



APRENDIZADO DE MÁQUINA

22 de agosto de 2025

Relatório - Trabalho Final

Nomes: Azzure A. do Carmo, Caroline O. Costa e Ivan C. Vieira

Matrículas: 2023100851, 2023101189 e 2023100838

Resumo

O presente relatório visa identificar os principais fatores que corroboram para o perfil de indiciados nos crimes de homicídio simples, homicídio qualificado, estupro, estupro de vulnerável e lesões corporais grave, gravíssima e seguida de morte no município de São Paulo - SP. Os dados foram retirados de boletins de ocorrência dos anos de 2007 a 2014 e estavam disponíveis no Kaggle (Kaggle n.d.) pelo membro Marco Zachi e outros dois colaboradores (Inquisitivecrow n.d.).

Os dados incluem variáveis de perfilamento como sexo, idade, cor, grau de instrução e profissão e detalhes sobre o crime, como o status e os desdobramentos. Neste trabalho utilizou-se técnicas de árvore de decisão e mecanismos para penalizar dados faltantes e balancear o *dataset*.

1 Introdução

A análise de dados criminais tem se consolidado como uma ferramenta fundamental para a compreensão de dinâmicas sociais e para o apoio a políticas públicas de segurança. No Brasil, o município de São Paulo representa um cenário particularmente relevante devido à magnitude de sua população e à diversidade dos registros de ocorrência, que permitem investigar diferentes perfis de indiciados e padrões na criminalidade brutal.

Neste trabalho, utilizou-se os registros oficiais de boletins de ocorrência disponibilizados publicamente na plataforma digital Kaggle (Inquisitivecrow n.d.), com o intuito de estruturar um estudo quantitativo sobre fatores associados a crimes violentos. A escolha da fonte se deu pelas possibilidades de identificação de tendências e correlações entre variáveis que são relevantes. O *dataset* possui colunas referentes à profissão e grau de escolaridade, além de cor, idade e sexo, que são mais comuns. Outras variáveis presentes são o status do crime, que classifica se houve tentativa ou consumação, neste caso, somente existe tentativa nos casos de homicídio e o desdobramento, que explica porque o caso foi enquadrado com aquele crime.

O tratamento da base de dados foi uma etapa essencial, dado o elevado número de registros despadronizados não informativos. Para lidar com esse problema, foi aplicada uma limpeza e padronização, seguida de análises descritivas para um panorama inicial das variáveis disponíveis. Posteriormente, empregou-se técnicas de aprendizado de máquina, mais especificamente árvores de decisão combinadas aos métodos de regularização LASSO e Ridge, com o objetivo de mitigar os efeitos da ausência de informações e destacar atributos mais relevantes para a caracterização dos indiciados. Quanto ao desbalanceamento dentro das categorias de crime (homicídio, estupro e lesão corporal), utilizou-se XGBoost de três maneiras: sem pesos, com pesos e SMOTE.

Dessa forma, este relatório não apenas organiza os dados em uma perspectiva analítica, mas também busca demonstrar o potencial de métodos estatísticos e computacionais no estudo da criminalidade. A abordagem adotada pretende contribuir para a compreensão das condições que se associam à prática de crimes graves, ao mesmo tempo em que evidencia os

desafios de bases de dados extensas e heterogêneas.

2 Metodologia

Para o pré-processamento dos dados, realizou-se uma limpeza extensa. Os dados iniciais continham informações de todo o estado de São Paulo e, por isso, o primeiro passo foi identificar as variáveis que se referiam a cidade de São Paulo. Após isso, deixar somente os indiciados nos boletins e remover as colunas sem utilidade. Procurando melhorar as análises realizadas posteriormente, realizou o agrupamento de algumas variáveis, como na coluna de dados `COR`, agrupando as cores Preta e Parda. Além disso, criou-se uma nova variável, chamada `FAIXA_ETARIA`, com faixas de dez em dez anos, utilizando as informações das idades presentes em `IDADE_PESSOA`.

Usando o R, criou-se novos arquivos no formato *csv* somente com as informações desejadas e depois juntados, em ordem cronológica. O Resultado foi um dataset com 17876 observações e 9 variáveis, sendo elas: id do crime, rubrica, status da ocorrência, desdobramento do crime, sexo, idade, cor, profissão e grau de escolaridade.

Tabela 2.1: Categorias das variáveis selecionadas

Variável	Categorias
Rubrica	Homicídio simples (art. 121) Homicídio qualificado (art. 121, §2º) Estupro (art. 213) Estupro de vulnerável (art. 217-A) Lesão corporal grave (art. 129, §1º) Lesão corporal gravíssima (art. 129, §2º) Lesão corporal seguida de morte (art. 129, §3º)
Status	T (Tentativa) C (Consumação)
Sexo	M (Masculino) F (Feminino)
Cor	Preta/parda Branca Amarela Vermelha
Grau de Instrução	Analfabeto 1 Grau incompleto 1 Grau completo 2 Grau incompleto 2 Grau completo Superior incompleto Superior completo

A Tabela 2.1 mostra as categorias das variáveis rubrica, status, sexo, cor e grau de instrução. As outras variáveis, idade e profissão possuem diversas categorias. Para melhorar o funcionamento da árvore, as idades foram transformadas em faixas etárias e os crimes foram agrupados em três: homicídio, estupro e lesão corporal.

Neste trabalho, foram aplicadas técnicas de aprendizado de máquina para a classificação de registros criminais com base em variáveis categóricas. A metodologia adotada está dividida em quatro etapas principais: pré-processamento e engenharia de atributos, divisão entre conjuntos de treino e teste, modelagem e avaliação, e interpretabilidade do modelo.

Pré-processamento e Engenharia de Atributos

Antes do treinamento do modelo, os dados foram preparados para que pudessem ser interpretados pelo algoritmo de aprendizado de máquina. Primeiramente, foram identificadas as colunas categóricas, como `SEXO_PESSOA` e `COR`, que contêm informações não numéricas.

As variáveis foram codificadas para uso nos modelos. A variável-alvo `RUBRICA` foi transformada em valores numéricos únicos utilizando *LabelEncoder*, garantindo compatibilidade com algoritmos de aprendizado de máquina que trabalham apenas com dados numéricos. As variáveis de entrada foram codificadas por *one-hot encoding*, gerando colunas binárias para cada categoria e evitando interpretações equivocadas de ordem entre categorias.

Divisão entre Treino e Teste

Os dados foram divididos em conjuntos de treino e teste, seguindo a proporção de 80% para treino e 20% para teste. A divisão estratificada (`stratify=y_encoded`) manteve a proporção de cada tipo de crime nos dois conjuntos, garantindo que o modelo fosse treinado e avaliado de forma representativa. Essa separação permite verificar se o modelo generaliza bem para novos dados, evitando *overfitting*.

Modelo e Avaliação

Para a tarefa de classificação, foi utilizado o algoritmo *XGBoost* (`XGBClassifier`), um método de *boosting* de gradiente que constrói sequencialmente árvores de decisão, em que cada nova árvore busca corrigir os erros das anteriores. Os parâmetros de regularização `reg_alpha` (L1) e `reg_lambda` (L2) foram utilizados para reduzir o *overfitting*, aplicando métodos de *Hyperparameter Tuning* para encontrar os valores dos parâmetros de regularização mais adequados aos modelos analisados.

O desempenho do modelo foi avaliado por meio do `classification_report`, que apresenta métricas de precisão, *recall* e *f1-score* para cada categoria de crime, permitindo uma análise detalhada da acurácia do modelo, além de matrizes de confusão para representação gráfica.

Interpretabilidade do Modelo

Para compreender o comportamento do modelo, foi analisada a importância dos atributos utilizando a função `feature_importances` do *XGBoost*, identificando quais variáveis tiveram maior influência na classificação.

3 Resultados

Estatísticas Descritivas

Inicialmente, para melhor entendimento do problema escolhido, foram realizadas análises descritivas das variáveis selecionadas, os resultados abaixo apresentam essas variáveis após a limpeza e preparação dos dados.

Profissão do Criminoso por Cada Crime Cometido

A variável profissão apresenta 10.532 dados faltantes, o que pode prejudicar os modelos realizados posteriormente. Além dos NAs, os criminosos apresentaram 276 profissões. Para entendermos sobre o perfil dessas pessoas, ranqueamos para cada crime quais as 10 profissões mais recorrentes. Espera-se que essas sejam as profissões que apareçam como resultado do modelo usado na realização do trabalho.

Pela Figura 3.1 é notável a discrepância das profissões ajudante e de pessoas desempregadas para o resto das profissões, o que pode significar um perfil mais definido para os autores de homicídios simples, que é a acusação de matar alguém sem nenhuma outra qualificadora. O assassino recebe pena abstrata de 6 a 20 anos de reclusão.

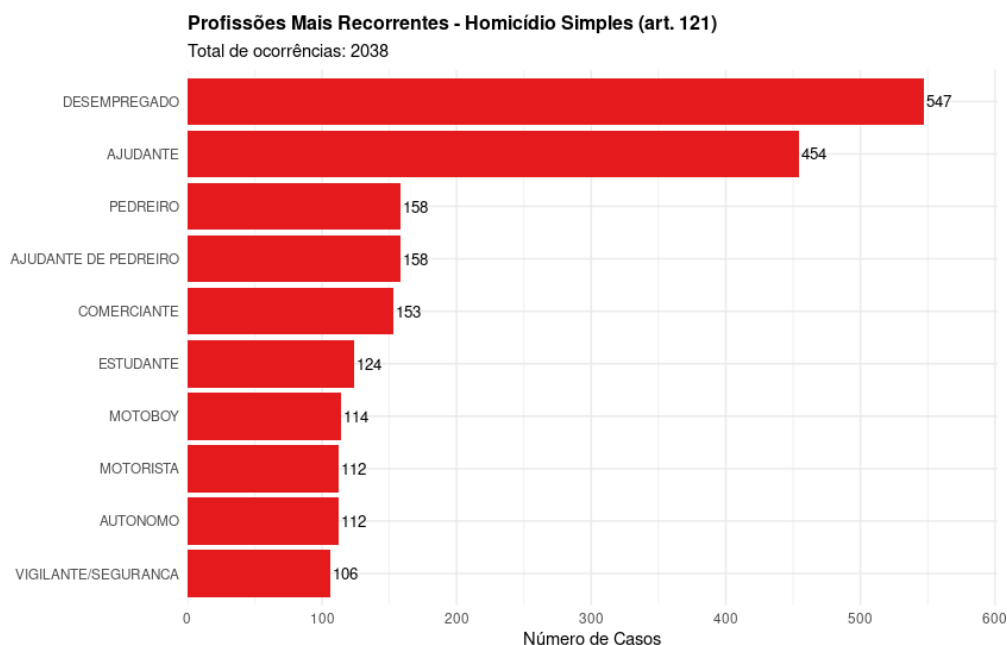


Figura 3.1: Profissões Mais Recorrentes - Homicídio Simples.

Já no homicídio qualificado, a pena varia de 12 a 30 anos e é definido como o ato de matar alguém agravado por algum outro motivo que o juiz julgue ser mais grave que apenas o ato, como no caso de feminicídio ou de matar um menor de 14 anos ou por motivo fútil.

Nesse caso, como é possível ver na Figura 3.2 as profissões mais recorrentes para o crime de homicídio qualificado são desempregado e ajudante, novamente.

