DETECCIÓN AUTOMÁTICA DEL NIVEL DE ESTRATIFICACIÓN SOCIOECONÓMICO URBANO USANDO REDES NEURONALES CONVOLUCIONALES SOBRE IMÁGENES SATELITALES CON INFORMACIÓN AUMENTADA

DANIEL ALCIDES CARVAJAL PATIÑO

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE INGENIERIAS FISICOMECANICAS
ESCUELA DE INGENIERIA DE SISTEMAS E INFORMÁTICA
BUCARAMANGA
2018

DETECCIÓN AUTOMÁTICA DEL NIVEL DE ESTRATIFICACIÓN SOCIOECONÓMICO URBANO USANDO REDES NEURONALES CONVOLUCIONALES SOBRE IMÁGENES SATELITALES CON INFORMACIÓN AUMENTADA

DANIEL ALCIDES CARVAJAL PATIÑO

TRABAJO DE GRADO PARA OPTAR POR EL TITULO DE INGENIERIA DE SISTEMAS

Director
FABIO MARTINEZ CARRILLO
Ph.D. EN INGENIERÍA DE SISTEMAS Y COMPUTACIÓN
Codirector
Ph.D RAUL RAMOS POLLAN

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE INGENIERIAS FISICOMECANICAS
ESCUELA DE INGENIERIA DE SISTEMAS E INFORMÁTICA
BUCARAMANGA
2018

ESPACIO PARA NOTA

ESPACIO PARA CARTA AUTORIZACIÓN USO DE DATOS

CONTENIDO

pág.

LISTA DE FIGURAS

pág.

LISTA DE TABLAS

pág.

RESUMEN

TITULO: DETECCIÓN AUTOMÁTICA DEL NIVEL DE ESTRATIFICACIÓN SOCIOECONÓMICO URBANO USANDO REDES NEURONALES CONVOLUCIONALES SOBRE IMÁGENES SATELITALES CON INFORMACIÓN AUMENTADA.

AUTORES: DANIEL ALCIDES CARVAJAL PATIÑO.

PALABRAS CLAVE: DANE, Machine learning, Deep Learning, Red Neuronal Convolucional, .

DESCRIPCION:

La finalidad de este proyecto de grado, es continuar con la investigación que el grupo Conuss ha venido desarrollando durante varios semestres en la creación de una infraestructura nube para la comunidad estudiantil, que permita su uso para el desarrollo de otros proyectos e investigaciones.

Debido a la necesidad de crear diplomados y laboratorios virtuales de computación para fomentar el avance en la formación de ingenieros de calidad, se decide utilizar soluciones de código abierto que permitan el fácil acceso y administración de los servidores físicos y virtualización. Entre las muchas soluciones, se decide optar por OpenStack gracias a su amplia gama de módulos y su abundante comunidad por el cual es respaldado, a su vez, integrando docker como solución para la creación de contenedores.

Todo esto con el fin de que la comunidad estudiantil tenga acceso a recursos de computo que no están al alcance de sus manos, otorgándoles la capacidad de conocer, además de disfrutar, las nuevas tecnologías que hoy por hoy están mejorando y automatizando los procesos de las grandes industrias tecnológicas.

ABSTRACT

TITLE: AUTOMATIC DETECTION OF THE URBAN SOCIOECONOMIC STRATIFICATION LEVEL USING CONVOLUTIONAL NEURAL NETWORKS ON SATELLITE IMAGES WITH INCREASED INFORMATION.

AUTHORS: DANIEL ALCIDES CARVAJAL PATIÑO.

KEYWORDS: Container, Cloud Computing, OpenStack, modules and services.

DESCRIPTION:

The purpose of this thesis is to continue with the research that the Conuss group has incorporated during several semesters in the creation of an infrastructure for the student community that allows its use for the development of other projects and research.

Due to the need to create certified courses and virtual computer labs to promote the advancement in the training of quality engineers, it is decided to use open source solutions that allow easy access and administration of physical servers and virtualization. Among the many solutions, it is decided to opt for OpenStack thanks to its wide range of modules and its abundant community for which it is backed, in turn, integrating docker as a solution for the creation of containers.

All this in order that the student community has access to computing resources that are not available to them, that they can know as well as enjoy the new technologies that are improving today and automating the processes of the large technological industries.

