Aplicación de Modelos de ML para lograr predecir el rendimientos de Arándanos Silvestres

- Mayumy Carrasco Huaccha -

- Mayumy Carrasco Huaccha -

Lic. Scientific Computing ||Datalover

Analytic Projections in Camposol

MayumyCH

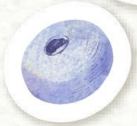
MayumyCH

Mayumy Carrasco Huaccha

01.

Comprensión del Negocio y los datos

El desafío mas grande que se tiene en el sector de la agricultura es lograr ***predecir el rendimiento de los cultivos.***

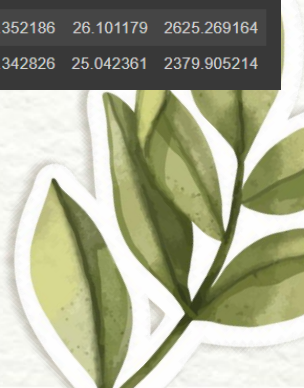


Obtención y carga de la data

Kaggle – Articulo ***"Wild blueberry yield prediction using a combination of computer simulation and machine learning algorithms"***

	clonesize	honeyBee	bumblesBee	andrenaBee	osmiaBee	MaxTempBS	MinTempBS	AverageTempBS	MaxTempBI	MinTempBI	AverageTempBI	RainingDays	AverageRainingDays	fruitset	fruitmass	seeds	yield
0	37.5	0.75	0.25	0.25	0.25	86.0	52.0	71.9	62.0	30.0	50.8	16.0	0.26	0.410652	0.408159	31.678898	3813.165795
1	37.5	0.75	0.25	0.25	0.25	86.0	52.0	71.9	62.0	30.0	50.8	1.0	0.10	0.444254	0.425458	33.449385	4947.605663
2	37.5	0.75	0.25	0.25	0.25	94.6	57.2	79.0	68.2	33.0	55.9	16.0	0.26	0.383787	0.399172	30.546306	3866.798965
3	37.5	0.75	0.25	0.25	0.25	94.6	57.2	79.0	68.2	33.0	55.9	1.0	0.10	0.407564	0.408789	31.562586	4303.943030
4	37.5	0.75	0.25	0.25	0.25	86.0	52.0	71.9	62.0	30.0	50.8	24.0	0.39	0.354413	0.382703	28.873714	3436.493543
5	37.5	0.75	0.25	0.25	0.25	86.0	52.0	71.9	62.0	30.0	50.8	34.0	0.56	0.309669	0.366284	27.345454	2825.003738
6	37.5	0.75	0.25	0.25	0.25	94.6	57.2	79.0	68.2	33.0	55.9	24.0	0.39	0.284443	0.352186	26.101179	2625.269164
7	37.5	0.75	0.25	0.25	0.25	94.6	57.2	79.0	68.2	33.0	55.9	34.0	0.56	0.246568	0.342826	25.042361	2379.905214

