PENERAPAN RANDOM FOREST CLASSIFIER UNTUK DETEKSI PDF MALWARE PADA LAYANAN AGREGATOR GARBA RUJUKAN DIGITAL (GARUDA) KEMENDIKBUD DIKTI

by 09011181823131 ALFIAH NUR FATMAWATI

Submission date: 19-Jan-2024 06:38PM (UTC+0700)
Submission ID: 2269742312
File name: kan_Digital_GARUDA_Kemendikbud_Dikti_-_alfiah_nur_fatmawati.docx (75.09K)
Word count: 4148
Character count: 25687

PENERAPAAN RANDOM FOREST CLASSIFIER UNTUK DETEKSI PDF MALWARE PADA LAYANAN AGREGATOR GARBA RUJUKAN DIGITAL (GARUDA) KEMENDIKBUD DIKTI)

BAB I
PENDAHULUAN

1.1 Latar Belakang

Portable Document Format yang biasanya di kenal PDF yaitu sistem format berkas yang di buat oleh adobe sytem pada tahun 1993 yang di gunakan sebagai pertukaran dokumen digital. Format PDF di manfaatkan untuk mempresentasikan dokumen dua dimensi yang meliputi grafik vektor dua dimensi ,huruf,teks,citra. PDF di kenalkan pertama kali di publikasikan pada tahun 1993, pada masa itu penggunaan format PDF relatif masih rendah.[1] Tidak dapat dipungkiri bahwa, seiring dengan kemajuan teknologi, masyarakat kini mengandalkan teknologi untuk menjalankan tugas sehari-hari maupun untuk terlibat dalam bidang politik, sosial, dan akademik [1]. Teknologi dimanfaatkan dalam berbagai macam hal seperti pada dunia Pendidikan, teknologi digunakan sebagai media belajar, untuk memahami perkembangan dunia digital penggunaan nya dengan cara memanfaatkan *e-book* , *e-learning* maupun pemanfaatan *Website* pemanfaatan dari teknologi yang berkembang ini tentunya memberikan keuntungan bagi kita yang mana hidup pada zaman modern yang begitu banyak memanfaatkan sarana teknologi digital [2].

Malicious Software atau yang biasa di sebut *malware* yaitu perangkat lunak secara eksplisit yang di desain untuk menjalankan berbagai macam perusak maupun aktifivas yang berbahaya bagi perangkat lunak lainya seperti *spyware, trojan, eksploit maupun virus*[3]. Oleh karena itu, untuk memastikan apakah aplikasi yang terdeteksi adalah malware dan untuk mengidentifikasi jenis malware, analisis dan deteksi merupakan langkah penting dalam menentukan potensi konsekuensi dari eksekusi sistem berbahaya. Mengenai hal-hal kebenaran yang dapat kita ketahui dari Malware yang telah di jelaskan pentingnya di lakukan metode Analisa agar mengetahui jenis-jenis dari serangan malware agar kita dapat mengetahui ciri atas malware itu sendiri.[2]

Garba Rujukan Digital yang di singkat GARUDA, merupakan wadah yang di gunakan sebagai tempat berkumpulnya informasi dan sarana pengetahuan yang menaungi sumber informasi yang melingkupi banyak aspek karya ilmiah yang ada di negara Indonesia, yang mana hal tersebut bertujuan untuk mengelolah dan mengakses karya ilmiah dengan lengkap dan mudah, aspek tersebut meliputi komputer, matematika dan prilaku[3].

Dataset *Imbalance* memungkinkan akan mengalami penurunan kesetabilan. Perolehan hasil yang di dapatkan lebih dominan akan memberikan lebih banyak pada mayoritas kelas. Dalam beberapa hal seperti pada *Multiclass classification*. Imbalance data akan memberikan representasi data sehingga mengakibatkan data yang lemah akan di acuhkan. Near Miss dan SMOTE yang merupakan kepanjangan dari *Syntetic Minorty Oversampling Technique* adalah metode undersampling dan Oversampling di gunakan oleh banyak orang untuk menghadapi masalah seperti dataset imbalance.

Randome Forest Classifier[3] adalah metode asembling yang di gunakan sebagai klasifikasi dan regresi dan tugas lainya yang beroperasi dengan membangun keputusan untuk pelatih dalam klasifikasi yang beroperasi dengan membangun banyak pohon keputusan pada waktu pelatihan untuk tugas klasifikasi hasil hutan acak untuk regresi prediksi rata-rata dari masing-masing pohon atau rata-rata dari masing-masing regresi untuk regresi masing-masing pohon di kembalikan [1][3][4].

