

**IMPLEMENTASI ALGORITMA K-MEANS DAN ALGORITMA NAÏVE
BAYES UNTUK IDENTIFIKASI TINGKAT RISIKO SAHAM PADA
SEKTOR BARANG KONSUMEN PRIMER**

SKRIPSI

ALFI DHARMAWAN

171401034



**PROGRAM STUDI S1 ILMU KOMPUTER
FAKULTAS ILMU KOMPUTER DAN TEKNOLOGI INFORMASI
UNIVERSITAS SUMATERA UTARA**

MEDAN

2024

**IMPLEMENTASI ALGORITMA K-MEANS DAN ALGORITMA NAÏVE
BAYES UNTUK IDENTIFIKASI TINGKAT RISIKO SAHAM PADA
SEKTOR BARANG KONSUMEN PRIMER**

SKRIPSI

Diajukan untuk melengkapi Tugas Akhir dan memenuhi syarat memperoleh
Ijazah Sarjana S-1 Ilmu Komputer

ALFI DHARMAWAN

171401034



**PROGRAM STUDI S1 ILMU KOMPUTER
FAKULTAS ILMU KOMPUTER DAN TEKNOLOGI INFORMASI
UNIVERSITAS SUMATERA UTARA
MEDAN
2024**

PERSETUJUAN

Judul : IMPLEMENTASI ALGORITMA K-MEANS DAN
ALGORITMA NAÏVE BAYES UNTUK
IDENTIFIKASI TINGKAT RISIKO SAHAM
PADA SEKTOR BARANG KONSUMEN PRIMER

Kategori : SKRIPSI

Nama : ALFI DHARMAWAN

Nomor Induk Mahasiswa : 171401034

Program Studi : SARJANA (S-1) ILMU KOMPUTER

Departemen : ILMU KOMPUTER

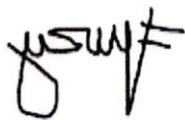
Fakultas : ILMU KOMPUTER DAN TEKNOLOGI
INFORMASI UNIVERSITASI SUMATERA
UTARA

Telah diuji dan dinyatakan lulus di Medan, 29 April 2024

Komisi Pembimbing :

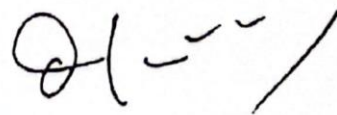
Pembimbing 2

Pembimbing 1



Fuzy Yustika Manik, S.Kom., M.Kom.

NIP. 198710152019032010

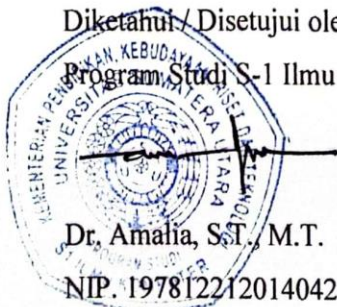


Dian Rachmawati S.Si., M.Kom.

NIP. 198307232009122004

Diketahui / Disetujui oleh

Program Studi S-1 Ilmu Komputer Ketua,



Dr. Amalia, S.T., M.T.

NIP. 197812212014042001

PERNYATAAN

IMPLEMENTASI ALGORITMA K-MEANS DAN ALGORITMA NAÏVE BAYES UNTUK IDENTIFIKASI TINGKAT RISIKO SAHAM PADA SEKTOR BARANG KONSUMEN PRIMER

SKRIPSI

Saya mengakui bahwa skripsi ini adalah hasil karya saya sendiri, kecuali beberapa kutipan dan ringkasan yang masing-masing telah disebutkan sumbernya.

Medan, 22 Maret 2024

Alfi Dharmawan

171401034

PENGHARGAAN

Alhamdulillah,

Segala puji bagi Allah Subhanahu Wa Ta'ala yang telah memberikan kekuatan dan kemudahan bagi penulis sehingga dapat menyelesaikan penyusunan tugas akhir dengan judul “Implementasi Algoritma K-Means dan Algoritma Naïve Bayes Untuk Identifikasi Tingkat Risiko Saham Pada Sektor Barang Konsumen Primer” sebagai syarat untuk memperoleh gelar Sarjana Komputer pada Program Studi S-1 Ilmu Komputer Fakultas Ilmu Komputer dan Teknologi Informasi Universitas Sumatera Utara.

Dengan segala kerendahan hati, penulis ingin menyampaikan rasa terima kasih dan syukur yang sebesar-besarnya kepada semua pihak yang terlibat baik secara langsung maupun tidak langsung dan pihak yang bertanggung jawab serta memberikan dukungan dan doa kepada penulis dalam proses pengerjaan tugas akhir ini. Terkhusus kepada orangtua, Ayahanda Heri Purnomo dan Ibunda Yunida serta kakak perempuan dan adik perempuan yang tak pernah luput dalam memberikan dukungan dan menguatkan penulis dalam menyelesaikan tugas akhir.

Tak lupa ucapan terima kasih juga penulis sampaikan kepada pihak terkait, antara lain :

1. Bapak Dr. Muryanto Amin, S.Sos., M.Si. selaku Rektor Universitas Sumatera Utara,
2. Ibu Dr. Maya Silvi Lydia, B.Sc., M.Sc. selaku Dekan Fakultas Ilmu Komputer dan Teknologi Informasi Universitas Sumatera Utara,
3. Ibu Dr. Amalia S.T., M.T. selaku Ketua Program Studi S-1 Ilmu Komputer, Fakultas Ilmu Komputer dan Teknologi Informasi Universitas Sumatera Utara,

4. Ibu Dian Rachmawati S.Si., M.Kom. selaku dosen pembimbing I yang selalu memberikan bimbingan, arahan, kritik, dan juga saran kepada penulis selama pengerjaan tugas akhir,
5. Ibu Fuzy Yustika Manik, S.Kom, M.Kom. selaku dosen pembimbing II yang senantiasa memberikan bimbingan, arahan, kritik, dan juga saran kepada penulis selama pengerjaan tugas akhir,
6. Seluruh dosen dan pegawai di Program Studi S-1 Ilmu komputer FasilkomTI Universitas Sumatera Utara,
7. Senior yang selalu memberikan dukungan dan motivasi, serta pengetahuan baru dan sebagai tempat bertukar pikiran kepada penulis, yakni, Arif Iskandar, Ilham Syahbandi, dan Annissa Kamila Mardhiyyah,
8. Sahabat-sahabat seperjuangan, Bagas Laska, Muhammad Rivaldi, Muhammad Fachrozy, Muhammad Ridho, Christian Tambunan, Andrew Wijaya, Henry, Niken Alfrido, Reksa, Ananda Pratama, Sari, Avinda Deviana yang memberikan semangat dan dukungan kepada penulis serta menjadi tempat bercerita dan kolaborasi,
9. Teman-teman Stambuk 2017 khususnya kom A yang sudah memberikan kenangan manis dan kebersamaan selama masa perkuliahan di Ilmu Komputer,
10. Semua pihak yang telah membantu penulis dalam penyelesaian skripsi ini yang tidak bisa disebutkan satu per satu.

IMPLEMENTASI ALGORITMA K-MEANS DAN ALGORITMA NAÏVE BAYES UNTUK IDENTIFIKASI TINGKAT RISIKO SAHAM PADA SEKTOR BARANG KONSUMEN PRIMER

ABSTRAK

Saham merupakan aset investasi yang memiliki karakteristik *high risk high gain*, yaitu aset yang menawarkan imbal hasil yang tinggi namun juga memiliki tingkat risiko tinggi. Data laporan tahunan Bursa Efek Indonesia menunjukkan bahwa sepanjang tahun 2022 terdapat peningkatan jumlah investor publik di pasar modal Indonesia yang mencakup saham, reksadana, dan obligasi, sebesar 37,68% menjadi 10,31 juta Single Investor Identification (SID). Dalam melakukan investasi, investor dihadapkan dengan dua potensi yaitu potensi keuntungan dan kerugian. Risiko berinvestasi dapat diminimalisir dengan memilih saham perusahaan yang memiliki kondisi fundamental yang baik. Penelitian ini bertujuan memudahkan investor dalam melakukan penyaringan kelompok saham yang berfundamental baik, khususnya pada sektor barang konsumen primer menggunakan algoritma K-Means *clustering* dengan optimasi K-Means++ dan algoritma klasifikasi Naïve Bayes. Penelitian menggunakan 82 data saham perusahaan sektor barang konsumen primer kategori *second liner* yang diambil dari sekuritas Stockbit dan aplikasi RTI Business pada tanggal 29 Februari 2024 dengan 6 rasio keuangan yang terdapat pada masing-masing perusahaan tersebut. Berdasarkan sistem yang telah dibangun didapatkan kesimpulan bahwa jumlah *k* optimal untuk kelas *clustering* adalah sebanyak 3 kelas menurut metode Elbow, dengan masing-masing kelasnya memiliki tingkat risiko rendah, sedang, dan tinggi. Sistem yang dibangun berbasis website ini juga mampu melakukan klasifikasi pada 10 saham di sektor yang sama namun tidak termasuk *second liner*. Dengan 4 skenario rasio data pelatihan dan pengujian, hasil pengujian sistem terhadap model yang telah dibangun berhasil memperoleh nilai akurasi tertinggi sebesar 94,44% dengan persentase 80% untuk data latih dan 20% untuk data pengujian, sedangkan rata-rata nilai akurasi yang didapat adalah sebesar 87,12%.

Kata kunci: *clustering*, klasifikasi, K-Means, Naïve Bayes, saham.

IMPLEMENTATION OF K-MEANS ALGORITHM AND NAÏVE BAYES ALGORITHM FOR IDENTIFICATION OF STOCK RISK LEVEL IN PRIMARY CONSUMER GOODS SECTOR

ABSTRACT

Stocks are investment assets that have high risk high gain characteristics, namely assets that offer high returns but also have a high level of risk. Data from the Indonesia Stock Exchange annual report shows that throughout 2022 there was an increase in the number of public investors in the Indonesian capital market, which includes stocks, mutual funds, and bonds, by 37.68% to 10.31 million Single Investor Identification (SID). In making an investment, investors are faced with two potentials, namely potential profits and losses. The risk of investing can be minimized by choosing company stocks that have good fundamental conditions. This research aims to facilitate investors in filtering groups of stock with good fundamental, especially in the primary consumer goods sector using the K-Means clustering algorithm with K-Means++ optimization and the Naïve Bayes classification algorithm. The study used 82 stock data of primary consumer goods sector companies in the second liner category taken from Stockbit securities and RTI Business applications on February 29, 2024 with 6 financial ratios contained in each of these companies. Based on the system that has been built, it is concluded that the optimal number of k for clustering classes is 3 classes according to the Elbow method, with each class having a low, medium and high risk level. This web-based system is also able to classify 10 stocks in the same sector but not including second liners. With 4 scenarios of training and testing data ratio, the results of system testing on the model that has been built successfully obtained the highest accuracy value of 94.44% with a percentage of 80% for training data and 20% for testing data, while the average accuracy value obtained is 87.12%.

Keywords: clustering, classification, K-Means, Naïve Bayes, stocks.

DAFTAR ISI

PERSETUJUAN	iii
PERNYATAAN.....	iv
PENGHARGAAN	v
ABSTRAK	vii
ABSTRACT.....	viii
DAFTAR ISI.....	ix
DAFTAR GAMBAR	xii
DAFTAR TABEL.....	xiii
BAB 1 PENDAHULUAN	1
1.1. Latar Belakang.....	1
1.2. Rumusan Masalah.....	6
1.3. Batasan Masalah	6
1.4. Tujuan Penelitian	7
1.5. Manfaat Penelitian	7
1.6. Metodologi Penelitian.....	7
1.7. Sistematika Penulisan	9
BAB 2 LANDASAN TEORI.....	10
2.1. Saham.....	10
2.2. Rasio Keuangan	11
2.2.1. Return on Equity (ROE).....	11
2.2.2. Net Profit Margin (NPM)	11
2.2.3. Debt to Equity Ratio (DER)	12
2.2.4. Dividend Yield	12

2.2.5.	Price Earning Ratio (PER)	13
2.2.6.	Price to Book Value (PBV)	13
2.3.	Sektor Barang Konsumen Primer	14
2.4.	Data Mining	15
2.5.	Machine Learning	17
2.5.1.	Supervised Learning.....	18
2.5.2.	Unsupervised Learning.....	19
2.5.3.	Semi-supervised Learning	19
2.6.	Data Scaling	20
2.7.	K-Means Clustering.....	21
2.8.	Metode Elbow	23
2.9.	Naïve Bayes Classifier.....	24
2.10.	Confusion Matrix	27
2.11.	Tinjauan Penelitian	29
BAB 3 ANALISIS DAN PERANCANGAN		31
3.1.	Analisis Sistem	31
3.1.1.	Analisis Masalah	31
3.2.	Arsitektur Umum	32
3.3.	Dataset	35
3.4.	Pre-processing Dataset.....	36
3.5.	Clustering.....	38
3.6.	Klasifikasi & <i>Training Model</i>	39
3.7.	Hasil Training	39
3.8.	Klasifikasi Dataset	39
3.9.	Pemodelan Sistem.....	40

3.9.1. Use Case Diagram	40
3.9.2. Activity Diagram	41
3.9.3. Sequence Diagram	43
3.10. Perancangan Sistem	44
3.10.1. Clustering dengan K-Means	44
3.10.2. Klasifikasi dengan Naïve Bayes	45
3.10.3. Integrasi K-Means dengan Naïve Bayes	46
3.11. Perancangan Antarmuka (<i>Interface</i>)	46
3.11.1. Rancangan Antarmuka Dashboard	47
3.11.2. Rancangan Antarmuka KMNB (K-Means Naïve Bayes)	48
3.11.3. Rancangan Antarmuka About	49
BAB 4 IMPLEMENTASI DAN PENGUJIAN SISTEM	51
4.1. Implementasi Sistem	51
4.2. Implementasi Data	51
4.3. Implementasi <i>Pre-processing</i> Data	56
4.4. Implementasi <i>Clustering</i>	57
4.5. Implementasi Klasifikasi Dataset	62
4.6. Pengujian Sistem	64
BAB 5 KESIMPULAN DAN SARAN	66
5.1. Kesimpulan	66
5.2. Saran	66
DAFTAR PUSTAKA	68
LAMPIRAN	71

DAFTAR GAMBAR

Gambar 2.1 Tahapan <i>Knowledge Discovery in Database</i>	15
Gambar 2.2 Grafik Metode Elbow	24
Gambar 2.3 <i>Confusion Matrix</i>	27
Gambar 3.1 Diagram Ishikawa	32
Gambar 3.2 Arsitektur Umum	34
Gambar 3.3 Dataset Saham Sektor Barang Konsumen Primer (Sumber : <i>Stockbit</i>)	36
Gambar 3.4 <i>Pre-processing</i> data (<i>filtering</i>).....	37
Gambar 3.5 Data setelah <i>min-max normalization</i>	38
Gambar 3.6 <i>Use Case Diagram</i>	40
Gambar 3.7 <i>Activity Diagram</i>	42
Gambar 3.8 <i>Sequence Diagram</i>	43
Gambar 3.9 <i>Flowchart</i> Algoritma K-Means	44
Gambar 3.10 <i>Flowchart</i> Algoritma Naïve Bayes	45
Gambar 3.11 Alur Integrasi K-Means dengan Naïve Bayes	46
Gambar 3.12 Rancangan Antarmuka <i>Dashboard</i>	47
Gambar 3.13 Rancangan Antarmuka KMNB.....	48
Gambar 3.14 Rancangan Antarmuka <i>About</i>	49
Gambar 4.1 <i>Upload</i> dataset setelah normalisasi.....	57
Gambar 4.2 Hasil Metode Elbow	58
Gambar 4.3 Hasil Akhir <i>Clustering</i>	61
Gambar 4.4 Data <i>Centroid</i> Akhir	62
Gambar 4.5 Proses Klasifikasi Dataset.....	63
Gambar 4.6 Nilai Akurasi Rasio 90% Data Pelatihan	64

DAFTAR TABEL

Tabel 4.1 Dataset Sistem.....	52
Tabel 4.2 Data Setelah Normalisasi	56
Tabel 4.3 Optimasi dengan K-Means++	59
Tabel 4.4 Hasil Klasifikasi	63
Tabel 4.5 Hasil Nilai Akurasi 4 Skenario Rasio	65

BAB 1

PENDAHULUAN

1.1. Latar Belakang

Dalam beberapa dekade terakhir, kemajuan teknologi informasi dan komputasi telah memberikan dampak signifikan terhadap analisis keuangan, terutama dalam bidang investasi. Investasi sendiri merupakan tindakan pembelian aset dan mengalokasikannya pada aset produktif yang direncanakan untuk dijual kembali di masa mendatang dengan nilai yang lebih tinggi sebagai strategi untuk memperoleh keuntungan dalam jangka waktu tertentu. Beberapa jenis aset investasi dapat berupa emas, deposito, properti, dan aset finansial pada pasar modal seperti saham, reksadana dan obligasi. Setiap jenis aset investasi memiliki karakteristik risiko dan imbal hasil yang berbeda, sehingga pemilihan aset-aset tersebut bergantung pada tujuan keuangan, jangka waktu investasi, dan toleransi risiko investor.

Saham termasuk salah satu aset investasi yang memiliki karakteristik *high risk* dan *high gain*, yaitu aset yang menawarkan potensi imbal hasil yang tinggi namun juga memiliki tingkat risiko yang tinggi (Daniswara & Daryanto, 2020). Saham adalah bagian kepemilikan yang dipegang seseorang (investor) dalam sebuah perusahaan yang didapatkan melalui investasi modal. Investasi modal ini kemudian memberikan hak kepada investor untuk dianggap sebagai salah satu pemilik dan memiliki hak pengawasan atas perusahaan sesuai dengan porsi kepemilikan mereka. Selayaknya suatu produk investasi, saham memiliki tempat perdagangan yang disebut sebagai pasar modal. Bursa Efek Indonesia (BEI) sebagai bursa efek utama di Indonesia berfungsi sebagai tempat pertemuan antara penawaran dari perusahaan-perusahaan yang ingin mendapatkan modal dari investor publik dengan permintaan dari investor yang mencari kesempatan untuk berinvestasi di instrumen keuangan dalam pasar modal (Sugiarti & Andreas, 2022).

Data laporan tahunan dari Bursa Efek Indonesia melalui situs resminya menunjukkan bahwa sepanjang tahun 2022 terdapat peningkatan

jumlah investor publik di pasar modal Indonesia yang mencakup saham, reksadana, dan obligasi, sebesar 37,68% menjadi 10,31 juta Single Investor Identification (SID) hingga akhir 2022. SID adalah kode pengenal identifikasi unik yang diberikan kepada setiap investor di pasar modal. Jumlah ini menandai peningkatan hingga hampir sembilan kali lipat dibandingkan lima tahun sebelumnya, dengan jumlah investor saham naik 28,64% menjadi 4,439 juta SID jika dibandingkan tahun sebelumnya.

Mengacu pada *press release* dari BEI (Bursa Efek Indonesia, 2023), terdapat lebih dari 900 perusahaan publik yang terdaftar dalam Bursa Efek Indonesia. Perusahaan publik yang terdaftar kemudian dikelompokkan ke dalam sektor-sektor yang sesuai dengan area konsentrasi perusahaan tersebut. Salah satu sektor yang terdapat dalam Bursa Efek Indonesia adalah sektor Barang Konsumen Primer (*Consumer Non-Cyclicals*). Menurut penjelasan dari Bursa Efek Indonesia, sektor Barang Konsumen Primer melibatkan perusahaan yang bertanggung jawab atas produksi atau pendistribusian barang dan jasa yang selalu dibutuhkan oleh konsumen, termasuk barang yang bersifat non-siklikal atau esensial, dimana kebutuhan atas produk-produk ini tetap konsisten tanpa memperhatikan kondisi ekonomi. Ini mencakup perusahaan-perusahaan seperti ritel makanan, obat-obatan, supermarket, produsen minuman dan makanan kemasan, penjual produk pertanian, produsen rokok, serta barang-barang kebutuhan rumah tangga dan perawatan pribadi. Perusahaan yang terdiri dari sub sektor *food & beverage*, *food & staples retailing*, *tobacco* dan *nondurable household products* ini memiliki bisnis yang menyediakan produk esensial yang dibutuhkan oleh seluruh masyarakat untuk pemenuhan kebutuhan sehari-hari mereka (Kesara, Zahro, & Susanti, 2023).

Beberapa saham perusahaan publik di sektor Barang Konsumen Primer memiliki kondisi fundamental yang berbeda-beda. Banyaknya jumlah perusahaan ini menawarkan variasi komposisi saham yang beragam pada portfolio investor, sehingga investor dituntut untuk dapat melakukan analisis

sebelum menentukan saham yang layak dibeli agar potensi keuntungan yang dapat diraih sesuai dengan risiko ataupun tingkat kerugian yang bisa dia hadapi. Risiko berinvestasi ini dapat diminimalisir dengan memilih saham-saham perusahaan yang memiliki kondisi fundamental yang baik. Fundamental perusahaan menggambarkan kondisi kesehatan perusahaan tersebut yang mempengaruhi fluktuasi harga sahamnya di bursa efek. Oleh karena itu, sebelum investor memutuskan untuk membeli saham, sangat disarankan untuk melakukan proses penyaringan saham atau *stock screening* yang bertujuan untuk menentukan saham-saham yang memenuhi kriteria tertentu sesuai dengan kebutuhan investor, seperti perusahaan yang secara konsisten memberikan dividen, memiliki rasio utang yang rendah, dan lain-lain (Sukamto, Setiawan, & Pratama, 2023).

Umumnya investor akan melakukan proses penyaringan ini dengan menelusuri dan membaca secara lengkap laporan keuangan tiap perusahaan yang tentu membutuhkan waktu cukup lama. Untuk mempersingkat waktu dalam proses penyaringan ini dibutuhkan pemahaman mengenai rasio keuangan pada perusahaan. Rasio keuangan yang sehat menunjukkan fundamental yang baik pada perusahaan, dan investor yang berfokus pada analisis fundamental perusahaan pada umumnya memiliki standar khusus dalam mengevaluasi perusahaan berdasarkan rasio keuangan tersebut, seperti besar dividen (pembagian laba) yang dibayarkan dibandingkan dengan harga saham (*Dividen Yield*), rasio utang perusahaan terhadap ekuitas yaitu *Debt to Equity Ratio* (DER), besar persentase keuntungan bersih dari pendapatan perusahaan yang diwujudkan sebagai *Net Profit Margin* (NPM), dan tingkat pengembalian atas ekuitas atau *Return on Equity* (ROE). Rasio keuangan lainnya berupa *Price Earning Ratio* (PER) yang mengukur valuasi berdasarkan perbandingan harga saham dengan laba bersih per lembar saham serta *Price to Book Value* (PBV) yang membandingkan harga saham dengan nilai buku perusahaan, juga menjadi parameter pilihan dalam memilih saham yang layak dibeli.

Untuk memudahkan proses pencarian dan pengelompokan saham berdasarkan rasio keuangan yang diinginkan investor, diperlukan suatu sistem yang dapat menampilkan daftar saham dengan kategori pilihan, dan hal ini dapat dilakukan dengan penerapan penggalian data (*data mining*) khususnya menggunakan algoritma klastering yaitu K-Means untuk pengelompokan tingkat risiko saham berdasarkan kesamaan kriteria rasio keuangan dan algoritma klasifikasi yakni Naïve Bayes untuk proses pengujian model dan klasifikasi saham. *Data mining* merupakan metode di mana data dalam skala besar dikumpulkan dengan melakukan ekstraksi dan mengidentifikasi pola-pola yang signifikan atau menelusuri informasi yang terdapat dalam basis data (Vulandari, 2017). Penelitian terkait pengelompokan saham dengan algoritma K-Means sebelumnya pernah dilakukan oleh (Sukamto, Setiawan, & Pratama, 2023) yang mengelompokkan saham pada sektor energi menggunakan metode K-Means. Dengan pertimbangan empat rasio keuangan yaitu *Dividend Yield*, *Gross Profit Margin*, *Return on Equity* dan *Debt Equity Ratio*, hasil penelitian mampu mengelompokkan 71 saham menjadi 4 kelas/*cluster* dengan karakteristik yang berbeda-beda sesuai dengan kondisi fundamental dan besar dividen yang diberikan perusahaan. Kesimpulan yang didapat adalah algoritma K-Means mampu membantu investor dalam melakukan pengelompokan saham berdasarkan kemiripan karakteristik yang diinginkan.

