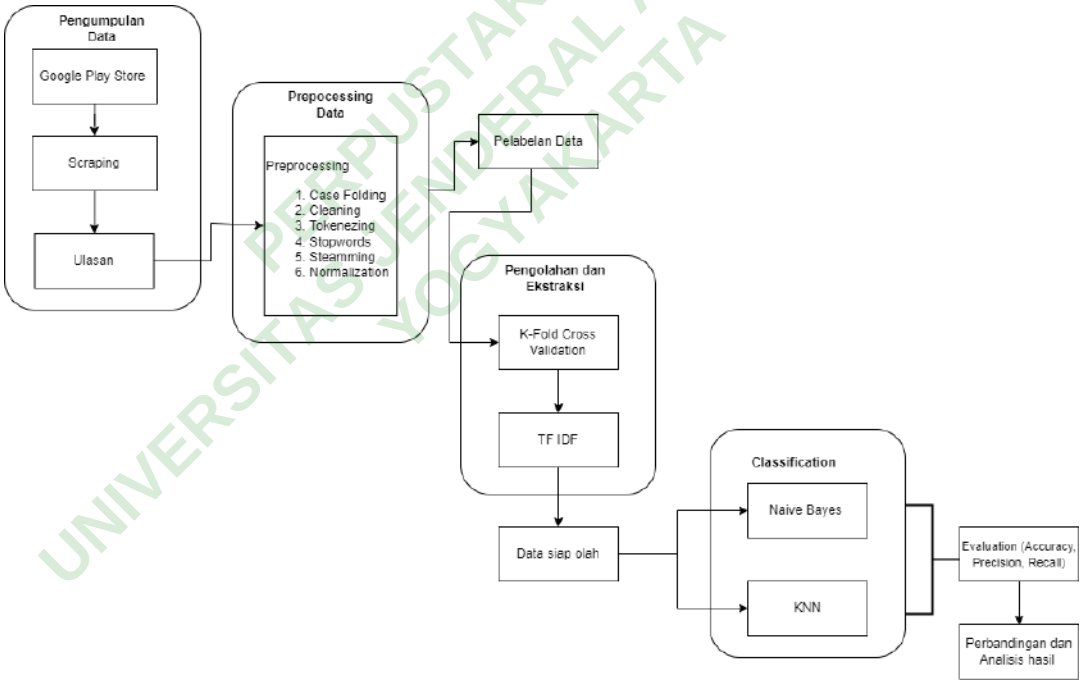
BAB 3
METODE PENELITIAN

Data untuk penelitian diperoleh dari ulasan pengguna langsung dalam aplikasi halodoc di Google *Play Store*. Kemudian data yang telah terkumpul diolah lebih lanjut melalui proses *preprocessing* , termasuk pembersihan dan penyesuaian data. Setelah itu, kinerja algoritma diuji dengan membagi data yaitu data latih dan data uji. Selanjutnya, dalam algoritma KNN, nilai k akan uji untuk menemukan nilai k yang optimal, dan kemudian membandingkan akurasinya dengan akurasi algoritma NB. Perbandingan kinerja kedua algoritma dilakukan menggunakan KKN dan NB. Pada Gambar 1.2 penjelasan tahapan metode penelitian.



