ارائه رویکردی به منظور شناسایی و پیش بینی وب‌سایت‌های فیشینگ به وسیله الگوریتمهای کلاس‌بندی براساس مشخصه‌های صفحات وب

مهدی دادخواه[1*]، محمد داورپناه جزی[1]، مجید سعیدی مبارکه[2]

اطلاعات مقاله

دریافت مقاله: ۱۳۹۳/۰۶/۱۴

پذیرش مقاله: ۱۳۹۴/۰۲/۱۵

واژگان کلیدی:

فیشینگ،

دزدی الکترونیکی،

امنیت تجارت الکترونیک،

ژورنال فیشینگ،

مدل‌سازی حملات،

مشخصه صفحات وب.

چکیده

امروزه مهمترین ریسک و چالش مورد توجه در تجارت و بانکدارای الکترونیک، خطر کلاهبرداری آنلاین و حملات فیشینگ است. حملات فیشینگ همواره به عنوان یکی از ابزارهای پرکاربرد برای مهاجمان، به منظور سرقت کلمه‌های عبور و رمزهای الکترونیک کاربران در مبادلات الکترونیک بوده است. در این نوع کلاهبرداری، مهاجمان نامه‌های الکترونیک با ادعاهای مختلف به قربانی ارسال می‌کند و با تکنیکهای مختلفی قربانی را به صفحه‌های جعلی خود هدایت می‌کند سپس اقدام به سرقت اطلاعات حساس کاربران مانند رمزهای عبور می‌نماید. صفحات وب، نامه‌های الکترونیک و آدرسهای فیشینگ دارای ویژگی‌هایی هستند که از آنها می‌توان برای شناسایی این حملات استفاده کرد. در این مقاله رویکردی جهت شناسایی و پیش‌بینی وب‌سایت‌های فیشینگ با استفاده از الگوریتمهای کلاس‌بندی بر اساس مشخصه‌های صفحات وب ارائه خواهد شد که نرخ خطای کمتری نسبت به سایر تکنیکهای مقابله با حملات فیشینگ، به خصوص تکنیکهای مشابه میتنی بر الگوریتمهای داده کاوی دارد. در رویکرد ارائه شده، ویژگی‌های قابل استفاده در شناسایی صفحات فیشینگ براساس میزان تاثیر در شناسایی این حملات وزن بندی شده سپس با اعمال الگوریتمهای کلاس بندی بر روی مجموعه داده‌های مرتبط، الگویی به منظور شناسایی این حملات استخراج می‌گردد که قادر به شناسایی حملات ژورنال فیشینگ بوده و نرخ خطای کمتری را نسبت به سایر روشهای مشابه پیشین نیز دارا می‌باشد.

۱- مقدمه

حمله فیشینگ به تلاش برای دستیابی به اطلاعات حساس افراد مانند نام کاربری، کلمه عبور، و اطلاعات کارتهای اعتباری، به‌وسیله تکنیکهای مهندسی اجتماعی گفته می‌شود [۱]. این حملات برای اولین بار با جزئیات در سال ۱۹۸۷ توضیح داده شدند و این واژه برای اولین بار در سال ۱۹۹۶ مورد استفاده قرار گرفت [۲]. واژهٔ فیشینگ مخفف عبارت Password Harvesting Fishing (شکار کردن گذرواژه کاربر از طریق یک طعمه) است که در آن حرف "Ph" به جای F برای القای مفهوم فریفتن جایگزین شده است [۳]. اولین قربانی حملات فیشینگ نیز یک فراهم کننده سرویس اینترنت به نام AOL بود. این حمله در سایت کمپانی AOL به تعداد زیاد اتفاق افتاد به صورتی که مسئولین این کمپانی مجبور شدند اطلاعیه‌های متعددی را به منظور آگاه ساختن کاربران خود منتشر نمایند. مهاجمان به منظور افزایش

* پست الکترونیک نویسنده مسئول: mdt@dr.com

۱. گروه مهندسی کامپیوتر و فناوری اطلاعات، موسسه آموزش عالی صنعتی فولاد، فولادشهر، اصفهان

۲. گروه کامپیوتر، واحد مبارکه، دانشگاه آزاد اسلامی، مبارکه، ایران

ضریب موفقیت حملات سعی می‌نمایند خود را به گونه‌ای عرضه نمایند که قربانیان به آنها اعتماد نموده و به عنوان نمایندگان قانونی مراکز معتبری نظیر بانک‌ها آنها را قبول نمایند. در این نوع حمله، فیشرها (مهاجمان یاکسانی که حمله فیشینگ را انجام می‌دهند) با طراحی یک سایت که شبیه به سایت مورد نظر است، کار خود را آغاز می‌کنند. پس از انجام این مرحله آنها باید روشی را پیدا کنند که قربانیان خود را مجبور کنند تا در سایت آنها وارد شده و اطلاعات محرمانه خود را وارد کنند. به عبارت دیگر، هدف اصلی یک حملهٔ فیشینگ انجام یک ارتباط جعلی است که معمولا با یک ایمیل که حاوی یک URL جعلی از سایت بانک یا نهاد دولتی است آغاز می‌گردد. فرد مهاجم یا طراح حمله فیشینگ سعی می‌کند از مواردی استفاده کند که برای قربانیان جالب بوده و بتواند توجه آنها را جلب کند. سپس سعی می‌کند نام، آدرس، شماره تلفن یا هر اطلاعاتی که بتواند بعدا برای پیشبرد اهدافش از آنها استفاده کند را بدست آورد.

۲- مطالعات صورت‌گرفته در حوزهٔ تکنیک‌های مقابله با حملات فیشینگ

مطالعات و تلاش‌های زیادی جهت ارائه راهکارهای مختلف برای مقابله با حملات فیشینگ مطرح شده است. از جملهٔ آنها می‌توان مقابله با حملات فیشینگ به وسیله نشانهٔ ورود [۴]، توسعه سیستم خبره مبتنی بر مشخصه-های صفحات وب جهت تشخیص وب‌سایت‌های فیشینگ [۵]، تشخیص ابرلینکهای فیشینگ با استفاده از الگوریتم ژنتیک [۶]، شناسایی حملات فیشینگ با دسته‌بندی ابرلینکها [۷]، شناسایی صفحات فیشینگ براساس ویژگی‌های پایه [۸]، تشخیص صفحات فیشینگ براساس محتوا [۹]، مقابله با حملات فیشینگ به وسیلهٔ احراز هویت دو مرحله‌ای [۱۰]، شناسایی صفحات فیشینگ براساس میزان شباهت محتوای صفحات وب با دسته‌بندی دامنه [۱۱]، مقابله با وب‌سایت‌های فیشینگ براساس روابط انجمنی [۱۲]، شناسایی صفحات فیشینگ با مقایسهٔ میزان تفاوت رشتهٔ آدرس با لیست سفید [۱۳]، شناسایی دامنه‌های فیشینگ بر اساس الگوریتم رتبه-بندی [۱۴] و استفاده از الگوریتمهای داده‌کاوی [۱۵] جهت پیش‌بینی وب‌سایت‌های فیشینگ را نام برد. در میان تکنیکهای نام برده شده تکنیکهای مشابه رویکرد ما نیز وجود دارند اما دارای نقطه ضعفهایی هستند که در ادامه توضیح داده می‌شود. در تکنیک مقابله با حملات فیشینگ به وسیله توسعهٔ سیستم خبره [۵]، لیستی حاوی ۲۷ مشخصه مورد توجه در صفحات فیشینگ معرفی شده سپس آنها را در شش دسته قرار می‌دهد و بر اساس آنها سیستم خبره را با مجموعه‌ای از قوانین آموزش می‌دهد اما در این تکنیک مجموعه وسیعی از قوانین تولید می‌شود. همچنین این مقاله مشخص نمی‌کند که مشخصه‌ها به خصوص ویژگی‌هایی که مرتبط با عاملهای انسانی هستند چگونه از وب‌سایت‌ها استخراج شده‌اند. در [۱۶، ۱۷، و ۱۸] از تکنیکهای داده‌کاوی جهت کشف الگوهای مربوط به صفحات فیشینگ استفاده شده و ۲۷ مشخصهٔ مربوط به صفحات فیشینگ را به منظور آموزش دادن طبقه‌بندی کننده مورد توجه قرار می‌دهد. اما تکنیکهای مبتنی بر داده‌کاوی با این شیوه معمولا نرخ خطای بالایی را شامل می‌شوند و چون به وسیله مجموعه داده‌های کم آموزش داده شده‌اند با گسترش حجم داده‌ها نرخ خطای آنها افزایش می‌یابد. در برخی از روشهای مبتنی بر داده کاوی لیستی حاوی چند مشخصه مطرح از ۲۷ مشخصه مورد توجه انتخاب شده و الگوریتمهای داده کاوی روی آنها اعمال می‌شود [۴، ۶، ۹ و ۱۱] که این رویکرد باعث تشخیص اشتباه در حملات فیشینگ متفاوت با مجموعه آزمایشی انتخاب شده می‌گردد.

