UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE INGENIERÍAS FISICOMECÁNICAS
ESCUELA DE INGENIERÍA DE SISTEMAS E INFORMÁTICA

PLAN DE TRABAJO DE GRADO

FECHA DE PRESENTACIÓN: Bucaramanga,

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

TRABAJO DE INVESTIGACIÓN

AUTOR: Daniel Carvajal Patiño - 2132817 Firma: _____

CODIRECTOR: Raúl Ramos Pollán - Profesor Firma: _____

DIRECTOR: Fabio Martinez Carrillo - Profesor Firma: _____

ENTIDAD INTERESADA: Universidad Industrial de Santander.

COMITÉ DE TRABAJOS DE GRADO:

EVALUADOR ASIGNADO:

CONCEPTO DEL EVALUADOR:

APROBACIÓN DEL COMITÉ´:

FECHA: _____

Acta No. _____

TABLA DE CONTENIDO.

1. PLANTEAMIENTO Y JUSTIFICACIÓN DEL PROBLEMA...……...…..……**1**
2. OBJETIVOS……....…….……...……...……**2**
2.1 Objetivo General……...……...……..…….....**2**
2.2 Objetivos Específicos……..……..……..**2**
3. MARCO DE REFERENCIA……..……..……..**3**
3.1. Estratificación Social.……....…….....**3**
3.2 Machine Learning……..……..**3**
3.3 Predicción o Clasificación de la Estratificación……...…….**5**
4. METODOLOGÍA……..……..……....**6**
5. CRONOGRAMA……...…..……....**8**
6. PRESUPUESTO……...……...**9**
7. BIBLIOGRAFÍA……...……**10**

1. PLANTEAMIENTO Y JUSTIFICACIÓN DEL PROBLEMA

La medición del nivel económico de una zona 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. Según la metodología que usa 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 es desarrollar métodos basados en 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.)

2. OBJETIVOS

2.1. OBJETIVO GENERAL

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

2.2 OBJETIVOS ESPECÍFICOS

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

3. MARCO DE REFERENCIA

3.1. 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”[4] y no es algo nuevo, se han hecho estratificaciones hasta en las sociedades antiguas, la antigua Mesopotamia contaba con una división social con diferentes estratos 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 social presenta diferente capacidad de acceso a ciertos “recursos” u oportunidades. Por ejemplo un esclavo le era más difícil conseguir el pan que para un escriba o un noble.

En la actualidad la estratificación sigue presente en la sociedad, siguen existiendo personas con mayor capacidad de acceso a bienes y servicios, que otras, por ende sigue existiendo una desigualdad. Pero dicha estratificación es necesaria para la tarea que están llevando los gobiernos en contra del hambre, la pobreza y parte a la misma desigualdad. Dado que se suelen crear programas que beneficien a las personas de menores estratos y se suele cobrar mayor cantidad de impuestos a las de mayores estratos.

En el caso de Colombia, se maneja una estratificación de 6 estratos. “De éstos, 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.”[1]

Dicha estratificación no es tarea fácil, para el cálculo de la misma (definir el estrato al que pertenece una persona u hogar) se requiere gran cantidad de variables por cada persona, como su nivel educativo, la zona en la que vive, el tamaño del hogar, materiales en que fue fabricada la casa en que habita, entre otros. Lo que conlleva a una tarea de recolección de datos bastante amplia. Las datos posteriormente son analizados o procesados para generar un resultado, casa por casa.

3.2 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.”[5]

Se han llevado muchos desarrollos y avances en distintos campos usando técnicas de machine learning, una 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 también usa técnicas de machine learning, en este caso se utiliza para detectar saboteos en su enciclopedia. Se han realizado trabajos en la detección de objetos, cáncer, patrones, etc. trabajos de predicción de tráfico urbano, precios de bienes, etc.

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 parámetros del algoritmo para que este "aprenda" lo que deba aprender y no exista un sobre aprendizaje o un infra aprendizaje que afecte la predicción o la detección.

Las "redes neuronales" son una de dichas técnicas que se pueden emplear en las tareas de machine learning. "Una **red neuronal** es un modelo simplificado que emula el modo en que el cerebro humano procesa la información: Funciona simultaneando un número elevado de unidades de procesamiento interconectadas que parecen versiones abstractas de neuronas"[6].

