BAB 2

TINJAUAN PUSTAKA

2.1 Tinjauan Non Statistik

2.1.1Tuberkulosis

2.1.1.1 Definisi Tuberkulosis

Tuberculosis merupakan penyakit kronik, menular, yang disebabkan oleh Mycobacterium tuberculosis, yang ditandai dengan jaringan granulasi nekrotik (perkijauan) sebagai respons terhdap kuman tersebut. Penyakit ini menular dengan cepat pada orang yang rentan dan daya tahan tubuh lemah. Diperkirakan seseorang penderita tuberculosis kepada 1 dari 10 orang disekitarnya. Tuberculosis adalah penyakit yang menganggu sumber daya manusia dan umumnya menyerang kelompok masyarakat dengan golongan sosial ekonomi rendah (Sejati & Sofiana, 2015).

Menurut Depkes RI dalam (Murtiwi, 2005) Penyakit TBC merupakan penyakit menular yang menyebabkan kematian, dan merupakan penyebab kematian ketiga di Indonesia. Hasil Survei Kesehatan Rumah Tangga (SKRT) tahun 2001 penyakit TBC merupakan penyebab kematian ketiga setelah penyakit kardiovaskular dan penyakit saluran pernafasan pada semua kelompok usia, bahkan peringkat pertama penyebab kematian penyakit menular. Jumlah pasiennya sekitar 500.000 orang/tahun dengan kematian sekitar 175.000/ tahun, khususnya di daerah pedesaan miskin dan daerah kumuh perkotaan yang rawan kuman.

Indonesia memiliki beban penyakit tuberculosis yang tinggi. Indonesia merupakan negara pertama diantara High Burden Country (HBC) di wilayah WHO South-East Asian yang mampu mencapai target global tuberculosis untuk deteksi kasus dan keberhasilan pengobatan pada tahun 2006. Pada tahun 2009, tercatat sejumlah 294.732 kasus tuberculosis telah ditemukan dan diobati (data awal Mei 2010) dan lebih dari 169.213 diantaranya terdeteksi BTA (+). Dengan demikian, cae notification rate untuk TB BTA (+) adalah 73 per 100.000 (case detection date 73%). Rerata pencapaian angka keberhasilan pengobatan selama 4 tahun terakhir adalah sekitar 90% dan pada kohort tahun 2008 mencapai 91%. Pencapaian target global tersebut merupakan tonggak pencapaian program pengendalian TB nasional yang utama (Sejati & Sofiana, 2015).

2.1.1.2 Gejala Tuberkulosis

Gejala utama pasien TB paru adalah batuk berdahak selama 2-3 minggu atau lebih. Batuk dapat diikuti dengan gejala tambahan yaitu dahak bercampur darah, batuk darah, sesak nafas, badan lemas, nafsu makan menurun, berat badan menurun, malaise, berkeringat malam hari tanpa kegiatan fisik,demam meriang lebih dari satu bulan. Gejalagejala tersebut diatas dapat dijumpai pula pada penyakit paru selain TB, seperti bronkiektasis, bronkitis kronis, asma, kanker paru, dan lain-lain. Mengingat prevalensi TB paru di Indonesia saat ini masih tinggi, maka setiap orang yang datang ke UPK dengan gejala tersebut diatas, dianggap sebagai seorang tersangka (suspek) pasien TB, dan perlu dilakukan pemeriksaan dahak secara mikroskopis langsung pada pasien remaja dan dewasa, serta skoring pada pasien anak. Pemeriksaan dahak berfungsi untuk menegakkan diagnosis, menilai

keberhasilan pengobatan dan menentukan potensi penularan. Pemeriksaan dahak untuk penegakan diagnosis pada semua suspek TB dilakukan dengan mengumpulkan 3 spesimen dahak yang dikumpulkan dalam dua hari kunjungan yang berurutan berupa dahak Sewaktu-Pagi-Sewaktu (SPS) (Groenewald, Baird, Verschoor, Minnikin, & Croft, 2014):

1. S(Sewaktu): Dahak dikumpulkan pada saat suspek TB datang berkunjung pertama kali. Pada saat pulang, suspek membawa sebuah pot dahak untuk mengumpulkan dahak pagi pada hari kedua.
2. P(Pagi): Dahak dikumpulkan di rumah pada pagi hari kedua, segera setelah bangun tidur. Pot dibawa dan diserahkan sendiri kepada petugas di UPK.
3. S(sewaktu): Dahak dikumpulkan di UPK pada hari kedua, saat menyerahkan dahak pagi.

