Aplicación de Machine Learning al análisis de espacios verdes en la Ciudad Autónoma de Buenos Aires

Universidad Tecnológica Nacional

Cátedra de Ciencia de Datos - 2020

Reinke, Brian **De Stefano, Tomas** **Kopyto, Pedro**

Abstract: A lo largo del informe analizaremos los datos obtenidos de los espacios verdes en la Ciudad Autónoma de Buenos Aires (CABA), como así también su población y el presupuesto que destinó espacios verdes

1.Introducción y objetivos

El trabajo surge del cuestionamiento e inquietud de saber cómo es la distribución de espacios verdes en la Ciudad Autónoma de Buenos Aires. La cantidad de metros cuadrados por persona de espacios verdes es una problemática que atraviesan la mayoría de las ciudades cosmopolitas del mundo. La OMS (Organización Mundial de la Salud)[1] recomienda un mínimo de 9 metros cuadrados por persona de áreas verdes. Estos espacios provocan beneficios en la salud de las personas. Debido a la importancia de lo mencionado, buscaremos analizar los datos y la situación de la Ciudad.

2.Descripción de los datasets

Para realizar el análisis utilizamos 3 datasets de la Ciudad Autónoma de Buenos Aires.

El dataset principal será el de espacios verdes, que contiene la información geográfica y clasificación de las áreas dentro de la Ciudad. En él se pueden encontrar distintos espacios separados por barrios como así también por comunas.

	WKT	clasificac	patio_de_j	ubicacion	barrio	comuna	id_ev_pub	area	perimeter
0	MULTIPOLYGON (((-58.4453556017045 -34.57924873...	PLAZOLETA	NO	CONDE - MATIENZO, BENJAMIN, TTE. - FREIRE, RAM...	COLEGIALES	13.0	2	1658.266	0.0
1	MULTIPOLYGON (((-58.4448145611193 -34.57991038...	PLAZOLETA	NO	CONDE - MATIENZO, BENJAMIN, TTE. - FREIRE, RAM...	COLEGIALES	13.0	5	3.984	0.0
2	MULTIPOLYGON (((-58.4448074253007 -34.57987067...	PARQUE	NO	CONDE - MATIENZO, BENJAMIN, TTE. - FREIRE, RAM...	COLEGIALES	13.0	6	4686.060	0.0

Por otro lado, el dataset de la población. En él se encuentra la evolución de los habitantes de la Ciudad desde 2015 a 2020, discriminado por comuna, por sexo y rango etario, como así tambien la estimación población propia del Gobierno de la Ciudad de la población hasta el año 2025.

	Total	Comuna 1	Comuna 2	Comuna 3	Comuna 4	Comuna 5	Comuna 6	Comuna 7	Comuna 8	Comuna 9	Comuna 10
Rango Etario											
Total	1426582	123030	66915	90417	113113	85451	83865	112134	106508	81816	79581
0-4	108023	8902	3819	6405	10301	5797	5806	9568	11768	6465	5743
5-9	105680	7939	3568	6198	10183	5868	5545	9274	11335	6865	6192
10-14	99209	7592	3347	5858	10054	5384	4955	8764	10616	6658	5488
15-19	94542	7654	4005	5626	8996	5262	4770	8086	9369	5845	5426

Por último, el de presupuestos que contiene los gastos de los distintos órganos del Gobierno de la Ciudad Autónoma de Buenos Aires a lo largo de los años.

UNIDAD_EJECUTORA	DIRECCION GENERAL DE ESPACIOS VERDES	MINISTERIO AMBIENTE Y ESPACIO PUBLICO	AGENCIA AMBIENTAL	DIRECCION GENERAL DE LIMPIEZA	UPE ECOPARQUE	DIRECCION GENERAL DE CONTROL
AÑO						
2010	$1.466.680.260	$380.848.800</td> <td>$149.234.030	$1.275.274.000</td> <td>-</td> <td>$107.669.410	-	$431.220.150</td> <td>-</td> <td>-</td> <td>$46.323.530	$44.346.200</td> <td>$21.104.120

Además, para realizar un análisis de la valoración de los espacios verdes por parte de los habitantes utilizaremos la información de Google Maps sobre los mismos. Esto se logró mediante un scrap de datos donde obtuvimos la cantidad de reviews y la calificación de las distintas áreas.

