

PROYECTO FINAL

CARRERA DATA SCIENTIST – CODERHOUSE – PIO CACERES

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

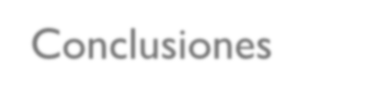
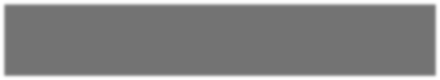
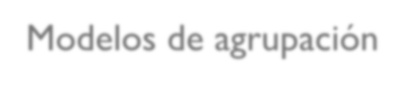
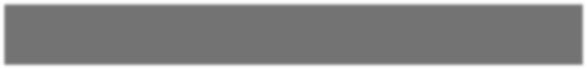
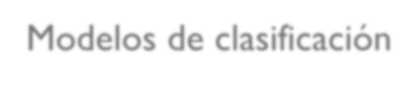
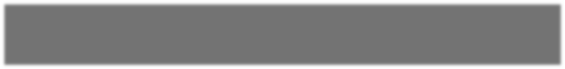
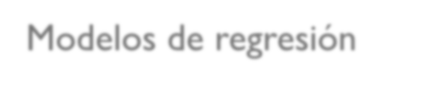
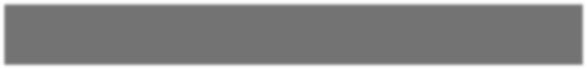
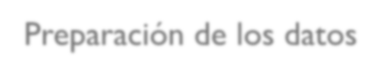
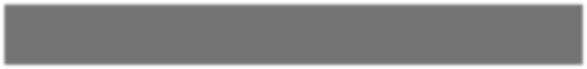
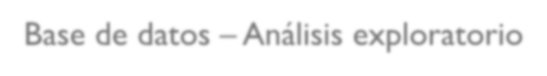
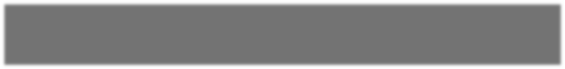
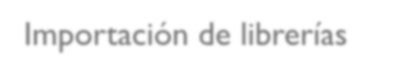
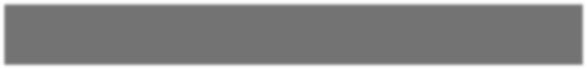
Importación de librerías

