

## **Informe Semillero Análisis de Datos (AD). Semestre 2021-2.**

### **Enfoque.**

El semillero está orientado a aprender, asimilar y aplicar todo lo relacionado con los programas Numpy, Pandas, Matplotlib y Scikit-Learn. Estos programas forman parte de las herramientas esenciales, creadas en Python para el análisis de datos.

### **Objetivo**

Desarrollar las competencias necesarias para realizar análisis de datos exploratorio e inferencial.

### **Estudiantes.**

El semillero está dirigido y puede ser de utilidad para todos los estudiantes de la Universidad EIA en sus diferentes escuelas. No tiene prerrequisitos.

### **Metodología.**

Encuentros semanales (este semestre: miércoles 14-16 y sábados 8-10) con presentación de los temas propuestos, formulación de tareas prácticas para cada tema. El semillero esta dividido en dos partes: herramientas básicas(Numpy, Pandas y Matplotlib) y bases de Machine Learning con Scikit-Learn. Cada una de estas partes esta acompañada de un proyecto de investigación para aplicar lo aprendido en el semillero.

### **Programa.**

- #1. 04.08.21 Python para el AD.
- #2. 11.08.21 Numpy I.
- #3. 18.08.21 Numpy II.
- #4. 25.08.21 Numpy III.
- #5. 01.09.21 Numpy IV.
- Parciales 04-08.09.21**
- #6. 15.09.21 Pandas I. Formulación del Proyecto I (EDA).
- #7. 22.09.21 Pandas II.
- #8. 25.09.21 Pandas III.**
- #9. 29.09.21 Pandas IV. Proyecto II (IDA).
- #10. 06.10.21 Series Temporales.
- #11. 09.10.21 Matplotlib I.**
- #12. 13.10.21 Matplotlib II.
- #13. 16.10.21 Scikit\_Learn I.**
- #14. 20.10.21 Scikit-Learn II.
- #15. 27.10.21 Hyperparámetros y Validación de modelos.
- #16. 03.11.21 Revisión de proyectos 1 y 2.

