# BAB II LANDASAN TEORI

## A. Tinjauan Pustaka

1. Penelitian Terkait

Tahun 2017 dilakukan penelitian dengan menggunakan metode KNN dan NBC dalam memprediksi ketepatan waktu lulus mahasiswa. Jurnal yang diterbitkan oleh Jurnal Komputer Terapan dan mengangkat judul "Prediksi Ketepatan Waktu Lulus Mahasiswa dengan *k-Nearest Neighbor* dan *Naïve Bayes Classifier* (Studi Kasus Prodi D3 Sistem Informasi Universitas Airlangga)". Hasil uji coba menunjukkan bahwa metode *k-Nearest Neighbor* menghasilkan akurasi lebih tinggi dibandingkan dengan *Naïve Bayes Classifier*. Akurasi tertinggi diperoleh dengan menggunakan metode *k-Nearest Neighbor* yaitu sebesar 98.7%[8].

Tahun selanjutnya dilakukan penelitian ditahun 2018 dengan judul "Perbandingan Klasifikasi Antara Knn Dan Naive Bayes Pada Penentuan Status Gunung Berapi Dengan *K-Fold Cross Validation*". Didalam penelitian tersebut diperoleh rata-rata akurasi sistem ketika menggunakan KNN sebesar 63,68 % dan standar deviasi 7,47. Sedangkan ketika diterapkan *Naive Bayes Classifier* dihasilkan rata-rata akurasi sistem sebesar 79,71 % dan standar deviasi 3,55%. Dengan demikian ketika diterapkan dengan *Naive Bayes Classifier* akurasi sistem dalam melakukan klasifikasi lebih baik dibandingkan dengan KNN[9].

Penelitian terbaru ditahun 2021 dilakukan klasifikasi nanas layak dijual dengan metode NBC dan KNN. Penelitian ini menyimpulkan bahwa metode NBC memiliki akurasi 73,3%, precision 73,3%, *recall* 73,3% dan AUC sebesar 0,704. Metode KNN menghasilkan akurasi 53,3%, precision 48,3%, *recall* 53,3% dan AUC sebesar 0,611[10].

Berdasarkan penelitian terkait dapat dilihat bahwa metode NBC dan metode KNN masih digunakan diberbagai bidang penelitian, hingga dalam penelitian ini dilakukan implementasi metode KNN dan NBC dalam melakukan klasifikasi pada dataset penyakit stroke. Perbedaan yang mendasar dalam penelitian ini selain melakukan klasifikasi dan membandingkan nilai akurasi, presisi dan *recall*, penelitian juga akan melakukan perbandingan dari sisi jumlah data testing yang digunakan dari masing-masing metode. Hingga dengan data testing yang berbeda-beda

dapat diketahui ada atau tidaknya pengaruh jumlah data testing yang digunakan dalam melakukan klasifikasi.

2. **Landasan Teori**

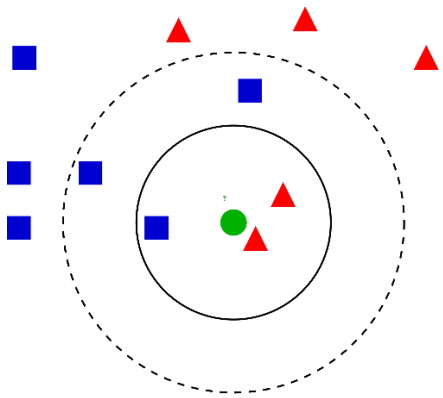
Dalam penelitian ini juga terdapat beberapa landasan-landasan teori yang digunakan serta dijadikan sebagai acuan dalam penelitian ini antara lain:

a. *Data Mining*

*Data mining* adalah proses mencari pola atau informasi menarik dalam data terpilih dengan menggunakan teknik atau metode tertentu. Teknik-teknik, metode-metode, atau algoritma dalam *Data mining* sangat bervariasi. Pemilihan metode atau algoritma yang tepat sangat bergantung pada tujuan dan proses *Knowledge Discovery in Database (KDD)* secara keseluruhan.

