GUSTAVO CORONEL
DESARROLLA SOFTWARE

CIENCIA DE DATOS: TALLER DE FUNDAMENTOS DE MACHINE LEARNING CON PYTHON

Módulo 01

FUNDAMENTOS DE CIENCIA DE DATOS: DEL CONCEPTO A LA PRÁCTICA CON PYTHON

Dr. Eric Gustavo Coronel Castillo
Docente UNI
gcoronel@uni.edu.pe

ÍNDICE

1. Presentación

La Ciencia de Datos se ha consolidado como una disciplina clave para comprender fenómenos y apoyar decisiones basadas en evidencia. En este módulo se presenta una visión general del campo, su contexto histórico, sus componentes y aplicaciones, así como una mirada práctica al mercado laboral. Además, se establece el entorno de trabajo que se usará en el curso (preferentemente con Anaconda y Jupyter), con el fin de garantizar una experiencia de aprendizaje accesible y reproducible. Finalmente, se desarrollan los primeros pasos con Python para cargar datos, explorar variables y generar visualizaciones básicas, sentando las bases para los módulos posteriores de machine learning.

2. Objetivo

Objetivo general

- Introducir los fundamentos de la Ciencia de Datos y preparar un entorno funcional con Python para realizar análisis exploratorio y visualización básica.

Objetivos específicos

- Definir la Ciencia de Datos, sus componentes y tipos de análisis, y reconocer su evolución y relevancia en el mercado laboral.
- Instalar y configurar el marco de trabajo del curso con Anaconda, incluyendo la gestión de entornos y paquetes.
- Ejecutar un primer flujo de trabajo en Python para cargar, inspeccionar, limpiar de forma básica y visualizar datos.
- Interpretar resultados elementales del EDA (distribuciones, relaciones simples y patrones iniciales) y documentar hallazgos de manera clara.

3. ¿Qué es Ciencia de Datos? | Mercado laboral

3.1 Definición de Ciencia de Datos

La ciencia de datos es un campo multidisciplinario que integra el conocimiento del dominio de aplicación (economía, medicina, finanzas, tecnología, etc.) con la estadística, el análisis de datos, la informática y las matemáticas para comprender y analizar fenómenos reales mediante datos (SAS Institute, 2025). Su objetivo principal es la resolución de problemas complejos que requieren el procesamiento y análisis avanzado de datos, aplicados a industrias de cualquier tipo.

La Ciencia de Datos combina datos, métodos estadísticos y computación para extraer conocimiento útil y apoyar decisiones (por ejemplo: describir fenómenos, explicar

relaciones, predecir resultados y proponer acciones). En términos prácticos, trabaja con el ciclo completo: obtener datos → prepararlos → analizarlos → comunicar hallazgos → desplegar soluciones cuando corresponde

Según SAS Institute (2025), la ciencia de datos implica convertir datos en información útil (insights) mediante el uso de inteligencia artificial, aprendizaje automático, deep learning, análisis, estadística y algoritmos. Este proceso permite a las organizaciones extraer valor de cantidades masivas de datos estructurados y no estructurados.

Una idea clave es que, cuando los datos crecen en volumen, variedad y velocidad, se requieren herramientas y arquitecturas (almacenamiento, cómputo, procesos) para gestionarlos y analizarlos de forma costo-efectiva; la ciencia de datos se apoya en ese ecosistema.

Contexto Histórico

Año / periodo	Hito	Aporte al concepto de Ciencia de Datos	Referencia
Década de 1960 (mencionado en 1974)	Peter Naur introduce el término “data science”	Usa el término como sustituto o alternativa a “ciencias computacionales”, vinculándolo al tratamiento y entendimiento de los datos.	Naur (1974)
1962	John W. Tukey publica *The Future of Data Analysis*	Anticipa la evolución del análisis de datos como un campo que se expande más allá de la estadística matemática tradicional.	Tukey (1962)
2009–2011 (aprox.)	Uso del título profesional “Data Scientist”	El término se consolida como rol profesional; se asocia a funciones reales en empresas tecnológicas (LinkedIn, Facebook).	Davenport & Patil (2012)
Octubre de 2012	Harvard Business Review publica *Data Scientist: The Sexiest Job of the 21st Century*	Populariza la profesión a nivel mundial y define al científico de datos como perfil híbrido (programación, análisis, comunicación y asesoría).	Davenport & Patil (2012)

El término fue acuñado en los años 60 por el científico danés Peter Naur como sustituto de las ciencias computacionales (Naur, 1974). En 1962, John W. Tukey precedió al concepto en su artículo The Future of Data Analysis, explicando una evolución de la estadística matemática (Tukey, 1962). Sin embargo, como campo reconocido, la ciencia de datos no existió hasta alrededor de 2009-2011, cuando DJ Patil y Jeff Hammerbacher acuñaron el término científico de datos como título profesional mientras trabajaban en LinkedIn y Facebook respectivamente (Davenport & Patil, 2012).

En octubre de 2012, la revista Harvard Business Review publicó el artículo Data Scientist: The Sexiest Job of the 21st Century, escrito por Thomas H. Davenport y DJ Patil (Davenport & Patil, 2012), que catapultó la profesión a la atención mundial. Este

artículo describió el perfil del científico de datos como un híbrido entre hacker de datos, analista, comunicador y consejero confiable.

3.2 Componentes y tipos de análisis en Ciencia de Datos

La ciencia de datos se encuentra en la intersección de tres disciplinas fundamentales:

- **Ciencia Computacional y Tecnologías de la Información:** Conocimiento de programación, manejo de bases de datos, arquitectura de datos y uso de herramientas tecnológicas para procesar grandes volúmenes de información.
- **Matemáticas y Estadística:** Fundamentos matemáticos para el modelado, análisis estadístico, inferencia y comprensión de los patrones presentes en los datos.
- **Conocimiento del Dominio:** Experiencia y comprensión del contexto específico de aplicación (agricultura, finanzas, salud, marketing, etc.) que permite formular las preguntas correctas y aplicar los análisis de manera útil.

Además de integrar programación, estadística y conocimiento del dominio, la Ciencia de Datos se aplica a distintos tipos de análisis. Estos tipos se diferencian por la pregunta que responden y el nivel de complejidad del trabajo.

En este curso se prioriza el análisis descriptivo y el análisis exploratorio (EDA), apoyados en Python, Pandas y visualización. Los enfoques predictivo y prescriptivo se mencionan solo como contexto.

3.3 ¿Qué problemas resuelve (y qué NO promete)?