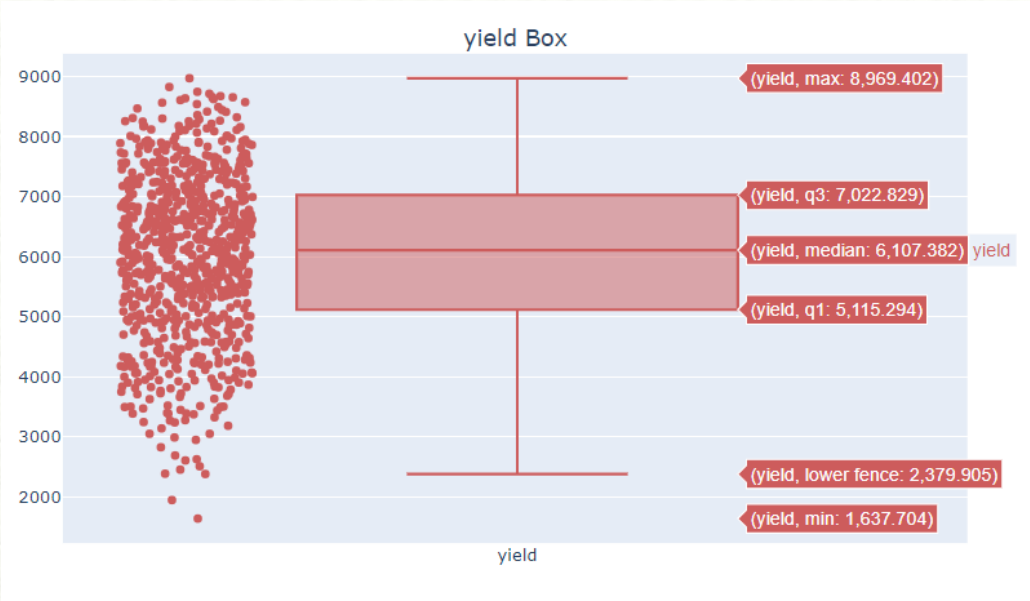
777 registros / 17 columnas



02.

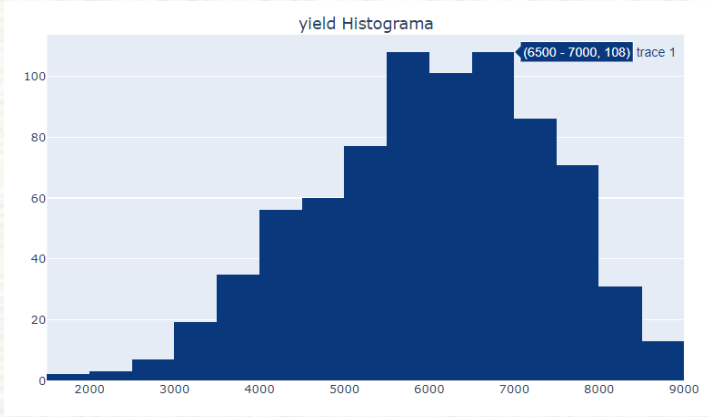
EDA y Preparación de la data

Variable target: "Yield"

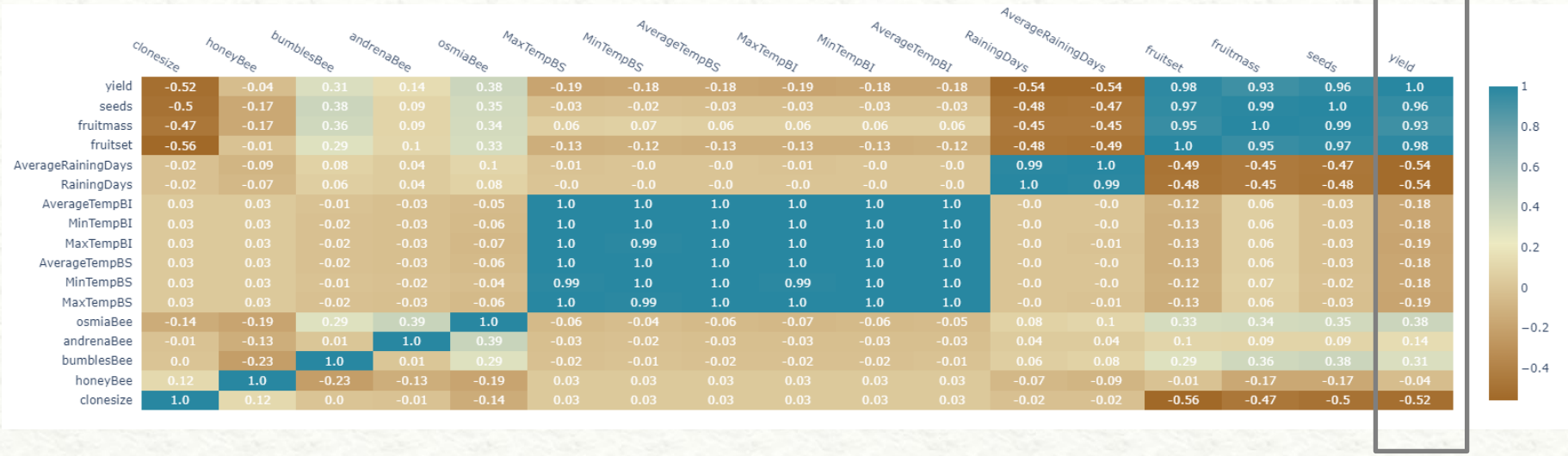


50%

Del rendimiento obtenido esta entre 5115 y 7022 kilos/Ha.



