



Resultados proyecto: Aplicación de Distribuciones de Probabilidad

David Alejandro Cajiao Lazt

Juan Andrés López Álvarez

Joan Mateo Bermudez Collazos

Introducción

En el presente informe, se aborda la aplicación de diversas distribuciones de probabilidad en una serie de escenarios reales, dando cumplimiento al proyecto 2 de la asignatura Estadística y Probabilidad 1 del programa de Ciencias Básicas. El objetivo principal de este proyecto es analizar datos y tomar decisiones fundamentadas en la teoría de probabilidad, desarrollando así habilidades esenciales para la solución de problemas en contextos de ingeniería y ciencia de datos.

A lo largo del proyecto, se explorarán conceptos relacionados con distribuciones de probabilidad como la normal, exponencial, gamma, Poisson y binomial. Mediante la generación de muestras aleatorias y el análisis de las mismas, se buscará comprender el comportamiento teórico y empírico de estas distribuciones, contrastando los resultados obtenidos con los valores esperados.

Adicionalmente, se abordará el análisis de un sistema de procesamiento de imágenes que opera bajo una tasa de llegada de imágenes modelada por una distribución de Poisson. Esto permitirá evaluar la probabilidad de ocurrencia de diversos escenarios de interés, brindando insumos clave para la planificación y gestión de la infraestructura tecnológica.

Por último, se estudiará la distribución normal aplicada a la duración de la gestación humana, lo cual tendrá implicaciones relevantes en un contexto académico. Mediante el análisis de la función de densidad y función de distribución acumulativa, se determinará la probabilidad de eventos relacionados con el parto de una docente embarazada.

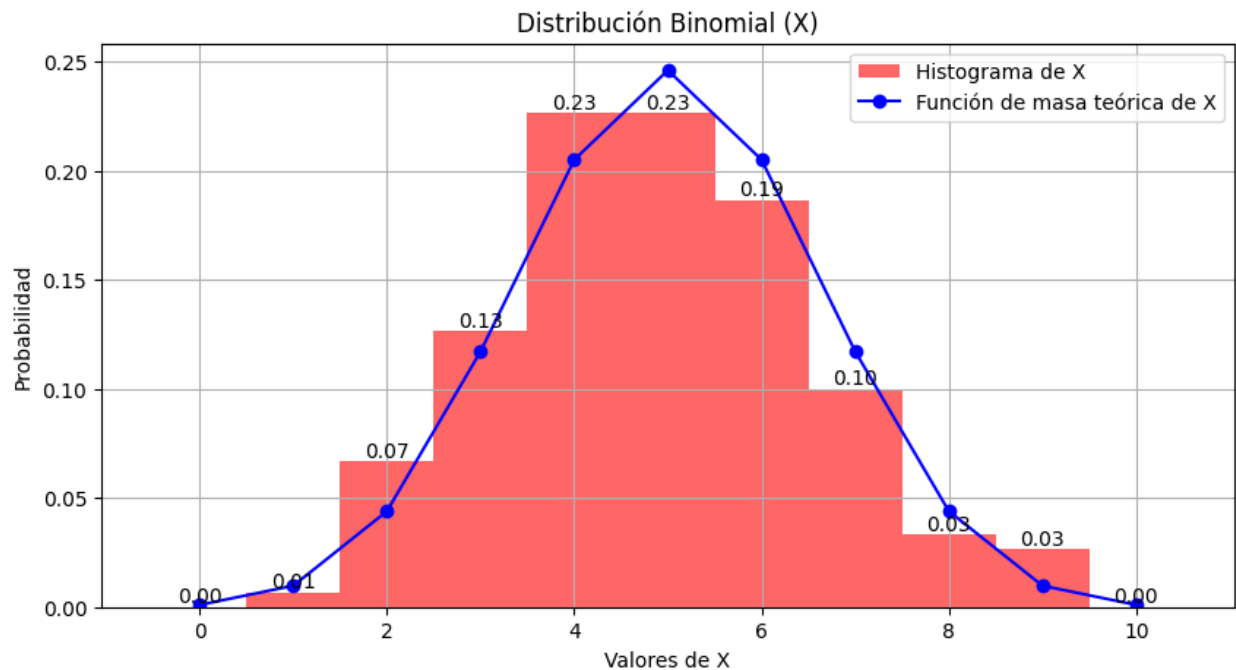
En resumen, este proyecto de estadística y probabilidad busca desarrollar en el estudiante la capacidad de aplicar herramientas y conceptos probabilísticos para la toma de decisiones informadas en escenarios de ingeniería y ciencia.