2.1.2 Tuberkulosis Aktif dan Laten

Menurut Blumberg dan Leonard dalam (Sinaga, Reviono, & Harsini, 2017) Infeksi Tuberkuosis terjadi karena inhalasi droplet nuclei yang mengandung kuman tuberkulosis. Setelah terpapar kuman Tuberkulosis ada empat keadaan yang bisa terjadi yaitu pertama tidak terjadi infeksi (ditandai dengan tes kulit tuberculin yang negative). Kedua terjadi infeksi kemudian menjadi Tuberkulosis yang aktif (TB primer), ketiga menjadi Tuberkulosis laten dimana mekanisme imun mencegah progresivitas penyakit menjadi Tuberkulosis aktif dan keempat menjadi Tuberkulosis laten tetapi kemudian terjadi reaktivitas dan berkembang menjadi Tuberkulosis aktif dalam beberapa bulan sampai beberapa tahun kemudian.

Menurut Jordao dalam (Setiawan & Nugraha, 2016) Infeksi tuberkulosis aktif adalah tuberkulosis yang dapat menularkan, penyakit tuberkulosis pada umumnya mengenai paru paru, yang menyebar melalui udara ketika penderita tuberkulosis tersebut bersin, batuk yang tidak ditutup dan berbicara. Sedangkan Infeksi Tuberkulosis laten didefinisikan sebagai kondisi dimana seseorang yang terinfeksi atau terpapar oleh Mycobacterium tuberculosis namun pada saat itu orang tersebut tidak sakit, tidak mempunyai gejala atau asymptomatic dan gambaran foto toraks normal. Menurut Kizza et al dalam (Setiawan & Nugraha, 2016) Infeksi TB laten didefinisikan sebagai keadaan respons imun persisten dipicu oleh antigen Mycrobacterium tuberculosis (Mtb) tanpa bukti manifestasi TB aktif secara klinis.

2.2 Tinjauan Statistik

2.2.1 Data Mining

Data Mining merupakan suatu proses penggalian data atau penyaringan data dengan memanfaatkan kumpulan data dengan ukuran yang cukup besar melalui serangkaian proses untuk mendapatkan informasi yang berharga dari data tersebut. Data mining ini juga dikenal dengan istilah pattern recognition (Sulastri & Gufroni, 2017). Salah satu teknik yang dibuat dalam data mining adalah adalah bagaimanana menelusuri data yang ada untuk membangun sebuah model, kemudian menggunakan model tersebut agar dapat mengenali pola data yang lain yang tidak berada dalam basis data yang tersimpan. Kebutuhan untuk prediksi juga dapat memanfaatkan teknik ini. Dalam data mining, pengelompokan data juga bisa dilakukan. Tujuannya adalah agar kita dapat mengetahui pola universal data-data yang ada (Haryati, Sudarsono, & Suryana, 2015). Dalam data mining terdapat dua

jenis pembelajaran yaitu *Supervised Learning* dan *Unsupervised Learning*, *Supervised learning* merupakan teknik dalam *data mining* yang digunakan untuk data yang sudah diketahui label kelasnya, sedangkan *Unsupervised Learning* merupakan teknik dalam *data mining* yang digunakan untuk data yang belum diketahui label kelasnya (Tan et al., 2006). *Data mining* memiliki lima peran yang dapat digunakan di berbagai bidang. Berikut ini adalah penjelasan dari lima peran utama *data mining*.

2.2.2 Estimasi

Estimasi hampir sama dengan klasifikasi, namun variabel target estimasi lebih ke arah numerik dari pada ke arah kategori (Pattipeilohy, Wibowo, & Utari, 2017). Model dibuat menggunakan record lengkap yang menyediakan nilai variabel dari target sebagai nilai prediksi, lalu pada peninjauan berikutnya estimasi nilai dari variabel target dibuat berdasarkan nilai variabel prediksi (Fikri, 2013).

2.2.3 Prediksi

Prediksi merupakan nilai dari hasil di masa mendatang yang akan dihasilkan (Pattipeilohy et al., 2017). Tujuan metode ini untuk membangun model prediksi suatu nilai yang mempunyai ciri-ciri tertentu (Kamagi & Hansun, 2014).

