PROGRAMA DE ESTUDIOS:
ANÁLISIS DE DATOS

KOICA HANDONG GLOBAL UNIVERSITY CIT UNA

Área: Big Data

Análisis de Datos

□ Información básica

[Información del curso]

1	Título	*Análisis de datos*
2	Año lectivo	*2025*
3	Semestre	*Segundo (Marzo a Junio)*
4	Departamento	*Centro de Innovación TIC (FPUNA)*
5	Nivel	*Avanzado*
6	Formato de clase (tipo)	*Clases interactivas, sesiones prácticas, laboratorios y actividades.*

[Horario y lugar]

1	Días	*Martes y Viernes*
2	Horario	*de 17:00 a 19:30*
3	Ubicación	*Online*

[Información del instructor/a]

1	Nombre	*Prof. MSc. María Teresa Chica Lange*
2	Oficina (si aplica)	*INE- Edificio Técnico-Avenida Boggiani esq. Cirilo Rivarola Nro. 6688.*
3	Contacto (correo)	*techifin@gmail.com*
4	Contacto (teléfono)	*(0971) 967448*

[Horario de oficina]

Si los estudiantes necesitan realizar consultas, a partir de las 16:00 hs a las 17:00 hs los días jueves.

[Notas adicionales]

Habilitación de Foros: Todos los estudiantes pueden tener retroalimentación y se responde una vez a preguntas recursivas de los mismos. Además, se tiene las evidencias de retroalimentación de parte del profesor a los alumnos.

□ Prerrequisitos

1. Conocimientos de Programación:

- Familiaridad con al menos uno de los lenguajes de programación como Python, R, SPSS para el análisis y procesamiento de datos.
- Manejo de Jupyter Notebook o Visual Code, Colab.

2. Fundamentos en Ciencia de Datos:

- Conocimientos sobre estadística, probabilidad y álgebra lineal.
- Familiaridad con técnicas de análisis de datos, limpieza y Modelado
- Estadístico, Machine Learning para extracción de insights.

3. Conocimientos Básicos:

Este curso está dirigido a profesionales de la informática, ingenieros de datos, científicos de dato s, analistas de negocios y cualquier persona interesada en adquirir habilidades avanzadas en el campo del Big Data. Se recomienda tener conocimientos básicos de programación, estadística y bases de datos.

- Programación: Dominio de al menos un lenguaje de programación como Python o R. Estos lenguajes son ampliamente utilizados en el mundo del análisis de datos y Big Data debido a sus potentes librerías y frameworks.
- Estadística y probabilidad: Conceptos básicos de estadística descriptiva e inferencial, así como probabilidad. Estos conocimientos son esenciales para entender y analizar los datos.
- Matemáticas: Nociones de álgebra lineal, cálculo y optimización. Estos conceptos son fundamentales para comprender algoritmos de machine learning y deep learning.

□ Descripción del curso

Este curso intensivo se enfoca en proporcionar a los estudiantes un dominio profundo de las tecnologías y metodologías de vanguardia en el campo del Big Data. A través de una combinación de clases teóricas interactivas y prácticas intensivas, los participantes adquirirán las habilidades necesarias para diseñar, implementar y gestionar soluciones de Big Data en entornos empresariales complejos.

Características Clave:

- **Contenido Teórico Sólido**: El curso abarca una amplia gama de temas, desde la extracción de datos, su limpieza y manipulación. Técnicas avanzadas de calidad de los datos. Detección y corrección de outliers Consistencia de datos. Transformación de datos. Normalización,

estandarización, codificación de variables categóricas. Reducción de dimensionalidad.

- **Prácticas Intensivas**: Los estudiantes trabajarán en proyectos reales, utilizando herramientas y tecnologías de última generación como Hadoop, Spark, PySpark . Desarrollarán habilidades de análisis exploratorio de datos, la construcción de modelos predictivos y la visualización de resultados.

- **Enfoque en la Resolución de Problemas**: El curso se centra en la aplicación práctica de los conocimientos adquiridos. Los participantes trabajarán en desafíos del mundo real, aprendiendo a identificar los problemas de negocio que pueden ser resueltos con Big Data, a diseñar soluciones escalables y eficientes, y a evaluar los resultados obtenidos.

- **Ética en el Manejo de Datos**: Se dedicará un módulo específico a la ética en el manejo de datos masivos. Los estudiantes aprenderán sobre los desafíos éticos asociados con el Big Data, como la privacidad, la seguridad, el sesgo algorítmico y la responsabilidad social. Se discutirán las mejores prácticas para garantizar un uso responsable y ético de los datos.

