FEP FACULDADE DE ECONOMIA
UNIVERSIDADE DO PORTO

Ano Letivo de 2015/2016

Mestrado em Modelação, Análise de Dados e Sistemas de Apoio à Decisão

Unidade Curricular de Extração de Conhecimento de Dados II

RELATÓRIO DO TRABALHO PRÁTICO II

"Modelação de um problema de Phising utilizando redes bayesianas"

Docentes: Professor Dr. João Gama
Professor Dr. Pavel Bradzil
Discentes: Rui Pedro Machado, 201300292
Hélder Filipe Russa, 201508409

INDICE

1. INTRODUÇÃO

No âmbito da disciplina de Extração de Conhecimento de Dados II, do Mestrado em Modelação, Análise de Dados e Sistemas de Apoio à Decisão da Faculdade de Economia da Universidade do Porto, 2015/2016, foi desenvolvido este trabalho, cujo tema central são as redes *Bayesianas* pretendendo-se modelar um determinado problema recorrendo a redes *Bayesianas* e efetuar algumas experiências de inferência (Diagnóstico e Predição) com a mesma. Para a concretização deste objetivo, foi escolhido pelo grupo gerar uma rede *Bayesiana* a partir de um conjunto de dados, por não considerarmos ter a base necessária de conhecimento de domínio para modelar o problema a partir do mesmo.

1.1 Redes Bayesianas

As redes Bayesianas são modelos gráficos, baseados em probabilidades, que permitem analisar e retirar conclusões em ambientes de incerteza de decisão, modelando os problemas através de variáveis e respetivas relações entre as mesmas. De um ponto de vista mais estatístico, as redes Bayesianas representam a probabilidade conjunta entre um conjunto de variáveis que caracterizam o problema em análise. (Wikipedia, 2016)

A forma típica de representação de uma rede Bayesiana é através de grafos direcionados acíclicos, também denominados DAG, sendo acíclicos por não possuírem ciclos. Estes grafos são composto por nós (vértices) e por arcos, onde cada nó representa uma variável e os arcos representam a dependência entre as variáveis. Cada nó além de caracterizar uma variável, guarda ainda as probabilidades condicionadas das várias combinações possíveis de valores dessa mesma variável com o nó imediatamente anterior (Relação entre variável explicativa/independente para com a variável explicada/dependente). (Koski & Noble, 2009)

A ilustração seguinte demonstra um caso típico de aplicação de redes Bayesianas da análise de uma determinada condição médica (Neste caso gravidez), dado um conjunto de variáveis explicativas (Sintomas). (Chan & Darwiche, 2002)

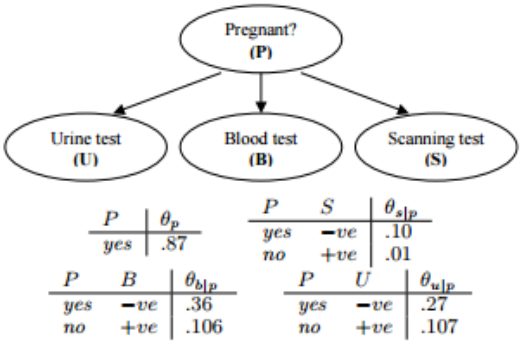

Figura 1 - Rede Bayesiana Simples

Tal como é possível verificar, para cada nó temos uma tabela de contingência que relacionam a probabilidade de uma mulher estar grávida sabendo um conjunto de sintomas.

As redes Bayesianas podem ter todas as distribuições de probabilidades nas suas variáveis, apoiando qualquer direção de raciocínio e, para isso, utilizam quatro tipos de raciocínio, nomeadamente:

- ***Diagnóstico -*** Permite a construção das probabilidades a partir dos sintomas até à causa;
- ***Preditivo -*** Verificar o impacto de novas informações sobre os sintomas, nas causas;
- ***Intercausal -*** Permite o raciocínio acerca das causas mútuas que determinam um efeito comum;
- ***Combinado -*** Conjuga o raciocínio por diagnóstico e o raciocínio preditivo;

Na elaboração deste trabalho prático serão considerados dois tipos de raciocínio: por diagnóstico e preditivo.

