**TECHNICAL REPORT**

| Aluno: Paulo Henrique Santos Marques | | |
| --- | --- | --- |

1. **Introdução**

*Violência Doméstica contra Mulheres*

*A violência doméstica contra mulheres é um problema significativo em muitos países em desenvolvimento. Compreender os fatores que contribuem para essa violência pode ajudar na criação de intervenções e políticas eficazes. Este conjunto de dados fornece uma base para análise e pesquisa nesta área crítica. Entre nossas features temos as seguintes:d*

| ***Colunas*** | ***Descrição*** |
| --- | --- |
| *Idade (Age)* | *Idade do respondente* |
| *Educação (Education)* | *Nível Educacional do Respondente* |
| *Emprego (Employment)* | *Status de emprego do Respondente* |
| *Renda (Income)* | *Nível de Renda do Respondente* |
| *Status Matrimonial (Marital status)* | *Estado Civil do Respondente* |
| *Violência (Violence)* | *Indica se o respondente sofreu violência ou não.* |
| *Número de Série do Registro (SL. No)* | *ID do Registro* |

*Temos como objetivo aplicar classificadores e responder as questões para verificar o desempenho de diferentes algoritmos de machine learning.*

1. **Observações**

*Alguns algoritmos apresentaram um desempenho mediano com acurácias por volta de 70% de taxa de acerto/acurácia. Baseado nisso, foram adotados diferentes estratégias para que o desempenho desses modelos apresentem uma melhor performance.*

1. **Resultados e discussão**

*Dado que entendemos o contexto da nossa problemática de dados e objetivos, vamos visualizar e discutir o que foi feito.*

*Questão 1 - Converta o dataset de imagem para um data frame e utilizando cálculo do índice de Gini e entropia, determine as duas possibilidades de nó raíz da árvore de decisão. A última coluna do dataset é a coluna alvo.*

*Resultados;*

| ***P. de Divisão para Gini*** | ***P. de Divisão para Entropy*** |
| --- | --- |
| Febre, Limite(Threshold): **0.5** | Febre, Limite(Threshold): **0.5** |
| Sem sabor ou Cheiro, Limite(Threshold): **0.5** | Sem sabor ou Cheiro, Limite(Threshold): **0.5** |
| Tosse, Limite(Threshold):**0,5** | Tosse, Limite(Threshold):**0,5** |

***Interpretação:***

Feature: Fever, Threshold: 0.5: Isso significa que a presença ou ausência de febre (codificada como 0 ou 1) é uma característica significativa para a divisão inicial.

Feature: No Taste or Smell, Threshold: 0.5: Após considerar a febre, a ausência ou presença de perda de paladar ou olfato é a próxima característica importante.

Feature: Cough, Threshold: 0.5: Finalmente, a tosse é considerada para decisões posteriores.

***Discussão;***

Os critérios de Gini e entropia são ambos métodos para medir a impureza dos nós em uma árvore de decisão. Ao verificar os resultados, a primeira coisa que chama atenção é o fato das duas métricas, tanto o Gini quanto entropia serem exatamente iguais e com as medidas resultantes de 0.5. Após analisar os dados e pesquisar como funciona esses índices, aparentemente essa igualdade de resultados se deve ao fato dos dados ser bastante limitados, visto que temos apenas 10 instâncias e 3 features. Dado esse fator nos dados é comum os pontos de divisão serem iguais.

*Questão 2 - Faça o download do dataset e realize os pré-processamentos adequados. Selecione as colunas que você acredita ser adequada de analisar, remova dados desnecessários e realize uma predição utilizando a árvore de decisão. Mostre números e formas adequadas de avaliar o desempenho do classificador. Mostre, inclusive, curvas que auxiliam na análise de desempenho.*

*Para resolver essa questão, além dos processos de pré processamento de dados, foi implementado duas configurações de modelos, onde foi obtido resultados para as duas configurações.*

***Resultados:***

***Config- 1***

| *Config-1* | *Precision* | *Recall* | *F1-Score* | *Support* |
| --- | --- | --- | --- | --- |
| *0* | *75%* | *87%* | *81%* | *77* |
| *1* | *38%* | *21%* | *27%* | *28* |
|  |  |  |  |  |
| *accuracy* |  |  | *70%* | *105* |
| *macro avg* | *56%* | *54%* | *54%* | *105* |
| *weighted avg* | *65%* | *70%* | *66%* | *105* |

***Config- 2***

| *Config-2* | *Precision* | *Recall* | *F1-Score* | *Support* |
| --- | --- | --- | --- | --- |
| *0* | *81%* | *86%* | *84%* | *77* |
| *1* | *54%* | *46%* | *50%* | *28* |
|  |  |  |  |  |
| *accuracy* |  |  | *70%* | *105* |
| *macro avg* | *68%* | *66%* | *67%* | *105* |
| *weighted avg* | *74%* | *75%* | *75%* | *105* |

