Universidad Santo Tomás
Sede Bogotá

Facultad de Ingeniería Ambiental

Informe final. **V01**

Propuesta metodológica basada en Redes Neuronales Artificiales para la determinación de la gestión adecuada de los residuos sólidos urbanos en zonas de recolección de la ciudad de Bogotá

Presentado por:
Jefferson David Rodríguez Chivatá
Gabriel Rodrigo Pedraza Ferreira
Fabio Martínez Carrillo

Escuela de Ingeniería de Sistemas e Informática (EISI)

Universidad Industrial de Santander

16 de octubre de 2018

1. Resumen Ejecutivo

Este documento presenta un informe detallado de los resultados alcanzados en la consultoría técnico-científica para la implementación de estrategias de aprendizaje de máquina que permitan la predicción de la generación de residuos sólidos en Bogotá. Durante la consultoría fueron organizados y tratados un conjunto de datos entregados por la USTA, los cuales fueron usados para desarrollar modelos de predicción de residuos sólidos desde diferentes perspectivas. Además estas predicciones permitieron un análisis de las cantidades y porcentajes de residuos en predicciones de 20 años. A continuación se detallan los objetivos planteados en la propuesta y alcances esperados. Luego se describe de forma detallada los logros alcanzados, así como también una descripción de las herramientas implementadas. Finalmente se presentan algunas conclusiones que le permitirán al equipo USTA analizar variables ambientales en términos de los modelos desarrollados.

1.1. Objetivo del proyecto

Brindar una consultoría técnico-científico para la implementación de modelos basados en redes neuronales para la predicción de la generación de residuos sólidos en las localidades de la ciudad de Bogotá.

1.2. Alcance del proyecto

En el desarrollo de la presente propuesta, la UIS proveerá una consultoría técnico-científico para la implementación de modelos basados en redes neuronales, teniendo en cuenta:

1. Tratamiento y organización de datos entregados por la USTA que permitirán entrenar los modelos
2. Estudio de la proyección de residuos según diferentes variables independientes proporcionadas por la USTA
3. Implementación de modelos utilizando librerías especializadas e implementado en ambientes computacionales interactivos, como los Notebooks.
4. Apoyo en el uso de las herramientas implementadas para el análisis e interpretación por parte de expertos ambientales en USTA

1.3. Productos a entregar

En el objeto del contrato en el cual se desarrolla el contrato, la escuela de ingeniería de sistemas se compromete a la entrega de un informe con el resultado del análisis de datos y un conjunto de archivos con la implementación computacional del modelo. Teniendo en cuenta los compromisos establecidos, este documento representa el informe final con los resultados de la consultoria técnico-científcia, mientras que los modelos de datos, datos procesados e implementación de modelos puede ser consultados en: https://goo.gl/mVxnL9. Particularmente en cuanto a la implementación de modelos y visualización de resultados se implementaron tres etapas:

- Exploración de datos
- Implementación de modelos de predicción
- Caracterización de residuos predichos

Índice

2. Estructura y organización de los datos

Teniendo en cuenta el objetivo de la consultoría, en cuanto al análisis descriptivo de los datos y la implementación de modelos de predicción sobre estos datos, fue necesario una re-estructuración de los mismo que facilite su trabajo computacional. Por lo tanto, en la primera fase de la consultoría técnico-científico se re-estructuraron los datos proporcionados por la USTA para poder llevar un procesamiento y uso de los mismo de forma algorítmica. A continuación se describen como fueron estructurados estos datos, los cuales se encuentran disponibles en el proyecto: https://goo.gl/mVxnL9

Los datos fueron re-estructurados en 4 archivos **csv** que permitirán una manipulación mas adecuadada para el análisis e implementación de modelos predictivos. A continuación se describe la nueva estructuracíón de los datos:

- **Archivo por Zonas**
  La información para cada zona fue organizada en una serie de filas contiguas correspondiente a cada año de observación, como se ilustra en la figura 1. El archivo estructurado puede ser visualizado en el siguiente enlace: goo.gl/K3VvJ4

  El orden de las zonas no necesariamente debe contener algún orden. Mientras que, los años y su respectiva información deben estar almacenados en orden ascendente. Es importante que cada zona contenga los mismos registros en años, de no ser posible se debe dejar filas en blanco. Los datos de Zonas a manejar en esta nueva estructura son presentados como sigue: Ver figura 2.

Figura 1: Estructura datos Zonas

	Zona	AÑO	Enero	Febrero	Marzo	Abril	Mayo	Junio	Julio	Agosto	Septiembre	Octubre	Noviembre	Diciembre	Población Por localidad
2	ASE 1	2012	36369	36184	35851	35439	39658	34214	37693	37058	34590	38308	37814	20321	1574318
3	ASE 1	2013	36858	35741	37391	38107	40381	37450	39755	38896	38542	39737	39609	43353	1605106
4	ASE 1	2014	38841	36245	40939	37575	39681	36077	39544	37550	38769	39142	0	0	1636511
5	ASE 1	2015	37682	36529	38848	36638	37254	37105	38331	37570	37207	36612	37017	37345	1668802
6	ASE 1	2016	33909	34727	37003	37751	39049	37270	36150	38006	36600	37808	40389	41838	1723642
7	ASE 2	2012	28469	28669	29259	27872	30314	27351	28507	29265	28172	31381	29929	22140	1205158
8	ASE 2	2013	27659	26472	27357	27596	26012	30391	31105	30717	29263	31576	30939	33043	1221102
9	ASE 2	2014	30254	27628	31184	29127	31961	28778	31354	29220	29937	30854	0	0	1237695
10	ASE 2	2015	29657	28206	30347	29251	29643	29450	30162	29867	29358	29269	28775	30328	1255208
11	ASE 2	2016	26501	27082	28575	28609	29919	29066	28562	29695	28667	28975	30811	32126	1276762
12	ASE 3	2012	24754	25069	26187	23340	26293	24860	24163	25873	24740	25916	25528	14291	750365
13	ASE 3	2013	26671	27030	31909	35801	31474	31194	27665	29339	28250	30367	30809	31558	754506
14	ASE 3	2014	28211	28174	31305	29018	31274	27024	29303	28976	29354	30094	0	0	758648
15	ASE 3	2015	27813	28105	29897	28189	28693	26934	28484	27972	28286	28253	27742	28193	762829
16	ASE 3	2016	24517	25876	27215	27868	28525	26499	26007	27251	26743	27113	28024	29551	744898

Figura 2: Ejemplo Archivo por Zonas

- **Archivo por Localidades**
  La estructura de datos para localidad es la misma que para zonas, la única diferencia es el identificador de cada localidad (ID Localidad), como se ilustra en la figura 3. El archivo estructurado puede ser visualizado en el siguiente enlace: goo.gl/JQba8N

  Los datos de cada localidad en esta nueva estructura son presentados como sigue: Ver figura 4.

Figura 3: Estructura datos Localidades

	ID Localidad	AÑO	Enero	Febrero	Marzo	Abril	Mayo	Junio	Julio	Agosto	Septiembre	Octubre	Noviembre	Diciembre	Población Por localidad
2	19	2012	12933	13870	13573	12767	14537	13240	13851	14254	12057	14797	14487	7086	479830
3	19	2013													484764
4	19	2014	14181	12820	14703	13431	14186	13367	13926	13395	14061	14466			489526
5	19	2015	12927	13002	13283	12681	13398	11064	13855	13339	12906	12809	13681	12995	494066
6	19	2016	12048	12890	12905	13574	13836	13370	12684	13609	12449	13515	14278	14869	472908
7	15	2012	23436	22314	22278	22672	25121	20973	23842	22804	22533	23511	23327	13234	1094488
8	15	2013													1120342
9	15	2014	24660	23426	26236	24144	25494	22710	25617	24156	24708	24676			1146985
10	15	2015	24755	23528	25564	23958	23856	26040	24476	24232	24300	23803	23336	24350	1174736
11	15	2016	21861	21836	24098	24177	25213	23900	23466	24396	24151	24294	26111	26969	1250734
12	7	2012	10014	10615	10430	9920	10725	9316	9591	10272	10071	11471	10862	7889	353859
13	7	2013													362167
14	7	2014	11893	10821	18640	17678	19632	17978	19715	18857	19060	19036			370976
15	7	2015	11455	10883	11779	11039	10743	10562	10825	10588	10357	10422	10703	10878	380453
16	7	2016	9383	9473	9817	10066	9800	9985	9625	10570	9965	10258	10651	11088	403519

Figura 4: Ejemplo Archivo por Localidades

- **Archivo Estratificación:**
  Este archivo concentra información general de los datos. Permite relacionar los archivos Zonas y Localidades con su correspondiente etiquetaje. Facilita la anáilitica, vizualización y procesamiento de los mismos. Para tal motivo se agregaron las columnas: **ID**, **Zona**, **Longitud**, **Latitud** (ver en figura 6). El archivo estructurado puede ser visualizado en el siguiente enlace: goo.gl/ZCUnHA

