Mineração de Dados com foco no suporte a decisão binária para contribuir com o combate a atividades terroristas na América Latina.

Lucas Rabelo
Pernambuco, CIN UFPE, Brasil
Email: lram2@cin.ufpe.br

Ivson Guedes
Pernambuco, CIN UFPE, Brasil
Email: igon2@cin.ufpe.br

Antonio Ferreira
Pernambuco, CIN UFPE, Brasil
Email: antoniolfnt@gmail.com

Carlos Medeiros
Pernambuco, CIN UFPE, Brasil
Email: chvm@cin.ufpe.br

***Abstract*—Este estudo aborda a aplicação de mineração de dados para apoiar decisões no combate ao terrorismo na América Latina. Utilizando o Global Terrorism Database, foram analisados eventos ocorridos desde 1970 em 19 países. A abordagem incluiu a construção de modelos, como regressão logística e árvores de decisão, para identificar fatores associados ao sucesso ou fracasso no combate a ataques terroristas. Diversos indicadores, como dias da semana, tamanho populacional e presença de líderes militares, foram explorados. A análise também destaca a relevância do dia da semana dos ataques. Essas descobertas podem orientar estratégias de segurança e alocar recursos de maneira mais eficiente para enfrentar ameaças terroristas na região estudada.**

1. Introdução

O terrorismo, é um fenômeno que ameaça a estabilidade de países e regiões. O START - Study of Terrorism and Responses to Terrorism (Consórcio Nacional para o Estudo do Terrorismo e Respostas ao Terrorismo) [1], localizado na Universidade de Maryland (EUA), construíu o GTD (Global Terrorism Database) um banco de dados, resultante de várias fases de coleta de dados, livros, utilizando fontes públicas, como mídia e documentos legais, baseado em eventos que engloba mais de 200.000 registros de incidentes terroristas ocorridos globalmente desde 1970.

A coleta, manutenção e aprimoramentos do GTD têm sido apoiados por diversas organizações como o departamento de Segurança Interno dos EUA, a Universidade de Maryland, o Ministério Federal das Relações Exteriores da Alemanha e outras organizações americanas e internacionais. Reconhecendo as diversas definições de terrorismo, o GTD adota uma abordagem inclusiva, oferecendo mecanismos de filtragem para que os usuários possam personalizar o conjunto de dados de acordo com seus critérios específicos de definição. Isso garante uma ampla usabilidade, permitindo que os usuários adaptem os dados às suas necessidades.

Neste contexto é importante pensar em como estruturar atividades antiterroristas, com o foco no combate ao terrorismo, tanto em suas esferas organizacionais, estratégicas e táticas [2]. A partir desse conjunto de informações do GTD, a mineração de dados pode ser utilizada em benefício da sociedade ao dar suporte à tomada de decisões sobre o que fazer para ter maior eficiência no combate ao terrorismo. Devido aos aspectos socioeconômicos e culturais, similaridades econômicas e subdivisão político-cultural consolidada por diversas organizações internacionais, o trabalho é contextualizado na América Latina.

Utilizado para coagir, ameaçar ou oprimir uma população, muitas vezes, grupos recorrem ao uso sistematizado do terror. Na América Latina não é diferente, apesar dos países não terem problemas notáveis com a guerra ao terror dos países do hemisfério Norte, o terrorismo doméstico é um fator considerável na região, que também é utilizada por muitos grupos terroristas internacionais como um campo para o avanço de suas idéias [3]. Este trabalho visa indicar como a mineração de dados e o processo do KDD (Descoberta do conhecimento em bases de dados), podem ser utilizados pelas autoridades antiterrorismo dos 19 países da América Latina estudados, para ajudar com o suporte à decisão das autoridades.

2. Materiais e métodos

A base de dados disponibilizada no website da START, é composta por um dataset onde encontram-se diversas informações, dentre as quais, destacam-se a data do evento, o tipo de ataque, o grupo perpetrador, o tipo de alvo e, especialmente, a coluna ”Success”. Esta última revela se os objetivos dos terroristas foram alcançados, possibilitando uma compreensão mais profunda da eficácia de suas ações. Nesta coluna ”Success” foi realizada uma modificação para a classe alvo constar como o sucesso do combate ao ataque terrorista, ou, o fracasso dos terroristas em perpetrar o ataque.