2. DESCRIÇÃO DO PROBLEMA

Na elaboração deste trabalho os elementos do grupo, dado terem formação académica na área da informática, decidiram conciliar um problema dessa área com o objetivo de construir uma rede Bayesiana. O tema escolhido foi o *Phishing* um problema informático real e atual que continua a trazer problemas quer a empresas com negócios *online*, quer aos utilizadores desses mesmos websites.

2.1 Phishing

O *Phishing* refere-se a uma técnica informática, maliciosa, que visa adquirir de forma ilegal dados confidenciais tais como morada ou número de cartão de crédito acerca de utilizadores de um dado website. Esta técnica funciona de uma forma simples, um determinado hacker informático cria uma cópia do website que pretende atacar (*Phishing website*), de tal forma que se assemelhe ao original e de seguida envia um email a um conjunto utilizadores, fazendo-se passar pela entidade a quem estão a roubar a identidade. Neste email tipicamente o texto pede que os utilizadores partilhem informação pessoal por forma a corrigir alguma lacuna de informação nas bases de dados ou até para o premiar com uma oferta irrecusável, sendo que para tal basta utilizar o endereço de uma página (*Phishing* website) e fazendo *login*. (Shinde, 2013)

Quando os utilizadores se autenticam, todos os dados estão a ser enviados para uma base de dados paralela, detida pelo hacker, que poderá por exemplo clonar um cartão de crédito e provocar danos financeiros aos utilizadores. Este fenómeno informático traz problemas quer aos utilizadores, vitimas diretas, quer às entidades cuja identidade é roubada, vitimas indiretas. (Shinde, 2013)

A motivação e objetivo deste trabalho é de através da análise de um conjunto de variáveis relacionadas com tráfego e pedidos de rede, construir uma rede Bayesiana que permita analisar de um ponto de vista de sintomas e causa, quais as variáveis que mais explicam um ataque deste género e perceber como é que com a monitorização e observação de valores de uma ou mais variáveis é possível prever se estamos perante um ataque de *Phishing* ou não.

Este tipo de estratégia de deteção de websites de *Phishing,* recorrendo *a datamining,* é uma das mais recentes formas de minimizar o impacto deste tipo de ataque, permitindo detetar e eliminar páginas na fonte, utilizando por exemplo a construção do endereço da página, ao invés de simplesmente filtrar emails considerados maliciosos. (Shinde, 2013)

2.2 Conjunto de dados utilizado

O conjunto de dados utilizado foi recolhido do repositório de dados da Universidade da Califórnia (*UCI Machine Learning Repository*), com o nome "Phishing Websites Data Set" e contém um conjunto de variáveis que os autores identificam como principais para a deteção de um website de *Phishing*. O conjunto de dados pode ser encontrado no endereço seguinte Conjunto de Dados UCI. No subcapítulo seguinte fazemos uma apresentação das variáveis contidas no conjunto.

2.2.1 Variáveis do Modelo

De seguida apresentamos as variáveis utilizadas sendo que todas á exceção da variável (*Phishing*) devem ser interpretadas como sintomas de se tratar de uma página maliciosa enquanto que a última deve ser interpretada como a causa. Foram utilizadas todas as variáveis na construção da rede *Bayesiana*.

- ***having_IP_Address*** - Se é utilizado um endereço IP como alterativa ao nome de domínio do website. (Ex: http://123.234.12.1/home.html em vez de http://www.madsad.com/home.html)
- ***URL_Length*** - Se o endereço do website é longo (Provável que seja *Phishing*) ou curto
- ***Shortining_Service*** - Se o endereço utiliza um serviço de encurtamento de endereços (Como o TinyURL)
- ***having_At_Symbol*** – Se o endereço contém o símbolo "@"
- ***double_slash_redirecting*** - Se o endereço contém o símbolo "//" o que pode significar que o utilizador vai ser redirecionado para uma outra página
- ***Prefix_Suffix*** - Se o endereço contém o símbolo "-" o que raramente acontece com endereços verdadeiros.
- ***having_Sub_Domain*** - Se o endereço contém "." além de na primeira parte do domínio