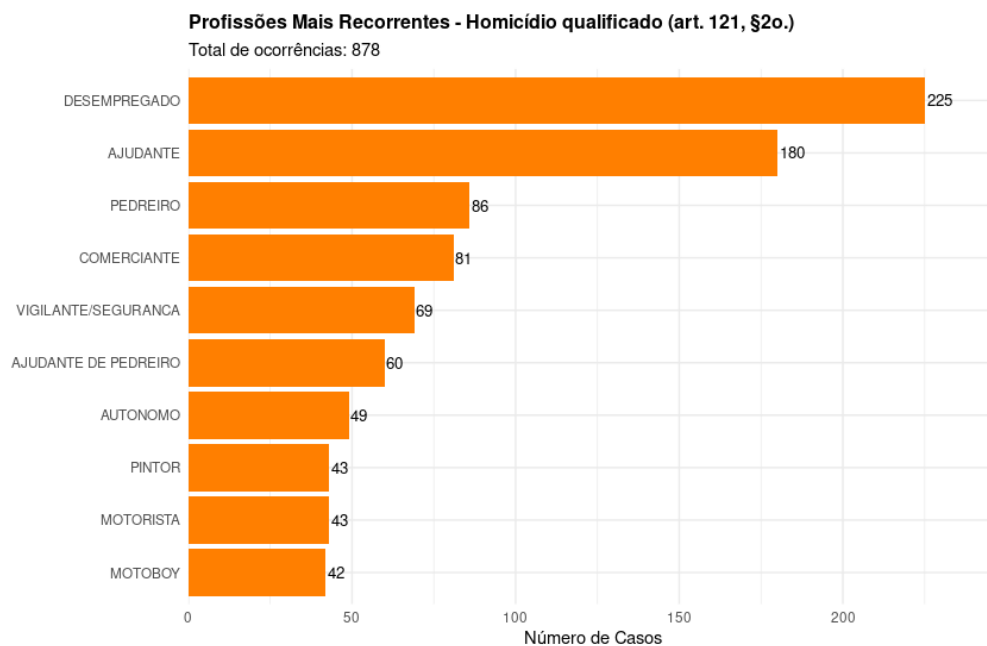


Figura 3.2: Profissões Mais Recorrentes - Homicídio Qualificado.

Para elaboração do modelo usado, aglomeramos os crimes, como foi dito anteriormente, porém é interessante observar separadamente cada crime. No caso de estupro, temos o estupro comum e o estupro de vulnerável.

O estupro comum se configura quando há violência ou grave ameaça para obrigar a vítima a praticar ou permitir um ato sexual. Já no estupro de vulnerável, a violência é presumida por lei, pois a vítima não tem capacidade de consentir com o ato. Isso ocorre quando a pessoa é menor de 14 anos, está inconsciente, ou não tem o necessário discernimento para o ato sexual devido a uma enfermidade ou deficiência mental. No caso do estupro de vulnerável, o crime se configura independentemente de a vítima ter consentido, ter experiência sexual anterior ou até mesmo ter um relacionamento com o agressor.

As Figuras 3.3 e 3.4 apresentam, assim como os demais, desempregado e ajudante sendo as profissões mais comuns, porém, no caso do estupro de vulnerável, a ordem é invertida e a distância para o 3 lugar, que é pedreiro, é reduzida.

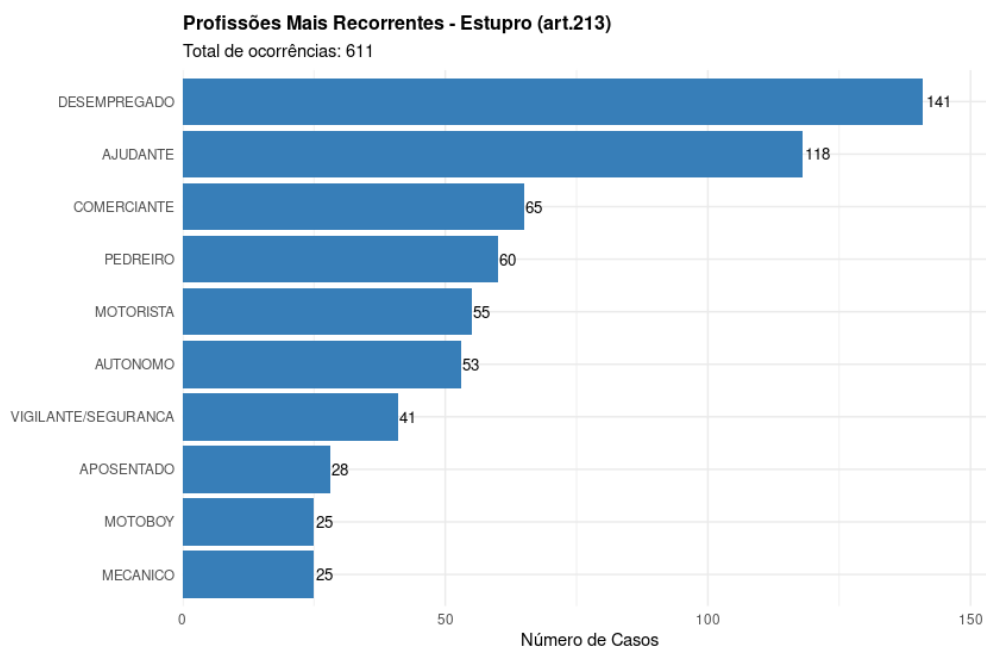


Figura 3.3: Profissões Mais Recorrentes - Estupro.

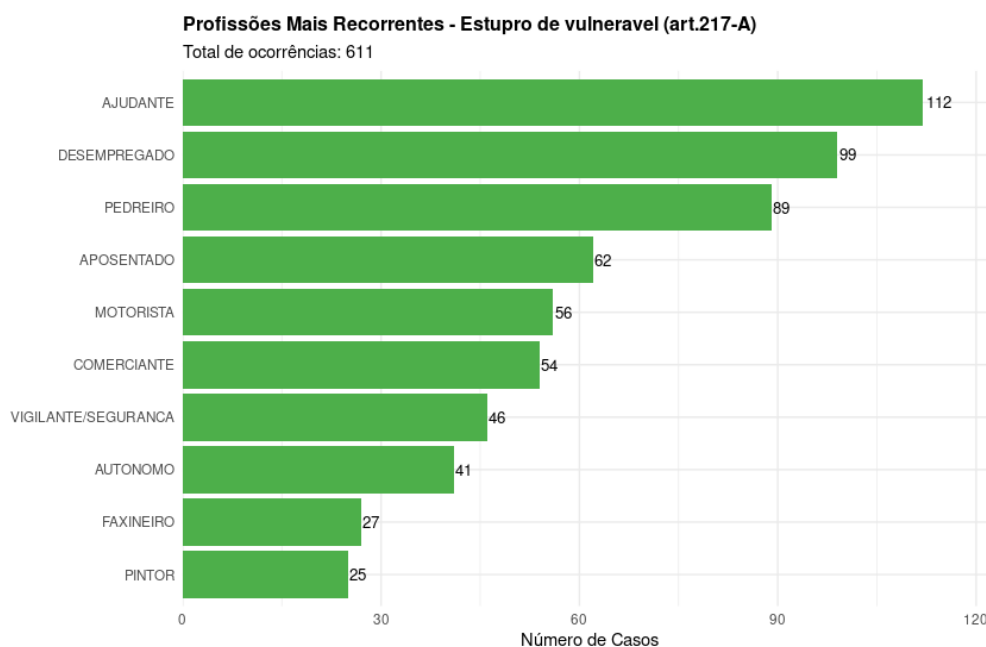


Figura 3.4: Profissões Mais Recorrentes - Estupro de Vulnerável.

A lesão corporal (Art. 129 do Código Penal) é um crime que se caracteriza pela ofensa à integridade física ou à saúde de outra pessoa. A gravidade do crime varia de acordo com a lesão:

- Lesão corporal grave: Ocorre quando a agressão resulta em incapacidade para o trabalho por mais de 30 dias, perigo de vida, debilidade permanente de membro, sentido ou função, ou aceleração de parto.

- Lesão corporal gravíssima: É a forma mais severa, que causa incapacidade permanente para o trabalho, enfermidade incurável, perda ou inutilização de membro, sentido ou função, deformidade permanente, ou aborto.

Devido a quantidade de dados presente, optamos pela junção de todos os tipos de lesões, para esse crime, como pode ser visto na Figura 3.5 as profissões mais recorrentes, assim como nos demais casos, são desempregado e ajudante.



Figura 3.5: Profissões Mais Recorrentes - Lesão Corporal.

Idade por Sexo do Criminoso

Outra variável presente na base e que supomos representar parte importante do perfil dos criminosos é o sexo. A primeira informação relevante são as quantidades de ocorrências de cada um, tendo 15.919 pessoas de sexo masculino e 678 do sexo feminino, além da variável apresentar 1.279 dados NA.

Quando analisamos a distribuição da idade por sexo, Figura 3.6 temos uma distribuição assimétrica a esquerda, onde se concentram os dados. A maior parte, tanto dos homens quanto das mulheres tem entre 20 a 35 anos e no caso das mulheres.

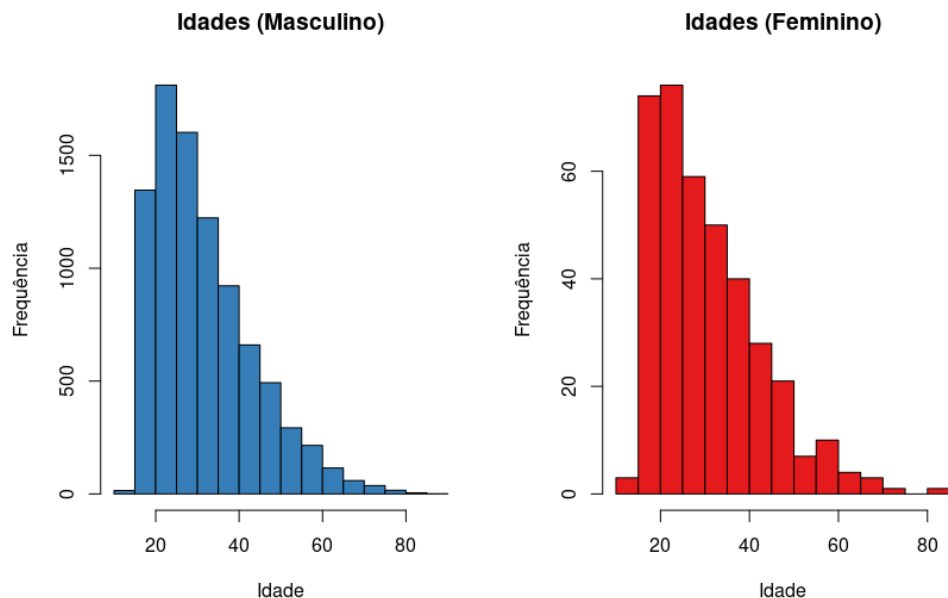


Figura 3.6: Distribuição das Idades por Sexo.

Distribuição das Idades por Cada Crime Cometido

Outra variável que pode apresentar alguma relevância na definição do perfil do criminoso é a idade, pensando nisso foi analisada a distribuição dessa variável par cada um dos crimes cometidos. Importante considerar a existência de 7.905 NAs nessa variável.

A Tabela 3.1 apresenta um resumo descritivos das distribuições das idades. Nota-se que a média de idade mais elevada são responsáveis por praticar o crime de estupro de vulnerável, com uma média de 37,9 anos e um mínimo de 14 anos.

Informações importantes e preocupantes é a idade mínima trazida na tabela, na maior parte dos crimes as pessoas mais novas a praticá-los são menores de idade, o único tipo de crime não cometido por menores foi lesão corporal.

