Machine Learning

Unidad # 3 - Aprendizaje Supervisado Avanzado y Aprendizaje No Supervisado

CC57 – 2019-1

Profesor
Andrés Melgar

Competencias a adquirir en la sesión

- Al finalizar la sesión el alumno comprenderá el funcionamiento del **aprendizaje inductivo**.
- Al finalizar la sesión el alumno implementará **modelos algorítmos** usando algoritmos no supervisados.
- Al finalizar la sesión el alumno **entenderá** el algoritmo de **k-means**.
- Al finalizar la sesión el alumno **aplicará** el algoritmo de **k-means** para obtener modelos algorítmicos.

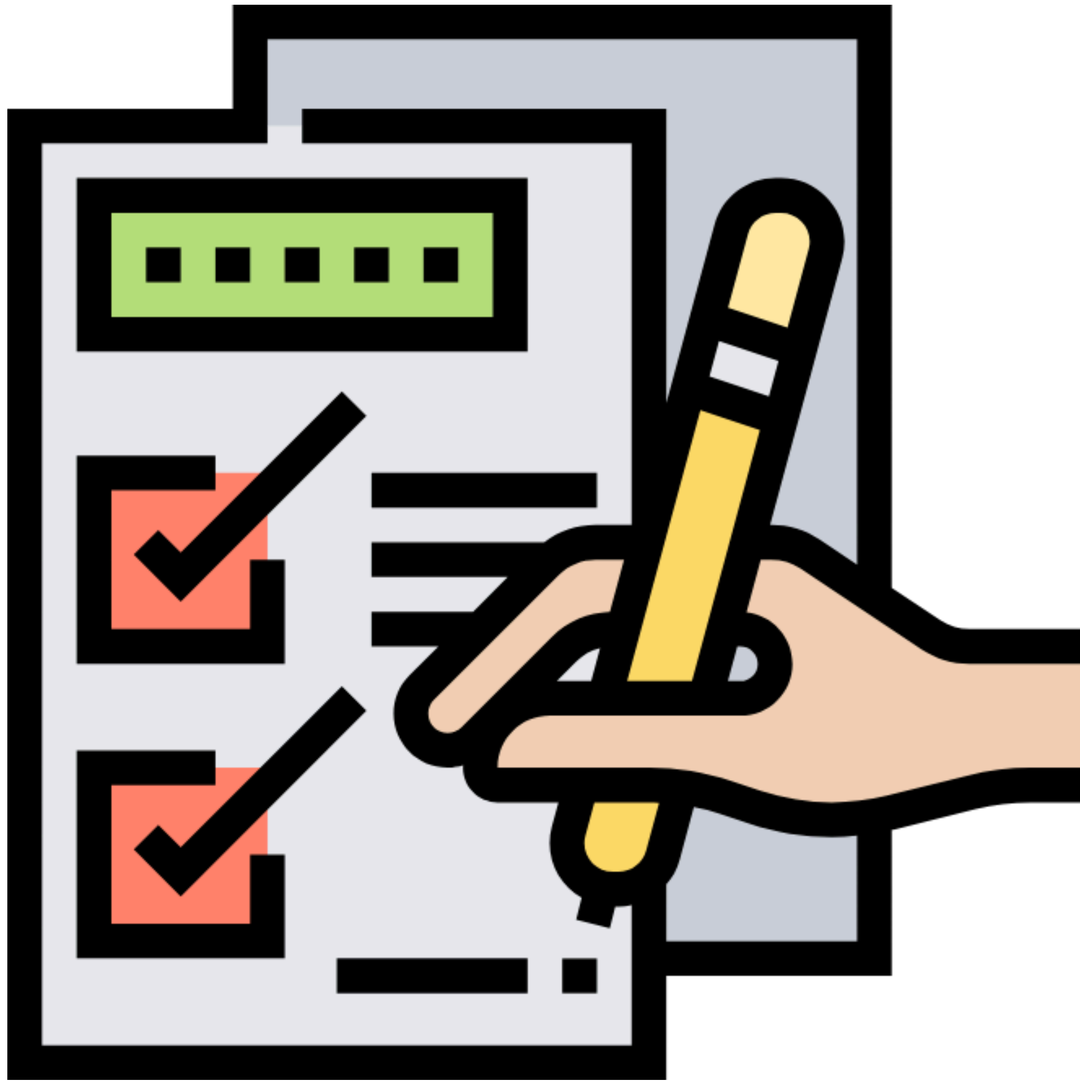


Imagen obtenida a través de la URL https://www.flaticon.com/free-icon/checklist_1467445

