E-ISSN : 2827-9425
P-ISSN : 2828-4666

Klasifikasi Malware Menggunakan Teknik Machine Learning

Evan Valdis Tjahjadi[1], Budy Santoso[2], Serwin[3]
Ilmu Komputer, Teknik Informatika, Universitas Ichsan Gorontalo, Gorontalo, Indonesia
Email: [1]evantjahjadi4@gmail.com , [2]budinho.jr@gmail.com,[3]erwindsn.ui@gmail.com

Abstrak - Jaringan komputer yang terhubung dengan internet dapat mengakses informasi dari seluruh dunia dengan sangat mudah. Namun, koneksi antara jaringan dan Internet justru meningkatkan potensi kegagalan sistem. Salah satu metode yang bisa digunakan pada machine learning merupakan metode algoritma random forest. Random forest merupakan salah satu metode pada machine learning yang digunakan untuk memecahkan masalah klarifikasi, Berdasarkan permasalahan tersebut perlu dilakukan klasifikasi malware yang datanya diambil dari dataset malware agar dapat memudahkan dalam mempelajari dan membedakan jenis malware. Proses terdiri dari pengumpulan dataset, pre proseccing, melatih machine learning dan melakukan pengujian performa model atau kinerja. Penelitian ini bertujuan untuk mengetahui performa atau kinerja Machine Learning menggunakanalgoritma random forest untuk klasifikasi malwarerandom forest. Pada proses ini pra pemrosesan data dilakukan dengan menginstal beberapa library python. Pandasadalah library python open source yang biasanya digunakan untuk kebutuhan data analisis. Model dilatih pada dataset dengan berbagai fitur dan hasilnya menunjukan akurasi yang tinggi sebesar 99%, yaitu berupa proses menganalisis sekumpulan data untuk meringkas karakteristik utamanya agar pengguna lebih memahami dataset yang akan digunakan. Model random forest memberikan hasil yang sangat baik tanpa preprocessing pada data. Hasilnya bagus meskipun datanya tidak seimbang. Tidak perlu menggunakan teknik apapun untuk menyeimbangkannya. Penskalaan/skaling tidak perlu dilakukan, model andom forest adalah model partisirekursif yang bergantung pada partisi data karena ia bekerja pada pemisahan nilai fitur dan tidak melkukan perhitungan di dalamnya. Hasil penelitian menunjukkan bahwa model memiliki presisi 0,99.

Kata kunci: klasifikasi malware, machine learning, metode Random Forest

Abstract *Computer networks connected to the Internet can access information from all over the world very easily. However, the connection between the network and the Internet increases the potential for system failure. One of the methods that can be used in machine learning is the random forest algorithm method. Random forest is one of the methods in machine learning that is used to solve clarification problems. Based on the problems, it is necessary to classify malware where data is taken from malware datasets to make it easier to learn and distinguish the types of malware. The process consists of collecting datasets, pre-processing, training machine learning, and testing model performance. This study aims to find out the performance of Machine Learning using a random forest algorithm for malware- random forest classification. In this process, pre-processing of data is done by installing several Python libraries. Pandas is an open-source Python library that is usually used for data analysis needs. The model is trained on a dataset with various features and the results show a high accuracy of 99%. The random forest model provides excellent results without preprocessing the data. The results are good even if the data is not balanced. There is no need to use any technique to balance it. Scaling is not necessary. The random forest model is a recursive partitioning model that depends on data partitioning as it works on splitting the feature values and does not perform any calculations in it. The results indicate that the model has a precision of 0.99.*

Keywords: malware classification, machine learning, Random Forest method

1. PENDAHULUAN

Jaringan komputer yang terhubung dengan internet dapat mengakses informasi dari seluruh dunia dengan sangat mudah. Namun, koneksi antara jaringan dan Internet justru meningkatkan potensi kegagalan sistem. Komputer menjadi mudah diakses dan ada risiko intrusi oleh orang yang ingin mengakses komputer Anda. Ini mengancam atau menyerang

E-ISSN : 2827-9425
P-ISSN : 2828-4666

sistem komputer. Ini sangat berbahaya untuk sistem komputer perusahaan yang berisi data sensitif dan hanya dapat diakses oleh orang tertentu. Ancaman virus dan malware yang dapat merusak komputer, server atau jaringan komputer harus diantisipasi. Internet digunakan sebagai media sosialisasi, dimana perilaku memiliki dampak yang besar. Indonesia sendiri merupakan salah satu negara dengan jumlah serangan malware tertinggi di Asia Pasifik [1].

