ISSN: 2962-3545
Prosiding SAINTEK: Sains dan Teknologi Vol.1 No.1 Tahun 2022
Call for papers dan Seminar Nasional Sains dan Teknologi Ke-1 2022
Fakultas Teknik, Universitas Pelita Bangsa, Juli 2022

Klasifikasi Aplikasi Malware Android Menggunakan Algoritma C5.0

Classification Of Android Malware Application Using C5.0 Algorithm

Muhammad Refhaldo[1], Eko Budiarto[2], Putri Anggun Sari[3], Sella Monica[4]
[1,2,3,4]Program Studi Teknik Informatika, Fakultas Teknik, Universitas Pelita Bangsa
[1]M.Refhaldo@mhs.pelitabangsa.ac.id, [2]ekobudiarto@pelitabangsa.ac.id*, [3]poetrispt@pelitabangsa.ac.id*, [4]sellamonica@pelitabangsa.ac.id*

Abstract

In this era of globalization, the use of cellular phones continues to grow from year to year, from what was originally large to now smaller in size and has an operating system or what is known as a smartphone. One of the most popular smartphones today is a smartphone with an Android-based operating system. Along with the number of android users and the emergence of applications for android, the impact on security threats is getting bigger. One of the security threats that occurs is the emergence of malicious software or commonly known as malware. This study aims to obtain the results of the analysis using the C5.0 method in classifying an application that is identified as malware. With the 80:20 split validation testing technique where 80% of the data will be used as training data and 20% of the data will be used as testing data, the accuracy results are 94.96% on training data and 94.23% on testing data. So it can be concluded that the C5.0 algorithm can be used properly in the process of classifying android malware.

Keywords: *Data Mining, Malware, Algorithm C5.0, RStudio*

Abstrak

Pada era globalisasi ini penggunaan telepon seluler terus berkembang dari tahun ke tahun, dari yang awalnya besar sekarang menjadi berukuran lebih kecil dan memiliki sistem operasi atau yang dikenal dengan smartphone. Salah satu smartphone yang popular saat ini adalah smartphone dengan sistem operasi berbasis android. Bersamaan dengan banyaknya pengguna android dan banyaknya bermunculan aplikasi-aplikasi untuk android, berdampak kepada ancaman keamanan yang semakin besar. Salah satu ancaman keamanan yang terjadi adalah munculnya malicious software atau yang biasa disebut dengan malware. Penelitian ini bertujuan untuk mendapatkan hasil analisi menggunakan metode C5.0 dalam mengklasifikasi sebuah aplikasi yang teridentifikasi sebagai malware. Dengan teknik pengujian split validation 80:20 dimana 80% data akan digunakan sebagai data training dan 20% data akan digunakan sebagai data testing maka hasil akurasi sebesar 94,96% pada data training dan 94,23% pada data testing. Sehingga dapat disimpulkan Algoritma C5.0 dapat digunakan dengan baik dalam proses pengklasifikasian malware android.

Kata kunci: Data Mining, Malware, Algoritma C5.0, RStudio

Pendahuluan

Pada era globalisasi ini penggunaan telepon seluler terus berkembang dari tahun ke tahun, dari yang awalnya besar sekarang menjadi berukuran lebih kecil dan memiliki sistem operasi atau yang dikenal dengan smartphone. Dengan hadirnya smartphone, kecanggihan fitur yang dulunya hanya ada di komputer, sekarang hampir semua dapat dilakukan oleh smartphone. Salah satu smartphone yang popular saat ini adalah smartphone dengan sistem operasi berbasis android. Berdasarkan data dari International Data Center (IDC) untuk kuartal 2 tahun 2022, pemasaran smartphone dengan sistem operasi android mencapai 84% di seluruh dunia [1].

