UNIVERSIDAD NACIONAL DE COLOMBIA
Facultad de Ingeniería, sede Bogotá
Curso virtual *Machine Learning and Data Science* (MLDS)

Glosario

A	**Agrupamiento** **"**Un algoritmo de agrupamiento (en inglés, clustering) es un procedimiento de agrupación de una serie de vectores de acuerdo con un criterio. Esos criterios son por lo general distancia o similitud." (Algoritmo de agrupamiento, s.f.). **Análisis de datos** **"**El análisis de datos es un proceso que consiste en inspeccionar, limpiar y transformar datos con el objetivo de resaltar información útil, para sugerir conclusiones y apoyo en la toma de decisiones. El análisis de datos tiene múltiples facetas y enfoques, que abarca diversas técnicas en una variedad de nombres, en diferentes negocios, la ciencia, y los dominios de las ciencias sociales. Los datos se coleccionan y analizan para indagar en cuestiones, probar conjeturas o probar la invalidez de teorías." (Análisis de datos, s.f.) **Análisis exploratorio de datos** "El análisis exploratorio de datos es una forma de analizar datos definido por John W. Tukey (EDA: Exploratory data analysis) como el tratamiento estadístico al que se someten las muestras recogidas durante un proceso de investigación en cualquier campo científico. Para mayor rapidez y precisión, todo el proceso suele realizarse por medios informáticos, con aplicaciones específicas para el tratamiento estadístico." (Análisis exploratorio de datos, s.f.). **Aprendizaje automático (*machine learning*)** **"**El aprendizaje automático o aprendizaje automatizado o aprendizaje de máquinas (del inglés, machine learning) es el subcampo de las ciencias de la computación y una rama de la inteligencia artificial, cuyo objetivo es desarrollar técnicas que permitan que las computadoras aprendan (…)" (Aprendizaje automático, s.f.). Este aprendizaje se genera a partir de datos o de la experiencia. **Aprendizaje no supervisado** **"**Aprendizaje no supervisado es un método de Aprendizaje Automático donde un modelo se ajusta a las observaciones. Se distingue del Aprendizaje supervisado por el hecho de que no hay un conocimiento a priori." (Aprendizaje no supervisado, s.f.). En este tipo de aprendizaje los datos de entrada no se dividen en datos de entrada y salida.

	Aprendizaje supervisado **“**En aprendizaje automático y minería de datos, el aprendizaje supervisado es una técnica para deducir una función a partir de datos de entrenamiento. Los datos de entrenamiento consisten de pares de objetos (normalmente vectores): una componente del par son los datos de entrada y el otro, los resultados deseados.” (Aprendizaje supervisado, s.f.).
B	***Big data*** Campo de la tecnología de la información que busca diferentes maneras de analizar y, de forma sistemática, extraer o tratar información de conjuntos de datos muy grandes o complejos, para ser tratados por aplicaciones de *software* de procesamiento de datos tradicionales. **Base de datos NoSQL** Bases de datos no relacionales que permiten almacenar, procesar y consultar datos que no tienen un esquema completamente definido.
C	**Clasificación** En aprendizaje automático y en estadística, la clasificación es el problema de identificar a cuál de un conjunto de categorías (subpoblaciones) pertenece una nueva observación, sobre la base de un conjunto de datos de formación que contiene observaciones (o instancias), cuya categoría de miembros es conocida. ***Cross Industry Standard Process for Data Mining*** (**CRISP-DM**) Se trata de un modelo estándar abierto del proceso que describe los enfoques comunes que utilizan los expertos en minería de datos. Es el modelo analítico más usado.
D	***Data science*** Campo interdisciplinario que usa métodos científicos, procesos, algoritmos y sistemas para extraer conocimiento de datos estructurados y no estructurados. Está relacionado con la minería de datos y con *big data*. **Datos estructurados** Datos que tienen una estructura (esquema) claramente definida. Típicamente se encuentran en las bases de datos situadas en el paradigma Entidad-Relación. **Datos no estructurados** Datos que no tienen una estructura definida, por ejemplo, imágenes, texto, videos, audio, entre otros. **Datos semiestructurados** Datos que tienen un híbrido entre los estructurados y no estructurados.