Penelitian lainnya yang melakukan pengelompokan saham berdasarkan *Price Earning Ratio* dan *Price to Book Value* menggunakan algoritma K-Means (Tohendry & Jollyta, 2023) memberikan kesimpulan bahwa dengan perbandingan dua rasio keuangan tersebut, algoritma K-Means dapat mengelompokkan 150 saham menjadi 3 *cluster* dimana jumlah *cluster* ini ditentukan melalui metode *Elbow*. *Cluster* tersebut terdiri atas kelompok saham dengan nilai murah yang cenderung berisiko rendah, saham bernilai sedang dengan risiko yang sedang, serta saham mahal yang cenderung berisiko tinggi. Sementara itu penelitian terkait integrasi algoritma K-Means dan Naïve Bayes pernah dilakukan oleh (Lestari, 2018) dengan judul

“Implementasi Algoritme K-Means Clustering Dan Naive Bayes Classifier Untuk Klasifikasi Diagnosa Penyakit Pada Kucing” untuk mengelompokkan diagnosa penyakit pada kucing menjadi 8 jenis penyakit dengan atribut gejala sebanyak 32 gejala. Hasil dari penelitian ini adalah dengan mengimplementasikan algoritma K-Means dan Naïve Bayes pada studi kasus diagnosa penyakit pada kucing berhasil mendapatkan rata-rata akurasi tertinggi sebesar 90% yang melebihi rata-rata akurasi dari metode K-Means konvensional sebesar 71,379%.

Penelitian selanjutnya yang menggabungkan kedua jenis algoritma tersebut juga pernah dilakukan oleh (Wahyuditomo, Cholissodin, & Sutrisno, 2021) dalam identifikasi tingkat risiko reksa dana yang memberi kesimpulan bahwa dengan mengkategorikan 250 data reksa dana menjadi 4 *cluster* yang optimal sesuai dengan hasil evaluasi optimal dari *Silhouette Coefficient* sebesar 0,46448 berhasil memberikan hasil klasifikasi dan evaluasi yang lebih baik dibanding menggunakan algoritma *Naïve Bayes* saja, dengan nilai rata-rata *precision* sebesar 0,9813, *recall* sebesar 0,9818, dan *F-measure* sebesar 0,9808. Penelitian terkait optimasi K-Means dengan K-Means++ untuk penentuan centroid awal yang lebih optimal sebelumnya pernah dilakukan oleh (Ferdiansyah & Chotijah, 2024) menunjukkan hasil evaluasi yang dilakukan mengindikasikan peningkatan kinerja dengan K-Means++ dibandingkan dengan penggunaan K-Means standar. Terlihat dari nilai Davies-Bouldin Index (DBI) yang lebih rendah dan Silhouette Coefficient yang lebih tinggi pada K-Means++, menunjukkan bahwa pengelompokan yang dihasilkan oleh K-Means++ lebih berkualitas dan memiliki batasan yang lebih jelas dalam pembentukan *cluster* untuk data penjualan bahan bangunan.

Sehingga berdasarkan latar belakang di atas penulis melakukan penelitian yang berjudul *“Implementasi Algoritma K-Means dan Algoritma Naïve Bayes Untuk Identifikasi Tingkat Risiko Saham Pada Sektor Barang Konsumen Primer”*

1.2. Rumusan Masalah

Melakukan proses pengelompokan saham pada sektor barang konsumen primer yang terdiri atas beberapa saham tentu membutuhkan waktu yang cukup lama. Seluruh saham tersebut juga memiliki likuiditas dan volatilitas yang berbeda-beda sehingga untuk meminimalisir pemilihan saham yang terlalu volatil dan kurang likuid, data saham yang akan dipilih dalam penelitian ini adalah saham perusahaan yang memiliki kapitalisasi pasar minimal 500 miliar yang tergolong ke dalam *second liner* hingga *first liner* (*blue chip*) yaitu perusahaan yang kapitalisasi pasarnya di atas 10 triliun sehingga data saham yang akan dikelompokkan menjadi 82 saham. Rumusan masalah dari penelitian ini mengarah kepada pengelompokan saham untuk memudahkan investor dalam memilih saham berfundamental baik berdasarkan rasio keuangan dan preferensi tingkat risiko yang terdapat pada tiap kelompok saham tersebut.

1.3. Batasan Masalah

Batasan masalah dari penelitian ini adalah sebagai berikut :

1. *Dataset* sebanyak 82 saham *second liner* pada sektor barang konsumen primer diperoleh dari sekuritas *Stockbit* berdasarkan data dan harga penutupan saham di akhir bulan Februari 2024 (29 Februari 2024), dalam format *xlsx* (Excel).
2. Algoritma yang diimplementasikan adalah algoritma K-Means *clustering* dengan optimasi K-Means++ untuk inisialisasi *centroid* awal dan algoritma klasifikasi Naïve Bayes.
3. Sistem yang dibangun berbasis *website* dengan menggunakan bahasa pemrograman PHP dan *framework* CodeIgniter versi 4.
4. Rasio keuangan yang dipilih sebagai atribut dalam penelitian ini sebanyak delapan atribut, yaitu *Return on Equity* (ROE), *Net Profit Margin* (NPM), *Debt to Equity Ratio* (DER), *Dividend Yield*, *Price Earning Ratio* (PER), dan *Price to Book Value* (PBV).

5. Data saham yang akan diidentifikasi menggunakan algoritma Naïve Bayes adalah yang tidak termasuk dalam kategori *second liner* sebanyak 10 saham teratas yang diurutkan berdasarkan besar kapitalisasi pasarnya.
6. Hasil akhir (*output*) dari penelitian adalah tingkat risiko pada saham sektor barang konsumen primer yang terbagi menjadi 3 kelas sesuai nilai optimum dari metode Elbow, yaitu risiko rendah, sedang dan tinggi.

1.4. Tujuan Penelitian

Tujuan penelitian ini adalah sebagai berikut :

1. Membangun sistem berbasis *website* untuk dapat menentukan pilihan saham dengan tingkat risiko sesuai rasio keuangan pada sektor barang konsumen primer.
2. Mengimplementasikan algoritma K-Means dengan optimasi K-Means++ dan algoritma Naïve Bayes dalam sistem untuk pembentukan kelompok dan klasifikasi saham dengan variasi tingkat risiko.
3. Menguji model dari proses *clustering* sebelumnya dengan data yang telah diberi label dan melakukan identifikasi tingkat risiko saham yang tidak termasuk kategori *second liner* menggunakan algoritma Naïve Bayes.

1.5. Manfaat Penelitian

Manfaat yang diharapkan dapat diperoleh dari penelitian ini adalah dapat memudahkan investor atau pelaku pasar modal dalam melakukan penyaringan saham di sektor barang konsumen primer berdasarkan rasio keuangan yang telah ditentukan sehingga mampu memberikan rekomendasi saham sesuai dengan tingkat risiko investor dan dapat melakukan identifikasi pada saham selain batasan data dalam penelitian ini melalui sistem berbasis *website*.

1.6. Metodologi Penelitian

Proses penelitian dilakukan menggunakan sejumlah metodologi sebagai berikut:

1. Studi Pustaka

Tahap awal dalam melakukan penelitian ini adalah dengan studi pustaka yaitu menelusuri referensi yang mendukung penelitian terkait dengan sumber informasi dari jurnal, skripsi, artikel ilmiah maupun buku-buku dan situs internet sehingga mendapatkan pemahaman yang lebih mendalam terkait algoritma K-Means *clustering* dan algoritma klasifikasi Naïve Bayes serta pemahaman dalam bidang analisis investasi yaitu saham dengan rasio keuangan yang terkait dalam penelitian.

2. Analisis dan Perancangan

Tahapan selanjutnya adalah melakukan analisis terkait kebutuhan penelitian dan sistem yang akan dibangun seperti pengumpulan data saham dari sekuritas, pra-pemrosesan data serta melakukan perancangan terhadap sistem sehingga sistem dapat berfungsi dengan baik.

3. Implementasi Sistem

Pada tahap ini dilakukan pembangunan atau pembuatan sistem berdasarkan analisis dan perancangan sebelumnya yang diimplementasikan dalam sistem berbentuk *website* menggunakan PHP dan *framework* CodeIgniter.

4. Pengujian Sistem

Setelah tahap implementasi sistem, perlu dilakukan evaluasi agar sistem dapat berjalan sesuai dengan yang diharapkan. Pada tahap ini akan dilakukan pengujian untuk mengevaluasi kinerja sistem dalam melakukan *clustering* dan klasifikasi pada data saham yang ditentukan.

5. Dokumentasi

Dokumentasi dilakukan sejak awal penelitian, mulai dari analisis hingga fase pengujian sistem, dan akan dirangkum dalam bentuk laporan penelitian atau skripsi.

1.7. Sistematika Penulisan

Sistematika dalam penulisan laporan penelitian ini terdiri dari beberapa bagian utama sebagai berikut :

BAB 1: PENDAHULUAN

Bab 1 menjelaskan latar belakang penulis dalam memilih topik penelitian terkait, rumusan masalah yang menjadi fokus penyelesaian, batasan masalah, tujuan penelitian, manfaat penelitian, metodologi penelitian, dan sistematika penulisan dalam penelitian.

BAB 2: LANDASAN TEORI

Bab 2 menjabarkan tentang konsep dan dasar teori yang dibutuhkan serta kajian pustaka dari penelitian sebelumnya untuk mendukung penelitian terkait permasalahan yang akan diselesaikan.

BAB 3: ANALISIS DAN PERANCANGAN SISTEM

Bab 3 membahas tentang analisis dan perancangan sistem, perhitungan manual, serta perancangan antarmuka terhadap sistem dalam penelitian yang dibangun sebagai solusi dalam masalah terkait.

BAB 4: IMPLEMENTASI DAN PENGUJIAN SISTEM

Bab 4 menjelaskan implementasi dari algoritma yang diterapkan dalam sistem dan menganalisis hasil pengujian untuk evaluasi kemampuan pada sistem dalam mengidentifikasi tingkat risiko saham dengan persentase rasio *data train* dan *data test* yang berbeda-beda.

BAB 5: KESIMPULAN DAN SARAN

Bab 5 berisi kesimpulan dari keseluruhan penelitian yang telah diuraikan dalam masing-masing bab dan saran yang diharapkan dapat menjadi pedoman berguna untuk penelitian mendatang serta potensi pengembangan lebih lanjut.

BAB 2

LANDASAN TEORI

2.1. Saham

Saham merupakan salah satu aset investasi yang diperdagangkan di pasar modal dengan karakteristik *high risk high gain*, dimana aset ini menawarkan potensi imbal hasil yang tinggi namun juga memiliki tingkat risiko yang tinggi (Daniswara & Daryanto, 2020). Aset investasi ini berbentuk selembarnya kertas atau lembaran yang menunjukkan bagian kepemilikan yang dipegang seseorang (investor) dalam sebuah perusahaan yang didapatkan melalui investasi modal. Investasi modal ini kemudian memberikan hak kepada investor untuk dianggap sebagai salah satu pemilik dan memiliki hak pengawasan atas perusahaan sesuai dengan porsi kepemilikan mereka.

Dalam setiap keputusan investasi tentu terdapat ekspektasi akan keuntungan yang dapat diperoleh dan kerugian yang harus dihadapi. Demikian halnya dengan saham, seorang investor yang memutuskan pilihannya untuk berinvestasi saham di pasar modal berharap untuk dapat memperoleh imbal hasil yang baik dan mampu meningkatkan portfolio keuangan mereka dalam berinvestasi. Pada investasi saham terdapat dua macam keuntungan yang bisa diperoleh yaitu *capital gain* dan dividen, dimana *capital gain* adalah keuntungan yang diperoleh dari perbedaan harga jual saham yang lebih tinggi dibandingkan dengan harga pembeliannya, sedangkan dividen adalah bagian dari laba yang didistribusikan oleh perusahaan kepada pemegang sahamnya. Namun selain potensi keuntungan juga terdapat potensi kerugian yang harus dihadapi investor berupa *capital loss* dan risiko likuidasi perusahaan. *Capital loss* terjadi ketika harga pembelian saham lebih tinggi daripada harga penjualannya sementara risiko likuidasi perusahaan adalah proses di mana sebuah perusahaan diumumkan bangkrut oleh pengadilan dan seluruh asetnya dijual guna melunasi utang perusahaan (Sugiarti & Andreas, 2022).

2.2. Rasio Keuangan

Rasio keuangan merupakan alat penting dalam analisis fundamental untuk mengevaluasi kesehatan finansial dan kinerja perusahaan, termasuk dalam pasar saham Indonesia. Terdapat beberapa jenis rasio seperti rasio aktivitas, rasio profitabilitas, rasio aktivitas, rasio solvabilitas, dan rasio pasar (Kamaludin & Indriani, 2012). Rasio-rasio ini membantu investor dan analis untuk membuat keputusan investasi yang berlandaskan prinsip fundamental dengan membandingkan informasi keuangan dari laporan keuangan perusahaan seperti neraca, laporan laba rugi, dan laporan arus kas. Beberapa rasio keuangan yang umum digunakan untuk menganalisis kesehatan perusahaan diantaranya sebagai berikut :

2.2.1. *Return on Equity (ROE)*

Salah satu rasio keuangan yang menunjukkan seberapa efektif perusahaan menghasilkan laba dari penjualan atau asetnya adalah *Return on Equity (ROE)* , dimana menurut (Gitman, 2000) *Return on Equity* berguna untuk mengukur seberapa besar nilai imbal hasil dan keuntungan yang mampu dihasilkan perusahaan untuk para pemegang saham terhadap modal yang telah diinvestasikan. Sehingga rasio ini dapat digunakan untuk mengukur tingkat efektivitas perusahaan dalam mengelola modal yang dimiliki untuk menghasilkan laba dalam perusahaan (Ristyan, 2019). Nilai rasio ini dihitung dengan persamaan sebagai berikut :

$$\text{Return On Equity (ROE)} = \frac{\text{Laba bersih perusahaan}}{\text{Total ekuitas}} \times 100\%$$

2.2.2. *Net Profit Margin (NPM)*

Net Profit Margin (NPM) termasuk ke dalam jenis rasio yang sama dengan *Return on Equity (ROE)*, yaitu rasio yang menunjukkan keefektifan perusahaan dalam menghasilkan laba dari penjualan (rasio profitabilitas). *Net Profit Margin* adalah indikator yang mengukur efisiensi perusahaan dalam menghasilkan rupiah laba bersih dari setiap pendapatan penjualan. Ini

mencerminkan efektivitas perusahaan dalam berbagai aspek, termasuk pemasaran, produksi, administrasi, pendanaan, kebijakan harga dan manajemen pajak (Sugiarti & Andreas, 2022). Oleh karena itu perusahaan dinilai semakin *profitable* dan efisien jika nilai rasio *Net Profit Margin* semakin tinggi. Rasio ini dinyatakan dalam persentase dengan rumus sebagai berikut :

$$\text{Net Profit Margin (NPM)} = \frac{\text{Laba bersih perusahaan}}{\text{Pendapatan}} \times 100\%$$

2.2.3. Debt to Equity Ratio (DER)

Selain rasio profitabilitas juga terdapat rasio solvabilitas, yaitu rasio yang mengukur kemampuan perusahaan dalam memenuhi kewajiban jangka panjang. *Debt to Equity Ratio* termasuk dalam rasio solvabilitas yang digunakan untuk menilai proporsi antara utang terhadap ekuitas yang dimiliki perusahaan (Brigham & Houston, 2019). Rasio ini mencerminkan kemampuan ekuitas perusahaan untuk menutupi kewajiban kepada pemberi pinjaman saat perusahaan harus dilikuidasi.

Jumlah utang yang lebih besar dibandingkan dengan ekuitas menunjukkan sikap agresif perusahaan dalam mengambil pinjaman, sehingga nilai *Debt to Equity Ratio* yang tinggi juga menunjukkan bahwa perusahaan memiliki risiko likuiditas yang tinggi. Rumus yang digunakan untuk mencari nilai *Debt to Equity Ratio* sebagai berikut :

$$\text{Debt to Equity Ratio (DER)} = \frac{\text{Laba bersih perusahaan}}{\text{Pendapatan}} \times 100\%$$

2.2.4. Dividend Yield

Dividen adalah bagian dari laba yang didistribusikan oleh perusahaan kepada pemegang sahamnya, dimana perbandingan persentase besaran dividen yang dibagikan terhadap harga sahamnya disebut sebagai *Dividend Yield*. Jika perusahaan memiliki nilai *Dividend Yield* yang tinggi maka dapat disimpulkan perusahaan mampu memberikan “bunga” dividen yang lebih

tinggi relatif terhadap harga sahamnya. *Dividend Yield* dapat dihitung dalam persentase sebagai berikut :

$$Dividend Yield = \frac{Dividen\ per\ lembar\ saham}{Harga\ saham\ per\ lembar} \times 100\%$$

2.2.5. Price Earning Ratio (PER)

Valuasi suatu perusahaan juga dapat diukur melalui rasio pasar, dimana rasio ini mengukur persepsi investor terhadap harga saham perusahaan. *Price Earning Ratio* menunjukkan seberapa besar investor bersedia membayar untuk setiap rupiah laba perusahaan yang dihasilkan. Rasio ini mengindikasikan ekspektasi pasar terhadap kinerja dan pertumbuhan laba masa depan perusahaan.

Dengan kata lain, *Price Earning Ratio* (PER) adalah ukuran yang membandingkan harga saham dengan pendapatan perusahaan atau laba per lembar saham. Ketika PER suatu perusahaan tinggi, investor cenderung menganggap perusahaan tersebut mahal, sebaliknya jika PER rendah maka perusahaan cenderung dianggap murah oleh investor (Tohendry & Jollyta, 2023). Namun nilai PER yang lebih tinggi juga dapat disebabkan oleh peningkatan harga saham yang melebihi kenaikan laba perusahaan secara signifikan, atau penurunan laba perusahaan yang lebih besar daripada penurunan harga saham. Rasio ini dapat dihitung dengan menggunakan rumus :

$$Price\ to\ Earning\ Ratio\ (PER) = \frac{Harga\ saham}{Laba\ per\ lembar\ saham}$$

2.2.6. Price to Book Value (PBV)

Price to Book Value menjelaskan valuasi saham terhadap nilai bukunya, dimana nilainya diperoleh melalui perbandingan harga saham dengan nilai buku perusahaan, sehingga rasio ini mengindikasikan seberapa tinggi pasar dalam menghargai nilai buku dalam suatu perusahaan (Darmadji & Fakhruddin, 2012). Nilai buku dalam rasio PBV menunjukkan nilai

keseluruhan aset perusahaan jika dilikuidasi dan penilaian terhadap rasio ini sering digunakan untuk perusahaan yang asetnya mudah dinilai jika dilihat dari laporan keuangan. Perhitungan untuk mencari nilai PBV dapat dirumuskan sebagai berikut :

$$\text{Price to Book Value (PBV)} = \frac{\text{Harga saham}}{\text{Nilai buku per lembar saham}}$$

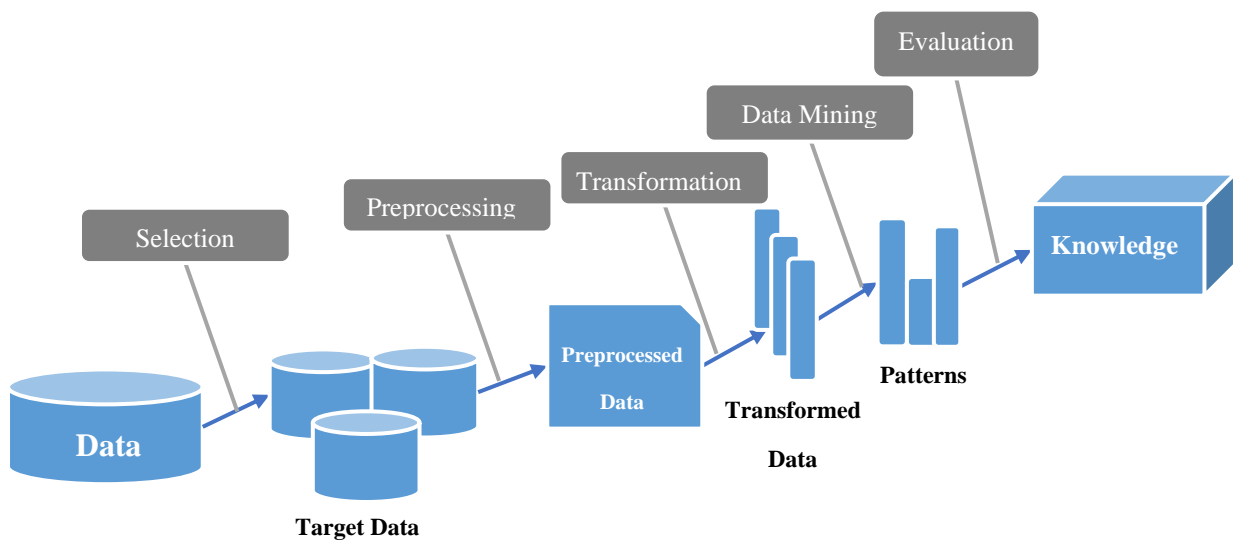
2.3. Sektor Barang Konsumen Primer

Sektor Barang Konsumen Primer merupakan salah satu sektor perusahaan yang terdapat di pasar modal Indonesia yang berisi daftar perusahaan dengan tanggung jawab atas produksi atau pendistribusian barang dan jasa yang selalu dibutuhkan oleh konsumen, termasuk barang yang bersifat non-siklikal atau esensial, dimana kebutuhan atas produk-produk ini tetap konsisten tanpa memperhatikan kondisi ekonomi. Ini mencakup perusahaan-perusahaan seperti ritel makanan, obat-obatan, supermarket, produsen minuman dan makanan kemasan, penjual produk pertanian, produsen rokok, serta barang-barang kebutuhan rumah tangga dan perawatan pribadi. Perusahaan yang terdiri dari sub sektor food & beverage, food & staples retailing, tobacco dan nondurable household products ini memiliki bisnis yang menyediakan produk esensial yang dibutuhkan oleh seluruh masyarakat untuk pemenuhan kebutuhan sehari-hari mereka (Kesara, Zahro, & Susanti, 2023).

Sektor ini memiliki karakteristik yang tahan terhadap berbagai kondisi ekonomi dikarenakan sifatnya yang anti-siklus. Pertumbuhan yang stabil dan moderat juga membuat saham di sektor Barang Konsumen Primer menjadi pilihan utama saat kondisi ekonomi menurun. Bahkan saat resesi, kinerja saham di dalam sektor ini tetap menunjukkan pertumbuhan yang meningkat pesat seperti perusahaan jamu yang mengalami lonjakan permintaan selama pandemi. Investor dengan preferensi investasi yang stabil dan kurang berisiko dapat memikirkan untuk investasi jangka panjang pada saham yang berada di sektor ini dikarenakan pendapatannya yang cenderung konsisten dan tidak terpengaruh oleh fluktuasi pasar yang signifikan.

2.4. Data Mining

Data mining adalah elemen penting dalam *proses Knowledge Discovery in Database* (KDD), yang memiliki peran utama dalam menarik informasi dari data yang akan diproses (Asroni, Fitri, & Prasetyo, 2018). Proses KDD ditunjukkan melalui langkah-langkah pada Gambar 2.1 berikut ini :



Gambar 2.1 Tahapan *Knowledge Discovery in Database*

Proses dalam *Knowledge Discovery in Database* (KDD) secara umum dapat diuraikan sebagai berikut:

1. Seleksi Data

Memilih data dari kumpulan data operasional sebagai langkah awal sebelum memulai proses ekstraksi informasi dalam KDD.

2. *Pre-processing* / Pembersihan

Langkah ini melibatkan pembersihan data yang menjadi pusat KDD sebelum memulai proses Data Mining. Tujuannya adalah untuk menghilangkan duplikasi, ketidakkonsistenan data, dan kesalahan ketik.

3. Transformasi

Proses ini mengubah data yang dipilih menjadi format yang sesuai untuk diterapkan dalam Data Mining, seringkali tahap ini melibatkan *coding* dalam proses transformasi data.

