DMC | INSTITUTE

Módulo Python for Analytics – Sesión 01

Diploma Data Scientist

Docente: Geider Nuñuvero

PERFIL PROFESIONAL

Educación:

- UNIVERSIDAD NACIONAL DE INGENIERIA : Ingeniero Estadístico
- edX MITx : MicroMasters Program Statistics and Data Science
- UDACITY : Machine Learning Engineer Nanodegree Program

Experiencia Laboral:

- Analytics Product Manager : Backus
- Lead Data Scientist: Interbank
- Consultor Senior en Data Science: Tottus y entidades públicas
- Senior Project Lead in Data Science : Rimac Seguros
- Sr Data Scientist : Belcorp
- Analista de Data Mining: Telefónica Ingeniería de Seguridad

REGLAS

Se requiere **puntualidad** para un mejor desarrollo del curso.

Para una mayor concentración **mantener silenciado el micrófono** durante la sesión.

Las preguntas se realizarán **a través del cha**t y en caso de que lo requieran **podrán activar el micrófono**.

Realizar las actividades y/o tareas encomendadas en **los plazos determinados**.

Identificarse en la sala Zoom con el primer nombre y primer apellido.

Objetivo

Este curso busca **enseñar los fundamentos de Python** para el análisis de datos, incluyendo la instalación de entornos, manejo de estructuras de datos y el uso de la librería Pandas para manipulación y análisis eficiente de datos

Silabo

- Introducción a Python.
- Instalación (descriptivo) , ambientes y librerías.
- Variables y tipos de datos.
- Estructuras de datos (Listas, tablas, diccionarios).
- Estructuras de control (Indentación, condicionales y bucles).
- Manipulación de datos (Librería Pandas).

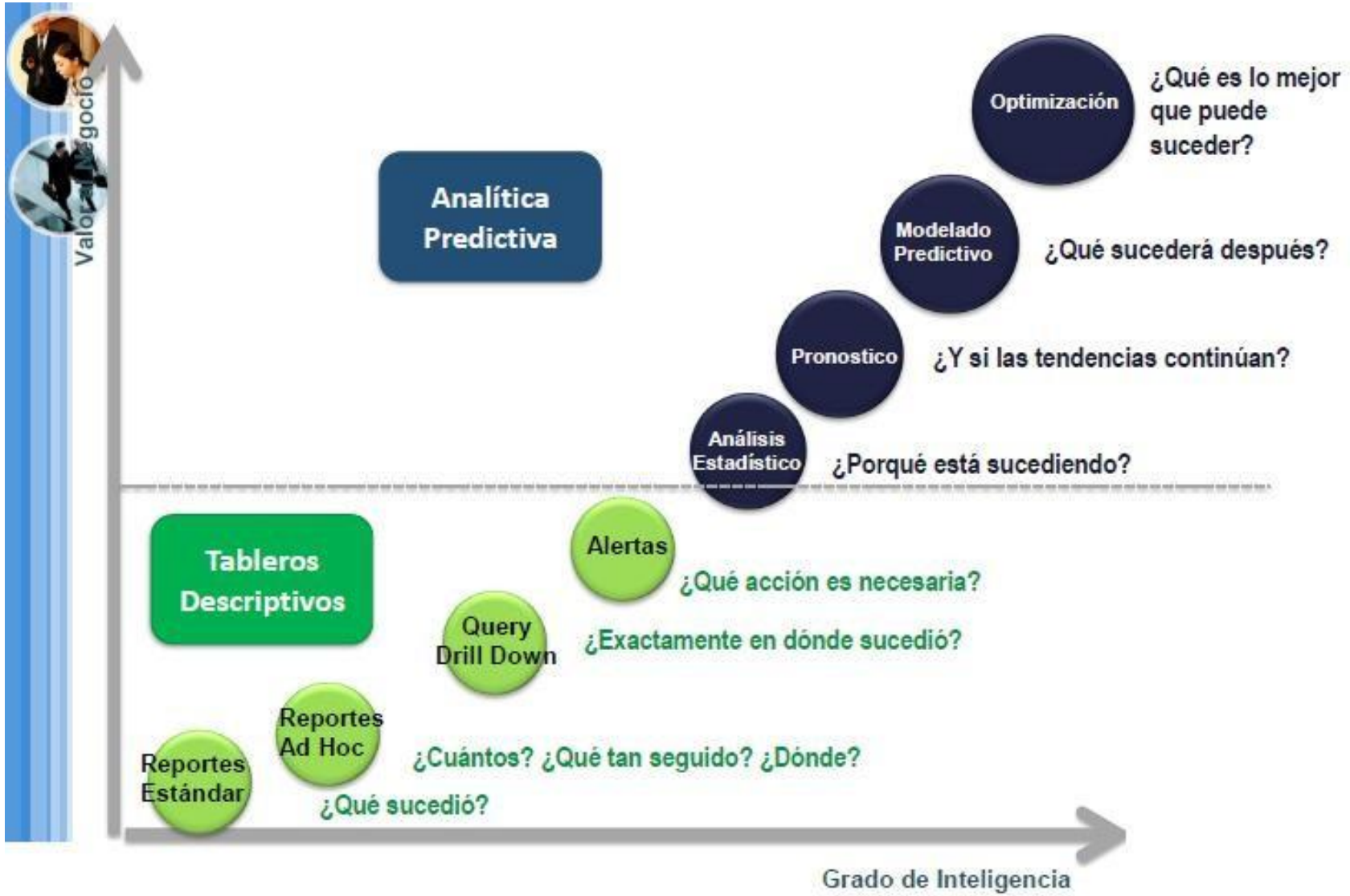
¿Qué es Analytics?

