Estimación de áreas a cultivar en Colombia

Jesus David Gomez Osorno

Trabajo de grado presentado para optar al título de

Especialista en Analítica y Ciencia de Datos

Asesor

Javier F. Botía,

Doctor (PhD)

Universidad de Antioquia

Facultad de Ingeniería

Especialización en Analítica y Ciencia de Datos

Medellín, Antioquia, Colombia

2022

Cita	Gomez Osorno [1]
Referencia Estilo IEEE (2020)	[1] J. D. Gomez Osorno, “Estimación de áreas a cultivar en Colombia”, Trabajo de grado especialización, Especialización en Analítica y Ciencia de Datos, Universidad de Antioquia, Medellín, Antioquia, Colombia, 2022.

Especialización en Analítica y Ciencia de Datos, Cohorte III.

Universidad de Antioquia - www.udea.edu.co

Rector: John Jairo Arboleda Céspedes.

Decano/Director: Jesús Francisco Vargas Bonilla.

Jefe departamento: Diego José Luis Botia Valderrama.

Agradecimientos

Al tutor y profesor Javier F. Botía.

TABLA DE CONTENIDO

LISTA DE TABLAS

LISTA DE FIGURAS

SIGLAS, ACRÓNIMOS Y ABREVIATURAS

MAE.	Media del error cuadrado
ISF.	Aislamiento forestal
LOF.	Factor atípico local
HGB	Histogram-based Gradient Boosting Regression Tree
MLP	Multi-layer Perceptron
R2.	Coeficiente de determinación (R cuadrado)

1. RESUMEN EJECUTIVO

En esta monografía, se centra en el escenario de diferentes cultivos entre el año 2007 y 2018 que está organizado en una base de datos del Ministerio de Agricultura y Desarrollo Rural, el cual son datos abiertos del Gobierno Colombiano. El principal problema que se enfoca en resolver la monografía es encontrar un modelo predictivo capaz de estimar la cantidad de área necesaria en un cultivo específico para lograr una cosecha objetivo. Debido a la alta dimensionalidad de los datos, se propone una estrategia para dividir los datos de acuerdo con una recomendación de Asobancaria, el cual menciona: ***"Respecto al tamaño de la tierra para la Food and Agriculture Organization - FAO (2012), la clasificación varía a lo largo de zonas geográficas y de producción - los pequeños productores son aquellos con -- menos de 1 hectárea productiva, los medianos rondan de 1 a 10 hectáreas y los grandes poseen más de 10 hectáreas" [0].*** Por consiguiente, al considerar que las hectáreas en la base de datos se manejan como un dato entero, no es posible trabajar con los pequeños agricultores. Lo anterior permitió generar dos bases de datos nuevos que representan los medianos (1/4 de los datos originales) y grandes productores (3/4 de los datos originales). A partir de la división de los datos, se realizó una exploración de datos para generar la mejor representación de las bases de datos antes de crear los modelos de regresión. Aplicando la estrategia de división de datos, se generaron mejores resultados, para la base de datos de medianos agricultores se logró un **MAE** de 0.001335 con un modelo de regresión de Huber, y para los grandes agricultores se logró un **MAE** de 0.001003 con el modelo de regresión de bosques aleatorios o random forest.

El repositorio de donde se encuentra el código y la base de datos que se usaron para el desarrollo de este proyecto se encuentra en el siguiente enlace: https://github.com/chuchodavidgomez/estimacion_areas_cultivar_colombia

Palabras clave — **Machine learning, regresión, agricultura, estimación, modelos, cosechas, análisis de datos.**

2. DESCRIPCIÓN DEL PROBLEMA

Actualmente hay pocos modelos que ayuden en la estimación de la cantidad de áreas que se pueden cosechar a partir de cierta cantidad de áreas sembradas en Colombia, debido a que en el país se cuenta múltiples zonas geográficas, y estas cuentan con diferentes aspectos climáticos, económicos geográficos, social, etc. Por lo cual se va a buscar el modelo óptimo para estimar estas áreas objetivo.

Inicialmente se va a tener un modelo general basado en características no tan específicas, pero si diferenciadoras como el departamento y municipio donde se quiere cultivar. Para lograr este fin, se buscará mediante diversos modelos, desde el más simple como regresión lineal hasta unos más complejos como redes neuronales simples. Además, no se tendrán en cuenta todas las características, debido a que la base de datos presenta un grado alto de redundancia y es necesario verificar en cuáles se da este problema.

Finalmente, el objetivo de esta primera investigación es sentar las bases para futuras investigaciones de modelos un poco más complejos, con información un poco más precisa como las precipitaciones y/o contaminación del municipio donde se quiere realizar la siembra.

2.1 PROBLEMA DE NEGOCIO

Colombia como país agricultor, le falta inversión e interés en la búsqueda de modelos estimadores que puedan ayudar tanto a los medianos campesinos como a los grandes campesinos en el cálculo de la cantidad de hectáreas a sembrar para obtener una mejor cosecha y así optimizar el espacio para otro tipo de cultivo que se pueda sembrar en la zona, esto partiendo de un conjunto de datos muy pequeño en características, pero grande en datos.

2.2 APROXIMACIÓN DESDE LA ANALÍTICA DE DATOS

Se creará diversos modelos que podrán ser de gran ayuda para los medianos y grandes campesinos a la hora de decidir la cantidad de áreas a sembrar para un cultivo en específico, es de aclarar, como se comentó anteriormente estos modelos no serán entrenados con datos muy específicos, por lo tanto es posible que haya algún sesgo en los resultados de estos, por lo que se recomienda utilizar los resultados del modelo como complemento a la hora de decidir.

2.3 ORIGEN DE LOS DATOS

La base de datos utilizada en este proyecto fue proporcionada por el Ministerio de Agricultura y Desarrollo Rural, a través de los datos abiertos de la gobernación de Colombia, contiene datos de la producción agrícola nacional, presentando los indicadores de áreas, producción y los rendimientos de los cultivos permanentes, transitorios y anuales entre los años 2007 y 2018.

2.4 MÉTRICAS DE DESEMPEÑO

Coeficiente de determinación o R-cuadrado: Esta medida de desempeño la trae por defecto la librería que se usó durante la investigación, el coeficiente de determinación mide la cantidad de varianza de la predicción, se obtiene de la resta de los cuadrados de nuestro vector Y verdadero y nuestro Y que se predijo. La mejor puntuación posible es 1,0 y puede ser negativa, debido a que el modelo puede ser arbitrariamente peor [3].

Media de la validación: Esta medida nos indica la media que se obtuvo en la validación.

Desviación estándar de la validación: Esta medida nos indica la desviación estándar que se obtuvo en la validación.

Varianza explicada: Mide la varianza entre nuestro vector Y verdadero y nuestro Y que se predijo, se obtiene a través del cuadrado de las desviaciones estándar, al igual que el coeficiente de determinación la mejor puntuación posible es 1,0

Error promedio absoluto (MAE): Esta medida calcula el error promedio absoluto, correspondiente al valor esperado de la pérdida del error absoluto entre la salida predicha y la salida de los datos originales. Esta métrica es la más importante, debido a que dependiendo de esta se selecciona el mejor modelo dependiendo si esta es grande o pequeña.

Pérdida media de regresión de la desviación de Poisson: Si la pérdida es 0, significa que el modelo es eficiente y es más sensible a la hora de predecir una subida o una bajada

Pérdida media de regresión de la desviación Gamma: Si la pérdida es 0, significa que el modelo es eficiente, pero es muy sensible a errores relativos.

3. DATOS

3.1 DATOS ORIGINALES

Originalmente la base de datos cuando se descarga pesa aproximadamente 25,7 MB, además contiene un total de 206068 de filas y 17 características, estas características están distribuidas entre categóricas, numéricas e identificadoras. Inicialmente se cuenta con algunos datos faltantes en algunas características, pero en la eliminación de características determinaremos si es necesario imputar estos

```
Data columns (total 17 columns):
 #   Column                  Non-Null Count   Dtype
---  ------                  --------------   -----
 0   CODIGO_DEPARTAMENTO     206068 non-null  object
 1   DEPARTAMENTO            206068 non-null  object
 2   CODIGO_MUNICIPIO        206068 non-null  object
 3   MUNICIPIO               206067 non-null  object
 4   GRUPO_CULTIVO           206068 non-null  object
 5   SUBGRUPO_CULTIVO        206068 non-null  object
 6   CULTIVO                 206068 non-null  object
 7   SISTEMA_PRODUCTIVO      206068 non-null  object
 8   ANIO                    206068 non-null  object
 9   PERIODO                 206068 non-null  object
 10  AREA_SEMBRADA           206068 non-null  object
 11  AREA_COSECHADA          206068 non-null  object
 12  PRODUCCION              206068 non-null  object
 13  RENDIMIENTO             202635 non-null  object
 14  ESTADO_FISICO_PRODUCCION 206068 non-null  object
 15  NOMBRE_CIENTIFICO       203211 non-null  object
 16  CICLO_CULTIVO           206068 non-null  object
```

Fig. 1. Descripción de las características

A continuación, se explica de forma detallada cada característica:

1. **CODIGO_DEPARTAMENTO** (Número): Código del departamento, según lo establecido por el DANE
2. **DEPARTAMENTO** (Texto simple): Departamento Colombiano
3. **CODIGO_MUNICIPIO** (Texto simple): Departamento Colombiano
4. **MUNICIPIO** (Texto simple): Departamento Colombiano
5. **GRUPO_CULTIVO** (Texto simple): Categoría del cultivo
6. **SUBGRUPO_CULTIVO** (Texto simple): Tipo de cultivo según categoría
7. **CULTIVO** (Texto simple): Nombre del cultivo

8. **SISTEMA_PRODUCTIVO** (Texto simple): Nombre genérico del cultivo
9. **ANIO** (Número): Año de producción
10. **PERIODO** (Texto simple): Periodo medico
11. **AREA_SEMBRADA** (Número): Área sembrada en hectáreas
12. **AREA_COSECHADA** (Número): Área cosechada en hectáreas
13. **PRODUCCION** (Número): Toneladas producidas
14. **RENDIMIENTO** (Número): Rendimiento de la cosecha (PRODUCCION/AREA_COSECHADA)
15. **ESTADO_FISICO_PRODUCCION** (Texto simple): Estado del producto
16. **NOMBRE_CIENTIFICO** (Texto simple): Nombre científico del cultivo
17. **CICLO_CULTIVO** (Texto simple): Ciclo del cultivo en el país

Para el acceso a los datos, se puede acceder por medio de la plataforma de la plataforma de datos abiertos de la gobernación de Antioquia, y no posee restricciones tal como lo dice en su página de términos y condiciones de la página **"El portal web de Datos Abiertos del Ministerio de Tecnologías de la Información y las Comunicaciones www.datos.gov.co (en adelante el Portal) tiene como función principal publicar de manera unificada, todos los datos producidos por las entidades públicas de Colombia, en formato abierto, con el fin de que éstos puedan ser usados de forma libre y sin restricciones por cualquier persona para desarrollar aplicaciones o servicios de valor agregado, hacer análisis e investigación, ejercer labores de control o para cualquier tipo de actividad comercial o no comercial." [1]**

3.2 CONJUNTO DE DATOS

3.2.1 ELIMINACIÓN DE CARACTERÍSTICAS

Luego de analizar la descripción de cada característica, se procedió a eliminar las características que generaban redundancia en la base de datos:

- **CODIGO_DEPARTAMENTO**: Esta característica es un identificador de la característica **DEPARTAMENTO**.
- **CODIGO_MUNICIPIO**: Esta característica es un identificador de la característica **MUNICIPIO**.

- **PERIODO**: Esta característica nos indica el periodo en el año, pero realizando una breve observación de los datos esta viene igual que la característica de año.
- **NOMBRE_CIENTIFICO**: Esta característica nos indica el nombre de cultivo, pero científico, ya está información ya nos la característica **CULTIVO**.
- **RENDIMIENTO**: Esta característica nos indica la cantidad de hectáreas obtenidas por área cosecha, es una operación entre dos columnas por lo tanto para evitar redundancia en la base de datos, se eliminará.

Lo que puede observar es que estas características no nos sirven para entrenar nuestros modelos, pero si es posible que se puedan agregar más características usando estos identificadores para relacionarla con otras bases de datos.

Adicionalmente, hubo 3 características **CULTIVO** con **SUBGRUPO_CULTIVO** y **SISTEMA_PRODUCTIVO** que posiblemente tenían redundancia, debido a que, durante el análisis, se identificó que estas características en alguno de sus datos se repetía el valor en estas características, por lo tanto, se procedió a realizar una prueba de hipótesis para determinar si estas características tenían una alta correlación. Para esto se usó un test chi-cuadrado entre la característica CULTIVO con las otras 2 características mencionadas anteriormente.

- Relación entre **CULTIVO** y **SUBGRUPO_CULTIVO**

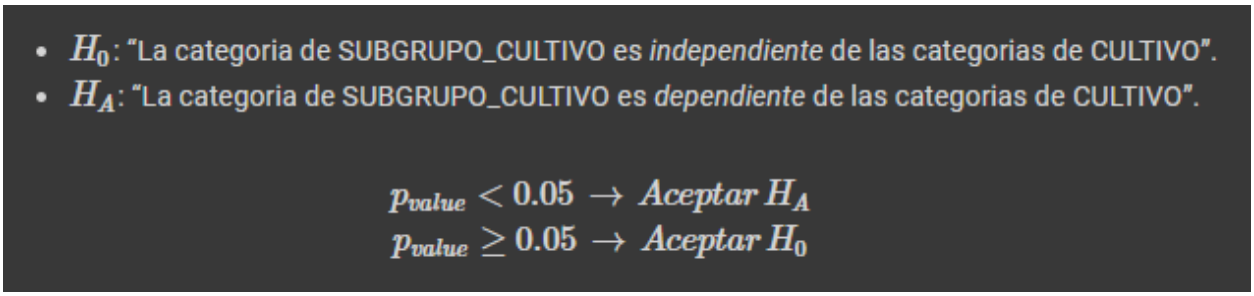

Fig. 2. Prueba de hipótesis entre CULTIVO y SUBGRUPO_CULTIVO

Considerando que todos casos de la tabla de resultados de la prueba chi-cuadrado generan un valor p igual o cercano 0, hay evidencias para rechazar **H0** y afirmar que las categorías de **SUBGRUPO_CULTIVO** son dependientes de las categorías de **CULTIVO**

- Relación entre **CULTIVO** y **SISTEMA_PRODUCTIVO**

- H_0: "La categoria de SISTEMA_PRODUCTIVO es *independiente* de las categorias de CULTIVO".
- H_A: "La categoria de SISTEMA_PRODUCTIVO es *dependiente* de las categorias de CULTIVO".

$$p_{value} < 0.05 \rightarrow Aceptar\ H_A$$
$$p_{value} \geq 0.05 \rightarrow Aceptar\ H_0$$

Fig. 3. Prueba de hipótesis entre **CULTIVO** y **SISTEMA_PRODUCTIVO**

Considerando que la mayoría de los casos de la tabla de resultados de la prueba chi-cuadrado generan un valor p igual o cercano 0, hay evidencias para rechazar **H0** y afirmar que las categorías de **SISTEMA_PRODUCTIVO** son dependientes de las categorías de **CULTIVO.**

Luego de los anteriores procedimientos, nuestra base de datos quedo con 9 características de las 17 que teníamos inicialmente, quedando configurada de la siguiente forma:

```
Data columns (total 9 columns):
 #   Column                  Non-Null Count   Dtype
---  ------                  --------------   -----
 0   DEPARTAMENTO            206068 non-null  object
 1   MUNICIPIO               206067 non-null  object
 2   GRUPO_CULTIVO           206068 non-null  object
 3   CULTIVO                 206068 non-null  object
 4   AREA_SEMBRADA           206068 non-null  object
 5   AREA_COSECHADA          206068 non-null  object
 6   PRODUCCION              206068 non-null  object
 7   ESTADO_FISICO_PRODUCCION 206068 non-null  object
 8   CICLO_CULTIVO           206068 non-null  object
```

Fig. 4. Descripción características seleccionadas

Como se observa hay un dato faltante en la característica **MUNICIPIO,** por lo tanto, se va intentar llenar este dato en posteriores procesos, primero se observará si el registro que contiene este dato faltante se retire en el apartado de eliminación de datos.

3.2.1 ELIMINACIÓN DE DATOS

Como nuestro objetivo es predecir la cantidad de áreas a sembrar y como se observa en esta característica se tiene registros donde la cantidad de áreas a sembrar son 0, esto debido a que como se tienen datos enteros no se puede manejar las fracciones, por lo tanto, se procede a eliminar estos,

además me mencionó al inicio que solo tendremos 2 tipos de agricultores, los que son los medianos agricultores y grandes agricultores.

Una vez que se eliminaron estos registros, la base de datos quedó con 205347 filas de las 206068 que se tenían, por lo tanto, se eliminaron 721 registros de la base de datos, lo cual no representa una cifra significativa a comparación de los registros que se tenían iniciales.

```
Data columns (total 9 columns):
 #   Column                  Non-Null Count   Dtype
---  ------                  --------------   -----
 0   DEPARTAMENTO            205347 non-null  object
 1   MUNICIPIO               205346 non-null  object
 2   GRUPO_CULTIVO           205347 non-null  object
 3   CULTIVO                 205347 non-null  object
 4   AREA_SEMBRADA           205347 non-null  float32
 5   AREA_COSECHADA          205347 non-null  float32
 6   PRODUCCION              205347 non-null  float32
 7   ESTADO_FISICO_PRODUCCION 205347 non-null  object
 8   CICLO_CULTIVO           205347 non-null  object
```

Fig. 5 Descripción característica con datos eliminados

Como se observa en la tabla si se redujo la cantidad de registros, pero aún permanece ese dato vacío en la característica **MUNICIPIO,** por lo tanto, se procederá a la imputación de este dato.

3.2.2 IMPUTACIÓN DE DATOS

Una vez que se observó este registro que contenía este dato faltante, se determinó que este contenía el dato de la característica **CODIGO_MUNICIPIO,** por lo que procedió a buscar qué municipio tenía este id, una vez se identificado, se observó que correspondía a **BAJO BAUDO**, por lo tanto, se procedió a rellenar este dato en el campo faltante, finalmente nuestra base de datos quedó sin datos faltantes.

```
Data columns (total 9 columns):
 #   Column                  Non-Null Count   Dtype
---  ------                  --------------   -----
 0   DEPARTAMENTO            205347 non-null  object
 1   MUNICIPIO               205347 non-null  object
 2   GRUPO_CULTIVO           205347 non-null  object
 3   CULTIVO                 205347 non-null  object
 4   AREA_SEMBRADA           205347 non-null  float32
 5   AREA_COSECHADA          205347 non-null  float32
 6   PRODUCCION              205347 non-null  float32
 7   ESTADO_FISICO_PRODUCCION 205347 non-null  object
 8   CICLO_CULTIVO           205347 non-null  object
```

Fig. 6. Descripción característica con datos imputados

3.3 DESCRIPTIVA

Hasta el momento nuestra base de datos luce de la siguiente manera:

TABLA I
DATOS DE LA BASE DE DATOS

	DEPARTAMENTO	MUNICIPIO	GRUPO_CULTIVO	CULTIVO	AREA_SEMBRADA	AREA_COSECHADA	PRODUCCION	ESTADO_FISICO_PRODUCCION	CICLO_CULTIVO
1	BOYACA	BUSBANZA	HORTALIZAS	ACELGA	2	1	1	FRUTO FRESCO	TRANSITORIO
2	CUNDINAMARCA	SOACHA	HORTALIZAS	ACELGA	82	80	1440	FRUTO FRESCO	TRANSITORIO
3	CUNDINAMARCA	COTA	HORTALIZAS	ACELGA	2	2	26	FRUTO FRESCO	TRANSITORIO
4	NORTE DE SANTANDER	LOS PATIOS	HORTALIZAS	ACELGA	3	3	48	FRUTO FRESCO	TRANSITORIO
5	NORTE DE SANTANDER	PAMPLONA	HORTALIZAS	ACELGA	1	1	5	FRUTO FRESCO	TRANSITORIO

3.3.1 CARACTERÍSTICAS:

3.3.1.1 Características numéricas:

Tenemos 3 características numéricas, las cuales poseen un rango de valores amplios, por lo tanto, es complejo ver su distribución, a partir de esta tabla se empieza a detectar el problema que se comentó inicialmente.

TABLA II
CARACTERÍSTICAS NUMÉRICAS

	AREA_SEMBRADA	AREA_COSECHADA	PRODUCCION
count	205347.00000	205347.00000	205347.00000
mean	292.09061	250.28154	2800.00391
std	1155.54773	982.07385	45143.62500
min	1.00000	0.00000	0.00000
25%	10.00000	8.00000	32.00000
50%	35.00000	30.00000	144.00000
75%	153.00000	130.00000	650.00000
max	47403.00000	38600.00000	4546116.00000

3.3.1.2 CARACTERÍSTICAS CATEGÓRICAS:

Tenemos 6 características categóricas, acá se observa otro problema y es cantidad de categorías que tiene cada variable, por lo tanto, se descarta el uso de la codificación one-hot, debido a que incrementa de forma razonable la cantidad de características:

TABLA III
CARACTERÍSTICAS CATEGORICAS

	DEPARTAMENTO	MUNICIPIO	GRUPO_CULTIVO	CULTIVO	ESTADO_FISICO_PRODUCCION	CICLO_CULTIVO
count	206068	206068	206068	206068	206068	206068
unique	32	1018	13	223	23	3
top	BOYACA	BOLIVAR	FRUTALES	MAIZ	FRUTO FRESCO	TRANSITORIO
freq	20576	1012	50236	24965	59682	108943

El resumen de nuestras características categóricas, Inicialmente se detectó que la distribución de los datos tenía rangos tan grandes que no se alcanzaba a generar la gráfica de distribución de los datos de forma correcta. Para poder observar la distribución de los datos de forma correcta, fue necesario escalonar los datos

3.3.2 Codificación de características

Retomando las características categóricas que tenemos en nuestra base de datos que son:

- **DEPARTAMENTO**
- **MUNICIPIO**
- **GRUPO_CULTIVO**
- **CULTIVO**
- **ESTADO_FISICO_PRODUCCION**
- **CICLO_CULTIVO**

Requerimos hacer una codificación a estas para poder entrenar el modelo, como se pudo ver anteriormente, se tiene una gran cantidad de categorías por cada característica, por lo tanto se optó por aplicar la codificación **label-encoder [26],** que básicamente le asigna un valor de 1 a n categoría a cada característica categórica, una vez que se aplicó este proceso nuestra base de datos quedó de la siguiente manera:

TABLA IV
BASE DE DATOS CON LABEL ENCODER

	DEPARTAMENTO	MUNICIPIO	GRUPO_CULTIVO	CULTIVO	AREA_SEMBRADA	AREA_COSECHADA	PRODUCCION	ESTADO_FISICO_PRODUCCION	CICLO_CULTIVO
1	5	114	6	0	2.0	1.0	1.0	7	2
2	13	849	6	0	82.0	80.0	1440.0	7	2
3	13	230	6	0	2.0	2.0	26.0	7	2
4	21	475	6	0	3.0	3.0	48.0	7	2
5	21	601	6	0	1.0	1.0	5.0	7	2

Las distribuciones de los datos quedaron de la siguiente forma:

TABLA V
INFORMACIÓN DE LA NUEVA BASE DE DATOS

	DEPARTAMENTO	MUNICIPIO	GRUPO_CULTIVO	CULTIVO	AREA_SEMBRADA	AREA_COSECHADA	PRODUCCION	ESTADO_FISICO_PRODUCCION	CICLO_CULTIVO
count	205347.00000	205347.00000	205347.00000	205347.00000	205347.00000	205347.00000	205347.00000	205347.00000	205347.00000
mean	14.96199	506.63010	5.79245	113.64990	292.09061	250.28154	2800.00391	10.34910	1.45741
std	9.11590	295.38992	3.79317	63.86404	1155.54773	982.07385	45143.62500	5.32256	0.62379
min	0.00000	0.00000	0.00000	0.00000	1.00000	0.00000	0.00000	0.00000	0.00000
25%	6.00000	253.00000	4.00000	52.00000	10.00000	8.00000	32.00000	7.00000	1.00000
50%	16.00000	505.00000	6.00000	128.00000	35.00000	30.00000	144.00000	8.00000	2.00000
75%	22.00000	760.00000	9.00000	163.00000	153.00000	130.00000	650.00000	12.00000	2.00000
max	31.00000	1017.00000	12.00000	220.00000	47403.00000	38600.00000	4546116.00000	22.00000	2.00000

Y gráficamente por medio de un diagrama de cajas podemos observar lo siguiente:

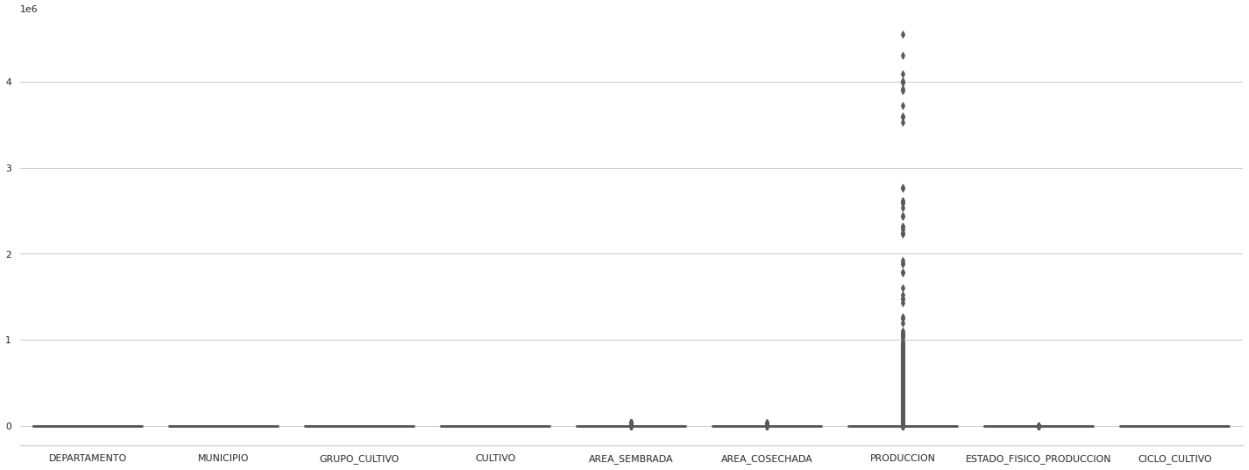

Fig. 7. Diagrama de cajas de la base de datos con label-encoder

Del gráfico anterior podemos observar que la diferencia entre escalas de los datos de cada característica es extremadamente amplia, problema que ya venía detectando, por lo tanto es necesario escalonar nuestra base de datos, pero antes de realizar este proceso, es necesario dividir la base de datos en cómo se propuso inicialmente, en medianos campesinos y grandes campesinos.

3.3.3 División de nuestra base datos en medianos agricultores y grandes agricultores:

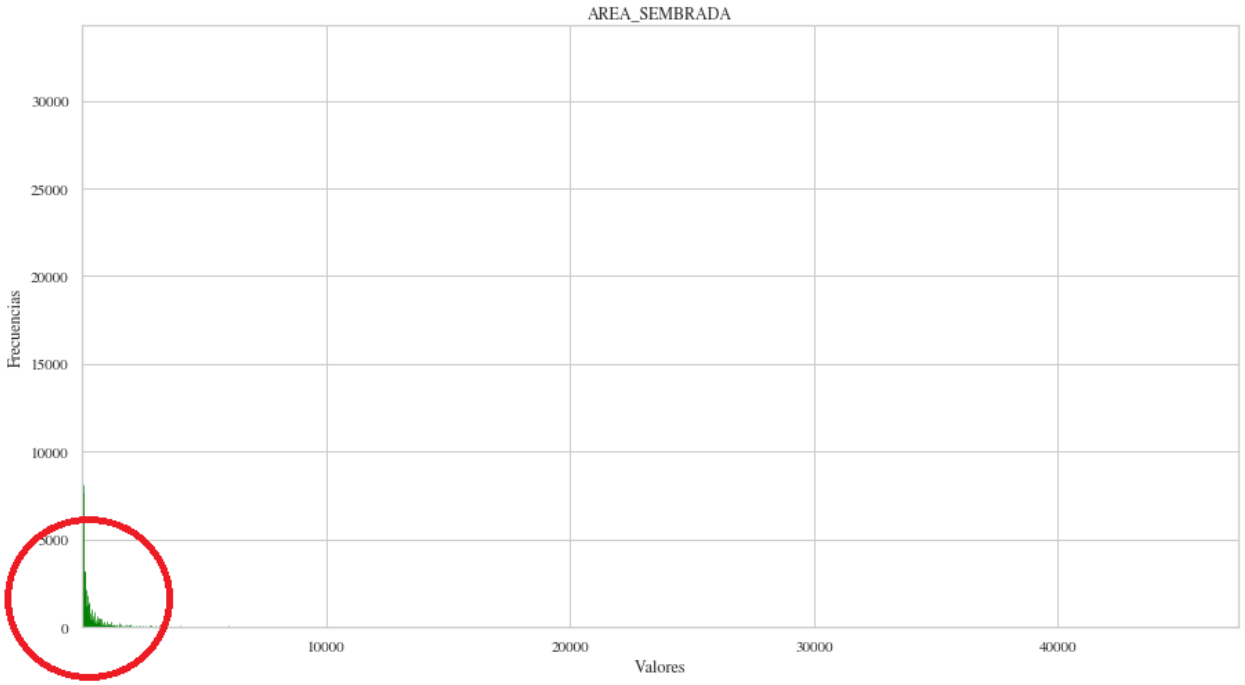

Fig. 8. Distribución de la característica objetivo

Tal como se observa en la distribución se observó que hay una gran cantidad de datos que se acumulan entre los rangos de 1 a 10 hectáreas, luego de realizar una vasta investigación se encontró el artículo ya mencionado [] en el que se define 3 tipos de agricultores:

- Pequeño agricultor: menor a 1 hectárea
- Medianos agricultores: entre 1 a 10 hectáreas.
- Grandes agricultores: de 10 hectáreas en adelante

Como se mencionó anteriormente, no se puede trabajar con los pequeños agricultores debido a que no se tiene una amplia cantidad de hectáreas sembradas menores a 1, además que la base de datos en esta característica es tipo entero por lo tanto no se tiene fraccionadas las hectáreas sembradas menores a 1.

