

Diploma

ADVANCED DATA SCIENTIST



Presentación

¿Sabías que desarrollar proyectos de ciencia de datos van mucho más allá que entrenar y evaluar modelos, y que se pueden combinar con otras disciplinas como la ingeniería de datos y el despliegue de soluciones?

El Diploma Advanced Data Science te capacitará en las herramientas y habilidades necesarias para complementar tus proyectos de machine learning, tanto al inicio del pipeline con técnicas de scraping para automatizar la captura de datos, como al final de este para el despliegue o puesta en producción de modelos.



Sobre este Diploma

37

sesiones

148

horas
académicas

65

talleres
prácticos

01

proyecto para
tu portafolio

¿Cómo impulsamos tu carrera?

- Sesiones 80% **enfocadas en la práctica.**
- Enfoque en **Casos Reales** enfrentando los retos del mercado.
- Énfasis en **habilidades técnicas y blandas.**
- **Mentoría especializada** con docentes praticioners.
- Acompañamiento **constante.**



¿Porqué estudiar este diploma?

Conviértete en un experto en las técnicas de ciencia de datos y su aplicación directa en los negocios.



Prepárate con una capacitación integral a todo el ciclo de vida de un proyecto de ciencia de datos.



Desarrolla un proyecto integrador con el acompañamiento y asesoramiento de un experto a partir de una casuística de negocio.



Objetivo del diploma

- Aplica estadística inferencial para selección de variables o muestras con base matemática, optimizando la predicción de modelos.
- Implementa scraping y captura automatizada desde webs, archivos en línea y APIs; además, construye modelos supervisados, realiza tuning y evalúa clasificación multiclase.
- Desarrolla soluciones de segmentación, RFM, asociación y forecasting con modelos de clusterización y redes neuronales como LSTM.

Objetivo Final

Diseña e implementa una solución integral de ciencia de datos, que incluye un modelo de machine learning y un módulo de scraping automatizado para la captura de datos, orientada a resolver necesidades de negocio como clasificación, segmentación, análisis RFM o asociación de productos.

¿A quién está dirigido?

1. Científicos de datos Jr.

Personas que desempeñen este rol, y busquen:

- Complementar sus habilidades en ciencia de datos con técnicas asociadas al scraping de datos y despliegue de modelos para producción.
- Aplicar las redes neuronales y el aprendizaje profundo a casuísticas específicas de negocio.



¿Cuáles son los requisitos?



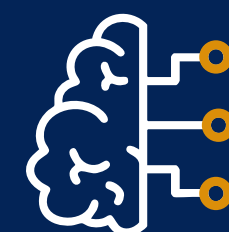
Conocimientos / Habilidades

- Dominio del lenguaje Python.
- Conocimiento de técnicas y algoritmos típicos de machine learning.



Experiencia Laboral

- Mínimo un año de experiencia trabajando con modelamiento de datos.



Tecnológicos

- Contar con una laptop o computadora de escritorio con disponibilidad de micrófono y cámara web.
- Tener instalado los softwares y herramientas señalados en la sección Contenidos.

Perfil del egresado

El egresado del Diploma Advanced Data Scientist estará en la capacidad de:

- Entrenar y evaluar modelos de clasificación multiclase, empleando tanto algoritmos tradicionales como ensembles.
- Entrenar y evaluar modelos clústering de aplicación directa a casuísticas negocios como agrupamiento de clientes, análisis RFM y asociación.
- Aplicar las redes neuronales y el aprendizaje profundo a casos de uso específicos de negocio.
- Poner en producción sus soluciones de machine learning tanto mediante el diseño de servidores propios como en plataformas cloud (AWS)..

Campo Laboral

Podrás laboral en puestos relacionados a:

- Científico de datos.
- Analista de ciencia de datos.
- Machine Learning Developer.

Herramientas



Python



Google Colab



Visual Studio Code



Git



AWS



DVC



Docker



Prometheus



Malla Curricular

I. Taller de Marca Personal

- Actividad de sociabilización y contacto.
- Marca personal ¿Qué es y cómo desarrollarla?
- Empleabilidad y ser empleable. Diferencias clave.
- ¿Cómo hacer más atractivo el curriculum?
- ¿Cómo afrontar una entrevista de trabajo?