Análisis de correlación

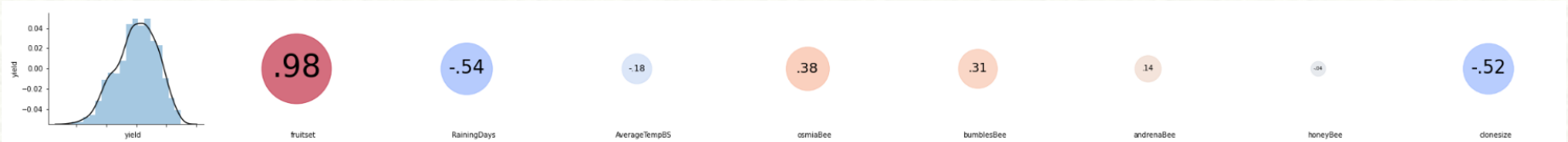


	clonesize	honeyBee	bumblesBee	andrenaBee	osmiaBee	MaxTempBS	MinTempBS	AverageTempBS	MaxTempBI	MinTempBI	AverageTempBI	RainingDays	AverageRainingDays	fruitset	fruitmass	seeds	yield
yield	-0.52	-0.04	0.31	0.14	0.38	-0.19	-0.18	-0.18	-0.19	-0.18	-0.18	-0.54	-0.54	0.98	0.93	0.96	1.0
seeds	-0.5	-0.17	0.38	0.09	0.35	-0.03	-0.02	-0.03	-0.03	-0.03	-0.03	-0.48	-0.47	0.97	0.99	1.0	0.96
fruitmass	-0.47	-0.17	0.36	0.09	0.34	0.06	0.07	0.06	0.06	0.06	0.06	-0.45	-0.45	0.95	1.0	0.99	0.93
fruitset	-0.56	-0.01	0.29	0.1	0.33	-0.13	-0.12	-0.13	-0.13	-0.13	-0.12	-0.48	-0.49	1.0	0.95	0.97	0.98
AverageRainingDays	-0.02	-0.09	0.08	0.04	0.1	-0.01	-0.0	-0.0	-0.01	-0.0	-0.0	0.99	1.0	-0.49	-0.45	-0.47	-0.54
RainingDays	-0.02	-0.07	0.06	0.04	0.08	-0.0	-0.0	-0.0	-0.0	-0.0	-0.0	1.0	0.99	-0.48	-0.45	-0.48	-0.54
AverageTempBI	0.03	0.03	-0.01	-0.03	-0.05	1.0	1.0	1.0	1.0	1.0	1.0	-0.0	-0.0	-0.12	0.06	-0.03	-0.18
MinTempBI	0.03	0.03	-0.02	-0.03	-0.06	1.0	1.0	1.0	1.0	1.0	1.0	-0.0	-0.0	-0.13	0.06	-0.03	-0.18
MaxTempBI	0.03	0.03	-0.02	-0.03	-0.07	1.0	0.99	1.0	1.0	1.0	1.0	-0.0	-0.01	-0.13	0.06	-0.03	-0.19
AverageTempBS	0.03	0.03	-0.02	-0.03	-0.06	1.0	1.0	1.0	1.0	1.0	1.0	-0.0	-0.0	-0.13	0.06	-0.03	-0.18
MinTempBS	0.03	0.03	-0.01	-0.02	-0.04	0.99	1.0	1.0	0.99	1.0	1.0	-0.0	-0.0	-0.12	0.07	-0.02	-0.18
MaxTempBS	0.03	0.03	-0.02	-0.03	-0.06	1.0	0.99	1.0	1.0	1.0	1.0	-0.0	-0.01	-0.13	0.06	-0.03	-0.19
osmiaBee	-0.14	-0.19	0.29	0.39	1.0	-0.06	-0.04	-0.06	-0.07	-0.06	-0.05	0.08	0.1	0.33	0.34	0.35	0.38
andrenaBee	-0.01	-0.13	0.01	1.0	0.39	-0.03	-0.02	-0.03	-0.03	-0.03	-0.03	0.04	0.04	0.1	0.09	0.09	0.14
bumblesBee	0.0	-0.23	1.0	0.01	0.29	-0.02	-0.01	-0.02	-0.02	-0.02	-0.01	0.06	0.08	0.29	0.36	0.38	0.31
honeyBee	0.12	1.0	-0.23	-0.13	-0.19	0.03	0.03	0.03	0.03	0.03	0.03	-0.07	-0.09	-0.01	-0.17	-0.17	-0.04
clonesize	1.0	0.12	0.0	-0.01	-0.14	0.03	0.03	0.03	0.03	0.03	0.03	-0.02	-0.02	-0.56	-0.47	-0.5	-0.52