Métricas de Evaluación

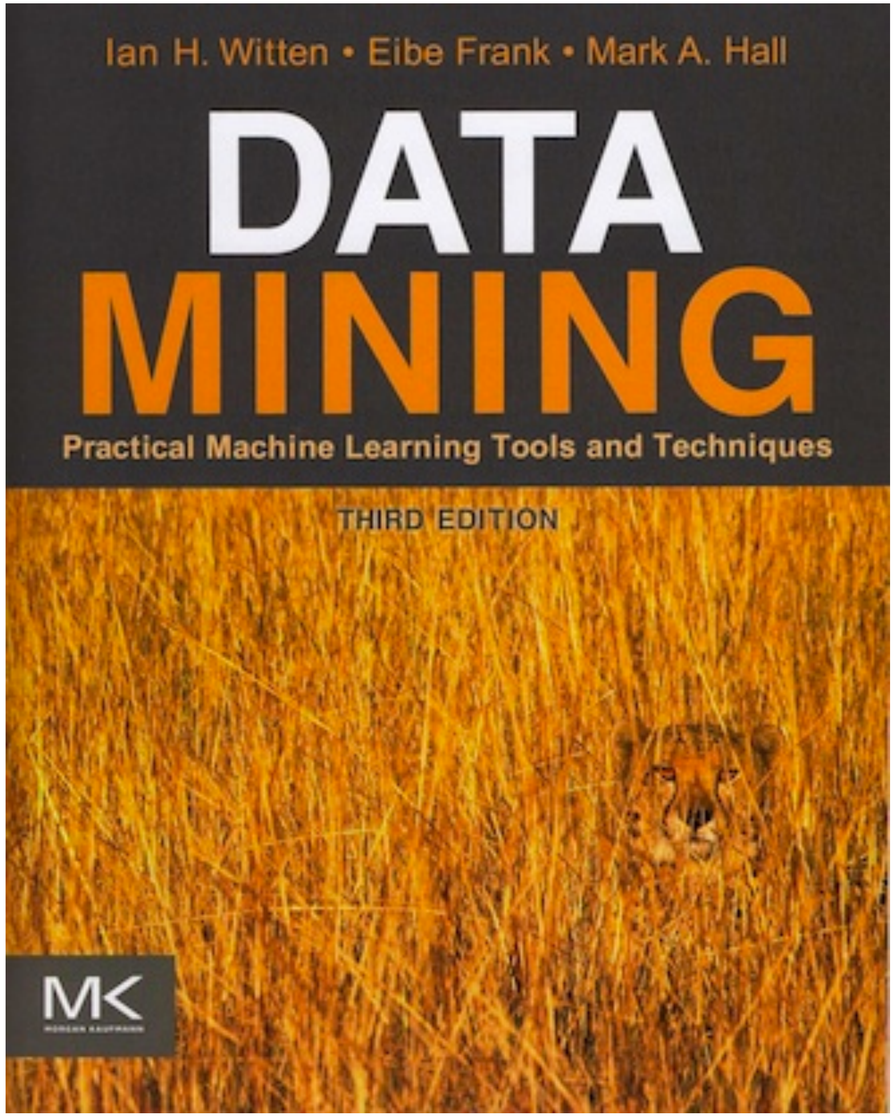
Texto guía

Witten, Ian H., Frank, Eibe, and Hall, Mark A.. 2011. ***Data Mining : Practical Machine Learning Tools and Techniques with Java Implementations.*** San Francisco: Elsevier Science & Technology.

