BAB II
TINJAUAN PUSTAKA

2.1 Pendidikan Tinggi

Pendidikan tinggi merupakan pendidikan jenjang untuk menuntut ilmu setelah pendidikan menengah (SMA/Sederajat). Pendidikan tinggi terdiri dari berbagai jenjang, yaitu jenjang diploma, sarjana, magister, doktor, dan program profesi. Satuan badan yang menjalankan pendidikan tinggi dan dikenal dengan Perguruan Tinggi Negeri (PTN) dan Perguruan Tinggi Swasta (PTS). Bentuk perguruan tinggi ada bermacam-macam, seperti universitas, institute, sekolah tinggi, politeknik, spesialis, dan akademi. (Undang-Undang Republik Indonesia Nomor 12 Pasal 1, 2012).

Fungsi pendidikan tinggi berdasarkan Undang-Undang Republik Indonesia Nomor 12 Pasal 4 tahun 2012 adalah sebagai berikut :

1. Mengembangkan kemampuan dan membentuk watak serta peradaban bangsa yang bermartabat dalam rangka mencerdaskan kehidupan bangsa.
2. Mengembangkan sivitas akademika yang inovatif, responsif, kreatif, terampil, berdaya saing, dan kooperatif melalui pelaksanaan tridharma, dan,
3. Mengembangkan ilmu pengetahuan dan Teknologi dengan memperhatikan dan menerapkan nilai humaniora.

Adapun tujuan pendidikan tinggi berdasarkan Undang-Undang Republik Indonesia Nomor 12 Pasal 5 adalah :

1. Berkembangnya potensi Mahasiswa agar menjadi manusia yang beriman dan bertakwa kepada Tuhan Yang Maha Esa dan berakhlak mulia, sehat, berilmu, cakap, kreatif, mandiri, terampil, kompeten, dan berbudaya untuk kepentingan bangsa.

2. Dihasilkannya lulusan yang menguasai cabang Ilmu Pengetahuan dan/atau Teknologi untuk memenuhi kepentingan nasional dan peningkatan daya saing bangsa.

3. Dihasilkannya Ilmu Pengetahuan dan Teknologi melalui Penelitian yang memperhatikan dan menerapkan nilai Humaniora agar bermanfaat bagi kemajuan bangsa, serta kemajuan peradaban dan kesejahteraan umat manusia.

4. Terwujudnya Pengabdian kepada Masyarakat berbasis penalaran dan karya Penelitian yang bermanfaat dalam memajukan kesejahteraan umum dan mencerdaskan kehidupan bangsa.

2.2 Data Mining

Data Mining adalah proses mengolah atau merangkum data yang berjumlah besar melalui proses analisis agar bisa mengambil kesimpulan data yang berharga. Selain itu, bisa diartikan dengan gabungan antara metode statistik dan *artificial intelligence*/kecerdasan buatan yang terus berkembang (Gorunescu, 2011).

2.2.1 Model Supervised Learning

Model ini digunakan untuk memprediksi hasil masa depan berdasarkan data historis untuk dipelajari menggunakan metode tertentu agar bisa memprediksi dengan akurat. Contoh penerapan metode *Supervised Learning* adalah untuk memprediksi kemungkinan terjadinya bahaya yang akan terjadi dengan melihat beberapa faktor sesuai dengan data historis yang telah dipelajari, seperti bencana gempa bumi, banjir, dan lain-lain (Jain, 2015).

Supervised Learning adalah sebuah pendekatan dengan cara melatih data yang sudah ada dan terdapat variabel yang ditargetkan sehingga tujuan dari pendekatan ini adalah mengkelompokkan suatu data ke data yang sudah ada. Metode analisis untuk *Supervised Learning* adalah *Decision Tree, Random Forest, Nearest-Neighbor Classifier, Naive Bayes Classifier, Artificial Neural Network, Support Vector Machine, Fuzzy K-Nearest Neighbor* (Chandra, 2017).

2.2.2 Model Unsupervised Learning

Unsupervised Learning tidak memiliki data latih, sehingga dari data yang ada, kita mengelompokkan data tersebut menjadi 2 bagian atau 3 bagian dan seterusnya. Contoh metode ini adalah seseorang belum pernah membeli buku sama sekali, namun dalam suatu hari, orang tersebut membeli banyak buku dan ingin membaginya kedalam beberapa kategori agar nantinya mudah dicari namun belum diketahui kategori dari buku tersebut. Maka pembagian buku

dengan cara mengidentifikasi buku mana yang mirip berdasarkan isinya. Metode analisis *Unsupervised Learning* adalah *K-Means, Hierarchical Clustering, DBSCAN, Fuzzy C-Means, Self-Organizing Map* (Chandra, 2017).