II. Estadística inferencial para ciencia de datos

1. Repaso de Estadística descriptivo

- Medidas de tendencia central
- **Taller:** Ejecución de un notebook sobre EDA de un dataset.
- **Taller:** Ejecución de un notebook sobre identificación y tratamiento de outliers y missing values.

2. Distribución de probabilidad

- Técnicas de muestreo.
- Distribuciones de probabilidad, sesgos, Teorema del límite central.
- **Taller:** Análisis visual y matemático del sesgo de variables cuantitativas de un dataset.
- **Taller:** Visualización de medias para diferentes muestras de una misma variable.
- **Taller:** Análisis y discusión de diferentes muestras (medias y sesgos) y sus impactos en la capacidad de predicción.

3. Prueba de hipótesis

- Formulación de hipótesis sobre diferencias entre clases en un dataset.
- **Taller:** Selección de variable a partir de análisis t-test o prueba Z.

4. Pruebas no paramétricas

- Aplicación de pruebas Mann-Whitney y Wilcoxon a variables sesgadas.
- **Taller:** Análisis de robustes de un dataset y su impacto en el modelo.

5. Análisis de variables categóricas (ANOVA)

- El análisis de varianza.
- **Taller:** Análisis e impacto del encoding categórico en la capacidad de predicción.

III. Web Scraping & API's

6. Introducción a Web Scraping

- Página web vs. Sitio Web vs. Aplicativo Web.
- Arquitectura HTML.
- Web Crawling vs. Web Scraping.
- El User-Agent.
- Scraping estático vs Scraping dinámico.

7. Implementación de Web Scraping

- **Taller:** Implementación de una solución para scrapear un catálogo web simple, con Python.
- **Taller:** Implementación de una solución para scrapear un catálogo web multipágina, con Python.
- **Taller:** Implementación de una solución para scrapear un archivo alojado en una fuente web pública, con Python.
- **Taller:** Implementación de una solución para scrapear una fuente API pública, con Python.

8. Serialización de dataset

- Formatos de dataset para ciencia de datos: csv, parquet, avro. ORC.
- **Taller:** Serialización de datos provenientes de un scraper.

IV. Aprendizaje supervisado multiclase

9. Técnicas de clasificación multiclase

- Árboles de decisión. Algoritmos para su estimación CART, C45.
- **Taller:** Entrenamiento y análisis de árboles de clasificación multiclase basados en CART y C45 con Python.
- Evaluación de modelos de clasificación multiclase. La matriz de confusión multiclase.
- **Taller:** Implementación y análisis de matrices de confusión para modelos multiclase.

10. Tuning de técnicas Ensembles

- Repaso técnicas Ensembles. Definición, algoritmos homogéneos y heterogéneos, descripción de Bagging y Boosting.
- Tipos de regularización L1 y L2.
- Exploración y descripción de hiperparámetros de un modelo RandomForest.
- **Taller:** Tuning de un RandomForest para clasificación multiclase con Python.
- Exploración y descripción de hiperparámetros de un modelo AdaBoost.
- **Taller:** Tuning de un AdaBoost para clasificación multiclase con Python.
- Exploración y descripción de hiperparámetros de un modelo GBM.
- **Taller:** Tuning de un GBM para clasificación multiclase con Python.
- Exploración y descripción de hiperparámetros de un modelo XGBoost.
- **Taller:** Tuning de un XGBoost para clasificación multiclase con Python.

11. Evaluación avanzada de modelos

- **Taller:** Búsqueda de mejores hiperparámetros con GridSearchCV.
- **Taller:** Bias-Variance Plot.
- **Taller:** Learning Rate Plot.

Malla Curricular

V. Aprendizaje No supervisado aplicado

12. Repaso técnicas de clusterización

- Clústering de datos. Definición, objetivo, técnicas y algoritmos para su implementación.
- **Taller:** Ejecución de un notebook para clusterización con K-Means y DBScan.

