Practicas 5- 8

Inteligencia de Negocio

Gustavo Sobrado Aller

UO286277

71777616K



TABLA DE CONTENIDO

[1. PRACTICA 5 4](#_Toc185555955)

[1.1. PARTE OBLIGATORIA 4](#_Toc185555956)

[1.1.1. EJERCICIO 9 4](#_Toc185555957)

[1.1.2. EJERCICIO 10 5](#_Toc185555960)

[1.1.3. EJERCICIO 11 5](#_Toc185555964)

[1.1.4. EJERCICIO 12 6](#_Toc185555969)

[1.1.5. EJERCICIO 13 6](#_Toc185555975)

[1.1.5. EJERCICIO 14 6](#_Toc185555981)

[1.2. PARTE OPCIONAL 7](#_Toc185555984)

[1.2.1. EJERCICIO 15 7](#_Toc185555985)

[2. PRACTICA 6 8](#_Toc185555986)

[2.1. PARTE 1 8](#_Toc185555987)

[2.1.1. EJERCICIO 1 8](#_Toc185555988)

[2.1.2. EJERCICIO 2 9](#_Toc185555989)

[2.1.3. EJERCICIO 3 (OPCIONAL) 9](#_Toc185555995)

[2.2. PARTE 2 10](#_Toc185555996)

[2.2.1. EJERCICIO 1 10](#_Toc185555997)

[2.2.2. EJERCICIO 2 11](#_Toc185556003)

[2.2.3. EJERCICIO 3 12](#_Toc185556010)

[2.2.4. EJERCICIO 4 (OPCIONAL) 12](#_Toc185556018)

[2.2.5. EJERCICIO 5 (OPCIONAL) 13](#_Toc185556027)

[2.2.6. EJERCICIO 6 (OPCIONAL) 13](#_Toc185556037)

[3. PRACTICA 7 14](#_Toc185556038)

[3.1. PARTE OBLIGATORIA 14](#_Toc185556039)

[3.1.1. EJERCICIO 1 14](#_Toc185556040)

[3.1.2. EJERCICIO 2 15](#_Toc185556046)

[3.1.3. EJERCICIO 3 16](#_Toc185556053)

[3.1.4. EJERCICIO 4 17](#_Toc185556061)

[3.1.5. EJERCICIO 5 17](#_Toc185556070)

[3.1.6. EJERCICIO 6 (OPCIONAL) 18](#_Toc185556080)

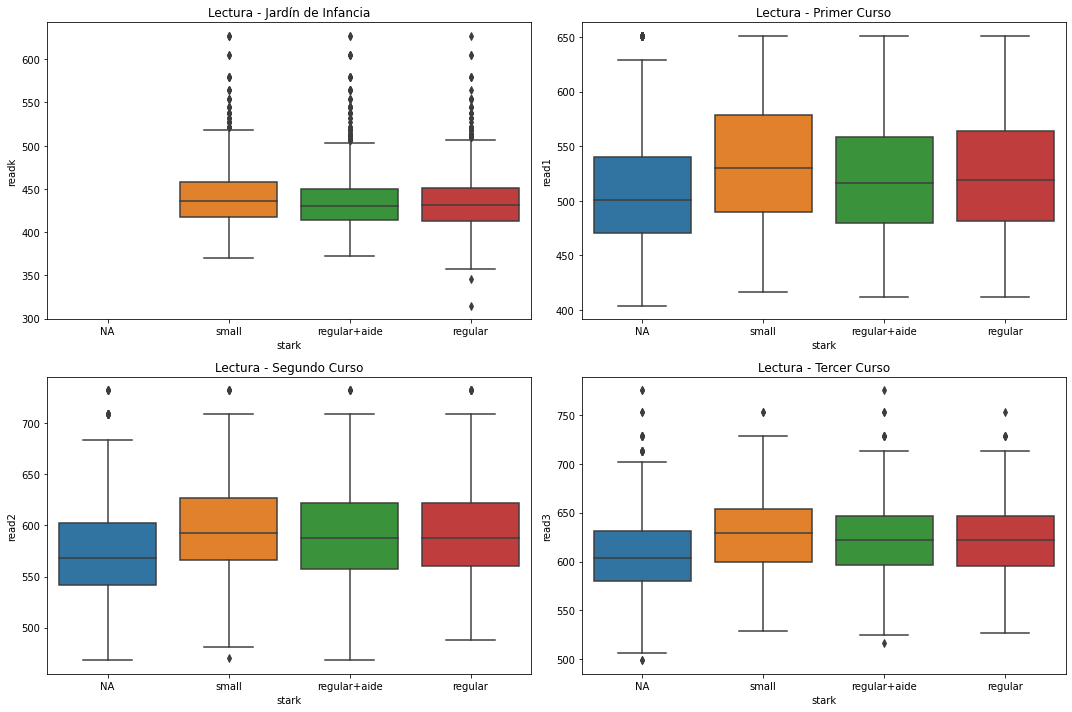
[4. PRACTICA 8 19](#_Toc185556081)

# PRACTICA 5

## PARTE OBLIGATORIA

### EJERCICIO 9

**Dibuja cuatro diagramas de cajas que relacionen el número de alumnos en el aula con los niveles de lectura para jardín de infancia, primer curso, segundo curso y tercer curso. A la vista de los diagramas, ¿crees que el tamaño de la clase influye en el nivel de lectura para todos los cursos?**



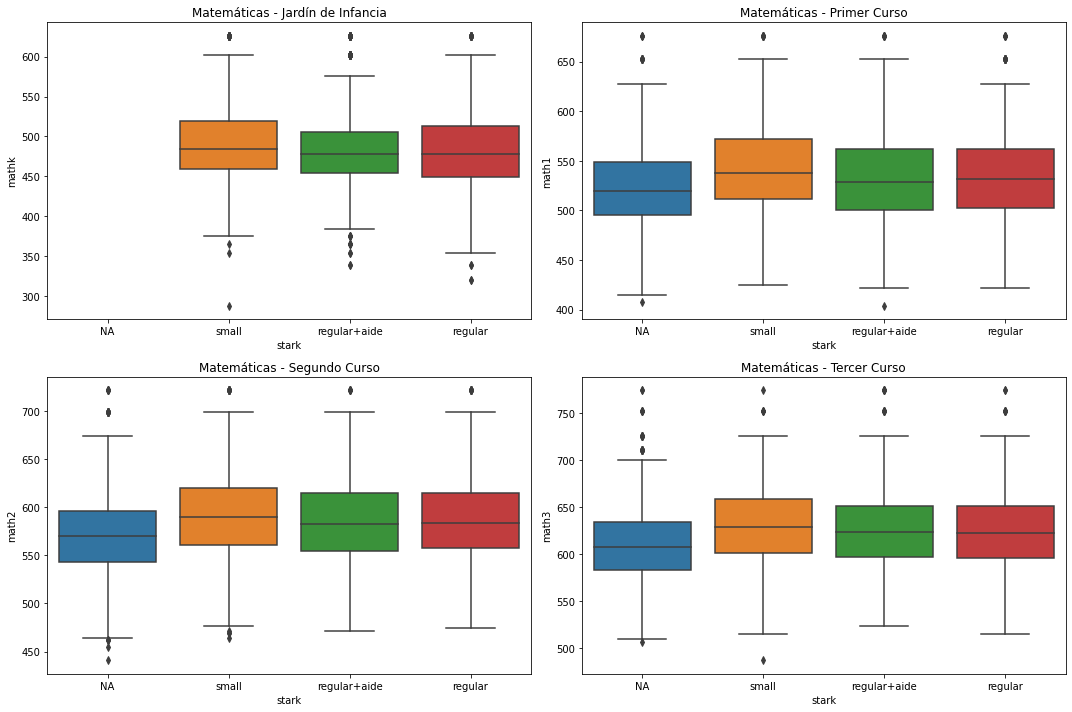
Basándome en los diagramas de caja, parece que hay algunas diferencias en los niveles de lectura según el tamaño de la clase. Sin embargo, no es consistente ni uniforme para todos los cursos.

