TC1002S: HERRAMIENTAS COMPUTACIONALES: EL ARTE DE LA ANALÍTICA

Actividades Día 2

SEMANA TEC

Edgar Asael Martinez Gutierrez

Datos abiertos

Datos abierto estado de Jalisco

https://datos.jalisco.gob.mx/

En esta plataforma encontrarás los datos de carácter público generados por las dependencias y entidades gubernamentales, están disponibles en formatos digitales en línea, y pueden ser usados, reutilizados y redistribuidos por cualquier interesado.

https://datos.jalisco.gob.mx/instituciones/iieg

INSTITUTO DE INFORMACIÓN ESTADÍSTICA Y GEOGRÁFICA DEL ESTADO DE JALISCO

https://datos.gob.mx/

Datos Abiertos de tu gobierno

Datos

- Los datos son representaciones simbólicas (vale decir: numéricas, alfabéticas, algorítmicas, etc.) de un determinado atributo o variable cualitativa o cuantitativa, o sea: la descripción codificada de un hecho empírico, un suceso, una entidad.

- Los datos son, así, la información (valores o referentes) que recibe el computador a través de distintos medios, y que es manipulada mediante el procesamiento de los algoritmos de programación. Su contenido puede ser prácticamente cualquiera: estadísticas, números, descriptores, que por separado no tienen relevancia para los usuarios del sistema, pero que en conjunto pueden ser interpretados para obtener una información completa y específica.

Tipos de datos

El análisis o analítica de datos (DA) es el proceso de examinar conjuntos de datos para encontrar tendencias y sacar conclusiones sobre la información que contienen. La analítica de datos se utiliza cada vez más con la ayuda de sistemas y software especializados. Las tecnologías y técnicas de análisis de datos se utilizan ampliamente en las industrias comerciales para permitir a las organizaciones tomar decisiones empresariales más informadas. También se utiliza por científicos e investigadores para verificar o refutar modelos, teorías e hipótesis científicas.

Las iniciativas de análisis de datos pueden ayudar a las empresas a aumentar los ingresos, mejorar la eficiencia operativa, optimizar las campañas de marketing y los esfuerzos de servicio al cliente. También puede utilizarse para responder rápidamente a las nuevas tendencias del mercado y obtener una ventaja competitiva sobre sus competidores. Sin embargo, el objetivo final de la analítica de datos es impulsar el rendimiento empresarial. Dependiendo de la aplicación concreta, los datos que se analizan pueden consistir en registros históricos o en información nueva que ha sido procesada para su análisis en tiempo real. Además, pueden proceder de una mezcla de sistemas internos y fuentes de datos externas.

Tipos de aplicaciones de analítica de datos

En un nivel alto, las metodologías de análisis de datos incluyen el análisis exploratorio de datos (EDA) y el análisis confirmatorio de datos (CDA). El EDA pretende encontrar patrones y relaciones en los datos, mientras que el CDA aplica técnicas estadísticas para determinar si las hipótesis sobre un conjunto de datos son verdaderas o falsas. El AED suele compararse con el trabajo de los detectives, mientras que el ACD se asemeja al trabajo de un juez o un jurado durante un juicio.

El análisis de datos también puede dividirse en análisis de datos cuantitativos y análisis de datos cualitativos. El primero implica el análisis de datos numéricos con variables cuantificables. Estas variables pueden compararse o medirse estadísticamente. El enfoque cualitativo es más interpretativo: se centra en comprender el contenido de los datos no numéricos, como texto, imágenes, audio y video, frases comunes, temas y puntos de vista.

Actividad: Carga de datos con Python

¿Qué voy a aprender con esta actividad?

En esta actividad aprenderás a leer un archivo de datos csv utilizando un programa creado por ti en Python.

¿Qué tengo que hacer?

- Busca el archivo .csv descargado desde el repositorio.
- Abre el archivo en excel y observa su contenido
- Abre el archivo con un editor de texto (gEdit, Notepad, Notepad++) o desde algun ide (Xcode, Visual Studio Code).
- Observa como se separan las columnas (variables) y las filas (objetos).
- Utilizando los comandos open, close, read, readline, readlines, rstrip y split, crea una función de Python que pueda leer los datos.