Serangan terhadap keamanan sistem informasi (security attacks) sering terjadi. Kejahatan komputer di dunia maya (cybercrime) dilakukan oleh kelompok dengan tujuan menembus keamanan sistem untuk menemukan, memperoleh, memodifikasi atau bahkan menghapus yang sudah ada. Informasi di sistem sangat membutuhkannya. Ada berbagai jenis serangan yang dapat dilakukan penyerang, termasuk pemadaman terjadi ketika informasi yang dikirimkan melalui jaringan dapat dihancurkan, terputus di tengah jalan, dan tidak dapat mencapai tujuannya. Serangan ini bertujuan untuk mendapatkan informasi [2]. Malware ini menyandang nama lengkap malware. Malware adalah istilah umum untuk setiap program atau perangkat lunak yang dirancang untuk menyusup atau merusak sistem komputer. Ada juga, beberapa pengguna internet tidak terbiasa dengan istilah malware. Secara rutin menyebut virus sebagai virus, dan media banyak menggunakan istilah itu, tetapi itu tidak sepenuhnya akurat. Program diklasifikasikan sebagai berbahaya karena tujuan pembuatannya, bukan karena properti khusus yang dimilikinya. Malware mencakup virus, worm, trojan horse, sebagian besar rootkit, spyware, adware, dan program berbahaya lainnya yang dapat membahayakan komputer. "Malware" adalah program komputer yang dibuat untuk maksud dan tujuan tertentu penciptanya, mencari kerentanan dalam program itu. Malware biasanya dibuat untuk memperkenalkan atau merusak perangkat lunak atau sistem operasi [3].

Berdasarkan permasalahan tersebut perlu dilakukan klasifikasi malware yang datanya diambil dari dataset malware agar dapat memudahkan dalam mempelajari dan membedakan jenis malware. Proses terdiri dari pengumpulan dataset, pre proseccing, melatih machine learning dan melakukan pengujian performa model atau kinerja. Data yang di ambil merupakan data public atau dataset malware dan di proses menggunakan machine learning dengan algoritma random forest. Parameter pengujian yang diukur berupa kecepatan dan akurasi dari machine learning dengan algoritma random forest dalam melakukan klasifikasi malware

Dengan demikian, berdasarkan kajian di atas maka peneliti bermaksud untuk melakukan penelitian ini diharapkan dapat memberikan sebuah solusi karena ingin mengetahui deteksi malware dengan menggunakan teknik machine learning Berdasarkan berbagai pemaparan di atas, dapat disimpulkan peneliti tertarik untuk mengangkat sebuah penelitian dengan judul: “Klasifikasi Malware menggunakan teknik Machine Learning”. Penelitian ini diharapkan membantu pengguna komputer dalam mengatasi Malware. Penelitian ini bertujuan untuk Untuk mengetahui performa atau kinerja Machine Learning menggunakan algoritma random forest untuk klasifikasi malware. Sehingga dapat mengetahui yang mana Malware yang mana bukan Malware dan dapat melakukan pencegahan jika terjadi serangan Malware.

2. TINJAUAN PUSTAKA

2.1 Keamanan Jaringan

Keamanan jaringan merupakan suatu cara pengamanan jaringan agar dapat terhindar dari berbagai ancaman yang berasal dari jaringan luar dan bertujuan untuk merusak atau mencuri data. Oleh karena itu, Anda harus mengambil tindakan pencegahan untuk menghadapi ancaman ini. Pertahanan dapat diimplementasikan melalui firewall, deteksi melalui IDS (Intrusion Detection System) dan kombinasi keduanya melalui IPS (Intrusion Prevention System).

2.2. Definisi Malware

E-ISSN : 2827-9425
P-ISSN : 2828-4666

Malware merupakan perangkat lunak berbahaya yang diprogram untuk merusak atau mengakses sistem komputer tanpa sepengetahuan pemilik sistem. Virus, worm, Trojan horse, keyloggers, spyware, dan ransomware merupakan contoh malware yang paling umum digunakan. Istilah seperti "worm", "virus", dan "Trojan horse" digunakan untuk mengkategorikan malware yang menunjukkan perilaku jahat serupa [8].

2.3 Klasifikasi Malware

Malware dapat dibagi menjadi beberapa kelas dan kategori. Secara umum, mereka dikategorikan berdasarkan proses dan respons berdasarkan desain dan evolusi malware [8].

2. 4 Analisis Malware

Analisis malware merupakan dasar untuk mengatur informasi, Informasi ini dapat dikembangkan sebagai tanda tangan untuk mendeteksi infeksi malware. Tujuan akhir dari analisis malware merupakan untuk menjelaskan dengan tepat bagaimana malware bekerja, menurut Adenansi & Novarina, 2017. Ada tiga teknik yang dapat dilakukan: analisis statis, dinamis, dan hybrid.

2. 5 Definisi Machine Learning

Machine Learning, juga dikenal sebagai pembelajaran mesin, merupakan ilmu komputer yang bekerja tanpa diprogram secara eksplisit. Banyak peneliti berpikir tentang bagaimana membuat kemajuan menuju AI pada skala manusia. Pembelajaran mesin merupakan kecerdasan buatan yang mempelajari cara membuat data. Pembelajaran mesin biasa disingkat ML. Hal ini diperlukan untuk menerapkan teknik yang cepat dan ampuh untuk menemukan masalah baru.

2. 6 Random Forest

Random Forest merupakan metode bagging yang, ketika membangun pohon selama pelatihan, menghasilkan pohon dalam jumlah besar dari data sampel secara independen dari pohon sebelumnya dan membuat keputusan berdasarkan suara terbanyak.