Android sebagai sistem operasi sendiri memiliki berbagai keunggulan, seperti sistem operasi bersifat open source, multitasking, kemudahan dalam penggunaan, hingga banyaknya aplikasi (software) yang dapat dinikmati. Akan tetapi, dibalik keunggulannya tersebut terdapat pula kelemahan. Keunggulan sistem operasi bersifat open access, dimana disediakan platform secara terbuka bagi para pengembang (developer) yang dimaksudkan agar dapat dengan mudah menciptakan dan mengembangkan suatu aplikasi sehingga dapat digunakan pada berbagai macam smartphone. Akan tetapi, hal ini juga yang malah menimbulkan kemudahan bagi pihak yang tidak bertanggung jawab untuk membangun dan mengembangkan aplikasi yang mengancam keamanan dan dapat dengan mudah masuk ke dalam suatu sistem di android.

Bersamaan dengan banyaknya pengguna android dan banyaknya bermunculan aplikasi-aplikasi untuk android, berdampak kepada ancaman keamanan yang semakin besar. Salah satu ancaman keamanan yang terjadi adalah munculnya malicious software atau yang biasa disebut dengan malware [2]. Malware Merupakan aplikasi yang diciptakan untuk menyusup atau merusak sistem komputer. Malware diciptakan dengan maksud tertentu yaitu melakukan aktivitas berbahaya yang berdampak sangat merugikan bagi para korbannya, antara lain seperti penyadapan serta pencurian informasi pribadi, hingga kasus perusakan sistem yang dilakukan oleh penyusup terhadap perangkat korban dengan berbagai alasan [3].

Berdasarkan permasalahan yang sudah dipaparkan sebelumnya, penulis mengusulkan penggunaan teknik data mining untuk menganalisa aplikasi android berdasarkan dataset android malware dari project Drebin. Dalam permasalahan yang ada penulis menggunakan algoritma C5.0, dengan mengklasifikasikan sebuah aplikasi menjadi 2 type yaitu malware atau benign. Dengan pengklasifikasian ini diharapkan dapat meningkatkan efektifitas dalam mendeteksi aplikasi yang terinfeksi malware. Maka dari itu, penulis melakukan sebuah penelitian yang berjudul "KLASIFIKASI APLIKASI MALWARE ANDROID MENGGUNAKAN ALGORITMA C5.0".

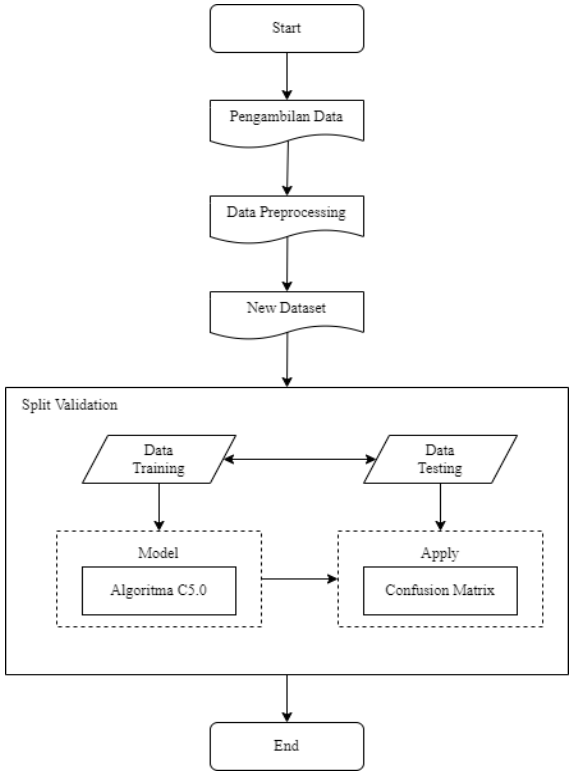
Metode Penelitian

Penelitian ini menggunakan objek malware dari The Drebin Dataset - Technische Universität Braunschweig. Dalam penelitian ini fokus penelitian hanya akan berfokus pada sampel data malware. Mengembangkan penelitian yang telah dilakukan oleh Suleiman Y. Yerima dan Sakir Sezer dengan judul "DroidFusion: A Novel Multilevel Classifier Fusion Approach for Android Malware Detection" menghasilkan sebuah dataset. Penelitian tersebut dilakukan di Institute of Electrical and Electronics Engineers, Manhattan, New York, U.S. Proses dilakukan kepada objek penelitian dengan cara memasang dan menjalankan aplikasi malware dan aplikasi benign pada perangkat smartphone. Untuk mendapatkan hasilnya perangkat smartphone juga telah dimodifikasi agar dapat digunakan untuk monitoring.