- **Analytics** puede ser definido como el proceso que abarca el uso de técnicas estadísticas, sistemas de información e investigación de operaciones para explorar, visualizar, descubrir; y comunicar patrones y tendencias en la data.

- **Analytics** puede convertir data en información útil, transformándola en un activo para las empresas.
- En pocas palabras, la **analytics** nos ayuda a ver información y datos significativos que de otro modo no podríamos detectar.

- La **analítica empresarial** se centra en el uso de conocimientos derivados de los datos para **tomar decisiones** más informadas que ayudarán a las organizaciones a aumentar las ventas, reducir los costes y realizar otras mejoras comerciales.

Niveles de Analytics



Importancia de tener un proyecto de Analytics

Mejor entendimiento del cliente.

- Lealtad, ciclo de vida, perfiles, comportamientos, …
- Mercado Saturado.

Estrategia efectiva en la toma de las decisiones.

- Rápida y de gran conocimiento.
- Memoria Inteligente.
- Más allá que aplicación de modelos.
- Conocimiento del negocio aplicado a los modelos.
- Resultados confiables, accionables y repetibles.

¿Qué es Python?

Python es un lenguaje de programación de alto nivel, de código abierto ampliamente usado en la actualidad debido a su versatilidad, simplicidad y popularidad en diversas aplicaciones.