- ***SSLfinal_State*** – Se a página tem certificado com emissor reconhecido SSL (*https* versus *http*)
- ***Domain_registeration_length*** **-** Antiguidade do domínio (Em anos)
- ***Favicon*** **-** Se a imagem do website no browser é carregada de uma fonte externa.
- ***port*** **-** Se utiliza a porta standard no Mercado para o protocolo em causa (Ex: FTP utiliza 21 e HTTP utiliza 443)
- ***HTTPS_token*** - Se é um pedido *http* mas contém "*https*" no endereço para enganar o utilizador
- ***Request_URL*** – Se o pedido contém objetos carregados de fontes externas. (Três estados possíveis: -1 = "< 22% fontes externas"; 0 = "Entre 22% a 61%"; 1 = "> 61 (Phishing)"
- ***URL_of_Anchor*** – Se as ancoras da página referenciam websites de domínios externos. (Três estados possíveis: -1 = "< 31% fontes externas"; 0 = "Entre 31% a 67%"; 1 = "> 67 (Phishing)"
- ***Links_in_tags*** - Se as links da página referenciam websites de domínios externos. (Três estados possíveis: -1 = "< 17% fontes externas"; 0 = "Entre 17% a 81%"; 1 = "> 81 (Phishing)"
- ***SFH*** – Se os "Server Form Handlers" contêm *strings* vazias ou se referenciam um domínio diferente.
- ***Submitting_to_email*** - Se a informação de um formulário está a ser enviada para um email. (Ex: utilização da função mail())
- ***Abnormal_URL*** **-** Se o endereço contém o "*host name*" ou não.
- ***Redirect*** **-** Número de redireccionamentos existentes na página. (Websites legítimos têm normalmente 1 redireccionamento no máximo)
- ***on_mouseover*** **-** Se utilizam JavaScript para mostrar endereços falsos na barra de estado.
- ***RightClick*** **-** Se o botão direito está desativado para evitar que se veja o código fonte da página.
- ***popUpWidnow*** **–** Se utilizam janelas *pop-up* para solicitar informação aos utilizadores.
- ***Iframe*** **-** Se utilizam *iframe* para mostrar websites dentro de websites. (Tipicamente sites legítimos não usam esta estratégia)

- ***age_of_domain*** **-** Se a idade do domínio é superior a 6 meses ou não. (Domínios recentes podem indicar *Phishing*)
- ***DNSRecord*** **–** Se existem registos em servidores *DNS* ou não.
- ***web_traffic*** **–** Se o website tem uma popularidade alta ou não em termos de tráfego. (Valores possiveis: -1 legitimo (<100.000) ; 1 valores suspeitos (>100.000))
- ***Page_Rank*** **–** Se o website tem popularidade ou não em termos de *PageRank***.**
- ***Google_Index*** – Se a página está indexada pelo Google ou não.
- ***Links_pointing_to_page*** **–** Se tem duas ou mais páginas a apontar para este endereço ou não. (Não ter referencias pode indicar *Phishing*)
- ***Statistical_report*** – Se existe alguma referencia em sites de deteção de *Phishing* ou não.
- ***PhisingAttack***– Varável alvo que indica se é uma página de Phishing ou não com todos os sintomas indicados anteriormente.

A descrição das variáveis é baseada na dada por Mohammad, Thabtah, & Mccluskey (2013).

3. DESENVOLVIMENTO

3.1 Construção da Rede Bayesiana

Para a construção da rede bayesiana foi utilizado o software Genie, plataforma abordada nas aulas e recomendada para a elaboração deste trabalho. Nesta plataforma existem algumas tarefas possíveis e que devem ser realizadas antes da construção da rede propriamente dita, nomeadamente (1) análise univariada das variáveis, (2) discretização das mesmas, (3) tratamento de casos omissos bem como (4) remoção ou tratamento de outliers. Por último dado que a nossa análise recai sobre uma variável "PhishingAttack", (5) é ainda realizada uma interdição de dependência entre esta variável e todas as outras e só posteriormente passamos à contrução da rede.

3.1.1 Análise univariada dos dados

Através desta análise preliminar posdemos não só observar a dispersão das variáveis como também identificar através da contagem de casos, valores minimo e máximos de cada variável e se exite algum caso omisso. Tal como é possivel constatar, tal não é o caso. Importa recordar que as variáveis são na sua maioria valores *booleanos*.

