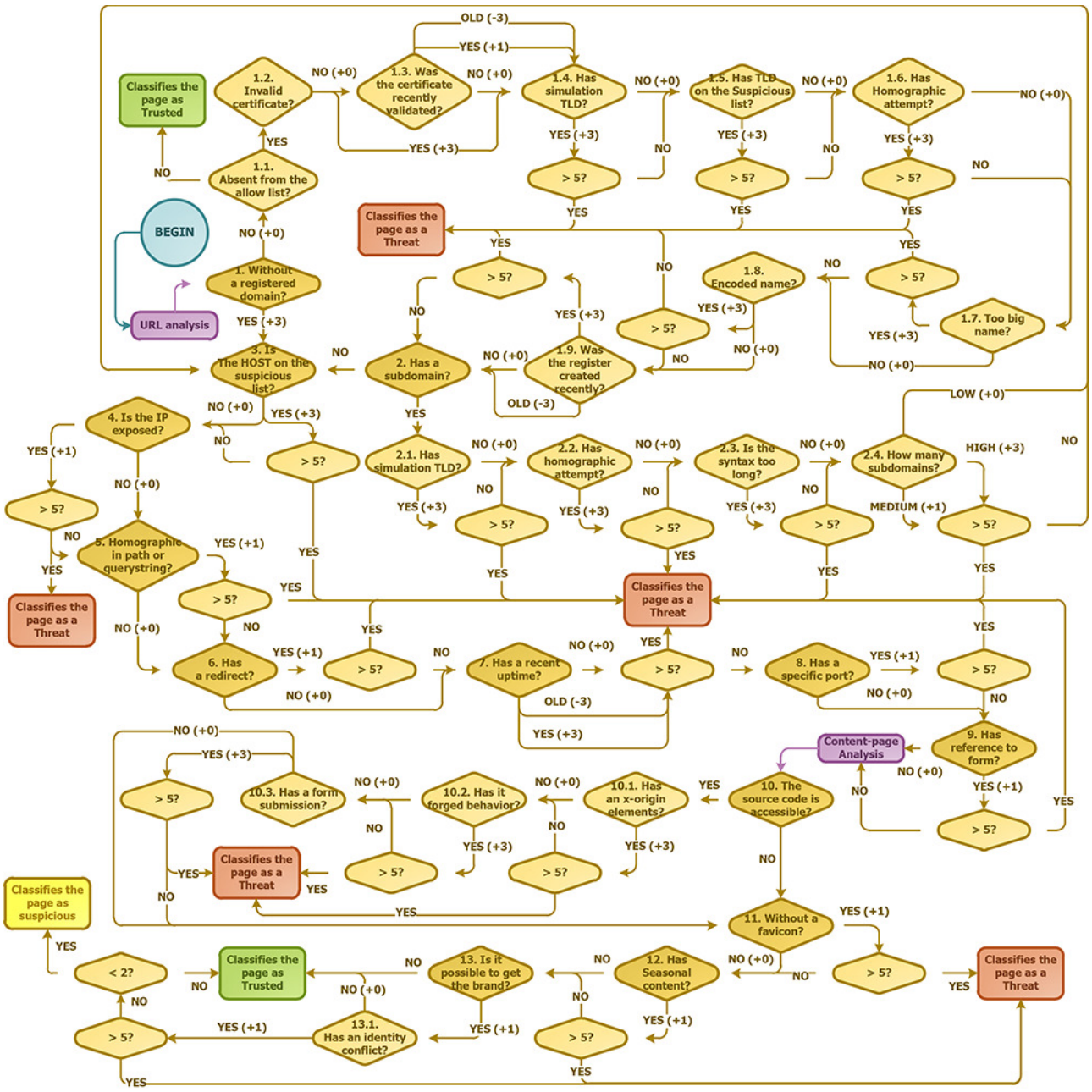

Figure 1: The rules-based model tree representation

A. Appendix

As a way of elucidating the steps of the inference engine, based on the rules-based tree illustrated in Figure 1, the description of each of the 30 flows of the rule tree follows.