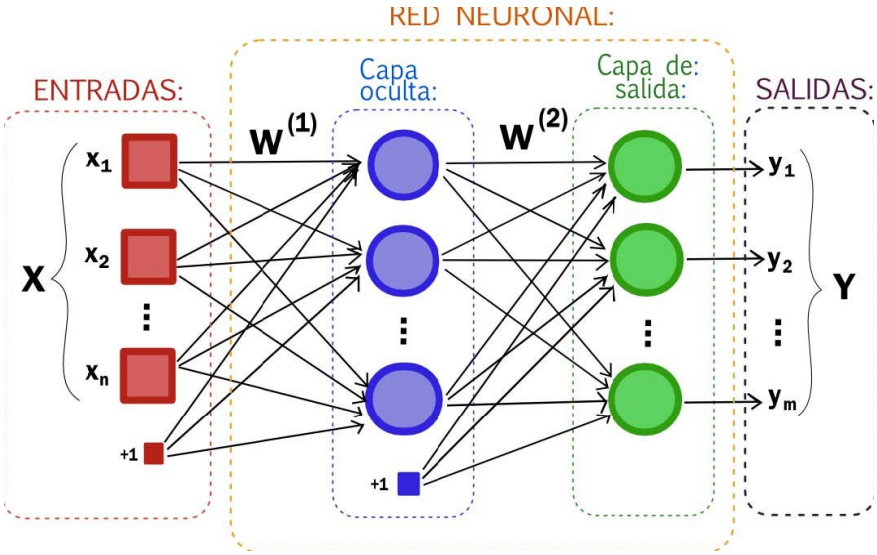


Fig 1. Diagrama de una red neuronal. imagen tomadada de <http://ceres.ugr.es/~alumnos/esclas/>

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.

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, con cambios en los parámetros y con 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”[7].

3.3 PREDICCIÓN O CLASIFICACIÓN DE LA ESTRATIFICACIÓN.

Aquellas tareas que requieran de predecir o clasificar son muy atractivas para las personas en el campo de machine learning y la detección o clasificación de la pobreza no deja de ser una de ellas, surge una pregunta y es ¿por que imágenes satelitales?, pues se podrían usar otro tipo de formato de datos para esta tarea, podrían usarse tablas de datos con información de la economía de un país, o datos sobre el índice de educación de los ciudadanos, entre otros datos. La necesidad de usar imágenes satelitales y como se mencionó como propósito de este trabajo nace en la dificultad de recolección de dichos datos, aparte de esto las imágenes satelitales son una buena fuente de información de niveles socioeconómicos de una zon urbana, pues en una imagen satelital se puede observar la aglomeración de las viviendas, la presencia de piscinas en las casas, el material de fabricación de los techos de las mismas, en imágenes nocturnas se puede observar la cantidad de energía eléctrica consumida, entre otros factores que indican un nivel social. la idea es que la red neuronal “aprenda” estos factores y arroje una buena predicción. No es la primera vez que se realiza una predicción del nivel social, incluso han habido varios. Neal Jean en colaboración con otras personas y varias instituciones realizó un modelo [8,9, 10,] para predecir la pobreza en cinco países de África usando imágenes satelitales y datos extra para dicha tarea, los datos extra que se usaron para ese trabajo fueron datos de indicadores de pobreza como los que se mencionaron anteriormente. Para la tarea de predicción utilizaron redes neuronales obteniendo un nivel de acierto del 75% . En Colombia, más específicamente en Medellín, también se han realizado modelos [11, 12] o estudios para determinar los índices de pobreza en dicha ciudad, uno fue elaborado por Miguel Noreña, aunque este estudio no se centra en la utilización de técnicas de machine learning si se centra en la predicción de la pobreza. El otro trabajo fue elaborado en medellín por Jorge Eduardo Patiño Quinchía utilizando imágenes satelitales. La idea del proyecto actual es lograr la predicción del “estrato social Colombiano”.

