MODEL PREDIKSI PENYAKIT DIABETES MENGGUNAKAN BAYESIAN CLASSIFICATION DAN INFORMATION GAIN UNTUK SELEKSI FITUR DAN ADAPTIVE BOOSTING UNTUK PEMBOBOTAN DATA

Ilham Susilo Bakti, Ivandari
STKIP NU Kabupaten Tegal, STMIK Widya Pratama Pekalongan
Email: ilham_susilo@stkipnutegal.ac.id, ivandarialkaromi@gmail.com

ABSTRAK

Diabetes merupakan salah satu penyakit berbahaya, penyakit yang dapat merusak organ tubuh dan memerlukan biaya yang mahal untuk mengobatinya. Mendiagnosis diabetes pada tahap awal sangat penting untuk membantu mencegah resiko diabetes yang lebih parah. Penelitian ini merupakan upaya untuk membantu meningkatkan akurasi dalam memprediksi dan mendiagnosis diabetes dengan menggunakan dataset Pima Indian Diabetes (PID). Hal ini juga merupakan upaya untuk membantu jutaan orang penderita diabetes agar dapat melakukan pencegahan dini. Naive Bayes adalah tekhnik machine learning yang dikenal untuk mengklasifikasi, karena sangat sederhana, efisien dan memiliki performa yang baik. Namun, Naive Bayes memiliki kekurangan ketika digunakan pada data yang memiliki fitur terlalu banyak, menyebabkan akurasi menjadi rendah. Oleh karena itu, dalam penelitian ini menggunakan Information Gain sebagai teknik seleksi fitur dan metode boosting untuk memberikan pembobotan data agar dapat meningkatkan akurasi pengklasifikasi Naive Bayes. Penelitian ini menghasilkan akurasi yang meningkat dalam mendiagnosis penyakit diabetes dibandingkan dengan penelitian sebelumnya. Pengukuran ini juga membandingkan akurasi Naive Bayes sebelum dan sesudah penambahan metode pemilihan fitur dan Adaboost. Validasi dilakukan dengan menggunakan 10 fold cross validation. Sedangkan pengukuran akurasi diukur dengan confusion matrix dan kurva ROC. Hasil penelitian menunjukkan peningkatan akurasi Naive Bayes dari 74.01% menjadi 79.10% dan nilai AUC 0.8722. Peningkatan akurasi dari penelitian sebelumnya yaitu dengan metode Fuzzy Decision Tree dari 75,8% dan pada penelitian ini menjadi 79,10%. Sehingga dapat disimpulkan bahwa integrasi metode Information Gain dan AdaBoost pada Pima Indian Diabetes ini mampu meningkatkan akurasi algoritma Naive Bayes.

Kata kunci: Pima Indian Diabetes (PID), bobotdata, Information Gain, Boosting, Naive Bayes.

1. PENDAHULUAN

Diabetes merupakan salah satu penyakit berbahaya, penyakit yang dapat merusak organ tubuh dan memerlukan biaya yang mahal untuk mengobatinya. Dengan menderita diabetes yang berat akan berdampak serius pada masalah keuangan dan kehidupan sosial para penderita diabetes (Zhang et al. 2009). Penyakit ini ditandai dengan kadar glukosa pada darah yang tinggi dan ada 2 tipe jenis penyakit. Diabetes tipe1 disebut juga *juvenile diabetes* yaitu terjadi gangguan sistem kekebalan tubuh, diabetes tipe 2 adalah yang paling sering terjadi ketika insulin yang dihasilkan pangkreas tidak cukup, atau sebaliknya tubuh yang tidak dapat menggunakan insulin yang dihasilkan dengan baik (States 2014). Meskipun tidak mungkin untuk menyembuhkan penyakit dengan sepenuhnya, tapi jika dicegah dengan baik atau dikendalikan maka orang dapat menjalani hidup dengan sehat.

Mendiagnosis diabetes pada tahap awal sangat penting untuk membantu mencegah resiko diabetes yang lebih parah. Seorang *diabetologist* harus mampu untuk secara kritis menganalisis beberapa faktor yang mempengaruhi untuk mendiagnosa diabetes. Ketidakmampuan untuk memahami data dalam jumlah yang besar dapat menyebabkan diagnosis yang salah. Untuk dapat mengatasi kebutuhan ini kita memerlukan efisiensi dan keefektifan waktu dan biaya dengan kecanggihan teknologi komputasi dan teknik menganalisis data untuk deteksi awal diabetes. Penelitian ini merupakan upaya untuk membantu meningkatkan akurasi dalam memprediksi dan mendiagnosis diabetes dengan menggunakan *dataset* Pima Indian Diabetes (PID), Hal ini juga merupakan upaya untuk membantu beberapa jutaan orang penderita diabetes agar dapat melakukan pencegahan dini.