*Data mining* secara umum dibagi menjadi beberapa kelompok berdasarkan tugas yang dapat dilakukan, yaitu Klasifikasi *(Classification)*, Pengklusteran *(Clustering)*, Asosiasi *(Association)*, Prediksi *(Prediction)* [11].

b. *K-Nearest Neighbor* (KNN)



Gambar 1. *K-Nearest Neighbor* (KNN)

KNN dilakukan dengan mencari k tetangga dalam data training yang paling dekat dengan data testing dengan mengukur jarak antara data training dan data testing. Dari k tetangga terdekat yang terpilih dilakukan voting dengan memilih kelas yang jumlahnya paling banyak sebagai label kelas hasil klasifikasi pada data testing.

Ada banyak cara untuk mengukur jarak kedekatan antara data testing dengan *data training*, diantaranya *euclidean distance dan manhattan distance (city block distance)*, namun yang paling sering digunakan adalah euclidean distance [12]. Persamaan euclidean distance ditunjukkan pada persamaan 1.

$$euc = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \cdots + (a_n - b_n)^2} \qquad (1)$$

Keterangan:

a : *data training* (data lama)

b : *data testing* (data baru)

n : nilai atribut

Langkah-langkah untuk menghitung metode *K-Nearest Neighbor* antara lain[13]:

1) Menentukan parameter *K* (jumlah tetangga paling dekat).
2) Menghitung kuadrat jarak Euclid (queri instance) masing-masing objek terhadap data sampel yang diberikan menggunakan persamaan 1.
3) Kemudian mengurutkan objek-objek tersebut ke dalam kelompok yang mempunyai jarak Euclid terkecil.
4) Mengumpulkan kategori *Y* (Klasifikasi Nearest Neighbor)
5) Dengan menggunakan kategori Nearest Neighbor yang paling mayoritas maka dapat diprediksi nilai query instance yang telah dihitung

Dalam menentukan nilai K sebaiknya dilihat dari jumlah klasifikasi bila jumlahnya genap maka sebaiknya menggunakan nilai K yang ganjil, dan sebaliknya jika jumlah klasifikasi jumlahnya ganjil maka sebaiknya dalam menggunakan nilai K yang genap, karena jika tidak begitu maka sistem kemungkinan tidak akan mendapatkan jawaban[14].

c. *Naive Bayes Classifier* (NBC)

NBC merupakan salah satu algoritma dalam teknik *Data mining* yang menerapkan teori Bayes dalam klasifikasi. Teorema keputusan bayes adalah pendekatan statistik yang fundamental dalam pengenalan pola *pattern recoginition* [6].

Metode NBC dapat menangani data berupa numerik dengan menggunkan asumsi distribusi normal. Persamaan distribusi normal ditunjukkan pada persamaan 2.

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (2)$$

Persamaan 3 merupakan persamaan untuk mendapatkan nilai rata-rata (*mean*) dan persamaan 4 untuk mendapatkan standar deviasi.

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i. \quad (3)$$

$$\sigma = \left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu)^2\right]^{0.5} \quad (4)$$

Keterangan:
μ : nilai rata-rata (*mean*)
σ : standar deviasi
x : sampel data
n : jumlah sampel

Naïve Bayesian Classifier mengasumsikan bahwa keberadaan sebah atribut (variabel) tidak ada kaitannya dengan beradaan atribut (variabel) yang lain karena asumsi atribut tidak saling terkait (conditionally independent). NBC merupakan salah satu algoritma klasifikasi yang sederhana namun memiliki kemampuan dan akurasi tinggi[15].

Keuntungan penggunaan Naive Bayes adalah bahwa metode ini hanya membutuhkan jumlah data pelatihan (Training Data) yang kecil untuk menentukan estimasi paremeter yang diperlukan dalam proses pengklasifikasian. Naive Bayes sering bekerja jauh lebih baik dalam kebanyakan situasi dunia nyata yang kompleks dari pada yang diharapkan.

d. *Confusion matrix*

Tabel 1. *Confusion matrix*

| | | **True Class** | |
|---|---|---|---|
| | | **Positive** | **Negative** |
| **Predicted Class** | **Positif** | True Positives (TP) | False Negative (FP) |
| | **Negatif** | False Positives (FN) | True Negatives (TN) |

*Confusion matrix* adalah sebuah metode dalam bentuk table yang biasa digunakan untuk perhitungan akurasi, presisi, *recall*, dan *f-measure*.

