BAB II

LANDASAN TEORI

2.1 Pengertian TOEFL

Test of English as a Foreign Language disingkat TOEFL adalah ujian kemampuan berbahasa Inggris (logat Amerika) yang diperlukan untuk mendaftar masuk ke Universitas di Amerika Serikat atau negara-negara lain. Ujian ini sangat diperlukan bagi pendaftar atau pembicara yang bahasa ibunya bukan bahasa Inggris. Jenis tes bahasa Inggris TOEFL ini pada umumnya diperlukan untuk persyaratan masuk kuliah pada hampir semua universitas di Amerika Serikat dan Kanada baik untuk program *undergraduate* (S-1) maupun *graduate* (S-2 atau S-3). Hasil tes TOEFL ini juga dipakai sebagai bahan pertimbangan mengenai kemampuan bahasa Inggris dari calon mahasiswa yang mendaftar ke universitas di negara lain, termasuk Universitas di Eropa dan Australia. TOEFL lebih berorientasi kepada *American English*, selain itu TOEFL pada dewasa ini sudah mulai digunakan dalam dunia kerja sebagai salah satu mekanisme jenjang kenaikan pangkat (Saifuddin, dkk. 2006).

Tes bahasa inggris TOEFL terdiri dari dua jenis yaitu *Computer-based Testing* dan *Paper-Based Testing*. Model *Computer-based Testing* adalah tes yang menggunakan media komputer. Skor penilaian model ini berada pada kisaran 216-677, sedangkan *Paper Based Testing* adalah tes yang menggunakan kertas sebagai media pengujiannya. Skor penilaian dengan model ini berada pada kisaran 450-550 ke atas (Saifuddin, dkk. 2006). Nilai hasil ujian TOEFL berkisar antara: 310 (nilai minimum) sampai 677 (nilai maximum) untuk versi PBT *(paper-based test).*

2.2 Komponen D1 dan TOEFL

Universitas Muhammadiyah Gresik telah menyelenggarakan program pembelajaran D1 bahasa Inggris untuk memperdalam kemampuan berbahasa Ingris mahasiswanya, khususnya untuk mahasiswa jurusan teknik informatika.

Program D1 bahasa Inggris diterapkan pada saat semester 1 dan 2. Pada akhir D1 pada semester 2 selalu diadakan TOEFL. Materi perkuliahan D1 bahasa Inggris mempelajari tentang *speaking, listening, writing and reading* yang juga mendasari soal TOEFL kecuali *speaking*, setelah berakhirnya program D1 mahasiswa akan diberi sertifikat berupa transkrip nilai. Tiga komponen pokok yang mendasari soal TOEFL *paper based testing* yaitu *listening comprehension, structure&written expression, and reading comprehension & vocab*. Hasil akhir TOEFL berupa jumlah skor dari tiga komponen pokok yang mendasari soal TOEFL. Prediksi kategori TOEFL menggunakan kriteria D1(*listening, reading and writing*) dan hasil skor TOEFL(*listening comprehension, structure & written expression, and reading comprehension & vocab).*

2.3 Data Mining

Data Mining adalah proses yang menggunakan teknik statistik, matematika, kecerdasan buatan dan *machine learning* untuk mengekstrasi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai database besar (Turban, dkk. 2005). Menurut (Han dan Kamber, 2006), *data mining* adalah proses menemukan pola yang menarik dan pengetahuan dari data dalam jumlah besar.

Dari beberapa pengertian diatas, dapat disimpulkan bahwa *data mining* merupakan proses ekstraksi informasi dari database yang berukuran besar untuk mendapatkan pengetahuan yang tersimpan dari data tersebut. Istilah *data mining* disebut juga *Knowledge Discovery in Database* (KDD). Istilah *data mining* sering dipakai, mungkin istilah ini lebih pendek dari *Knowledge Discovery in Database*. Sebenarnya kedua istilah tersebut memiliki konsep yang berbeda, tetapi berkaitan satu sama lain. Data *mining* dianggap hanya sebagai suatu langkah penting dalam KDD. Pekerjaan yang berkaitan dengan data mining dapat dibagi menjadi empat kelompok, yaitu model prediksi (*prediction modelling*), analisis kelompok (*cluster analysis*), analisis asosiasi (*association analysis*) dan deteksi anomali (*anomaly detection*) (Prasetyo, 2012).