3.3.3.1 Medianos agricultores

Luego de realizar esta partición vemos que esta base de datos quedo con 56393 de los 205347 que se tenían, aproximadamente 1/4 de la base de datos originales

Nuestra distribución de datos para esta base de datos quedó:

TABLA VI
INFORMACIÓN BASE DE DATOS MEDIANOS AGRICULTORES

	DEPARTAMENTO	MUNICIPIO	GRUPO_CULTIVO	CULTIVO	AREA_SEMBRADA	AREA_COSECHADA	PRODUCCION	ESTADO_FISICO_PRODUCCION	CICLO_CULTIVO
count	56393.00000	56393.00000	56393.00000	56393.00000	56393.00000	56393.00000	56393.00000	56393.00000	56393.00000
mean	15.27606	503.83734	5.71449	119.69009	5.25893	4.61274	48.80558	10.35363	1.55661
std	9.14997	295.84908	2.83887	64.18226	2.93314	2.89016	89.88070	4.27847	0.59254
min	0.00000	0.00000	0.00000	0.00000	1.00000	0.00000	0.00000	0.00000	0.00000
25%	6.00000	254.00000	4.00000	66.00000	3.00000	2.00000	7.00000	7.00000	1.00000
50%	16.00000	487.00000	6.00000	125.00000	5.00000	4.00000	20.00000	8.00000	2.00000
75%	23.00000	759.00000	7.00000	169.00000	8.00000	7.00000	57.00000	12.00000	2.00000
max	31.00000	1017.00000	12.00000	220.00000	10.00000	26.00000	1900.00000	22.00000	2.00000

Por otro lado, con el diagrama de cajas aún podemos observar que aún se presenta una gran cantidad de datos atípicos en la característica **PRODUCCION**

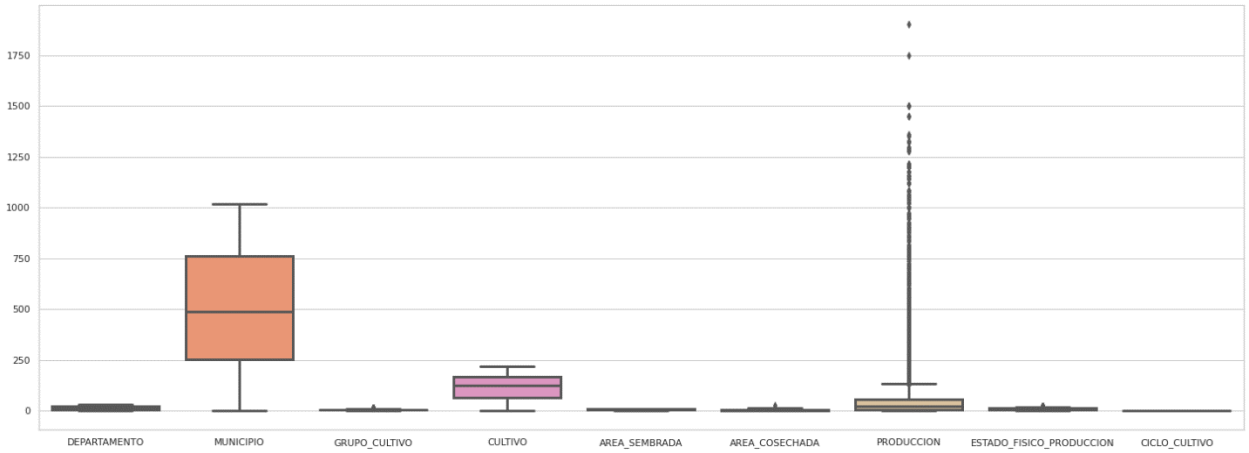

Fig. 9. Diagrama de cajas de la base de datos medianos agricultores

Y por último se puede observar en la variable objetivo ya posee una distribución mucho más visible

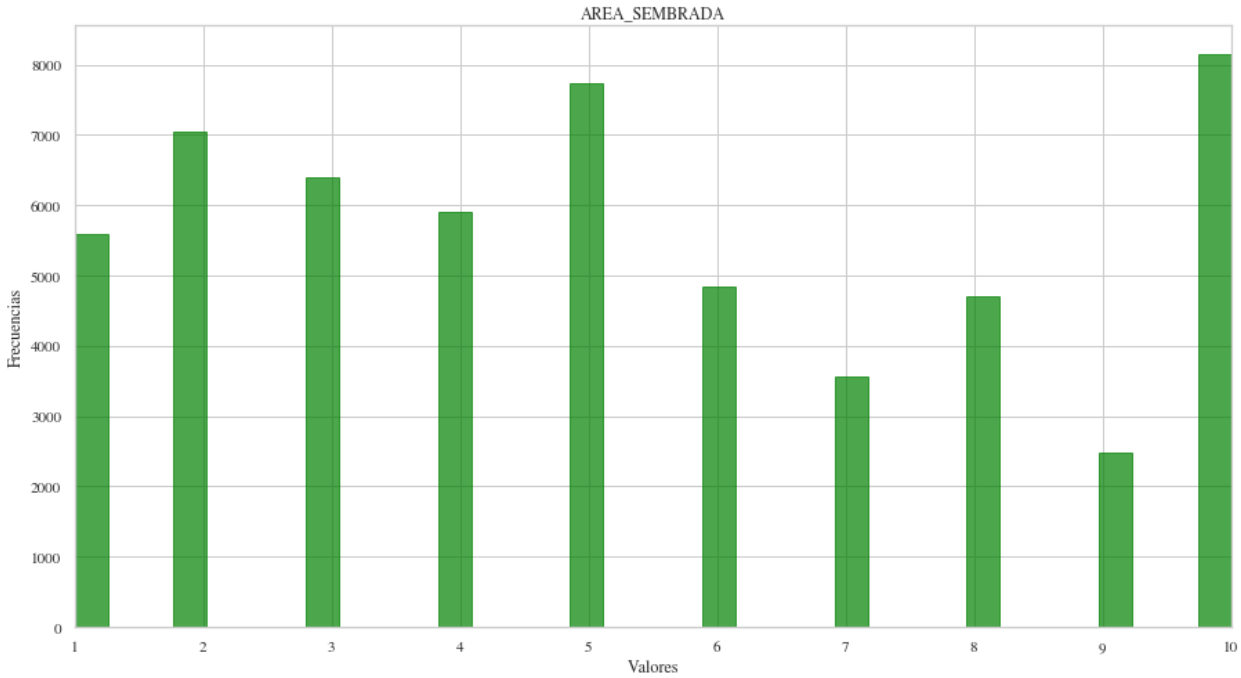

Fig. 10. Diagrama de barras de la característica objetivo de los medianos agricultores

3.3.3.1 Grandes agricultores

Por otro lado, en esta partición vemos que esta base de datos quedo con 148954 de los 205347 que se tenían, aproximadamente 3/4 de la base de datos originales. Nuestra distribución de datos para esta base de datos quedó muy similar a la que se tenía originalmente:

TABLA VII
INFORMACIÓN BASE DE DATOS GRANDES AGRICULTORES

	DEPARTAMENTO	MUNICIPIO	GRUPO_CULTIVO	CULTIVO	AREA_SEMBRADA	AREA_COSECHADA	PRODUCCION	ESTADO_FISICO_PRODUCCION	CICLO_CULTIVO
count	148954.00000	148954.00000	148954.00000	148954.00000	148954.00000	148954.00000	148954.00000	148954.00000	148954.00000
mean	14.84309	507.68743	5.82197	111.36312	400.72195	343.32376	3842.20752	10.34739	1.41985
std	9.10017	295.21000	4.09648	63.59382	1340.79749	1139.24622	52953.03125	5.66789	0.63116
min	0.00000	0.00000	0.00000	0.00000	11.00000	0.00000	0.00000	0.00000	0.00000
25%	6.00000	252.00000	4.00000	50.00000	30.00000	23.00000	90.00000	7.00000	1.00000
50%	13.00000	510.00000	6.00000	130.00000	79.00000	63.00000	300.00000	8.00000	1.00000
75%	22.00000	760.00000	9.00000	160.00000	269.00000	221.75000	1122.00000	12.00000	2.00000
max	31.00000	1017.00000	12.00000	220.00000	47403.00000	38600.00000	4546116.00000	22.00000	2.00000

Además, observando el diagrama de cajas de esta base de datos, se presenta una gran cantidad de datos atípicos en la característica **PRODUCCION**, presentando un problema muy similar observado en la base de datos original, por lo tanto, en posteriores apartados se va a realizar la eliminación de estos datos atípicos:

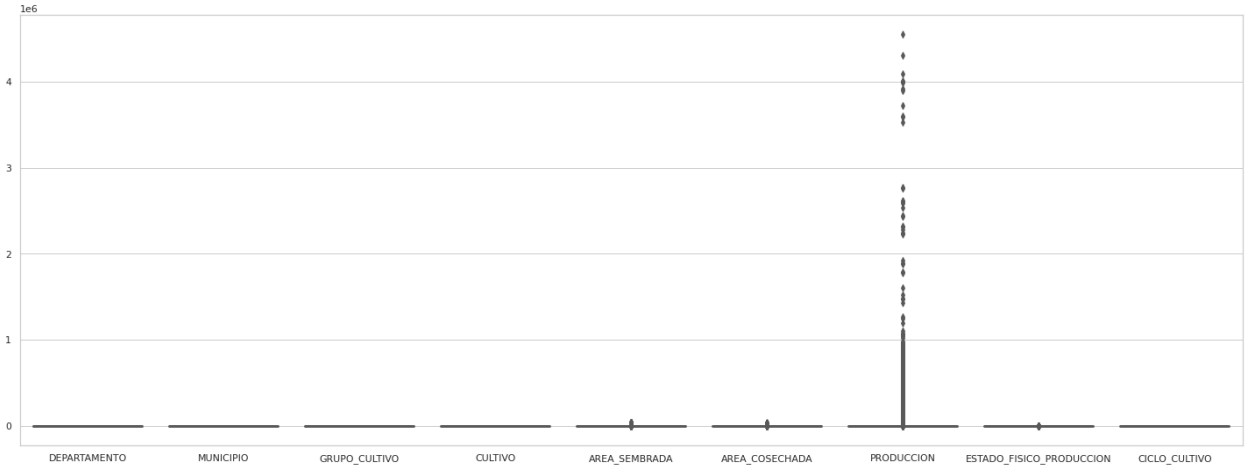

Fig. 11. Diagrama de cajas de la base de datos grandes agricultores

Y observando la característica objetivo **AREA_SEMBRADA**, se observa que aún hay un gran problema de escalas, por lo tanto, en el próximo apartado se realizará este procedimiento de escalado

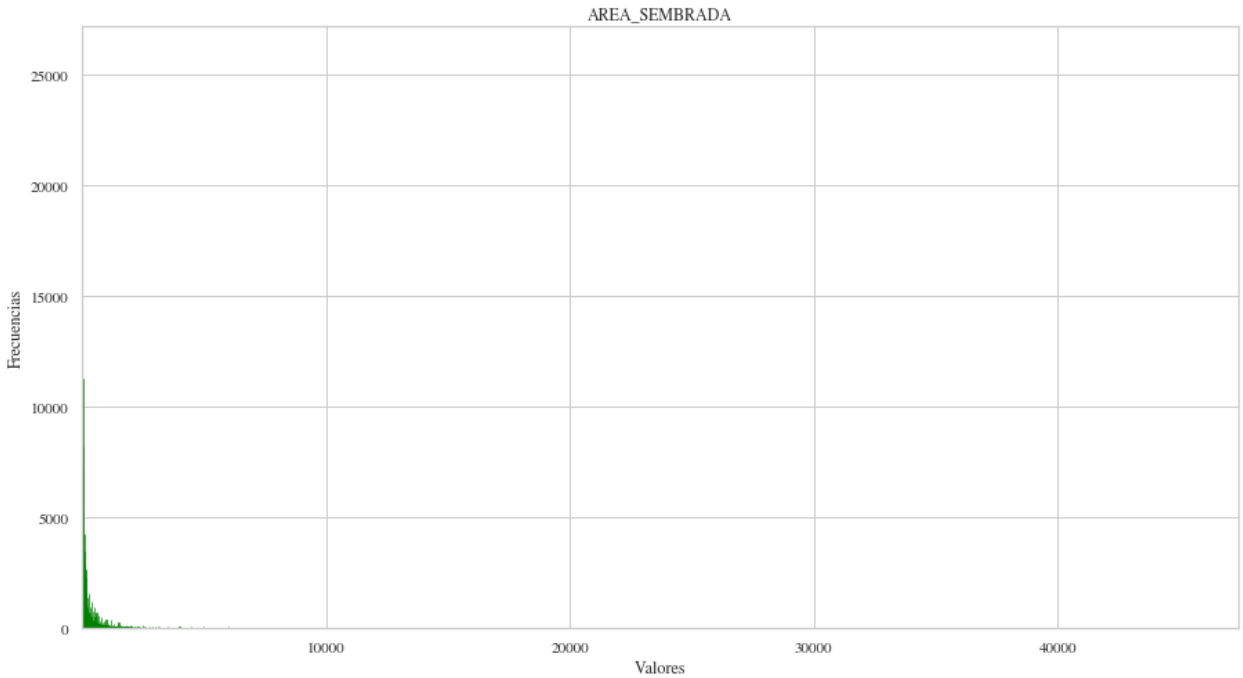

Fig. 12. Diagrama de barras de la característica objetivo de los grandes agricultores

4. PROCESO DE ANALÍTICA

4.1 PIPELINE PRINCIPAL

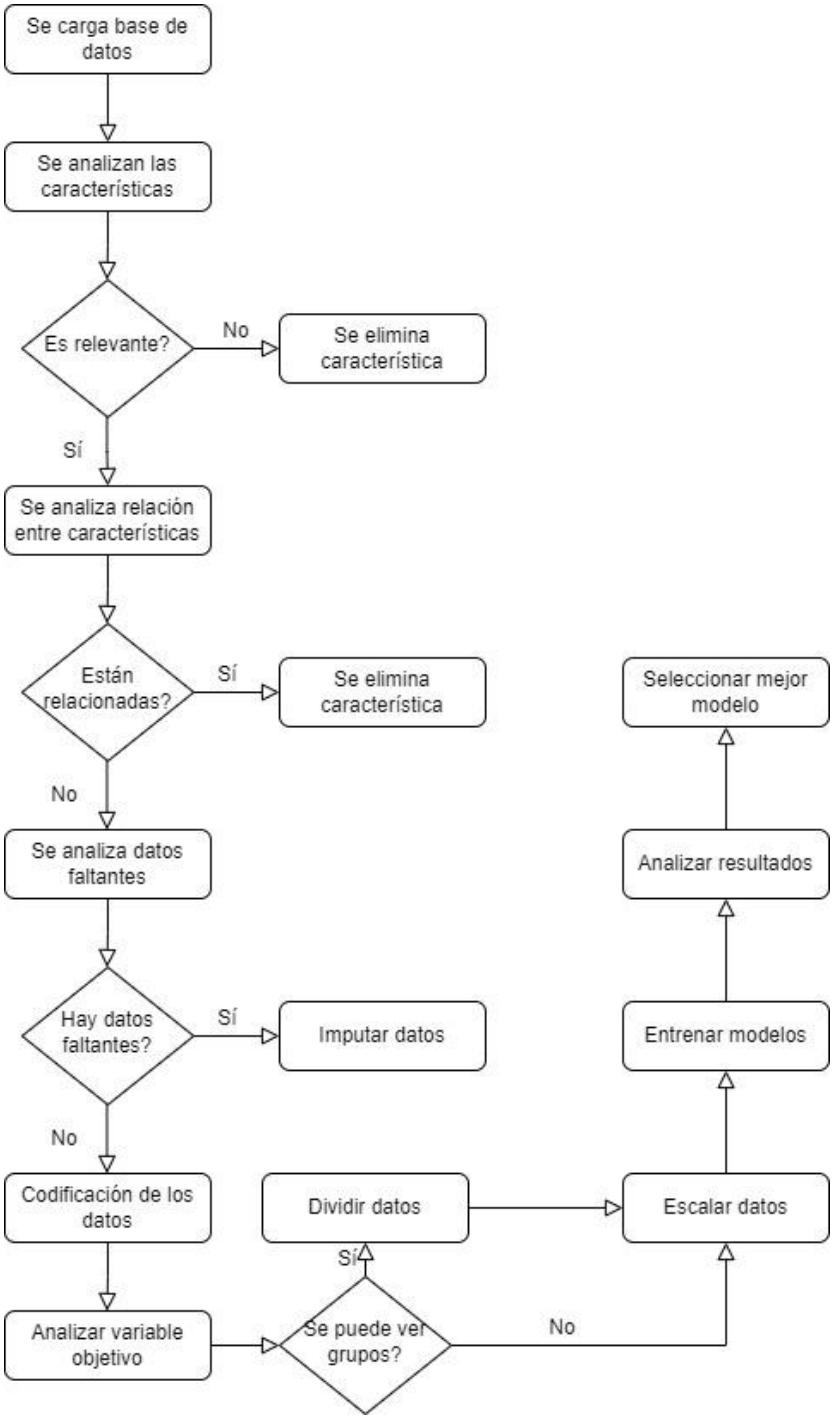

Fig. 13. Diagrama de flujo del procesamiento de los datos

4.2 PREPROCESAMIENTO

Se generaron múltiples bases de datos que surgieron luego de la aplicación de varias técnicas de codificación, además de estas las que se generaron a partir de varios métodos para eliminar los datos atípicos, donde se tuvo diferentes resultados de para entrenar los modelo y también se dejo la data sin eliminar datos atípicos

4.2.1 PREPROCESAMIENTO EN BASE DE DATOS DE MEDIANOS AGRICULTORES

4.2.1.1 Escalamiento robusto

En este parte de proceso se buscó una transformación de datos, con este método, debido a que se detectó una gran presencia de datos atípicos. En nuestro caso para esta transformación se escaló los datos en un rango entre el primer cuartil y el tercer cuartil, esto permitió escalar los datos independientemente de las características, ya que esta transformación computa las estadísticas más importantes de las muestras de los datos, como la media y la desviación estándar.

4.2.1.1.1 Medianos agricultores:

Una vez escalados los datos en la base de datos de los medianos agricultores con esta transformación, se observa que aún se presentan datos atípicos en la característica **PRODUCCION.**

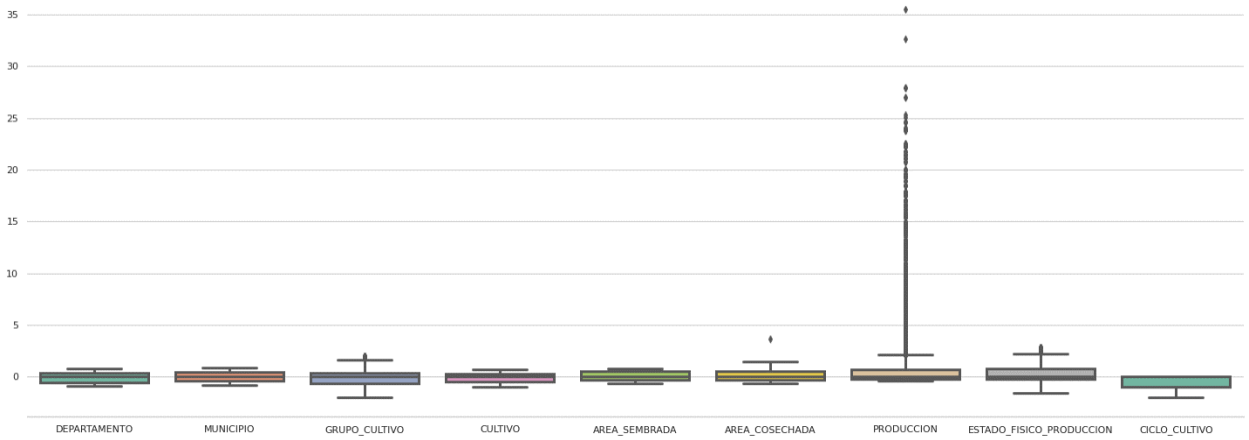

Fig. 14. Diagrama de cajas de los datos de medianos agricultores con escalamiento robusto

4.2.1.1.2 Grandes agricultores:

Para la base de datos de los grandes agricultores con esta transformación, se observa que aún se presentan datos atípicos en la característica **PRODUCCION,** además que en **AREA_COSECHA** y **AREA_SEMBRADA** también se detectaron algunos datos atípicos.

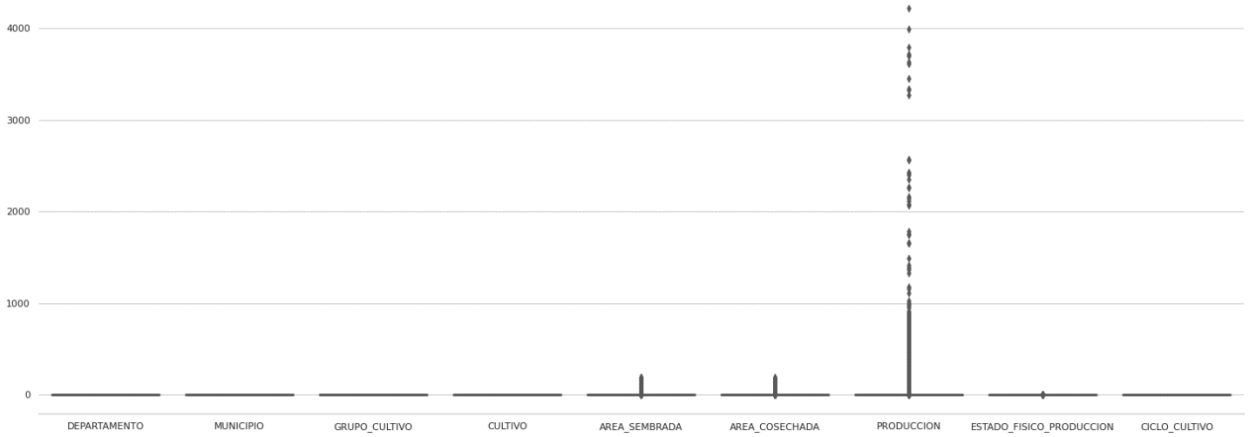

Fig. 15. Diagrama de cajas de los datos de grandes agricultores con escalamiento robusto

De este escalamiento se puede observar que aún los datos siguen estando es escalas que son muy diferentes por lo tanto no se apreció muy bien en los diagramas de cajas la distribución de cada característica.

4.2.1.2 Escalamiento estándar

Es una transformación es la más usada a la hora de escalar los datos antes del entrenamiento, básicamente consta de restar los datos por la media de cada característica y dividido por la varianza de cada característica.

4.2.1.2.1 Medianos agricultores:

Con esta transformación también podemos apreciar que hay una gran cantidad de datos atípicos en la característica de **PRODUCCION**, con esta transformación se puede observar un poco más claro la distribución de los datos en cada característica.

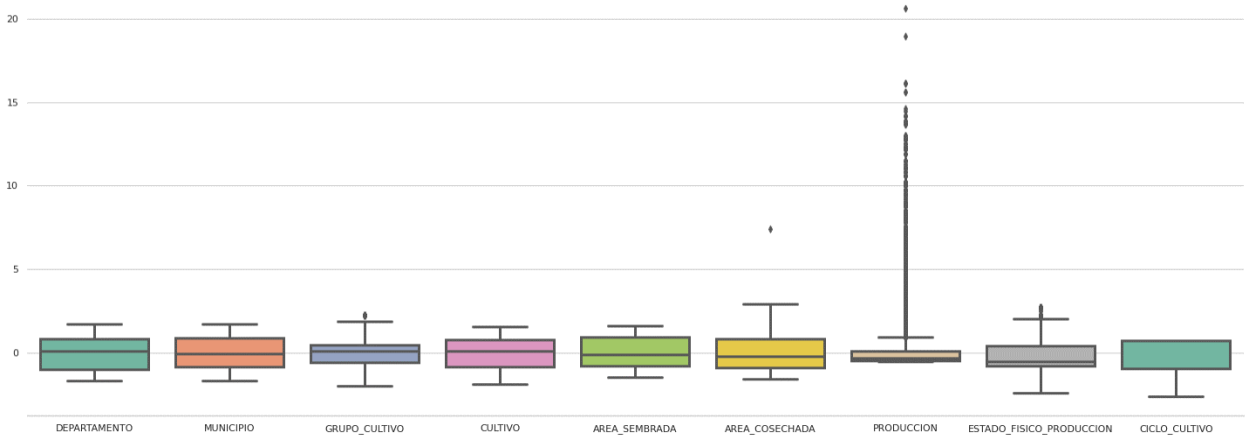

Fig. 16. Diagrama de cajas de los datos de medianos agricultores con escalamiento estándar

4.2.1.2.2 Grandes agricultores:

Por el lado de los grandes agricultores, se puede observar un poco mejor la distribución de los datos en cada característica, al igual que con la anterior transformación, acá se observa más datos atípicos en las características de **AREA_SEMBRADA** y **AREA_COSECHADA.**

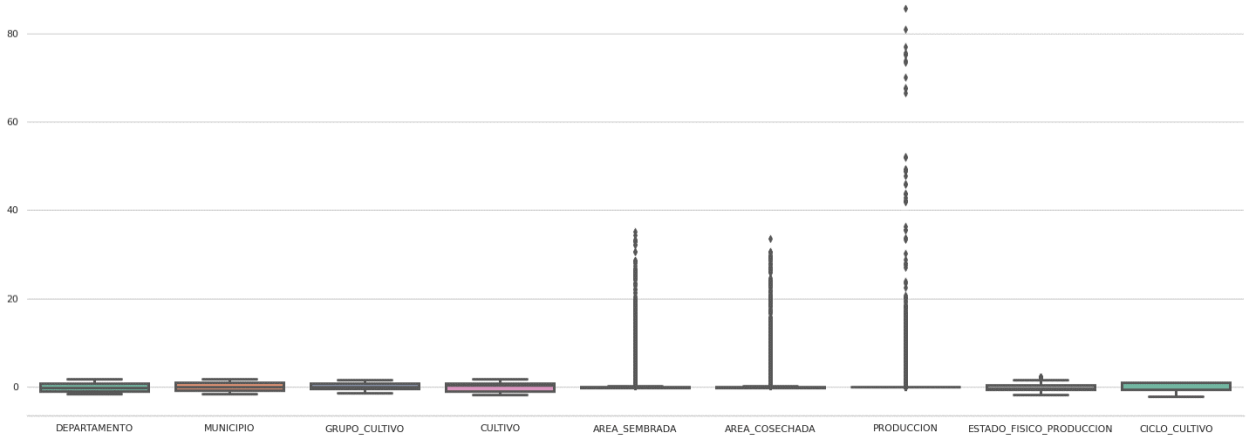

Fig. 17. Diagrama de cajas de los datos de grandes agricultores con escalamiento estándar

4.2.1.2 Escalamiento normalizado MIN - MAX

Con esta transformación, buscamos normalizar los datos a una escala en el rango entre 0 y 1, estrechando todos los datos y permitiendo observar de forma más clara la distribución de cada característica.

4.2.1.2.1 Medianos agricultores:

Con esta transformación podemos observar de forma mucho más clara la distribución de cada característica para los medianos agricultores, además se evidencio datos atípicos en otras características en las que los anteriores métodos no había detectado.

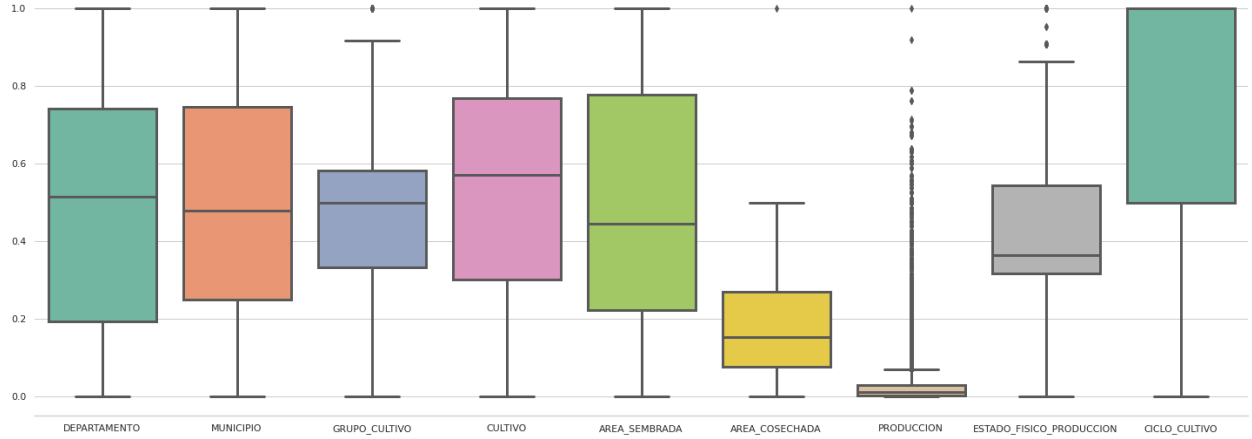

Fig. 18. Diagrama de cajas de los datos de medianos agricultores con escalamiento min-max

4.2.1.2.2 Grandes agricultores:

En esta transformación con los grandes agricultores, podemos observar que aumentaron los datos atípicos en las características **AREA_SEMBRADA** y **AREA_COSECHADA**, y por otro lado disminuyó la cantidad de datos atípicos que tenía la característica **PRODUCCION**.

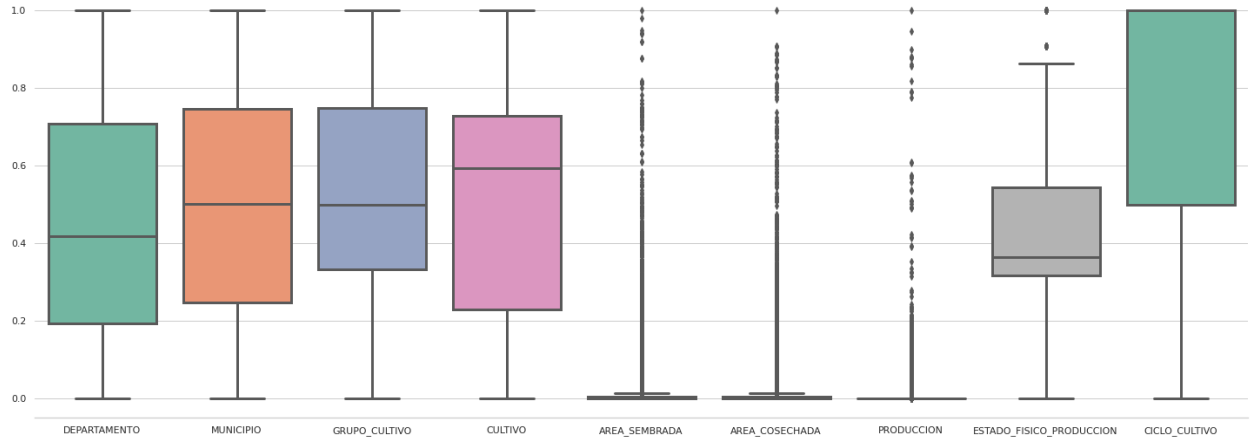

Fig. 19. Diagrama de cajas de los datos de grandes agricultores con escalamiento min-max

Esta transformación se comportó muy bien con ambas bases de datos, permitiendo observar de una forma más clara cada una de las variables.

4.2.1.2 Escalamiento Máxima Normalización

En esta transformación se buscará relacionar los datos con respecto al máximo valor de los datos por cada característica:

4.2.1.2.1 Medianos agricultores:

A comparación de las anteriores transformaciones, en este se observa una gran cantidad de datos atípicos que se generaron a partir de esta, la única característica que se ve sin datos atípicos es **CULTIVO**

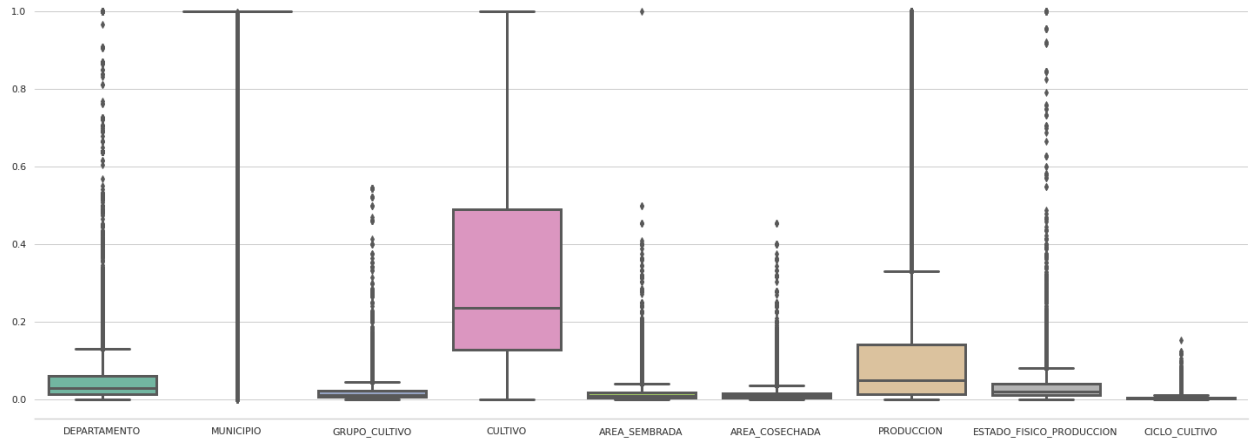

Fig. 20. Diagrama de cajas de los datos de medianos agricultores con máxima normalización

4.2.1.2.1 Grandes agricultores:

Al realizar esta transformación en esta base de datos, se observa que tuvo el mismo comportamiento que los medianos agricultores, a diferencia de que en esta las 2 únicas características sin datos atípicos, son **MUNICIPIO** y **PRODUCCION**

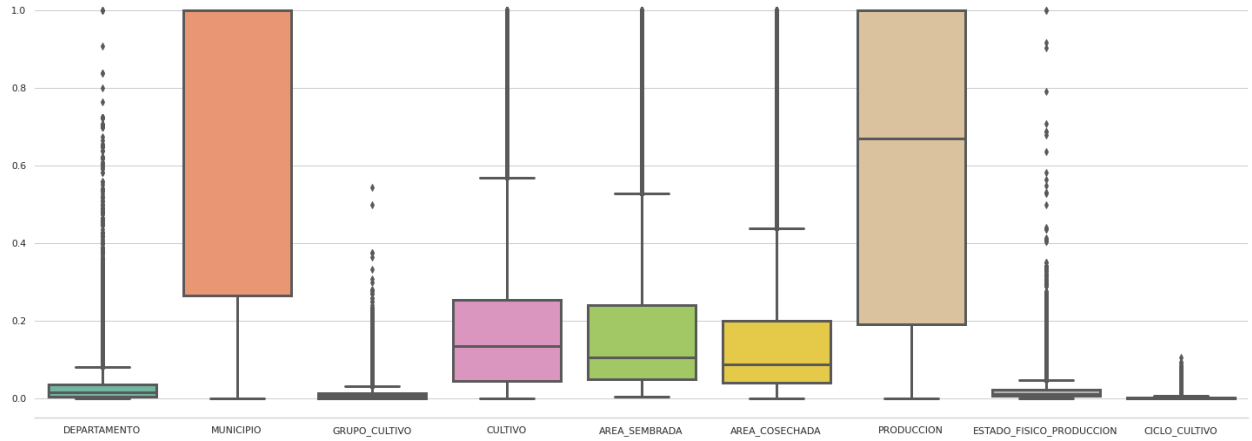

Fig. 21. Diagrama de cajas de los datos de grandes agricultores con máxima normalización

A partir de las anteriores transformaciones, podemos deducir que el método menos efectivo fue el escalamiento con máxima normalización y el escalado más efectivo fue con el método min-max, posiblemente esto se podrá observar mejor en el desempeño que tengan estas bases para el entrenamiento de los modelos. Además de los diagramas de cajas mostrados anteriormente podemos observar una gran cantidad de datos atípicos en diferentes características, por lo tanto, se procede a detectar y eliminar los datos atípicos por medio de dos métodos.

4.2.2.1 Detección de datos atípicos no supervisado

Este método se basa en una búsqueda local de datos atípicos, llamado factor atípico local o LOF, calculando la desviación local de la densidad de cada muestra con respecto a sus vecinos más cercanos. Por lo tanto, se debe utilizar una métrica de distancia entre una muestra y los vecinos más cercanos de forma local y a partir de la distancia se puede decir si es un dato atípico o no, donde una distancia más grande de un vecino con respecto a una muestra se considera como dato atípico.

Para este proceso que va a usar una función de la librería sklearn **LocalOutlierFactor [22]** para esto se variara 2 de sus parámetros, el número de vecinos y la distancia que se va tomar:

- Número de vecinos, se variaron entre 5, 7, 9 y 11 vecinos.
- Métricas de distancias, se usaron las métricas euclidean, minkowski y manhattan.

Cada una de estas variaciones junto con la base de datos escaladas generó una matriz diferente que se usarán para entrenar los diferentes modelos, a continuación, se mostrará el resultado de las bases de datos obtenidas.

