

PENERAPAN ALGORITMA ANONIMISASI PADA LINGKUNGAN BIG DATA

STEPHEN JORDAN—2016730018

1 Data Skripsi

Pembimbing utama/tunggal: **Mariskha Tri Adithia, P.D.Eng**

Pembimbing pendamping: **Dr. Veronica Sri Moertini, Ir., M.T.**

Kode Topik : **MTA4801**

Topik ini sudah dikerjakan selama : **1 semester**

Pengambilan pertama kali topik ini pada : **Semester 48 - Genap 19/20**

Pengambilan pertama kali topik ini di kuliah : **Skripsi 1**

Tipe Laporan : **B -** Dokumen untuk reviewer pada presentasi dan **review Skripsi 1**

2 Latar Belakang

Perkembangan penggunaan internet dan teknologi informasi mengakibatkan pertumbuhan data yang sangat besar dan terjadi secara terus-menerus sehingga data sulit untuk dikelola, diproses, maupun dianalisis menggunakan teknologi pengolahan data biasa. Data yang terus bertumbuh menyebabkan basis data konvensional menjadi kurang efektif untuk mengolah data. Teknologi saat ini telah menemukan sebuah cara untuk mengurangi biaya penyimpanan dan komputasi data, sehingga kapasitas data dapat ditingkatkan dan data menjadi lebih mudah diolah.

Big data adalah data dalam jumlah sangat besar dikumpulkan, disimpan, diolah, dan dianalisis agar menghasilkan informasi yang bermanfaat sebagai dasar pengambilan keputusan atau kebijakan. *Data mining* adalah teknik ekstraksi informasi terhadap sekumpulan data dalam jumlah besar. *Data mining* efektif menggantikan pemrosesan kueri pada basis data dengan data berukuran besar. Masalah yang umum terjadi adalah data yang tersimpan banyak mengandung data yang bersifat privasi sehingga perlu adanya perlindungan privasi pada data yang akan diolah.

Perlindungan privasi dicapai dengan metode enkripsi dan anonimisasi. Enkripsi adalah metode yang memanfaatkan pola atau kunci tertentu. Anonimisasi adalah metode yang menyamarkan satu atau lebih nilai atribut data. Pada kasus tertentu, keamanan enkripsi dapat ditembus melalui penalaran nilai atribut. Penalaran ini sangat berbahaya karena menghubungkan nilai atribut data secara tidak langsung, dapat mengungkapkan entitas pemilik data. Dengan menerapkan konsep anonimisasi diharapkan nilai keterhubungan antar atribut data dapat diperkecil.

Dengan melakukan anonimisasi pada sebagian nilai atribut, bobot informasi yang diperoleh akan semakin kecil. Permasalahan *K-anonymity* adalah pencarian solusi untuk menyeimbangkan nilai informasi yang diperoleh dengan nilai informasi yang disamarkan. Permasalahan *K-anonymity* diuji dengan pendekatan generalisasi dan supresi. Hasilnya dinilai kurang efektif karena tingginya jumlah informasi yang hilang. Berdasarkan penelitian, permasalahan *K-anonymity* tercapai melalui penerapan *K-member clustering*. Penerapan *K-member clustering* pada algoritma *Greedy K-member clustering* dinilai baik karena dapat meminimalkan jumlah informasi yang hilang.

Spark adalah *framework* yang tepat untuk memproses data dengan ukuran yang relatif besar seperti *big data*, dengan membagi data tersebut ke sistem terdistribusi. Penggunaan Spark menggeser penggunaan Map Reduce pada Hadoop yang dinilai cukup lambat. Kelebihan Spark memiliki proses komputasi yang lebih cepat karena sebagian besar pemrosesan Spark berada pada RAM. Selain itu, Spark mampu melakukan pemrosesan *data mining* menggunakan *library* tambahan Spark MLlib. Kekurangannya adalah Spark masih

tetap bergantung pada mekanisme penyimpanan Hadoop, agar hasil pemrosesan data dapat tersimpan di dalam *hardisk* komputer.

Pada skripsi ini, akan dibuat sebuah perangkat lunak yang dapat memproses data semi terstruktur menjadi data anonimisasi menggunakan konsep *K-anonymity*. Perangkat lunak ini berjalan di atas Spark untuk memudahkan proses anonimisasi pada lingkungan *big data*. Algoritma *Greedy K-member clustering* dinilai tepat untuk melakukan anonimisasi data karena meminimalkan jumlah informasi yang hilang saat proses *data mining* di penelitian sebelumnya. Penelitian ini bertujuan membandingkan hasil *data mining* sebelum dan setelah dilakukan anonimisasi.

3 Rumusan Masalah

Berdasarkan latar belakang di atas, rumusan masalah pada skripsi ini adalah sebagai berikut:

1. Bagaimana cara kerja algoritma *Greedy K-member clustering* ?
2. Bagaimana implementasi algoritma *Greedy K-member clustering* pada Spark?
3. Bagaimana hasil *data mining* sebelum dan setelah dilakukan anonimisasi?

4 Tujuan

Berdasarkan rumusan masalah di atas, tujuan pada skripsi ini adalah sebagai berikut:

1. Mempelajari cara kerja algoritma *Greedy K-member clustering*.
2. Mengimplementasikan algoritma *Greedy K-member clustering* pada Spark.
3. Menganalisis hasil *data mining* sebelum dan setelah dilakukan anonimisasi.

5 Detail Perkembangan Pengerjaan Skripsi

Detail bagian pekerjaan skripsi sesuai dengan rencana kerja/laporan perkembangan terakhir:

1. **Melakukan studi literatur mengenai konsep privasi, konsep PII, contoh informasi yang bersifat sensitif dan non-sensitif menurut standar PII**

Status : Ada sejak rencana kerja skripsi.

Hasil : Privasi adalah suatu keadaan dimana kehidupan pribadi seseorang atau sekelompok orang terbebas dari pengawasan atau gangguan orang lain. Privasi juga dapat berarti kemampuan satu atau sekelompok individu untuk menutupi atau melindungi kehidupan dan urusan personalnya dari publik dengan mengontrol sumber-sumber informasi mengenai diri mereka. Untuk melakukan publikasi data dari satu perusahaan ke perusahaan lain, digunakan teknik anonimisasi data untuk melindungi dan menyamarkan atribut sensitif untuk setiap data.

Personally Identifiable Information (PII) adalah standar yang digunakan untuk menentukan apakah informasi yang ada dapat melakukan identifikasi entitas individu secara langsung atau tidak langsung. PII menjelaskan bahwa identifikasi entitas secara langsung dapat dilakukan menggunakan atribut sensitif. Sedangkan identifikasi entitas secara tidak langsung dapat dilakukan menggunakan penggabungan beberapa atribut non-sensitif. PII adalah atribut yang biasanya terjadi pelanggaran data dan pencurian identitas. Jika data perusahaan atau organisasi terungkap, maka sangat mungkin data pribadi seseorang akan terungkap. Informasi yang diketahui dapat dijual dan digunakan untuk melakukan pencurian identitas, menempatkan korban dalam risiko.

