Studi Perbandingan Prediksi Penyakit Diabetes dengan menggunakan *Logistic Regression* dan *Decision Trees*

Andri Permana Wicaksono [*1], Tessy Badriyah[#2], Achmad Basuki[#3]

[#]*Teknik Informatika dan Multimedia Broadcasting, Politeknik Elektronika Negeri Surabaya Surabaya*

[2]tessy.badriyah@pens.ac.id

[3]basuki@pens.ac.id

[*] *Teknik Informatika dan Komputer, Politeknik Elektronika Negeri Surabaya Jember*

[2]andripermana@pasca.student.pens.ac.id

Abstract

Diabetes Melitus adalah penyakit yang disebabkan oleh tingginya kadar gula dalam darah akibat gangguan sekresi insulin. Penelitian ini membandingkan dua metode dalam teknik Data Mining yaitu *Logistik Regression* dan *Decission Tree* untuk memprediksi tingkat resiko penderita diabetes. Data yang digunakan dalam penelitian ini ada 1450 data pasien yang diambil dari RSD BALUNG JEMBER, dengan pengambilan data mulai 26 september 2014 sampai 30 april 2015. Pengukuran performansi dari kedua metode menggunakan nilai diskriminasi dengan kurva ROC (*Receiver Operating Characteristic*) atau c-index. Pada hasil percobaan menunjukkan bahwa kedua metode LR dan DT memiliki nilai performansi yang sama-sama bagus, dan dalam beberapa kasus Logistik Regression menunjukkan hasil yang lebih baik daripada *Decision Trees*. Akan tetapi kelebihan *Decision Trees* dibandingkan *Logistik Regression* adalah representasi pemodelannya yang intuitif dan mudah dipahami.

Keywords: **diabetes, logistic regression, decission tree**.

I. PENDAHULUAN

Teknik Data mining dapat digunakan untuk sistem pendukung keputusan yang salah satunya dapat digunakan pada bidang kesehatan untuk memprediksi tingkat resiko terhadap suatu penyakit. Metode dalam Data Mining digunakan untuk mengekstraksi dan menemukan pola dari kumpulan informasi yang berharga. Dalam teori data mining terdapat bermacam metode pembelajaran yang dapat digunakan untuk perbandingan dua model data pasien diabetes, sehingga dalam bidang kesehatan dapat digunakan untuk memprediksi suatu penyakit diabetes di Kabupaten Jember, tiap metode pembelajaran memiliki karakteristik model yang berbeda pula.

Dalam penelitian ini digunakan dua metode yaitu metode Logistic Regression dan Decission Tree untuk membedakan karakteristik pemodelan dan mengetahui metode mana yang tepat untuk prediksi penyakit diabetes dengan cara membuat klasifikasi data penderita diabetes menggunakan logistik regression yang diterapkan dengan ROC dan decission tree yang diterapkan dengan penggunaan algoritma C4.5. Dari dua metode tersebut diklasifikasikan dengan cara yang berbeda. pada metode logistik regression diaplikasikan pada sistem berbasis web, sedangkan metode decission tree diaplikasikan menggunakan aplikasi SPSS.

II. METODE

A. Metode Logistik Regression.

Logistik Regresi adalah metode yang paling umum digunakan dalam pendekatan untuk membuat model prediksi probabilitas kejadian suatu peristiwa seperti halnya regresi linear. Logistik Regression ini hanya digunakan jika variabel output dari model yang digunakan didefinisikan sebagai kategori biner. Perbedaannya metode logistik regression ini yaitu memprediksi variabel terikat yang berskala dikotomi. Yang dimaksud dengan skala dikotomi adalah skala nominal

yang mempunyai dua kategori, misalnya: Ya dan Tidak, atau Tinggi dan Rendah. Dalam persamaan rumus, $\boldsymbol{Pb_j}$ adalah probabilitas yang diprediksi dengan cara dikodekan yaitu 1, dan $(\mathbf{1} - \boldsymbol{Pb_j})$adalah probabilitas yang diprediksi keputusan lain dengan cara dikodekan dengan 0.

$$\log\left(\frac{\boldsymbol{Pb_j}}{\mathbf{1}-\boldsymbol{Pb_j}}\right) = \boldsymbol{\alpha} + \boldsymbol{\beta_1}.\boldsymbol{X_{1j}} + \boldsymbol{\beta_2}.\boldsymbol{X_{2j}} + \cdots + \boldsymbol{\beta_n}.\boldsymbol{X_{nj}}$$