4.2.2.1.1 Escalamiento robusto

4.2.2.1.1.1 Medianos agricultores:

TABLA VIII
BASES DE DATOS CREADAS A PARTIR DEL ESCALAMIENTO ROBUSTO Y ALGORITMO LOF APLICADO A MEDIANOS AGRICULTORES

	nombre	#vecinos	metrica	#muestras	#atipicos	#muestras_sin_atipicos
9	robusto_lof_euclidean_11	11	euclidean	56393	2064	54329
10	robusto_lof_minkowski_11	11	minkowski	56393	2064	54329
6	robusto_lof_euclidean_9	9	euclidean	56393	2951	53442
7	robusto_lof_minkowski_9	9	minkowski	56393	2951	53442
11	robusto_lof_manhattan_11	11	manhattan	56393	3117	53276
3	robusto_lof_euclidean_7	7	euclidean	56393	3989	52404
4	robusto_lof_minkowski_7	7	minkowski	56393	3989	52404
8	robusto_lof_manhattan_9	9	manhattan	56393	4185	52208
5	robusto_lof_manhattan_7	7	manhattan	56393	5492	50901
0	robusto_lof_euclidean_5	5	euclidean	56393	5905	50488
1	robusto_lof_minkowski_5	5	minkowski	56393	5905	50488
2	robusto_lof_manhattan_5	5	manhattan	56393	7797	48596

4.2.2.1.1.2 Grandes agricultores:

TABLA IX
BASES DE DATOS CREADAS A PARTIR DEL ESCALAMIENTO ROBUSTO Y ALGORITMO LOF APLICADO A GRANDES AGRICULTORES

	nombre	#vecinos	metrica	#muestras	#atipicos	#muestras_sin_atipicos
9	robusto_lof_euclidean_11	11	euclidean	148954	6612	142342
10	robusto_lof_minkowski_11	11	minkowski	148954	6612	142342
11	robusto_lof_manhattan_11	11	manhattan	148954	7903	141051
6	robusto_lof_euclidean_9	9	euclidean	148954	8147	140807
7	robusto_lof_minkowski_9	9	minkowski	148954	8147	140807
8	robusto_lof_manhattan_9	9	manhattan	148954	9654	139300
3	robusto_lof_euclidean_7	7	euclidean	148954	10507	138447
4	robusto_lof_minkowski_7	7	minkowski	148954	10507	138447
5	robusto_lof_manhattan_7	7	manhattan	148954	12429	136525
0	robusto_lof_euclidean_5	5	euclidean	148954	14632	134322
1	robusto_lof_minkowski_5	5	minkowski	148954	14632	134322
2	robusto_lof_manhattan_5	5	manhattan	148954	17186	131768

4.2.2.1.2 Escalamiento estándar

4.2.2.1.2.1 Medianos agricultores:

TABLA X
BASES DE DATOS CREADAS A PARTIR DEL ESCALAMIENTO ESTÁNDAR Y ALGORITMO LOF APLICADO A MEDIANOS AGRICULTORES

	nombre	#vecinos	metrica	#muestras	#atipicos	#muestras_sin_atipicos
9	estandar_lof_euclidean_11	11	euclidean	56393	2815	53578
10	estandar_lof_minkowski_11	11	minkowski	56393	2815	53578
6	estandar_lof_euclidean_9	9	euclidean	56393	3902	52491
7	estandar_lof_minkowski_9	9	minkowski	56393	3902	52491
11	estandar_lof_manhattan_11	11	manhattan	56393	3954	52439
8	estandar_lof_manhattan_9	9	manhattan	56393	5284	51109
3	estandar_lof_euclidean_7	7	euclidean	56393	5444	50949
4	estandar_lof_minkowski_7	7	minkowski	56393	5444	50949
5	estandar_lof_manhattan_7	7	manhattan	56393	6989	49404
0	estandar_lof_euclidean_5	5	euclidean	56393	7587	48806
1	estandar_lof_minkowski_5	5	minkowski	56393	7587	48806
2	estandar_lof_manhattan_5	5	manhattan	56393	9312	47081

4.2.2.1.2.2 Grandes agricultores:

TABLA XI
BASES DE DATOS CREADAS A PARTIR DEL ESCALAMIENTO ESTÁNDAR Y ALGORITMO LOF APLICADO A GRANDES AGRICULTORES

	nombre	#vecinos	metrica	#muestras	#atipicos	#muestras_sin_atipicos
9	estandar_lof_euclidean_11	11	euclidean	148954	20177	128777
10	estandar_lof_minkowski_11	11	minkowski	148954	20177	128777
11	estandar_lof_manhattan_11	11	manhattan	148954	20478	128476
6	estandar_lof_euclidean_9	9	euclidean	148954	22289	126665
7	estandar_lof_minkowski_9	9	minkowski	148954	22289	126665
8	estandar_lof_manhattan_9	9	manhattan	148954	22517	126437
3	estandar_lof_euclidean_7	7	euclidean	148954	24726	124228
4	estandar_lof_minkowski_7	7	minkowski	148954	24726	124228
5	estandar_lof_manhattan_7	7	manhattan	148954	25262	123692
0	estandar_lof_euclidean_5	5	euclidean	148954	27894	121060
1	estandar_lof_minkowski_5	5	minkowski	148954	27894	121060
2	estandar_lof_manhattan_5	5	manhattan	148954	28694	120260

4.2.2.1.3 Escalamiento normalizado MIN - MAX

4.2.2.1.3.1 Medianos agricultores:

TABLA XII
BASES DE DATOS CREADAS A PARTIR DEL ESCALAMIENTO MIN-MAX Y ALGORITMO LOF APLICADO A MEDIANOS AGRICULTORES

	nombre	#vecinos	metrica	#muestras	#atipicos	#muestras_sin_atipicos
9	min_max_lof_euclidean_11	11	euclidean	56393	3508	52885
10	min_max_lof_minkowski_11	11	minkowski	56393	3508	52885
11	min_max_lof_manhattan_11	11	manhattan	56393	4526	51867
6	min_max_lof_euclidean_9	9	euclidean	56393	4830	51563
7	min_max_lof_minkowski_9	9	minkowski	56393	4830	51563
8	min_max_lof_manhattan_9	9	manhattan	56393	5939	50454
3	min_max_lof_euclidean_7	7	euclidean	56393	6711	49682
4	min_max_lof_minkowski_7	7	minkowski	56393	6711	49682
5	min_max_lof_manhattan_7	7	manhattan	56393	7960	48433
0	min_max_lof_euclidean_5	5	euclidean	56393	9279	47114
1	min_max_lof_minkowski_5	5	minkowski	56393	9279	47114
2	min_max_lof_manhattan_5	5	manhattan	56393	10604	45789

4.2.2.1.3.2 Grandes agricultores:

TABLA XIII
BASES DE DATOS CREADAS A PARTIR DEL ESCALAMIENTO MIN-MAX Y ALGORITMO LOF APLICADO A GRANDES AGRICULTORES

	nombre	#vecinos	metrica	#muestras	#atipicos	#muestras_sin_atipicos
11	min_max_lof_manhattan_11	11	manhattan	148954	27120	121834
9	min_max_lof_euclidean_11	11	euclidean	148954	27162	121792
10	min_max_lof_minkowski_11	11	minkowski	148954	27162	121792
6	min_max_lof_euclidean_9	9	euclidean	148954	27863	121091
7	min_max_lof_minkowski_9	9	minkowski	148954	27863	121091
8	min_max_lof_manhattan_9	9	manhattan	148954	27915	121039
3	min_max_lof_euclidean_7	7	euclidean	148954	29472	119482
4	min_max_lof_minkowski_7	7	minkowski	148954	29472	119482
5	min_max_lof_manhattan_7	7	manhattan	148954	29472	119482
0	min_max_lof_euclidean_5	5	euclidean	148954	32037	116917
1	min_max_lof_minkowski_5	5	minkowski	148954	32037	116917
2	min_max_lof_manhattan_5	5	manhattan	148954	32245	116709

4.2.2.1.4 Escalamiento Máxima Normalización

4.2.2.1.4.1 Medianos agricultores:

TABLA XIV
BASES DE DATOS CREADAS A PARTIR DEL ESCALAMIENTO MÁXIMA NORMALIZACIÓN Y ALGORITMO LOF APLICADO A MEDIANOS AGRICULTORES

	nombre	#vecinos	metrica	#muestras	#atipicos	#muestras_sin_atipicos
9	max_normalizacion_lof_euclidean_11	11	euclidean	56393	1765	54628
10	max_normalizacion_lof_minkowski_11	11	minkowski	56393	1765	54628
6	max_normalizacion_lof_euclidean_9	9	euclidean	56393	2466	53927
7	max_normalizacion_lof_minkowski_9	9	minkowski	56393	2466	53927
11	max_normalizacion_lof_manhattan_11	11	manhattan	56393	2581	53812
8	max_normalizacion_lof_manhattan_9	9	manhattan	56393	3496	52897
3	max_normalizacion_lof_euclidean_7	7	euclidean	56393	3528	52865
4	max_normalizacion_lof_minkowski_7	7	minkowski	56393	3528	52865
5	max_normalizacion_lof_manhattan_7	7	manhattan	56393	5063	51330
0	max_normalizacion_lof_euclidean_5	5	euclidean	56393	5345	51048
1	max_normalizacion_lof_minkowski_5	5	minkowski	56393	5345	51048
2	max_normalizacion_lof_manhattan_5	5	manhattan	56393	7361	49032

4.2.2.1.4.2 Grandes agricultores:

TABLA XV
BASES DE DATOS CREADAS A PARTIR DEL ESCALAMIENTO MÁXIMA NORMALIZACIÓN Y ALGORITMO LOF APLICADO A GRANDES AGRICULTORES

	nombre	#vecinos	metrica	#muestras	#atipicos	#muestras_sin_atipicos
9	max_normalizacion_lof_euclidean_11	11	euclidean	148954	2305	146649
10	max_normalizacion_lof_minkowski_11	11	minkowski	148954	2305	146649
6	max_normalizacion_lof_euclidean_9	9	euclidean	148954	2835	146119
7	max_normalizacion_lof_minkowski_9	9	minkowski	148954	2835	146119
11	max_normalizacion_lof_manhattan_11	11	manhattan	148954	3488	145466
3	max_normalizacion_lof_euclidean_7	7	euclidean	148954	3838	145116
4	max_normalizacion_lof_minkowski_7	7	minkowski	148954	3838	145116
8	max_normalizacion_lof_manhattan_9	9	manhattan	148954	4362	144592
5	max_normalizacion_lof_manhattan_7	7	manhattan	148954	5678	143276
0	max_normalizacion_lof_euclidean_5	5	euclidean	148954	6400	142554
1	max_normalizacion_lof_minkowski_5	5	minkowski	148954	6400	142554
2	max_normalizacion_lof_manhattan_5	5	manhattan	148954	9038	139916

4.2.2.2 Detección de datos atípicos basado en el algoritmo de aislamiento forestal

Este método se basa en aislar las muestras por selectividad aleatoria de una característica de los datos, estableciendo un valor de referencia o umbral que divide entre los valores máximos y mínimos de una característica. Por lo tanto, para realizar la división, se generan particiones de los datos que es representado mediante una estructura de un árbol de decisión, donde cada partición aísla una o más muestras que representen datos atípicos. Para mayor efectividad en la detección de datos atípicos, el algoritmo genera un bosque de árboles aleatoriamente donde se calcula las longitudes de las ramificaciones de los árboles. Al final para determinar qué datos son atípicos, se debe buscar aquellas longitudes que sean más cortas para una o más muestras, estas se considerarán como datos atípicos.

Para este proceso que va a usar una función de la librería sklearn **IsolationForest [23]** para esto se variara 2 de sus parámetros, el número de estimadores o árboles y la contaminación:

- **Número de estimadores,** se variaron entre 100, 200, 300 y 400 árboles.
- **Contaminación,** se variaron entre 0.05, 0.10, 0.15, 0.20 y auto, donde la contaminación auto, se calculará de forma automática en cada iteración.

Cada una de estas variaciones junto con la base de datos escaladas generó una matriz diferente que se usarán para entrenar los diferentes modelos, a continuación, se mostrará el resultado de las bases de datos obtenidas.

4.2.2.2.1 Escalamiento robusto

4.2.2.2.1.1 Medianos agricultores:

TABLA XVI
BASES DE DATOS CREADAS A PARTIR DEL ESCALAMIENTO ROBUSTO Y ALGORITMO ISF APLICADO A MEDIANOS AGRICULTORES

	nombre	#estimadores	contaminacion	#muestras	#atipicos	#muestras_sin_atipicos
11	robusto_isf_300_0_05	300	0.05	56393	2818	53575
6	robusto_isf_200_0_05	200	0.05	56393	2819	53574
1	robusto_isf_100_0_05	100	0.05	56393	2820	53573
16	robusto_isf_400_0_05	400	0.05	56393	2820	53573
2	robusto_isf_100_0_1	100	0.1	56393	5640	50753
17	robusto_isf_400_0_1	400	0.1	56393	5640	50753
7	robusto_isf_200_0_1	200	0.1	56393	5640	50753
12	robusto_isf_300_0_1	300	0.1	56393	5640	50753
18	robusto_isf_400_0_15	400	0.15	56393	8458	47935
3	robusto_isf_100_0_15	100	0.15	56393	8459	47934
8	robusto_isf_200_0_15	200	0.15	56393	8459	47934
13	robusto_isf_300_0_15	300	0.15	56393	8459	47934
9	robusto_isf_200_0_2	200	0.2	56393	11279	45114
14	robusto_isf_300_0_2	300	0.2	56393	11279	45114
19	robusto_isf_400_0_2	400	0.2	56393	11279	45114
4	robusto_isf_100_0_2	100	0.2	56393	11279	45114
10	robusto_isf_300_auto	300	auto	56393	17870	38523
15	robusto_isf_400_auto	400	auto	56393	18423	37970
5	robusto_isf_200_auto	200	auto	56393	18864	37529
0	robusto_isf_100_auto	100	auto	56393	19080	37313

4.2.2.2.1.2 Grandes agricultores:

TABLA XVII
BASES DE DATOS CREADAS A PARTIR DEL ESCALAMIENTO ROBUSTO Y ALGORITMO ISF APLICADO A GRANDES AGRICULTORES

	nombre	#estimadores	contaminacion	#muestras	#atipicos	#muestras_sin_atipicos
1	robusto_isf_100_0_05	100	0.05	148954	7448	141506
16	robusto_isf_400_0_05	400	0.05	148954	7448	141506
6	robusto_isf_200_0_05	200	0.05	148954	7448	141506
11	robusto_isf_300_0_05	300	0.05	148954	7448	141506
7	robusto_isf_200_0_1	200	0.1	148954	14895	134059
2	robusto_isf_100_0_1	100	0.1	148954	14896	134058
17	robusto_isf_400_0_1	400	0.1	148954	14896	134058
12	robusto_isf_300_0_1	300	0.1	148954	14896	134058
3	robusto_isf_100_0_15	100	0.15	148954	22343	126611
8	robusto_isf_200_0_15	200	0.15	148954	22343	126611
18	robusto_isf_400_0_15	400	0.15	148954	22343	126611
13	robusto_isf_300_0_15	300	0.15	148954	22343	126611
14	robusto_isf_300_0_2	300	0.2	148954	29790	119164
9	robusto_isf_200_0_2	200	0.2	148954	29791	119163
19	robusto_isf_400_0_2	400	0.2	148954	29791	119163
4	robusto_isf_100_0_2	100	0.2	148954	29791	119163
5	robusto_isf_200_auto	200	auto	148954	30727	118227
10	robusto_isf_300_auto	300	auto	148954	31749	117205
15	robusto_isf_400_auto	400	auto	148954	31773	117181
0	robusto_isf_100_auto	100	auto	148954	32600	116354

4.2.2.2.2 Escalamiento estándar

4.2.2.2.2.1 Medianos agricultores:

TABLA XVIII
BASES DE DATOS CREADAS A PARTIR DEL ESCALAMIENTO ESTÁNDAR Y ALGORITMO ISF APLICADO A MEDIANOS AGRICULTORES

	nombre	#estimadores	contaminacion	#muestras	#atipicos	#muestras_sin_atipicos
11	estandar_isf_300_0_05	300	0.05	56393	2818	53575
6	estandar_isf_200_0_05	200	0.05	56393	2819	53574
1	estandar_isf_100_0_05	100	0.05	56393	2820	53573
16	estandar_isf_400_0_05	400	0.05	56393	2820	53573
2	estandar_isf_100_0_1	100	0.1	56393	5640	50753
17	estandar_isf_400_0_1	400	0.1	56393	5640	50753
7	estandar_isf_200_0_1	200	0.1	56393	5640	50753
12	estandar_isf_300_0_1	300	0.1	56393	5640	50753
18	estandar_isf_400_0_15	400	0.15	56393	8458	47935
3	estandar_isf_100_0_15	100	0.15	56393	8459	47934
8	estandar_isf_200_0_15	200	0.15	56393	8459	47934
13	estandar_isf_300_0_15	300	0.15	56393	8459	47934
9	estandar_isf_200_0_2	200	0.2	56393	11279	45114
14	estandar_isf_300_0_2	300	0.2	56393	11279	45114
19	estandar_isf_400_0_2	400	0.2	56393	11279	45114
4	estandar_isf_100_0_2	100	0.2	56393	11279	45114
10	estandar_isf_300_auto	300	auto	56393	17870	38523
15	estandar_isf_400_auto	400	auto	56393	18423	37970
5	estandar_isf_200_auto	200	auto	56393	18864	37529
0	estandar_isf_100_auto	100	auto	56393	19080	37313

4.2.2.2.2.2 Grandes agricultores:

TABLA XIX
BASES DE DATOS CREADAS A PARTIR DEL ESCALAMIENTO ESTÁNDAR Y ALGORITMO ISF APLICADO A GRANDES AGRICULTORES

	nombre	#estimadores	contaminacion	#muestras	#atipicos	#muestras_sin_atipicos
1	estandar_isf_100_0_05	100	0.05	148954	7448	141506
16	estandar_isf_400_0_05	400	0.05	148954	7448	141506
6	estandar_isf_200_0_05	200	0.05	148954	7448	141506
11	estandar_isf_300_0_05	300	0.05	148954	7448	141506
7	estandar_isf_200_0_1	200	0.1	148954	14895	134059
2	estandar_isf_100_0_1	100	0.1	148954	14896	134058
17	estandar_isf_400_0_1	400	0.1	148954	14896	134058
12	estandar_isf_300_0_1	300	0.1	148954	14896	134058
3	estandar_isf_100_0_15	100	0.15	148954	22343	126611
8	estandar_isf_200_0_15	200	0.15	148954	22343	126611
18	estandar_isf_400_0_15	400	0.15	148954	22343	126611
13	estandar_isf_300_0_15	300	0.15	148954	22343	126611
14	estandar_isf_300_0_2	300	0.2	148954	29790	119164
9	estandar_isf_200_0_2	200	0.2	148954	29791	119163
19	estandar_isf_400_0_2	400	0.2	148954	29791	119163
4	estandar_isf_100_0_2	100	0.2	148954	29791	119163
5	estandar_isf_200_auto	200	auto	148954	30727	118227
10	estandar_isf_300_auto	300	auto	148954	31749	117205
15	estandar_isf_400_auto	400	auto	148954	31773	117181
0	estandar_isf_100_auto	100	auto	148954	32600	116354

4.2.2.2.3 Escalamiento normalizado MIN - MAX

4.2.2.2.3.1 Medianos agricultores:

TABLA XX
BASES DE DATOS CREADAS A PARTIR DEL ESCALAMIENTO MIN-MAX Y ALGORITMO ISF APLICADO A MEDIANOS AGRICULTORES

	nombre	#estimadores	contaminacion	#muestras	#atipicos	#muestras_sin_atipicos
11	min_max_isf_300_0_05	300	0.05	56393	2818	53575
6	min_max_isf_200_0_05	200	0.05	56393	2819	53574
1	min_max_isf_100_0_05	100	0.05	56393	2820	53573
16	min_max_isf_400_0_05	400	0.05	56393	2820	53573
2	min_max_isf_100_0_1	100	0.1	56393	5640	50753
17	min_max_isf_400_0_1	400	0.1	56393	5640	50753
7	min_max_isf_200_0_1	200	0.1	56393	5640	50753
12	min_max_isf_300_0_1	300	0.1	56393	5640	50753
18	min_max_isf_400_0_15	400	0.15	56393	8458	47935
3	min_max_isf_100_0_15	100	0.15	56393	8459	47934
8	min_max_isf_200_0_15	200	0.15	56393	8459	47934
13	min_max_isf_300_0_15	300	0.15	56393	8459	47934
9	min_max_isf_200_0_2	200	0.2	56393	11279	45114
14	min_max_isf_300_0_2	300	0.2	56393	11279	45114
19	min_max_isf_400_0_2	400	0.2	56393	11279	45114
4	min_max_isf_100_0_2	100	0.2	56393	11279	45114
10	min_max_isf_300_auto	300	auto	56393	17870	38523
15	min_max_isf_400_auto	400	auto	56393	18423	37970
5	min_max_isf_200_auto	200	auto	56393	18864	37529
0	min_max_isf_100_auto	100	auto	56393	19080	37313

4.2.2.2.3.2 Grandes agricultores:

TABLA XXI
BASES DE DATOS CREADAS A PARTIR DEL ESCALAMIENTO MIN-MAX Y ALGORITMO ISF APLICADO A GRANDES AGRICULTORES

	nombre	#estimadores	contaminacion	#muestras	#atipicos	#muestras_sin_atipicos
1	min_max_isf_100_0_05	100	0.05	148954	7448	141506
16	min_max_isf_400_0_05	400	0.05	148954	7448	141506
6	min_max_isf_200_0_05	200	0.05	148954	7448	141506
11	min_max_isf_300_0_05	300	0.05	148954	7448	141506
7	min_max_isf_200_0_1	200	0.1	148954	14895	134059
2	min_max_isf_100_0_1	100	0.1	148954	14896	134058
17	min_max_isf_400_0_1	400	0.1	148954	14896	134058
12	min_max_isf_300_0_1	300	0.1	148954	14896	134058
3	min_max_isf_100_0_15	100	0.15	148954	22343	126611
8	min_max_isf_200_0_15	200	0.15	148954	22343	126611
18	min_max_isf_400_0_15	400	0.15	148954	22343	126611
13	min_max_isf_300_0_15	300	0.15	148954	22343	126611
14	min_max_isf_300_0_2	300	0.2	148954	29790	119164
9	min_max_isf_200_0_2	200	0.2	148954	29791	119163
19	min_max_isf_400_0_2	400	0.2	148954	29791	119163
4	min_max_isf_100_0_2	100	0.2	148954	29791	119163
5	min_max_isf_200_auto	200	auto	148954	30727	118227
10	min_max_isf_300_auto	300	auto	148954	31749	117205
15	min_max_isf_400_auto	400	auto	148954	31773	117181
0	min_max_isf_100_auto	100	auto	148954	32600	116354

4.2.2.2.4 Escalamiento Máxima Normalización

4.2.2.2.4.1 Medianos agricultores:

TABLA XXII
BASES DE DATOS CREADAS A PARTIR DEL ESCALAMIENTO MÁXIMA NORMALIZACIÓN Y ALGORITMO ISF APLICADO A MEDIANOS AGRICULTORES

	nombre	#estimadores	contaminacion	#muestras	#atipicos	#muestras_sin_atipicos
1	min_max_isf_100_0_05	100	0.05	56393	2819	53574
6	min_max_isf_200_0_05	200	0.05	56393	2820	53573
11	min_max_isf_300_0_05	300	0.05	56393	2820	53573
16	min_max_isf_400_0_05	400	0.05	56393	2820	53573
2	min_max_isf_100_0_1	100	0.1	56393	5640	50753
17	min_max_isf_400_0_1	400	0.1	56393	5640	50753
7	min_max_isf_200_0_1	200	0.1	56393	5640	50753
12	min_max_isf_300_0_1	300	0.1	56393	5640	50753
18	min_max_isf_400_0_15	400	0.15	56393	8455	47938
3	min_max_isf_100_0_15	100	0.15	56393	8459	47934
8	min_max_isf_200_0_15	200	0.15	56393	8459	47934
13	min_max_isf_300_0_15	300	0.15	56393	8459	47934
5	min_max_isf_200_auto	200	auto	56393	9822	46571
15	min_max_isf_400_auto	400	auto	56393	9839	46554
10	min_max_isf_300_auto	300	auto	56393	9906	46487
0	min_max_isf_100_auto	100	auto	56393	10177	46216
9	min_max_isf_200_0_2	200	0.2	56393	11278	45115
4	min_max_isf_100_0_2	100	0.2	56393	11278	45115
14	min_max_isf_300_0_2	300	0.2	56393	11279	45114
19	min_max_isf_400_0_2	400	0.2	56393	11279	45114

4.2.2.2.4.2 Grandes agricultores:

TABLA XXIII
BASES DE DATOS CREADAS A PARTIR DEL ESCALAMIENTO MÁXIMA NORMALIZACIÓN Y ALGORITMO ISF APLICADO A GRANDES AGRICULTORES

	nombre	#estimadores	contaminacion	#muestras	#atipicos	#muestras_sin_atipicos
11	min_max_isf_300_0_05	300	0.05	148954	7447	141507
1	min_max_isf_100_0_05	100	0.05	148954	7448	141506
16	min_max_isf_400_0_05	400	0.05	148954	7448	141506
6	min_max_isf_200_0_05	200	0.05	148954	7448	141506
7	min_max_isf_200_0_1	200	0.1	148954	14895	134059
2	min_max_isf_100_0_1	100	0.1	148954	14896	134058
17	min_max_isf_400_0_1	400	0.1	148954	14896	134058
12	min_max_isf_300_0_1	300	0.1	148954	14896	134058
3	min_max_isf_100_0_15	100	0.15	148954	22343	126611
8	min_max_isf_200_0_15	200	0.15	148954	22343	126611
18	min_max_isf_400_0_15	400	0.15	148954	22343	126611
13	min_max_isf_300_0_15	300	0.15	148954	22343	126611
9	min_max_isf_200_0_2	200	0.2	148954	29791	119163
14	min_max_isf_300_0_2	300	0.2	148954	29791	119163
19	min_max_isf_400_0_2	400	0.2	148954	29791	119163
4	min_max_isf_100_0_2	100	0.2	148954	29791	119163
15	min_max_isf_400_auto	400	auto	148954	29958	118996
10	min_max_isf_300_auto	300	auto	148954	30122	118832
5	min_max_isf_200_auto	200	auto	148954	30314	118640
0	min_max_isf_100_auto	100	auto	148954	31268	117686

Una vez que se obtuvieron los resultados, se puede observar que los datos atípicos detectados por aislamiento forestal varían mucho dependiendo de la contaminación, por otro lado, con el algoritmo LOF la cantidad de atípicos no varía tanto con los parámetros ingresados.

4.3 MODELOS

Debido a que el objetivo de este primer acercamiento a la estimación de **AREAS_SEMBRADAS**, sólo se modificará el hiperparámetro más relevante del modelo en caso de que este posea alguno.

4.3.1 Modelo de regresión lineal

Consta de un modelo lineal con coeficientes o pesos para minimizar la suma residual de cuadrados entre los objetivos observados en el conjunto de datos y los objetivos predichos por la aproximación lineal, para esto se hará uso del algoritmo **LinearRegression [3]** de sklearn.

4.3.2 Modelo de regresión robusta

Es un algoritmo iterativo que permite trabajar un estimador o modelo de regresión, por medio de iteraciones subsiguientes después de dividir el conjunto de datos sin y con influencia de datos atípicos. Por lo tanto, permite trabajar dos tipos de problema que evita el aumento del error de regresión. para esto se hará uso del algoritmo **RANSACRegresor [4]** de sklearn y variando el residual máximo para que una muestra de datos se clasifique como un dato erróneo. Los puntos cuyos residuos son estrictamente iguales al umbral se consideran datos erróneos, se variará con 0.05, 0.1, 0.15, 0.2 y 0.25, con un máximo de iteraciones de 1000.

4.3.3 Modelo de regresión lineal con características polinómicas

La regresión polinómica es una técnica que consiste en usar modelos lineales cuando el conjunto de datos tiene una alta no-linealidad. Por lo tanto, de busca adicionar alguna variable extra que es computado desde las variables existente, para esto se volver a usar el algoritmo **LinearRegression** junto a un generador de matrices de características de las combinaciones de polinomios de las características con un grado menor o igual que el grado especificado, para esto se usa el algoritmo **PolynomialFeatures [5]** de sklearn, en nuestro caso se varió con 2, 3, 4, 5 y 6 exponentes.

4.3.4 Modelo de regresión basada en bosques aleatorios o random forest

La regresión basada en bosques aleatorios se ajusta a una serie de árboles de decisión de clasificación en varias submuestras del conjunto de datos y utiliza el promedio para mejorar la

precisión predictiva y controlar el sobreajuste. Para esto se va a usar el algoritmo **RandomForestRegressor [6]** de la librería sklearn. Además, se configuraron la profundidad máxima del árbol igual número de características que en nuestro caso son 9 características y se variara la cantidad de arbole de 20 en 20 hasta llegar a 100 árboles como máximo

4.3.5 Modelo de regresión MLP

En este apartado se va a hacer uso de una red neuronal simple que nos provee la librería sklearn, dicho algoritmo es **MLPRegresor [7]** la configuración de esta red neuronal consta de dos posibles configuraciones 2, 100 neuronas en 1 capa y 100 neuronas con 2 capas, resto se dejará las configuraciones que vienen por defecto:

- La función de activación va a ser una relu
- El optimizador va a ser un Adam
- El tamaño del lote será seleccionado de forma automática
- La ratio de aprendizaje inicial será de 0.01
- El número máximo de iteraciones será de 100

4.3.6 Modelo de regresión HGB

Esta regresión durante el entrenamiento, el cultivador de árboles aprende en cada punto de división si las muestras con valores faltantes deben ir al hijo izquierdo o derecho, según la ganancia potencial. Al predecir, las muestras con valores faltantes se asignan al hijo izquierdo o derecho en consecuencia. Si no se encontraron valores faltantes para una característica determinada durante el entrenamiento, las muestras con valores faltantes se asignan al hijo que tenga la mayor cantidad de muestras. Para usar este algoritmo se usa el modelo **HistGradientBoostingRegressor [21]** de la librería sklearn

4.3.7 Modelo de regresión de Huber

La regresión Huber funciona optimizando la pérdida cuadrática para las muestras donde y la pérdida absoluta para las muestras. Este modelo posee un parámetro **sigma** que variamos 1.05, 1.25 y 1.45 estos nos aseguran de que, si y aumenta o disminuye por un determinado factor, no es necesario volver a escalar épsilon para lograr la misma robustez. Esto nos aporta la facilidad que

la función de pérdida no se vea muy influenciada por los valores atípicos sin ignorar por completo su efecto. Para usar este algoritmo se usa el modelo **HuberRegressor [20]** de la librería sklearn.

4.3.8 Modelo de regresión de Theil Sen

La regresión Theil Sen, permite calcular soluciones de mínimos cuadrados en subconjuntos variados de las muestras en X. Dependiendo del número de submuestras entre el número de características y muestras conduce a un estimador con un compromiso entre robustez y eficiencia.

Se basa en el principio que el número de soluciones de mínimos cuadrados es "un número de muestras eligen un número de submuestras", puede ser extremadamente grande, pero se puede limitar por medio del hiper parámetro de máxima subpoblación. Además, si se alcanza este límite, los subconjuntos se eligen aleatoriamente. Al final, se calcula la mediana espacial (o mediana L1) de todas las soluciones de mínimos cuadrados. Para usar este algoritmo se usa el modelo **TheilSenRegressor [19]** de la librería sklearn.

4.4 MÉTRICAS

Para calcular las métricas de desempeño se usó diferentes algoritmos presentes en la librería de sklearn, principalmente se tiene el R2 que es el score por defecto que nos arroja los modelos sklearn que se usaron para entrenar las diferentes bases de datos, se tienen 9 métricas de desempeño que son las siguientes:

- **R2 del modelo (R2_MODEL)**: esta métrica es arrojada por el estimador o modelo una vez se realizaba su predicción, esta métrica viene por defecto en los estimadores de sklearn.

Para las siguientes 4 métricas, hay que tener en cuenta que estas provienen de las 2 validaciones cruzadas que se hicieron. La primera validación cruzadas con el score de error cuadrático medio:

- **Error cuadrático medio obtenido con la validación cruzada (CROSS_ECM)**: esta métrica se obtuvo a partir del vector resultado de la validación cruzada, por lo tanto, esta métrica sería la media de los errores cuadrados que se obtuvo en cada validación.

- **Desviación estándar de los errores cuadráticos obtenidos en la validación cruzada (CROSS_ECM_DE)**: esta métrica se obtuvo a partir del vector resultado de la validación cruzada, por lo tanto, esta métrica sería la desviación estándar de los errores cuadrados obtenidos en cada validación.

Y por otro lado tenemos la otra validación cruzada con el score del R2:

- **Coeficiente de determinación o R2 obtenido con la validación cruzada (CROSS_R2)**: esta métrica se obtuvo a partir del vector resultado de la validación cruzada, por lo tanto, esta métrica sería la media de los coeficientes de determinación o los R2 que se obtuvo en cada validación.
- **Desviación estándar de los coeficientes de determinación o R2 obtenido con la validación cruzada (CROSS_R2_DE)**: esta métrica se obtuvo a partir del vector resultado de la validación cruzada, por lo tanto, esta métrica sería la desviación estándar de los coeficientes de determinación o los R2 que se obtuvo en cada validación.

Las otras métricas que se tuvieron en cuenta fueron:

- **Varianza Explicada (VARIANZA_EXPL)**: esta métrica se obtuvo por medio del algoritmo **explained_variance_score [8]** que provee las métricas de la librería sklearn.
- **Error promedio absoluto (MAE):** esta métrica se obtuvo por medio del algoritmo **mean_absolute_error [9]** que provee las métricas de la librería sklearn.

Para las dos siguientes métricas se necesita que tanto las muestras estimadas y las muestras original tengan valores positivos, por lo tanto, se realizó un escalamiento de estos dos conjuntos de datos para evitar errores

- **Pérdida media de regresión de la desviación de Poisson (MAE_POISSON)**: Para calcular esta métrica se usará el algoritmo **mean_poisson_deviance [10]** de la biblioteca sklearn
- **Pérdida media de regresión de la desviación Gamma (MAE_GAMMA)**: Al igual que los anteriores esta métrica se usará el algoritmo **desviación_gamma_media [10]** de la biblioteca sklearn.

5. METODOLOGÍA

5.1 BASELINE

La primera iteración, en resumidas cuentas, se realizó de la siguiente manera:

1. Inicialmente comenzamos analizando la base de datos, observando cuantos datos teníamos, el número de características, el tipo de estas características.
2. Posteriormente se procedió a eliminar las características que fuesen identificadores, en este caso notamos que **CODIGO_DEPARTAMENTO**, **CODIGO_MUNICIPIO** y **NOMBRE_CIENTIFICO** eran identificadores de otras características **DEPARTAMENTO**, **MUNICIPIO** Y **CULTIVO** respectivamente. Además, se eliminaron las columnas **ANIO** Y **PERIODO**, esto debido a que estas características nos indica el año de la cosecha, nos limitará los futuros modelos a saber la fecha de la para poder hacer la predicción, no obstante, esta característica nos servirá para agregar otras características que nos ayuden a mejorar el modelo. Por último, se procedió a eliminar la característica **RENDIMIENTO** que genera redundancia en la base de datos.
3. Debido a que del análisis inicial se observó que podría haber una relación de la variable **CULTIVO** con **SUBGRUPO_CULTIVO** y **SISTEMA_PRODUCTIVO**, debido a que estas variables se identificaron durante el análisis que poseían los mismo valores en algunos registros, se realizó una pruebas de hipótesis con el test chi-cuadrado para saber si eliminamos estas características que tampoco nos iban aportar valor a nuestro modelo, al final el resultado que obtuvo

del test nos dio la confianza de poder estar las características **SUBGRUPO_CULTIVO** y **SISTEMA_PRODUCTIVO** de nuestros datos.

4. Luego se procedió a la imputación de datos faltantes, se identificó en las características **MUNICIPIO**.
5. Una vez que se tenía claro que columnas se usarían para entrenar nuestro modelo, se procedió a hacer la codificación de las características categóricas de nuestra base de datos, se observó que estas características poseían una gran cantidad de categorías, por lo tanto, se optó por un label encoder para codificar estas variables.
6. Se genera un diagrama de distribución para nuestra característica objetivo y se observa que hay una gran cantidad de datos entre 1 a 10 hectáreas sembradas las cuales correspondían a los medianos agricultores, posterior a esta observación se divide la base de datos en 2.
7. Se realiza el escalamiento min-max a las 2 bases de datos
8. Se revisaron los datos atípicos de nuestros datos con un diagrama de cajas, se observa que hay una gran cantidad de datos atípicos en la característica **PRODUCCION**, por lo tanto se procedió a eliminar los datos atípicos de las 2 bases de datos.
9. Una vez tenemos los datos limpios se procede a verificar la importancia de nuestras características, donde se corroboró que la característica que más aportaba al modelo era la característica **AREA_COSECHADA**.
10. Finalmente se procedió a estimar con una regresión lineal simple para ambas bases de datos.

Los resultados obtenidos en ambas bases de datos fueron buenos, debido a que se obtuvieron métricas medianamente buenas. No obstante, las mejores métricas las arrojó la base de datos de los grandes agricultores, pero a pesar de estos resultados, aún no se lograba el objetivo de un MAE menor a 0.01 para ambas bases de datos. Por lo tanto, se optó por usar varios métodos de escalamiento y varios métodos de eliminación de datos atípicos, esto para aumentar la posibilidad de llegar a esa métrica objetivo, estas bases de datos se guardaron y posteriormente se usarán para entrenar cada modelo con estos distintos datos, para que al final podamos tener un gran repertorio de posibilidades.