2. 7 Confusion Matrix

Confusion matrix merupakan sebuah metode yang dapat digunakan untuk mengukur kinerja dari sebuah metode klasifikasi. Pada dasarnya, Confusion Matrix berisi informasi yang membandingkan hasil klasifikasi yang dilakukan oleh sistem dengan hasil klasifikasi sebagaimana mestinya. Saat mengukur kinerja menggunakan Confusion matrix, ada empat istilah yang menjelaskan hasil dari proses klasifikasi. Keempat suku tersebut merupakan true positive (TP), true negative (TN), false positive (FP), dan false negative (FN). Nilai True Negative (TN) merupakan jumlah data negatif yang dikenali dengan benar. Positif palsu (FP), di sisi lain, mengenali data negatif sebagai data positif. Confusion matrix merupakan metode yang digunakan untuk mengukur atau melakukan perhitungan akurasi untuk konsep data mining. Sebuah Confusion matrix terdiri dari kumpulan data uji yang diprediksi benar atau salah oleh model klasifikasi.

3. METODE PENELITIAN

3.1 Jenis, Metode, Subjek,Objek, Waktu dan Lokasi Penelitian

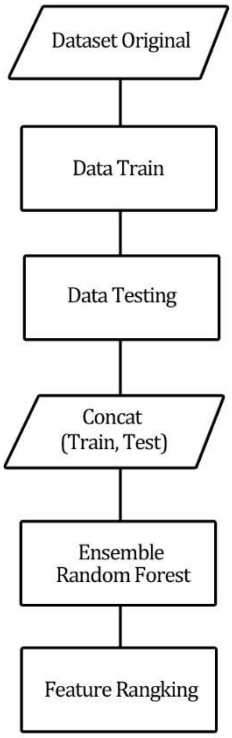
Pada Penelitian ini Objek yang dilakukan penelitian ini ialah mendeteksi Malware menggunakan Machine Learning dengan metode Algoritma Random Forest, maka penelitian ini merupakan penelitian kualitatif. Maka jenis dari penelitian ini merupakan penelitian model uji coba metode – eksperimental.

3.2 Pengumpulan Data

Dalam memulai sebuah penelitian tentunya kita memerlukan sebuah data sebagai pondasi masalah untuk diolah agar mencapai tujuan penelitian yang sukses. dataset yang diambil dalam penelitian ini diperoleh dari internet yaitu pada website https://www.kaggle.com/datasets/amauricio/pe-files-malwares. Terdiri dari 19611 baris dan 79 kolom.

E-ISSN : 2827-9425
P-ISSN : 2828-4666

3.3 Pemodelan



Gambar 1. Flowchart Pemodelan Data

3.4 Machine Learning

Cara kerja machine learning dalam praktik bergantung pada teknik dan metode pembelajaran apa yang Anda gunakan dengan ML. Namun pada dasarnya, prinsip kerja machine learning merupakan sama: pengumpulan data, eksplorasi data, pemilihan model atau metode, pemberian pelatihan untuk model yang dipilih, dan evaluasi hasil ML.

3.5 Pra Pengolahan Data

Preprocessing adalah bagian selanjutnya dari sebuah kegiatan pengumpulan dataset yang diperoleh dalam bentuk spreadsheet berformat excel maka dari itu belum dapat langsung diolah atau di gunakan karena dataset masih dalam bentuk mentah tidak teratur bahkan hanya di pisahkan tanda koma. data yang diambil merupakan DATASET MALWARE yang bersumber dari kaggle sehingga dijadikan sebagai dataset dengan melalui proses tahapan pengolahan data.

3.5.1. Analysis/Model Malware

Menyajikan taksonomi tentang bagaimana pembelajaran mesin digunakan untuk analisis malware. Identifikasi tiga dimensi utama yang memudahkan pengorganisasian pekerjaan yang diteliti. Yang pertama mencirikan tujuan akhir dari analisis, seperti deteksi malware. Dimensi kedua menjelaskan fitur yang mendasari analisis dalam hal bagaimana mereka diekstraksi, misalnya dengan analisis dinamis, dan fitur apa yang diperhitungkan, misalnya register CPU. Terakhir, dimensi ketiga menentukan jenis algoritme pembelajaran mesin (mis. pembelajaran terawasi) yang digunakan untuk analisis.

3.5.2 Pengujian Algoritma Random Forest

Algoritma Random Forest (RF) merupakan perluasan dari metode klasifikasi dan pohon regresi (CART) dengan menerapkan bootstrap aggregation (bagging) dan metode pemilihan fitur secara acak (Breiman 2001). Algoritma RF merupakan algoritma yang baik untuk mengklasifikasikan data dalam jumlah besar, dan algoritma RF tidak memiliki pemangkasan atau pemangkasan variabel seperti algoritma pohon keputusan. Metode RF menggabungkan banyak pohon (tree) untuk membuat kelas klasifikasi dan prediksi,

E-ISSN : 2827-9425
P-ISSN : 2828-4666

berbeda dengan pohon yang hanya terdiri dari satu pohon. Dalam RF, pembangunan pohon dilakukan dengan melatih sampel data. Sampling-by-penggantian merupakan metode yang digunakan untuk sampel data. Pemilihan variabel yang digunakan untuk pemisahan merupakan acak. Setelah semua pohon terbentuk, dilakukan klasifikasi. Keputusan klasifikasi dalam RF ini didasarkan pada voting dari masing-masing pohon, dengan voting terbanyak menjadi pemenangnya.

