Pipeline de Ciencia de Datos

Leon Felipe Davis Coropuna

Orientador: Prof Ana Maria Cuadros Valdivia

Plan de Tesis presentado la Escuela Profesional Ciencia de la Computación como paso previo a la elaboración de la Tesis Profesional.

UNSA - Universidad Nacional de San Agustín de Arequipa
Junio de 2025

Índice

1. Datasets

1.1. Crimes - 2001 to Present

Este dataset proporciona un registro de incidentes delictivos en la ciudad de Chicago desde el año 2001 hasta la actualidad. Los datos son proporcionados por el sistema **CLEAR** (Citizen Law Enforcement Analysis and Reporting) del Departamento de Policía de Chicago, y son actualizados diariamente.

En total, el conjunto de datos contiene más de 8.3 millones de registros y 22 columnas, incluyendo datos espaciales disponible en [Chicago Police Department, 2025]. Para el análisis, se tomó una muestra de datos desde el año 2020.

1.2. Tweets sobre Criminalidad

Este dataset fue obtenido mediante técnicas de *scraping* en la plataforma X.com (antes Twitter), utilizando palabras clave relacionadas con crímenes. Se recolectaron aproximadamente 15,033 registros, cada uno representando un tweet individual junto con sus respectivos metadatos.

La información fue inicialmente almacenada en archivos JSON organizados cronológicamente, y posteriormente unificada y convertida en un archivo CSV con 106 atributos por registro.

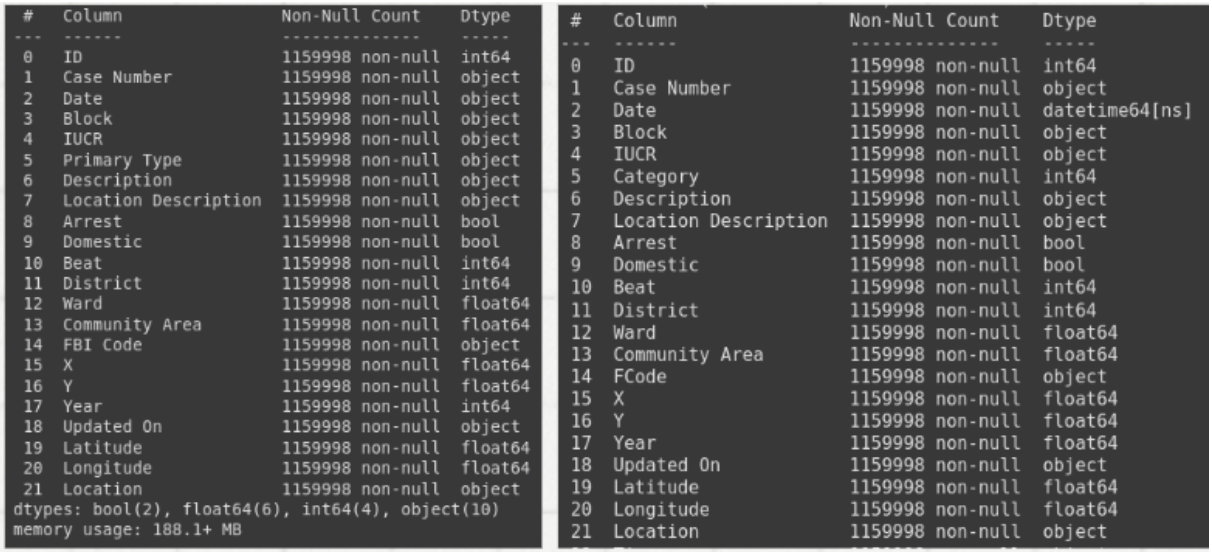
Periodo de los tweets: Enero a abril de 2020.

