Solucionario BanBif

Grupo Race Againsta The Machine

Integrantes :

- Victor Villacorta Plasencia

- Joseph Vega Alaluna

- Carlos Rodriguez

#Pregunta 1

Select “select\_Columns” from tableA A full outer join tableB B on A.key=B.key where A.key is null or B.Key is null

#Pregunta 2

Select “select\_columns” from tableA A left join tableB B on A.key=B.key where B.key is null

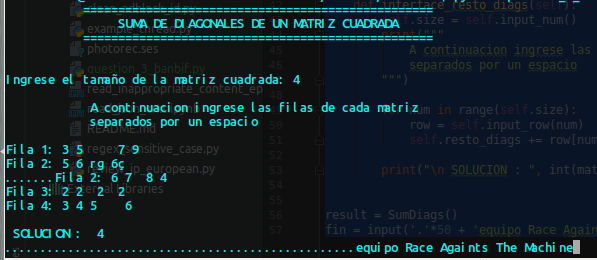
#Pregunta 3

[SELECT](http://localhost:81/phpmyadmin/url.php?url=https://dev.mysql.com/doc/refman/5.5/en/select.html) \* FROM ventas WHERE (Region,ventas) [IN](http://localhost:81/phpmyadmin/url.php?url=https://dev.mysql.com/doc/refman/5.5/en/comparison-operators.html#function_in) ( [SELECT](http://localhost:81/phpmyadmin/url.php?url=https://dev.mysql.com/doc/refman/5.5/en/select.html) Region, [MAX](http://localhost:81/phpmyadmin/url.php?url=https://dev.mysql.com/doc/refman/5.5/en/group-by-functions.html#function_max)(ventas) FROM ventas GROUP BY Region )

#Pregunta 4

import re  
import math  
  
  
class SumDiags:  
 def \_\_init\_\_(self):  
 self.resto\_diags = 0  
 self.size = 0  
  
 self.cabecera()  
 self.interface\_resto\_diags()  
  
 def cabecera(self):  
 print(' '\*10, '='\*50)  
 print(' '\*15, 'SUMA DE DIAGONALES DE UN MATRIZ CUADRADA')  
 print(' '\*10, '='\*50, '\n\n')  
  
 def input\_num(self):  
 try:  
 n = int(input('Ingrese el tamaño de la matriz cuadrada: '))  
 if n <= 1:  
 raise  
  
 return n  
 except Exception:  
 return self.input\_num()  
  
 def format\_row(self, cadena):  
 return re.sub(r'[ /t/s]', ' ', cadena).split()  
  
 def input\_row(self, orden, extra=''):  
 try:  
 row = input('{}Fila {}: '.format(extra, orden + 1))  
 row = [int(\_) for \_ in self.format\_row(row)]  
 if len(row) != self.size:  
 raise  
  
 return row  
 except Exception:  
 return self.input\_row(orden, extra='.'\*7)  
  
 def interface\_resto\_diags(self):  
 self.size = self.input\_num()  
 print("""  
 A continuacion ingrese las filas de cada matriz   
 separados por un espacio  
 """)  
  
 for num in range(self.size):  
 row = self.input\_row(num)  
 self.resto\_diags += row[num] - row[self.size - (1 + num)]  
  
 print("\n SOLUCION : ", int(math.fabs(self.resto\_diags)))  
  
  
result = SumDiags()  
fin = input('.'\*50 + 'equipo Race Againts The Machine')

Inclusion de tolerancia a fallos



**#Pregunta 5**

El algoritmo XGBoost sólo trabaja con datos numéricos. ¿Qué tratamiento hace a sus

variables categóricas antes de emplearlo?

**Respuesta**:

A diferencia del modelo Catboost o LightGBM, el algoritmo XGBoost no puede trabajar con variables categórcias directamente, sino que solo aceptan inputs numéricos.

En ese sentido, es necesario realizar algún tipo de transformación o “encoding” de dichas variables categóricas antes de ser incorporadas en el modelo. Por ejemplo, se puede comenzar realizando un one-hot encoding o un label encoding a las var. categóricas, y posteriormente ir probando con formas de encoding más inteligentes (i.e que aporten mayor poder predictivo) como “mean target” encoding, por ejemplo.

