

BAB 1

PENDAHULUAN

1.1 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*.

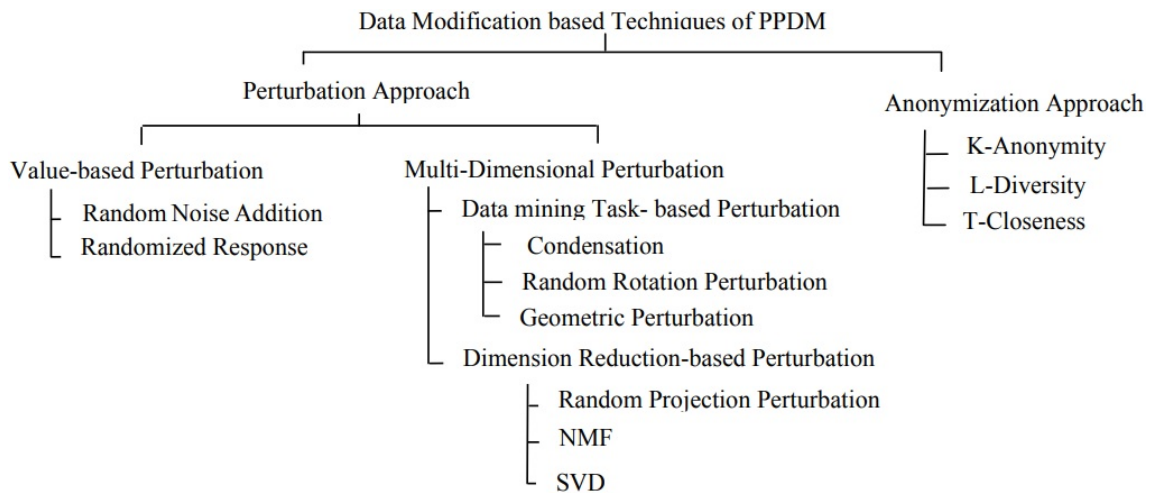
Ada kesulitan dalam menentukan data seperti apa yang dapat disebut sebagai privasi. Privasi dapat dikatakan adalah sebuah informasi personal seseorang yang dapat mengidentifikasi suatu hal pada orang tersebut. Konsep yang sering kali digunakan untuk mendeskripsikan informasi personal adalah *Personally 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.

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 2.1, 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 privasi pada data tersebut terlindungi, 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



Gambar 1.1: Berbagai macam teknik modifikasi data untuk *privacy preserving data mining*

serta membandingkan hasil dari kedua teknik tersebut.

1.2 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*?

1.3 Tujuan

Berdasarkan rumusan masalah, maka tujuan dari penelitian 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*

1.4 Batasan Masalah

Batasan-batasan masalah untuk penelitian ini adalah sebagai berikut.

1. «TODO»

Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Donec odio elit, dictum in, hendrerit sit amet, egestas sed, leo. Praesent feugiat sapien aliquet odio. Integer vitae justo. Aliquam vestibulum fringilla lorem. Sed neque lectus, consectetur at,

consectetur sed, eleifend ac, lectus. Nulla facilisi. Pellentesque eget lectus. Proin eu metus. Sed porttitor. In hac habitasse platea dictumst. Suspendisse eu lectus. Ut mi mi, lacinia sit amet, placerat et, mollis vitae, dui. Sed ante tellus, tristique ut, iaculis eu, malesuada ac, dui. Mauris nibh leo, facilisis non, adipiscing quis, ultrices a, dui.

1.5 Metodologi

Metodologi yang digunakan dalam penelitian ini adalah sebagai berikut.

1. Melakukan studi literatur dasar-dasar privasi data
2. Melakukan studi literatur teknik *Random Rotation Perturbation* dan *Random Projection Perturbation* untuk *privacy preserving data mining*
3. Melakukan studi literatur teknik penambangan data yang akan digunakan
4. Melakukan analisis terhadap teknik *Random Rotation Perturbation* dan *Random Projection Perturbation* serta bagaimana penerapannya dengan teknik penambangan data yang akan digunakan
5. Melakukan perancangan perangkat lunak yang mengimplementasikan teknik *Random Rotation Perturbation* dan *Random Projection Perturbation*
6. Membangun perangkat lunak yang mengimplementasikan teknik *Random Rotation Perturbation* dan *Random Projection Perturbation*
7. Menguji perangkat lunak secara fungsional dan eksperimental dengan menggunakan *real data*
8. Menerapkan teknik penambangan data terhadap data yang telah diproses untuk menganalisis hasil dari teknik *Random Rotation Perturbation* dan *Random Projection Perturbation*
9. Melakukan analisis dan pengujian untuk membandingkan dan mengukur hasil dari teknik *Random Rotation Perturbation* dan *Random Projection Perturbation*
10. Menarik kesimpulan berdasarkan hasil eksperiment yang telah dilakukan

1.6 Sistematika Pembahasan

Laporan penelitian tersusun ke dalam enam bab secara sistematis sebagai berikut.