ID	Nombre Localidad	Total de Viviendas	Str 0 (%)	Str 1 (%)	Str 2 (%)	Str 3 (%)	Str 4 (%)	Str 5 (%)	Str 6 (%)	Longitud	Latitud	Zona
1	Antonio Nariño	27.774,00	1,79	0	4,49	93,72	0	0	0	-74.1	4.59	ASE 5
2	Barrios Unidos	57.196,00	0,16	0	0	54,6	41,89	3,35	0	-74.08	4.68	ASE 3
3	Bosa	133.097,00	4,51	4,83	87,18	3,48	0	0	0	-74.19	4.62	ASE 6
4	Chapinero	55.919,00	0,21	3,38	9,41	6,3	33,18	10,45	37,07	-74.06	4.65	ASE 3
5	Ciudad Bolívar	152.266,00	1,21	60,02	34,56	4,22	0	0	0	-74.16	4.58	ASE 4
6	Engativá	232.205,00	1,04	0,74	23,58	70,63	4	0	0	-74.11	4.70	ASE 2
7	Fontibón	116.233,00	1,94	0	18,76	44,66	33,75	0,88	0	-74.14	4.68	ASE 2
8	Kennedy	269.028,00	0,75	0,58	49,37	46,8	2,5	0	0	-74.15	4.64	ASE 6
9	La Candelaria	7.857,00	0,71	0,45	53,3	45,54	0	0	0	-74.07	4.59	ASE 3
10	Los Mártires	27.497,00	0,18	0	8,79	83,29	7,74	0	0	-74.09	4.60	ASE 3
11	Puente Aranda	70.682,00	1,05	0	0,32	98,62	0	0	0	-74.11	4.61	ASE 4
12	Rafael Uribe	104.433,00	1,12	8,53	47,72	42,63	0	0	0	-74.07	4.66	ASE 5
13	San Cristóbal	112.721,00	0,39	7,74	77,4	14,47	0	0	0,01	-74.08	4.56	ASE 5
14	Santa Fe	36.163,00	1,51	7,49	55,47	25,13	9,39	0,49	0,51	-74.07	4.61	ASE 3
15	Suba	288.568,00	1,61	0,21	30,56	33,59	18,78	13,71	1,54	-74.08	4.75	ASE 1
16	Sumapaz	1.743,00	0	54,91	28,06	9,7	3,61	1,61	2,12	-74.1	4.61	ASE 4
17	Teusaquillo	57.972,00	0,16	0	0	14,12	80,71	5,02	0	-74.08	4.64	ASE 3
18	Tunjuelito	49.168,00	0,49	0	56,48	43,04	0	0	0	-74.13	4.58	ASE 4
19	Usaquen									-74.03	4.70	ASE 1
20	Usme									-74.12	4.47	ASE 5

Figura 5: Ejemplo archivo Estratificación

- **Caracterización de residuos sólidos:**

  También se procedió a hacer una re-estructuración de la información entregada en cuanto a la caracterización de los residuos solidos residenciales, generados en la ciudad de Bogotá D.C. Estos archivos tienen una media ponderada de componentes físicos de los R.S dimiciliarios, definidos por porcentaje y estrato socioeconómico en Bogotá. El archivo estructurado puede ser visualizado en el siguiente enlace: goo.gl/usuiwV

ID	CATEGORIA	ID SUBCATEGORIA	SUBCATEGORIA	Estrato 1	Estrato 2	Estrato 3	Estrato 4	Estrato 5	Estrato 6	GLOBAL
1	Alimentos	1	Alimentos preparados	6.33	8.68	9.56	7.82	4.96	7.85	8.56
1	Alimentos	2	Alimentos no preparados	55.09	53.75	49.88	52.76	48.83	47.46	52.00
2	Residuos de jardineria	1	Residuos de jardineria	0.53	0.43	0.81	1.82	1.57	5.15	0.87
3	Residuos papel y carton	1	Residuos papel y carton	4.97	4.48	8.16	10.12	17.93	8.65	7.1
4	Residuos plastico	1	Polietileno	6.86	6.56	6.02	5.48	5.34	5.60	6.20
4	Residuos plastico	2	Policarbonato	0.17	0.03	0.04	0.02	0.01	0.03	0.04
4	Residuos plastico	3	Poliestileno rigido	0.29	0.31	0.37	0.31	0.45	0.44	0.34
4	Residuos plastico	4	Policloruro de vinilo	0.09	0.03	0.04	0.02	0.00	0.03	0.04
4	Residuos plastico	5	Pet transparente	1.21	0.96	1.76	1.36	1.22	1.06	1.33
4	Residuos plastico	6	Pet ambar	0.09	0.09	0.07	0.18	0.05	0.06	0.09
4	Residuos plastico	7	Pet verde	0.03	0.08	0.06	0.06	0.23	0.18	0.07
4	Residuos plastico	8	Polipropileno rigido	0.26	0.31	0.28	0.33	0.43	0.65	0.31
4	Residuos plastico	9	Polietileno de alta dens	0.93	0.71	0.65	0.67	0.90	0.38	0.70
4	Residuos plastico	10	Polipropileno flexible	1.17	0.81	0.95	0.69	0.56	0.85	0.87
4	Residuos plastico	11	Icopor	0.23	0.25	0.29	0.51	0.35	0.45	0.30
4	Residuos plastico	12	Otros	0.24	0.21	0.12	0.18	0.02	0.01	0.16
5	Residuos caucho y cuero	1	Residuos caucho y cuero	1.33	0.42	0.24	0.51	0.13	0.12	0.42

Figura 6: Ejemplo archivo Caracterización de datos

3. Exploración y análisis de datos

El análisis de datos, así como también la implementación de modelos de predicción fueron desarrollados en **Python**, utilizando diferentes librerías para en análisis de datos (ver anexo A). Además para la visualización y despliegue de los datos se utilizó la tecnología de ***Notebooks*** que permiten la codificación de archivos digitales interactivos que envuelven tanto el código, descripción de los modelos y la visualización de los resultados obtenidos.

En la segunda fase de la consultoría técnico-científico se desarrolló un archivo un **Notebook** que contiene la lectura,descripción y visualización de los datos. la información obtenida es desplegada de forma dinámica, lo cual facilita el análisis y manipulación de expertos ambientales en cuanto a la información contenida de los datos. la información puede ser analizada por zonas y localidades de manera independiente y también correlacionada entre ellas. A continuación detallamos los resultados que se pueden observar en la implementación desarrollada. El notebook desarrollado puede ser consultado en la siguiente localización: goo.gl/VKsFSk

3.0.1. Análisis por zona

Inicialmente se realizó una exploración de los datos para determinar relaciones y patrones de las cifras almacenadas. La mejor manera para visualizar y entender los datos es transformar los valores almacenados a series de tiempo. En la figura 7 se muestran las respectivas series de tiempo para cada una de las zonas. En este apartado se puede observar cuales son las zonas en Bogotá que más generan residuos y cuales generan menos. A su vez, se puede identificar algunos picos y rangos periódicos pequeños. Por ejemplo: Las Zonas que generan más residuos son la 1 y la 6, y las zonas que generan menos residuos son la 3 y 4. Además, se presentan algunas caídas en los últimos meses de cada año.

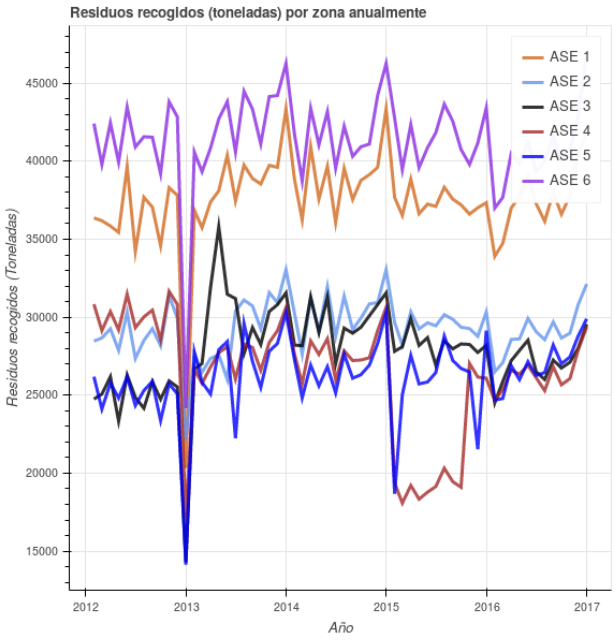

Figura 7: Residuos recogidos por zona anualmente

3.0.2. Análisis por localidad

Además, se decidió analizar las series de tiempo de las localidades por cada uno de sus meses. En la figura 8 se muestra el comportamiento de la generación de residuos en los meses de cada año para la localidad seleccionada. Y en la figura 9 se muestra la relación de la población con la generación de residuos a través de los años registrados. Es fácil observar que la población de cada localidad crece de manera suave sin cambios bruscos y que la localidad de Suba presenta un crecimiento en la generación de residuos. Para este paso se presenta una visualización alternativa en función del año y mes seleccionado. Ver figura: 11