Berikut adalah contoh informasi yang bersifat sensitif menurut standar PII:

- Identitas diri
Nama lengkap, tempat tanggal lahir, alamat rumah, alamat email.
- Nomor identitas diri
NIK, nomor passport, nomor SIM, nomor wajib pajak, nomor rekening, nomor telepon, dan nomor kartu kredit.
- Karakteristik pribadi
Foto diri, sidik jari, dan tulisan tangan.
- Data biometrik
Pemindaian retina, jenis suara, dan geometri wajah.
- Aset informasi lainnya
IP Address dan Media Access Control (MAC).

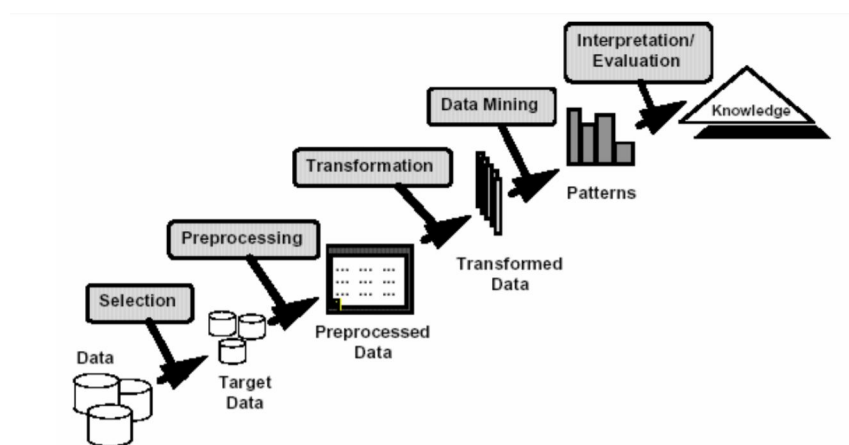
Berikut adalah contoh informasi yang bersifat non-sensitif menurut standar PII:

- Rekaman medis
- Riwayat pendidikan
- Riwayat pekerjaan
- Informasi finansial
- Letak geografis

2. Melakukan studi literatur mengenai teknik-teknik dasar data mining

Status : Ada sejak rencana kerja skripsi.

Hasil : Data yang dikumpulkan bertambah banyak, sehingga perlu adanya cara untuk melakukan proses ekstraksi informasi pada sekumpulan data yang sangat banyak. Menurut Gartner, Data Mining adalah proses menemukan korelasi, pola, dan tren baru yang bermakna dengan menyaring sejumlah besar data yang disimpan menggunakan teknologi pengenalan pola serta teknik statistik dan matematika. Data mining merupakan bagian dari Knowledge Discovery in Databases (KDD). KDD adalah proses transformasi sekumpulan data yang disimpan pada basis data menjadi informasi yang berguna.



Gambar 1: Tahapan pada KDD

Berikut ini adalah penjelasan tahapan pada KDD pada Gambar 1 sebagai berikut:

- (a) Selection: proses mengambil data yang relevan terhadap analisis.
- (b) Preprocessing: proses pembersihan data dari data yang tidak konsisten dan integrasi data saat penggabungan data.
- (c) Transformation: proses manipulasi data menggunakan konsep agregasi, generalisasi, normalisasi, dan reduksi untuk kebutuhan analisis.
- (d) Data Mining: proses ekstraksi informasi menggunakan metode pengenalan pola seperti klasifikasi, pengelompokan/clustering.
- (e) Interpretation/Evaluation: proses interpretasi hasil pengolahan data menjadi sebuah grafik yang dapat dimengerti.

Berikut adalah beberapa jenis tipe data terkait teknik data mining:

- Binary: tipe data alphabet/numerik yang hanya memiliki 2 kemungkinan nilai.
Contoh: diadakan survei evaluasi beberapa produk pakaian untuk mengetahui produk yang diminati dan tidak diminati. Penilaian produk dapat diwakilkan nilai True atau False. True atau False termasuk jenis binary.
- Nominal: tipe data alphabet/numerik yang memiliki lebih dari 2 kemungkinan nilai.
Contoh: seseorang memilih beberapa bahan dari warna yang berbeda. Warna yang mungkin adalah kuning, hijau, hitam, merah. Warna termasuk jenis nominal.

Tujuan dari penggunaan teknik data mining adalah sebagai berikut:

- Prediksi: proses menggunakan nilai dari beberapa atribut yang sudah ada untuk memprediksi nilai atribut di masa yang akan datang. Contoh: klasifikasi.
- Deskripsi: proses menemukan pola yang dapat merepresentasikan kelompok dari sebuah data. Contoh: pengelompokan/clustering.

(a) *Classification*

Tujuan utama *Classification* (klasifikasi) adalah membuat model yang dalam kasus ini disebut *classifier* yang mana dapat mengidentifikasi nilai kelas dari suatu data. Dalam kata lain, sebuah *classifier* dibuat dari sebuah *training set* dan model ini digunakan untuk mengklasifikasi data tidak diketahui ke dalam salah satu kelas. Ada dua tahap dalam proses klasifikasi yaitu tahap latihan dan tahap klasifikasi.

Pada tahap latihan, model akan dibuat dengan menggunakan *training set*. *Training set* yang dimaksud adalah data yang sudah diketahui kelasnya sehingga model yang ada melatih dirinya. Setelah *classifier* terbentuk, barulah tahap klasifikasi dapat dilakukan dengan menggunakan *classifier* yang tadi sudah dibuat. *Classifier* akan memprediksi data yang kelasnya tidak diketahui. *Classifier* akan semakin baik performanya seiring dengan banyaknya tahap latihan yang dilakukan.

Teknik *machine learning* yang paling dikenal untuk klasifikasi antara lain *K-nearest Neighbors*, *Decision Tree*, dan *Naive Bayes*. Dalam penelitian ini, hanya teknik *K-nearest Neighbors* yang digunakan untuk pengujian sehingga berikutnya hanya akan dijelaskan teknik *K-nearest Neighbors* saja.

Teknik *K-nearest Neighbors* adalah teknik penambangan data klasifikasi yang mencari label terbanyak pada sejumlah tetangga terdekatnya. Teknik ini bergantung pada jarak Euclidean antara titik yang mana adalah data yang akan diprediksi dengan tetangga-tetangganya. Setiap rekord

pada data dipetakan ke bidang Euclidean dengan beberapa atribut yang menentukan letaknya pada bidang Euclidean.

Berikut langkah kerja dari teknik *K-nearest Neighbors*.