Kinerja algoritma *Random Forest* di bandingkan dengan *Naive Bayes dan K-Nerest* setelah melalukan proses klasifikasi dengan data set yang telah di kumpulkan menunjukkan bahawa *Random Forest* memiliki akurasi dan daya ingat yang lebih tinggi di antara metode yang lain. Hal ini menunjukkan lompatan akurasi dibandingkan penelitian

Penelitian dalam hal ini akan berkonsentrasi pada sejumlah contoh yang telah kami kumpulkan, seperti file malware PDF. Dari 10.000 kumpulan data yang diambil, hanya 197 yang berbahaya; fitur kumpulan data ini belum diketahui saat ini, oleh karena itu diperlukan penyelidikan lebih lanjut. Analisis dan deteksi ini dilakukan untuk mencegah berbagai ancaman dan serangan yang dapat merugikan korbannya. Kerugian yang dapat ditimbulkan antara lain pencurian informasi secara kasar, peretasan, dan masuknya virus berbahaya ke dalam perangkat pengguna.

Dengan memberikan konteks, menjadi jelas bahwa selain manfaat yang bisa kita rasakan, ada juga kelemahan yang harus kita waspadai. Dengan pengetahuan yang kita miliki, kita dapat menangkis serangan virus dengan lebih efektif. Oleh karena itu, kami mempunyai harapan yang besar agar penelitian ini dapat membantu masyarakat dan memberikan pengaruh yang baik.

1.2. Perumusan Masalah

Berdasarkan latar belakang yang telah di jabarkan maka di perolehlah perumusan sebagai berikut:

1. Bagaimana teknik yang digunakan dalam mengekstrak Raw PDF dari PDF Malware Layanan Agregator GARUDA menjadi dataset.
2. Bagaimana cara penerapan untuk mengklasifikasi PDF Malware berupa *mal-pdf,benign, mal-html* menggunakan Algoritma *Random Forest Classifier*
3. Bagaimana pengaruh hasil performasi dan evaluasi dari kinerja algoritma *Random Forest Classifier* dalam dataset Imbalance pada dataset PDF Malware Layanan Agregator GARUDA.

1.3. Batasan Masalah

Batasan masalah tersebut antara lain merupakan batasan yang penulis miliki saat ini, yang bertujuan agar pembahasan tetap pada topik.

1. Data set untuk penelitian ini berasal dari PDF Malware di agregator garba rujukan digital (GARUDA) kemdikbud Dikti berjumlah 10.000
2. Menganalisa karakteristik PDF Malware hanya menggunakan metode yang diusulkan yaitu metode *Random Forest Classifier* pada PDF malware di agregator GARUDA kemdikbud Dikti
3. Tidak membahas bagaimana cara masuk dan mencegah serangan Malware pada file PDF.

1.4 Tujuan

Berdasarkan hasil penelitian ini tujuan yang akan dicapai adalah sebagai berikut :

1. Memudahkan dalam mengekstrak Raw data pada PDF malware di agregator Garba Rujukan Digital (GARUDA) kemdikbud Dikti menjadi data yang di olah agar menjadi dataset.
2. Penerapan Random Forest Classifier untuk klasifikasi PDF malware berupa *mal-pdf,benign, mal-html*.
3. Pengaruh hasil performasi dan evaluasi dari kinerja algoritma Random Forest dalam dataset Imbalance pada dataset PDF malware GARUDA.

1.5 Manfaat

Adapun manfaatnya yang dapat di peroleh dari penelitian Skripsi ini sebagai berikut;

1) Kumpulan data Raw PDF malware yang telah di ekstraksi akan di ubah menjadi dataset
2) Memahami dan mampu mengidentifikasi malware berdasarkan kategorisasi PDF Malware Referensi Digital Kementerian Pendidikan dan Kebudayaan. Agregator GARUDA Garba menggunakan teknik Random Forest Classifier
3) Memperoleh hasil yang tepat dan optimal dalam mendekteksi PDF malware

1.6. Manfaat bagi kampus

Adapun manfaatnya yang dapat di peroleh dari penelitian Skripsi ini sebagai berikut;

1. Memahami permasalahan mengenai penyerangan Malware.
2. Menganalisa karakteristik malware yang ada pada PDF Malware di agregator GARUDA kemdikbud Dikti dengan Metode *Random Forest Classifier*
3. Menambah wawasan dan pengetahuan yang bisa digunakan sebagai acuan dalam Deteksi Malware pada suatu instansi, serta kampus yang mana bisa bermanfaat sebagai bahan ajar agar ilmu secara teori.
4. Tidak membahas tentang bagaimana mencegah serangan Malware.

BAB II
TINJAUAN PUSTAKA

2.1 Pendahuluan

Tahap selanjutnya yang di lakukan adalah mencari penelitian yang berkaitan dengan tugas akhir yang kerjakan, seoperti penelitian apa yang telah di lakukan dan apa yang di hasilkan pada penelitian sebelumnya serta dapat mengetahui metode apa saja yang di gunakan untuk menyelsaikan permasalahan yang terkait.