5.2 VALIDACIÓN

Como se mencionó anteriormente, de la base de datos original se derivaron 2 bases de datos, las cuales son los medianos agricultores y los grandes agricultores, luego de realizar el respectivo escalado de los datos para ambas bases de datos, se dividió los datos en 20% para las pruebas y 80% para el entrenamiento.

Posteriormente se entrenó cada modelo con los datos de entrenamiento, luego se procedió a realizar la predicción con los datos de prueba, con estos datos estimados se procedió a sacar las métricas. Por otro lado, también se procedió a hacer 2 validaciones cruzadas con ayuda de algoritmo **cross_val_score [**16**]** que provee la librería sklearn

5.2.1 Validación cruzada con error cuadrático medio

Para esta validación se utilizó un puntaje de error cuadrático medio, además de que se utilizaron 5 pliegues o dobleces

5.2.2 Validación cruzada con coeficiente de determinación o R2

Para esta validación se utilizó un puntaje del coeficiente de determinación o R2, además de que se utilizaron 10 pliegues o dobleces

5.3 ITERACIONES y EVOLUCIÓN

Una vez se tiene claro cómo se va hacer la validación de los diferentes modelo a entrenar, se procedió a crear diferentes bases de datos con diferentes escalamientos y diferentes métodos de eliminación de datos atípicos, en total se crearon 224 bases de datos, 112 de medianos agricultores y 112 de grandes agricultores, con las diferentes configuraciones que se hicieron en la detección de atípicos, incluyendo la base de datos original, posteriormente se procedió a entrenar el primer modelo con estas 224 bases de datos, adicional a este se agregaron 7 modelos más, los cuales en su mayoría solo se les varió un hiper parámetro, debido a que en esta investigación se busca un modelo inicial simple, posteriormente si el lector lo ve conveniente puede modificar algunos de estos hiper parámetros para obtener mejores resultados.

En cada modelo que se entrenaba, se observaba que había modelos que funcionaban muy bien con un tipo de base de datos escalonada, además que algunos modelos más sencillos ofrecían mejores métricas que otros modelos más complejos. Se obtuvo un buen resultado tanto para los

pequeños agricultores, como para los grandes agricultores, pero en cada iteración se observaba que se obtenía mejores métricas con la base de datos de los grandes agricultores, se probaron modelos tan sencillos como el de regresión lineal y tan complejos como el random forest.

Se pudo evidenciar que el enfoque que se le dio a la investigación mostraba muy buenos resultados, pero cada vez que pasaba una iteración el costo computacional era fluctuante, donde había modelos que no requerían tanto procesamiento y no tomaba mucho tiempo, como había otros que requerían muchas horas de procesamiento.

Al final se obtuvo la siguiente evolución:

- Iteración 0: Exploración de los datos y primer modelo.
- Iteración 1: Algoritmos de escalamiento de datos y algoritmo de eliminación de datos atípicos.
- Iteración 2: Generación de datos medianos agricultores.
- Iteración 3: Generación de datos de grandes agricultores.
- Iteración 4: Modelo de regresión lineal.
- Iteración 5: Modelo de regresión robusta.
- Iteración 6: Modelo de regresión simples con características polinómicas.
- Iteración 7: Modelo de regresión de árboles aleatorios o random forest.
- Iteración 8: Modelo de regresión MLP.
- Iteración 9: Modelo de regresión HGB.
- Iteración 10: Modelo de regresión de Huber.
- Iteración 11: Modelo de regresión de Theil Sen.
- Iteración 12: Selección del modelo.

5.4 HERRAMIENTAS

5.4.1 Entorno de ejecución y desarrollo

Para esta investigación se usó **Colab** que es un servicio de Google Research de notebooks de Jupyter que permite escribir y ejecutar código de Python desde el navegador, dichos archivos pueden estar alojados tanto en repositorios como en la nube y pueden ser ejecutados localmente, colab permite iniciar de forma rápida y replicar el proceso de investigación realizado para esta investigación, debido a que no requiere instalación para usarlo y brinda acceso con recursos computacionales limitados[12].

5.4.2 Librerías de procesamiento

La librería que más se usó tanto por sus algoritmos de procesamiento, como para los modelos fue **skleran** [13] debido a que es una de las librerías más útiles y sólidas para el aprendizaje automático en Python, ya que proporciona una selección de herramientas eficientes para el aprendizaje automático y el modelado estadístico, la regresión y la reducción de la dimensionalidad.

- **LabelEncoder,** para hacer la codificación de las características categóricas.
- **RobustScaler, StandardScaler, MinMaxScaler y Normalizer,** se usaron para escalar los datos antes del procesamiento.
- **LocalOutlierFactor y IsolationForest,** se usaron para la detección de datos atípicos.
- **LinearRegression, RANSACRegressor, PolynomialFeatures, RandomForestRegressor, MLPRegressor, HistGradientBoostingRegressor**, **HuberRegressor** y **TheilSenRegressor**, estos fueron los modelos que se usaron con las diferentes bases de datos
- **Train_test_split,** se usó para la partición de datos de entrenamiento y prueba.
- **Cross_val_score, explained_variance_score, mean_absolute_error, mean_gamma_deviance y mean_poisson_deviance,** estas se usaron para calcular las métricas de desempeño del modelo.

Por otro lado, las librerías para la gestión de los datos fueron:

- **Pandas:** Proporciona estructuras de datos rápidas, flexibles y expresivas, diseñadas para trabajar con datos de manera fácil y realizar análisis prácticos de estos [15].
- **Numpy:** Proporciona un objeto de matriz multidimensional, varios objetos derivados (como matrices y matrices enmascaradas) y una variedad de rutinas para operaciones rápidas en matrices, que incluyen manipulación matemática, lógica, de formas, clasificación, selección, transformadas discretas de Fourier, álgebra lineal básica, operaciones estadísticas básicas, simulación aleatoria y mucho más [14].

5.4.3 Librerías de gráficas

Para graficar los diferentes análisis estadísticos realizados se usaron dos librerías para graficar:

- **Seaborn:** Permite realizar gráficos 2D de matrices. Está basado en MATLAB, pero es independiente de este. Permite el uso de la librería como NumPy, debido a que está construido con esta, esto para proporcionar un buen rendimiento incluso para arreglos grandes, es una librería fácil de usar de ahí su reconocimiento [17].
- **Matplotlib:** Permite la visualización de datos. Proporciona una serie de algoritmos para dibujar gráficos estadísticos atractivos e informativos, además que está basada en la librería matplot, ya explicada anteriormente [18].

6. RESULTADOS

6.1 MÉTRICAS

A continuación, se mostrarán los resultados obtenidos de las diferentes iteraciones que se hicieron con diferentes modelos, con diferentes configuraciones y diferentes bases de datos escalonadas, esto con el fin de identificar la mejor composición para seleccionar el mejor modelo. Se hará bastante énfasis en saber cuáles algoritmos de escalamiento y eliminación de datos atípicos dieron buenos resultados y cuáles no. Además en la parte inferior de las tablas podrá ver la cantidad de bases de datos con las que se entreno ese modelo.

6.1.1 Modelo de regresión lineal

6.1.1.1 Medianos agricultores

Los mejores resultados se obtuvieron con el escalado min-max, además que el algoritmo de eliminado de datos atípicos era el mismo solo variaba la cantidad de estimadores y se observa que el parámetro automático junto con la contaminación de 0.2 arroja los mejores resultados. Por otro lado, vemos que los peores resultados se obtuvieron con el escalado estándar y con el algoritmo LOF solo varia la cantidad de vecinos y la distancia

TABLA XXIV
RESULTADOS DEL MODELO REGRESIÓN LINEAL CON LAS MULTIPLES BASES DE DATOS GENERADAS DE LOS MEDIANOS AGRICULTORES

	NOMBRE	CONFIG	R2_MODEL	CROSS_ECM	CROSS_ECM_DE	CROSS_R2	CROSS_R2_DE	VARIANZA_EXPL	MAE	MAE_POISSON	MAE_GAMMA
111	min_max_isf_400_0_2.csv	NaN	0.835214	-0.000010	1.489862e-06	0.833337	0.034245	0.835248	0.001844	0.040983	1.511002
96	min_max_isf_100_0_2.csv	NaN	0.837917	-0.000010	5.012124e-07	0.837448	0.015195	0.837920	0.001861	0.029109	4.900262
101	min_max_isf_200_0_2.csv	NaN	0.830008	-0.000010	1.360899e-06	0.829567	0.031746	0.830023	0.001882	0.035346	0.966900
106	min_max_isf_300_0_2.csv	NaN	0.835896	-0.000010	6.190353e-07	0.835618	0.029403	0.835902	0.001892	0.036218	1.114131
92	min_max_isf_100_auto.csv	NaN	0.856191	-0.000010	1.047104e-06	0.855848	0.029072	0.856222	0.001902	0.028779	0.900309
97	min_max_isf_200_auto.csv	NaN	0.845894	-0.000012	1.630905e-06	0.844599	0.032388	0.845894	0.001953	0.028980	1.193272
107	min_max_isf_400_auto.csv	NaN	0.841609	-0.000011	1.652493e-06	0.841077	0.028991	0.841611	0.001964	0.029858	1.913480
102	min_max_isf_300_auto.csv	NaN	0.834277	-0.000012	2.325841e-06	0.833573	0.032186	0.834312	0.001975	0.031354	0.981105
95	min_max_isf_100_0_15.csv	NaN	0.841044	-0.000015	2.118156e-06	0.840332	0.028973	0.841046	0.002104	0.031065	12.883957
105	min_max_isf_300_0_15.csv	NaN	0.825556	-0.000016	3.096461e-06	0.825651	0.059128	0.825561	0.002120	0.031210	0.880914
...	...	...	...	...	...	...	...	...	...	...	...
83	estandar_isf_300_0_05.csv	NaN	0.790165	-0.206723	1.798559e-02	0.789355	0.029739	0.790184	0.293812	0.089772	1.557036
26	estandar_lof_minkowski_11.csv	NaN	0.793052	-0.207456	1.060636e-02	0.792610	0.017867	0.793064	0.295768	0.086255	1.434113
25	estandar_lof_euclidean_11.csv	NaN	0.793052	-0.207456	1.060636e-02	0.792610	0.017867	0.793064	0.295768	0.086255	1.434113
24	estandar_lof_manhattan_9.csv	NaN	0.790910	-0.207571	1.154410e-02	0.790482	0.016006	0.790920	0.296520	0.096120	1.477667
22	estandar_lof_euclidean_9.csv	NaN	0.788851	-0.208580	1.858760e-02	0.788243	0.022954	0.788865	0.300212	0.087039	1.436291
23	estandar_lof_minkowski_9.csv	NaN	0.788851	-0.208580	1.858760e-02	0.788243	0.022954	0.788865	0.300212	0.087039	1.436291
17	estandar_lof_minkowski_5.csv	NaN	0.791663	-0.206944	1.117524e-02	0.791460	0.017895	0.791734	0.300280	0.088116	1.426095
16	estandar_lof_euclidean_5.csv	NaN	0.791663	-0.206944	1.117524e-02	0.791460	0.017895	0.791734	0.300280	0.088116	1.426095
18	estandar_lof_manhattan_5.csv	NaN	0.777261	-0.222478	1.226224e-02	0.777220	0.022078	0.777275	0.302322	0.090809	1.363732
21	estandar_lof_manhattan_7.csv	NaN	0.767325	-0.228192	1.077981e-02	0.767638	0.018266	0.767582	0.305223	0.097016	3.657875

112 rows × 11 columns

Como se puede observar este modelo se adapta muy bien a los datos, pero no al punto de generar un sobreajuste, además se observa que la predicción es un poco más alta que los datos originales.

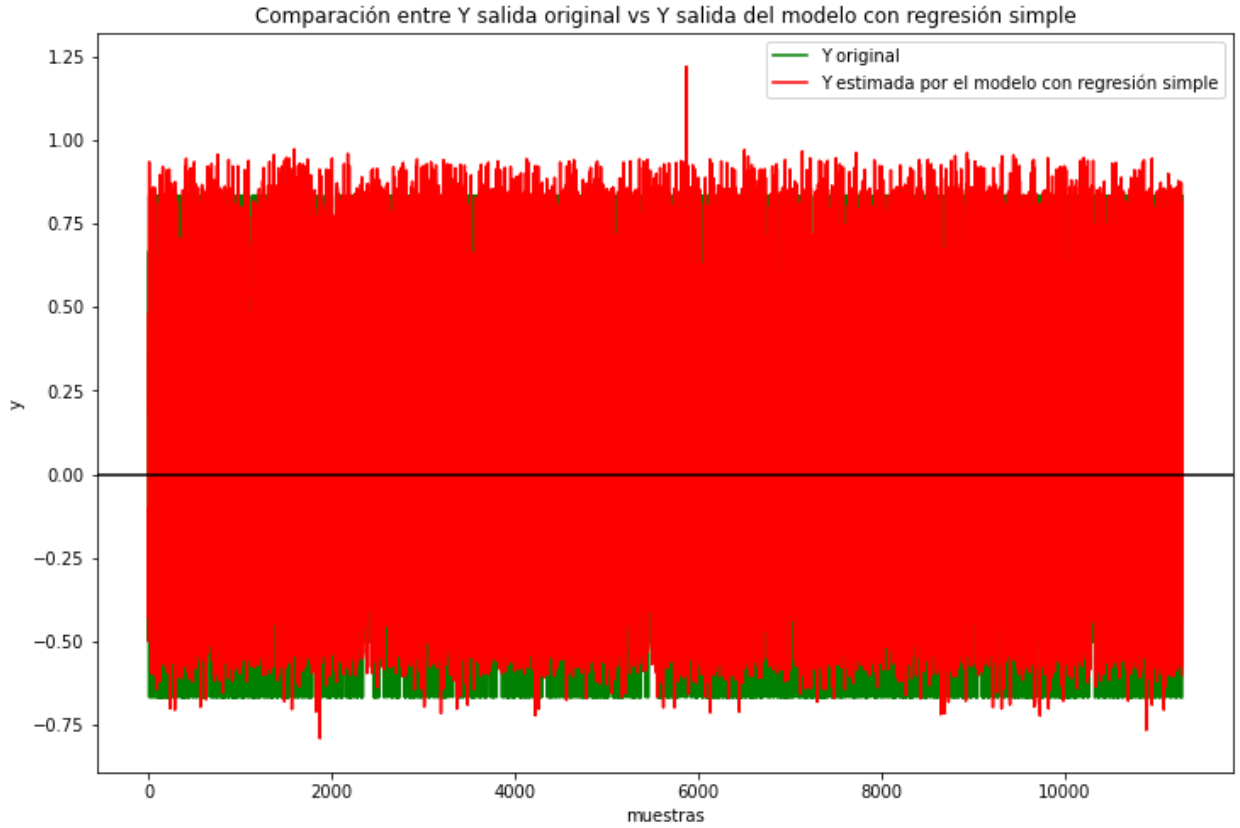

Fig. 22. Comparación de los resultados estimados por el mejor modelo de regresión simple y los resultados verdaderos para los medianos agricultores.

6.1.1.2 Grandes agricultores

Los mejores resultados se obtuvieron con el escalado min-max al igual que en los medianos agricultores, pero a diferencia que el algoritmo de eliminado de datos atípicos es LOF el cual es diferente, pero este se repite en los mejores resultados solo variaba la cantidad de vecinos y distancias. Por otro lado, vemos que los peores resultados se obtuvieron con el escalado robusto y con el algoritmo de atípicos LOF que es similar al obtenido en los mejores resultados.

TABLA XXV
RESULTADOS DEL MODELO REGRESIÓN LINEAL CON LAS MULTIPLES BASES DE DATOS GENERADAS DE LOS GRANDES AGRICULTORES

	NOMBRE	CONFIG	R2_MODEL	CROSS_ECM	CROSS_ECM_DE	CROSS_R2	CROSS_R2_DE	VARIANZA_EXPL	MAE	MAE_POISSON	MAE_GAMMA
37	min_max_lof_euclidean_11.csv	NaN	0.958669	-0.000031	0.000015	0.958005	0.027739	0.958669	0.001388	0.001891	1.115753
38	min_max_lof_minkowski_11.csv	NaN	0.958669	-0.000031	0.000015	0.958005	0.027739	0.958669	0.001388	0.001891	1.115753
29	min_max_lof_minkowski_5.csv	NaN	0.954161	-0.000040	0.000013	0.957298	0.032225	0.954162	0.001388	0.002522	1.304483
28	min_max_lof_euclidean_5.csv	NaN	0.954161	-0.000040	0.000013	0.957298	0.032225	0.954162	0.001388	0.002522	1.304483
30	min_max_lof_manhattan_5.csv	NaN	0.964887	-0.000031	0.000016	0.966605	0.018884	0.964889	0.001391	0.001942	1.261576
34	min_max_lof_euclidean_9.csv	NaN	0.955094	-0.000039	0.000009	0.949360	0.032898	0.955095	0.001403	0.002201	1.171780
35	min_max_lof_minkowski_9.csv	NaN	0.955094	-0.000039	0.000009	0.949360	0.032898	0.955095	0.001403	0.002201	1.171780
39	min_max_lof_manhattan_11.csv	NaN	0.959592	-0.000035	0.000024	0.962075	0.023371	0.959592	0.001409	0.002368	1.238260
32	min_max_lof_minkowski_7.csv	NaN	0.952003	-0.000041	0.000009	0.954734	0.023106	0.952004	0.001419	0.002378	1.268572
31	min_max_lof_euclidean_7.csv	NaN	0.952003	-0.000041	0.000009	0.954734	0.023106	0.952004	0.001419	0.002378	1.268572
...	...	...	...	...	...	...	...	...	...	...	...
8	robusto_lof_minkowski_7.csv	NaN	0.968124	-0.890574	0.196041	0.966255	0.012554	0.968128	0.272143	0.002580	1.415873
7	robusto_lof_euclidean_7.csv	NaN	0.968124	-0.890574	0.196041	0.966255	0.012554	0.968128	0.272143	0.002580	1.415873
11	robusto_lof_minkowski_9.csv	NaN	0.955385	-1.188741	0.471036	0.953457	0.021553	0.955385	0.274170	0.003139	1.477498
10	robusto_lof_euclidean_9.csv	NaN	0.955385	-1.188741	0.471036	0.953457	0.021553	0.955385	0.274170	0.003139	1.477498
6	robusto_lof_manhattan_5.csv	NaN	0.960174	-1.068501	0.483774	0.961516	0.016937	0.960175	0.277581	0.002635	1.322873
12	robusto_lof_manhattan_9.csv	NaN	0.960466	-1.160600	0.290809	0.961381	0.019039	0.960466	0.277708	0.002503	1.290395
0	robusto_original.csv	NaN	0.962931	-1.120656	0.339174	0.962574	0.009854	0.962932	0.283739	0.002848	1.481060
9	robusto_lof_manhattan_7.csv	NaN	0.957610	-1.341678	0.279144	0.956077	0.023732	0.957614	0.285743	0.002224	1.277880
5	robusto_lof_minkowski_5.csv	NaN	0.947016	-1.804409	0.812510	0.948411	0.025603	0.947017	0.286748	0.002200	1.263738
4	robusto_lof_euclidean_5.csv	NaN	0.947016	-1.804409	0.812510	0.948411	0.025603	0.947017	0.286748	0.002200	1.263738

112 rows × 11 columns

Como se puede observar este modelo se adapta muy bien a los datos y la distribución sigue los valores reales muy bien, pero no llegando al punto de ser igual a estos, evitando el sobreajuste de este.

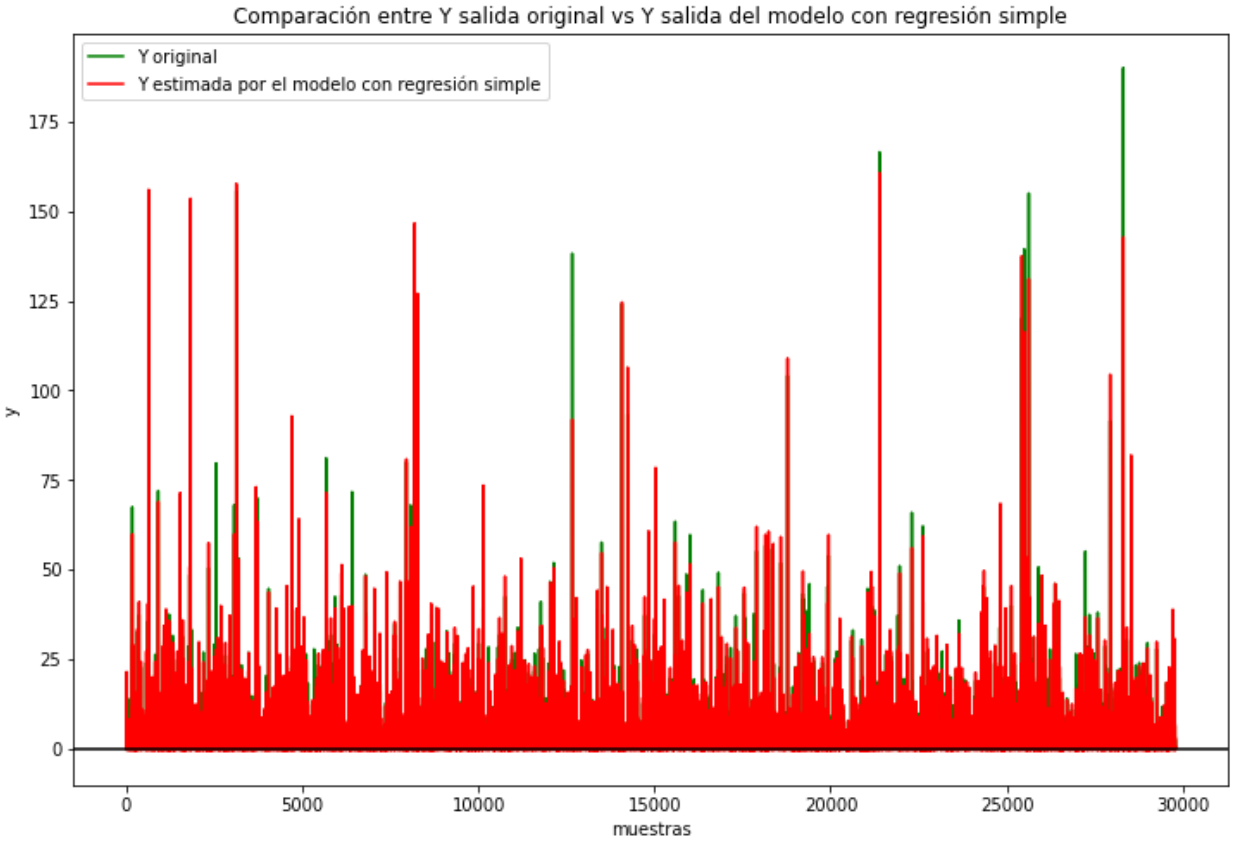

Fig. 23. Comparación de los resultados estimados por el mejor modelo de regresión simple y los resultados verdaderos para los grandes agricultores

6.1.2 Modelo de regresión robusta

6.1.2.1 Medianos agricultores

Estos resultados si varían un poco en comparación a los resultados obtenidos anteriormente, se mantiene a la cabeza el escalamiento min-max, pero en este caso el algoritmo que se usó para eliminar los atípicos fue el ISF con 400 y 100 estimadores, además con una contaminación del 20%. Por otro lado, los peores resultados obtenidos, se dio con el escalamiento estándar con un algoritmo de atípicos LOF que varía entre 5 y 7 vecinos, por último, comparten el parámetro de distancia.

TABLA XXVI
RESULTADOS DEL MODELO REGRESIÓN ROBUSTA CON LAS MULTIPLES BASES DE DATOS GENERADAS DE LOS MEDIANOS AGRICULTORES

	NOMBRE	CONFIG	R2_MODEL	CROSS_ECM	CROSS_ECM_DE	CROSS_R2	CROSS_R2_DE	VARIANZA_EXPL	MAE	MAE_POISSON	MAE_GAMMA
555	min_max_isf_400_0_2.csv	0.05	0.835222	-0.000010	1.489848e-06	0.833289	0.034275	0.835255	0.001844	0.041001	1.526411e+00
559	min_max_isf_400_0_2.csv	0.25	0.835214	-0.000010	1.489862e-06	0.833337	0.034245	0.835248	0.001844	0.040983	1.511002e+00
558	min_max_isf_400_0_2.csv	0.20	0.835214	-0.000010	1.489852e-06	0.833337	0.034245	0.835248	0.001844	0.040983	1.511002e+00
557	min_max_isf_400_0_2.csv	0.15	0.835214	-0.000010	1.489862e-06	0.833337	0.034245	0.835248	0.001844	0.040983	1.511002e+00
556	min_max_isf_400_0_2.csv	0.10	0.835214	-0.000010	1.489862e-06	0.833337	0.034245	0.835248	0.001844	0.040983	1.511002e+00
480	min_max_isf_100_0_2.csv	0.05	0.837924	-0.000010	5.016961e-07	0.837455	0.015207	0.837928	0.001860	0.029248	3.663281e+00
481	min_max_isf_100_0_2.csv	0.10	0.837910	-0.000010	5.012124e-07	0.837448	0.015195	0.837914	0.001861	0.029049	4.991114e+00
482	min_max_isf_100_0_2.csv	0.15	0.837917	-0.000010	5.012124e-07	0.837448	0.015195	0.837920	0.001861	0.029109	4.900262e+00
483	min_max_isf_100_0_2.csv	0.20	0.837917	-0.000010	5.011872e-07	0.837448	0.015195	0.837920	0.001861	0.029109	4.900262e+00
484	min_max_isf_100_0_2.csv	0.25	0.837917	-0.000010	5.012124e-07	0.837448	0.015195	0.837920	0.001861	0.029109	4.900262e+00
...	...	...	...	...	...	...	...	...	...	...	...
91	estandar_lof_manhattan_5.csv	0.10	0.696098	-0.303241	2.041070e-02	0.696204	0.034107	0.749160	0.230102	0.989634	1.864204e+14
90	estandar_lof_manhattan_5.csv	0.05	0.696098	-0.303241	2.041070e-02	0.696204	0.034107	0.749160	0.230102	0.989634	1.864204e+14
94	estandar_lof_manhattan_5.csv	0.25	0.696098	-0.300751	2.330514e-02	0.698439	0.034347	0.749160	0.230102	0.989634	1.864204e+14
93	estandar_lof_manhattan_5.csv	0.20	0.696098	-0.303241	2.041070e-02	0.696204	0.034107	0.749160	0.230102	0.989634	1.864204e+14
92	estandar_lof_manhattan_5.csv	0.15	0.696098	-0.303241	2.041070e-02	0.696204	0.034107	0.749160	0.230102	0.989634	1.864204e+14
109	estandar_lof_manhattan_7.csv	0.25	0.678888	-0.309668	2.057727e-02	0.678626	0.026485	0.736891	0.238825	0.999631	7.900334e+13
106	estandar_lof_manhattan_7.csv	0.10	0.678888	-0.315396	1.592418e-02	0.678626	0.026485	0.736891	0.238825	0.999631	7.900334e+13
105	estandar_lof_manhattan_7.csv	0.05	0.678888	-0.315396	1.592418e-02	0.678626	0.026485	0.736891	0.238825	0.999631	7.900334e+13
107	estandar_lof_manhattan_7.csv	0.15	0.678888	-0.315396	1.592418e-02	0.678626	0.026485	0.736891	0.238825	0.999631	7.900334e+13
108	estandar_lof_manhattan_7.csv	0.20	0.678888	-0.315396	1.592418e-02	0.678626	0.026485	0.736891	0.238825	0.999631	7.900334e+13

560 rows × 11 columns

Como se puede observar en la gráfica de comparación entre la salida original y la estimada, se puede observar que los datos superiores se adaptan muy bien, pero los datos inferiores si se desfasa un poco a comparación de los datos originales, por lo tanto, no se evidencia un sobreajuste de este modelo.

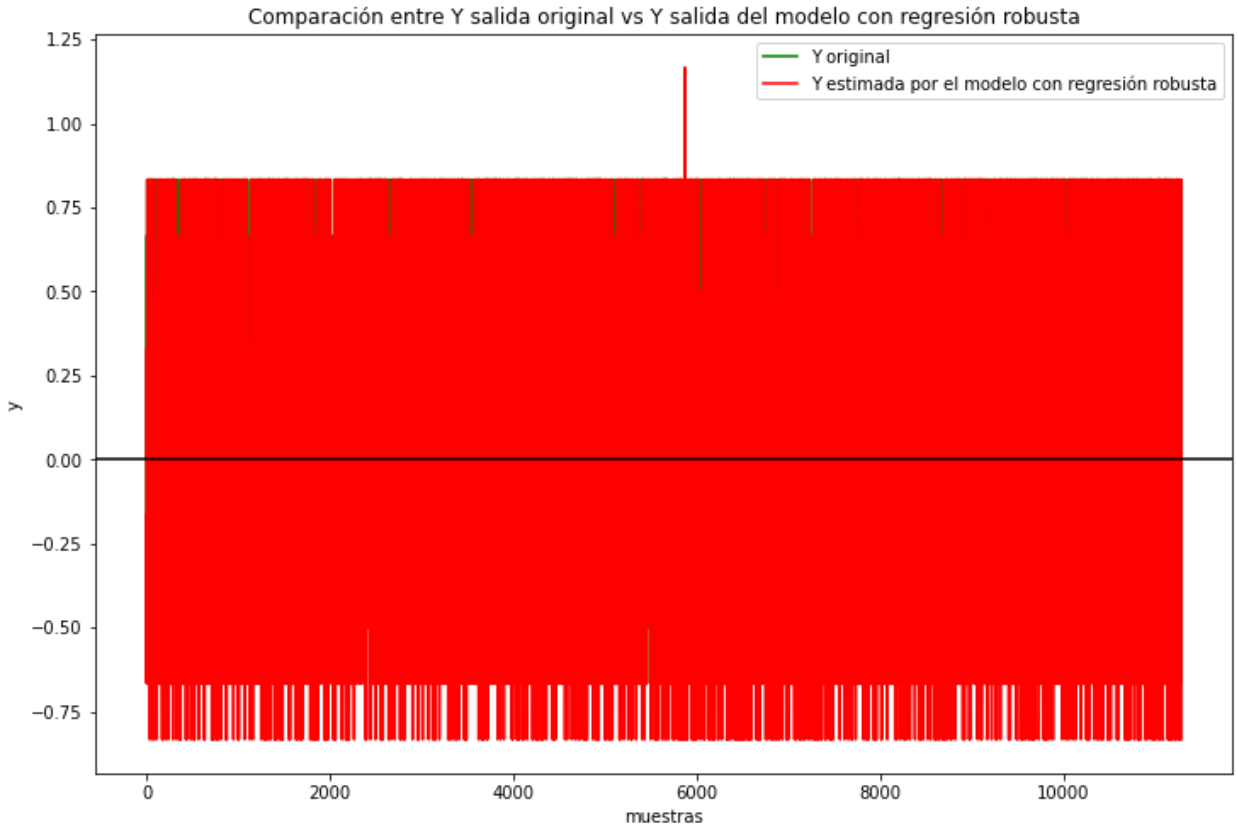

Fig. 24. Comparación de los resultados estimados por el mejor modelo de regresión robusta y los resultados verdaderos para los medianos agricultores

6.1.2.2 Grandes agricultores

Con estos datos, se puede evidenciar que tanto los como mejores como los peores se obtuvieron con el algoritmo de LOF, pero sus vecinos y el factor de contaminación muy variado, la diferencia es el escalamiento usado, en los mejores resultados, se observa que vuelve a estar a la cabeza el escalamiento min-max y para los menos favorables, vuelve a salir el escalamiento robusto.

TABLA XXVII
RESULTADOS DEL MODELO REGRESIÓN ROBUSTA CON LAS MULTIPLES BASES DE DATOS GENERADAS DE LOS GRANDES AGRICULTORES

	NOMBRE	CONFIG	R2_MODEL	CROSS_ECM	CROSS_ECM_DE	CROSS_R2	CROSS_R2_DE	VARIANZA_EXPL	MAE	MAE_POISSON	MAE_GAMMA
150	min_max_lof_manhattan_5.csv	0.05	0.959440	-0.000054	0.000020	0.945680	0.034046	0.959520	0.001238	0.001288	0.960675
190	min_max_lof_minkowski_11.csv	0.05	0.955222	-0.000034	0.000016	0.921915	0.078536	0.955282	0.001254	0.001423	0.894379
155	min_max_lof_euclidean_7.csv	0.05	0.946050	-0.000049	0.000010	0.940867	0.039014	0.946121	0.001292	0.001772	1.026473
186	min_max_lof_euclidean_11.csv	0.10	0.958341	-0.000037	0.000021	0.953837	0.027671	0.958348	0.001309	0.001655	1.007318
151	min_max_lof_manhattan_5.csv	0.10	0.962974	-0.000036	0.000023	0.949611	0.046330	0.962985	0.001314	0.001805	1.206280
192	min_max_lof_minkowski_11.csv	0.15	0.958581	-0.000033	0.000014	0.953623	0.027594	0.958586	0.001318	0.001677	1.017216
160	min_max_lof_minkowski_7.csv	0.05	0.951655	-0.000097	0.000078	0.922971	0.041098	0.951659	0.001319	0.001502	0.908349
140	min_max_lof_euclidean_5.csv	0.05	0.947307	-0.000064	0.000045	0.935955	0.069080	0.947401	0.001323	0.001946	1.082750
193	min_max_lof_minkowski_11.csv	0.20	0.958529	-0.000057	0.000056	0.956859	0.027999	0.958533	0.001325	0.001706	1.028643
195	min_max_lof_manhattan_11.csv	0.05	0.957111	-0.000037	0.000024	0.954139	0.032182	0.957131	0.001325	0.002083	1.122802
...	...	...	...	...	...	...	...	...	...	...	...
66	robusto_lof_euclidean_11.csv	0.10	0.922476	-1.865536	0.946030	0.945866	0.025261	0.922980	0.239143	0.001402	0.933962
23	robusto_lof_euclidean_5.csv	0.20	0.925635	-2.514092	0.985256	0.922491	0.028711	0.926927	0.239175	0.001124	0.862701
47	robusto_lof_manhattan_7.csv	0.15	0.938402	-3.279384	2.637196	0.939921	0.030167	0.939894	0.239183	0.001179	0.835122
25	robusto_lof_minkowski_5.csv	0.05	0.925250	-2.546399	0.850998	0.926399	0.032008	0.926822	0.240046	0.001230	2.558352
49	robusto_lof_manhattan_7.csv	0.25	0.940208	-1.786006	0.520143	0.938581	0.029594	0.941590	0.240396	0.001186	0.919377
20	robusto_lof_euclidean_5.csv	0.05	0.924990	-2.568448	0.868805	0.920082	0.040224	0.926565	0.240759	0.001241	2.804033
57	robusto_lof_minkowski_9.csv	0.15	0.872756	-2.261947	1.633844	0.936741	0.022542	0.874768	0.262924	0.001648	1.005716
59	robusto_lof_minkowski_9.csv	0.25	0.845326	-3.286519	3.510784	0.922567	0.042954	0.846952	0.269172	0.001928	0.909677
29	robusto_lof_minkowski_5.csv	0.25	0.904277	-2.554595	0.940905	0.911761	0.051274	0.904960	0.271095	0.001531	0.910523
54	robusto_lof_euclidean_9.csv	0.25	0.873639	-1.913655	0.645018	0.935656	0.016267	0.873945	0.273382	0.001897	0.834977

560 rows × 11 columns

En esta gráfica podemos observar que los datos estimados, se acomodan muy bien a los datos originales, al igual que el anterior modelo, se ve muy similar la distribución que se generó en este.