- i. Tentukan nilai k yang menentukan seberapa banyak tetangga yang digunakan
- ii. Lakukan perulangan dengan iterasi sebanyak record yang ada selain record yang ingin diprediksi labelnya
 - A. Hitung jarak Euclidean antara record iterasi sekarang dengan record yang ingin diprediksi labelnya
 - B. Catat jarak Euclidean dari record yang ingin diprediksi dan indeks record iterasi sekarang
- iii. Urutkan jarak Euclidean titik-titik yang sudah dihitung pada perulangan pada langkah sebelumnya secara menaik
- iv. Pilih record teratas (jarak Euclidean yang paling kecil) sebanyak k dari urutan pada langkah sebelumnya
- v. Ambil label dari semua record yang terpilih pada langkah sebelumnya. Label terbanyak adalah hasil prediksi label pada record yang ingin diprediksi

(b) *Clustering*

Clustering adalah proses mengelompokkan kumpulan objek ke dalam sebuah kelompok (*cluster*) sedemikian rupa sehingga objek-objek dari suatu *cluster* memiliki lebih banyak kemiripan dari pada objek-objek dari *cluster* lainnya.

Salah satu contoh teknik *clustering* adalah *K-means*. Teknik *k-means* adalah teknik penambangan data *clustering* yang memanfaatkan jarak Euclidean antara titik-titik yang ada untuk menentukan titik mana saja yang masuk ke kluster mana.

Berikut langkah kerja dari teknik *K-means*.

- i. Tentukan nilai k yang menentukan seberapa banyak kluster yang diinginkan dan sebuah *threshold* untuk menentukan batas perubahan nilai centroid
- ii. Tentukan secara acak sebuah centroid sebanyak k untuk setiap kluster
- iii. Lakukan perulangan sampai nilai fitur-fitur semua centroid (titik tengah kluster) relatif tidak berubah atau dengan kata lain perubahannya kurang dari *threshold*
 - A. Menghitung jarak Euclidean tiap titik dari centroid ke titik tersebut dengan menggunakan beberapa fitur yang dipilih
 - B. Kluster yang memiliki jarak Euclidean paling kecil dengan sebuah titik adalah kluster titik tersebut
 - C. Tentukan kembali centroid setiap kluster dengan cara menghitung rata-rata tiap fitur seluruh data pada kluster tersebut

3. Melakukan studi literatur mengenai *privacy preserving data mining*

Status : Ada sejak rencana kerja skripsi.

Hasil : Aktivitas penambangan data melibatkan jumlah data yang sangat masif. Data-data yang digunakan memiliki privasi banyak individu di dalamnya. Hal ini berpotensi menyebabkan pelanggaran privasi dalam kasus tidak adanya proteksi yang cukup dan penyalahgunaan privasi data untuk tujuan lain. Faktor utama pelanggaran privasi pada penambangan data adalah penyalahgunaan data sehingga hal ini dapat merugikan seorang individu maupun sebuah organisasi. Oleh karena itu, ada kebutuhan untuk menghindari penyebaran informasi pribadi yang rahasia maupun pengetahuan lainnya yang dapat diambil dari data yang digunakan untuk aktivitas penambangan data.

Konsep privasi sering kali lebih kompleks dari pada yang dibayangkan. Dalam kasus penambangan data, definisi dari menjaga privasi masih tidak jelas. Ada sebuah paper yang mendefinisikan *privacy*

preserving data mining sebagai “getting valid data mining results without learning the underlying data values”, mendapatkan hasil penambangan data yang valid tanpa nilai pada data. Tetapi pada saat ini setiap teknik *privacy preserving data mining* yang ada memiliki definisi privasinya masing-masing.

Salah satu cara untuk melakukan *privacy preserving data mining* adalah dengan melakukan modifikasi data yang ada sebelum diberikan kepada pihak lain. Berbagai macam pendekatan modifikasi data untuk *privacy preserving data mining* telah dikembangkan antara lain *Perturbation Approach* dan *Anonymization Approach*, selengkapnya dapat dilihat pada Gambar ???. *Perturbation Approach* adalah pendekatan untuk *privacy preserving data mining* dengan cara mengacaukan data yang ada, tetapi hasil data yang dikacaukan masih tetap dapat ditambang. Sedangkan pada *Anonymization Approach*, data diterapkan de-identifikasi di mana dataset mentah disebarluaskan setelah menghapus inti dari identitas setiap record.

Perturbation Approach dapat dibagi menjadi dua jenis lagi yaitu *Value-based Perturbation Techniques* dan *Multi-Dimensional Perturbation*. *Value-based Perturbation Techniques* adalah teknik yang bekerja dengan cara menyisipkan *random noise* pada data. Sedangkan terdapat dua jenis teknik *Multi-Dimensional Perturbation* yaitu *Data mining Task-based Perturbation* dan *Dimension Reduction-based Perturbation*. *Data mining Task-based Perturbation* adalah teknik yang bekerja dengan cara modifikasi data sehingga properti yang bertahan pada data yang telah dimodifikasi spesifik hanya properti yang digunakan oleh suatu teknik penambangan data tertentu. Sedangkan *Dimension Reduction-based Perturbation* adalah teknik yang bekerja dengan cara modifikasi data sekaligus mengurangi dimensi dari data asli.

Hal yang sering kali diperhatikan pada teknik-teknik *Perturbation Approach* adalah perbandingan antara jumlah privasi yang hilang dan jumlah informasi yang hilang. Idealnya teknik *Perturbation Approach* yang baik adalah teknik yang fokus meminimalkan jumlah privasi yang hilang dan jumlah informasi yang hilang sehingga hasil penambangan dan akurasi sama baiknya dengan tanpa menerapkan teknik *Perturbation Approach*. Setiap teknik penambangan data memakai properti yang berbeda-beda pada data yang ditambang. Oleh karena itu, properti yang terjaga pun sebaiknya berdasarkan properti yang digunakan pada teknik penambangan data yang digunakan. Pada saat ini, teknik modifikasi data yang ada sering kali memiliki perbedaan pada properti-properti yang terjaga. Teknik-teknik modifikasi data tertentu sering kali memiliki fungsi yang berbeda atau teknik penambangan data yang dapat digunakan berbeda karena properti yang terjaga pada teknik-teknik tersebut berbeda juga.

4. Melakukan studi literatur mengenai metode *Randomization*

Status : Ada sejak rencana kerja skripsi.

Hasil : Dari berbagai macam teknik modifikasi data untuk *privacy preserving data mining* yang dapat dilihat pada Gambar ??, terdapat empat teknik yang menggunakan metode *Randomization* yaitu *Random Noise Addition*, *Randomized Response*, *Random Rotation Perturbation*, dan *Random Projection Perturbation*.

Berbagai macam teknik dengan metode randomisasi umumnya menerapkan merusak nilai pada data. Salah satu teknik yang pertama kali menggunakan metode randomisasi untuk *privacy preserving data mining* adalah teknik *Random Noise Addition* yang dikemukakan oleh Agrawal dan Srikant pada paper berikut. Teknik *Random Noise Addition* ini dilakukan dengan cara menambahkan nilai random (*noise*) pada data. Nilai random tersebut diambil dari sebuah distribusi. Untuk menambang data yang telah ditambahkan *noise* ini perlu dilakukan rekonstruksi distribusi untuk mendapatkan distribusi yang asli. Oleh karena itu, teknik *Random Noise Addition* ini hanya menjaga distribusi data asli sehingga hanya teknik penambangan data yang bergantung pada distribusi data saja yang dapat digunakan. Penyesuaian pada algoritma penambangan data yang digunakan pun perlu dilakukan agar teknik *Random Noise Addition* ini dapat digunakan dan mendapatkan hasil penambangan data yang

hampir sama dengan tanpa menggunakan teknik *Random Noise Addition*.

