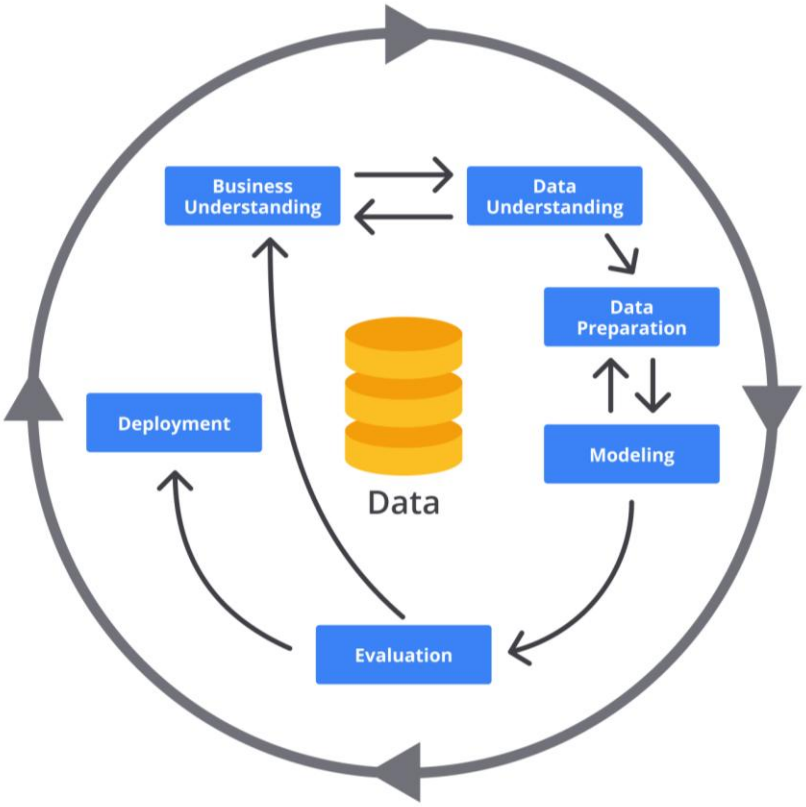
BAB III
METODE PENELITIAN

3.1 Alur Penelitian

Penelitian ini menggunakan pendekatan kuantitatif dengan metode eksperimen komputasional untuk mengevaluasi kinerja algoritma *Support Vector Machine* (SVM) dalam memprediksi kanker paru-paru. Tahapan penelitian mengikuti kerangka kerja CRISP-DM *(Cross Industry Standard Process for Data Mining)* yang terdiri dari enam tahap utama: *business understanding*, *data understanding*, *data preparation*, *modeling*, *evaluation*, dan *deployment*[22]:



Gambar 3. 1 Bagan Alur Penelitian

Sumber : www.dicoding.com

3.2 Sumber Data

Data yang digunakan dalam penelitian ini merupakan *dataset* publik tentang gejala dan risiko penyakit paru-paru yang diperoleh dari platform Kaggle. *Dataset* ini terdiri dari 5000 entri dan 19 atribut (18 fitur prediktor dan 1 kelas target).

Tabel 3.1 menyajikan deskripsi atribut yang digunakan dalam penelitian ini.

Tabel 3. 1 Metadata

No.	Nama Kolom	Tipe Data	Penjelasan Data	Contoh Data
1	Age	Integer	Usia pasien dalam tahun.	68
2	Gender	Integer (0/1)	Jenis kelamin pasien (0 = Perempuan, 1 = Laki-laki).	1
3	Smoking	Integer (0/1)	Apakah pasien adalah perokok (1 = Ya, 0 = Tidak).	1
4	Finger_Discoloration	Integer (0/1)	Apakah terdapat perubahan warna pada jari (1 = Ya, 0 = Tidak).	1
5	Mental_Stress	Integer (0/1)	Apakah pasien mengalami stres mental (1 = Ya, 0 = Tidak).	1
6	Exposure_To_Pollution	Integer (0/1)	Terpapar polusi dalam jangka waktu tertentu (1 = Ya, 0 = Tidak).	1
7	Long_Term_Illness	Integer (0/1)	Mengidap penyakit kronis lain (1 = Ya, 0 = Tidak).	0
8	Energy_Level	Float	Tingkat energi pasien (skala numerik).	57.83
9	Immune_Weakness	Integer (0/1)	Kondisi sistem imun lemah (1 = Ya, 0 = Tidak).	0
10	Breathing_Issue	Integer (0/1)	Masalah pernapasan yang dialami pasien (1 = Ya, 0 = Tidak).	0
11	Alcohol_Consumption	Integer (0/1)	Apakah pasien mengonsumsi alkohol (1 = Ya, 0 = Tidak).	1
12	Throat_Discomfort	Integer (0/1)	Mengalami ketidaknyamanan di tenggorokan (1 = Ya, 0 = Tidak).	1
13	Oxygen_Saturation	Float	Persentase saturasi oksigen dalam darah (%).	95.98
14	Chest_Tightness	Integer (0/1)	Dada terasa sesak atau berat (1 = Ya, 0 = Tidak).	1

