Programa de Captura de Datos

MÓDULO 1: TÉRMINOS BÁSICOS

En el presente modulo se abordarán los términos básicos asociados a la Ciencia de Datos, garantizando así un correcto punto de partida para la comprensión del resto de los módulos. Por ello el contenido es el siguiente:

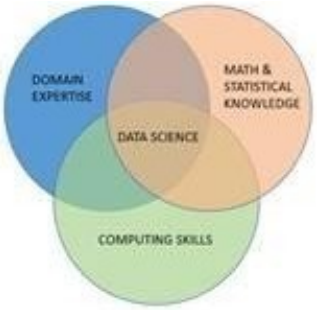
1. Ciencia de Datos.
2. BI.
3. Machine Learning.
4. Deep Learning.
5. Big Data y su ámbito de aplicación.
6. IOT.
7. Inteligencia Artificial.

MÓDULO 2: CONCEPTOS BÁSICOS

Este módulo proporcionará conocimiento acerca de que datos recoger y como almacenar esta información. Para ello se mostrarán los siguientes conceptos:

1. Almacenamiento de Datos
2. Dataset.
    a. Posibles fuentes (texto, Excel, logs, bases de datos, Salesforce…)
    b. Fila, columna, identificador único (como organizar una base de datos)
3. Consistencia, Restricciones (primary key, foreign key…)
4. Otras fuentes: API, Web Scrapping, NLP, RRSS
5. Calidad del Dato:
    a. Tipos de variables.
    b. Instancias.
    c. Sesgos.
    d. Dimensiones de la calidad del dato.
    e. Las 5 "V" del Dato.
6. Conceptos estadísticos básicos:
    a. Máximo.
    b. Mínimo.
    c. Rango.
    d. Media.
    e. Mediana.
    f. Cuantiles.
    g. Rango Inter-cuartil.
    h. Modelo.
    i. Concepto de preprocesamiento.
    j. EDA
    k. Train/Test/Validación.
    l. Balanceo de los datos.
    m. Algoritmos supervisados y no supervisados.

n. Modelos de Regresión y Clasificación
o. Matriz de Confusión y métricas asociadas.
p. Coeficiente de Determinación y Error Cuadrático Medio.
q. Expertos del dominio:
    i. Necesidad de comprender el dato y tener una visión de "negocio".
    ii. Habilidades necesarias para la correcta realización de la ciencia de datos:



Interacción entre áreas para la ciencia de datos.

7. Gráficos básicos:

    a. Histograma.
    b. Gráfico de Barras.
    c. Scatterplot.
    d. Boxplot.
    e. Pairplot
    f. Otros.

Visualización de ejemplos a través de Power BI, donde se mostrarán ejemplos visuales de los conceptos anteriormente citados.

MÓDULO 3: BASES DE DATOS

En este módulo se mostrarán los principales tipos de bases de datos, así como las principales Bases de Datos de cada tipo y sus características, por ello se incluirán los siguientes aspectos:

1. Bases de Datos Relacionales.
2. Bases de Datos No Relacionales
3. Ejemplos de cada caso.
4. Ventajas y Desventajas.
5. Conexión a bases de datos.

MÓDULO 4: SISTEMAS DE ALMACENAMIENTO

En este apartado se explicarán los principales sistemas de almacenamiento existentes en la actualidad, lo que incluye los siguientes sistemas:

1. CRM.
2. ERP.
3. Data Lake.
4. Data WareHouse.
5. Cluod/On Premise.
6. Ejemplos Prácticos.

MÓDULO 5: PREPROCESAMIENTO

El preprocesamiento de los datos es una de las tareas que más tiempo requieren en cualquier proceso de ciencia de datos. Pese a que esta fase es necesaria en todos los proyectos, el 90% del tiempo empleado viene provocado por una mala planificación en la recogida de datos o una mala ejecución de la misma. En caso de que se logre planificar de una manera adecuada la recogida de datos, siendo esta coherente con los posteriores análisis que se van a realizar se puede reducir notablemente el tiempo de ejecución.

Por ello se mostrarán las principales tareas que se realizan en el preprocesamiento de los datos y como una correcta planificación de la recogida afectaría a estas tareas. Los conceptos a desarrollar son los siguientes:

1. Homogeneización y estandarización de las variables.
2. Nivel de granularidad.
3. Corrección de errores.
4. Outliers.
5. Valores Perdidos.
6. Selección de Variables:
    a. Correlación.
    b. Variabilidad.
    c. Tabla ANOVA.
    d. Filtros, Wrapper y Embebidas.
    e. Selección mediante modelos de Regresión y Árbol de Decisión.
    f. Selección de Instancias.
    g. Etiquetado de Datos.
    h. Trazabilidad.

MÓDULO 6: GOBERNANZA DE DATO

En este módulo se mostrará la evolución que debe seguir una compañía para ser una compañía que explota el dato de manera óptima. Por ello se abordarán los siguientes aspectos:

1. Introducción al Gobierno del Dato.
2. Órganos, Roles y Responsabilidades.
3. Funciones y Actividades.
4. Herramientas.
5. Buenas Prácticas.
6. Seguridad Física/Lógica.
7. RGPD.

MÓDULO 7: APLICACIONES PRÁCTICAS

En este módulo se mostrarán aplicaciones prácticas, con ejemplo accesibles a los alumnos donde puedan observar, sin disponer de conocimientos de programación, diferentes casos de uso asi como de un análisis EDA, previo al desarrollo de cualquier modelo. Junto a ello se dispone de una serie de códigos de Python donde se resuelven los casos de uso y una serie de videos donde se explica como instalar Anaconda y los códigos incluidos en este módulo. Por lo que el contenido es el siguiente:

1. Posibles casos de aplicación en distintos sectores
2. Descarga, instalación y configuración de Anaconda y entorno.
3. Desarrollo de casos prácticos
4. Ejemplo de EDA.

Garantizando así la existencia de ejemplos de las diferentes empresas que se han presentado a anteriores ediciones del laboratorio de ciencia de datos y atendiendo a la estrategia s4.

MÓDULO 8: MODELOS Y CASOS DE USO

En este módulo se realizarán diversos sistemas que permitirán a los asistentes visualizar las principales salidas y representaciones graficas de los modelos más importantes. Para cada uno de ellos existirá un video explicativo, una documentación y ejemplos prácticos donde podrán interactuar con los modelos si necesidad de disponer de conocimientos de programación:

Los modelos a visualizar son:

1. Regresiones.
2. Series Temporales.
3. Arboles de decisión.
4. Proyectos de tratamiento de imágenes.
5. Automatización de procesos de ML.
6. Automatización de las acciones a realizar.
7. Interacción de las personas con Sistemas Predictivos

MODULO 9: PROBLEMAS RELACIONADOS CON LA MALA RECOGIDA

1. Problemas de Sesgo y Varianza.
2. Sobre-entrenamiento.
3. Imposibilidad de realizar modelos complejos.
4. Falta de fiabilidad.
5. Imbalanceo de los Datos.

MÓDULO 10: CASOS PRÁCTICOS

En el último modulo del presente curso se mostrarán aplicaciones prácticas donde los alumnos podrán ver un sistema de BI, a través de Power BI, y una aplicación Web donde podrán entrenar y visualizar modelos de Machine Learning a "golpe de click". Además de ello se incluyen una serie de videos explicativos y un ejercicio sobre el desarrollo de un Power BI.

Los contenidos del curso son los siguientes:

1. Ejemplo y ejercicio Power BI.
2. Shiny para visualizar la influencia de la calidad del dato en el modelado.