8-10-2022

Predicción de la generación municipal de residuos orgánicos. Una aproximación desde el aprendizaje de máquinas.

Estudiante: Juan Luis Orrego Henao

Director: Paula María Almonacid Hurtado
Departamento de Finanzas
palmona1@eafit.edu.co
Investigadora Junior (Minciencias)
Profesora Asociada (EAFIT)

Palabras clave
Residuos orgánicos; Residuos sólidos urbanos; Depósito de basura; Relleno sanitario; Aprendizaje automático.
Keywords
Organic Waste; Municipal Solid Waste; Waste Disposal; Landfill; Machine Learning.

Contenido

Resumen

Una de las grandes problemáticas que tiene cualquier ciudad en el mundo es la disposición y el tratamiento de los residuos generados por los hogares, comercios e industrias, y más aún cuando hay diferentes iniciativas a nivel mundial y Objetivos de Desarrollo Sostenible encaminadas a la protección del medio ambiente. El reto en torno a este tema es lograr que los diferentes usuarios del sistema separen adecuadamente los residuos inservibles, reciclables y orgánicos. Sin embargo, una vez logrado esto se tiene que buscar la disposición y el tratamiento independiente de estos tres.

Alrededor del mundo se han desarrollado diferentes estrategias entorno al tratamiento del material reciclable y el uso de rellenos sanitarios para la disposición final de los residuos inservibles, pero aún son pocas las entidades territoriales que tienen un progreso significativo en el tratamiento de los residuos orgánicos. En la mayoría de los casos, estos últimos terminan en los rellenos sanitarios junto con todo el material inservible, generando una serie de líquidos lixiviados los cuales son sumamente nocivos para el medio ambiente, en especial para las fuentes hídricas.

En este proyecto se utilizan técnicas de aprendizaje de máquinas para modelar y predecir la generación municipal de residuos orgánicos, que será la materia prima de una planta de residuos orgánicos. Al ser este un servicio público y operar con recursos públicos tiene la obligación de ser lo más eficiente posible, de tal manera que los recursos que se utilicen en la planta para tratar los residuos orgánicos, como mano de obra u otros insumos químicos u orgánicos, no sean excesivos y se malgaste el presupuesto; o por el contrario sean pocos y no se tenga la capacidad de tratar el 100% de los residuos orgánicos, y que estos terminen en el relleno sanitario generando un sinnúmero de complicaciones. El modelado y predicción de la generación de residuos orgánicos brinda información que permita la operación eficiente de una planta de este tipo. Como resultados, se analizaron las métricas de modelos lineales y no lineales, univariados y multivariados, donde el mejor ajuste lo logró el Modelo autorregresivo integrado de media móvil (ARIMA).

Introducción

Usualmente, una planta de transformación industrial, como es el caso del acero o del plástico, tiene su ritmo de trabajo basado en la cantidad de producto demandado. Por el contrario, en las plantas de Residuos Orgánicos el ritmo de trabajo está dado por la oferta, es decir la cantidad de Residuos Orgánicos que reciba, producto de los desechos generados por los ciudadanos, comercios e industrias de cada municipio, la cual es la materia prima que se transformará para obtener un producto final aprovechable que permita aportar a una economía circular.

Predecir con la mayor exactitud posible la cantidad de Residuos Orgánicos que genera un municipio día tras día, brinda información estratégica y de utilidad al área administrativa de la planta, para así poder conocer cuál será el ritmo de trabajo necesario y planificar los recursos humanos, técnicos y de insumos para operar con la mayor eficiencia posible.

Las Empresas Públicas de La Ceja ESP se convirtió en la única empresa colombiana en ganar el proyecto FASEP, estrategia convocada por el Gobierno francés a través de su embajada en Colombia y que busca brindar recursos económicos, con el fin de optimizar y mejorar con tecnología e innovación el proceso de separación de residuos en la localidad. Para llevar a cabo

esta iniciativa, el gobierno francés aportará $2.141 millones de pesos, mientras que la empresa aportará una inversión de $2.993 millones de pesos, para un total de $5.134 millones de pesos.

Mediante este trabajo se espera brindar luces para reducir el impacto ambiental y devolver el residuo al posconsumo, a través de la reducción de un 50% de la cantidad de desechos dispuestos en el relleno sanitario, y con esto aprovechar el 90% de los orgánicos producidos en la localidad haciendo compostaje y combustibles, que sea un referente a nivel nacional para tener plantas innovadoras y proyectos con tecnología.

El proyecto de las EEPP fue elegido junto a dos más en América Latina. En el caso de la Ceja, los funcionarios de la Embajada de Francia en Colombia serán los encargados de inspeccionar y hacer vigilancia a este proyecto.

Por todo la anterior, se vuelve sumamente relevante la capacidad de las Empresas Públicas de La Ceja de operar de la manera más eficiente posible esta nueva planta, y para lograrlo es necesario predecir con la mayor exactitud posible la cantidad de residuos orgánicos que se generarían día tras día en el municipio y que será la materia prima de la planta.

Si la asignación de recursos, acorde a la planeación, es inferior a la necesidad de tratamiento de los residuos que genera el municipio y llegan a la planta no se podrá tratar el 100% de los residuos, y los que no tengan la posibilidad de ingresar a la planta terminarán en el relleno sanitario, impactando negativamente al medio ambiente a través de líquidos lixiviados que se producen. Y lo que busca esta gran inversión de dinero es evitar este efecto.

Por el otro lado, si la planeación asigna más recursos de los necesarios para tratar los residuos orgánicos que lleguen a la planta, se podrá tratar el 100% de los residuos, sin embargo, se habrán gastado recursos, que son de carácter público, ineficiente e inoficiosamente y que no se podrán recuperar. No sobra aclarar en este punto que, los recursos públicos son propiedad de todos los ciudadanos y que son limitados y vigilados. En este sentido, buscando predecir la cantidad de Residuos Orgánicos que producirá un municipio, a partir de sus datos históricos, en este trabajo se aplicó ingeniería de características para determinar las mejores variables que permitan predecir el comportamiento de la generación de residuos orgánicos en el municipio, acompañada de la estadística descriptiva necesaria. Se definieron los modelos de aprendizaje de máquinas adecuados para la predicción de generación de residuos orgánicos, junto con las métricas adecuadas para evaluar el desempeño y la eficacia de los diferentes modelos; y se modeló y validó la opción que presentó el mejor comportamiento para la predicción de la generación de residuos orgánicos en el municipio.

Estado del Arte y Marco Teórico

Cuando se refiere al tema de predicción de la generación de residuos sólidos a nivel municipal, se puede encontrar una literatura que incrementa cada vez más, y la componen modelos sumamente heterogéneos. Solo en el periodo de 1970 hasta el 2014 se puede hallar más de 80 estudios publicados que abordan esta problemática y se pueden agrupar en 5 categorías diferentes: métodos de estadística descriptiva, análisis de regresiones, análisis de flujo de materiales, análisis de series de tiempo y modelos de inteligencia artificial. Sin embargo, todas estas aproximaciones tienen sus fortalezas y debilidades.

Modelo autorregresivo integrado de media móvil (ARIMA)

En estadística y econometría, y en particular en el análisis de series de tiempo, un modelo de promedio móvil integrado autorregresivo (ARIMA) es una generalización de un modelo de promedio móvil autorregresivo (ARMA). Ambos modelos se ajustan a datos de series de tiempo para comprender mejor los datos o para predecir puntos futuros en la serie. Los modelos ARIMA se aplican en algunos casos donde los datos muestran evidencia de no estacionariedad en el sentido de media, donde un paso inicial de diferenciación se puede aplicar uno o más veces para eliminar la no estacionariedad de la función media. Cuando la estacionalidad se muestra en una serie temporal, se podría aplicar la diferenciación estacional para eliminar el componente estacional.

La parte AR de ARIMA indica que la variable de interés en evolución se retrocede sobre sus propios valores rezagados. La parte MA indica que el error de regresión es en realidad una combinación lineal de términos de error cuyos valores ocurrieron simultáneamente y en varios momentos en el pasado. La I (de "integrado") indica que los valores de los datos han sido reemplazados con la diferencia entre sus valores y los valores anteriores. El propósito de cada una de estas características es hacer que el modelo se ajuste a los datos lo mejor posible.

Los modelos ARIMA no estacionales generalmente se denominan ARIMA(p,d,q) donde los parámetros p, d y q son números enteros no negativos, p es el orden (número de retrasos) del modelo autorregresivo, d es el grado de diferenciación (la cantidad de veces que se restaron los valores anteriores de los datos), y q es el orden del modelo de promedio móvil. Los modelos ARIMA estacionales generalmente se denominan ARIMA(p,d,q)(P,D,Q)m, donde m se refiere al número de períodos en cada temporada, y las mayúsculas P,D,Q se refieren a la diferenciación autorregresiva, y términos de promedio móvil para la parte estacional del modelo ARIMA.

K Vecinos más cercanos (KNN)

Esta es una técnica que suele usarse en problemas de clasificación, a partir del aprendizaje supervisado, es decir que obligatoriamente necesita un conjunto de datos de entrenamiento, que, matemáticamente hablando, sirven para estimar las funciones de densidad.

El concepto sobre el cual está basado es bastante sencillo. En primer lugar, se le debe asignar un valor a K. Con el fin de ilustrar un ejemplo, se supondrá que existen datos divididos en dos categorías (A y B) y se asignará a K el valor de 3. Los datos de entrenamiento están previamente clasificados, y a la hora de evaluar un nuevo dato, se tomarán los k vecinos más cercanos a ese dato en el espacio para determinar a cuál categoría pertenece. En este caso se tomarán los 3 vecinos más cercanos. El nuevo dato pertenecerá a la categoría que pertenezcan la mayoría de estos 3 vecinos.

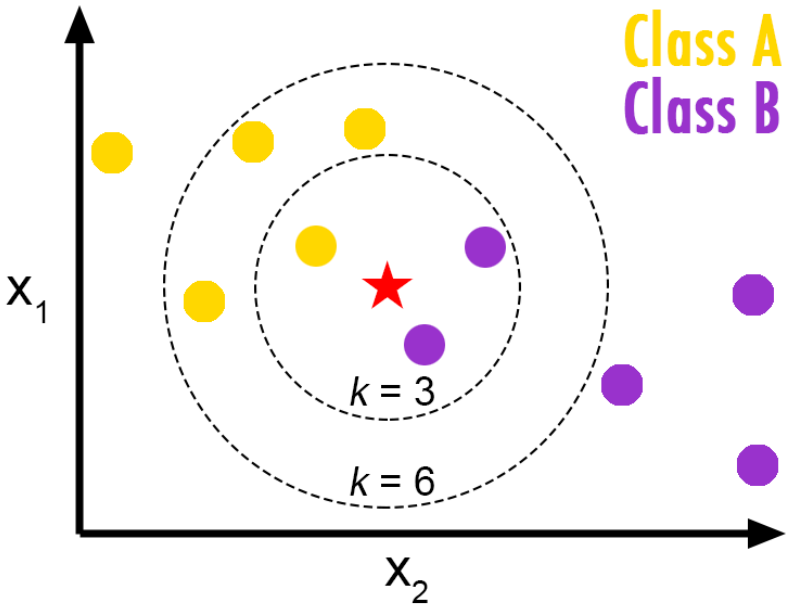

Ilustración 1. Ejemplo de KNN

Esta práctica se vio aplicada en Logan City (Abassi, 2016) y en Vietnam (Nguyen, 2021). En ninguno de los casos presentó mejores métricas que los algoritmos con los que se comparaba. Sin embargo, es de resaltar que en Vietnam tuvo MAE=121.5, RMSE=202.3 y R2=0.96.

Árboles de regresión, random forest y potenciación de gradiente

Los árboles, dentro del aprendizaje automático son sumamente famosos debido a su simplicidad. Cuando la variable explicada puede tomar valores continuos se le conoce como árbol de regresión. El objetivo es crear un modelo que predice el valor de una variable de destino en función de diversas variables de entrada. Un árbol puede ser "aprendido" mediante el fraccionamiento del conjunto inicial en subconjuntos basados en una prueba de valor de atributo. Este proceso se repite en cada subconjunto derivado de una manera recursiva llamada particionamiento recursivo. La recursividad termina cuando el subconjunto en un nodo tiene todo el mismo valor de la variable objetivo, o cuando la partición ya no agrega valor a las predicciones.

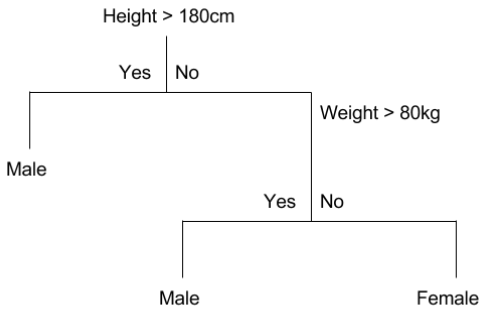

Ilustración 2. Ejemplo de Árbol de Decisión (Desicion Tree)

Los Random Forest, conocidos en español como '"Bosques Aleatorios"' es una combinación de árboles predictores tal que cada árbol depende de los valores de un vector aleatorio probado independientemente y con la misma distribución para cada uno de estos.