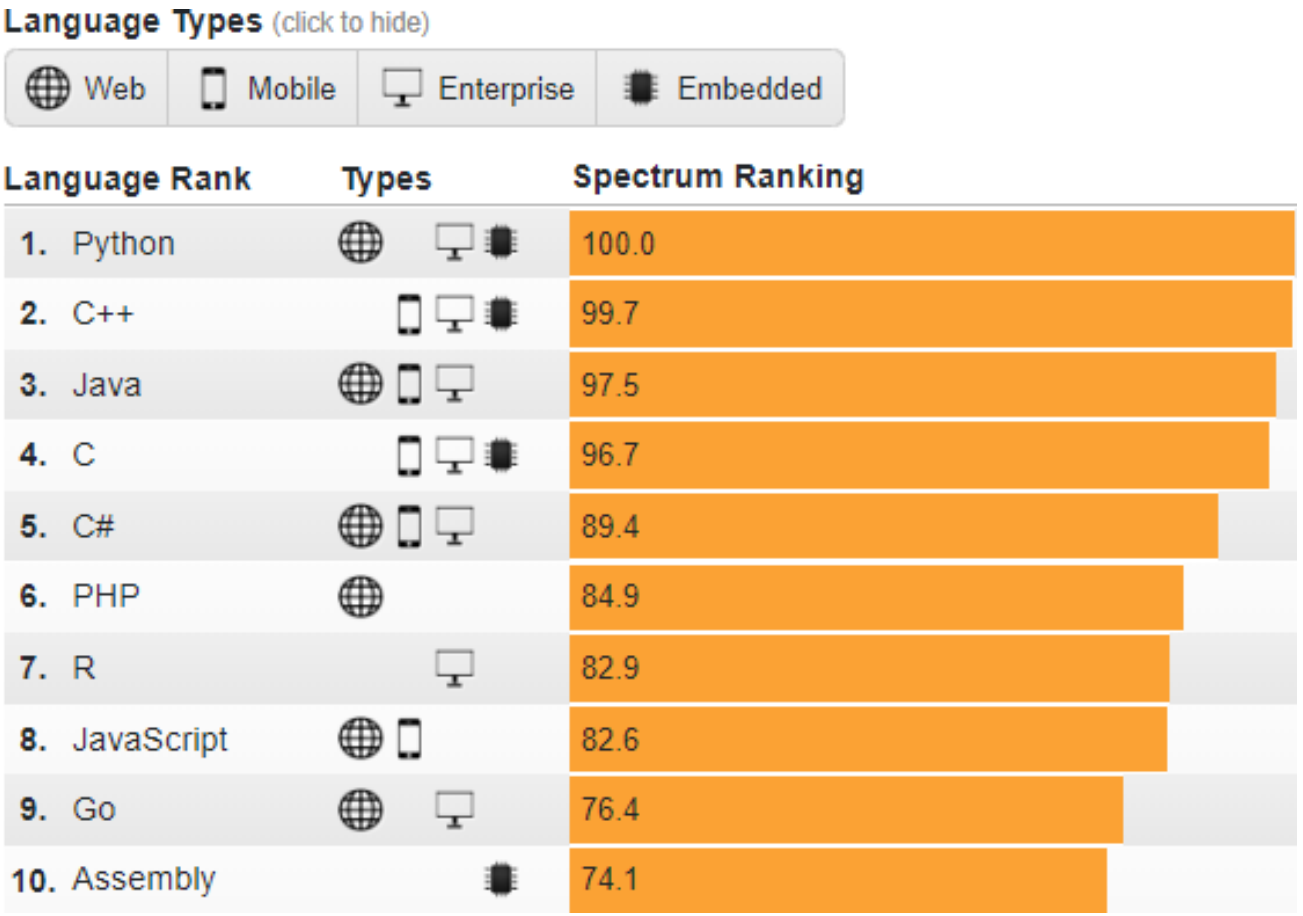
Veamos algunas de las características que lo hacen famoso

¿Por qué Python?

- ✓ Es un lenguaje de programación interpretado. ¡Bye compilador!
- ✓ Posee un tipado dinámico, es decir no requiere que se declare el tipo de dato de cada variable creada y además puede cambiar conforme se le vaya asignando valores.
- ✓ Recomendado para aprender el programa, sintaxis muy sencilla y legible (como si estuviéramos hablándole al ordenador)
- ✓ ¡Código abierto! Completamente gratis, libre de usar y distribuir sin perder presencia en ámbitos comerciales.
- ✓ Es multiplataforma, se puede utilizar y ejecutar en Windows, Linux, Mac, etc.
- ✓ Enorme cantidad de módulos y paquetes respaldados por la comunidad .

Ventajas de Python sobre otros lenguajes

- ✓ IEEE Spectrum, la organización profesional más grande del mundo dedicada a la ingeniería y las ciencias aplicadas, elabora anualmente un ranking de los lenguajes de programación más usados, preferidos y relevantes, que este año ubicó a Python en el primer lugar.



Ventajas de Python sobre otros lenguajes

1. Está desarrollado bajo una licencia de código abierto, por lo que es de libre uso y distribución, incluso para uso comercial.
2. Es uno de los lenguajes de programación más versátiles que existen, puede ser usado en muchos campos diferentes. Es decir, permite programar desde videojuegos hasta aplicaciones móviles.
3. Es fácil de aprender. Si comprendes Python, podrás entender más fácilmente otros lenguajes de programación. Esto quiere decir que es una excelente opción si apenas incursionas en el mundo de los desarrolladores.
4. Gracias a su popularidad, cuenta con una amplia comunidad que organiza eventos, conferencias, reuniones y colabora en materia de códigos e información.
5. El Python Package Index (PyPI) aloja miles de módulos de terceros para Python. Tanto la biblioteca estándar de Python como los módulos aportados por la comunidad permiten infinitas posibilidades.

