**1. Define Ciencia de Datos**

Es la obtención de conocimiento proveniente de grandes volúmenes de datos estructurados y/o no estructurados. Además, permite la creación de productos de datos: una aplicación que adquiere su valor a partir de datos, y crea más datos como resultado.

**2. Define Big Data y las 3 V’s**

El término Big data se refiere a información que no puede ser procesada o analizada utilizando métodos tradicionales

– Una base relacional sería muy lenta y costosa

• Se describe en términos de:

– Volumen: enormes cantidades de datos estructurados, no estructurados y semiestructurados

– Variedad: gran diversidad en el Apo de datos, tales como correos, tweets, audio, videos, etc.

– Velocidad: respuestas rápidas para obtener la información necesaria en el Aempo preciso

• Big data requiere de la ciencia de datos para transformar la información en conocimiento

**3. Menciona y describe los tipos de preguntas que se pueden responder en un proyecto de ciencia de datos**

**Descriptiva**

• Recolecta, presenta y caracteriza un conjunto de datos

**Exploratoria**

• InvesAgación realizada para ganar familiaridad con un fenómeno

**Predictiva**

• Análisis de hechos actuales o pasados con el fin de predecir eventos futuros desconocidos

**Causal**

• Determinar las causas detrás de un evento

**Deductiva**

• Inferir a parAr de una muestra aleatoria el comportamiento de la población.

**Mecanicista**

• Comprender los mecanismos detrás de un proceso

• ¿Cuál es el nivel de ingreso de mis clientes?

• ¿Qué clientes están en riesgo de irse con la competencia y como puedo retenerlos?

• ¿Qué productos debo tener en existencia en cada una de mis Aendas?

• ¿A qué equipo debo darle mantenimiento?

• ¿Cuál es el nivel adquisiAvo de mis clientes?

• ¿Qué opinan los clientes en redes sociales de nuestros productos o servicios?

**4. Menciona y describe las etapas de la metodología CRISP-DM**

****

The life cycle of a data mining project consists of six phases, shown in Figure 2. The sequence of the phases is not rigid. Moving back and forth between different phases is always required. The outcome of each phase determines which phase, or particular task of a phase, has to be performed next. The arrows indicate the most important and frequent dependencies between phases.

The outer circle in Figure 2 symbolizes the cyclical nature of data mining itself. Data mining does not end once a solution is deployed. The lessons learned during the process and from the deployed solution can trigger new, often more-focused business questions. Subsequent data mining processes will benefit from the experiences of previous ones. In the following, we briefly outline each phase:

**Business understanding**

This initial phase focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives.

**Data understanding**

The data understanding phase starts with initial data collection and proceeds with activities that enable you to become familiar with the data, identify data quality problems, discover first insights into the data, and/or detect interesting subsets to form hypotheses regarding hidden information.

**Data preparation**

The data preparation phase covers all activities needed to construct the final dataset [data that will be fed into the modeling tool(s)] from the initial raw data. Data preparation tasks are likely to be performed multiple times and not in any prescribed order. Tasks include table, record, and attribute selection, as well as transformation and cleaning of data for modeling tools.

**Modeling**

In this phase, various modeling techniques are selected and applied, and their parameters are calibrated to optimal values. Typically, there are several techniques for the same data mining problem type. Some techniques have specific requirements on the form of data. Therefore, going back to the data preparation phase is often necessary.

**Evaluation**

At this stage in the project, you have built a model (or models) that appears to have high quality from a data analysis perspective. Before proceeding to final deployment of the model, it is important to thoroughly evaluate it and review the steps executed to create it, to be certain the model properly achieves the business objectives. A key objective is to determine if there is some important business issue that has not been sufficiently considered. At the end of this phase, a decision on the use of the data mining results should be reached.

**Deployment**

Creation of the model is generally not the end of the project. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it. It often involves applying “live” models within an organization’s decision making processes—for example, real-time personalization of Web pages or repeated scoring of marketing databases. Depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process across the enterprise. In many cases, it is the customer, not the data analyst, who carries out the deployment steps. However, even if the analyst will carry out the deployment effort, it is important for the customer to understand up front what actions need to be carried out in order to actually make use of the created models.

**5. Contrasta los métodos estadísticos tradicionales con el aprendizaje de máquina**

• **Inferencia**: el proceso de sacar conclusiones sobre la población a parAr de muestras

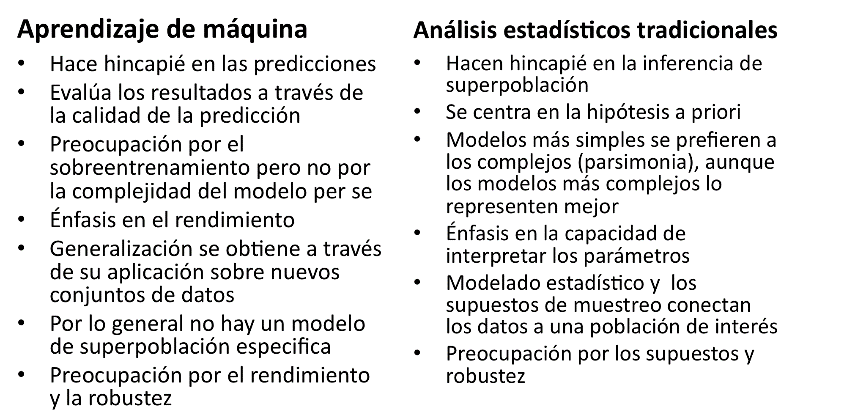
– estimación, muestreo, variabilidad, definición de poblaciones

La estadística busca modelos simples que expliquen el por qué de los fenómenos

**• Predicción:** el proceso de adivinar un resultado a partir de predictores

– clasificación, regresión y más recientemente aprendizaje de máquina, aprendizaje supervisado, deep learning, boosAng, árboles de decisión, bosques aleatorios, etc.

El aprendizaje de máquina busca que las predicciones sean lo más certero possible



**6. ¿Cómo se determina el éxito de un proyecto de ciencia de datos?**

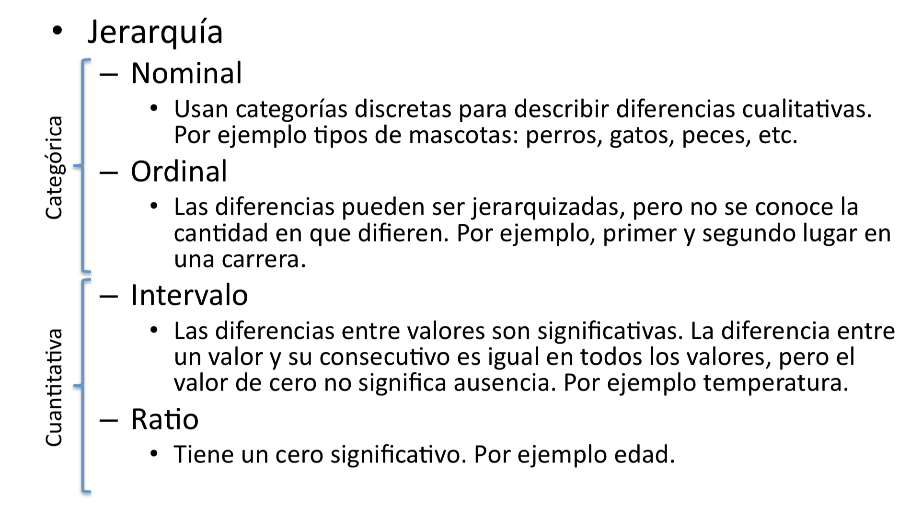
• Si se crea nuevo conocimiento.

• Si decisiones se toman o políAcas se determinan en función de los resultados del experimento

• Si se crea un informe, presentación o aplicación con impacto

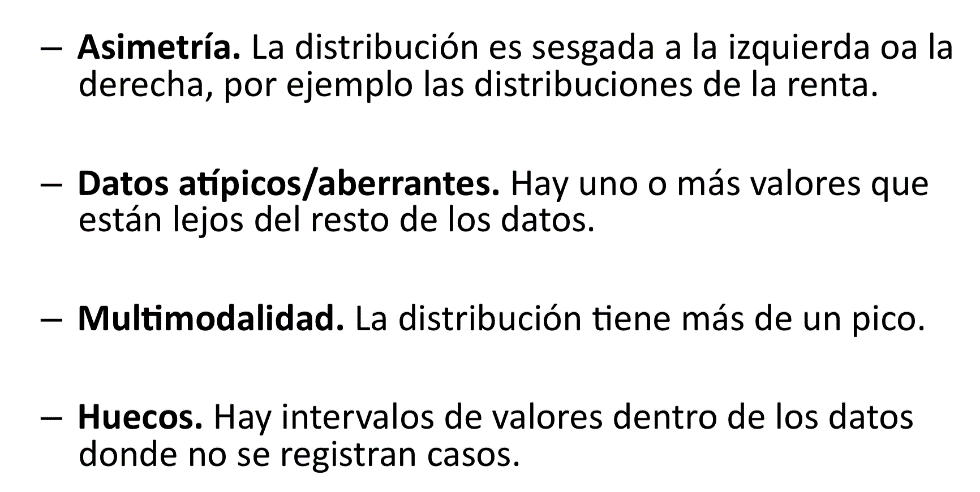
• Si se determina que los datos no pueden responder a la pregunta planteada

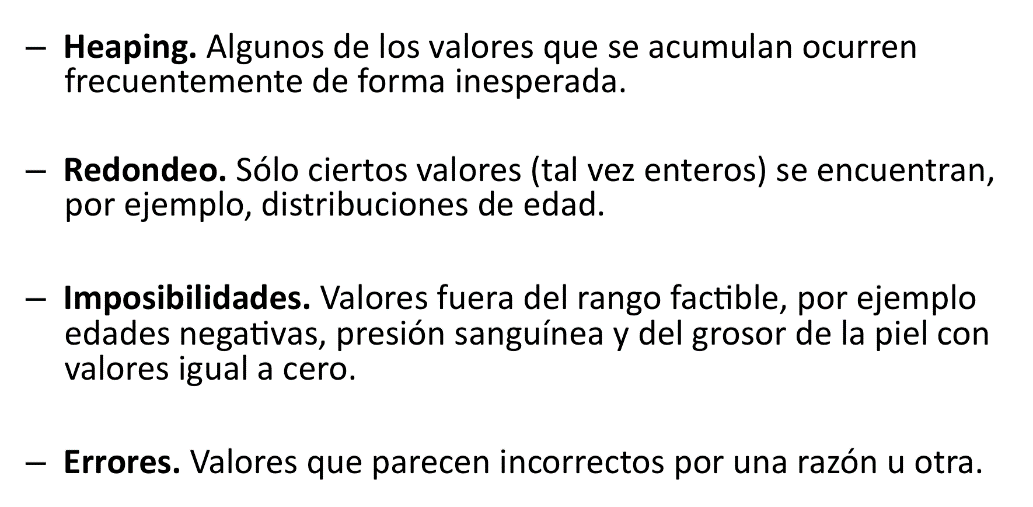
**7. Menciona y describe la jerarquía de escalas de medida**