- Bab 1 Pendahuluan
Berisi latar belakang, rumusan masalah, tujuan, batasan masalah, metodologi penelitian, dan sistematika pembahasan.
- Bab 2 Dasar Teori
Berisi dasar teori tentang dasar-dasar privasi data, *Random Rotation Perturbation*, *Random Projection Perturbation*, dan teknik penambangan data.
- Bab 3 Analisis
Berisi analisis masalah, studi kasus, dan diagram aliran proses.
- Bab 4 Perancangan
Berisi perancangan perangkat lunak yang dibangun meliputi perancangan antarmuka dan diagram kelas yang lengkap.

- Bab 5 Implementasi dan Pengujian
Berisi implementasi antarmuka perangkat lunak, pengujian fungsional, pengujian eksperimental, dan kesimpulan dari pengujian.
- Bab 6 Kesimpulan dan Saran
Berisi kesimpulan dari awal hingga akhir penelitian dan saran untuk pengembangan selanjutnya.

BAB 2

DASAR TEORI

Dalam menjaga privasi data, perlu adanya definisi privasi yang konkrit untuk menentukan data seperti apa yang menjadi privasi. Pada penambahan data, perlu ada teknik yang baik untuk menjaga privasi tidak tersebar kepada orang yang tidak berhak. Ada beberapa teknik untuk menjaga privasi pada penambahan data antara lain modifikasi data dengan metode randomisasi yaitu teknik *Random Rotation Perturbation* dan *Random Projection Perturbation*

2.1 Privasi Data

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.

2.1.1 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. [1] 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. [1] 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.

2.1.2 *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. [2]

Informasi yang termasuk membedakan individu adalah informasi yang dapat mengidentifikasi seorang individu. Informasi seperti ini adalah data privasi yang secara langsung bisa didapatkan. 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.

Dari sebuah data, bisa saja data tersebut secara tidak langsung mengandung privasi, identitas seseorang bisa didapatkan tanpa data tersebut memberikan langsung identitas orang tersebut. Mengusut identitas seseorang adalah proses dari membuat perkiraan tentang aspek spesifik dari aktivitas atau status seseorang. Jika sebuah data dapat dianalisis datanya sampai identitas seseorang dapat diakses, berarti data tersebut secara tidak langsung mengandung privasi. 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. Informasi tersebut secara tidak langsung mengandung privasi dan dapat diolah agar identitas seseorang bisa didapatkan. 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.

2.2 Penambangan Data

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. [3] 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.

2.2.1 Classification

Tujuan utama *Classification* (klasifikasi) adalah membuat model yang dalam kasus ini disebut *classifier* yang mana dapat mengidentifikasi nilai kelas dari suatu data. [3] 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 penambahan 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 record pada data dipetakan ke bidang Euclidean dengan beberapa atribut yang menentukan letaknya pada bidang Euclidean [4].

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

1. Tentukan nilai k yang menentukan seberapa banyak tetangga yang digunakan
2. 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
3. Urutkan jarak Euclidean titik-titik yang sudah dihitung pada perulangan pada langkah sebelumnya secara menaik
4. Pilih record teratas (jarak Euclidean yang paling kecil) sebanyak k dari urutan pada langkah sebelumnya
5. Ambil label dari semua record yang terpilih pada langkah sebelumnya. Label terbanyak adalah hasil prediksi label pada record yang ingin diprediksi

2.2.2 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. [3]

Salah satu contoh teknik *clustering* adalah *K-means*. Teknik *k-means* adalah adalah teknik penambahan 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* [4].

1. Tentukan nilai k yang menentukan seberapa banyak kluster yang diinginkan dan sebuah *threshold* untuk menentukan batas perubahan nilai centroid
2. Tentukan secara acak sebuah centroid sebanyak k untuk setiap kluster

