**Predicción del resultado económico de proyectos ganaderos de cría**

**Objetivos**

En el siguiente trabajo se realizó un modelo para predecir el resultado neto expresado en pesos por hectárea de la producción de carne bovina en argentina, más específicamente en la actividad de cría. Luego de varios intentos, y comparaciones se llegó a un modelo de regresión lineal con un grado de exactitud aceptable, alcanzando un valor de 0.299 de MAE.

**Introducción**

Históricamente, la actividad ganadera ha sido uno de los sectores más pujantes de la República Argentina. Actualmente, el país se encuentra entre los principales productores y exportadores a nivel mundial de carne bovina. En el 2017 se exportaron 681 toneladas de productos de origen ganadero, lo cual visualiza la importancia de la producción bovina para argentina.

**Dataset**

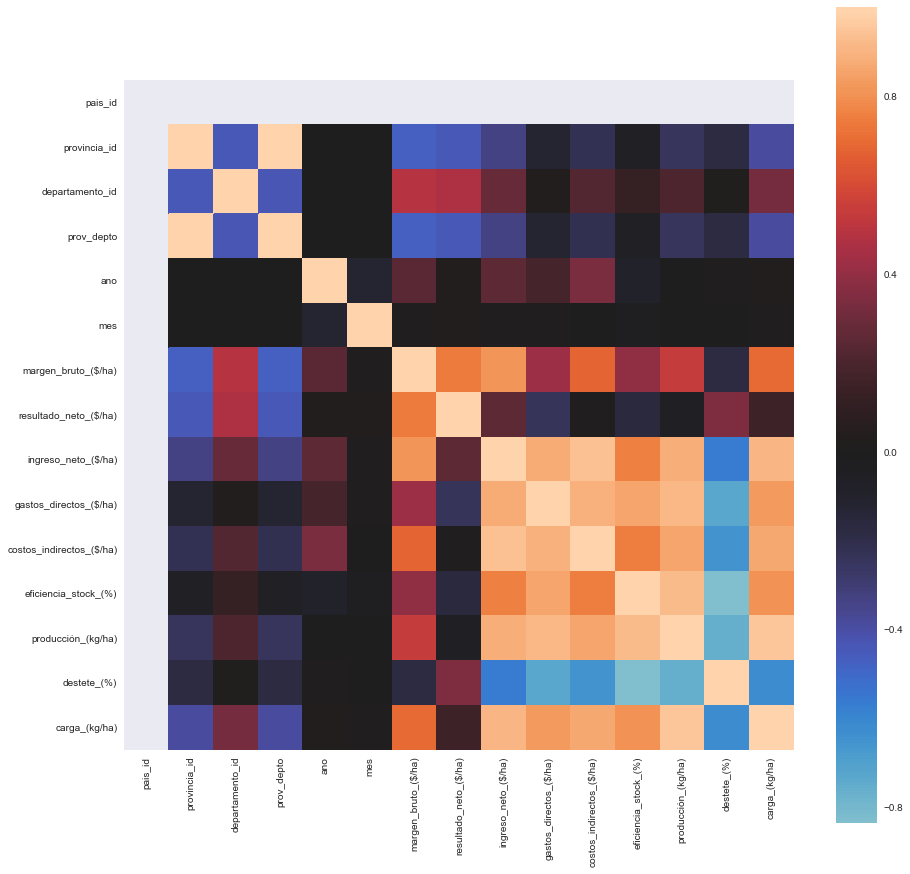
El dataset “producción de carne bovina” fue obtenido desde la página “datos.gob.ar” que es la página oficial de datos de la nación. El mismo, cuenta con datos de 13 provincias a lo largo de 4 años para las actividades de cría, invernada y ciclo completo con un total de 2598 registros y 21 features. Cada fila contiene los datos de un desarrollo ganadero con su respectiva localización, es decir, país, provincia, departamento; y además la fecha, la clasificación de la actividad y los distintos datos económicos como: “margen bruto ($/ha)” “resultado neto ($/ha)” “ingreso neto ($/ha)” “gastos directos ($/ha)” “costos indirectos\_($/ha)” “eficiencia stock (%)” “producción (kg/ha)”. En un principio se notó que había algunas columnas que no tenían información relevante, por lo que se decidió descartarlas luego. De todas formas, se realizó un análisis exploratorio de datos para poder definir mejor cuales serían las columnas que serían de utilidad.