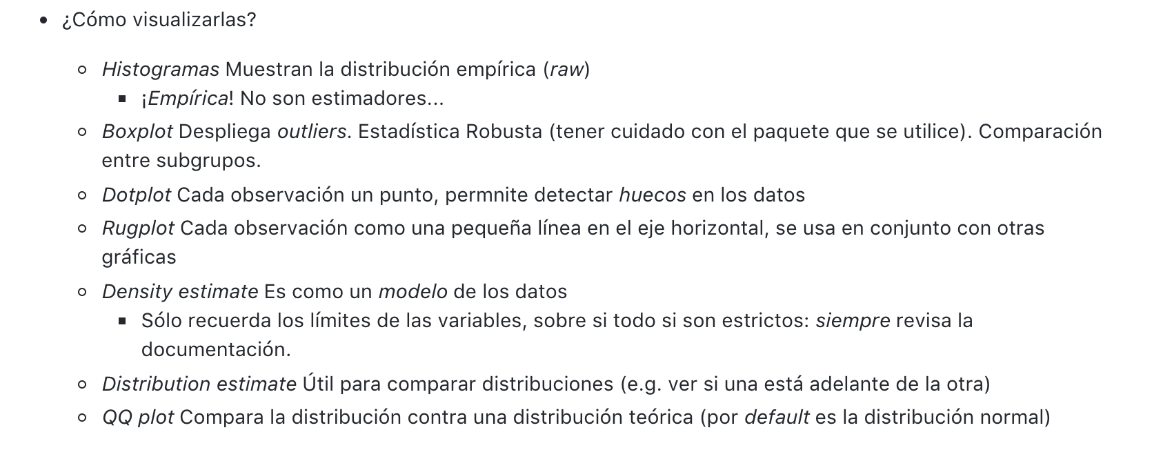
****

**8. ¿Qué podemos encontrar durante el análisis univariado cuantitativo y en el categórico?**

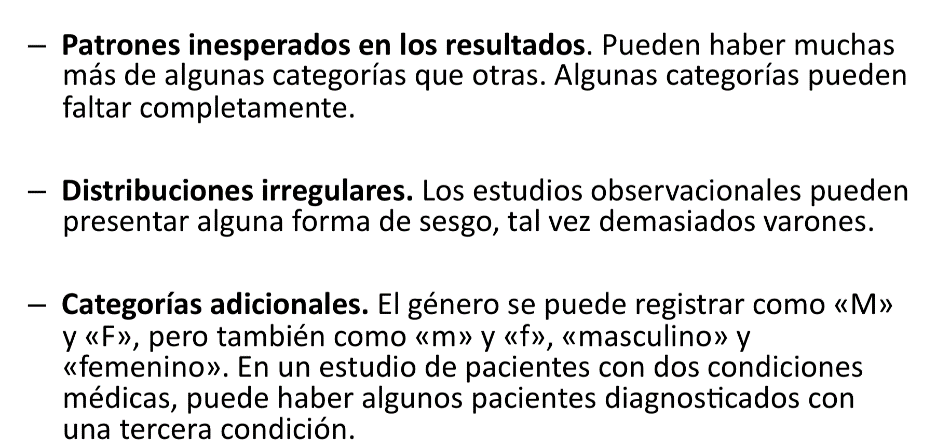
**Cuantitativas**

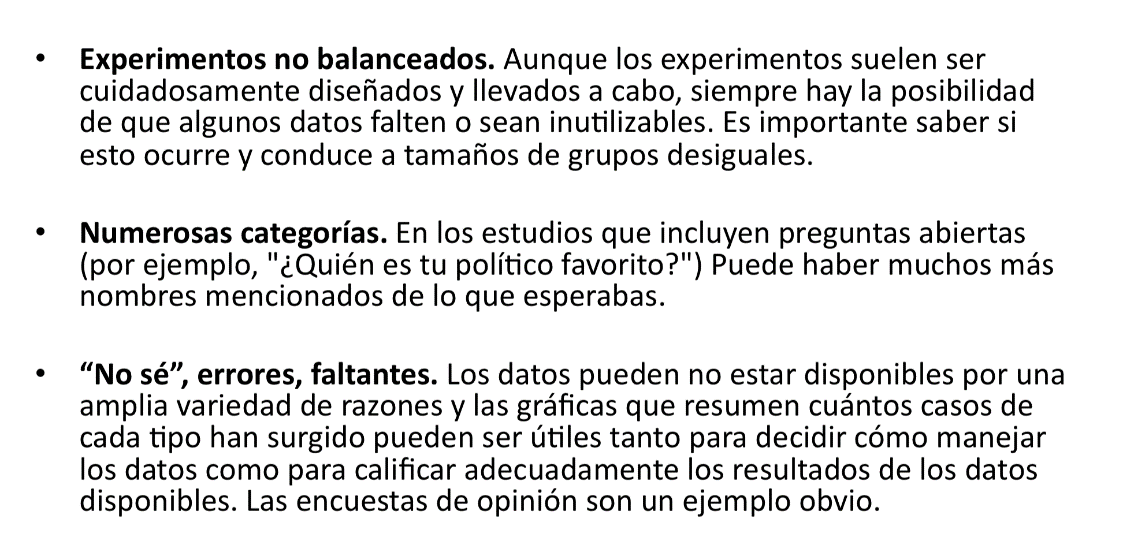
****

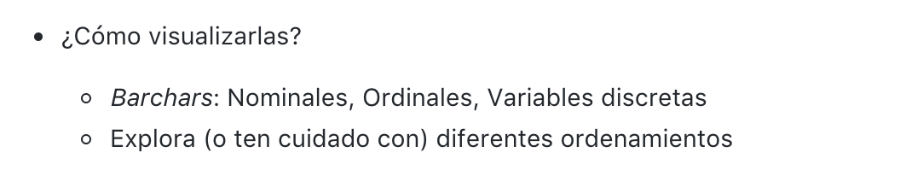
****

****

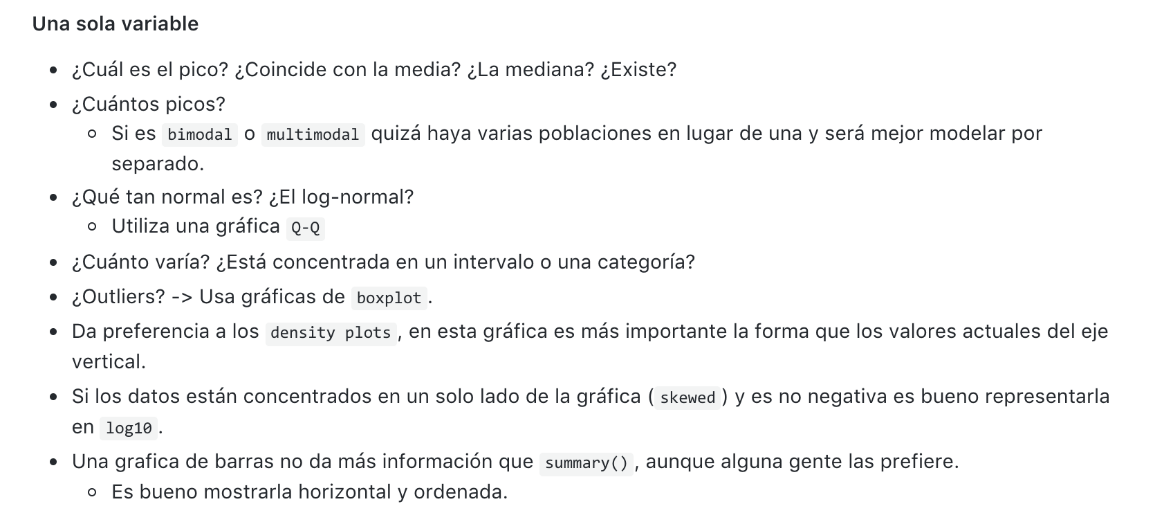
**Categoricas**

****

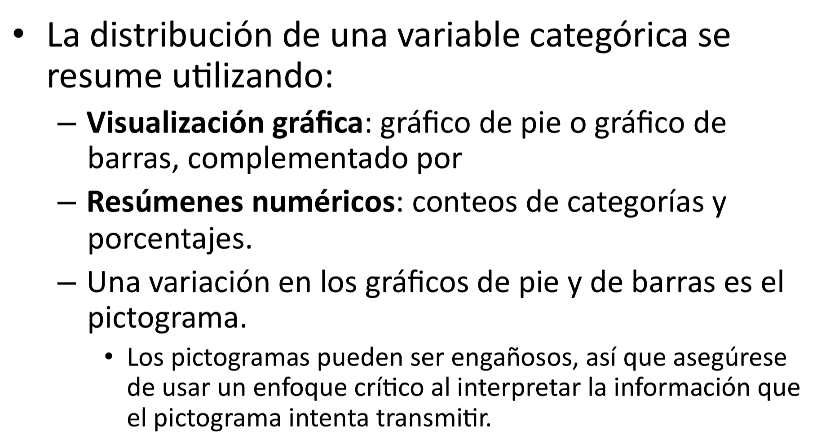
****

****

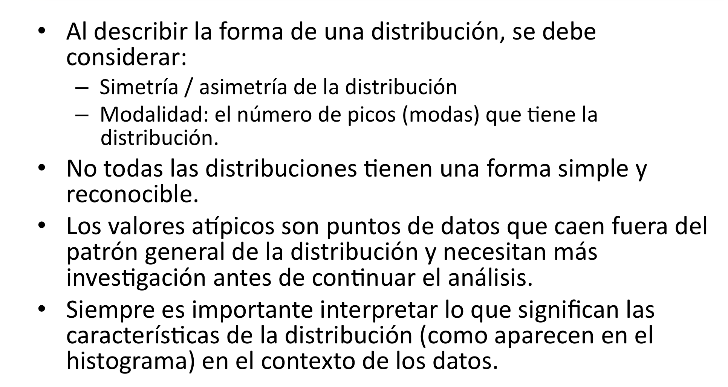
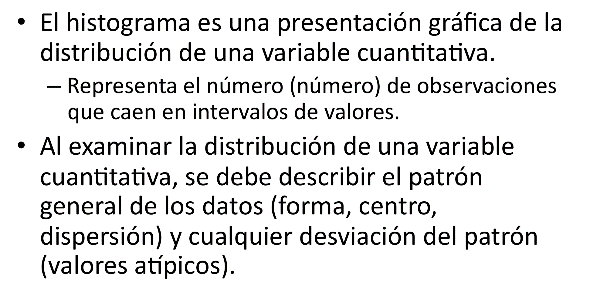
**9. ¿Qué tipo de gráficas se utilizan comúnmente en el análisis univariado cuantitativo y categórico y con qué fin?**

