

BAB 1

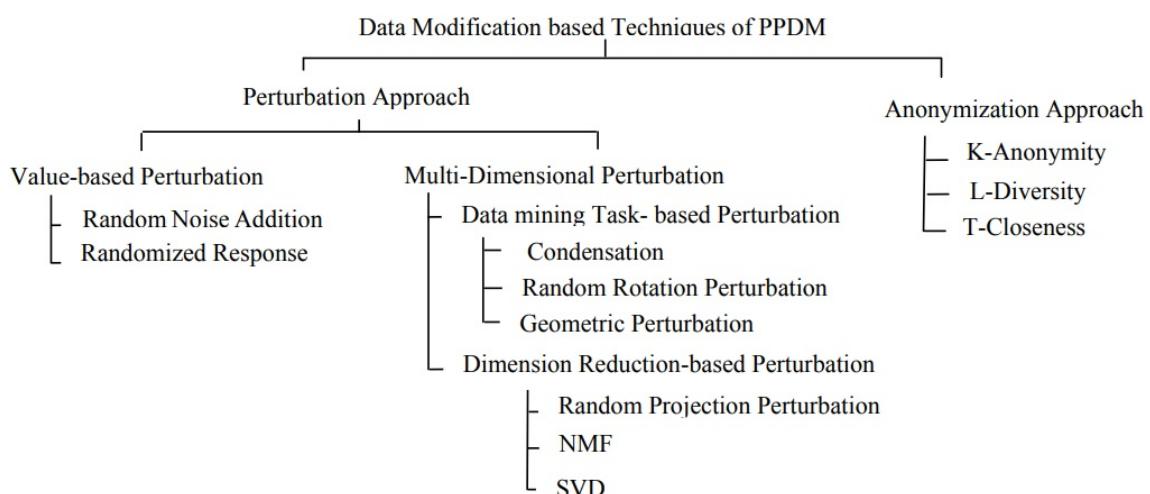
PENDAHULUAN

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 tersebut 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*, 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 dapat ditambah. *Perturbation Approach* dapat dibagi menjadi dua jenis yaitu *Value-based Perturbation Techniques* dan *Multi-Dimensional Perturbation*.



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

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 1.1, 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 ditambah menambah menjadi data yang telah dimodifikasi dengan metode *Randomization* sehingga privasi pada data tersebut terlindungi, tetapi masih dapat ditambah. 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.

1.2 Rumusan Masalah

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

1. Bagaimana cara kerja 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, consectetuer at, consectetuer 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*, *Random Rotation 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 konkret untuk menentukan data seperti apa yang menjadi privasi. Pada penambangan data, perlu ada teknik yang baik untuk menjaga privasi tidak tersebar kepada orang yang tidak berhak. Ada beberapa teknik untuk menjaga privasi pada penambangan 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 dapat 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 dapat 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 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 [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 rekord yang ada selain rekord yang ingin diprediksi labelnya
 - (a) Hitung jarak Euclidean antara rekord iterasi sekarang dengan rekord yang ingin diprediksi labelnya
 - (b) Catat jarak Euclidean dari rekord yang ingin diprediksi dan indeks rekord iterasi sekarang
3. Urutkan jarak Euclidean titik-titik yang sudah dihitung pada perulangan pada langkah sebelumnya secara menaik
4. Pilih rekord teratas (jarak Euclidean yang paling kecil) sebanyak k dari urutan pada langkah sebelumnya
5. Ambil label dari semua rekord yang terpilih pada langkah sebelumnya. Label terbanyak adalah hasil prediksi label pada rekord yang ingin diprediksi

2.2.2 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 [3].

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*. [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 1.1 [5]. *Perturbation Approach* adalah pendekatan untuk *privacy preserving data mining* dengan cara mengacaukan data yang ada, tetapi hasil data yang dikacaukan masih tetap dapat ditambah. Sedangkan pada *Anonymization Approach*, data diterapkan de-identifikasi di mana dataset mentah disebarluaskan setelah menghapus inti dari identitas setiap rekord [5].

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 akurasinya sama baiknya dengan tanpa menerapkan teknik *Perturbation Approach*. Setiap teknik penambangan data memakai properti yang berbeda-beda pada data yang ditambah. 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 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.

2.4 Metode *Randomization*

Dari berbagai macam teknik modifikasi data untuk *privacy preserving data mining* yang dapat dilihat pada Gambar 1.1, 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 menambah 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. Masing-masing dari dua teknik tersebut akan dijelaskan lebih detil pada subbab-subbab berikutnya.

2.4.1 *Random Noise Addition*

Ide utama dari teknik *Random Noise Addition* [7] 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 2.1.