- **Metodología de Aprendizaje Interactivo**: El curso se basa en una metodología de aprendizaje activo, que fomenta la participación y el intercambio de conocimientos entre los estudiantes. Se utilizarán diversas herramientas pedagógicas, como casos prácticos, estudios de casos, debates y proyectos colaborativos.

- **Herramientas y software**: Código abierto preferentemente tanto para el análisis como procesamiento.

□ Objetivos del curso

Al finalizar con éxito este curso los estudiantes serán capaces de:

1. Diseñar y desarrollar arquitecturas de Big Data escalables y robustas.
2. Seleccionar y aplicar las herramientas y tecnologías adecuadas para cada escenario.
3. Extraer valor de grandes volúmenes de datos estructurados y no estructurados.
4. Construir modelos predictivos y realizar análisis avanzados.
5. Comunicar de forma efectiva los resultados de sus análisis a audiencias técnicas y no técnicas.
6. Actuar de manera ética y responsable en el manejo de datos.
    - Aplicar técnicas de análisis de datos para resolver problemas a través de estudios de casos.
    - Hadoop y ecosistema.

- Spark: procesamiento de datos a gran escala, PySpark casos de usos.
- Visualización de datos (Tableu, PowerBI).
- Ética en el Big Data (Gobierno Dato, anonimización).
- Casos prácticos y proyectos.

□ Método de evaluación

Calificación absoluta de 100%:

- Actividades de laboratorios y tareas: 50%
- Exámenes:
  - De proceso: 30%
  - Final: 20%.

□ Libros de texto y otros materiales necesarios

- Chakrabarti, S. (2003). *Mining the web: discovering knowledge from hypertext data* (pp. 17-43). Morgan Kaufmann.
- Debenham, J. (1998). *Knowledge engineering. Unifying knowledge base and database design* (pp. 15-22). Springer.
- Matallah, H., Belalem, G., & Bouamrane, K. (2021). *Comparative study between the MySQL relational database and the MongoDB NoSQL database*. International Journal of Software Science and Computational Intelligence (IJSSCI), 13(3), 38-63.
- Meier, A., & Kaufmann, M. (2019). *SQL & NoSQL databases*. Springer Fachmedien Wiesbaden.
- de Oliveira, V. F., Pessoa, M. A. D. O., Junqueira, F., & Miyagi, P. E. (2021). *SQL and NoSQL Databases in the Context of Industry 4.0. Machines*, 10(1), 20.
- Redman, T. C. (1996). *Data quality for the information age* (pp. 245-267). Artech House.
- Shafranovich, Y. (2005). *Common format and MIME type for Comma-Separated Values (CSV) file*s. Internet Engineering Task Force IETF RFC 4180.
- Sumalatha, A., Vookanti, R., & Vannala, S. (2021). *Study on Applications of SQL and Not only SQL Databases used for Big Data Analytics. International Journal For Research & Development In Technology*, 15, 127-130.
- Wanumen, L. F. (2018). *Bases de datos en SQL server*. Ecoe Ediciones.
- Matallah, H., Belalem, G., & Bouamrane, K. (2020). *Evaluation of NoSQL databases: MongoDB, Cassandra, HBase, Redis, Couchbase, OrientDB*. International Journal of Software Science and Computational Intelligence (IJSSCI), 12(4), 71-91.

- Bansal, H. (2019, octubre 18). *Best Languages For Machine Learning in 2020!* Medium. Disponible en https://becominghuman.ai/best-languages-for-machinelearning- in-2020-6034732dd242
- *Best Python libraries for Machine Learning*. (2019, enero 18). GeeksforGeeks. Disponible en https://www.geeksforgeeks.org/best-python-libraries-for-machinelearning/
- Puget, J.F. (2016) *The Most Popular Language For Machine Learning and Data Science Is…* KDnuggets. Recuperado de https://www.kdnuggets.com/the-mostpopular-l anguage-for-machine-learning-and-data-science-is.html/
- Raschka, S. & Mirjalili, V. (2019). *Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2*. Packt Publishing Ltd.

