

# PENERAPAN TEOREMA BAYES DALAM PENGKLASIFIKASIAN DATA NASABAH ASURANSI

**Puji S.M. Napitupulu (12S15003), Melani  
Tambun(12S15013), Bonggal B.  
Siahaan(12S15023), Astri D.  
Pangaribuan(12S15036), Anggi Frecelia  
(12S15048), Dimpu T.M. Hutasoit (12S15059)**

*Fakultas Teknik Informatika dan Elektro  
Program Studi Sistem Informasi  
Institut Teknologi Del*

**ABSTRAK** *Data mining adalah teknik yang memanfaatkan data dalam jumlah yang besar untuk memperoleh informasi berharga yang sebelumnya tidak diketahui dan dapat dimanfaatkan untuk pengambilan keputusan penting. Pada penelitian ini, penulis berusaha menambang data nasabah sebuah perusahaan asuransi untuk mengetahui lancar, kurang lancar, atau tidak lancarnya nasabah tersebut. Penelitian Naive Bayes bertujuan untuk melakukan klasifikasi data pada kelas tertentu. Pola tersebut dapat digunakan untuk memperkirakan nasabah yang bergabung, sehingga perusahaan bisa mengambil keputusan menerima atau menolak calon nasabah tersebut.*

**Keywords** : *data mining, asuransi, klasifikasi, algoritma Naive Bayes*

## 1. PENDAHULUAN

Premi merupakan pendapatan bagi perusahaan asuransi, yang jumlahnya ditentukan dalam suatu presentase atau tarif tertentu dari jumlah yang dipertanggungkan.

Permasalahan yang sering muncul dalam perusahaan asuransi adalah banyaknya nasabah yang menunggak dalam membayar premi, sehingga diperlukan suatu sistem yang dapat mengklasifikasikan nasabah mana yang masuk ke dalam kelompok lancar, kelompok kurang lancar, dan nasabah mana yang masuk ke dalam membayar iuran premi. Sehingga pihak asuransi dapat mengatasi permasalahan sejak dini.

Penggunaan teknik data mining akan diharapkan mampu memberikan informasi yang berguna tentang teknik klasifikasi data nasabah yang akan bergabung dalam kelompok lancar, kurang lancar, atau tidak lancar dalam membayar premi.

## 2. LANDASAN TEORI

### 2.1 Data Mining

Data mining adalah suatu metode dalam menemukan sebuah informasi baru dengan cara mencari pola atau aturan tertentu dari sejumlah data yang sangat besar. Data mining juga dapat kita gunakan dalam mencari suatu nilai tambah dalam bentuk pengetahuan yang mungkin selama ini tidak kita ketahui secara langsung dari suatu kumpulan data.

Knowledge Discovery adalah keseluruhan proses non-trivial untuk mencari dan mengidentifikasi pola (pattern) dalam data, dimana pola yang ditemukan bersifat sah, baru, dapat bermanfaat dan dapat dimengerti.

Tahap- tahap dari proses *Knowledge Discovery* dalam Basis Data adalah :

#### a. *Selection*

- 1.) Menciptakan himpunan data target , pemilihan himpunan data, atau memfokuskan pada subset variabel atau sampel data, dimana penemuan (discovery) akan dilakukan.
- 2.) Pemilihan (seleksi) data dari sekumpulan data operasional perlu dilakukan sebelum tahap penggalian informasi dalam KDD dimulai. Data hasil seleksi yang akan digunakan untuk proses data mining, disimpan dalam suatu berkas, terpisah dari basis data operasional.

#### b. *Pre-Processing / Cleaning*

- 1.) Pemrosesan pendahuluan dan pembersihan data merupakan operasi dasar seperti penghapusan noise dilakukan. Sebelum proses data mining dapat dilaksanakan, perlu dilakukan proses cleaning pada data yang menjadi fokus KDD.
- 2.) Proses cleaning mencakup antara lain membuang duplikasi data, memeriksa data yang inkonsisten, dan memperbaiki kesalahan pada data, seperti kesalahan cetak (tipografi). Dilakukan proses enrichment, yaitu proses “memperkaya” data yang sudah ada dengan data atau informasi lain yang relevan dan diperlukan untuk KDD, seperti data atau informasi eksternal.

### c. Transformation

Pencarian fitur-fitur yang berguna untuk mempresentasikan data bergantung kepada goal yang ingin dicapai.

