# BAB II
# LANDASAN TEORI

## 2.1 Pemilihan Jurusan di SMA

Penjurusan merupakan upaya untuk membantu siswa dalam memilih jenis sekolah atau program pengajaran khusus atau program studi yang akan diikuti oleh siswa dalam pendidikan lanjutannya. Tujuan dari penjurusan siswa adalah agar siswa dapat memperoleh informasi yang lengkap dan jelas tentang berbagai kemungkinan pilihan yang ada bagi kelanjutan pendidikannya. Sehingga dengan upaya tersebut peserta didik dapat memilih dengan tepat jenis sekolah atau program pengajaran khusus, atau program studi yang ada itu sesuai dengan kemampuan dasar umum (kecerdasan), bakat, minat, kecenderungan pribadi dan hal-hal yang dapat mempengaruhi kelanjutan pendidikannya itu.

Perbedaan individual antara siswa di sekolah di antaranya meliputi perbedaan kemampuan kognitif, motivasi berprestasi, minat dan kreativitas (Snow 1986). Lebih lanjut Snow mengemukakan bahwa oleh karena adanya perbedaan individu tersebut, maka fungsi pendidikan tidak hanya dalam proses belajar mengajar, tetapi juga meliputi bimbingan/konseling, pemilihan dan penempatan siswa sesuai dengan kapasitas individual yang dimiliki, rancangan sistem pengajaran yang sesuai dan strategi mengajar yang disesuaikan dengan karakteristik individu siswa.

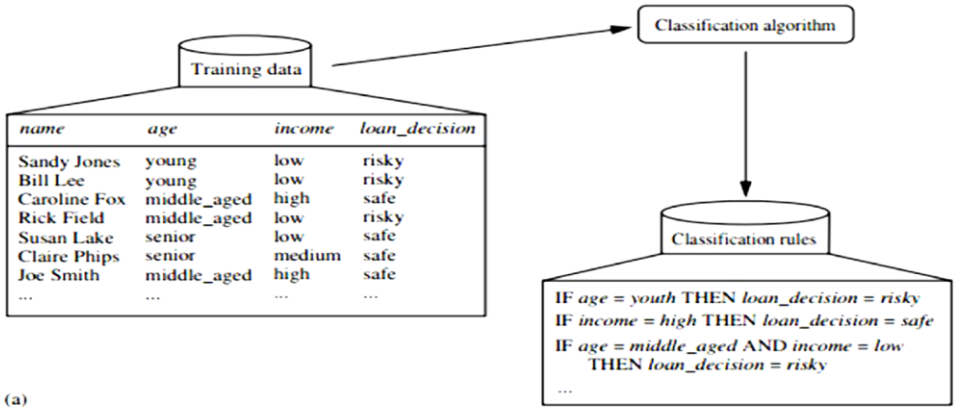
Oleh karena itu, sekolah memegang peranan penting untuk dapat mengembangkan potensi diri yang dimiliki siswa. Kemungkinan yang akan terjadi jika siswa mengalami kesalahan dalam penjurusan adalah rendahnya prestasi belajar siswa atau dapat menyebabkan terjadinya kegamangan dalam aktualisasi diri. Tak jarang siswa tidak mengerti alasan pemilihan jurusan tersebut, hendak kemana setelah tamat sekolah dan apa cita-citanya.

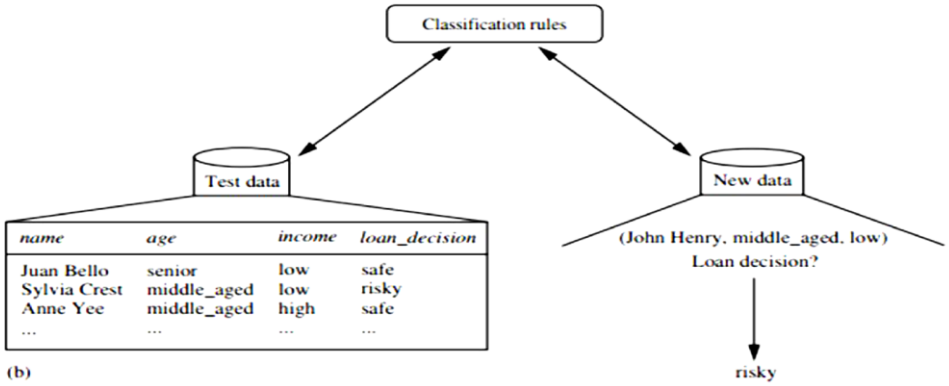
Psikolog UI, Indri Savitri, mengemukakan bahwa penjurusan siswa di sekolah menengah tidak saja ditentukan oleh kemampuan akademik tetapi juga harus didukung oleh faktor minat, karena karakteristik suatu ilmu menuntut karakteristik yang sama dari yang mempelajarinya. Dengan demikian, siswa

yang mempelajari suatu ilmu yang sesuai dengan karakteristik kepribadiannya (minat terhadap suatu ilmu tertentu) akan merasa senang ketika mempelajari ilmu tersebut [Gupta et.al. 2006].

