PROYECTO FINAL

CARRERA DATA SCIENTIST – CODERHOUSE – EMILSE BOVER- JULIO 2023

ANÁLISIS DE PRODUCCIÓN DE POLLOS PARRILLEROS

Definición de objetivo

- El objetivo del presente trabajo es detectar el impacto de los factores que afectan al crecimiento de los pollos parrilleros para tomar decisiones que mejoren los resultados productivos.

Contexto comercial

- El retorno económico de la empresa productora de pollos parrilleros depende de la obtención de mejores resultados productivos. A mayor ganancia de peso diaria, por ejemplo, el consumo total de alimento será menor disminuyendo el costo. Otro impacto positivo del aumento de la ganancia de peso diaria es que se reduce la edad a faena de los animales ya que se alcanza el peso deseado en menor cantidad de días. Esta reducción de días permite: liberación de superficie de producción (metros cuadrados de galpón, rotación), menor propensión a sufrir enfermedades y accidentes (por ejemplo, cortes de luz). También el dueño de la granja tiene menores costos de luz y gas y por lo tanto obtiene mayor retorno económico al final de la crianza.

Motivación y audiencia

- La motivación del presente trabajo es hallar relaciones entre los datos provistos por la empresa para mejorar los resultados productivos, está dirigido a los directivos y los mandos superiores encargados de ejecutar las acciones de cada área productiva.

Problema comercial

- La empresa plantea los siguientes interrogantes: ¿Por qué la ganancia de peso no es similar en todas las granjas? ¿Por qué tampoco es uniforme a lo largo del año? ¿Es posible predecir el crecimiento en el futuro? ¿El consumo de alimento es el esperado? ¿Qué tiene mayor impacto, el consumo de alimento o las condiciones del ambiente? ¿Es posible mejorar los valores de ganancia de peso manteniendo iguales las condiciones de genética, nutrición y ambiente (en el sentido estricto de la localización geográfica de las granjas)?

Contexto analítico

- La empresa ha provisto un archivo Excel con información de resultados y otros parámetros con localización geográfica, tipo de granja y zona de producción.

