran una mayor precisión los sistemas que utilizan Inteligencia Artificial (Mora Esperanza, 2004).

En el 2004 en Nueva Zelanda se realizó una comparación empírica sobre el poder de predicción del modelo hedónico frente a las redes neuronales artificiales. Para el estudio se utilizó una muestra de 200 casas en Christchurch (Nueva Zelanda), se analizaron factores tales como la superficie de la casa, edad, tipo de casa, número de habitaciones, número de cuartos de baño, el número de garajes, se consideraron servicios alrededor de la casa y la ubicación geográfica.

En conclusión las redes neuronales son mejores en la predicción que los modelos hedónicos. También los modelos hedónicos muestran resultados más pobres en el pronóstico fuera de la muestra. El resultado del estudio en el modelo 1 (Vivienda con y sin jardín) muestra para los métodos hedónicos un error coeficiente de determinación R2 de 0,6192 frente a 0,9000 de las redes neuronales y un error de desviación cuadrático medio (RMSE) de 876.215,63 frente a 449.111,46 de las redes neuronales (Limbosunchai, 2004).

En el 2008 se realizó un estudio en la ciudad de Córdoba (España) en el que se examinaron dos métodos de valoración aplicables al mercado inmobiliario: la Metodología de Precios Hedónicos (MPH), en la cual se analiza el precio del bien vivienda en función de sus principales características, frente a las Redes Neuronales Artificiales (RNA), las cuales tratan de superar la inflexibilidad y linealidad de los modelos hedónicos tradicionales. Para el estudio se tomó una muestra de 2.888 registros en la que se utilizaron 16 características del inmueble, tanto internas como externas. Entre las variables internas se evaluaron: superficie de construcción, dormitorios, baños, aseos, terraza, teléfono, armarios empotrados, garaje, trastero, climatización, calidad, reformas, exterior y variables externas: año de edificación, ascensor, tendero, piscina, tenis, jardines y zonas de ubicación.

El proceso de estimación mediante el uso de redes neuronales ofreció resultados más satisfactorios que la estimación mediante modelos hedónicos. Con la red se ha conseguido un grado de ajuste del 86% (R2) frente al 77% alcanzado por el modelo hedónico de regresión. Además, al aplicar la red, la raíz del error cuadrático medio (RMSE) disminuye de 41.645,43 a 39.540,36, también se observa una clara disminución de la desviación típica residual, del error medio absoluto y del error medio relativo (Villamandos, Caridad, & Núñez, 2008).

En el 2009 Selim, en una ciudad de Turquía, comparó la regresión hedónica y las técnicas de redes neuronales, para determinar los precios de la vivienda. Se documentó que las redes neuronales

podían ser una mejor alternativa técnica de modelado en la determinación del precio de la vivienda en Turquía, en razón de la no linealidad de la regresión hedónica (Selim, 2009).

En Taiwán por ese mismo año se desarrolló un modelo híbrido basado en algoritmos genéticos y máquinas de soporte vectorial (HGA-SVR) y en la teoría del Feng Shui que hace parte de la cultura China. Este trata sobre la disposición del espacio y orientación de las partes de la vivienda según los días del calendario. En el estudio se realizaron dos modelos: modelo 1 sin Feng Shui y modelo 2 con variables que afectan el Feng Shui. Para el modelo 1 se utilizaron 12 variables, en el modelo 2 se utilizaron 16 variables. Se realizó la comparación del modelo (HGA-SVR) con los modelos BPN Back-Propagation Neuronal Network y FFN-Feedforward, para los datos de entrenamiento y prueba se evaluaron 190 viviendas de los distritos norte y oeste de la ciudad de Taichung. El estudio dio como resultado que el modelo HGA-SVR fue mejor que los otros (Wu, Li, Fang, Hsu, & Ling, 2009).

En Bogotá en el 2009 se utilizaron las máquinas de soporte vectorial para realizar regresión y aplicarlo al cálculo del metro cuadrado construido en esa ciudad. El estudio estuvo enfocado a explorar las diferentes alternativas para encontrar mecanismos automáticos que produzcan los avalúos catastrales de la manera más precisa posible. Para el estudio se utilizaron 2.627 registros. Este estudio es la continuación de dos trabajos previos: Oficina de Extensión y Asesoría de la Universidad Nacional de Colombia. Elaboración de Model os Econométricos División de Actualización DACD. Facultad de Ciencias. Departamento de Esta dística, Bogotá. 2002. (Oficina de extensión y asesoría de la Universidad Nacional de Colombia, 2002) y (Ávila & Robayo, 2003) .

El resultado en el que se desarrolló el modelo econométrico fue una regresión lineal basada en características del predio, dio como resultado un error cuadrático medio de 1,97 que representa $3’546.180.oo pesos colombianos por metro cuadrado, y el resultado del segundo estudio de la red neuronal con el uso de la función de activación, Elliott dio como resultado un error de 0,26 que representa $473.940.oo pesos colombianos por metro cuadrado. Al utilizar la función de activación logística dio como resultado 0,41 que representa $751.320.oo pesos colombianos por metro cuadrado.

En la aplicación de las máquinas con vectores de soporte se utilizó la función kernel polinomial y un modelo de regresión ridge con función kernel de grado polinomio 5 y un valor lambda de 0,005. Como resultado dio un error de 0,23 en el cual se usaron los mismos indicadores prediales del modelo econométrico y la red neuronal. En conclusión los métodos de regresión de las máquinas con vectores de soporte ofrecieron buenos resultados en comparación con los obtenidos en el modelo econométrico, pero no fue mejor resultado que las redes neuronales, cuando se utilizaron todos los indicadores ya que dio un error cuadrático medio de 0.10 que corresponde a $ 179.521 pesos colombianos por metro cuadrado frente a 0.01 error de la red neuronal que corresponde a $28.700 pesos colombianos por metro cuadrado (Morales & Hernández, 2009).

Otro estudio fue realizado en el 2010 en Eskis ehir (Turquía) en el que se desarrolló un nuevo modelo con lógica difusa para predecir el precio de las construcciones en la ciudad. Para el modelo de predicción se ingresaron como entrada los planes de la ciudad, la cercanía a la diversidad cultural, hospitales, escuelas, desarrollo tecnológico, transporte público, además, una aplicación de cuestionario incluyó estos factores para determinar los valores de la formación y las pruebas de

los conjuntos difusos. Las entradas del modelo se dividen en varios factores que son variables relacionadas a la casa: 1. Factores de la residencia (RF), 2. Factores de construcción, 3. Factores del piso 4. Factor de conformidad de la región, criterio de planeación de la ciudad, 5. Factor de ruido y contaminación, 6. Factor de transporte, 7. Factor de cercanía a centros socio culturales, 8. Factor de cercanía al comercio, 9. Factor de transporte público y 10. Factor de región socio económica. En el estudio se seleccionó una muestra de 200 residencias en 40 regiones urbanas. Los resultados muestran que la predicción se aproxima a los valores reales. El error de desviación del cuadrático medio (RMSE) dio 4,86 X 10^{-4}, el coeficiente de determinación (R^2) dio como resultado 0,99 y el error porcentual absoluto medio (MAPE) 0,007% (Kusan, Aytekin, & Özdemir, 2010).

En el 2014 en el Condado de Fairfax Virginia se partió de una muestra de 5.359 viviendas. Fueron evaluadas inicialmente 76 características de las cuales se seleccionaron 49 aplicando una prueba t student sobre las variables. Posteriormente se seleccionaron 28 variables que fueron determinadas por el método de regresión logística, en el experimento se aplicaron 4 métodos de clasificación. La idea del experimento fue clasificar el valor de la vivienda en dos grupos: alto precio, bajo precio, los algoritmos de aprendizaje de máquina utilizados fueron: C4.5, RIPPER, Naïve Bayesian, y AdaBoost, para el análisis de los algoritmos se utilizó el software WEKA que es una librería de algoritmos de aprendizaje de máquina para minería de datos desarrollado en el lenguaje de programación JAVA en la Universidad de Waikato (Weka 3: Data Mining Software in Java, s.f.). En todas las pruebas realizadas el modelo de RIPPER supera a los otros modelos, tanto en desempeño, como en precisión en la clasificación, en comparación con los datos reales (Park & Kwon, 2015).

En el 2014 en la ciudad de New York se realizó un estudio en el que fueron comparados el desempeño entre el modelo de regresión múltiple y el de redes neuronales para la estimación del valor de las casas, en este estudio se seleccionó una muestra aleatoria de 1.047 casas, las características principales utilizadas fueron el área habitable, número de cuartos, área del lote, antigüedad de la casa, el resultado del estudio fue que el coeficiente de determinación r^2 en la red neuronal fue superior al modelo de regresión lineal múltiple, y el error cuadrático medio la red neuronal fue inferior comparado con el otro modelo, lo que determinó que la red neuronal es mejor en la estimación del precio de viviendas (Khamis & Binti Kamarudin, 2014).

En Colombia se realizó un estudio en Fusagasugá en el año 2014 mediante Redes Neuronales Artificiales (RNA) para la determinación del avalúo de un predio urbano. Para el estudio se utilizaron 991 predios urbanos. En total se utilizaron 14 variables incluida la variable geográfica. Para el entrenamiento de la red neuronal se tomó 90,1 % de los datos y 9,9 % para validación. Finalmente, para la implementación de la red, se utilizó el programa estadístico SPSS (Fajardo, 2014, págs. 32-36).

ESTUDIO	Nueva Zelanda (2004)		Córdoba España (2008)		Taiwán Modelo 1 (2009)			Bogotá 2009		Eskis ehir (Turquía)(2010)		Condado de Fairfax Virginia(2014)				New York (2014)		Fusagasugá Colombia (2014)
MODELO	MPH	RNA	MPH	RNA	BNP	FNN	HGA-SVR	ME	RNA	MSV	LOGICA DIFUSA	C.4.5	IPPER	N. BAYESIAN	ADA BOOST	RM	RNA	RNA
COEFICIENTE DE DETERMINACIÓN (R^2)	61,92%	90,00%	77,38%	86,05%							99.9 %					64,60%	81,70%	
RAÍZ DEL ERROR CUADRÁTICO MEDIO(RMSE)	876215,63	449111,46	41.645,43	39.540,36				19.70	0,26	0,2263	4.86 X 10-4					$1,633 X 10^9$	1,293 X 10-9	
DESVIACIÓN TÍPICA RESIDUAL			41.911,91	39.102,13														
ERROR MEDIO ABSOLUTO			30.579,18	28.551,34														
ERROR MEDIO RELATIVO			14,45%	13,69%														13,60%
PROCENTAJE ERROR MEDIO ABSOLUTO (MAPE)					12.5%	9.80%	8.99%				0.007%							
FE(ERROR DE PREDICCIÓN)< 5%					70.59%	68.38%	70.21%											
FE(ERROR DE PREDICCIÓN) 5% ~ 15%					9.80%	11.11%	14.89%											
FE(ERROR DE PREDICCIÓN) >15%					19.61%	20.51%	13.82%											
ERROR VALIDACION CRUZADA (PROMEDIO)												0,28	0,25	0,30	0,26			
NUMERO DE OBSERVACIONES	200	200	2888	2888	130	117	94	2627	2627	2627	200	5359	5359	5359	5359	1047	1047	981
NUMERO DE VARIABLES	13	12	26	26	12	12	12	17	17	17	28	28	28	28	28	5	5	14

Tabla 1. Resumen de los estudios realizados en la estimación de precios de vivienda

6. Materiales y métodos

6.1 Descripción del territorio

Entorno Físico y Geográfico: El Municipio de Pereira está localizado a 4º 49' de latitud norte, 75º 42' de longitud y 1.411 metros sobre el nivel del mar; en el centro de la región occidental del territorio colombiano, en un pequeño valle formado por la terminación de un contrafuerte que se desprende de la cordillera central. Su estratégica localización central dentro de la región cafetera, la ubica en el panorama económico nacional e internacional. Está unida vialmente con los tres centros urbanos más importantes del territorio nacional y con los medios tanto marítimos como aéreos de comunicación internacionales. Su población consta de 488.839 personas de las cuales 410.535 se encuentran en el área urbana localizadas en 19 comunas y 78.304 en el área rural en 12 corregimientos.

En su geografía cuenta con pisos térmicos que van desde las nieves perpetuas (Nevado de Santa Isabel a 5.200 mts sobre el nivel del mar) en límites con el Departamento del Tolima, hasta pisos cálidos a 900 mts / sobre el nivel del mar y a orillas del rio Cauca. Por lo tanto, presenta distintas alternativas de uso agrícola.

De hecho, existen áreas de bosques para protección de cuencas, zonas de diversificación y zonas medias conocidas como la zona cafetera y zonas cálidas con actividad ganadera y agrícola (piña, caña de azúcar, caña panelera y pasto).

La extensión geográfica municipal de Pereira es de 702 km2 y se encuentra a una altura promedio de 1.411 mts. sobre el nivel del mar y cuenta con una temperatura promedio de 21ºC.

Límites: Al Sur, con los municipios de Ulloa (Departamento del Valle), Filandia y Salento (Departamento del Quindío). Al Oriente, con el Departamento del Tolima, con Anzoátegui, Santa Isabel, Ibagué y zona de los nevados. Al Occidente, con los municipios de Cartago, Anserma Nuevo (Departamento del Valle), Balboa, La Virginia (Departamento de Risaralda) (Pereira, 2017).

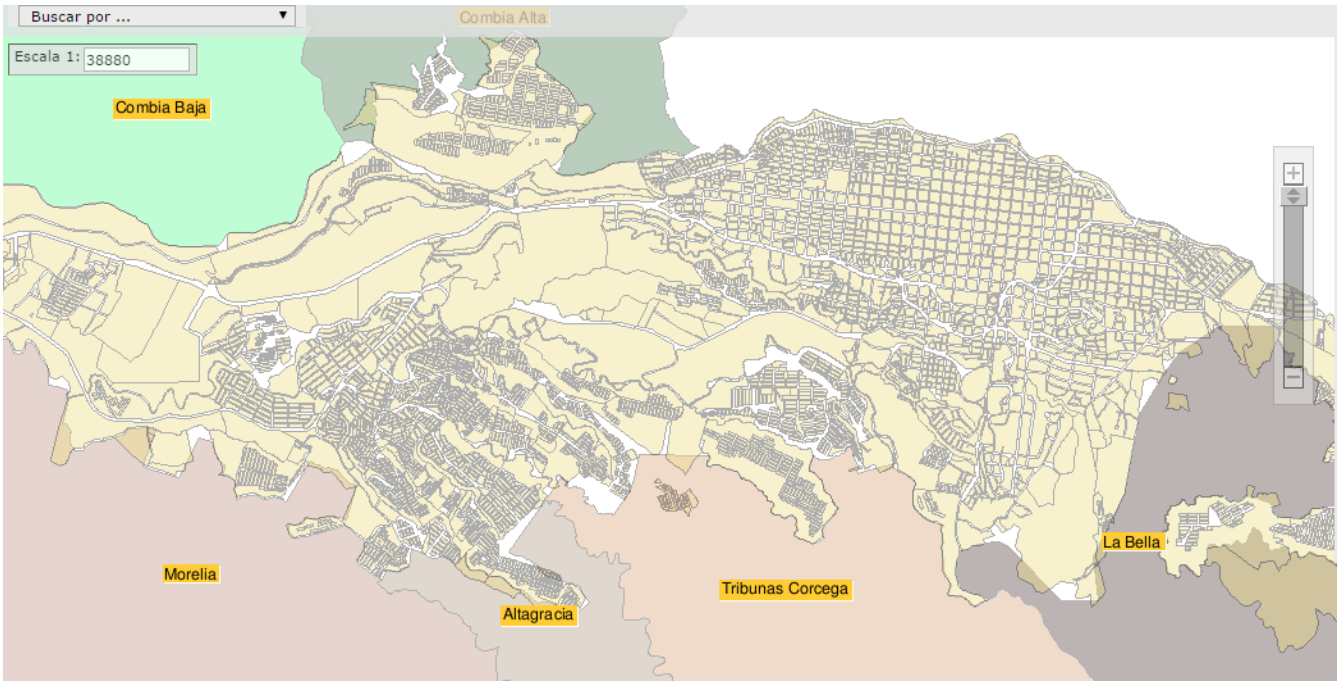


Figura 1. Límites de la ciudad de Pereira, fuente: elaboración propia

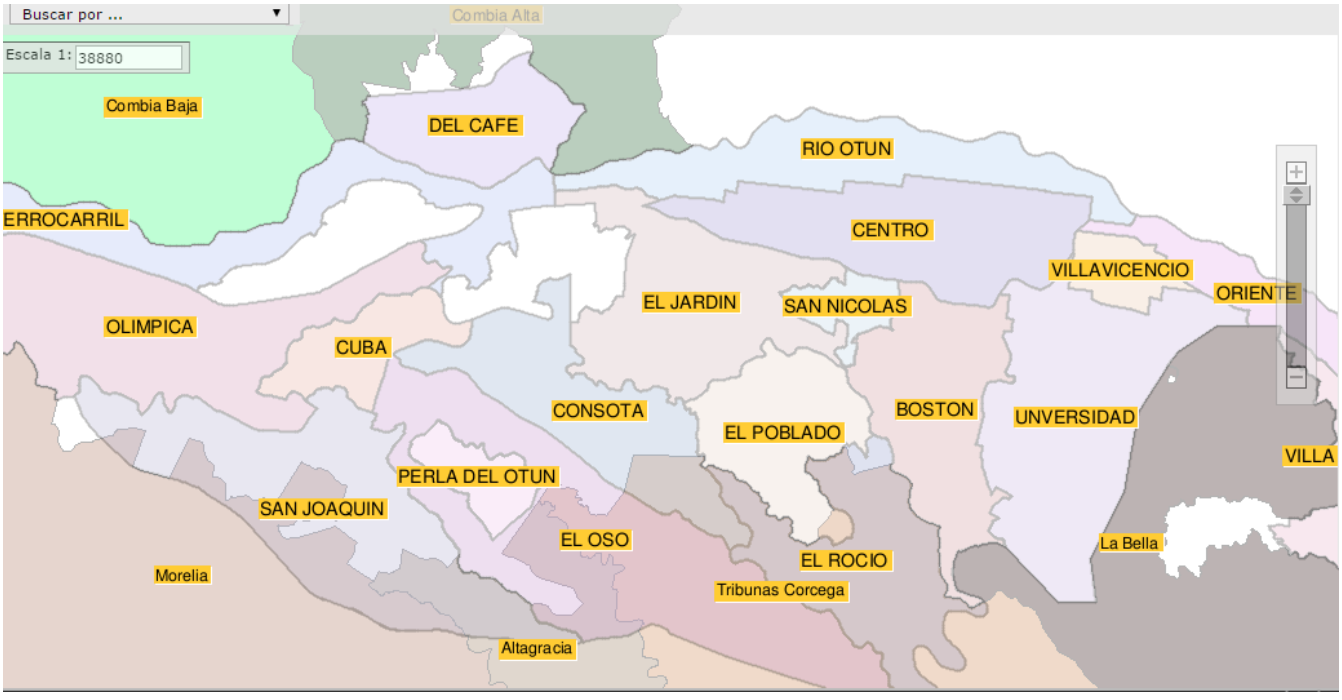


Figura 2. División por comunas ciudad de Pereira, fuente: elaboración propia

6.2 Información Predial

La base de datos catastral año 2013 del municipio cuenta con 73.078 predios urbanos, y un avalúo catastral global urbano de $6.497.099.291.200. A continuación se presenta una distribución por rango de avalúos.

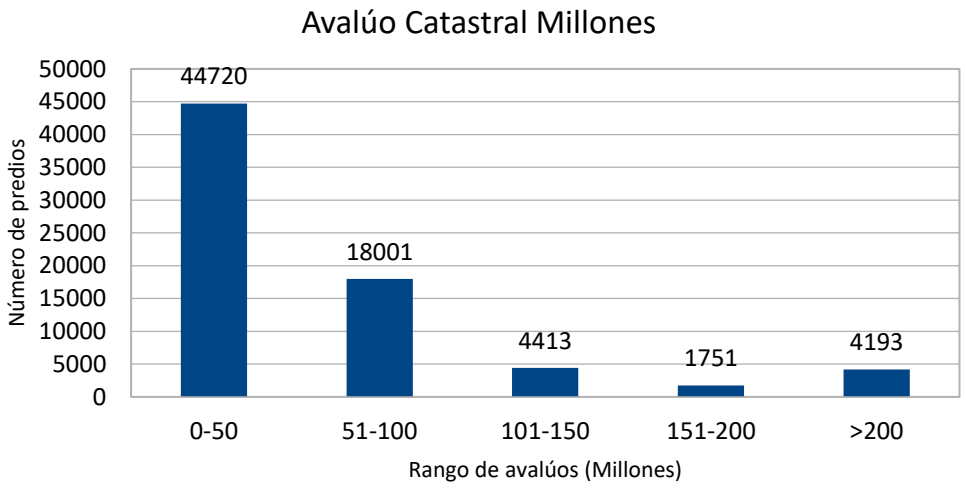


Figura 3. Avalúo catastral y número de predios

Como se observa en la figura 3, el 61 % de los predios esta entre 0 y 50 millones, el 25% entre 51 y 100 millones, el 6% entre 101 y 150 millones, el 2% entre 151 y 200 millones y 6% mayores a 200 millones. Podemos afirmar que la mayor concentración del avaluó se encuentra entre 0 y 100 millones que representa un 86 % del total de los datos.

6.3 Materiales

Tabla 2. Materiales utilizados en el desarrollo de la investigación

Material	Fuente	Año	Descripción
Base de datos información predial y avalúos catastrales	IGAC	2013	Información catastral proporcionados por el Instituto Agustín Codazzi seccional Pereira (Risaralda)
Weka 3.6	http://www.cs.waikato.ac.nz/ml/weka/	2017	Software y Librería para minería de datos y aprendizaje de máquina desarrollado en el lenguaje de java

scikit-learn 0.18	http://scikit-learn.org/stable/	2017	Librerías de aprendizaje de máquina escritas en el lenguaje Python, que se enfoca en algoritmos de clasificación, regresión y agrupamiento en el cual se puede aplicar aprendizaje de máquina supervisado y no supervisado (Brownlee Jason, Machine Learning Mastery With Python 2016)
SciPy 0.18.1	https://www.scipy.org/	2017	Es un ecosistema de librerías en Python para matemáticas, ciencia e ingeniería, está compuesto por los siguientes módulos: Numpy 1.11.2 Permite trabajar eficientemente con arreglos de datos. Matplotlib 1.5.1 Permite crear gráficas estadísticas en 2D Pandas 0.18.0 Permite analizar y estructurar los datos.
Python 2.7.11	https://www.python.org/	2017	Lenguaje de programación que permite integrar varias librerías eficientemente.
Mlxtend 0.6.0	http://rasbt.github.io/mlxtend/	2017	Mlxtend es una librería de Python para aprendizaje de máquina, en este proyecto se utilizó específicamente para la selección de características utilizando Sequential Forward Selection (SFS)

6.4 Métodos

6.4.1 Selección de la muestra

Los datos fueron obtenidos de las memorias del Estudio Zonas Homogéneas Geoeconómicas de la Zona Urbana del Municipio de Pereira, realizado por el Instituto Geográfico Agustín Codazzi – Territorial Risaralda, obtenidos durante la fase de campo y puestos en vigencia en el año 2013.

6.4.2 Análisis exploratorio de los datos

6.4.2.1 Detección de valores atípicos y ausentes de la muestra

El estudio se inició con una muestra de 73.078 predios urbanos, los cuales fueron sometidos a un cuidadoso proceso de depuración, que consistió en la eliminación de datos atípicos, faltantes y extremos. Esta labor de filtrado se realizó a partir de la del análisis exploratorio de datos y el estudio de los gráficos de dispersión que fue realizado para cada variable.

Con relación a los predios con datos ausentes y por datos atípicos estos fueron eliminados debido a que se contaba con una muestra de tamaño suficiente, por lo tanto solo fueron utilizadas observaciones con datos completos. Fue así, como se eliminaron o descartaron 14.555 predios. Se logró finalmente una muestra efectiva de 58523 predios.

6.4.2.2 Preparación de los datos

Se realizó una transformación y normalización de los datos con el fin de ingresarlos a los algoritmos: RNA (Red Neuronal Artificial), MSV (Maquinas de Soporte Vectorial), PG (Proceso Gaussiano), RLM (Regresión Lineal Múltiple).

Para cada una de las variables se realizaron normalización de las variables cuantitativas y codificación de las cualitativas, para que de esta forma pudieran ser analizadas e incluidas en el modelo. Este tiene el propósito de ajustar los datos de alguna manera tal que los algoritmos puedan procesar la información eficientemente, es decir proceder a su normalización (Casas Fajardo A.E. Propuesta Metodológica para Calcular el Avalúo Catastral de un Predio Utilizando Redes Neuronales Artificiales Tesis Universidad Nacional de Colombia p. 32-36, 2014) (Azoff, 1995).