## 2.2 Klasifikasi

Menurut Mike Chapple, klasifikasi adalah teknik data mining yang dilakukan untuk memprediksi kelas atau properti dari setiap *instance* data.




**Gambar 2.1** Tahapan Klasifikasi Data Mining

Tahapan dari klasifikasi dalam data mining terdiri dari :

a. Pembangunan Model

   Pada tahapan ini dibuat sebuah model untuk menyelesaikan masalah klasifikasi class atau attribut dalam data. Tahap ini merupakan fase pelatihan, dimana data latih dianalisis menggunakan algoritma klasifikasi, sehingga model pembelajaran direpresentasikan dalam bentuk aturan klasifikasi.

b. Penerapan Model

   Pada tahapan ini model yang sudah dibangun sebelumnya digunakan untuk menentukan attribut atau class dari sebuah data baru yang attribut atau *class*nya belum diketahui sebelumnya. Tahap ini digunakan untuk memperkirakan keakuratan aturan klasifikasi terhadap data uji. Jika model dapat diterima, maka aturan dapat diterapkan terhadap klasifikasi data baru.

## 2.3 Teorema *Bayes*

Ide dasar aturan *bayes* adalah hasil dari hipotesis atau peristiwa (H) dapat diperkirakan berdasarkan pada beberapa *evidence* (E) yang diamati.

Teori keputusan *bayes* adalah pendekatan statistik yang fundamental dalam pengenalan pola (*pattern recognition*). Pendekatan ini didasarkan pada kuantifikasi *trade-off* antara berbagai keputusan klasifikasi dengan menggunakan probabilitas dan ongkos yang ditimbulkan dalam keputusan-keputusan tersebut.

Hal penting dalam *bayes* adalah

a. Sebuah probabilitas awal atau priori H atau P(H), adalah probabilitas dari suatu hipotesis sebelum bukti diamati.
b. Sebuah probabilitas posterior H atau P(H|E), adalah probabilitas dari suatu hipotesis setelah bukti-bukti yang diamati ada.

$$P(H \mid E) = \frac{P(E \mid H) \times P(H)}{P(E)} \quad \text{.....(2.1)}$$

Keterangan :

P(H|E) : Probabilitas posterior bersyarat (*Conditional Probability*) suatu hipotesis H terjadi jika diberikan evidence/bukti E terjadi.

P(E|H) : Probabilitas sebuah evidence E terjadi akan mempengaruhi hipotesis H.

P(H) : Probabilitas awal (priori) hipotesis H terjadi tanpa memandang evidence apapun.

P(E) : Probabilitas awal (priori) *evidence* E terjadi tanpa memandang hipotesi/evidence yang lain.

Contoh Teorema Bayes adalah sebagai berikut :

Peramalan cuaca untuk memperkirakan terjadinya hujan, misal ada faktor yang mempengaruhi terjadinya hujan yaitu mendung. Jika diterapkan dalam *naive bayes* maka probabilitas terjadinya hujan jika bukti mendung sudah diamati :

$$P(Hujan \mid Mendung) = \frac{P(Mendung \mid Hujan) \times P(Hujan)}{P(Mendung)} \quad \text{.....(2.2)}$$

Keterangan :

P(Hujan|Mendung) : nilai probabilitas hipotesis hujan terjadi jika bukti mendung sudah diamati.

P(Mendung|Hujan) : probabilitas bahwa mendung yang diamati akan mempengaruhi terjadinya hujan.

P(Hujan) : probabilitas awal hujan tanpa memandang bukti apapun

P(Mendung) : probabilitas terjadinya mendung.

