BAB II

TINJAUAN PUSTAKA

Suwarningsih (2012) melakukan penelitian mengenai perancangan dan pembangunan sebuah sistem (aplikasi) untuk pengawasan terhadap bahan baku obat. Dikarenakan bahan baku obat yang sangat banyak, maka dipilih suatu cara yaitu pengelompokan data. Pengelompokan data berguna untuk mempermudah monitoring ketersedian bahan baku obat. Tujuan dari pengelompokan data, yaitu untuk memudahkan proses pencarian, sehingga dapat mempercepat waktu yang digunakan user dalam suatu proses pencarian. Hasil dari penelitain yang didapat, bahwa sistem (aplikasi) yang dirancang dan dibangun ini dapat memudahkan user dalam pencarian dan monitoring bahan baku obat yang ada, duna pelaporan ke badan POM.

Kotsiantis et al (2006) mengatakan bahwa banyak faktor yang mempengaruhi keberhasilan *Machine Learning* (*ML*) dalam merepresentasikan suatu tugas yang diberikan. Hal tersebut dikarenakan data pertama yang tidak relevan. Jika data yang ada tidak relevan, maka data tersebut tidak dapat diandalkan. Data yang ada tidak dapat diandalkan dikarenakan adanya data atau informasi yang sama (*redundant information*), data atau informasi tidak lengkap, ataupun data memuat error, maka dibutuhkannya *data preprocessing*, agar dapat mencapai kinerja yang terbaik dalam merepresentasikan suatu tugas.

Agushinta & Irfan (2008) menggunakan *preprocessing data* yaitu pembersihan data (*cleaning*) sebelum proses data mining pada penelitiannya dilaksanakan. Proses

pembersihan data dilakukan pada data yang menjadi fokus *knowledge discovery in database (*KDD). Proses pembersihan data mencakup mengenai membuang duplikasi data, memeriksa data yang inkonsistensi, dan memperbaiki kesalahan pada data. Selain itu juga dapat dilakukan proses *enrichment*. Proses *enrichment* merupakan proses "memperkaya" data yang sudah ada dengan data atau informasi lain yang relevan dan diperlukan untuk KDD. Data atau informasi lain yang relevan bisa diambil dari data atau informasi eksternal atau berasal dari luar.

Hidayat et al (2013) melakukan penelitian dengan memanfaatkan *preprocessing data* sebagai salah satu langkah yang digunakan untuk membuat sistem yang berguna untuk memprediksi faktor-faktor yang mempengaruhi mahasiswa *drop out* (DO). *Prepocessing data* dilakukan untuk memangani data yang bersifat *noise* dan *missing value data*. Data *noise* dan informasi yang tidak relevan dapat berkurang dengan melakukan pembersihan data dan pengintegrasian data. Pembersihan data (*cleaning data*) ataupun transformasi data merupakan salah satu bagian dari *preprocessing data*. Pembersihan data adalah suatu proses yang digunakan untuk menghapus data ganda, memeriksa data yang tidak konsisten, penanganan data *missing* (data yang hilang) dan merapikan data *noise*. Transformasi data adalah suatu proses mengubag data ke dalam model analitik.

Junedy (2014) menggunakan algoritma dari metode *levensthein distance* (*edit distance*) yang dapat digunakan untuk mendeteksi kemiripan dokumen teks. Hal tersebut dapat membantu dalam menentukan *plagiarisme*.

Sebelum menggunakan algoritma *levensthein distance* dilakukan proces *preprocessing* terhadap dokumen teks yang akan di proses. *Preprocessing* dilakukan untuk meningkatkan kerja algoritma. Dari uji coba yang dilakukan *preprocessing* terbukti dapat meningkatkan kinerja algoritma meskipun menambah waktu pemrosesan dokumen. Selain itu algoritma *levensthein distance* juga mampu mendeteksi kemiripan dokumen teks.

Adriyani et al (2012) menggunakan implementasi agoritma *levenshtein distance* untuk menampilkan saran perbaikan kesalahan penulisan dokumen berbahasa Indonesia. Kesalahan penulisan pada umumnya disebabkan karena beberapa hal, seperti kedekatan letak *keyboard*, slip jari ataupun karena dua karakter yang letaknya tertukar. Algoritma *levenshtein distance* digunakan untuk menghitung keterkaitan antar *string* dan menghitung jumlah keterbedaan antar dua *string*.

Berdasarkan hasil penelitian diatas, maka dibuatlah Sistem Pengelolaan Data Mahasiswa (SIPEMA) dengan menggunakan metode *levenshtein distance* (*edit distance*), yang mana data akan dikenakan proses pembersihan data pada *preprocessing data* karena adanya ketidakkonsistenan data dan adanya data yang tidak lengkap. Data mahasiswa yang ada dikelompokan berdasarkan sekolahnya. Pengelompokan data mahasiswa berdasarkan sekolahnya bertujuan untuk mempermudah dalam proses pencarian data mahasiswa berdasarkan asal sekolahnya. Diharapkan dari pemilihan metode yang ada, dapat membantu staf Kantor Kerjasama dan Promosi (KKP) yang bertugas.