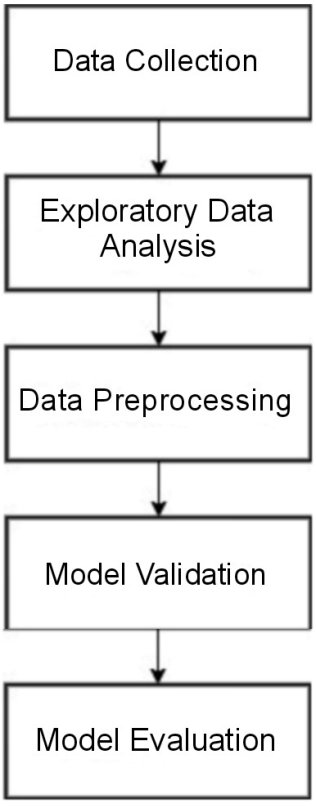
BAB III
METODOLOGI PENELITIAN

3.1 Alur Penelitian

Metode penelitian yang digunakan pada penelitian ini menggunakan 3 algoritma yaitu *Random Forest, C.5.0, Support Vector Machine*. Adapun tahapan penelitian yang dilakukan oleh penulis sebagai berikut :



Gambar 3.1 Alur Penelitian

3.1.1 *Data Collection*

Penelitian ini menggunakan dataset penyakit *liver* yang diperoleh *Kaggle repository*. Dataset yang digunakan berjumlah 30.692 data dalam bentuk *comma separated values* (csv). Pada data tersebut terdiri dari 11 *column* untuk Kolom sebagai *predicted* adalah usia pasien, jenis kelamin pasien, jumlah bilirubin total, jumlah bilirubin langsung, tingkat alkali fosfatase, tingkat alanin aminotransferase, tingkat aspartat aminotransferase, jumlah protein total, jumlah albumin dan rasio albumin dan globulin, Sedangkan *column* sebagai target adalah *column Result*. Berikut lima contoh entri dari dataset yang digunakan dalam penelitian.

Tabel 3.1 *Data Sample*

	Age of the patient	Gender of the patient	Total Bilirubin	Direct Bilirubin	Alkphos Alkaline Phosphotase	Sgpt Alamine Aminotransferase	Sgot Aspartate Aminotransferase	Total Protiens	ALB Albumin	A/G Ratio Albumin and Globulin Ratio	Result
0	65.0	Female	0.7	0.1	187.0	16.0	18.0	6.8	3.3	0.90	1
1	62.0	Male	10.9	5.5	699.0	64.0	100.0	7.5	3.2	0.74	1
2	62.0	Male	7.3	4.1	490.0	60.0	68.0	7.0	3.3	0.89	1
3	58.0	Male	1.0	0.4	182.0	14.0	20.0	6.8	3.4	1.00	1
4	72.0	Male	3.9	2.0	195.0	27.0	59.0	7.3	2.4	0.40	1

Tabel 3.2 *Metadata*

Atribute	*Description*	*Value*	*Type Data*
Age of the Patient	Usia pasien (dalam tahun)	*4 - 90*	*Numerical*
Gender of the Patient	Jenis kelamin pasien	Male dan Female	*Categorical*

Total Bilirubin	Jumlah bilirubin total (mg/dl)	0.4 – 75.0	*Numerical*
Direct Bilirubin	Jumlah bilirubin langsung (mg/dl)	0.1 – 19.7	*Numerical*
Alkphos Alkaline Phosphotase	Tingkat Alkali Fosfatase (IU/L)	63.0 - 2110.0	*Numerical*
Sgpt Alamine Aminotransferase	Tingkat Alanin Aminotransferase (IU/L)	10.0 - 2000.0	*Numerical*
Sgot Aspartate Aminotransferase	Tingkat Aspartat Aminotransferase (IU/L)	10.0 - 4929.0	*Numerical*
Total Protiens	Jumlah protein total (g/dL))	2.7 - 9.6	*Numerical*
ALB Albumin	Jumlah albumin (g/dL)	0.9 - 5.5	*Numerical*
A/G Ratio Albumin and Globulin Ratio	Rasio Albumin dan Globulin	0.3 - 2.8	*Numerical*

Result	Hasil diagnostik (1 = liver disease, 0 = healthy)	1 atau 2	*Categorical*

3.1.2 *Exploratory Data Analysis*

Dalam penelitian ini terdiri dalam beberapa langkah pengerjaan yaitu tahapan *analyzing* dataset kemudian dilanjutkan dengan deskripsi dataset menggunakan EDA *(Exploratory Data Analysis),* dengan melakukan EDA, akan sangat berguna dalam mendeteksi kesalahan dari awal, dapat mengidentifikasi *outlier,* mengetahui hubungan antar data serta dapat menggali faktor-faktor penting dari data [1]. Data yang telah dikumpulkan akan dianalisis terlebih dahulu untuk mengetahui karakteristik data yang memilikki nilai yang duplikat, null, jumlah data dan tipe data yang tidak konsisten, fungsi lain EDA adalah mengenali kesalahan dataset dengan menguasai pola suatu data dan menemukan hubungan antar variabel**.**[1]

Dalam melakukan analisa data yang terkait dengan resiko stunting pada balita dengan menerapkan *Exploratory data analysis* (EDA) seperti teknik visualisasi sebagai berikut :

a. *Data Understanding*

   Proses EDA diawali dengan memahami struktur dan tipe data yang tersedia, termasuk pemeriksaan dimensi dataset serta klasifikasi variabel sebagai numerik atau kategorikal. Pemahaman ini penting untuk memberikan wawasan awal mengenai karakteristik data yang akan dianalisis.