* Lo primero seria revisar la presencia de valores nulos , lo cual puede tomar estos 3 casos:

|  |  |
| --- | --- |
| Caso de Valores Nulos | Alternativa |
| Poco porcentaje de nulos. | Imputacion por la moda. |
| Porcentaje de nulos considerable, considerando comportamiento con otros features y con el mismo target. | Creacion de una nueva categoria asociada a los nulos, e inclusive posibilidad de crear 2 o mas en analisis al comportamiento mencionado en el caso. |
| Poco porcentaje de nulos que a su vez coninciden a nivel fila con nulos en varias columnas. | Posibilidad de eliminar la fila tras encontrarse que la conincidencia con otros nulos supera la mitad de cantidad de columnas. Por motivos de no aportar a la prediccion. |
|  |  |

* Una vez limpio los nulos, se procede a analizar la mejor manera de que mis datos categoricos aporten mas a la prediccion:

|  |  |
| --- | --- |
| Caso de mejor aportacion | Alternativa |
| Cantidad de categorias no muy elevada | Separar la columna categorica por columnas dummies, analizando correlacion de Kendall entre todas las columnas para descartar las mas correlacionada. |
| Cantidad de categorias muy elevada | En caso de un target categorico, optar por reemplazar en la misma columna las categorias por su ratio de conversion en referencia al target. |
|  |  |

#Pregunta 6

|  |
| --- |
| **A continuación, se muestran los resultados de una campaña telefónica dirigida a 2** |
| **segmentos de clientes diferentes. ¿Explique para Ud. cuál de las 2 campañas fue más** |
| **exitosa? ¿por qué?** |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Por la cantidad de llamadas respondidas el segmento A está enfocado en clientes de un nivel de ingreso alto. | | | | | | | |  |  |  |  |  |  |  |  |
| Mientras que el segmento B está enfocado en clientes de nivel intermedio/bajo | | | | | | | |  |  |  |  |  |  |  |  |
|  | | | | | | | |  |  |  |  |  |  |  |  |
| **Llamadas respondidas** | | | | | | | |  |  |  |  |  |  |  |  |
|  | | | | | | | | **Seg A** | **Seg B** | **Total** |  |  |  |  |  |
|  | | | | | | | | 11000 | 8000 | 19000 |  |  |  |  |  |
| **Llamadas respondidas por nivel de ingreso** | | | | | | | |  |  |  |  |  |  |  |  |
|  | | | | | | | | **Seg A** | **Seg B** | **Total** |  |  |  |  |  |
| Alto | | | | | | | | 8750 | 2500 | 11250 |  |  |  |  |  |
| Medio | | | | | | | | 1750 | 4000 | 5750 |  |  |  |  |  |
| Bajo | | | | | | | | 500 | 1500 | 2000 |  |  |  |  |  |
|  | | | | | | | | **11000** | **8000** | **19000** |  |  |  |  |  |
| **Aceotacion de tarjetas por nivel de ingreso** | | | | | | | |  |  |  |  |  |  |  |  |
|  | | | | | | | | **Seg A** | **Seg B** | **Total** |  |  |  |  |  |
| Alto | | | | | | | | 80% | 31% | 59% |  |  |  |  |  |
| Medio | | | | | | | | 16% | 50% | 30% |  |  |  |  |  |
| Bajo | | | | | | | | 5% | 19% | 11% |  |  |  |  |  |
|  | | | | | | | | **100%** | **100%** | **100%** |  |  |  |  |  |
|  |  |  |  |  |  |  |  |