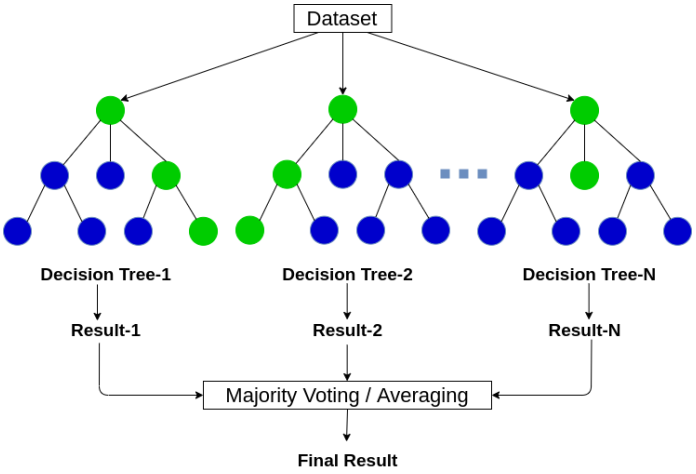

Ilustración 3. Ejemplo de Random Forest

Finalmente, dentro de este grupo de algoritmos se encuentran los árboles de regresión con potenciación de gradiente el cual produce un modelo predictivo en forma de un conjunto de árboles de decisión. Construye el modelo de forma escalonada como lo hacen otros métodos de potenciación, y los generaliza permitiendo la optimización arbitraria de una función de pérdida diferenciable. La potenciación puede ser interpretada como un algoritmo de optimización en una función de coste adecuada.

Para analizar el desempeño de este grupo de algoritmos es necesario traer a colación el estudio realizado en la República Checa, donde se implementaron los 3 modelos a nivel municipal. Pero los bosques aleatorios y la potencialización del gradiente al ser estructuras más complejas derivadas del mismo árbol de regresión, presentan un mejor desempeño. Por un lado, los bosques aleatorios presentaron un desempeño que se puede resumir con las siguientes métricas: RMSE= 0.02, MAE= 0.02, MAPE=8.52, R2=0.56, mientras que la potencialización del gradiente muestra los siguientes resultados: RMSE= 0.01, MAE= 0.01, MAPE=85.66, R2=0.18 (Rosecký, 2021). Sin embargo, en el estudio que se llevó a cabo en Vietnam, los bosques aleatorios presentaron estadísticas aún más alentadoras, donde MAE= 125, RMSE=201.6 y R2=0.96 (Nguyen, 2021).

Metodología

Para abordar el desarrollo de este proyecto, se propuso la incorporación de la metodología CRISP-DM (Cross Industry Standard Process for Data Mining), más para la ejecución de este proyecto solo se utilizaron 5 de las 6 fases que la metodología propone:

- Comprensión del Negocio
- Comprensión de los datos
- Preparación de los datos
- Modelado
- Evaluación

Comprensión del negocio

El dueño de los datos, y a quien se considera el sponsor del proyecto es la oficina de Manejo Integral de Residuos Sólidos de las Empresas Públicas de La Ceja ESP. En esta fase se buscó formular la descripción general del problema y exponer las expectativas acerca de la solución. Su participación e involucramiento a lo largo de todo el proyecto fue de suma importancia, pues su conocimiento y dominio del tema son esenciales, en especial, las particularidades del entorno que afectan directa o indirectamente el comportamiento de la generación de Residuos Orgánicos.

Comprensión de los datos

En este proyecto se utilizaron datos reales de generación de residuos sólidos recibidos por la empresa desde agosto del 2016 hasta julio del 2021, por lo cual desde esta etapa se permite inferir que una aproximación de series de tiempo puede ser eficiente. La empresa, dentro su Sistema de Gestión de Calidad establece que, se debe registrar la llegada de los vehículos de recolección de residuos al sitio de disposición final, donde se especifique las siguientes variables:

- NUSD: Número único del sitio de disposición final.
- NUAP: Número único del área de prestación del servicio de recolección de residuos.
- Placa: Placa de identificación del vehículo recolector.
- Fecha en formato dd/mm/aaaa
- Hora de entrada del vehículo: Hora de entrada del vehículo al sitio de la disposición final, en formato de 24 h.
- Hora de salida del vehículo: Hora de salida del vehículo del sitio de la disposición final.
- Macroruta: Tipo de residuos que está recolectando el vehículo (reciclaje, orgánico, material vegetal, inservible, entre otros).
- Toneladas recogidas en suelo urbano asociadas al barrido y limpieza provenientes del área de prestación, con dos decimales de exactitud.
- Toneladas recogidas en suelo no urbano asociadas al barrido y limpieza provenientes del área de prestación, con dos decimales de exactitud.
- Toneladas dispuestas en suelo urbano asociadas a la limpieza y corte de zonas verdes provenientes del área de prestación, con dos decimales de exactitud.
- Toneladas dispuestas en suelo urbano asociadas a la recolección y disposición provenientes del área de prestación, con dos decimales de exactitud.
- Toneladas dispuestas en suelo no urbano asociadas a la recolección y disposición provenientes del área de prestación, con dos decimales de exactitud.
- Toneladas dispuestas del servicio ordinario provenientes del área de prestación, con dos decimales de exactitud.

Tabla 1. Registros aleatorios del DataFrame original.

	NUSD	NUAP	Placa	Fecha	Hora de entrada del vehículo	Hora de salida del vehículo	Macroruta	Toneladas recogidas en suelo urbano asociadas al barrido y limpieza provenientes del área de prestación	Toneladas recogidas en suelo no urbano asociadas al barrido y limpieza provenientes del área de prestación	Toneladas dispuestas en suelo urbano asociadas a la recoleccion y disposicion provenientes del área de prestación	Toneladas dispuestas en suelo no urbano asociadas a la recoleccion y disposicion provenientes del área de prestación	Toneladas dispuestas en suelo urbano asociadas a la limpieza y corte de zonas verdes provenientes del área de prestación	Toneladas dispuestas del servicio ordinario provenientes del área de prestación
5803	552.0	1044.0	LAK726	2018-05-24 00:00:00	11:35:59	11:36:15	VEGETAL	0.0	0.0	0.0	0.00	0.72	0.72
11808	552.0	1044.0	TTU439	2020-01-15 00:00:00	20:00:00	20:10:00	RECICLAJE	0.0	0.0	0.0	1.14	0.00	1.14

De la misma manera, es en este momento del proceso donde se incorporaron los datos externos de carácter poblacional y demográfico, provenientes de las bases de datos del Departamento Administrativo Nacional de Estadística (DANE). Esta dependencia cuenta con 10 registros a lo largo de la historia de Colombia, desde el 1912 hasta el 2018, donde se puede encontrar el censo poblacional del municipio de La Ceja.

Tabla 2. Censos poblacionales del Municipio de La Ceja

Año	Habitantes
1912	7878
1938	10115
1951	10568
1964	16906
1973	22157
1985	29939
1993	38709
2005	46268
2017	59386
2018	64889

Preparación de los datos

La primera situación por solucionar es la frecuencia de muestreo de los dos sets de datos. Por un lado, la recepción de residuos sólidos, en el campo “Fecha”, tiene una frecuencia diaria, mientras que los censos poblaciones no tiene una frecuencia anual constante. Se procedió a realizar una interpolación de los datos para obtener un aproximado de la cantidad de habitantes diaria a través de una ecuación de tercer orden, la cual tuvo un R2 de 0,9948.

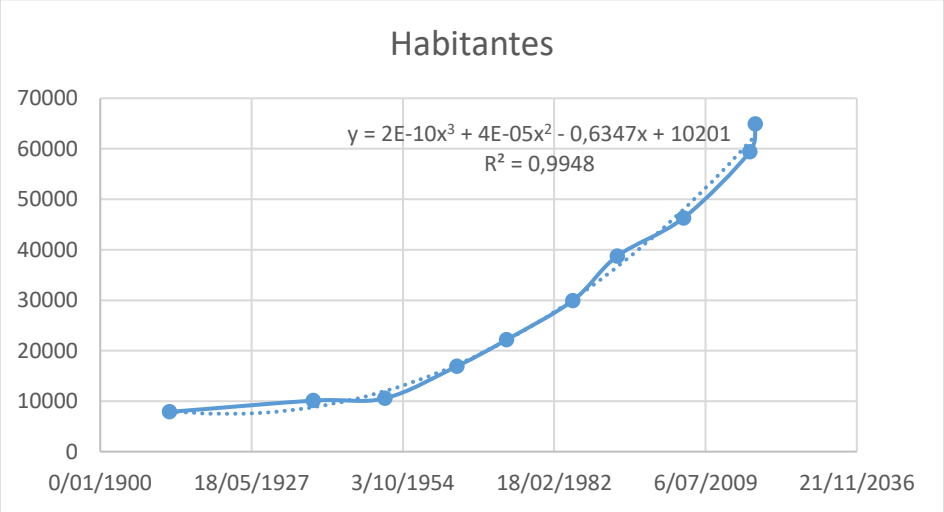

Ilustración 4. Interpolación de los habitantes del Municipio de La Ceja.

Adicionalmente, en esta etapa, se llevaron a cabo actividades de limpieza de datos para mejorar su calidad, tales como fechas erróneas o toneladas por fuera de los rangos aceptables. Esta limpieza es necesaria ya que los datos son ingresados manualmente y son muy susceptibles a errores a la hora de la digitación. Es entendible que los sets de datos, sobre todo aquellos con una gran cantidad de registros, no sean 100% limpios; sin embargo, al ser el insumo principal de los algoritmos de inteligencia artificial se entiende que se debe incluir un esfuerzo para brindarla tan limpia como sea posible, ya que de esto dependerá los resultados finales de las predicciones.

Una vez limpios los datos, la primera transformación a la que se sometió el set de datos fue un agrupamiento diario de los registros y la eliminación de columnas que presentaban valores constantes o valores repetidos, y lograr así tener en un mismo dataframe la información de la recepción de residuos sólidos y la interpolación de la cantidad de habitantes. Sin embargo, se tomó la decisión, en consenso con la empresa, de trabajar con un muestreo semanal. Lo anterior debido al comportamiento de la generación de residuos orgánicos, que en su mayoría solo se recolecta en dos de los siete días de la semana, sin embargo, la realidad es que se genera todos los días. Como índice de cada registro semanal se tomó la fecha del lunes de cada semana.

Tabla 3. Transformación del DataFrame

Fecha	Año	Semana	INORGANICO	ORGANICO	RECICLAJE	VEGETAL	Habitantes
2016-08-01	2016	31	119.565	100.120	0.00	0.000	60305
2016-08-08	2016	32	124.960	104.175	0.00	0.000	60328
2016-08-15	2016	33	124.730	104.150	0.00	0.000	60351
2016-08-22	2016	34	141.650	105.110	0.00	0.000	60374
2016-08-29	2016	35	150.138	90.070	0.00	0.000	60397

En este punto, la empresa indicó que durante el tiempo de confinamiento por el COVID-19 se logró apreciar un comportamiento atípico en las recolecciones, lo que incentivó la incorporación

de una variable Dummy llamada "COVID" donde toma el valor de 1 en aquellas fechas comprendidas entre el 20/03/2020 y el 01/06/2020.

Y, como insumo adicional, se realiza una última transformación a este set de datos, escalándolos a través de la herramienta MinMaxScaler de la librería de Scikit-Learn, para que la digestión de estos por parte de los algoritmos de inteligencia artificial sea más sencilla.

Una vez organizado el set de datos y transformado acorde a las necesidades, se continuó realizando un Análisis Exploratorio de Datos, conocido como EDA por sus siglas en inglés (Exploratory Data Analysis).

Tabla 4. Descripción estadística del DataFrame transformado sin escalar.

	Año	Semana	INORGANICO	ORGANICO	RECICLAJE	VEGETAL	Habitantes	COVID
count	261.000000	261.000000	261.000000	261.000000	261.000000	261.000000	261.000000	261.000000
mean	2018.582375	26.601533	159.428344	103.625690	14.639117	5.192360	63345.049808	0.038314
std	1.500614	15.097702	38.650372	23.798345	14.707810	3.613001	1781.846527	0.192322
min	2016.000000	1.000000	92.770000	27.570000	0.000000	0.000000	60305.000000	0.000000
25%	2017.000000	14.000000	133.640000	88.020000	0.000000	2.550000	61805.000000	0.000000
50%	2019.000000	27.000000	145.820000	104.580000	18.766000	4.830000	63351.000000	0.000000
75%	2020.000000	40.000000	176.067000	118.080000	27.340000	8.020000	64874.000000	0.000000
max	2021.000000	53.000000	334.090000	176.100000	75.815000	18.120000	66443.000000	1.000000

Tabla 5. Descripción estadística del DataFrame transformado escalado.