La Ciencia de Datos se aplica cuando se necesita:

- **Describir:** "¿Qué está pasando?" (reportes, indicadores, distribución, variación).
- **Explicar:** "¿Por qué pasa?" (relaciones, segmentaciones, hipótesis razonables).
- **Predecir:** "¿Qué podría pasar?" (modelos predictivos con incertidumbre).
- **Recomendar:** "¿Qué conviene hacer?" (reglas, optimización, simulaciones).

Lo que NO promete por sí sola:

- **Verdades absolutas:** siempre hay incertidumbre.
- **Modelos mágicos:** sin calidad de datos, un modelo suele fallar (regla de oro: basura entra, basura sale)

3.4 Habilidades del Científico de Datos

La ciencia de datos aplicada requiere el desarrollo de habilidades en cuatro áreas principales (Davenport & Patil, 2012):

- **Programación:** Capacidad de usar lenguajes como Python, R, SQL y herramientas de análisis para automatizar procesos y resolver problemas que serían imprácticos abordar manualmente.
- **Matemáticas y Estadística:** Dominio de regresión lineal, análisis de probabilidad, pruebas de hipótesis y técnicas de modelado predictivo.
- **Comunicación:** Habilidad para explicar procesos complejos, traducir hallazgos técnicos a términos comprensibles para audiencias diversas y crear visualizaciones efectivas.
- **Conocimiento de Dominio:** Experiencia acumulada en el campo específico de aplicación que permite contextualizar los análisis y generar valor empresarial.

3.5 Proceso típico de trabajo (visión general)

En un proyecto real, este flujo se repite en iteraciones:

1. **Defina el problema**

   Incluye la decisión a tomar y el criterio de éxito (métrica).

2. **Obtener datos**

   Incluye fuentes: bases de datos, archivos, APIs, encuestas, sensores.

3. **Prepare y limpie**

   Incluye corregir nulos, duplicados, tipos de datos, outliers, consistencia.

4. **EDA (Análisis Exploratorio)**

   Incluye explorar la distribución de los datos, relaciones, patrones, segmentaciones.

5. **Modelado y evaluación de modelos (cuando aplique)**

   Este paso se aplica cuando el objetivo requiere predicción, clasificación o detección automática. Si el objetivo es descriptivo (reportes, tendencias, segmentación simple) o se resuelve con reglas claras, puede bastar con EDA y visualización. Cuando se modele, se deben comparar enfoques simples y complejos y validar con métricas adecuadas y un esquema de validación (por ejemplo, partición entrenamiento/prueba o validación cruzada).

6. **Comunicación y operacionalización**

   Presente hallazgos, limitaciones, recomendaciones y próximos pasos.

3.6 Mercado Laboral Global

Crecimiento y Demanda

El mercado laboral de ciencia de datos continúa experimentando un crecimiento robusto en 2025. Según los datos más recientes:

- El mercado global de ciencia de datos alcanzó los $178.5 mil millones en 2025, con una tasa de crecimiento anual compuesta (CAGR) del 26.5% desde 2023 (Statista, 2025).

- La Oficina de Estadísticas Laborales de Estados Unidos proyecta un crecimiento del 21% en empleos de ciencia de datos entre 2021 y 2031, más de cuatro veces la tasa de crecimiento promedio para todas las ocupaciones (U.S. Bureau of Labor Statistics, 2025).
- McKinsey estimó que las organizaciones que utilizan analítica de datos mejoran su desempeño e incrementan sus ganancias hasta en un 126% (McKinsey Global Institute, 2022).
- Según McKinsey Global Institute (2022), las empresas que implementan tecnologías de ciencia de datos e inteligencia artificial duplican sus ingresos en cinco años en comparación con las que no las utilizan.

Rangos Salariales en 2025

Los salarios en ciencia de datos han experimentado incrementos significativos. Según datos de 365 Data Science (2025) y Glassdoor (2025):

Nivel de Experiencia	Rango Salarial (USD/año)	Promedio
Nivel Inicial (0-2 años)	$80,000 - $120,000	$152,000
Nivel Intermedio (2-6 años)	$120,000 - $160,000	$140,000
Nivel Senior (6+ años)	$160,000 - $200,000+	$180,000

Nota. Los datos corresponden al mercado estadounidense. El salario promedio para científicos de datos en Estados Unidos alcanzó los $166,000 según Glassdoor (2025).

Evolución del Mercado Laboral

El mercado laboral de ciencia de datos ha experimentado transformaciones significativas:

- **Diversificación de Roles:** Mientras que en 2022 se observó una disminución del 26% en posiciones tradicionales de 'científico de datos', hubo un incremento en roles especializados como analista de datos, ingeniero de machine learning e ingeniero de datos (365 Data Science, 2025).
- **Preferencia por Experiencia:** El mercado ha evolucionado hacia profesionales con mayor experiencia. Las posiciones de nivel inicial (0-2 años) son ahora menos comunes que en 2024, reflejando una maduración de la industria (365 Data Science, 2025).

- Profesionales Versátiles: El 57% de las ofertas laborales buscan profesionales versátiles con experiencia en múltiples dominios, mientras que solo el 5% requieren científicos de datos de ciclo completo (full-stack) (365 Data Science, 2025).
- **Alternativas Educativas:** Entre el 18-26% de las ofertas laborales están abiertas a candidatos sin títulos formales, enfatizando habilidades prácticas y experiencia sobre la educación tradicional (365 Data Science, 2025).

Tendencias Clave en 2025

- **Inteligencia Artificial Generativa:** El 77% de las ofertas laborales relacionadas con IA requieren habilidades en machine learning. Las herramientas de IA generativa se están integrando como complemento al trabajo del científico de datos, no como reemplazo (365 Data Science, 2025).
- **Analítica en Tiempo Real:** Crecimiento proyectado del 23.8% CAGR hasta 2028. Las organizaciones requieren dashboards que se actualicen en tiempo real en lugar de reportes por lotes (Gartner, 2025).
- **Trabajo Híbrido:** La mayoría de empresas ha adoptado modelos híbridos post-pandemia, aunque algunas posiciones pueden requerir reubicación para roles de alto nivel.
- **Habilidades en Computación Distribuida:** Incremento en la demanda de conocimientos en Hadoop, Spark y plataformas de streaming como Kafka para manejar grandes volúmenes de datos (365 Data Science, 2025).

3.7 Mercado Laboral en América Latina y Perú

Panorama Regional

América Latina se ha convertido en un hub importante para profesionales de ciencia de datos, con salario promedio de $96,346 USD anuales para posiciones remotas según Remote Rocketship (2025), basado en 526 ofertas laborales. El mercado regional se beneficia de:

- Alineación de zonas horarias con empresas estadounidenses, facilitando colaboración en tiempo real.
- Costos competitivos que atraen outsourcing de compañías internacionales.
- Talento altamente capacitado con formación técnica sólida.

