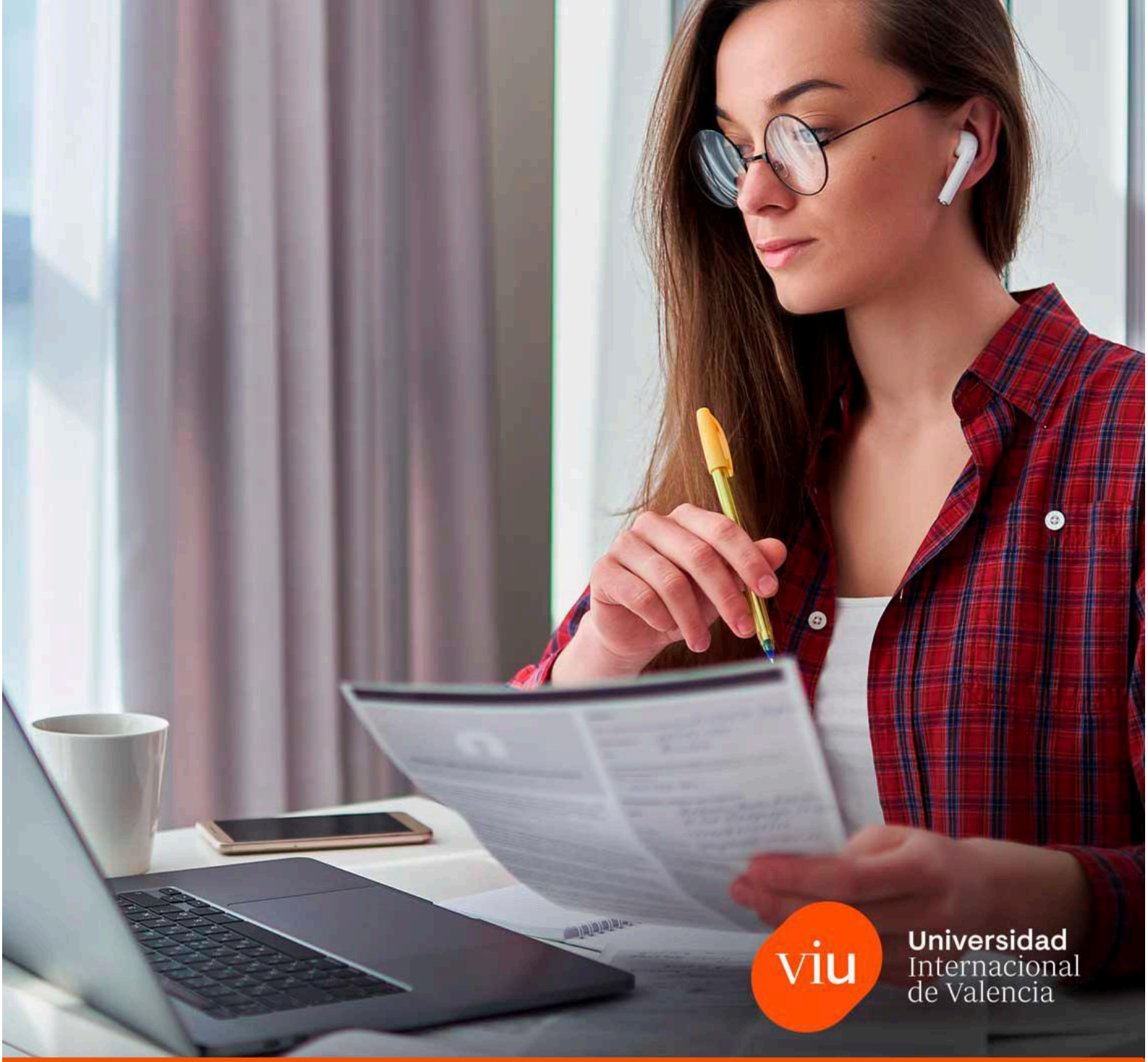

Actividad Práctica I

Metodologías de Gestión y Diseño de Proyectos Big Data

Título: Máster en Big Data y Data Science
Módulo: "Metodologías de Gestión y diseño de proyectos Big Data"
Créditos: 6 ECTS
Código: 13MBID
Curso: ABR 2025-26