2.3 Klasifikasi

Klasifikasi adalah metode untuk memprediksi suatu kejadian atau keputusan yang akan datang berada di suatu titik. Klasifikasi merupakan suatu pekerjaan yang melakukan penelitian terhadap suatu objek data untuk masuk dalam suatu kelas tertentu dari sejumlah kelas yang tersedia (Prasetyo, 2012). Tahapan klasifikasi ada beberapa, yaitu :

2.3.1 Data Collection

Data Collection adalah proses mengumpulkan dan memastikan informasi pada *variable of interest* (subjek yang akan dilakukan uji coba), dengan cara yang sistematis yang memungkinkan seseorang dapat menjawab pertanyaan dari permasalahan yang ada dan mengevaluasi hasil. Komponen pengumpulan data dari penelitian ini bersifat umum, bisa dilakukan untuk semua bidang studi termasuk ilmu fisik dan sosial, humaniora, bisnis, dan lainnya (Redaksi, 2014).

2.3.2 Data Cleaning

Data yang baik dan berkualitas adalah kunci dasar untuk menghasilkan data yang berkualitas, dengan cara menghapus data *error*, menghapus data *incomplete* atau data yang nilai atributnya hilang atau datanya kosong, dan

menyamakan satuan atau nilai dari data (Subhan, 2017). *Data Cleaning* adalah proses analisa kualitas dari suatu data dengan cara mengubah, mengoreksi, atau menghapus data-data yang salah, tidak lengkap, tidak akurat, atau memiliki format yang salah dalam basis data guna menghasilkan data berkualitas tinggi (Tawakal, 2015). Contoh :

Tabel 2.1 *Data Sebelum Cleaning*

No	Jenis Klmn	Prodi	Fakultas	Tempat Lahir	Kabupaten	Propinsi
1	P	Farmasi	MATEMATIKA DAN ILMU PENGETAHUA	WERU	KAB.INDRAMAYU	JAWA BARAT
2	L	Ilmu Komunikasi	PSIKOLOGI DAN ILMU SOSIAL BUDAYA	Bumi Agung	KAB.SLEMAN	D.I. YOGYAKARTA
3	L	Manajemen	EKONOMI	Tuban	KODYA YOGYAKARTA	D.I. YOGYAKARTA
4			TEKNIK SIPIL DAN PERENCANAAN	Lamongan		
5				Semarang		
6	L		TEKNIK SIPIL DAN PERENCANAAN	BLORA		
7	P	Psikologi	PSIKOLOGI DAN ILMU SOSIAL BUDAYA	Jakarta	KODYA JAKARTA UTARA	DKI JAKARTA
8	L	Teknik Sipil	TEKNIK SIPIL DAN PERENCANAAN	Bantul	KAB.BANTUL	D.I. YOGYAKARTA
9	L	Farmasi	MATEMATIKA DAN ILMU PENGETAHUA	Klaten	KAB.KLATEN	JAWA TENGAH
10	L	Teknik Industri	TEKNOLOGI INDUSTRI	Kediri	KAB.KEDIRI	JAWA TIMUR
11	L	Informatika	TEKNOLOGI INDUSTRI	Tuban	KAB.TUBAN	JAWA TIMUR
12	L	Teknik Kimia	TEKNOLOGI INDUSTRI	Kendal	KAB.KENDAL	JAWA TENGAH
13	P	Hukum	HUKUM	Madiun		
14	P	Teknik Kimia	TEKNOLOGI INDUSTRI	Klangenan Cirebon	KAB.CIREBON	JAWA BARAT
15	P	Psikologi	PSIKOLOGI DAN ILMU SOSIAL BUDAYA	Serang	X kab.pandegelang	JAWA BARAT
16	L	Pendidikan Agama Islam	ILMU AGAMA ISLAM	Sleman	KAB.SLEMAN	D.I. YOGYAKARTA
17	L	Manajemen	EKONOMI	Serang		
18	L	Manajemen	EKONOMI	Yogyakarta	KAB.KENDAL	JAWA TENGAH
19	L	Hukum	HUKUM	Sanggau	KAB.SANGGAU	KALIMANTAN BARAT
20	L	Hukum	HUKUM	Samarinda	KODYA YOGYAKARTA	D.I. YOGYAKARTA