INTRODUCCION

La medición del nivel económico de una zona urbana, actualmente, conlleva un trabajo extenso, como lo expresa el DANE, "en el caso de las revisiones generales urbanas, así como en la estratificación rural se apoya en censos de vivienda" [1]. Es decir, se requiere la elaboración de una encuesta de gran tamaño, la cual consume mucho tiempo y personal. Posteriormente, si la encuesta no se realizó usando software de recolección de datos, es necesario realizar su tipeo, lo cual tambien requiere tiempo. Luego, como lo indica el DANE [2] [3], el cálculo final del estrato se realiza mediante modelos estadísticos y económicos especialmente calibrados para esta tarea.

En este contexto surgen varias interrogantes respecto a la capacidad de actualización de esta metodología: ¿Qué sucede cuando una ciudad tiene una alta tasa de desarrollo urbano?, ¿Cómo mantiene el gobierno actualizada la información de los estratos ante éstas circunstancias?, ¿Que tan efectiva es la metodología actual ante estos casos de alto desarrollo urbano?

Por tanto, el objetivo de este trabajo consiste en seleccionar redes neuronales convolucionales y evaluar su capacidad para determinar automáticamente el estrato socioeconómico usando imágenes satelitales e información adicional (información catastral, presencia y consumo de servicios, etc), con el fin de presentar una alternativa que haga frente a las inquietudes planteadas.

No es la primera vez que se realiza una predicción del nivel socioeconomico utilizando tecincas de machine learning o deep learning. Neal Jean en colaboración con varias

[1] DANE. Estratificación - Preguntas frecuentes. [en línea]. <https://www.dane.gov.co/files/geoestadistica/Preguntas_frecuentes_estratificacion.pdf>[citado en 25 de Mayo de 2018]

[2] DANE. Metodología de estratificación. [en línea]. <http://www.dane.gov.co/index.php/servicios-al-ciudadano/servicios-de-informacion/estratificacion-socioeconomica>[citado en 25 de Mayo de 2018].

[3] DANE. Procedimiento del cálculo. [en línea]. <http://www.dane.gov.co/files/geoestadistica/estratificacion/procedimientoDeCalculo.pdf>[citado en 25 de Mayo de 2018].

personas e instituciones realizó un modelo [4] [5] [6] capaz de predecir el nivel de pobreza en cinco países de África, usando imágenes satelitales y datos extra para dicha tarea. En Colombia, más específicamente en Medellín, también se han realizado modelos [7] para determinar niveles socioeconomicos de una zona urbana. En la Universidad EAFIT, usando tecnicas de Machine Learning e imagenes satelitales logran medir los indices de pobreza de dicha ciudad.

[4]NEAL jean. Combining satellite imagery and machine learning to predict poverty. [en Linea]. <http://sustain.stanford.edu/predicting-poverty/>[citado en 25 de Mayo de 2018]

[5]NEAL jean. Combining satellite imagery and machine learning to predict poverty. [en linea]. <https://github.com/nealjean/predicting-poverty>[citado en 25 de Mayo de 2018]

[6]NEAL Jean, MARSHALL Burke, † MICHAEL Xie, W. Matthew Davis, DAVID B. Lobell, STEFANO Ermon. Combining satellite imagery and machine learning to predict poverty. Science 353 (6301), p. 790-794. 2016

[7]EAFIT. Con imágenes satelitales miden los índices de pobreza en Medellín. [en linea]. <http://www.eafit.edu.co/investigacion/revistacientifica/edicion-167/Paginas/con-imagenes-satelitales-miden-los-indices-de-pobreza-en-medellin.aspx>[citado en 25 de Mayo de 2018]

1. OBJETIVOS

1.1. OBJETIVO GENERAL

1. Seleccionar y evaluar redes convolucionales para la determinación del nivel socio económico urbano mediante el uso de imágenes satelitales e información adicional.

1.2. OBJETIVOS ESPECIFICOS

1. Identificar fuentes de datos de imágenes satelitales e información adicional.
2. Diseñar y construir datasets integrando los datos obtenidos de las fuentes identificadas.
3. Seleccionar entre distintas arquitecturas de redes neuronales convolucionales existentes en la literatura y repositorios tecnológicos .
4. Entrenar las redes convolucionales probando configuraciones de datasets.
5. Evaluar el desempeño de las redes convolucionales con el uso de los distintos dataset.
6. Elegir la mejor configuración tanto de red convolucional como de conjunto de datos, teniendo en cuenta el desempeño obtenido.