Nilai TP (*True Positive*) dan TN (*True Negative*) menunjukan tingkat ketepatan klasifikasi. Umumnya semakin tinggi nilai TP dan TN semakin baik pula tingkat klasifikasi dari akurasi, presisi, dan *recall*. Jika label prediksi keluaran bernilai benar (*true*) dan nilai sebenarnya bernilai salah (*false*) disebut sebagai *False Positive* (FP). Sedangkan jika prediksi label keluaran bernilai salah (*false*) dan nilai sebenarnya bernilai benar (*true*) maka hal ini disebut sebagai *False Negative* (FN) [16].

e. Performa

Evaluasi klasifikasi didasarkan pengujian pada objek yang benar dan objek yang salah. validasi ini digunakan untuk menentukan jenis model yang terbaik dari hasil klasifikasi. Evaluasi ini menggunakan *confusion matrix*. Dari hasil *confusion matrix* akan ditentukan nilai akurasi, presisi, *recall*, dan f-measure [17]. Akurasi adalah ketetapan sistem dalam melakukan proses klasifikasi dengan benar. Persamaan akurasi ditunjukkan pada p ersamaan 5.

$$\text{Akurasi} = \frac{TP+TN}{TP+FP+TN+FN} \quad (5)$$

Presisi adalah rasio jumlah data yang relevan dengan total jumlah dokumen yang ditemukan pada sistem klasifikasi. Persamaan presisi ditunjukkan pada persamaan 6.

$$\text{Presisi} = \frac{TP}{TP+FP} \quad (6)$$

*Recall* adalah rasio jumlah data yang ditemukan kembali oleh sistem klasifikasi dengan total jumlah data yang relevan. Persamaan *recall* ditunjukkan pada persamaan 7.

