SKRIPSI

PEMBANGUNAN APLIKASI PENDETEKSI ANOMALI SISTEM LOG APACHE MENGGUNAKAN PEMBELAJARAN MESIN

Ferdian Benyamin

NPM: 2016730086

PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS TEKNOLOGI INFORMASI DAN SAINS
UNIVERSITAS KATOLIK PARAHYANGAN
2020

UNDERGRADUATE THESIS

DEVELOPING APPLICATION FOR ANOMALY DETECTION IN APACHE SYSTEM LOG USING MACHINE LEARNING

Ferdian Benyamin

NPM: 2016730086

DEPARTMENT OF INFORMATICS
FACULTY OF INFORMATION TECHNOLOGY AND SCIENCES
PARAHYANGAN CATHOLIC UNIVERSITY
2020

LEMBAR PENGESAHAN

PEMBANGUNAN APLIKASI PENDETEKSI ANOMALI SISTEM LOG APACHE MENGGUNAKAN PEMBELAJARAN MESIN

Ferdian Benyamin

NPM: 2016730086

Bandung, 22 Juni 2020

Menyetujui,

Pembimbing

Chandra Wijaya, M.T.

Ketua Tim Penguji

Anggota Tim Penguji

Elisati Hulu, M.T.

Pascal Alfadian, Nugroho, M.Comp.

Mengetahui,

Ketua Program Studi

Mariskha Tri Adithia, P.D.Eng

PERNYATAAN

Dengan ini saya yang bertandatangan di bawah ini menyatakan bahwa skripsi dengan judul:

PEMBANGUNAN APLIKASI PENDETEKSI ANOMALI SISTEM LOG APACHE MENGGUNAKAN PEMBELAJARAN MESIN

adalah benar-benar karya saya sendiri, dan saya tidak melakukan penjiplakan atau pengutipan dengan cara-cara yang tidak sesuai dengan etika keilmuan yang berlaku dalam masyarakat keilmuan.

Atas pernyataan ini, saya siap menanggung segala risiko dan sanksi yang dijatuhkan kepada saya, apabila di kemudian hari ditemukan adanya pelanggaran terhadap etika keilmuan dalam karya saya, atau jika ada tuntutan formal atau non-formal dari pihak lain berkaitan dengan keaslian karya saya ini.

Dinyatakan di Bandung,
Tanggal 22 Juni 2020

Ferdian Benyamin
NPM: 2016730086

PERNYATAAN

Dengan ini saya yang bertandatangan di bawah ini menyatakan bahwa skripsi dengan judul:

PEMBANGUNAN APLIKASI PENDETEKSI ANOMALI SISTEM LOG APACHE MENGGUNAKAN PEMBELAJARAN MESIN

adalah benar-benar karya saya sendiri, dan saya tidak melakukan penjiplakan atau pengutipan dengan cara-cara yang tidak sesuai dengan etika keilmuan yang berlaku dalam masyarakat keilmuan.

Atas pernyataan ini, saya siap menanggung segala risiko dan sanksi yang dijatuhkan kepada saya, apabila di kemudian hari ditemukan adanya pelanggaran terhadap etika keilmuan dalam karya saya, atau jika ada tuntutan formal atau non-formal dari pihak lain berkaitan dengan keaslian karya saya ini.

Dinyatakan di Bandung,
Tanggal 22 Juni 2020

Ferdian Benyamin
NPM: 2016730086

ABSTRAK

Kemajuan teknologi membuat internet merupakan salah satu sarana untuk membantu pekerjaan setiap orang. Website adalah salah satu kategori layanan yang populer digunakan di internet. Website ini dijalankan oleh webserver yang beroperasi selama 24 jam tanpa henti. Salah satu webserver yang populer digunakan adalah apache. Apache akan mencatat setiap permintaan dari client, sehingga seluruh kejadian akan dapat ditelusuri. Banyaknya data membuat proses pengecekan menjadi lama dan kurang akurat jika dilakukan secara manual. Solusi permasalahan ini adalah menggunakan pembelajaran mesin untuk memeriksa log akses, sehingga dapat ditemukan, apabila terdapat kejadian yang bersifat anomali. Teknik yang digunakan melibatkan pembelajaran terarah (supervised learning) dan juga pembelajaran tak terarah (unsupervised learning). Sebelum diproses, file log ini akan diperiksa dan akan diperoleh atribut - atribut yang menjelaskan karakteristik dari log tersebut. Kemudian akan ditentukan 2 kategori data, yaitu data latihan dan data pengujian. Data latihan merupakan data yang telah diberi label dan dijadikan masukan untuk system agar dapat mempelajari karakteristik dari file log. Sedangkan data pengujian adalah data yang akan dikenali oleh system berdasarkan data latihan yang telah dimasukan sebelumnya. Pengujian yang dilakukan akan bertujuan untuk mengevaluasi baik buruknya performa algoritma dari model pembelajaran. Akurasi akan menjadi pertimbangan dalam menentukan hal tersebut. Hasil dari penelitian ini adalah sebuah aplikasi yang dapat menentukan ada atau tidaknya anomali yang terjadi berdasarkan dari file log apache yang menjadi masukkan.