Tabel 2.2 *Data Setelah Cleaning*

No	Jenis Klmn	Prodi	Fakultas	Tempat Lahir	Kabupaten	Propinsi
1	P	Farmasi	MATEMATIKA DAN ILMU PENGETAHUA	WERU	KAB.INDRAMAYU	JAWA BARAT
2	L	Ilmu Komunikasi	PSIKOLOGI DAN ILMU SOSIAL BUDAYA	Bumi Agung	KAB.SLEMAN	D.I. YOGYAKARTA
3	L	Manajemen	EKONOMI	Tuban	KODYA YOGYAKARTA	D.I. YOGYAKARTA
7	P	Psikologi	PSIKOLOGI DAN ILMU SOSIAL BUDAYA	Jakarta	KODYA JAKARTA UTARA	DKI JAKARTA
8	L	Teknik Sipil	TEKNIK SIPIL DAN PERENCANAAN	Bantul	KAB.BANTUL	D.I. YOGYAKARTA
9	L	Farmasi	MATEMATIKA DAN ILMU PENGETAHUA	Klaten	KAB.KLATEN	JAWA TENGAH
10	L	Teknik Industri	TEKNOLOGI INDUSTRI	Kediri	KAB.KEDIRI	JAWA TIMUR
11	L	Informatika	TEKNOLOGI INDUSTRI	Tuban	KAB.TUBAN	JAWA TIMUR
12	L	Teknik Kimia	TEKNOLOGI INDUSTRI	Kendal	KAB.KENDAL	JAWA TENGAH
14	P	Teknik Kimia	TEKNOLOGI INDUSTRI	Klangenan Cirebon	KAB.CIREBON	JAWA BARAT
15	P	Psikologi	PSIKOLOGI DAN ILMU SOSIAL BUDAYA	Serang	X kab.pandegelang	JAWA BARAT
16	L	Pendidikan Agama Islam	ILMU AGAMA ISLAM	Sleman	KAB.SLEMAN	D.I. YOGYAKARTA
18	L	Manajemen	EKONOMI	Yogyakarta	KAB.KENDAL	JAWA TENGAH
19	L	Hukum	HUKUM	Sanggau	KAB.SANGGAU	KALIMANTAN BARAT
20	L	Hukum	HUKUM	Samarinda	KODYA YOGYAKARTA	D.I. YOGYAKARTA

2.3.3 Data Reduction

Pada tahap data *Reduction* adalah tahap dimana data yang telah berkumpul, kemudian di bersihkan, proses selanjutnya adalah memiliki variabel atau atribut yang akan digunakan dalam penelitian. Tahap ini dilakukan untuk mengurangi atribut yang tidak digunakan akan tetapi tetap bersifat informatif (Subhan, 2017).

Tabel 2.3 Data Sebelum Dan Sesudah *Reduce Variabel*

Variabel Asli		**Variabel Setelah Reduce**
No Mhs		Jenis Klmn
Nama Mhs		Prodi
Jenis Klmn		Provinsi
Prodi		Pendidikan Ayah
Provinsi		Pendidikan Ibu
Pendidikan Ayah		
Pendidikan Ibu		

2.3.4 *Data Training* dan *Data Testing*

Data *training* digunakan oleh algoritma untuk membentuk sebuah model klasifikasi. Model ini merupakan representasi pengetahuan yang akan digunakan untuk mengukur sejauh mana klasifikasi berhasil melakukan prediksi dengan benar. Karena itu, data yang ada pada data *testing* seharusnya tidak boleh ada pada data *training* sehingga dapat diketahui apakah model klasifikasi dapat melakukan klasifikasi dengan baik. Proporsi antara data

training dan data *testing* tidak mengikat tetapi agar variasi dalam model tidak terlalu besar maka disarankan data *training* lebih besar dibandingkan data *testing*. Biasanya ¾ dari total data dijadikan data *training* sedangkan sisanya dijadikan data *testing*. Selain itu, ada pula penelitian yang menghasilkan keakuratan model klasifikasi optimum dengan proporsi 75 : 25 untuk data *training* dan data *testing* (Rachman dan Purnami, 2012).