Contenidos de Dataset:

- https://www.openml.org/d/37
- https://www.kaggle.com/datasets

Lecturas Complementarias:

- Amón, J. (1984). *Estadística para Psicólogos. Vol. 2: Probabilidad y Estadística Inferencial*. Madrid: Pirámide.
- Lipschutz, S. (1971). *Teoría y problemas de probabilidad*. México: McGraw-Hill.
- Martín, A. (2004). *Bioestadística para las ciencias de la salud* (1ª ed.). Madrid: Norma-Capitel.
- Moore, D. S. (2006). *Introduction to the practice of statistics* (5th. ed.). New York: Freeman and Company.
- Ríus, F. (1998). *Bioestadística: Métodos y aplicaciones*. Málaga: Universidad de Málaga. Versión electrónica: https://www.bioestadistica.uma.es/baron/apuntes/clase/apuntes/pdf/

□ Tarea(s) y examen(es)

Las tareas están diseñadas para reforzar los conceptos teóricos tratados en las clases y proporcionar experiencia práctica. Los exámenes están estructurados para evaluar la comprensión general de los conceptos, las teorías y las aplicaciones relacionadas.

Tareas:

Se realizarán con software libres instalados y servirán para que los alumnos aprendan a desarrollar habilidades.

- Objetivo: Garantizar que los estudiantes puedan realizar tareas de manejo de datos y análisis con

las herramientas recomendadas a través de prácticas en laboratorios.

- Frecuencia: Semanal de acuerdo a la profundidad de las tareas.
- Formato: Realización de las tareas y entrega con capturas de pantallas en Word de los procesos realizados acompañados de programas según el caso. Las sentencias utilizadas deberán estar en formato texto con un editor de manera que pueda replicarse en las Bases para validar contenido.

Exámenes:

1. Examen parcial: Cubre el material de la primera mitad del curso y se centra en conceptos fundamentales dictados con el uso de las herramientas aprendidas. Consta de una parte teórica y una parte practica.

2. Examen final: Es integral, cubre todo el contenido del curso con énfasis en la integración de conceptos. Consta de una parte teórica y una parte práctica que englobe el contenido del curso.

- Objetivo: Evaluar la comprensión de los estudiantes del material del curso y su capacidad para aplicar conocimientos teóricos para resolver problemas.
- Formato: Online, dependiendo del formato de enseñanza del curso.

Proyecto Final:

- Objetivo: Simular un proyecto de análisis de datos del mundo real, desde el preprocesamiento de datos, el análisis, estructurar base de datos y difusión de resultados utilizando las herramientas estudiadas y la generación de informes.
- Alcance: Los estudiantes pueden realizar el tema de su proyecto en base a uno de los retos que presente el instructor. Los proyectos deben ser individuales y deben mostrar la capacidad de limpiar datos, aplicar análisis estadísticos, crear visualizaciones e interpretar hallazgos.
- Formato: Proyectos individuales, que culminan con un informe escrito.

Criterios de Evaluación:

- Tareas: Evaluadas en función de la precisión, la integridad y la aplicación de métodos y herramientas apropiados.
- Exámenes: Se califican según la exactitud de las respuestas, la aplicación de conceptos y la capacidad de razonar a través de problemas.

□ Actividades del curso

Las actividades del curso están diseñadas para que los estudiantes se involucren activamente con los materiales, refuercen su comprensión y desarrollen habilidades prácticas. Estas actividades van desde conferencias y debates interactivos hasta laboratorios prácticos y proyectos en grupo.

- Las clases interactivas incluyen, entre otras cosas, sesiones de preguntas y respuestas.
- Las sesiones de debate incluyen, entre otras cosas, estudios de casos.
- Actividades a desarrollar con casos presentados por el instructor y que deben ser entregados en un documento.
- Talleres de temas especializados y actividades grupales.
- Seminarios y las ponencias de invitados con charlas sobre la industria y preguntas y respuestas con expertos.

Cada una de estas actividades está diseñada para complementar los conocimientos teóricos adquiridos en las clases, profundizar en la comprensión mediante el debate y la aplicación, y prepa rar a los estudiantes para tareas relacionadas con el mundo real. Mediante la participación en di versas actividades del curso, los estudiantes desarrollan un conjunto completo de habilidades que incluyen la competencia técnica, el razonamiento ético y las experiencias de trabajo en colaboración.