Índice

Actividad Práctica - Automatización de un proyecto de ciencia de datos - Parte I

La gestión efectiva de un proyecto de ciencia de datos tiene por objetivo lograr que el mismo produzca los resultados esperados por los interesados en el mismo. Para ello, en la actualidad, la automatización se ha convertido en una práctica habitual, incorporando técnicas y herramientas a las fases de un proyecto para garantizar su escalabilidad y la capacidad de reproducción de la experimentación asociada.

Se comenzará por el establecimiento de lineamientos técnicos en torno al entorno operativo de trabajo. Posteriormente, se definirán las acciones a ejecutar para mejorar diferentes aspectos de las tareas del proyecto. Estas tareas serán ordenadas en iteraciones a ejecutar (que representan las partes de la actividad práctica) y entregables a presentar.

Si bien no se trata de un escenario real, se utilizarán técnicas de gestión ágil de proyectos y diversas herramientas y librerías para lograr la automatización buscada y así cumplir con los objetivos del curso.

Resumen

DESCRIPCIÓN	
Introducción	En la presente Actividad Práctica (AP) se abordará el inicio y la definición general de la gestión del proyecto de ciencia de datos a desarrollar a lo largo de la práctica de la asignatura. Concretamente, se comenzará por la definición y replicación del entorno tecnológico, se proseguirá con la especificación del proyecto y se dará inicio a la ejecución de las acciones de automatización sobre el mismo.
Objetivo	El objetivo principal de esta primera actividad será aplicar enfoques ágiles de gestión al desarrollo del proyecto de ciencia de datos, incorporando además prácticas iniciales de DataOps orientadas a la automatización y estandarización del trabajo. Las tareas estarán centradas en: ● La revisión y reformulación de los objetivos de negocio del proyecto. ● La comprensión de los datos disponibles. ● La evaluación de la calidad de los datos mediante herramientas de validación automatizada. ● La implementación de pipelines que permitan ejecutar partes del proceso de análisis y transformación de los datos de manera independiente de los recursos actuales (*notebooks*). Complementariamente, se espera que se incorpore al flujo de trabajo un conjunto de herramientas propias de proyectos de ciencia de datos modernos, aplicadas con criterios basados en las buenas prácticas de la industria. La adquisición de estas competencias contribuye a la formación de los maestrandos, ya que combina la dimensión metodológica de CRISP-DM y Scrum con prácticas técnicas de DataOps, fortaleciendo la capacidad de abordar proyectos de forma sistemática, escalable y colaborativa.

Trabajo previo	Lectura del material docente de la parte específica que se encuentra disponible desde el comienzo del curso en la carpeta: **Contenidos de la asignatura >** **Materiales del docente.** Visualización de las videoconferencias asociadas a la teoría de este apartado de la asignatura que se encontrarán disponibles en: **Videoconferencias >** **Acceso a clases virtuales >** **Grabaciones.**
Metodología	En las videoconferencias teóricas (VC) se expondrán al alumno conocimientos, material e indicaciones suficientes para que pueda elaborar una unidad didáctica basada en el aprendizaje y enseñanza por competencias en matemáticas e informática. Las actividades de cada sesión práctica (SP) se centrarán en poner en práctica y asentar los conocimientos adquiridos en la videoconferencia teórica anterior.
Tarea para el e-portfolio	Al finalizar esta primera parte de la AP, los estudiantes deberán elaborar un reporte con los avances logrados en la ejecución del proyecto utilizando tanto los *datasets* como las plantillas de documentación y código provistos por el docente. La ejecución deberá realizarse utilizando la plantilla de repositorio provista por la cátedra y los recursos de código que deberán ser ejecutados previamente y sincronizados con la copia remota del repositorio en uso antes de ser enviados para su corrección por parte del docente. Los elementos mínimos a desarrollar se encuentran definidos en el presente documento; además, se establecen otros de carácter opcional para los estudiantes que podrán incorporarse si así lo creen conveniente. Complementariamente, se solicita elaborar un registro de la actividad desarrollada (informe / memoria) que consta de una descripción de alto nivel de las actividades desarrolladas sobre los datos.