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 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

```

(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 2.1: Algoritma rekonstruksi

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*.

2.4.2 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 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 [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 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.

2.4.3 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 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.

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

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 $(0, 1)$.

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

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

Pada bab ini akan dijelaskan analisis yang telah dilakukan terhadap *privacy preserving data mining*, teknik randomisasi, teknik penambangan data, dan gambaran umum perangkat lunak. Algoritma teknik randomisasi akan dijelaskan secara rinci beserta studi kasus yang telah dilakukan. Perancangan perangkat lunak akan dijelaskan melalui diagram aktivitas dan diagram kelas yang telah dibuat berdasarkan analisis yang telah dilakukan.

3.1 *Privacy Preserving Data Mining*

Dengan banyaknya data yang disebarluaskan untuk tujuan penambangan data, timbul masalah yaitu adanya potensi data tersebut memiliki data privasi seseorang sehingga privasi tersebut tersebar kepada pihak yang sebenarnya tidak berhak mengetahuinya. Oleh karena itu, perlu ada cara agar data yang ingin ditambah bisa dihilangkan privasinya. Privasi yang perlu dijaga antara lain mengenai identitas seseorang atau hal yang dapat dikaitkan terhadap identitas seseorang. Metode randomisasi menjadi salah satu solusi untuk menghilangkan privasi yang ada pada data dengan mengacak data tersebut 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 yaitu 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. Oleh karena itu, kedua teknik randomisasi akan diuji terhadap data *real* dan dilakukan penambangan data dengan teknik klasifikasi yaitu *K-nearest neighbors* dan teknik *clustering* yaitu *K-means* terhadap data *real* tersebut yang telah dirandomisasi.

Teknik *Random Rotation Perturbation* akan mengacak data dengan cara merotasikan seluruh data yang direpresentasikan sebagai titik pada bidang Euclidean sehingga nilainya akan berubah tetapi jarak Euclidean antara setiap titik tidak akan berubah. Teknik *Random Projection Perturbation* akan mengacak data dengan cara memproyeksikan data ke dimensi yang lebih kecil dengan menggunakan matriks acak. Teknik ini berdasarkan pada lemma *Johnson-Lindenstrauss* yang menyatakan bahwa data yang cukup besar dapat direduksi dimensinya sedemikian rupa tanpa merusak jarak Euclidean secara signifikan. Oleh karena itu, dipilih teknik penambangan data yang hanya bergantung pada jarak Euclidean yaitu *K-nearest neighbors* dan *K-means*. Kedua teknik penambangan data tersebut akan dipakai untuk menguji coba keberhasilan kedua teknik randomisasi dan untuk membandingkan kualitas hasil dari kedua teknik randomisasi tersebut.

Teknik randomisasi dan penambangan data akan diimplementasikan dengan bantuan bahasa pemograman Python beserta berbagai *library* seperti Pandas¹, Numpy², dan Scikit-learn³. Perangkat lunak randomisasi akan dibangun beserta antarmukanya. Perangkat lunak untuk menerapkan

¹<https://pandas.pydata.org/>

²<https://numpy.org/>

³<https://scikit-learn.org/stable/>

penambangan data akan dijalankan dengan bantuan perangkat lunak Spyder⁴ dan Anaconda⁵ untuk menampilkan visualisasi dan teks yang dihasilkan oleh perangkat lunak penambangan data yang berbahasa pemrograman Python.

3.2 Random Rotation Perturbation

Seperti yang telah dijelaskan di bab 2, ide dari teknik *Random Rotation Perturbation* adalah merotasi seluruh data yang direpresentasikan sebagai titik pada bidang Euclidean sehingga jarak antara titik-titik yang ada tidak berubah walaupun nilai tiap titik berubah secara drastis. Berikut akan dijelaskan analisis, algoritma, dan studi kasus yang telah dilakukan pada teknik *Random Rotation Perturbation*.

3.2.1 Analisis

Teknik *Random Rotation Perturbation* menggunakan matriks orthogonal untuk mengacak dataset tanpa mengubah jarak Euclidean dataset tersebut. Matriks orthogonal relatif sulit dibuat pada dimensi yang tinggi dan kompleksitasnya tidak kecil yaitu $O(n^2)$ [8]. Untuk dataset yang sangat besar, masih dapat dilakukan randomisasi dengan teknik ini walaupun mungkin akan sedikit memakan waktu. Hal ini akan diuji pada bab 5 saat pengujian eksperimental.

Walaupun seperti itu, teknik ini dapat menjamin bahwa tidak akan ada *error* yang terjadi pada jarak Euclidean pada seluruh data atau dengan kata lain jarak Euclidean tidak berubah sama sekali. Hal ini disebabkan oleh metode yang dipakai hanya dengan merotasi seluruh data yang direpresentasikan sebagai titik dalam bidang Euclidean. Transformasi rotasi akan merubah nilai setiap titik pada bidang tersebut tetapi karena pada dasarnya setiap titik bergerak dengan gerakan yang sama yaitu sebuah rotasi yang sama sehingga tidak akan ada perubahan pada jarak Euclidean antara seluruh titik-titik yang ada.

Transformasi rotasi mempunyai kelemahan yaitu apabila ada titik pada bidang Euclidean yang nilainya kecil mendekati 0 maka walaupun dirotasi dengan rotasi yang berbeda-beda, bagaimanapun juga titik tersebut akan tetap bernilai kecil atau berada dekat dengan nilai 0. Hal ini berpotensi menjadi sebuah faktor yang lemah untuk melindungi privasi [6]. Oleh karena itu transformasi translasi perlu dilakukan sebelum melakukan rotasi agar titik tersebut tidak selalu mendekati nilai 0 saat dirotasi dengan berbagai macam rotasi. Transformasi translasi akan dilakukan dengan melakukan translasi yang nilainya diambil secara acak pada rentang [0, 100] sehingga translasi yang dilakukan tidak bisa diketahui oleh orang yang tidak berhak.

3.2.2 Algoritma

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

1. Dataset yang memiliki atribut sebanyak d dan rekord sebanyak n direpresentasikan dalam bentuk matriks berukuran $n \times d$
2. Buatlah matriks translasi acak yang diambil mengikuti distribusi *uniform* dengan rentang [0, 100] berdimensi $(d + 1) \times (d + 1)$
3. Untuk keperluan transformasi translasi, matriks dataset perlu ditambahkan sebuah kolom dengan nilai 1 pada seluruh barisnya.
4. Lakukan transformasi translasi dengan cara mengkalikan matriks dataset dengan matriks translasi yang telah dibuat pada langkah kedua

⁴<https://www.spyder-ide.org/>

⁵<https://www.anaconda.com/>

Tabel 3.1: 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

5. 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 $n \times d$
6. Buatlah *random rotation matrix* berukuran $d \times d$ dengan membuat matriks *orthogonal* acak. Matriks *orthogonal* memiliki sifat yaitu determinannya sebesar 1 dan hasil perkalian matriks tersebut dengan transposenya adalah matriks identitas
7. Lakukan transformasi rotasi dengan cara mengkalikan matriks dataset dengan *random rotation matrix* yang telah dibuat pada langkah keenam
8. Hasil matriks yang telah dirotasi sudah dapat langsung digunakan untuk penambangan data

3.2.3 Studi Kasus

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 3.1. 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 3.1.

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

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

2. Membuat matriks translasi yang diambil mengikuti distribusi *uniform* dengan rentang [0,100]

dengan dimensi sesuai dimensi matriks dataset

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

3. Untuk keperluan translasi, matriks dataset ditambahkan sebuah kolom dengan nilai 1 pada setiap barisnya

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

4. Dilakukan transformasi translasi dengan matriks translasi yang telah dibuat pada langkah sebelumnya dengan cara mengkalikan matriks dataset dengan matriks translasi

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

5. Berikut adalah hasil translasi pada matriks dataset. Dapat dilihat kolom terakhir adalah kolom yang harus dibuang karena kolom tersebut ada hanya untuk melakukan transformasi translasi yang sudah dilakukan pada langkah sebelumnya

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

6. Berikutnya matriks rotasi dibuat dengan cara membuat matriks spesial orthogonal yang berdimensi sesuai dimensi matriks dataset. Matriks rotasi berikut dibuat dengan menggunakan

library Scipy⁶ pada bahasa pemrograman Python

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

7. Dilakukan transformasi rotasi dengan matriks rotasi yang telah dibuat pada langkah sebelumnya dengan cara mengkalikan matriks dataset dengan matriks rotasi
8. Berikut adalah hasil rotasi pada matriks dataset. Hasil teknik *Random Rotation Perturbation* pada dataset *iris* ini sudah dapat langsung digunakan untuk dilakukan penambangan data

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

3.3 Random Projection Perturbation

Ide dari teknik *Random Projection Perturbation* seperti yang telah dijelaskan pada bab 2 adalah mereduksi dimensi dataset sehingga nilai pada setiap kolom akan berubah bahkan kolomnya akan berkurang yang mengakibatkan kolom-kolom yang ada tidak bisa diketahui kolom tersebut adalah kolom apa. Berikut akan dijelaskan analisis, algoritma, dan studi kasus yang telah dilakukan pada teknik *Random Projection Perturbation*.

3.3.1 Analisis

Teknik *Random Projection Perturbation* didasarkan pada lemma *Johnson-Lindenstrauss* yang mengatakan bahwa sejumlah titik pada bidang Euclidean berdimensi tertentu dapat diproyeksikan ke dimensi yang lebih kecil tanpa mengubah secara signifikan jarak Euclidean antara titik-titik tersebut. *Error* terburuk yang dapat terjadi pada jarak Euclidean setelah proyeksi diterapkan ditentukan oleh nilai variabel Epsilon dengan pertidaksamaan sebagai berikut menunjukkan rentang jarak Euclidean data setelah diproyeksi.

$$(1 - \epsilon) \|u - v\|^2 < \|p(u) - p(v)\|^2 < (1 + \epsilon) \|u - v\|^2 \quad (3.1)$$

Pertidaksamaan di atas menyatakan bahwa kuadrat dari jarak Euclidean antara dua buah titik setelah diproyeksi tidak akan kurang dari kuadrat jarak Euclidean aslinya dikalikan $(1 - \epsilon)$ dan tidak akan lebih dari kuadrat jarak Euclidean aslinya dikalikan $(1 + \epsilon)$. Lemma *Johnson-Lindenstrauss* menjamin bahwa jarak Euclidean pada data setelah diproyeksi hanya akan berada pada rentang tersebut dan bisa saja jarak Euclidean setelah diproyeksi sangat dekat dengan jarak Euclidean aslinya karena belum tentu jarak Euclidean data setelah diproyeksi menyentuh batas pertidaksamaan tersebut. Tetapi pertidaksamaan ini juga dapat menyatakan bahwa jarak Euclidean setelah diproyeksi dengan aslinya tidak mungkin sama persis.

⁶http://scipy.github.io/devdocs/generated/scipy.stats.special_ortho_group.html

Agar pertidaksamaan sebelumnya berlaku ada persyaratan yang harus dipenuhi yaitu nilai minimal variabel k atau dengan kata lain target dimensi terkecilnya proyeksi dilakukan harus memenuhi persamaan berikut.

$$k \geq 4(\epsilon^2/2 - \epsilon^3/3)^{-1} \ln n \quad (3.2)$$

Variabel n pada persamaan tersebut adalah jumlah titik pada bidang Euclidean atau dengan kata lain jumlah baris pada dataset yang ingin dirandomisasi. Dengan melihat persamaan tersebut maka bisa disimpulkan bahwa dimensi minimal sebuah dataset diproyeksikan berbanding lurus dengan jumlah baris pada dataset tersebut. Apabila sebuah dataset diproyeksikan ke dimensi paling minimal dan diterapkan teknik penambangan data sehingga menghasilkan sebuah model penambangan data dengan data latihannya adalah dataset tersebut maka model tersebut akan melanggar persamaan di atas jika model tersebut dilatih kembali dengan data yang baru karena artinya bidang Euclidean yang disebutkan tadi akan memiliki titik yang lebih banyak, nilai n yang lebih besar.

Oleh karena itu, berdasarkan analisis di atas teknik *Random Projection Perturbation* tidak relevan untuk model penambangan data yang memiliki sangat banyak data tetapi berdimensi relatif kecil dibandingkan dengan jumlah datanya. Tetapi masalah ini dapat dihindari dengan tidak mereduksi dimensi ke dimensi yang sangat kecil atau paling minimal sehingga tidak terlalu cepat nilai minimal variabel k meningkat sampai menyulut dimensi dataset setelah proyeksi. Selain itu, jarak Euclidean setelah diproyeksi juga belum tentu pasti menyentuh batas pertidaksamaan yang di atas sehingga belum tentu hasil randomisasi memiliki *error* yang sangat besar apabila besar dimensi melebihi nilai minimal variabel k . Hal ini akan diuji pada bab 5 saat melakukan pengujian eksperimental.

Teknik *Random Projection Perturbation* memiliki kompleksitas yang relatif kecil yaitu $O(dkn)$ [11]. Dengan kompleksitas yang cukup rendah maka teknik ini dapat bekerja secara cepat untuk penambangan data dengan dataset yang besar sekalipun. Oleh karena itu, teknik ini berpotensi dapat berkerja dengan baik dalam lingkungan *big data* karena kompleksitas yang relatif kecil dan dapat bekerja pada dataset yang sangat besar. Hal ini disebabkan oleh metode yang digunakan oleh teknik ini adalah melakukan proyeksi dengan cara mengkalikan dataset yang sudah berbentuk matriks dengan matriks proyeksi yang berupa matriks acak saja, tidak ada sifat spesial pada matriks acak tersebut seperti teknik *Random Rotation Perturbation* sehingga pembuatan matriks proyeksinya dapat dilakukan dengan cepat.

3.3.2 Algoritma

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

1. Dataset yang memiliki atribut sebanyak d dan rekord sebanyak n direpresentasikan dalam bentuk matriks berukuran $n \times d$
2. Tentukan nilai variabel ϵ (epsilon) yang diinginkan dan berada pada rentang $(0, 1)$
3. Hitung nilai minimal variabel k (dimensi minimal) dengan rumus berikut $k \geq 4(\epsilon^2/2 - \epsilon^3/3)^{-1} \ln n$
4. Tentukan nilai variabel k yang diinginkan dengan memenuhi persyaratan pada langkah ketiga dan nilai variabel k yang dipilih harus lebih kecil dari d (dimensi dataset aslinya)
5. Buatlah matriks proyeksi berukuran $d \times k$ dengan cara membuat matriks acak yang diambil mengikuti distribusi normal dengan rata-rata bernilai 0 dan standar deviasi bernilai $1/\sqrt{k}$
6. Lakukan proyeksi dengan cara mengkalikan matriks dataset dengan matriks proyeksi yang telah dibuat pada langkah kelima
7. Hasil matriks yang telah diproyeksi sudah dapat langsung digunakan untuk penambangan data

3.3.3 Studi Kasus

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 3.1. 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 3.1.

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

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

2. Ditentukan nilai ϵ (epsilon) yang diinginkan adalah 0.5. Dengan rumus berikut dapat dihitung untuk melihat seberapa besar distorsi yang terjadi pada jarak Euclidean sebuah titik/baris data dengan menunjukkan rentang jarak Euclidean data setelah diproyeksi. Contoh yang akan diberikan berikut menguji rentang jarak Euclidean antara baris ke-6 dan ke-9 setelah diproyeksi

$$\begin{aligned} \|u - v\|^2 &= (5.4 - 4.4)^2 + (3.9 - 2.9)^2 + (1.7 - 1.4)^2 + (0.4 - 0.2)^2 \\ &= 2.13 \end{aligned}$$

$$\begin{aligned} (1 - \epsilon) \|u - v\|^2 &< \|p(u) - p(v)\|^2 < (1 + \epsilon) \|u - v\|^2 \\ (1 - 0.5)2.13 &< \|p(u) - p(v)\|^2 < (1 + 0.5)2.13 \\ 1.065 &< \|p(u) - p(v)\|^2 < 3.195 \end{aligned}$$

3. Nilai minimal variabel k (dimensi minimal) dihitung dengan rumus berikut

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

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

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

6. Dilakukan proyeksi dengan cara mengkalikan matriks dataset dengan matriks proyeksi yang telah dibuat pada langkah sebelumnya

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

7. Berikut adalah hasil matriks yang telah diproyeksi. Hasil dari teknik *Random Projection Perturbation* pada dataset *iris* ini sudah dapat langsung digunakan untuk dilakukan penambangan data

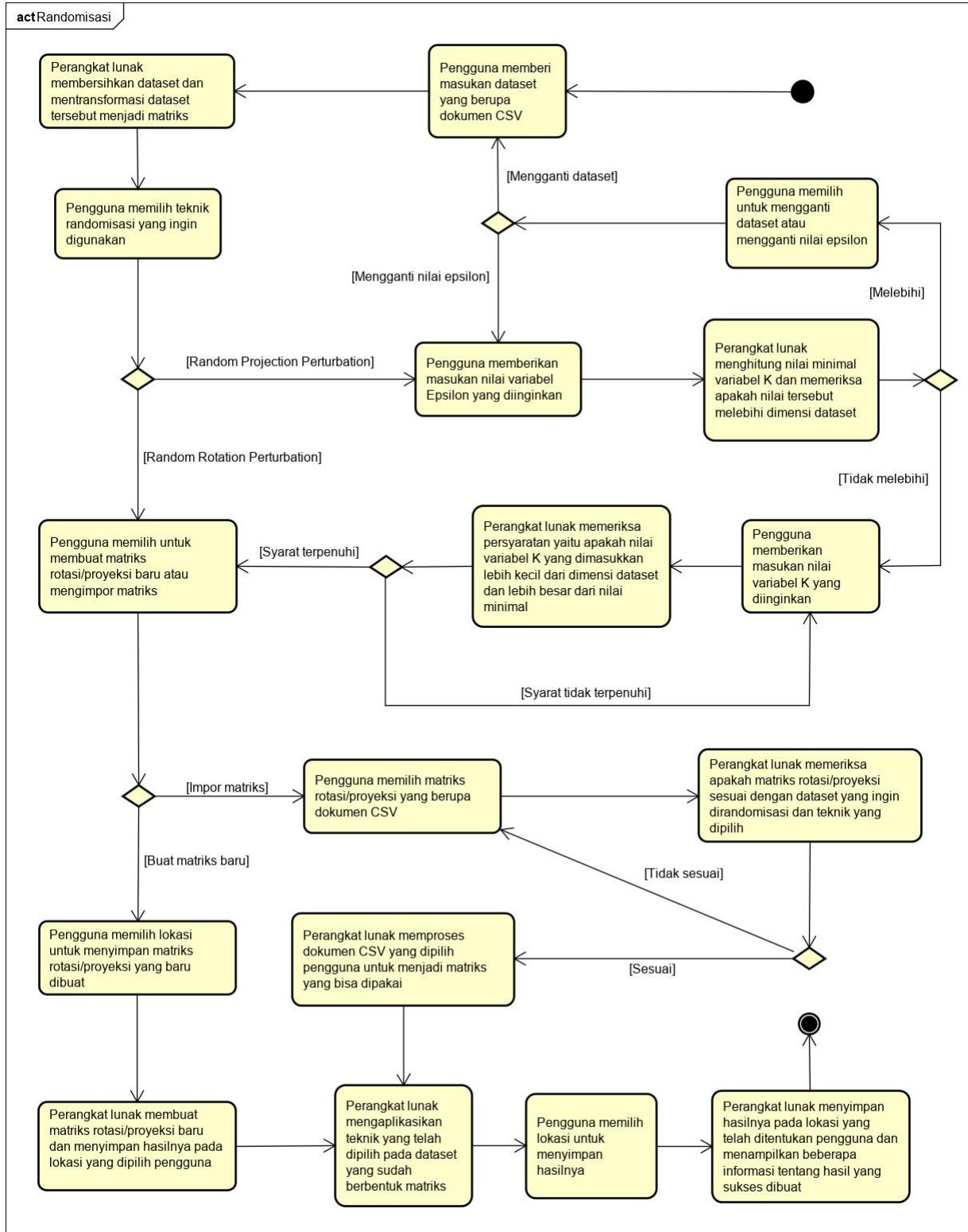
$$\begin{bmatrix} 0.65684027 & -1.0035495 & 0.72820632 \\ 0.60194004 & -0.90822018 & 0.66416944 \\ 0.60217727 & -0.92172316 & 0.66620218 \\ 0.56344827 & -0.88887716 & 0.65931422 \\ 0.6517441 & -1.00838106 & 0.7317004 \\ 0.67922914 & -1.09633771 & 0.76805051 \\ 0.58987895 & -0.9446188 & 0.66701567 \\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

3.4 Perangkat Lunak

Ada beberapa persyaratan yang harus dipenuhi pada implementasi perangkat lunak seperti alur, masukan, keluaran, dan antarmuka perangkat lunak. Hal-hal tersebut harus disesuaikan dengan kebutuhan dan tujuan dibuatnya perangkat lunak ini. Oleh karena itu, perlu adanya perancangan terlebih dahulu untuk menjadi gambaran bagaimana perangkat lunak yang akan dibuat akan berfungsi. Pada subbab ini akan ditunjukkan gambaran singkat perancangan perangkat lunak yang akan dibuat berupa diagram aktivitas dan diagram kelas.

3.4.1 Diagram Aktivitas

Perangkat lunak randomisasi adalah perangkat lunak yang digunakan untuk memodifikasi data dengan metode randomisasi. Perangkat lunak ini akan memiliki fungsi untuk mengubah nilai



Gambar 3.1: Diagram aktivitas perangkat lunak randomisasi

setiap data yang dimasukan agar privasinya terjaga tetapi masih dapat dilakukan penambangan 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 3.1. Detail dari diagram aktivitas tersebut adalah sebagai berikut.

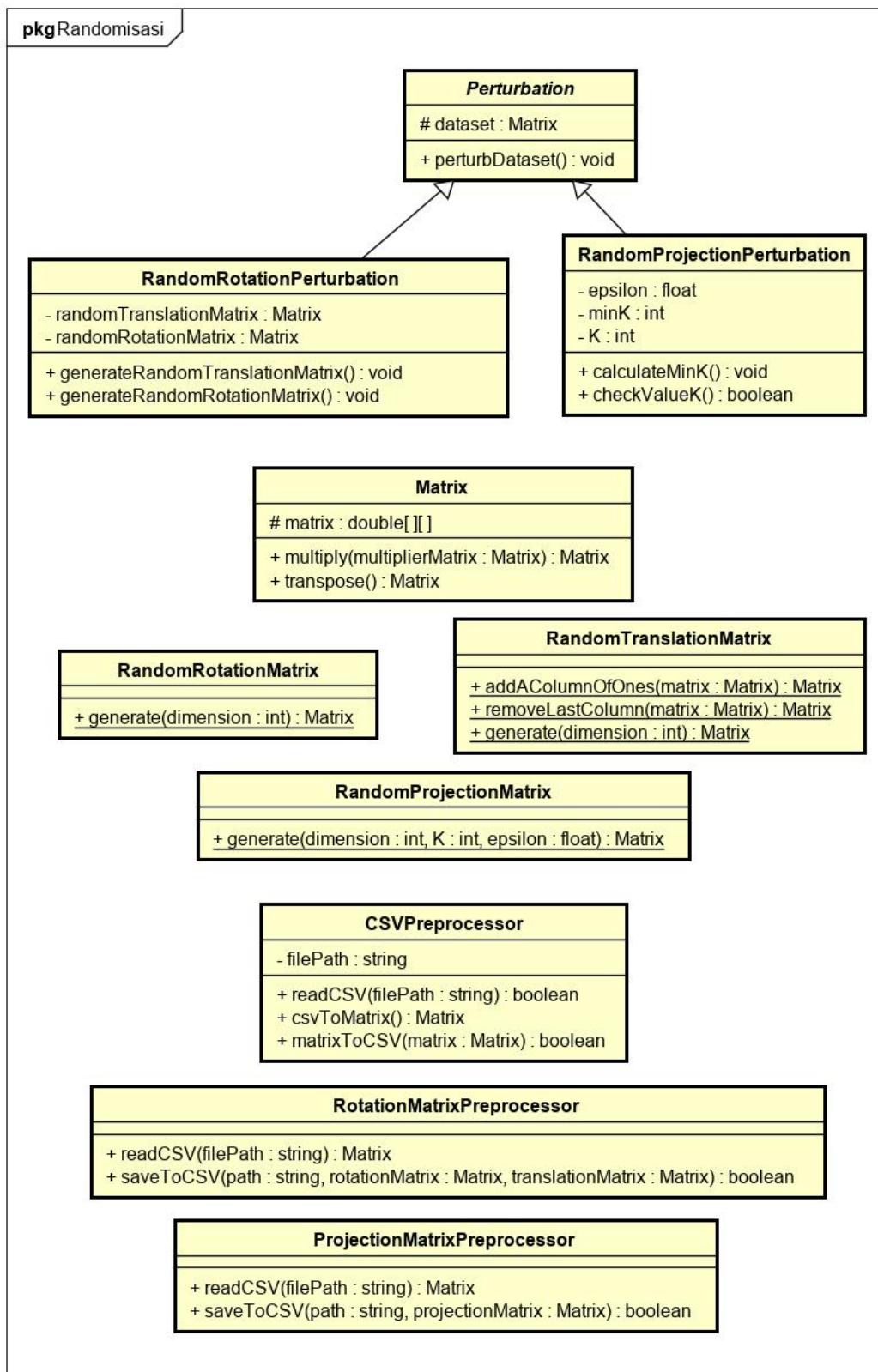
1. Pengguna memberikan masukan berupa dataset yang berupa dokumen berjenis *comma-separated values* (CSV). Dokumen ini harus berisi tiap rekord pada barisnya dan tiap fitur pada kolomnya. Selain itu, pada baris pertama dalam dokumen tersebut harus berupa nama setiap fitur pada setiap kolomnya
2. Perangkat 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
3. 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.
 - (a) Pengguna memberi masukan nilai variabel Epsilon yang diinginkan. Variabel ini akan menentukan *error* terburuk yang dapat terjadi pada hasil randomisasi dan menentukan nilai minimal variabel K
 - (b) Perangkat lunak memeriksa persyaratan yang harus dipenuhi oleh matriks dataset dengan menghitung menggunakan algoritma tertentu dan menghitung nilai minimal variabel K dan menampilkannya kepada pengguna. Pengecekan yang dilakukan adalah pengecekan jumlah kolom/dimensi pada matriks dataset apakah cukup untuk matriks tersebut direduksi dimensinya.
 - (c) Jika persyaratan tidak terpenuhi maka pengguna harus memilih untuk mengganti datasetnya atau mengganti nilai Epsilon
 - (d) Jika persyaratan terpenuhi maka berikutnya pengguna memberikan masukan nilai variabel K yang diinginkan. Nilai yang diberikan pengguna harus lebih dari nilai minimal variabel K yang sudah dihitung oleh perangkat lunak pada langkah sebelumnya
 - (e) Perangkat lunak memeriksa persyaratan teknik yaitu nilai variabel K yang diberikan harus lebih kecil dari dimensi dataset dan lebih besar dari nilai minimal variabel K.
 - (f) Jika persyaratan tidak terpenuhi maka pengguna harus mengganti nilai variabel K kembali.
4. Pengguna memilih untuk membuat matriks rotasi/proyeksi baru sesuai teknik yang dipilih atau pengguna dapat mengimpor matriks yang telah dibuat dan disimpan sebelumnya
5. Berikut ini langkah-langkah tambahan yang harus dilakukan pengguna apabila pengguna memilih untuk membuat matriks rotasi/proyeksi baru
 - (a) Pengguna memilih lokasi direktori pada komputer pengguna sebagai lokasi untuk menyimpan matriks rotasi/proyeksi yang akan dibuat
 - (b) Perangkat lunak membuat matriks rotasi/proyeksi baru dan menyimpan hasilnya pada lokasi yang telah dipilih pengguna dalam bentuk dokumen berjenis CSV
6. Berikut ini langkah-langkah tambahan yang harus dilakukan pengguna apabila pengguna memilih untuk mengimpor matriks rotasi/proyeksi yang pernah dibuat sebelumnya

- (a) Pengguna memilih matriks rotasi atau proyeksi sesuai teknik yang dipilih berupa dokumen berjenis CSV pada sebuah direktori komputer pengguna yang dipilih
 - (b) Perangkat lunak memeriksa dokumen CSV yang dipilih berisi matriks rotasi/proyeksi yang sesuai dengan dataset dan teknik randomisasi yang dipilih
 - (c) Jika tidak sesuai maka pengguna harus memilih kembali matriks rotasi/proyeksi dengan benar
 - (d) Jika sudah sesuai dan memenuhi persyaratan maka perangkat lunak akan memproses dokumen CSV yang dipilih pengguna untuk menjadi matriks rotasi/proyeksi yang dapat dipakai untuk teknik randomisasi yang dipilih
7. Perangkat lunak mengaplikasikan teknik yang telah dipilih pada dataset yang sudah berbentuk matriks dengan memakai matriks rotasi/proyeksi yang telah dipilih
8. Pengguna memilih lokasi direktori pada komputer pengguna untuk menyimpan hasil randomisasi dari teknik yang dipilih. Hasilnya adalah sebuah dokumen *comma-separated values* yang berisi dataset yang sudah diterapkan teknik randomisasi
9. Perangkat lunak menyimpan hasilnya pada lokasi yang telah ditentukan pengguna dalam bentuk dokumen berjenis *comma-separated values* dan menampilkan beberapa informasi tentang hasil yang sukses dibuat

3.4.2 Diagram Kelas

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 3.2 adalah sebagai berikut.

- Kelas *Perturbation* adalah kelas abstrak yang akan di-*extend* oleh kelas yang mengimplementasikan algoritma *Random Rotation Perturbation* dan *Random Projection Perturbation*. Kelas ini mempunyai sebuah atribut yaitu *dataset* yang berfungsi untuk menyimpan dataset yang ingin dirandomisasi. Ada sebuah fungsi pada kelas ini yaitu *perturbDataset* yang berfungsi untuk menerapkan teknik randomisasi pada dataset yang ingin dirandomisasi dengan kata lain atribut *dataset*
- Kelas *RandomRotationPerturbation* adalah kelas turunan dari kelas *Perturbation* yang bertujuan untuk mengimplementasikan algoritma *Random Rotation Perturbation*. Kelas ini memiliki dua tambahan atribut yaitu *randomTranslationMatrix* dan *randomRotationMatrix* yang masing-masing memiliki fungsi untuk menyimpan matriks translasi dan matriks rotasi. Ada dua buah fungsi tambahan juga pada kelas ini yaitu *generateRandomTranslationMatrix* dan *generateRandomRotationMatrix* yang masing-masing memiliki fungsi untuk membuat matriks translasi dan matriks rotasi baru.
- Kelas *RandomProjectionPerturbation* adalah kelas turunan dari kelas *Perturbation* yang bertujuan untuk mengimplementasikan algoritma *Random Projection Perturbation*. Kelas ini memiliki tiga tambahan atribut yaitu *epsilon*, *minK*, dan *K* yang masing-masing memiliki fungsi untuk menyimpan nilai variabel epsilon, nilai minimal variabel K, dan nilai variabel K yang akan dipakai untuk menerapkan algoritma *Random Projection Perturbation*. Ada dua buah fungsi tambahan juga pada kelas ini yaitu *calculateMinK* dan *checkValueK* yang masing-masing memiliki fungsi untuk menghitung nilai minimal variabel K dan memeriksa persyaratan pada nilai atribut *K*.



Gambar 3.2: Diagram kelas perangkat lunak randomisasi

- Kelas *Matrix* adalah kelas untuk merepresentasikan matriks dan menyimpan nilai-nilai pada setiap elemen matriks. Kelas ini memiliki sebuah atribut yaitu *matrix* yang berfungsi untuk menyimpan setiap elemen matriks yang ada. Kelas ini juga memiliki dua buah fungsi yaitu *multiply* dan *transpose* yang masing-masing berfungsi untuk melakukan operasi perkalian dan transpose matriks yang akan digunakan untuk implementasi kedua algoritma teknik randomisasi
- Kelas *RandomTranslationMatrix* adalah kelas yang mempunyai tujuan utama untuk membuat matriks translasi secara acak. Kelas ini adalah kelas statis sehingga tidak bisa diinstansiasi dan hanya memiliki atribut atau fungsi statis saja. Ada tiga buah fungsi statis yaitu *addAColumnOfOnes*, *removeLastColumn*, dan *generate*. Fungsi *addAColumnOfOnes* memiliki fungsi untuk menambahkan sebuah kolom yang pada setiap baris memiliki nilai berupa angka 1 kepada sebuah matriks. Fungsi *removeLastColumn* memiliki fungsi untuk menghapus kolom terakhir pada suatu matriks. Fungsi *generate* adalah fungsi yang bertujuan untuk membuat matriks translasi baru secara acak
- Kelas *RandomRotationMatrix* adalah kelas yang mempunyai tujuan utama untuk membuat matriks rotasi secara acak. Kelas ini adalah kelas statis sehingga tidak bisa diinstansiasi dan hanya memiliki atribut atau fungsi statis saja. Hanya ada sebuah fungsi pada kelas ini yaitu *generate* yang memiliki fungsi untuk membuat matriks rotasi baru secara acak
- Kelas *CSVPreprocessor* adalah kelas untuk menangani masukan dataset berupa dokumen *comma-separated values* yang akan direpresentasikan menjadi matriks. Kelas ini berguna untuk mengkonversi dokumen berjenis *comma-separated values* menjadi matriks dan sebaliknya. Hanya ada sebuah atribut pada kelas ini yaitu *filePath* yang berfungsi untuk menyimpan lokasi dokumen yang ingin diproses menjadi matriks. Ada tiga buah fungsi pada kelas ini yaitu *readCSV*, *csvToMatrix*, dan *matrixToCSV* yang masing-masing berfungsi untuk membaca dokumen berjenis *comma-separated values*, mengkonversi dokumen *comma-separated values* menjadi matriks, dan mengkonversi matriks menjadi dokumen *comma-separated values*.
- Kelas *RotationMatrixPreprocessor* adalah kelas untuk menangani masukan matriks rotasi yang digabung dengan matriks translasi berupa sebuah dokumen *comma-separated values* yang akan dikonversi menjadi dua buah matriks (rotasi dan translasi) dan sebaliknya. Ada dua buah fungsi pada kelas ini yaitu *readCSV* dan *saveToCSV*. Fungsi *readCSV* berfungsi untuk membaca dokumen berjenis *comma-separated values* untuk dikonversi menjadi matriks rotasi dan matriks translasi. Fungsi *SaveToCSV* berfungsi untuk mengkonversi matriks rotasi dan matriks translasi menjadi sebuah dokumen berjenis *comma-separated values*.
- Kelas *ProjectionMatrixPreprocessor* adalah kelas untuk menangani masukan matriks proyeksi berupa sebuah dokumen *comma-separated values* yang akan dikonversi menjadi sebuah matriks. Ada dua buah fungsi pada kelas ini yaitu *readCSV* dan *saveToCSV*. Fungsi *readCSV* berfungsi untuk membaca dokumen berjenis *comma-separated values* untuk dikonversi menjadi matriks proyeksi. Fungsi *SaveToCSV* berfungsi untuk mengkonversi matriks proyeksi menjadi sebuah dokumen berjenis *comma-separated values*.

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 dokumen berjenis *comma-separated values* yang ingin dirandomisasi. Pertama-tama pengguna wajib untuk mengisi kolom tersebut apapun teknik randomisasi yang akan digunakan nanti.

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

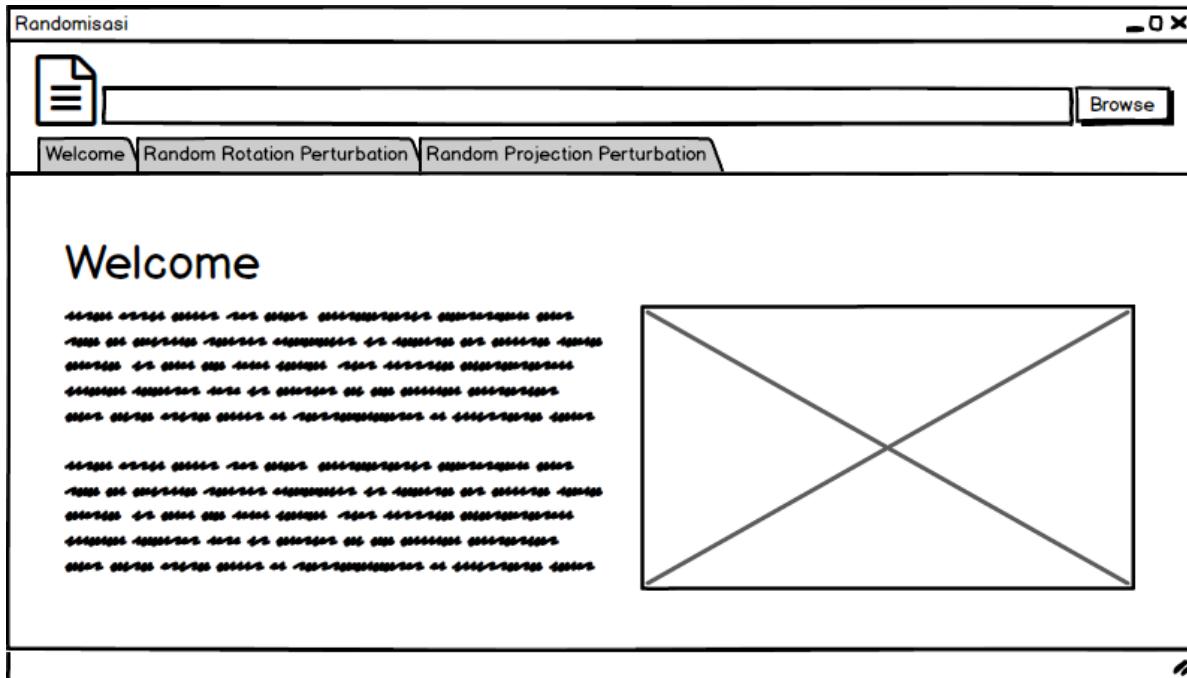
4.1.1 Halaman *Random Rotation Perturbation*

Halaman ini memiliki fungsi untuk melakukan teknik *Random Rotation Perturbation*. Setelah user memasukkan dokumen 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.

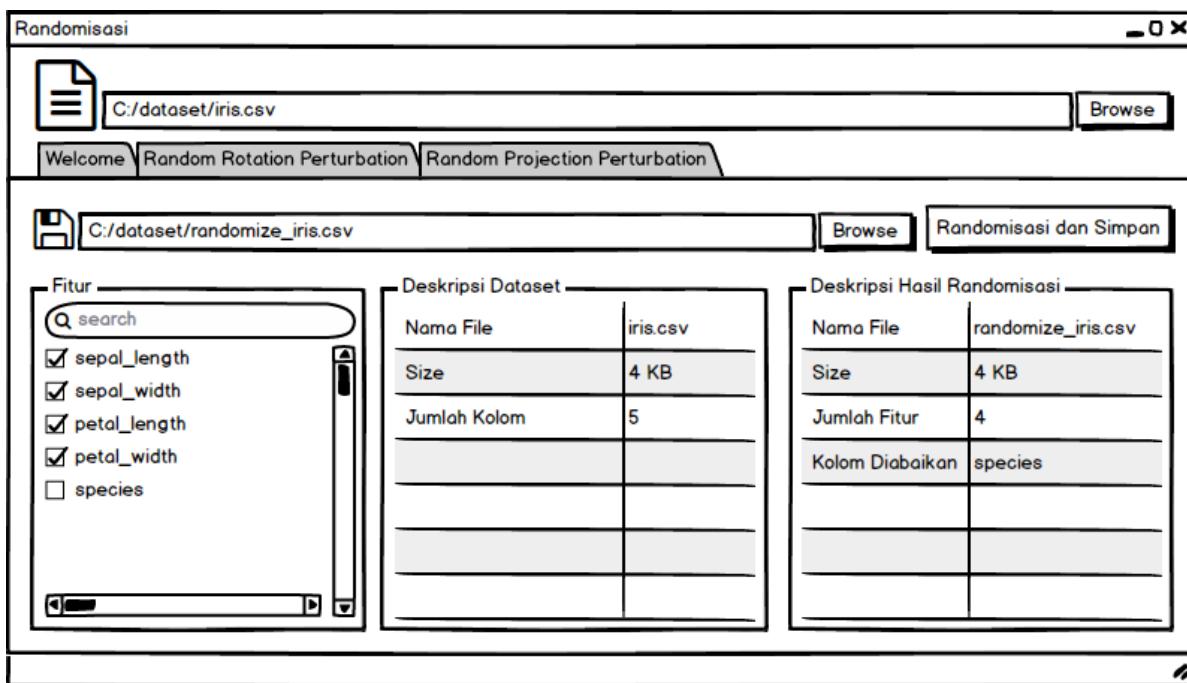
Sebelum melakukan randomisasi, pengguna perlu memeriksa fitur mana saja yang ingin dirandomisasi karena bisa saja pada dataset tersebut ada kolom yang berupa label. Perangkat lunak akan menyimpan hasil randomisasi pada dokumen dataset pengguna yang telah diubah nilainya dan perangkat lunak akan menyimpan dokumen tersebut di suatu lokasi yang pengguna harus tentukan. Pengguna bisa menentukan lokasi dokumen 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.

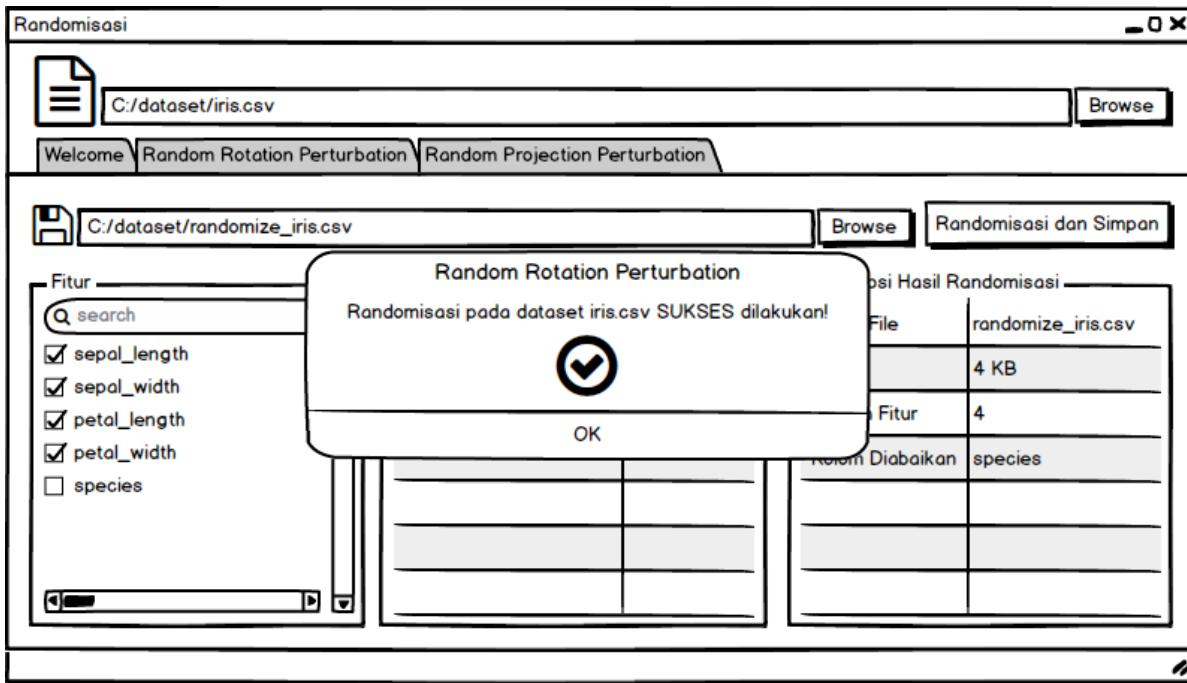
Perangkat lunak akan menampilkan deskripsi hasil dari randomisasi yang dilakukan. Informasi



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 dokumen, ukuran dokumen, 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.

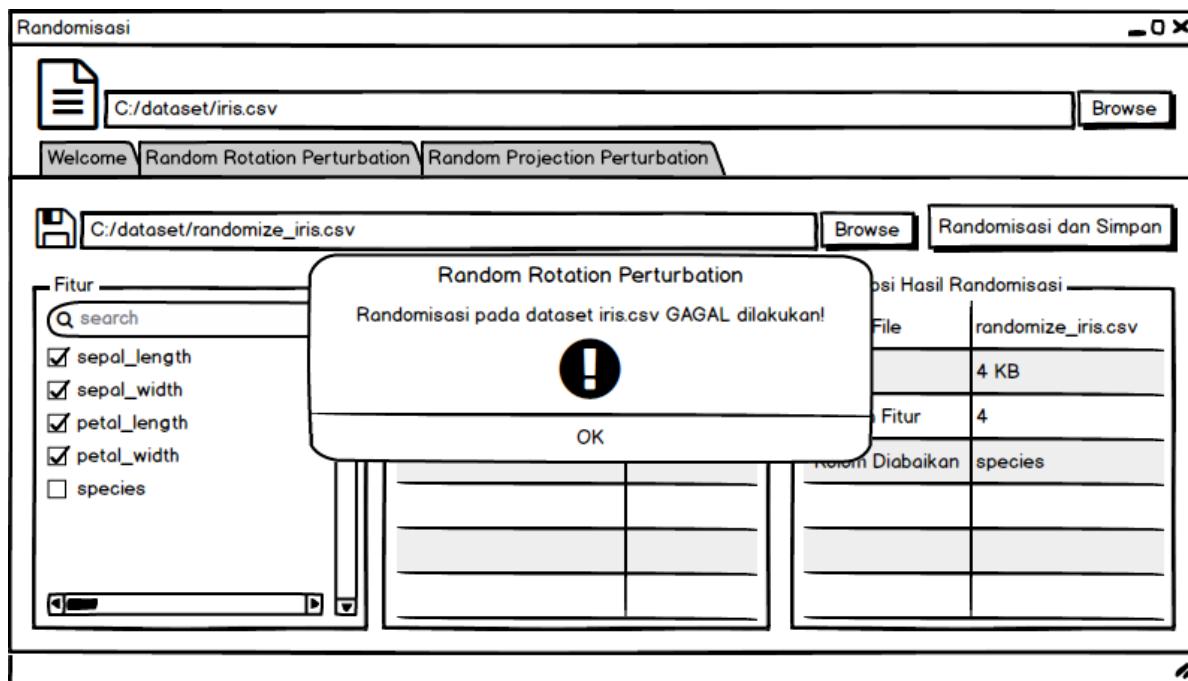
Apabila ada masalah-masalah tertentu dan randomisasi gagal dilakukan, maka perangkat lunak akan memberitahukan bahwa randomisasi telah gagal dilakukan dengan menampilkan *popup* yang berisi peringatan bahwa randomisasi gagal dilakukan. Hasil randomisasi tidak akan terbuat dan perangkat lunak tidak akan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

4.1.2 Halaman *Random Projection Perturbation*

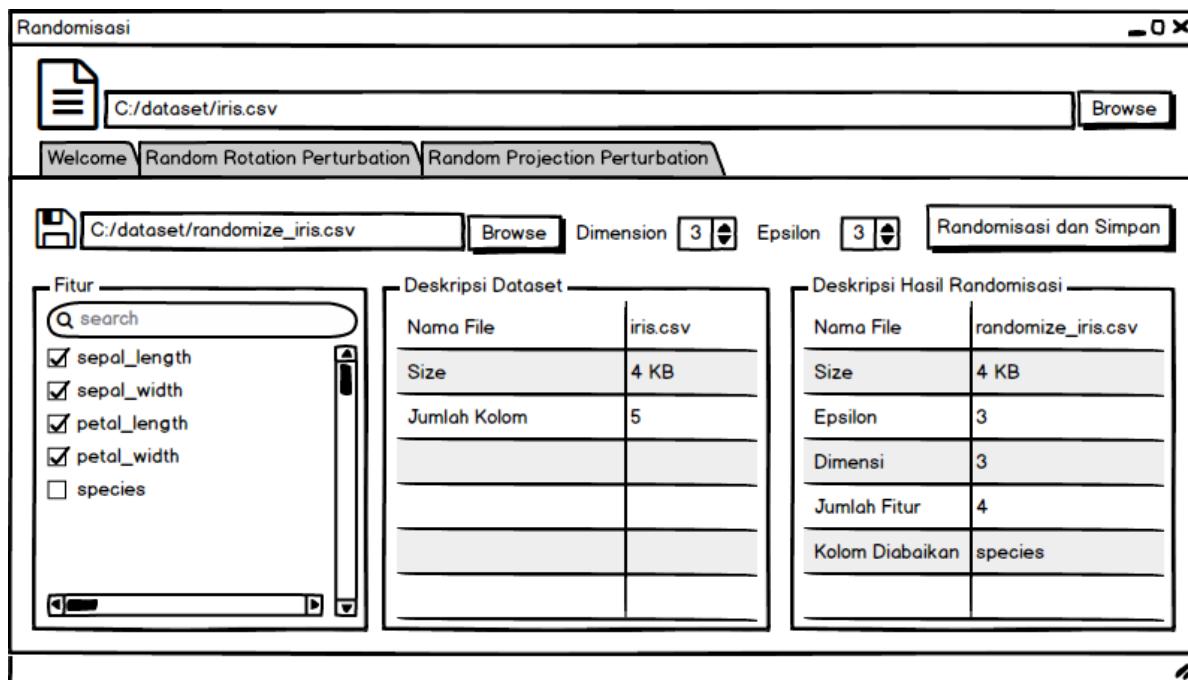
Halaman ini memiliki fungsi untuk melakukan teknik *Random Projection Perturbation*. Rancangan dari halaman ini dapat dilihat pada Gambar 4.5. Sebelum pengguna melakukan randomisasi, pengguna harus mengatur beberapa pengaturan yang ada. Sama seperti halnya pada halaman *Random Rotation Perturbation*, 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.

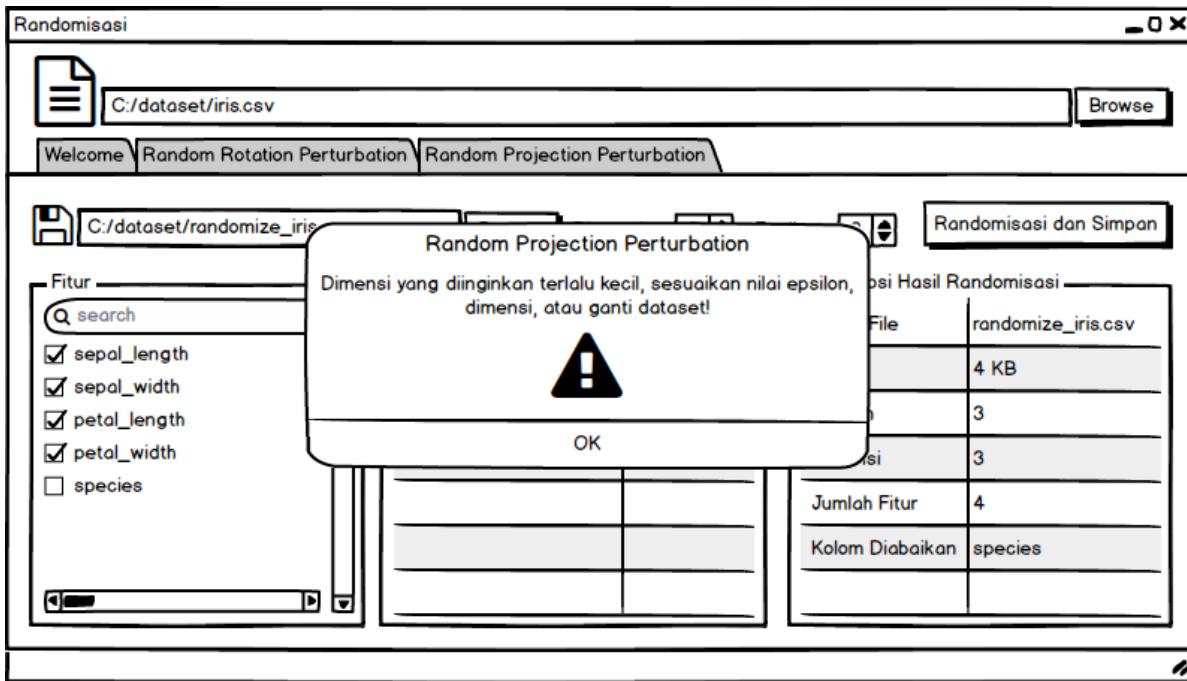
Pada teknik *Random Projection Perturbation*, ada persyaratan yang harus dipenuhi mengenai dimensi akhir yang diinginkan dan nilai Epsilon. Kedua nilai tersebut perlu sesuai dengan dataset yang ada sehingga pengguna tidak bisa sembarang menentukan dimensi dan nilai Epsilon. Perangkat lunak akan membuat aturan mengenai minimal dimensi yang bisa digunakan sehingga pengguna tidak bisa memasukan 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



Gambar 4.5: Halaman untuk melakukan teknik *Random Projection Perturbation*



Gambar 4.6: *Popup* yang akan ditampilkan apabila persyaratan pada dataset tidak terpenuhi

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 dokumen hasil randomisasi, ukuran dokumen, 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 dokumen berjenis *comma-separated values* di lokasi yang telah ditentukan sebelumnya oleh pengguna.

