|  |  |  |
| --- | --- | --- |
|  | **FORMATO DE GUÍA DE PRÁCTICA DE LABORATORIO / TALLERES / CENTROS DE SIMULACIÓN – PARA DOCENTES** | |
|  | | |
| **COMÚN** | | INTELIGENCIA ARTIFICIAL  MACHINE LEARNING |
| **TÍTULO PRÁCTICA**: Análisis exploratorio de datos, Extracción de Características y Presentación de Resultados | | |
| **OBJETIVOS:**  - Reforzar los conocimientos adquiridos sobre análisis exploratorio de datos  - Aplicar técnicas para extraer características a partir de un conjunto de datos  - Presentar los resultados y conclusiones más relevantes | | |
| **INSTRUCCIONES**: | 1. Revisar el contenido teórico del tema | |
| 1. Profundizar los conocimientos revisando los libros guías, el material de aprendizaje (diapositivas y cuadernos de Jupyter), y la documentación disponible en fuentes académicas en línea | |
| 1. Desarrollar las actividades de la sección “**ACTIVIDADES POR DESARROLLAR**” | |
| 1. Subir al AVAC el cuaderno de jupyter y la presentación (.pdf o .pptx) | |
|  | | |
| **ACTIVIDADES POR DESARROLLAR**: En el lenguaje de programación (Python, R, Java, C++, etc.) o herramienta de análisis de datos de su preferencia (Orange3, SPSS, Matlab, Power BI, Oracle BI, etc.) realizar las actividades: | | |
| **- Revisar los fundamentos del lenguaje Python en la guía con ejemplos del archivo “0.1.AprendiendoPythonFundamentos.ipynb”**  **- Seguir los pasos y reproducir los resultados de la guía de ejemplo práctico del archivo “0.2.AnalisisExploratorio.ipynb”. Las actividades del análisis exploratorio son las siguientes:**   * **Carga de Datos:** Importar un conjunto de datos de interés. * **Exploración de Datos Iniciales:** indicar el tamaño, la cantidad de variables y observaciones, tipos de variables y la calidad de los datos (verificar si hay valores faltantes, duplicados o atípicos). * **Resumen Estadístico:** Calcular estadísticas descriptivas básicas, como la media, desviación estándar, percentiles, para comprender la variabilidad de los datos. A partir de las estadísticas presentar algunas conclusiones. * **Visualización de Datos:** Crear gráficos y visualizaciones para representar los datos de manera efectiva. Esto puede incluir histogramas, diagramas de dispersión, gráficos de barras y diagramas de caja, entre otros. * **Análisis de Variables Categóricas:** Si existen variables categóricas, realizar análisis de frecuencia para comprender la distribución de categorías. * **Análisis de Correlación:** Evaluar las relaciones entre las variables mediante el cálculo de correlaciones. A partir de los valores de correlación presenta las variables más relevantes y genera conclusiones. * **Manejo de Datos Faltantes:** Decidir cómo manejar los valores faltantes, que puede incluir imputación (sustitución), eliminación o análisis de su impacto en los resultados. * **Análisis de Outliers:** Identificar y comprender los valores atípicos que pueden afectar el análisis y la interpretación de los resultados. * **Segmentación de Datos:** Si es relevante, dividir el conjunto de datos en segmentos o grupos con características similares para un análisis más detallado. * **Generación de Hipótesis:** Basado en la exploración de datos, formular hipótesis iniciales sobre relaciones o patrones que se puedan investigar más adelante. * **Presentación de Resultados:** Esta actividad consiste en comunicar hallazgos iniciales a través de informes o presentaciones que resuman las observaciones y los posibles pasos siguientes en el análisis. En este caso, con una presentación exponer lo siguiente:   1. Descripción del dataset: número de variables y observaciones, la variable de estudio Y (objetivo)  2. Conclusiones adicionales del resumen estadístico  3. Conclusiones sobre las figuras que considere más relevantes en la visualización  4. Conclusiones sobre el análisis de frecuencia de categorías  5. Las 3 variables más correlacionadas con la variable objetivo Y, y conclusiones de estas correlaciones  6. Análisis de la figura de análisis de Outliers y explicar si es conveniente o no eliminar datos  7. Indicar hipótesis (supuestos) iniciales para determinar un valor adecuado en Y con nuevos ejemplos (observaciones)  **Realizar el análisis exploratorio con un dataset distinto de interés particular. Algunos repositorios de datasets públicos son:**  - <https://archive.ics.uci.edu/ml/index.php>  - <https://www.kaggle.com/datasets>  **Algunos datasets son:**  - El conjunto de datos "**Heart Disease**" (Enfermedad Cardíaca) es una recopilación de datos clínicos relacionados con la salud cardiovascular de pacientes. Este conjunto de datos se utiliza comúnmente para tareas de análisis de datos y aprendizaje automático con el objetivo de predecir la presencia o ausencia de enfermedades cardíacas en pacientes: <https://archive.ics.uci.edu/dataset/45/heart+disease>  - Dataset de **Covid**: <https://www.covid19survivalcalculator.com/download>  - Dataset