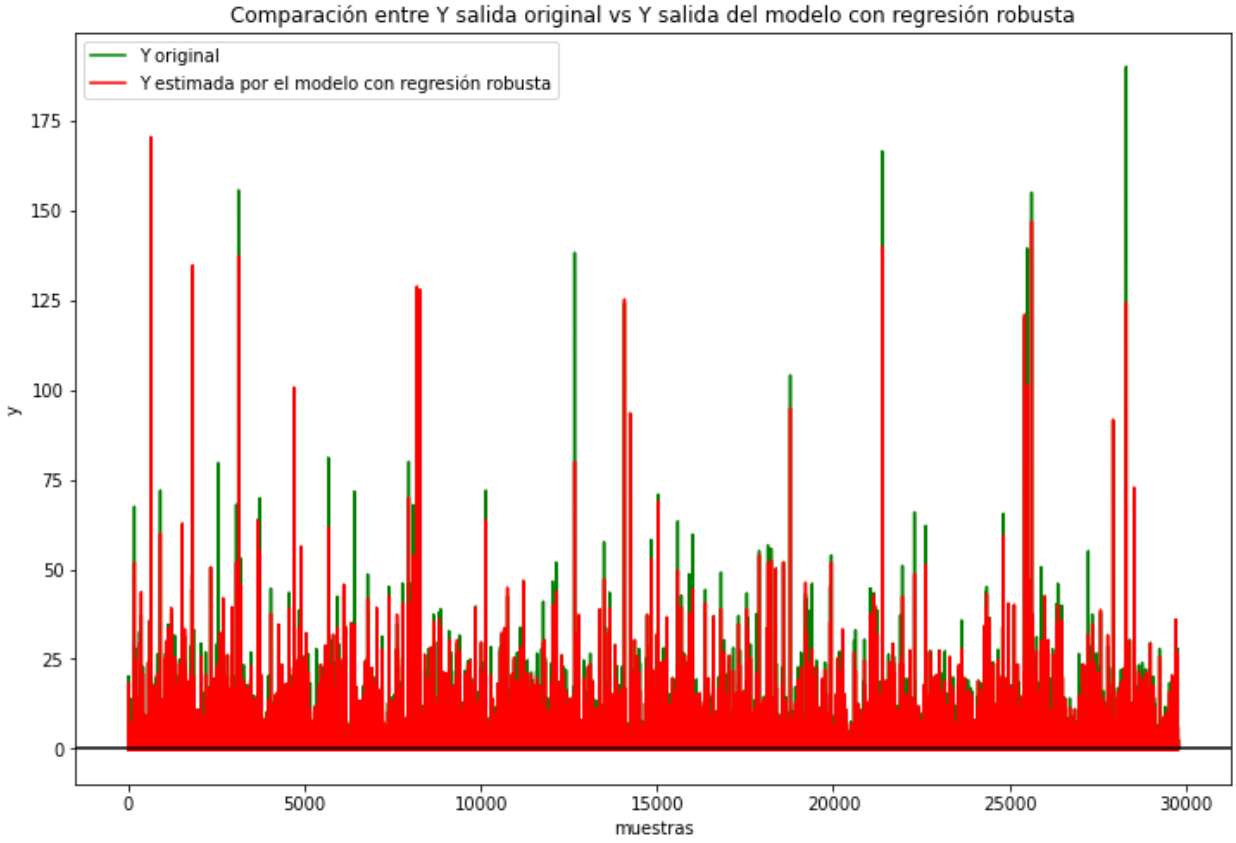

Fig. 25. Comparación de los resultados estimados por el mejor modelo de regresión robusta y los resultados verdaderos para los grandes agricultores

6.1.3 Modelo de regresión lineal con características polinómicas

De estos resultados podemos ver que sigue la tendencia del min-max como el algoritmo de escalamiento que mejores resultados arroja y el algoritmo de detección de atípicos, sigue siendo ISF con sus parámetros variados. Por el de los peores resultados se puede ver que hay 3 algoritmos de escalamiento diferentes y 2 de detección de atípicos diferentes.

6.1.3.1 Medianos agricultores

TABLA XXVIII
RESULTADOS DEL MODELO REGRESIÓN LINEAL CON CARACTERÍSTICAS POLINÓMICAS CON LAS MULTIPLES BASES DE DATOS GENERADAS DE LOS MEDIANOS AGRICULTORES

	NOMBRE	CONFIG	R2_MODEL	CROSS_ECM	CROSS_ECM_DE	CROSS_R2	CROSS_R2_DE	VARIANZA_EXPL	MAE	MAE_POISSON	MAE_GAMMA
555	min_max_isf_400_0_2.csv	2	8.372660e-01	-9.790712e-06	1.414868e-06	8.355645e-01	3.279193e-02	8.373139e-01	1.837465e-03	0.037255	1.013880e+00
480	min_max_isf_100_0_2.csv	2	8.410863e-01	-1.022073e-05	5.583332e-07	8.388218e-01	1.580671e-02	8.410932e-01	1.842470e-03	0.026234	1.365947e+00
507	min_max_isf_200_0_2.csv	4	8.350124e-01	-4.665617e+07	6.090759e+07	-1.245054e+13	3.735161e+13	8.350457e-01	1.849489e-03	0.034060	6.030431e-01
481	min_max_isf_100_0_2.csv	3	8.416550e-01	-6.095483e+01	1.140361e+02	-2.539594e+09	7.594080e+09	8.416550e-01	1.849509e-03	0.024960	1.283484e+00
506	min_max_isf_200_0_2.csv	3	8.354685e-01	-4.214004e+08	8.333034e+08	-2.849313e+14	8.547940e+14	8.354905e-01	1.852149e-03	0.030385	6.206161e-01
505	min_max_isf_200_0_2.csv	2	8.338159e-01	-1.025252e-05	1.242143e-06	8.320054e-01	2.953264e-02	8.338401e-01	1.860559e-03	0.030226	1.187206e+00
509	min_max_isf_200_0_2.csv	6	8.188191e-01	-1.071631e+08	2.115251e+08	-1.162661e+11	3.487963e+11	8.188382e-01	1.860966e-03	0.068945	1.777704e+00
533	min_max_isf_300_0_2.csv	5	8.395176e-01	-3.148231e+13	6.296463e+13	-1.028930e+14	3.086789e+14	8.395193e-01	1.862428e-03	0.029562	5.360296e-01
532	min_max_isf_300_0_2.csv	4	8.412680e-01	-2.473773e+13	4.947546e+13	-2.835195e+17	8.505585e+17	8.412688e-01	1.864109e-03	0.031270	6.133708e-01
531	min_max_isf_300_0_2.csv	3	8.410130e-01	-3.203174e+12	6.406349e+12	-4.753366e+17	1.426010e+18	8.410170e-01	1.866484e-03	0.029439	7.365900e-01
...	...	...	...	...	...	...	...	...	...	...	...
259	max_normalizacion_lof_manhattan_11.csv	6	-1.611428e+10	-3.284518e+11	6.356522e+11	-1.677816e+14	5.032612e+14	-1.611109e+10	4.014482e+01	1.560806	5.879105e+00
234	max_normalizacion_lof_euclidean_9.csv	6	-7.601048e+10	-2.048727e+12	4.093527e+12	-4.534369e+11	1.360311e+12	-7.600230e+10	5.916347e+01	0.842939	9.478179e+05
239	max_normalizacion_lof_minkowski_9.csv	6	-7.601048e+10	-2.048727e+12	4.093527e+12	-4.534369e+11	1.360311e+12	-7.600230e+10	5.916347e+01	0.842939	9.478179e+05
233	max_normalizacion_lof_euclidean_9.csv	5	-2.276666e+11	-1.572569e+11	2.928972e+11	-4.161097e+12	1.248329e+13	-2.276421e+11	1.024384e+02	0.880410	1.649504e+06
238	max_normalizacion_lof_minkowski_9.csv	5	-2.276666e+11	-1.572569e+11	2.928972e+11	-4.161097e+12	1.248329e+13	-2.276421e+11	1.024384e+02	0.880410	1.649504e+06
339	robusto_isf_400_auto.csv	6	-3.192055e+13	-1.601572e+17	1.998436e+17	-1.755001e+17	2.303711e+17	-3.191507e+13	4.536455e+04	0.513277	3.960593e+00
338	robusto_isf_400_auto.csv	5	-9.502013e+16	-1.738312e+17	2.310563e+17	-4.065931e+16	9.755151e+16	-9.498930e+16	2.420142e+06	0.661163	4.121370e+00
414	estandar_isf_300_auto.csv	6	-2.387849e+18	-1.181231e+21	1.255478e+21	-2.818475e+21	6.217579e+21	-2.387400e+18	1.890194e+07	0.665709	2.385551e+00
438	estandar_isf_400_auto.csv	5	-1.043621e+21	-1.083179e+21	1.382417e+21	-1.401458e+21	3.467575e+21	-1.043309e+21	4.861368e+08	0.675724	2.676547e+00
439	estandar_isf_400_auto.csv	6	-3.342044e+21	-6.267935e+21	7.986959e+21	-7.126881e+21	1.187990e+22	-3.341438e+21	9.401873e+08	0.543090	3.994956e+00

560 rows × 11 columns

De esta gráfica podemos observar que los valores estimados varían más que los valores originales, que vemos tienen un límite tanto superior como inferior, vemos que no se logra adaptar muy bien los valores predichos estos se salen del techo superior y muy pocos salen del límite inferior.

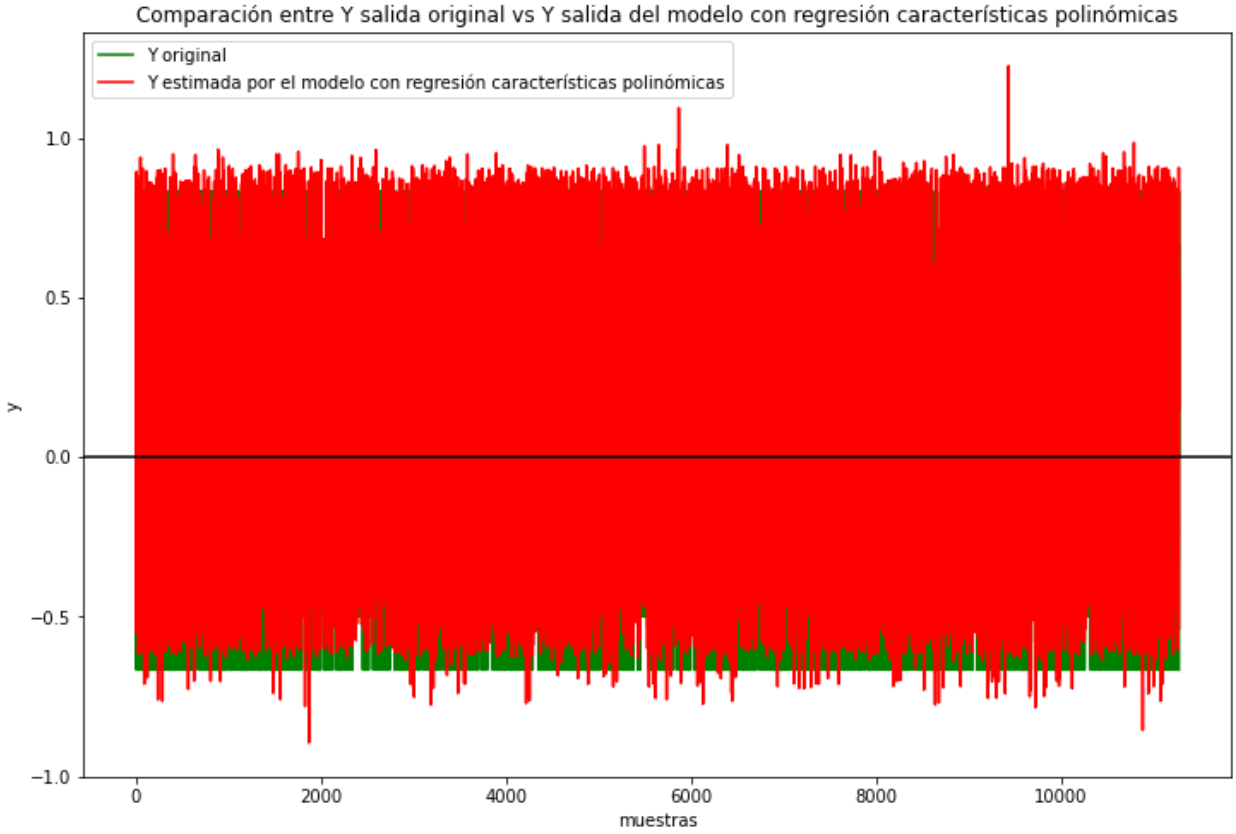

Fig. 26. Comparación de los resultados estimados por el mejor modelo de regresión lineal con características polinómicas y los resultados verdaderos para los medianos agricultores

6.1.3.2 Grandes agricultores

Para los grandes agricultores podemos observar que tanto los mejores resultados como los peores se obtuvieron con el escalamiento min-max, y el algoritmo de detección de atípicos en los medianos es LOF y en los grandes es ISF, los parámetros de estos algoritmos si muestran una variación en cuanto a su configuración.

TABLA XXIX
RESULTADOS DEL MODELO REGRESIÓN LINEAL CON CARACTERÍSTICAS POLINÓMICAS CON LAS MULTIPLES BASES DE DATOS GENERADAS DE LOS GRANDES AGRICULTORES

	NOMBRE	CONFIG	R2_MODEL	CROSS_ECM	CROSS_ECM_DE	CROSS_R2	CROSS_R2_DE	VARIANZA_EXPL	MAE	MAE_POISSON	MAE_GAMMA
145	min_max_lof_minkowski_5.csv	2	9.604706e-01	-3.516857e-05	1.120606e-05	9.617588e-01	2.797403e-02	9.604711e-01	0.001155	0.001444	0.982993
140	min_max_lof_euclidean_5.csv	2	9.604706e-01	-3.516857e-05	1.120606e-05	9.617588e-01	2.797403e-02	9.604711e-01	0.001155	0.001444	0.982993
150	min_max_lof_manhattan_5.csv	2	9.701099e-01	-2.743672e-05	1.388078e-05	9.710866e-01	1.774176e-02	9.701124e-01	0.001156	0.001122	0.915578
189	min_max_lof_euclidean_11.csv	6	9.625704e-01	-4.246745e-05	2.285585e-05	9.473425e-01	2.661404e-02	9.625714e-01	0.001163	0.019614	3.318598
194	min_max_lof_minkowski_11.csv	6	9.625704e-01	-4.246745e-05	2.285585e-05	9.473425e-01	2.661404e-02	9.625714e-01	0.001163	0.019614	3.318598
185	min_max_lof_euclidean_11.csv	2	9.616523e-01	-3.079324e-05	1.536875e-05	9.611762e-01	2.527731e-02	9.616524e-01	0.001166	0.001088	0.727311
190	min_max_lof_minkowski_11.csv	2	9.616523e-01	-3.079324e-05	1.536875e-05	9.611762e-01	2.527731e-02	9.616524e-01	0.001166	0.001088	0.727311
153	min_max_lof_manhattan_5.csv	5	9.711439e-01	-3.564909e-05	1.865987e-05	9.646689e-01	1.860944e-02	9.711464e-01	0.001168	0.016009	3.231103
188	min_max_lof_euclidean_11.csv	5	9.622570e-01	-3.932466e-05	2.517825e-05	9.545999e-01	2.305677e-02	9.622578e-01	0.001169	0.017700	3.199446
193	min_max_lof_minkowski_11.csv	5	9.622570e-01	-3.932466e-05	2.517825e-05	9.545999e-01	2.305677e-02	9.622578e-01	0.001169	0.017700	3.199446
...	...	...	...	...	...	...	...	...	...	...	...
513	min_max_isf_300_auto.csv	5	-1.345321e+07	-1.535129e+12	2.506094e+12	-2.686721e+15	8.060164e+15	-1.345274e+07	3.643713	0.146269	2.167793
509	min_max_isf_200_0_2.csv	6	-1.468176e+07	-4.265908e+10	8.531817e+10	-3.750615e+11	1.125184e+12	-1.468012e+07	5.791233	1.321344	4.160230
514	min_max_isf_300_auto.csv	6	-9.733101e+07	-2.646631e+11	5.007891e+11	-2.679849e+15	7.721992e+15	-9.732756e+07	9.713769	0.152856	2.254314
512	min_max_isf_300_auto.csv	4	-3.973146e+08	-9.100035e+11	1.816549e+12	-4.110323e+12	1.233097e+13	-3.973006e+08	19.629966	0.150665	2.224935
508	min_max_isf_200_0_2.csv	5	-1.617705e+08	-4.983243e+10	9.966485e+10	-1.324205e+11	3.972615e+11	-1.617502e+08	20.378456	1.321657	4.162662
507	min_max_isf_200_0_2.csv	4	-2.458237e+08	-4.705203e+10	9.410406e+10	-2.624615e+10	7.873844e+10	-2.457965e+08	23.548082	3.042529	473879.705298
486	min_max_isf_200_auto.csv	3	-2.986149e+08	-1.681491e+12	3.234327e+12	-4.637424e+13	1.048308e+14	-2.986018e+08	26.556549	0.857615	3.662808
488	min_max_isf_200_auto.csv	5	-2.097475e+10	-1.571203e+11	1.914368e+11	-4.169882e+13	1.176699e+14	-2.097371e+10	234.606031	0.802040	3.584008
487	min_max_isf_200_auto.csv	4	-2.558194e+10	-3.396842e+11	4.844167e+11	-4.729045e+13	1.344526e+14	-2.558103e+10	258.507007	0.701926	3.431543
489	min_max_isf_200_auto.csv	6	-2.355996e+11	-6.616136e+11	7.637889e+11	-9.423520e+12	2.574301e+13	-2.355898e+11	789.440472	0.754523	3.513397

560 rows × 11 columns

Como era de esperar, para los grandes agricultores los modelos se adaptan muy bien tal cual como se ve en la gráfica, donde se observa que hay puntos máximos en donde coincide la estimación con los datos originales, como hay otros puntos máximos que se observan diferentes.

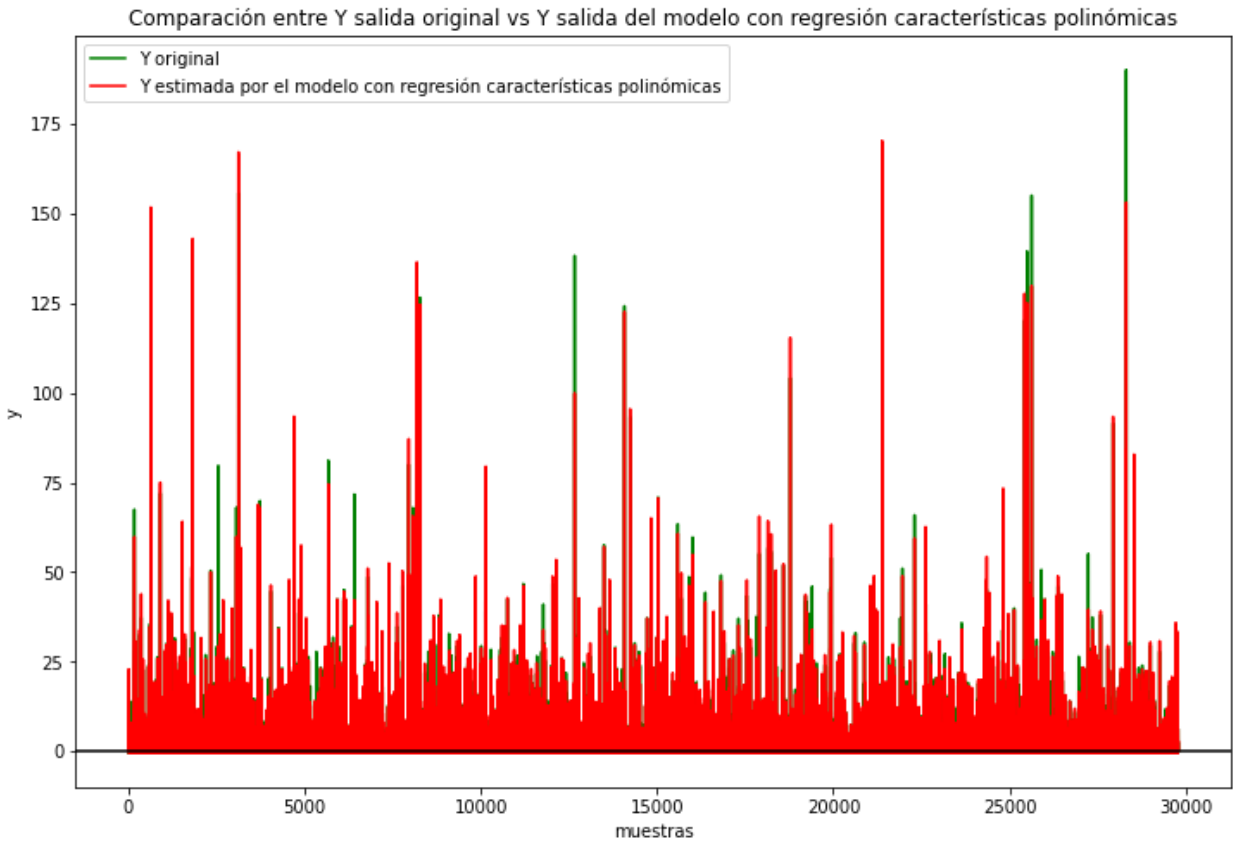

Fig. 27. Comparación de los resultados estimados por el mejor modelo de regresión lineal con características polinómicas y los resultados verdaderos para los grandes agricultores

6.1.4 Modelo de regresión basada en bosques aleatorios

6.1.4.1 Medianos agricultores

Para este modelo se observa que los mejores resultados, muestran el mismo algoritmo de detección de atípicos con parámetros muy similares, la cantidad de estimadores son 100 y 400, pero el factor de contaminación si es el mismo. Por otro lado, los peores resultados vemos un algoritmo LOF en donde si varían sus parámetros, pero el algoritmo de codificación es el mismo al igual que en los mejores resultados.

TABLA XXX
RESULTADOS DEL MODELO REGRESIÓN RANDOM FOREST CON LAS MULTIPLES BASES DE DATOS GENERADAS DE LOS MEDIANOS AGRICULTORES

	NOMBRE	CONFIG	R2_MODEL	CROSS_ECM	CROSS_ECM_DE	CROSS_R2	CROSS_R2_DE	VARIANZA_EXPL	MAE	MAE_POISSON	MAE_GAMMA
558	min_max_isf_400_0_2.csv	80	0.882498	-0.000008	8.668667e-07	0.864358	0.023085	0.882506	0.001543	0.020769	1.043576
559	min_max_isf_400_0_2.csv	100	0.882337	-0.000008	8.568142e-07	0.864652	0.022928	0.882345	0.001543	0.020241	1.029158
557	min_max_isf_400_0_2.csv	60	0.882288	-0.000008	8.657822e-07	0.863846	0.022990	0.882296	0.001544	0.020227	0.982944
556	min_max_isf_400_0_2.csv	40	0.881761	-0.000008	9.113319e-07	0.863417	0.022766	0.881768	0.001545	0.021096	0.813888
555	min_max_isf_400_0_2.csv	20	0.880075	-0.000008	8.334840e-07	0.862385	0.023383	0.880085	0.001552	0.020194	0.739812
481	min_max_isf_100_0_2.csv	40	0.880696	-0.000009	5.973486e-07	0.866121	0.018015	0.880711	0.001560	0.014953	0.464127
482	min_max_isf_100_0_2.csv	60	0.881093	-0.000009	5.573987e-07	0.866124	0.017698	0.881105	0.001560	0.014947	0.574869
480	min_max_isf_100_0_2.csv	20	0.879832	-0.000009	6.050863e-07	0.865156	0.017566	0.879859	0.001560	0.015148	0.818675
483	min_max_isf_100_0_2.csv	80	0.881188	-0.000009	5.311951e-07	0.866183	0.017751	0.881203	0.001561	0.014963	0.464565
484	min_max_isf_100_0_2.csv	100	0.881102	-0.000009	5.412832e-07	0.866089	0.017615	0.881116	0.001562	0.014975	0.418477
...	...	...	...	...	...	...	...	...	...	...	...
117	estandar_lof_minkowski_9.csv	60	0.839753	-0.167934	1.721422e-02	0.830810	0.020082	0.839778	0.251491	0.053421	1.314160
111	estandar_lof_euclidean_9.csv	40	0.839598	-0.168598	1.680626e-02	0.830722	0.019497	0.839623	0.251686	0.053268	1.242752
116	estandar_lof_minkowski_9.csv	40	0.839598	-0.168598	1.680626e-02	0.830722	0.019497	0.839623	0.251686	0.053268	1.242752
115	estandar_lof_minkowski_9.csv	20	0.839050	-0.170185	1.789965e-02	0.829512	0.019620	0.839078	0.251895	0.054823	2.170982
110	estandar_lof_euclidean_9.csv	20	0.839050	-0.170185	1.789965e-02	0.829512	0.019620	0.839078	0.251895	0.054823	2.170982
109	estandar_lof_manhattan_7.csv	100	0.827893	-0.177003	1.161595e-02	0.821892	0.013642	0.828064	0.252857	0.055585	1.109008
107	estandar_lof_manhattan_7.csv	60	0.827649	-0.177331	1.164588e-02	0.821972	0.013453	0.827815	0.252957	0.055710	1.115455
108	estandar_lof_manhattan_7.csv	80	0.827671	-0.177076	1.160555e-02	0.822036	0.013439	0.827841	0.252964	0.055752	1.119063
106	estandar_lof_manhattan_7.csv	40	0.827556	-0.178063	1.164940e-02	0.821808	0.013312	0.827720	0.252965	0.055858	1.116247
105	estandar_lof_manhattan_7.csv	20	0.826756	-0.179027	1.156151e-02	0.820222	0.013559	0.826921	0.253027	0.056323	1.126162

560 rows × 11 columns

Como se puede observar, se tuvo muy buen desempeño de este modelo, hasta el momento es el que modelo que no se sale del límite superior y tanto los límites de los datos estimados como los originales coinciden, por otro lado, el límite inferior los datos estimados les falto un poco para alcanzar al límite de los datos originales, con esto se observa que hay si cierta diferencia y nos asegura que el modelo no aprendió el comportamiento de estos.

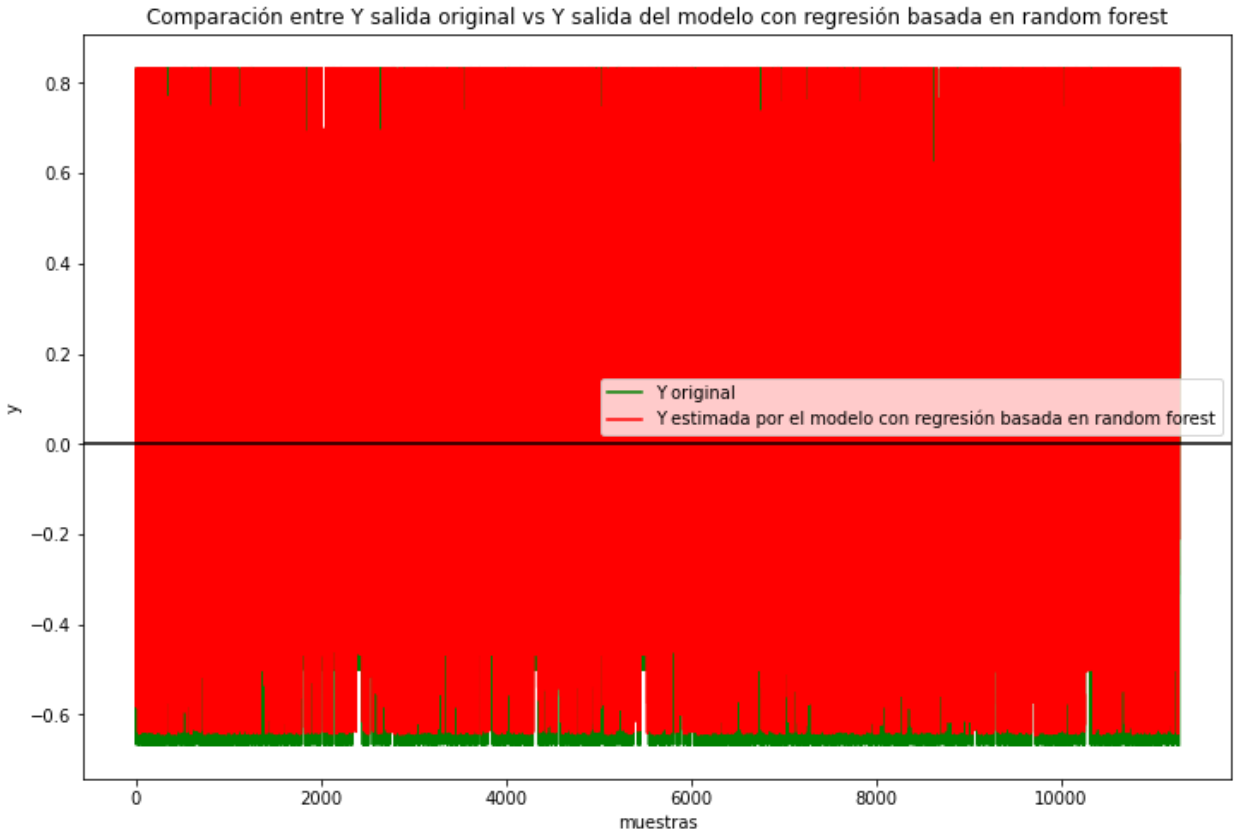

Fig. 28. Comparación de los resultados estimados por el mejor modelo de regresión random forest con características polinómicas y los resultados verdaderos para los medianos agricultores

6.1.4.2 Grandes agricultores

Con esta base de datos también se observa que, tanto para los buenos como los malos resultados, se usó el mismo algoritmo de detección de atípicos, donde el parámetro de vecinos es igual, pero la distancia si varía, con respecto al algoritmo de escalado vemos que los mejores resultados se siguen obteniendo con el min-max y para los peores es el robusto, escalamiento que ya se había visto en los peores resultados.

TABLA XXXI
RESULTADOS DEL MODELO REGRESIÓN RANDOM FOREST CON LAS MULTIPLES BASES DE DATOS GENERADAS DE LOS GRANDES AGRICULTORES

	NOMBRE	CONFIG	R2_MODEL	CROSS_ECM	CROSS_ECM_DE	CROSS_R2	CROSS_R2_DE	VARIANZA_EXPL	MAE	MAE_POISSON	MAE_GAMMA
154	min_max_lof_manhattan_5.csv	100	0.973404	-0.000041	0.000023	0.959706	0.028319	0.973404	0.001003	0.000754	14.667231
152	min_max_lof_manhattan_5.csv	60	0.974021	-0.000041	0.000023	0.960722	0.027018	0.974021	0.001004	0.000606	0.937795
153	min_max_lof_manhattan_5.csv	80	0.973406	-0.000041	0.000023	0.959814	0.028093	0.973406	0.001004	0.000750	11.550295
151	min_max_lof_manhattan_5.csv	40	0.973053	-0.000041	0.000023	0.960409	0.027950	0.973054	0.001011	0.000614	1.128173
150	min_max_lof_manhattan_5.csv	20	0.973313	-0.000042	0.000023	0.959582	0.029106	0.973313	0.001012	0.000605	1.065993
142	min_max_lof_euclidean_5.csv	60	0.964394	-0.000047	0.000018	0.951149	0.033550	0.964395	0.001022	0.001050	3.456511
147	min_max_lof_minkowski_5.csv	60	0.964394	-0.000047	0.000018	0.951149	0.033550	0.964395	0.001022	0.001050	3.456511
141	min_max_lof_euclidean_5.csv	40	0.965167	-0.000049	0.000018	0.949220	0.033917	0.965169	0.001023	0.001012	1.558455
146	min_max_lof_minkowski_5.csv	40	0.965167	-0.000049	0.000018	0.949220	0.033917	0.965169	0.001023	0.001012	1.558455
143	min_max_lof_euclidean_5.csv	80	0.964212	-0.000046	0.000017	0.950716	0.033611	0.964213	0.001024	0.001051	4.482694
...	...	...	...	...	...	...	...	...	...	...	...
29	robusto_lof_minkowski_5.csv	100	0.958531	-1.657499	0.726593	0.951164	0.024498	0.958532	0.217678	0.000936	0.955538
24	robusto_lof_euclidean_5.csv	100	0.958531	-1.657499	0.726593	0.951164	0.024498	0.958532	0.217678	0.000936	0.955538
23	robusto_lof_euclidean_5.csv	80	0.959208	-1.644443	0.708371	0.951232	0.024800	0.959209	0.217780	0.000924	0.730981
28	robusto_lof_minkowski_5.csv	80	0.959208	-1.644443	0.708371	0.951232	0.024800	0.959209	0.217780	0.000924	0.730981
22	robusto_lof_euclidean_5.csv	60	0.960061	-1.679122	0.704743	0.950189	0.025933	0.960061	0.217885	0.000922	0.766694
27	robusto_lof_minkowski_5.csv	60	0.960061	-1.679122	0.704743	0.950189	0.025933	0.960061	0.217885	0.000922	0.766694
21	robusto_lof_euclidean_5.csv	40	0.959594	-1.688466	0.721750	0.948940	0.028294	0.959594	0.219243	0.001034	0.733797
26	robusto_lof_minkowski_5.csv	40	0.959594	-1.688466	0.721750	0.948940	0.028294	0.959594	0.219243	0.001034	0.733797
20	robusto_lof_euclidean_5.csv	20	0.958604	-1.777035	0.653193	0.948532	0.025687	0.958604	0.221084	0.001132	0.695160
25	robusto_lof_minkowski_5.csv	20	0.958604	-1.777035	0.653193	0.948532	0.025687	0.958604	0.221084	0.001132	0.695160

560 rows × 11 columns

Con respecto a la gráfica se observa que los valores máximos obtenidos en la predicción quedan por debajo de los valores originales, por otro lado, si vemos los valores mínimos estos se ven muy similares entre los reales y estimados, con una pequeña diferencia, pero no sobrepasando el valor original.

