EGADE Business School
Tecnológico de Monterrey

Aplicaciones de analítica de datos a los negocios II

PROF: JUAN C. BUSTAMANTE

JUCBUSTAM@TEC.MX

Normas para la conexión síncrona:

1. La clases tiene un back-up garantizado (Grabación disponible en la nube de Zoom).
2. Ingresar a la clase con la cámara del equipo de computo encendida.
3. La **cámara deberá permanecer encendida a lo largo de la clase**.
4. Al ingresar a la clase deben silenciar el micrófono del equipo de computo.
5. Levantar la mano es una opción cuando se quiere preguntar algo durante la sesión de clase, pero les recomiendo que mejor hagamos uso intensivo del chat del canal general para hacer preguntas**.**
6. En caso de necesitar hacer una pregunta, puede interrumpir la clase sin problema, activando el micrófono de vuestro equipo de computo, luego de la pregunta desactívelo nuevamente**.**
7. Para una buena clase online es indispensable **el debate, así que foméntelo!!!**.
8. Toda la información se gestiona en CANVAS LMS.

EGADE Business School
Tecnológico de Monterrey

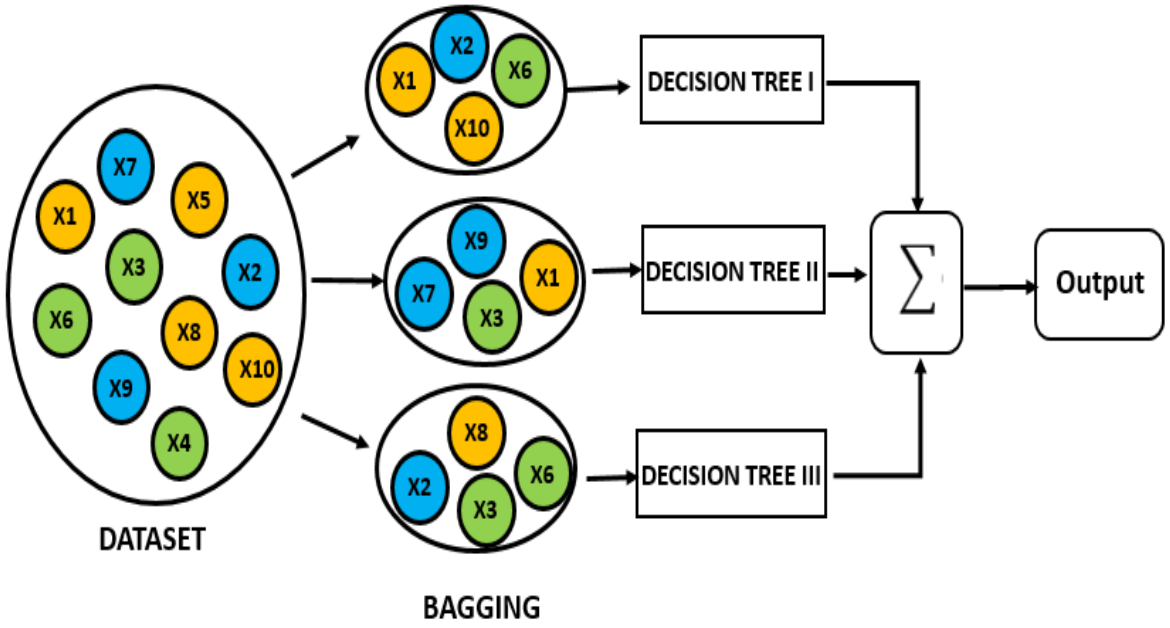
Cronograma de trabajo:

Sesiones	Contenidos	Actividad		Fecha
1	Información general del curso	Utility of classification algorithms		Martes 18/04
2	Algoritmo de regresión logística	Ejecutar script	**Solución caso:** Retention modelling at Scholastic Travel Company (A) and (B)	Martes 25/04
3	Algoritmo Naïve Bayes	Ejecutar script		Martes 02/05
4	Algoritmo k-nearest-neighbors (KNN)	Ejecutar script		Martes 09/05
5	Algoritmo Support vector machine	Ejecutar script		Martes 16/05
6	Algoritmo Decision Trees	Ejecutar script		Martes 23/05
7	Algoritmo Random Forest	Ejecutar script		Martes 30/05
8	Modelo RFM	Ejecutar script	**Solución caso**: CD Now	Martes 06/06
9	Modelo valor de vida del cliente (I)	Ejecutar script		Martes 13/06
10	Modelo valor de vida del cliente (II)	Ejecutar script		Martes 20/06
11	Análisis de series de tiempo	Ejecutar script		Martes 27/06
12	**Proyecto final**	Presentación en equipos		Martes 04/07
	Evaluación final			