Forma de entrega	La memoria de cada actividad se subirá al sitio de la asignatura en formato PDF (ningún otro formato será admitido), a la misma deberá acompañarse el enlace al repositorio de GitHub que se haya generado para el desarrollo del proyecto. El desarrollo de la AP podrá ser realizado en forma grupal. Los informes/memorias deben incluir una página principal con el título de actividad, fecha y nombre de los autores del trabajo. En cualquier caso, las entregas se realizarán dentro de los plazos establecidos en el calendario de la asignatura o en las fechas establecidas para las convocatorias a examen. Las entregas solo serán válidas si se realizan a través del espacio del Campus Virtual de la asignatura: **Actividades y Evaluación Final >** **Actividades Formativas >** **Espacio de entrega - Actividad Práctica 1.**
Criterios de evaluación	• Formato (10%) • Completitud de los requerimientos (40%) • Adecuado uso de las herramientas (40%) • Integración de elementos extra en las acciones requeridas (10%)
Fecha de entrega	
1ª Entrega AP I	*Martes 04 de noviembre de 2025 hasta las 23:59*
2ª Entrega AP I	*Viernes 14 de noviembre de 2025 hasta las 23:59*
2ª Instancia de Entrega de las AP	*Lunes 02 de marzo de 2026 hasta las 23:59* *(válido para las dos AP de la asignatura)*

Actividades a desarrollar

Archivos y recursos

Los archivos de datos para el desarrollo de la presente AP se encuentran disponibles en el campus virtual de la asignatura en la carpeta **Contenidos de la asignatura > Materiales del docente > AP1**. En tal directorio encontrarán el enlace al repositorio de tipo plantilla (*template*) para ejecutar el proyecto, los archivos de datos de origen y las plantillas para la documentación.

Escenario de trabajo

El desarrollo de las actividades prácticas que inician en el presente documento se basan en el siguiente escenario:

Las autoridades de una entidad financiera disponen de una base de datos con los resultados de sus acciones de marketing sobre su cartera de clientes. En una instancia anterior del proyecto se ha desarrollado una **versión inicial (MVP[1]) de un producto de datos para realizar la predicción** de la efectividad de tales campañas a partir de analizar si resultan en la constitución de un depósito a plazos.

El objetivo principal será incorporar al proyecto diferentes cuestiones de DataOps y MLOps para lograr automatizar y estandarizar el trabajo a realizar en el mismo. Para ello, se proponen las siguientes acciones:

- Versionar todos los elementos del proyecto (código y datos).
- Incorporar automatización en las validaciones de calidad de los datos.
- Generar pipelines para las tareas de limpieza y transformación de los datos.
- Incorporar registro de las experimentaciones realizadas y los modelos generados.
- Generar componentes para poder desplegar el producto generado a los usuarios finales.
- Actualizar la documentación del proyecto.

Adicionalmente, se podrá evaluar la inclusión de acciones de automatización en la fase de despliegue del producto generado hacia algún proveedor en la nube a partir de acciones específicas en el repositorio del proyecto.

En las próximas secciones se presentan las operaciones a realizar en la presente Actividad Práctica.

[1] MVP, sigla para Mínimo Producto Viable (en inglés).

Actividad 1: replicación del entorno

Sobre el repositorio *template* generado para la asignatura, realizar una copia del mismo para obtener los siguientes elementos:

- Estructura del directorio del proyecto
- Archivos de configuración
- Archivos de datos
- Archivos de desarrollo de la versión inicial del proyecto
- Directorio de ejemplos para algunas de las herramientas de gestión a emplear

De esta manera, se obtiene un repositorio funcional en el usuario de GitHub de cada estudiante que podrá ser clonado a su equipo personal para poder trabajar.