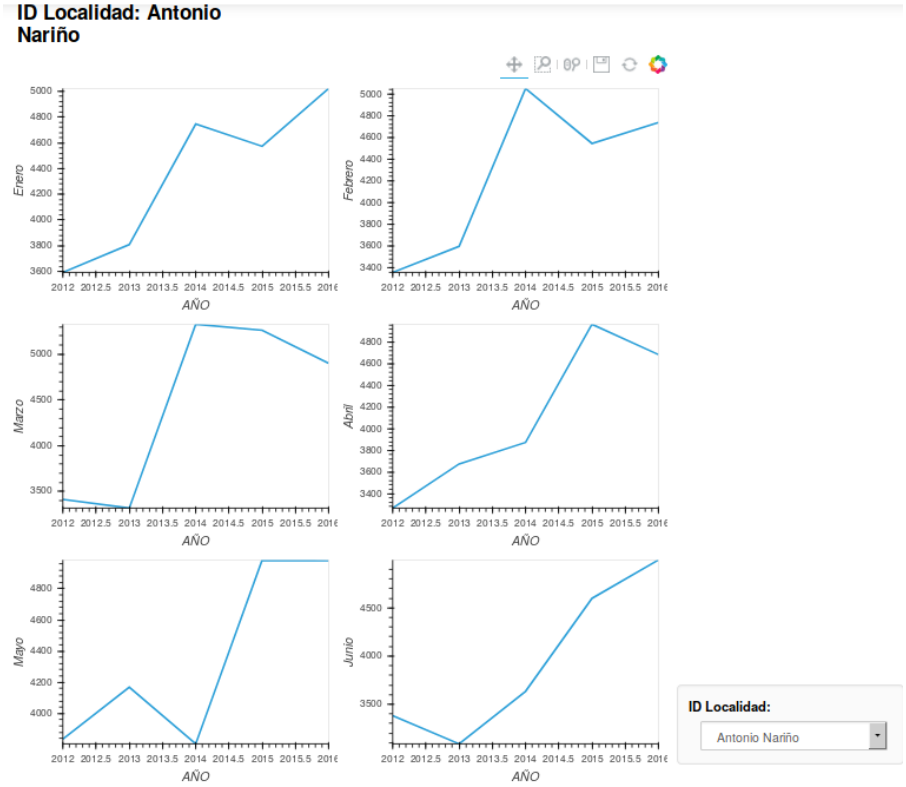

Figura 8: Residuos recogidos por localidad anualmente especificado por meses

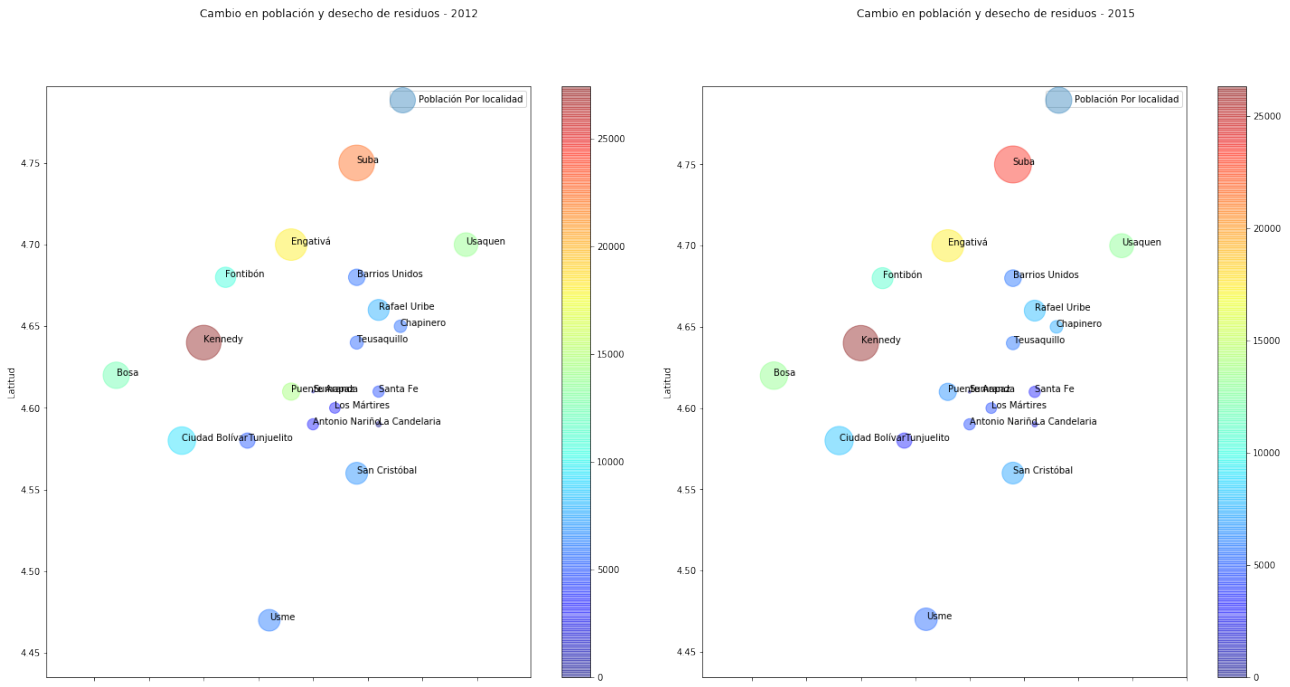

Figura 9: Crecimiento y decrecimiento en la generación de residuos y población anualmente (mes Febrero)

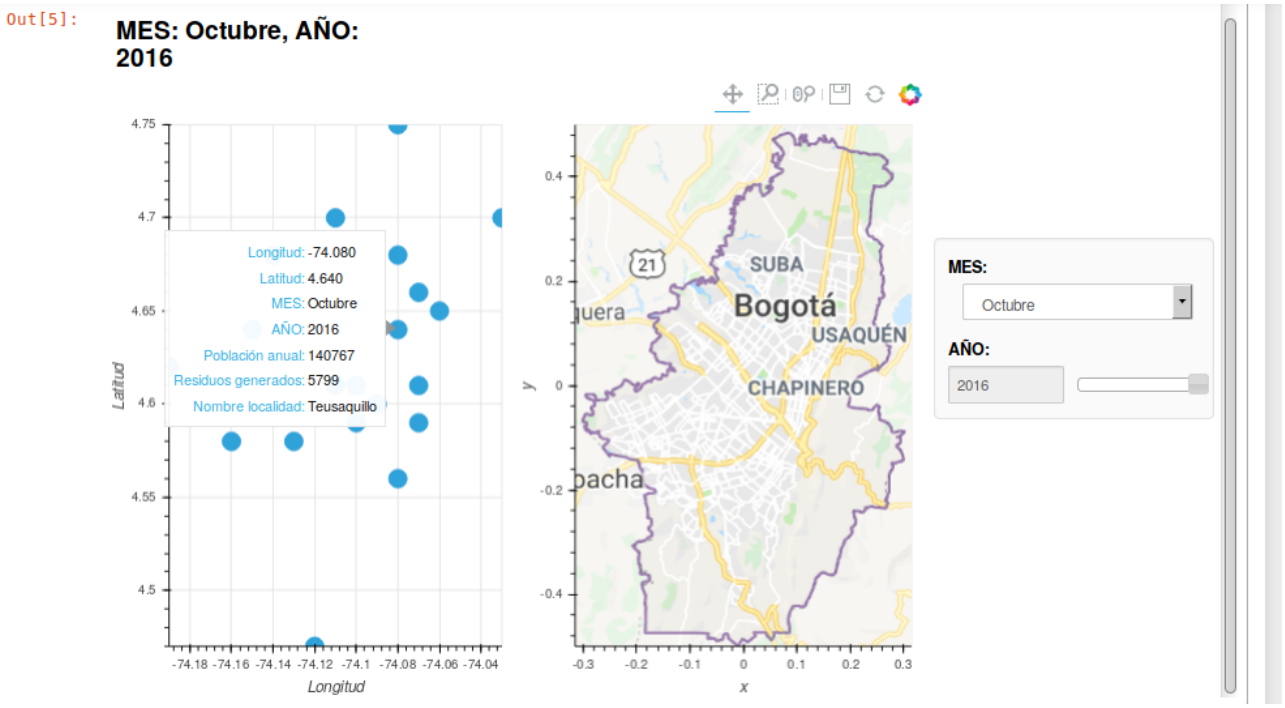

Figura 10: Visualización alternativa en función del año y mes

3.0.3. Análisis conjunto por zona y localidad

En la figura 11 se muestra el porcentaje de participación que tiene cada localidad en cada zona correspondiente por cada mes en todos los años registrados.

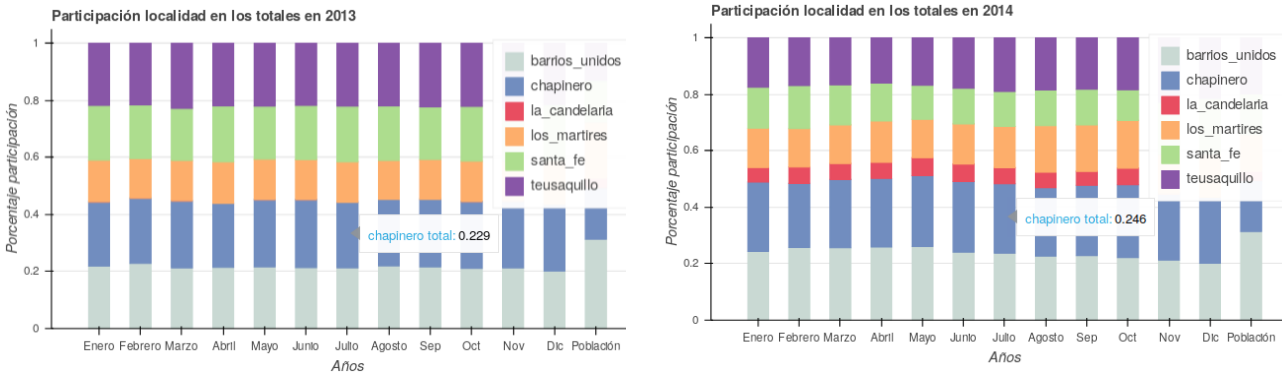

Figura 11: Participación portcentual de las localidad en la generación de residuos en su respectiva localidad

La visualización y analisis generados pueden ser ejecutados por el usiario y modificados en una plataforma de Notebooks. Esto le permitirá a los expertos en el proyecto analizar diferentes variables de interes y tener una primera perspectíva visual de los datos que trataran en su problema.