CHAPTER 4

Algorithms: The Basic Methods

4.8 CLUSTERING



Clustering

- Video inducción
- Vea el video e indique:
  - ¿Qué entiende por clustering?
  - ¿Para qué sirve?
  - ¿Qué tipos de clustering existen?

https://www.youtube.com/watch?v=2QTeuO0C-fY

clustering - Microsoft PowerPoint

Home Insert Design Animations Slide Show Review View Add-Ins

Slides Outline

1 Clustering

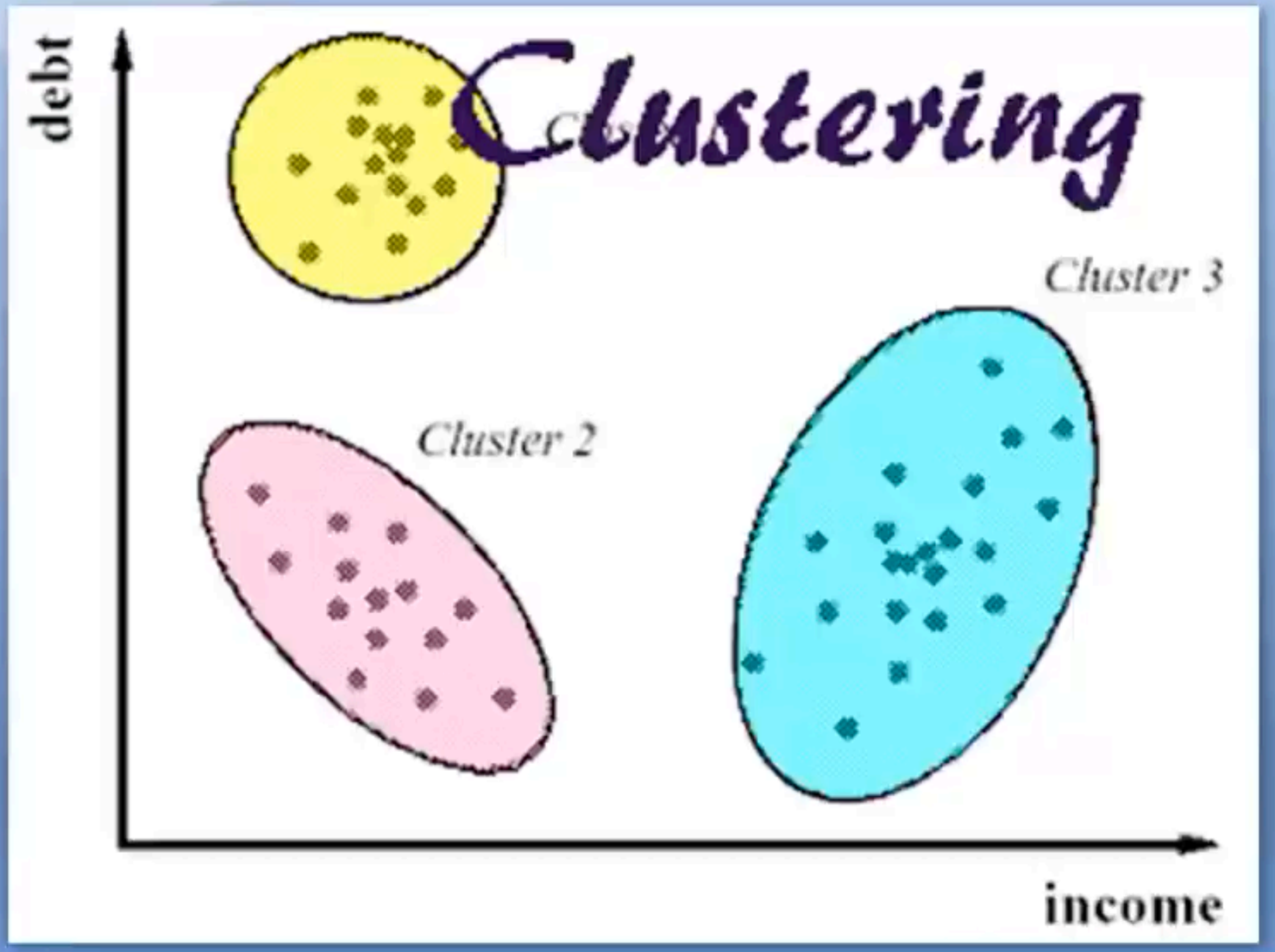
2 WHAT IS CLUSTERING

3 STAGES OF CLUSTERING

4 DIFFERENT REPRESENTATIONS

5 APPLICATIONS OF CLUSTERING

6 EXAMPLES OF CLUSTERING



Click to add notes

Custom Animation

Add Effect Remove

Modify effect

Start:

Property:

Speed:

Select an element of the slide, then click "Add Effect" to add animation.