Desarrollo del proyecto

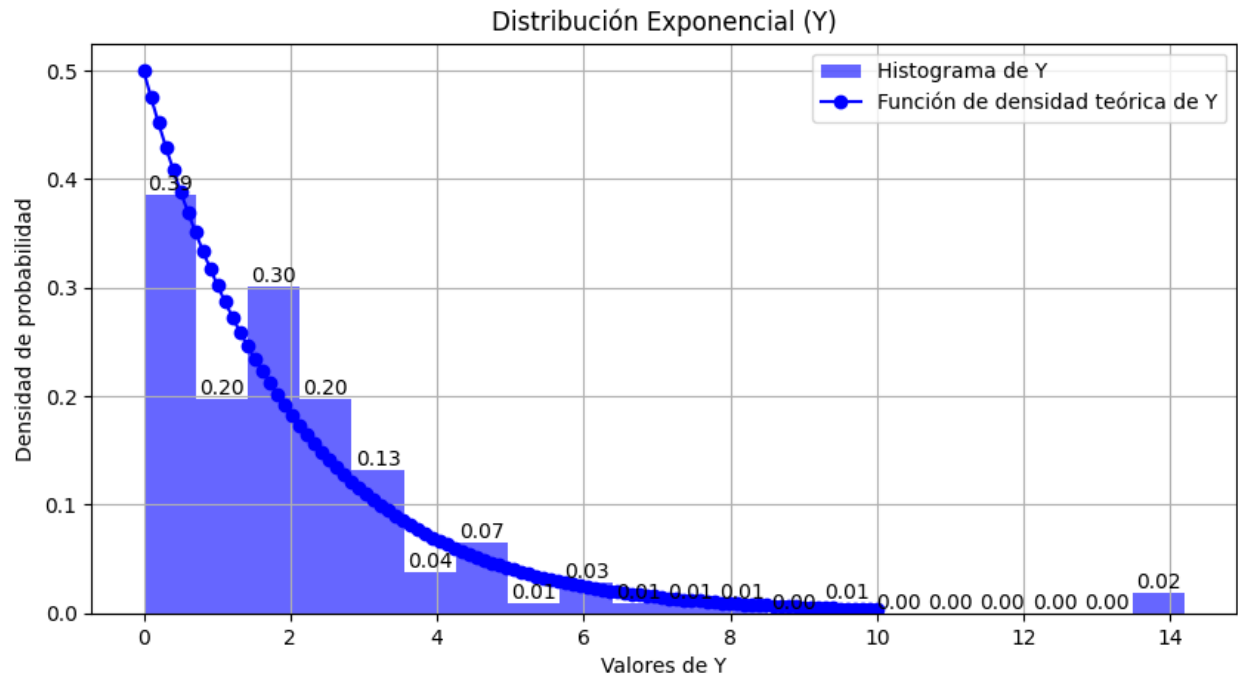
Interpretación de la Pregunta 1

En primer lugar, se abordó la generación de muestras aleatorias para una distribución binomial y una distribución exponencial, con el fin de comparar los resultados empíricos con los valores teóricos esperados.

Para la distribución binomial, con parámetros $n=10$ y $p=0.5$, se generó una muestra aleatoria de tamaño 150. A partir de esta muestra, se construyó un histograma de masa de probabilidad, que representa la frecuencia de ocurrencia de cada posible valor de la variable aleatoria X . Este histograma fue comparado con la función de masa de probabilidad teórica, obtenida mediante la función `binom.pmf` de la biblioteca `scipy.stats`. La comparación entre el histograma y la función teórica revela una concordancia notable entre los resultados empíricos y los esperados teóricamente, lo que sugiere que la distribución binomial es una buena aproximación para el fenómeno modelado.



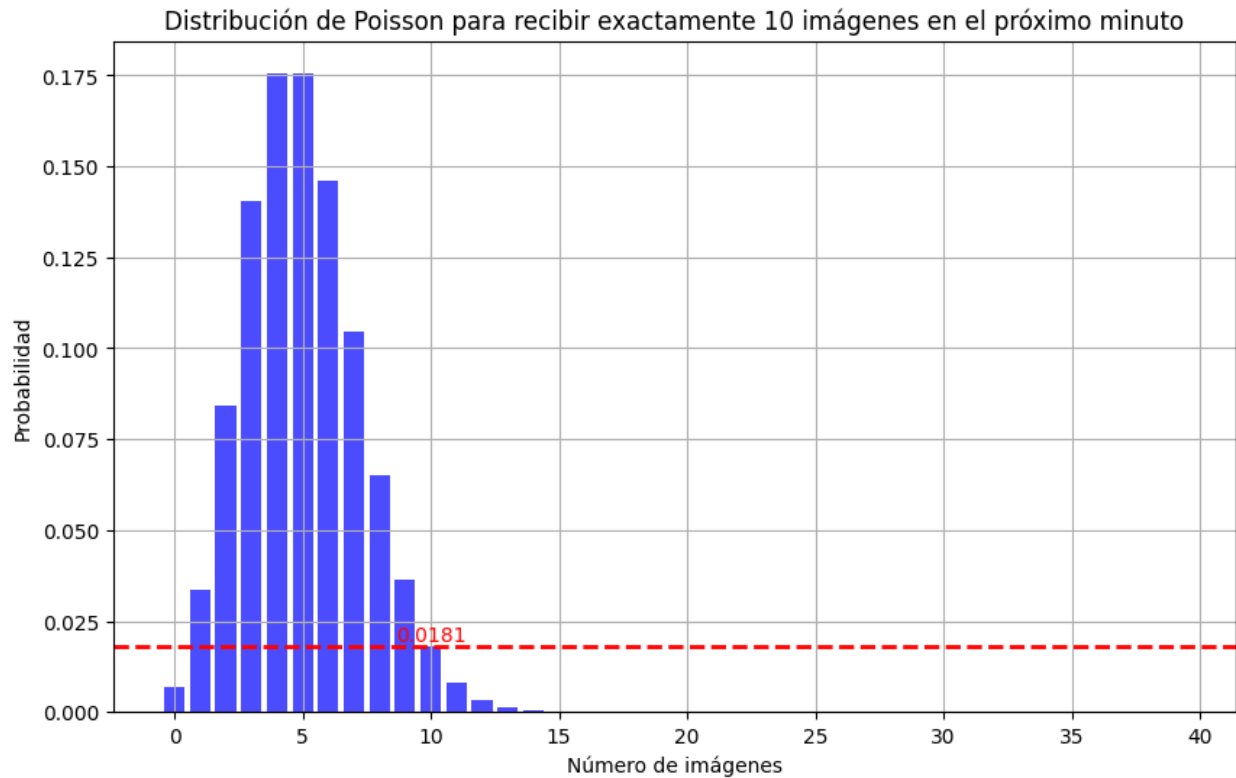
Por otro lado, para la distribución exponencial, con parámetro $\lambda=0.5$, también se generó una muestra aleatoria de tamaño 150. Se construyó un histograma de densidad de probabilidad, que muestra la distribución de los datos generados. Este histograma fue contrastado con la función de densidad de probabilidad teórica, obtenida mediante la función `expon.pdf` de la biblioteca `scipy.stats`. Al igual que en el caso de la distribución binomial, se observa una consistencia entre los resultados empíricos y los teóricos, lo que sugiere que la distribución exponencial es una buena aproximación para el fenómeno modelado.



