Privacy Preserving Data Mining dengan Metode Randomization

Chris Eldon–2016730073

1 Data Skripsi

Pembimbing utama/tunggal: **Mariskha Tri Adithia, P.D.Eng**
Pembimbing pendamping: -
Kode Topik : **MTA4703**
Topik ini sudah dikerjakan selama : **1** semester
Pengambilan pertama kali topik ini pada : Semester **47 - Ganjil 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

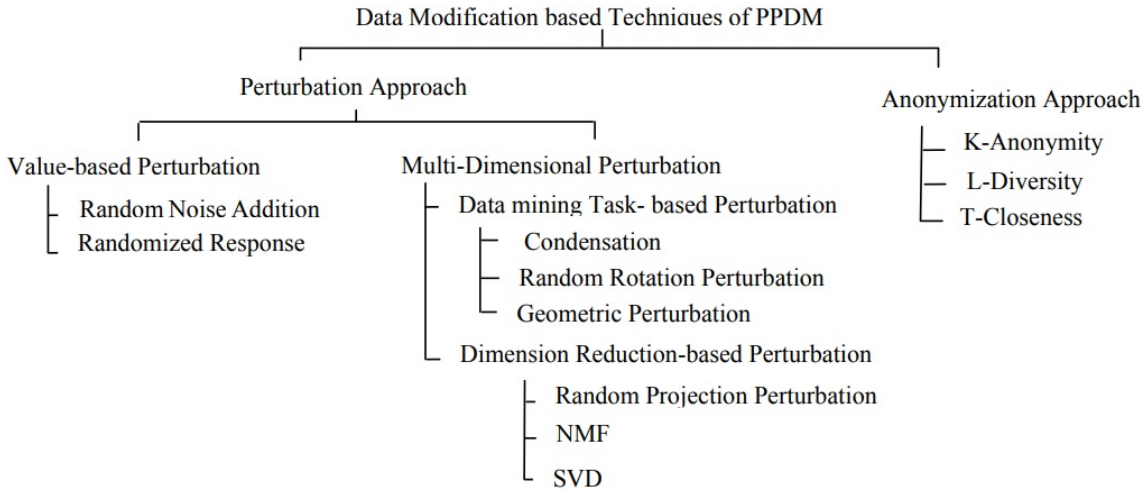
Dengan semakin banyaknya penambangan data yang dilakukan dan data yang digunakan juga semakin banyak, semakin banyak juga privasi di dalam data tersebut yang tersebar kepada pihak yang melakukan penambangan data. Data privasi tersebut dapat tersebar kepada pihak yang tidak bertanggung jawab dan disalahgunakan. Oleh karena itu perlu adanya suatu cara untuk mencegah privasi tersebar pada proses penambangan data, menjaga privasi pada data tersebut. Istilah untuk hal tersebut adalah *privacy preserving data mining*.

Salah satu cara untuk melakukan *privacy preserving data mining* adalah dengan melakukan modifikasi data yang ada sebelum diberikan kepada pihak lain. Ada macam-macam teknik dan algoritma yang bertujuan modifikasi data untuk *privacy preserving data mining* yang bisa dibagi menjadi dua jenis yaitu *Perturbation Approach* dan *Anonymization Approach*. *Perturbation Approach* adalah pendekatan untuk *privacy preserving data mining* dengan cara mengacaukan data yang ada, tetapi hasil data yang dikacaukan masih tetap bisa ditambang. *Perturbation Approach* bisa dibagi menjadi dua jenis 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.

Dari berbagai macam teknik modifikasi data untuk *privacy preserving data mining* yang dapat dilihat pada Gambar 1, terdapat empat teknik yang menggunakan metode *Randomization* yaitu *Random Noise Addition*, *Randomized Response*, *Random Rotation Perturbation*, dan *Random Projection Perturbation*.

