ANÁLISIS DE DATOS

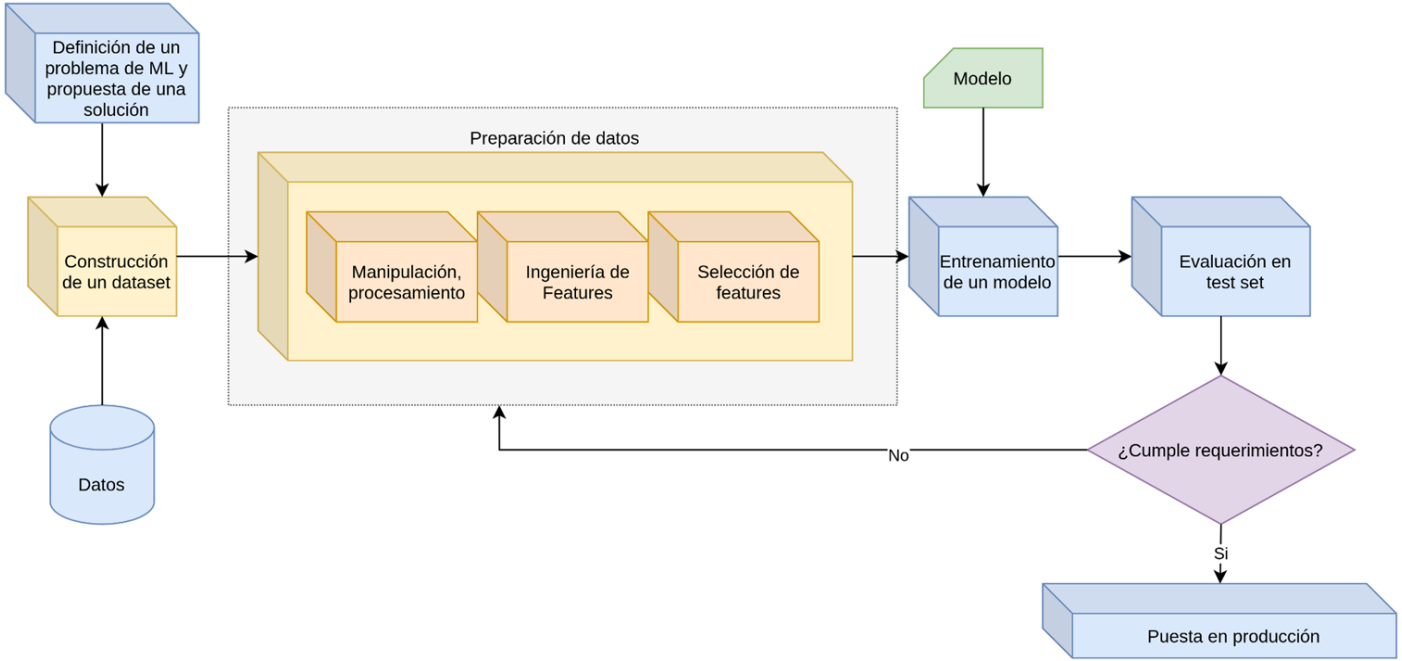
Clase 3
Taller de preparación de datos

Temario

1. Introducción a la preparación de datos.
2. Caracterización de las variables.
3. Imputación de datos faltantes.
4. Desbalance de clases.

1. INTRODUCCIÓN A LA PREPARACIÓN DE DATOS

Flujo de trabajo de un problema de ML
Tareas de preparación de datos



Motivación

- El primer objetivo de la preparación de datos es generar un dataset apto para entrenar un modelo de aprendizaje automático.
- El segundo objetivo es optimizar su representación para mejorar el desempeño de un modelo.

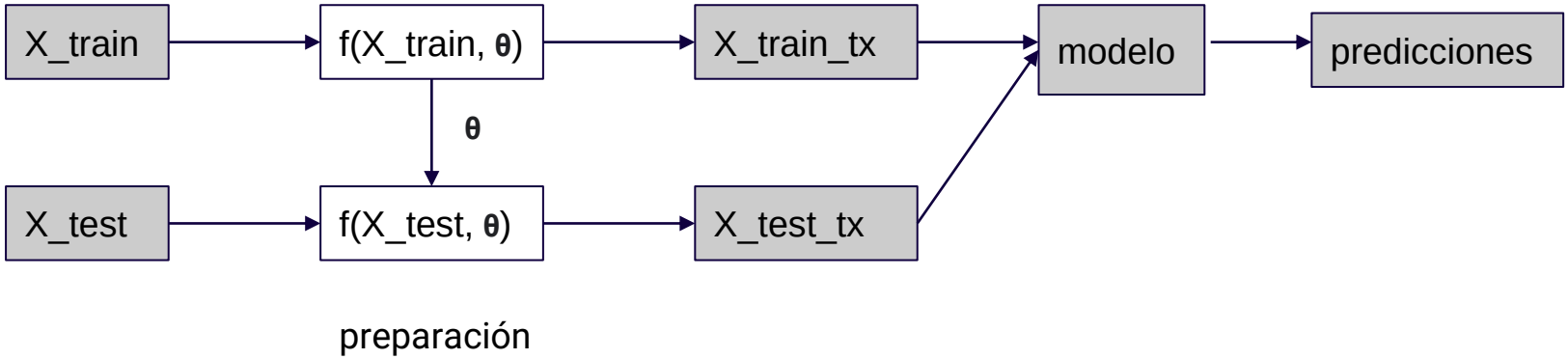
Tareas de preparación de datos

Generalmente, implica realizar una o más de las siguientes tareas:

- **Limpieza de datos**: identificar y corregir errores en los datos.
- **Transformación de datos**: modificar la escala o distribución de los valores.
- **Ingeniería de features**: utilizar información del dominio del problema para seleccionar o generar variables relevantes.
- **Selección de features**: identificar las variables de entrada de mayor relevancia para la tarea.
- **Reducción de dimensiones**: crear proyecciones compactas de los datos.

Reproducibilidad de las tareas de preparación

Los procesos de transformación que se apliquen durante el entrenamiento (train set) deben ser reproducibles para datos no vistos (test set).