4. Predicción de residuos: Implementación de algoritmos

En una tercera fase de la consultoría, se utilizaron los datos proporcionados por la USTA y re-estructurados para el desarrollo e implementación de modelos de predicción utilizando herramientas de aprendizaje de máquina. El archivo que implementa los diferentes modelos de predicción pueden ser consultados en el siguiente enlace: goo.gl/oxQnTP. Teniendo en cuenta el tamaño limitado de sus datos, así como su relación temporal se deicidió utilizar tres diferentes modelos:

- Árboles de decisión (DT)
- Máquinas de soporte vectoral (SVM)
- Redes neuronales recurrentes (LSTM)

Para cuantificar el error asociado a cada una de las predicciones realizadas por este método, en este trabajo se utilizó la raíz del error medio cuadrático, definido como:

$$e = \sqrt{\frac{1}{M}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$$

donde $\hat{y}_i$ es el valor predicho por el método implementado y y_i es el valor real. Este error mide las diferencias locales y nos da una estimación en terminos de la variable cuantificada, es decir, cantidad de residuos.

4.0.1. Predicción con Árboles de decisión

El árbol de decisión, en aprendizaje de máquina, es un algoritmo no paramétrico que permite modelar fronteras de separación de datos basados en reglas de decisión aprendidos sobre las características de entrada del modelo. Según se requiera la sensibilidad del modelo, se definen el profundidad del árbol aprendido, que contiene un conjunto de umbrales de decisión que separan los datos contenidos en cada característica. Por ejemplo, en nuestro problema particular la característica de entrada son los residuos por año. Si se define un árbol de decisión de dos, el algoritmo de aprendizaje intentará separar los residuos en dos subgrupos y solo existirá un valor de separación. A medida que se define un mayor número de profundidad, la separación se vuelve más fina y se pueden separar los datos de una mejor forma. Sin embargo, Un valor muy alto en el árbol de profundidad hace líneas de separación sobre ajuste lo que limitan la generalidad del modelo para nuevos valores. Esto quiere decir que el modelo aprendió incluso datos ruidosos de la entrada y no podrá predecir valores coherentes para valores con cierta varianza con respecto a los datos de entrenamiento.

En términos técnicos, los árboles de decisión recursivamente particionan el espacio de características $\left\{\mathbf{x_j}\right\}_{j_{1\ldots m}}$ que en nuestro caso particular corresponde al porcentaje de residuos. Cada partición candidata es definida como $\theta = (j, \tau_m)$ donde j es el valor en términos de residuos y τ_m es el umbral de predicción. Cada nodo del árbol Q es por lo tanto definido como la partición $Q_{left}(\theta)$ and $Q_{right}(\theta)$. La implementación utilizó una versión optimizada del algoritmo CART. En el notebook desarrollado (ver en goo.gl/oxQnTP) se illustra el uso arbol de decisión para la zona 4 utilizando árboles de profundidad de 3 y 6, como se ilustra en las siguientes figuras:

La principal ventaja de este método es la raídez en el cálculo de del arbol representativo y la posibilida de visualizar y entender los cortes realizados por el modelo. Sin embargo, si los datos tienen una alta variabilidad, este método puede ser restringido en dar unos apropiados resultados.

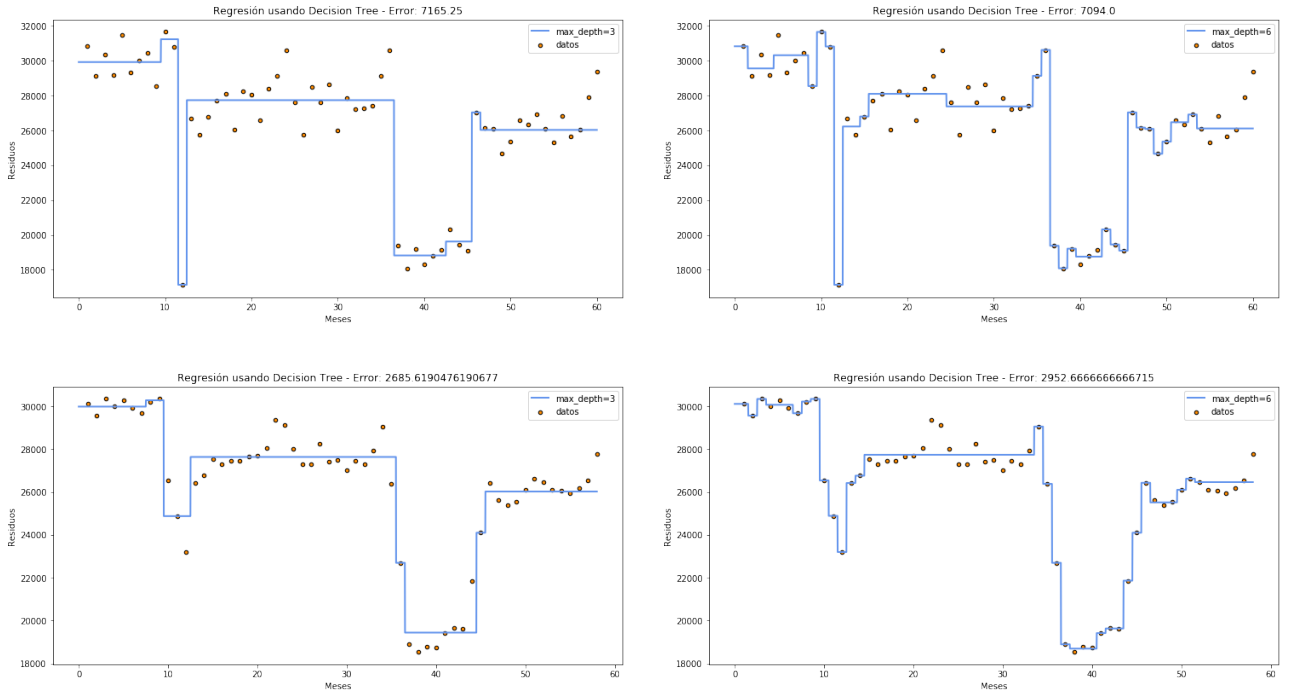

Figura 12: Comparación del rendimiento de los árboles de decisión usando profundidad de nivel 3 y 6 vs usando una ventana deslizante

4.0.2. Predicción con Máquinas de Soporte Vectorial (SVM)

Un segundo método implementado como modelo de predicción fueron las máquinas de soporte vectorial (comúnmente conocidas por su nombre en inglés: *support vector machine*). Este método está basado en funciones de separación local de los datos, denominados: "kernels". Estas funciones interponen un plano de separación entre un conjunto de datos cercanos (localmente) y su flexibilidad de separación depende del tipo de función. Los kernels más utilizados en la literatura son los del tipo lineal, polinomial y las funciones de base radial. Para el ajuste de estas funciones, sobre un conjunto de datos se seleccionan aleatoriamente un conjunto de datos denominado *vectores de soporte*. Una vez seleccionado los vectores de soporte se intenta maximizar la distancia entre estos vectores, denominado margen de clasificación. En el notebook implementado (ver en goo.gl/oxQnTP) se calculo un modelo de regresión utilizando como kernel una función de base radial. El comportamiento se ilustra en las siguientes figuras. Como se puede observar este método se ajusta apropiadamente a los datos seleccionados y logra curvas coherentes de regresión pese a que los datos de entrenamiento son muy limitados.

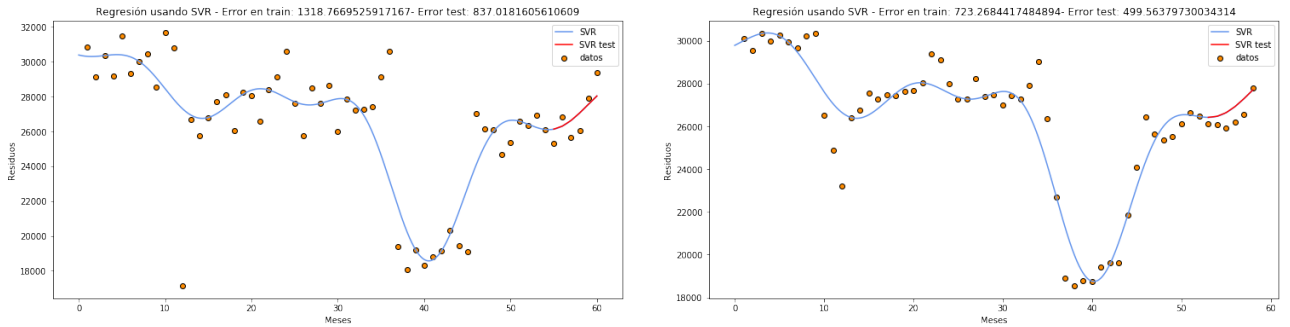

Figura 13: Comparación del rendimiento de las máquinas de soporte vectorial entre modelos con ventana deslizante y sin ventana deslizante

La principal ventaja de las máquinas de soporte vectorial es su apropiado ajuste a los datos a pesar de la naturaleza variable de estos o frente a problemas que contienen pocos datos de entrenamiento. En la literatura estos métodos de regresión son muy utilizados debido a los resultados alcanzados. Por otra parte, existe la posibilidad de usar diferentes kernels que pueden interpretar mejor los datos de entrenamiento para efectuar una mejor predicción.