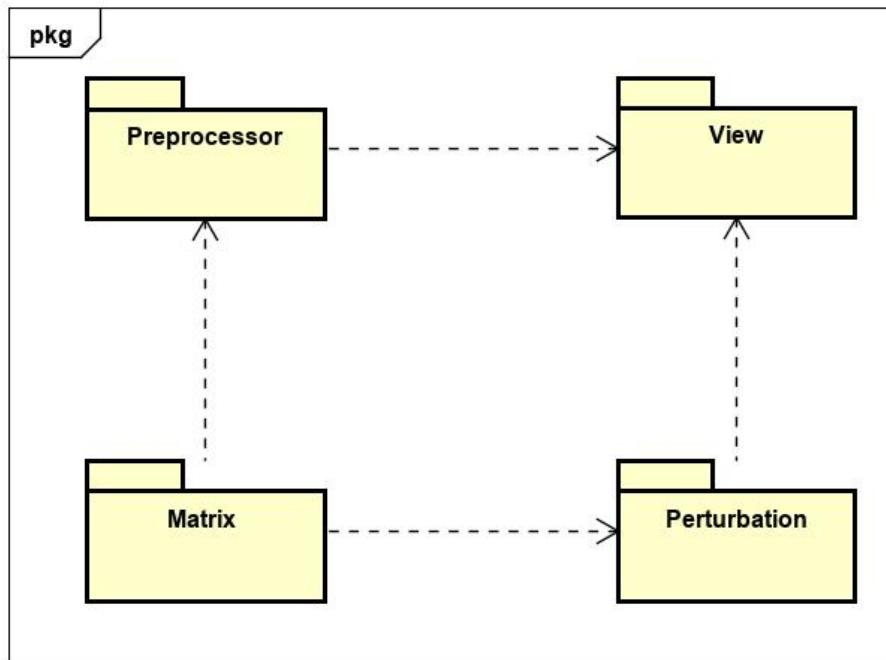
Oleh karena adanya persyaratan yang harus dipenuhi oleh pengguna untuk melakukan randomisasi dengan teknik *Random Projection Perturbation*, maka randomisasi bisa saja gagal dilakukan karena persyaratan yang ada tidak dipenuhi oleh pengguna. Persyaratan yang disebutkan adalah persyaratan jumlah dimensi 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

Dalam pengembangan perangkat lunak, perlu adanya perancangan perangkat lunak secara menyeluruh untuk menghindari kesulitan dan kebingungan pada waktu pengembangan. Perangkat lunak yang akan dibuat pada penelitian ini akan berorientasi objek sehingga perlu adanya perancangan kelas-kelas yang akan dibuat. Pada subbab ini akan dijelaskan perancangan kelas perangkat lunak dan apa saja kegunaannya.

4.2.1 Diagram *Package*

Perangkat lunak akan mempunyai 4 buah *package* yaitu *Perturbation*, *Matrix*, *Preprocessor*, dan *View*. Keempat *package* ini mempunyai fungsinya masing-masing untuk mendukung perangkat lunak berjalan yang akan dijelaskan pada subbab ini. Diagram *Package* perangkat lunak dapat dilihat pada Gambar 4.7



Gambar 4.7: Diagram *package* perangkat lunak

Package Perturbation

Package Perturbation merupakan *package* yang menangani implementasi algoritma dari teknik randomisasi yang dipakai pada penelitian ini yaitu *Random Rotation Perturbation* dan *Random Projection Perturbation*. Kedua algoritma ini masing-masing akan menjadi sebuah kelas terpisah dan memiliki fungsinya masing-masing. Kedua kelas ini akan berada di dalam *package Perturbation*. Satu kelas lagi akan ada di dalam *package* ini yang berperan sebagai kelas *super* bersifat abstrak untuk kedua kelas lainnya.

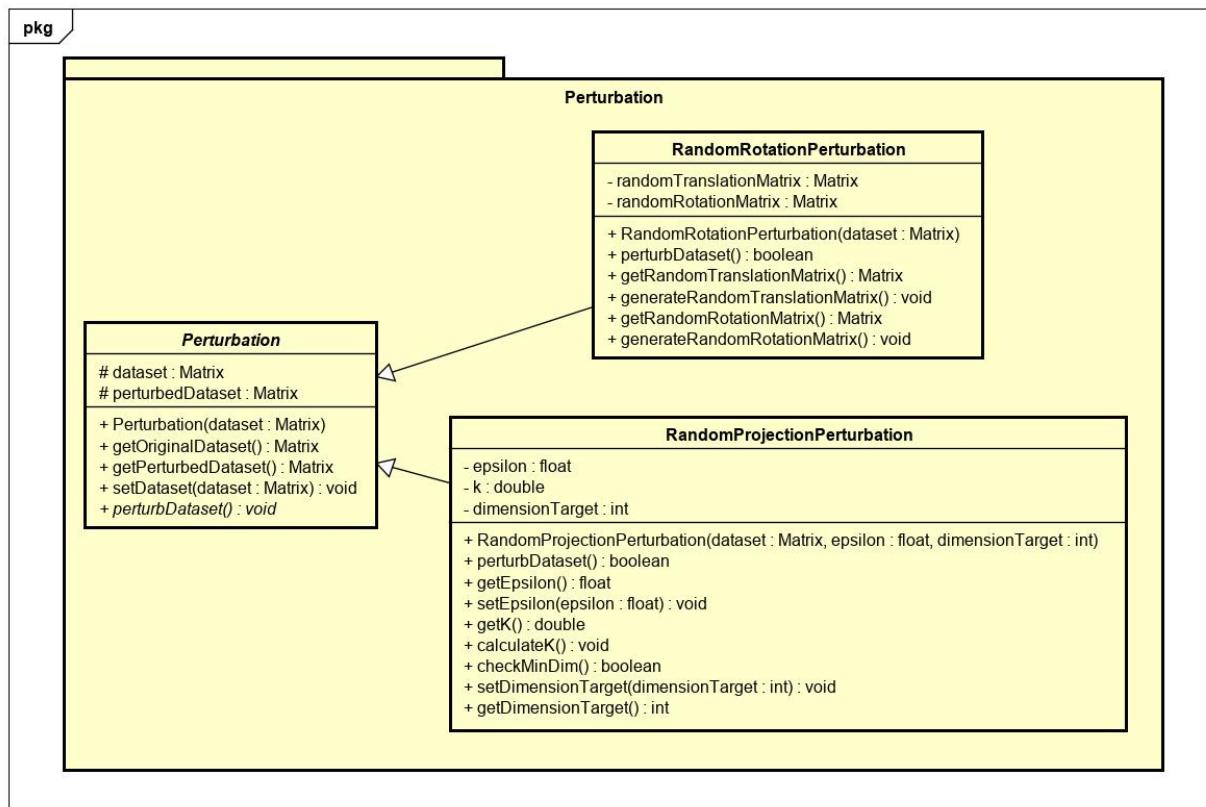
Dalam penerapan algoritma teknik yang dipakai, *package* ini membutuhkan komponen atau fungsi lain untuk membuat algoritma yang diimplementasikan bekerja. Fungsi yang dibutuhkan antara lain adalah membuat matriks dengan sifat tertentu. Oleh karena itu *package Perturbation* membutuhkan *package Matrix* untuk membantu dalam pembuatan matriks yang khusus digunakan pada algoritma. *Package Matrix* akan dijelaskan setelah ini.

Package Matrix

Package Matrix merupakan *package* yang menangani segala jenis fungsi yang berkaitan dengan matriks. Semua fungsi yang terkait tentang matriks diimplementasikan pada *package* ini, fungsi tersebut antara lain adalah implementasi struktur data matriks yang diimplementasikan pada sebuah kelas dan pembuatan matriks khusus yang akan dipakai untuk implementasi algoritma pada *package Perturbation* yang diimplementasikan menjadi tiga buah kelas. *Package* ini juga mengimplementasikan berbagai macam operasi matriks seperti perkalian, transpose matriks, dan menghitung determinan. Operasi-operasi ini diimplementasikan karena adanya kebutuhan operasi-operasi tersebut untuk membantu implementasi dari algoritma pada *package Perturbation*.

Package Preprocessor

Package Preprocessor merupakan *package* yang berfungsi sebagai *preprocessor* untuk masukan perangkat lunak. Tentunya masukan yang diberikan pengguna kepada perangkat lunak tidak bisa langsung diolah begitu saja. Perlu adanya pengolahan terlebih dahulu sebelum masukan yang



Gambar 4.8: Diagram kelas pada *package Perturbation*

diterima dipakai pada perangkat lunak. Oleh karena itu, *package* ini mempunyai fungsi untuk menyelesaikan masalah tersebut yaitu mengolah masukan yang diterima menjadi struktur data yang sesuai untuk digunakan pada perangkat lunak

Pada penelitian ini, masukan perangkat lunak yang dimaksud adalah dataset yang akan dirandomisasi. Pengguna perlu mengikuti persyaratan masukan seperti apa yang dapat menjadi masukan perangkat lunak pada penelitian ini. Pada penelitian ini, perangkat lunak dirancang untuk hanya menerima dokumen berjenis *comma-separated values*. Oleh karena itu, *package* *preprocessor* ini memiliki sebuah fungsi untuk mengolah masukan berupa dokumen berjenis *comma-separated values* menjadi struktur data matriks dengan menggunakan *package* *Matrix*.

Package View

Package View merupakan *package* yang menangani bagian antarmuka pada perangkat lunak di penelitian ini. Antarmuka perangkat lunak akan diimplementasikan menggunakan *framework* antarmuka grafis berbasis bahasa pemrograman Python yang bernama Kivy¹. *Package* ini akan berisi kelas-kelas dan seluruh fungsi yang bertujuan untuk menampilkan antarmuka pada perangkat lunak.

4.2.2 Diagram Kelas pada *Package Perturbation*

Package Perturbation memiliki tiga buah kelas yang bertujuan untuk mengimplementasikan algoritma teknik *Random Rotation Perturbation* dan *Random Projection Perturbation*. Ketiga kelas tersebut adalah *Perturbation*, *RandomRotationPerturbation*, dan *RandomProjectionPerturbation*

¹<https://kivy.org/#home>

yang akan dijelaskan secara detail pada subbab ini. Diagram kelas pada package *Perturbation* dapat dilihat pada Gambar 4.8.

Kelas *Perturbation*

Kelas *Perturbation* berperan sebagai kelas abstrak yang akan menjadi kelas *super* dari kedua kelas lainnya dalam package *Perturbation* yaitu kedua kelas yang mengimplementasikan algoritma teknik *Random Rotation Perturbation* dan *Random Projection Perturbation*. Kelas ini mendeklarasikan atribut dan fungsi apa saja yang seharusnya diimplementasikan oleh kelas *RandomRotationPerturbation*, dan *RandomProjectionPerturbation*. Beberapa atribut dan fungsi tersebut masih kosong pada kelas *Perturbation* sehingga berbagai atribut dan fungsi tersebut perlu didefinisikan pada kelas-kelas turunannya. Selanjutnya akan dijelaskan secara rinci tiap atribut dan fungsi pada kelas ini.

Berikut adalah deskripsi setiap atribut pada kelas *Perturbation*.

- *dataset* adalah atribut untuk menampung dataset yang diambil dari masukan perangkat lunak yang ingin dirandomisasi dan sudah berbentuk matriks. Tipe data atribut ini adalah *Matrix*.
- *perturbedDataset* adalah atribut untuk menampung hasil dari dataset yang telah dirandomisasi. Tipe data atribut ini adalah *Matrix*.

Berikut adalah deskripsi setiap fungsi pada kelas *Perturbation*.

- *Perturbation* adalah *constructor* dari kelas *Perturbation*. Tujuan utama fungsi ini adalah mendefinisikan atribut-atribut yang ada. Fungsi ini memiliki sebuah masukan yang dinamakan *dataset* yang bertipe data *Matrix* dan berfungsi untuk mendefinisikan atribut *dataset*.
- *getOriginalDataset* adalah fungsi untuk mendapatkan dataset asli yang belum dirandomisasi atau dengan kata lain mendapatkan atribut *dataset*. Fungsi ini tidak memiliki masukan apapun dan mempunyai tipe data kembalian berupa *Matrix*.
- *getPerturbedDataset* adalah fungsi untuk mendapatkan hasil dari dataset yang telah dirandomisasi atau dengan kata lain mendapatkan atribut *perturbedDataset*. Fungsi ini tidak memiliki parameter apapun. Tipe data kembalian pada fungsi ini berupa *Matrix*.
- *setDataset* adalah fungsi untuk mendefinisikan ulang atribut *dataset* dengan dataset yang baru. Fungsi ini memiliki sebuah masukan yang dinamakan *dataset* yang bertipe data *Matrix* dan berfungsi untuk mendefinisikan atribut *dataset*. Tidak ada kembalian pada fungsi ini.
- *perturbDataset* adalah fungsi untuk melakukan randomisasi pada atribut *dataset* dan menyimpan hasilnya pada atribut *perturbedDataset*. Fungsi ini bersifat abstrak yang berarti fungsi ini belum didefinisikan sehingga fungsi ini harus didefinisikan pada setiap kelas turunannya. Tidak ada kembalian ataupun masukan pada fungsi ini.

Kelas *RandomRotationPerturbation*

Kelas *RandomRotationPerturbation* adalah kelas yang mempunyai tujuan utama melakukan randomisasi dengan teknik *Random Rotation Perturbation* pada dataset yang menjadi masukan perangkat lunak. Kelas ini merupakan kelas turunan dari kelas *Perturbation* sehingga kelas ini pun mewarisi semua atribut dan fungsi yang dimiliki oleh kelas *Perturbation*. Oleh karena itu, fungsi *perturbDataset* yang diturunkan dari kelas *Perturbation* harus didefinisikan oleh kelas *RandomRotationPerturbation*. Kelas ini akan mengimplementasikan algoritma dari teknik *Random Rotation Perturbation* untuk melakukan randomisasi pada fungsi *perturbDataset*. Selanjutnya akan dijelaskan secara rinci tiap atribut dan fungsi pada kelas ini.

Berikut adalah deskripsi setiap atribut pada kelas *RandomRotationPerturbation*.

- *randomTranslationMatrix* adalah atribut untuk menampung matriks translasi yang akan dipergunakan untuk implementasi algoritma *Random Rotation Perturbation*. Fungsi *generateRandomTranslationMatrix* akan mendefinisikan atribut ini dengan membuat matriks translasi acak, yang akan dijelaskan kemudian. Atribut ini mempunyai tipe data *Matrix*.
- *randomRotationMatrix* adalah atribut untuk menampung matriks rotasi yang akan dipergunakan untuk implementasi algoritma *Random Rotation Perturbation*. Fungsi *generateRandomRotationMatrix* akan mendefinisikan atribut ini dengan membuat matriks translasi acak, yang akan dijelaskan kemudian. Atribut ini mempunyai tipe data *Matrix*.

Berikut adalah deskripsi setiap fungsi pada kelas *RandomRotationPerturbation*.

- *RandomRotationPerturbation* adalah *constructor* untuk mendefinisikan atribut-atribut yang ada dan memanggil *constructor* dari kelas *super* milik kelas *RandomRotationPerturbation* yaitu *Perturbation*. Tujuan utama fungsi ini adalah memanggil *constructor* dari kelas *super* yaitu *Perturbation* sehingga fungsi ini berguna untuk mendefinisikan atribut *dataset*. Selain itu, fungsi ini juga akan mendefinisikan atribut *randomTranslationMatrix* dan *randomRotationMatrix*.
- *perturbDataset* adalah fungsi turunan dari kelas *Perturbation* yang pada kelas *RandomRotationPerturbation* diimplementasikan algoritma teknik *Random Rotation Perturbation*. Fungsi ini akan mengimplementasikan algoritma teknik *Random Rotation Perturbation* kepada atribut *dataset* dan menyimpan hasilnya kepada atribut *perturbedDataset*. Tidak ada masukan pada fungsi ini tetapi memiliki kembalian berupa *boolean* yang menyatakan apakah randomisasi berhasil dilakukan.
- *getRandomTranslationMatrix* adalah fungsi untuk mendapatkan matriks translasi yang digunakan untuk implementasi algoritma teknik *Random Rotation Perturbation* atau dengan kata lain mendapatkan atribut *randomTranslationMatrix*. Fungsi ini tidak memiliki masukan apapun tetapi mempunyai kembalian berupa matriks translasi yang bertipe data *Matrix*.
- *generateRandomTranslationMatrix* adalah fungsi untuk membuat matriks translasi secara acak yang akan digunakan untuk implementasi algoritma teknik *Random Rotation Perturbation*. Fungsi ini akan membuat matriks translasi acak tersebut dan menyimpan hasilnya pada atribut *randomTranslationMatrix*. Tidak ada masukan ataupun kembalian pada fungsi ini.
- *getRandomRotationMatrix* adalah fungsi untuk mendapatkan matriks rotasi yang digunakan untuk implementasi algoritma teknik *Random Rotation Perturbation* atau dengan kata lain mendapatkan atribut *randomRotationMatrix*. Fungsi ini tidak memiliki masukan apapun tetapi mempunyai kembalian berupa matriks rotasi yang bertipe data *Matrix*.
- *generateRandomRotationMatrix* adalah fungsi untuk membuat matriks rotasi secara acak yang akan digunakan untuk implementasi algoritma teknik *Random Rotation Perturbation*. Fungsi ini akan membuat matriks rotasi acak tersebut dan menyimpan hasilnya pada atribut *randomRotationMatrix*. Tidak ada masukan ataupun kembalian pada fungsi ini.

Kelas *RandomProjectionPerturbation*

Kelas *RandomProjectionPerturbation* adalah kelas yang mempunyai tujuan utama melakukan randomisasi dengan teknik *Random Projection Perturbation* pada dataset yang menjadi masukan perangkat lunak. Kelas ini merupakan kelas turunan dari kelas *Perturbation* sehingga kelas ini pun mewarisi semua atribut dan fungsi yang dimiliki oleh kelas *Perturbation*. Oleh karena itu, fungsi *perturbDataset* yang diturunkan dari kelas *Perturbation* harus didefinisikan oleh kelas *RandomProjectionPerturbation*. Kelas ini akan mengimplementasikan algoritma dari teknik *Random*

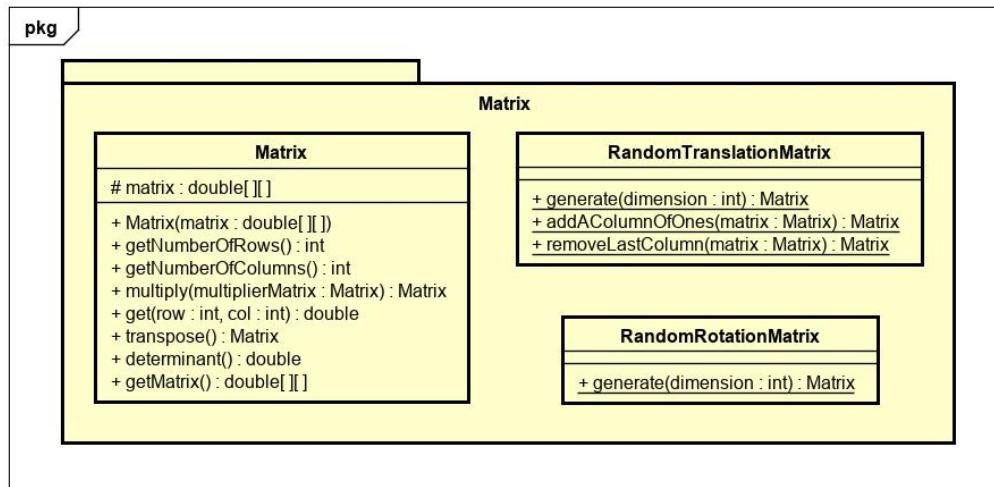
Projection Perturbation untuk melakukan randomisasi pada fungsi *perturbDataset*. Selanjutnya akan dijelaskan secara rinci tiap atribut dan fungsi pada kelas ini.

Berikut adalah deskripsi setiap atribut pada kelas *RandomProjectionPerturbation*.

- *epsilon* adalah atribut yang berguna untuk menentukan seberapa besar batas maksimal distorsi yang dapat terjadi pada hasil randomisasi. Atribut *epsilon* ini akan digunakan untuk menghitung nilai dari atribut *k* yang akan dijelaskan nanti. Tipe data atribut ini adalah *float*, bilangan riil yang nantinya hanya akan diisi dengan rentang nilai (0, 1). Pengguna akan menentukan seberapa besar batas maksimal distorsi yang dapat terjadi pada hasil randomisasi dengan cara mendefinisikan atribut *epsilon* ini.
- *k* adalah atribut yang berguna untuk menentukan dimensi terkecil untuk sebuah dataset yang ingin dirandomisasi dapat direduksi dengan distorsi yang terkontrol sesuai atribut *epsilon*. Atribut ini juga menentukan apakah dataset memenuhi persyaratan untuk direduksi yaitu memiliki jumlah kolom yang lebih besar dari nilai *k*. Tipe data atribut ini adalah *double*, bilangan riil. Atribut ini akan dihitung oleh perangkat lunak sesuai nilai *epsilon* yang telah ditentukan pengguna.
- *dimensionTarget* adalah atribut untuk menyimpan besar dimensi akhir yang diinginkan pengguna dimiliki oleh hasil dataset yang telah dirandomisasi. Pengguna harus mendefinisikan atribut ini dengan nilai yang lebih besar dari atau sama dengan nilai *k*, jika tidak maka distorsi yang terjadi pada hasil randomisasi akan lebih besar dari nilai *epsilon* yang diinginkan. Tipe data atribut ini adalah *integer*, bilangan bulat.

Berikut adalah deskripsi setiap fungsi pada kelas *RandomProjectionPerturbation*.

- *RandomProjectionPerturbation* adalah *constructor* untuk mendefinisikan atribut-atribut yang ada dan memanggil *constructor* dari kelas *super* milik kelas *RandomProjectionPerturbation* yaitu *Perturbation*. Tujuan utama fungsi ini adalah memanggil *constructor* dari kelas *super* yaitu *Perturbation* sehingga fungsi ini berguna untuk mendefinisikan atribut *dataset*. Selain itu, fungsi ini juga akan mendefinisikan atribut *epsilon*, *k*, dan *dimensionTarget*.
- *perturbDataset* adalah fungsi turunan dari kelas *Perturbation* yang pada kelas *RandomProjectionPerturbation* diimplementasikan algoritma teknik *Random Projection Perturbation*. Fungsi ini akan mengimplementasikan algoritma teknik *Random Projection Perturbation* kepada atribut *dataset* dan menyimpan hasilnya kepada atribut *perturbedDataset*. Tidak ada masukan pada fungsi ini tetapi memiliki kembalian berupa *boolean* yang menyatakan apakah randomisasi berhasil dilakukan.
- *getEpsilon* adalah fungsi untuk mendapatkan nilai batas maksimal distorsi yang dapat terjadi pada hasil randomisasi atau dengan kata lain mendapatkan atribut *epsilon*. Fungsi ini tidak memiliki masukan apapun tetapi mempunyai kembalian berupa nilai dari atribut *epsilon* yang bertipe data *float*, bilangan riil.
- *setEpsilon* adalah fungsi untuk mengubah nilai dari atribut *epsilon* dengan nilai baru yang menjadi masukan fungsi ini. Fungsi *setEpsilon* memiliki sebuah masukan *epsilon* yang bertipe data *float*, bilangan riil dan berfungsi untuk mendefinisikan atribut *epsilon*. Parameter ini akan menjadi nilai baru dari atribut *epsilon*. Tidak ada kembalian pada fungsi ini.
- *getK* adalah fungsi untuk mendapatkan besar dimensi minimal hasil dataset yang telah dirandomisasi atau dengan kata lain mendapatkan atribut *k*. Fungsi ini tidak memiliki masukan apapun tetapi mempunyai kembalian berupa nilai dari atribut *k* yang bertipe data *double*, bilangan riil.

Gambar 4.9: Diagram kelas pada *package Matrix*

- *calculateK* adalah fungsi untuk menghitung nilai k sesuai dengan atribut *epsilon* dan jumlah baris pada dataset yang ingin dirandomisasi. Fungsi ini akan menghitung nilai k tersebut dan menyimpan hasilnya pada atribut k . Tidak ada masukan ataupun kembalian pada fungsi ini.
- *checkMinDim* adalah fungsi untuk memeriksa apakah dimensi dari *dataset* yang ingin dirandomisasi lebih besar dari nilai k , besar dimensi minimal. Selain itu, fungsi ini memeriksa apakah nilai *dimensionTarget* lebih besar dari atau sama dengan nilai k . Intinya fungsi ini menguji apakah dataset dan keinginan pengguna memenuhi persyaratan untuk mengaplikasikan teknik *Random Projection Perturbation*. Fungsi ini tidak memiliki masukan apapun tetapi mempunyai kembalian berupa sebuah *boolean* yang menandakan apakah *dataset* dan *dimensionTarget* yang diberikan oleh pengguna memenuhi persyaratan untuk mengaplikasikan teknik *Random Projection Perturbation* dengan *dataset* dan *dimensionTarget* tersebut.
- *setDimensionTarget* adalah fungsi untuk mengubah nilai dari atribut *dimensionTarget* dengan nilai baru yang menjadi masukan fungsi ini. Fungsi *setDimensionTarget* memiliki sebuah masukan *dimensionTarget* yang bertipe data *integer*, bilangan bulat dan berfungsi untuk mendefinisikan atribut *dimensionTarget*. Parameter ini akan menjadi nilai baru dari atribut *dimensionTarget*. Tidak ada kembalian pada fungsi ini.
- *getDimensionTarget* adalah fungsi untuk mendapatkan besar dimensi akhir yang diinginkan pengguna dimiliki oleh hasil dataset yang telah dirandomisasi atau dengan kata lain mendapatkan atribut *dimensionTarget*. Fungsi ini tidak memiliki masukan apapun tetapi mempunyai kembalian berupa nilai dari atribut *dimensionTarget* yang bertipe data *integer*, bilangan bulat.

4.2.3 Diagram Kelas pada *Package Matrix*

Package Matrix memiliki empat buah kelas yang bertujuan untuk menangani segala jenis fungsi yang berkaitan dengan matriks maupun pembuatan matriks yang khusus digunakan untuk implementasi algoritma. Keempat kelas tersebut adalah *Matrix*, *RandomTranslationMatrix*, dan *RandomRotationMatrix* yang akan dijelaskan secara detail pada subbab ini. Diagram kelas pada *package Matrix* dapat dilihat pada Gambar 4.9.

Kelas *Matrix*

Kelas *Matrix* adalah kelas yang menangani struktur data matriks serta segala fungsi atau operasi yang berkaitan dengan matriks. Kelas ini akan mempunyai semua kebutuhan terkait dengan urusan

struktur data matriks yang ada untuk implementasi algoritma *Random Rotation Perturbation* dan *Random Projection Perturbation*, antara lain seperti perkalian, transpose matriks, dan mencari determinan. Selanjutnya akan dijelaskan secara rinci tiap atribut dan fungsi pada kelas ini.

Berikut adalah deskripsi setiap atribut pada kelas *Matrix*.

- *matrix* adalah atribut untuk menyimpan matriks dengan cara menyimpannya memakai *array* 2 dimensi dengan tipe data *double*, bilangan riil.

Berikut adalah deskripsi setiap fungsi pada kelas *Matrix*.

- *Matrix* adalah *constructor* dari kelas *PerturbMatrixation*. Tujuan utama fungsi ini adalah mendefinisikan atribut-atribut yang ada. Fungsi ini memiliki sebuah masukan berbentuk array yang dinamakan *matrix* yang bertipe data *double* dan berfungsi untuk mendefinisikan atribut *matrix*.
- *getNumberOfRows* adalah fungsi untuk mendapatkan jumlah baris yang ada pada matriks kelas ini atau dengan kata lain mendapatkan ukuran dari atribut *matrix*. Fungsi ini tidak memiliki masukan apapun tetapi mempunyai kembalian berupa jumlah baris yang bertipe data *integer*, bilangan bulat.
- *getNumberOfColumns* adalah fungsi untuk mendapatkan jumlah kolom yang ada pada matriks kelas ini atau dengan kata lain mendapatkan ukuran dari atribut *matrix*. Fungsi ini tidak memiliki masukan apapun tetapi mempunyai kembalian berupa jumlah baris yang bertipe data *integer*, bilangan bulat.
- *multiply* adalah fungsi yang berguna untuk mengkalikan matriks yang ada pada atribut *matrix* dengan suatu matriks lain. Tentu saja matriks lain tersebut harus sudah berbentuk objek kelas *Matrix* sehingga mempunyai tipe data yang sama dan dapat dikalikan. Fungsi ini mempunyai sebuah parameter yaitu *multiplierMatrix* bertipe data objek kelas *Matrix* yang berguna sebagai pengali matriks yang ingin dikalikan. Kembalian pada fungsi ini adalah hasil kali antara kedua matriks dan kembalian ini bertipe data objek kelas *Matrix*.
- *get* adalah fungsi untuk mendapatkan sebuah elemen pada matriks atau dengan kata lain sebuah nilai pada atribut *matrix*. Fungsi ini memiliki dua buah masukan yaitu *row* dan *col* yang masing-masing berguna sebagai penunjuk baris dan kolom mana yang ingin didapatkan. Kembalian pada fungsi ini adalah elemen matriks pada baris dan kolom yang diinginkan dan bertipe data *double*, bilangan riil.
- *transpose* adalah fungsi untuk mendapatkan transpose dari matriks kelas ini atau dengan kata lain transpose dari atribut *matrix*. Fungsi ini memiliki kembalian berupa matriks yang merupakan hasil transpose dan bertipe data objek kelas *Matrix*. Tidak ada masukan apapun pada fungsi ini.
- *determinant* adalah fungsi untuk mendapatkan determinan dari matriks kelas ini. Fungsi ini memiliki kembalian berupa determinan matriks yang bertipe data *double*, bilangan riil. Tidak ada masukan apapun pada fungsi ini.
- *getMatrix* adalah fungsi untuk mendapatkan matriks pada kelas ini atau dengan kata lain mendapatkan atribut *matrix*. Fungsi ini tidak memiliki masukan apapun tetapi mempunyai kembalian berupa *array* yang bertipe data *double*, bilangan riil.

Kelas *RandomTranslationMatrix*

Kelas *RandomTranslationMatrix* adalah kelas statis sehingga tidak bisa diinstansiasi dan hanya memiliki atribut atau fungsi statis saja. Tujuan utama dari kelas ini adalah menangani pembuatan

matriks translasi yang akan digunakan untuk melakukan operasi translasi yang akan diimplementasikan di kelas *RandomRotationPerturbation*. Kelas ini tidak memiliki atribut apapun tetapi memiliki tiga buah fungsi statis yang memiliki fungsi seputar pembuatan matriks translasi dan fungsi yang mendukung operasi translasi. Selanjutnya akan dijelaskan secara rinci tiap fungsi pada kelas ini.

Berikut adalah deskripsi setiap fungsi pada kelas *RandomTranslationMatrix*.

- *generate* adalah fungsi statis yang berguna untuk membuat matriks translasi. Fungsi ini memiliki sebuah parameter yaitu *dimension* yang akan menentukan ukuran atau dimensi matriks translasi yang ingin dibuat. Kembalian pada fungsi ini tentu saja sebuah matriks translasi yang bertipe data objek kelas *Matrix*.
- *addAColumnOfOnes* adalah fungsi statis yang berguna untuk menambahkan sebuah kolom pada suatu matriks di posisi terakhir (setelah kolom terakhir). Kolom tersebut akan berisi angka satu pada setiap barisnya. Alasan dibuatnya fungsi ini dikarenakan kebutuhan persyaratan yang harus dipenuhi sebuah matriks apabila ingin diterapkan operasi translasi. Fungsi ini memiliki sebuah parameter yaitu matriks yang diinginkan bernama *matrix* dan bertipe data objek kelas *Matrix*. Kembalian pada fungsi ini adalah matriks yang telah ditambahkan sebuah kolom dan bertipe data objek kelas *Matrix*.
- *removeLastColumn* adalah fungsi statis yang berguna untuk menghapus kolom terakhir pada suatu matriks. Alasan dibuatnya fungsi ini dikarenakan kebutuhan persyaratan yang harus dipenuhi sebuah matriks apabila ingin diterapkan operasi translasi. Setelah matriks ditambahkan dengan menggunakan fungsi *addAColumnOfOnes* dan diterapkan operasi translasi, kolom terakhir pada matriks tersebut perlu dibuang dikarenakan kolom tersebut adalah kolom hasil penambahan yang hanya digunakan untuk operasi translasi dan bukan kolom asli matriks tersebut.

Kelas *RandomRotationMatrix*

Kelas *RandomRotationMatrix* adalah kelas statis sehingga tidak bisa diinstansiasi dan hanya memiliki atribut atau fungsi statis saja. Tujuan utama dari kelas ini adalah menangani pembuatan matriks rotasi yang akan digunakan untuk melakukan operasi rotasi yang akan diimplementasikan di kelas *RandomRotationPerturbation*. Kelas ini tidak memiliki atribut apapun tetapi memiliki sebuah fungsi statis yang memiliki fungsi untuk pembuatan matriks rotasi. Selanjutnya akan dijelaskan secara rinci tiap fungsi pada kelas ini.

Berikut adalah deskripsi setiap fungsi pada kelas *RandomRotationMatrix*.

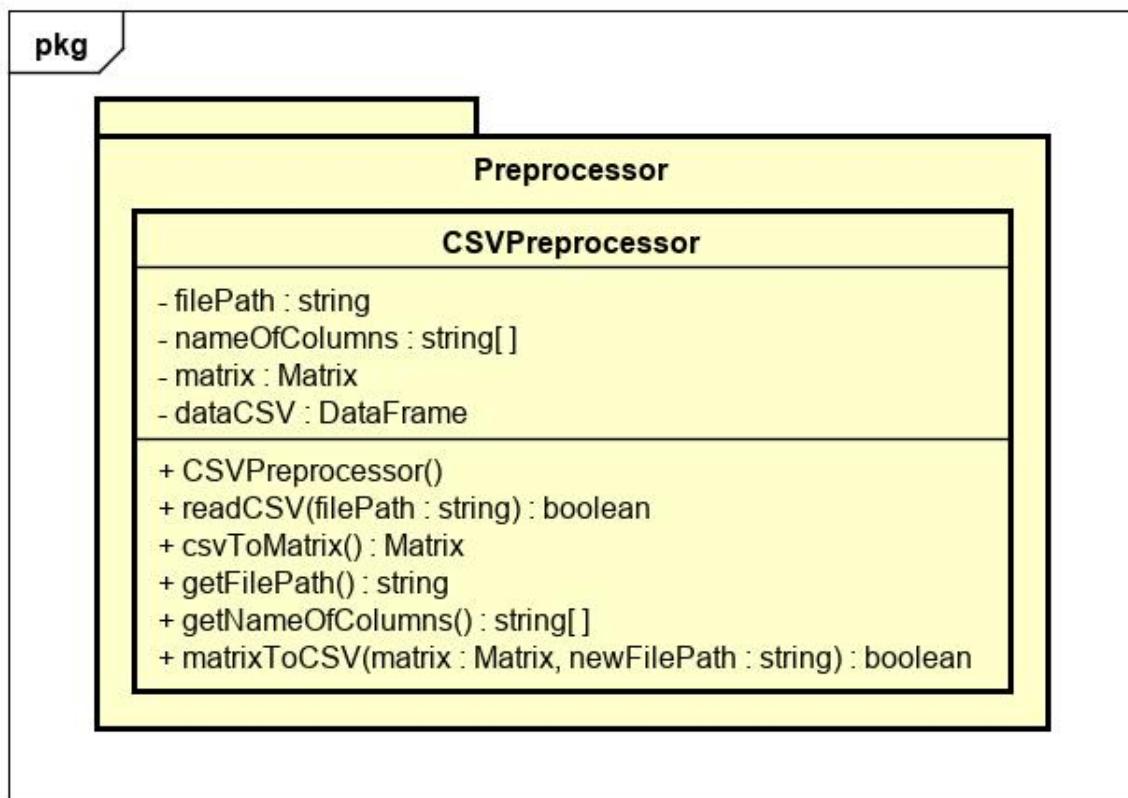
- *generate* adalah fungsi statis yang berguna untuk membuat matriks rotasi. Pembuatan matriks rotasi dilakukan dengan mengikuti distribusi Haar [8]. Fungsi ini memiliki sebuah parameter yaitu *dimension* yang akan menentukan ukuran atau dimensi matriks rotasi yang ingin dibuat. Kembalian pada fungsi ini tentu saja sebuah matriks rotasi yang bertipe data objek kelas *Matrix*.

4.2.4 Diagram Kelas pada *Package Preprocessor*

Package Preprocessor memiliki sebuah kelas yang bertujuan untuk mengolah masukan perangkat lunak agar dapat diterapkan teknik randomisasi. Kelas tersebut bernama *CSVPreprocessor* yang akan dijelaskan secara detail pada subbab ini. Diagram kelas pada *package Preprocessor* dapat dilihat pada Gambar 4.10.

Kelas *CSVPreprocessor*

Kelas *CSVPreprocessor* berguna untuk mengolah masukan perangkat lunak yang berjenis *comma-separated values* sebelum diterapkan teknik randomisasi agar masukan tersebut sesuai dengan



Gambar 4.10: Diagram kelas pada *package Preprocessor*

persyaratan teknik randomisasi. Tujuan utama dari kelas ini adalah mengubah masukan berjenis *comma-separated values* menjadi sebuah objek kelas *Matrix* dan sebaliknya. Kelas ini memiliki tiga buah atribut dan enam buah fungsi yang selanjutnya akan dijelaskan secara rinci.

Berikut adalah deskripsi setiap atribut pada kelas *CSVPreprocessor*.

- *filePath* adalah atribut untuk menyimpan lokasi masukan yang berjenis *comma-separated values* yang ingin diolah pada komputer pengguna. Atribut ini memiliki tipe data *string*.
- *nameOfColumns* adalah atribut untuk menyimpan nama seluruh kolom pada dataset yang telah diolah. Atribut ini berupa array yang bertipe data *string*.
- *matrix* adalah atribut untuk menyimpan objek kelas *Matrix* yang didapatkan dari hasil pengolahan masukan yang berjenis *comma-separated values*. Atribut ini memiliki tipe data objek kelas *Matrix*.
- *dataCSV* adalah atribut untuk menyimpan data dokumen *comma-separated values* yang telah dibaca dalam bentuk *DataFrame*². Atribut ini memiliki tipe data objek kelas *DataFrame*.

Berikut adalah deskripsi setiap fungsi pada kelas *CSVPreprocessor*.

- *CSVPreprocessor* adalah *constructor* kelas ini yang berfungsi untuk membuat sebuah objek kelas *CSVPreprocessor* dan menginisialisasi semua atribut pada kelas ini. Fungsi ini tidak memiliki masukan apapun.
- *readCSV* adalah fungsi untuk membaca dokumen berjenis *comma-separated values* yang ingin diolah oleh kelas ini. Tujuan utama dari fungsi ini adalah menyimpan lokasi dokumen pada

²<https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.html>

atribut *filePath* dan menyimpan data dokumen *comma-separated values* yang telah dibaca dalam bentuk *DataFrame* pada atribut *dataCSV*. Fungsi ini memiliki sebuah masukan bernama *filePath* yang berguna untuk menentukan lokasi dokumen yang ingin diolah. Kembalian pada fungsi ini adalah sebuah *boolean* yang menyatakan apakah fungsi ini berhasil membaca dokumen yang menjadi masukan fungsi *readCSV*.

- *csvToMatrix* adalah fungsi untuk mengolah dokumen *comma-separated values* yang telah dibaca menjadi sebuah objek kelas *Matrix*. Fungsi ini tidak memiliki masukan apapun tetapi mempunyai kembalian berupa objek kelas *Matrix* yang didapatkan dari hasil pengolahan dokumen *comma-separated values*.
- *getFilePath* adalah fungsi untuk mendapatkan lokasi dokumen *comma-separated values* yang telah dibaca atau dengan kata lain mendapatkan atribut *filePath*. Fungsi ini tidak memiliki masukan apapun tetapi mempunyai kembalian berupa lokasi dokumen yang bertipe data *string*.
- *getNameOfColumns* adalah fungsi untuk mendapatkan nama semua kolom yang ada pada dokumen *comma-separated values* yang telah dibaca atau dengan kata lain mendapatkan atribut *nameOfColumns*. Fungsi ini tidak memiliki masukan apapun tetapi mempunyai kembalian berupa *array* yang bertipe data *string*.
- *matrixToCSV* adalah fungsi untuk mengolah sebuah objek kelas *Matrix* menjadi sebuah dokumen *comma-separated values* atau dengan kata lain mengkonversi sebuah matriks menjadi dokumen berjenis *comma-separated values*. Fungsi ini bisa dikatakan kebalikan dari fungsi *csvToMatrix*. Ada dua buah masukan pada fungsi ini yaitu sebuah matriks dan lokasi penyimpanan dokumen yang baru. Kembalian pada fungsi ini adalah sebuah *boolean* yang menyatakan apakah konversi berhasil dilakukan.

4.3 Masukan Perangkat Lunak

Perangkat lunak akan mempunyai sebuah masukan yang berupa dataset. Dataset yang ingin dirandomisasi memiliki persyaratan yaitu harus berupa sebuah matriks. Tetapi mayoritas dataset yang ada didistribusikan bukan sebagai matriks melainkan dokumen berjenis tertentu antara lain seperti dokumen berjenis *comma-separated values*. Oleh karena itu, perangkat lunak ini hanya dapat menerima masukan berupa dokumen berjenis *comma-separated values*. Berikut akan dijelaskan secara rinci masukan yang dapat diterima oleh perangkat lunak.

Dokumen *comma-separated values* adalah dokumen berisi teks yang menggunakan koma untuk memisahkan antara tiap nilai dengan nilai lainnya. Berikut adalah contoh isi dari dokumen berjenis *comma-separated values*.

