Saturdays.AI
LATAM

Saturdays.AI Latam Online 2da. Edición

MACHINE LEARNING

EDA

Contenido

¿Qué y porqué EDA?

Objetivos y consideraciones

Cómo realizar un EDA

Herramientas

Saturdays.AI
LATAM

¿Qué es EDA?

Exploratory Data Analysis

Conjunto de procedimientos cuyo objetivo general es proporcionar una **visión detallada y precisa de las variables estudiadas.**

descriptivo
visual
analítico
sin ideas preconcebidas

} “exploratorio”

Saturdays.AI
LATAM

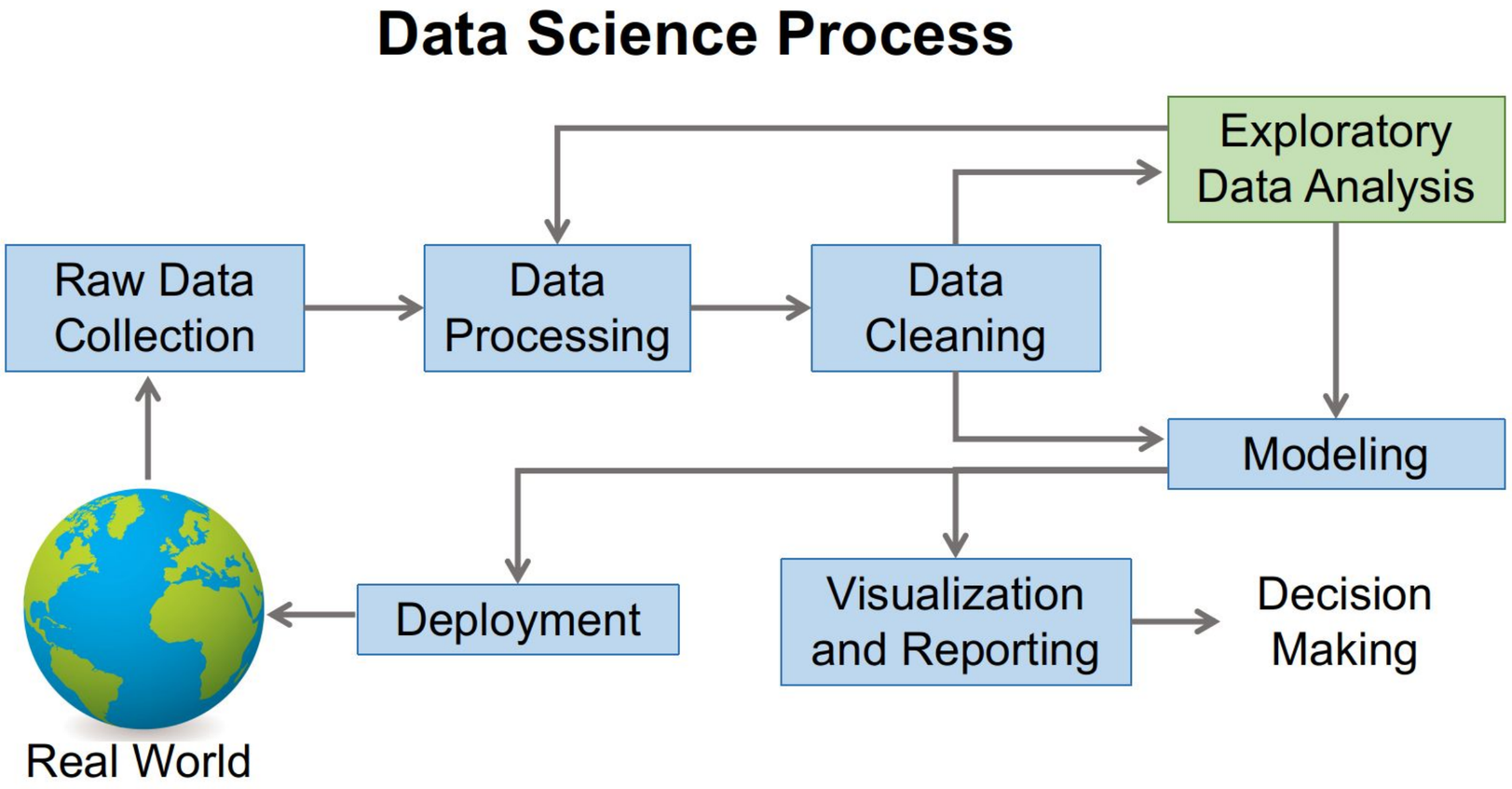
¿Porqué EDA?

El cerebro no puede procesar la gran cantidad de datos generados

Saturdays.AI
LATAM

¿Porqué EDA?