4. METODOLOGÍA

Para el desarrollo de este proyecto se realizará una serie de subtareas. Se llevará un control de la ejecución de dichas subtareas en una serie de reuniones con el director del proyecto. Dichas tareas serán planteadas a favor del cumplimiento de los objetivos en el plazo requerido. En primer lugar se tienen 6 tareas principales, (los objetivos específicos) que se llevarán a cabo mediante la realización de subtareas, quedando de la siguiente manera.

1. Identificar fuentes de datos de imágenes satelitales e información adicional.

- **1.1.** Consultar la disponibilidad y formato de datos de estratificación en colombia o en ciudades de colombia en distintas posibles fuentes.
- **1.2.** Determinar la fuente y formato de datos de estratificación para su uso en el proyecto.
- **1.3.** Consultar disponibilidad y capacidad de obtención de imágenes satelitales con herramientas, instituciones, empresas o grupos.
- **1.4.** Seleccionar la fuente de imágenes satelitales para su uso en el proyecto.
- **1.5.** Consultar disponibilidad de información adicional, como el nivel de riesgo o seguridad de las zonas urbanas a estudiar en distintas posibles fuentes.
- **1.6.** Escoger la fuente y formato de la información extra a usar.

2. Diseñar y construir datasets integrando los datos obtenidos de las fuentes identificadas.

- **2.1.** verificar Accesibilidad a los distintos datos a usar.
- **2.2.** Obtener y organizar los datos de tal manera que facilite su acceso desde los algoritmos o herramientas usadas en la elaboración del proyecto.
- **2.3.** verificar el acceso de los datos desde las herramientas a usar para el proyecto, herramientas como Jupyter notebook.
- **2.4.** Realizar un pequeño estudio y análisis al dataset.

3. Diseñar distintas arquitecturas de redes convolucionales

- **3.1.** Consultar sobre la configuración de redes convolucionales y como estas se ven afectadas con dichas configuraciones.
- **3.2.** probar distintas configuraciones de redes neuronales, haciendo combinaciones en cambios de parámetros, profundidad y neuronas, en búsqueda de la que mejor se adapte al objetivo principal.
- **3.3.** Escoger la mejor red teniendo en cuenta desempeño, rendimiento, etc.

4. Entrenar las redes convolucionales probando configuraciones de datasets.

- **4.1.** Entrenar las redes convolucionales probando configuraciones de datasets.
- **4.2.** Obtener resultados de la red con los datasets.

5. Evaluar el desempeño de las redes convolucionales con el uso de los distintos dataset.

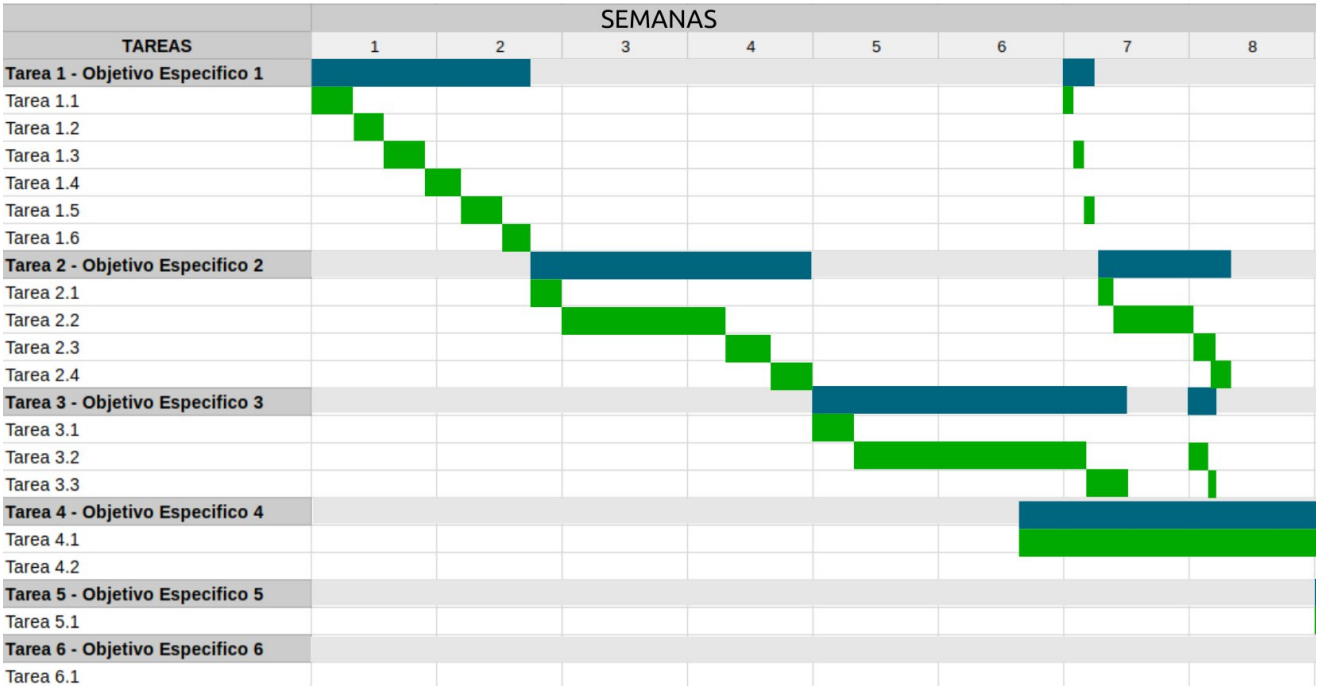
- **5.1.** Comparar el desempeño de la red con cada uno de los distintos dataset. haciendo análisis donde se muestran posibles ventajas y desventajas de cada uno de los datasets.

