

Fecha: 24 FEB 2026

**PROGRAMA ACADÉMICO:** Ingeniería de Sistemas y Computación

**SEMESTRE:** DECIMO

**ASIGNATURA:** ELECTIVA IV - GESTIÓN DE DATOS CON PYTHON

**CÓDIGO:** 8108282

**NÚMERO DE CRÉDITOS:** 3

### PRESENTACIÓN

La Gestión de Datos se refiere a explorar y preprocesar los datos matricialmente, para hacer tratamiento y visualización de datos antes de abordar un proyecto de Ciencia de Datos y Big Data.

La limpieza de datos es el proceso de detectar y corregir o eliminar registros corruptos o inexactos de un conjunto de registros, tabla o base de datos y se refiere a la identificación de partes incompletas, incorrectas, inexactas o irrelevantes de los datos, para su posterior sustitución, modificación o eliminación de los datos si es el caso.

Por tanto, el objetivo del preprocesamiento de datos es utilizar técnicas computacionales y estadísticas que garanticen los procesos de carga, descripción, limpieza y transformación de datos en estructuras matriciales requeridas para crear modelos de Machine Learning.

Igualmente se analizan datos utilizando herramientas computacionales y estadísticas para conocimiento de variables y comportamiento. Se realiza una descripción estadística de las variables y análisis gráfico de las mismas.

### JUSTIFICACIÓN

Los avances de la tecnología en todas las áreas, el aumento de dispositivos conectados en tiempo real a la red y el uso de diferentes tipos de dispositivos (celulares, relojes, computadores, cámaras, televisores en general dispositivos tipo IoT) han provocado que se generen grandes cantidades de datos, que requieren de técnicas computacionales apropiadas para extraerlos, almacenarlos, procesarlos, limpiarlos, transformarlos y visualizarlos en estructuras de datos que permitan manipularlos para encontrar patrones ocultos.

Por lo anterior, un científico de datos necesita realizar una serie de operaciones previas antes de aplicar modelos. Una de esas actividades previas se refiere a conocer y preparar los datos, tarea que suele ser de vital importancia para corregir múltiples deficiencias en los datos y extraer conocimiento.

La calidad del conocimiento extraído depende en gran medida de la calidad de los datos. Infortunadamente, estos datos se ven afectados por factores negativos como: ruido, valores perdidos, inconsistencias, datos superfluos y/o un tamaño grande en cualquier dimensión (número de atributos e instancias). Está demostrado que una baja calidad de los datos conduce en la mayoría de los casos



a una baja calidad del conocimiento extraído de los modelos de Machine Learning.

Por lo anterior, la gestión de datos utiliza técnicas que los profesionales de cualquier área requieren dominar para enfrentar el reto de las organizaciones en convertir datos en información y luego en conocimiento útil para la toma de decisiones.

Finalmente, en esta asignatura electiva se centrará en la aplicación del lenguaje de programación Python para el tratamiento de datos, ya que nos proporciona herramientas muy interesantes que se aplican en el área de la Ciencia de Datos y Big Data. Se iniciará por conceptos de programación para posteriormente suministrar estrategias para hacer tratamiento y visualización de datos, uso de técnicas estadísticas para el análisis de variables, que serán insumos para las siguientes etapas que contemplan la construcción de modelos en Machine Learning.

### COMPETENCIAS

- Aplicar las diferentes técnicas computacionales y estadísticas para realizar los procesos de exploración, preprocesamiento, procesamiento y transformación de datos.
- Implementar aplicaciones sobre los conjuntos de datos.
- Aplicar técnicas de limpieza de datos para preparación de información que pueda ser utilizada en procesos de Gestión de datos y para algoritmos de Machine Learning.
- Presentar y comunicar los resultados obtenidos de los proyectos realizados de gestión de datos de forma gráfica a través de las visualizaciones.

### RESULTADOS DE APRENDIZAJE

Tiene la capacidad de ejecutar flujos de trabajo de exploración Analítica de Datos (EDA) utilizando Python (Pandas/NumPy) para identificar patrones, anomalías y distribuciones estadísticas, fundamentando las decisiones de transformación de datos en métricas cuantitativas.

Desarrolla scripts o aplicaciones funcionales que consuman, procesen y almacenen conjuntos de datos de diversas fuentes (CSV, JSON, SQL), asegurando la eficiencia computacional y la escalabilidad de la solución tecnológica.

Aplica procesos de limpieza y transformación de datos (ETL), al igual que identificar y manejo de *outliers*, valores faltantes y codificación de variables categóricas, garantizando que los datasets resultantes cumplan con los estándares de calidad requeridos por algoritmos de aprendizaje automático.

Construye visualizaciones de datos interactivas y estáticas utilizando librerías como Plotly o Matplotlib para comunicar hallazgos complejos de manera clara y técnica, facilitando la toma de

decisiones con los datos.

### METODOLOGÍA

El profesor presentará de forma magistral los conceptos a través de libros de trabajo (notebooks) en los cuales los participantes observarán de forma práctica la implementación de dichos conceptos.