Gambar 3.1 Metode Penelitian

3.1 BAHAN DAN ALAT PENELITIAN

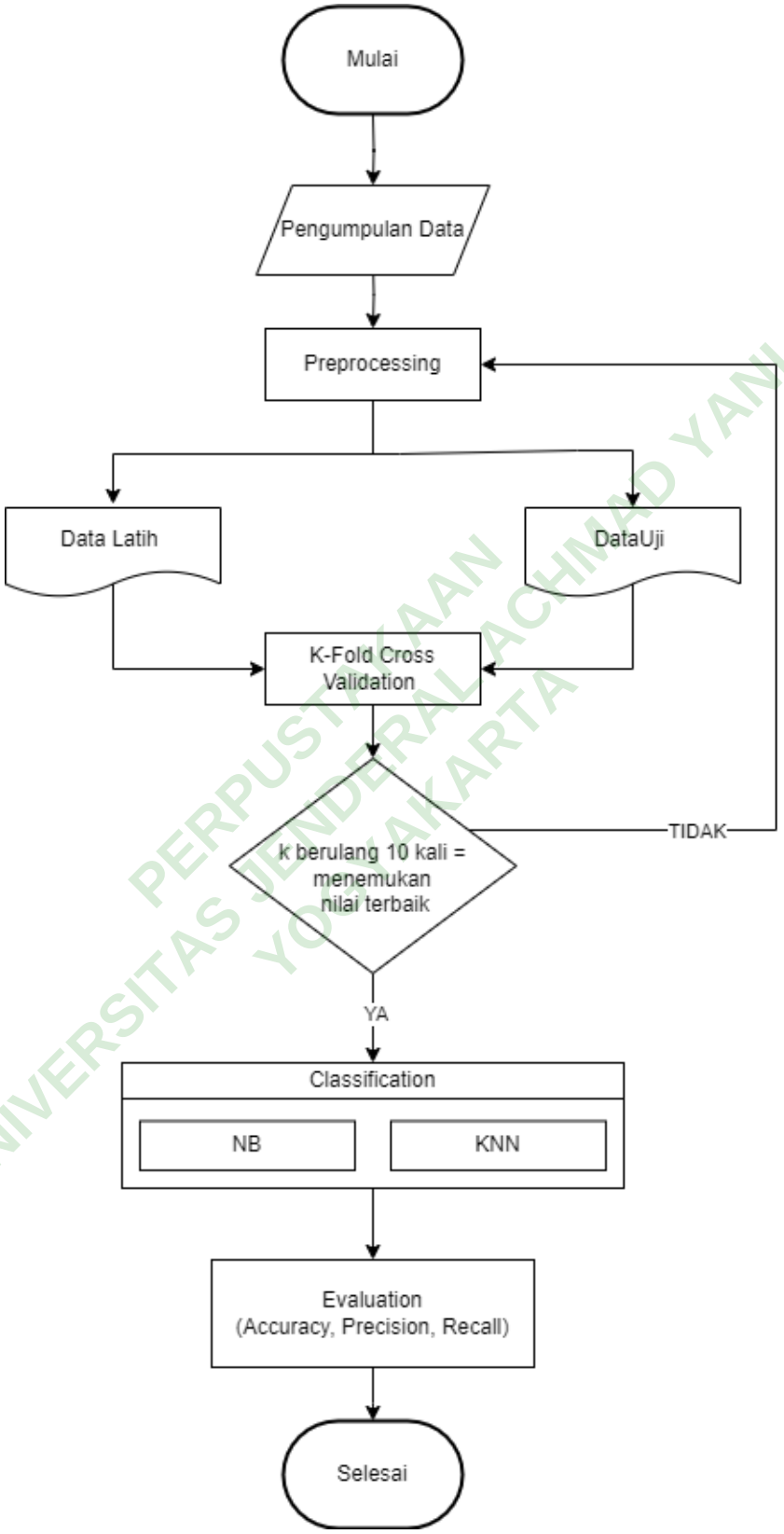
Bahan dari penelitian ini terdiri dari ulasan dan feedback pengguna yang diperoleh dari platform *Google Play Store* tentang aplikasi Halodoc. Data dikumpulkan dengan melakukan *scraping* ulasan pengguna secara langsung dari platform tersebut dari tahun 2021 hingga 2024.

Dalam penelitian ini, digunakan sistem operasi, perangkat lunak, dan akses internet untuk menjalankan aplikasi yang dikembangkan. Berikut beberapa contoh yang digunakan :

1. Sistem Operasi: Windows 11
2. Google colab
3. Visual Studio Code
4. Microsoft Excel 2021
5. *Google Play Store*

3.2 JALAN PENELITIAN

Dalam penelitian ini, Bahasa *phyton,Anaconda 3* dan *Jupyter Notebook* digunakan untuk pemrograman. Presentasi data dilakukan menggunakan *Microsoft Office Excel*, dan dimodelkan dengan dukungan berbagai *library Python*. Proses penelitian dapat dilihat pada Gambar 3.1.*Microsoft Office Excel*, dan dimodelkan dengan dukungan berbagai *library Python*. Alur penelitian dapat dilihat pada Gambar 1.3 dengan dukungan berbagai *library Python*. Proses penelitian dapat dilihat pada Gambar 1.3.



