KLASIFIKASI PENYAKIT DIABETES PADA WANITA MENGGUNAKAN METODE NAÏVE BAYES

M. Subaeri, Triawan A. C, Ilham Saifudin
Jurusan Teknik Informatika Fakultas Teknik
Universitas Muhammdiyah Jember
Muhammadsubaeri2811@gmail.com, triawanac@unmuhjember.ac.id, ilham.saifudin@unmuhjember.ac.id

ABSTRAK

Diabetes melitus (DM) merupakan suatu kelompok penyakit metabolik dengan karakteristik hiperglikemia yang terjadi karena kelainan sekresi insulin, kerja insulin, atau kedua-duanya. *Naïve Bayes* merupakan suatu metode klasifikasi yang menggunakan perhitungan probabilitas. Penentuan kelas dari suatu data pada dataset dilakukan dengan cara membandingkan nilai probabilitas suatu sampel berada di kelas yang satu dengan nilai probabilitas suatu sampel berada di kelas yang lain. Peneliti tertarik mengimplementasikan metode *naive bayes* terhadap klasifikasi penderita diabetes dengan sumber data UCI *Machine Learning*. Penelitian dengan implementasi metode *naive bayes* dalam klasifikasi penderita diabetes dengan data latih berjumlah 584 dan data uji berjumlah 146 dengan skenario pengujian menggunakan *K-Fold Cross Validation* sebanyak 5 kali memiliki hasil tertinggi akurasi 78,1% dan presisi 66,7% pada uji coba ke 3.

Kata Kunci : Penyakit, Diabetes, *Naive bayes*.

1. Latar Belakang

Diabetes mellitus merupakan suatu penyakit yang ditandai oleh kadar glukosa darah melebihi normal dan gangguan
metabolisme karbohidrat, lemak dan protein yang disebabkan kekurangan hormon insulin secara relatif maupun absolut. Bila hal ini dibiarkan tak terkendali dapat terjadi komplikasi metabolik akut maupun komplikasi vaskuler jangka panjang, baik mikroangiopati maupun makroangiopati.

Klasifikasi diabetes melitus mengalami perkembangan dan perubahan dari waktu ke waktu. Dahulu diabetes diklasifikasikan berdasarkan waktu munculnya (time of onset). Diabetes yang muncul sejak masa kanak-kanak disebut "juvenile diabetes", sedangkan yang baru muncul setelah seseorang berumur di atas 45 tahun disebut "adult diabetes". Namun klasifikasi ini sudah tidak layak dipertahankan lagi, sebab banyak sekali kasus-kasus diabetes yang muncul pada usia 20-39 tahun, yang menimbulkan kebingungan untuk mengklasifikasikannya.

Naïve Bayes merupakan suatu metode klasifikasi yang menggunakan perhitungan probabilitas. Penentuan kelas dari suatu dokumen dilakukan dengan cara membandingkan nilai probabilitas suatu sampel berada di kelas yang satu dengan nilai probabilitas suatu sampel berada di kelas yang lain. Metode klasifikasi Naïve Bayes adalah metode pembelajaran Bayesian yang ditemukan sangat berguna dalam berbagai aplikasi. Naïve Bayes merupakan salah satu metode supervised document classification. Metode ini dikenal memiliki tingkat akurasi yang tinggi dengan perhitungan sederhana.

2. Tinjauan Pustaka

a. Diabetes

Diabetes melitus merupakan suatu kelompok penyakit metabolik dengan karakteristik hiperglikemia yang terjadi karena kelainan sekresi insulin, kerja insulin atau kedua-duanya (Henderina, 2010). Menurut PERKENI (2011) seseorang dapat didiagnosa diabetes melitus apabila mempunyai gejala klasik diabetes melitus seperti poliuria, polidipsi dan polifagi disertai dengan kadar gula darah sewaktu ≥200 mg/dl dan gula darah puasa ≥126 mg/dl.

b. Data Mining

Data mining adalah kegiatan menemukan pola yang menarik dari data dalam jumlah besar, data dapat disimpan dalam database, data warehouse, atau penyimpanan informasi lainnya. Data mining berkaitan dengan bidang ilmu – ilmu lain, seperti database system, data warehousing, statistik, machine learning, information retrieval, dan komputasi tingkat tinggi. Selain itu, data mining didukung oleh ilmu lain seperti neural network, pengenalan pola, spatial data analysis, image database, signal processing (Han, et al., 2006).