Notasi dalam Rumus Logistik, dimana:

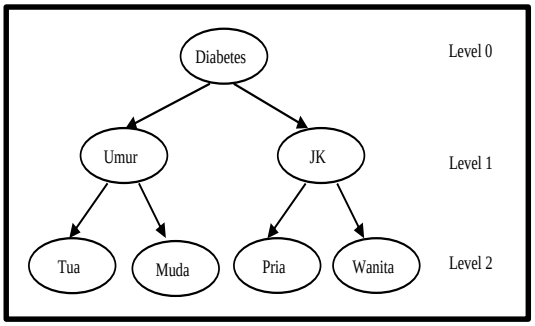
- α adalah Intercept,
- $X_{1j} \ldots X_{nj}$ adalah atribut indepedent dalam catatan –j,
- $\beta_1 \ldots \beta_n$ adalah atribut independent penurunan,
- n adalah jumlah atribut independent,
- j adalah jumlah record dalam dataset.

Pada karakteristik kurva ROC adalah mengukut permodelan yang akan dibuat untuk menentukan probabilitas. Pada kurva ROC terdapat dua sumbu Y yang disebut sebagai True Positif dan sumbu X yang disebut dengan False Positif. Untuk menghitung probabilitasnya menggunakan AUROC (*Area Under the ROC Curve*). Dengan cara menghitung data, predicted dan hasil dari data yang akan di prediksi. Pada AUROC memiliki nilai antara 0,0 sampai 1,0, dikarenakan nilai AUROC akan semakin kuat digunakan untuk klasifikasi.

B. Metode Decision Tree.

Decission Tree adalah salah satu metode klasifikasi data mining yang paling populer, Analisis metode decission tree yang kompleks dengan ketidakpastian dapat membingungkan karena:

1) Jumlah yang besar dan berbeda harus dipertimbangkan ketika membuat keputusan,
2) Kosekuensi tersebut memiliki prediksi yang tidak pasti, dan
3) **bisa berguna menambah data yang tidak pasti**.



Gambar 1. Decission Tree

III. HASIL DAN PEMBAHASAN

Pada percobaan kali ini akan menampilkan histori data pasien yang terkena penyakit diabetes dan tidak terkena penyakit diabetes. Penelitian menggunakan beberapa atribut dalam percobaan meliputi: Jenis Kelamin, Umur, data *Hemogloin*(g/dl), data *White Cell Count*(10^3/ul), data Gula Darah Sewaktu(mg/dl), data *Creatin serum*(mg/dl), Urea(mg/dl), data *Cholesterol total*(gr/dl), data *Trigliserida*(gr/dl), dan untuk atribut diabetes digunakan untuk *variabel input*.

A. *Analisa metode Logistik Regression*

Logistik Regreession merupakan salah satu analisis *multivariate*, yang berguna untuk memprediksi *dependent variabel* berdasarkan *variabel independen*.

1) Data

Pada *logistic regresi*, *dependen variabel* adalah variabel dikotomi (kategori). Ketika kategori *variabel dependennya* berjumlah dua kategori maka digunakan *binary logistic*, dan ketika dependen variabelnya lebih dari dua kategori maka digunakan multinominal *Logistic Regression*. Lalu ketika dependen variabelnya berbentuk ranking, maka disebut dengan *Ordinal Logistic Regression*. Sedangkan pada penelitian ini menggunakan *binary logistik*

2) Konsep Logistik Regression

Untuk menggambarkan bagaimana menghitung *c-indeks*, pada percobaan ini memiliki 1450 catatan data, di mana nilai adalah atribut prediksi dan hasil adalah atribut yang diamati. Skor diskrit dengan sejumlah kisaran dari 0,0 sampai 1,0, Hasil biner dengan 0 dan 1 sebagai nilai-nilai.

TABEL 1.
DATASET SAMPEL MENUNJUKKAN AREA DIBAWAH KURVA ROC.