De acordo com a legislação brasileira, especificamente as leis federais que regem os direitos da criança e do adolescente (Brasil 1990) no Brasil, quando um menor comete um crime, ele não é julgado como maior, o objetivo é utilizar de medidas socioeducativas para a ressocialização do menor, e as medidas são definidas com base na gravidade do crime. Em casos mais graves, como os mencionados no trabalho, o menor é internado por um período máximo de 3 anos, outras medidas socioeducativas são advertência, obrigação de reparar o dano, prestação de serviços à comunidade, liberdade assistida ou regime de semiliberdade.

Tabela 3.1: Estatísticas de Idade por Tipo de Crime.

RUBRICA	Média	Mediana	Mínimo	Máximo	D.V.
Homicídio simples	30.4	28	12	83	11.1
Homicídio qualificado	31.7	29	12	83	11.4
Estupro	32.9	30	14	83	11.0
Estupro de vulnerável	37.9	36	14	86	13.9
Lesão corporal GRAVE	31.5	30	15	60	10.5
Lesão corporal GRAVÍSSIMA	32.6	33	18	52	10.6
Lesão corporal seguida de morte	27.9	23	18	48	9.3

Quando olhamos para os crimes de homicídios, Figuras 3.7, notamos que em média, a idade é maior no crime de homicídio qualificado, 31,7 anos e como é apresentado na Tabela 3.1 os máximos e mínimos são os mesmos (12 e 83). O crime de homicídio é o que apresenta uma idade mínima registrada mais baixa, o que pode é preocupante e pode ser estudado mais a fundo futuramente.

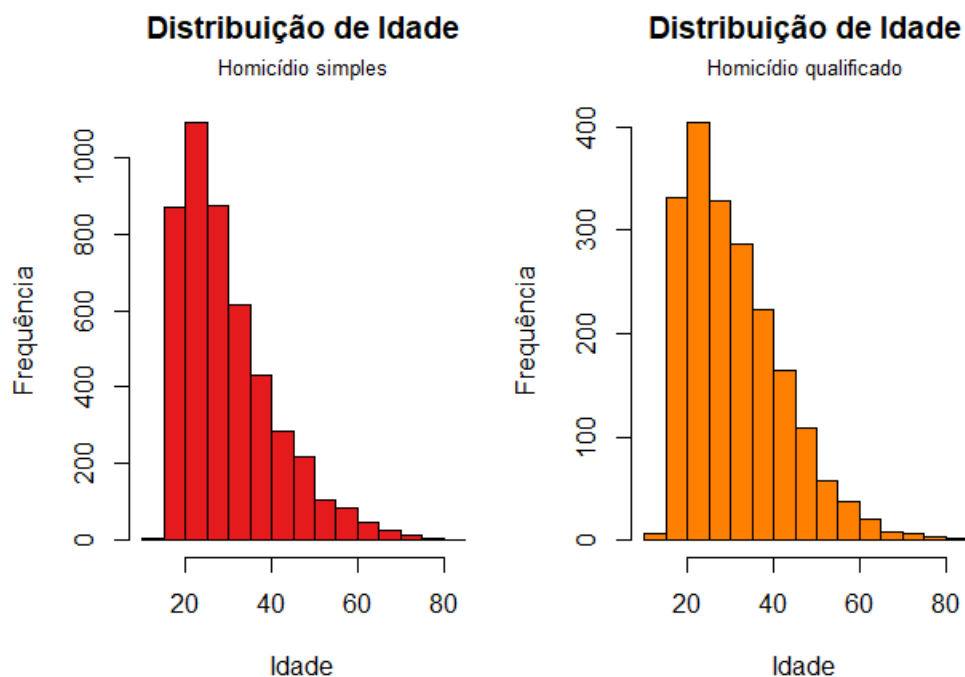


Figura 3.7: Distribuições das Idades por Crimes de Homicídio.

Diferente do caso de homicídio, no crime de estupro as distribuições apresentam maiores divergências. O estupro simples segue uma distribuição similar a dos crimes de homicídio, com uma média de 32,9 anos e desvio padrão de 11. Porém, o mesmo não ocorre como estupro de vulnerável, nota-se um deslocamento da distribuição para a direita, indicando pessoas mais velhas cometendo esse crime, isso é refletido na média, os criminosos que cometem esse crime tem em média 37.9 anos e um desvio padrão de 13.9. A pessoa mais velha registrada tinha 86 anos, e a mais nova 14.

Uma informação importante é sobre a idade de consentimento no Brasil, segundo a legislação brasileira (Jusbrasil [Ano]), com base no Código Penal e no Estatuto da Criança e do Adolescente (ECA), estabelece a idade de consentimento em 14 anos, portanto, qualquer ato sexual praticado com uma pessoa menor de 14 anos é considerado estupro de vulnerável (Art. 217-A do Código Penal), a partir dos 14 anos, a pessoa é considerada capaz de consentir. No entanto, o consentimento deve ser livre e espontâneo.

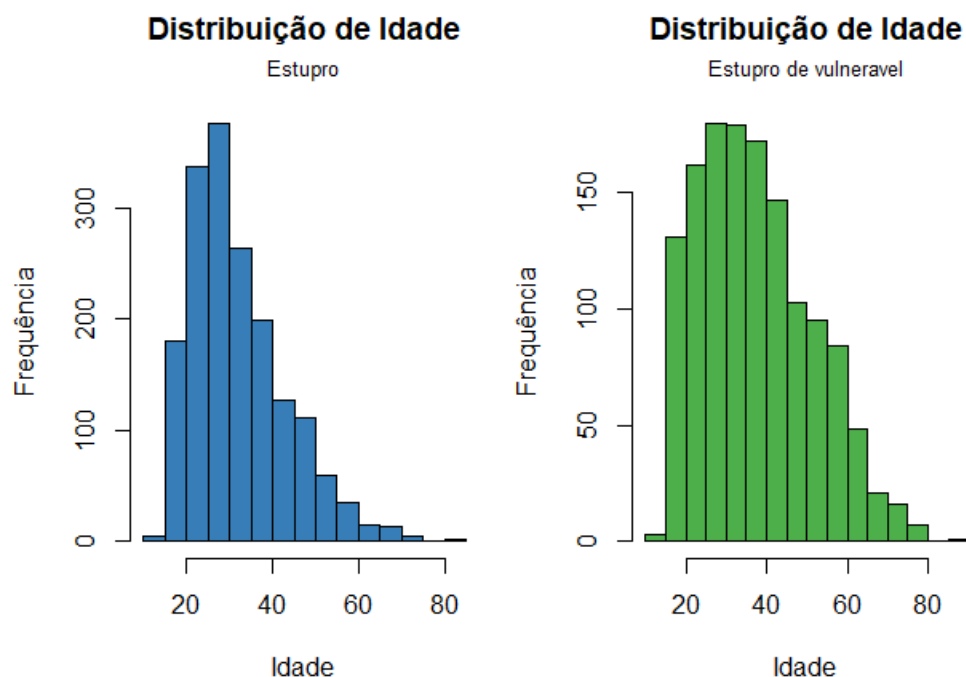


Figura 3.8: Distribuições das Idades por Crimes de Estupro.

Quando observamos a Figura 3.9 notamos que ocorre uma concentração maior em idades até 40 anos. O gráfico mostra a junção de todas as lesões corporais, porém, olhando de forma separada, como apresentada na Tabela 3.1, a menor média registrada é de lesão corporal seguida de morte, com uma média de 27,9 anos, mas quando olhamos para lesão corporal gravíssima, a média aumenta para 32,6 anos. Outro apontamento curioso é a idade máxima registrada, em ambos os outros crimes, a idade máxima é acima de 80 anos, enquanto no crime de lesão corporal não passa de 60.

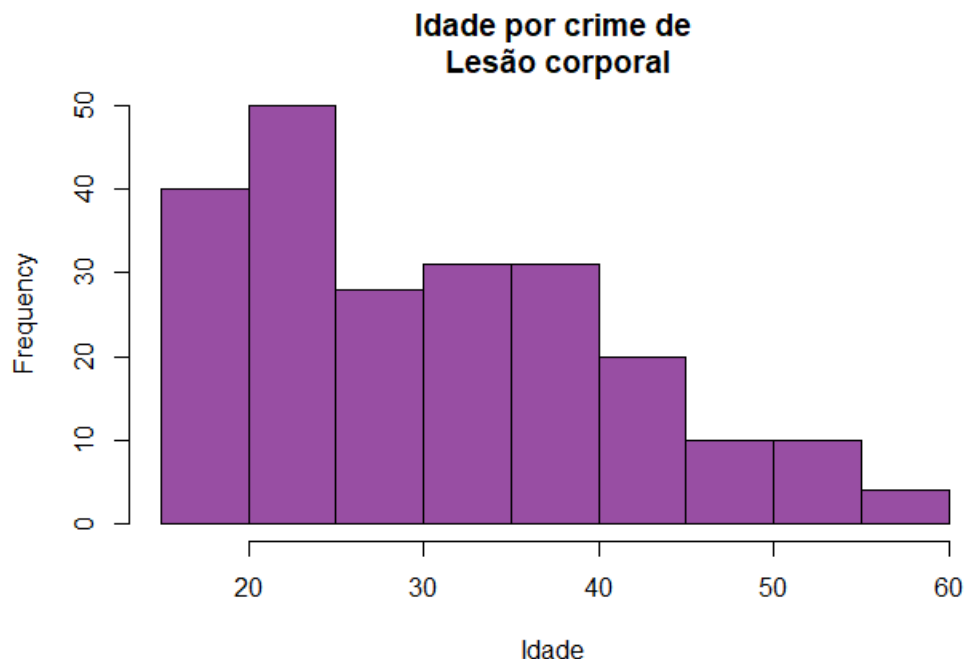


Figura 3.9: Distribuições das Idades por Crimes de Lesão Corporal.

Tentativa e Consumação

Um crime é consumado quando todas as etapas de sua execução foram concluídas, ou seja, o resultado desejado pelo criminoso foi alcançado. No caso do homicídio, a consumação ocorre quando a morte da vítima se concretiza. Já a tentativa ocorre quando o agente inicia a execução do crime, mas ele não se consuma por circunstâncias alheias à sua vontade. No homicídio, isso significa que a pessoa agiu com a intenção de matar, mas o resultado (a morte) não aconteceu, seja porque a vítima sobreviveu, a polícia interveio ou por qualquer outro motivo.

Análise de Tentativa e Consumação em Homicídios O conceito de tentativa e consumação são fundamentais no direito penal. Um crime é consumado quando todas as etapas de sua execução foram concluídas, ou seja, o resultado desejado pelo criminoso foi alcançado. No caso do homicídio, a consumação ocorre quando a morte da vítima se concretiza.

Já a tentativa ocorre quando o agente inicia a execução do crime, mas ele não se consuma por circunstâncias alheias à sua vontade. No homicídio, isso significa que a pessoa agiu com a intenção de matar, mas o resultado (a morte) não aconteceu, seja porque a vítima sobreviveu, a polícia interveio ou por qualquer outro motivo.

A Figura 3.10 apresenta a proporção entre tentativas e consumações em ambos os crimes de homicídio. No Homicídio Qualificado (Art. 121, §2º), os dados mostram que 63,1% dos casos foram consumados, enquanto 36,9% foram tentativas. Já no Homicídio Simples (Art. 121), ocorre similar ao homicídio qualificado, porém a diferença é ainda mais notável: 70,8% dos casos foram consumados e apenas 29,2% tentados. Isso indica que, visualmente, em ambos os tipos de homicídio, a consumação é mais comum do que a tentativa.

