**UNIVERSIDADE FEDERAL DO CEARÁ**

**CENTRO DE CIÊNCIAS**

**DEPARTAMENTO DE ESTATÍSTICA E MATEMÁTICA APLICADA**

**PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM E MÉTODOS QUANTITATIVOS**

**MESTRADO ACADÊMICO EM MODELAGEM E MÉTODOS QUANTITATIVOS**

**ANA CAROLINA NEPOMUCENO COSTA**

# AVALIAÇÃO DE PREDIÇÃO DE VIOLÊNCIA CONTRA A MULHER ATRAVÉS DE ESTRATÉGIAS DE APRENDIZADO DE MÁQUINAS

**FORTALEZA**

**2023**

ANA CAROLINA NEPOMUCENO COSTA

# AVALIAÇÃO DE PREDIÇÃO DE VIOLÊNCIA CONTRA A MULHER ATRAVÉS DE ESTRATÉGIAS DE APRENDIZADO DE MÁQUINAS

Dissertação apresentada ao Curso de Mestrado Acadêmico em Modelagem e Métodos Quantitativos do Programa de Pós-Graduação em Modelagem e Métodos Quantitativos do Centro de Ciências da Universidade Federal do Ceará, como requisito parcial à obtenção do título de mestre em Modelagem e Métodos Quantitativos. Área de Concentração: Modelagem e Métodos Quantitativos

Orientador: Prof. Dr. Charles Casimiro Cavalcante

Coorientador: Prof. Dr. Guilherme de Alencar Barreto

FORTALEZA

2023

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Sistema de Bibliotecas
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

C87a Costa, Ana Carolina Nepomuceno.
Avaliação de predição de violência contra a mulher através de estratégias de aprendizado de máquinas / Ana Carolina Nepomuceno Costa. – 2023.
65 f. : il. color.

Dissertação (mestrado) – Universidade Federal do Ceará, Centro de Ciências, Programa de Pós-Graduação em Modelagem e Métodos Quantitativos, Fortaleza, 2023.
Orientação: Prof. Dr. Charles Casimiro Cavalcante.
Coorientação: Prof. Dr. Guilherme de Alencar Barreto.

1. Algoritmos supervisionados. 2. Aprendizado de máquinas. 3. Seleção de atributos. 4. Violência contra as mulheres. 5. Classificação. I. Título.

CDD 510

ANA CAROLINA NEPOMUCENO COSTA

# AVALIAÇÃO DE PREDIÇÃO DE VIOLÊNCIA CONTRA A MULHER ATRAVÉS DE ESTRATÉGIAS DE APRENDIZADO DE MÁQUINAS

Dissertação apresentada ao Curso de Mestrado Acadêmico em Modelagem e Métodos Quantitativos do Programa de Pós-Graduação em Modelagem e Métodos Quantitativos do Centro de Ciências da Universidade Federal do Ceará, como requisito parcial à obtenção do título de mestre em Modelagem e Métodos Quantitativos. Área de Concentração: Modelagem e Métodos Quantitativos

Aprovada em: 16/06/2023

BANCA EXAMINADORA

---

Prof. Dr. Charles Casimiro Cavalcante (Orientador)
Universidade Federal do Ceará (UFC)

---

Prof. Dr. Guilherme de Alencar Barreto (Coorientador)
Universidade Federal do Ceará (UFC)

---

Prof. Dr. José Raimundo Carvalho
Universidade Federal do Ceará (UFC)

## AGRADECIMENTOS

Agradeço principalmente a Deus por tudo.

À pessoa que me ajudou, teve confiança, paciência e me fez acreditar que eu podia realizar este trabalho, meu professor e orientador, o Dr. Charles Casimiro, me dando todo o apoio, ensinamentos nas reuniões de orientandos e orientação nessa dissertação.

À banca de avaliadores que aceitou participar deste momento tão importante na minha vida: o Dr. José Raimundo, coordenador do Programa de Pós-Graduação em Economia da Universidade Federal do Ceará (CAEN), que forneceu as informações necessárias para a realização deste trabalho, e o Dr. Guilherme de Alencar Barreto, com seu conhecimento técnico em aprendizado de máquina, para oferecer suas sugestões.

Aos meus pais, que, nos momentos de minha ausência dedicados ao estudo superior, sempre me fizeram entender que o futuro é construído a partir da constante dedicação no presente.

Ao meu namorado e amigo Luís Fernando, pelo amor, atenção e por me ajudar a acreditar que posso superar as dificuldades.

Agradeço a todos os professores por me proporcionarem o conhecimento não apenas racional, mas a manifestação do caráter e afetividade da educação no processo de formação profissional. Agradeço também por terem se dedicado a mim, não apenas por me ensinarem, mas por terem me feito aprender.

E por fim, a todas as pessoas que conheci nesta jornada, seja com um "bom dia" ou palavras de incentivo, que de alguma forma me estimularam a continuar, que contribuíram para que este sonho se torne realidade.

## RESUMO

A violência contra a mulher é um problema de educação, saúde e segurança públicas que afeta uma a cada três mulheres no mundo, segundo a Organização Mundial de Saúde (OMS). Além disso, percebe-se que a violência contra a mulher inicia-se prematuramente, pois uma a cada quatro mulheres entre 15 e 24 anos já relataram violência cometida pelo seu parceiro quando estiveram em um relacionamento. Ademais, 1.48 milhões de mulheres reportaram violência entre 2010 e 2018, segundo o Instituto Igarapé. Diante desses números preocupantes, esse trabalho objetiva na predição de violência contra a mulher e encontrar padrões de violência com o intuito de reduzir essas taxas alarmantes e auxiliar na mensuração dos dados. Para isso, foram feitos modelos para obter uma métrica de classificação e identificação de características que resultam em violência. Portanto, essa pesquisa é focada em identificar quais atributos são mais importantes para um bom resultado do modelo, utilizando *Feature Selection*, e reconhecer padrões de violência através da análise exploratória dos dados obtidos pelo Programa de Pós-Graduação em Economia (CAEN/UFC), no qual os questionários são da Pesquisa de Condições Socioeconômicas e Violência Doméstica e Familiar contra a Mulher (PCSVDF-Mulher). Os resultados indicam a obtenção de percentuais altos, da ordem de 81%, na capacidade de classificação de eventos de violência, demonstrando uma possibilidade real de contribuir com modelos preditivos aos questionários de informações sobre violência.



**Palavras-chave:** algoritmos supervisionados; aprendizado de máquinas; seleção de atributos; violência contra as mulheres; classificação.

## ABSTRACT

Violence against women is a educational, public health and safety problem that affects one in three women worldwide, according to the World Health Organization (WHO). In addition, it is clear that violence against women begins prematurely, as one in four women between 15 and 24 years have already reported violence committed by their partner when they were in a relationship. In addition, 1.48 million women reported violence between 2010 and 2018, according to the Igarapé Institute. In view of these worrying numbers, this work aims to predict violence against women and find patterns of violence in order to reduce these alarming rates and assist in results measurement. For this, models were proposed to obtain metrics for classifying and identifying characteristics that result in violence. Therefore, this research is focused on identifying which attributes are most important for a good result of the model, using *Feature Selection* and recognizing patterns of violence through exploratory analysis of data obtained by the Graduate Program in Economics (CAEN/ UFC), in which the questionnaires are from the Survey of Socioeconomic Conditions and Domestic and Family Violence against Women (PCSVDF-Mulher, in Portuguese).The results indicate high performance, around 81%, on the classification of violence events, showing a real possibility to contribute with predictive models to formularies of violence information.



**Keywords:** supervised algorithms; machine learning; feature selection; violence against women; classification.

## LISTA DE FIGURAS

## LISTA DE TABELAS

**SUMÁRIO**

# 1 INTRODUÇÃO

## 1.1 Motivação

A violência contra a mulher é um problema de saúde e segurança públicas que afeta uma a cada três mulheres no mundo, segundo a Organização Mundial de Saúde (OMS). Além disso, percebe-se que a violência contra a mulher inicia-se bem cedo. Dessa forma, uma a cada quatro mulheres jovens, que tem entre 15 e 24 anos, já sofreu violência pelo parceiro quando estavam em um relacionamento. O mais alarmante é que esse número permanece praticamente o mesmo (se não, aumenta, principalmente com o impacto da pandemia gerada da COVID-19) por mais de uma década segundo a OMS. E essa violência praticada pelo parceiro é identificada como a predominante violência em todas as mulheres do mundo (ORGANIZAÇÃO MUNDIAL DE SAÚDE, 2021).

Desse modo, sabe-se que a violência, de todas as formas cometidas, tem consequências negativas na saúde e bem-estar da mulher pelo resto de sua vida, mesmo que ela passe um longo período sem ser vítima, mas que já tenha sofrido algum tipo de violência. Isso se reflete na sociedade em sua totalidade, sejam por problemas psicológicos (depressão, ansiedade e outros), como por outros problemas como contrair doenças sexualmente transmissíveis, gravidez não planejada e até ocasionar em feminicídio.

Posto isso, para combater esse tipo de violência, há uma urgência em capacitar profissionais de saúde, de educação e de segurança pública a fim de encarar atitudes e crenças que discriminam os gêneros para reduzir o machismo estrutural e proporcionar a igualdade de gênero (ORGANIZAÇÃO MUNDIAL DE SAÚDE, 2021).

Todavia, como nesta pesquisa ainda não se pode promover esse tipo de capacitação, por se tratar de um tema amplo e complexo, busca-se fazer uma predição da violência, de forma quantitativa, através de levantamento de dados amostrais para fortalecer e mensurar a pesquisa sobre esse tema, por ser bastante significativo, uma vez que os danos que podem causar às mulheres violentadas, principalmente jovens, são inúmeros, podendo até ser irreversível.

Dito isso, se mostra relevante desenvolver um modelo computacional para auxiliar que sejam tomadas e realizadas ações para diminuição dos casos e reforçar as medidas das organizações responsáveis, além de melhorar a medição das diferentes formas de violência sofridas por mulheres de todas as idades, incluindo as bastante jovens.

## 1.2 Objetivos

O objetivo geral desta pesquisa é desenvolver uma abordagem baseada no aprendizado de máquinas supervisionado de alguns algoritmos capazes de fazerem uma boa predição de identificar a chance de uma mulher ser vítima de violência com base numa avaliação dos hiperparâmetros que constituem o espaço de observação dos dados de coleta de informações, além de analisar os dados a fim de identificar padrões que indicam uma possível violência. Essa predição permitirá a elaboração de futuras ações para diminuição da violência contra a mulher.

Para atingir o objetivo geral dessa pesquisa, alguns objetivos específicos serão de fundamental importância:

**a)** Análise dos dados recebidos para obter percepções acerca da problemática;

**b)** Realização de uma revisão da literatura sobre o contexto em que o projeto está inserido de violência contra a mulher e a predição de violência, tendo como objetivo identificar os algoritmos que são candidatos à resolução do problema apresentado;

**c)** Identificação das medidas relevantes para os modelos apresentados;

**d)** Criação dos modelos e posterior verificação das medidas, com o intuito de verificar se o modelo proposto atende à expectativa de predizer os eventos de violência de forma satisfatória e;

**e)** Utilização de Matriz de Confusão e análise dos valores encontrados utilizando as medidas de classificação para fins de comparação entre os modelos avaliados.

Com efeito, a predição encontrada pelo modelo proposto nesse trabalho poderá influenciar na situação da mulher, no sentido de classificar se a mulher tem ou não chances de sofrer violência com base nos elementos abordados no questionário feito pelo Programa de Pós-Graduação em Economia (CAEN/UFC) da Pesquisa de Condições Socioeconômicas e Violência Doméstica e Familiar contra a Mulher (PCSVDF-Mulher).

## 1.3 Estrutura do trabalho

O restante desse trabalho é organizado em 4 capítulos. No Capítulo 2 é feita uma revisão da literatura sobre os aspectos da violência contra a mulher, Aprendizado de Máquina e suas técnicas usadas nos modelos para classificação além das medidas usadas por cada modelo. Já no Capítulo 3 é discutido a descrição do problema em si e sua modelagem matemática para a realização do experimento. Neste capítulo, encontram-se ainda as informações sobre a proposta

e a metodologia para realização do trabalho.

As etapas do experimento e as etapas do processo de predição de eventos de violência são detalhadas no Capítulo 4, no qual também são mostrados os resultados obtidos para os diferentes modelos de aprendizado de máquinas.

Por fim, no Capítulo 5 são apresentadas as conclusões e contribuições feitas até o momento acerca do processo e mostrados os resultados preliminares. Também são relatadas algumas limitações e possíveis melhorias que podem ser feitas para os avanços na temática abordada pela dissertação.

# 2 FUNDAMENTAÇÃO TEÓRICA

Para atingir os objetivos definidos, nesse capítulo é apresentada uma revisão de importantes estudos na área, através de uma busca bibliográfica em diferentes bases de dados, utilizando as palavras-chaves: "violência contra a mulher", "feminicídio no Brasil", "*Machine Learning*", "Algoritmos supervisionados de Aprendizado de Máquina", "Regressão Logística", "Random Forest" e outros, sendo determinado como critério de exclusão os materiais com enfoque apenas com conteúdos jurídico, penal e de medicina. Além disso, são excluídos os trabalhos mais antigos, sendo consideradas apenas as referências a partir de 2011. Além da base de dados recebida pelo CAEN, buscou-se outras bases de dados para complementar este trabalho.

Para tanto, a presente seção é separada em duas partes. Primeiro é discutido sobre o contexto de violência contra a mulher no qual o problema se insere. Em seguida são apresentados alguns algoritmos de aprendizado de máquina supervisionados.

## 2.1 Violência contra a mulher

A violência contra a mulher é uma questão de saúde, segurança públicas educação no mundo, no qual muitas mulheres sofrem variados tipos de violência, seja emocional, moral, sexual, física, doméstica, dentre outros podendo até trazer o caso mais grave que é a morte, chamado comumente de feminicídio. Portanto, a coleta de dados sobre esse assunto é fundamental para procurar entender padrões da violência contra a mulher e ir no cerne dessa problemática para levantar ações visando diminuir os casos de violência. Porém, a coleta e divulgação desses dados enfrentam uma série de desafios, desde a falta de iniciativa de vários órgãos responsáveis, seja de saúde ou segurança, até encontrar muitos dados ausentes quando são coletados. Este último caso se remete provavelmente pela própria falta de segurança e apoio para a mulher em informar a sua realidade.

Para fundamentar mais essa pesquisa serão utilizadas referências de leis relacionadas a violência contra a mulher e trabalhos que contém informações sobre o assunto. Além disso, buscam-se organizações e institutos que também se debruçam sobre essa temática e dados com o intuito de coletar percepções acerca do tema.

Diante disso, no Brasil, existe a Lei Maria da Penha (BRASIL, 2006) que define a agressão contra a mulher e suas formas de evitar, enfrentar e punir tal violência, além de apontar

quais órgãos são responsáveis para auxiliar a mulher em situação de risco.

Segundo a Lei Maria da Penha (BRASIL, 2006) existem 5 formas de violência: violência física, psicológica, sexual, patrimonial e moral. Para a primeira, a Lei entende como violência física aquela que refere-se a "[...] qualquer conduta que ofenda sua integridade ou saúde corporal" (BRASIL, 2006). Já a segunda, que é a violência psicológica, teve sua redação modificada em 2018 adicionando à lista "violação de sua intimidade" da sequência de condutas que degradam e controlam como mostra na Lei Maria da Penha (BRASIL, 2006):

> II - a violência psicológica, entendida como qualquer conduta que lhe cause dano emocional e diminuição da autoestima ou que lhe prejudique e perturbe o pleno desenvolvimento ou que vise degradar ou controlar suas ações, comportamentos, crenças e decisões, mediante ameaça, constrangimento, humilhação, manipulação, isolamento, vigilância constante, perseguição contumaz, insulto, chantagem, violação de sua intimidade, ridicularização, exploração e limitação do direito de ir e vir ou qualquer outro meio que lhe cause prejuízo à saúde psicológica e à autodeterminação; (Redação dada pela Lei nº 13.772, de 2018) (BRASIL, 2006).

A terceira, dita como violência sexual, que resumidamente envolve a relação sexual causada quando não é intencionada pela mulher. Esse tipo de violência é informada na Lei Maria da Penha (BRASIL, 2006) a seguir:

> III - a violência sexual, entendida como qualquer conduta que a constranja a presenciar, a manter ou a participar de relação sexual não desejada, mediante intimidação, ameaça, coação ou uso da força; que a induza a comercializar ou a utilizar, de qualquer modo, a sua sexualidade, que a impeça de usar qualquer método contraceptivo ou que a force ao matrimônio, à gravidez, ao aborto ou à prostituição, mediante coação, chantagem, suborno ou manipulação; ou que limite ou anule o exercício de seus direitos sexuais e reprodutivos;(BRASIL, 2006).

O quarto tipo tipo de violência é a patrimonial que envolve os bens materiais e outras tipos semelhantes da mulher. Esse tipo de violência é melhor explicada na Lei Maria da Penha (BRASIL, 2006) abaixo:

> IV - a violência patrimonial, entendida como qualquer conduta que configure retenção, subtração, destruição parcial ou total de seus objetos, instrumentos de trabalho, documentos pessoais, bens, valores e direitos ou recursos econômicos, incluindo os destinados a satisfazer suas necessidades;(BRASIL, 2006).

E por último, a quinta violência é a definida como violência moral "[...] entendida como qualquer conduta que configure calúnia, difamação ou injúria" (BRASIL, 2006).

Percebe-se que, com frequência, as mulheres são retratadas como as principais vítimas de todos esses tipos de violência citados e que vem aumentando cada vez mais esses relatos ao longo dos anos (INSTITUTO IGARAPÉ, ).

