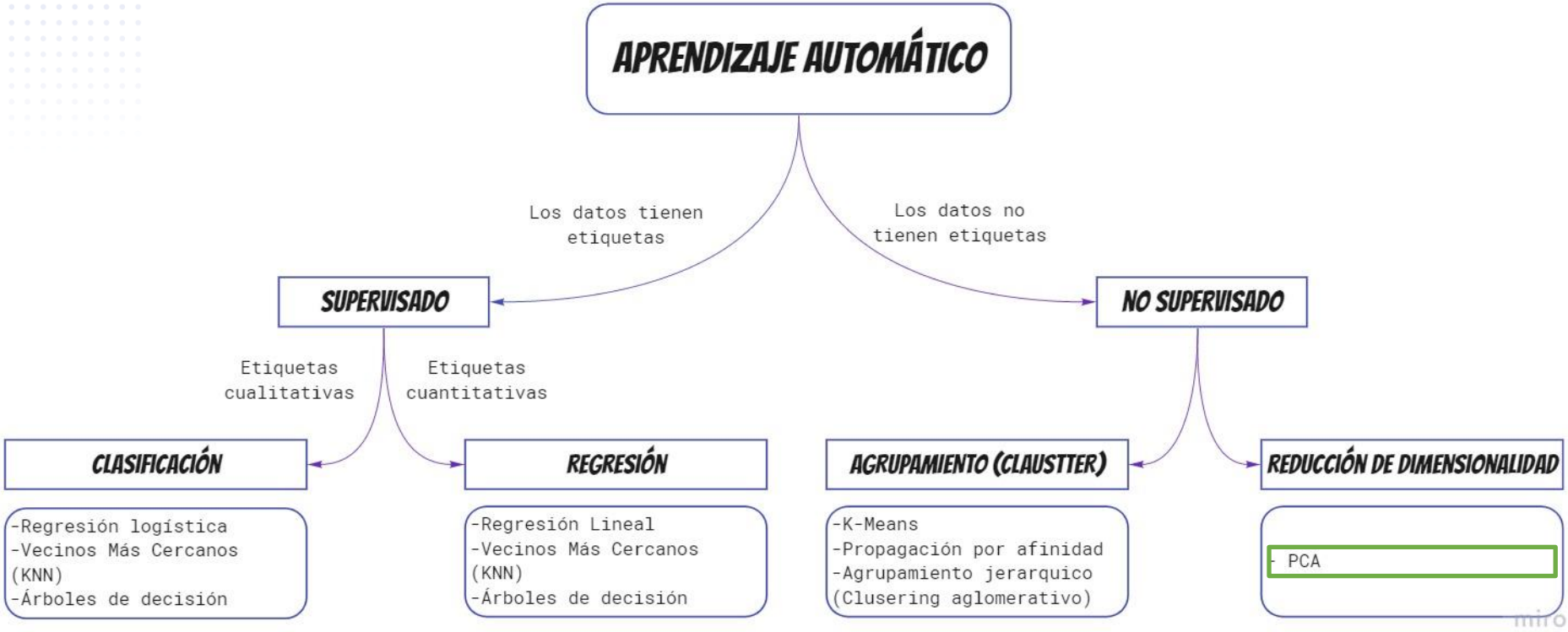
Ciencia de Datos

- **Módulo 4**

PCA

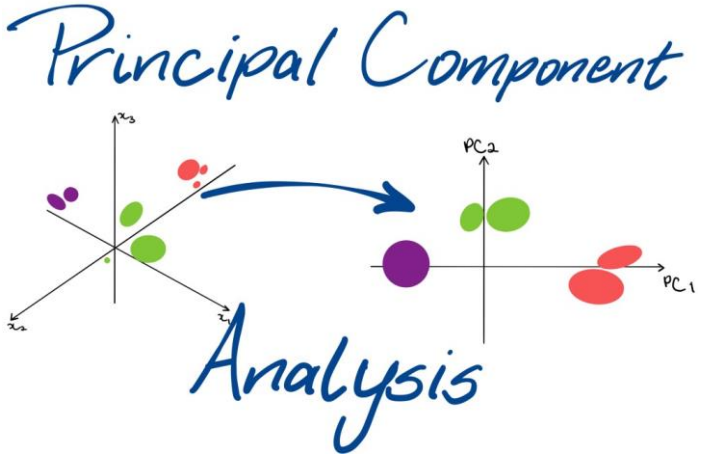
UTNFra GYE Gestión y Evaluación Educativa «aprendé programando Virtual» SECUNDARIA del futuro