Setelah teknik *Random Noise Addition* ditemukan, berbagai macam teknik lain pun dikembangkan terinspirasi dari teknik *Random Noise Addition* ini. Teknik *Random Rotation Perturbation* dan *Random Projection Perturbation* adalah teknik adalah salah satunya, tetapi teknik tersebut tidak dilakukan dengan cara menambahkan *noise* melainkan mengkalikan data asli dengan nilai random. Bagaimanapun juga, inti dari teknik-teknik randomisasi yang telah disebutkan di atas masih sama yaitu merusak data sehingga data yang dirilis bukanlah data asli melainkan data yang sudah rusak sehingga data yang dirilis tidak mengandung privasi dan privasi pun terjaga.

5. Melakukan studi literatur dan mempelajari teknik *Random Noise Addition*

Status : Ada sejak rencana kerja skripsi tetapi tidak dilanjutkan.

Hasil : Ide utama dari teknik *Random Noise Addition* adalah mendistorsi nilai pada data dengan cara menambahkan *random noise* yang diambil dari distribusi *Uniform* atau *Gaussian* dan memiliki rata-rata bernilai 0. Tetapi menurut penelitian yang telah dilakukan, distribusi *Gaussian* lebih baik digunakan untuk teknik ini. *Random noise* yang digunakan memiliki nilai yang berbeda untuk setiap nilai pada data.

Dengan teknik *Random Noise Addition*, dari data yang sudah didistorsi bisa didapatkan kembali distribusi data asli dengan merekonstruksi distribusinya tanpa mendapatkan setiap nilai-nilai yang ada pada data asli. Metode rekonstruksi yang digunakan berdasarkan pada aturan *Bayes*. Algoritma rekonstruksi untuk mendapatkan distribusi dari data asli dapat dilihat pada Gambar ??.

Algoritma ini berhenti sampai kriteria berhentinya terpenuhi. Kriteria tersebut adalah perbedaan estimasi distribusi iterasi sekarang dengan yang sebelumnya sangat kecil. Algoritma ini akan menghasilkan estimasi distribusi data asli dengan menggunakan data yang telah terdistorsi tanpa menggunakan nilai-nilai pada data asli, sehingga nilai-nilai pada data asli tidak tersebar. Oleh karena teknik *Random Noise Addition* hanya menjaga distribusi pada data maka teknik penambahan data yang dapat digunakan hanya teknik-teknik yang bergantung pada distribusi data saja.

Modifikasi pada algoritma penambahan data yang digunakan pun perlu dilakukan. Contohnya apabila algoritma pohon keputusan digunakan, maka perlu modifikasi pada algoritma pohon keputusan tersebut. Hal ini menimbulkan masalah pada aplikasi pada dunia nyata karena tidak efisien dan memakan waktu untuk memodifikasi setiap algoritma yang ingin digunakan untuk menyesuaikan dengan teknik *Random Noise Addition*. Masalah mengenai algoritma yang dapat digunakan pun menjadi perhatian karena teknik *Random Noise Addition* hanya dapat digunakan untuk algoritma yang bergantung pada distribusi saja sedangkan teknik randomisasi lain tidak menjaga distribusi pada data. Ada juga penelitian yang mengatakan bahwa teknik *Random Noise Addition* ini memiliki kualitas yang kurang baik dalam menjaga privasi data karena banyaknya celah yang dapat diserang pada teknik ini. Oleh karena masalah-masalah tersebut, akhirnya teknik ini pun tidak akan digunakan untuk diuji kualitas hasilnya. Teknik *Random Projection Perturbation* akan digunakan untuk menggantikan teknik *Random Noise Addition*.

6. Melakukan studi literatur dan mempelajari teknik *Random Rotation Perturbation*

Status : Ada sejak rencana kerja skripsi.

Hasil : Ide utama dari teknik *Random Rotation Perturbation* adalah jika data direpresentasikan sebagai matrix $X_{n \times d}$, *rotation perturbation* dari dataset X didefinisikan sebagai berikut.

$$G(X) = X_{n \times d} R_{d \times d} \quad (1)$$

Dimana $R_{d \times d}$ adalah *random rotation matrix*. *Random rotation matrix* berukuran d dimensi dapat dibuat dengan cara membuat matriks *special orthogonal* acak karena matriks rotasi memiliki sifat

special orthogonal. Matriks *special orthogonal* adalah matriks yang memiliki sifat *orthogonal* dan determinannya bernilai +1, yang mana matriks *orthogonal* adalah matriks yang menghasilkan matriks identitas apabila dikalikan dengan transposenya sendiri. Matriks rotasi ini dapat dibuat secara efisien mengikuti distribusi Haar. Dari definisi di atas dapat disimpulkan transformasi rotasi tersebut menjaga jarak Euclidean.

Teknik ini menjaga beberapa properti pada data antara lain yaitu jarak Euclidean, *inner product*, dan *geometric shape hyper* pada bidang multi-dimensi. Oleh karena itu, beberapa teknik penambangan data tidak berpengaruh (dapat digunakan) terhadap teknik *Random Rotation Perturbation* antara lain yaitu *K-nearest Neighbors*, *Support Vector Machines*, dan *Perceptrons*. Teknik ini dipercaya dapat memberikan hasil penambangan yang maksimal, hasil penambangan data yang telah dirusak persis sama dengan hasil penambangan data aslinya. Sehingga jumlah informasi yang hilang tidak ada, tetapi jumlah privasi yang hilangnya tinggi. Walaupun demikian ada beberapa penelitian yang mengatakan bahwa karena teknik *Random Rotation Perturbation* ini memiliki sifat demikian sehingga teknik ini dikatakan tidak aman dan dapat diserang dengan beberapa teknik untuk mendapatkan data asli yang lengkap.

Transformasi translasi juga perlu dilakukan agar rotasi yang dilakukan merusak data secara menyeluruh. Apabila tidak dilakukan translasi, nilai pada data yang mendekati nilai nol akan menghasilkan nilai yang mendekati nol juga setelah dirotasi. Implikasi dari hal tersebut adalah lemahnya dalam menjaga privasi. Translasi dapat dilakukan dengan cara membuat matriks translasi yang acak lalu kalikan dengan matriks data asli. Translasi dapat dilakukan karena translasi tidak mengubah properti geometris dari matriks yang ditranslasi sehingga jarak Euclidean dan properti lainnya pun terjaga dan hasil penambangan data pun tetap sama.