Tarjetas Aceptadas

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| se podria pensar que el segmento B fue mas exitoso por tener un mayor % de tajetas aceptadas, sin embargo | | | | | | | | |
| al revisar los numeros absolutos podemos ver que las tarjetas aceptadas son muy parecidas para ambos segmentos. | | | | | | | | |
| \*\*\* los totales de tarjetas aceptadas por segmento no cuadran. | | | | |  |  |  |  |
|  | **Seg A** | **Seg B** | **Total** |  |  |  |  |  |
| % | 11% | 14% | 12% |  | Peso por segmento, basado en sus clientes foco | | | |
| # | **1210** | **1120** | **2330** | \*\*\* |  | **Seg A** | **Seg B** |  |
|  |  |  |  |  | Alto | 3 | 1 |  |
|  | **Seg A** | **Seg B** | **Total** |  | Medio | 2 | 3 |  |
| Alto | 9% | 8% |  |  | Bajo | 1 | 2 |  |
| Medio | 18% | 16% |  |  |  |  |  |  |
| Bajo | 26% | 17% |  |  | Calificacion OK | |  |  |
|  | **53%** | **41%** |  |  |  |  |  |  |
|  | **Seg A** | **Seg B** | **Total** |  |  | **Seg A** | **Seg B** | **Total** |
| Alto | 788 | 200 | 988 |  | Alto | 26,250 | 2,500 | 28,750 |
| Medio | 315 | 640 | 955 |  | Medio | 3,500 | 12,000 | 15,500 |
| Bajo | 130 | 255 | 385 |  | Bajo | 500 | 3,000 | 3,500 |
|  | **1,233** | **1,095** | **2,328** | \*\*\* |  | **30,250** | **17,500** | **47,750** |
|  |  |  |  |  |  |  |  |  |
|  |  | **Calculo de alcance de campaña** | | |  |  |  |  |
|  |  |  | **Seg A** | **Seg B** | **Total** |  |  |  |
|  |  | Alto | 2,363 | 200 | 2,563 |  |  |  |
|  |  | Medio | 630 | 1,920 | 2,550 |  |  |  |
|  |  | Bajo | 130 | 510 | 640 |  |  |  |
|  |  |  | **3,123** | **2,630** | **5,753** |  |  |  |
|  |  |  | **Seg A** | **Seg B** | **Total** |  |  |  |
|  |  | Alto | 9% | 8% |  |  |  |  |
|  |  | Medio | 18% | 16% |  |  |  |  |
|  |  | Bajo | 26% | 17% |  |  |  |  |
|  |  | **Alcance** | 10% | 15% |  |  |  |  |

|  |
| --- |
| Conclusión: |
| Teniendo en cuenta el enfoque de cada segmento, podríamos decir que la **campaña B** fue más exitosa en el sentido que |
| se está ofreciendo el producto/campaña al grupo de clientes correctos. |
|  |
| al cumplir con el objetivo de desarrollar los productos/campañas en los clientes que necesitamos podríamos tener un mejor |
| Performance de experiencia con el cliente, el cual a largo plazo se traduce en mayores ingresos. |