13. Clústering aplicado

- **Taller:** Clusterización de un dataset de clientes y caracterización de grupos.
- **Taller:** Clusterización de un dataset de productos y caracterización de grupos.
- **Taller:** Clusterización de un dataset de sales-transaction y caracterización de grupos.

14. Análisis RFM

- **Taller:** Clusterización de un dataset de cliente basado en la técnica RFM.

15. Técnicas de asociación

- Association Rule Mining. Definición, casos de uso, algoritmos disponibles para su implementación.
- Conceptos asociados: Support, Confidence, Lift.
- **Taller:** Obtención y análisis de Frequent-Itemset a partir de un dataset. Configuración de diferentes umbrales (support) y número de items (Python).
- **Taller:** Obtención de reglas de asociación (Python).

VI. Fundamentos de aprendizaje profundo

16. Fundamentos de Aprendizaje Profundo

- Deep Learning vs. Machine Learning
- Deep Learning vs. generative AI.
- Arquitectura de una neurona artificial
- Función de activación. Definición y tipos.
- **Taller:** Diseño y proceso de cálculo de una neurona que simula el comportamiento de una compuerta lógica. (En Excel).

17. Proceso de aprendizaje de una red neuronal

- De la neurona artificial a la red neuronal.
- Arquitectura de una red neuronal.
- Algoritmos Feedforward y Backpropagation.
- Gradiente del descenso.
- Función de coste.
- **Taller:** Diseño y proceso de cálculo de una red neuronal que hace clasificación multiclase. (En Excel).

18. Training de redes neuronales básicas

- TensorFlow vs. Keras. Casos de uso independientes, y por separado.

- **Taller:** Implementación y evaluación de una red neuronal básica para clasificación binaria. Uso y configuración de layers.
- **Taller:** Implementación y evaluación de una red neuronal para clasificación multiclase. Configuración de la capa de salida.

VII. Pronóstico de series de tiempo (forecasting)

19. Fundamentos del análisis de series de tiempo

- Tipos de series temporales y sus aplicaciones.
- Estructura: tendencia, estacionalidad, ciclo e irregularidad.
- Estacionariedad y transformaciones básicas.
- **Taller:** Preprocesamiento de dataset de serie temporal para forecasting.

20. Preparación de datos para modelos neuronales

- Transformaciones: escalado, diferenciación, generación de lags.
- Proceso de construcción de ventanas de entrada y salida.
- **Taller:** Construcción de un dataset supervisado para serie temporal.

21. Aprendizaje profundo para Forecasting

- Arquitecturas de redes neuronales: MLP, RNN, LSTM, GRU, TCN, Transformer.
- **Taller:** Entrenamiento de una modelo LSTM para pronóstico de una serie univariada.

22. Forecasting multivariado y modelos avanzados

- Modelos con múltiples entradas y salidas. Casos de uso.
- Variables exógenas y calendáricas.
- **Taller:** Forecasting multivariado basado en Transformers.

VIII. Clasificación de textos y lenguaje natural (sentiment analysis)

23. Fundamentos de procesamiento de lenguaje natural

- Introducción a NLP. Aplicaciones y desafíos.
- Flujo típico de un proyecto de análisis de texto.
- Técnicas de preprocesamiento de textos: limpieza, tokenización, lematización, expresiones regulares y stop-words.
- **Taller:** Preprocesamiento de un dataset de textos.

Malla Curricular

24. Exploración y análisis léxico de texto

- **Taller:** Representación y análisis de frecuencia de palabras en nubes de palabras.
- **Taller:** Análisis de textos basados en n-gramas.

25. Representación vectorial de texto

- Introducción a bolsas de palabras y TF-IDF.
- Introducción a representaciones densas: Word2Vec, GloVe y embeddings.
- **Taller:** Vectorización de texto y preparación de características para clasificación de sentimientos.

26. Introducción a los transformers

- De las RNN a los Transformers. Arquitecturas y principales diferencias.
- **Taller:** Repaso de un Notebook sobre un transformer básico para traducción de texto.

27. Aprendizaje profundo para clasificación de textos y análisis de sentimiento

- Modelos modernos con redes neuronales y Transformers (BERT).
- **Taller:** Construcción de un modelo de análisis de sentimiento empleando BERT.
- **Taller:** Comparación del clasificador con técnicas tradicionales (regresión logística, RandomForest, SVM).