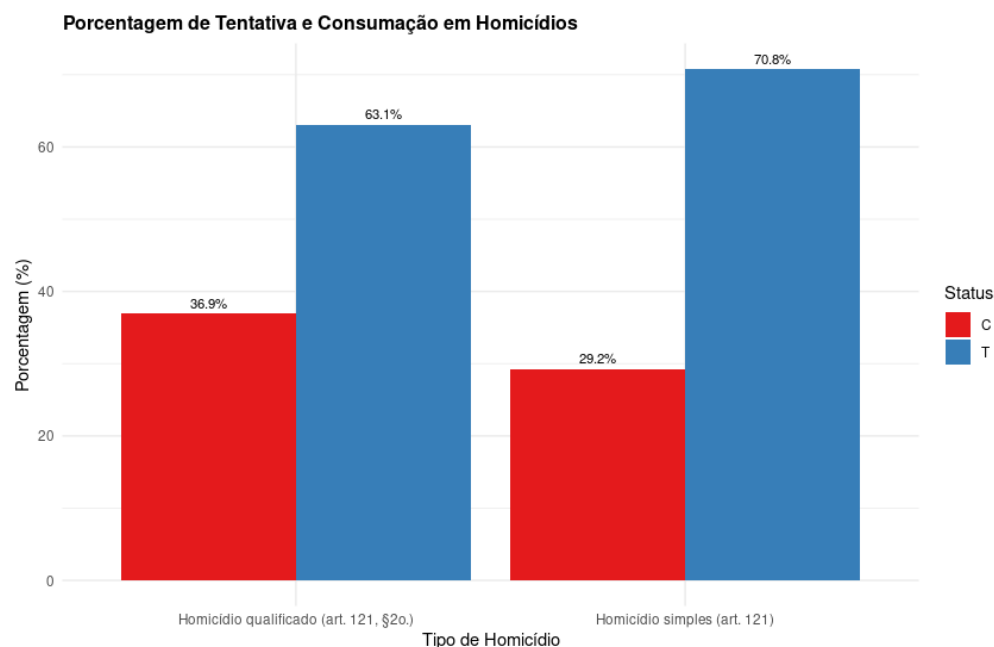


Figura 3.10: Porcentagem de Tentativas e Consumações nos Casos de Homicídio.

Grau de Escolaridade por Cada Tipo de Crime Cometido

Outra variável que acredita-se explicar o perfil dos criminosos é a escolaridade, porém, essa variável também apresentou muitos NAs, (10.020) além de informações inconsistentes. Ao longo do processo de limpeza, foi imputado alguns dados para preservar o tamanho amostral e melhorar a qualidade dos resultados.

Na seção atual os gráficos estão aglomerando as subcategorias de cada crime. Na Figura 3.11 é apresentado o que em disparada com 3.057 ocorrências, os criminosos possuem apenas o 1 Grau completo, o que indica um baixo nível de escolaridade.

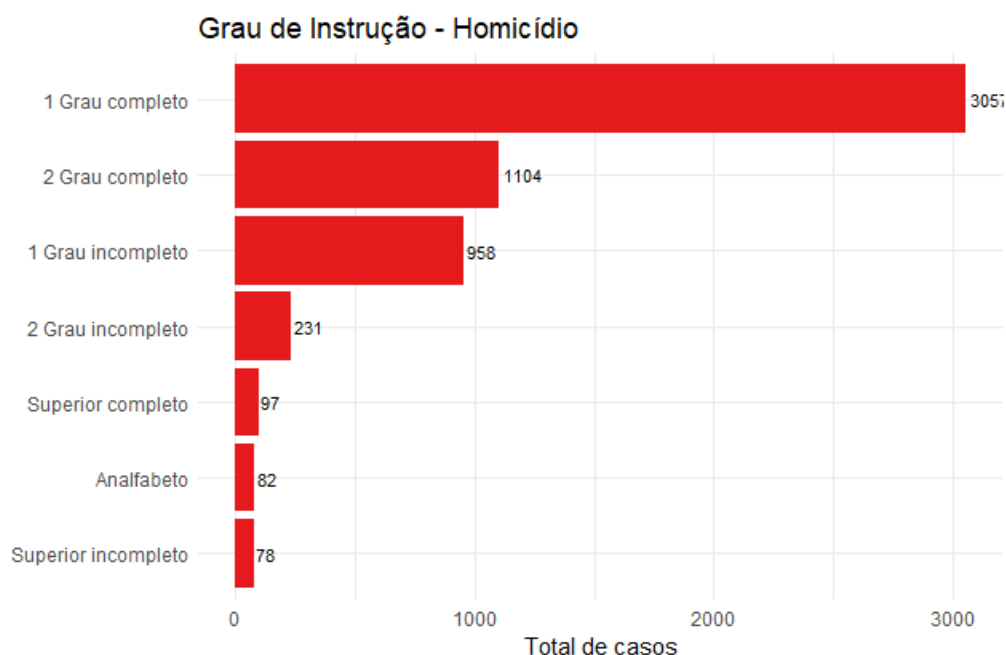


Figura 3.11: Grau de Instrução por Crime de Homicídio.

No caso dos crimes de estupro, como visto na Figura 3.12 a diferença entre os primeiros lugares é menor que para os crimes de homicídio, entretanto os criminosos também apresentam baixo nível de escolaridade, tendo a maior parte no máximo o segundo grau completo (419) muitos nem o primeiro completo (560).

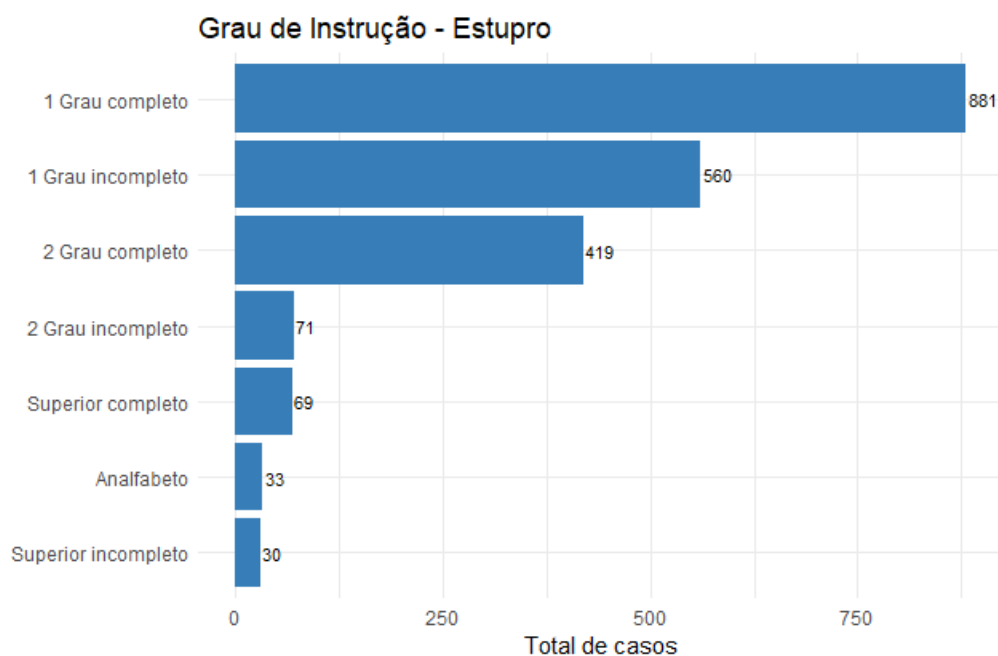


Figura 3.12: Grau de Instrução por Crime de Estupro.

Para o caso de lesão corporal, o gráfico se mostra visivelmente similar ao gráfico de homicídio, como pode-se observar na Figura 3.13, onde o grau de instrução mais frequente é apenas o 1 grau completo (76) e a diferença para os demais é consideravelmente elevada.

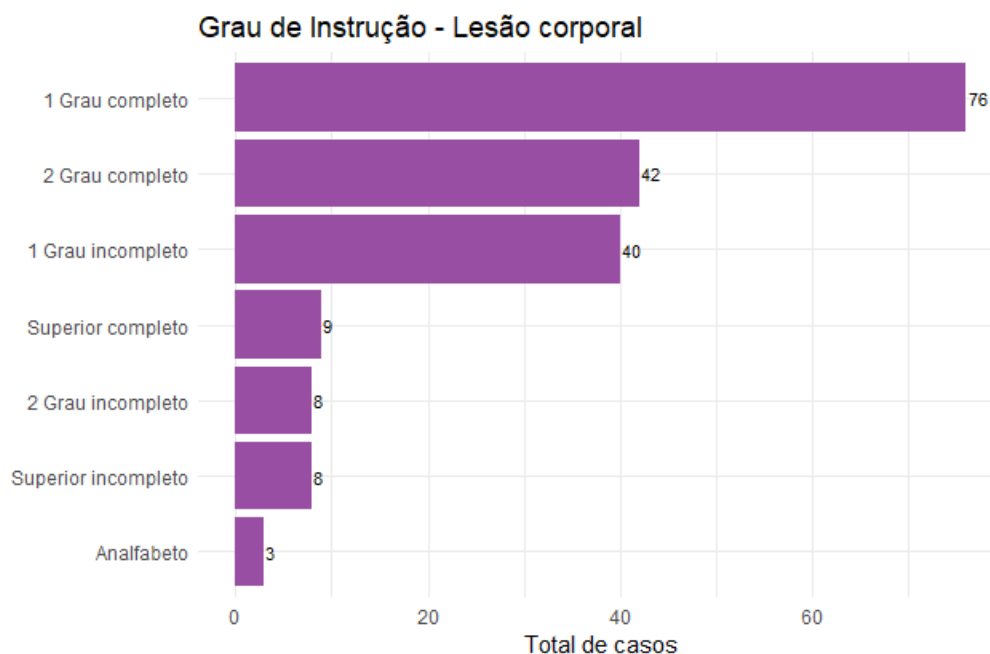


Figura 3.13: Grau de Instrução por Crime de Lesão Corporal.

Distribuição de Cor por Cada Tipo de Crime Cometido

Quando falamos sobre cor, é preciso considerar é que algo alto declarada e isso pode influenciar na veracidade dos resultados. Além disso, para melhorar a visualização, foi realizada a junção das cores preta e parda. Para todos os três tipos de crimes, as únicas cores que apresentaram frequências consideráveis foram branca e preta/parda. Para essa variável, a base apresentou 3.062 informações faltantes.

Para os crimes de homicídio, Figura 3.14, a cor mais frequente foi preto/pardo com 54,9% seguido de branca com 45%.

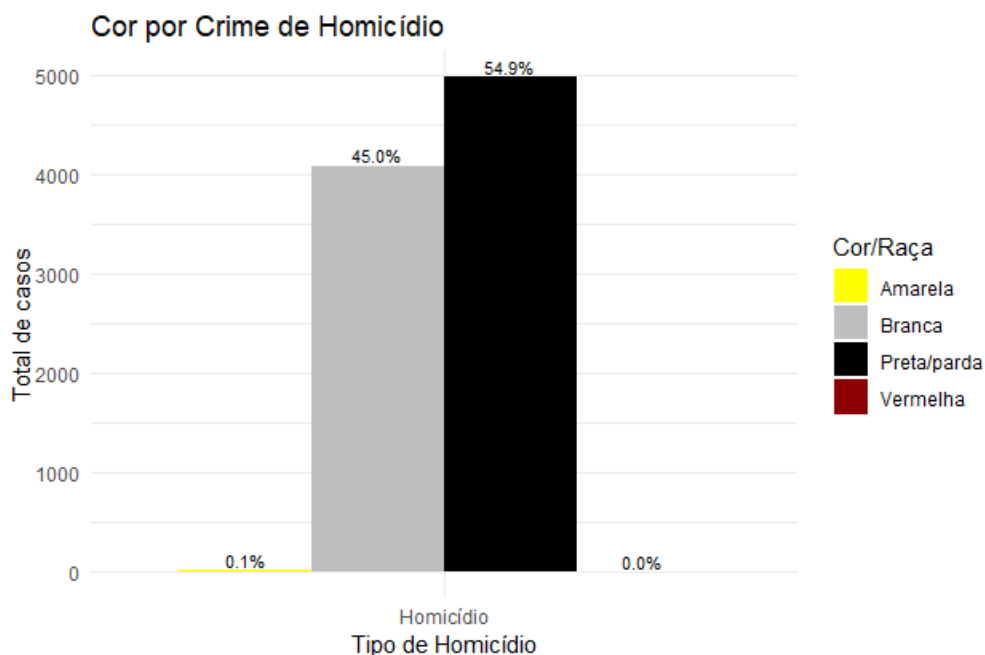


Figura 3.14: Cor por Crime de Homicídio.