Aunque hay una tendencia hacia mejores resultados en clases pequeñas, esta no se mantiene en todos los casos. Por ejemplo, los promedios en cursos posteriores ("read2" y "read3") no siempre reflejan una mejora significativa para clases más pequeñas.



### EJERCICIO 10

**Dibuja cuatro diagramas de cajas que relacionen el número de alumnos en el aula con los niveles de matemáticas para jardín de infancia, primer curso, segundo curso y tercer curso. A la vista de los diagramas, ¿crees que el tamaño de la clase influye en el nivel de matemáticas para todos los cursos?**



A partir de los datos analizados, parece que el tamaño de la clase puede influir ligeramente en el nivel de matemáticas, pero no de manera consistente en todos los cursos. El tamaño de la clase podría influir marginalmente, pero no es el único factor que determina los resultados en matemáticas.



### EJERCICIO 11

**Haz un modelo de regresión lineal que relacione el nivel de lectura en el jardín de infancia con las variables de la lista siguiente.**

Todo el código que he utilizado está en el archivo 2024-10-25 - DBFS Example.ipynb. Aquí solo se presentarán los resultados obtenidos.





### EJERCICIO 12

**Repite el modelo para primer curso, segundo y tercero, usando las variables necesarias para ello.**









### EJERCICIO 13

**Haz cuatro modelos lineales que relacionen el nivel de matemáticas con las variables correspondientes en los cuatro cursos.**





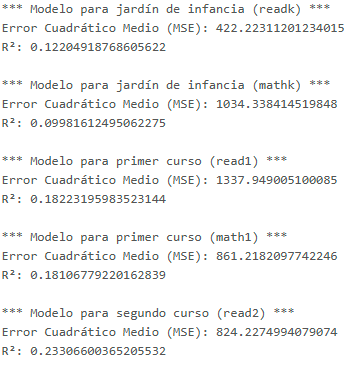
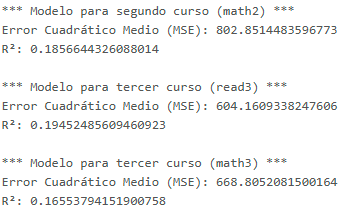






### EJERCICIO 14

**Repite los apartados 11 a 13 con Random Forest y compara los resultados de la regresión lineal y del random forest.**

**** ****

Para que quede mas claro a nivel visual lo presento en formato tabla:

|  |  |  |
| --- | --- | --- |
| Curso | Error Cuadrático Medio (MSE) | R² |
| Jardín de infancia (readk) | 422.223112 | 0.122049 |
| Jardín de infancia (mathk) | 1034.338415 | 0.099816 |
| Primer curso (read1) | 1337.949005 | 0.182232 |
| Primer curso (math1) | 861.218210 | 0.181068 |
| Segundo curso (read2) | 824.227499 | 0.233066 |
| Segundo curso (math2) | 802.851448 | 0.185664 |
| Tercer curso (read3) | 604.160934 | 0.194525 |
| Tercer curso (math3) | 668.805208 | 0.165538 |



## PARTE OPCIONAL

### EJERCICIO 15

**¿Podrías ordenar las variables según su importancia en relación con las capacidades de lectura y escritura en jardín de infancia, primero, segundo y tercero?**

He utilizado un modelo de Random Forest en PySpark para calcular la importancia de variables relacionadas con las capacidades de lectura y escritura en distintos cursos. Seleccioné las columnas relevantes, imputé valores faltantes con Imputer, ensamblé las variables de entrada con VectorAssembler, y entrené un RandomForestRegressor para cada curso y habilidad. Finalmente, ordené las variables por su importancia extraída del modelo y repetí el proceso para jardín de infancia, primero, segundo y tercero.

A lo largo de los cursos, lunch (programa de almuerzo), school (escuela), y birth (fecha de nacimiento) destacan como las variables con mayor impacto, especialmente en jardín de infancia y primer curso.

# PRACTICA 6

## PARTE 1

### EJERCICIO 1

**¿Cuáles son las asociaciones más relevantes entre estos productos? Pruebe con diferentes combinaciones de confianza, soporte y lift.**

El código de este ejercicio está recogido en asociacionesRelevantes.py.

1. Mejor asociación por lift:

* (citrus fruit, root vegetables) → (other vegetables) tiene un lift de 3.03, lo que indica que la probabilidad de comprar "other vegetables" aumenta significativamente cuando se compran "citrus fruit" y "root vegetables" juntos.

1. Alto soporte y confianza:

* (curd, yogurt) → (whole milk) tiene una confianza del 58.2%, mostrando una fuerte dependencia entre estos productos.
* (root vegetables, tropical fruit) → (other vegetables) tiene soporte de 1.23% y confianza del 58.4%, mostrando frecuencia y asociación robusta.

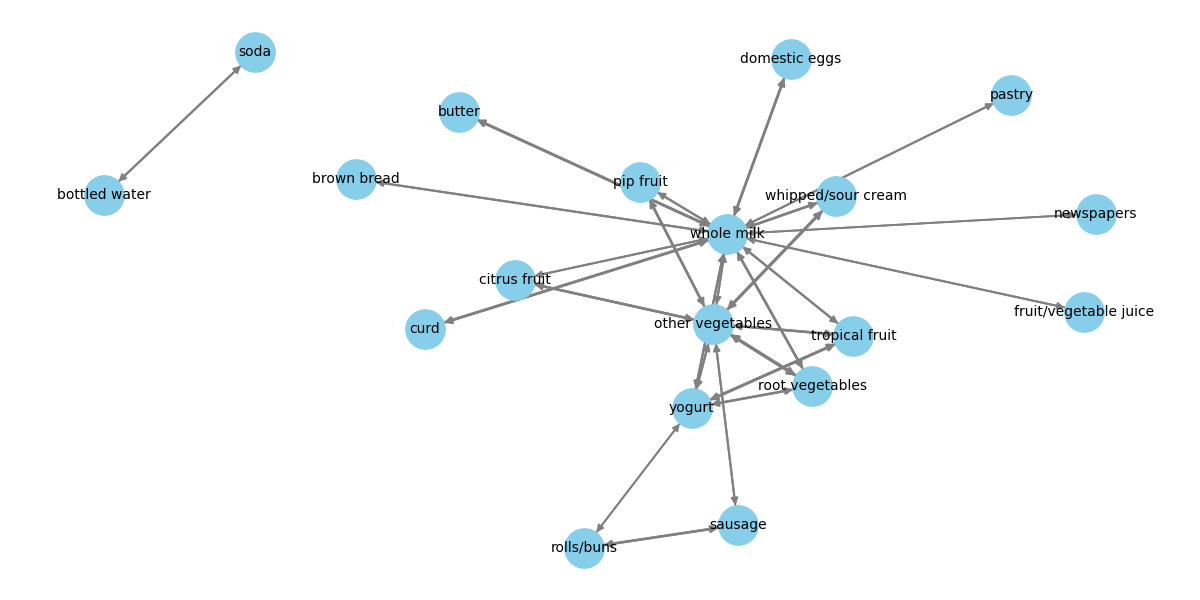
1. Patrones generales:

* "Whole milk" es un producto que aparece frecuentemente como consecuencia en múltiples asociaciones, especialmente con combinaciones de productos como (butter, other vegetables), (yogurt, other vegetables), y (rolls/buns, root vegetables), indicando que es un producto central en las compras.
* "Other vegetables" también tiene asociaciones frecuentes con combinaciones que incluyen "root vegetables" y "tropical fruit".

### EJERCICIO 2

**Visualice un grafo con las reglas más relevantes (en el grafo deben mostrarse más de 10 reglas y menos de 100).**

El código de este ejercicio está recogido en Grocery\_V2.py.





### EJERCICIO 3 (OPCIONAL)

**Dé un ejemplo de cómo usaría esta información para decidir un producto en el que realizar un descuento y un producto asociado a este anterior al que se le pueda elevar el precio y compensar el margen comercial perdido en el descuento.**

1. Producto con descuento: "Whole milk".

* Es un producto que aparece frecuentemente como consecuencia en varias asociaciones. Ofrecer un descuento en "Whole milk" atraerá a clientes y generará un mayor volumen de ventas.

1. Producto con aumento de precio: "Yogurt".

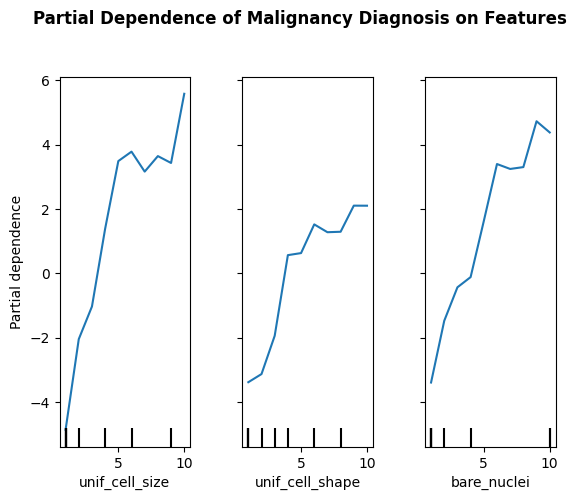
* Razón: La regla (yogurt, other vegetables) → (whole milk) muestra una fuerte asociación con un lift de 2.01 y una confianza del 51.2%, lo que indica que los clientes que compran "whole milk" suelen comprar también "yogurt".
* Dado que "yogurt" tiene una relación fuerte con "whole milk", es probable que los clientes sigan comprándolo incluso con un pequeño aumento de precio, lo que compensará el margen comercial perdido por el descuento en "whole milk".

## PARTE 2

### EJERCICIO 1

**Gráfico de dependencia parcial mostrando cómo influyen en el diagnóstico las variables “Uniformity of Cell Size”, “Uniformity of Cell Shape” y “Bare Nuclei”.**

El código de este ejercicio está recogido en interpretableAI-negocio.ipynb.

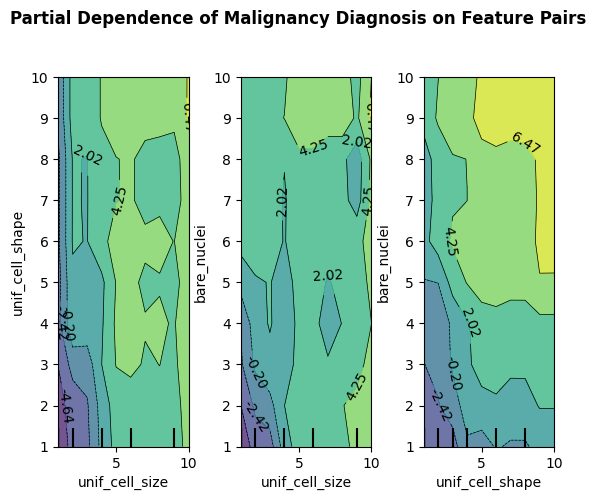




### EJERCICIO 2

**Gráfico de dependencia parcial mostrando la dependencia del diagnóstico con las combinaciones de variables anteriores, tomadas dos a dos.**

El código de este ejercicio está recogido en interpretableAI-negocio.ipynb.

