Tratamiento y análisis de datos

Curso de Experto en Inteligencia Artificial y Machine Learning

1. LECTURA dE dATOS

```
pd.read_csv()
pd.read_excel()
pd.read_sql()
pd.read_json()
….
```

4. dATOS CUANTITATIVOS

```
df[“col”].describe()
df[“col”].mean()
df[“col”].median()
df[“col”].quantile(0,95)
df[“col”].mode()[0]
df[“col”].hist(bins=X)
…
```

5. dATOS CUALITATIVOS

```
df[“col”].value_counts()
df[“col”].value_counts().plot.bar()
df[“col”].unique()
df[“col”].nunique()
df[“col”].mode()[0]
```

2. PRIMEROS PASOS

```
df.head() o  df.tail()
df.info()
df.rename(columns={‘col_antes’:’col_ahora’}, inplace = True)
df.drop(‘col’, axis=1) es lo mismo que del(df[‘col’])
df[“col”] = df[“col”].astype(int)
```

6. ÍNdICES

```
df.set_index(“col”, inplace=True)
df.reset_index(inplace=True)
```

3. ANÁLISIS dE NULOS

```
df.isnull().any() o df.isnull().all()
df.isnull().sum()
df.notnull().sum()
df_sin_nulls = df[df[‘col’].notnull()] es lo mismo que df_sin_nulls = df.dropna()
df[‘col’].fillna(valor, inplace=True)
```

7. dATOS dUPLICAdOS

```
df[“col”].duplicated()
df[“col”].duplicated().sum()
df.drop_duplicates(subset=[columnas], keep=“first”)
```

8. CONSULTAS Y FILTROS

```
Consultas simples:
df[df[“col1”]==‘k’]
Consultas compuestas:
df.query(“col1==‘ok’ & col2>5)
```

9. AGREGACIONES

```
df.groupby([columnas_datos_cualitativos])[columna_datos_cuantitativos].mean()
df.groupby([columnas_datos_cualitativos])[columna_datos_cuantitativos].aggregate([“mean”, “max”])
df.groupby([columnas_datos_cualitativos])[columna_datos_cuantitativos].mean().reset_index()
df.groupby([columnas_datos_cualitativos])[columna_datos_cuantitativos].transform(‘mean’)
df.groupby([columnas_datos_cualitativos])[columna_datos_cuantitativos].apply(lambda x: x-x.min())
```

10. PIVOT TABLES

```
df.pivot_table(values=“col1",index=“col2", columns=“col3",aggfunc="sum")
df.pivot_table(values=“col1",index=[“col2“, “col3”], columns=“col4",aggfunc="sum")
df.pivot_table(values=“col1",index=[“col2“, “col3”], columns=“col4",aggfunc=["sum“, “mean”))
```