****

**GUÍA PRÁCTICA**

**EXPLORAR, MANIPULAR, VISUALIZAR Y ANALIZAR DATOS, CON LA CARGA DE LOS DATOS DESDE DIFERENTES FORMATOS DE FICHEROS.**

1. **Datos Generales**

|  |  |
| --- | --- |
| **Carrera:** | **Tecnología Superior en Big Data** |
| **Período académico:** | **Abril – Agosto 2024** |
| **Asignatura:** | **Minería de Datos - Laboratorio** |
| **Unidad Nº:** | **2.Preprocesamiento de Datos con Python:** |
| **Tema:** | **Explorar, manipular, visualizar y analizar datos, con la carga de los datos desde diferentes formatos de ficheros.** |
| **Ciclo-Paralelo:** | **M3A** |
| **Fecha de inicio de la Unidad:** |  |
| **Fecha de fin de la Unidad** |  |
| **Práctica Nº**: | **2** |
| **Horas:** | **10** |
| **Docente:** | **Ing. Verónica Chimbo. Mgtr.** |

1. **Contenido** 
   1. **Fundamentos**

|  |
| --- |
| Las actividades que llevaremos a cabo en esta práctica se realizan en las fases iniciales de un proyecto de minería de datos. Tienen como objetivo obtener un dominio de los datos con las que construiremos el modelo de minería. Tenemos que conocer los datos profundamente tanto en su formato como contenido. Tareas típicas en esta fase pueden ser la selección de características o variables, la preparación del juego de datos para posteriormente ser consumido por un algoritmo e intentar extraer el máximo conocimiento posible de los datos.  Desarrollaremos un subconjunto de tareas mínimas y de ejemplo. Podemos incluir muchas más y mucho más profundas, como hemos visto al material docente.  Como ejemplo, trabajaremos con el juego de datos “Titanic.csv” que recoge datos sobre el famoso crucero. Para la práctica seleccionarán una data de los diferentes repositorios dados en clase. |

* 1. **Objetivo de la Guía**

|  |
| --- |
| * Explorar, manipular, visualizar y analizar datos, con la carga de los datos desde diferentes formatos de ficheros. * Leer ficheros de tipo CSV, Excel, TXT, JSON i ZIP, cargarlos en un DataFrame y volverlos a guardar en otro fichero. * Manipular datos con la librería pandas * Limpiar conjunto de datos con la librería pandas * Visualización de datos |

* 1. **Evaluación del Aprendizaje**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Rúbrica de Evaluación de la Guía Práctica   |  |  | | --- | --- | | **Criterios de Evaluación** | **Puntuación Máxima** | | Paso 1: Lectura de ficheros | /2,5 | | Paso 2: Exploración de los datos | /2,5 | | Paso 3: Limpieza de datos | /2,5 | | Paso 4: Visualización de datos | /2,5 | | Puntuación Total | /10 | |