En el equipo de destino deberá contar con una instalación de Python 3.x y una de las siguientes herramientas (a elección, pudiendo utilizarse otras según preferencias de cada estudiante) para la confección de un entorno virtual sobre el cual estarán instaladas las librerías a utilizar: *venv*, *conda* o *uv*. Opcionalmente, se podrá hacer uso de un conjunto de comandos mediante *make* para agilizar este paso.

Una vez disponible el directorio del proyecto en la instancia local (el equipo personal de cada estudiante), se podrán ejecutar los ejemplos del caso siguiendo las guías disponibles en el mismo repositorio.

Entregable esperado: repositorio generado en la plataforma y URL de acceso.

Actividad 2: adaptación de la visión del proyecto

Con base en lo establecido en el escenario detallado anteriormente, se pasará a conformar la visión y los objetivos del proyecto. A nivel general se reconoce el siguiente objetivo principal del proyecto:

- Poder predecir con un margen de confianza del 80% la efectividad de las acciones de marketing sobre la cartera de clientes de la entidad.

Considerando que el mismo está cumplido, inicialmente, con el desarrollo disponible del producto de datos que conforma el punto de partida de la presente actividad, se deberá avanzar con el abordaje de cuestiones como:

- ¿Qué riesgos enfrenta el proyecto para pasar a producción el producto generado?
- ¿Qué expectativas tiene el negocio respecto de escalabilidad, actualización y mantenimiento de la solución provista?
- ¿Quiénes serán los responsables del monitoreo del producto en uso?
- ¿Cuáles métricas serán utilizadas para monitorear al producto?

Estas, y otras actividades, serán las que van a ejecutarse a lo largo de las dos actividades prácticas previstas para la asignatura. La base metodológica a emplear será

CRISP-DM a nivel de ciencia de datos, siendo el estándar de la industria desde hace años. Para la gestión de las actividades se propone un enfoque ágil basado en Scrum. Y las acciones que se desean integrar al proyecto se relacionan con las prácticas de DataOps y MLOps que, actualmente, son aplicables a este tipo de proyectos (ver figura 1).