3. 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

2.3 *Privacy Preserving Data Mining*

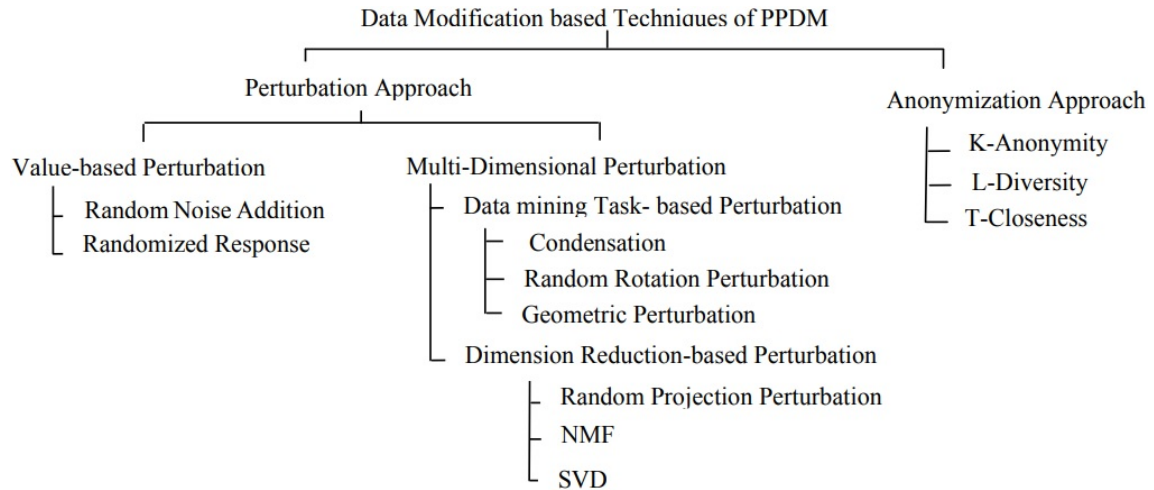
Aktivitas penambangan data melibatkan jumlah data yang sangat masif. Data-data yang digunakan memiliki privasi banyak individu di dalamnya. Hal ini berpotensi menyebabkan pelanggaran privasi dalam kasus tidak adanya proteksi yang cukup dan penyalahgunaan privasi data untuk tujuan lain. [5] Faktor utama 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 2.1. [5] *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. [5]

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 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. [6] 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



Gambar 2.1: Berbagai macam teknik modifikasi data untuk *privacy preserving data mining*

atau teknik penambangan data yang dapat digunakan berbeda karena properti yang terjaga pada teknik-teknik tersebut berbeda juga.

2.4 Metode *Randomization*

Dari berbagai macam teknik modifikasi data untuk *privacy preserving data mining* yang dapat dilihat pada Gambar 2.1, 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 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 [7]. 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. Masing-masing dari dua teknik tersebut akan dijelaskan lebih detil pada subbab berikut.

2.4.1 Random Rotation Perturbation

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.

$$G(X) = X_{n \times d} R_{d \times d} \quad (2.1)$$

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. [8] Dari definisi di atas dapat disimpulkan transformasi rotasi tersebut menjaga jarak Euclidean. [6]

Teknik ini menjaga beberapa properti pada data antara lain yaitu jarak Euclidean, *inner product*, dan *geometric shape hyper* pada bidang multi-dimensi. [5] 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*. [6] 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 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.

2.4.2 Random Projection Perturbation

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* [9].

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. [10]

Projection perturbation dari dataset X didefinisikan sebagai berikut. [10]

$$G(X) = X_{n \times d} R_{d \times k} \quad (2.2)$$

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.

$$k \geq 4(\epsilon^2/2 - \epsilon^3/3)^{-1} \ln n \quad (2.3)$$

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 $[0, 1]$.

$$(1 - \epsilon) \|u - v\|^2 < \|p(u) - p(v)\|^2 < (1 + \epsilon) \|u - v\|^2 \quad (2.4)$$

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.

BAB 3

ANALISIS

«TODO»

BAB 4

PERANCANGAN PERANGKAT LUNAK

Pada bab ini akan dijabarkan perancangan perangkat lunak untuk penelitian ini. Perancangan perangkat lunak tersebut meliputi perancangan antarmuka dan perancangan kelas untuk penelitian ini.

4.1 Perancangan Antarmuka

Perancangan antarmuka yang dibuat disesuaikan dengan diagram kelas dan diagram aktivitas yang dibuat sesuai analisis perangkat lunak dilakukan. Perangkat lunak randomisasi akan memiliki 3 halaman antarmuka yang diimplementasikan dengan desain antarmuka tabs. Rancangan antarmuka halaman utama dapat dilihat pada Gambar 4.1, halaman ini akan berisi pesan selamat datang berupa deskripsi singkat perangkat lunak beserta cara kerjanya dan petunjuk singkat cara penggunaan perangkat lunak. Pada bagian atas antarmuka terdapat sebuah kolom untuk memasukkan file csv yang ingin dirandomisasi. Pertama-tama pengguna wajib untuk mengisi kolom tersebut apapun teknik randomisasi yang akan digunakan nanti.

Setelah pengguna memasukkan file yang diinginkan, perangkat lunak akan menampilkan berbagai informasi mengenai dataset yang ada di file tersebut seperti ukuran filenya, nama filenya, dan jumlah kolom. Deskripsi ini bertujuan untuk memberitahukan pengguna bahwa file yang dimasukkan telah benar dan bagaimana sifat dari dataset yang dimasukkan.