Teorema Bayes juga bisa menangani beberapa evidence, misalnya ada $E_1$, $E_2$, dan $E_3$, maka probabilitas posterior untuk hipotesis hujan:

$$P(H \mid E_1, E_2, E_3) = \frac{P(E_1, E_2, E_3 \mid H) \times P(H)}{P(E_1, E_2, E_3)} \quad \text{.....(2.3)}$$

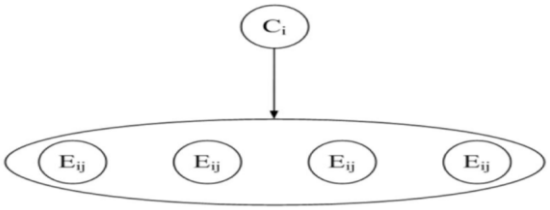
$$P(H \mid E_1, E_2, E_3) = \frac{P(E_1 \mid H) \times P(E_2 \mid H) \times P(E_3 \mid H) \times P(H)}{P(E_1, E_2, E_3)} \quad \text{.....(2.4)}$$

Bentuk diatas dapat diubah menjadi:

Untuk contoh diatas, jika ditambahkan *evidence* suhu udara dan angin

$$P(Hujan \mid Mendung, Suhu, Angin) = \frac{P(Mendung \mid Hujan) \times P(Suhu \mid Hujan) \times P(Angin \mid Hujan) \times P(Hujan)}{P(Mendung, Suhu, Angin)} \quad \text{.....(2.5)}$$

Pengklasifikasian menggunakan teorema *bayes* ini membutuhkan biaya komputasi yang mahal (waktu prosessor dan ukuran *memory* yang besar) karena kebutuhan untuk menghitung nilai probabilitas untuk tiap nilai dari perkalian kartesius untuk tiap nilai atribut dan tiap nilai kelas.




**Gambar 2.2** Ilustrasi Teorema Bayes

## 2.4 *Naive Bayes Classifier*

Klasifikasi *naive bayes* adalah metode yang berdasarkan probabilitas dan teorema *bayes* dengan asumsi bahwa setiap variabel bersifat bebas (*independence*) dan mengasumsikan bahwa keberadaan sebuah fitur tidak ada kaitannya dengan keberadaan fitur yang lain. Asumsi keindependenan atribut akan menghilangkan kebutuhan banyaknya jumlah data latih dari perkalian kartesius seluruh atribut yang dibutuhkan untuk mengklasifikasikan suatu data. Formulasi *naive bayes* untuk klasifikasi adalah

$$P(Y|X) = \frac{P(Y)\prod_{i=1}^{q} P(X_i|Y)}{P(X)} \quad \text{.....(2.6)}$$

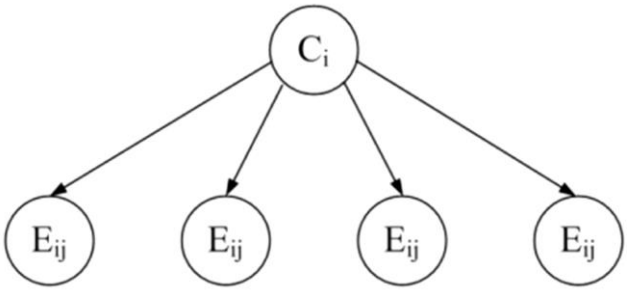
Keterangan :

P(Y|X) : Probabilitas data dengan vektor X pada kelas Y

P(Y) : Probabilitas awal kelas Y

$\prod_{i=1}^{q} P(X_i|Y)$ : Probabilitas independen kelas Y dari semua fitur dalam vektor X

Karena P (X) selalu tetap, sehingga dalam perhitungan prediksi nantinya cukup hanya dengan menghitung $P(Y)\prod_{i=1}^{q} P(X_i|Y)$.




**Gambar 2.3** Ilustrasi Naive Bayes.

Umumnya, *naive bayes* mudah dihitung untuk fitur bertipe kategoris seperti pada contoh diatas. Namun untuk tipe numerik (kontinu), ada perlakuan khusus sebelum dimasukkan dalam Naïve Bayes, yaitu :

1. Melakukan diskretisasi pada setiap fitur kontinu dan mengganti nilai fitur kontinu tersebut dengan nilai interval diskret. Pendekatan ini dilakukan dengan mentransformasi fitur kontinu ke dalam fitur ordinal.
2. Mengasumsikan bentuk tertentu dari distribusi probabilitas untuk fitur kontinu dan memperkirakan parameter distribusi dengan data pelatihan. Distribusi *Gaussian* biasanya dipilih untuk mempresentasikan probabilitas bersyarat dari fitur kontinu pada sebuah kelas P(Xi|Y). Untuk setiap kelas yj, probabilitas bersyarat kelas yj untuk fitur Xi adalah

$$P(X_i = x_i \mid Y = y_j) = \frac{1}{\sqrt{2\pi}\sigma_{ij}} \exp^{-\frac{(x_i-\mu_{ij})^2}{2\sigma_{ij}^2}} \quad \text{.....(2.7)}$$

Keterangan :

$\mu_{ij}$ : mean sampel Xi ( $\bar{x}$ ) dari semua data latih.

