BAB II
LANDASAN TEORI

2.1 Posisi Penelitian

Posisi penelitian ini merupakan referensi penelitian terlebih dahulu. Dengan referensi penelitian terdahulu menjadi pelajaran dalam menggunakan metode – metode dalam *data cleansing*. Sehingga pada penelitian ini diharapkan dapat menghasilkan data yang bersih dan siap untuk dilakukan proses *data mining.*

Penelitian terdahulu dilakukan oleh (Ambedkar & Babu, 2015) menggunakan *Remove Duplicates* pada RapidMiner untuk menghilangkan duplikasi data pada *training data set*. Dalam penelitan ini menyajikan analisis komprehensif tentang serangan *Probe*, dengan menerapkan berbagai teknik pembelajaran mesin populer seperti *Naïve Bayes, SVM, Multilayer Perceptron, Decision Trees* dll dan menghasilkan duplikasi data dapat diatasi sehingga memaksimalkan teknik Neural Net yang memberikan akurasi tertinggi 99,44%. AutoMLP memberikan akurasi tertinggi kedua yaitu 99,06%.

Sedangkan penelitian yang dilakukan oleh (Islam, Mamun, & Rahman, 2014) ialah menggunakan metode *Missing Value Imputation* guna menghilangkan data kosong yang terdapat pada koleksi data dari *Wireless Sensor Networks.* Penelitian ini menghasilkan bahwa dengan menggunakan metode *Missing Value Imputation* saat nilai yang hilang dimasukkan, lalu selanjutnya melakukan analisis *data mining* pada kumpulan data yang diperhitungkan dan menunjukkan peningkatan dalam akurasi prediksi dari pengklasifikasian yang dibangun dari kumpulan data yang diimputasi.

Selain itu (Zhou, Chen, Gao, & Guo, 2011) meneliti mengenai penyaringan data kotor pada bidang penelitian *Wireless Sensor Networks*. Data kotor ini disebabkan oleh *noise* lingkungan sekitarnya, gangguan yang disebabkan oleh perangkat keras itu sendiri, suhu lingkungan, dan di sisi lain kegagalan sensor simpul yang disebabkan oleh energi yang habis atau perangkat keras yang rusak akan menghasilkan data kesalahan abnormal / tidak konsisten. Maka penelitian ini menggunakan teknik penyaringan data kotor berdasarkan korelasi temporal-spasial. Teknik ini dapat mengurangi data transmisi untuk menghemat energi, memperpanjang siklus hidup jaringan, dan memastikan keakuratan data transmisi pada saat yang sama.

Lalu terdapat penelitian yang dilakukan oleh (Krishnan et al., 2015) telah mengembangkan proyek bernama *SampleClean* guna meminimalisir data yang mengalami duplikasi.

SampleClean mempelajari integrasi *Sample-based Approximate Query Processing* dan pembersihan data untuk memberi para analis membersihkan seluruh dataset. Hasil yang ditemukan adalah terdapat 72 query yang eror dengan menggunakan proyek *SampleClean.*

Penelitian terdahulu juga dilakukan oleh (Riezka et al., 2011). Dalam penelitian tersebut menggunakan metode *Multi-Pass Neighborhood* (MPN) dalam mengilangkan duplikasi data. Berdasarkan pengujian yang sudah dilakukan, metode tersebut menghasilkan nilai *recall* dan *false-positive* yang cukup baik dengan parameter ukuran lebar *window*, kombinasi *rule* dan jumlah *passes* yang digunakan. Berdasarkan pengujian yang sudah dilakukan, metode tersebut menghasilkan nilai *recall* dan *false-positive* yang cukup baik dengan parameter ukuran lebar *window*, kombinasi *rule* dan jumlah *passes* yang digunakan.