Pada penelitian ini, akan dibuat sebuah perangkat lunak yang dapat memproses data yang akan ditambang menjadi data yang telah dimodifikasi dengan metode *Randomization* sehingga tidak mengandung privasi, tetapi masih dapat ditambang. Dari berbagai macam teknik dengan metode *Randomization* yang ada, dipilih dua buah teknik yaitu *Random Rotation Perturbation* dan *Random Projection Perturbation* untuk diimplementasikan pada perangkat lunak serta membandingkan hasil dari kedua teknik tersebut.



Gambar 1: Berbagai macam teknik modifikasi data untuk *privacy preserving data mining*

3 Rumusan Masalah

Berdasarkan latar belakang, rumusan masalah pada penelitian ini adalah sebagai berikut.

1. Bagaimana cara kerja dari teknik *Random Rotation Perturbation* dan *Random Projection Perturbation* untuk *privacy preserving data mining*?
2. Bagaimana implementasi dari teknik *Random Rotation Perturbation* dan *Random Projection Perturbation* pada perangkat lunak?
3. Bagaimana perbandingan antara hasil dari teknik *Random Rotation Perturbation* dan *Random Projection Perturbation*?

4 Tujuan

Berdasarkan rumusan masalah, maka tujuan dari penilitian ini adalah sebagai berikut.

1. Mempelajari cara kerja dari teknik *Random Rotation Perturbation* dan *Random Projection Perturbation* untuk *privacy preserving data mining*
2. Mengimplementasikan teknik *Random Rotation Perturbation* dan *Random Projection Perturbation* pada perangkat lunak
3. Melakukan analisis dan pengujian untuk membandingkan dan mengukur hasil dari teknik *Random Rotation Perturbation* dan *Random Projection Perturbation*

5 Detail Perkembangan Pengerjaan Skripsi

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

1. **Melakukan studi literatur mengenai dasar-dasar privasi data**
   Status : Ada sejak rencana kerja skripsi.
   Hasil : Pada umumnya sebuah data bisa dikatakan privasi apabila data tersebut dapat dikaitkan dengan identitas seseorang. Tetapi setiap orang memiliki kepentingan privasi yang berbeda-beda sehingga definisi dari privasi sulit untuk dijelaskan secara eksak. Oleh karena itu, perlu adanya konsep privasi yang dapat menjadi acuan untuk menentukan data seperti apa yang termasuk privasi atau bukan.

(a) Privasi

Dalam mendefinisikan privasi, sulit untuk mendapatkan definisi yang tepat untuk privasi karena setiap individu memiliki kepentingan yang berbeda-beda sehingga privasi pada setiap individu dapat berbeda-beda juga. Beberapa definisi privasi telah dikemukakan dan definisi tersebut bermacam-macam berdasarkan konteks, budaya, dan lingkungan. Menurut Warren dan Brandeis pada papernya, mereka mendefinisikan privasi sebagai "*the right to be alone.*", hak untuk menyendiri. Lalu pada papernya, Westin mendefinisikan privasi sebagai "*the desire of people to choose freely under what circumstances and to what extent they will expose themselves, their attitude, and their behavior to others*", keinginan orang untuk memilih secara bebas dalam segala situasi dan dalam hal mengemukakan diri mereka, sikap mereka, dan tingkah laku mereka pada orang lain. Schoeman mendefinisikan privasi sebagai "*the right to determine what (personal) information is communicated to others*", hak untuk menentukan informasi pribadi apa saja yang dikomunikasikan kepada yang lain, atau "*the control an individual has over information about himself or herself.*", kendali seorang individu terhadap informasi tentang dirinya sendiri. Lalu baru-baru ini, Garfinkel menyatakan bahwa "*privacy is about self-possession, autonomy, and integrity.*", privasi adalah tentang penguasaan diri sendiri, otonomi, dan integritas. Di samping itu, Rosenberg berpendapat bahwa privasi sebenarnya bukan sebuah hak tetapi sebuah rasa: "*If privacy is in the end a matter of individual taste, then seeking a moral foundation for it – beyond its role in making social institutions possible that we happen to prize – will be no more fruitful than seeking a moral foundation for the taste for truffles.*", intinya setiap orang memiliki perhatian yang berbeda-beda terhadap privasi mereka sendiri sehingga hal tersebut tergantung apa yang dirasakan oleh setiap individu.