La lista y descripción de variables se anexan al final de este documento. Para transformar la variable del avalúo se realizó una operación matemática de dividir el valor entre el área de terreno y luego entre 1.000.000, con el fin de realizar la estimación del valor por metro cuadrado.

6.4.3 Comparación de los algoritmos de estimación RNA, MSV, PG, RLM

Para comparar los algoritmos primero se debe seleccionar el mejor modelo con los hiperparámetros óptimos, para lo anterior se utiliza validación cruzada con cinco grupos y el estimador del error cuadrático medio (ECM), los datos se dividen 4 grupos para entrenamiento y un grupo de validación que corresponde al 80% (entrenamiento) y al 20% (prueba).

Con el fin de optimizar los hiperparámetros y seleccionar el mejor algoritmo con base en el error cuadrático medio (ECM) y coeficiente de determinación (r^2), se utiliza validación cruzada anidada. Esta consiste en combinar validación cruzada con búsqueda en grilla basada en un conjunto de hiperparámetros iniciales del algoritmo.

En la validación cruzada anidada tenemos un ciclo de validación cruzada de k-grupos externo para dividir los datos en conjuntos de entrenamiento y prueba, y un ciclo interno se utiliza para seleccionar el modelo que usa la validación cruzada de k-grupos en el conjunto de entrenamiento.

Después de la selección del modelo, el conjunto de prueba se utiliza para evaluar el rendimiento. La figura siguiente explica el concepto de validación cruzada anidada con cinco conjuntos externos y dos internos, lo que puede ser útil para conjuntos de datos grandes donde el rendimiento computacional es importante; este tipo particular de validación cruzada anidada también se conoce como validación cruzada 5x2 (Raschka, 2015, págs. 187-188)

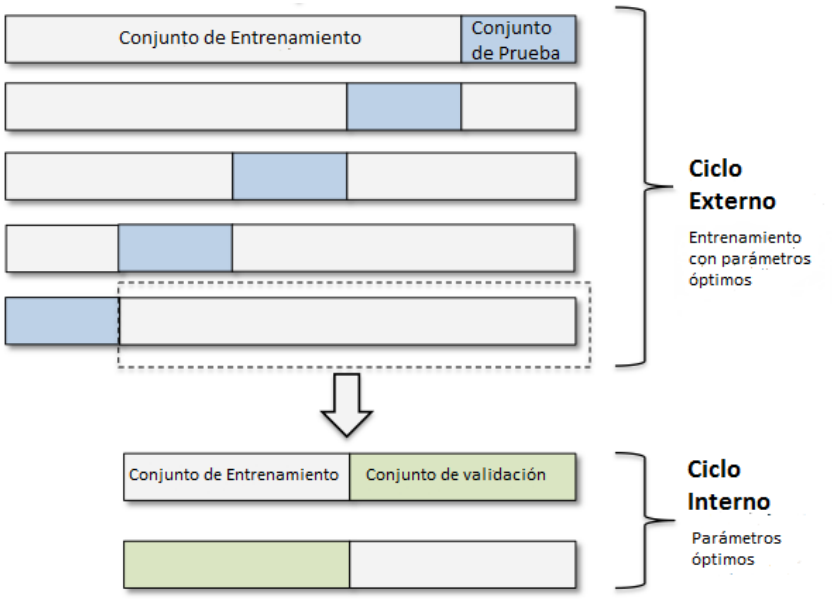


Figura 4. Validación cruzada anidada (Raschka , 2015 ,pág. 188).

6.4.4 Selección de características

Para cada algoritmo se realizó selección de características en la que se utilizó Sequential Forward Selection (SFS) la cual consiste en determinar el conjunto de características que tenga el mejor desempeño con base en el estimador del error cuadrático medio (ECM). El conjunto de características se determinó en un rango de 5 – 10 variables, donde se evaluaron todas las posibles combinaciones (Raschka, Recuperado de http://rasbt.github.io/mlxtend/user_guide/feature_selection/SequentialFeatureSelector, 2017)

Luego de tener el conjunto de características se realizó una selección de cinco muestras aleatorias de 10.000 registros, a cada muestra aleatoria se le realizó validación cruzada anidada con cinco grupos. Finalmente se estimó el error cuadrático medio (ECM) y el coeficiente de determinación (r^2) para cada algoritmo.

7. Marco Teórico

7.1 Regresión Lineal Múltiple – Modelo Hedónico

El análisis de regresión es definido como una herramienta estadística para predecir los valores de una variable continua dependiente con base en los valores de una variable independiente. La regresión lineal es utilizada para explorar los datos, con el fin de comprender la naturaleza de las relaciones entre las diferentes variables. Al ser una relación matemática entre la variable de respuesta y la variable explicativa, se asume una correlación lineal entre las dos.

La regresión múltiple permite más variables independientes que tienen un efecto positivo o negativo sobre la variable dependiente (Murphy, 2012). La relación lineal entre la variable dependiente y la variable independiente se presenta por medio de la línea de ajuste óptimo, también llamada recta de regresión, la cual es una línea que se acerca tanto como sea posible a todos los puntos en el diagrama de dispersión.

La relación entre la variable y (variable dependiente) y de m variables independientes x_i (i=i hasta m) se representa por la siguiente ecuación:

$$Y = w_0 + w_1 x_1 + w_2 x_2 + w_2 x_2 + e$$

Los coeficientes w_0, w_1, w_2 w_3 son cantidades desconocidas que necesitan ser determinadas matemáticamente minimizando la suma de los errores y e es un término de ruido aleatorio normalmente distribuido con media cero y desviación estándar desconocida δ.

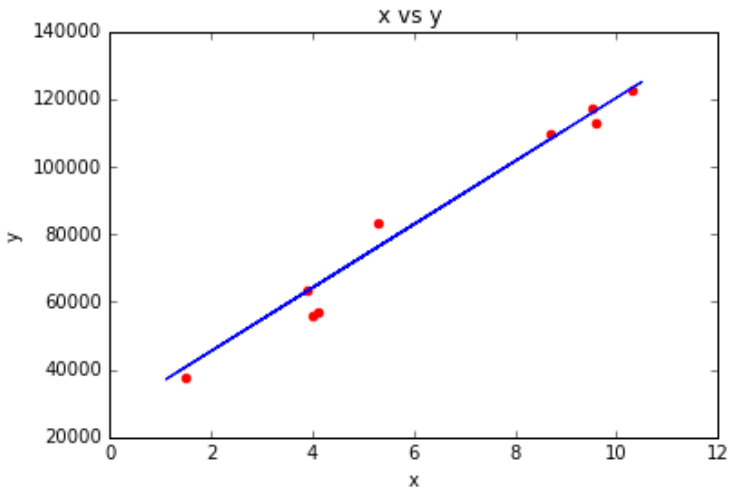


Figura 5. Regresión lineal fuente elaboración propia

Para minimizar la suma de la distancia al cuadrado de cada punto de entrenamiento (denotado en el punto rojo de la figura) al punto estimado representado por la línea azul, se debe minimizar el error residual.

$$RSS(w) = \|\in\|_2^2 = \sum_{i=1}^{m} \in_i{}^2$$

$$donde\ \in_i = (y_i - w^T x_i)$$

El RSS es también llamado la suma de los errores al cuadrado y divido entre el número de datos es llamado error cuadrático medio (ECM), la estimación de la máxima verosimilitud para w es minimizar la función RSS.

El RSS mide la precisión de los valores previstos entre la variable dependiente indicando que tan lejos o cerca están los valores reales de los estimados, cuanto menor sea el RSS mayor será la precisión con respecto a la recta de regresión. Al asumir que los datos de entrenamiento son independientes e idénticamente distribuidos, se puede escribir la función de verosimilitud como:

$$\ell(\theta) \triangleq \log\big(P(D|\theta)\big) = \sum_{i=1}^{N} \log\big(P(y_i|x_i,\theta)\big)$$

Dado que maximizar la verosimilitud es equivalente a minimizar el negativo de la misma o NLL.

$$NLL(\theta) \triangleq -\sum_{i=1}^{N} \log\big(P(y_i|x_i,\theta)\big)$$

Dado

$$\ell(\theta) = \sum_{i=1}^{N} \log\left[\frac{1}{2\pi\sigma^2}^{\ 2}\ e^{-\frac{1}{2\delta^2}(y_i - w^T x_i)^2}\right]$$

$$\ell(\theta) = -\frac{1}{2\sigma^2} RSS(w) - \frac{N}{2}\log(2\pi\delta^2)$$

y $RSS(w) \triangleq \|\epsilon\|_2^2 = \sum_{i=1}^{N} \epsilon_i^2$ donde $\epsilon_i = (y_i - w^t x_i) = (y_i - \hat{y}_i)$

(Murphy, 2012, pág. 217).

7.2 Procesos Gaussianos

En aprendizaje supervisado tenemos un histórico de entradas x_i y sus correspondientes salidas y_i, estos son utilizados como datos de entrenamiento, se asume que las salidas están relacionadas con las entradas por medio de una función f, tal que $y_i = f(x_i)$ siendo f una función posiblemente afectada por el ruido; un enfoque óptimo es introducir una distribución de probabilidad condicional sobre las funciones de los datos P(f(x,y)) y realizar predicciones sobre los nuevos valores de entrada (Murphy, 2012, pág. 515).

$$P(y_*|x_*, X, y) = \int P((y_*|f, x_*)p(f|X, y)df$$

Un proceso gaussiano (PG) es una colección infinita de variables aleatorias escalares indexadas por un espacio de entradas X de tal forma que para cada combinación finita de puntos X= $\{X_{i...}.....X_n\}$, todas las funciones $f \triangleq [f(X_i) \dots \dots \dots \dots \dots f(X_n)]^T$ siguen una distribución gaussiana multivariada.

Donde un proceso gaussiano está especificado por una función media m(X) y una función de covarianza k(x, x'), esto es una generalización natural de la distribución gaussiana donde la media es un vector y la covarianza es una matriz.

$$f \approx PG(m, k)$$

Esto significa que la función f es distribuida como un PG con función media m y función de covarianza k (Rasmussen & Williams, 2006, pág. 13)

7.2.1 Procesos gaussianos para regresión

En procesos gaussianos para regresión, se define un prior sobre la función de regresión que puede ser denotado como:

$$f \approx PG(m(x), k(x, x'))$$

$$m(x) = \mathbb{E}[f(x)]$$

$$k(x, x') = \mathbb{E}[f(x) - m(x)(f(x') - m(x'))^T]$$

Donde k() debe ser una función kernel semidefinida positiva con la cual se determina la matriz de covarianza, para todo el conjunto de puntos este proceso se define como una mezcla gaussiana:

$$p(f|x) = \mathcal{N}\big((f|\mu), K\big)$$

Donde $K_{ij} = K\big(x_i, x_j\big)$ y $\mu = (m(x_1)\dots m(x_n))$

Debido a que muchas veces no existe información a priori sobre la media de las funciones, suele asumirse m(x)=0.

7.2.3 Predicciones usando observaciones libres de ruido

Dado un conjunto de datos de entrenamiento $D = \{(x_i, x_j), i = 1{:}N\}$ donde f=f(x_i) y la f libre de ruido, se tiene un conjunto de datos de pruebas x_* de tamaño N_* x D, se quiere predecir la función de salida f_*.
Por definición el PG tiene la siguiente forma:

$$\begin{pmatrix} f \\ f_* \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu \\ \mu_* \end{pmatrix}, \begin{pmatrix} K & K_* \\ k_*^T & K_{**} \end{pmatrix}\right)$$

Donde K=k(X,X) es de tamaño N x N, K_* =k(X,X_*) es de tamaño N x N_*, y K_{**} =k(X_*,X_*) es de tamaño N_* x N_*.

Dado el prior se obtiene una distribución del posterior condicionado a las observaciones que utilizan propiedades gaussianas, las cuales toman la forma:

$$p(f_*|X_*, X, f) = \mathcal{N}\big((f_*|\mu_*, \Sigma_*)\big)$$

$$\mu_* = \mu(X_*) + k_*^T K^{-1}(f - \mu(X))$$

$$\Sigma_* = K_{**} - K_*^T K^{-1} K_*$$

Otra representación sería:

$$(f_*|X_*, X, f \sim \mathcal{N}((K(X_*, X)K(X, X)^{-1}f), K(X_*, X_*) - K(X_*, X)K(X, X^{-1})K(X, X_*))$$

Donde el término $K(X_*, X)K(X, X)^{-1}f$ representa la media y $K(X_*, X_*) - K(X_*, X)K(X, X^{-1})K(X, X_*)$ Representa la matriz de covarianza.
Si asumimos la media igual a cero tenemos:

$$\begin{pmatrix} f \\ f_* \end{pmatrix} \sim \mathcal{N}\left(0, \begin{pmatrix} K(X, X) & K(X, X_*) \\ K(X_*, X) & K(X_*, X_*) \end{pmatrix}\right)$$

Un kernel utilizado comúnmente es el RBF kernel o exponencial cuadrático dado por:

$$K(X, X') = \delta^2 + e^{\left(-\frac{1}{2\ell^2}(X - X\prime)^2\right)}$$

Donde ℓ controla la longitud de escala horizontal y δ^2 controla la variación vertical.
Cuando tenemos situaciones donde las observaciones presentan ruido tenemos $y = f(x) + \epsilon$ donde $\epsilon \sim \mathcal{N}(0, \delta_y^2)$ siendo δ_y^2 la varianza del ruido, la covarianza toma la forma de:

$$Cov(y[X]) = k + \delta_y^2 I_N \triangleq K_y$$

Donde I es la matriz identidad.
Al haber asumido el ruido independientemente de la distribución de probabilidad conjunta a los datos y la función a predecir tenemos:

$$\begin{pmatrix} f \\ f_* \end{pmatrix} \sim \mathcal{N}\left(0, \begin{pmatrix} K(X, X) + \delta_y^2 I & K(X, X_*) \\ K(X_*, X) & K(X_*, X_*) \end{pmatrix}\right)$$

$$(f_*|X_*, X, f \sim \mathcal{N}(f_*, Cov(f_*))$$

Donde

$f_* \triangleq \mathbb{E}(f_*|X, y, X_*) = K(X_*, X)\,[K(X, X) + \delta^2 I]^{-y}$

$Cov(f_*) = k(X_*, X_*) - K(X_*, X)[K(X, X) + \delta^2 I]^{-1} K(X, X_*)$ (Murphy, 2012, pág. 518).

Con el fin de introducir la verosimilitud marginal o evidencia, p(y|x) la cual es la integral del producto de la verosimilitud y el prior.

$$P(y|x) = \int P((y|f, x)p(f|x)df$$

La cual se refiere a la marginalización sobre los valores de la función f bajo el modelo del proceso gaussiano en el cual el prior es gaussiano,

$$P(f|x) = \mathcal{N}(0, K)$$

$$\log\left(P\left((f|x)\right)\right) = -\frac{1}{2} f^T K^{-1} f - \frac{1}{2}\log|K| - \frac{n}{2}\log(2\pi)$$

$$\mathrm{L} = \log\left(P\left((y|x, \theta)\right)\right) = -\frac{1}{2} y^T K_y^{-1} y - \frac{1}{2}\log\left|K_y\right| - \frac{n}{2}\log(2\pi)$$

Donde $K_y = K_f + \delta^2 I$ y K_f es la matriz de covarianza de la función sin ruido, θ son los avalúos de la función de la covarianza.
Se pueden estimar los valores óptimos de los hiperparámetros maximizando la verosimilitud marginal basada en derivadas parciales de cada uno de los avalúos.

$$\frac{\partial \log(P\left((y|x, \theta)\right))}{\partial \theta_j} = \frac{1}{2} y^T K_y^{-1} \frac{\partial K_y}{\partial \theta_j} K_y^{-1} y - \frac{1}{2} tr(K_y^{-1} \frac{\partial K_y}{\partial \theta_j})$$

$$= \frac{1}{2} tr(\frac{\alpha\alpha^T - K_y^{-1}\,\partial K_y}{\partial \theta_j})$$

Donde $\alpha = K_y^{-1}\mathrm{y}$.
Para encontrar una buena configuración de hiperparámetros es conveniente utilizar una rutina de optimización numérica como el gradiente conjugado, también cabe anotar que la optimización de los hiperparámetros no es un problema convexo entonces puede darse que no exista un único máximo local.

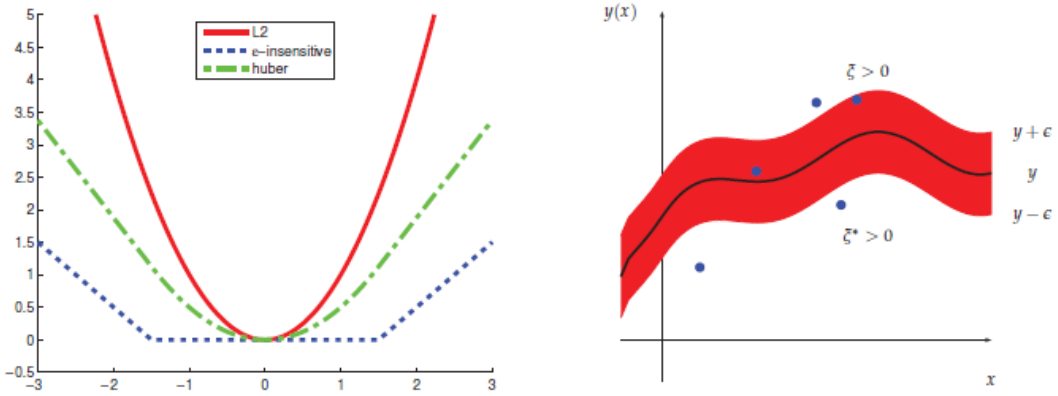
7.3 Máquinas de soporte vectorial para regresión

En las máquinas de vectores de soporte para regresión, la función objetivo se define como:

$$J = C\sum_{i=1}^{N} L_e\left(y_{i,}\hat{y}_i\right) + \frac{1}{2}\|w\|^2$$

Donde $\hat{y}_i = f(x_i) = w^T x_i + w_0$ y $C = \frac{1}{\lambda}$ que es la constante de regularización, la función $L_e(y_{i,}\hat{y}_i)$ se conoce como la función de pérdida insensible épsilon la cual está dada por:

$$L_e\left(y_{i,}\hat{y}_i\right) \triangleq \begin{cases} 0 & si\ |y-\hat{y}| < \xi \\ |y-\hat{y}| - \xi & de\ otra\ forma \end{cases}$$



Figuras 6 y 7. Función de perdida insensible y e-tube usado en modelos de regresión Fuente (Murphy, 2012, pág 497)

El objetivo es minimizar la función J, para ello se debe tener en cuenta que la función $L_e(y_{i,}\hat{y}_i)$ no es diferenciable por el valor absoluto, el problema se formula como un problema de optimización con restricciones, se introducen las variables slack para representar el grado para el cual cada punto está por fuera del e-tubo, para cada dato x_n se necesitan dos variables slack $\xi_n \geq 0$ y $\hat{\xi}_n \geq 0$, donde $\xi_n \geq 0$ corresponde a un punto para el cual $y_{i,} > y(x_i) + \epsilon$ y $\xi_n > 0$ corresponde a un punto para el cual $y_{i,} < y(x_i) - \epsilon$, la condición para que un punto objetivo se encuentre dentro de un e-tubo es que :

$$y(x_i) - \epsilon \leq y_{i,} \leq y(x_i) + \epsilon$$

Al introducir las variables slack se permite que los puntos estén por fuera del e-tubo.

$$y_{i,} \leq y(x_i) + \epsilon + \hat{\xi}_i$$

$$y_{i,} \geq y(x_i) - \epsilon + \xi_i$$

Dado lo anterior podemos reescribir el objetivo como:

$$J = C\sum_{i=1}^{N}(\widehat{\xi}_i + \xi) + \frac{1}{2}\|w\|_2^2$$

Que es la función cuadrática de w y debe ser minimizada dada las condiciones:

$$\hat{\xi} \geq 0$$
$$\xi \geq 0$$
$$y_{i,} \leq y(x_i) + \epsilon + \hat{\xi}_i$$
$$y_{i,} \geq y(x_i) - \epsilon + \xi_i$$

Se puede minimizar introduciendo los multiplicadores de lagrange:

$$a_i \geq 0$$
$$\hat{a}_i \geq 0$$
$$\mu_i \geq 0$$
$$\hat{\mu}_i \geq 0$$

Y luego optimizando el lagrangiano.

$$L = C\sum_{i=1}^{N}(\widehat{\xi}_i + \xi_i) + \frac{1}{2}\|w\|_2^2 - \sum_{i=1}^{N}(\hat{\mu}_i\hat{\xi}_i + \mu_i\xi_i) - \sum_{i=1}^{N} a_i(\epsilon + \xi_i + y(x_i) - y_{i,})$$

$$-\sum_{i=1}^{N}\hat{a}_i(\epsilon + \hat{\xi}_i - y(x_i) + y_{i,})$$

Realizando los reemplazos necesarios se puede demostrar que el problema dual implica maximizar:

$$\tilde{L}(a,\hat{a}) = -\frac{1}{2}\sum_{i=1}^{N}\sum_{n=1}^{N}(a_i - \hat{a}_i)(a_n - \hat{a}_n)K(x_i, x_j) - \epsilon\sum_{n=1}^{N}(a_n + \hat{a}_n) + \sum_{n=1}^{N}(a_n - \hat{a}_n)y_i$$

$sujeto\ a:$

$$0 \leq a_i \leq C$$
$$0 \leq \hat{a}_i \leq C$$
$$\sum_{n=1}^{N}(a_n - \hat{a}_n) = 0$$

La regresión toma la forma de:

$$y(x) = \sum_{i=1}^{N}(a_i - \hat{a}_i)K(x, x_i) + w_0$$

Los x_i para los cuáles $a_i > 0$ o $\hat{a}_i > 0$ son llamados vectores de soporte, que son los puntos donde el error está en los límites del e-tubo o por fuera. Los puntos dentro del tubo tienen $a_i = \hat{a}_i = 0$. En la predicción solo se evalúan los términos que involucran vectores de soporte. El parámetro w_0 se puede encontrar con un x_i para el cual $0 \leq \hat{a}_i \leq C$.

$w_o = y_i - \epsilon - \sum_{i=1}^{N}(a_i - \hat{a}_i)K(x, x_i)$ (Murphy, 2012, págs. 496-497).

7.4 Redes Neuronales de perceptrón multicapa

Las redes neuronales utilizadas en la investigación de propagación hacia adelante están aplicadas a la regresión.

Se consideran modelos para regresión y clasificación que representan una combinación lineal de funciones base fijas. Es necesario adaptar las funciones base y los parámetros a los datos. Estos son adaptados durante la fase de entrenamiento. El más exitoso en el contexto de aprendizaje automático es el perceptrón multicapa o redes de propagación hacia adelante.

El término de redes neuronales tiene sus orígenes en una representación matemática de cómo los sistemas biológicos procesan la información (McCulloch & Pitts, 1943, Windrow & Huff 1960, Rosenblatt, 1962, Rumelhart et. al, 1986) citados por (Bishop, 2006)

Un modelo de regresión lineal está basado en las combinaciones lineales de funciones de base no lineales $\varphi_i(x)$ de la forma:

$$y(x, w) = f \sum_{j=1}^{M} w_i \, \phi_j(x)$$

Donde f es una función de activación no lineal. La meta es extender el modelo con el fin que las funciones base $\varphi_i(x)$ dependan de los parámetros y sean ajustadas con los coeficientes w_i durante la fase de entrenamiento.

Los modelos básicos de redes neuronales pueden describirse por una serie de transformaciones funcionales donde se construyen M combinaciones lineales de las variables de entrada $X_{i...}X_D$ de la forma:

$$a_j = \sum_{i=1}^{D} w_{ji}^{(1)} \, x_i + w_{j0}^{(1)}$$

Donde j=1,….., M y el superíndice (1) indica los parámetros correspondientes a la primera capa de la red, que refiere los parámetros $w_{ji}^{(1)}$ como los pesos y los parámetros $w_{j0}^{(1)}$ como las bases, las cantidades a_j son conocidas como activaciones, cada una de ellas se transforma utilizando una función de activación no lineal h que por lo general se utiliza la función sigmoidal, sigmoidal logística o tangente hiperbólica.

$$Z_j = h(a_j)$$

Estas cantidades corresponden a las salidas de la función base que en el contexto de redes neuronales se conocen como nodos ocultos.

Los Z_j se combinan de nuevo linealmente para dar activaciones de salida:

$$a_k = \sum_{j=1}^{D} w_{k,j}^{(2)} Z_j + w_{k,0}^{(2)}$$

Donde k=1,….,K y K es el número total de salidas, esta transformación corresponde a la segunda capa de la red y el superíndice (2) indica los parámetros correspondientes.