Data mining adalah kegiatan menemukan pola yang menarik dari data dalam jumlah besar, data dapat disimpan dalam database, data warehouse, atau penyimpanan informasi lainnya. Data mining berkaitan dengan bidang ilmu – ilmu lain, seperti database system, data warehousing, statistik, machine learning, information retrieval, dan komputasi tingkat tinggi. Selain itu, data mining didukung oleh ilmu lain seperti neural network, pengenalan pola, spatial data analysis, image database, signal processing.

c. Naive Bayes

Naïve Bayes merupakan pengklasifikasian dengan metode probabilitas dan statistik yang dikemukakan oleh ilmuwan Inggris Thomas Bayes, yaitu memprediksi peluang dimasa depan berdasarkan pengalaman di masa sebelumnya sehingga dikenal sebagai teorema Bayes. Teorema tersebut dikombinasikan dengan "naive" dimana diasumsikan kondisi antar atribut saling bebas. Pada sebuah dataset, setiap baris/dokumen diasumsikan sebagai vector dari nilai-nilai atribut dimana tiap nilai-nilai menjadi peninjauan atribut (Kusumadewi, 2003).

$$P\ C_i|X = \frac{P\ X|C_i\ P(C_i)}{P(X)}$$

Keterangan :
P(Ci|X) : peluang dokumen X pada kategori Ci.
P(X|Ci) : peluang pada kategori Ci, dimana kata pada dokumen X muncul pada kategori tersebut.
P(Ci) : peluang dari kategori yang diberikan, dibandingkan dengan kategori-kategori lainnya yang dianalisa.
P(X) : peluang dari dokumen tersebut secara spesifik. Pada pengembangannya, P(X) dapat dihilangkan karena nilainya tetap, sehingga saat dibandingkan dengan tiap kategori, nilai ini dapat dihapus.

d. Confusion Matrix

Confusion Matrix memberikan keputusan yang diperoleh dalam traning dan testing, confusion matrix memberikan penilaian performance klasifikasi berdasarkan objek dengan benar atau salah. Confusion matrix berisi informasi aktual (actual) dan prediksi (predicted) pada sistem klasifikasi. (Girja, Bhargava & Mathuria, 2013).
Berikut merupakan persamaan model confusion matrix untuk menghitung akurasi, presisi dan recall.
Akurasi= (TP+TN)/(TP+FP+TN+FN)
Precison= TP/(TP+FP)
Recall= TP/(TP+FN)
Keterangan:
TP = True Positives
TN = True Negatives
FP = False Positives
FN = False Negatives

3. Metodologi Penelitian

1. Tahapan Penelitian

a. Study Literatur
Tahapan studi literatur mempelajari mengenai informasi yang terkait dengan penyakit diabetes serta dataset sebagai bahan penelitian. Dengan mendapatkan hasil yang diperoleh maka dapat merumuskan cara untuk menentukan pemeriksaan lanjut yang terbaik terhadap pasien penderita penyakit diabetes serta mempelajari algoritma Naive bayes.

b. Penyediaan dataset
Dataset diperoleh dari UCI Machine Learning Repository pada laman https://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes dengan atribut jumlah kehamilan, konsentrasi kandungan glukosa, tekanan diastol, Ketebalan Lipatan Kulit Trisep, masa tubuh, Silsilah keturunan diabetes dan umur.

c. Implementasi naive bayes
Pada tahap ini terbagi menjadi beberapa proses yaitu:
- Preprocessing
- Pembagian data latih dan uji
- Pembangunan model
- Hasil model yang didapatkan

d. Pengujian
Pengujian disini menggunakan metode dari confusion matrix. Dengan mengukur tingkat akurasi, presisi dan recall klasifikasi terhadap data asli. Dari hasil ini akan dipertimbangkan atau ditarik kesimpulan.

e. Kesimpulan
Kesimpulan yang dihasilkan diambil dari hasil pengukuran metode ini terhda data pasien diabetes menggunakan confusion matrix. Selanjutnya dapat diketahui apakah metode ini cocok atau tidak sehingga kedepannya dapat diketahui atau dikembangkan menjadi sistem dan sebagainya.

4. Implementasi dan Pengujian

a. Deskripsi data
Tahap ini adalah tahap awal dalam pengolahan data untuk menemukan hasil dalam penelitian. Dataset yang diambil dari Website UCI dengan jumlah 730 data terbagi menjadi 80% data latih dan 20% data uji. Data latih berjumlah 584 record dan data uji berjumlah 146 record. Data awal dari data ini berupa hasil pemeriksaan pada pasien wanita berupa data nominal.

b. Preprocessing
Dari data awal hasil pemeriksaan pasien selanjutnya akan dilakukan pengubahan data yang sebelumnya berbentuk nominal dijadikan kategori. Dalam pengubahan data ini sesuai ketentuan atau dasar yang diberikan UCI sebagai penyedia dataset. Sebagaimana aturan preprocessing yang dijelaskan pada bab tiga penulis mengulas kembali aturan preprocessing pada bab empat ini. Aturan preprocessing dapat dilihat pada tabel dibawah ini.