Jenis Data :
Data penelitian ini bersifat kuantitatif. Data tersimpan dengan format Comma Separated Values (.csv). Dengan nama artikel Android Malware Dataset for Machine Learning. Dalam website tersebut penulis mendapatkan dataset malware dengan nama drebin-215-dataset-5560malware-9476-benign.csv. Dari dataset ini didapatkan 216 attribut termasuk class.

Data Yang Digunakan :
Data yang digunakan pada penelitian ini merupakan data sekunder. Data sekunder merupakan data yang didapatkan tidak secara langsung dari objek atau subjek penelitian. Data ini didapat dari website Kaggle.com dari penelitian yang pernah dilakukan pada Januari 2018. Dataset ini terdiri atas 215 atribut dan berisi 15.036 aplikasi, dengan jumlah aplikasi malware sebanyak 5560 dan aplikasi benign sebanyak 9476 [18].

Metode Yang Digunakan :
Metode penelitian merupakan cara ilmiah untuk mendapatkan data dengan tujuan dan kegunaan tertentu. Metode yang digunakan penulis pada penelitian ini merupakan data mining menggunakan algoritma C5.0 dengan teknik klasifikasi. Penelitian ini akan melalui beberapa proses. Alur urutan proses pada penelitian ini akan dijelaskan pada metode penelitian ini, metode penelitian ini dapat dilihat sebagaimana pada gambar berikut ini:



Gambar 3.1 Metode yang Digunakan

Pengambilan Data :
Dataset didapatkan melalui website penyedia dataset terbuka bernama Kaggle.com, pada halaman yang berjudul "Android Malware Dataset for Machine Learning". Halaman tersebut dibuat oleh user Bernama Shashwat Tiwari pada September 2021. Dataset yang digunakan sudah memiliki lisensi Attribution 4.0 International (CC BY 4.0) yang berarti diperbolehkan untuk disebarluaskan dan digunakan untuk kepentingan umum.

Data Preprocessing :
Dataset yang sudah didapatkan kemudian dilakukan preprocessing, dimana dataset akan dibersihkan dan diolah agar dapat digunakan dalam proses penelitian. Pada tahapan ini dilakukan pembersihan data yang tidak terpakai dan dilakukan beberapa perubahan agar dataset dapat terbaca oleh program yang digunakan.

New Dataset :
Dataset yang sudah melalui tahapan preprocessing akan akan dibuatkan dataset baru agar lebih mudah dalam melakukan penelitian.

Proses Pengujian :
Dataset baru yang sudah melalui tahapan preprocessing kemudian dilakukan proses pengujian. Dataset akan diuji menggunakan teknik split validation. Data akan dipecah menjadi dua bagian dengan perbandingan 80:20, dimana 80% dari data akan digunakan sebagai data training dan 20% data akan digunakan sebagai data testing. Data training yang sudah dibuat akan diterapkan pada algoritma C5.0. Setelah penerapan data algoritma dilakukan, kemudian hasil tersebut akan diujikan dengan data testing sehingga akan didapatkan sebuah hasil berbentuk confusion matrix sebagai hasil akhir.

Evaluasi dan Validasi Hasil :
Evaluasi dilakukan dengan cara mengamati dan menganalisa hasil dari algoritma yang digunakan untuk memastikan bahwa hasil pengujian itu benar atau tidak sesuai dengan pembahasan. Sedangkan, validasi dilakukan dengan mengukur hasil prediksi untuk mengetahui tingkat akurasi, presisi, dan recall.