2.4 Balancing Data

Balancing data adalah merubah data yang tidak seimbang (*imbalance* data) menjadi data yang seimbang (*balance*). *Imbalance* data adalah kondisi ketidakseimbangan dalam jumlah data *training* antara dua kelas yang berbeda, salah satu kelasnya mempresentasikan jumlah data yang sangat kecil (*minority class)* (Sastrawan, Baizal, dan Bijaksana, 2010). Metode ini secara luas dikenal sebagai 'Metode Sampling'. Umumnya, metode ini bertujuan untuk memodifikasi data yang tidak seimbang ke dalam distribusi yang seimbang menggunakan beberapa mekanisme. Modifikasi terjadi dengan mengubah ukuran kumpulan data asli dan menyediakan proporsi keseimbangan yang sama.

Menurut *Analytics Vidhya Content Team* (2016), menjelaskan bahwa *balancing* data ada beberapa metode, yaitu diantaranya adalah *under sampling.* Metode *under sampling* ini bekerja dengan kelas mayoritas (*majority class),* dengan mengurangi jumlah observasi dari kelas mayoritas

untuk membuat kumpulan data dengan jumlah yang seimbang. Metode ini paling baik digunakan ketika kumpulan data sangat besar dan mengurangi jumlah sampel pelatihan yang membantu meningkatkan waktu operasi dan masalah penyimpanan. Metode *Under Sampling* secara acak memilih observasi dari kelas mayoritas yang dieliminasi sampai set data menjadi seimbang antar masing-masing kelas.

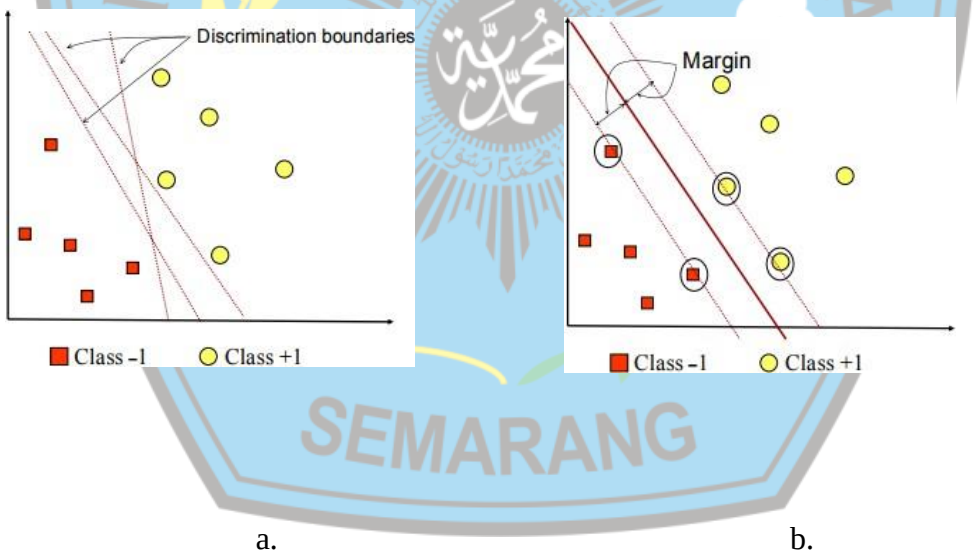
2.5 *Support Vector Machine* (SVM)

Vapnik memperkenalkan SVM untuk pertama kali pada tahun 1992 sebagai rangkaian konsep unggulan pada bidang *pattern recognition*. Usia SVM sebagai salah satu metode *pattern recognition* masih terbilang relatif muda. Dewasa ini SVM merupakan salah satu metode yang berkembang pesat. SVM merupakan salah satu metode *machine learning* yang bekerja atas prinsip *Structural Risk Minimization* (SRM) yang bertujuan untuk menemukan *hyperlane* terbaik untuk memisahkan dua kelas data. SVM bekerja dengan memaksimalkan *margin* yang merupakan jarak pemisah antara kedua kelas data tersebut (Pratiwi, 2017). Pada dasarnya SVM mempunyai prinsip linear, akan tetapi kini SVM telah berkembang sehingga dapat bekerja pada masalah *non*-linear. Cara kerja SVM pada masalah *non*-linear adalah dengan memasukkan konsep kernel pada ruang berdimensi tinggi. Dalam ruang yang berdimensi ini, nantinya akan dicari pemisah atau yang sering disebut *Hyperlane.*

Hyperlane dapat memaksimalkan jarak atau *margin* antara kelas data. *Hyperlane* terbaik antara kedua kelas dapat ditemukan dengan mengukur *margin* dan kemudian mencari titik maksimalnya. Usaha dalam mencari *Hyperlane* yang terbaik sebagai pemisah kelas-kelas adalah inti dari proses dalam metode SVM (Assaffat, 2015).