El Caso de Perú

Perú presenta un ecosistema tecnológico en expansión acelerada:

- **Crecimiento del Sector:** La industria de software crece al 9% anual, con proyecciones de alcanzar $1.2 mil millones. Se estiman 40,000 nuevas ofertas laborales en tecnología, especialmente en desarrollo de software y ciberseguridad (INEI, 2025).
- **Oportunidades Laborales:** Glassdoor (2025) reporta 131 posiciones en ciencia de datos, 24 para científicos de datos específicamente, y 52-63 para analistas de datos en Perú.
- **Rangos Salariales:** Los científicos de datos experimentados pueden alcanzar hasta $80,000 USD anuales. El salario promedio en ciberseguridad es de $25,406 USD. Desarrolladores de software ganan entre $35,600 y $54,000 USD anuales (INEI, 2025).
- **Inversión Gubernamental:** El gobierno peruano está construyendo 11 Parques de Ciencia y Tecnología y realizando mejoras sustanciales en instituciones educativas. Se han introducido iniciativas como el primer proyecto de ley para regular la IA (INEI, 2025).
- **Formación Académica:** La Universidad Nacional de Ingeniería (UNI) lidera la formación especializada con carreras profesionales de pregrado en Ingeniería de Ciberseguridad (desde 2022), Ingeniería de Inteligencia Artificial (desde 2024) y Ciencia de la Computación. Además, ofrece programas de posgrado incluyendo la Maestría en Inteligencia Artificial en la Facultad de Ingeniería Industrial y de Sistemas (UNI, 2024). Otras instituciones como la PUCP también cuentan con programas especializados en IA, ciberseguridad y ciencia de datos.
- **Ecosistema Startup:** 159 startups activas con $250 millones en inversiones esperadas para 2024, creando oportunidades en múltiples sectores tecnológicos (INEI, 2025).

3.8 Sectores de Aplicación

La ciencia de datos se aplica transversalmente en múltiples industrias:

- **Banca y Finanzas:** Detección de fraude, gestión de riesgos, análisis de crédito, recopilación de datos de clientes y trading algorítmico.
- **Salud y Farmacéutica:** Análisis de imágenes médicas, predicción de reingresos, identificación de objetivos farmacéuticos, diseño de ensayos clínicos y farmacovigilancia.
- **Retail y E-commerce:** Sistemas de recomendación, optimización de precios, predicción de demanda, análisis de comportamiento del consumidor y personalización de ofertas.
- **Marketing y Ventas:** Segmentación de clientes, análisis de campañas, automatización de marketing. El 34% de recursos de ciencia de datos se asignan a departamentos GTM - Go-To-Market (McKinsey Global Institute, 2022).
- **Telecomunicaciones:** Optimización de redes, predicción de churn (cancelación de servicios), análisis de patrones de uso.
- **Logística y Cadena de Suministro:** Optimización de rutas, predicción de demanda, gestión de inventarios, planificación de capacidad.
- **Sector Público:** Análisis de políticas públicas, optimización de servicios, detección de fraude fiscal, planificación urbana.
- **Energía:** Predicción de demanda energética, optimización de fuentes renovables (eólica y solar), gestión de redes inteligentes.
- **Agricultura:** Modelado de compensación de carbono, agricultura de precisión, predicción de rendimientos.

3.9 Perfiles Profesionales

El ecosistema de ciencia de datos incluye diversos roles especializados:

Rol	Responsabilidades Principales
Analista de Datos	Limpieza y transformación de datos, análisis exploratorio, visualización de datos, identificación de tendencias, comunicación de hallazgos. Rango salarial: $73,000-$92,000 USD.
Científico de Datos	Desarrollo de modelos predictivos, implementación de algoritmos de machine learning, análisis estadístico avanzado, experimentación A/B, interpretación de resultados. Rango salarial: $80,000-$200,000+ USD.

Ingeniero de Datos	Construcción de pipelines de datos, diseño de arquitectura de datos, ETL (Extract, Transform, Load), gestión de bases de datos, optimización de consultas, infraestructura de big data.
Ingeniero de Machine Learning	Diseño de arquitectura de IA, productivización de modelos de ML, desarrollo de APIs para inferencia, implementación de MLOps, optimización de rendimiento de modelos en producción.
Analista de Business Intelligence	Preparación de datasets, creación de dashboards, comunicación de hallazgos con visualizaciones, reporting ejecutivo, análisis de KPIs.

3.10 Desafíos y Oportunidades

Desafíos Actuales

- Brecha de Talento: Según Gartner (2025), más del 50% de los líderes empresariales reportan falta de experiencia interna en ciencia de datos.
- Calidad de Datos: Datos incompletos, inconsistentes o no estructurados dificultan la generación de insights confiables.
- Consideraciones Éticas: Privacidad de datos, sesgos algorítmicos, transparencia en modelos de IA y cumplimiento regulatorio (GDPR, HIPAA, SOC 2).
- Integración Organizacional: Muchas empresas no comprenden completamente cómo los datos y el análisis pueden transformar sus negocios.

Oportunidades

- Democratización de la Educación: Incremento de cursos online, bootcamps y certificaciones que permiten el acceso a la profesión sin títulos formales tradicionales.
- Especialización: Oportunidades en nichos como deep learning, procesamiento de lenguaje natural (NLP), computer vision, y data engineering.
- Transformación Digital: La pandemia aceleró la adopción de tecnologías digitales, incrementando la demanda de profesionales de datos en todos los sectores.
- Impacto Social: Aplicaciones en salud pública, conservación ambiental, agricultura sostenible y mejora de servicios gubernamentales ofrecen oportunidades para generar impacto positivo.

3.11 Reflexiones Finales

La ciencia de datos se ha consolidado como una disciplina fundamental en la economía digital del siglo XXI. Más allá de ser simplemente una profesión con alta demanda y buenos salarios, representa una herramienta transformadora que permite a las organizaciones tomar decisiones basadas en evidencia, optimizar procesos, descubrir nuevas oportunidades de negocio y crear valor tanto económico como social.

El mercado laboral, aunque ha evolucionado desde su explosión inicial, continúa presentando oportunidades robustas para profesionales que combinen habilidades técnicas sólidas con capacidad de comunicación, pensamiento crítico y conocimiento del dominio de negocio. La tendencia hacia profesionales versátiles que puedan moverse entre diferentes etapas del pipeline de datos refleja la maduración del campo y la necesidad de equipos que comprendan el ciclo completo de vida de los proyectos de datos.