15	Family_History	Integer (0/1)	Riwayat keluarga dengan penyakit paru-paru (1 = Ya, 0 = Tidak).	0
16	Smoking_Family_History	Integer (0/1)	Riwayat merokok dalam keluarga (1 = Ya, 0 = Tidak).	0
17	Stress_Immune	Integer (0/1)	Gangguan sistem imun akibat stres (1 = Ya, 0 = Tidak).	0
18	Pulmonary_Disease **(Label Target)**	Kategori (YES/NO)	Target: Apakah pasien didiagnosis dengan penyakit paru-paru. (YES atau NO).	NO

3.3 Data Understanding

Tahap ini bertujuan untuk mengenal karakteristik data, memeriksa kualitasnya, dan memperoleh wawasan awal melalui Analisis Data Eksploratif (EDA) sebelum data diproses lebih lanjut.

3.3.1 Pemeriksaan Kualitas Data (Rencana)

Pada tahap ini, akan dilakukan pemeriksaan kualitas data, meliputi pengecekan nilai hilang (*missing values*), data duplikat (*duplicate data*), dan inkonsistensi data. Hasil temuan detail dari pemeriksaan ini akan disajikan pada Bab IV.

3.3.2 Analisis Data Eksploratif (EDA)

Analisis Data Eksploratif (EDA) direncanakan untuk mendapatkan wawasan awal mengenai karakteristik data dan hubungan antar variabel. EDA meliputi:

a. Analisis Distribusi Variabel Target

Akan dihitung proporsi (persentase) kelas target (*Pulmonary_Disease*: YES dan NO) untuk mengidentifikasi adanya masalah ketidakseimbangan kelas (*imbalanced data*).

b. Analisis Statistika Deskriptif

Melakukan perhitungan statistik dasar (*mean*, *median*, *modus*, standar deviasi, *min*, *max*) untuk variabel numerik (*Energy_Level* dan *Oxygen_Saturation*) untuk mengidentifikasi sebaran data dan potensi *outlier*.

c. Analisis Bivariat Awal

Menyelidiki hubungan antara setiap variabel prediktor dengan variabel target. Untuk variabel kategorikal, akan digunakan visualisasi seperti *bar chart* untuk membandingkan persentase kasus penyakit paru-paru pada setiap kategori fitur.

3.4 Data Preparation

Tahap ini mencakup semua aktivitas yang dilakukan untuk membersihkan, mengubah, dan memformat data mentah menjadi *input* yang sesuai untuk algoritma SVM.

3.4.1 Transformasi Data

a. Normalisasi Data

Fitur numerik akan diubah skalanya menggunakan metode Normalisasi *Min-Max* untuk menghasilkan nilai dalam rentang [0, 1]. Normalisasi *Min-Max* dipilih karena model SVM sangat sensitif terhadap perbedaan skala fitur, sehingga normalisasi memastikan kontribusi yang setara dari semua fitur.

b. *Encoding* Data

Fitur kategorikal (*Pulmonary_Disease*) dan semua fitur biner akan diubah menjadi representasi numerik.

3.4.2 Seleksi Fitur (*Feature Selection*)

Akan dilakukan pemilihan fitur optimal menggunakan metode *Recursive Feature Elimination* (RFE) yang dikombinasikan dengan algoritma SVM. RFE dipilih karena secara iteratif melatih model dan menghilangkan fitur yang kurang penting, sehingga menghasilkan subset fitur yang paling berpengaruh.