***Discussão;***

*Com intuito de responder a questão, resolvi quebrar a questão em pequenas tarefas, de modo que fosse mais fácil concluí-la. As tarefas foram as seguintes:*

Análise Exploratória;

Pré processamento dos dados;

Implementar o modelo;

Reconfigurar e Aplicar novamente o modelo;

**Vamos começar pela análise descritiva dos dados.**

Nosso dataset inicial contém um total de 7 colunas são elas:

SL. No Income

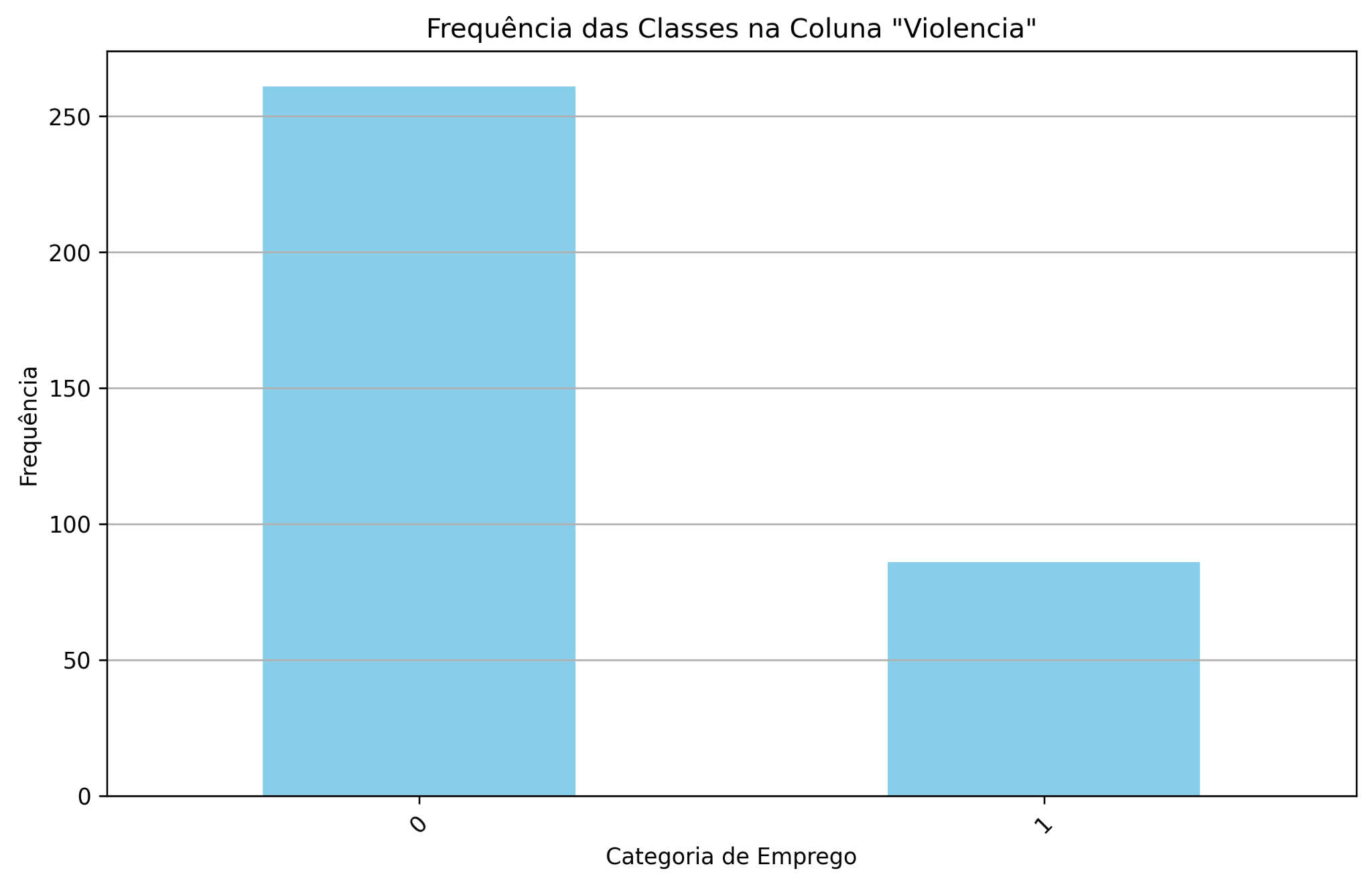
*Age Marital Status*

*Education Violence*

*Employment*

*De resumo, sobre os tipos das variáveis, temos SL No sendo um identificador único de registro das respondentes as outras variáveis apresentam* ***dados numéricos sendo essas as colunas: Age, Income,*** *as outras variáveis apresentam dados do tipo objetos, mas como se trata de variáveis categóricas, foi aplicado o algoritmo de codificação LabelEncoder para transformar essas variáveis em numéricas. Feito isso, tendo um dataset totalmente adequado para aplicar os modelos, foi realizada a primeira aplicação. para visualização dos resultados basta consultar na seção de resultados a tabela config-1.*