۳- مشخصه‌های مورد استفاده در شناسایی صفحات فیشینگ

به طور کلی ۲۷ مشخصه برای تشخیص حملات فیشینگ مطرح شده است [۱۹] که در رویکرد ارائه شده در این مقاله، سه مشخصه رتبه‌بندی، رنکینگ در موتور جستجو

و وجود عنوان در لیست JCR به منظور شناسایی ژورنال-فیشینگها و افزایش نرخ دقت به این مشخصه‌ها اضافه شده‌اند. این مشخصه‌ها در قالب ۳۰ مشخصه و شش بخش قابل طبقه بندی هستند. این شش بخش مختلف شامل مشخصه‌های دامنه وب سایت، امنیت و رمزنگاری مورد استفاده، کدهای اسکریپتی موجود در صفحه، ظاهر صفحه و محتوای آن، آدرس صفحه وب و ویژگی‌های رفتاری سایت می‌شوند. لیست این مشخصه‌ها که به عنوان پارامتر برای طبقه بندی کننده تعریف می‌گردند در جدول ۱ نشان داده شده است. در ادامه مشخصه‌های جدید به کار رفته در این مقاله جهت شناسایی حملات فیشینگ به طور خلاصه توضیح داده شده‌اند سایر مشخصه‌ها نیز در تحقیقات پیشین به تفصیل معرفی شده‌اند [۵ و ۱۹]. اگر هر یک از این مشخصه‌ها به تنهایی استفاده شوند نمی-توانند معیار خوبی برای تشخیص صفحات فیشینگ باشند بلکه به کارگیری این مشخصه‌ها با یکدیگر به تشخیص این صفحات کمک می‌کنند.

وجود عنوان صفحه در لیست JCR: این پارامتر به منظور تشخیص ژورنال-فیشینگها به مجموعه ویژگیها اضافه شده است و زمانی که عنوان یک سایت در لیست اعلام شده از طرف سایت تامسون رویترز باشد مقدار منطقی ۱ را دریافت می‌کند.

ترتیب در نتایج موتور جستجو: این ویژگی در رویکرد مطرح شده به منظور افزایش دقت در شناسایی صفحات فیشینگ افزوده شده است به این صورت که عنوان وب-سایت مورد نظر در موتور جستجو مورد بررسی قرار گرفته و نتیجه بر گردانده می‌شود. در بسیاری از موارد مهاجمان با نفوذ به سایتهایی که دارای رتبه بالا در موتورهای جستجو می‌باشند صفحه فیشینگ خود را در آنها قرار می‌دهند بنابراین بالا بودن رنکینگ دامنه به تنهایی نمی-تواند دلیلی برای قانونی بودن وب‌سایت مورد بررسی باشد.

رتبه دامنه سایت در موتور جستجو: چون وب‌سایت‌های فیشینگ کپی سایت اصلی هستند و معمولا مدت کوتاهی پیش از حمله راه اندازی می‌شوند رتبه بالایی در موتورهای جستجو به خود اختصاص نمی‌دهند ضمن آنکه با وجود سایت اصلی معمولا موتورهای جستجو رتبه‌ای به سایتهای جعلی اختصاص نخواهند داد. در تحقیق انجام شده با بررسیهایی که روی ۱۰ سایت مربوط به ژورنالهای جعلی انجام شد این نتیجه حاصل گشت که ژورنال-فیشینگ‌ها دارای رتبه صفر در موتورهای جستجو هستند. موتورهای جستجوی گوناگونی وجود دارند ولی موتور جستجوگر گوگل به دلیل دارا بودن قابلیت Page Rank در رویکرد ارائه شده مورد استفاده قرار می‌گیرد. جدول ۲ چند ژورنال-فیشینگ به همراه رتبه آنها را در موتور جستجوگر گوگل نشان می‌دهد.

جدول ۱- لیست مشخصه‌های مورد استفاده در طبقه‌بندی کننده برای تشخیص وب سایتهای فیشینگ

مشخصه	شماره	معیار
استفاده از آدرس IP در آدرس وب‌سایت	۱	مشخصه دامنه و آدرس وب‌سایت‌ها
آدرس غیر عادی	۲	
میزان وجود اختلال در آدرس دهی پایه	۳	
وجود اطلاعات غیر عادی در سرور	۴	
آدرس درخواستی	۵	
رتبه‌دامنه سایت در موتور جستجو	۶	
استفاده از گواهینامه امنیت سایت	۷	امنیت و رمزنگاری
اعتبار داشتن گواهینامه	۸	
وجود فایلهای کوکی غیرعادی	۹	
وجود اطلاعات شناسایی گواهی امنیت	۱۰	
وجود صفحات انتقال	۱۱	کدهای اسکریپتی موجود در صفحه
وجود حمله استرادلینگ	۱۲	
وجود حمله فارمینگ	۱۳	
مخفی نمودن آدرس واقعی صفحه	۱۴	
بررسی فرمهای کنترلی در سرور	۱۵	
اشکالات املایی در محتوا	۱۶	ظاهر صفحه و محتوای آن
سایت کپی شده از سایت اصلی	۱۷	
استفاده از فرم با دکمه ارسال	۱۸	
استفاده از پنجره پاپ آپ	۱۹	
غیرفعال شدن کلیک راست موس	۲۰	
وجود عنوان صفحه در لیست JCR	۲۱	
وجود آدرس وب طولانی	۲۲	آدرس صفحه وب
جایگذاری کاراکترهای مشابه در آدرس	۲۳	
افزودن پسوند و پیشوند در آدرس	۲۴	
استفاده از کاراکتر @ در آدرس	۲۵	
استفاده از کاراکترهای هگزا دسیمال	۲۶	
ترتیب در نتایج موتور جستجو	۲۷	
میزان تاکید بر امنیت	۲۸	ویژگی‌های رفتاری سایت
نوع خوشامدگویی عمومی مورد استفاده	۲۹	
صرف زمان بیش از حد در دریافت اطلاعات	۳۰	