Ainda que tenha sido uma vitória as conquistas e os avanços sociais assegurados pela Lei Maria da Penha (BRASIL, 2006), a violência continua sendo uma triste realidade para muitas mulheres no Brasil. Não obstante, deve-se considerar que o aumento dos dados apontados pelas estatísticas podem também mostrar o crescimento das denúncias pela popularização da lei e pelo auxílio dos movimentos da sociedade civil organizada, sobretudo, do Movimento de Mulheres. Este movimento tem sido agente para a garantia das medidas protetivas e de urgência, junto aos órgãos do poder público que são os responsáveis em prevenir, proteger e coibir a violência (ESCORSIM, ).

A fim de diminuir esses números alarmantes, se torna necessário obter dados sobre os casos de violência contra a mulher, na tentativa de reconhecer seus padrões e saber onde deve-se atentar nas referências achadas para prevenção, redução e eliminação dos casos de violência contra mulheres e meninas.

Tendo em vista esses fatos citados, foi elaborado um questionário pelo Programa de Pós-Graduação em Economia (CAEN/UFC) nomeado de PCSVDF-Mulher em diferentes anos (chamados de ondas), em que são perguntadas algumas informações sobre a pessoa entrevistada, tais como: onde ela mora, sua idade, se há algum tipo de violência na sua rua, se tem filhos e outros. Ainda, são feitas vários questionamentos sobre a violência em si e se a entrevistada tem conhecimento sobre a Lei Maria da Penha e outros meios de apoio e segurança para caso aconteça algo com ela ou outra pessoa que ela conheça.

Ademais, os dados obtidos em diferentes anos contêm as mesmas mulheres entrevistadas para serem acompanhadas a fim de analisar os dados e entender padrões. De fato, o recorte dos dados trabalhados na nossa pesquisa é de mulheres que foram entrevistadas em 2016 e consequentemente em 2017.

Além desses dados, também foram pesquisados outros dados públicos, nomeados de "Evidências sobre Violência e Alternativas para mulheres e meninas" (EVA) do Instituto Igarapé (INSTITUTO IGARAPÉ, ). Esta fonte apresenta três bases de dados que se integram e apresentam as formas de violência ao longo dos anos contra a mulher. Para tanto, são mostrados:

1. O que é registrado no sistema de saúde, após a atenção às mulheres que procuram atendimento médico ou para atestar uma morte violenta.

> 2. O que é conhecido no sistema de segurança, através de denúncias ou conhecimento das forças de segurança.
>
> 3. O que as mulheres revelam sobre seus níveis de vitimização pessoal, através de pesquisas de vitimização. (INSTITUTO IGARAPÉ, ).

Portanto, tais dados da EVA são imprescindíveis para complementar o presente trabalho de dissertação a fim de encontrar *insights* sobre os dados, pois a base dos dados principal (PCSVDF-Mulher) contém muitos dados faltantes ou não informados. Apesar das bases de dados da EVA também possuírem vários dados nulos/faltantes, eles encontram formas de completarem tais dados analisando conjuntamente o sistema de saúde, como apontado:

> A análise dos dados provenientes do sistema de saúde é fundamental para complementar os do sistema de segurança pública, uma vez que nem todas as mulheres denunciam as violências sofridas aos órgãos competentes. A comparação entre eles nos permite identificar os gargalos no que diz respeito à notificação desse tipo de crime (INSTITUTO IGARAPÉ, ).

Dessa forma, dos dados oriundos do sistema de saúde, onde tem-se informações sobre quais tipos de violência estão sendo atendidos em hospitais e outras unidades de mesma natureza, tem-se que 1.48 milhões de mulheres reportaram violência entre 2010 e 2018 no Brasil. Dessas, 27,3% de todas as formas de violência contra a mulher são cometidas por companheiros e ainda são 59% causado por violência física do que foi registrado. Ademais, 39,2% das vítimas de feminicídio têm entre 45 e 65 anos (INSTITUTO IGARAPÉ, ).

Nesse panorama, em 2016 a proporção por tipo de arma vem sendo em sua maioria por força corporal (55,8%), seguido por "não especificado" (14,2%). Em 2017 esses dados continuam praticamente o mesmo: "força corporal" com 56,3% e "não especificado" com 13,5% (INSTITUTO IGARAPÉ, ).

Também foram observadas as estatísticas da Secretaria de Segurança Pública e Defesa Social do Ceará (SSPDS CE) (CEARÁ. Secretaria de Segurança Pública e Defesa Social, ), que trazem diversos dados sobre segurança pública do estado, incluindo sobre violência contra a mulher e seus tipos.

Trazendo esse problema para o momento pós-pandemia causada pela COVID-19, o número de casos aumentou pelo fato de que as pessoas foram obrigadas a ficarem em casa no período de isolamento social e, por conseguinte, aumentou o número de violência doméstica e casos de feminicídio no país, chegando esse aumento a 300%, levando em conta os casos no primeiro trimestre de 2020 em relação ao mesmo período de 2019 (OKABAYASHI *et al.*, 2020).

Com o avanço da COVID-19, a questão de violência contra a mulher foi pouco discutida por causa do enfrentamento da pandemia, mas foi observado o aumento no número de registros de violência. Este cenário torna-se ainda mais expressivo porque em casos de violência doméstica contra a mulher, na maior parte das vezes, também há violência contra crianças e adolescentes (MARQUES *et al.*, ).

De fato esse é um problema desde que não é particular dessa época atual, pois a violência contra a mulher é consequência de uma sociedade patriarcal e sexista, mas que sempre está em voga e que deve-se ter atenção e tomar atitudes para diminuir os casos de violência contra a mulher.

## 2.2 Técnicas de aprendizado de máquina para classificação

A classificação tem muitas atribuições, tais como identificação de falhas e analisa os dados para discernir previsões e padrões (CHENG; GREINER, 2013). Essa tarefa (classificação) busca encontrar uma função que seja capaz de explicar as classes a partir de variáveis observadas (exemplos disponíveis). Desse modo, um problema que interfere no aprendizado de máquina é a inferência de classificadores com os dados pré-classificados. Desse ponto tem-se várias representações úteis tais como Àrvores de Decisão, *Random Forest*, Regressão Logística e outros. (ALPAYDIN, 2020).

### *2.2.1 Aprendizado de máquina*

O aprendizado de máquina é um sistema capaz de aprender e que, depois de determinados treinamentos, faz as mesmas atividades que lhe foi ensinado automaticamente de maneira eficaz. Existem dois tipos de aprendizado de máquina: o supervisionado e o não-supervisionado. No primeiro é criado um modelo que analisa os atributos de treinamento tendo conhecimento de quais são as respostas das classes dos atributos de teste, já no segundo não tem as respostas para elaborar o modelo. Nesta pesquisa serão utilizados os modelos supervisionados, especificamente os de classificação (ALPAYDIN, 2020) (MOHRI *et al.*, 2018).

Em (RASCHKA; MIRJALILI, 2019) e (ALBON, 2018) são apresentados modelos para a construção de boas bases de dados utilizando técnicas de pré-processamento, mostrando como tratar os dados para que se adequem à utilização de técnicas de aprendizagem de máquina supervisionada. Além disso, essas referências abordam um roteiro para a construção de sistemas

de aprendizado de máquina, ensinando a treinar e selecionar um modelo preditivo, utilizando da linguagem de programação *Python* para realizar a aprendizagem de máquina. Elas também mostram como fazer o treinamento de algoritmos de aprendizado de máquina para classificação.

Em relação à análise estatística dos dados faltantes, (LITTLE; RUBIN, 2019) abordam sobre o problema de *missing data* e como resolvê-los com experimentos e métodos. Ademais, (ALPAYDIN, 2020) e (MOHRI *et al.*, 2018) apresentam desde conceitos básicos até avançados sobre aprendizado de máquina.

As técnicas de Aprendizado de Máquina que abordam a problemática sobre os classificadores desta pesquisa pode ser encontrada em (MAIMON; ROKACH, 2014). Eles aprofundam o conhecimento focado em Árvores de Decisão. Já (CHRISTODOULOU *et al.*, ) se concentra em Regressão Logística. Além disso, (GÉRON, 2019) aborda sobre Árvore de Decisão, *Random Forest* e Regressão Logística.

A seguir, são apresentadas brevemente algumas técnicas bem definidas e conhecidas que podem ser usadas para previsão binária, já que consideramos duas classes (violência e não violência), levando em consideração a velocidade, a robustez, a interpretabilidade, a confiabilidade, a eficiência e por ser acessível no ambiente de pesquisa para escolher o melhor modelo que corresponde às características e requisitos da problemática em questão.

### *2.2.2 Algoritmos de predição*

Nesta subseção são mostrados breves conceitos sobre os algoritmos de classificação baseados em técnicas de Aprendizado de Máquina que são utilizados no modelo de predição.

#### 2.2.2.0.1 Árvores de decisão

Árvores de Decisão são estruturas em forma de árvore bem intuitiva e de entendimento simplificado. Elas representam conjuntos de decisões capazes de gerar uma árvore probabilística para um conjunto de dados específico ou uma estrutura que pode ser usada para dividir uma grande coleção de registros em conjuntos sucessivamente menores, aplicando uma sequência de regras de decisão simples. Assim, a árvore pode ser convertida em um conjunto de regras simples que são fáceis de entender. Outra possibilidade é aprender uma base de regras diretamente (MAIMON; ROKACH, 2014).

A Árvore de Decisão é composta por nós, chamados nós de decisão, que implementa uma função com resultados que rotulam os ramos, tomando decisões de acordo com sua entrada.

Esse processo começa na raiz e se repete até que um nó folha seja atingido que é a saída do modelo (ALPAYDIN, 2020).

Na classificação, a Árvore de decisão é quantificada por uma medida de impureza. Por isso, uma divisão é pura se após a divisão, para todas as ramificações, todas as instâncias que decidem uma ramificação pertencem à mesma classe. Essa medida de impureza comumente usada é a chamada Gini que pode ser vista na Equação (2.1) a seguir (MAIMON; ROKACH, 2014):

$$G_i = 1 - \sum_{k=1}^{n} (p_{i,k})^2. \tag{2.1}$$

Em que, $p_{i,k}$ é a razão de instâncias de classe $k$ entre as instâncias de treinamento no $i$-ésimo nó.

Nela, pode-se observar que o objetivo da impureza Gini é saber qual a probabilidade de classificar uma observação corretamente e diminuir pelo total, que é 1, para saber sua impureza.

Por padrão, a medida de impureza Gini é usada, mas é possível utilizar a medida de impureza de entropia que é uma medida de desordem. Quanto mais próxima de zero, mais idênticos os dados serão, ou seja, quando os dados são da mesma classe (GÉRON, 2019). A fórmula da entropia é mostrada na Equação (2.2):

$$H_i = -\sum_{k=1}^{n} (p_{i,k}).log_2(p_{i,k}). \tag{2.2}$$

A diferença das duas medidas é que a Gini é mais rápida para calcular, pois sua curva de custo é mais suave e ela calcula a probabilidade do modelo classificar errado, daí se torna o padrão, porém essa última tende a isolar a classe mais frequente no galho da árvore. Já a entropia tende a produzir árvores mais balanceadas, porque a sua curva de custo é mais punitiva, ou seja, ela calcula a probabilidade de ter um dado inesperado numa amostra de uma determinada classe (ALPAYDIN, 2020).

Se o nó for impuro, as instâncias devem ser divididas para diminuir a impureza, e existem vários atributos possíveis nos quais pode-se dividir. Dessa forma, procura-se a divisão que minimize a impureza após a divisão porque é desejado gerar a menor árvore. Se os subconjuntos após a divisão estiverem mais próximos do puro, menos divisões serão necessárias posteriormente. Porém esse método é localmente ótimo e não se tem garantia de encontrar a menor árvore de decisão (ALPAYDIN, 2020).

As vantagens do uso dessa técnica é que o seu resultado é facilmente explicável e ele é paralelizável, ou seja, executa uma única instrução sobre múltiplos dados. Ademais, esse algoritmo lida bem com dados faltantes e apresenta os atributos mais importantes na seleção. Já suas desvantagens é que ela é suscetível a mudanças nos dados e, com isso, pode causar *overfitting*, que é o sobreajuste dos dados. Para impedir o *overfitting*, é possível fazer a regularização dos hiperparâmetros (GÉRON, 2019).

Essa técnica pode ser usada tanto para regressão como para classificação e esta última terá mais enfoque neste trabalho. As Árvores de Decisão também são os componentes fundamentais do *Random Forest*, que está entre os algoritmos de Aprendizado de Máquina disponíveis atualmente e que será assunto da próxima subseção (GÉRON, 2019).

#### 2.2.2.0.2 Random forest

*Random Forest* (RF) é uma técnica de classificação inserida dentro do contexto de *Ensemble Learning*, também chamado de aprendizado por agrupamento. Ele se apoia na ideia de juntar vários modelos de predição mais simples, nesse caso de várias árvores, treiná-los para uma mesma tarefa e produzir a partir desses um modelo agrupado mais complexo no qual a técnica do *Random Forest* utiliza-se de várias árvores de decisão para ser construída (GÉRON, 2019).

Dessa forma, quando são combinados diversos modelos mais fracos, diminui a sensibilidade deles, o *bias* e a variância, tornando-os mais robustos. A partir disso, normalmente é escolhido apenas um modelo base para ser treinado em conjuntos diferentes e posteriormente é feita a combinação, formando assim um modelo homogêneo. Se o *Ensemble* final for formado por modelos diferentes dizemos que é um preditor heterogêneo.

O método RF, traduzido como Floresta Randômica, é uma técnica que constrói diversas Árvores de Decisão que são utilizadas para classificação. Essas árvores são treinadas utilizando diferentes subconjuntos aleatórios de amostras e de atributos. Na fase de teste, cada uma das árvores escolhe uma das classes e a que tiver mais escolhas é a classe prevista. Apesar de sua simplicidade, este é um dos algoritmos de aprendizado de máquina mais poderosos disponíveis atualmente (GÉRON, 2019).

O motivo do uso dessa técnica no presente trabalho é que ela fornece estimativas internas úteis de erro, força, correlação e a importância da variável. Além disso, o RF é facilmente paralelizado e relativamente robusto para *outliers* e ruídos, que são dados discrepantes. Ademais, o algoritmo *Random Forest* introduz aleatoriedade extra ao cultivar árvores, em vez de procurar

o melhor recurso ao dividir um nó (GÉRON, 2019). Porém, essa técnica comparada com a Regressão Logística tem maior custo computacional nos experimentos, como será vista no Capítulo 4.

2.2.2.0.3 Algoritmo XGBoost

O algoritmo *XGBoost* é um sistema de aprendizado de máquina escalável para aumento de árvores. Dessa forma, *Boosting* é uma técnica de aprendizado de máquina que pode ser usada para problemas de regressão e classificação. Ela gera um aprendiz fraco a cada passo e acumula no modelo total. Se o aprendiz fraco para cada etapa for baseado na direção do gradiente da função de perda, ele pode ser chamado de *Gradient Boosting Machines (GBM)* (PAN, ).

De maneira geral, o *Gradient Boosting*, que é a estrutura utilizada pelo *XGBoost* baseado em árvore de decisão, funciona adicionando preditores sequencialmente a um conjunto, cada um corrigindo seu predecessor. No entanto, em vez de ajustar os pesos da instância a cada iteração, esse método tenta ajustar o novo preditor aos erros residuais cometidos pelo preditor anterior (GÉRON, 2019).

A principal diferença entre *Random Forest* e *Gradient Boosted Machines* é que, enquanto no RF, as árvores são construídas independentemente umas das outras, o GBM adiciona uma nova árvore para complementar as já construídas (PAN, ).

Para a função objetivo, suponha que o conjunto de dados $D = (x_i, y_i) : i = 1...n, x_i \in \mathbb{R}^m, y_i \in \mathbb{R}$, então obtemos $n$ observações com $m$ *features* e com uma variável correspondente $y$. Enquanto $\hat{y_i}$ pode ser definido como o resultado dado por um conjunto representado pelo modelo generalizado como é mostrado na Equação (2.3):

$$\hat{y_i} = \Phi(x_i) = \sum_{k=1}^{k} f_k(x_i), \tag{2.3}$$

no qual $f_k$ é a árvore de decisão e $f_k(x_i)$ representa o escore obtido pela $k$-ésima árvore para a $i$-ésima observação nos dados. Para a função $f_k$ a função objetiva regularizadora (Equação (2.4)) pode ser otimizada:

$$\kappa(\Phi) = \sum_{i} \iota(y_i, \hat{y_i}) + \sum_{k} \Omega(f_k), \tag{2.4}$$

em que $\iota$ é a função de perda. Para evitar uma complexidade muito grande do modelo, o termo de penalidade $\Omega$ é incluída na Equação (2.5):

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda ||w|| \tag{2.5}$$

nesta equação, $\lambda$ e $\gamma$ são parâmetros de controle de penalidade para o número de folhas $T$ e magnitude dos pesos das folhas $w$ respectivamente. O intuito de $\Omega(f_k)$ é prevenir o *overfitting* e simplificar o modelo.

Dessa maneira, um método iterativo é usado para minimizar a função objetivo. A função objetivo que minimiza na j-ésima iteração que se quer adicionar a $f_j$ é dada na Equação seguinte:

$$\kappa^j = \sum_{i=1}^{n} \iota(y_i, \hat{y_i}^{j-1} + f_j(x_i) + \Omega(f_i). \tag{2.6}$$

Essa função 2.6 pode ser simplificada usando a expansão de Taylor. Então a fórmula pode ser derivada pela redução de perda depois da cortar a árvore. Por fim são resultadas nas funções $g_i$ e $h_i$ a seguir:

$$g_i = \partial_{\hat{y_i}^{j-1}} . \iota(y_i, \hat{y_i}^{j-1}), \tag{2.7}$$

$$h_i = \partial_{\hat{y_i}^{j-1}} . \iota(y_i, \hat{y_i}^{j-1}). \tag{2.8}$$

Existem três razões pelas quais o *XGBoost* tem um desempenho melhor do que outros métodos de aumento de árvores. Eles são:

- A introdução da função de perda regularizada;
- Os pesos de cada nova árvore podem ser reduzidos por uma constante $\eta$, o que reduz a influência de uma única árvore na pontuação final;
- A amostragem de coluna que funciona de maneira semelhante ao RF (PAN, ).