Variável	Mean	Variance	StdDev	CV	Min	Max	Count
having_IP_Address	0,314	0,902	0,950	105%	-1	1	11055
URL_Length	-0,633	0,587	0,766	131%	-1	1	11055
Shortining_Service	0,739	0,454	0,674	148%	-1	1	11055
having_At_Symbol	0,701	0,509	0,714	140%	-1	1	11055
double_slash_redirecting	0,741	0,450	0,671	149%	-1	1	11055
Prefix_Suffix	-0,735	0,460	0,678	147%	-1	1	11055
having_Sub_Domain	0,064	0,668	0,818	122%	-1	1	11055
SSLfinal_State	0,251	0,832	0,912	110%	-1	1	11055
Domain_registeration_length	-0,337	0,887	0,942	106%	-1	1	11055
Favicon	0,629	0,605	0,778	129%	-1	1	11055
port	0,728	0,470	0,685	146%	-1	1	11055
HTTPS_token	0,675	0,544	0,738	136%	-1	1	11055
Request_URL	0,187	0,965	0,982	102%	-1	1	11055
URL_of_Anchor	-0,077	0,511	0,715	140%	-1	1	11055

Links_in_tags	-0,118	0,584	0,764	131%	-1	1	11055
SFH	-0,596	0,576	0,759	132%	-1	1	11055
Submitting_to_email	0,636	0,596	0,772	130%	-1	1	11055
Abnormal_URL	0,705	0,503	0,709	141%	-1	1	11055
Redirect	0,116	0,102	0,320	313%	0	1	11055
on_mouseover	0,762	0,419	0,647	154%	-1	1	11055
RightClick	0,914	0,165	0,406	246%	-1	1	11055
popUpWidnow	0,613	0,624	0,790	127%	-1	1	11055
Iframe	0,817	0,333	0,577	173%	-1	1	11055
age_of_domain	0,061	0,996	0,998	100%	-1	1	11055
DNSRecord	0,377	0,858	0,926	108%	-1	1	11055
web_traffic	0,287	0,685	0,828	121%	-1	1	11055
Page_Rank	-0,484	0,766	0,875	114%	-1	1	11055
Google_Index	0,722	0,479	0,692	144%	-1	1	11055
Links_pointing_to_page	0,344	0,325	0,570	175%	-1	1	11055
Statistical_report	0,720	0,482	0,694	144%	-1	1	11055
Result	0,114	0,987	0,994	101%	-1	1	11055

Figura 2 Análise univariada dos dados

3.1.1.1 Discretização de Variáveis

No nosso conjunto de dados as variáveis já se econtravam discretizadas, pelo que tal tarefa não irá ser realizada. Poderiamso no entanto fazê-lo facilmente através da ferramenta Genie, tal como demonstrado na figura seguinte.

Figura 3 - Ferramenta de discretização do Genie

3.1.1.2 *Tratamento de Casos Omissos*

Esta tarefa pode igualmente ser realizada através da ferramenta Genie, na opção "*Missing Values -> Select/Replace -> Select all rows with missing values"* e no nosso conjunto de dados não foi encontrado qualquer valor omisso.

3.1.1.3 *Tratamento de Outliers*

Dada a natureza das nossas variáveis, não existem *outliers*, pelo que não houv necessidade de realizar esta tarefa.

3.1.2 Construção da Rede Bayesiana

De seguida construimos a rede bayseana utilizando a opção "*Data -> Learn New Network*". Nesta janela do Genie é necessário tomar algumas decisões, nomeadamente (1) conhecimento prévio de relação de variáveis, (2) as variáveis a utilizar, (3) o algoritmo de treino bem como (4) configuração de parametros do algoritmo selecionado.

Tal como já foi previamente dito para a decisão (1) e dado que o output desejado para este estudo é informação respeitante à variável PhishongAttack foi necessário colocar interdições de dependência desta variável para qualquer uma das outras variáveis. A figura seguinte demonstra construção de tal restrição no *software Genie*.