Base de datos – Análisis exploratorio

Preparación de los datos

Modelos de regresión

Modelos de clasificación Modelos de agrupación



Conclusiones

ANÁLISIS EXPLORATORIO DE DATOS

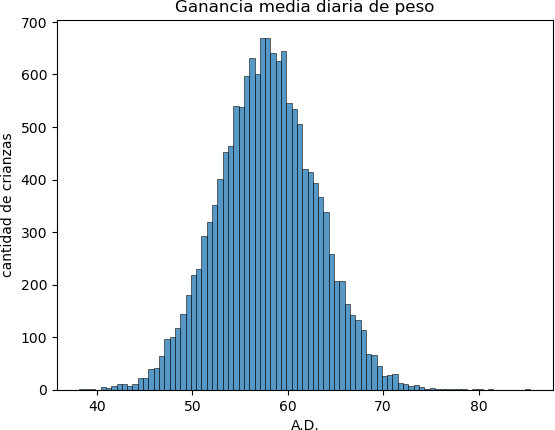


#### La base de datos posee 14.844 filas, cada una de las cuáles representa una crianza de una granja. La base tiene

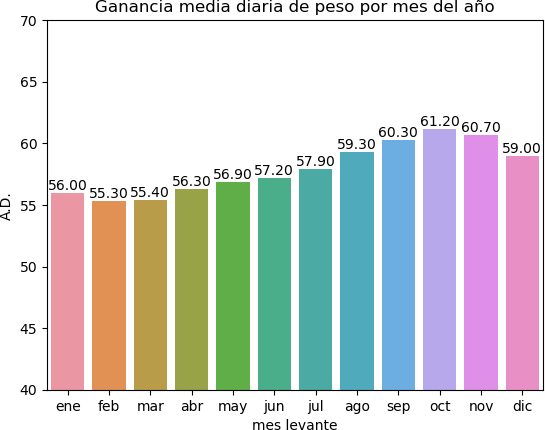
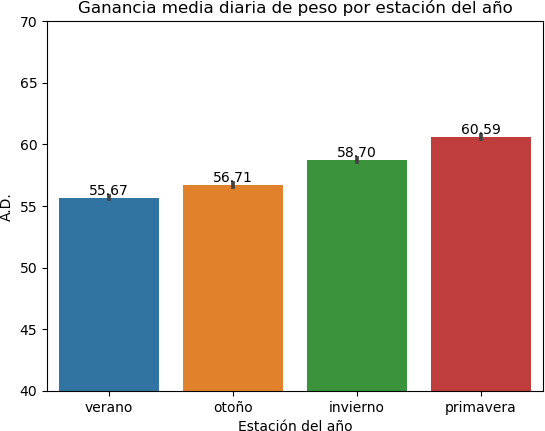
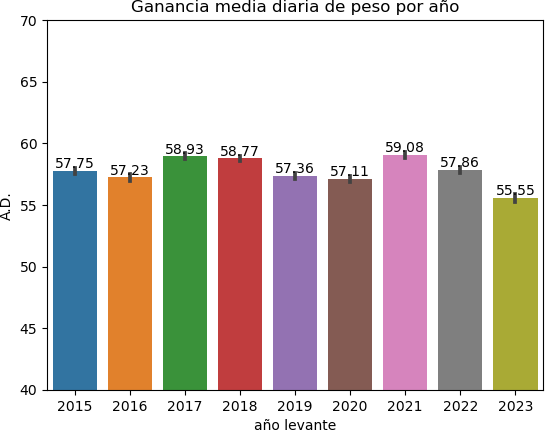
61 columnas con información sobre las características de la granja, su ubicación y los resultados de producción

#### obtenidos. Posee valores desde el año 2015 hasta el primer semestre del año 2023.

ANÁLISIS EXPLORATORIO DE DATOS

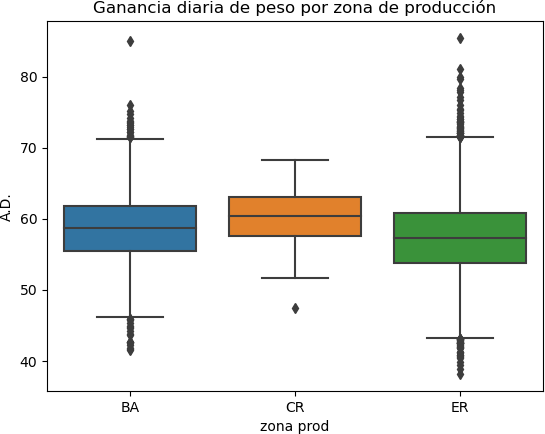
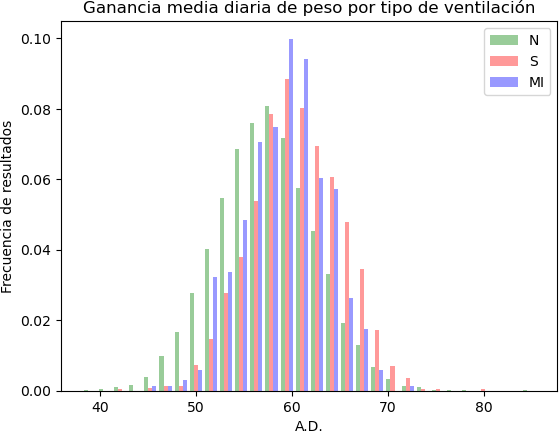
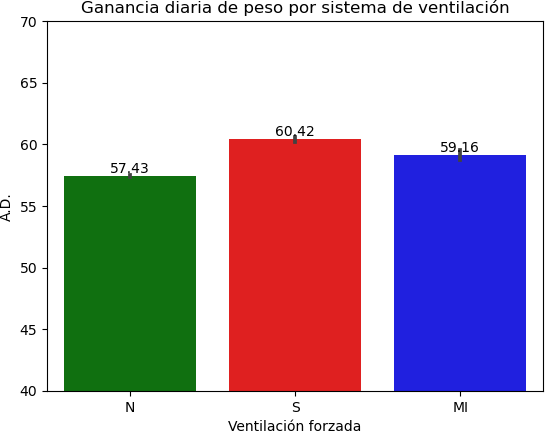
La **ganancia media diaria de peso** es una medida que permite comparar resultados entre granjas y además permite ver la evolución de los resultados globales de la empresa a lo largo del tiempo. A mayor valor, mejor será el resultado. El valor medio es de **57.84 gramos** de peso ganados promedio por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

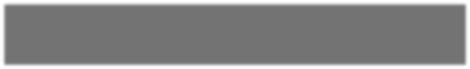
**(g)** como variable dependiente en los modelos de machine learning de **regresión** que se utilizarán. A continuación, se confeccionarán también modelos de **clasificación** y **agrupación** para relacionar la variable **ganancia de peso** con otros indicadores productivos y con las características 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

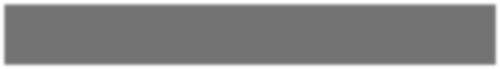
Learning utilizados.