Além disso, ele é aceito por uma ampla variedade de aplicações, podendo ser usado para resolver problemas de regressão, classificação, ranqueamento e outros, funciona em várias linguagens (incluindo *python* que é a utilizada nesta pesquisa) e oferece suporte de vários ecossistemas de integração na nuvem, em caso de *deploy* do modelo.

#### 2.2.2.0.4 Regressão logística

A Regressão Logística (RL) é um algoritmo de classificação supervisionado, utilizado para estimar a probabilidade de uma instância pertencente a uma classe específica. Esse algoritmo tem três tipos, sendo eles: RL Binomial, RL Ordinal e RL Multinomial. O primeiro se refere a uma classificação de dois grupos ou categorias, o segundo é para três ou mais classes que têm uma ordem predeterminada e a terceira é para três ou mais categorias que não tem ordem entre elas. Geralmente ele é usado para classificação binária e sua análise resultante está contida num intervalo entre zero e um. (GÉRON, 2019).

A análise de regressão é um processo estatístico para estimar as relações entre as variáveis e é bem definida na tarefa de classificação binária de uma variável categórica. Desse modo, ela inclui muitas técnicas para modelar e analisar várias variáveis, quando o foco está na relação entre uma variável dependente e uma ou mais variáveis independentes. Como já dito, a Regressão Logística é um tipo de modelo de classificação estatística probabilística.

A Regressão Logística é muito semelhante à Regressão Linear, exceto pela forma como são usadas. A Linear é usada para resolver problemas de regressão, enquanto a Logística é usada para resolver os problemas de classificação e de regressão (por resultar de valores contidos no intervalo [0,1]).

Para tanto, as aplicações desse algoritmo são bem variadas. Pode-se usar para previsão de risco, identificar se um e-mail é spam ou não, ele é usado também para diagnóstico médico (se uma pessoa tem risco de ter certa doença ou não e sua probabilidade, por exemplo). Segmentação e categorização de imagens também é muito utilizado por esse algoritmo, além de previsão do tempo e vários outros.

A função Sigmoide ou função Logística é dada pela Fórmula 2.9:

$$\sigma(t) = \frac{1}{1+e^{-t}}. \tag{2.9}$$

Em que:

- $\sigma(t)$ é a probabilidade de que o *output* seja um número entre 0 e 1;
- t é um número real dado pela combinação linear dos atributos usados na previsão dado pela Equação (2.10) a seguir:

$$t = \theta_0 + \theta_1 X_1 + ... + \theta_n X_n. \tag{2.10}$$

Nessa Equação (2.10), quando $t \longrightarrow \infty$, então $\sigma(t) \longrightarrow 1$ e quando $t \longrightarrow -\infty$, então $\sigma(t) \longrightarrow 0$.

Para a variável de resposta resultar em 0 ou 1, existe o chamado *threshold* que nada mais é um limiar que irá definir a partir de qual valor ele será definido para a classe 1 e abaixo desse certo valor, ele será associado à classe zero. Normalmente esse valor é definido como 0,5 ou 50%, porém não existe regra para isso. A Equação (2.11) mostra o resultado requerido y:

$$y = \begin{cases} 0 \; if \; \sigma(t) < 0.5 \\ 1 \; if \; \sigma(t) \geq 0.5. \end{cases} \tag{2.11}$$

Esse limiar é interessante, pois consegue-se ver a capacidade da resposta em relatar a segurança ou incerteza associada à classificação em questão.

Para entender melhor como se chega ao resultado da Função Sigmoide, reescreve-se a função Sigmoide da Equações (2.9) e (2.10), resultando na Equação (2.12) a seguir:

$$\sigma(t) = \frac{1}{1 + e^{-(\theta_0 + \theta_1 X_1 + ... + \theta_n X_n)}}. \tag{2.12}$$

Pode-se reorganizar os termos e reescrever a Equação (2.12) para a Equação (2.13):

$$e^{\theta_0 + \theta_1 X_1 + ... + \theta_n X_n} = \frac{\sigma(t)}{1 - \sigma(t)} \Leftrightarrow \qquad \theta_0 + \theta_1 X_1 + ... + \theta_n X_n = ln\left(\frac{\sigma(t)}{1 - \sigma(t)}\right). \tag{2.13}$$

Mostrando com mais clareza, essa é a função inversa da função Logística, chamada também de Função *logit*. Nela conseguimos ver que, no domínio da probabilidade, tem-se a expressão da Equação (2.14):

$$odds(p) = \frac{p}{1 - p} \tag{2.14}$$

Que é a chance da ocorrência de um evento. Daí, vê-se o porquê o algoritmo de Regressão Logística resulta numa probabilidade. (GÉRON, 2019).

De fato, como foi mencionado anteriormente, a Regressão Logística usa a função Sigmoide para retornar a probabilidade de um rótulo. Ele mapeia qualquer valor real em outro valor dentro de um intervalo de 0 e 1. Para isso, uma função de custo é usada para calcular o erro e esta, por sua vez, é a diferença entre o valor previsto e o valor real. (GÉRON, 2019).

Tal função de custo encontra a função Sigmoide que se ajusta aos dados de treinamento para encontrar os parâmetros $\theta_n$ que minimizam os erros. Essa função de custo de todo

o conjunto de treinamento é o custo médio de todas as instâncias de treinamento. Ele pode ser escrito em uma única expressão chamada *Log Loss* ou ainda *Binary Cross Entropy*, mostrada na Equação (2.15) para um dado somente.

$$H_i = -(y_i.ln(\sigma(t)) + (1-y_i).ln(1-\sigma(t)) \tag{2.15}$$

E na Equação (2.16) para todo o conjunto de dados.

$$J(\theta) = -\frac{1}{m}\sum_{m}^{i=1} H_i = -\frac{1}{m}\sum_{m}^{i=1}[y_i.ln(\sigma(t)) + (1-y_i).ln(1-\sigma(t))] \tag{2.16}$$

Em que:

- $y_i$ é o valor de um rótulo/classe já conhecido;
- $\sigma(t)$ é a probabilidade do resultado.

Para minimizar a função de custo, pode-se usar o *Gradient Descent* que estima os parâmetros ou pesos do modelo. Além disso, se aumentar o número de épocas para treinar o modelo, a função de custo diminuirá para o valor ótimo.

Dessa forma, as vantagens para usar a RL são inúmeras, dentre elas é que ela tem facilidade em lidar com variáveis independentes categóricas e ela fornece os resultados em termos de probabilidade. Além disso, o modelo é relativamente fácil de implementar, interpretar e é rápido computacionalmente. Ademais, os coeficientes dão uma ideia de quão relevante um atributo é para a previsão, pois ele aplica "pesos" para cada atributo, informando qual atributo importa mais no modelo (ZOU *et al.*, ).

Todavia, existem algumas desvantagens e observações a serem feitas quando se usa Regressão Logística, dentre elas pode-se citar que se o número de observações for menor que os atributos, este algoritmo não deve ser usado, caso contrário pode levar a *overfitting* em que o modelo tem um ótimo resultado nos dados de treinamento, mas ele acaba não dando bons resultados nos dados de teste, isto é, ele apenas "decorou" os dados (SUR; CANDÈS, 2019).

Outra questão a se colocar é que problemas não lineares não podem ser resolvidos com RL, pois, como já foi mostrado, ele utiliza de uma função linear para ser resolvido. Além disso, precisa-se que as variáveis independentes sejam linearmente relacionadas ao *log* das probabilidades. Portanto, a RL não atribui uma relação linear entre as variáveis dependentes e independentes. Ela assume uma relação linear entre as probabilidades logarítmicas da variável dependente e as variáveis independentes, como foi mostrado na Equação (2.13).

Tendo em vista essas informações mencionadas sobre Regressão Logística, opta-se por usá-la no intuito de saber as probabilidades de uma mulher sofre ou não violência.

### *2.2.3 Medidas de avaliação de desempenho*

Existem várias Medidas de Avaliação de Desempenho. para avaliar os modelos de classificação no Aprendizado de Máquina. Dentre elas, existem duas divisões que são para dados balanceados e as Medidas de Avaliação de Desempenho. para dados desbalanceados. Para a primeira, pode-se metrificar os dados através da Matriz de Confusão, Erro do tipo I, Erro do tipo II e Acurácia. Já para a segunda, dados desbalanceados podem ser mensurados por Matrizes de Confusão, *F1-Score*, Erro do tipo I, Erro do tipo II, *recall* e Precisão.

Dito isso, a Matriz de Confusão é uma das técnicas de medição de desempenho do modelo mais populares e amplamente utilizadas para algoritmos de classificação. Nele observa-se o valor predito e o valor real, mostrando quantos dados foram calculados corretamente e erroneamente em cada classe.

Na matriz de confusão, pode-se identificar o verdadeiro positivo (*True Positive - TP*) e verdadeiro negativo (*True Negative - TN*) que são os que foram previstos corretamente para 0 e 1. O Erro do Tipo I são os rótulos que predizem para a classe 1, mas o certo seria para a classe 0, chamado de falso positivo (*False Positive - FP*). Já o Erro do Tipo II considera classe 0, mas na realidade seria classe 1, chamado de falso negativo (*False Negative - FN*) . Nos dados deste trabalho pode-se dizer que 1 corresponde a mulher ter sofrido violência e 0 caso contrário.

A matriz de confusão e as outras Medidas de Avaliação de Desempenho. estão intrinsecamente relacionadas, onde pode-se fazer os seguintes cálculos:

- Acurácia: (TP + TN)/total;
- Precisão: TP/(TP + FP);
- Recall: TP/(TP + FN);
- F1-score: 2.(precisão. recall)/(precisão + recall);
- Especificidade: TN/(TN + FP).

Em que TP é verdadeiro positivo, FP é falso positivo, TN é verdadeiro negativo, FN é falso negativo e o total é a soma de TP, FP, TN e FN.

A Acurácia diz quantos dos dados foram de fato classificados corretamente, porém ela não é um bom índice para dados desbalanceados, pois pode mostrar um resultado alto, mas o modelo ter uma performance inadequada porque ela atribui o mesmo peso para todos os erros. Já a Precisão dá ênfase maior para erros falso positivo, então dos classificados como positivo ele mostra quantos são realmente positivos.

Em contrapartida, o *Recall* é a sensibilidade ou também chamado de taxa verdadeiro

positivo. Ele dá maior ênfase para erros por falso negativo. Então, de todos os positivos ele mostra a taxa de quantos foram classificados corretamente como positivos. Enquanto isso, o *F1-Score* faz um resumo melhor da qualidade do modelo já que ele é a média harmônica entre a precisão e o *Recall*.

Por outro lado, a Especificidade avalia a eficácia do método em detectar resultados negativos, ou seja, a capacidade do sistema em predizer corretamente a falta da condição para dados que realmente não têm. Pode-se dizer que o *Recall* e a Especificidade variam em direções opostas.

# 3 PROBLEMA ABORDADO E MODELAGEM MATEMÁTICA

O presente problema se inicia na identificação dos dados, os quais foram recebidos pelo Programa de Pós-Graduação em Economia (CAEN/UFC) os questionários da PCSVDF-Mulher(Pesquisa de Condições Socioeconômicas e de Violência Doméstica e Familiar contra a Mulher - Survey of Socioeconomic Conditions and Domestic and Family Violence against Women) e alguns dos seus dados de respostas.

A Pesquisa de Condições Socioeconômicas e Violência Doméstica e Familiar contra a Mulher (PCSVDFMulher), realizada em parceria pelo Instituto Maria da Penha e a Universidade Federal do Ceará (UFC), é um estudo abrangente que visa compreender a violência de gênero e suas implicações socioeconômicas. Financiada pela Secretaria de Políticas Públicas para as Mulheres (SPM) e com apoio técnico-institucional do Banco Mundial, a pesquisa tem desempenhado um papel crucial na geração de conhecimento inédito e no desenvolvimento de políticas públicas para combater esse grave problema.

Apesar de a violência intrafamiliar ser mais comum em regiões em progresso e constituir um importante dilema socioeconômico e de saúde coletiva, diversos países ainda enfrentam a escassez de dados relevantes para compreender as origens e ramificações dessa violência, a fim de implementar políticas públicas mais eficazes com foco no bem-estar feminino (CARVALHO *et al.*, ). Isto é o caso do Brasil (o quinto país mais prevalente em DV no mundo) onde as melhores estatísticas sobre violência doméstica são escassas e remontam ao Estudo Multipaíses da Organização Mundial da Saúde de 2005 (ORGANIZAÇÃO MUNDIAL DE SAÚDE, 2021).

Dessa maneira, é apresentado um conjunto único de dados longitudinais sobre violência doméstica no Brasil: o PCSVDFMulher, que é um esforço interdisciplinar para construir evidências empíricas que permitam estudar a violência, alocação de recursos no domicílio, mulheres, saúde infantil e desenvolvimento infantil, e as inter-relações entre eles por meio de uma abordagem interdisciplinar, no qual o enfoque principal neste trabalho será de violência contra a mulher. O projeto reuniu informações de mais de 10000 mulheres de 15 a 49 anos que viveram nas capitais do nordeste do Brasil, em duas ondas: 2016 e 2017 (CARVALHO *et al.*, ).

Além dos dados relacionados à violência contra mulheres, o projeto disponibiliza informações acerca da saúde feminina, capacidade de negociação e recursos intrafamiliares, distribuição, costumes culturais e sociais, conhecimento sobre direitos civis e utilização de

medidas legais de proteção contra a violência doméstica, bem como dados sobre casais (por exemplo, educação, comportamento arriscado para a saúde, tonalidade da pele, status ocupacional, etc.), expectativas subjetivas das mulheres e crenças em relação ao bem-estar e ao abuso por parte do parceiro, entre outros aspectos (CARVALHO *et al.*, ).

De fato, os números mostram que cerca de 30% das mulheres sofreram violência doméstica (ou seja, violência emocional, física ou sexual) ao longo da vida, 14% relataram que aconteceu nos últimos 12 meses. Também em (CARVALHO *et al.*, ) mostra que a educação das mulheres é um fator preventivo na violência, enquanto o risco de sofrer abuso do parceiro está aumentando entre as mulheres jovens e não brancas no Brasil. Esses dados geraram relatórios importantes, revelando os diferentes aspectos da violência doméstica no seio das famílias (CARVALHO *et al.*, ).

Por meio de sua consultoria pedagógica, o Instituto Maria da Penha tem desempenhado um papel ativo na elaboração do questionário da pesquisa e na capacitação das entrevistadoras que coletam os dados de campo. Essa colaboração enriquece a qualidade e a abrangência da PCSVDFMulher, proporcionando uma compreensão mais aprofundada das questões relacionadas à violência doméstica.

A importância dessa pesquisa é inegável, uma vez que o banco de dados gerado pela PCSVDFMulher é inédito na América Latina. Sua amplitude informativa revela não apenas as violações dos direitos humanos, problemas de saúde e segurança pública, mas também as ramificações econômicas e sociais diretas da violência doméstica sobre as mulheres, suas famílias e o país como um todo. Com essa abordagem holística, a pesquisa contribui para o fortalecimento de iniciativas voltadas para a erradicação da violência de gênero e para a construção de uma sociedade mais justa e igualitária.

Em suma, a PCSVDFMulher se destaca como uma pesquisa pioneira, fornecendo subsídios valiosos para a comunidade científica e organizações envolvidas na luta contra a violência de gênero. Seu enfoque abrangente e sua abordagem longitudinal são fundamentais para o avanço do conhecimento e para a implementação de políticas públicas efetivas na prevenção e no combate à violência doméstica.

Para tanto, o problema se insere no contexto da violência contra a mulher, no qual foram feitos questionários entre várias mulheres de vários estados do Brasil, com o intuito de identificar possíveis relações com a violência contra a mulher e as respostas desses questionários.

A partir disso, no questionário encontram-se respostas na escala quantitativa e

qualitativa. Algumas dessas variáveis são:

- Localização em geral (cidade, estado, endereço, GPS);
- Quantas pessoas vivem nesta casa (incluindo homens e mulheres);
- Nome, idade, sexo, orientação sexual, raça, religião, escolaridade, tipo de domicílio, altura, peso;
- Relação com o chefe de família;
- Rua: brigas, assaltos, patrulhamento policial, coleta de lixo, iluminação, segurança noturna;
- Em relação aos parentes em geral;
- Se recebeu algum benefício do governo, pensão alimentícia ou doação, seguro-desemprego;
- Mercado de trabalho e horário de trabalho (tempo de trabalho, localização);
- Tarefas de casa;
- Saúde geral e reprodutiva;
- Se a mulher já sofreu algum tipo de violência e se ela tem conhecimento dos órgãos responsáveis para auxiliá-la;
- Se a mulher tem conhecimento sobre as leis que a cobrem sobre violência contra a mulher.

## 3.1 Proposta

Com essas informações, a proposta será de elaborar um modelo de Aprendizado de Máquina que consiga predizer qual a chance ou probabilidade de uma mulher ser vítima de seu companheiro ou do homem que mora com ela, com base nas suas respostas. Portanto, pretende-se estimar uma pontuação ou probabilidade de que uma mulher saiba se será violentada ou não e metrificar o quão boa essa mulher é em prever essas informações.