2.2 Penelitian Terkait

Penelitian telah memanfaatkan malware secara ekstensif. Berikut ini adalah teknik yang digunakan para peneliti:

Penelitian yang berjudul "Analisis dan Deteksi Malware Menggunakan Metode Malware Analisis Dinamis dan Malware Analisis Statis" [5] Tujuan dari penelitian ini adalah untuk menentukan apakah keluaran yang dihasilkan oleh kedua metode dinamis dan statis untuk menganalisis malware Posion Ivy RAT (Remote Access Trojan) adalah sama. Kemampuan untuk mengidentifikasi malware Poison Ivy RAT merupakan fungsi dari penerapan pendekatan ini dan pemahaman metode konsep Analisis Statis [5].

Penelitian yang berjudul "Analisis malware menggunakan Metode Dynamic Analysis pada jaringan Universitas Sam Ratulangi" [2]. Tujuan dari penelitian ini adalah untuk mengetahui jenis malware yang ditemukan pada jaringan Universitas Sam Ratulangi. Cockoo Sandbox merupakan instrumen yang digunakan dalam penelitian ini, dan pendekatan yang digunakan adalah metode Analisis Dinamis. Sehingga terhindar dari serangan infeksi *malware*. Berdasarkan hasil analisa yang telah di lakukan kepada malware sehingga dapat di ketahui karakteristik *malware*, dan juga dapat di terangkan bahwasanya beberapa terdapat *string*, *signiture* dan perubahan *value resgistrasi* padanya. Kaitanya dalam metode ini adalah kesepahaman tentang bagaiamana mengidentifikasi dan mengetahui karakteristik yang ada pada sebuah

malware .[2]

Penelitian yang berjudul “Tools and Techniques for colletion and Analysis of Internet-of-Things malware ” DarckComet digunakan dalam penelitian ini bersama dengan pendekatan analitis dan statis. Temuan penelitian ini menunjukkan bahwa pendekatan ini dijelaskan dengan cara yang efektif dan efisien, sehingga dapat meningkatkan kinerja deteksi dan memastikan bahwa virus tersebut dihilangkan.[6]

Penelitian dengan Tema “Random Forest Classifier” [7] Pada penelitian tersebut tentang membahas Analisa malware Trojan dengan menggunakan Metode Dinamis dan Statis pada operasi sistem windows. Kaitanya dengan penelitian karena kesepahaman konsep Analysis Static untuk mengetahui karakter sebuah malware[7]

Penelitian yang berjudul “Server Analysis Malware pembangunan menggunakan Cucook Sanbox pada Sistem operasi berbasis Linux”. Penelitian ini menggunakan alat analisis yang sama Cockook Sanbox dan membahas tentang analisis malware yang memanfaatkan teknik dinamis. [8]

Beberapa Penelitian yang dilakukan oleh peneliti diatas dapat di lihat pada tabel 1. Berkaitan dengan metode mendasar yang akan diterapkan, maka temuan penelitian yang bertajuk “Penerapan Random Forest Classifier untuk Deteksi PDF Malware di Agregator Garba Rujukan Dgital (GARUDA) Kemendikbud Dikti” telah diperoleh. sebagian besar memenuhi persyaratan penelitian yang akan dilakukan.

2.3 Landasan Teori

2.3. 1 Malware

Malware, atau perangkat lunak berbahaya, adalah perangkat lunak yang sengaja dibuat dan diciptakan untuk merusak, menyusupi, atau mendapatkan akses ke sistem komputer tanpa sepengetahuan pemiliknya sehingga dapat menimbulkan risiko pada sistem dan menimbulkan dampak buruk yang berbeda-beda.[6] Malware hadir dalam berbagai bentuk, termasuk virus dan lainnya, seperti "trojan", "keylonggers", dan "spyware", yang dapat berperilaku berbahaya.[2]

2.4 PDF Malware

File PDF Malaware[2] adalah sebuah bentuk file yang mudah di akses dan juga di manupulasi karena di dalamnya terdapat teks dan juga hanya sedikit batasan untuk para oknum hacker untuk melakukan peretasan file tersebut. Pada adobe memberikan format yang berisikan penambahan format algoritma enkripsi, scripting, multi media serta support. [5]

Dengan adanya file PDF Malware maka dengan mudah para oknum peretas untuk menanamkan sebuah malware yang berbahaya yang tidak di ketahui para user, dalam PDF Malware tedapat beberapa tools-tools yang dapat mengidentifikasi ciri-ciri malware yang sangat bermanfaat dalam membantu untuk mendeteksi karakteristik dari malware tersebut di deteksi dengan menggunakan Metode Random Forest Classifier.[3]