Hasil dan Pembahasan

Berdasarkan perhitungan dari confusion matrix data training yang berjumlah 11932 data, 7510 data diklasifikasikan sebagai aplikasi aman dan 35 data diklasifikasikan sebagai aplikasi malware ternyata merupakan aplikasi aman. Selain itu, berdasarkan confusion matrix yang sama terdapat 4288 data yang diklasifikasikan sebagai aplikasi malware dan 99 data diklasifikasikan sebagai aplikasi aman ternyata merupakan aplikasi malware. Berdasarkan hal tersebut, maka diciptakanlah hasil nilai accuracy sebesar 98,88%. Data training juga menghasilkan nilai precision sebesar 98,70% dan nilai recall sebesar 99,54%.

Tabel 4.4 Tabel Accuracy, Precision, dan Recall Data Training

Accuracy	***Precision***	***Recall***
0,949631244	0,952790819	0,968323393
94,96%	95,28%	96,83%

Berdasarkan perhitungan dari confusion matrix data testing yang berjumlah 3103 data, 1896 data diklasifikasikan sebagai aplikasi aman dan 35 data diklasifikasikan sebagai aplikasi malware ternyata merupakan aplikasi aman. Selain itu, berdasarkan confusion matrix yang sama terdapat 1126 data yang diklasifikasikan sebagai aplikasi malware dan 46 data diklasifikasikan sebagai aplikasi aman ternyata merupakan aplikasi malware. Berdasarkan hal tersebut, maka diciptakanlah hasil nilai accuracy sebesar 97,39%. Data training juga menghasilkan nilai precision sebesar 97,63% dan nilai recall sebesar 98,19%.

Tabel 4.5 Tabel Accuracy, Precision, dan Recall Data Testing

Accuracy	***Precision***	***Recall***
0,94231389	0,943768997	0,964785085
94,23%	94,38%	96,48%

Kesimpulan

Berdasarkan hasil penelitian ini, klasifikasi malware android menggunakan algoritma C5.0 disertai dengan Teknik Split Validation 80:20 memberikan hasil yang baik. Menghasilkan tingkat akurasi dalam mengklasifikasi sebuah aplikasi android yang mengandung malware sebesar 94,96% pada training dan 94,23% pada data testing. Hal ini menunjukan Algoritma C5.0 yang disertai dengan Split Validation 80:20 dapat digunakan dengan baik dalam proses pengklasifikasian malware android.

Saran

Saran peneliti berdasarkan hasil penelitian yang telah diperoleh untuk dilakukan penelitian selanjutnya perlu dilakukan perbandingan dengan algoritma lain untuk menguji sejauh mana algoritma C5.0 dapat diandalkan dalam mengklasifikasi suatu aplikasi android. Selain itu perlu dilakukan pengujian menggunakan berbagai tools untuk menganalisa hasil yang didapat agar mendapat hasil yang terbaik. Dan juga perlunya dilakukan penelitian menggunakan dataset dengan topik yang sama dengan penggunaan data dalam jumlah yang lebih banyak.

Daftar Pustaka

[1] N. Popal and R. Reith, "Smartphone Market Share," IDC, 4 August 2022. [Online]. Available: https://www.idc.com/promo/smartphone-market-share. [Accessed 10 August 2022].

[2] R. R. Ramli and A. Ika, "Bank Indonesia Akui Diretas, Kena Serangan "Ransomware", Data Kritikal Dipastikan Aman," 21 January 2022. [Online]. Available: https://money.kompas.com/read/2022/01/21/101728026/bank- indonesia-akui-diretas-kena-serangan-ransomware-data-kritikal- dipastikan?page=all. [Accessed 10 August 2022].

[3] T. A. Cahyanto, V. Wahanggara and D. Ramadana, Analisis dan Deteksi Malware Menggunakan Metode Analisis Dinamis, JUSTINDO, Jurnal Sistem & Teknologi Informasi Indonesia, 2017.