3.5.3 Evaluasi Confusion Matrix

Untuk mengevaluasi hasil algoritma klasifikasi, teknik evaluasi model algoritma klasifikasi yang digunakan dalam penelitian ini merupakan fusion matrix. Matriks konfusi merupakan cara paling sederhana untuk menilai hasil kinerja algoritma pengklasifikasi dengan membandingkan contoh positif yang diklasifikasikan sebagai benar atau salah dengan contoh negatif yang diklasifikasikan sebagai benar atau salah. Ada banyak pandangan berbeda tentang matriks kebingungan, dan matriks kebingungan memainkan peran mendasar dalam mengevaluasi kinerja algoritma klasifikasi. Dalam matriks kebingungan, baris mewakili label yang benar dan kolom mewakili prediksi algoritma pengklasifikasi.

3.5.4 Pengujian

Pada tahap pengujian akan dilakukan pengumpulan data terlebih dahulu, lalu mengumpulkan alat dan bahan yang akan digunakan. Kemudian mendeteksi malware pada laptop yang sudah disiapkan lalu melakukan pengujian malware menggunakan teknik machine learning dengan metode algoritma random forest, sehingga dapat mengetahui apakah teknik ini dapat berguna untuk mendeteksi malware yang ada pada setiap laptop dan bisa mengatasi terjadinya malware pada laptop.

4. HASIL DAN PEMBAHASAN

4.1. Hasil Pengumpulan Data

Hasil pengumpulan data ini penulis menggunakan data public https://www.kaggle.com/datasets/amauricio/pe-files-malwares kemudian dataset tersebut akan diproses dan di klasifikasi berdasarkan pemodelan yang telah diolah oleh peneliti. Pengumpulan data berasal dari perangkat/alat dari hasil penelitian orang lain dataset public, dataset public adalah data yang sudah ada yang digunakan para peneliti sebelumnya. Malware yang menyusup di browser. Dalam satu baris ada satu serangan yang terjadi, dan setiap serangan malware pasti akan meninggalkan jejak yang disebut loc.

<class 'pandas.core.frame.DataFrame'>
Dataset yang digunakan pada contoh ini adalah dataset malware, tahapan-tahapan yang akan dilakukan adalah analisis deskriptif dan penanganan data , pembagian data latih dan data uji, pemodelan dengan Random Forest serta evaluasi model.

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 19611 entries, 0 to 19610
Data columns (total 79 columns):

#	Column	Non-Null Count	Dtype
0	Name	19611 non-null	object
1	e_magic	19611 non-null	int64
2	e_cblp	19611 non-null	int64
3	e_cp	19611 non-null	int64
4	e_crlc	19611 non-null	int64
5	e_cparhdr	19611 non-null	int64
6	e_minalloc	19611 non-null	int64
7	e_maxalloc	19611 non-null	int64
8	e_ss	19611 non-null	int64
9	e_sp	19611 non-null	int64
10	e_csum	19611 non-null	int64

E-ISSN : 2827-9425
P-ISSN : 2828-4666

```
11 e_ip                     19611 non-null  int64
12 e_cs                     19611 non-null  int64
13 e_lfarlc                 19611 non-null  int64
14 e_ovno                   19611 non-null  int64
15 e_oemid                  19611 non-null  int64
16 e_oeminfo                19611 non-null  int64
17 e_lfanew                 19611 non-null  int64
18 Machine                  19611 non-null  int64
19 NumberOfSections         19611 non-null  int64
20 TimeDateStamp            19611 non-null  int64
21 PointerToSymbolTable     19611 non-null  int64
22 NumberOfSymbols          19611 non-null  int64
23 SizeOfOptionalHeader     19611 non-null  int64
24 Characteristics          19611 non-null  int64
25 Magic                    19611 non-null  int64
26 MajorLinkerVersion       19611 non-null  int64
27 MinorLinkerVersion       19611 non-null  int64
28 SizeOfCode               19611 non-null  int64
29 SizeOfInitializedData    19611 non-null  int64
30 SizeOfUninitializedData  19611 non-null  int64
31 AddressOfEntryPoint      19611 non-null  int64
32 BaseOfCode               19611 non-null  int64
33 ImageBase                19611 non-null  int64
34 SectionAlignment         19611 non-null  int64
35 FileAlignment            19611 non-null  int64
36 MajorOperatingSystemVersion  19611 non-null  int64
37 MinorOperatingSystemVersion  19611 non-null  int64
38 MajorImageVersion        19611 non-null  int64
39 MinorImageVersion        19611 non-null  int64
40 MajorSubsystemVersion    19611 non-null  int64
41 MinorSubsystemVersion    19611 non-null  int64
42 SizeOfHeaders            19611 non-null  int64
43 CheckSum                 19611 non-null  int64
44 SizeOfImage              19611 non-null  int64
45 Subsystem                19611 non-null  int64
46 DllCharacteristics       19611 non-null  int64
47 SizeOfStackReserve       19611 non-null  int64
48 SizeOfStackCommit        19611 non-null  int64
49 SizeOfHeapReserve        19611 non-null  int64
50 SizeOfHeapCommit         19611 non-null  int64
51 LoaderFlags              19611 non-null  int64
52 NumberOfRvaAndSizes      19611 non-null  int64
53 Malware                  19611 non-null  int64
54 SuspiciousImportFunctions  19611 non-null  int64
55 SuspiciousNameSection    19611 non-null  int64
56 SectionsLength           19611 non-null  int64
57 SectionMinEntropy        19611 non-null  float64
58 SectionMaxEntropy        19611 non-null  int64
59 SectionMinRawsize        19611 non-null  int64
60 SectionMaxRawsize        19611 non-null  int64
61 SectionMinVirtualsize    19611 non-null  int64
62 SectionMaxVirtualsize    19611 non-null  int64
63 SectionMaxPhysical       19611 non-null  int64
64 SectionMinPhysical       19611 non-null  int64
65 SectionMaxVirtual        19611 non-null  int64
66 SectionMinVirtual        19611 non-null  int64
67 SectionMaxPointerData    19611 non-null  int64
68 SectionMinPointerData    19611 non-null  int64
69 SectionMaxChar           19611 non-null  int64
```