No caso do crime de estupro, Figura 3.15, ocorreu algo similar, a proporção de preto.pardo ficou em 54,7 e de brancos 44,8%.

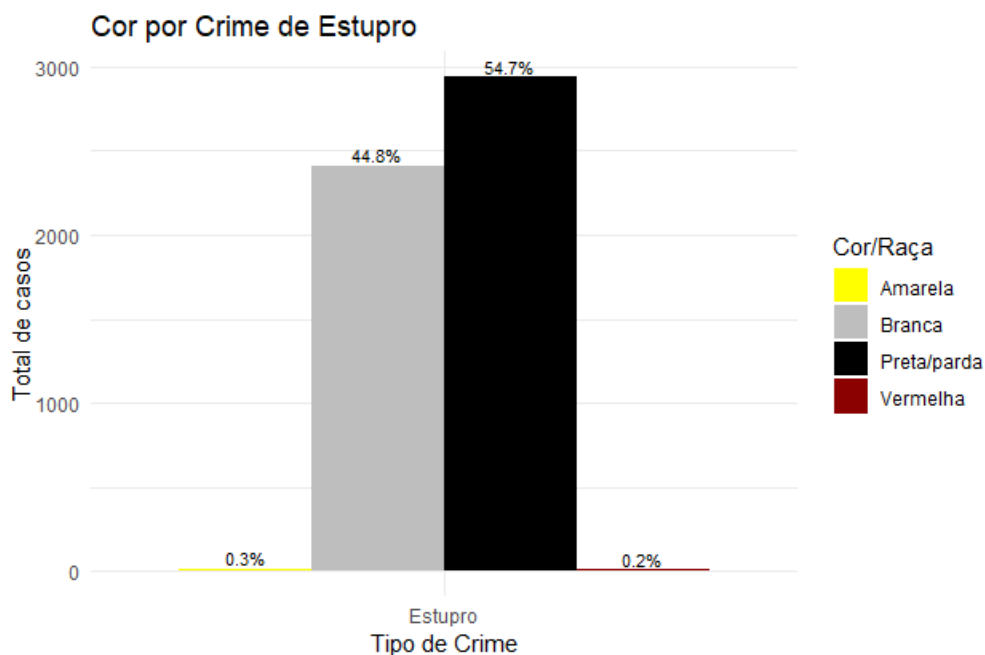


Figura 3.15: Cor por Crime de Estupro.

Acerca dos crimes de lesão corporal, é importante lembrar que ele apresenta menor número de casos e por conta disso, os resultados tendem a ter menor precisão. A Figura 3.16 mostrou que a frequência de criminosos brancos e pretos para esse crime foi a mesma (49,6%).

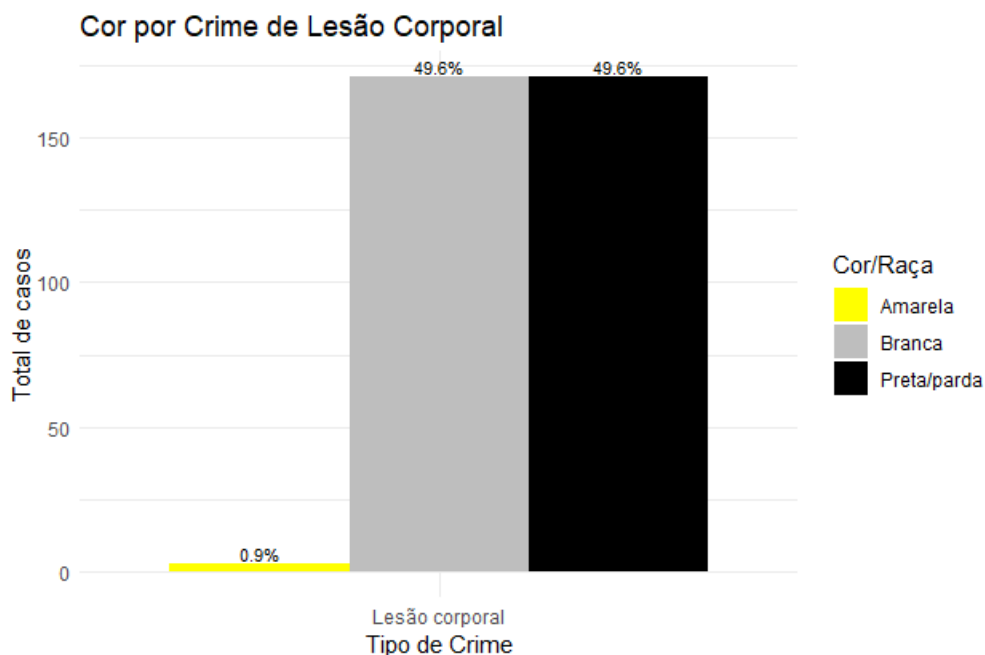


Figura 3.16: Cor por Crime de Lesão Corporal.

Modelo *XGBoost*

Foram realizadas algumas modelagens com o *XGBoost*, algoritmo que combina vários modelos de aprendizado mais simples e fracos para criar um modelo final muito mais forte e preciso. Esse algoritmo foi escolhido pela sua capacidade de lidar com dados faltantes de maneira satisfatória quando comparado com outros algoritmos relacionados à árvores de decisão. Os modelos gerados trabalharam com diferentes dados do *dataset* descrito anteriormente, sendo eles: os dados agrupados dos tipos de crimes (homicídio, estupro e lesão corporal) e mais uma modelagem para cada tipo de crime para classificar suas subdivisões específicas. Juntamente ao modelo *XGBoost*, realizou-se também métodos de balanceamento de classes, como o SMOTE e a adição de pesos às classes com o intuito de procurar equilibrar os dados e aperfeiçoar a análise realizada pelo modelo.

Os dados, nesta etapa, foram preparados novamente para que pudessem passar pelo processo da modelagem. Inicialmente, os dados categóricos, referentes às colunas de sexo, cor, profissão, grau de instrução e faixa etária de cada indiciado, além do crime cometido, foram codificados para valores numéricos, em vista de que o modelo *XGBoost* trabalha apenas com dados numéricos. Os dados então foram separados entre conjuntos de treino (80%) e de teste (20%), mantendo uma proporção adequada entre cada tipo de crime em todos os modelos aplicados.

Para prosseguir com o processo de aprendizado de máquina, realizou-se um cálculo de "pesos" para cada classe de crime, visando um melhor equilíbrio entre os crimes com menor quantidade de registros, como os crimes de lesão corporal, que receberam maior peso, e os crimes com maior número de registros, como os crimes de homicídio, que receberam, por sua vez, menor peso. Em outra instância, foi realizado um balanceamento com a técnica

SMOTE, com a intenção de comparar os resultados obtidos por cada um dos métodos e pela aplicação do algoritmo sem ajustes de balanceamento de classes. Além disso, para melhor otimização do modelo, utilizou-se de algoritmos de ajuste de parâmetros, como o *RandomizedSearchCV*, para encontrar melhores combinações de parâmetros para os modelos *XGBoost*.

Após a aplicação de todos esses algoritmos e técnicas, realizou-se então os treinamentos e testes de cada um dos modelos. Os resultados seguem na seção adiante. Por conta da limitação de número de dados de casos de lesão corporal e grave desbalanceamento de dados, não houve a aplicação de modelos para análise de suas subdivisões pela sua ineficácia.

Para melhor visualização dos resultados, eles serão apresentados em formatos de tabelas e matrizes de confusão.

Modelo para todos os crimes agrupados

No caso dos crimes agrupados, foram apresentadas as seguintes distribuições de classes (número de ocorrências em cada classe), pesos e redistribuições via SMOTE (Tabela 3.2):

Tabela 3.2: Distribuição de Classes, SMOTE e Pesos - Crimes Agrupados.

	Classes	SMOTE	Pesos
Homicídio	11581	9264	0.5145
Estupro	5868	9264	1.0155
Lesão Corporal	427	9264	13.9376

A aplicação do algoritmo com e sem os métodos de balanceamento de classes citados anteriormente levou aos resultados que podem ser observados nas Figuras 3.18 à 3.20 e Tabelas 3.3 à 3.5. Além disso, foram identificadas as variáveis mais explicativas para o modelo *XGBoost*, como pode ser visto na Figura 3.17. O perfil geral dos indiciados é dominado por gênero e cor, seguido por idade, educação e profissão. Isso sugere que fatores demográficos básico são os principais discriminadores criminais de forma geral, com características socioeconômicas desempenhando um papel secundário, porém relevante.

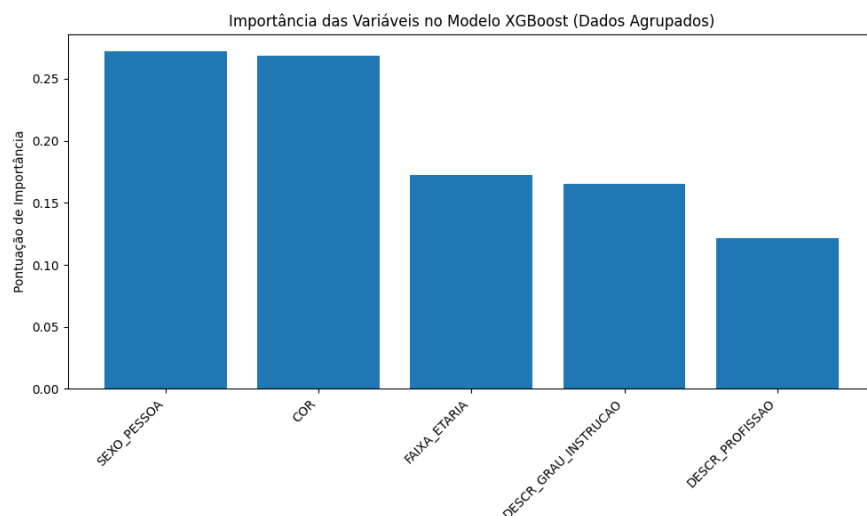


Figura 3.17: Variáveis Explicativas para o Modelo com Peso - Dados Agrupados.