4.0.3. Predicción con Redes Neuronales Recurrentes (LSTM)

Teniendo en cuenta el principal objetivo del proyecto y la consultoría, con respecto a la caracterización de los datos utilizando redes neuronales artificiales, se implementaron modelos recurrentes de redes neuronales para la predicción de los datos suministrados por la USTA. Existe una amplia variedad de redes neuronales, con diferentes propósitos y diversas arquitecturas según la tarea de predicción establecida. En el caso, de series temporales las redes neuronales recurrentes han mostrado resultados favorables y el diseño de su arquitectura resulta útil para explorar correlaciones temporales entre los datos. En este trabajo se decidió por usar estas redes debido a su aplicación directa como predicción de series temporales. Una de las principales ventajas de estas redes (comúnmente conocidas como LSTM por su nombre en inglés : *Long Short-Term Memory network*), es la capacidad para ajustar comportamientos no lineales de los datos y mantener estados de memoria y olvido que tienen en cuenta información temporal pasada. En este caso las neuronas son denominadas bloques de memoria, conectados a través de diferentes capas. Cada bloque tiene tres compuertas, definidas como:

- Compuerta de olvido: decide la información relevante que debe mantenerse para predecir nuevos valores
- Compuerta de entrada: en esta compuerta se decide que valores serán actualizados para actualizar la mamoria de la red
- Compuerta de salida: Condicionalmente decide la salida que debe enviarse a un siguiente bloque de memoria.

Los algoritmos basados en redes neuronales profundas utilizan un método de gradiente descendente para aprender los pesos entre capas y conexiones de las neuronas, en este caso los bloques de memoria. Debido a que su uso es normalmente en conjunto con grandes volúmenes de datos, existen nuevos algoritmos de aprendizaje que intentan aprender estos parámetros de forma recursiva sobre pequeños conjuntos de datos. Estos algoritmos requieren entonces la definición de nuevos parámetros como: *batch* y el *epochs* and. El *batch* se refiere a un subconjunto de datos seleccionado para encontrar el gradiente de optimización hacia donde convergerá el algoritmo. Por otra parte el *epochs* se refiere al número de iteraciones que se ejecuta el gradiente descendiente en la función objetivo para minimizar el problema y encontrar la solución[1]. para mejorar el comportamiento de estos métodos, se implementó previamente una ventana deslizante que permite filtrar temporalmente los datos para obtener un suavizado secuencial según sus vecinos locales. En la siguiente Figura se ilustran los resultados obtenidos.

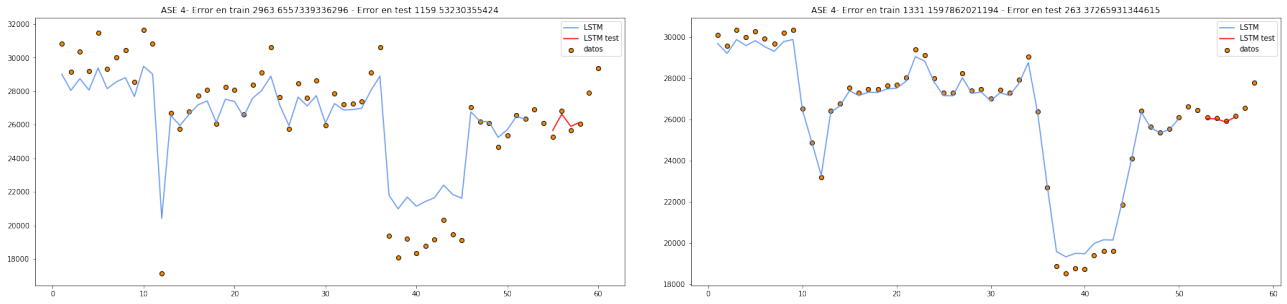

Figura 14: Comparación del rendimiento de una Red LSTM básica entre modelos con ventana deslizante y sin ventana deslizante en pre-procesado

[1]Los algoritmos de *deep learning* tienen una función objetivo no-convexa y por lo tanto no existe certeza de encontrar la mejor solución de forma analítica

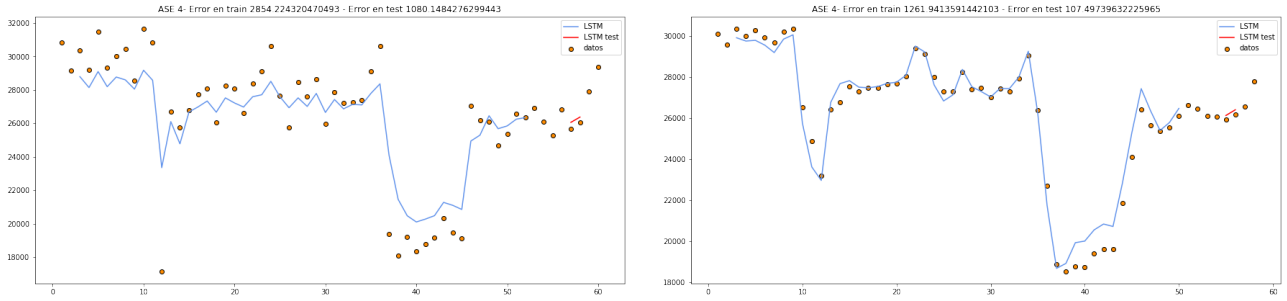

Figura 15: Comparación del rendimiento de una Red LSTM con ventana entre modelos con ventana deslizante y sin ventana deslizante en pre-procesado

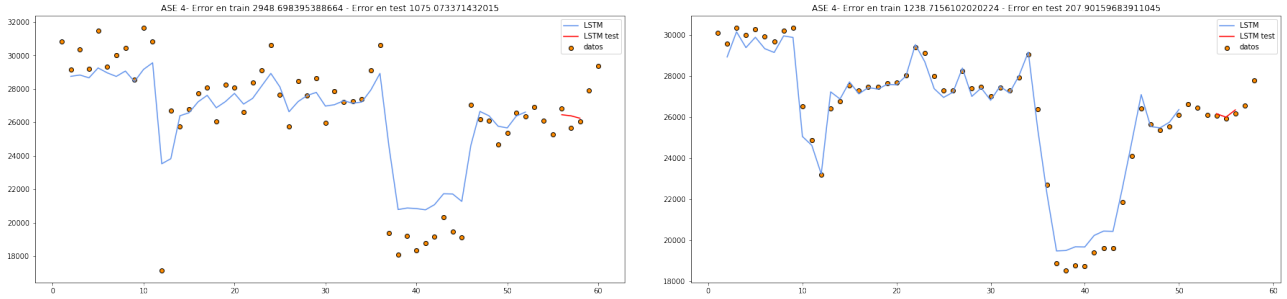

Figura 16: Comparación del rendimiento de una Red LSTM con pasos de tiempo entre modelos con ventana deslizante y sin ventana deslizante en pre-procesado

Teniendo en cuenta que estas redes LSTM fueron diseñadas para problemas de predicción temporal, existen parámetros propios que pueden ser adaptados para un mejor comportamiento de la predicción. Por ejemplo, en señales escalonadas que tienen periodos propios de frecuencia, se puede definir ajustar el tiempo de paso. En un ejemplo concreto, si para nuestra señal es por meses y tenemos varios años registrados temporales, se esperaría que un año sea el periodo ideal de la señal y la señal tiene un comportamiento cada 12 meses, entonces podría ajustar el tiempo de paso en 12. A continuación se muestran los resultados ajustando el tiempo de paso.

La principal ventaja de los métodos basados en redes neuronales es su bien conocido resultados frente a diversos problemas y en diferentes áreas de aplicación. Debido a los sobresaliente resultados, en la comunidad científica estos métodos han llamado la atención y han avanzado en los últimos años de forma intensa. Las redes neuronales recurrentes utilizadas en este trabajo logran un descripción de los datos coherente, pero sin embargo, en algunos casos es muy sensible a ajustar los datos. Una de las principales dificultades para la

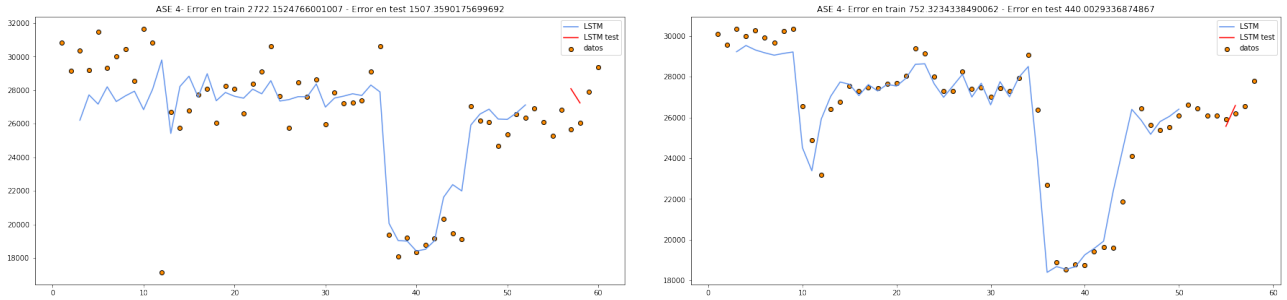

Figura 17: Comparación del rendimiento de una Red LSTM memoria por lotes entre modelos con ventana deslizante y sin ventana deslizante en pre-procesado

predicción con redes neuronales es el requerimiento de una basta cantidad de datos para lograr predicciones robustas y apropiadas con respecto a los datos.