Con esta información buscaremos analizar si existe una relación correlación entre las características y distribución de cada espacio verde (ya sea tipo de espacio, superficie, área, ubicación, etc.) y las valoraciones de los mismos (calificación o reviews).

* . Tomamos como supuesto que la cantidad de reviews está directamente relacionada con la frecuencia de visitantes que tiene cada espacio verde

3. Análisis Exploratorio de Datos

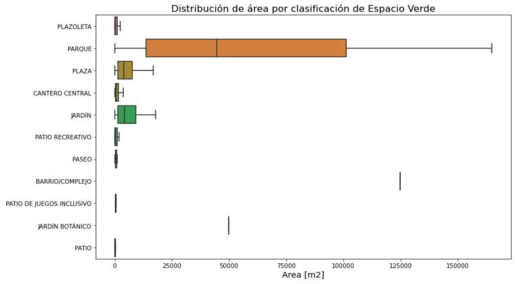
Espacios Verdes

El dataset contempla 1.736 filas y 37 columnas, donde cada fila representa un espacio verde y cada columna una característica.

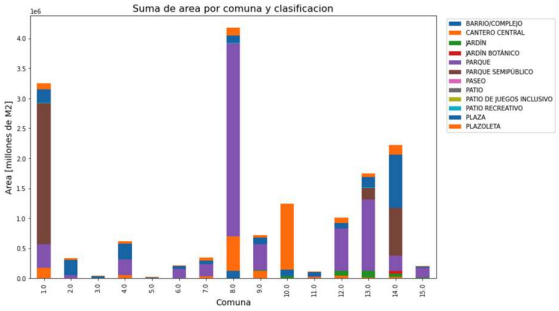
Debido a una mala calidad de algunas features, ya sea por Nulls o datos irrelevantes, se redujeron a las siguientes: Nombre, clasificación, barrio, comuna, área, perímetro y geometría. Este último es el que nos permitirá graficar los espacios dentro del mapa con la extensión de geopandas. Para comenzar a explorar los datos se realizó un boxplot de las áreas de los EV según su clasificación y se obtuvo lo siguiente:

[1] OECD iLibrary | Environmental sustainability in metropolitan areas

2013-English-OECD

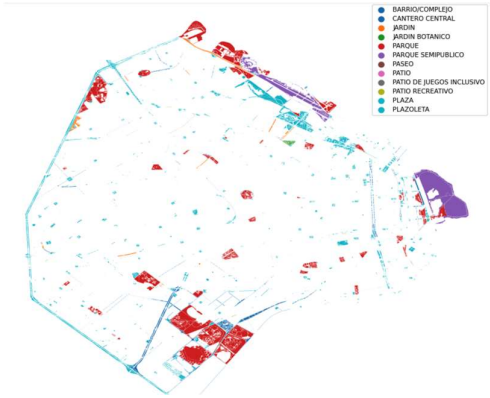


(*) Para el siguiente gráfico se eliminaron los outliers y la categoría "Parque semipúblico" para que la escala sea apreciable visualmente. Luego, para entender cuál es la distribución por comuna se procedió a realizar un barplot de la suma total de M2 de EV por comuna subdividido por clasificación:



Se observan grandes diferencias entre las comunas.

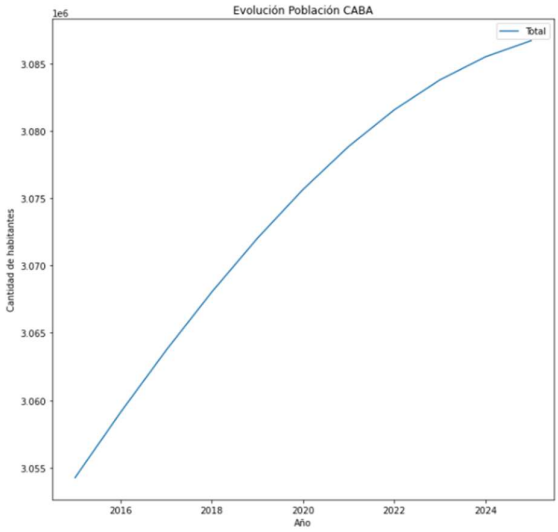
.Las clasificaciones que más M2 aportas son aquellas que menos cantidad hay. Finalmente se procedió a generar un gráfico del mapa para visualizar dónde se encuentran y qué volumen ocupan los distintos EV dentro de la ciudad, según su clasificación.