4. *Data Mining*

Langkah ini mencakup proses menemukan pola atau informasi dalam data menggunakan metode tertentu. Pilihan metode atau algoritma sangat bergantung pada tujuan dan proses KDD secara keseluruhan.

5. Evaluasi / Interpretasi

Setelah pola atau informasi ditemukan dalam data kemudian dilakukan pengecekan apakah informasi tersebut sesuai dengan hipotesa ataupun fakta sebelumnya.

Hasil penemuan pola atau informasi dalam proses *data mining* dapat dimanfaatkan sebagai pertimbangan atas pengambilan keputusan untuk berbagai tujuan di berbagai bidang seperti pendidikan, kesehatan, bisnis dan bidang lainnya. Proses *data mining* sendiri melibatkan dua jenis data dalam pengolahannya, yaitu data terstruktur dan data tidak terstruktur. Data tidak terstruktur memiliki pengertian bahwa informasi atau pola dalam data tersebut tidak mudah diproses seperti pada data terstruktur yang disimpan dalam suatu *database* yang berhubungan, seperti gambar, video atau teks. Data-data dalam bentuk tersebut diolah dengan runtutan proses yang menghabiskan waktu lebih lama dibandingkan dengan data terstruktur seperti data yang berbentuk numerik.

Proses penggalian informasi ini melibatkan beberapa algoritma yang digunakan sesuai dengan tujuan akhir yang ingin diperoleh seperti klasifikasi, klusterisasi, prediksi numerik, asosiasi dan *neural networks* (Ramageri, 2010).

2.5. Machine Learning

Machine Learning adalah salah satu bidang ilmu dari Kecerdasan Buatan (*Artificial Intelligence*) yang pertama kali diperkenalkan oleh peneliti dari IBM (*International Business Machines*) Corporation bernama Arthur Samuel pada tahun 1959, yang berfokus pada pengembangan algoritma komputer untuk melakukan pengolahan data sehingga menghasilkan respons / tindakan yang cerdas (*intelligent action*). Bidang ini melakukan pembelajaran dari model matematis yang telah dibentuk melalui sekumpulan data yang berisi fitur-fitur tertentu dengan tujuan akhirnya berupa penggambaran karakteristik umum dari keseluruhan data (deskriptif), prediksi kejadian yang akan datang berdasarkan kesimpulan data sebelumnya (prediktif), serta penemuan pengetahuan atau informasi baru dari data yang telah dilatih (*knowledge discovery*).

Proses dalam *machine learning* sendiri berkaitan erat dengan data yang akan dikumpulkan. Hal ini dikarenakan *output* yang dihasilkan oleh *machine learning* berbanding lurus dengan kualitas data dalam proses pembelajarannya, sehingga semakin baik kualitas data yang dipersiapkan maka akan semakin akurat pula *output* dari proses pembelajaran *machine learning* tersebut. Dalam langkah-langkah pembelajaran ini, proses penentuan data pelatihan / sampel menghabiskan waktu yang lebih banyak dibandingkan langkah-langkah lainnya. Data yang dikumpulkan selanjutnya akan dibagi menjadi data pelatihan dan data pengujian berdasarkan rasio tertentu, oleh karena itu sampel yang ditentukan pada langkah sebelumnya harus mampu merepresentasikan keseluruhan data yang ada sehingga model pembelajaran mampu berfungsi dengan optimal dan memberikan *output* yang baik untuk tujuan akhir proses pembelajaran.

Setelah proses penentuan dan pembagian data menjadi data pelatihan dan data pengujian berdasarkan rasio tertentu, maka dilakukan proses pelatihan model melalui data tersebut dengan algoritma yang tersedia dalam *machine learning*. Model ini kemudian akan diuji dan dievaluasi kembali untuk menghindari kemungkinan bias / kecenderungan informasi yang

terbentuk. Pengevaluasian performa model ini dilakukan melalui data uji yang dibagi secara persentase dari keseluruhan data berdasarkan rasio yang telah ditentukan sebelumnya untuk dapat mencapai *output* yang optimal (Lantz, 2013).

Secara umum terdapat tiga macam algoritma dalam *machine learning*, yaitu *Supervised Learning* (algoritma pembelajaran terarah dengan label yang akan diprediksi / dideskripsikan sudah diketahui sebelumnya), *Unsupervised Learning* (algoritma pembelajaran tidak terarah dengan label tidak diketahui sebelumnya), serta *Semi-supervised Learning* (berada di antara *supervised* dan *unsupervised learning*).

2.5.1. *Supervised Learning*

Supervised Learning merupakan algoritma pembelajaran terarah dan terstruktur dalam *machine learning* dengan sifat data inputannya yang telah memiliki label untuk memperoleh hasil yang diharapkan sebagai tujuan akhir proses pembelajaran baik itu untuk tujuan deskriptif maupun prediktif. Label yang telah diberikan pada data sebelumnya menggambarkan sifat umum dan karakteristik penyebaran setiap data untuk selanjutnya memudahkan model pembelajaran dalam mengenali pola dan hubungan dalam data tersebut.

Terdapat dua *output* yang umumnya dihasilkan dalam metode *supervised learning*, yaitu klasifikasi dan regresi. Klasifikasi memberikan hasil berupa tipe data kategorikal dari proses pelatihan dan pengujian data selama proses pembelajaran, sementara regresi menghasilkan variabel berupa nilai *riil* yang bersifat kontinyu dengan fokus utamanya adalah mencari hubungan dari setiap variabel yang ada. Beberapa jenis metode yang populer digunakan pada tipe metode pembelajaran terarah ini seperti Naïve Bayes, *Support Vector Machine* (SVM), *Neural Networks*, dan *k-Nearest Neighbor* untuk tujuan klasifikasi, serta *Linear Regression* yang umum digunakan untuk menghasilkan *output* dalam bentuk regresi.

2.5.2. *Unsupervised Learning*

Metode pembelajaran *unsupervised learning* disebut sebagai metode pembelajaran tidak terarah dikarenakan sifat dataset sebagai inputannya yang tidak memiliki label sehingga hasilnya bukanlah suatu prediksi dari atribut-atribut yang tersedia dalam data. Hasil dari metode ini berupa pengelompokkan yang didasarkan pada kemiripan karakteristik setiap atribut dari kumpulan data tersebut, dengan demikian pemodelan data dalam proses pembelajaran diperoleh dari pola atau *cluster* pada data.

Pengelempokkan data dilakukan dengan aturan-aturan dan pengamatan tertentu yang menghasilkan beberapa *cluster* untuk setiap data, dan data yang berhasil dikategorikan dalam satu *cluster* tertentu memiliki perbedaan karakteristik dengan data yang terdapat pada *cluster* lainnya, sehingga setiap data memiliki pemberian label yang berbeda untuk setiap kelompok / *cluster* sesuai dengan kemiripan sifat atribut yang dimiliki.

2.5.3. *Semi-supervised Learning*

Perpaduan antara metode *supervised learning* dan *unsupervised learning* disebut sebagai *semi-supervised learning* (SSL). Jenis pembelajaran ini mengurangi kemungkinan kesalahan dari kedua metode pembelajaran sebelumnya, dimana *supervised learning* sangat terkait dengan kebutuhan terhadap dataset yang harus memiliki label pada proses pembelajarannya, sehingga kualitas *output* yang dihasilkan tergantung pada ada tidaknya label pada dataset, sedangkan *unsupervised learning* memiliki kemungkinan pengelompokkan data yang tidak akurat dikarenakan adanya ketidakseimbangan atribut atau sifat dari kumpulan data tersebut.

Oleh karena itu untuk mengatasi kekurangan tersebut, metode *semi-supervised learning* dikembangkan dengan mempelajari sebagian data yang memiliki label (*supervised*) dan sebagian data lainnya yang tidak memiliki label (*unsupervised*). Model dalam algoritma ini kemudian dibangun dengan memproses data yang berlabel sebagai data *training* dan data yang tidak berlabel sebagai *data testing*.

2.6. Data Scaling

Pre-processing data, yang juga dikenal sebagai pra-pemrosesan data, adalah tahap dalam teknik *data mining* dan *machine learning*. Tahap ini diperlukan untuk mengatasi kemungkinan data yang tidak lengkap, tidak konsisten, atau bercampur (seperti ketika ada variabel numerik dan kategorik dalam satu dataset) yang akan sangat berpengaruh pada hasil prediksi yang salah dan menjadi tidak signifikan. *Data scaling* merupakan salah satu tahap dari pra-pemrosesan yang bertujuan untuk mengurangi kesalahan prediksi sebagai tujuan akhir dari pengolahan data, dengan menyesuaikan skala nilai yang terdapat pada kumpulan data ke dalam suatu rentang atau standar tertentu. Tahapan ini juga membuat proses pembelajaran mesin menjadi lebih efisien.

Salah satu teknik dalam *data scaling* adalah *min-max normalization*, yang menormalisasikan data dengan mengubah semua rentang nilai dalam fitur data menjadi rentang atau skala baru yang umum diterapkan, antara 0 hingga 1. Proses *min-max normalization* dilakukan dengan persamaan sebagai berikut :

$$X_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

dengan keterangan :

X_{norm} = nilai setelah normalisasi data

X = nilai asli data

X_{min} = nilai minimum (terkecil) pada dataset

X_{max} = nilai maksimum (terbesar) pada dataset

Dengan melakukan proses normalisasi, data akan memiliki rentang distribusi dalam proporsi yang sama sehingga meminimalisir kemungkinan adanya variabel tidak seimbang yang memengaruhi hasil *machine learning*.

2.7. K-Means Clustering

Dalam metode pembelajaran tanpa arah (*unsupervised learning*) terdapat salah satu teknik pengelompokan yang disebut *clustering*. *Clustering* merupakan teknik pemilahan data menjadi beberapa kelompok atau *cluster* dimana proses ini didasarkan pada kesamaan atau kemiripan karakteristik antar data untuk selanjutnya data dikelompokkan ke dalam *cluster* yang sesuai dengan kemiripan tersebut. Algoritma *k-means* adalah salah satu algoritma *clustering* yang melakukan fungsinya dalam membagi data menjadi kelompok-kelompok tertentu namun bukan berdasarkan kemiripan atribut, melainkan dengan pertimbangan jarak terdekat dari pusat data (*centroid*) ke tiap-tiap titik data yang tersedia. Untuk dapat mengetahui cara kerja algoritma *k-means* secara komprehensif maka berikut adalah langkah-langkah dari algoritma tersebut :

1. Menentukan jumlah *cluster* untuk pengelompokan data baik itu sesuai keinginan *user* maupun melalui metode evaluasi tertentu seperti *Elbow Method*, *Davies Bouldin Index*, atau *Silhouette Coefficient*. Jumlah *k* (*cluster*) yang telah ditentukan menjadi pedoman jumlah data sebagai *centroid* (pusat *cluster*) selama algoritma dijalankan.
2. Setelah mengetahui jumlah *k*, selanjutnya dilakukan inisialisasi *centroid* awal dari data yang tersedia baik secara acak atau melalui optimasi *k-means++*, dimana *k-means++* bertujuan untuk meminimalisir keacakan data dalam inisialisasi *centroid* awalnya.
3. Melakukan perhitungan selisih jarak tiap data ke masing-masing *centroid* dengan persamaan *Euclidean Distance* yaitu sebagai berikut :

$$D(x_i, y_j) = \sqrt{\sum_{k=1}^n (x_{ik} - y_{jk})^2}$$

dengan keterangan sebagai berikut :

$D(x_i, y_j)$ = jarak dari data i ke *centroid* j pada *cluster* data terkait.

x_{ik} = nilai data ke i pada atribut ke k

y_{jk} = nilai *centroid* (titik pusat) ke j pada atribut ke k

4. Perhitungan dilakukan secara berulang sampai keseluruhan data telah diketahui selisih jaraknya, kemudian dilakukan pengelompokkan data ke dalam *cluster* yang ditentukan berdasarkan jarak terdekat tiap data ke *cluster-cluster* tersebut (nilai paling kecil / minimal).
5. Menetapkan nilai *centroid* baru pada iterasi selanjutnya dengan mencari nilai rata-rata dari tiap data dalam masing-masing atribut yang termasuk dalam *cluster* yang dipilih dengan rumus :

$$C_k = \frac{\sum_{i=1}^n x_{ik}}{n_{ik}}$$

dimana :

C_k = *centroid* baru pada *cluster* k

x_{ik} = data atribut i pada *cluster* k

n_k = banyak data atribut i pada *cluster* k

6. Apabila iterasi selanjutnya menghasilkan *cluster* dengan data yang berbeda dibandingkan data *cluster* di iterasi sebelumnya maka proses dilakukan ulang dengan kembali ke langkah nomor 3.
7. Apabila iterasi selanjutnya menghasilkan *cluster* dengan data yang sama seperti iterasi sebelumnya maka proses selesai dan hasil pengelompokkan telah didapatkan.

Untuk memperoleh hasil yang lebih efisien dan optimal dalam algoritma *k-means* dapat dilakukan optimasi dengan metode *k-means++* pada inisialisasi *centroid* awalnya, dimana langkah-langkahnya adalah sebagai berikut :

1. Memilih *centroid* awal (inisialisasi *centroid*) secara acak dari kumpulan data yang tersedia.

2. Melakukan perhitungan untuk nilai *centroid* baru dengan mengecualikan titik data yang telah dipilih sebagai *centroid* awal sebelumnya berdasarkan probabilitas terbesar jarak kuadrat antara nilai data x ke *centroid* sebelumnya dengan persamaan berikut :

$$K = \frac{D(x)^2}{\sum_{x \in X} D(x)^2}$$

dengan keterangan :

$D(x)^2$ = jarak *Euclidean Distance* dari data x ke *centroid*

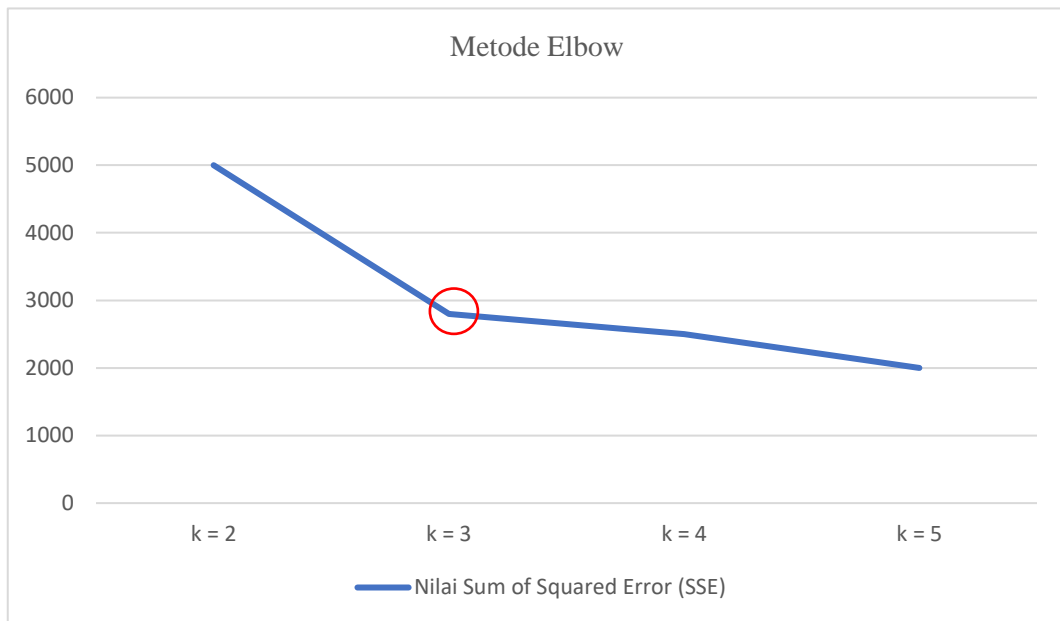
$\sum_{x \in X} D(x)^2$ = total jumlah jarak *Euclidean Distance*

3. Iterasi langkah 2 sampai *centroid* sebanyak k telah didapatkan.
4. Melanjutkan proses pengelompokkan data dengan algoritma *k-means* standar dengan *centroid* yang telah ditentukan dari algoritma *k-means++*

Algoritma *k-means++* akan secara berurutan memilih sekelompok K pusat *cluster* dengan mengambil titik pusat berikutnya dari distribusi di mana setiap titik memiliki probabilitas yang proporsional terhadap jarak kuadratnya ke pusat terdekat (Lattanzi & Sohler, 2019).

2.8. Metode Elbow

Untuk mendapatkan jumlah *cluster* optimal dalam algoritma *k-means*, terdapat salah satu metode yang menggunakan perbandingan hasil setiap k (*cluster*) dengan melihat nilai kesalahan kuadrat (*sum of squared error*) yang dihasilkan yang disebut metode Elbow (siku). Metode ini menghitung nilai *sum of squared error* (SSE) yang dihasilkan dari setiap rentang jumlah k dan ditampilkan dalam bentuk grafik untuk selanjutnya dibandingkan hasilnya. Apabila nilai SSE pada suatu *cluster* dengan *cluster* berikutnya membentuk siku atau nilainya mengalami penurunan yang signifikan maka nilai pada *cluster* tersebut adalah nilai *cluster* yang optimum. Representasi grafis untuk metode Elbow dapat dilihat pada Gambar 2.2 berikut :



Gambar 2.2 Grafik Metode Elbow

Untuk mencari nilai SSE dari setiap k dapat menggunakan persamaan sebagai berikut :

$$SSE = \sum_{K=1}^K \sum_{x_i} \| X_i - C_k \|^2$$

dimana :

K = jumlah *cluster*

X_i = nilai data ke- i

C_k = nilai *centroid cluster*

2.9. Naïve Bayes Classifier

Metode pembelajaran terarah (*supervised learning*) merupakan metode yang melakukan pelatihan dan pengujian pada data yang telah memiliki label sebelumnya, sehingga metode ini akan melakukan prediksi atau klasifikasi pada data baru yang belum memiliki label. Contoh dari metode ini adalah pada proses klasifikasi, dimana tujuan klasifikasi adalah untuk melakukan pengenalan / prediksi suatu data baru yang selanjutnya akan dinilai untuk dimasukkan pada label tertentu dari keseluruhan label yang tersedia. Proses

pengenalan ini didasarkan pada model yang telah dibangun sebelumnya menggunakan data pelatihan yang telah memiliki label.

Naïve Bayes adalah salah satu algoritma klasifikasi yang umum digunakan karena prosesnya yang tidak terlalu rumit berdasarkan teorema Bayes dimana teorema ini menjelaskan konsep probabilitas suatu kejadian berdasarkan pengalaman kejadian lain sebelumnya, atau dapat dirumuskan sebagai berikut :

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

dengan keterangan tiap notasinya adalah :

A = hipotesis data B merupakan suatu kelas tertentu

B = data dengan label atau kelas yang belum diketahui

$P(A|B)$ = probabilitas hipotesis A bersyarat kejadian bukti (*evidence*) B
(probabilitas posterior)

$P(B|A)$ = probabilitas bukti B yang dipengaruhi hipotesis A

$P(A)$ = probabilitas awal (prior) hipotesis A tanpa bukti apapun

$P(B)$ = probabilitas bukti B tanpa melihat hipotesis yang lain

Dalam teorema Naïve Bayes, untuk menentukan kelas tertentu yang cocok bagi sampel data yang dianalisis dibutuhkan sejumlah petunjuk agar dapat dilakukan proses klasifikasi. Sehingga teorema tersebut dapat disesuaikan seperti berikut :

$$P(C|F_1 \dots F_n) = \frac{P(C) \times P(F_1 \dots F_n|C)}{P(F_1 \dots F_n)}$$

dimana penjelasan formulanya adalah sebagai berikut :

C	=	kelas dari sampel data
$F_1 \dots F_n$	=	karakteristik petunjuk dari data untuk kebutuhan klasifikasi
$P(C F_1 \dots F_n)$	=	probabilitas hipotesis C bersyarat kejadian bukti $F_1 \dots F_n$ (probabilitas posterior)
$P(F_1 \dots F_n C)$	=	probabilitas bukti $F_1 \dots F_n$ yang dipengaruhi hipotesis C
$P(C)$	=	probabilitas awal (prior) hipotesis C tanpa bukti apapun
$P(F_1 \dots F_n)$	=	probabilitas bukti $F_1 \dots F_n$ tanpa melihat hipotesis lain
n	=	jumlah hipotesis yang terjadi

Persamaan ini menjelaskan probabilitas sampel data untuk masuk ke suatu kelas C tertentu (posterior) berdasarkan karakteristik pada sampel tersebut. Probabilitas ini dicari dengan menghitung peluang muncul kelas C sebelum sampel tersebut masuk (prior) kemudian dikali dengan probabilitas karakteristik-karakteristik sampel muncul pada kelas C (*likelihood*) lalu dibagi dengan probabilitas karakteristik sampel muncul secara global (*evidence*). Nilai probabilitas posterior selanjutnya akan dibandingkan dengan nilai probabilitas posterior dari kelas-kelas lainnya untuk menentukan kelas sampel yang diklasifikasikan.

Algoritma Naïve Bayes umumnya digunakan untuk data yang bersifat kategorik, dan oleh karena itu pencarian nilai probabilitas posterior pada algoritma untuk tipe data kontinyu yang nilainya berubah-ubah seperti data numerik memerlukan fungsi Gaussian atau sebaran normal dalam prosesnya. Algoritma Naïve Bayes untuk tipe data seperti ini selanjutnya disebut sebagai algoritma Gaussian Naïve Bayes dengan persamaan sebagai berikut :

$$P(X|Y) = \frac{1}{\sqrt{2\pi}\sigma} \exp \frac{-(x-\mu)^2}{2\sigma^2}$$

dimana :

σ = nilai simpangan baku / standar deviasi

μ = nilai rata-rata

dan notasi lainnya yaitu π (Phi) yang bernilai 3,14, exp yang bernilai 2,7183 dan x adalah nilai data pada sampel terkait.

2.10. Confusion Matrix

Performa model klasifikasi umumnya diukur dengan nilai-nilai yang direpresentasikan dalam bentuk matriks yang disebut sebagai *confusion matrix*. *Confusion Matrix* menjelaskan performa dari model yang telah dibuat dengan membandingkan nilai prediksi dan aktual dari sampel data dalam bentuk matriks. Setiap baris merepresentasikan hasil prediksi dari data sedangkan setiap kolom merepresentasikan nilai aktual dari data sebenarnya. *Confusion Matrix* dapat diilustrasikan seperti pada Gambar 2.3 berikut :

		NILAI PREDIKSI	
		1 (Positive)	0 (Negative)
NILAI AKTUAL	1 (Positive)	TP (True Positive)	FN (False Negative) <i>Type II Error</i>
	0 (Negative)	FP (False Positive) <i>Type I Error</i>	TN (True Negative)

Gambar 2.3 *Confusion Matrix*

dengan keterangan sebagai berikut :

1. *True Positive* (TP) merupakan jumlah data yang aktualnya positif dan model memprediksi dengan benar (positif).
2. *True Negative* (TN) merupakan jumlah data yang aktualnya negatif dan model memprediksi dengan benar (negatif).
3. *False Positive* (FP) – atau Type error II merupakan jumlah data yang aktualnya negatif namun model memprediksi positif.
4. *False Negative* (FN) – atau Type error I merupakan jumlah data yang aktualnya positif namun model memprediksi negatif.