2.5 PDF ID

Ekstraksi dataset file PDF Malware di lakukan dengan menggunakan PDF ID [7]sebagai alat yang di gunakan dalam menyelsaikan tugas akhir ini. PDF ID merupakan sebuah tools yang di rancang oleh seseorang Bernama Didier Stevens. Analisa secara statis di lakukan menggunakan PDF id karena PDF id adalah Script phyton. File PDF tersebut akan di tinjau secara langsung menggunakan script yang telah di rancang. Setiap fitur melakukan perhitungan terhadap nilai yang ada. Satu fitur yang umumnya di temukan yang terasa mencurigakan dari sekian banyaknya. Nilai dan fitur yang telah di dapatkan maka akan di konversikan ke dalam bentuk CSV.[9]

Gambar 2.1 Antarmuka PDFiD

2.6 Ekstraksi Dataset

Ekstraksi dataset di lakukan untuk memperoleh hasil Analisa secara statis sebagai penunjang dalam menyelsaikan tugas Akhir secara parsing untuk semua file PDF terhadap dataset agar memperoleh hasil headeryang bermanfaat untuk fitur dataset.

2.7 Dataset PDF Malware

Untuk saat ini sudah ada sebanyak 10.000 PDF Malware yang telah kami analisa menggunakan virustotal, sejauh ini kami telah mendapat penambahan 4 malicious jenis dokumen PDF sehingga total dari malicious pdf menjadi 197. Dataset yang digunakan adalah dataset PDF Malware GARUDA.

Dataset PDF Malware GARUDA diperoleh untuk saat ini sudah ada sebanyak 10.000 PDF Malware yang telah kami analisa menggunakan virustotal, sejauh ini kami telah mendapat penambahan 4 malicious jenis dokumen pdf sehingga total dari malicious pdf menjadi 197. Dengan melihat perbandingan total pdf benign(8350) dan PDF malicious(197) yang ada.

2.8 SMOTE

Ketidak seimbangan pada pengklasifikasian suatu bahan akan melibatkan model prediktif sebagai bentuk pengembangan pada data yang mengalami ketidak seimbangan

Kumpulan data yang tidak seimbang merupakan tantangan dalam bekerja sehingga pembelajaran mesin akan sedikit di abaikan. Pendekatan yang dapat di lakukan salah satunya adalah menggunakn SMOTE[6] dengan cara mengambil sample minoritas dengan berlebiha. Duplikasi merupakan salah satu contoh pendekatan paling sederhana. Ada lanjutanya

2.9 Machine Learning

Machine learning adalah salah satu dari cabang Artificial Intelegent (Kecerdasan Buatan). Machine learning di kembangkan agar secara langsung dapat di pelajari itulah mengapa disebut mesin pembelajar, berdasarkan ini machine learning berisi ilmu-ilmu statistika, matematika dan juga lainya.[10]

2.10 Random Forest Classifier

Untuk mengumpulkan hasil tugas klasifikasi, Randome Forest Classifier[3] membuat keputusan untuk pelatih dalam klasifikasi dengan membangun banyak pohon

keputusan selama pelatihan. Rata-rata prediksi setiap pohon atau rata-rata setiap regresi untuk setiap pohon regresi dikembalikan dalam hasil Random Forest untuk regresi.

2.11 Metode Malware Analisis

Pada umumnya malware ialah sebuah program yang di kelompokkan tentu berdasarkan tujuan tertentu algoritma dan logika yang di gunakan yang relavan dengannya. Hasilnya, model analitik yang digunakan untuk menyelidiki malware memiliki ikatan yang kuat dengan ilmu komputer dasar. seperti struktur data, algoritma, bahasa pemrograman, dan rekayasa perangkat lunak.

Secara umum, sebuah perangkat lunak menggunakan salah satu dari tiga bentuk analisis untuk menentukan apakah sesuatu yang terhubung dengannya adalah malware atau bukan. Oleh karena itu, ketiga model tersebut yang masing-masing akan diberikan dan dibahas sebagai berikut merupakan strategi yang dapat diterapkan.[5]

2.12 Malware Analisis Statis

Berbeda dengan pendekatan analisis dinamis, analisis statis melibatkan mempelajari perangkat lunak tanpa menjalankannya. File malware tidak akan langsung terpicu selama analisis statis; sebaliknya, kode sumber tertulis akan diteliti dan ditelusuri, menciptakan kembali kode sumber dan algoritma yang telah dikembangkan oleh program. Data yang dikumpulkan bersifat komprehensif dan dapat memberikan gambaran yang sangat jelas tentang fungsi sistem malware secara keseluruhan. Debugger, assembler, dan program analisis semuanya dapat digunakan untuk analisis statis. Berikut adalah beberapa contoh metode analisis statis yang berbeda.