Foi considerada uma amostra de 10.000 incidentes na região, para os países considerados na análise (Argentina, Bolívia, Brasil, Chile, Colômbia, Costa Rica, Cuba, República Dominicana, Equador, El Salvador, Guatemala, Honduras, México, Nicarágua, Panamá, Paraguai, Peru, Uruguai e Venezuela). Dos incidentes registrados, a classe

alvo contém 1588 ataques (15,88%) em que o combate ao ataque foi efetivo, ou seja, a variável "Success" possui o valor 1.

O trabalho foi realizado levando em consideração o processo do CRISP-DM (CRoss Industry Standard Process for Data Mining) [3]. Este modelo é padronizado para projetos de mineração de dados, desenvolvido por uma coalizão de líderes da indústria. É organizado em fases, como compreensão do negócio, compreensão dos dados, modelagem, avaliação e implementação, e também, oferece uma estrutura independente do setor e da tecnologia. O modelo visa tornar os projetos de mineração de dados eficientes, replicáveis e gerenciáveis, destacando a importância da compreensão do negócio, da qualidade dos dados e da avaliação rigorosa ao longo do processo, o fluxograma da **Figura 2**, detalha o processo que foi inspirado no CRISP-DM aplicado para a construção do trabalho.

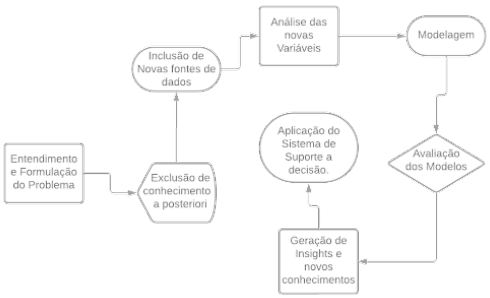

Figure 1. Fluxo de Processo do Trabalho

No desenvolvimento do trabalho, foram utilizadas as linguagens de programação R, para a limpeza, seleção da amostra, coleta de dados de diversas fontes (que será detalhada na subseção 2.1), a linguagem Python para a construção de um dashboard de análises estatísticas no dataset com o pacote `pandas-profiling` e a linguagem Javascript para a coleta de informações referentes aos feriados públicos dos países da América Latina.

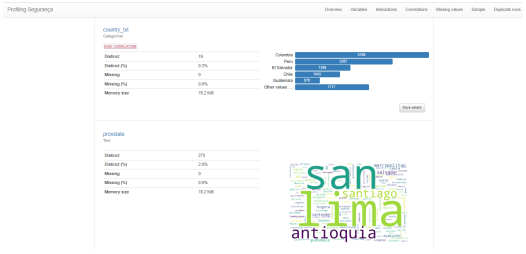

Figure 2. Overview do Dashboard construído pelo Pandas Profiling

2.1. Coleta de Informações

Para a coleta de algumas informações como o crescimento populacional do país, o crescimento populacional da América Latina e tamanho populacional, foram utilizadas as bibliotecas `wbstats` e `WDI` que possuem uma integração com a API do Banco Mundial. Referentes aos dias de feriado público foi utilizada a biblioteca `date-holidays` do javascript, já quanto a outras informações mais específicas sobre o país, como se houve eleição direta nacional no ano analisado, relação profissional militar do chefe de governo e quais são as cidades grandes e capital (ou capitais) do país, foram feitas pesquisas individual em cada país e, assim, foi construído um Datamart (como mostra a **Figura 3**) para incluir essas informações no dataset. Cabe mencionar que para a definição das cidades grandes foi utilizado o treshold da CIA de pelo menos 750.000 habitantes [5].

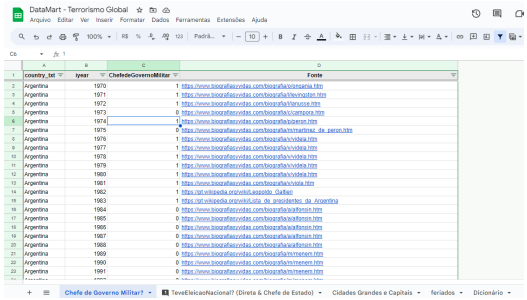

Figure 3. Overview do Dashboard construído pelo Pandas Profiling

Na tabela 1 estão sumarizadas as variáveis do dataset:

TABLE 1. DESCRIÇÃO DAS VARIÁVEIS

Variável	Descrição
country_txt	Nome do país
provstate	Nome do Estado
city	Nome da Cidade
success	Houve sucesso no combate ao ataque? (Dummy 1 - sim; 0 não)
nkill	Número de Mortes
nkillter	Número de Terroristas Mortos
nwound	Número de Feridos
nperpcap	Número de terroristas presos
nhostkid	Número de reféns
date	Data do atentado
weekday	Dia da semana do atentado
ChefedeGovernoMilitar	Chefe de estado tem relação militar (Dummy 1 - sim; 0 não)
eleicaonacionaldireta	Eleição nacional direta para chefe de estado no ano? (Dummy 1 - sim; 0 não)
CidadeOuCapital	Cidade Grande ou Capital (Dummy 1 - sim; 0 não)
TimeDifference_city_event	Tempo de diferença (em dias) em relação ao último ataque na cidade
TimeDifference_country_event	Tempo de diferença (em dias) em relação ao último ataque no país
TimeDifference_province_event	Tempo de diferença (em dias) em relação ao último ataque na província
feriado_oucomemo	Feriado público no dia (Dummy 1 - sim; 0 não)
SP.POP.GROW	Crescimento populacional do país no ano
popgrow_higher_than_latam	Crescimento populacional do país no ano maior que o da América Latina (Dummy 1 - sim; 0 não)
country_population	Tamanho da população no ano (no país)
tamanho_pais	Discretização com base nos quartis do tamanho populacional (em escala de milhões)
tempo_rel_ultimo_atk_city_ano	Variável categórica construída com base nos quartis que informa a quantidade de ataques na cidade (ano)
ataque_ult_ano_provincia	Variável binária que informa se houve ataque no último ano na província (Dummy 1 - sim; 0 não)
ataque_ult_ano_pais	Variável binária que informa se houve ataque no último ano no país (Dummy 1 - sim; 0 não)

As variáveis `nkill`, `nkillter`, `nwound`, `nperpcap` e `nhostkid`, são utilizadas como KPIs para avaliar conjuntamente com os escores de propensão o impacto do modelo no suporte à decisão.

2.2. Modelos Utilizados

Com o objetivo de extração do conhecimento do dataset, foi utilizado o modelo de árvores de decisão e o algoritmo JRip. O JRip é um algoritmo indutor de regras proposto por William W. Cohen como uma versão otimizada do algoritmo IREP. [6]. Já o modelo de árvore de decisão é um algoritmo de aprendizado supervisionado que possui uma estrutura em forma de fluxograma, com cada folha da árvore representando uma regra específica, as quais podem ser facilmente convertidas em regras de classificação, a árvore de decisão é uma ferramenta gráfica e intuitiva para tomar decisões com base em condições específicas, a validade das regra induzidas foi avaliada por meio das métricas Lift, Confiança e Cobertura [7].

Com o objetivo de dar suporte a decisão foi implementado um modelo de regressão logística. O que diferencia um modelo de regressão logística de um modelo de regressão linear é que, na regressão logística, a variável de resultado é binária ou dicotômica. Levando em conta essa distinção, os princípios gerais seguidos em uma análise que emprega regressão logística são os mesmos usados na regressão linear [8].

O modelo de regressão logística calcula a probabilidade de uma variável de resposta y pertencer a uma categoria específica com base em uma ou mais variáveis preditoras. A forma geral da equação é definida em 1.

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k \qquad (1)$$

- p é a probabilidade da variável dependente (y) pertencer à classe de interesse.
- $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ são os coeficientes estimados.
- $x_1, x_2, \ldots, x_k$ são as variáveis independentes.

Apenas variáveis com significância estatística ao nível de 5% foram incluídas no modelo.

2.3. Knime

A ferramenta utilizada para implementação dos modelos foi o KNIME que é uma plataforma open-source para análise de dados, mineração de dados, aprendizado de máquina e integração de dados. Ele oferece uma interface gráfica intuitiva que permite aos usuários criar fluxos de trabalho (workflows) analíticos, combinando diferentes etapas, desde a preparação e manipulação de dados até a construção de modelos preditivos.

O Workflow final implementado no Knime teve a configuração da Figura 4, seguindo um fluxo de modelagem com a regressão logística indicado pela seta vermelha e o fluxo de regras de indução com a função de adquirir conhecimento indicado pela seta azul.