2.5.1 Linear Separable Data

Metode SVM dengan *hyperlane* yang berbentuk garis lurus disebut dengan *linear saparable*. **Gambar 2.1** merupakan ilustrasi dari *hyperlane linear separable* data ;



a. b.

Gambar 2.1 *Garis Linear Pemisah Dua Kelas (Sumber : Nugroho, 2003)*

Dapat dilihat ilustrasi pada **Gambar 2.1** adalah beberapa *pattern* yang merupakan anggota dari dua buah kelas yaitu kelas +1 dan kelas -1. Simbol

untuk *pattern* pada kelas -1 adalah kotak yang berwarna merah, sedangkan simbol untuk *pattern* pada kelas +1 adalah lingkaran dengan warna kuning. Dalam SVM yang telah disebutkan diatas menemukan garis (*hyperlane*) yang dapat memisahkan antara kedua kelompok tersebut. Berbagai macam garis pemisah (*discrimination boundaries*) *alternative* yang ditunjukkan **Gambar 2.1** bagian a. Dalam menemukan *hyperlane* yaitu dengan cara mengukur *Margin hyperlane* tersebut dan kemudian mencari titik maksimalnya. Jarak antara *hyperlane* dengan *pattern* pada masing-masing kelas biasa disebut dengan *margin*. Untuk *pattern* paling dekat disebut dengan *support vector*. Pada **Gambar 2.1** bagian b garis yang berada di tengah menunjukkan *hyperlane* yang terbaik, karena terletak tepat pada tengah-tengah antar kelas, sedangkan *Support Vector* adalah titik merah dan kuning yang berada dalam lingkaran hitam. Usaha dalam mencari lokasi *hyperlane* ini merupakan proses inti dari SVM.

Pada penelitian ini data yang tersedia dapat dinotasikan sebagai $x \in R^d$, sedangkan label untuk masing-masing kelas dinotasikan $y_i \in \{-1, +1\}$ untuk i = 1,2,3, ... n.

Pada kedua kelas dapat diasumsikan terpisah secara sempurna oleh *Hyperlane* berdimensi *d*, yang didefinisikan sebagai berikut:

$$\vec{w} \cdot \vec{x} + b = 0 \qquad \text{...(2.1)}$$

Untuk *pattern* x_i yang termasuk kelas -1 dapat dirumuskan sebagai *pattern* yang memenuhi pertidaksamaan:

$$\vec{w} \cdot \vec{x} + b \leq -1 \quad \text{...(2.2)}$$

Sedangkan untuk *pattern* x_i yang termasuk kelas +1 dapat dirumuskan sebagai berikut:

$$\vec{w} \cdot \vec{x} + b \geq +1 \quad \text{...(2.3)}$$

dimana

R^d = ruang vektor

d = dimensi

n = banyak data

$\vec{w}$ = vektor bobot

$\vec{x}$ = vektor data (*input*)

b = bias

Margin terbesar dapat diperoleh dengan cara memaksimalkan nilai jarak antara jarak dan titik terdekatnya, yaitu $\frac{1}{\|w\|}$. Hal ini dapat dirumuskan sebagai masalah *Quadratic Programming* (QP), yaitu mencari titik minimal persamaan 3.4, dengan memperlihatkan *constraint* persamaan 3.5. (Pratiwi, 2017)

$$min_w = \tau(w) = \frac{1}{2}\|w\|^2 \quad ...(2.4)$$

$$y_i(x_i \,.w + b) - 1 \geq 0 \quad ...(2.5)$$

dimana $\|w\|$ adalah vektor normal.

