Proyecto Final – Data Science

Producción de Leche en Argentina

Base de dato - Año 2013 - 2023

CODERHOUSE

By José Buschiazzo

Introducción

Este proyecto tiene como objetivo proporcionar una visión integral de los patrones y tendencias en la producción láctea, con el fin de mejorar la calidad y eficiencia en la industria.

Motivación del Proyecto
Argentina, conocida por sus vastas llanuras y su rica tradición agrícola, se destaca en la producción láctea. En este contexto, la motivación detrás de este proyecto es esclarecer no solo las métricas numéricas sino también las historias detrás de los números. ¿Qué impulsa las fluctuaciones estacionales? ¿Cómo varían los niveles de grasa y proteína entre provincias? Estas son las preguntas que este análisis se propone responder.

Importancia Económica y Nutricional
La producción de leche no es simplemente una estadística; es el resultado tangible de la labor diaria de innumerables agricultores y ganaderos. La leche, una fuente primordial de nutrición, es fundamental para la dieta de millones. Además, su impacto económico, desde la generación de empleo hasta la exportación, subraya su importancia en el tejido socioeconómico de Argentina.

Alcance del Análisis
Este proyecto no se limita a simples números y tendencias. Explora la conformidad con normas de calidad, compara el rendimiento entre provincias y, finalmente, se sumerge en el mundo del modelado predictivo para desvelar pronósticos futuros. Es una exploración integral que abarca desde el pasado hasta el futuro de la producción láctea en Argentina.

Enfoque Metodológico
Desde el data wrangling meticuloso hasta la aplicación de modelos predictivos avanzados, este proyecto sigue una metodología sólida y equilibrada. Cada paso está diseñado para extraer conocimientos significativos y garantizar la robustez de las conclusiones.