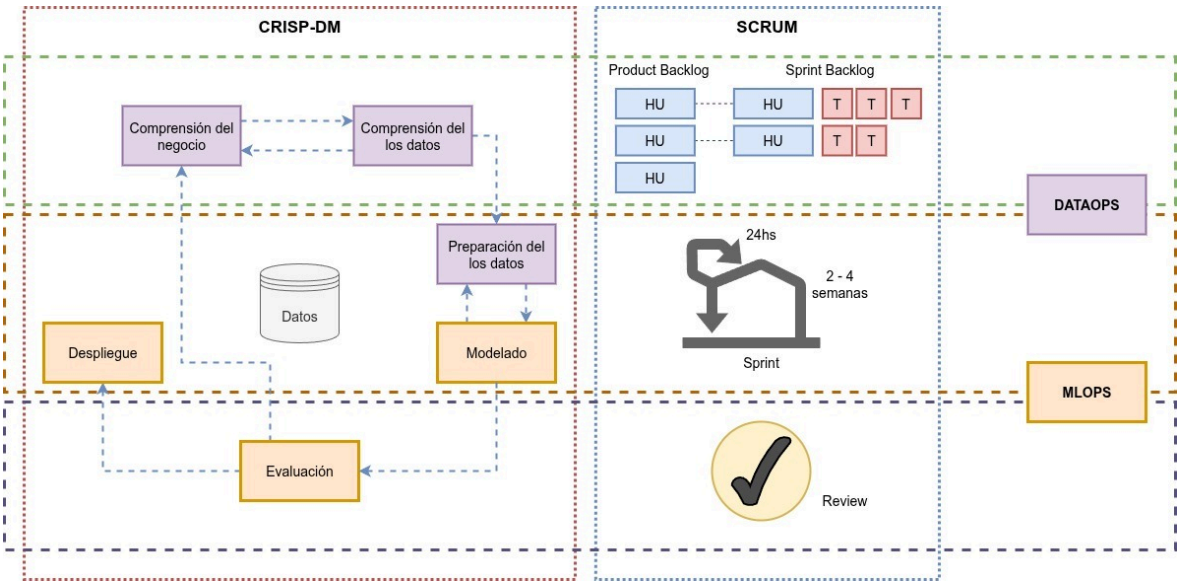

***Figura 1**. Integración de CRISP-DM con Scrum y acciones de DataOps y MLOps.*

La gestión del proyecto será realizada con la herramienta online **GitHub - Projects**. A través de ella se generará un proyecto nuevo en el cual se reflejará el trabajo a realizar desde este punto hasta el final de la Actividad Práctica II.

Se comenzará a especificar una visión inicial del **Product Backlog** de esta fase del proyecto a fin de contar con una planificación de alto nivel que sirva como base para realizar la planificación detallada de las iteraciones a ejecutar.

Entregable esperado: Documentación del proyecto (memoria de trabajo actualizada en las secciones involucradas)

Actividad 3: definición de la primera iteración

En función del Product Backlog y de la planificación de versiones del proyecto, se podrá definir una **1ra. Iteración** con su correspondiente **backlog** de tareas.

Las historias de usuario / ítems de trabajo seleccionados deberán ser resolubles en una iteración (*sprint*), salvo excepciones que el equipo de trabajo podrá considerar. Con el sprint backlog definido, se iniciará el sprint en la herramienta para poder pasar a la etapa de ejecución del mismo.

A partir de este punto, se podrá comenzar a desarrollar el trabajo correspondiente a la iteración, registrando el progreso a través del movimiento de las diferentes historias de

usuario entre las columnas del tablero del proyecto. De esta manera, a simple vista se podrá tener dimensión de sus avances y del trabajo que resta completar para lograr los objetivos previstos en cada iteración.

Entregable esperado: Enlace al proyecto generado en la plataforma online (con una descripción básica de los ítems de trabajo y definición de las iteraciones a ejecutar)

Actividad 4: comprensión de los datos

Se tomarán los datos del escenario junto al producto generado hasta el momento (MVP) y se realizará una exploración inicial sobre los mismos a fin de identificar estructuras, características básicas de la implementación realizada y detectar con qué elementos se deberá trabajar en la continuidad del proyecto.

En esta misma actividad, para la actividad prevista de evaluación de calidad de datos, se va a tomar lo establecido en el **Capítulo 13 del DMBoK** y la **norma ISO 25012**. En función de lo establecido en ambas fuentes, se analizará la evaluación preexistente y se modificará la misma para que pueda ser automatizada, además de evaluar la incorporación de otras verificaciones que puedan ser de utilidad para el negocio.

En este apartado se trabajará con librerías de *testing* como PyTest que permiten generar *tests* mediante código que es capaz de ser incorporado en flujos de trabajo automatizables.

Entregable esperado: desarrollo y documentación de los casos de prueba mediante las librerías utilizadas (al menos una) para automatizar la evaluación de calidad de los datos.

Actividad 5: preparación de los datos

Considerando las modificaciones que pudieran haber sido realizadas previamente sobre los datos del escenario, se generarán scripts para su implementación. Con estos archivos nuevos se pasará a definir pipelines a fin de hacer que tales operaciones puedan ser reproducidas de forma automática cuando se requiera.

Estos nuevos elementos pasarán a versionarse y determinar, si existen, restricciones para su aplicación. Además, se podrá determinar si es necesario generar algún tipo de salida o reporte intermedio que también podría ser versionado.

De esta manera, las fases iniciales del proyecto modificarán su documentación para incorporar las consideraciones que derivaron en la automatización de algunas acciones y las decisiones que hubieran sido tomadas en el proceso.

Entregable esperado: desarrollo y documentación de los *scripts* de transformación de los datos a aplicar para automatizar las acciones de esta fase del proyecto. Adicionalmente, se espera la implementación y documentación de los *pipelines* de las actividades ejecutadas hasta este punto del proyecto.