Además, se calcularon estadísticas descriptivas como el promedio, los cuartiles y la desviación estándar para ambas distribuciones. Estos valores fueron comparados con los valores teóricos esperados, encontrándose una correspondencia satisfactoria entre ellos. En resumen, los resultados obtenidos en esta pregunta refuerzan la utilidad de las distribuciones de probabilidad para modelar fenómenos aleatorios en diversos contextos.

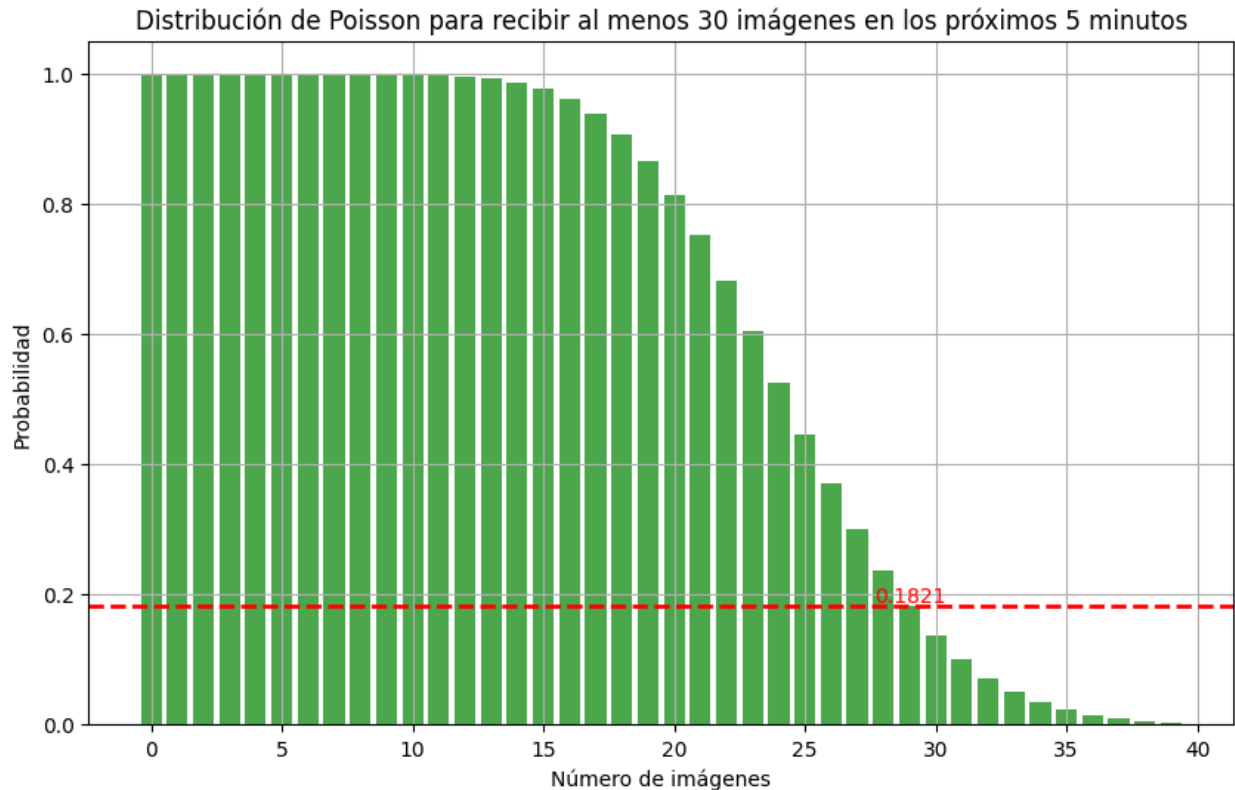
Interpretación de la Pregunta 2

En primer lugar, se calculó la probabilidad de que el sistema de procesamiento de imágenes reciba exactamente 10 imágenes en el próximo minuto. Este cálculo se realizó utilizando la función de masa de probabilidad de la distribución de Poisson, obtenida mediante la función `poisson.pmf` de la biblioteca `scipy.stats`. La probabilidad resultante fue de aproximadamente 0.0181.



Posteriormente, se calculó la probabilidad de que el sistema reciba al menos una imagen en el próximo minuto. Esta probabilidad se obtuvo de manera similar utilizando la función de masa de probabilidad de la distribución de Poisson, y resultó ser aproximadamente **0.033**.

Además, se determinó la probabilidad de que el sistema reciba al menos 30 imágenes en los próximos 5 minutos. Para ello, se utilizó la función de distribución acumulativa (`poisson.cdf`) y se restó el valor de la probabilidad acumulada hasta 29 imágenes de la unidad. La probabilidad resultante fue de aproximadamente 0.18, lo que indica una alta probabilidad de que el sistema reciba al menos 30 imágenes en dicho intervalo de tiempo.