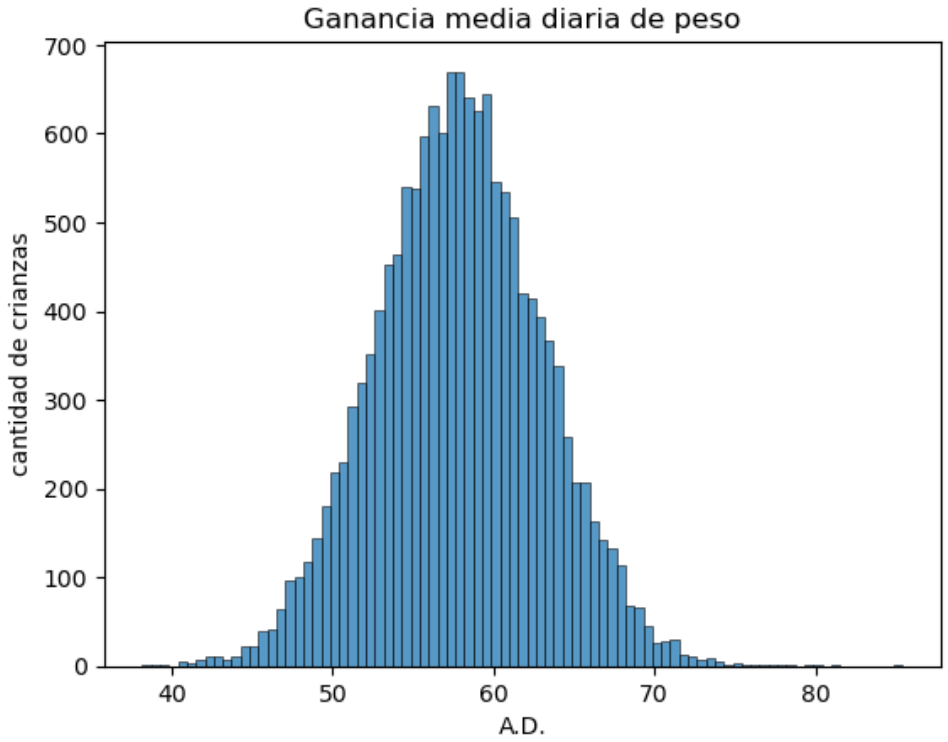
ÍNDICE DE CONTENIDO

ANÁLISIS EXPLORATORIO DE DATOS

	granja	Primer BB	Cantid. BB	A Faena	% a 7 Días	% Mortan.	% Fal.	Kg. Pollo	P. Prom.	Kg. Alim.	...	dada de baja	Cuartil EFS 2019	% desvío consumo std	EFS según ranking 19	Índice	zona climática	Cuartil EFS 2021	consumo total	Pes conv 2,7	zona prod
0	2544	2014-11-13	36000	34964	0.55	2.88	-0.33	93360	2.670175	198673	...	NaN	I	-10.389276	NaN	6	Arrecifes	NaN	5.682216	1.262878	BA
1	2558	2014-11-07	57500	54217	0.83	5.71	-0.26	145260	2.679233	335800	...	NaN	IV	-14.867094	NaN	7	Arrecifes	NaN	6.193629	1.164476	BA
2	2761	2014-11-13	39000	36483	0.92	6.45	-0.48	100360	2.750870	228500	...	NaN	I	-1.227077	NaN	8	25 de mayo	NaN	6.263191	1.194771	BA
3	2524	2014-11-14	30000	28231	1.35	5.90	1.07	77514	2.745705	176157	...	NaN	I	-1.595289	NaN	15	Arrecifes	NaN	6.239843	1.196094	BA
4	2741	2014-11-14	42000	38912	0.76	7.35	0.19	120920	3.107525	274466	...	x	NaN	3.530095	NaN	21	Arrecifes	NaN	7.053505	1.265246	BA

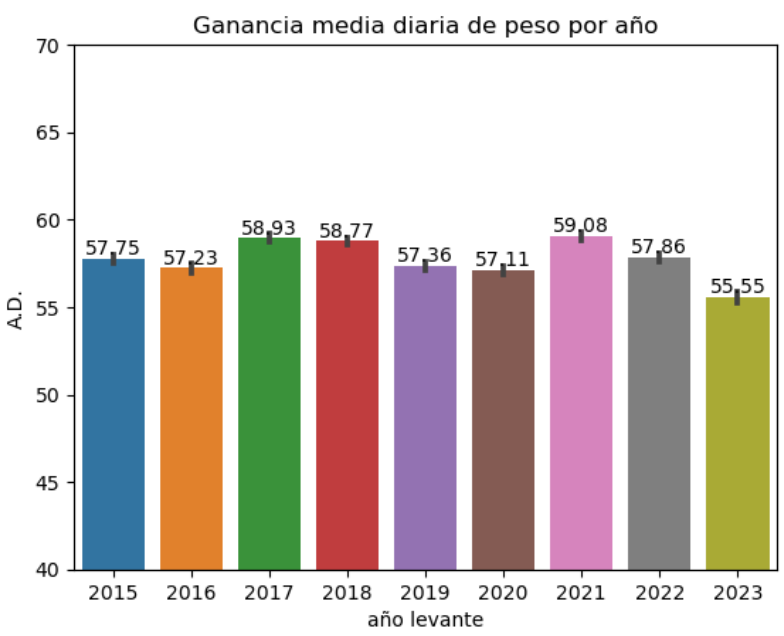
La base de datos posee 14.844 filas, cada una de las cuáles representa una crianza de una granja. La base tiene 61 columnas con información sobre las características de la granja, su ubicación y los resultados de producción obtenidos. Posee valores desde el año 2015 hasta el primer semestre del año 2023.