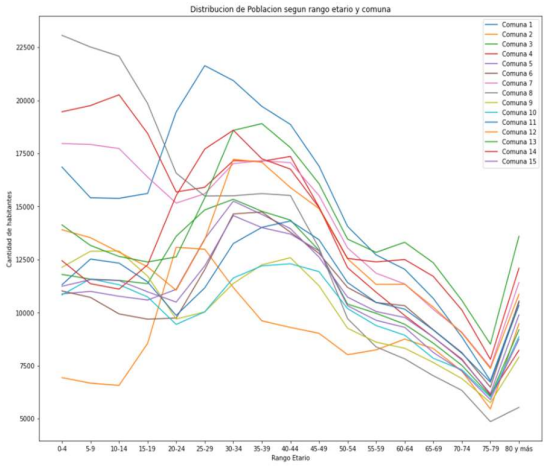
Población

El dataset contemplaba 396 filas y 19 columnas. Para comenzar la limpieza analizamos si existían Nulls.

Luego modificamos los nombres de las columnas de las comunas del dataset para poder identificar de mejor manera a cada una de ellas. Eliminamos las filas que representaban los subtotales de cada comuna en cada año. Mediante un plot observamos la evolución de la población en la ciudad.



Luego filtramos únicamente la información del año 2020, ya que es la que nos interesaba para conclusiones actuales sobre la situación de la Ciudad con relación a los espacios verdes. Pudimos observar particularidades de las comunas como por ejemplo la cantidad de personas jóvenes, en un rango etario de 20 a 34 años, que habitan en la comuna 1 en comparación a las otras.



Presupuesto Sancionado

El dataset contempla 401.068 filas y 22 columnas y tiene datos desde 2010.

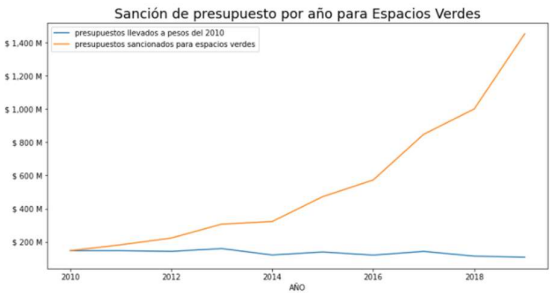
La base contiene todos los gastos que demande el desenvolvimiento de los órganos del gobierno central, de los entes descentralizados y comunas, el servicio de la deuda pública, las inversiones patrimoniales y los recursos para cubrir tales erogaciones. Los recursos publicados

detallan la distribución de los créditos aprobados al máximo nivel de desagregación.

De todas las áreas buscamos analizar la destinada a espacios verdes. Por lo tanto, filtrando el dataset obtuvimos una cantidad de 4226 registros distribuidos entre los años 2010 y 2019.

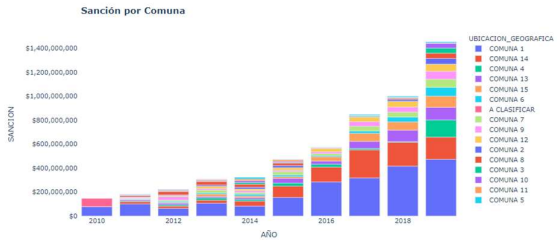
El paso siguiente fue analizar cuánto de ese presupuesto se sancionó para obras relacionadas a espacios verdes. Usando herramientas gráficas de Python obtuvimos los siguientes resultados:

Se observa una tendencia creciente y exponencial de los sancionado en pesos argentinos. Pero como se sabe esta moneda ha ido perdiendo valor a lo largo de los años. Por lo tanto, para un correcto análisis de la tendencia en lo presupuestado afectamos las inversiones llevándolas todas a un valor de referencia. En este caso a los valores de la moneda en 2010. Nuevamente graficamos y obtuvimos los siguientes resultados:



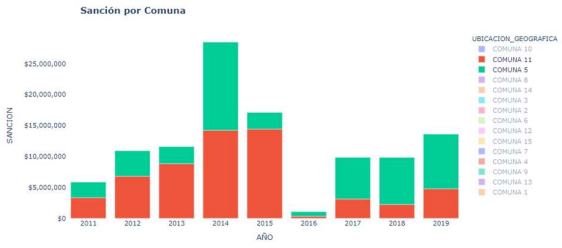
Como se puede observar, las inversiones, a niveles reales, caen año a año para obras relacionadas a espacios verdes. Además, se puede ver que para los años impares suele haber picos de presupuesto, lo que nos induce a pensar que se relaciona con temas electorales.