Record	Predicted	Observed
1	1.0	1
2	1.0	1
3	1.0	1
4	1.0	1
5	1.0	1
6	1.0	1
7	1.0	1
8	1.0	1
9	1.0	1
10	1.0	1

Dari Tabel 1, untuk setiap hasil, dapat memilih skor N dimana nilai prediksinya adalah "1" jika skor lebih besar dari atau sama dengan N (N adalah nilai dari berbagai atribut skor). Maka akan menghasilkan tabel baru, berdasarkan pada Tabel 2 untuk menghitung: *False Positive, False Negative, True Positive, True Negative, sensitivitas* dan *(1-spesifisitas)* seperti yang ditunjukkan pada Tabel 3.

TABEL 2.

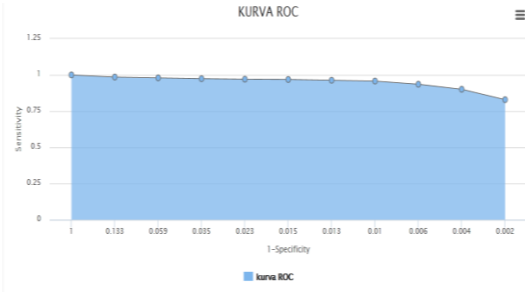
	True Positif	False Positif	True Negative	False Negative
0.0	545	905	0	0
0.1	537	120	785	8
0.2	534	63	852	11
0.3	531	32	873	14
0.4	529	21	884	16
0.5	528	14	891	17
0.6	525	12	893	20
0.7	522	9	896	23

0.8	510	5	900	35
0.9	491	4	901	54
1.0	452	2	903	93

TABEL 3.
Set Titil Dalam Sensitivity dan 1-Spesificity untuk Membentuk Kurva ROC

Sensitivity	Specitity	1-Specitity	ROC
1	0	1	0.861
0.985	0.867	0.133	0.073
0.98	0.941	0.059	0.023
0.974	0.965	0.035	0.012
0.971	0.977	0.023	0.008
0.969	0.985	0.015	0.002
0.963	0.987	0.013	0.003
0.958	0.99	0.01	0.004
0.936	0.994	0.006	0.002
0.901	0.996	0.004	0.002
0.829	0.998	0.002	0.988

Pada tabel 3, diatas, terlihat bahwa nilai *ROC* yang terdapat blok warna kuning menunjukkan jumlah data *ROC* ialah 0,988, data ini hampir mendekati tingkat nilai akurasi yang sempurna yaitu 1.

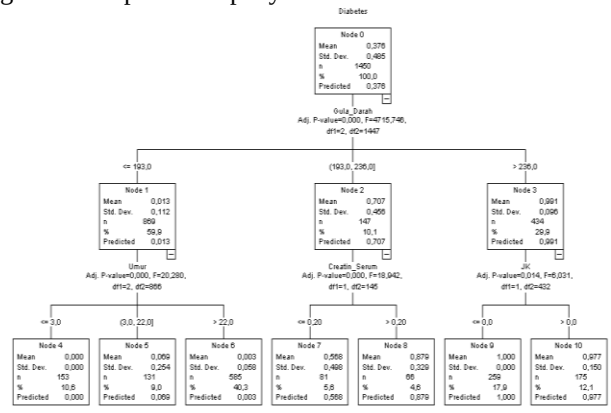


Gambar 2. Kurva *ROC*

Pada gambar 2, diatas, menunjukkan kurva *ROC* yang setiap *sensitivitynya* menunjukkan kurva semakin menurun. Dari hasil kurva tersebut menunjukkan nilai akurasi yang sangat bagus

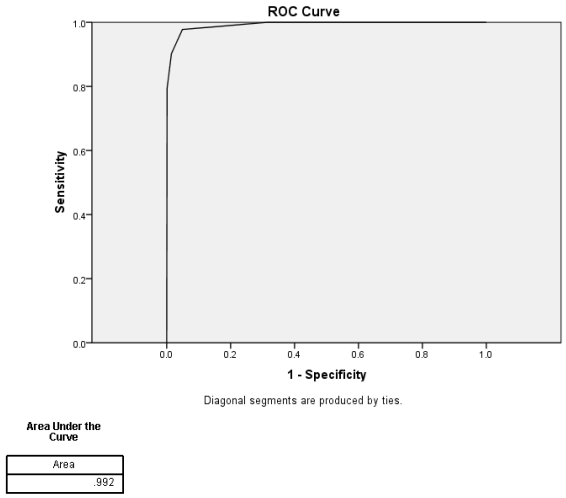
B. *Analisa metode Decission Tree*

Pohon keputusan adalah model prediksi menggunakan struktur pohon untuk melihat keputusan yang mempengaruhi tingkat resiko penderita penyakit diabetes.



