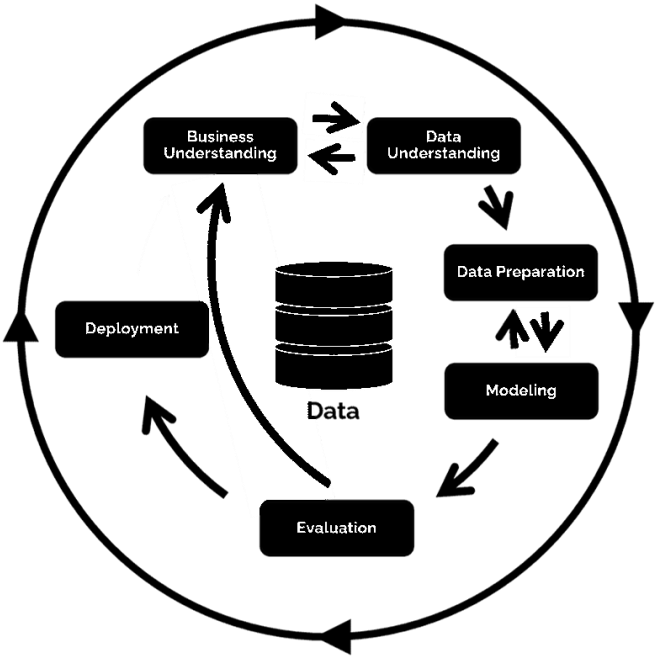
BAB III
METODOLOGI PENELITIAN

3.1 Alur Penelitian

Metode penelitian yang digunakan pada penelitian ini menggunakan algoritma *Random Forest.* Adapun tahapan penelitian yang dilakukan untuk mengembangkan model *Machine Learning* menggunakan model *Cross Industry Standard Process For Data Minning*, tahapannya sebagai berikut [27].



Gambar 3. 1 *CRISP for Data Minning*

3.2 *Business Understanding*

a) *Problem Statement*

Berdasarkan uraian latar belakang diatas, permasalahan yang dapat diselesaikan pada proyek ini ialah.

1) Berdasarkan fitur yang tersedia, fitur manakah yang paling berpengaruh terhadap prediksi risiko penyakit *stroke*?
2) Bagaimana model *Machine Learning* yang dapat memprediksi risiko penyakit *stroke* secara akurat dan tepat, sehingga dapat membantu tenaga medis dalam pendeteksian dini dan upaya pencegahan dimasa depan?

b) Goals

1) Mendapatkan analisa yang cukup terkait prediksi risiko penyakit *stroke*.
2) Memprediksi seseorang terkena penyakit *stroke* dengan nilai akurasi >85%.

c) Solutions Statement

Solusi yang dapat diterapkan untuk menyelesaikan permasalahan tersebut adalah:

1. Melakukan Analisa data terkait data penyakit *stroke* dengan menerapkan *Exploratory data analysis* (EDA) seperti teknik visualisasi. Adapun analisa yang dapat dilakukan yitu
   1) Melakukan *Pre-processing*.
   2) mengeksplor korelasi variabel atau data fitur terhadap variable target.
   3) menangani *outlier*.

2. Melakukan persiapan data, seperti
   1) *Encoding* fitur kategori
   2) Reduksi dimensi
   3) *Splitting* data
   4) Melakukan normalisasi data atau standarisasi data untuk dapat digunakan dalam model *Machine Learning*.

3.3 *Data Understanding*

Penelitian ini menggunakan *dataset* penyakit *stroke* yang diperoleh dari repositori *Kaggle*. *Dataset* yang digunakan berjumlah 4981 data dalam format *comma separated values* (csv). Pada data tersebut terdiri dari 11 *column feature* dan satu *column* target. Adapun *column feature* adalah jenis kelamin, usia, hipertensi, penyakit jantung, status menikah, jenis pekerjaan, jenis tempat tinggal, kadar rata-rata glukosa dalam darah, BMI, dan status merokok. Sedangkan, *column* target bernama *stroke*.

