BAB I

PENDAHULUAN

1.1. **Latar Belakang Masalah**

Knowledge Discovery in Database (KDD) merupakan suatu proses dalam melakukan identifikasi pola yang *valid*, baru dan berguna serta dapat dipahami dari sekumpulan data yang besar dan kompleks (Maimon & Rokach, 2010). *Data Mining* adalah inti dari proses KDD yang melibatkan kesimpulan dari algoritma yang mengeksplorasi data, mengembangkan model serta menemukan pola yang sebelumnya tidak diketahui. Proses penambangan atau *mining* dapat dilakukan dengan metode statistik, matematika hingga teknologi kecerdasan buatan (*Artificial Intelligent*) dan *Machine Learning* yang bertujuan untuk mengekstrak serta mengidentifikasi informasi dan pengetahuan yang potensial yang terkandung dalam suatu *database* besar. (Turban, *et al.* 2004).

Salah satu metode yang sering digunakan dalam *Data Mining* adalah *K-Mean Clustering*, yaitu pengelompokan sejumlah data atau objek ke dalam klaster berdasarkan atribut tertentu sehingga setiap anggota di dalam klaster mempunyai kemiripan satu dengan yang lain. K-Mean merupakan metode yang cukup populer dan banyak digunakan para peneliti karena merupakan algoritma yang sederhana dan mudah dalam implementasi. K-Mean merupakan *Unsupervised Learning Algorithm* yang bermakna algoritma klasterisasi dilakukan pada kelompok data yang tidak mempuyai label. Algoritma ini tidak membutuhkan *training* dalam melakukan pengelompokkan data. Data-data dikelompokan berdasarkan kemiripan nilai atribut yang ada. Algoritma ini cukup efektif dalam melakukan klasterisasi data tanpa label, sebagaimana yang dilakukan oleh Mohd Ariffin, *et al.* (2020b) dalam penelitiannya menggunakan metode K-Mean untuk menemukan profil penggunaan trafik internet. Dalam penelitian yang dilakukan, Mohd Ariffin membagi penggunaan trafik internet dalam tiga klaster yaitu *high*, *medium* dan *low*. Hasil dari penelitian tersebut

kemudian dapat direkomendasikan sebagai pedoman dalam melakukan *management bandwidth* serta memperkuat kebijaksanaan keamanan jaringan (*Network Security Policy*).

Penggunaan metode K-Mean juga dilakukan oleh Wang, *et al.* (2020) sebagaimana dijelaskan dalam artikel yang menjelaskan deteksi *domain* berbahaya. Pada salah satu tahap yang dilalui, Wang melakukan klasterisasi sejumlah besar data trafik *DNS* untuk menemukan *domain* berbahaya menggunakan metode K-Mean ini.

Klasterisasi juga dapat dimanfaatkan dalam mengurangi beban penggunaan trafik internet dan akses pengguna seperti yang dijelaskan Nasser (2018) dalam artikelnya mengenai pengelompokan pengguna *web*. Artikel tersebut memuat penjelasan mengenai klasterisasi penggunaan trafik internet berdasarkan *domain* yang dikunjungi serta waktu yang dihabiskan dalam kunjungan tersebut.

Dalam sebuah penelitian yang dijalankan oleh Ruan, *et al.* (2013) yang melakukan klasterisasi berupa *Time Series DNS Query* juga menggunakan metode K-Mean untuk dapat menemukan pola perilaku pengguna *web* pada *domain* yang berbeda.