```
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
4.9,3.1,1.5,0.1,setosa
```

Baris pertama pada contoh di atas merupakan nama semua kolom yang ada pada dataset. Baris lainnya adalah nilai-nilai setiap kolom pada suatu rekord. Pada dataset di atas, kolom terakhir

adalah kolom label. Selain kolom tersebut adalah kolom fitur yang akan dirandomisasi. Kolom fitur tersebut untuk perangkat lunak randomisasi ini mempunyai persyaratan yaitu nilai kolom tersebut harus berjenis numerik.

BAB 5

IMPLEMENTASI DAN PENGUJIAN

Pada bab ini akan ditunjukkan tampilan dari implementasi perangkat lunak dan juga bagaimana perangkat lunak diimplementasikan. Pengujian juga akan diterapkan pada perangkat lunak secara fungsional dan eksperimental. Hasil dari pengujian akan dijelaskan secara rinci dan sistematis serta akan dibuat kesimpulan untuk pengujian yang telah dilakukan.

5.1 Implementasi Antarmuka

Antarmuka perangkat lunak diimplementasikan dengan memakai *framework* antarmuka grafis berbasis bahasa pemrograman Python yang bernama Kivy¹. Implementasi antarmuka disesuaikan dengan rancangan antarmuka perangkat lunak yang telah dibuat pada bab 4. Gambar 5.1 adalah tampilan antarmuka dari implementasi perangkat lunak.

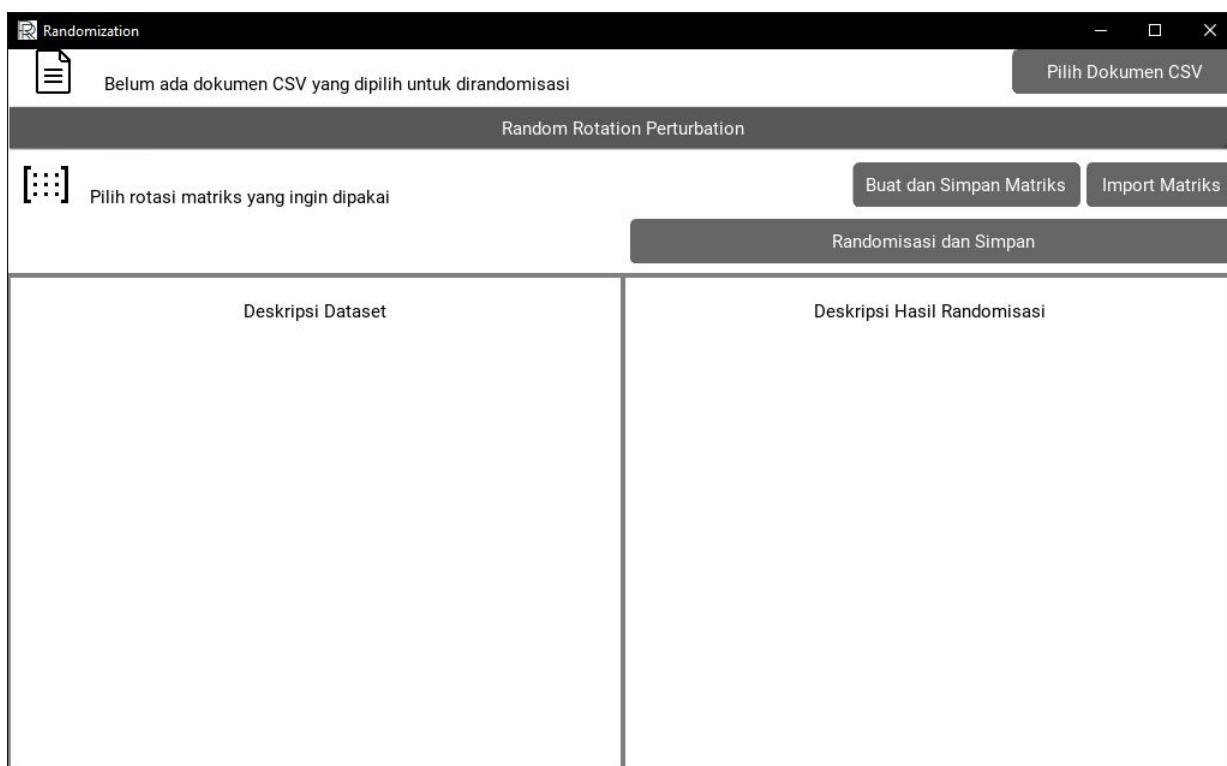
Antarmuka perangkat lunak mempunyai tiga buah bagian yang mempunyai fungsinya masing-masing. Ketiga bagian ini dapat dilihat pada Gambar 5.2 Pertama adalah bagian masukan dan pengaturan, terdapat pada bagian atas yang bernomor satu dan dikelilingi kotak merah. Kedua adalah bagian deskripsi dataset, terdapat pada bagian bawah sebelah kiri yang bernomor dua dan dikelilingi kotak biru. Terakhir adalah bagian deskripsi hasil randomisasi, terdapat pada bagian bawah sebelah kanan yang bernomor tiga dan dikelilingi kotak hijau. Ketiga bagian ini akan dijelaskan secara rinci pada subbab-subbab berikutnya.

Perangkat lunak randomisasi ini mengimplementasikan dua buah teknik randomisasi yang berbeda yaitu *Random Rotation Perturbation* dan *Random Projection Perturbation*. Oleh karena itu, antarmuka perangkat lunak akan menyesuaikan dengan teknik yang dipilih oleh pengguna. Ketiga bagian antarmuka yang telah disebutkan tadi dengan otomatis akan berubah sesuai dengan teknik yang dipilih. Pada setiap subbab akan dijelaskan juga sekaligus perbedaan antarmuka teknik randomisasi satu dengan yang lainnya.

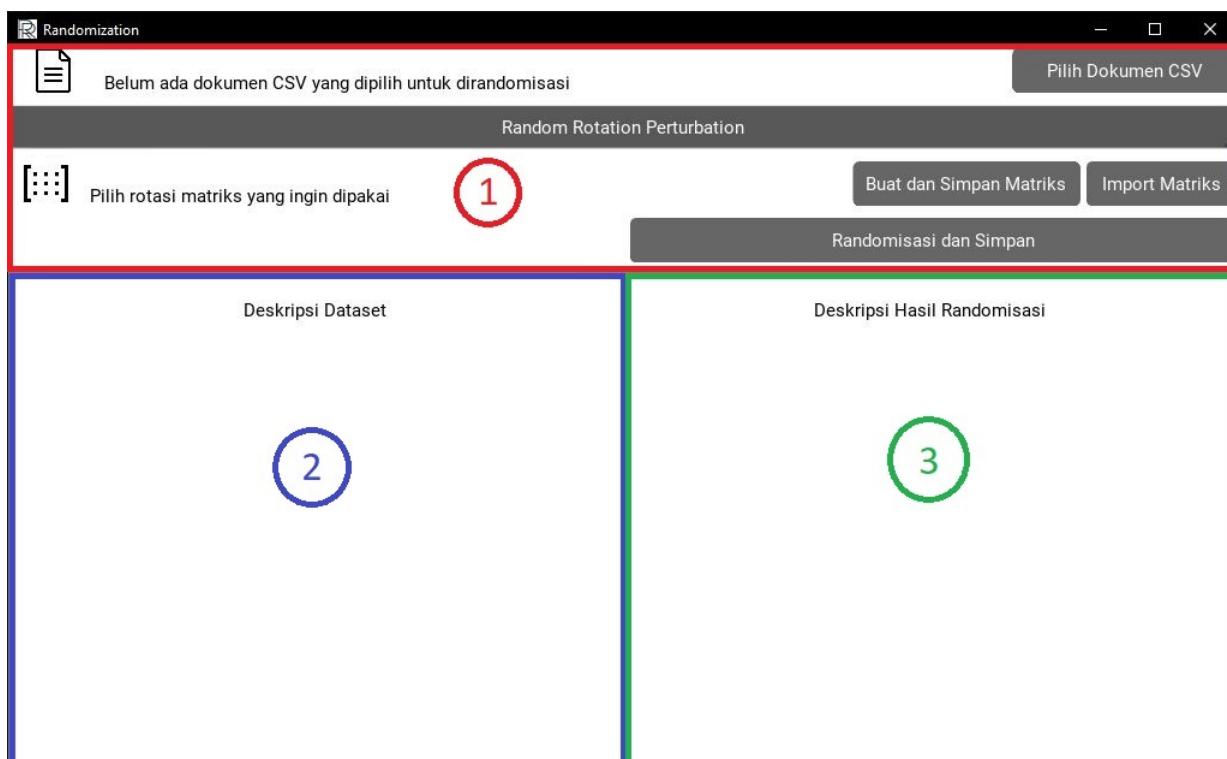
5.1.1 Masukan dan Pengaturan

Bagian masukan dan pengaturan menyediakan berbagai interaksi untuk pengguna dapat mengatur masukan yang perlu diberikan kepada perangkat lunak dan menerapkan teknik randomisasi yang diinginkan. Ada beberapa fungsi inti pada bagian ini yaitu masukan dataset berupa file *comma-separated values* yang ingin dirandomisasi, memilih teknik randomisasi yang ingin digunakan, membuat baru dan memilih matriks rotasi atau proyeksi yang ingin digunakan, masukan nilai variabel Epsilon dan nilai variabel K untuk teknik *Random Projection Perturbation*, dan sebuah tombol untuk menerapkan teknik randomisasi dan menyimpan hasilnya. Berikut akan dijelaskan secara rinci dengan gambar setiap fungsi tersebut yang dapat dilihat pada Gambar 5.3 dan cara pemakaianya yang benar secara berturut.

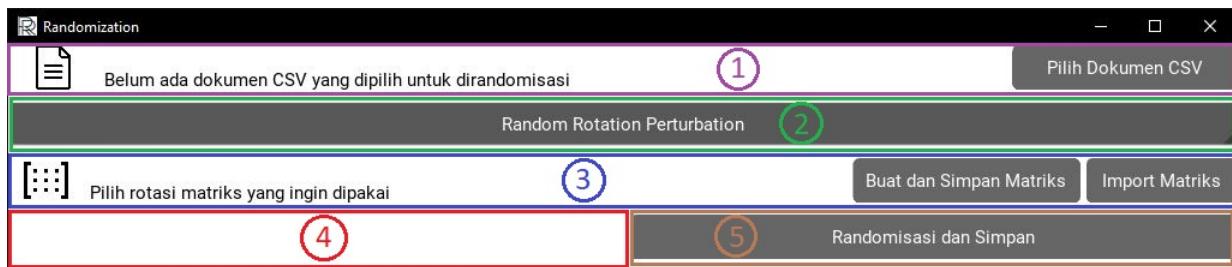
¹<https://kivy.org/#home>



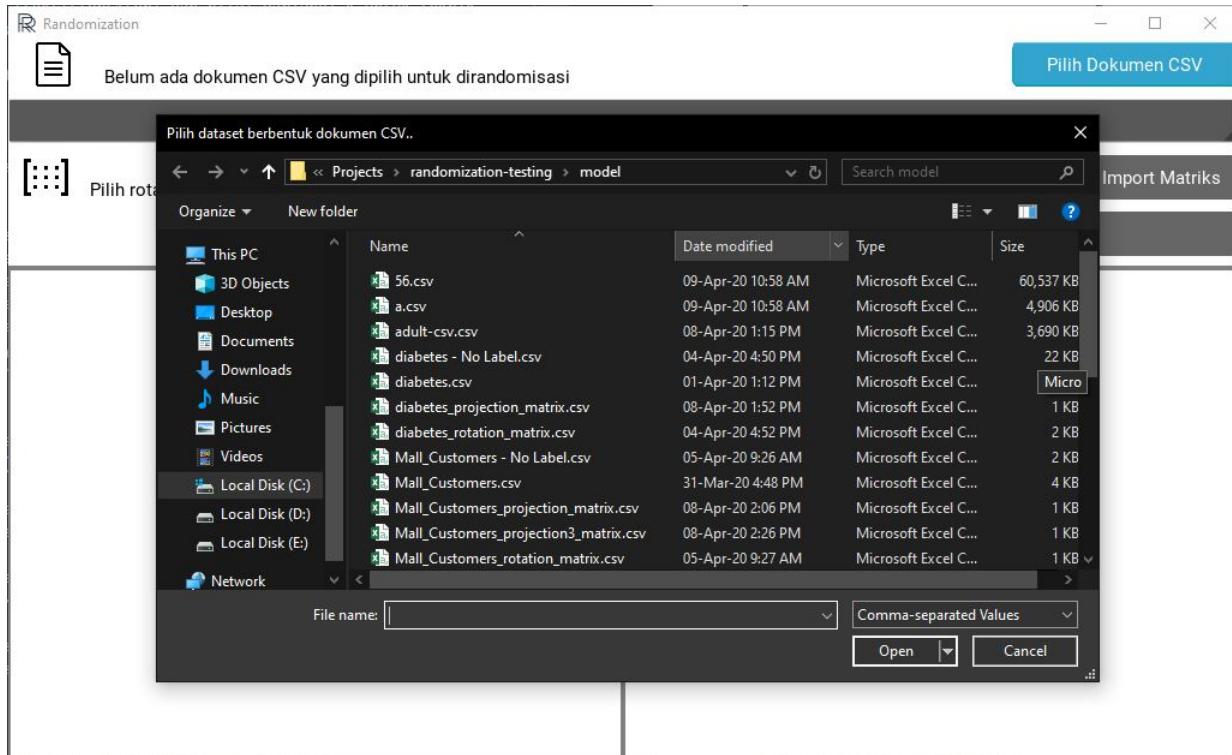
Gambar 5.1: Tampilan perangkat lunak yang pertama ditampilkan saat perangkat lunak baru dibuka



Gambar 5.2: Bagian-bagian pada antarmuka perangkat lunak



Gambar 5.3: Bagian antarmuka masukan dan pengaturan perangkat lunak

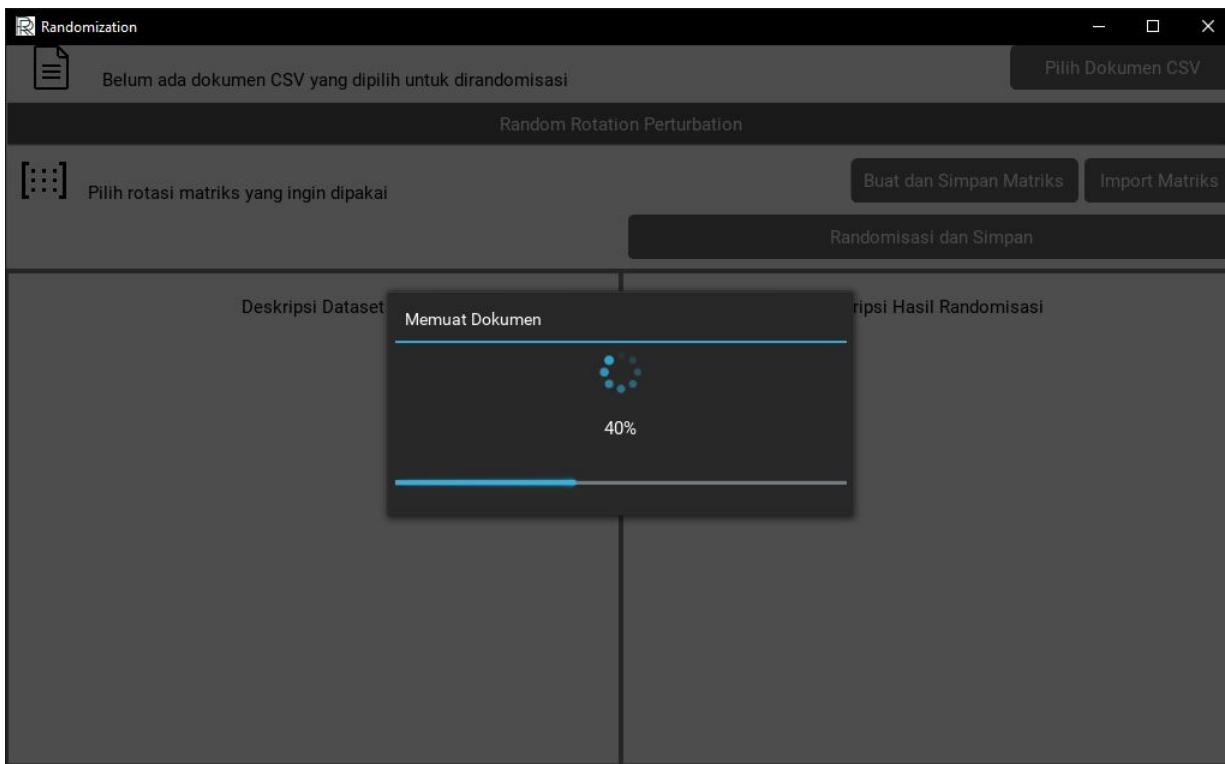


Gambar 5.4: Jendela untuk memilih dataset yang berupa dokumen CSV

Masukan Dataset

Pertama pengguna perlu memberikan masukan dataset yang ingin dirandomisasi berupa dokumen berjenis *comma-separated values*. Perangkat lunak menyediakan fitur tersebut yang dapat dilihat pada Gambar 5.3 yang terdapat pada bagian yang dikelilingi kotak berwarna merah dan bernomor satu. Pengguna dapat menekan tombol “Pilih Dokumen CSV” yang terletak pada ujung sebelah kanan. Tombol ini bertujuan untuk memilih dokumen yang ingin dirandomisasi pada direktori pengguna. Ketika tombol ditekan, perangkat lunak akan membuka jendela baru untuk memilih dokumen yang dapat dilihat pada gambar 5.4.

Setelah pengguna memilih dataset yang diinginkan, perangkat lunak akan otomatis menuliskan lokasi dokumen yang dipilih berada. Perangkat lunak akan menampilkan lokasi dokumen tersebut pada bagian tengah sebelah kanan simbol dokumen dan sebelah kiri tombol “Pilih Dokumen CSV”. Jika belum ada dataset yang dipilih maka perangkat lunak akan menampilkan label yang berupa kalimat “Belum ada dokumen CSV yang dipilih untuk dirandomisasi” yang menunjukkan bahwa belum ada dokumen yang dipilih oleh pengguna. Jika pengguna memilih ulang dokumen, maka secara otomatis juga perangkat lunak akan memperbaharui lokasi dokumen sesuai dokumen yang dipilih pengguna.



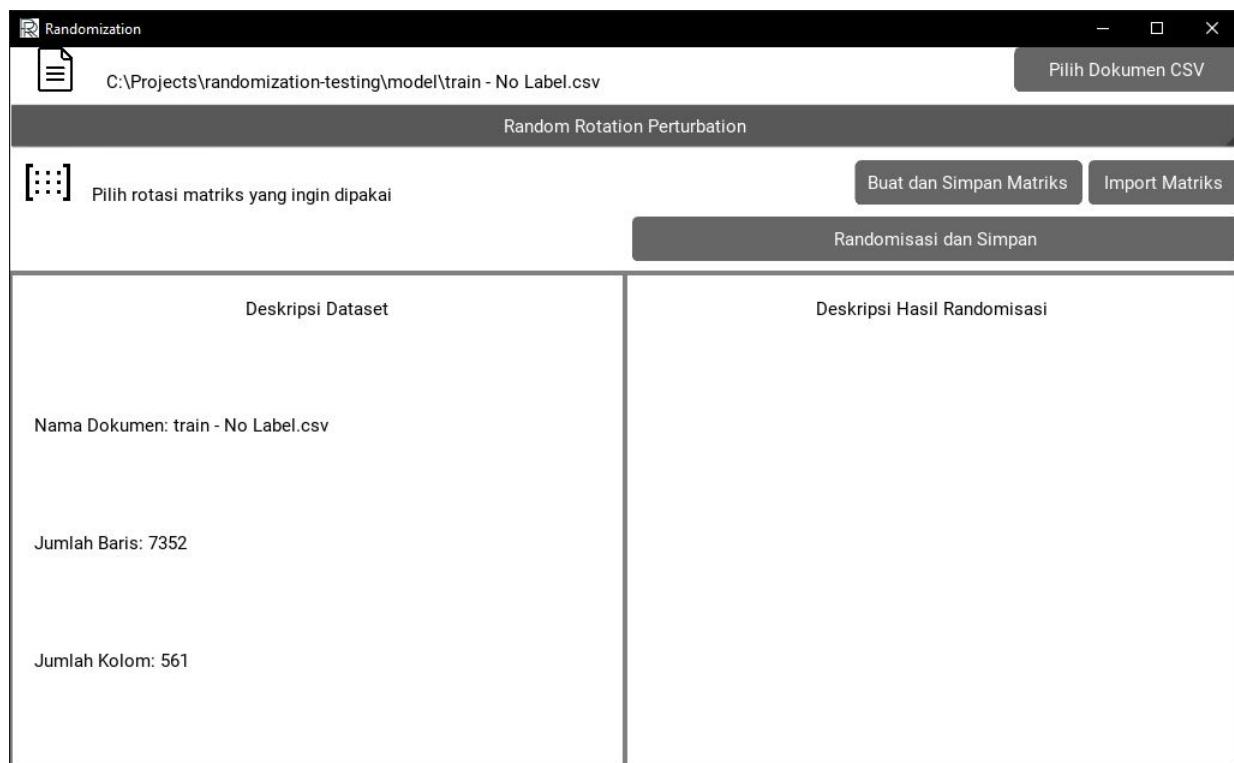
Gambar 5.5: Tampilan *popup* yang ditampilkan saat proses berlangsung

Apabila dokumen yang dipilih berukuran besar, maka perangkat lunak akan memakan sedikit waktu yang lebih lama. Dalam rangka memberitahukan kepada pengguna bahwa perangkat lunak sedang melakukan proses pemilihan dokumen, perangkat lunak akan menampilkan sebuah *popup* yang memberitahukan bahwa proses pemilihan sedang berjalan dan perangkat lunak tidak berhenti bekerja maupun *error* sehingga pengguna tidak bingung apabila perangkat lunak memakan waktu yang lebih lama untuk memproses dokumen yang dipilih. Tampilan antarmuka *popup* tersebut dapat dilihat pada Gambar 5.5. Setelah dokumen dipilih pengguna dan perangkat lunak berhasil memproses dokumen tersebut, perangkat lunak akan memperbarui lokasi dokumen dan menampilkan beberapa informasi dataset yang dipilih pada bagian deskripsi dataset yang akan dijelaskan pada subbab berikutnya. Tampilan antarmuka setelah pengguna memilih dokumen dapat dilihat pada Gambar 5.6

Selain itu setelah pengguna memilih dokumen CSV, perangkat lunak akan membaca dokumen tersebut dan memproses isi dari dokumen tersebut menjadi dataset yang berupa matriks. Proses ini dilakukan sekali saja tepat setelah pengguna memilih dokumen dengan menekan tombol ‘‘Pilih Dokumen CSV’’. Oleh karena itu, apabila sebuah dokumen CSV diubah isinya setelah dokumen tersebut dipilih oleh pengguna maka perangkat lunak tetap akan menggunakan isi dari dokumen tersebut yang belum diubah. Pengguna harus berhati-hati apabila isi dokumen diubah maka pengguna juga harus memilih kembali dokumen yang sama tersebut walaupun perangkat lunak sudah menunjukkan lokasi dokumen yang digunakan adalah dokumen yang pengguna inginkan.

Pemilihan Teknik Randomisasi

Setelah pengguna memilih dataset yang ingin dirandomisasi, pengguna juga harus memilih teknik randomisasi apa yang ingin diterapkan terhadap dataset yang sudah dipilih. Pada awal perangkat lunak dibuka, secara otomatis teknik *Random Rotation Perturbation* yang dipilih. Apabila pengguna ingin mengganti teknik yang ingin diterapkan pada dataset, pengguna dapat menekan tombol *dropdown* yang bertuliskan nama teknik randomisasi. Tombol ini dapat dilihat pada Gambar 5.3



Gambar 5.6: Tampilan antarmuka setelah sebuah dokumen dipilih

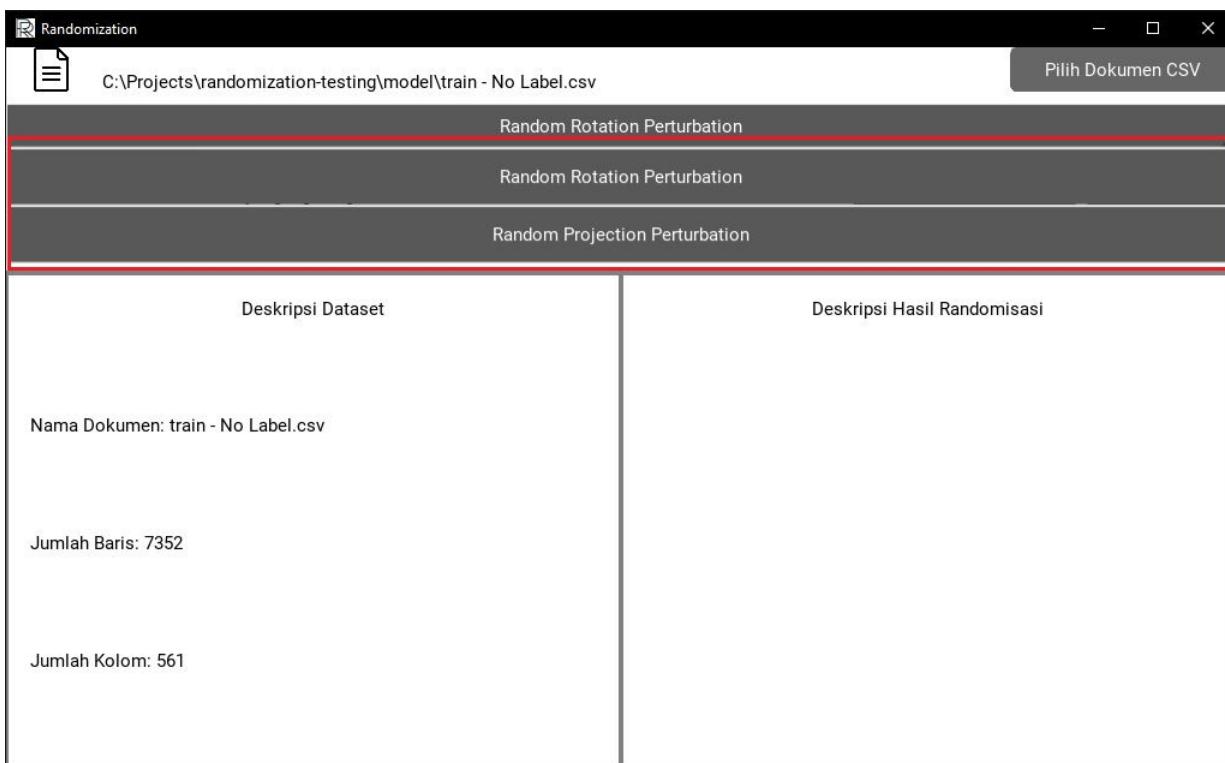
yang dikelilingi kotak berwarna hijau dan bernomor dua.

Apabila pengguna menekan tombol ini maka perangkat lunak akan menampilkan *dropdown* yang mempunyai dua buah opsi teknik randomisasi yaitu “Random Rotation Perturbation” dan “Random Projection Perturbation”. Antarmuka tersebut dapat dilihat pada Gambar 5.7 yang dikelilingi oleh kotak merah. Pemilihan teknik ini juga akan memicu beberapa perubahan pada tampilan antarmuka perangkat lunak menyesuaikan dengan teknik yang dipilih. Beberapa perubahan pada perangkat lunak tersebut melengkapi bagian pembuatan dan pemilihan matriks, parameter teknik randomisasi, dan bagian randomisasi dan simpan yang akan dijelaskan setiap perubahan tersebut pada subbab berikutnya.

Pembuatan dan Pemilihan Matriks

Setelah pengguna memilih teknik yang ingin diterapkan, pengguna harus memilih matriks yang diinginkan atau membuat baru. Matriks yang dimaksudkan adalah matriks rotasi atau matriks proyeksi sesuai teknik randomisasi yang dipilih. Apabila teknik *Random Rotation Perturbation* yang dipilih maka perangkat lunak akan mengubah fungsi pembuatan dan pemilihan matriks ini menjadi matriks rotasi. Apabila teknik *Random Projection Perturbation* yang dipilih maka perangkat lunak akan mengubah fungsi pembuatan dan pemilihan matriks ini menjadi matriks proyeksi. Perubahan ini dapat terlihat pada label yang berada di sebelah kanan simbol matriks apabila belum memilih atau membuat matriks maka label tersebut akan menampilkan kalimat “Pilih matriks rotasi yang ingin digunakan” atau “Pilih matriks proyeksi yang ingin digunakan”. Bagian ini dapat dilihat pada Gambar 5.3 yang dikelilingi oleh kotak berwarna hijau dan bernomor dua.

Ada dua buah tombol pada bagian ini yaitu “Buat dan Simpan Matriks” dan “Import Matriks”. Tombol “Buat dan Simpan Matriks” mempunyai fungsi untuk membuat matriks rotasi atau proyeksi baru sesuai teknik randomisasi yang dipilih dan menyimpan matriks tersebut pada sebuah dokumen CSV baru yang dibuat oleh perangkat lunak pada direktori tertentu yang akan dipilih oleh pengguna. Pada saat perangkat lunak sedang memproses matriks tersebut, perangkat lunak akan menampilkan



Gambar 5.7: Tampilan antarmuka saat pengguna memilih teknik

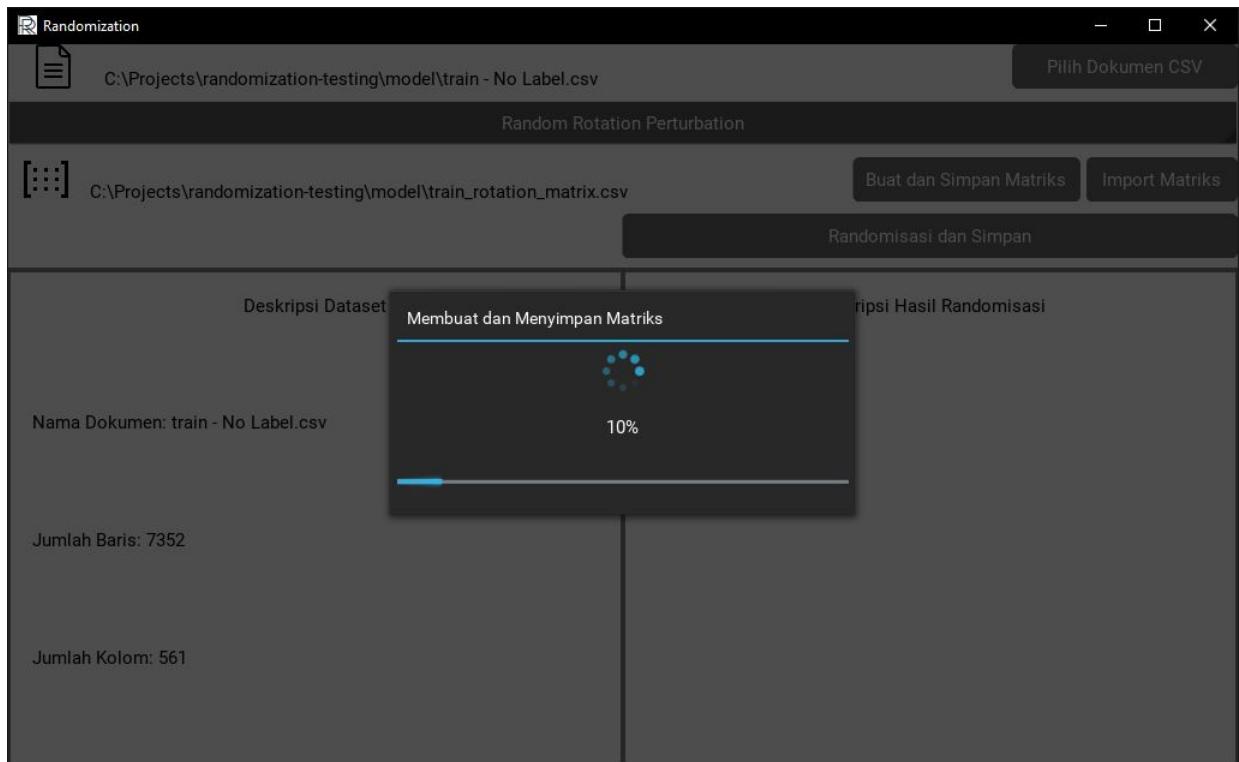
popup memuat yang dapat dilihat pada Gambar 5.8. *Popup* ini juga akan tampil saat proses impor matriks dilakukan. Hasil matriks yang dibuat oleh perangkat lunak dapat digunakan kembali untuk lain kali sehingga rotasi atau proyeksi yang diterapkan akan sama dengan yang sebelumnya.

Pengguna dapat melakukan impor matriks dengan cara menekan tombol “Import Matriks” untuk memilih matriks rotasi atau proyeksi yang diinginkan untuk diterapkan pada dataset. Matriks yang dipilih harus sesuai dengan dataset yang ingin dirandomisasi, misalnya apabila matriks rotasi yang dipilih memiliki dimensi yang berbeda dengan dataset maka perangkat lunak akan melarang impor matriks dilakukan karena randomisasi tidak dapat dilakukan. Perangkat lunak akan menampilkan *popup* peringatan untuk pengguna yang dapat dilihat pada Gambar 5.9. Apabila pengguna memilih teknik *Random Projection Perturbation* dan pengguna mengimpor matriks proyeksi maka parameter variabel K akan terisi secara otomatis sesuai dengan matriks proyeksi yang diimporkan.

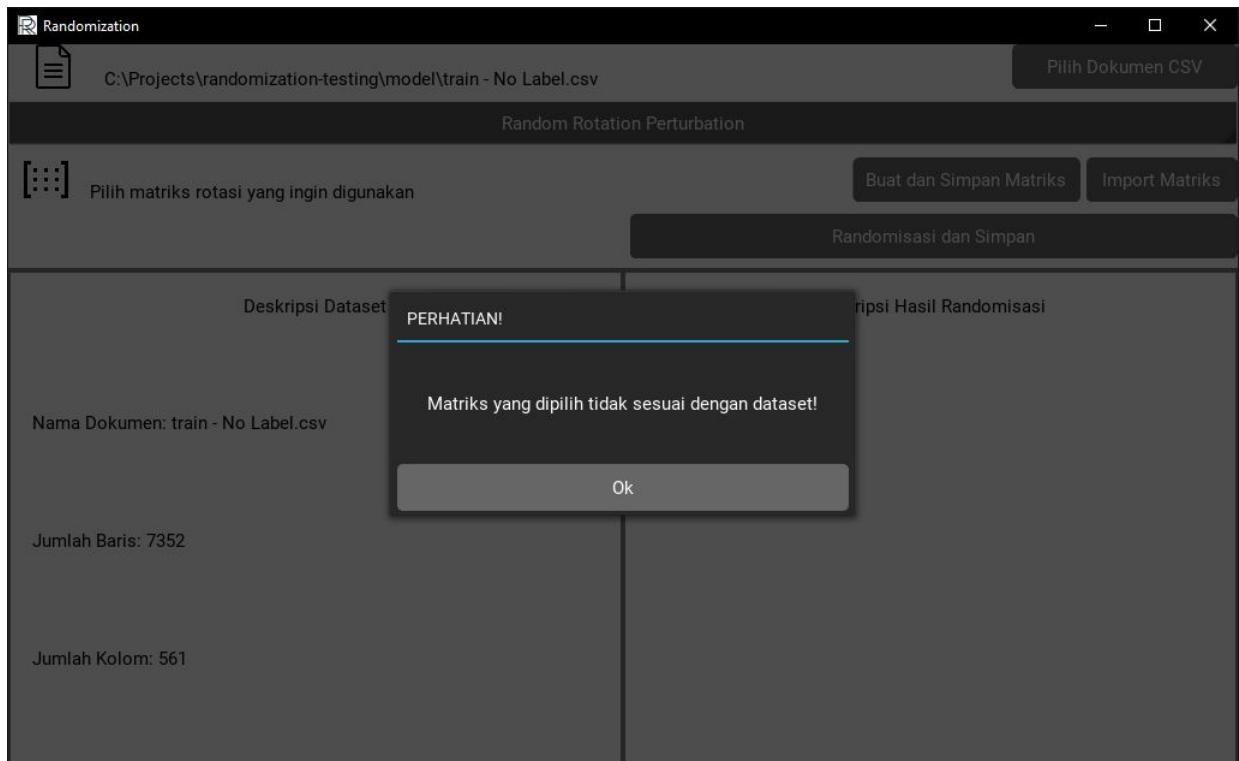
Apabila pengguna belum memilih dataset yang ingin dirandomisasi, pengguna tidak dapat membuat maupun impor matriks terlebih dahulu. Hal ini dikarenakan perlu ada proses pengecekan terlebih dahulu yang dilakukan perangkat lunak untuk memastikan dataset yang ingin dirandomisasi sudah sesuai persyaratan dan sesuai dengan matriks yang akan dipilih. Perangkat lunak akan melarang pengguna membuat maupun impor matriks dengan menampilkan sebuah *popup* peringatan yang dapat dilihat pada Gambar 5.10. Pada teknik *Random Projection Perturbation*, pengguna baru bisa membuat matriks proyeksi apabila sudah memenuhi persyaratan yang diminta yaitu mengisi parameter teknik tersebut yang mana adalah variabel Epsilon dan variabel K.

Parameter Teknik Randomisasi

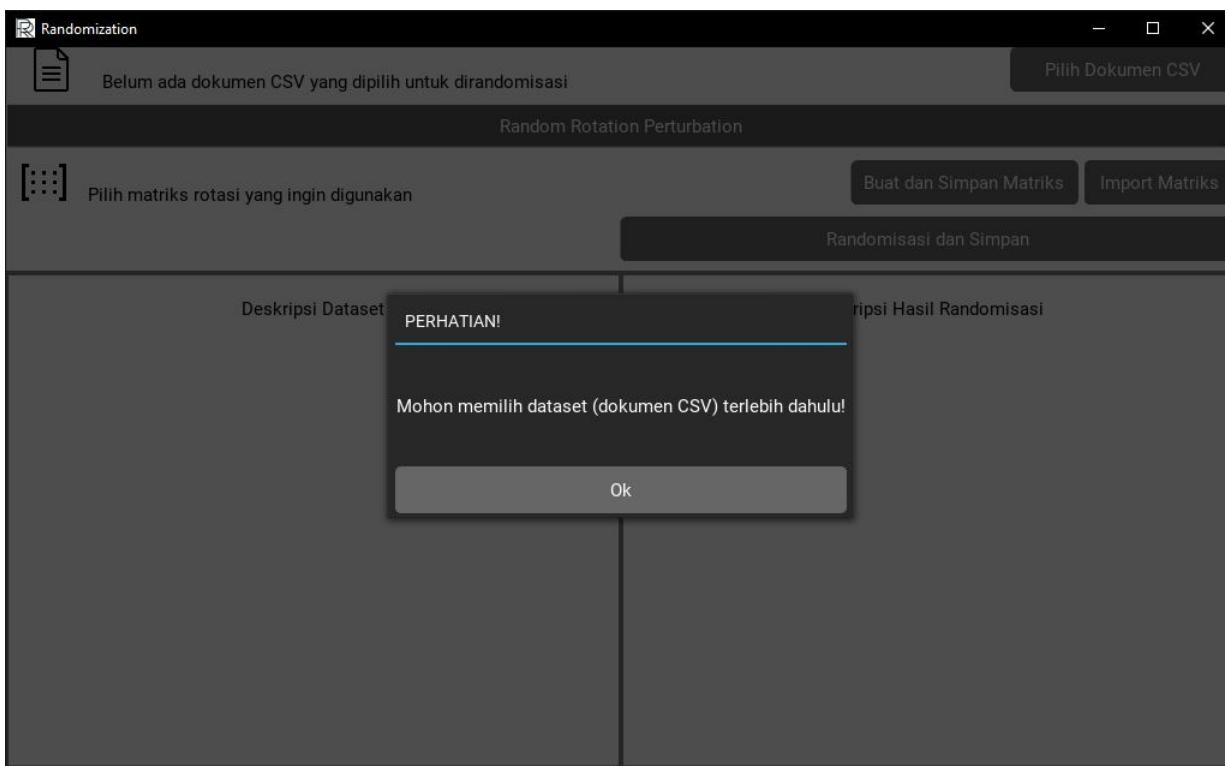
Perangkat lunak hanya meminta kepada pengguna parameter untuk teknik *Random Projection Perturbation* saja apabila pengguna memilih teknik tersebut. Pada teknik *Random Rotation Perturbation* tidak ada parameter yang perlu pengguna berikan. Ada dua buah parameter yang perlu pengguna berikan yaitu variabel Epsilon dan variabel K. Pengguna dapat memasukkan nilai kedua variabel tersebut dengan menekan kolom variabel tersebut masing-masing. Kedua buah



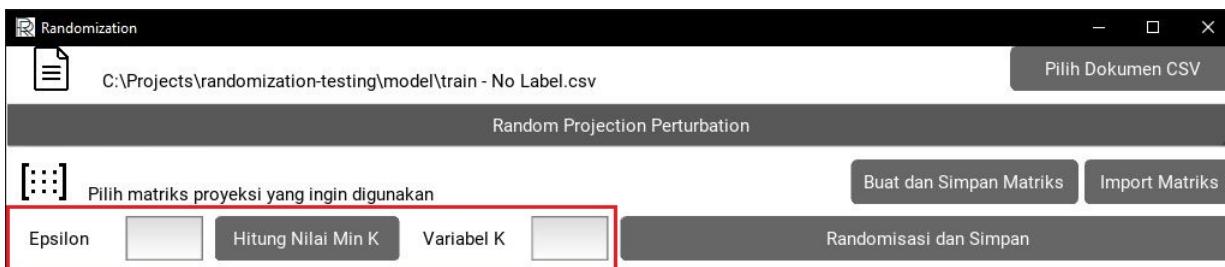
Gambar 5.8: Tampilan antarmuka saat perangkat lunak membuat dan menyimpan matriks



Gambar 5.9: Tampilan *popup* yang ditampilkan apabila matriks yang ingin diimpor tidak sesuai dengan dataset



Gambar 5.10: Tampilan *popup* yang ditampilkan apabila pengguna belum memilih dataset yang ingin dirandomisasi



Gambar 5.11: Tampilan antarmuka parameter teknik randomisasi *Random Projection Perturbation*

parameter tersebut dapat dilihat antarmukanya pada Gambar 5.11

Seperti yang disinggung pada subbab sebelumnya antarmuka perangkat lunak akan menyesuaikan secara otomatis sesuai teknik yang dipilih pengguna. Pada bagian parameter teknik randomisasi, perangkat lunak akan menyembunyikan antarmuka parameter *Random Projection Perturbation* apabila pengguna memilih teknik *Random Rotation Perturbation*. Antarmuka tersebut dapat dilihat pada Gambar 5.3 yang dikelilingi oleh kotak berwarna merah dan bernomor empat, dapat dilihat tidak ada parameter apapun yang tampil apabila teknik *Random Rotation Perturbation* yang dipilih.

Selain dua buah parameter, pada bagian ini juga ada sebuah tombol yaitu “Hitung Nilai Min K” yang memiliki fungsi untuk menghitung nilai minimal variabel K yang pengguna berikan. Pada teknik *Random Projection Perturbation*, ada beberapa persyaratan yang harus dipenuhi oleh pengguna dan salah satunya adalah variabel K yang diberikan harus melebihi sebuah nilai minimal yang dihitung berdasarkan ukuran dataset dan nilai variabel Epsilon. Oleh karena itu, sebelum tombol ini dapat berfungsi, pengguna harus memilih terlebih dahulu dataset yang ingin dirandomisasi dan memberikan masukan nilai variabel Epsilon yang sesuai dengan persyaratan variabel Epsilon yaitu nilainya lebih besar dari 0 dan kurang dari 1. Apabila pengguna belum memenuhi kedua persyaratan tersebut, tombol tidak akan berfungsi dan perangkat lunak akan



Gambar 5.12: Tampilan *popup* peringatan tombol ‘Hitung Nilai Min K’

menampilkan *popup* peringatan yang dapat dilihat pada Gambar 5.12. Nilai minimal variabel K akan ditampilkan pada bagian antarmuka deskripsi dataset yang akan dijelaskan pada subbab berikutnya.

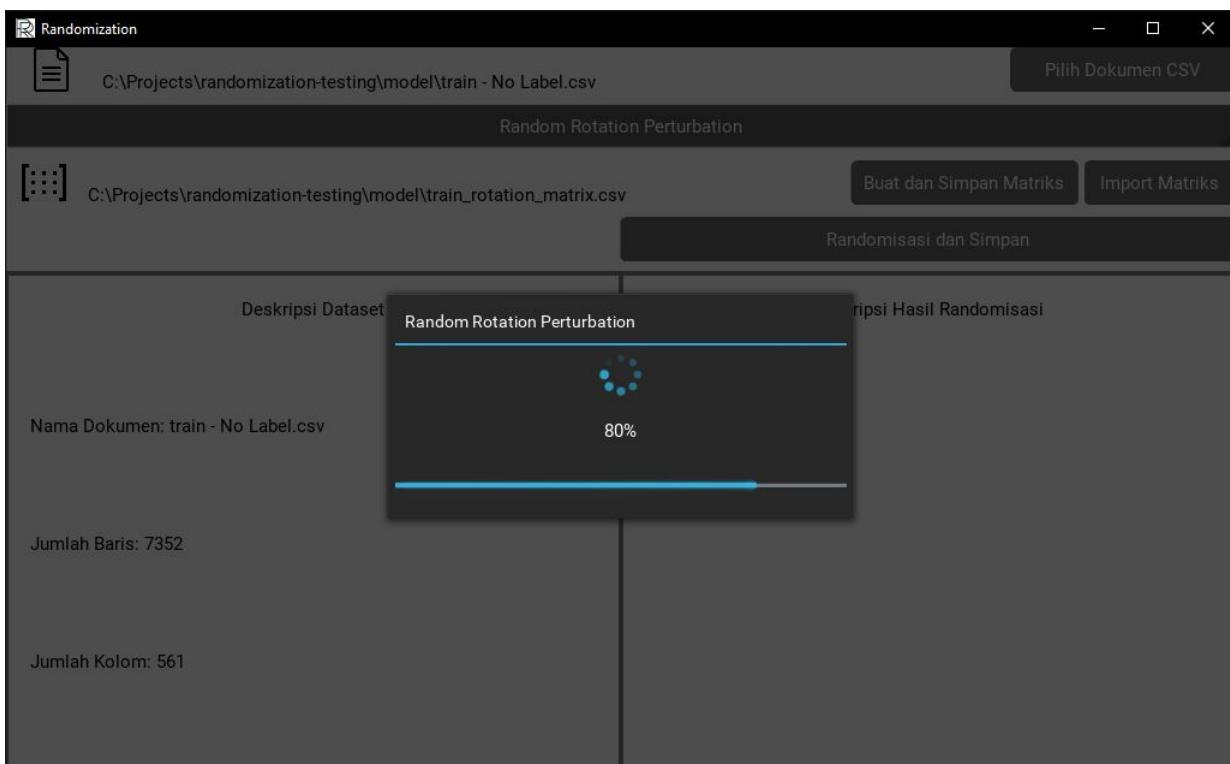
Perangkat lunak juga akan memberikan peringatan apabila nilai minimal K melebihi dimensi dari dataset yang ingin dirandomisasi karena salah satu persyaratan dari teknik *Random Projection Perturbation* adalah nilai variabel K harus lebih kecil daripada jumlah dimensi pada dataset yang ingin dirandomisasi. Apabila pengguna melakukan impor matriks maka variabel K akan terisi secara otomatis dan pengguna harus menyesuaikan nilai variabel Epsilon dengan variabel K yang tidak boleh diubah oleh pengguna.

Randomisasi dan Simpan

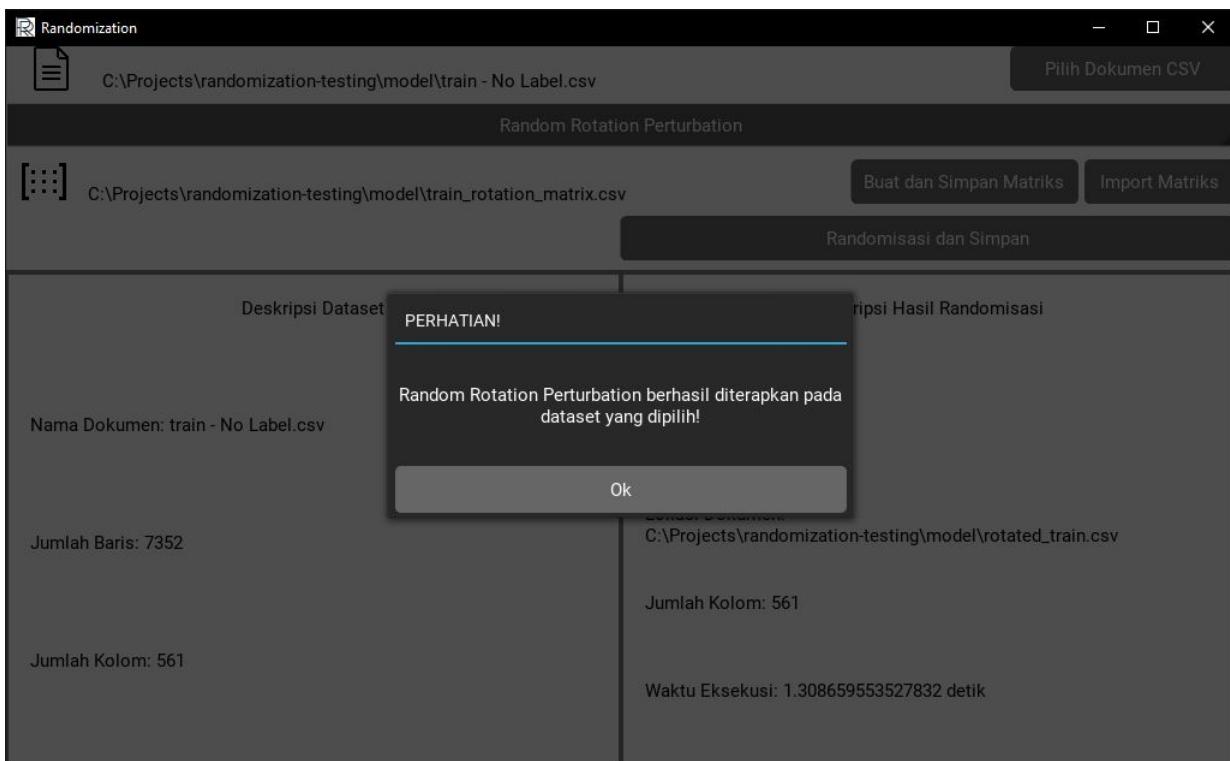
Setelah pengguna memberikan masukan yang sesuai dan mengatur pengaturan yang diinginkan maka pengguna telah dapat melakukan randomisasi dengan menekan tombol ‘Randomisasi dan Simpan’. Tombol ini akan menerapkan teknik randomisasi yang dipilih oleh pengguna terhadap dataset yang ingin dirandomisasi menggunakan matriks yang telah dibuat atau dipilih oleh pengguna dan parameter-parameter yang pengguna berikan. Tampilan antarmuka saat proses randomisasi dilakukan perangkat lunak dapat dilihat pada Gambar 5.13.

Setelah perangkat lunak berhasil melakukan randomisasi, perangkat lunak akan meminta pengguna untuk memilih direktori tempat penyimpanan dan nama dokumen hasil randomisasi. Perangkat lunak akan menyimpan hasil randomisasi dalam bentuk dokumen berjenis *comma-separated values*. Jendela baru untuk memilih direktori penyimpanan akan ditampilkan perangkat lunak, apabila pengguna membatalkan atau dengan kata lain menutup jendela tersebut tanpa memilih direktori penyimpanan maka perangkat lunak tidak akan melanjutkan proses randomisasi dan dianggap gagal. Tampilan antarmuka *popup* yang akan tampil setelah perangkat lunak berhasil melakukan proses randomisasi dan menyimpan hasilnya pada direktori yang pengguna pilih dapat dilihat pada Gambar 5.14. Perangkat lunak juga akan menampilkan berbagai informasi hasil randomisasi pada bagian antarmuka deskripsi hasil randomisasi yang akan dijelaskan pada subbab berikutnya.

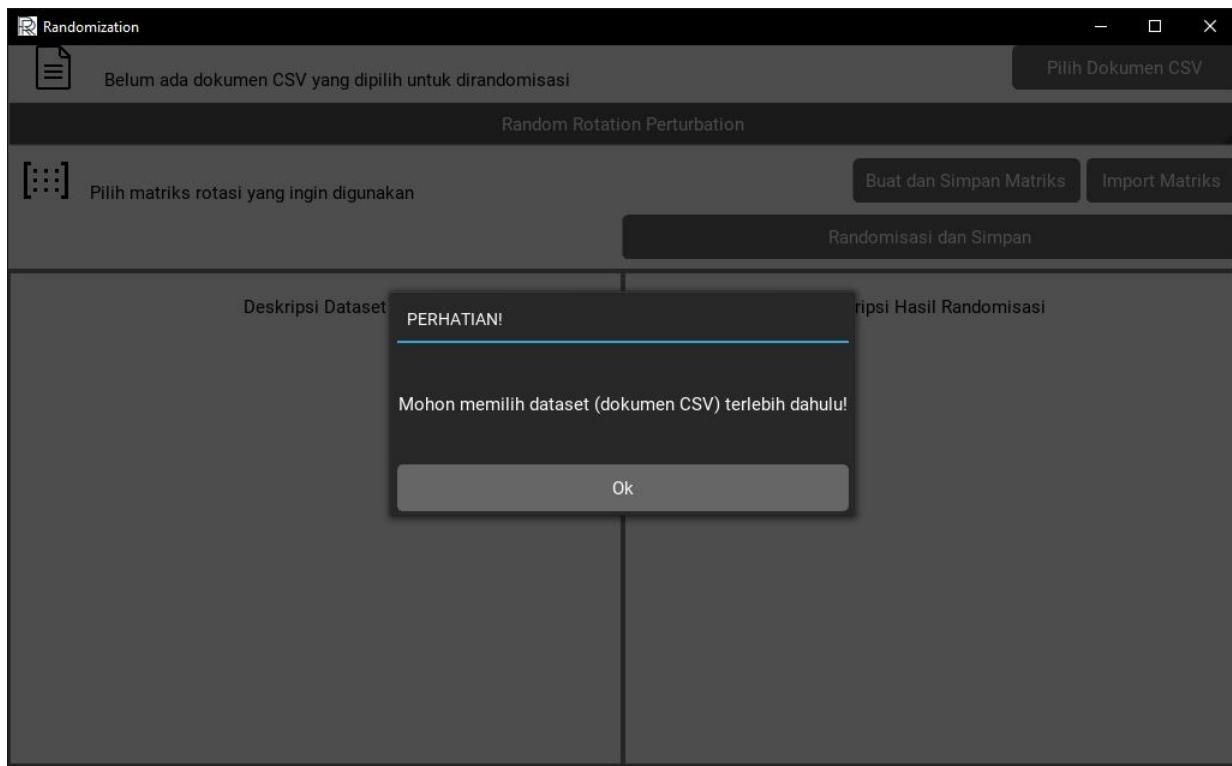
Ada beberapa persyaratan yang harus dipenuhi oleh pengguna sebelum melakukan randomisasi yaitu memilih dataset yang ingin dirandomisasi, memilih teknik randomisasi yang diinginkan, membuat atau memilih matriks rotasi atau proyeksi, dan memberikan masukan nilai yang sesuai persyaratan yang ada kepada parameter-parameter teknik randomisasi. Apabila ada persyaratan yang tidak dipenuhi oleh pengguna maka perangkat lunak akan menampilkan *popup* untuk



Gambar 5.13: Tampilan saat perangkat lunak sedang melakukan proses randomisasi



Gambar 5.14: Tampilan *popup* untuk memberitahukan pengguna bahwa randomisasi berhasil dilakukan



Gambar 5.15: Tampilan *popup* peringatan apabila pengguna belum memilih dataset yang diinginkan untuk dirandomisasi

memberikan peringatan kepada pengguna dan perangkat lunak tidak akan melanjutkan proses randomisasi. Perangkat lunak akan menampilkan *popup* peringatan terhadap pelanggaran masing-masing persyaratan tersebut, salah satu contoh tampilan antarmuka *popup* tersebut dapat dilihat pada Gambar 5.15.

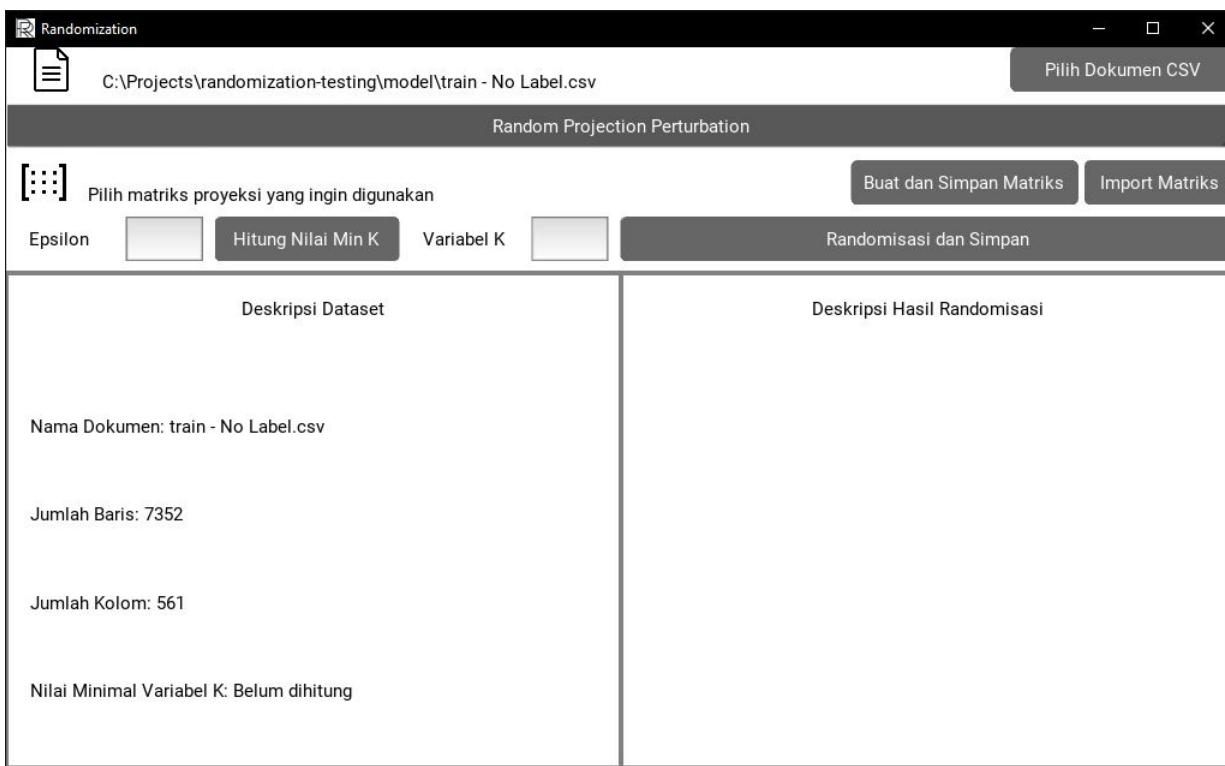
5.1.2 Deskripsi Dataset

Pengguna dapat melihat berbagai informasi dokumen *comma-separated values* yang dipilih sebagai dataset yang ingin dirandomisasi pada bagian antarmuka deskripsi dataset. Perangkat lunak akan menampilkan berbagai informasi dataset yaitu nama dokumen, jumlah baris dataset, jumlah kolom dataset, dan nilai minimal variabel K. Seperti yang disinggung pada subbab sebelumnya, pada awalnya nilai minimal variabel K belum diketahui karena belum dihitung. Pengguna harus mengisi variabel Epsilon dan menekan tombol “Hitung Nilai Min K” agar perangkat lunak menghitung nilai minimal variabel K dan dapat menampilkannya pada deskripsi dataset.

Bagian antarmuka deskripsi dataset ini akan selalu secara otomatis diperbarui setiap pengguna memilih dataset baru. Tampilan antarmuka bagian deskripsi dataset dapat dilihat pada Gambar 5.2 yang dikelilingi oleh kotak berwarna biru dan bermotor dua. Apabila pengguna telah memilih dataset yang diinginkan untuk dirandomisasi maka perangkat lunak secara otomatis akan memperbarui tampilan antarmuka deskripsi dataset yang dapat dilihat pada Gambar 5.16

5.1.3 Deskripsi Hasil Randomisasi

Perangkat lunak akan menampilkan isi dari deskripsi hasil randomisasi setelah pengguna menekan tombol “Randomisasi dan Simpan” dan perangkat lunak melakukan proses randomisasi. Bagian deskripsi hasil randomisasi ini akan menampilkan informasi-informasi yang berkaitan dengan hasil randomisasi dan deskripsi dataset yang baru. Informasi-informasi tersebut adalah status, lokasi dokumen, lokasi dokumen matriks yang dipakai, jumlah kolom, waktu eksekusi, nilai variabel



Gambar 5.16: Tampilan antarmuka deskripsi dataset setelah pengguna memilih dataset yang ingin dirandomisasi

Epsilon yang digunakan, dan nilai variabel K yang digunakan. Dua informasi terakhir tersebut hanya tampil apabila pengguna memilih teknik randomisasi “Random Projection Perturbation”.

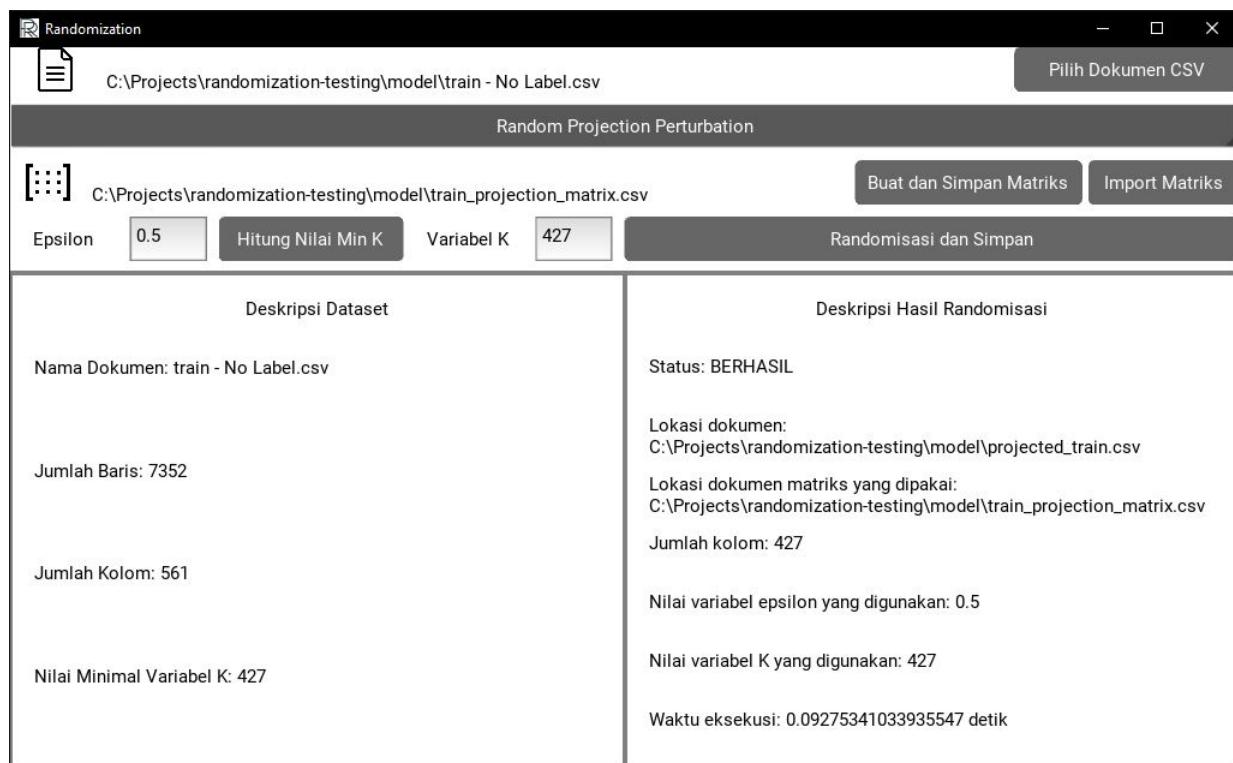
Bagian antarmuka deskripsi dataset ini akan selalu secara otomatis diperbaharui setiap pengguna memilih dataset baru. Tampilan antarmuka bagian deskripsi dataset dapat dilihat pada Gambar 5.2 yang dikelilingi oleh kotak berwarna biru dan bermotor dua. Apabila pengguna telah memilih dataset yang diinginkan untuk dirandomisasi maka perangkat lunak secara otomatis akan memperbaharui tampilan antarmuka deskripsi dataset yang dapat dilihat pada Gambar 5.17

5.2 Pengujian Perangkat Lunak Fungsional

Pengujian fungsional bertujuan untuk memastikan perangkat lunak randomisasi dapat menerapkan kedua teknik randomisasi yaitu *Random Rotation Perturbation* dan *Random Projection Perturbation* dengan baik terhadap dataset yang memenuhi syarat. Proses pengujian akan dilakukan dengan cara menerapkan kedua teknik tersebut dari awal memasukkan dataset sampai menghasilkan dataset yang telah dirandomisasi. Berikut pengujian pada setiap teknik randomisasi dengan menerapkan teknik penambangan data. Berikut pengujian pada setiap teknik randomisasi dengan menerapkan teknik penambangan data.

5.2.1 Teknik *Random Rotation Perturbation* dengan teknik penambangan data *K-Nearest Neighbors*

Pengujian teknik *Random Rotation Perturbation* akan menggunakan dataset *diabetes* yang berisi data kesehatan beberapa orang yang memiliki penyakit *diabetes* dan tidak. Dataset ini memiliki 8 buah fitur dan sebuah label. Delapan buah fitur tersebut akan dirandomisasi dan diharapkan hasilnya akan mengacak dataset sehingga nilai tiap fitur tersebut berbeda dari aslinya. Selain itu, matriks rotasi harus dapat disimpan dan digunakan kembali untuk lain kali. Kemudian diterapkan



Gambar 5.17: Tampilan antarmuka deskripsi hasil randomisasi setelah perangkat berhasil melakukan randomisasi

teknik klasifikasi dengan algoritma *K-nearest Neighbors* untuk menguji apakah dataset mempunyai akurasi model klasifikasi yang sama persis. Pengujian tersebut didasarkan pada sifat teknik *Random Rotation Perturbation* yang menjamin jarak Euclidean setiap titik tidak berubah sama sekali.

Label pada dataset *diabetes* yang ingin dirandomisasi harus dihilangkan terlebih dahulu agar hanya fitur dataset tersebut saja yang terrandomisasi. Berikut akan ditampilkan 20 baris pertama dataset asli dan dataset yang telah dirandomisasi masing-masing pada Gambar 5.18 dan Gambar 5.19. Dapat dilihat pada gambar tersebut, dataset setelah dirandomisasi memiliki nilai yang sangat berbeda dengan aslinya, bahkan nilai yang sama pada beberapa baris di dataset asli tidak sama dengan dataset yang telah dirandomisasi pada baris yang sama seperti pada kolom *insulin* baris satu sampai tiga memiliki nilai 0 pada dataset asli tetapi pada dataset yang telah dirandomisasi ketiga baris tersebut memiliki nilai yang berbeda antara satu dengan yang lainnya.

Teknik penambangan data *K-nearest Neighbors* diterapkan menggunakan delapan buah fitur yang ada. Dataset akan dibagi dua menjadi *train set* dan *test set* yang masing-masing berguna untuk melatih model dan menguji akurasi model. Akurasi akan dihitung dengan jumlah tetangga (nilai *k*) dari 1 sampai 30. Nilai akurasi pada dataset asli dan dataset yang telah dirandomisasi masing-masing dapat dilihat pada Listing 5.1 dan Listing 5.2. Implementasi kode teknik penambangan data ini diterapkan dengan bahasa pemrograman Python dan dibantu oleh *library Scikit-learn*.

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
6	148	72	35	0	33.6		0.627	50
1	85	66	29	0	26.6		0.351	31
8	183	64	0	0	23.3		0.672	32
1	89	66	23	94	28.1		0.167	21
0	137	40	35	168	43.1		2.288	33
5	116	74	0	0	25.6		0.201	30
3	78	50	32	88	31		0.248	26
10	115	0	0	0	35.3		0.134	29
2	197	70	45	543	30.5		0.158	53
8	125	96	0	0	0		0.232	54
4	110	92	0	0	37.6		0.191	30
10	168	74	0	0	38		0.537	34
10	139	80	0	0	27.1		1.441	57
1	189	60	23	846	30.1		0.398	59
5	166	72	19	175	25.8		0.587	51
7	100	0	0	0	30		0.484	32
0	118	84	47	230	45.8		0.551	31
7	107	74	0	0	29.6		0.254	31
1	103	30	38	83	43.3		0.183	33
1	115	70	30	96	34.6		0.529	32

Gambar 5.18: Dua puluh baris pertama dataset *diabetes* asli

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
-91.37651756	-80.18724466	23.05889315	-18.9693363	110.5088496	-133.4205069		86.0698456	-48.84586837
-75.53749369	-36.05193919	13.61201036	-17.05774302	96.33252668	-90.82560108		76.21980243	-40.35699091
-92.42942058	-94.56928839	1.851857108	3.981465625	114.1469917	-163.8819462		58.28394087	-55.95014122
-126.993899	-19.97252557	41.95044912	12.71359672	146.366212	-80.48590451		42.87623651	-14.4243376
-175.912903	-54.87009029	87.61024212	41.12103859	178.1742884	-93.93255265		23.24115845	15.4702748
-77.49366851	-53.89318148	-6.042582122	-6.564720711	113.553306	-114.0705412		64.92332441	-50.40619642
-120.6678891	-23.70836905	48.43553354	7.584019346	132.5262512	-69.33207772		43.64006616	-4.804724932
-69.83804655	-86.30230284	6.576880008	10.61500227	73.50342656	-98.42959311		36.60831617	-13.21252366
-380.5808875	-22.67446556	243.749478	117.0593059	410.6210675	-85.54982523		-87.46332942	95.24128787
-66.68175924	-53.19432898	4.407546648	-29.43172457	136.7787079	-128.4024803		61.29313373	-67.66505132
-82.77052893	-48.11261586	-14.37485848	-7.899390021	122.7354016	-107.9874489		80.23207042	-54.51218446
-95.0159058	-89.40851965	-6.621076283	0.876107771	119.2961308	-148.640664		71.21720253	-53.07414301
-75.73544771	-79.23855094	1.003430106	-18.05831711	130.9163815	-126.3571694		70.26494314	-54.22574441
-521.8891047	17.70297409	349.3454371	195.9590596	583.3955701	-27.00658802		-199.2092795	183.302024
-177.2238609	-63.14743279	84.04348195	29.40445346	215.1435288	-120.5701915		17.77020272	-6.772259965
-62.0408439	-77.4456268	9.358955118	7.376454038	73.79847691	-87.95326333		34.66589639	-13.57824067
-215.9630713	-15.85940033	104.9031479	39.45825766	228.6949467	-77.25266936		29.06237766	18.15464029
-76.61265022	-51.2057761	-8.491558255	-8.751231878	113.380821	-105.9266292		67.03349362	-46.88195403
-124.2450241	-53.51904364	56.4471758	15.2220361	121.7116444	-79.50822308		47.21844068	0.938260398
-135.5651967	-39.46936443	49.87545502	11.63635561	153.8601703	-96.05165177		51.95779491	-17.56987501

Gambar 5.19: Dua puluh baris pertama dataset *diabetes* setelah dirandomisasi

Listing 5.1: Akurasi Dataset Asli

Akurasi setiap K pada training set dataset asli:

```

1: 1.0
2: 0.9887105549510338
3: 0.9919749727965179
4: 0.9862622415669206
5: 0.984221980413493
6: 0.984221980413493
7: 0.9817736670293797
8: 0.9790533188248096
9: 0.9771490750816104
10: 0.9764689880304679
11: 0.9741566920565833
12: 0.9734766050054406
13: 0.9721164309031556
14: 0.9713003264417845
15: 0.9691240478781284
16: 0.9685799782372143
17: 0.9674918389553863
18: 0.9681719260065288
19: 0.9659956474428727
20: 0.9665397170837867
21: 0.9635473340587595
22: 0.9642274211099021
23: 0.9624591947769314
24: 0.9642274211099021
25: 0.9609630032644179
26: 0.9615070729053319
27: 0.9604189336235038
28: 0.9604189336235038
29: 0.9593307943416758
30: 0.9594668117519043

```

Akurasi setiap K pada test set dataset asli:

```

1: 0.8785205293518833
2: 0.8659653885307091
3: 0.8907363420427553
4: 0.8934509670851714
5: 0.9002375296912114
6: 0.9019341703427214
7: 0.9022734984730234
8: 0.9060061079063454
9: 0.9053274516457415
10: 0.9066847641669494
11: 0.9039701391245334
12: 0.9029521547336274
13: 0.9056667797760435
14: 0.9049881235154394
15: 0.9046487953851374
16: 0.9077027485578555
17: 0.9070240922972514
18: 0.9077027485578555
19: 0.9070240922972514
20: 0.9056667797760435

```

Listing 5.2: Akurasi Dataset Randomisasi

Akurasi setiap K pada training set dataset randomisasi:

```

1: 1.0
2: 0.984221980413493
3: 0.9896626768226333
4: 0.9835418933623504
5: 0.9828618063112078
6: 0.9797334058759521
7: 0.9782372143634385
8: 0.9759249183895539
9: 0.9736126224156693
10: 0.9737486398258978
11: 0.9721164309031556
12: 0.9717083786724701
13: 0.9691240478781284
14: 0.9683079434167573
15: 0.9666757344940152
16: 0.9658596300326442
17: 0.9640914036996736
18: 0.9640914036996736
19: 0.9623231773667029
20: 0.9623231773667029
21: 0.9605549510337323
22: 0.9602829162132753
23: 0.9586507072905331
24: 0.9586507072905331
25: 0.9570184983677911
26: 0.9578346028291621
27: 0.9566104461371056
28: 0.956474428726877
29: 0.9547062023939065
30: 0.9549782372143635

```

Akurasi setiap K pada test set dataset randomisasi:

```

1: 0.8649474041398032
2: 0.8578215134034611
3: 0.8775025449609772
4: 0.8781812012215813
5: 0.8812351543942993
6: 0.8870037326094333
7: 0.8903970139124533
8: 0.8897183576518494
9: 0.8917543264336614
10: 0.8941296233457754
11: 0.8907363420427553
12: 0.8914149983033594
13: 0.8934509670851714
14: 0.8951476077366813
15: 0.8954869358669834
16: 0.8954869358669834
17: 0.8982015609093994
18: 0.8975229046487954
19: 0.8992195453003053
20: 0.9002375296912114

```

CustomerID	Gender	Age	Annual Income (k\$)	Spending Score (1-100)
1	Male	19	15	39
2	Male	21	15	81
3	Female	20	16	6
4	Female	23	16	77
5	Female	31	17	40
6	Female	22	17	76
7	Female	35	18	6
8	Female	23	18	94
9	Male	64	19	3
10	Female	30	19	72
11	Male	67	19	14
12	Female	35	19	99
13	Female	58	20	15
14	Female	24	20	77
15	Male	37	20	13
16	Male	22	20	79
17	Female	35	21	35
18	Male	20	21	66
19	Male	52	23	29
20	Female	35	23	98

Gambar 5.20: Dua puluh baris pertama dataset *mall_customers* asli

Dapat dilihat pada kedua listing tersebut akurasi *training set* dan *test set* pada kedua dataset sama persis. Hal ini dikarenakan jarak Euclidean kedua buah dataset tidak berubah sama sekali. Oleh karena itu, perangkat lunak berhasil menerapkan dengan baik teknik *Random Rotation Perturbation* untuk teknik penambangan data klasifikasi.

5.2.2 Teknik *Random Rotation Perturbation* dengan teknik penambangan data *K-means*

Pengujian teknik *Random Rotation Perturbation* akan menggunakan dataset *mall_customers* yang berisi informasi pribadi pelanggan sebuah mall. Dataset ini memiliki empat buah fitur yaitu jenis kelamin, umur, penghasilan, dan skor pengeluaran. empat buah fitur tersebut akan dirandomisasi dan diharapkan hasilnya akan mengacak dataset sehingga nilai tiap fitur tersebut berbeda dari aslinya. Selain itu, matriks rotasi harus dapat disimpan dan digunakan kembali untuk lain kali. Kemudian diterapkan teknik *clustering* dengan algoritma *K-means* untuk menguji apakah dataset menghasilkan kluster dan bentuk yang sama dengan sudut yang berbeda saat divisualisasikan. Pengujian tersebut didasarkan pada sifat teknik *Random Rotation Perturbation* yang menjamin jarak Euclidean setiap titik tidak berubah sama sekali.

Berikut akan ditampilkan 20 baris pertama dataset asli dan dataset yang telah dirandomisasi masing-masing pada Gambar 5.20 dan Gambar 5.21. Dapat dilihat pada gambar tersebut, dataset setelah dirandomisasi memiliki nilai yang sangat berbeda dengan aslinya.

Teknik penambangan data *K-means* diterapkan menggunakan empat buah fitur yang ada. Akurasi akan dihitung dengan jumlah tetangga (nilai k) dari 1 sampai 20. Visualisasi *cluster* pada dataset asli dan dataset yang telah dirandomisasi masing-masing dapat dilihat pada Gambar 5.22 dan Gambar 5.23. Implementasi kode teknik penambangan data ini diterapkan dengan bahasa pemrograman Python dan dibantu oleh *library* Scikit-learn.

Dapat dilihat pada kedua *scatter plot* tersebut *cluster* yang dihasilkan sama yaitu 6 *cluster* dan bentuk dari *cluster-cluster* tersebut sama walaupun titik-titik yang ada memiliki lokasi yang

CustomerID	Gender	Age	Annual Income (k\$)	Spending Score (1-100)
1	Male	30.80293576	-74.70371227	-150.6802393
2	Male	11.64235717	-100.965712	-177.34819
3	Female	46.00730281	-52.26243313	-131.8065483
4	Female	13.72109486	-96.94089629	-176.6555807
5	Female	28.78173046	-66.65349299	-159.6305856
6	Female	15.29164496	-97.12815467	-175.9496722
7	Female	43.6079501	-41.3215056	-141.9819702
8	Female	7.89446913	-108.3817053	-187.9762839
9	Male	37.85168599	-17.88714487	-158.3739244
10	Female	16.58136376	-88.74788781	-179.4292919
11	Male	32.15552849	-22.92463185	-166.8696594
12	Female	3.246982854	-102.8696182	-198.8398822
13	Female	35.0299167	-30.37999029	-162.4929572
14	Female	16.86176701	-96.61595268	-179.3071067
15	Male	41.66545616	-44.66880168	-148.4644332
16	Male	16.52297528	-99.4240021	-179.2982406
17	Female	33.31692523	-60.79510937	-161.0836341
18	Male	23.6853071	-92.42640702	-170.7113514
19	Male	33.02903428	-44.40340067	-168.8443689
20	Female	7.10459415	-102.6273334	-200.2751986
21	Male	35.87775807	-61.108908	-162.6145376

Gambar 5.21: Dua puluh baris pertama dataset *mall_customers* setelah dirandomisasi

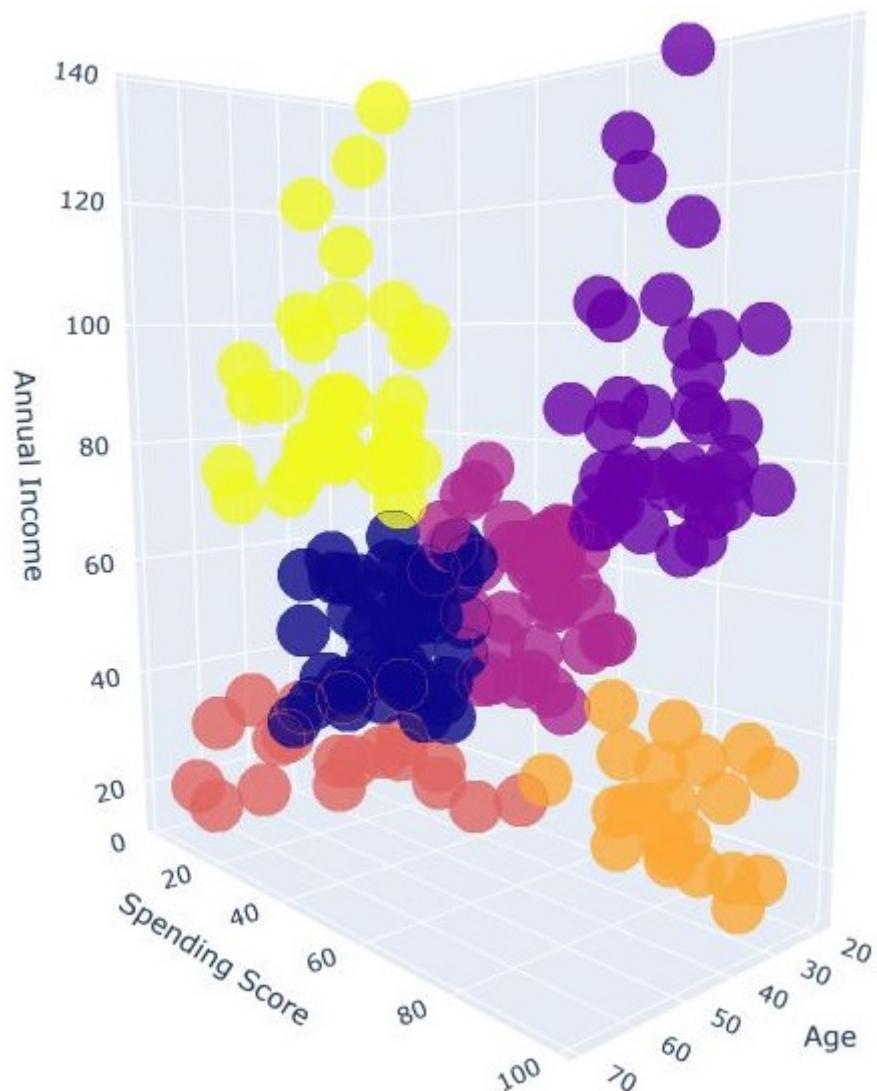
berbeda. *Adjusted Rand Index* pada kedua model tersebut memiliki nilai 1 yang berarti anggota *cluster* pada setiap cluster antara kedua model tersebut sama persis. Hal ini dikarenakan jarak Euclidean kedua buah dataset tidak berubah sama sekali. Oleh karena itu, perangkat lunak berhasil menerapkan dengan baik teknik *Random Rotation Perturbation* untuk penambangan data *clustering*.

5.2.3 Teknik *Random Projection Perturbation* dengan teknik penambangan data *K-Nearest Neighbors*

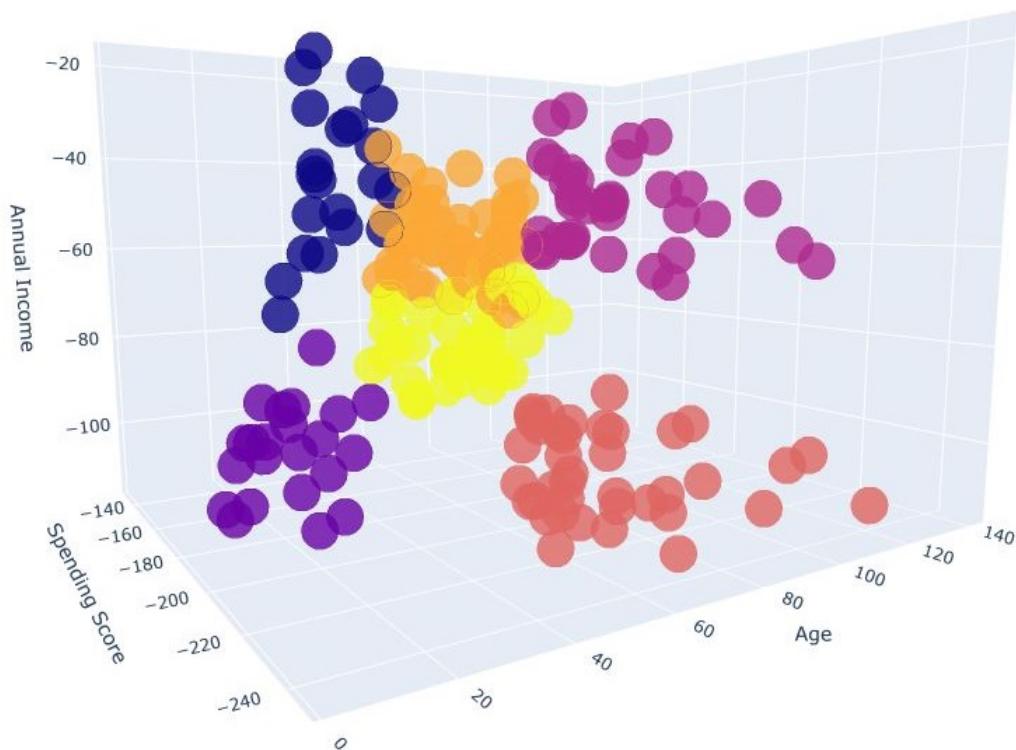
Pengujian teknik *Random Projection Perturbation* akan menggunakan dataset *mobile_sensor* yang berisi data sensor *smartphone* banyak orang yang sedang melakukan aktivitas tertentu seperti berdiri, duduk, berjalan, berjalan menanjak, berjalan menurun, dan lain-lain. Dataset ini memiliki 561 buah fitur dan sebuah label. Seluruh fitur tersebut akan dirandomisasi dan diharapkan hasilnya akan mengacak dataset sehingga nilai tiap fitur tersebut berbeda dari aslinya. Selain itu, matriks proyeksi harus dapat disimpan dan digunakan kembali untuk lain kali. Kemudian diterapkan teknik klasifikasi dengan algoritma *K-nearest Neighbors* untuk menguji apakah dataset mempunyai akurasi model klasifikasi yang hampir sama. Pengujian tersebut didasarkan pada sifat teknik *Random Projection Perturbation* yang menjamin jarak Euclidean setiap titik tidak berubah jauh dengan besar *error* yang ditentukan pengguna.

Label pada dataset *mobile_sensor* yang ingin dirandomisasi harus dihilangkan terlebih dahulu agar hanya fitur dataset tersebut saja yang terrandomisasi. Berikut akan ditampilkan 7 fitur terakhir serta label pada 20 baris pertama dataset asli dan dataset yang telah dirandomisasi masing-masing pada Gambar 5.24 dan Gambar 5.25. Dapat dilihat pada gambar tersebut, dataset setelah dirandomisasi memiliki nilai yang sangat berbeda dengan aslinya.

Teknik penambangan data *K-nearest Neighbors* diterapkan menggunakan delapan buah fitur yang ada. Dataset akan dibagi dua menjadi *train set* dan *test set* yang masing-masing berguna untuk melatih model dan menguji akurasi model. Akurasi akan dihitung dengan jumlah tetangga (nilai *k*) dari 1 sampai 20. Nilai akurasi pada dataset asli dan dataset yang telah dirandomisasi masing-masing dapat dilihat pada Listing 5.3 dan Listing 5.4. Implementasi kode teknik penambangan data ini diterapkan dengan bahasa pemrograman Python dan dibantu oleh *library Scikit-learn*.



Gambar 5.22: Visualisasi *cluster* pada dataset yang asli

Gambar 5.23: Visualisasi *cluster* pada dataset yang telah dirotasi

angle(tBodyAcc	angle(tBodyGyr	angle(tBodyGyr	angle(X,gravity	angle(Y,gravit	angle(Z,gravity)	Activity
0.030400372	-0.46476139	-0.018445884	-0.84124676	0.17994061	-0.058626924	STANDING
-0.007434566	-0.73262621	0.70351059	-0.8447876	0.18028889	-0.054316717	STANDING
0.17789948	0.10069921	0.80852908	-0.84893347	0.18063731	-0.049117815	STANDING
-0.012892494	0.64001104	-0.48536645	-0.84864938	0.18193476	-0.047663183	STANDING
0.12254196	0.69357829	-0.61597061	-0.84786525	0.18515116	-0.043892254	STANDING
-0.14343901	0.27504075	-0.36822404	-0.84963158	0.18482251	-0.042126383	STANDING
-0.23062193	0.01463669	-0.18951153	-0.85215025	0.18216997	-0.043009987	STANDING
0.59399581	-0.56187067	0.46738333	-0.85101671	0.18377851	-0.041975833	STANDING
0.080936389	-0.23431263	0.11779701	-0.84797148	0.18898248	-0.037363927	STANDING
-0.12773018	-0.48287054	-0.070670135	-0.84829438	0.19031033	-0.034417291	STANDING
0.59579055	-0.47580245	0.11593062	-0.85156175	0.18760932	-0.034681168	STANDING
-0.065980273	0.57886112	-0.65194513	-0.8527234	0.18605036	-0.035852089	STANDING
-0.10122189	0.63908399	0.76548488	-0.85065446	0.18761054	-0.035997955	STANDING
-0.090277545	-0.1324028	0.49881419	-0.84977267	0.1888122	-0.035063399	STANDING
-0.05871932	0.031207971	-0.26879133	-0.73093729	0.28315855	0.036443909	STANDING
-0.029076714	-0.013034217	-0.056927156	-0.76110079	0.26311858	0.02417211	STANDING
-0.048109773	-0.34047349	-0.22915457	-0.75917219	0.26432447	0.027014344	STANDING
0.092367048	-0.82223861	0.36755744	-0.75936327	0.26403279	0.029664019	STANDING
-0.033006915	-0.24057155	0.78819291	-0.76105187	0.26288599	0.029345734	STANDING
0.10256897	0.066134774	-0.41172948	-0.76062023	0.26316935	0.029573033	STANDING

Gambar 5.24: Dua puluh baris pertama dataset *mobile_sensor* yang asli

420	421	422	423	424	425	426	Activity
-0.060976974	-0.28720261	0.749987819	-0.314597972	0.862413791	-0.184264921	0.30054771	STANDING
-0.118304471	-0.367147618	0.968242315	-0.635217584	0.657331951	0.213839293	0.471794176	STANDING
-0.192180445	-0.31118681	0.978629636	-0.633964644	0.39068659	0.152568399	0.62044912	STANDING
0.025086451	-0.504801202	0.7858091	-0.907280086	0.869665518	0.3148417	0.468902009	STANDING
-0.102949398	-0.323268879	0.952716391	-0.899688251	0.629212652	0.223750593	0.684361819	STANDING
-0.147008408	-0.447760471	1.123065848	-0.637227607	0.67689574	0.228867562	0.747769951	STANDING
-0.163893961	-0.357135652	0.87956909	-0.938537783	0.615377996	0.334211022	0.427323639	STANDING
-0.247640685	-0.365958546	0.873331126	-0.736313605	0.73465538	0.324703704	0.46087866	STANDING
-0.102040114	-0.272211976	0.990322769	-0.927303583	0.551974442	0.130154901	0.945745752	STANDING
-0.040080611	-0.182285976	0.865031503	-0.913260332	0.679558281	0.227692766	0.44382714	STANDING
-0.048529732	-0.202250452	0.723406522	-0.960376711	0.583341661	0.331778206	0.516310888	STANDING
0.044383724	-0.357253613	0.893175211	-0.769930023	0.684675053	0.269595542	0.705982695	STANDING
0.116948369	-0.345136512	0.930098548	-0.6464027	0.467892769	0.262751838	0.603885473	STANDING
0.248042906	-0.303948701	0.737485374	-0.859802493	0.647402395	0.296905238	0.514994603	STANDING
-0.709711844	0.032196412	1.157554755	-0.647494103	0.548833741	0.301928493	0.474742466	STANDING
-0.137785364	-0.085822981	0.85368708	-0.63551871	0.987405426	0.477568422	0.590664492	STANDING
-0.337459958	-0.243026448	1.095896604	-0.707716697	0.739239517	0.04972716	0.494796784	STANDING
-0.160292217	-0.477462394	0.904414643	-0.587028222	0.662549199	0.143131645	0.361362174	STANDING
0.277181657	-0.236343455	0.672690154	-0.756133708	0.562502489	0.395388677	0.518869726	STANDING
-0.005303415	-0.333661577	0.813756698	-0.928377447	0.7866421	0.402249818	0.458687589	STANDING

Gambar 5.25: Dua puluh baris pertama dataset *mobile_sensor* setelah dirandomisasi

Listing 5.3: Akurasi Dataset Asli

Akurasi setiap K pada training set dataset asli :

1: 1.0
 2: 0.9887105549510338
 3: 0.9919749727965179
 4: 0.9862622415669206
 5: 0.984221980413493
 6: 0.984221980413493
 7: 0.9817736670293797
 8: 0.9790533188248096
 9: 0.9771490750816104
 10: 0.9764689880304679
 11: 0.9741566920565833
 12: 0.9734766050054406
 13: 0.9721164309031556
 14: 0.9713003264417845
 15: 0.9691240478781284
 16: 0.9685799782372143
 17: 0.9674918389553863
 18: 0.9681719260065288
 19: 0.9659956474428727
 20: 0.9665397170837867

Akurasi setiap K pada test set dataset asli :

1: 0.8785205293518833
 2: 0.8659653885307091
 3: 0.8907363420427553
 4: 0.8934509670851714
 5: 0.9002375296912114
 6: 0.9019341703427214
 7: 0.9022734984730234
 8: 0.9060061079063454

Listing 5.4: Akurasi Dataset Randomisasi

Akurasi setiap K pada training set dataset randomisasi :

1: 1.0
 2: 0.984221980413493
 3: 0.9896626768226333
 4: 0.9835418933623504
 5: 0.9828618063112078
 6: 0.9797334058759521
 7: 0.9782372143634385
 8: 0.9759249183895539
 9: 0.9736126224156693
 10: 0.9737486398258978
 11: 0.9721164309031556
 12: 0.9717083786724701
 13: 0.9691240478781284
 14: 0.9683079434167573
 15: 0.9666757344940152
 16: 0.9658596300326442
 17: 0.9640914036996736
 18: 0.9640914036996736
 19: 0.9623231773667029
 20: 0.9623231773667029

Akurasi setiap K pada test set dataset randomisasi :

1: 0.8649474041398032
 2: 0.8578215134034611
 3: 0.8775025449609772
 4: 0.8781812012215813
 5: 0.8812351543942993
 6: 0.8870037326094333
 7: 0.8903970139124533
 8: 0.8897183576518494

Dapat dilihat pada kedua listing tersebut akurasi *test set* pada kedua dataset berbeda. Akurasi *test set* tertinggi pada dataset asli adalah sebesar 0.9077027485578555 dengan nilai k sebesar 16. Sementara pada dataset yang telah dirandomisasi, akurasi *test set* tertingginya adalah sebesar 0.9002375296912114 dengan nilai k sebesar 20. Jika dihitung perbedaan akurasinya adalah sebesar 0.0074652188666441 yang mana relatif kecil tetapi ada perbedaan pada nilai k yang memiliki akurasi tertinggi. Perbedaan nilai k ini akan diuji lebih lanjut pada pengujian eksperimental. Dengan perbedaan akurasi yang relatif kecil dan dataset yang telah dirandomisasi teracak dengan baik maka dapat disimpulkan perangkat lunak berhasil menerapkan teknik *Random Projection Perturbation* untuk teknik penambangan data *clustering*.

5.2.4 Teknik *Random Projection Perturbation* dengan teknik penambangan data *K-means*

Pengujian teknik *Random Projection Perturbation* akan menggunakan dataset *mobile_sensor* yang sama seperti sebelumnya dan dapat dilihat pada Gambar 5.24 dan Gambar 5.25. Dataset ini akan diterapkan teknik *clustering* dengan algoritma *K-means* untuk menguji apakah dataset menghasilkan kluster dan bentuk yang sama dengan sudut yang berbeda saat divisualisasikan. Pengujian tersebut didasarkan pada sifat teknik *Random Projection Perturbation* yang menjamin jarak Euclidean setiap titik tidak berubah jauh dengan besar *error* yang ditentukan pengguna.

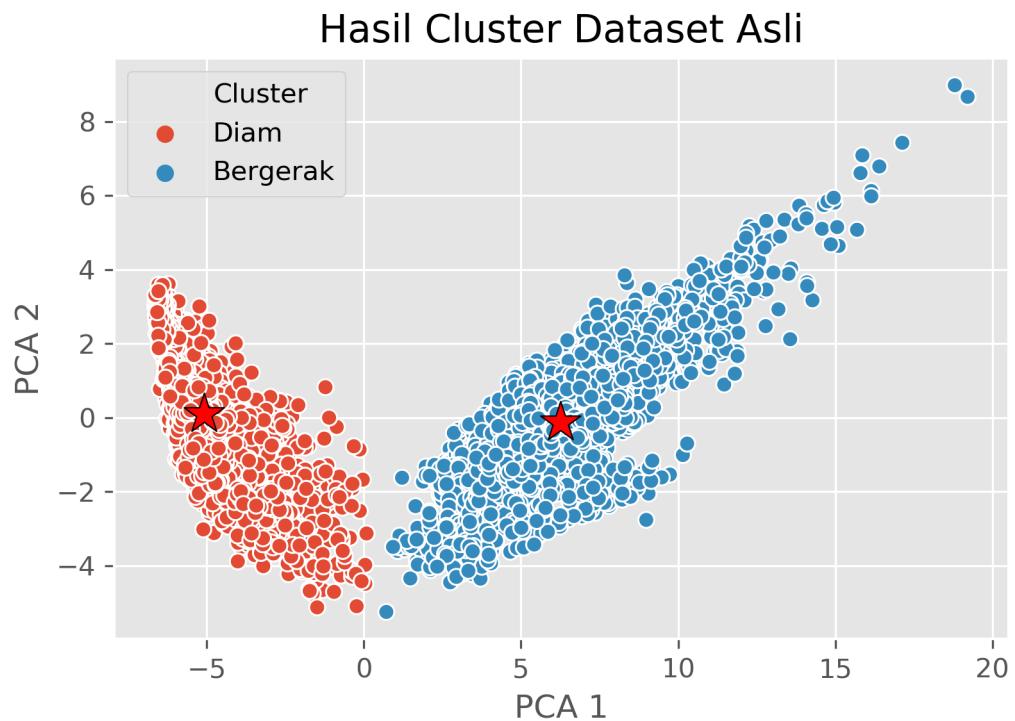
Sebelum melakukan teknik *clustering* dengan algoritma *K-means*, dataset yang memiliki fitur yang sangat banyak tersebut dimensinya harus direduksi terlebih dahulu agar model dapat divisualisasikan dengan mudah. Dataset yang akan *dicluster* akan direduksi dimensinya sampai hanya memiliki 2 dimensi. Reduksi dimensi dilakukan dengan menerapkan teknik *Principal Component Analysis* yang sudah umum digunakan saat penambangan data dan hasilnya relatif baik.

Teknik penambangan data *K-means* diterapkan menggunakan dua buah fitur yang ada hasil dari teknik *Principal Component Analysis*. Model *K-means* akan dibuat dengan nilai k yang terbaik dihitung menggunakan metode Elbow dan nilai k yang memiliki *Sillhouette Score* tertinggi. Visualisasi *cluster* pada dataset asli dan dataset yang telah dirandomisasi masing-masing dapat dilihat pada Gambar 5.26 dan Gambar 5.27. Implementasi kode teknik penambangan data ini diterapkan dengan bahasa pemrograman Python dan dibantu oleh *library Scikit-learn*.

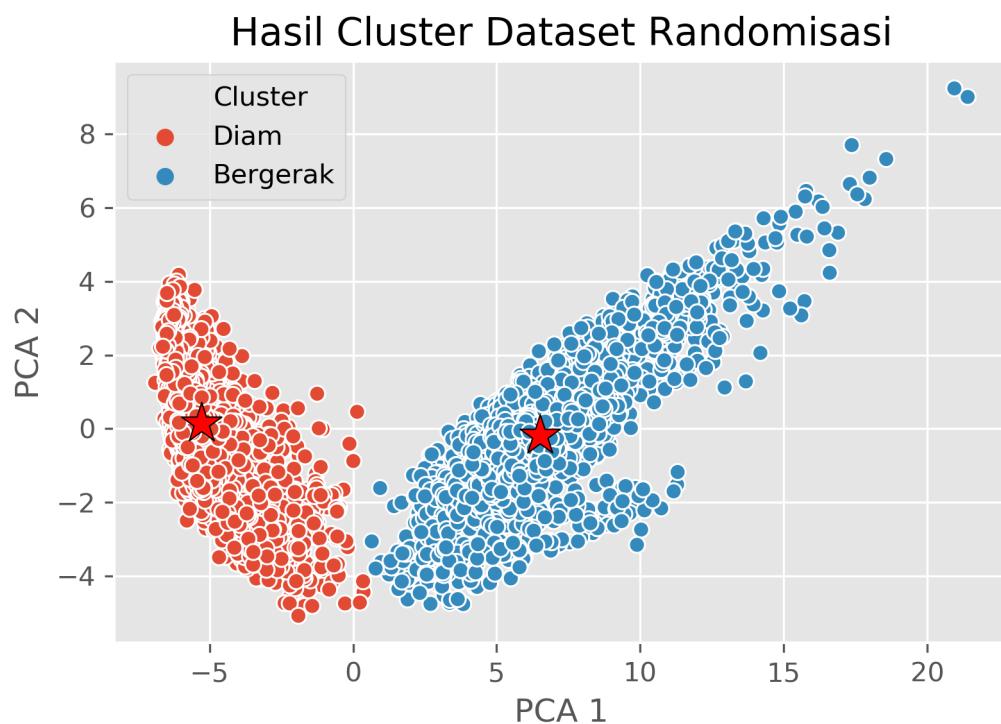
Hasil *clustering* menyatakan bahwa ada dua buah *cluster* yaitu kelompok orang-orang yang diam dan kelompok orang-orang yang bergerak. Dapat dilihat pada kedua *scatter plot* tersebut *cluster* yang dihasilkan sama yaitu 2 *cluster* walaupun ada sedikit perbedaan lokasi titik-titik yang ada pada kedua *scatter plot* tersebut. *Adjusted Rand Index* pada kedua model tersebut memiliki nilai sebesar 0.9994558701020273, mendekati angka 1. Dengan *Adjusted Rand Index* sebesar itu dapat diartikan anggota *cluster* pada setiap cluster antara kedua model tersebut hampir sama persis. Hal ini dikarenakan jarak Euclidean kedua buah dataset tidak rusak secara signifikan dan *errornya* terkendali sesuai yang pengguna inginkan. Oleh karena itu, perangkat lunak berhasil menerapkan dengan baik teknik *Random Projection Perturbation* untuk penambangan data *clustering*.

5.3 Pengujian Eksperimental

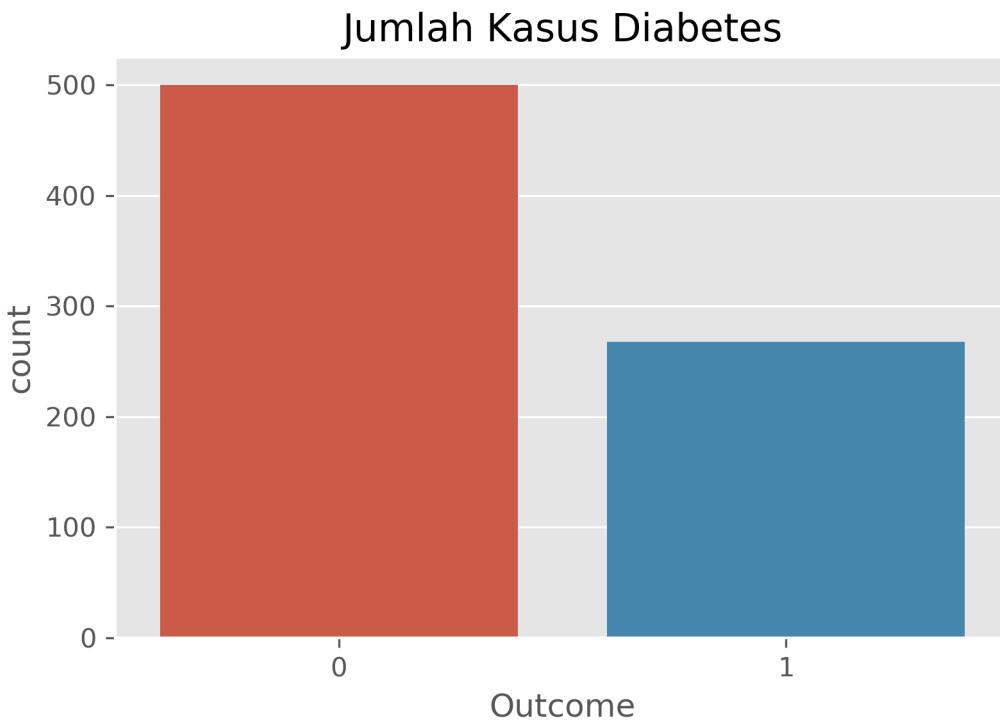
Pengujian eksperimental bertujuan untuk menguji kualitas hasil dari perangkat lunak randomisasi pada kedua teknik randomisasi dan membandingkan kualitas hasil randomisasi dari kedua teknik tersebut pada penambangan data. Pengujian akan dilakukan dalam dua bagian yaitu pengujian kualitas teknik randomisasi untuk penambangan data klasifikasi dan pengujian kualitas teknik randomisasi untuk penambangan data *clustering*. Pada setiap bagian tersebut akan dibagi lagi menjadi tiga bagian yaitu pengujian teknik *Random Rotation Perturbation*, pengujian teknik *Random Projection Perturbation*, dan pengujian untuk membandingkan kedua teknik randomisasi tersebut. Pengujian akan dilakukan dengan menggunakan program pengujian yang menerapkan teknik penambangan data yang telah dibuat pada bahasa pemrograman Python dan didukung oleh



Gambar 5.26: Visualisasi *cluster* pada dataset yang asli



Gambar 5.27: Visualisasi *cluster* pada dataset yang telah diproyeksi



Gambar 5.28: Histogram distribusi label dataset *diabetes*

perangkat lunak *Spyder* untuk menampilkan visualisasi hasil penambangan data.

5.3.1 Penambangan Data Klasifikasi

Pengujian dengan penambangan data klasifikasi akan berpusat pada pembuatan model dengan dataset asli dan dataset yang telah dirandomisasi dan membandingkan kualitas model tersebut. Teknik penambangan data klasifikasi yang digunakan adalah *K-nearest neighbors*. Berikut hasil pengujian eksperimental yang telah dilakukan.

Random Rotation Perturbation

Pengujian teknik *Random Rotation Perturbation* untuk penambangan data klasifikasi dengan algoritma *K-nearest neighbors* akan dilakukan dengan dataset *diabetes* yang dapat dilihat 20 baris pertamanya pada Gambar 5.18. Dataset ini dirandomisasi dengan teknik *Random Rotation Perturbation* dan hasil randomisasi dapat dilihat pada Gambar 5.19. Dengan kedua dataset tersebut, dilakukan penambangan data terhadap kedua dataset tersebut dan dihasilkan berbagai macam informasi yang dapat dibandingkan.

Pada dataset *diabetes*, jumlah kasus diabetes atau dengan kata lain distribusi label pada dataset ini dapat dilihat pada Gambar 5.28. Dataset *diabetes* asli dan dataset *diabetes* yang telah dirandomisasi tentunya memiliki distribusi label yang sama karena label tidak dirandomisasi sehingga tidak berubah.