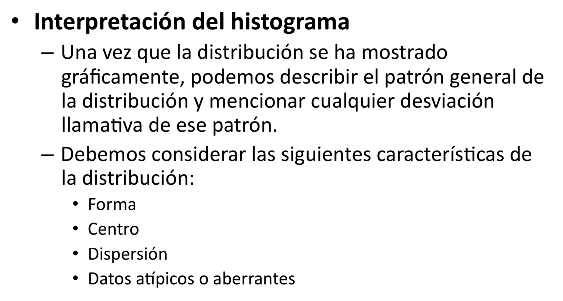
****

**Categóricas**

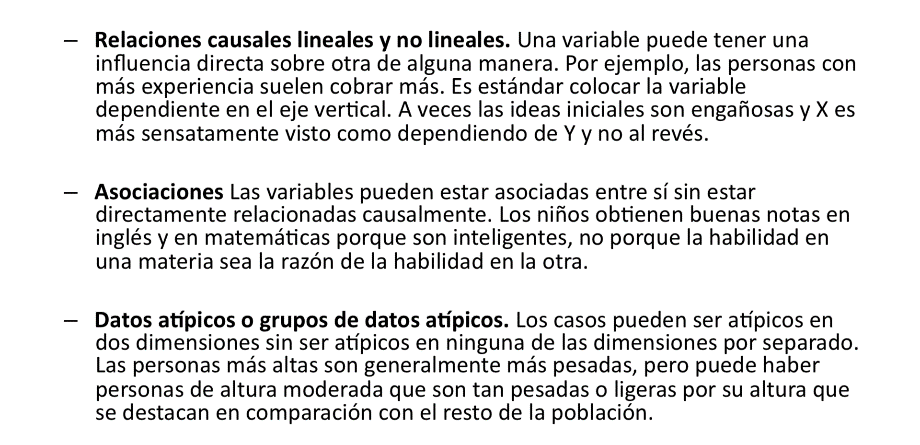
****

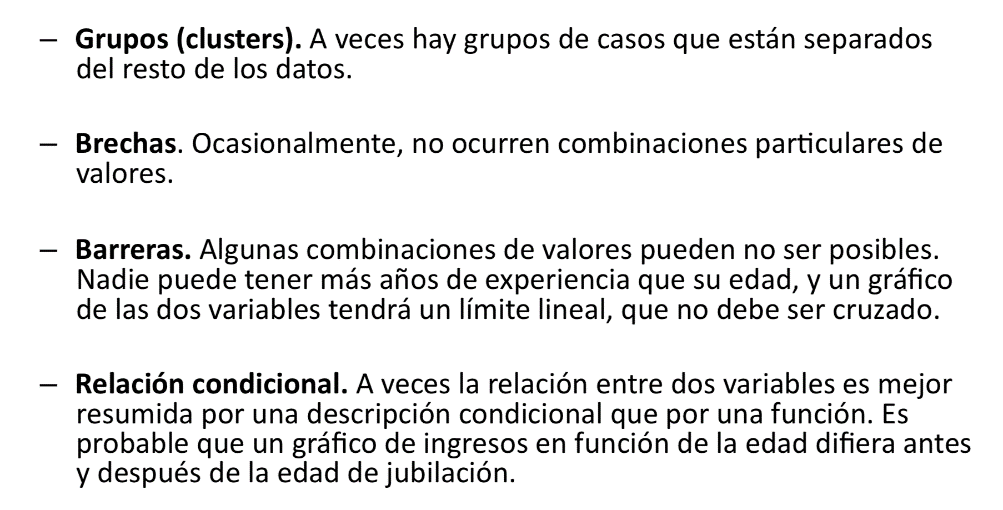
**Cuantitativas**

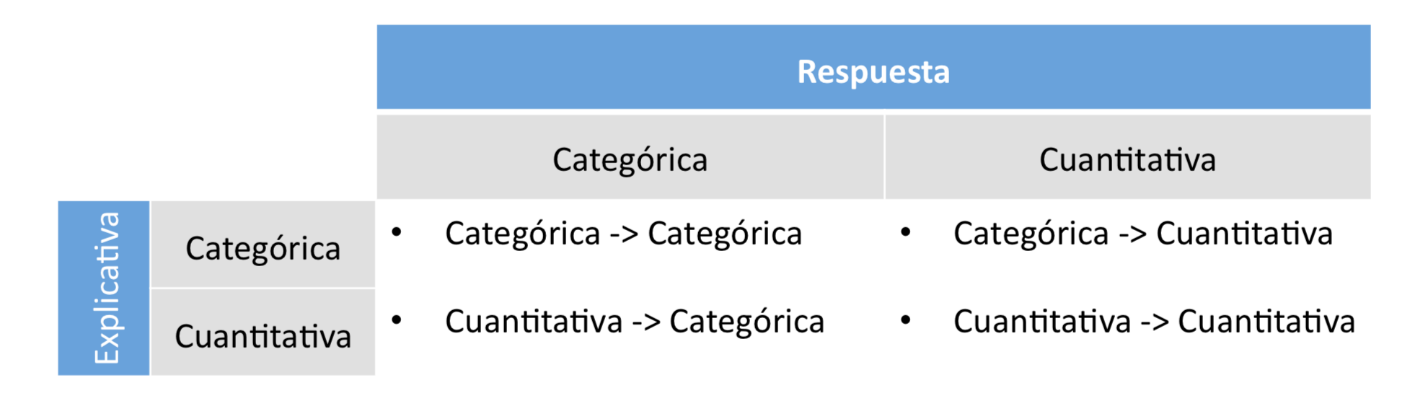
****

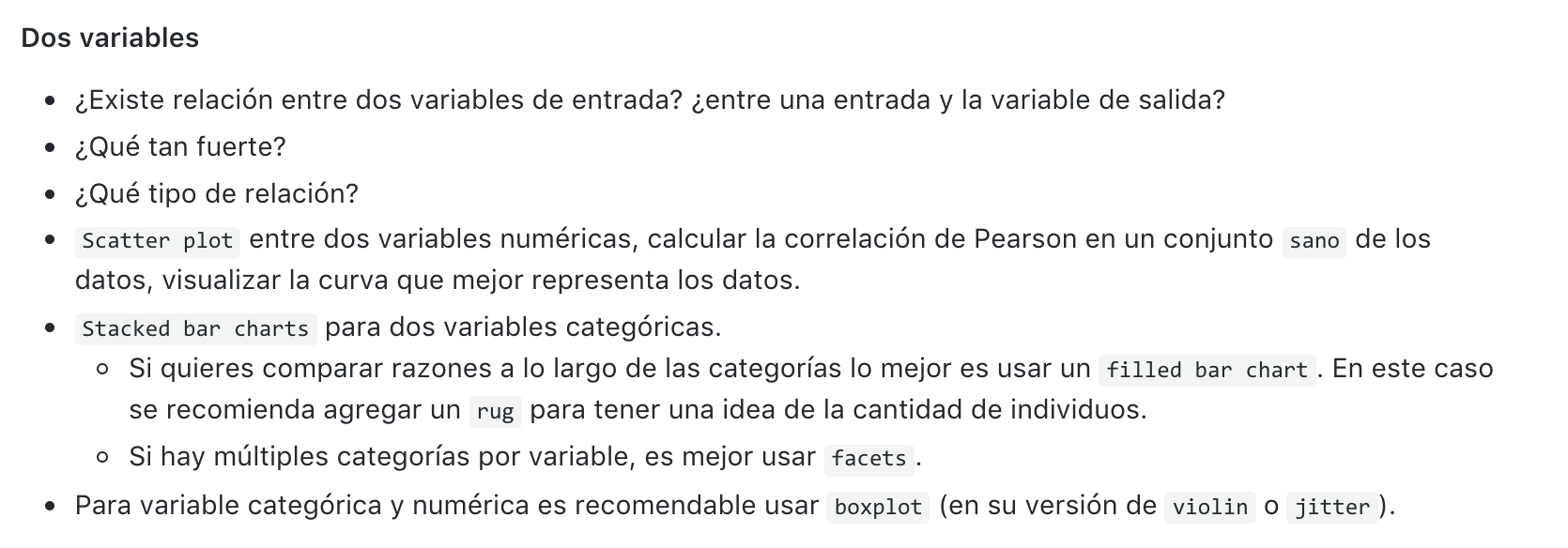
****

**10. ¿Qué podemos encontrar durante el análisis bivariado cuantitativo y en el categórico?**

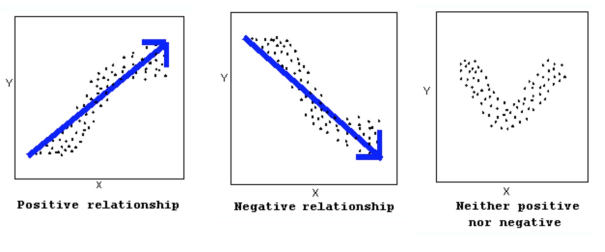
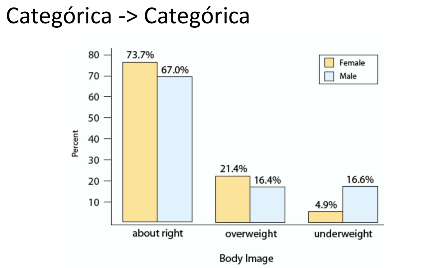
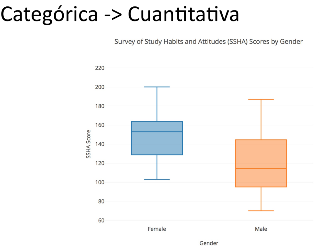
****