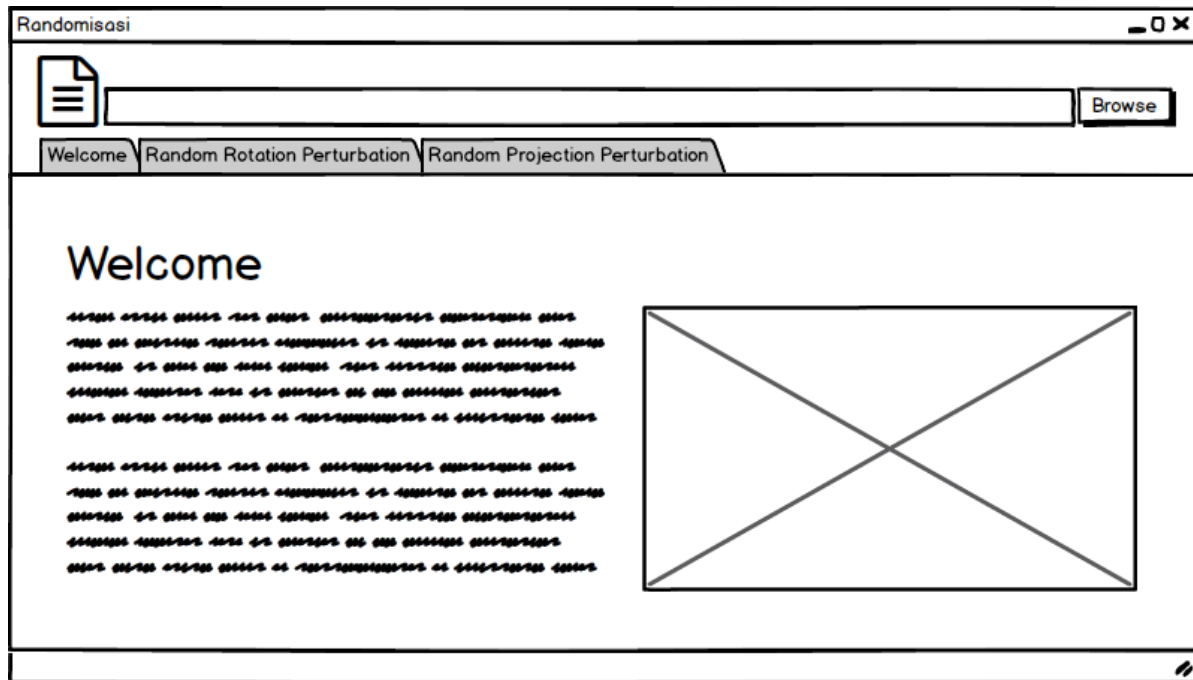
4.1.1 Halaman *Random Rotation Perturbation*

Halaman ini memiliki fungsi untuk melakukan teknik *Random Rotation Perturbation*. Setelah user memasukkan file yang diinginkan pada kolom yang berada pada bagian atas antarmuka, pengguna baru bisa mengakses halaman ini. Pada halaman ini terdapat berbagai informasi dataset yang dimasukkan dan hasil dari randomisasi jika berhasil dilakukan, serta pengaturan apa saja fitur yang akan dirandomisasi pada dataset yang pengguna masukkan. Rancangan dari halaman ini dapat dilihat pada Gambar 4.2.

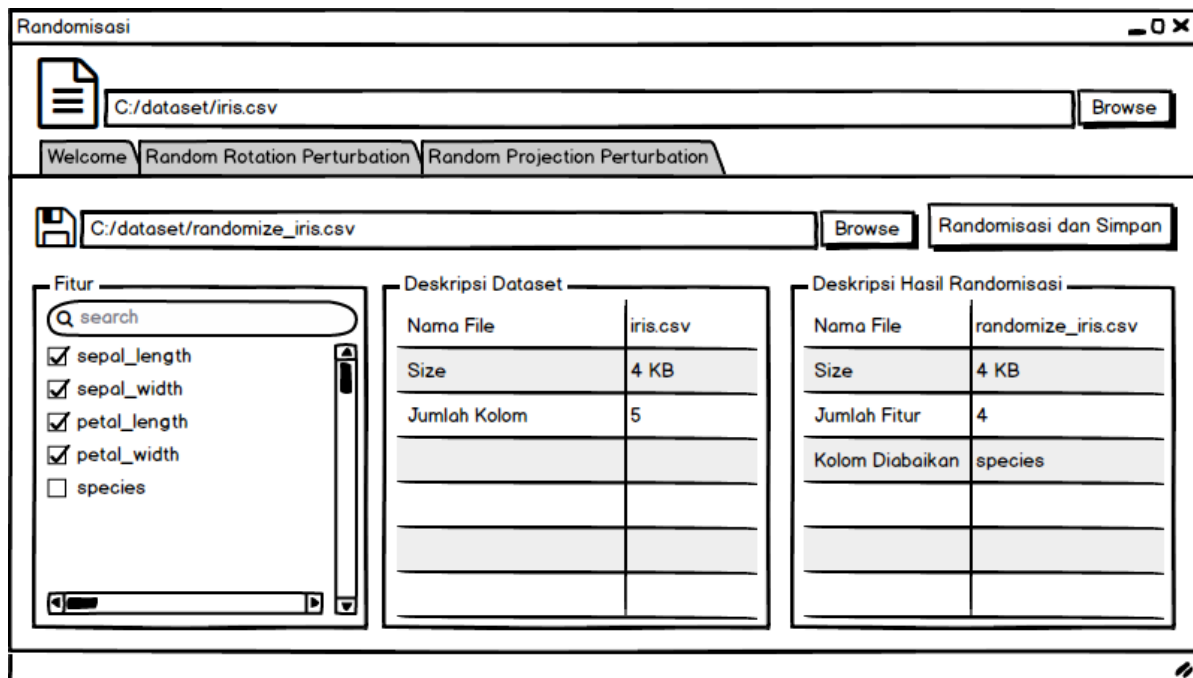
Sebelum melakukan randomisasi, pengguna perlu memeriksa fitur mana saja yang ingin dirandomisasi karena bisa saja pada dataset tersebut ada kolom yang berupa label. Perangkat lunak akan menyimpan hasil randomisasi pada file dataset pengguna yang telah diubah nilainya dan perangkat lunak akan menyimpan file tersebut di suatu lokasi yang pengguna harus tentukan. Pengguna bisa menentukan lokasi file hasil randomisasi pada kolom yang terdapat di sebelah kanan ikon simpan dan sebelah kiri tombol "Randomisasi dan Simpan".

