****

**GUÍA PRÁCTICA**

**Limpieza, Transformación y Selección de Datos para Análisis Avanzado**

1. **Datos Generales**

|  |  |
| --- | --- |
| **Carrera:** | Tecnología Superior en Big Data |
| **Período académico:** | 2025-I |
| **Asignatura:** | Minería de Datos I |
| **Unidad Nº:** | 1. Preparación de Datos |
| **Tema:** | Limpieza, Transformación y Selección de Datos para Análisis Avanzado |
| **Ciclo-Paralelo:** | M2A |
| **Fecha de inicio de la Unidad:** | 22/04/2025 |
| **Fecha de fin de la Unidad** | 15/05/2025 |
| **Práctica Nº**: | 1 |
| **Horas:** | 16 |
| **Docente:** | Mgtr. Verónica Paulina Chimbo Coronel |

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

|  |
| --- |
| En esta guía práctica, se analizarán los pasos esenciales para la limpieza, transformación y selección de datos.  **1. Limpieza de Datos**   * **Definición**: La limpieza de datos es el proceso de identificar y corregir errores o inconsistencias en un conjunto de datos. Esto incluye la eliminación de duplicados, el manejo de valores nulos y la normalización de formatos. * **Importancia**: Datos limpios son esenciales para obtener resultados precisos y confiables en cualquier análisis. Los errores en los datos pueden llevar a conclusiones erróneas.   **2. Transformación de Datos**   * **Definición**: La transformación de datos implica modificar los datos para hacerlos más útiles para el análisis. Esto incluye la agregación, la creación de nuevas variables y la reestructuración de los datos. * **Importancia**: La transformación permite a los analistas preparar los datos para modelos estadísticos o de machine learning, facilitando la extracción de información valiosa.   **3. Selección de Datos**   * **Definición**: La selección de datos es el proceso de elegir subconjuntos específicos de un conjunto de datos para un análisis más profundo. Esto se puede hacer mediante filtrado, agrupamiento y condiciones específicas. * **Importancia**: Seleccionar los datos correctos es crucial para enfocar el análisis en las áreas más relevantes, lo que mejora la eficiencia y la efectividad del proceso analítico.   **4. Herramientas y Bibliotecas en Python**   * **Pandas**: Una biblioteca fundamental para la manipulación y análisis de datos. Permite realizar operaciones de limpieza, transformación y selección de manera eficiente. * **NumPy**: Utilizada para operaciones numéricas y manipulación de arreglos, complementando las funcionalidades de Pandas.   **5. Prácticas Recomendadas**   * **Documentación**: Mantener un registro de los pasos realizados durante la limpieza y transformación para asegurar la reproducibilidad. * **Visualización**: Usar gráficos y visualizaciones para entender mejor los datos y detectar anomalías o patrones.   Estos fundamentos son esenciales para cualquier análisis de datos efectivo y proporcionan la base sobre la cual los estudiantes pueden construir habilidades más avanzadas en el uso de Python para el análisis de datos. |

* 1. **Objetivos de la Guía**

|  |
| --- |
| 1. Desarrollar habilidades prácticas en el uso de Python para limpiar y transformar datos reales. 2. Comprender la importancia de la calidad de los datos para análisis avanzados. 3. Aplicar técnicas de filtrado y agrupamiento sobre datos personales. 4. Visualizar datos limpios para apoyar la toma de decisiones. |

* 1. **Evaluación del Aprendizaje**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Rúbrica de Evaluación de la Guía Práctica   | **Criterio** | **2 pts (Excelente)** | **1.5 pts (Bueno)** | **1 pt (Regular)** | **0.5 pt (Deficiente)** | **0 pt (No presentado)** | | --- | --- | --- | --- | --- | --- | | Dominio conceptual | Explica las metodologías con ejemplos claros | Explica una metodología correctamente | Confunde conceptos | No identifica diferencias | No participa | | Aplicación práctica | Implementa todas las fases en el archivo | Aplica 3 fases correctamente | Aplica 1-2 fases | No logra aplicar | No entrega | | Resultados y análisis | Presenta conclusiones válidas con métricas | Conclusiones básicas sin métricas | Análisis superficial | Sin análisis | Sin entregables | |

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

|  |
| --- |
| * Archivo: personas\_registradas.xlsx * Jupyter Notebook o Google Colab * Python con pandas, matplotlib, seaborn * Extensión: openpyxl (para abrir Excel en pandas) |