2.3.1 Model Prediksi

Model prediksi berkaitan dengan pembuatan sebuah model yang dapat melakukan pemetaan dari setiap himpunan variable ke setiap targetnya, kemudian menggunakan model tersebut untuk memberikan nilai target pada himpunan baru yang didapat. Ada dua jenis model prediksi (Prasetyo, 2012):

1. Klasifikasi

   Klasifikasi digunakan untuk variable target diskret, hanya beberapa jenis kemungkinan nilai target yang didapatkan dan tidak ada nilai deret waktu (*time series*) untuk mendapatkan target nilai akhir.
2. Regresi

   Regresi untuk variable bersifat target kontinu, ada nilai deret waktu yang harus dihitung untuk mendapatkan nilai target akhir yang diinginkan.

2.3.2 Klasifikasi (*Classification)*

Klasifikasi (*Classification*) merupakan proses untuk menemukan sekumpulan model yang menjelaskan dan membedakan kelas data, sehingga model tersebut dapat digunakan untuk memprediksi nilai suatu kelas yang belum diketahui pada sebuah objek. Untuk mendapatkan model, kita harus melakukan analisis terhadap data latih (*training set*), sedangkan data uji (*test set*) digunakan untuk mengetahui tingkat akurasi dari model yang dihasilkan (Tang, 2005). Klasifikasi dapat digunakan untuk memprediksi nama atau nilai kelas dari suatu obyek data. Metode klasifikasi diantaranya, adalah *Artificial Neural Network* (ANN), *Support Vector Machine* (SVM), *Decision Tree*, Bayesian dan sebagainya

2.4 Teorema Bayes

Bayes merupakan teknik klasifikasi untuk prediksi probabilitas berbasis sederhana yang berdasar pada penerapan teorema Bayes (atau aturan Bayes) dengan asumsi independensi yang kuat pada fitur, maksudnya adalah sebuah fitur pada sebuah data tidak berkaitan dengan ada atau tidaknya fitur lain dalam data yang sama. Dalam Bayes masing-masing fitur seolah tidak memiliki hubungan

apa pun. Prediksi Bayes didasarkan pada teorema Bayes dengan formula umum sebagai berikut (Prasetyo, 2012):

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

Keterangan :

P(H | E) = Probabilitas akhir bersyarat (*conditional probability*) suatu hipotesis H terjadi jika diberikan bukti (*evidence*) E terjadi.

P(E | H) = Probabilitas sebuah bukti E terjadi akan mempengaruhi hipotesis H.

P(H) = Probabilitas awal (priori) hipotesis H terjadi tanpa memandang bukti apa pun.

P(E) = Probabilitas awal (priori) bukti E terjadi tanpa memandang hipotesis atau bukti yang lain.

Ide dasar dari aturan Bayes adalah bahwa hasil dari hiotesis atau peristiwa (H) dapat diperkirakan berdasarkan pada beberapa bukti (E) yang diamati. Ada beberapa hal penting dari aturan Bayes tersebut, yaitu:

1. Sebuah probabilitas awal/priori H atau P(H) adalah probabilitas dari suatu hipotesis sebelum bukti diamati.
2. Sebuah probabilitas akhir H atau P(H|E) adalah probabilitas dari suatu hipotesis setelah bukti diamati.

2.5 *Naive Bayes Classifier*

Klasifikasi *Naive Bayes* adalah metode yang berdasarkan probabilitas dan teorema Bayes dengan asumsi bahwa setiap variabel bersifat bebas (*independence*) dan mengasumsikan bawa keberadaaan sebuah fitur tidak ada kaitannya dengan keberadaan fitur yang lain. 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 (Prasetyo, 2012):

$$P(Y|X) = \frac{P(Y)\prod_{i=1}^{q} P(X_i|Y)}{P(X)} \quad \text{.....(2.2)}$$

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 $P(Y)\prod_{i=1}^{q} P(X_i|Y)$. Umumnya *Naive Bayes* mudah dihitung untuk fitur bertipe kategoris seperti pada contoh diatas. Namun untuk tipe numerik, ada perlakuan khusus sebelum dimasukkan *Naive 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 kedalam 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 pada sebuah kelas $P(X_i|Y)$. Untuk setiap kelas yj, probabilitas bersyarat kelas yj untuk fitur X_i adalah :

   $$P(X_{i = Xi}|Y = ij) = \frac{1}{\sqrt{2\pi}\,\sigma_{ij}} exp^{-\frac{(x_i-\mu_{ij})^2}{2\sigma_{ij}^2}} \quad \text{.....(2.3)}$$

   Keterangan :