Los participantes seguirán el material al mismo tiempo que el profesor con el fin de poder apropiarse los conocimientos a través de la práctica.

El profesor propondrá talleres prácticos en los que el participante ponga en práctica lo aprendido en las sesiones magistrales.

### INVESTIGACIÓN

La asignatura por sus condiciones particulares, promueve en el estudiante la investigación formativa, de modo sea posible obtener los fundamentos con los que podrá posteriormente vincularse en el tratamiento de áreas específicas de la disciplina Informática, por medio de los grupos y semilleros de investigación dinamizados al interior de la Escuela de Ingeniería de Sistemas.

### MEDIOS AUDIOVISUALES

La asignatura es eminentemente práctica, de manera que se hace necesario que todas las sesiones sean realizadas utilizando un computador.

Además, se utilizará de Internet y aplicaciones de streaming (Google Meet y/o Zoom).

### EVALUACIÓN

#### EVALUACIÓN COLECTIVA

Por grupos desarrollan los talleres de finalización de cada capítulo.

Realizar un taller final de asignatura donde aplique todo lo aprendido en un proyecto de gestión de datos con información real de plataformas Open Data.

#### EVALUACIÓN INDIVIDUAL

La evaluación en la asignatura está orientada a determinar el nivel de desarrollo de los procesos lógicos en el estudiante, junto con su capacidad para abstraer problemas y generar soluciones informáticas.

## RANGOS DE VALORACIÓN (%)

El cálculo de la nota final es de 100% y se hará de la siguiente manera:

- Proyecto: 35% (40% Desarrollo del proyecto + 30% Sustentación (escrita o verbal) + 30% Informe escrito del proyecto)
- Participación: 30%
- Trabajos/Talleres: 20%
- Exposiciones/Quiz: 15%

## CONTENIDOS TEMÁTICOS CENTRALES

Unidad I: Estructuras que permiten trabajar con colección de Datos

- 1.1 Cadena, Tuplas y conjuntos.
- 1.2 Listas y Diccionarios.

Unidad II: Programación Funcional, expresiones y operadores.

- 2.1 Uso de expresiones lambda y list comprehensions.
- 2.2 Operadores list, map, filter, reduce

Unidad III: Conocimiento, exploración y cargue de datos.

- 3.1 Numpy, Series y Pandas
- 3.2 Tipos de datos estructurados, semiestructurados y no estructurados.
- 3.3 Tipos de formatos (csv, Excel, Bases de Datos) para consumir datos.
- 3.4 Leer datos In Situ o en la Nube.
- 3.5 Limitar el consumo de datos tanto en variables como en registros.
- 3.6 Crear nuevas columnas.
- 3.7 Observar cada una de las funcionalidades de pandas en las diferentes graficas.

Unidad IV: Procesamiento de datos en estructuras matriciales.

- 4.1 Bases de datos relacionales. Archivo CSV en MariaDB o MySQL.
- 4.2 Fusionar (merge) y unir (join) los conjuntos de datos.
- 4.3 Emplear segmentación e indexación a los datos.
- 4.4 Analizar datos con groupby.
- 4.5 Examinar los datos manipulando, cortando y aplicando funciones agregadas.
- 4.6 Combinar conjuntos de datos (concatenar).
- 4.7 Resúmenes numéricos y gráficos.

Unidad V: Preprocesamiento de Datos (Conocer, Limpiar, fusionar y gestionar datos).

- 5.1 Conocer los datos (caracterización de variables)
- 5.2 Operaciones validas sobre los diferentes tipos de variables (Numéricas (cuantitativas) y Categóricas (cualitativas))
- 5.3 Analizando y comprendiendo las gráficas con respecto a los valores y tipos de variables.
- 5.4 Matriz de correlación
- 5.5. PCA – Reducción de la dimensionalidad.
- 5.3 Manejo de datos categóricos - comprender el tratamiento de las variables categóricas
- 5.4 Crear nuevas columnas.

- 5.5 Manejo de valores faltantes y fechas.
- 5.6 Escala y normalización
- 5.7 Gráficas y Dashboards para visualización de datos.

Unidad VI: Apache Spark con Python PySpark

- 6.1. Spark RDD
- 6.2. SparkContext vs SparkSessions
- 6.3. DataFrames
- 6.4. SQL
- 6.5. Streaming
- 6.6. MLlib
- 6.7. GraphFrames y Resource

**LECTURAS MÍNIMAS**

Jake VanderPlas (2016), Python Data Science Handbook, O'Reilly,2022

**BIBLIOGRAFÍA**

- [1] Jake VanderPlas (2016), Python Data Science Handbook, O'Reilly, 2016
- [2] Hastie T., Tibshirani R., Friedman J. (2017) The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition (Springer Series in Statistics).
- [3] Jason W. Osborne - Best Practices in Data Cleaning\_ A Complete Guide to Everything You Need to Do Before and After Collecting Your Data-SAGE Publications, Inc (2012)
- [4] Q. Ethan McCallum - Bad Data Handbook-O'Reilly Media (2012)

Nombre del docente responsable: Jorge Enrique Quevedo Reyes