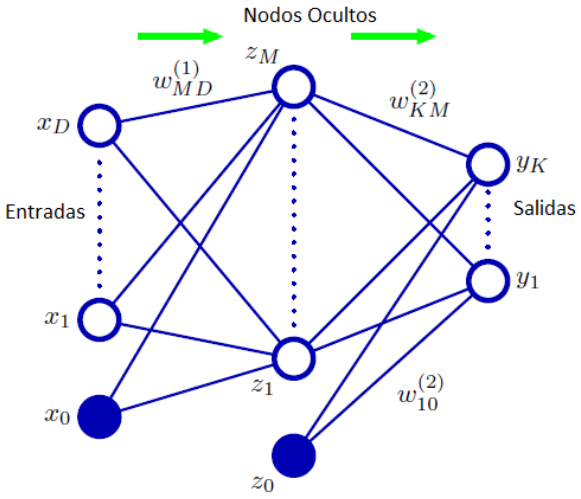


Figura 8. Red Neuronal de dos capas Fuente (Bishop, 2006, pág. 228)

Para un problema estándar de regresión, la función de activación es la identidad, entonces:

$$y_k = a_k$$

Se pueden combinar varias etapas que dan como resultado una función general de redes neuronales de la forma:

$$y_k(x, w) = \delta\left(\sum_{j=1}^{M} w_{k,j}^{(2)}\, h\left(\sum_{i=1}^{D} w_{ji}^{(1)}\, x_i + w_{j0}^{(1)}\right) + w_{k,0}^{(2)}\right)$$

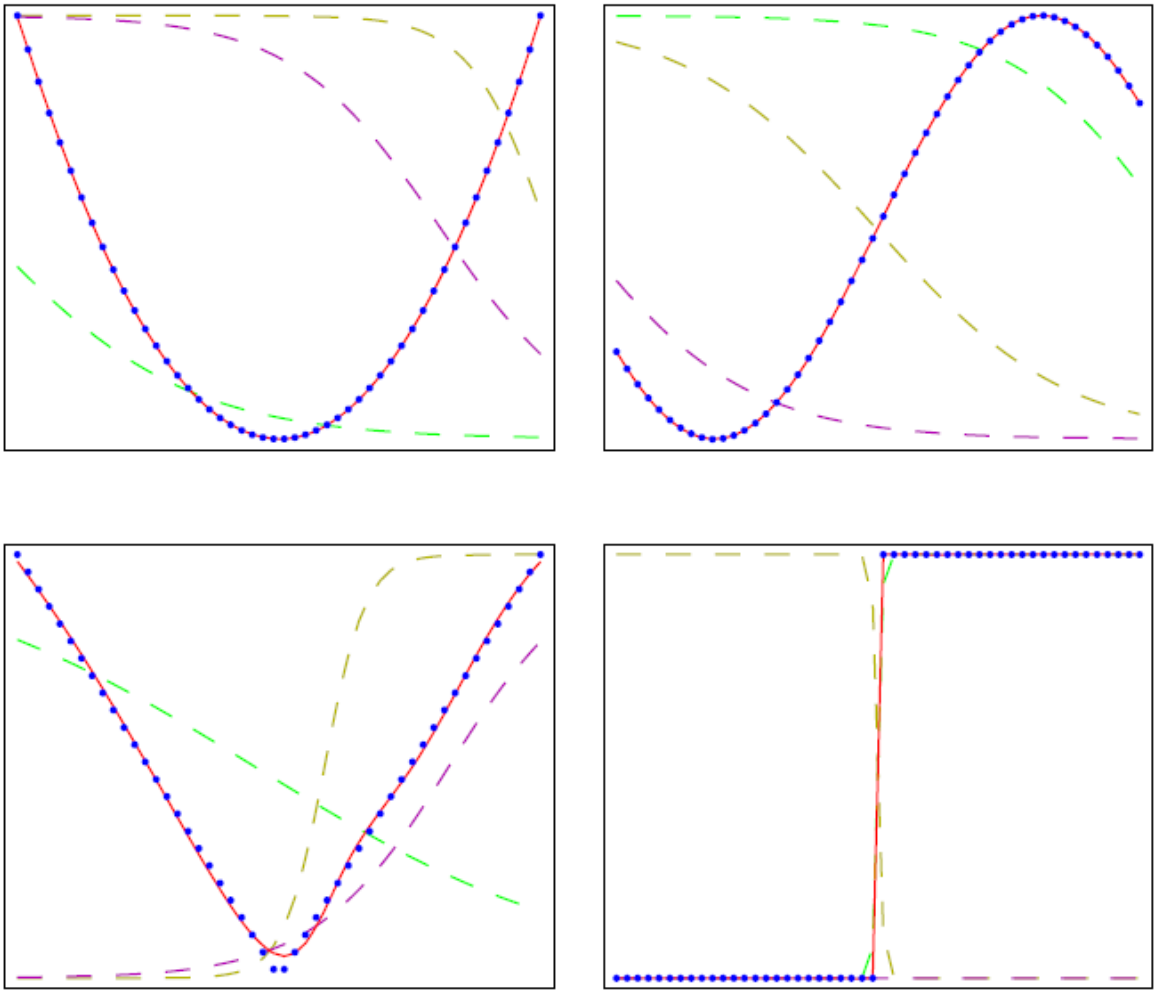
Con $x_0 = 1$ toma la forma:

$$= \delta\left(\sum_{j=0}^{M} w_{k,j}^{(2)}\, h\left(\sum_{i=1}^{D} w_{ji}^{(1)}\, x_i\right)\right)$$

Donde todos los pesos y parámetros bias son agrupados en un vector w. El proceso anterior puede ser interpretado como una propagación hacia adelante.

7.4.1 Propiedades de aproximación

Las propiedades de aproximación de la red de la propagación hacia adelante son ampliamente estudiadas, y son llevadas a aproximaciones universales.



Figuras 9,10,11,12. Ilustra la capacidad del perceptrón multicapa de aproximar diferentes funciones fuente (Bishop, 2006, pág. 231)

7.4.2 Entrenamiento de la red

El problema principal consiste en encontrar los valores de los parámetros adecuados w, dado un conjunto de datos de entrenamiento para minimizar la función de error cuadrático medio.

$$E(w) = \frac{1}{2}\sum_{n=1}^{N} \|y(X_n, w) - t_n\|^2$$

7.4.3 Optimización de los parámetros

El objetivo principal es encontrar un vector de pesos w que minimice la función $E(w)$ para ello se puede utilizar la optimización por gradiente descendente, donde se busca encontrar una solución analítica de la ecuación:

$$\nabla E(w) = 0$$

Esta técnica selecciona un valor inicial w(0) y se va moviendo a través del espacio sucesivamente de la forma:

$$w^{(T+1)} = w^T + \Delta w^T$$

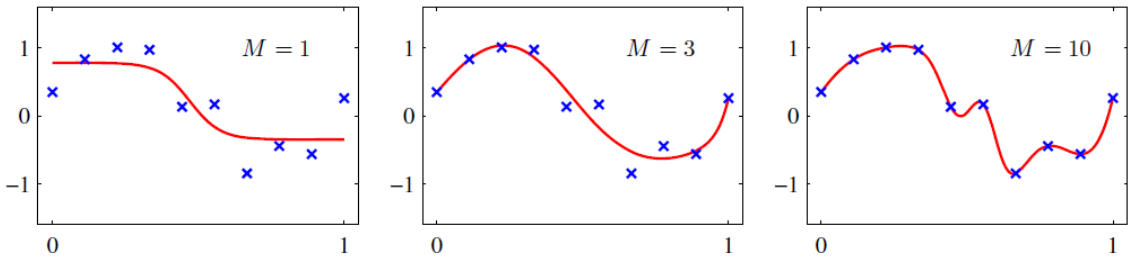
$$\Delta w_{i,j} = -\eta \frac{d\epsilon}{dw_{i,j}}$$

Donde T es el paso de la iteración y $\eta > 0$ se conoce como la razón de aprendizaje. Otros métodos de optimización incluyen gradiente conjugado o los métodos de Cuasi-Newton.

7.4.4 Regularización de redes neuronales

El número de unidades de entrada y salida de una red neuronal está determinado por la dimensionalidad de los datos N y por el número M de nodos ocultos que es un parámetro que se puede ajustar para mejorar el desempeño.

El parámetro M determina el número de pesos y bias de la red.



Figuras 13,14,15. Entrenamiento de la red con 1, 3 y 10 nodos ocultos Fuente (Bishop, 2006, pág. 257)

Sin embargo la generalización no es una función simple de M, dada la presencia de mínimos locales en la función de error. Existen otras maneras de controlar la complejidad del entrenamiento de la red neuronal con la intención de evitar el sobreentrenamiento. La regularización más simple es de la forma:

$$\tilde{E}(w) = E(w) + \frac{\lambda}{2} w^T w$$

La regularización es también conocida como “weigth decay”. La complejidad del modelo es determinada por la escogencia del coeficiente de regularización λ . Se requiere que la

regularización sea invariante a transformaciones lineales de las entradas o salidas. Está dada por:

$$\frac{\lambda_1}{2}\sum_{w\epsilon w_1} w^2 + \frac{\lambda_2}{2}\sum_{w\epsilon w_2} w^2$$

Donde w^1 denota el conjunto de pesos de la primera capa y w^2 denota el conjunto de pesos de la segunda capa y el bias es excluido.

7.4.5 Detención o criterio de parada

Una alternativa de regularización es controlar el criterio de parada de los algoritmos de optimización utilizados en el entrenamiento. Estos dan un error que es una función no creciente del índice de la iteración, sin embargo, en un conjunto de validación, el error medio decrece al principio y empieza a crecer cuando la red identifica un sobreentrenamiento. El entrenamiento puede detenerse en el punto de menor error con respecto al conjunto de validación con el fin de obtener una red con buena capacidad de generalización (Bishop, 2006, págs. 225-284).

7.6 Validación cruzada para selección de modelos

En orden de encontrar el mejor modelo teniendo un conjunto de datos de validación y un conjunto de datos de entrenamiento se puede utilizar validación cruzada (Bishop, 2006, pág. 33).

La técnica de validación cruzada consiste en dividir los datos de entrenamiento en k grupos, entonces para cada k que pertenece a (1,…,k), se realiza el entrenamiento sobre todos los grupos exceptuando el k-esimo grupo que se utiliza para la validación.

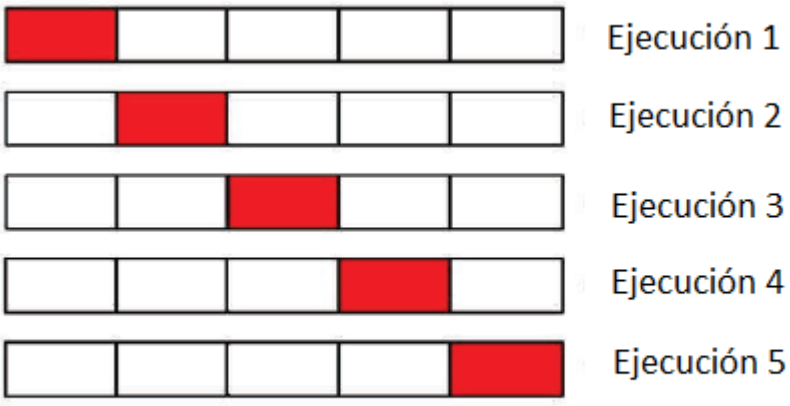


Figura 16. Validación cruzada con 5 grupos Fuente (Murphy, 2012, pág. 24)

Como la figura anterior se realizan las mismas n iteraciones variando los grupos de validación y los grupos de entrenamiento, se calcula el promedio de error de todos los grupos y se selecciona el modelo con mejor desempeño (Murphy, 2012, pág. 24).

8. Marco experimental

Para la selección del mejor algoritmo se tomaron como punto de comparación las siguientes métricas:

Error Absoluto Medio: Es la sumatoria de la diferencia del avalúo catastral estimado $\hat{Y}_i$ con el avalúo catastral real Y_i para las n observaciones del conjunto de datos.

$$EAM = \frac{1}{n}\sum_{i=1}^{n}\left|(\hat{Y}_i - Y_i\right|$$

Error Cuadrático Medio: Es la sumatoria de la diferencia del avalúo catastral estimado $\hat{Y}_i$ con el avalúo catastral real Y_i , y elevando al cuadrado la diferencia para las n observaciones del conjunto de datos.

$$ECM = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2$$

Coeficiente de determinación: Es uno menos la razón de la sumatoria de la diferencia al cuadrado del avalúo catastral estimado $\hat{Y}_i$ con el avalúo catastral real Y_i al cuadrado contra la sumatoria de la diferencia del avalúo catastral estimado $\hat{Y}_i$ con el valor medio del avalúo real $\bar{Y}$ al cuadrado.

$$r^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2}{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}$$

Donde $\hat{Y}_i$ es el valor del avalúo catastral estimado por el algoritmo de aprendizaje de máquina, Y_i es el valor real del avalúo catastral y $\bar{Y}$ es el valor promedio del avalúo catastral real. El valor de r^2 toma valores entre 0 y 1, y un valor de r^2 alto y valores bajos de EAM y ECM indican un mejor algoritmo. (Khamis & Binti Kamarudin, 2014).

Para determinar el mejor modelo de cada algoritmo se utilizó la métrica de comparación ECM seleccionando el modelo con menor valor del indicador.

Con el fin de ejecutar el procedimiento se partió de la base de datos predial de 58.000 predios y 54 características de los predios en la ciudad de Pereira y se realizaron dos tipos de experimentos que se describen a continuación.

8.1 Experimento 1

Se realizó una selección de cinco muestras aleatorias de 10.000 registros. A cada muestra aleatoria se le desarrolló validación cruzada anidada con cinco grupos, después fue estimado el error cuadrático medio (ECM) y el coeficiente de determinación (r2) para cada algoritmo, para la

estimación de parámetros se utilizó búsqueda en grilla con la siguiente combinación de hiperparámetros:

Tabla 3. Valores de los hiperparámetros de los algoritmos

Algoritmo	Parámetro	Conjunto de Valores
MSV(Máquina de soporte vectorial)	Constante de regularización $C = \frac{1}{\lambda}$	0.001, 0.001, 0.1, 1, **10**, 100
	Coeficiente Gamma de los kernel: RBF, poly, sigmoid	0.001, **0.01** ,0.1, 1, 10, 100
	Kernel	**RBF** $\boldsymbol{k(x, x') = e^{-(\frac{\|x-x'\|^2}{2\alpha^2})}}$ Poly $k(x, x') = (\alpha x^T x' + c)^d$ Sigmoid $k(x, x') = \tanh(\alpha x^T x' + r)$
RNA(Red Neuronal Artificial) Perceptrón multicapa	Función de Activación para la capa oculta	Función de Identidad: $f(x) = x$ **Función Logística sigmoidea :** $\boldsymbol{f(x) = \frac{1}{1 + e^{-x}}}$ Función Tangente Hiperbólica: $f(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$
	Número de neuronas	10, 20, 30, 40, 50, **80**,100
	Número de capas ocultas	1
	Alpha término de regularización	**0.001**,0.01,0.1,1
	Tasa de aprendizaje	0.001,0.01,0.1,1
RLM(Regresión Lineal Múltiple)	N.A.	N.A.
PG(Proceso Gaussiano)	Rational quadratic kernel	$\boldsymbol{k(xi, xj) = (1 + \frac{d(xi, xj)^2}{2\alpha l^2})}$

Los resaltados en color rojo fueron los parámetros más óptimos del algoritmo. En el algoritmo de procesos gaussianos no se hizo optimización utilizando combinación de parámetros porque el algoritmo realiza esta tarea durante la fase de entrenamiento.

8.2 Experimento 2

Se realizó una selección de cinco muestras aleatorias de 10.000 registros, a cada muestra aleatoria, a cada muestra se le realizó validación cruzada anidada con cinco grupos. Después se estimó el ECM y el r^2 para cada algoritmo y se guardó el mejor modelo con los datos de entrenamiento, para la estimación de parámetros se utiliza búsqueda en grilla con la combinación de hiperparámetros de la tabla anterior. Al terminar el proceso se realizó la validación de cada algoritmo con los 58.000 predios. Se excluyeron los datos de entrenamiento identificados en el mejor modelo y finalmente se calculó el ECM y el r^2.

Para los dos experimentos anteriores se calculó el promedio de ECM y r^2 de todas las iteraciones, el cual sirvió como indicador para determinar el mejor algoritmo. Entre mayor es el r^2 y menor es el ECM mejor es el desempeño del algoritmo.

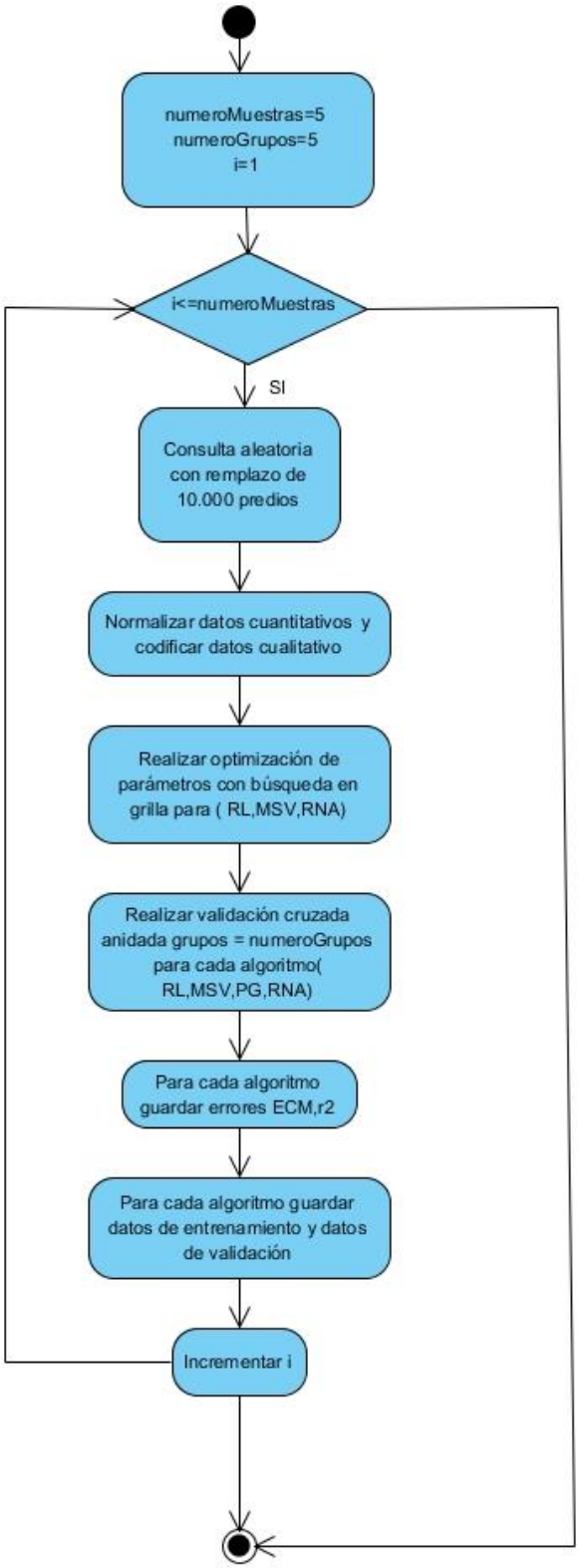


Figura 17. Diagrama de flujo experimento 1

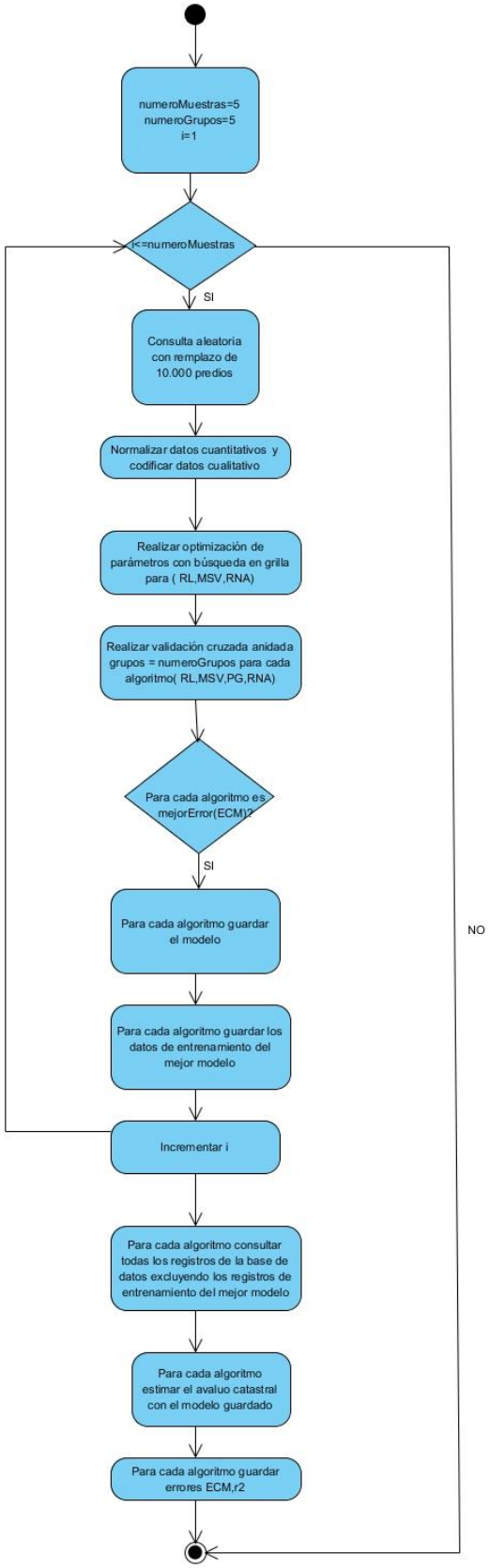


Figura 18. Diagrama de flujo experimento 2

Capítulo IX

9. Análisis de Resultados

9.1 Análisis Exploratorio de las variables

9.1.1 Descripción de las variables

Las variables seleccionadas para el estudio fueron proporcionados por el Instituto Agustín Codazzi seccional Pereira (Risaralda), al final del documento se presenta un cuadro con su respectiva descripción (Anexo).
El total de la población es de 73.078 registros y 82 variables, que corresponden a la totalidad de los predios urbanos y sus características registrados en el censo catastral año 2013.

9.1.2 Detección de valores atípicos, depuración de variables y de la base de datos

La muestra inicial para el estudio es de 73.078 predios urbanos que fueron sometidos a un proceso de depuración donde se descartaron variables repetidas y predios con datos ausentes o información incompleta y se eliminaron registros con datos atípicos. El proceso se inició con una transformación matemática del avalúo del predio, en el que se dividió el avalúo en 1.000.000 y luego se dividió entre el área del terreno del predio, lo anterior para obtener el avalúo por metro cuadrado, dado que se afectó el avalúo del predio que es nuestra variable objetivo con una variable independiente que es el área del terreno, se eliminó esta variable y las variables similares del estudio.

De esta forma se eliminaron las siguientes variables: perímetro del predio, área calculada, área del terreno. Luego se procedió a identificar y eliminar las variables que sean similares, por ejemplo el código del barrio y nombre del barrio que representan la misma identidad. Con base en lo anterior se eliminaron las siguientes variables con el mismo comportamiento: nombre del barrio, nombre de la comuna, nombre del plan parcial, nombre de la vía más cercana, nombre de la estación de transporte más cercana, nombre de la centralidad urbana más cercana, nombre del equipamiento urbano más cercano, nombre de la estación de servicio más cercano, nombre de la instalación crítica más cercana.

También se identificaron y eliminaron las siguientes variables que no disponían de suficiente información: censo_arboles_dist_min (Distancia mínima a los árboles), censo_arboles_gid (Distancia mínima a los árboles).

A continuación se presentan los previos (registros) eliminados por datos ausentes o información incompleta y datos atípicos, para la eliminación de datos atípicos se utilizó la técnica de estadística paramétrica donde los datos se ajustan a una función de distribución de probabilidad, en caso de

que el valor tenga una probabilidad muy baja será considerado como un dato atípico, basados en el teorema de la desigualdad de Chebyshev la probabilidad de que los valores aleatorios se dispersen de la media se distribuye de la siguiente forma: 68% de los datos se encuentran dentro de una desviación estándar de la media, el 95% de los valores se encuentran dentro de dos desviaciones estándar de la media y el 99,7 % de los valores se encuentran dentro de tres desviaciones estándar de la media (Knuth, 1997, pág. 107).

Para eliminar los datos atípicos del avalúo catastral se determinó una media de 71,53 millones y una desviación estándar de 117,19 millones, se eliminaron los datos de valores mayores a la media más dos veces la desviación estándar que corresponde a valores mayores a 305,92 millones.

$$Datos\ atípicos > \mu + 2\sigma$$

Tabla 4. Resumen de depuración de registros base de datos predial

Total de registros eliminados	Descripción
3266	Por ausencia del barrio
1749	Por ausencia de la comuna
72	Por ausencia del sector normativo
159	Por ausencia de delimitación de zona física
7147	Por ausencia del estrato
179	Por ausencia de información general
1983	Por datos atípicos el valor medio del predio mayor a 2 veces la desviación estándar
14555	**Total**

Después de la depuración de variables y datos se finalizó con un total 58.523 registros y 45 variables. Ver anexo B

9.1.3 Estadística descriptiva de variables más relevantes

A continuación se presentan los histogramas de las variables más importantes según el ranking del algoritmo Sequential Forward Selection (SFS). (Raschka, Recuperado de http://rasbt.github.io/mlxtend/user_guide/feature_selection/SequentialFeatureSelector, 2017), el cual se explica más adelante.