5. Caracterización de residuos predichos

La última fase de la consultoría consta de las predicciones resultantes realizadas con los modelos entrenados y la caracterización de los residuos predichos por estrato y categoría de residuo en cada localidad o zona. Inicialmente, en este entorno interactivo se debe configurar el tipo de modelo y los parámetros deseados para iniciar el proceso de entrenamiento ver figura 18. Cabe mencionar que diferentes parámetros llevan a diferentes resultados. La opción de configurar los modelos permite analizar y estudiar el impacto que tiene los diferentes valores de los parámetros sobre el problema de predicción tratado. A continuación se describen este proceso de configuración inicial con más detalle:

- Se debe escoger sobre qué tipo de datos se hará la predicción, si sobre las zonas en general o sobre las localidades.
- Seguidamente el tamaño de la ventana de suavizado. Es decir, promediar o suavizar la serie de valores por medio de una ventana deslizante donde el tamaño será el parámetro a escoger.
- Escoger el modelo a entrenar. Disponibles SVR y LSTM, cada modelo tiene diferentes parámetros.
- Si se escoge SVR se deben configurar las siguientes variables:
  - Número de predicciones a realizar. Es decir, a partir del último mes registrado, cuántas predicciones se desea realizar.
  - Se debe configurar el número de entrenamientos, o el número de iteraciones que el algoritmo debe utilizar para entrenarse.
  - Finalmente se fija el número de validaciones cruzadas o el número de subconjuntos en el que se dividen los datos para validar el proceso de entrenamiento.
- Si el modelo escogido son las LSTM, se debe configurar:
  - Las capas que tendrá la red neuronal recurrente, esto permite aprender detalles más finos o ajustarse más a los datos de entrenamiento.
  - Número de entrenamientos, descrito en el modelo anterior.
  - Internamente podemos decirle a nuestro modelo, cuantos valores (meses), tenga en cuenta para realizar la predicción del siguiente valor. Esto se refiere al parámetro de pasos de tiempo a configurarse.

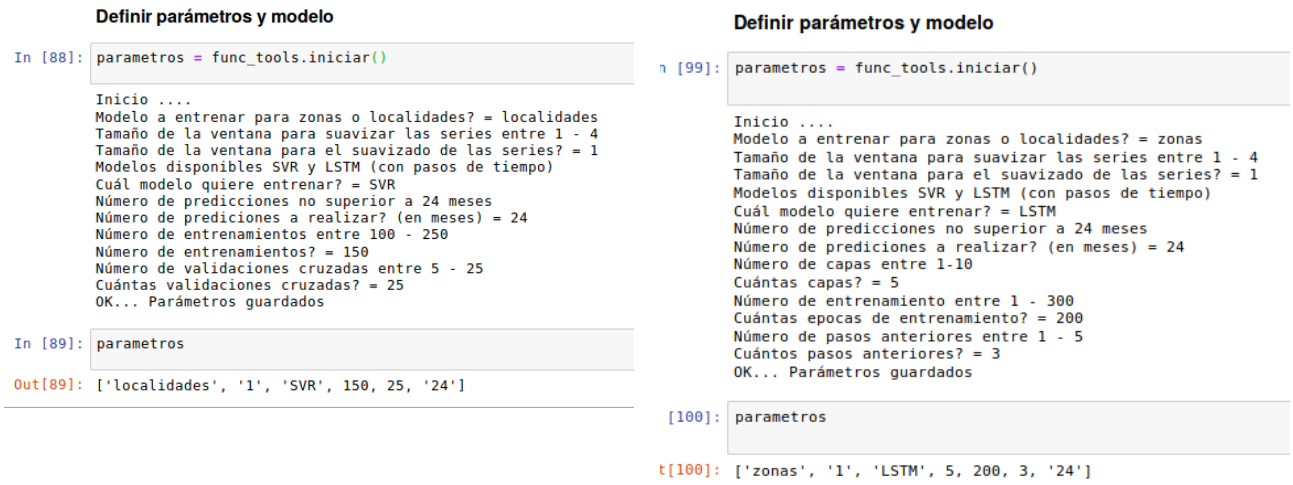

Figura 18: Ejemplo de parametrización los modelos para predicción y caracterización de residuos por localidad y zonas

Una vez entrenados los modelos, se realizan las predicciones de los residuos que se generarán en los próximos meses. En las figuras 19 y 20 se muestra las predicciones realizadas tanto para el conjunto de entrenamiento como para las predicciones realizadas por los modelos en el futuro en las localidades de Chapinero, San Cristóbal y en las zonas ASE 3 y ASE 6, utilizando SVR y LSTM respectivamente. Podemos ver que la predicción realizada para la localidad de Chapinero es mucho más suave y coherente con los patrones en el conjunto de datos. Pero para la localidad de San Cristóbal el modelo se sobre ajusta a los datos y las predicciones realizadas no parecen seguir la misma tendencia. Esto principalmente se debe a la naturaleza del SVR.

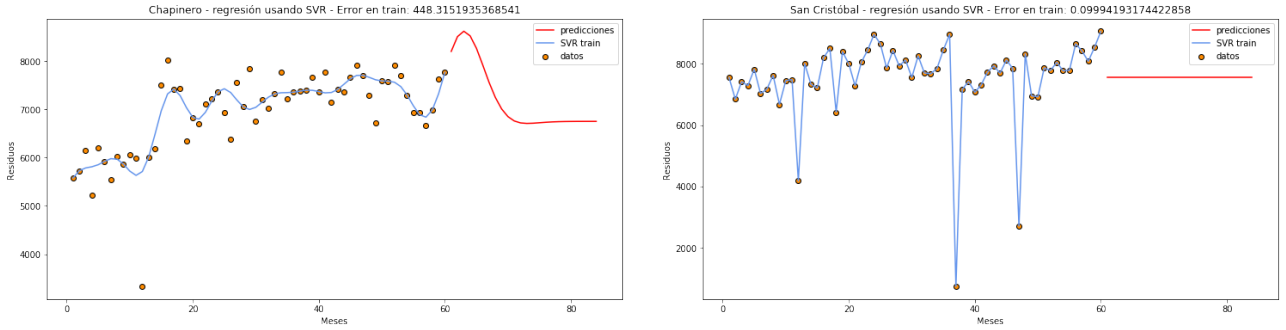

Figura 19: Gráficas de entrenamientos y predicciones para el modelo SVR en las localidades de Chapinero y San Cristóbal

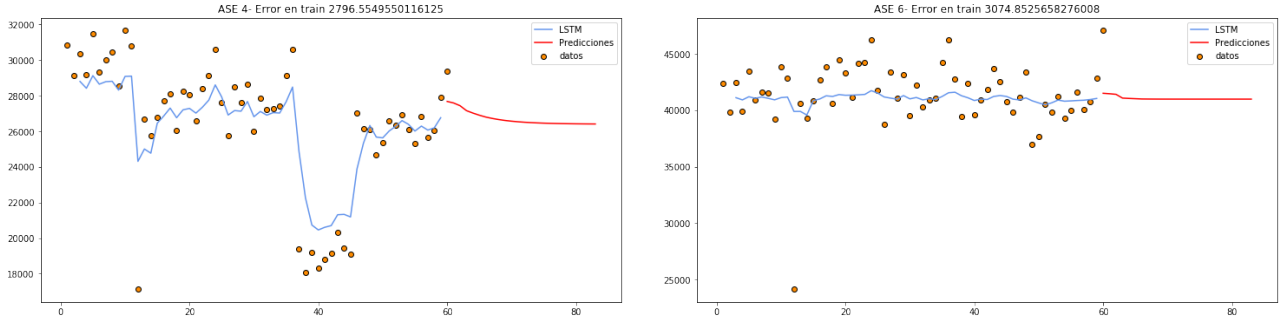

Figura 20: Gráficas de entrenamientos y predicciones para el modelo LSTM en las zonas ASE 4, ASE 6

En cuanto a las predicciones realizadas en las LSTM, se puede mencionar que tienden a seguir un promedio, esto es debido principalmente a que las predicciones se realizan de manera recursiva, es decir se realizan predicciones en base a predicciones anteriores. Por lo general estos modelos necesitan grandes cantidades de datos para mantener un histórico de patrones más fino, y alimentar la red con observaciones reales que se vayan obteniendo.