1

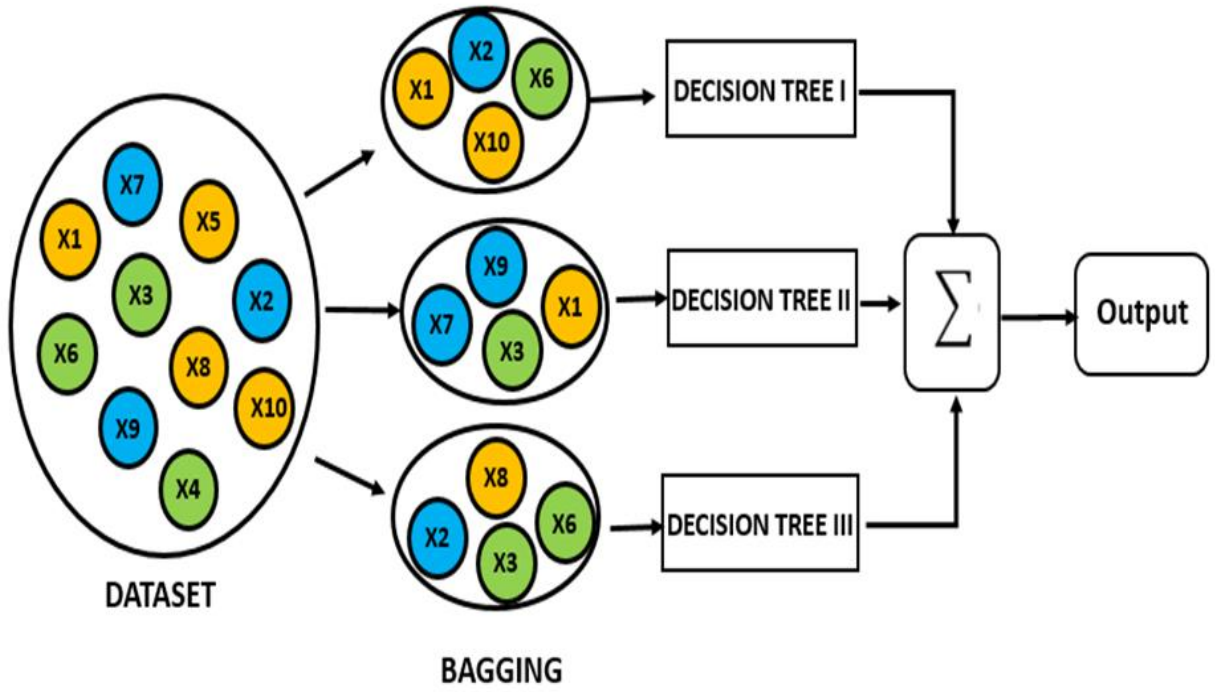
〈 Random forest

RANDOM FOREST



Random Forest?

Un Random Forest es un conjunto (ensemble) de árboles de decisión combinados con bagging. Al usar bagging, lo que en realidad está pasando, es que distintos árboles ven distintas porciones de los datos. Ningún árbol ve todos los datos de entrenamiento. Esto hace que cada árbol se entrene con distintas muestras de datos para un mismo problema. De esta forma, al combinar sus resultados, unos errores se compensan con otros y tenemos una predicción que generaliza mejor.



Proceso iterativo de un modelo Random Forest

- Dado que el número de casos en el conjunto de entrenamiento es N. Una muestra de esos N casos se toma aleatoriamente, pero **CON REEMPLAZO**. Esta muestra será el conjunto de entrenamiento para construir el árbol i.
- Si existen M varibles de entrada, un número m<M se especifica tal que para cada nodo, m variables se seleccionan aleatoriamente de M. La mejor división de estos m atributos es usado para ramificar el árbol. El valor m se mantiene constante durante la generación de todo el bosque.
- Cada árbol crece hasta su máxima extensión posible y **NO hay proceso de poda**.