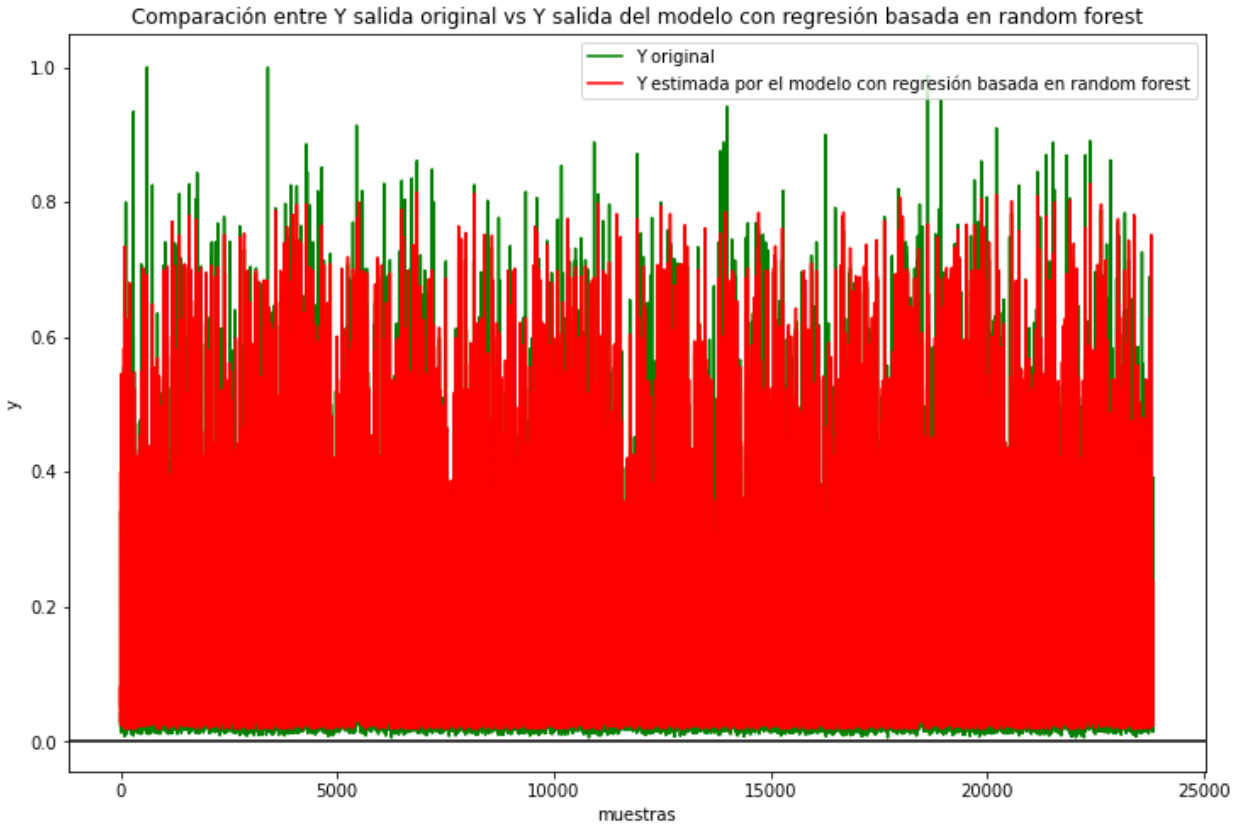

Fig. 29. Comparación de los resultados estimados por el mejor modelo de regresión random forest con características polinómicas y los resultados verdaderos para los grandes agricultores

6.1.5 Modelo de regresión MLP

6.1.5.1 Medianos agricultores

En los resultados del modelo con estos datos, se puede observar que tanto para los mejores resultados como para los peores el algoritmo de detección de atípicos que más prevalece es ISF con los parámetros de contaminación y estimadores muy aleatorios, por otro el min-max también prevalece aquí como el algoritmo de escalamiento que mejor resultados arrojo y estándar sigue apareciendo en los peores resultados.

TABLA XXXII
RESULTADOS DEL MODELO REGRESIÓN MLP CON LAS MULTIPLES BASES DE DATOS GENERADAS DE LOS MEDIANOS AGRICULTORES

	NOMBRE	CONFIG	R2_MODEL	CROSS_ECM	CROSS_ECM_DE	CROSS_R2	CROSS_R2_DE	VARIANZA_EXPL	MAE	MAE_POISSON	MAE_GAMMA
223	min_max_isf_400_0_2.csv	100_2	6.939605e-01	-0.000046	0.000009	0.288577	0.242406	0.702056	0.001831	0.117531	2.025766
215	min_max_isf_400_auto.csv	100_2	8.362265e-01	-0.000083	0.000035	0.242489	0.502493	0.836323	0.001961	0.033603	2.054826
213	min_max_isf_300_0_2.csv	100_2	7.036153e-01	-0.000036	0.000003	0.447683	0.213386	0.732757	0.001989	0.080320	1.839866
209	min_max_isf_300_0_1.csv	100_2	8.412787e-01	-0.000045	0.000015	0.646366	0.086313	0.854407	0.002066	0.035344	0.612263
195	min_max_isf_200_auto.csv	100_2	8.318973e-01	-0.000094	0.000015	0.236516	0.491039	0.832194	0.002096	0.062701	1.041028
191	min_max_isf_100_0_15.csv	100_2	7.935315e-01	-0.000047	0.000037	0.625288	0.368686	0.814775	0.002136	0.080176	1.095509
221	min_max_isf_400_0_15.csv	100_2	7.887023e-01	-0.000066	0.000047	0.312299	0.378229	0.789967	0.002140	0.139559	1.319247
205	min_max_isf_300_auto.csv	100_2	8.068266e-01	-0.000082	0.000012	-0.037703	0.283378	0.806981	0.002159	0.074654	1.135530
193	min_max_isf_100_0_2.csv	100_2	7.380107e-01	-0.000032	0.000012	0.389059	0.223926	0.738802	0.002230	0.078718	0.942941
219	min_max_isf_400_0_1.csv	100_2	8.066107e-01	-0.000033	0.000004	0.751666	0.041356	0.827526	0.002340	0.048327	1.056697
...	...	...	...	...	...	...	...	...	...	...	...
0	robusto_original.csv	100_1	-7.727405e-04	-0.238325	0.004122	-0.001163	0.001381	0.000000	0.418297	7.314308	9434.900307
22	robusto_lof_minkowski_9.csv	100_1	-1.108382e-03	-0.238162	0.004428	-0.000749	0.000860	0.000000	0.418949	7.374881	9506.529184
20	robusto_lof_euclidean_9.csv	100_1	-1.108382e-03	-0.238162	0.004428	-0.000749	0.000860	0.000000	0.418949	7.374881	9506.529184
174	estandar_isf_400_auto.csv	100_1	-1.374030e-08	-0.756979	0.023064	-0.000714	0.000526	0.000000	0.717937	6.192055	8143.584696
154	estandar_isf_200_auto.csv	100_1	-2.507609e-05	-0.759811	0.023270	-0.003353	0.005567	0.000000	0.722779	6.209492	8164.034278
164	estandar_isf_300_auto.csv	100_1	-2.638067e-04	-0.775084	0.021106	0.245999	0.378878	0.000000	0.730427	6.242571	8199.368617
144	estandar_isf_100_auto.csv	100_1	-5.894062e-04	-0.793554	0.018002	0.499844	0.408700	0.000000	0.736290	6.530343	8550.333356
172	estandar_isf_300_0_2.csv	100_1	-8.057494e-05	-0.304184	0.293320	0.078144	0.237938	0.000000	0.782944	6.687987	8707.706916
162	estandar_isf_200_0_2.csv	100_1	-1.852528e-04	-0.895692	0.019892	-0.002058	0.002195	0.000000	0.785890	6.741688	8772.382591
182	estandar_isf_400_0_2.csv	100_1	-5.718170e-07	-0.748556	0.287006	0.168953	0.339528	0.000000	0.788909	6.684520	8700.803617

224 rows × 11 columns

En la siguiente gráfica se muestra como el límite superior de los datos estimados coincide con el límite superior de los datos originales, pero en los límites inferiores no se observa este comportamiento, se observa que los valores estimados varían en la parte inferior, pasando el límite inferior que tiene los datos originales.

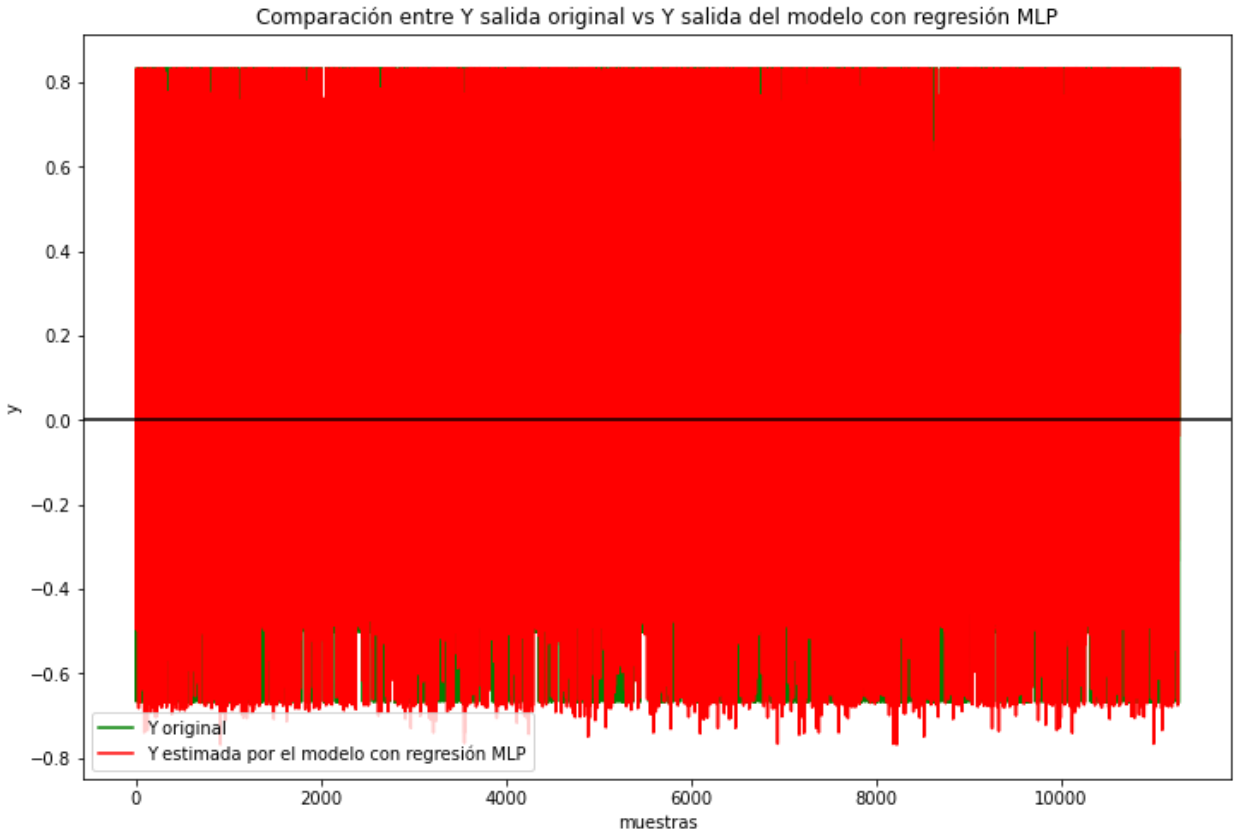

Fig. 30. Comparación de los resultados estimados por el mejor modelo de regresión MLP con características polinómicas y los resultados verdaderos para los medianos agricultores

6.1.5.2 Grandes agricultores

Acá podemos observar que la configuración de 2 capas y 100 neuronas tuvo un mejor desempeño a las que solo tenían 1 capa con 100 neuronas, además que podemos observar que en los peores resultados aparece la base de datos original, sin aplicarle ningún tipo de eliminación de datos atípicos, pero si con un escalamiento robusto.

TABLA XXXIII
RESULTADOS DEL MODELO REGRESIÓN MLP CON LAS MULTIPLES BASES DE DATOS GENERADAS DE LOS GRANDES AGRICULTORES

	NOMBRE	CONFIG	R2_MODEL	CROSS_ECM	CROSS_ECM_DE	CROSS_R2	CROSS_R2_DE	VARIANZA_EXPL	MAE	MAE_POISSON	MAE_GAMMA
5	min_max_original.csv	100_2	0.966476	-0.000034	0.000010	0.958612	0.011125	0.966929	0.001665	0.008123	2.429899
57	min_max_lof_euclidean_5.csv	100_2	0.957178	-0.000048	0.000015	0.952129	0.030739	0.957374	0.001735	0.008616	2.434258
59	min_max_lof_minkowski_5.csv	100_2	0.957178	-0.000048	0.000015	0.952129	0.030739	0.957374	0.001735	0.008616	2.434258
71	min_max_lof_minkowski_9.csv	100_2	0.953891	-0.000041	0.000010	0.936089	0.031918	0.954609	0.002300	0.008761	2.348388
69	min_max_lof_euclidean_9.csv	100_2	0.953891	-0.000041	0.000010	0.936089	0.031918	0.954609	0.002300	0.008761	2.348388
79	min_max_lof_manhattan_11.csv	100_2	0.955665	-0.000047	0.000022	0.959485	0.021738	0.959003	0.002625	0.006731	2.141868
61	min_max_lof_manhattan_5.csv	100_2	0.948000	-0.000036	0.000016	0.956037	0.014300	0.965530	0.004125	0.008946	2.536267
77	min_max_lof_minkowski_11.csv	100_2	0.939281	-0.000040	0.000014	0.946446	0.025950	0.960826	0.004612	0.005117	1.866019
75	min_max_lof_euclidean_11.csv	100_2	0.939281	-0.000040	0.000014	0.946446	0.025950	0.960826	0.004612	0.005117	1.866019
63	min_max_lof_euclidean_7.csv	100_2	0.925879	-0.000044	0.000009	0.952284	0.019747	0.953829	0.005521	0.006031	2.083448
...	...	...	...	...	...	...	...	...	...	...	...
28	robusto_lof_minkowski_11.csv	100_1	-0.000009	-14.363447	10.818188	0.813099	0.281518	0.000000	2.007808	0.087041	166.369389
26	robusto_lof_euclidean_11.csv	100_1	-0.000009	-14.363447	10.818188	0.813099	0.281518	0.000000	2.007808	0.087041	166.369389
16	robusto_lof_minkowski_7.csv	100_1	-0.000026	-12.186312	13.971316	0.869916	0.290604	0.000000	2.008715	0.089419	170.751115
14	robusto_lof_euclidean_7.csv	100_1	-0.000026	-12.186312	13.971316	0.869916	0.290604	0.000000	2.008715	0.089419	170.751115
30	robusto_lof_manhattan_11.csv	100_1	-0.000222	-1.636174	0.953316	0.946735	0.036045	0.000000	2.021278	0.089237	169.617023
0	robusto_original.csv	100_1	-0.000045	-15.401192	12.728425	0.451721	0.455251	0.000000	2.034700	0.083298	159.421171
12	robusto_lof_manhattan_5.csv	100_1	-0.000072	-9.465555	8.776055	0.588518	0.362491	0.000000	2.048502	0.096112	182.143515
10	robusto_lof_minkowski_5.csv	100_1	-0.000070	-5.666701	1.698621	0.518459	0.426369	0.000000	2.096645	0.089384	169.347995
8	robusto_lof_euclidean_5.csv	100_1	-0.000070	-5.666701	1.698621	0.518459	0.426369	0.000000	2.096645	0.089384	169.347995
18	robusto_lof_manhattan_7.csv	100_1	-0.000049	-6.964469	9.527800	0.864583	0.092664	0.000000	2.147730	0.093464	176.435953

224 rows × 11 columns

Por el lado de la gráfica, podemos observar que los datos estimados llegan hasta un límite, al contrario de los datos originales que sí tiene unos picos altos que sobrepasan el límite superior de los datos estimados, en cuanto al límite inferior, podemos observar que estos sí coinciden

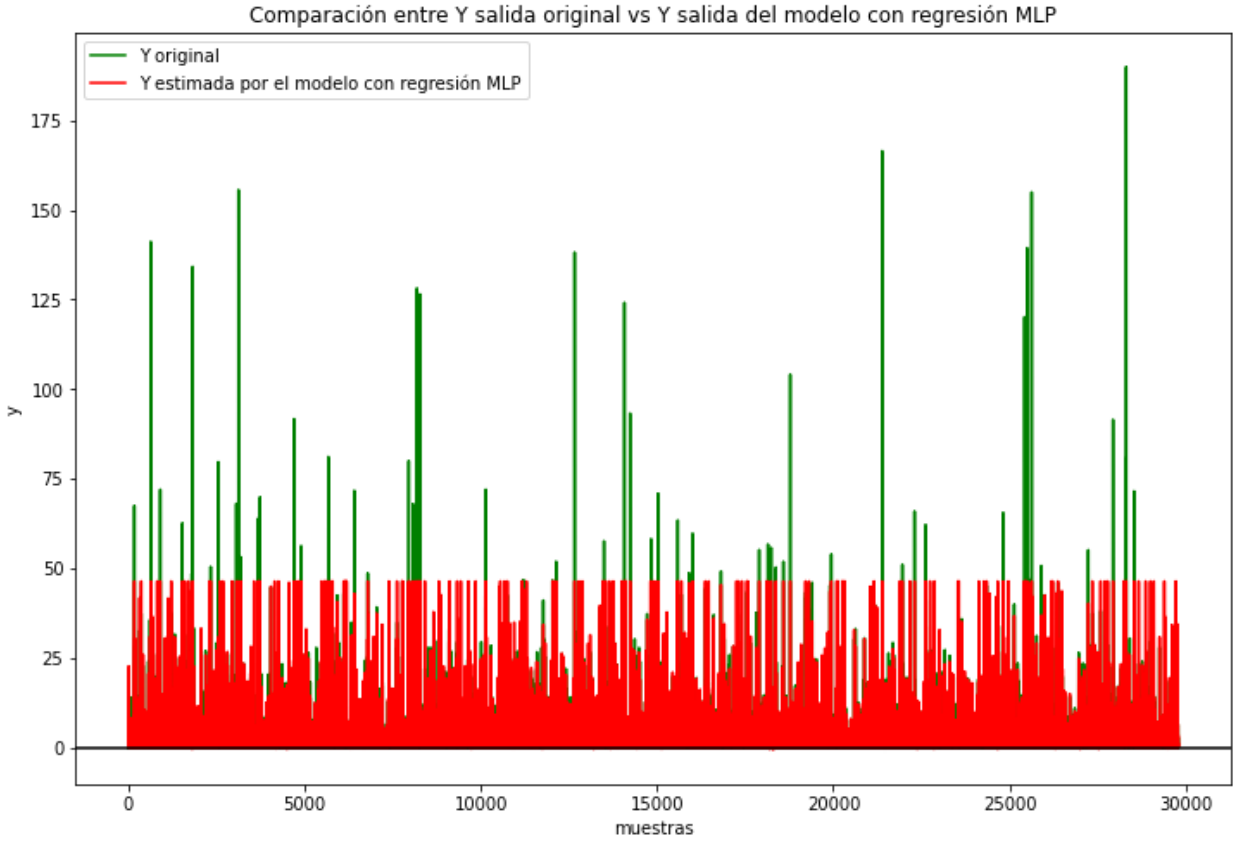

Fig. 31. Comparación de los resultados estimados por el mejor modelo de regresión MLP con características polinómicas y los resultados verdaderos para los grandes agricultores

6.1.6 Modelo de regresión HGB.

6.1.6.1 Medianos agricultores

Con esta base de datos observamos que las bases de datos que obtuvieron los mejores resultados repiten mucho los resultados obtenidos con anteriores bases de datos, el mejor escalamiento sigue siendo min-max y para los peores vuelve aparecer el estándar, además para los peores que se observa 2 algoritmos diferentes es eliminación de datos atípicos, con configuraciones aleatorias sin observar una configuración que se repite.

TABLA XXXIV
RESULTADOS DEL MODELO REGRESIÓN HGB CON LAS MULTIPLES BASES DE DATOS GENERADAS DE LOS MEDIANOS AGRICULTORES

	NOMBRE	CONFIG	R2_MODEL	CROSS_ECM	CROSS_ECM_DE	CROSS_R2	CROSS_R2_DE	VARIANZA_EXPL	MAE	MAE_POISSON	MAE_GAMMA
111	min_max_isf_400_0_2.csv	NaN	0.880080	-0.000008	8.568620e-07	0.871315	0.021513	0.880101	0.001582	0.029520	0.569178
96	min_max_isf_100_0_2.csv	NaN	0.878271	-0.000008	4.795193e-07	0.872229	0.015928	0.878276	0.001584	0.019978	0.288541
106	min_max_isf_300_0_2.csv	NaN	0.879503	-0.000008	5.771161e-07	0.875492	0.022458	0.879507	0.001617	0.026142	0.293535
101	min_max_isf_200_0_2.csv	NaN	0.872375	-0.000008	8.916480e-07	0.865223	0.017613	0.872408	0.001625	0.025078	0.607318
92	min_max_isf_100_auto.csv	NaN	0.888348	-0.000008	7.931310e-07	0.886155	0.019601	0.888362	0.001660	0.024108	0.558355
97	min_max_isf_200_auto.csv	NaN	0.886602	-0.000009	6.970697e-07	0.882471	0.018933	0.886603	0.001683	0.020640	0.841949
107	min_max_isf_400_auto.csv	NaN	0.887182	-0.000009	8.729417e-07	0.880747	0.017084	0.887182	0.001685	0.018519	0.580262
102	min_max_isf_300_auto.csv	NaN	0.882786	-0.000009	1.132206e-06	0.874983	0.014487	0.882810	0.001702	0.016578	0.223338
105	min_max_isf_300_0_15.csv	NaN	0.875585	-0.000012	1.775777e-06	0.863194	0.035649	0.875614	0.001800	0.025705	0.283215
95	min_max_isf_100_0_15.csv	NaN	0.883369	-0.000012	1.637341e-06	0.874828	0.021850	0.883369	0.001805	0.023613	0.490046
...	...	...	...	...	...	...	...	...	...	...	...
25	estandar_lof_euclidean_11.csv	NaN	0.845870	-0.165354	1.142401e-02	0.834946	0.016345	0.845872	0.244087	0.052720	1.145482
16	estandar_lof_euclidean_5.csv	NaN	0.849981	-0.163432	5.128985e-03	0.836399	0.012998	0.850028	0.244383	0.053069	1.184391
17	estandar_lof_minkowski_5.csv	NaN	0.849981	-0.163432	5.128985e-03	0.836399	0.012998	0.850028	0.244383	0.053069	1.184391
78	estandar_isf_200_0_05.csv	NaN	0.842722	-0.163788	1.383880e-02	0.830461	0.015892	0.842729	0.245546	0.052278	1.194888
90	estandar_isf_400_0_15.csv	NaN	0.837243	-0.164843	1.085325e-02	0.826746	0.014372	0.837467	0.246099	0.053179	1.197816
83	estandar_isf_300_0_05.csv	NaN	0.845741	-0.163078	6.947816e-03	0.836172	0.017429	0.845769	0.246152	0.052502	1.180923
18	estandar_lof_manhattan_5.csv	NaN	0.841343	-0.168230	8.156581e-03	0.833648	0.015131	0.841345	0.246250	0.052724	1.089186
23	estandar_lof_minkowski_9.csv	NaN	0.847796	-0.160265	1.350974e-02	0.838726	0.016756	0.847815	0.247183	0.053020	1.166210
22	estandar_lof_euclidean_9.csv	NaN	0.847796	-0.160265	1.350974e-02	0.838726	0.016756	0.847815	0.247183	0.053020	1.166210
21	estandar_lof_manhattan_7.csv	NaN	0.835372	-0.171107	1.274728e-02	0.828572	0.013958	0.835535	0.248229	0.053926	1.055458

112 rows × 11 columns

En la gráfica podemos observar que los datos estimados varían tanto en el límite superior como inferior, donde en el límite superior, se observa que varían alejándose y acercándose al límite superior de los datos originales. Por el lado del límite inferior se puede observar que tiene una variación más alta y pocas veces los datos estimados pueden pasar el límite de los datos originales.

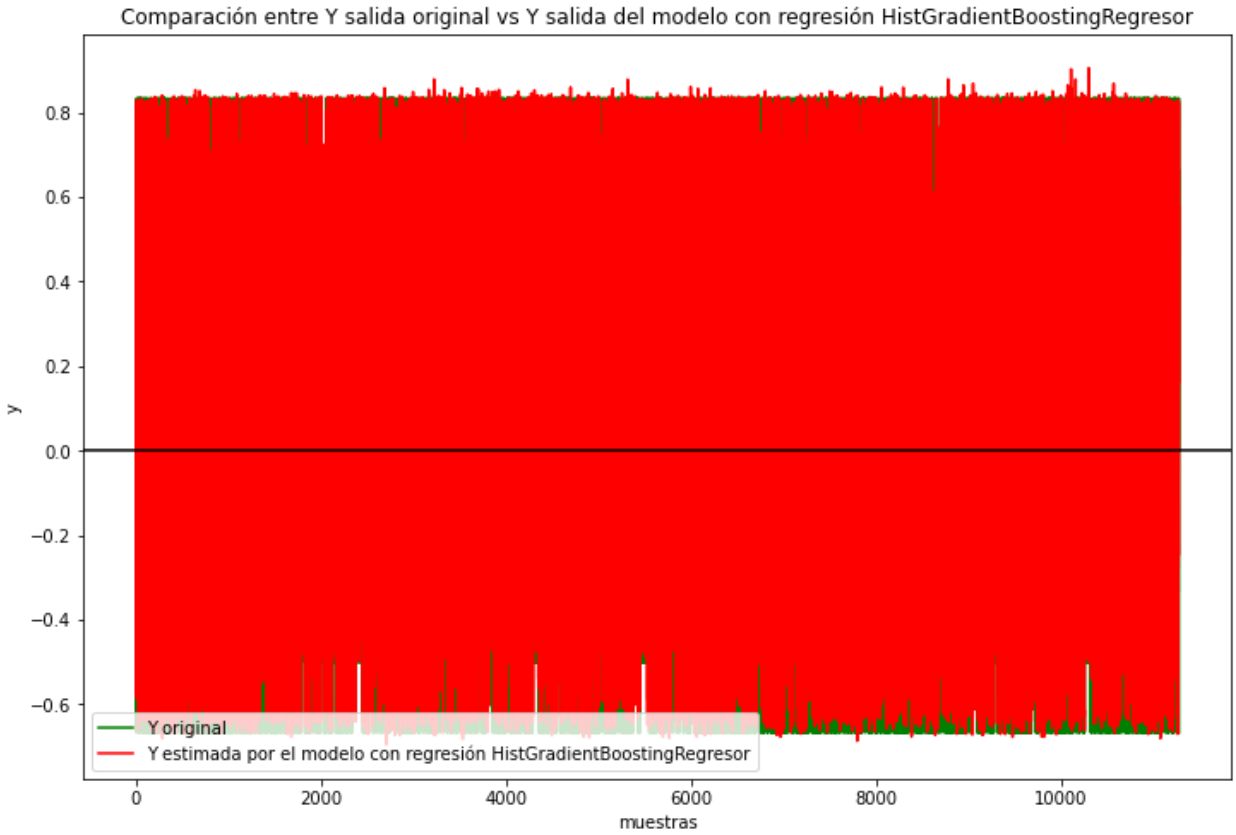

Fig. 32. Comparación de los resultados estimados por el mejor modelo de regresión HGB con características polinómicas y los resultados verdaderos para los medianos agricultores

6.1.6.2 Grandes agricultores

Para este conjunto de datos, tanto los mejores como los peores resultados se obtuvieron con las bases de datos que se les aplico el algoritmo de detección de datos atípicos LOF, con sus parámetros variados, por el lado de los algoritmos de escalamiento, vemos que no hay mucha diferencia a los anteriores resultados, el min-max sigue estando en el top, mientras que el robusto vuelve a aparecer en los peores resultados.

TABLA XXXV
RESULTADOS DEL MODELO REGRESIÓN HGB CON LAS MULTIPLES BASES DE DATOS GENERADAS DE LOS GRANDES AGRICULTORES

	NOMBRE	CONFIG	R2_MODEL	CROSS_ECM	CROSS_ECM_DE	CROSS_R2	CROSS_R2_DE	VARIANZA_EXPL	MAE	MAE_POISSON	MAE_GAMMA
30	min_max_lof_manhattan_5.csv	NaN	0.958215	-0.000101	0.000036	0.903765	0.021283	0.958215	0.001119	0.000644	0.332907
29	min_max_lof_minkowski_5.csv	NaN	0.959979	-0.000092	0.000046	0.915255	0.053368	0.959979	0.001129	0.000748	0.434208
28	min_max_lof_euclidean_5.csv	NaN	0.959979	-0.000092	0.000046	0.915255	0.053368	0.959979	0.001129	0.000748	0.434208
39	min_max_lof_manhattan_11.csv	NaN	0.962276	-0.000082	0.000041	0.902807	0.038033	0.962276	0.001136	0.000986	0.310533
33	min_max_lof_manhattan_7.csv	NaN	0.958511	-0.000083	0.000021	0.915087	0.040538	0.958513	0.001138	0.000685	0.360186
34	min_max_lof_euclidean_9.csv	NaN	0.946729	-0.000091	0.000024	0.912032	0.030238	0.946729	0.001143	0.001069	0.599300
35	min_max_lof_minkowski_9.csv	NaN	0.946729	-0.000091	0.000024	0.912032	0.030238	0.946729	0.001143	0.001069	0.599300
36	min_max_lof_manhattan_9.csv	NaN	0.940462	-0.000078	0.000027	0.914085	0.041290	0.940464	0.001181	0.001641	0.994254
37	min_max_lof_euclidean_11.csv	NaN	0.945265	-0.000113	0.000046	0.867232	0.056444	0.945265	0.001197	0.000820	0.678219
38	min_max_lof_minkowski_11.csv	NaN	0.945265	-0.000113	0.000046	0.867232	0.056444	0.945265	0.001197	0.000820	0.678219
...	...	...	...	...	...	...	...	...	...	...	...
11	robusto_lof_minkowski_9.csv	NaN	0.949455	-3.253185	1.891198	0.904506	0.047750	0.949457	0.221692	0.002958	1.567529
13	robusto_lof_euclidean_11.csv	NaN	0.950985	-2.856631	0.600856	0.915295	0.034952	0.950985	0.221714	0.001167	0.712259
14	robusto_lof_minkowski_11.csv	NaN	0.950985	-2.856631	0.600856	0.915295	0.034952	0.950985	0.221714	0.001167	0.712259
15	robusto_lof_manhattan_11.csv	NaN	0.956735	-2.592355	1.882646	0.921416	0.040717	0.956735	0.222101	0.004562	1.948806
6	robusto_lof_manhattan_5.csv	NaN	0.938943	-3.522309	1.088353	0.892733	0.043017	0.938943	0.223158	0.004026	1.768614
12	robusto_lof_manhattan_9.csv	NaN	0.950778	-2.133523	0.601814	0.940708	0.026340	0.950778	0.227126	0.001070	0.657764
9	robusto_lof_manhattan_7.csv	NaN	0.954027	-2.565534	1.077309	0.902525	0.041845	0.954027	0.231137	0.000786	0.303385
5	robusto_lof_minkowski_5.csv	NaN	0.946202	-3.124169	0.565231	0.921419	0.015950	0.946202	0.238962	0.001152	0.387268
4	robusto_lof_euclidean_5.csv	NaN	0.946202	-3.124169	0.565231	0.921419	0.015950	0.946202	0.238962	0.001152	0.387268
0	robusto_original.csv	NaN	0.939240	-3.172429	1.031592	0.892256	0.031754	0.939245	0.242616	0.001384	0.935455

112 rows × 11 columns

En esta gráfica podemos ver que los picos máximos tanto de los datos estimados como los datos originales coinciden en la muestra, pero en altura no, se ve que los máximos alcanzados por los datos originales no son alcanzados por los picos máximo de los datos estimados, por otro lado, como se observado en modelos anteriores el límite inferior de ambos datos siempre coinciden.

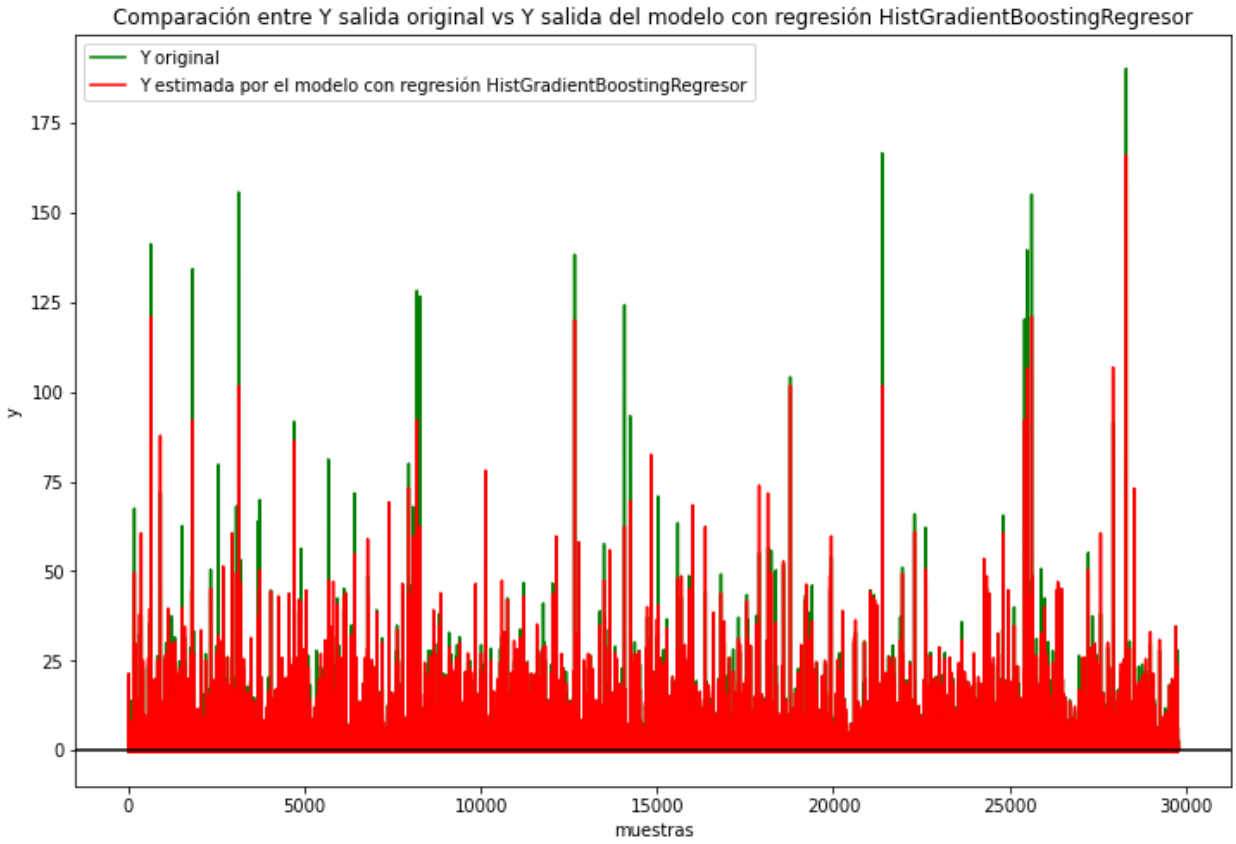

Fig. 33. Comparación de los resultados estimados por el mejor modelo de regresión HGB con características polinómicas y los resultados verdaderos para los grandes agricultores

*4.3.7 Modelo de regresión **Huber***

6.1.7.1 Medianos agricultores

El resultado obtenido de estas iteraciones muestra unos resultados muy similares a los anteriores en cuanto a los mejores conjuntos de datos, donde el algoritmo de eliminación de atípicos en los mejores fue ISF, en los peores fue LOF acá se observa una gran variación de los diferentes parámetros de estos dos métodos, por otros lados los métodos de escalamiento siguen teniendo resultados similares.