Para la variable área construida de la tabla 5. la mayoría de los predios se encuentra entre 76 y 125 metros cuadrados que corresponden a 19.742, seguido de 16.215 predios entre 26 y 75 metros

cuadrados y 10.049 entre 126 y 175 metros cuadrados, en total 46.006 predios que son el 78.6 % del total de datos.

El estrato del municipio de la figura 19. 21.634 predios están en estrato 2, seguido de 16.140 en estrato 3 y 8.904 estrato 1, los estratos 1, 2 y 3 representan el 79,8 % del total de datos.

El número de habitaciones de la primera construcción aparece en la tabla 6. 14.515 predios tienen 3 habitaciones, seguido de 11.827 predios con 4 habitaciones luego 9.713 predios con 2 habitaciones y 8.446 sin habitaciones, esto en total suma 35.515 predios que corresponde a 75,1 % del total de los datos

El número de pisos de la primera construcción se puede observar en la tabla 7. 26.585 predios tienen 2 pisos, 20.162 predios tienen 1 piso, seguido de 5.179 predios con 3 pisos. En total suma 51.926 que corresponde al 88,7 % de los datos.

Tabla 5. Histograma área construida del predio

Rango en metros cuadrado	Total de predios
0 - 25	5311
26 -75	16215
76-125	19742
126-175	10049
176-225	3561
226-275	1603
276-325	739
326-375	393
376-425	254
426-475	162
476-525	122
526-575	89
576-625	46
626-675	39
676-725	28
726-775	22
776-900	33
900-1500	82
1501-2500	28
>=2051	5

ÁREA CONSTRUIDA CATASTRO

Variable: Área Construida

Mínimo: 0

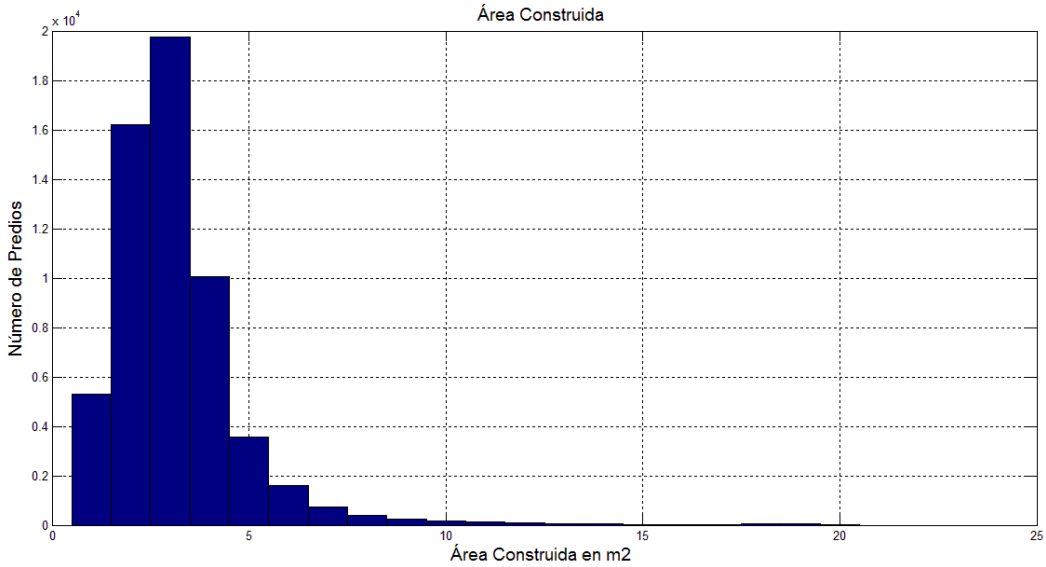
Máximo: 2884

Media: 108,626

Desviación. Estándar.: 98,85

Unidad: m2

Total = 58523



ESTRATO MUNICIPIO
Variable: Estrato municipio
Unidad: Categórica
Total = 58523

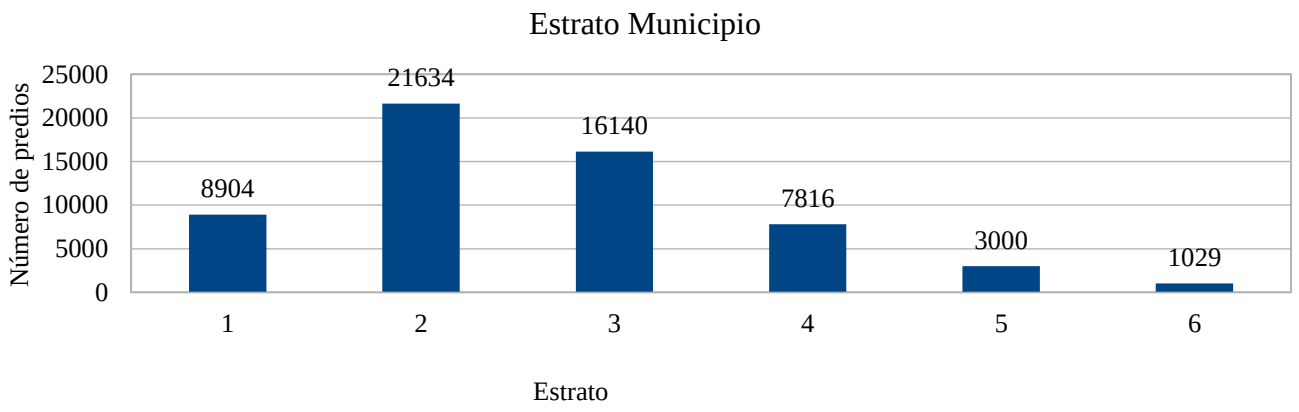


Figura 19. Gráfico de barras estrato del predio

Tabla 6. Histograma número de habitaciones del predio

Número de Habitaciones	Total de Predios
0	8446
1	3157
2	9713
3	14515
4	11827
5	5019
6	3513
7	981
8	723
9	307
10	143
11	34
12	116
20	22
40	5
60	1
80	1
100	0

NÚMERO DE HABITACIONES PRIMERA CONSTRUCCIÓN
Variable: Número de habitaciones primera construcción
Mínimo: 0
Máximo: 80
Media: 3,061
Desviación Estándar: 2,064
Unidad: Cantidad
Total = 58523

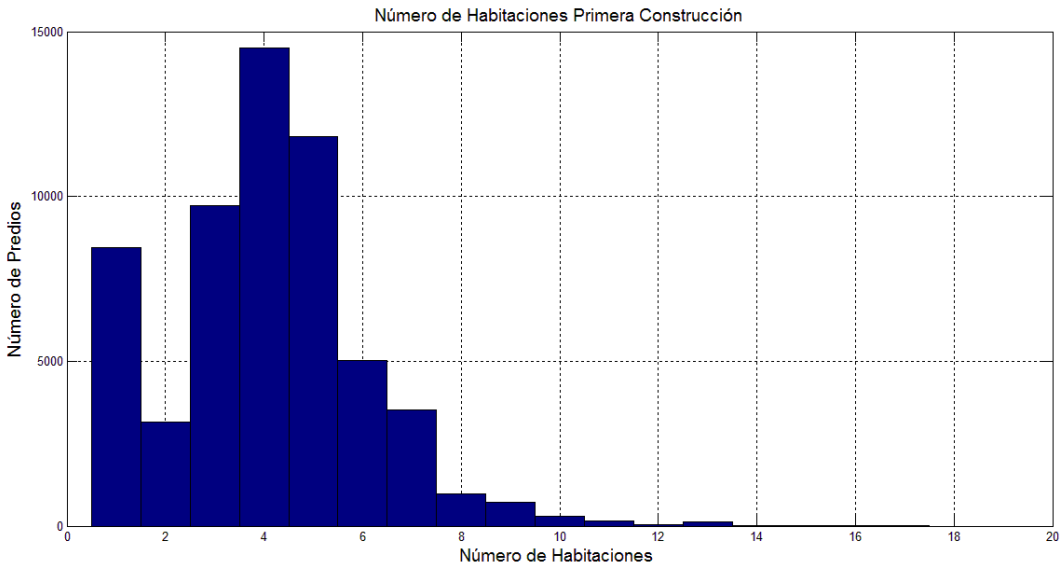


Tabla 7. Histograma número de pisos de la primera habitación

Número de Pisos	Total de Predios
0	6400
1	20162
2	26585
3	5179
4	175
5	13
6	2
7	2
8	0
9	0
10	0
11	1
12	1
20	2
40	1

NÚMERO DE PISOS PRIMERA CONSTRUCCIÓN
Variable: Número de pisos primera construcción
Mínimo: 0
Máximo: 33
Media: 1,534
Desviación Estándar: 0,838
Unidad: Cantidad
Total = 58523

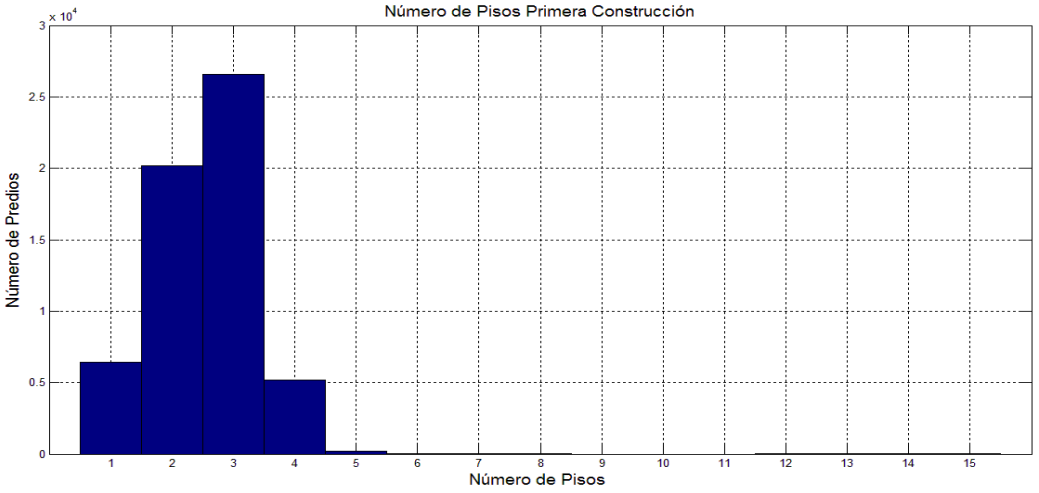
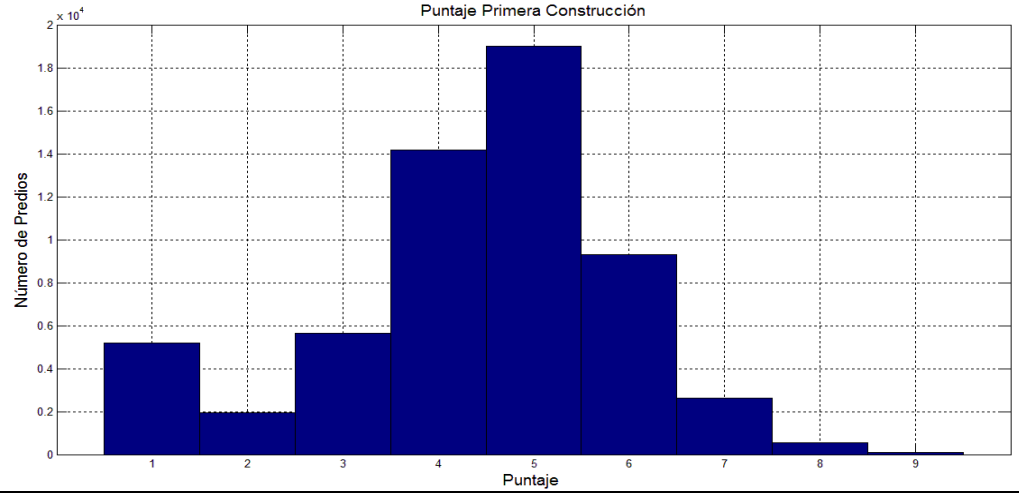


Tabla 8. Histograma puntaje de la primera construcción del predio

Rango Puntaje	Total Predios
0-5	5186
6-15	1955
16-25	5658
26-35	14186
36-45	19001
46-55	9306
56-65	2613
66-75	541
76-85	77

PUNTAJE DE LA PRIMERA CONSTRUCCIÓN
Variable: Puntaje primera construcción
Mínimo: 0
Máximo: 85
Media: 34,459
Desviación Estándar: 15,342
Unidad: Cantidad
Total = 58523



En la tabla 8 se observa el puntaje de la primera construcción con un máximo puntaje de 85 y un mínimo puntaje entre 0 y 5 puntos, la mayoría de predios están entre 26 y 55 puntos, donde el valor medio es de 34,4 puntos.

Ver el anexo C, para los histogramas de las otras variables relacionadas.

9.1.4 Resultados selección del mejor algoritmo y modelo para la estimación del avalúo catastral

El objetivo principal es realizar la comparación de modelos para la estimación del avalúo catastral en la ciudad de Pereira. Los algoritmos seleccionados para la comparación son los siguientes:

- Modelo MSV (Máquinas de Soporte Vectorial)
- Modelo RLM (Regresión Lineal Múltiple)
- Modelo PG (Procesos Gaussianos)
- Modelo RNA (Red Neuronal Artificial Perceptrón Multicapa)

Para realizar la selección del modelo se utilizó validación cruzada anidada en la que se realizó la combinación de cinco grupos y se optimizaron los avalúos por búsqueda en grilla (Raschka, Capitulo 6 Mejores prácticas para evaluación del modelo y optimización de hiperparametros, 2015, págs. 187-188).

9.1.5 Resultado experimento 1

Del total de registros 58.523 se tomaron cinco muestras aleatorias de 10.000 registros con reemplazo, a los cuales se les calculó el error cuadrático medio (ECM) y el error absoluto medio (EAM), finalmente se determinó el promedio de los cinco errores.

Tabla 9. Resultados de selección del modelo experimento 1 con la base de datos depurada sin datos atípicos, fuente elaboración propia

Algoritmo	ECM	ECM Desv. Est.	EAM	EAM Desv. Est	EAM(pesos)	Desv. Est	r2	Desv. Est
Modelo MSV	0,022	0,002	0,097	0,003	$ 97.113	0,003	85%	0,012
Modelo RLM	0,043	0,002	0,159	0,002	$ 159.320	0,002	70%	0,016
Modelo PG	**0,020**	**0,001**	**0,085**	0,002	**$ 84.847**	0,002	**86%**	**0,010**
Modelo RNA	0,024	0,002	0,103	0,003	$ 102.575	0,003	83%	0,015

Como se observa en la tabla anterior al comparar el error cuadrático medio, los errores de los algoritmos PG, MSV, RNA son muy parecidos con diferencias muy pequeñas. El mejor es el modelo PG. Para el error absoluto medio EAM y el coeficiente de determinación r^2, se ve claramente que el mejor algoritmo con una diferencia más significativa es el PG (Proceso Gaussiano).

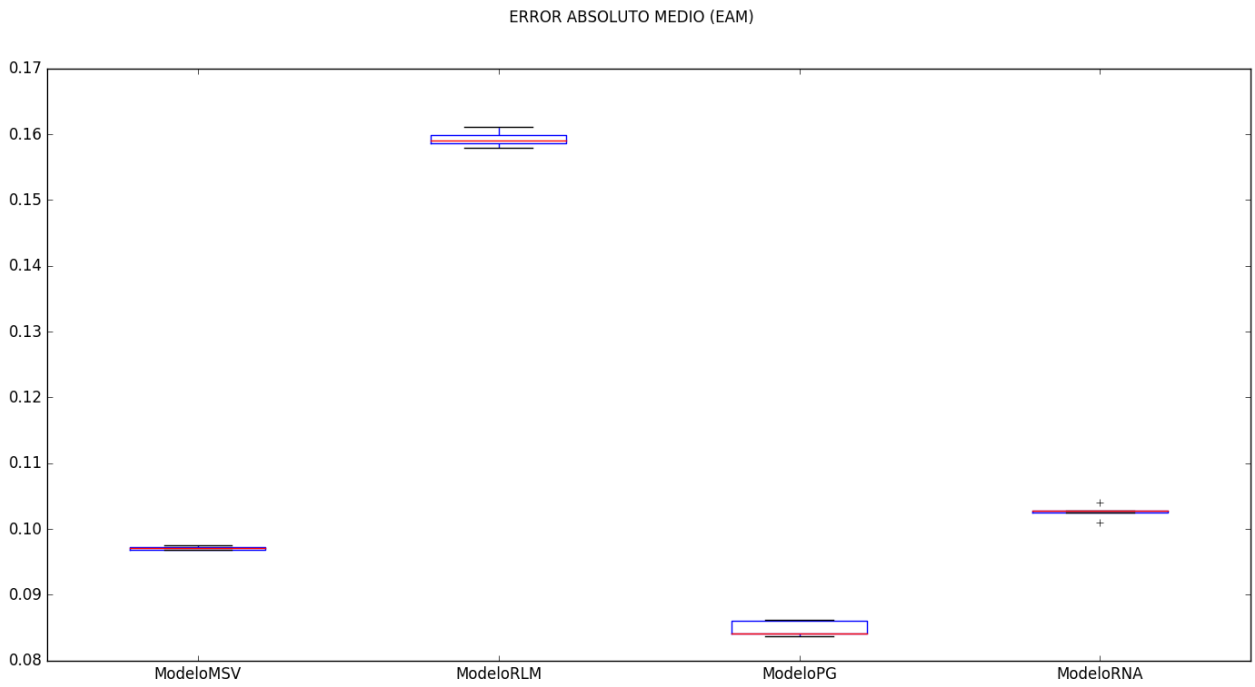


Figura 20. EAM (Error Absoluto Medio) con la base de datos depurada sin outliers, fuente elaboración propia

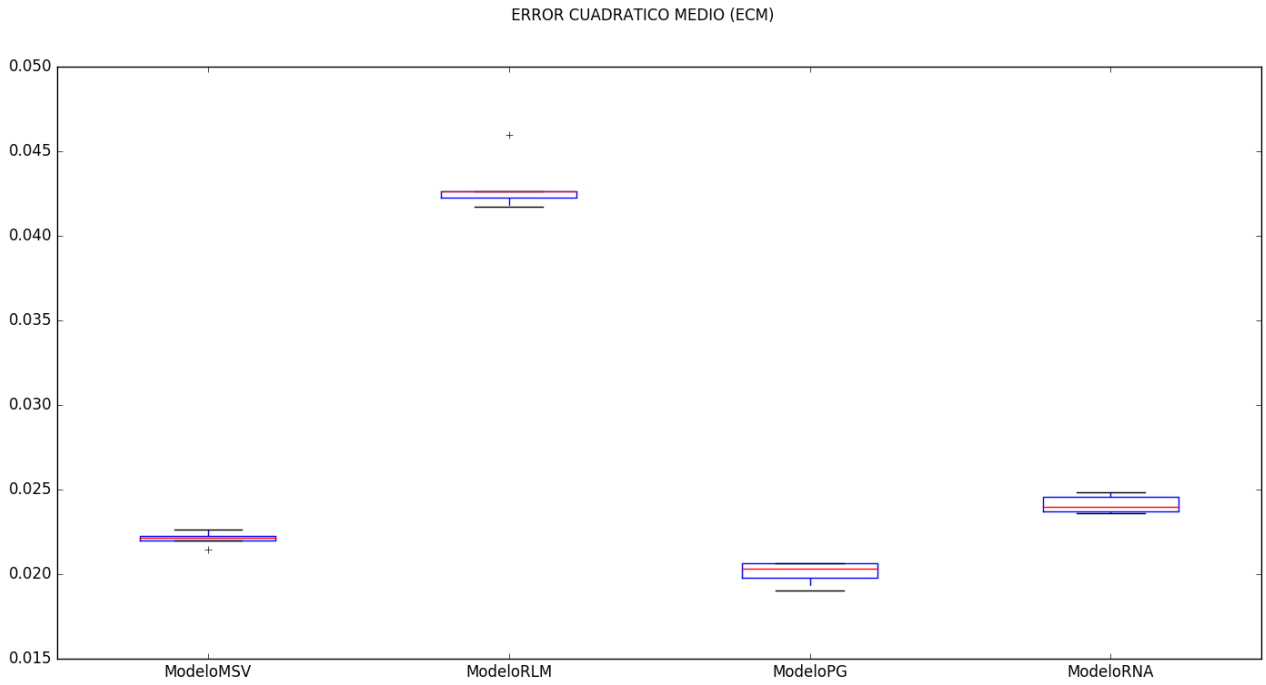
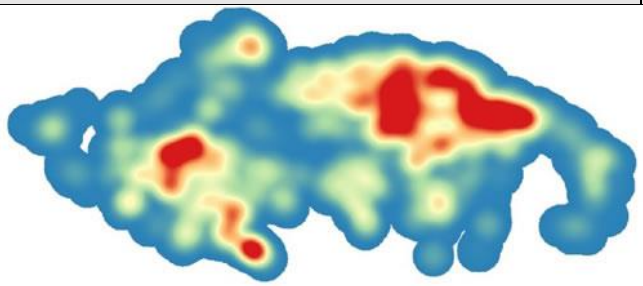
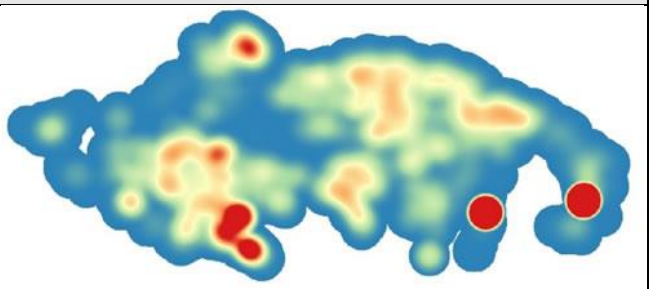
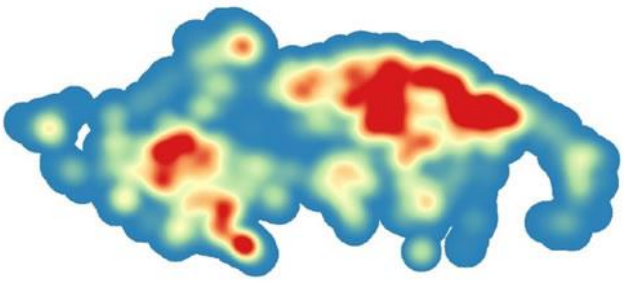
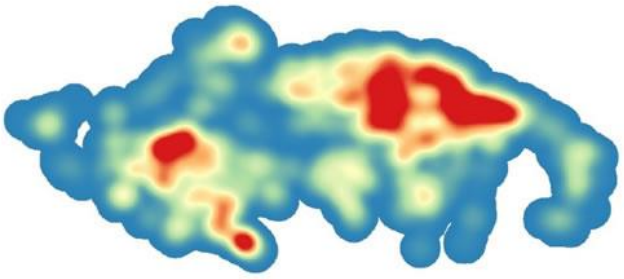
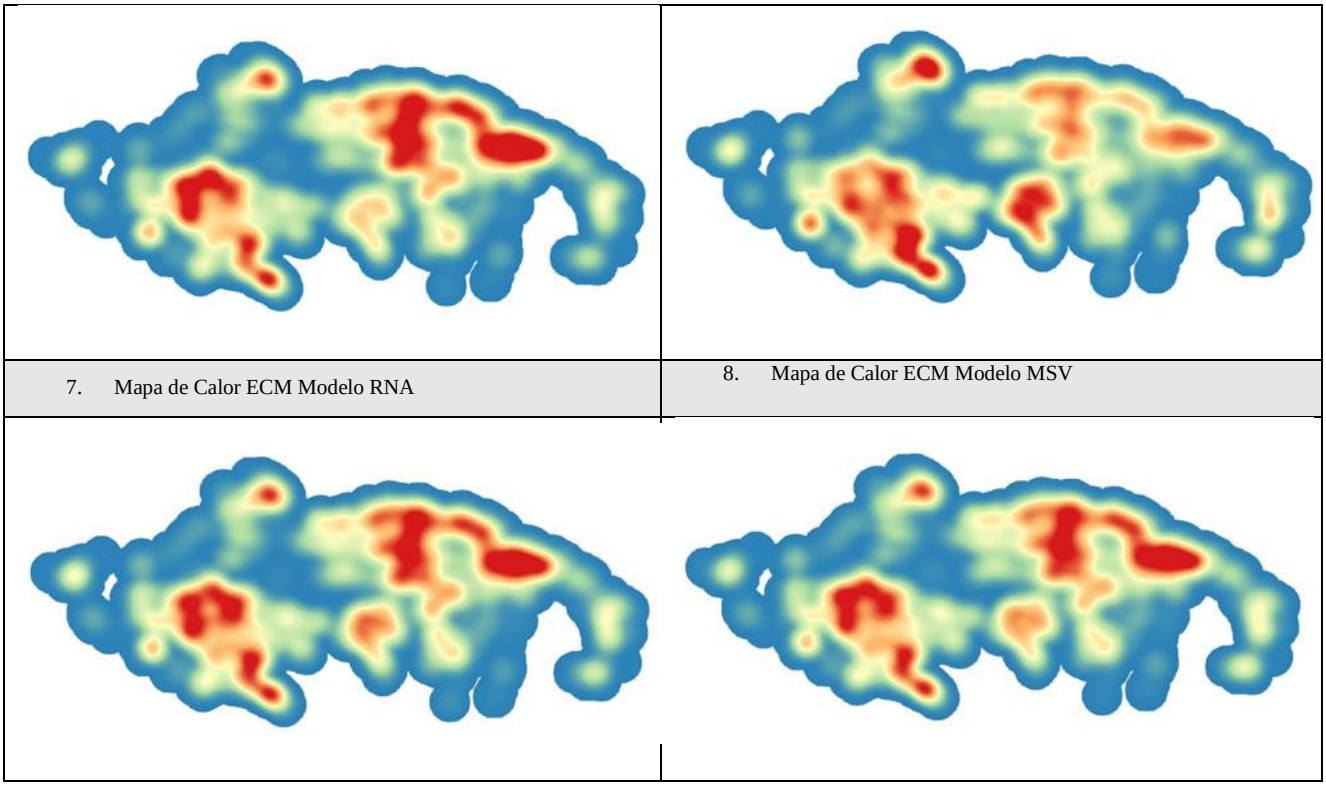


Figura 21. ECM (Error Cuadrático Medio) con la base de datos depurada sin outliers, fuente elaboración propia

1. Mapa de Calor EAM Modelo PG	2. Mapa de Calor EAM Modelo RLM
	
3. Mapa de Calor EAM Modelo RNA	4. Mapa de Calor EAM Modelo MSV
	
5. Mapa de Calor ECM Modelo PG	6. Mapa de Calor ECM Modelo RLM



7. Mapa de Calor ECM Modelo RNA

8. Mapa de Calor ECM Modelo MSV

Figura 22. Mapas de calor.