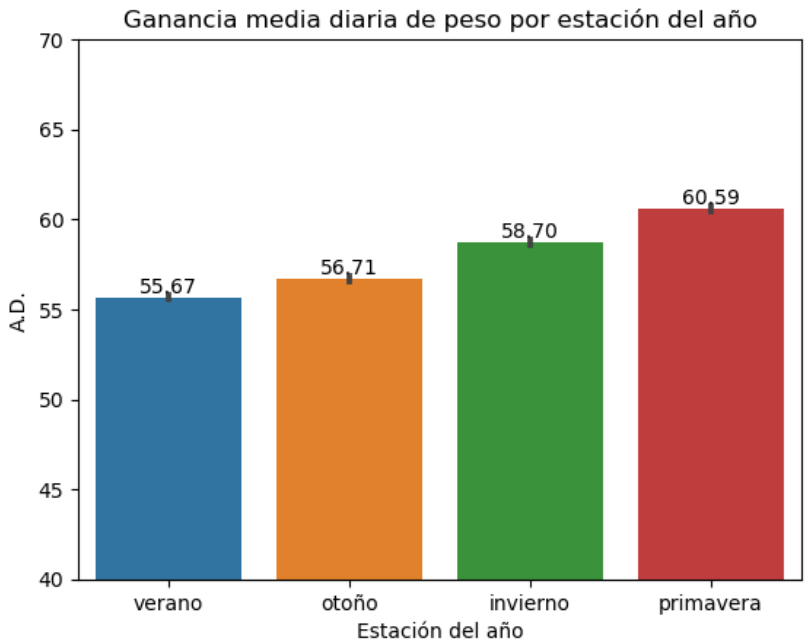
ANÁLISIS EXPLORATORIO DE DATOS

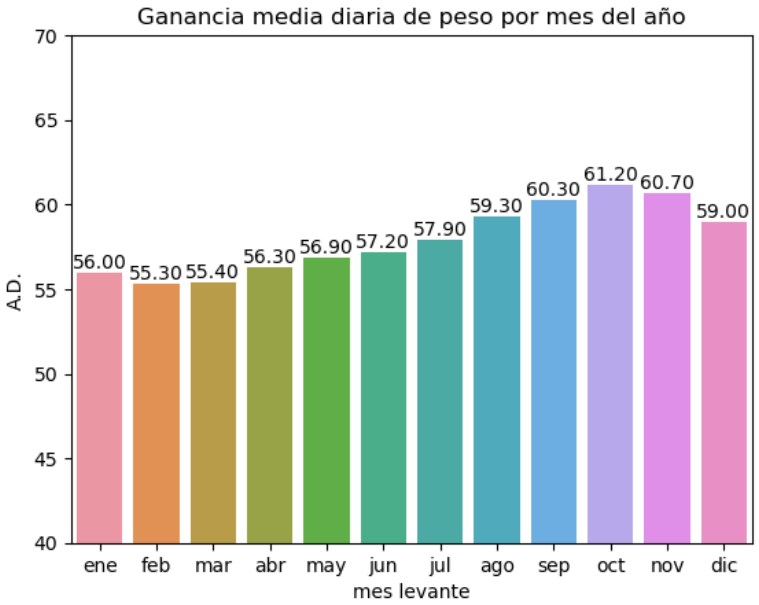


La **ganancia media diaria de peso** es una medida que permite comparar resultados entre granjas y además permite ver la evolución de los resultados globales de la empresa a lo largo del tiempo. A mayor valor, mejor será el resultado. El valor medio es de **57.84 gramos** de peso ganados promedio por día y los valores se distribuyen de manera normal.