Para quienes inician su camino en ciencia de datos, el mensaje es claro: el aprendizaje continuo, la adaptabilidad y el enfoque en resolver problemas reales son tan importantes como el dominio técnico. Python, como lenguaje de programación predominante en este ecosistema, representa el punto de partida ideal para desarrollar estas competencias y acceder a las múltiples oportunidades que ofrece este campo dinámico y en constante evolución.

4. Instalación y configuración del marco de trabajo

4.1 Introducción

Antes de comenzar a trabajar con Python para ciencia de datos, es fundamental configurar correctamente nuestro entorno de desarrollo. Un entorno bien configurado facilita la gestión de paquetes, evita conflictos entre dependencias y garantiza la reproducibilidad de nuestros análisis (Anaconda Team, 2025b). En este capítulo, exploraremos las mejores prácticas para instalar y configurar las herramientas esenciales que utilizaremos a lo largo del curso.

4.2 Distribuciones de Python para Ciencia de Datos

Existen varias opciones para trabajar con Python en ciencia de datos. A continuación, presentamos las principales alternativas, incluyendo tanto entornos locales como basados en la nube.

Opción 1: Google Colaboratory (Colab)

Google Colaboratory, o simplemente Colab, es un entorno de notebook Jupyter gratuito basado en la nube que no requiere instalación y se ejecuta completamente en el navegador (Google, 2025). Es especialmente popular en entornos educativos debido a su accesibilidad y características avanzadas.

Características Principales:

- **Cero Configuración:** Solo requiere una cuenta de Google. No es necesario instalar Python, librerías ni configurar el entorno (DataCamp, 2022).
- **Acceso Gratuito a GPUs y TPUs:** Proporciona acceso sin costo a procesadores GPU (NVIDIA Tesla K80, P4, T4) y TPU, recursos de hardware especializados que aceleran significativamente el cálculo, especialmente útiles para entrenamiento de modelos de machine learning y deep learning (Kartaca, 2025).
- **Librerías Preinstaladas:** Incluye NumPy, Pandas, Matplotlib, Seaborn, TensorFlow, PyTorch, Keras y Scikit-learn, entre otras, listas para usar sin instalación (DataCamp, 2022).
- **Integración con Google Drive:** Los notebooks se almacenan en Google Drive, permitiendo acceso desde cualquier dispositivo y facilitando el respaldo automático (Kartaca, 2025).
- **Colaboración en Tiempo Real:** Permite compartir notebooks y trabajar simultáneamente con otros usuarios, similar a Google Docs (DataCamp, 2022).
- **Data Science Agent con Gemini:** Desde marzo 2025, incluye un agente de IA que puede generar notebooks completos a partir de descripciones en lenguaje natural, automatizando tareas de importación de librerías, carga de datos y generación de código (Google Developers Blog, 2025a).

- **Modo Presentación:** Permite convertir notebooks en presentaciones interactivas donde se puede ejecutar código en vivo durante la clase (Google, 2025).
- **Control de Versiones de Runtime:** Permite congelar versiones específicas del entorno de ejecución para garantizar reproducibilidad a largo plazo (Google Developers Blog, 2025b).

Colab Pro para Educación (Gratuito):

Google ofrece suscripciones gratuitas de Colab Pro por un año a estudiantes y profesores de instituciones educativas en Estados Unidos (Google, 2025). Aunque esta oferta específica es para EE.UU., todos los usuarios tienen acceso a la versión gratuita estándar de Colab con recursos GPU/TPU limitados pero suficientes para aprendizaje y proyectos educativos.

Ventajas de Google Colab:

- **Ideal para principiantes:** Elimina barreras técnicas de instalación y configuración.
- **Accesibilidad universal:** Funciona en cualquier dispositivo con navegador (computadoras, tablets, Chromebooks).
- **Sin limitaciones de hardware local:** Permite entrenar modelos complejos sin necesidad de hardware potente.
- **Perfecto para educación:** Facilita el compartir material de clase y garantiza que todos los estudiantes tengan el mismo entorno.
- **Rápido inicio de proyectos:** Se puede empezar a codificar en menos de un minuto.

Limitaciones:

- **Requiere conexión a internet:** No se puede trabajar sin conexión.
- **Tiempo de ejecución limitado:** Las sesiones gratuitas tienen límites de tiempo (12 horas máximo) y pueden desconectarse tras inactividad.
- **Recursos compartidos:** En la versión gratuita, el acceso a GPU/TPU está sujeto a disponibilidad.
- **Menor control sobre el entorno:** Algunas configuraciones avanzadas del sistema no son accesibles.

Opción 2: Instalación Local con Anaconda

Para quienes prefieren trabajar localmente con control total sobre el entorno, Anaconda es la opción más recomendada debido a que incluye Python, un gestor de paquetes (conda), un gestor de entornos virtuales y más de 300 paquetes científicos preinstalados y probados para trabajar en conjunto (Anaconda Team, 2025a).

Comparación de Opciones

Criterio	Google Colab	Anaconda	Miniconda
Instalación	No requiere. Solo cuenta de Google.	~500 MB. Completa.	~50 MB. Mínima.
Conexión	Requiere internet.	Funciona offline.	Funciona offline.
Hardware	GPU/TPU gratuitos en la nube.	Usa recursos locales.	Usa recursos locales.
Colaboración	Nativa, en tiempo real.	Requiere herramientas adicionales (Git).	Requiere herramientas adicionales (Git).
Costo	Gratuito (con límites). Pro: $9.99/mes.	Gratuito.	Gratuito.
Mejor para	Principiantes, educación, proyectos colaborativos, ML/DL.	Trabajo local, proyectos complejos, desarrollo profesional.	Usuarios avanzados, entornos específicos.

Recomendación para este curso: Utilizaremos principalmente Google Colab por su facilidad de acceso y configuración cero, lo que garantiza que todos los estudiantes tengan el mismo entorno desde la primera clase. Sin embargo, también enseñaremos la instalación de Anaconda para quienes deseen trabajar localmente o en entornos sin conexión constante a internet.

4.3 Instalación de Anaconda

Paso 1: Descarga del Instalador

1. Visitar el sitio oficial de Anaconda en https://www.anaconda.com/download (Anaconda, 2025a).
2. El sitio detectará automáticamente tu sistema operativo (Windows, macOS o Linux).
3. Descargar el instalador correspondiente a la versión más reciente de Python 3 (actualmente Python 3.13 en Anaconda Distribution 2025.06).

4. El archivo de instalación es considerable (~500 MB), por lo que la descarga puede tomar varios minutos dependiendo de tu conexión a internet.

Paso 2: Verificación de Integridad (Opcional pero Recomendado)

Antes de ejecutar el instalador, es recomendable verificar su integridad mediante el hash SHA-256 para asegurar que el archivo no ha sido alterado o corrompido durante la descarga (Conda Development Team, 2025).