Se observa que hay features que sean correlacionados entre sí:
Ejemplo: Clima de las banda superior e inferior y features del fruto.

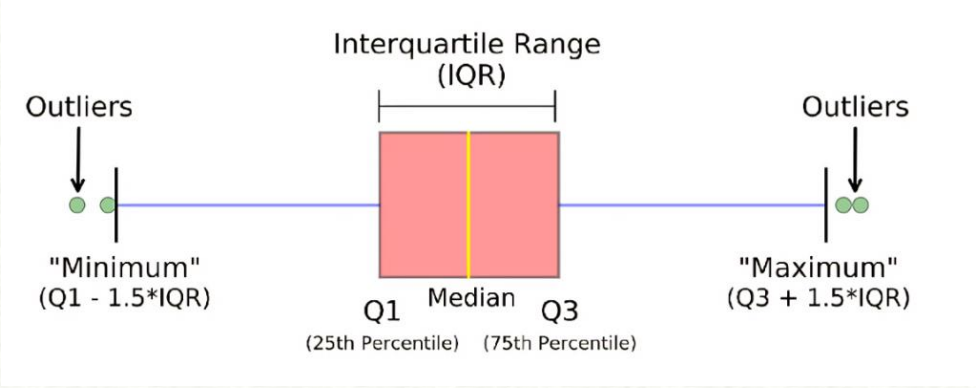
Variables Predictoras

	fruitset	RainingDays	AverageTempBS	osmiaBee	bumblesBee	andrenaBee	honeyBee	clonesize
0	0.410652	16.0	71.9	0.25	0.25	0.25	0.75	37.5
1	0.444254	1.0	71.9	0.25	0.25	0.25	0.75	37.5
2	0.383787	16.0	79.0	0.25	0.25	0.25	0.75	37.5
3	0.407564	1.0	79.0	0.25	0.25	0.25	0.75	37.5
4	0.354413	24.0	71.9	0.25	0.25	0.25	0.75	37.5
5	0.309669	34.0	71.9	0.25	0.25	0.25	0.75	37.5

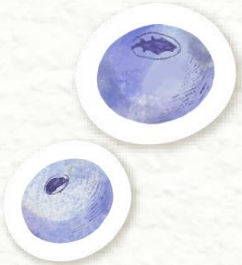


Eliminar los outliers de nuestra data

```
df_blueberry = df_blueberry[~((df_blueberry < (q1 - 1.5 * iqr)) |(df_blueberry > (q3 + 1.5 * iqr))).any(axis=1)]
```



752 registros / 9 columnas



03.