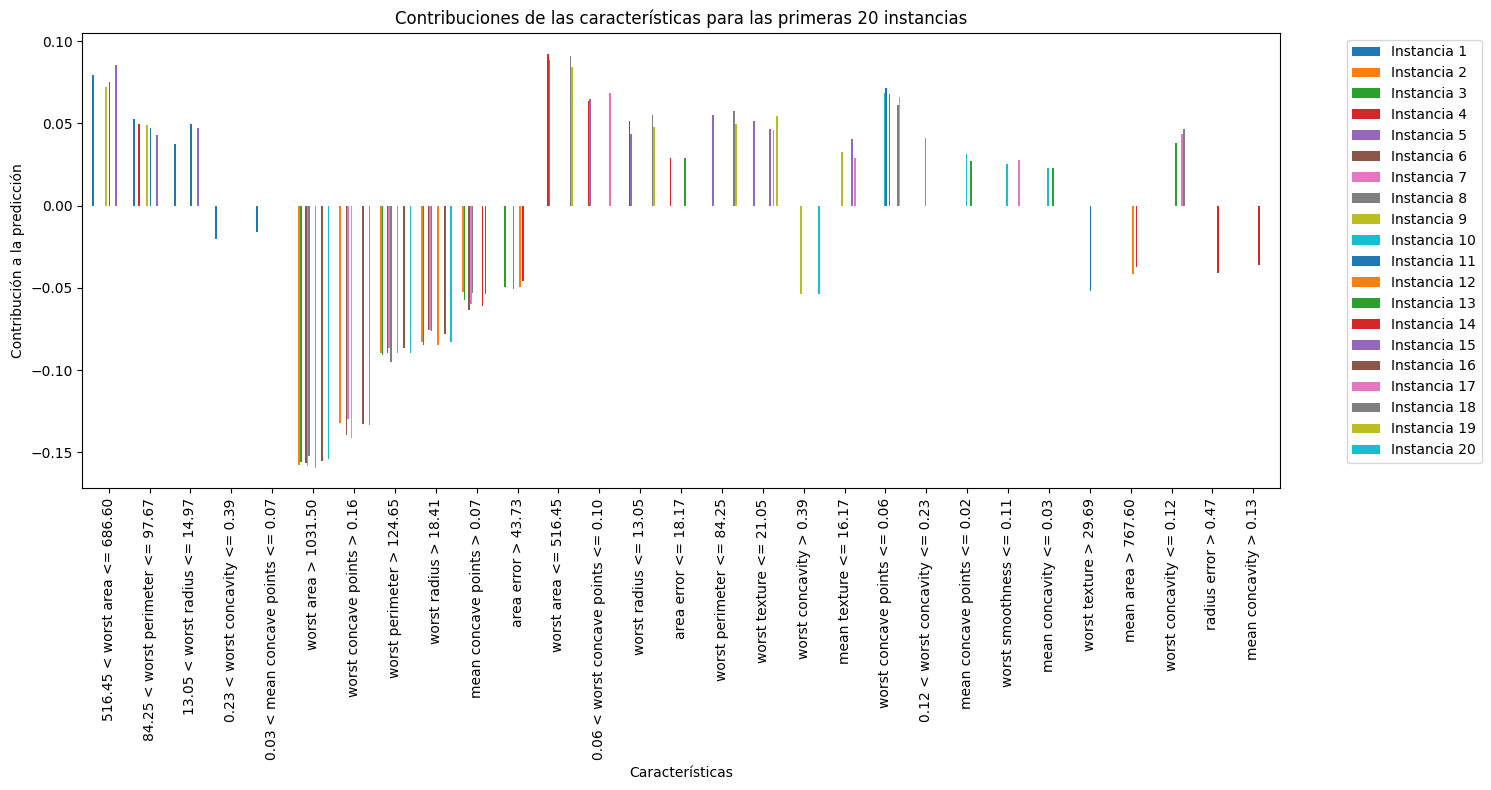
****



### EJERCICIO 3

**Explicación LIME del diagnóstico de las primeras 20 instancias del conjunto de test.**

El código de este ejercicio está recogido en interpretableAI-negocio.ipynb.

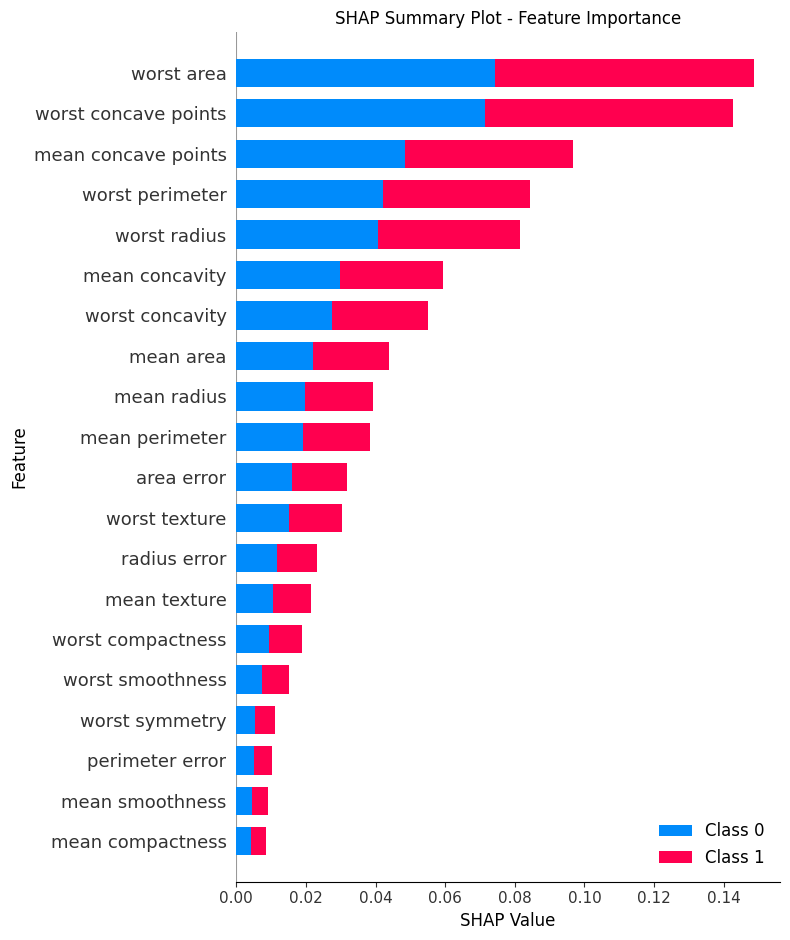
****

También se puede ver en el notebook un DataFrame sobre esto y más datos relevantes.



### EJERCICIO 4 (OPCIONAL)

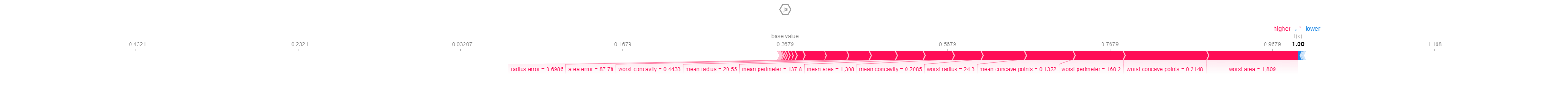
**Gráfico SHAP resumiendo la importancia de cada variable en el diagnóstico.**

