BAB III

METODELOGI PENELITIAN

3.1 Gambaran Umum Objek Penelitian

Objek penelitian kali ini adalah data ulasan atau *review* dari *Google Play Store* terhadap aplikasi Pencari Kerja untuk menentukan bagaimana perkembangan pendapat para pengguna aplikasi KitaLulus. Data diambil dari periode Januari 2020- April 2025 kemudian diproses terlebih dahulu sebelum membagi sentimen pengguna ke dalam 2 kategori yaitu positif dan negatif. Berikutnya, analisis sentimen akan dilakukan dengan menggunakan metode *Machine Learning*. Penelitian ini menggunakan 5 algoritma *supervised learning* diantaranya adalah *Naïve Bayes*, *Decision Tree*, *K-Nearest Neighbors* (KNN), *Support Vector Machine* (SVM) dan *Random Forest* dalam memproses data ulasan pengguna aplikasi KitaLulus, Jobstreet, LinkedIn. Lalu membandingkan manakah yang memiliki akurasi tertinggi. Hasil dari analisis sentimen ini nantinya dapat digunakan sebagai acuan untuk aplikasi tersebut.

3.2 Metode Penelitian

Berikut merupakan perbandingan metode yang digunakan yaitu KDD:

KDD (Knowledge Discovery in Databases) dan CRISP-DM (Cross-Industry Standard Process for Data Mining).

Tabel 3 1Perbandingan KDD dan CRISP-DM. Sumber:[40]

Aspek	**KDD (Knowledge Discovery in Databases)**	**CRISP-DM (Cross-Industry Standard Process for Data Mining)**
Fleksibilitas	Memberikan fleksibilitas dalam pendekatan analisis dan pemodelan data.	Struktur yang lebih terstandarisasi dengan panduan yang jelas.
Pendekatan	Fokus pada ekstraksi pengetahuan dari data besar dan kompleks.	Menyediakan langkah-langkah yang jelas dalam proses data mining.

Aspek	KDD (Knowledge Discovery in Databases)	CRISP-DM (Cross-Industry Standard Process for Data Mining)
Tahapan	Selection, Preprocessing, Transformation, Data Mining, Interpretation/Evaluation.	Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, Deployment.
Iterasi	Proses iteratif, data dapat diproses lebih dari satu kali.	Proses berulang dengan evaluasi model yang terus menerus.
Metodelogi	Lebih fleksibel dan tidak terikat pada metodelogi khusus.	Lebih terstruktur dan berbasis pada standar industri.
Kesesuaian untuk Industri	Digunakan terutama untuk analisis dan pengolahan data besar.	Dapat digunakan di berbagai industri dengan pendekatan yang konsisten.
Pengembangan Model	Model dapat berkembang seiring waktu berdasarkan hasil analisis baru.	Menekankan pembuatan model yang siap digunakan dalam waktu yang lebih singkat.
Pemahaman Bisnis	Kurang menekankan pemahaman bisnis, lebih pada teknis data mining.	Memiliki tahapan khusus untuk pemahaman bisnis di awal proses.
Evaluasi	Lebih fokus pada evaluasi hasil data mining secara teknis.	Evaluasi melibatkan pemahaman bisnis dan kegunaan model di dunia nyata.

Tabel ini menunjukkan perbedaan antara **KDD** yang lebih fleksibel dan berfokus pada pengolahan data besar, serta **CRISP-DM** yang lebih terstruktur dengan fokus pada langkah-langkah yang jelas dalam implementasi data mining di berbagai industri.

3.2.1 Metode KDD

Metode KDD (Knowledge Discovery in Databases) adalah proses yang digunakan untuk mengekstrak pengetahuan atau pola yang bermanfaat dari data besar yang tidak terstruktur atau terorganisir. KDD melibatkan beberapa tahapan yang berurutan dan iteratif untuk mengidentifikasi pola, hubungan, dan informasi tersembunyi dalam dataset yang besar dan kompleks. Tujuan utamanya adalah untuk menemukan pengetahuan yang dapat digunakan untuk pengambilan keputusan, analisis, dan pemodelan data.