Gambar 3.2 Flowchart Jalan Penelitian

3.2.1 Pengumpulan Data

Tahap pengumpulan data adalah proses memperoleh data dari ulasan aplikasi Halodoc di *Google Play Store* dengan memanfaatkan tools *google colab* dan dieksekusi pada Jupyter Notebook yang kemudian ditampilkan di *Microsoft Excel*.

3.2.1.1 Web Scraping

Web scraping yaitu metode untuk mengambil data dari internet atau sosial media seperti *Google Play Store*, pengambilan data melalui *web scrping* menghasilkan dokumen yang bersifat semi-terstruktur. Tujuan dari web *scraping* yaitu untuk mengekstraksi informasi dari sumber online, baik secara menyeluruh maupun sebagian (Wahyudi & Kusumawardana, 2021). Setelah melakukan *web scraping* untuk mengambil dataset, data tersebut disimpan dan dikonversi ke dalam format CSV dengan jumlah ulasan aplikasi sesuai kebutuhan. Gambar 3.3 menunjukkan gambaran proses *web scraping*.

Gambar 3.3 Proses Web Scraping

(Setya Ananto & Hasan, 2023)

3.2.2 Preprocessing

Preprocessing adalah Langkah untuk membersihkan dataset yang dipilih kemudian diproses dengan mengubah beberapa data yang tidak terstruktur menjadi data yang terstruktur, Langkah ini bertujuan untuk menyiapkan dataset sebelum melakukan klasifikasi (Septianingrum et al., 2021). Dalam *preprocessing* dataset, ada beberapa langkah yang perlu dilakukan seperti *case folding*, *cleaning*,

normalization, steaming, tokenizing, stopword, dan melakukan pembobotan dengan TF IDF.

1. C*ase folding* yaitu proses merubah seluruh huruf kapital menjadi huruf kecil dalam teks.
    a. Special Removal merupakan kode atau algoritma yang digunakan untuk menghapus karakter-karakter khusus atau symbol-simbol tertentu dari teks atau data.
    b. Number Removal untuk menghapus angka dalam teks.
    c. Punctuation Removal menghapus tanda baca dari teks, kode ini membantu membersihkan teks dari karakter seperti titik, koma, tanda tanya, dan tanda seru.
    d. Whitespaces Removal menghapus spasi tambahan, tab yang ada pada awal dan akhir kalimat.
    e. Single Car Removal menghapus karakter tunggal atau karakter individu dalam teks atau data. Misalnya, “@”, “$”, dan “1”.
2. *Cleaning* digunakan untuk menghilangkan *noise* atau simbol
3. *Tokenizing* digunakan untuk membagi kalimat dalam data menjadi beberapa bagian kata yang lebih kecil.
4. *Stopword* adalah langkah dalam penggunaan daftar stopwords Kaggle untuk menghilangkan kata-kata yang dianggap tidak berguna atau penting. Namun, ada beberapa kata-kata tertentu seperti "kok", "lah", dan sejenisnya yang kembali dimasukkan ke dalam data, yang dianggap tidak relevan untuk penelitian ini.
5. *Stemming* untuk menghapus imbuhan kata sehingga diperoleh kata dasarnya.
6. *Normalization* digunakan untuk memperbaiki pengejaan kata yang salah atau singkat, contohnya mengubah kata ‘sama’ menjadi ‘sm’, atau ‘kamu’ menjadi ’km’.

3.2.3 Labelling

Data *scraping* yang telah di preprocessing akan dilakukan pelabelan. Pada tahap ini, data yang digunakan akan di labeling berdasarkan rating. *Review* dengan rating 1, 2, atau 3 akan diberi label negatif , sedangkan *review* dengan rating 4 atau 5 diberi label positif (Syaripah et al., 2024) .

3.2.4 K-Fold Cross Validation

Cross validation adalah metode yang diterapkan membagi data menjadi dua bagian, yaitu data pelatihan dan pengujian. Data pelatihan digunakan untuk membangun model, sedangkan data pengujian digunakan untuk mengevaluasi kinerja model. Dalam validasi silang (*cross-validation*), data disilangkan berulang kali, memastikan setiap titik data diuji (Marutho, 2019).