[4] S. Y. Yerima and S. Sezer, "DroidFusion: A Novel Multilevel Classifier Fusion Approach for Android Malware Detection," IEEE Transactions on Cybernetics, vol. 49, no. 2, pp. 453-466, 2019.

[5] D. Utomo and M. Mesran, "Analisis Komparasi Metode Klasifikasi Data Mining dan Reduksi Atribut Pada Data Set Penyakit Jantung," JURNAL MEDIA INFORMATIKA BUDIDARMA, vol. 4, p. 437, 2020.

[6] A. H. Lashkari, A. F. A. Kadir and L. a. G. A. A. Taheri, "Toward Developing a Systematic Approach to Generate Benchmark Android Malware Datasets and Classification," in 2018 International Carnahan Conference on Security Technology (ICCST), 2018 , pp. 1-7.

[7] I. Id, Machine Learning : Teori, Studi Kasus dan Implementasi Menggunakan Python, Pekanbaru: Badan Penerbit Universitas Riau , 2021.

[8] Y. Setianto, K. Kusrini and H. Henderi, "Penerapan Algoritma K-Nearest Neighbour Dalam Menentukan Pembinaan Koperasi Kabupaten Kotawaringin Timur," Creative Information Technology Journal, vol. 5, p. 232, 2019.

[9] S. Faisal, "KLASIFIKASI DATA MINNING MENGGUNAKAN ALGORITMA C4.5 TERHADAP KEPUASAN PELANGGAN SEWA KAMERA CIKARANG," Techno Xplore : Jurnal Ilmu Komputer dan Teknologi Informasi, vol. 4, pp. 1-8, 2019.

[10] R. Rian Putra and C. Wadisman, "Implementasi Data Mining Pemilihan Pelanggan Potensial Menggunakan Algoritma K Means," INTECOMS: Journal of Information Technology and Computer Science, no. 1, pp. 72-77, 2018.

[11] R. T. Vulandari, Data mining : teori dan aplikasi rapidminer, Yogyakarta : Gava Media, 2017.

[12] Y. Mardi, "Data Mining : Klasifikasi Menggunakan Algoritma C4.5," Edik Informatika, vol. 2, no. 2, pp. 213-219, 2016.

[13] Y. Agustin, K. Kusrini and E. Luthfi, "Klasifikasi Penerimaan Mahasiswa Baru Menggunakan Algortima C4.5 Dan Adaboost (Studi Kasus : STMIK XYZ)," CSRID (Computer Science Research and Its Development Journal), vol. 9, pp. 1-11, 2017.

[14] I. Iskandar, L. Hiryanto and J. Hendryli, "PREDIKSI KELULUSAN MAHASISWA MENGGUNAKAN ALGORITMA DECISION TREE C4.5 DENGAN TEKNIK PRUNING," Jurnal Ilmu Komputer dan Sistem Informasi, 2018.

[15] U. Riyanto, "Analisis Perbandingan Algoritma Naive Bayes Dan Support Vector Machine Dalam Mengklasifikasikan Jumlah Pembaca Artikel Online," JIKA (Jurnal Informatika), vol. 2, no. 2, p. 62–72, 2019.

[16] P. U. Gio and A. R. Effendie, Belajar Bahasa Pemrograman R (Dilengkapi Cara Membuat Aplikasi Olah Data Sederhana dengan R Shiny), Medan: USU Press , 2017.

[17] T. A. Nengsih, F. Mubarak and V. Y. Sundara, Pemograman R Dasar, Lombok Tengah: Penerbit : Forum Pemuda Aswaja, 2020.

[18] S. Tiwari, "Android Malware Dataset for Machine Learning," Kaggle, 2021. [Online]. Available: https://www.kaggle.com/datasets/shashwatwork/android-malware-dataset-for-machine-learning. [Accessed 21 June 2022].