E-ISSN : 2827-9425
P-ISSN : 2828-4666

```
 70  SectionMainChar              19611 non-null  int64
 71  DirectoryEntryImport         19611 non-null  int64
 72  DirectoryEntryImportSize     19611 non-null  int64
 73  DirectoryEntryExport         19611 non-null  int64
 74  ImageDirectoryEntryExport    19611 non-null  int64
 75  ImageDirectoryEntryImport    19611 non-null  int64
 76  ImageDirectoryEntryResource  19611 non-null  int64
 77  ImageDirectoryEntryException 19611 non-null  int64
 78  ImageDirectoryEntrySecurity  19611 non-null  int64
dtypes: float64(1), int64(77), object(1)
memory usage: 11.8+ MB
```

Dataset malware terdiri dari 19611 baris. Terdapat 79 kolom, dimana seluruhnya bertipe numerik, Berikut penjelasan variabel nama kolom.

4.2. **Pra Pemrosesan Data**

Pada proses ini pra pemrosesan data dilakukan dengan menginstal beberapa library python. Pandas adalah library python open source yang biasanya digunakan untuk kebutuhan data analisis. Pandas membuat python supaya dapat bekerja dengan data yang berbentuk tabular seperti spreadsheet dengan cara pemuatan data yang cepat, manipulasi data, mengggabungkan data, serta ada berbagai fungsi yang lain. Untuk memanggil library pandas, library numpy juga ikut dipanggil. Selanjutnya import pickle untuk implementasi protocol biner untuk serializing dan de-serializing dari struktur objek pada python. Selanjutnya import seaborn untuk melakukan visualisasi data. Selanjutnya import matplotlib.pyplot as plt untuk kumpulan fungsi yang membuat beberapa perubahan pada gambar, membuat area plot pada gambar, menambah label di plot dan lainnya.
Eksplorasi Data adalah bagian penting dariproses data science, yaitu berupa proses menganalisis sekumpulan data untuk meringkas karakteristik utamanya agar pengguna lebih memahami dataset yang akan digunakan.

Head() digunakan untuk menampilkan data awal atau data teratas pada dataframe. Default-nya jika kita tidak memberikan argument di dalam tanda kurung (), data yang akan ditampilkan adalah 5 baris data teratas

```
data = pd.read_csv("dataset_malwares.csv")
data.head()
```

✓ 2.5s Python

	Name	e_magic	e_cblp	e_cp	e_crlc	e_cparhdr	e_minalloc	e_maxalloc	e_ss	e_sp	...	SectionMaxChar	SectionMainChar	Directo
0	VirusShare_a878ba26000edaac5c98eff4432723b3	23117	144	3	0	4	0	65535	0	184	...	3758096608	0	
1	VirusShare_ef9130570fddc174b312b2047f5f4cf0	23117	144	3	0	4	0	65535	0	184	...	3791650880	0	
2	VirusShare_ef84cdeba22be72a69b198213dada81a	23117	144	3	0	4	0	65535	0	184	...	3221225536	0	
3	VirusShare_6bf3608e60ebc16cbcff6ed5467d469e	23117	144	3	0	4	0	65535	0	184	...	3224371328	0	
4	VirusShare_2cc94d952b2efb13c7d6bbe0dd59d3fb	23117	144	3	0	4	0	65535	0	184	...	3227516992	0	

5 rows × 79 columns

Gambar 2. Hasil Eksplorasi Data

Data.loc merupakan salah satu cara yang efektif untuk memilih baris dan kolom pada dataframe sesuai dengan nama index baris atau kolom. Seleksi kolom yang kita gunakan terdiri dari “Nama”,”Machine”,”TimeDateStamp”,”Malware”.