#Pregunta 7

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **‘Firstmilk’ es una empresa productora de leche y sus derivados. Sus clientes son** | | | | | | |  |
| **supermercados principalmente. La Tabla 1 muestra los ingresos generados en el último año** | | | | | | |  |
| **y el crecimiento promedio obtenido en los últimos 5 años para las categorías de productos.** | | | | | | |  |
|  |  |  |  |  |  |  |  |
|  |  |  | **Ingresos** | **Crecimiento Prom** | **Ingresos** |  |  |
|  | **expresado en MM** | | **Presente Año** | **5 Ult años** | **5 Years ago** |  |  |
|  | Leche evaporada | | S/ 370.00 | 4.20% | 355.09 |  |  |
|  | Leche en polvo | | S/ 65.00 | 1.50% | 64.04 |  |  |
|  | Yogurt |  | S/ 50.00 | 1.20% | 49.41 |  |  |
|  | Quesos |  | S/ 12.00 | -7% | 12.90 |  |  |
|  |  |  |  |  |  |  |  |
| ‘Firstmilk’ ha sido adquirida recientemente por ‘Dairy Products Co.’ con el objetivo de impulsar | | | | | | | |
| las ventas. El gerente de ‘Dairy Products Co’. ha trazado como meta que al cabo de 5 años | | | | | | |  |
| quiere ver un crecimiento de por lo menos el 10% en las ventas de los 2 años últimos. En la | | | | | | |  |
| Gráfica 1 se muestran 4 posibles escenarios de crecimiento para ‘Firstmilk’ con el año 0 como | | | | | | |  |
| el año actual de partida. | |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  | **Escenarios** | **0** | **1** | **2** | **3** | **4** | **5** |
|  | A | 100% | 111% | 119% | 129% | 138% | 148% |
|  | B | 100% | 102% | 101% | 104% | 117% | 130% |
|  | C | 100% | 101% | 102% | 105% | 120% | 129% |
|  | D | 100% | 101% | 102% | 114% | 125% | 129% |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  | **Escenarios** | **0** | **1** | **2** | **3** | **4** | **5** |
|  | A | S/ 370.00 | S/ 410.70 | S/ 440.30 | S/ 477.30 | S/ 510.60 | S/ 547.60 |
|  | B | S/ 65.00 | S/ 66.30 | S/ 65.65 | S/ 67.60 | S/ 76.05 | S/ 84.50 |
|  | C | S/ 50.00 | S/ 50.50 | S/ 51.00 | S/ 52.50 | S/ 60.00 | S/ 64.50 |
|  | D | S/ 12.00 | S/ 12.12 | S/ 12.24 | S/ 13.68 | S/ 15.00 | S/ 15.48 |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
| **a) ¿cuál de los escenarios cumple los requerimientos de ‘Dairy Products Co.’? Desarrolle** | | | | | | |  |
| **su respuesta.** |  |  |  |  |  |  |  |
| **Escenarios** | **0** | **1** | **2** | **3** | **4** | **5** |  |
| A |  | 11% | 7% | 8% | 7% | 7% |  |
| B |  | 2% | -1% | 3% | 13% | 11% |  |
| C |  | 1% | 1% | 3% | 14% | 8% |  |
| D |  | 1% | 1% | 12% | 10% | 3% |  |
|  |  |  |  |  |  |  |  |
| como se puede aprenciar en el grafico los escenarios muestran un creciento mayor de 10% | | | | | | |  |
|  |  |  |  |  |  |  |  |
| **b) ¿cuál es el nivel de ventas que tenía la categoría ‘Quesos’ hace 5 años? Desarrolle su** | | | | | | |  |
| **respuesta.** |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  | **Ingresos** | **Crecimiento Prom** | **Ingresos** |  |  |  |
| **expresado en MM** | | **Presente Año** | **5 Ult años** | **5 Years ago** |  |  |  |
| Quesos |  | S/ 12.00 | -7% | 12.90 |  |  |  |
|  |  |  |  |  |  |  |  |
| **c) ¿cuál es el valor estimado aproximado de las ventas en el Año 3 bajo el escenario C?** | | | | | | |  |
| **Desarrolle su respuesta.** | |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
| **Escenarios** | **0** | **1** | **2** | **3** | **4** | **5** |  |
| C | S/ 50.00 | S/ 50.50 | S/ 51.00 | S/ 52.50 | S/ 60.00 | S/ 64.50 |  |
|  |  |  |  |  |  |  |  |
| El gerente de ‘Dairy Products Co.’ quiere probar con un piloto en una ciudad. El piloto consiste | | | | | | |  |
| en reducir el precio de venta al consumidor final en 6% acompañando la campaña con una | | | | | | |  |
| estrategia de marketing en diversos medios. El gerente de ventas de ‘First Milk’ le da la | | | | | | |  |
| siguiente información: | |  |  |  |  |  |  |
| • La campaña de marketing tendrá un costo de S/.2.1 millones. | | | | |  |  |  |
| • ‘First Milk’ tiene un mercado de 1 millón de compradores quienes compran una caja | | | | | | |  |
| de leche evaporada por mes. | | |  |  |  |  |  |
| • El precio de venta a los supermercados es de S/.10. | | | |  |  |  |  |
| • El precio de venta al consumidor final es de S/.11. | | | |  |  |  |  |
| • ‘First Milk’ obtiene un 20% de margen de utilidad en el producto de leche evaporada. | | | | | | |  |
| • La campaña no deberá afectar el margen de ganancia que tienen los supermercados | | | | | | |  |
| actualmente. |  |  |  |  |  |  |  |
| **d) ¿cuál es el monto de utilidad promedio por caja de leche evaporada que obtiene ‘First** | | | | | | |  |
| **Milk’ antes de la campaña? Detalle su respuesta.** | | | |  |  |  |  |
|  |  |  |  |  |  |  |  |
| el margen es de 2 soles por caja | | |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
| Costo Campaña | 2.1 | MM |  |  |  |  |  |
| Mercado | 1 | MM |  |  |  |  |  |
| cj x año | 12 |  |  |  |  |  |  |
| PV SM | 10 |  |  |  |  |  |  |
| PV Cons | 11 |  |  |  |  |  |  |
| Mg | 20% |  |  |  |  |  |  |
| Mg Abs | 2 |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
| ‘Dairy Products Co.’ requiere que toda inversión se pague al primer año de implementada. | | | | | | |  |
| e) ¿En qué porcentaje se deben incrementar las ventas para cumplir con el | | | | | |  |  |
| requerimiento? Detalle su respuesta. | | |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
|  |  | **sin campaña** | **con campaña** | |  |  |  |
| Costo Supermercado | | S/ 10.00 | S/ 10.00 |  |  |  |  |
| PV Consumidor |  | S/ 11.00 | S/ 10.34 |  |  |  |  |
| Reduccion de precio | | 0% | 6% |  |  |  |  |
| Mg Supermercado | | S/ 1.00 | S/ 0.34 |  |  |  |  |
| % Mg Supermercado | | 9% | 3% |  |  |  |  |
| Crecimiento necesario | |  | 294% |  |  |  |  |
|  |  |  | S/ 1.00 |  |  |  |  |