Modelamiento y Evaluación

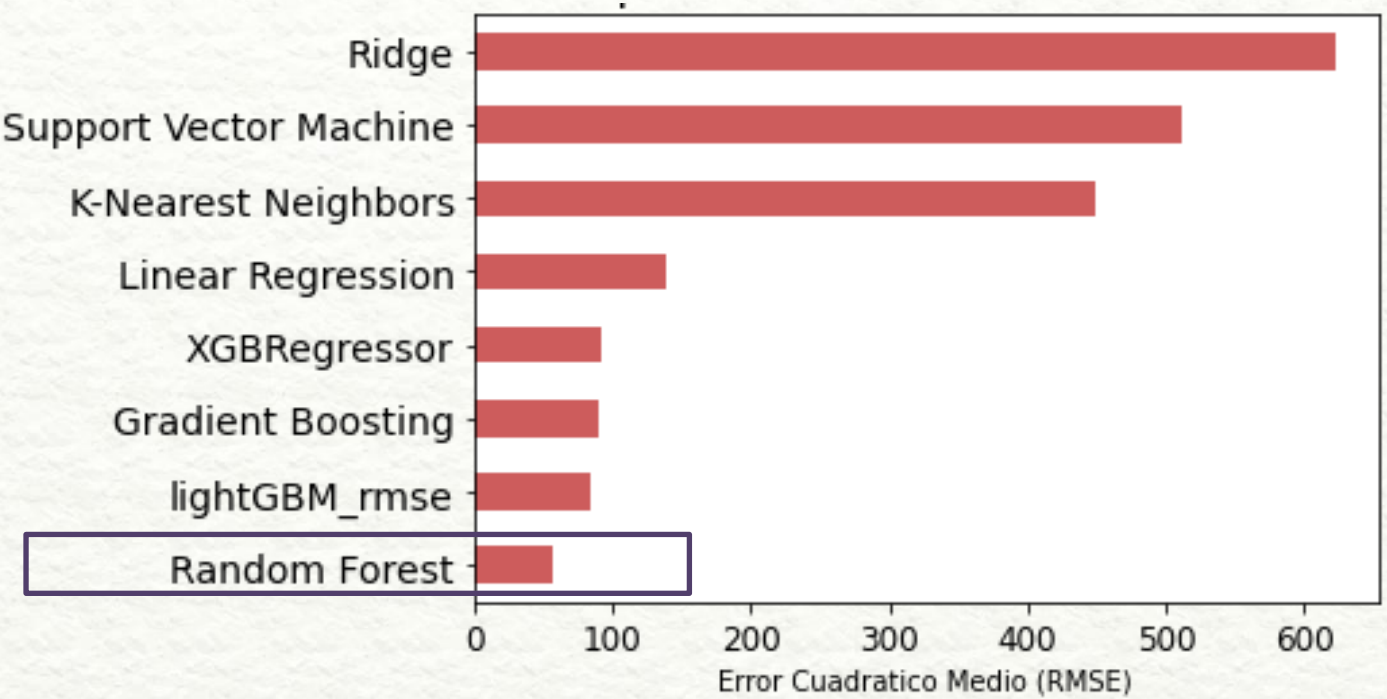
Datos de Entrenamiento y Test

```
# Separar Variables
X = df_blueberry.drop(columns="yield") # Variables predictoras
y = df_blueberry["yield"] # Variables Target o Objetivo

X_train, X_test, y_train, y_test = train_test_split(X, y,
                                                    test_size=0.3, random_state=88)
```

```
print("Tamaño del X_train",X_train.shape, "\nTamaño del X_test ",X_test.shape)
print("--")
print("Tamaño del y_train",y_train.shape, "\nTamaño del y_test ",y_test.shape)
```

```
Tamaño del X_train (526, 8)
Tamaño del X_test  (226, 8)
--
Tamaño del y_train (526,)
Tamaño del y_test  (226,)
```