En la figura 22 se pueden observar gráficamente los resultados del experimento en una visualización en mapas de calor, los cuáles son una representación geográfica de la concentración de información en puntos geográficos específicos, en estos mapas se muestra la densidad de los puntos en un área determinada, a mayor densidad de puntos, el área quedará de un color más intenso, en este caso se vuelve más intenso el color rojo. A continuación se hace una descripción de los resultados, contextualizando en la realidad del territorio:

1. Mapa de Calor EAM Modelo PG, 3. Mapa de Calor EAM Modelo RNA, 4. Mapa de Calor EAM Modelo MSV

Distribución del Error Absoluto Medio (EAM) del modelo de Procesos Gaussianos (PG), Redes Neuronales Artificiales (RNA) y Máquinas de Soporte Vectorial (MSV) en el territorio, el error se ve con mayor concentración, a más color rojo mayor cantidad de error, se evidencian concentraciones bajas en los sectores Circunvalar, Villa santana y Cuba. Estos no son los lugares donde existe una mayor concentración de propiedades horizontales o construcciones en altura, lo que evidencia un fallo en el modelo aplicado para la estimación del avalúo catastral

2. Mapa de Calor EAM Modelo RLM

Distribución del Error Absoluto Medio (EAM) del modelo de Regresión lineal Múltiple (RLM) en el territorio, el error se ve con mayor concentración, a más color rojo mayor cantidad de error, se evidencian concentraciones en el sector centro tradicional de la ciudad, en el sector de

la circunvalar, sector de Cuba. Estos son los lugares donde mayor concentración de propiedades horizontales o construcciones en altura existen, lo que provoca una mayor concentración de avalúos por metro cuadrado. Este era el resultado esperado en una estimación de avalúos eficiente y veraz en el sector urbano de la ciudad de Pereira.

5. Mapa de Calor ECM Modelo PG, 6. Mapa de Calor ECM Modelo RLM,7. Mapa de Calor ECM Modelo RNA, 8. Mapa de Calor ECM Modelo MSV

Se evidencia la distribución del Error Cuadrático Medio (ECM), a mayor color rojo en el mapa, mayor es la concentración de error. Se evidencia una distribución del error similar para los modelos de PG, RNA y MSV concentrándose en los sectores de Centro, Cuba, Circunvalar y Parque Industrial. Esto debido a que en estos lugares es donde está la mayor concentración de propiedad horizontal y por consiguiente mayor cantidad de viviendas por metro cuadrado, estos tres modelos presentan una distribución del error de forma similar y es el resultado esperado si el modelo está haciendo una buena estimación del avalúo catastral.

El modelo de RLM presenta poca variación de ECM no concentrada, esto debido a que el modelo no está haciendo estimaciones correctas de avalúos catastrales en los sectores donde existe mayor concentración de viviendas en propiedad horizontal.

9.1.6 Resultado experimento 2

Del total de registros 58.523 se tomaron cinco muestras aleatorias de 10.000 registros con reemplazo a los cuales se les calculó el error cuadrático medio (ECM) y para cada algoritmo se guardó el mejor modelo y los datos de entrenamiento basado en el ECM. Luego se realizó una validación con toda la base de datos de la cual se excluyeron los datos de entrenamiento. Total datos de entrenamiento 8.000, datos de validación 50.523.

Tabla 10. Resultados de selección del modelo experimento 2 con la base de datos depurada sin datos atípicos,

Algoritmo	ECM	ECM Desv. Est	EAM	EAM Desv. Est.	EAM(pesos)	EAM (Pesos) DESV. EST.	r2
Modelo MSV	0,0523	0.0915	0,1742	0,1482	$ 174.123	$ 148.299	63,29%
Modelo RLM	0,0652	0.1262	0,2017	0,1565	$ 201.777	$ 156.543	54,94%
Modelo PG	**0,0412**	**0.0847**	**0,1465**	**0,1406**	**$ 146.520**	**$ 140.620**	**71,27%**
Modelo RNA	**0,0405**	**0.0990**	**0,1401**	**0,1444**	**$ 140.155**	**$ 144.477**	**71,59%**

Como se observa en los resultados anteriores el ECM de los modelos GP y RNA es igual y tienen diferencia significativa de los otros dos modelos. Para el error absoluto medio (EAM) y el coeficiente de determinación (r^2), se ve claramente que los mejores algoritmos son RNA y el PG a diferencia de los otros dos. Este experimento demuestra la capacidad de generalización de los algoritmos pues se estima el valor de 50.523 predios.

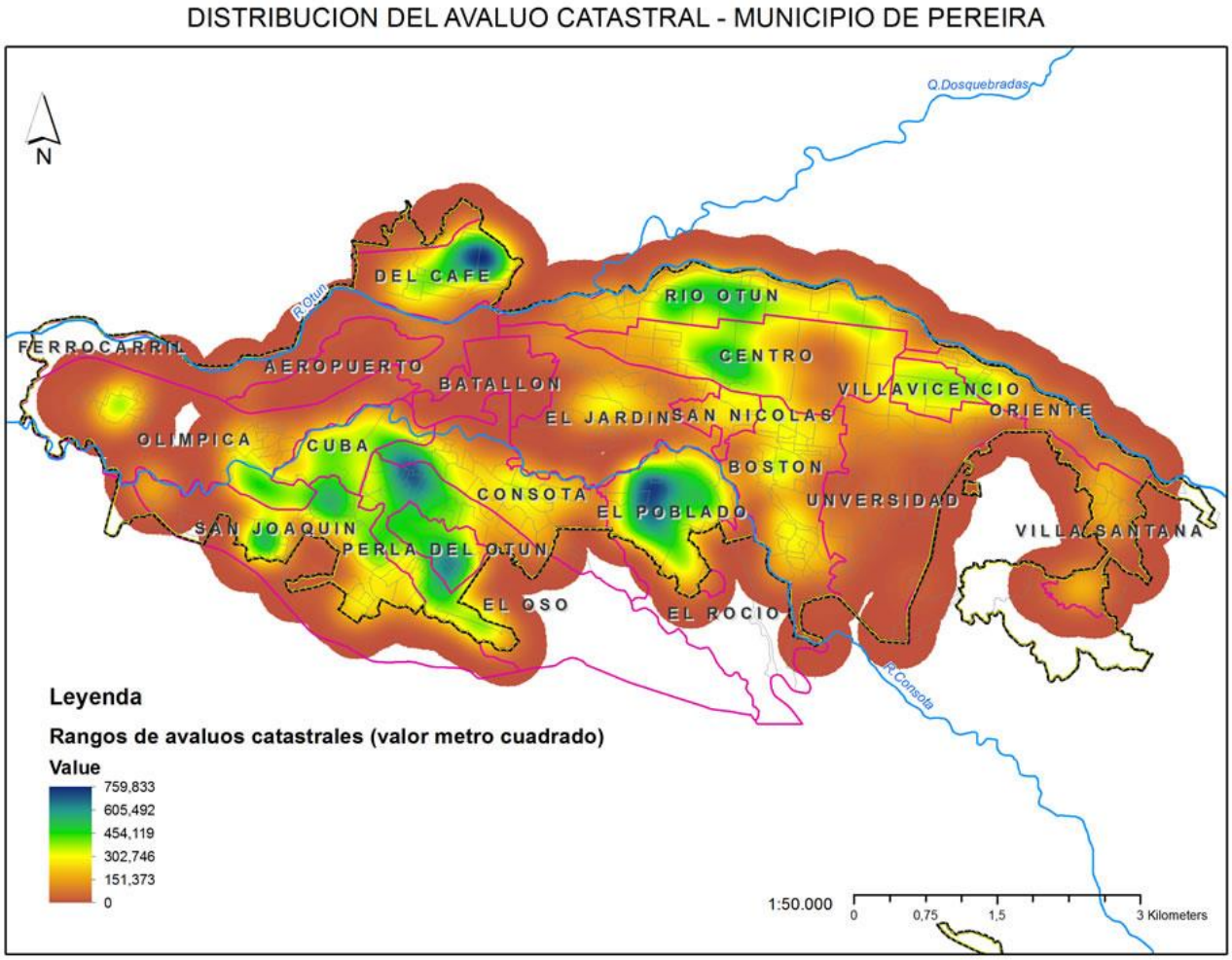


Figura 23. Mapa de calor avalúo catastral real ciudad de Pereira Fuente: Elaboración propia

En la figura 23 se observa la distribución de los avalúos en el municipio de Pereira, los puntos más azules representa los sectores de la ciudad donde el valor de las propiedades está más concentrado, estos son los resultados del proceso de avalúos realizado en el año 2013 por parte del IGAC utilizando las técnicas tradicionales de la metodología IGAC para la estimación de avalúos. Se observan concentraciones en el sector de Comuna del Café o Parque Industrial, el Poblado y Cuba y concentraciones más bajas en los sectores centro tradicional y sector Circunvalar.

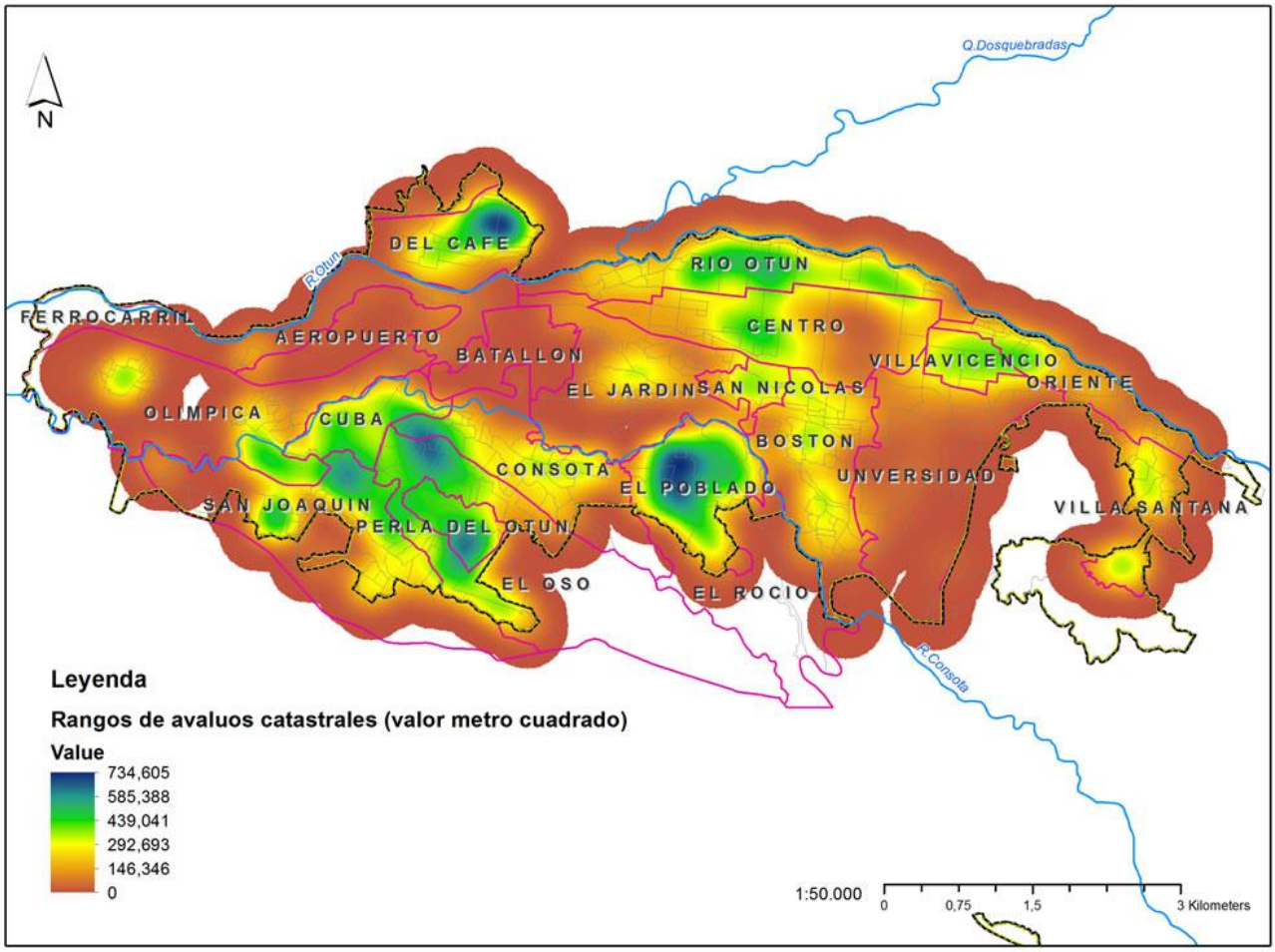


Figura 24. Mapa de calor avalúo catastral estimado modelo PG Fuente: Elaboración propia

En la figura 24 se observa la concentración de avalúos resultado de aplicar el modelo GP para la estimación de avalúos, el resultado es muy similar al de la figura 23, lo que significa que el modelo está haciendo una estimación acertada del avalúo catastral con concentraciones similares a las de los resultados reales de los avalúos hechos por el IGAC.

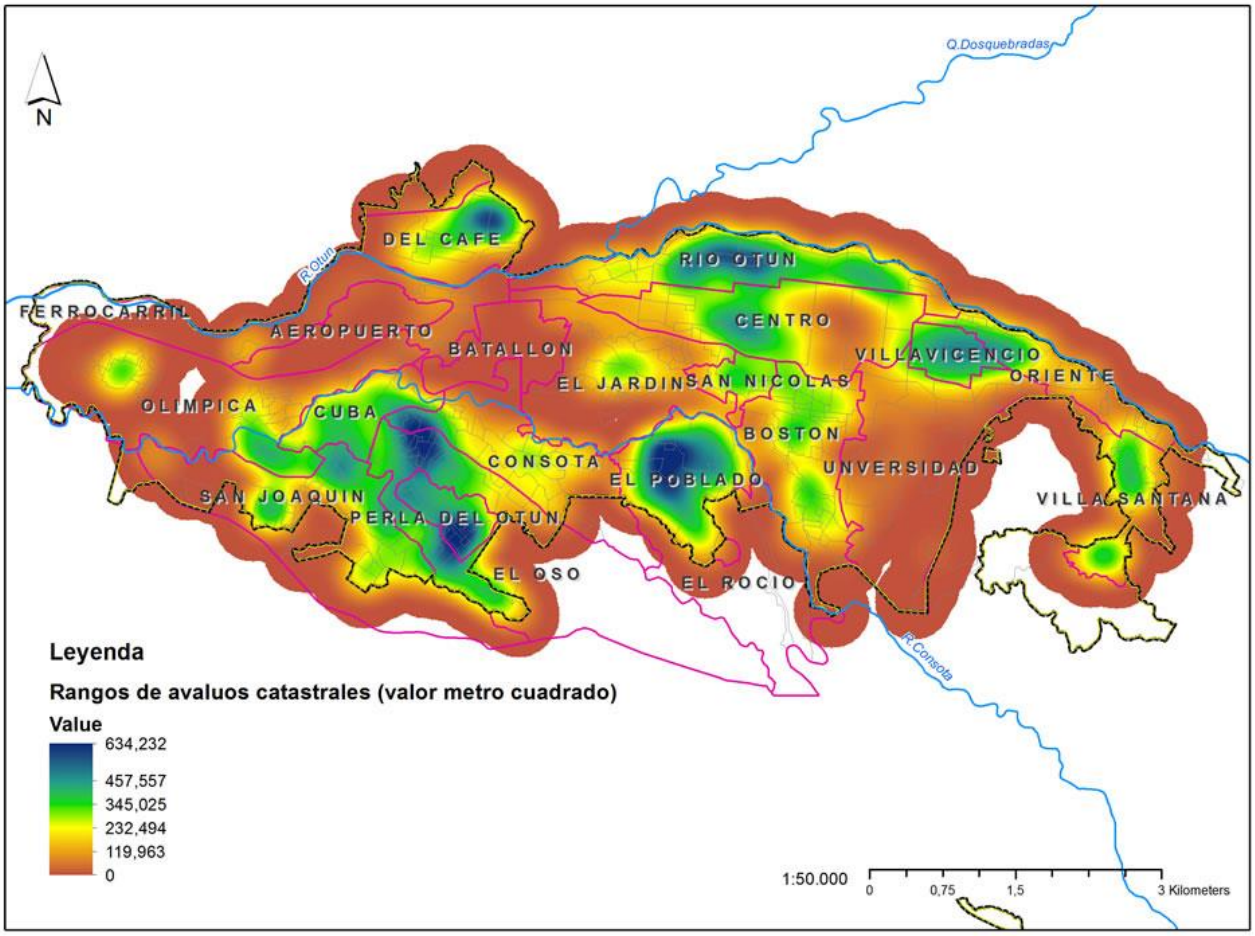


Figura 25. Mapa de calor avalúo catastral estimado modelo MSV Fuente: Elaboración propia

En la figura 25 se observa la concentración de avalúos resultado de aplicar el modelo MSV para la estimación de avalúos, el resultado es similar al de la figura 23, con algunas concentraciones pronunciadas en el sector de Cuba y el Poblado, lo que significa que el modelo no es tan acertado a la realidad de los avalúos expresada geográficamente en el mapa de calor de la figura 23.

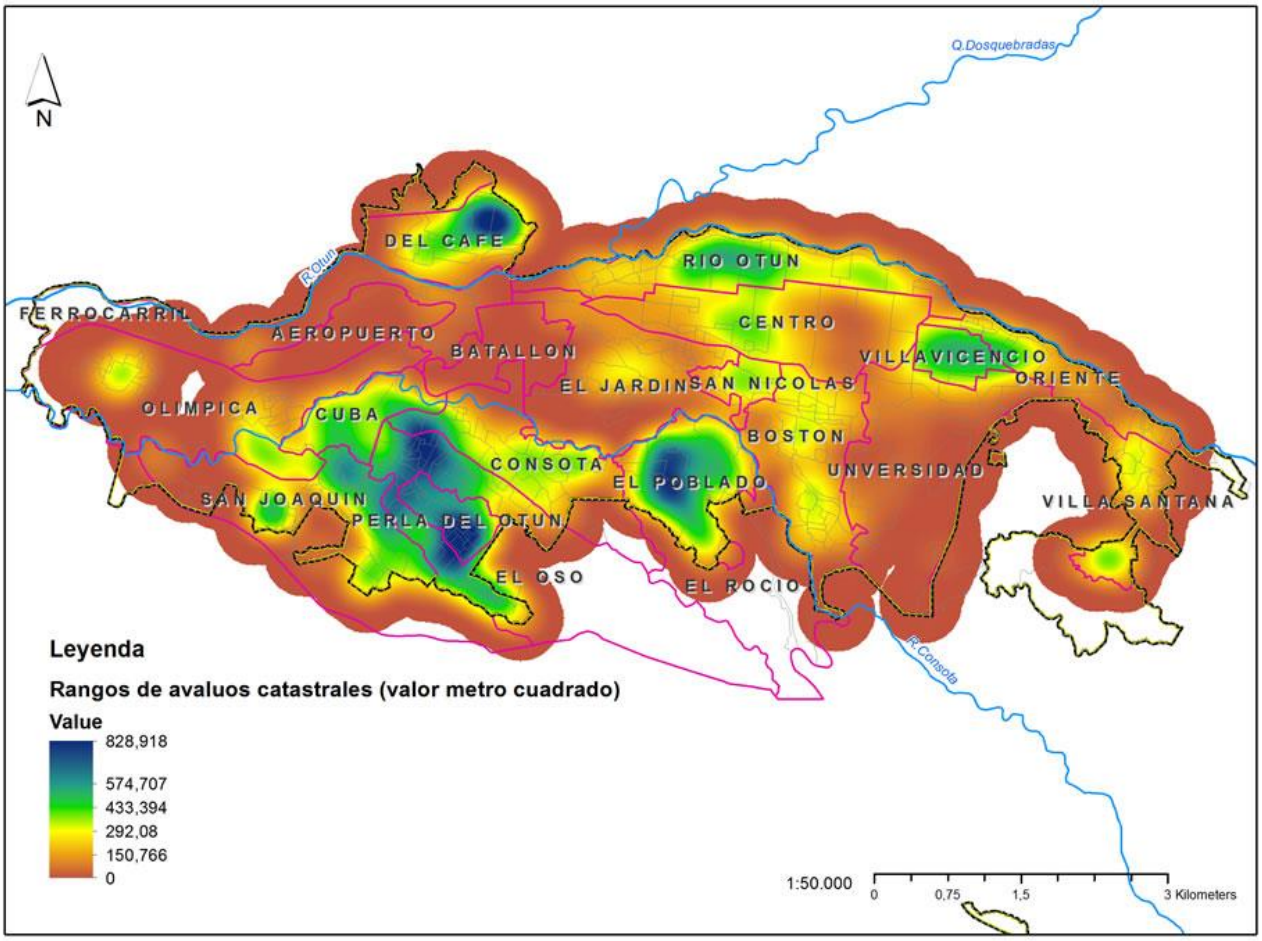


Figura 26. Mapa de calor avalúo catastral estimado modelo RL Fuente: Elaboración propia

En la figura 26 se observa la concentración de avalúos resultado de aplicar el modelo Regresión Lineal Múltiple (RLM) para la estimación de avalúos, el resultado es parecido al de la figura 23, con algunas concentraciones pronunciadas en el sector de Cuba, el Poblado y Parque Industrial, lo que significa que el modelo no es tan acertado a la realidad de los avalúos expresada geográficamente en el mapa del calor de la figura 23.