Tais dados coletados dizem à respeito de 9330 dados brutos de mulheres com 266 colunas que têm características dessas mulheres que sofreram ou não violência. Nesses dados, a mesma mulher se repete duas vezes, pois foram aplicados dois questionários com a mesma mulher, sendo um em 2016 e o outro em 2017.

Ademais, nesse projeto é proposto aplicar uma técnica de classificação que será possível predizer se um novo dado de outra mulher, isto é, aquela que não está nos dados pré estabelecidos, tem chance ou não (ou ainda de saber a probabilidade) de sofrer uma violência. É importante mencionar que a intenção é não especificar qual o tipo de violência, pois serão abordadas todas.

## 3.2 Metodologia

Para o presente trabalho é proposto a revisão bibliográfica a fim de fazer uma pesquisa qualitativa para buscar informações que ocorrem sobre a violência contra a mulher, ao mesmo passo em que também é feita uma abordagem quantitativa para fazer um estudo de caso e levantamento e saber dos resultados da amostra para constituir um retrato real de toda a população alvo desta pesquisa.

Para isso será utilizada da linguagem matemática e computacional para descrever com melhor precisão a avaliação do número de previsão de violência contra a mulher, usando de técnicas de Aprendizado de Máquina. Com relação aos classificadores, estão sendo utilizadas as seguintes abordagens: *Random Forest* e Regressão Logística.

Sua natureza será de pesquisa aplicada com o intuito de gerar conhecimentos para aplicação prática para o problema em específico. Quanto aos objetivos, a pesquisa é descritiva com a intenção de descrever características e padrões da problemática em questão.

Dito isto, é feito um levantamento de dados elaborado e executado pelo Programa de Pós-Graduação em Economia (CAEN/UFC) os questionários PCSVDF-Mulher e autorizada para essa pesquisa de fazer análises e modelos com os objetivos citados.

Para entender mais sobre os dados, buscou-se outra fonte de dados do EVA (INSTITUTO IGARAPÉ, ) para entender mais padrões das mulheres e buscar por mais estatísticas sobre o assunto de violência contra a mulher. Nele foram considerados vários dados novos acerca do problema em questão. Nesses novos dados tem atributos de localização, ano, ocorrência, tipo de violência, sexo, faixa etária, raça, arma, agressor, se teve violência doméstica, quantidade de casos, taxa e população.

Desses dados são encontrados 3 bases separadas, sendo elas oriundas da saúde e segurança (essa está separada em duas por ser uma por estado e outra por município). Então, opta-se por filtrar esses dados para colocar os anos de 2016 e 2017 e os municípios retratados na base de dados principal que é do PCSVDF-Mulher a fim de enriquecer mais o material recebido.

Dessa forma, a primeira parte é a análise exploratória dos dados a fim de encontrar percepções acerca do problema e onde se fazem visualizações e algumas estatísticas para entender melhor a temática da pesquisa.

Após analisar o conjunto de dados, a próxima etapa é a de realizar o pré-processamento. Nele é encontrado o desafio de lidar com dados faltantes. Percebe-se que de todos os dados, pelo menos um tem algum atributo nulo e apenas excluir não será o suficiente. Outro desafio

encontrado é sobre os dados estarem desbalanceados. De início não é feito tal tratamento nos dados, mas depois de observar os resultados com os modelos, provavelmente seja necessário balancear. Após esse tratamento, observam-se alguns *insights* em torno dos dados sobre os comportamentos das respostas do questionário.

Após aplicar todas as técnicas no conjunto de dados baixado, o algoritmo de Regressão Logística e o de *Random Forest* são utilizado nos dados. Nesta fase, os resultados são apresentados e comparados com os resultados anteriores do experimento. Se houver alguma melhoria, a decisão final será baseada na combinação dos hiperparâmetros que possuir melhor precisão no conjunto de dados fornecido.

Por fim, os modelos são verificados e avaliados com índices de classificação para comparar suas técnicas e parâmetros para análise dos resultados obtidos.

# 4 EXPERIMENTOS E RESULTADOS

Este Capítulo trata-se da execução de fato do modelo para fazer análise exploratória dos dados a fim de encontrar algum padrão e predizer se a mulher será violentada ou não com base nos atributos concedidos. Nele serão descritos todos os processos feitos, bem como seus resultados e avaliações.

## 4.1 Pré-processamento

Inicialmente, os arquivos do PCSVDF-Mulher são baixados, o primeiro com 9330 linhas e 16 colunas e o segundo com 254 novos atributos e as mesmas 9330 linhas. Os arquivos foram recebidos separadamente, então foram feitas análises exploratórias tanto nos arquivos separados como juntos.

Na etapa do pré-processamento dos dados, é observado que os dados não estão balanceados usando a característica "violence" que significa se a mulher sofreu qualquer tipo de violência pelo menos alguma vez na vida. Usando esse atributo, observa-se que tem 1807 dados para sim (valor 1 do *output*) e 3517 no caso contrário (valor 0 do *output*). Além disso, nessa mesma informação tem 4006 dados nulos (quase metade dos dados recebidos). Já analisando a característica "violence12" que significa se a mulher sofreu qualquer tipo de violência nos últimos 12 meses, tem 766 dados para sim (valor 1 do *output*) e 4390 no caso contrário (valor 0 do *output*). Além disso, nesse mesmo atributo tem 4174 dados nulos. Essas duas variáveis são as possíveis respostas para os modelos futuros.

Dito isso, constata-se que nesse conjunto de dados contém muitos dados nulos. Nele é verificado que de cada dado pelo menos uma característica tem valores faltantes, portanto não basta somente excluir todos os dados nulos. Isso ocorre porque no questionário, dependendo da pergunta anterior, as questões posteriores podem ser modificadas, portanto cada questionário é personalizado de acordo com a pessoa entrevistada.

Dessa forma, opta-se, inicialmente, por usar a característica "violence12" como *output* do modelo, então são excluídas todas as linhas que estão ausentes dessa coluna em específico, sobrando então 5156 linhas do *dataset*.

Após isso, são excluídas algumas colunas que contém formatações diferentes de tempo e pouco importantes nesse momento inicial com dados do tipo categóricos em numéricos no mesmo atributo. Em seguida, é feita a Correlação de Pearson com o *target* do modelo. Esta

correlação mede a ligação que dois atributos podem ter, estando no intervalo fechado de [-1,1], sendo quanto mais próximo de 1 resulta em alta correlação, enquanto mais próximo de zero corresponde a baixa correlação e quanto mais próximo de -1 tem alta correlação, porém negativa, ou seja, sempre que uma aumenta a outra diminui.

Ademais, existe uma interpretação para o resultado da correlação: acima de 0,7 é considerado correlação forte, entre 0,5 e 0,7 é considerado correlação moderada. Já no intervalo de 0,3 a 0,5 indica uma correlação fraca e entre 0 e 0,3 indica correlação ínfima, portanto, é desconsiderada a presença de correlação. O mesmo ocorre para o intervalo de -1 a zero, porém invertido.

Com essa informação, observa-se que altas correlações, em sua maioria, são com os atributos referentes a violência. Portanto, ainda não é possível observar se essa variável de resposta tem alguma relação com alguma característica que não seja relacionado à violência.

Além do problema ter muitos dados nulos, nos dados categóricos contém "Não sabe/ Não respondeu" como opção (valor: 88888). Dito isto, dos dados que sobraram que tiver essa opção e ainda ter dados nulos são colocados os valores faltantes como "não sabe/ não respondeu" a fim de preencher o *NaN* dos dados. A partir daí, ainda sobram alguns atributos que contém algum valor faltante e é optado por excluir os atributos que contiver pelo menos 25% de dados nulos, ficando então de 264 para 188 colunas no total, restando 5 atributos que ainda têm valores faltantes. Excluindo os dados que ainda contém ausência de dados, sobram, no final, 4940 linhas.

Posto isso, é feito um *ranking* de correlação com o *target* "violence12". Sabendo que acima de 0,7 há correlação alta, o único atributo que teve este tipo de correlação foi a que pergunta: "Insultou você ou te fez sentir mal consigo mesma? Nos últimos 12 meses." (retirado do questionário), com 6 respostas, entre elas:

- (1) Nunca
- (2) Raramente
- (3) Às vezes
- (4) Frequentemente
- (5) Sempre
- (88888) Não sabe/Não respondeu.

A maioria dos dados categóricos tem essa configuração de resposta listada acima. Assim, na Tabela 1 vê-se o *ranking* dos atributos e suas definições. Vale ressaltar que nessas tabelas com *ranking* de correlações, as correlações estão com valor absoluto.

Tabela 1 – Correlação com o Target "violence12" depois de excluir dados ausentes

| **Atributo** | **Definição** | **Correlação** |
|---|---|---|
| violence12 | Target - se ela sofreu violência física (últ. 12m) | 1,000000 |
| q708_c1 | Insultou ela ou a fez se sentir mal (últ. 12m) | 0,820917 |
| violence | Se ela já sofreu alguma violência na vida | 0,601081 |
| q738 | Se ela já agrediu o parceiro quando ele não a estava agredindo | 0,575927 |
| vio_fis_12 | Se ela sofreu violência física nos últimos 12 meses | 0,568273 |
| q708_c4 | Fez coisas para assustá-la ou intimidá-la de propósito (últ. 12m) | 0,555180 |
| q708_c3 | Menosprezou humilhou ela na frente de outras pessoas (últ. 12m) | 0,534442 |
| q708_c2 | Menosprezou humilhou ela na frente da família dela (últ. 12m) | 0,516857 |
| q709_c1 | Deu um tapa ou jogou algo que poderia machucá-la (últ. 12m) | 0,446567 |
| vio_fis | Se ela já sofreu alguma violência física na vida | 0,356206 |

Fonte: Elaborado pela autora

Outra consideração a se fazer é sobre os atributos que foram retirados do *dataset*, por conterem muitos dados nulos (cerca de 80%). Tais dados dizem respeito ao que ocorreu, como por exemplo: "O seu PARCEIRO ATUAL, EX-PARCEIRO (MAIS RECENTE) OU QUALQUER OUTRO EX-PARCEIRO já..."(retirado do questionário) completando as perguntas do questionário como no parágrafo anterior (se já insultou, ou algo da mesma variedade). Visto que essas são perguntas interessantes do questionário a fim de identificar padrões, porém com muitos dados nulos, optou-se por antes de excluir as linhas e colunas, fazer uma correlação com tais dados. Tal correlação pode ser vista na Tabela 2.

Tabela 2 – Correlação com o Target "violence12" antes de excluir dados ausentes

| Atributo | Definição | Correlação |
|---|---|---|
| q726_b10 | Pequenos cortes, furos,mordidas (últ. 12m) | 1.000000 |
| violence12 | Target - se ela sofreu violência física (últ. 12m) | 1,000000 |
| q708_b1 | Vio. Emocional - Insultou ela ou a fez se sentir mal (últ. 12m) | 0,913137 |
| q708_b4 | Fez coisas para assustá-la ou intimidá-la de propósito (últ. 12m) | 0,864937 |
| q708_b2 | Menosprezou humilhou ela na frente da família dela (últ. 12m) | 0,857967 |
| q708_b3 | Menosprezou humilhou ela na frente de outras pessoas (últ. 12m) | 0,850180 |
| q726_b5 | Ferimentos profundos,cortes (últ. 12m) | 0,843274 |
| q708_b5 | Ameaçou ferí-la ou ferir alguém importante pra ela | 0,826741 |
| q726_b6 | Ruptura do tímpano, lesões oculares (últ. 12m) | 0,786796 |
| q711_b3 | Forçou a fazer algo durante uma relação sexual | 0,756913 |
| q726_b8 | Dentes quebrados (últ. 12m) | 0,745356 |
| q726_b11 | Perda da consciência (últ. 12m) | 0,728869 |
| q709_b3 | Deu um soco ou fez algo que poderia machucá-la (últ. 12m) | 0,717317 |
| q709_b2 | Empurrou-a ou puxou seu cabelo (últ. 12m) | 0,712815 |
| q711_b2 | Teve relação sexual com medo,mas sem violência(últ. 12m) | 0,712775 |
| q709_b5 | Estrangulou-a (últ. 12m) | 0,710905 |
| q709_b7 | Ameaçou usar uma arma contra ela (últ. 12m) | 0,710179 |
| q709_b1 | Deu um tapa ou jogou algo que poderia machucá-la (últ. 12m) | 0,698658 |
| q709_b8 | Chegou a usar uma arma contra ela (últ. 12m) | 0,688165 |

Fonte: Elaborado pela autora

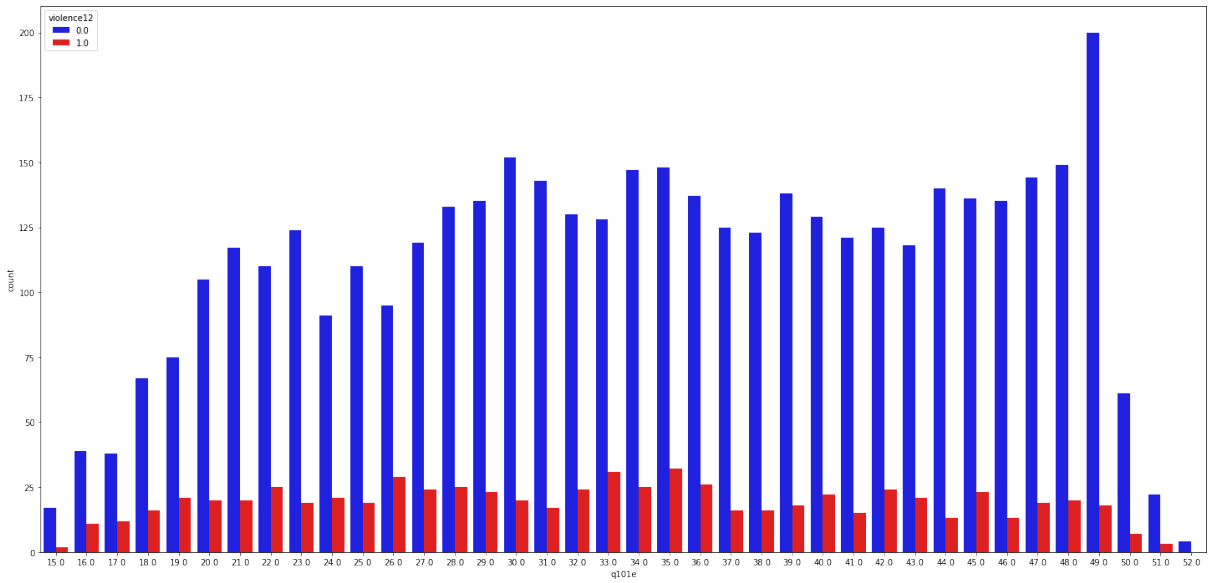
Dito isso, mesmo contendo poucos dados realmente respondidos que possam contri-

buir com esse trabalho, tais *insights* ajudam a perceber que existem mulheres que não responderam (deixaram nulo) a pergunta sobre se ela já sofreu violência, mas que respondeu que já sofreu algum tipo de violência quando especificado no questionário.

Considerando os atributos da Tabela 2 acima, das mulheres que responderam "sim" para algum desses tipos de violência, 17 deles contém NaN para o *target* "violence12" e consequentemente "violence". Percebe-se que ela já sofreu ou sofre algum tipo de violência, mas por algum motivo (seja de medo ou por não achar que era de fato violência, ou outro caso) ela não respondeu. Apesar de ser interessante colocar neste trabalho que existem dados dessa natureza não respondidos, tais dados foram excluídos por estar faltante na variável de resposta e não foi optado por substituir o *NaN* por 1, por questões éticas.

Depois de ter algumas percepções acerca dos dados, foram plotados alguns gráficos para ver a distribuição dos dados. Um desses gráficos é a observação do *target* "violence12" na distribuição da idade das entrevistadas como mostrada na Figura 1.