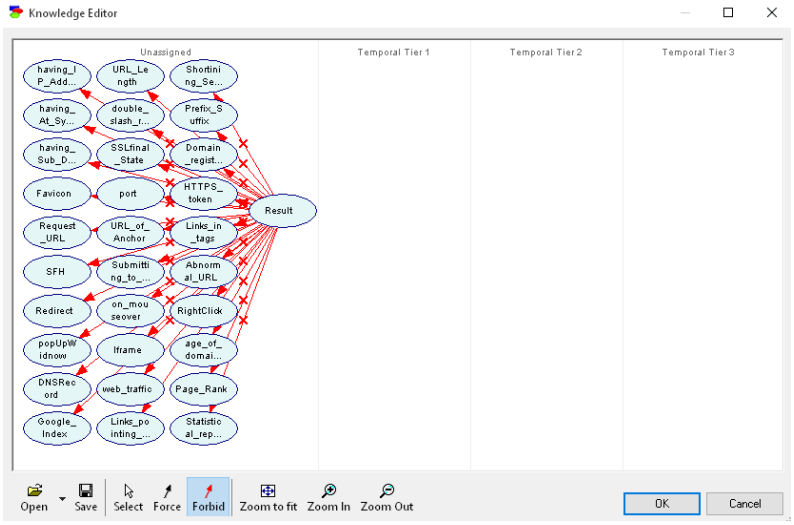

Figura 4 - Aprendizagem supervisada

Relativamente à decisão (2), variáveis a utilizar, a decisão foi recorrer a todas para que possamos analisar de uma forma holistica todas as propagação de evidência que acharmos relevantes e identificar as variáveis mais e menos afetadas por tal propagação.

A decisão (3) foi provavelmente a mais complicada dado o conjunto de algoritmos disponiveis na plataforma, no entanto optamos por recorrrer ao algoritmo *Greedy Thick Thinning* com todos os parametros (Decisão 4) com os valores default.

Este algoritmo é baseado em restrição determinística e o seu método é simples, inicia o processo com um grafo vazio e, repetidamente, adiciona o próximo arco que maximize a métrica da pontuação bayesiana, de modo a atingir um Máximo Local. Importa salientar que este algoritmo tem um parametro importante (*Priors*), respeitante à forma de pesquisa e ranking. Este método, K2, inicia com uma ordenação dos vértices, que é utilizada para a sua maximização, e depois pesquisa directamente as muitas estruturas existentes no espaço dos grafos para encontrar a melhor pontuação. A figura seguinte demonstra o resultado obtido com este algoritmo.