Merupakan proses transformasi pada data yang telah dipilih, sehingga data tersebut sesuai untuk proses data mining. Proses ini merupakan proses kreatif dan sangat tergantung pada jenis atau pola informasi yang akan dicari dalam basis data

### d. Data mining

- 1.) Pemilihan tugas data mining; pemilihan goal dari proses KDD misalnya klasifikasi, regresi, clustering, dll.
- 2.) Pemilihan algoritma data mining untuk pencarian (searching)
- 3.) Proses Data mining yaitu proses mencari pola atau informasi menarik dalam data terpilih dengan menggunakan teknik atau metode tertentu. Teknik, metode, atau algoritma dalam data mining sangat bervariasi. Pemilihan metode atau algoritma yang tepat sangat bergantung pada tujuan dan proses KDD secara keseluruhan.

### e. Interpretation/ Evaluation

Penerjemahan pola-pola yang dihasilkan dari data mining.

Pola informasi yang dihasilkan dari proses data mining perlu ditampilkan dalam bentuk yang mudah dimengerti oleh pihak yang berkepentingan.

Tahap ini merupakan bagian dari proses KDD yang mencakup pemeriksaan apakah pola atau informasi yang ditemukan bertentangan dengan fakta atau hipotesa yang ada sebelumnya.

## 2.2 Metode Klasifikasi

Klasifikasi adalah proses untuk menemukan model atau fungsi untuk menjelaskan atau membedakan konsep atau kelas data, tujuannya untuk memperkirakan kelas dari suatu objek yang labelnya tidak diketahui. Dalam mencapai tujuan tersebut, proses klasifikasi membentuk suatu model yang mampu membedakan data ke dalam kelas-kelas yang berbeda berdasarkan aturan atau fungsi tertentu. Model tersebut bisa berupa aturan 'jika-maka' (implikasi), berupa pohon keputusan atau formula matematis.

## 2.3 Algoritma Naive Bayes

Algoritma *Naive Bayes* merupakan salah satu algoritma yang terdapat pada teknik klasifikasi. *Naive*

*Bayes* merupakan pengklasifikasian dengan metode probabilitas dan statistik yang dikemukakan oleh ilmuwan Inggris *Thomas Bayes*, yaitu memprediksi peluang di masa depan berdasarkan pengalaman dimasa sebelumnya sehingga dikenal sebagai *Teorema Bayes*. Teorema tersebut dikombinasikan dengan *Naive* dimana diasumsikan dengan kondisi antar atribut yang saling bebas. Klasifikasi *Naive Bayes* diasumsikan bahwa ada atau tidak ciri tertentu dari sebuah kelas tidak ada hubungannya dengan ciri dari kelas lainnya. Untuk menjelaskan teorema *Naive Bayes*, perlu diketahui bahwa proses klasifikasi memerlukan sejumlah petunjuk untuk menentukan kelas apa yang cocok bagi sampel yang dianalisis tersebut. Karena itu, teorema *Bayes* di atas disesuaikan sebagai berikut:

$$P(C|F_1 \dots F_n) = \frac{P(C)P(F_1 \dots F_n|C)}{P(F_1 \dots F_n)}$$

Dimana:

C = representasi kelas

F = merepresentasikan karakteristik petunjuk yang dibutuhkan untuk melakukan klasifikasi

Rumus tersebut menjelaskan bahwa peluang masuknya sampel karakteristik tertentu dalam kelas C (*Posterior*) adalah peluang munculnya kelas C (sebelum masuknya

sampel tersebut, seringkali disebut *prior*), dikali dengan peluang kemunculan karakteristik karakteristik sampel pada kelas C (disebut juga *likelihood*), dibagi dengan peluang kemunculan karakteristik karakteristik sampel secara global (disebut juga *evidence*). Karena itu, rumus diatas dapat pula ditulis secara sederhana sebagai berikut :

$$\text{Posterior} = \frac{\text{Prior} \times \text{likelihood}}{\text{evidence}}$$

Dari *posterior* tersebut nantinya akan dibandingkan dengan nilai nilai *posterior* kelas lainnya untuk menentukan ke kelas apa suatu sampel akan diklasifikasikan. Penjabaran lebih lanjut rumus *Bayes* tersebut dilakukan dengan menjabarkan menggunakan aturan perkalian sebagai berikut :

$$P(C|F_1, \dots, F_n) = P(C) P(F_1, \dots, F_n|C)$$

Dapat dilihat bahwa hasil penjabaran tersebut menyebabkan semakin banyak dan semakin kompleksnya faktor faktor syarat yang mempengaruhi nilai probabilitas, yang hampir mustahil untuk dianalisa satu persatu. Akibatnya, perhitungan tersebut menjadi sulit untuk dilakukan. Maka digunakan asumsi independensi yang sangat tinggi (*naif*), bahwa masing masing petunjuk saling bebas (*independen*) satu sama lain.