3.2.5 TF IDF

Pembobotan TF (*Term Frequency*) - IDF (*Inverse Documen Frequency*), proses pembobotan TF IDF ini menghitung bobot kata yang sering digunakan dalam layanan informasi, metode ini lebih efesien dan menghasilkan hasil yang lebih akurat dengan menggabungkan perhitungan TF-IDF. Metode ini digunakan untuk menilai hubungan antara sebuah kata dan dokumen dengan menghitung frekuensi kemunculan kata dalam dokumen serta pentingnya kata tersebut dalam konteks dokumen. Konsep ini mengintegrasikan kedua metode untuk menilai relevansi kata dalam dokumen secara holistik (Agtira et al., 2023).

3.2.6 Klasifikasi

Data yang selesai diproses selanjutnya mengolah data menggunakan KNN dan NB.

Beberapa langkah dalam tahap klasifikasi menggunakan motode KNN adalah sebagai berikut:

1. Mencari jumlah jarak antara sampel yang tidak diketahui dengan semua sampel pada data latih menggunakan formula jarak tertentu, seperti jarak Euclidean atau jarak Manhattan.

2. Memilih k tetangga terdekat dari sampel yang tidak diketahui berdasarkan jarak yang telah dihitung.
3. Menghitung label kelas mayoritas dari k tetangga terdekat. Dalam klasifikasi biner, label mayoritas diperoleh dengan menghitung frekuensi setiap kelas pada k tetangga terdekat dan memilih kelas dengan frekuensi terbanyak. Sedangkan dalam klasifikasi multi kelas, label mayoritas ditentukan melalui voting, yaitu dengan menghitung jumlah suara dari setiap kelas pada k tetangga terdekat dan memilih kelas dengan jumlah suara terbanyak.
4. Mengembalikan label kelas mayoritas sebagai hasil klasifikasi untuk sampel yang tidak diketahui.

Tahapan klasifikasi menggunakan metode NB adalah sebagai berikut:

1. Sebelum melakukan klasifikasi, langkah pertama adalah menangani ketidakseimbangan data dan menerapkan pembobotan TF-IDF. Tujuannya adalah memastikan dataset yang dihasilkan dapat digunakan secara efektif dalam proses pemodelan klasifikasi.
2. Pembagian Dataset

   Setelah memperoleh dataset yang sudah diberi label, dataset dipisahkan menjadi dua bagian,yaitu data latih dan data uji. pembagian dilakukan dengan perbandingan 90:10 dari keseluruhan dataset yang telah diberi label.
3. Pembuatan Model Klasifikasi *Naïve Bayes*

   Setelah dataset training dan testing terbagi, langkah selanjutnya adalah membuat model klasifikasi *Naïve Bayes* menggunakan dataset *training*.

Pengujian model klasifikasi NB dan KNN dilakukan menggunakan dataset pengujian yang bertujuan untuk menguji model yang telah dibangun. Proses ini melibatkan perhitungan probabilitas kata dalam setiap kelas untuk melakukan prediksi data.

3.2.7 Evaluasi

Evaluasi model klasifikasi melibatkan beberapa metrik penting seperti akurasi, *precision*, dan *recall* (*sensitivitas*). Akurasi mengukur seberapa tepat model dalam memprediksi kelas data secara keseluruhan. *Precision* mengukur seberapa baik model dalam mengidentifikasi kelas positif yang sebenarnya positif, sedangkan sensivitas mengukur kemampuan model untuk mengidentifikasi kelas positif dengan benar (Rinanda et al., 2022). Dengan metrik-metrik ini, evaluasi model memberikan pemahaman yang jelas tentang kinerja dan keandalan prediksi model secara keseluruhan. *F1-score* adalah matrik yang mengevaluasi keseimbangan antara *precision* (seberapa banyak prediksi positif yang benar-benar positif) dan *recall* (seberapa baik model dalam mendeteksi semua contoh positif) memberikan gambaran komprehensif tentang performa model dalam mendeteksi contoh positif secara akurat. *F1-score* dihitung menggunakan rumus :

$$\text{f1} - \text{score} = 2X\frac{precision\ X\ \ recall}{precision + recall} \quad (3)$$

(Fikri et al., 2020)