	Año	Semana	INORGANICO	ORGANICO	RECICLAJE	VEGETAL	Habitantes	COVID
count	261.000000	261.000000	261.000000	261.000000	261.000000	261.000000	261.000000	261.000000
mean	0.516475	0.492337	0.276224	0.512056	0.193090	0.286554	0.495283	0.038314
std	0.300123	0.290340	0.160162	0.160226	0.193996	0.199393	0.290298	0.192322
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.200000	0.250000	0.169360	0.406988	0.000000	0.140728	0.244379	0.000000
50%	0.600000	0.500000	0.219833	0.518481	0.247524	0.266556	0.496253	0.000000
75%	0.800000	0.750000	0.345172	0.609372	0.360615	0.442605	0.744379	0.000000
max	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000

En esta descripción inicial de los datos, se aprecia a primera vista el impacto del Scaler en comparación a los datos sin escalar, donde, en todas las variables, los datos pueden tomar un valor entre 0 y 1.

Sin escalar, la variable ORGANICO, que es la variable de interés, la variable dependiente, presenta un valor mínimo de 27.75 toneladas semanales y un valor máximo de 176.1, con una media de 103.62 toneladas semanales, una mediana de 104.58 y una desviación estándar de 23.8.

A través de la herramienta "Profiling Report" de la librería de Pandas, se obtiene mayor información aportante al Análisis Exploratorio de Datos. Toda esta información generada se encuentra anexa a este documento.

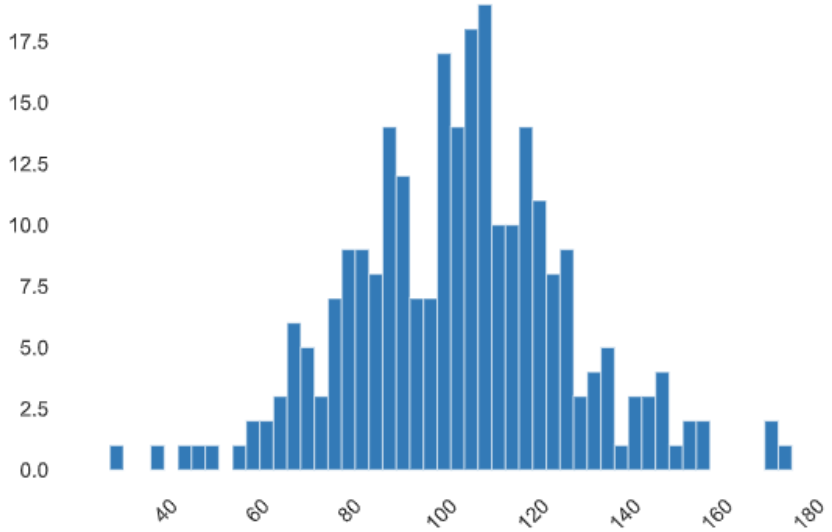

Ilustración 5. Histograma de la variable 'ORGANICO' sin escalar.

En el histograma de la variable ORGANICO se puede apreciar que su comportamiento se asemeja al de la campana normal, con su centro en la mediana de 104 toneladas como se mostraba en la descripción anterior.

Desde la etapa de la Comprensión del Negocio se tenía la hipótesis que la cantidad de habitantes podía ser un estimador importante de la generación de residuos orgánicos, y en este EDA se puede analizar la interacción que tienen estas dos variables y todas las otras, sin embargo, no se logra ver un comportamiento correlacional. Lo que si llama la atención es en la gráfica de correlaciones de Pearson, la relación negativa que este entre ORGANICO e INORGANICO, y la relación positiva entre INORGANICO y Habitantes.

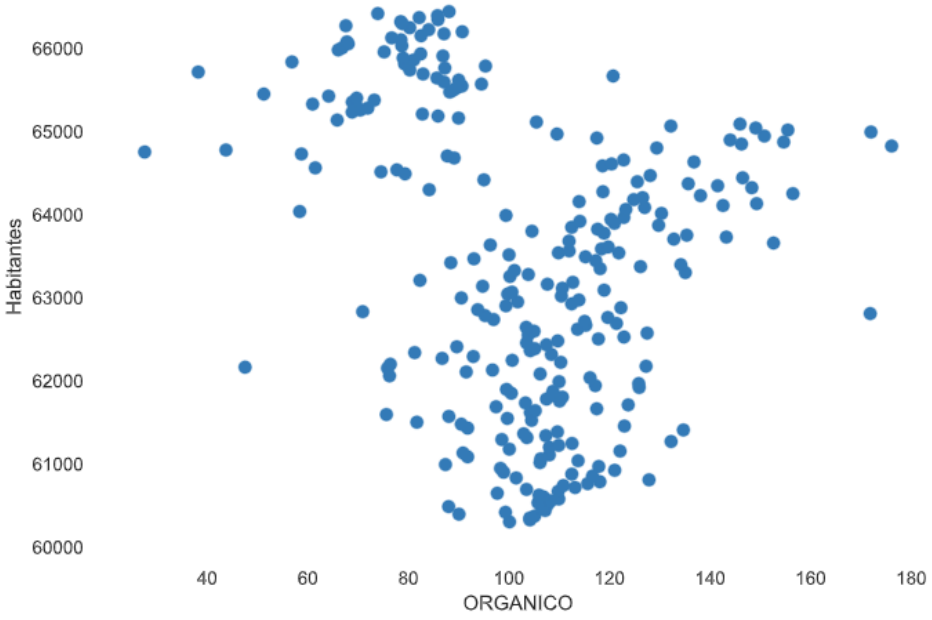

Ilustración 6. Interacción entre las variables 'ORGANICO' y 'Habitantes'

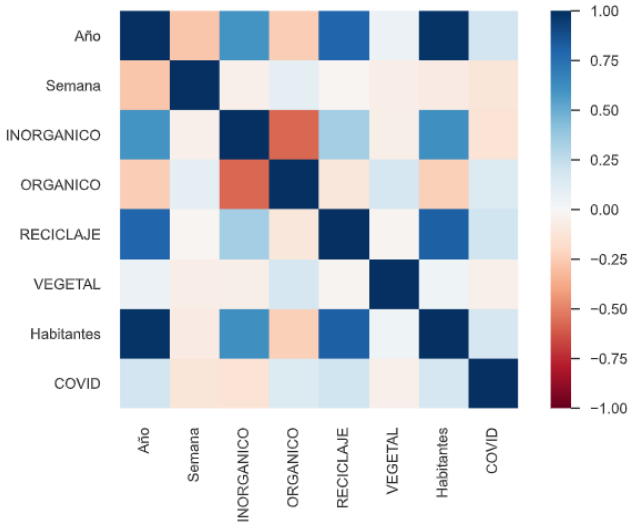

Ilustración 7. Correlaciones de Pearson

Modelado

A continuación, se presentan los diferentes modelos implementamos para la predicción de residuos orgánicos, explorando diferentes alternativas desde modelos lineales hasta modelos mucho más flexibles.

Regresión Lineal

Tabla 6. Regresión Lineal por Mínimos Cuadrados Ordinarios.

OLS Regression Results

Dep. Variable:	ORGANICO	R-squared:	0.373
Model:	OLS	Adj. R-squared:	0.356
Method:	Least Squares	F-statistic:	21.54
Date:	Fri, 07 Oct 2022	Prob (F-statistic):	1.02e-22
Time:	21:41:06	Log-Likelihood:	169.11
No. Observations:	261	AIC:	-322.2
Df Residuals:	253	BIC:	-293.7
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.3757	0.243	1.548	0.123	-0.102	0.854
Habitantes	-1.6626	2.029	-0.819	0.413	-5.658	2.333
INORGANICO	-0.6467	0.076	-8.523	0.000	-0.796	-0.497
RECICLAJE	-0.0665	0.078	-0.852	0.395	-0.220	0.087
VEGETAL	0.0930	0.042	2.200	0.029	0.010	0.176
COVID	0.0323	0.045	0.721	0.472	-0.056	0.121
Semana	0.4068	0.400	1.017	0.310	-0.381	1.194
Año	1.7875	2.024	0.883	0.378	-2.198	5.773

Omnibus:	21.966	Durbin-Watson:	1.751
Prob(Omnibus):	0.000	Jarque-Bera (JB):	70.107
Skew:	-0.227	Prob(JB):	5.98e-16
Kurtosis:	5.498	Cond. No.	519.

El punto de partida para la exploración y modelación de los datos en este caso específico es un modelo de regresión lineal múltiple. Los resultados de este modelo permitieron determinar que la mayoría de las variables explicativas propuestas a la luz de la teoría y la lógica económica no son determinantes de la variable objetivo, debido a que no resultaron ser estadísticamente significativas, tomando un nivel de significancia del 5%; y esto a pesar de que el modelo presenta un buen ajuste en términos del R-cuadrado (considerando el contexto del problema que se está trabajando). Solamente las variables "Inorgánico" y "vegetal", resultaron ser estadísticamente significativas, por lo que se podría inferir que para este caso en específico podría ser más pertinente realizar un proceso de modelación univariado de tipo lineal.

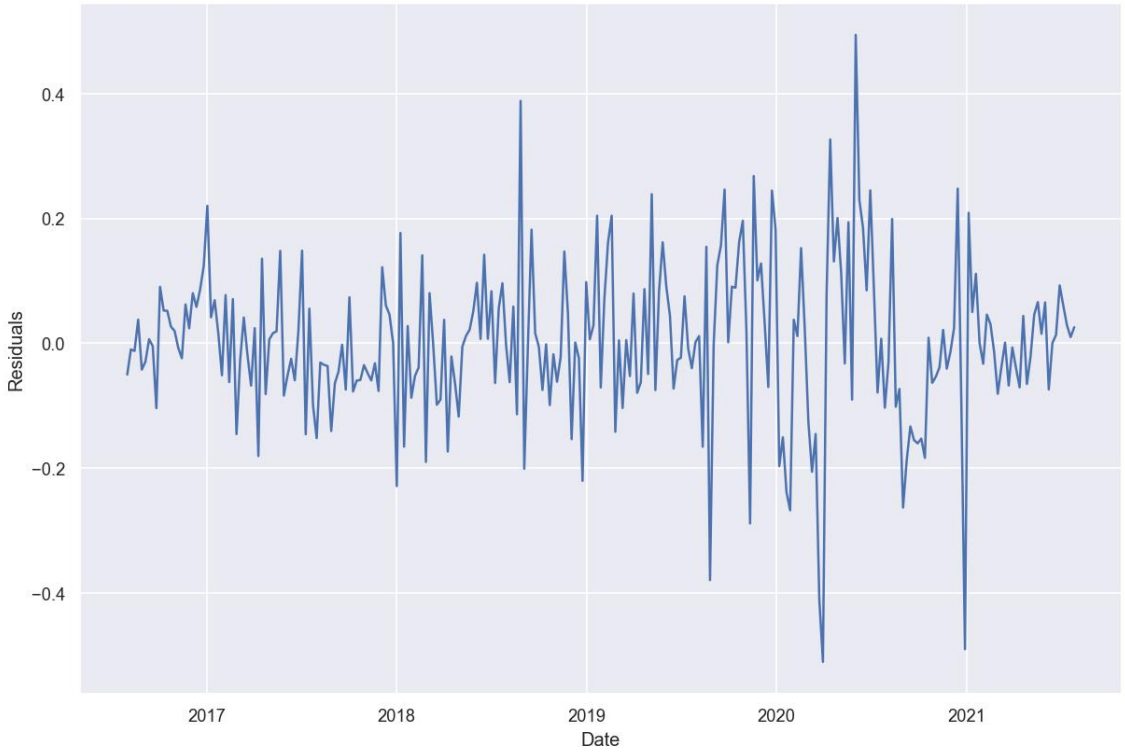

Ilustración 8. Residuales de la Regresión Lineal

La anterior hipótesis se puede apoyar mediante el análisis de los supuestos de la regresión lineal. Por ejemplo, los resultados obtenidos al aplicar el test Breusch-Pagan, permiten rechazar la hipótesis nula de homoscedasticidad, por lo que se podría concluir que los residuales presentan son heteroscedasticidad, con una mayor variabilidad presentada en la época de COVID. Por otra parte, los resultados obtenidos mediante los tests de Durbin-Watson, y de Breusch-Godfreydan se pudo concluir que hay aucorrelación entre los residuales y por tanto un modelo de series de tiempo podría contribuir a una mejor modelación de los datos. Seguidamente, en cuanto al supuesto de normalidad, el test de Jarque-Barra arroja un Chi2 de 6e-16, con lo que se infiere que los residuales no tienen un comportamiento normal, aunque su gráfica se asemeje mucho a la campana normal, con problemas de sesgo y kurtosis. Finalmente, al llevarse a cabo un test reset se obtiene un Pvalue de 0.0498 que lleva al rechazo de la hipótesis nula, lo cual permite afirmar que hay indicios de omisión de variables.

