



FUNDAMENTOS DE INGENIERÍA DE DATOS CON PYTHON

PREPARACIÓN DE DATOS PARA
ANALÍTICA, CIENCIA DE DATOS E
INTELIGENCIA ARTIFICIAL

Ramírez ● Araiza ● Salazar ● Rodríguez



APREND A
EDICIONES

¿Sabes preparar datos para analítica e IA?

Los datos rara vez están listos para ser utilizados. Antes de construir reportes, entrenar modelos o desarrollar soluciones de inteligencia artificial, es necesario comprender, limpiar y transformar la información correctamente.

Fundamentos de analítica de datos con Python presenta una metodología práctica para preparar datos utilizando Python y pandas, abordando las principales actividades desarrolladas en proyectos de analítica e ingeniería de datos, y estableciendo las bases necesarias para trabajos posteriores de ciencia de datos.

A través de ejercicios y laboratorios, el lector aprenderá a explorar conjuntos de datos, corregir problemas de calidad, transformar estructuras, generar variables derivadas y automatizar procesos mediante pipelines de preparación de datos.

Más allá de la programación, esta obra desarrolla criterio analítico: la capacidad de convertir datos en información consistente, confiable y lista para generar valor.

Dirigido a estudiantes, analistas, profesionistas e ingenieros, este libro constituye una guía práctica para quienes buscan utilizar Python como herramienta para el procesamiento y preparación de datos.

Aprenderás a:



Convertir
fenómenos
reales en
datos útiles



Clasificar
datos
mediante
DTXC



Limpiar y
transformar
datos con
pandas



Tratar datos
ausentes y
atípicos



Construir
pipelines para
analítica e IA



Forma parte del cuerpo de conocimiento para la certificación **CDE** (**Certified Data Expert**).



Escanea para acceder a
recursos descargables.



APRENDA
EDICIONES



Índice del libro

A continuación encontrarás el índice del libro, para que conozcas a detalle su contenido y estructura.

Ir a librería



Conoce todos los títulos disponibles en APRENDA. Aquí también podrás descargar archivos y recursos complementarios que acompañan a los títulos.

Comprar



Si el libro es de tu interés y deseas comprarlo, aquí puedes hacerlo.

Contenido

1.	INTRODUCCIÓN, RECURSOS Y DINÁMICA DE TRABAJO	9
1.1.	¿Para quién es este libro?.....	9
1.1.1.	Audiencia específica	9
1.1.2.	¿Por qué leerlo?	10
1.1.3.	Estructura del libro	11
1.2.	Archivos y recursos de trabajo.....	12
1.2.1.	Formato de los archivos de trabajo	12
1.2.2.	Plataforma de trabajo	12
1.2.3.	Archivos de trabajo y material adicional.....	13
1.3.	LAB: Verificar el acceso a los datos de prueba.....	13
1.3.1.	Uso de los LABS	15
2.	FUNDAMENTOS DE ANALÍTICA, DATOS E INFORMACIÓN	17
2.1.	Conceptos	17
2.1.1.	Dato.....	17
2.1.2.	Información.....	18
2.1.3.	Analítica de datos	20
2.2.	Tipos de análisis	22
2.3.	Técnicas para definir el tipo de objetivo	23
2.4.	Analítica de datos, Big Data y Data Science	24
3.	FUNDAMENTOS DEL PROCESO ETL (EXTRACT, TRANSFORM, LOAD).....	27
3.1.	Ambientes.....	27
3.2.	Repositorios de datos.....	29
3.2.1.	Almacén de datos (Data Warehouse).....	29
3.2.2.	Master de datos	29
3.3.	La ruta de ETL.....	31
3.3.1.	Fase de extracción (Extract)	31
3.3.2.	Fase de transformación (Transform)	33
3.3.3.	Fase de carga (Load)	34
4.	COMPRESIÓN DEL CASO DE ESTUDIO Y DEFINICIÓN DE ANÁLISIS.....	37
4.1.	Inmersión	39
4.2.	LAB: Analizar el contexto del caso	40
4.3.	Observación informal	41
4.4.	LAB: Identificar los supuestos del caso	42
4.4.1.	Diferencia entre supuestos e hipótesis	43