جدول ۲-چند نمونه ژورنال فیشینگ به همراه رتبه آنها در موتور جستجوگر گوگل

نام ژورنال	وب سایت اصلی	وب‌سایت جعلی
Afinidad (ISSN: 0001-9704)	http://www.aiqs.es PR:4	http://www.afinidad.org PR:0
Archives des sciences (ISSN: 1661-464X)	http://www.unige.ch/sphn PR:5	http://www.archiveofscience.com PR:0
Bradleya (ISSN: 0265-086X)	http://www.bcss.org.uk/brad.php PR:4	http://www.britishedu.org.uk PR:0
Bothalia (ISSN: 0006-8241)	http://www.abcjournal.org PR:0	http://www.bothalia.com PR:0
Ciencia e tecnica vitivinicola (ISSN: 0254-0223)	http://www.scielo.oces.mctes.pt PR:7	http://www.ciencia-e-tecnica.org PR:0
Nautilus	http://www.shellmuseum.org/nautilus/index.html PR:1	http://www.nautilusjournal.net PR:0
Texas Journal of Science (ISSN: 0040-4403)	http://www.texasacademyofscience.org PR:5	http://www.texassciences.us PR:0 http://www.texassciences.com PR:0 http://www.texasacademyofscience.com PR:0

۴- رویکرد ارائه شده جهت مقابله با حملات فیشینگ

در رویکرد ارائه شده جهت مقابله با حملات فیشینگ ضمن مورد توجه قرار گرفتن ۲۷ مشخصه شناخته شده برای این حملات، سه مشخصه جدید نیز جهت شناسایی ژورنال-فیشینگها و افزایش نرخ دقت بیان شده است. سپس این مشخصه‌ها بر اساس نوع در شش دسته متفاوت قرار گرفته و با توجه به فراوانی آنها در حملات فیشینگ میزان وزن مرتبط با هر دسته طبق مجموعه آموزشی محاسبه می‌گردد. سپس با تشکیل مجموعه آموزشی دیگری شامل وب‌سایت‌های قانونی و فیشینگ، دسته‌های مرتبط باهم شناسایی شده و الگوریتمهای کلاس‌بندی بر روی داده‌های مرتبط با آنها اعمال می‌گردد و بهترین الگوریتم با توجه به نرخ خطا برای هر دسته انتخاب می‌شود و در آخر بر اساس وزن هر دسته جدول تصمیم نهایی شکل می‌گیرد. در تکنیکهایی که تاکنون معرفی شده‌اند فقط از ویژگی‌های مربوط به یکی از دسته‌ها جهت شناسایی حملات فیشینگ استفاده می‌شود و یا همه ویژگیها با هم در نظر گرفته شده و الگوریتمهای داده‌کاوی بر روی آنها اعمال می‌شود ولی در رویکرد مطرح شده همه ویژگیها به نسبت وزنی که دارند در شناسایی وب‌سایت‌های فیشینگ موثر هستند و این خود باعث کاهش نرخ خطا می‌شود. ابتدا در بخش ۴-۱ مجموعه داده مورد استفاده و نحوه آماده سازی آن تشریح شده است سپس در بخش ۴-۲ وزن هر دسته محاسبه گردیده و نهایتا در بخش ۴-۳ الگوریتم مناسب هر دسته و جدول تصمیم نهایی ارائه شده است.

۴-۱ داده‌های مورد استفاده و آماده سازی آنها

سایتهای بسیاری وجود دارند که اطلاعاتی راجع به وب‌سایت‌های فیشینگ در اختیار قرار می‌دهند که از جمله آنها می‌توان به وب‌سایت ابزارهای ضد فیشینگ مانند Comod Web Inspector [۲۰] را نام برد. اما اطلاعاتی که این قبیل سایتها در اختیار قرار می‌دهند مربوط به وب‌سایت‌های فیشینگ خنثی شده هستند. بنابراین امکان پژوهش روی آنها وجود ندارد. منابع دیگری مانند Phishtank [۲۱] و Millersmiles [۲۲] وجود دارند که به طور روزانه لیست وب‌سایت‌های فیشینگ فعال را منتشر می‌نمایند که قابلیت استفاده از آنها برای انجام پژوهشهای مورد نظر وجود دارد.

اطلاعاتی که این وب‌سایت‌ها درباره حملات فیشینگ منتشر می‌کنند عمدتا شامل آدرس صفحه فیشینگ، تاریخ شناسایی آن و یک تصویر از وب‌سایت مربوطه است که برای پژوهش ما کافی نیست به همین دلیل باید وب‌سایت‌های فیشینگ منتشر شده توسط این منابع در بازه‌های زمانی خاص مورد بررسی قرار گرفته و اطلاعات مربوط به ۳۰ پارامتر مطرح شده از آنها استخراج گردد تا براساس آنها مجموعه آزمایشی مورد نظر شکل گیرد.

۲-۴ محاسبه وزن هر دسته از مشخصه‌ها

به منظور افزایش دقت در شناسایی وب‌سایت‌های فیشینگ و کاهش نرخ خطای مثبت به جای درنظر گرفتن کلیه ویژگیها به صورت برابر باید هر دسته از مشخصه‌ها بر اساس وزنی که دارند در شناسایی صفحات فیشینگ مدنظر قرار گیرند. برای محاسبه این وزنها می‌توان از وب‌سایت‌های فیشینگ شناسایی شده استفاده نمود. فرآیند محاسبه وزن هر دسته به این صورت است که ابتدا لیستی حاوی ۴۰ وب‌سایت فیشینگ از منابع توضیح داده شده در فوق تهیه گردیده سپس ۳۰ پارامتر موثر در شناسایی صفحات فیشینگ به صورت یک پارامتر بولین با دو مقدار صفر و یک در نظر گرفته می‌شود که یک به معنای موثر بودن پارامتر مورد بررسی در شناسایی صفحه فیشینگ فعلی و صفر به معنای عدم تاثیر آن است. سپس تعداد یکهای مربوط به صفتهای هر دسته اندازه گیری شده و نسبت به تعداد کل یکها در مجموعهٔ آموزشی سنجیده می‌شود این به آن معناست که چند درصد از صفات موجود در هر دسته در شناسایی صفحات فیشینگ موثر هستند و به این ترتیب وزن هر دسته محاسبه می‌گردد. جدول ۳ وزن‌های محاسبه شده برای هر دسته را نشان می‌دهد.

جدول ۳-وزنهای مربوط به مشخص‌ها در هر دسته.

نام دسته	تعداد یکها (از ۲۵۲)	وزن
مشخصه دامنه و آدرس وب‌سایت‌ها	۸۳	۳۳٪
امنیت و رمزنگاری	۳۸	۱۵٪
کدهای اسکریپتی موجود در صفحه	۸	۴,۵٪
ظاهر صفحه و محتوای آن	۵۵	۲۳٪
آدرس صفحه وب	۵۷	۲۴٪
ویژگی‌های رفتاری	۱	۰,۵٪

صفات متعلق به دستهٔ ویژگی‌های رفتاری به دلیل داشتن تاثیر کم در شناسایی صفحات فیشینگ در رویکرد ما مورد استفاده قرار نخواهند گرفت. همچنین در مجموعه آموزشی مورد استفاده قرار گرفته مقادیر برخی از صفات صفر است ولی این صفات به دلیل اینکه جز ویژگی‌های شناخته شده در بسیاری از حملات فیشینگ هستند در رویکرد ما نیز مورد توجه قرار می‌گیرند. ویژگی‌های مربوط به حملات فارمینگ و استرادلینگ چون در وب‌سایت‌ها صورت نمی‌گیرد و معمولا در سیستم عامل کاربر یا شبکه محلی رخ می‌دهد در پیش‌بینی وب‌سایت‌های فیشینگ موثر نخواهد بود. به علاوه ویژگی غلطهای نگارشی در محتوا نیز نمی‌تواند شاخص مناسبی جهت شناسایی وب‌سایت‌های فیشینگ باشد زیرا امروزه صفحات فیشینگ با کپی کردن از سایت اصلی ایجاد می‌گردند بنابراین اگر سایت اصلی فاقد غلط املایی باشد سایت جعلی هم دارای غلطهای نگارشی و املایی نخواهد بود ضمن آنکه در مجموعه آموزشی مورد بررسی چنین ویژگی در سایتهای فیشینگ مشاهده نگردید. همچنین وجود کوکی‌های غیر عادی طبق بررسی‌های انجام شده روی ۱۰۰ وب‌سایت فیشینگ نمی‌تواند مشخصه مناسبی برای تشخیص صفحات فیشینگ باشد. جدول ۴ ویژگی‌های مورد استفاده در رویکردی که ارائه خواهد شد را نشان می‌دهد.