Contenido

Estructura del Proyecto

1. Data Wrangling
1.1 Limpieza de Datos
1.2 Promedios y Sustitución

2. Análisis Exploratorio de Datos (EDA)
2.1 Tendencias Anuales
2.2 Comparación con Normas de Calidad

3. Modelado Predictivo
3.1 Selección de Algoritmo
3.2 Validación Cruzada y Despliegue

4. Recomendaciones.

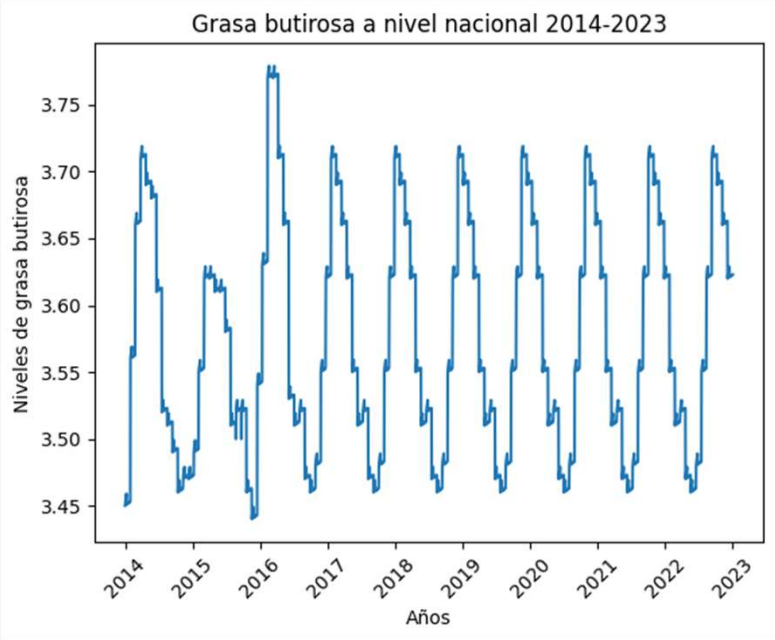
1. Data Wrangling

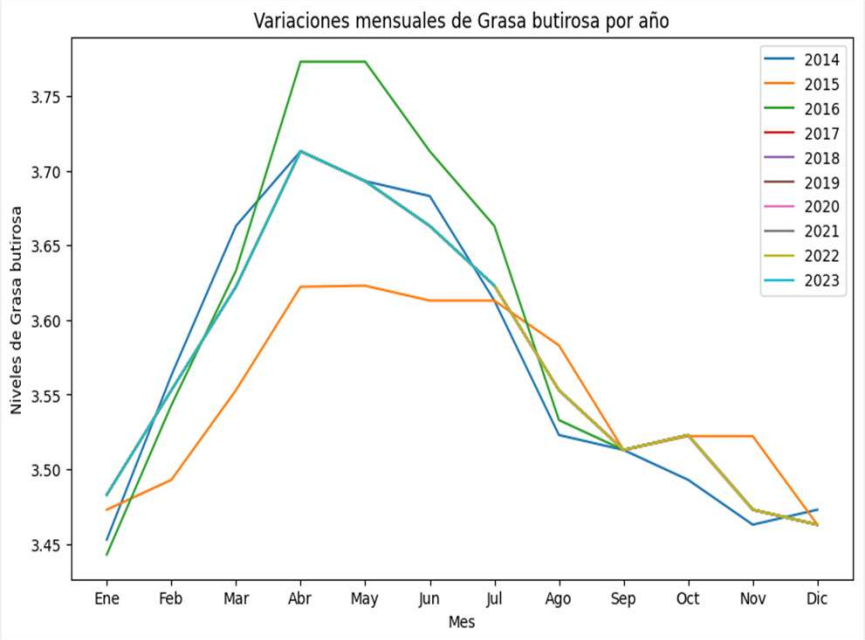
Orígenes del Análisis: El proyecto comenzó con una simple pregunta: ¿Dondé podemos tener una leche de mejor calidad? ¿Cumple con las normativas establecidas todo el año? Para lograr esto se tomó los datos oficiales de la web oficial de base de datos de Arg, un data set que proporcionó información valiosa.