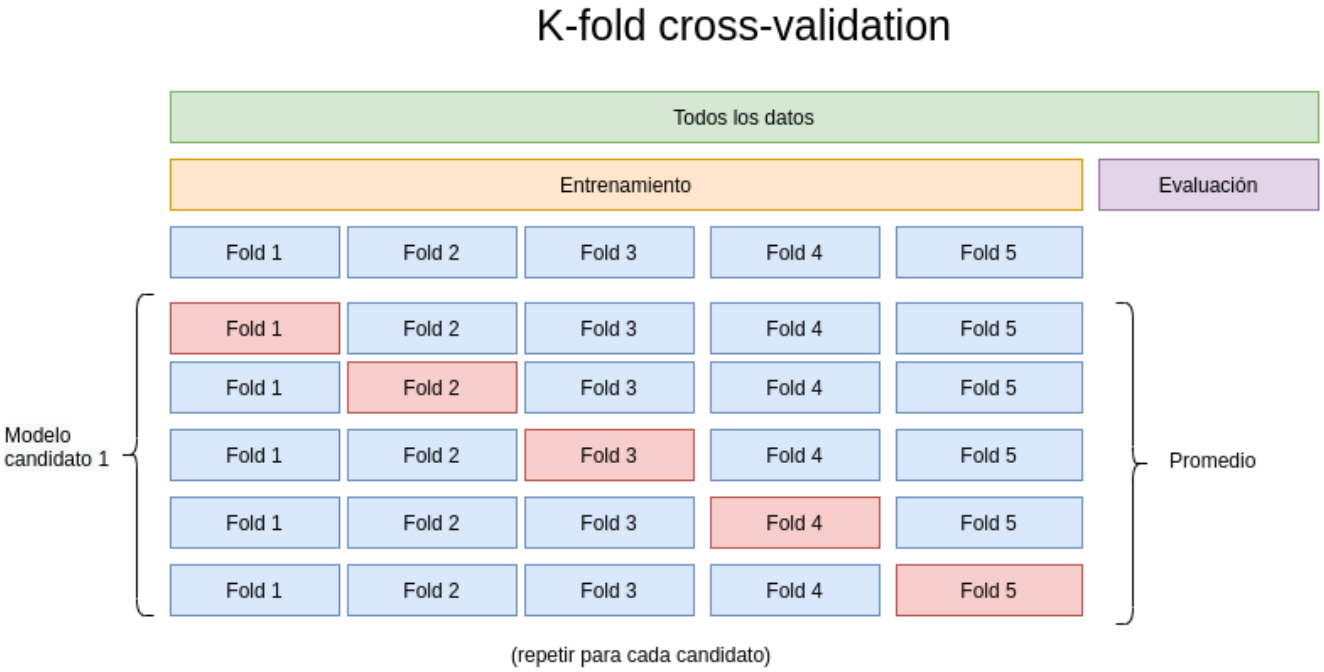


Preparación de datos y validación de resultados

- Al preparar datos para el entrenamiento de modelos es importante **calcular los parámetros de las transformaciones solo sobre el set de entrenamiento** y **nunca** sobre los datos de evaluación.
- En SKLearn típicamente utilizaremos el siguiente flujo de trabajo:
  - Durante el entrenamiento/validación cruzada:
    - Calcular parámetros óptimos en train set con ***fit()/fit_transform()***.
    - Exportar estos parámetros (por ejemplo, con *pickle*).
  - Durante la evaluación/producción (datos no vistos):
    - Importar los parámetros y aplicarlos antes de realizar inferencias.
    - En algunos casos, aplicar una transformación inversa con ***inverse_transform()*** sobre el resultado de la inferencia.

Preparación de datos y validación de resultados
Validación cruzada

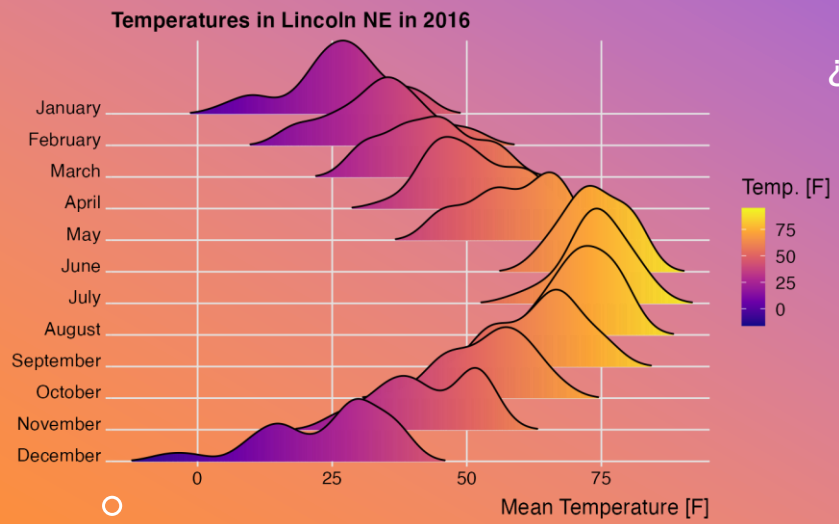
Preparación de datos y validación de resultados.
Cross validation