Properti-properti pada dataset *diabetes* asli dapat dilihat pada Tabel 5.1 dan Tabel 5.2. Sementara untuk dataset *diabetes* yang telah dirandomisasi dapat dilihat pada Tabel 5.3 dan Tabel 5.4. Jika dilihat pada keempat tabel tersebut, seluruh properti pada dataset yang telah dirandomisasi mempunyai nilai yang berbeda kecuali jumlah baris (*count*) dan kolom label (*Outcome*). Hal ini menunjukkan selain nilai pada setiap data, teknik *Random Rotation Perturbation* juga mengacak bermacam properti data seperti rata-rata, standar deviasi, batas bawah dan batas atas nilai pada data.

Tabel 5.1: Properti-properti pada dataset *diabetes* asli

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin
count	768.000000	768.000000	768.000000	768.000000	768.000000
mean	3.845052	120.894531	69.105469	20.536458	79.799479
std	3.369578	31.972618	19.355807	15.952218	115.244002
min	0.000000	0.000000	0.000000	0.000000	0.000000
25%	1.000000	99.000000	62.000000	0.000000	0.000000
50%	3.000000	117.000000	72.000000	23.000000	30.500000
75%	6.000000	140.250000	80.000000	32.000000	127.250000
max	17.000000	199.000000	122.000000	99.000000	846.000000

Tabel 5.2: Properti-properti pada dataset *diabetes* asli

	BMI	DiabetesPedigreeFunction	Age	Outcome
count	768.000000	768.000000	768.000000	768.000000
mean	31.992578	0.471876	33.240885	0.348958
std	7.884160	0.331329	11.760232	0.476951
min	0.000000	0.078000	21.000000	0.000000
25%	27.300000	0.243750	24.000000	0.000000
50%	32.000000	0.372500	29.000000	0.000000
75%	36.600000	0.626250	41.000000	1.000000
max	67.100000	2.420000	81.000000	1.000000

Tabel 5.3: Properti-properti pada dataset *diabetes* yang telah dirandomisasi

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin
count	768.000000	768.000000	768.000000	768.000000	768.000000
mean	-125.296036	-46.928128	38.476869	8.970500	149.05599
std	65.385296	24.314318	48.951102	29.325358	64.087311
min	-521.889105	-130.730268	-21.449526	-40.186125	60.899387
25%	-156.640887	-61.988042	4.264140	-9.412375	109.984343
50%	-102.169672	-44.481520	21.426488	2.512069	127.946623
75%	-79.991828	-30.399185	61.569078	20.287410	171.374079
max	-43.084371	29.654467	349.345437	195.959060	583.395570

Tabel 5.4: Properti-properti pada dataset *diabetes* yang telah dirandomisasi

	BMI	DiabetesPedigreeFunction	Age	Outcome
count	768.000000	768.000000	768.000000	768.000000
mean	-103.101329	50.477502	-23.129061	0.348958
std	24.146554	35.533157	32.695887	0.476951
min	-176.332307	-199.209279	-74.182020	0.000000
25%	-116.872237	35.049918	-46.748389	0.000000
50%	-101.458852	58.320138	-28.463990	0.000000
75%	-86.757613	73.320823	-9.757361	1.000000
max	-22.156712	116.151884	183.302024	1.000000

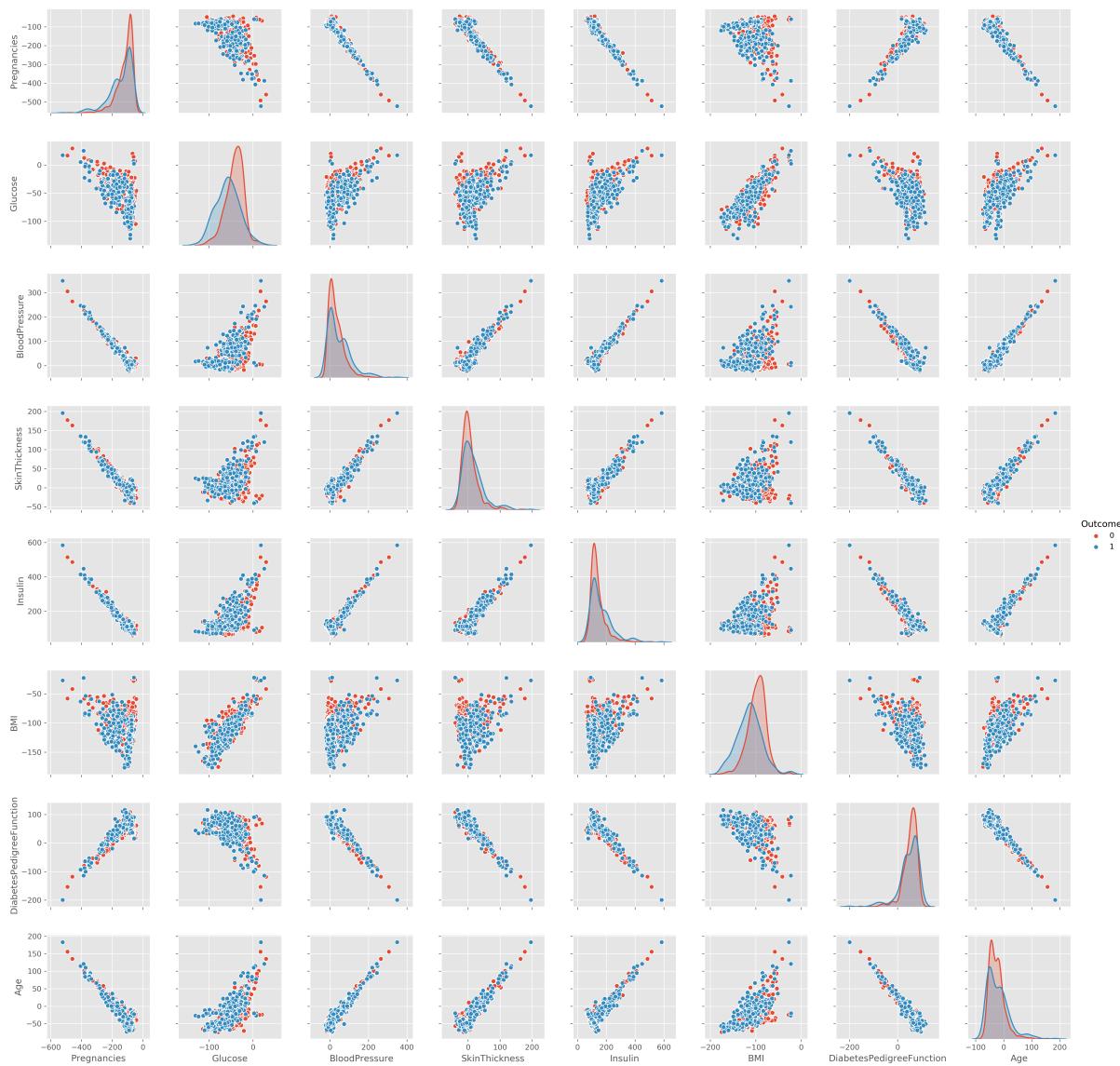


Gambar 5.29: *Scatter plot* antara seluruh fitur pada dataset *diabetes* yang asli

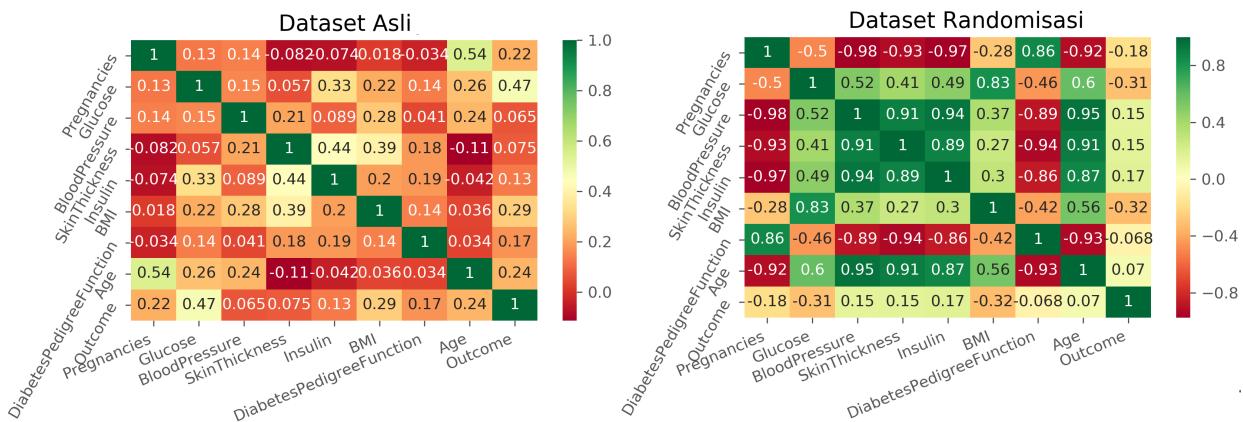
Pada Gambar 5.29 terdapat *scatter plot* antara satu fitur dengan fitur lainnya pada dataset asli. Sementara pada dataset yang telah dirandomisasi dapat dilihat pada Gambar 5.30. Dapat terlihat pada kedua gambar tersebut, lokasi pada titik-titik pada seluruh scatter plot berbeda. Hal ini menunjukkan teknik *Random Rotation Perturbation* berhasil mengacak dengan baik setiap nilai yang ada dengan rotasi sehingga visualisasi seperti ini akan terlihat acak karena dataset dirotasi pada dimensi yang lebih besar.

Pada Gambar 5.31 dapat dilihat korelasi antara satu fitur dengan fitur lainnya pada dataset asli dan dataset yang telah dirandomisasi. Apabila dibandingkan korelasi antar fitur pada dataset randomisasi sangat berbeda dengan dataset asli, bahkan tidak terlihat ada indikasi kesamaan atau ada properti pada dataset asli yang sama dengan dataset randomisasi. Hal ini menunjukkan teknik *Random Rotation Perturbation* mengacak setiap nilai pada data tanpa menjaga properti lain selain jarak Euclidean.

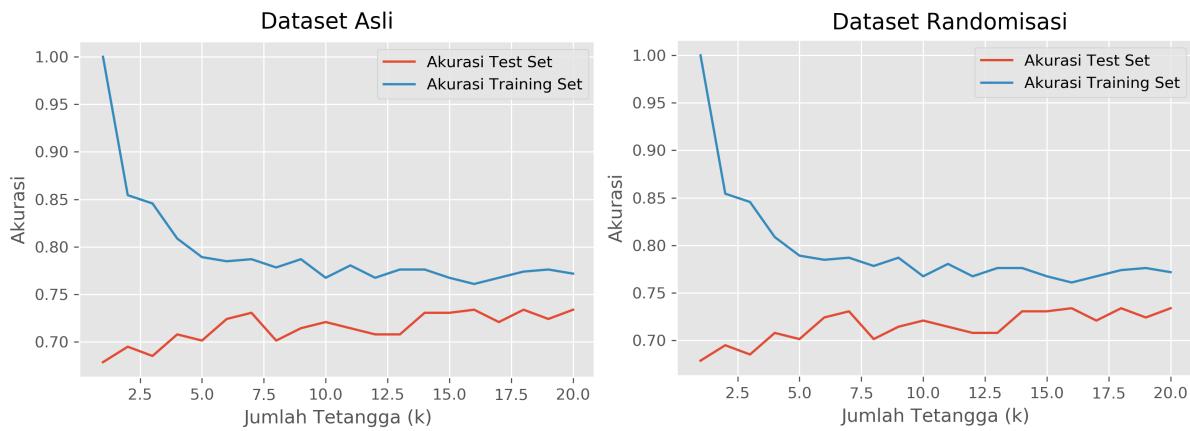
Pada Gambar 5.32 dapat dilihat grafik akurasi model *K-nearest Neighbors* dalam memprediksi label *training set* dan *test set* pada dataset asli dan dataset yang telah dirandomisasi. Apabila dibandingkan, kedua grafik tersebut memiliki nilai akurasi yang sama persis untuk setiap nilai *k*. Hal ini dapat menjadi bukti bahwa teknik *Random Rotation Perturbation* menjaga jarak Euclidean



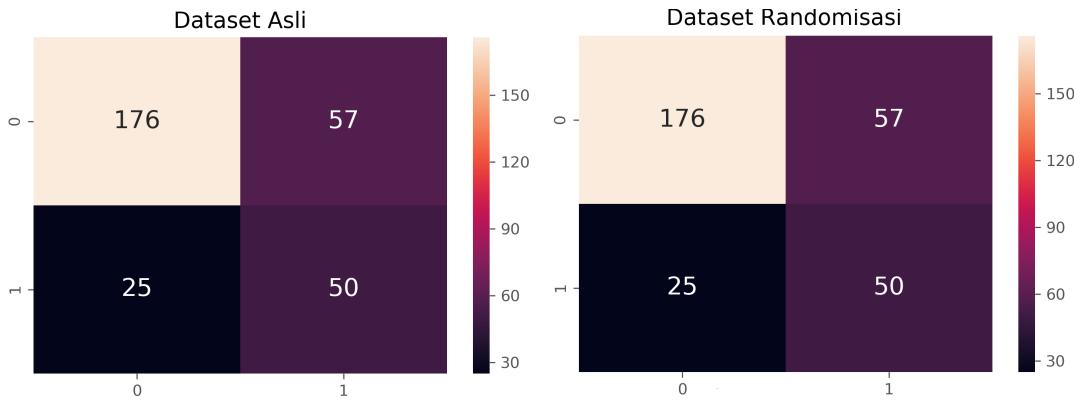
Gambar 5.30: *Scatter plot* antara seluruh fitur pada dataset *diabetes* yang telah dirandomisasi



Gambar 5.31: *Heatmap* korelasi antar fitur pada dataset *diabetes*



Gambar 5.32: Grafik akurasi model klasifikasi pada *training set* dan *test set* dataset *diabetes*



Gambar 5.33: *Confusion matrix* pada hasil prediksi *test set* dataset *diabetes*

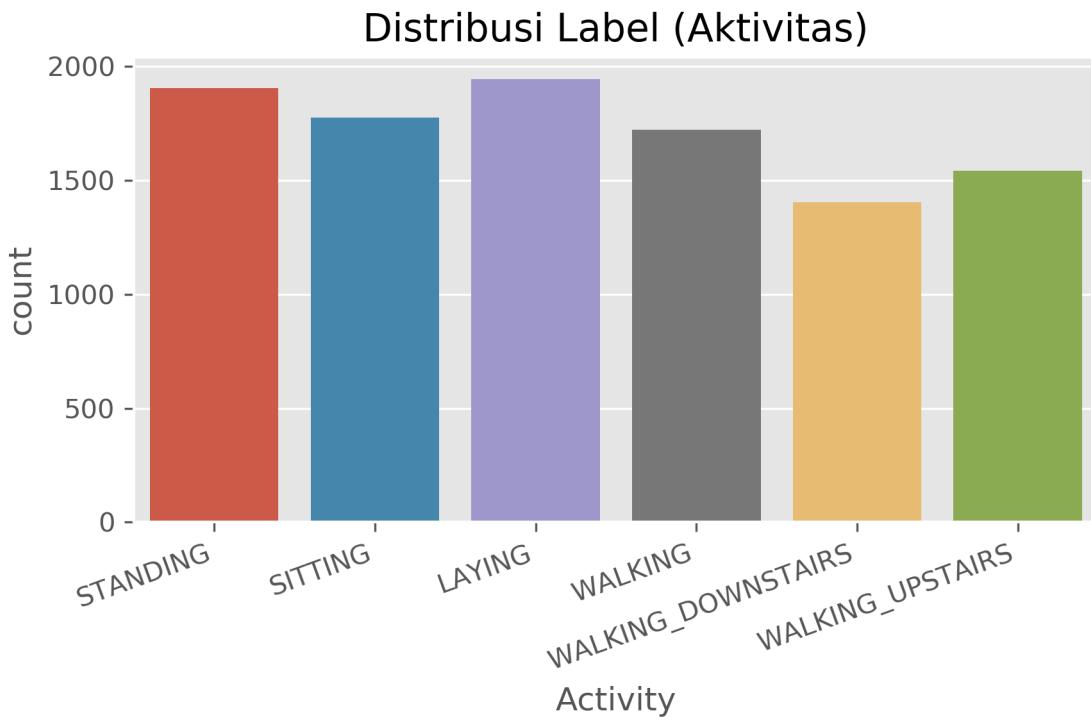
dengan sempurna, tidak ada perubahan sama sekali pada jarak Euclidean antara seluruh titik. Oleh karena itu model *K-nearest Neighbors* yang terbuat memiliki hasil yang sama persis.

Akurasi model untuk setiap nilai k pada dataset asli dan dataset yang telah dirandomisasi dapat dilihat masing-masing pada Listing 5.1 dan Listing 5.2. Akurasi tertinggi pada model *K-nearest Neighbors* dengan dataset asli adalah sebesar 0.7337662337662337 dengan nilai k sebesar 16. Nilai yang sama juga muncul pada dataset yang telah dirandomisasi. Waktu eksekusi yang dibutuhkan untuk melatih model klasifikasi dengan nilai k sebesar 16 memakai dataset asli adalah sebesar 0.0009965896606445312 detik dan waktu yang dibutuhkan untuk memprediksi *test set* sebesar 0.02593088150024414 detik. Sementara untuk dataset yang telah dirandomisasi membutuhkan waktu eksekusi untuk melatih model klasifikasi sebesar 0.0019948482513427734 detik dan waktu yang dibutuhkan untuk melakukan prediksi adalah sebesar 0.009980201721191406 detik. Hal ini menunjukkan tidak ada pengaruh yang signifikan terhadap durasi waktu eksekusi untuk melatih model dan memprediksi dengan dataset asli dan dataset yang telah dirandomisasi.

Pada Gambar 5.33 dapat dilihat *confusion matrix* pada model klasifikasi dengan nilai k sebesar 16 memakai *test set* pada dataset asli dan dataset yang telah dirandomisasi. Apabila dibandingkan antar keduanya, *confusion matrix* yang ada memiliki nilai yang sama persis. Hal ini juga menguatkan kesimpulan bahwa teknik *Random Rotation Perturbation* menjaga jarak Euclidean dengan sempurna.

Random Projection Perturbation

Pengujian teknik *Random Projection Perturbation* untuk penambangan data klasifikasi dengan algoritma *K-nearest neighbors* akan dilakukan dengan dataset *mobile_sensor* yang dapat dilihat 20 baris pertamanya pada Gambar 5.24. Dataset ini dirandomisasi dengan teknik *Random Projection*



Gambar 5.34: Histogram distribusi label dataset *mobile_sensor*

Tabel 5.5: Properti-properti pada dataset *mobile_sensor* asli

	tBodyAcc-mean()-X	tBodyAcc-mean()-Y	angle(Y,gravityMean)	angle(Z,gravityMean)
count	10299.000000	10299.000000	10299.000000	10299.000000
mean	0.274347	-0.017743	0.063255	-0.054284
std	0.067628	0.037128	0.305468	0.268898
min	-1.000000	-1.000000	-1.000000	-1.000000
25%	0.262625	-0.024902	0.002151	-0.131880
50%	0.277174	-0.017162	0.182028	-0.003882
75%	0.288354	-0.010625	0.250790	0.102970
max	1.000000	1.000000	1.000000	1.000000

Perturbation dan hasil randomisasi dapat dilihat pada Gambar 5.25. Dengan kedua dataset tersebut, dilakukan penambangan data terhadap kedua dataset tersebut dan dihasilkan berbagai macam informasi yang dapat dibandingkan.

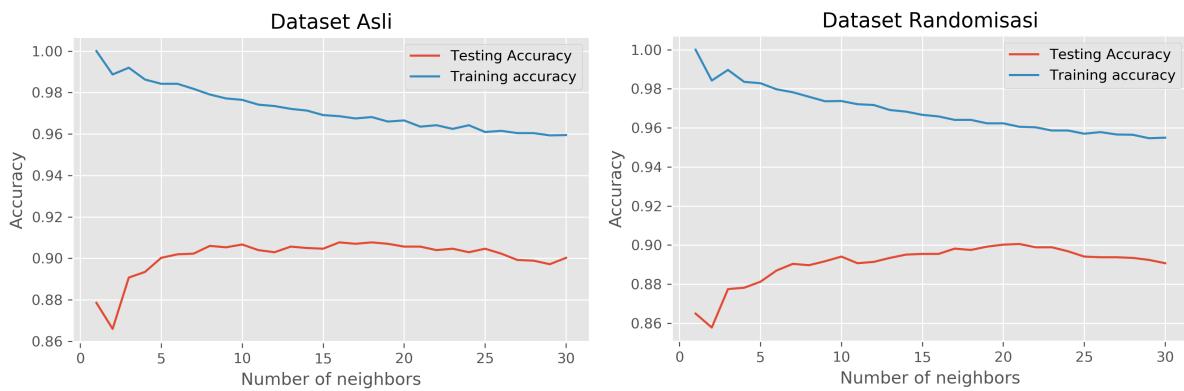
Pada dataset *mobile_sensor*, jumlah orang yang melakukan aktivitas tertentu atau dengan kata lain distribusi label pada dataset ini dapat dilihat pada Gambar 5.34. Dataset *mobile_sensor* asli dan dataset *mobile_sensor* yang telah dirandomisasi tentunya memiliki distribusi label yang sama karena label tidak dirandomisasi sehingga tidak berubah.

Properti-properti pada dataset *mobile_sensor* asli dapat dilihat pada Tabel 5.6. Sementara untuk dataset *mobile_sensor* yang telah dirandomisasi dapat dilihat pada Tabel 5.6. Jika dilihat pada kedua tabel tersebut, seluruh properti pada dataset yang telah dirandomisasi mempunyai nilai yang berbeda kecuali jumlah baris (*count*). Hal ini menunjukkan selain nilai pada setiap data, teknik *Random Projection Perturbation* juga mengacak bermacam properti data seperti rata-rata, standar deviasi, batas bawah dan batas atas nilai pada data.

Pada Gambar 5.35 dapat dilihat grafik akurasi model *K-nearest Neighbors* dalam memprediksi label *training set* dan *test set* pada dataset asli dan dataset yang telah dirandomisasi. Apabila

Tabel 5.6: Properti-properti pada dataset *mobile_sensor* yang telah dirandomisasi

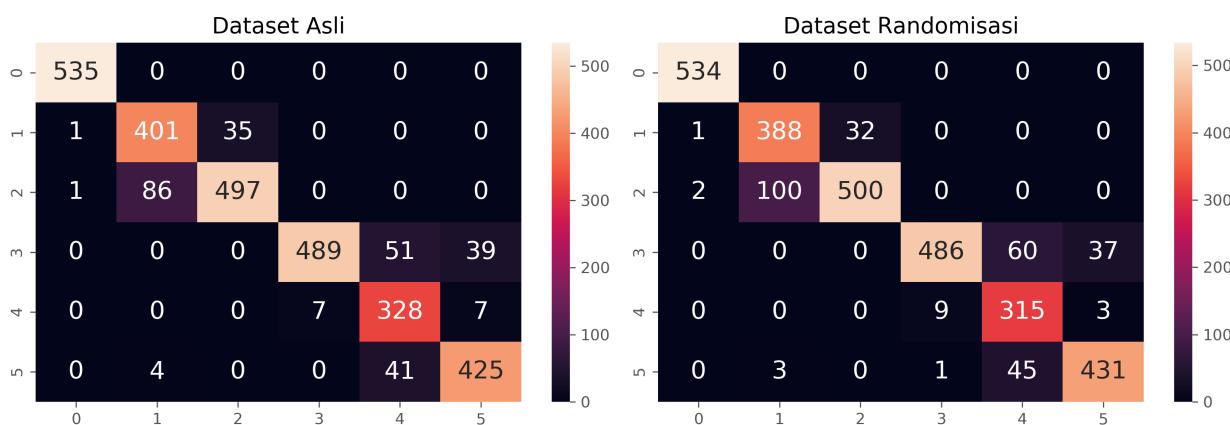
	0	1	425	426
count	10299.000000	10299.000000	10299.000000	10299.000000
mean	0.136942	-0.289681	0.173610	0.353280
std	0.528322	0.266849	0.214992	0.475443
min	-1.433025	-1.206428	-1.437424	-0.976616
25%	-0.340174	-0.482918	0.033424	-0.049233
50%	0.308314	-0.270461	0.179048	0.322172
75%	0.582374	-0.097266	0.325611	0.688616
max	1.283867	0.726498	0.978304	1.786145

Gambar 5.35: Grafik akurasi model klasifikasi pada *training set* dan *test set* dataset *mobile_sensor*

dibandingkan, kedua grafik tersebut memiliki nilai akurasi yang mirip. Walaupun begitu, teknik *Random Projection Perturbation* tidak menjamin jarak Euclidean terjaga dengan sempurna maka ada sedikit perbedaan yang lumayan terlihat khususnya pada akurasi *test set* dan nilai k yang memiliki akurasi tertinggi. Tetapi perbedaan nilai akurasinya masih mirip dengan dataset asli. Hal ini dapat menjadi bukti bahwa teknik *Random Projection Perturbation* menjaga jarak Euclidean dengan baik dan terkontrol *error*-nya sesuai yang pengguna inginkan. Oleh karena itu model *K-nearest Neighbors* yang terbuat memiliki hasil yang sangat mirip.

Akurasi model untuk setiap nilai k pada dataset asli dan dataset yang telah dirandomisasi dapat dilihat masing-masing pada Listing 5.3 dan Listing 5.4. Akurasi tertinggi pada model *K-nearest Neighbors* dengan dataset asli adalah sebesar 0.9077027485578555 dengan nilai k sebesar 16. Sementara pada dataset yang telah dirandomisasi, akurasi tertingginya adalah sebesar 0.9005768578215134 dengan nilai k sebesar 21. Waktu eksekusi yang dibutuhkan untuk melatih model klasifikasi dengan nilai k sebesar 16 memakai dataset asli adalah sebesar 0.2872335910797119 detik dan waktu yang dibutuhkan untuk memprediksi *test set* sebesar 17.754546642303467 detik. Sementara untuk dataset yang telah dirandomisasi membutuhkan waktu eksekusi untuk melatih model klasifikasi dengan nilai k sebesar 21 adalah 0.5630350112915039 detik dan waktu yang dibutuhkan untuk melakukan prediksi adalah sebesar 12.86760687828064 detik. Hal ini menunjukkan ada pengaruh yang signifikan terhadap durasi waktu eksekusi untuk melatih model dan memprediksi dengan dataset asli dan dataset yang telah dirandomisasi. Pada dataset yang telah dirandomisasi waktu prediksi lebih cepat dikarenakan fitur-fitur yang ada pun lebih sedikit daripada dataset asli.

Pada Gambar 5.36 dapat dilihat *confusion matrix* pada model klasifikasi dengan nilai k sebesar 16 memakai *test set* pada dataset asli dan nilai k sebesar 21 memakai dataset yang telah dirandomisasi. Apabila dibandingkan antar keduanya, *confusion matrix* yang ada memiliki nilai yang sangat mirip. Oleh karena itu, akurasi modelnya juga mempunyai nilai yang sangat mirip. Hal ini juga menguatkan kesimpulan bahwa teknik *Random Rotation Perturbation* menjaga jarak Euclidean dengan baik dan *error*-nya terkontrol sesuai yang pengguna inginkan.



Gambar 5.36: *Confusion matrix* pada hasil prediksi *test set* dataset *mobile_sensor*

Perbandingan

Berdasarkan pengujian sebelumnya, dapat disimpulkan bahwa teknik *Random Rotation Perturbation* mengacak data dengan baik dan menjamin jarak Euclidean pada dataset terjaga dengan sempurna. Teknik *Random Projection Perturbation* juga mengacak data dengan baik tetapi hanya menjamin jarak Euclidean dapat terjaga dengan *error* yang ditentukan oleh pengguna dan teknik ini hanya dapat diterapkan terhadap dataset yang relatif besar, mempunyai dimensi yang cukup besar untuk diproyeksi ke dimensi yang lebih kecil. Oleh karena itu, teknik *Random Rotation Perturbation* memiliki keunggulan yaitu dataset apapun dapat dirandomisasi dengan teknik ini, sementara teknik *Random Projection Perturbation* hanya bisa menerapkan dataset yang relatif cukup besar.

Pengujian untuk membandingkan juga akan dilakukan dengan memakai dataset *mobile_sensor*. Dataset akan dirandomisasi dengan kedua teknik tersebut dan dibandingkan model klasifikasi yang dilatih oleh kedua dataset yang telah dirandomisasi. Pengujian ini bertujuan untuk melihat

5.3.2 Penambangan Data *Clustering*

Pengujian dengan penambangan data *clustering* akan berpusat pada pembuatan model *clustering* dengan dataset asli dan dataset yang telah dirandomisasi dan membandingkan kualitas model tersebut. Teknik penambangan data *clustering* yang digunakan adalah *K-means*. Berikut hasil pengujian eksperimental yang telah dilakukan.

Random Rotation Perturbation

Pengujian teknik *Random Rotation Perturbation* untuk penambangan data *clustering* dengan algoritma *K-means* akan dilakukan dengan dataset *mall_customers* yang dapat dilihat 20 baris pertamanya pada Gambar 5.20. Dataset ini dirandomisasi dengan teknik *Random Rotation Perturbation* dan hasil randomisasi dapat dilihat pada Gambar 5.21. Dengan kedua dataset tersebut, dilakukan penambangan data terhadap kedua dataset tersebut dan dihasilkan berbagai macam informasi yang dapat dibandingkan.

Properti-properti pada dataset *mall_customers* asli dapat dilihat pada Tabel 5.7. Sementara untuk dataset *diabetes* yang telah dirandomisasi dapat dilihat pada Tabel 5.8. Jika dilihat pada kedua tabel tersebut, seluruh properti pada dataset yang telah dirandomisasi mempunyai nilai yang berbeda kecuali jumlah baris (*count*). Hal ini menunjukkan selain nilai pada setiap data, teknik *Random Rotation Perturbation* juga mengacak bermacam properti data seperti rata-rata, standar deviasi, batas bawah dan batas atas nilai pada data.

Pada Gambar 5.37 terdapat *scatter plot* antara satu fitur dengan fitur lainnya pada dataset asli dan dataset yang telah dirandomisasi. Dapat terlihat pada kedua *scatter plot* tersebut, lokasi pada titik-titik pada seluruh scatter plot berbeda. Hal ini menunjukkan teknik *Random Rotation*

Tabel 5.7: Properti-properti pada dataset *mall_customers* asli

	Age	Annual Income (k\$)	Spending Score (1-100)
count	200.000000	200.000000	200.000000
mean	38.850000	60.560000	50.200000
std	13.969007	26.264721	25.823522
min	18.000000	15.000000	1.000000
25%	28.750000	41.500000	34.750000
50%	36.000000	61.500000	50.000000
75%	49.000000	78.000000	73.000000
max	70.000000	137.000000	99.000000

Tabel 5.8: Properti-properti pada dataset *mall_customers* yang telah dirandomisasi

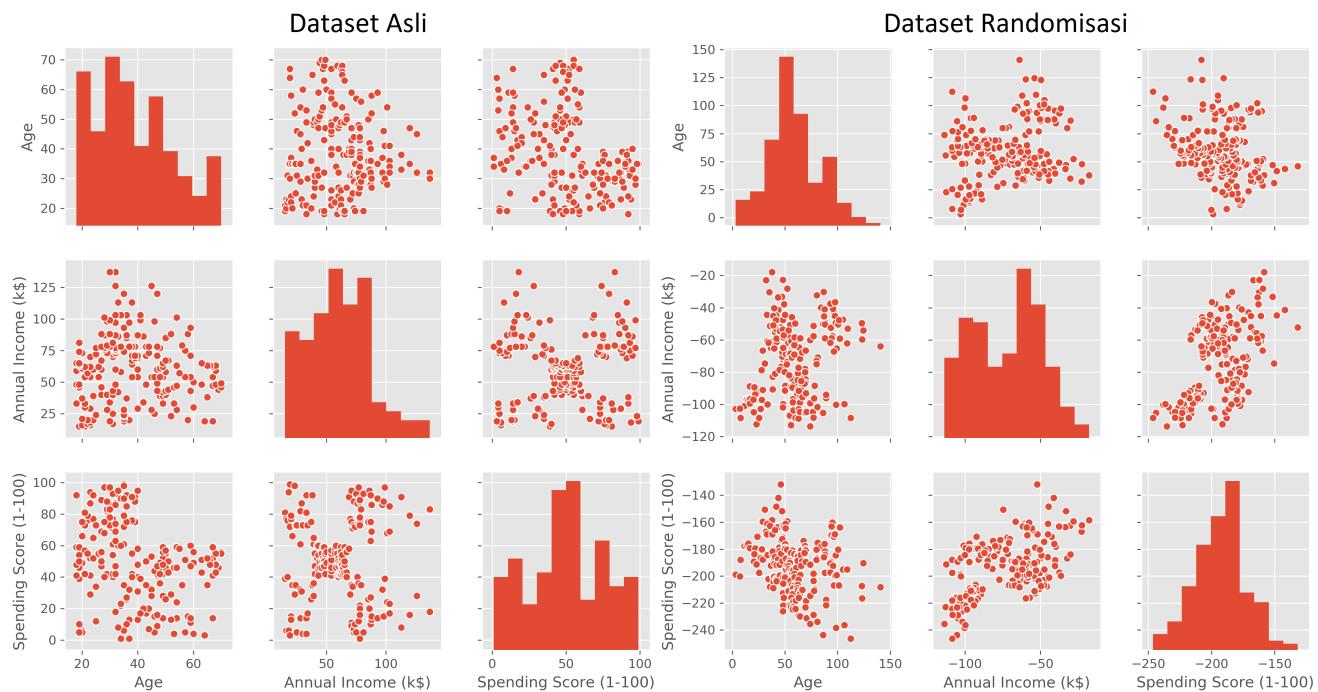
	Age	Annual Income (k\$)	Spending Score (1-100)
count	200.000000	200.000000	200.000000
mean	59.295607	-72.113582	-192.830387
std	24.832342	22.891719	20.276760
min	3.246983	-113.632925	-246.309601
25%	43.595492	-92.541715	-206.989917
50%	56.497882	-68.628508	-190.676923
75%	73.795114	-53.968669	-180.098255
max	140.690958	-17.887145	-131.806548

Perturbation berhasil mengacak dengan baik setiap nilai yang ada dengan rotasi sehingga visualisasi seperti ini akan terlihat acak karena dataset dirotasi pada dimensi yang lebih besar.

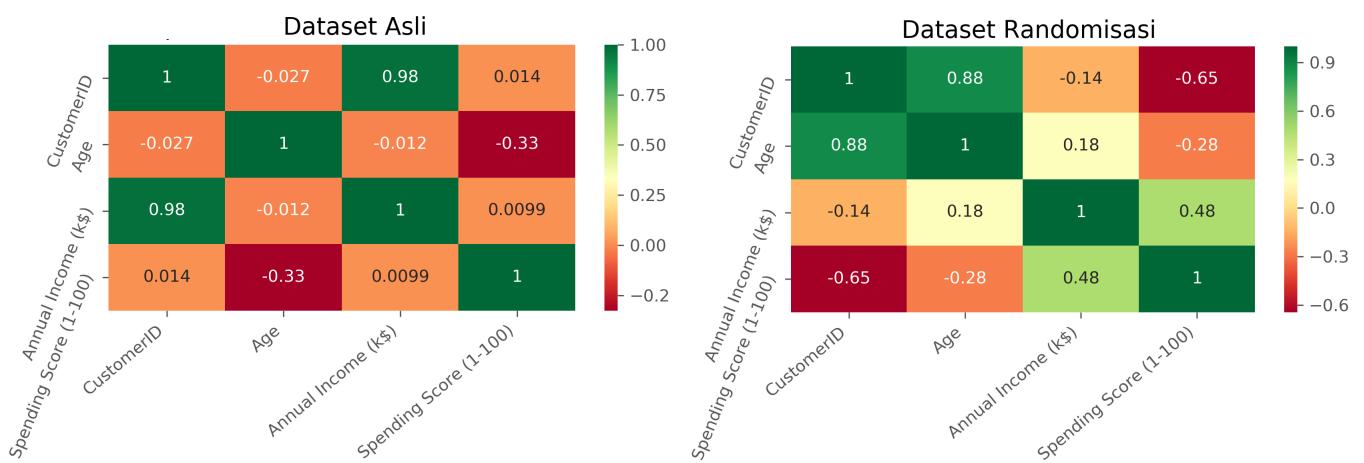
Pada Gambar 5.38 dapat dilihat korelasi antara satu fitur dengan fitur lainnya pada dataset asli dan dataset yang telah dirandomisasi. Apabila dibandingkan korelasi antar fitur pada dataset randomisasi sangat berbeda dengan dataset asli, bahkan tidak terlihat ada indikasi kesamaan atau ada properti pada dataset asli yang sama dengan dataset randomisasi. Hal ini menunjukkan teknik *Random Rotation Perturbation* mengacak setiap nilai pada data tanpa menjaga properti lain selain jarak Euclidean.

Dalam menentukan jumlah *cluster* atau nilai *k* yang terbaik untuk membuat model *clustering* dengan algoritma *k-means* perlu ada metode untuk menentukan nilai tersebut. Metode Elbow menjadi salah satu metode yang digunakan untuk menentukan nilai *k* yang terbaik untuk dipakai. Pada Gambar 5.39 terdapat grafik distorsi untuk menggunakan metode Elbow pada dataset asli dan dataset yang telah dirandomisasi. Terlihat nilai *k* 5, 6, dan 7 adalah kandidat yang terbaik untuk menjadi nilai *k* yang dipakai pada dataset asli maupun dataset randomisasi. Pada kedua dataset tersebut grafik distorsinya terlihat sama persis dikarenakan teknik *Random Rotation Perturbation* menjaga jarak Euclidean dengan sempurna. Dalam menentukan nilai *k* yang terbaik antara ketiga nilai tersebut, *Sillhouette Score* dapat digunakan untuk metode alternatif untuk menentukan nilai *k* yang terbaik.

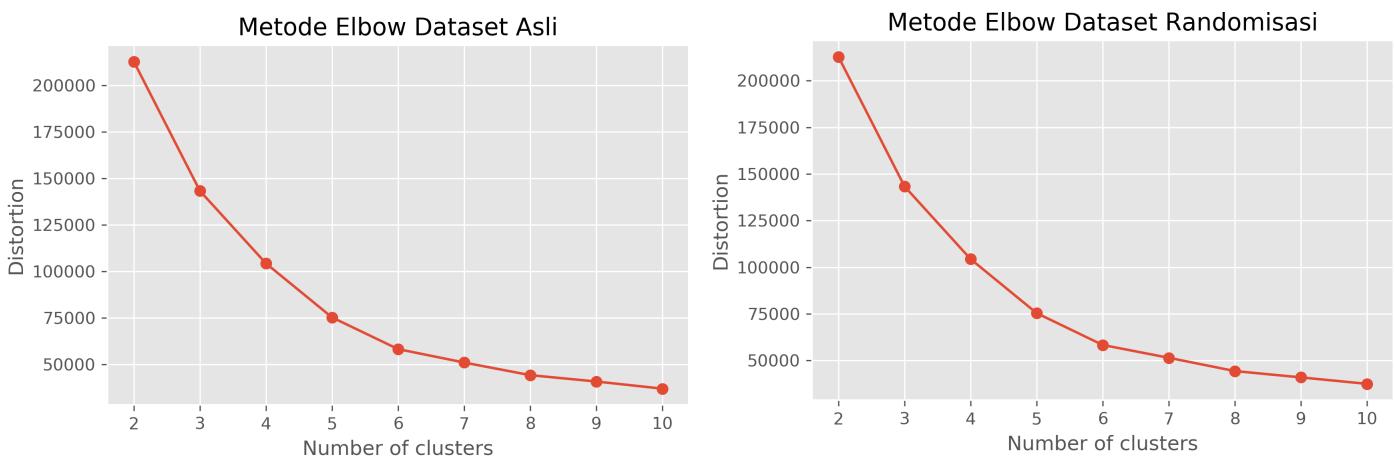
Pada Gambar 5.40 terdapat grafik *Sillhouette Score* dari dataset *mall_customers* asli dan yang telah dirandomisasi. Dapat terlihat kedua grafik *Sillhouette Score* terlihat sama persis dan dapat dilihat nilai-nilai pada setiap *k* tersebut di Listing 5.5 dan Listing 5.6 ada sedikit perbedaan yang tidak terlalu signifikan. Hal ini mungkin dikarenakan oleh nilai pada setiap data yang berbeda dan mempengaruhi sedikit algoritma pada *Sillhouette Score*. Dapat terlihat *Sillhouette Score* pada nilai *k* 6 adalah nilai paling besar yaitu 0.4523443947724053 pada dataset asli dan 0.4523443947780976 pada dataset yang telah dirandomisasi. Hal ini menunjukkan teknik *Random Rotation Perturbation* tidak mempengaruhi secara signifikan nilai *Sillhouette Score* pada setiap *k*.



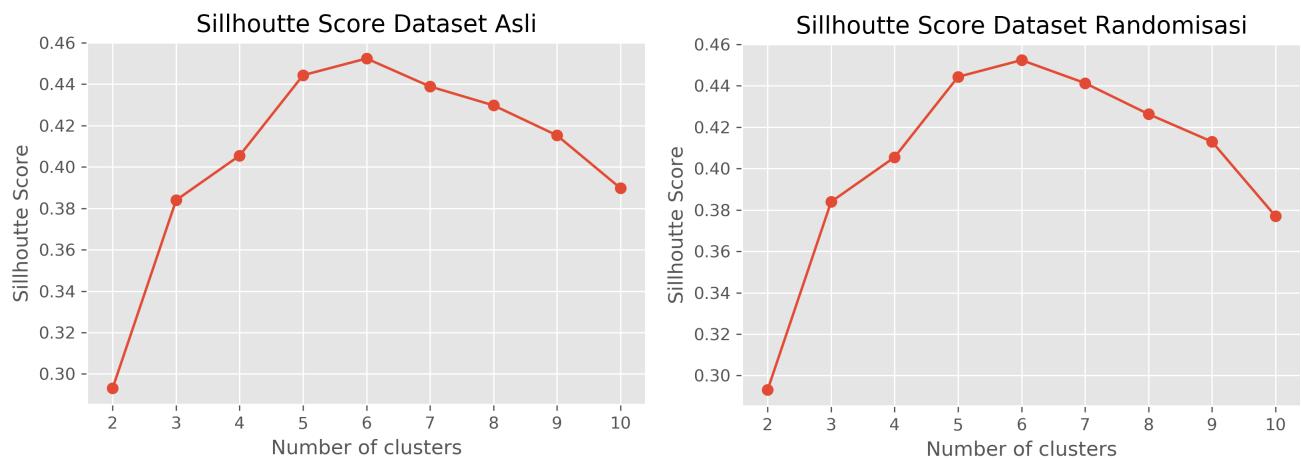
Gambar 5.37: *Scatter plot* antara seluruh fitur pada dataset *mall_customers* yang asli



Gambar 5.38: *Heatmap* korelasi antar fitur pada dataset *mall_customers*



Gambar 5.39: Grafik distorsi untuk menggunakan metode Elbow untuk mencari nilai *k* terbaik

Gambar 5.40: Grafik *Sillhouette Score* model *clustering* pada dataset *mall_customers*

Listing 5.5: Sillhouette Score Dataset Asli

Sillhouette Score setiap K pada dataset asli :