* 1. **Procedimientos a emplear**

|  |
| --- |
| **import** **pandas** **as** **pd**  # 1. Cargar datos  df = pd.read\_excel("personas\_registradas.xlsx")  # 2. Visualizar y revisar datos  **print**(df.head())  **print**(df.info())  #2.1.Agregar ydataprofile para las estadisticas  # 3. Eliminar duplicados  df.drop\_duplicates(subset=["Correo", "Celular"], inplace=True)  # 4. Normalizar nombres y apellidos  df["Nombres"] = df["Nombres"].str.title()  df["Apellidos"] = df["Apellidos"].str.title()  # 5. Crear columna Nombre\_Completo  df["Nombre\_Completo"] = df["Nombres"] + " " + df["Apellidos"]  # 6. Extraer dominio del correo  df["Dominio\_Correo"] = df["Correo"].str.extract(r'@([\w\.-]+)')  # 7. Agrupar por Institución  instituciones = df["Institución"].value\_counts()  # 8. Agrupar por Dominio  dominios = df["Dominio\_Correo"].value\_counts()  # 9. Guardar archivo limpio  df.to\_csv("personas\_limpias.csv", index=False)  **✅ GRÁFICAS POSIBLES**  **1. 📊 Cantidad de registros por institución**   * **Tipo**: Gráfico de barras (Bar chart) * **Objetivo**: Ver qué institución tiene más personas registradas.   **import** **matplotlib.pyplot** **as** **plt**  instituciones = df['Institución'].value\_counts()  instituciones.plot(kind='bar', figsize=(**10**,**6**), color='skyblue')  plt.title('Registros por Institución')  plt.xlabel('Institución')  plt.ylabel('Cantidad de Personas')  plt.xticks(rotation=**45**)  plt.tight\_layout()  plt.show()  **2. 🥧 Distribución por dominio de correo**   * **Tipo**: Gráfico de torta (Pie chart) * **Objetivo**: Ver qué proveedores de correo predominan (gmail.com, ups.edu.ec, etc.)   dominios = df['Dominio\_Correo'].value\_counts()  dominios.head(**5**).plot(kind='pie', autopct='%1.1f%%', figsize=(**6**,**6**), shadow=True)  plt.title('Distribución de Dominios de Correo (Top 5)')  plt.ylabel('')  plt.show()  **3. 📊 Top 10 nombres más comunes**   * **Tipo**: Gráfico de barras horizontales * **Objetivo**: Ver los nombres más frecuentes en la base.   nombres = df['Nombres'].value\_counts().head(**10**)  nombres.plot(kind='barh', color='lightgreen')  plt.title('Top 10 Nombres más Comunes')  plt.xlabel('Frecuencia')  plt.gca().invert\_yaxis()  plt.tight\_layout()  plt.show()  **4. 🧑‍🤝‍🧑 Número de personas por institución que usan correo institucional vs. personal**   * **Tipo**: Gráfico de barras agrupadas * **Objetivo**: Comparar tipos de correo dentro de instituciones.   df['Tipo\_Correo'] = df['Dominio\_Correo'].apply(**lambda** x: 'Institucional' **if** 'edu' **in** x **else** 'Personal')  agrupado = df.groupby(['Institución', 'Tipo\_Correo']).size().unstack().fillna(**0**)  agrupado.plot(kind='bar', stacked=False, figsize=(**10**,**6**))  plt.title('Tipo de Correo por Institución')  plt.ylabel('Cantidad')  plt.xticks(rotation=**45**)  plt.tight\_layout()  plt.show()   1. **Mapa de palabras**   **#**pip install pandas wordcloud matplotlib openpyxl  **import** **pandas** **as** **pd**  **from** **wordcloud** **import** WordCloud  **import** **matplotlib.pyplot** **as** **plt**  # Cargar el archivo Excel  df = pd.read\_excel('personas\_registradas.xlsx')  # Unir los campos 'Nombres' y 'Apellidos' en un solo texto  texto = ' '.join(df['Nombres'].astype(str)) + ' ' + ' '.join(df['Apellidos'].astype(str))  # Crear el mapa de palabras  wordcloud = WordCloud(width=**800**, height=**400**, background\_color='white', colormap='viridis').generate(texto)  # Mostrar la gráfica  plt.figure(figsize=(**12**, **6**))  plt.imshow(wordcloud, interpolation='bilinear')  plt.axis('off')  plt.title("Mapa de Palabras - Nombres y Apellidos", fontsize=**16**)  plt.tight\_layout(pad=**0**)  plt.show()  **Adicional**  Utilizar la libreria Streamlit para mostrar los pasos anteriores  Ejemplo:    **import** **streamlit** **as** **st**  **import** **pandas** **as** **pd**  **from** **wordcloud** **import** WordCloud  **import** **matplotlib.pyplot** **as** **plt**  st.set\_page\_config(page\_title="Mapa de Palabras", layout="centered")  st.title("🧠 Generador de Mapa de Palabras")  st.write("Sube un archivo Excel con columnas `Nombres` y `Apellidos` para generar un WordCloud.")  # Subir archivo  archivo = st.file\_uploader("📂 Cargar archivo Excel", type=["xlsx"])  **if** archivo **is** **not** None:  **try**:  df = pd.read\_excel(archivo)  **if** "Nombres" **in** df.columns **and** "Apellidos" **in** df.columns:  # Concatenar Nombres y Apellidos  texto = ' '.join(df['Nombres'].astype(str)) + ' ' + ' '.join(df['Apellidos'].astype(str))  # Generar mapa de palabras  wordcloud = WordCloud(width=**800**, height=**400**, background\_color='white', colormap='plasma').generate(texto)  # Mostrar gráfico  st.subheader("🔤 Mapa de Palabras Generado")  fig, ax = plt.subplots(figsize=(**10**, **5**))  ax.imshow(wordcloud, interpolation='bilinear')  ax.axis('off')  st.pyplot(fig)  **else**:  st.warning("El archivo debe contener las columnas 'Nombres' y 'Apellidos'.")  **except** **Exception** **as** e:  st.error(f"❌ Error al procesar el archivo: {e}") |