2. MARCO TEÓRICO

2.1. ESTRATIFICACIÓN SOCIAL

"La estratificación social es un fenómeno presente en todas las sociedades. Los miembros se clasifican a sí mismos y a los otros basándose en jerarquías que vienen dadas por diversos factores" [8] y no es algo nuevo, la antigua Mesopotamia contaba con una división social cuyos miembros iban desde el rey y su familia, en el estrato más alto, hasta los esclavos en el más bajo.

Más que una simple división, la estratificación representa la desigualdad existente en una sociedad. Cada uno de los estratos, niveles o grupos sociales indica la capacidad de acceso a recursos, oportunidades, bienes y servicios por parte de las personas pertenecientes a cada nivel. Dicha estratificación es necesaria para la tarea que llevan acabo los gobiernos en contra de la desigualdad, dado que se suelen crear programas que beneficien a las personas de los niveles más bajos y cobrar mayores impuestos a las personas de los niveles más altos.

En el caso de Colombia, se maneja una estratificación de 6 niveles.

> "De éstos 6, los estratos 1, 2 y 3 corresponden a estratos bajos que albergan a los usuarios con menores recursos, los cuales son beneficiarios de subsidios en los servicios públicos domiciliarios; los estratos 5 y 6 corresponden a estratos altos que albergan a los usuarios con mayores recursos económicos, los cuales deben pagar sobrecostos (contribución) sobre el valor de los servicios públicos domiciliarios. El estrato 4 no es beneficiario de subsidios, ni debe pagar sobrecostos, paga exactamente el valor que la empresa defina como costo de prestación del servicio."[9]

Dicha estratificación, definir el estrato al que pertenece una vivienda, no es tarea fácil. Se requiere gran cantidad de variables por cada una, como las caracteristicas de la zona en la que se ubica, el tamaño, materiales en que fue fabricada, entre otras. Lo que

[8] GONZÁLEZ Vanessa. ¿Qué es la estratificación social? [en línea]. <`https://www.lifeder.com/estratificacion-social/`>[citado en 31 de Mayo del 2018].

[9] DANE, Estratificación - Preguntas frecuentes. Op. cit., p. 1.

conlleva a una tarea de recolección de datos bastante amplia.

2.2. MACHINE LEARNING

> "Machine learning es una disciplina científica del ámbito de la Inteligencia Artificial que crea sistemas que aprenden automáticamente. Aprender en este contexto quiere decir identificar patrones complejos en millones de datos. La máquina que realmente aprende es un algoritmo que revisa los datos y es capaz de predecir comportamientos futuros. Automáticamente, también en este contexto, implica que estos sistemas se mejoran de forma autónoma con el tiempo, sin intervención humana."[10]

Se han llevado muchos desarrollos y avances en distintos campos usando técnicas de machine learning. Uno de las más recientes y populares fue la máquina de Google AlphaGo que venció al mejor jugador a nivel mundial de GO! Ke Jie. Wikipedia usa técnicas de machine learning para detectar saboteos en su enciclopedia. Otros usos de machine learning consisten en la detección de objetos, patrones o enfermedades incluso predicción de tráfico urbano y precios de bienes.