2.2.4 Klaster

Pengklusteran (Clustering), merupakan pengelompokan record, pengamatan atau memperhatikan dan membentuk kelas objek-objek yang memiliki kemiripan (Pattipeilohy et al., 2017). Berbeda dengan klasifikasi, pada pengklasteran tidak ada variable target. Pengklasteran tidak melakukan klasifikasi, estimasi, maupun

prediksi pada nilai variable target, akan tetapi pengklasteran membagi kesuluruhan data menjadi beberapa kelompok yang memiliki kemiripan (Gunadi & Sensuse, 2012).

2.2.5 Asosiasi

Menurut (Pattipeilohy et al., 2017) asosiasi adalah menemukan atribut yang muncul dalam satu waktu. Aturan asosisai adalah teknik dalam data mining yang digunakan untuk menemukan pola, aturan assosiatif, korelasi atau struktur sederhana dalam kumpulan item atau objek yang terdapat pada database transaksi, database relasional, dan repositori informasi (Nursafa'ah, 2018).

2.2.6 Klasifikasi

Klasifikasi, adalah proses untuk menemukan model atau fungsi yang menjelaskan atau membedakan konsep atau kelas data, dengan tujuan untuk mendapat memperkirakan kelas dari suatu objek yang labelnya tidak diketahui (Pattipeilohy et al., 2017). Klasifikasi (taksonomi) merupakan proses penempatan objek atau konsep tertentu ke dalam satu set kategori berdasarkan objek yang digunakan. Terdapat enam algoritma klasifikasi dari sepuluh algoritma terbaik yang dipilih oleh peneliti data mining enam diantaranya adalah algoritma klasifikasi yaitu C4.5, *Support Vector Machines* (SVM), *AdaBoost, k-Nearest Neighbor* (k-NN), *Naïve Bayes* dan CART (Wu, Ouyang, Chen, & Lu, 2015).

2.2.7 Algoritma C4.5

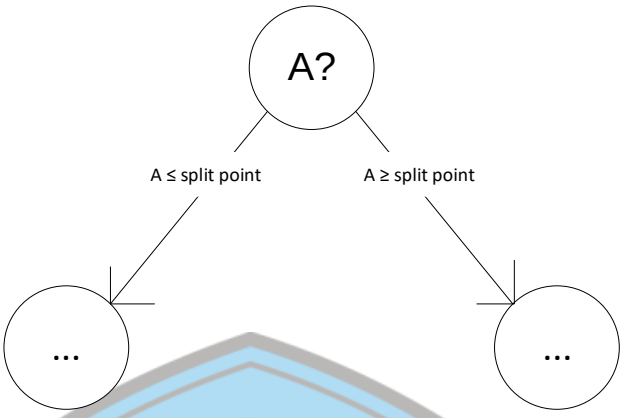
Algoritma C4.5 merupakan generasi baru dari algoritma ID3 yang dikembangkan oleh J. Ross Quinlan pada tahun 1983 (Salzberg, 1994). Algoritma

C4.5 merupakan algoritma klasifikasi data dengan teknik pohon keputusan yang dapat mengolah data numerik (kontinyu) dan diskrit, dapat menangani nilai atribut yang hilang, menghasilkan aturan-aturan yang mudah diinterpretasikan dan tercepat diantara algoritma-algoritma lain. Secara umum algoritma C4.5 untuk membangun pohon keputusan adalah sebagai berikut:

1. Pilih kriteria sebagai akar.
2. Buat cabang untuk tiap-tiap nilai.
2. Bagi kasus dalam cabang.
3. Ulangi proses untuk setiap cabang sampai semua kasus pada cabang memiliki kelas yang sama.

Dalam proses pengujian atribut, cabang baru yang terbentuk akan diperhatikan dari tipe atribut. Berikut 2 jenis cabang (*branches*) yang mungkin muncul dalam pohon keputusan (*Decision Tree*):

1. Data kontinyu adalah data yang dapat digunakan untuk operasi hitung, data yang diperoleh dari hasil penghitungan atau pengukuran, sehingga data tidak hanya berupa bilangan bulat, tetapi juga bisa dalam bentuk decimal (Sugianto, 2016). Contoh dari data ini adalah jumlah benar atau salah dalam suatu tes, skor nilai, ranking, tinggi badan, berat badan, panjang, jarak dll. Jika atribut bernilai kontinyu, maka cabang yang terbentuk akan mempunyai 2 kemungkinan dengan kondisi nilai kategori atribut ≤ *split point* dan nilai kategori atribut > *split point.* Dimana *split point* merupakan bagian dari *splitting criterion*.



Gambar 0.1 Cabang Pohon yang Terbentuk dari Atribut Bernilai Kontinyu.

