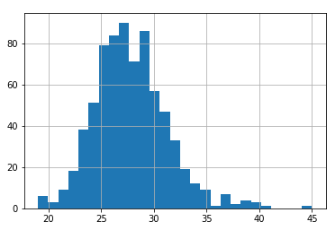
DATATHON UNI 2019

PRE-SELECCIÓN

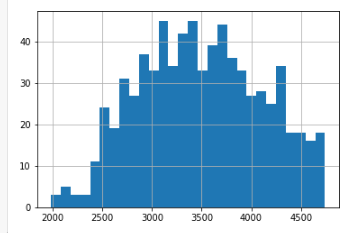
Pregunta 02:

1. Realizar un análisis descriptivo de las variables edad, peso al nacer, tipo de parto y uso de oxitocina. Describir los resultados.

Histograma de Edades



Histograma de Peso al nacer



Tipo de parto

inducido 326

instrumental 235

cesarea 170

inducido 0.445964

instrumental 0.321477

cesarea 0.232558

Uso de oxitocina

no 617

si 114

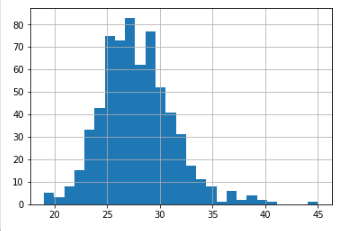
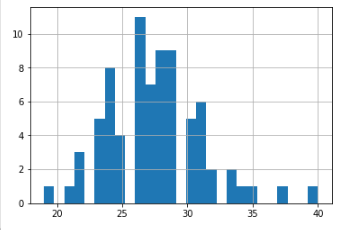
no 0.844049

si 0.155951

1. Realizar el mismo análisis distinguiendo dos grupos: uno que utiliza anestesia epidural y otro que no. Describir los resultados

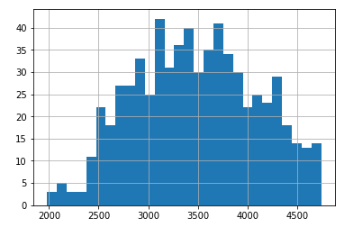
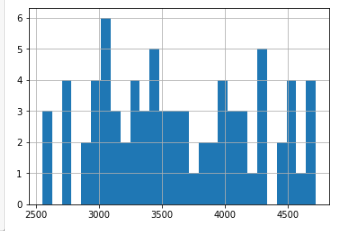
Histograma de Edades

Con epidural Sin epidural



Histograma de Peso al nacer

Con epidural Sin epidural



Tipo de parto

Con epidural

inducido 36

instrumental 21

cesarea 20

inducido 0.467532

instrumental 0.272727

cesarea 0.259740

Sin epidural

inducido 290

instrumental 214

cesarea 150

inducido 0.443425

instrumental 0.327217

cesarea 0.229358

Uso de oxitocina

Con epidural

no 67

si 10

no 0.87013

si 0.12987

Sin epidural

no 550

si 104

no 0.840979

si 0.159021

1. Estudiar el porcentaje de mujeres que utiliza anestesia epidural

Sin epidural

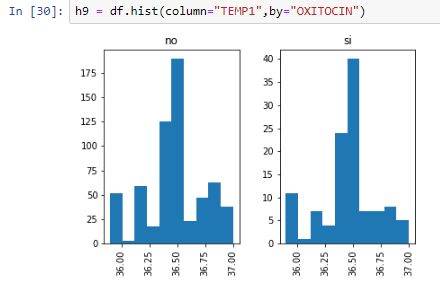
no 0.894665

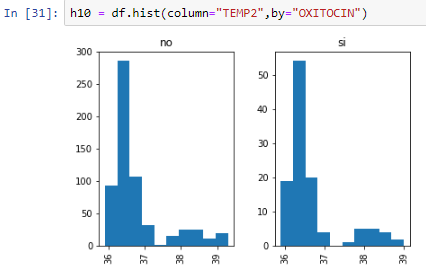
si 0.105335

1. Estudiar la temperatura media de la madre antes y después del parto. ¿Existen en general diferencias significativas?, ¿Y si distinguimos por grupos según el uso de anestesia epidural?