No.	Atribut	Aturan perubahan	sumber
1	jumlah kehamilan	0 = belum pernah hamil 1= primigravida 2= multigravida	Prawirohardjo, 2008, p.180 dan media.neliti.com/peublications
2	konsentrasi kandungan glukosa	normal=Below 7.8 mmol/l Below 140 mg/dl Prediabetes=7.8 to 11.0 mmol/l 140 to 199 mg/dl Diabetes=11.1 mmol/l or more 200 mg/dl or more	https://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes
3	tekanan darah	<60=rendah 60-90=normal >90=indikasi tinggi	https://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes
4	ketebalan lipatan kulit trisep	<=23=normal >23=terindikasi diabetes	https://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes

5	masa tubuh	18.5-25=normal 26-30=gemuk >30=obesitas	https://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes
6	silsilah keturunan diabetes	<1=cenderung rendah >=1=cenderung tinggi	https://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes
7	Umur	17-25=remaja akhir 26-35=dewasa awal 36-45=dewasa akhir 46-55=lansia awal 56-65=lansia akhir >65= manula	Kemenkes RI

c. Implementasi

Dari penghitungan menggunakan metode naive bayes didapat nilai probabilitas untuk 2 output yaitu diabetes dan tidak disetiap parameter pada atribut. Nilai tersebut yang akan diambil dan diletakkan pada data uji. Berikut adalah model probabilitas dari hasil hitung setiap atribut

Atribut	Parameter	C1	C2
Total keseluruhan		0.666	0.334
jumlah kehamilan	584		
	primigravida	0.26	0.128
	multigravida	0.602	0.703
	belum pernah	0.139	0.169
konsentrasi kandungan glukosa	584		
	normal	0.856	0.482
	prediabetes	0.144	0.518
tekanan darah	584		
	rendah	0.162	0.067
	normal	0.812	0.877
	tinggi	0.026	0.056
ketebalan lipatan kulit trisep	584		
	normal	0.388	0.149
	terindikasi obesitas	0.612	0.851
masa tubuh	584		
	normal	0.175	0.015
	gemuk	0.27	0.149
	obesitas	0.555	0.836
silsilah keturunan diabetes	584		
	cenderung rendah	0.954	0.872
	cenderung tinggi	0.046	0.128
umur	584		
	remaja akhir	0.491	0.195
	dewasa awal	0.296	0.318
	dewasa akhir	0.141	0.272
	manula	0.003	0.005
	lansia awal	0.036	0.154
	lansia akhir	0.033	0.056

d. K-Fold Cross Validation

Dalam penelitian ini penulis menggunakan metode *Cross Fold Validation* dalam mencari model terbaik dengan rata-rata tertinggi dari *confusion matrix* yang dihasilkan. Dengan total data 730, penulis membuat 5 skenario pembagian data. Berikut gambaran skenarionya.

	1 - 146	147 - 292	293 - 438	438 - 584	585 - 730
Skenario 1	Uji	Latih	Latih	Latih	Latih
Skenario 2	Latih	Uji	Latih	Latih	Latih
Skenario 3	Latih	Latih	Uji	Latih	Latih
Skenario 4	Latih	Latih	Latih	Uji	Latih
Skenario 5	Latih	Latih	Latih	Latih	Uji

Dari skenario K-Fold *Cross Validation* diatas terhadap dataset ini menghasilkan nilai akurasi dan presisi. Dari ketiga nilai tersebut akan dirata-rata dan rata-rata tertinggi akan menjadi model yang digunakan. Berikut hasil yang didapat.

	AKURASI	PRESISI	RATA-RATA
Skenario 1	69.9	61.7	65.8
Skenario 2	78.1	59.5	68.8
Skenario 3	78.1	66.7	72.4
Skenario 4	76.7	63.5	70.1
Skenario 5	74.0	61.7	67.8

e. Pengujian

Tahap selanjutnya adalah pengujian dengan data latih. Nilai model yang dihasilkan diatas akan diujikan terhadap 146 data uji. Berikut adalah gambaran data uji.