****

****

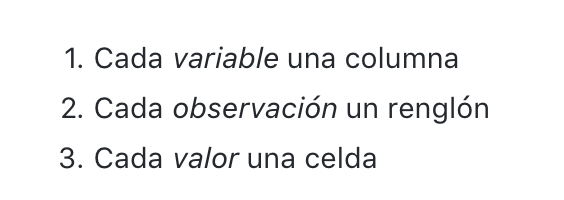
****

**11. ¿Qué tipo de gráficas se utilizan comúnmente en el análisis bivariado cuantitativo y categórico y con qué fin?**

****

**12. Describe las características de un conjunto de datos tidy**

Los datos Tidy, tienen tres caracteristicas:



**13. Describe el proceso de validación cruzada. ¿Para qué sirve?**

El proceso de validación cruzada consiste en tomar de un mismo set de datos, muestras aleatorias de n cantidad de registros cada una, y de esta manera hacer las pruebas de los modelos con combinaciones diferentes de esos sets de datos.

Por Ejemplo: si mi set de datos X = {x1, x2, x3, x4}, donde cada subconjunto xi tiene por ejemplo 5,000 registros. Voy a armar mi set de datos entrenamiento con combinaciones de 3 de esos sets, y el restante seria el set de pruebas. Se toman métricas del o de los modelos ajustando con cada set de datos, y de esta manera podemos sacar el mayor provecho de un modelo. Este proceso además, ayuda a que no sobre ajustemos o sub ajustemos un modelo. Se hace uso de los datos de la mejor manera posible, a mi criterio, ya que siempre los datos son un bien escaso.

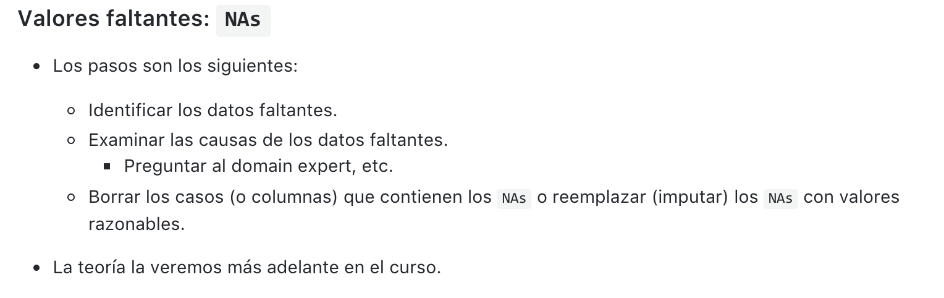
**14. ¿Qué es la maldición de la dimensionalidad?**

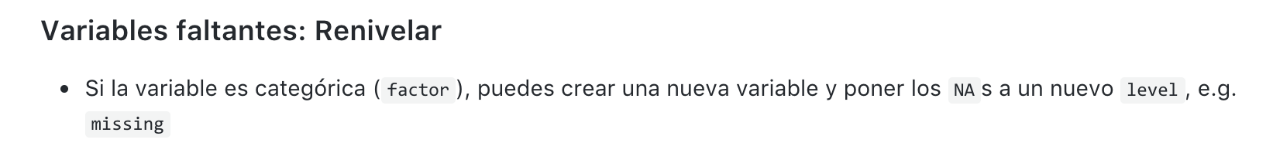
La maldición de la dimensionalidad se refiere a que en muchas oportunidades, debemos evaluar con mucho cuidado la cantidad de variables que vamos a utilizar en un modelo de datos, ya que a veces nos vemos tentados a usar más de las necesarias, y esto trae aparejado varios problemas, como por ejemplo, el conjunto de datos representativo de cada “unico” ejemplo de cada registro. Esto a su vez tiene grandes consecuencias en el tamaño del set de datos y por ende, en su procesamiento.

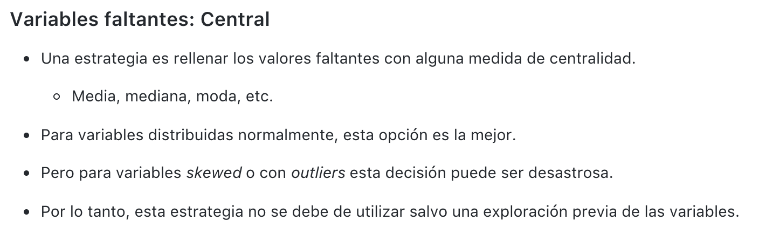
**15. ¿En qué consiste el efecto Rashomon? ¿Qué método puede evaluar de mejor forma la importancia de las variables?**

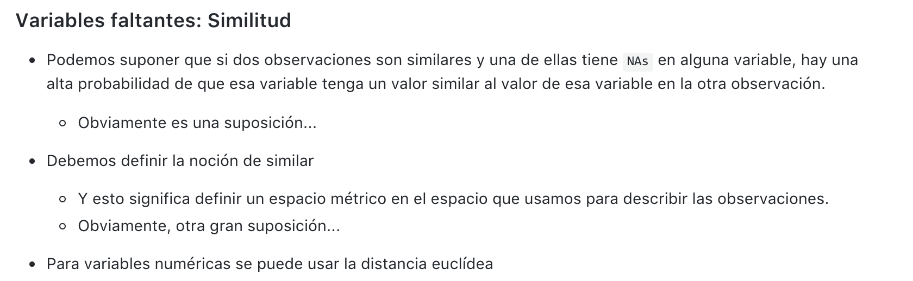
Many different models describe a situation equally accurate.

**16. Define tres formas de resolver el problema de variables faltantes indicando cuándo es conveniente utilizar cada una de ellas.**

****

****

****

****

**17. ¿Qué es aprendizaje supervisado, no supervisado y por refuerzo? ¿qué diferencia existe entre cada uno de ellos? ¿cuándo los uso?**

El aprendizaje se cataloga como supervisado o no supervisado dependiendo de si el algoritmo debe contar o no con información especifica de datos satisfactorios para el objetivo del aprendizaje

**No Supervisado**

Encargado de detectar factores no fácilmente observables dentro de los datos. Se utiliza principalmente en minería de datos. Sospechamos que existe algún tipo de relación entre los datos, pero es muy complejo intentar adivinarla.

Algoritmos

– Agrupamiento (clustering) – Componentes principales

– Modelo de mezclas (mixture models)

**Supervisado**

Utilizan una serie de variables y resultados observables asociados para predecir los resultados cuando estos no se conocen. Se utilizan ampliamente en analítica predictiva

Algoritmos

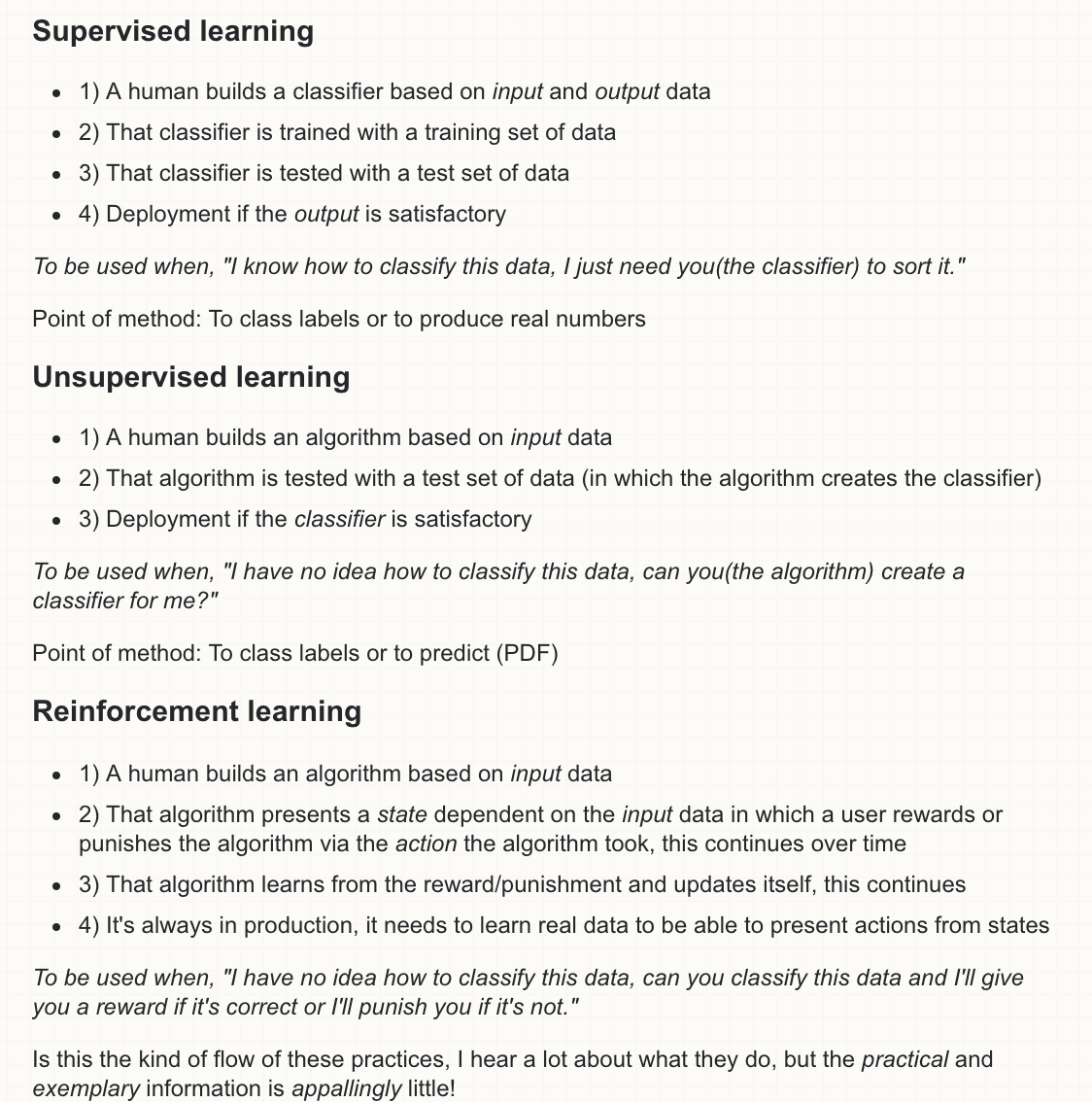
– Método de Bayes básico

– Regresiones lineales y no lineales – Redes neuronales

– Máquinas de soporte vectorial

– Árboles de decisión

– K-vecinos cercanos



**18. ¿Qué diferencia existe entre PCA y t-SNE? ¿Cuándo los utilizo?**

Ambos son algoritmos para reducer la dimensionalidad de los datos que voy a modelar.