□ Cronograma del curso

Semana	Tema	Tipo de clases	Materiales
1	**Tema 1.** Fundamentos estadísticos para el análisis y métodos de captura de información. **Tema 2.** Herramientas para uso de análisis de datos y Análisis Exploratorio de Datos (EDA Parte I).	Clase interactiva y Laboratorio.	Jupyter Notebook Visual Code, R, Python
2	**Tema 3.** Análisis Exploratorio de Datos (EDA Parte II).	Clase interactiva y Laboratorio.	Jupyter Notebook Visual Code, R, Python
3	**Tema 4.** Aprendizaje automático de datos.	Clase interactiva y Laboratorio.	Jupyter Notebook, Python y librerías NTLK, Scikit-Learn, TensorFlow
4	Examen parcial.	Laboratorios prácticos y examen parcial.	Jupyter Notebook, Python y librerías NTLK, Scikit-Learn, TensorFlow

5	**Tema 5.** HDFS: Almacenamiento grandes cantidad de datos. **Tema 6.** Apache Spark.	Clase interactiva y Laboratorio.	Apache Hadoop, Apache Spark
6	**Tema 7.** Big data y protección de datos personales. **Tema 8.** La disociación de datos personales y técnicas de anonimización.	Clase interactiva.	
7	Uso de Big data en el INE.	Ponencia invitada. Caso práctico.	
8	Examen final.	Taller. Examen final.	Jupyter Notebook, Python y librerías NTLK, Scikit-Learn, TensorFlow

□ Contenidos del curso

Semana 1:

- **Tema 1: Fundamentos estadísticos para el análisis y métodos de captura de información**
  - ¿Qué es estadística? Utilidad y objetivos de uso.
  - Organización de los datos. Población y muestra. Estadística descriptiva e inferencial.
  - Tipos de variables estadísticas.
  - Razonamiento estadístico
  - Métodos de captura de la información.
  - Transformación y calidad del Dato.
  - Tipos de Ficheros, csv, Json, xml.
- Referencias bibliográficas:
  - Moore, D. S. (2006). Introduction to the practice of statistics (5th. ed.). New York: Freeman and Company.
  - Ríus, F. (1998). Bioestadística: Métodos y aplicaciones. Málaga: Universidad de Málaga. Versión electrónica: https://www.bioestadistica.uma.es/baron/apuntes/clase/apuntes/pdf/
  - Triola, M. F. (2009). Estadística (10ª ed.). México D.F.: Pearson Educación

- **Tema 2: Herramientas para uso de análisis de datos**
  - Herramientas: R, Python, Anaconda, Jupyter Notebook.
  - Análisis Exploratorio de Datos (EDA Parte I).
  - Origen y calidad de los datos. Detección y corrección de outliers.
  - Consistencia de datos.

- Herramientas de visualización.
- Herramientas de análisis estadístico. R, Python, Jupyter Notebook.

- Referencias bibliográficas:
  - https://link.springer.com/book/10.1007/978-3-319-24277-4 https://ggplot2-book.org/
  - https://github.com/rstudio/cheatsheets/blob/main/data-visualization.pdf
  - The R Foundation. (S. f.). The R Project for Statistical Computing. https://www.rproject. org/

Semana 2:

- **Tema 3: Análisis Exploratorio de Datos (EDA). Práctica con un conjunto de datos a fin de descubrir la estructura de datos, patrones, identificar anomalías y generar hipótesis**
  - Descubrimiento de anomalías
  - Validación de supuestos
  - Generación de hipótesis
  - Limpieza de datos
  - Visualizaciones (histogramas, Box plots, Scatter plots, Pair plots, heatmaps)
  - Estadísticas descriptivas
    - Medidas de tendencia central (media, mediana, moda)
    - Medidas de dispersión
    - Cuartiles
  - Tablas de frecuencia
  - Visualización de datos. Dashboard. Tableu. PowerBI
- Referencias bibliográficas:
  - https://datos.gob.es/es/documentacion/guia-practica-de-introduccion-al-analisis-exploratorio-de- datos
  - Tableau: Software de análisis e inteligencia de datos. https://www.tableau.com/

Semana 3:

- **Tema 4: Aprendizaje automático de datos**
  - Técnicas Supervisadas (Regresión y Clasificación).
  - Técnicas No Supervisadas (Clustering).
  - Procesamiento de lenguaje Natural (NLP).
  - Transformación y re-escalamiento de variables.
    - Binning. Encoder. Scaler. Power Transformer.
  - Scaling. Métodos. Normalización. Estandarización.
  - Encoding en variables categóricas.

- Discretización en k-Bins.