Domain Name System (*DNS*) menyediakan data yang kaya dan menarik, serta dapat diekstrak untuk mengungkap informasi yang dapat dianalisis bagi berbagai keperluan seperti tindakan keamanan, pembatasan *bandwidth* hingga kebijakan lain yang diterapkan dalam suatu jaringan. Penelitian yang sedang dilakukan ini menggunakan data primer berupa *dataset* yang bersumber dari *log file* yang dihasilkan oleh *DNS Server*. Sebuah artikel yang ditulis oleh Snyder (2009) menjelaskan langkah-langkah awal yang dapat dilakukan dalam menambang data dari suatu *log file* yang dihasilkan oleh *DNS Server*. Meskipun artikel tersebut berusia cukup lama, namun tahapan *pre-processing* serta gagasan yang dijelaskan berupa langkah-langkah dasar masih tetap relevan dijadikan sebagai pedoman dalam menyelesaikan penelitian ini.

Penelitian ini dilakukan untuk membuat klasterisasi terhadap penggunaan trafik jaringan *internet*, sehingga diharapkan memberikan manfaat yang dapat digunakan untuk meningkatkan layanan jaringan (*QoS*) serta melakukan efisiensi terhadap pemakaian *bandwidth*. Objek penelitian ini mengambil lokasi pada jaringan Pemerintah Kota Pekanbaru, dengan target pengguna pegawai yang bekerja pada lingkungan pemerintahan tersebut dalam memanfaatkan trafik internet yang telah disediakan. Hasil klasterisasi tersebut diharapkan dapat bermanfaat untuk membuat

kebijakan dalam mengelola pemakaian *bandwidth* sehingga menjadi lebih efisien dan tepat sasaran.

1.2. **Perumusan Masalah**

Berdasarkan latar belakang yang telah dijelaskan, maka dapat ditarik rumusan masalah sebagai berikut:

1. Bagaimana melakukan *redirect* permintaan *DNS* (*DNS query*) pada satu *Server* terpusat sehingga dapat dilakukan *logging* terhadap seluruh *query* pada jaringan?
2. Bagaimana melakukan ekstraksi *DNS log file* untuk mendapatkan *dataset* awal serta melakukan transformasi terhadap *dataset* yang sudah diperoleh tersebut?
3. Bagaimana menggunakan algoritma K-Mean untuk mendapatkan klasterisasi trafik penggunaan *internet* berdasarkan *domain* yang dikunjungi?

1.3. **Batasan Masalah**

Untuk menghasilkan pembahasan yang *optimal* dan berada pada *scope* yang sesuai dengan topik, penelitian ini dibatasi pada hal berikut:

1. Penelitian mengambil objek jaringan *internet* Pemerintah Kota Pekanbaru (Pemko Pekanbaru), dengan *dataset* yang mencerminkan aktivitas penggunaan trafik internet yang dilakukan oleh Pegawai Negeri (ASN), Tenaga Harian Lepas (THL), Tenaga Ahli serta Tenaga Pendukung dalam jajaran Pemerintah Kota.
2. Pengambilan *dataset* dilakukan dalam rentang waktu 5 hari kerja 13-Jul-2022 s/d 19-Jul-2022 selama waktu jam kerja 08:00 s/d 17:00, tidak termasuk waktu istrirahat (12:00 s/d 13:00).
3. *Dataset* yang dianalisis merupakan ekstraksi informasi yang berasal dari *log file* aplikasi *DNS Server*.
4. Penelitian ini dilakukan dengan fokus utama klasterisasi penggunaan trafik jaringan internet berdasarkan jumlah *DNS Request* terhadap situs dikunjungi.
5. Trafik yang menjadi target penelitian berupa aliran *downstream* yang merupakan trafik yang di-*request* oleh pengguna internet di dalam jaringan. Trafik dimaksud telah dipisahkan dari trafik *upstream* (*request* terhadap *Authoritative Nameserver* yang melewati *router*/*gateway* yang sama).

6. Tidak termasuk analisis terhadap komponen transmisi seperti *protocol*, nomor *port*, *header* dan komponen protokol TCP/IP lainnya.