1. metode pendeteksian yang mengandalkan metode tanda tangan yang biasa disebut dengan metode sidik jari, masker, atau metode pencocokan string atau pola. Tanda tangan adalah serangkaian program yang dimasukkan oleh pemrogram malware ke dalam aplikasi untuk mengidentifikasi bagian malware tertentu. Pendeteksi malware mencari tanda tangan yang telah ditentukan sebelumnya dalam kode untuk mengidentifikasi konten berbahaya.
2. Teknik deteksi heuristik Nama lain metode ini adalah metode proaktif. Metode

ini sebanding dengan metode yang mengandalkan tanda tangan kode tertentu; pendeteksi malware sekarang mencari perintah atau instruksi yang tidak ada dalam perangkat lunak aplikasi. Hasilnya, identifikasi jenis malware yang baru teridentifikasi menjadi lebih mudah. Berikut ini adalah banyak metode analisis heuristik.

a). File based heuristic analysis

Analisis file adalah jenis file yang mencakup analisis heuristik. Metode analisis file ini memeriksa secara menyeluruh konten file, pemrosesan, dan tujuan penggunaan, serta perintah apa pun yang mungkin disertakan untuk menghancurkan atau merusak file lain.

b). Weight based heuristic analysis

Analisis heuristik berbasis bobot adalah metode lama. Setiap permohonan diberi bobot berdasarkan potensi bahayanya. Program diduga mempunyai kode bahaya jika nilai bobotnya lebih besar dari nilai ambang batas yang terdeteksi.

c). Rule based heuristic analysis

Dalam hal ini, analisis menghasilkan aturan yang menentukan penerapannya. Setelah kriteria tersebut digabungkan dengan aturan yang telah ditetapkan sebelumnya, program dianggap mengandung malware jika aturannya tidak sesuai.

d). Generic signature analysis

Meskipun virus yang dimaksud adalah variasi malwar, virus ini memiliki kemiripan dengan “kembar identik” dalam hal perilaku. Teknik ini mencari varian malware baru dengan memanfaatkan definisi antivirus yang sudah ada.

e).Keuntungan dari metode analisis statis.

Lebih aman dan cepat menggunakan analisis statistik untuk mengumpulkan struktur kode program untuk analisis tertentu. apakah tindakan dan perilaku tindakan keamanan di masa depan dapat dihitung menggunakan analisis statis.

f). Kerugian dari metode analisis statis

Menganalisis malware yang tidak dikenal masih diperlukan selain analisis statis. Banyak kode sumber aplikasi yang sulit ditemukan, oleh karena itu melakukan analisis statis memerlukan peneliti untuk memiliki pemahaman yang lebih baik tentang cara

kerja sistem.

2.13 Malware Analisis Dinamis

Analisis dinamis adalah metode yang digunakan dalam teknik ini untuk memeriksa perilaku atau operasi yang dilakukan program saat sedang berjalan. [5] Salah satu cara untuk melakukan analisis dinamis adalah dengan mengikuti panggilan, memantau fungsi, dan mengumpulkan informasi tentang dampak malware yang teridentifikasi saat dilakukan. Akhrinya dapat kita ketahui apa saja dan kegiatan apa yang telah malware lakukan pada saat berhasil menginfeksi sebuah perangkat komputer. Pemerikasaan akan di lakukan secara keseluruhan pada tahapan analisis dinamis. Untuk penelitian ini, mesin virtual atau sandbox biasanya digunakan. Aplikasi yang meragukan biasanya dijalankan di lingkungan virtual; jika aplikasi bekerja aneh, maka dianggap berbahaya atau terinfeksi malware.

1. Salah satu manfaat analisis dinamis adalah memudahkan pengguna mengidentifikasi virus yang tidak dikenal hanya dengan melihat perilaku program.

2.Kerugian dari analisis dinamis.

Pemeriksaan ini memerlukan waktu karena program mungkin berjalan lambat atau tidak aman dalam beberapa situasi. Aplikasi yang menunjukkan variasi perubahan perilaku dengan berbagai situasi pemicu tidak tercakup dalam pemeriksaan ini. misalnya, tidak mengidentifikasi malware multipath.

2.5.1 Analisis Hybrid

Teknik analitik statis dan dinamis digunakan dalam penyelidikan ini. Metode. Manfaat dari dua strategi sebelumnya digabungkan dengan cara ini.

2.14 Hasil Studi Pustaka

Tujuan dari studi pustaka ini mengumpulkan tinjauan dari penellitian sebelumnya yang telah telah di lakukan dengan yang akan di kerjakan oleh peneliti untuk mengetahui keterkaitan

kan terlihat pada gambar 3.14

BAB IV
HASIL DAN ANALISA

4.1 Pendahuluan

Pada bab 4 ini akan membahas tentang tahapan-tahapan pengujian yang akan di lakukan dalam penelitian sesuai dengan prosedur perancangan sistem berupa *pre-processing* dan *processing*, dan selanjutnya akan di lakukan pelabelan data lalu melakukan *Oversampling* di karenakan dataset pada penelitian ini data imbalance. Pengujian data pada PDF malware dengan menggunakan *Random Forest* pada tahap *processing* lalu di lakukan SMOTE yang di gunakan untuk menyeimbangkan data yang imbalance dan proses *K-fold cross validation* pada tahap processing klasifikasi PDF malware di lakukan dengan *Random Forest* dan melakukan validasi kepada data pengujian.