Salah satu teknis komputasi yaitu *lagrange multiplier* yang dapat memecahkan masalah ini dapat dinyatakan pada persamaan 2.6.

$$L(w, b, a) = \frac{1}{2\|w\|^2}\, 2 - \sum \alpha_i \left(y_i\left((\vec{x}.\vec{w} + b) - 1\right)\right);\ \text{i} = 1,2, ...,\text{n} \quad ...(2.6)$$

Dimana α merupakan *lagrange multiplier*, yang bernilai $\alpha_i \geq 0$. Nilai optimal pada persamaan 2.6 dapat dihitung dengan meminimalkan nilai L terhadap w dan b, dan memaksimalkan nilai L terhadap α_i, dengan memperhatikan sifat bahwa pada titik optimal *gradient* L = 0. Pada persamaan 2.6 dapat dimodifikasi sebagai maksimalisasi *problem* yang hanya mengandung α_i, seperti yang terlihat pada persamaan 2.7 dan persamaan 2.8 dibawah ini.

$$\sum_i^1 = 1\, \alpha_1 - \frac{1}{2}\sum_i^1 j = 1\, \alpha_i \alpha_j y_i y_j \vec{x}_i.\vec{x}_i \quad ...(2.7)$$

$$\text{dimana } \alpha_i \geq 0 (i = 1, 2, \dots, 1) \sum_i^1 = 1\, \alpha_i y_i = 0 \quad ...(2.8)$$

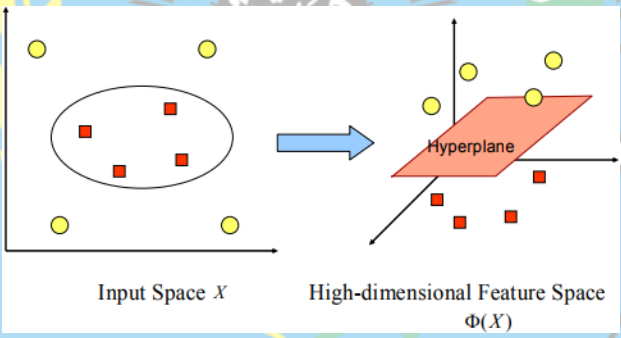
Dengan demikian, maka akan diperoleh α_i yang kebanyakan bernilai positif yang disebut sebagai *support vector* dan juga memperoleh persamaan 2.9 dan persamaan 2.10 sebagai solusi pemisah. (Pratiwi, 2017)

$$w = \sum \alpha_i \, y_i x_i \quad \text{...(2.9)}$$

$$b = y_k - w^T x_k \quad \text{...(2.10)}$$

2.5.2 Non-Linear Separable Data

Dalam dunia nyata (*real world problem)* pada umumnya masalah data yang diperoleh jarang yang bersifat *linear*, banyak yang bersifat *non linear*. Pada SVM terdapat sebuah fungsi kernel, yaitu fungsi yang digunakan untuk menyelesaikan *problem non linear*. Kernel berfungsi memungkinkan untuk mengimplementasikan suatu model pada ruandimensi lebih tinggi (ruang fitur).



Gambar 2.2 *Hyperplane* (Sumber : Nugroho, 2003)

a. Kernel *Radian* Basis *Function* (RBF)

Kernel RBF adalah populer fungsi kernel yang digunakan dalam berbagai *kernelized* algoritma pembelajaran. Secara khusus, ini biasanya digunakan dalam klasifikasi mesin vektor dukungan. (Wikipedia)

Kernel RBF pada dua sampel X dan X', direpresentasikan sebagai vektor

fitur di beberapa ruang input, didefinisikan sebagai berikut :

$$K(\vec{X}_i, \vec{X}_j = \exp(\frac{\left\|\overrightarrow{x_i} - \vec{x}_j\right\|^2}{2\sigma^2}) \quad ...(2.11)$$

b. Kernel *Sigmoid*

Fungsi *Sigmoid* adalah fungsi real yang dibatasi, dapat dibedakan, yang didefinisikan untuk semua nilai input nyata dan memiliki turunan non- negatif di setiap titik dan tepat satu titik belok. (Wikipedia)

$$K(\vec{x}_i, \vec{x}_j = \tan(\sigma x_j^t \, x_j) \quad ...(2.12)$$

2.6 Random Forest

Random Forest pertama kali dikenalkan oleh Breiman pada Tahun 2001. Dalam penelitiannya menunjukkan kelebihan *Random Forest* antara lain dapat menghasilkan *error* yang lebih rendah, memberikan hasil yang bagus dalam klasifikasi, dapat mengatasi data *training* dalam jumlah sangat besar secara efisien, dan metode yang efektif untuk mengestimasi *missing* data (Breiman, 2001).