Kata-kata kunci: Sistem Log Apache, Deteksi Anomali, Pembelajaran Machine, Apache, Log Website, Akses Log

ABSTRACT

The advance of technology in this modern era makes humans use the internet as a means of doing their jobs. Website services are one of the most popular categories used at this time. This service provide 24 hours of operating without stop. One of the most used web service brand is called Apache's. Apache's will write up every request from client, therefore it can track it on. A tons of data's that must be proceed affect the investigation takes longer and least accurate. The solution of the problem is using machine learning as the key of investigate the logs access, so it can be found, if there is a problem or case that anomaly. The technique that will used concluded supervised learning and unsupervised learning. Before proceeding even further, the data will be extracted some attributes that will explain there characteristic of the logs. Then from that point will be determined 2 types of data's, there is training and testing. Train data is the data that already classified and will be used as program input to learn the characteristic of the data. The other is test data, the data that will be used to be recognized by system from the train data. This experiment conclude to evaluated the model of algorithm and the perform from the task. Accuracy will be the main index to determined the algorithm.The result of this experiment is a program that can determined anomaly in the log from the input log file that will be used as input data.

Keywords: System Log Apache, Anomali Detection, Machine Learning, Apache, Website Logs, Access Log

Dipersembahkan untuk Tuhan YME, keluarga, negara, almamater dan teman-teman yang telah memberi dukungan dalam pembuatan skripsi ini.

KATA PENGANTAR

Puji syukur penulis panjatkan kepada Tuhan Yang Maha Esa, karena dengan rahmat dan karunia-Nya, penulis dapat menyelesaikan penyusunan skripsi berjudul "Pembangunan Aplikasi Pendeteksi Anomali Sistem Log Apache Menggunakan Pembelajaran Mesin". Skripsi ini dibuat dan diajukan untuk memenuhi salah satu syarat untuk memperoleh gelar Sarjana pada Program Studi Teknik Informatika, Universitas Katolik Parahyangan. Selain itu, penulisan skripsi ini bertujuan untuk memberikan pengetahuan kepada pembaca mengenai anomali apa saja pada log dan bagaimana cara mengetahuinya dengan efektif. Selama penulisan skripsi ini, penulis menyadari bahwa penulisan skripsi ini dapat selesai karena bantuan dan dukungan beberapa pihak. Oleh karena itu, penulis mengungkapkan rasa terima kasih kepada:

1. Bapak Chandra Wijaya, M.T.selaku dosen pembimbing yang telah membimbing dan mendukung penulis selama proses penyusunan skripsi ini.
2. Bapak Elisati Hulu, M.T.dan Bapak Pascal Alfadian, Nugroho, M.Comp.selaku dosen penguji yang telah memberikan kritik dan saran yang membangun.
3. Keluarga yang selalu mendukung dan membantu penulis
4. Komunitas Sel *Youth Teens Bethesda* yang selalu mengingatkan dan mendoakan penulis untuk tetap fokus dan menyelesaikan skripsi ini.
5. Cornelius David sebagai kakak tingkat yang sudah membantu penulis dalam mengerjakan skripsi ini.
6. Teman - Teman perjuangan skripsi (Cantika, Joshua, Jason, Kevin) dan para admin LAB FTIS UNPAR (Febrian, Yehezkiel, Cahyadi).

Penulis menyadari bahwa penelitian ini masih jauh dari kata sempurna. Oleh karena itu, penulis memohon maaf jika terdapat kekurangan pada penelitian ini. Penulis juga mengharapkan kritik dan saran yang membangun untuk menyempurnakan penelitian ini. Semoga penelitian ini dapat bermanfaat bagi segenap pihak yang berkepentingan.