Fuente: X.com

2. Problemas de los Datasets

2.1. Crimes - 2001 to Present

- **Datos faltantes y duplicados:** Algunos registros contenían valores nulos o estaban duplicados.
  - Registros originales: 1,183,866
  - Registros después de limpieza: 1,159,998
  - Registros eliminados: 23,868 (2.02 %)
- **Variables categóricas:** No numéricas, como “District”, no pueden ser procesadas directamente por modelos. Por ello se aplica un labelencoding y se obtuvieron nuevos tipos de datos como se ve en la figura 1.



#	Column	Non-Null Count	Dtype
0	ID	1159998 non-null	int64
1	Case Number	1159998 non-null	object
2	Date	1159998 non-null	object
3	Block	1159998 non-null	object
4	IUCR	1159998 non-null	object
5	Primary Type	1159998 non-null	object
6	Description	1159998 non-null	object
7	Location Description	1159998 non-null	object
8	Arrest	1159998 non-null	bool
9	Domestic	1159998 non-null	bool
10	Beat	1159998 non-null	int64
11	District	1159998 non-null	int64
12	Ward	1159998 non-null	float64
13	Community Area	1159998 non-null	float64
14	FBI Code	1159998 non-null	object
15	X	1159998 non-null	float64
16	Y	1159998 non-null	float64
17	Year	1159998 non-null	int64
18	Updated On	1159998 non-null	object
19	Latitude	1159998 non-null	float64
20	Longitude	1159998 non-null	float64
21	Location	1159998 non-null	object

dtypes: bool(2), float64(6), int64(4), object(10)
memory usage: 188.1+ MB

#	Column	Non-Null Count	Dtype
0	ID	1159998 non-null	int64
1	Case Number	1159998 non-null	object
2	Date	1159998 non-null	datetime64[ns]
3	Block	1159998 non-null	object
4	IUCR	1159998 non-null	object
5	Category	1159998 non-null	int64
6	Description	1159998 non-null	object
7	Location Description	1159998 non-null	object
8	Arrest	1159998 non-null	bool
9	Domestic	1159998 non-null	bool
10	Beat	1159998 non-null	int64
11	District	1159998 non-null	int64
12	Ward	1159998 non-null	float64
13	Community Area	1159998 non-null	float64
14	FCode	1159998 non-null	object
15	X	1159998 non-null	float64
16	Y	1159998 non-null	float64
17	Year	1159998 non-null	float64
18	Updated On	1159998 non-null	object
19	Latitude	1159998 non-null	float64
20	Longitude	1159998 non-null	float64
21	Location	1159998 non-null	object

Figura 1: Label encoding y transformación de tipos de datos

- **Rangos de valores diferentes:** Algunas variables numéricas estaban en diferentes escalas usando StandartScale() o MinMaxScaler() según corresponda, el resultado se puede observar en 2
- **Información temporal subutilizada:** La columna Date como texto limita el análisis temporal. Por ello se uso la columna Date y de ella se derivaron atributos de tiempo, los nuevos atributos son mostrados en la figura 3

2.2. Tweets de X.com

- **Columnas inservibles:** Algunas columnas con valores totalmente vacíos, registros duplicados 4.
- **Información temporal subutilizada:** Fecha en formato texto limita la detección de patrones 5.
- **Normalización de texto:** Presencia de emojis, menciones, hashtags, etc 6.
- **Falta de etiquetas:** No se indica explícitamente el tipo de crimen 7.