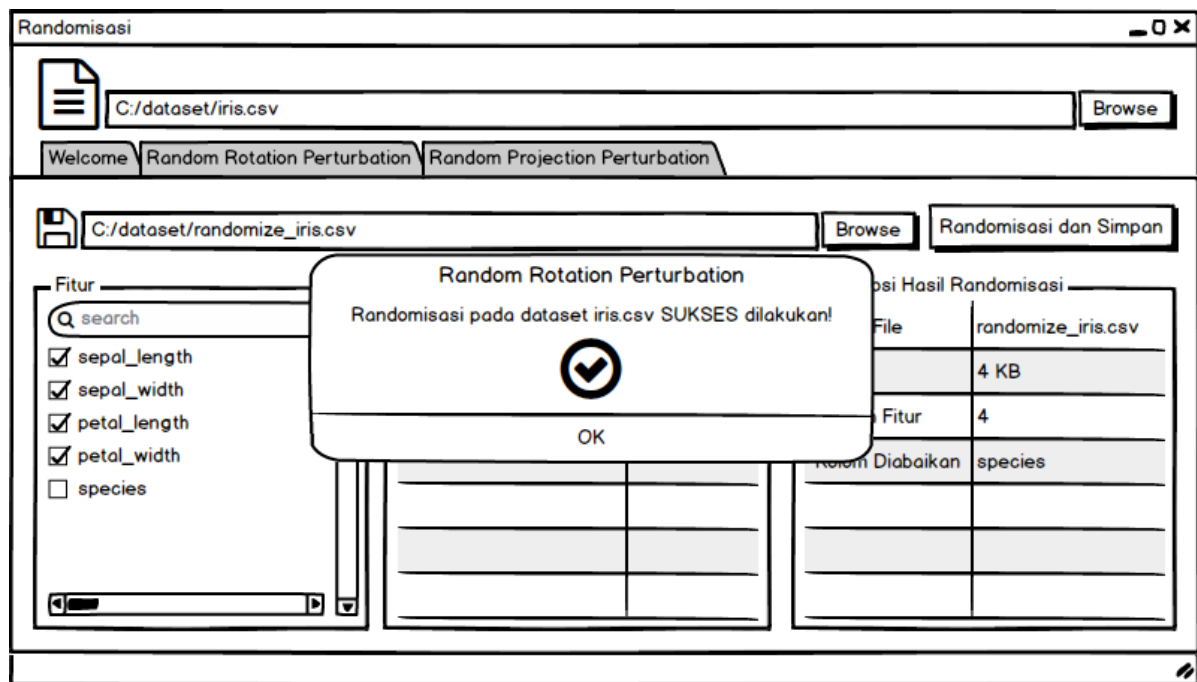
Setelah pengguna melakukan berbagai pengaturan, pengguna dapat melakukan randomisasi dengan menekan tombol "Randomisasi dan Simpan". Apabila randomisasi berhasil dilakukan, perangkat lunak akan menampilkan *popup* yang memberitahukan bahwa randomisasi dengan teknik *Random Rotation Perturbation* pada dataset yang dimasukkan berhasil dilakukan. *Popup* tersebut dapat dilihat pada Gambar 4.3. Selain itu, perangkat lunak juga akan menampilkan berbagai informasi beserta deskripsi tentang hasil randomisasi yang berhasil dilakukan.