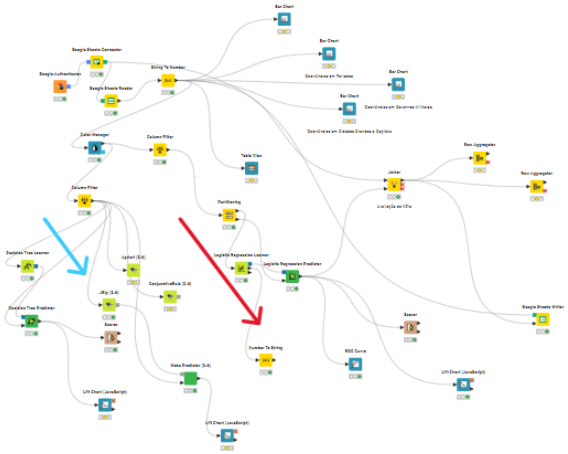

Figure 4. Overview do Workflow Knime

3. Resultados

Na regressão logística as variáveis incluídas no modelo estão presentes na Tabela 2, a variável "Chefe de Governo Militar = 1" possui um coeficiente negativo significativo (-0.29) com um p-valor de 0.006, indicando uma associação estatisticamente significativa com uma redução de aproximadamente 25.17% na chance de sucesso ao combate do ataque terrorista quando o chefe de governo é militar. Da mesma forma, a variável "tamanho_pais = mais de 30mi" (mais de 30 milhões de habitantes) apresenta um coeficiente positivo significativo (0.458) com um p-valor muito baixo (0.0001), sugerindo uma associação estatisticamente robusta com um aumento substancial de 58.09% na chance do sucesso ao combate do ataque.

Por outro lado, a variável "TimeDifference_city_event" possui um coeficiente próximo de zero (-0.0001) com um p-valor de 0.025, indicando uma associação estatística mais fraca, porém ainda significativa, com uma diminuição de 1.00% na chance de sucesso de combate ao ataque, para cada unidade de aumento na diferença de tempo entre o último ataque terrorista da cidade e o ataque atual.

Variável	Coeff.	P-valor	Chance
Chefe de Governo Militar = 1	-0.29	0.006	-25.17%
TimeDifference_city_event	-0.0001	0.025	-1.00%
SP.POP.GROW	-0.151	0.035	-14.02%
tamanho_pais = entre 8mi e 20mi	0.29	0.004	33.64%
tamanho_pais = mais de 30mi	0.458	0.0001	58.09%
tamanho_pais = menor ou igual a 8mi	-0.252	0.021	-22.28%
ataque_ult_ano_provincia = 1	0.157	0.036	17.00%

TABLE 2. COEFICIENTES DA REGRESSÃO LOGÍSTICA

O gráfico de lift de regressão logística apresentado (Figura 5) mostra que o lift acumulado é monotonicamente decrescente. Isso significa que, à medida que o percentual de observações aumenta, o lift acumulado diminui. Esse comportamento é esperado, pois o modelo está começando a classificar mais observações negativas como positivas (Ou

seja, afirmar que o combate será efetivo quando realmente o combate não é efetivo). Isso ocorre porque o modelo está sendo treinado para minimizar o erro de classificação, e isso pode levar a um aumento no número de falsos positivos.

Em geral, o gráfico mostra que o modelo é eficaz na classificação das observações positivas. No entanto, o limiar do escore de propensão utilizado para classificação deve ser escolhido com cautela, pois pode aumentar o número de falsos positivos.

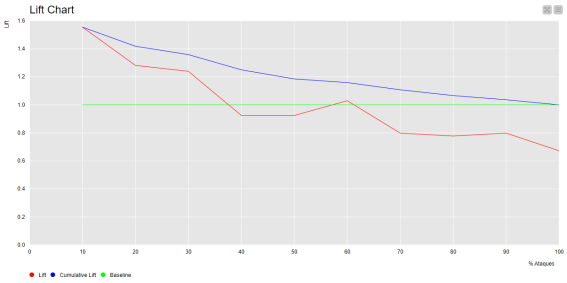

Figure 5. Gráfico da Curva de Lift

A árvore de decisão ficou com muitos ramos e folhas, portanto para representação do conhecimento obtido em um dos ramos da árvore foi escolhida a Figura 6, o lift em cada folha (calculado com base na classe alvo) é sempre acima de 1, foi escolhido um ramo em que existe uma maior probabilidade de sucesso do combate ao ataque em sua última folha.