3. Descubrimientos Relevantes

- El análisis de **mutual information** permitió evaluar la importancia de cada variable respecto a la categoría del crimen 8.
- La cantidad de **outliers** es baja, pero eliminarlos directamente puede ser contraproducente 10.

```
from sklearn.preprocessing import LabelEncoder

categorical_cols = ['District', 'Category', 'Hour_Zone', 'Season']
label_encoders = {}

for col in categorical_cols:
    le = LabelEncoder()
    df[col] = le.fit_transform(df[col])
    label_encoders[col] = le
```

```
from sklearn.preprocessing import MinMaxScaler

to_scale = ['Hour', 'Minute', 'Year', 'Month', 'dayOfWeek', 'dayOfMonth',
            'dayOfYear', 'weekOfMonth', 'weekOfYear']

scaler = MinMaxScaler()
df[to_scale] = scaler.fit_transform(df[to_scale])
```

Figura 2: Escalado de variables con MinMaxScaler y StandartScale

```
22  Month                1159998 non-null  int32
23  dayOfWeek            1159998 non-null  int32
24  dayOfMonth           1159998 non-null  int32
25  dayOfYear            1159998 non-null  float64
26  weekOfMonth          1159998 non-null  int64
27  weekOfYear           1159998 non-null  UInt32
28  Hour                 1159998 non-null  int32
29  Minute               1159998 non-null  int32
30  Hour_Zone            1159998 non-null  category
31  BusinessHour         1159998 non-null  int64
32  Weekend              1159998 non-null  int64
33  Season               1159998 non-null  object
34  Holiday              1159998 non-null  bool
35  Rot30_X              1159998 non-null  float64
36  Rot30_Y              1159998 non-null  float64
37  Cluster              1159998 non-null  int32
```

Figura 3: Atributos derivados de Date

```
Valores nulos por columna (y su porcentaje):

                              Nulos  Porcentaje (%)
viewCount                     15033          100.00
retweetedTweet                15033          100.00
quotedTweet                   15033          100.00
place                         15033          100.00
coordinates                   14981           99.65
...                             ...             ...
quotedTweet_card              14974           99.61
quotedTweet_possibly_sensitive 14501          96.46
quotedTweet__type             14263           94.88
card_options                  15020           99.91
card_finished                 15020           99.91

[67 rows x 2 columns]
```

Figura 4: Columnas con muchos vacíos

```python
df["Year"] = df["Date"].dt.year
df["Month"] = df["Date"].dt.month
df["dayOfWeek"] = df["Date"].dt.dayofweek  # lunes=0, domingo=6
df["dayOfMonth"] = df["Date"].dt.day
df["dayOfYear"] = df["Date"].dt.dayofyear
df["weekOfYear"] = df["Date"].dt.isocalendar().week
df["weekOfMonth"] = df["Date"].apply(lambda x: int(np.ceil(x.day / 7.0)))
df["Hour"] = df["Date"].dt.hour
df["Minute"] = df["Date"].dt.minute
df["Time"] = df["Date"].dt.strftime("%H:%M:%S")
df["Weekend"] = df["dayOfWeek"].isin([5, 6]).astype(int)
df["BusinessHour"] = df["Hour"].apply(lambda x: 1 if 9 <= x <= 17 else 0)

# Zona horaria simple según la hora
df["Hour_Zone"] = pd.cut(df["Hour"],
                         bins=[-1, 5, 11, 17, 21, 24],
                         labels=[0, 1, 2, 3, 4]).astype(int)
```

Figura 5: Atributos derivados de Date en el dataset de Tweets

```
Woman shot in head on Far South Side, man crit.
@JoeStreckert If you want the basics in a much.
"How to Get Away with Murder" writers created .
@dude_vol @PatriotAlways2 @prettynikkivar You .
Man Killed In Bishop Ford Freeway Shooting htt.
```

```
woman shot head far south side man critically .
want basic muchmore digestible format rick gea.
get away murder writer created gabriel maddox .
understand rare compared weekly death toll cit.
          man killed bishop ford freeway shooti
```

Figura 6: Textos normalizados con regex para quitar emojis, etc

```python
model_name = "Luna-Skywalker/BERT-crime-analysis"
tokenizer = AutoTokenizer.from_pretrained(model_name)
model = AutoModelForSequenceClassification.from_pretrained(model_name)
```

```python
classifier = pipeline("zero-shot-classification", model="facebook/bart-large-mnli")
```