   μ_{ij} = rata-rata X_i ($\bar{x}$) dari semua data latih

   σ_{ij}^2 = varian dari data latih

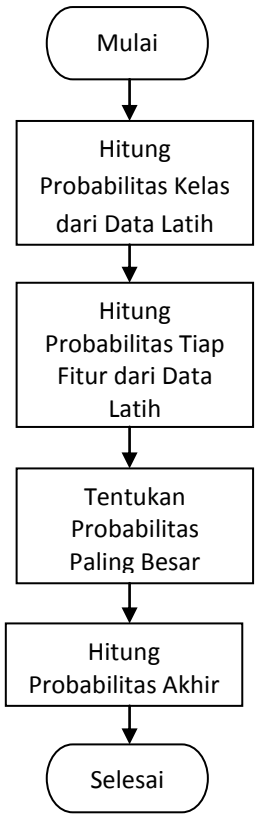
2.5.1 Algoritma Klasifikasi Naive *Bayes*

Algoritma Klasifikasi *Naive Bayes* dihitung sesuai dengan rumus $P(Y) \prod_{i=1}^{q} P(X_i|Y)$, tahapan perhitungannya dijelaskan sebagai berikut (Prasetyo, 2012):

1. Menghitung nilai probabilitas kelas berdasarkan data latih → P(Y)…..(2.4)
2. Menghitung nilai probabilitas tiap fitur berdasarkan data latih → $\prod_{i=1}^{q} P(X_i|Y)$…… (2.5)
   - Untuk fitur tiap data uji yang bertipe numerik menggunakan rumus :

     $P(X_{i=Xi}|Y = yj) = \frac{1}{\sqrt{2\pi}\,\sigma_{ij}} exp^{-\frac{(x_i-\mu_{ij})^2}{2\sigma_{ij}^2}}$

3. Menghitung probabilitas akhir
   - Mengalikan hasil dari P(X) dan $\prod_{i=1}^{q} P(X_i|Y)$ pada masing-masing kelas dan data uji.....(2.6)
4. Data uji akan diklasifikasikan pada kelas dengan nilai probabilitas akhir terbesar sesuai **gambar 2.1**



Gambar 2.1 *Flowchart Naive Bayes*

2.5.2 Contoh Perhitungan

Algoritma *Naive Bayes* dapat digunakan untuk proses pengklasifikasian menggunakan pelatihan data dalam jumlah kecil. Tahapan perhitungan algoritma klasifikasi *Naive Bayes* telah dijelaskan diatas dan berikut ini adalah contoh perhitungannya pada **tabel 2.1** (Prasetyo, 2012):

Tabel 2.1 Contoh perhitungan *Naive Bayes* klasifikasi hewan

- 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

- Jika ditambahkan sebuah data uji hewan musang dengan nilai fitur: penutup kulit = rambut, melahirkan = ya, berat = 15. Masuk ke kelas manakah untuk hewan musang tersebut?

Tabel 2.2 Perhitungan probabilitas fitur dan kelas

Penutup Kulit		Melahirkan	
Mamalia	**Reptil**	**Mamalia**	**Reptil**
Sisik = 0 Bulu = 1 Rambut = 5 Cangkang = 0	Sisik = 2 Bulu = 0 Rambut = 0 Cangkang = 2	Ya = 6 Tidak = 0	Ya = 1 Tidak = 3

Lanjutan Tabel 2.2

Penutup Kulit		**Melahirkan**	
Mamalia	**Reptil**	**Mamalia**	**Reptil**
P(Kulit = Sisik| Mamalia) = 0 P(Kulit = Bulu | Mamalia) = 1/6 P(Kulit = Rambut | Mamalia) = 5/6 P(Kulit = Cangkang | Mamalia) = 0	P(Kulit = Sisik | Reptil) = 0.5 P(Kulit = Bulu | Reptil) = 0 P(Kulit = Rambut | Reptil) = 0 P(Kulit = Cangkang | Reptil) = 0.5	P(Lahir = Ya | Mamalia) = 1 P(Lahir = Tidak | Mamalia) = 0	P(Lahir = Ya | Reptil) = 0.25 P(Lahir = Tidak | Reptil) = 0.75
Berat		**Kelas**	
Mamalia	**Reptil**	**Mamalia**	**Reptil**
$\bar{x}_{mamalia} = 38.55$ $s^2_{mamalia} = 1960.255$ $s_{mamalia} = 44.275$	$\bar{x}_{reptil} = 4.425$ $s^2_{reptil} = 23.6425$ $s_{reptil} = 4.8624$	Mamalia = 6 P(Mamalia) = 6/10 = 0.6	Reptil = 4 P(Reptil) = 4/10 = 0.4

- Hitung nilai probabilitas untuk fitur dengan tipe numerik yaitu berat.

