SKRIPSI

ANALISIS DATA BERDIMENSI BESAR MENGGUNAKAN MODEL *RANDOM FOREST* DENGAN PENERAPAN METODE ANALISIS KOMPONEN UTAMA

ANNISA ALIANA KOSASIH

NPM: 6161901060

PROGRAM STUDI MATEMATIKA
FAKULTAS TEKNOLOGI INFORMASI DAN SAINS
UNIVERSITAS KATOLIK PARAHYANGAN

2024

FINAL PROJECT

ANALYSIS OF HIGH DIMENSIONAL DATA USING RANDOM FOREST MODEL WITH THE IMPLEMENTATION OF PRINCIPAL COMPONENT ANALYSIS METHOD

ANNISA ALIANA KOSASIH

NPM: 6161901060

DEPARTMENT OF MATHEMATICS
FACULTY OF INFORMATION TECHNOLOGY AND SCIENCES
PARAHYANGAN CATHOLIC UNIVERSITY
2024

LEMBAR PENGESAHAN

ANALISIS DATA BERDIMENSI BESAR MENGGUNAKAN MODEL *RANDOM FOREST* DENGAN PENERAPAN METODE ANALISIS KOMPONEN UTAMA

Annisa Aliana Kosasih

NPM: 6161901060

Telah lulus ujian skripsi pada 18 Januari 2024 dengan penguji:
Farah Kristiani, Ph.D. dan Dr. Erwinna Chendra

Bandung, 31 Januari 2024

Menyetujui,

Pembimbing 1	**Pembimbing 2**
Felivia Kusnadi, M.Act.Sc.	**Robyn Irawan, M.Sc.**

Mengetahui,

Ketua Program Studi

Jonathan Hoseana, Ph.D.

PERNYATAAN

Dengan ini saya yang bertandatangan di bawah ini menyatakan bahwa skripsi dengan judul:

ANALISIS DATA BERDIMENSI BESAR MENGGUNAKAN MODEL *RANDOM FOREST* DENGAN PENERAPAN METODE ANALISIS KOMPONEN UTAMA

adalah benar-benar karya saya sendiri, dan saya tidak melakukan penjiplakan atau pengutipan dengan cara-cara yang tidak sesuai dengan etika keilmuan yang berlaku dalam masyarakat keilmuan.

Atas pernyataan ini, saya siap menanggung segala risiko dan sanksi yang dijatuhkan kepada saya, apabila di kemudian hari ditemukan adanya pelanggaran terhadap etika keilmuan dalam karya saya, atau jika ada tuntutan formal atau non-formal dari pihak lain berkaitan dengan keaslian karya saya ini.

Dinyatakan di Bandung,
31 Januari 2024

Annisa Aliana Kosasih
NPM: 6161901060

ABSTRAK

Data banyak berperan dalam kehidupan manusia serta memiliki beragam jenis, karakteristik, dan teknik pengumpulannya. Salah satu karakteristik data yang kerap kali dijumpai yakni data yang memiliki dimensi besar. Karakteristik tersebut terkadang dapat menjadi tantangan dalam proses analisis data. Salah satu pilihan untuk menangani hal tersebut yakni memanfaatkan metode dalam *machine learning*. Penelitian ini bertujuan untuk menerapkan dan melihat pengaruh suatu metode *machine learning* yang terpilih untuk penelitian, dalam pembentukan model prediksi, dengan melibatkan proses penanganan karakteristik data berdimensi besar. Model prediksi yang dipilih ialah *Random Forest*. Metode yang digunakan dalam mendukung pembentukan model prediksi tersebut antara lain adalah analisis komponen utama (*Principal Component Analysis*). Fokus utama dalam penelitian yakni melihat pengaruh dari penerapan reduksi dimensi data dengan metode analisis komponen utama, untuk suatu model prediksi jenis klasifikasi dengan *Random Forest*. Data yang digunakan dalam penelitian memiliki dimensi yang besar. Pada penelitian ini, digunakan himpunan data dengan topik kanker payudara dan kebangkrutan perusahaan. Harapannya adalah dapat menambah literasi terkait pengaruh penerapan metode yang dilakukan terhadap suatu model prediksi *Random Forest*. Hasil pembahasan skripsi ini menunjukkan bahwa penerapan reduksi dimensi dengan metode analisis komponen utama tidak menunjukkan hasil yang lebih signifikan dibandingkan model tanpa adanya reduksi dimensi. Artinya, dalam hal ini ternyata model *Random Forest* sudah cukup untuk mengolah data. Meskipun demikian, hal yang dapat dipastikan yakni waktu pelatihan data (*training time*) lebih cepat dan ukuran dimensi data yang menjadi lebih kecil. Selain itu, di dalam analisis hasil *Random Forest* juga dikaji *variable importance* model yang memberikan hasil bahwa adanya penerapan analisis komponen utama tidak mengganggu esensi informasi yang dimiliki data asli meskipun dimensi data lebih kecil. Penelitian ini berkontribusi pada literatur mengenai dampak metode yang diajukan terhadap model *Random Forest*, serta memberikan wawasan tentang efektivitasnya dalam menangani data berdimensi besar.