$2\sigma_{ij}^2$ : varian sampel ($s^2$) dari data latih.

$\sqrt{2\sigma_{ij}^2}$ : standart deviasi sampel ($\sqrt{\sigma}$) dari data latih.

Untuk mengukur nilai akurasi yang didapat dari hasil pengujian, menggunakan rumus 2.8. Sedangkan untuk mengukur tingkat kesalahannya menggunakan rumus 2.9.

$$Akurasi = \frac{Jumlah\ data\ yang\ diprediksi\ secara\ benar}{Jumlah\ prediksi\ yang\ dilakukan} \quad (2.8)$$

$$Laju\ error = \frac{Jumlah\ data\ yang\ diprediksi\ secara\ salah}{Jumlah\ prediksi\ yang\ dilakukan} \quad (2.9)$$

Sensitivitas akan mengukur proporsi positif asli yang dikenali (diprediksi) secara benar sebagai positif asli. Rumus perhitungannya menggunakan rumus 2.10. Sedangkan spesifisitas akan mengukur proporsi negatif asli yang dikenali (diprediksi) secara benar sebagai negatif asli. Rumus perhitungannya menggunakan rumus 2.10.

Sensitivitas = $\frac{TP}{TP+FN}$ …… (2.10)

Keterangan:

TP : Hasil data penjurusan siswa dengan kelas ipa yang diklasifikasikan secara benar mempunyai kelas ipa

FN : Hasil data penjurusan siswa dengan kelas ipa yang diklasifikasikan secara salah mempunyai kelas ips.

Spesifisitas = $\frac{TN}{FP+TN}$ …… (2.11)

Keterangan:

TN : Hasil data penjurusan siswa dengan kelas ips yang diklasifikasikan secara salah mempunyai kelas ipa.

FP : Hasil data penjurusan siswa dengan kelas ips yang diklasifikasikan secara benar mempunyai kelas ips.

### 2.4.1 Algoritma Klasifikasi *Naive Bayes*

Algoritma klasifikasi *naive bayes* dihitung sesuai dengan rumus *naive bayes* $P(Y)\prod_{i=1}^{q} P(X_i|Y)$, yang langkah-langkah perhitungannya dijelaskan sebagai berikut :

1. Menghitung nilai probabilitas kelas berdasarkan data latih $\rightarrow P(Y)$
2. Menghitung nilai probabilitas tiap fitur berdasarkan data latih$\rightarrow \prod_{i=1}^{q} P(X_i|Y)$

   Untuk fitur bertipe numerik menggunakan rumus berikut :

$$P(X_i = x_i \mid Y = y_j) = \frac{1}{\sqrt{2\pi}\sigma_{ij}} \exp^{-\frac{(x_i-\mu_{ij})^2}{2\sigma_{ij}^2}}$$

   Fitur numerik berikut ini dihitung tiap data uji.
3. Menghitung nilai probabilitas akhir

   Mengalikan hasil dari $P(Y)$ dan $\prod_{i=1}^{q} P(X_i|Y)$ pada masing-masing kelas dan data uji.
4. Data uji akan diklasifikasikan pada kelas dengan nilai probabilitas akhir terbesar.

Berikut flowchart perhitungan naïve bayes berdasarkan penjelasan diatas.


**Gambar 2.4** *Flowchart Naive Bayes*

### 2.4.2 Karateristik *Naive Bayes*

Karakteristik *naive bayes* bekerja berdasarkan teori probabilitas yang memandang semua fitur dari data sebagai bukti dalam probabilitas. Hal ini memberikan karakteristik *naive bayes* sebagai berikut :

1. Metode *naive bayes* teguh (*robust*) terhadap data-data yang terisolasi yang biasanya merupakan data dengan karakteristik berbeda (*outlier*). *naive bayes* juga bisa menangani nilai atribut yang salah dengan mengabaikan data latih selama proses pembangunan dan prediksi.
2. Tangguh menghadapi atribut yang tidak relevan.
3. Atribut yang mempunyai korelasi bisa mendegradasi kinerja klasifikasi *naive bayes* karena asumsi independensi tersebut sudah tidak ada.