Re-Order

Play Slide Show

AutoPreview

Slide 1 of 9 "Office Theme"

81%

1:39 PM
7/19/2015

Clustering

- Responda
  - ¿Qué entiende por clustering?
  - ¿Para qué sirve?
  - ¿Qué tipos de clustering existen?

https://www.youtube.com/watch?v=2QTeuO0C-fY

Clustering

- Las técnicas de clustering se aplican cuando **no existe una clase** para predecir pero las instancias naturalmente **se dividen en grupos**
  - Estos grupos, reflejan algún **mecanismo** que actúa en el dominio de las instancias, un mecanismo que hace que algunos casos sean más parecidos entre sí que lo restantes.
  - Clustering requiere naturalmente de **diferentes técnicas** a los métodos de aprendizaje de clasificación que hemos estudiado hasta ahora.

Clustering

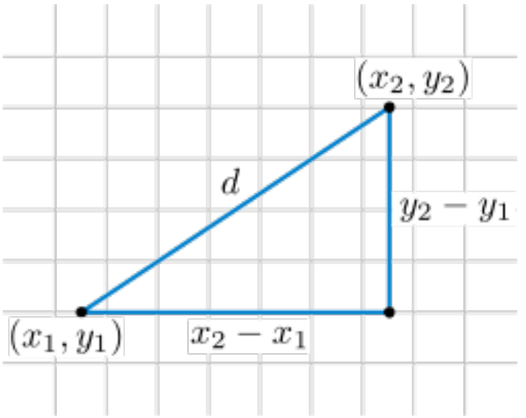
- Hay diferentes formas de expresar el resultado de la agrupación.
  - Los grupos que se identifican pueden ser **exclusivos**. Cualquier instancia pertenece a un solo grupo.
  - O pueden estar **superpuestos**. Una instancia puede pertenecer a varios grupos.
  - O pueden ser **probabilísticos**. Una instancia pertenece a cada grupo con una cierta probabilidad.
  - O pueden ser **jerárquicos**. Una división aproximada de casos en grupos en el nivel superior, siendo refinados en los niveles inferiores hasta llegar a las instancias individuales.

Clustering

- En realidad, la elección entre estas posibilidades debe ser dictada por la naturaleza de los mecanismos que se cree que subyacen en el fenómeno de agrupamiento en particular.

- Sin embargo, debido a que estos mecanismos son raramente conocidos - la existencia misma de los grupos es, después de todo, algo que estamos tratando de descubrir- y por razones pragmáticas también, la elección suele estar dictada por las herramientas de agrupación que están disponibles.

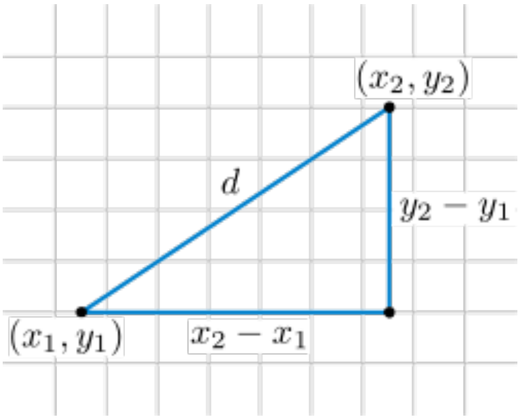
k-means



- K-means es la técnica de agrupación clásica
  - En primer lugar, se especifica de antemano la cantidad de grupos que se desean obtener: este es el parámetro k.
  - Luego, k puntos son elegidos al azar como centros de grupo.
  - Todos las distancias son asignadas a su centro más cercano de acuerdo con la distancia euclidiana.
  - A continuación, se calcula el centro de gravedad, o media (means) de los instancias en cada grupo.
  - Estos centroides se toman como nuevos valores de centro en sus respectivos grupos.
  - Finalmente, todo el proceso se repite con los nuevos centros de los grupos.
  - La iteración continúa hasta que los mismos puntos se asignan a cada grupo en bucles consecutivos.