Dari definisi-definisi privasi yang telah disebutkan di atas, dapat disimpulkan bahwa privasi dilihat sebagai konsep sosial dan budaya. Konsep privasi pada suatu lingkungan dapat berbeda dari lingkungan lainnya dan hal ini menyebabkan sulitnya menentukan apakah sebuah data termasuk privasi atau bukan. Oleh karena itu, perlu adanya sebuah standar privasi untuk menentukan data mana yang dapat disebut sebuah privasi. Organisasi National Institute of Standards and Technology dari Amerika Serikat, membuat standar mereka sendiri untuk menentukan informasi seperti apa yang dapat disebut sebagai privasi. Mereka mengemukakan konsep *Personally Identifiable Information* sebagai informasi yang dapat dikatakan personal untuk setiap individu.

(b) *Personally Identifiable Information*

Privasi dapat dikatakan adalah sebuah informasi personal seseorang yang dapat mengidentifikasi suatu hal pada orang tersebut. Konsep yang sering kali digunakan untuk mendeskripsikan informasi personal adalah*Personally Identifiable Information* yang disingkat PII. PII adalah segala informasi mengenai individu yang dikelola oleh sebuah instansi, termasuk segala informasi yang dapat digunakan untuk membedakan atau mengusut identitas seseorang dan juga segala informasi yang berhubungan atau dapat dihubungkan kepada suatu individu, seperti informasi medis, pendidikan, finansial, dan pekerjaan seseorang.

Informasi yang termasuk membedakan individu adalah informasi yang dapat mengidentifikasi seorang individu. Beberapa contoh informasi yang mengidentifikasi seorang individu adalah nama, nomor KTP, tempat tanggal lahir, nama ibu kandung, atau catatan medis. Sedangkan, data yang hanya berisi misalkan saldo tabungan tanpa ada informasi lain mengenai identitas seseorang yang berkaitan tidak menyediakan informasi yang cukup untuk mengidentifikasi seorang individu.

Mengusut identitas seseorang adalah proses dari membuat perkiraan tentang aspek spesifik dari aktivitas atau status seseorang. Contohnya adalah sebuah catatan finansial seseorang dapat digunakan untuk memperkirakan aktivitas dari individu tersebut.

Informasi yang berhubungan dapat didefinisikan sebagai informasi yang berkaitan dengan seorang

individu yang mana terkait secara logis dengan informasi lain tentang individu tersebut. Contohnya adalah apabila ada dua buah basis data yang memiliki data berbeda dari seorang individu, maka seseorang yang memiliki akses pada 2 basis data tersebut berpotensi dapat mengaitkan data-data tersebut lalu mengidentifikasi individu yang ada pada data tersebut.

2. **Melakukan studi literatur mengenai penambangan data dan tekniknya**
   Status : Ada sejak rencana kerja skripsi.
   Hasil : Pada era teknologi informasi, sangat banyak data terkumpul pada basis data. Data yang masif ini dapat dimanfaatkan untuk menggali informasi penting yang berguna untuk pembuatan keputusan. Proses pada aktivitas ini secara kasar dapat disebut dengan penambangan data.

   Penambangan data adalah proses mengekstrak sebuah pola atau sebuah pengetahuan dari kumpulan data yang besar, yang mana dapat direpresentasikan dan diinterpretasikan. Pada penambangan data, teknik *machine learning* dan *pattern recognition* intensif digunakan untuk mendapatkan pola maupun pengetahuan baru dari data. Tujuan utama dari penambangan data adalah untuk membentuk model deskriptif dan prediktif dari suatu data. Model deskriptif berusaha untuk mengubah pola-pola yang ada pada data menjadi deskripsi yang bisa dimengerti oleh orang awam. Sedangkan model prediktif digunakan untuk memprediksi data yang tidak diketahui atau data yang berpotensi muncul di kemudian hari.