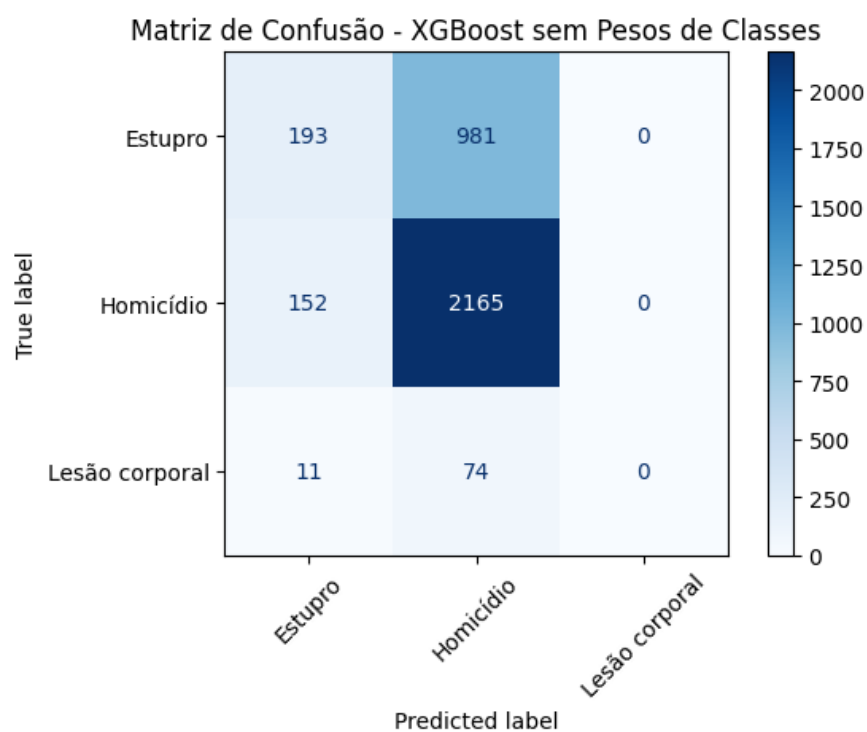


Figura 3.18: Matriz de Confusão - sem Pesos.

Tabela 3.3: Resultados sem Peso de Classes.

Classe	Precision	Recall	F1-Score
Estupro	0.55	0.20	0.29
Homicídio	0.68	0.92	0.78
Lesão corporal	0.50	0.01	0.02
Acurácia	0.66		

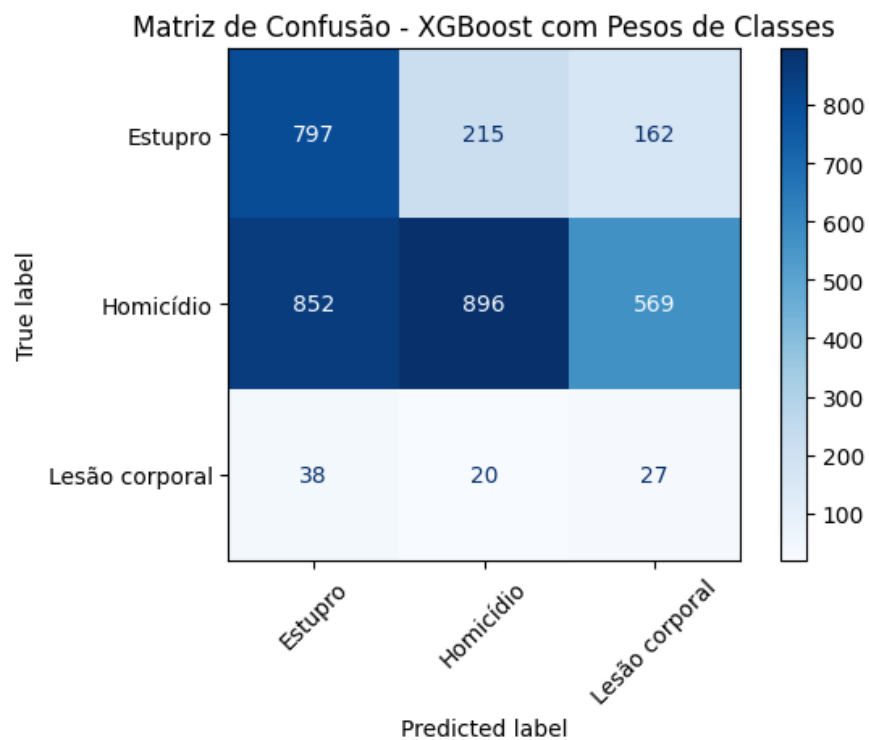


Figura 3.19: Matriz de Confusão - com Pesos.

Tabela 3.4: Resultados com Peso de Classes.

Classe	Precision	Recall	F1-Score
Estupro	0.47	0.68	0.56
Homicídio	0.79	0.39	0.52
Lesão corporal	0.04	0.32	0.06
Acurácia	0.48		

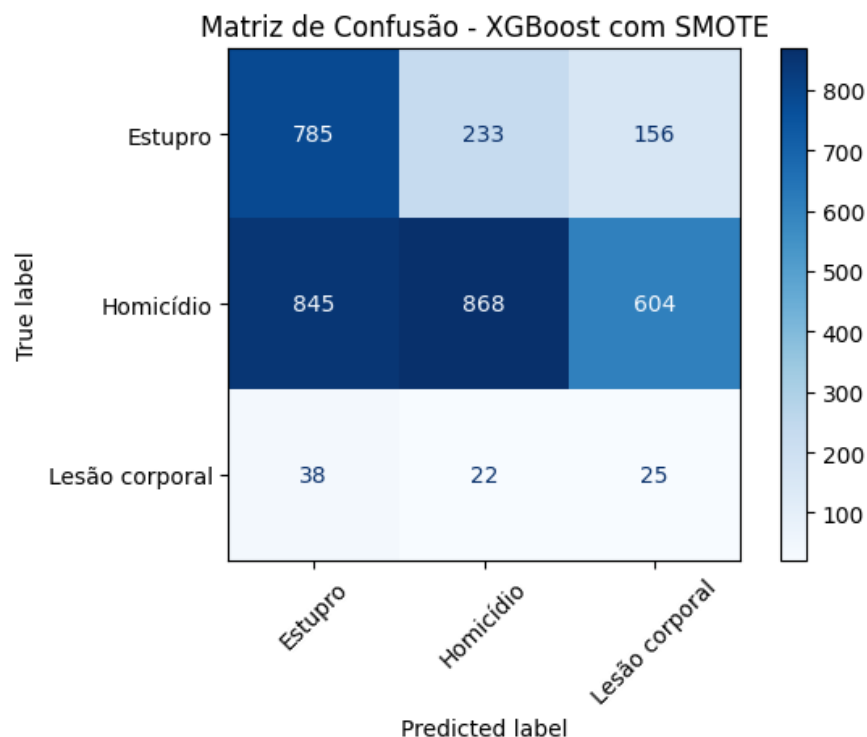


Figura 3.20: Matriz de Confusão - SMOTE sem Pesos.

Tabela 3.5: Resultados com SMOTE.

Classe	Precision	Recall	F1-Score
Estupro	0.47	0.67	0.55
Homicídio	0.77	0.37	0.50
Lesão corporal	0.03	0.29	0.06
Acurácia	0.47		

A análise dos resultados revela problemas consistentes em todos os modelos testados. O conjunto de dados apresenta um desbalanceamento grave, onde a classe de crime de homicídio representa a grande maioria das amostras, enquanto lesão corporal possui pouquíssimos exemplos. Este desequilíbrio impacta diretamente o desempenho de todos os modelos. Sem nenhum tipo de balanceamento, é possível observar que o modelo desenvolveu um forte viés pela classe majoritária, apresentando boa performance para homicídio mas falhando completamente nas classes minoritárias, especialmente em lesão corporal que demonstrou-se indetectável para o modelo, como pode ser observado na matriz de confusão pela concentração de predições na classe de homicídio, estando de acordo com o esperado pelo funcionamento dos algoritmos de aprendizado de máquina.

A aplicação de métodos de balanceamento trouxe melhorias, mas não resolveu os problemas. Tanto o uso de pesos de classes quanto a técnica SMOTE conseguiram melhorar o reconhecimento da classe de crimes de estupro de maneira significativa, porém às custas de uma redução no desempenho da classe de homicídio. A classe de lesão corporal continuou

com performance muito baixa em todas as abordagens, independentemente da técnica utilizada, sugerindo que o problema não reside apenas no desbalanceamento das classes, mas provavelmente também é inerente às características inadequadas e quantidade insuficiente de dados para esta classe específica. Portanto, mesmo após as técnicas de correção, foi observado que todos os modelos mantiveram dificuldade em classificar adequadamente as categorias.

As principais conclusões indicam que os métodos de balanceamento cumpriram seu papel de redistribuir de alguma forma o poder preditivo do modelo de maneira mais "equilibrada" entre as classes, porém, realizaram esse papel às custas da qualidade de classificação da classe majoritária, não resolvendo de maneira satisfatória os problemas causados pelo desbalanceamento inicial. A classe de crimes de lesão corporal, ainda por cima, permanece como um problema particular que requer abordagens que vão além de balanceamento de classes.

Modelo para crimes de homicídio

Partindo para o caso dos crimes de homicídio simples e homicídio qualificado, foram apresentadas as seguintes distribuições de classes, pesos e redistribuições via SMOTE (Tabela 3.6).

Notavelmente, ao analisarmos as variáveis mais explicativas para homicídios (Figura 3.21), o grau de instrução torna-se a variável mais importante, superando até mesmo gênero e cor. Isso pode indicar que homicídios estão mais fortemente associados a questões educacionais e de capital humano, com a profissão também ganhando maior peso relativo em comparação com a análise geral.

Tabela 3.6: Distribuição de Classes, SMOTE e Pesos - Homicídios.

	Classes	SMOTE	Pesos
Homicídio simples	8790	7031	0.6588
Homicídio qualificado	2791	7031	2.0743

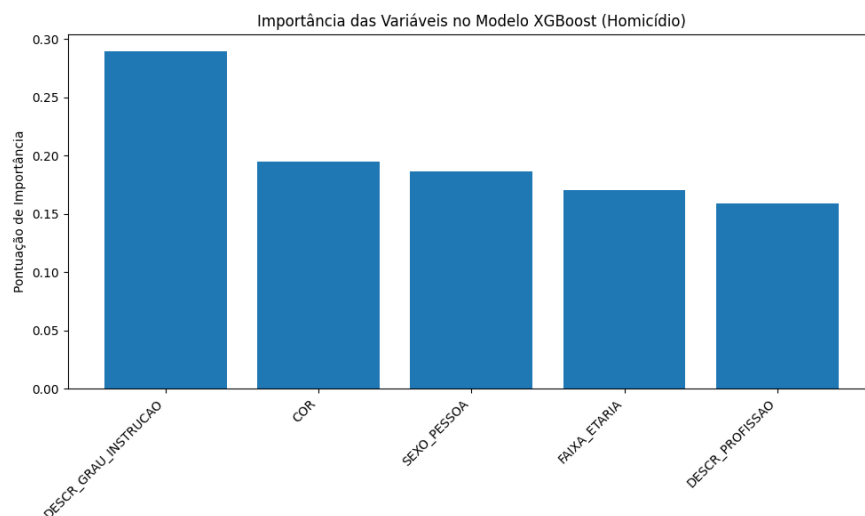


Figura 3.21

A aplicação do algoritmo com e sem os métodos de balanceamento de classes levou aos resultados que podem ser observados nas Figuras 3.22 à 3.24 e Tabelas 3.7 à 3.9. Novamente, percebeu-se um grave problema de desbalanceamento no conjunto de dados, mesmo dentro da classe de homicídios que, anteriormente, se mostrou como a classe majoritária em relação aos demais crimes. Observando-a internamente, é possível reparar que a classe de homicídio simples apresenta mais que três vezes a quantidade de dados da classe de homicídio qualificado. Quando denotamos esse problema e o associamos aos resultados apresentados nas matrizes de confusão, percebemos que o modelo mais uma vez aprendeu a priorizar a classe com mais exemplos em detrimento da classe menos representada. A alta acurácia de 0.75 no modelo sem pesos e sem SMOTE mascara este problema, pois é inflada pelo bom desempenho na classe majoritária. Reparamos também no *recall* da classe minoritária, que indica uma péssima capacidade de identificação correta dos exemplos reais de homicídio qualificado.