APRENDIZAJE AUTOMÁTICO
Los datos tienen etiquetas
Los datos no tienen etiquetas
SUPERVISADO
NO SUPERVISADO
Etiquetas cualitativas
Etiquetas cuantitativas
CLASIFICACIÓN
REGRESIÓN
AGRUPAMIENTO (CLAUSTTER)
REDUCCIÓN DE DIMENSIONALIDAD
-Regresión logística
-Vecinos Más Cercanos (KNN)
-Árboles de decisión
-Regresión Lineal
-Vecinos Más Cercanos (KNN)
-Árboles de decisión
-K-Means
-Propagación por afinidad
-Agrupamiento jerarquico (Clusering aglomerativo)
- PCA

PCA

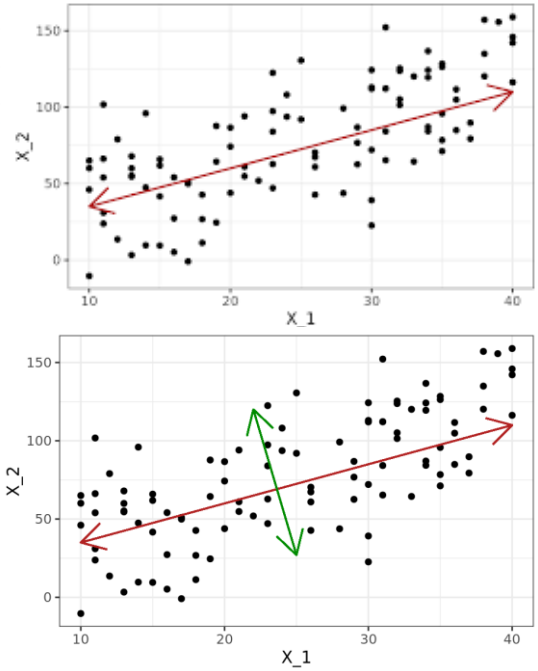
Más conocido como PCA (por sus siglas en inglés Principal Components Analysis), es un algoritmo de aprendizaje no supervisado, comúnmente usado para lo que llamamos reducción de dimensionalidad.

En muchos problemas de la vida real se tienen una gran cantidad de variables. En estos casos puede no ser conveniente incluir todas las variables, y decidir cuáles eliminar es un problema en si, este problema es el que trata de resolver PCA.



Fuente: https://laptrinhx.com/understanding-principal-component-analysis-1459875749/

PCA



Fuente: https://rpubs.com/Joaquin_AR/287787

PCA busca eliminar variables afectando lo menos posible la variabilidad de los datos, esto es, mantener la mayor cantidad de casos existentes, para esto lo que se hace es aplicar **transformaciones** a las variables (descomponerlas), convirtiendo a las mismas en lo que llamamos **componentes**.

Cada componente aporta un grado de variabilidad al conjunto de datos, estas componentes están ordenadas de forma tal que la primera componente es aquella que mantiene la mayor variabilidad del conjunto, y la última la que menor variabilidad aporta, luego nos quedamos con las N primeras componentes.

Además estos componentes son independientes entre sí, siendo que puede suceder que algunas variables originales no lo sean (esto es muy importante en algunos modelos)

PCA

Fuente: https://www.seoperu.com/los-mejores-procesadores-para-servidores/

PCA se utiliza para:

- Comprimir los datos lo que es importante en términos de memoria y procesamiento. Además sirve para reducir el número de variables sin identificar cuales son las más explicativas, permite descartar información redundante o ruido lo que puede ser importante para la generalización
- Visualizar en 2 dimensiones de conjuntos de datos.

El mayor inconveniente con la utilización de PCA es que limita de un manera significativa la interpretabilidad de los datos

PCA – Código

En Python con la librería Scikit-Learn se utiliza:

- Método *.fit* para entrenar el modelo
- Método *.transform* para transformar los datos.

Es posible utilizar todo en una misma línea utilizando el método *.fit_transform*

La cantidad de componentes las podemos seleccionar de forma directa definiéndolo como **hiperparámetros** (decir cuántas componentes se quieren trabajar) o seleccionar la cantidad de componentes a mantener tal que se mantenga cierto porcentaje de variabilidad del conjunto de datos.

El máximo de componentes es el mínimo entre la cantidad de registros o la cantidad de variables.

Solamente se utiliza para **variables numéricas**. Para variables categóricas se utilizan otras técnicas (análisis de correspondencias y categorical PCA)

Este es un modelo NO supervisado por lo tanto no se desea predecir ninguna etiqueta. Solamente se **transforman** los datos que se están utilizando.

PCA

Vídeo

- https://www.youtube.com/watch?v=_UVHneBUBW0

Documentación Scikit-Learn

- https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html

Adicional

- https://www.youtube.com/watch?v=HMOI_lkzW08
- https://empresas.blogthinkbig.com/python-para-todos-que-es-el-pca/
- https://www.jacobsoft.com.mx/es_mx/analisis-del-componente-principal/