Secara sederhana, algoritma pembentukan RF dapat disebutkan sebagai berikut. Andaikan gugus data *training* yang kita miliki berukuran n dan terdiri atas *d* perubah penjelas (*predictor)*. Tahapan penyusunan dan pendugaan menggunakan RF adalah :

a. (tahapan *bootstrap)* tarik contoh acak dengan permulihan berukuran n dari gugus data *training*.
b. (tahapan *random sub-setting)* susun pohon berdasarkan data tersebut, namun pada setiap proses pemisah pilih secara acak jumlah *variable* prediktor (*m)* < *d* peubah penjelas, dan lakukan pemisahan terbaik.
c. Ulangi langkah a-b sebanyak k kali sehingga diperoleh k buah pohon acak.
d. Lakukan pendugaan gabungan berdasarkan *k* buah pohon tersebut (misal menggunakan *majority vote* untuk kasus klasifikasi, atau rata-rata untuk kasus regresi.

Perhatikan bahwa pada setiap kali pembentukan pohon, kandidat peubah penjelas yang digunakan untuk melakukan pemisahan bukanlah seluruh peubah yang terlibat namun hanya sebagian saja hasil pemilihan secara acak. Bisa dibayangkan bahwa proses ini menghasilkan kumpulan pohon tunggal dengan ukuran dan bentuk yang berbeda-beda. Hasil yang diharapkan adalah kumpulan pohon tunggal memiliki korelasi yang kecil antar pohonnya. Korelasi kecil ini mengakibatkan ragam dugaan hasil RF menjadi kecil (Hastie *et al*, 2008) dan lebih kecil dibandingkan ragam dugaan hasil *bagging* (Zhu, 2008).

Jika melihat secara seksama algoritma pembentukan RF, salah satu yang bisa diubah adalah nilai *m*, yaitu banyaknya peubah penjelas yang digunakan sebagai kandidat pemisah dalam pembentukan pohon. Nilai *m* yang semakin

besar akan menyebabkan korelasi (ρ) semakin besar. Contoh ekstrim adalah jika kita gunakan $m = d$ yang menyebabkan setiap kali pengulangan akan menghasilkan pohon yang sama sehingga nilai korelasi akan menjadi maksimum yaitu sebesar 1. Namun, jika nilai *m* kita buat sekecil mungkin yaitu hanya 1 peubah penjelas saja yang dijadikan kandidat pemisah, maka pohon yang diperoleh akan menjadi pohon dengan akurasi yang sangat rendah. Dengan demikian jelas bahwa pemilihan *m* memegang peranan dalam menentukan kebaikan RF yang dihasilkan.

Pohon keputusan dimulai dengan cara menghitung nilai *entropy* sebagai penentu tingkat ketidakmurnian atribut dan nilai *information gain*. Untuk menghitung nilai *entropy* digunakan rumus seperti pada persamaan 2.12, sedangkan nilai *information gain* menggunakan persamaan 2..13 (Nugroho, 2017).

$$Entropy\ (Y) = -\sum p(c|Y) log^2 p(c|Y) \quad ...(2.13)$$

Dimana Y adalah himpunan kasus dan p(c|Y) merupakan proporsi nilai Y terhadap kelas c.

$$Information\ gain\ (Y, a) = Entropy\ (Y) - \sum v \in$$

$$Values(a) \left|\frac{Yv}{Ya}\right| Entropy(Yv) \quad ...(2.14)$$

Dimana *Values(a)* merupakan semua nilai yang mungkin dalam himpunan

kasus a. Yv adalah subkelas dari Y dengan kelas v yang berhubungan dengan kelas a, dan Ya adalah semua nilai yang sesuai dengan a.

2.7 Confusion Matrix

Berikut ini merupakan hasil dari *confusion matrix* (Sasongko, 2016).

Tabel 2.4 *Tabel Confusionan Matrix*

Prediksi	Aktual	
	True	*False*
True	TP	FN
False	FP	TN

Keterangan :

TP = Jumlah prediksi yang tepat bersifat positif (*True Positive).*

TN = Jumlah prediksi yang tepat bersifat negatif (*True Negative).*

FP = Jumlah prediksi yang salah bersifat positif (*False Positive).*

FN = Jumlah prediksi yang salah bersifat negatif (*False Negative).*

Accuracy merupakan proporsi jumlah prediksi benar. Rumus akurasi adalah:

$$Akurasi = \frac{TP+TN}{TP+FP+TN+FN} \quad \text{...(2.15)}$$