* 1. **Preparación previa, materiales, herramientas, equipos y software**

|  |
| --- |
| Computador personal, Jupyter. |

* 1. **Procedimientos a emplear**

|  |
| --- |
| De cada uno de los notebooks a realizar se debe de proporcionar un pdf generado del mismo con la ejecución de cada una de las líneas de código con la data seleccionada.   1. Realizar Lectura y escritura de Fichero de Datos, para ello se debe de acceder a los recursos del siguiente repositorio:<https://drive.google.com/drive/folders/1hzzkMe3VuFGw_txy_fWb-K2q58CcIynE?usp=share_link> o descargar directamente desde la plataforma virtual.        Realizar la Manipulación de datos con la librería pandas Para la siguiente sección se debe descargar del siguiente link: <https://drive.google.com/drive/folders/1EVfhbrb4X1Fud4voAKKVF-8AffbEVYLW?usp=share_link> , el archivo contine los siguiente comandos:   1. Realizar la limpieza del conjunto de datos descargando los siguientes recursos: <https://drive.google.com/drive/folders/1OnFuODfl9CGZPthaS9Qv0LFVRraTIIZk?usp=share_link> |
| 1. Finalmente realizar la visualización de datos para ello descargar los recursos del siguiente link: <https://drive.google.com/drive/folders/1MmdojCoR6I20TskChXAOB43iI7LwvbX6?usp=share_link>   Al finalizar se debe de enviar los notebooks generados con las datas correspondientes y el documento adjunto con las evidencias de manipulación de los datos.  Responder a las siguientes preguntas:  1. ¿Qué es el preprocesamiento de datos?  a. El proceso de análisis de datos después de que hayan sido procesados  b. El proceso de limpieza y preparación de datos para el análisis  c. El proceso de visualización de datos usando Python  d. El proceso de recopilación de datos sin procesar.  2. ¿Cuál de los siguientes NO es un paso en el preprocesamiento de datos?  a. Limpieza de datos  b. Transformación de datos  c. Visualización de datos  d. Integración de datos  3. ¿Por qué es importante el preprocesamiento de datos en el análisis de datos?  a. Para facilitar la recopilación de datos  b. Mejorar la precisión de los resultados del análisis  c. Para omitir el paso de análisis de datos  d. Para hacer los datos más confusos  4. ¿Cuál de las siguientes técnicas se utiliza para la detección de valores atípicos en el preprocesamiento de datos?  a. Estandarización  b. Normalización  c. Imputación  d. puntuación Z  5. ¿En qué ayuda el escalado de datos en el preprocesamiento de datos?  a. Reducir el tamaño del conjunto de datos  b. Mejorar la interpretabilidad de los datos  c. Garantizar que todas las características tengan la misma escala  d. Agregar ruido a los datos  6. ¿Qué técnica se utiliza para manejar los datos faltantes en el preprocesamiento de datos?  a. Imputación  b. Normalización  c. Estandarización  d. Codificación en caliente  7. ¿Cuál es el propósito de codificar variables categóricas en el preprocesamiento de datos?  a. Eliminar todas las variables categóricas del conjunto de datos  b. Convertir variables categóricas a formato numérico  c. Para eliminar todas las variables numéricas  d. Para visualizar las variables categóricas  8. ¿Cuál de las siguientes es una técnica de reducción de dimensionalidad utilizada en el preprocesamiento de datos?  a. Escalado de características  b. Análisis de Componentes Principales (PCA)  c. Codificación One-Hot  d. Normalización de la puntuación Z  9. ¿Cuándo se debe realizar el preprocesamiento de datos en el proceso de análisis de datos?  a. Al final del análisis  b. Al inicio del análisis  c. Nunca  d. Sólo si hay tiempo  10. ¿Qué biblioteca de Python se utiliza habitualmente para tareas de preprocesamiento de datos?  a. Matplotlib  b. Nacido en el mar  c. Pandas  d. NumPy  11. ¿Cuál es el propósito de la transformación de datos en el preprocesamiento de datos?  a. Convertir datos a un formato adecuado para el análisis  b. Para hacer los datos más confusos  c. Para eliminar todos los valores faltantes  d. Para omitir el paso de análisis  12. ¿Cuál de las siguientes NO es una técnica de preprocesamiento de datos?  a. Escalado de características  b. Limpieza de datos  c. Visualización de datos  d. barajado de datos  13. ¿Cuál es el objetivo de la normalización de datos en el preprocesamiento de datos?  a. Para convertir datos en un rango de [0, 1]  b. Eliminar todos los valores atípicos de los datos  c. Para aumentar el tamaño del conjunto de datos  d. Para disminuir el número de funciones  14. ¿Cuál de las siguientes es una tarea de limpieza de datos en el preprocesamiento de datos?  a. Manejo de valores faltantes  b. Convertir variables categóricas a numéricas  c. Escalar los datos  d. Realizar ingeniería de características  15. ¿Cuál es el propósito de la ingeniería de características en el preprocesamiento de datos?  a. Crear nuevas funciones a partir de las existentes  b. Para eliminar todas las características del conjunto de datos  c. Visualizar las características  d. Para estandarizar las características  16. ¿Cuál de los siguientes es un beneficio del preprocesamiento de datos?  a. Reducir la precisión del análisis  b. Hacer que los datos sean más difíciles de entender  c. Mejorar la calidad de los resultados del análisis  d. Saltarse el paso de análisis de datos  17. ¿Cuál es el papel del preprocesamiento de datos en los modelos de aprendizaje automático?  a. No tiene ningún impacto en el modelo  b. Es crucial para el éxito del modelo  c. Sólo afecta el tiempo de entrenamiento  d. es opcional  18. ¿Qué paso viene primero en el flujo de trabajo de preprocesamiento de datos?  a. Limpieza de datos  b. Transformación de datos  c. Visualización de datos  d. Escalado de datos  19. ¿Cómo ayuda el preprocesamiento de datos a mejorar la eficiencia de los modelos de aprendizaje automático?  a. Haciendo los datos más complejos  b. Aumentando el número de funciones  c. Reduciendo el ruido y las inconsistencias en los datos  d. Saltándose el paso de entrenamiento  20. ¿Cuál de los siguientes es un desafío común de preprocesamiento de datos?  a. Tener muy pocas funciones  b. Manejo de datos faltantes  c. Ignorar los valores atípicos  d. Realizar una limpieza mínima de datos |

* 1. **Resultados esperados**

|  |
| --- |
| Al finalizar la practica el estudiante:   * Explorar, manipular, visualizar y analizar datos, con la carga de los datos desde diferentes formatos de ficheros. * Leer ficheros de tipo CSV, Excel, TXT, JSON i ZIP, cargarlos en un DataFrame y volverlos a guardar en otro fichero. * Manipular datos con la librería pandas * Limpiar conjunto de datos con la librería pandas * Visualización de datos |

* 1. **Bibliografía**

|  |
| --- |
| Hall, M. A., Frank, E., & Witten, I. H. (2011). Data Mining: Practical Machine Learning Tools and Techniques. Editorial. |

1. **Firmas de Responsabilidad**

|  |  |  |
| --- | --- | --- |
| **ESTUDIANTE** | **DOCENTE** | **DIRECTORA DE CARRERA** |
| **Firma** | **Nombre: Ing. Verónica Chimbo. Mgtr.**  **Firma** | **Nombre: Ing. Verónica Chimbo. Mgtr.**  **Firma** |
| **Fecha:** | **Fecha:** | **Fecha:** |