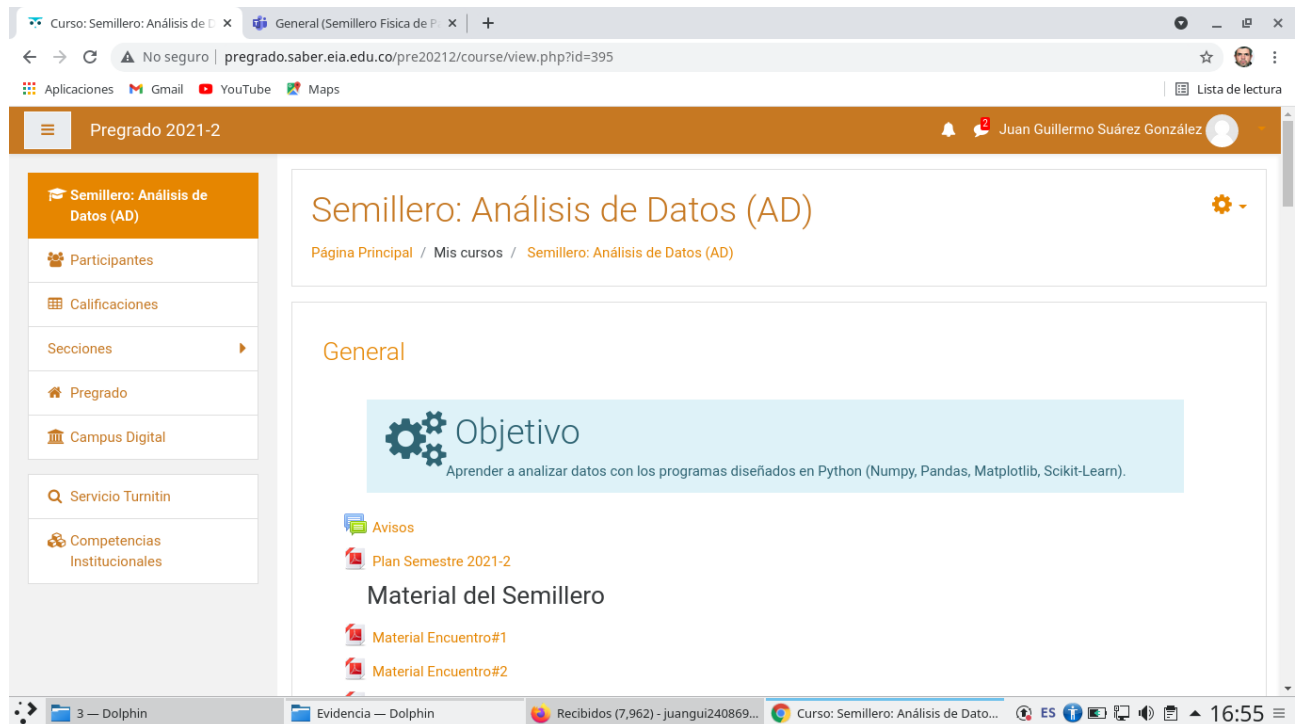
### **06.11.21 Fin de clases.**

Cada encuentro está acompañado de un notebook de Jupyter. Los notebooks están disponibles en el aula en Moodle del semillero.

## Aula en Moodle.

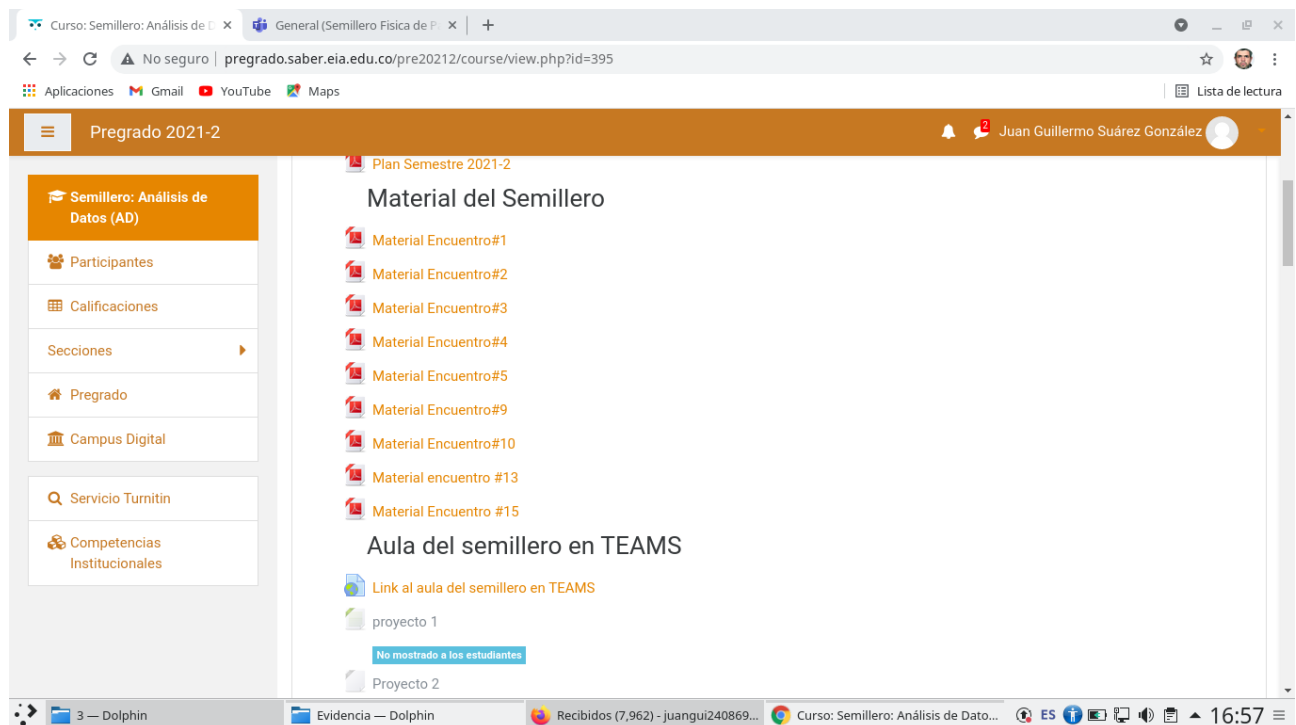
El aula en Moodle incluye:

### 1. Objetivo del semillero y plan.



The screenshot shows a Moodle course page for 'Semillero: Análisis de Datos (AD)'. The page title is 'Semillero: Análisis de Datos (AD)'. The breadcrumb trail is 'Página Principal / Mis cursos / Semillero: Análisis de Datos (AD)'. The main content area is titled 'General' and features a blue box with the heading 'Objetivo' and the text 'Aprender a analizar datos con los programas diseñados en Python (Numpy, Pandas, Matplotlib, Scikit-Learn)'. Below this, there is a section for 'Material del Semillero' with a list of items: 'Avisos', 'Plan Semestre 2021-2', 'Material Encuentro#1', and 'Material Encuentro#2'. The left sidebar contains navigation options: 'Participantes', 'Calificaciones', 'Secciones', 'Pregrado', 'Campus Digital', 'Servicio Turnitin', and 'Competencias Institucionales'. The top navigation bar shows 'Pregrado 2021-2' and the user's name 'Juan Guillermo Suárez González'.

### 2. Materiales complementarios.



The screenshot shows the same Moodle course page, but with the 'Material del Semillero' section expanded to show a list of 15 'Material Encuentro' items (Material Encuentro#1 through Material Encuentro #15). Below this list, there is a section titled 'Aula del semillero en TEAMS' with a link 'Link al aula del semillero en TEAMS'. There are also two project items listed: 'proyecto 1' and 'Proyecto 2'. A blue box indicates 'No mostrado a los estudiantes'. The left sidebar and top navigation bar are the same as in the previous screenshot.

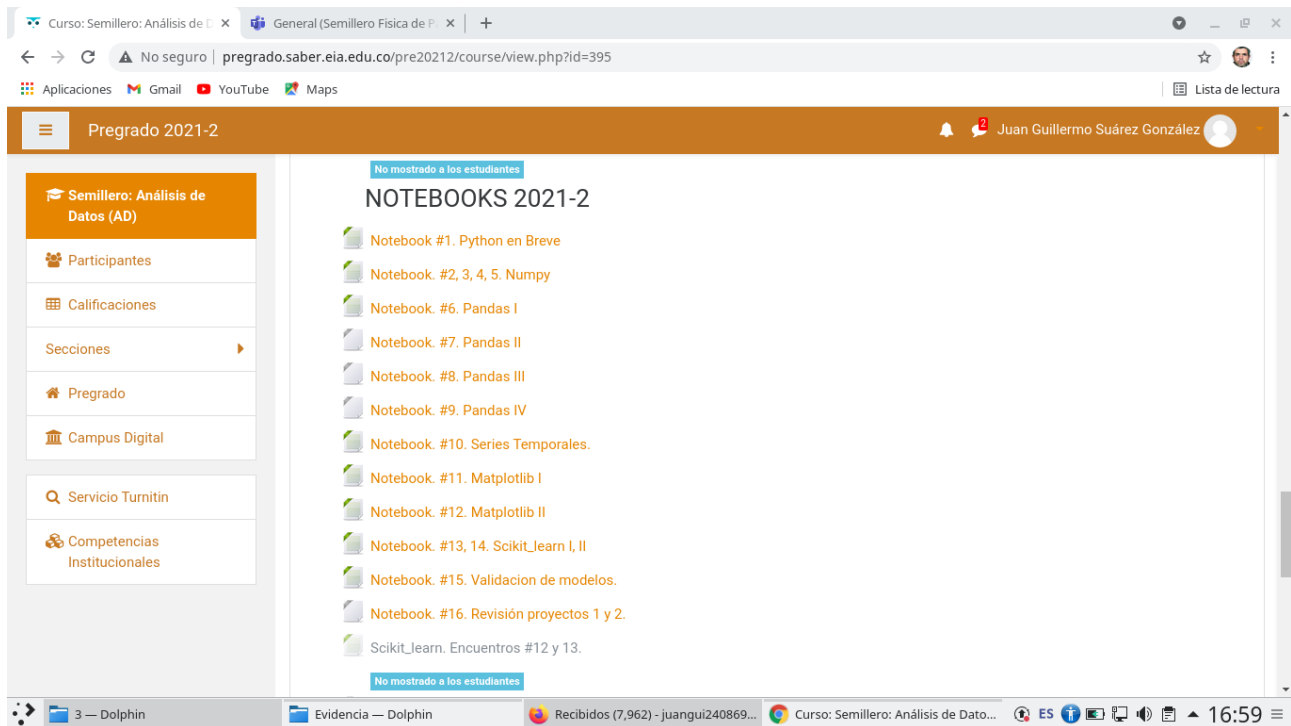
### 3. Proyectos.