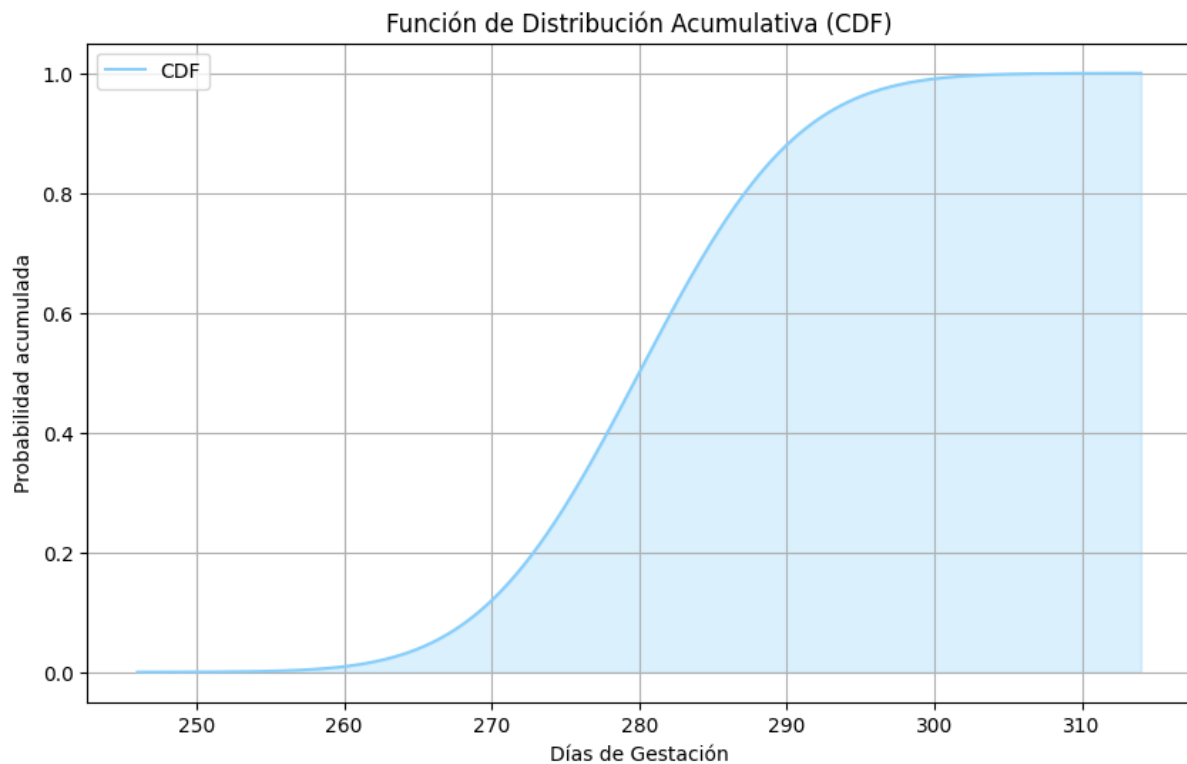
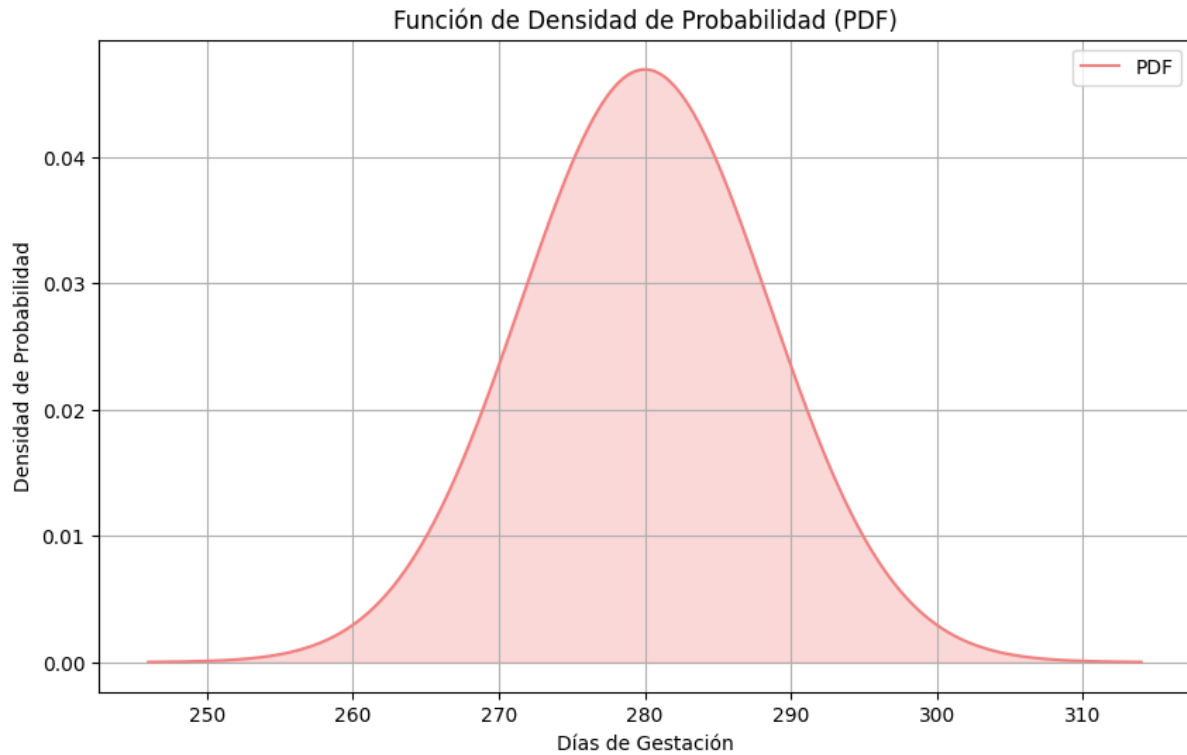


Finalmente, se calculó la probabilidad de que el sistema reciba un total de 150 imágenes o más durante un período de 30 minutos. Esta probabilidad se obtuvo nuevamente utilizando la función de distribución acumulativa de la distribución de Poisson, y el resultado fue aproximadamente 0.51. Esto sugiere que existe una probabilidad considerablemente alta de que el sistema reciba un número considerable de imágenes durante este intervalo de tiempo.

Interpretación de la Pregunta 3

Para abordar esta pregunta se creó una función para generar la distribución normal, obteniendo valores para el rango, la función de densidad de probabilidad (PDF) y la función de distribución acumulativa (CDF) sobre la variable aleatoria de días de gestación.