6. Elegir la mejor configuración tanto de red convolucional como de conjunto de datos, teniendo en cuenta el desempeño obtenido.

6.1. Escoger el dataset con el que la red presentó mejor desempeño, basándose en los análisis realizados.

5. CRONOGRAMA

Las reuniones para el seguimiento de las tareas serán de una a dos horas semanales. y la elaboración



de las tareas se llevará a cabo de la siguiente manera.

Fig 2. Cronograma de las primeras 8 semanas.

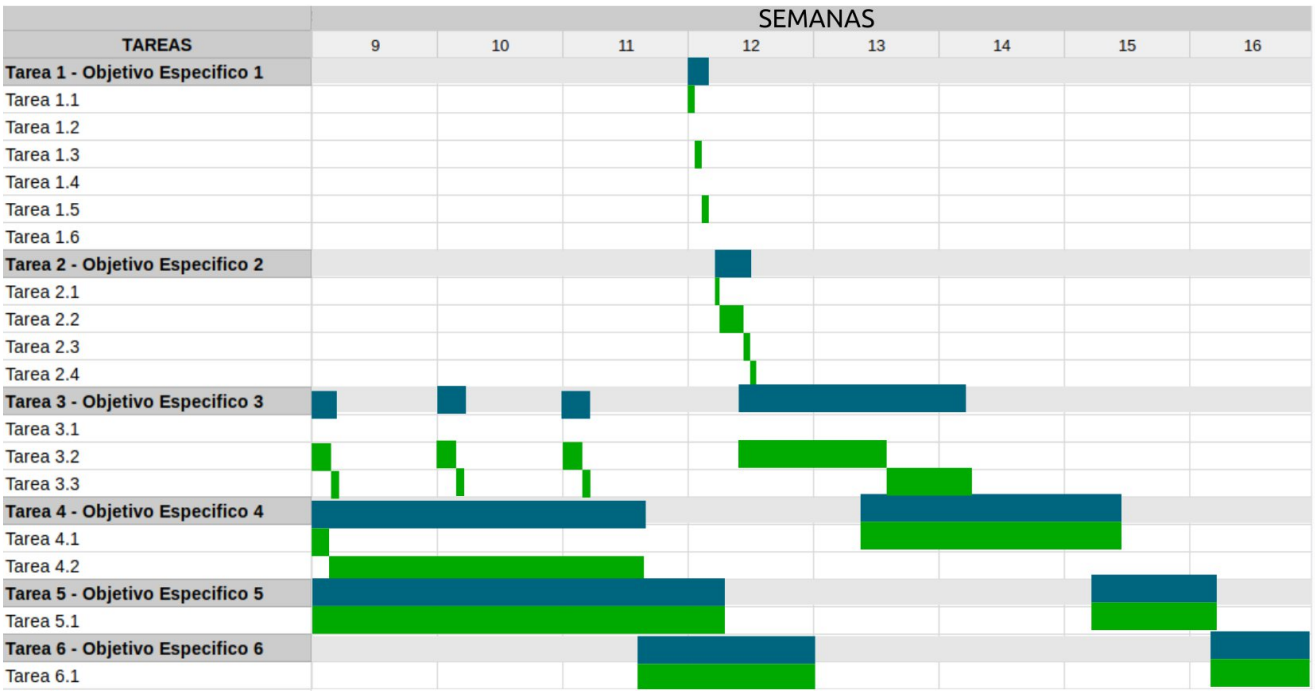


Fig 3. Cronograma de las segundas 8 semanas.

6. PRESUPUESTO

Para llevar a cabo el presente proyecto se requieren de ciertos recursos tanto físicos como humanos. Se quiere de material de cómputo para la elaboración de las consultas. diseños, construcción, entrenamientos y análisis de los distintos datos y algoritmos a usar. También se requiere de un tiempo de dedicación y esfuerzo tanto de parte del director del proyecto como del estudiante a realizar el proyecto. En la siguiente tabla se enuncian de mejor manera los recursos a usar y su respectivo costo.

EQUIPO		
Descripción	Cantidad	Valor
Computador	1	$2.500.000,00

Tabla 1. Presupuesto para el equipo a usar en el proyecto

OTROS		
Descripción	Valor pasaje	Total en 16 semanas
transporte	2100	$336.000,00

Tabla 2. Presupuesto para otros costos del proyecto

RECURSOS HUMANOS			
Cargo	Valor hora	Horas Semana	Total en 16 semanas
Director Proyecto	$200.000,00	2	$6.400.000,00
Autor Proyecto	$12.000,00	30	$5.760.000,00
Total			$12.160.000,00

Tabla 3. Presupuesto para el recurso humano del proyecto

TOTAL	
Descripción	Valor
Recursos Humano	$12.160.000,00
Equipo	$2.500.000,00
Otros	$336.000,00
TOTAL	$14.996.000,00

Tabla 3. Presupuesto Total

7. BIBLIOGRAFÍA

