

Projeto Prático II (SAD)

Mineração de Dados em Crimes de Chicago

Classificação da variável Arrest

Seu Nome

UFOP – Sistemas de Apoio à Decisão

25 de fevereiro de 2026

Contexto e Enunciado

Tema: aplicação de mineração de dados em base real para atender ao Projeto Prático II da disciplina SAD.

Exigências do enunciado contempladas:

- Caracterização da base de dados selecionada.
- Definição do problema e tarefa de mineração escolhida.
- Análises exploratórias e resultados obtidos.
- Avaliação com precisão, revocação, F1-score e matriz de confusão.

Tempo do seminário: 15 min + 5 min de discussão.

Problema e Motivação

Base: Chicago Crimes (arquivo `Crimes_-_2001_to_Present.csv`).

Problema proposto (classificação binária):

- Dado um registro de ocorrência, prever se houve prisão (`Arrest = 1`).

Justificativa:

- Alvo supervisionado bem definido.
- Relevância para análise operacional e identificação de padrões.
- Permite uso das métricas exigidas no enunciado.

Hipótese inicial: tipo de crime, local, horário, distrito e indicador de violência doméstica influenciam a chance de prisão.

Base de Dados e Recorte de Atributos

Estratégia de leitura no notebook:

- Seleção de colunas relevantes ao problema (12 atributos + alvo).
- Conversão de datas (Date) e padronização de booleanos (Arrest, Domestic).
- Nesta execução final: **base completa**.

Print sugerido: saída da leitura da base (df.head() + linhas/colunas lidas).

Saída da leitura da base (celula 5)

Arquivo: Crimes - 2001_to_Present.csv
Linhas lidas: 7,784,664
Colunas selecionadas: 12

	Date	Primary Type	Description	\
0	2015-09-05 13:30:00	BATTERY	DOMESTIC SIMPLE	
1	2015-09-04 11:30:00	THEFT	POCKET-PICKING	
2	2018-09-01 08:01:00	THEFT	OVER \$500	
3	2015-09-05 12:45:00	NARCOTICS	POSS: HEROIN(BRN/TAN)	
4	2015-09-05 13:00:00	ASSAULT	SIMPLE	

	Location	Description	Arrest	Domestic	Beat	District	Community Area	\
0	RESIDENCE	False	False	924	0.0000	51	0.0000	
1	CTA BUS	False	False	1511	15.0000		25.0000	
2	RESIDENCE	False	True	631	6.0000	44	0.0000	
3	SIDEWALK	True	False	1412	14.0000		21.0000	
4	APARTMENT	False	True	1522	15.0000		25.0000	

	Year	Latitude	Longitude
0	2015	41.8151	-87.6700
1	2015	41.8951	-87.7654
2	2018	NaN	NaN
3	2015	41.9374	-87.7166
4	2015	41.8819	-87.7551

Caracterização da Base

- Verificação de dimensão da base e tipos de dados.
- Análise de valores ausentes por coluna.
- Distribuição da variável alvo Arrest (desbalanceamento de classes).

Ponto de atenção: a taxa da classe positiva (prisão) impacta a interpretação das métricas.

Taxa de casos com prisão (base completa): XX,XX%.

Caracterização da base (celula 7)

```
Shape da amostra: (7784664, 12)
class: 'pandas.core.DataFrame'
Rows: 7784664
Data columns (total 12 columns):
 #   Column          Dtype  
 0   Date            datetime64[us]
 1   Primary Type    str    
 2   Description     str    
 3   Location         str    
 4   Arrest           bool   
 5   Domestic         bool   
 6   Beat             int64  
 7   District         float64
 8   Community Area  int64  
 9   Latitudes       float64
 10  Longitudes      float64
dtypes: bool(2), datetime64[us](1), float64(4), int64(2), str(3)
memory usage: 888.8 MB
```