Los Datos Clave: Utilizamos conjuntos de datos históricos que abarcan varias provincias y años, con un enfoque en los niveles de grasa y proteína.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 3511 entries, 1 to 3511
Data columns (total 18 columns):
 #   Column                                Non-Null Count  Dtype  
---  ------                                --------------  -----  
 0   indice_tiempo                         3511 non-null   object 
 1   anio                                  3511 non-null   int64  
 2   mes                                   3511 non-null   int64  
 3   dia                                   3511 non-null   int64  
 4   grasa_butirosa_nivel_nacional         3511 non-null   float64
 5   proteina_nivel_nacional               3511 non-null   float64
 6   grasa_butirosa_bs_as                  3499 non-null   float64
 7   grasa_butirosa_cordoba                3499 non-null   float64
 8   grasa_butirosa_entre_rios             3499 non-null   float64
 9   grasa_butirosa_la_pampa               3499 non-null   float64
 10  grasa_butirosa_santa_fe               3499 non-null   float64
 11  grasa_butirosa_santiago_del_estero    3499 non-null   float64
 12  proteina_bs_as                        3499 non-null   float64
 13  proteina_cordoba                      3499 non-null   float64
 14  proteina_entre_rios                   3499 non-null   float64
 15  proteina_la_pampa                     3499 non-null   float64
 16  proteina_santa_fe                     3499 non-null   float64
 17  proteina_santiago_del_estero          3499 non-null   float64
dtypes: float64(14), int64(3), object(1)
memory usage: 521.2+ KB
```

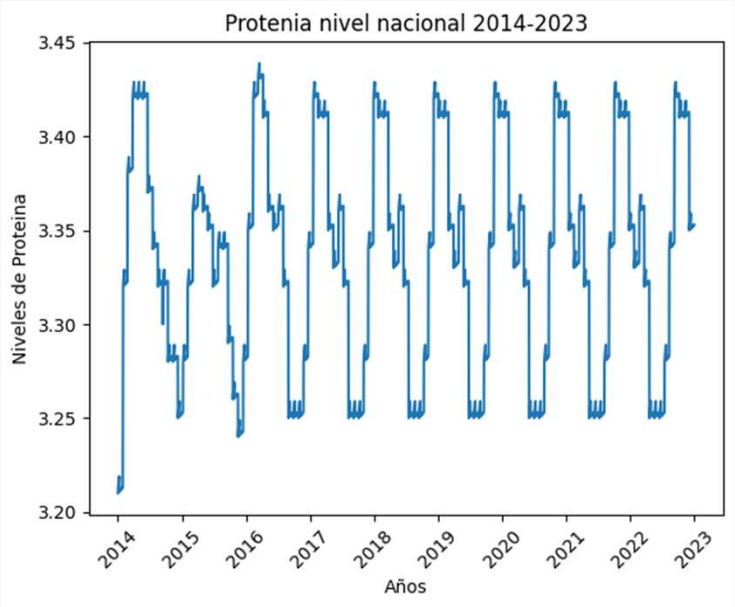
2.1 Niveles de grasa Nacional anual

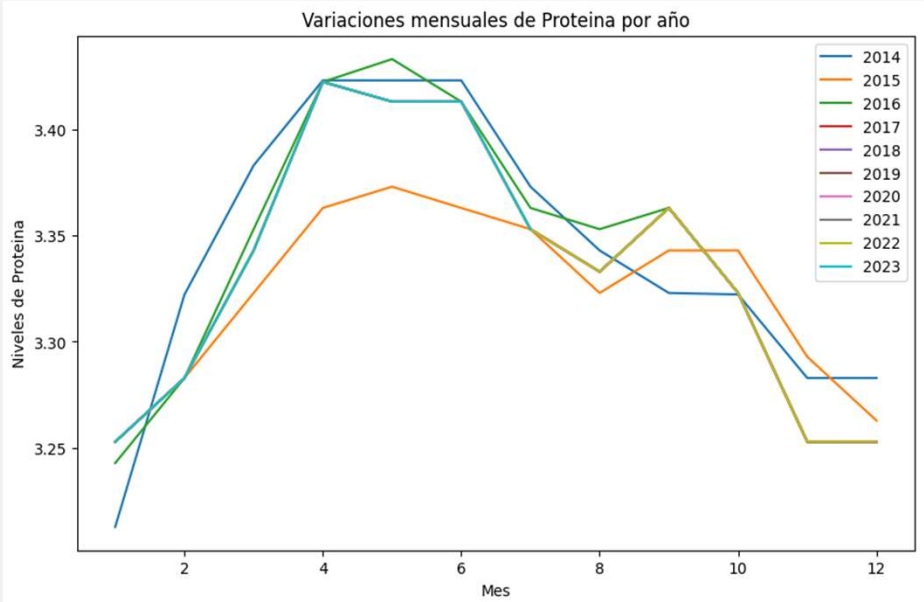