Instalación de la Suite Anaconda Navigator

Como primer paso para manejar nuestros conjuntos de datos (data sets) utilizaremos la plataforma Anaconda Navigator, que nos permite trabajar con el lenguaje Python de una forma más amigable, pero además gracias a las Jupyter Notebook, podremos ir creando bloques de programas en Python e irlos comprobando bloque por bloque de forma interactiva, así podremos ir haciendo paso a paso nuestros ejercicios y nuestros propios proyectos de aprendizaje automático.

Anaconda Navigator es un ambiente de trabajo para la ciencia de datos que permite hacer funcionar aplicaciones y administrar fácilmente distintos paquetes. Así, Anaconda Navigator puede buscar paquetes en Anaconda Cloud o en otros repositorios, y está disponible para ambientes Windows, macOS y Linux.

Actividad: Carga de datos con Pandas

¿Qué voy a aprender con esta actividad?

En esta actividad aprenderás a leer un archivo de datos csv utilizando la librería de Pandas,

¿Qué tengo que hacer?

1. Importa pandas
2. Lee el archivo .csv usando la función read_csv
3. Observa las primeras líneas usando la función head
4. Obtén diferentes combinaciones de columnas y datos.
5. Revisa los datos obtenidos por la función describe

Comandos a utilizar

```
import pandas as pd       #Libreria pandas

df = pd.read_csv("peso_estatura_genero.csv")   #open archivo de datos

df   #Imprime datos
```

Actividad Evaluable: Obtención de estadísticas descriptivas

¿Qué voy a aprender con esta actividad?

En esta actividad evaluable, harás un análisis inicial exploratorio de tus datos

¿Qué tengo que hacer?

En esta actividad trabajaras con el conjunto de datos asignado para el reto. Tienes que replicar los pasos vistos durante las horas de salón en un diferente conjunto de datos.

1. Carga los datos usando tu lector de csv o con pandas. Es recomendable hacerlo con pandas.
2. Verifica la cantidad de datos que tienes, las variables que contiene cada vector de datos e identifica el tipo de variables.
3. Analiza las variables para saber que representa cada una y en que rangos se encuentran. Si la descripción del problema no te lo indica, utiliza el máximo y el mínimo para encontrarlo.
4. Basándose en la media, mediana y desviación estándar de cada variable, que conclusiones puedes entregar de los datos.

Entregables

1. En github deben quedar todos los códigos que realices, así como un pequeño manual de como correrlos.
2. Reporte con la respuesta a las preguntas.

Actividad: Visualizaciones de estadística basica

¿Qué voy a aprender con esta actividad?

En esta actividad aprenderás a crear e interpretar diferentes graficas de los datos que te permitirán dar una descripción de estos.

¿Qué tengo que hacer?

Sigue a tu profesor en la creación de las diferentes graficas usando el conjunto de datos que se asigno el primer día. Los pasos para seguir serán:

1. Cargar las librerías pandas, seaborn y matplotlib
2. Con los datos cargados, utiliza la función boxplot para obtener un diagrama de cajas y bigotes. En este diagrama podrás observar para una variable la diferencia entre el mayor y el menor, como están distribuidos los datos, donde se encuentra la mediana y si existen outliers, datos que se salen de los valores esperados a partir de los demás.
3. Con los datos utiliza la función hist para crear un histograma. Esta grafica te dirá información acerca de como esta distribuida la variable y en que rangos se encuentran más valores.
4. Utiliza la función corr para obtener un diagrama de correlación. Este análisis te sirve para indicar si hay variables que se parecen o pueden explicar otras. Generalmente conviene eliminar variables con alta correlación, ya que estas pueden meter ruido y hacer más difícil el problema de manejar.
5. Utilizando la función heatmap de seaborn, harás un mapa de calor. Este te indica de manera grafica la correlación entre dos variables, donde entre más fuerte el color, significa que la correlación es más grande.