ANÁLISIS EXPLORATORIO DE DATOS

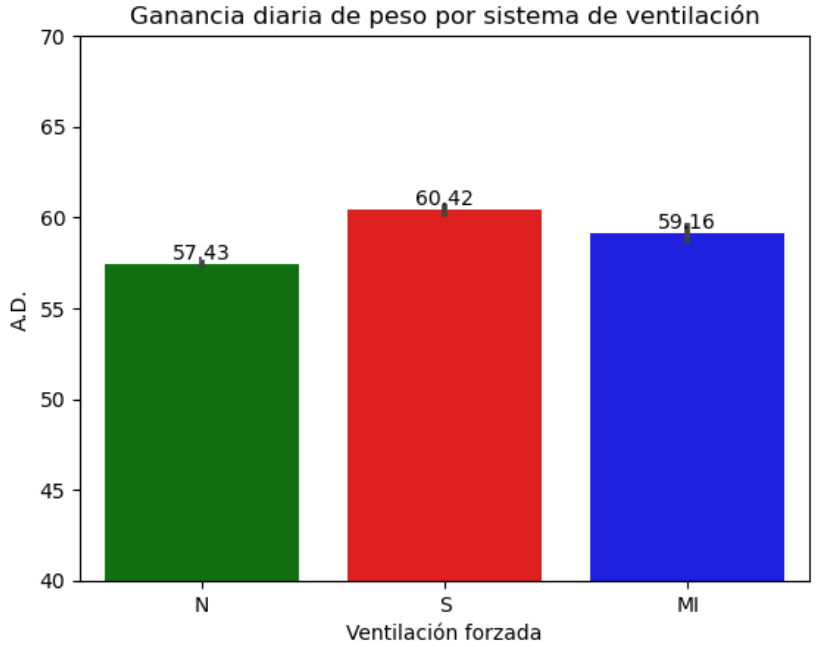


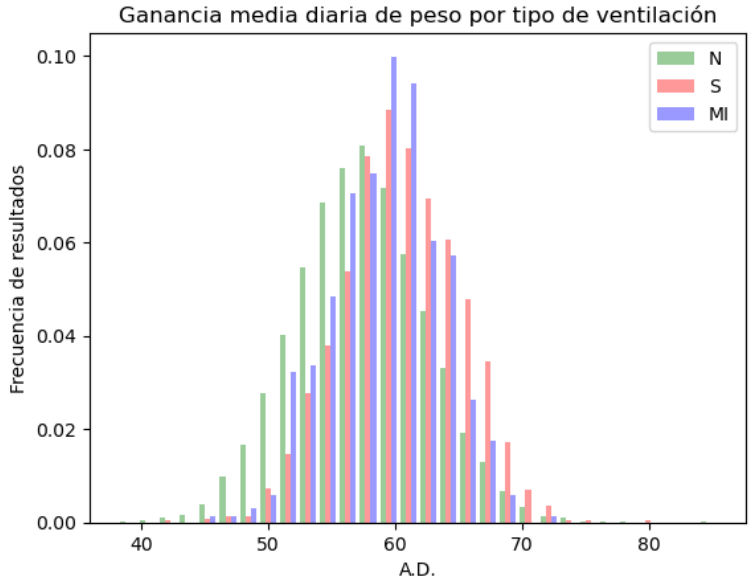


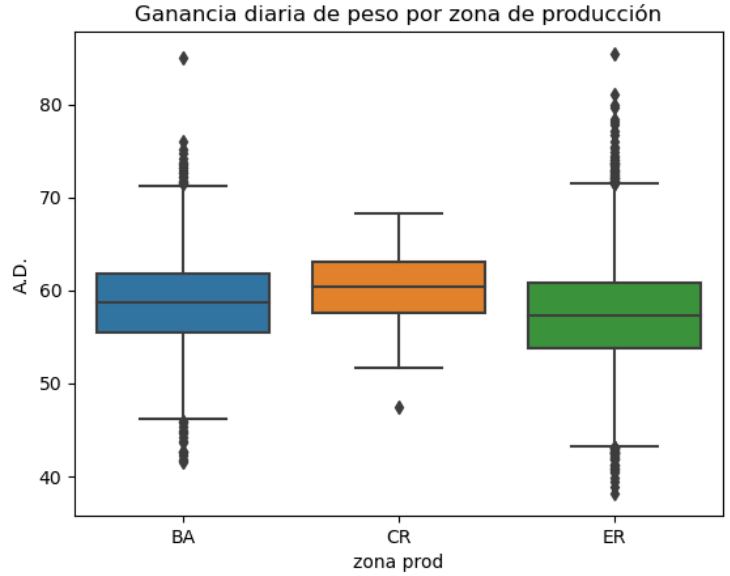


La **ganancia media diaria de peso** promedio es diferente entre años, estaciones del año y meses del año en que fueron criados los animales dando indicios del impacto del **clima** sobre el desempeño de las aves.

ANÁLISIS EXPLORATORIO DE DATOS

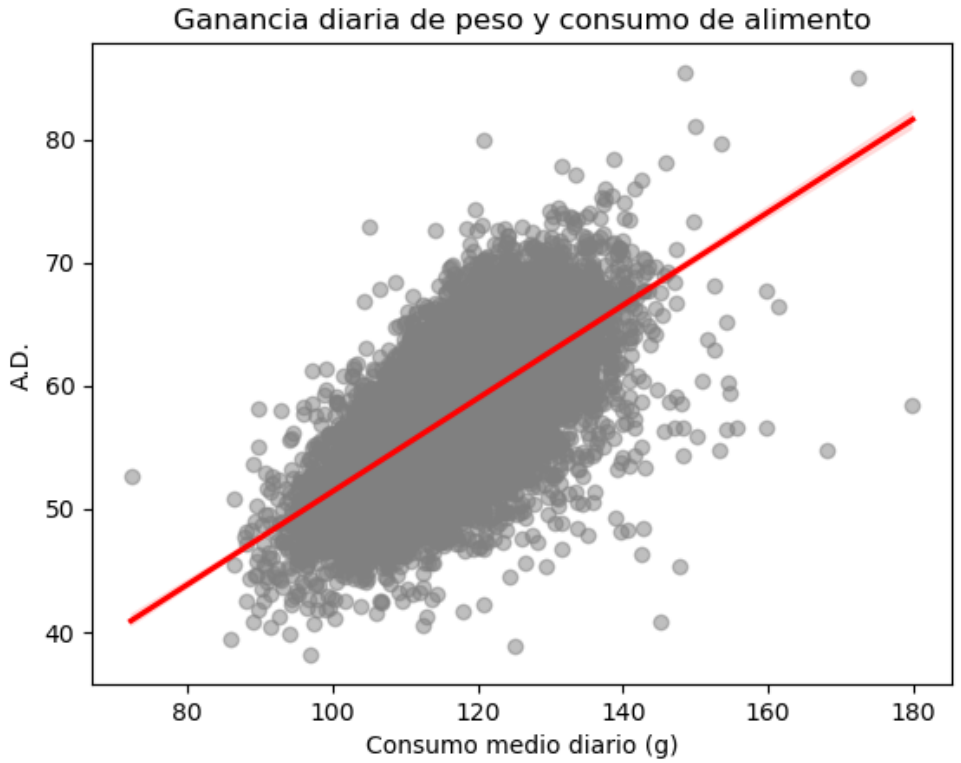






La **ganancia media diaria de peso** promedio también varía según el sistema de ventilación de los **galpones** y la **zona geográfica** de producción.

