PYTHON DATA PRO

Análisis, Visualización y Machine Learning

Propósito: Desarrollar competencias prácticas y empleables en Ciencia de Datos: importación y limpieza de datos, análisis exploratorio, visualización de impacto, estadística aplicada y machine learning con buenas prácticas industriales (pipelines, evaluación, tuning y despliegue).

Estructura general

- **Duración total:** 36 horas sincrónicas + **Módulo 0** (2 horas online asincrónicas).
- **Modalidad sugerida:** Teórico-práctica (demos + laboratorio guiado + retos).
- **Producto final:** Proyecto capstone (limpieza + EDA + modelado + evaluación + despliegue básico).

MÓDULO 0: ORIENTACIÓN PREVIA (2 horas online asincrónicas)

- **Preparación del entorno:** instalación guiada de Anaconda, configuración de VS Code / JupyterLab.
- **Primeros pasos:** introducción a GitHub, creación de repositorio personal del curso.
- **Kit de supervivencia:** cheatsheet de Python básico, dataset de práctica, tutorial de Markdown.

MÓDULO 1: FUNDACIONES SÓLIDAS — EL ECOSISTEMA PYTHON (9 horas)

Sesión 1: Lanzamiento Explosivo — Python para la Revolución de Datos

Objetivo: instalar, ejecutar y dominar el núcleo Python + NumPy + primer contacto con Pandas en un flujo real.

Bloque	Temas principales	Actividad	Evidencia
3h (1–3)	▪ Por qué Python domina el Data Science (contexto actual). ▪ Configuración profesional: entornos virtuales con conda/venv. ▪ Núcleo para analistas: variables, estructuras de datos, lambdas. ▪ NumPy: arrays vs listas, vectorización, broadcasting. ▪ Pandas desde el minuto 1: Series y DataFrames.	▪ Crear un DataFrame desde cero con datos ficticios de ventas. ▪ Aplicar operaciones vectorizadas para métricas de negocio. ▪ Exportar a CSV y Excel automáticamente.	Notebook con **10 operaciones** sobre dataset creado + exportación exitosa.

Sesión 2: Data Wrangling Ninja — Transformación Profesional

Objetivo: limpiar, transformar y resumir datos con criterios de calidad listos para análisis y dashboards.

Bloque	Temas principales	Actividad	Evidencia
3h (4–6)	▪ Indexación avanzada: `.loc[]`, `.iloc[]`, `.query()`. ▪ Filtrado inteligente: máscaras booleanas y condiciones complejas. ▪ Transformación masiva: `.apply()`, `.map()`, `.transform()`. ▪ Agregaciones: `groupby()` con múltiples funciones. ▪ Encadenamiento profesional (method chaining).	▪ Dataset real de e-commerce. ▪ Limpieza: duplicados, outliers, formatos, nulos. ▪ KPI dashboard con `groupby`.	Script que transforma dataset “sucio” en dataset **listo para análisis** + KPIs.

Sesión 3: Conexiones Élite — Fuentes Multiplataforma

Objetivo: construir un pipeline de ingesta desde archivos, APIs, SQL y web.

Bloque	Temas principales	Actividad	Evidencia
3h (7–9)	▪ Formatos: Parquet, Feather, HDF5 (rendimiento vs CSV). ▪ APIs: `requests` + JSON para datos en vivo. ▪ SQL profesional: SQLAlchemy (PostgreSQL/MySQL) y buenas prácticas. ▪ Web scraping básico: BeautifulSoup para extraer tablas. ▪ Merges complejos: concatenación vertical/horizontal; joins.	▪ Extraer datos de una API financiera en tiempo real. ▪ Almacenar en base de datos SQL local. ▪ Cruzar con dataset interno usando joins avanzados.	Pipeline completo de ingesta: API $\rightarrow$ SQL $\rightarrow$ join con dataset interno.

MÓDULO 2: VISUALIZACIÓN DE IMPACTO Y EDA (6 horas)

Sesión 4: Arte con Datos — Visualización que Convence

Objetivo: dominar visualización profesional y storytelling para comunicar hallazgos.

Bloque	Temas principales	Actividad	Evidencia
3h (10–12)	▪ Matplotlib a fondo: Figure/Axes, subplots profesionales. ▪ Seaborn avanzado: pairplots, jointplots, heatmaps. ▪ Plotly Express: gráficos interactivos para dashboards. ▪ Personalización: estilos, paletas, anotaciones, leyendas. ▪ Storytelling: narrativa visual y secuencia de gráficos.	▪ Recrear visualizaciones estilo The Economist/Bloomberg. ▪ Dashboard unificado con 4 gráficos interrelacionados. ▪ Exportar a HTML interactivo.	Portafolio visual con **5 gráficos** y 1 dashboard HTML.

Sesión 5: Detective de Datos — EDA Automatizado

Objetivo: acelerar el análisis exploratorio con automatización, anomalías y geodatos.

Bloque	Temas principales	Actividad	Evidencia
3h (13–15)	▪ EDA automático: ydata-profiling (antes pandas-profiling). ▪ Detección: outliers con Isolation Forest, DBSCAN. ▪ Geodatos: geopandas + folium (mapas interactivos). ▪ Series temporales con pandas. ▪ Correlaciones avanzadas y clustering jerárquico.	▪ Dataset COVID-19 o cambio climático. ▪ Profiling completo + detección de estacionalidad/anomalías. ▪ Identificación de hallazgos accionables.	Reporte EDA automatizado + insights principales (negocio/política pública).