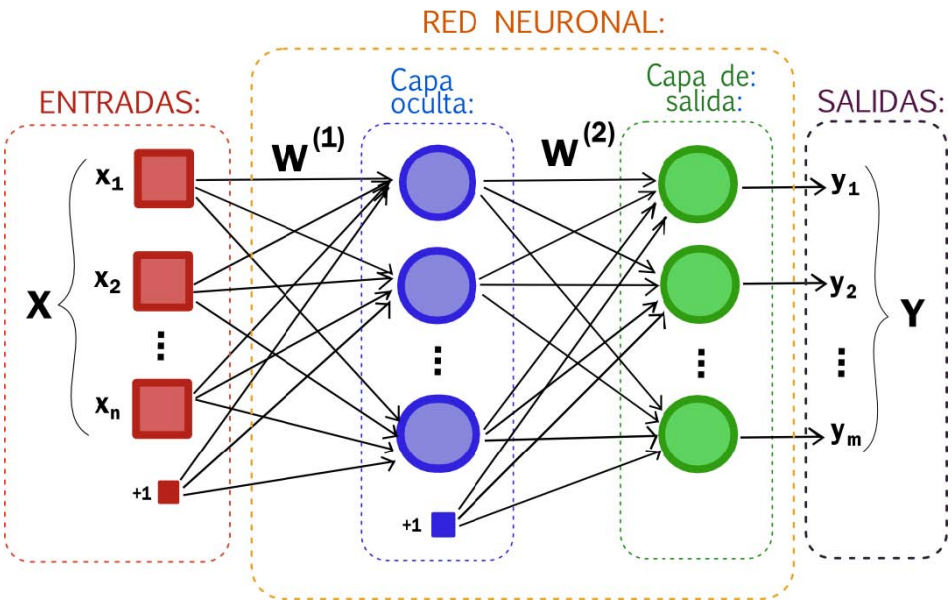
Aunque en el fondo, independientemente del campo en el que se trabaje, las técnicas de machine learning son las mismas, existen tareas muchos más complejos que otras. Por ejemplo, hoy en día, es mucho más fácil predecir el valor de una casa a predecir el valor del dólar o predecir terremotos. Esta dificultad se da debido a que, para cada tarea que se realice se deben "ajustar" los datos y los algortimos para que estos aprendan y puedan realizar predicciónes o detecciónes con un nivel de tolerancia aceptable.

2.2.1. redes neuronales es una de las técnicas o algortimos de machine learning que se pueden emplear en las tareas de predicción o detección. "Una red neuronal es un modelo simplificado que emula el modo en que el cerebro humano procesa la

[10]GONZÁLEZ Andrés. ¿Qué es Machine Learning? [en línea]. (Recuperado en 10 oct 2017) `http://cleverdata.io/que-es-machine-learning-big-data/`

información: Funciona simultaneando un número elevado de unidades de procesamiento interconectadas que parecen versiones abstractas de neuronas"[11]

Figura 1: Diagrama de red neuronal



VÍLCHEZ GARCÍA, Víctor Gabriel.*Estimación y clasificación de daños en materiales utilizando modelos AR y redes neuronales para la evaluaciónno destructiva con ultrasonidos.* [en linea]. (Recuperado en 24 may 2018) `http://ceres.ugr.es/~alumnos/esclas/`

Una red neuronal está constituida por una serie de capas que se activan con determinadas entradas generando determinadas salidas que podrían ser tomadas o no por otras capas, dependiendo de la "profundidad" de la red. Utilizar redes neuronales consiste en hacer que la misma aprenda examinando las entradas, prediciendo las salidas y haciendo ajustes a los distintos parámetros de la misma, este proceso se repite muchas veces hasta que la red sea capaz de predecir o clasificar con un margen de error tolerable.

[11]IBM. El modelo de redes neuronales [en línea]. <`https://www.ibm.com/support/knowledgecenter/es/SS3RA7_18.0.0/modeler_mainhelp_client_ddita/components/neuralnet/neuralnet_model.html`>[citadoen31deMayode2018]

La idea del proyecto es obtener una red neuronal capaz de predecir el nivel socioeconómico de una zona urbana, para esto se planea alimentar la red con imágenes satelitales e información extra de dichas zonas urbanas. Con esta información, ajustes en los parámetros y bastantes iteraciones la red neuronal aprenderá. Los cambios en los parámetros, profundidad y datos usados en la red neuronal son realizados para encontrar la red neuronal que mejor desempeño presente para la tarea propuesta. Dado que los datos a utilizar son imágenes, es recomendable utilizar "redes neuronales convolucionales" (CNN), las cuales son un tipo de red neuronal que se adapta mejor al uso de imágenes dado que "las CNN eliminan la necesidad de una extracción de características manual, por lo que no es necesario identificar las características utilizadas para clasificar las imágenes. La CNN funciona mediante la extracción de características directamente de las imágenes. Las características relevantes no se entrenan previamente; se aprenden mientras la red se entrena con una colección de imágenes"[12]

[12]MATHWORKS. Aprendizaje profundo [en línea]. <`https://es.mathworks.com/discovery/deep-learning.html`>[citado en 31 de Mayo de 2018]

3. DESARROLLO DEL PROYECTO

Esta claro que la recoleccion de datos medieante encuestas a vivienda es una metodologia que conlleva bastane tiempo y que el calculo del estrato debe tener en cuenta un sin numero de variables. Como alternativa se propone usar imagenes satelitales dado que se puede obtener mucha información socioeconomica con el analizis de las mismas. Se pueden identificar patrones, detectar construcciones especificas, clasificar materiales de construccion en los tejados y más, todos estos analizis son variables usadas en el calculo del estrato.