No existen diferencias significativas cuando se toma en cuenta el total de los datos, pero sí cuando se separan según el uso de anestesia epidural, se puede notar que para las mujeres que usan anestesia epidural la diferencia de temperatura entre TEMP1 y TEMP2 es mayor y signigicativa

1. ¿Existe relación entre el uso de oxitocina y la aparición de fiebre?





Se puede notar en los gráficos que no existe diferencia notable entre las pacientes que usan oxitocina y las que no

Pregunta 04:

¿Qué metodologías usas para un proyecto de Data Science?, y de la metodología escogida, ¿Qué parte del proceso consideras más importante? Explicar

Nuestro equipo está formado por físicos, así que usamos la Data Science como herramienta para comprender los diferentes fenómenos de la naturaleza. La metodología que usamos se base principalmente en:

1) Encontrar una fuente de datos del fenómeno que queremos analizar, por ejemplo, una fuente de datos de la dispersión de partículas elementales son los detectores.

2) Una vez obtenida la fuente procedemos a la recopilación de datos, en esta parte, para el mismo fenómeno, hacemos que se dé el fenómeno a través de simulaciones o en la realización misma del fenómeno.

3) Tratamiento de datos, sabemos que los datos vienen con ruido. Se desarrollan diferentes filtros para el limpiado de los datos; ya que el fenómeno es natural, la mayoría de los cortes que se hacen tienen justificación física.

4) Generalmente se desarrolla el experimento o la simulación teniendo en claro el fenómeno que se quiere analizar, así que se desarrollan diferentes técnicas estadísticas para la verificación del fenómeno.

5) Visualización de las variables del fenómeno a través de gráficos que relacionen las variables que lo caracterizan.

6) Explicación del fenómeno.

Según los tipos de modelos, ¿Cuáles son los indicadores de performance para datos desbalanceados? Explicar.

A los datos desbalanceados los entendemos como un conjunto de datos que no se ajustan a las coclusiones o hipótesis predichas. Los indicadores que nos pueden ayudar a entenderlos pueden ser:

1) exactitud: entendido como la probabilidad de clasificar bien un registro.

2) precición: probabilidad de que la clasificación anterior sea buena.

3) sensibilidad: probabilidad de detectar un registro positivo. Al registro positivo se entiende como el hecho que se lleve a cabo el fenómeno predicho.

4) Curvas ROC: Curvas en función de los registros positivos y negativos.

¿Qué estrategias utilizarías para tratar problemas de overfitting? Explicar.

Los problemas de overffiting los entendemos como los relacionados a las malas predicciones que se hacen en función a los datos estudiados. Por ejemplo, en la detección de nuevas partículas, siempre se da el problema de que la señal analizada la relación a los datos de background, pues con los datos que se manejan no se puede hacer una buena predicción para la caracterización de las nuevas partículas. Algunas maneras de solucionar el problema podrían ser:

1) Obtener mayor cantidad de eventos que mejoren la caracterización de las partículas, para así obtener nuevas propiedades que eran indetectables con pocos eventos.

2) Mejorar el filtro de datos utilizando mejores métodos estadísticos.

3) Leímos que el CERN utiliza Inteligencia Artificial para el tratamiento de datos, pues muchas veces hay funciones de ajuste que son muy difíciles de hacer.

Se cuenta con un dataset de 2500 registros y 10 variables. ¿Considera necesario el uso de Feature Engineering?

Entendemos al Featuring Engineering como un método de obtención de características de un conjunto de datos. Ya que en esta técnica se considera a los datos como materia prima y haciendo una analogía con la detección de partículas, consideramos: que los datos con las variables propuestas no son suficientes para obtener nuevas características a través del FE.

Pregunta 05:

1. 21790 hombres and 10771 mujeres
2. Edad promedio 36.86
3. 99 horas/semana, 7763 personas, 9.5% son ricos
4. 41 y 48