4.5.	Declaración del objetivo de análisis.....	44
4.6.	LAB: Redactar el objetivo del análisis del caso	47
4.7.	Desarrollo de hipótesis	48
4.7.1.	Concepto de hipótesis	48
4.8.	LAB: Redactar las hipótesis del caso	49
4.9.	Análisis de coherencia de hipótesis.....	51
4.10.	LAB: Revisar la coherencia de las hipótesis del caso.....	51
4.11.	Enumeración de fuentes de datos.....	53
4.12.	LAB: Documentar información de la fuente.....	54
4.13.	Identificación de variables	55
4.14.	LAB: Identificar variables dependientes e independientes	56
5.	ANÁLISIS PRELIMINAR DE DATOS USANDO PANDAS	59
5.1.	Análisis censal y muestral	59
5.1.1.	Concepto de muestra estadística	59
5.1.2.	Cálculo del tamaño de la muestra estadística	61
5.2.	LAB: Calcular el tamaño de la muestra para el caso	62
5.3.	Series y DataFrames.....	63
5.4.	Listas, diccionarios y tuplas.....	65
5.4.1.	Listas	66
5.4.2.	Diccionarios	66
5.4.3.	Tuplas.....	67
5.4.4.	Iterabilidad	67
5.4.5.	Creación de Serie y DataFrame a partir de listas y diccionarios ..	67
5.5.	LAB: Trabajar con listas, diccionarios y tuplas	68
5.6.	Carga de datos a un DataFrame	72
5.6.1.	Cargando datos desde CSV.....	72
5.6.2.	Cargando datos desde Libros de Excel	73
5.6.3.	Cargando datos desde texto sin delimitadores	74
5.7.	Forma de los datos (shape)	76
5.8.	LAB: Cargar datos a un DataFrame desde un archivo CSV	77
6.	ANÁLISIS SEMÁNTICO Y TÉCNICO DE VARIABLES	81
6.1.	Tipos de datos en pandas	81
6.2.	Semántica de los datos	82
6.3.	Análisis de la estructura de datos	82
6.4.	Verificación de relevancia de variables.....	83
6.5.	LAB: Análisis semántico de las variables	86
6.6.	Clasificación de los datos	90
6.6.1.	Categorías de los datos.....	90
6.6.2.	Data Taxonomic Code (DTXC)	91

6.7.	LAB: Clasificar los datos usando código DTXC	93
7.	EXPLORACIÓN, VISUALIZACIÓN Y FILTRADO DE DATOS	95
7.1.	Ver los datos del DataFrame	95
7.1.1.	Ver todos los datos del Data Frame	95
7.1.2.	Ver los encabezados y colas de datos	96
7.1.3.	Filtrado de columnas.....	96
7.2.	LAB: Ver datos del DataFrame	97
7.3.	Técnicas de filtrado de filas.....	100
7.3.1.	Operadores comparativos en pandas.....	100
7.3.2.	Operadores lógicos en pandas.....	101
7.3.3.	Filtrando filas usando loc[]	101
7.3.4.	Filtrando filas usando where()	102
7.4.	LAB: Técnicas de filtrado de filas	102
8.	LIMPIEZA INICIAL Y PREPARACIÓN DE DATAFRAMES	107
8.1.	Limpieza de datos (data cleaning)	107
8.2.	Ingeniería de datos (feature engineering).....	109
8.3.	Tareas generales con DataFrame	111
8.3.1.	Reducción de dimensionalidad	111
8.3.2.	Eliminación de filas duplicadas	113
8.3.3.	Eliminación de filas vacías.....	114
8.3.4.	Creación de un identificador ficticio	116
8.3.5.	Verificación de unicidad en columnas	116
8.3.6.	Escribir un CSV a partir de un DataFrame	117
8.3.7.	Escribir un Archivo de Excel a partir de un DataFrame	119
8.4.	LAB: Tareas generales con DataFrames	120
8.5.	Tareas generales con columnas	124
8.5.1.	Reordenar las columnas de un DataFrame	124
8.5.2.	Cambiar el nombre de las columnas en un DataFrame	124
8.6.	LAB: Trabajo con columnas.....	125
8.7.	¿Qué sigue? Transformación de datos	131
9.	CONVERSIÓN Y NORMALIZACIÓN DE TIPOS DE DATOS	133
9.1.	Conversión a diferente tipo de dato	134
9.1.1.	Conversiones usando astype() y to_datetime().....	134
9.1.2.	Conversiones al momento de cargar datos	137
9.2.	Conversiones de unidades de medida.....	138
9.3.	Conversiones de unidades monetarias	138
9.4.	LAB: Ejecutando conversiones de tipo y de moneda	140