**Análisis de los datos**

En primer lugar, se verifico la existencia de valores nulos en los registros y se comprobó que el dataset no contenía valores nulos. Se realizó una primera visualización en donde se observa mucha dispersión y no se detectan a simple vista patrones claros, mas allá de una cierta correlación entre “costos indirectos” “gastos directos” “ingreso neto”.



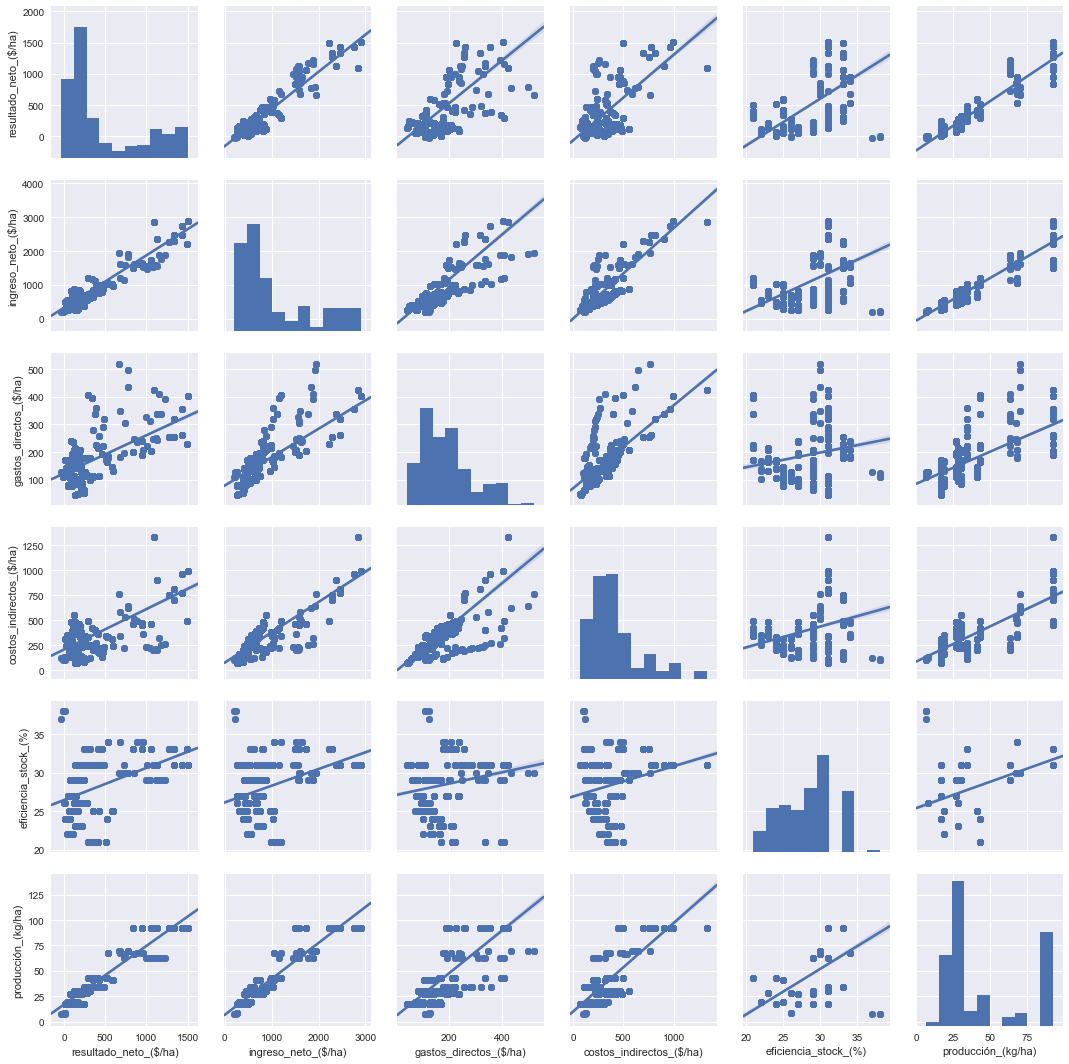
Se continuó realizando la matriz de correlación detectando tanto correlaciones positivas como negativas entre las features. De dicha matriz, se obtuvo como resultado más importante las columnas que no presentan correlación con las demás, tales como “año” y “mes”.