*Naive Bayes* memiliki beberapa keuntungan dan kekurangan yaitu sebagai berikut :

1. Keuntungan *Naive Bayes*
   a. Cepat dan efisiensi ruang.
   b. Kokoh terhadap atribut yang tidak relevan.
   c. Hanya memerlukan sejumlah kecil data pelatihan untuk mengestimasi parameter (rata-rata dan variansi dari variabel) yang dibutuhkan untuk klasifikasi.
   d. Menangani kuantitatif dan data diskrit.
2. Kekurangan *Naive Bayes*
   a. Tidak berlaku jika *probabilitas* kondisionalnya adalah nol, apabila nol maka *probabilitas* prediksi akan bernilai nol juga.
   b. Mengasumsikan variabel bebas.

Contoh perhitungan *naive bayes* adalah sebagai berikut :

1. Jika ada sebuah data uji berupa hewan musang dengan nilai fitur: penutup kulit = rambut, melahirkan = ya, berat = 15, masuk kelas manakah untuk hewan musang tersebut ?

**Tabel 2.1** Contoh Data Latih Klasifikasi Hewan

| **Nama hewan** | **Penutup kulit** | **Melahirkan** | **Berat** | **Kelas** |
|---|---|---|---|---|
| Ular | Sisik | Ya | 10 | Reptil |
| Tikus | Bulu | Ya | 0.8 | Mamalia |
| Kambing | Rambut | Ya | 21 | Mamalia |
| Sapi | Rambut | Ya | 120 | Mamalia |
| Kadal | Sisik | Tidak | 0.4 | Reptil |
| Kucing | Rambut | Ya | 1.5 | Mamalia |
| Bekicot | Cangkang | Tidak | 0.3 | Reptil |
| Harimau | Rambut | Ya | 43 | Mamalia |
| Rusa | Rambut | Ya | 45 | Mamalia |
| Kura-Kura | Cangkang | Tidak | 7 | Reptil |

**Tabel 2.2** Contoh Perhitungan Nilai Probabilitas Tiap Fitur

| **Penutup kulit** | | **Melahirkan** | |
|---|---|---|---|
| Mamalia | Reptil | Mamalia | Reptil |
| Sisik = 0<br>Bulu = 1<br>Rambut = 5<br>Cangkang = 0 | Sisik = 2<br>Bulu = 0<br>Rambut = 0<br>Cangkang = 2 | Ya = 6<br>Tidak = 0 | Ya = 1<br>Tidak = 3 |

| | | | |
|---|---|---|---|
| P(Kulit = Sisik \| Mamalia) = 0<br>P(Kulit = Bulu\| Mamalia) = 1/6<br>P(Kulit = Rambut \| Mamalia) = 5/6<br>P(Kulit = Cangkang \| Mamalia) = 0 | P(Kulit = Sisik \| Reptil) = 0.5<br>P(Kulit = Bulu \| Reptil) = 0<br>P(Kulit = Rambut \| Reptil) = 0<br>P(Kulit = Cangkang \| Reptil) = 0.5 | P(Lahir = Ya \| Mamalia) = 1<br>P(Lahir = Tidak \| Mamalia) = 0 | P(Lahir = Ya \| Reptil) = 0.25<br>P(Lahir = Tidak \| Reptil) = 0.75 |

| **Berat** | | **Kelas** | |
|---|---|---|---|
| Mamalia | Reptil | Mamalia | Reptil |
| $\bar{x}_{mamalia} = 38.55$<br>$s^2_{mamalia} = 1960.255$<br>$s_{mamalia} = 44.275$ | $\bar{x}_{reptil} = 4.425$<br>$s^2_{reptil} = 23.6425$<br>$s_{reptil} = 4.8624$ | Mamalia = 6<br>P(Mamalia) = 6/10<br>= 0.6 | Reptil = 4<br>P(Reptil) = 4/10<br>= 0.4 |

Hitung nilai probabilitas untuk fitur dengan tipe numerik yaitu berat.

$$P(\text{Berat} = 15 \mid \text{Mamalia}) = \frac{1}{\sqrt{2\pi}\,44.275} \exp^{-\frac{(15-38.55)^2}{2\times 1960.255}} = 0.0078$$

$$P(\text{Berat} = 15 \mid \text{Reptil}) = \frac{1}{\sqrt{2\pi}\,4.8624} \exp^{-\frac{(15-4.425)^2}{2\times 23.6425}} = 0.0077$$