En conclusión, los resultados obtenidos de los análisis de los residuales de la regresión lineal indican que una modelación mediante algoritmos de series de tiempo puede ajustarse bastante bien al comportamiento de los datos.

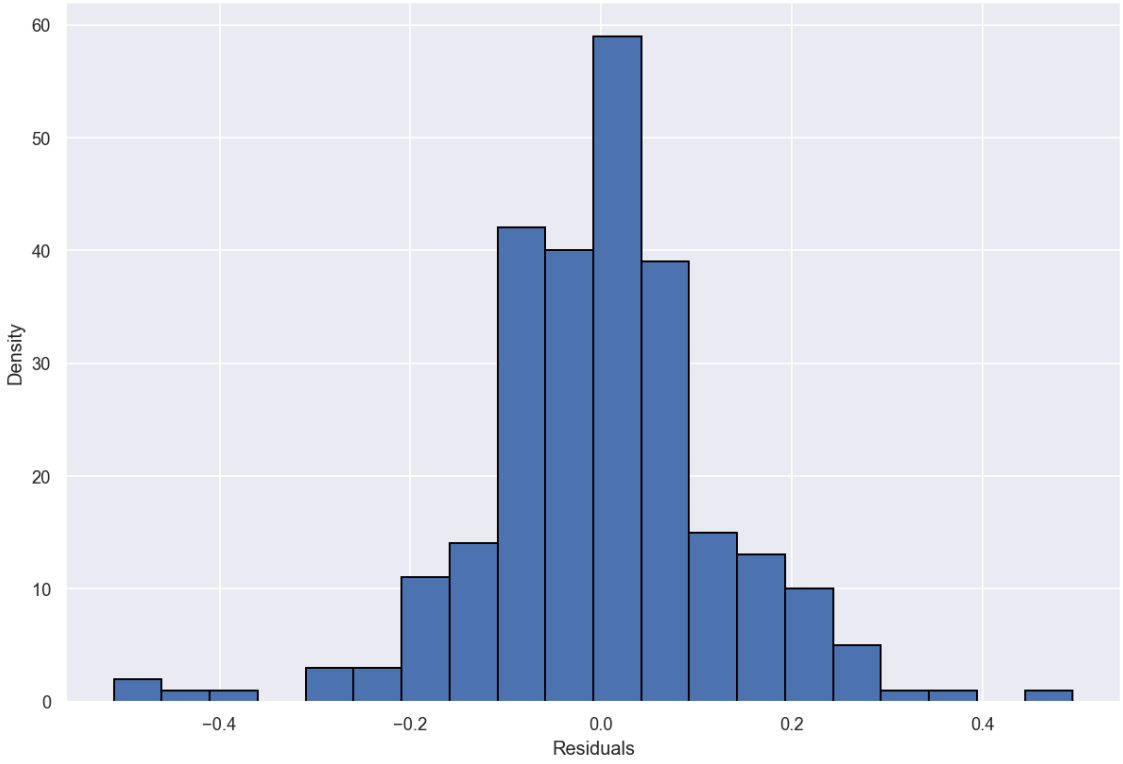

Ilustración 9. Histograma de los Residuales de la Regresión Lineal.

Series de Tiempo

A través de un analisis de series de tiempo se pueden extraer caracterísiticas como la tendencia y la estacionalidad para lograr mejorar la predicción, y estacionarizar la serie con el mismo objetivo. A primera vista se logra detectar el comportamiento atípico en los tiempos de COVID al que hacía mensión la empresa en un principio.

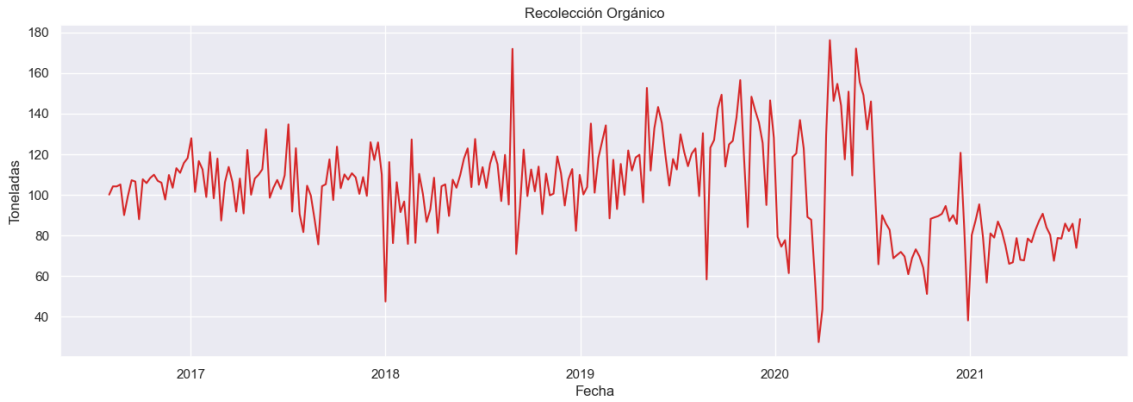

Ilustración 10. Serie de tiempo de la Recolección de Residuos Orgánicos

Al graficar el promedio de cada mes a lo largo de cada año, no se logra apreciar un comportamiento temporal o cíclico que se repita cada año.

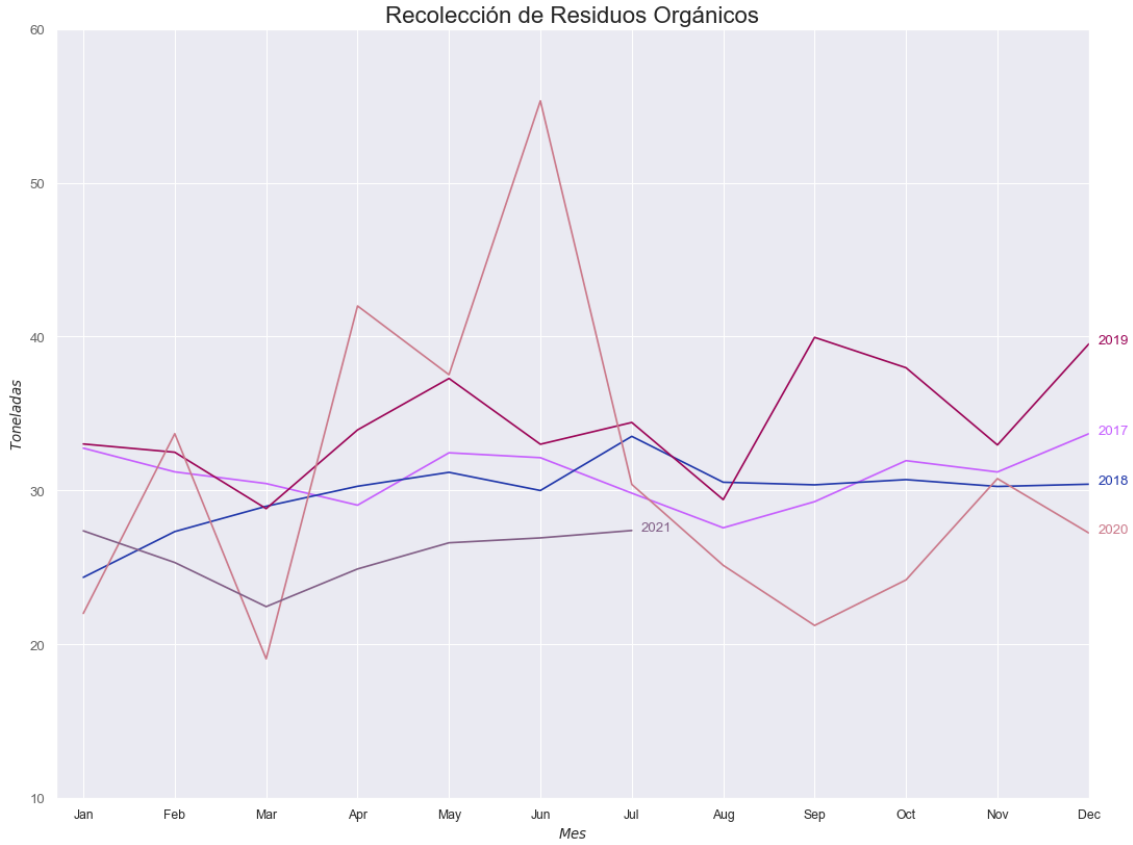

Ilustración 11. Recolección de residuos orgánicos año a año para la detección de patrones.

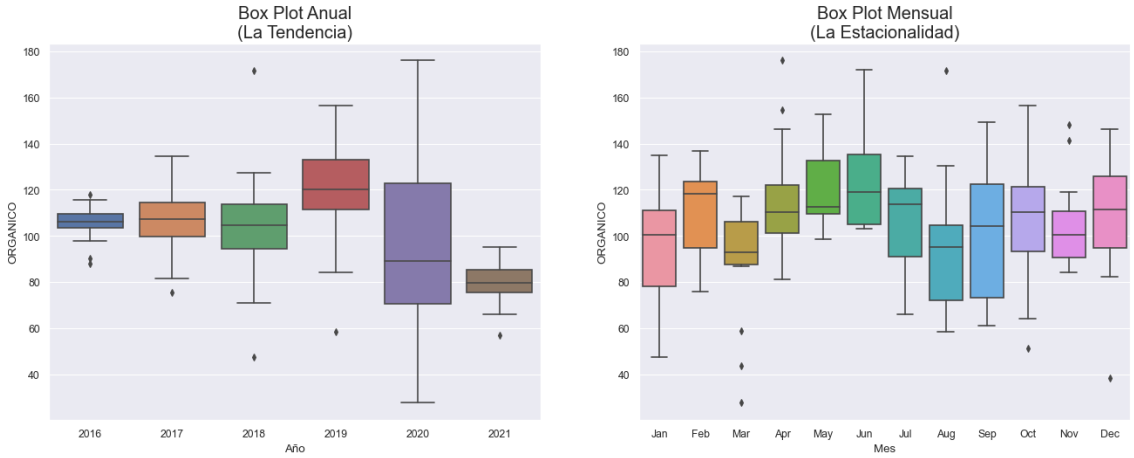

Ilustración 12. Diagramas de cajas y bigotes (boxplot) de tendencia y estacionalidad.

En este tipo de análisis grafico los diagramas de cajas y bigotes pueden brindar una idea general del comportamiento de cada periodo. En este caso, si se grafica cada año se podría apreciar la tendencia que lleva la seria de datos (al alza o a la baja) y si se grafican los meses o las semanas del año se podría ver la estacionalidad.

Para este caso, del 2016 al 2019 hay una tendencia alcista en los datos, sin embargo, 2020 y 2021 presentaron comportamientos a la baja. En cuanto en la estacionalidad, se puede apreciar un pico en el mes de junio y lo que parece ser un valle en el mes de agosto. Sin embargo, estos hallazgos gráficos no son concluyentes.

La serie de tiempo se puede descomponer, entonces, en la base inicial, la tendencia, la estacionalidad y el error. Y surgen dos caminos: sumar estos componentes para obtener el valor de cada periodo con un modelo aditivo, o multiplicarlos para obtener un modelo multiplicativo. Dependerá de los residuales de cada modelo la decisión de cual modelo se puede adaptar mejor para esta serie de tiempo.

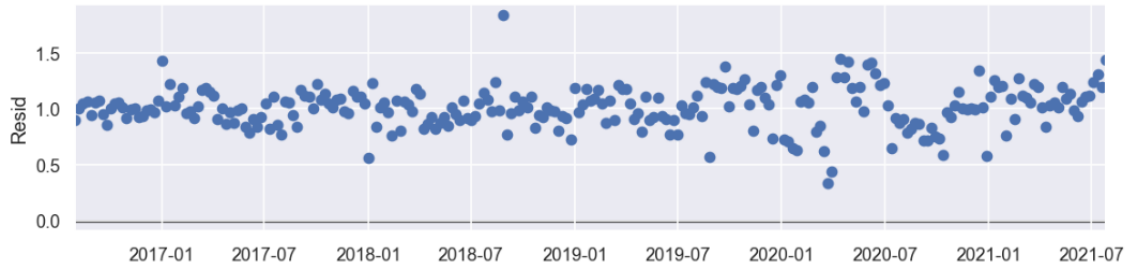

Ilustración 13. Residuales de la descomposición multiplicativa.

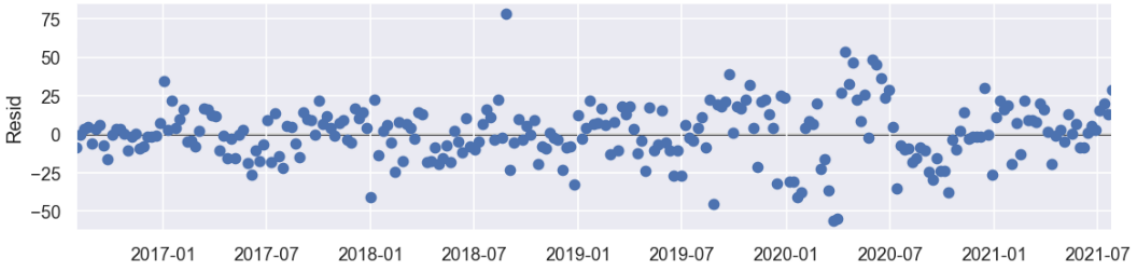

Ilustración 14. Residuales de la descomposición aditiva.