Como se mencionó existe un par de trabajos sobre el nivel socioeconomico usando tecnicas de machine learning. El trabajo de Neal Jean usa redes neuronales convolucionales y Transfer Learning para lograr una prediccion de la pobreza en 5 paises de africa. El trabajo de la EAFIT usa tecnicas de machine learning para detectar caracteristicas que puedan dar a conocer el nivel socioeconomico de una zona. Hay que tener en cuenta que en ambos trabajos la principal fuente de datos son las imagenes satelitales. En este proyecto tambien se usaron imagenes satelitales.

A grandes rasgos lo que se planteo realizar con las imagenes satelitales, la informacion aumentada, la informacion de los estratos y las redes neuronales, se muestra en las siguientes figuras.

Figura 2: Datos a usar en el proyecto

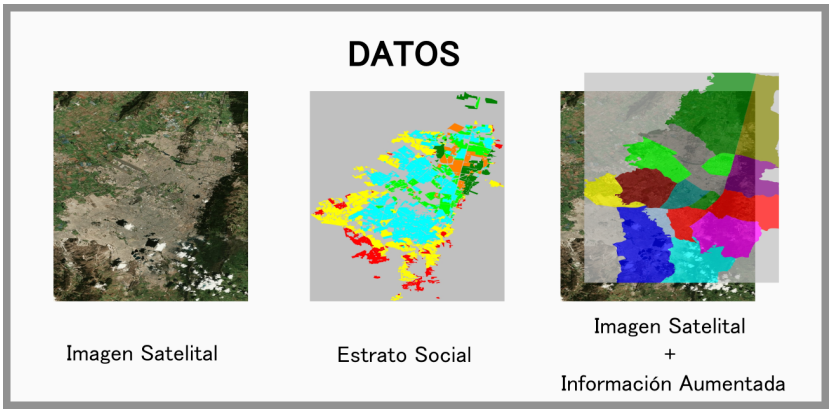


Imagen Propia

Figura 3: Modelo a usar en el proyecto

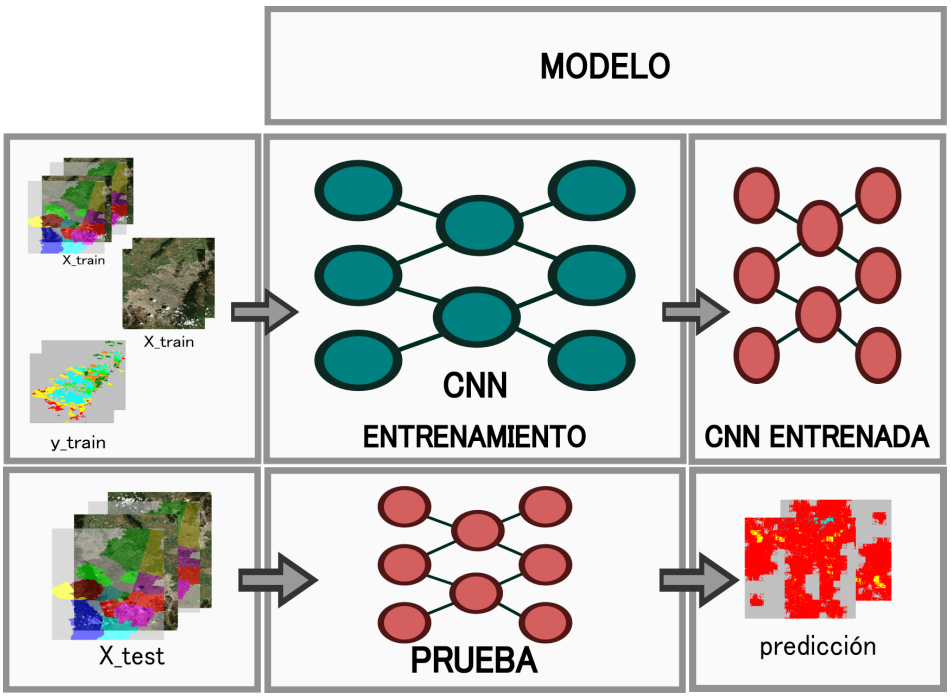


Imagen Propia

El proyecto se desarrollo encontrando fuentes de datos, generando dataset con imagenes satelitales, informacion aumentada y distintos niveles de zoom, seleccionando redes neuronales y realizando pruebas para determinar la mejor manera de predecir el estrato social. Hay que mencionar que el desarrollo del proyecto esta documentado en el sigueinte repositorio de Github `https://github.com/DaielChom/proyecto_uis`. El desarrollo se llevo de la siguiente manera.