EJEMPLOS EN JUPYTER

Clase 3.1 - Preparación de datos - Esquemas de validación.ipynb

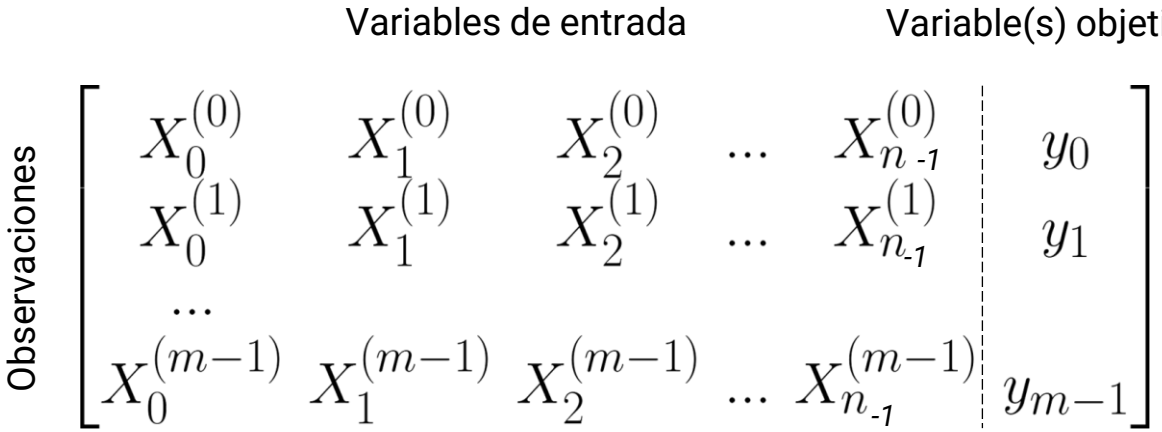
CARACTERIZACIÓN DE LAS VARIABLES



¿Qué son las variables de nuestro Dataset?

Dataset (para un problema de aprendizaje supervisado)

- Llamamos ***dataset*** a un conjunto de m observaciones, cada una de ellas conformada por n variables de entrada y, en el caso de un problema de aprendizaje supervisado, r variables objetivo (target). Para los ejemplos de esta clase r=1.



¿Qué es una variable?

Una **variable** es una característica que puede fluctuar y cuya variación es susceptible a adoptar diferentes valores, los cuales pueden medirse u observarse.

Ejemplos de variables

- Edad (18, 23,70, …)
- Género (masculino, femenino, …)
- Ingreso ($40.000, $50.0000, …)
- País de nacimiento (Argentina, Perú, México)
- Color de ojos (marrón, verde)
- Vehículo (Ford, Volkswagen)

Tipos de variables

Numéricas

- Discretas
- Continuas

Categóricas

- Nominales
- Ordinales

Fecha/hora

Compuestas

¿Qué aspectos de las variables es importante considerar para un modelo de aprendizaje supervisado?

Datos faltantes

Variables categóricas. Cardinalidad y etiquetas raras

Cumplimiento de los supuestos de linealidad

Distribución de los valores de las independientes

Valores extremos (outliers)

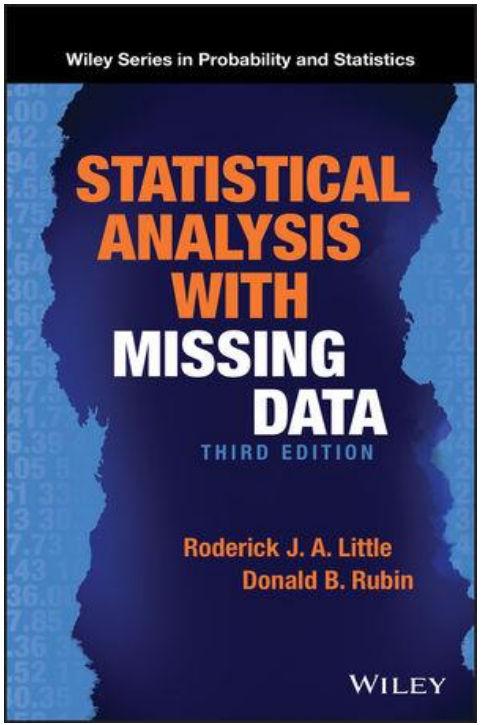
Magnitud/escala

2.1 Datos faltantes: Definición

- Los datos o valores faltantes ocurren cuando una observación está incompleta.
- Tienen un efecto en las conclusiones que puedan establecerse a partir de esos datos.

ID	Color	Weight	Broken	Class
1	Black	80	Yes	1
2	Yellow	100	No	2
3	Yellow	120	Yes	2
4	Blue	90	No	2
5	Blue	85	No	2
6	**?**	60	No	1
7	Yellow	100	**?**	2
8	**?**	40	**?**	1

2.1 Datos faltantes: Clasificación



Entender el motivo de ausencia de los datos permite seleccionar una estrategia adecuada para su tratamiento.

2.1 Datos faltantes: Clasificación

Clasificación de Rubin de valores faltantes (1976):

- **Missing Completely at Random (MCAR)**: significa que la razón por la cual el dato no está es completamente aleatoria, y que probablemente no podamos predecir el valor a partir de otro valor en los datos.

- **Missing at Random (MAR)**: los datos faltantes pueden ser explicados por valores en las otras columnas, pero no por valores de esa columna. Por ejemplo, si cada columna representa una elección excluyente.

- **Missing not at Random (MNAR)**: significa que es probable que la falta de ese dato no sea al azar. En este caso tenemos que investigar la causa por la que falta ese valor.

2.1.1 Missing Data Completely at Random (MCAR)

- La probabilidad de valores faltantes es la misma para todas las observaciones.
- No existe ninguna relación entre los datos faltantes y otros valores observados o faltantes en el dataset.
- Omitir estas observaciones no implicaría un sesgo en las inferencias.

2.1.2 Missing Data at Random (MAR)

- La probabilidad de datos faltantes en una observación no está relacionada con los datos faltantes, pero puede depender de la información que sí está disponible. Es decir, el azar no es la única causa por la que faltan esos datos.
- Ejemplo:
    - En un estudio para un nuevo tratamiento, algunos sujetos se retiran cuando empiezan a tener efectos secundarios.

2.1.3 Missing Data not at Random (MNAR)

- Existe una explicación por la cual hay valores faltantes en un dataset.
- Ejemplos:
    - En una encuesta sobre salarios, los participantes con menor nivel educativo se ven menos inclinados a reportar sus ingresos.
    - En un estudio sobre depresión, los pacientes con depresión son menos propensos a responder algunas preguntas.

¿Qué aspectos de las variables es importante considerar para un modelo de aprendizaje supervisado?