MÓDULO 3: ESTADÍSTICA APLICADA Y MODELADO (9 horas)

Sesión 6: Estadística Inferencial — Decisiones con Certeza

Objetivo: inferencia moderna con intervalos, pruebas, poder y simulación.

Bloque	Temas principales	Actividad	Evidencia
3h (16–18)	▪ Distribuciones: Normal, t-Student, binomial, Poisson. ▪ Intervalos de confianza: Bootstrap para cualquier estadístico. ▪ Pruebas de hipótesis: t-test (1 y 2 muestras), proporciones. ▪ p-value, poder estadístico y lectura correcta. ▪ Monte Carlo: resolver problemas con aleatoriedad.	▪ Simulación completa de A/B testing para marketing. ▪ Tamaño muestral mínimo y significancia real.	Reporte estadístico con conclusiones ejecutivas (A/B test completo).

Sesión 7: ANOVA y Regresión — Relaciones Complejas

Objetivo: modelar y diagnosticar ANOVA y regresión con buenas prácticas.

Bloque	Temas principales	Actividad	Evidencia
3h (19–21)	▪ ANOVA multifactorial: efectos e interacciones. ▪ Post-hoc: Tukey, Bonferroni, Games-Howell. ▪ Regresión múltiple: statsmodels vs scikit-learn. ▪ Diagnósticos: VIF, heterocedasticidad, autocorrelación. ▪ Selección de variables: Stepwise, Lasso, criterios AIC/BIC.	▪ Dataset de precios de propiedades. ▪ Modelar precio con 10+ variables. ▪ Selección automática de predictores.	Modelo de regresión con diagnóstico completo + selección de variables justificada.

Sesión 8: No Paramétrica y Reportes Automáticos

Objetivo: ejecutar pruebas robustas y automatizar reportes reproducibles.

Bloque	Temas principales	Actividad	Evidencia
3h (22–24)	▪ Tests de normalidad (Shapiro-Wilk) y cuándo no paramétricas. ▪ Mann-Whitney, Kruskal-Wallis, Wilcoxon. ▪ Chi-cuadrado avanzado (tablas múltiples). ▪ Reportes: Jupyter $\rightarrow$ PDF/HTML con nbconvert. ▪ Automatización: papermill (notebooks parametrizados).	▪ Dataset no normal del mundo real. ▪ Suite completa de pruebas no paramétricas. ▪ Generación de reporte ejecutivo automático.	Reporte automático PDF/HTML con resultados + interpretación técnica.

MÓDULO 4: MACHINE LEARNING PROFESIONAL (12 horas)

Sesión 9: Fundamentos de ML Industrial con Scikit-Learn

Objetivo: construir pipelines reproducibles y métricas correctas por problema.

Bloque	Temas principales	Actividad	Evidencia
3h (25–27)	▪ Flujo: train/test split estratificado. ▪ Pipelines + ColumnTransformer (preprocesamiento avanzado). ▪ Feature engineering: polynomial, binning, interacciones. ▪ Métricas: RMSE/MAE, Accuracy, Precision/Recall, F1, ROC-AUC. ▪ CV: k-fold, stratified, time series split.	▪ Dataset churn. ▪ Pipeline completo de preprocesamiento + 5 métricas.	Pipeline reproducible listo para producción + evaluación completa.

Sesión 10: Algoritmos Supervisados — Más Allá de lo Básico

Objetivo: entrenar modelos potentes, interpretar y mejorar performance.

Bloque	Temas principales	Actividad	Evidencia
3h (28–30)	▪ Árboles: visualización e interpretación. ▪ Ensambles: Random Forest, Gradient Boosting (XGBoost/LightGBM). ▪ SVM: kernels no lineales. ▪ Redes neuronales básicas: MLPClassifier. ▪ Stacking models: combinación de algoritmos.	▪ Competencia estilo Kaggle (Titanic / House Prices). ▪ Entrenar 5 modelos + ensamble.	Modelo ensemble con mejor performance + comparación de modelos.

Sesión 11: Optimización Hiperparamétrica — Máximo Rendimiento

Objetivo: optimizar de forma eficiente y diagnosticar sobre/infraajuste.

Bloque	Temas principales	Actividad	Evidencia
3h (31–33)	▪ GridSearchCV eficiente; RandomizedSearchCV. ▪ Optimización bayesiana: Optuna. ▪ Feature selection: RFE, SelectFromModel, importancias. ▪ Curvas de aprendizaje: bias-variance, over/underfitting.	▪ Dataset complejo (texto/imagen simplificado). ▪ Optimizar Random Forest con 10+ hiperparámetros. ▪ Analizar trade-off bias-variance.	Modelo optimizado + justificación de hiperparámetros + diagnóstico.

Sesión 12: Proyecto Capstone y Despliegue

Objetivo: resolver un caso completo y presentar un MVP desplegable.

Bloque	Temas principales	Actividad	Evidencia
3h (34–36)	▪ Capstone (2h): dataset sorpresa; limpieza, EDA, modelado, evaluación. ▪ Presentación lightning: 5 min por participante. ▪ Despliegue (1h): MLflow (tracking/registro), FastAPI (API REST), Streamlit (app en 20 líneas), monitoreo (drift y retraining).	▪ Desplegar el mejor modelo como API (concep- tual/demostrativo). ▪ Crear dashboard Streamlit para interactar con el modelo.	Notebook final + demo de despliegue (API/Streamlit) + pitch de resultados.