

BAB III

METODOLOGI PENELITIAN

3.1 Jenis Penelitian

Penelitian ini dirancang dengan pendekatan kuantitatif bersifat eksperimental komparatif, di mana data dihimpun melalui penyebaran kuesioner terstruktur untuk kemudian diolah secara statistik. Tujuan utamanya adalah membandingkan tingkat performa tiga algoritma *machine learning* yaitu Naïve Bayes, K-Nearest Neighbor (KNN), dan Random Forest dalam memprediksi tingkat kecemasan sosial pada Gen Z. Pendekatan ini bersifat terapan karena hasilnya diarahkan untuk pengembangan sistem skrining kesehatan mental yang aplikatif, dan sekaligus tergolong cross-sectional karena pengumpulan data dilakukan pada satu waktu tertentu dari reponden Gen Z di Indonesia.

3.2 Fokus Penelitian

Fokus penelitian ini mencakup tiga aspek utama yang saling berkaitan untuk menjawab permasalahan penelitian secara komprehensif yang meliputi:

1. Membandingkan performa ketiga model klasifikasi melalui metrik akurasi, presisi, recall, F1-score, dan waktu komputasi.
2. Mengeksplorasi hubungan antara faktor gaya hidup digital (seperti durasi penggunaan media sosial, FoMO, perbandingan sosial daring dan lain lain) dengan tingkat kecemasan sosial.
3. Mengidentifikasi variabel prediktor yang paling berpengaruh melalui analisis korelasi dan *feature importance*.

3.3 Sumber Data

Sumber data penelitian ini berasal dari kuesioner daring yang dibagikan kepada Generasi Z di Indonesia berusia 13-28 tahun dengan jumlah total data yang terkumpul sebanyak 302 responden. Kuesioner disebarakan menggunakan Google

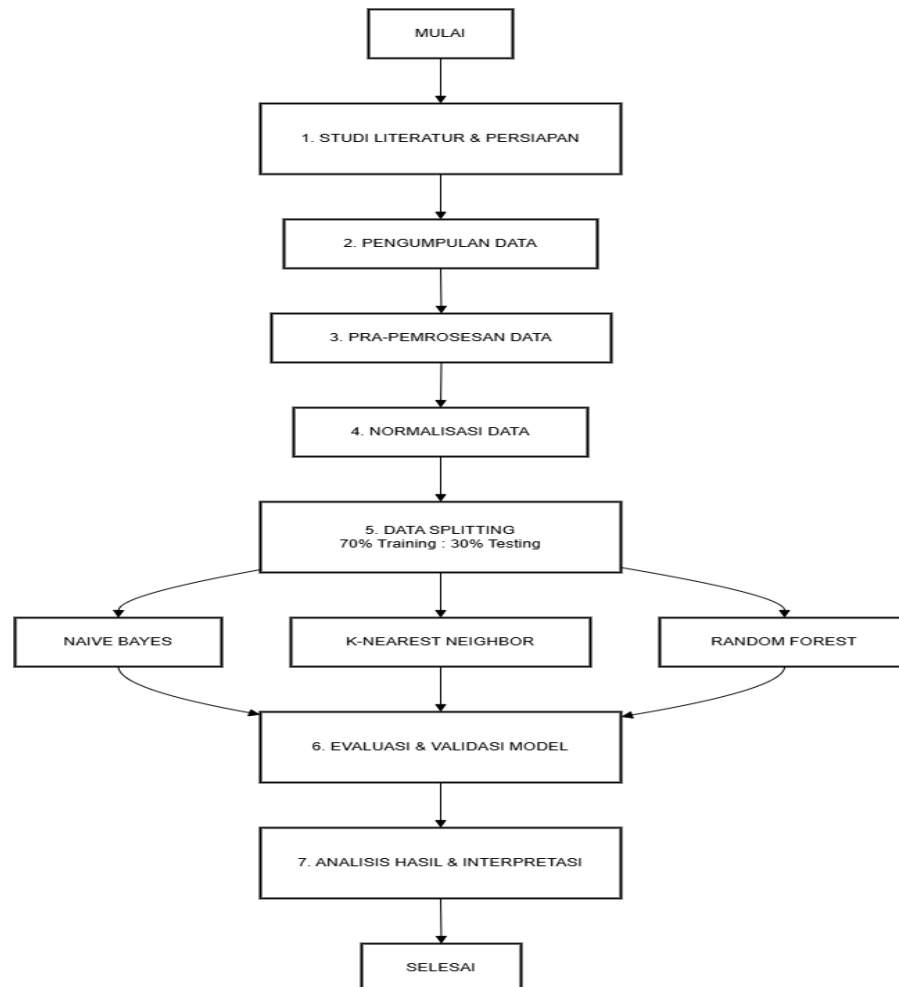
Forms dengan teknik purposive sampling, kuesioner ini dirancang untuk mengukur beberapa variabel independen meliputi lamanya waktu mengakses media sosial per hari (dihitung dalam jam), tingkat *Fear of Missing Out* (FoMO), serta seberapa sering seseorang membandingkan dirinya dengan pihak lain di media sosial. Secara operasional, ketiga aspek tersebut terangkum dalam satu instrument kuesioner adaptasi Liebowitz Social Anxiety Scale (LSAS) berjumlah 20 item, tanpa dipisahkan menjadi sub-skala tersendiri.