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

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

- 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) = α x P(Mamalia) x P(X | Mamalia) = α x 0.0065 x 0.6 = 0.0039α
  - P(Reptil | X) = α x P(Reptil) x P(X | Reptil) = α x 0.4 x 0 = 0
- α = 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 (0.0039α)**, maka data uji musang diprediksi sebagai kelas **mamalia**.

2.6 Tinjauan Penelitian Sebelumnya

Naive Bayes merupakan metode popular yang banyak digunakan untuk klasifikasi. Beberapa riset yang telah dilakukan berkaitan dengan kasus prediksi yang menggunakan metode *Naive Bayes*, antara lain :

1. Penelitian dengan judul "Sistem Prediksi Prestasi (IPK) Mahasiswa Berdasarkan Latar Belakang Sekolah Asal dan Atribut Mahasiswa Ketika Awal Masuk Kuliah Menggunakan *Naive Bayes*" oleh Meinggan Vilian Sari Teknik Informatika Universitas Muhammadiyah Gresik. Penelitian dilakukan untuk memprediksi mahasiswa yang dalam dua kategori, IPK tinggi dan IPK rendah. Data sampel yang digunakan adalah data mahasiswa angkatan 2010 semester 6 sejumlah 103 mahasiswa. Variabel penetu yang digunakan adalah instansi sekolah, status sekolah, jurusan sekolah, motivasi kuliah, status kerja dan nilai danem. Penelitian menghasilkan sebuah sistem yang dapat membantu pihak Kaprodi dan mahasiswa untuk mengetahui kategori IPK secara dini.
2. Penelitian dengan judul "*Penerapan Algoritma Naive Bayes Untuk Mengklasifikasi Data Nasabah Asuransi*" oleh Bustami Dosen Teknik Informatika Universitas Malikussaleh. Penelitian dilakukan untuk mengklasifikasikan nasabah yang lancar dan tidak lancar dalam pembayaran premi asuransi. Adapun data sampel yang digunakan adalah data 20 orang nasabah asuransi. Variabel penentu yang digunakan adalah jenis kelamin, usia,

status, pekerjaan, penghasilan, masa asuransi dan cara pembayaran. Penelitian menghasilkan sebuah sistem klasifikasi yang dapat menjadi pertimbangan untuk pihak asuransi mengetahui calon nasabah yang lancar dan tidak lancar dalam membayar premi nantinya.

3. Penelitian yang berjudul "*Data mining classification untuk prediksi lama masa studi mahasiswa berdasarkan jalur penerimaan dengan metode naive bayes*" oleh John Fredrik Ulysses Teknik Informatika Universitas Atmajaya Yogyakarta. Data sampel yang digunakan adalah data dari 57 alumni mahasiswa STMIK Palangkaraya jurusan D3 Manajemen Informatika tahun kelulusan 2006-2008. Variabel yang digunakan adalah lama studi/semester dan dibagi menjadi 2 kelas, yaitu jalur khusus dan SPMB. Hasil penelitian menunjukkan bahwa mahasiswa yang masuk melalui jalur khusus memiliki kecenderungan untuk lulus lebih cepat dibandingkan mahasiswa melalui jalur SPMB.
4. Penelitian yang berjudul "*Penerapan Data Mining untuk Evaluasi Kinerja Akademik Mahasiswa Menggunakan Algoritma Naive Bayes*" oleh Mujib Ridwan, dkk. Penelitian dilakukan untuk mengklasifikasikan kelulusan mahasiswa dengan cara mengevaluasi kinerja pada tahun pertama dan tahun kedua. Data training yang digunakan adalah data mahasiswa angkatan 2005-2009 yang sudah dinyatakan lulus berdasarkan variabel yang digunakan adalah NIM, jenis kelamin, asal sekolah, jalur masuk, nilai ujian nasional, gaji orangtua, IP semester 1-4, IPK semester 1-4, keterangan lulus dan data riwayat kuliah. Data testing berupa data akademik mahasiswa angkatan 2010-2011 yang belum lulus berdasarkan variabel NIM, jenis kelamin, asal sekolah, jalur masuk, nilai ujian nasional, gaji orangtua, IP semester 1-4, IPK semester 1-4 dan data riwayat kuliah. Hasil penelitian menunjukkan pengujian data mahasiswa angkatan 2005-2009 menghasilkan nilai *precision, recall* dan *accuracy* masing-masing 83%, 50% dan 70%. Hasil tersebut akan dijadikan sebagai rule untuk menetukan kelas pada data testing.