   Model tersebut biasanya dibuat dengan menggunakan teknik *machine learning*, yang mana terdapat dua teknik *machine learning* yang paling sering digunakan yaitu *classification* dan *clustering*. Subbab berikutnya akan menjelaskan secara singkat kedua teknik tersebut dan contoh algoritmanya.

   (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.

   (b) *Clustering*

   Clustering adalah proses mengelompokan 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 «TODO»

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 1. *Perturbation Approach* adalah pendekatan untuk *privacy preserving data mining* dengan cara mengacaukan data yang ada, tetapi hasil data yang dikacaukan masih tetap bisa ditambang. Sedangkan pada *Anonymization Approach*, data diterapkan de-identifikasi di mana dataset mentah disebarluaskan setelah menghapus inti dari identitas setiap rekord.

   Perturbation Approach bisa 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 akurasinya 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 mempunyai perbedaan pada properti-properti yang terjaga. Teknik-teknik modifikasi data tertentu sering kali mempunyai 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 1, terdapat empat teknik yang menggunakan metode *Randomization* yaitu *Random Noise Addition*, *Randomized Response*, *Random Rotation Perturbation*, dan *Random Projection*

(1) f_X^0 := Uniform distribution
(2) $j := 0$ // Iteration number
repeat
(3) $f_X^{j+1}(a) := \frac{1}{n} \sum_{i=1}^{n} \frac{f_Y(w_i - a) f_X^j(a)}{\int_{-\infty}^{\infty} f_Y(w_i - z) f_X^j(z) dz}$
(4) $j := j + 1$
until (stopping criterion met)

Gambar 2: Algoritma rekonstruksi

Perturbation.

Berbagai macam teknik dengan metode randomisasi umumnya menerapkan perusakan 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 bisa 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 mempunyai 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 2.

   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 distribusi lah yang

terjaga oleh teknik *Random Noise Addition* maka teknik penambangan data yang dapat digunakan hanya teknik-teknik yang bergantung pada distribusi data saja.

Modifikasi pada algoritma penambangan 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 bisa 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 mempunyai 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.

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

   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 mempunyai sifat special orthogonal. Matriks *special orthogonal* adalah matriks yang mempunyai 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 tranformasi 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 mempunyai sifat demikian sehigga 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* $f : \mathbb{R}^m \to \mathbb{R}^k$ *such that, for all* $x, y \in S$*,* $(1 - eps)||u - v||^2 < ||p(u) - p(v)||^2 < (1 + eps)||u - v||^2$*, where* $||.||$ *denotes the vector 2-norm.*

«TEOREMA» 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.

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

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 attribut d, yang mana d akan menjadi dimensi matriks. Oleh karena reduksi dimensi lah yang diinginkan 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 mempunyai 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.

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

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* de-

ngan ϵ (eps) yang ditentukan berada pada rentang [0, 1].

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

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 analisa 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-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 analisa terhadap teknik *Random Rotation Perturbation***
   Status : Ada sejak rencana kerja skripsi.
   Hasil : Algoritma *Random Rotation Perturbation* mempunyai beberapa langkah yaitu sebagai berikut.

   (a) Dataset yang mempunyai atribut sebanyak d dan rekord 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* mempunyai 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 penambangan data

10. **Melakukan analisa terhadap teknik *Random Projection Perturbation***
    Status : Ada sejak rencana kerja skripsi.
    Hasil : Algoritma *Random Projection Perturbation* mempunyai beberapa langkah yaitu sebagai berikut.

    (a) Dataset yang mempunyai atribut sebanyak d dan rekord sebanyak n direpresentasikan dalam bentuk matriks berukuran $n \times d$

    (b) Tentukan nilai ϵ (eps) 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 syarat 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 penambangan data

11. **Melakukan analisa 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. Diagram aktivitas untuk perangkat lunak randomisasi dapat dilihat pada Gambar 3. Detail dari diagram aktivitas tersebut adalah sebagai berikut.