de **Covid**: <https://www.covidanalytics.io/>  - Dataset de **Covid**: <https://data.europa.eu/euodp/en/data/dataset/covid-19-coronavirus-data>  - Conjunto de datos de **Censo de Ingresos**: Un conjunto de datos del censo de EE. UU. que contiene información sobre los ingresos de las personas, incluyendo características categóricas como ocupación, educación y estado civil. Es adecuado para problemas de clasificación de ingresos: <https://archive.ics.uci.edu/dataset/20/census+income>  - Conjunto de datos de **Titanic**: Este conjunto de datos contiene información sobre los pasajeros del Titanic, incluyendo características como la clase del boleto, edad, género y si sobrevivieron o no. Es excelente para prácticas de clasificación y visualización de datos: <https://www.kaggle.com/c/titanic/data>  - Conjunto de datos de **Diabetes**: Este conjunto de datos se centra en la predicción de la progresión de la diabetes en pacientes. Contiene características médicas y personales, incluyendo algunas variables categóricas. <https://www4.stat.ncsu.edu/~boos/var.select/diabetes.html>  - Conjunto de datos de **Tarjetas de Crédito**: Contiene información sobre transacciones de tarjetas de crédito, incluyendo características categóricas como el tipo de transacción y el comerciante. <https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud>  - Conjunto de datos de **Evaluación de Seguros**: Contiene datos de seguros de automóviles, con variables categóricas como tipo de póliza y características del vehículo. <https://archive.ics.uci.edu/dataset/125/insurance+company+benchmark+coil+2000>  - Conjunto de datos de **Precios de Viviendas en California**: Contiene información sobre precios de viviendas en diferentes áreas de California, con categorías como proximidad al océano y tamaño de la población. <https://www.kaggle.com/datasets/camnugent/california-housing-prices>  **Rúbrica:**   |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | Actividad | Insatisfactorio (0-2) | Aceptable (3-5) | Bueno (6-8) | Excelente (9-10) | Puntaje Obtenido | | 1.Descripción del Dataset | No se presentó información sobre el dataset. | Presentación básica del dataset, pero incompleta. | Descripción adecuada del dataset y variable de estudio. | Descripción detallada y clara del dataset y variable Y. |  | | 2.Conclusiones del Resumen Estadístico | No se presentaron conclusiones del resumen estadístico. | Conclusiones limitadas, falta de profundidad. | Conclusiones claras y efectivas sobre el resumen. | Conclusiones detalladas y profundas con justificación. |  | | 3.Conclusiones de Visualización | No se discutieron figuras relevantes. | Discusión superficial de algunas visualizaciones. | Análisis efectivo de figuras relevantes. | Análisis exhaustivo de las visualizaciones y su impacto. |  | | 4.Análisis de Frecuencia de Categorías | No se presentó análisis de frecuencias. | Análisis básico y poco profundo. | Análisis efectivo y comprensible de frecuencias. | Análisis completo y bien justificado de las categorías. |  | | 5.Correlaciones con Variable Y | No se identificaron variables correlacionadas. | Identificación básica, falta de conclusiones. | Identificación clara de variables correlacionadas y conclusiones. | Análisis profundo de correlaciones y sus implicaciones. |  | | 6.Análisis de Outliers | No se presentó análisis de outliers. | Identificación superficial de outliers. | Análisis efectivo de outliers y sus implicaciones. | Análisis completo y justificado sobre la eliminación de outliers. |  | | 7.Hipótesis Iniciales | No se formularon hipótesis. | Hipótesis limitadas o sin base en el análisis. | Hipótesis formuladas de manera efectiva. | Hipótesis sólidas y bien justificadas con ejemplos. |  | | 8.Comunicación y Presentación | Presentación desorganizada y difícil de seguir. | Presentación aceptable pero con falta de claridad. | Presentación clara y bien estructurada. | Presentación excepcionalmente clara, estructurada y convincente. |  | | | |
| **RESULTADO(S) OBTENIDO(S)**:  Aplica fundamentos de estadística mediante un análisis exploratorio  Aplica técnicas para extraer variables a partir de un conjunto de datos  Realiza una presentación con análisis de resultados  **RESULTADO(S) DE APRENDIZAJE:**  Comprende la importancia del análisis exploratorio de datos  Genera conclusiones basadas en los fundamentos teóricos | | |
| **CONCLUSIONES**:   * Los estudiantes comprenden los fundamentos de análisis exploratorio * Los estudiantes aplican técnicas para extraer variables a partir de un conjunto de datos | | |
| **RECOMENDACIONES**:   * Revisar la información proporcionada por el docente previo a la práctica. * Haber asistido a las sesiones de clase. * Consultar con el docente las dudas que puedan surgir al momento de realizar la práctica. | | |

***Docente*:** Ing. Remigio Hurtado, PhD.

***Firma*: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**