جدول ۴-ویژگی‌های انتخاب شده جهت پیش بینی وب سایت‌های فیشینگ.

مشخصه	شماره	معیار
استفاده از آدرس IP در آدرس وب‌سایت	۱	مشخصه دامنه و آدرس وب‌سایت‌ها
آدرس غیر عادی	۲	
میزان وجود اختلال در آدرس دهی پایه	۳	
وجود اطلاعات غیر عادی در سرور	۴	
آدرس درخواستی	۵	
رتبه‌دامنه سایت در موتور جستجو	۶	
وجود عنوان صفحه در لیست JCR	۷	
استفاده از گواهینامه امنیت سایت	۸	امنیت و رمزنگاری
وجود اطلاعات شناسایی گواهی امنیت	۹	
وجود اطلاعات شناسایی گواهی امنیت	۱۰	
وجود صفحات انتقال	۱۱	کدهای اسکریپتی موجود در صفحه
مخفی نمودن آدرس واقعی صفحه	۱۲	
بررسی فرمهای کنترلی در سرور	۱۳	
سایت کپی شده از سایت اصلی	۱۴	ظاهر صفحه و محتوای آن
استفاده از فرم با دکمه ارسال	۱۵	
استفاده از پنجره پاپ آپ	۱۶	
غیرفعال شدن کلیک راست موس	۱۷	
وجود آدرس وب طولانی	۱۸	آدرس صفحه وب
جایگذاری کاراکترهای مشابه در آدرس	۱۹	
افزودن پسوند و پیشوند در آدرس	۲۰	
استفاده از کاراکتر @ در آدرس	۲۱	
استفاده از کاراکترهای هگزا دسیمال	۲۲	
ترتیب در نتایج موتور جستجو	۲۳	

۳-۴ انتخاب الگوریتم مناسب برای هر دسته

به منظور انتخاب الگوریتم مناسب برای هر دسته از صفات ابتدا باید مجموعه آموزشی از منابعی که در بخشهای قبلی توضیح داده شد تهیه گردد و مقدار ۲۳ پارامتر انتخاب شده برای هر وب‌سایت فیشینگ اندازه گیری شود، سپس بر روی مجموعه آموزشی تهیه شده الگوریتمهای کلاس‌بندی اعمال شده و با توجه به نرخ خطا مناسب‌ترین الگوریتم برای هر دسته انتخاب شود. مجموعه آموزشی به کار رفته علاوه بر وب‌سایت‌های فیشینگ باید شامل وب‌سایت‌های قانونی و اصلی نیز باشد تا رویکردی که ارائه خواهد شد توانایی تشخیص وب‌سایت‌های اصلی را نیز داشته باشد. صفات مربوط به دستهٔ مشخصه دامنه و آدرس وب‌سایت‌ها با سه مقدار شامل کم، متوسط و زیاد اندازه‌گیری می‌شوند (به استثنای استفاده از IP در آدرس و صفت موجود بودن عنوان در لیست JCR) به صورتی که اگر تعداد لینکهای مربوط به دامنه خارجی در سایت مورد بررسی صفر بود صفت اختلال در آدرس‌دهی پایه مقدار کم، اگر تعداد لینکهای خارجی بین یک تا دو لینک بود مقدار متوسط و اگر بیشتر از دو لینک بود مقدار زیاد را به خود تخصیص می دهد. مقدار این مرزبندیها با توجه به مطالعات صورت گرفته روی سایتهای قانونی و سایتهای متعلق به حملات فیشینگ همچنین مطالعات پیشین صورت گرفته در [۵ و ۱۹] تعیین شده است. برای صفت رتبه صفحه در موتور جستجو برای رتبه صفر مقدار کم، رتبه بین یک تا سه مقدار متوسط و رتبه بیشتر از سه مقدار زیاد به آن احتصاص داده می‌شود. مقادیر بقیه

صفات این دسته نیز به صورت بصری قابل تعیین هستند. صفات مربوط به دسته‌های امنیت و رمزنگاری، کدهای اسکریپتی موجود در صفحه، ظاهر صفحه و محتوای آن و آدرس صفحه وب نیز (به غیر از ترتیب در نتایج موتور جستجو) به صورت یک پارامتر بولین در نظر گرفته شده و با دو مقدار صفر و یک اندازه‌گیری می‌گردند که یک به معنای وجود صفت مورد بررسی و صفر به معنای عدم وجود آن است. همچنین وب‌سایت‌های فیشینگی که از صفحات انتقال استفاده می‌نمایند صفات مربوط به مشخصه دامنه و آدرس وب‌سایت‌ها باید در آدرس سایت قبل از انتقال مورد بررسی قرار گیرند ولی صفات سایر دسته‌ها پس از انتقال به صفحه دوم اندازه‌گیری می‌شوند. این روش اندازه‌گیری مقادیر صفات، میزان تشخیص وب‌سایت‌های فیشینگ را افزایش می‌دهد زیرا مهاجمان معمولا از صفحات انتقالی استفاده می‌کنند که تعداد کمی از مشخصه‌های مربوط به صفحات فیشینگ را دارا می‌باشند و بسیاری از ابزارهای فیشینگ فقط توانایی بررسی صفحه اول را داشته و صفحه دوم را نمی‌توانند تشخیص دهند بنابراین قادر به شناسایی وب‌سایت‌های فیشینگ دارای صفحه انتقال نخواهند بود.

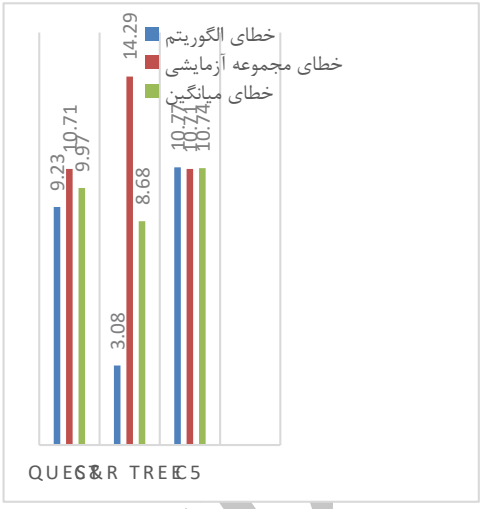
پس از آماده سازی مجموعه آموزشی و بررسی آن مشاهده می‌شود که دسته صفات امنیت و رمزنگاری و مشخصه دامنه وب‌سایت‌ها با یکدیگر مرتبط هستند همچنین به منظور تعیین اعتبار گواهی امنیت باید از اطلاعات دامنه استفاده گردد و بهتر است الگوریتمها روی مجموعه آنها اعمال شود. به علاوه مجموعه صفات مربوط به کدهای اسکریپتی موجود در صفحه و ظاهر صفحه و محتوای آن مرتبط با محتوای داخلی صفحه بوده و در نظر گرفتن انها با یکدیگر به تشخیص حملات فیشینگ کمک خواهد نمود ضمن آنکه باتوجه به وزنهای محاسبه شده برای دسته‌ها در بخش قبل مشاهده می‌گردد که وزن مجموعه صفات بسیار کم است و در نظر گرفتن آن با سایر دسته‌ها به شناسایی بهتر وکاهش نرخ خطا کمک خواهد کرد. بنابراین الگوریتمهای داده‌کاوی باید بر روی سه دسته جدید مشخصه دامنه و رمزنگاری، اسکریپتها و محتوای صفحه و دامنه وب‌سایت مورد بررسی اعمال گردند. جدول ۵ دسته بندیهای جدید به همراه وزنشان را نشان می‌دهد. اکثر رویکردهای معرفی شده جهت مقابله با حملات فیشینگ معمولا صفات موجود در یکی از دسته‌هارا مورد توجه قرار می‌دهند.