Kata-kata kunci: Data Dimensi Besar; Analisis Komponen Utama; Model Prediksi; *Random Forest*.

ABSTRACT

Data has significant role in human life and comes in various types, characteristics, and collection techniques. Characteristics of data that is often found is data that has large dimensions. This characteristics could sometimes pose a challenge in data analysis. One of the solution to address these challenge is using methods in machine learning. The primary objective of this research is to apply and see the influence of a machine learning method selected for this research, in forming a prediction model, particularly when dealing with high-dimensional data. The chosen model is Random Forest. Methods employed to support the formation of this predictive model is Principal Component Analysis. The main focus of this research is on examining the influence of dimensionality reduction through Principal Component Analysis on a classification model using Random Forest. The research data encompasses large dimensions, involving datasets related to breast cancer and corporate bankruptcy case. The aim is to enhance literacy regarding the impact of applied techniques on a Random Forest predictive model. The results of this thesis indicates that the implemetation of dimensionality reduction through Principal Component Analysis did not yield significantly improved results. In this context, it appears that the Random Forest model alone is sufficient for processing the data. However, notable outcomes include faster training times and reduced data dimensions. Additionally, in the analysis of the Random Forest results, the variable importance of the model was examined. The result suggest that the implementation of principal component analysis does not disrupt the essential information inherent in the original data, even though the data dimensions are reduced. This research contributes to the literature on the impact of this method on Random Forest predictive models, offering insights into its effectiveness in handling high-dimensional data.

Keywords: High-dimensional Data; Principal Component Analysis; Predictive Model; Random Forest.

Percaya bahwa setelah kesulitan akan selalu ada kemudahan meskipun waktunya belum dapat diketahui. Setidaknya, dengan hal tersebut, 'semangat' akan selalu membersamai perjalanan hidup.

KATA PENGANTAR

Segala puji serta syukur penulis panjatkan kepada Tuhan yang Maha Esa dan atas dukungan serta doa dari orang-orang terkasih, skripsi ini dapat dirampungkan dengan baik. Skripsi dengan judul "Analisis Data Berdimensi Besar Menggunakan Model *Random Forest* dengan Penerapan Analisis Komponen Utama" disusun untuk memenuhi salah satu syarat untuk menyelesaikan studi Strata-1 Program Studi Matematika, Fakultas Teknologi Informasi dan Sains, Universitas Katolik Parahyangan, Bandung. Adapun harapannya skripsi ini dapat memberi manfaat bagi pembaca.

Banyak tantangan serta rintangan dalam penulisan skripsi ini. Oleh karenanya, penulis ingin mengucapkan terima kasih secara khusus kepada:

1. Ibu Felivia Kusnadi, MActSc, ASA, ASAI dan Bapak Robyn Irawan, M.Sc selaku dosen pembimbing yang telah berbagi ilmu, bantuan dan motivasi, serta senantiasa memberikan waktu, arahan, dan bimbingan yang bermanfaat dalam penyelesaian skripsi ini.
2. Ibu Farah Kristiani, Ph.D dan Dr. Erwinna Chendra selaku dosen penguji yang telah memberikan saran serta kritik untuk skripsi ini menjadi lebih baik.
3. Kedua orang tua penulis, mamah dan papah, yang senantiasa selalu memberikan dukungan, semangat, serta doa terbaik untuk kelancaran proses perkuliahan hingga penulisan skripsi ini.
4. Keluarga besar penulis yang selalu memberikan semangat, dukungan, serta doa kepada penulis.
5. Rifva Putri Abie S. sebagai sahabat yang tak henti-hentinya dalam memberikan dukungan serta semangat penuh selama proses perkuliahan dan penulisan skripsi ini.
6. Sahabat dan teman tersayang, Grup Misteri dan Grup Stress Lintas Kampus, yang selalu menjadi tempat terbaik untuk berbagi cerita, memberi dukungan dan doa untuk penulis.
7. Teman-teman yang memberi semangat serta berjuang bersama selama proses perkuliahan, Antonius, Richard, Manzo, dan seluruh anggota Grup RT 19 RW 20.
8. Teman-teman Matematika 2019 dan keluarga besar UKM LISTRA yang sama-sama berjuang, berbagi ilmu dan pengalaman selama masa perkuliahan.
9. Seluruh dosen, tata usaha, pekarya, dan pihak lainnya yang tidak dapat disebutkan satu per satu. Terima kasih banyak atas semua bantuan dan dukungannya selama ini.