4.2 Dataset

Dataset yang telah di olah berasal dari GARUDA *repository* yang telah di ekstraksi. Dalam data tersebut terdiri dari 3 kelompok yaitu kelas *benign, mal-pdf, dan non-pdfmal* dataset ini berjumlah 10.000 yang mana data tesebut terdiri dari 9.800 data benign, 194 data non-malpdf dan 6 data mal-pdf. Dataset tersebut memiliki 21 atribut dan 1 class label. Dapat dilihat pada tabel 4.1 untuk gambar dataset PDF Malware GARUDA dan Tabel Dataset sebelum di Normalisasi pada tabel 4.1

Gambar 4.1 Gambar File PDF Malware

Tabel 4.1 Tabel Dataset GARUDA

4.2.1 Pelabelan Data

Agar mudah di proses maka label sebelumnya di lakukan perubahan menjadi 0 untuk benign, 1 untuk non-malpdf dan 2 untuk mal-df

Gambar 4.2 Pelabelan Data

4.2.2 Analisa Statis PDF GARUDA

Pada tahap ini di lakukan untuk langkah awal tugas akhir. Dataset berupa file PDF akan di ekstraksi menggunakan sistem operasi Kali Linux yang telah di instal pada VirtualBox. Pada tahap ini di lakukan secara Statis di lakukan pengecekan secara manual satu persatu pada setiap file PDF pada dataset. Pada prosesnya di bagi menjadi dua macam yaitu;

1. Analisa menggunakan *Website* VirusTotal

   Pada tahap ini yaitu kami memiliki file PDF malware berjumlah 10.000 file, yang akan di ekstraksi menggunakan virus total untuk di analisis dan cek secara manua;l satu persatu pada *website*, selanjutnya VirusTotal akan secara otomatis menjalankan sistem untuk medeteksi dan mengidentifikasi apakah file tersebut menyisipkan malware atau tidak. Hasil dari 10.000 file PDF yang telah di ekstraksi adalah 9800 benign, 200 file pdf malware. Analisa tersebut dapat di lihat pada gambar 4.2.

Gambar 4.3 Analisa Statis VirusTotal

2. Analisa Menggunakan Terminal Linux

   Proses analisis ini menggunakan PDFiD setelah melakukan proses tersebut dapat di peroleh hasil dari 200 file PDF malware yang telah di analisa di hasilkan 194 file mal-html dan 6 file pdf-mal setelah di lakukan analisa menggunakan PDFiD menghasilkan nilai dari 21 atribut dataframe yang di miliki, berikut gambar nya setelah di lakukan normalisasi terlihat pada gambar 4.3

Gambar 4.2 Analisa PDFiD

Hasil dari analisa menggunakan PDFiD dapat di ketahui file PDF malware yang berjumlah 10.000 dan telah di analisa menghasilkan 9800 file *benign*, 194 file *non-pdfmal* dan enam *pdf-mal*. Sehingga hasil yang di dapatkan adalah kumpulan-kumpulan dataframe dari nilai tersebut yang selanjutnya akan di ubah menjadi *dataset* dalam bentuk *.csv (Comma Separated Value)* hal tersebut dapat terlihat pada tabel 4.2

3. Fitur Atribut Dataframe

   1) Obj

      Yaitu fitur yang berfungsi untuk menunjukkan nilai yang sesuai dengan jumlah objek yang terdapat pada file

   2) EndObj

      Yaitu sebuah fitur yang mana nilainya wajib sama dengan Obj karena mereka berpasangan

   3) Stream

      Yaitu fitur yang nilainya dapat di hitung sesuai dengan instruksi yang urut dan menjelaskan tampilanya.

4) EndStream

   Yaitu fitur yang merupakan pasangan dari Stream karena pada EndStream nilainya harus sama dengan Stream

5) Xref

   Yaitu fitur yang mana nilainya dapat mengetahui apakah file PDF merupakan file yang berbahaya atau tidak. Jika nilai yang di hasilkan dari fitur ini sama dengan 0, maka tidak sempurna file PDF tersebut dengan kata lain tidak aman

6) Trailer

   Yaitu fitur yang mana dapat mengetahui apakah file PDF tersebut adalah file yang membahayakan atau tidak, jika nilai dari fitur tersebut sama dengan 0, maka file tersebut tidak aman

7) Starxref

   Yaitu fitur yang merupakan pasangan dari Xref, yang mana nilainya wajib sama dengan fitur dari Xref untuk dapat mengetahui apakah file PDF berbahaya atau tidak