3.1. FUENTES DE DATOS

Era necesario tener imagenes satelitales, imagenes con los estratos sociales e imagenes con informacion extra que correspondan geograficamente y que tengan distintos niveles de zoom, como se muestran en la Figura 2. Se usan distintos niveles de zoom para determinar si esto influye en la prediccion del estrato social. Existen muchas plataformas para obtener imagenes satelitales tanto de alta como baja resolucion, varias de estas plataformas tienen la opcion de insertar información en dichos mapas de forma de marcador, poligono o linea. Esta informacion por lo general son archivos en formato .kml o .shp. Aprochando esta opcion se realizaron busquedas de archivos kml con distintos tipos de informacion, incluyendo el estato social. Gracias a la politica de datos abiertos[13] es psobile encontrar archivos kml o shp en distintas paginas web del gobierno. sin embargo no todos los departamentos o ciudades cuentan con la misma cantidad de datos disponibles y menos con datos del estrato social. El unico portal web donde se encontro una buena cantidad de informacion, incluyendo la del estrto social fue en el portal de mapas de Bogota, disponible en `http://mapas.bogota.gov.co`.

No fue posible descargar los archivos kml mostrados en la plataforma, por ende, y aprovechando la licencia Creative Commons que presentan los datos, se crearon nuevos archivos kml usando los mapas de la plataforma como guia. Los archivos kml realizados estan disponibles en `https://drive.google.com/open?id=15VnvN6ZRTbsqqd9kl3ukNWBj3oqADsy0`. se crearon dos kml, uno con informacion de los estratos sociales y otro con los indices de condiciones de seguridad nocturna por localidad que tienen las mujeres,

[13]GOBIERNO DIGITAL. Datos Abiertos [en linea] <`http://estrategia.gobiernoenlinea.gov.co/623/w3-article-9407.html`>[citado en 30 de Mayo de 2018]

mas especificamente los datos de la categoria *riesgo alto* de dicho mapa.

3.2. DATASET

Como en la mayoria de trabajos de machine learning se debe tener un dataset, un conjunto de datos organizados y con una estrucura dividida en train (datos para entrenamiento) y test (datos para pruebas) con los cuales se entrena o prueba un algortimo de aprendizaje. El dataset que se construyo lleva por nombre OVERLAECOBO y cuenta con 3 tipos imagenes (Figura 2) geograficamente correspondidas y de distintos niveles de zoom o acercamiento satelital. Las imagenes del dataset fueron obtenidas usando el software Ruso SASplanet[14] y los archivos kml diseñados. Usando *Bing maps* como fuente de imagenes satelitales, fuente ofrecida en SASplanet, se tomo el mapa de Bogota y se traso una linea que dividiera la ciudad en dos partes, de tal manera que en ambas hubiera informacion de todos los estratos. Una parte se escogio para train y otra la para test, como se muestra en la sigueinte imagen.

Figura 4: Division de Train y Test

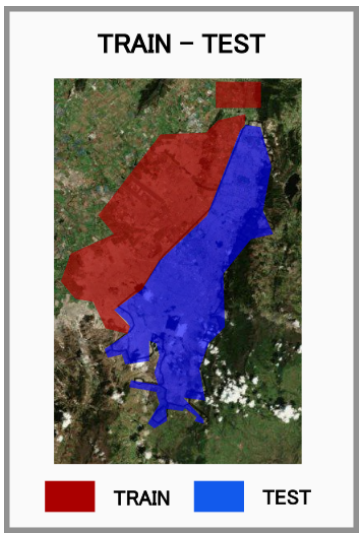


Imagen Propia

[14]Disponible en http://www.sasgis.org/

Se extrayeron imagenes de 7 Zooms diferentes, del 13 al 20, Cada nivel de zoom cuenta con imagenes satelitales, de informacion extra y de estrato social. Cada una de estas dividida en train y test. OVERLAECOBO cuenta con la siguente estructura.