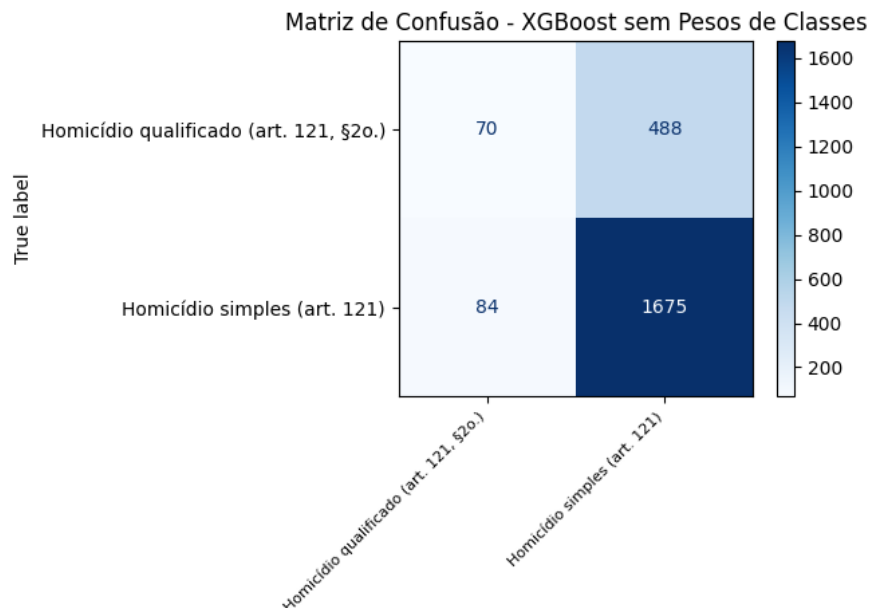


Figura 3.22: Matriz de Confusão - sem Pesos.

Tabela 3.7: Resultados sem Peso de Classes.

Classe	Precision	Recall	F1-Score
Homicídio qualificado	0.45	0.13	0.20
Homicídio simples	0.77	0.95	0.85
Acurácia	0.75		

Quando partimos para os modelos com balanceamento, percebemos uma clara melhora em relação aos dados minoritários. Apesar da classe majoritária apresentar certo decaimento em relação aos resultados do modelo sem balanceamento, os modelos balanceados compensam na melhora evidente da capacidade de acertar os exemplos reais da classe minoritária, com *recall* de 0.44 e 0.41 para os modelos com pesos e SMOTE, respectivamente. É evidente que o balanceamento de classes, nesse caso, não soluciona o problema e ainda demonstra baixa eficácia no equilíbrio global dos dados, uma vez que a acurácia decaiu consideravelmente, de 0.75 para 0.66, em ambos os casos de balanceamento.

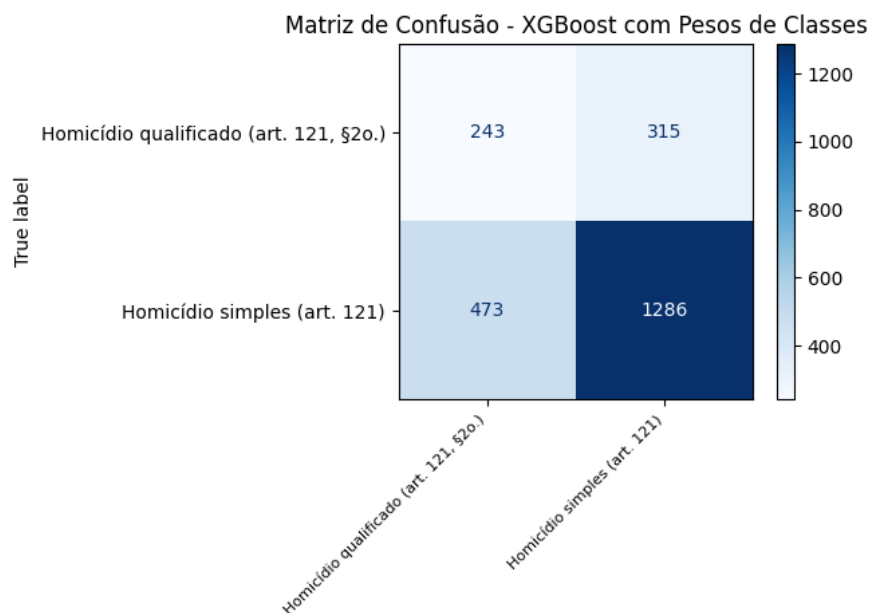


Figura 3.23: Matriz de Confusão - com Pesos.

Tabela 3.8: Resultados com Peso de Classes

Classe	Precision	Recall	F1-Score
Homicídio qualificado	0.34	0.44	0.38
Homicídio simples	0.80	0.73	0.77
Acurácia	0.66		

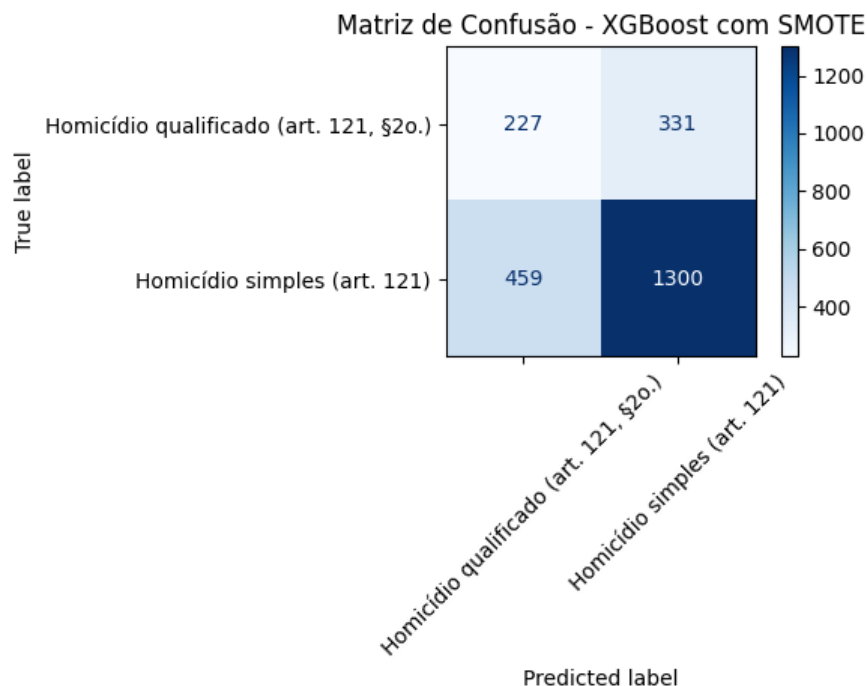


Figura 3.24: Matriz de Confusão - SMOTE sem Pesos.

Tabela 3.9: Resultados com SMOTE

Classe	Precision	Recall	F1-Score
Homicídio qualificado	0.33	0.41	0.36
Homicídio simples	0.80	0.74	0.77
Acurácia	0.66		

Modelo para crimes de estupro

Por fim, as últimas análises realizadas envolveram o caso dos crimes de estupro e estupro de vulnerável, que apresentaram as seguintes distribuições de classes, pesos e redistribuições via SMOTE (Tabela 3.10):

Tabela 3.10: Distribuição de Classes, SMOTE e Pesos - Estupros.

	Classes	SMOTE	Pesos
Estupro	4008	3206	0.7321
Estupro de vulnerável	1860	3206	1.5773

É possível reparar que, entre os dados analisados, os de crime de estupro são os mais internamente balanceados, apesar de ainda apresentarem uma certa distância. No entanto, esse comportamento não é corroborado pelos resultados obtidos no modelo sem balanceamento de dados. Como pode ser observado na Figura 3.26 e na Tabela 3.11, o modelo não detecta caso algum da classe minoritária, demonstrando que é sim estritamente necessário, apesar

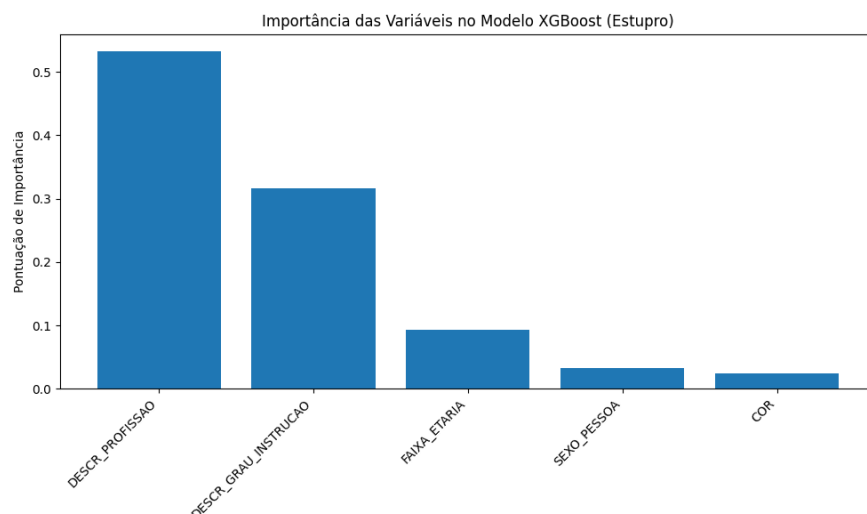


Figura 3.25: Variáveis Explicativas para o Modelo com Peso - Estupro.

dos resultados pouco satisfatórios em todos os modelos anteriores, que seja realizado um balanceamento de classes de alguma maneira.

Além disso foram identificadas as variáveis mais explicativas para o modelo dos dados de estupro, como pode ser visto na Figura 3.25, e eles indicam um padrão claramente diferente em relação aos demais dados, onde profissão e educação dominam completamente o poder explicativo, respondendo por uma boa parcela da importância total. Gênero e cor, que eram predominantes na análise geral, tornam-se quase irrelevantes. Isso sugere que estupros podem estar mais relacionados a contextos socioeconômicos específicos e oportunidades, mais do que características demográficas básicas.

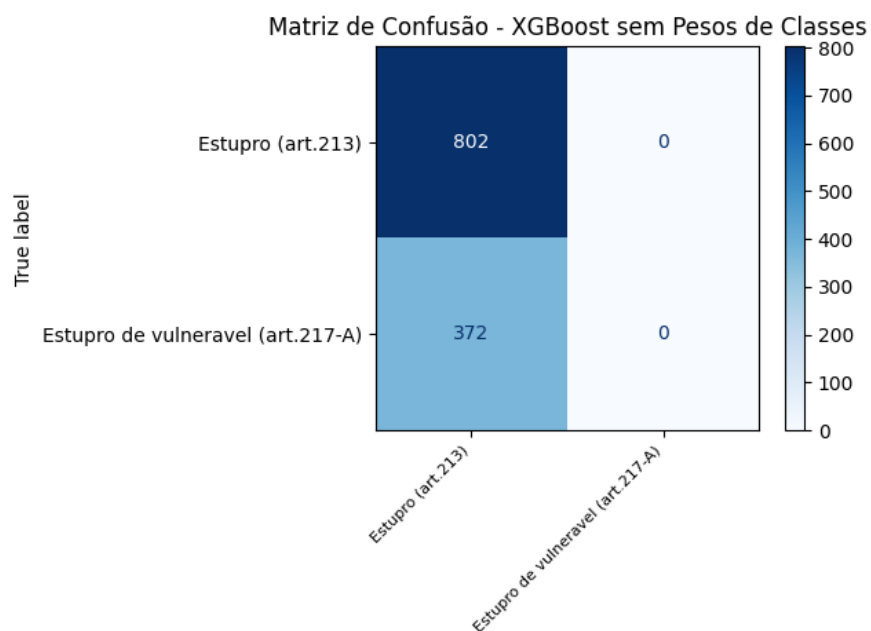


Figura 3.26: Matriz de Confusão - sem Pesos

Tabela 3.11: Resultados sem Peso de Classes.