IX. Despliegue de modelos y MLOps

28. Fundamentos para servir modelos con Python

- Diferencia entre experimentación y producción.
- Formatos de serialización de modelos. Pickle, Joblib, ONNX.
- **Taller:** Serialización y carga de un modelo desde un script Python.

29. Desarrollo de APIs REST para predicción uno-a-uno

- Web serving como concepto. Introducción a Flask y FastAPI.
- API como concepto.
- **Taller:** Exposición de un modelo de clasificación mediante FastAPI. (Input vía JSON)

30. Desarrollo de interfaz web para predicción por lotes

- Estructura de una aplicación web ligera con FastAPI y HTML.
- **Taller:** Implementación de una interfaz web que recibe un Excel, aplica un modelo, y devuelve el archivo con una columna de predicción.

31. Fundamentos de MLOps

- Definición y ciclo de vida de MLOps.
- Introducción teórica a buenas prácticas de MLOps: versionado, logging, pruebas.
- Requerimientos organizacionales y técnicos para implementar MLOps.

X. Del desarrollo de modelos a MLE

32. Introducción a MLE y repaso de conceptos útiles

- Rol del Científico de Datos vs ML Engineer.
- Ciclo de vida de un proyecto de Data Science. Experimentación vs Producción.
- Notebook de desarrollo vs Código de producción. Buenas prácticas para su integración.
- **Taller:** Análisis y diagnóstico de un notebook de desarrollo.

33. Arquitectura de un proyecto de MLE

- Organización del código de un proyecto de MLE: src/, config/, test/ y main.py.
- Separación de responsabilidades: data, features, models, predictions.
- **Taller:** Introducción a Visual Studio Code y organización de nuevo proyecto de MLE.
- **Taller:** Paso de un notebook de desarrollo a módulos py.
 - Módulo data_loader.py
 - Módulo preprocessing.py
 - Módulo config.py

34. Pipelines y reproducibilidad

- Pipelines de Machine Learning. ¿Por qué en MLE? ¿con qué herramientas?
- ¿Qué es la reproducibilidad de un modelo?
- Arquitectura de un pipeline típico de machine learning.
- **Taller:** Implementación de un pipeline reproducible desde la carga hasta la evaluación.
 - Módulo feature_engineering.py
 - Módulo train.py
 - Módulo evaluate.py

35. Testing de funcionalidad

- ¿Qué es Testing en el contexto de MLE?
- El framework pytest y ventajas de su uso.
- **Taller:** Implementación de testing simples para las etapas de preprocesamiento y predicción.
 - Módulo test_processing.py
 - Módulo test_train.py
 - Módulo dvc.yaml

Malla Curricular

36. Fundamentos de Git Básico aplicado a proyectos de ML

- ¿Por qué usar Git en ciencia de datos?
- El flujo clone-branch-commit-push-PR.
- Pull-request y el trabajo colaborativo.
- **Taller:** Gestión de repos y ramas, commit y push.

37. Fundamentos de Data-Version-Control (DVC)

- ¿Por qué no versionar datasets directamente con Git?
- **Taller:** Uso de comandos DVC init, add, remote, push.
- **Taller:** Integración con Git para seguimiento conjunto.

38. Manejo de versiones en un proyecto MLE

- **Taller:** Trackear un dataset (CSV), re-entrenar y guardar una nueva versión del modelo.

XI. Despliegue y mantenimiento automatizado de modelos

39. Introducción a API REST

- ¿Qué es y cómo funciona una API web?
- Solicitud y recepción de datos desde una API web. Los métodos GET y POST.
- **Taller:** Probando API desde el navegador y desde Postman.

40. Despliegue de modelos en un mini servidor web

- **Taller:** Implementación de un mini servidor web con FastAPI y testeado de respuesta.
- **Taller:** Alojar el modelo en el servidor y configurarlo para atender solicitudes individuales mediante JSON.
- **Taller:** Alojar el modelo en el servidor y configurarlo para atender solicitudes batch (masivas) mediante CSV.