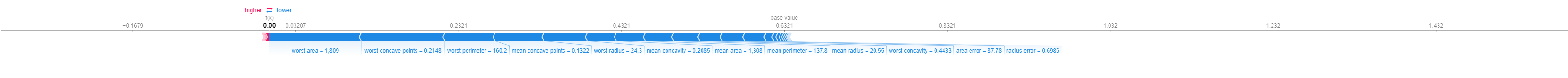




### EJERCICIO 5 (OPCIONAL)

**Gráficos de fuerza SHAP de una instancia benigna y de una instancia maligna del** **conjunto de test.**

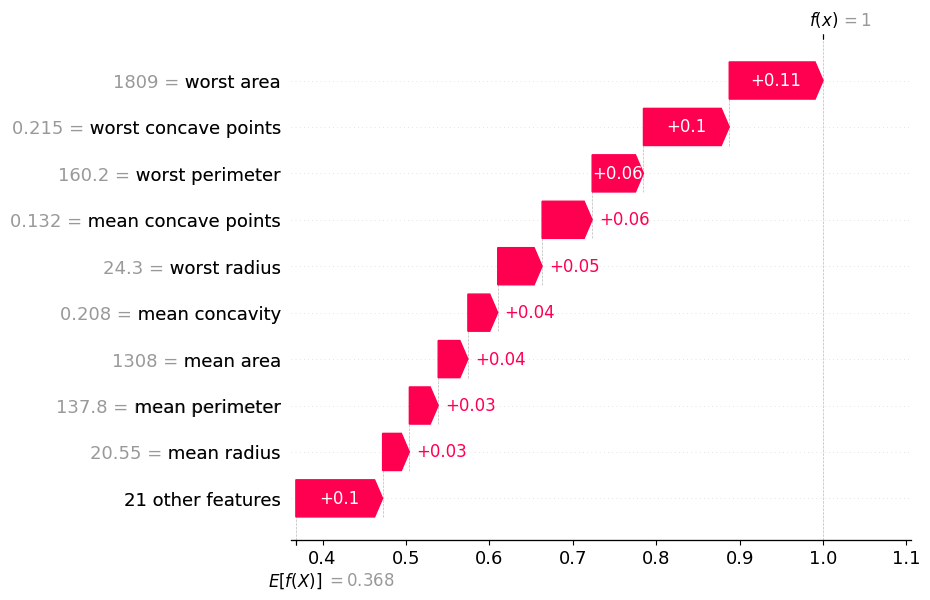


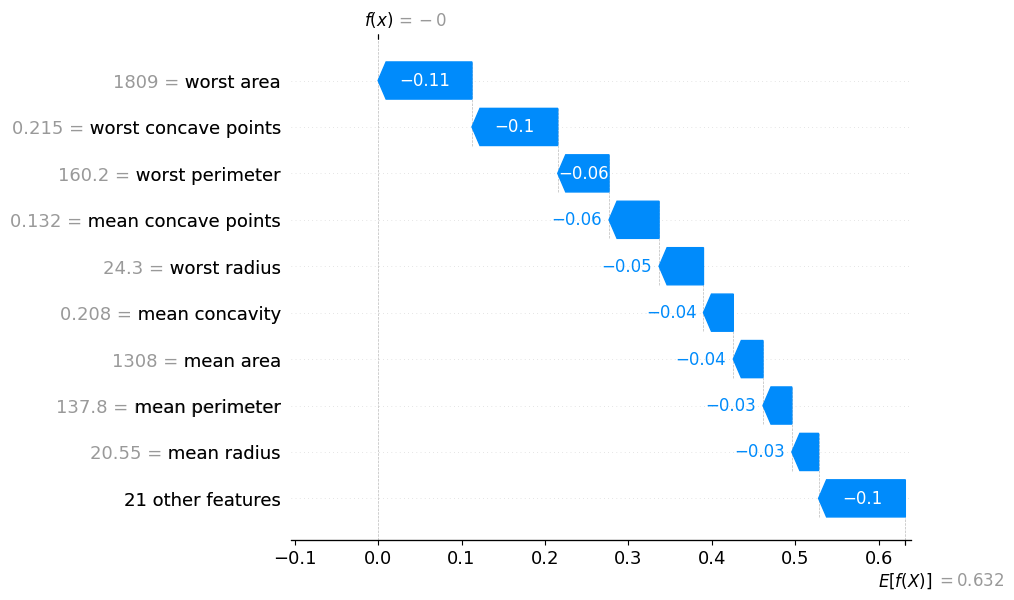




### EJERCICIO 6 (OPCIONAL)

**Gráfico de cascada (waterfall) de las mismas instancias que el apartado 5.**

****

****

# PRACTICA 7

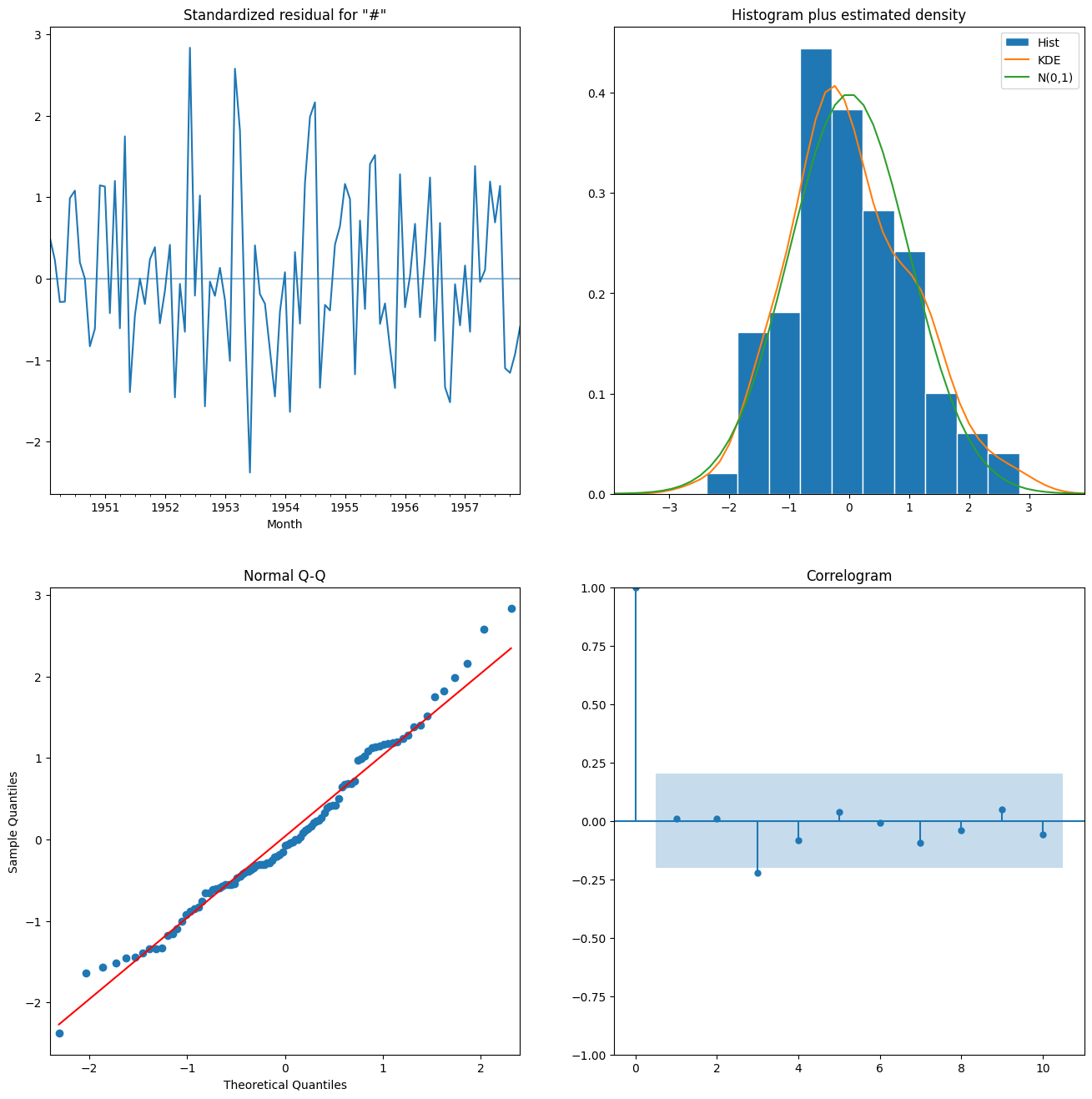
## PARTE OBLIGATORIA

### EJERCICIO 1

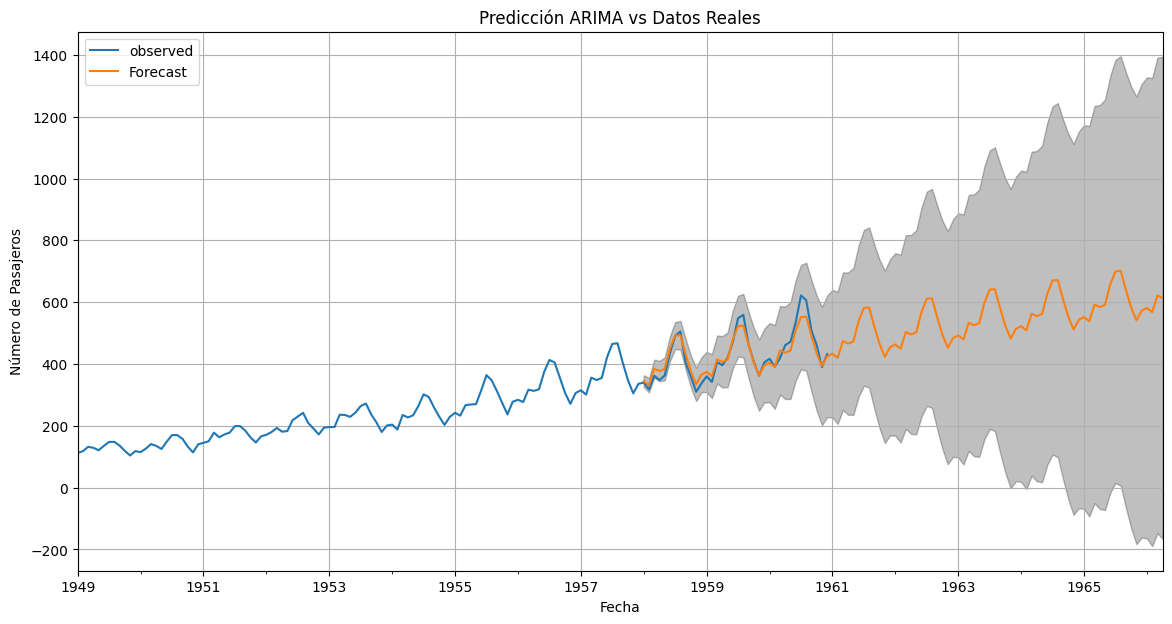
**Determinar los parámetros del modelo ARIMA con los que se obtenga el mejor ajuste en la serie AirPassengers. Valida el modelo en un periodo comprendido entre los años 1958, 1959 y 1960 (es decir, no deben emplearse los datos de estos tres años para aprender los parámetros del modelo).**

El código de este ejercicio está recogido en LeeAirPassengers.ipynb.