Classe	Precision	Recall	F1-Score
Estupro	0.68	1.00	0.81
Estupro de vulneravel	0.00	0.00	0.00
Acurácia		0.68	

A aplicação de métodos de balanceamento transformou radicalmente o cenário apresentado no modelo sem balanceamento. Como pode ser visto nas Figuras 3.27 e 3.28 e nas Tabelas 3.12 e 3.13, ambos os métodos conseguiram fazer o modelo aprender a detectar a classe minoritária, com a classe de estupro de vulnerável alcançando *f1-scores* de 0.58 e 0.57, respectivamente. O *recall* para esta classe saltou de zero para 0.66 com peso de classes e 0.60 com SMOTE, representando uma melhoria surpreendente na capacidade de identificação de casos. É importante notar que essa melhoria veio com uma redução moderada no desempenho da classe majoritária (o *f1-score* da classe de crimes de estupro caiu de 0.81 para 0.76-0.78), mas, nesse caso, é uma perda irrisória e necessária para um modelo mais justo, uma vez que o modelo não balanceado apresentava comportamento completamente enviesado em relação à classe majoritária.

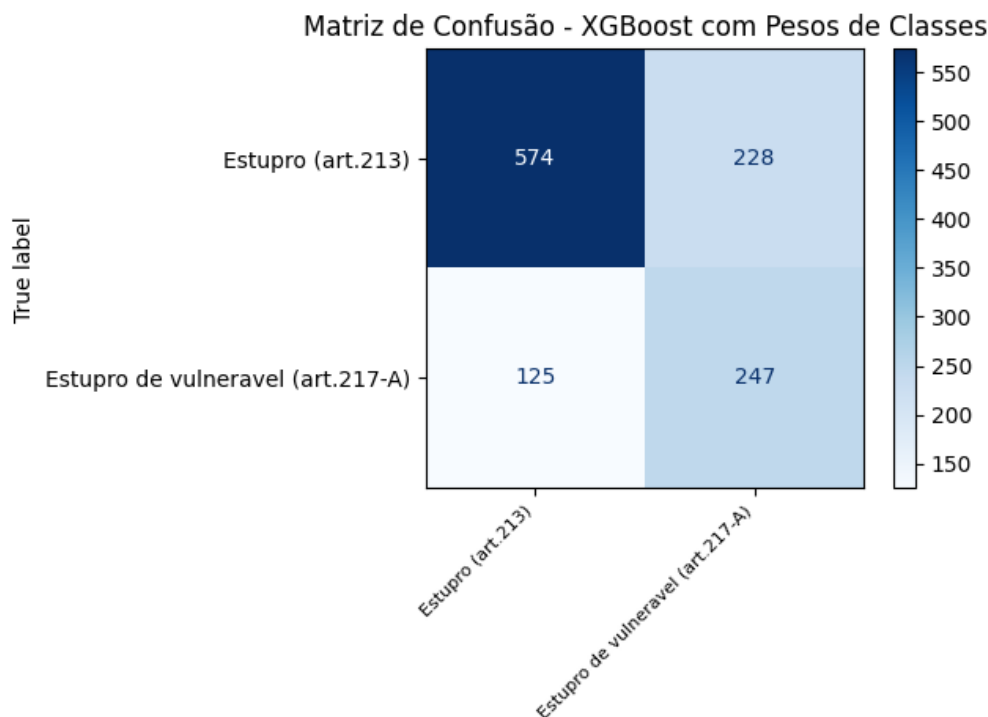


Figura 3.27: Matriz de Confusão - com Pesos.

Tabela 3.12: Resultados com Peso de Classes.

Classe	Precision	Recall	F1-Score
Estupro	0.82	0.72	0.76
Estupro de vulneravel	0.52	0.66	0.58
Acurácia	0.70		

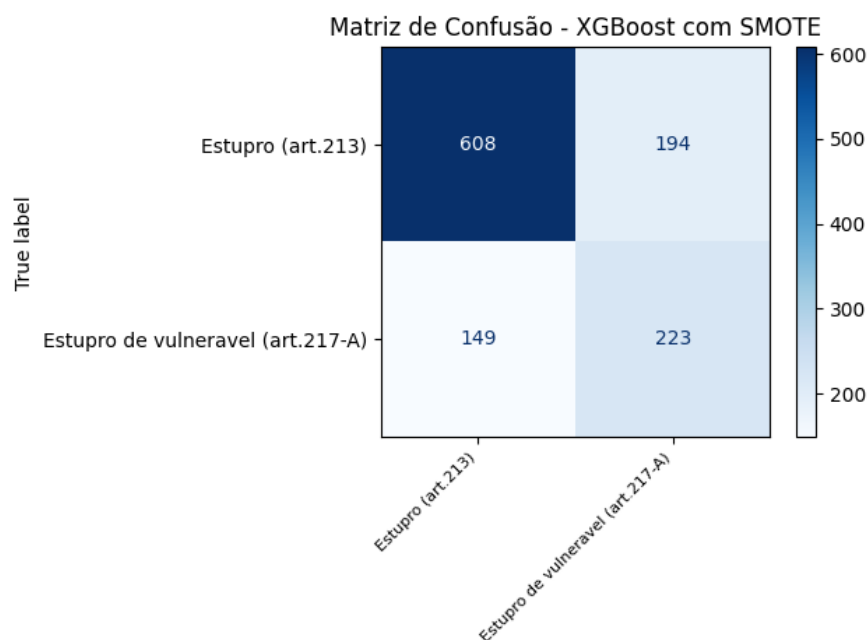


Figura 3.28: Matriz de Confusão - SMOTE sem Pesos.

Tabela 3.13: Resultados com SMOTE.

Classe	Precision	Recall	F1-Score
Estupro	0.80	0.76	0.78
Estupro de vulneravel	0.53	0.60	0.57
Acurácia	0.71		

As técnicas de balanceamento mostraram-se essenciais na modelagem para os casos da classe de crimes de estupro, com ambos os métodos (peso de classes e SMOTE) produzindo resultados semelhantes. A acurácia geral também melhorou modestamente (de 68% para 70-71%) e, além disso, passou a representar um valor que corresponde verdadeiramente ao equilíbrio dos dados, sem estar mascarada pelo viés da classe majoritária, indicando que, além de mais equilibrado, o modelo se tornou ligeiramente mais preciso globalmente. Os resultados demonstram que para aplicações reais onde todas as classes são importantes, o uso de técnicas de balanceamento não é opcional, mas sim necessário para evitar que crimes menos frequentes sejam completamente negligenciados pelo sistema de classificação.

4 Conclusão

A análise inicial, apesar das limitações e da alta taxa de dados faltantes, revelou padrões importantes presentes nos dados de indiciados na cidade de São Paulo entre os anos de 2007 e 2014. A maioria dos indiciados por homicídio, estupro e lesão corporal foram homens, enquanto a faixa etária predominante para a maioria dos crimes foi entre 20 e 35 anos.

A análise descritiva também destacou a alta frequência de indiciados desempregados, sugerindo uma possível correlação entre a falta de ocupação formal e a criminalidade. Em relação à escolaridade, a maioria dos indiciados possui apenas o 1º grau completo, indicando também um baixo nível educacional por parte daqueles que foram indiciados por crimes violentos. No que diz respeito à cor, a população preta/parda e a população branca apresentaram distribuições similares de ocorrências para os crimes analisados, enquanto as demais cores apresentaram pouquíssimos casos associados.

A análise das variáveis mais explicativas revelou que diferentes crimes possuem dinâmicas distintas: enquanto fatores demográficos (gênero e cor) predominam na análise geral, homicídios mostram forte associação com educação, e estupros estão profundamente ligados a fatores socioeconômicos (profissão e educação). Essa heterogeneidade sugere que políticas de prevenção devem ser específicas para cada tipo de crime, mais que abordagens genéricas.

A etapa de modelagem, com o uso do algoritmo *XGBoost*, demonstrou vários desafios quanto ao intuito de trabalhar com dados reais e severamente desbalanceados. Nos modelos que tentaram classificar os crimes agrupados (homicídio, estupro e lesão corporal), a classe de lesão corporal, que era a menos representada, foi totalmente negligenciada pelo modelo sem balanceamento. Embora as técnicas de balanceamento de classes (pesos e SMOTE) tenham melhorado a detecção de estupro, a performance para lesão corporal permaneceu bastante insatisfatória, o que sugere que a quantidade e a qualidade dos dados para esta categoria são insuficientes para uma classificação eficaz. Neste caso, preferiu-se por não realizar aplicações de modelos voltados exclusivamente para os dados de lesão corporal.

A modelagem interna para crimes de homicídio apresentou um problema similar de desbalanceamento entre homicídio simples e qualificado. O modelo sem balanceamento se mostrou tendencioso, com alta acurácia, mas baixo *recall* para homicídio qualificado. Novamente, as técnicas de balanceamento melhoraram a capacidade de identificação da classe minoritária, mas com o custo de uma queda na acurácia geral, evidenciando a necessidade de buscar um equilíbrio entre a precisão geral e a capacidade de detectar classes menos frequentes. Assim como citado para o caso de lesão corporal, esse fato evidencia que a qualidade dos dados ou a falta de mais variáveis explicativas para esta categoria pode ser um fator limitante para a performance do modelo.

No entanto, a classificação dos crimes de estupro apresentou uma melhora considerável em relação aos demais, onde a aplicação das técnicas de balanceamento foi decisiva. O modelo sem balanceamento não conseguiu detectar casos de estupro de vulnerável, a classe minoritária dentro dos crimes de estupro. Após a aplicação de pesos de classes e do SMOTE, o

modelo não apenas aprendeu a identificar essa classe minoritária de forma significativa, como a acurácia geral aumentou. Isso demonstra que para problemas reais de classificação, o balanceamento de classes é estritamente necessário e uma etapa importantíssima para evitar que o modelo seja enviesado e para garantir que todas as categorias de interesse sejam consideradas.

Em suma, este estudo evidenciou a complexidade de analisar e modelar dados criminais reais, que frequentemente apresentam problemas de qualidade e desequilíbrio, com muitos dados faltantes, poucos registros de alguns tipos de casos e erros no armazenamento desses dados. A análise descritiva forneceu entendimentos valiosos sobre o perfil dos indiciados, e a modelagem com aprendizado de máquina, embora enfrentando muitos desafios, ressaltou a importância de técnicas de pré-processamento para obter resultados mais justos e confiáveis. Os padrões distintos encontrados nas variáveis explicativas para diferentes crimes reforçam a necessidade de abordagens diferenciadas em políticas de segurança pública. Os resultados podem servir de base para políticas públicas, destacando a relevância da escolaridade e da ocupação na prevenção da criminalidade, ao mesmo tempo que apontam para a necessidade de dados mais completos, detalhados e com mais possibilidades de variáveis explicativas para análises futuras.

Bibliografia

- Brasil (jul. de 1990). *Estatuto da Criança e do Adolescente*. Lei nº 8.069, de 13 de julho de 1990. Dispõe sobre o Estatuto da Criança e do Adolescente e dá outras providências. Brasília, DF.
- Inquisitivecrow (n.d.). *Crime Data in Brazil*. Kaggle. URL: <https://www.kaggle.com/datasets/inquisitivecrow/crime-data-in-brazil> (acesso em 17/08/2025).
- Jusbrasil ([Ano]). *[Título do artigo ou conteúdo específico]*. Disponível em: <https://www.jusbrasil.com.br/>. Acesso em: [dia, mês. ano].
- Kaggle (n.d.). *Kaggle: Your Machine Learning and Data Science Community*. URL: <https://www.kaggle.com> (acesso em 17/08/2025).