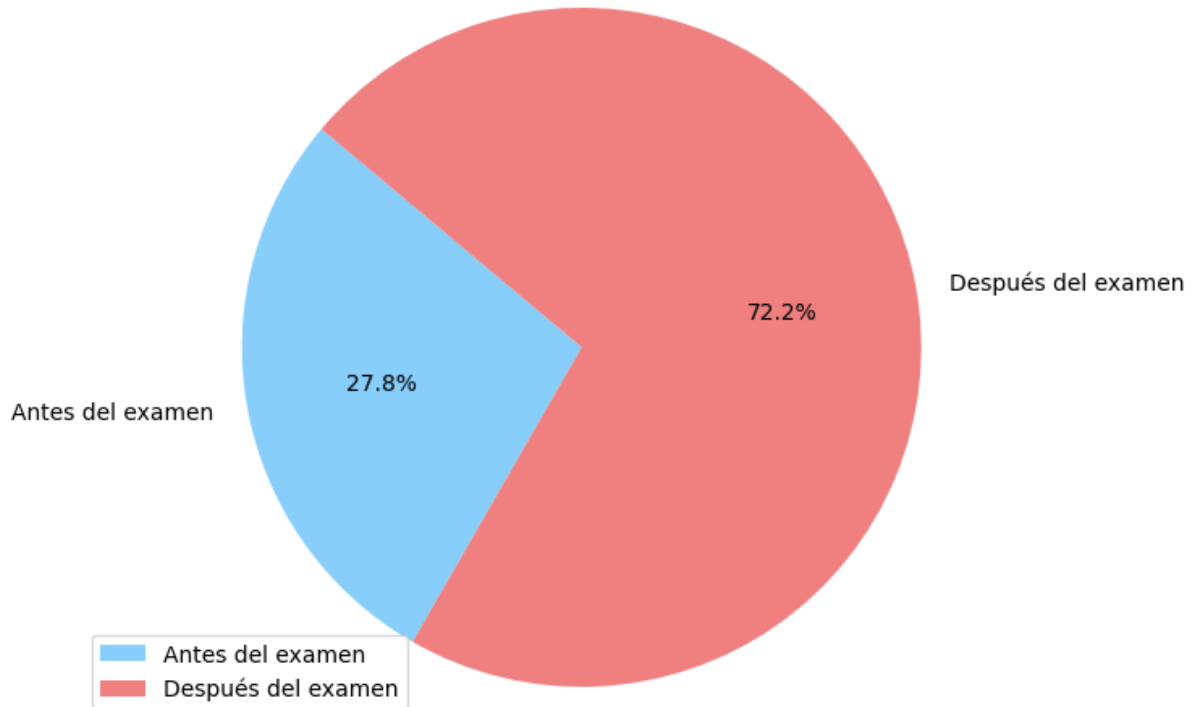
La línea `x = np.linspace(mu - 4*sigma, mu + 4*sigma, 1000)` se utiliza para generar un rango de valores de la variable aleatoria. En este caso, se elige un rango que se extiende hasta cuatro desviaciones estándar por encima y por debajo de la media. La distribución normal tiene la propiedad de que aproximadamente el 99.7% de los datos están dentro de tres desviaciones estándar de la media. Al extenderlo a cuatro desviaciones estándar, se cubre prácticamente la totalidad de la distribución.



Con el examen final programado para el 19 de abril y una fecha de parto prevista para el 24 de abril. Encontramos la probabilidad de que dé a luz el mismo día del examen final o antes utilizando la función de distribución

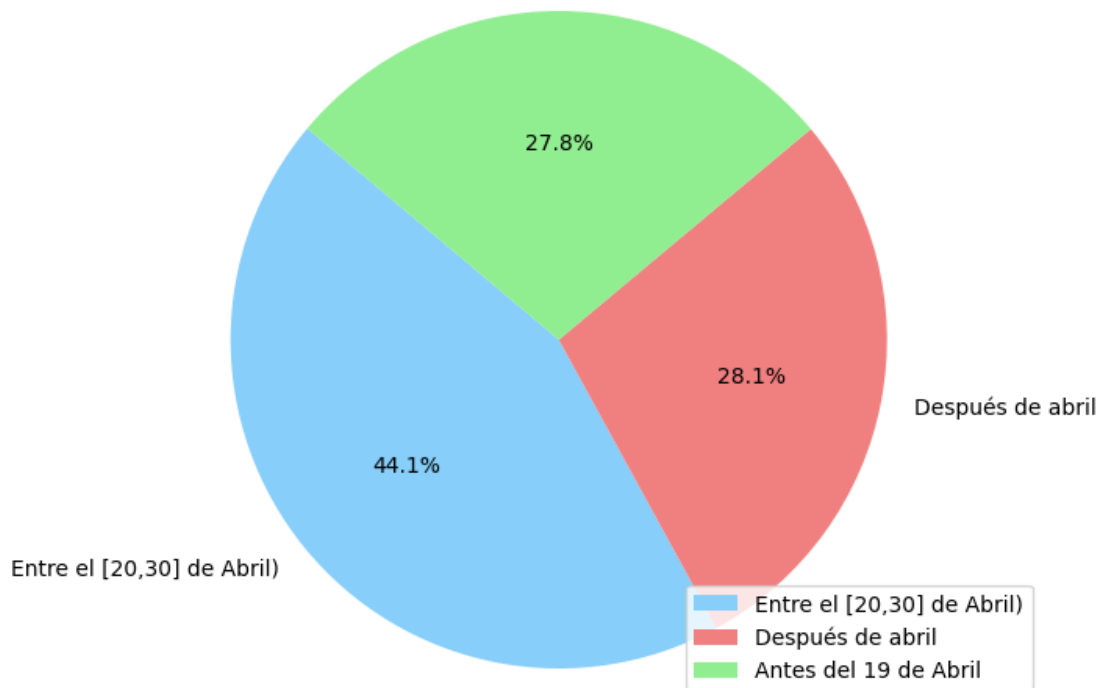
acumulativa (CDF) de la distribución normal. El resultado obtenido fue de aproximadamente 27.82%, lo que indica una probabilidad significativa de que el parto ocurra antes del examen final.