Dalam melakukan analisa data yang terkait dengan penyakit *stroke* dengan menerapkan :

a) *Exploratory data analysis* (EDA):
   1) Menangani *Missing Values*

      Pada tahapan ini menangani data yang hilang atau *null* dan nilai duplikat pada data, menangani *missing values* untuk meningkatkan akurasi kinerja pada model guna meningkatkan keakuratan pada data.
   2) Menangani *Outlier*

      Mengecek *outlier* untuk dapat mempengaruhi hasil analisis dan pemodelan serta meningkatkan akurasi model dan mencegah *overfitting*. Mengecek *outlier* menggunakan *box plot*, *scatter plot*, dan *histogram*.

b) *Corelations Analysis*
   1) *Univariate Analysis*

      Univariate Analysis bertujuan untuk mengeksplorasi data dan memahami karakteristik individu dari satu variabel.

   2) *Multivariate Analysis*

      Multivariate Analysis bertujuan untuk mengeksplorasi hubungan antar variabel numerik yang memiliki korelasi. Korelasi yang

ditemukan akan membantu dalam mengetahui *feature* terpenting dalam membangun *model*.

3.4 Data *Preparation*

Pada tahapan ini akan dilakukan tahapan persiapan data, yaitu :

a) *Encoding* Fitur kategori

Encoding feature kategori bertujuan untuk mengubah nilai ke dalam bentuk data nominal agar dapat diproses oleh model.

b) Reduksi dimensi dengan PCA

Principal Component Analysis (PCA) digunakan untuk mengurangi dimensi data, yang membantu dalam mengurangi kompleksitas model dan meningkatkan efisiensi komputasi. Pada analisis ini akan menentukan komponen utama yang diinginkan serta menampilkan fitur numerik saja. Fitur yang saling berkorelasi akan diagregasi dengan metode PCA.

c) *Splitting* Data

Pembagian data menjadi data latih (*train*) dan data uji (*test*) berfungsi untuk mengevaluasi performa model secara objektif. Menggunakan fungsi 'train_test_split' untuk membagi set pelatihan dan pengujiannya. Proporsi pembagian data menjadi fitur (x) dan target (y) dengan membagi data menjadi data latih (*train*) (80%) dan data uji (*test*) (20%).

d) Normalisasi Data

Normalisasi data dilakukan untuk memastikan bahwa semua fitur memiliki skala yang sama, yang dapat membantu algoritma pembelajaran mesin bekerja lebih efisien, dengan standarisasi data 0-1. Normalisasi data bertujuan untuk mempermudah dan mempercepat model mencapai konvergen dan mengurangi *cost computation* model.

3.5 *Modeling*

Tahapan *modelling* berperan dalam perancangan sebuah model *Machine Learning* yang dapat memprediksi data masukan menjadi sebuah hasil prediksi. Pembangunan model dengan menggunakan algoritma *Random Forest* dikenal dengan kehandalannya dalam menangani dataset yang besar dan kompleks serta mengatasi masalah overfitting yang dapat terjadi pada pohon hutan keputusan tunggal dan dapat menjaga stabilitas kinerja yang tinggi dan baik.

3.6 *Evaluation*

Melakukan penilaian kinerja model dan memastikan model valid dengan mengukur performa model dengan menggunakan *confusion* matrix, tahapan ini berfungsi untuk mengidentifikasi kekuatan dan kelemahan model dalam memprediksi kelas tertentu. Kinerja model dievaluasi menggunakan metrik seperti akurasi, *precision*, *recall* dan F1-*score*.

3.7 *Deployment*

Model akan diuji langsung dengan data masukan yang baru. Model yang telah dievaluasi akan diuji coba untuk memprediksi data baru langsung melalui *google colaboratory* dan platform streamlit.