Presentación

CIDæN

Luis de la Ossa

Máster en Ciencia de Datos e Ingeniería de Datos en la Nube

CIDaeN: Máster en Ciencia de Datos e Ingeniería de Datos en la Nube

- Máster **propio** de la **Universidad de Castilla-La Mancha**
- Modalidad **on-line**
- 60 Créditos ECTS (48 en módulos + 12 en Trabajo Fin de Grado)
- Impartición: 3 de octubre de 2022 a 26 de mayo de 2023
- Fecha límite para la conclusión y entrega de trabajos: 8 de Septiembre de 2023

En realidad, la carga del TFM es menor. Lo veremos después.

Programa

Ciencia de Datos

1. Herramientas básicas

3. Análisis exploratorio de datos

4. Fundamentos de *Machine Learning*

5. Técnicas avanzadas de *Machine Learning*

6. *Deep Learning*

7. Series Temporales

8. Texto, redes y sistemas de recomendación

9. Visualización y BI

Ingeniería de Datos

2. Adquisición, almacenamiento y preparación de datos

10. Servicios en la nube (Introducción a *AWS*)

11. Servicios avanzados en la nube (*AWS*)

12. Arquitecturas y procesos *Big Data*

14. Creación y despliegue de servicios

A partir del módulo 4 los contenidos **no** se impartirán en orden: intercalaremos ciencia e ingeniería.

1. Introducción. Herramientas Básicas

Contenidos

- Presentación
- Herramientas: Github, Docker, entornos virtuales, Jupyter Notebook
- Introducción a la programación con Python
- Numpy
- Manipulación de datos con Pandas
- Elaboración de gráficas con Matplotlib

Aunque Docker corresponde a este módulo, se verá después, en un seminario que se impartirá en primavera.

2. Adquisición, almacenamiento, y preparación de datos

Contenidos

- Arquitecturas y tecnologías para el almacenamiento de datos
- Serialización
- Bases de datos relacionales (*SQL*)
- Modelado multidimensional
- Bases de datos *NoSQL*
- Expresiones regulares y *Web Scraping*
- *Tidy Data*

Durante el máster se verán otras bases de datos. Será posteriormente, cuando se comience con los contenidos de AWS.

3. Análisis exploratorio de datos

Contenidos

- Introducción a la estadística descriptiva
- Introducción a la estadística inferencia (test estadísticos)
- Regresión lineal
- Análisis exploratorio de datos
- Visualización con *Seaborn*

4. Fundamentos de *machine learning*

Contenidos

- Introducción al aprendizaje automático
- Aprendizaje supervisado
  - Regresión lineal y logística
  - Evaluación y validación
  - Algoritmos de aprendizaje supervisado
- Aprendizaje no supervisado
  - Agrupamiento
- *scikit-learn*

5. Técnicas avanzadas de *machine learning*

Contenidos

- Aprendizaje supervisado
  - Ensembles
  - Aprendizaje sobre grandes volúmenes de datos
  - Conceptos teóricos
  - Técnicas avanzadas: Datos no balanceados, AutoML, etc.
- Aprendizaje no supervisado
  - Reducción de dimensionalidad
  - Detección de anomalías
  - Etc
- Otras técnicas

6. *Deep learning*

Contenidos

- Introducción a Deep Learning
- Redes Convolucionales
- Manejo de datos masivos y herramientas
- Uso de modelos preentrenados y ajuste de modelos
- AutoKeras
- Aplicaciones

7. Series temporales

Contenidos

- Métodos clásicos de pronóstico
- Machine *learning* y *deep learning* sobre series temporales
- Prophet
- Bases de datos para series temporales
- Herramientas para ingestión y visualización
- AWS Forecast

8. Text mining, redes sociales y sistemas de recomendación

Contenidos

- Introducción a NLP y *machine/deep* learning sobre texto.
- Análisis de grafos y redes
- Sistemas de recomendación
- Recuperación de la información

9. Visualización y BI

Contenidos

- Visualización de la información
- Herramientas de visualización y BI
- Plotly Dash
- Procesamiento de datos espaciales con GeoPandas
- AWS QuickSight

10. Servicios en la nube

Contenidos

- Introducción a AWS
- Infraestructura
- Seguridad
- Redes y entrega de contenido
- Almacenamiento
- Bases de datos
- Arquitectura en la nube
- Monitorización y escalado automático

En este módulo no se hará proyecto, sino que se seguirá una formación oficial de AWS en la que os apoyaremos.