Los residuales del modelo multiplicativo se comportan en un rango entre 0.3 y 1.8, mientras que en el modelo aditivo el rango va desde -60 hasta 80, por lo cual un modelo multiplicativo se puede adaptar mejor a esta serie de tiempo.

Estacionaridad

La propiedad de estacionariedad en una serie de tiempo facilita su modelación y, por tanto, contribuye a mejorar su predicción. Por tanto, en caso de que la serie no sea estacionaria, se sugiere estacionarizarla, y para verificar que la serie es estacionaria se pueden efectuar los tests ADF y KPSS. En este caso específico se pudo concluir que la series no es estacionaria (pues se obtuvieron valores-p para ADF de 0.07 y para KPSS de 0.02), y en este sentido se realizaron las transformaciones respectivas para lograr estacionarizar la serie.

Sustracción de los componentes de tendencia y estacionalidad

Ilustración 15. Serie de tiempo de ORGANICO una vez sustraído el componente de tendencia.

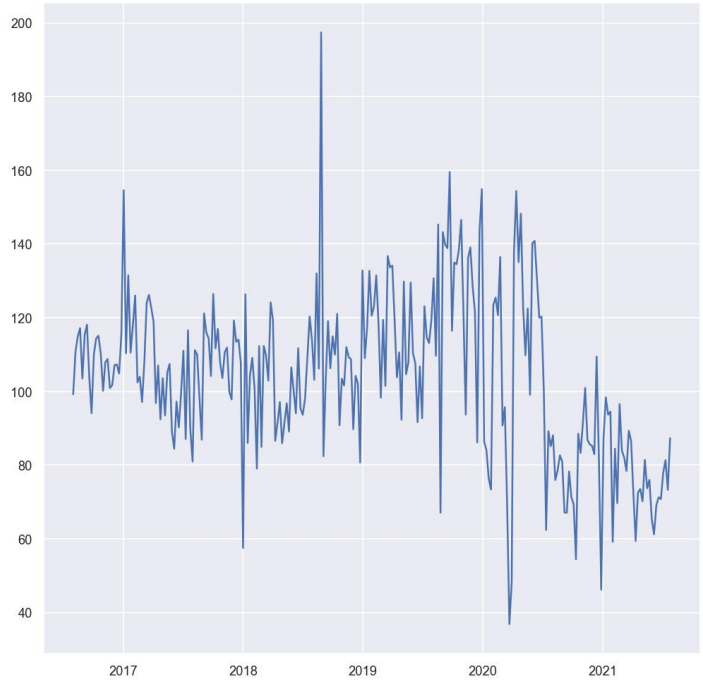

Ilustración 16. Serie de tiempo de ORGANICO una vez sustraído el componente de estacionalidad.

Al intentar extraer estos componentes de la serie no se obtienen los resultados esperados.

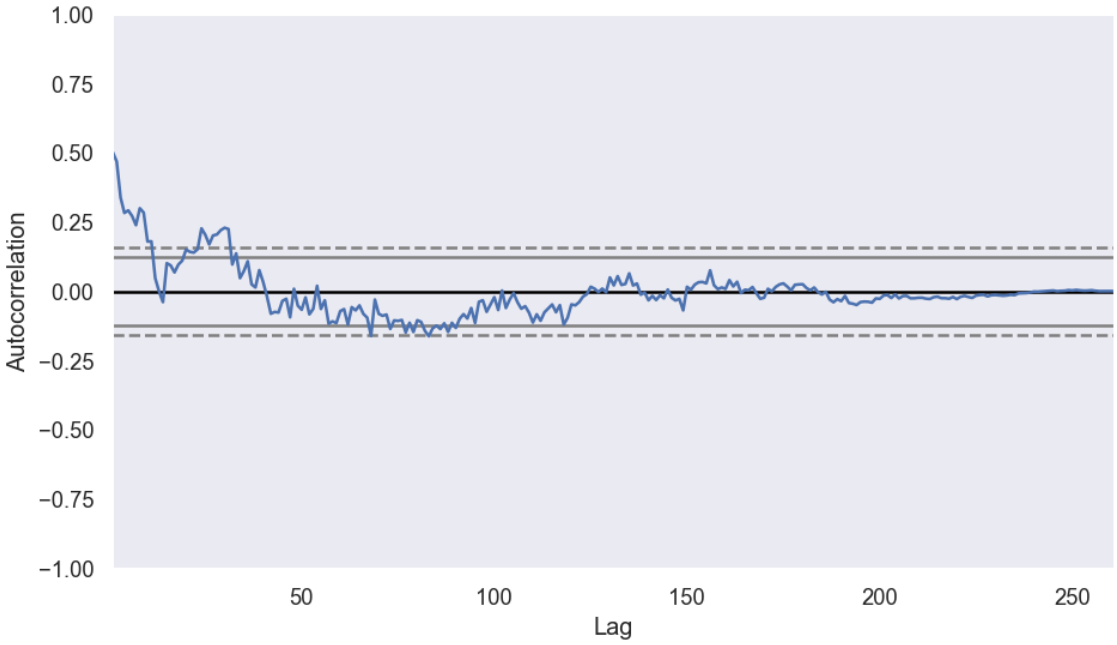

Ilustración 17. Autocorrelación de la serie de tiempo con sus valores previos.

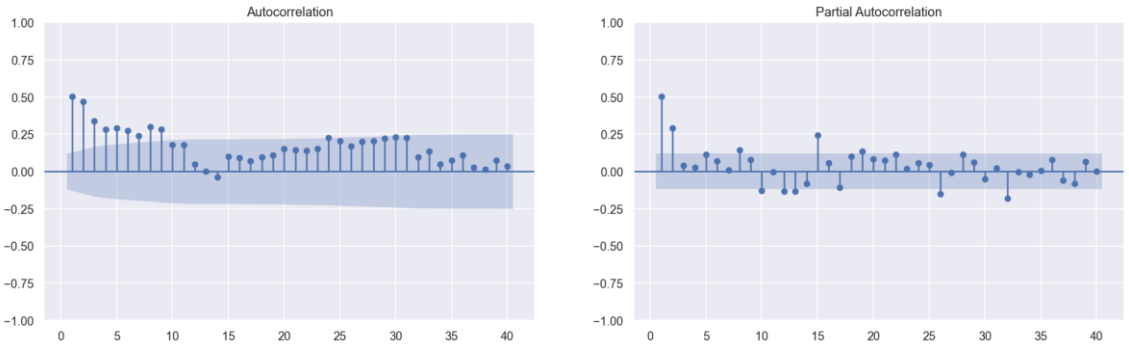

Ilustración 18. Autocorrelación (ACF) y Autocorrelación Parcial (PACF)

Por otro lado, si se analiza la correlación de la serie con valores previos de la misma se puede apreciar que los primeros 10 valores previos son estadísticamente significativos para la modelación. La grafica de autocorrelación ACF muestra una caída suavizada en los primeros 15 lags mientras que la gráfica de autocorrelación parcial PACF muestra picos muy elevados en los primeros dos lags, lo que da un indicio que un modelo AR se puede ajustar muy bien.

Las ventanas de tiempo de 100 periodos para graficas la media móvil y la varianza móvil muestran igualmente que la serie no es estacionaria.

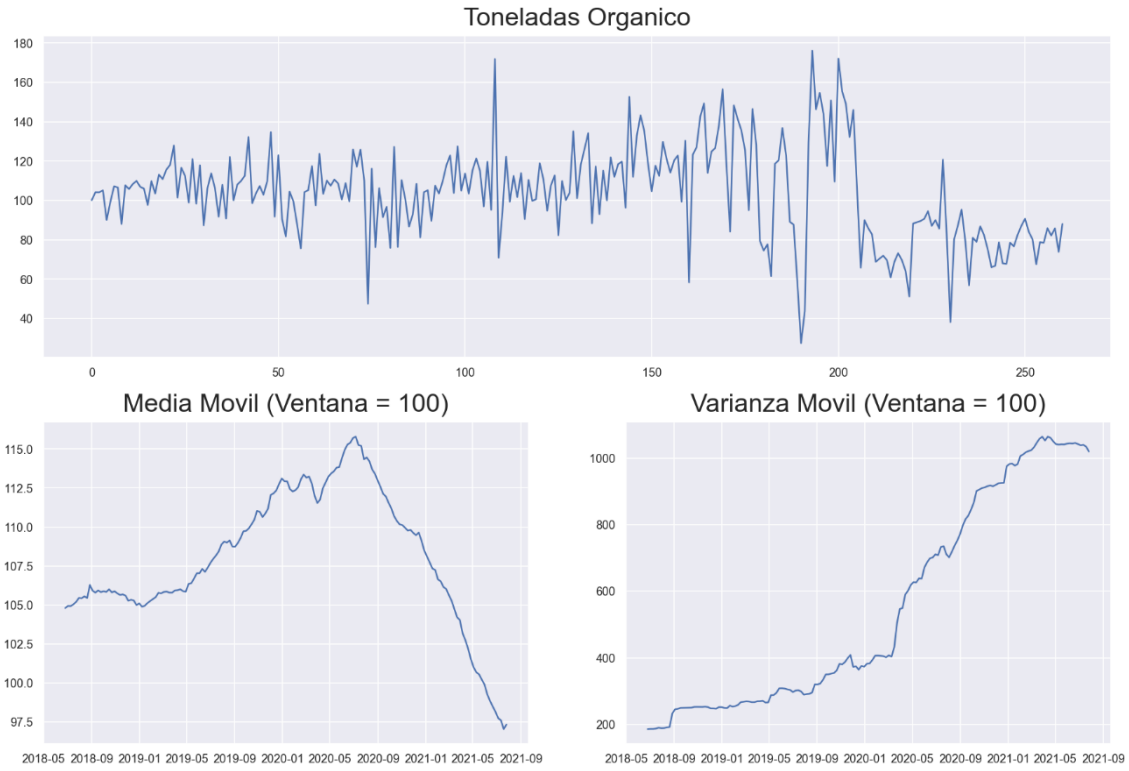

Ilustración 19. Media Movil y Varianza Movil de la serie de tiempo 'ORGANICO'.

Para estacionarizar la serie se aplicó un logaritmo seguido de una diferenciación, que en el mundo de los logaritmos es equivalente a una división:

$$y_{nueva_t} = \log\left(\frac{y_t}{y_{t-1}}\right)$$

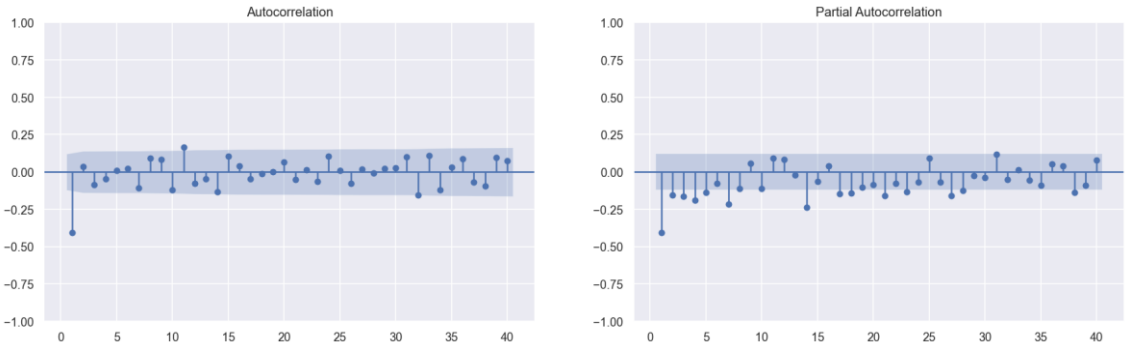

Ilustración 20. Autocorrelación y Autocorrelación Parcial de la serie de tiempo una vez aplicada la transformación.

Al aplicarle un test de estacionaridad a esta nueva serie de datos entrega un valor-p de 1.5e-07 lo que permite rechazar la hipótesis nula, y ahora si afirmar que la serie está estacionarizada.

ARIMA

Gracias a la herramienta AutoArima, se pudieron evaluar varios escenarios y determinar cual tenía los mejores hiperparámetros para iniciar con el diseño del ARIMA.