[1]. DANE. Estratificación - Preguntas frecuentes. [en línea]. <https://www.dane.gov.co/files/geoestadistica/Preguntas_frecuentes_estratificacion.pdf> [citado: 05 oct 2017].

[2]. DANE. Metodología de estratificación. [en línea]. <http://www.dane.gov.co/index.php/servicios-al-ciudadano/servicios-de-informacion/estratificacion-socioeconomica#metodolog%C3%ADa> [citado: 05 oct 2017].

[3]. DANE. Procedimiento del cálculo. [en línea]. <http://www.dane.gov.co/files/geoestadistica/estratificacion/procedimientoDeCalculo.pdf> [citado: 05 oct 2017].

[4]. GONZÁLEZ Vanessa. ¿Qué es la estratificación social? [en línea]. <https://www.lifeder.com/estratificacion-social/> [citado: 10 oct 2017].

[5]. GONZÁLEZ Andrés. ¿Qué es Machine Learning? [en línea]. < http://cleverdata.io/que-es-machine-learning-big-data/ > [citado: 10 oct 2017].

[6]. 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> [citado: 15 oct 2017]

[7]. MATHWORKS. Aprendizaje profundo [en línea]. <https://es.mathworks.com/discovery/deep-learning.html> [citado: 15 oct 2017]

[8]. NEAL jean. Combining satellite imagery and machine learning to predict poverty. [en Linea] <http://sustain.stanford.edu/predicting-poverty/ > [citado: 10 oct 2017]

[9]. NEAL jean. Combining satellite imagery and machine learning to predict poverty. [en linea] <https://github.com/nealjean/predicting-poverty> [citado: 10 oct 2017]

[10]. 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

[11]. 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: 10 oct 2017]

[12]. NOREÑA Miguel. Deteccion y caracterizacion de zonas marginales en la ciudad de Medellın mediante el analisis exploratorio de datos espaciales [en linea] <http://www.banrep.gov.co/sites/default/files/eventos/archivos/TesisMiguelNorena_0.pdf> [citado: 10 oct 2017]