- Datos faltantes
- **Variables categóricas. Cardinalidad y etiquetas raras**
- Cumplimiento de los supuestos de linealidad
- Distribución de los valores de las independientes
- Valores extremos (outliers)
- Magnitud/escala

2.2 Cardinalidad: Definición

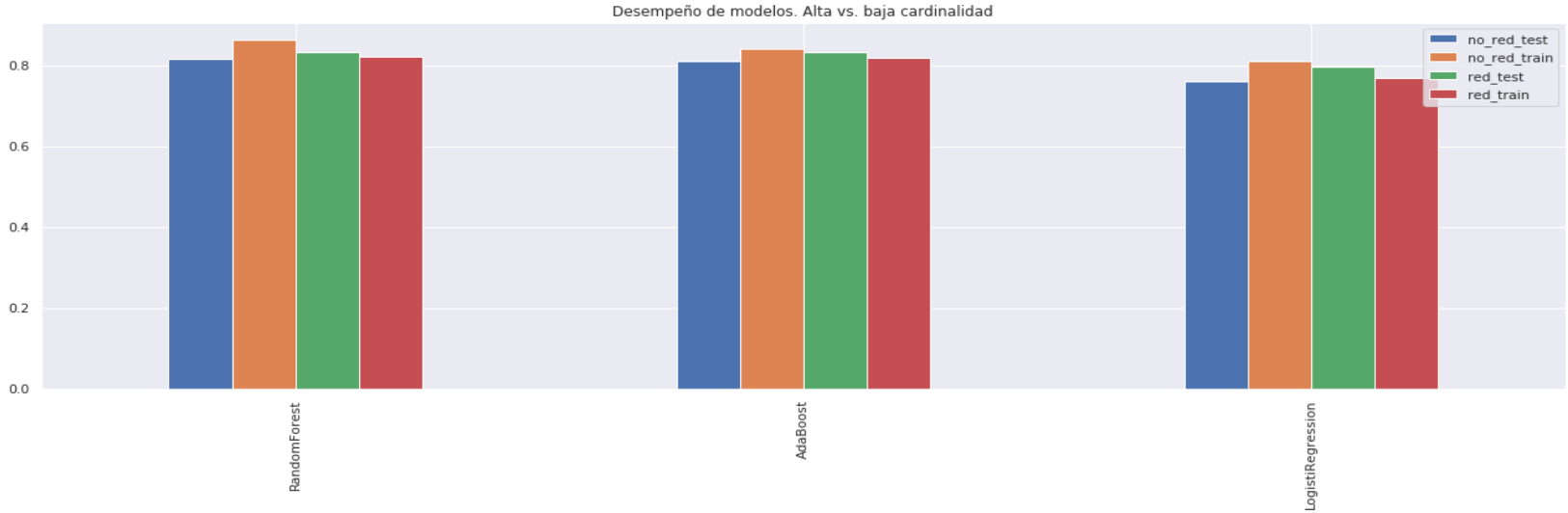
- Los valores que puede tomar una **variable categórica** forman un grupo de **categorías** (también llamadas etiquetas).
- Se denomina **cardinalidad** al número de categorías existentes.

2.2 Cardinalidad: Problemas

- La mayoría de los modelos de aprendizaje automático no aceptan cadenas de texto como entradas, por lo tanto, ***las categorías deben ser codificadas numéricamente***.
- Las técnicas de codificación pueden tener un efecto secundario indeseado, como aumentar la dimensión del espacio de las variables de entrada.
- Pueden ocurrir errores en la partición de entrenamiento y validación:
  - Valores que sólo estén disponibles en train set → overfitting.
  - Valores que sólo estén disponibles en test set → el modelo no podrá interpretarlos.

2.2 Cardinalidad: Overfitting

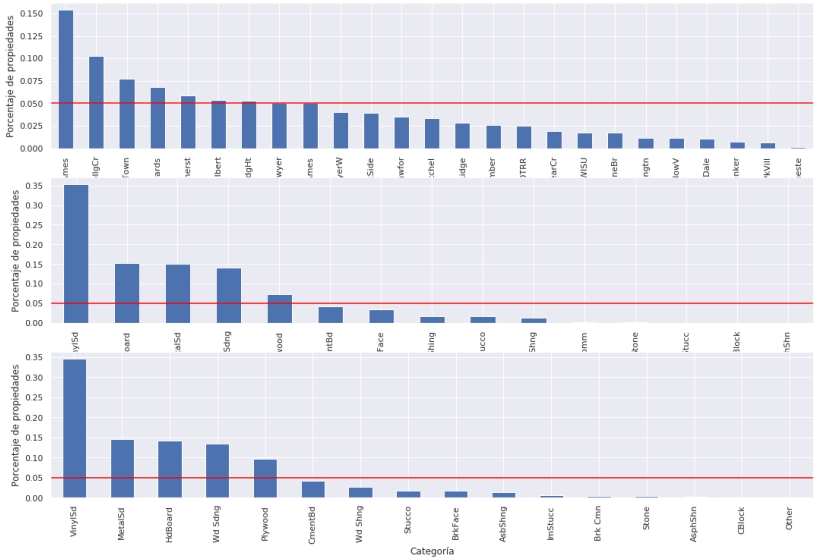
- Las variables con muchas etiquetas dominarán sobre las que tengan menos etiquetas, especialmente en los algoritmos basados en árboles de decisión.
- Un número grande de etiquetas empeora la relación señal/ruido.
- Reducir la cardinalidad puede contribuir a mejorar el desempeño de algunos modelos.



Entrenamiento de modelos con dos versiones del dataset (original vs con reducción de categorías) (ejemplo completo en notebook)

2.2 Etiquetas poco frecuentes

- Son aquellas que aparecen en una pequeña proporción de observaciones de todo el dataset.
- Por ejemplo en un censo nacional, si la variable es “ciudad de residencia”:
  - “Buenos Aires” puede ser un valor muy frecuente
  - “Faro” probablemente será una categoría poco frecuente (tiene ~14 habitantes) (*).



Categorías de propiedades (ejemplo completo en notebook)

(*) https://es.wikipedia.org/wiki/Faro_(Buenos_Aires)

¿Qué aspectos de las variables es importante considerar para un modelo de aprendizaje supervisado?