¿Dónde encontrar los códigos hash oficiales?

Los códigos hash SHA-256 oficiales se publican en dos ubicaciones (Anaconda, 2025b):

5. **Archivo de Anaconda:** `https://repo.anaconda.com/archive/` - Esta página lista todos los instaladores con sus hashes SHA-256 correspondientes. Buscar el archivo descargado y copiar el hash que aparece junto a él.

Generación del hash local en Windows (PowerShell):

```
Get-FileHash .\Anaconda3-2025.12-2-Windows-x86_64.exe -Algorithm SHA256
```

Comparar el hash generado con el hash oficial publicado en el sitio web de Anaconda. Si ambos hashes coinciden exactamente, el archivo es seguro y no ha sido alterado (Anaconda, 2025b).

Ejemplo de verificación:

Hash generado localmente: 09C8A69E7A717A963A9F2516974...

Hash oficial de Anaconda: 09c8a69e7a717a963a9f2516974...

✓ Coinciden - El archivo es seguro

Nota importante: Los hashes SHA-256 no distinguen entre mayúsculas y minúsculas. Los valores “`09C8A69E7A717A963A9F`” y “`09c8a69e7a717a963a9f`” son idénticos y ambos válidos. Lo importante es que todos los caracteres alfanuméricos coincidan, independientemente de si están en mayúsculas o minúsculas.

Paso 3: Ejecución del Instalador - para Windows:

6. Hacer doble clic en el archivo .exe descargado.
7. Hacer clic en 'Next' en la pantalla de bienvenida.

8. Leer y aceptar los términos de servicio.
9. **Seleccionar el tipo de instalación:** *'Just Me (Recommended)*' para instalar solo para el usuario actual, o 'All Users' si se tienen privilegios de administrador.
10. Seleccionar la carpeta de destino (se recomienda usar la ubicación predeterminada).
11. En las opciones avanzadas:
    - **NO** marcar 'Add Anaconda to my PATH environment variable' (esto puede causar conflictos con otras instalaciones de Python) (Anaconda, 2025b).
    - **SÍ** marcar 'Register Anaconda3 as my default Python 3.13' (recomendado).
12. Hacer clic en 'Install' y esperar a que complete el proceso (puede tomar varios minutos).
13. Hacer clic en 'Finish' para completar la instalación.

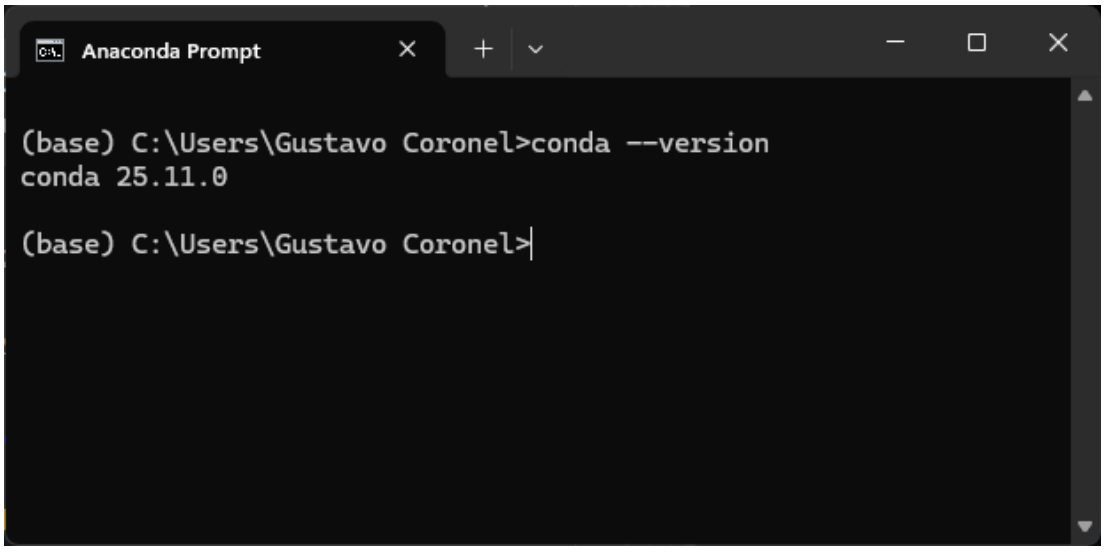
Paso 4: Verificación de la Instalación

Una vez completada la instalación, es importante verificar que todo funcione correctamente:

14. **Windows:** Buscar y abrir 'Anaconda Prompt' desde el menú Inicio.
15. Ejecutar el siguiente comando para verificar la versión de conda:

```
conda --version
```

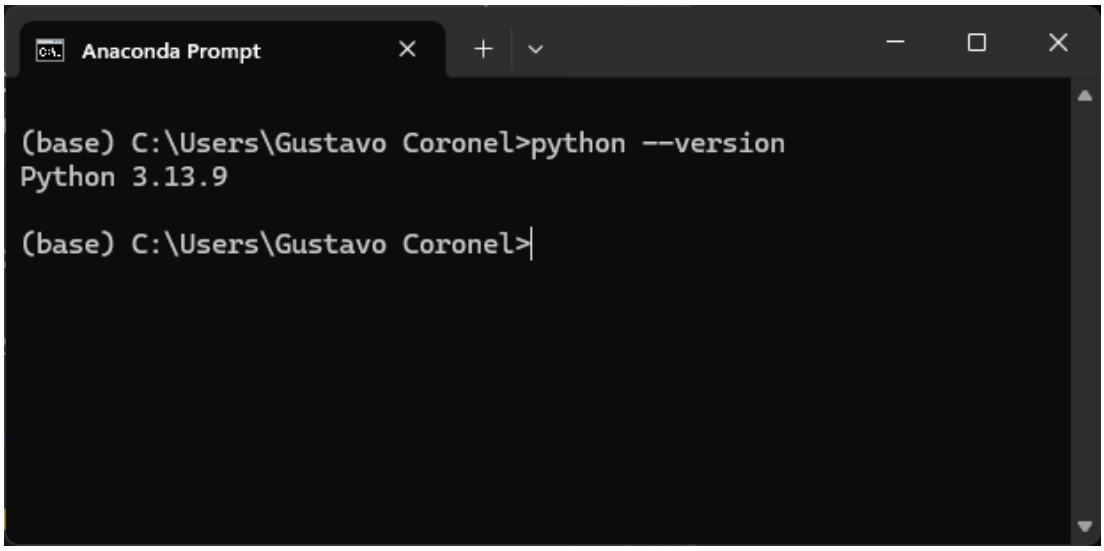
Deberías ver la versión correspondiente a tu instalación:



16. Verificar la versión de Python:

```
python --version
```

Deberías ver la versión correspondiente a tu instalación:



4.5 Gestión de Paquetes

Anaconda viene con dos gestores de paquetes principales: `conda` y `pip`. Aunque ambos pueden instalar paquetes de Python, conda es el gestor preferido en entornos Anaconda porque maneja dependencias de forma más robusta y puede instalar paquetes no escritos en Python (Anaconda Team, 2025b).