PCA es una técnica matemática de reducción de dimensionalidad, mientras que TSNE convierte afinidad de puntos en probabilidades:

* PCA (linear)
* t-SNE (non-parametric/ nonlinear) usa hiperparámetros

The best way to use the algorithm is to use it for exploratory data analysis. It will give you a very good sense of patterns hidden inside the data. It can also be used as an input parameter for other classification & clustering algorithms.

**19. ¿Cómo puedo utilizar singular value decomposition para realizar clustering?**

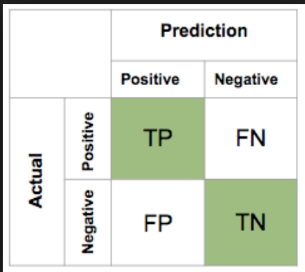
Clustering with Singular Value Decomposition (CSVD) method, which combines clustering and singular value decomposition (SVD) to reduce the number of index dimensions, while maintaining a reasonably high precision for a given value of recall. In the proposed CSVD method, homogeneous points are grouped into clusters such that the points in each cluster are more amenable to dimensionality reduction than the original dataset.

Reducir dimensiones de la matriz manteniendo las distamcias.

**20. ¿Para qué se utiliza una matriz de confusión? ¿Sólo sirve para clasificación binaria?**

Cada columna de la matriz representa el número de predicciones de cada clase, mientras que cada fila representa a las instancias en la clase real. Uno de los beneficios de las matrices de confusión es que facilitan ver si el sistema está confundiendo dos clases.

No, se puede extender para más de dos clases.

****

**21. ¿Qué es un micro-average?¿un macro-average? ¿Qué diferencia hay entre ellos? ¿Cuándo se utilizan?**

Evaluación de modelos de clasificación: **Accuracy**

¿Qué tan frecuente el clasificadorhace la predicción correcta?

El accuracy es un ejemplo de **micro-average**

La Matriz de confusion permite calcular el **accuracy de cada clase,** un ejemplo de un **macro-average**

A macro-average will compute the metric independently for each class and then take the average (hence treating all classes equally), whereas a micro-average will aggregate the contributions of all classes to compute the average metric. In a multi-class classification setup, micro-average is preferable if you suspect there might be class imbalance (i.e you may have many more examples of one class than of other classes). If your dataset varies in size, micro-average is the useful tool.

**22. ¿Por qué es importante separar los datos en entrenamiento y prueba antes de realizar el modelo? ¿Qué es el conjunto de validación?**

Para evitar el Overfitting. The models become very specific to the training data, so far as they even get trained on the occasional erroneous labels present in the training data. As a result, over fitted models don't work at well on the other data outside those in your actual data.

Training dataset

A training dataset is a dataset of examples used for learning, that is to fit the parameters (e.g., weights) of, for example, a classifier. Most approaches that search through training data for empirical relationships tend to over fit the data, meaning that they can identify apparent relationships in the training data that do not hold in general.

Test dataset

A test dataset is a dataset that is independent of the training dataset, but that follows the same probability distribution as the training dataset. If a model fit to the training dataset also fits the test dataset well, minimal overfitting has taken place. A better fitting of the training dataset as opposed to the test dataset usually points to overfitting.

Training Dataset: The sample of data used to fit the model.

Validation Dataset: The sample of data used to provide an unbiased evaluation of a model fit on the training dataset **while tuning model hyperparameters.** The evaluation becomes more biased as skill on the validation dataset is incorporated into the model configuration.

Test Dataset: The sample of data used to provide an unbiased evaluation of a final model fit on the training dataset.

**23. ¿Por qué se dice que algoritmos como random forest, redes neuronales y máquina de vectores de soporte son cajas negras?**

Model interpretation is one of the key aspects of the model evaluation process. The explanation of the relationship between model variables and outputs is easy for statistical models, such as linear regressions, thanks to the availability of model parameters and their statistical significance.

For “black box” models, such as random forest, this information is hidden inside the model structure

**24. ¿Cómo puedo comprender los modelos generados por este tipo de algoritmos?**

Como una perturbacion en los datos de entrada afecta las salidas o analisis de importancia relative

A traves de un analisis de la importancia de variables var inplot R para árboles

**25. ¿Qué tipo de modelo usarías si tienes pocas observaciones y un número mayor de variables? ¿Por qué? ¿Cómo sé que puede funcionar?**

SVM reduce la maldición de la dimensionalidad y con kernels que relajan el requerimiento de la separabilidad lineal

**26. ¿Qué es el mean decrease accuracy? ¿Para que sirve?**

Es una medida de la importancia de cada variable en el modelo, y se prueba excluyendo cada variable y viendo cuanto se ve afectada el accuracy. Cuanto más afecta, mas importante es.

27.- ¿Qué es caret y para qué funciona?

*R.-* caret por **C**lassification **A**nd **R**egression **T**rainning. Es un conjunto de funciones que buscan simplificar el proceso para generar modelos predictivos. El paquete contiene herramientas para :

* Separación de datos
* Pre-procesamiento
* Selección de características
* Ajustado del modelo con resampleo
* Estiamcion del orden de importancia de las variables
* entre otras

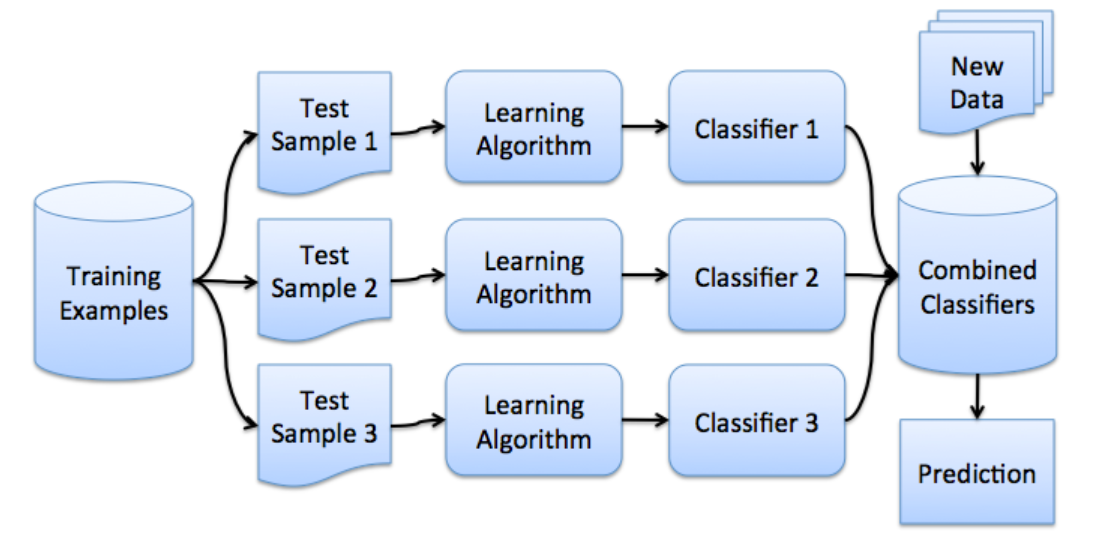
Nota.- esta paquetería empezo como una forma de proveer una interface uniforma de funciones así como de actividades estandarízadas (por ejemplo: ajuste de parámetros e importancia de las variables)

28.- ¿Qué es doSNOW y para qué funciona?

*R.-* doSNOW por **S**imple **N**etwork **O**f **W**orkstations. Es una paquetería en R con un conjunto de funciones que permiten hacer computo en paralelo a travez de la interface de R. En particular la paquetería permite registrar *clusters* de objetos que seran usados para la ejecución en paralelo

29. ¿Qué es un método de ensamble? ¿por qué y cómo funciona? ¿Sólo pueden hacerse ensambles del mismo tipo de algoritmo? Si/no por qué

Los métodos combinados (métodos de ensamble) utilizan múltiples algoritmos de aprendizaje para obtener un rendimiento predictivo que mejore el que podría obtenerse por medio de cualquiera de los algoritmos de aprendizaje individuales que lo constituyen.



El término de métodos de ensamble se suele reservar para aquellas combinaciones que hacen uso de múltiples hipótesis pertenecientes a una misma familia, mientras que se usa el término más general de sistemas de aprendizaje múltiples cuando las hipótesis que se combinan provienen de diversas familias.