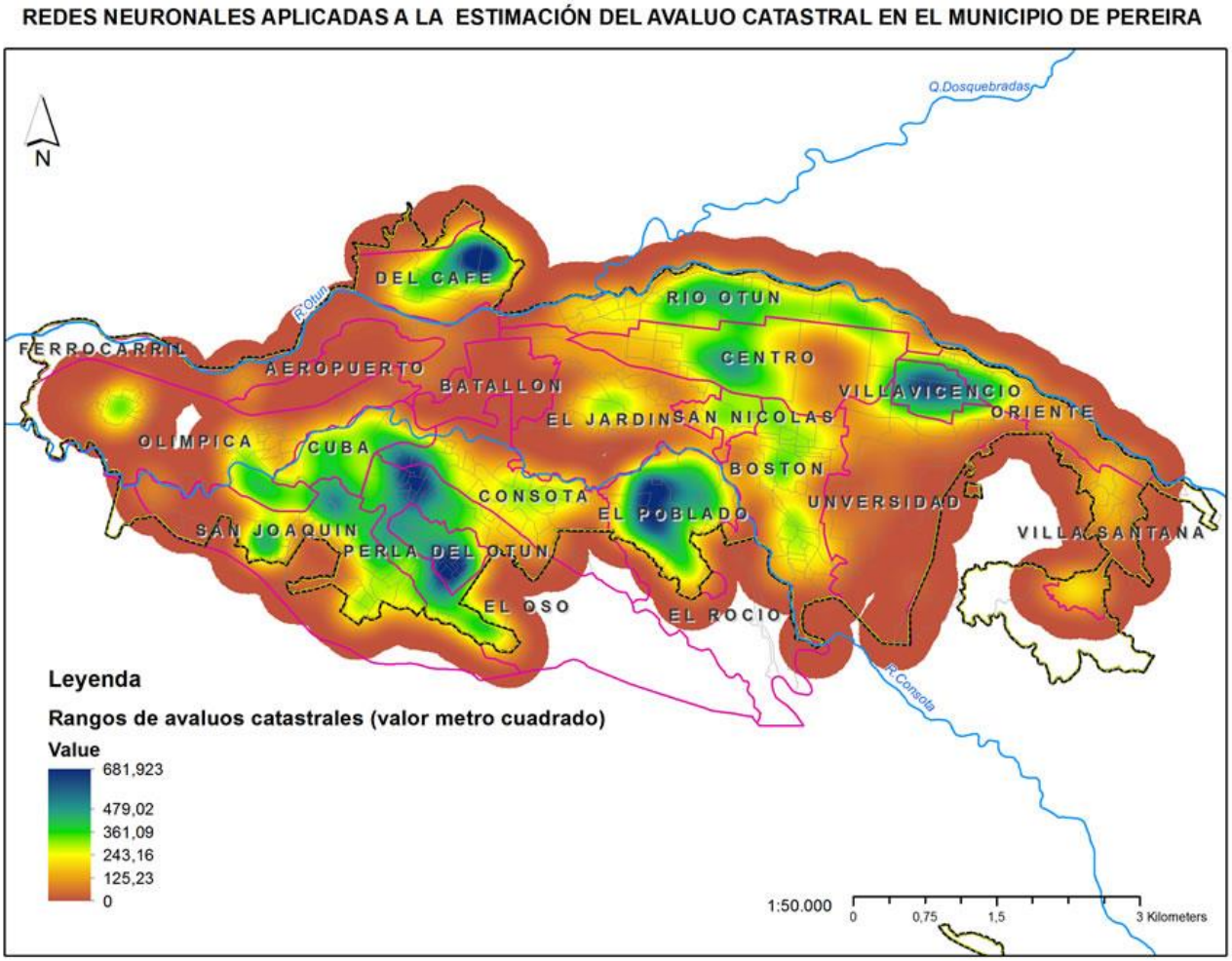


Figura 27. Mapa de calor avalúo catastral estimado modelo RNA Fuente: Elaboración propia

En la figura 27 se observa la concentración de avalúos resultado de aplicar el modelo RNA para la estimación de avalúos, el resultado es muy similar al de la figura 23, lo que significa que el modelo está haciendo una estimación acertada del avalúo catastral con concentraciones similares al de los resultados reales de los avalúos hechos por el IGAC, este modelo al igual que el de PG presentan resultados similares acertados.

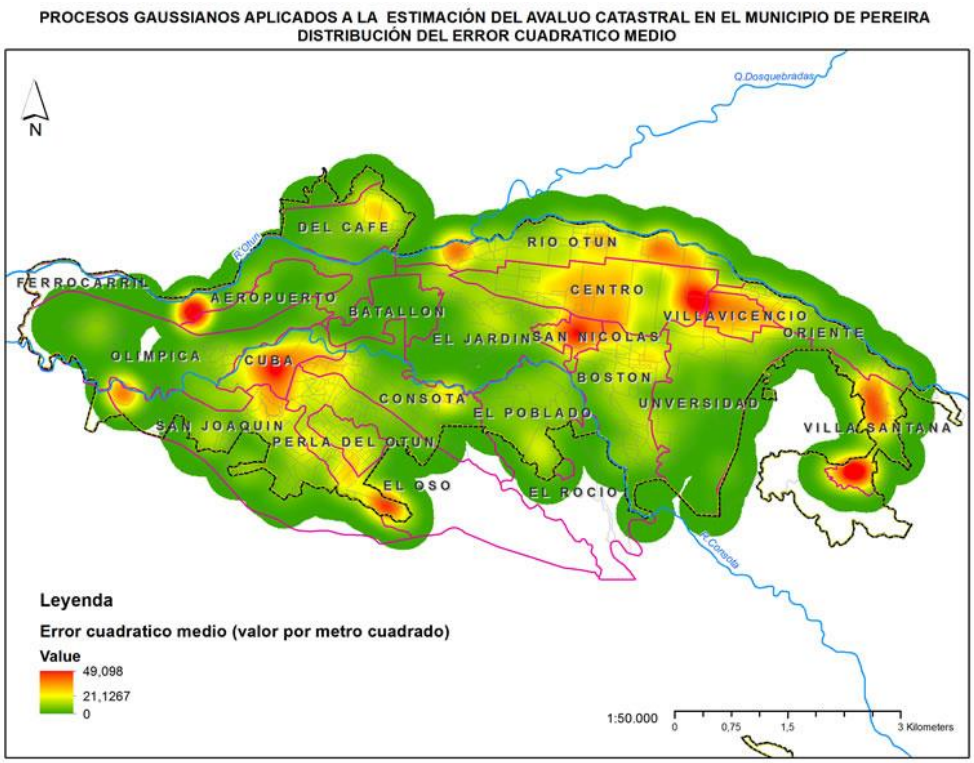


Figura 28. Mapa de calor Error Cuadrático Medio Modelo PG Fuente: Elaboración propia

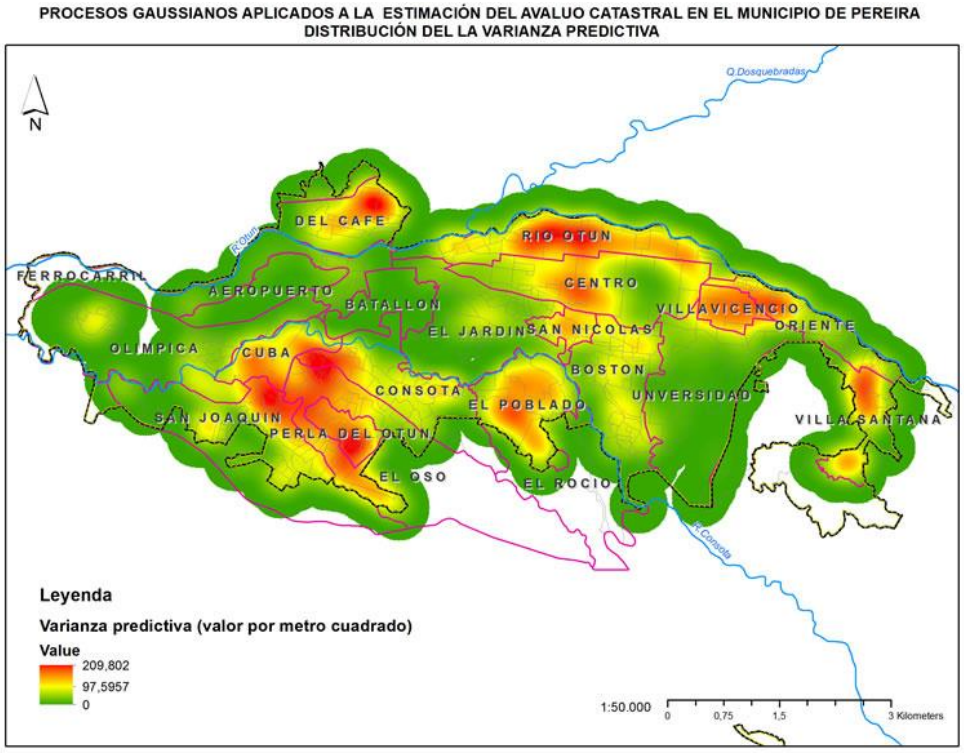


Figura 29. Mapa de calor Varianza Predictiva Modelo PG Fuente: Elaboración propia

En las figuras 28 y 29 se observa la concentración del error cuadrático medio ECM y la varianza predictiva resultado de la aplicación del modelo de GP. Se observa una concentración de error en el sector de Villa Santana y en un sector cerca al Aeropuerto, Cuba y el sector Circunvalar, en general el modelo está haciendo una buena estimación del avalúo catastral. La varianza predictiva tiene concentraciones que no son muy pronunciadas en los lugares donde el avalúo catastral está más concentrado, en los sectores de Cuba, Parque Industrial, Centro tradicional, el Poblado y la Circunvalar, estos serían los resultados esperados para un modelo que esté haciendo buenas estimaciones del avalúo catastral.

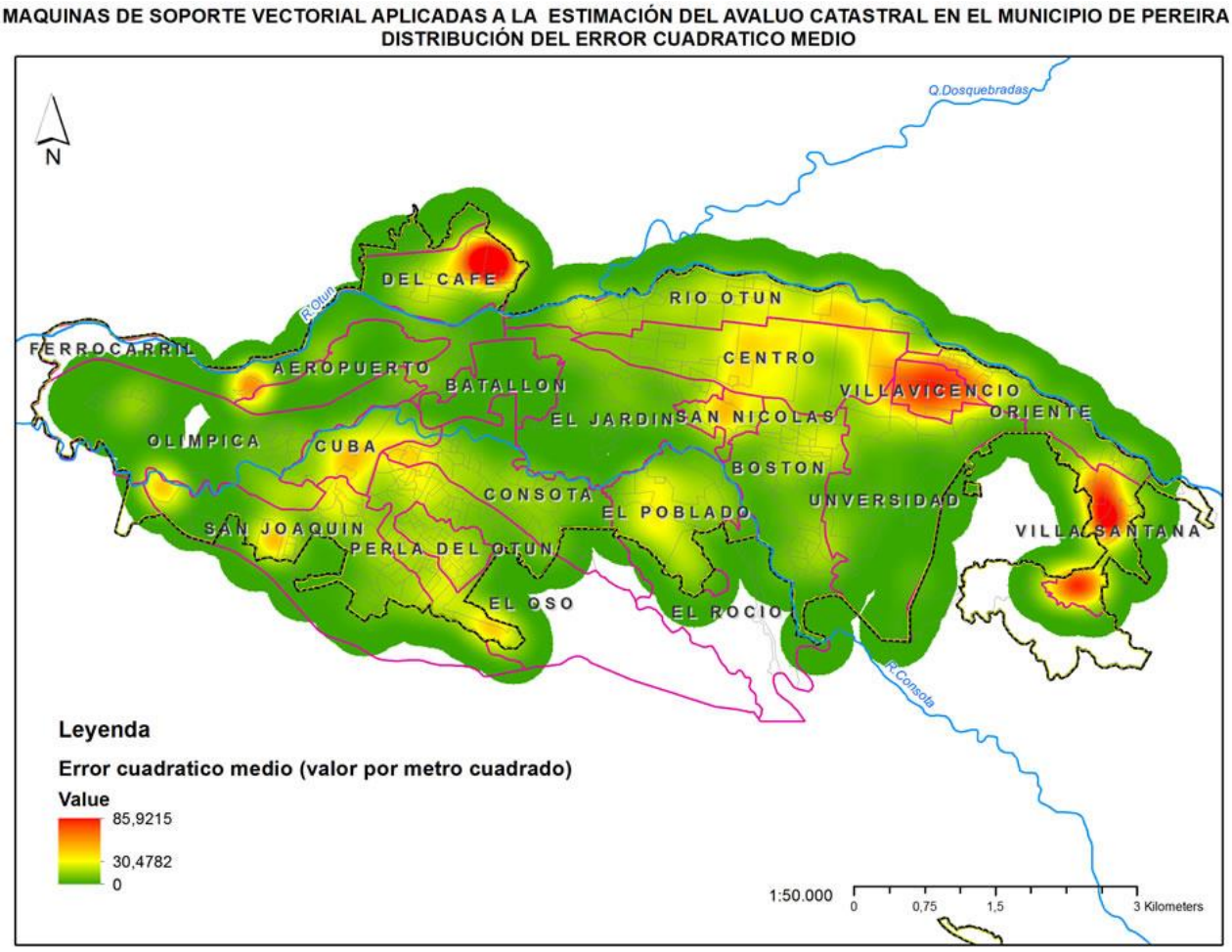


Figura 30. Mapa de calor Error Cuadrático Medio Modelo MSV Fuente: Elaboración propia

En la figura 30 se observa la concentración del error cuadrático medio ECM resultado de la aplicación del modelo de MSV, se observa una concentración de error en el sector de Villa Santana, la Circunvalar y el Parque Industrial. En general el modelo no presenta concentraciones de error en lugares donde está la mayor cantidad de avalúos por metros cuadrados. El modelo está haciendo una buena generalización de la estimación de avalúos catastrales.

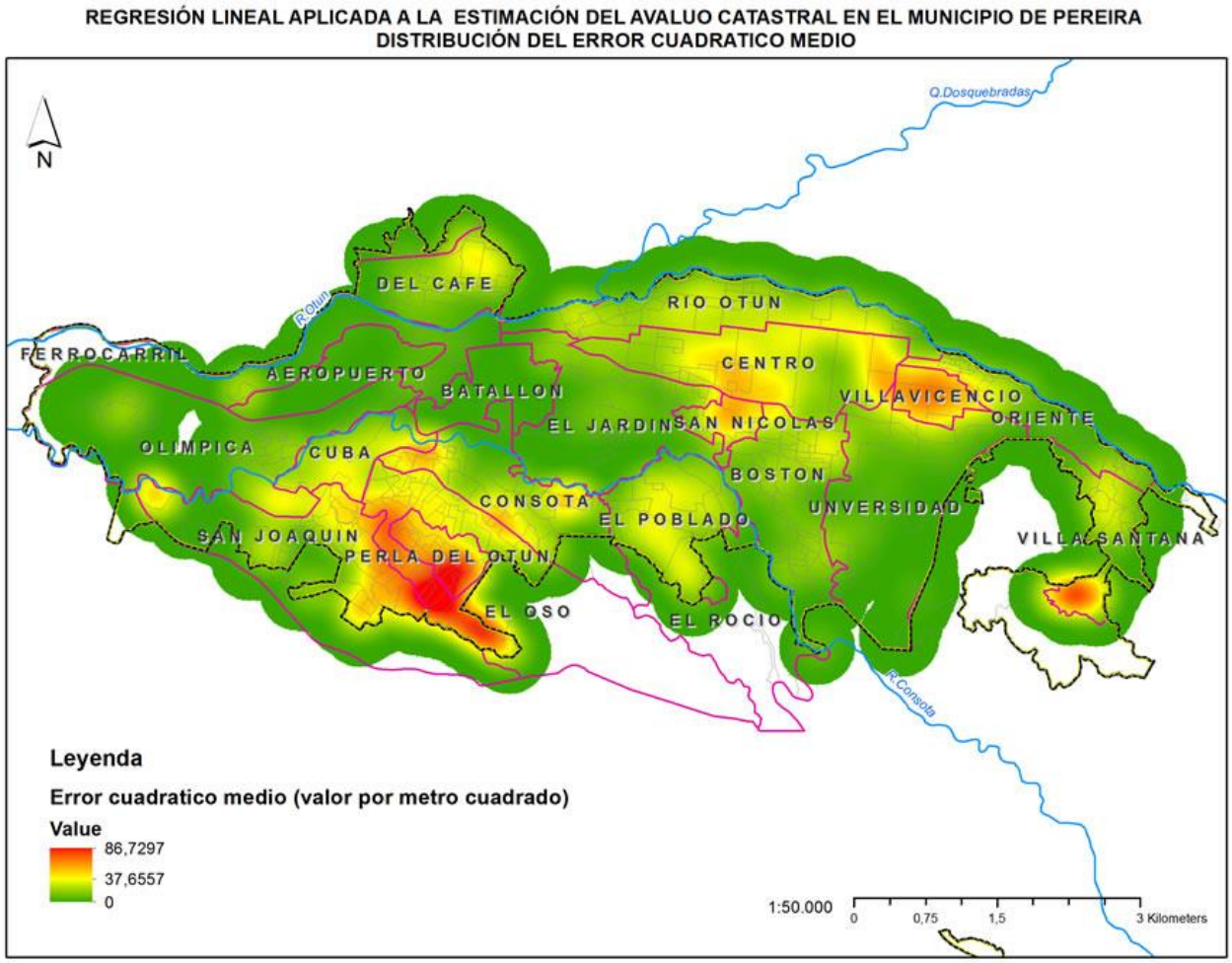


Figura 31. Mapa de calor Error Cuadrático Medio Modelo RL Fuente: Elaboración propia

En la figura 31 se observa la concentración del error cuadrático medio ECM resultado de la aplicación del modelo de RLM. Se observa una concentración de error en el sector de Cuba hacia la parte Sur Oriental. En general el modelo no presenta concentraciones de error en lugares donde está la mayor cantidad de avalúos por metros cuadrados. El modelo está haciendo una buena generalización de la estimación de avalúos catastrales.

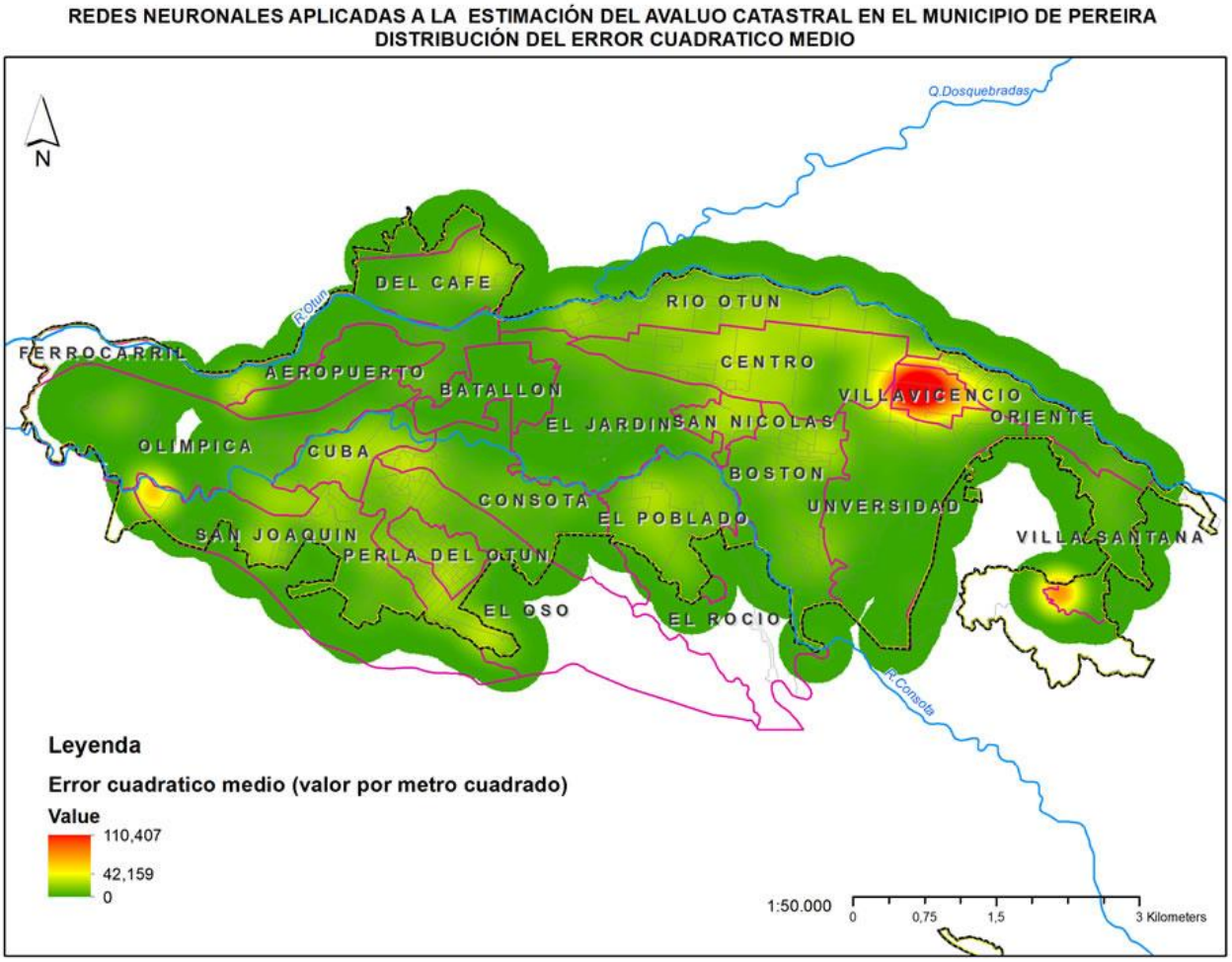


Figura 32. Mapa de calor Error Cuadrático Medio Modelo RNA Fuente: Elaboración propia

En la figura 32 se observa la concentración del error cuadrático medio ECM resultado de la aplicación del modelo de RNA. Se observa una concentración de error en el sector de la Circunvalar, muy concentrado en un radio de 20 manzanas. En general el modelo no presenta concentraciones en otros lugares y el error es casi imperceptible en las otras zonas del territorio. Se ve claramente que este modelo está haciendo la mejor generalización para la estimación del avalúo catastral, presenta solo una concentración de error relativamente pequeña.

El modelo de RNA es el que ha presentado mejores resultados en la estimación del avalúo catastral del experimento 2. Se evidencia una concentración de avalúos similar a los del avalúo IGAC, además se puede observar una distribución del error indicadora de una mejor capacidad de generalización con los datos reales del experimento 2.

9.1.7 Resultado selección de características

Para cada algoritmo se realiza selección de características en la que se utiliza Sequential Forward Selection (SFS). (Raschka, Recuperado de http://rasbt.github.io/mlxtend/user_guide/feature_selection/SequentialFeatureSelector, 2017).

Tabla 11. Resultados de selección de características

Algoritmo	Variables más importantes	ECM	ECM Desv. Est.
Modelo MSV	Tipo Predio, Suelo de Protección, Distancia a Centralidades Urbanas, Estrato, Número de Habitaciones Construcción 1, Número de Pisos Construcción 1, Puntaje Construcción 1, Puntaje Construcción 2	0,031	0,001
Modelo RLM	Tipo Predio, Área Construida, Plan parcial, Sector normativo, Estrato, Número de Pisos Construcción 1, Puntaje Construcción 1, Puntaje Construcción 2, Distancia a vías principales	0,041	0.002
Modelo PG	Tipo Predio, Estrato, Plan parcial, Suelo de Protección, Número de Pisos Construcción 1, Puntaje Construcción 1	0,033	0.001
Modelo RNA	Destino económico, Área Construida, Plan parcial, Sector Normativo, Estrato, Número de Habitaciones Construcción 1, Número de Puntaje Construcción 1, Puntaje Construcción 2, Longitud, Latitud	0,027	0.002

Como se observa en los resultados anteriores aunque el ECM es un buen error no es mejor que los resultados en el experimento 1, por lo que se decidió utilizar todas las variables para el entrenamiento y selección del mejor modelo de predicción.

9.2 Aplicación Web para la estimación de avalúos

Se desarrolló una aplicación Web con SIG (Sistemas de Información Geográfico), la cual realiza la estimación del avalúo catastral con el empleo de aprendizaje de máquina supervisado. Esta implementa el algoritmo de Procesos Gaussianos como primera opción, Redes Neuronales Artificiales y Máquinas de Soporte Vectorial, Regresión Lineal Múltiple como opciones alternativas.

La plataforma se encuentra disponible en línea y se puede acceder a ella utilizando la siguiente dirección web: http://aperion.csacolombia.com:3080/mlavaluos

Figura 33. Aplicación Web – selección de predios Fuente: Elaboración propia

En la figura número 33 se observa como en la aplicación web se selecciona un predio el cual no se tuvo en cuenta en la fase de entrenamiento de los modelos. Dicho predio tiene una avalúo catastral real de $24.109.000, a partir de este predio se calculan y se extraen las variables con las que los modelos hacen una estimación del avalúo.

En la imagen 34 se observan las 44 características de entrada para aplicar cada uno de los modelos: Procesos Gaussianos, Redes Neuronales, Máquinas de Soporte Vectorial y Regresión Lineal Múltiple. Con las variables calculadas o indicadas por el usuario se procede a aplicar el modelo seleccionado.

La plataforma web se encarga de ejecutar el algoritmo con los datos de entrada y cargar los parámetros del mejor modelo encontrado con los datos de entrenamiento. En la imagen 35 se muestran los resultados de aplicar cada uno de los algoritmos al predio seleccionado. Se observa que la red neuronal tuvo mejor desempeño en este caso puntual, con un error del 3.13% seguido de procesos gaussianos con un error del 3.73%, luego las máquinas de soporte vectorial con un error del 8.96% y finalmente la regresión lineal con un error del 19.90%.

Este resultado es muy acorde a las evaluaciones y pruebas con todo el conjunto de datos de evaluacion en los procesos de entrenamiento y validación.