Tabla 7. Resultados del AutoArima

```
ARIMA(2,0,2)(1,0,1)[12] intercept   : AIC=inf, Time=0.91 sec
ARIMA(0,0,0)(0,0,0)[12] intercept   : AIC=-332.783, Time=0.08 sec
ARIMA(1,0,0)(1,0,0)[12] intercept   : AIC=-376.090, Time=0.53 sec
ARIMA(0,0,1)(0,0,1)[12] intercept   : AIC=-402.192, Time=0.23 sec
ARIMA(0,0,0)(0,0,0)[12]             : AIC=343.425, Time=0.04 sec
ARIMA(0,0,1)(0,0,0)[12] intercept   : AIC=-400.799, Time=0.19 sec
ARIMA(0,0,1)(1,0,1)[12] intercept   : AIC=inf, Time=0.62 sec
ARIMA(0,0,1)(0,0,2)[12] intercept   : AIC=-403.491, Time=0.64 sec
ARIMA(0,0,1)(1,0,2)[12] intercept   : AIC=inf, Time=1.28 sec
ARIMA(0,0,0)(0,0,2)[12] intercept   : AIC=-332.724, Time=0.32 sec
ARIMA(1,0,1)(0,0,2)[12] intercept   : AIC=-416.776, Time=1.09 sec
ARIMA(1,0,1)(0,0,1)[12] intercept   : AIC=-414.625, Time=0.73 sec
ARIMA(1,0,1)(1,0,2)[12] intercept   : AIC=inf, Time=1.26 sec
ARIMA(1,0,1)(1,0,1)[12] intercept   : AIC=inf, Time=0.72 sec
ARIMA(1,0,0)(0,0,2)[12] intercept   : AIC=-377.444, Time=0.55 sec
ARIMA(2,0,1)(0,0,2)[12] intercept   : AIC=-406.422, Time=0.96 sec
ARIMA(1,0,2)(0,0,2)[12] intercept   : AIC=-399.839, Time=1.03 sec
ARIMA(0,0,2)(0,0,2)[12] intercept   : AIC=-411.809, Time=0.65 sec
ARIMA(2,0,0)(0,0,2)[12] intercept   : AIC=-382.495, Time=0.61 sec
ARIMA(2,0,2)(0,0,2)[12] intercept   : AIC=inf, Time=1.65 sec
ARIMA(1,0,1)(0,0,2)[12]             : AIC=inf, Time=0.75 sec

Best model:  ARIMA(1,0,1)(0,0,2)[12] intercept
```

Tabla 8. Resumen del ARIMA una vez establecidos los parámetros.

Dep. Variable:	ORGANICO_log_diff	**No. Observations:**	182
Model:	SARIMAX(1, 0, 1)x(0, 0, [1, 2], 12)	**Log Likelihood**	64.591
Date:	Sat, 08 Oct 2022	**AIC**	-119.181
Time:	13:14:39	**BIC**	-103.161
Sample:	08-08-2016	**HQIC**	-112.687
	- 01-27-2020		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.2101	0.085	-2.460	0.014	-0.378	-0.043
ma.L1	-0.7402	0.056	-13.234	0.000	-0.850	-0.631
ma.S.L12	-0.1195	0.103	-1.161	0.246	-0.321	0.082
ma.S.L24	0.0985	0.086	1.141	0.254	-0.071	0.268
sigma2	0.0286	0.002	17.640	0.000	0.025	0.032

Ljung-Box (L1) (Q):	0.01	**Jarque-Bera (JB):**	255.23
Prob(Q):	0.93	**Prob(JB):**	0.00
Heteroskedasticity (H):	3.31	**Skew:**	-1.47
Prob(H) (two-sided):	0.00	**Kurtosis:**	8.00

El enfoque de *Machine Learning* para el entrenamiento de series de tiempo presenta tres alternativas diferentes:

- Entrena una vez – Predice una vez
- Entrena una vez – Predicciones Continuas
- Entrenamiento Continuo – Predicciones Continuas

La diferencia entre ellas radica en si cada predicción generada a lo largo del tiempo es utilizada para un nuevo entrenamiento o no.

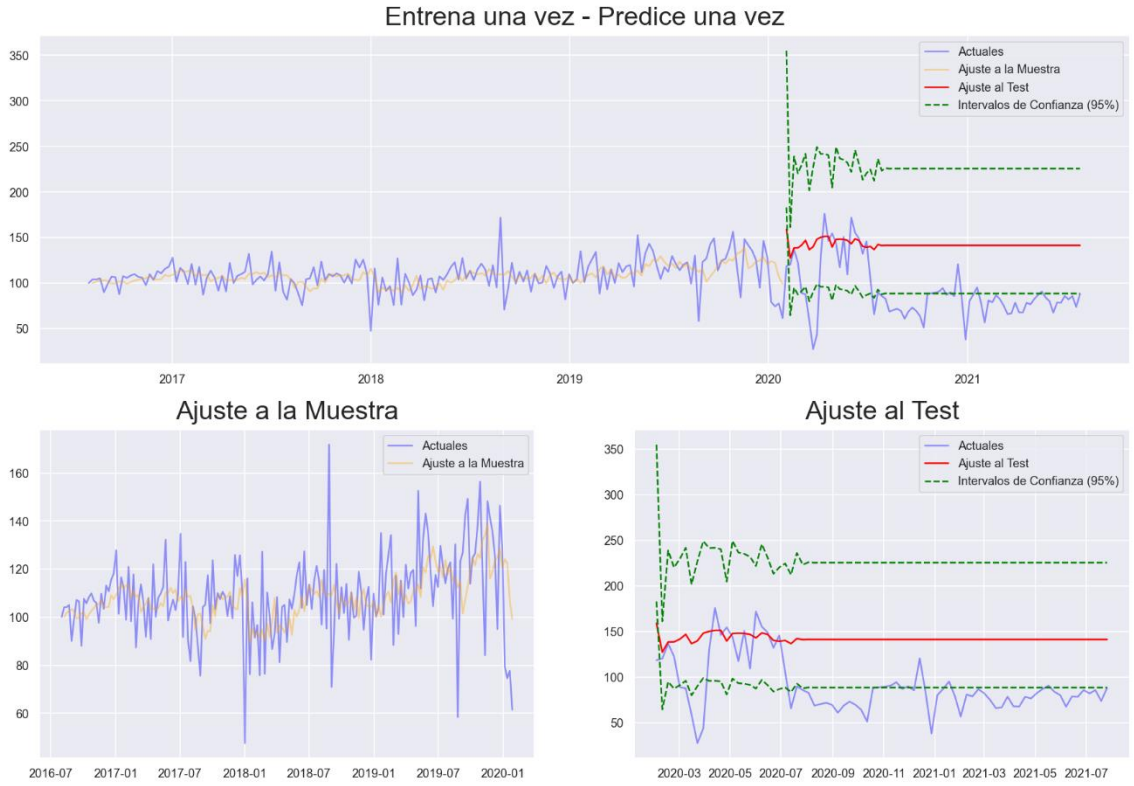

Ilustración 21. Resultados gráficos del ARIMA (Entrena una vez - Predice una vez)

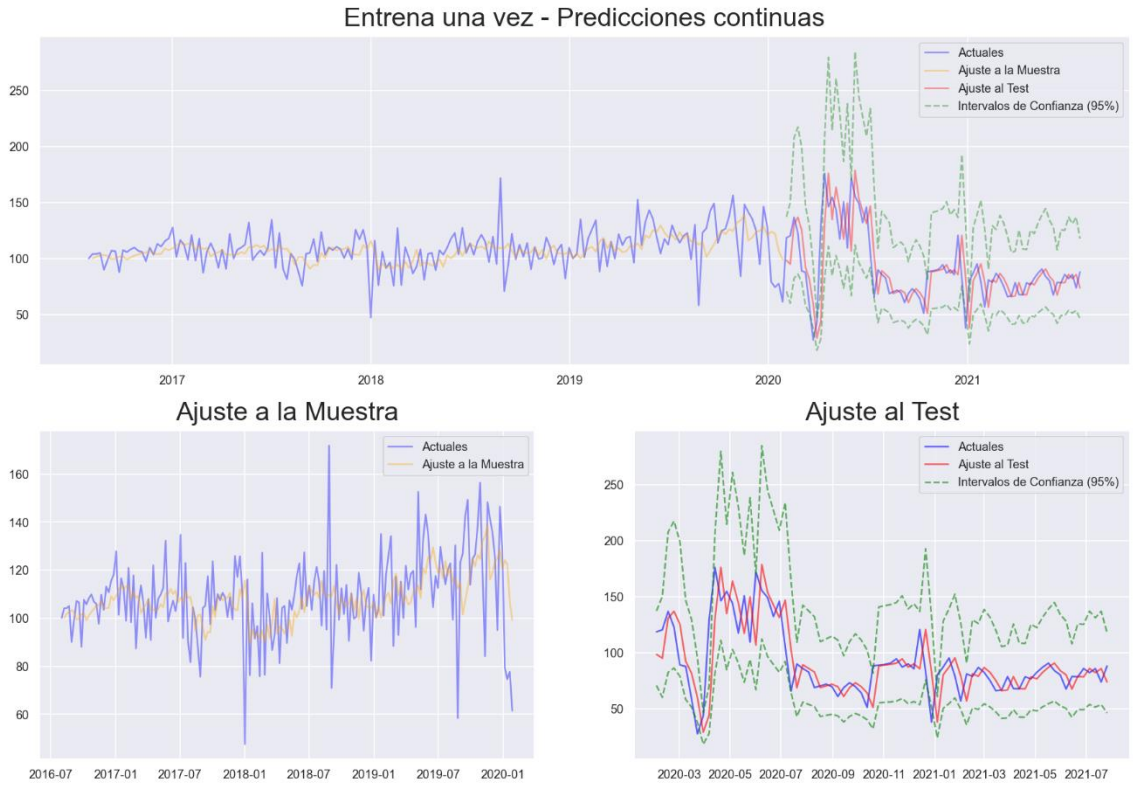

Ilustración 22. Resultados Gráficos del ARIMA (Entrena una vez - Predicciones Continuas)

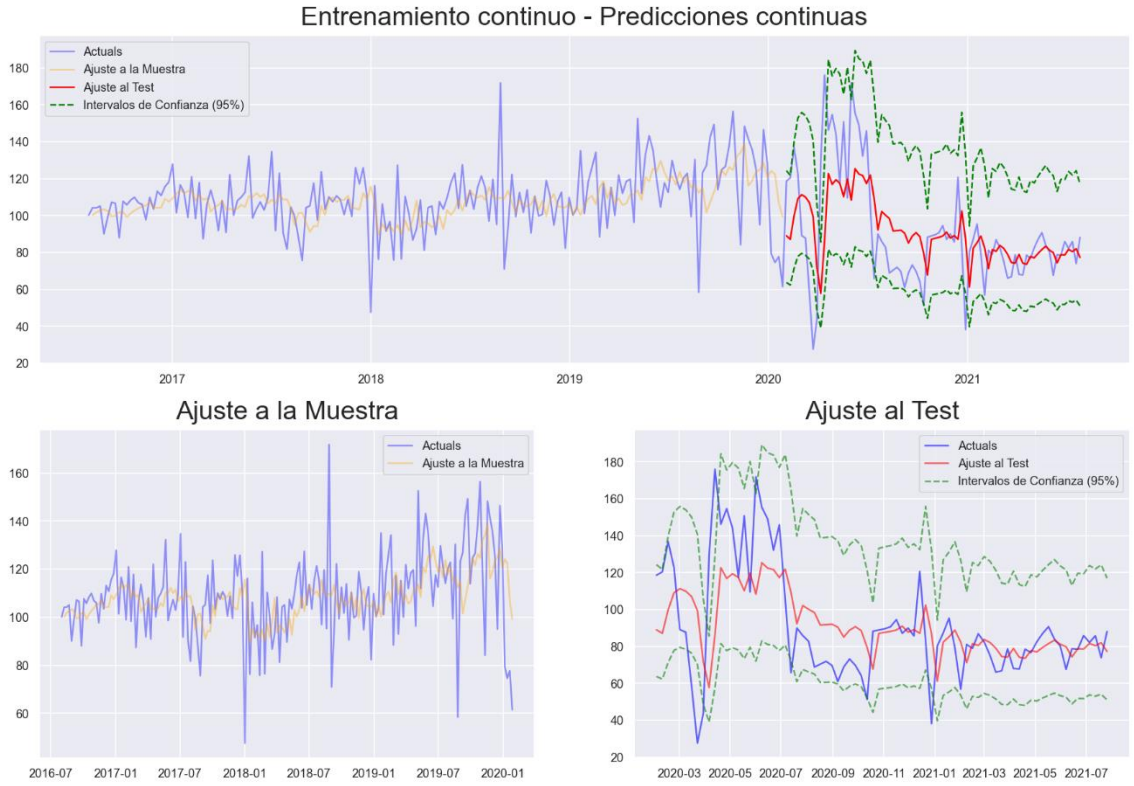

Ilustración 23. Resultados gráficos del ARIMA (Entrenamiento Continuo - Predicciones Continuas)

Tabla 9. Métricas de los modelos ARIMA (RMSE)

Tipo	Ajuste a la Muestra	Ajuste al Test
Entrena una vez - Predice una vez	17.78	58.71
Entrena una vez - Predicciones continuas	17.18	22.34
Entrenamiento continuo - Predicciones continuas	17.18	25.81

En los 3 casos, se puede apreciar el desempeño del modelo en términos de predicción para los datos de entrenamiento 70%, así como para los datos de prueba 30%. El entrenamiento, ilustrado en el "Ajuste a la Muestra" muestra la comparación entre la serie de tiempo real y las predicciones generadas durante este periodo de aprendizaje. Por otro lado, en el "Ajuste al Test", donde se pone a prueba el modelo, se encuentran dos elementos muy interesantes. En primer, la predicción para cada periodo de tiempo, y segundo, pero no menos importante, los intervalos de confianza al 95%, que es equivalente a tener un Alpha del 5%. Esto se puede interpretar que, con una confianza del 95% el valor de la serie de tiempo para una fecha específica va a estar dentro de ese rango. Y si se analiza gráficamente, se puede observar que el comportamiento actual, el real, se comportó en su mayoría dentro de estos intervalos de confianza, con la excepción de unos movimientos muy bruscos, que le cuesta al modelo predecir, tales como los del tiempo de COVID, que mencionaba la empresa desde la etapa de "Comprensión del Negocio".