Gambar 3. Diagram Decission Tree

Pada gambar 3, diatas, menunjukkan diagram yang paling dominan untuk prediksi diabetes terdapat pada atribut gula darah dan yang kedua terdapat pada atribut umur, *creatin serum*, dan jenis kelamin. Konsep dari pohon keputusan adalah mengubah data menjadi pohon keputusan dan aturan-aturan keputusan.



Gambar 2. Kurva ROC

Hasil representasi pengetahuan yang ada pada pohon keputusan cukup menggambarkan variabel yang paling dominan dalam mempengaruhi tingkat resiko penyakit diabetes yaitu gula darah, sedangkan variabel yang berpengaruh pada tingkat nomer dua salah satunya pada creatin serum.

IV. KESIMPULAN

Penelitian ini membahas tentang perbandingan teknik data mining dengan menggunakan dua metode yang berbeda yaitu metode Logistik Regression dan metode Decission Tree menggunakan nilai diskriminasi dengan kurva ROC (Receiver operating characteristic) atau c-index untuk mengukur performansinya. Pada kedua metode, menghasilkan pengukuran c-index yang bagus dengan nilai lebih besar dari 0.8. Dimana dengan menggunakan dataset yang sama, pada Logistik Regression didapatkan besaran c-index sebesar 0.988, sementara pada Decision Trees didapatkan nilai c-index sebesar 0.992.

Dengan demikian dapat disimpulkan bahwa kedua metode memiliki nilai diskriminasi yang bagus untuk memprediksi penyakit diabetes. Perbedaan yang menonjol pada kedua metode adalah pada bentuk pemodelannya. Dimana pemodelan pada Decision Trees memiliki bentuk pemodelan yang intuitif dan dapat dipahami untuk melihat variabel yang paling berpengaruh pada pengambilan keputusan yaitu gula darah, disusul oleh variabel kedua yang berpengaruh yang salah satunya adalah variabel creatin. Sedangkan pada pemodelan Logistik Regression, pemodelan yang dihasilkan hanya menunjukkan nilai besaran intercept dan slope untuk setiap variabel yang ada yang belum bisa secara intuitif dipahami untuk memahami variabel yang

secara dominan dan berpengaruh dalam penentuan tingkat resiko penyakit diabetes.

UCAPAN TERIMA KASIH

Dengan terselesaikan paper ini, penulis mengucapkan terima kasih yang sedalam-dalamnya kepada :

1. Allah S.W.T atas limpahan karunia dan hidayahnya sehingga penulis dapat menyelesaikan paper ini.
2. Ibu Tessy Badriyah, S.Kom, MT, Ph.D selaku Dosen Pembimbing atas bimbingan, arahan dan koreksinya selama penyusunan dan penulisan paper ini.
3. Bapak Drs. Achmad Basuki, M.Kom, Ph.D selaku Dosen Pembimbing atas bimbingannya, membantu dan mendukung saya dalam mengerjakan paper ini.
4. Kedua Orang Tua saya yang telah mendukung dan mendoakan saya dalam mengerjakan paper ini.

DAFTAR PUSTAKA

[1] Maroco J, Silva D, Rodrigues A, Guerreiro M, Santana I, de Mendonca A. Data mining methods in the prediction of Dementia: a real-data comparison of the accuracy, sensitivity and specificityof linear discriminant analysis, logistic regression, neural networks, support vector machines, classificationtrees and random forests. BMC Res Notes 2011; 4:299

[2] Kim S, Kim W, Park RW. A comparison of intensive care unit mortality prediction models through the use of data mining techniques. Healthc InformRes2011;17(4):232-43.

[3] Han, J. dan Kamber, M., 2006, *"Data mining: Concepts and Techniques (2)", Elsevier Inc*.

[4] Cessie, L. dan Houwelingen, J.C., (1994), Logistic Regression for Correlated Binary Data, *Applied Statistics*, 42, hal. 95-108.

[5] Larose, Daniel T, 2006, "Data Mining Methods and Models" Hoboken New Jersey : Jhon Wiley & Sons, Inc