Figura 34. Aplicación Web – Variables y estimación. Fuente: Elaboración propia

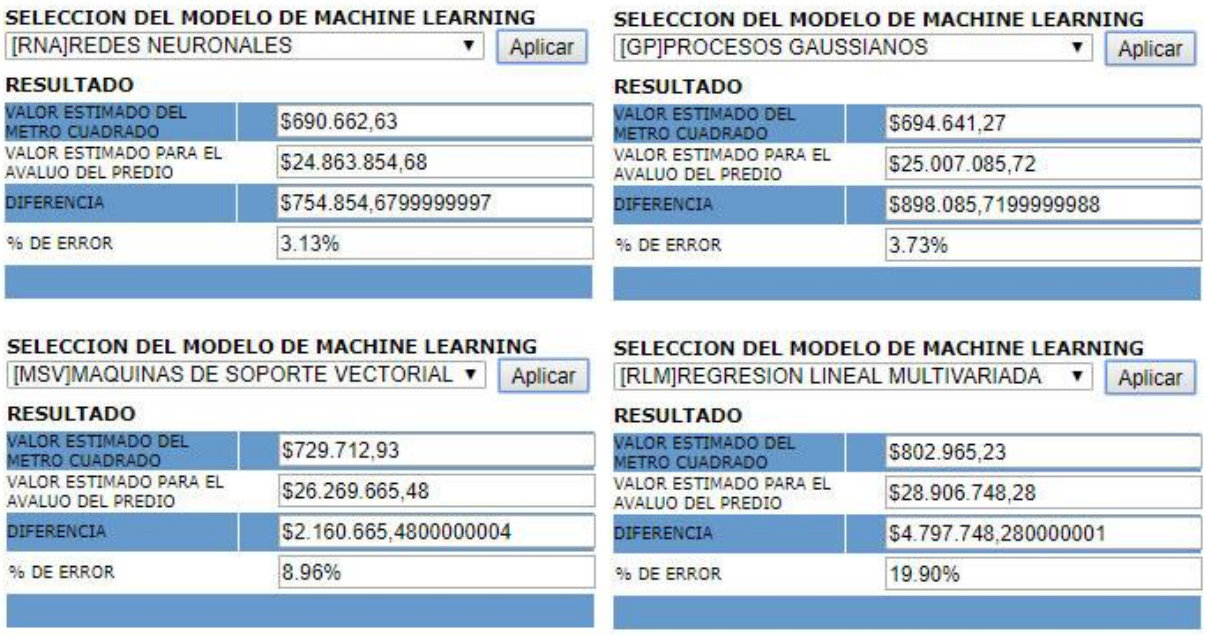


Figura 35. Aplicación Web – Variables y estimación. Fuente: Elaboración propia

En las figuras número 36, y 37 se observan las variables y ubicación de otro caso cerca al aeropuerto Matecaña del municipio de Pereira, este predio no fue incluido en los datos de entrenamiento de los algoritmos, el predio tiene una avalúo catastral real de $575.000 se considera en la zona como un avalúo que comercialmente no tiene concordancia con el avalúo catastral, debido a que el avalúo del IGAC está muy reducido. Al realizar la estimación con los algoritmos de inteligencia artificial se aprecia una mejor estimación del avalúo, más coherente con la realidad del predio dada su ubicación y características.

SELECCION DEL MODELO DE MACHINE LEARNING

[RNA]REDES NEURONALES ▼ Aplicar

RESULTADO

VALOR ESTIMADO DEL METRO CUADRADO	$541.736,1699999999
VALOR ESTIMADO PARA EL AVALUO DEL PREDIO	$10.834.723,399999999
DIFERENCIA	$10.239.723,399999999
% DE ERROR	1720.96%

SELECCION DEL MODELO DE MACHINE LEARNING

[GP]PROCESOS GAUSSIANOS ▼ Aplicar

RESULTADO

VALOR ESTIMADO DEL METRO CUADRADO	$785.929,8999999999
VALOR ESTIMADO PARA EL AVALUO DEL PREDIO	$15.718.597,999999998
DIFERENCIA	$15.123.597,999999998
% DE ERROR	2541.78%

SELECCION DEL MODELO DE MACHINE LEARNING

[MSV]MAQUINAS DE SOPORTE VECTORIAL ▼ Aplicar

RESULTADO

VALOR ESTIMADO DEL METRO CUADRADO	$811.812,49
VALOR ESTIMADO PARA EL AVALUO DEL PREDIO	$16.236.249,8
DIFERENCIA	$15.641.249,8
% DE ERROR	2628.78%

SELECCION DEL MODELO DE MACHINE LEARNING

[RLM]REGRESION LINEAL MULTIVARIADA ▼ Aplicar

RESULTADO

VALOR ESTIMADO DEL METRO CUADRADO	$127.112,47
VALOR ESTIMADO PARA EL AVALUO DEL PREDIO	$2.542.249,4
DIFERENCIA	$1.947.249,4
% DE ERROR	327.27%

INFORMACIÓN PREDIAL

BUSQUEDA ESPECIALIZADA

Por ejemplo:Matricula Nro "290-64761","64761", Ficha catastral de *12* (010600020002),*17* (66001010600020002) digitos

FICHA CATASTRAL: 66001010907850059

MATRICULA:

Consultar

INDENTIFICACION DEL PREDIO

CODIGO PREDIAL:	66001010907850059
MATRICULA:	
DIRECCION:	K 11 86 95 Br MATECANA
AREA CATASTRAL	20
AVALUO CATASTRAL	$595.000

VARIABLES

1	TIPO PREDIO	TERRENO ▼
2	DESTINO ECONOMICO	A ▼
3	AREA CONSTRUIDA	0
4	CODIGO PLAN PARCIAL	0 ▼
5	CODIGO SECTOR NORMATIVO	16 ▼
6	SUELOS DE PROTECCIÓN	NO ▼
7	CENTRALIDADES URBANAS	6 ▼
8	ESTRATO	1 ▼
9	TIENE ESTABLECIMIENTO COMERCIAL	NO ▼
10	USO SEGUN CENSO EMP.	VIVIENDA ▼
11	1 HABITACIONES	0
12	1 BAÑOS	0
13	1 LOCALES	0
14	1 PISOS	0
15	1 PUNTAJE	0
16	1 AREA CONST	0

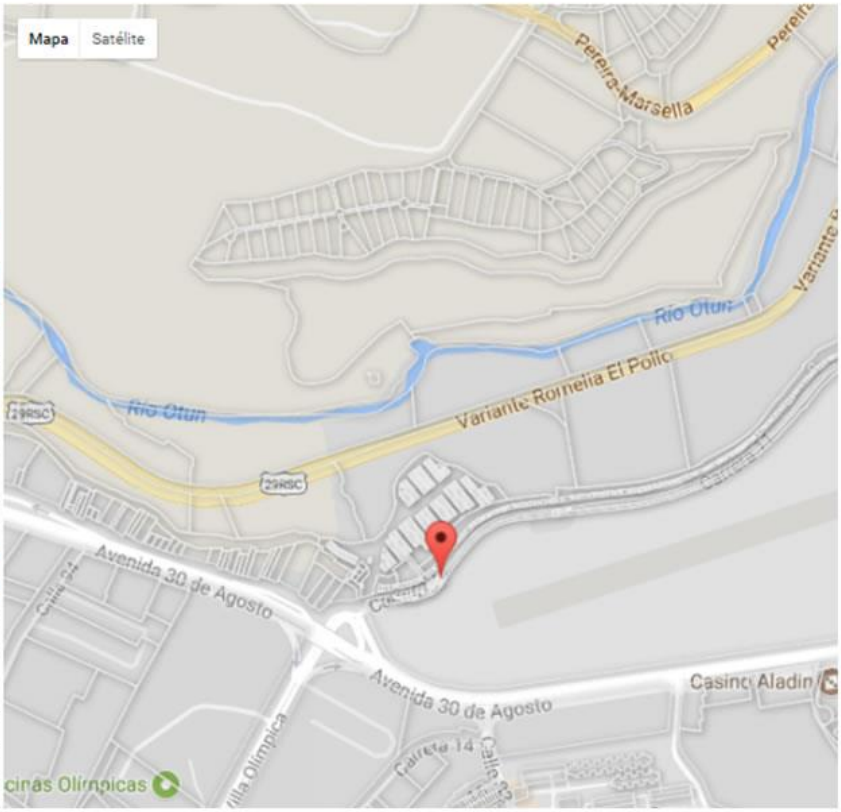


Figura 36. Aplicación Web – Variables y estimación. Fuente: Elaboración propia

17	2 HABITACIONES	0
18	2 BAÑOS	0
19	2 LOCALES	0
20	2 PISOS	0
21	2 PUNTAJE	0
22	2 AREA CONST	0
23	3 HABITACIONES	0
24	3 BAÑOS	0
25	3 LOCALES	0
26	3 PISOS	0
27	3 PUNTAJE	0
28	3 AREA CONST	0
29	ISTP DIST MINIMA	1267.922037
30	CENTRALIDADES URBANAS DISTANCIA MIN	915.630725
31	EQUIPAMIENTO URNANO DISTANCIA MINIMA	9.008118
32	ESTACION DE SERVICIO DISTANCIA MINIMA	1276.935609
33	ESTRUCTURAS MASIVA DIST MIN	369.200572
34	INSTALACIONES CRITICAS DIST MINIMA	9.138274
35	LATITUD (CENTROIDE)	4.810498
36	LONGITUD (CENTROIDE)	-75.748851
37	VIA DISTANCIA MINIMA	5.652535
38	COD BARRIO GEO	66001011207
39	VIA GID	846
40	SITP GID	402
41	EQUIPAMIENTO URBANO	153
42	ESTACION DE SERVICIO GID	24
43	ESTRUCTURAS DE CONCENTRACION MASIVA GID	38
44	INSTALACIONES CRITICAS - GID	9

Figura 37. Aplicación Web – Variables y estimación. Fuente: Elaboración propia

9.3 Comparación de resultados con otras investigaciones

A continuación se presenta una tabla comparativa con los estudios relacionados al tema de investigación que se plantearon en el estado del arte. Se incluyen las investigaciones que utilizaron el indicador r^2, ECM y los mismos algoritmos con los resultados finales.

Tabla 12. Comparación de resultados con otros estudios

Estudio	**Algoritmo**	**r^2**	**ECM**	**Total de registros**	**Número de Variables**
Christchurch Nueva Zelanda. Junio 2004	MPH	61,92%	876.215,63	200	13
Christchurch Nueva Zelanda. Junio 2004	RNA	90,00%	449.111,46	200	12
Córdoba España, Enero 2008.	MPH	77,38%	41.645,43	2888	26
Córdoba España, Enero 2008.	RNA	86,05%	39.540,36	2888	26
Bogotá Colombia, Septiembre 2009	MPH		1,9700	2627	17
Bogotá Colombia, Septiembre 2009	RNA		0,2633	2627	17
Bogotá Colombia, Septiembre 2009	MSV		0,2263	2627	17
New York EEUU. Diciembre 2014	RLM	64,50%	1.66E+09	1047	5
New York EEUU. Diciembre 2014	RNA	81,70%	1.29E+09	1047	5
Pereira Colombia, Mayo 2017 (Experimento 1)	MSV	84,62%	0,022	58523	45

Pereira Colombia, Mayo 2017 (Experimento 1)	RLM	69,85%	0,043	58523	45
Pereira Colombia, Mayo 2017 (Experimento 1)	PG	85,93%	0,020	58523	45
Pereira Colombia, Mayo 2017 (Experimento 1)	RNA	83,10%	0,024	58523	45
Pereira Colombia, Mayo 2017 (Experimento 2)	MSV	63,29%	0,052	58523	45
Pereira Colombia, Mayo 2017 (Experimento 2)	RLM	53,83%	0,065	58523	45
Pereira Colombia, Mayo 2017 (Experimento 2)	PG	71,06%	0,041	58523	45
Pereira Colombia, Mayo 2017 (Experimento 2)	RNA	71,57%	0,041	58523	45
Pereira Colombia, Mayo 2017 (Sel Car)	MSV		0,031	58523	8
Pereira Colombia, Mayo 2017 (Sel Car)	RLM		0,041	58523	9
Pereira Colombia, Mayo 2017 (Sel Car)	PG		0,033	58523	5
Pereira Colombia, Mayo 2017 (Sel Car)	RNA		0,027	58523	10

CONVENCIONES:

Modelo de precio hedónico	MPH
Red Neuronal Artificial	RNA
Máquinas de Soporte Vectorial	MSV
Regresión Lineal Múltiple	RLM

Al comparar esta investigación con otras se puede observar que los resultados del coeficiente de determinación (r^2) del experimento 1 de los algoritmos PG y RNA estuvieron acordes con las investigaciones en los otros países y se mantuvo la premisa según la cual los métodos de inteligencia artificial son mejores que los modelos tradicionales hedónicos y que el método de regresión lineal múltiple.

También cabe resaltar la comparación de la cantidad de registros y número de variables utilizados. En esta investigación se utilizó mayor cantidad de registros y mayor número de variables que en los otros estudios.

10. Conclusiones y recomendaciones

10.1 Conclusiones

En todos los estudios se demostró que los métodos de inteligencia artificial son más efectivos que los métodos tradicionales, modelos hedónicos y de regresión.

Se evidencia la importancia de incorporar las características geográficas y las distancias del predio a diferentes sitios de interés urbanos como por ejemplo centralidades urbanas, vías principales, puntos de transporte y otros.

Se destaca la necesidad de incorporar en el desarrollo de aplicaciones para el cálculo del avalúo catastral, las variables geográficas inherentes al territorio de estudio, razón por la cual se recomienda el uso de Sistemas de información Geográfica (SIG) para la consulta y procesamiento de la información que involucra el cálculo del avalúo catastral.

En el análisis exploratorio de datos se pudieron identificar los datos atípicos que al ser excluidos mejoraron el desempeño del error. Se evidenció la importancia de homogenizar, normalizar, codificar las entradas y depurar los datos ausentes para las entradas de los algoritmos y garantizar el correcto desempeño.

En la etapa de comparación de los algoritmos de inteligencia artificial con respecto a los métodos tradicionales, se mantiene el mismo resultado de los artículos expuestos en el estado del arte, donde los algoritmos de inteligencia artificial son mejores que los métodos tradicionales en la estimación del avalúo. Se tomaron como base de comparación el error cuadrático medio (ECM), el error absoluto medio (EAM) y el coeficiente de determinación (r2).

En el primer experimento de selección del mejor algoritmo y modelo de estimación del avalúo catastral, presentó mejor rendimiento el proceso gaussiano (PG), el valor de coeficiente de determinación (r^2) fue superior en 1% frente al modelo de máquinas de soporte vectorial (MSV), un 3% al modelo de redes neuronales y un 16% al modelo de regresión lineal multivariada (RLM), también se observó mejor desempeño del PG en el ECM y EAM.

El segundo experimento permitió garantizar la capacidad de generalización de los algoritmos. El resultado del r^2 en la RNA superó por 1% al PG y en el estimador ECM se comportaron de la misma manera, los dos algoritmos fueron superiores a los modelos MSV y RLM en 9% y 19% aproximadamente.

En el proceso de selección de las variables explicativas de los diferentes algoritmos se evidenciaron características importantes como lo son los puntajes de las construcciones, distancia

a centralidades urbanas, distancia a vías principales, estrato, latitud, longitud, a pesar de que el resultado del ECM no fue mejor que el de todas las variables.

En la comparación con otros estudios planteados en el estado del arte de este documento, se pudo comprobar la coherencia de los resultados en lo concerniente al estimador r^2 y ECM, también se observó una importante diferencia en el aporte de este estudio dado que se utilizó una mayor cantidad de variables y registros lo que hace más difícil realizar los procesos de entrenamiento, validación y pruebas pero da una mayor capacidad de generalización en la estimación del avalúo.

Los modelos de inteligencia artificial han sido entrenados para operar en un sector geográfico y con características específicas, estos modelos han sido ajustados para operar con las características de estos lugares, no es factible utilizar este modelo en lugares alejados geográficamente del territorio objetivo de entrenamiento y estimación, se presentan variaciones de estimaciones de avalúos muy grandes.

Luego de tener los resultados de los modelos optimizados, desarrollar la aplicación web permitió la consulta y la variación de parámetros en línea para obtener resultados rápidos de cálculos de avalúos aplicando los modelos obtenidos.

Utilizar herramientas de software libre integradas con las librerías de Python para aprendizaje de máquina permitió el desarrollo de la aplicación web, a bajo costo y con buenos desempeños de cálculo.

Las técnicas de inteligencia artificial representan una alternativa para la modernización de las técnicas y metodologías aplicadas por el IGAC para la estimación del avalúo catastral.

10.2 Recomendaciones

Para trabajos futuros se espera que con base en esta investigación se puedan comparar los resultados con otras técnicas de aprendizaje de máquina supervisado que puedan o no obtener mejores resultados como: Random Forest Regressor (RFR), AdaBoost Regressor, Deep Learning o métodos diferentes como lógica difusa, sistemas basados en reglas y otros algoritmos de inteligencia artificial.

Realizar un estudio donde se incorporen nuevas variables que no fueron tenidas en cuenta en esta investigación, tanto externas al predio que puedan afectar su valor, como por ejemplo: el factor de contaminación ambiental, ruido, índice de criminalidad y variables internas como: piscina, tina, acabados, tipo de materiales y otras.

Aplicar esta misma metodología para la predicción del avalúo catastral en otro municipio o localidades para validar la capacidad de generalización del aprendizaje del algoritmo y la generalidad de la técnica.

Finalmente, se espera que el Estado promueva más este tipo de investigaciones y las aplique en el cálculo del avalúo catastral para que no haya intervención humana y sea el sistema que dé el valor del predio, con el fin de tener una mayor trasparencia y equidad en el proceso.

Aplicar esta metodología para el entrenamiento y posterior validación en la estimación de avalúos comerciales de predios utilizando como insumos los avalúos comerciales de observatorios inmobiliarios municipales.

Se pueden implementar técnicas de entrenamiento de forma automática utilizando nuevos datos de entrada para actualizar los modelos y poder corregir errores de estimación en lugares específicos.

Las técnicas de muestreo para la selección de datos de entrenamiento y validación se pueden mejorar utilizando los métodos de cadenas de Markov Monte Carlo (MCMC), Algoritmo Metropolis Hastings

Se recomienda utilizar los modelos y técnicas abordadas en este estudio para realizar el entrenamiento con muestras aleatorias de avalúos comerciales, luego hacer una estimación de forma automática sobre el 100% de los predios del municipio, posteriormente calcular el avalúo catastral como el 60% del avalúo comercial, esto permitiría realizar rápidamente actualizaciones catastrales de municipios completos de una forma más ágil y económica.

Referencias Bibliográficas

Ávila, L., & Robayo, V. (2003). Tesis. *Red neuronal para determinar el valor del metro cuadrado de construcción*. Colombia.

Ayan, E., & Erkin, C. (2013). Hedonic modeling for a growing housing market: valuation of apartments in complexes. *International journal of economics and finance* .

Bishop, C. (2006). *Pattern Recognition and Machine Learning.* Springer.

Bozic, Milicevic, Perica, & Marosan. (2013).

Brownlee, J. (2016). *Machine learning Mastery with Python.*

EL TIEMPO. (06 de 02 de 2016). Abecé de la polémica por los avalúos de vehículos en Colombia. *EL TIEMPO*, págs. http://www.eltiempo.com/economia/finanzas-personales/tabla-de-avaluos-de-vehiculos-en-colombia-polemica-datasoft/16502011?ts=89.

Fajardo, A. C. (2014). Tesis Universidad Nacional de Colombia. *Propuesta metodológica para calcular el avalúo catastral de un predio utilizando Redes Neuronales Artificiales*. Colombia.

IGAC. (2017). *http://www.igac.gov.co/igac*. Obtenido de http://www.igac.gov.co/igac: http://www.igac.gov.co/igac

Khamis, A., & Binti Kamarudin, K. (2014). Comparative study on estimate house pricing using statistical and neural network model. *International Journal of scientific & technology research*.

Kim, K., & Park, J. (2005). Segmentation of the housing market and its determinants: Seoul and its neighboring new towns in Korea. *Australian Geographer*, 221-232.

Knuth, D. (1997). *El arte de la programación de computadores Volumen 1 (3ª edición).* Massachusetts: Addison–Wesley.

Kusan, H., Aytekin, & Özdemir, I. (2010). The use of fuzzy logic in predicting house selling price. *Expert systems with applications* , 1808-1813.

Limbosunchai, V. (junio de 2004). Artículo presentado en NZARES conference. *House price prediction: Hedonic Price model vs. Artificial Neural Network*. Blenheim Country, Nueva Zelandia: New Zealand Agricultural and resource economics society.

Mora Esperanza, J. (2004). La inteligencia artificial aplicada a la valoración de inmuebles. Un ejemplo para valorar Madrid. *CT Catastro*.

Morales, G. M., & Hernández, G. (2009). Utilización de máquinas con vectores de soporte para regresión m2 de construcción en Bogotá. *Revista Avances en Sistemas e Informática*.

Murphy, K. P. (2012). *Machine learning a probabilistic perspective.* Massahusetts: MIT.

Oficina de extensión y asesoría de la Universidad Nacional de Colombia. (2002). Tesis. *Elaboración de modelos econométricos*. Bogotá, Colombia: Facultad de Ciencias. Departamento de estadística.

Pagourtzi, E., Assimakopoulos, V., Hatzichristos, T., & French, N. (2003). Real estate appraisal: a review of valuation methods. *Journal of property investment and finance*, 21(4),383-401.

Park, B., & Kwon, B. J. (2015). Using machine learning algorithms for housing price prediction, The case of Fairfax County, Virginia housing data. *Expert Systems with Applications*.

Pereira, M. d. (2017). *http://www.pereira.gov.co/MiMunicipio/Paginas/Informacion-del-Municipio.aspx*. Obtenido de http://www.pereira.gov.co/MiMunicipio/Paginas/Informacion-del-Municipio.aspx.

Raschka, S. (2015). Capitulo 6 Mejores prácticas para evaluación del modelo y optimización de hiperparametros. En S. Raschka, *Python machine learning* (pág. 185).

Raschka, S. (18 de Octubre de 2017). *Recuperado de http://rasbt.github.io/mlxtend/user_guide/feature_selection/SequentialFeatureSelector*. Recuperado el 18 de Octubre de 2017, de mlxtend: http://rasbt.github.io/mlxtend/user_guide/feature_selection/SequentialFeatureSelector

Rasmussen, C. E., & Williams, C. K. (2006). *Gaussian processes for machine learning.* Massachusetts: MIT Press.

Romero, J. J., Dafonte, C., & Penousal, F. J. (2007). Inteligencia artificial y computación avanzada. *Fundación Alfredo Brañas Colección Informática número 13*, 11-13.

Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics 7*.

Villamandos, N. C., Caridad, M. J., & Núñez, T. J. (2008). Metodología de precios hedónicos vs. redes neuronales artificiales como alternativas a la valoración de inmuebles. *CT Catastro*.

Weka 3: Data Mining Software in Java. (s.f.). Obtenido de UNIVERSIDAD DE WAIKATO: http://www.cs.waikato.ac.nz/ml/weka/

Wu, C. H., Li, C. H., Fang, I. C., Hsu, C. C., & Ling, W. T. (2009). Hybrid genetic-based support vector regression with feng shui theory for appraising real estate price. *IEEE Computer society DOI 10 . 1109*.

