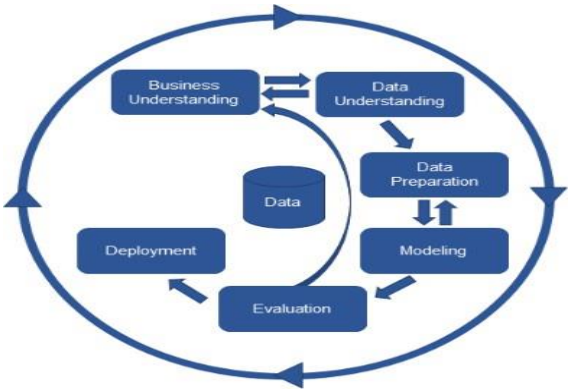
# BAB III
# METODOLOGI PENELITIAN

## 3.1 Alur Penelitian

Penelitian ini dilakukan untuk merancang sebuah model yang dapat dikembangkan untuk meningkatkan akurasi dalam memprediksi *alzheimer disease (AD)* berdasarkan algoritma *random forest*. Berikut adalah tahapan penelitian yang dilaksanakan untuk mengembangkan model *machine learning* dengan memanfaatkan model *Cross Industry Standard Process For Data Mining* [33].




Gambar 3. 1 Tahapan metode crisp-dm

## 3.2 *Business Understanding*

a. *Problem Statement*

Berdasarkan uraian latar belakang di atas, permasalahan yang dapat diselesaikan pada proyek ini adalah:

1. Dari fitur yang tersedia, fitur mana yang paling berpengaruh dalam memprediksi risiko *alzheimer disease (AD)*?
2. Bagaimana model *machine learning* dapat digunakan untuk memprediksi risiko alzheimer secara akurat dan tepat sehingga dapat membantu tenaga medis dalam deteksi dini dan upaya pencegahan di masa depan?

b. *Goals*

1. Mendapatkan analisis yang mendalam mengenai prediksi risiko *alzheimer disease (AD)*.

2. Memprediksi seseorang berisiko terkena alzheimer dengan tingkat akurasi di atas 85%.

*c. Solution Statement*

Solusi yang dapat diterapkan untuk menyelesaikan permasalahan ini meliputi:

1. Melakukan analisis data terkait *alzheimer disease (AD)* dengan menerapkan *Exploratory Data Analysis* (EDA), seperti teknik visualisasi. Adapun langkah-langkah analisis yang dilakukan adalah:
   a) Melakukan pre-processing data.
   b) Mengeksplorasi korelasi antara fitur dengan variabel target.
   c) Menangani *outlier*.
2. Melakukan persiapan data, seperti:
   a) Encoding fitur kategori.
   b) Splitting data untuk pelatihan dan pengujian.
   c) Normalisasi atau standarisasi data agar dapat digunakan dalam model *machine learning.*

### 3.3 *Data Understanding*

Penelitian ini menggunakan dataset *alzheimer disease (AD)* yang diperoleh dari repositori *Kaggle*. Dataset yang digunakan berjumlah 2149 data dalam format *comma separated values* (csv). Pada data tersebut terdiri dari 32 *column feature* dan satu *column target*. Adapun *column feature* adalah usia, jenis kelamin, etnisitas, tingkat pendidikan, indeks massa tubuh (imt), merokok, konsumsi alkohol, aktivitas fisik, kualitas pola makan, kualitas tidur, riwayat keluarga alzheimer, penyakit kardiovaskular, diabetes, depresi, cedera kepala, hipertensi, tekanan darah sistolik, tekanan darah diastolik, kolesterol total, kolesterol ldl, kolesterol hdl, trigliserida, mmse, penilaian fungsional, keluhan memori, masalah perilaku, adl, kebingungan, disorientasi, perubahan kepribadian, kesulitan menyelesaikan tugas, lupa. Sedangkan, *column target* bernama diagnosis.