Penulis menyadari bahwa skripsi ini belum sempurna dan memiliki kekurangan. Penulis terbuka akan kritik maupun saran. Meskipun demikian, semoga skripsi ini dapat memberikan manfaat. Akhir kata, penulis ucapkan terima kasih.

Bandung, 31 Januari 2024

Penulis

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

BAB 1

PENDAHULUAN

1.1 Latar Belakang

Data merupakan salah satu hal yang sangat berpengaruh dalam kehidupan manusia saat ini. Berbagai bidang memerlukan data untuk membuat analisis dan kajian terhadap sebagian besar fenomena yang terjadi dalam kehidupan. Teknik pengumpulan data juga beragam tergantung pada bidang dan wilayah kajian masing-masing. Selain itu, data juga memiliki ukuran bervariasi mulai dari data berukuran kecil hingga besar. Oleh karena itu, pengetahuan akan pengolahan data menjadi hal yang dapat membawa kebaikan dan diperlukan untuk mengikuti perkembangan jaman.

Data memiliki beragam jenis, karakteristik, dan teknik pengumpulannya. Terdapat data yang memiliki banyak variabel numerik karena dikumpulkan berdasarkan penelitian kuantitatif, sementara yang lain bersifat kategorik karena dikumpulkan melalui kuesioner, ada pula data yang memiliki sampel observasi terbatas karena hanya dihimpun dalam lingkup kecil (misalnya siswa dalam 1 kelas saja), dan lain sebagainya. Semuanya tergantung pada situasi dan kondisi asal-usul data tersebut diperoleh. Terkadang, karakteristik data yang ditemui dapat menjadi tantangan dalam pengolahan dan analisis data. Tidak semua data yang dijumpai mudah untuk diolah dan dianalisis.

Salah satu karakteristik yang kerap kali dijumpai adalah data berdimensi besar. Data berdimesi besar mengacu pada kondisi data di mana proses pendataannya melibatkan banyak aspek atau variabel bebas yang dievaluasi dalam data. Karakteristik semacam ini juga dapat menimbulkan tantangan dalam pengolahan dan analisis data. Tantangan yang dapat muncul contohnya seperti waktu yang diperlukan untuk pemrosesan data lebih lama, meningkatkan kesulitan dalam visualisasi dan analisis, kebutuhan akan memori penyimpanan yang besar, dan lain sebagainya. Oleh karenanya, dibutuhkan metode untuk menangani tantangan tersebut.

Beberapa metode pembelajaran mesin (*machine learning*) banyak diajukan untuk mengatasi hal-hal tersebut. Terlebih lagi, apabila dihadapkan dengan suatu masalah prediksi atau pembuatan keputusan. Model-model *machine learning* yang terkenal antara lain seperti *Decision Tree*, *Random Forest*, *Boosting*, *Neural Network*, dan lain sebagainya. Berdasarkan literasi, model *Random Forest* adalah pilihan yang baik saat berhadapan dengan data yang memiliki banyak variabel prediktor [1]. Selain itu, ditemukan beberapa penelitian terdahulu yang melakukan pra-pemrosesan data sebelum model prediksi memanfaatkan metode reduksi dimensi dalam *machine learning*. Metode analisis komponen utama (*Principal Component Analysis*) merupakan salah satu strategi efektif untuk mereduksi dimensi data. Selain itu, berdasarkan beberapa penelitian terdahulu, penerapan metode analisis komponen utama memiliki tingkat akurasi yang baik dalam model-model prediktif, contohnya seperti pada penelitian Zhu, dkk [2].

Dalam penelitian yang akan dilakukan, dimanfaatkan sebanyak dua alat atau metode dalam *machine learning*. Metode pertama yakni metode yang diajukan untuk digunakan dalam mereduksi dimensi data yaitu analisis komponen utama. Selain itu, digunakan juga model *Random Forest* untuk membuat prediksi atau keputusan jenis klasifikasi dalam penelitian ini. Dalam mengkaji hasil *Random Forest* ini juga disajikan analisis *variable importance* yang merupakan alat untuk melihat seberapa besar pengaruh setiap variabel bebas terhadap hasil prediksi dari modelnya. Hal ini penting untuk dilakukan karena ingin mengevaluasi kinerja reduksi dimensi terhadap hasil prediksi, serta menilai potensi pengaruh signifikan atau gangguan yang mungkin timbul dari proses tersebut. Dalam penelitian, akan dibandingkan model dengan ataupun tanpa dilakukan reduksi dimensi data. Tujuan utama penelitian ini yaitu melihat serta mengevaluasi penerapan metode analisis komponen utama tersebut terhadap model prediksi dengan *Random Forest* pada data yang berdimensi besar.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan, penulis merumuskan masalah yang dikaji pada penelitian ini yaitu:

1. Bagaimana penerapan analisis komponen utama atau *Principal Component Analysis* (PCA) terhadap data penelitian yang berdimensi besar?
2. Bagaimana pengaruh metode analisis komponen utama untuk hasil prediksi menggunakan *Random Forest*?
3. Bagaimana pengaruh analisis komponen utama terhadap analisis *variable importance* dari model *Random Forest* yang terbentuk?