$$Recall = \frac{TP}{TP+FN} \quad (7)$$

f. Bahasa Pemrograman Python

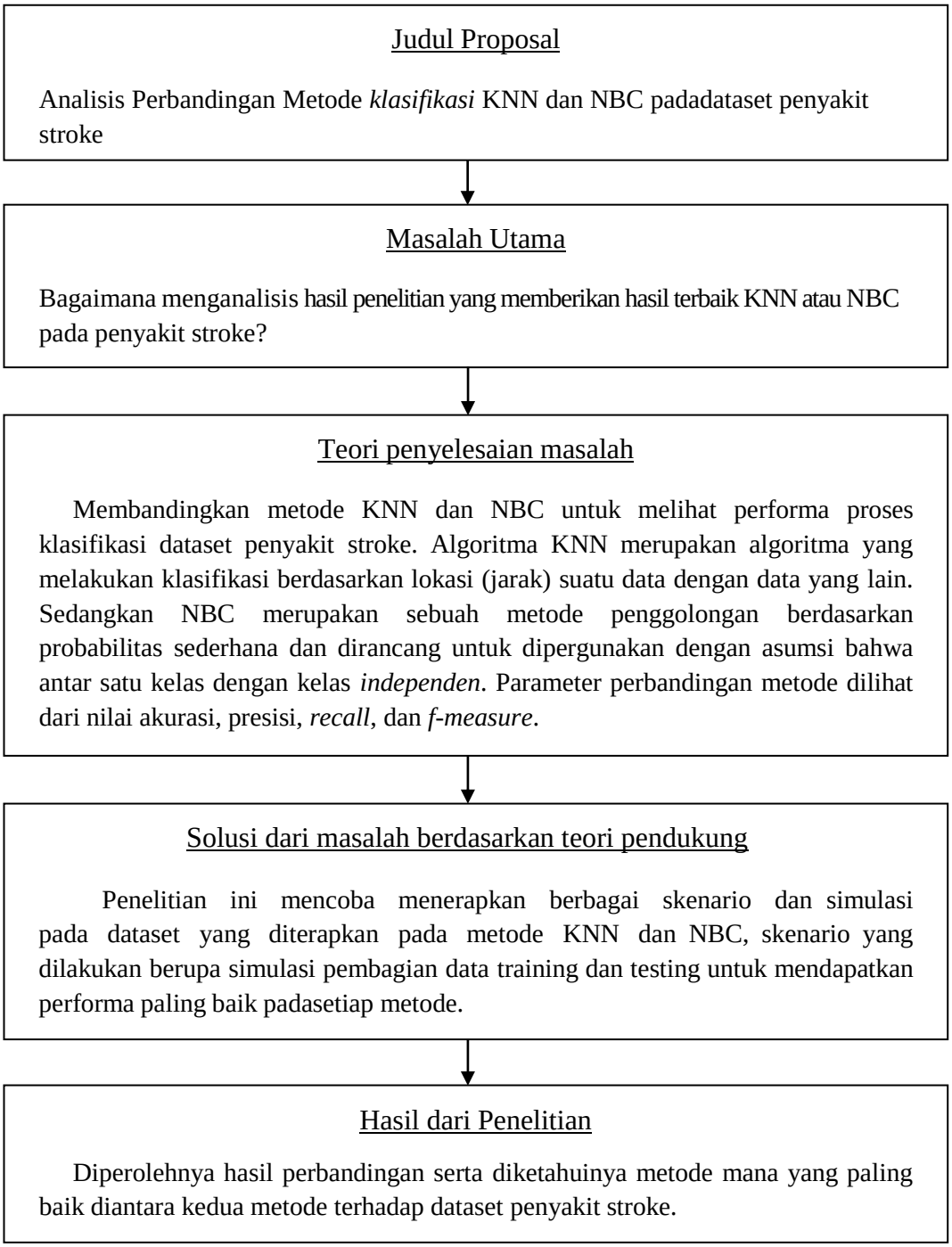
Python merupakan bahasa pemrograman yang berorientasi objek dinamis, dapat digunakan untuk bermacam-macam pengembangan perangkat lunak. Bahasa pemrograman Python dipilih karena bahasa ini mempunyai banyak keunggulan khususnya untuk pemrograman berbasis machine learning. Python memiliki banyak library yang dapat dengan mudah dipergunakan untuk machine learning, seperti numpy

untuk operasi vektor dan matrik, scikit-learn untuk data analisisi dan statistik, pandas data frame untuk pengolahan data, matplotlib untuk visualisasi data grafik, dan lain sebagainya [18].

g. *Scikit-Learn*

Scikit-learn adalah modul Python yang mengintegrasikan berbagai algoritma pembelajaran machine learning. *Scikit-learn* mencakup banyak algoritma yang digunakan untuk memproses data, seperti *K-Nearest Neighbor* (KNN), *Naive Bayes Classifier* (NBC), *Support Vector Machine* (SVM), Random Forest, dan lain-lain. *Scikit learn* adalah modul berbasis Python yang mengintegrasikan machine learning algoritma permasalahan supervised dan unsupervised. Modul *Scikit learn* sangat efisien untuk digunakan dalam *data mining* dan analisis data[19][20].