		C1 (tidak)	C2 (Diabetes)
jumlah kehamilan	belum pernah	0.138817	0.169231
konsentrasi kandungan glukosa	prediabetes	0.144	0.518

tekanan darah	normal	0.812339	0.876923
ketebalan lipatan kulit trisep	terindikasi obesitas	0.611825	0.851282
masa tubuh	obesitas	0.55527	0.835897
silsilah keturunan diabetes	cenderung rendah	0.953728	0.871795
umur	dewasa awal	0.29563	0.317949
PRIOR PROBABILITY		0.666096	0.333904
TOTAL		0.001036	**0.005062**
KLASIFIKASI	diabetes		
hasil	diabetes		
KRITERIA	TP		

f. Pengukuran menggunakan confusion matrix
Dari hasil klasifikasi naive bayes selanjutnya akan dibandingkan dengan hasil data asli.

Kriteria	Jumlah
TP	34
TN	80
FP	17
FN	14

Terakhir, untuk mengetahui tingkat akurasi, presisi dan recall metode naive bayes terhadap data uji ini menggunakan persamaan pada bab 2 poin 2.2, 2.3 dan 2.4. Sehingga diperolehlah hasil seperti dibawah ini.

Akurasi	78,1%
Presisi	66,7%

Dari hasil penghitungan confusion matrix, menunjukkan bahwa akurasi yang diperoleh sebesar 78,1% dan presisi diperoleh sebesar 66,7%.

5. Penutup

a. Kesimpulan
Dari hasil penelitian ini, didapat beberapa hal yang dapat disimpulkan yaitu dari percobaan penelitian klasifikasi penderita diabetes menggunakan metode naive bayes dimana atribut yang digunakan antara lain Glucose, Blood Pressure, Skin Thickness, Body Mass Indeks, Diabetes dan Pedigree Function yang dilakukan, bahwa metode naive bayes bisa digunakan dalam mengklasifikasi penderita diabetes dengan ditunjukkan hasil matrix confusion. Dari perbandingan data latih dan data uji sebesar 80% : 20% dengan jumlah data latih 584 dan data uji 146 serta 5 kali skenario uji coba menggunakan Cross Fold Validation menghasilkan nilai rata-rata paling tinggi akurasi sebesar 78,1% dan presisi sebesar 66,7% pada skenario uji ke 3 dengan rata-rata 72,4%.

b. Saran
Penulis sebagai peneliti disini sadar bahwa penelitian ini jauh dari kesempurnaan, untuk itu penulis membuka lebar bagi pengembang yang ingin mengembangkan penelitian ini agar lebih baik kedepannya. Berikut saran yang dapat diberikan oleh penulis:
1) Pengembang dapat menemukan atribut baru yang lebih modern atau yang lebih update.
2) Pengembang dapat menggunakan metode lain yang lebih cocok.
3) Pengembang dapat membandingkan dengan metode yang lain agar mendapatkan metode yang paling cocok.

Daftar Pustaka

Purnamasari, D. (2009). Diagnosis dan klasifikasi diabetes melitus. Sudoyo, Aru W., Bambang Setyohadi, Idrus Alwi, Marcellus Simadibrata, Siti Setiati. Buku Ajar Ilmu Penyakit Dalam Jilid, 5, 1880-1883.

Darmono, S. T., Pemayun, T. G., & Padmomartono, F. S. (2007). Naskah lengkap diabetes melitus ditinjau dari berbagai aspek penyakit dalam. Semarang: Badan Penerbit Universitas Diponegoro.

Perkeni, P. E. I. (2011). Konsensus Pengendalian dan Pencegahan Diabetes Melitus Tipe 2 di Indonesia (The Consensus of Control and Prevention of Type 2 Diabetes Mellitus). Jakarta: Perkeni (Indonesian Society of Endocrinology).

Budi, S. (2007). Data Mining: "Data Mining Teknik Pemanfaatan Data Untuk Keperluan Bisnis".

Sitompul, O. S. (2008). Data Warehouse dan Data Mining untuk Sistem Pendukung Manajemen. Data Warehouse dan Data Mining untuk Sistem Pendukung Manajemen.

Larose, D. T. (2005). An introduction to data mining. Traduction et adaptation de Thierry Vallaud.

Kulkarni, S., Singh, A., Ramakrishnan, G., & Chakrabarti, S. (2009, June). Collective annotation of Wikipedia entities in web text. In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 457-466). ACM.

Rokach, L., & Maimon, O. Z. (2008). Data mining with decision trees: theory and applications (Vol. 69). World scientific.

Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A., & Huber, W. (2005). BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics, 21(16), 3439-3440.

Chakrabarti, S., Ester, M., Fayyad, U., Gehrke, J., Han, J., Morishita, S., ... & Wang, W. (2006). Data mining curriculum: A proposal (Version 1.0). Intensive Working Group of ACM SIGKDD Curriculum Committee, 140.