Con un RMSE de 22.132 los mejores parámetros son: ((1, 1, 0), (0, 1, 0, 12)) con AIC: 704.001. Ahora entrenamos el modelo con los mejores parámetros.



Ahora, realizamos la predicción a múltiples pasos:



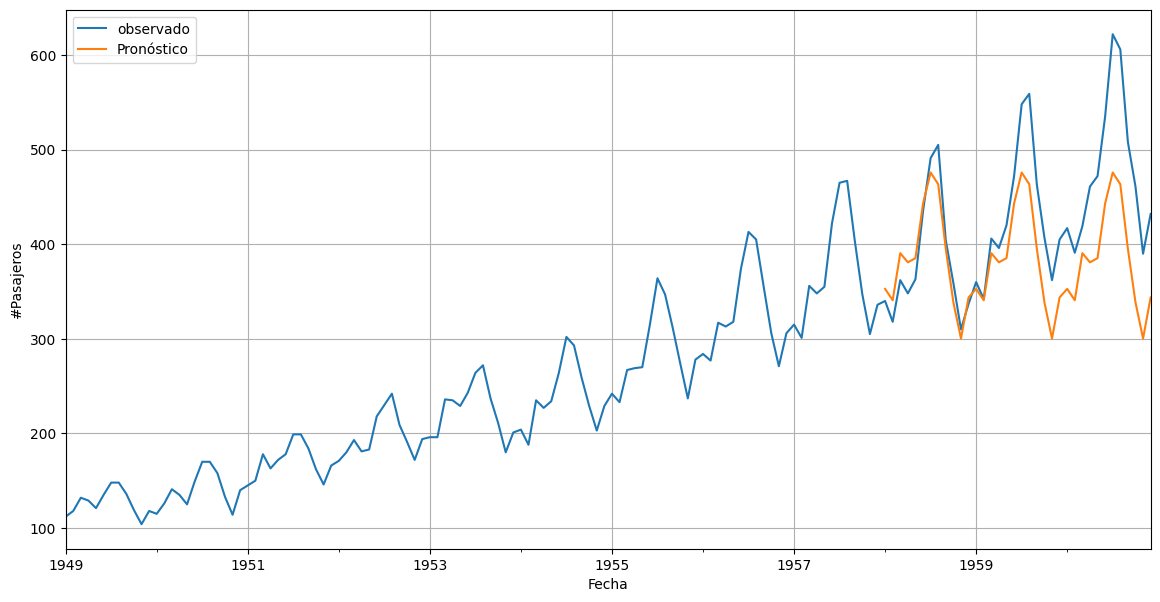
Y, por último, evaluamos los resultados:

* Error cuadrático medio ARIMA 536.28
* Raíz cuadrada de ECM ARIMA 23.16
* Error porcentual medio ARIMA 4.14



### EJERCICIO 2

**Ajustar el modelo Holt-Winters a esta serie y comparar sus resultados en el mismo conjunto de test.**

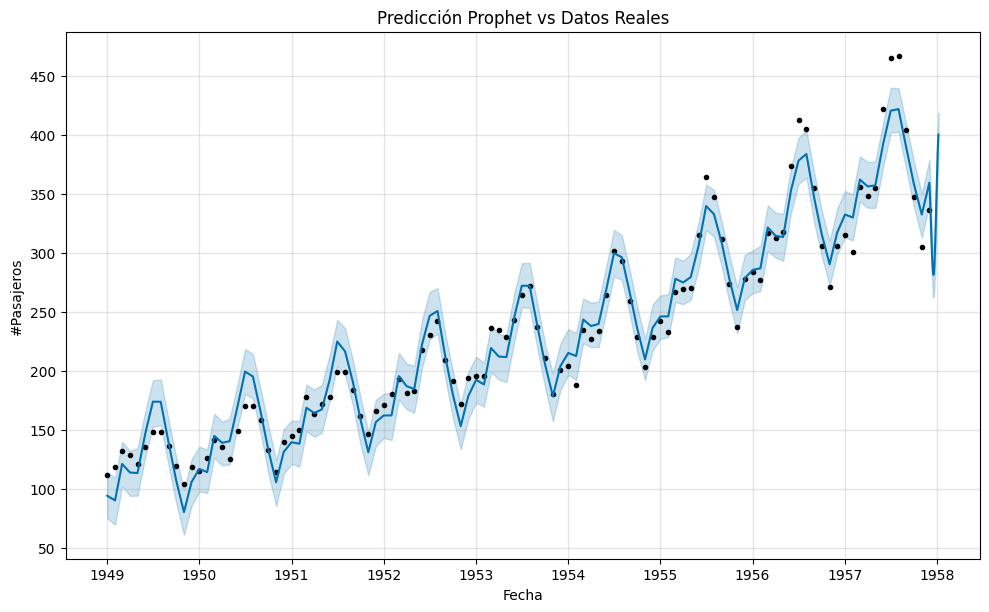
****

* Error cuadrático medio Holt-Winters 4295.53
* Raíz cuadrada de ECM Holt-Winters 65.54
* Error porcentual medio Holt-Winters 11.28