None

```
Community Area    missing  pct
013476 7.000000
Latitude        0.000000
Longitude       0.000000
Longitude       0.000000
Location Description  0.000000
District         0.000000
Description      0.000000
Date             0.000000
Primary Type    0.000000
Beat             0.000000
Domestic         0.000000
Arrest            0.000000
Year              0.000000

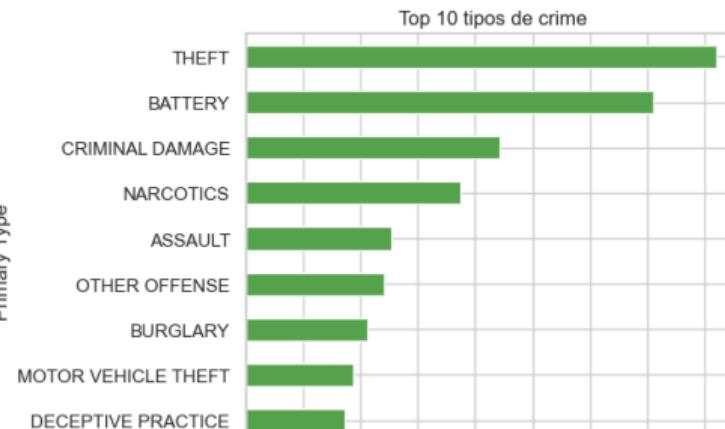
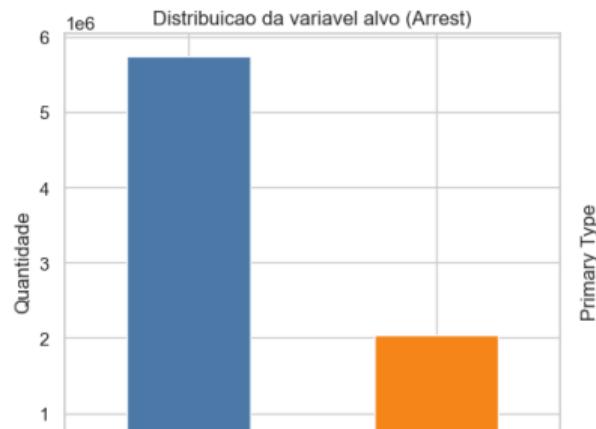
Arrest   Count  Pct
0   False  5793866 73.00000
1   True   2034766 26.1381
```

Análise Exploratória Geral (EDA)

Objetivo: entender a composição da base antes da modelagem.

- Distribuição do alvo (Arrest)
- Top 10 tipos de crime
- Top 10 locais de ocorrência
- Distribuição de Domestic

Leitura esperada do gráfico: identificar frequência, concentração e possíveis variáveis com poder preditivo.



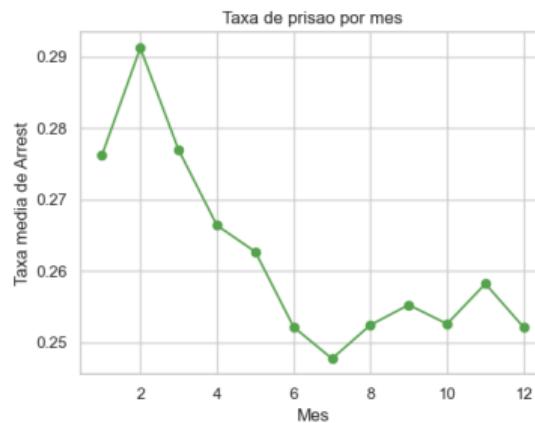
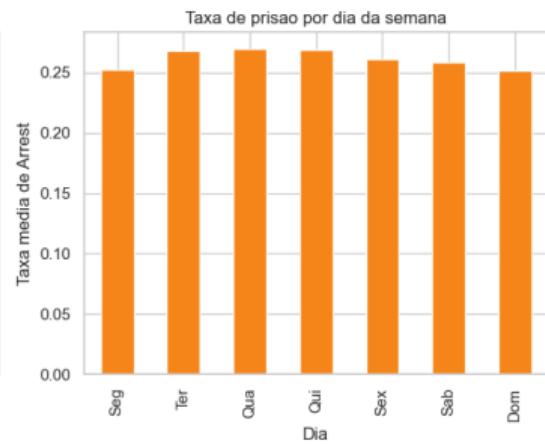
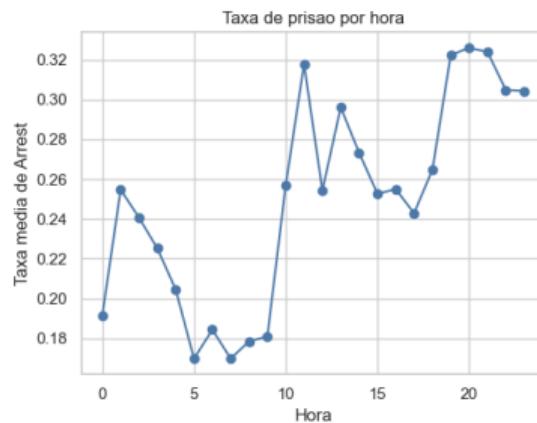
Análise Temporal e Relação com Prisão