E-ISSN : 2827-9425
P-ISSN : 2828-4666

	Name	Machine	TimeDateStamp	Malware
0	VirusShare_a878ba26000edaac5c98eff4432723b3	34404	1236512358	1
1	VirusShare_ef9130570fddc174b312b2047f5f4cf0	332	1365109591	1
2	VirusShare_ef84cdeba22be72a69b198213dada81a	332	1438777028	1
3	VirusShare_6bf3608e60ebc16cbcff6ed5467d469e	332	1354629311	1
4	VirusShare_2cc94d952b2efb13c7d6bbe0dd59d3fb	332	1386631250	1
...	...	...	...	...
19606	clip.exe	332	1377143713	0
19607	VNC-Server-6.2.0-Windows.exe	332	1501777476	0
19608	Microsoft.GroupPolicy.Management.ni.dll	332	1377135839	0
19609	cryptuiwizard.dll	332	1377141725	0
19610	winhttp.dll	332	1377139145	0

19611 rows × 4 columns

Gambar 3. Hasil data loc

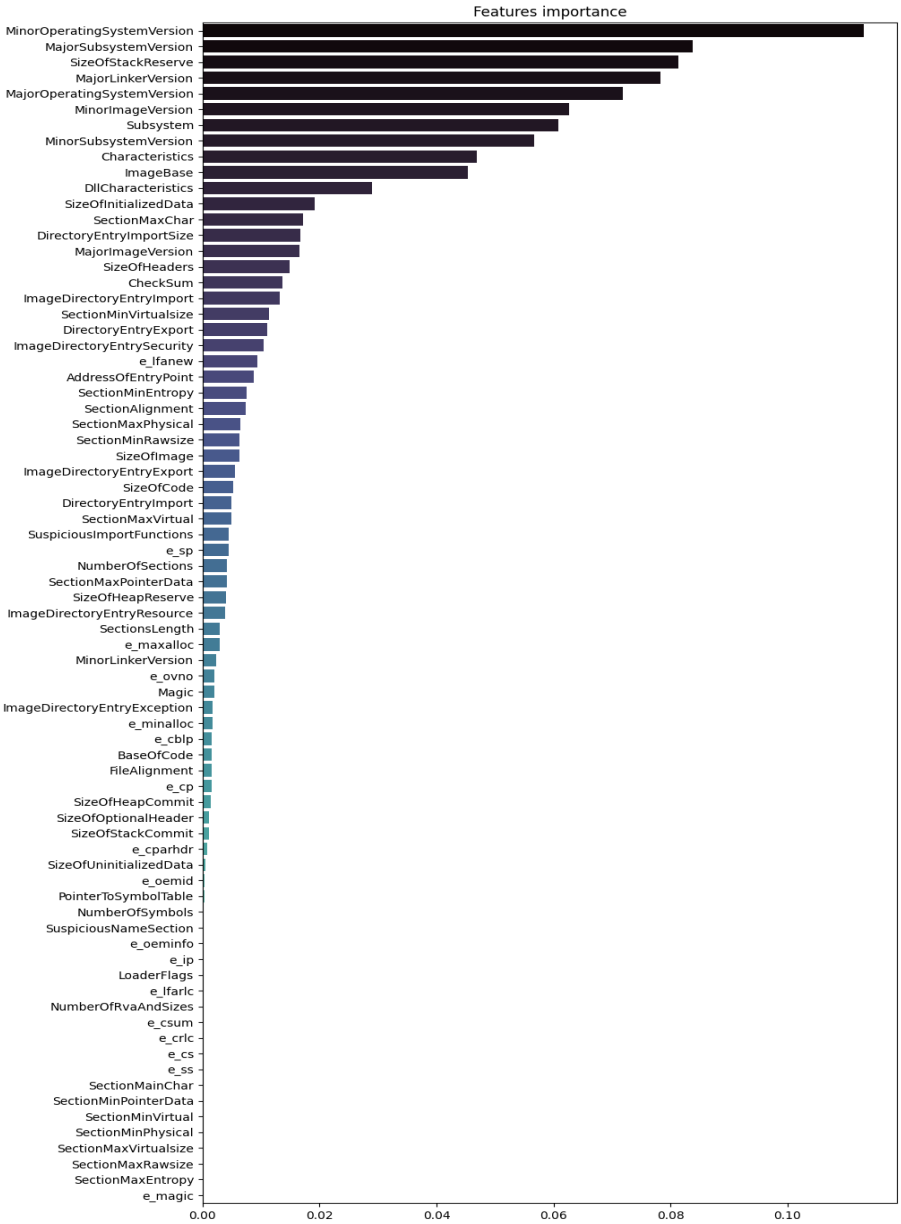
data.columns

```
Index(['Name', 'e_magic', 'e_cblp', 'e_cp', 'e_crlc', 'e_cparhdr',
       'e_minalloc', 'e_maxalloc', 'e_ss', 'e_sp', 'e_csum', 'e_ip', 'e_cs',
       'e_lfarlc', 'e_ovno', 'e_oemid', 'e_oeminfo', 'e_lfanew', 'Machine',
       'NumberOfSections', 'TimeDateStamp', 'PointerToSymbolTable',
       'NumberOfSymbols', 'SizeOfOptionalHeader', 'Characteristics', 'Magic',
       'MajorLinkerVersion', 'MinorLinkerVersion', 'SizeOfCode',
       'SizeOfInitializedData', 'SizeOfUninitializedData',
       'AddressOfEntryPoint', 'BaseOfCode', 'ImageBase', 'SectionAlignment',
       'FileAlignment', 'MajorOperatingSystemVersion',
       'MinorOperatingSystemVersion', 'MajorImageVersion', 'MinorImageVersion',
       'MajorSubsystemVersion', 'MinorSubsystemVersion', 'SizeOfHeaders',
       'CheckSum', 'SizeOfImage', 'Subsystem', 'DllCharacteristics',
       'SizeOfStackReserve', 'SizeOfStackCommit', 'SizeOfHeapReserve',
       'SizeOfHeapCommit', 'LoaderFlags', 'NumberOfRvaAndSizes', 'Malware',
       'SuspiciousImportFunctions', 'SuspiciousNameSection', 'SectionsLength',
       'SectionMinEntropy', 'SectionMaxEntropy', 'SectionMinRawsize',
       'SectionMaxRawsize', 'SectionMinVirtualsize', 'SectionMaxVirtualsize',
       'SectionMaxPhysical', 'SectionMinPhysical', 'SectionMaxVirtual',
       'SectionMinVirtual', 'SectionMaxPointerData', 'SectionMinPointerData',
       'SectionMaxChar', 'SectionMainChar', 'DirectoryEntryImport',
       'DirectoryEntryImportSize', 'DirectoryEntryExport',
       'ImageDirectoryEntryExport', 'ImageDirectoryEntryImport',
       'ImageDirectoryEntryResource', 'ImageDirectoryEntryException',
       'ImageDirectoryEntrySecurity'],
      dtype='object')
```