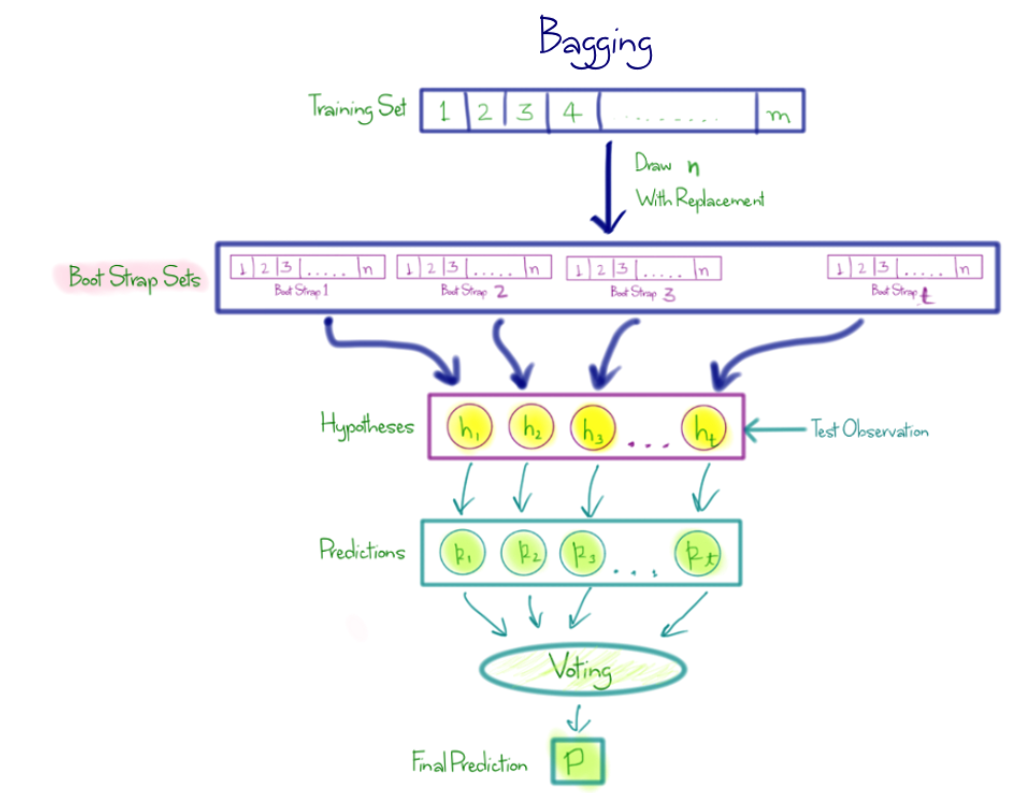
Puede resultar muy útil dado que cada modelo que compone el ensamble puede ser muy bueno para clasificar ciertas características y al unirlos se convierten en un modelo más potente. Muchos modelos tontos y convertirlos en uno más potente. Ej Decision tree sencillos -> Random Forest. Bootstraping y cambio de predictores.

30. ¿Cómo se resuelve un ensamble de regresión? Tratas de minimizar la pérdida cuatrática ¿Cómo uno de clasificación? Intentas maximizar accuracy

La idea de los métodos combinados es considerar múltiples hipótesis simultáneamente para formar una hipótesis que, con suerte, se comporte mejor.

Al final, el objetivo es obtener una predicción, entre ellas para el problema de clasificación.

Un ejemplo de método de ensamble es Bootstrap (también conocida como Bagging) es realmente un meta-algoritmo diseñado para conseguir combinaciones de modelos a partir de una familia inicial, provocando una disminución de la varianza y evitando el sobreajuste.



Otro ejemplo es el boosting que utiliza el conjunto completo de entrada, y se manipulan los pesos de los datos para generar modelos distintos. La idea es que en cada iteración se incremente el peso de los objetos mal clasificados por el predictor en esa iteración, por lo que en la construcción del próximo predictor estos objetos serán más importantes y será más probable clasificarlos bien.

31.- ¿Qué es ICE y para qué funciona?

*R.-* ICE por **I**ndividual **C**onditional **E**xpectation. Es un conjunto de funciones que buscan hacer más interpretables los resultados de cualquier modelo de predicción supervisado

32.- ¿Qué es LOCO y para qué funciona?

*R.-*

33.- ¿Cuando usar Bayes Ingenuo y por qué?

*R.-* El enfoque de Bayes Ingenuo para clasificación da buenos resutlados cuando se tienen multiples clases. Las ventajas son:

* Es más computacionalmente más facil y si el supuesto de independencia se sostiene, un modelo de Bayes ingenuo, convergirá más rápido que cualquier otro modelo de discrimanción a la regresion logística.
* Necesita menos cantidad de información
* Requiere menos tiempo de entrenamiento

34.- ¿Para qué sirve el truco del kernel en las máquinas de vectores de soporte?

*R.-* Permite añadir un número arbitrariamente largo de dimensiones sin tener que pagar un precio computacionalmente alto. Pues, genera un mecanismo que permite calcular el producto punto de los datos de forma eficiente. Computar el kernel toma un tiempo **proporcional al numero de atributos**, mientras el número de características generadas es **proporcional al cuadrado del número de atributos**

35.- Compara la cantidad de parámetro s a estimar para la predicción de una regresión lineal, una red neuronal y k-vecinos cercanos

*R.-*

* Regresión Lineal.- Sea **n** el número de características, el número de parámetros a estimare es igual a **n** o los βi′sβi′s o w′iswi′s asociados
* Red Neurona.- Todos los w′iswi′s que conectan a una neurona con otra en cada una de sus capas (cuando la red cumple con ciertas características se puede aplicar el *metodo de backpropagation*)
* K-nearest neighbors.- Sea **K** el número de instacias y **n** el número de datos de entrenamiento, es necesario calcular la distancia entre las instancias y cada uno de los ejemplos del conjunto de entrenamieno **K/n**

36.- Menciona dos metodos para representar numericamente las variables categóricas y cuando se usan

*R.-* *One-hot-encoding* o *Dummy variables*. El primero genera una variable indicativa para cada uno de los valores de la variable de interés, el segundo, genera una variable indicativa para cada uno de los valores de la varaible de interés **menos uno**

37.- ¿Cuándo es conveniente convertir variables continuas a categóricas?

*R.-* Transformar variables continuas a categoricas puede ayudar a construi modelos más complejos con regresiones lineales (aumentan su expresividad). Los arboles de decisión lo hacen automaticamente

38. ¿En qué consiste la ingeniería de características?

Es el proceso de determinar que variables productivas contribuyen mejor al poder predictivo del algoritmo. FE es, quizá, la parte más importante del proceso de minería de datos. Con buenas variables, un modelo simple puede ser mejor que un modelo complicado con malas variables. Es el elemento humano en el modelado: El entendimiento de los datos, más la intuición y la creatividad, hacen toda la diferencia. Es más un arte que una ciencia. Regularmente es un proceso iterativo con el EDA. Un domain expert puede ser de mucha utilidad en esta etapa. Feature Learning es lo que hace Deep Learning y no se verá en este curso.

39. Describe el proceso manual para generar características.

• Brainstorming

o No juzguen en esta etapa

o Permitan y promuevan ideas muy locas.

o Construyan en las ideas de otros

o No divaguen

o No mantengan conversaciones en paralelo

o Sean visuales

o Vayan por cantidad, la calidad se verá luego

o Otros consejos se pueden consultar aquí (https://challenges.openideo.com/blog/seven-tips-on-better-brainstorming, son los mismos)

• Decidir que features crear

o No hay tiempo infinito

• Crear esos features

• Estudiar el impacto de los features en el modelo

• Iterar

40. Describe el proceso automático para generar características.

• Interacción multiplicativa

o (C = A \cdot B)

o Hacer para todas las posibles combinaciones.

o Es importante mencionar que estas nuevas variables benefician mucho a los regresores lineales, pero no afectan mucho a árboles de decisión (e.g. el Random Forest)

• Interacción de razón

o (C = A / B)

o Tener cuidado con dividir por cero (\to) hay que tomar una decisión

o Hacer para todas las posibles combinaciones.

• Transformar una variable numérica en una binaria.

o Se trata de encontrar el cut-off que maximize tu variable dependiente.

o Muy parecido a lo que hacen algoritmos como el J48 (en su versión comercial se conoce como C5).

o Hay un paquete de R que lo implementa: C50.

• Numérica (\to) bin.

• Otras

o (X^2)

o (\log X)

o etc.

41. ¿Cómo puedo crear una variable binaria a partir de una continua? ¿Cómo decido el punto de corte?

Se trata de encontrar el cut-off que maximize tu variable dependiente. Buscar el valor que divida los datos como me interesan con datos continuos y de ahi puedo hacer feature engineer.

Muy parecido a lo que hacen algoritmos como el J48 (en su versión comercial se conoce como C5).

Hay un paquete de R que lo implementa: C50. Numérica (\to) bin

42. ¿Para qué sirve la selección de características?

El proceso de seleccionar variables antes que ejecutar los algoritmos.

• Realiza cross-validation

o Realizar cross-validation sólo en una parte del proceso (i.e. el modelo) es hacer trampa.

• ¡Cuidado! No hagas feature selection en todos tus datos antes de construir el modelo.

o Aumenta el riesgo de over-fitting.

o Aún realizando cross-validation.

• Hay todo un paquete en sklearn : sklearn.feature\_selection

43. ¿Qué es el filtrado basado en las propiedades de la distribución? Da un ejemplo (Luis)

Filtrado basado en las propiedades de la distribución

• Si hay poca variabilidad, no pueden ser usados para distinguir entre clases.

• Podemos utilizar como medidas de variabilidad a la mediana y al inter-quartile range IQR.

• En sklearn puedes utilizar VarianceThreshold

Filtrado basado en las propiedades de la distribución (Algoritmo)

• Obtenga para cada variable su mediana.

• Obtenga para cada variable sus quartiles, en particular, reste el tercer quartil del primero, para obtener el IQR.

• Realice un scatter-plot entre ambas variables, esta gráfica nos da una visión de la distribución de las variables.

• Eliminemos las variables que tengan "baja variabilidad" i.e. que sean menores que un porcentaje del IQR global.

o e.g. (< 1/5) ó (< 1/6).

• ¡Cuidado! Que las variables individuales tengan baja variabilidad, no significa que unidas con otras variables la tengan. Para una posible solución ver "A practical approach to Feature Selection" de Kira and Rendell, 1992.

44. ¿En qué consiste el filtrado por correlación? (Luis)

• Descrito en "Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution de Yu & Liu ICML 2003 (https://static.aminer.org/pdf/PDF/000/335/746/feature\_selection\_for\_high\_dimensional\_data\_a\_fast\_correlation\_based.pdf)

• Obtienes un conjunto de variables no muy relacionado entre sí, pero altamente relacionado a la variable de salida.

45. ¿En qué consiste el Fast-correlation filtering? Describe el algoritmo

• Encuentra una medida de relación entre cada par de variables.

o Aquí usaremos la correlación, el artículo usa otra cosa.

• Encuentra la correlación de cada variable con la variable de salida.

• Ordena las variables según su correlación con la variable de salida.

• Elige la mejor variable (la de hasta arriba).

• Tira las variables muy correlacionadas con esta.

• Repite el proceso.