جدول ۵- دسته بندی جدید صفات جهت اعمال الگوریتمهای داده کاوی

وزن	نام دسته
۴۸٪	مشخصه دامنه و رمزنگاری
۲۷,۵٪	کدهای اسکریپتی و محتوای صفحه
۲۴,۵٪	آدرس صفحه وب

۴-۳-۱ مشخصه دامنه و رمزنگاری

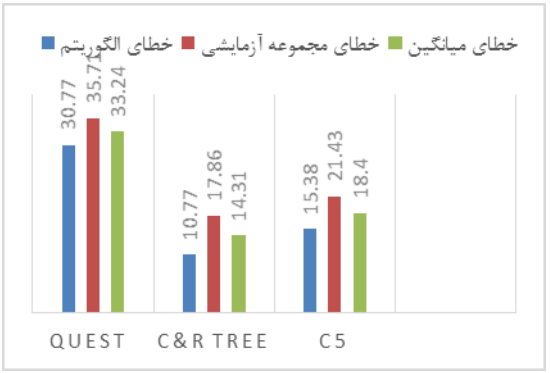
در رویکرد مطرح شده از سه الگوریتم C۵ [۲۳]، C&R Tree [۲۴] و Quest [۲۵] جهت کلاس بندی استفاده می‌شود. شکل ۱ میزان خطای هر الگوریتم را پس از اعمال روی داده‌های مرتبط با مجموعه صفات دستهٔ مشخصه دامنه و رمزنگاری نشان می‌دهد. میزان دقت هر الگوریتم به وسیله مجموعه داده آزمایشی تهیه شده در نرم افزار Clementine [۲۶] بدست آمده است. این نرم‌افزار دو نرخ خطا یکی برای کلاس شکل گرفته و دیگری برای مجموعه دادهٔ آزمایشی اعمال شده روی الگوریتم در اختیار قرار می‌دهد بنابراین مناسب‌ترین الگوریتم برای صفات این دسته C&R Tree می‌باشد.



شکل ۱-میزان نرخ خطای الگوریتمهای به کار رفته در مجموعه صفات مشخصه دامنه و رمزنگاری

۴-۳-۲ کدهای اسکریپتی و محتوای صفحه

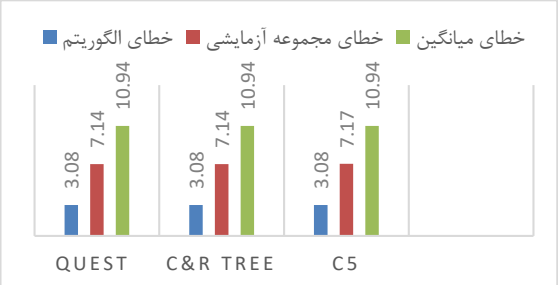
مطابق آنچه در بخش قبلی توضیح داده شد میزان خطای الگوریتمها پس از اعمال روی داده‌های مرتبط با مجموعه صفات دستهٔ کدهای اسکریپتی و محتوای صفحه در شکل ۲ نشان داده شده است. الگوریتم مناسب این دسته نیز C&R Tree می‌باشد.



شکل ۲-میزان نرخ خطای الگوریتمهای به کار رفته در مجموعه صفات کدهای اسکریپتی و محتوای صفحه

۴-۳-۳ آدرس صفحه وب

شکل ۳ نرخ خطای الگوریتمهای اعمال شده روی مجموعه داده‌های مرتبط با آدرس صفحه وب را نشان می‌دهد. همان‌طور که مشاهده می‌گردد این بخش مستقل از نوع الگوریتم کلاس‌بندی می‌توان عملیات پیش‌بینی وب‌سایت‌های فیشینگ را انجام داد. بنابراین الگوریتم مناسب این مرحله را با توجه به مطالعات مقایسه کننده‌ای که بر روی الگوریتمهای کلاس‌بندی انجام شده‌اند انتخاب می‌شود. طبق مطالعات صورت گرفته از بین الگوریتمهای کلاس‌بندی دو الگوریتم C۵ و Quest کمترین نرخ خطا را دارا می‌باشند ولی الگوریتم Quest دارای سرعت بیشتری است [۲۷] بنابراین الگوریتم Quest به عنوان الگوریتم مناسب این مرحله انتخاب می‌شود.



شکل ۳-میزان نرخ خطای الگوریتمهای به کار رفته در مجموعه صفات آدرس صفحه وب

۴-۴ جدول تصمیم نهایی

همان‌طور که در بخشهای قبلی بیان شد مشخصه‌های مرتبط با صفحات فیشینگ در سه دسته کلی جای گرفته و در هر دسته الگوریتم مناسب جهت تشخیص حملات فیشینگ روی مجموعه داده‌ها اعمال می‌شود. با توجه به وزنهای بدست آمده در جدول ۵، صفحه مورد بررسی زمانی فیشینگ تشخیص داده می‌شود که حداقل دو الگوریتم آن را به عنوان صفحه فیشینگ شناسایی کنند. جدول ۶ نحوه تصمیم گیری در مورد جعلی یا اصلی بودن صفحه مورد بررسی را نشان می‌دهد. یک در جدول به معنای فیشینگ تشخیص داده شدن صفحه مورد بررسی در الگوریتم مورد استفاده است.

جدول ۶-جدول تصمیم نهایی جهت شناسایی وب‌سایت‌های فیشینگ

نتیجه نهایی	نتیجه الگوریتم در هر مرحله		
	آدرس صفحه وب	کدهای اسکریپتی و محتوای صفحه	مشخصه دامنه و رمزنگاری
سایت قانونی	۰	۰	۰
سایت قانونی	۱	۰	۰
سایت قانونی	۰	۱	۰
صفحه فیشینگ	۱	۱	۰
وب‌سایت مشکوک	۰	۰	۱
صفحه فیشینگ	۱	۰	۱
صفحه فیشینگ	۰	۱	۱
صفحه فیشینگ	۱	۱	۱

۵- اندازه گیری نرخ خطا

به منظور اندازه گیری نرخ خطا، مجموعهٔ آموزشی حاوی رکوردهای مرتبط با سایتهای قانونی و فیشینگ را مورد استفاده قرار می‌دهیم. جدول ۷ نتایج در هر مرحله را نشان می‌دهد. رکوردها با بررسی کلاسهای شکل گرفته توسط الگوریتمها به گونه‌ای انتخاب شده‌اند که بیشترین نرخ خطا حاصل گردد. در صورتی که از رویکرد مطرح شده جهت مقابله با حملات فیشینگ استفاده گردد،

میزان نرخ خطا برابر ۱۴٪ و اگر مشابه سایر رویکردهایی که الگوریتمهای داده کاوی به صورت یکجا روی مجموعه صفات اعمال می‌شوند [۱۵، ۱۷ و ۱۸] انجام گردد میزان نرخ خطا برابر ۲۱٪ خواهد بود.