(a) Pengguna memberikan masukan berupa dataset yang berupa file berjenis *comma separated value* (CSV). File ini harus berisi tiap rekord pada barisnya dan tiap fitur pada kolomnya. Adanya nama kolom diperbolehkan pada baris pertama dalam file tersebut

(b) Perangkat lunak akan membersihkan file yang berisi dataset tersebut dan mentransformasi datasetnya menjadi sebuah matriks. Matriks tersebut akan berisi nilai-nilai pada dataset saja tanpa nama kolom

(c) Pengguna memilih teknik randomisasi yang ingin digunakan antara *Random Rotation Perturbation* dan *Random Projection Perturbation*. Jika *Random Projection Perturbation* yang dipiilih maka akan ada beberapa langkah yang harus dipenuhi yaitu sebagai berikut.

i. Pengguna memberi masukan nilai epsilon yang diinginkan

ii. Perangkat lunak memeriksa syarat 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 syarat terpenuhi maka langkah selanjutnya adalah perangkat lunak mengaplikasikan teknik yang dipilih

iii. Jika syarat tidak terpenuhi maka pengguna harus memilih untuk mengganti datasetnya atau mengganti nilai epsilon

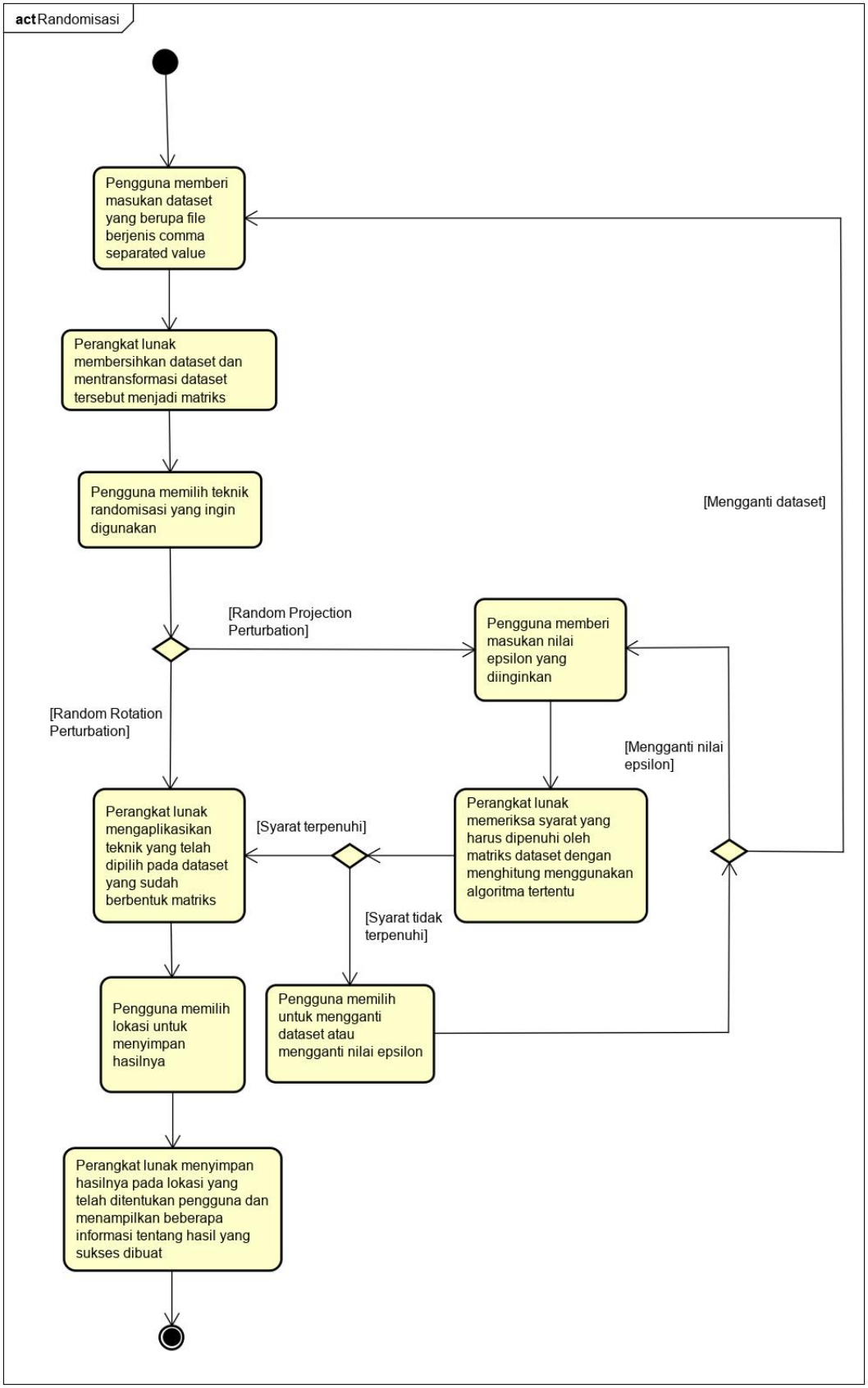
(d) Perangkat lunak mengaplikasikan teknik yang telah dipilih pada dataset yang sudah berbentuk matriks