7. Melakukan studi literatur dan mempelajari teknik *Random Projection Perturbation*

Status : Ditambahkan untuk menggantikan teknik *Random Noise Addition*.

Hasil : Ide utama dari teknik *Random Projection Perturbation* adalah mereduksi dimensi dari representasi matriks data asli dengan syarat dimensi matriks tersebut cukup besar. Dasar dari teknik *Random Projection Perturbation* berdiri pada *Johnson-Lindenstrauss Lemma*.

Lemma 1 (JOHNSON-LINDENSTRAUSS LEMMA) *For any $0 < \epsilon < 1$ and any integer s , let k be a positive integer such that $k \geq 4(\epsilon^2/2 - \epsilon^3/3)^{-1} \ln n$. Then, for any set S of $s = |S|$ data points in \mathbb{R}^m , there is a map $f : \mathbb{R}^m \rightarrow \mathbb{R}^k$ such that, for all $x, y \in S$, $(1 - \epsilon)s||u - v||^2 < ||p(u) - p(v)||^2 < (1 + \epsilon)s||u - v||^2$, where $||\cdot||$ denotes the vector 2-norm.*

Inti dari Lemma ini menunjukkan bahwa titik pada bidang Euclidean d -dimensi dapat diproyeksikan ke bidang Euclidean berdimensi lebih kecil dari d , sedemikian rupa sehingga jarak antara dua titik tetap konsisten dengan *error* yang terkontrol tetapi dengan syarat d harus cukup besar. Oleh karena adanya *error* yang muncul, properti-properti pada data pun relatif sedikit berubah dan hal ini menyebabkan akurasi pada model yang dibuat dengan data tersebut berkurang dibandingkan data aslinya.

Projection perturbation dari dataset X didefinisikan sebagai berikut.

$$G(X) = X_{n \times d} R_{d \times k} \quad (2)$$

Dimana $R_{d \times k}$ adalah *random projection matrix* yang dihasilkan mengikuti distribusi normal, dengan rata-rata bernilai 0 dan standar deviasi bernilai $1/\sqrt{k}$. Ukuran matriks $R_{d \times k}$ disesuaikan dengan matriks $X_{n \times d}$ yang mana dataset asli dengan jumlah rekord n dan jumlah atribut d , yang mana d akan menjadi dimensi matriks. Oleh karena yang ingin dilakukan adalah reduksi dimensi maka k harus lebih kecil dari pada d , yang mana k adalah dimensi dari matriks baru yang dihasilkan dari *Random Projection Perturbation* ini.

Jika *random projection matrix* yang digunakan dihasilkan secara acak saja, hasil dari *random projection perturbation* akan terlalu merusak nilai pada data sehingga akurasi pada model yang akan dibuat kemungkinan berkurang drastis. Cara menanggulangi hal tersebut adalah menggunakan matriks *orthogonal* sebagai *random projection matrix*. Tetapi membuat matriks *orthogonal* yang berdimensi tinggi memiliki kompleksitas yang tinggi sehingga memerlukan *cost* yang besar. Pada observasi yang dilakukan Hecht-Neilsen menunjukkan bahwa “*that in a high-dimensional space, vectors with random directions are almost orthogonal*”. Dapat disimpulkan bahwa dalam kasus matriks berdimensi tinggi apabila sebuah matriks dihasilkan secara acak mengikuti suatu distribusi, matriks tersebut akan kurang lebih hampir *orthogonal*. Oleh karena itu, matriks yang dibuat untuk *Random Projection Perturbation* cukup matriks acak yang mengikuti suatu distribusi saja.

Menurut *Johnson-Lindenstrauss Lemma*, reduksi dimensi pada matriks berdimensi tinggi minimal berdimensi k , yang mana k didefinisikan sebagai berikut.

$$k \geq 4(\epsilon^2/2 - \epsilon^3/3)^{-1} \ln n \quad (3)$$

Sebuah matriks yang akan diproyeksikan ke dimensi yang lebih kecil akan memiliki nilai *error* pada jarak Euclidean yang dimiliki oleh titik-titik (setiap elemen dari matriks) pada bidang Euclidean tersebut. Nilai *error* tersebut ditentukan oleh variabel ϵ , yang mana ϵ menjadi ukuran seberapa baik proyeksi dilakukan. Semakin kecil nilai ϵ maka semakin besar k , yang mana k adalah dimensi minimal matriks yang dihasilkan. Semakin titik-titik pada bidang Euclidean diproyeksikan ke dimensi lebih kecil, semakin besar kerusakan yang timbul pada jarak Euclidean titik-titik tersebut.

Persamaan berikut menyatakan rentang *error* yang terjadi pada *Random Projection Perturbation* dengan ϵ (*eps*) yang ditentukan berada pada rentang $(0, 1)$.

$$(1 - \epsilon) \|u - v\|^2 < \|p(u) - p(v)\|^2 < (1 + \epsilon) \|u - v\|^2 \quad (4)$$

Pada hasil proyeksi, jarak Euclidean antara suatu titik dengan suatu titik lainnya dapat dipastikan berada pada rentang tersebut dan tidak akan melebihi *error* yang ditentukan.

8. Melakukan analisis terhadap *privacy preserving data mining* dan metode randomisasi

Status : Ada sejak rencana kerja skripsi.

Hasil : Privasi yang perlu dijaga antara lain mengenai identitas seseorang atau hal yang dapat dikaitkan terhadap identitas seseorang. Pada data yang digunakan untuk penambangan data sangat banyak sekali data privasi tersebut sehingga perlu adanya cara untuk menjaga privasi tersebut. Metode randomisasi dapat digunakan untuk menghilangkan privasi pada data tetapi masih dapat dilakukan penambangan data. Metode yang dipilih untuk diimplementasikan adalah teknik *Random Rotation Perturbation* dan *Random Projection Perturbation*. Tetapi ada kekurangan pada kedua teknik tersebut. kekurangan tersebut adalah nilai setiap fitur yang ada pada data harus bersifat numerik dan kedua teknik tersebut hanya menjaga jarak Euclidean sehingga hanya teknik penambangan data yang bergantung pada jarak Euclidean saja yang dapat digunakan.

9. Melakukan analisis terhadap teknik *Random Rotation Perturbation*

Status : Ada sejak rencana kerja skripsi.

Hasil : Algoritma *Random Rotation Perturbation* memiliki beberapa langkah yaitu sebagai berikut.