Figura 1 – Distribuição da violência por idade

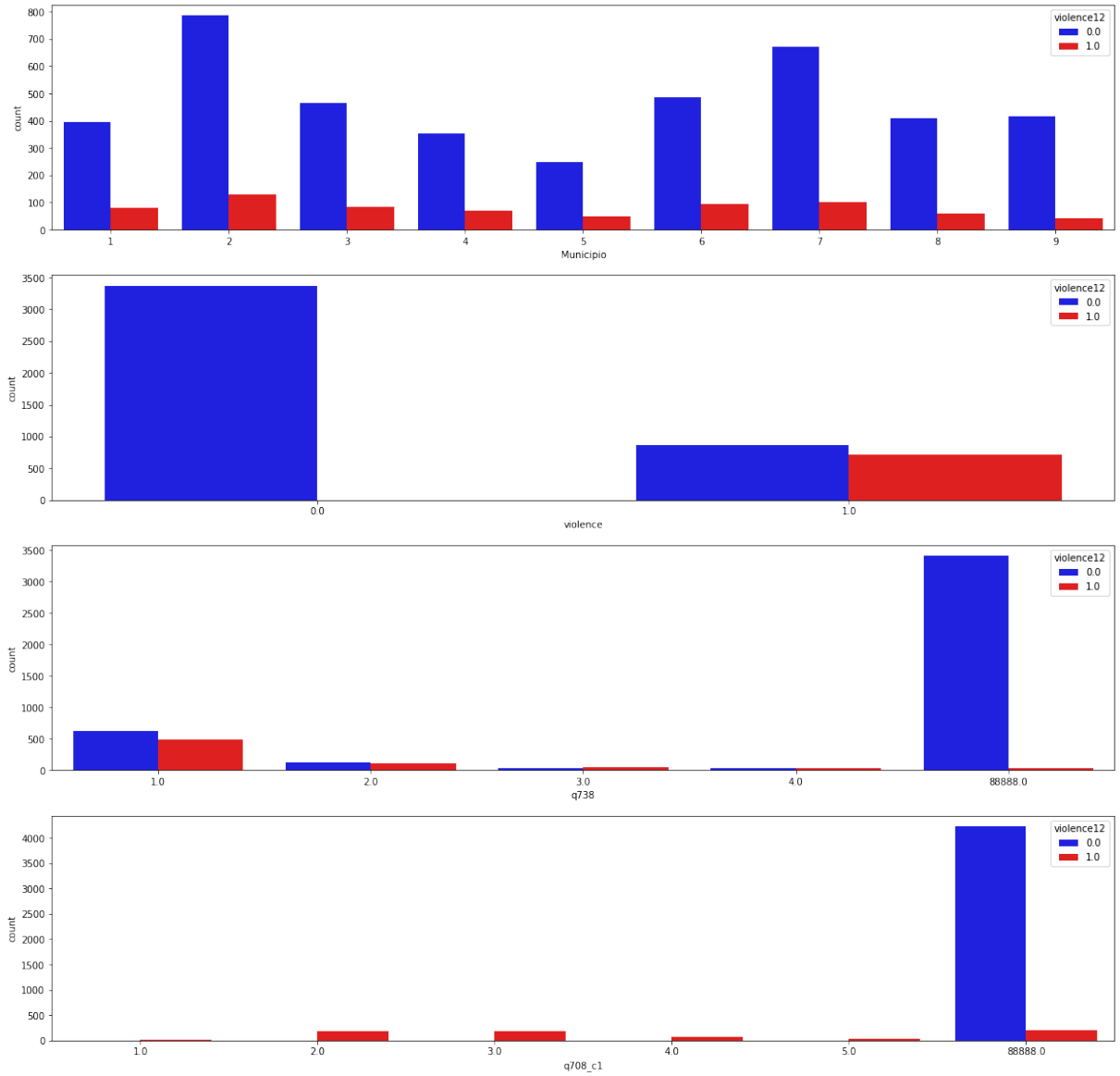



Fonte: Elaborado pela autora com as Bibliotecas *Seaborn* e *Matplotlib* do *Python*

As barras em vermelho indicam 1, respondendo que foram violentadas nos últimos 12 meses e em azul indicam 0, respondendo que não foram violentadas nos últimos 12 meses. Com isso, percebe-se que em quase todas as idades mencionadas nas entrevistas, exceto as de 52 anos, tiveram registro de violência nos últimos 12 meses e pode-se dizer que os dados que informaram 1 (em vermelho) estão bem distribuídos ao longo das idades.

Outros gráficos foram plotados na tentativa de ter mais *insights* acerca dos dados como colocado na Figura 2 a seguir.

Figura 2 – Distribuição da violência pelos atributos




Fonte: Elaborado pela autora com as Bibliotecas *Seaborn* e *Matplotlib* do *Python*

Com esses gráficos, observa-se que sobre no atributo "Municipio", Fortaleza (2), Recife (6) e Salvador (7) registram mais violência (*target* "violence12" de classe 1), mas tem muitos dados no caso contrário também (*target* "violence12" de classe 0). Nota-se também que o atributo "violence" é muito parecido com "violence12" em relação à classe 1, mostrando que nos últimos 12 meses (de 2016 para 2017) houveram muitos casos de violência contra a mulher comparado aos anos anteriores e essa informação é bem significativa.

Em relação ao atributo "q738" que pergunta:"Você já agrediu seu parceiro quando ele não a estava agredindo?" está bem distribuída ao longo das respostas (1 para "Nunca", 2 para "Uma/Duas vezes", 3 para "Várias (3-5) vezes", 4 para "Muitas (mais de 5) vezes"e 88888 para "Não sabe/Não respondeu"), exceto a resposta 88888, em relação às classes de violência.

Pode-se dizer que não afetam os valores indicar se a mulher agride o marido ou não, ela está sendo violentada da mesma forma (no caso da classe 1).

Sobre o atributo "q708_c1" que significa: "Insultou você ou te sentir mal consigo mesma?" há mais violência (classe 1) do que não-violência (classe 0) ao longo das respostas, exceto na resposta 88888.

Após algumas visualizações, é feita a separação dos dados em treino, teste e validação para a construção dos primeiros modelos usando os algoritmos mencionados no Capítulo 2. Nessa parte foram separados os dados de 2016 e 2017, sendo os dados de 2016 e início de 2017 somente para treino. Para os dados de teste e validação foram somente os dados restantes de 2017, ficando assim 3270 dados para treino, 835 dados para teste e 836 dados para validação. Além disso, foram retirados os atributos "violence12", "year" e "ID" do X_train, X_test e X_valid e "year" e "ID" do y_train, y_test e y_valid.

## 4.2 Modelos da 1ª fase: dados socioeconômicos e de violência

Nessa etapa são feitos alguns modelos com o intuito de identificar como os algoritmos estão performando e encontrar padrões nesses exemplos. Nesta primeira fase são feitos 5 modelos, testando diferentes técnicas para um melhor entendimento dos dados. Ao fim será observado uma certa dúvida sobre o resultado dos dados ser incomum.

### *4.2.1 Primeiro modelo: usando todos os atributos*

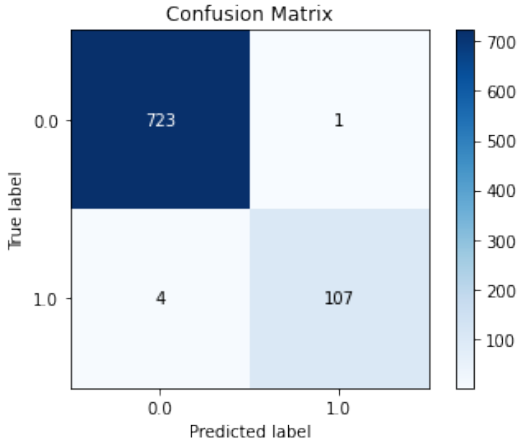
A partir da separação dos dados, a primeira aplicação do algoritmo de Regressão Logística foi aplicada. Para isso, usou-se os dados separados como explicado anteriormente, usando todos os atributos e feito o redimensionamento dos dados para o intervalo fechado [0, 1]. Dessa forma, foi usada a biblioteca *Sklearn* (PEDREGOSA *et al.*, 2011) e utilizados os seguintes parâmetros:

- penalty='l2',
- dual=True,
- verbose = 1,
- random_state= 42,
- solver='liblinear',
- C=5.0.

Dito esses parâmetros, foi optado usar a penalidade L2, pois a L1 penaliza a soma dos valores absolutos dos pesos, enquanto a L2 penaliza a soma dos quadrados dos pesos. Outro fato a ser mencionado é que a solução da penalidade L1 é esparsa, enquanto a da de penalidade L2 não é. Além disso, a regularização L2 não executa a seleção de recursos, visto que os pesos são reduzidos apenas a valores próximos a 0 em vez de 0, enquanto a regularização L1 possui seleção de recursos incorporada. Ademais, a regularização L1 é robusta para *outliers*, em contrapartida a regularização L2 não é. Já o parâmetro *dual=True* é feito, pois foi escolhida a penalidade L2 e com o solucionador *Liblinear* que também suporta a penalidade L2.

Após algumas tentativas, os resultado desse modelo gira em torno de 99% para dados de treino e 99% para os dados de teste. Além disso, sua acurácia é em torno de 99% também e plotando a matriz de confusão ele só erra 4 dados indicando zero quando é um e 1 dado indicando 1 quando é zero, como mostrado na Figura 3 a seguir.

Figura 3 – Matriz de confusão para o primeiro teste




Fonte: Elaborado pela autora com a biblioteca *Scikitplot* do *Python*

Além disso, a precisão também ficou em torno de 99%, a sensibilidade ficou em na faixa de 96% e o *F1-Score* ficou em torno de 98%. Esses resultados são melhores observados na Tabela 3 a seguir:

Tabela 3 – Sumarização da Classificação com Regressão Logística

| | precisão | sensibilidade | f1-score | support |
|---|---|---|---|---|
| 0.0 | 0.99 | 1.00 | 1.00 | 724 |
| 1.0 | 0.99 | 0.96 | 0.98 | 111 |
| accuracy | | | 0.99 | 835 |
| macro avg | 0.99 | 0.98 | 0.99 | 835 |
| weighted avg | 0.99 | 0.99 | 0.99 | 835 |

Fonte: Elaborado pela autora

Como mostrado na Tabela 3 e explicado anteriormente, a acurácia é um índice que indica a proporção de dados que foram classificados corretamente. No entanto, ela pode ser enganosa quando os dados estão desbalanceados, pois não leva em consideração o peso dos erros.

A precisão enfatiza os erros de falso positivo, ou seja, quantos dos dados classificados como positivos são realmente positivos.

A sensibilidade, também conhecido como taxa verdadeiro positivo, enfatiza os erros de falso negativo. Ela mostra a proporção de dados positivos que foram classificados corretamente como positivos.

O F1-Score é uma medida que combina a precisão e a sensibilidade em uma único índice, fornecendo um resumo da qualidade do modelo. É calculado como a média harmônica entre a precisão e a sensibilidade.

A Especificidade avalia a capacidade do modelo em detectar resultados negativos corretamente, ou seja, a capacidade de prever corretamente a falta de uma condição quando ela não está presente nos dados. A sensibilidade e a especificidade variam em direções opostas.

Esses índices são utilizadas para avaliar diferentes aspectos da performance de um modelo de classificação, levando em consideração os tipos de erros que são mais relevantes para o problema em questão.

Sobre a classe 0, a precisão é de 0,99, significando que quando o modelo prevê que a mulher não será violentada, 99% das vezes está correto. Já a sensibilidade é 1,0, ou seja, significa que o modelo prevê corretamente que 100% das mulheres não sofrerão violência. A pontuação F1 é 100% também.

Sobre a classe 1, a precisão é 1,0, ou seja, significa que quando o modelo prevê que que a mulher será violentada, 100% das vezes está correto. A sensibilidade é de 0,95, ou seja, significa que o modelo prediz corretamente que 95% das mulheres sofrerão violência. A pontuação F1 é de 97% e a precisão deste modelo é de 99%. A pontuação F1 é de 98% para todo o modelo (macro avg).

Tendo em vista esses resultados mencionados, percebe-se que esse valor é significativamente alto, então podemos pensar na possibilidade de ter dado *overfitting* apesar dos dados de teste terem dado alto *score*. Assim serão ajustados os parâmetros nos próximos modelos.

### *4.2.2 Segundo modelo: usando seleção de atributos*

A partir desse resultado é optado fazer a seleção dos melhores atributos que irá selecionar menos quantidade de atributos para a elaboração do modelo para observar se o *score* diminui.

A Seleção de atributos, também conhecida como seleção de variáveis, seleção de atributos ou seleção de subconjuntos de variáveis, é um processo de seleção de um subconjunto de recursos relevantes para uso na construção de modelos. Ele permite que o algoritmo de aprendizado de máquina treine mais rápido, reduzindo a complexidade de um modelo e facilita a interpretação. Ademais, ele melhora a precisão de um modelo se o subconjunto correto for escolhido.

Dessa forma, será também utilizada a biblioteca *Sklearn* (PEDREGOSA *et al.*, 2011) e importando *SelectKBest, SelectFromModel* e *RFE*. Com isso, são construídas 3 listas de melhores atributos para serem usados no modelo. Após, essas listas de atributos juntam-se em uma só para contar quantas vezes um único atributo foi contado e por fim é feito um *ranking* com o total de atributos. As variáveis selecionadas são aquelas que aparecem pelo menos duas vezes, ou seja, que foi mostrado em pelo menos duas listas. Por fim é selecionada uma lista contendo 22 atributos considerados mais importantes.

A lista dos 22 melhores atributos selecionados foram os seguintes:

- ’q101e’ - "Qual a idade da entrevistada?";
- ’q703_b1’ - "Você pode me dizer o que causou o desentendimento?";
- ’q705_b’ - "O seu %sitpat% bateu ou jogou algum objeto em você?";
- ’q702’ - "No seu relacionamento com o seu %sitpat% com que frequência você diria que vocês discutem/discutiam?";
- ’q704_b’ - "Seu %sitpat% gritou com você?";
- ’q706_d’ - "Seu %sitpat% a ignorou e/ou a tratou com indiferença";
- ’q707_d’ - "Ignorou e/ou tratou seu %sitpat% com indiferença?";
- ’q708_c1’ - "Seu %sitpat% a-Insultou você ou te fez sentir mal consigo mesma? - Nos últimos 12 meses,você diria que isto aconteceu:";
- ’q707_f’ - "Suspeitou de que o seu %sitpat% seja infiel a você?";
- ’q708_c2’ - "Seu %sitpat% b-Menosprezou ou te humilhou na frente da sua família? - Nos últimos 12 meses,você diria que isto aconteceu:";
- ’q708_c3’ - "Seu %sitpat% c-Te menosprezou ou te humilhou na frente de outras pessoas?

- Nos últimos 12 meses,você diria que isto aconteceu:";

- ’q708_c4’ - "Seu %sitpat% d-Fez coisas para te assustar ou te intimidar de propósito (ex: pela forma como ele te olhou, por gritar ou quebrar coisas) ? - Nos últimos 12 meses,você diria que isto aconteceu:";
- ’q709_c1’ - "Seu %sitpat% a-Te deu um tapa ou jogou algo em você que poderia machucá-la? - Nos últimos 12 meses,você diria que isso aconteceu:";
- ’q738’ - "Você já agrediu seu parceiro quando ele não a estava agredindo?",
- ’q726_b2’ - "Arranhões, abrasões, hematomas - Isso aconteceu nos últimos 12 meses?";
- ’q901_b’ - "Como se sente com relação a alguns aspectos da sua vida. Quão satisfeita você está com a sua vida afetiva e sexual?";
- ’q901_e’ - Como se sente com relação a alguns aspectos da sua vida. Quão satisfeita você está com o respeito e a cortesia que seus parentes e vizinhos demonstram por você?";
- ’q913_ph_verbal’ - "Qual é a chance (ou probabilidade) de você ser vítima de uma agressão física cometida pelo seu %sitpat% nos próximos 12 (doze) meses?",
- "seq_id_quest- "Código da entrevistada";
- ’vio_fis’ - "Já sofreu violência doméstica física em qualquer momento da sua vida?";
- ’vio_fis_12’ - "Já sofreu violência doméstica física nos últimos 12 meses?";
- ’violence’ - "Já sofreu violência doméstica (física, emocional ou sexual) em qualquer momento de sua vida".

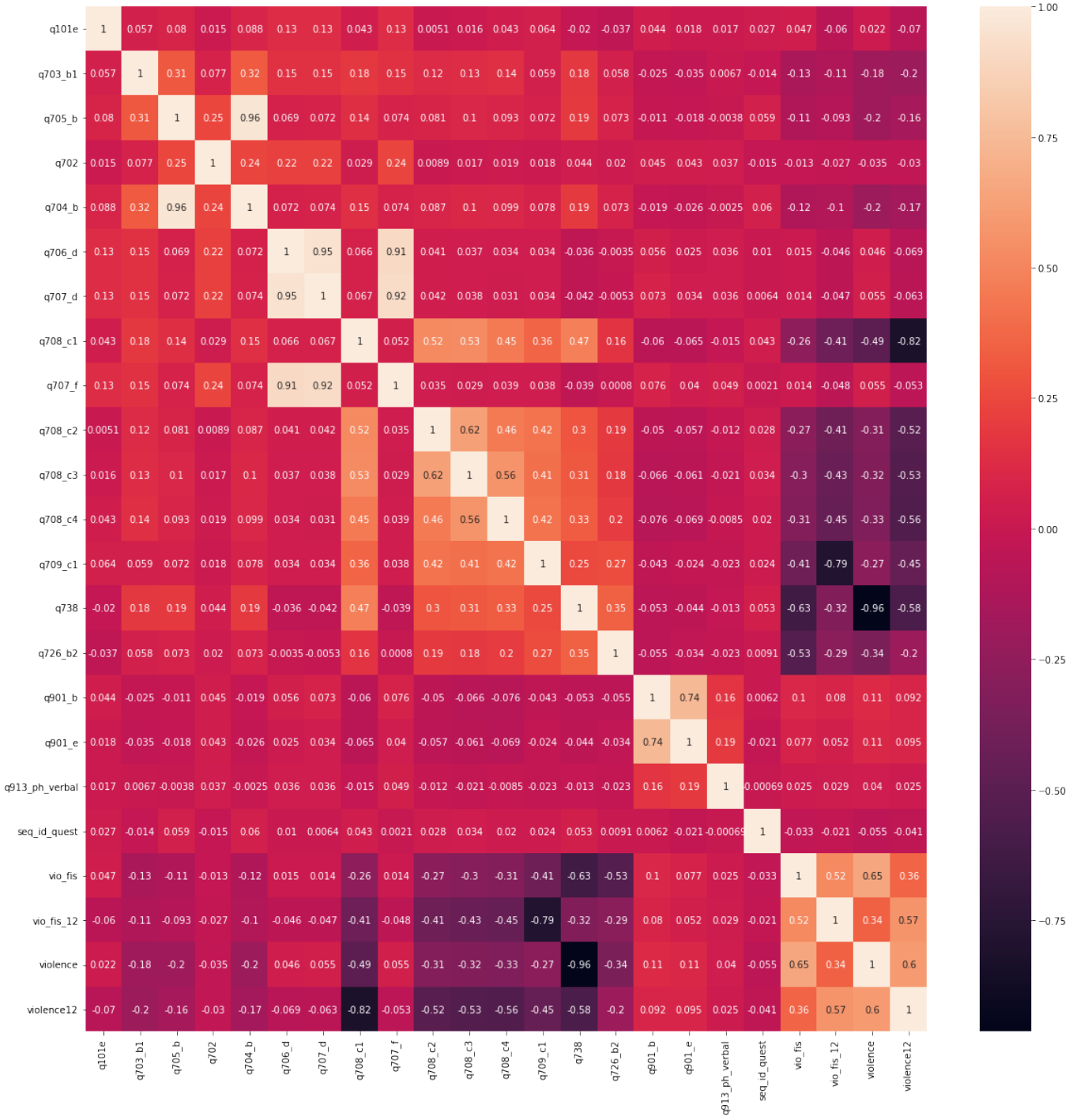
Sendo o "%sitpat%" o parceiro atual, ou o ex-parceiro (mais recente) ou ainda qualquer outro ex-parceiro da entrevistada.