Instalación de Paquetes Individuales

Usando conda:

```
conda install numpy pandas matplotlib
```

Usando pip (cuando el paquete no está disponible en conda):

```
pip install seaborn
```

Paquetes Esenciales para Ciencia de Datos

Para trabajar eficientemente en ciencia de datos, es necesario instalar ciertos paquetes fundamentales:

Paquete	Descripción y Uso
NumPy	Librería fundamental para computación numérica. Proporciona arrays multidimensionales y funciones matemáticas de alto rendimiento.
Pandas	Herramienta principal para manipulación y análisis de datos. Ofrece estructuras de datos (DataFrame, Series) y operaciones para trabajar con datos tabulares.
Matplotlib	Librería de visualización 2D. Permite crear gráficos estáticos, animados e interactivos de alta calidad.
Seaborn	Construida sobre Matplotlib, proporciona una interfaz de alto nivel para crear visualizaciones estadísticas atractivas con menos código.
Scikit-learn	Librería de machine learning que incluye algoritmos de clasificación, regresión, clustering y preprocesamiento de datos.
Jupyter	Entorno de desarrollo interactivo basado en navegador. Permite combinar código ejecutable, visualizaciones y texto narrativo en un mismo documento.

Todos estos paquetes ya vienen preinstalados en Anaconda Distribution, lo cual es una de sus principales ventajas.

Actualización de Paquetes

Actualizar un paquete específico:

```
conda update numpy
```

Actualizar todos los paquetes:

```
conda update –all
```

Actualizar conda:

```
conda update conda
```

4.6 Jupyter Notebook

Jupyter Notebook es el entorno de desarrollo interactivo más popular para ciencia de datos. Permite combinar código ejecutable, visualizaciones ricas, ecuaciones

matemáticas y texto narrativo en un mismo documento, facilitando la exploración de datos y la comunicación de resultados (DataCamp, 2018).

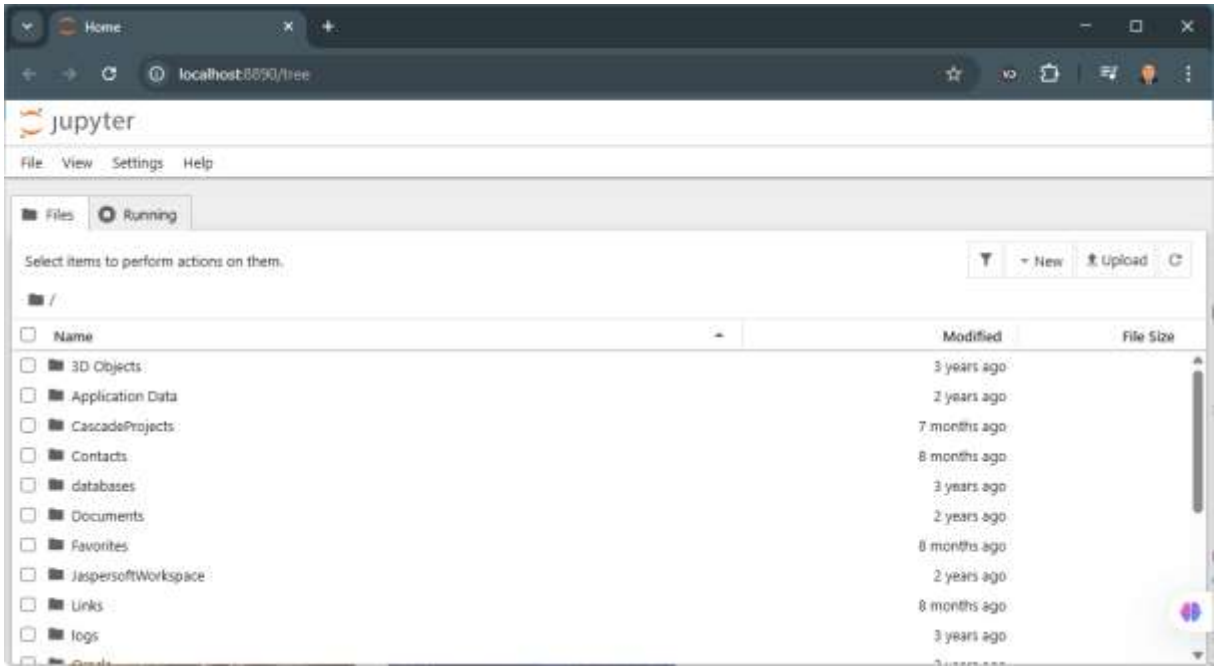
Lanzar Jupyter Notebook

Desde Anaconda Prompt o Terminal, ejecutar:

```
jupyter notebook
```

Esto abrirá automáticamente el navegador web predeterminado mostrando el Dashboard de Jupyter, que es un explorador de archivos desde donde se pueden crear, abrir y gestionar notebooks.