#Pregunta 8

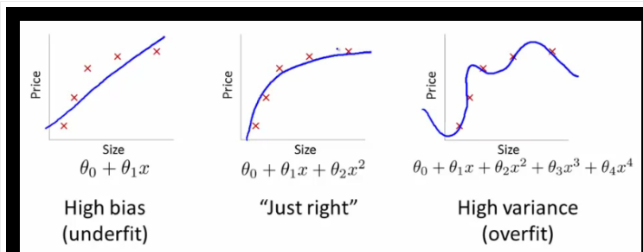
Explique la disyuntiva entre sesgo y varianza al evaluar los resultados de un modelo

Predictivo y por qué se da.

El trade-off entre sesgo y varianza es inherente a todo algoritmo de aprendizaje supervisado de Machine Learning.

El aumento en la complejidad del modelo:

Para obtener un buen resultado en un modelo, buscamos que la diferencia entre los datos reales y la predicción del modelo sea el menor posible, es decir, buscamos minimizar el error de predicción fuera de la muestra. A esta diferencia entre el valor real y el valor predicho se le conoce como Sesgo.

Para entender el trade off entre sesgo y varianza veamos el siguiente ejemplo

En el gráfico anterior, tenemos 3 modelos alternativos. El modelo del extremo izquierdo se trata de una Regresión Lineal o también puede verse como un polinomio de grado 1. Vemos que este es un modelo simple, sin embargo tiene un sesgo grande, pues no logra ajustarse muy bien a los datos (puntos rojos).

Para reducir el sesgo del modelo, podemos aumentar la complejidad del mismo, en el caso de una regresión polinomial esto se traduce en aumentar el polinomio a uno de grado 2. Podemos ver que este modelo se ajusta más a los datos, es decir tiene un menor error de predicción.

Podemos aumentar el grado del polinomio aún más, aumentando a 4 el grado del polinomio (gráfico del extremo derecho). Este nuevo modelo tendrá poco sesgo o un error de predicción dentro de la muestra (in sample) muy bajo, sin embargo, tendrá una varianza muy grande y no será capaz de generalizar a datos fuera de la muestra, produciéndose así el fenómeno conocido como Overfitting.

Generalización del Modelo:

Cuando se aumenta la complejidad del modelo este se vuelve sensible a pequeñas variaciones en los datos de entrada, es ahí cuando la varianza aumenta.

En conclusión, es necesario encontrar un balance entre sesgo y varianza, o visto de otra forma, entre error y complejidad; ya que este balance nos permitirá aplicar el modelo a diversos datos y que el performance no se vea muy alterado.

# Pregunta 9

9. La gerencia de producto le solicita realizar un modelo predictivo para retención de sus

Clientes ya que necesita identificar qué clientes son los más propensos a dejar el servicio.

Explique qué ventajas encuentra en utilizar una regresión logística en lugar de bosques

Aleatorios para dicha tarea.

Comenzaremos deja en claro:

* Se tiene un target categórico dicotómico y la posibilidad de tener variables categóricas y continuas. Por lo tanto estamos hablando de un problema de clasificación.
* Independientemente del modelo a elegir, la mejora de la predicción se basa principalmente en el  reprocesamiento y el manejo de los variables de la mano con la lógica de mercado.
* Denotaremos a la regresión logística como LR y a los bosques aleatorios como BA

Una vez en claro los puntos expuestos,  de cara al modelo, cabe resaltar que una ventaja a primera mano de la RL es su poca complejidad y su  aceptable robustez, siendo muy similar a la acción de un Perceptron de un sola capa. Mientras que un bosque necesita calcular la  predicción de todos sus árboles para luego calcular la suya en base a todos ellos. (MENOS COMPLEO).