## B. Kerangka Pikir

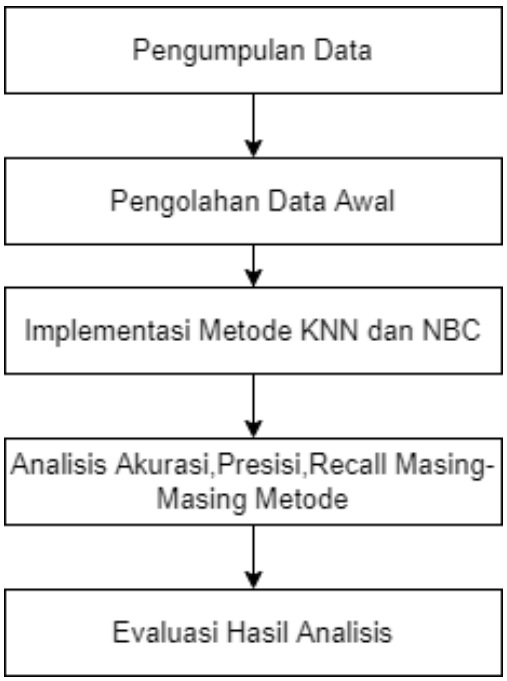



Gambar 2. Kerangka Pikir

# BAB III METODOLOGI PENELITIAN

## A. Tahapan Penelitian

Pada penelitian ini, data yang digunakan adalah data sekunder dari data kesehatan stroke. Data nilai tersebut akan diolah menggunakan KNN dan NBC. Berikut tahapan dalam melakukan penelitian perbandingan metode KNN dan NBC dalam melakukan klasifikasi.




Gambar 3. Tahapan Penelitian

Tahap awal dalam penelitian ini adalah melakukan pengumpulan data, data penelitian dikumpulkan dengan Instrumen yang valid dan reliabel. Data yang dikumpulkan adalah data sekunder yang diperoleh dari situs penyedia data yang valid. Selanjutnya melakukan pengolahan data awal yaitu melakukan analisis data penelitian yang relevan dengan tujuan penelitian. Disamping itu dalam pengolahan data awal dilakukan pengecekan dan analisis data hingga data yang dapat menimbukan error saat diimplementasikan kedalam metode akan dihilangkan.

Setelah data telah siap maka akan melangkah ke tahap selanjutnya yaitu menginputkan data masuk kedalam masing-masing metode yaitu KNN dan NBC. Tahap implementasi ini dilakukan pada platform Pyhton dengan membuat sistem yang mengacu pada perhitungan metode KNN dan NBC. Perhitungan sistem ini akan mengolah data berupa nilai yang sudah

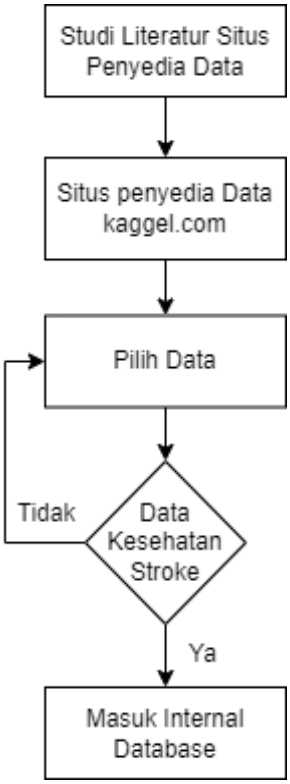
dimasukkan dan akan menghasilkan nilai-nilai yang sesuai dengan tahapan metode.

Tahapan selanjutnya adalah melakukan analisis hasil implementasi metode KNN dan NBC. Hasil akhir dari implementasi metode adalah nilai akurasi, presisi dan *recall* masing-masing metode, hasil tersebut dianalisis hingga menemukan hasil yang sesuai dengan tujuan penelitian. Langkah akhir dalam penelitian ini adalah melakukan evaluasi hasil dari masing-masing metode hingga menghasilkan hasil analisis yang baik dan akurat.

## B. Data Penelitian

Informasi mengenai kesehatan stroke akan diperoleh dari berbagai sumber baik data primer maupun sekunder. Data tersebut memiliki makna yang berguna untuk peningkatan pengetahuan dalam bidang kesehatan. Data kesehatan telah banyak disediakan oleh pihak penyedia data. Penelitian ini mengunakan data sekunder yaitu data didapat dari situs penyedia data yang resmi. Adapun hal yang berkaitan di dalamnya adalah data kesehatan stroke.

Model data yang digunakan untuk melakukan pelatihan program sebagai data latih serta data uji diambil dengan mempertimbangkan kesesuaian data dengan karakteristik yang telah ditentukan. Karakteristik tersebut adalah berupa data kesehatan stroke yang memiliki beberapa variabel.