Gambar 4. Hasil Data Kolom

Berikut seluruh daftar nama kolom 79 data column yang ada pada dataset malware ini. Data.columns digunakan untuk menampilkan nama nama kolom pada dataframe. Hasil dari tahap Feature Importance merupakan sekumpulan fitur beserta ukuran tingkat kepentingannya. Setelah pentingnya fitur ditentukan, fitur dapat dipilih dengan tepat. Berikut merupakan deretan hasil feature importance dari yang terpenting sampai terendah.

E-ISSN : 2827-9425
P-ISSN : 2828-4666



Gambar 5. Hasil Feature Importance

Dengan informasi tersebut, kita dapat memahami dengan lebih baik bagaimana fitur-fitur pada model Random Forest Classifier berpengaruh dalam memprediksi hasilnya. Plot tersebut dapat membantu dalam mengevaluasi kualitas model dan memberikan insight pada proses feature selection atau reduksi dimensi yang mungkin diperlukan paca model tersebut. Dari plot diatas kita dapat melihat bahwa fitur minoroperatingsystemversion yang paling membantu model kita untuk membedakan malware dan bukan malware.

4.3. Pembahasan Model

Pada analisis ini dibangun model pembelajaran mesin menggunakan random forest clasifir untuk mengklasifikasikan malware. Model dilatih pada dataset dengan berbagai fitur dan hasilnya menunjukan akurasi yang tinggi sebesar 99%. Laporan confusion matrix dan klasifikasi juga menunjukkan presisi tinggi, daya ingat, dan score f1. Fitur importance fitur menunjukkan pentingnya setiap fitur dalam klasifikasi malware. Model terakhir yang dilatih disimpan menggunakan perpustakaan acar. Analisis ini menyoroti potensi penggunaan algoritme pembeljaran mesin dalam klasifikasi malware.

E-ISSN : 2827-9425
P-ISSN : 2828-4666

4.4. Klasifikasi Menggunakan Random Forest

Selanjutnya fungsi classification_report digunakan untuk menghasilkan evaluasi komprehensif dari kinerja model pada data uji. Variabel y_test dan y_pred masing-masing adalah nilai target aktual dan nilai prediksi. Parameter target_names digunakan untuk menentukkan label kelas untuk variabel target. Dalam kasus ini, target variabel memiliki dua kelas: “Bukan Malware” dan “Malware”.

Hasil prediksi RandomForestClassifier untuk membuat prediksi guna menguji keakuratan model

```
y_pred = rfc.predict(X_test)

print(classification_report(y_test, y_pred, target_names=['Benign', 'Malware']))
```

Tabel 1. Hasil Akurasi Random Forest

	Precision	Recall	F1-Score	Support
Benign	0,99%	96%	97%	1004
Malware	0,99%	100%	99%	2919
Accuracy			99%	3923
Marco avg	0,99%	98%	98%	3923
Weighted avg	99%	99%	99%	3923

.

Ini menampilkan skor presisi, recall, f1. Pada eksperimen yang telah dilakukan terdapat beberapa factor yang menjadi indikator berjalannya tahap pembelajaran dan klasifikasi untuk mendapatkan akurasi model, balancing terhadap dataset dan visualisasi dataset malware. Kemudian ini hasil accuracy nya 99%, sebagai bentuk evaluasi terhadap model random forest.

5. KESIMPULAN

Model random forest memberikan hasil yang sangat baik tanpa preprocessing pada data. Hasilnya bagus meskipun datanya tidak seimbang. Tidak perlu menggunakan teknik apapun untuk menyeimbangkannya. Penskalaan/skaling tidak perlu dilakukan, model random forest adalah model partisi rekursif yang bergantung pada partisi data karena ia bekerja pada pemisahan nilai fitur dan tidak melakukan perhitungan di dalamnya. Hasil penelitian menunjukkan bahwa model memiliki presisi 0,99 untuk kelas ”Bukan Malware” dan “Malware”, recall 0,96 untuk “Bukan Malware” dan 1,00 untuk “Malware”, serta f1-score 0,98 untuk “Bukan Malware” dan 0,99 untuk “Malware”. Akurasi modelnya adalah 0,99 yang cukup baik rata-rata tertimbang dari presisi,recall, dan f1-score juga 0,99.