Tabel 3. 1 Metadata

| No | Fitur | Tipe Data | Keterangan | Rentang Nilai |
|---|---|---|---|---|
| 1 | *Age* | *Continues* | Usia Pasien | 60 - 90 (Tahun) |
| 2 | *Gender* | *Categorical* | Jenis Kelamin Pasien | (0 = Pria, 1 = Wanita) |
| 3 | *Ethnicity* | *Categorical* | Etnisitas Pasien | Dikodekan Sebagai Berikut:<br>• 0: Caucasian<br>• 1: African American<br>• 2: Asian<br>• 3: *Other* |
| 4 | *Education Level* | *Categorical* | Tingkat Pendidikan Pasien | Dikodekan Sebagai Berikut:<br>• 0: *None*<br>• 1: *High School*<br>• 2: *Bachelor's*<br>• 3: *Higher* |
| 5 | *BMI* | *Continues* | Indeks Massa Tubuh Pasien | Mulai Dari 15 Hingga 40 |
| 6 | *Smoking* | *Categorical* | Status Merokok | (0 = Tidak, 1 = Ya) |
| 7 | *Alcohol Consumption* | *Continues* | Konsumsi Alkohol Mingguan Dalam Satuan, | Mulai Dari 0 Hingga 20 |
| 8 | *Physical Activity* | *Continues* | Aktivitas Fisik Mingguan Dalam Jam | Mulai Dari 0 Hingga 10 |
| 9 | *Diet Quality* | *Continues* | Skor Kualitas Diet | Mulai Dari 0 Hingga 10 |
| 10 | *Sleep Quality* | *Continues* | Skor Kualitas Tidur | Mulai Dari 4 Hingga 10 |
| 11 | *Family History Alzheimers* | *Categorical* | Riwayat Keluarga Dengan Penyakit Alzheimer | ( 0 = Tidak, 1 = Ya) |
| 12 | *Cardiovascular Disease* | *Categorical* | Adanya Penyakit Kardiovaskular | ( 0 = Tidak, 1 = Ya) |
| 13 | *Diabetes* | *Categorical* | Adanya Diabetes | (0 = Tidak, 1 = Ya) |
| 14 | *Depression* | *Categorical* | Adanya Depresi | (0 = Tidak, 1 = Ya) |
| 15 | *HeadInjury* | *Categorical* | Riwayat Cedera Kepala | ( 0 = Tidak, 1 = Ya) |

| | | | | |
|---|---|---|---|---|
| 16 | *Hypertension* | *Categorical* | Adanya Hipertensi, | (0 = Tidak, 1 = Ya) |
| 17 | *Systolic BP* | *Continues* | Tekanan Darah Sistolik | Mulai Dari 90 Hingga 180 Mmhg |
| 18 | *Diastolic BP* | *Continues* | Tekanan Darah Diastolik | Berkisar Antara 60 Hingga 120 Mmhg. |
| 19 | *Cholesterol Total* | *Continues* | Kadar Kolesterol Total | Mulai Dari 150 Hingga 300 Mg/Dl. |
| 20 | *Cholesterol LDL* | *Continues* | Kadar Kolesterol Lipoprotein Densitas Rendah | Berkisar Antara 50 Hingga 200 Mg/Dl |
| 21 | *Cholesterol HDL* | *Continues* | Kadar Kolesterol Lipoprotein Densitas Tinggi | Berkisar Antara 20 Hingga 100 Mg/Dl. |
| 22 | *CholesterolTriglycerides* | *Continues* | Kadar Trigliserida | Mulai Dari 50 Hingga 400 Mg/Dl. |
| 23 | *MMSE* | *Continues* | Skor Mini-Mental State Examination | Mulai Dari 0 Hingga 30. Skor Yang Lebih Rendah Menunjukkan Gangguan Kognitif. |
| 24 | *FunctionalAssessment* | *Continues* | Skor Penilaian Fungsional, Mulai Dari 0 Hingga 10 | Skor Yang Lebih Rendah Menunjukkan Gangguan Yang Lebih Besar. |
| 25 | *MemoryComplaints* | *Categorical* | Adanya Keluhan Memori | ( 0 = Tidak, 1 = Ya) |
| 26 | *BehavioralProblems* | *Categorical* | Adanya Masalah Perilaku | (0 = Tidak, 1 = Ya) |
| 27 | *ADL* | *Continues* | Skor Aktivitas Kehidupan Sehari-Hari | Mulai Dari 0 Hingga 10. Skor Yang Lebih Rendah Menunjukkan Gangguan Yang Lebih Besar. |
| 28 | *Confusion* | *Categorical* | Adanya Kebingungan | ( 0 = Tidak, 1 = Ya) |
| 29 | *Disorientation* | *Categorical* | Adanya Disorientasi | ( 0 = Tidak, 1 = Ya) |
| 30 | *PersonalityChanges* | *Categorical* | Adanya Perubahan Kepribadian | ( 0 = Tidak, 1 = Ya) |