Este gráfico muestra cómo los niveles de grasa butirosa han variado a lo largo de los meses para diferente año. Concluyo que Existen Patrones Estacionales porque hay variación mensual y anual de grasa butirosa a nivel nacional.

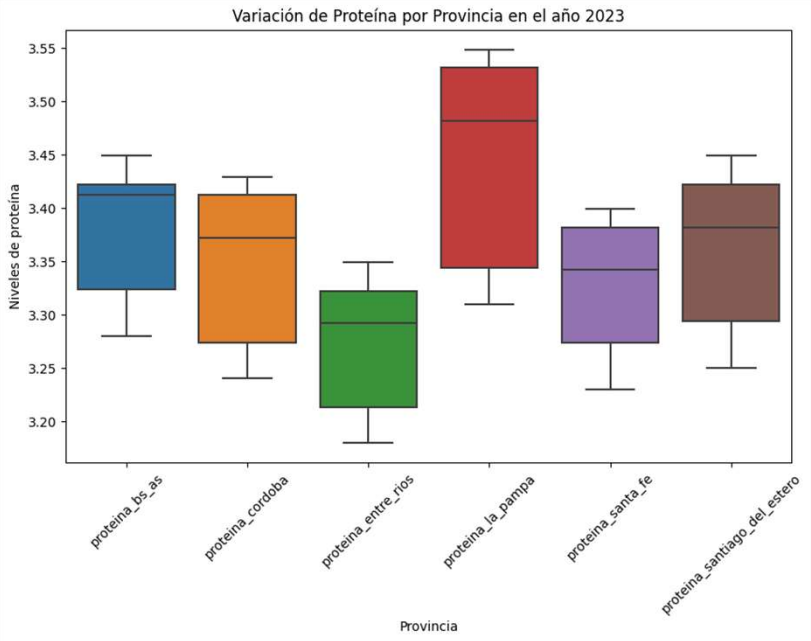
2.1 Niveles de Proteína Nacional anual

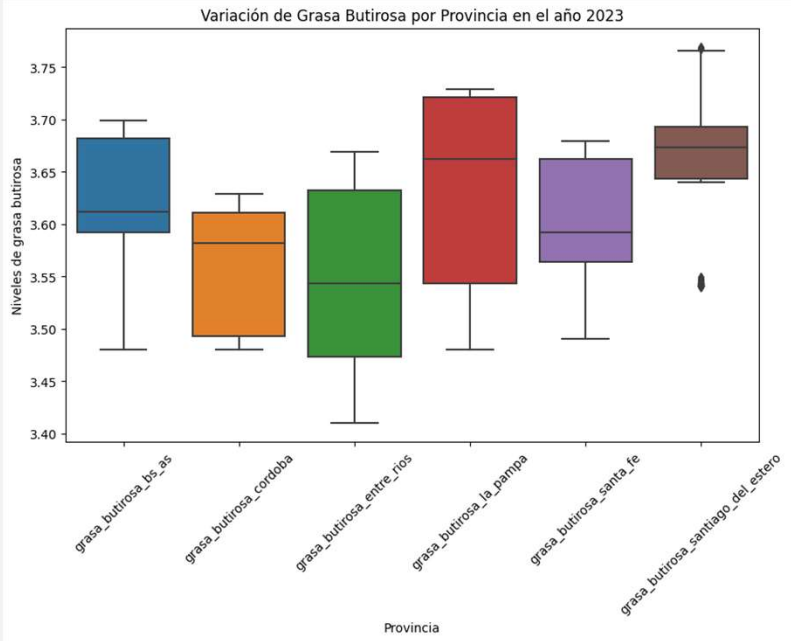




Este gráfico muestra cómo los niveles de proteína han variado a lo largo de los años para diferentes meses. Concluyo que también Existen Patrones de Estacionales por la variación mensual y anual de proteína a nivel nacional.

2.1 Variación de Proteína y grasa





Se evidencia que la Pampa en el 2023 se destaca. Si bien ya se ha visto que año a año no varía tanto, solo la variación es mes a mes, pero lo vamos a visualizar para verlo con respecto a Santiago del Estero que se ve un valor de grasa positivo

2.2 Comparación con Normas de Calidad

Comparación con Normas de Calidad:
Para evaluar la conformidad de los niveles de grasa y proteína en la leche con los estándares de calidad en Argentina, se han considerado las siguientes normas y reglamentos:

Código Alimentario Argentino (CAA): El CAA establece requisitos mínimos de calidad para la leche cruda. Según el CAA, la leche cruda debe tener un contenido mínimo de grasa del 3,5% y un contenido mínimo de proteína del 3,2%.

Reglamento Técnico Mercosur para leche y productos lácteos (Resolución GMC N° 56/97): Este reglamento aplica en el ámbito del Mercosur e establece requisitos mínimos de calidad para la leche cruda. De acuerdo con esta resolución, la leche cruda debe tener un contenido mínimo de grasa del 3,5% y un contenido mínimo de proteína del 3,2%.

Norma IRAM 171:2003 - Leche cruda: Esta norma nacional establece los requisitos mínimos de calidad para la leche cruda en Argentina. Según la norma IRAM 171:2003, la leche cruda debe tener un contenido mínimo de grasa del 3,5% y un contenido mínimo de proteína del 3,2%.