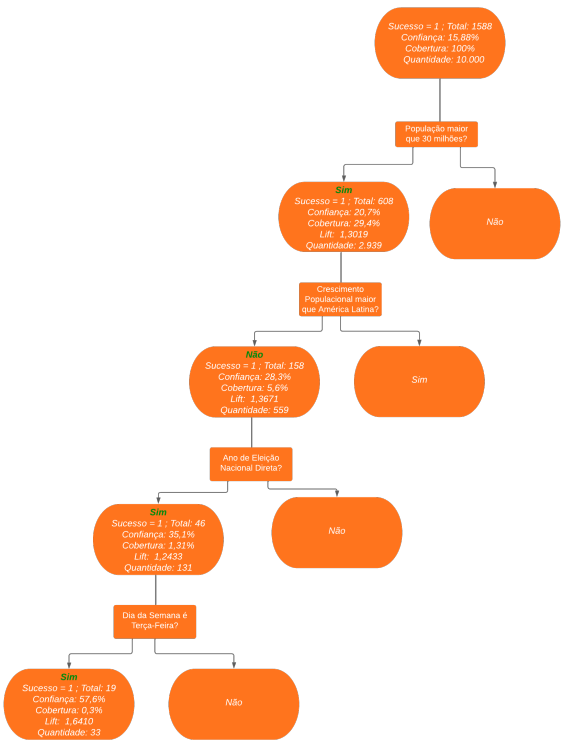

Figure 6. Ramo da Árvore de Decisão

A árvore de decisão começa com uma pergunta inicial: a população de um determinado país é maior do quê 30 milhões? Se a resposta for sim, então a probabilidade de sucesso é de 20,7%. Se a resposta for não, então a árvore de decisão prossegue para outra ramificação, mas seguindo a resposta positiva a árvore traz uma segunda pergunta: o crescimento populacional do país é maior do quê o da América Latina? Se a resposta for não, então a probabilidade de sucesso é de 28,3%. Se a resposta for sim, então a árvore de decisão prossegue para uma terceira pergunta: o país está em ano de eleição nacional direta? Se a resposta for sim, então a probabilidade de sucesso é de 35,1%. Se a resposta for sim, então a árvore de decisão prossegue para uma quarta pergunta: o dia da semana é terça-feira? Se a resposta for sim, então a probabilidade de sucesso de combate ao ataque terrorista é de 57,6%.

4. Discussão

Em geral, a árvore de decisão indica que a probabilidade de sucesso é maior em países com população maior do que 30 milhões, crescimento populacional menor do que o da América Latina e em ano de eleições nacionais diretas. A probabilidade de sucesso é ainda maior se o dia de semana for uma terça-feira, a tendência referente aos dias de semana também apareceu em algumas threads de regras do algoritmo JRip, que apesar de ter produzido algumas regras inconsistentes em ramos diferentes, produziu regras similares a ilustrada na thread da Figura 7.

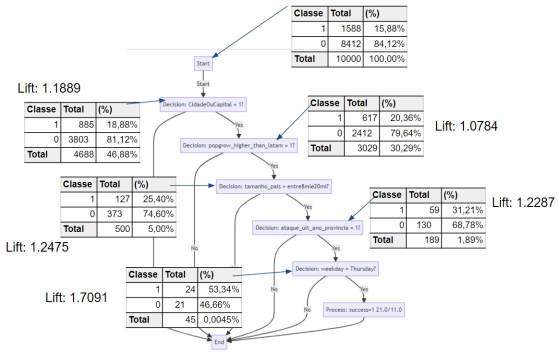

Figure 7. Thread de regras do algoritmo JRip

A recorrência dos dias da semana aparecendo como um fator que pode ajudar a definir ou não o sucesso do combate ao ataque terrorista faz sentido, apesar do contexto socio-econômico e cultural diferente da América Latina, em um estudo [9] no Paquistão é possível encontrar uma maior recorrência de ataques em determinados dia da semana, o trabalho afirma que: ”Quinta-feira e sexta-feira, sendo dias de visitas em larga escala a santuários e orações, respectivamente, são uma escolha racional para terroristas que desejam causar vítimas em massa” (FEYYAZ, 2023, p. 73, tradução nossa). Devido ao contexto diferenciado, esse fator pode ser melhor avaliado pelos especialistas do setor de segurança pública e autoridades de antiterrorismo.