10.	TRANSFORMACIÓN DE DATOS DE TEXTO	145
10.1.	Columnas con diferente casing.....	145
10.2.	División de cadenas.....	146
10.3.	Eliminar espacios en los extremos	148
10.4.	Concatenación de cadenas	148
10.5.	Extracción de subcadenas.....	149
10.6.	LAB: Derivaciones a partir de cadenas	151
11.	CREACIÓN DE COLUMNAS DERIVADAS.....	161
11.1.	Columnas derivadas usando fórmulas	161
11.2.	Columnas derivadas usando UDF (user defined functions)	162
11.3.	Columnas derivadas usando lamda	164
11.4.	LAB: Cálculos con columnas	165
12.	CODIFICACIÓN Y TRATAMIENTO DE VARIABLES CATEGÓRICAS.....	169
12.1.	Equivalencias categóricas.....	169
12.1.1.	Tipos de categóricos	170
12.1.2.	Generando equivalencia usando map()	171
12.1.3.	Generando equivalencia usando merge()	172
12.2.	LAB: Generación de categóricos descriptivos equivalentes	175
12.3.	Equivalencia de intervalos	180
12.4.	LAB: Generación de categóricos de intervalo.....	181
13.	TRATAMIENTO DE DATOS AUSENTES (MISSING).....	189
13.1.	Identificación de datos ausentes	189
13.1.1.	Causas de datos ausentes.....	189
13.1.2.	Tipos de datos ausentes	190
13.2.	Estrategias para tratar datos ausentes	194
13.2.1.	Revisión inicial de posibles faltantes	196
13.2.2.	Eliminación de ausentes	197
13.2.3.	Sustitución por omisión	199
13.2.4.	Sustitución de un valor específico	200
13.2.5.	Cálculo basado en requeridos indirectos.	200
13.2.6.	Cálculo máxima verosimilitud.....	201
13.3.	LAB: Tratamiento de datos ausentes	202
14.	TRATAMIENTO DE DATOS ATÍPICOS (OUTLIERS)	209
14.1.	Identificación de datos atípicos (outliers)	209
14.1.1.	Datos atípicos	209
14.1.2.	Percentiles y cuartiles.....	210

14.1.3.	Regla de Tukey	211
14.2.	Estrategias para tratar datos atípicos	216
14.2.1.	Dejar como están	217
14.2.2.	Eliminación de atípicos	217
14.2.3.	Truncamiento de atípicos.....	217
14.2.4.	Transformación de atípicos.....	217
14.2.5.	Binning de atípicos	218
14.2.6.	Imputación de atípicos	218
14.3.	LAB: Tratamiento de datos atípicos	218
14.4.	LAB: Tratamiento de datos atípicos y ausentes para el Titanic	232
15.	TÉCNICAS DE MUESTREO.....	237
15.1.	Muestra aleatoria simple	237
15.2.	Muestra estratificada	239
15.3.	LAB: Muestra aleatoria simple y muestra estratificada	240
16.	PIPELINES DE PREPARACIÓN DE DATOS.....	249
16.1.	¿Qué son los Pipelines?.....	249
16.2.	¿Cómo se implementan los Pipelines?	251
16.3.	Función base para Pipeline	252
16.4.	LAB: Uso de pipelines para preparación de datos	252
17.	CASO PRÁCTICO DE LIMPIEZA E INGENIERÍA DE DATOS	259
17.1.	Preparación integral de datos para analítica	259
17.1.1.	Análisis inicial de la estructura de datos	260
17.1.2.	2. Selección y organización de variables.....	260
17.1.3.	3. Validación de calidad de los datos	260
17.1.4.	4. Limpieza de datos	261
17.1.5.	5. Transformación de datos	261
17.1.6.	6. Ingeniería de datos (Feature Engineering)	261
17.1.7.	7. Generación de categóricos e integración final	262
17.1.8.	¿Por qué seguir esta secuencia?	262
17.1.9.	Resultado esperado.....	263
17.2.	Nota sobre datos ausentes, datos atípicos y muestreo	263
17.3.	Sobre el modelado de datos	264
17.4.	LAB: Preparación de datos usando Python y pandas.....	265
18.	ÍNDICE	286