Gambar 4. Tahap Pengumpulan Data

Dalam proses pengumpulan data akan melakukan pemilihan data yang sesuai. Dalam proses pemilihan data dilakukan dengan melihat data tersebut adalah data kesehatan dan memiliki atribut. Pemilihan dilakukan pula untuk megatasi *missing value* dalam data. Pemilihan dilakukan dengan menghapus data yang mengandung *missing value*.

## C. Metode Penelitian

1. Waktu dan Lokasi

   Waktu penelitian pada bulan Oktober 2021 hingga dengan bulan Februari 2022. Penelitian ini dilakukan di Fakultas Ilmu Komputer Universitas Muslim Indonesia, Makassar.

2. Bahan dan Alat

   Instrumen penelitian adalah alat bantu yang dipilih dalam kegiatannya agar sistematis dan mempermudah peneliti selama melakukan penelitian. Instrumen ini terbagi menjadi dua yaitu:

   a. Perangkat Keras (*Hardware*)

      Perangkat keras yang digunakan adalah:

      1). Laptop

      2). Processor Intel® Core ™ i7.

      3). RAM 6,00 GB.

   b. Perangkat Lunak (*Software*)

      Perangkat lunak atau software yang digunakan adalah:

      1). Microsoft Windows 10 Professional 64-bit, sebagai sistem operasi.

      2). Python sebagai bahasa pemrograman.

      3). Browser

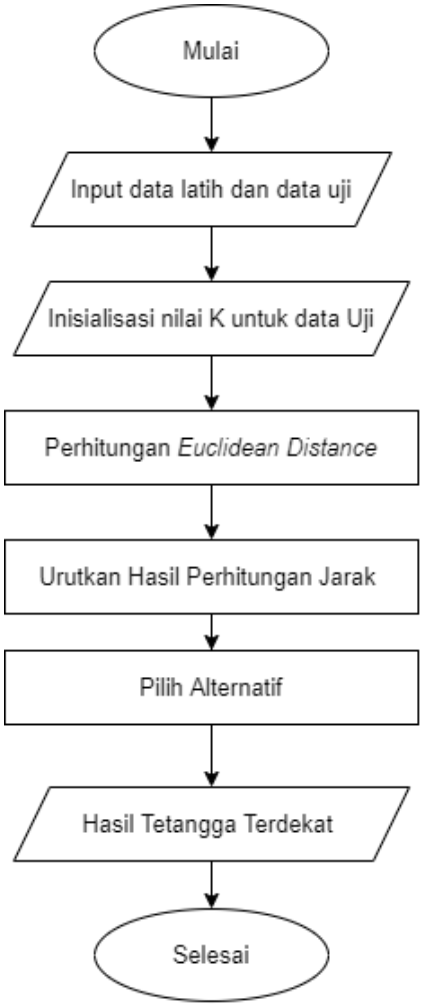
      4). Scikit-learn sebagai library machine learning.

      5). Repository Kaggle sebagai tempat penyimpanan data.

      6). Google Colab sebagai tempat mengolah data

3. Metode *K-Nearest Neighbor* (KNN)

   *K-Nearest Neighbor* atau KNN merupakan salah satu metode pengenalan pola yang umum dan sering digunakan untuk proses klasifikasi sekelompok data karena tekniknya yang sederhana,.

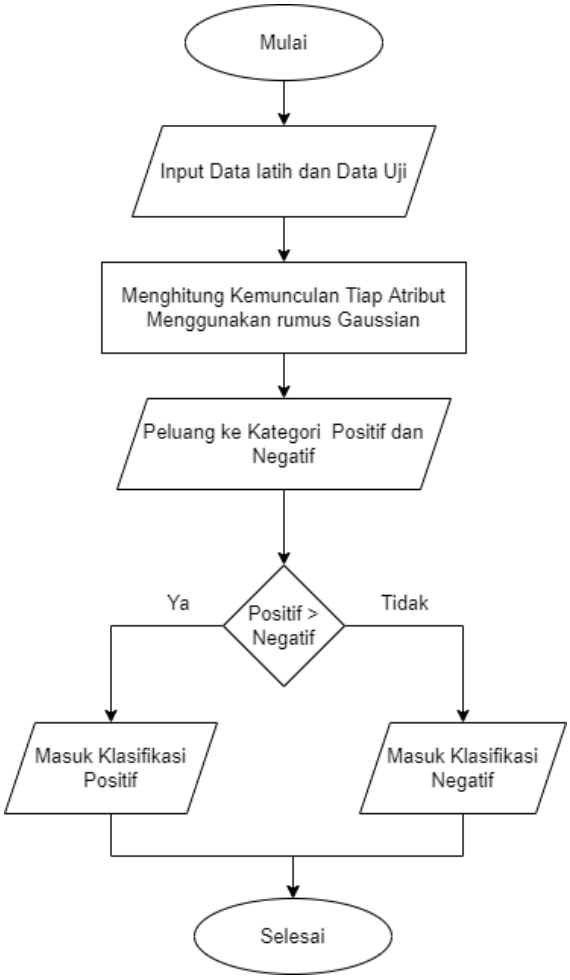



Gambar 5. *Flowchart* Metode KNN

Klasifikasi KNN mengkategorikan sebuah sampel data tidak berlabel dengan menggunakan label mayoritas dari sampel data tetangga yang paling terdekat (paling mirip) dalam data training Semakin mirip suatu dokumen maka semakin tinggi peluang untuk dikelompokan menjadi satu dokumen, sebaliknya semakin tidak mirip maka semakin rendah peluang untuk dikelompokkan menjadi satu dokumen.

4. Metode NBC

*Naïve Bayes Classifier* merupakan metode klasifikasi dengan probabilitas yang cepat dan sederhana, *Naïve Bayes Classifier* berasal dari teorema bayes dan hipotesis kebebasan yang dapat menghasilkan klasifier statistik berdasarkan peluang.




Gambar 6. *Flowchart* Metode *Naïve Bayes Classifier*

Metode *Naive Bayes Classifier* memiliki dua tahapan dalam proses klasifikasi teks yaitu tahap pelatihan dan tahap klasifikasi, pada tahap pelatihan dilakukan proses terhadap sampel data yang akan menjadi representasi data tersebut, selanjutnya adalah tahapan penentuan probabilitas prior untuk setiap kategori berdasarkan sampel data. Sedangkan pada tahapan klasifikasi ditentukan nilai kategori dari suatu data berdasarkan trem yang muncul.

## D. Evaluasi

Melakukan pengecekan terhadap setiap nilai atribut dan model yang sudah dibangun, kemudian melakukan evaluasi terhadap hasil dengan melakukan analisis pada metode KNN dan NBC. Tahap ini juga merupakan tahapan dimana dilakukan perbaikan kembali bila terjadi kekurangan pada tahapan ini bisa saja kembali lagi ke tahap yang pertama dan kemudian ke tahap berikutnya dengan tujuan perbaikan, sampai sesuai kebutuhan.

## E. Definisi Operasional Variabel

1. Variabel data penelitian

   Variabel dalam penelitian ini terdiri dari beberapa variabel dan salah satunya menjadi variabel target dalam melakukan klasifikasi. Adapun beberapa variabel tersebut terdiri dari *gender, age, hypertension, heart_disease, ever_married, work_type, residence_type, avg_glucose_level,* bmi, *smoking_status* dan variabel target berupa data dapat dikategorikan stroke atau tidak.

2. Variabel Analisis Metode

   Dalam penelitian ini akan dilakukan implementasi metode KNN dan NBC pada data yang sama. Adapun akurasi, presisi, dan *recall* sebagai variabel pengukur dalam melakukan analisis. Analisis metode akan dilakukan dengan memperhatikan jumlah data testing yang digunakan dalam melakukan klasifikasi. Percobaan jumlah data yang digunakan adalah data training 50% dan data testing 50%, selanjutnya data training 40%, hingga pengujian berakhir pada jumlah data testing 10%.