جدول ۷- الگوریتمها به همراه نرخ خطا در تشخیص حملات فیشینگ

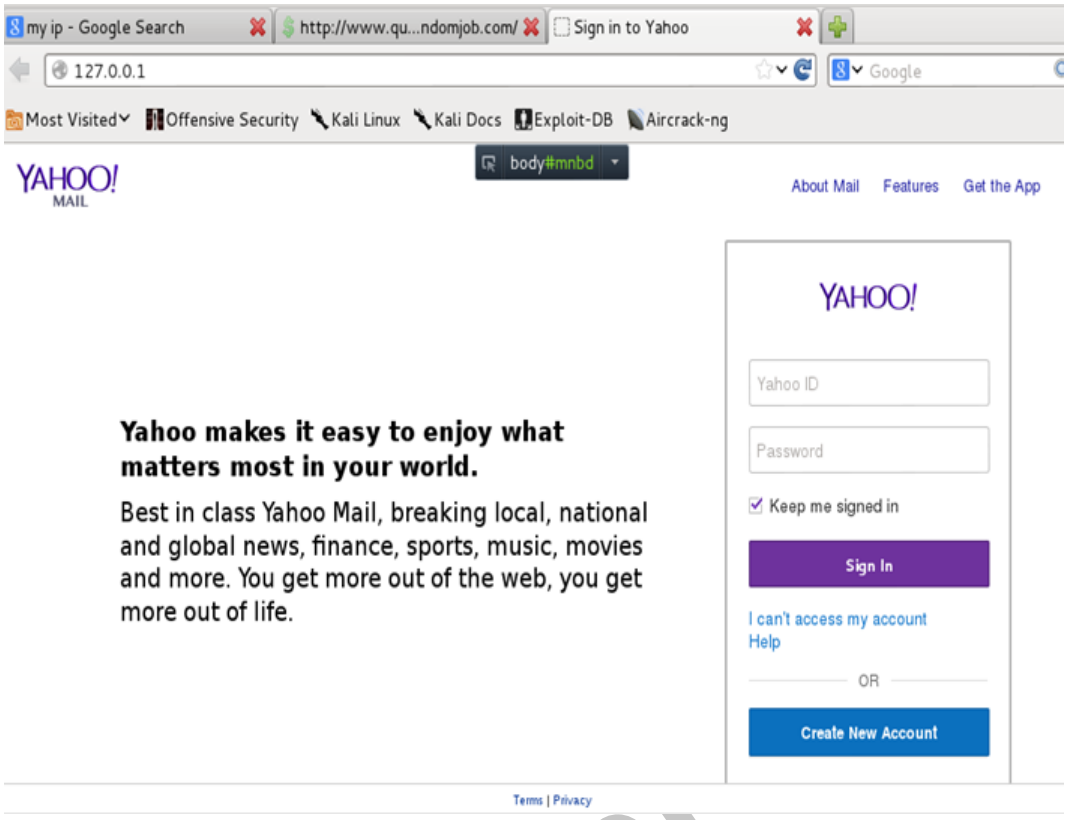
نام الگوریتم	نرخ خطا
رویکرد معرفی شده	۱۴٪
C&R Tree	۲۱،۵٪
Quest	۲۱،۵٪
C5	۲۱،۵٪

۶- آزمایش تجربی

به منظور آزمایش تجربی رویکرد ارائه شده از ابزار مهندسی اجتماعی [۲۸] موجود در سیستم عامل لینوکس Kali استفاده می‌کنیم. این سیستم عامل به منظور تست امنیت مورد استفاده قرار می‌گیرد و شامل پیشرفته‌ترین ابزارهای تست امنیت از جمله انجام حملات فیشینگ است که امروزه مهاجمان نیز آن را مورد استفاده قرار می-دهند زیرا امکان ساخت یک وب‌سایت فیشینگ را به راحتی فراهم می‌کند. برای این منظور ابتدا با رفتن به آدرس زیر این ابزار را فراخوانی می‌کنیم:

Go to Applications -> Kali Linux -> Exploitation Tools -> Social Engineering Tools ->se-toolkit

بعد از فراخوانی ابزار، آدرس سایتی که قصد ساخت صفحه جعلی از آن را داریم وارد نموده و صفحه جعلی به طور اتوماتیک ساخته می‌شود. آدرس صفحه جعلی ساخته شده آدرس آی پی مهاجم است که باید قربانیان را مجاب نماید که در آن نام کاربری و رمز عبور خود را وارد نمایند. شکل ۴ صفحه جعلی ساخته شده را نشان می‌دهد.



شکل ۴- صفحه فیشینگ ساخته شده توسط لینوکس Kali

جدول ۸-مشخصه‌های صفحهٔ فیشینگ ساخته شده توسط لینوکس Kali

مشخصه	پارامتر	مشخصه	پارامتر	مشخصه	پارامتر
استفاده از آدرس IP در آدرس وب-سایت	۱	وجود اطلاعات شناسایی گواهی امنیت	۰	غیرفعال شدن کلیک راست موس	۰
آدرس غیر عادی	H	وجود اطلاعات شناسایی گواهی امنیت	۰	وجود آدرس وب طولانی	۰
میزان وجود اختلال در آدرس دهی پایه	H	وجود صفحات انتقال	۰	جایگذاری کاراکترهای مشابه در آدرس	۰
وجود اطلاعات غیر عادی در سرور	H	مخفی نمودن آدرس واقعی صفحه	۰	افزودن پسوند و پیشوند در آدرس	۰
آدرس درخواستی	L	بررسی فرمهای کنترلی در سرور	۰	استفاده از کاراکتر @ در آدرس	۰
رتبه دامنه سایت در موتور جستجو	L	سایت کپی شده از سایت اصلی	۱	استفاده از کاراکترهای هگزا دسیمال	۰
وجود عنوان صفحه در لیست JCR	۰	استفاده از فرم با دکمه ارسال	۱	ترتیب در نتایج موتور جستجو	L
استفاده از گواهینامه امنیت سایت	۰	استفاده از پنجره پاپ آپ	۰		

برای تست تجربی رویکرد ارائه شده نیاز است تا مشخصه-های این صفحه استخراج شده سپس این مشخصه‌ها در الگوریتمهای انتخاب شده در بخشهای قبلی در نرم‌افزار Clementine وارد شوند و صحت نتایج بررسی گردد. مشخصه‌های این صفحه فیشینگ در جدول ۸ نشان داده شده‌اند.

با وارد نمودن این اطلاعات در رکوردهای دیتاست و آزمایش نمودن آن در نرم‌افزار Clementine، رویکرد مطرح شده صفحه مورد نظر را فیشینگ شناسایی نمود. لازم به ذکر است که پس از ساخت صفحات جعلی متعدد به وسیلهٔ ابزار نام برده تمامی آنها توسط رویکرد ارائه شده در این مقاله شناسایی گردیدند زیرا این ابزار از الگویی مشابه برای ساخت تمامی صفحات فیشینگ استفاده می-نماید.