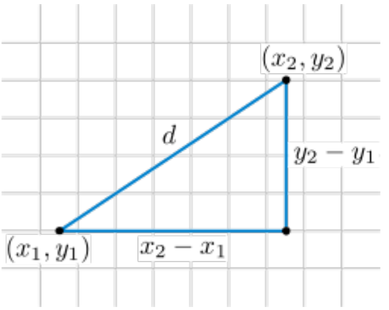
Machine Learning

k-means



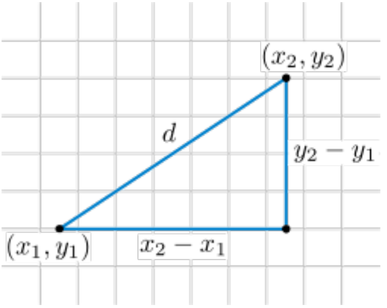
- Este método de agrupamiento es simple y eficaz.
  - Es fácil demostrar que el centro del cluster **minimiza la distancia al cuadrado total** de cada uno de los puntos del cluster a su centro.
  - Una vez que la iteración se ha estabilizado, a cada punto se le asigna su **centro más cercano**, por lo que el efecto general es minimizar la distancia al cuadrado total de todos los puntos a sus centros.

k-means



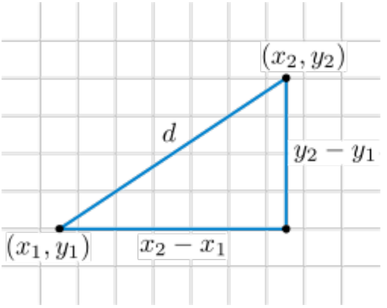
- Sin embargo, esta **distancia mínima es local**; no hay ninguna garantía de que sea global
  - Los grupos finales son bastante sensibles a los **centros iniciales**.
  - Arreglos de datos completamente diferentes pueden originarse debido a pequeños cambios en la elección aleatoria inicial.
- Para aumentar la probabilidad de encontrar una solución global, a menudo se ejecuta el algoritmo varias veces con diferentes opciones iniciales y se elije el mejor resultado final, el que tiene la distancia al cuadrado total más corta.

k-means



- Es fácil imaginar situaciones en las que k-means no encuentra un buen agrupamiento.
  - Considere por ejemplo cuatro instancias dispuestas en los vértices de un rectángulo en el espacio de dos dimensiones.
  - Hay dos grupos naturales, formados mediante la agrupación de los dos vértices en cada extremo de un lado corto.
  - Pero supongamos que los dos centros iniciales pasan a caer en los puntos medios de los lados largos.
  - Esto forma una configuración estable. Los dos grupos contienen cada uno las dos instancias en cada extremo de un lado largo, no importa cuán grande es la diferencia entre el largo y los lados cortos.

k-means



- k-means puede mejorar drásticamente mediante una cuidadosa selección de los centros iniciales, a menudo llamados semillas (seed)

- En lugar de comenzar con un conjunto arbitrario de semillas, que aquí hay una mejor procedimiento.

  - Elija la semilla inicial al azar de todo el espacio, con una distribución de probabilidad uniforme.

  - A continuación, elija la segunda semilla con una probabilidad que es proporcional al cuadrado de la distancia de la primera.

  - Proceder, en cada etapa de la elección siguiente con una semilla con una probabilidad proporcional al cuadrado de la distancia desde la semilla más cercana que haya sido elegida.

k-means

- Use el algoritmo de k-means para encontrar grupos en los conjuntos de datos
  - iris
  - weather
  - vote

Competencias a adquirir en la sesión

- Al finalizar la sesión el alumno comprenderá el funcionamiento del **aprendizaje inductivo**.
- Al finalizar la sesión el alumno implementará **modelos algorítmos** usando algoritmos no supervisados.
- Al finalizar la sesión el alumno **entenderá** el algoritmo de **k-means**.
- Al finalizar la sesión el alumno **aplicará** el algoritmo de **k-means** para obtener modelos algorítmicos.