Dengan asumsi tersebut, maka berlaku suatu kesamaan sebagai berikut:

$$P(P_i|F_j) = \frac{P(F_i \cap F_j)}{P(F_j)} = \frac{P(F_i)P(F_j)}{P(F_j)} = P(F_i)$$

Dari persamaan diatas dapat disimpulkan bahwa asumsi independensi naif tersebut membuat syarat peluang menjadi sederhana, sehingga perhitungan menjadi mungkin untuk dilakukan. Selanjutnya, penjabaran

dapat disederhanakan menjadi :

$$P(C|F_1, \dots, F_n) = P(C)P(F_1|C)P(F_2|C)P(F_3|C) \dots$$

Persamaan diatas merupakan model dari teorema *Naive Bayes* yang selanjutnya akan digunakan dalam proses klasifikasi. Untuk klasifikasi dengan data kontinyu digunakan rumus *Densitas Gauss* :

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

Keterangan :

$P$  : Peluang

$X_i$  : Atribut ke  $i$

$x_i$  : Nilai atribut ke  $i$

$Y$  : Kelas yang dicari

$y_j$  : Sub kelas  $Y$  yang dicari

$\mu$  : Mean, menyatakan rata rata dari seluruh atribut

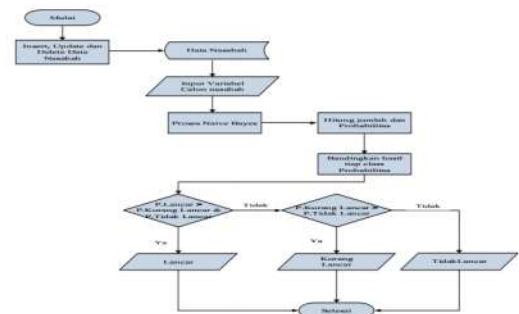
$\sigma$  : Deviasi standar, menyatakan varian dari seluruh atribut

Adapun alur dari metode *Naive Bayes* adalah sebagai berikut :

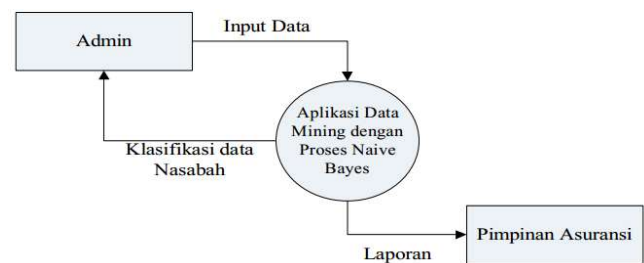
1. Baca data training
2. Hitung Jumlah dan probabilitas, namun apabila data numerik maka:
  - a. Cari nilai mean dan standar deviasi dari masing masing parameter yang merupakan data numerik.
  - b. Cari nilai probabilitas dengan cara menghitung jumlah data yang sesuai dari kategori yang sama dibagi dengan jumlah data pada kategori tersebut.
3. Mendapatkan nilai dalam tabel mean, standart deviasi dan probabilitas.