Modelos Univariados

Todos las pruebas y resultados que se han obtenido hasta este momento permiten inferir que un modelo univariado es el que mejor se puede ajustar para generar las predicciones más acertadas en esta serie de datos. Si bien ARIMA ya arrojó algunos resultados de predicción, es necesario implementar otros modelos para evaluar cual obtiene las predicciones más acertadas o si se presentan relaciones no lineales en las observaciones, también de manera univariada, para lo cual se evaluaron los siguientes modelos:

- Random Forest (rf)
- K-Nearest Neighbors (knn)
- Elastic Net
- XGBoost
- Multiple Linear Regression (mlr)
- Multiple Layer Perceptron (mlp)
- GradientBoost (gbt)

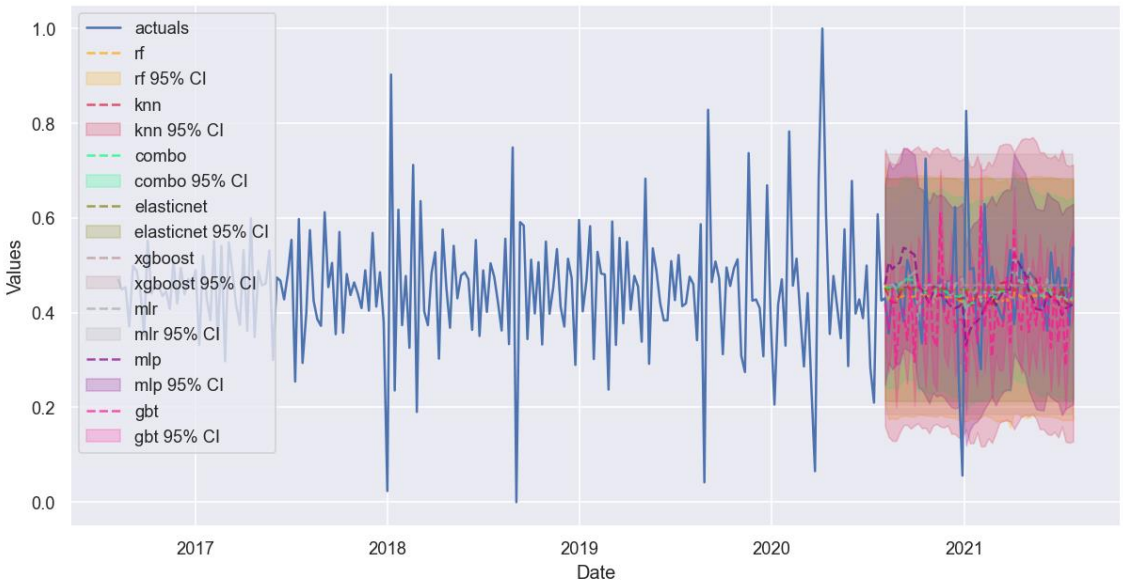

Ilustración 24. Resultados gráficos de los Modelos Univariados.

Tabla 10. Resultados numéricos de los Modelos Univariados.

	ModelNickname	Series	Integration	LevelTestSetRMSE	InSampleRMSE	best_model
0	elasticnet	Organico	0	0.1086	0.1273	True
1	knn	Organico	0	0.1089	0.1252	False
2	xgboost	Organico	0	0.1091	0.1277	False
3	rf	Organico	0	0.1102	0.1129	False
4	combo	Organico	0	0.1123	0.1183	False
5	mlr	Organico	0	0.1167	0.1112	False
6	mlp	Organico	0	0.1246	0.1272	False
7	gbt	Organico	0	0.1578	0.0502	False

Modelos No Lineales Multivariados.

El hecho de que hasta ahora se hayan se hayan implementado solamente modelo univariados se debe a la hipótesis que se ha planteado a partir de la inferencia de los test aplicados. Para verificar esta hipótesis es necesario comparar también evaluar los modelos multivariados, con las variables que se tienen disponibles. Y con los resultados de las métricas, compararlos con las métricas de los modelos univariados para así poder aceptar o rechazar la hipótesis planteada.

El modelo univariado, solo considera la serie de tiempo de la variable que va a predecir, el comportamiento de los valores anteriores de la misma para estimar un valor futuro. Por el contrario, el modelo multivariado toma en cuenta no solo la serie de tiempo de la variable a predecir, sino que además considera los valores previos y actuales de las otras variables. Antes de iniciar a modelar de manera multivariada, se aplicó un análisis de correlación de la variable dependiente ‘Orgánico’ con los 13 valores previos de las variables las variables ‘Inorgánico’, ‘Reciclaje’ y ‘Habitantes’.

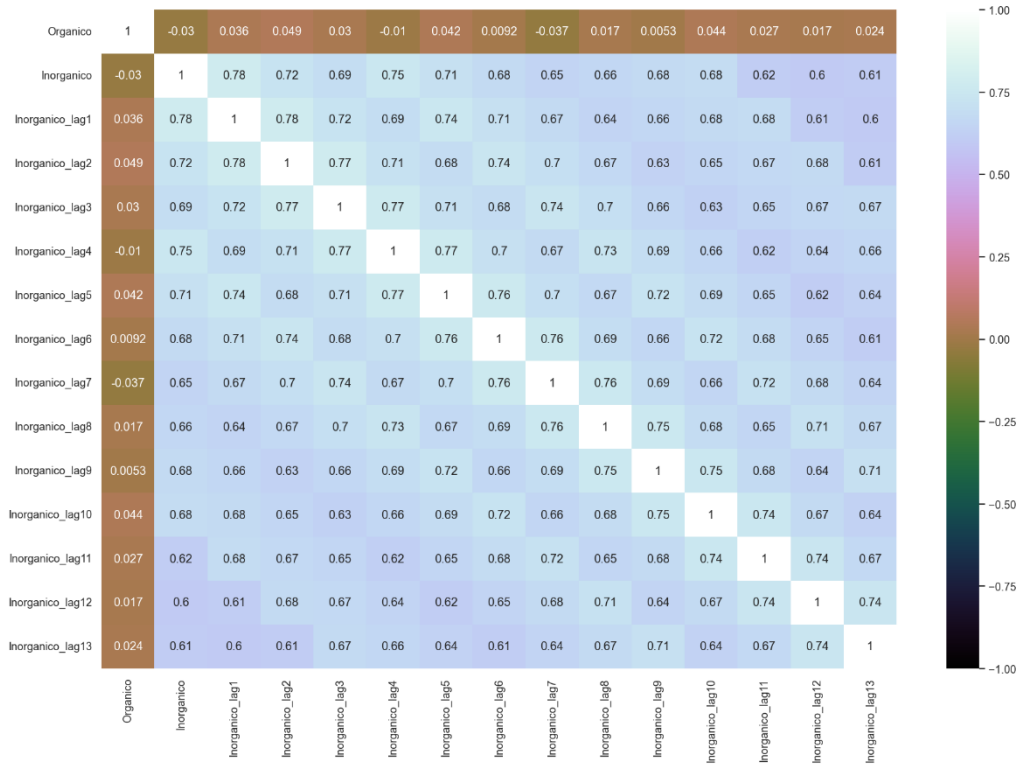

Ilustración 25. Autocorrelación de la variable 'ORGANICO' con los 13 valores previos de la variable 'INORGANICO'

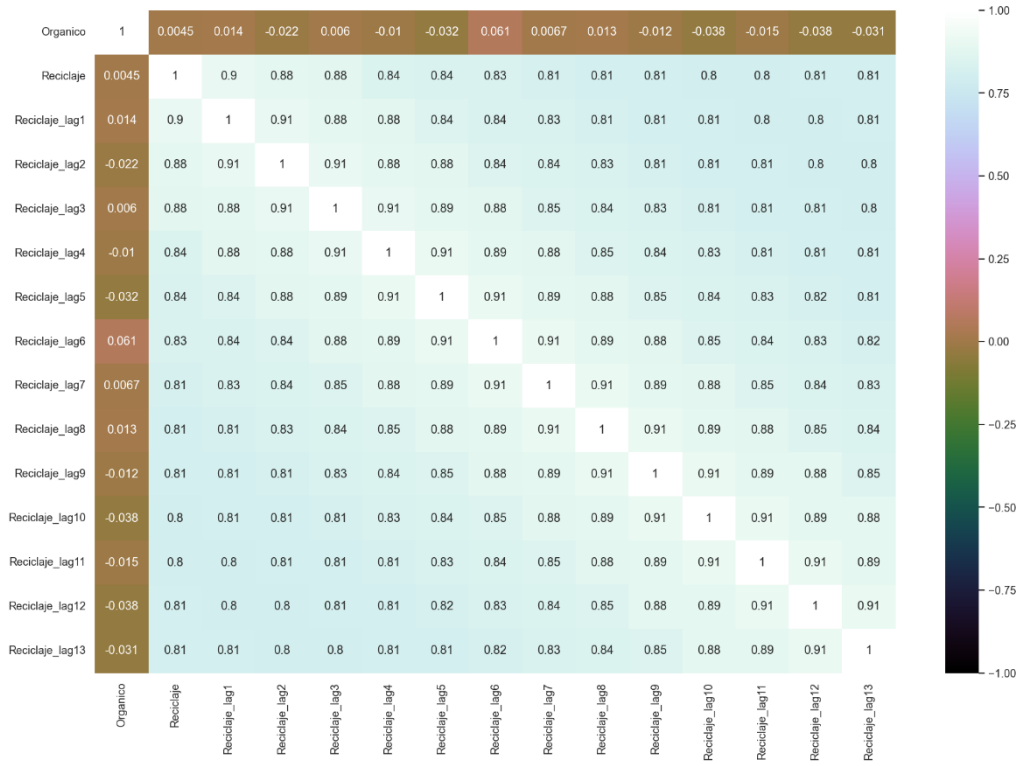

Ilustración 26. Autocorrelación de la variable 'ORGANICO' con los 13 valores previos de la variable 'RECICLAJE'

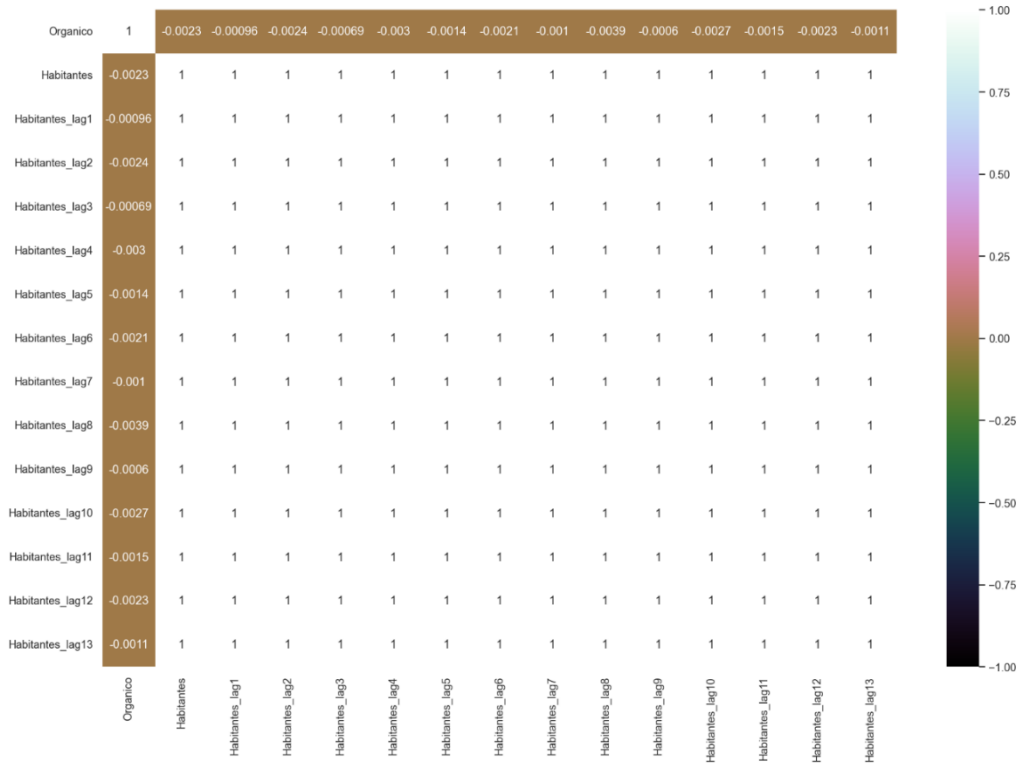

Ilustración 27. Autocorrelación de la variable 'ORGANICO' con los 13 valores previos de la variable 'Habitantes'.