The screenshot shows a Moodle course page for 'Pregrado 2021-2'. The left sidebar contains navigation options: Semillero: Análisis de Datos (AD), Participantes, Calificaciones, Secciones, Pregrado, Campus Digital, Servicio Turnitin, and Competencias Institucionales. The main content area is titled 'Proyectos' and lists several items, each with a 'No mostrado a los estudiantes' label: Proyecto 3, proyectos 2020-2, Proyecto I. 2021-2, and Proyecto II. 2021-2. Below this is a section titled 'NOTEBOOKS' with a list of IPyNB notebooks: Encuentro #6, #7, #8, and Matplotlib Encuentro #9. The browser address bar shows 'pregrado.saber.eia.edu.co/pre20212/course/view.php?id=395' and the system clock shows 16:57.

### 4. Vídeos.

The screenshot shows the same Moodle course page for 'Pregrado 2021-2', but the main content area is titled 'Vídeos Semestre 2021-2'. It lists 13 video encounters, each with a 'No mostrado a los estudiantes' label and a video icon: Encuentro #1 (04.08.21 Python en breve), Encuentro #2 (11.08.21 Numpy I), Encuentro #3 (18.08.21 Numpy II), Encuentro #4 (25.08.21 Numpy III), Encuentro #5 (31.08.21 Numpy IV), Encuentro #6 (15.09.21 Pandas I), Encuentro #7 (22.09.21 Pandas II), Encuentro #8 (25.09.21 Pandas III), Encuentro #9 (29.09.21 Pandas IV), Encuentro #10 (05.10.21 Series Temporales), Encuentro #11 (09.10.21 Matplotlib I), Encuentro #12 (13.10.21 Matplotlib II), and Encuentro #13 (16.10.21 Scikit-Learn I). The system clock shows 16:58.

## 5. Notebooks 2021-2



The screenshot shows a web browser window with the URL `pregrado.saber.eia.edu.co/pre20212/course/view.php?id=395`. The page title is "Pregrado 2021-2" and the user is identified as "Juan Guillermo Suárez González". The main content area is titled "NOTEBOOKS 2021-2" and lists 16 notebooks:

- Notebook #1. Python en Breve
- Notebook #2, 3, 4, 5. Numpy
- Notebook #6. Pandas I
- Notebook #7. Pandas II
- Notebook #8. Pandas III
- Notebook #9. Pandas IV
- Notebook #10. Series Temporales.
- Notebook #11. Matplotlib I
- Notebook #12. Matplotlib II
- Notebook #13, 14. Scikit\_Learn I, II
- Notebook #15. Validacion de modelos.
- Notebook #16. Revisión proyectos 1 y 2.
- Scikit\_Learn. Encuentros #12 y 13.

### Certificado.

Para obtener el certificado del semillero es necesario participar en las actividades del semillero durante dos semestres consecutivos y presentar los respectivos proyectos.

### Resultados.

Logramos completar en su totalidad el plan propuesto al inicio del semillero. Durante el semillero realizamos dos proyectos de investigación: el primero dedicado al análisis exploratorio de datos y el segundo dedicado al análisis inferencial de datos. Para los proyectos utilizamos la base de datos **UCI\_Credit\_Card.csv**. Esta base de datos contiene información sobre el historial de pagos de las tarjetas de crédito en Taiwan durante abril – septiembre 2005. El objetivo del proyecto 2 consistía en predecir si un cliente estaría o no en condición de cumplir con las obligaciones de pago de sus cuotas de tarjeta de crédito en meses posteriores con base en los pagos hechos en meses anteriores. Se utilizaron varios algoritmos para las predicciones. La precisión alcanzada fue del 80% en promedio. El cambio de algoritmo permitió reducir el número de falsos negativos (clientes que no cumplieron con sus obligaciones pero fueron clasificados por los algoritmos como personas al orden del día). Los resultados de los proyectos están detallados en un archivo que se agrega a la evidencia del semillero.

### Conclusiones.

Al finalizar el semillero logramos:

1. considerar todos los temas planeados,
2. realizar algunas tareas para la asimilación del material expuesto,
3. realizar dos proyectos de investigación con aplicación de los temas estudiados,
4. para el próximo semestre planeamos continuar con el trabajo aplicando métodos estadísticos en computación con ayuda de los programas Scipy y Statsmodels.