Atributos temporais derivados de Date:

- Hora (Hour)
- Dia da semana (DayOfWeek)
- Mês (Month)

Motivação: variações temporais podem aumentar o poder preditivo do modelo.

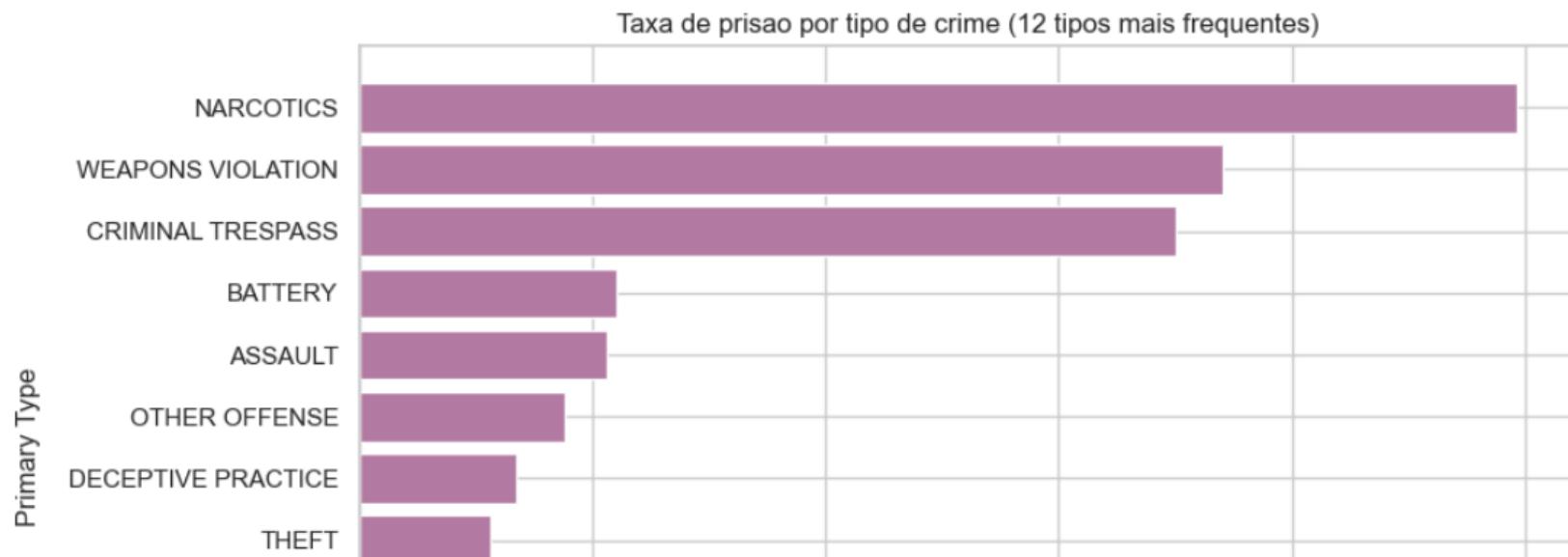
Fale no seminário: destaque 2 padrões visuais observados (picos/vales por hora, dia ou mês).



Tipo de Crime x Taxa de Prisão

- Cálculo da taxa média de prisão por Primary Type.
- Foco nos tipos mais frequentes para evitar distorções por categorias raras.
- Evidência de relação entre categoria do crime e probabilidade de prisão.

Mensagem principal: o tipo de crime é um atributo informativo para a classificação.