1.3 Tujuan

Tujuan dari penulisan skripsi ini adalah sebagai berikut:

1. Menerapkan analisis komponen utama pada data penelitian yang memiliki dimensi besar.
2. Melihat pengaruh penerapan metode analisis komponen utama terhadap kinerja prediksi menggunakan model *Random Forest*.
3. Melihat pengaruh metode analisis komponen utama terhadap analisis *variable importance* dari model *Random Forest* yang terbentuk.

1.4 *State of the Art*

Penelitian terdahulu memiliki fungsi untuk memperkaya analisis dan pembahasan serta sebagai pembanding dengan hal yang sedang diteliti. Dalam penelitian ini disertakan beberapa penelitian terdahulu terkait dengan hal yang sedang diteliti untuk memperjelas posisi penelitian penulis. Beberapa penelitian sebelumnya yang terkait antara lain:

1. Penelitian [3] dengan judul *Dimensionality Reduction Using PCA and K-Means Clustering for Breast Cancer Prediction* menggunakan metode analisis komponen utama dan *K-Means* untuk mereduksi sejumlah variabel yang memengaruhi kanker payudara sebelum dilakukan klasifikasi. Berdasarkan penelitian tahun 2018 tersebut, metode analisis komponen utama dan klasterisasi *K-Means* menampilkan hasil yang hampir sama baiknya dalam mereduksi dimensi data.

2. Penelitian [2] pada tahun 2019 dari *Lanzhou University of Technology* merupakan penelitian dengan menggabungkan metode analisis komponen utama, *K-Means*, dan regresi logistik di mana menghasilkan tingkat akurasi yang baik. Dalam hal ini analisis komponen utama meningkatkan algoritma klasterisasi *K-Means* yang selanjutnya memiliki pengaruh meningkatkan hasil prediksi dari model regresi logistik.

3. Sebuah penelitian selanjutnya [4] merupakan penelitian di India yang merupakan modifikasi dari penelitian [2] yang terdapat dalam buku prosiding *Innovations in Computer Science and Engineering: Procedings of 8th ICICSE*. Penelitian ini meneliti peningkatan hasil prediksi melalui *voting classifier* yang merupakan gabungan dari *Random Forest, Naive Bayes* dan *Multilayer Perceptron* dengan bantuan metode analisis komponen utama dan K-Means. Hasil yang diperoleh adalah tingkat akurasi prediksi yang lebih besar daripada penelitian [2].

Berdasarkan penelitian-penelitian terdahulu tersebut, dalam skripsi ini penulis meneliti topik analisis prediktif yang melibatkan reduksi dimensi dengan analisis komponen utama untuk suatu klasifikasi dengan *Random Forest* yang diterapkan pada data penelitian. Khususnya ketika berhadapan dengan karakteristik data berdimensi besar dan melihat pengaruhnya. Pembeda dengan penelitian sebelumnya yakni dalam penelitian ini diteliti dan dikaji lebih lanjut terkait penerapan metode analisis komponen utama terhadap suatu model prediksi, dalam hal ini dengan *Random Forest*. Hal ini karena ditemukan beberapa penelitian yang melakukan reduksi dimensi terlebih dahulu sebelum dilakukan metode *Random Forest*. Himpunan data yang digunakan dalam penelitian ialah data yang memiliki karakteristik berdimensi besar.

1.5 Sistematika Pembahasan

Skripsi ini terdiri dari 5 bab. Berikut ini merupakan sistematika pembahasan dari penelitian, tidak termasuk bab pertama, yaitu

Bab 2: Landasan Teori

Pada bab ini memuat penjelasan teori-teori dasar yang digunakan dalam penelitian. Teori yang dibahas yakni analisis komponen utama yang juga melibatkan dasar aljabar dan statistika, klasifikasi dengan *Random Forest* secara umum, beserta penjelasan metrik evaluasi.

Bab 3: Data dan Metodologi Penelitian

Bab ini memuat penjelasan data serta metode yang digunakan pada permasalahan dalam bab 4, yaitu analisis komponen utama dan *Random Forest*.

Bab 4: Hasil dan Pembahasan

Bab keempat ini memuat hasil penelitian serta pembahasannya memanfaatkan model yang disinggung pada Bab 3 terhadap masalah yang diteliti.

Bab 5: Kesimpulan dan Saran

Pada bab ini dimuat kesimpulan dari penelitian serta saran penelitian untuk dipertimbangkan dalam topik serupa kedepannya.