Comparando los modelos con métrica RMSE

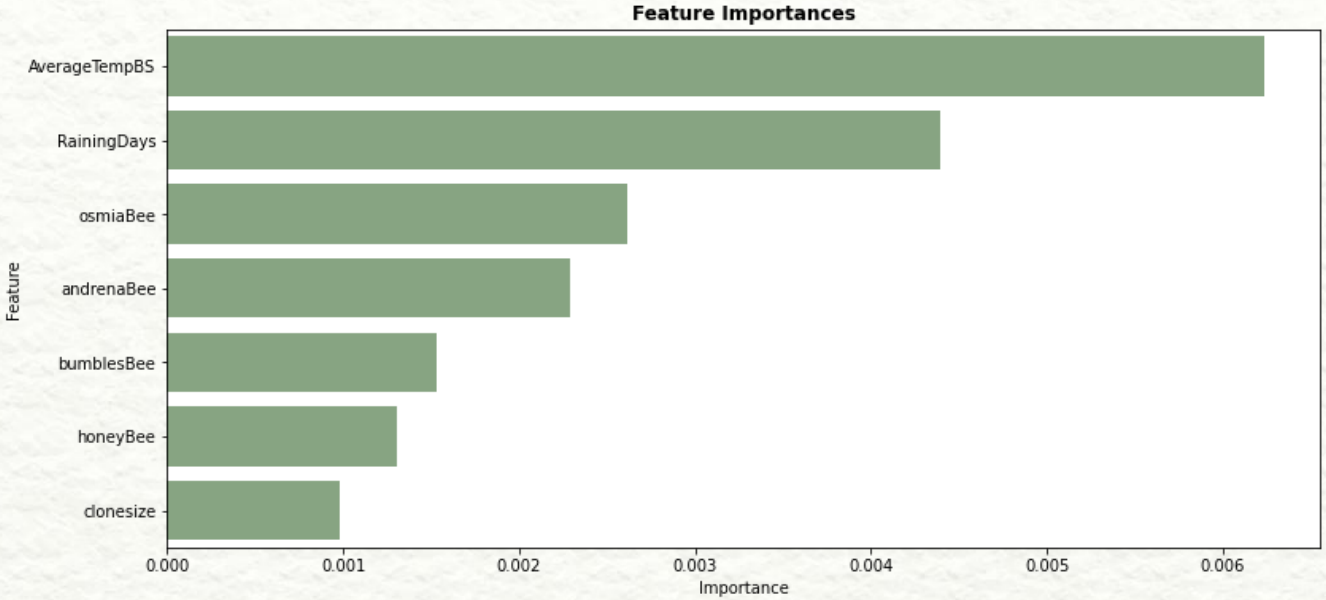


Modelo escogido

```
# Los mejores parametros:
random_forest_regresor = RandomForestRegressor(random_state=0,n_estimators=200,max_depth=10)
random_forest_rmse = fit_and_evaluate(random_forest_regresor,True)

model_pred_train = random_forest_regresor.predict(X_train)
model_pred_test = random_forest_regresor.predict(X_test)
```

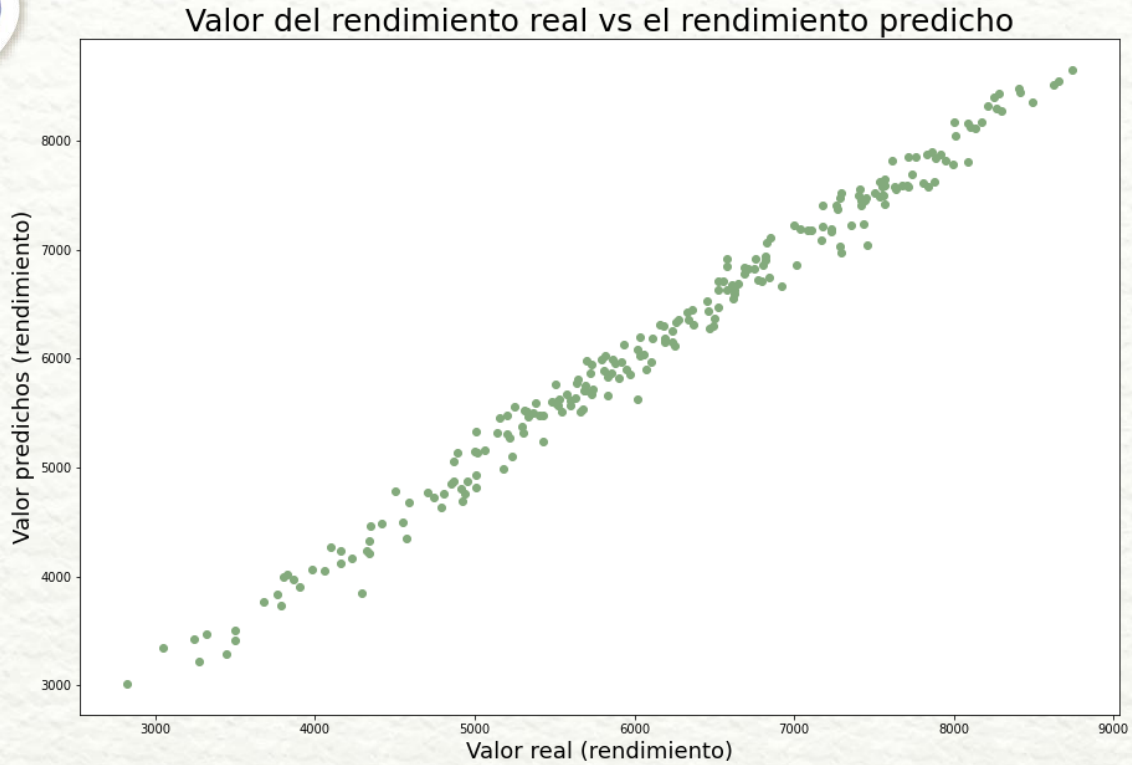
```
Calculando los errores en la data del TRAIN
-------------------------------------------------
Linear Regression Performance on the test set: MAE = 45.59
Linear Regression Performance on the test set: MSE = 3349.45
Linear Regression Performance on the test set: RMSE = 57.87
r2_square= 1.00

Calculando los errores en la data del TEST
-------------------------------------------------
Linear Regression Performance on the test set: MAE = 116.42
Linear Regression Performance on the test set: MSE = 21264.28
Linear Regression Performance on the test set: RMSE = 145.82
r2_square= 0.99
```