A.1. 1. Without a registered domain?

This flow checks if the page in question is devoid of a registered domain by the owner. A major difficulty in analyzing this issue occurs in cases where the malicious host frauds through a hosting service, such as *sites.google.com* or *000webhostapp*, since the WHOIS protocol result will indicate that the page has a domain registration, even though the fraud owner did not do it. For these cases, an exception list has been created. In addition to the list, the domain registration time was longer than 3 years to minimize the mentioned problem. If this behavior appears, +3 will be added to the page's score. The flow will proceed to item 3 since item 2 analyzes the subdomain, an element for which a registered domain would be a prerequisite. If a domain exists, other features will be evaluated, namely:

A.1.1. 1.1. Absent from the allow list?

As it has a domain, the URL may have been authorized. This process performs a query on the allow list. If confirmed, the page is marked as trusted, and the gradual analysis process is aborted. Otherwise, the flow proceeds normally.

A.1.2. 1.2. Invalid certificate?

Checks whether the domain has a valid digital certificate, that is, issued and not expired. Domains that have their certificate expired will have the same reputation as a domain without a digital certificate. It is possible to obtain this information because every digital certificate has an expiration date. If this behavior appears, +3 will be added to the score.

A.1.3. 1.3. Was the certificate recently validated?

If it is valid, it is checked if the issuance of the certificate was recent. It is possible to obtain this information because every digital certificate has an issue date and an expiration date. It is important to mention that **temporal features** refer to the volatile nature of *phishing*. As it is generally short-lived, uptime becomes an important ally in distinguishing the page's reputation. This strategy is also adopted by many SEO evaluators (like Alexa Page Rank), which is based on uptime to assess the reputation of a particular website. Therefore, any temporal feature considers something recent with a term and up to 20 days or defines something older than 6 months old. This average time was based on the results obtained in (Silva et al., 2019). If this behavior appears, +3 will be added for newly created ones or -3 for old issues, indicating that the page has a reputation.

A.1.4. 1.4. Does it have simulations TLD?

It is a malicious behavior that occurs when a given domain makes a character distribution to simulate a top domain (TLD), usually preceded by a hyphen ("-"), for example, *HTTP: // magazine luiza -com .info*. This type of attack aims to offer trustworthiness through the nomenclature of the page's domain, and in many cases, the URL is displayed truncated in the browser's status bar. In such cases, the user may end up not observing the ".info" excerpt, giving the impression that the page's TLD is ".com". If this behavior appears, +3 will be added to the page's score.

A.1.5. 1.5. Does it have TLD on the suspicious list?

There is a list of free domains on the internet (like .tk), which makes them very targeted for crime. In the table 1 there is a list of the domains considered in this feature. If this behavior appears, +3 will be added to the page's score.

A.1.6. 1.6. Does it have homographic attempt?

In these cases, they are domains with keywords to represent a brand, for example, *http://auxilio-emergencial.tk* to convey the idea that it is of the *Caixa Econômica* bank.

A.1.7. 1.7. Does it have a long name?

Cases of *phishing* with many characters in the domain name composition are not uncommon. Internationally, a domain has a minimum size of 3 and a maximum of 63 characters (including 4 referring to extensions, such as .com, .net, .gov), so the attacker can arbitrarily enter keywords with a freedom. The study considered over 40 characters as a large size. If this behavior appears, +3 will be added to the score

A.1.8. 1.8. Does it have Encoded name?

This feature refers to encoding the domain name in a different charset instead of the conventional one, such as attacks based on *punycode* or URL *encoding*. If this behavior appears, +3 will be added to the page's score

A.1.9. 1.9. Was the register created recently?

This is another time feature that refers to the volatile nature of *phishing*. As it is temporal, this feature follows the rule described in item 1.3, but it is based on the page creation time through the WHOIS protocol. It is possible to obtain this information because WHOIS informs the entry creation date, update date, and expiration date. If this behavior appears, +3 will be added for newly created or -3 for old registers, indicating that the page has a reputation.

A.2. 2. Does it have a subdomains?