- **Penjelasan Tahapan KDD:**

1. **Selection (Pemilihan Data)**

   Tahap pertama dalam KDD adalah pemilihan data yang relevan untuk analisis. Ini mencakup proses pengumpulan data dari berbagai sumber, seperti database, file, atau sumber eksternal. Data yang dipilih harus memiliki kualitas dan keterkaitan dengan tujuan analisis yang diinginkan. Pemilihan data juga dapat melibatkan pengambilan subset data untuk fokus pada area yang lebih spesifik.

2. **Preprocessing (Prapemprosesan Data)**

   Setelah data dipilih, tahap berikutnya adalah prapemrosesan data, yang bertujuan untuk membersihkan dan mempersiapkan data untuk analisis lebih lanjut. Proses ini meliputi:

   - **Pembersihan data**: Menghapus data yang hilang, duplikasi, atau data yang tidak relevan.
   - **Transformasi data**: Mengonversi data ke dalam format yang lebih sesuai atau lebih mudah untuk diproses. Ini termasuk normalisasi, standar, atau encoding data.

- **Penyusunan ulang data**: Mengatur data dalam bentuk yang lebih mudah dianalisis.

3. **Transformation (Transformasi Data)**

   Data yang telah dipersiapkan kemudian ditransformasikan menjadi format yang lebih sesuai untuk proses analisis. Tahap ini melibatkan teknik seperti pengurangan dimensi, ekstraksi fitur, atau penggunaan metode statistik untuk memperkaya data. Transformasi bertujuan untuk menyederhanakan data dan memperjelas pola atau hubungan yang ada.

4. **Data Mining (Penambangan Data)**

   Data mining adalah inti dari proses KDD, di mana teknik-teknik statistik, pembelajaran mesin, atau algoritma lainnya digunakan untuk mengekstrak pola atau pengetahuan dari data yang telah diproses. Di sini, berbagai algoritma seperti klasifikasi, clustering, regresi, atau asosiasi diterapkan untuk menemukan pola atau hubungan yang tersembunyi dalam data.

5. **Interpretation/Evaluation (Interpretasi/Evaluasi)**

   Setelah pola ditemukan, hasil tersebut perlu dievaluasi untuk menentukan seberapa bermakna dan bermanfaatnya pengetahuan yang diperoleh. Evaluasi dilakukan dengan membandingkan hasil model dengan tujuan atau kriteria yang telah ditetapkan. Di tahap ini, hasil-hasil yang relevan dieksplorasi, ditafsirkan, dan diubah menjadi informasi yang berguna untuk pengambilan keputusan.

6. **Knowledge Presentation (Penyajian Pengetahuan)**

   Tahap terakhir adalah penyajian pengetahuan yang ditemukan. Hasil dari KDD perlu disajikan dalam bentuk yang mudah dipahami oleh pengguna akhir, seperti dalam bentuk laporan, visualisasi data, atau dashboard interaktif. Pengetahuan yang diperoleh dari proses ini dapat digunakan untuk perencanaan strategi, pengambilan keputusan, atau pengembangan model lebih lanjut.

3.3 Variabel Penelitian

Terdapat 2 variabel penelitian dalam penelitian ini, yaitu independen dan dependen. Diantaranya adalah:

3.3.1 Variabel Independen

Variabel independent terdiri dari berbagai fitur yang diekstrak dari teks ulasan pengguna. Beberapa fitur yang digunakan antara lain adalah representasi teks dalam bentuk TF-IDF yang mengubah teks ulasan menjadi format numerik agar bisa diproses oleh model machine learning.

3.3.2 Variabel Dependen

Dalam penelitian ini, variabel dependen yang digunakan adalah sentiment ulasan pengguna, yang dikategorikan menjadi positif dan negatif Berdasarkan analisis teks ulasan yang diperoleh dari aplikasi KitaLulus, Jobstreet dan Linkedin. Sentimen ini menjadi target utama yang ingin diprediksi dalam penelitian

3.4 Teknik Pengumpulan Data

Teknik pengumpulan data ulasan pengguna aplikasi berbasis pada Google Play Store dilakukan menggunakan metode web scraping. Pengumpulan dilakukan dengan bantuan platform Google Colaboratory dan menggunakan library Google Play Scraper untuk mengakses dan mengambil data ulasan dari aplikasi KitaLulus, Jobstreet, dan LinkedIn. Data yang dikumpulkan

berjumlah 1000 ulasan untuk masing-masing aplikasi, dengan rentang waktu Januari 2020 hingga April 2025. Metode scraping ini dipilih karena mampu mengambil data secara cepat dan efisien tanpa perlu intervensi manual. Proses dilakukan dengan menargetkan parameter-parameter seperti nama pengguna, waktu review, rating bintang, dan isi ulasan (teks).