### 3. PERANCANGAN SISTEM

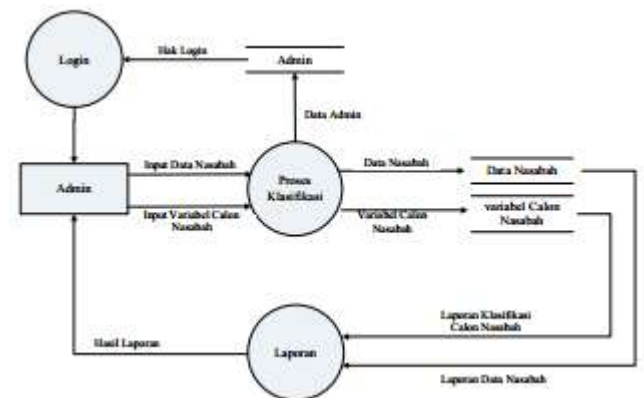
#### 3.1 Flowchart Sistem



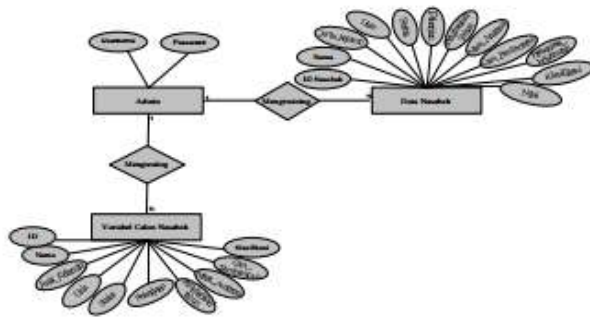
#### 3.2 Diagram Konteks



#### 3.3 Data Flow Diagram (DFD)



### 3.4 Entity Relationship Diagram (ERD)



## 4. PERANCANGAN BASIS DATA

### 4.1 Desain Tabel Admin

Perancangan tabel yang digunakan untuk menyimpan data admin.

Nama Field: Username (Nama User), Password (Password User).

### 4.2 Desain Tabel Data Nasabah

Perancangan tabel yang akan menyimpan data nasabah yang akan digunakan dalam system.

Nama Field: ID Nasabah, Nama, Jenis Kelamin, Usia, Status, Pekerjaan, Penghasilan/tahun, Masa asuransi, Cara pembayaran, Persentase kelancaran, Klasifikasi, Nilai.

### 4.3 Desain Tabel Variabel Calon Nasabah

Perancangan tabel yang akan menyimpan data calon nasabah yang akan digunakan dalam system.

Nama: ID (Calon Nasabah), Nama, Jenis Kelamin, Usia, Status, Pekerjaan, Penghasilan/tahun, Masa asuransi, Cara Pembayaran, Klasifikasi.

## 5. IMPLEMENTASI DENGAN PERHITUNGAN NAIVE BAYES

Metode Naïve Bayes merupakan suatu model probabilistik yang digunakan dalam membantu dalam mengambil keputusan. Algoritma Naïve Bayes merupakan algoritma teknik klasifikasi yang apabila atribut yang terdapat di dalamnya saling bebas (*independence*), maka nilai probabilitas sebagai berikut:

$$P(x_1, \dots, x_k | C) = P(x_1 | C) \times \dots \times P(x_k | C)$$

Tahap awal perhitungan dilakukan dengan pengambilan atau klasifikasi data training dan data

nasabah asuransi, dengan variable penentu yang digunakan adalah:

1. Jenis Kelamin

Terdiri dari: laki-laki dan perempuan.

2. Usia

Terdiri dari: 20-29 tahun, 30-40 tahun, dan 40 tahun keatas.

3. Status

Terdiri dari: kawin dan belum kawin.

4. Pekerjaan

Terdiri dari: PNS, Pegawai Swasta, dan Wiraswasta.

5. Penghasilan

Terdiri dari: 0-25 juta, 25-50 juta, dan 50 juta ke atas.

6. Cara pembayaran premi

Terdiri dari: bulanan, triwulan, semesteran, dan tahunan.

7. Masa pembayaran premi

Terdiri dari: 5-10 tahun, 11-15 tahun, dan lebih dari 15 tahun.

Berdasarkan data input di atas, apabila diberikan input baru, klasifikasi dapat dilakukan dengan cara:

- 1) Menghitung jumlah class / label

$$P(Y) = \frac{\text{Jumlah data}}{\text{Jumlah keseluruhan}}$$

Jumlah data lancar/kurang lancar/tidak lancar dibagi dengan jumlah keseluruhan data.

- 2) Menghitung jumlah kasus yang sama dengan class yang sama

$$P(A|Y) = \frac{\text{Jumlah data}}{\text{Jumlah keseluruhan}}$$

A= Jenis kelamin, Usia, Status, Pekerjaan, Penghasilan, Masa\_Asuransi, dan Cara Pembayaran.

Y= Lancar, Kurang Lancar, dan Tidak Lancar.