1.4. **Tujuan Penelitian**

Tujuan yang ingin dicapai dalam pelaksanaan penelitian ini dapat diuraikan sebagai berikut:

1. Menemukan klaster penggunaan trafik internet sebagai media untuk menunjang pelaksanaan Sistem Pemerintahan Berbasis Elektronik (SPBE).
2. Merancang metodologi yang dapat digunakan untuk melakukan klasterisasi penggunaan trafik internet.
3. Menerapkan algoritma K-Means dalam upaya melakukan klasterisasi penggunaan trafik internet berdasarkan *DNS request* terhadap *domain* yang dikunjungi.
4. Menguji metodologi digunakan sehingga dinyatakan layak untuk diterapkan dalam menentukan penggunaan trafik internet.

1.5. **Manfaat Penelitian**

Penelitian ini dilakukan dengan memanfaatkan sumber data primer yang cukup akurat serta metode yang tepat. Sehingga diharapkan dapat memberi manfaat yang baik dan sesuai dengan tujuan penelitian ini. Adapun manfaat yang dapat diambil dari hasil penelitian ini adalah:

1. Sebagai wacana tambahan bagi dunia akademis dalam melakukan penelitian terkait klasterisasi penggunaan trafik/*bandwidth* internet.
2. Mempelajari dan memanfaatkan *log file* yang tersedia pada *DNS Server* (*dnsmasq*).
3. Menjadi salah satu pertimbangan bagi *Administrator* jaringan untuk menentukan besaran *bandwidth* yang diberikan sehingga dapat mengoptimalkan penggunaan *bandwidth* (QoS).
4. Untuk melakukan efisiensi terhadap penggunaan trafik internet sehingga pemakaiannya menjadi tepat dan sesuai sasaran.

5. Dalam pengembangannya diharapkan menjadi pedoman bagi pemangku kebijakan untuk mengoptimalkan rencana anggaran biaya sewa/beli bandwidth internet.

1.6. **Sistematika Penulisan**

Sistematika penulisan ini berguna untuk memberikan gambaran yang jelas dan terarah sehingga tidak menyimpang dari pokok permasalahan yang telah ditetapkan. Adapun sistematika penulisan yang diterapkan pada penelitian ini adalah:

BAB I - PENDAHULUAN

Bab ini merupakan tahapan awal yang dilakukan untuk menjelaskan latar belakang, rumusan masalah, batasan, tujuan dan manfaat penelitian serta sistematika penulisan.

BAB II - LANDASAN TEORI

Bab ini membahas teori-teori yang menjadi landasan dalam melakukan penelitian terkait dengan *Knowledge Discovery in Database* (KDD) dan *Data Mining*, Klasterisasi, K-Mean, Bahasa Pemrograman Python, serta teori pendukung lain yang digunakan dalam melakukan analisis dan pembahasan.

BAB III - METODOLOGI PENELITIAN

Pada bagian ini diuraikan hal-hal yang menjadi metode dalam melakukan penelitian ini, di antaranya waktu dan tempat penelitian, kerangka kerja, identifikasi dan analisis masalah, pre-processing, serta kesimpulan yang akan dirumuskan.

BAB IV – ANALISA DAN PERANCANGAN

Bagian yang menjelaskan rencana kerja dan tahap-tahap yang dilalui dalam melakukan penelitian ini. Pada tahap ini juga dilakukan analisis terhadap objek penelitian serta hal-hal lain yang menjadi bagian dari objek dimaksud, termasuk teknik pengumpulan dan penyaringan data, transformasi data, kode program pendukung, algoritma.

BAB V – IMPLEMENTASI DAN HASIL

Bab ini berisi ulasan mengenai hasil yang diperoleh dari penelitian, disertai analisis yang dilakukan untuk menyimpulkan apakah sudah sesuai dengan yang diharapkan serta sesuai dengan hipotesis yang ada.

BAB VI - KESIMPULAN DAN SARAN

Bagian terakhir penulisan tesis ini merupakan kesimpulan dari penelitian serta saran-saran yang diharapkan menjadi panduan guna pengembangan di masa datang.