Hitung probabilitas akhir untuk setiap kelas:

P(X | Mamalia) = P(Kulit = Rambut | Mamalia) x P(Lahir = Ya | Mamalia) x P(Berat = 15 | Mamalia)

= 5/6 x 1 x 0.0078 = 0.0065

P(X | Reptil) = P(Kulit = Rambut | Reptil) x P(Lahir = Ya | Reptil) x P(Berat = 15 | Reptil)

= 0 x 0.25 x 0.0077 = 0

Nilai tersebut dimasukkan untuk mendapatkan probabilitas akhir:

P(Mamalia | X) = $\alpha$ x P(Mamalia) x P(X | Mamalia)

= $\alpha$ x 0.6 x 0.0065

= 0.0039$\alpha$

P(Reptil | X) = $\alpha$ x P(Reptil) x P(X | Reptil) = $\alpha$ x 0.4 x 0 = 0

Untuk $\alpha$ = 1/P(X) pasti nilainya konstan sehingga tidak perlu diketahui karena terbesar dari dua kelas tersebut tidak dapat dipengaruhi P(X).

Karena nilai probabilitas akhir (posterior probability) terbesar ada di kelas **mamalia (0.0039$\alpha$)**, maka data uji musang diprediksi sebagai kelas **mamalia**.

## 2.5 Penelitian Sebelumnya

*Naive bayes* merupakan metode popular yang banyak digunakan untuk klasifikasi. Beberapa riset yang telah dilakukan berkaitan dengan kasus klasifikasi yang menggunakan metode *naive bayes* dan penelitian tentang bantuan langsung sementara masyarakat, antara lain :

Penelitian oleh Gresika Duita pada tahun 2016 yang berjudul "*Klasifikasi Penerimaan Bantuan Langsung Sementara Masyarakat (BLSM) Dengan Metode Naive Bayes (Studi Kasus: Kelurahan Pekauman Kecamatan Gresik)"*. Adapun permasalahan yang diambil dalam penelitian ini yaitu Bagaimana mengklasifikasikan masyarakat yang berhak menerima Bantuan Langsung Sementara Masyarakat (BLSM) di Kelurahan Pekauman Kecamatan Gresik. Penelitian ini bertujuan untuk membuat sistem yang dapat mengetahui masyarakat yang berhak menerima Bantuan Langsung Sementara Masyarakat (BLSM) di Kelurahan Pekauman Kecamatan Gresik. Atribut yang digunakan sebagai data latih sistem adalah penghasilan, kondisi rumah, makan/hari, jumlah keluarga, tanggungan anak sekolah dan kelas. Dari 458 data dari Kelurahan Pekauman, data tersebut diambil 40% yang akan dijadikan sebagai data uji dan 60% akan menjadi data training. Jadi jumlah pembagiannya adalah 275 data sebagai data training dan 183 data untuk data uji. Keakurasian ketepatan hasil klasifikasi dengan *naive bayes* untuk mengetahui tingkat kecenderungan penyelesaian studi mahasiswa sistem memiliki akurasi kebenaran sistem sebesar 99.45% dengan nilai error 0.55%.

Sedangkan penelitian lain yang digunakan sebagai rujukan dalam tugas akhir ini adalah penelitian dilakukan oleh Friday Aries L pada tahun 2015

dengan judul "*Aplikasi Klasifikasi Penentuan Penerimaan Beras Miskin (Raskin) di Desa Sidomulyo Kecamatan Deket Kabupaten Lamongan Dengan Metode Klasifikasi Naive Bayes".* Adapun permasalahan yang diambil dalam penelitian ini yaitu bagaimana cara menentukan masyarakat yang berhak menerima beras miskin (Raskin) berdasarkan regulasi Pemerintah di Ds. Sidomulyo Kec. Deket Kab. Lamongan. Penelitian ini bertujuan untuk membuat sistem klasifikasi yang dapat menentukan pembagian beras miskin (Raskin) kepada masyarakat Ds. Sidomulyo Kec. Deket Kab. Lamongan dengan lebih tepat sasaran. Atribut yang digunakan adalah pekerjaan, penghasilan, harta benda (kendaraan), kondisi rumah, jumlah keluarga dan kelas. Penelitian ini menggunakan 60 data yang diperoleh dari desa sidomulyo kecamatan deket kabupaten lamongan dalam percobaan dengan pembagian 50 data latih dan 10 data uji. Tingkat akurasi prediksi 85 %, laju error 15%, sensitivitas 99,17% dan spesifisitas 70,59%.