	PredictedCrime	Confidence
0	assault	0.529124
1	murder	0.644339
2	murder	0.606229
3	homicide	0.271090
4	homicide	0.484051

Figura 7: Modelos para etiquetar comentarios de tweets de acuerdo al crimen mencionado

	Feature	MI_Score
0	ICode	2.396172
1	Rot30_Y	0.216835
2	Radius	0.216441
3	Rot60_Y	0.214921
4	Rot60_X	0.213040
5	Rot30_X	0.211844
6	Angle	0.211687
7	Y	0.192693
8	Rot45_X	0.189836
9	Rot45_Y	0.184477
10	X	0.169264
11	Minute	0.081840
12	Cluster	0.061339
13	District	0.053931
14	Hour	0.046175
15	BusinessHour	0.035876
16	Hour_Zone	0.034600
17	Year	0.011988
18	Weekend	0.009897
19	Season	0.009304
20	dayOfYear	0.008485
21	weekOfMonth	0.006682
22	dayOfWeek	0.006550
23	weekOfYear	0.006334
24	Month	0.005412
25	dayOfMonth	0.002081
26	Holiday	0.000425

Figura 8: Atributos seleccionados según la selección de características

ICode	2.468959
FCode	2.445897
Angle	0.289704
Rot30_Y	0.287521
Rot45_Y	0.286572
Rot60_X	0.286064
Rot45_X	0.285376
Rot30_X	0.284664
Rot60_Y	0.284502
Radius	0.283441
Y	0.204331
ID	0.174396
X	0.170194
BusinessHour	0.126087
Minute	0.098302
Cluster	0.077721
Hour_Zone	0.074456
District	0.067808

Figura 9: Atributos seleccionados de un paper que usa un dataset similar

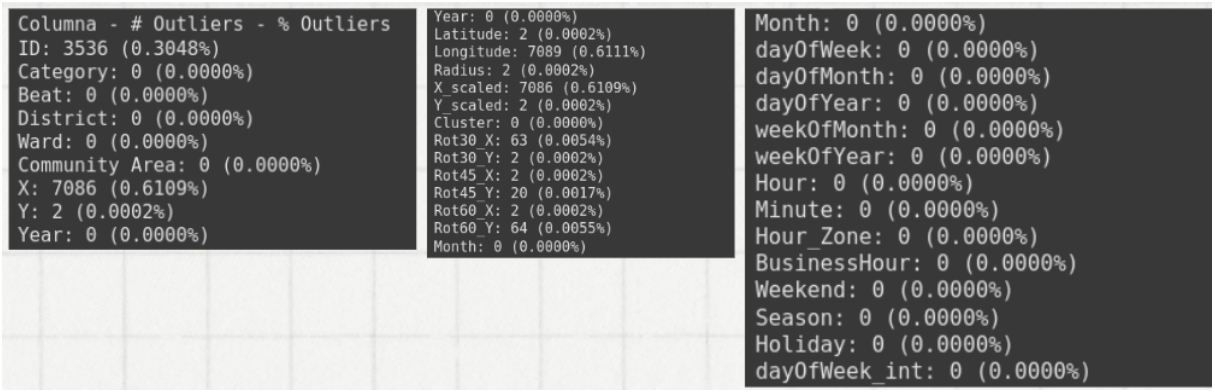


Figura 10: Outliers por porcentaje en las columnas del dataset de tweets

- **Las estaciones del año** influyen significativamente en la frecuencia de crímenes.
- Los crímenes más comunes son hurto, agresión física y daño a la propiedad.
- Los tweets tienden a mencionar más homicidios o drogas, a diferencia del dataset gubernamental donde predominan los robos.

Referencias

[Chicago Police Department, 2025] Chicago Police Department (2025). Crimes - 2001 to present. https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2/about$_data. Datasetactualizadoal29demayode2025. ProporcionadoporChicagoPoliceDepartmen$