4.1. Kpis e Escores da Regressão Logística

O gráfico das KPIs, com informações referentes às vítimas dos ataques, construído com base no escore de propensão da regressão logística (organizado em decis por ordem decrescente do escore), demonstra na Figura 8 que há um aumento nas instâncias 50 até 70 da quantidade de reféns de um ataque terrorista (vale ressaltar que na curva de lift, a partir dessas instâncias o modelo já estava inferior ao baseline), enquanto o número de mortes sofre flutuações ao longo das instâncias mas sobe rapidamente da instância 90 até a 100, já o número de feridos sofre flutuações ao longo dos decis.

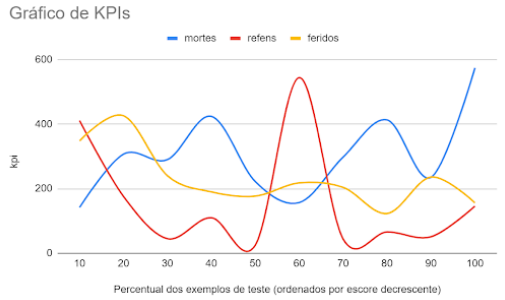

Figure 8. KPIs (quantidade de Mortes, Reféns e Feridos) ordenados por Escores de Propensão decrescente

O gráfico das mesmas KPIs com informações acumuladas (Figura 9), evidencia que há um crescimento acelerado na quantidade de reféns do decil 50 até o 60, justificando o salto no número de reféns, já o número de feridos e de mortes parece crescer de maneira linear porém constante ao longo das instâncias.

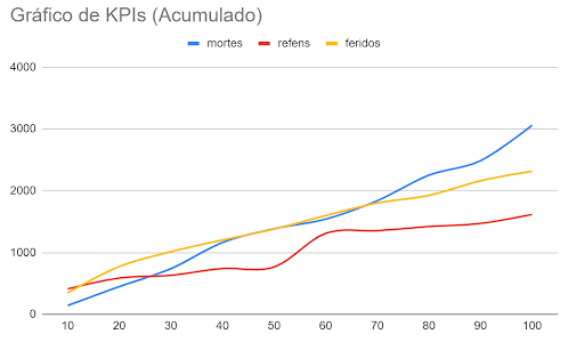

Figure 9. KPIs (acumulado de Mortes, Reféns e Feridos) ordenados por Escores de Propensão decrescente

A quantidade de terroristas capturados ou mortos (Figura 10) é maior nos primeiros decis, onde há uma maior propensão ao sucesso do combate a atividade terrorista, havendo uma bruca queda na quantidade de terroristas mortos ou capturados a partir da instância 60, voltando a crescer um pouco até o decil 100, mas ainda sendo inferior as instâncias iniciais.

Figure 10. KPIs (quantidade de terroristas mortos ou capturados) ordenados por Escores de Propensão decrescente

Já o gráfico acumulado da KPI de terroristas mortos e capturados (Figura 11) mostra uma estagnação da instância 50 até a 70, onde a propensão ao sucesso do combate ao ataque terrorista já é menor, ou seja a evolução da captura e neutralização da ameaça nessas instâncias já não parece ser interessante para o tomador de decisões.

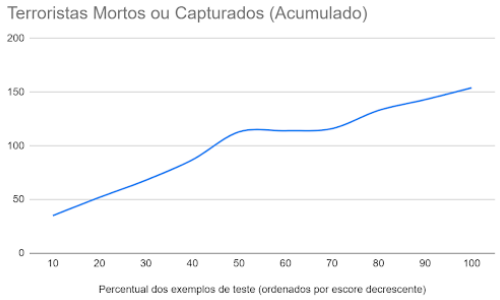

Figure 11. KPIs (acumulado de terroristas mortos ou capturados) ordenados por Escores de Propensão decrescente

O gráfico da KPI principal (Figura 12), que é a taxa de vítimas mortas por terroristas mortos no ataque, mostra um aumento repentino da taxa de mortalidade de vítimas por terroristas nos decis 50 até o 70, evidenciando que quando a propensão de sucesso ao combate do ataque é menor, o número de vítimas tende a aumentar.