¿Cuál es la probabilidad de dar a luz el mismo día o antes del examen final? (Día 19)



Se determinó la probabilidad de que la profesora dé a luz después del examen final pero antes de que termine abril. Esta probabilidad se calculó utilizando la función de distribución acumulativa (CDF) en el rango del 20 al 30 de abril. El resultado mostró que la probabilidad de que el parto ocurra después del examen pero en abril es del 44.09%, lo que sugiere una alta probabilidad de que el parto ocurra durante ese período.

¿Cuál es la probabilidad de dar a luz luego del examen en Abril? (Entre el [20,30] de Abril)
Antes del 19 de Abril



Se calculó el día de gestación correspondiente al percentil 5 utilizando la función inversa de distribución acumulativa (ppf) de la distribución normal. Luego, se convirtió este día de gestación en una fecha respecto al 24 de abril (fecha de parto estimada) para determinar la fecha adecuada para el examen. Se encontró que la fecha del examen debe ser reagendada para el 10 de abril, lo que garantiza un 95% de probabilidad de que el parto ocurra después del examen.

Output:

Para que la probabilidad de que la profesora de a luz antes del examen sea
El examen debe ser reagendado para el: 10 de abril

Interpretación de la Pregunta 4

El análisis comienza con la identificación de patrones de tráfico web a partir de un conjunto de datos históricos que registran la tasa de llegada de solicitudes a un servidor web durante intervalos de 5 minutos. Estos patrones son esenciales para comprender y gestionar la carga del servidor, así como para planificar la infraestructura y garantizar un rendimiento óptimo.

Sobre el análisis y patrones de demanda del servidor

Para este punto se planteó una analítica bastante profunda partiendo de la base de analizar el rango de tiempo en la que los datos fueron tomados. Al tener más de 100k registros, y establecer que los registros han sido tomados cada 5 minutos, significa que:



1 registro = 5 minutos

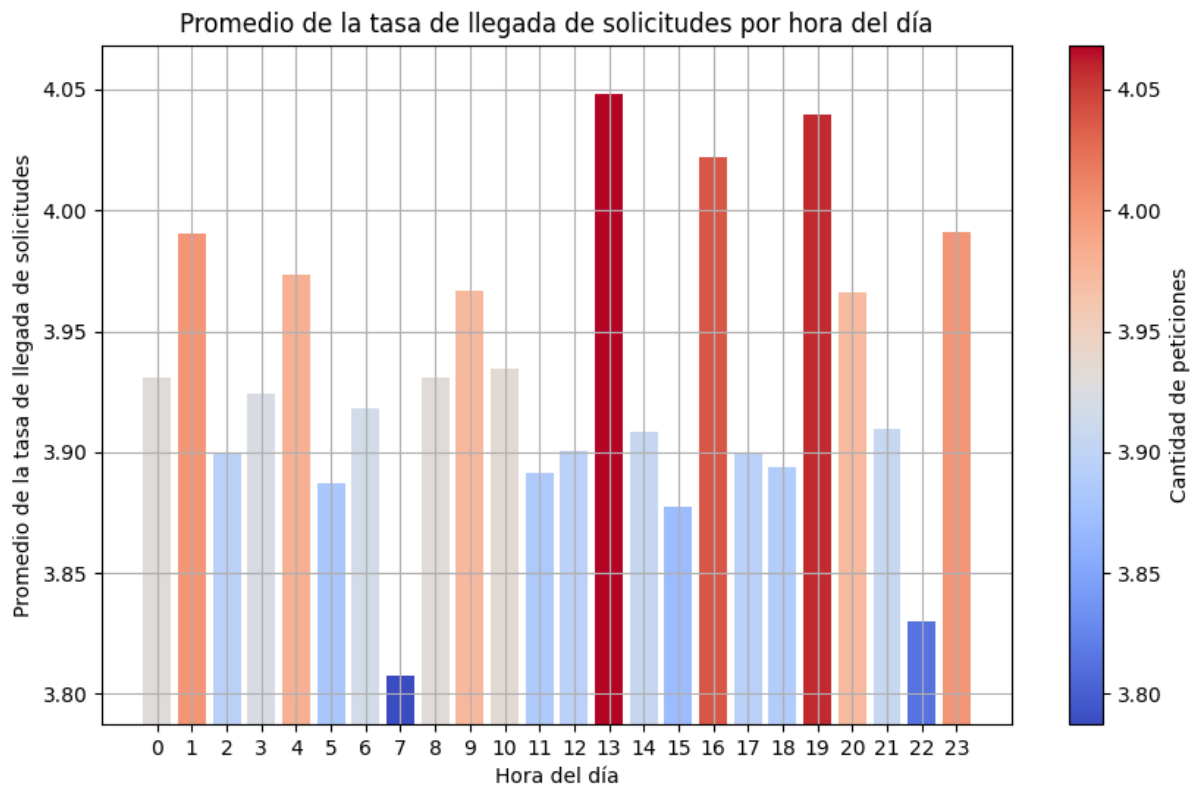
12 registros = 1 hora

288 registros = 1 día

8640 registros = 1 mes

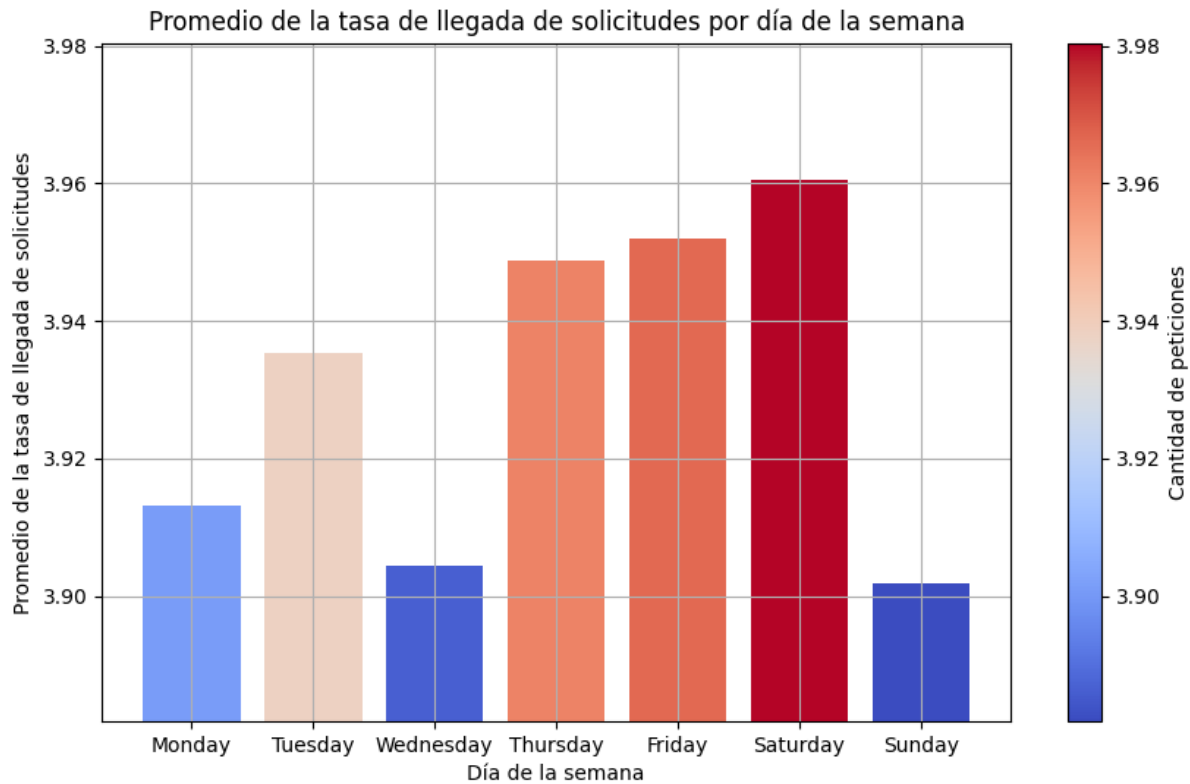
Teniendo esta información, podemos realizar análisis de tráfico en diferentes marcas temporales (horas al día y días a la semana, meses al año)

Promedio de peticiones por cada hora del día



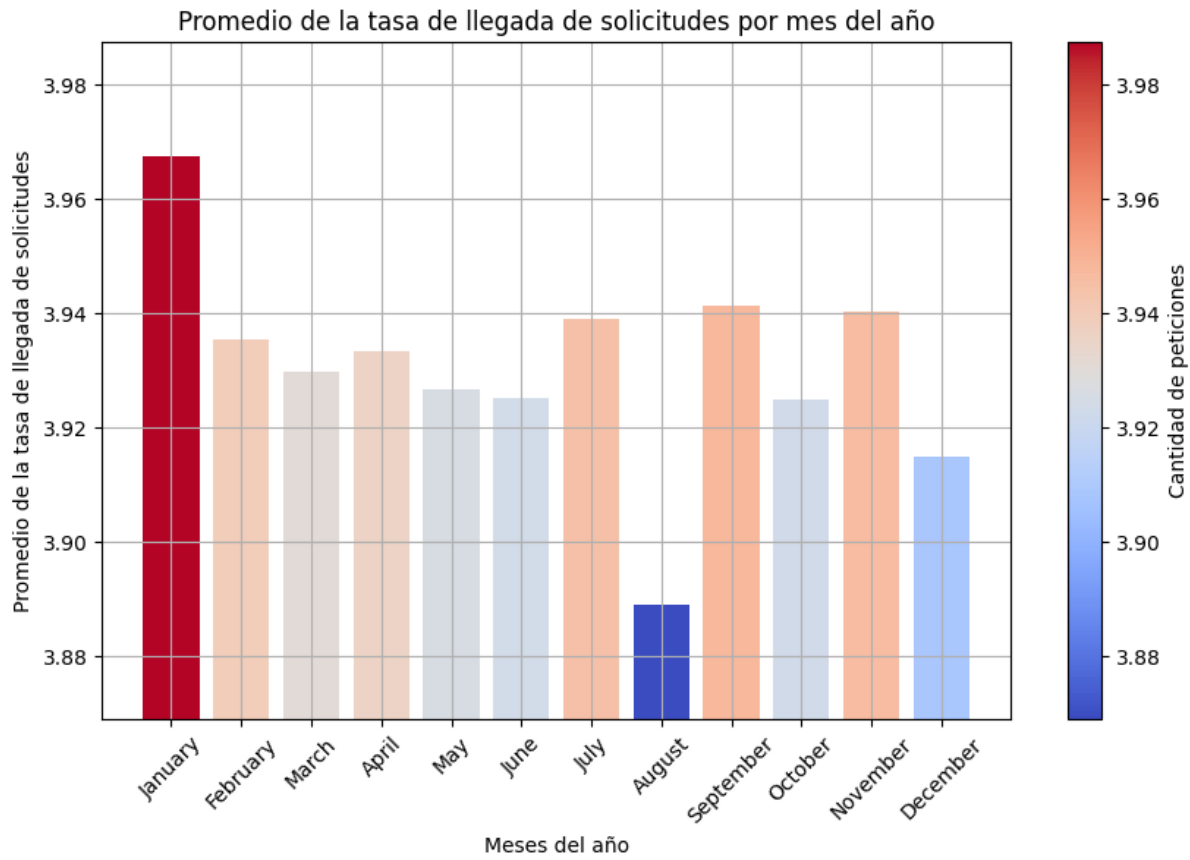
Como podemos ver, las horas del día en el que se presenta mayor número de peticiones el servidor en promedio es a las 13:00, 19:00, 16:00. Mientras que las horas en las que el servidor recibe menos peticiones es en las horas 7:00 y 22:00 ¿Qué podemos decir con esto? si se requiere aplicar actualizaciones al servidor, estas son las mejores horas para esa tarea ya que el tráfico es el menor durante todo el día

Promedio de peticiones por cada día de la semana



En términos de la carga por cada día de la semana, podemos ver que antes de finalizar la semana la tendencia es creciente desde el miércoles hasta el sábado. Mientras que el domingo y el miércoles son los días en los que menos peticiones en promedio le llegan al servidor respecto a los demás días de la semana, de manera que estos son los mejores días en caso de que se requiera hacer mantenimiento físico al servidor.