H	***Hadoop*** *Framework* de acceso abierto que implementa el modelo de procesamiento *MapReduce*. Este utiliza un sistema de archivos distribuido llamado HDFS (por sus siglas en inglés: *Hadoop Distributed File System*). ***HBase*** Base de datos orientada a columnas que se realiza en HDFS y soporta operaciones *MapReduce.* ***Hadoop Distributed File System* (HDFS)** Sistema distribuido de archivos típicamente utilizado por sistemas basados en *Hadoop*. ***Hive*** Sistema de almacenamiento de datos basado en HDFS que sirve para manejar grandes volúmenes de datos. Provee características que permiten, mediante sentencias similares a las de *Structured Query Language* (SQL), la manipulación de los datos.
M	**Mapas coropléticos** Un mapa coroplético, coropleto o de coropletas es un mapa temático en el que las regiones se colorean de un motivo que muestra una medida estadística, como puede ser la densidad de población o el ingreso por habitante. Este tipo de mapa facilita la comparación de una medida estadística de una región con la de otra o muestra la variabilidad de esta para una región dada. ***MapReduce*** Modelo de procesamiento distribuido de datos en el que participan varios nodos. Un programa que se ejecuta bajo este modelo se compone de dos partes: la función *map,* que se encarga de filtrar y ordenar los datos, y la función reduce, que realiza operaciones de agregación sobre los datos. ***Matplotlib*** Biblioteca para la generación de gráficos, a partir de datos contenidos en listas o *arrays* en el lenguaje de programación *Python* y su extensión matemática *NumPy*. **Motor de búsqueda** Sistema que permite crear un índice a partir del contenido de un repositorio de datos, con el cual se pueden realizar búsquedas por el contenido de dichos datos. Por ejemplo, Google, Bing, Yahoo son motores de búsqueda de texto.
N	***NumPy***

	Extensión de *Python* que le agrega mayor soporte para vectores y matrices, con lo cual se constituye una biblioteca de funciones matemáticas de alto nivel para operar con dichos vectores o matrices.
P	***Plotly*** Biblioteca de gráficos para *Python* que permite crear gráficos interactivos con calidad de publicación. Posibilita la construir de gráficos de líneas, dispersión, áreas, barras, histogramas, mapas de calor, subtramas, ejes múltiples, gráficos polares y gráficos de burbujas. ***Python*** Lenguaje de programación interpretado cuya filosofía hace énfasis en la legibilidad de su código. Se trata de un lenguaje de programación multiparadigma, ya que soporta orientación a objetos, programación imperativa y, en menor medida, programación funcional. Este lenguaje es interpretado, dinámico y multiplataforma; posee una licencia de código abierto denominada *Python Software Foundation License*.
S	***Seaborn*** Biblioteca de visualización de datos de *Python* basada en *matplotlib.* Proporciona una interfaz de alto nivel para dibujar gráficos estadísticos atractivos e informativos. **Sobreajuste** En aprendizaje automático, el sobreajuste u *overfitting* es el efecto de sobreentrenar un algoritmo de aprendizaje con unos ciertos datos para los que se conoce el resultado deseado.
R	**Reducción de dimensionalidad** En aprendizaje automático y estadísticas, la reducción de dimensionalidad o reducción de la dimensión es el proceso de reducción del número de variables aleatorias que se trate. **Regresión** El análisis de la regresión es un proceso estadístico para estimar las relaciones entre variables. Incluye técnicas para el modelado y análisis de diversas variables, cuando la atención se centra en la relación entre una variable dependiente y una o más variables independientes o predictoras
V	**Validación cruzada** También denominada *cross-validation* es una técnica utilizada para evaluar los resultados de un análisis estadístico y garantizar que son independientes de la partición entre datos de entrenamiento y prueba. **Visualización de datos** Proceso de búsqueda, interpretación, contraste y comparación de datos que permite un conocimiento en profundidad y detalle de los mismos, de tal forma que se transformen en información comprensible para el usuario.

Referencias

Algoritmo de agrupamiento. (2019, 30 de julio). En *Wikipedia*. https://es.wikipedia.org/w/index.php?title=Algoritmo_de_agrupamiento&oldid=117867304

Análisis de datos. (2020, 26 de abril). En *Wikipedia.* https://es.wikipedia.org/w/index.php?title=An%C3%A1lisis_de_datos&oldid=125538114

Aprendizaje automático. (2020, 21 de abril). En *Wikipedia.* https://es.wikipedia.org/w/index.php?title=Aprendizaje_autom%C3%A1tico&oldid=125365936

Análisis exploratorio de datos. (2019, 12 de julio). En *Wikipedia.* https://es.wikipedia.org/w/index.php?title=An%C3%A1lisis_exploratorio_de_datos&oldid=117373471.

Aprendizaje no supervisado. (2019, 20 de noviembre). *Wikipedia.* https://es.wikipedia.org/w/index.php?title=Aprendizaje_no_supervisado&oldid=121459999.Apr

Aprendizaje supervisado. (2020, 19 de enero). En *Wikipedia.* https://es.wikipedia.org/w/index.php?title=Aprendizaje_supervisado&oldid=122881109.