* Se eliminaron valores outliers



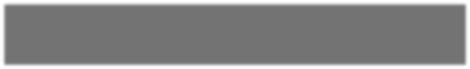
Modelos de regresión

* OneHotEncoder para variables categóricas

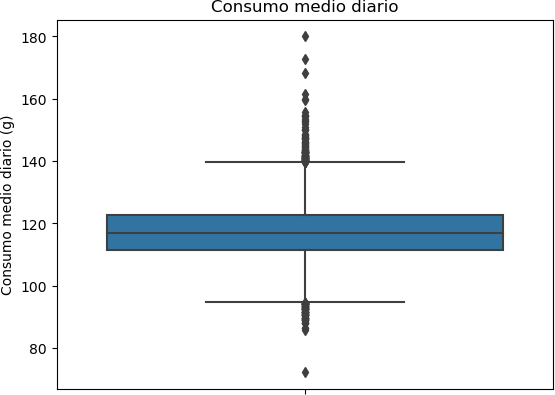


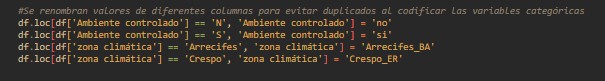
Modelos de clasificación

* Creación de variable “Ganancia”
* Se eliminaron valores nulos
* LabelEncoder para variables categóricas
* Oversampling para balancear cantidad de muestras
  + Se crearon subset diferentes para cada zona de producción



Modelos de agrupación

* Se renombraron valores en columnas de la base original



Regresión

Clasificación

Agrupación

* LinearRegression
* KNN
* RandomForestRegressor
* XGBoost
* SVR
* RandomForestClassifier 4n
* RandomForestClassifier 7n
* BRFC
* KNN Classifier
* BSCAN
* KMeans

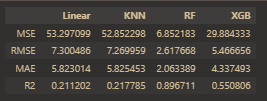
MACHINE LEARNING

**Modelos de regresión**

* + Métricas de los modelos para los

datos de prueba

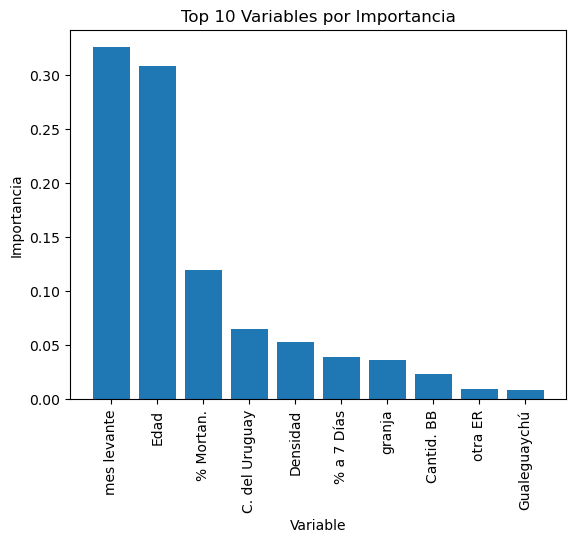
* + Métricas de los modelos para los datos de entrenamiento