Nilai-nilai pada matriks tersebut selanjutnya dapat digunakan untuk mengukur performa model yang telah dibangun, dengan penjelasan sebagai berikut :

1. *Accuracy*, yang menunjukkan kinerja model dalam memprediksi benar (positif atau negatif) dibandingkan dengan total keseluruhan data. Nilai ini ditunjukkan dalam bentuk perbandingan sebagai berikut :

$$Accuracy = \frac{TP + TN}{n}$$

2. *Recall* (*sensitivity / True Positive Rate*), yang menunjukkan kinerja model dalam memprediksi benar positif dibandingkan dengan keseluruhan data yang aktualnya juga positif. Persamaannya adalah :

$$Recall = \frac{TP}{FN + TP}$$

3. *Precision*, yang menunjukkan kinerja model dalam memprediksi benar positif dibandingkan dengan total keseluruhan prediksi positif pada data, dengan persamaan sebagai berikut :

$$Precision = \frac{TP}{FP + TP}$$

4. *F1 Score*, yang menunjukkan rata-rata atau keseimbangan antara nilai *precision* dan *recall*. Apabila nilai *F1 Score* tinggi maka semakin baik pula performa model yang telah dibangun.

$$F1\ Score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

2.11. Tinjauan Penelitian

Beberapa penelitian terkait topik algoritma K-Means dan Naïve Bayes untuk studi kasus pengelompokkan saham maupun implementasi integrasi kedua algoritma untuk studi kasus lainnya pernah dilakukan di penelitian sebelumnya seperti :

1. Berdasarkan penelitian oleh (Sukanto, Setiawan, & Pratama, 2023) dengan judul “*Data Mining untuk Pengelompokan Saham pada Sektor Energi dengan Metode K-Means*” terkait pengelompokan saham pada sektor energi menggunakan algoritma K-Means. Dengan pertimbangan empat rasio keuangan yaitu *Dividend Yield*, *Gross Profit Margin*, *Return on Equity* dan *Debt Equity Ratio*, hasil penelitian mampu mengelompokkan 71 saham menjadi 4 kelas/*cluster* dengan karakteristik yang berbeda-beda sesuai dengan kondisi fundamental dan besar dividen yang diberikan perusahaan. Kesimpulan yang didapat adalah algoritma K-Means mampu membantu investor dalam melakukan pengelompokan saham berdasarkan kemiripan karakteristik yang diinginkan.
2. Berdasarkan penelitian (Tohendry & Jollyta, 2023) berjudul “*Penerapan Algoritma K-Means Clustering Untuk Pengelompokkan Saham Berdasarkan Price Earning Ratio Dan Price To Book Value*” memberikan kesimpulan bahwa dengan perbandingan dua rasio keuangan tersebut, algoritma K-Means dapat mengelompokkan 150 saham menjadi 3 *cluster* dimana jumlah *cluster* ini ditentukan melalui metode *Elbow*. *Cluster* tersebut terdiri atas kelompok saham dengan nilai murah yang

cenderung berisiko rendah, saham bernilai sedang dengan risiko yang sedang, serta saham mahal yang cenderung berisiko tinggi

3. Berdasarkan penelitian oleh (Lestari, 2018) dengan judul “*Implementasi Algoritme K-Means Clustering Dan Naive Bayes Classifier Untuk Klasifikasi Diagnosa Penyakit Pada Kucing*” untuk mengelompokkan diagnosa penyakit pada kucing menjadi 8 jenis penyakit dengan atribut gejala sebanyak 32 gejala. Hasil dari penelitian ini adalah dengan mengimplementasikan algoritma K-Means dan Naïve Bayes pada studi kasus diagnosa penyakit pada kucing berhasil mendapatkan rata-rata akurasi tertinggi sebesar 90% yang melebihi rata-rata akurasi dari metode K-Means konvensional sebesar 71,379%.
4. Berdasarkan penelitian (Wahyuditomo, Cholissodin, & Sutrisno, 2021) berjudul “*Implementasi Integrasi K-Means dan Naïve Bayes dalam Identifikasi Tingkat Risiko Reksa Dana*” dalam identifikasi tingkat risiko reksa dana yang memberi kesimpulan bahwa dengan mengkategorikan 250 data reksa dana menjadi 4 *cluster* yang optimal sesuai dengan hasil evaluasi optimal dari *Silhouette Coefficient* sebesar 0,46448 berhasil memberikan hasil klasifikasi dan evaluasi yang lebih baik dibanding menggunakan algoritma *Naïve Bayes* saja, dengan nilai rata-rata *precision* sebesar 0,9813, *recall* sebesar 0,9818, dan *F-measure* sebesar 0,9808.
5. Berdasarkan penelitian oleh (Ferdiansyah & Chotijah, 2024) dengan judul “*Implementasi Algoritme K-Means++ Untuk Clustering Penjualan Bahan Bangunan*” menunjukkan hasil evaluasi yang dilakukan mengindikasikan peningkatan kinerja dengan K-Means++ dibandingkan dengan penggunaan K-Means standar. Terlihat dari nilai Davies-Bouldin Index (DBI) yang lebih rendah sebesar 0,5890 dan *Silhouette Coefficient* yang lebih tinggi pada K-Means++ sebesar 0,6652, menunjukkan bahwa pengelompokan yang dihasilkan oleh K-Means++ lebih berkualitas dan memiliki batasan yang lebih jelas dalam pembentukan *cluster* untuk data penjualan bahan bangunan.

BAB 3

ANALISIS DAN PERANCANGAN

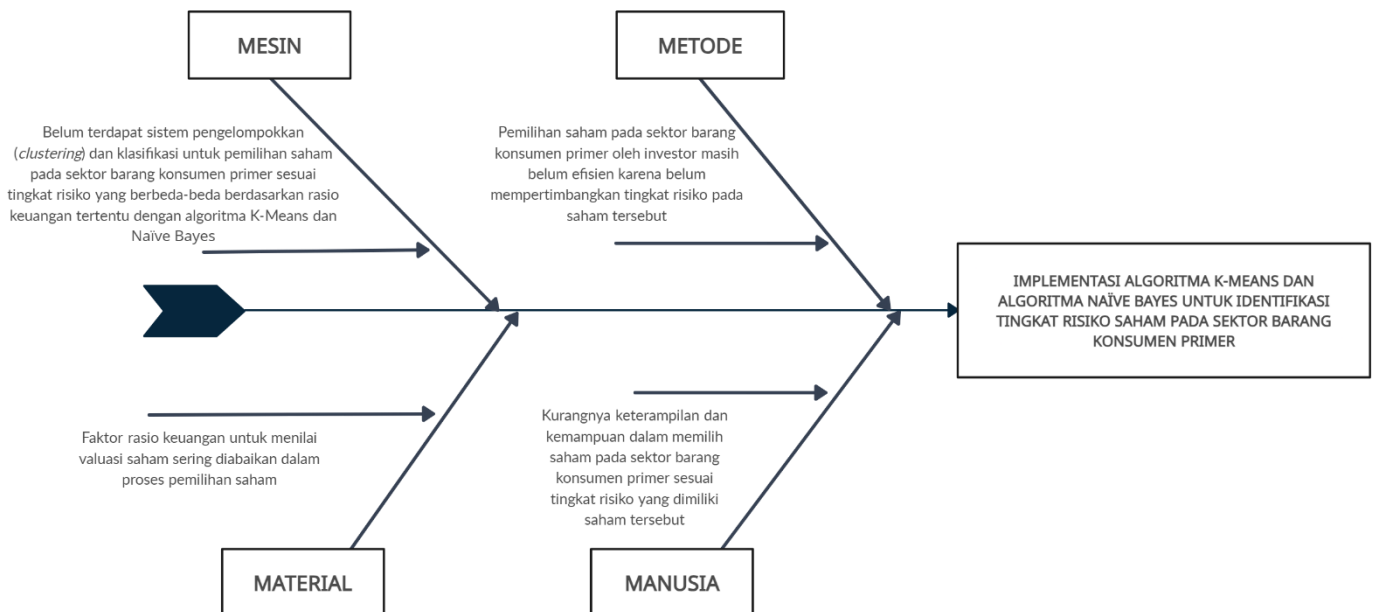
3.1. Analisis Sistem

Untuk mencapai tujuan tertentu dari suatu permasalahan dalam sistem dibutuhkan suatu proses analisa kebutuhan dalam sistem itu sendiri yang disebut sebagai analisis sistem. Analisis ini melibatkan proses penguraian sistem menjadi elemen-elemen pembentuknya untuk memahami cara kerja dan interaksi antar elemen dalam mencapai tujuan sistem yang diinginkan.

3.1.1. Analisis Masalah

Analisis masalah merupakan proses analisa dalam sistem dengan fokus utama adalah untuk mengidentifikasi masalah yang terdapat dalam sistem dengan mengetahui faktor penyebab terjadinya masalah tersebut lalu menguraikan faktor-faktor tersebut untuk selanjutnya ditemukan solusi terbaiknya. Penelitian ini membahas permasalahan mengenai bagaimana membangun suatu sistem yang dapat mengidentifikasi tingkat risiko pada saham sektor barang konsumen primer berdasarkan rasio keuangan yang telah ditentukan untuk kemudian dapat dijadikan dasar atau landasan dalam klasifikasi jenis saham lainnya (saham baru) di sektor yang sama.

Masalah penelitian ini dapat dinyatakan melalui sebuah diagram yang menjelaskan faktor penyebab permasalahan melalui hubungan sebab-akibat yang dikenal sebagai diagram Ishikawa. Melalui representasi dalam bentuk garis dan simbol, diagram Ishikawa mampu menunjukkan hubungan sebab dan akibat yang melatarbelakangi suatu permasalahan. Diagram yang berbentuk seperti tulang ikan (*fishbone*) ini dapat dilihat pada Gambar 3.1.

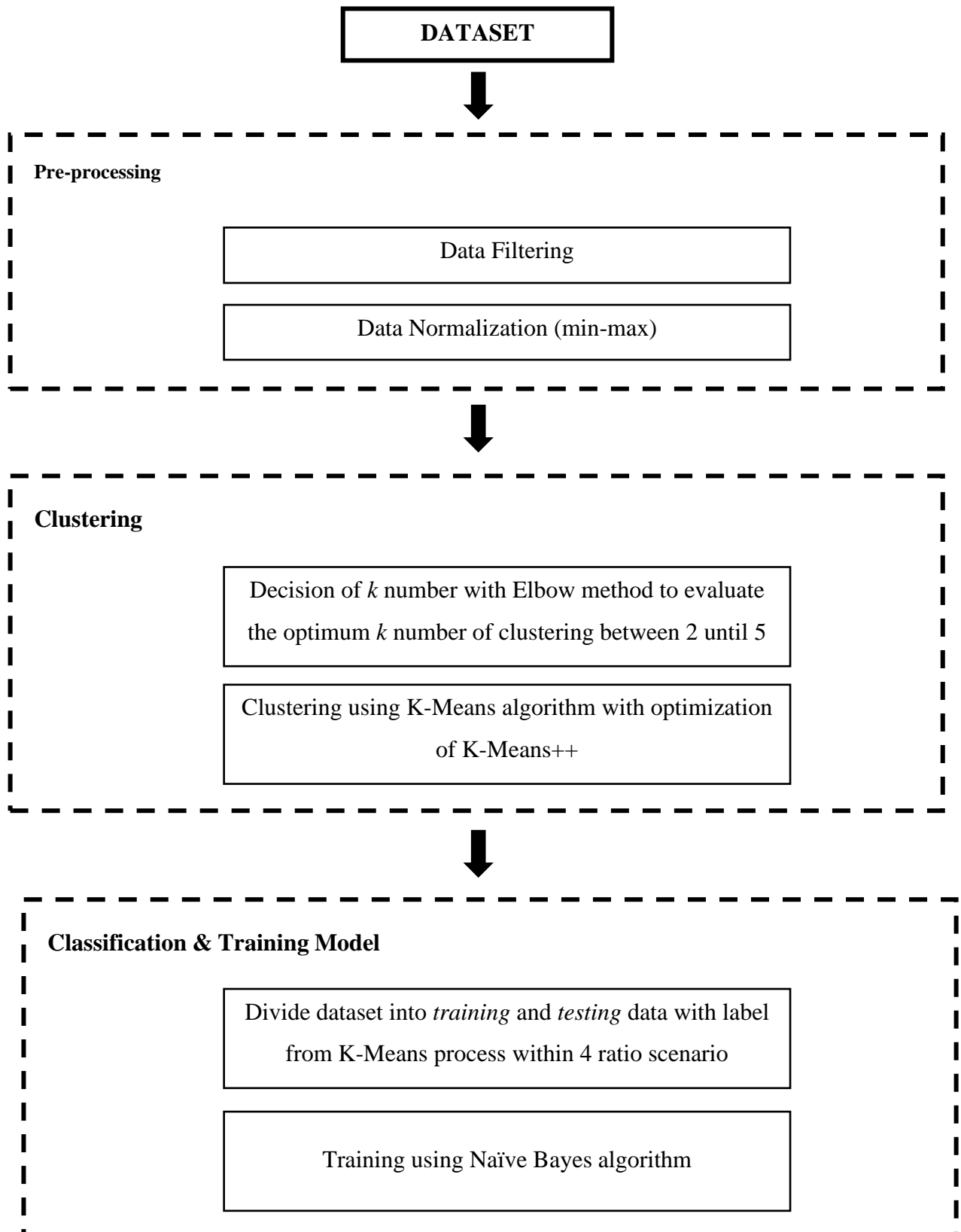


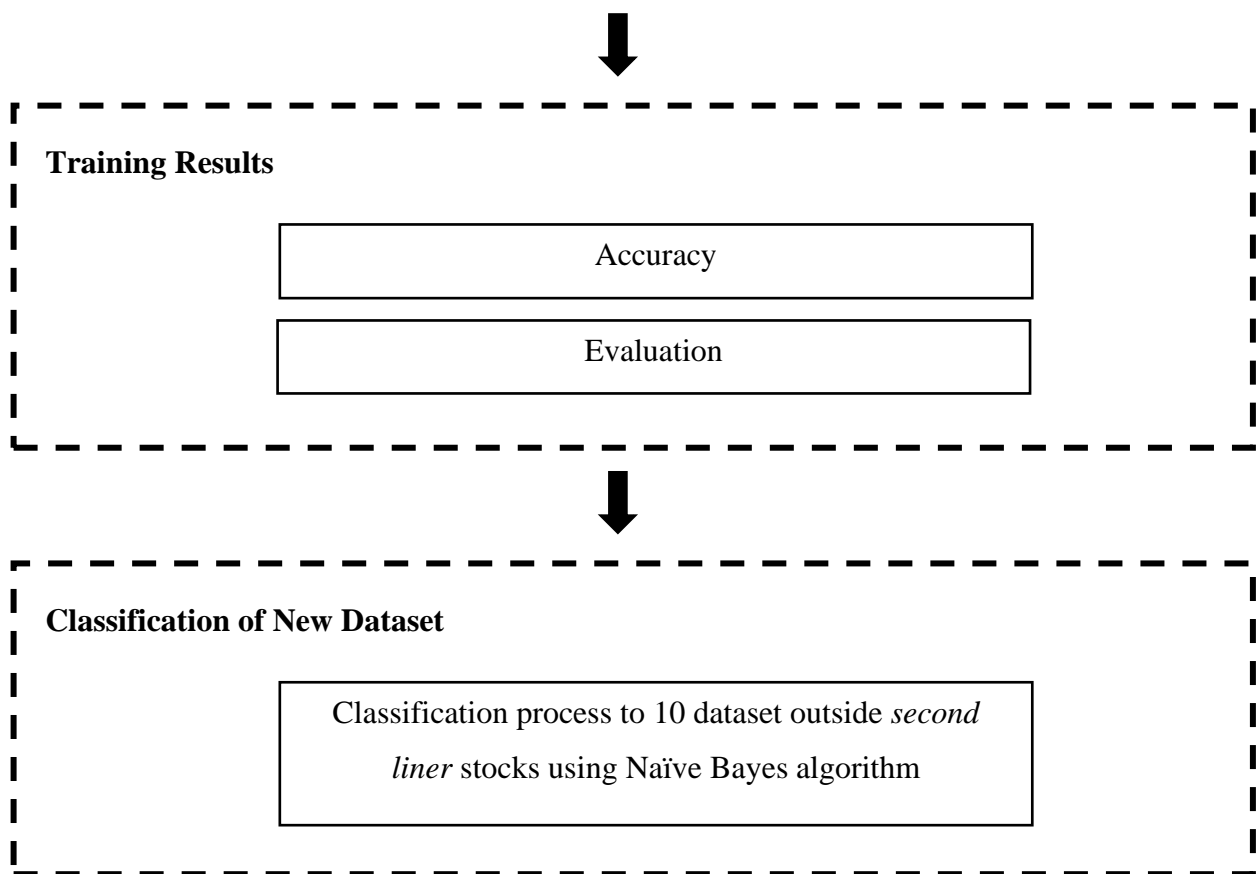
Gambar 3.1 Diagram Ishikawa

Dengan representasi simbol dan garis yang mirip seperti tulang ikan, bagian kepala ikan dalam diagram tersebut menjelaskan permasalahan dalam sistem yang terjadi, sementara tulang ikan yang terhubung ke kepala ikan tersebut menjelaskan penyebab dari masalah terkait.

3.2. Arsitektur Umum

Seluruh proses dan alur yang terjadi dalam sistem harus direpresentasikan agar sistem memiliki proses yang sistematis dalam setiap langkah-langkah pengerjaannya untuk mencapai tujuan yang diharapkan. Representasi sistem yang dimaksud ini disebut sebagai arsitektur umum. Arsitektur umum akan membantu dalam memahami keseluruhan alur sistem yang dibangun sehingga setiap prosesnya memiliki penjelasan yang lengkap mengenai komponen-komponen yang dibutuhkan dan *output* yang dihasilkan untuk selanjutnya memperoleh tujuan besar dalam sistem tersebut. Arsitektur umum pada sistem dapat dilihat pada Gambar 3.2 berikut :





Gambar 3.2 Arsitektur Umum

Pada arsitektur umum yang terlampir, tahap awal dalam penelitian ini adalah menyiapkan dataset saham pada sektor barang konsumen primer yang bersumber dari sekuritas *Stockbit*. Kemudian proses *data filtering* dilakukan pada dataset dengan hanya memilih data saham yang memiliki *Market Capital* (kapitalisasi pasar) minimal 500 miliar atau yang termasuk dalam kategori *second liner*. Sebanyak 82 dataset yang terpilih selanjutnya dinormalisasi dengan menggunakan *min-max normalization*, sehingga tidak ada ketidakseimbangan nilai pada keseluruhan sampel data.

Setelah dilakukan normalisasi data, proses berikutnya adalah melakukan *clustering* pada dataset dengan menentukan jumlah k (*cluster*) terlebih dahulu dari evaluasi metode Elbow dengan rentang k dari 2 sampai dengan 5 untuk mengetahui nilai k optimum yang akan digunakan dalam proses *clustering* dengan algoritma K-Means. Saat proses *clustering*

dilakukan juga optimasi dengan K-Means++ pada inisiasi nilai *centroid* (titik pusat data) awal untuk menghindari pemilihan data yang acak untuk nilai titik pusat data.

Algoritma K-Means akan memberikan hasil berupa data-data yang telah memiliki *cluster* (kelas) atau label. Langkah selanjutnya adalah melibatkan algoritma klasifikasi Naïve Bayes untuk melakukan pelatihan (*training*) dan pengujian (*testing*) pada data yang telah memiliki label dari proses K-Means dengan 4 skenario rasio presentase data latih dan data uji. Hasil pengujian kemudian akan dievaluasi dan dilihat apakah model memperoleh nilai akurasi yang baik sehingga tujuan penelitian dapat tercapai dan mampu memberikan kinerja yang baik.

Tahap akhir dalam penelitian ini adalah menjalankan model yang telah diuji pada dataset saham di sektor yang sama yang tidak termasuk ke dalam *second iner* untuk diklasifikasi apakah saham-saham tersebut termasuk ke dalam *cluster* atau label-label yang telah ditentukan.

3.3. Dataset

Dataset dalam penelitian ini diperoleh dari sekutiras *Stockbit* dengan mengumpulkan data saham yang berada pada sektor barang konsumen primer. Hasil eksplorasi dan pengumpulan data pada *Stockbit* memperoleh beberapa data saham yang termasuk pada sektor tersebut dengan beragam karakteristik fundamental dalam perusahaan. Dataset diambil pada akhir Februari tahun 2024 tepatnya di tanggal 29 Februari 2024 seperti terlampir pada Gambar 3.3 berikut ini :

Stockbit

Stock Screener

Screen Name Screener Konsumen Primer

Results 1 - 25 of 125 Equities

First Previous Next Last

Symbol	Market Cap
FLMC	21.88 B
SOUL	31.39 B
AMMS	39.60 B
IKAN	41.67 B
KPAS	47.62 B
OILS	52.22 B
NAYZ	53.55 B
PCAR	58.33 B
NASI	62.17 B
GRPM	64.89 B
FOOD	74.75 B
NANO	81.42 B
AGAR	90.00 B
MBTO	99.51 B
DPUM	100.20 B
TAYS	104.40 B
DSFI	107.71 B
ALTO	109.59 B
ENZO	123.27 B
DEWI	130.00 B
WAPO	135.26 B
COCO	140.60 B
MRAT	153.22 B
PSDN	154.08 B
GOLL	183.25 B

First Previous Next Last

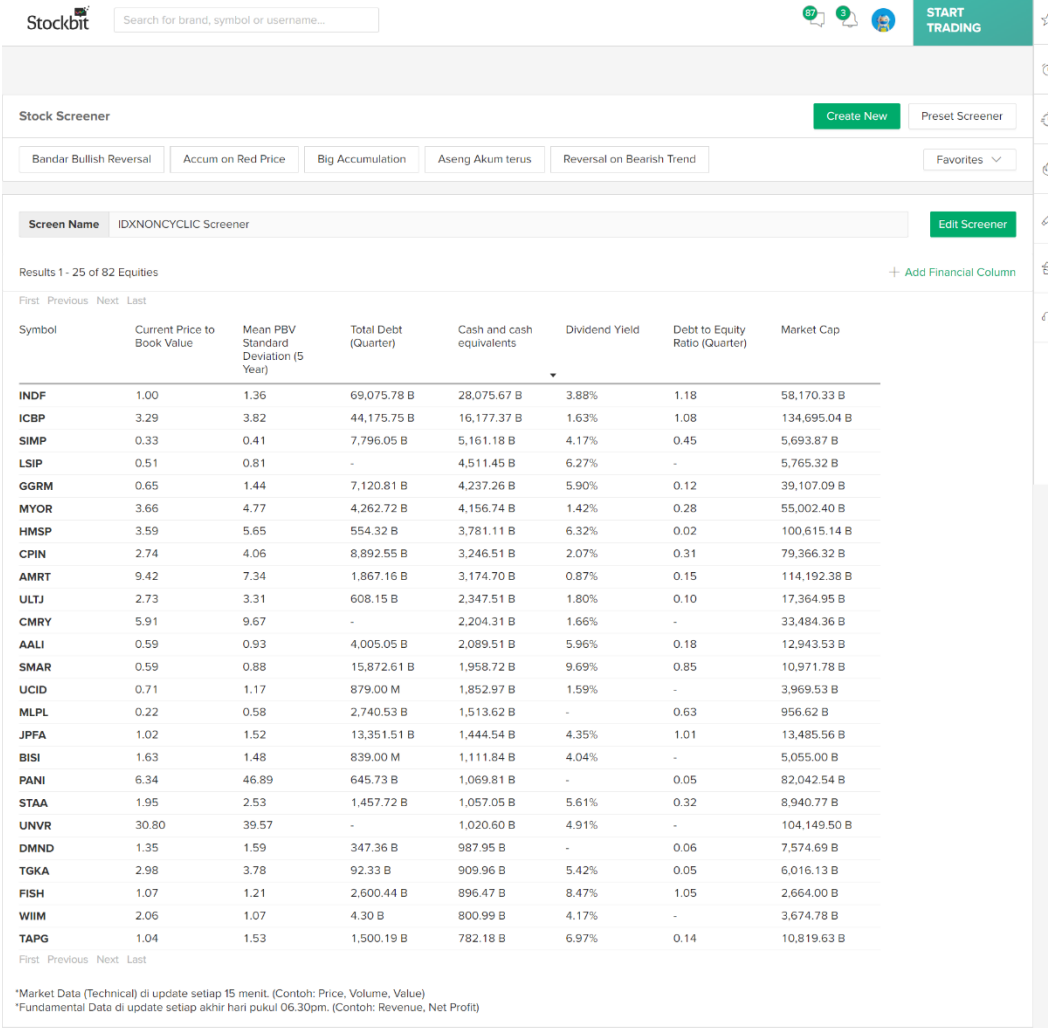
*Market Data (Technical) di update setiap 15 menit. (Contoh: Price, Volume, Value)
*Fundamental Data di update setiap akhir hari pukul 06.30pm. (Contoh: Revenue, Net Profit)

Gambar 3.3 Dataset Saham Sektor Barang Konsumen Primer
(Sumber : *Stockbit*)

3.4. Pre-processing Dataset

Pada tahap *pre-processing*, dataset yang telah terkumpul diolah dengan melakukan *filtering* dan normalisasi dengan *min-max normalization* agar data yang dihasilkan semakin optimal dan memiliki keseragaman antar data sehingga tidak terdapat ketidakseimbangan nilai pada data-data tersebut. Proses ini memastikan data yang akan digunakan dalam algoritma harus sesuai dengan tujuan yang diharapkan dengan menghilangkan nilai pada data

yang tidak relevan dan memastikan keseluruhan nilai pada data memiliki rentang yang sama. Dataset beberapa saham yang telah diperoleh sebelumnya akan difilter menjadi data saham yang memiliki kapitalisasi pasar minimal 500 miliar, sehingga total keseluruhan dataset menjadi 82 data seperti ditunjukkan pada Gambar 3.4 berikut :



Stockbit Search for brand, symbol or username...

START TRADING

Stock Screener Create New Preset Screener

Bandar Bullish Reversal Accum on Red Price Big Accumulation Aseng Akum terus Reversal on Bearish Trend Favorites

Screen Name IDNONCYCLIC Screener Edit Screener

Results 1 - 25 of 82 Equities + Add Financial Column

Symbol	Current Price to Book Value	Mean PBV Standard Deviation (5 Year)	Total Debt (Quarter)	Cash and cash equivalents	Dividend Yield	Debt to Equity Ratio (Quarter)	Market Cap
INDF	1.00	1.36	69,075.78 B	28,075.67 B	3.88%	1.18	58,170.33 B
ICBP	3.29	3.82	44,175.75 B	16,177.37 B	1.63%	1.08	134,695.04 B
SIMP	0.33	0.41	7,796.05 B	5,161.18 B	4.17%	0.45	5,693.87 B
LSIP	0.51	0.81	-	4,511.45 B	6.27%	-	5,765.32 B
GGRM	0.65	1.44	7,120.81 B	4,237.26 B	5.90%	0.12	39,107.09 B
MYOR	3.66	4.77	4,262.72 B	4,156.74 B	1.42%	0.28	55,002.40 B
HMSP	3.59	5.65	554.32 B	3,781.11 B	6.32%	0.02	100,615.14 B
CPIN	2.74	4.06	8,892.55 B	3,246.51 B	2.07%	0.31	79,366.32 B
AMRT	9.42	7.34	1,867.16 B	3,174.70 B	0.87%	0.15	114,192.38 B
ULTJ	2.73	3.31	608.15 B	2,347.51 B	1.80%	0.10	17,364.95 B
CMRY	5.91	9.67	-	2,204.31 B	1.66%	-	33,484.36 B
AALI	0.59	0.93	4,005.05 B	2,089.51 B	5.96%	0.18	12,943.53 B
SMAR	0.59	0.88	15,872.61 B	1,958.72 B	9.69%	0.85	10,971.78 B
UCID	0.71	1.17	879.00 M	1,852.97 B	1.59%	-	3,969.53 B
MLPL	0.22	0.58	2,740.53 B	1,513.62 B	-	0.63	956.62 B
JPFA	1.02	1.52	13,351.51 B	1,444.54 B	4.35%	1.01	13,485.56 B
BISI	1.63	1.48	839.00 M	1,111.84 B	4.04%	-	5,055.00 B
PANI	6.34	46.89	645.73 B	1,069.81 B	-	0.05	82,042.54 B
STAA	1.95	2.53	1,457.72 B	1,057.05 B	5.61%	0.32	8,940.77 B
UNVR	30.80	39.57	-	1,020.60 B	4.91%	-	104,149.50 B
DMND	1.35	1.59	347.36 B	987.95 B	-	0.06	7,574.69 B
TGKA	2.98	3.78	92.33 B	909.96 B	5.42%	0.05	6,016.13 B
FISH	1.07	1.21	2,600.44 B	896.47 B	8.47%	1.05	2,664.00 B
WIIM	2.06	1.07	4.30 B	800.99 B	4.17%	-	3,674.78 B
TAPG	1.04	1.53	1,500.19 B	782.18 B	6.97%	0.14	10,819.63 B

First Previous Next Last

*Market Data (Technical) di update setiap 15 menit. (Contoh: Price, Volume, Value)
*Fundamental Data di update setiap akhir hari pukul 06.30pm. (Contoh: Revenue, Net Profit)

Gambar 3.4 Pre-processing data (filtering)

Setelah *data filtering* selesai dilakukan, proses selanjutnya adalah memastikan agar keseluruhan nilai pada data memiliki rentang yang sama. Hal ini dapat diperoleh dengan melakukan normalisasi pada dataset melalui

min-max normalization. Perhitungan manual untuk proses ini dilakukan di Microsoft Excel seperti pada Gambar 3.5 berikut :

simbol	return_on_equity	net_profit_margin	debt_equity_ratio	dividend_yield	per	pbv
INDF	0.8132	0.3001	0.0126	0.2601	0.1254	0.0062
ICBP	0.8218	0.3217	0.0114	0.1092	0.1343	0.0242
SIMP	0.7983	0.2846	0.0061	0.2795	0.1271	0.0009
GGRM	0.8054	0.2880	0.0033	0.3954	0.1259	0.0034
LSIP	0.8014	0.3396	0.0004	0.4202	0.1270	0.0023
HMSP	0.8300	0.2948	0.0080	0.4236	0.1320	0.0266
CPIN	0.8084	0.2890	0.0034	0.1387	0.1430	0.0199
MYOR	0.8196	0.3070	0.0044	0.0952	0.1375	0.0271
AMRT	0.8232	0.2771	0.0129	0.0583	0.1615	0.0726
ULTJ	0.8176	0.3281	0.0013	0.1206	0.1355	0.0198
CMRY	0.8214	0.3341	0.0009	0.1113	0.1471	0.0449
AALI	0.7990	0.2866	0.0020	0.0818	0.1321	0.0029
SMAR	0.7977	0.2703	0.0089	1.0000	0.1359	0.0029
UCID	0.8028	0.2831	0.0040	0.1066	0.1287	0.0039
MLPL	0.7985	0.2731	0.0161	0.0000	0.1241	0.0000
JPFA	0.8049	0.2760	0.0129	0.2916	0.1305	0.0063
BISI	0.8066	0.3360	0.0004	0.2708	0.1350	0.0111
STAA	0.8121	0.3174	0.0035	0.3760	0.1337	0.0137
UNVR	0.9707	0.3163	0.0330	0.1548	0.1424	0.2413
DMND	0.7997	0.2793	0.0020	0.0000	0.1464	0.0089
TGKA	0.8215	0.2789	0.0095	0.0308	0.1330	0.0218
EISH	0.8085	0.2718	0.0177	0.5677	0.1282	0.0067

Gambar 3.5 Data setelah *min-max normalization*

Terlihat pada dataset yang telah dinormalisasi bahwa atribut rasio keuangan untuk masing-masing data juga telah disertakan seperti *Return on Equity*, *Net Profit Margin*, *Debt to Equity Ratio*, *Dividend Yield*, PER, serta PBV. Langkah normalisasi ini dilakukan sehingga dataset untuk proses pelatihan dan pengujian memiliki rentang nilai yang sama dan tidak ada ketimpangan antar nilai pada data tersebut.

3.5. Clustering

Langkah selanjutnya setelah memastikan keseluruhan nilai pada data seimbang dan relevan maka dilakukan proses *clustering* dengan algoritma K-Means. Dalam prosesnya, nilai *k* harus diketahui terlebih dahulu agar hasil *clustering* sesuai dengan tujuan yang diinginkan. Penentuan nilai *k* pada penelitian ini menggunakan evaluasi dari metode Elbow dengan rentang *k* dari 2-5 sehingga pada proses algoritma K-Means nantinya akan menggunakan nilai *k* dari metode tersebut untuk setiap prosesnya.

Untuk menghindari nilai pemilihan yang acak pada inisialisasi data untuk pemilihan *centroid* awal pada proses K-Means, optimisasi dengan K-Means++ dibutuhkan sehingga data-data yang digunakan sebagai *centroid*

dalam algoritma K-Means tidak sepenuhnya dipilih secara acak. Hasil dari proses K-Means ini akan memberikan kesimpulan berupa dataset yang telah memiliki label atau kelas pada keseluruhan datanya.

3.6. Klasifikasi & Training Model

Dataset yang telah memiliki label atau kelas dari proses sebelumnya kemudian akan dilatih dan diuji menggunakan algoritma klasifikasi Naïve Bayes. Pelatihan dan pengujian ini bisa dilakukan karena dataset telah memiliki label, dan pada penelitian ini proses pemisahan data latih dan data uji dari keseluruhan dataset akan dibagi menjadi 4 skenario rasio, yaitu 90%, 80%, 70%, dan 60% untuk data latih. Sementara data pengujian juga akan dibagi menjadi 10%, 20%, 30%, dan 40%. Hasil dari masing-masing persentase rasio pelatihan dan pengujian kemudian akan dicari nilai rata-ratanya sehingga didapatkan kesimpulan nilai akurasi rata-rata dan evaluasi dari data tersebut.

3.7. Hasil Training

Untuk dapat mengetahui kesimpulan performa model selama proses pembelajaran maka dilakukan evaluasi pada model yang telah dibangun dengan menggunakan *confusion matrix* sehingga dapat diperoleh nilai akurasi dari model tersebut. Proses ini penting dilakukan agar pada klasifikasi data di tahap selanjutnya dapat diketahui apakah model yang akan digunakan relevan dan mampu melakukan tugasnya untuk mendapatkan tujuan yang diharapkan.

3.8. Klasifikasi Dataset

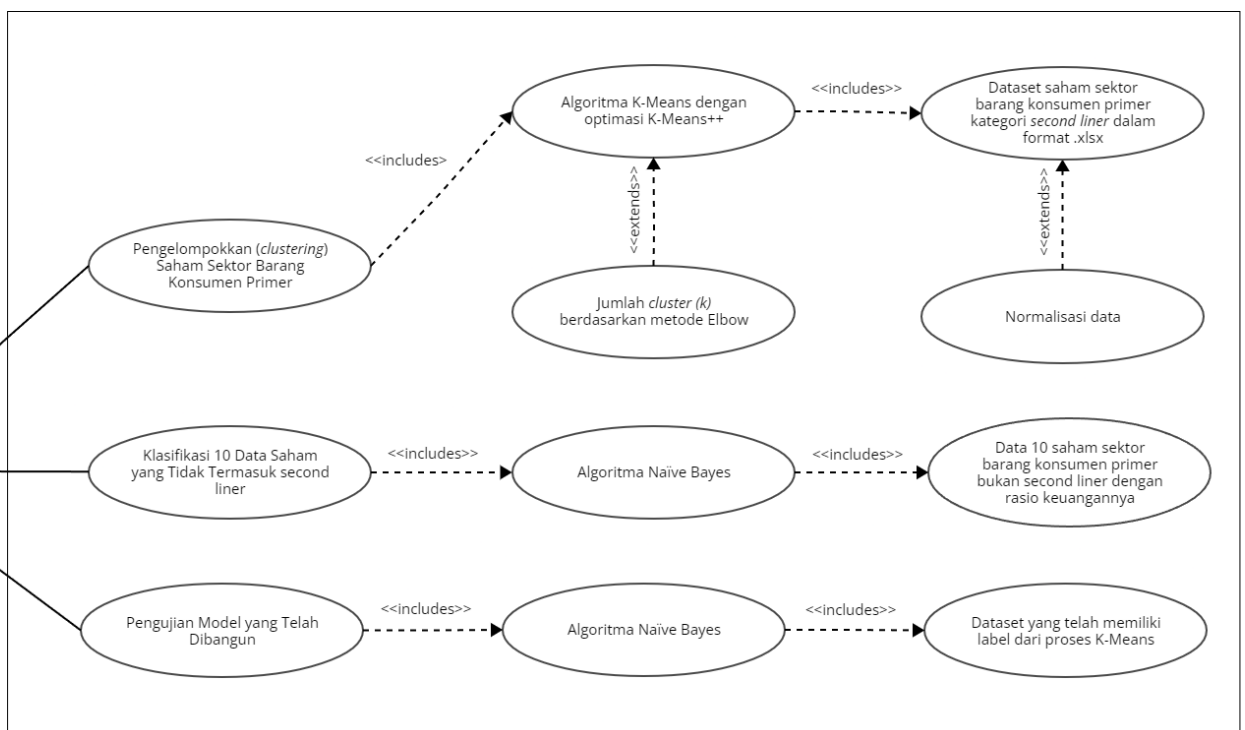
Model yang telah dibangun dan dikembangkan selanjutnya akan digunakan untuk klasifikasi pada 10 data saham yang diluar kategori *second liner* sehingga dapat disimpulkan apakah saham-saham tersebut termasuk ke dalam kelas-kelas tertentu sesuai dengan *output* yang telah ditentukan dari proses pembelajaran. Tiap kelas ini akan memiliki tingkat risiko yang berbeda-beda sesuai dengan atribut rasio keuangan yang dimiliki kelas-kelas tersebut.

3.9. Pemodelan Sistem

Pengetahuan tentang hubungan antar komponen atau objek dalam sistem dan bagaimana interaksi antar komponen tersebut dapat diketahui melalui pemodelan sistem yang dibangun. Pemodelan sistem dalam penelitian ini menggunakan *Use Case Diagram*, *Activity Diagram*, dan *Sequence Diagram*.

3.9.1. Use Case Diagram

Use Case Diagram merupakan jenis diagram yang termasuk dalam *Unified Modeling Language (UML)* yaitu bahasa pemodelan sistem yang menggambarkan hubungan dan interaksi antara aktor (pengguna atau sistem lain) dengan sistem yang dirancang. Umumnya aktor dideskripsikan sebagai komponen atau objek di luar sistem yang memiliki interaksi dengan sistem tersebut, sedangkan *use case* menjelaskan tindakan aktor yang dapat dilakukan pada sistem. *Use case diagram* sistem dapat dilihat pada Gambar 3.6 berikut :



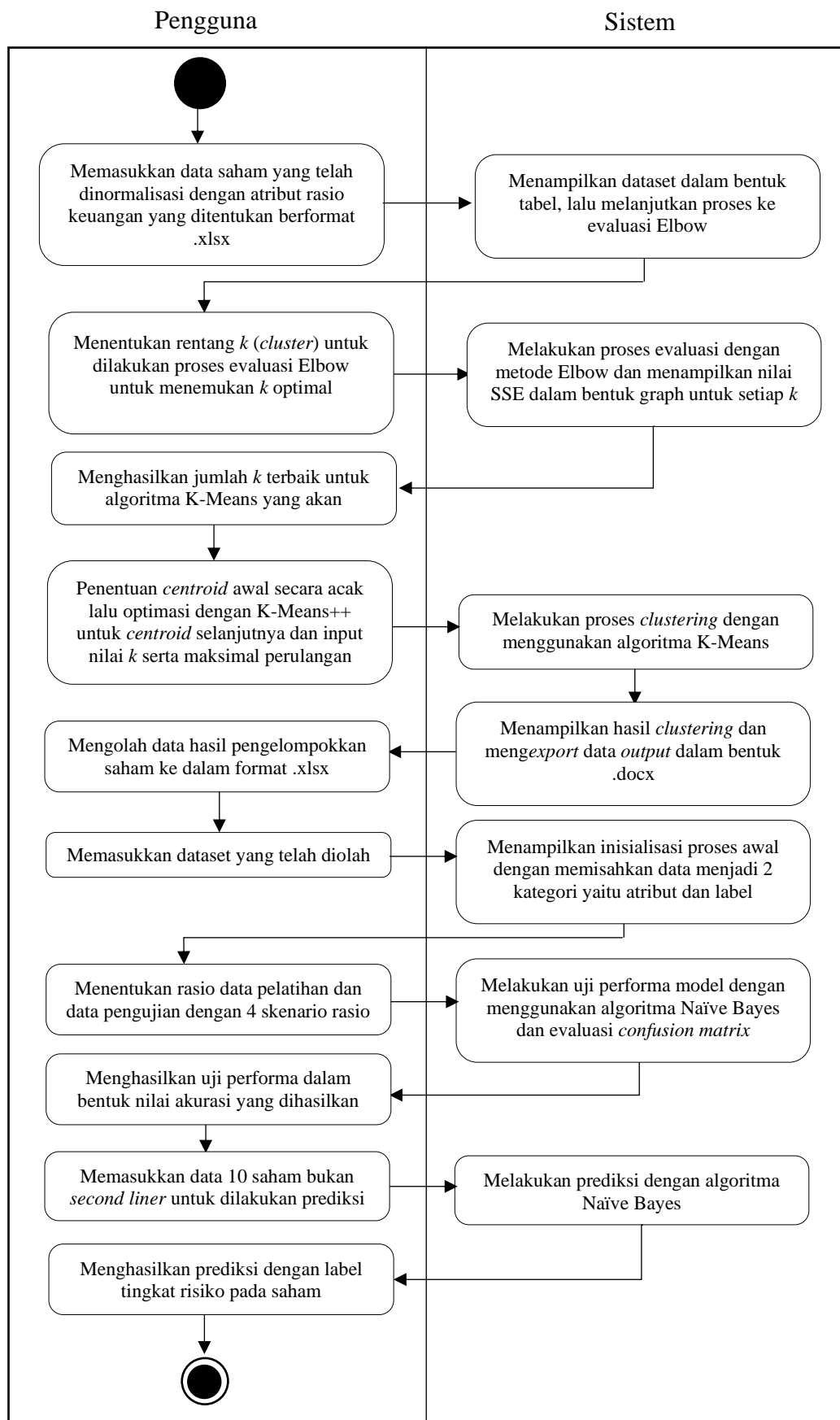
Gambar 3.6 *Use Case Diagram*

Representasi hubungan antara aktor dengan *use case* digambarkan dalam bentuk garis penghubung, yang berupa garis lurus, sementara hubungan antara satu *use case* dengan *use case* lain yang independen dihubungkan dengan garis putus-putus yang disebut sebagai *extend*. Jika suatu *use case* membutuhkan kondisi tertentu untuk dijalankan agar dapat terhubung ke *use case* lainnya maka penghubung interaksi ini direpresentasikan dengan garis putus-putus yang dinamakan *include*.

Pada gambar yang terlampir dijelaskan bahwa sistem dalam penelitian dapat digunakan oleh aktor (pengguna) yang selanjutnya aktor dapat memulai proses pada sistem dengan memasukkan dataset saham pada sektor barang konsumen primer yang termasuk dalam kategori *second iner* dalam format .xlsx untuk kemudian diproses dengan algoritma K-Means dan algoritma Naïve Bayes sehingga diperoleh hasil akhir berupa identifikasi tingkat risiko pada saham-saham dalam penelitian.

3.9.2. Activity Diagram

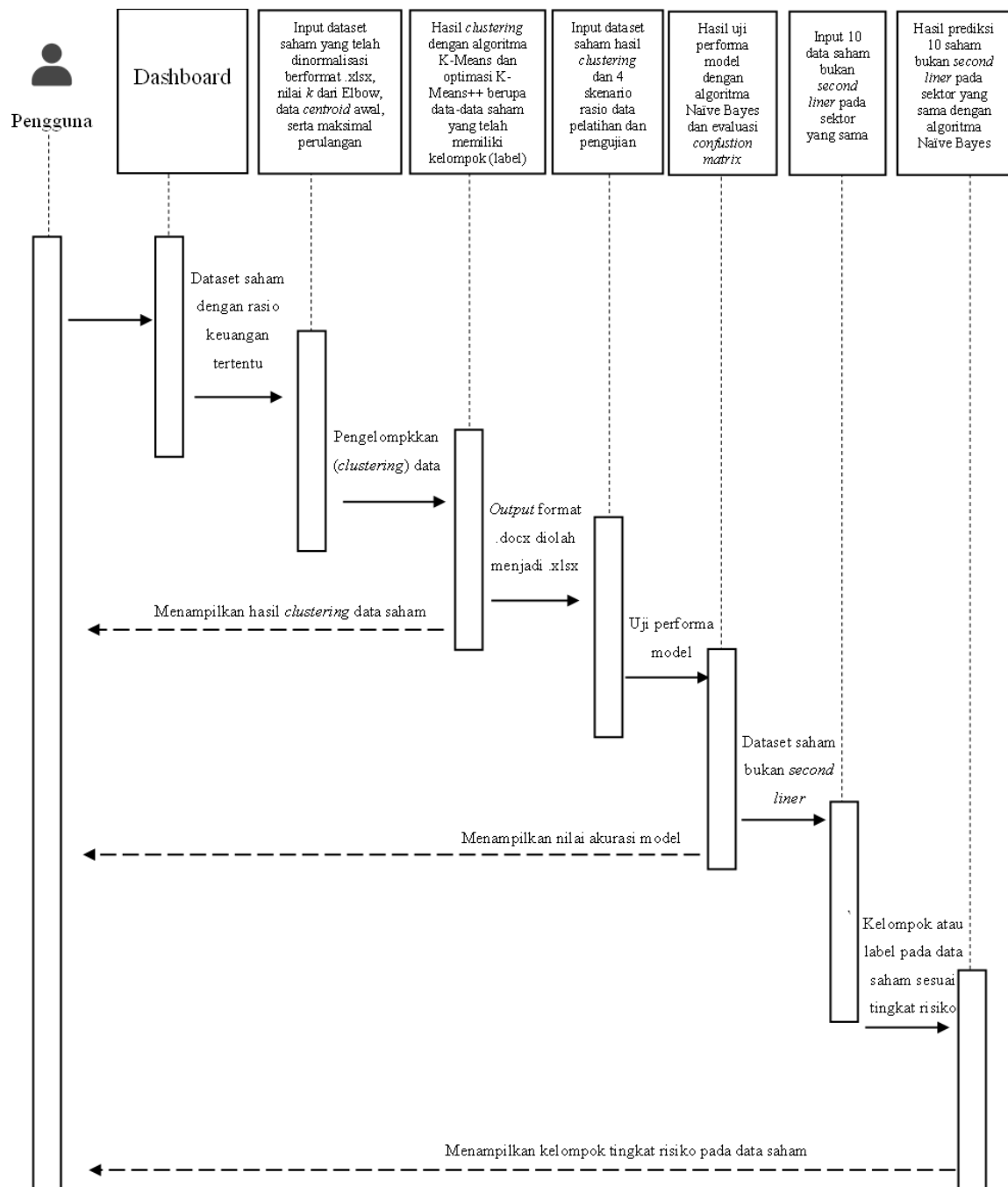
Keseluruhan alur kerja dan urutan aktivitas dalam sistem digambarkan oleh diagram UML melalui *activity diagram*. Diagram ini menjelaskan langkah-langkah yang dijalankan oleh objek dalam sistem tersebut sesuai alur kerjanya dari awal sampai akhir. *Activity Diagram* memberikan pemahaman yang lengkap dan terstruktur mengenai keputusan yang diambil, alur kontrol serta urutan aktivitas dalam sistem. Representasi diagram ini dapat dilihat pada Gambar 3.7 :



Gambar 3.7 Activity Diagram

3.9.3. Sequence Diagram

Jenis diagram UML lain yang merepresentasikan interaksi antar objek dalam sistem adalah *sequence diagram*. Diagram ini merepresentasikan serangkaian pesan yang saling dikirim dan dipertukarkan oleh objek-objek dalam sistem dalam urutan waktu tertentu. *Sequence Diagram* sistem dalam penelitian dapat dilihat pada Gambar 3.8 berikut :



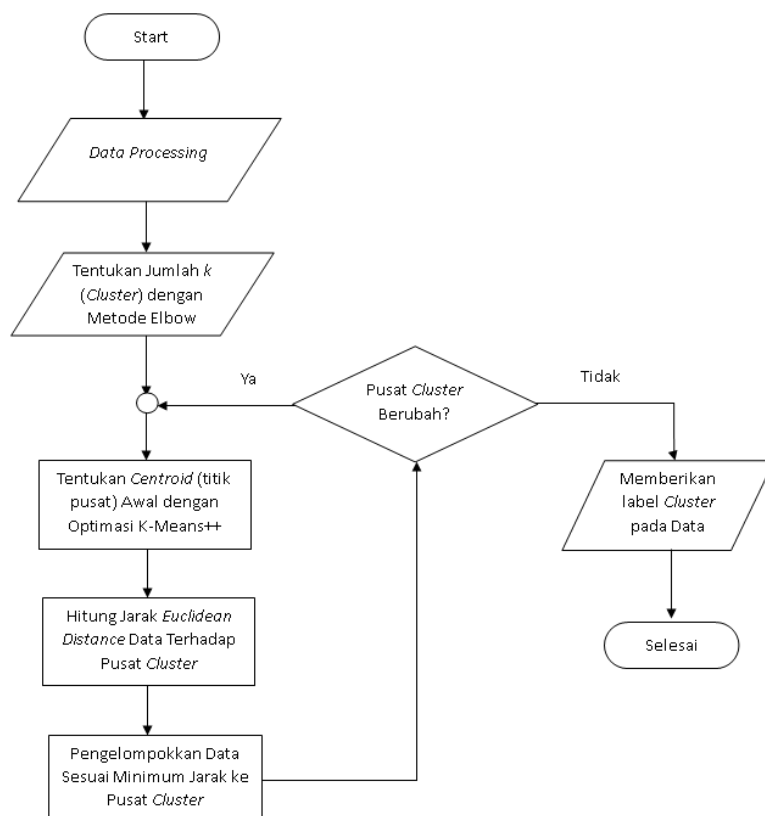
Gambar 3.8 *Sequence Diagram*

3.10. Perancangan Sistem

Perancangan sistem bertujuan untuk memberikan gambaran secara menyeluruh mengenai komponen dan objek yang akan saling berinteraksi dalam sistem sehingga menciptakan struktur yang sistematis pada aplikasi yang dibangun. Dalam proses perancangan sistem di penelitian ini melibatkan proses pengolahan data dengan algoritma yang telah ditentukan yaitu algoritma *clustering* menggunakan K-Means dengan optimasi K-Means++ serta algoritma klasifikasi menggunakan Naïve Bayes, sehingga proses pengolahan akan dibedakan menjadi 3 bagian yaitu proses pengelompokkan menggunakan K-Means, klasifikasi menggunakan Naïve Bayes, serta keseluruhan proses integrasi K-Means dengan Naïve Bayes.

3.10.1. Clustering dengan K-Means

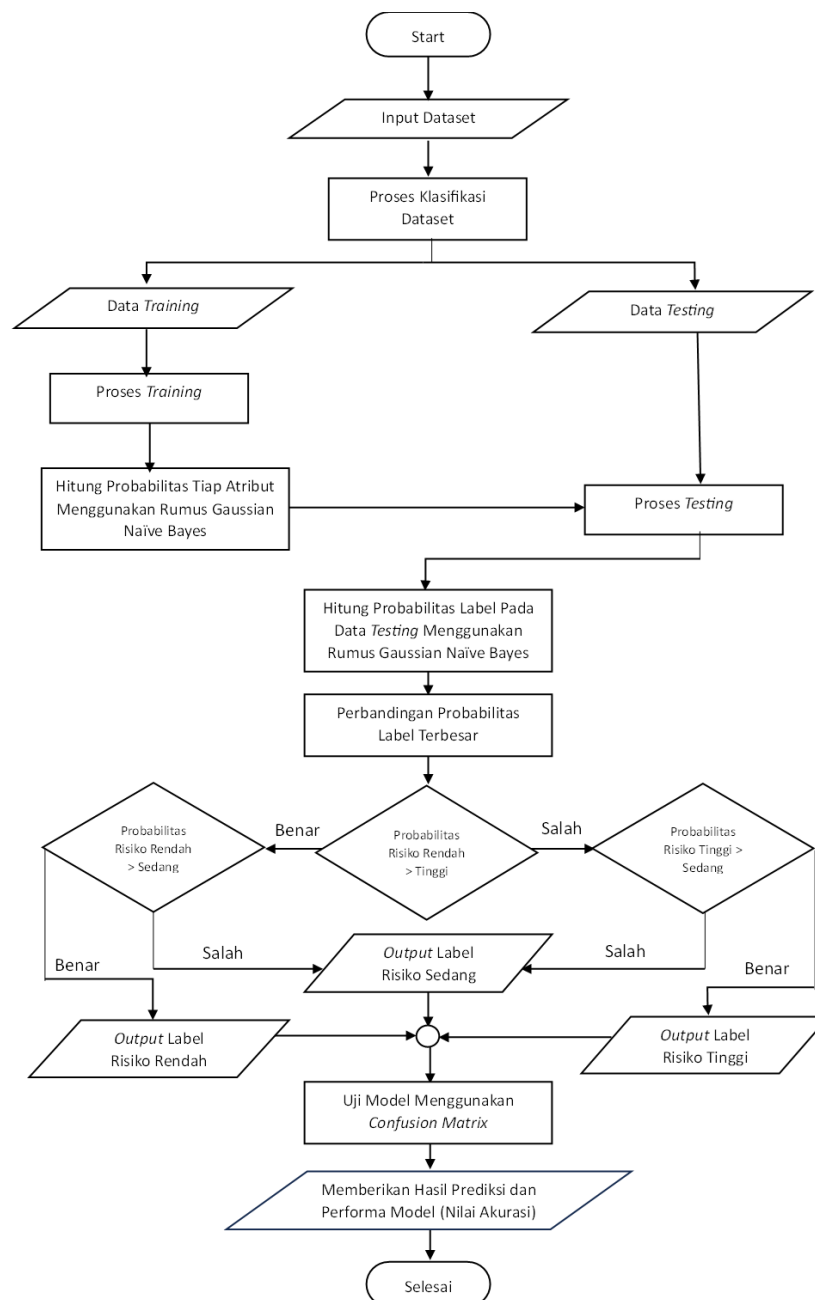
Proses pengelompokkan data dengan menggunakan K-Means pada penelitian adalah membentuk sebanyak k cluster berdasarkan hasil evaluasi dari metode Elbow. Penjelasan proses ini dapat dilihat pada Gambar 3.9 berikut :



Gambar 3.9 Flowchart Algoritma K-Means

3.10.2. Klasifikasi dengan Naïve Bayes

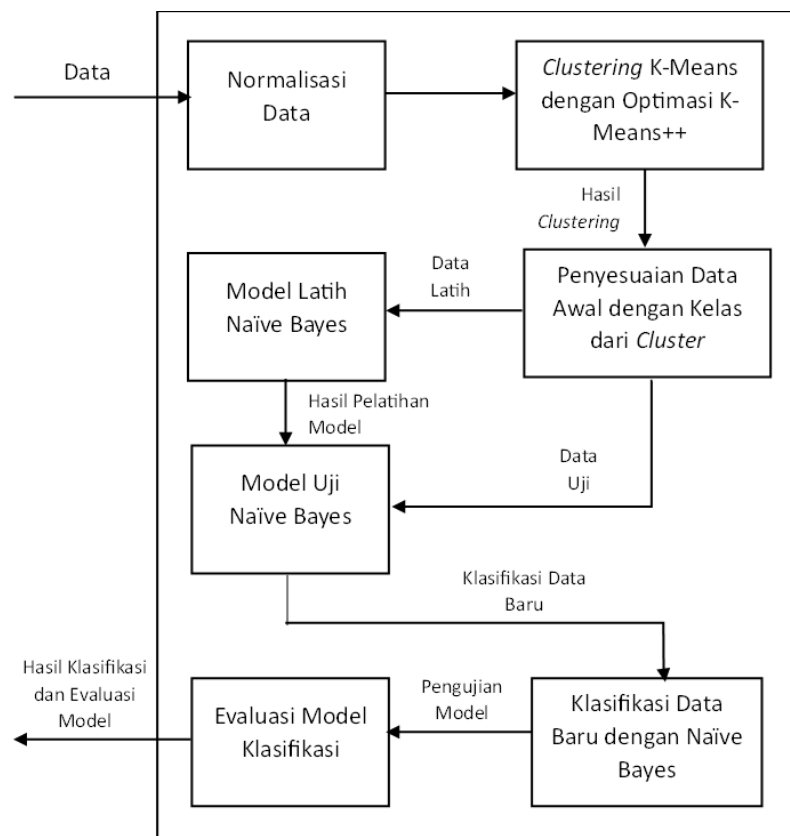
Setelah dilakukan clustering selanjutnya data akan diklasifikasi menggunakan algoritma klasifikasi Naïve Bayes. Proses klasifikasi ini digunakan untuk 10 data saham yang berada di luar kategori second liner sehingga data-data tersebut dapat terklasifikasi ke dalam label tingkat risiko yang telah ditentukan selama proses pembelajaran. Proses algoritma ini dapat dilihat pada Gambar 3.10 berikut :



Gambar 3.10 Flowchart Algoritma Naïve Bayes

3.10.3. Integrasi K-Means dengan Naïve Bayes

Keseluruhan proses ini kemudian diintegrasikan untuk didapatkan hasil penelitian yang sesuai dengan tujuan akhirnya, dimana proses pengelompokkan data yang dilakukan dengan algoritma K-Means selanjutnya dilatih dan diuji dengan algoritma Naïve Bayes untuk kemudian dilakukan klasifikasi pada 10 data saham yang berada diluar kategori *second liner* namun tetap berada pada sektor yang sama. Proses ini dapat dilihat pada Gambar 3.11 berikut :



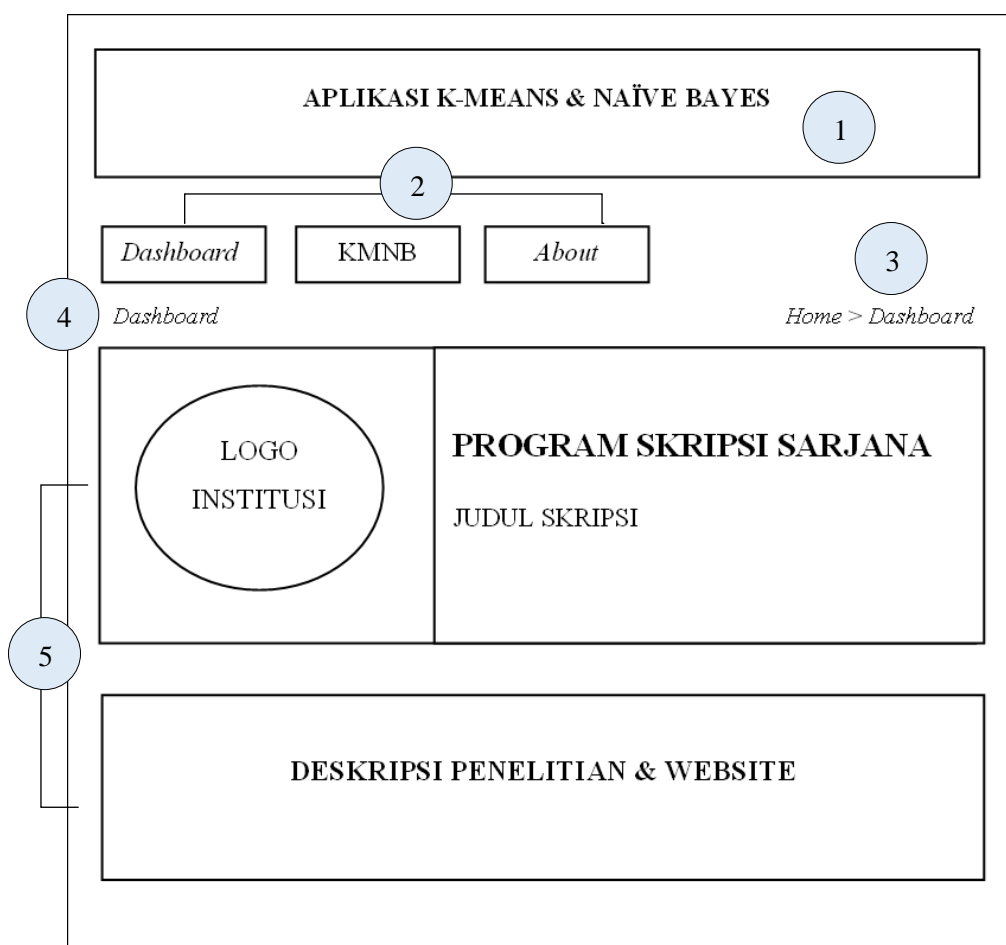
Gambar 3.11 Alur Integrasi K-Means dengan Naïve Bayes

3.11. Perancangan Antarmuka (Interface)

Selanjutnya perancangan sistem juga memiliki rancangan dalam bentuk antarmuka (*interface*) yang akan menggambarkan tampilan sistem serta komponen-komponen yang terdapat dalam sistem . Melalui rancangan antarmuka ini pengguna akan lebih mudah dalam memahami dan menjalankan sistem yang telah dibangun.

3.11.1. Rancangan Antarmuka Dashboard

Halaman *dashboard* adalah halaman pertama yang akan muncul saat pengguna mengakses website. Rancangan antarmuka untuk halaman *dashboard* ini akan menampilkan logo institusi pendidikan penulis, judul penelitian, serta deskripsi singkat mengenai website dan penelitian yang dilakukan. Rancangan untuk halaman ini dapat dilihat pada Gambar 3.12 berikut :



Gambar 3.12 Rancangan Antarmuka *Dashboard*

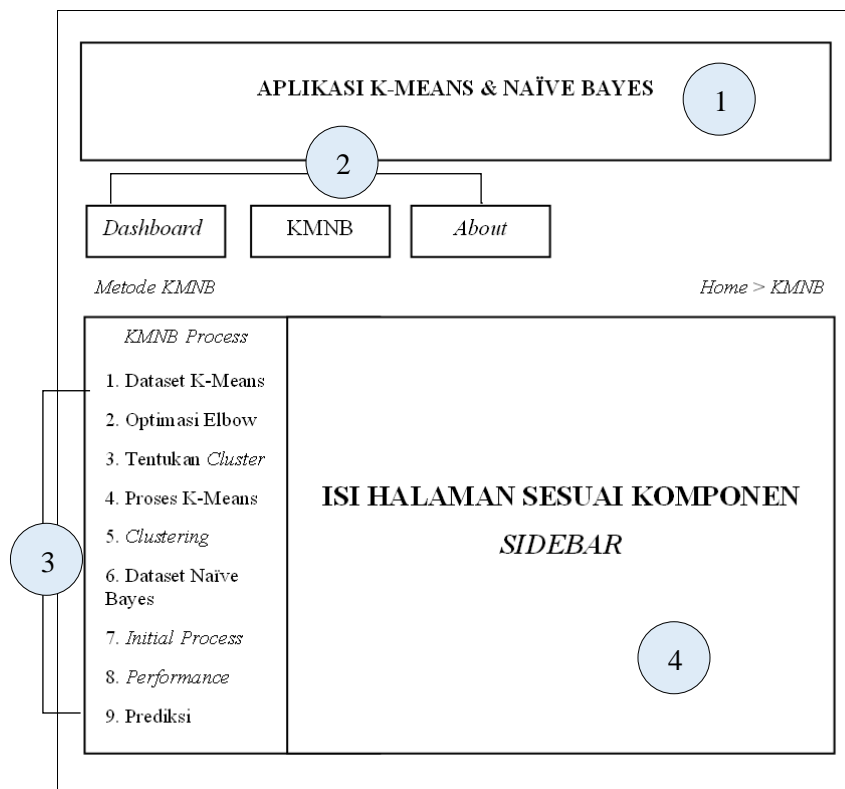
Keterangan tiap komponen adalah sebagai berikut :

1. *Container*, menampilkan judul (*title*) aplikasi yang dibangun,

2. *Navigation bar*, terdiri dari *dashboard*, KMNB (K-Means Naïve Bayes), dan *about*, untuk navigasi menuju halaman lainnya dalam website melalui tombol *navigation* tersebut,
3. *Breadcrumb*, menunjukkan struktur halaman apa saja yang telah dibuka oleh pengguna / pengunjung website,
4. *Page title*, menampilkan judul atau teks bagian halaman yang sedang dikunjungi,
5. *Card body*, menampilkan judul penelitian terkait dan deskripsi singkat tentang penelitian serta sistem.

3.11.2. Rancangan Antarmuka KMNB (K-Means Naïve Bayes)

Bagian inti dari program yang akan dirancang berada pada halaman KMNB, dimana halaman ini menampilkan keseluruhan proses algoritma dan inisialisasi dataset dalam aplikasi yang dibangun. Rancangan antarmuka untuk halaman KMNB dengan komponen terkait tertera pada Gambar 3.13 berikut :



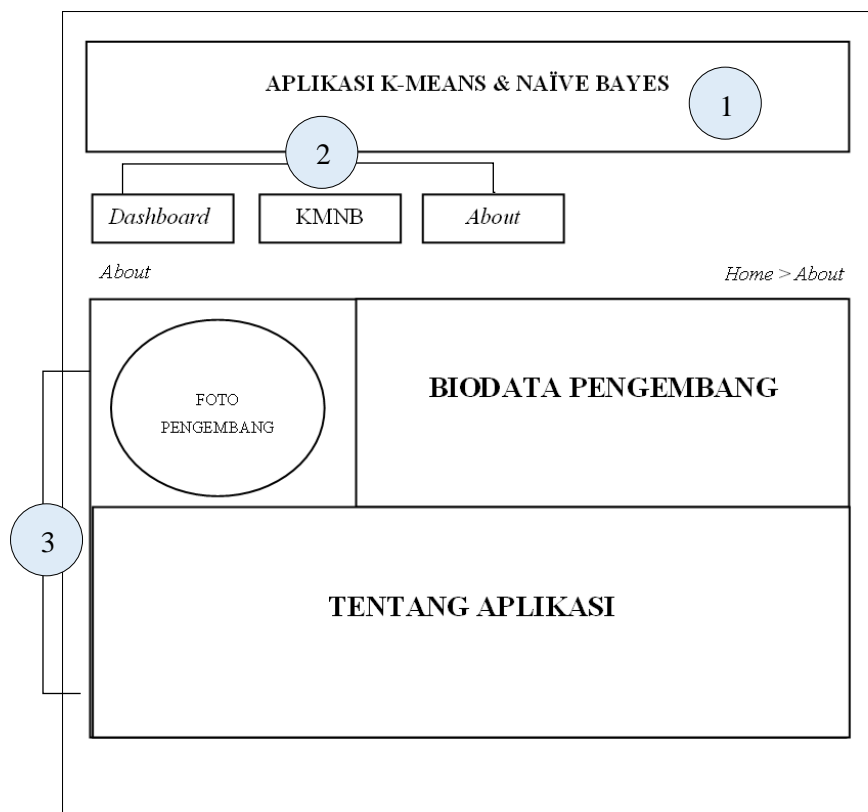
Gambar 3.13 Rancangan Antarmuka KMNB

Dengan keterangan sebagai berikut :

1. *Container*, menampilkan judul (*title*) aplikasi yang dibangun,
2. *Navigation bar*, terdiri dari *dashboard*, KMNB (K-Means Naïve Bayes), dan *about*, untuk fungsi navigasi menuju halaman lainnya dalam website melalui tombol *navigation* tersebut,
3. *Sidebar*, menampilkan daftar (*list*) navigasi tambahan yang akan menampilkan langkah-langkah dalam proses KMNB,
4. *Card body*, menampilkan isi halaman berdasarkan *list* pada *sidebar* yang dipilih pengguna.

3.11.3. Rancangan Antarmuka About

Halaman *about* menjelaskan informasi singkat mengenai pengembang aplikasi dan tentang aplikasi yang dibangun dalam penelitian terkait. Rancangan antarmuka untuk halaman ini dapat dilihat pada Gambar 3.14 berikut :



Gambar 3.14 Rancangan Antarmuka *About*

Keterangan :

1. *Container*, menampilkan judul (*title*) aplikasi yang dibangun,
2. *Navigation bar*, terdiri dari *dashboard*, KMNB (K-Means Naïve Bayes), dan *about*, untuk fungsi navigasi menuju halaman lainnya dalam website melalui tombol *navigation* tersebut,
3. *Card body*, menampilkan biodata pengembang dan informasi tentang aplikasi yang dibangun

BAB 4

IMPLEMENTASI DAN PENGUJIAN SISTEM

4.1. Implementasi Sistem

Metode penelitian yang telah dirancang dalam pembangunan sistem selanjutnya diimplementasikan pada tahap implementasi sistem. Implementasi sistem membutuhkan beberapa komponen perangkat lunak (*software*) dan perangkat keras (*hardware*) agar sistem yang telah dirancang dapat diimplementasikan dengan baik. Perangkat yang dibutuhkan dalam tahapan ini adalah sebagai berikut :

4.1.1 Perangkat Keras (hardware)

Tahap implementasi sistem dalam penelitian ini dilakukan dengan spesifikasi perangkat keras sebagai berikut :

1. Processor : AMD Ryzen 5 PRO 3500U Radeon Vega 2.10 GHz
2. Sistem Operasi : Windows 11 Home 64-bit
3. SSD : 512 GB
4. RAM : 16 GB

4.1.2 Perangkat Lunak (software)

Perangkat lunak yang digunakan dalam pembangunan sistem pada penelitian adalah sebagai berikut :

1. Windows 11 Pro 64-bit
2. PHP 8.0.30
3. CodeIgniter v3.1.11
4. XAMPP Control Panel v3.3.0
5. Visual Studio Code
6. Microsoft Excel 2019

4.2. Implementasi Data

Sampel data yang digunakan dalam penelitian ini adalah 82 saham perusahaan terbuka di bursa efek Indonesia yang termasuk dalam sektor

barang konsumen primer dengan minimal kapitalisasi pasar 500 miliar (*second liner*). Data yang diambil juga melibatkan nilai rasio keuangan yang telah ditentukan dalam penelitian ini seperti nilai *Return on Equity* (ROE), *Net Profit Margin* (NPM), *Debt to Equity Ratio* (DER), *Dividend Yield*, *Price Earning Ratio* (PER), dan *Price to Book Value* (PBV). Sumber dari sampel data ini berasal dari sekuritas *Stockbit* dan aplikasi *RTI Business* yang menyajikan informasi keuangan perusahaan terbuka di bursa efek Indonesia, dengan tanggal pengambilan data adalah pada 29 Februari 2024. Data dilampirkan pada Tabel 4.1 berikut :

Tabel 4.1 Dataset Sistem

simbol	return_on_e quity	net_profit_m argin	debt_equity_ ratio	dividend_ yield	per	pbv
INDF	16.17	8.44	153.18	3.88	6.16	1.00
ICBP	22.99	13.76	138.53	1.63	14.3 1	3.29
SIMP	4.25	4.60	76.67	4.17	7.73	0.33
GGRM	9.91	5.45	44.45	5.9	6.58	0.65
LSIP	6.71	18.19	10.28	6.27	7.57	0.51
HMSP	29.55	7.11	99.17	6.32	12.1 6	3.59
CPIN	12.31	5.68	44.56	2.07	22.2 5	2.74
MYOR	21.24	10.14	57.11	1.42	17.2 2	3.66
AMRT	24.11	2.74	156.82	0.87	39.1	9.42
ULTJ	19.68	15.35	19.96	1.8	15.4 1	2.73
CMRY	22.73	16.83	16.17	1.66	25.9 8	5.91
AALI	4.8	5.09	28.53	1.22	12.2 6	0.59

SMAR	3.73	1.07	109.83	14.92	15.7 6	0.59
UCID	7.81	4.24	52.57	1.59	9.13	0.71
MLPL	4.37	1.75	194.5	0	4.99	0.22
JPFA	9.49	2.48	157.19	4.35	10.7 9	1.02
BISI	10.91	17.29	9.55	4.04	14.9 4	1.63
STAA	15.24	12.71	46.19	5.61	13.7 2	1.95
UNVR	141.99	12.43	392.84	2.31	21.6 9	30.8
DMND	5.32	3.30	28.92	0	25.3 5	1.35
TGKA	22.8	3.19	117.06	0.46	13.0 6	2.98
FISH	12.35	1.45	213.17	8.47	8.65	1.07
PANI	2.09	12.51	113.07	0	303. 82	6.34
WIIM	32.97	11.87	58.63	4.17	6.24	2.06
TAPG	14.19	18.27	26.67	6.97	7.35	1.04
GOOD	14.79	4.53	102.9	1.42	33.0 2	4.88
TCID	-0.55	-0.50	28.73	1.67	- 108. 49	0.59
PSGO	9.55	8.22	121.62	0	16.8 4	1.61
MLBI	73.95	30.51	161.73	1.55	16.2	11.98
ROTI	13.31	8.10	78.39	8.92	24.1 1	3.21
ADES	23.92	26.64	19.3	0	16.2 2	3.88
DLTA	23.39	28.89	31.79	9.62	12.9 4	3.03
TLDN	13.95	8.12	142.74	2.93	19.1 6	2.67
CAMP	15.99	12.85	14.8	6.21	12.6 7	2.02

EPMT	10.15	2.56	51.67	8.07	9.55	0.97
CEKA	10.09	2.59	21.09	5.33	6.85	0.69
DSNG	9.64	8.84	83.63	5.83	6.5	0.63
FAPA	3.93	3.23	118.3	0	128.10	5.03
SGRO	10.91	10.14	87.17	14.40	6.46	0.92
CBUT	18.09	1.67	242.66	0	35.01	6.33
MIDI	14.27	3.02	99.79	0.1	27.36	3.9
SIPD	0.75	0.15	175.49	0	196.76	1.47
MAIN	2.88	0.51	156.27	0	18.32	0.53
MKTR	7.04	6.49	142.51	0.78	39.27	2.77
MGRO	-41.47	-3.43	391.67	0	-10.93	4.53
NSSS	-4.41	-3.81	243.32	0	-97.08	4.29
KINO	5.51	2.14	194.18	0	23.82	1.31
SSMS	11.03	11.61	110.14	13.15	15.8	1.74
MPPA	-121.09	-4.65	1140.79	0	-2.04	2.47
DAYA	-633.45	-2.66	11742.9	0	-20.04	126.95
PMMP	7.36	3.15	247.24	0	8.1	0.59
BUAH	21.11	1.95	102.47	0.77	52.45	11.07
JARR	5.28	1.54	180.29	0	40.99	2.17
TBLA	7.65	3.67	215.3	8.96	6.95	0.53
BUDI	6.99	2.25	123.87	2.44	11.58	0.81
CPRO	12.25	4.75	101.13	0	7.23	0.89
STTP	19.92	18.92	13.74	0	14.22	2.83
KEJU	11.98	7.88	23.5	6.44	21.75	2.61
ANJT	0.62	1.11	44.53	3.86	59.88	0.37
SKLT	11.52	3.96	70.04	1.34	19.41	2.24

MSJA	11.58	9.75	44.85	0	17.2 7	2.01
RANC	-30.36	-3.6	252.13	0	- 9.22	2.8
CSRA	14.67	17.59	69.87	5.56	5.81	0.85
HERO	2.71	0.51	576.71	0	131. 81	3.57
AISA	2.4	1.10	119.14	0	64.6 4	1.55
BWPT	7.51	3.73	378.5	0	10.0 4	0.75
GZCO	1.17	1.76	50.77	0	44.3 4	0.52
BEER	5.85	24.28	5.22	0.46	80.7 1	4.73
WMPP	-22.87	-35.7	247.39	0	- 4.19	0.96
WINE	18.99	19.89	42.45	0.49	21.8 5	4.15
BTEK	-9.88	-65.7	264.56	0	- 20.9 5	2.07
STRK	165.36	181.32	40.08	0	6.81	11.2 6
HOKI	-2.59	-1.42	35.24	0	- 105. 01	2.71
VICI	21	13.97	29.76	0.56	21.8 4	4.59
IBOS	2.56	9.99	12.34	0	805. 57	20.6 7
BEEF	52.11	14.03	446.48	0	23.5 1	12.2 5
PGUN	9.41	17.66	65.33	0	16.5 7	1.56
PTPS	44.40	43.28	40.77	0	40.8 7	18.1 4
JAWA	-391.27	-35.68	4250.2	0	- 6.17	24.1 6
CLEO	19.92	17.76	36.96	0.22	33.1 8	6.61
WMUU	-7.93	-24.38	130.24	0	- 7.25	0.57
TRGU	0.61	0.12	240.58	0.57	276. 03	1.70

4.3. Implementasi *Pre-processing* Data

Sebelum data diproses oleh algoritma untuk dilakukan pengelompokkan dan klasifikasi serta uji model, data perlu dipersiapkan terlebih dulu agar dataset yang digunakan dalam proses pembelajaran berikutnya menjadi lebih baik dan tidak memiliki ketimpangan nilai dalam data tersebut. Pada tahapan ini dilakukan normalisasi data menggunakan *min-max normalization* agar keseluruhan nilai berada pada rentang yang sama sehingga didapatkan data setelah proses normalisasi seperti pada Tabel 4.2 berikut :

Tabel 4.2 Data Setelah Normalisasi

simbol	return_on_e quity	net_profit_m argin	debt_equity_ ratio	dividend_ yield	per	pbv
INDF	0.8132	0.3001	0.0126	0.2601	0.125 4	0.0062
ICBP	0.8218	0.3217	0.0114	0.1092	0.134 3	0.0242
SIMP	0.7983	0.2846	0.0061	0.2795	0.127 1	0.0009
GGRM	0.8054	0.2880	0.0033	0.3954	0.125 9	0.0034
LSIP	0.8014	0.3396	0.0004	0.4202	0.127 0	0.0023
...
PTPS	0.8486	0.4412	0.0030	0.0000	0.163 4	0.1414
JAWA	0.3032	0.1215	0.3617	0.0000	0.111 9	0.1889
CLEO	0.8179	0.3379	0.0027	0.0147	0.155 0	0.0504
WMUU	0.7831	0.1673	0.0107	0.0000	0.110 8	0.0028
TRGU	0.7938	0.2665	0.0201	0.0382	0.420 7	0.0117

4.4. Implementasi *Clustering*

Proses pengelompokkan data pada penelitian ini menggunakan algoritma K-Means dengan optimasi K-Means++ dan jumlah k (*cluster*) pada algoritma ditentukan dari hasil evaluasi metode Elbow dengan rentang nilai k dari 2-5 dan maksimal perulangan 10 kali sehingga dapat ditemukan jumlah k yang optimal dalam penelitian. Metode Elbow selanjutnya diterapkan pada dataset yang telah dinormalisasi dan diupload pada website dengan format .xlsx seperti terlampir pada Gambar 4.1. Hasil evaluasi dari metode Elbow mendapatkan hasil yang terbaik pada $k = 3$ dengan sudut siku yang menonjol dan selisih nilai SSE yang signifikan seperti terlihat pada Gambar 4.2.

APLIKASI K-MEANS & NAÏVE BAYES

Dashboard

KMNB

About

Metode KMNB

Home > KMNB

K-Means & Elbow Menu

1. Dataset Kmeans

2. Optimasi Elbow

3. Tentukan Cluster

4. Proses K-Means

5. Clustering

6. Dataset Naive Bayes

7. Initial Process

8. Performance

9. Prediksi

Pilih Data Excel

Download contoh Format.xlsx

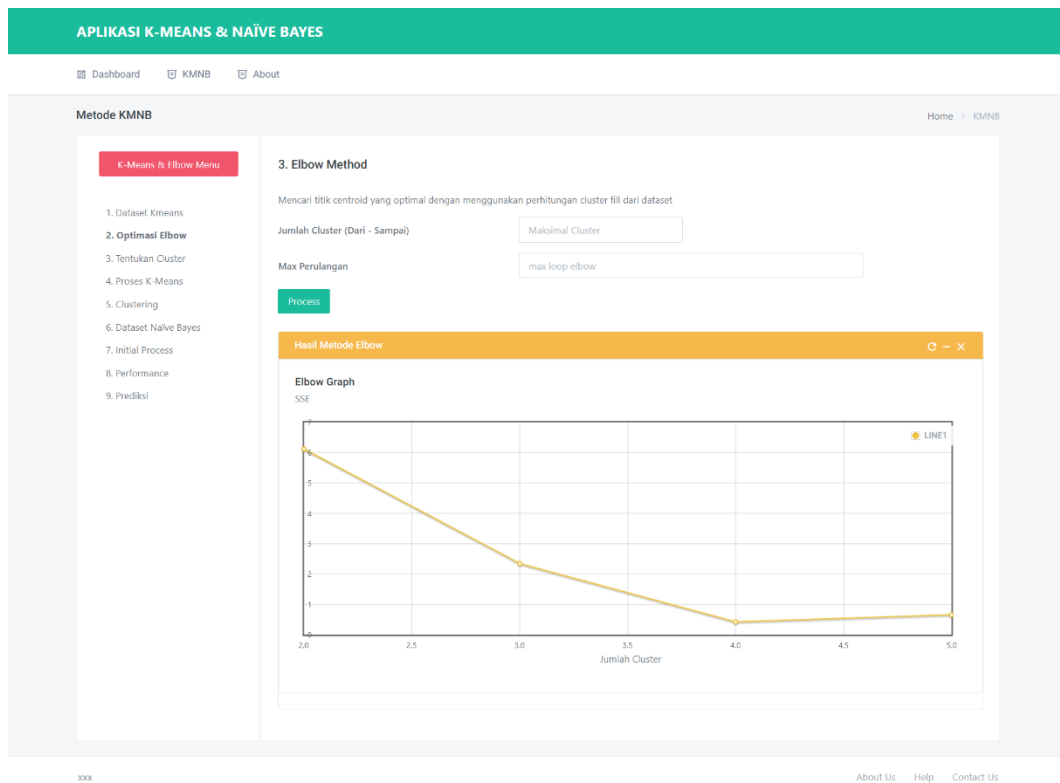
Choose File

No file chosen

Dataset K-Means & Elbow

simbol	return_on_equity	net_profit_margin	debt_equity_ratio	dividend_yield	per	pbv
INDF	0.813234687	0.300137641	0.012605557	0.260053619	0.125429403	0.006154817
ICBP	0.821772386	0.321674358	0.01135744	0.10924933	0.134345667	0.02422473
SIMP	0.79831249	0.284592341	0.006087234	0.279490617	0.127147014	0.000867987
GGRM	0.80539803	0.288033358	0.003342228	0.395442359	0.125888891	0.00339304

Gambar 4.1 Upload dataset setelah normalisasi



Gambar 4.2 Hasil Metode Elbow

Setelah nilai k optimum ditemukan, selanjutnya adalah proses inisialisasi *centroid* awal pada K-Means dengan optimasi K-Means++. Proses optimasi ini dilakukan di Microsoft Excel 2019 dengan pemilihan data awalnya adalah pada saham CMRY yang memiliki kinerja operasional dan potensi pertumbuhan yang tinggi berdasarkan rasio keuangannya, Dari data CMRY kemudian ditentukan *centroid* awal selanjutnya melalui optimasi K-Means++ yaitu dengan menghitung rata-rata selisih kuadrat jarak setiap atribut data dengan data *centroid* pertama (CMRY) menggunakan Microsoft Excel sehingga didapatkan data akhir untuk *centroid* awal seperti pada Tabel 4.3 berikut :

Tabel 4.3 Optimasi dengan K-Means++

x						$\min(d(x, z_i)^2)$	<i>simbol</i>	<i>prob</i>
0.813 2	0.300 1	0.012 6	0.260 1	0.125 4	0.006 2	0.0255	INDF	0.003
0.821 8	0.321 7	0.011 4	0.109 2	0.134 3	0.024 2	0.0009	ICBP	0.000
0.798 3	0.284 6	0.006 1	0.279 5	0.127 1	0.000 9	0.0337	SIMP	0.003
0.805 4	0.288 0	0.003 3	0.395 4	0.125 9	0.003 4	0.0853	GGRM	0.009
0.801 4	0.339 6	0.000 4	0.420 2	0.127 0	0.002 3	0.0981	LSIP	0.010
0.830 0	0.294 8	0.008 0	0.423 6	0.132 0	0.026 6	0.0998	HMSP	0.010
0.808 4	0.289 0	0.003 4	0.138 7	0.143 0	0.019 9	0.0036	CPIN	0.000
0.819 6	0.307 0	0.004 4	0.095 2	0.137 5	0.027 1	0.0014	MYOR	0.000
0.823 2	0.277 1	0.012 9	0.058 3	0.161 5	0.072 6	0.0072	AMRT	0.001
0.817 6	0.328 1	0.001 3	0.120 6	0.135 5	0.019 8	0.0009	ULTJ	-
0.821 4	0.334 1	0.000 9	0.111 3	0.147 1	0.044 9	0.0000	CMRY	0.000
0.799 0	0.286 6	0.002 0	0.081 8	0.132 1	0.002 9	0.0056	AALI	0.001
0.797 7	0.270 3	0.008 9	1.000 0	0.135 9	0.002 9	0.7964	SMAR	0.080
0.802 8	0.283 1	0.004 0	0.106 6	0.128 7	0.003 9	0.0050	UCID	0.000
0.798 5	0.273 1	0.016 1	0.000 0	0.124 1	0.000 0	0.0194	MLPL	0.002
0.804 9	0.276 0	0.012 9	0.291 6	0.130 5	0.006 3	0.0381	JPFA	0.004
0.806 6	0.336 0	0.000 4	0.270 8	0.135 0	0.011 1	0.0270	BISI	0.003
0.812 1	0.317 4	0.003 5	0.376 0	0.133 7	0.013 7	0.0716	STAA	0.007
0.970 7	0.316 3	0.033 0	0.154 8	0.142 4	0.241 3	0.0641	UNVR	0.006
0.799 7	0.279 3	0.002 0	0.000 0	0.146 4	0.008 9	0.0172	DMND	0.002
0.821 5	0.278 9	0.009 5	0.030 8	0.133 0	0.021 8	0.0103	TGKA	0.001
0.808 5	0.271 8	0.017 7	0.567 7	0.128 2	0.006 7	0.2145	FISH	0.021

0.795 6	0.316 6	0.009 2	0.000 0	0.451 1	0.048 3	0.1058	PANI	0.011
0.834 3	0.314 0	0.004 6	0.279 5	0.125 5	0.014 5	0.0303	WIIM	0.003
0.810 8	0.339 9	0.001 8	0.467 2	0.126 7	0.006 5	0.1287	TAPG	0.013
0.811 5	0.284 3	0.008 3	0.095 2	0.154 8	0.036 8	0.0030	GOOD	0.000
0.792 3	0.263 9	0.002 0	0.111 9	0.000 0	0.002 9	0.0292	TCID	0.003
0.804 9	0.299 2	0.009 9	0.000 0	0.137 1	0.011 0	0.0152	PSGO	0.002
0.885 6	0.389 5	0.013 3	0.103 9	0.136 4	0.092 8	0.0098	MLBI	0.001
0.809 7	0.298 8	0.006 2	0.597 9	0.145 1	0.023 6	0.2386	ROTI	0.024
0.822 9	0.373 8	0.001 2	0.000 0	0.136 4	0.028 9	0.0143	ADES	0.001
0.822 3	0.382 9	0.002 3	0.644 8	0.132 8	0.022 2	0.2877	DLTA	0.029
0.810 5	0.298 8	0.011 7	0.196 4	0.139 7	0.019 3	0.0094	TLDN	0.001
0.813 0	0.318 0	0.000 8	0.416 2	0.132 6	0.014 2	0.0945	CAMP	0.009
0.805 7	0.276 3	0.004 0	0.540 9	0.129 1	0.005 9	0.1900	EPMT	0.019
0.805 6	0.276 5	0.001 4	0.357 2	0.126 2	0.003 7	0.0662	CEKA	0.007
0.805 1	0.301 8	0.006 7	0.390 8	0.125 8	0.003 2	0.0817	DSNG	0.008
0.797 9	0.279 0	0.009 6	0.000 0	0.258 8	0.038 0	0.0286	FAPA	0.003
0.806 6	0.307 0	0.007 0	0.965 1	0.125 8	0.005 5	0.7321	SGRO	0.073
0.815 6	0.272 7	0.020 2	0.000 0	0.157 0	0.048 2	0.0167	CBUT	0.002
0.810 9	0.278 2	0.008 1	0.006 7	0.148 6	0.029 0	0.0145	MIDI	0.001
0.793 9	0.266 6	0.014 5	0.000 0	0.333 9	0.009 9	0.0540	SIPD	0.005
0.796 6	0.268 0	0.012 9	0.000 0	0.138 7	0.002 4	0.0194	MAIN	0.002
0.801 8	0.292 2	0.011 7	0.052 3	0.161 7	0.020 1	0.0066	MKTR	0.001
0.741 1	0.252 1	0.032 9	0.000 0	0.106 7	0.034 0	0.0283	MGRO	0.003
0.787 5	0.250 5	0.020 3	0.000 0	0.012 5	0.032 1	0.0392	NSSS	0.004

0.799 9	0.274 6	0.016 1	0.000 0	0.144 7	0.008 6	0.0179	KINO	0.002
0.806 8	0.313 0	0.008 9	0.881 4	0.136 0	0.012 0	0.5950	SSMS	0.059
0.641 4	0.247 1	0.096 7	0.000 0	0.116 5	0.017 8	0.0632	MPPA	0.006
0.000 0	0.255 2	1.000 0	0.000 0	0.096 8	1.000 0	2.6063	DAYA	0.260

Tabel berwarna kuning adalah data *centroid* pertama yang diinisiasi berdasarkan pertimbangan rasio keuangan yang baik, sedangkan *centroid* selanjutnya dipilih dari nilai probabilitas terbesar yaitu DAYA dengan nilai 0.260 dan SMAR dengan nilai 0.080.

Setelah menentukan *centroid* awalnya, proses *clustering* dilanjutkan dengan algoritma K-Means dengan *centroid* yang sudah ditentukan dari proses K-Means++ sehingga didapatkan hasil akhir proses *clustering* pada sistem seperti pada Gambar 4.3 berikut :

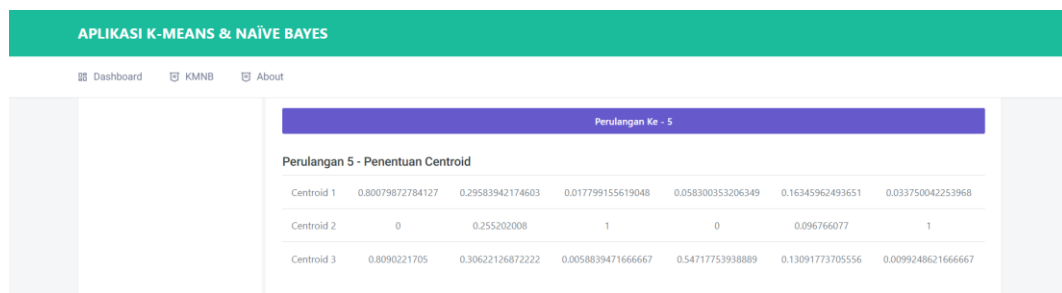
APLIKASI K-MEANS & NAÏVE BAYES							
Dashboard	KMNB	About					
SSMS	0.806800115	0.31297061	0.008938734	0.881367292	0.135975757	0.011994003	3
TBLA	0.802568821	0.280827463	0.017897915	0.600536193	0.126293679	0.002446145	3
KEJU	0.807989384	0.297870618	0.001557378	0.431635389	0.142485176	0.018858992	3
CSRA	0.811356893	0.337179176	0.005507903	0.372654155	0.125046496	0.004971199	3
Jumlah Cluster							
Cluster	Jumlah						
1	63						
3	18						
2	1						

Gambar 4.3 Hasil Akhir *Clustering*

Proses *clustering* telah selesai dijalankan dengan pengelompokkan masing-masing data ke dalam *cluster* terdekat, seperti pada *cluster* 1 yang memiliki 63 data, *cluster* 2 dengan 1 data dan *cluster* 3 yang memiliki 18 data. Dilihat dari rasio keuangan yang dimiliki tiap data dari *cluster* tersebut, dapat disimpulkan bahwa *cluster* 1 merupakan saham yang memiliki tingkat risiko sedang dikarenakan rasio keuangannya yang tidak terlalu baik juga tidak terlalu buruk, sedangkan *cluster* 2 merupakan saham berisiko tinggi,

disebabkan rasio keuangannya yang buruk bahkan memiliki nilai negatif yang besar daripada saham lainnya, sementara itu *cluster* 3 merupakan kelompok saham berisiko rendah karena memiliki rasio keuangan yang lebih baik daripada *cluster* lainnya.

Perulangan pada proses *clustering* yang berjalan di sistem ini berhenti pada perulangan ke-5, dengan data akhir *centroid* ditunjukkan pada Gambar 4.4 berikut :

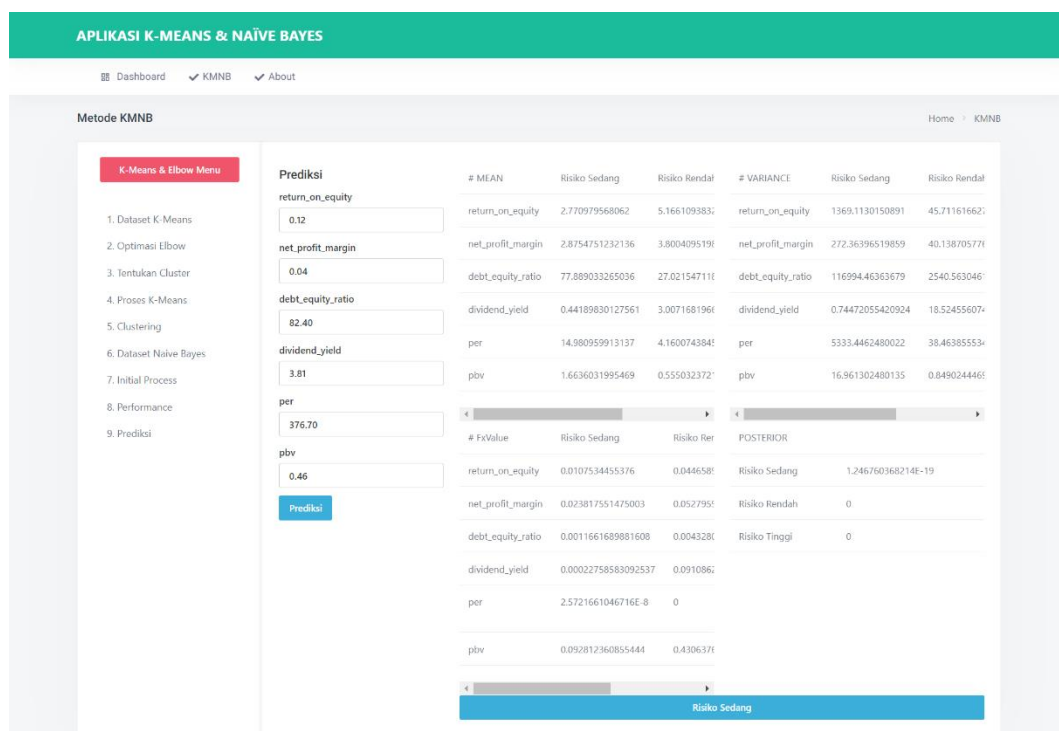


Perulangan Ke - 5						
Perulangan 5 - Penentuan Centroid						
Centroid 1	0.80079872784127	0.29583942174603	0.017799155619048	0.058300353206349	0.16345962493651	0.033750042253968
Centroid 2	0	0.255202008	1	0	0.096766077	1
Centroid 3	0.8090221705	0.30622126872222	0.0058839471666667	0.54717753938889	0.13091773705556	0.0099248621666667

Gambar 4.4 Data *Centroid* Akhir

4.5. Implementasi Klasifikasi Dataset

Dataset yang telah memiliki label dari proses K-Means selanjutnya diproses melalui algoritma Naïve Bayes untuk dilakukan klasifikasi terhadap dataset baru yang tidak termasuk ke dalam dataset pelatihan dan pengujian. Dataset ini merupakan daftar saham perusahaan yang berada pada sektor yang sama namun tidak memenuhi syarat sebagai saham *second liner* atau tidak memiliki kapitalisasi pasar minimal 500 miliar. Terdapat 10 saham yang akan diklasifikasi untuk ditentukan tingkat risikonya masing-masing dan proses ini akan dijalankan pada sistem. Salah satu saham yang akan diklasifikasi adalah saham SKBM yang memiliki kapitalisasi pasar 480,97 miliar. Pengguna akan memasukkan data-data rasio keuangan yang dimiliki SKBM dan sistem akan melakukan identifikasi tingkat risiko seperti terlampir pada Gambar 4.5 berikut ini :