Dessa forma, são feitas análises, exploração e visualização dos dados com os 22 atributos. A partir daí é feita a correlação de Pearson com os atributos na Figura 5 a seguir. Dito isso, os tipos de correlação de interesse desse trabalho serão as correlações fortes e moderadas.

As correlações mostradas na Figura 5 que são consideradas fortes e moderadas são:

- ’q705_b’ e ’q704_b’ (r = 0.96);
- ’q706_d’ e ’q707_d’ (r = 0.95);
- ’q706_d’ e ’q707_f’ (r = 0.91);
- ’q707_d’ e ’q707_f’ (r = 0.92);
- ’q901_e’ e ’q901_b’ (r = 0.74);
- ’violence’ e ’vio_fis’ (r = 0.65);
- ’q708_c3’ e ’q708_c2’ (r = 0.62);

Figura 4 – Correlação dos 22 atributos da Seleção de atributos




Fonte: Elaborado pela autora

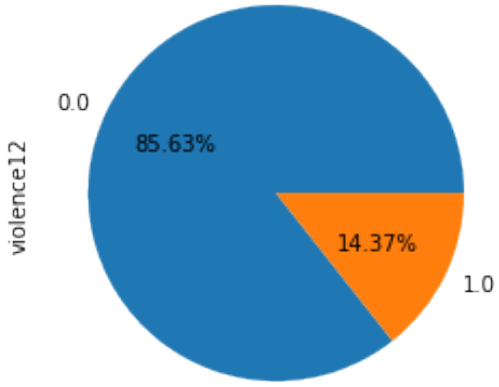
- ’violence12’ e ’violence’ (r = 0.6);
- ’vio_fis_12’ e ’violence12’ (r = 0.57);
- ’q708_c3’ e ’q708_c4’ (r = 0.56);
- ’vio_fis_12’ e ’vio_fis’ (r = 0.52);
- ’vio_fis’ e ’q736_b2’ (r = -0.53);
- ’violence12’ e ’q708_c3’ (r = -0.53);

- ’violence12’ e ’q708_c4’ (r = -0.56);
- ’violence12’ e ’q738’ (r = -0.58);
- ’vio_fis’ e ’q738’ (r = -0.63);
- ’vio_fis_12’ e ’q709_c1’ (r = -0.79);
- ’violence12’ e ’q708_c1’ (r = -0.82);
- ’violence’ e ’q738’ (r = -0.96)

Visto esses itens acima, é curioso observar as correlações negativas, principalmente com o *target* do modelo. O atributo que pergunta se a entrevistada já agrediu alguma vez seu parceiro, mostra que quanto menos ocorrências de agressão com o parceiro maiores são as chances dela sofrer violência e vice-versa. Outra correlação negativa interessante de se observar com o *output* do modelo é com o atributo ’q708_c1’ que pergunta se o "%sitpat%" insultou ou a fez sentir mal consigo mesma nos últimos 12 meses, indicando que essa é a mais importante correlação negativa com ’violence12’. Outras correlações fortes e moderadas com o *target* são com os atributos: violence (r = 0.6), q738 (r = - 0.58), vio_fis_12 (r = 0.57), q708_c4 (r = -0.56), q708_c3 (r = -0.53) e q708_c2 (r = -0.52).

Além da correlação, é plotada a porcentagem de dados em relação à variável ’violence12’ para ver a porcentagem da contagem de dados da classe 1 e classe 0. Esse resultado é mostrado no gráfico circular da Figura a seguir.

Figura 5 – Gráfico circular com a porcentagem de cada classe do target ’violence12’



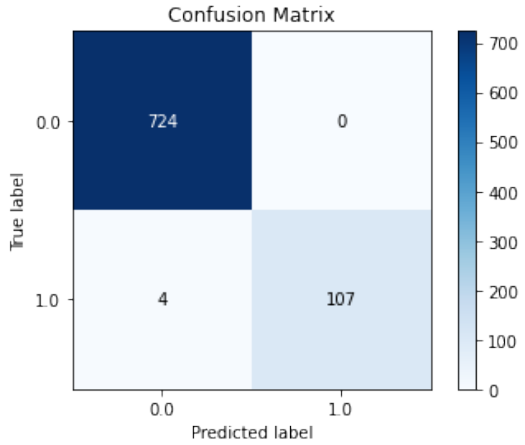

Fonte: Elaborado pela autora com a biblioteca *Matplotlib* do *Python*

Com esse gráfico observa-se que a variável de saída é bem desbalanceada em relação à classe 1 por ter muito menos do que a classe 0. Porém, como mostrado nos modelos anteriores, os resultados dos algoritmos não erraram, por isso não se optou por fazer *undersampling* (remover

os dados com maior quantidade da classe de maiores dados) ou *oversampling* (aumentar os dados com menor quantidade da classe de menores dados).

A partir disso, roda-se o modelo de Regressão Logística e *Random Forest* somente com esses 22 atributos mais a variável de saída. Com esses dois algoritmos, o F1-score dos dois resultou em torno de 98% mostrando que não é necessário ter mais de 200 atributos, como no início dos experimentos para se obter um bom resultado. A matriz de confusão da Figura 6 mostra o quanto os dois modelos acertaram e erraram.

Figura 6 – Matriz de Confusão para o segundo teste (Regressão Logística e *random Forest*)




Fonte: Elaborado pela autora com a biblioteca *Scikitplot* do *Python*

Com esse resultado, observa-se que a matriz de confusão e os índices são exatamente iguais para os algoritmos de Regressão Logística e *Random Forest*, por isso na legenda da Figura 6 estão as duas técnicas. Além disso, constata-se que além do resultado quase permanecer o mesmo do modelo rodado com todos os atributos, diminuir o número de perguntas torna o questionário menos invasivo se o objetivo deste for somente predizer se a mulher será violentada ou não. Além disso, os *scores* estão listados na Tabela 4 a seguir.

Tabela 4 – Sumarização da Classificação com Regressão Logística e *Random Forest* para 22 atributos

| | precisão | sensibilidade | f1-score | support |
|---|---|---|---|---|
| 0.0 | 0.99 | 1.00 | 1.00 | 724 |
| 1.0 | 1.00 | 0.96 | 0.98 | 111 |
| accuracy | | | 1.00 | 835 |
| macro avg | 1.00 | 0.98 | 0.99 | 835 |
| weighted avg | 1.00 | 1.00 | 1.00 | 835 |

Fonte: Elaborado pela autora

Como os dados selecionados na Seleção de atributos são, em sua maioria, relacionados diretamente com a violência, é interessante observar se com os dados que não tem tanta

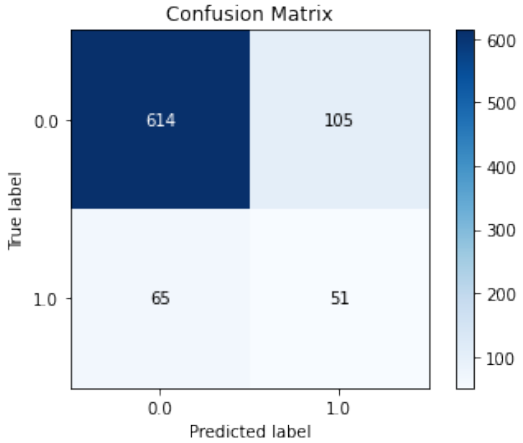
relação com a violência (por exemplo: idade, raça, educação e outros) o modelo irá acertar tanto quanto com os dados que são majoritariamente de violência.

### *4.2.3 Terceiro modelo: sem seleção de atributos*

Tendo em vista do modelo estar 'acertando' demais com as *atributos* selecionadas, testa-se o modelo sem os 22 atributos selecionados na Seleção de atributos, com 166 atributos e retirando o *target*, ano e ID, sobram 163 características. Nele, é feito um *ranking* com os atributos que sobraram para ver a correlação com a variável de entrada e percebe-se que nenhum atributo tem correlação alta ou moderada, mesmo que negativa. A maior correlação seria com o atributo 'q727' que pergunta: "Ao longo de sua vida,você alguma vez já foi gravemente ferida por parceiro atual, ex-parceiro (mais recente) ou qualquer outro ex-parceiro a ponto de precisar de cuidados médicos (mesmo que não os tenha recebido)?", que continua sendo sobre violência, mas que tem 4818 dados com o valor 88888 que significa "Não sabe/ Não respondeu" dos 4990 dados totais.

O modelo em questão é treinado com os algoritmos de Regressão Logística e *Random Forest* resultando em torno de 56% de *F1-Score* para a Regressão Logística e 63% de *F1-Score* para o *Random Forest*. Com isso, na Figura 7 pode-se observar a matriz de confusão dos resultados obtidos pelo algoritmo de Regressão Logística.

Figura 7 – Matriz de Confusão para o terceiro teste (Regressão Logística)




Fonte: Elaborado pela autora com a biblioteca *Scikitplot* do *Python*

Observa-se nessa matriz de confusão da Figura 7 que o algoritmo continua acertando bastante para a classe 0 (valor de 614 dados), mas já aparecem mais erros (valor de 65). Além disso, erra bem mais para a classe 1 (valor de 105 dados). Na Tabela 5 pode-se ver os resultados

dos índices obtidos.

Tabela 5 – Sumarização da Classificação com Regressão Logística e para 166 atributos

| | precisão | sensibilidade | f1-score | support |
|---|---|---|---|---|
| 0.0 | 0.90 | 0.85 | 0.88 | 719 |
| 1.0 | 0.33 | 0.44 | 0.37 | 116 |
| accuracy | | | 0.80 | 835 |
| macro avg | 0.62 | 0.65 | 0.63 | 835 |
| weighted avg | 0.82 | 0.80 | 0.81 | 835 |

Fonte: Elaborado pela autora

Tendo em vista o terceiro modelo rodando sem os dados da Seleção de atributos é possível observar que os atributos relacionados a violência realmente tem muita importância principalmente com a classe 1 da variável de resposta. É relevante mencionar que esse modelo foi feito apenas para observações dos resultados e ter mais *insights* sobre o que pode-se ser feito em seguida.

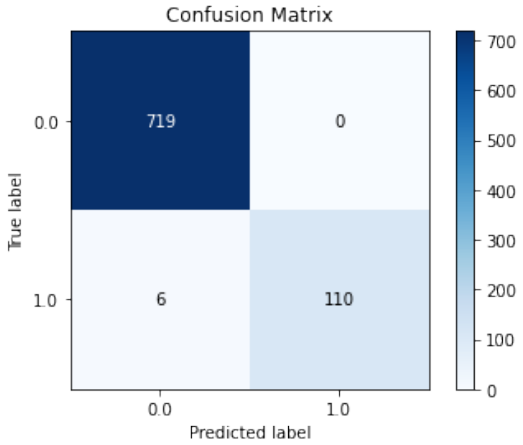
### *4.2.4 Quarto modelo: usando somente os atributos de forte e moderada correlação*

Após os resultados do terceiro modelo, opta-se por fazer outro teste, agora somente com os dados de correlação forte e moderada, sobrando então 7 *atributos* e 1 *target* sendo essas variáveis as que tem correlação acima de 0,5, sendo elas:

- 'violence12': Se a entrevistada já sofreu algum tipo de violência nos últimos 12 meses(*target*);
- 'q708_c1': a-Insultou fez sentir mal consigo mesma? Nos últimos 12 meses;
- 'violence': Se a entrevistada já sofreu algum tipo de violência alguma vez na vida;
- 'q738':Se já agrediu o parceiro quando ele não a estava agredindo (correlação negativa);
- 'vio_fis_12': Se a entrevistada já sofreu violência física nos últimos 12 meses;
- 'q708_c4': d-Fez coisas para assustar ou intimidar de propósito (ex: pela forma como ele te olhou, por gritar ou quebrar coisas) - Nos últimos 12 meses;
- 'q708_c3': c- menosprezou ou humilhou na frente de outras pessoas - Nos últimos 12 meses;
- 'q708_c2': b-Menosprezou ou humilhou na frente da sua família - Nos últimos 12 meses.

O modelo é treinado com as técnicas de Regressão Logística e *Random Forest*. Com a Regressão Logística, pode-se ver na Figura 8 que o modelo ainda continua muito bom, diminuindo de 22 para 7 atributos

Figura 8 – Matriz de Confusão para o quarto teste (Regressão Logística)




Fonte: Elaborado pela autora com a biblioteca *Scikitplot* do *Python*

Tabela 6 – Sumarização da Classificação com Regressão Logística para 166 atributos

| | precisão | sensibilidade | f1-score | support |
|---|---|---|---|---|
| 0.0 | 0.99 | 1.00 | 1.00 | 719 |
| 1.0 | 1.00 | 0.95 | 0.97 | 116 |
| accuracy | | | 0.99 | 835 |
| macro avg | 1.00 | 0.97 | 0.98 | 835 |
| weighted avg | 0.99 | 0.99 | 0.99 | 835 |

Fonte: Elaborado pela autora

Observa-se com a matriz de confusão da Figura 8 e dos índices da Tabela 6 que com 7 atributos somente, relacionados à violência, resolvem o problema de classificação dessa problemática em questão. Então, como seu resultado ainda foi bastante significativo, opta-se por testar somente com os atributos de forte correlação com a variável de saída para ver se os resultados continuam ou mudam.

### *4.2.5 Quinto modelo: usando somente os atributos de forte correlação*

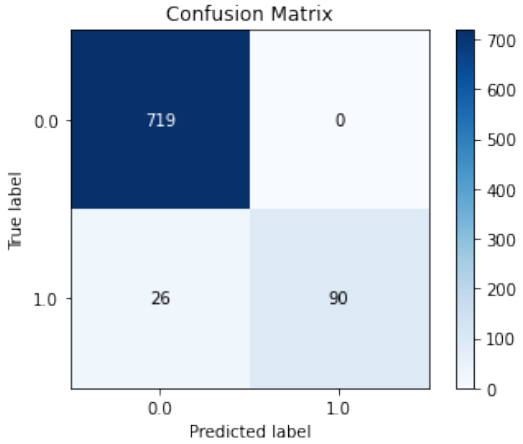
Para tanto, ainda foi testado usar somente os atributos que tem forte correlação com o *target*. Com isso, só tem um atributo que é 'q708_c1' que pergunta: "Insultou ou fez sentir mal consigo mesma? Nos últimos 12 meses". Com esse atributo, que tem -0.82 de correlação com 'violence12'. Esse valor é negativo, pois quando se faz a contagem dos valores para a classe 0 aparace somente o valor categórico 88888, com 4230 valores exclusivos e mais nenhuma outra categoria foi usada para a classe 0.

Já para a classe 1, tem que o valor 88888 aparece 208 vezes, sendo o maior valor mostrado para a classe 1 e ficando em segundo lugar o valor 2, com 189 valores exclusivos e o terceiro sendo o valor 3, com 189 valores exclusivos também. O valor 4 ficou com 73 dados, 5

com 34 dados e 1 somente com 17 dados. Todas essas observações são possíveis ver na Figura 2, no último gráfico.

O seu resultado foi bastante significativo, pois somente com um atributo, o modelo conseguiu prever muito bem as classes. Com 90% de *F1-Score macro*, o modelo não erra nenhum dado para a classe 0 e erra somente 44 dados para a classe 1, como podemos ver na matriz de confusão da Figura 9 a seguir.

Figura 9 – Matriz de Confusão para o quinto teste (Regressão Logística)




Fonte: Elaborado pela autora com a biblioteca *Scikitplot* do *Python*

Dito isso, pode-se observar na Tabela 7 a seguir os índices de classificação do modelo de Regressão Logística.

Tabela 7 – Sumarização da Classificação com Regressão Logística para 1 atributo

| | precisão | sensibilidade | f1-score | support |
|---|---|---|---|---|
| 0.0 | 0.97 | 1.00 | 0.98 | 719 |
| 1.0 | 1.00 | 0.78 | 0.87 | 116 |
| accuracy | | | 0.97 | 835 |
| macro avg | 0.98 | 0.89 | 0.93 | 835 |
| weighted avg | 0.97 | 0.97 | 0.97 | 835 |

Fonte: Elaborado pela autora

Observa-se que a diferença do quarto modelo para o quinto foi de apenas 20 dados a mais errados na classe 1. Ou seja, apenas com um atributo pode-se inferir na condição da mulher em predizer, somente com uma pergunta se ela será violentada ou não. A partir daí, serão feitos mais testes com mais modelos, porém sem nenhum dado de violência, apenas com dados referentes a características que não tem a ver com violência.

Pode-se achar parecido com o terceiro modelo, porém serão descartados todos os atributos de violência e somente serão considerados os atributos com outras características.

Dentro desses ainda será feita a Seleção de atributos para determinar um novo modelo.

Além disso, é planejado fazer mais buscas sobre dados de violência contra a mulher, explorar os dados internos e externos, ter *insights* acerca dos dados a fim de conseguir encontrar padrões além do que já foi mencionado.