Por último, en este EDA se analizaron las inversiones por comuna.

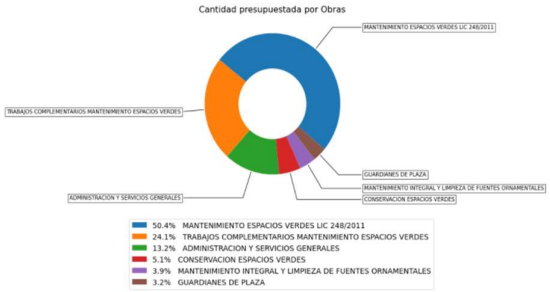


En este análisis obtuvimos resultados dispares donde las comunas que más dinero presupuestan para las obras son las comunas 1 y 14 y caso contrario las que menos inversión tienen son las comunas 11 y 5.

Por último, cabe aclarar que individualmente por comuna, hay años donde se presencian picos de inversión debido a paquetes presupuestarios que se distribuyen a todas las comunas.



Dentro de las categorías de lo presupuestado lo que más dinero se lleva son los programas de mantenimiento de los espacios verdes.



Conclusiones parciales:

Internamente dentro de la Ciudad no es equitativa la distribución del presupuesto destinado en espacios verdes. Lo que consideramos principal motivo de la disparidad en m2 de espacios verdes por comunas.

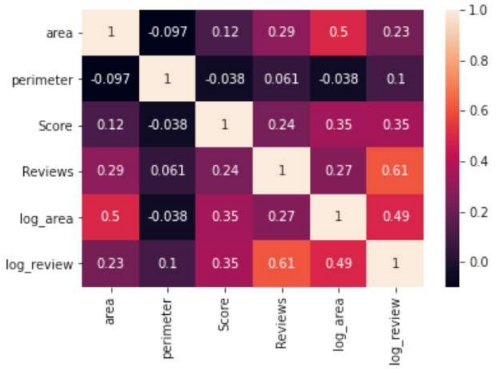
La ciudad en comparación a los centros urbanos de otros países está muy por debajo de lo recomendado y la tendencia en los presupuestos indica que lo seguirá estando.

Merge EV + Scrap Google Maps

Se realizó el merge entre el dataset de los espacios verdes y los datos scrapeados de Google Maps mediante la variable “nombre” del espacio verde. De las 1736 samples que se tenían originalmente se logró hacer match con 235, por lo que se buscó estimar el resto mediante una regresión.

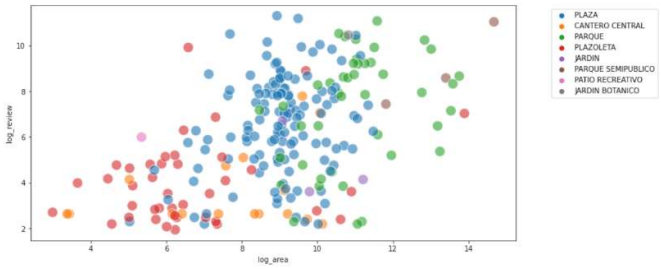
Previo a las pruebas de regresión se realizaron algunos gráficos para entender la situación.

Heatmap entre variables continuas de los espacios verdes y las clasificaciones (Score) y opiniones (reviews):



(*) *Se transformaron las features mediante un logaritmo ya que había mucha variación entre las mismas*

Se tomo como hipótesis de que existe una relación lineal entre el Área de cada espacio verde y sus Reviews, debido a que si se tiene mayor espacio, más personas concurrirán y más opiniones recibirá, para visualizarlo se realizó un scatter plot entre las variables Área y Reviews.



A grandes rasgos pareciera existir una relación lineal entre estas variables.