Perangkat lunak akan menampilkan deskripsi hasil dari randomisasi yang dilakukan. Informasi



Gambar 4.1: Halaman utama yang berisi petunjuk cara penggunaan perangkat lunak

Gambar 4.2: Halaman untuk melakukan teknik *Random Rotation Perturbation*



Gambar 4.3: *Popup* yang ditampilkan apabila randomisasi sukses dilakukan

yang ditampilkan oleh perangkat lunak antara lain nama file, ukuran file, jumlah fitur, dan kolom yang diabaikan. Informasi ini ditampilkan oleh perangkat lunak bertujuan untuk memberitahukan pengguna properti-properti dataset yang telah dirandomisasi dan pengguna dapat memeriksa hasil yang dihasilkan oleh perangkat lunak apakah sesuai dengan keinginan pengguna.

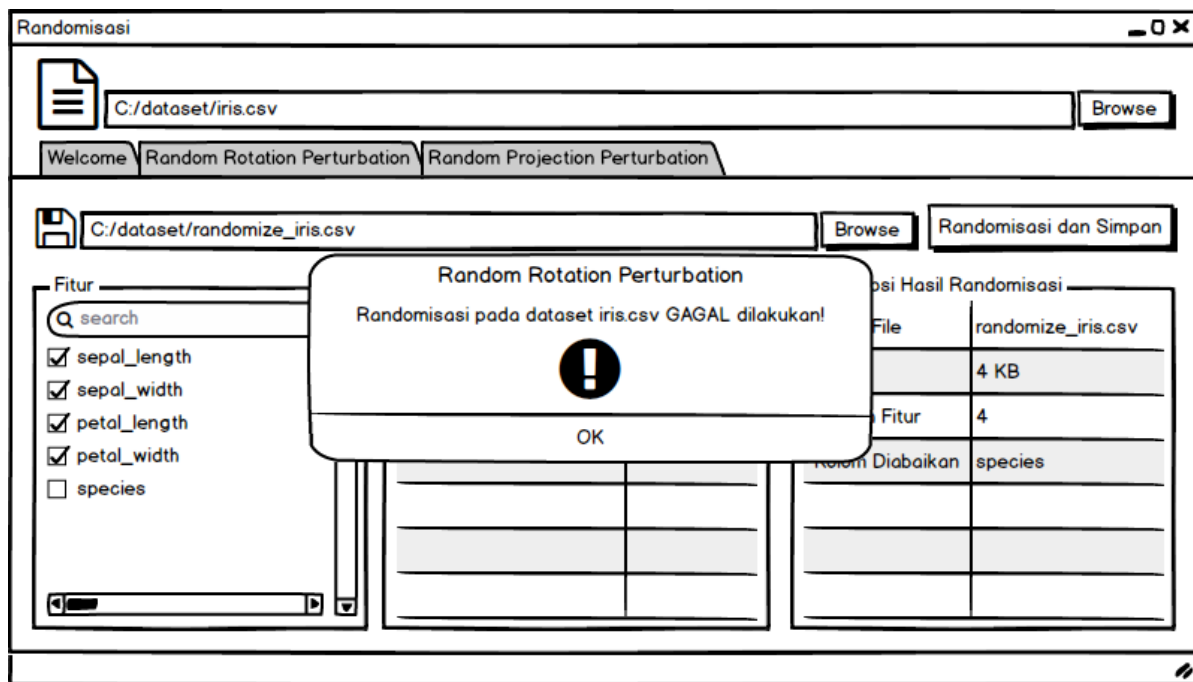
Apabila ada masalah-masalah tertentu dan randomisasi gagal dilakukan, maka perangkat lunak akan memberitahukan bahwa randomisasi telah gagal dilakukan dengan menampilkan *popup* yang berisi peringatan bahwa randomisasi gagal dilakukan. Hasil randomisasi tidak akan terbuat dan perangkat lunak tidak akan menampilkan informasi apapun tentang hasil randomisasi. *Popup* tersebut dapat dilihat pada Gambar 4.4.

4.1.2 Halaman *Random Projection Perturbation*

Halaman ini memiliki fungsi untuk melakukan teknik *Random Projection Perturbation*. Rancangan dari halaman ini dapat dilihat pada Gambar 4.5. Sebelum pengguna melakukan randomisasi, pengguna harus mengatur beberapa pengaturan yang ada. Sama seperti halnya pada halaman *Random Rotation Perturbation*, pengguna perlu menentukan lokasi penyimpanan hasil randomisasi yang akan dilakukan dan kolom apa saja pada dataset yang menjadi fitur. Pada bagian pemilihan fitur pada dataset, perangkat lunak mempunyai fitur kolom pencarian untuk mencari fitur pada dataset secara mudah dengan mengetikkan nama fitur yang diinginkan.

Selain itu, pada halaman ini yang khusus untuk teknik *Random Projection Perturbation* pengguna harus menentukan dimensi yang diinginkan dan nilai epsilon. Setelah pengaturan-pengaturan tersebut diatur, pengguna dapat melakukan randomisasi menggunakan teknik *Random Projection Perturbation* dengan menekan tombol "Randomisasi dan Simpan", perangkat lunak akan melakukan randomisasi dan menyimpan hasil randomisasi di lokasi yang telah pengguna tentukan.