46. ¿En qué consiste el Forward Selection?

El cual inicia sin variables y va agregando una a una las variables, hasta que no mejora la metrica de evaluación.

• Ejecuta el algoritmo con cada variable (i.e. de manera individual)

• Si tienes (x) número de variables, ejecutas el algoritmo (x) veces.

• Como siempre, usando cross-validation.

• Elige el mejor modelo y quédate con esa variable.

• Ahora, ejecuta el modelo de nuevo, pero ahora con la variable recién seleccionada y con cada variable restante.

• Elige el mejor modelo y quédate con esas dos variables.

• Repite hasta que no mejore el modelo agregando más variables.

47. ¿En qué consiste el Backward Selection?

Empieza con todas las variables en el modelo, y se van removiendo.

• Backward selection es el mismo algoritmo, pero invertido

• kind-of …

• En sklearn, este algoritmo está implementado con el nombre RFE (Recursive Feature Elimination)

48. ¿Cómo podemos usar los filtros ANOVA para seleccionar variables?

• Si la variable tiene una distribución similar para los posibles valores de la variable a predecir, seguramente no sirve para discriminar.

• Compararemos la media condicionada a los valores de la variable de salida.

• Para las variables que tengamos una confianza estadística elevada de que son iguales a lo largo de los valores de la variable dependiente, serán descartados.

• Para eso usaremos métodos ANOVA.

• En sklearn este método está implementado en la sección de "Univariate feature selection" el cuál contiene los métodos SelectKBest , SelectPercentile, SelectFpr (False positive rate test), SelectFdr (False discovery rate test), entre otros. Estos objetos reciben como parámetro la prueba estadística a utilizar, en particular ANOVA está implementado mediante f\_classif.

43. ¿Qué es el filtrado basado en las propiedades de la distribución? Da un ejemplo

*Feature selection* se refiere al proceso de seleccionar variables antes de ejecutar los algoritmos. En este sentido, filtrar basado en las propiedades de la distribución se refiere a seleccionar las variables explicativas considerando su distribución (Si hay poca variabilidad, no pueden ser usados para distinguir entre clases).

Se puede utilizar como medidas de variabilidad a la mediana y al inter-quartile range IQR, por ejemplo:

* Obtenga para cada variable su mediana.
* Obtenga para cada variable sus quartiles, en particular, reste el tercer quartil del primero, para obtener el IQR.
* Realice un scatter-plot entre ambas variables, esta gráfica nos da una visión de la distribución de las variables.
* Eliminemos las variables que tengan "baja variabilidad" i.e. que sean menores que un porcentaje del IQR global: e.g. <1/5 ó <1/6.

44. ¿En qué consiste el filtrado por correlación?

El filtrado por correlación se refiere a eliminar la variable que estén muy correlacionadas.

Problema: ¿Cuál tiras? No hay criterio establecido, a veces se puede tirar la mejor.

45. ¿En qué consiste el Fast-correlation filtering? Describe el algoritmo

Fast correlation-based filtering (Algoritmo):

* Encuentra una medida de relación entre cada par de variables.
* Encuentra la correlación de cada variable con la variable de salida.
* Ordena las variables según su correlación con la variable de salida.
* Elige la mejor variable (la de hasta arriba).
* Tira las variables muy correlacionadas con esta.
* Repite el proceso.

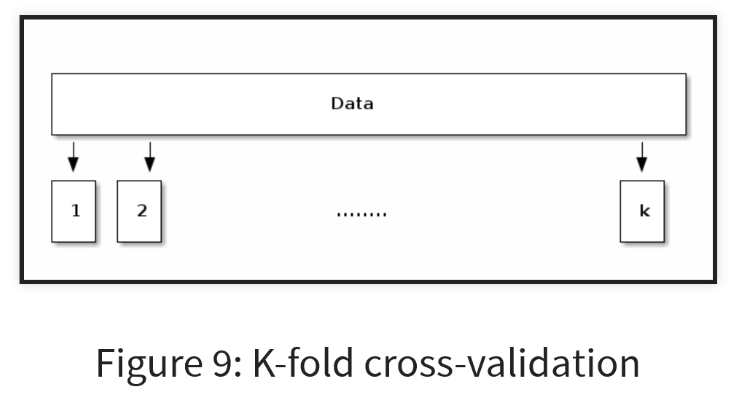
58. ¿Cómo puedo determinar si un modelo de clustering es adecuado?

La evaluación de modelos de clustering debe considerar la siguiente información:

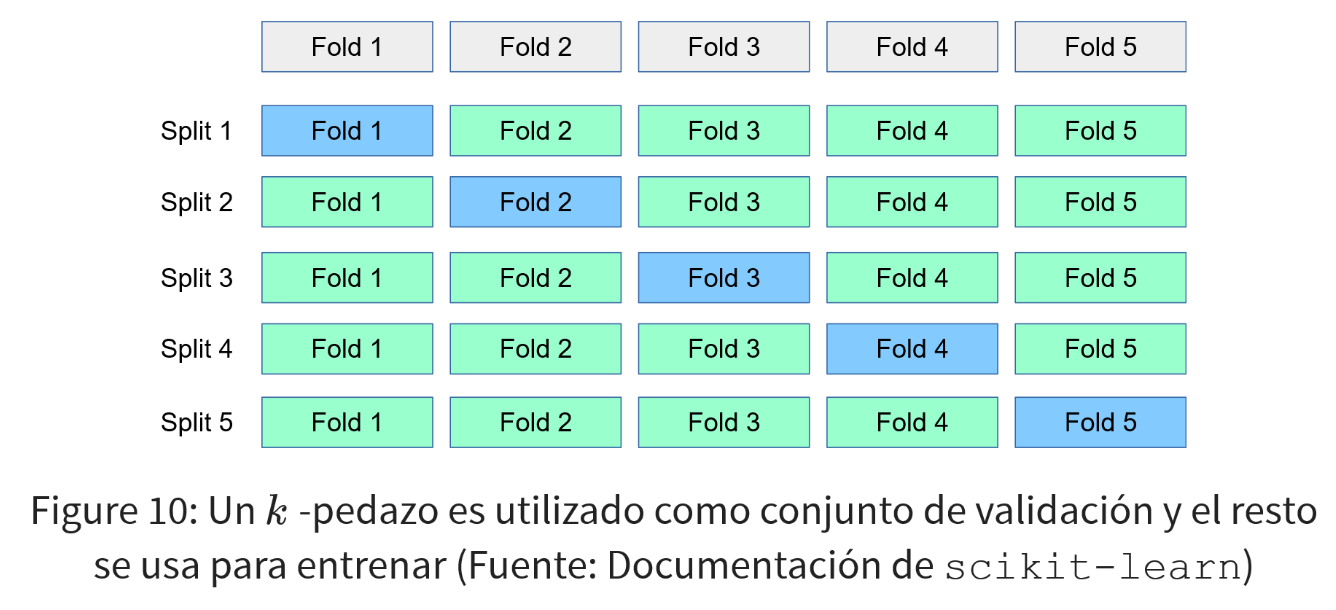
* Verificar resúmenes de la clusterización.
* Número de clusters.
* Número de observaciones por cluster:
  + hair clusters : Muy pocas observaciones.
  + waste clusters: Muchos puntos
* Comparar la distancia entre dos puntos en el cluster con la distancia típica entre dos clusters.

59. ¿En qué consiste el K-fold cross-validation? ¿Para qué sirve?

El *Cross-validation* es un método estadístico para evaluar la generalización del desempeño del modelo. Hay muchas variantes, la más común es k-fold cross-validation (Figura 9).



En k-fold cross-validation se divide el conjunto de entrenamiento en k pedazos. Para cada conjunto de hiperparámetros, un k -pedazo será usado como el conjunto de validación y los k−1pedazos restantes serán usados para entrenamiento (Figura 10).



El desempeño del modelo se toma como el promedio del desempeño en los k pedazos.

Una de las ventajas es la siguiente: Si hacemos hold-out tendremos, por ejemplo, 80% de los datos para entrenar y 20% para probar, si hacemos 10-fold cross-validation tendremos 90% de los datos para entrenar y 10% para probar.

La desventaja más grande es el costo computacional

############

60. ¿Se crean los modelos finales con cross-validation? Si/no y por qué.

############

61. ¿En qué consiste el leakage? Da un ejemplo

No utilizar validation set en hyperparameter tunning.

(Data Leakage is the creation of unexpected additional information in the training data, allowing a model or machine learning algorithm to make unrealistically good predictions.)

############

62. Da un ejemplo de cómo Caret puede ayudar en la selección de modelos

############

63. ¿Cuál es la filosofía UNIX para procesamiento de información?

############

64. Principales comando en UNIX

pwd - print working directory

ls - list

man - manual

> - redirect

cd - change directory

mkdir - make directory

rmdir - remove directory

############

65. ¿En que se parecen los “|” en UNIX con “%>%” en R

Llamadas a funciones de tal forma que se vea con un pipeline de ejecuciones.

############

66. ¿Qué diferencia hay entre una contenedor y una imagen en docker?

La imagen contiene las características de la máquina virtual.

Un ejecutable aislado y autónomo que contiene todo lo necesario para correr algún \_software\_, como código, librerías, binarios, variables de ambiente, etc. El \_template\_ de un contenedor.

Un contenedor es una instancia de una imagen. Una imagen una vez ejecutada resulta en un contenedor que corre la pieza de \_software\_ especificada en ella. Un contenedor está completamente aislado del anfitrión y de otros contenedores, accediendo sólo lo que se le comparte explícitamente.

############

67. ¿Para qué sirve docker?

- En desarrollo e investigación

\* Instalar y configurar es costoso

\* Manejar dependencias es doloroso

\* Un equipo puede tener diferencias vastas en sus ambientes de trabajo

\* "En mi máquina corre"

\* A pesar de herramientas como `virtualenv` y `packrat`, diferencias a nivel de sistema pueden causar dolores de cabeza

\* Diferencias en versiones de herramientas y tecnologías pueden causar problemas de compatibilidad

- En producción

\* Compartir y distribuir configuraciones y dependencias es difícil

\* "En mi máquina corre"

\* El ambiente de desarrollo o investigación nunca es exactamente igual al de producción

\* Distribuir cambios puede ser extremadamente doloroso

Docker nos permite resolver estos problemas, pues provee una manera sencilla y portable de \*\*construir y compartir ambientes completos\*\*, además de que corre nativamente en cualquier servidor con Linux (y algunos con Windows), permitiendo su \*\*completo acceso a los recursos del sistema\*\*. Es posible también crear una red virtual para instancias de Docker, permitiendo aislar y escalar fácilmente cada parte de la infraestructura.