```

2: 0.293166070535953
3: 0.3839349967742105
4: 0.40546302077733304
5: 0.44504314844253573
6: 0.4523443947724053
7: 0.43978902692261157
8: 0.42790288922594905
9: 0.4137641526186506
10: 0.3750147687842441

```

Listing 5.6: Sillhouette Score Dataset Randomisasi

Sillhouette Score setiap K pada dataset randomisasi :

```

2: 0.29316607053507854
3: 0.383934996807901
4: 0.40546302082487856
5: 0.44428597567883826
6: 0.4523443947780976
7: 0.44128075766857394
8: 0.42815090435529995
9: 0.3861502477348431
10: 0.3897532214988177

```

TODO VISUALISASI CLUSTER DAN WAKTU EKSEKUSI

Random Projection Perturbation

Perbandingan

DAFTAR REFERENSI

- [1] Oliveira, S. R. M. dan Zaïane, O. R. (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] MENDES, R. dan VILELA, J. P. (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] Keyvanpour, M. dan Moradi, S. S. (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] Chen, K. dan Liu, L. (2005) A random rotation perturbation approach to privacy preserving data classification. Technical Report GIT-CC-05-12. Georgia Institute of Technology, Georgia.
- [7] Agrawal, R. dan Srikant, R. (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, 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.
- [11] Ella Bingham, H. M. (2001) Random projection in dimensionality reduction: Applications to image and text data. In *Proceedings of the ACM SIGKDD*, **7**, 245–250.