BAB I
PENDAHULUAN

1.1 Latar Belakang Masalah

Pada era industri 4.0 yang semakin berkembang pada saat ini. Dengan banyaknya aplikasi yang diciptakan untuk menyelesaikan pekerjaan pada penggunanya, aplikasi menjadi salah satu media penyebaran malware untuk menyerang targetnya[1]. Penyebaran malware saat ini begitu mudah dilakukan melalui berbagai media seperti *website, software*, iklan-iklan tertentu, maupun dari fisik seperti *usb flashdisk*.

Malware adalah *software* yang diciptakan untuk menyerang dan merusak sistem operasi pada sebuah komputer. Ada beberapa jenis *malware* yang terdiri dari *spyware*, bot, *rootkit*, *trojan*, dan *virus* yang dirancang untuk melakukan tugas-tugas seperti mengganggu kinerja sistem operasi pada sebuah komputer[2].Oleh karena itu, *malware* menjadi ancaman keamanan sangat penting di saat ini [3]. Dalam beberapa tahun terakhir, banyak malware muncul dengan berbagai jenis yang sangat produktif yaitu dengan mengambil alih kendali komputer yang ada di seluruh dunia untuk waktu yang lama. Menurut data pada perusahaan *Symantec*, varian *malware* yang diperuntukkan untuk mobile meningkat 54% pada tahun 2017 sementara untuk perangkat Iot meningkat 600% [4].

Untuk menemukan adanya indikasi serangan malware, pada penelitian terdahulu yang dijadikan referensi sebagai perbandingan dan acuan pada penelitian ini menggunakan berbagai metode analisis *malware* yaitu statis, dinamis, dan kombinasi dalam analisis *malware*. Pada analisis statis, *malware* tidak akan dijalankan atau diaktifkan melainkan diteliti serta menganalisis terhadap *source code* yang tertulis pada *malware* tersebut. Tidak seperti metode analisis statis, pada analisis dinamis malware yang diteliti akan diaktifkan dalam sebuah wadah atau mesin virtual untuk melihat tingkah laku dari malware. Dari proses tersebut pengumpulan informasi mengenai tingkah laku dari sebuah *malware*. Tahapan analisis melibatkan keseluruhan sistem pada komputer atau mesin virtual dengan melihat perubahan pada registry. Sedangkan analisis kombinasi, penggabungan antara dua metode yaitu dinamis dan statis[5].

Berbagai metode yang digunakan dalam melacak keberadaan malware pada sebuah komputer telah banyak bermunculan, salah satunya dengan menggunakan *artificial intelligence* (AI). Beberapa tahun belakangan ini, *artificial intelgent*(AI) menjadi pilihan dalam melakukan analisis di beberapa bidang termasuk keamanan siber salah satunya menggunakan teknik *machine learning* (ML) [6]. *Machine Learning* (ML) adalah suatu bidang studi atau bidang keilmuan yang mencakup perencanaan dan pengembangan algoritma yang memampukan sebuah komputer dapat mengembangkan perilaku berdasarkan data-data yang diterimanya[7]. Pada *machine learning* (ML) terdapat tiga algortima, yaitu *unsupervised learning*, *supervised learning*, dan *reinforcement learning*. *Unsupervised learning* adalah algoritma yang tidak menggunakan data label atau tidak terstruktur dan mampu belajar dari data dengan menemukan pola implisit. *Reinforcement learning* adalah algoritma yang menggunakan kinerja mesin untuk berinteraksi langsung dengan objek secara dinamis untuk mencapai tujuannya. Sedangkan *supervised learning* adalah algoritma yang menggunakan data label atau terstruktur dan mampu mengidentifikasi fitur secara eksplisit[8].

Pada penelitian ini,peneliti menggunakan dua algoritma dari *supervised learning* yaitu *Decision Tree* (DT) dan *Random Forests*(RF) yang nantinya mendukung dalam mengklasifikasi dan memprediksi adanya serangan malware. *Decision Tree* (DT) adalah model analisis dengan menggunakan struktur pohon atau struktur hirarki[9]. *Random Forest* (RF) adalah suatu algoritma yang menggabungkan pohon-pohon (trees) dengan melakukan pelatihan pada sampel data yang ada[10]. Tujuan dari penelitian ini untuk memprediksi tingkat *accuracy* , *recall*, *F score* dan *confusion matrix*yang akan menyimpulkan *True Positive Rate (TPR)* pada dua model algoritma yang dipilih untuk mengklasifikasi malware. Hasil dari penelitian ini yaitu menjadi acuan dalam pemilihan algoritma yang tepat dalam penelitian-penelitian selanjutnya.