| | | | | |
|---|---|---|---|---|
| 31 | *DifficultyCompletingTasks* | *Categorical* | Adanya Kesulitan Dalam Menyelesaikan Tugas | (0 = Tidak, 1 = Ya) |
| 32 | *Forgetfulness* | *Categorical* | Adanya Kelupaan | (0 = Tidak, 1 = Ya) |
| 33 | *Diagnosis* | *Categorical* | Status Diagnosis Untuk Penyakit Alzheimer | (0 = Tidak, 1 = Ya) |

| Age | Gender | Ethnicity | EducationLevel | BMI | Smoking | AlcoholConsumption | PhysicalActivity | DietQuality | SleepQuality | FamilyHistoryAlzheim | CardiovascularDisea |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 73 | 0 | 0 | 2 | 22.92774923 | 0 | 13.29721773 | 6.327112474 | 1.347214306 | 9.025678666 | 0 | 0 |
| 89 | 0 | 0 | 0 | 26.82768119 | 0 | 4.542523818 | 7.61988454 | 0.5187671387 | 7.151292743 | 0 | 0 |
| 73 | 0 | 3 | 1 | 17.79588244 | 0 | 19.55508453 | 7.844987791 | 1.826334665 | 9.673574158 | 1 | 0 |
| 74 | 1 | 0 | 1 | 33.80081704 | 1 | 12.20926555 | 8.42800135 | 7.43560414 | 8.392553685 | 0 | 0 |
| 89 | 0 | 0 | 0 | 20.71697383 | 0 | 18.45435609 | 6.310460689 | 0.7954975089 | 5.597237678 | 0 | 0 |
| 86 | 1 | 1 | 1 | 30.62688555 | 0 | 4.140143784 | 0.2110616307 | 1.584922011 | 7.261952505 | 0 | 0 |
| 68 | 0 | 3 | 2 | 38.38762186 | 1 | 0.6460472705 | 9.25769491 | 5.897387927 | 5.477685594 | 0 | 0 |
| 75 | 0 | 0 | 1 | 18.77600941 | 0 | 13.72382571 | 4.649450668 | 8.341903192 | 4.213209925 | 0 | 0 |
| 72 | 1 | 1 | 0 | 27.83318838 | 0 | 12.16784763 | 1.531359788 | 6.736882044 | 5.748223869 | 0 | 0 |
| 87 | 0 | 0 | 0 | 35.45630173 | 1 | 16.02868824 | 6.440772687 | 8.086019121 | 7.551773444 | 0 | 1 |
| 89 | 0 | 3 | 1 | 39.46303422 | 0 | 9.811292129 | 8.819950351 | 0.4340202778 | 7.6440973 | 0 | 0 |
| 78 | 0 | 0 | 2 | 22.46338265 | 1 | 19.30018298 | 3.834639382 | 8.279189504 | 8.312325537 | 0 | 0 |
| 84 | 1 | 0 | 1 | 26.770946 | 0 | 10.97802164 | 3.978078872 | 7.024417381 | 8.253004551 | 1 | 0 |
| 78 | 1 | 0 | 1 | 28.87065239 | 1 | 10.1947063 | 0.6312807271 | 1.653281417 | 7.333235623 | 1 | 0 |
| 64 | 1 | 0 | 2 | 27.94286273 | 0 | 2.17577965 | 9.714565829 | 5.317231744 | 9.087141195 | 0 | 1 |
| 69 | 0 | 0 | 1 | 18.04591747 | 0 | 8.116831616 | 2.956494729 | 7.570632794 | 6.736796642 | 0 | 0 |
| 63 | 1 | 1 | 2 | 22.82289624 | 1 | 4.433961006 | 7.182894555 | 7.929485772 | 4.65482804 | 0 | 1 |
| 65 | 1 | 0 | 1 | 16.33328275 | 1 | 4.161794911 | 1.306320412 | 2.888935528 | 5.436422671 | 0 | 0 |
| 72 | 0 | 0 | 2 | 37.93246903 | 0 | 9.385602765 | 7.127939893 | 3.314982664 | 6.790196088 | 0 | 0 |
| 68 | 0 | 0 | 3 | 20.04140036 | 0 | 18.42636447 | 4.060713917 | 3.361536133 | 7.393126211 | 0 | 0 |
| 82 | 1 | 0 | 0 | 36.2230988 | 0 | 4.19289551 | 6.381502407 | 7.971127069 | 9.521026999 | 0 | 1 |
| 65 | 0 | 1 | 2 | 37.54394317 | 1 | 12.06395931 | 9.126038218 | 3.531206847 | 9.574004795 | 0 | 0 |

Gambar 3. 2 *Data Sample* (*Sebagian data terpotong*)

Dalam menganalisis data yang berkaitan dengan *alzheimer disease (AD)*, dilakukan melalui beberapa tahapan sebagai berikut:

a) Penerapan *Exploratory Data Analysis* (EDA)

1. Penanganan *Missing Values*

   Tahap ini bertujuan untuk menangani data yang hilang atau bernilai *null* serta menghapus data duplikat guna meningkatkan kualitas dan keakuratan data. Penanganan *missing values* dilakukan agar model yang digunakan memiliki performa yang lebih optimal serta dapat meningkatkan akurasi hasil analisis.