ANÁLISIS EXPLORATORIO DE DATOS



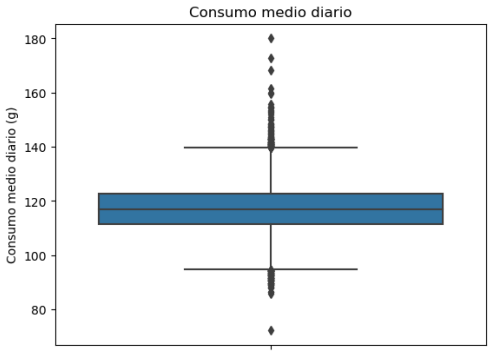
Se observa una alta correlación entre la ganancia media diaria de peso y el consumo de alimento. Debido a ello se utilizará la variable **Consumo medio diario de alimento (g)** como variable dependiente en los modelos de machine learning de **regresión** que se utilizarán. A continuación, se confeccionarán también modelos de **clasificación** y **agrupación** para relacionar la variable **ganancia de peso** con otros indicadores productivos y con las características de las granjas, su ubicación y el efecto del clima.

PREPARACIÓN DE LOS DATOS

Se confeccionaron diferentes subset de la base original para los distintos tipos modelos de Machine Learning utilizados.

Modelos de regresión

- ✓ Se eliminaron valores outliers
- ✓ OneHotEncoder para variables categóricas



Modelos de agrupación

- ✓ Se crearon subset diferentes para cada zona de producción

Modelos de clasificación

- ✓ Creación de variable “Ganancia”
- ✓ Se eliminaron valores nulos
- ✓ LabelEncoder para variables categóricas
- ✓ Oversampling para balancear cantidad de muestras

- ✓ Se renombraron valores en columnas de la base original

```
#Se renombran valores de diferentes columnas para evitar duplicados al codificar las variables categóricas
df.loc[df['Ambiente controlado'] == 'N', 'Ambiente controlado'] = 'no'
df.loc[df['Ambiente controlado'] == 'S', 'Ambiente controlado'] = 'si'
df.loc[df['zona climática'] == 'Arrecifes', 'zona climática'] = 'Arrecifes_BA'
df.loc[df['zona climática'] == 'Crespo', 'zona climática'] = 'Crespo_ER'
```

Regresión

- ✓ LinearRegression
- ✓ KNN
- ✓ RandomForestRegressor
- ✓ XGBoost
- ✓ SVR

Clasificación

- ✓ RandomForestClassifier 4n
- ✓ RandomForestClassifier 7n
- ✓ BRFC
- ✓ KNN Classifier

Agrupación

- ✓ BSCAN
- ✓ KMeans

MACHINE LEARNING

Modelos de regresión

✓ Métricas de los modelos para los datos de prueba

	Linear	KNN	RF	XGB
MSE	53.297099	52.852298	6.852183	29.884333
RMSE	7.300486	7.269959	2.617668	5.466656
MAE	5.823014	5.825453	2.063389	4.337493
R2	0.211202	0.217785	0.896711	0.550806

✓ Métricas de los modelos para los datos de entrenamiento

	Linear	KNN	RF	XGB
MSE	52.980669	53.953175	6.978970	30.350884
RMSE	7.278782	7.345282	2.641774	5.509164
MAE	5.768913	5.852635	2.025253	4.291011
R2	0.224899	0.210671	0.899753	0.562796

Comparando las métricas MSE, RMSE, MAE y R2 de los 4 modelos los mejores valores son para RandomForest en primer lugar y XGB en segundo lugar.

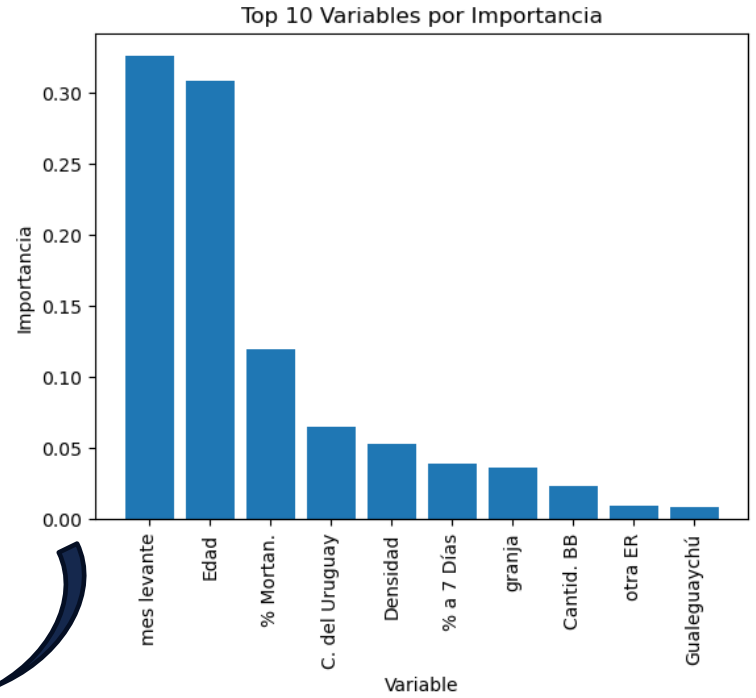
MACHINE LEARNING

Stratified-K fold para Random Forest Regressor

Scores cross-validation

```
array([-5.90380813, -6.23930999, -5.9131547 , -5.92677953, -5.92431458,
       -6.0307267 , -6.06706368])
```

- ✓ Se observó que el modelo es estable ya que al variar la conformación del grupo de datos de entrenamiento los valores de negative MAE obtenidos son similares.