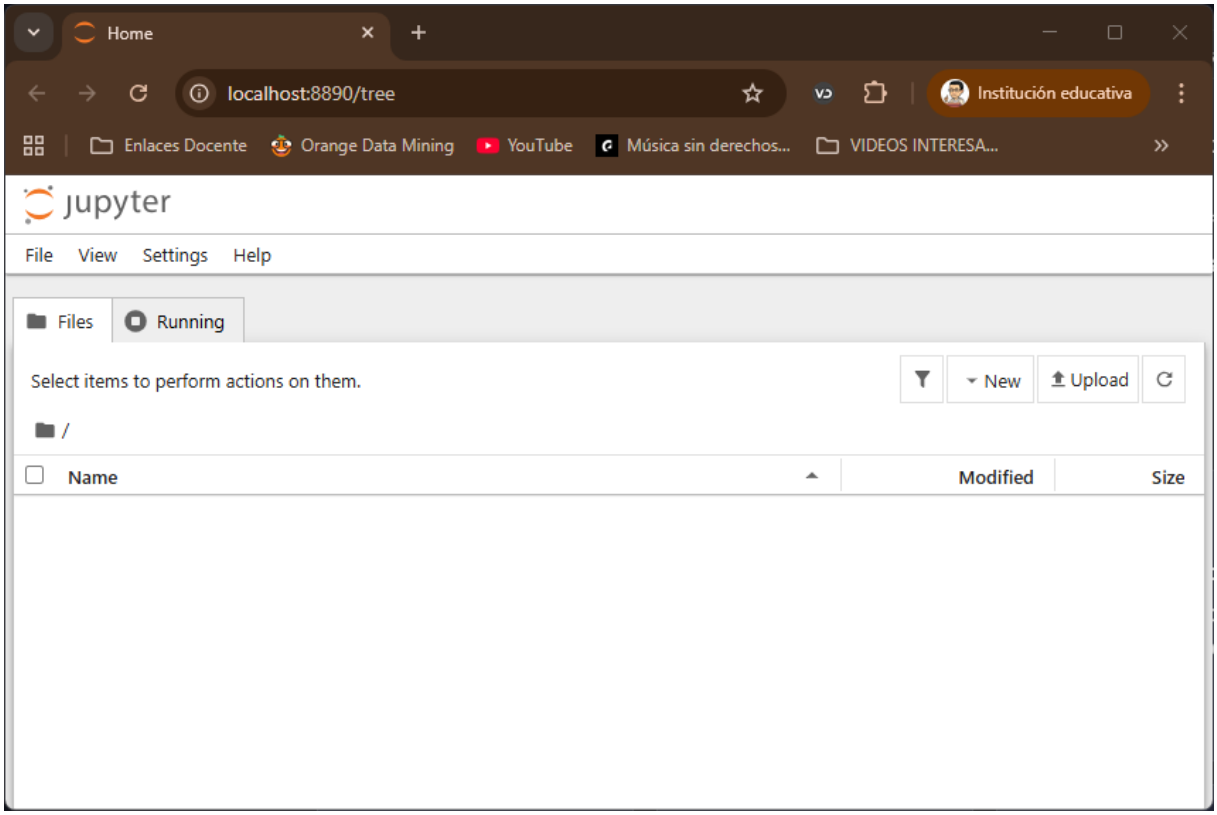
Por defecto la carpeta de trabajo es: `C:\users\[Tu Usuario]`



Si lo que quieres es cargar Jupyter NoteBook en una carpeta de trabajo diferente, por ejemplo “C:\CienciaDatos\Curso70160”, entonces debes ejecutar el siguiente comendo:

```
jupyter notebook --notebook-dir="E:\ProyectosCienciaDeDatos\Curso70160"
```

El resultado es el siguiente:



Crear un Nuevo Notebook

17. En el Dashboard, hacer clic en 'New' (esquina superior derecha).
18. Seleccionar 'Python 3' (o el nombre de tu entorno virtual si está activado).
19. Se abrirá un nuevo notebook en una pestaña del navegador.

Operaciones Básicas

- Ejecutar una celda: Presionar Shift + Enter
- Agregar celda debajo: Presionar B (en modo comando)
- Agregar celda arriba: Presionar A (en modo comando)
- Eliminar celda: Presionar DD (dos veces D en modo comando)
- Cambiar a celda de Markdown: Presionar M (en modo comando)
- Cambiar a celda de código: Presionar Y (en modo comando)

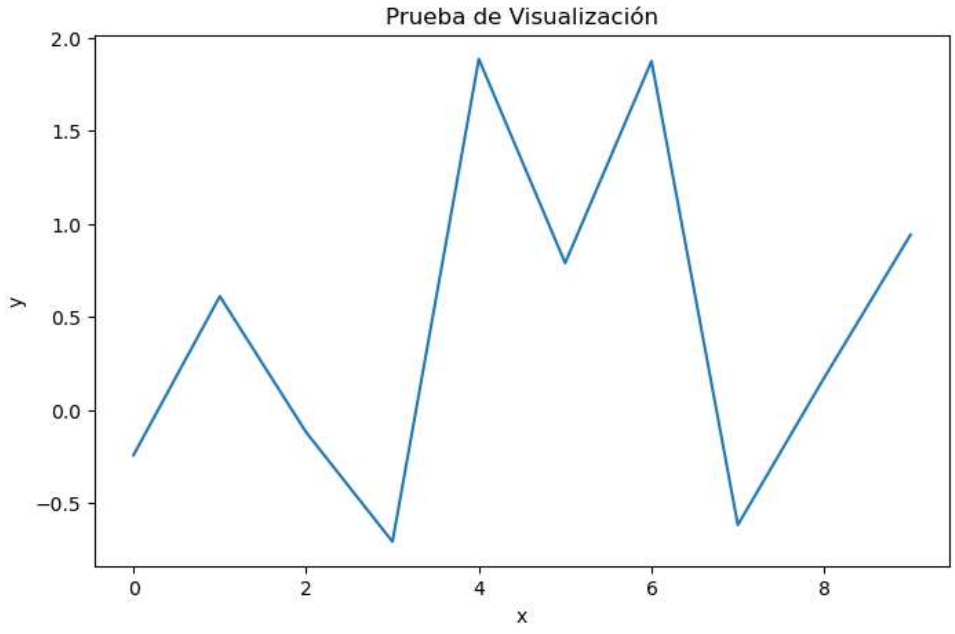
4.7 Verificación Final del Entorno

Para asegurar que todo está funcionando correctamente, en una celda de un notebook de Jupyter debes copiar y ejecutar el siguiente código de prueba:

```
# Librerías
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

#Versión de las librerías
print('NumPy version:', np.__version__)
print('Pandas version:', pd.__version__)
print('Matplotlib version:', plt.matplotlib.__version__)
print('Seaborn version:', sns.__version__)

# Crear un gráfico de prueba
data = pd.DataFrame({'x':range(10),'y':np.random.randn(10) })
plt.figure(figsize=(8, 5))
sns.lineplot(data=data, x='x', y='y')
plt.title('Prueba de Visualización')
plt.show()

print('\n¡El entorno está configurado correctamente!')
```

Debes obtener un resultado similar al que se muestra a continuación:

```
NumPy version: 2.3.5
Pandas version: 2.3.3
Matplotlib version: 3.10.6
Seaborn version: 0.13.2
```




```
¡El entorno está configurado correctamente!
```

Si el código se ejecuta sin errores y muestra un gráfico, el entorno está completamente configurado y listo para comenzar a trabajar con ciencia de datos.

4.8 Conclusiones

La correcta instalación y configuración del entorno de trabajo es un paso fundamental que determina en gran medida la eficiencia y calidad del trabajo en ciencia de datos. Anaconda Distribution proporciona una solución integral que simplifica este proceso, permitiendo a los científicos de datos concentrarse en el análisis y modelado en lugar de resolver problemas de dependencias y configuración.

Con el entorno correctamente configurado, estamos listos para comenzar nuestro viaje en el análisis de datos con Python, explorando sus poderosas librerías y aplicándolas a problemas reales del mundo de los datos.