- Datos faltantes
- Variables categóricas. Cardinalidad y etiquetas raras
- **Cumplimiento de los supuestos de linealidad**
- Distribución de los valores de las independientes
- Valores extremos (outliers)
- Magnitud/escala

2.3 Supuestos de linealidad

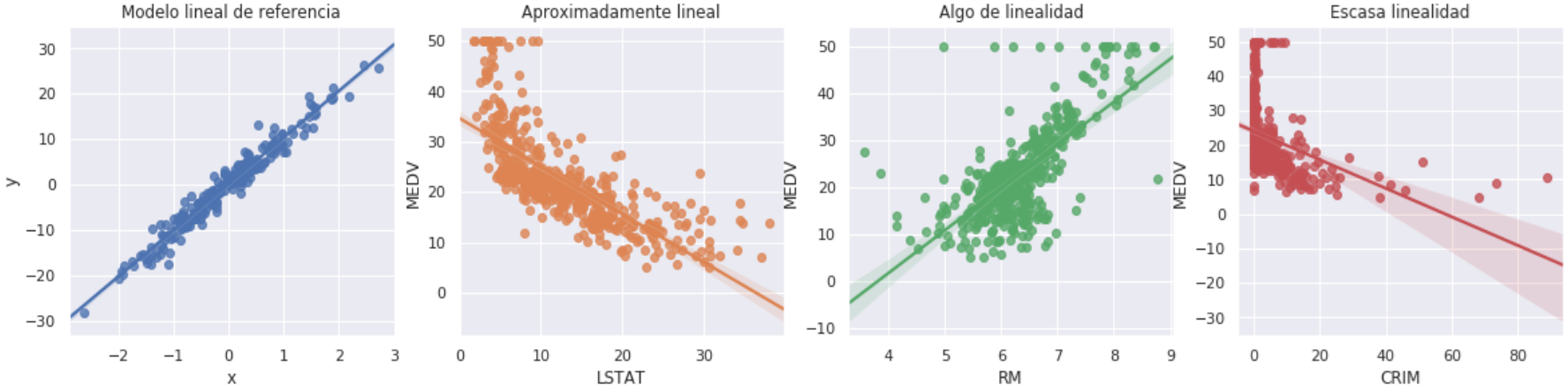
Los modelos lineales(*) realizan los siguientes supuestos sobre las variables independientes:

- Existe una relación lineal entre variables de entrada y de salida.
- Normalidad.
- Ausencia total de colinealidad, o muy baja.
- Homocedasticidad (homogeneidad de la varianza).

(*) Ejemplos: regresión lineal, regresión logística, PCA.

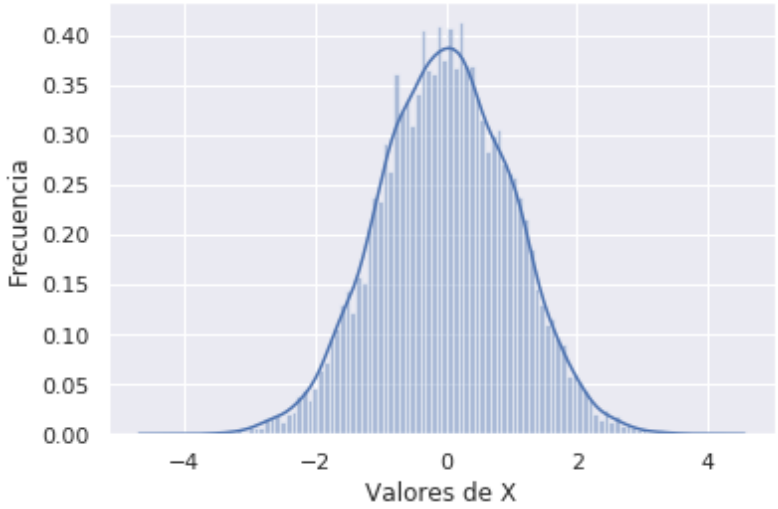
2.3.1 Supuestos de linealidad: Relación lineal

- y ~= b0 + b1X1 + b2X2 + … bnXn
- Un método de verificación es mediante scatter plots.
- A veces, aplicar transformaciones no lineales a las variables puede mejorar la relación lineal.



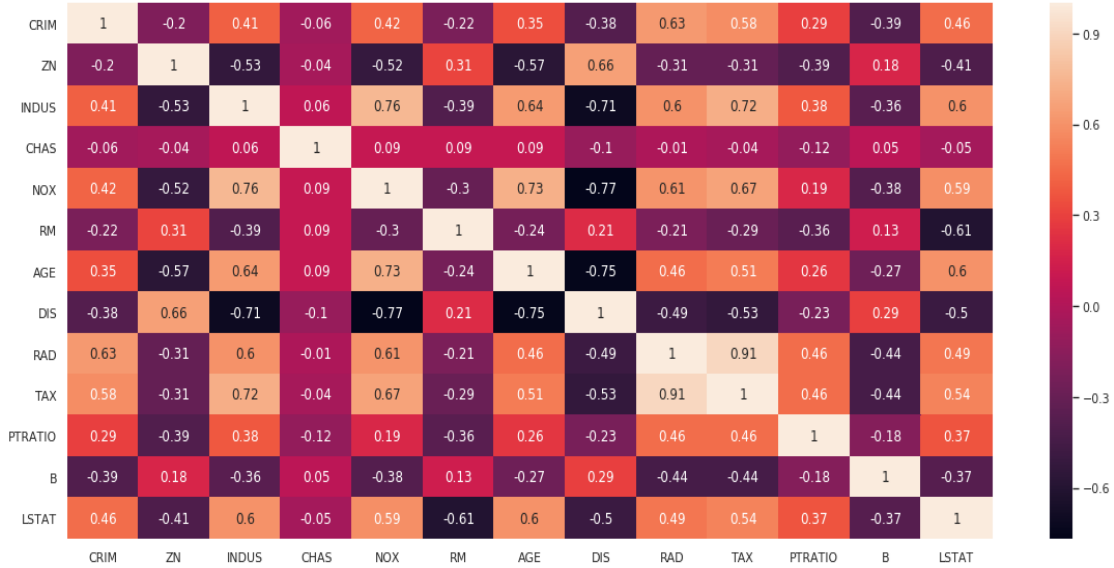
2.3.2 Supuestos de linealidad: Normalidad

- Las variables independientes obedecen una distribución gaussiana.
- Puede evaluarse con gráficos Q-Q.
- También puede ser evaluada con tests estadísticos, como el test de Kolmogorov-Smirnov.
- A veces, cuando una variable no tiene una distribución normal, puede aplicarse una transformación (por ejemplo log) para asemejarse a una distribución normal.