Las variables de mayor importancia en la regresión fueron el mes de levante, la edad y la mortalidad de las aves

MEJORA DE LOS MODELOS

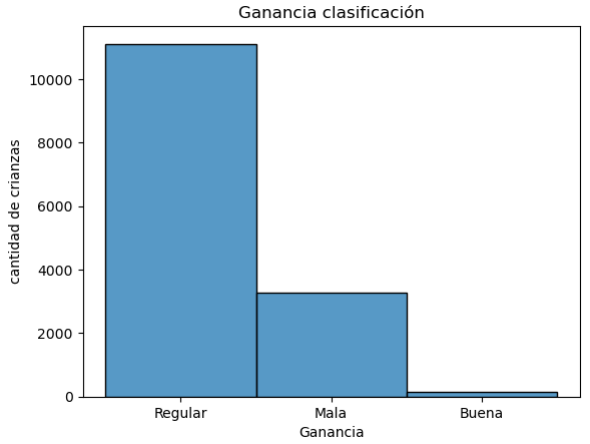
Modelos de clasificación

✓ RandomForestClassifier 4n

	precision	recall	f1-score	support
0	0.82	0.81	0.81	2785
1	0.92	0.82	0.86	2835
2	0.77	0.87	0.82	2782
accuracy			0.83	8402
macro avg	0.84	0.83	0.83	8402
weighted avg	0.84	0.83	0.83	8402

✓ BRFC

	precision	recall	f1-score	support
0	0.62	0.57	0.59	2785
1	0.68	0.63	0.65	2835
2	0.68	0.79	0.73	2782
accuracy			0.66	8402
macro avg	0.66	0.66	0.66	8402
weighted avg	0.66	0.66	0.66	8402



✓ RandomForestClassifier 7n

	precision	recall	f1-score	support
0	0.96	0.96	0.96	2785
1	0.98	0.95	0.96	2835
2	0.92	0.95	0.94	2782
accuracy			0.95	8402
macro avg	0.95	0.95	0.95	8402
weighted avg	0.95	0.95	0.95	8402

✓ KNN Classifier

	precision	recall	f1-score	support
0	0.99	1.00	0.99	2785
1	0.99	1.00	0.99	2835
2	1.00	0.97	0.99	2782
accuracy			0.99	8402
macro avg	0.99	0.99	0.99	8402
weighted avg	0.99	0.99	0.99	8402

Mejores valores de F1 y *accuracy*

MACHINE LEARNING

✓ KNN Classifier

- Base original sin oversampling: Accuracy 76,5 %
- Base original agrupada con PCA: Acurracy 72,3 %

GridSearchCV

```
Mejores hiperparámetros: {'n_neighbors': 20, 'p': 1, 'weights': 'distance'}
Precisión del modelo: 0.7816513761467889
```

Boosting

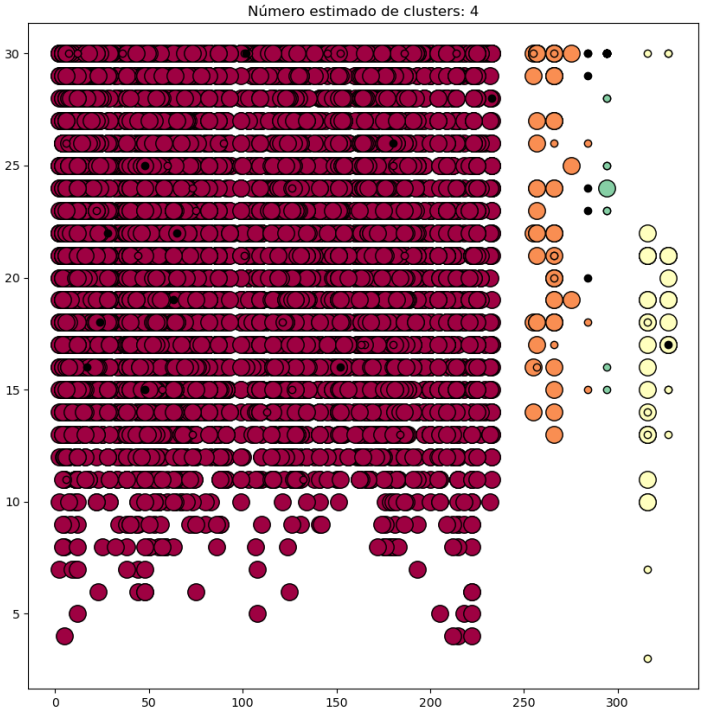
✓ XGBoost Classifier

	precision	recall	f1-score	support
0	0.50	0.06	0.10	54
1	0.58	0.37	0.45	967
2	0.82	0.92	0.87	3339
accuracy			0.79	4360
macro avg	0.64	0.45	0.47	4360
weighted avg	0.77	0.79	0.77	4360