¿Quiénes utilizan Python?

- Google & YouTube.
- Instagram desarrollado en Django (framework de Python).
- Recomendación de series y películas en Netflix.
- Casi la cuarta parte de Facebook está escrito en Python.

Proceso de instalación Python con Jupyter

1. Descargar Anaconda (https://www.anaconda.com/download/success#download) desde la página oficial, selecciona la versión para Windows (64-bit).

2. Ejecutar el instalador y aceptar términos.

3. Elegir instalación 'Just Me' y ruta por defecto.

4. No agregar Anaconda al PATH.

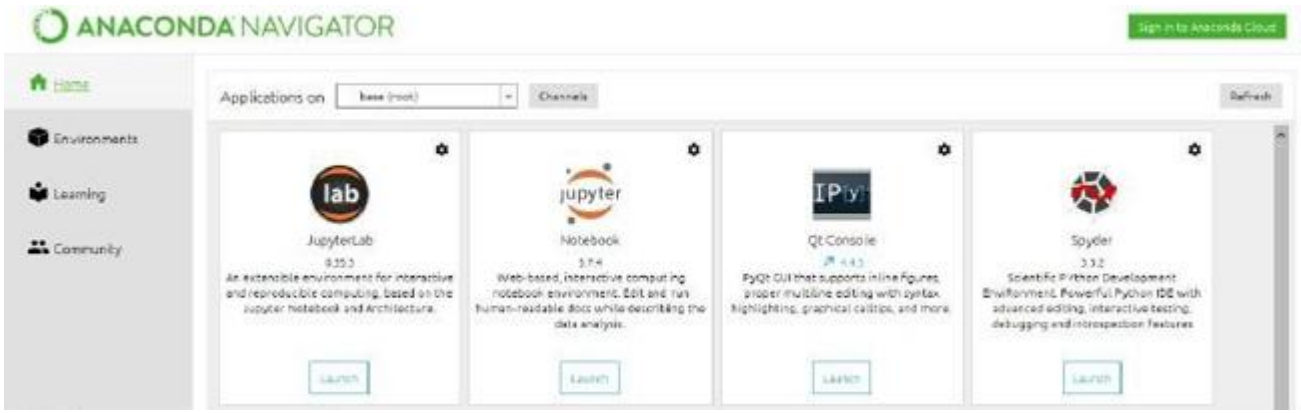
5. Registrar Anaconda como Python por defecto