Además, a través de DockerHub, Docker provee la capacidad de utilizar ambientes creados por la comunidad, si la herramienta o tecnología es relativamente popular, seguramente existe en DockerHub.

Por último, orquestadores de contenedores como Kubernetes y Swarm nos permiten \*\*manejar y escalar una infraestructura\*\* basada en Docker.

###########

68. ¿Cómo deep learning puede apoyar en el proceso de ingeniería de características?

###########

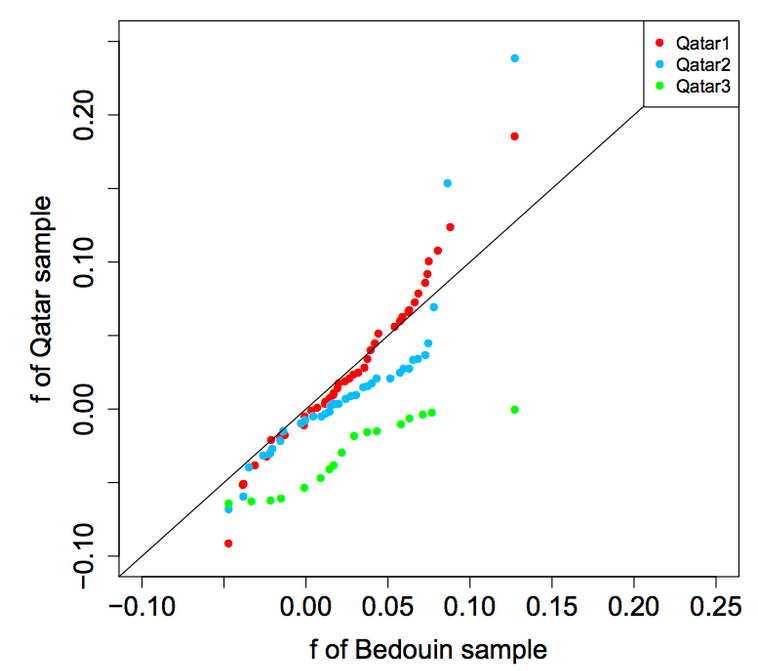
69. ¿Qué es lo que realizan las capas intermedias de las redes neuronales?

70.   Revisar el contenido de los artículos para realizar presentaciones ejecutivas y cómo contar una historia en ellas

71.   ¿Puedo utilizar un q-q plot para determinar si las distribuciones de dos variables continuas son semejantes? ¿Cómo?

Los gráficos Cuantil-Cuantil (Q-Q plots) permite observar cuan cerca está la distribución de un conjunto de datos a alguna distribución ideal ó comparar la distribución de dos conjuntos de datos.

Para comparar si las distribuciones de dos variables continuas son semejantes es necesario realizar un gráfico q-q plot de ambas variables sobre el mismo gráfico, por ejemplo:



**Loco** (Leave one coovariate out) : conjunto de reglas para separar cada árbol

Mecanismo para calcular la importancia de cada característica en el modelo al quitar cada una generalización del mean accuracy decrease.

49. Describe el método de selección Epsilon tanto para variables categóricas como numéricas

50. ¿Cómo puedo utilizar los métodos de clustering para realizar ingeniería de características.

**¿Qué son precision, recall, f1 score?**

Son medidas de desempeño del modelo ¿Cómo se calculan? ¿Cuándo utilizo cada una de esas medidas? Entender lo que el modelo debe hacer para

tener un desempeño aceptable es importante

**Precisión**:

El accuracy es un ejemplo de micro-average

(TP + TN)/(TP+FP+TN+FN)

-No sirve para dataset no balanceados (por ejemplo en detección de fraude).

-En este caso el modelo nulo es muy preciso (very accurate), pero obviamente esto no lo hace el mejor modelo.

-Cuando el modelo dice que el data point pertenece a la clase, que tan frecuentemente le atina.

-La precisión se ve afectada por la cantidad de falsos positivos (el sistema clasificó una observación erróneamente)

Especificidad **dism tasa de falsos neg**

1-Especificidad: 1-[TN/Neg=(FP+TN)]

**Recall/Sensibilidad pos bien clasificados**

TP/(TP+FN)=Pos

* ¿Qué fracción que están en la clase fueron detectadas por el modelo?
* Qué tan frecuentemente el clasificador encuentra lo que debe de encontrar.
* La proporción de observaciones clasificadas como \(C\) de todas las posibles que podían ser \(C\).
* Mide la capacidad del sistema de encontrar todas las observaciones relevantes
* El recall se ve afectado por la cantidad de falsos negativos (el sistema falló al clasificar una observación relevante)

**f1 score:** medida resumen armónica de Prec and recall

Muchas veces es importante contar con un solo número para poder optimizar el modelo

2\*(Precision\*Recall)/(Precision+Recall)

* Se usa en conjunto con precision y recall.
* Mide el sacrificio de recall y/o precision uno respecto al otro.
* Es una manera de resumir la curva de precision-recall
  + Análogo al AUC para la curva ROC

Mide el sacrificio de recall y/o precision uno respecto al otro.

* Es una manera de resumir la curva de precision-recall
  + Análogo al AUC para la curva ROC

**52. ¿Qué es el área bajo la curva? ¿Qué mide?**

La medida AUC (area under the curve) para un clasificador es el área bajo la curva generada por los pares sensibilidad-especificidad de la curva ROC.

Para examinar el desempeño del modelo en cuanto a su sensibilidad se utiliza una curva ROC (Receiver Operating Characteristic). El eje de las x representa el porcentaje (o proporción) de FPs y el eje de las y el porcentaje de TPs.

Cada punto en el gráfico representa la proporción FPs y TPs para una calificación dada. Notese que es acumulativo.

En base a esto podemos escoger el umbral

**53. ¿Cómo se construye la curva ROC?**

•La curva ROC muestra la sensibilidad del clasificador, al dibujar la tasa de verdaderos positivos (TRP) contra la tasa de falsos positivos (FPR). Para distintos valores de punto de corte

Muestra simultáneamente sensibilidad y especificidad

En lugar de examinar cada punto de corte por separado, podemos hacer el análisis de todos los posibles puntos de corte mediante la curva ROC (receiver operating characteristic, de ingeniería).

**54. ¿Cuándo es importante utilizar distintos puntos de corte para clasificar?  da dos ejemplos.**

**En qué te interesa equivocarte menos FN Cáncer**

En ocasiones los modelos de clasifiación dan una calificación (o probabilidad) de pertenencia a una

clase y por tanto la pertenencia de clase depende de un punto de corte (de la umbralización)

Por ejemplo:

* En la detección de fraudes por lo general se asigna una calificación entre cero y uno a cada transacción. El operador del sistema debe de decidir a partir de que valor se considera algo como fraude
* En el caso de detección de fraude se debe definir a partir de que “probabilidad” se recomienda tratamiento

**55. ¿Cómo interpreto la gráfica de doble densidad?**

Con la curva de doble densidad determinas la ponderación que le das a cada uno de los errores tomando como medida el p-value. Qué tanto estás poderando un error sobre otro. Donde se cruzan las dos distribucionesse asigna el mismo porcentaje de error a los errores tipo 1 y tipo 2

Si el modelo clasifica muy bien vemos dos distribuciones sin traslape y la curva ROC Cubriría un área muy grande

Entre más se traslapen las curvas, más se parecen los grupos ( e. sanos y enfermos) por tanto más baja la curva roc y el AUC

Mientras menos traslape mejor clasifica el modelo

**56. ¿Qué mide la R^2?**

La proporción de la varianza explicada por el modelo, entre mayor es mejor.

Qué tanto tu variable a predicir se explica por las variables explicadas

**¿Qué problemas tiene?**

Siempre que incluyes variables adicionales crece, no penaliza la inclusión de variables explicativas que son no significativas

Es igual a uno menos cuanta varianza no estamos explicando por el modelo

**57. ¿Qué mide el criterio de Akaike?**

El criterio de información de Akaike (AIC) es una medida de la calidad relativa de un modelo estadístico, para un conjunto dado de datos. Como tal, el AIC proporciona un medio para la selección del modelo. AIC maneja un trade-off entre la bondad de ajuste del modelo y la complejidad del modelo.

**58. ¿Cómo puedo determinar si un modelo de clustering es adecuado?**

* Son difíciles de evaluar \(\to\) verificar resumenes de la clusterización.
* Número de clusters
* Número de observaciones por cluster.
  + hair clusters : Muy pocas observaciones
  + waste clusters: Muchos puntos
* Compactos
  + Comparar la distancia entre dos puntos en el cluster con la distancia típica entre dos clusters.

**59. ¿En qué consiste el K-fold cross-validation? ¿Para qué sirve?**

* Es un método estadístico para evaluar la generalización del desempeño del modelo.
* Hay muchas variantes, la más común es k-fold cross-validation (Figura [9](file:///C:\ITAM\DataMining\ItroTodataScience\intro-to-data-science-2017-master\models\models.html#/slide-orgbde6935))
* En k-fold cross-validation se divide el conjunto de entrenamiento en \(k\) pedazos. Para cada conjunto de hiperparámetros, un \(k\) -pedazo será usado como el conjunto de validación y los \(k-1\) -pedazos restantes serán usados para entrenamiento (Figura [10](file:///C:\ITAM\DataMining\ItroTodataScience\intro-to-data-science-2017-master\models\models.html#/slide-orgf1d760c))
* El desempeño del modelo se toma como el promedio del desempeño en los \(k\) -pedazos
* Una de las ventajas es la siguiente: Si hacemos hold-out tendremos, por ejemplo, 80% de los datos para entrenar y 20% para probar, si hacemos 10-fold cross-validation tendremos 90% de los datos para entrenar y 10% para probar.
* La desventaja más grande es el costo computacional

**60. ¿Se crean los modelos finales con cross-validation? Si/no y por qué.**

No cross validation me sirve para encontrar los hiperparámetros del modelo.