### EJERCICIO 3

**Ajustar Prophet a los mismos datos y comparar los resultados (la parte de series múltiples no es aplicable a este problema).**

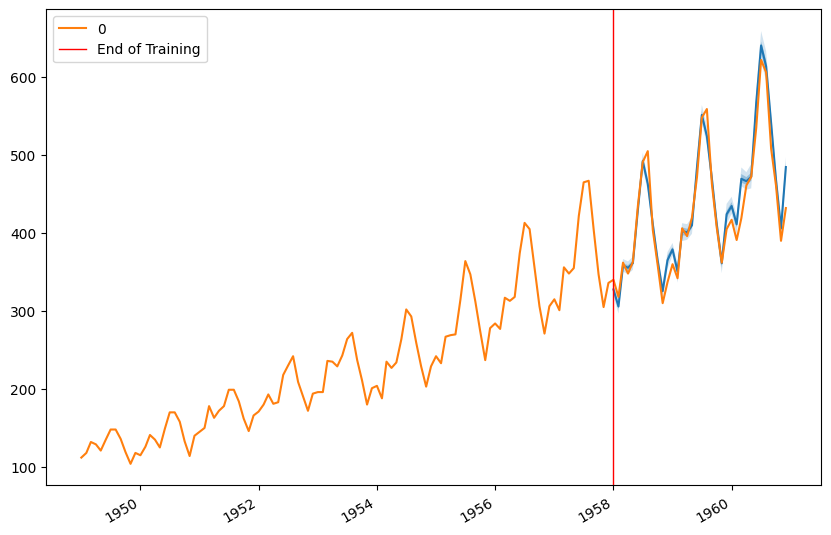


* Error cuadrático medio Prophet 1019.78
* Raíz cuadrada de ECM Prophet 31.93
* Error porcentual medio Prophet 9.39



### EJERCICIO 4

**Ajustar DeepAR a los mismos datos y comparar los resultados (la parte de datos múltiples no es aplicable a este problema).**

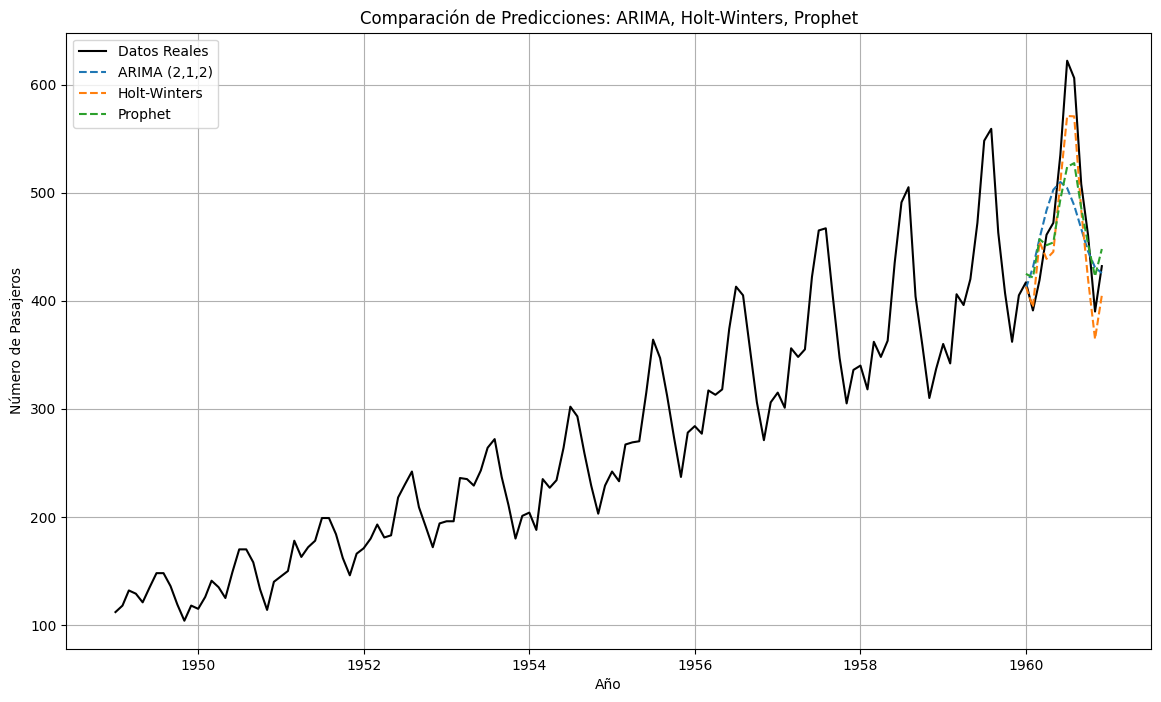
****

* Error cuadrático medio DEEPAR: 12620.82
* Raíz cuadrada de ECM DEEPAR: 112.34
* Error porcentual medio DEEPAR: 21.22%