The subdomain is another element of the URL where the attacker has a certain arbitrariness in its composition, and similar exploitation cases carried out in the domain are not uncommon. If the URL has no subdomains, the flow goes to item 3 without changes to the page's score; otherwise, some details will be analyzed, namely:

A.2.1. 2.1. Does it have simulation TLD?

Like item 1.4, but the attacker does not need to use a separator; it is even possible to combine a composition in the domain, making the attack more elaborate, like the URL *http://paypal. com .secure-transaction.tk*. If this behavior appears, +3 will be added to the page's score.

Table 1
Suspicious list for the hosts and TLDs

HOST	
000webhostapp.com	webcindario.com
sharepoint.com	wixsite.com
myfreesites.net	blogspot.com
beget.tech	qponn.net
hol.es	drive.google.com
vineyard-garden.com	godaddysites.com
oreo-e-assistance.com	1drv.com
umbler.com	igotrip.info
sym-global.com	stcroixlofts.com
day-giftcard.com	wefbee.com
greendatainfo.com	kylelierman.com
dermrefresh.com	bayandtools.com
munozbr.com	avaksystems.com

TLD			
.tk	.ml	.ga	.cf
.gq	.nom.za	.tt	.2ya.com
.vze.com	1sta.com	24ex.com	.xyz

A.2.2. 2.2. Does it have homographic attempt?

Like item 1.6, however, with subdomains, the attacker has greater freedom in exploring keywords due to the greater capacity of characters than the domain's composition. For example, *http://bradesco.net.empresas . security - transactions.tk*. If this behavior appears, +3 will be added to the page's score. It is also a feature that can generate *bypass* in some allow list policy that authorizes URLs with known domain terms, such as "facebook.com" or "paypal.com".

A.2.3. 2.3. Is the syntax too long?

Internationally, a domain can have up to 127 subdomains, and each subdomain can have up to 63 characters. However, the study defines that its full syntax, i.e., the joining of the existing subdomains including the periods ("."), Above 14 characters, is considered a long syntax. If this behavior appears, +3 will be added to the page's score.

A.2.4. 2.4. How many subdomains?

A domain can hold up to 127 subdomains; however, the study considered up to 1 subdomain a low quantity, up to 2 an average quantity, and 3 onwards a high quantity. If this behavior appears, +3 will be added to the score.

A.2.5. 3. Is the HOST on the suspicious list?

As well as item 1.5, the study also developed a host dictionary that is highly targeted for crime, like *000webhostapp*. They are attractive hosting services, as they have a very low or even free monthly fee. In the appendix, In the table 1 there is a list of HOSTs considered in this feature. If it appears in the list, +3 will be added to the page's score, and the flow will go to item 4; otherwise, the flow will go to item 4 without changes in the score.

A.2.6. 4. Is the IP exposed?

There are cases in which, in addition to not having a registered domain, the page is still accessed using the public IP of the respective server hosting the page, as this type of information does not offer any identity. The solution considers this suspicious feature, assigning +1 to the page's score in existing situations.

A.2.7. 5. Homographic in path *or* querystring*?*

These are cases in which the exploration is carried out in the *path* or *querystring* of the URL. These values are arbitrarily attributed by the malicious ones, serving as elements that can guarantee trustworthiness or even propagation. Trustworthiness is the result of homographic attacks that use keywords that can bring some confidence. As for the propagation aspect, the malicious can generate different URLs for the same page. This is for cases where block list mechanisms generate the URL *hash* considering all the character content, which would result in a hash different for each URL with different values in its *path* or *querystring*, enabling a *bypass* in this type of solution.

A.2.8. 6. Has a redirect?

These are cases where the application, when loading a URL, expects that in some valued sector of the same, such as a *path* or *querystring*, another URL appears as a value, making a redirect. Many attackers use a given domain's prestige, with

low SEO, to make this type of onslaught. If the behavior is listed, +1 will be added to the page's score, and the flow will proceed to item 6.

A.2.9. 7. Has a recent uptime?