TABLA XXXVI
RESULTADOS DEL MODELO REGRESIÓN HUBER CON LAS MULTIPLES BASES DE DATOS GENERADAS DE LOS MEDIANOS AGRICULTORES

	NOMBRE	CONFIG	R2_MODEL	CROSS_ECM	CROSS_ECM_DE	CROSS_R2	CROSS_R2_DE	VARIANZA_EXPL	MAE	MAE_POISSON	MAE_GAMMA
333	min_max_isf_400_0_2.csv	1.05	0.801617	-0.000012	1.732563e-06	0.799464	0.040152	0.829517	0.001292	0.088298	4.336807e+02
335	min_max_isf_400_0_2.csv	1.45	0.801697	-0.000012	1.715777e-06	0.799874	0.040091	0.829536	0.001293	0.141100	5.350367e+05
334	min_max_isf_400_0_2.csv	1.25	0.801431	-0.000012	1.716401e-06	0.799508	0.039912	0.829440	0.001295	0.066582	2.563104e+01
278	min_max_isf_100_auto.csv	1.45	0.825290	-0.000013	1.213365e-06	0.825338	0.033429	0.850171	0.001335	0.111238	1.051941e+05
277	min_max_isf_100_auto.csv	1.25	0.825316	-0.000013	1.215953e-06	0.825175	0.033566	0.850179	0.001336	0.142188	9.365399e+06
276	min_max_isf_100_auto.csv	1.05	0.825333	-0.000013	1.216771e-06	0.825275	0.033428	0.850184	0.001336	0.129226	1.148778e+06
289	min_max_isf_100_0_2.csv	1.25	0.803483	-0.000012	8.048686e-07	0.803162	0.017851	0.831937	0.001344	0.144001	6.227423e+05
288	min_max_isf_100_0_2.csv	1.05	0.803558	-0.000012	8.109204e-07	0.803240	0.017948	0.831936	0.001345	0.169198	1.146747e+07
290	min_max_isf_100_0_2.csv	1.45	0.803310	-0.000012	7.974071e-07	0.803502	0.017901	0.831875	0.001346	0.151806	1.627087e+06
305	min_max_isf_200_0_2.csv	1.45	0.791481	-0.000013	1.786320e-06	0.791405	0.039774	0.821609	0.001358	0.076653	8.368319e+01
...	...	...	...	...	...	...	...	...	...	...	...
67	estandar_lof_euclidean_9.csv	1.25	0.713420	-0.282723	2.540857e-02	0.712920	0.032041	0.766029	0.227949	0.470740	6.350895e+08
70	estandar_lof_minkowski_9.csv	1.25	0.713420	-0.282723	2.540857e-02	0.712920	0.032041	0.766029	0.227949	0.470740	6.350895e+08
68	estandar_lof_euclidean_9.csv	1.45	0.713435	-0.282138	2.496185e-02	0.713945	0.031613	0.766033	0.227956	0.302960	4.304499e+03
71	estandar_lof_minkowski_9.csv	1.45	0.713435	-0.282138	2.496185e-02	0.713945	0.031613	0.766033	0.227956	0.302960	4.304499e+03
54	estandar_lof_manhattan_5.csv	1.05	0.696098	-0.303241	2.041069e-02	0.696204	0.034107	0.749160	0.230102	0.522993	3.196029e+06
55	estandar_lof_manhattan_5.csv	1.25	0.696098	-0.303241	2.041102e-02	0.696204	0.034107	0.749160	0.230102	0.524657	3.390842e+06
56	estandar_lof_manhattan_5.csv	1.45	0.696249	-0.303201	2.044001e-02	0.696468	0.034067	0.749208	0.230165	0.259176	2.154828e+02
63	estandar_lof_manhattan_7.csv	1.05	0.678888	-0.315396	1.592418e-02	0.678626	0.026485	0.736891	0.238825	0.540049	4.341912e+06
64	estandar_lof_manhattan_7.csv	1.25	0.678888	-0.315375	1.591675e-02	0.678626	0.026484	0.736891	0.238825	0.502946	1.113327e+06
65	estandar_lof_manhattan_7.csv	1.45	0.678898	-0.296070	3.094433e-02	0.692962	0.038774	0.736894	0.238829	0.361465	7.903178e+03

336 rows × 11 columns

De esta gráfica podemos observar que el límite inferior de los datos estimados sobrepasó los límites de los datos originales en gran medida, en cuanto al límite superior vemos que ambos límites coinciden, a excepción de un pico que se observa en los datos estimados.

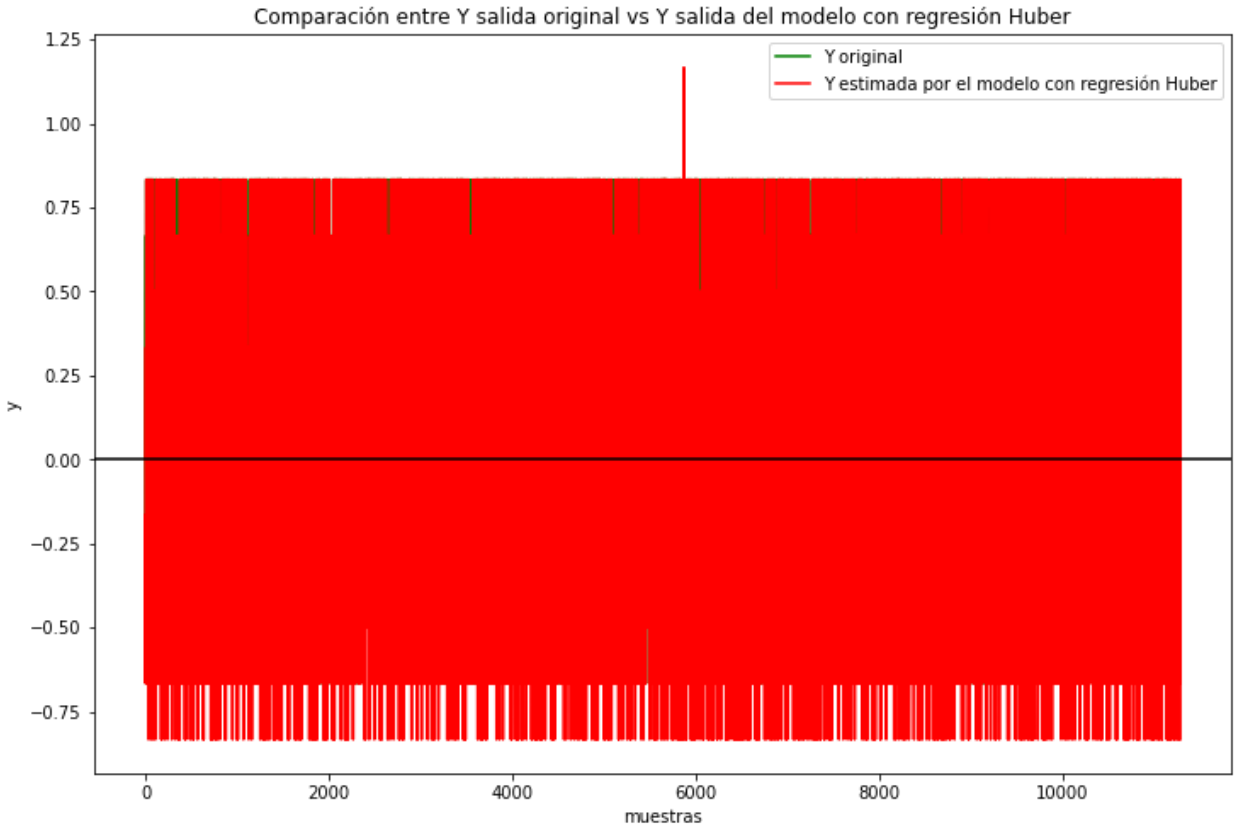

Fig. 34. Comparación de los resultados estimados por el mejor modelo de regresión Huber con características polinómicas y los resultados verdaderos para los medianos agricultores

6.1.7.2 Grandes agricultores

Tanto para los mejores resultados como para los peores resultados el algoritmo de eliminación de atípicos fue LOF donde sus vecinos y distancias tenían diferentes combinaciones, en cuanto a los algoritmos de escalamiento para los mejores resultados es min-max tal como se esperaba y para los peores fue con un escalamiento robusto.

TABLA XXXVII
RESULTADOS DEL MODELO REGRESIÓN HUBER CON LAS MULTIPLES BASES DE DATOS GENERADAS DE LOS GRANDES AGRICULTORES

	NOMBRE	CONFIG	R2_MODEL	CROSS_ECM	CROSS_ECM_DE	CROSS_R2	CROSS_R2_DE	VARIANZA_EXPL	MAE	MAE_POISSON	MAE_GAMMA
90	min_max_lof_manhattan_5.csv	1.05	0.958431	-0.000037	0.000019	0.961087	0.023277	0.958924	0.001044	0.000828	0.769607
91	min_max_lof_manhattan_5.csv	1.25	0.958616	-0.000037	0.000019	0.961086	0.023155	0.959082	0.001045	0.000833	0.744070
111	min_max_lof_euclidean_11.csv	1.05	0.954056	-0.000035	0.000016	0.953008	0.029328	0.954626	0.001045	0.000941	0.793603
114	min_max_lof_minkowski_11.csv	1.05	0.954056	-0.000035	0.000016	0.953008	0.029328	0.954626	0.001045	0.000941	0.793603
112	min_max_lof_euclidean_11.csv	1.25	0.954137	-0.000035	0.000016	0.953130	0.029333	0.954681	0.001046	0.000937	0.716783
115	min_max_lof_minkowski_11.csv	1.25	0.954137	-0.000035	0.000016	0.953130	0.029333	0.954681	0.001046	0.000937	0.716783
92	min_max_lof_manhattan_5.csv	1.45	0.958755	-0.000037	0.000019	0.961320	0.023056	0.959189	0.001048	0.000848	0.729580
113	min_max_lof_euclidean_11.csv	1.45	0.954312	-0.000035	0.000016	0.953387	0.029319	0.954813	0.001050	0.000940	0.664309
116	min_max_lof_minkowski_11.csv	1.45	0.954312	-0.000035	0.000016	0.953387	0.029319	0.954813	0.001050	0.000940	0.664309
108	min_max_lof_manhattan_9.csv	1.05	0.962022	-0.000033	0.000017	0.959088	0.035343	0.962511	0.001075	0.001043	0.717081
...	...	...	...	...	...	...	...	...	...	...	...
2	robusto_original.csv	1.45	0.958098	-1.269111	0.383325	0.957369	0.009824	0.958700	0.212276	0.001238	0.843454
27	robusto_lof_manhattan_7.csv	1.05	0.950416	-1.522252	0.380205	0.949911	0.027829	0.951180	0.219864	0.001182	0.825595
28	robusto_lof_manhattan_7.csv	1.25	0.950511	-1.519894	0.379999	0.949991	0.027813	0.951246	0.220056	0.001181	0.794233
12	robusto_lof_euclidean_5.csv	1.05	0.937546	-2.142371	0.871389	0.939612	0.030127	0.938235	0.220205	0.001139	1.044522
15	robusto_lof_minkowski_5.csv	1.05	0.937546	-2.142371	0.871389	0.939612	0.030127	0.938235	0.220205	0.001139	1.044522
13	robusto_lof_euclidean_5.csv	1.25	0.937643	-2.135646	0.870255	0.939789	0.030050	0.938306	0.220397	0.001129	0.851258
16	robusto_lof_minkowski_5.csv	1.25	0.937643	-2.135646	0.870255	0.939789	0.030050	0.938306	0.220397	0.001129	0.851258
29	robusto_lof_manhattan_7.csv	1.45	0.950698	-1.514676	0.379017	0.950157	0.027783	0.951387	0.220661	0.001187	0.774097
14	robusto_lof_euclidean_5.csv	1.45	0.937844	-2.124134	0.868833	0.940090	0.029982	0.938464	0.221032	0.001123	0.780108
17	robusto_lof_minkowski_5.csv	1.45	0.937844	-2.124134	0.868833	0.940090	0.029982	0.938464	0.221032	0.001123	0.780108

336 rows × 11 columns

En esta gráfica podemos observar cómo los datos estimados tienen una distribución muy similar a la original, pero los picos máximos no han sido alcanzados por los datos estimados vemos que en sus muestras coinciden muy, pero en la altura se queda un poco.

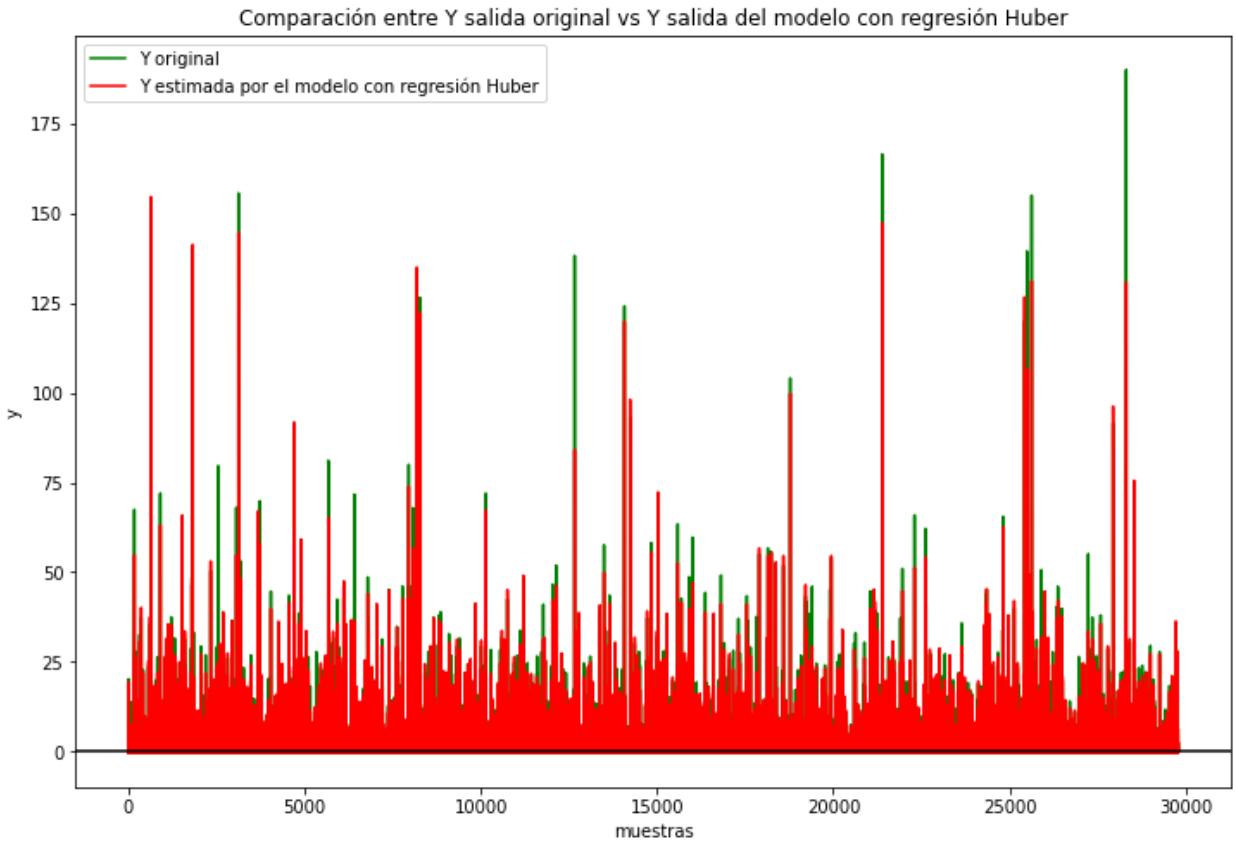

Fig. 35. Comparación de los resultados estimados por el mejor modelo de regresión Huber con características polinómicas y los resultados verdaderos para los grandes agricultores

4.3.8 Modelo de regresión Theil Sen

6.1.8.1 Medianos agricultores

Con este último modelo, podemos observar los resultados obtenidos anteriormente con los diferentes modelos, se asemejan tanto a los algoritmos de escalamiento y de eliminación de datos atípicos, con algunas diferencias en los parámetros, pero no tan relevantes, además que en todos modelos los mejores resultados se obtuvieron con el escalamiento min-max

TABLA XXXVIII
RESULTADOS DEL MODELO REGRESIÓN THEIL SEN CON LAS MULTIPLES BASES DE DATOS GENERADAS DE LOS MEDIANOS AGRICULTORES

	NOMBRE	CONFIG	R2_MODEL	CROSS_ECM	CROSS_ECM_DE	CROSS_R2	CROSS_R2_DE	VARIANZA_EXPL	MAE	MAE_POISSON	MAE_GAMMA
111	min_max_isf_400_0_2.csv	NaN	0.823982	-0.000011	1.635960e-06	0.820809	0.037402	0.832719	0.001520	0.043816	2.167053
96	min_max_isf_100_0_2.csv	NaN	0.825208	-0.000011	7.007064e-07	0.826256	0.017077	0.835626	0.001545	0.031885	1.478217
106	min_max_isf_300_0_2.csv	NaN	0.821995	-0.000011	7.615111e-07	0.823251	0.032859	0.833127	0.001554	0.039004	1.378199
92	min_max_isf_100_auto.csv	NaN	0.845187	-0.000011	1.188473e-06	0.845781	0.031947	0.853707	0.001558	0.029854	1.998413
101	min_max_isf_200_0_2.csv	NaN	0.815051	-0.000011	1.615539e-06	0.815717	0.036398	0.825823	0.001568	0.038930	1.404984
107	min_max_isf_400_auto.csv	NaN	0.830195	-0.000012	1.789578e-06	0.832232	0.030424	0.840657	0.001593	0.030089	2.126502
102	min_max_isf_300_auto.csv	NaN	0.819600	-0.000013	2.698134e-06	0.819673	0.037044	0.831022	0.001618	0.033625	1.086078
97	min_max_isf_200_auto.csv	NaN	0.834644	-0.000012	1.688376e-06	0.833359	0.035645	0.843613	0.001619	0.032607	1.524619
95	min_max_isf_100_0_15.csv	NaN	0.825696	-0.000016	2.378954e-06	0.826248	0.031272	0.836427	0.001675	0.032251	0.939783
100	min_max_isf_200_0_15.csv	NaN	0.815985	-0.000017	3.676486e-06	0.815570	0.043038	0.826916	0.001695	0.033514	2.012749
...	...	...	...	...	...	...	...	...	...	...	...
26	estandar_lof_minkowski_11.csv	NaN	0.774743	-0.229462	1.406855e-02	0.770394	0.022275	0.786135	0.253120	0.096711	1.696716
27	estandar_lof_manhattan_11.csv	NaN	0.774352	-0.223094	9.911193e-03	0.773683	0.018559	0.786005	0.254435	0.096666	1.679712
24	estandar_lof_manhattan_9.csv	NaN	0.766971	-0.228818	1.721187e-02	0.769312	0.021576	0.781682	0.254664	0.114951	1.865716
16	estandar_lof_euclidean_5.csv	NaN	0.772113	-0.226098	1.258402e-02	0.771106	0.020939	0.784489	0.255441	0.093634	1.567693
17	estandar_lof_minkowski_5.csv	NaN	0.772113	-0.226098	1.258402e-02	0.771106	0.020939	0.784489	0.255441	0.093634	1.567693
83	estandar_isf_300_0_05.csv	NaN	0.769929	-0.226077	2.405012e-02	0.770021	0.037801	0.783033	0.255475	0.100143	2.053155
22	estandar_lof_euclidean_9.csv	NaN	0.766583	-0.227869	2.237314e-02	0.767370	0.028681	0.780416	0.259804	0.098432	1.728240
23	estandar_lof_minkowski_9.csv	NaN	0.766583	-0.227869	2.237314e-02	0.767370	0.028681	0.780416	0.259804	0.098432	1.728240
18	estandar_lof_manhattan_5.csv	NaN	0.756043	-0.245532	1.785216e-02	0.754283	0.028589	0.768180	0.261769	0.098326	1.523484
21	estandar_lof_manhattan_7.csv	NaN	0.739268	-0.252277	1.345296e-02	0.743722	0.021704	0.755117	0.268468	0.110058	3.922420

112 rows × 11 columns

En esta gráfica podemos observar que este es de los pocos modelos que sobrepasan los límites tanto el inferior como el superior de los datos originales, además que estos datos estimados tienen cierta variación a comparación de los datos originales que no posee una variabilidad razonable. Por último, podemos observar un pequeño pico que se generó en los datos estimados.

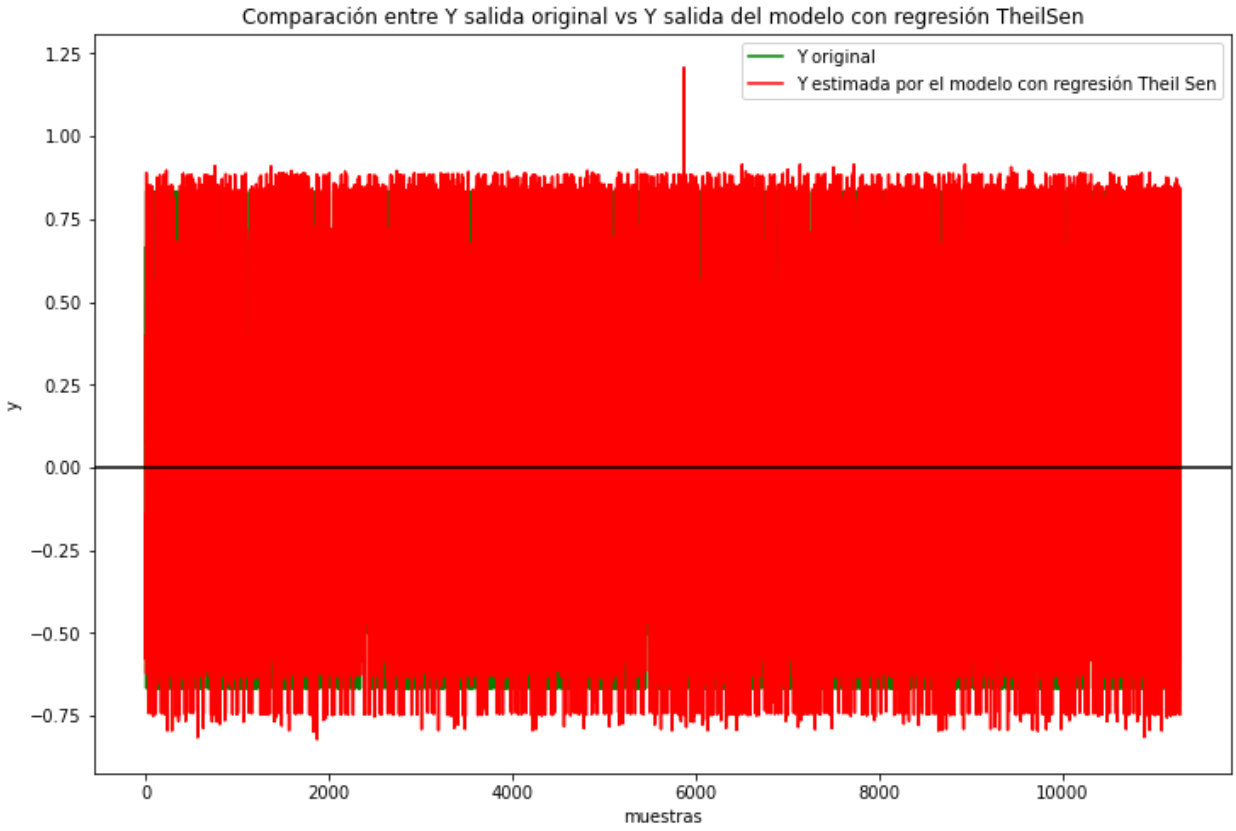

Fig. 36. Comparación de los resultados estimados por el mejor modelo de regresión Theil Sen con características polinómicas y los resultados verdaderos para los medianos agricultores

6.1.8.2 Grandes agricultores

Los resultados obtenidos con el conjunto de datos creadas a partir de los datos de grandes agricultores, podemos observar que el algoritmo de eliminación de atípicos coincide tanto en los mejores resultados como en los peores, por otro lado, vemos que en los peores resultados se encuentra una base de datos escalonada con escalamiento robusto, dicho resultado sólo se observó 2 veces.

TABLA XXXIX
RESULTADOS DEL MODELO REGRESIÓN THEIL SEN CON LAS MULTIPLES BASES DE DATOS GENERADAS DE LOS GRANDES AGRICULTORES

	NOMBRE	CONFIG	R2_MODEL	CROSS_ECM	CROSS_ECM_DE	CROSS_R2	CROSS_R2_DE	VARIANZA_EXPL	MAE	MAE_POISSON	MAE_GAMMA
38	min_max_lof_minkowski_11.csv	NaN	0.945931	-0.000039	0.000018	0.948279	0.029409	0.946495	0.001132	0.001053	0.679068
37	min_max_lof_euclidean_11.csv	NaN	0.945931	-0.000039	0.000018	0.948279	0.029409	0.946495	0.001132	0.001053	0.679068
30	min_max_lof_manhattan_5.csv	NaN	0.952024	-0.000040	0.000017	0.956476	0.021880	0.952410	0.001156	0.000871	0.744459
36	min_max_lof_manhattan_9.csv	NaN	0.956496	-0.000038	0.000016	0.952614	0.032705	0.956926	0.001162	0.001088	0.706411
32	min_max_lof_minkowski_7.csv	NaN	0.940163	-0.000049	0.000013	0.947111	0.030074	0.940634	0.001164	0.001186	0.763762
31	min_max_lof_euclidean_7.csv	NaN	0.940163	-0.000049	0.000013	0.947111	0.030074	0.940634	0.001164	0.001186	0.763762
34	min_max_lof_euclidean_9.csv	NaN	0.943835	-0.000049	0.000010	0.937158	0.035390	0.944325	0.001177	0.001229	0.730332
35	min_max_lof_minkowski_9.csv	NaN	0.943835	-0.000049	0.000010	0.937158	0.035390	0.944325	0.001177	0.001229	0.730332
29	min_max_lof_minkowski_5.csv	NaN	0.938358	-0.000050	0.000017	0.944173	0.037636	0.938899	0.001188	0.001229	0.765421
28	min_max_lof_euclidean_5.csv	NaN	0.938358	-0.000050	0.000017	0.944173	0.037636	0.938899	0.001188	0.001229	0.765421
...	...	...	...	...	...	...	...	...	...	...	...
15	robusto_lof_manhattan_11.csv	NaN	0.950939	-1.379581	0.661083	0.953849	0.015299	0.951618	0.219112	0.001334	0.800665
12	robusto_lof_manhattan_9.csv	NaN	0.945287	-1.602095	0.369925	0.942662	0.023582	0.946001	0.219564	0.001419	0.835906
14	robusto_lof_minkowski_11.csv	NaN	0.938451	-1.629798	0.365442	0.944597	0.020748	0.939168	0.220227	0.001176	0.747117
13	robusto_lof_euclidean_11.csv	NaN	0.938451	-1.629798	0.365442	0.944597	0.020748	0.939168	0.220227	0.001176	0.747117
10	robusto_lof_euclidean_9.csv	NaN	0.938932	-1.693419	0.559408	0.938092	0.019168	0.939653	0.221501	0.001521	0.867953
11	robusto_lof_minkowski_9.csv	NaN	0.938932	-1.693419	0.559408	0.938092	0.019168	0.939653	0.221501	0.001521	0.867953
0	robusto_original.csv	NaN	0.946570	-1.567995	0.620567	0.948986	0.013285	0.947302	0.225810	0.001323	0.836373
5	robusto_lof_minkowski_5.csv	NaN	0.933078	-2.383334	0.931118	0.934846	0.030063	0.933739	0.233579	0.001119	0.762385
4	robusto_lof_euclidean_5.csv	NaN	0.933078	-2.383334	0.931118	0.934846	0.030063	0.933739	0.233579	0.001119	0.762385
9	robusto_lof_manhattan_7.csv	NaN	0.941733	-1.694919	0.245483	0.942371	0.027577	0.942522	0.235534	0.001187	0.782912

112 rows × 11 columns

De esta gráfica podemos concluir que las distribuciones obtenidas tanto en los datos originales como en los datos estimados en los diferentes modelos tienen a comportarse similar con la base de datos de los grandes agricultores, en esta gráfica podemos remarcar el hecho hubo al menos un pico de los datos estimados que pasó a un pico de los datos originales.

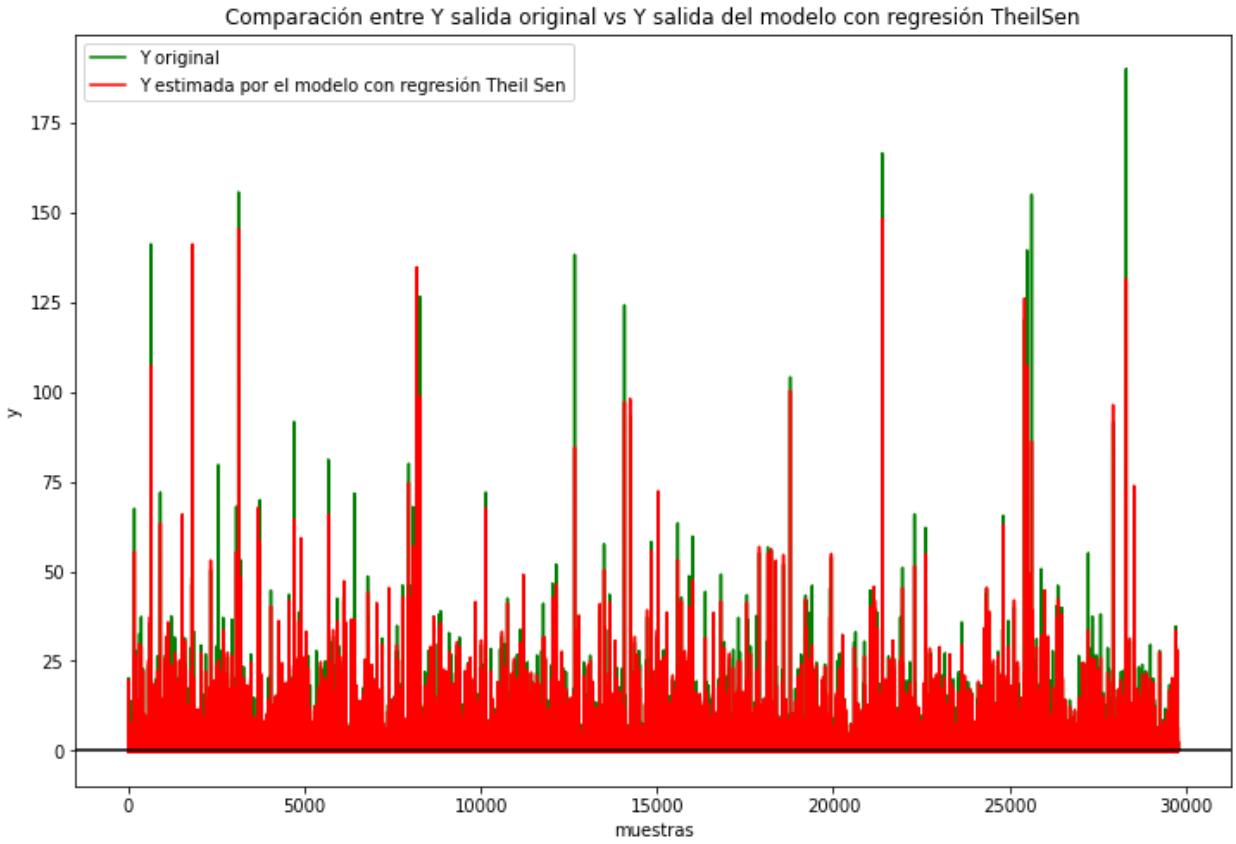

Fig. 37. Comparación de los resultados estimados por el mejor modelo de regresión Theil Sen con características polinómicas y los resultados verdaderos para los grandes agricultores

6.2 EVALUACIÓN CUALITATIVA

Tal cual como se comentó al principio, nuestra métrica para determinar el modelo que se usará fue el **MAE**, a continuación, se presentará el resultado teniendo en cuenta los 3 mejores resultados de los 8 modelos entrenados, tanto de los pequeños productores como los grandes agricultores y a partir de estos seleccionar el mejor modelo. En este apartado nos centraremos en observar cuáles fueron los modelos que mejor se comportaron en cada base de datos.

6.2.1 Selección de los mejores modelos teniendo en cuenta los 3 mejores resultados.

A continuación, se mostrará qué posición ocupa cada modelo dependiendo de la métrica MAE obtenida.

6.2.1.1 Medianos agricultores

1. Regresión de Huber
2. Regresión de Theil Sen
3. Regresión de bosque aleatorio (Random forest)
4. Regresión HGB
5. Regresión con características polinómicas
6. Regresión robusta
7. Regresión simple
8. Regresión MLP

TABLA XL
MEJORES MODELOS TENIENDO EN CUENTA LOS 3 MEJORES RESULTADOS PARA LOS MEDIANOS AGRICULTORES.