RANDOM FOREST (HYPERPARÁMETROS)

Los Hyperparámetros esenciales son:

- ntree: número de árboles en el bosque
- nVar: número de variables candidatas a seleccionar

Otros Hyperparámetros:

- mtry: número de variables aleatorias como candidatas a cada ramificación
- sampsize: número de muestras sobre las cuales entrenar.
- Nodesize: mínimo número de muestras dentro de los nodos terminales

Random Forest

VENTAJAS

- Existen muy pocas suposiciones y por lo tanto la preparación de los datos es mínima.
- Puede resolver ambos tipos de problemas, es decir, clasificación y regresión, y realiza una estimación decente en ambos casos.
- Puede manejar hasta miles de variables de entrada e identificar las más significativas. Método de reducción de dimensionalidad.
- Una de las salidas del modelo es la importancia de variables.
- Es posible usarlo como método no supervisado (clustering) y detección de outliers.

DESVENTAJAS

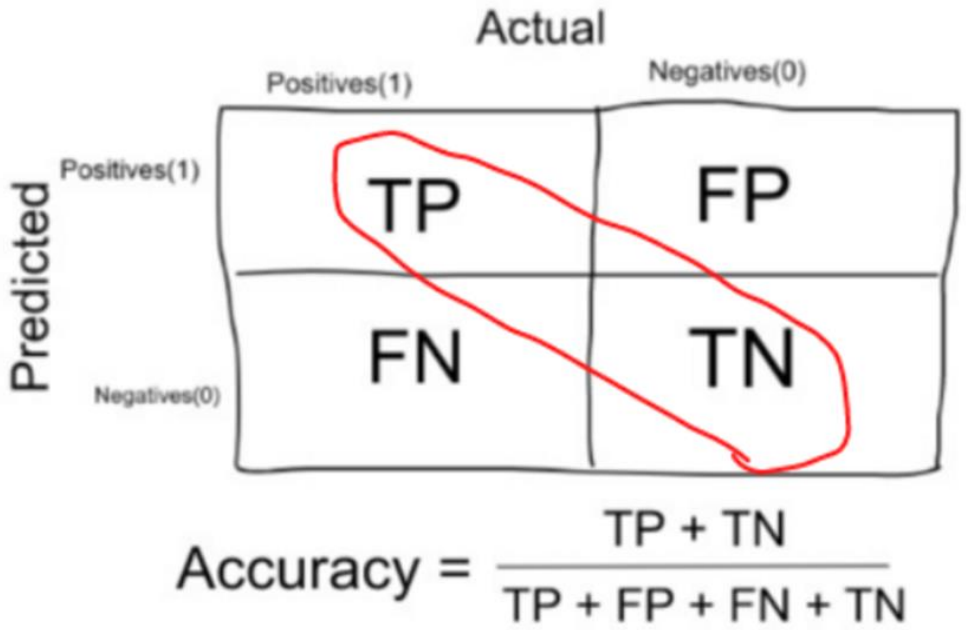
- Pérdida de interpretación.
- Bueno para clasificación, no tanto para regresión. Las predicciones no son de naturaleza continua.
- En regresión, no puede predecir más allá del rango de valores del conjunto de entrenamiento.
- Poco control en lo que hace el modelo (modelo caja negra para modeladores estadísticos).