## 4.3 Modelos da 2ª fase: somente dados socioeconômicos

Observa-se que os modelos citados acima tiveram um desempenho bem satisfatório e até surpreendente: com apenas 1 atributo o modelo consegue prever muito bem a variável de saída. Contudo, isso foi notado e explicado pelo criador do questionário: os dados estão realmente correlacionados porque quando a entrevistada responde alguma pergunta relacionada à violência de forma que indique que ela sofreu de alguma forma, o *target* já é acionado a ter a resposta como 1, que indica que sim, que ela sofreu violência. Portanto, não se pode utilizar as perguntas do questionário relacionadas à violência.

Dito isso, o novo modelo se dará em cima das perguntas relacionadas às questões socioeconômicas da entrevistada. Tais questões dizem respeito à idade, número de filhos, município que mora, trabalho remunerado e outros. Logo, são feitas mais análises exploratórias dos dados para um melhor compreendimento destes.

### *4.3.1 Sexto modelo: regressão logística*

Diante dos fatos mencionados, é refeito o modelo *baseline* para essa nova referência dos dados. Então é utilizado o algoritmo de Regressão Logística com os mesmos hiperparâmetros do modelo 1 para ter uma noção dos resultados e, a partir dele, melhorá-lo. Para tanto, seu resultado foi desfavorável como já era esperado, já que não se tem dados sobre violência. Com 103 atributos, o resultado desse modelo pode ser visto na Figura 10 a seguir:

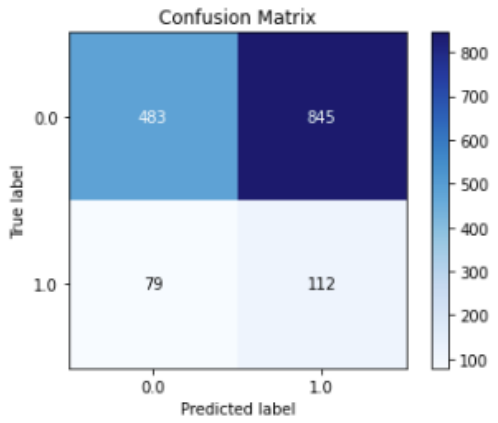
Dito isso, pode-se observar na Tabela 8 a seguir os índices de classificação do modelo de Regressão Logística.

Tabela 8 – Sumarização da Classificação com Regressão Logística para 103 atributos

| | precisão | sensibilidade | f1-score | support |
|---|---|---|---|---|
| 0.0 | 0.86 | 0.36 | 0.51 | 1328 |
| 1.0 | 0.12 | 0.59 | 0.20 | 191 |
| accuracy | | | 0.39 | 1519 |
| macro avg | 0.49 | 0.48 | 0.35 | 1519 |
| weighted avg | 0.77 | 0.39 | 0.47 | 1519 |

Fonte: Elaborado pela autora

Figura 10 – Matriz de Confusão para o sexto teste (Regressão Logística)

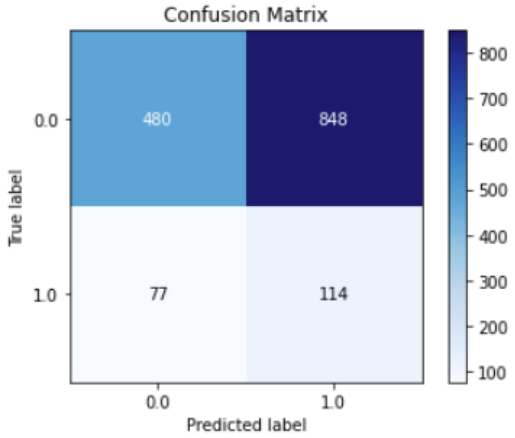



Fonte: Elaborado pela autora com a biblioteca *Scikitplot* do *Python*

Diante desses resultados, observa-se que o índice *F1-Score* que é o índice utilizado como base para todos os modelos anteriores, baixou consideravelmente para 35%.

Com esse resultado, faz-se o *Grid Search* para saber quais hiperparâmetros melhor resultam no modelo, porém dos hiperparâmetros colocados, muda-se somente o "C" de 0,5 para 0,1, no qual esse hiperparâmetro é o inverso da força de regularização, em que valores menores especificam uma regularização mais forte. Porém o resultado deste é insuficiente para mudar o índice que está sendo usado como referência. Os resultados são mostrados na Figura 11 a seguir.

Figura 11 – Matriz de Confusão para o sexto teste (Regressão Logística)




Fonte: Elaborado pela autora com a biblioteca *Scikitplot* do *Python*

Dito isso, pode-se observar na Tabela 9 a seguir os índices de classificação do modelo de Regressão Logística com *Grid Search*.

Observa-se que no índice avaliado, nada é mudado, portanto, considerado sem melhoras significativas.

Tabela 9 – Sumarização da Classificação com Regressão Logística para 103 atributos com *Grid Search*

| | precisão | sensibilidade | f1-score | support |
|---|---|---|---|---|
| 0.0 | 0.86 | 0.36 | 0.51 | 1328 |
| 1.0 | 0.12 | 0.60 | 0.20 | 191 |
| accuracy | | | 0.39 | 1519 |
| macro avg | 0.49 | 0.48 | 0.35 | 1519 |
| weighted avg | 0.77 | 0.39 | 0.47 | 1519 |

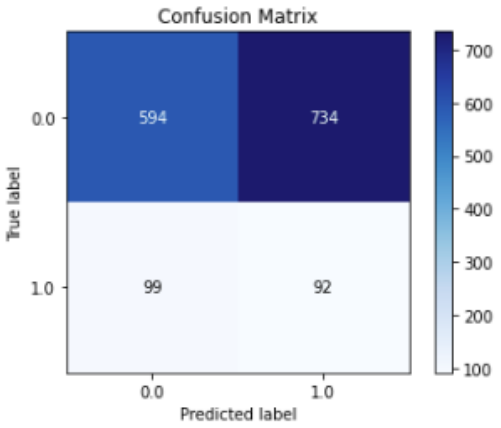
Fonte: Elaborado pela autora

### *4.3.2 Sétimo modelo: regressão logística com seleção de atributos*

Com os resultados insatisfatórios do modelo *baseline*, o objetivo agora é melhorá-lo para obter um índice convincente para a sua possível utilização na vida real. Para tanto, é feita a seleção das melhores *atributos* dentre as 103 perguntas do questionário. A partir disso, são feitas três listas de atributos na Seleção de atributos com os algoritmos *SelectKBest*, *RFE* e *SelectFromModel* com o estimador do *Random Forest*. Dessa forma, a partir dessas três listas é feita uma pegando somente aquele atributo que se repete pelo menos duas vezes nas listas criadas. Então, a lista da Seleção de atributos é formada com 34 atributos.

O modelo de Regressão Logística é feito, depois de separar em treino e teste, e o seus erros e acertos são mostrados na Matriz de Confusão da Figura 12 a seguir:

Figura 12 – Matriz de Confusão para o sétimo teste (Regressão Logística)




Fonte: Elaborado pela autora com a biblioteca *Scikitplot* do *Python*

Dito isso, pode-se observar na Tabela 10 a seguir os índices de classificação do modelo de Regressão Logística com Seleção de atributos.

Tabela 10 – Sumarização da Classificação com Regressão Logística para 34 atributos com Seleção de atributos

| | precisão | sensibilidade | f1-score | support |
|---|---|---|---|---|
| 0.0 | 0.86 | 0.45 | 0.59 | 1328 |
| 1.0 | 0.11 | 0.48 | 0.18 | 191 |
| accuracy | | | 0.45 | 1519 |
| macro avg | 0.48 | 0.46 | 0.38 | 1519 |
| weighted avg | 0.76 | 0.45 | 0.54 | 1519 |

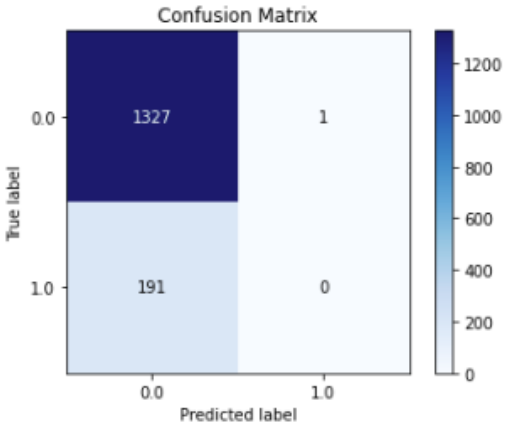
Fonte: Elaborado pela autora

A vista disso, observa-se que o modelo melhorou 3% em relação ao modelo anterior, comparando com o índice de *F1-Score*. A partir desses resultados opta-se por testar outros algoritmos.

### *4.3.3 Oitavo modelo: MLP com seleção de atributos e balanceamento*

Com os resultados da Regressão Logística, é escolhido fazer outros modelos com outros algoritmos. Então é escolhido o algoritmo de redes neurais, chamado de Perceptron Múltiplas Camadas (PMC ou MLP — *Multi Layer Perceptron*). Nele, usa-se 500 épocas com os 34 atributos da Seleção de atributos. O resultado é dado a seguir na Figura 13:

Figura 13 – Matriz de Confusão para o oitavo teste (MLP)




Fonte: Elaborado pela autora com a biblioteca *Scikitplot* do *Python*

Dito isso, pode-se observar na Tabela 11 a seguir os índices de classificação do modelo de MLP com Seleção de atributos.

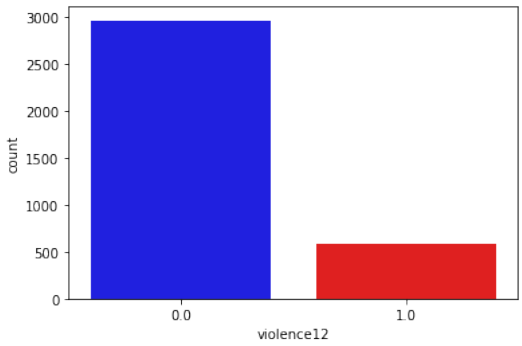
Tabela 11 – Sumarização da Classificação com MLP para 34 atributos com Seleção de atributos

| | precisão | sensibilidade | f1-score | support |
|---|---|---|---|---|
| 0.0 | 0.87 | 1.00 | 0.93 | 1328 |
| 1.0 | 0.00 | 0.00 | 0.00 | 191 |
| accuracy | | | 0.87 | 1519 |
| macro avg | 0.44 | 0.50 | 0.47 | 1519 |
| weighted avg | 0.76 | 0.87 | 0.82 | 1519 |

Fonte: Elaborado pela autora

Nota-se que no índice de referência o modelo melhorou para 47%, porém observando a Tabela 11, a classe 1 que é o que indica a chance de ter violência, está com 0, ou seja ele erra para exatamente todos os registros de classe 1. Além disso, os dados estão desbalanceados. Na Figura 14 é quantificado esse desbalanceamento em um gráfico de barras.

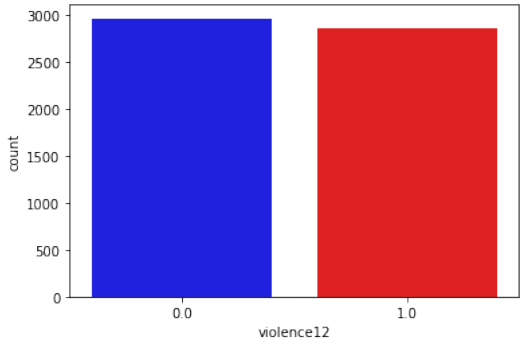
Figura 14 – Gráfico de desbalanceamento do *target*




Fonte: Elaborado pela autora com a biblioteca *Seaborn* do *Python*

Portanto é decidido fazer o balanceamento dos dados com o algoritmo *ADASYN* nos dados somente de treino. Com o balanceamento aplicado, os dados ficam como na Figura 15 a seguir:

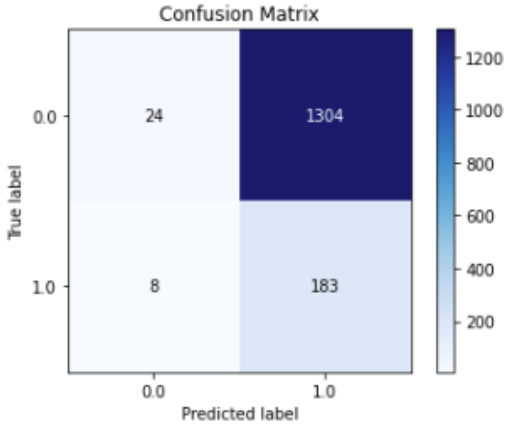
Figura 15 – Gráfico após o balanceamento com ADASYN do *target*




Fonte: Elaborado pela autora com a biblioteca *Seaborn* do *Python*

Após o balanceamento, é feito novamente o mesmo modelo MLP. Assim sendo, o resultado se dá com a matriz de confusão da Figura 16 a seguir:

Figura 16 – Matriz de Confusão para o oitavo teste (MLP)




Fonte: Elaborado pela autora com a biblioteca *Scikitplot* do *Python*

Dito isso, pode-se observar na Tabela 12 a seguir os índices de classificação do modelo de MLP com Seleção de atributos.

Tabela 12 – Sumarização da Classificação com MLP para 34 atributos com Seleção de atributos e dados balanceados

| | precisão | sensibilidade | f1-score | support |
|---|---|---|---|---|
| 0.0 | 0.75 | 0.02 | 0.04 | 1328 |
| 1.0 | 0.12 | 0.96 | 0.22 | 191 |
| accuracy | | | 0.14 | 1519 |
| macro avg | 0.44 | 0.49 | 0.13 | 1519 |
| weighted avg | 0.67 | 0.14 | 0.06 | 1519 |

Fonte: Elaborado pela autora

Apesar de mostrar melhora para a classe 1, o índice de *F1 Score* piorou bastante. Portanto, mais uma vez opta-se por utilizar outro algoritmo para comparar os resultados.

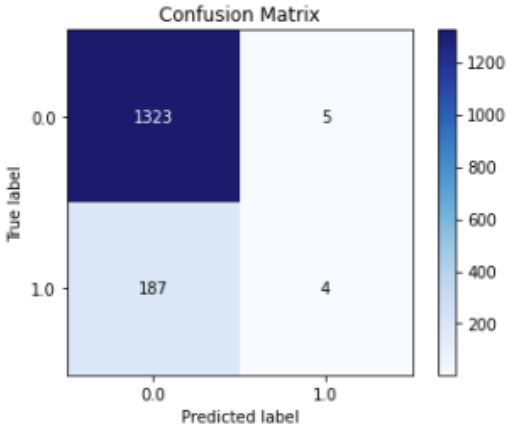
### *4.3.4 Nono modelo: random forest com seleção de atributos e balanceamento*

Com as mesmas mudanças dos algoritmos anteriores (seleção de mehores atributos, balancemanto, separação treino/teste) o modelo com o algoritmo de *Random Forest* é aplicado. Além disso, é feito o *Grid Search* para selecionar os melhores hiperparâmetros e poder comparar com os algortimos anteriores. Os melhores hiperparâmetros são listados a seguir:

- criterion = 'gini';
- n_estimators = 50;
- oob_score = True;
- warm_start = False.

O resultado do *Random Forest* pode ser visto na Matriz de Confusão da Figura a seguir:

Figura 17 – Matriz de Confusão para o nono teste (*Random Forest*)




Fonte: Elaborado pela autora com a biblioteca *Scikitplot* do *Python*

Dito isso, pode-se observar na Tabela 13 a seguir os índices de classificação do

modelo de *Random Forest* com Seleção de atributos.

Tabela 13 – Sumarização da Classificação com MLP para 34 atributos com Seleção de atributos e dados balanceados

| | precisão | sensibilidade | f1-score | support |
|---|---|---|---|---|
| 0.0 | 0.88 | 1.00 | 0.93 | 1328 |
| 1.0 | 0.44 | 0.02 | 0.04 | 191 |
| accuracy | | | 0.87 | 1519 |
| macro avg | 0.66 | 0.51 | 0.49 | 1519 |
| weighted avg | 0.82 | 0.87 | 0.82 | 1519 |

Fonte: Elaborado pela autora

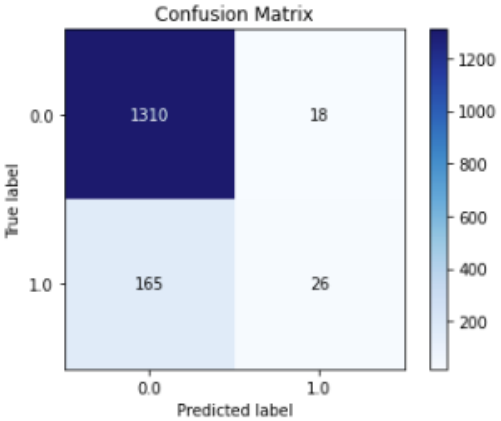
Verifica-se que apesar desse ser o melhor resultado até agora, o modelo continua errando bastante para a classe 1, que seria a classe considerada a chance de ter violência.

### *4.3.5 Décimo modelo: XGBoost*

Neste último modelo são encontrados resultados bastante significativos para o trabalho. Porém, para chegar no resultado final, precisou-se de algumas mudanças na técnica para sua excelência. Para tanto, o primeiro teste se deu no uso da técnica de forma padrão, ou seja, sem modificar os hiperparâmetros para observar seu resultado. Vale salientar que a partir do primeiro modelo usando o *XGBoost* também foi utilizada as técnicas de balanceamento e de seleção dos melhores atributos.

À vista disso, o primeiro resultado se deu na Matriz de Confusão da Figura 18 a seguir:

Figura 18 – Matriz de Confusão para o nono teste (*XGBoost*)




Fonte: Elaborado pela autora com a biblioteca *Scikitplot* do *Python*.

Observa-se na Tabela 14 a seguir os índices de classificação do modelo de *XGBoost*.