Bandung, Juni 2020

Penulis

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

BAB 1

PENDAHULUAN

1.1 Latar Belakang

Di masa sekarang ini, dengan majunya industri teknologi tidak aneh jika internet menjadi salah satu sarana untuk menyelesaikan permasalahan didalam kehidupan sehari - hari. Pada umumnya perangkat yang digunakan adalah komputer penyedia layanan yang beroperasi selama 24 jam penuh. Selama masa operasi sistem, semua aktifitas akan dicatat disuatu tempat didalam komputer tersebut. Nama catatan tersebut adalah *system log* atau disingkat dengan *syslogs* atau *log file.Log file* ini ada bermacam - macam misalnya di sistem operasi *windows log file* dapat dilihat menggunakan *windows event viewer*, didalam sistem operasi linux dapat dilihat di *var/log/*, didalam *web server* seperti apache dan nginx yang memiliki *access log file* dan *error log*. Hal ini dibuat agar saat masalah yang membuat sistem tidak berfungsi dengan semestinya, tim ahli dapat mengolah data yang tersedia berdasarkan sumber masalahnya dan nantinya dapat memberikan solusi juga penanganan agar hal yang sama tidak terulang kembali. Banyaknya data yang perlu diolah dan dianalisis membuat proses penanganan relatif lama terlebih ada data yang terlewat sehingga hasil dari penangannya kurang akurat. Maka diperlukan sistem yang mampu mengolah data dan menganalisis data dengan cepat, sehingga proses penanganan dapat diatasi dengan lebih cepat dan akurat. *Machine Learning* adalah salah satu solusi yang mungkin untuk menyelesaikan permasalahan tersebut.

Machine Learning adalah teknik yang membuat sistem mampu belajar dari data. Ketika menggunakan *machine learning* biasanya data tidak dapat langsung diproses, karena data tersebut harus disesuaikan terlebih dahulu. Proses menyesuaikan ini disebut *preprocessing*. Pada proses ini data akan disesuaikan agar lebih terstruktur. Data yang sudah terstruktur ini nantinya akan diproses (*learn*) oleh sistem dan mengeluarkan representasi data yang disebut *model*. Data yang digunakan untuk menciptakan model disebut dengan *traning sets* atau *data train*. Semakin variatif data yang digunakan semakin baik model yang tercipta. Setelah model tercipta, *model* akan diuji performanya menggunakan set data lain yang disebut dengan data *test set*. Setelah proses ini nantinya data dan *model* akan dianalisis baik buruknya. Baik buruknya *model* bergantung terhadap variasi datanya saat di *training*, jika datanya tidak bervariasi maka akan terjadi *overfitting* jika datanya kurang representatif maka terjadi *underfitting*.

Pada skripsi ini akan dilakukan penelitian bagaimana membuat sistem sampai membuat aplikasi yang mampu mendeteksi anomali yang terdapat di *log file*, untuk *log file* yang akan diteliti adalah *log file access* apache. Pembangunan sistem dan aplikasi ini akan menggunakan *Machine Learning* untuk mendeteksi anomali didalam *log file*. Pada skripsi ini juga akan dijelaskan tahap per tahap bagaimana data mentah dari *log file* diproses menjadi data terstruktur yang nantinya diolah oleh algortima *machine learning*. Alasan utama penulis menggunakan teknik *machine learning* adalah, karena *machine learning* bekerja dengan baik dengan data yang banyak dan data *log file* banyak. Diharapkan dengan adanya skripsi ini maka ketika terjadi anomali yang berhasil dideteksi oleh perangkat lunak, admin dapat dengan cepat memberikan antisipasi sehingga dapat tetap menjaga keamanan pada data itu sendiri.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang sudah dibuat, maka rumusan masalah yang akan dibahas antara lain:

1. Berapa akurasi yang dapat dihasilkan dengan menggunakan algortima *Machine Learning*
2. Bagaimana membangun model yang dapat menemukan anomali tersebut?
3. Bagaimana membangun aplikasi yang dapat menemukan anomali?

1.3 Tujuan

Berdasarkan rumusan masalah yang sudah dibuat, maka tujuan penelitian ini adalah sebagai berikut:

1. Mengolah data *access log file* agar dapat diproses didalam *machine learning*
2. Membangun model *machine learning* yang dapat menemukan anomali
3. Membangun aplikasi yang dapat mendeteksi anomali dengan memanfaatkan teknik *machine learning*

1.4 Batasan Masalah

Beberapa batasan yang dibuat terkait penelitian ini antara lain :

1. Sistem yang dibangun hanya diuji dengan *operating system* windows 10.
2. Sistem yang dibangun hanya dapat menerima input berupa *file* dengan ekstensi *.txt*
3. Sistem yang dibangun memerlukan proses *train* dan *test*.