6. Finalizar e iniciar Anaconda Navigator.

Software y editores de código para ciencia de datos

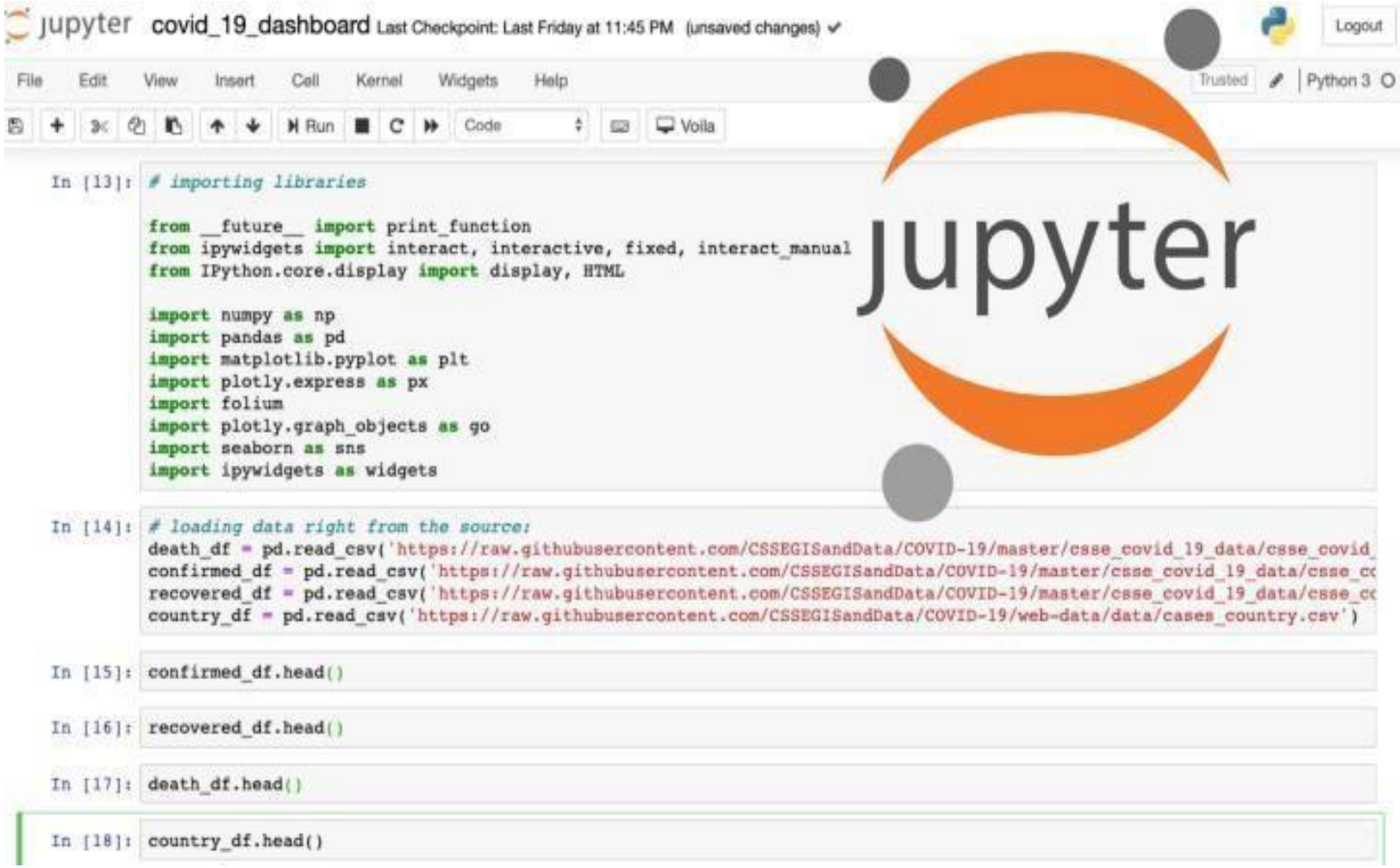
Elegir el computador y editor de código a utilizar:

- ✓ Kernel de Anaconda (requiere instalación en propia PC) y utilizar un editor de los tantos que ofrece (Visual Studio Code, Jupyter Notebook, Spyder, etc.).
- ✓ Utilizar una solución en la nube que provea el computador y/o editor de código como Google Colab (.ipynb).