- (a) Dataset yang memiliki atribut sebanyak d dan record sebanyak n direpresentasikan dalam bentuk matriks berukuran $n \times d$
- (b) Buatlah matriks translasi acak yang diambil mengikuti distribusi *uniform* dengan rentang $[0, 100]$ berdimensi $(d + 1) \times (d + 1)$

- (c) Untuk keperluan transformasi translasi, matriks dataset perlu ditambahkan sebuah kolom dengan nilai 1 pada seluruh barisnya.
- (d) Lakukan transformasi translasi dengan cara mengkalikan matriks dataset dengan matriks translasi yang telah dibuat pada langkah kedua
- (e) Oleh karena keperluan transformasi translasi, hasil translasi akan berupa matriks berdimensi $n \times (d + 1)$ dengan kolom terakhir berisi nilai 1 pada setiap barisnya. Oleh karena itu, kolom tersebut perlu dibuang agar dimensi matriks dataset kembali sesuai aslinya
- (f) Buatlah *random rotation matrix* dengan membuat matriks *orthogonal* acak. Matriks *orthogonal* memiliki sifat yaitu determinannya sebesar 1 dan hasil perkalian matriks tersebut dengan transposenya adalah matriks identitas
- (g) Lakukan transformasi rotasi dengan cara mengkalikan matriks dataset dengan *random rotation matrix* yang telah dibuat pada langkah keenam
- (h) Hasil matriks yang telah dirotasi sudah dapat langsung digunakan untuk penambahan data

10. Melakukan analisis terhadap teknik *Random Projection Perturbation*

Status : Ada sejak rencana kerja skripsi.

Hasil : Algoritma *Random Projection Perturbation* memiliki beberapa langkah yaitu sebagai berikut.

- (a) Dataset yang memiliki atribut sebanyak d dan rekord sebanyak n direpresentasikan dalam bentuk matriks berukuran $n \times d$
- (b) Tentukan nilai ϵ (epsilon) yang diinginkan dan berada pada rentang $(0, 1)$
- (c) Hitung nilai k (dimensi minimal) dengan rumus berikut $k \geq 4(\epsilon^2/2 - \epsilon^3/3)^{-1} \ln n$
- (d) Tentukan nilai k yang diinginkan atau tentukan secara acak dengan persyaratan pada langkah ketiga terpenuhi dan k harus lebih kecil dari d
- (e) Buatlah matriks proyeksi dengan cara membuat matriks acak yang diambil mengikuti distribusi normal dengan rata-rata bernilai 0 dan standar deviasi bernilai $1/\sqrt{k}$ berdimensi $d \times k$
- (f) Lakukan proyeksi dengan cara mengkalikan matriks dataset dengan matriks proyeksi yang telah dibuat pada langkah kelima
- (g) Hasil matriks yang telah diproyeksi sudah dapat langsung digunakan untuk penambahan data

11. Melakukan studi kasus terhadap teknik *Random Rotation Perturbation*

Status : Ada sejak rencana kerja skripsi.

Hasil : Untuk lebih memahami bagaimana cara kerja teknik *Random Rotation Perturbation*, studi kasus dilakukan pada dataset *iris*, tetapi untuk memudahkan perhitungan pada studi kasus ini data yang dipakai hanya sebagian kecil saja dari seluruh data pada dataset *iris*. Data tersebut dapat dilihat pada Tabel 1. Dataset *iris* adalah dataset yang berisi data tentang korelasi antara ukuran bunga dengan spesiesnya. Dataset ini memiliki empat buah fitur dan satu buah label. Fitur-fitur pada dataset *iris* adalah kolom *sepal_length*, *sepal_width*, *petal_length*, dan *petal_width*. Label pada dataset *iris* adalah kolom *species*

Berikut langkah-langkah teknik *Random Rotation Perturbation* yang diaplikasikan pada dataset *iris* pada Tabel 1.

- (a) Fitur-fitur pada dataset tersebut yang berbentuk tabel akan direpresentasikan sebagai matriks.

Tabel 1: Tabel dataset *iris* yang digunakan sebagai contoh kasus

sepal_length	sepal_width	petal_length	petal_width	species
5.1	3.5	1.4	0.2	setosa
4.9	3	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa
4.6	3.4	1.4	0.3	setosa
5	3.4	1.5	0.2	setosa
4.4	2.9	1.4	0.2	setosa

Labelnya tidak diikutsertakan

$$\begin{bmatrix} 5.1 & 3.5 & 1.4 & 0.2 \\ 4.9 & 3 & 1.4 & 0.2 \\ 4.7 & 3.2 & 1.3 & 0.2 \\ 4.6 & 3.1 & 1.5 & 0.2 \\ 5 & 3.6 & 1.4 & 0.2 \\ 5.4 & 3.9 & 1.7 & 0.4 \\ 4.6 & 3.4 & 1.4 & 0.3 \\ 5 & 3.4 & 1.5 & 0.2 \\ 4.4 & 2.9 & 1.4 & 0.2 \end{bmatrix}_{9 \times 4}$$

- (b) Membuat matriks translasi yang diambil mengikuti distribusi uniform dengan rentang $[0,100]$ dengan dimensi sesuai dimensi matriks dataset

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 71.35281261 & 93.96479736 & 77.16763568 & 27.88189356 & 1 \end{bmatrix}_{5 \times 5}$$

- (c) Untuk keperluan translasi, matriks dataset ditambahkan sebuah kolom dengan nilai 1 pada setiap barisnya

$$\begin{bmatrix} 5.1 & 3.5 & 1.4 & 0.2 & 1 \\ 4.9 & 3 & 1.4 & 0.2 & 1 \\ 4.7 & 3.2 & 1.3 & 0.2 & 1 \\ 4.6 & 3.1 & 1.5 & 0.2 & 1 \\ 5 & 3.6 & 1.4 & 0.2 & 1 \\ 5.4 & 3.9 & 1.7 & 0.4 & 1 \\ 4.6 & 3.4 & 1.4 & 0.3 & 1 \\ 5 & 3.4 & 1.5 & 0.2 & 1 \\ 4.4 & 2.9 & 1.4 & 0.2 & 1 \end{bmatrix}_{9 \times 5}$$

- (d) Dilakukan transformasi translasi dengan matriks translasi yang telah dibuat pada langkah sebe-

lunnya dengan cara mengkalikan matriks dataset dengan matriks translasi

$$\begin{bmatrix} 5.1 & 3.5 & 1.4 & 0.2 & 1 \\ 4.9 & 3 & 1.4 & 0.2 & 1 \\ 4.7 & 3.2 & 1.3 & 0.2 & 1 \\ 4.6 & 3.1 & 1.5 & 0.2 & 1 \\ 5 & 3.6 & 1.4 & 0.2 & 1 \\ 5.4 & 3.9 & 1.7 & 0.4 & 1 \\ 4.6 & 3.4 & 1.4 & 0.3 & 1 \\ 5 & 3.4 & 1.5 & 0.2 & 1 \\ 4.4 & 2.9 & 1.4 & 0.2 & 1 \end{bmatrix} \times \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 71.35281261 & 93.96479736 & 77.16763568 & 27.88189356 & 1 \end{bmatrix}$$