* 1. **Normas de Seguridad**

|  |
| --- |
| Las normas de seguridad se han tomado del reglamento general de seguridad para el uso de los talleres, aulas y laboratorios del Instituto Superior Universitario Tecnológico del Azuay.  El estudiante, al ingresar a los talleres o laboratorios, está sujeto a este reglamento; y, tendrá la supervisión del profesor y del personal técnico; será responsable de:  a) Usar los EquipoS de Protección Personal(EPP) de acuerdo con lo establecido en la “Matriz de equipos de protección individual (EPP`s) requeridos para el ingreso de estudiantes y profesores a los laboratorios y talleres del INSTITUTO”;  b) Al inicio de cada práctica, recibir y revisar el material y herramientas requeridas para la Práctica, serán responsables de su buen uso.  c) La operación de los equipos por los estudiantes deberá ser con el conocimiento de su funcionamiento y bajo las directrices del profesor o personal técnico del laboratorio o taller; bajo ninguna circunstancia el estudiante podrá trabajar solo y sin vigilancia;  d) Seguir las instrucciones dadas por el docente o el personal técnico de apoyo;  e) Al término de la práctica, entregar limpio tanto el material como su área de trabajo;  f) Informar inmediatamente al profesor o personal técnico de apoyo, cualquier desperfecto que se localice en los equipos e instalaciones. |

* 1. **Resultados esperados**

|  |
| --- |
| * Archivo limpio sin duplicados y nombres normalizados. * Columna de nombre completo y dominio extraído. * Estadísticas y visualizaciones de instituciones y dominios. |

* 1. **Bibliografía**

|  |
| --- |
| **Descripción en norma APA** |
| * Chapman, P., et al. (2000). CRISP-DM 1.0: Step-by-step data mining guide. CRISP-DM Consortium. * SAS Institute. (2018). SEMMA: Data mining methodology. SAS Publishing. * Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. 4th Int. Conf. on KDD, 29-39. |

1. **Firmas de Responsabilidad**

|  |  |  |
| --- | --- | --- |
| **ESTUDIANTE** | **DOCENTE** | **DIRECTORA DE CARRERA** |
| **Nombre:**  **Firma** | **Nombre:** Mgtr. Verónica Chimbo  **Firma** | **Nombre:**  Mgtr. Priscila Bernal  **Firma** |
| **Fecha: ( )** | **Fecha:(**18/04/2025**)** | **Fecha: ( )** |