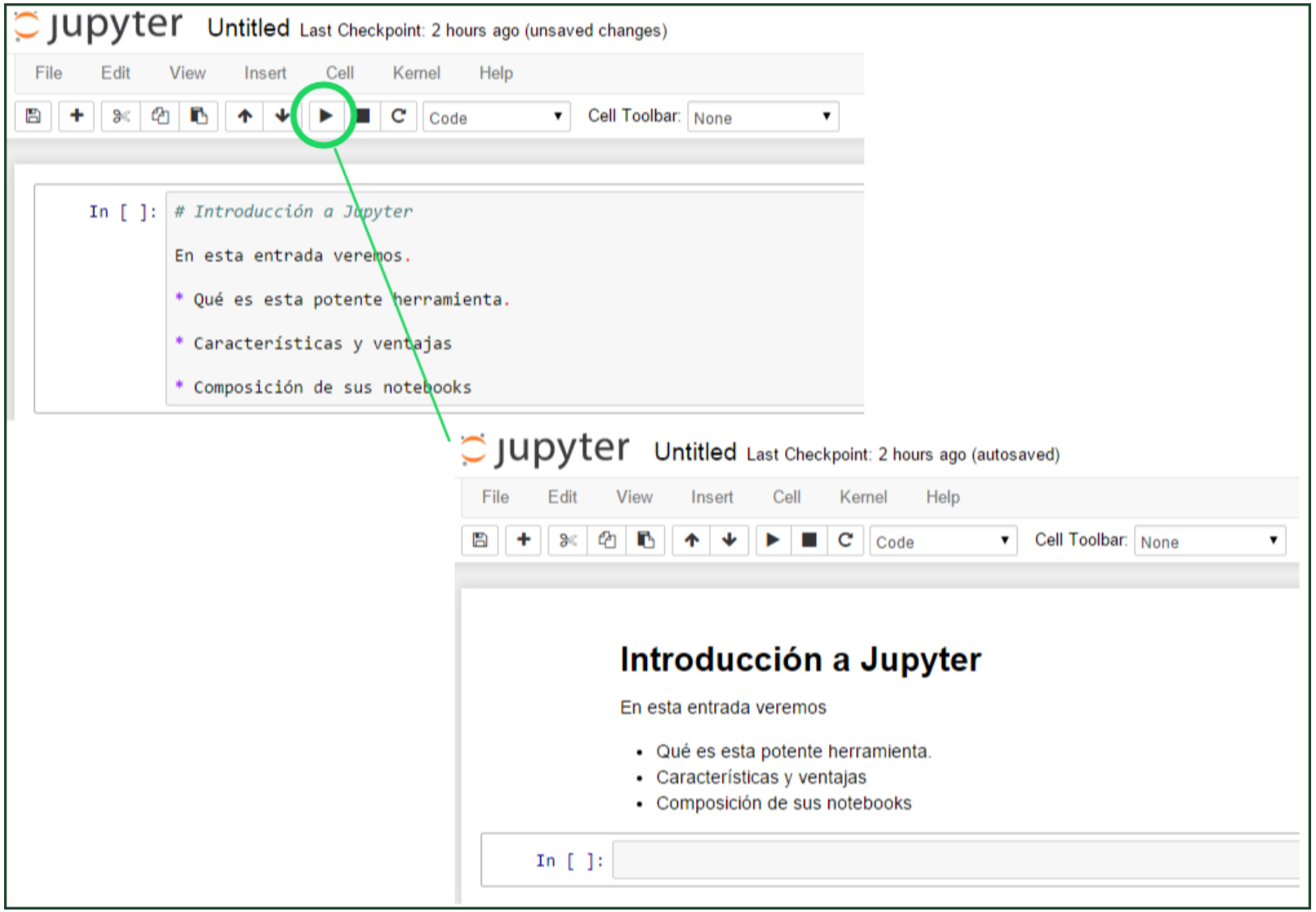



ANACONDA NAVIGATOR
Upgrade Now
Sign in to Anaconda Cloud
Home
Environments
Learning
Community
Documentation
Developer Blog
Feedback
Applications on
base (root)
Channels
Refresh
jupyterlab
0.31.12
An extensible environment for interactive and reproducible computing, based on the Jupyter Notebook and Architecture.
Launch
notebook
5.4.0
Web-based, interactive computing notebook environment. Edit and run human-readable docs while describing the data analysis.
Launch
qtconsole
4.3.1
PyQt GUI that supports inline figures, proper multiline editing with syntax highlighting, graphical calltips, and more.
Launch
spyder
3.3.0
Scientific PYthon Development EnviRonment. Powerful Python IDE with advanced editing, interactive testing, debugging and introspection features
glueviz
0.13.3
Multidimensional data visualization across files. Explore relationships within and among related datasets.
orange3
3.13.0
Component based data mining framework. Data visualization and data analysis for novice and expert. Interactive workflows with a large toolbox.