El color marrón simboliza una correlación cercana a cero, mientras que el blanco se refiere a una correlación perfecta positiva y el negro a una correlación perfecta negativa. En los 3 casos, se puede apreciar que, gráficamente hablando, prácticamente, la correlación de la variable dependiente 'Orgánico' con los valores previos de las otras 3 variables es nula.

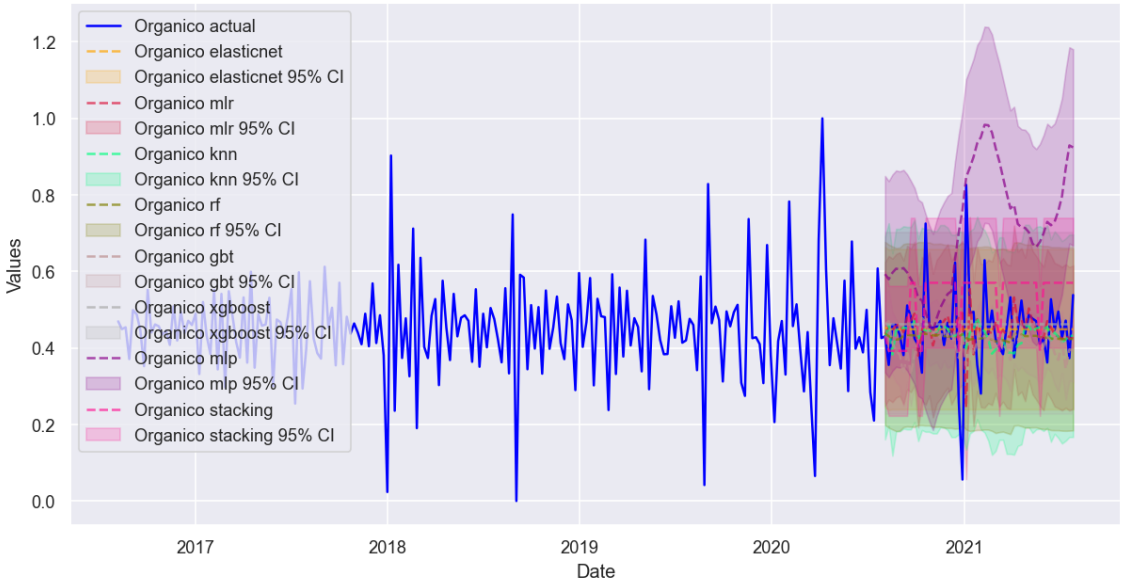

Ilustración 28. Resultados gráficos de los Modelos Multivariados.

Tabla 11. Resultados numéricos de los Modelos Multivariados.

ModelNickname	Series	LevelTestSetRMSE	InSampleRMSE	Lags	best_model
elasticnet	Organico	0.1086	0.1194	3	True
mlr	Organico	0.1322	0.1113	1	False
knn	Organico	0.1168	0.1258	3	False
rf	Organico	0.1116	0.1129	1	False
gbt	Organico	0.1249	0.0419	6	False
xgboost	Organico	0.1095	0.1275	1	False
mlp	Organico	0.3213	0.1389	1	False
stacking	Organico	0.1414	0.1014	13	False

Evaluación

Se decidió utilizar la métrica del error cuadrático medio (RMSE) para determinar cual es el modelo que mejor se puede ajustar a la predicción. Esto debido a que dicha métrica le puede brindar a la empresa una idea cercana de la producción de desechos en toneladas semanales, puesto que se encuentra en la misma escala de la variable analizada. Se tomará aquel modelo que tenga el menor RMSE en el ajuste al test.

Tabla 12. RMSE de todos los modelos aplicados.

Tipo	Ajuste a la Muestra	Ajuste al Test
Entrena una vez - Predice una vez	17.78	58.71
Entrena una vez - Predicciones continuas	17.18	22.34
Entrenamiento continuo - Predicciones continuas	17.18	25.81
Elasticnet univariado	0.1273	0.1086
KNN univariado	0.1252	0.1089
XGBoost univariado	0.1277	0.1091
RandomForest univariado	0.1129	0.1102
combo univariado	0.1183	0.1123
Multiple Linear Regression univariado	0.1112	0.1167
Multiple Layer Perceptron univariado	0.1272	0.1246
Gradient Boost univariado	0.0502	0.1578
Elasticnet multivariado	0.1194	0.1086
Multiple Linear Regression multivariado	0.1113	0.1322
KNN multivariado	0.1258	0.1168
RandomForest multivariado	0.1129	0.1116
Gradient Boost multivariado	0.0419	0.1249
XGBoost multivariado	0.1275	0.1095
Multiple Layer Perceptron multivariado	0.1389	0.3213
Stacking Multivariado	0.1014	0.1414

Resultados y Discusión

De los primero 3 modelos, “Entrena una vez - Predicciones continuas” tiene el mejor RMSE con un valor de 22.34 toneladas semanales. Del resto de modelos, el cual el RMSE está escalado, el mejor resultado lo tiene “Elasticnet univariado” con un valor de 0.1086, que si se desescala, corresponde a 43.7 toneladas semanales.

También es relevante mostrar que algunos modelos como “Entrena una vez - Predice una vez” presentan sobreajuste, al tener un RMSE en la muestra muy bajo y en el test muy alto.

En cuanto a la hipótesis que se ha planteado a lo largo de todo el documento, que un modelo univariado puede presentar mejores predicciones que un modelo multivariado para esta serie de tiempo, se acepta la hipótesis. Esto a debido a que lose errores cuadráticos medios de los modelos multivariados fueron iguales o mayores a los de los modelos univariados en el “Ajuste al Test”.

Conclusiones

Para la serie de tiempo de la recolección de residuos orgánicos con un muestreo semanal, se concluye a partir de las pruebas efectuadas que, los modelos univariados se ajustan a la necesidad de predicción.

Entre todos los modelos evaluados, el ARIMA en su variante “Entrena una vez - Predicciones continuas” presentó el mejor desempeño con un RMSE de 22.34 toneladas semanales. Este es un valor aceptado por la empresa y permitirá darles una predicción semanal de la cantidad de residuos orgánicos a tratar.

Es tipo de análisis son importantes, tanto para las empresas industriales de carácter público, y la ciudadanía en general, ya que permiten optimizar los recursos, provenientes de los impuestos pagados por los contribuyentes, llevándolas a ser eficientes y competitivas, sin malgastar los dineros públicos y causando un impacto social acorde a las necesidades de la ciudadanía.

Derivado de lo anterior, los resultados obtenidos en este trabajo son de suma relevancia ya que servirán, en primera instancia, como garantía para los diferentes inversionistas del proyecto que los recursos destinados cumplirán con su misión, y por el otro lado, pero no menos importante, se mitigará el impacto ambiental que hoy generan estos residuos.

Es importante traer a colación que, si se destinan más recursos de los necesarios para tratar estos residuos orgánicos, en cuestión de materia prima y mano de obra, se estarán despilfarrando los recursos públicos. Por el contrario, si no destinan los recursos necesarios, los residuos no tratados tendrán como destino final un relleno sanitario, causando un impacto ambiental enorme, sobre todo a las fuentes hídricas de las cuales depende el municipio.

Para trabajos futuros, se recomienda abordar con una mayor anticipación la generación de datos desde su origen, y generar los cambios necesarios para garantizar que sean tan limpios como sea posible, y de la misma manera que sean aptos para generar las transformaciones necesarias. También se deja constancia que es muy importante estar en permanente contacto con el sponsor del proyecto, la empresa, e ir analizando con cierta frecuencia los avances del proyecto, ya que ellos como dueños de los datos y conocedores del proceso pueden brindar visiones, inferencias o correcciones necesarias que llevan finalmente a unos resultados útiles.

Referencias Bibliográficas

Abbasi, M., & El Hanandeh, A. (2016). Forecasting municipal solid waste generation using artificial intelligence modelling approaches. Waste management, 56, 13-22.

Abbasi, M., Rastgoo, M. N., & Nakisa, B. (2019). Monthly and seasonal modeling of municipal waste generation using radial basis function neural network. Environmental Progress & Sustainable Energy, 38(3), e13033.

Ayeleru, O. O., Fajimi, L. I., Oboirien, B. O., & Olubambi, P. A. (2021). Forecasting municipal solid waste quantity using artificial neural network and supported vector machine techniques: A case study of Johannesburg, South Africa. Journal of Cleaner Production, 289, 125671.

Ceylan, Z. (2020). Estimation of municipal waste generation of Turkey using socio-economic indicators by Bayesian optimization tuned Gaussian process regression. Waste Management & Research, 38(8), 840-850.

Dissanayaka, D. M. S. H., & Vasanthapriyan, S. (2019, December). Forecast municipal solid waste generation in Sri Lanka. In 2019 International Conference on Advancements in Computing (ICAC) (pp. 210-215). IEEE.

Fan, L., Abbasi, M., Salehi, K., Band, S. S., Chau, K. W., & Mosavi, A. (2021). Introducing an evolutionary-decomposition model for prediction of municipal solid waste flow: application of intrinsic time-scale decomposition algorithm. Engineering Applications of Computational Fluid Mechanics, 15(1), 1159-1175.

Fasano, F., Addante, A. S., Valenzano, B., & Scannicchio, G. (2021). Variables Influencing per Capita Production, Separate Collection, and Costs of Municipal Solid Waste in the Apulia Region (Italy): An Experience of Deep Learning. International Journal of Environmental Research and Public Health, 18(2), 752.

Ghanbari, F., Kamalan, H., & Sarraf, A. (2021). An evolutionary machine learning approach for municipal solid waste generation estimation utilizing socioeconomic components. Arabian Journal of Geosciences, 14(2), 1-16.

Guo, H. N., Wu, S. B., Tian, Y. J., Zhang, J., & Liu, H. T. (2021). Application of machine learning methods for the prediction of organic solid waste treatment and recycling processes: A review. Bioresource Technology, 319, 124114.

Huang, J., & Koroteev, D. D. (2021). Artificial intelligence for planning of energy and waste management. Sustainable Energy Technologies and Assessments, 47, 101426.

Kannangara, M., Dua, R., Ahmadi, L., & Bensebaa, F. (2018). Modeling and prediction of regional municipal solid waste generation and diversion in Canada using machine learning approaches. Waste Management, 74, 3-15.

Kontokosta, C. E., Hong, B., Johnson, N. E., & Starobin, D. (2018). Using machine learning and small area estimation to predict building-level municipal solid waste generation in cities. Computers, Environment and Urban Systems, 70, 151-162.

Lin, K., Zhao, Y., Tian, L., Zhao, C., Zhang, M., & Zhou, T. (2021). Estimation of municipal solid waste amount based on one-dimension convolutional neural network and long short-

term memory with attention mechanism model: A case study of Shanghai. Science of The Total Environment, 791, 148088.

Magazzino, C., Mele, M., Schneider, N., & Sarkodie, S. A. (2021). Waste generation, wealth and GHG emissions from the waste sector: Is Denmark on the path towards circular economy?. Science of the Total Environment, 755, 142510.

Nguyen, X. C., Nguyen, T. T. H., La, D. D., Kumar, G., Rene, E. R., Nguyen, D. D., ... & Nguyen, V. K. (2021). Development of machine learning-based models to forecast solid waste generation in residential areas: A case study from Vietnam. Resources, Conservation and Recycling, 167, 105381.

Rosecký, M., Šomplák, R., Slavík, J., Kalina, J., Bulková, G., & Bednář, J. (2021). Predictive modelling as a tool for effective municipal waste management policy at different territorial levels. Journal of Environmental Management, 291, 112584.

Índice de Tablas e Ilustraciones

Anexos

- Jupyter Notebook “Predicción de la generación municipal de residuos orgánicos. Una aproximación desde el aprendizaje de máquina.”
- OutputSemanal.jpeg: Resultado de la herramienta Pandas Profiling que contiene datos útiles para el análisis exploratorio de datos
- OutputSemanalEscalado.: Resultado de la herramienta Pandas Profiling que contiene datos escalados útiles para el análisis exploratorio de datos.