Pada teknik *Random Projection Perturbation*, ada persyaratan yang harus dipenuhi mengenai dimensi akhir yang diinginkan dan nilai epsilon. Kedua nilai tersebut perlu sesuai dengan data-set yang ada sehingga pengguna tidak bisa sembarangan menentukan dimensi dan nilai epsilon. Perangkat lunak akan membuat aturan mengenai minimal dimensi yang bisa digunakan sehingga pengguna tidak bisa memasukkan dimensi yang lebih kecil dari minimal yang telah ditentukan. Minimal dimensi ini bergantung pada banyaknya baris pada dataset dan nilai epsilon seperti yang



Gambar 4.4: *Popup* yang ditampilkan apabila randomisasi gagal dilakukan

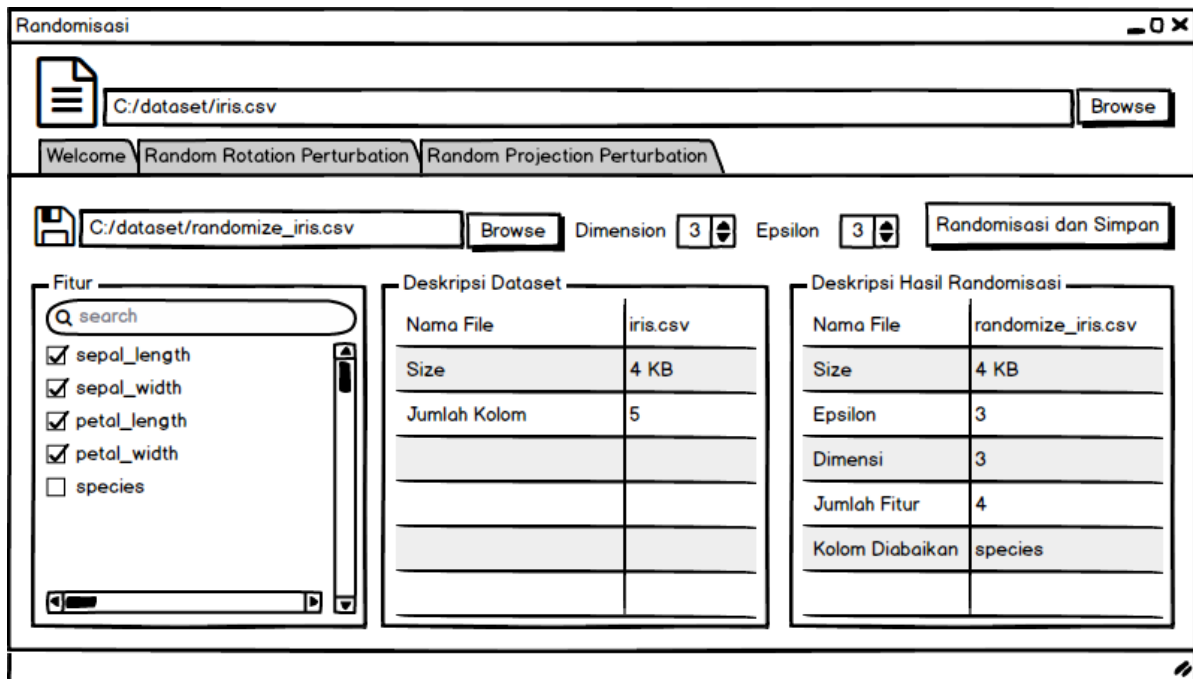
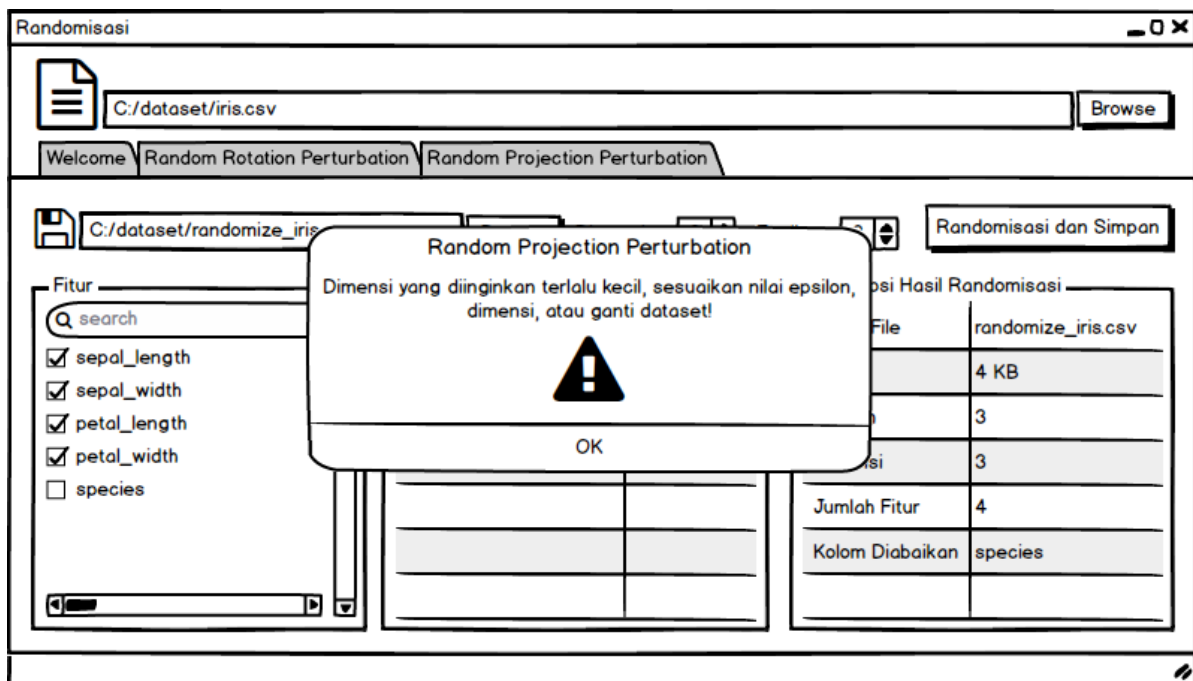
telah dijelaskan pada bagian analisis bab 3.