MEJORA DE LOS MODELOS

Modelos de agrupación

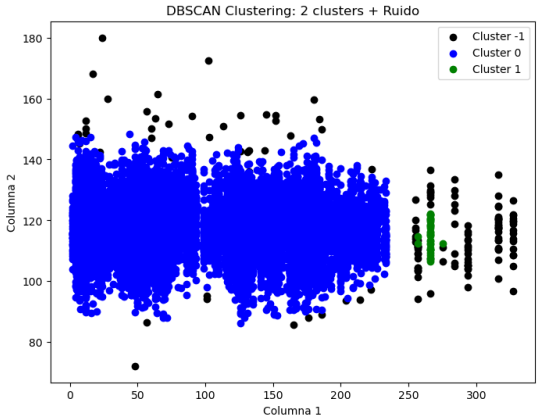
✓ BSCAN



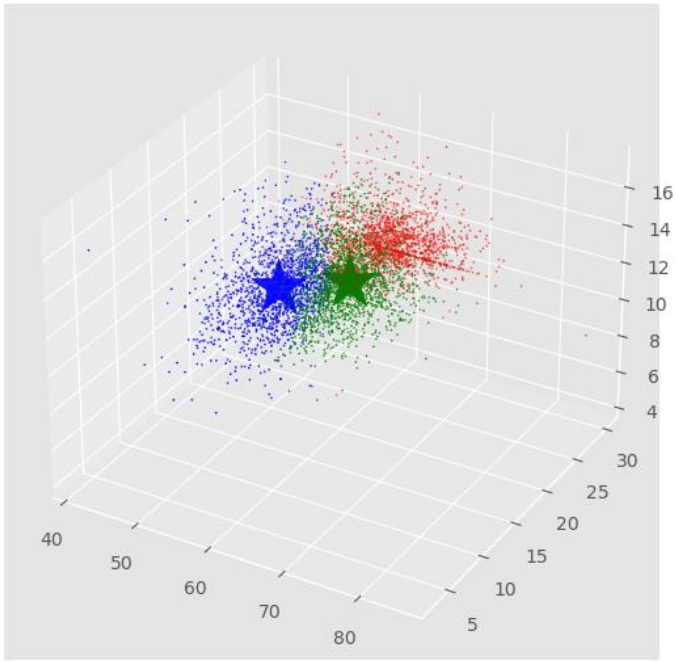
✓ Mejora hiperparámetros

```
best_dbscan

DBSCAN(eps=10, min_samples=20)
```