Tambien se realizó un gráfico con Geopandas para entender la distribución geográfica y la cantidad de reviews por cada EV, se obtuvo lo siguiente:

4. Materiales y Métodos

Regresión lineal

La regresión es su forma más sencilla se llama regresión lineal simple. Se trata de una técnica estadística que analiza la relación entre dos variables cuantitativas, tratando de verificar si dicha relación es lineal.

Su objetivo es explicar el comportamiento de una variable Y, que se denomina variable explicada (o dependiente o endógena), a partir de otra variable X, que se la llama variable explicativa (o independiente o exógena).

Mediante las técnicas de regresión se inventa una variable $\hat{Y}$ como función de otra variable X (o viceversa).

El criterio para construir esta función es que la diferencia entre Y e $\hat{Y}$, denominada error o residuo, sea pequeña.

$$\hat{Y} = f(X), \qquad Y - \hat{Y} = \text{error},$$

Los residuos o errores ei son la diferencia entre los valores observados (verdadero valor de Y) y los valores pronosticados por el modelo: ei =Y-$\hat{Y}$. Recogen la parte de

la variable Y que no es explicada por el modelo de regresión.

A partir de la definición de residuo, podemos escribir Y = f(X) + error. El término llamado error debe ser tan pequeño como sea posible. El objetivo será buscar la función (modelo de regresión) Ŷ= f(X) que lo minimice. [2]

Regresión Ridge

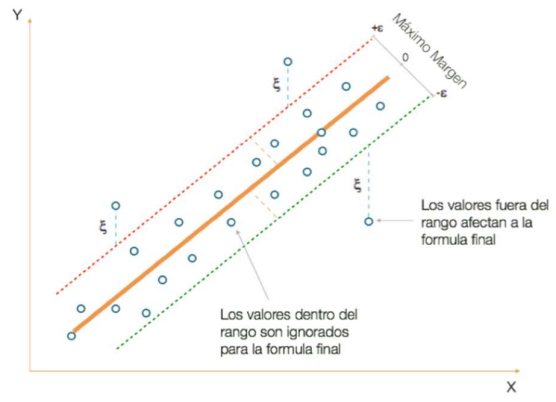
La Regresión Rigde regulariza el modelo resultante imponiendo una penalización al tamaño de los coeficientes de las características predictivas y la variable objetivo. En este caso, los coeficientes calculados minimizan la suma de los cuadrados de los residuos penalizada al añadir el cuadrado de la norma L2 del vector formado por los coeficientes:

$$RSS_{ridge} = \sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda \sum_{j=1}^{p} {\beta_j}^2$$

...donde λ es un parámetro que controla el grado de penalización: cuanto mayor éste, los coeficientes serán menores resultando más robustos a la colinealidad. Cuando α es igual a cero, Ridge es equivalente a la regresión lineal.[3]

Support Vector Regression (SVR)

Este algoritmo se basa en buscar la curva o hiperplano que modele la tendencia de los datos de entrenamiento y según ella predecir cualquier dato en el futuro. Esta curva siempre viene acompañada con un rango (máximo margen), tanto del lado positivo como en el negativo, el cual tiene el mismo comportamiento o forma de la curva.



Todos los datos que se encuentren fuera del rango son considerados errores por lo que es necesario calcular la distancia entre el mismo y los rangos. Esta distancia lleva por nombre epsilon y afecta la ecuación final del modelo.[4]

5. Resultados

Empleamos modelos de regresión con el objetivo de poder predecir para los espacios verdes su valoración de Google Maps y si existe una relación entre el área de los mismos y su reviews correspondientes en Google Maps.

Regresión lineal

Dividimos nuestro train (90%) y test (10%) para poder realizar la regresión.

El algoritmo nos brindó el siguiente resultado de R2 (coeficiente de determinación):

R2=0,24

Con Polynomial Features el número fue:

R2=0,22

Regresión Ridge

Utilizando los mismos porcentajes de train y test que en la regresión lineal, el resultado que obtuvimos fue:

R2=0.169

Support Vector Regression (SVR)

Por último, utilizando nuevamente los porcentajes de la regresión lineal para nuestros datos de entrenamiento y de testeo el resultado fue:

R2=0,21

6. Discusión y conclusiones

Debido a los bajos rendimientos de todos los métodos de regresión aplicados creemos que no es correcto realizar la regresión para las samples que no obtuvieron match.