Apabila randomisasi berhasil dilakukan, maka perangkat lunak akan menampilkan beberapa deskripsi tentang hasil randomisasi yang berhasil dilakukan seperti nama file hasil randomisasi, ukuran file, nilai epsilon yang dipakai, jumlah dimensi pada hasil randomisasi, dan kolom-kolom apa saja yang diabaikan. Perangkat lunak juga akan menyimpan hasil randomisasi yang telah dilakukan dalam bentuk file csv di lokasi yang telah ditentukan sebelumnya oleh pengguna.

Oleh karena adanya persyaratan yang harus dipenuhi oleh pengguna untuk melakukan randomisasi dengan teknik *Random Projection Perturbation*, maka randomisasi bisa saja gagal dilakukan karena persyaratan yang ada tidak dipenuhi oleh pengguna. Persyaratan yang disebutkan adalah persyaratan jumlah dimensi dan nilai epsilon yang telah dijelaskan di atas. Perangkat lunak akan menampilkan *popup* yang memberitahukan bahwa persyaratan tidak dipenuhi dan pengguna harus mengatur kembali pengaturan atau mengganti dataset. Rancangan ini dapat dilihat pada Gambar 4.6.

4.2 Perancangan Kelas

«TODO»

Gambar 4.5: Halaman untuk melakukan teknik *Random Projection Perturbation*Gambar 4.6: *Popup* yang akan ditampilkan apabila syarat pada dataset tidak terpenuhi

DAFTAR REFERENSI

- [1] dan Osmar R. Zaïane, S. R. M. O. (2004) Towards standardization in privacy-preserving data mining. *ACM SIGKDD 3rd Workshop on Data Mining Standards*, **3**, 862–870.
- [2] NIST Special Publication 800-122 (2010) *Guide to Protecting the Confidentiality of Personally Identifiable Information (PII)*. National Institute of Standards and Technology, U.S. Department of Commerce, Erika McCallister, Tim Grance, Karen Scarfone. Gaithersburg, Maryland.
- [3] dan JOAO P. VILELA, R. M. (2017) Privacy-preserving data mining: Methods, metrics, and applications. *IEEE Access*, **5**, 10562–10582.
- [4] Han, J., Kamber, M., dan Pei, J. (2012) *Data Mining: Concepts and Techniques*, 3rd edition. Morgan Kaufmann, Waltham.
- [5] dan Somayyeh Seifi Moradi, M. K. (2011) Classification and evaluation the privacy preserving data mining techniques by using a data modification-based framework. *International Journal on Computer Science and Engineering*, **3**, 862–870.
- [6] dan L. Liu, K. C. (2005) A random rotation perturbation approach to privacy preserving data classification. Technical Report GIT-CC-05-12. Georgia Institute of Technology, Georgia.
- [7] dan R. Srikant, R. A. (2000) Privacy preserving data mining. *In Proceedings of the ACM SIGMOD*, **3**, 439–450.
- [8] STEWART, G. W. (1980) The efficient generation of random orthogonal matrices with an application to condition estimators. *SIAM Journal on Numerical Analysis*, **17**, 403–409.
- [9] Johnson, W. B. dan Lindenstrauss, J. (1984) Extensions of lipschitz mappings into a hilbert space. *Contemporary Mathematics*, **26**, 189–206.
- [10] Kun Liu, d. J. R., Hillol Kargupta (2006) Random projection-based multiplicative data perturbation for privacy preserving distributed data mining. *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*, **18**, 92–106.