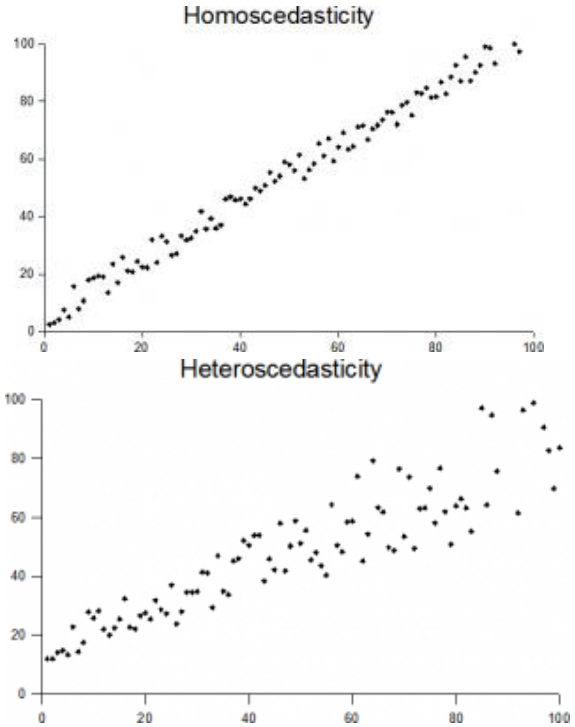
2.3.3 Supuestos lineales. Ausencia de colinealidad

La multi-colinealidad ocurre cuando las variables independientes están correlacionadas entre sí.
Puede evaluarse con una matriz de correlación.



2.3.4 Supuestos de linealidad: Homocedasticidad (homogeneidad de la varianza)

- Las variables independientes tienen la misma varianza finita.
- También conocida como homogeneidad de la varianza.
- Existen gráficos y tests que permiten determinarla, por ejemplo:
  - Gráfico de residuales.
  - Test de Levene.
  - Test de Barlett.
  - Test de Goldfeld-Quandt.
- A veces, la homogeneidad de la varianza puede mejorarse aplicando transformaciones no lineales.



¿Qué aspectos de las variables es importante considerar para un modelo de aprendizaje supervisado?

- Datos faltantes
- Variables categóricas. Cardinalidad y etiquetas raras
- Cumplimiento de los supuestos de linealidad
- **Distribución de los valores de las independientes**
- Valores extremos (outliers)
- Magnitud/escala

2.4 Distribución

- En la segunda clase se presentaron distribuciones:
  - Simétricas:
    - La media, mediana y moda coinciden.
  - Con oblicuidad
    - La media está influenciada por la cola.
- Los modelos lineales asumen que las variables independientes obedecen a una distribución normal.
- Otros modelos no realizan una suposición de cómo están distribuidas las variables, pero a veces una mejor distribución puede mejorar su desempeño.

2.4 Distribución: Normalización mediante transformación y discretización

Se estudiarán dos grupos de métodos para llevar una distribución a una forma normal:

- Transformaciones
  - ln(x)
  - exp(x)
  - 1/x
  - Box-Cox, Yeo Johnson
- Discretización
  - Bins de frecuencia fija (suelen mejorar la distribución)
  - Bins de ancho fijo (por lo general no mejoran la distribución)

¿Qué aspectos de las variables es importante considerar para un modelo de aprendizaje supervisado?

- Datos faltantes
- Variables categóricas. Cardinalidad y etiquetas raras
- Cumplimiento de los supuestos de linealidad
- Distribución de los valores de las independientes
- **Valores extremos (outliers)**
- Magnitud/escala

2.5 Valores extremos

- Un valor extremo es un punto significativamente diferente del resto de los datos.

  “An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism.” [D. Hawkins. Identification of Outliers, Chapman and Hall , 1980.]

2.5 Valores extremos

- En algunos casos, como detección de anomalías, el mayor interés está en la identificación de los valores extremos.
- En esta clase, sin embargo, nos enfocaremos en cómo tratarlos para mejorar el desempeño general de un modelo regresivo o de clasificación.

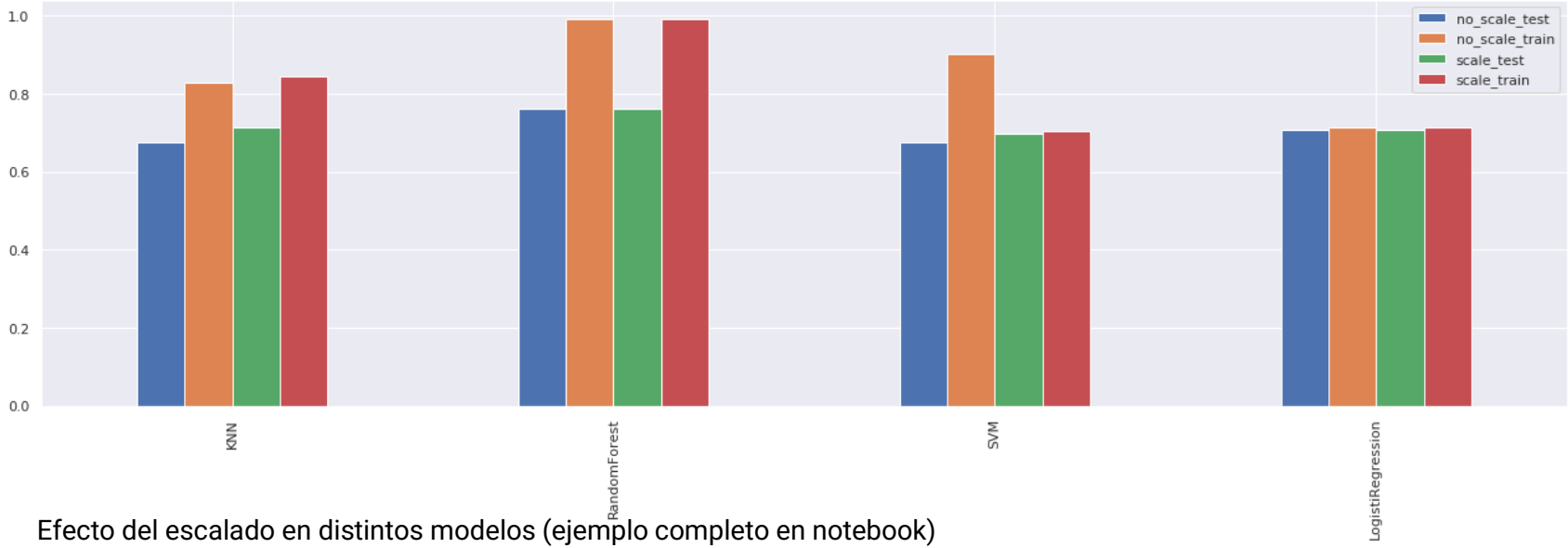
2.6 Magnitud de las variables de entrada