	MODEL	NOMBRE	CONFIG	R2_MODEL	CROSS_ECM	CROSS_ECM_DE	CROSS_R2	CROSS_R2_DE	VARIANZA_EXPL	MAE	MAE_POISSON	MAE_GAMMA
278	huber_regressor.csv	min_max_isf_100_auto.csv	1.45	0.825290	-1.252414e-05	1.213365e-06	8.253379e-01	3.342905e-02	0.850171	0.001335	0.111238	1.051941e+05
277	huber_regressor.csv	min_max_isf_100_auto.csv	1.25	0.825316	-1.251933e-05	1.215953e-06	8.251753e-01	3.356635e-02	0.850179	0.001336	0.142188	9.365399e+06
276	huber_regressor.csv	min_max_isf_100_auto.csv	1.05	0.825333	-1.250476e-05	1.216771e-06	8.252749e-01	3.342798e-02	0.850184	0.001336	0.129226	1.148778e+06
96	theil_sen_regressor.csv	min_max_isf_100_0_2.csv	NaN	0.825208	-1.099621e-05	7.007064e-07	8.262562e-01	1.707703e-02	0.835626	0.001545	0.031885	1.478217e+00
106	theil_sen_regressor.csv	min_max_isf_300_0_2.csv	NaN	0.821995	-1.116523e-05	7.615111e-07	8.232510e-01	3.285890e-02	0.833127	0.001554	0.039004	1.378199e+00
92	theil_sen_regressor.csv	min_max_isf_100_auto.csv	NaN	0.845187	-1.108243e-05	1.188473e-06	8.457813e-01	3.194696e-02	0.853707	0.001558	0.029854	1.998413e+00
481	random_forest.csv	min_max_isf_100_0_2.csv	40.0	0.880696	-8.653518e-06	5.973486e-07	8.661208e-01	1.801453e-02	0.880711	0.001560	0.014953	4.641269e-01
482	random_forest.csv	min_max_isf_100_0_2.csv	60.0	0.881093	-8.648913e-06	5.573987e-07	8.661238e-01	1.769829e-02	0.881105	0.001560	0.014947	5.748687e-01
480	random_forest.csv	min_max_isf_100_0_2.csv	20.0	0.879832	-8.773238e-06	6.050863e-07	8.651562e-01	1.756639e-02	0.879859	0.001560	0.015148	8.186749e-01
96	hgbr_regressor.csv	min_max_isf_100_0_2.csv	NaN	0.878271	-8.182953e-06	4.795193e-07	8.722294e-01	1.592790e-02	0.878276	0.001584	0.019978	2.885411e-01
106	hgbr_regressor.csv	min_max_isf_300_0_2.csv	NaN	0.879503	-7.987427e-06	5.771161e-07	8.754916e-01	2.245752e-02	0.879507	0.001617	0.026142	2.935348e-01
101	hgbr_regressor.csv	min_max_isf_200_0_2.csv	NaN	0.872375	-8.218902e-06	8.916480e-07	8.652233e-01	1.761288e-02	0.872408	0.001625	0.025078	6.073181e-01
480	caracteristicas_polinomicas.csv	min_max_isf_100_0_2.csv	2.0	0.841086	-1.022073e-05	5.583332e-07	8.388218e-01	1.580671e-02	0.841093	0.001842	0.026234	1.365947e+00
507	caracteristicas_polinomicas.csv	min_max_isf_200_0_2.csv	4.0	0.835012	-4.665617e+07	6.090759e+07	-1.245054e+13	3.735161e+13	0.835046	0.001849	0.034060	6.030431e-01
481	caracteristicas_polinomicas.csv	min_max_isf_100_0_2.csv	3.0	0.841655	-6.095483e+01	1.140361e+02	-2.539594e+09	7.594080e+09	0.841655	0.001850	0.024960	1.283484e+00
480	regresion_robusta.csv	min_max_isf_100_0_2.csv	0.05	0.837924	-1.030233e-05	5.016961e-07	8.374546e-01	1.520662e-02	0.837928	0.001860	0.029248	3.663281e+00
481	regresion_robusta.csv	min_max_isf_100_0_2.csv	0.1	0.837910	-1.030286e-05	5.012124e-07	8.374475e-01	1.519508e-02	0.837914	0.001861	0.029049	4.991114e+00
96	regresion_simple.csv	min_max_isf_100_0_2.csv	NaN	0.837917	-1.030286e-05	5.012124e-07	8.374475e-01	1.519508e-02	0.837920	0.001861	0.029109	4.900262e+00
482	regresion_robusta.csv	min_max_isf_100_0_2.csv	0.15	0.837917	-1.030286e-05	5.012124e-07	8.374475e-01	1.519508e-02	0.837920	0.001861	0.029109	4.900262e+00
101	regresion_simple.csv	min_max_isf_200_0_2.csv	NaN	0.830008	-1.038802e-05	1.360899e-06	8.295668e-01	3.174550e-02	0.830023	0.001882	0.035346	9.669000e-01
106	regresion_simple.csv	min_max_isf_300_0_2.csv	NaN	0.835896	-1.046877e-05	6.190353e-07	8.356184e-01	2.940328e-02	0.835902	0.001892	0.036218	1.114131e+00
215	mlp_regressor.csv	min_max_isf_400_auto.csv	100_2	0.836226	-8.346774e-05	3.451386e-05	2.424888e-01	5.024929e-01	0.836323	0.001961	0.033603	2.054826e+00
213	mlp_regressor.csv	min_max_isf_300_0_2.csv	100_2	0.703615	-3.562299e-05	3.159294e-06	4.476827e-01	2.133857e-01	0.732757	0.001989	0.080320	1.839866e+00
209	mlp_regressor.csv	min_max_isf_300_0_1.csv	100_2	0.841279	-4.485374e-05	1.472426e-05	6.463663e-01	8.631289e-02	0.854407	0.002066	0.035344	6.122626e-01

6.2.1.2 Grandes agricultores

1. Regresión de bosque aleatorio (Random forest)
2. Regresión de Huber
3. Regresión HGB
4. Regresión de Theil Sen
5. Regresión con características polinómicas
6. Regresión robusta
7. Regresión simple
8. Regresión MLP

TABLA XLI
MEJORES MODELOS TENIENDO EN CUENTA LOS 3 MEJORES RESULTADOS PARA LOS GRANDES AGRICULTORES

	MODEL	NOMBRE	CONFIG	R2_MODEL	CROSS_ECM	CROSS_ECM_DE	CROSS_R2	CROSS_R2_DE	VARIANZA_EXPL	MAE	MAE_POISSON	MAE_GAMMA
154	random_forest.csv	min_max_lof_manhattan_5.csv	100	0.973404	-0.000041	0.000023	0.959706	0.028319	0.973404	0.001003	0.000754	14.667231
152	random_forest.csv	min_max_lof_manhattan_5.csv	60	0.974021	-0.000041	0.000023	0.960722	0.027018	0.974021	0.001004	0.000606	0.937795
153	random_forest.csv	min_max_lof_manhattan_5.csv	80	0.973406	-0.000041	0.000023	0.959814	0.028093	0.973406	0.001004	0.000750	11.550295
90	huber_regressor.csv	min_max_lof_manhattan_5.csv	1.05	0.958431	-0.000037	0.000019	0.961087	0.023277	0.958924	0.001044	0.000828	0.769607
91	huber_regressor.csv	min_max_lof_manhattan_5.csv	1.25	0.958616	-0.000037	0.000019	0.961086	0.023155	0.959082	0.001045	0.000833	0.744070
111	huber_regressor.csv	min_max_lof_euclidean_11.csv	1.05	0.954056	-0.000035	0.000016	0.953008	0.029328	0.954626	0.001045	0.000941	0.793603
30	hgbr_regressor.csv	min_max_lof_manhattan_5.csv	NaN	0.958215	-0.000101	0.000036	0.903765	0.021283	0.958215	0.001119	0.000644	0.332907
28	hgbr_regressor.csv	min_max_lof_euclidean_5.csv	NaN	0.959979	-0.000092	0.000046	0.915255	0.053368	0.959979	0.001129	0.000748	0.434208
29	hgbr_regressor.csv	min_max_lof_minkowski_5.csv	NaN	0.959979	-0.000092	0.000046	0.915255	0.053368	0.959979	0.001129	0.000748	0.434208
38	theil_sen_regressor.csv	min_max_lof_minkowski_11.csv	NaN	0.945931	-0.000039	0.000018	0.948279	0.029409	0.946495	0.001132	0.001053	0.679068
37	theil_sen_regressor.csv	min_max_lof_euclidean_11.csv	NaN	0.945931	-0.000039	0.000018	0.948279	0.029409	0.946495	0.001132	0.001053	0.679068
145	caracteristicas_polinomicas.csv	min_max_lof_minkowski_5.csv	2.0	0.960471	-0.000035	0.000011	0.961759	0.027974	0.960471	0.001155	0.001444	0.982993
140	caracteristicas_polinomicas.csv	min_max_lof_euclidean_5.csv	2.0	0.960471	-0.000035	0.000011	0.961759	0.027974	0.960471	0.001155	0.001444	0.982993
30	theil_sen_regressor.csv	min_max_lof_manhattan_5.csv	NaN	0.952024	-0.000040	0.000017	0.956476	0.021880	0.952410	0.001156	0.000871	0.744459
150	caracteristicas_polinomicas.csv	min_max_lof_manhattan_5.csv	2.0	0.970110	-0.000027	0.000014	0.971087	0.017742	0.970112	0.001156	0.001122	0.915578
150	regresion_robusta.csv	min_max_lof_manhattan_5.csv	0.05	0.959440	-0.000054	0.000020	0.945680	0.034046	0.959520	0.001238	0.001288	0.960675
190	regresion_robusta.csv	min_max_lof_minkowski_11.csv	0.05	0.955222	-0.000034	0.000016	0.921915	0.078536	0.955282	0.001254	0.001423	0.894379
155	regresion_robusta.csv	min_max_lof_euclidean_7.csv	0.05	0.946050	-0.000049	0.000010	0.940867	0.039014	0.946121	0.001292	0.001772	1.026473
37	regresion_simple.csv	min_max_lof_euclidean_11.csv	NaN	0.958669	-0.000031	0.000015	0.958005	0.027739	0.958669	0.001388	0.001891	1.115753
38	regresion_simple.csv	min_max_lof_minkowski_11.csv	NaN	0.958669	-0.000031	0.000015	0.958005	0.027739	0.958669	0.001388	0.001891	1.115753
28	regresion_simple.csv	min_max_lof_euclidean_5.csv	NaN	0.954161	-0.000040	0.000013	0.957298	0.032225	0.954162	0.001388	0.002522	1.304483
5	mlp_regressor.csv	min_max_original.csv	100_2	0.966476	-0.000034	0.000010	0.958612	0.011125	0.966929	0.001665	0.008123	2.429899
57	mlp_regressor.csv	min_max_lof_euclidean_5.csv	100_2	0.957178	-0.000048	0.000015	0.952129	0.030739	0.957374	0.001735	0.008616	2.434258
59	mlp_regressor.csv	min_max_lof_minkowski_5.csv	100_2	0.957178	-0.000048	0.000015	0.952129	0.030739	0.957374	0.001735	0.008616	2.434258

Podemos observar que el orden de los mejores modelos por cada base de datos varía un poco, debido a que vemos que los últimos 4 modelos coinciden en ambas bases de datos.

6.2.2 Selección de los mejores modelos teniendo en cuenta todos los resultados.

En este apartado se mostrará los 10 mejores resultados y los 10 peores resultados, teniendo en cuenta las 2553 iteraciones totales de todos los modelos y configuraciones probadas en ambas bases de datos, esto para identificar a cuáles modelos les fue bien y a cuáles no.

6.2.2.1 Medianos agricultores

De todos los resultados podemos observar que hay un modelo que se queda con estos 10 lugares que es la regresión de Huber, vemos que son diferentes configuraciones del modelo, con bases de datos diferentes. En cuanto a los 10 peores vemos que la regresión con características polinómicas se lleva estos últimos 10 puestos, a diferencia de los mejores modelos vemos, que estos resultados se dan por dos configuraciones en específico, exponentes entre 5 y 6.

TABLA XLII
MEJORES 10 RESULTADOS Y PEORES 10 RESULTADOS PARA LOS MEDIANOS AGRICULTORES

	MODEL	NOMBRE	CONFIG	R2_MODEL	CROSS_ECM	CROSS_ECM_DE	CROSS_R2	CROSS_R2_DE	VARIANZA_EXPL	MAE	MAE_POISSON	MAE_GAMMA
278	huber_regressor.csv	min_max_isf_100_auto.csv	1.45	8.252905e-01	-1.252414e-05	1.213365e-06	8.253379e-01	3.342905e-02	8.501713e-01	1.335380e-03	0.111238	1.051941e+05
277	huber_regressor.csv	min_max_isf_100_auto.csv	1.25	8.253163e-01	-1.251933e-05	1.215953e-06	8.251753e-01	3.356635e-02	8.501791e-01	1.335617e-03	0.142188	9.365399e+06
276	huber_regressor.csv	min_max_isf_100_auto.csv	1.05	8.253328e-01	-1.250476e-05	1.216771e-06	8.252749e-01	3.342798e-02	8.501838e-01	1.335773e-03	0.129226	1.148778e+06
289	huber_regressor.csv	min_max_isf_100_0_2.csv	1.25	8.034826e-01	-1.246628e-05	8.048686e-07	8.031620e-01	1.785094e-02	8.319371e-01	1.343664e-03	0.144001	6.227423e+05
288	huber_regressor.csv	min_max_isf_100_0_2.csv	1.05	8.035579e-01	-1.244892e-05	8.109204e-07	8.032397e-01	1.794762e-02	8.319361e-01	1.344718e-03	0.169198	1.146747e+07
290	huber_regressor.csv	min_max_isf_100_0_2.csv	1.45	8.033097e-01	-1.245675e-05	7.974071e-07	8.035020e-01	1.790126e-02	8.318749e-01	1.346273e-03	0.151806	1.627087e+06
305	huber_regressor.csv	min_max_isf_200_0_2.csv	1.45	7.914815e-01	-1.272648e-05	1.786320e-06	7.914055e-01	3.977408e-02	8.216090e-01	1.357991e-03	0.076653	8.368319e+01
320	huber_regressor.csv	min_max_isf_300_0_2.csv	1.45	8.008408e-01	-1.270035e-05	9.283950e-07	8.007826e-01	3.477644e-02	8.296793e-01	1.358678e-03	0.072861	6.901356e+01
319	huber_regressor.csv	min_max_isf_300_0_2.csv	1.25	8.011887e-01	-1.268524e-05	9.138437e-07	8.007233e-01	3.436744e-02	8.298677e-01	1.359104e-03	0.165879	4.960292e+06
318	huber_regressor.csv	min_max_isf_300_0_2.csv	1.05	8.012276e-01	-1.268258e-05	9.142663e-07	8.007492e-01	3.443367e-02	8.298812e-01	1.359317e-03	0.159632	1.552762e+07
...	...	...	...	...	...	...	...	...	...	...	...	...
259	caracteristicas_polinomicas.csv	max_normalizacion_lof_manhattan_11.csv	6.0	-1.611428e+10	-3.284518e+11	6.356522e+11	-1.677816e+14	5.032612e+14	-1.611109e+10	4.014482e+01	1.560806	5.879105e+00
234	caracteristicas_polinomicas.csv	max_normalizacion_lof_euclidean_9.csv	6.0	-7.601048e+10	-2.048727e+12	4.093527e+12	-4.534369e+11	1.360311e+12	-7.600230e+10	5.916347e+01	0.842939	9.478179e+05
239	caracteristicas_polinomicas.csv	max_normalizacion_lof_minkowski_9.csv	6.0	-7.601048e+10	-2.048727e+12	4.093527e+12	-4.534369e+11	1.360311e+12	-7.600230e+10	5.916347e+01	0.842939	9.478179e+05
233	caracteristicas_polinomicas.csv	max_normalizacion_lof_euclidean_9.csv	5.0	-2.276666e+11	-1.572569e+11	2.928972e+11	-4.161097e+12	1.248329e+13	-2.276421e+11	1.024384e+02	0.880410	1.649504e+06
238	caracteristicas_polinomicas.csv	max_normalizacion_lof_minkowski_9.csv	5.0	-2.276666e+11	-1.572569e+11	2.928972e+11	-4.161097e+12	1.248329e+13	-2.276421e+11	1.024384e+02	0.880410	1.649504e+06
339	caracteristicas_polinomicas.csv	robusto_isf_400_auto.csv	6.0	-3.192055e+13	-1.601572e+17	1.998436e+17	-1.755001e+17	2.303711e+17	-3.191507e+13	4.536455e+04	0.513277	3.960593e+00
338	caracteristicas_polinomicas.csv	robusto_isf_400_auto.csv	5.0	-9.502013e+16	-1.738312e+17	2.310563e+17	-4.065931e+16	9.755151e+16	-9.498930e+16	2.420142e+06	0.661163	4.121370e+00
414	caracteristicas_polinomicas.csv	estandar_isf_300_auto.csv	6.0	-2.387849e+18	-1.181231e+21	1.255478e+21	-2.818475e+21	6.217579e+21	-2.387400e+18	1.890194e+07	0.665709	2.385551e+00
438	caracteristicas_polinomicas.csv	estandar_isf_400_auto.csv	5.0	-1.043621e+21	-1.083179e+21	1.382417e+21	-1.401458e+21	3.467575e+21	-1.043309e+21	4.861368e+08	0.675724	2.676547e+00
439	caracteristicas_polinomicas.csv	estandar_isf_400_auto.csv	6.0	-3.342044e+21	-6.267935e+21	7.986959e+21	-7.126881e+21	1.187990e+22	-3.341438e+21	9.401873e+08	0.543090	3.994956e+00

2553 rows × 12 columns

6.2.2.2 Grandes agricultores

De todos los resultados podemos observar que hay un modelo que se queda con estos 10 lugares que es la regresión de random forest, vemos que son diferentes configuraciones del modelo donde el mejor resultado se obtuvo con la cantidad máxima de árboles 100 y de ahí va descendiendo, además se puede ver que se utiliza una misma base de datos, solo que en algunos casos variaba la distancia del algoritmo con el que se eliminó los datos atípicos. En cuanto a los 10

peores vemos que la regresión con características polinómicas se lleva estos últimos puestos, a diferencia de los mejores modelos vemos, que estos resultados se dan por diferentes configuraciones vemos que el exponente más pequeño es 3 y el exponente más grande es 6.

TABLA XLIII
MEJORES 10 RESULTADOS Y PEORES 10 RESULTADOS PARA LOS GRANDES AGRICULTORES

	MODEL	NOMBRE	CONFIG	R2_MODEL	CROSS_ECM	CROSS_ECM_DE	CROSS_R2	CROSS_R2_DE	VARIANZA_EXPL	MAE	MAE_POISSON	MAE_GAMMA
154	random_forest.csv	min_max_lof_manhattan_5.csv	100	9.734039e-01	-4.064800e-05	2.253480e-05	9.597061e-01	2.831870e-02	9.734043e-01	0.001003	0.000754	14.667231
152	random_forest.csv	min_max_lof_manhattan_5.csv	60	9.740208e-01	-4.097875e-05	2.255946e-05	9.607219e-01	2.701817e-02	9.740212e-01	0.001004	0.000606	0.937795
153	random_forest.csv	min_max_lof_manhattan_5.csv	80	9.734056e-01	-4.090709e-05	2.264772e-05	9.598144e-01	2.809317e-02	9.734060e-01	0.001004	0.000750	11.550295
151	random_forest.csv	min_max_lof_manhattan_5.csv	40	9.730531e-01	-4.113174e-05	2.304303e-05	9.604094e-01	2.794993e-02	9.730537e-01	0.001011	0.000614	1.128173
150	random_forest.csv	min_max_lof_manhattan_5.csv	20	9.733128e-01	-4.198299e-05	2.298606e-05	9.595816e-01	2.910582e-02	9.733134e-01	0.001012	0.000605	1.065993
147	random_forest.csv	min_max_lof_minkowski_5.csv	60	9.643937e-01	-4.726830e-05	1.767840e-05	9.511495e-01	3.355021e-02	9.643955e-01	0.001022	0.001050	3.456511
142	random_forest.csv	min_max_lof_euclidean_5.csv	60	9.643937e-01	-4.726830e-05	1.767840e-05	9.511495e-01	3.355021e-02	9.643955e-01	0.001022	0.001050	3.456511
146	random_forest.csv	min_max_lof_minkowski_5.csv	40	9.651671e-01	-4.875294e-05	1.832381e-05	9.492198e-01	3.391732e-02	9.651688e-01	0.001023	0.001012	1.558455
141	random_forest.csv	min_max_lof_euclidean_5.csv	40	9.651671e-01	-4.875294e-05	1.832381e-05	9.492198e-01	3.391732e-02	9.651688e-01	0.001023	0.001012	1.558455
143	random_forest.csv	min_max_lof_euclidean_5.csv	80	9.642115e-01	-4.646394e-05	1.678967e-05	9.507158e-01	3.361124e-02	9.642132e-01	0.001024	0.001051	4.482694
...	...	...	...	...	...	...	...	...	...	...	...	...
513	caracteristicas_polinomicas.csv	min_max_isf_300_auto.csv	5.0	-1.345321e+07	-1.535129e+12	2.506094e+12	-2.686721e+15	8.060164e+15	-1.345274e+07	3.643713	0.146269	2.167793
509	caracteristicas_polinomicas.csv	min_max_isf_200_0_2.csv	6.0	-1.468176e+07	-4.265908e+10	8.531817e+10	-3.750615e+11	1.125184e+12	-1.468012e+07	5.791233	1.321344	4.160230
514	caracteristicas_polinomicas.csv	min_max_isf_300_auto.csv	6.0	-9.733101e+07	-2.646631e+11	5.007891e+11	-2.679849e+15	7.721992e+15	-9.732756e+07	9.713769	0.152856	2.254314
512	caracteristicas_polinomicas.csv	min_max_isf_300_auto.csv	4.0	-3.973146e+08	-9.100035e+11	1.816549e+12	-4.110323e+12	1.233097e+13	-3.973006e+08	19.629966	0.150665	2.224935
508	caracteristicas_polinomicas.csv	min_max_isf_200_0_2.csv	5.0	-1.617705e+08	-4.983243e+10	9.966485e+10	-1.324205e+11	3.972615e+11	-1.617502e+08	20.378456	1.321657	4.162662
507	caracteristicas_polinomicas.csv	min_max_isf_200_0_2.csv	4.0	-2.458237e+08	-4.705203e+10	9.410406e+10	-2.624615e+10	7.873844e+10	-2.457965e+08	23.548082	3.042529	473879.705298
486	caracteristicas_polinomicas.csv	min_max_isf_200_auto.csv	3.0	-2.986149e+08	-1.681491e+12	3.234327e+12	-4.637424e+13	1.048308e+14	-2.986018e+08	26.556549	0.857615	3.662808
488	caracteristicas_polinomicas.csv	min_max_isf_200_auto.csv	5.0	-2.097475e+10	-1.571203e+11	1.914368e+11	-4.169882e+13	1.176699e+14	-2.097371e+10	234.606031	0.802040	3.584008
487	caracteristicas_polinomicas.csv	min_max_isf_200_auto.csv	4.0	-2.558194e+10	-3.396842e+11	4.844167e+11	-4.729045e+13	1.344526e+14	-2.558103e+10	258.507007	0.701926	3.431543
489	caracteristicas_polinomicas.csv	min_max_isf_200_auto.csv	6.0	-2.355996e+11	-6.616136e+11	7.637889e+11	-9.423520e+12	2.574301e+13	-2.355898e+11	789.440472	0.754523	3.513397

2553 rows × 12 columns

En ambas bases de datos, el modelo que obtuvo muy malas métricas fue el de regresión con características polinómicas, esto a pesar de que obtuvo el puesto quinto en ambas bases de datos como el modelo que tuvo mejor desempeño.

6.2.3 Mejor modelo

6.2.3.1 Medianos agricultores

TABLA XLIV
MEJOR RESULTADO PARA LOS MEDIANOS AGRICULTORES

	MODEL	NOMBRE	CONFIG	R2_MODEL	CROSS_ECM	CROSS_ECM_DE	CROSS_R2	CROSS_R2_DE	VARIANZA_EXPL	MAE	MAE_POISSON	MAE_GAMMA
278	huber_regressor.csv	min_max_isf_100_auto.csv	1.45	0.82529	-0.000013	0.000001	0.825338	0.033429	0.850171	0.001335	0.111238	105194.122269

La base de datos con la que se entrenó este modelo fue una de las que más datos atípicos eliminó tal cual como se puede observar en la tabla [**TABLA XXII**], con 100 estimadores y la contaminación automática. Por otro lado, vemos que el modelo que mejor se adaptó a estos datos fue el Huber con una épsilon de 1.45, por lo que vemos la épsilon es medianamente alto por lo cual

el proceso de eliminación de datos atípicos si fue efectivo, debido a que cuanto más pequeño es la épsilon, más robusto es para los valores atípicos.

6.2.3.2 Grandes agricultores

TABLA XLV
MEJOR RESULTADO PARA LOS GRANDES AGRICULTORES

	MODEL	NOMBRE	CONFIG	R2_MODEL	CROSS_ECM	CROSS_ECM_DE	CROSS_R2	CROSS_R2_DE	VARIANZA_EXPL	MAE	MAE_POISSON	MAE_GAMMA
154	random_forest.csv	min_max_lof_manhattan_5.csv	100	0.973404	-0.000041	0.000023	0.959706	0.028319	0.973404	0.001003	0.000754	14.667231

Al igual que en los pequeños agricultores, para los grandes agricultores la base de datos con la que se entrenó este modelo fue la que más datos atípicos eliminó con el método de factor atípico local (LOF) tal cual como se puede observar en la tabla [**TABLA XIII**], con una distancia de manhattan y 5 vecinos, en cuanto al modelo, el que mejor se comportó con la base de datos fue un random forest con 100 árboles. El MAE que se obtuvo con este modelo es menor al modelo obtenido con los medianos agricultores y ambas bases de datos coincidieron con el método de escalamiento.

6.2.4 Peor modelo

6.2.4.1 Medianos agricultores

TABLA XLVI
DATOS DE LA BASE DE DATOS

	MODEL	NOMBRE	CONFIG	R2_MODEL	CROSS_ECM	CROSS_ECM_DE	CROSS_R2	CROSS_R2_DE	VARIANZA_EXPL	MAE	MAE_POISSON	MAE_GAMMA
439	caracteristicas_polinomicas.csv	estandar_isf_400_auto.csv	6.0	-3.342044e+21	-6.267935e+21	7.986959e+21	-7.126881e+21	1.187990e+22	-3.341438e+21	940187345.8	0.5	4.0

Al observar la configuración con la que esté modelos tuvo tan malas métricas, se observa que la base de datos con la que fue entrenada se encuentra casi en la antepenúltima posición de las bases de datos que más datos atípicos eliminaron, tal como se observa en la tabla [**TABLA XVIII**], con 400 estimadores y la contaminación automática. Por el lado del modelo vemos que aumentando a un exponente de grado 6 este modelo con característica polinómicas da como resultado el modelo con las peores métricas con la base de datos de medianos agricultores. Adicionalmente vemos que el MAE llega a 9 cifras, un error extremadamente alto.

6.2.4.2 Grandes agricultores

TABLA XLVII
DATOS DE LA BASE DE DATOS

	MODEL	NOMBRE	CONFIG	R2_MODEL	CROSS_ECM	CROSS_ECM_DE	CROSS_R2	CROSS_R2_DE	VARIANZA_EXPL	MAE	MAE_POISSON	MAE_GAMMA
489	caracteristicas_polinomicas.csv	min_max_isf_200_auto.csv	6.0	-2.355996e+11	-6.616136e+11	7.637889e+11	-9.423520e+12	2.574301e+13	-2.355898e+11	789.4	0.8	3.5

Con respecto a los resultados obtenidos con los grandes agricultores, vemos que la base de datos con la que se obtuvo tan mal desempeño ocupa la penúltima posición de las bases de datos que más datos atípicos eliminaron en el método de aislamiento forestal con 200 estimadores y la contaminación automática, tal como se observa en la tabla [**TABLA XXI**]. Ahora si vamos a ver con cual modelo se entrenó, resulta ser el mismo y con la misma configuración que en los medianos agricultores. A comparación de las métricas obtenidas en los medianos, vemos que el mae en este caso solo tiene 3 a comparación de los 9 que resultó en los medianos agricultores.

6.3 CONSIDERACIONES DE PRODUCCIÓN

Inicialmente se buscará poner el modelo que mejor métrica obtuvo en cada base de datos, para esto se va a usar un framework que nos ayude a desarrollar la aplicación de una forma más sencilla y rápida, aquí entra **Flask [24]** y se alojará en un host de **Heroku [25]** aprovechando su sistema de almacenamiento fácil de implementar y además económico.

En esta página web se tendrá un apartado en donde cada usuario podrá ingresar los resultados que obtuvo a partir de la estimación y lo resultados verdaderos que obtuvo durante la cosecha, esto para poder evaluar el desempeño del modelo de una forma más real y a medida que se va recibiendo esta retroalimentación se irá alimentando y ajustando el modelo, de tal forma que al cabo de unos años, se tenga un modelo confiable y que ayude a los campesinos en la selección de la cantidad de hectáreas a sembrar.

En paralelo para los administradores del sistema, se tendrá una pantalla de monitoreo, que se estará alimentando de la retroalimentación que haga el campesino, por otro lado, el administrador podrá agregar año por año cuáles factores afectan o benefician la cosecha, aspecto como el abono, el clima, la forma de sembrado, entre otros; Esto con el fin de aumentar el número de características que puedan darle un poco más de robustez al modelo.

7. CONCLUSIONES

- Es de suma importancia abarcar un gran número de horas conocer los datos, debido a que estos tendrán un gran impacto a la hora de entrenar nuestros modelos.
- Analizar a fondo cada una de las características y en especial la característica objetivo, debido a que esta nos puede arrojar posibles soluciones a los problemas de escalabilidad de los datos.
- Con respecto a la división de datos, se deja iniciado el proceso, para que la variable de salida, se le realice un proceso de agrupamiento para observar la cantidad de grupos óptimos que podrían salir.
- A pesar de entrenar un modelo sencillo de deep learning, no se obtuvieron los resultados esperados, esto debido a que es necesario construir un modelo mucho más robusto con otras herramientas como keras de tensor-flow.
- Los mejores resultados se obtuvieron con la base de datos de los grandes agricultores, pero a pesar de que los medianos agricultores tenían el tercio de los grandes, sus resultados fueron buenos.
- Es necesario agregar más información que nos ayude a entrenar los modelos de una forma más real, esto se puede hacer asociando cada uno de las variables categóricas con valores que me aportan consistencia a los modelos como las precipitaciones por municipio.
- El objetivo con el que se inició esta investigación se completó de forma exitosa, por lo tanto se invita al lector a realizar múltiples entrenamientos con otros modelos y series temporales con estos datos, debido a que demuestra que aporta mucha información necesaria para la estimación de las áreas.
- Variar los escaladores y eliminadores de datos atípicos, permite tener múltiples bases de datos que permite un mejor estudio, en nuestro caso, esto ayudó a comprender que la base de datos de medianos campesinos hay más presencia de datos atípicos, esto debido a que se tuvo un mejor resultado cuando se eliminó un 25% de la base de datos original, las cuales correspondía a datos atípicos, caso contrario en los grandes donde no hubo la necesidad de eliminar muchos registros de la base de datos.

- A pesar de que los medianos agricultores y grandes agricultores provienen de la misma base de datos, ambos tuvieron desempeños muy diferentes, no obstante, sí coincidieron en algunos métodos de preprocesamiento a la hora de obtener el mejor resultado.
- El mejor escalado aplicado a ambas bases de datos fue el min-max, debido a que con este escalado se obtuvo el mejor desempeño en ambas bases de datos, a pesar de que el peor resultado obtenido por un modelo también se haya escalado con este método, se observó que este mal resultado se obtuvo por la configuración del modelo, mas no del método de escalamiento seleccionado.
- Con esta investigación se puede observar claramente que la selección de la base de datos es fundamental a la hora de encontrar un modelo con unas buenas métricas.

8. REFERENCIAS

[0] BIBLIOTECA DE DOCUMENTOS – PILOTOS DE INNOVACIÓN FINANCIERA. Asobancaria. [En línea]. Disponible en: https://www.asobancaria.com/2016/02/01/pilotos-de-innovacion/ [Último acceso: 2022].

[1] Términos y condiciones | Datos Abiertos. (2022). Términos y condiciones | Datos Abiertos. [En línea]. Disponible en: https://herramientas.datos.gov.co/terminos [Último acceso: 2022].

[3] "sklearn.linear_model.LinearRegression". scikit-learn. [En línea]. Disponible en: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html [Último acceso: 2022].

[4] sklearn.linear_model.RANSACRegressor. (s. f.). scikit-learn. [En línea]. Disponible en: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.RANSACRegressor.html#sklearn.linear_model.RANSACRegressor [Último acceso: 2022].

[5] sklearn.preprocessing.PolynomialFeatures. (s. f.). scikit-learn. [En línea]. Disponible en: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.PolynomialFeatures.html [Último acceso: 2022].

[6] sklearn.ensemble.RandomForestRegressor. (s. f.). scikit-learn. [En línea]. Disponible en: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html [Último acceso: 2022].

[7] sklearn.neural_network.MLPRegressor. (s. f.). scikit-learn. [En línea]. Disponible en: https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPRegressor.html [Último acceso: 2022].

[8] sklearn.metrics.explained_variance_score. (s. f.). scikit-learn. [En línea]. Disponible en: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.explained_variance_score.html [Último acceso: 2022].

[9] sklearn.metrics.mean_absolute_error. (s. f.). scikit-learn. [En línea]. Disponible en: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_absolute_error.html [Último acceso: 2022].

[10] sklearn.metrics.mean_poisson_deviance. (s. f.). scikit-learn. [En línea]. Disponible en: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_poisson_deviance.html [Último acceso: 2022].

[11] sklearn.metrics.mean_gamma_deviance. (s. f.). scikit-learn. [En línea]. Disponible en: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_gamma_deviance.html [Último acceso: 2022].

[12] Google Colab. (s. f.). Google Research. [En línea]. Disponible en: https://research.google.com/colaboratory/faq.html#:~:text=Colaboratory,%20or%20“Colab”%20for,learning,%20data%20analysis%20and%20education. [Último acceso: 2022].

[13] About us. (s. f.). scikit-learn. [En línea]. Disponible en: https://scikit-learn.org/stable/about.html#citing-scikit-learn [Último acceso: 2022].

[14] What is NumPy? — NumPy v1.24.dev0 Manual. (s. f.). NumPy. [En línea]. Disponible en: https://numpy.org/devdocs/user/whatisnumpy.html [Último acceso: 2022].

[15] Package overview — pandas 1.4.2 documentation. (s. f.). pandas - Python Data Analysis Library. [En línea]. Disponible en: https://pandas.pydata.org/docs/getting_started/overview.html [Último acceso: 2022].

[16] sklearn.model_selection.cross_val_score. (s. f.). scikit-learn. [En línea]. Disponible en: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.cross_val_score.html [Último acceso: 2022].

[17] seaborn: statistical data visualization — seaborn 0.11.2 documentation. (s. f.). seaborn: statistical data visualization — seaborn 0.11.2 documentation. [En línea]. Disponible en: https://seaborn.pydata.org/ [Último acceso: 2022].

[18] History — Matplotlib 3.5.2 documentation. (s. f.). Matplotlib — Visualization with Python. [En línea]. Disponible en: https://matplotlib.org/stable/users/project/history.html [Último acceso: 2022].

[19] sklearn.linear_model.TheilSenRegressor. (s. f.). scikit-learn. [En línea]. Disponible en: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.TheilSenRegressor.html [Último acceso: 2022].

[20] sklearn.linear_model.HuberRegressor. (s. f.). scikit-learn. [En línea]. Disponible en: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.HuberRegressor.html [Último acceso: 2022].

[21] sklearn.ensemble.HistGradientBoostingRegressor. (s. f.). scikit-learn. [En línea]. Disponible en: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.HistGradientBoostingRegressor.html [Último acceso: 2022].

[22] sklearn.neighbors.LocalOutlierFactor. (s. f.). scikit-learn. [En línea]. Disponible en: https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.LocalOutlierFactor.html [Último acceso: 2022].

[23] sklearn.ensemble.IsolationForest. (s. f.). scikit-learn. [En línea]. Disponible en: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.IsolationForest.html [Último acceso: 2022].

[24] Welcome to Flask — Flask Documentation (2.1.x). (s. f.). Welcome to Flask — Flask Documentation (2.1.x). [En línea]. Disponible en: https://flask.palletsprojects.com/en/2.1.x/ [Último acceso: 2022].

[25] About Heroku | Heroku. (s. f.). Cloud Application Platform | Heroku. [En línea]. Disponible en: https://www.heroku.com/about [Último acceso: 2022].

[26] sklearn.preprocessing.LabelEncoder. (s. f.). scikit-learn. [En línea]. Disponible en: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html [Último acceso: 2022].