*Foi observado que na primeira configuração, o algoritmo teve um desempenho superior a classe 0 (Não Violência), forte indicativo de que o modelo poderia estar enviesado por conta de desbalanceamento. Para confirmar essa tese, foi feito um plot da frequência das classes da variável target, violência, o que nos ajudou a confirmar a hipótese.*

**

*Para lidar com esse problema foi aplicado um algoritmo de SMOTE para aplicar esse balanceamento através de geração de dados sintéticos. feito esse balanceamento, na tabela* ***Config- 2*** *na seção* ***Resultados***, é notório a evolução alcançada por essa estratégia.

*Questão 3 - Em uma atividade de casa você implementou manualmente o random forest. Esse algoritmo é exclusivo para aplicação de variação de árvore de decisão. Implemente manualmente uma generalização, Random Classifier, em que em vez de unicamente a árvore de decisão, o algoritmo possa trabalhar com o classificador KNN ou árvore de decisão.*

*Antes de tudo, é importante ressaltar que os dados utilizados nessa questão é o mesmo definido na questão 2. Com intuito de reaproveitamento o dataset resultante do pré processamento implementado na questão 2, foi salvo e utilizado nas próximas questões. Sabendo disso, vamos progredir com os exercícios.*

***Resultados:***

*Decision Tree: 0.44…*

*KNN = 0.51…*

***Discussão:***

A priori, é inegável que o desempenho dos classificadores desenvolvidos e implementados nessa questão, não apresentaram o melhor dos desempenhos, e é possível inferir algumas razões para estes resultados. Primeiro, o dataset é bem limitado e talvez pelo tipo de dado apresentado não seja tão adequado as aplicação dos modelos. Alguns requisitos são importantes para que os modelos consiga bom desempenho na generalização dos resultados. É um fato, relativo ao que acontece nos pontos de divisão trabalhados na questão 1, com o índice de Gini e Entropia. Ou a outra alternativa, é a estratégia de implementação da classe, é possível que outros tipos de normalização e estratégias de manipulação de dados faça com que o modelo seja mais eficaz. Bom, acredito que esse desempenho baixo dos modelos pode ser devido a essas razões, por tanto, fica registrado essa questão.

*Questão 4 -* Em um único script faça uma implementação otimizada que compare para o seu dataset, o desempenho de árvore de decisão, random forest, adaboost, gradiente Boost e SGB. Mostre números de forma organizada que seja possível interpretar a melhor forma de realizar a classificação.

***Resultados:***

|  | ***Accuracy*** | ***Precision*** | ***Recall*** | ***F1 Score*** |
| --- | --- | --- | --- | --- |
| ***Decision Tree*** | ***0.74*** | ***0.52*** | ***0.42*** | ***0.47*** |
| ***Random F*** | ***0.75*** | ***0.54*** | ***0.46*** | ***0.50*** |
| ***AdaBoost*** | ***0.66*** | ***0.40*** | ***0.50*** | ***0.44*** |
| ***Gradient Boost*** | ***0.71*** | ***0.46*** | ***0.42*** | ***0.44*** |
| ***SGB*** | ***0.70*** | ***0.44*** | ***0.42*** | ***0.43*** |

***Discussão:***

Os algoritmos ficaram na sua média. Caso a gente compare o desempenho geral dos algoritmos de machine learning, é possível inferir que nenhum deles teve um desempenho muito discrepante quando comparado aos outros. As métricas de avaliação escolhidas foram as mesmas da questão dois com intuito de padronizar a medida de desempenho. Novamente ressaltando que temos dados balanceados através do SMOTE, também vale ressaltar que foi testado algumas técnicas como validação cruzada e outras alternativas com intuito de melhorar o desempenho, entretanto, como não teve um desempenho significativo, permaneceu apenas o balanceamento da target.

1. **Conclusões**

Com base nas análises feitas, acredito que a estratégia de implementação utilizada, não foi das melhores, o que refletiu diretamente nos resultados dos modelos. Acredito que com base no que foi afirmado antes, o desempenho poderia ser melhor. Tirando isso, as questões foram resolvidas como planejado e prontamente executadas como deve ser.

1. **Próximos passos**

*Explorar diferentes configurações de modelos, além de utilizar mais métricas que possa representar melhor os resultados das tais.*