Tabel 2.1 Tabel perbandingan penelitian terdahulu

No	Judul	Tahap Penelitian	Metode Penelitian	Hasil yang didapatkan
1	*Detection of Probe Attacks Using Machine Learning Techniques* (Ambedkar & Babu, 2015)	Implementasi	*Remove Duplicates*	Duplikasi data dapat diatasi sehingga memaksimalkan teknik Neural Net yang memberikan akurasi tertinggi 99,44%. AutoMLP memberikan akurasi tertinggi kedua yaitu 99,06%
2	*Data Cleansing during Data Collection from Wireless Sensor Networks* (Islam et al., 2014)	Implementasi	*Missing Value Imputation*	Data kosong dapat diganti dan hasil yang didapatkan menunjukkan peningkatan dalam akurasi prediksi dari pengklasifikasian yang dibangun dari kumpulan data yang diimputasi
3	*A Technique of Filtering Dirty Data Based on Temporal-Spatial Correlation in Wireless Sensor Network* (Zhou et al., 2011)	Implementasi	Korelasi temporal-spasial	Berhasil dalam mengatasi data tidak konsisten sehingga dapat mengurangi data transmisi untuk menghemat energi, memperpanjang siklus hidup jaringan, dan memastikan keakuratan data transmisi pada saat yang sama
4	*SampleClean : Fast and Reliable Analytics on Dirty Data* (Krishnan et al., 2015)	Implementasi	Sample-based Approximate Query Processing	Sukses dalam mengatasi 72 *query* yang eror dengan menggunakan proyek SampleClean dan dapat membersihkan data yang mengalami duplikasi, data kosong, dan data tidak konsisten.
5	Analisis dan Implementasi Data-Cleaning dengan Menggunakan Metode Multi-Pass Neighborhood (Mpn) (Riezka et al., 2011)	Implementasi	Multi-Pass Neighborhood	Duplikasi data dapat diatasi dan menghasilkan nilai *recall* dan *false-positive* yang cukup baik dengan parameter ukuran lebar *window*, kombinasi *rule* dan jumlah *passes* yang digunakan.

2.2 Pengertian Data Kotor

Permasalahan pada data seringkali terjadi dalam kumpulan data tunggal seperti *file* dan *database.* Permasalahan tersebut seperti kesalahan ejaan, informasi yang hilang atau data tidak valid lainnya. Biasanya disebabkan oleh petugas dalam entri data, model data dan desain skema sistem yang berbeda, atau data yang tumpang tindih, kontradiktif, dan tidak konsisten. Beberapa sumber data perlu diintegrasikan karena sumber data sering berisi data yang berlebihan dalam representasi yang berbeda. Sehingga diperlukan untuk akses ke data yang akurat dan konsisten, konsolidasi berbagai representasi dan penghapusan data informasi duplikat (Rahm & DO, 2014).

2.3 Indikasi Data yang Berkualitas

Pemilihan metode yang cocok dalam mengatasi permasalahan pada data kotor sangat dibutuhkan dalam proses *data cleansing.* Pemilihan metode *data cleansing* digunakan agar mendapatkan data yang berkualitas. Data yang berkualitas harus memiliki indikasi sebagai berikut :

1. Akurasi: hasil bagi dari jumlah nilai yang benar dalam data koleksi dan jumlah keseluruhan nilai.
2. Validitas: Tingkat kepatuhan terhadap aturan atau batasan bisnis yang ditetapkan
3. Kelengkapan: Sejauh mana semua tindakan yang diperlukan diketahui
4. Konsistensi: Sejauh mana seperangkat tindakan setara di seluruh sistem
5. Keseragaman: Tingkat pengukuran data yang ditetapkan dengan menggunakan satuan ukuran yang sama di semua sistem (Müller & Freytag, 2003)

2.4 *Duplicate Elimination*

Metode ini menentukan apakah dua data atau lebih merupakan representasi rangkap dari entitas yang sama. Cara kerja metode ini ialah dengan membandingkan setiap data yang ada berdasarkan atribut yang ditentukan. Maka hasil didapatkan adalah berapa banyak data yang mengalami duplikasi data. Lalu metode ini dapat menghapus duplikasi data sehingga hanya satu dari semua data duplikat yang disimpan (Müller & Freytag, 2003).

2.5 *Incosistency Detection*

Penerapan metode ini dilakukan ketika data tersedia dari beberapa sumber yang berbeda dan dalam format yang berbeda. Tidak konsistennya data dapat disebabkan oleh kesalahan

manusia (*human error*) dalam pengentrian data (Aggarwal, 2015). Misalnya pada atribut NO_POLIS seharusnya diisikan dengan nomor 6 digit. Namun karena adanya kesalahan dalam pengentrian data, ada data yang disimpan dengan 4 digit atau diisi dengan "-". Metode ini akan menemukan anomali yang terjadi pada data dan mengubahnya menyesuaikan format yang telah ditentukan.

2.6 *Handling Missing Entries*

Dalam mengatasi data yang hilang (*missing entries*), metode ini dapat menghilangkan data apapun yang berisi entri yang hilang (Aggarwal, 2015). Semisal ada kasus pasien yang rawat inap pada suatu rumah sakit. Namun didata tersebut menunjukkan bahwa ID_BED tidak dientri sehingga data pasien tersebut dapat dihapus.