(e) Pengguna memilih lokasi pada komputer pengguna untuk menyimpan hasil dari teknik yang dipilih. Hasilnya adalah sebuah file 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

12. **Melakukan analisa dan merancang diagram kelas perangkat lunak randomisasi**
Status : Ada sejak rencana kerja skripsi.
Hasil : Detail dari diagram kelas perangkat lunak randomisasi pada Gambar 4 adalah sebagai berikut.

- Kelas *Perturbation* adalah kelas abstrak yang akan di-*extend* oleh kelas yang mengimplementasikan algoritma *Random Rotation Perturbation* dan *Random Projection Perturbation*



Gambar 3: Diagram aktivitas perangkat lunak randomisasi

- Kelas *RandomRotationPerturbation* adalah kelas yang mengimplementasikan algoritma *Random Rotation Perturbation*. Kelas ini merupakan *subclass* dari kelas *Perturbation*
- Kelas *RandomProjectionPerturbation* adalah kelas yang mengimplementasikan algoritma *Random Projection Perturbation*. Kelas ini merupakan *subclass* dari kelas *Perturbation*
- Kelas *Matrix* adalah kelas untuk merepresentasikan matriks dan menyimpan nilai-nilai pada setiap elemen matriks. Kelas ini juga memiliki fungsi perkalian yang akan digunakan untuk implementasi algoritma *Random Rotation Perturbation* dan *Random Projection Perturbation*
- Kelas *RandomTranslationMatrix* adalah kelas untuk membuat matriks translasi acak. Kelas ini adalah *subclass* dari kelas *Matrix*
- Kelas *RandomRotationMatrix* adalah kelas untuk membuat matriks rotasi acak. Kelas ini adalah *subclass* dari kelas *Matrix*
- Kelas *RandomProjectionMatrix* adalah kelas untuk membuat matriks proyeksi acak. Kelas ini adalah *subclass* dari kelas *Matrix*
- Kelas *CSVPreprocessor* adalah kelas untuk menangani masukan dataset berupa file CSV yang akan direpresentasikan menjadi matriks. Kelas ini berguna untuk mengkonversi file CSV menjadi matriks dan sebaliknya.