```
# Valores mínimos de grasa según el Reglamento Técnico Mercosur
valor_minimo_grasa = 3.5

provincias = [
    'bs_as',
    'cordoba',
    'entre_rios',
    'la_pampa',
    'santa_fe',
    'santiago_del_estero'
]

provincias_meses_no_cumplen_grasa = []
for provincia in provincias:
  nombre_columna = f'grasa_butirosa_{provincia}'
  valores_grasa = DS_filtrado[nombre_columna]
  meses_no_cumplen_grasa = DS_filtrado.loc[valores_grasa < valor_minimo_grasa, 'mes'].unique()
  if len(meses_no_cumplen_grasa) > 0:
        provincias_meses_no_cumplen_grasa.append((provincia, meses_no_cumplen_grasa))

# Imprimo los resultados
for provincia, meses_no_cumplen_grasa in provincias_meses_no_cumplen_grasa:
    print(f"En la provincia {provincia}, los valores de grasa no cumplen con los requisitos en los meses:", meses_no_cumplen_grasa)
```

```
En la provincia bs_as, los valores de grasa no cumplen con los requisitos en los meses: [1]
En la provincia cordoba, los valores de grasa no cumplen con los requisitos en los meses: [ 1  9 10 11 12  2  3]
En la provincia entre_rios, los valores de grasa no cumplen con los requisitos en los meses: [ 1  2  8  9 10 11 12  3  4  6]
En la provincia la_pampa, los valores de grasa no cumplen con los requisitos en los meses: [11 12  1]
En la provincia santa_fe, los valores de grasa no cumplen con los requisitos en los meses: [ 1 10 11 12]
En la provincia santiago_del_estero, los valores de grasa no cumplen con los requisitos en los meses: [ 1 11 12]
```

```
[ ]  # valores mínimos de proteína según el Reglamento Técnico Mercosur
     valor_minimo_proteina = 3.2
     provincias_meses_no_cumplen_proteina = []

     for provincia in provincias:

         nombre_columna = f'proteina_{provincia}'
         valores_proteina = DS_filtrado[nombre_columna]
         meses_no_cumplen_proteina = DS_filtrado.loc[valores_proteina < valor_minimo_proteina, 'mes'].unique()
         if len(meses_no_cumplen_proteina) > 0:
             provincias_meses_no_cumplen_proteina.append((provincia, meses_no_cumplen_proteina))

     # Imprime los resultados
     for provincia, meses_no_cumplen_proteina in provincias_meses_no_cumplen_proteina:
         print(f"En la provincia {provincia}, los valores de proteína no cumplen con los requisitos en los meses:", meses_no_cumplen_proteina)
```

```
En la provincia entre_rios, los valores de proteína no cumplen con los requisitos en los meses: [11  3]
En la provincia santa_fe, los valores de proteína no cumplen con los requisitos en los meses: [1]
En la provincia santiago_del_estero, los valores de proteína no cumplen con los requisitos en los meses: [1]
```

3. Modelado Predictivo

Resultado de los Modelados:

```
Error Absoluto Medio (MAE): 0.050602132961813 66
Error Cuadrático Medio (MSE): 0.0042654144457221 98
Raíz del Error Cuadrático Medio (RMSE): 0.065310140450945 27
Coeficiente de Determinación (R²): 0.542959815751253
```

```
Modelo de Regresión Polinómica:
Error Cuadrático Medio (MSE): 0.0022451285095118 95
Coeficiente de Determinación (R²): 0.7594339399589821
```