(e) Berikut adalah hasil translasi pada matriks dataset

$$\begin{bmatrix} 76.45281261 & 97.46479736 & 78.56763568 & 28.08189356 & 1 \\ 76.25281261 & 96.96479736 & 78.56763568 & 28.08189356 & 1 \\ 76.05281261 & 97.16479736 & 78.46763568 & 28.08189356 & 1 \\ 75.95281261 & 97.06479736 & 78.66763568 & 28.08189356 & 1 \\ 76.35281261 & 97.56479736 & 78.56763568 & 28.08189356 & 1 \\ 76.75281261 & 97.86479736 & 78.86763568 & 28.28189356 & 1 \\ 75.95281261 & 97.36479736 & 78.56763568 & 28.18189356 & 1 \\ 76.35281261 & 97.36479736 & 78.66763568 & 28.08189356 & 1 \\ 75.75281261 & 96.86479736 & 78.56763568 & 28.08189356 & 1 \end{bmatrix}_{9 \times 5}$$

(f) Berikutnya matriks rotasi dibuat dengan cara membuat matriks spesial orthogonal yang berdimensi sesuai dimensi matriks dataset. Matriks rotasi berikut dibuat dengan *library* Scipy pada bahasa pemrograman Python

$$\begin{bmatrix} -0.45126938 & -0.70425922 & 0.32389616 & 0.44211556 \\ -0.43989334 & 0.70728617 & 0.39249528 & 0.39011226 \\ -0.17797534 & 0.06110969 & -0.83056872 & 0.52416218 \\ 0.75576092 & 0.00555185 & 0.22626167 & 0.61449187 \end{bmatrix}_{4 \times 4}$$

(g) Dilakukan transformasi rotasi dengan matriks rotasi yang telah dibuat pada langkah sebelumnya dengan cara mengkalikan matriks dataset dengan matriks rotasi

(h) Berikut adalah hasil rotasi pada matriks dataset. Hasil teknik *Random Rotation Perturbation* pada dataset *iris* ini sudah dapat langsung digunakan untuk dilakukan penambahan data

$$\begin{bmatrix} -70.13483265 & 20.05005561 & 4.11528068 & 130.26146931 \\ -69.8246321 & 19.83726437 & 3.85425381 & 129.97799007 \\ -69.80455936 & 20.11346248 & 3.95103051 & 129.91517319 \\ -69.75103816 & 20.12538172 & 3.71327762 & 129.93678284 \\ -70.13369505 & 20.19121015 & 4.12214059 & 130.25626898 \\ -70.34841122 & 20.14113559 & 4.16552936 & 130.83029591 \\ -69.78963253 & 20.33201179 & 3.93670924 & 130.06284949 \\ -70.06351391 & 20.05586388 & 3.96058467 & 130.23066274 \\ -69.55500808 & 20.11866536 & 3.65305621 & 129.71792106 \end{bmatrix}_{9 \times 4}$$

12. Melakukan studi kasus terhadap teknik *Random Projection Perturbation*

Status : Ada sejak rencana kerja skripsi.

Hasil : Untuk lebih memahami bagaimana cara kerja teknik *Random Projection Perturbation*, studi kasus dilakukan pada dataset *iris*. Teknik *Random Projection Perturbation* memiliki persyaratan pada dataset agar teknik ini menghasilkan hasil yang baik yaitu dataset tersebut harus memiliki dimensi yang cukup besar. Sebetulnya dataset *iris* tersebut tidak memenuhi persyaratan untuk mendapatkan hasil yang baik, tetapi untuk memudahkan perhitungan pada studi kasus ini data yang dipakai adalah dataset *iris* yang memiliki dimensi yang kecil dan hanya sebagian kecil saja data yang dipakai. Data tersebut dapat dilihat pada Tabel 1. Dalam menghitung nilai k juga, pada studi kasus ini menggunakan jumlah rekord dan atribut yang tidak sesuai dengan dataset *iris* untuk keperluan kemudahan dalam melakukan studi kasus dan juga agar memenuhi persyaratan teknik *Random Projection Perturbation*. Jumlah rekordnya adalah 1000 dan jumlah atributnya adalah 500.

Berikut langkah-langkah teknik *Random Projection Perturbation* yang diaplikasikan pada dataset *iris* pada Tabel 1.

- (a) Fitur-fitur pada dataset tersebut yang berbentuk tabel akan direpresentasikan sebagai matriks. Labelnya tidak diikutsertakan

$$\begin{bmatrix} 5.1 & 3.5 & 1.4 & 0.2 \\ 4.9 & 3 & 1.4 & 0.2 \\ 4.7 & 3.2 & 1.3 & 0.2 \\ 4.6 & 3.1 & 1.5 & 0.2 \\ 5 & 3.6 & 1.4 & 0.2 \\ 5.4 & 3.9 & 1.7 & 0.4 \\ 4.6 & 3.4 & 1.4 & 0.3 \\ 5 & 3.4 & 1.5 & 0.2 \\ 4.4 & 2.9 & 1.4 & 0.2 \end{bmatrix}_{9 \times 4}$$

- (b) Ditentukan nilai ϵ (epsilon) yang diinginkan adalah 0.5
(c) Nilai k (dimensi minimal) dihitung dengan rumus berikut

$$\begin{aligned} k &= \frac{4 \ln n}{\frac{\epsilon^2}{2} - \frac{\epsilon^3}{3}} \\ &= \frac{4 \ln 1000}{\frac{0.5^2}{2} - \frac{0.5^3}{3}} \\ &= \frac{27.63}{0.125 - 0.041666} \\ &= 331.57 \end{aligned}$$

- (d) Nilai k dipilih sesuai keinginan, dalam kasus ini dipilih nilai k sebesar 332
(e) Membuat matriks proyeksi dengan cara membuat matriks acak yang diambil mengikuti distribusi normal dengan rata-rata bernilai 0 dan standar deviasi bernilai $1/\sqrt{k}$ berdimensi $d \times k$. Untuk keperluan kemudahan dalam melakukan studi kasus, dataset *iris* akan direduksi dimensinya menjadi 3 dimensi

$$\begin{bmatrix} 0.11483014 & -0.10167359 & 0.06652355 \\ 0.0638684 & -0.1499892 & 0.10146435 \\ -0.10429573 & 0.03839861 & 0.04955419 \\ -0.0315941 & -0.06905021 & -0.17782438 \end{bmatrix}_{4 \times 3}$$

- (f) Dilakukan proyeksi dengan cara mengkalikan matriks dataset dengan matriks proyeksi yang telah

dibuat pada langkah sebelumnya

$$\begin{bmatrix} 5.1 & 3.5 & 1.4 & 0.2 \\ 4.9 & 3 & 1.4 & 0.2 \\ 4.7 & 3.2 & 1.3 & 0.2 \\ 4.6 & 3.1 & 1.5 & 0.2 \\ 5 & 3.6 & 1.4 & 0.2 \\ 5.4 & 3.9 & 1.7 & 0.4 \\ 4.6 & 3.4 & 1.4 & 0.3 \\ 5 & 3.4 & 1.5 & 0.2 \\ 4.4 & 2.9 & 1.4 & 0.2 \end{bmatrix} \times \begin{bmatrix} 0.11483014 & -0.10167359 & 0.06652355 \\ 0.0638684 & -0.1499892 & 0.10146435 \\ -0.10429573 & 0.03839861 & 0.04955419 \\ -0.0315941 & -0.06905021 & -0.17782438 \end{bmatrix}$$