Promedio de peticiones por cada mes del año



Si hablamos de la tasa promedio de peticiones en función de los lapsos temporales mensuales, estaríamos observando que al inicio del año es cuando más se dispara la tasa promedio de peticiones, ya que ese mes se registra el promedio más alto, mientras que durante un poco más de la mitad del año, se reporta el índice de peticiones promedio más bajo de todo el año. Con esto podemos inferir que el momento adecuado del año para realizar cambios grandes en la infraestructura del servidor, es adecuado hacerlo durante el mes de agosto.

Ajuste de la Distribución Gamma

Se ajusta una distribución gamma a los datos de la tasa de llegada de solicitudes para comprender mejor su comportamiento. Los parámetros de la distribución gamma ajustada son:

- **Forma (Shape):** 1.28632
- **Locación (Loc):** -0.000102199
- **Escala (Scale):** 3.05596

A continuación, se plotea un histograma de los datos junto con la función de densidad de probabilidad (PDF) de la distribución gamma ajustada. Este gráfico permite visualizar cómo la distribución gamma se ajusta a los datos observados.

Cálculo de Probabilidades

1. **Probabilidad de recibir más de 8 solicitudes durante el próximo intervalo:** Se calcula la probabilidad de que el servidor reciba más de 8 solicitudes durante el próximo intervalo, utilizando la distribución gamma ajustada. El resultado muestra que la probabilidad es del 11.64%.
 2. **Probabilidad de que la tasa de llegada supere 15 solicitudes por minuto en el próximo intervalo:** Se determina la probabilidad de que la tasa de llegada supere un umbral crítico de 15 solicitudes por minuto en el próximo intervalo. La probabilidad calculada es del 1.36%.
 3. **Probabilidad de superar la capacidad máxima de procesamiento del servidor (25 solicitudes por intervalo) en el siguiente intervalo:** Se estima la probabilidad de que la tasa de llegada de solicitudes supere la capacidad máxima de procesamiento del servidor, que es de 25 solicitudes por intervalo. El resultado revela una probabilidad muy baja del 0.06%.
-

Conclusiones

En este proyecto, hemos explorado de manera exhaustiva la aplicación de diversas distribuciones de probabilidad en contextos reales, demostrando su utilidad en la modelación y análisis de fenómenos aleatorios. A través de la generación de muestras aleatorias y el análisis de datos empíricos, pudimos contrastar los resultados obtenidos con las expectativas teóricas, encontrando una notable concordancia entre ambos.

La distribución binomial y exponencial emergieron como herramientas efectivas para modelar fenómenos discretos y continuos, respectivamente. La consistencia entre los resultados empíricos y teóricos resalta la robustez de estas distribuciones en la aproximación de eventos reales.

Además, abordamos el análisis de un sistema de procesamiento de imágenes utilizando la distribución de Poisson, lo que nos permitió evaluar la probabilidad de ocurrencia de diversos escenarios de interés. Este enfoque resulta invaluable en la planificación y gestión de la infraestructura tecnológica, donde la precisión en las predicciones es crucial.

La aplicación de la distribución normal en el contexto de la duración de la gestación humana evidenció su relevancia en situaciones prácticas, como la planificación de eventos académicos. Mediante el análisis de probabilidades, pudimos tomar decisiones fundamentadas para la reprogramación de eventos, minimizando riesgos y asegurando una gestión eficiente del tiempo.

Por último, el análisis de patrones de tráfico web y el ajuste de la distribución gamma a los datos nos proporcionaron insights valiosos sobre el comportamiento de sistemas complejos. La capacidad para prever y gestionar la carga del servidor es crucial para garantizar un rendimiento óptimo y una experiencia del usuario satisfactoria.

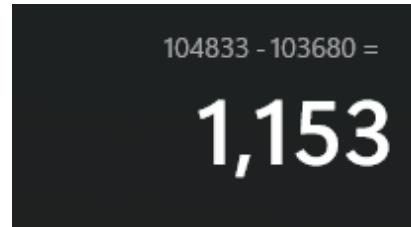
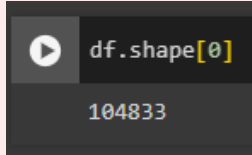
Sobre la cantidad de datos

Si cada registro se tomó en un lapso de 5 minutos eso significa que

- 1 registro = 5 minutos
- 12 registros = 1 hora

¿Qué pasó con esos 1.153 registros de diferencia?

- 288 registros = 1 día
- 8.640 registros = 1 mes
- 103.680 registros = 1 año



- Siguiendo la misma lógica anterior, debemos entender que:

- 288 registros = 1 día


$$1153 \div 288 \approx 4 \text{ días}$$

- Para el análisis dimos por sentado que cada mes tiene 30 días, supuesto que al ser incorrecto nos generó un margen de error de 4 días (1153 registros).

Links

<https://github.com/DCajiao/Distribuciones-de-Probabilidad>

Proyecto 2: Distribuciones de Probabilidad - Google Drive

 <https://drive.google.com/drive/folders/1Zn4z8eorVPT4GbRCxGxqIeIPfcHyxJRM?usp=sharing>