1.5 Metodologi

Langkah langkah yang akan dilakukan dalam melakukan penelitian ini adalah:

1. Melakukan studi literatur tentang *system log* apache
2. Melakukan studi literatur tentang *machine learning* dan *decision tree*
3. Menentukan model yang digunakan untuk *preprocessing*
4. Merancang alur *preprocessing* dan sistem
5. Merancang perangkat lunak dan mengimplementasikan teknik *machine learning*
6. Melakukan pengujian terhadap aplikasi yang telah dibuat
7. Melakukan analisis terhadap model dan perangkat lunak yang telah dibuat

1.6 Sistematika Pembahasan

Dokumentasi dari penelitian ini dibangun dalam enam bab dengan sistematika dan penjelasan sebagai berikut:

1. Bab 1. Pendahuluan
   Bab 1 berisi latar belakang mengapa topik 'Pembangunan Aplikasi Pendeteksi Anomali Log *Apache* Menggunakan *Machine Learning*' dibuat sebagai judul skripsi ini. Setelah itu dijelaskan juga Rumusan masalah dan tujuan dari topik ini, serta metodoli dan sistematika pembahasan dalam skripsi ini.

2. Bab 2. Landasan Teori
   Bab 2 berisi tentang teori sebagai landasan utama dalam skripsi ini. Konsep yang akan dibahas adalah Log secara umum dan *Access Log Apache* secara spesifik seperti format dan arti dari log tersebut. Selanjutnya menjelaskan konsep dasar dari *machine learning* dan beberapa contoh *machine learning* yang ada. Contoh yang dijelaskan antara lain adalah *Decision Tree*, K-NN, DBSCAN dan K- Means. Selanjutnya akan menjelaskan bagaimana melakukan evaluasi dari model *machine learning* yang dibuat serta menjelaskan konsep dari *preprocessing* dan mengapa hal tersebut dibutuhkan ditopik ini.

3. Bab 3. Analisis
   Bab 3 berisi hasil analisi berdasarkan landasan teori yang digunakan. Seperti bagaimana memilih fitur yang akan digunakan, agar mendapatkan hasil yang baik. Juga hasil analisis untuk memilih model *machine learning* yang akan digunakan untuk mencapai target yang diinginkan. Pada bagian ini juga akan dijelaskan attribut attribut yang dibutuhkan saat membangun aplikasi, serta contoh kasus yang akan terjadi didalam aplikasi. Pada bab ini juga akan dijelaskan mengenai *use case* diagram yang berupa analisis dari aplikasi agar pengguna dapat menggunakannya untuk mencapai tujuan yang diiinginkan.

4. Bab 4. Perancangan
   Bab 4 berisi penjelasan dari percanganan yang merupakan hasil dari analisi masalah di Bab 3. Pada bagian ini akan dijelaskan mengenai penjelasan spesifik mengenai masukkan dan keluaran. Perancangan dari kelas dalam aplikasi yan akan digunakan berserta pejelasan fungsi - fungsi untuk mencapai tujuan dari skripsi ini. Selanjutnya akan dijelaskan mengenai rancangan tampilan antarmuka yang akan dibangun. Akan dijelaskan juga mengenai alur program bagaimana data log yang diolah dapat diproses dan mencapai tujuan dari skripsi ini.

5. Bab 5. Pengujian dan Eksperimen
   Bab 5 berisi tentang pengujian dari aplikasi dalam mendeteksi log dengan berbagai macam skenario yang sudah dibuat. Pada bagian ini akan dijelaskan hasil dari pengujian dari beberapa eksperimen yang dikerjakan juga hasil kesimpulan analisis dari eksperimen ini. Pada bagian akhir dari bab ini akan dijelaskan hasil analisis dari keseluruhan pengujian dan eksperimen yang digunakan dari berbagai macam skenario yang ada.

6. Bab 6. Kesimpulan dan Saran
   Bab 6 akan berisi tentang kesimpulan dari hasil analisis secara keseluruhan saat melakukan pembangunan aplikasi ini. Juga akan memuat tentang saran untuk peneliti selanjutnya agar dapat mengembangkan aplikasi ini kedepannya.