Teknik pengumpulan data ulasan pengguna aplikasi *based on* Google Play Store bisa dilakukan dengan *Google Colab* dan *Selenium*.

Tabel 3 2 Perbandingan Google Colab dengan Selenium[27]

	Web Scraping dengan Google Colab	**Web Scraping dengan Selenium**
Kompleksitas	Lebih sederhana dan mudah dipelajari	Lebih kompleks dan membutuhkan pemahaman tentang web browser
Kecepatan	Cenderung lebih cepat karena tidak memerlukan rendering browser	Cenderung lebih lambat karena melibatkan browser dan rendering halaman web
Kemudahan Penggunaan	Lebih mudah karena tersedianya library seperti BeautifulSoup atau Request	Memerlukan pemahaman lebih lanjut dan lebih rumit karena melibatkan browser dan interaksi dinamis
Ketersediaan	Gratis dan tersedia secara online	Berbayar dan harus di install di komputer lokal
Fitur tambahan	Mendukun pemrosesan data di cloud, sehingga memiliki skalabilitas tinggi	Tidak mendukung pemrosesan data di cloud

3.5 Teknik Analisis Data

Pada tahap pengumpulan data, saya menggunakan metode kuantitatif dimana mengumpulkan data dengan cara *web scraping* untuk mendapatkan data ulasan Kita lulu,Jobstreet dan Linkedin di Playstore pada periode 2021- 2025.

Berikut merupakan tahap tahap Teknik analisis data pada penelitian kali ini:

- Tahap *Pre Processing*

  akan dilakukan pembersihan data dari duplikasi, normalisasi, stopwords removal, tokenization dan stemming.

- Tahap Labeling sentiment

  Pada tahap ini nantinya akan dilakukan secara manual dengan memberikan label sentiment pada setiap data seperti positif dan negatif.

- Tahap Pembagian data

  Data akan dibagi menjadi dua set, yaitu data *training* dan data *testing.*

- Tahap Ekstraksi fitur

  Tahap ini nantinya akan menggunakan metode TF-IDF untuk mengubah teks ulasan menjadi representasi vector numerik yang dapat digunakan oleh algoritma klasifikasi.

- Tahap Pembuatan 5 model yang berbeda.
- Evaluasi Model

  Pada tahapan ini model akan dievaluasi menggunakan metrik seperti *accuracy,precision,recall,f-I score* dan juga optimasi model.

- Membandingan Hasil dari semua model

  Tahapan ini untuk melihat manakal model algoritma dengan akurasi terbaik.

Tabel 3.5 1 Perbandingan teori model algoritma

Naïve Bayes	***K-Nearest Neighbors* (KNN)**	***SVM***	***Decision Tree***	***Random Forest***
Naive Bayes menganalisis sentimen dengan menghitung probabilitas suatu teks(misalnya positif atau negatif) berdasarkan frekuensi kata-kata yang muncul dalam data latih.	Menggunakan model KNN untuk menentukan kelas sentimen dengan membandingkan kemiripan ulasan terhadap ulasan lain di sekitarnya.	Memisahkan ulasan berdasarkan vektor di ruang multidimensi untuk menemukan hyperplane terbaik yang membagi sentimen positif dan negatif.	Membangun pohon keputusan yang membagi data berdasarkan fitur kata tertentu untuk mengklasifikasikan ulasan ke dalam kategori sentimen.	Menggunakan beberapa pohon keputusan secara bersamaan untuk meningkatkan akurasi dalam menentukan sentimen berdasarkan fitur teks yang ada.