Perhitungan *split point* pada gambar 2.1 didapat dari persamaan

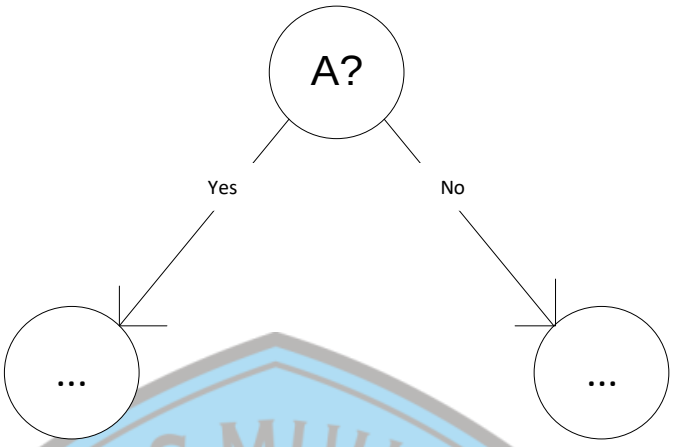
$$Split\ Point = \frac{a_1 + a_{i+1}}{2}$$

Berdasarkan persamaan diatas, berikut merupakan keterangan:

a_1 = nilai saat ini

a_{i+1}= nilai selanjutnya

2. Data diskret adalah data pengkategorian atau membedakan atau mengelompokkan jenis tertentu, data yang termasuk dalam data diskret adalah data bilangan bulat. Bilangan bulat adalah bilang yang tidak dalam bentuk pecahan/ decimal (Sugianto, 2016). Contohnya adalah Jumlah siswa laki-laki dan perempuan, responden yang menjawab ya atau tidak, pengelompokkan benda berdasarkan bentuk dan ukurannya, jawaban benar atau salah. Jika atribut bernilai diskret dan bernilai biner, maka cabang yang terbentuk akan selalu dua dengan nilai *Yes* atau *No*.



Gambar 0.2 Cabang Pohon yang Terbentuk Berdasarkan Atribut Diskret dan Biner.

Algoritma C4.5 adalah pengembangan dari algoritma ID3. Oleh karena pengembangan tersebut algoritma C4.5 mempunyai prinsip dasar kerja yang sama dengan algoritma ID3 (Defiyanti & D. L. Crispina Pardede, 2010). Ada beberapa tahap dalam membuat sebuah pohon keputusan dengan algoritma C4.5 (Haryati et al., 2015), yaitu:

1. Menyiapkan data training. Data training biasnya dari data histori yang pernah terjadi sebelumnya dan sudah dikelompokkan ke dalam kelas-kelas tertentu.
2. Menentukan akar dari pohon akar akan diambil dari atibut yang terpilih dengan cara menghitung nilai *gain* dari masing-masing atribut, nilai *gain* yang paling tinggi yang akan menjadi akar pertama. Sebelum menghitung nilai *gain* dari atribut, hitung dahulu nilai *entropy* yaitu:

$$Entropy\ (S) = \sum_{i-1}^{n} -pi \times log_2 pi$$

Keterangan:

S: himpunan kasus

A: atribut

N: jumlah partisi

pi: proporsi dari Si terhadap S

3. Kemudian hitung nilai *gain*, Untuk menghitung *gain* digunakan rumus seperti berikut:

$$Gain\ (S, A) = Entropy\ (s) - \sum_{i}^{m} \frac{|Si|}{|S|} \times Entropy\ (S_i)$$

Keterangan:

S: himpunan dataset

A: kriteria atau atribut

m: jumlah nilai yang mungkin pada kriteria A (jumlah kelas)

Si: himpunan dataset untuk nilai m

|Si|: jumlah dataset untuk nilai m

|S|: jumlah dataset dalam S

4. Untuk memilih kriteria sebagai akar, didasarkan pada nilai *ratio gain* tertinggi dari kriteria kriteria yang ada. *Split information* menyatakan entropy gain informasi potensial. Untuk menghitung *Split Info* dilakukan dengan rumus:

$$Split\ Informations\ (S, A) = -\sum_{i=1}^{n} \frac{|Si|}{|S|} log_2 \frac{|Si|}{S}$$

Setelah menghitung nilai *Entropy*, *Gain*, dan *Split Information* maka selanjutnya adalah menghitung nilai *Gain Ratio.* Untuk menghitung *Gain Ratio* dilakukan dengan rumus:

$$Gain\ Ratio\ (S, A) = \frac{\text{Gain (S, A)}}{\text{Split Info (S, A)}}$$

2.2.8 Validasi

Tahap validasi pada model klasifikasi adalah hal yang paling penting (EL-HABI, 2014). Validasi adalah proses yang membantu untuk mengetahui seberapa baik model dalam hal kesalahan pengujian (Najafi, 2011).