- Referencias bibliográficas:
  - Han, Jiawei, Kamber, Micheline, & Pei, Jian. (2012). Data Mining: Concepts and Techniques. Mo rgan Kaufmann. Un texto completo sobre minería de datos, que abarca desde la recopilación ha sta el análisis de datos.
  - Witten, Ian H., Frank, Eibe, & Hall, Mark A. (2016). Data Mining: Practical Machine Learning Too ls and Techniques. Morgan Kaufmann. Otro libro de referencia en minería de datos, con un enfo que práctico y ejemplos.
  - Russell, Stuart J., & Norvig, Peter. (2016). Inteligencia artificial: un enfoque moderno. Pearson Educación. (Aunque abarca un campo más amplio, este libro dedica capítulos a temas como el aprendizaje automático y el procesamiento del lenguaje natural, que son fundamentales para el procesamiento de datos.)

Semana 4: Examen Parcial

- Objetivo: Evaluar el aprendizaje de los estudiantes a través de actividades teóricas y prácticas de laboratorios.
- El examen consistirá de dos partes:
  - 1) Preguntas que deberá marcar el alumno con la respuesta correcta.
  - 2) Planteamiento de un problema y con las herramientas disponibles realizar las actividades solicitadas. El documento deberá ser enviado con sus nombres, apellidos y cédula de identidad.
- Duración: 2 horas.
- Uso de los materiales y referencias de todo lo que se dio durante el curso.

Semana 5:

- **Tema 5: HDFS: Almacenamiento grandes cantidad de datos**
  - Introducción a HDFS.
  - Almacenamiento datos HDFS.
  - Hive (SQL), Pig (scripting).
  - Conectando Hadoop con R.

- **Tema 6: Apache Spark**
  - Ecosistema completo
  - Transformaciones y acciones
  - Spark Streaming y Spark SQL
  - Aprendizaje automático con Spark MLlib

- Referencias bibliográficas:
    - Dean, J. y Ghemawat, S. (2008). MapReduce: simplified data processing on large clusters. Communications of the ACM, 51(1), 107-113. https://doi.org/10.1145/1327452.1327492
    - Ghemawat, S., Gobioff, H. y Leung, S-T. (2003). The Google File System. A CM SIGOPS Operating Systems Review, 37(5), 29-43. https://doi.org/10.1145/1165389.945450

Semana 6:

- **Tema 7: Big data y protección de datos personales**
    - Cómo cumplir con la protección de datos en el big data
    - Evaluaciones de impacto (PIA/EIPD)
    - Derecho de información
    - Protección de datos en EE.UU. y otros países

- **Tema 8: La disociación de datos personales y técnicas de anonimización**
    - Introducción y objetivos
    - Definiciones
    - La disociación y anonimización de datos
    - Técnicas de anonimización
    - K-anonimato y sus variantes
    - Riesgos asociados a las técnicas de anonimización
    - Principios a la hora de construir un data warehouse

- Referencias bibliográficas:
    - Article 29 Working Party. (2013). Opinion 03/2013 on purpose limitation [Archivo PDF]. https://ec.europa.eu/justice/article-29/documentation/opinionrecommendation/files/2013/wp203_en.pdf
    - Directiva 95/46/CE del Parlamento Europeo y del Consejo, de 24 de octubre de 1995, relativa a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos. Diario Oficial de las Comunidades Europeas L 281, 23 de noviembre de 1995, pp. 31-50.
    - Google. (s.f.). Términos del servicio de Google. https://policies.google.com/terms?hl=es
    - ISACA. (2013). Big data: impactos y beneficios. ISACA.
    - MediaBuzz. (agosto de 2013). The 7 foundational principles of privacy by design https://www.mediabuzz.com.sg/best-practices-aug-13/the-7-foundational-principles-of-privacy-by-design
    - Norwegian Data Protection Authority. (13 de octubre de 2014). Resolution Big Data. 36th

International Conference of Data Protection and Privacy Commissioners, Mauricio. http://globalprivacyassembly.org/wp-content/uploads/2015/02/Resolution-Big-Data.pdf

Semana 7: Uso de Big Data en el INE

- Captura de datos de las Encuestas a través de dispositivos móviles con consistencias de datos interactivas

Semana 8: Examen Final

- Objetivo: Evaluar el aprendizaje de los estudiantes a través de actividades teóricas y prácticas de laboratorios.
- El examen consistirá en dos partes:
    - 1) Preguntas que deberá marcar el alumno con la respuesta correcta.
    - 2) Planteamiento de un problema y con las herramientas disponibles realizar las actividades solicitadas. El documento deberá ser enviado con sus nombres, apellidos y cédula de identidad.
- Duración: 2 horas.
- Uso de los materiales y referencias de todo lo que se dio durante el curso.