Analisis ini dibangun model pembelajaran mesin menggunakan random forest clasifir untuk mengklasifikasikan malware. Model dilatih pada dataset dengan berbagai fitur dan hasilnya menunjukan akurasi yang tinggi sebesar 99%. Laporan confusionmatrix dan klasifikasi juga menunjukkan presisi tinggi, daya ingat, dan score f1. Plot kepentinga fitur menunjukkan pentingnya setiap fitur dalam klasifikasi malware. Model terakhir yang dilatih disimpan menggunakan perpustakaan acar. Analisis ini menyoroti potensi penggunaan algoritme pembeljaran mesin dalam klasifikasi malware.

Setelah penelitian ini selesai, peneliti memberikan beberapa saran untuk dilakukan pengembangan penelitian selanjutnya:

1. Pada penelitian selanjutnya dapat mempertimbangkan dataset apa yang akan di klasifikasikan atau di olah, juga dapat menambahkan metode atau algoritma yang lain sebagai pembanding.
2. Pada penelitian selanjutnya juga dapat mengembangkan penelitian serupa dengan melakukan pembangunan sistem.

DAFTAR PUSTAKA

E-ISSN : 2827-9425
P-ISSN : 2828-4666

[1] Togu Novriansyah Turnip, Chatrine Febriyanti Manurung, Yogi Septian Lubis 2023 "Klasifikasi Malware Android Aplikasi Menggunakan Random Forest Berdasarkan Fitur Statik," Jurnal Teknik Informatika dan Sistem Informasi, Vol. 10, No.1, Maret 2023.

[2] Edward Tansen, Deris Wahyu Nurdianto "Program studi teknik komputer, Universitas amikom yogyakarta", Jurnal Teknologi Informasi Vol.4, No.2, Desember 2020.

[3] Triawan Adi Cahyono, Victor Wahanggara, Darmawan Ramadana "Analisis dan Deteksi Malware Menggunakan Metode Malware Analisis Dinamis dan Malware Analisis Statis," Jurnal Sistem & Teknologi Informasi Indonesia, Vol. 2, No. 1, Februari 2017

[4] Zalavadiya & Priyanka, Klasifikasi Malware: http://repository.uin-suska.ac.id/16332/8/7.%20BAB%20II%20LANDASAN%20TEORI.pdf.

[5] Raden Budiarto Hadiprakoso, Wahyu Rendra Aditya, Febriora Novia Pramitha," Analisis Statis Deteksi Malware Android Menggunakan Algoritma Supervised Machine Learning,"CyberSecurity dan forensik Digital, Vol. 5, No. 1, Mei 2022.

[6] Henny Wahyu Sulistyo, Hardian Oktavianto "Analisis klasifikasi Kanker Payudara Menggunakan Algoritma Random Forest,"Jurnal Informatika Vol.8 No.2 Februari 2020

[7] Yitsak Wanli Sitorus, Parman Sukarno, Satria Mandala, "Analisis Deteksi Malware Android menggunakan Metode Support Vector Machine & Random Forest,"Jurnal e-Proceeding of Engineering, Vol. 8, No. 6,, Desember 2021.

[8] Devi Rizky Septani, Nur Widiyasono, Husni Mubarok, "Investigasi Serangan Malware Njrat Pada PC," Jurnal Edukasi dan Penelitian Informatika (JEPIN), Vol. 2, No. 2, 2016.

[9] Robi Aziz Zuama, Syaifur Rahmatullah, Yuri Yulian "Analisis Performa Algoritma Machine Learning pada Prediksi Penyakit Cerebrovascular Accidents," Jurnal Media Informatika Budidarma, Vol. 6, No. 1, Januari 2022

[10] Fikri Bahriar, Nur Widyasono, Aldy Putra Aldya, "Memory Volatile Forensik Untuk Deteksi Malware Menggunakan Algoritma Machine Learning,"Jurnal Teknik Informatika dengan Sistem Informasi, Vol. 4, No. 2, Agustus 2018.

[11] Syafrial Fachri Pane, Jenly Ramdan, "Pemodelan Machine Learning : Analisis Sentimen Masyarakat Terhadap Kebijakan PPKM Menggunakan Data Twitter,"Jurnal Sistem Cerdas (2022), Vol. 5, No. 1 eISSN : 2622-8254.

[12] Fajar Mu Alim, Rahmi Hidayati, "Implementasi Metode Random Forest Untuk Penjurusan Siswa Di Madrasah Aliyah Negeri Sintang,"Jurnal Jupiter, Vol. 14 No. 1 Bulan April, Tahun 2022.

[13] http://library.binus.ac.id/eColls/eThesisdoc/Bab2/2010-1-00247-IF%20BAB%202.pdf

[14] Karsito, Santi Susanti,"Klasifikasi Kelayakan Peserta Pengajuan Kredit Rumah Dengan Algoritma Naïve Bayes Di Perumahan Azzura Residencia,"Jurnal Teknologi Pelita Bangsa, Vol. 9, No. 3, Maret 2019 ISSN : 2407-3903