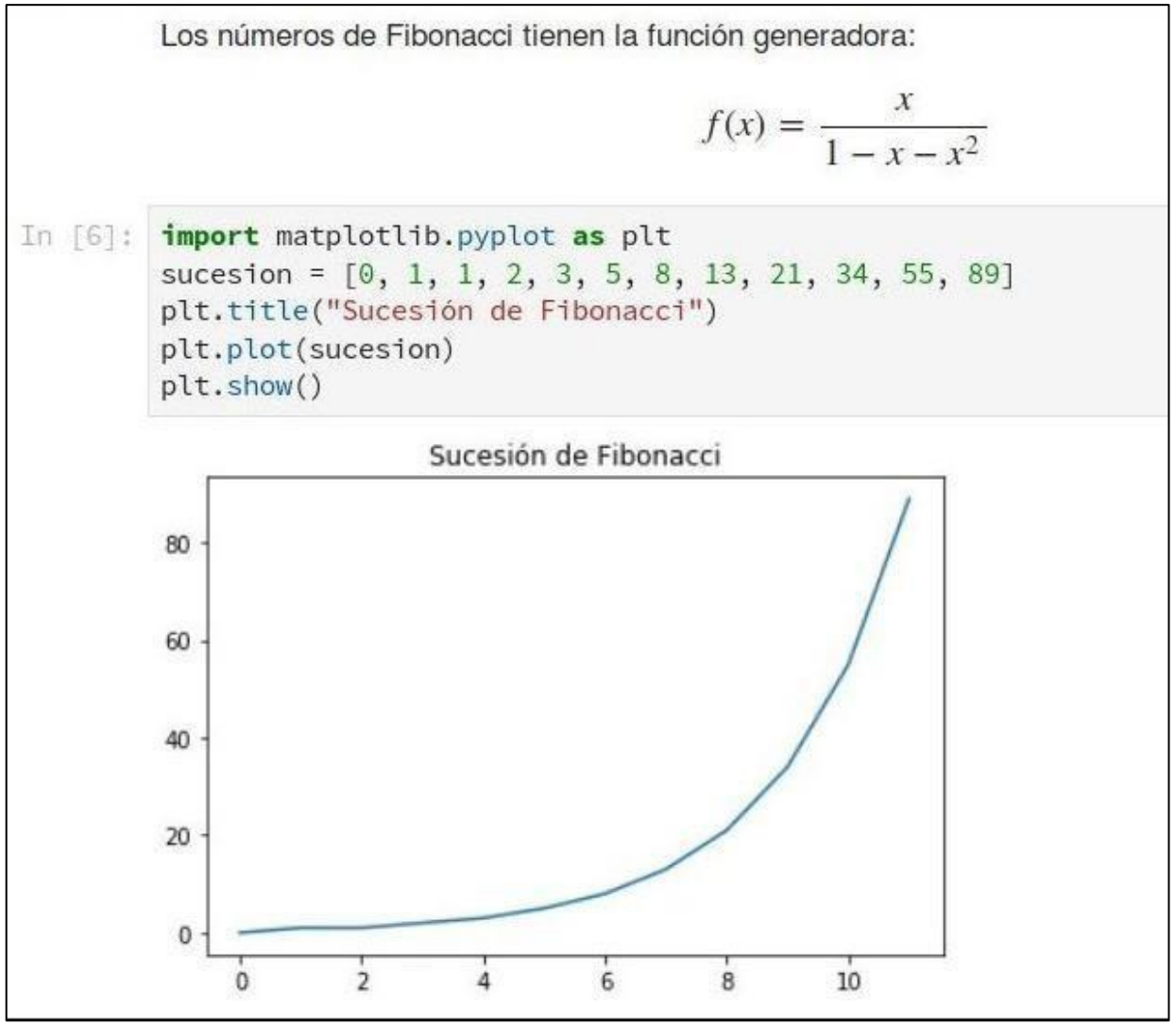
Jupyter Notebook



Jupyter Notebook



Jupyter Notebook



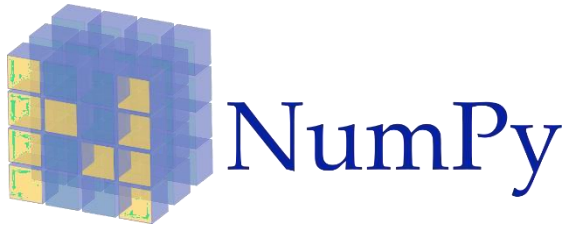
Google Colab

Es la solución de Google para el desarrollo de proyectos de Ciencia de Datos alojados completamente en la Nube, entre sus ventajas tenemos:

- ✓ Brinda una máquina virtual con 13 gb de ram y 50 gb de disco.
- ✓ Es completamente gratuito aunque pronto habrá una versión de paga.
- ✓ Se sincroniza con nuestros archivos de Google Drive y nuestros notebooks se guardan automáticamente en una carpeta llamada Colab Notebooks.