- 3) Kalikan semua variabel Lancar, Kurang Lancar, Tidak Lancar

$$P(A_1 | Lancar) \times P(A_2 | Lancar) \times \dots \times P(A_n | Lancar)$$

- a)

- b) 
$$\frac{P(A_1|Kurang Lancar) \times P(A_2|Kurang Lancar) \times \dots \times P(A_n|Kurang Lancar)}{P(A_1|Tidak Lancar) \times P(A_2|Tidak Lancar) \times \dots \times P(A_n|Tidak Lancar)}$$
- c) 4) Bandingkan hasil class Lancar, Kurang Lancar, Tidak Lancar

Dari data hasil perkalian, jika nilai probabilitas tinggi maka dapat disimpulkan status calon nasabah tersebut apakah lancar, kurang lancar, atau tidak lancar.

## 6. IMPLEMENTASI SISTEM

Setelah perancangan sistem selesai, database selanjutnya ialah implementasi sistem. Implementasi sistem adalah bagian akhir dari perancangan sistem yang dibangun yang sekaligus merupakan testing program.

### a. Form Login

Form login adalah form untuk masuk ke program yang ingin diakses dengan cara menginput pasangan *username* dan *password* yang berfungsi sebagai form keamanan. Jika hak akses telah diberikan oleh sistem, maka *user* dapat mengakses menu utama aplikasi.

### b. Form Menu Utama

Form ini berfungsi untuk mengakses segala perintah yang terdapat pada aplikasi dan dapat diakses setelah *user* telah *login* terlebih dahulu. Form ini memiliki beberapa menu yaitu Menu File Data yang berisi submenu data nasabah dan cek presentase kelancaran, menu admin, laporan, dan exit.

### c. Form Data Nasabah

Form ini berfungsi untuk mencari data nasabah, menambah, menghapus, dan menyimpan data nasabah. Data nasabah ini selanjutnya akan digunakan untuk data pelatihan pada proses klasifikasi.

### d. Form Cek Presentasi Kelancaran

Form ini adalah form data testing yang digunakan untuk mengecek tingkat kelancaran dari calon nasabah.

### e. Form Hasil Input Data Calon Nasabah

Form ini menampilkan hasil output dari penginputan data calon nasabah yang sebelumnya telah diproses dengan algoritma *Naive Bayes*. Proses klasifikasi sangat dipengaruhi oleh atribut-atribut terpilih yang mendukung penentuan kelas nasabah yang lancar, kurang lancar, maupun tidak lancar.

### f. Form Laporan Akhir

Form ini adalah output dari proses klasifikasi data yang menampilkan hasil akhir dari proses yang telah dilakukan yaitu *menginput* data calon nasabah dengan algoritma *Naive Bayes*.

## 7. KESIMPULAN

Berdasarkan hasil pembahasan maka dapat diambil beberapa kesimpulan antara lain :

- 1.) Sistem klasifikasi data nasabah ini digunakan untuk menampilkan informasi klasifikasi lancar, kurang lancar atau tidak lancarnya calon nasabah dalam membayar premi asuransi dengan menggunakan algoritma *Naive Bayes*. Algoritma *Naive Bayes* untuk Klasifikasi Nasabah Asuransi 145.
- 2.) Dengan adanya sistem ini maka mempermudah pihak asuransi dalam memperkirakan nasabah yang bergabung, sehingga perusahaan bisa mengambil keputusan untuk menerima atau menolak calon nasabah tersebut.
- 3.) Algoritma *Naive Bayes* di dukung oleh ilmu Probabilistik dan ilmu statistika khususnya dalam penggunaan data petunjuk untuk mendukung keputusan pengklasifikasian. Pada algoritma *Naive Bayes*, semua atribut akan memberikan kontribusinya dalam pengambilan keputusan, dengan bobot atribut

yang sama penting dan setiap atribut saling bebas satu sama lain.

4.) Variabel penentu yang digunakan dalam penelitian ini adalah jenis kelamin, usia, status, pekerjaan, penghasilan per tahun, masa pembayaran asuransi, dan cara pembayaran asuransi.

#### **DAFTAR PUSTAKA**

Bustami. “Penerapan Algoritma *Naïve Bayes* untuk Mengklasifikasi Data Nasabah Asuransi”. 18 September 2016. <https://www.scribd.com/doc/239130383/Penerapan-Algoritma-Naive-Bayes-Untuk-Mengklasifikasi-Data-Nasabah-Asuransi>.