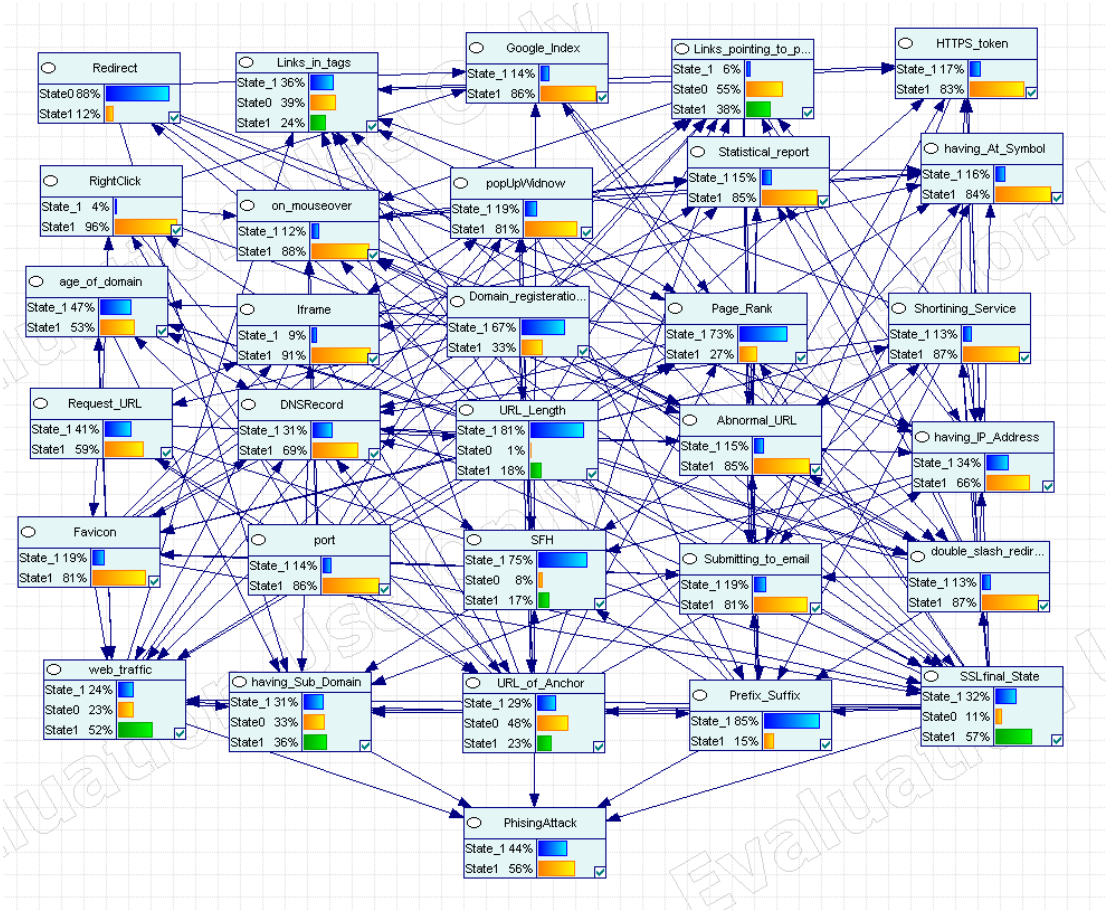

Figura 5 - Rede Bayesiana Gerada

3.2 Diagnóstico

Um raciocínio por diagnóstico permite tal como foi dito observar um determinado valor da variável de saída (*Output*) e propagar tal evidencia nas variável de entrada para que desta forma consigamos compreender quais as variáveis de entrada (Sintomas) que poderão ter um valor explicativo maior na variável de saída (Causa).

No nosso problema fixamos um valor da variável booleana *Phishing* (Sim ou Não) e propagamos tal evidência na rede.

Assim sendo na ilustração seguinte podemos observar o resultado de ficar na nossa rede a evidência de se ter observado um website de *Phishing* (Sim) versus ser um website legitimo.

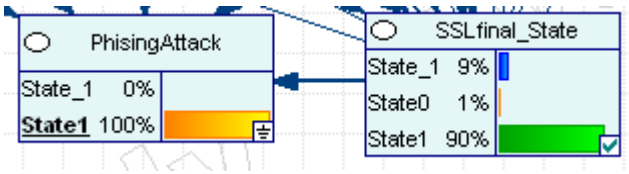

Figura 6 Obervação de um website de phishing

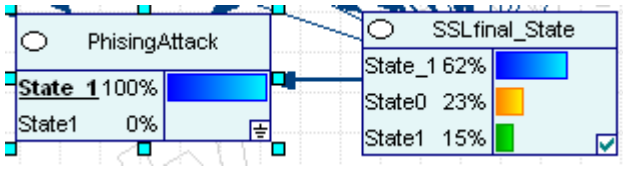

Figura 7 - Obervação de um website normal

Tal como é possível constatar neste exemplo a variável *SSLfinal_state,* que identifica se a página tem um certificado SSL valido é bastante afetada mediante a evidencia obbservada da variável de saída. Neste caso especifico podemos verificar que quando observamos um website de Phishing, 90% destes não têm um certificado valido, sendo assim uma boa variável de diagnóstico.

Podemos tambem analisar outra variável, *web_traffic,* relacionada com a popularidade do website e ver como se comporta quando temos uma observação de um website de *Phishing* versus ser legitimo.

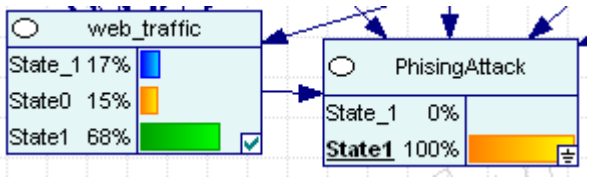

Figura 8 - 6 Obervação de um website de phishing

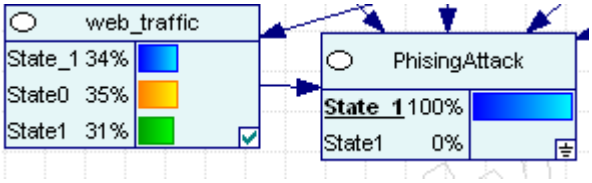

Figura 9 - Obervação de um website normal

Olhando para esta variável é possivel verificar que quando se trata de um website normal, as distribuições de probabilidade são bastante equilibradas mas quando obervamos um website de *Phishing*, passa a existir 68% de valores demasiado elevados para serem considerado um website normal. Desta forma esta também é uma variável bastante explicativa.