۷- نتیجه گیری

در این مقاله به اختصار حملات فیشینگ معرفی شدند و تکنیکهایی که تاکنون جهت مقابله با این حملات ابداع شده‌اند مطرح گردید. سپس رویکردی جهت پیش‌بینی وب‌سایت‌های فیشینگ بر اساس وزن مشخصه‌های موجود در صفحات وب ارائه گشت. رویکرد ارائه شده میزان تاثیر هر مشخصه بر اساس وزن آن را در نظر می‌گیرد و با تقسیم مشخصه‌ها به دسته‌های مرتبط با یکدیگر تکنیکهای کلاس‌بندی را بر روی آنها اعمال می‌کند. مزیت عمده این تکنیک در نظر گرفتن همهٔ پارامترهای ممکن و موثر در شناسایی حملات فیشینگ است ضمن آنکه نسبت به سایر تکنیکهای مشابه که از الگوریتمهای کلاس‌بندی استفاده می‌کنند توانایی شناسایی ژورنال فیشینگها را دارا بوده و از نرخ خطای کمتری نیز بر خوردار است.

۸- منابع و مراجع

[1]. Wikipedia (July 2014), “Phishing”, Online Document Available at: http://en.wikipedia.org/wiki/Phishing

[2]. San Martino A and Perramon X (2010), “Phishing Secrets: History, Effects, and Countermeasures”, International Journal of Network Security, ۱۱(۳), ۱۶۳-۱۷۱.

[3]. McRae C.M and Vaughn R.B (2007), “Phighting the Phisher: Using Web Bugs and Honeytokens to Investigate the Source of Phishing Attacks”, Proceedings of the 40th Annual Hawaii International Conference on System Sciences (IEEE), 1-7, Waikoloa.

[4]. Agarwal N, Renfro S and Bejar A (2009), “Yahoo Sign-In Seal and Current Anti-Phishing Solutions”, eCrime Researchers Summit, 1-4.

[5]. Aburrous M, Hossain M. A, Dahal, K and Thabatah, F (2010), "Intelligent Phishing Detection System for E-Banking Using Fuzzy Data Mining”, Expert Systems with Applications, no. 37, 7913–7921.

[6]. Shreeram V, Suban M, Shanthi P and Manjula K (2010), “Anti-phishing detection of phishing attacks using genetic algorithm”, IEEE International Conference on Communication Control and Computing Technologies (ICCCCT), 447 – 450, Ramanathapuram, 7-9 Oct.

[7]. Chen J and Guo C (2006), “Online Detection and Prevention of Phishing Attacks”, First International Conference on Communications and Networking (IEEE), 1 – 7, China, 25-27 Oct.

[8]. Atighetchi M and Pal P (2009), “Attribute-based Prevention of Phishing Attacks”, Eighth International Symposium on Network Computing and Applications (IEEE), 266 – 269, Cambridge, MA, 9-11 July.

[9]. Dunlop M, Groat S and Shelly D (2010), “Gold Phish: Using Images for Content-Based Phishing Analysis”, the Fifth International Conference on Internet Monitoring and Protection (IEEE), 123 – 128, Barcelona, 9-15 May.

[10]. Mishra M, Gaurav and Jain A (2012), “A Preventive Anti-Phishing Technique using Code word”, International Journal of Computer Science and Information Technologies, 3(3), 2012, 4248 – 4250.

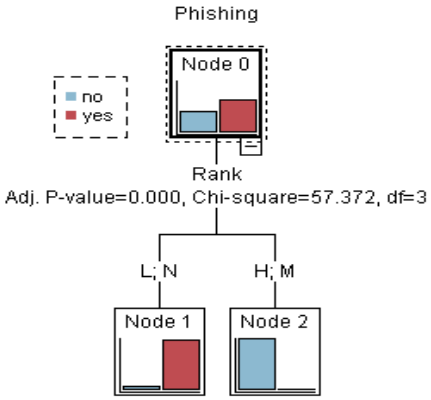
[11]. Sanglerdsinlapachai N and Rungsawang A (2010), “Using Domain Top-page Similarity Feature in Machine Learning-Based Web Phishing Detection”, Third International Conference on Knowledge Discovery and Data Mining (IEEE), 187 – 190, Phuket, 9-10 Jan.

[12]. Liu G, Qiu B and Wenyin L (2010), "Automatic Detection of Phishing Target from Phishing Webpage", International Conference on Pattern Recognition (IEEE), 4153-4156, Istanbul, 23-26 Aug.
[13]. Reddy V.P, Radha V and Jindal M (2011), "Client Side protection from Phishing attack", International Journal of Advanced Engineering Sciences and Technologies, 3(1), 39-45.
[14]. Khonji M, Jones A and Iraqi Y (2011), "A Novel Phishing Classification Based On URL Features', GCC Conference and Exhibition (IEEE), 221 – 224, Dubai, 19-22 Feb.
[15]. Ruth Ramya K, Priyanka K, Anusha K, Jyosthna Devi CH and Siva Prasad Y.A (2011), "An Effective Strategy for Identifying Phishing Websites using Class-Based Approach", International Journal of Scientific & Engineering Research, 2(12), 1-7.
[16]. Damodaram A, Phil M.C and Valarmathi M.L (2012), "Phishing website detection and optimization using Modified bat algorithm", International Journal of Engineering Research and Applications, 2(1), 870-876.
[17]. Aburrous M, Hossain M.A, Dahal K and Thabtah F (2010), "Associative Classification Techniques for predicting e-Banking Phishing Websites", International Conference on Multimedia Computing and Information Technology (IEEE), 9 – 12, Sharjah, 2-4 March.
[18]. Damodaram R and Valarmathi M.L (2011), "Fake Website Detection: Association Classification Algorithm with Ant Colony Optimization Technique", International Journal of Advanced Research in Computer Science, 2(1), 568-577.
[19]. Abdelhamid N, Ayesh A and Thabtah F (2014), "Phishing detection based Associative Classification data mining", Expert Systems with Applications, No. 41, 5948–5959.
[20]. COMODO (Aug 2014), "Comod Web Inspector", Online Tools Available at:
https://app.webinspector.com/
[21]. Phishtank (Aug 2014), "Comod Web Inspector", Online Document Available at:
http://www.phishtank.com/
[22]. Millersmiles (Aug 2014), "Comod Web Inspector", Online Document Available at:
http://www.millersmiles.co.uk/
[23]. Patil N, Lathi R and Chitre V (2012), "Customer Card Classification Based on C5.0 & CART Algorithms", International Journal of Engineering Research, 2(4), 164-167.
[24]. StatSoft Inc (Aug 2014), "Popular Decision Tree: Classification and Regression Trees", Online Document Available at:
http://www.statsoft.com/Textbook/Classification-and-Regression-Trees
[25]. Suknovic M, Delibasic B, Jovanovic M, Vukicevic M, Becejski-Vujaklija D and Obradovic Z (2011), "Reusable components in decision tree induction algorithms", Computational Statistics, 27(1), 127-148.
[26]. IBM (April 2015), "Clementine 11.1- Data Mining", Online Document Available at: http://www-304.ibm.com/partnerworld/gsd/solutiondetails.do?solution=10387&expand=true
[27]. Lim T.S, Loh W.Y and Shih Y.S (2000), "A Comparison of Prediction Accuracy, Complexity, and Training Time of Thirty-three Old and New Classification Algorithms", Machine Learning, No. 40, 203-229.
[28]. Kali Linux (April 2015), "Tools Included in the Set Package", Online Document Available at: http://tools.kali.org/information-gathering/set