**La principal ventaja es la fácil intepretabilidad del modelo,**  que nos da en base a sus coeficientes, y los exponenciales de los mismos, con los cuales se puede determinar la relación directa o indirecta y la propensión de cada variable (feature) con respecto al target. Es decir, es con una Regresión Logística podemos explicar de manera sencilla a las áreas no técnicas de la empresa, cuáles son las variables que impactan más en la fuga de clientes y gracias a su coeficiente podemos incluso saber cuánto eleva/disminuye la probabilidad de fuga una variable para un individuo dado. Mientras que en los BA solo se puede lograr tener un aproximado de árbol de reglas de su mejor árbol, el cual fue seleccionado tomando una  muestra aleatoria de nuestra data (MAS CONTROL). Si bien es cierto, un Bosque Aleatorio también nos permite obtener un ranking de Importancia de Variables (Feature Importance) de las variables que permitieron hacer un mayor número de “splits” en los árboles, sin embargo, esto no es directamente interpretable como el caso de los coeficientes de una regresión logística.

# Pregunta 10

Ud. ha elaborado un modelo de clasificación y ha conseguido un alto valor de AUC con un

algoritmo de machine learning. Cuando le piden la probabilidad de aceptación de un grupo de clientes encuentra que las probabilidades que obtiene con el modelo difieren de las probabilidades reales. ¿Qué pasos y técnicas emplearía para calibrar las probabilidades obtenidas?

La respuesta a esta pregunta es depende de la causa que está detrás del problema de que tengamos un alto valor de AUC en la etapa de entrenamiento, pero nuestras probabilidades predichas para datos no vistos sean diferentes de las probabilidades reales.

Esto se puede deber esencialmente a 2 razones: (1) clases no balanceadas, (2) Overfitting ocasionado por i) mal corte entre Train y Test Set, ii) mala optimización de parámetros del algoritmo, iii) otros.

Si lo que causa el problema es (1), se debe tratar el problema de clases no balanceadas con diferentes métodos como oversampling, undersampling, smote, entre otros, antes de entrenar el modelo. Si la causa del problema se debe a (2), se tiene que volver a optimizar los hiperparámetrps del modelo a fin de reducir la complejidad del mismo para evitar incurrir en overfitting y que el modelo no sea capaz de generalizar a datos no vistos durante el entreanmiento.

# Pregunta 11.

Los algoritmos de boosting siguen un procedimiento secuencial donde los errores de una

clasificación previa sirven de input a la siguiente. Explique cómo es posible que algoritmos de boosting como el XGBoost sean “paralelizables”.

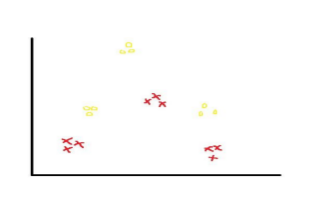
En términos generales, existen dos formas de paralizar un algoritmo de boosting como XgBoost o Gradient Boosting. La “paralelización” de un algoritmo de boosting es de alguna manera similar a la paralización de un árbol de decisión. En ese sentido, a continuación describimos las dos formas de paralelizar un árbol de decisión:

1. Hacer la “paralelización” a nivel de los nodos: Se pueden construir los nodos de un árbol en paralelo
2. Hacer la “paralelización” a nivel del Feature Split: Se calcula y se encuentra el mejor feature split (punto de corte) en paralelo

# Pregunta 12.

Calcule y explique el número de capas y el número de neuronas que usaría para resolver el

siguiente problema de clasificación binaría, utilizando un ANN (artificial neural network) del tipo MP (multiple perceptron).



Al tener un target binario , el numero de neuronas en la capa de salida es solo una,

Suponiendo 6 neuronas en la capa de entrada ya aplicando el principio del triangulo geometrico optariamos por una sola capa oculta con 3 neuronas.

Total de neuronas = 6 + 3 + 1 = 10 neuronas

En 3 capas