3.3 Previsão

Tal como foi enunciado previamente, este tipo de raciocínio em redes *Bayesianas* permite que através da observação do valor de um dado sintoma (Variável de Entrada) possamos propagar tal evidência e observar o impacto nas causas (Variáveis de saída). Comecemos por analisar o impacto de observar que a idade do dominion apresenta um valor inferior a 6 meses, o que de acordo com a descrição feita previamente da variável, poderá indicar (Sintoma) um website de *Phishing,* dado que existe maior probabilidade de tal ocorrer em dominios recentes.

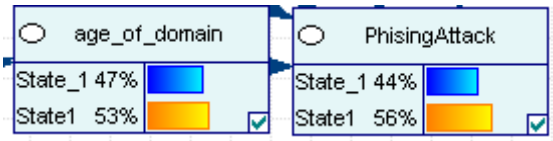

Figura 10 - Estado inicial

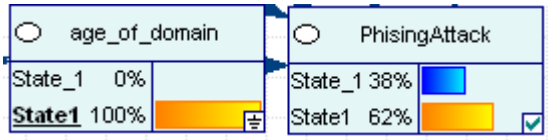

Figura 11 - Propagação da evidência de ser um endereço com idade inferio r a 6 meses

Tal como se pode constatar através da propagação de tal evidência, a probabilidade de ser um website de *Phishing* subiu 6 pontos percentuais, permitindo constatar que se se observar que um dado endereço tem idade inferior a 6 meses, a probabilidade de se tratar de um website ilegal aumenta.

Passemos agora para um outro exemplo. Neste caso vamos supor que observamos duas evidências, nomeadamente (1) que se trata de um endereço com idade inferior a 6 meses tal como o anterior mas acrescentando outra evidência, neste caso, (2) que a variável *Request_URL,* que representa a percentagem de objectos que contém fontes externas, apresenta o estado 1 (> 61% objectos de fontes externas).

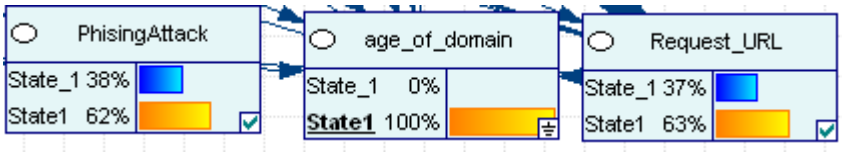

Figura 12 - Estado original

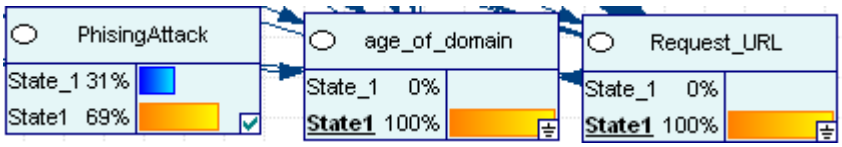

Figura 13 - Propagação da evidência de ser um endereço com idade inferio r a 6 meses e Request_URL > 61%

Tal como é possível verifiicar nas imagens seguintes, simplesmente observando estas duas variáveis e propagando tal evidência na rede, é possível prever que se trata, com 69% de probabilidade, de um website de *Phishing*. Face à previsão anterior, melhoramos 7 pontos percentuais à nossa previsão.

Por último realizamos uma outra experiência de previsão que consistiu em propagar a evidência de a variável *Favicon,* que representa se a imagem do website no browser é carregada ou não de uma fonte externa, ser de facto de uma fonte externa, acrescida às duas evidências anteriormente propagadas.