This is another time feature that refers to the volatile nature of *phishing*. As it is temporal, this feature follows the same rule described in item 1.3; however, it is based on the HOST activity time, information obtained through the WHOIS protocol. It is possible to obtain this information because WHOIS informs the date of creation (start of activity) and the last update (last uptime). If this behavior appears, +3 will be added for newly active or -3 for old activities, indicating that the page has a certain reputation.

A.2.10. 8. Has a specific port?

These are cases of pages that do not run on standard port 80 or 443. This feature raises suspicions because the page service is often made available using ports adopted as a standard by certain tools, such as tomcat port 8080. If this behavior exists, the flow will assign +1 to the page's score.

A.2.11. 9. Has reference to form?

This feature analyzes the URL suffix, which often has a reference to the file that is loaded when accessing it, such as suffixes such as index.html. Often, the attack uses a form to submit sensitive data, and eventually, the page containing that form has expressions with suggestive names, such as cart, form, login, or auth. If this behavior exists, the flow will assign +1 to the score.

A.3. 10. It the source code accessible?

There may be little or no content in the page's source code when it loads. In many cases, the explanation is that the content is dynamically displayed to the end-user. These are rendered using functions encapsulated in external files, such as *javascript* functions that design the page layout and appear in files with extension ".js". Another behavior is pages with their aesthetics created through images as a background through HTML tags such as <*map*>. If the source code is significant, some features will be analyzed, namely:

A.3.1. 10.1. Does it have an x-origin element?

A feature present in *phishing* attacks is a large number of links that offer navigability to the end-user. A particularity that deserves suspicion is cross-domain referrals, referring from one HOST to another distinct one too much. The study considers more than 25% of the links on the page to be too many. This item also suggests looking at elements such as <*form*>, <*iframe*> and *clickjacking*, which are click thefts through elements that appear in overlap with others. If the behavior exists, the flow will assign +3.

A.3.2. 10.2. Does it have forged behavior?

A common onslaught in targeted *phishing* is to simulate HTTP errors or display an interface to the user if based on the device's resolution. It is not uncommon for cases of pages that normally open on mobile devices (via shared links via SMS, the feature of *SMiShing*), but that if it is opened in a desktop browser, a forged error is displayed to the user. Therefore, manipulations *javascript* through *navigator.userAgent* aiming to detect the resolution of the device based on the header *User-Agent* are invested that characterize this feature. Once present, the flow will assign +3.

A.3.3. 10.3. Has a form submission?

This feature analyzes whether the page in question submits form data; if so, if it does not have a valid certificate, the solution raises suspicions about the page, assigning +1.

A.4. 11. Without a favicon?

In the same way, described in item 7, some malicious people are not so concerned with details, leaving their application with the standard favicon of the tool used in the page's development. This detail could also strengthen the visual identity of the brand. Therefore, if the behavior exists, the flow will assign +1 to the page's score.

A.4.1. 12. Has seasonal content?

This feature refers to the phenomena that can occur in a certain period of the calendar, such as Christmas and *blackfriday*, which significantly move *e-commerce*. In addition to a defined calendar, these seasonal events also have keywords that facilitate their identification. However, as described in item 9.3, this feature is not only common in fraud, so the flow assigns +1 to the total of the page.

A.5. 13. Is it possible to get the brand?

Theoretically, many of the previous flows make it possible to obtain information about the brand involved, such as item 1 and its children, item 2 and its children, item 9 and its children, and also items 10 and 11. Therefore, the last step of the analysis will be to observe particularities regarding the attack's target brand. If the brand is possible to be recognized, the flow goes to item 12.1; otherwise, the process is closed with a verdict.

A.5.1. 13.1. Does it have an identity conflict?

After obtaining the brand information through the previous items, it is observed if there is any identity conflict, for example, a domain not corresponding to the company name in the certificate. Besides, if the host in question is on the suspicious list, it is observed if the title, certificate, or registered domain references any brand. In case of such divergence, +3 will be assigned.

References

Silva, C.M.R., Feitosa, E.L., Garcia, V.C., 2019. Heuristic-based strategy for phishing prediction: A survey of urlbased approach. Computers & Security .