3.4 Alat dan Bahan Penelitian

Untuk mendukung pelaksanaannya, penelitian ini memanfaatkan perangkat keras dan perangkat lunak dengan rincian spesifikasi sebagai berikut:

1. Perangkat Keras (Hardware): Laptop dengan Processor Intel Core i5, RAM 8 GB, dan SSD 256 GB.
2. Perangkat Lunak (Software): Bahasa pemrograman Python, IDE Google Colab, serta pustaka pendukung seperti Pandas, Scikit-learn, Matplotlib, dan Seaborn.

3.5 Desain Sistem Penelitian



Gambar 3. 1 Flowchart Alur Penelitian

3.5.1 Studi Literatur Dan Persiapan

Studi literature untuk membangun landasan yang kuat sebelum penelitian dilaksanakan. Pada tahap ini, peneliti mengumpulkan referensi ilmiah berupa jurnal, prosiding, dan artikel yang membahas algoritma Naïve Bayes dan K-Nearest Neighbor (KNN), intrusmen pengukuran kecemasan sosial, serta penelitian-penelitian terdahulu yang relevan sebagai bahan perbandingan metodologi. Hasil studi literature digunakan untuk menyusun kisi-kisi kuesioner, memilih instrument pengukuran kecemasan sosial yang sesuai, dan merancang prosedur preprocessing serta evaluasi model.

3.5.2 Pengumpulan Data

Proses pengumpulan data ditempuh melalui kuesioner online (*Google Forms*) dengan teknik *purposive sampling* dan *snowball sampling*. *Purposive sampling* adalah teknik penentuan sampel berdasarkan kriteria tertentu yang telah ditetapkan peneliti, sedangkan *snowball sampling* merupakan cara penjarangan sampel yang berlangsung secara beruntun, yakni responden pertama diminta memberikan rekomendasi responden berikutnya yang memenuhi kriteria penelitian. Kedua teknik tersebut menyasar Gen Z usia 13-28 tahun yang tergolong aktif mengakses media sosial dengan durasi minimal 1 jam setiap harinya. Sebelum mulai mengisi kuesioner, setiap responden terlebih dahulu diminta untuk membaca serta menyetujui informed consent yang memuat penjelasan mengenai tujuan penelitian dan jaminan atas kerahasiaan data yang diberikan.

Kuesioner dikonfigurasi agar seluruh pertanyaan wajib dijawab sehingga tidak ada data kosong saat dikirimkan. Pemantauan berkala dilakukan selama proses pengumpulan data hingga jumlah responden sesuai target.