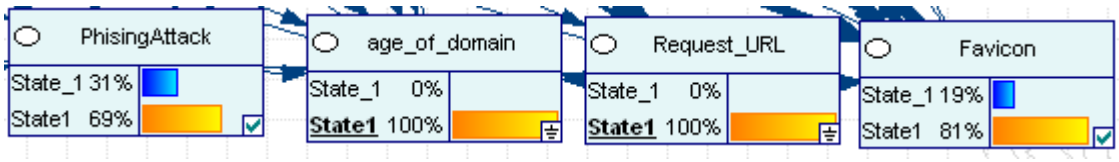

Figura 14 - Estado inicial

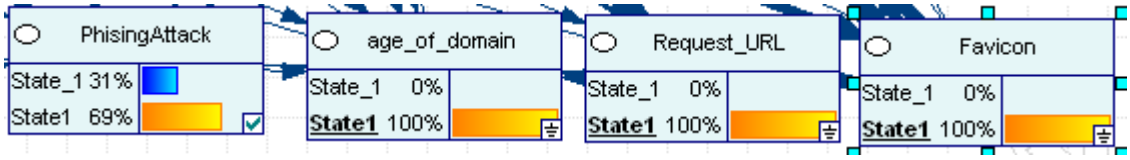

Figura 15 - Propagação da evidência de ser um endereço com idade inferio r a 6 meses e Request_URL > 61% e Favicon ser de uma fonte externa

Tal como se pode observar esta variável não ter qualquer impacto nas probabilidade do facto de ser um website de *Phishing* ou não. Isto indica que as variáveis *Favicon* e *PhishingAttack* não se encontram correlacionadas e tal pode ser observado olhando para a matriz de correlações dada igualmente pela plataforma Genie. Na ilustração seguinte é demonstrado o resultado da analise de correlação entre as quatro variáveis do exemplo anterior.

	Favicon	Request_URL	age_of_domain
PhishingAttack	-0,00028	**0,25337**	**0,12150**

Figura 16 - Coeficiente de correlação linear

Analisando a tabela anterior, constatamos o facto previamente identificado, de que de facto existe uma correlação quase inexistente entre a variável *Favicon* e *PhishingAttack*, daí o valor da última não variar com a propagação de evidência da primeira. Por outro lado explica-se através desta tabela o porquê de o contrário acontecer quando propagamos as duas variáveis seguintes. A variável request_URL tem um coeficiente positivo o que indica que quando o Request_URL varia de forma positiva, a variável PhishingAttack varia no mesmo sentido, tal como acontece com a variável age_of_domain, sendo que a correlação desta última é consideravelmente inferior à primeira. Por outras palavras se constarmos que o request_URL ou o age_of_domain apresenta estado 1, a propabibilidade de ser um ataque de Phishing aumenta de forma relevante.

4. CONCLUSÃO

A construção de redes Bayesianas provou ser um método simples e eficaz de perceção de relação entre variáveis e de que forma a previsão de um dado acontecimento pode ser gerido em termos de evidências. Com isto referimos que, tendo o nosso caso como exemplo, caso uma empresa queira desenvolver um mecanismo de deteção de de websites de *Phishing* com todas as variáveis de identificamos, o esforço pode ser elevado e ter um custo de tempo demasiado elevado para o objetivo da empresa.

Assim através de uma rede bayesiana é possivel identificar qual os sintomas que mais explicam uma dada causa e investir por ordem de prioridades em ferramentas ou até agentes que possam assegurar a deteção de tais evidências e numa ótica de previsão inserir tais evidências numa rede previamente desenvolvida e verificar com um determinado nível de probabilidade se é ou não um website de *Phishing*.

Por último importa realçar que o software Genie, com uma curva de aprendizagem reduzida, permite fazer diversas experiências e tratamento de dados de forma simples, permitindo antecipar o retorno que se pode obter com uma automatização de uma rede para diganóstico e previsão.

Por último de referir que a única limitiação identificada pelo grupo quanto a esta ferramente foi a impossibilidade de se efetuar uma avaliação à rede produzida. Por exemplo não foi possivel fazer uma validação cruzada entre dados de treino e teste, tarefa que seria facil de realizar numa outra plataforma como KNIME ou até mesmo com a linguagem R ou Python.

5. BIBLIOGRAFIA

Chan, H., & Darwiche, A. (2002). Reasoning about Bayesian Network Classifiers.

Koski, T., & Noble, J. M. (2009). *Bayesian networks: an introduction*. *Wiley Interdisciplinary Reviews: Computational …*.

Mohammad, R. M., Thabtah, F., & Mccluskey, L. (2013). Phishing Websites Features.

Shinde, S. K. (2013). Detection of Phishing Websites Using Data Mining Techniques, *2*(12), 3725–3729.

Wikipedia. (2016). Bayesian Networks. Retrieved from https://pt.wikipedia.org/wiki/Rede_bayesiana