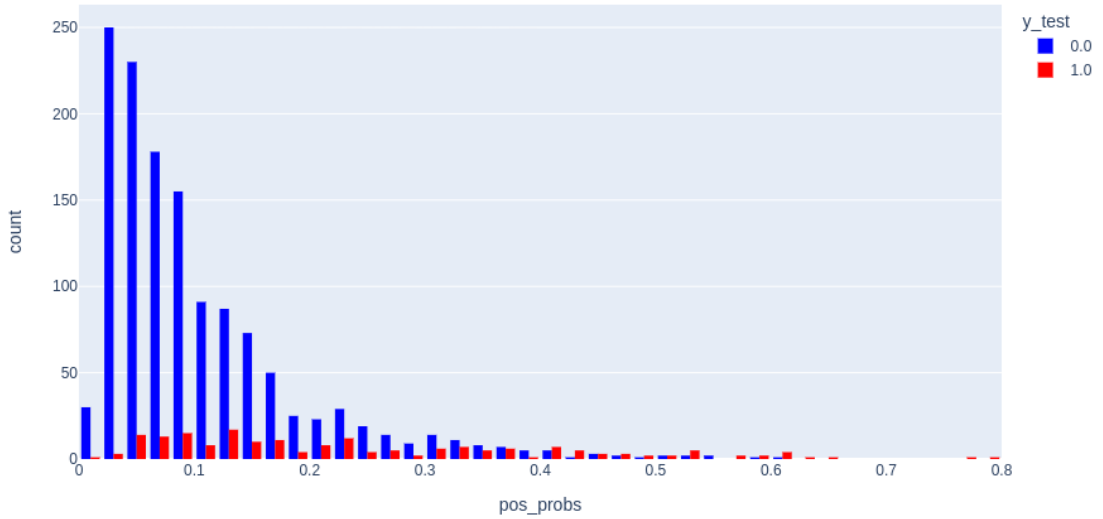
Tabela 14 – Sumarização da Classificação com *XGBoost* para 34 atributos com Seleção de atributos e dados balanceados

| | precisão | sensibilidade | f1-score | support |
|---|---|---|---|---|
| 0.0 | 0.89 | 0.99 | 0.93 | 1328 |
| 1.0 | 0.59 | 0.14 | 0.22 | 191 |
| accuracy | | | 0.88 | 1519 |
| macro avg | 0.74 | 0.56 | 0.58 | 1519 |
| weighted avg | 0.85 | 0.88 | 0.85 | 1519 |

Fonte: Elaborado pela autora

De acordo com a Tabela 14 verifica-se que este é o melhor resultado obtido até o momento, com *F1 Score* de 58%, mas que ainda está abaixo do esperado. Para isso, é feita a prediçao dos dados de teste, depois de treinar o modelo com os dados de treino, mas dessa vez com probabilidades para observar sua distribuição. Tal distribuição é vista na Figura 19 seguinte.

Figura 19 – Distribuição de Probabilidade usando o *XGBoost*



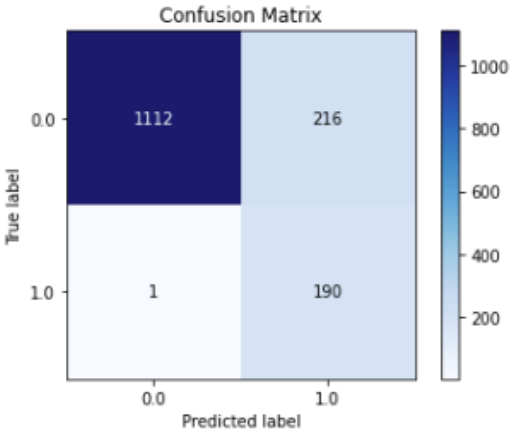

Fonte: Elaborado pela autora com a biblioteca *Plotly* do *Python*

Por consequência, detecta-se que os dados de classe 0 estão concentrados nas probabilidades entre 0 e 0,2 (barras em azul), porém existe uma quantidade significativa de dados da classe 1 (em vermelho) na mesma região, mostrando que o limiar de 0,5, que é a metade e consequentemente o padrão de todas as técnicas citadas acima, não é o suficiente para tornar-se um bom resultado.

Diante dessa informação, o modelo do *XGBoost* é testado novamente, mas com a mudança do seu limiar, comumente chamado de *threshold*. Com o modelo posterior são testados vários *threshold* de forma empírica para observar seus resultados e fazer comparações. A Figura

20 mostra a Matriz de Confusão do modelo.

Figura 20 – Distribuição de Probabilidade usando o *XGBoost*




Fonte: Elaborado pela autora com a biblioteca *Scikitplot* do *Python*

Depois de treinado o modelo, os índices são calculados com base nas probabilidades das predições que serão mostradas nas tabelas seguintes com seus *thresholds* no título, sendo este limitado em um intervalo de 0 a 1. O primeiro teste é feito com o limiar maior ou igual a 0,1 e seu resultado cai de 58% para 56% do índice do *F1 Score*. Já quando aumenta para 0,11 o índice fica expressivamente maior que o último resultado, aumentando de 58% para 77% de *F1 Score Macro AVG*.

Tabela 15 – Sumarização da Classificação com *XGBoost* com *threshold* >= 0.11

| | precisão | sensibilidade | f1-score | support |
|---|---|---|---|---|
| 0.0 | 1.00 | 0.84 | 0.91 | 1328 |
| 1.0 | 0.47 | 0.99 | 0.64 | 191 |
| accuracy | | | 0.86 | 1519 |
| macro avg | 0.73 | 0.92 | 0.77 | 1519 |
| weighted avg | 0.93 | 0.86 | 0.88 | 1519 |

Fonte: Elaborado pela autora

Ainda é feito outro teste com o limiar de 0,2, como é mostrado na Tabela 17 a seguir, porém o seu resultado, apesar de melhor do que o modelo *baseline*, ele ainda é menor que o resultado do limiar de 0,11.

Ainda é feito outro teste com o *threshold* maior ou igual a 0,3, porém o índice resulta no mesmo valor quando é colocado a 0,2. Dessa forma, como esses resultados são empíricos, há a possibilidade de existir algum limiar que tenha o melhor resultado e que não foi colocado até agora nesta pesquisa.

Tabela 16 – Sumarização da Classificação com *XGBoost* com *threshold* >= 0.2

| | precisão | sensibilidade | f1-score | support |
|---|---|---|---|---|
| 0.0 | 0.92 | 0.87 | 0.90 | 1328 |
| 1.0 | 0.36 | 0.49 | 0.41 | 191 |
| accuracy | | | 0.83 | 1519 |
| macro avg | 0.64 | 0.68 | 0.65 | 1519 |
| weighted avg | 0.85 | 0.83 | 0.84 | 1519 |

Fonte: Elaborado pela autora

Portanto, é optado por fazer um estudo sobre as medidas de tendência central nos valores observados da Figura 19 de distribuição das probabilidades. Isto posto, os seus resultados estão na Tabela17.

Tabela 17 – Análise descritiva da Classificação com *XGBoost*

| | **yhat** | **pos_probs** | **y_test** |
|---|---|---|---|
| **count** | 1519 | 1519 | 1519 |
| **mean** | 0.017775 | 0.121717 | 0.125741 |
| **std** | 0.132176 | 0.112098 | 0.331666 |
| **min** | 0 | 0.009144 | 0 |
| **25%** | 0 | 0.046861 | 0 |
| **50%** | 0 | 0.083862 | 0 |
| **75%** | 0 | 0.150752 | 0 |
| **max** | 1 | 0.794248 | 1 |

Fonte: Elaborado pela autora

Sendo o *y_hat* os valores somente de 0 e 1 dos resultados da predição, o *pos_probs* as probabilidades compreendidas no intervalo entre 0 e 1 e o *y_test* o resultado real dos dados de teste. Então, como o desvio padrão dos dados de teste está com o valor de aproximadamente de 0,33, tal valor é colocado como o *threshold* para observar se o resultado obtém melhora. E, satisfatoriamente, ele atinge o resultado significativo de 81% de *F1 Score*. A Tabela 18 mostra esta e outros índices.

Tabela 18 – Sumarização da Classificação com *XGBoost* com *threshold* >= 0.33

| | precisão | sensibilidade | f1-score | support |
|---|---|---|---|---|
| 0.0 | 0.98 | 0.91 | 0.94 | 1328 |
| 1.0 | 0.56 | 0.84 | 0.68 | 191 |
| accuracy | | | 0.90 | 1519 |
| macro avg | 0.77 | 0.87 | 0.81 | 1519 |
| weighted avg | 0.92 | 0.90 | 0.91 | 1519 |

Fonte: Elaborado pela autora

# 5 CONCLUSÕES E SUGESTÕES

Neste trabalho, foi realizada a classificação de mulheres que serão violentadas ou não com base no questionário feito em vários lugares do país. Após identificar a base de dados foram feitos vários modelos com o intuito de classificar bem através de algoritmos de aprendizado de máquina e de observar quais padrões podem-se ver com os dados recebidos.

Os experimentos realizados visam avaliar os modelos e observar seus índices para entender seus resultados. Porém, tendo em vista os fatos e modelos mencionados, é interessante destacar que entender os dados é extremamente importante para um bom entendimento do modelo e para obter boas inferências, não apenas o resultado de seus índices.

Os resultados obtidos reforçam que o uso da predição de violência contra a mulher demonstrou ser bastante satisfatório em predizer corretamente e mostrar poucos erros na matriz de confusão, porém percebe-se que com apenas um atributo, o modelo já se destaca com 93% de *F1-Score macro avg* errando bem pouco nos resultados. Porém este resultado mostrou-se equivocado quando percebe-se que estão altamente correlacionados com o atributo de violência. Dito isso houve uma segunda fase que possibilitou a execução correta dos modelos utilizando somente dados socioeconômicos e não correlacionados diretamente.

Também é fundamental destacar que ainda, nos experimentos realizados, constata-se que a mensuração dos modelos preditores utilizando os índices de classificação devem ser utilizadas com certa cautela. Pois em determinadas situações alguns índices mostradas não são suficientes para indicar o valor real do acerto/erro do modelo.

Além disso, esta pesquisa contribuiu de forma significativa em três aspectos principais, trazendo avanços importantes para o campo do combate à violência contra a mulher e para a própria pesquisa do PCSVDF-Mulher.

Primeiramente, a criação de modelos de previsão de violência contra a mulher proporciona uma ferramenta valiosa para identificar e prevenir casos de violência doméstica. Esses modelos podem ajudar a sinalizar os casos de maior risco, direcionando recursos e apoio às mulheres que mais precisam, além de facilitar a intervenção precoce por parte das autoridades competentes.

Em segundo lugar, a disponibilização dos índices de classificação obtidas a partir dos modelos desenvolvidos durante esta pesquisa é de extrema relevância para a comunidade científica, organizações não governamentais e formuladores de políticas públicas. Esses índices fornecem uma visão mais precisa sobre a eficácia dos modelos e podem auxiliar na tomada de

decisões embasadas em evidências para o combate à violência contra a mulher.

Por fim, a utilização do melhor resultado obtido por meio dos modelos preditivos contribui para identificar a probabilidade de uma mulher ser vítima de violência, permitindo uma atuação preventiva e o desenvolvimento de estratégias de redução das taxas de violência. Essa abordagem baseada em dados contribui para uma compreensão mais aprofundada dos fatores socioeconômicos e contextuais que influenciam a violência doméstica, possibilitando a formulação de políticas públicas mais efetivas e direcionadas.

Dessa forma, essa pesquisa representa um avanço importante no campo da prevenção e combate à violência contra a mulher, fornecendo *insights* valiosos e embasamento científico para a tomada de decisões e a implementação de políticas públicas que visem a segurança e o bem-estar das mulheres em nossa sociedade.

É importante ressaltar que, embora tenham sido obtidos resultados significativos nessa pesquisa, esses resultados podem ter sido afetados por diferentes fatores, como a presença de dados faltantes, principalmente em informações como "Não sabe/Não respondeu". Isso impacta a obtenção de percepções concretas sobre certos *insights* relacionados aos dados, especialmente quando se trata de atributos ligados à violência em si.

Outro desafio encontrado foi o desequilíbrio dos dados, o que pode prejudicar o aprendizado dos modelos. Felizmente, encontrou-se uma maior quantidade de mulheres não violentadas nessa pesquisa, o que permitiu que os modelos aprendessem bem em relação à classe majoritária (classe 0), sendo utilizado técnicas de *oversampling* para abordar a outra classe.

Por fim, é importante destacar que o trabalho obteve um resultado final impressionante de 81% de *F1-Score macro avg*, utilizando técnicas de seleção de atributos e *oversampling* nos dados de treinamento durante a fase final de testes dos modelos. Esse resultado é relevante para fins de comparação e demonstra um desempenho promissor em relação à previsão da violência contra a mulher.

# REFERÊNCIAS


ALBON, C. **Machine learning with python cookbook**: practical solutions from preprocessing to deep learning. [S.l.]: O'Reilly Media, Inc., 2018.

ALPAYDIN, E. **Introduction to machine learning**. *[S.l.]*: MIT Press, 2020.

BRASIL. Lei nº 11.340, de 07 de agosto de 2006: Cria mecanismos para coibir a violência doméstica e familiar contra a mulher [...]. **Diário Oficial [da] República Federativa do Brasil**, Brasília, DF, 2006. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2004-2006/2006/lei/l11340.htm. Acesso em: 12 dez. 2022.

CARVALHO, J. R.; OLIVEIRA, V. H. de; SILVA, A. B. R. da. **The PCSVDF study**: new data, prevalence and correlates of domestic violence in brazil. [S.l.: s.n.]. Fortaleza: CAEN, 2018. (Série Estudos Econômicos do CAEN).

CEARÁ. Secretaria de Segurança Pública e Defesa Social. **Estatísticas**. Fortaleza: SSPDS, 2022. Disponível em: https://www.sspds.ce.gov.br/estatisticas-2/. Acesso em: 29 jun. 2022.

CHENG, J.; GREINER, R. Comparing bayesian network classifiers. **arXiv.org**, [Ithaca, N. Y.], 2013. Disponível em: https://arxiv.org/abs/1301.6684. Acesso em: 10 dez. 2022.

CHRISTODOULOU, E.; MA, J.; COLLINS, G. S.; STEYERBERG, E. W.; VERBAKEL, J. Y.; CALSTER, B. V. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. **Journal of Clinical Epidemiology**, United States. v. 110, p. 12–22, 2019.

ESCORSIM, S. M. Violência de gênero e saúde coletiva: um debate necessário. **Revista Katálysis**, Brasil. v. 17, p. 235–241, 2014.

GÉRON, A. **Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow**: concepts, tools, and techniques to build intelligent systems. *[S.l.]*: O'Reilly Media, Inc., 2019.

INSTITUTO IGARAPÉ. **Metodologia EVA**. Rio de Janeiro. Instituto Igarapé, 2020. Disponível em: https://eva.igarape.org.br/metodologia_eva_pt.pdf. Acesso em: 20 jun. 2022.

LITTLE, R. J.; RUBIN, D. B. **Statistical analysis with missing data**. *[S.l.]*: John Wiley & Sons, 2019. v. 793.

MAIMON, O. Z.; ROKACH, L. **Data mining with decision trees: theory and applications**. *[S.l.]*: World Scientific, 2014. v. 81.

MARQUES, E. S.; MORAES, C. L. d.; HASSELMANN, M. H.; DESLANDES, S. F.; REICHENHEIM, M. E. A violência contra mulheres, crianças e adolescentes em tempos de pandemia pela covid-19: panorama, motivações e formas de enfrentamento. **Cadernos de Saúde Pública**, Brasil. v. 36, 2020.

MOHRI, M.; ROSTAMIZADEH, A.; TALWALKAR, A. **Foundations of machine learning**. *[S.l.]*: MIT Press, 2018.

OKABAYASHI, N. Y. T.; TASSARA, I. G.; CASACA, M. C. G.; FALCÃO, A. de A.; BELLINI, M. Z. Violência contra a mulher e feminicídio no brasil-impacto do isolamento social pela covid-19. **Brazilian Journal of Health Review**, Brasil, v. 3, n. 3, p. 4511–4531, 2020.

ORGANIZAÇÃO MUNDIAL DE SAÚDE. **OMS**: uma em cada 3 mulheres em todo o mundo sofre violência. Brasília, DF: ONU Brasil, 2021. Disponível em: https://brasil.un.org/pt-br/115652-oms-uma-em-cada-3-mulheres-em-todo-o-mundo-sofre-violencia. Acesso em: 29 jun. 2022.

PAN, B. Application of XGBoost algorithm in hourly PM2.5 concentration prediction. *In:* **IOP conference series:** Earth and environmental science. *[S. l.]:* IOP, 2018. v. 113, n. 1, p. 012127.

PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V.; VANDERPLAS, J.; PASSOS, A.; COURNAPEAU, D.; BRUCHER, M.; PERROT, M.; DUCHESNAY, E. Scikit-learn: Machine learning in Python. **Journal of Machine Learning Research**, United States, v. 12, p. 2825–2830, 2011.

RASCHKA, S.; MIRJALILI, V. **Python machine learning**: machine learning and deep learning with python, scikit-learn, and tensorflow 2. *[S.l.]*: Packt Publishing Ltd, 2019.

SUR, P.; CANDÈS, E. J. A modern maximum-likelihood theory for high-dimensional logistic regression. **Proceedings of the National Academy of Sciences**, United States, v. 116, n. 29, p. 14516–14525, 2019.

ZOU, X.; HU, Y.; TIAN, Z.; SHEN, K. Logistic regression model optimization and case analysis. *In:* INTERNATIONAL CONFERENCE ON COMPUTER SCIENCE AND NETWORK TECHNOLOGY, 7., 2019, Dalian, China. **Proceedings** [...]. *[S. l.]*: IEEE, 2019. p. 135–139.