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

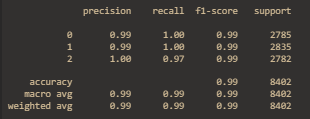
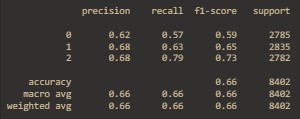
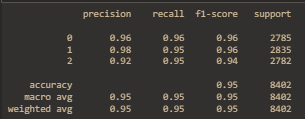
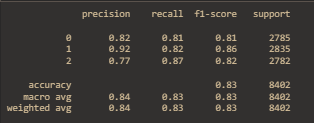
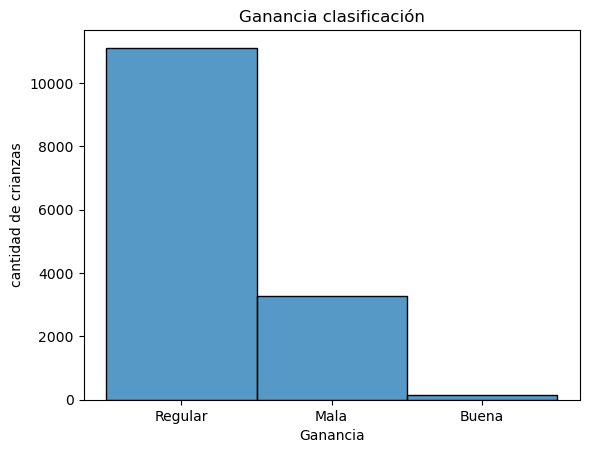
### Se observó que el modelo es estable ya que al variar la conformación del grupo de datos de entrenamiento los valores de negative MAE obtenidos son similares.

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

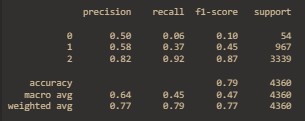
## Modelos de clasificación

* RandomForestClassifier 4n
  + RandomForestClassifier 7n
* BRFC
  + KNN Classifier



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

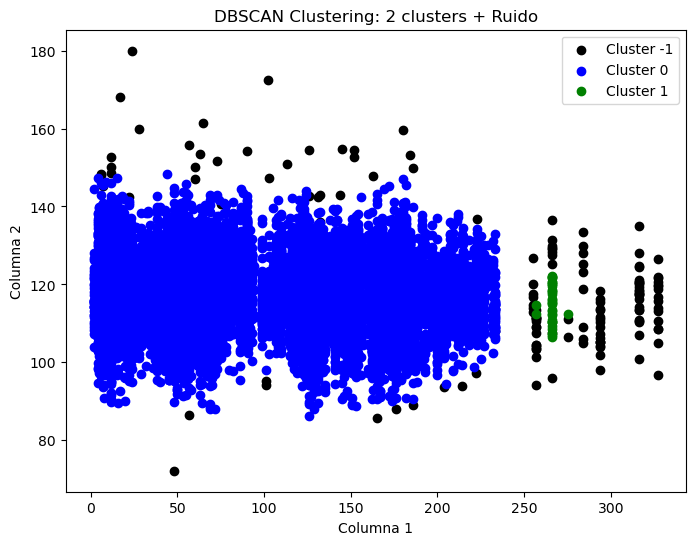
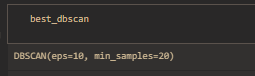
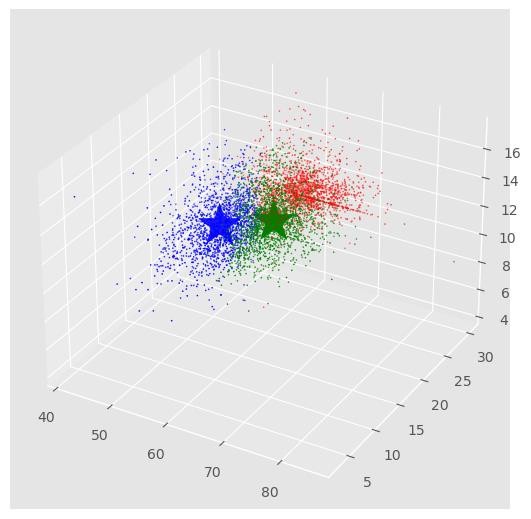
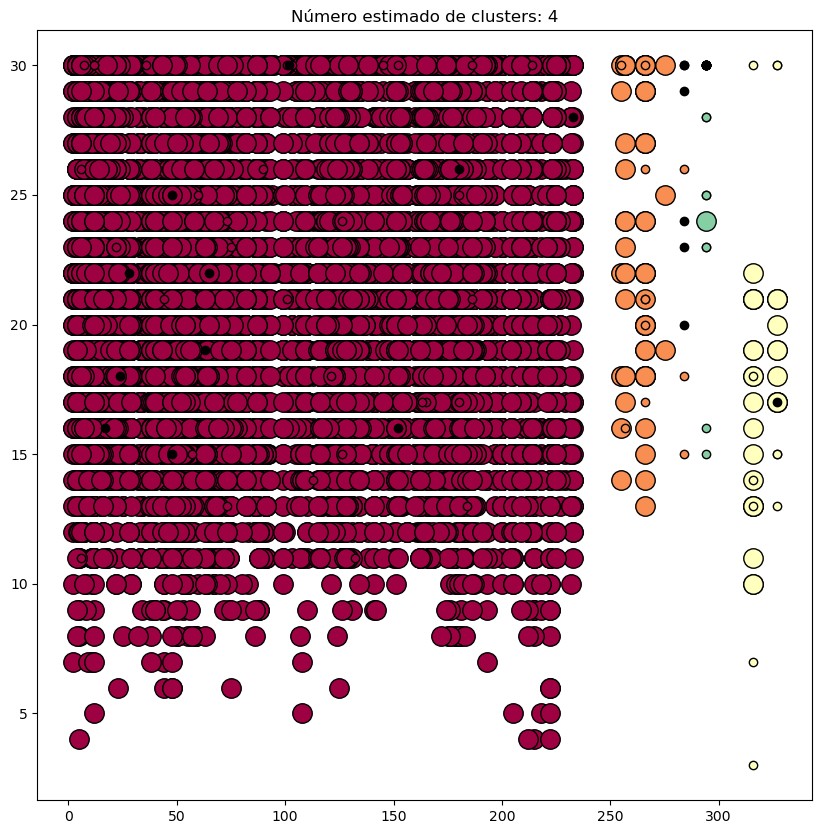
**Boosting**