Figura 5: Estrucutra OVERLAECOBO

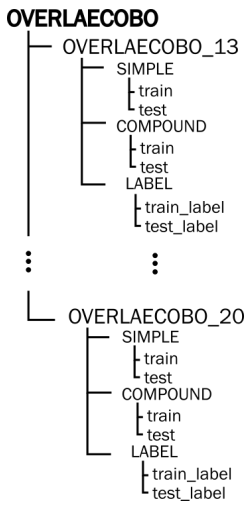


Imagen Propia

OVERLAECOBO es el dataset con el que se realizaron las pruebas en busca de la mejor manera de detectar el estrato social.

3.3. REDES NEURONALES

3.4. DETECCION DEL ESTRATO SOCIAL

3.5. RESULTADOS

4. CONCLUSIONES

Tabla 1: Red Conuss

Red Conuss			
	VLAN 1	VLAN 2	VLAN3
Direccion	10.6.100.0	10.6.101.0	10.6.102.0
Mascara	255.255.255.0	255.255.255.0	255.255.255.0
Gateway	10.6.100.1	10.6.101.1	10.6.102.1
Uso	Administración	Usuarios	Pruebas

Las interfaces de los nodos de OpenStack están definidas como en la **TABLA 7**.

Tabla 2: Servidores e Interfaces

Interfaces de Red		
	Interfaz 1	Interfaz 2
Labroides	10.6.100.2	Proveedor de 10.6.101.0/24
Lactoria	10.6.100.3	Proveedor de 10.6.101.0/24
Nautilus	10.6.100.5	Proveedor de 10.6.101.0/24
Sistemas	10.6.100.4	N/A

5. RECOMENDACIONES Y TRABAJO FUTURO

6. LIMITACIONES Y PROBLEMAS

BIBLIOGRAFIA

DANE. Estratificación - Preguntas frecuentes. [en línea]. <https://www.dane.gov.co/files/geoestadistica/Preguntas_frecuentes_estratificacion.pdf>[citado en 25 de Mayo de 2018]

DANE. Metodología de estratificación. [en línea]. <http://www.dane.gov.co/index.php/servicios-al-ciudadano/servicios-de-informacion/estratificacion-socioeconomica>[citado en 25 de Mayo de 2018]

DANE. Procedimiento del cálculo. [en línea]. <http://www.dane.gov.co/files/geoestadistica/estratificacion/procedimientoDeCalculo.pdf>. [citado en 25 de Mayo de 2018].

NEAL jean. Combining satellite imagery and machine learning to predict poverty. [en Linea]. <http://sustain.stanford.edu/predicting-poverty/>[citado en 25 de Mayo de 2018].

NEAL jean. Combining satellite imagery and machine learning to predict poverty. [en linea]. <https://github.com/nealjean/predicting-poverty>[citado en 25 de Mayo de 2018]

NEAL Jean, MARSHALL Burke, † MICHAEL Xie, W. Matthew Davis, DAVID B. Lobell, STEFANO Ermon. Combining satellite imagery and machine learning to predict poverty. Science 353 (6301), p. 790-794. 2016

EAFIT. Con imágenes satelitales miden los índices de pobreza en Medellín. [en linea]. <http://www.eafit.edu.co/investigacion/revistacientifica/edicion-167/Paginas/con-imagenes-satelitales-miden-los-indices-de-pobreza-en-medellin.aspx>[citado en 25 de Mayo de 2018]

GONZÁLEZ Vanessa. ¿Qué es la estratificación social? [en línea]. <https://www.lifeder.com/estratificacion-social/>[citado en 31 de Mayo del 2018].

IBM. El modelo de redes neuronales [en línea]. <https://www.ibm.com/support/knowledgecenter/es/SS3RA7_18.0.0/modeler_mainhelp_client_ddita/components/neuralnet/neuralnet_model.html>[citadoen31deMayode2018]

MATHWORKS. Aprendizaje profundo [en línea]. <https://es.mathworks.com/discovery/deep-learning.html>[citado en 31 de Mayo de 2018]

GOBIERNO DIGITAL. Datos Abiertos [en linea] <http://estrategia.gobiernoenlinea.gov.co/623/w3-article-9407.html>[citado en 30 de Mayo de 2018]