2.2.8.1 K-Fold Cross Validation

Cross Validation adalah teknik validasi dengan membagi data secara acak kedalam k bagian dan masing-masing bagian akan dilakukan proses klasifikasi (Badrul, 2014). Dengan menggunakan cross validation akan dilakukan percobaan sebanyak k. Data yang digunakan dalam percobaan ini adalah data training untuk mencari nilai error rate secara keseluruhan. Secara umum pengujian nilai k dilakukan sebanyak 10 kali untuk memperkirakan akurasi estimasi seperti pada table 2.1 ilustrasi 10-*fold Cross Validation.*

Tabel 0.1 *10-Fold Cross Validation*

	Dataset								
Split 1	**Split 2**	**Split 3**	**Split 4**	**Split 5**	**Split 6**	**Split 7**	**Split 8**	**Split 9**	**Split 10**
				Train					Test
			Train					Test	
		Train					Test		
	Train					Test			
Train					Test				
				Test					Train
			Test					Train	
		Test					Train		
	Test					Train			
Test					Train				

2.2.9 Evaluasi

Evaluasi hasil dari eksperimen merupakan sebuah alat ukur yang dapat digunakan untuk menilai seberapa baik metode. Kriteria utama dalam evaluasi model klasifikasi adalah kriteria akurasi keseluruhan yang diperoleh dari metode validasi model.Salah satu metode paling terkenal dari evaluasi model klasifikasi adalah *Confusion Matrix* (EL-HABI, 2014).

2.2.9.1 Confusion Matrix

Confusion matrix adalah tabel yang dipergunakan sebagai alat ukur yang berguna untuk melakukan analisis seberapa baik pengklasifikasian benar dan salah dari prediksi yang dilakukan dalam kelas-kelas yang berbeda. Tabel ini dipergunakan untuk menentukan kinerja suatu model klasifikasi (Pattipeilohy et al., 2017). Metode ini menggunakan table matrik seperti pada Tabel 2.2.

Tabel 0.2 Confusion Matrix

		Actual Class	
		Positive	*Negative*
Predicted Class	*Positive*	*True Positive (TP)*	*False Positive (FP)*
	Negative	*False Negative (FN)*	*True Negative (TN)*

2.2.10 Accuracy

Menurut Han, Kamber, & Pei, 2012, hasil klasifikasi dapat diuji dengan menggunakan *Confusion Matrix.* Matriks ini terdiri atas jumlah kasus yang diklasifikasikan secara tepat. Metode ini menggunakan level matriks seperti pada tabel 2.2. Untuk mengetahui hasil kinerja dari klasifikasi yang dikonstruksikan dapat diukur menggunakan akurasi. Akurasi dalam klasifikasi adalah presentase

ketetapan record data yang diklasifikasikan secara benar setelah dilakukan pengujian pada hasil klasifikasi (Han, Kamber, & Pei, 2012). Dirumuskan dengan:

$$\text{Accuracy} = \frac{TP+TN}{TP+FN+FP+TN}$$

2.2.11 Precision

Precision adalah jumlah kasus yang diprediksi positif yang juga positif benar pada data yang sebenarnya (Andriani, 2012).

$$\text{Presis (Prec)} = \frac{TP}{TP+FP}$$

2.2.12 Recall

Recall atau *sensitivity* adalah jumlah kasus yang sebenarnya yang diprediksi positif secara benar (Andriani, 2012).

$$\text{Sensitivitas (Sens)} = \frac{TP}{TP+FN}$$

2.2.13 F-Measure

F-measure merupakan perhitungan evaluasi yang mengkombinasikan sensitivitas dan presisi. Nilai sensitivitas dan presisi dapat memiliki bobot yang berbeda pada suatu keaadan. Ukuran yang menampilakn timbal balik sensitivitas dan presisi adalah F-measure yang merupakan bobot dari sensitivitas dan presisi (Espíndola & Ebecken, 2005).

$$\text{F - Measure} = \frac{2\times sens\times prec}{sens+prec}$$

Keterangan:

TP: True Positive FP: False Positive

TN: True Negative FN: False Negative