Anexos

Anexo A. Definición de Variables

Núm.	Variable	Convención	Unidades	Tipo Variable	Descripción
1	Tipo de Predio	tipo_predio	Categórica	Nominal	Define la clasificación del predio en : Terreno, Condominio Urbano o Pent House (PH).
2	Perímetro calculado	shape_leng	M	Numérica	Perímetro calculado por el SIG en metros
3	Área calculada	shape_area	M^2	Numérica	Área calculada por el SIG en metros cuadrados
4	Destino Económico	destino_ec	Categórica	Nominal	Agrupa los previos según su actividad económica o residencial por ejemplo: comercial, residencial, estatal
5	Área del terreno	area_terreno_catrastro	M^2	Numérica	Área del predio en metros cuadrados calculada por catastro
6	Área construida	area_construida_catastro	M^2	Numérica	Área construida del predio en metros cuadrados calculada por catastro
7	Latitud	latitud	Planares	Numérica	Coordenada geográfica que representa la distancia del punto sobre el eje x
8	Longitud	longitud	Planares	Numérica	Coordenada geográfica que representa la distancia del punto sobre el eje y
9	Código del barrio	cod_barrio_geo	Categórica	Nominal	Código del barrio asignado según la cartografía del IGAC
10	Nombre del barrio	nombre_barrio	Categórica	Nominal	Nombre que representa el barrio en la ciudad
11	Código de la comuna	cod_comuna_geo	Categórica	Nominal	Código de la comuna asignado según la cartografía del IGAC
12	Nombre de la comuna	nombre_comunad	Categórica	Nominal	Nombre que representa la comuna en la ciudad
13	Código plan parcial	cod_plan_parcial_geo	Categórica	Ordinal	Código del plan parcial asignado según la cartografía del IGAC
14	Plan parcial	nombre_plan_parcial	Categórica	Nominal	Nombre del plan parcial asignado a una zona basado en el plan

					de ordenamiento territorial (POT) de la ciudad.
15	Código sector normativo	cod_sector_normativo_geo	Categórica	Ordinal	Código de norma que aplica la zona basado en el POT de la ciudad.
16	Suelo de Protección	suelo_proteccion	Categórica	Nominal	Si es o no suelo de protección según el POT de la ciudad.
17	Código de zona física 1	zfis_1	Categórica	Ordinal	Código zona física 1.
18	Código zona económica 1	z_econ1	Categórica	Ordinal	Código zona económica 1.
19	Área del terreno en la zona física 1	aterreno1	M^2	Numérica	Área del terreno en la zona física 1.
20	Código de zona física 2	zfis_2	Categórica	Ordinal	Código zona física 2
21	Código zona económica 2	z_econ2	Categórica	Ordinal	Código zona económica 2
22	Área del terreno en la zona física 2	aterreno2	M^2	Numérica	Área del terreno de la zona física 2.
23	Habitaciones de la construcción 1	habitaciones1	Cantidad	Numérica	Número de habitaciones de la primera construcción del predio
24	Baños de la construcción 1	banos1	Cantidad	Numérica	Número de baños de la primera construcción del predio
25	Locales de la construcción 1	locales1	Cantidad	Numérica	Número de locales de la primera construcción del predio
26	Pisos de la construcción 1	pisos1	Cantidad	Numérica	Número de pisos de la primera construcción
27	Estrato de la construcción 1	estrato1	Categórica	Ordinal	Estrato de la primera construcción (0,1,2,3,4,5,6,7,8,9)
28	Destino económico de la construcción 1	destino1	Categórica	Ordinal	Destino económico de la primera construcción.
29	Puntaje de la construcción 1	puntaje1	Categórica	Ordinal	Puntaje de la primera construcción.
30	Área construcción 1	area_cons1	M^2	Numérica	Área de la primera construcción
31	Habitaciones de la construcción 2	habitaciones2	Cantidad	Numérica	Número de habitaciones de la segunda construcción
32	Baños de la construcción 2	banos2	Cantidad	Numérica	Número de baños de la segunda construcción del predio
33	Locales de la construcción 2	locales2	Cantidad	Numérica	Número de locales de la segunda construcción del predio

34	Pisos de la construcción 2	Pisos2	Cantidad	Numérica	Número de pisos de la segunda construcción
35	Estrato de la construcción 2	Estrato2	Categórica	Ordinal	Estrato de la segunda construcción (0,1,2,3,4,5,6,7,8,9)
36	Destino económico de la construcción 2	destino2	Categórica	Ordinal	Destino económico de la segunda construcción.
37	Puntaje de la construcción 2	puntaje2	Categórica	Ordinal	Puntaje de la segunda construcción.
38	Área construcción 2	area_cons2	M^2	Numérica	Área de la segunda construcción
39	Habitaciones de la construcción 3	habitaciones3	Cantidad	Numérica	Número de habitaciones de la tercera construcción
40	Baños de la construcción 3	banos3	Cantidad	Numérica	Número de baños de la tercera construcción del predio
41	Locales de la construcción 3	locales3	Cantidad	Numérica	Número de locales de la tercera construcción del predio
42	Pisos de la construcción 3	pisos3	Cantidad	Numérica	Número de pisos de la tercera construcción
43	Estrato de la construcción 3	Estrato3	Categórica	Ordinal	Estrato de la tercera construcción (0,1,2,3,4,5,6,7,8,9)
44	Destino económico de la construcción 3	destino3	Categórica	Ordinal	Destino económico de la tercera construcción.
45	Puntaje de la construcción 3	puntaje3	Categórica	Ordinal	Puntaje de la tercera construcción.
46	Área construcción 3	area_cons3	M^2	Numérica	Área de la tercera construcción
47	Distancia mínima a las vías	via_dist_minima	M	Numérica	Distancia en metros mínima a una vía de carretera.
48	Código de la via mas cercana	via_gid	Categórica	Ordinal	Código de la vía más cercana al predio.
49	Nombre de la vía mas cercana	via_nombre	Categórica	Nominal	Nombre de la vía más cercana al predio.
50	Distancia mínima a una estación de sistema integrado de transporte	sitp_dist_minima	M	Numérica	Distancia en metros a la estación de sistema de integrado de transporte
51	Código de la estación de sistema de integrado de transporte mas cercano	stip_gid	Categórica	Ordinal	Código de la estación de sistema de integrado de transporte más cercana al predio
52	Nombre de la estación de sistema de integrado de transporte más cercana	stip_nombre	Categórica	Nominal	Nombre de la estación de sistema de integrado de transporte más cercana al predio
53	Distancia mínima a una centralidad	centralidades_urbanas_dist_minima	M	Numérica	Distancia en metros mínima a una

	urbana definida en el POT				centralidad urbana definida en el POT de la ciudad.
54	Código de la centralidad urbana más cercana	centralidades_urbanas_gid	Categórica	Ordinal	Código de la centralidad urbana mas cercano al predio
55	Nombre de la centralidad urbana mas cercana	centralidades_urbanas_nombre	Categórica	Nominal	Nombre de la centralidad mas cercana al predio
56	Distancia mínima al equipamiento urbano	equipamiento_urbano_dist_minima	M	Numérica	Distancia en metros mínima a un equipamiento urbano.
57	Código del equipamiento urbano más cercano	equipamiento_urbano_gid	Categórica	Ordinal	Código del equipamiento urbano más cercano al predio
58	Nombre del equipamiento urbano más cercano	equipamiento_urbano_nombre	Categórica	Nominal	Nombre del equipamiento urbano más cercano al predio
59	Distancia mínima a estación de servicio de combustible	estacion_servicio_combustible_dist_minima	M	Numérica	Distancia en metros mínima a una estación de servicio de combustible
60	Código de la estación de servicio de combustible más cercana	estaciones_servicio_combustible_gid	Categórica	Ordinal	Código de la estación de servicio de combustible más cercana al predio
61	Nombre de la estación de servicio de combustible más cercana	estaciones_servicio_combustible_nombre	Categórica	Nominal	Nombre de la estación de servicio de combustible más cercana al predio
62	Distancia mínima a una estructura de concentración masiva	estructuras_de_concentracion_masiva_dist_minima	M	Numérica	Distancia en metros mínima a una estructura de concentración masiva
63	Código de estructura de concentración masiva	estructuras_de_concentracion_masiva_gid	Categórica	Ordinal	Código de la estructura de concentración masiva más cercana al previo
64	Nombre de estructura de concentración masiva	estructuras_de_concentracion_masiva_nombre	Categórica	Nominal	Nombre de la estructura de concentración masiva más cercana al previo
65	Distancia mínima a instalaciones criticas	instalaciones_criticas_dist_minima	M	Numérica	Distancia en metros mínima a una instalación critica
66	Código instalación critica	instalaciones_criticas_gid	Categórica	Ordinal	Código de la instalación critica mas cercana al previo
67	Nombre instalaciones críticas	instalaciones_criticas_nombre	Categórica	Nominal	Nombre de la instalación crítica más cercana al predio
68	Uso del suelo según el POT.	suelo_pot2015_nombre	Categórica	Nominal	Nombre del uso del suelo según el POT de la ciudad.

69	Estrato general del predio	estrato_mun	Categórica	Ordinal	Estrato global del predio (1, 2 ,3 , 4, 5, 6)
70	CIIU Código de identificación de la actividad económica del predio	censo_ciiu	Categórica	Ordinal	Código de actividad económica del predio
71	Tipo Actividad Económica	censo_tipo_actividad	Categórica	Nominal	Tipo de actividad económica (Comercio, Industria, Servicio,Vivienda)
72	Estrato Censo comercial	censo_estrato	Categórica	Ordinal	Estrato establecido por el censo empresarial de la cámara de comercio (1,2,3,4,5,6)
73	Cantidad de establecimientos en un predio	censo_cantidad_est	Cantidad	Numérica	Total de establecimientos que existen en un predio
74	Distancia mínima a un árbol	censo_arboles_dist_min	Metros	Numérica	Distancia en metros mínima a un árbol desde un predio.
75	Código del árbol más cercano al predio	censo_arboles_gid	Categórica	Ordinal	Código del árbol más cercano al predio
76	Zona económica 1	z_eco_1	Categórica	Ordinal	Zona económica 1 asignada al predio
77	Área Zona económica 1	z_eco1_area	M^2	Numérica	Área en metros cuadrados de la zona económica 1
78	Zona económica 2	z_eco_2	Categórica	Ordinal	Zona económica 2 asignada al predio
79	Área Zona económica 2	z_eco2_area	M^2	Numérica	Área en metros cuadrados de la zona económica 2
80	Zona económica 3	z_eco_3	Categórica	Ordinal	Zona económica 3 asignada al predio
81	Área Zona económica 3	z_eco3_area	M^2	Numérica	Área en metros cuadrados de la zona económica 3
82	Establecimiento comercial	establecimiento_comercial	Categórica	Nominal	Define si un predio es o no un establecimiento comercial si(1), no (0)
83	Avalúo catastral	Avalúo	Pesos	Numérica	Avalúo Catastral en pesos colombianos del predio

Anexo B. Características del predio depuradas

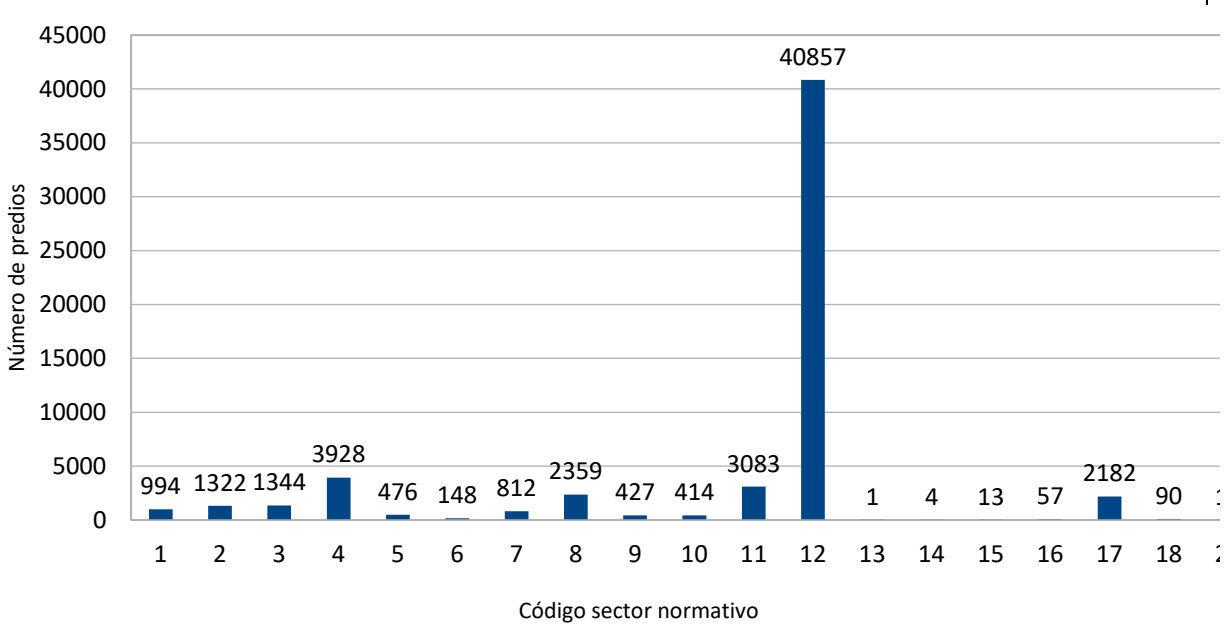
Núm.	Variable	Convención	Unidades	Tipo Variable	Descripción
1	Tipo de Predio	tipo_predio	Categórica	Nominal	Define la clasificación del predio en: Terreno, Condominio Urbano o Penthouse (PH).
2	Destino Económico	destino_ec	Categórica	Nominal	Agrupa los previos según su actividad económica o residencial por ejemplo: comercial, residencial, estatal
3	Área construida	area_construida_catastro	m^2	Numérica	Área construida del predio en metros cuadrados calculada por catastro
4	Latitud	latitud	Planares	Numérica	Coordenada geográfica que representa la distancia del punto sobre el eje x
5	Longitud	longitud	Planares	Numérica	Coordenada geográfica que representa la distancia del punto sobre el eje y
6	Código del barrio	cod_barrio_geo	Categórica	Nominal	Código del barrio asignado según la cartografía del IGAC
7	Código plan parcial	cod_plan_parcial_geo	Categórica	Ordinal	Código del plan parcial asignado según la cartografía del IGAC
8	Código sector normativo	cod_sector_normativo_geo	Categórica	Ordinal	Código de norma que aplica la zona basado en en el POT de la ciudad.
9	Suelo de protección	suelo_proteccion	Categórica	Nominal	Si es o no suelo de protección según el POT de la ciudad.
10	Habitaciones de la construcción 1	habitaciones1	Cantidad	Numérica	Número de habitaciones de la primera construcción del predio
11	Baños de la construcción 1	banos1	Cantidad	Numérica	Número de baños de la primera construcción del predio
12	Locales de la construcción 1	locales1	Cantidad	Numérica	Número de locales de la primera construcción del predio
13	Pisos de la construcción 1	pisos1	Cantidad	Numérica	Número de pisos de la primera construcción
14	Puntaje de la construcción 1	puntaje1	Categórica	Ordinal	Puntaje de la primera construcción.
15	Área construcción 1	area_cons1	M^2	Numérica	Área de la primera construcción
16	Habitaciones de la construcción 2	habitaciones2	Cantidad	Numérica	Número de habitaciones de la segunda construcción
17	Baños de la construcción 2	banos2	Cantidad	Numérica	Número de baños de la segunda construcción del predio
18	Locales de la construcción 2	locales2	Cantidad	Numérica	Número de locales de la segunda construcción del predio
19	Pisos de la construcción 2	Pisos2	Cantidad	Numérica	Número de pisos de la segunda construcción
20	Puntaje de la construcción 2	puntaje2	Categórica	Ordinal	Puntaje de la segunda construcción.
21	Área construcción 2	area_cons2	M^2	Numérica	Área de la segunda construcción
22	Habitaciones de la construcción 3	habitaciones3	Cantidad	Numérica	Número de habitaciones de la tercera construcción
23	Baños de la construcción 3	banos3	Cantidad	Numérica	Número de baños de la tercera construcción del predio

24	Locales de la construcción 3	locales3	Cantidad	Numérica	Número de locales de la tercera construcción del predio
25	Pisos de la construcción 3	pisos3	Cantidad	Numérica	Número de pisos de la tercera construcción
26	Puntaje de la construcción 3	puntaje3	Categórica	Ordinal	Puntaje de la tercera construcción.
27	Área construcción 3	area_cons3	M^2	Numérica	Área de la tercera construcción
28	Distancia mínima a las vías	via_dist_minima	M	Numérica	Distancia en metros mínima a una vía de carretera.
29	Código de la vía más cercana	via_gid	Categórica	Ordinal	Código de la vía más cercana al predio.
30	Distancia mínima a una estación de sistema integrado de transporte	sitp_dist_minima	M	Numérica	Distancia en metros a la estación de sistema integrado de transporte
31	Código de la estación de sistema de integrado de transporte más cercano	stip_gid	Categórica	Ordinal	Código de la estación de sistema de integrado de transporte más cercano al predio
32	Distancia mínima a una centralidad urbana definida en el POT	centralidades_urbanas_dist_minima	M	Numérica	Distancia en metros mínima a una centralidad urbana definida en el POT de la ciudad.
33	Código de la centralidad urbana más cercana	centralidades_urbanas_gid	Categórica	Ordinal	Código de la centralidad urbana más cercana al predio
34	Distancia mínima al equipamiento urbano	equipamiento_urbano_dist_minima	M	Numérica	Distancia en metros mínima a un equipamiento urbano.
35	Código del equipamiento urbano más cercano	equipamiento_urbano_gid	Categórica	Ordinal	Código del equipamiento urbano más cercano al predio
36	Distancia mínima a estación de servicio de combustible	estacion_servicio_combustible_dist_minima	M	Numérica	Distancia en metros mínima a una estación de servicio de combustible
37	Código de la estación de servicio de combustible más cercana	estaciones_servicio_combustible_gid	Categórica	Ordinal	Código de la estación de servicio de combustible más cercana al predio
38	Distancia mínima a una estructura de concentración masiva	estructuras_de_concentracion_masiva_dist_minima	M	Numérica	Distancia en metros mínima a una estructura de concentración masiva
39	Código de estructura de concentración masiva	estructuras_de_concentracion_masiva_gid	Categórica	Ordinal	Código de la estructura de concentración masiva más cercana al predio
40	Distancia mínima a instalaciones críticas	instalaciones_criticas_dist_minima	M	Numérica	Distancia en metros mínima a una instalación crítica
41	Código instalación crítica	instalaciones_criticas_gid	Categórica	Ordinal	Código de la instalación crítica más cercano al previo

42	Estrato general del predio	estrato_mun	Categórica	Ordinal	Estrato global del predio (1, 2 ,3 , 4, 5, 6)
43	Tipo Actividad Económica	censo_tipo_actividad	Categórica	Nominal	Tipo de actividad económica (Comercio, Industria, Servicio,Vivienda)
44	Establecimiento comercial	establecimiento_comercial	Categórica	Nominal	Define si un predio es o no un establecimiento comercial si(1), no (0)
45	Avalúo catastral	Avalúo	Pesos	Numérica	Avalúo Catastral en pesos colombianos del predio

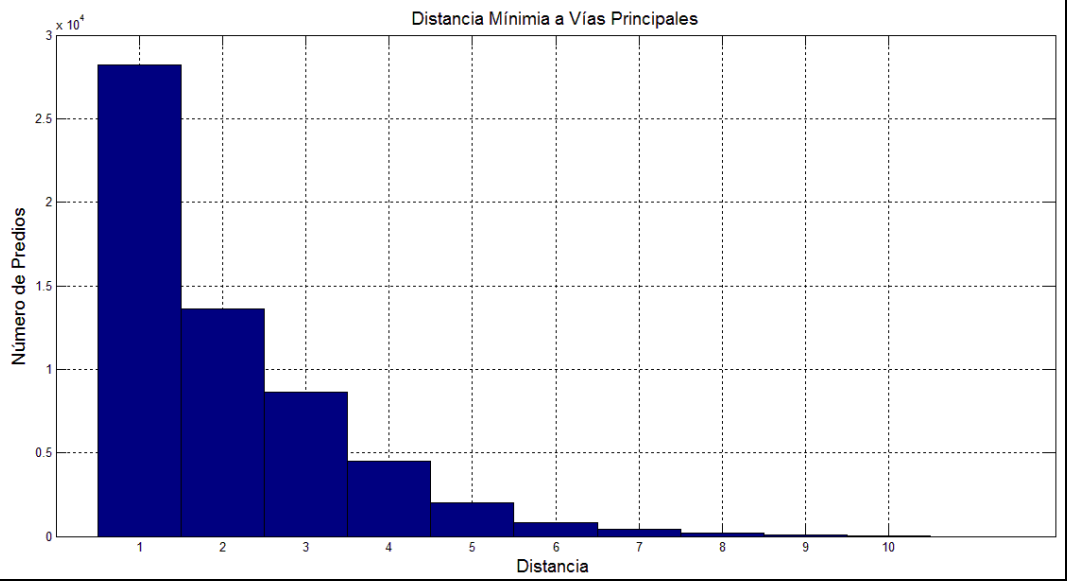
Anexo C. Histogramas

CÓDIGO SECTOR NORMATIVO
Variable: Código sector normativo
Unidad: Categórica
Total = 58523



Rango de distancia	Total Predios
0-15	28218
16-45	13620
46-75	8645
76-105	4477
76-135	2017
136-165	838
166-195	412
196-225	176
226-255	113
256-285	7

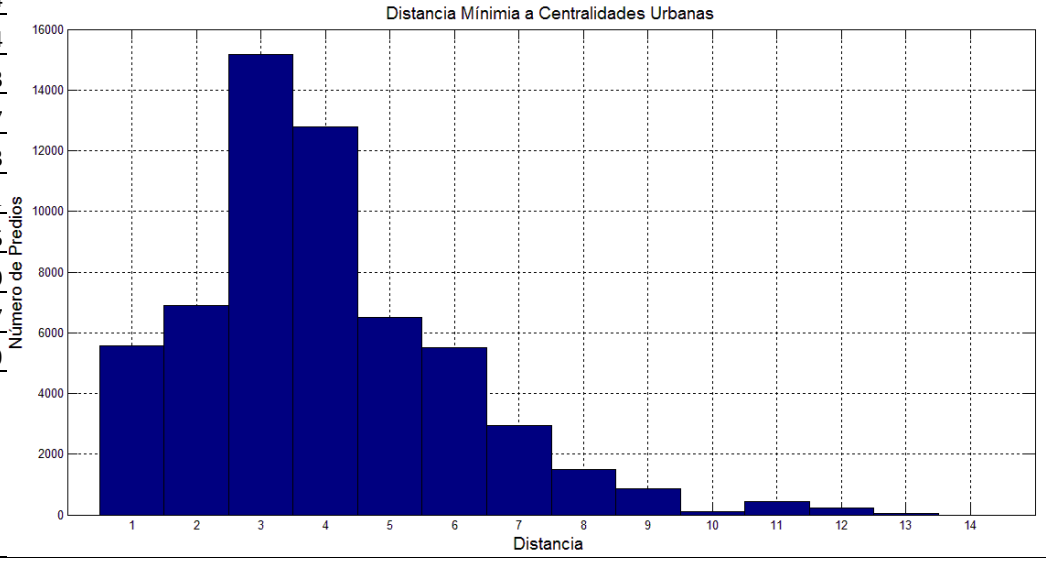
DISTANCIA MÍNIMA A VÍAS PRINCIPALES
Mínimo: 0
Máximo: 262,955
Media: 33,43
Desviación Estándar: 38,22
Unidad: Metros
Total = 58523



Rango de distancia	Total de Predios
0-100	11124
101-300	10330
301-500	11349
501-700	8637
701-800	7104
801-1100	4334
1101-1300	2473
1301-1500	1507
1501-1700	863
1701-1900	101
1901-2100	425
2101-2300	239
2301-2500	37
2501-2700	0

DISTANCIA MÍNIMA A CENTRALIDADES URBANAS

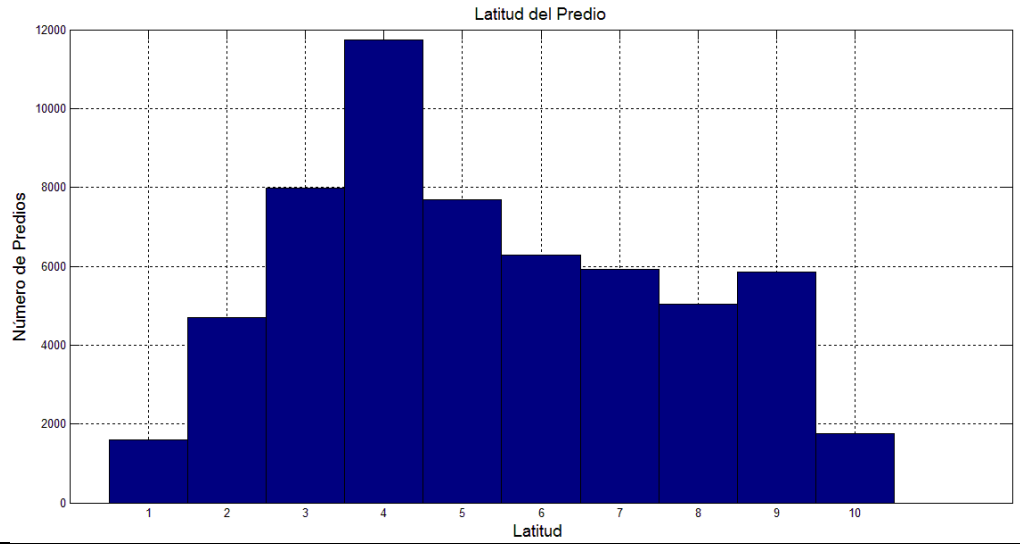
Variable: Distancia mínima a centralidades urbanas
Mínimo: 0
Máximo: 2489,261
Media: 513,079
Desviación Estándar.: 427,578
Unidad: Metros
Total = 58523



Coordenada Latitud	Total de predios
47.884	1594
47.924	4706
47.965	7978
48.005	11737
48.045	7677
48.086	6289
48.126	5920
48.167	5031
48.207	5843
48.247	1748

LATITUD

Variable: Latitud
Mínimo: 4,786
Máximo: 4,827
Media: 4,806
Desviación Estándar: 0,009
Unidad: grados
Total = 58523



Coordenada Longitud	Total de Predios
-75.76	831
-75.75	2468
-75.74	7868
-75.729	13435
-75.719	4143
-75.709	10113
-75.699	8580
-75.689	4411
-75.679	2825
-75.669	3849

LONGITUD
Variable: Longitud
Mínimo: -75,765
Máximo: -75,664
Media: -75,714
Desviación Estándar: 0,022
Unidad: grados
Total = 58523