b. Duplicate

Dilakukan pemeriksaan pada dataset yang bertujuan untuk mengidentifikasi adanya data duplikat. Data duplikat dapat menyebabkan bias pada model dan memengaruhi akurasi prediksi, sehingga penting untuk memastikan bahwa setiap baris data bersifat unik. Setelah dilakukan pemeriksaan, dari 30691 data ditemukan duplikat berjumlah 11323 data yang kemudian dilakukan pengahapusan dan tersisa 19368 data unik.

c. Identifikasi *Missing Values*

memeriksa apakah terdapat nilai yang hilang atau null dan nilai duplikat pada data, untuk itu diperlukan identifikasi *missing values* untuk meningkatkan akurasi kinerja model guna meningkatkan keakuratan data. Hasil pemeriksaan menunjukkan bahwa dataset memiliki 5.425 *missing values.* Oleh karena itu, dilakukan penghapusan data yang mengandung nilai kosong, sehingga dari total 19.368 data, tersisa 16.389 data yang akan digunakan untuk analisis lebih lanjut.

d. Outlier

Dalam penelitian ini, pemeriksaan *outlier* menggunakan IQR *(Inter Quartile Range)* digunakan untuk mendeteksi *outlier* dalam data dengan cara menentukan batas bawah *(lower)* dan batas atas *(upper)* dimana data dianggap sebagai *outlier.* Dalam memprediksi kasus penyakit *liver* ini *outlier* dapat memengaruhi performa model secara negative dan memungkinkan adanya *overfitting*, sehingga diputuskan untuk menghapusnya agar model dapat mengenali pola data secara lebih optimal. Setelah dilakukan pemeriksaan, dari 16389 data ditemukan *Outlier* berjumlah 5241 data yang kemudian dilakukan pengahapusan dan tersisa 11148 data.

3.1.3 *Data Preprocessing*

Pada tahapan ini akan dilakukan persiapan data yaitu :

a. Encoding fitur kategori

   Proses mengubah suatu fitur yang memiliki nilai kategori, dimana nilai dari setiap kolom akan diganti menjadi angka. Pada tahapan ini adanya pengubahan format numerik agar dapat digunakan dalam algoritma mesin. Dengan mengubah kategori menjadi 0 dan 1.[17]

b. Data *Splitting*

   proses memisahkan data latih dan data uji, yaitu dataset menjadi dua bagian yaitu data latih (training) dan data uji (testing). Data latih digunakan untuk melatih model agar mengenali pola dan hubungan dalam data. Dalam penelitian ini, proses split ini ditentukan data *training* 70% dan data *testing* 30%. Langkah ini penting agar model dapat menggeneralisasi dengan baik dan tidak hanya berfokus pada data latih, sehingga menghasilkan prediksi yang akurat saat digunakan.[14]

3.1.4 *Model Validation*

Validasi model dilakukan untuk mengevaluasi performa algoritma yang digunakan dalam penelitian ini, yaitu *random forest*, C 5.0 , dan SVM. Langkah-langkah validasi yang dilakukan adalah sebagai berikut:

1. *Training*

   Dalam proses pelatihan model, parameter disesuaikan untuk mengurangi kesalahan dengan memanfaatkan data pelatihan guna mengenali pola atau hubungan antara fitur dan target yang berjumlah 7.803 *record.* Setelah model selesai dilatih, metrik evaluasi digunakan untuk mengukur kinerjanya pada data pelatihan serta data pengujian.

2. *Cross validation*

   Pada tahap ini, evaluasi kinerja model dilakukan menggunakan teknik 10-fold *cross-validation* di RapidMiner dengan proporsi pembagian data 70% data digunakan sebagai data pelatihan, sementara 30% sisanya digunakan sebagai data pengujian. Proses ini diulang sebanyak 10 kali, di mana setiap bagian dataset akan menjadi data uji satu kali, sedangkan sembilan bagian lainnya digunakan sebagai data pelatihan. Pendekatan ini memastikan bahwa model diuji secara menyeluruh pada berbagai bagian dataset, sehingga menghasilkan evaluasi yang lebih stabil dan mengurangi kemungkinan bias akibat pembagian data yang kurang representatif.

3. Metrik Evaluasi Untuk mengukur performa setiap model, digunakan beberapa metrik evaluasi berikut:

- Akurasi: Mengukur persentase prediksi yang benar terhadap total data testing.
- *Precision*: Tingkat keakuratan model dalam memprediksi kelas positif (kasus stunting).
- *Recall*: Kemampuan model dalam mendeteksi seluruh kasus positif (sensitivitas).

3.1.5 *Model Evaluation*

Pada fase ini dilakukan proses pembentukan keluaran yang mudah dimengerti, dimana tahapan ini melakukan penilaian kinerja model dan memastikan model dengan mengukur performa model dengan menggunakan confusion matrix, yang dapat membantu mengidentifikasi model dalam memprediksi kelas. Kinerja model dievaluasi menggunakan matrix seperti akurasi*, precision*, dan *recall*[18].

3.1.6 Perhitungan Manual

https://drive.google.com/drive/folders/18tFvomLxLz56wXlYTDGnurEWLqgsZD7?usp=sharing