2. Penanganan *Outlier*

   Pada tahap ini, dilakukan identifikasi dan penanganan *outlier* karena keberadaannya dapat memengaruhi hasil analisis dan pemodelan. Langkah

ini bertujuan untuk meningkatkan akurasi model serta mencegah *overfitting*. Identifikasi *outlier* dilakukan menggunakan metode visualisasi seperti *box plot, scatter plot*, dan *histogram* untuk memastikan distribusi data yang lebih representatif.

b) *Corelations Analysis*

1. *Univariate Analysis*

   *Univariate Analysis* digunakan untuk mengeksplorasi data serta memahami karakteristik masing-masing individu dari satu variabel.

2. *Multivariate Analysis*

   *Multivariate Analysis* digunakan untuk mengeksplorasi hubungan antar variabel numerik yang memiliki korelasi. Korelasi yang ditemukan dapat membantu dalam mengidentifikasi fitur yang memiliki peran penting dalam membangun model.

*c) Feature Selection*

Perhitungan PCor salah satu metode *feature selection filter based* digunakan secara umum untuk mengidentifikasi kebergantungan di antara dua atau lebih variabel acak. Perhitungan PCor antara atribut independen (fitur) dan atribut dependennya (kelas) menghasilkan pengukuran kesamaan (*similarity*), yang mengevaluasi penting atau tidaknya suatu atribut. Pada bagian ini, perhitungan metode korelasi pearson akan digunakan untuk mengeleminasi fitur yang tidak penting.

### 3.4 *Data Preparation*

Pada tahapan ini akan dilakukan tahap persiapan data, yaitu :

a. *Encoding* Fitur kategori

   *Encoding feature* kategori diterapkan untuk mengubah nilai ke dalam bentuk data nominal sehingga dapat diproses oleh model.

d) *Splitting* Data

   Pembagian data menjadi data latih (*train*) dan data uji (*test*) bertujuan untuk mengevaluasi performa model secara objektif. Pada proses ini menggunakan fungsi ‘train_test_split’ untuk membagi set pelatihan dan

pengujian. Proporsi pembagian data menjadi fitur (x) dan target (y) dengan membagi data menjadi data latih (*train*) (80%) dan data uji (*test*) (20%).

e) Normalisasi Data

Normalisasi data diterapkan untuk memastikan semua fitur memiliki skala yang sama, sehingga dapat membantu algoritma pembelajaran mesin bekerja lebih optimal, dengan menggunakan standarisasi data 0-1. Normalisasi data bertujuan untuk memudahkan dan memastikan model lebih cepat mencapai konvergen dan mengurangi *cost computation model.*

### 3.5 *Modelling*

Tahap *modelling* berperan dalam merancang model *machine learning* yang dapat memprediksi data masukan menjadi sebuah hasil prediksi. Model dikembangkan menggunakan algoritma *random forest* dengan kombinasi *feature selection*. Algoritma *random forest* dikenal andal dalam menangani dataset yang berukuran besar dan kompleks. selain itu, dapat mengatasi masalah *overfitting* yang sering terjadi pada pohon hutan keputusan tunggal serta mampu mempertahankan kinerja yang stabil dan optimal.

### 3.6 *Evaluation*

Penilaian kinerja model diterapkan untuk memastikan model valid dengan mengukur performa model menggunakan *confusion matrix*, tahap ini bertujuan untuk mengidentifikasi keunggulan dan kelemahan model dalam memprediksi kelas tertentu. Kinerja model dievaluasi menggunakan metrik seperti akurasi, *precision, recall* dan F1-*score*.

### 3.7 *Deployment*

Pada tahap ini dilakukan pengimplementasian dengan diterapkan model yang akan diuji dengan data masukan yang baru untuk mengukur kinerja secara langsung. Model yang telah dievaluasi akan diuji coba untuk memprediksi data baru melalui *google colaboratory* dan *platform streamlit.*