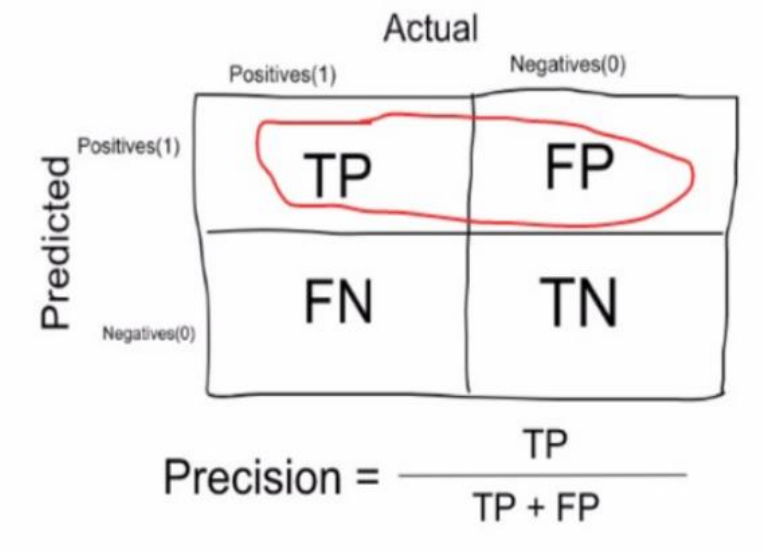
Aplicaciones del algoritmo Random Forest

- ✓ **Finanzas**: Se puede utilizar para evaluar clientes con alto riesgo crediticio, para detectar fraudes y problemas de opciones de precios.
- ✓ **Comercio electrónico**: se puede utilizar para motores de recomendación con fines de venta cruzada.
- ✓ **Cuidado de la salud**: le permite a los médicos abordar problemas como la clasificación de la expresión génica, el descubrimiento de biomarcadores y la anotación de secuencias. Como resultado, los médicos pueden hacer estimaciones sobre las respuestas de los medicamentos a medicamentos específicos.

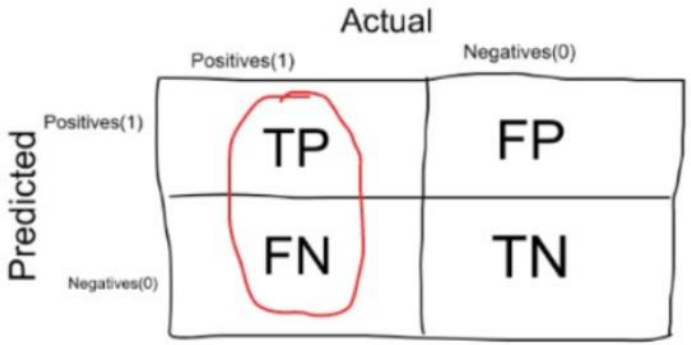
Métricas

1. **True Positives (TP):** cuando la clase real del punto de datos era 1 (Verdadero) y la predicha es también 1 (Verdadero)
2. **Verdaderos Negativos (TN):** cuando la clase real del punto de datos fue 0 (Falso) y el pronosticado también es 0 (Falso).
3. **False Positives (FP):** cuando la clase real del punto de datos era 0 (False) y el pronosticado es 1 (True).
4. **False Negatives (FN):** Cuando la clase real del punto de datos era 1 (Verdadero) y el valor predicho es 0 (Falso).