Ya con estas predicciones se es posible realizar una caracterización por estratos y por residuos en cada zona o localidad con base a la tabla de estratificación y caracterización. Un aspecto importante para mencionar, es que estas tablas se consideraron como recientes y corresponden a datos actualizados del último año.

En las figuras 22 y 21, se muestra las predicciones de residuos en toneladas por estrato y mes de predicción para la localidad Antonio Nariño y la zona ASE 3. Estos resultados permiten cuantificar la cantidad de residuos generados por los diferentes estratos presentes.

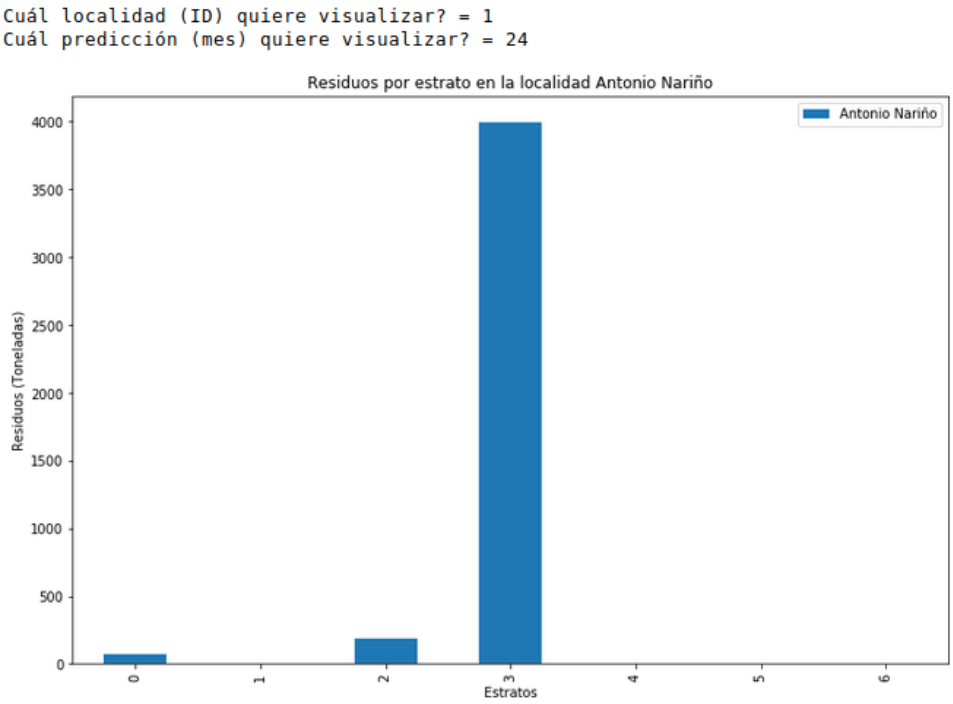

Figura 21: Caracterización de las predicciones por estratos en la localidad Antonio Nariño usando el modelo SVR

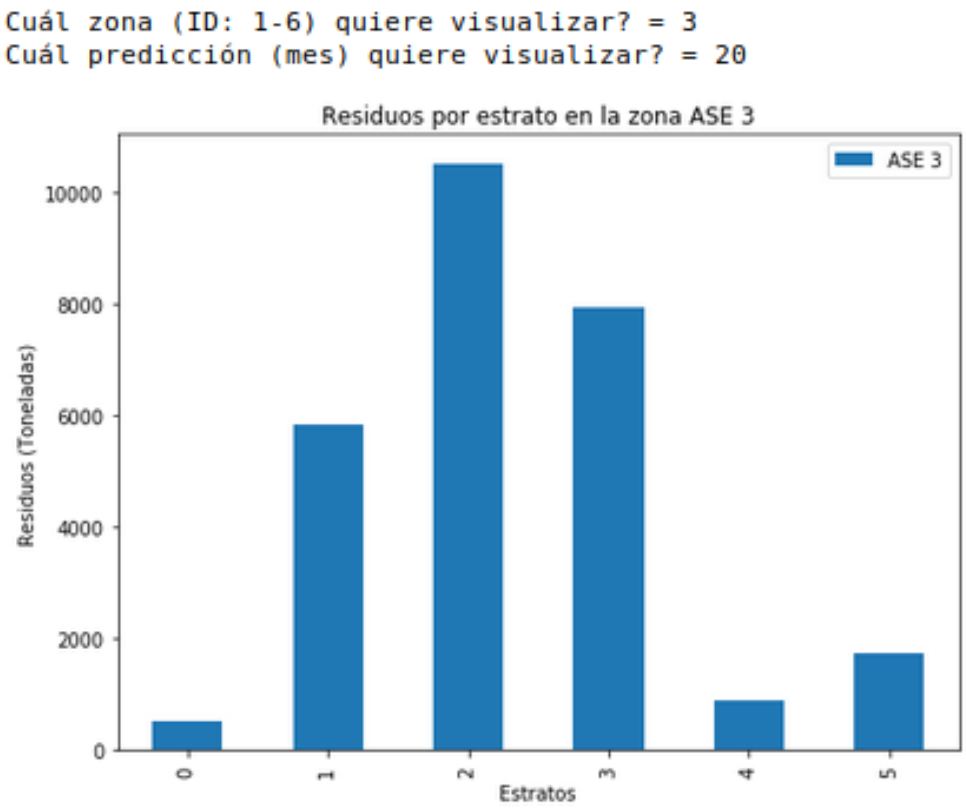

Figura 22: Caracterización de las predicciones por estratos en la zona ASE 3 usando el modelo LSTM

Yendo más allá de los anteriores resultados, podemos caracterizar por cada estrato cuales residuos (categoría) y en qué cantidades se generarán. Ver figuras 23 y 24.

```
Cuál zona (1-6) quiere visualizar? = 3
Cuál predicción (mes) quiere visualizar? = 20
Cuál es el estrato (1-6) que quiere visualizar? = 3
```

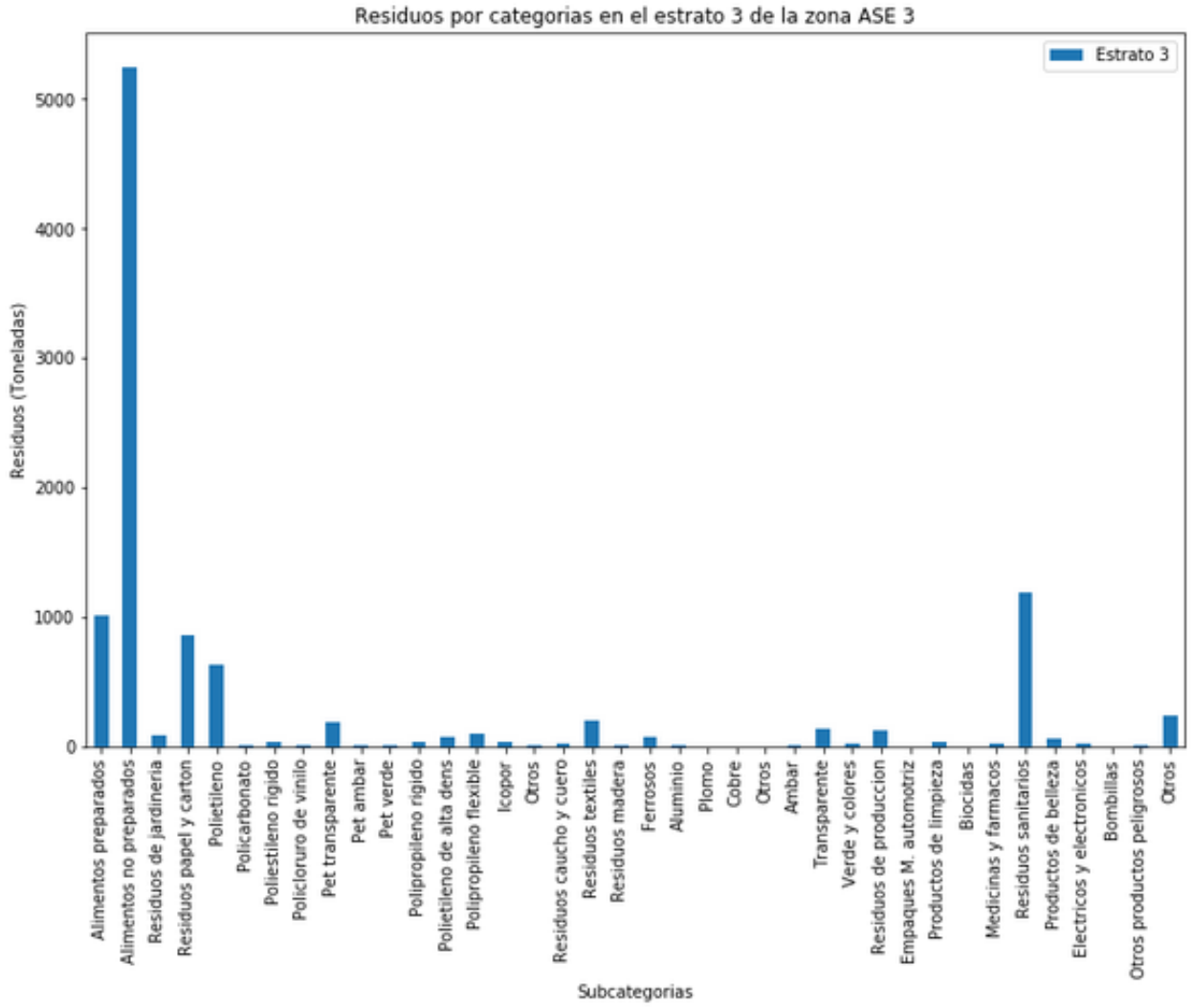

Figura 23: Caracterización de las predicciones por categorías en la zona ASE 3 - estrato 3 usando el modelo LSTM

```
Cuál localidad (ID) quiere visualizar? = 1
Cuál predicción (mes) quiere visualizar? = 24
Cuál es el estrato (1-6) que quiere visualizar? = 3
```