41. Manejo de Containers (Docker)

- ¿Qué es un Container y cómo se emplea en el contexto de MLE?
- **Taller:** Configuración de un container y construcción de imágenes.
- **Taller:** Empaquetado de un proyecto MLE en un container.

42. Despliegue de modelos contenerizados en Cloud (AWS)

- **Taller:** Creación y configuración de los servicios necesarios en AWS.
- **Taller:** Alojamiento del container con el modelo en AWS y prueba de funcionalidad.

43. Monitoreo y mantenimiento del modelo

- Conceptos clave: Data-drift y concept-drift.
- Herramientas para monitoreo de inferencias.
- Estrategias de reentrenamiento y rollback.
- **Taller:** Simulación de monitoreo y detección de degradación de un modelo (Prometheus).

XII. Proyecto integrador

- Sesión 1. Lineamientos generales.
- Sesiones 2, 3. Revisión de avance y feedback.
- Sesión 4. Presentación final y sustentación.

Certificación DMC

Por aprobación del Diploma Advanced Data Scientist, por un total de 148 horas académicas.



Nuestra Propuesta de Capacitación

Las metodologías que aplicamos



Desarrollo de competencias clave en el mundo de los datos

Analiza • Innova • Transforma



Aprendizaje Secuencial

- Descubre conocimiento de vanguardia
- Explora con la guía del experto
- Aplica lo aprendido



Aprendizaje basado en práctica (Learning by Doing)

- Resuelve retos
- Aprende en base a proyectos
- Analiza casos



Docentes Expertos

Aprende con los líderes de las mejores empresas de Latam.



Claudia Aguilar

Managing Director
en **OREKAN**



Cesar Quezada

Científico de Datos
en **KONECTA**



Edwin Aldana

Scientific Researcher
en **CONACYT**



Jimmy Salazar

Lead Data Scientist
en **RIMAC SEGUROS**



Luis Cajachahua

Artificial Intelligence
Strategy Manager
en **CREDICORP**



Luis Garayar

Chapter Leader Data
Scientist Senior
en **BCP PERÚ**



Angel Tintaya

Senior Data Engineer
en **TRANZACT**



Nilton Yanac

Gerente de Desarrollo
de Negocios
en **BUSINESS ANALYTICS S.A.C.**

*En caso de contingencias podría cambiar alguno de los docentes por otro profesional de similar perfil.

¿Por qué elegirnos?

+15

Más de 15 años de experiencia.

+300

Más de 300 empresas asesoradas en Perú, Ecuador y Bolivia.

35k

35 mil profesionales capacitados en más de 20 países de América Latina.



Propuesta integral en formación en Data & AI.

+150

Más de 150 docentes expertos de Latinoamérica, España y Estados Unidos.



Comunidad más grande en Data & AI con beneficios exclusivos: Networking, empleabilidad, habilidades blandas.



Excelente nivel de servicio.



Nuestros Partners

CertiProf® | Partner

Google Partners



Estas empresas confían en nosotros



BBVA



ANTAMINA



PROM PERÚ



SCOTIABANK



PACÍFICO
SEGUROS



SUNAT



CAJA
HUANCAYO



BUENAVENTURA



PRONABEC



CAJA
AREQUIPA



RIMAC



BCRP



MIBANCO



MAPFRE



ONCOSALUD



LOS ANDES

Métodos de pago

J&J DATA MINING CONSULTING S.A.C.
RUC: 20520972740

1. Depósito en cuenta BCP

- Corriente soles BCP: **193-225-1181-0-01**
- CCI BCP: **00219300225118100116**
- Corriente BCP dólares: **193-2318515-1-52**
- CCI BCP dólares: **002-193-002318515152-11**

2. Depósito en cuenta BBVA

- Ahorros BBVA soles: **0011-0177-02-00180473**
- CCI BBVA: **011-177-000200180473-37**

2. Pago Online

Generamos un link de pago online donde se acepta todas las tarjetas.

3. Pago con Yape

A nombre de J J Data
Mining Consulting Sac



4. Pago online por PayPal



06 CUOTAS SIN INTERESES pagando con:





Visita nuestra web

www.dmc.pe

Síguenos en:

