BAB III

METODOLOGI PENELITIAN

3.1 Gambaran Umum Objek Penelitian

Penelitian dilakukan untuk mendapatkan hasil rekomendasi produk *fashion* XYZ berdasarkan kemiripan visual. *Fashion* XYZ merupakan perusahaan yang berasal dari Singapura dan menjual produk-produknya di Indonesia melalui *offline* dan *online store*. Produk *fashion* XYZ memiliki tiga lini produk, yaitu wanita, pria, dan anak-anak. Objek penelitian ini merupakan produk *fashion* XYZ dengan jenis pelanggan wanita dan pria. Produk anak-anak tidak disertakan dalam penelitian ini karena penjualan lini tersebut belum difokuskan di Indonesia. Produk wanita dan pria terbagi kembali menjadi beberapa jenis, mulai dari tas hingga aksesoris. Produk wanita terdiri atas dompet, tas, dan sepatu, sedangkan produk pria terdiri atas dompet, tas, sepatu, dan ikat pinggang. Seluruh data gambar pada masing-masing kategori digunakan dalam penelitian ini sebagai *input* model.

Jumlah gambar produk yang dimiliki sebanyak 2287, namun tidak seluruh gambar dapat digunakan. Oleh karena itu, dilakukan penyaringan gambar hingga mendapatkan hasil akhir sebanyak 1918 gambar produk. Penyaringan gambar dilakukan secara manual dengan menghilangkan dokumen *corrupt*, kategori yang terlalu sedikit, dan gambar produk yang kurang jelas. Gambar produk yang diambil merupakan produk yang bersifat unik dari sisi model dan warna. Pemilihan dan pengambilan sejumlah data tersebut juga didasari pada ketentuan perusahaan yang membatasi penggunaan data. Data gambar memiliki format .jpeg dan didapatkan dari *website e-commerce* XYZ pada Januari 2023 melalui metode *scraping*.

3.2 Metode Penelitian

Penelitian ini termasuk ke dalam jenis penelitian kuantitatif, yaitu pendekatan pengolahan data menggunakan perhitungan rumus dan model matematis [77]. Klasifikasi tersebut didasarkan pada penggunaan fitur gambar dalam bentuk *array* angka dan perhitungan rumus matematika dalam algoritma *machine learning* untuk

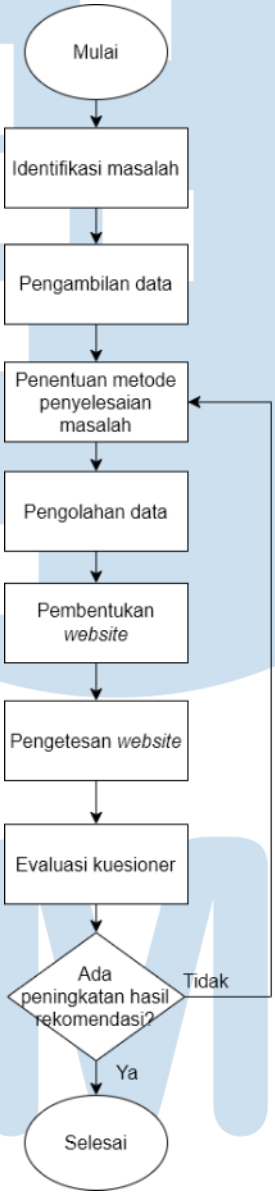
mencari kedekatan gambar. Evaluasi kemiripan gambar juga dilaksanakan berdasarkan perhitungan rumus yang termasuk ke dalam proses kuantitatif.

3.2.1 Alur Penelitian

Alur penelitian merupakan susunan rencana kerja yang akan dilaksanakan selama penelitian berlangsung. Alur ini berfungsi sebagai acuan penelitian dan menjaga agar proses penelitian dilaksanakan secara terstruktur. Berdasarkan gambar 3.1, penelitian dimulai dengan mengidentifikasi permasalahan yang ada pada sistem *e-commerce* perusahaan XYZ. Berdasarkan identifikasi tersebut, ditemukan bahwa fitur rekomendasi pada *e-commerce* belum bekerja secara optimal. Oleh karena itu, peningkatan rekomendasi *e-commerce* tersebut menjadi tujuan utama dalam penelitian ini. Setelah mengetahui permasalahan, selanjutnya dilakukan pengambilan data gambar menggunakan metode *scraping* dari *e-commerce* resmi XYZ. Selain pengambilan gambar, dilakukan juga pembuatan data *dummy* yang berisi informasi dari setiap produk sebagai data pendukung. Berdasarkan data yang ada, dilakukan penentuan metode penyelesaian masalah menggunakan *Content-Based Image Retrieval*, VGG16, dan K-Nearest Neighbor untuk membentuk model rekomendasi gambar yang baru. Penentuan tersebut didasarkan pada studi literatur dan disesuaikan dengan kebutuhan serta data yang ada. Setelah menentukan metode, proses pengolahan data dimulai dengan tahapan berdasarkan *framework* CRISP DM.

Proses penelitian tidak hanya berhenti hingga pembentukan model rekomendasi saja, tetapi juga dilanjutkan hingga pembentukan *website* untuk pengetesan model. Pembentukan model dilakukan dengan HTML, CSS, PHP, JavaScript, dan Flask. Setelah *website* berhasil dibuat, selanjutnya dilakukan *testing* untuk mengevaluasi kerja model pada *website*. Tahap akhir dalam penelitian ini adalah evaluasi kepada pengguna secara acak untuk membandingkan hasil rekomendasi sebelum dan sesudah perbaikan sesuai dengan *User Acceptance Test* (UAT). Penelitian ini selesai apabila hasil evaluasi menyatakan bahwa ada peningkatan hasil rekomendasi dari sebelumnya.

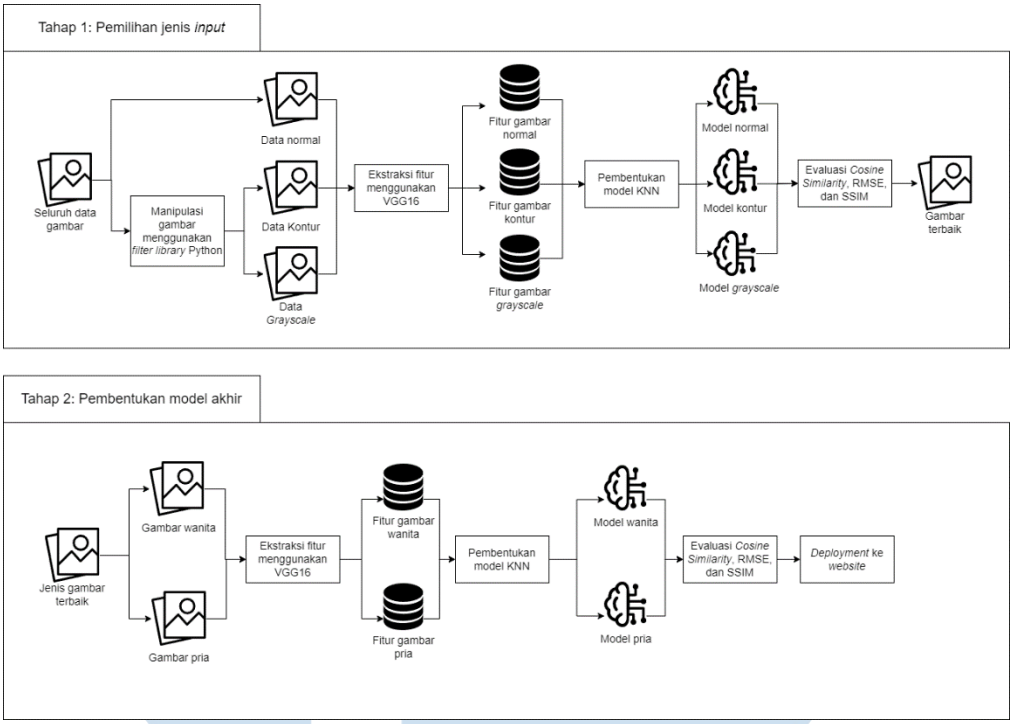
Namun, apabila hasil rekomendasi masih belum memuaskan pengguna, maka diperlukan evaluasi terhadap metode yang digunakan dan pengolahan data ulang.



Gambar 3.1 Alur Penelitian

Secara lebih jelas, proses pengolahan data dan evaluasi model pada penelitian ini digambarkan melalui gambar 3.2. Pengolahan data terbagi menjadi dua tujuan, yaitu pemilihan jenis *input* dan pembentukan model akhir. Tahap pemilihan *input* dimulai dengan pembentukan data gambar ke dalam tiga jenis,

yaitu normal, *contour*, dan *grayscale*. Ketiga jenis *input* dibentuk ke dalam model rekomendasi sehingga mendapatkan jenis yang terbaik berdasarkan hasil evaluasi. Proses pembentukan model dimulai dengan ekstraksi fitur menggunakan VGG16 sehingga tercipta 3 jenis fitur yang berbeda. Setiap fitur merupakan *input* utama dalam model K-Nearest Neighbors. Berdasarkan *input* tersebut, didapatkan 3 jenis model KNN untuk mencari gambar terdekat. Selanjutnya, setiap hasil model dievaluasi dengan pengukuran RMSE, *Cosine Similarity*, SSIM, dan perhitungan kesalahan rekomendasi. Hasil *input* terbaik akan diambil kembali untuk pembetukan model pria dan wanita (tahap 2). Pembagian model pria dan wanita didasari oleh kebutuhan perusahaan untuk merekomendasikan produk sesuai dengan lini yang sedang dilihat. Model wanita hanya dibentuk dari data produk wanita, sedangkan model pria hanya dibentuk dari data produk pria. Pemisahan tersebut menjamin bahwa tidak ada rekomendasi silang antar lini. Hal ini disebabkan model wanita hanya mencari produk dari sekumpulan data wanita, begitu juga sebaliknya. Pembentukan model dimulai dengan ekstraksi fitur menggunakan VGG16 sehingga menciptakan fitur pria dan wanita yang terpisah. Selanjutnya, model rekomendasi dibentuk kembali menggunakan algoritma KNN. Hasil model tersebut dievaluasi menggunakan metode evaluasi yang sama seperti tahap 1. Model yang telah dievaluasi kemudian akan memasuki tahap *deployment* ke *website* untuk keperluan *testing*.



Gambar 3.2 Kerangka Proses Pengolahan Data

3.2.2 Metode *Data Mining*

Penelitian ini berfokus untuk membentuk model rekomendasi produk menggunakan ilmu pengolahan data melalui *data mining*. Terdapat beberapa kerangka kerja yang dapat digunakan untuk membantu pelaksanaan proses *data mining*, seperti *Cross-Industry Standard Process for Data Mining* (CRISP-DM), *Knowledge Discovery in Database* (KDD), dan *Obtain, Scrub, Explore, Model, and Interpret* (OSEMN).

Tabel 3.1 Perbandingan Tahapan *Framework Data Mining*

Indikator	CRISP-DM	KDD	OSEMN
Tahapan	1. Pemahaman bisnis 2. Pemahaman data 3. Persiapan data 4. Pembuatan model 5. Evaluasi 6. *Deployment*	1. Pengambilan dan seleksi data 2. *Preprocessing* atau persiapan data 3. Transformasi data	1. *Obtain* atau pengumpulan data 2. *Scrub* atau pembersihan data 3. Eksplorasi data 4. Pembentukan model

Indikator	CRISP-DM	KDD	OSEMN
		4. *Data mining* 5. Evaluasi 6. Pengambilan pengetahuan	5. *Interpret* dan penggunaan hasil model
Proses	Iteratif	Iteratif	Non-iteratif
Kekurangan	• *Documentation heavy* dan dapat memperlambat kinerja tim dalam proses dokumentasi • Tidak mendukung kolaborasi tim dan *big data*	• *Outdated* dan kurang sesuai untuk diterapkan dalam proyek *data science* modern • Tidak dapat diterapkan pada arsitektur *big data* • Memerlukan penyimpanan sebelum data dapat diproses • Tidak memiliki pemahaman kebutuhan bisnis dan *deployment* untuk penyebaran informasi	• Tidak memiliki pemahaman terhadap kondisi bisnis dan kebutuhan proyek • Tidak memiliki proses *deployment* dan mengasumsikan bahwa hasil hanya digunakan satu kali • Tidak mendukung kolaborasi tim dan *big data*
Kelebihan	• Memiliki tahapan yang ringkas namun menyeluruh • Berfokus kepada tujuan bisnis sehingga hasil analisis selaras dengan kebutuhan • Pendekatan iteratif memberikan	• Bekerja dengan baik untuk kasus prediksi dan pengenalan kebutuhan pelanggan • Pendekatan iteratif memberikan kesempatan evaluasi pada setiap proses *data mining*	• Memiliki tahapan yang ringkas namun menyeluruh • Setiap tahapan mampu merepresentasikan *data science life cycle* secara logis • Memiliki susunan taksonomi untuk mendefinisikan kemajuan dari

Indikator	CRISP-DM	KDD	OSEMN
	kesempatan evaluasi proses-proses dalam *data science* • Dapat diterapkan ke dalam berbagai jenis teknologi untuk beragam kebutuhan analisis data	• Memberikan hasil akhir berupa pengetahuan dari hasil analisis data	proyek *data science*

Berdasarkan perbandingan *framework* pada tabel 3.1, setiap kerangka kerja memiliki tahapan dan tujuan implementasi yang berbeda. CRISP-DM bekerja sebagai *framework* yang menerjemahkan permasalahan atau tujuan bisnis ke dalam proyek *data science* dan diimplementasi menjadi sebuah produk melalui proses *deployment* [78]. Berbeda dengan CRISP-DM, kerangka kerja KDD dan OSEMN tidak memiliki tahapan yang berfokus dalam pemahaman bisnis. KDD merupakan *framework* yang berfokus pada pencarian *pattern* dan informasi dari data, oleh karena itu tidak ada tahapan untuk *deployment* model.

Pada penelitian ini, *framework* yang dipilih adalah CRISP-DM. Pemilihan ini berdasarkan tujuan penelitian, yaitu pembentukan model rekomendasi *similar product* untuk kepentingan bisnis dan direalisasikan dalam bentuk *website*. Melalui tujuan tersebut, diperlukan *framework* yang mencangkup pemahaman akan bisnis, *deployment*, dan bersifat iteratif karena koleksi produk selalu berubah. *Framework* terbaik untuk penelitian ini adalah CRISP-DM karena mencangkup kebutuhan penelitian. Tahapan dalam proses CRISP-DM dalam penelitian ini secara lebih jelas dijabarkan pada poin-poin berikut:

3.2.2.1 *Business Understanding*

Tahap pertama dalam kerangka kerja CRISP-DM adalah pemahaman bisnis. Pada penelitian ini, kasus yang diangkat adalah perusahaan *fashion*

XYZ yang membutuhkan model rekomendasi untuk fitur *similar product* pada *e-commerce* khusus milik XYZ. Hasil rekomendasi tersebut digunakan untuk menawarkan produk yang sesuai dengan minat pelanggan, sehingga meningkatkan pengetahuan pelanggan terhadap produk yang tersedia. Perusahaan sudah memiliki fitur rekomendasi, namun fitur tersebut belum mampu bekerja secara optimal untuk menampilkan produk serupa. Produk-produk yang direkomendasikan umumnya datang dari departemen produk yang beragam dan tidak sesuai dengan produk yang sedang dilihat oleh pelanggan. Oleh karena itu, perbaikan dan peningkatan kualitas rekomendasi perlu dilakukan. Rekomendasi ini juga dapat digunakan sebagai salah satu strategi pemasaran produk melalui *e-commerce* dengan menampilkan berbagai desain dari produk yang diminati pelanggan.

3.2.2.2 *Data Understanding*

Tahap kedua pada CRISP-DM merupakan pemahaman terhadap data, sehingga pemrosesan yang dilakukan sesuai dengan karakteristik data dan kebutuhan pembentukan model. Data yang digunakan dalam penelitian ini terbagi menjadi dua jenis, yaitu berupa gambar dan data tabel. Data gambar merupakan foto produk yang dijual oleh *fashion* XYZ, sedangkan data tabel berisi deskripsi gambar seperti kode produk, nama produk, kategori, dan harga. Data gambar digunakan sebagai data utama untuk pembentukan model rekomendasi, sedangkan data tabel merupakan data tambahan untuk deskripsi produk pada *website*.

3.2.2.3 *Data Preparation*

Tahap ketiga dalam CRISP-DM adalah persiapan data. Data yang diperoleh harus melewati tahap pembersihan, transformasi, dan serangkaian proses persiapan lainnya sebelum digunakan dalam pembentukan model. Hal ini dilakukan untuk memperoleh model dan hasil rekomendasi yang lebih baik. Data yang digunakan dalam penelitian ini terbagi menjadi dua

jenis, yaitu data gambar dan data tabel. Berikut merupakan tahap persiapan data yang dilewati:

1) Penanganan nilai yang hilang (*missing value*)

   Tahap pertama dalam data *preparation* adalah pengecekan *missing value* untuk data tabel. Analisis *missing value* dilakukan untuk mengetahui proporsi *missing value* dan jenis data yang hilang, kemudian mengisi kembali baris tersebut dengan data baru yang sesuai.

2) Pemotongan gambar

   VGG16 menerima *input* berupa gambar dalam rasio 1:1 (kotak), sedangkan gambar yang didapatkan memanjang secara vertikal. Oleh karena itu, gambar perlu dipotong menjadi ukuran kotak secara seragam. Proses pemotongan gambar dilakukan secara otomatis menggunakan Python, namun pengecekan ulang secara manual diperlukan untuk memastikan tidak ada objek yang terpotong.

3) Manipulasi gambar

   Manipulasi gambar merupakan tahapan untuk mengubah gambar menjadi dua tipe, yaitu *grayscale* dan *edge contour*. Selain itu, gambar juga dibuat menjadi lebih tajam dengan fungsi *sharpen*. Pada data normal, data gambar tidak dimanipulasi dengan pemberian *filter*. Data *grayscale* merupakan data gambar yang diubah menjadi warna hitam-putih, sedangkan data *edge contour* merupakan gambar yang diubah sehingga menyisakan garis pinggiran dan garis lekukan objek saja. Kedua metode manipulasi data tersebut umum digunakan untuk menekankan bentuk dan tekstur dari suatu objek. Setiap data tersebut dibentuk ke dalam model KNN yang berbeda-beda untuk dilakukan komparasi sehingga mendapatkan jenis *input* terbaik.

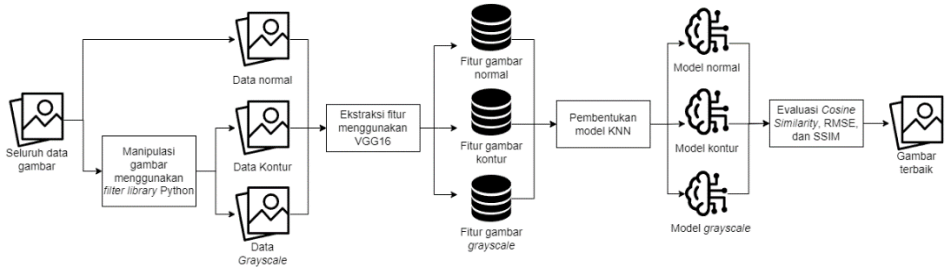
4) *Feature Extraction*

   Feature extraction merupakan proses pengubahan gambar menjadi angka dalam bentuk *array*. Proses pemisahan fitur ini dilakukan dengan bantuan arsitektur VGG16. VGG16 merupakan *pre-trained network*

yang terdiri atas 16 lapisan *layer* dan telah dilatih menggunakan miliaran data dalam *database* ImageNet. Arsitektur tersebut mampu memahami data gambar dengan baik, sehingga banyak digunakan untuk proses klasifikasi gambar, *feature extraction*, dan *transfer learning*. Fitur yang didapatkan dari tahap ini merupakan komponen yang digunakan untuk pembentukan model.

3.2.2.4 *Modeling*

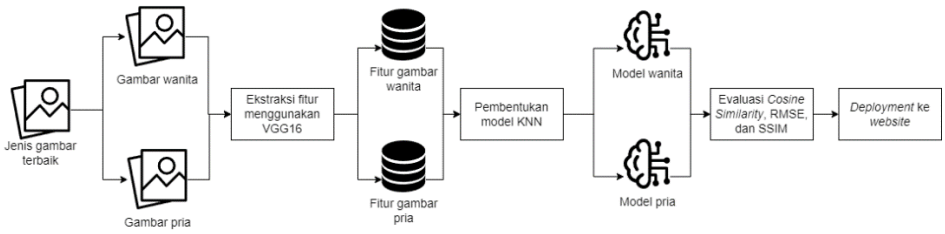
Data yang telah dibersihkan dan dipersiapkan dengan baik dapat melanjutkan ke tahap *modeling*. KNN digunakan untuk mendeteksi gambar yang paling dekat berdasarkan perhitungan *cosine distance* atas fitur-fitur gambar yang telah diekstrak. Keseluruhan proses dalam tahap ini dilakukan pada *environment* Jupyter menggunakan bahasa pemrograman Python. Secara umum, pembentukan model KNN tidak hanya dilakukan satu kali, tetapi dilakukan ke dalam dua tujuan besar, yaitu pemilihan *input* dan pemisahan lini.



Gambar 3.3 Proses Pemilihan Jenis *Input*

Proses pemilihan *input* membandingkan hasil rekomendasi 3 model dengan jenis data gambar yang berbeda, yaitu data normal, kontur, dan *grayscale*. Pembentukan model dilakukan untuk mendapatkan jenis *input* terbaik untuk model selanjutnya. Tahapan pada proses ini terdapat pada gambar 3.3. Proses pembentukan model didahului oleh proses ekstraksi fitur menggunakan VGG16. Setiap jenis *input* dibentuk ke dalam *array* yang

berbeda-beda dan disimpan dalam dokumen dengan nama yang berbeda. Setiap fitur tersebut dipanggil kembali sebagai *input* dalam model KNN. Berdasarkan proses tersebut, terdapat 3 jenis model KNN yang dievaluasi. Proses evaluasi dilakukan untuk menentukan jenis *input* terbaik.



Gambar 3.4 Proses Pembentukan Model Wanita dan Pria

Setelah penentuan jenis *input* berhasil dilakukan, selanjutnya model dibentuk kembali untuk memisahkan produk berdasarkan lini seperti yang tercantum dalam gambar 3.4. Lini merupakan kategori produk yang meliputi jenis pria dan wanita. Pemberian rekomendasi produk harus sesuai dengan lini produk yang sedang dilihat, dengan demikian dilakukan pemisahan model untuk memastikan bahwa hasil rekomendasi sesuai dengan lini yang diharapkan. Pembentukan model pria dan wanita diawali dengan pemisahan data dan ekstraksi fitur secara terpisah untuk dua data tersebut. Pembentukan model pada penelitian ini tidak dipisahkan menjadi data *training* dan *testing*. Seluruh data dalam *database* dimuat untuk *feature extraction* sebagai sumber untuk pencarian data gambar. *Split training* dan *testing* akan memperkecil ketepatan hasil rekomendasi karena konten atau objek produk yang dibandingkan berkurang jumlahnya. Setiap fitur yang dihasilkan kemudian digunakan kembali sebagai *input* model KNN. Selanjutnya, hasil model pria dan wanita dievaluasi untuk menentukan apakah model tersebut layak untuk disebarkan ke dalam *e-commerce* atau tidak.

Proses pembentukan model wanita dan pria dapat dilakukan kembali apabila terdapat penambahan atau perubahan data produk XYZ. Apabila produk *fashion* mengalami pergantian musim, tren, atau stok produk, sistem

rekomendasi tetap dapat digunakan, namun diperlukan ekstraksi ulang fitur dan pembentukan ulang model berdasarkan tahapan serta metode yang sama. Ekstraksi fitur dan pembentukan model hanya perlu dijalankan satu kali setelah perubahan atau penamabahan data.

3.2.2.5 *Evaluation*

Evaluasi merupakan tahap penting yang berfungsi untuk mengetahui seberapa baik performa model dalam memberikan rekomendasi *similar product*. Evaluasi pada penelitian ini dilakukan menggunakan nilai *Cosine Similarity*, *Root Mean Square Error* (RMSE), dan *Structure Similarity Index Method* (SSIM). Evaluasi ini bertujuan untuk mengetahui kualitas model serta mengetahui kemiripan *input* dengan hasil rekomendasi. Semakin kecil nilai *error*, maka semakin baik rekomendasi yang diberikan. Sebaliknya, semakin besar nilai SSIM, maka hasil rekomendasi yang diperoleh semakin baik. Evaluasi dilakukan menggunakan data yang sama dengan data pembentuk model. Jumlah yang digunakan yaitu satu sampel dari setiap kategori produk.

Setelah sistem rekomendasi berhasil dibentuk, selanjutnya dilaksanakan evaluasi kepada pengguna *e-commerce* melalui survei. Proses evaluasi tersebut dilaksanakan berdasarkan *User Acceptance Test* dengan cara membagikan kuesioner yang berisi pernyataan dan dijawab dalam skala likert (rentang 1-5). UAT dipilih karena mampu mengevaluasi sistem berdasarkan tujuan pembentukan sistem dan tujuan bisnis, hal tersebut merupakan fokus utama dalam evaluasi penelitian ini. Semakin tinggi nilai yang dipilih, maka responden semakin setuju dengan pernyataan yang diajukan. Kuesioner dibagikan untuk mengetahui pendapat responden terhadap sistem rekomendasi sebelum diperbaiki dan setelah diperbaiki. Pernyataan yang diajukan berasal dari tujuan pembentukan sistem rekomendasi. Rekomendasi diberikan untuk menampilkan produk yang mirip secara visual sehingga meningkatkan minat pengguna dalam

mengeksplorasi produk, membeli produk, membantu memperkenalkan koleksi produk, dan membantu pelanggan dalam menemukan produk yang sesuai dengan minatnya. Berdasarkan tujuan tersebut, evaluasi dilaksanakan dengan 6 pernyataan, yaitu:

- Hasil rekomendasi terbaru lebih sesuai dengan produk utama yang dilihat
- Hasil rekomendasi terbaru lebih sesuai dengan harapan saya sebagai pengunjung *e-commerce*
- Hasil rekomendasi terbaru lebih baik dalam memperkenalkan koleksi produk XYZ
- Hasil rekomendasi terbaru lebih membantu dalam menemukan produk yang diminati
- Hasil rekomendasi terbaru meningkatkan ketertarikan penguna untuk melihat dan mengeksplorasi koleksi produk yang ditawarkan
- Hasil rekomendasi terbaru meningkatkan ketertarikan untuk membeli produk yang ditawarkan

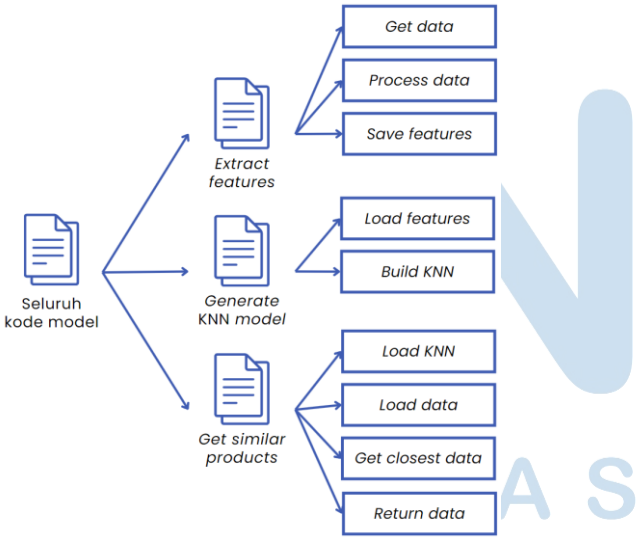
Pernyataan tersebut dapat dinilai dari rentang 1-5 berdasarkan skala likert. Nilai 1 merepresentasikan pendapat sangat tidak setuju, nilai 2 berarti tidak setuju, nilai 3 merupakan nilai netral, nilai 4 berarti setuju, dan 5 berarti sangat setuju. Hasil evaluasi ini merupakan salah satu penentu untuk menyatakan bahwa penelitian terlah berhasil dilaksanakan.

3.2.2.6 *Deployment*

Tahap terakhir pada CRISP-DM adalah penyebaran informasi kepada pengguna akhir atau disebut juga sebagai *deployment*. Tahap ini dapat dilakukan apabila evaluasi model sudah menghasilkan kesimpulan yang baik. *Deployment* dapat dilakukan melalui beberapa cara, seperti visualisasi, *dashboard*, *website*, aplikasi *mobile*. Sistem rekomendasi pada penelitian ini dibentuk berbasis *web* sebagai simulasi penerapan model ke dalam *e-commerce fashion* XYZ. Proses *deployment* terdiri atas 3 tahapan, yaitu

pembentukan *script* Python, *interface website*, dan API sebagai komunikator antara model dengan *interface.*

Pembentukan *script* Python bertujuan untuk memisahkan fungsi-fungsi besar pada tahapan pemrosesan data dan pemberian rekomendasi sehingga mempermudah *maintenance* dan mempercepat pemanggilan rekomendasi. Proses pemisahan *script* ditunjukkan pada gambar 3.5. *Script* dibagi menjadi 3 fungsi utama, yaitu *extract features*, *generate model*, dan *get similar products*. *Extract features* berfungsi untuk mengubah seluruh gambar menjadi *array* untuk diproses oleh komputer. *Generate model* digunakan untuk membentuk model KNN dari *input* fitur yang telah diekstrak. *Get similar products* digunakan untuk mengambil *query image*, mencarinya ke dalam *database*, dan mengembalikan hasil rekomendasi dalam bentuk JSON. *Script extract features* dan *generate model* hanya perlu dijalankan satu kali atau saat terdapat perubahan data dalam *database*. *Get similar products* merupakan bagian yang dipanggil setiap kali rekomendasi akan ditampilkan dalam *website*.



Gambar 3.5 Proses Pembentukan *Script*

Website yang dibuat dalam penelitian ini hanya digunakan untuk keperluan pengetesan model. Oleh karena itu, *website* tidak memiliki fungsi

yang menyeluruh. Halaman yang dibentuk terdiri atas halaman utama, katalog produk, dan detail produk. Hasil rekomendasi akan ditampilkan pada halaman detail produk. Proses *deployment* terakhir adalah pembentukan API menggunakan Flask untuk menghubungkan hasil model dengan *interface website*.

3.3 Teknik Pengumpulan Data

Penelitian ini menggunakan dua jenis data, yaitu gambar produk dan deskripsi produk. Gambar produk merupakan objek utama dalam proses analisa dan pembentukan model, sedangkan deskripsi produk merupakan dokumen dalam bentuk tabel yang berisi data tambahan dari gambar, seperti kode produk, kategori produk, nama produk, harga, dan lain-lain. Deskripsi tersebut berfungsi sebagai data tambahan untuk ditampilkan ke halaman *website*. Data gambar didapatkan melalui proses *scraping* pada *website fashion* XYZ. Data deskripsi produk merupakan data *dummy* yang dibuat untuk proses *testing* pada *website*. Penilaian hasil rekomendasi didukung dengan validasi oleh pengguna acak melalui kuesioner. Kuesioner disebarkan secara *online* menggunakan Google Formulir kepada responden berusia minimal 17 tahun di daerah Jabodetabek dan pernah mengakses *e-commerce* apapun untuk mencari barang *fashion*.

3.3.1 Populasi dan Sampel

Pengambilan data gambar pada penelitian ini menggunakan teknik *non-probability sampling*, yaitu sampel yang diambil tidak berdasarkan peluang acak. Secara spesifik, metode pengambilan sampel yang digunakan adalah *purposive sampling*. *Purposive sampling* merupakan teknik pengambilan data dari sebuah populasi berdasarkan kriteria tertentu yang ditetapkan oleh peneliti dan sampel tersebut dianggap mampu mewakili hasil dari penelitian yang diharapkan [79]. Sampel penelitian ini adalah gambar produk berdasarkan kategori yang sudah ditentukan, yaitu produk tas, sepatu, dompet, dan aksesoris untuk wanita serta pria dari merek XYZ. Sampel yang digunakan sebanyak 1918 gambar produk dengan berbagai macam kategori dan diharapkan dapat mewakili

populasi data yang mencangkup keseluruhan produk *fashion* XYZ dari berbagai lini.

Tahap validasi akhir melalui kuesioner dilakukan dengan teknik *cluster random sampling* atau pengambilan sampel secara acak berdasarkan geografis (area). Area yang dipilih pada penelitian ini adalah wilayah Jabodetabek (Jakarta, Bogor, Depok, Tangerang, dan Bekasi). Jabodetabek dipilih karena pusat penjualan produk XYZ berada pada wilayah tersebut. Setiap wilayah diwakilkan oleh 4 responden, sehingga validasi akan dilakukan oleh 20 responden sebagai sampel dari keseluruhan populasi pengguna *e-commerce*.

3.3.2 Periode Pengambilan Data

Proses pengambilan data dari *website fashion* XYZ dilakukan pada bulan Januari 2023, sedangkan data kuesioner dilakukan pada bulan Mei 2023. Data gambar yang dimiliki merupakan produk-produk yang masih dijual di Indonesia hingga bulan Januari 2023. Data berbentuk tabel yang berisi deskripsi produk tidak memiliki keterangan periode waktu dan dibentuk sebagai data *dummy*. Berdasarkan hal tersebut, seluruh data gambar yang dimiliki bersifat *up-to-date* hingga penelitian ini dimulai, yaitu pada Januari 2023. Data kuesioner mulai disebarkan setelah model selesai dibentuk, yaitu pada Mei 2023.

3.4 Implementasi Metode

Penelitian ini dilaksanakan berdasarkan *framework* CRISP-DM dengan bantuan algoritma *deep learning* dan *machine learning*. *Deep learning* yang digunakan adalah model dari CNN VGG16 untuk ekstraksi fitur gambar. Selanjutnya, model *machine learning* digunakan untuk mencari kemiripan gambar dengan bantuan algoritma K-Nearest Neighbor. Proses analisa dan pengolahan data menggunakan pemrograman Python pada *environment* Jupyter dengan tools Visual Studio Code.

Pre-trained model VGG16 digunakan karena model tersebut telah dilatih menggunakan jutaan data dari *database* ImageNet sehingga mampu mengenali

gambar dengan baik. Predikat tersebut didapatkan karena model VGG16 berhasil menjadi salah satu pemenang dalam kompetisi pengolahan gambar ILSVRC (ImageNet *Large Scale Visual Recognition Challenge*) pada tahun 2014. Model ini mampu mendapatkan akurasi 92.7% dalam memprediksi 14 juta gambar dalam *database* ImageNet. Model ini juga dipilih berdasarkan penelitian terdahulu yang membandingkan performa model VGG16, VGG19, dan ResNet 50 dalam mengenali gambar medis. Hasil penelitian tersebut menyebutkan bahwa VGG16 memiliki tingkat akurasi pengenalan gambar yang tertinggi, yaitu mencapai 88% [15]. Oleh karena itu, VGG16 digunakan dalam penelitian ini untuk mengekstrak fitur-fitur pada gambar produk XYZ.

Berdasarkan tinjauan teori, pencarian *similar image* umumnya dilakukan dengan menghitung jarak dari fitur gambar dan mengambil gambar dengan jarak terdekat sebagai hasil. Oleh karena itu, dibutuhkan algoritma yang mampu menghitung jarak dari data gambar. K-Nearest Neighbor merupakan algoritma *machine learning* yang bekerja dengan menghitung jarak antar data dan mencari data dengan jarak terdekat. Berdasarkan hal tersebut, algoritma KNN dipilih karena sesuai dengan kebutuhan pengolahan data. Algoritma ini bekerja dengan berbagai rumus perhitungan yang beragam, salah satunya adalah *cosine*. Metode perhitungan *cosine* sendiri telah dibuktikan pada penelitian [18] sebagai pengukuran terbaik untuk menghitung jarak pada data gambar.

Proses analisis data dan pembentukan model dapat dilakukan menggunakan bahasa pemrograman atau *software* khusus *data mining* yang tidak memerlukan penulisan kode. Bahasa pemrograman yang umum digunakan adalah R dan Python, sedangkan *tools data mining* meliputi RapidMiner dan SAS Visual Analytics. *Tools* tanpa kode tidak digunakan dalam penelitian ini karena kemampuan *custom* model yang terbatas. Dengan demikian, penelitian ini dilaksanakan menggunakan *tools* berupa bahasa pemrograman. Perbandingan antara Python dan R dijabarkan dalam tabel 3.2 [80][81].

Tabel 3.2 Perbandingan Python dan R

Indikator	Python	R
Definisi	*High-level programming language* dengan domain penggunaan yang luas dan terdiri atas sekumpulan *library* untuk pengolahan data, *networking*, *database*, dan lain-lain.	Bahasa pemrograman yang digunakan untuk mengolah data secara statistik dan menampilkan *output* grafis, didukung oleh *open source repository* CRAN untuk analisa data.
Lisensi	*Open source*	*Open source license, commercial license*
Kegunaan	*General-purpose programming language* yang dapat dimodifikasi dan digunakan sesuai kebutuhan, seperti pembuatan *website*, *scripting*, hingga analisa dan pengolahan data.	*Statistical purpose*, digunakan dengan tujuan spesifik analisis data.
Kekurangan	• Python tidak cocok untuk menjalankan *multithreaded parallel code*. • Terdapat dependensi dari beberapa *library* dan *package*	• R lebih sulit untuk dipahami bagi pemula atau pengguna yang tidak memiliki latar belakang statistik. • *Package* yang tersedia sangat banyak sehingga pencarian *package* akan memakan waktu yang lama. • Terdapat dependensi dari setiap *package*
Kelebihan	• Lebih mudah dipahami dan digunakan bagi pemula. • Mampu digunakan untuk berbagai macam keperluan, mulai dari pengolahan data hingga pembentukan *website*. • Pengguna tidak perlu menggunakan bahasa pemrograman lain karena kegunaan Python beragam. • Fleksibilitas bahasa pemrograman yang lebih tinggi dibandingkan R.	• R secara spesifik memberikan penyelesaian terhadap permasalahan yang berkaitan dengan statistik. • R memiliki *library* tidyverse yang membuat proses *data science* menjadi lebih sederhana dan cepat. • R memiliki sistem RMarkdown, sehingga pengguna dapat

Indikator	Python	R
	• Ukuran Python tergolong kecil dan baik digunakan sebagai solusi untuk *cost reduction*. • Python dapat digunakan untuk implementasi dan proses *deploy machine learning* dalam skala yang besar, serta lebih mudah dalam *maintenance*.	menuliskan dokumentasi dan kode sekaligus pada tempat yang sama. • Dokumentasi dapat disimpan dalam berbagai format, seperti html, pdf, atau docx.

Berdasarkan komparasi di atas, dapat disimpulkan bahwa Python memiliki kegunaan yang beragam dan lebih fleksibel, sedangkan R berfokus untuk menyelesaikan permasalahan statistik dan didukung dengan *package* yang beragam. Kedua bahasa pemrograman memiliki keunggulannya masing-masing, namun penelitian ini akan menggunakan Python sebagai *tools* utama. Pemilihan Python didasari oleh kemampuan Python yang lebih luas dibandingkan R. Selain itu, Python mendukung komponen utama yang diperlukan dalam penelitian ini, yaitu *scripting* untuk *deployment* model ke *website*.

Proses *deployment website* dibantu dengan Flask sebagai pembentuk API (*Application Programming Interface*). Flask digunakan sebagai komunikator dari *website* ke model rekomendasi untuk mengoper *input* untuk diproses dan menerima hasil kembali rekomendasi. Flask ditulis dalam bahasa Python, sehingga sama dengan bahasa yang digunakan dalam pembentukan model. *Framework* ini digunakan karena kompleksitas yang lebih rendah dibandingkan *framework* sejenisnya, seperti Django. Selain itu, *framework* ini juga digunakan dalam perusahaan XYZ. Oleh karena itu, Flask digunakan kembali dalam penelitian ini sebagai pembentuk API.