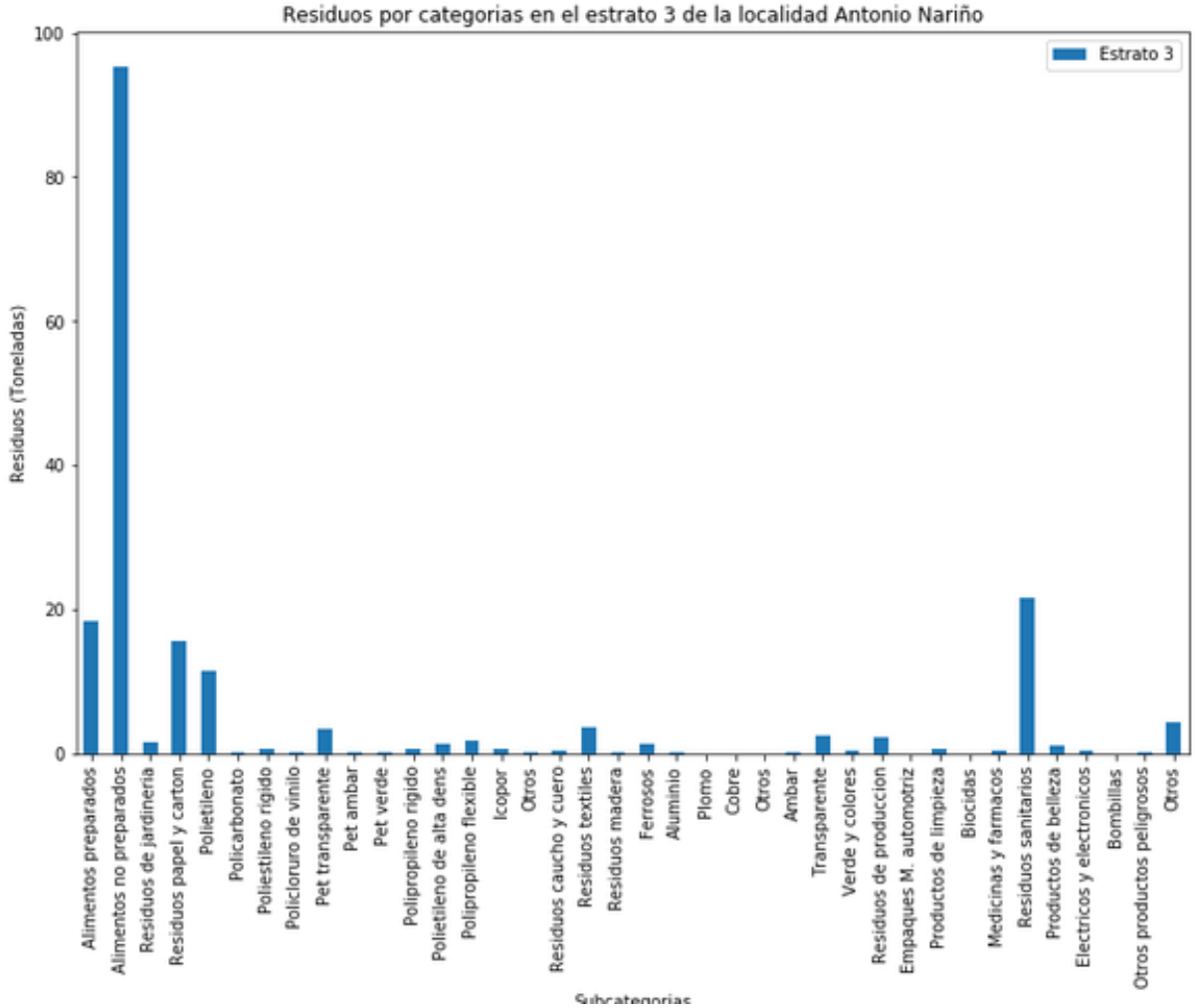

Figura 24: Caracterización de las predicciones por categorías en la localidad Antonio Nariño - estrato 3 usando el modelo SVR

6. Conclusiones

Durante la consultoría técnico-científica desarrollada por la escuela de ingeniería de sistemas, se trabajó específicamente en tres ejes principales 1) estructuración y exploración de los datos suministrados, 2) implementación de modelos de residuos y 3) caracterización de los residuos.

En cuanto a la estructuración de los datos, durante el desarrollo del proyecto se modificó la organización de los datos suministrados para lograr una lectura automatizada de los mismos y poder evidenciar la calidad de los mismos, en cuanto a elementos faltantes. Estos archivos tienen extensión *cvs* y pueden ser consultados en la carpeta del proyecto. Los elementos (datos) faltantes fueron estimados a partir de observaciones con las mismas características, por ejemplo, valores en años precedentes o valores vecinos en la columna de datos. Una vez organizados y estructurado los datos, se pasó al desarrollo de una herramienta que permitiera un análisis visual de los mismos. Estos análisis pueden ser modificados por el usuario, o calculados a partir de nuevos datos estructurados en los archivos *cvs*. En las reuniones definidas durante el proyecto se brindó asesoría de su funcionamiento y manipulación, para que los expertos en el problema ambiental pudieran manipularlos de forma efectiva.

En una segunda etapa, se implementaron tres diferentes métodos de predicción de datos. Teniendo en cuenta el número limitado de éstos, se inició por una exploración de los datos utilizando árboles de decisión, con diferentes profundidades. También se logró la implementación de máquinas de soporte vectorial para el cálculo de modelos de regresión basados en funciones radiales locales, calculadas a partir de puntos (vectores de soporte) en un vecindario específico. Finalmente se implementaron métodos basados en redes neuronales para la estimación de trayectorias de puntos. Estos métodos son conocidos como redes LSTM. En la implementación de las redes LSTM se tuvo en cuenta diferentes configuraciones, filtrados temporales y cálculo de periodos anuales de los desechos. Como herramienta de cálculo del error se implementó la raíz del error medio cuadrático. En los ejemplos ilustrados, se visualiza un apropiado comportamiento de las máquinas de soporte vectorial y las redes LSTM. Por ejemplo, se evidencia un error promedio de: 898 toneladas para las máquinas de soporte vectorial en la zona 4, mientras que para el LSTM se evidencia un error de 1444 toneladas en la misma zona. Teniendo en cuenta esta evidencia cuantitativa se puede concluir que en términos de tendencia local de los puntos, y asumiendo una alta confiabilidad en los valores registrados, que el método de mejor comportamiento es: SVR. También se aclara, que se deben hacer más experimentos y la observación de expertos ambientales para la definición de la herramienta de predicción.

En cuanto a la caracterización de residuos se implementó un *notebook* adicional que permite el uso de tablas de conocimiento ambiental, en cuanto a estadísticas y distribuciones generales por porcentaje de residuos. Estas tablas pueden ser utilizadas junto con los modelos de predicción para discriminar los porcentajes y tipos de residuos que se generaran según las tendencias predicas. Esta herramienta permite al usuario introducir diferentes valores para la predicción, como los años en los que desea hacer la predicción. También le permite al usuario seleccionar el método de regresión y dar un mayor panorama de análisis para los expertos ambientales.

Referencias

Apéndices

A. Anexos A: Requerimientos técnicos y dependencia de paquetes

El desarrollo de las soluciones presentadas fueron desarrolladas en un entorno de trabajo específico. Por lo tanto, requiere de un conjunto de herramientas tecnologícas para su correcto funcionamiento. A continuación se listan las herraminetas necesarias:

- Python v3.6 (Lenguaje de programación)
- Anaconda v3.6 (Plataforma para ciencia de datos y Machine Learning)
- Librería Nodejs v8.11.3 sobre Anaconda 3.6 (Librería adicional)
- Librería bokeh v0.12.16 sobre Anaconda 3.6 (Librería adicional)
- Librería Holoviews v1.10.7 sobre Anaconda 3.6 (Librería adicional)
- keras v2.2.2
- tensorflow v1.9.0

Una vez instalados los requerimientos en su totalidad, el sistema de notebooks debería ejecutarse correctamente. Adicionalmente se presenta una guia rápida de instalación en sistemas operativos basados en GNU-Linux.

- **GNU-Linux**

```
$wget https://repo.anaconda.com/archive/Anaconda3-5.2.0-Linux-x86_64.sh
$bash Anaconda3-5.2.0-Linux-x86_64
$conda install nodejs
$conda install bokeh
$conda install holoviews
$conda install keras
```

Nota: El archivo **Anaconda3-5.2.0-Linux-x86-64** es entregado por si en el futuro no se encuentra disponible en internet.