11. Servicios avanzados en la nube

Contenidos

- S3 y DynamoDB
- Computación serverless con Lambda
- API Gateway
- Servicios de mensajería SNS y SQS
- Servicios cognitivos

12. Arquitecturas y procesos *Big Data*

Contenidos

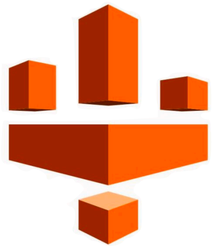
- Arquitecturas Big Data
- Ecosistema Hadoop
- Apache Spark
  - Spark Core
  - Spark DataFrames y Spark SQL
  - Spark Streaming
  - Spark Lib
- Apache Spark en producción (AWS)
- Productivización de Machine Learning

databricks

13. Almacenes de datos: *Datalakes*

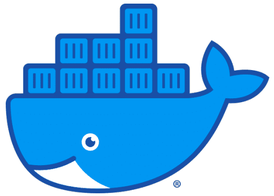
Contenidos

- Arquitecturas *Datalake* en AWS
- Change Data Capture
- Orquestación de ETLs
- Prefect cloud



14. Despliegue de servicios basados en datos

Contenidos



- Docker en AWS
- Despliegue de modelos IA en AWS
- Infraestructura como código
- CI/CD

Metodología

- 12 clases de 45 minutos por semana
- Lunes, martes y miércoles de 17:00 a 20:15 (descanso de 18:30 a 18:45)
- Algunos jueves tendremos clase de 17:00 a 18:30.
- Clases teóricas y seminarios prácticos.
- Ocasionalmente, las sesiones incluirán alguna actividad
- Las clases se grabarán, y los enlaces están disponibles en campus virtual.

Sabemos que muchos de vosotros trabajáis y haréis el máster por las noches o fines de semana… os lo pondremos fácil!

Evaluación

- En cada módulo (menos el 10) se elaborará un proyecto (***capstone***)
- Los proyectos se harán de forma autónoma
- Se establecerán plazos de entrega
- En algunos módulos se propondrán también hojas de ejercicios y/o cuestionarios.
- Ninguna actividad de evaluación se hará en tiempo real
  - Todas las podéis hacer en casa a vuestro ritmo
- Colgaremos las notas en campus virtual, y os enviaremos feedback

Trabajo fin de máster

En realidad el trabajo no tiene que requerir 12 ECTS, ya que a lo largo del curso vais a desarrollar mucho trabajo práctico.

- 3 líneas:
  - Podéis proponer trabajos relacionados con vuestra actividad profesional, académica o intereses
    - Estas propuestas deberán ser concretas, bien definidas
    - En muchos casos, la materialización de una propuesta dependerá de la disposición de datos
  - Se propondrán algunos trabajos entre los que podéis elegir, relacionados con las distintas temáticas
  - Otra posibilidad consiste en hacer y entregar un **portfolio** con ejercicios adicionales a partir de cada capstone
    - Estos ejercicios serán extensiones propuestas por cada uno.

Esto supone una gran ventaja, porque luego se echa el tiempo encima.

- Para optar a un sobresaliente, habrá que defender el TFM
- Habrá varios plazos de defensa, desde antes de navidad hasta septiembre
  - Quienes vengan del curso anterior o tengan formación, pueden empezar desde octubre a hacerlo.

Comunicación

- Os daremos todo el material, enlaces a clases, etc a través de campus virtual.
- Todas las semanas os enviaremos un correo con la planificación de las clases.
- Estaremos en permanente contacto a través de Slack.
  - La comunicación se hará a través de los canales de cada módulo
  - Hemos de intentar participar, preguntando, comentando y respondiendo
  - A la vez tenemos que intentar no saturar los canales con información o comentarios innecesarios.
  - La comunicación en privado con los profesores está restringida:
    - Participaremos ocasionalmente en los canales
    - Contestaremos a los privados en franjas de tiempo establecidas.

Debéis tener en cuenta que sois muchos alumnos (entre antiguos y nuevos), con distintos horarios, y es imposible atender Slack constantemente

¿Preguntas?