- Los coeficientes de regresión dependen de la magnitud de la variable.
- Las variables de entrada con mayor magnitud pueden dominar a las de menor magnitud.

2.6 Magnitud de las variables de entrada

- Algoritmos afectados
  - Regresión lineal y logística.
  - Redes neuronales.
  - Support Vector Machines.
  - KNN.
  - K-means.
  - Principal Component Analysis (PCA).
- No afectados (por ejemplo, los basados en árboles)
  - Árboles de clasificación y regresión.
  - Random Forest
  - Gradient Boosted Trees.

2.6 Magnitud de las variables de entrada



Efecto del escalado en distintos modelos (ejemplo completo en notebook)

Ejemplos en jupyter + recursos de interés

Clase 3.2 - Preparación de datos - Caracterización de variables.ipynb

QQPlots (visualización interactiva): https://xiongge.shinyapps.io/QQplots/

3. IMPUTACIÓN DE DATOS FALTANTES

Imputación de datos faltantes

- La imputación es el acto de reemplazar datos faltantes con estimaciones estadísticas de los valores ausentes.
- El objetivo de cualquier técnica de imputación es producir un dataset completo que permita entrenar un modelo de aprendizaje automático.
- Pueden agruparse las técnicas de imputación en dos categorías:
  - **Univariada**: cuando la imputación se realiza de manera independiente para cada variable de entrada, sin considerar las otras variables.
  - **Multivariada**: cuando el valor imputado para cada variable es una función de dos o más variables.

Técnicas de imputación univariada

- Algunas técnicas de imputación univariada (por tipo de variable al que aplican):
  - **Variables numéricas:**
    - Imputación por promedio/mediana.
    - Imputación por valor arbitrario.
    - Imputación de “fin de cola” (end of tail).
  - **Variables categóricas:**
    - Imputación por categoría frecuente.
    - Agregar categoría “FALTANTE”.
  - **Ambas:**
    - Análisis de caso completo.
    - Agregar indicador “FALTANTE”.
    - Imputación por muestreo aleatorio.

Análisis de caso completo (CCA)

- CCA (por sus siglas en inglés: *Complete Case Analysis*) consiste en descartar todas las observaciones en las que cualquiera de las variables están ausentes.
- Por lo tanto, se analizan sólo las observaciones para las cuales están todos los datos disponibles.

Supuestos de CCA

- MCAR (Missing completely at random)

Ventajas de CCA

- Simple.
- No se requiere manipulación previa de los datos.
- Preserva la distribución de las variables.

Limitaciones de CCA

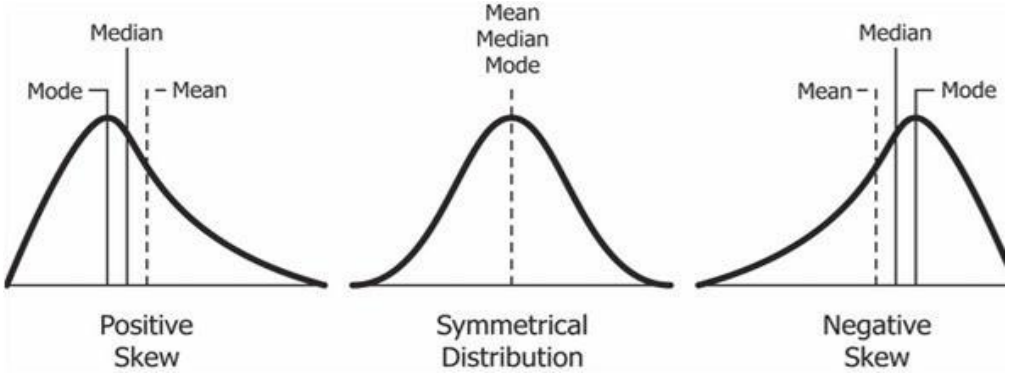
- Puede excluir una gran parte del dataset original.
- Las observaciones excluidas pueden ser informativas para el análisis (si no se cumple el supuesto MCAR, por ejemplo).
- CCA creará un dataset sesgado si los casos completos difieren de aquellos en los que falta información (MAR/NMAR).
- El modelo puesto en producción no podrá manejar observaciones incompletas.

¿Cuando aplicar CCA?

- Se cumple el supuesto MCAR.
- Las observaciones a eliminar no superan el 5% del total del dataset.

Imputación por media o mediana

- Consiste en reemplazar las ocurrencias de valores faltantes por la media o la mediana de esa variable.
- Si la variable está normalmente distribuida la media y la mediana son similares.
- Si en cambio, la distribución tiene oblicuidad, la mediana es una mejor representación de los datos faltantes.



Imputación por media o mediana. Supuestos.

- MCAR, MAR.
- Las observaciones faltantes se asemejan a la mayoría (por eso podemos reemplazarlas por la media o la mediana).

Imputación por media o mediana. Ventajas.

- Fácil de implementar.
- Se puede integrar en producción.

Imputación por media o mediana. Limitaciones.

- Distorsiona la distribución original.
- Distorsiona la varianza original.
- Distorsiona la covarianza con las variables restantes.
- A mayor cantidad de datos faltantes, mayor distorsión.