* XGBoost Classifier

MEJORA DE LOS MODELOS

## Modelos de agrupación

* + BSCAN
* Mejora hiperparámetros
* KMeans

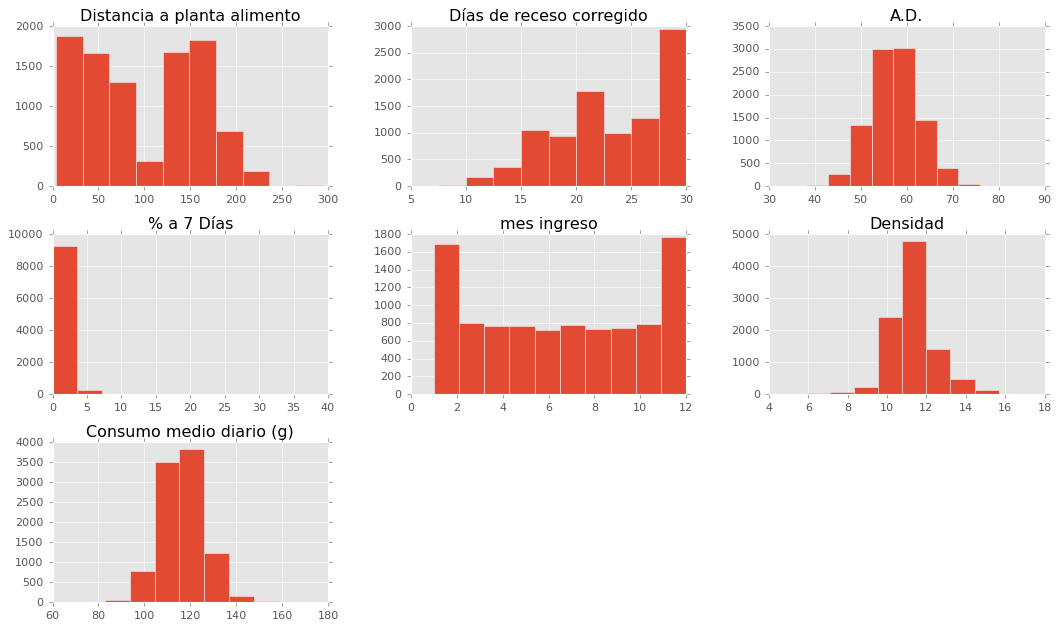
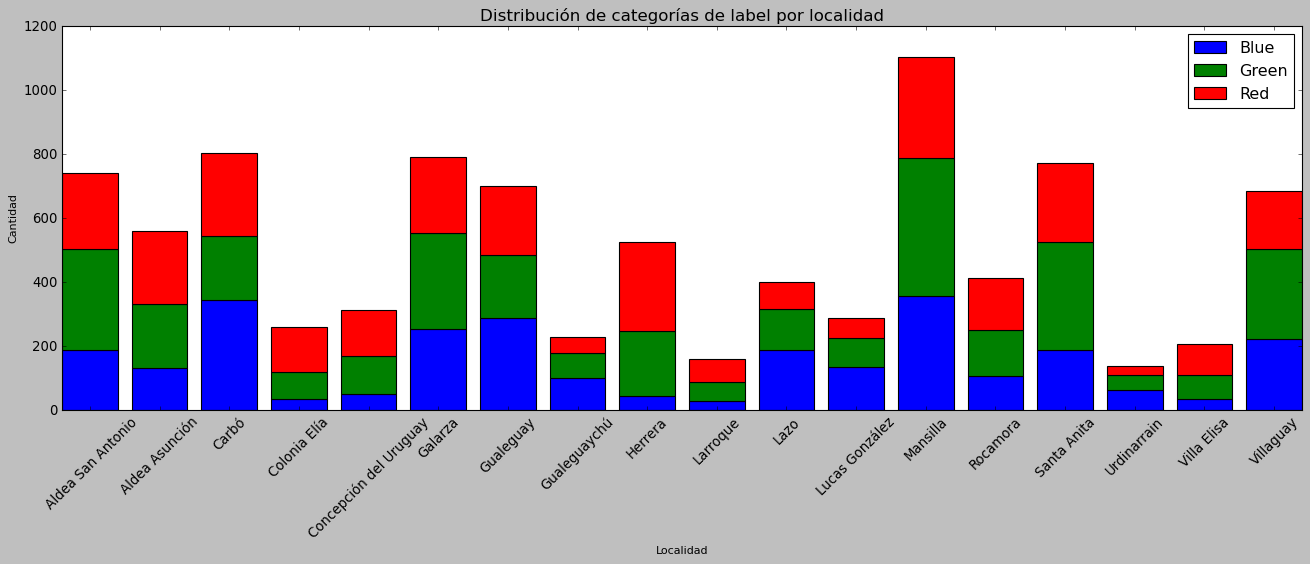