Importancia de los features



El feature *fruitset* es el que tiene mas importancia.

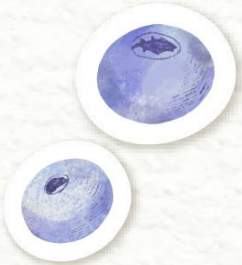
Modelo escogido



El modelo predice con un error de

$\pm$ 57.10 kilos/ha

(rendimiento de los arándanos silvestres)

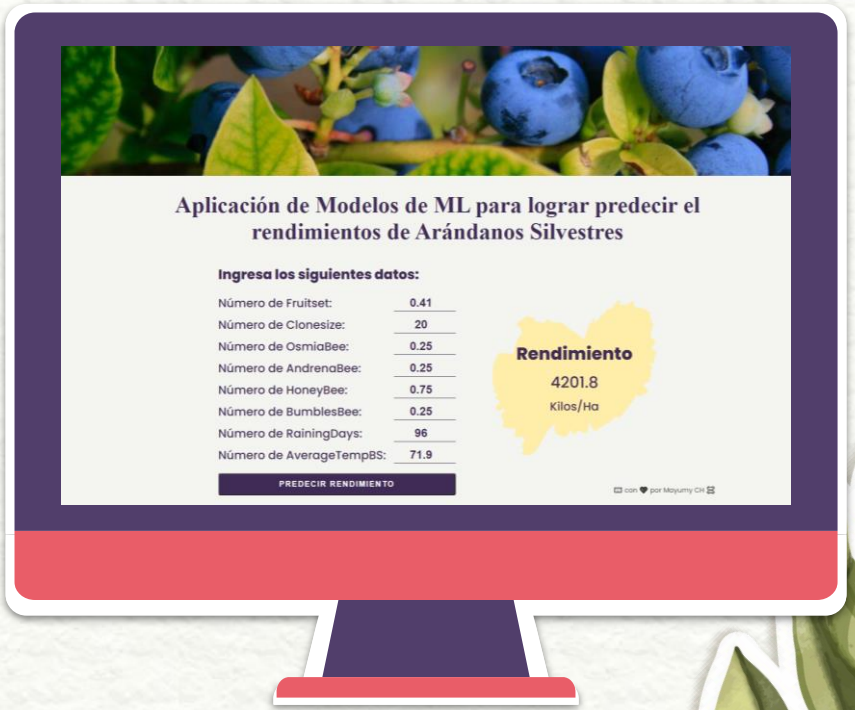


04.

Despliegue

DEMO del Modelo

- Api creado con **FastApi** para consumir el modelo.
- Frontend con **Angular** para correr el modelo



Gracias totales

Agradecimiento especial

Bootcamp de Codigo Facilito