Todo problema de ML inicia con EDA



Saturdays.AI LATAM

Objetivos del EDA

1. Naturaleza de los datos
2. Características de la distribución de las variables
3. Relaciones entre variables
4. Valores atípicos (aberrante, *outlier*)
5. Extraer variables importantes
6. Probar y definir supuestos

Saturdays.AI
LATAM

¿Qué debo considerar antes de realizar un EDA?

1. Hazte las preguntas correctas
2. Ten conocimientos básicos sobre el dominio del tema / Haz un *research* / pregunta a algún experto
3. Nunca olvides tu objetivo

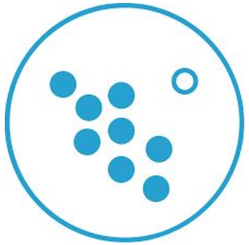
Saturdays.AI
LATAM

¿Cómo realizar el EDA?

Tipos de datos
Todas las variables son del mismo tipo?
Las fechas están siendo consideradas como fechas?

Datos faltantes
Hay datos faltantes? Son muchos? Faltan sistemáticamente? Los puedo eliminar?

Rango de lo datos
Están todos los valores en el rango de datos esperado?



Datos atípicos
Hay datos extremos?
Son relevantes en el análisis?

Estructura de los datos
Cada variable forma una columna
Cada observación forma una fila
Cada celda es una medición

Análisis univariado, bi y multivariado
Qué distribución tienen los datos? Existe relación entre diferentes datos? Qué patrones hay? Qué relación hay? Tiempo-espacio?

Saturdays.AI
LATAM

Herramientas de Python

EDA

Manipulación de datos y estadística

pandas

Manipulación de datos, estadística y matemática

NumPy

SciPy

Visualización

matplotlib

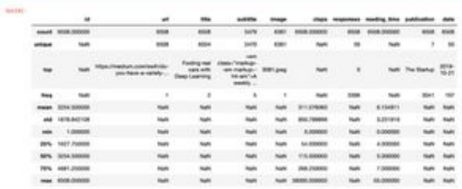
seaborn

Saturdays.AI
LATAM

Herramientas de Python

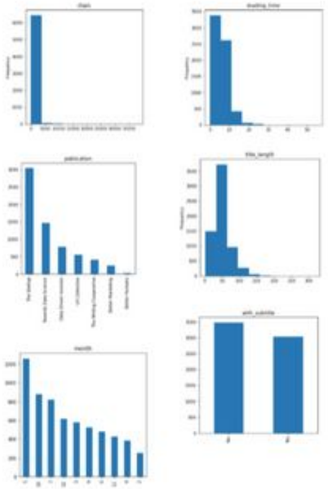
EDA CHEATSHEET

Non-graphical Analysis



df.info( )
df.describe( )
df.isnull( )

Univariate Analysis

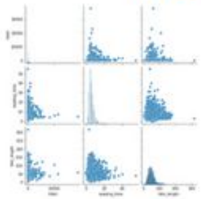


Numerical:
df[column].plot(kind = “hist”)

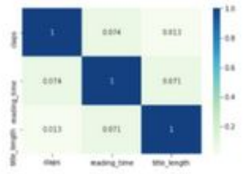
Categorical:
df[column].plot(kind = “bar”)

Multivariate Analysis

Numerical vs. Numerical

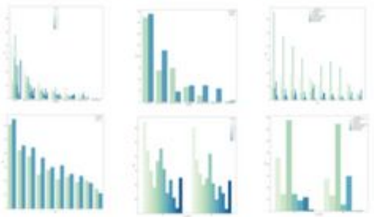


sns.pairplot( )



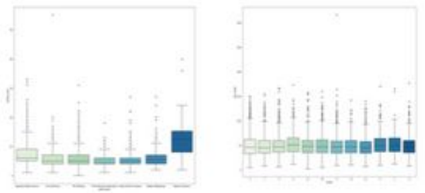
sns.heatmap( )

Categorical vs. Categorical

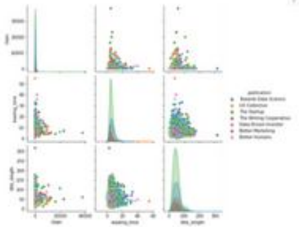


sns.countplot(hue = …)

Categorical vs. Numerical



sns.boxplot( )



sns.pairplot(hue = …)

Saturdays.AI
LATAM

Saturdays.AI
LATAM

Conclusiones

No omitas el EDA en tu proyecto de inteligencia artificial, te ayudará a tener un mejor entendimiento de tus datos y cómo abordarlos.

Un EDA poco exhaustivo puede disminuir el performance de tu modelo.

Saturdays.AI
LATAM

Rosa Sunum

rosa.sunum@saturdays.ai

Saturdays.AI
LATAM