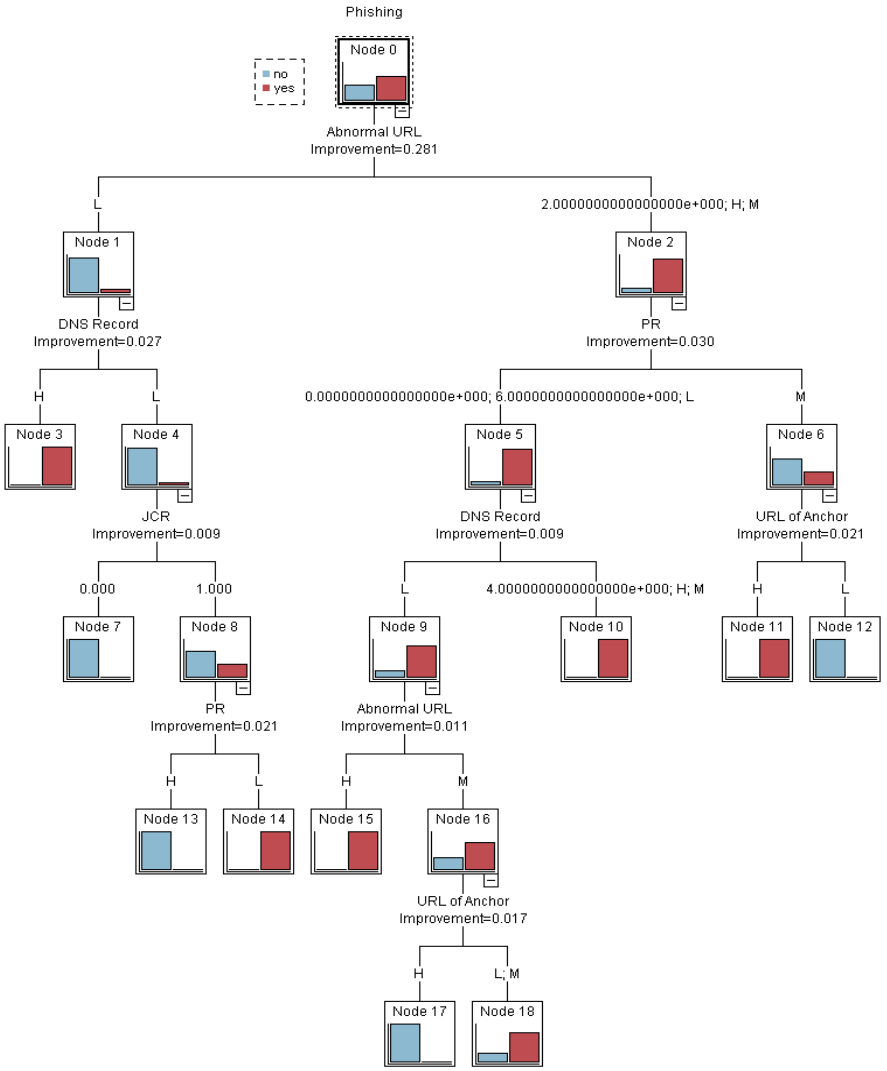
۹- پیوست‌ها

در ادامه خروجی الگوریتمهای مربوط به مشخصه دامنه و رمزنگاری، کدهای اسکریپتی و آدرس صفحه آورده شده‌اند.

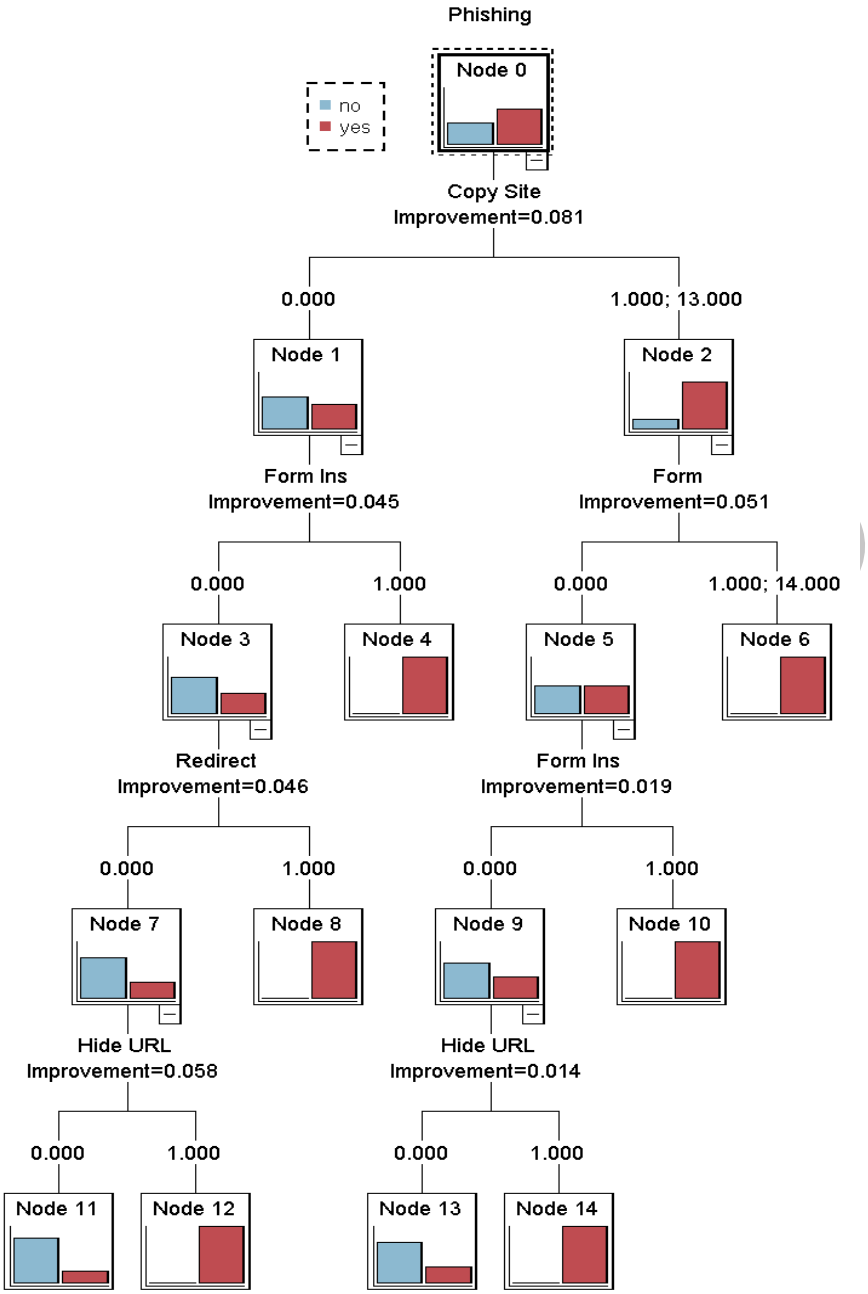
الف) خروجی الگوریتم آدرس صفحه وب



ب) خروجی الگوریتم مشخصه دامنه و رمزنگاری



ج) خروجی الگوریتم کدهای اسکریپتی و محتوای صفحه


Phishing
no
yes
Node 0
Copy Site
Improvement=0.081
0.000
1.000; 13.000
Node 1
Node 2
Form Ins
Improvement=0.045
Form
Improvement=0.051
0.000
1.000
0.000
1.000; 14.000
Node 3
Node 4
Node 5
Node 6
Redirect
Improvement=0.046
Form Ins
Improvement=0.019
0.000
1.000
0.000
1.000
Node 7
Node 8
Node 9
Node 10
Hide URL
Improvement=0.058
Hide URL
Improvement=0.014
0.000
1.000
0.000
1.000
Node 11
Node 12
Node 13
Node 14