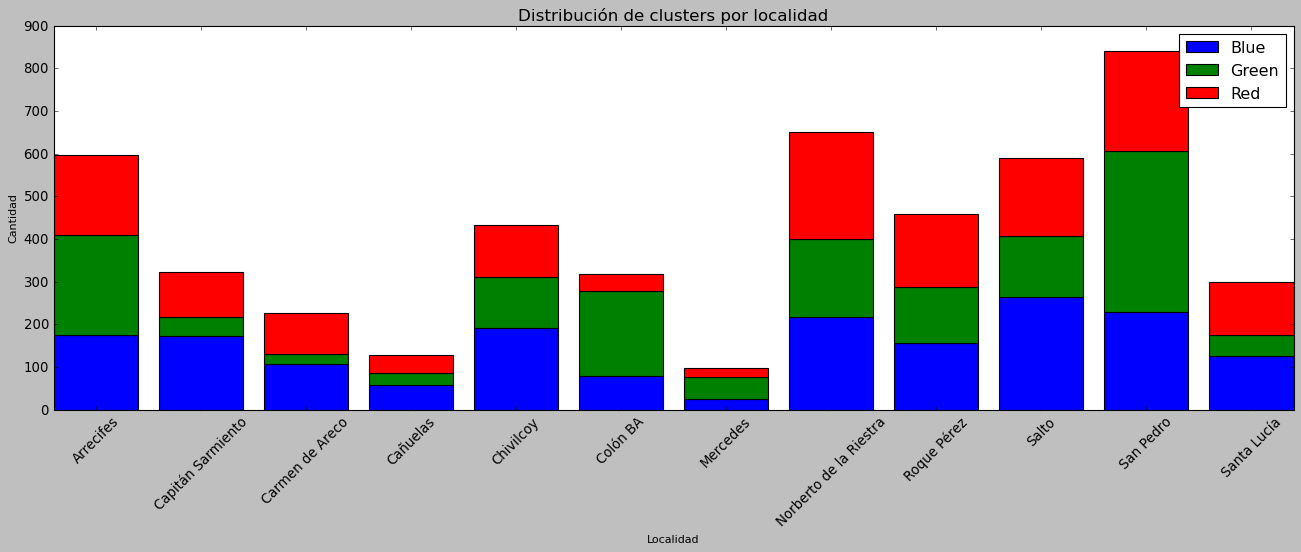


Se analizaron localidades con KNN ya que los resultados de BSCAN no

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

PIO CACERES