Luego, se agruparon los datos según la etiqueta “actividad” y se generó un nuevo dataframe solamente con los datos de la actividad ***cría*** y se volvió a visualizar, observándose una mejora sustancial en la dispersión de los datos. Se nota claramente una disminución del ruido y una distribución de los datos mucho más lineal. Siendo entonces, un conjunto de datos óptimo para ser trabajado con la herramienta de regresión lineal.



Seguido a esto, se generó un gráfico de correlación donde se observa buena correlación entre ciertas features, principalmente, entre “resultado neto”, nuestra variable a predecir, y las etiquetas “ingreso neto” y “producción”.



**Preprocesamiento**

El preprocesamiento comenzó estandarizando las variables, creando un nuevo dataframe con las 8 features mas importantes a nivel económico, las cuales se obtuvieron del análisis de correlación. Seguido a esto, se realizó un análisis de las componentes principales (PCA) donde la primera explica el 80% de la varianza y se visualizaron las mismas.



A continuación, se realizó una clusterizacion con Kmeans, una herramienta de clusterización que utiliza centroides minimizando la distancia euclidia entre los diversos puntos y dichos centroides, obteniendo dos grupos definidos en los cuales no se profundizo el análisis, pero si se determino qué registros se encuentran dentro de cada cluster. Siendo posible ahondar en los mismos a futuro.



Luego, fueron eliminadas del dataframe las columnas que no aportaban información relevante y se generaron los dummies para las features “Provincia”, “departamento”. Se identificó que no era posible generar los dummies correspondientes ya que existían dos con el mismo nombre, por lo que surgió la necesidad de reemplazar el departamento “Formosa” por “FormosaDep”. Estos dummies generados se anexaron al dataframe y con el mismo se generó la “x” y la “y” que es el resultado neto (pesos/Ha) que se desea predecir. Las variables fueron divididas en train y test con un tamaño de test de 40% y luego se aplicó un auto scalling para luego realizar una feature selection.

La feature selection fue realizada para reducir el numero de variables y utilizar solamente las más representativas. Se utilizó un *threshold* de 0.5. Del mismo se obtuvo como resultado 7 variables. Las cuales son “margen bruto” “gastos directos” “costos indirectos” “eficiencia stock” “producción” “destete” “carga”.

Con las variables seleccionadas se realizó una regresión lineal obteniendo un error cuadrático medio y absoluto bajo del orden de las 30 decimas

Para contrastar el modelo anterior se hizo una regresión KNN con la cual obtuvimos un error cuadrático medio con valores mucho mas elevados.

Como conclusión el modelo de regresión lineal es el que mejor predice el resultado económico. Se asume que esta mejora comparativa se da a partir de que las relaciones entre features son muy lineales, de modo que la regresión lineal se adapta mejor que la regresión KNN. El modelo podría ser utilizado en caso de tener datos históricos de un negocio para prever futuros resultados y adaptar en nivel de ciertos parámetros, como los gastos directos y costos indirectos, de modo de maximizar el resultado neto.