6. Cierre del módulo

En este módulo se estableció una base conceptual y práctica para iniciar el trabajo en Ciencia de Datos. Primero, se revisó qué es la disciplina, qué problemas aborda y cuáles son sus principales componentes, conectando el enfoque técnico con la necesidad de comprender el dominio y comunicar resultados. Luego, se configuró el entorno de trabajo con Python (Anaconda/Jupyter) para asegurar una ejecución ordenada y reproducible. Con el entorno listo, se realizaron los primeros pasos del análisis con Python: carga de datos, revisión inicial, exploración descriptiva y visualizaciones básicas. Con estas competencias, el estudiante queda preparado para avanzar hacia análisis más estructurados y, posteriormente, hacia técnicas de modelado en los siguientes módulos.

Recursos para Continuar Aprendiendo

Documentación Oficial y Tutoriales:

- **Python Tutorial Oficial:** `https://docs.python.org/3.13/tutorial/` - La guía oficial de Python, completa y siempre actualizada.
- **Real Python:** https://realpython.com - Tutoriales de alta calidad sobre Python y ciencia de datos.
- **Kaggle Learn:** https://www.kaggle.com/learn - Cursos cortos y prácticos sobre Python, Pandas, y Machine Learning.

Libros Recomendados:

- VanderPlas, J. (2016). Python Data Science Handbook. O'Reilly Media. Disponible gratuitamente en https://jakevdp.github.io/PythonDataScienceHandbook/
- McKinney, W. (2022). Python for Data Analysis. O'Reilly Media. Escrito por el creador de Pandas.

Comunidades y Foros:

- **Stack Overflow:** https://stackoverflow.com - La comunidad más grande para resolver dudas de programación.
- **Reddit r/datascience:** https://reddit.com/r/datascience - Discusiones, recursos y consejos sobre ciencia de datos.

- **Python Discord:** https://pythondiscord.com - Comunidad activa para aprender Python.

Mensaje Final

Has completado el primer módulo de un taller que te transformará en un profesional capaz de extraer valor de los datos. El camino que tienes por delante es desafiante pero extraordinariamente gratificante. Cada línea de código que escribas, cada error que corrijas, cada insight que descubras, te acercará más a la maestría.

7. Referencias

Anaconda. (2025a). *Download Anaconda Distribution*. https://www.anaconda.com/download

Anaconda. (2025b). *Installing Anaconda Distribution*. https://www.anaconda.com/docs/getting-started/anaconda/install

Anaconda Team. (2025a). *New Release: Anaconda Distribution 2025.06*. Anaconda. https://www.anaconda.com/blog/new-release-anaconda-distribution-2025-06

Anaconda Team. (2025b). *Python Packages: Installation & Management Best Practices*. Anaconda. https://www.anaconda.com/guides/python-packages

Conda Development Team. (n.d.). *Installing conda*. Conda. https://docs.conda.io/projects/conda/en/latest/user-guide/install/index.html

DataCamp. (2018). *Setup a Data Science Environment on your Computer*. Tutorials. https://www.datacamp.com/tutorial/setup-data-science-environment

DataCamp. (2022). *Google Colab Tutorial for Data Scientists*. Tutorials. https://www.datacamp.com/tutorial/tutorial-google-colab-for-data-scientists

Davenport, T. H., & Patil, D. (2012). Data scientist: The sexiest job of the 21st century. *Harvard Business Review, 90*(10), 70-76. https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

Gartner. (2025). *Gartner Identifies Top Trends in Data and Analytics for 2025*. Gartner, Inc. https://www.gartner.com/en/newsroom/press-releases/2025-03-05-gartner-identifies-top-trends-in-data-and-analytics-for-2025

Glassdoor. (2025). *Data scientist salaries*. Glassdoor. https://www.glassdoor.com/Salaries/data-scientist-salary-SRCH_KO0,14.htm

Google. (2025). *New Google Colab features for higher education*. Google for Education Blog. https://blog.google/products-and-platforms/products/education/colab-higher-education/

Google Developers Blog. (2025a). *Data Science Agent in Colab: The future of data analysis with Gemini*. Developers. https://developers.googleblog.com/en/data-science-agent-in-colab-with-gemini/

Google Developers Blog. (2025b). *Google Colab Adds More Back to School Improvements!* Developers. https://developers.googleblog.com/en/google-colab-adds-more-back-to-school-improvements/

INEI. (2025). *Perú: Comportamiento de los Indicadores del Mercado Laboral a nivel nacional y 27 ciudades - Cuarto trimestre 2024.* https://www.gob.pe/institucion/inei/informes-publicaciones/6474256-peru-comportamiento-de-los-indicadores-del-mercado-laboral-a-nivel-nacional-y-de-27-ciudades-cuarto-trimestre-2024

Kartaca. (2025). *Google Colab: A Deep Dive into Its Features and Applications*. Kartaca Blog. https://kartaca.com/en/google-colab-a-deep-dive-into-its-features-and-applications/

Magnet, S. (2025). *Data Scientist Job Outlook 2025 [Research on 1,000 Job Postings]*. 365 DataScience. https://365datascience.com/career-advice/career-guides/data-scientist-job-outlook-2025/

McKinsey Global Institute. (2022). The data-driven enterprise of 2025. *McKinsey & Company*. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-data-driven-enterprise-of-2025

Naur, P. (1974). *Concise survey of computer methods.* Studentlitteratur. https://archive.org/details/concisesurveyofc0000naur/page/n1/mode/2up

Remote Rocketship. (2025). *Remote data science jobs in Latin America*. Remote Rocketship. https://www.remoterocketship.com/country/latin-america/jobs/data-scientist/?page=1&sort=DateAdded&jobTitle=Data+Scientist&locations=Latin+America

SAS Institute. (2025). *Data Science: What it is and why it matters*. SAS Institute. https://www.sas.com/en_us/insights/analytics/data-science.html

Statista. (2025). *Data science platform market size worldwide 2023-2032*. Statista. https://www.statista.com/statistics/1356509/data-science-platform-market-size-worldwide/

Tukey, J. W. (1962). The future of data analysis. *The Annals of Mathematical Statistics, 33*(1), 1-67. https://doi.org/https://doi.org/10.1214/aoms/1177704711

U.S. Bureau of Labor Statistics. (2025). *Data scientists*. Occupational Outlook Handbook. https://www.bls.gov/ooh/math/data-scientists.htm

Universidad Nacional de Ingeniería. (2024). *Maestría en Inteligencia Artificial*. Programa Maestrías de Especialización. https://www.fiis.uni.edu.pe/posgrado/programa-maestrias-especializacion/maestria-en-inteligencia-artificial