2.7 Kakas Analisis

Penelitian ini menggunakan perangkat lunak RapidMiner dalam melakukan data cleansing pada data sistem rumah sakit. Perangkat lunak RapidMiner adalah sebuah perangkat lunak yang memiliki kecerdasan buatan melalui platform sains data dan sebagai mesin data mining. RapidMiner dapat memberikan solusi dari persiapan data hingga pembelajaran mesin dan penerapan model prediksi (RapidMiner, 2019). Dalam penggunaan data mining, RapidMiner menyediakan prosedur data mining dan machine learning termasuk : ETL (*Extraction, Transformation, Loading*), *data preprocessing*, visualisasi, *modelling* dan evaluasi (Wicaksana, 2013).

RapidMiner memiliki keunggulan diantaranya:

a. Merupakan *freeware*. Sehingga tidak perlu mengeluarkan biaya lebih untuk instalasinya.
b. RapidMiner menggunakan bahasa pemrograman java sehingga dapat digunakan di berbagai sistem operasi.
c. Memungkinkan untuk eksperimen data dalam skala besar.
d. Dapat menampilkan grafis yang canggih seperti diagram batang, *3D Scatter plots*, dan lain – lain.
e. Memiliki beberapa *plug in* sehingga dapat menggunakan berbagai teknik deskriptif dan prediksi dalam memberikan wawasan kepada pengguna sehingga dapat membuat keputusan yang paling baik (Tamilselvi & Gifta, 2011).

RapidMiner digunakan oleh perusahaan besar seperti :

a. Deloitte Consulting LLP
b. Domino's Pizza Inc
c. SLALOM, LLC
d. InFocus Corporation
e. Boston Consulting Group (Enlyft, 2019)

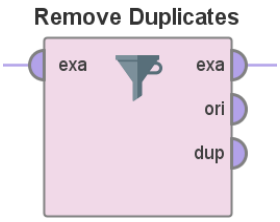
Fitur dalam RapidMiner yang akan digunakan :

1. *Read Excel*

Gambar 2.1 Fitur *Read Excel*

Pada aplikasi RapidMiner terdapat fitur Read Excel yang berfungsi sebagai pembaca file excel. Read Excel merupakan operator yang paling dasar digunakan sebelum memulai sebeuah proses. Operator ini dapat digunakan untuk memuat data dari spreadsheet Microsoft Excel. Operator ini membaca data dari Excel 95, 97, 2000, XP, dan 2003. Dalam penelitian ini, file excel yang digunakan berformat .xls.

2. *Remove Duplicates*

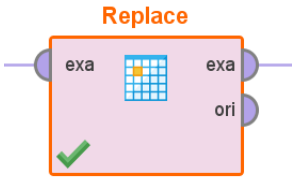


Gambar 2.2 Fitur *Remove Duplicates*

Fitur Remove Duplicates membandingkan keseluruhan dataset satu sama lain berdasarkan atribut yang ditentukan. Dua contoh data dianggap duplikat jika atribut yang dipilih

memiliki nilai yang sama di dalamnya. Operator ini menghapus contoh duplikat sehingga hanya satu dari semua contoh duplikat disimpan.

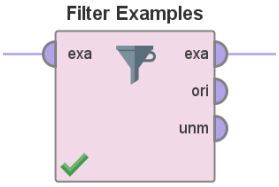
3. *Replace*



Gambar 2.3 Fitur *Replace*

Fitur Replace berfungsi untuk melakukan penggantian data yang terpilih dan menggantinya sesuai karakter yang akan menggantikannya. Nilai atribut dari atribut yang dipilih yang cocok dengan data yang dimasukkan nantinya akan digantikan oleh penggantian yang ditentukan.

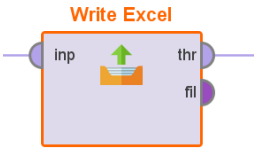
4. *Filter Example*



Gambar 2.4 Fitur *Filter Example*

Penelitian ini menggunakan fitur Filter Example dalam menghilangkan data kosong. *File* data yang di import ke dalam RapidMiner menggunakan fitur Read Excel dapat ditampilkan secara keseluruhan. Data yang terdapat pada file ini apabila memiliki data kosong (missing value) akan ditampilkan dengan tanda "?". Dengan fitur Filter Example ini, data yang bertanda "?" akan dihapus. Nilai pengisian apa pun juga dapat ditentukan sebagai pengganti nilai yang hilang.

5. *Write Excel*



Gambar 2.5 Fitur *Write Excel*

Fitur Write Excel dapat digunakan untuk menulis ExampleSets ke dalam file spreadsheet Microsoft Excel. Fitur ini membuat file Excel yang dapat dibaca oleh Excel 95, 97, 2000, XP, 2003 dan versi yang lebih baru. Dengan begitu hasil dari penelitian ini dapat di *export* menjadi *file excel*.