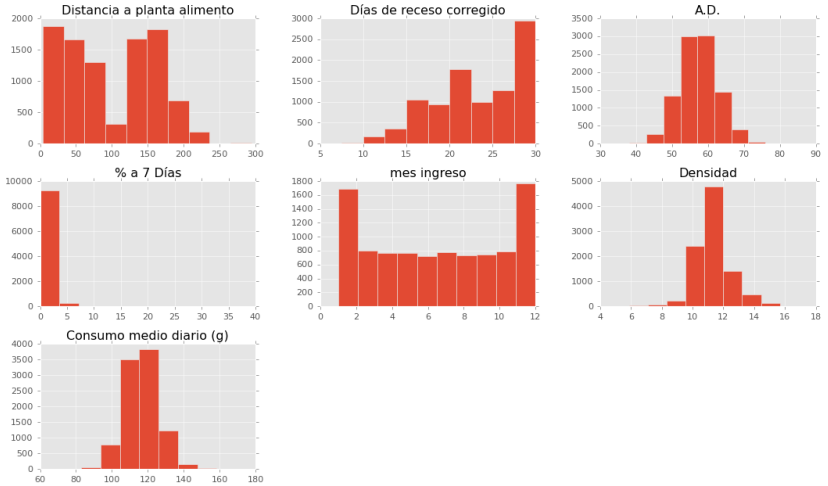
✓ KMeans



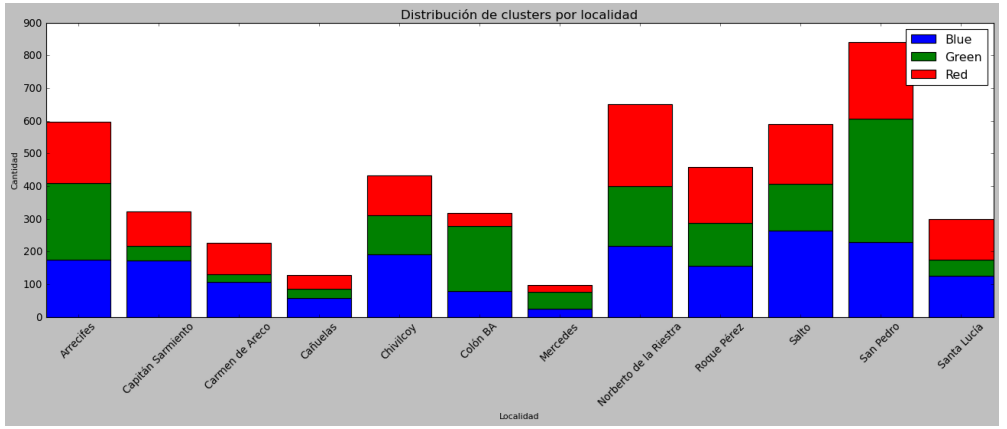
Se analizaron localidades con KNN ya que los resultados de BSCAN no resultaron adecuados para la base de datos

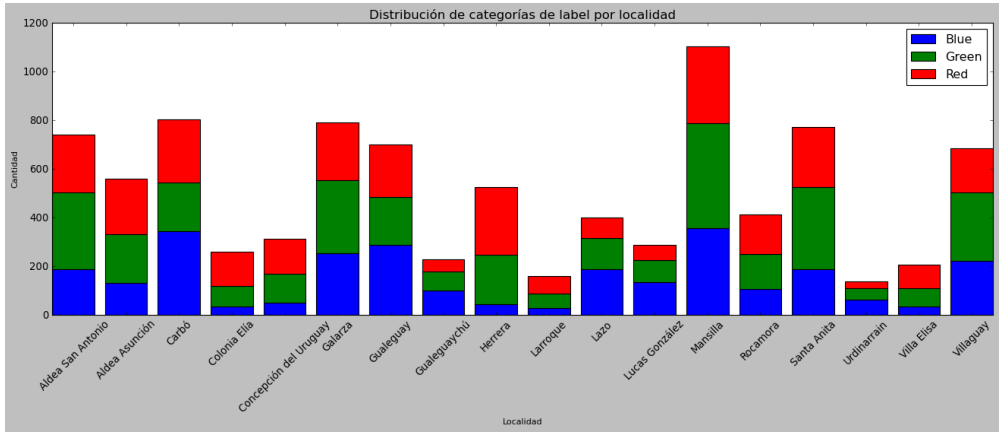
MACHINE LEARNING

- ✓ KMeans



Se crearon 3 clústers con las variables: Días de receso corregido, A.D. (ganancia media diaria de peso) y Densidad. Los 3 están representados en todas las localidades de ambas zonas de producción





MACHINE LEARNING

CONCLUSIONES Y CONSIDERACIONES FINALES

- Del análisis del presente trabajo surge que el **crecimiento de los pollos parrilleros depende de múltiples factores y no de una única variable**. El mes de ingreso de las aves a la granja y la estación del año en que se desarrolla la crianza indican que hay un impacto del **clima** sobre el desempeño.
- El **sistema de ventilación de los galpones** de tipo ventilación forzada y/o ambiente controlado permite obtener, en promedio, mayores valores de ganancia media diaria de peso.
- La **zona de producción** Crespo tiene mejores resultados productivos pero la cantidad de datos de crianza es mucho menor que el resto de las zonas y, como se observó también, los resultados son diferentes comparando los distintos años productivos, por lo que no es posible concluir que esta zona es mejor por falta de datos.
- El **consumo de alimento** fue la variable que tuvo mayor correlación con la ganancia media diaria de peso, debido a ello se utilizaron modelos de regresión para predecir este indicador, el **mes de levante** fue la variable de mayor importancia.
- Se utilizaron modelos de clasificación creando la variable ganancia que divide en tres partes iguales los valores de ganancia media diaria de peso: buena, regular y mala. Dadas las métricas obtenidas por los diferentes modelos, es posible asumir que las variables relacionadas con **la ubicación geográfica, los días de receso, la densidad, la mortalidad de primera semana y la época del año** son determinantes para obtener una mejor o peor ganancia diaria de peso.
- Finalmente evaluando las localidades geográficas por zona de producción utilizando modelos de agrupación se observó que **no hay localidades que tengan resultados de ganancia de peso uniformes** (siempre buenos, regulares o malos) si no que todas las posibilidades de resultados están presentes en todas ellas.

MUCHAS GRACIAS

EMILSE BOVER

EMILSEBOVER@GMAIL.COM