### EJERCICIO 5

**Comparar entre sí las predicciones a largo plazo (años 1960 a 1970) de ARIMA, Holt Winters y Prophet.**

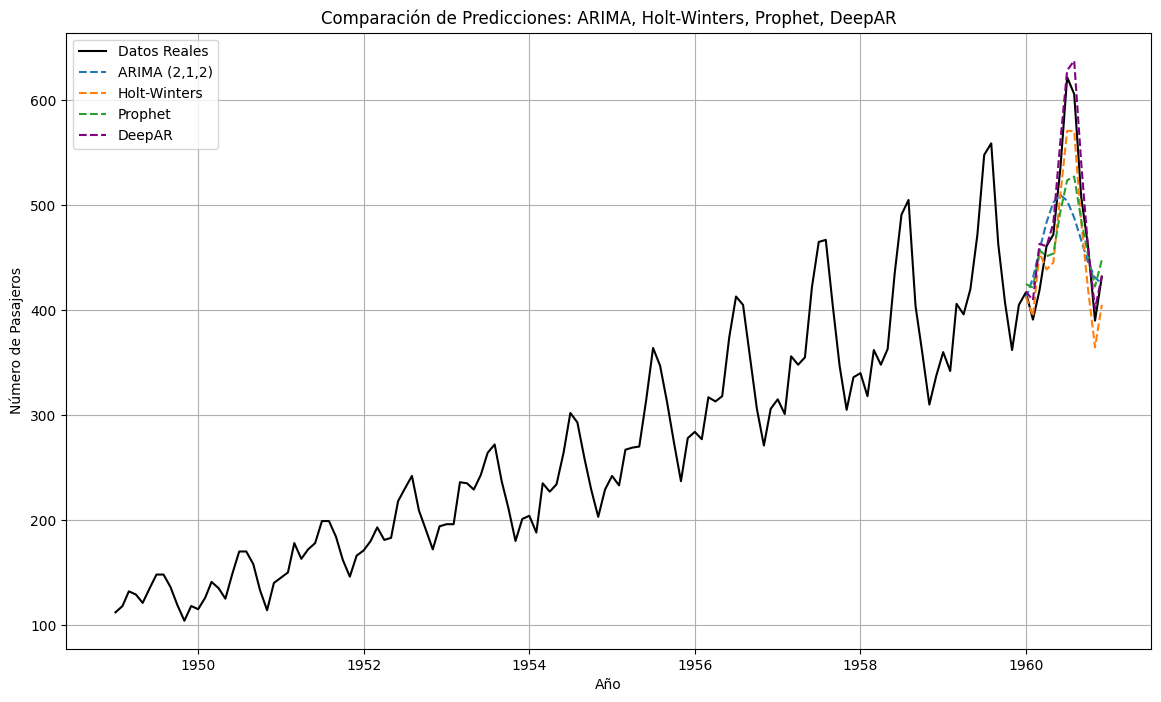


* RMSE ARIMA: 55.22
* RMSE Holt-Winters: 29.74
* RMSE Prophet: 43.07



### EJERCICIO 6 (OPCIONAL)

**Incluir las predicciones a largo plazo de DeepAR en la comparativa anterior.**



* RMSE ARIMA: 55.22
* RMSE Holt-Winters: 29.74
* RMSE Prophet: 43.07
* RMSE DeepAR: 21.05

# PRACTICA 8

**Elegir una tarea (diferente de “text classification”). En el siguiente enlace, en la sección de “Natural Language Processing” se pueden inspeccionar diferentes tareas de minería de texto.**

Mi elección ha sido Question Answering.

He elegido estos dos modelos, básicamente por su buen desempeño en castellano:

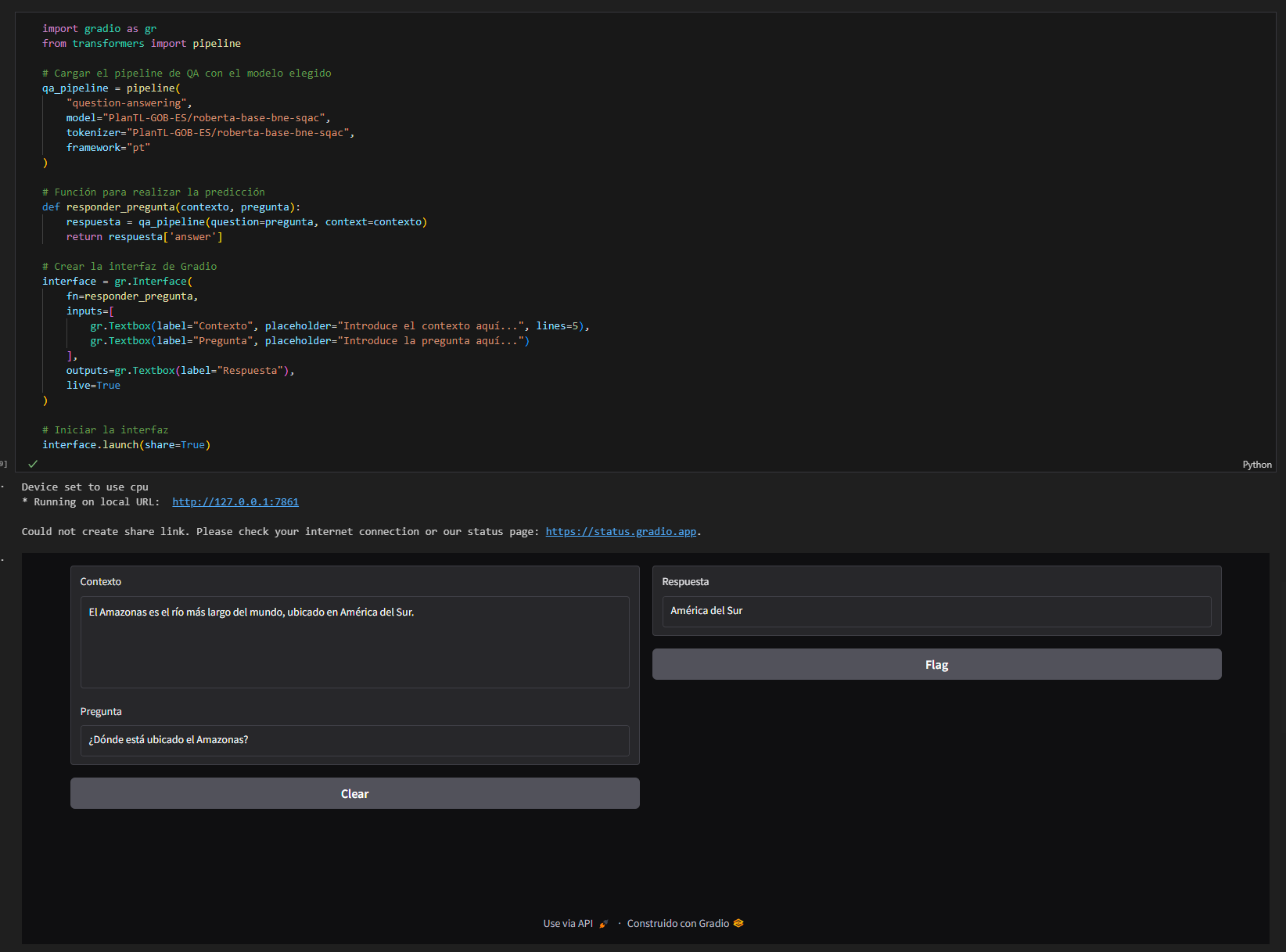
* bert-large-uncased-whole-word-masking-finetuned-squad
* PlanTL-GOB-ES/roberta-base-bne-sqac

**Evaluar sobre el dataset elegido y hacer una comparativa de los modelos. Cada modelo puede estar pre-entrenado con diferentes datasets, por lo que es de interés investigar el impacto tanto del modelo elegido como de los datos con los que fue entrenado.**



Podemos ver que con el contexto dado, el modelo PlanTL-GOB-ES/roberta-base-bne-sqac se comporta bastante mejor.

**Elegir el mejor modelo y crear una demo para desplegarlo.**

****