Preparação dos Dados e Modelagem

Pré-processamento:

- Remoção de registros sem alvo (Arrest).
- Engenharia temporal a partir de Date e remoção da data bruta.
- Separação treino/teste estratificada (80/20, random_state=42).
- Pipeline com tratamento de atributos numéricos e categóricos.

Modelos comparados:

- **Baseline:** DummyClassifier (classe majoritária)
- **Modelo principal:** Regressão Logística (com balanceamento de classe)

Split e atributos (celula 13)

```
Treino: (6227731, 13) | Teste: (1556933, 13)
Classe positiva (Arrest=1) - treino: 0.2614
Classe positiva (Arrest=1) - teste : 0.2614
Atributos categóricos: ['Primary Type', 'Description', 'Location Description', 'Domestic', 'Beat', 'District', 'Community Area']
Atributos numéricos : ['Year', 'Latitude', 'Longitude', 'Hour', 'Month', 'DayOfWeek']
```

Resultados: Comparação entre Modelos

Métricas exigidas no enunciado: precisão, revocação, F1-score (e acurácia como apoio).

Resultados da Regressão Logística (base completa)

- Acurácia: XX,XX%
- Precisão: XX,XX%
- Revocação: XX,XX%
- F1-score: XX,XX%

Interpretação: comparar com o baseline e destacar ganho em F1/revocação para a classe positiva.

Tabela de resultados (celula 17)

	modelo	accuracy	precisao	revocacao	f1 score
1	Regressao Logistica	0.8566	0.7179	0.7434	0.7304
0	Dummy (classe majoritaria)	0.7386	0.0000	0.0000	0.0000

<Figure size 900x450 with 0 Axes>

<Figure size 1000x500 with 1 Axes>

Grafico de metricas

Comparacao de metricas entre modelos



Relatório de Classificação e Matriz de Confusão

Análise por classe:

- Classe 0: ocorrências sem prisão
- Classe 1: ocorrências com prisão

O que discutir:

- Quantidade de falsos positivos e falsos negativos.
- Trade-off entre precisão e revocação para Arrest=1.
- Impacto do desbalanceamento na avaliação.

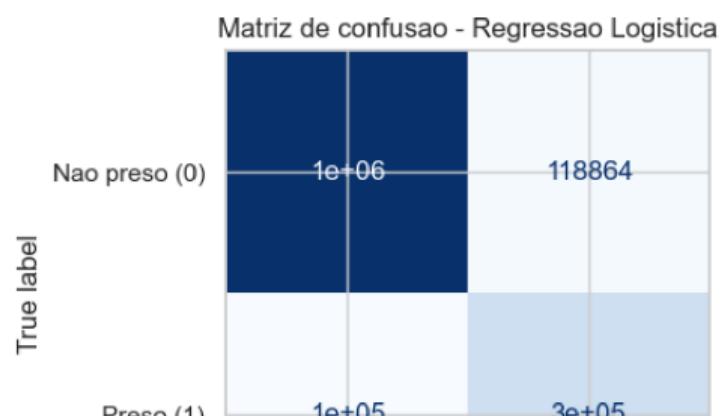
Relatorio de classificacao (celula 18)

```
Relatorio de classificacao - Regressao Logistica
precision    recall   f1-score   support
          0       0.9880    0.8966    0.9023    1149988
          1       0.7179    0.7434    0.7384    400953

accuracy                           0.8566
macro avg                           0.8130
weighted avg                         0.8583

<Figure size 500x400 with 1 Axes>
```

Matriz de confusao



Discussão Crítica e Limitações

Pontos fortes da abordagem:

- Pipeline reproduzível de pré-processamento + modelagem.
- Comparação com baseline simples.
- Atendimento às métricas exigidas no enunciado.

Limitações:

- Desbalanceamento da classe Arrest.
- Possível ruído/ausência em atributos (ex.: localização).
- Modelo linear pode não capturar relações não lineares complexas.

Melhorias futuras: tuning de hiperparâmetros, modelos baseados em árvore, balanceamento e validação cruzada.

Conclusão

- O trabalho implementa uma solução de **classificação** para prever Arrest em crimes de Chicago.
- A análise exploratória mostrou que variáveis de tipo, local e tempo são informativas.
- O modelo de Regressão Logística foi comparado a um baseline e avaliado com métricas adequadas.

Obrigado(a)!

Perguntas?