13. **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

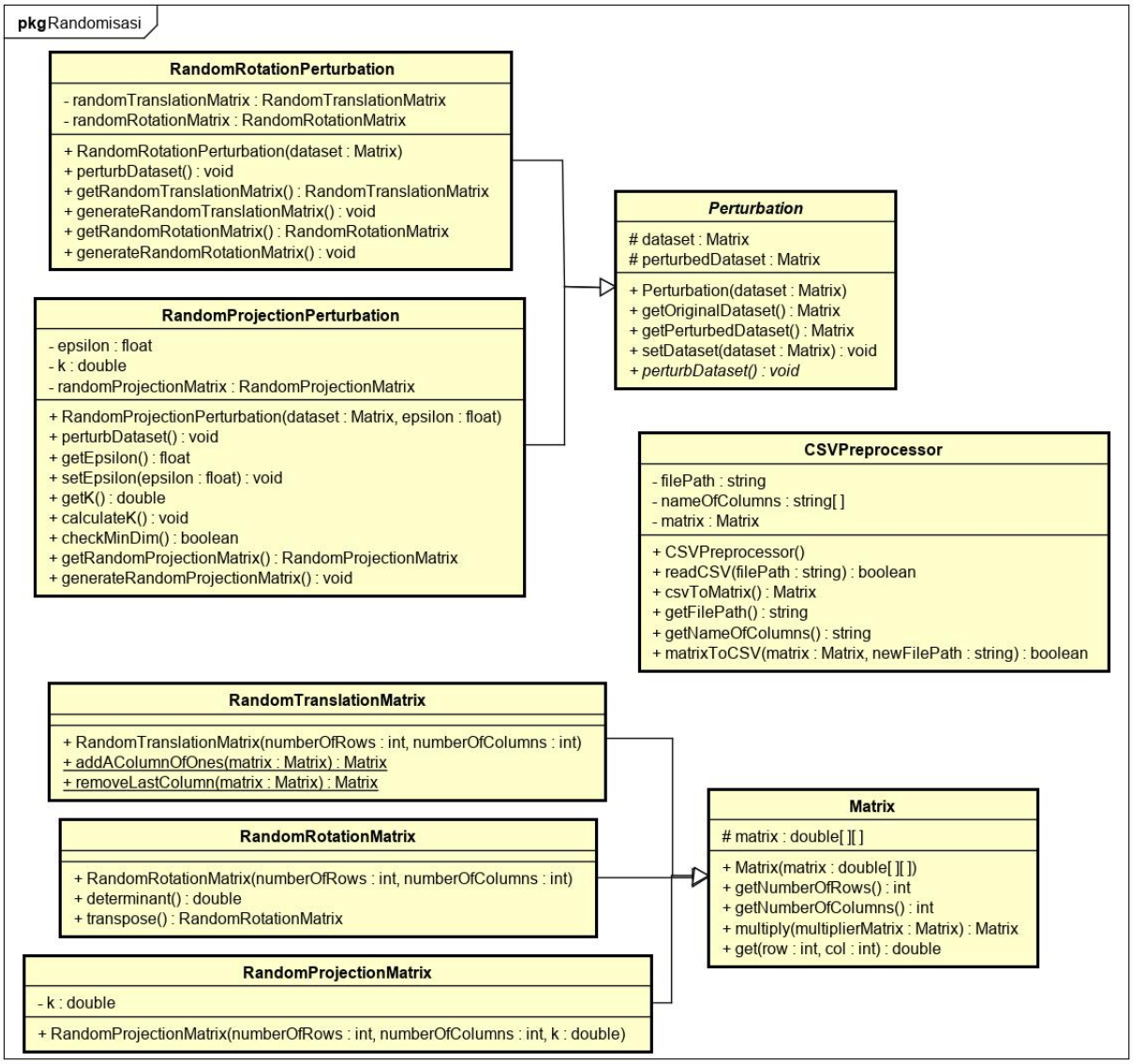
Langkah-langkah kerja yang berhasil diselesaikan dalam Skripsi 1 ini adalah sebagai berikut:

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

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



Gambar 4: Diagram kelas perangkat lunak randomisasi

- 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, 19/11/2019

Chris Eldon

Menyetujui,

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