- (g) Berikut adalah hasil matriks yang telah diproyeksi. Hasil dari teknik *Random Projection Perturbation* pada dataset *iris* ini sudah dapat langsung digunakan untuk dilakukan penambahan data

$$\begin{bmatrix} 0.65684027 & -1.0035495 & 0.72820632 \\ 0.60194004 & -0.90822018 & 0.66416944 \\ 0.60217727 & -0.92172316 & 0.66620218 \\ 0.56344827 & -0.88887716 & 0.65931422 \\ 0.6517441 & -1.00838106 & 0.7317004 \\ 0.67922914 & -1.09633771 & 0.76805051 \\ 0.58987895 & -0.9446188 & 0.66701567 \\ 0.62854085 & -0.97454336 & 0.71636295 \\ 0.53813813 & -0.84238446 & 0.62076123 \end{bmatrix}_{9 \times 3}$$

13. Melakukan analisis dan merancang diagram aktivitas perangkat lunak randomisasi

Status : Ada sejak rencana kerja skripsi.

Hasil : Perangkat lunak randomisasi adalah perangkat lunak yang digunakan untuk memodifikasi data dengan metode randomisasi. Perangkat lunak ini akan memiliki fungsi untuk mengubah nilai setiap data yang dimasukkan agar privasinya terjaga tetapi masih dapat dilakukan penambahan data. Algoritma *Random Rotation Perturbation* dan *Random Projection Perturbation* akan diimplementasikan pada perangkat lunak ini untuk fungsi utama yaitu mengubah nilai pada setiap data. Dengan mempertimbangkan studi literatur, analisis masalah, dan studi kasus yang telah dilakukan pada kedua teknik tersebut, perangkat lunak akan memiliki berbagai pilihan dan parameter yang pengguna harus masukan dan juga ada beberapa persyaratan atau batasan agar perangkat lunak ini berjalan dengan semestinya. Diagram aktivitas untuk perangkat lunak randomisasi dapat dilihat pada Gambar ?? . Detail dari diagram aktivitas tersebut adalah sebagai berikut.

- Pengguna memberikan masukan berupa dataset yang berupa dokumen berjenis *comma-separated values* (CSV). dokumen ini harus berisi tiap record pada barisnya dan tiap fitur pada kolomnya. Adanya nama kolom diperbolehkan pada baris pertama dalam dokumen tersebut
- Perangkat lunak akan membersihkan dokumen yang berisi dataset tersebut dan mentransformasi datasetnya menjadi sebuah matriks. Matriks tersebut akan berisi nilai-nilai pada dataset saja tanpa nama kolom
- Pengguna memilih teknik randomisasi yang ingin digunakan antara *Random Rotation Perturbation* dan *Random Projection Perturbation*. Jika *Random Projection Perturbation* yang dipilih maka akan ada beberapa langkah yang harus dipenuhi yaitu sebagai berikut.
 - Pengguna memberi masukan nilai Epsilon yang diinginkan

- ii. Perangkat lunak memeriksa persyaratan yang harus dipenuhi oleh matriks dataset dengan menghitung menggunakan algoritma tertentu. Pengecekan ini adalah pengecekan jumlah kolom pada matriks dataset apakah cukup untuk matriks tersebut direduksi dimensinya. Jika persyaratan terpenuhi maka langkah selanjutnya adalah perangkat lunak mengaplikasikan teknik yang dipilih
- iii. Jika persyaratan tidak terpenuhi maka pengguna harus memilih untuk mengganti datasetnya atau mengganti nilai Epsilon
- (d) Perangkat lunak mengaplikasikan teknik yang telah dipilih pada dataset yang sudah berbentuk matriks
- (e) Pengguna memilih lokasi pada komputer pengguna untuk menyimpan hasil dari teknik yang dipilih. Hasilnya adalah sebuah dokumen CSV yang berisi dataset yang sudah diaplikasikan teknik randomisasi
- (f) Perangkat lunak menyimpan hasilnya pada lokasi yang telah ditentukan pengguna dan menampilkan beberapa informasi tentang hasil yang sukses dibuat

14. Melakukan analisis dan merancang diagram kelas perangkat lunak randomisasi

Status : Ada sejak rencana kerja skripsi.

Hasil : Perancangan diagram kelas didasarkan pada analisis terhadap algoritma yang ingin diimplementasikan yaitu *Random Rotation Perturbation* dan *Random Projection Perturbation*, serta berdasarkan pada diagram aktivitas yang telah dibuat, dan dengan mempertimbangkan studi literatur, analisis masalah, dan studi kasus yang telah dilakukan pada kedua algoritma randomisasi yang ingin diimplementasikan. Detail dari diagram kelas perangkat lunak randomisasi pada Gambar ?? adalah sebagai berikut.

15. Menulis dokumen skripsi

Status : Ada sejak rencana kerja skripsi.

Hasil : Penulisan dokumen skripsi telah dilakukan sampai hasilnya sekarang sudah ada bab 1 yang berisi pendahuluan, bab 2 yang berisi dasar teori, dan bab 3 yang berisi analisis masalah. Tetapi masih belum 100% selesai, perlu adanya proses finalisasi.

6 Pencapaian Rencana Kerja

Langkah-langkah kerja yang berhasil diselesaikan dalam Skripsi 1 ini adalah sebagai berikut:

1. Mempelajari dasar-dasar privasi data
2. Mempelajari teknik *Random Rotation Perturbation* dan *Random Projection Perturbation* untuk *privacy preserving data mining*
3. Mempelajari teknik penambahan data yang akan digunakan
4. Melakukan analisis terhadap teknik *Random Rotation Perturbation* dan *Random Projection Perturbation* serta bagaimana penerapannya dengan teknik penambahan data yang akan digunakan
5. Menulis dokumen skripsi bab 1, 2, dan 3

7 Kendala yang Dihadapi

Kendala - kendala yang dihadapi selama mengerjakan skripsi :

- Kesibukan lain yang menghabiskan banyak waktu seperti kuliah, kerja praktek, tugas-tugas mata kuliah lain, masalah pribadi dan hal-hal lainnya
- Kesulitan dalam melakukan studi literatur dengan membaca paper yang berbahasa Inggris dan penuh dengan rumus matematika
- Kesulitan dalam memahami konsep matematika yang ada pada teknik-teknik yang digunakan
- Kesulitan dalam mengetahui apa saja yang perlu ada di dokumen skripsi dan apa saja analisis yang harus dilakukan
- Kesulitan dalam menjaga mental untuk tetap semangat mengerjakan skripsi di kala masa-masa sulit

Bandung, 02/05/2020

Stephen Jordan

Menyetujui,

Menyetujui,

Nama: Mariskha Tri Adithia, P.D.Eng
Pembimbing Utama

Nama: Dr. Veronica Sri Moertini, Ir., M.T.
Pembimbing Pendamping