8) /Page

   Yaitu fitur yang memiliki nilai yang sama dengan jumlah yang ada di file PDF

9) /Encrypt

   Yaitu fitur yang nilainya dapat mengetahui adanya kode yang di pin di dalam file PDF

10) /ObjStm

    Fitur tersebut dapat di gunakan untuk menyamarkan sebuah objek yang dapat menuju pada file PDF berbahaya

11) /JS

    Fitur tersebut adalah tanda terdapatnya JavaScript yang terdapat pada file PDF nilai pada fitur tersebut di hitung dari jumlah pernyataan pada blok JavaScript

12) /JavaScript

Fitur tersebut juga salah satu tanda adanya JavaScript yang terdapat pada sebuah File PDF, tetapi nilai dalam fitur ini di hitung dari jumlah blok pada pernyataan JavaScript

13) /AA

Nilai yang fitur ini tunjukkan yaitu jumlah tindakan yang akan di lakukan ketika file di akses, menunjukkan respon

14) /OpenAction

Nilai dari fitur ini menunjukkan banyaknya tindakan yang akan di lakukan ketika file PDF di akses tetapi tidak membutuhkan respon dari pengguna untuk menjalankanya, karena otomatis

15) / Acroform

Nilai fitur tersebut di hitung dari jumlah konten yang ada pada sebuah file PDF dan berjalan secara otomatis

16) /JBIG2Decode

Menunjukkan berapa jumlah pengguna yang mengakses filter JBIG2Decode pada file PDF

17) /Launch

Nilai fitur tersebut di hitung dari jumlah tempat yang dapat digunakan untuk mendemokan program Script pada file PDF.

18) /Embeddedfile

Nilai pada Fitur tersebut dapat berfungsi untuk mengetahui apakah adanya file yang dapat membahayakan di dalam PDF. Fitur ini akan menginformasikan jumlah kode yang di tanam pada file PDF.

4.3 Pre-Processing

4.3.1 Analisa Dataset

Tahap ini di lakukan untuk analisa dataset PDF Malware untuk mengetahui informasi mengenai dataset tersebut

1. Dataset PDF malware terdiri dari 10.000 file PDF
2. Dataset memiliki dua tipe yaitu data objek dan data integer
3. Dataset memiliki 3 kelas yaitu kelas *benign, mal-pdf, dan non-malpdf*
4. Dataset tidak memiliki ruang kosong
5. Dataset termasuk kedalam golongan kelas Imbalance karena data tersebut berjumlah 9800 untuk *file benign*, 194 untuk *file non-pdfmal* dan 6 untuk *file mal-pdf*

Seperti yang terlihat pada gambar 4.3 merupakan hasil dari dataset dengan imbalance.

Gambar 4.3 *Dataset Imbalance*

4.3.2 Normalisasi

Proses Normalisasi bertujuan untuk menyamakan nilai angka yang terdapat pada dataset dalam skala yang sama sehingga angka-angka tersebut memiliki selisih minimum. Nilai dari dataset akan di ubah dari skala 0 menjadi skala 1. Hasil normalilasi dapat di lihat pada tabel 4.2

4.4 Processing

4.4.1 *Resampling*

Proses *resampling* di bagi menjadi 2 tahap yang berbeda dalam penelitian ini, yang pertama yaitu penerapan *Oversampling* dan yang kedua yaitu penggabungan

Oversampling dan undersampling dalam hal ini pembagian tersebut bertujuan untuk mengetahui perbedaan yang di hasilkan dari masiing-masing teknik yang selanjutnya akan di bandingkan untuk memperoleh hasil terbaik

Proses *resampling* yang akan di gunakan adalah sebagai berikut;

Resampling dengan SMOTE yang di gunakan pada metode untuk resampling dengan SMOTE ini adalah penerapan dari Oversampling, proses SMOTE ini akan menghasilkan data baru secara sintetis. Data yang akan dibuat bisa mirip dengan data yang asli pada kelas minoritas.

Proses dari Oversampling ini akan menghasilkan jumlah data pada setiap kelasnya dengan jumlah yang sama data mayoritas yaitu sebanyak 9800 pada setiap kelasnya. Hasil dari proses *Resampling* menggunakan SMOTE dan undersampling menggunakan *NearMiss* bahwa jumlah data yang di hasilkan benign, mal-pdf, dan non-pdfmal. Yang telah terdistribusi telah *balance* dengan jumlah data 3000 dari hasil resampling dari masing-masing teknik dan pada masing-masing kelas. Hasil dapat di lihat pada gambar yang ada di bawah yaitu terlihat pada gambar 4.4

Gambar 4.3 *Dataset Balance*

4.4.2 Split Data

Split data training dan juga data testing merupakan proses yang di lakukan untuk melakukan pengujian model pada penelitian. Yang di lakukan yaitu proses pembagian data pada penelitian yaitu 20% data testing dan 80% data training.