La baja cantidad de match entre el nombre de los EV del dataset del Gobierno de la Ciudad y el scrap de Google se debe a que el gobierno tiene en cuenta muchos espacios como canteros o plazoletas y Google Maps tenía en su mayoría plazas y parques, debido que son los más concurridos.

La relación lineal que creíamos ver en el scatter plot no era lo suficientemente representativa y cuando se intentó separarlo por clasificación de EV el resultado fue incluso menos acertado.

---

[2] Laguna, C. (2014). Correlación y regresión lineal. Instituto Aragonés de Ciencias de la Salud, 1-18.

[3] Regresión Ridge | Interactive Chaos

Spanish

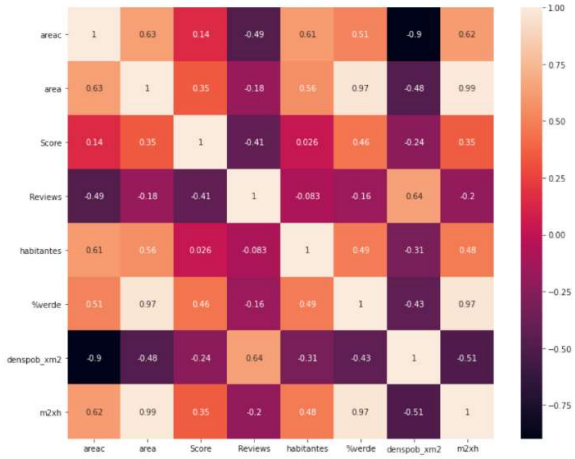
[4] Aprendizaje Supervisado: Support Vector Regression 2020SpanishL. Gonzalez-Aprende IA

Debido a que las features del dataset de EV no presentaban ningún tipo de relación, proseguimos a mergear la información de los distintos EDA con la variable “comuna” y buscar nuevamente la existencia de alguna relación entre las variables de la población y presupuestos (orientados a los EV) con las valoraciones de los habitantes de la ciudad.

Para ello, luego de hacer el merge, se crearon nuevas features cuantificables que son de relevancia a la hora de analizar esta información, los mismos son:

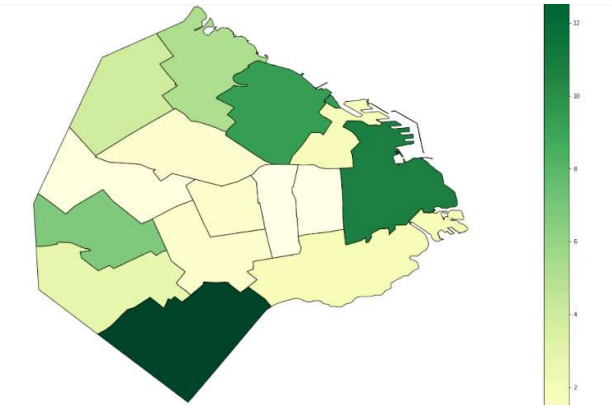
- Porcentaje de M2 de EV sobre M2 totales en cada comuna
- Densidad poblacional (en M2) en cada comuna.
- M2 de EV por habitante en cada comuna.

Luego se generó un en heatmap para analizar las features de los distintos EDA sumado a las recientemente creadas y medir la relación lineal con las valoraciones:



Debido a que es un conjunto de datos de tan solo 15 samples (ya que hay 15 comunas) no tendría sentido aplicar algún algoritmo de ML, pero si se puede entender cuán relacionadas están las variables. Como por ejemplo Mientras más densidad poblacional en la comuna, mayor será la cantidad de reviews en sus EV, por eso el valor obtenido es de 0,64.

Finalmente se realizó un gráfico en Geopandas de la cantidad de M2 de EV por comunas donde se puede ver cuales cumplen con la recomendación de la ONU, siendo esta mayor a 9 M2 de EV por habitante:



Se puede observar que muy pocas comunas consiguen alcanzar o superar el nivel recomendado por la ONU, creemos que hay mucho trabajo por hacer en este aspecto ya que los espacios verdes son muy beneficiosos para salud y cultura de las personas. la Ciudad Autónoma de Buenos Aires es una de las ciudades cosmopolitas más grandes de Sudamérica y su demografía crece a niveles exponenciales, es por eso que creemos que el análisis de estos datos y la propuesta de nuevas ideas y proyectos son fundamentales y urgentes.