Imputación por media o mediana. ¿Cuando aplicar?.

- Se cumple el supuesto MCAR, MAR.
- No más del 5% de observaciones incompletas del total del dataset.
- Por lo general, se utiliza agregando una variable binaria indicando si ese valor estaba ausente.

Imputación por valor arbitrario. Definición

- Consiste en reemplazar las ocurrencias de los valores faltantes (NA) por una variable de un valor arbitrario.
- Los valores típicos son:
  - En distribuciones numéricas: 0, 999, -999 o -1 (para distribuciones positivas).
  - En variables categóricas: “MISSING”/”FALTANTE”.

Imputación por valor arbitrario. Supuestos

- MNAR.
- En este caso, sospechamos que existe una razón por la cual estos valores están ausentes, por lo cual la media o mediana no necesariamente son representativos.

Imputación por valor arbitrario. Ventajas.

- Fácil de implementar.
- Método rápido para obtener datasets completos.
- Puede integrarse en producción.
- Captura la importancia de los datos “faltantes”, en caso de haberla.

Imputación por valor arbitrario. Limitaciones.

- Distorsiona la distribución original de los datos.
- Distorsiona la varianza original.
- Distorsiona la covarianza con las variables restantes.
- Si el valor arbitrario está al final de la distribución, puede introducir outliers.
- Se debe tener cuidado de no elegir un valor arbitrario demasiado similar a la media o mediana.
- Cuanto mayor porcentaje de NA, mayor la distorsión.

Imputación por valor arbitrario. Cuando aplicar.

- Se cumple MNAR.

Imputación de “fin de cola”. Definición

- Es equivalente a la imputación por valor arbitrario, pero en este caso automáticamente se seleccionan valores arbitrarios al final de la distribución.
- Si la variable tiene una distribución normal se puede utilizar el promedio +- 3 veces el desvío estándar.
- Si la variable tiene una distribución con oblicuidad, se puede utilizar la regla de proximidad IQR.
- Sólo aplica para tipos numéricos.

Imputación por categoría frecuente o moda. Definición.

- La imputación por moda o categoría frecuente consiste en reemplazar los valores faltantes (NA) por la moda o valor más frecuente.
- Por lo general, esta técnica se utiliza para variables categóricas.

Imputación por categoría frecuente o moda. Supuestos.

- MCAR o MAR.
- Las observaciones faltantes se asemejan a la mayoría.

Imputación por categoría frecuente o moda. Ventajas.

- Fácil de implementar.
- Método rápido para obtener datasets completos.
- Puede integrarse en producción.

frecuente o moda. Limitaciones.

- Distorsiona la relación entre la etiqueta más frecuente con otras variables del dataset.
- Puede exagerar la presencia de la etiqueta más frecuente si el número de NAs es alto.
- A mayor cantidad de NA, mayor distorsión.

frecuente o moda. Cuando aplicar.

- MCAR.
- No más del 5% del total de observaciones faltantes.

Imputación por categoría faltante. Definición.

- Este método consiste en crear una categoría adicional (típicamente “MISSING” o “FALTANTE”) y asignarla a los valores con NA.
- Es el método más utilizado para imputación categórica.

Imputación por categoría faltante. Ventajas.

- Fácil de implementar.
- Método rápido para obtener datasets completos.
- Puede integrarse en producción.
- Captura la importancia de los tipos faltantes, en caso de haberla.
- No realiza supuestos sobre los datos.

Imputación por categoría faltante. Limitaciones.

- Si la cantidad de NA es baja, introduce un nuevo problema: agrega una nueva variable poco frecuente.

Imputación por muestreo aleatorio. Definición.

- Consiste en tomar observaciones aleatorias de las disponibles, y utilizarlas para completar los valores ausentes.
- Este método es válido tanto para variables numéricas como categóricas.

Imputación por muestreo aleatorio. Supuestos.

- MCAR.
- Se espera que conservar la distribución de la variable original.

Imputación por muestreo aleatorio. Ventajas.

- Fácil de implementar.
- Método rápido para obtener un dataset completo.
- Puede integrarse en producción.
- Preserva la varianza de la variable original.

Imputación por muestreo aleatorio. Limitaciones.

- Aleatoriedad.
- La relación entre las variables imputadas con otras variables puede verse afectada si el número de NAs es grande.
- Si bien es apto para producción, implica disponer del dataset original para extraer y reemplazar valores NA en las nuevas observaciones.

3. DESBALANCE DE CLASES

Desbalance de clases. Definición.

- En el mundo real no siempre vamos a tener una distribución “bonita” de nuestros datos.
- Muchas veces nos vamos a topar con categorías que inherentemente son poco frecuentes, ya sea por construcción del problema, sampleo o naturaleza del sistema.
- Ejemplos:
  - Detección de fraudes
  - Detección de anomalías
  - Diagnóstico médico.
  - Biométrica (Reconocimiento facial, huella digital, etc.)

Desbalance de clases. Definición.

- En general, vamos a considerar que tenemos un desbalance de datos cuando el peso de la (o las) clases mayoritaria(s) superan ampliamente a la(s) clase(s) minoritaria(s).
- El ratio puede variar, pero en general vamos a pensar que relaciones del tipo 100:1 ya puede ser un problema desbalanceado dependiendo de lo que utilicemos.

Desbalance de clases. Efectos.

- Todos los modelos clásicos de ML están preparados para poder trabajar con datos que tienen representatividad del modelo. Esto nos lleva naturalmente a que la cantidad de datos nos da una información latente.
- Esto, también nos puede jugar en contra si tenemos desbalance podemos tener respuestas *sesgadas* a la clase mayoritaria

Desbalance de clases. Técnicas posibles.

- *Under sampling* (reducir la cantidad de muestras de la muestra mayoritaria)
  - Sampleo aleatorio
  - Sampleo estratificado
  - etc.
- *Over sampling* (aumentar la cantidad de muestras de la muestra minoritaria)
  - Repetición de datos random.
- Generación de datos sintéticos (*SMOTE*, *ADASYN*)

DUDAS?

ENCUESTA