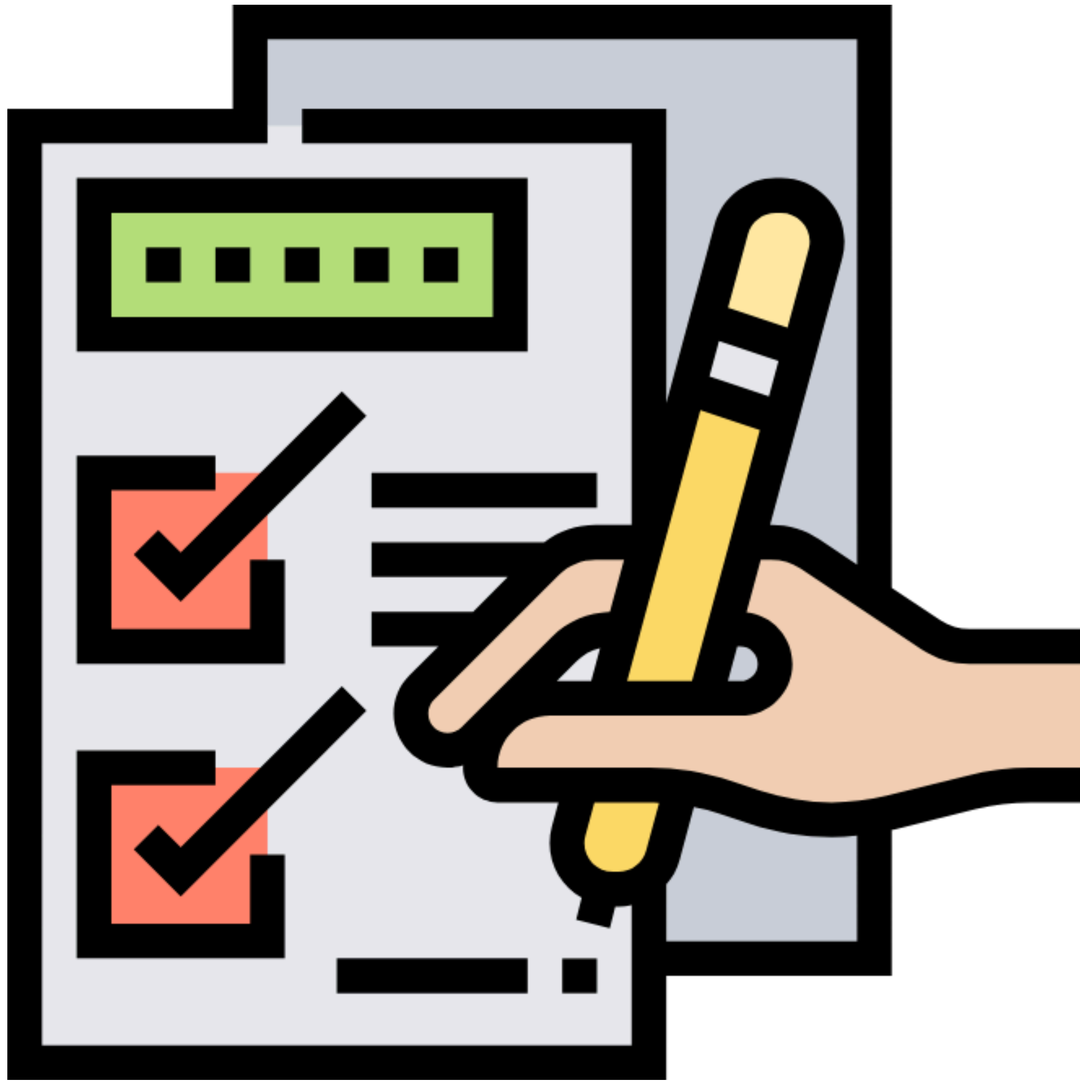


Imagen obtenida a través de la URL https://www.flaticon.com/free-icon/checklist_1467445