Figure 12. KPIs (taxa de mortalidade de vitimas por terroristas) ordenados por Escores de Propensão decrescente

O gráfico acumulado da taxa de mortalidade de vítimas

por terroristas (Figura 13) é relativamente baixa no primeiro decil, crescendo timidamente até a instância 50, mas aumenta rapidamente nos decis seguintes. De fato, levando em conta as Kpis apresentadas com base no escore de propensão conjuntamente com o gráfico da curva de lift (Figura 5), o ponto de operação a ser escolhido deveria estar entre as instâncias 30 ou 40, onde seria possível direcionar mais recursos, visto que, a propensão de sucesso ao combate ao ataque é maior nesses cenários, o modelo ainda possui um lift consideravelmente maior que uma chance aleatória e não há um crescimento acelerado da taxa de vítimas por terroristas ao mesmo tempo que a quantidade de terroristas capturados ou mortos é elevada.

Figure 13. KPIs (taxa acumulada de mortalidade de vitimas por terroristas) ordenados por Escores de Propensão decrescente

5. Conclusão

As informações que podem ser levantadas a partir dos dados, utilizando metodologias de mineração de dados e conhecimento, são valiosas, porque junto com um sistema de suporte a decisão binária podem ajudar no gerenciamento eficaz de recursos para o combate a ataques terroristas, evidenciando cenários em que há uma alta propensão ao sucesso do combate antiterrorista, com isso ajudando a salvar vidas de inocentes e mitigando danos materiais ou sociais que um ataque pode causar na sociedade.

Esses resultados sugerem que a regressão logística é um modelo eficaz para prever a probabilidade de morte em ataques terroristas. À medida que o escore de propensão aumenta, a probabilidade de uma vítima morrer em um ataque terrorista também aumenta. É importante ressaltar que esses resultados são apenas uma estimativa, pois são baseados em um conjunto de dados específico.

Para obter resultados mais precisos, é necessário realizar testes com outros conjuntos de dados. Além disso, é importante que o tomador de decisões e os encarregados de realizar políticas antiterroristas considerem outros fatores que podem influenciar a probabilidade de morte em ataques terroristas, como o tipo de ataque, o local do ataque e as condições políticas e sociais do país onde o ataque ocorre.

References

[1] Global Terrorism Database. Codebook. Disponível em: https://www.start.umd.edu/gtd/downloads/Codebook.pdf. Acesso em: 07/12/2023.

[2] KRAUSE, Peter. When terrorism works: success and failure across different targets and goals. Revista CIDOB d'Afers Internacionals, v. 112, n. 1, p. 69-91, 2016. DOI: 10.24241/rcai.2016.112.1.69.

[3] SULLIVAN, Mark P. Latin America: Terrorism Issues. In: LINDEN, Edward V. (Ed.). Focus on Terrorism, Volume 9. UK ed. Nova Science Pub Inc, 2007. 328 p. ISBN: 978-1600217098.

[4] Wirth, Rüdiger and Jochen Hipp. "Crisp-dm: towards a standard process model for data mining." (2000).

[5] CIA. The World Factbook. Major urban areas - population. Disponível em: https://www.cia.gov/the-world-factbook/field/major-urban-areas-population/. Acesso em: 07/12/2023.

[6] Weka. (n.d.). JRip (Weka 3.6). Weka Documentation. https://weka.sourceforge.io/doc.dev/weka/classifiers/rules/JRip.html. Acesso em: 10/12/2023.

[7] Han, Jiawei; Kamber, Micheline; Pei, Jian. Mineração de Dados: Conceitos e Técnicas, 3. ed. The Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann Publishers, julho de 2011. ISBN 978-0123814791.

[8] HOSMER, D. W.; LEMESHOW, S. Applied logistic regression. [S.l.]: John Wiley and Sons, 2000. ISBN 0471356328, 9780471356325

[9] Feyyaz, M. (2023). "Conceptualising Terrorism Trend Patterns in Pakistan - an Empirical Perspective." Perspectives on Terrorism, 7(1), 73-102. Disponível em: http://escholar.umt.edu.pk:8080/jspui/bitstream/123456789/872/1/243-1637-2-PB.pdf