3.5.3 Pra-pemrosesan Data

Tahap preprocessing sangat krusial untuk memastikan kualitas data sebelum digunakan dalam proses training model. Data mentah hasil kuesioner harus melalui serangkaian proses pembersihan dan transformasi untuk menghasilkan

dataset yang siap digunakan dalam algoritma machine learning. Tahapan preprocessing meliputi lima sub-tahapan utama:

1. Data Cleaning (Pembersihan Data): Menghapus data responden yang tidak lengkap dan menghapus data duplikat.
2. Data Transformation (Transformasi Data): Mengonversi jawaban responden yang bersifat teks menjadi format numerik.
3. Data Labeling (Pelabelan Data): Menentukan kelas target untuk setiap responden, yaitu tingkat kecemasan sosial (misalnya: Non-klinis, Ringan, Sedang dan Berat).
4. Normalisasi Data: Penerapan metode *Min-Max Scaling* dilakukan untuk menyetarakan seluruh rentang skor variabel ke dalam skala $[0, 1]$. Hal ini dilakukan agar setiap fitur gaya hidup digital memiliki bobot yang seimbang dan mencegah variabel angka besar mendominasi perhitungan jarak pada algoritma KNN.

3.5.4 Data Splitting

Setelah data diproses, data kemudian dibagi menjadi dua kelompok melalui teknik data splitting, dengan komposisi 70% dialokasikan sebagai data training dan 30% sisanya sebagai data testing. Data training berfungsi untuk melatih model agar mampu mengenali pola yang terkandung dalam data, sementara data testing dimanfaatkan untuk menguji sejauh mana kemampuan generalisasi model terhadap data baru yang belum pernah ditemuinya sebelumnya.

3.5.5 Pembentukan Model Klasifikasi

Pada tahap ini, dua algoritma klasifikasi dijalankan secara paralel menggunakan data training yang telah disiapkan sebelumnya.

1. Naïve Bayes diterapkan untuk menyusun model klasifikasi yang didasarkan pada probabilitas kemunculan setiap fitur pada masing-masing kelas. Model tersebut terlebih dahulu dilatih menggunakan data training, sebelum akhirnya dimanfaatkan untuk memprediksi kelas pada data uji.
2. K-Nearest Neighbor (KNN) diterapkan dengan cara mengukur jarak antara masing-masing data uji terhadap seluruh data training

menggunakan *Euclidean distance*. Sejumlah K tetangga terdekat kemudian diidentifikasi, dan kelas yang paling sering muncul diantara tetangga tersebut ditetapkan sebagai hasil prediksi.

3. Random Forest bekerja dengan cara membentuk sejumlah pohon keputusan (decision tree) secara bersamaan, di mana masing-masing pohon dilatih menggunakan sebagian data training beserta sebagian fitur yang dipilih secara acak melalui teknik bootstrap aggregating (bagging). Tiap pohon tersebut menghasilkan prediksi kelasnya masing-masing secara terpisah, sebelum akhirnya hasil klasifikasi akhir ditentukan berdasarkan suara terbanyak (majority voting) dari keseluruhan pohon yang telah dibangun. Cara kerja semacam ini berkontribusi dalam menekan risiko overfitting sekaligus meningkatkan akurasi dan kestabilan model dibandingkan apabila hanya mengandalkan satu pohon keputusan saja.

3.5.6 Evaluasi & Validasi Model

Setelah model selesai, dilakukan tahap evaluasi dan validasi menggunakan data testing. Pada tahap ini, performa masing-masing model diukur melalui beberapa metrik evaluasi yang lazim dipakai dalam klasifikasi, antara lain:

1. Akurasi: proporsi prediksi yang benar dari keseluruhan data uji.
2. Presisi (*Precision*): sejauh mana model mampu meminimalkan kesalahan pada prediksi positif.
3. Recall: sejauh mana model mampu menangkap seluruh kelas positif yang sebenarnya ada.
4. F1-Score: nilai rata-rata harmonis dari presisi dan recall yang menggambarkan performa secara menyeluruh.
5. Confusion Matrix: tabel yang merangkum hasil prediksi model secara visual untuk setiap kelas.

Validasi dilakukan untuk memastikan bahwa model tidak mengalami *overfitting* maupun *underfitting*, sehingga hasil yang diperoleh dapat diandalkan dan bersifat general.

3.5.1 Analisis Hasil dan Interpretasi

Pada tahap ini, peneliti membandingkan secara mendalam hasil evaluasi dari ketiga algoritma, yaitu: Naïve Bayes, KNN dan Random Forest, untuk menentukan model mana yang menunjukkan performa terbaik dalam menyelesaikan permasalahan klasifikasi yang diteliti.