4.5 Processing

Proses ini merupakan tahap yang di lakukan untuk klasifikasi PDF Malware dengan menggunakan Random Forest.

4.6 Hasil Percobaan pada klasifikasi Random Forest

Proses percobaan pada Klasifikasi ini menggunakan metode Random Forest

menggunakan *GridSearchCV*. Dari hasil nya dapat di ketahui *confussion matrix* dapat di lihat pada gambar 4.4

Tabel 4.4 Perbandingan Hasil Performa

Tabel dari hasil percobaan maka di peroleh hasil akurasi yang terbaik ketika jumlah pada pohon adalah 40% yaitu menunjukkan sebesar 86,22%, kemudian max_dipth yang di gunakan yaitu 10, dan min_samples_split yaitu 3.

Gambar 4.3 Confussion Matrix

Tabel 4. Akurasi

BAB V

KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan hasil penelitian yang telah di lakukan dapat di simpulkan sebagai berikut;

1. Dataset yang di miliki berjumlah 10.000 dataset dengan perolehan 9800 file *benign*, 194 *file non-pdfmal* dan 6 *file pdf-mal*. Dari proses ektraksi yang telah di lakukan, menggunakan tools dari PDFiD menggunakan sistem operasi kali linux. Yang selanjutnya data yang telah di ekstraksi di ubah kedalam bentuk dataset *.cvs (Comma Separated Value)*. Dataset yang di miliki merupakan kelas imbalance atau data tidak seimbang hal tersebut merupakan permasalahan data maining yang perlu di lakukanya proseses keseimbangan menggunakan proses *Resampling* dengan penerapan *Oversampling* SMOTE dan *Undersampling*
2. Proses untuk menghasilkan Performasi dan evaluasi pada penelitian ini menggunakan metode *Random Forest Classifier*. Selanjutnya hasil performa tersebut akan di lakukanya proses Validasi
3. Perolehan Hasil akurasi sebelum Resampling adalah sebesar 100% setelah di lakukan Proses Resampling dengan Oversampling maka nilai akurasi 86,22%

PENERAPAN RANDOM FOREST CLASSIFIER UNTUK DETEKSI PDF MALWARE PADA LAYANAN AGREGATOR GARBA RUJUKAN DIGITAL (GARUDA) KEMENDIKBUD DIKTI

ORIGINALITY REPORT

1%	1%	0%	0%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	jurnal.sar.ac.id Internet Source	1%

Exclude quotes	On	Exclude matches	< 1%
Exclude bibliography	On		

SURAT KETERANGAN PENGECEKAN
SIMILARITY

Saya yang bertanda tangan di bawah ini

Nama	: Alfiah Nur Fatmawati
Nim	: 09011181823131
Prodi	: Sistem Komputer
Fakultas	: Fakultas Ilmu Komputer

Menyatakan bahwa benar hasil pengecekan similarity Skripsi/Tesis/Disertasi/Lap. Penelitian yang berjudul **PENERAPAN RANDOM FOREST CLASSIFIER UNTUK DETEKSI PDF MALWARE PADA LAYANAN AGREGATOR GARBA RUJUKAN DIGITAL (GARUDA) KEMENDIKBUD DIKTI** adalah 1%.

Dicek oleh operator *: 1. Dosen Pembimbing

(2) UPT Perpustakaan

Demikianlah surat keterangan ini saya buat dengan sebenarnya dan dapat saya pertanggung jawabkan.

Indralaya, Januari 2024

Menyetujui Dosen pembimbing I	Dosen pembimbing II	Yang menyatakan,
Prof. Deris Stiawan, M.T.,Ph.D	Nurul Afifah, M.Kom.	Alfiah Nur Fatmawati
NIP. 197806172006041002	NIP.199211102023212049	NIM. 09011181823131

PENERAPAN RANDOM FOREST CLASSIFIER UNTUK DETEKSI PDF MALWARE PADA LAYANAN AGREGATOR GARBA RUJUKAN DIGITAL (GARUDA) KEMENDIKBUD DIKTI

by 09011181823131 ALFIAH NUR FATMAWATI

Submission date: 19-Jan-2024 06:38PM (UTC+0700)
Submission ID: 2269742312
File name: kan_Digital_GARUDA_Kemendikbud_Dikti_-_alfiah_nur_fatmawati.docx (75.09K)
Word count: 4148
Character count: 25687

PENERAPAN RANDOM FOREST CLASSIFIER UNTUK DETEKSI PDF MALWARE PADA LAYANAN AGREGATOR GARBA RUJUKAN DIGITAL (GARUDA) KEMENDIKBUD DIKTI

ORIGINALITY REPORT

1%	1%	0%	0%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	jurnal.sar.ac.id Internet Source	1%

Exclude quotes	On	Exclude matches	< 1%
Exclude bibliography	On		