Gambar 4.5 Proses Klasifikasi Dataset

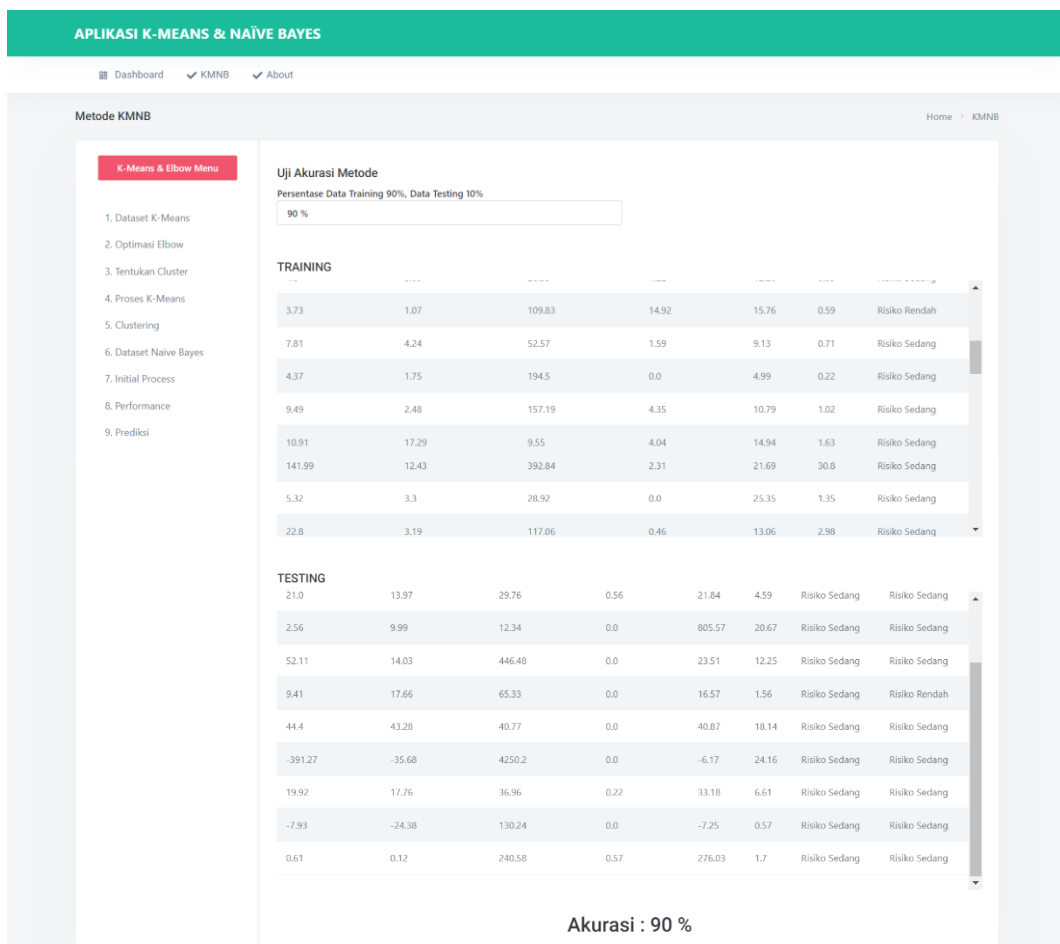
Dapat dilihat pada gambar bahwa saham SKBM berhasil diklasifikasi oleh sistem dan memiliki tingkat risiko sedang berdasarkan nilai probabilitas terbesar yang diperoleh dari nilai rasio keuangan yang telah diinput pengguna. Klasifikasi untuk 9 saham lainnya dapat dilihat pada Tabel 4.4 berikut :

Tabel 4.4 Hasil Klasifikasi

simbol	return_on_eq uity	net_profit_ma rgin	debt_equity_ ratio	dividend_y ield	per	pbv	cluster
SKBM	0.12	0.04	82.4	3.81	376.7	0.46	Risiko Sedang
MAXI	1.05	1.81	29.66	0	302.3 4	3.19	Risiko Sedang
ANDI	-13.4	-12.82	90.34	0	- 16.09	2.16	Risiko Sedang
GULA	3.15	1.27	50.12	0	100.8 2	3.17	Risiko Sedang
MAGP	-57.48	-32.77	1038.37	0	- 10.64	6.11	Risiko Sedang
CRAB	6.37	2.05	68.91	0.43	38.14	2.43	Risiko Sedang
AYAM	10.84	3.15	127.74	0.00	34.03	3.69	Risiko Sedang
KMDS	25.64	15.53	26.28	3.52	6.89	1.77	Risiko Rendah
TGUK	3.16	4.15	12.81	0	61.84	1.95	Risiko Sedang
WICO	-926.61	-9.34	2285.13	0	-2.12	19.6 3	Risiko Sedang

4.6. Pengujian Sistem

Setelah model berhasil melakukan klasifikasi pada dataset baru yang bukan *second liner*, selanjutnya akan dilakukan pengujian model pada sistem melalui 4 skenario rasio persentase data pelatihan dan pengujian yang berbeda-beda. Tujuan dari proses ini adalah untuk mengevaluasi apakah model yang telah dibangun sistem mampu menghasilkan model yang baik untuk proses klasifikasi pada data-data baru berikutnya. Proses pengujian ini dilakukan pada sistem dengan skenario pertama adalah pembagian data latih 90% dan data uji 10%. Pemilihan sampel untuk data latih ini akan dilakukan secara acak sehingga hasil evaluasi berupa nilai akurasi model dapat dilihat pada Gambar 4.6 berikut :



Gambar 4.6 Nilai Akurasi Rasio 90% Data Pelatihan

Hasil nilai akurasi dari 4 skenario rasio data pelatihan dan pengujian dapat dilihat pada Tabel 4.5 berikut ini :

Tabel 4.5 Hasil Nilai Akurasi 4 Skenario Rasio

Persentase Data Latih dan Data Uji	Akurasi
60%:40%	79.41%
70%:30%	84.61%
80%:20%	94.44%
90%:10%	90%
Rata-rata Akurasi	87.12%

Dari 4 skenario rasio persentase yang berbeda-beda untuk setiap data pelatihan dan pengujian didapatkan kesimpulan bahwa nilai akurasi terbaik berada di rasio persentase 80% untuk data pelatihan dan 20% untuk data pengujian, dengan rata-rata keseluruhan nilai akurasi yang didapat adalah sebesar 87,12%. Hasil ini menunjukkan bahwa model mampu memberikan nilai akurasi yang baik selama proses pembelajaran untuk dapat digunakan pada proses klasifikasi selanjutnya.

BAB 5

KESIMPULAN DAN SARAN

5.1. Kesimpulan

Kesimpulan yang diperoleh berdasarkan penelitian yang telah dilakukan untuk identifikasi tingkat risiko saham pada sektor barang konsumen primer adalah sebagai berikut :

1. Sistem berbasis website berhasil dibangun dengan implementasi algoritma *clustering* yaitu K-Means dan algoritma klasifikasi Naïve Bayes untuk identifikasi tingkat risiko saham sektor barang konsumen primer.
2. Pengelompokkan data berhasil dilakukan pada 82 data saham sektor barang konsumen primer dalam kategori *second liner* dengan memisahkan tiap data ke dalam 3 kelompok (*cluster*) berdasarkan jumlah *k* yang optimal pada metode Elbow dengan tingkat risiko berbeda-beda yaitu rendah, sedang dan tinggi.
3. Model yang telah dibangun mampu melakukan klasifikasi pada 10 data saham sektor barang konsumen primer yang bukan *second liner* dengan tingkat risiko yang berbeda-beda.
4. Berdasarkan hasil pengujian sistem untuk mengevaluasi model yang telah dibangun dengan 4 skenario rasio persentase diperoleh hasil bahwa nilai akurasi terbaik yaitu sebesar 94,44% berada di persentase 80% untuk data latih dan 20% untuk data pengujian, dengan rata-rata nilai akurasi yang didapat adalah sebesar 87,12%.

5.2. Saran

Beberapa saran yang dapat diberikan untuk pertimbangan dan bahan pengembangan serta perbaikan penelitian selanjutnya adalah :

1. Menerapkan integrasi algoritma K-Means dan Naïve Bayes untuk klasifikasi saham pada sektor yang berbeda.

2. Proses optimasi *centroid* dengan K-Means++ sebaiknya mampu dilakukan pada sistem sehingga pengguna tidak perlu melakukan perhitungan secara manual.
3. Melakukan proses seleksi data saham sebagai dataset penelitian dengan lebih baik lagi seperti tidak memasukkan data saham yang nilai pasarnya berada di angka Rp50 (batas bawah nilai saham di bursa efek Indonesia) dan tidak melibatkan data saham yang berada di papan akselerasi dan pantauan khusus yang nilai batas bawah sahamnya dapat mencapai Rp1.
4. Mengimplementasikan metode evaluasi lain selain metode Elbow seperti *Silhouette Coefficient Index* dan Davies Bouldin dalam penentuan jumlah k yang optimal.
5. Mengintegrasikan algoritma *clustering* selain K-Means dan algoritma klasifikasi selain Naïve Bayes untuk mengetahui kinerja model yang lebih baik dalam proses *clustering* dan klasifikasi pada data yang tidak berlabel / tidak memiliki kelas

DAFTAR PUSTAKA

- Asroni, A., Fitri, H., & Prasetyo, E. (2018). Penerapan Metode Clustering dengan Algoritma K-Means pada Pengelompokan Data Calon Mahasiswa Baru di Universitas Muhammadiyah Yogyakarta (Studi Kasus: Fakultas Kedokteran dan Ilmu Kesehatan, dan Fakultas Ilmu Sosial dan Ilmu Politik). *Semesta Teknika*, 21(1), 60-64. doi:<https://doi.org/10.18196/st.211211>
- Brigham, E. F., & Houston, J. F. (2019). *Dasar-Dasar Manajemen Keuangan* (14th ed.). Jakarta: Salemba Empat.
- Bursa Efek Indonesia. (2021). Indeks. Retrieved from <https://www.idx.co.id/media/9594/idx-industrial-classification.zip>
- Bursa Efek Indonesia. (2022). Laporan Tahunan. 2. Retrieved from <https://www.idx.co.id/Media/ccrhy5gi/laporan-tahunan-pt-bursa-efek-indonesia-tahun-2022-230627.pdf>
- Bursa Efek Indonesia. (2023). Melalui Berbagai Pencapaian Tahun 2023, Pasar Modal Indonesia Tunjukkan Optimisme Hadapi Tahun 2024. Retrieved from <https://www.idx.co.id/en/news/press-release/2080>
- Daniswara, H. P., & Daryanto, W. M. (2020). Earning Per Share (EPS), Price Book Value (PBV), Return on Asset (ROA), Return on Equity (ROE), and Indeks Harga Saham Gabungan (IHSG) Effect on Stock Return. *South East Asia Journal of Contemporary Business, Economics and Law*, 20(1), 11-27.
- Darmadji, T., & Fakhrudin, H. M. (2012). *Pasar Modal Di Indonesia* (3rd ed.). Jakarta: Salemba Empat.
- Ferdiansyah, M., & Chotijah, U. (2024). Implementasi Algoritme K-Means++ Untuk Clustering Penjualan Bahan Bangunan. *Jurnal Ilmiah Teknik Informatika dan Komunikasi*, 181-193. doi:<https://doi.org/10.55606/juitik.v4i1.767>

- Gitman, L. J. (2000). *Principles Of Managerial Finance* (9th ed.). America: Addison-Wesley.
- Kamaludin, & Indriani, R. (2012). *Manajemen Keuangan: Konsep Dasar dan Penerapannya*. Bandung: CV Mandar Maju.
- Kesara, R., Zahro, N. I., & Susanti, D. A. (2023). Profitabilitas Perusahaan Sektor Consumer Non-cyclical: Pengaruh Leverage, Ukuran Perusahaan, dan Likuiditas. *Jurnal Akuntansi Syariah*, 6(1), 56-69. doi:<https://doi.org/10.21043/aktsar.v6i1.20548>
- Lantz, B. (2013). *Machine Learning with R*. Birmingham: Packt Publishing.
- Lattanzi, S., & Sohler, C. (2019). A Better k-means++ Algorithm via Local Search. *International Conference on Machine Learning*. 97, pp. 6521-6530. California: PMLR.
- Lestari, P. I. (2018). *Implementasi Algoritme K-Means Clustering Dan Naive Bayes Classifier Untuk Klasifikasi Diagnosa Penyakit Pada Kucing*. Universitas Brawijaya.
- Ramageri, B. M. (2010). Data Mining Techniques and Applications. *Indian Journal of Computer Science and Engineering*, 1(4), 301-305.
- Ristyawan, M. R. (2019). Pengaruh Return On Equity (ROE), Debt To Assets Ratio (DAR), Price To Book Value (PBV) Dan Net Profit Margin (NPM) Terhadap Return Saham Perusahaan Sektor Perkebunan Yang Terdaftar Di Bursa Efek Indonesia Periode 2011-2017. *Jurnal Ekonomi Bisnis dan Kewirausahaan*, 1-17. doi:<https://doi.org/10.26418/jebik.v8i1.26966>
- RTI-Business. (2024). PT RTI Infokom. Retrieved from <https://rti.co.id>
- Stockbit. (2024). PT Stockbit Sekuritas Digital. Retrieved from <https://stockbit.com>
- Sugiarti, S., & Andreas. (2022). Pengaruh Price Book Value (PBV) dan Net Profit Margin (NPM) terhadap Return Saham pada P.T. Elnusa, Tbk. *Formosa*

Journal of Multidisciplinary Research, 1(3), 395-420.
doi:<https://doi.org/10.55927/fjmr.v1i3.594>

Sukanto, A. S., Setiawan, W., & Pratama, E. E. (2023). Data Mining untuk Pengelompokan Saham pada Sektor Energi dengan Metode K-Means. *Jurnal Edukasi dan Penelitian Informatika*, 8(4), 76-81.
doi:<https://doi.org/10.26418/jp.v9i1.62509>

Tohendry, D., & Jollyta, D. (2023). Penerapan Algoritma K-Means Clustering Untuk Pengelompokkan Saham Berdasarkan Price Earning Ratio dan Price to Book Value. *Jurnal Mahasiswa Aplikasi Teknologi Komputer dan Informasi*, 1-7.

Vulandari. (2017). *Data Mining*. Jakarta: Gava Media.

Wahyuditomo, K. W., Cholissodin, I., & Sutrisno, S. (2021). Implementasi Integrasi K-Means dan Naïve Bayes dalam Identifikasi Tingkat Risiko Reksa Dana. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, 5(7), 3094-3102.

LAMPIRAN

Lampiran : Listing Program

1. Algoritma *clustering* : K-Means

```
File Edit Selection View Go ... c3kmeans
KmeansFlow_Model.php
104 public function execute()
105 {
106     for($i=0;$i<$n;$i++){
107         $this->cluster[$i] = array();
108         if($i==0){
109             $this->data_cluster[$i] = $this->centroid;
110             $this->centroid_allor[$i] = $this->data_cluster[$i];
111         }else{
112             $this->data_cluster[$i] = $this->data_cluster[$i-1];
113             $this->centroid_allor[$i] = $this->data_cluster[$i];
114         }
115     }
116     //Carilah Jarak
117     foreach ($this->data as $d) {
118         $distance_data = array();
119         foreach ($this->data_cluster[$i] as $d0) {
120             $cluster_data = array();
121             for ($k=0;$k<$this->dimensi;$k++) {
122                 $cluster_data[$k] = pow(abs($d[$k] - $d0[$k]),2);
123             }
124             $distance_data[] = sqrt(array_sum($cluster_data));
125         }
126         $this->distance[$i][] = $distance_data;
127     }
128     //cluster
129     $s=0;
130     foreach ($this->distance[$i] as $key) {
131         $min=$key;
132         $c = array_search($min,$key);
133         $this->cluster[$i][$s] = $c;
134         $s++;
135     }
136     //repoint
137     $this->data_cluster[$i] = array();
138     foreach ($this->cluster[$i] as $key => $value) {
139         $this->data_cluster[$i][$value][] = $this->data[$key];
140     }
141     $this->data_cluster_temp = $this->data_cluster[$i];
142     $this->data_cluster[$i] = array();
143     //
144     foreach ($this->data_cluster_temp as $key => $value) {
145         $temp=array();
146         foreach ($value as $keys => $values) {
147             foreach ($values as $keys => $values) {
148                 $temp[$keys][] = $this->data_cluster_temp[$key][$keys][$values];
149             }
150         }
151         for($s1=0;$s1<sizeof($temp);$s1++){
152             $this->data_cluster[$i][$key][$s1] = array_sum($temp[$s1])/count($temp[$s1]);
153         }
154     }
155     $temp2[$i] = array();
156     $max_temp2 = array_keys($this->data_cluster[$i]);
157     $max_temp2 = max($max_temp2);
158     for($s=0;$s<=$max_temp2;$s++){
159         if(isset($this->data_cluster[$i][$s])){
160             $temp2[$i][$s] = $this->data_cluster[$i][$s];
161         }
162     }
163     $this->data_cluster[$i] = $temp2[$i];
164     if($i>0){
165         if($this->stoploop[$i]){break;}
166     }
167 }
```

```
File Edit Selection View Go ... c3kmeans
KmeansFlow_Model.php
104 $this->distance[$i][] = $distance_data;
105 }
106 //cluster
107 $s=0;
108 foreach ($this->distance[$i] as $key) {
109     $min=$key;
110     $c = array_search($min,$key);
111     $this->cluster[$i][$s] = $c;
112     $s++;
113 }
114 //repoint
115 $this->data_cluster[$i] = array();
116 foreach ($this->cluster[$i] as $key => $value) {
117     $this->data_cluster[$i][$value][] = $this->data[$key];
118 }
119 $this->data_cluster_temp = $this->data_cluster[$i];
120 $this->data_cluster[$i] = array();
121 //
122 foreach ($this->data_cluster_temp as $key => $value) {
123     $temp=array();
124     foreach ($value as $keys => $values) {
125         foreach ($values as $keys => $values) {
126             $temp[$keys][] = $this->data_cluster_temp[$key][$keys][$values];
127         }
128     }
129     for($s1=0;$s1<sizeof($temp);$s1++){
130         $this->data_cluster[$i][$key][$s1] = array_sum($temp[$s1])/count($temp[$s1]);
131     }
132 }
133 $temp2[$i] = array();
134 $max_temp2 = array_keys($this->data_cluster[$i]);
135 $max_temp2 = max($max_temp2);
136 for($s=0;$s<=$max_temp2;$s++){
137     if(isset($this->data_cluster[$i][$s])){
138         $temp2[$i][$s] = $this->data_cluster[$i][$s];
139     }
140 }
141 $this->data_cluster[$i] = $temp2[$i];
142 if($i>0){
143     if($this->stoploop[$i]){break;}
144 }
145 }
146 }
147 }
```

2. Algoritma klasifikasi : Naïve Bayes

```
nb > naiveBayes.gaussian > gaussianclass.php
45 function meanAttribute(){
46     $sum = array();
47     foreach($this->labelInq as $key){
48         foreach($this->dataset as $index->$keys){
49             if($key===$keys[$this->class]){
50                 //Berarti perhitungan
51                 foreach($this->attribute as $s){
52                     if(isset($sum[$s])){
53                         $sum[$s] += $this->dataset[$index][$s];
54                     }
55                     $sum[$s] += $this->dataset[$index][$s];
56                 }
57             }
58         }
59     }
60     $result = array();
61     foreach($sum as $key->$val1){
62         foreach($val1 as $keys->$val2){
63             //Rumus mean = data-rata
64             $result[$key][$keys] = $val2/$this->labelsum[$key];
65         }
66     }
67     return $result;
68 }
69
70 function varianceAttribute(){
71     $variance_collect = array();
72     foreach($this->dataset as $index->$key) {
73         foreach($this->labelInq as $keys) {
74             if($key==$keys[$this->class]){
75                 foreach($this->attribute as $attr){
76                     $variance_collect[$keys][$attr] += $this->dataset[$index][$attr];
77                 }
78             }
79         }
80     }
81     //Berarti nilai variance
82     $variance = array();
83     foreach($variance_collect as $index->$value1) {
84         foreach($value1 as $index2->$value2) {
85             $sum = $this->meanAttribute($index1[$index2]);
86             $temp_sum = 0;
87             foreach($value2 as $value3) {
88                 $temp_sum += pow($value3 - $sum,2);
89             }
90             $temp_sum /= (sizeof($value2)-1);
91             if(sizeof($value2)>1){
92                 $variance[$index1][$index2] = $temp_sum/(sizeof($value2)-1);
93             }
94             else{
95                 $variance[$index1][$index2] = 0;
96             }
97         }
98     }
99     return $variance;
100 }
101
102 function fx(){
103     $fx = array();
104     foreach($this->labelInq as $value1) {
105         foreach($this->attribute as $value2) {
106             if($this->varianceAttribute[$value1][$value2]==0){
107                 $fx[$value1][$value2] = 0;
108             }
109             else{
110                 $fx[$value1][$value2] = 1/sqrt(2*3.142*$this->varianceAttribute[$value1][$value2])*exp(-1*(pow($this->predict[$value2]-$this->meanAttribute[$value1][$value2],2)/(2*$this->varianceAttribute[$value1][$value2])));
111             }
112         }
113     }
114     return $fx;
115 }
116
117 function posterior(){
118     $posterior = array();
119     foreach($this->labelInq as $value1) {
120         foreach($this->attribute[$value1] as $index2->$value2) {
121             if(isset($posterior[$value1])){
122                 $posterior[$value1] += $value2;
123             }
124             else{
125                 $posterior[$value1] = $value2;
126             }
127         }
128         $posterior[$value1] *= $this->piLabel[$value1];
129     }
130     return $posterior;
131 }
132
133 function generateResult(){
134     $result = array();
135     foreach($this->dataset as $index->$keys){
136         $posterior = $this->posterior();
137         $result[$index] = $keys;
138     }
139     return $result;
140 }
```

```
nb > naiveBayes.gaussian > gaussianclass.php
107     $temp_sum = 0;
108     foreach($value2 as $value3) {
109         $temp_sum += pow($value3 - $sum,2);
110     }
111     $temp_sum /= (sizeof($value2)-1);
112     if(sizeof($value2)>1){
113         $variance[$index1][$index2] = $temp_sum/(sizeof($value2)-1);
114     }
115     else{
116         $variance[$index1][$index2] = 0;
117     }
118 }
119
120 function fx(){
121     $fx = array();
122     foreach($this->labelInq as $value1) {
123         foreach($this->attribute as $value2) {
124             if($this->varianceAttribute[$value1][$value2]==0){
125                 $fx[$value1][$value2] = 0;
126             }
127             else{
128                 $fx[$value1][$value2] = 1/sqrt(2*3.142*$this->varianceAttribute[$value1][$value2])*exp(-1*(pow($this->predict[$value2]-$this->meanAttribute[$value1][$value2],2)/(2*$this->varianceAttribute[$value1][$value2])));
129             }
130         }
131     }
132     return $fx;
133 }
134
135 function posterior(){
136     $posterior = array();
137     foreach($this->labelInq as $value1) {
138         foreach($this->attribute[$value1] as $index2->$value2) {
139             if(isset($posterior[$value1])){
140                 $posterior[$value1] += $value2;
141             }
142             else{
143                 $posterior[$value1] = $value2;
144             }
145         }
146         $posterior[$value1] *= $this->piLabel[$value1];
147     }
148     return $posterior;
149 }
150
151 function generateResult(){
152     $result = array();
153     foreach($this->dataset as $index->$keys){
154         $posterior = $this->posterior();
155         $result[$index] = $keys;
156     }
157     return $result;
158 }
```

BIODATA

Nama : Alfi Dharmawan
Tempat/Tgl Lahir : Binjai, 26 Februari 2000
Alamat : Jalan Candra Kirana No. 23 Binjai Kota

E-Mail : alfidarmawan79@gmail.com

Agama : Islam

Nama Orang Tua :

Ayah : Heri Purnomo
Ibu : Yunida

Jumlah Saudara :3
Anak Ke :2

Riwayat Pendidikan : SDN 020254 BINJAI (2005-2011)
SMP NEGERI 1 BINJAI (2011-2014)
SMA NEGERI 2 BINJAI (2014-2017)
FAKULTAS ILMU KOMPUTER DAN TEKNOLOGI
INFORMASI UNIVERSITAS SUMATERA UTARA
(2017 - 2024)

Medan, 22 Maret 2024

Alfi Dharmawan