Principales librerías o módulos en Python

Principales librerías o módulos en Python



NumPy: el nombre viene de numerical Python, Es una de las librerías más útiles en Python para el cálculo matricial, implementando una gran diversidad de funciones y transformaciones algébricas.

SciPy: el nombre corresponde a la expresión Scientific Python. SciPy evoluciona las funciones de NumPy añadiendo, por ejemplo la posibilidad de calcular series de Fourier y ejecutar procesos de optimización.además facilita trabajar con matrices dispersas.

Pandas: se trata una de las librerías estrella de Python, Agrupa una serie de funcionalidades clave que facilita el siempre indispensable y engorroso trabajo de limpiar, formatear y procesar adecuadamente los datos.

Principales librerías o módulos en Python

Matplotlib: indispensable a la hora de generar gráficos que nos ayuden a comprender mejor las características de nuestros datos, análisis y modelos.

Sckit Learn for Machine Learming: partiendo de la base de las funcionalidades ofrecidas por NumPy, SciPy y Matplotlib, ofrece utilidades que permite ejecutar de forma sencilla algoritmos de aprendizaje automático y modelos estadísticos, incluyendo regresión, clasificación, clustering y reducción de dimensiones.

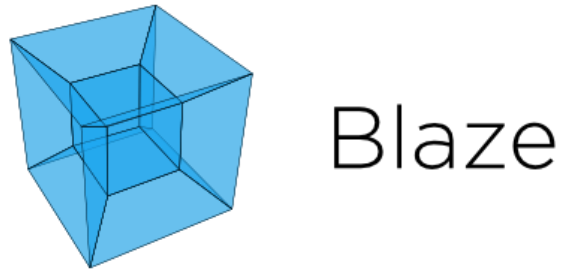
Statsmodels: librería enfocada en el modelo estadístico, La exploración de los datos, creación de modelos estadísticos y realización de test estadísticos es el objetivo de esta librería. StatsModels proporciona un listado considerable de estimadores estadísticos complementados con funcionalidades para la visualización del trabajo realizado.

Principales librerías o módulos en Python

Seaborn: extraordinaria librería de visualización basada en MatploLib. El objetivo de Seaborn es ayudar a las tareas de exploración de los datos, proporciona amplia y diversa información estadística de forma gráfica, lo que facilita su comprensión.

Bokeh: librería que ofrece gráficos iterativos y dinámicos. Bokeh es una librería eficiente que facilita la creación de cuadros de mando visualmente atractivos.

Blaze: esta librería parte de la base proporcionada tanto por NumPy como Pandas para facilitar el tratamiento de datos distribuidos y generados de forma continua (streaming). Permite acceder a bases de datos en distintas ubicaciones y repositorios variados cmo bases de datos NoSQL o ficheros de archivos distribuidos. Blaze proporciona herramientas para la visualización de datos en este contexto.



Software y componentes utilizados en la industria para machine learning (ML)

Plataforma	Enfoque	Lo más destacado	Ecosistema / Integración
Azure Machine Learning	ML end-to-end + MLOps	AutoML, notebooks, pipelines, generación IA	Azure completo: redes, seguridad, CI/CD
AWS SageMaker	ML en nube robusta	ML incorporado, frameworks gestionados + integración AWS	Amplia integración con servicios AWS
Vertex AI	ML + Generative AI	Gemini, AutoML, Agent Builder, vídeo generativo	Modelos Google, infraestructura GCP
databricks	DL + Big Data + ML	Lakehouse, Spark, MLflow, LLM propio, Mosaic AI	Multi-nube, fuerte para data engineering + ML
sas viya	Analítica avanzada ML	Automatización, interpretabilidad, módulos sectoriales	Multi-nube, fuerte en gobernanza y cumplimiento

¡MUCHAS GRACIAS!