1.2 Rumusan Masalah

Pada rumusan masalah yang akan diteliti pada penelitian ini. Berdasarkan latar belakang diatas, rumusan masalah dalam penelitian ini yaitu :

1 Bagaimana menganalisis serangan *malware* dengan menggunakan metode klasifikasi dengan menggunakan *Random Forest* dan *Decision Tree*?

2 Bagaimana hasil perbandingan pada algoritma *Decision Tree* dan *Random Forest* dalam mengklasifikasi terhadap serangan *malware* ?

1.3 Batasan Masalah

Agar penelitian lebih terarah dan sesuai dengan rumusan masalah yang telah dipaparkan sebelumnya, peneliti membuat batasan masalah. Batasan masalah yang ditetapkan dalam penelitian adalah sebagai berikut:

a. Pelaksanaan penelitian ini untuk menganalisa menggunakan algoritma pada *supervised learning* yaitu *Random Forest* dan *Decision Tree* untuk mengklasifikasi dan memprediksi adanya serangan malware.

b. Pemilihan fitur pada dataset menggunakan metode feature selection berdasarkan score tertinggi dari keseluruhan fitur.

c. Data yang digunakan berasal dari website databases yaitu Kaggle.com yang bersifat *open source*.

d. Data malware yang digunakan berasal dari hasil ekstraksi beberapa sampel malware X dan Y.

e. Data dibagi menjadi 2 jenis yaitu data training 80% dan data testing 20%.

f. Bahasa pemrograman yang digunakan untuk membuat model Machine Learning penelitian adalah menggunakan bahasa Python versi 3.9.

1.4 Tujuan Penelitian

Berdasarkan rumusan masalah sebelumnya, tujuan yang ingin dicapai dari penelitian ini adalah :

a. Mengimplementasikan dua algoritma pada *machine learning* (ML) dalam melakukan klasifikasi menghadapi serangan malware.

b. Mencari dan menemukan hasil terbaik dari implementasi *decision tree* dan *random forest* menghadapi serangan *malware*.
c. Memperkaya khasanah penelitian selanjutnya dalam menentukan model *machine learning* terbaik untuk menangkal serangan *malware*.

1.5 Manfaat Penelitian

Manfaat yang diharapkan pada penelitian ini berdasarkan latar belakang , rumusan masalah , batasan masalah dan tujuan penelitian adalah sebagai berikut :

a. Hasil penelitian ini diharapakan dapat memahami kelebihan dan kekurangan masing-masing algoritma yang digunakan dalam memproses data dengan banyak fitur.
b. Hasil dari penelitian ini dapat membantu dalam mengembangkan teknik baru dalam feature selection yang bertujuan untuk memecahkan masalah klasifikasi.

1.6 Sistematika Penulisan

Sistematik penulisan berisikan garis besar atau gambaran secara umum penelitian ini sehingga mempermudah pemahaman alur isi. Adapun garis besar isi skripsi ini adalah sebagai berikut :

Bab I Pendahuluan, tahapan ini merupakan bab awal yang menjelaskan tentang latar belakang , masalah penelitian, rumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian dan sistematika penyajian.

Bab II Landasan Teori, bab ini menjelaskan tinjauan kepustakaan dari penelitian-penelitian terkait yang membahas beberapa teori antara lain analisis *malware, machine learning (ml), Decision Tree (dt) , Random Forest (rf) , cross industry standard process for data mining (crisp-dm),* dan tool yang akan digunakan dalam proses analisis *malware*.

Bab III Metodologi Penelitian, bab ini berisikan gambaran umum tentang alur dari penelitian, prosedur, dan mekanisme metode analisis yang diterapkan pada penelitian.

Bab IV Pembahasan / Analisa, pada tahapan ini membahas implementasi dan analisa hasil dari yang ditemukan menggunakan metode . Bab ini juga menyampaikan pembahasan secara teknis dari hasil analisa.

Bab V Penutup, bab ini menjelaskan tahapan terakhir yang dilakukan peneliti dan memuat kesimpulan dan keseluruhan uraian dari bab-bab sebelumnya. Tahapan ini juga memaparkan kekurangan serta saran untuk pengembangan penelitian berikutnya.

Daftar Referensi, berisi referensi terkait dengan penelitian ini, baik melalui publikasi jurnal dan artikel situs yang dapat menunjang proses penelitian.