3.4.3 Pembagian Data (*Data Splitting*)

Dataset yang telah disiapkan akan dibagi menjadi data latih dan data uji dengan rasio [80:20]. Rasio [80:20] dipilih untuk menyeimbangkan kebutuhan data yang cukup untuk melatih model (*training*) dan data yang independen untuk pengujian kinerja (*testing*).

3.5 Modeling

Tahap pemodelan melibatkan penerapan algoritma *Support Vector Machine* (SVM) dan mengoptimalkannya menggunakan *Hyperparameter Tuning*.

3.5.1 Algoritma Support Vector Machine (SVM)

SVM adalah algoritma klasifikasi yang bertujuan menemukan optimal *hyperplane* (batas keputusan) yang memiliki *margin* terbesar antara kelas-kelas data. Titik data terdekat dengan *hyperplane* disebut Support Vector. Untuk data non-linear, digunakan Fungsi *Kernel* yang memetakan data ke ruang dimensi yang lebih tinggi agar dapat dipisahkan secara linier (trik *kernel*).

3.5.2 Pemilihan Fungsi Kernel dan Hyperparameter Tuning

a. Pemilihan Fungsi *Kernel*

Fungsi *Kernel* adalah teknik krusial dalam SVM yang memungkinkan algoritma melakukan klasifikasi non-linear. Fungsi *kernel* bertugas untuk memetakan data tersebut ke ruang dimensi yang lebih tinggi. Di ruang dimensi baru ini, data diharapkan dapat dipisahkan secara linier, sehingga *hyperplane* dapat ditemukan.

Beberapa jenis fungsi *kernel* yang umum digunakan meliputi:

1. *Linear Kernel:* (Digunakan dalam penelitian ini) Digunakan ketika data dapat dipisahkan secara linier.
2. *Polynomial Kernel:* Cocok untuk data dengan batas keputusan melengkung.

3. *Radial Basis Function (RBF) Kernel:* (Digunakan dalam penelitian ini). RBF adalah pilihan *default* yang sangat populer dan efektif untuk kasus non-linear, karena mampu memetakan data ke ruang fitur berdimensi tak hingga.

b. *Hyperparameter Tuning*

Optimalisasi model SVM RBF akan dilakukan dengan mencari nilai optimal untuk *hyperparameter* C (parameter regularisasi) dan Gamma menggunakan metode Grid Search Cross-Validation (GridSearchCV). Rencananya rentang nilai *hyperparameter* yang akan diuji adalah: C: [misal: 0.1, 1, 10, 100] dan gamma: [misal: $0.001, 0.01, 0.1, 1$]. Akan digunakan K-fold, yaitu 10-fold *Cross-Validation*.

3.6 Evaluation

Evaluasi dilakukan untuk mengetahui seberapa baik model dalam mengklasifikasikan data. Metrik yang digunakan adalah:

- *Accuracy*: Tingkat prediksi benar terhadap total data.
- *Precision*: Kemampuan model dalam mengidentifikasi kelas positif secara tepat.
- *Recall:* Kemampuan model dalam menangkap seluruh kelas positif.
- *F1-Score:* Rata-rata harmonis dari precision dan recall.
- *Confusion Matrix*: Perbandingan prediksi model terhadap label aktual dari data.

Metrik *Precision, Recall, dan F1-Score* dipilih selain Akurasi, karena dapat memberikan gambaran yang lebih akurat mengenai kinerja model, terutama dalam konteks penyakit di mana *Recall* (kemampuan model untuk menemukan semua kasus positif/penyakit) sangat penting. Matriks kebingungan (*Confusion Matrix*) akan digunakan sebagai dasar perhitungan semua metrik ini.

3.7 Deployment

Tahap deployment merupakan proses penerapan hasil model ke dalam sistem nyata atau bentuk dokumentasi untuk dapat digunakan dalam pengambilan keputusan atau aplikasi operasional. Dalam penelitian ini, deployment dilakukan secara terbatas dalam bentuk:

- Dokumentasi model terbaik (SVM dengan kernel RBF dan hyperparameter tuning)
- Visualisasi hasil evaluasi model (classification report, confusion matrix, learning curve)
- Penyimpanan model dalam bentuk file .pkl menggunakan Python untuk keperluan implementasi selanjutnya.

Meski tidak dilakukan integrasi ke sistem informasi klinik atau aplikasi web, hasil model dapat digunakan oleh tenaga medis atau peneliti sebagai referensi awal dalam proses diagnosa dini berbasis data.