```
Modelo de Regresión Lasso:
Error Cuadrático Medio (MSE): 0.009357686374361 69
Coeficiente de Determinación (R²): -0.0026783467593671784
```

Conclusión:

El modelo de regresión polinómica parece ser el mejor de los tres, tiene el menor MSE y el mayor coeficiente de determinación (R^2). Esto significa que es el modelo que mejor se ajusta a tus datos y tiene la mejor capacidad predictiva.

El modelo de regresión Ridge tiene un MSE y R^2 intermedios. Ridge es útil cuando se sospecha que hay multicolinealidad en los datos o se quiere evitar la sobreajuste (overfitting). En este caso, Ridge parece ofrecer un buen equilibrio entre ajuste y capacidad predictiva.

El modelo de regresión Lasso tiene el peor rendimiento de los tres modelos. El MSE es significativamente mayor y el coeficiente de determinación R^2 es negativo, esto indica un rendimiento muy pobre. Lasso no es la mejor elección para este conjunto de datos.

Por otro lado, carece de sentido analizarlo con modelos no lineales, por la información obtenida anteriormente que brinda un indicio de naturaleza lineal de los datos.

4. Recomendaciones

Recomendaciones:

- Mejorar el monitoreo de niveles de grasa en enero.
- Enfocarse en prácticas para cumplir estándares de calidad.

Implicaciones para la Industria Lechera

- Impacto positivo en la toma de decisiones para productores y reguladores.
- Potencial para optimizar producción y calidad de la leche en Argentina.

Audiencia y Conclusión

6. Audiencia Objetivo

- Productores, reguladores y actores de la industria láctea.

7. Conclusión

En resumen, este análisis proporciona una perspectiva esclarecedora sobre la producción de leche en Argentina. Los hallazgos y recomendaciones ofrecen un camino hacia una producción más eficiente y de mayor calidad.

.

Conclusión Final

Conclusión del Análisis Exploratorio de Datos de Niveles de Grasa y Proteína en la Leche

En el transcurso de este análisis exploratorio de datos centrado en los patrones de composición de la leche a lo largo de un período temporal extenso, se ha llevado a cabo un análisis exhaustivo de los datos recopilados. Mediante la aplicación de técnicas de visualización, estadísticas descriptivas y análisis de tendencias, se han identificado varios aspectos clave que arrojan luz sobre la evolución de los componentes de la leche.

Se observó una clara tendencia ascendente en los niveles de proteína a lo largo de los años, lo que podría indicar mejoras en las prácticas de cría y alimentación del ganado. En contraste, los niveles de grasa parecen haber experimentado fluctuaciones estacionales, lo que sugiere posibles influencias ambientales o estacionales en la dieta y la producción de leche.

Como así también se observaron varias conclusiones clave a partir de este análisis:

Distribución de Niveles: Se visualizaron las distribuciones de los niveles de proteína en La Pampa y a nivel nacional. Se encontró que las distribuciones eran relativamente similares, con algunos valores atípicos presentes en la variable proteina_la_pampa.

Correlación y Relaciones: Se exploraron las correlaciones entre diferentes variables. La matriz de correlación reveló relaciones tanto positivas como negativas entre las variables de interés. Se identificaron algunas correlaciones más fuertes entre variables específicas.

Tendencias Anuales: Se analizaron las tendencias anuales de los niveles de grasa y proteína en Argentina. Se observó que los promedios anuales de grasa y proteína tendieron a fluctuar, y se pudo visualizar cómo estos niveles variaron a lo largo del tiempo.

Variaciones por Provincia: Se compararon los niveles de proteína en diferentes provincias. Se detectaron diferencias en las distribuciones y se identificaron valores atípicos en algunos casos.

Valores Atípicos: Se aplicaron técnicas de detección de valores atípicos para la variable proteina_la_pampa. No se encontraron valores atípicos según la definición basada en el rango intercuartil (IQR) en este caso específico.

En resumen, este análisis exploratorio de datos proporcionó una visión integral de los niveles de grasa y proteína en la leche en Argentina. Las visualizaciones, hipótesis y resúmenes numéricos revelaron patrones, correlaciones y diferencias provinciales en los datos. Si bien no se encontraron valores atípicos en la variable La Pampa para esta definición particular, se destacó la importancia de explorar diferentes aspectos de los datos para obtener una comprensión completa, detallada y precisa.

Este análisis servirá como base para futuros pasos analíticos, estudios y decisiones relacionadas con la producción y calidad de la leche y sus derivados en Argentina.