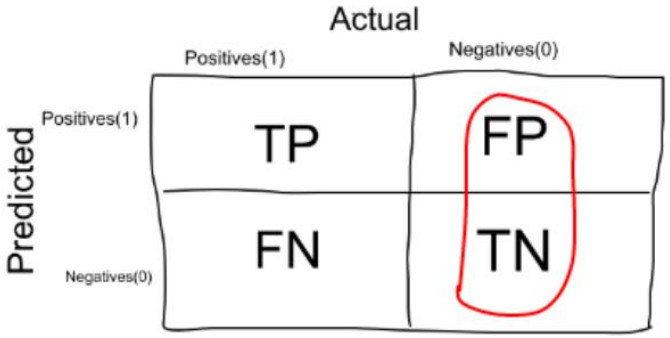




$$\text{Precision} = \frac{TP}{TP + FP}$$



$$\text{Recall} = \frac{TP}{TP + FN}$$



$$\text{Specificity} = \frac{TN}{TN + FP}$$

Métricas

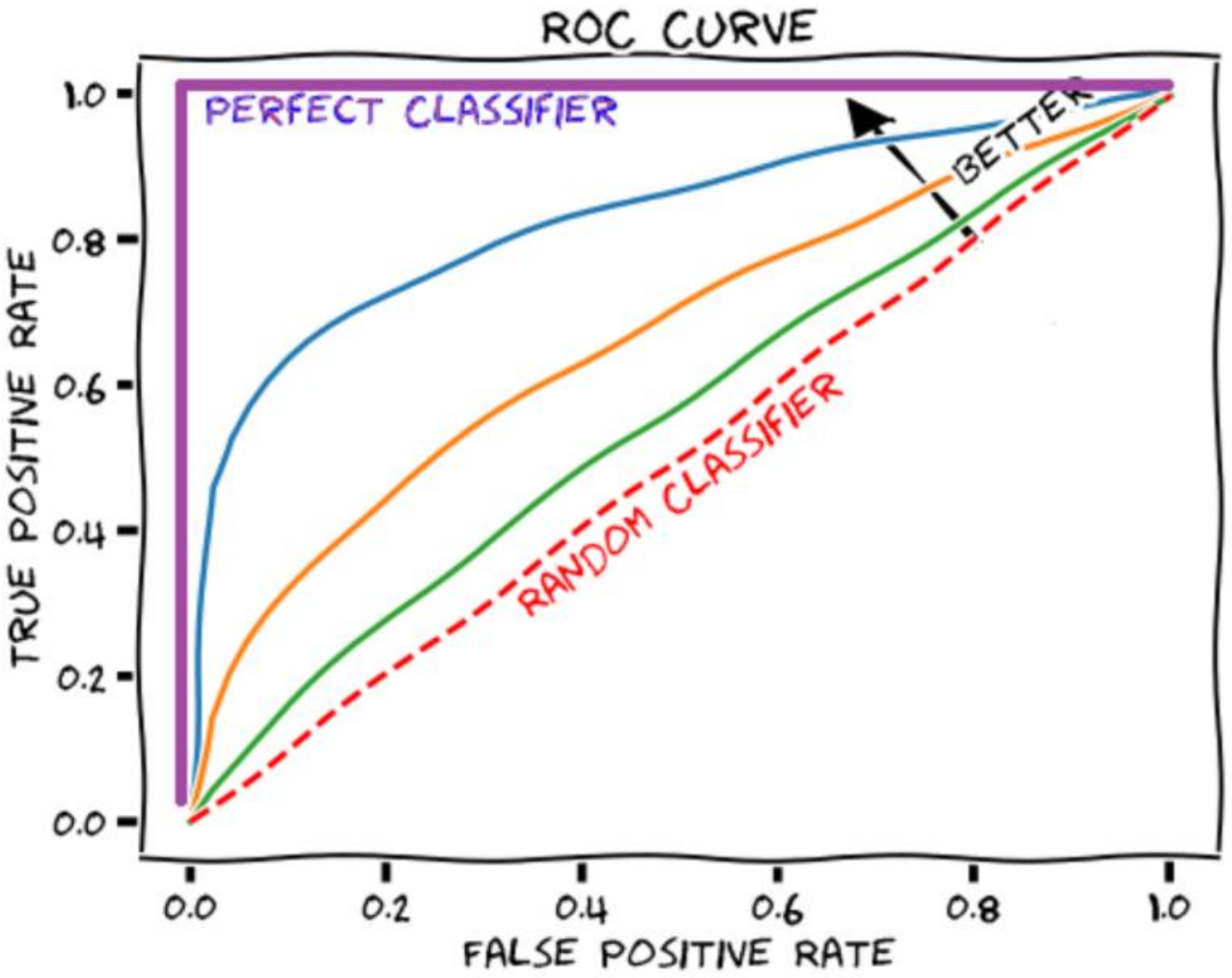
Métricas

Curva ROC

ROC es un acrónimo para Receiver Operating Characteristic (Característica Operativa del Receptor). Es una gráfica que enfrenta el ratio de falsos positivos (eje x) con el ratio de falsos negativos (eje y).

La curva ROC es útil por dos principales motivos:

- Permite comparar diferentes modelos para identificar cual otorga mejor rendimiento como clasificador
- El área debajo de la curva (AUC) puede ser utilizado como resumen de la calidad del modelo



Bibliotecas

Manipulación y análisis de datos

Creación de vectores y matrices
Colección de funciones matemáticas

Generación de gráficos

Biblioteca de visualización basada en Matplotlib

Biblioteca de aprendizaje automático