Beberapa penelitian yang dilakukan untuk mendiagnosis penyakit diabates diantaranya Jia Zhu, Qing Xie, & Kai Zheng (Zhu, Xie, and Zheng

2015), dimana meningkatkan akurasi diagnosis dini untuk mengurangi tingkat *error* data dengan mengkomparasi beberapa metode *Data Mining*. Ada juga penelitian terhadap *dataset* Pima Indian Diabetes (PID) oleh Kamadi V.S.R.P.Varma[a]Allam AppaRao[b]T.Sita Maha Lakshmi[a]P.V.Nageswara Rao[a](Varma et al. 2014) pada tahun 2014, yang menggunakan *fuzzy* pada algoritma decision tree. Dalam penelitian tersebut menggunakan fungsi *fuzzy* untuk meminimalkan *index gini* dan hanya mencapai akurasi sebesar 75.8 %. Dari penelitian ini metode decision tree dan ini akan meningkatkan akurasi jauh lebih baik dibanding dengan penelitian sebelumnya. Penelitian sebelumnya yang menggunakan metode decision tree hanya mencapai akurasi sebesar 65.06 % (Bluma and Langley 1997). Decession Tree memiliki kekurangan dalam mengakumulasi jumlah *error* dalam setiap level pada sebuah data yang besar (Diponegoro and Fatoni 2014). Berbeda dengan Naïve Bayes yang dapat mengatasi masalah Decession tree yaitu dapat menangani data yang besar dengan meminimalisir dalam mengakumulasi jumlah *error*. (Seminar and Aplikasi 2012). Dan dalam Penelitian yang dilakukan pada tahun 2011 dengan menggunakan Naïve Bayes oleh G. Parthiban, A. Rajesh & S.K.Srivatsa [8] memperoleh akurasi sebesar 74 %. Dengan membandingkan antara decision tree dengan Naïve Bayes dapat disimpulkan bahwa dengan meningkatkan metode klasifikasi Naïve Bayes diperkirakan mampu menghasilkan akurasi yang lebih tinggi dibanding dengan fuzzy decision tree.

Naïve Bayes adalah metode klasifikasi yang efektif, paling mudah dan banyak diuji untuk induksi *probabilistik* yang dikenal dengan nama lain *bayesian classifier* [9]. Meskipun paling banyak digunakan tapi Naïve Bayes masih memiliki kelemahan yaitu hasil probabilitas kurang berjalan secara optimal dan sering salah pada attribut. Untuk mengatasi kelemahan Naïve Bayes yaitu dengan cara metode pembobotan data agar akurasi dari Naïve Bayes meningkat [10].

Tujua dasar AdaBoost adalah untuk membentuk klasifikasi yang kuat dengan menggabungkan beberapa klasifikasi yang lemah (Hu and Hu 2005). AdaBoost merupak Algoritma yang paling aktif menghasilkan klasifikasi baru untuk digabungkan menjadi klasifikasi utama (Kim 2000). AdaBoost jika digunakan pada klasifikasi Naïve Bayes akan meningkatkan kerja dari klasifikasi Naïve Bayes dengan cara mengklasifikasikan data yang masuk, ke dalam *class* yang masih tidak seimbang dengan semua atribut yang ada didalam dataset (Korada, Kumar, and Deekshitulu 2012). Kemudian dataset diproses oleh AdaBoost, dengan koefesien untuk mengatur bobot, dengan memberikan bobot yang lebih tinggi pada data yang masih salah dalam klasifikasi. Selanjutnya dilakukan proses dengan mengkombinasikan ke dalam bentuk *committee* dengan menggunakan koefisien untuk memberikan bobot yang berbeda pada data dengan klasifikasi yang berbeda pula (Xindong Wu 2009). Jadi AdaBoost dapat digunakan untuk meningkatkan akurasi Naïve Bayes dengan melakukan pembobotan data, dan ini akan mengatasi masalah pada data yang salah klasifikasi.

Tidak hanya masalah pada hasil probabilitasnya tetapi juga memiliki masalah utama untuk klasifikasi data yang berdimensi tinggi pada sebuah fitur. Hal ini sering terjadi pada data yang memiliki puluhan ribu fitur. Justru mungkin ada fitur yang mengurangi akurasi dari sebuah klasifikasi, dan bahkan akan memperlambat proses klasifikasi [15]. Pemilihan fitur dengan mengurangi jumlah data yang akan dianalisis akan membuat hasil klasifikasi lebih baik, lebih efesien dan efektif. Selain menggunakan cara tersebut juga bisa yaitu dengan cara mengidentifikasi fitur yang sesuai [16].

Untuk mengidentifikasi fitur dan mengurangi jumlah data yang dianalisis membutuhkan dua jenis model pemilihan fitur yaitu wrapper dan filter. Wrapper menggunakan algoritma sebagai fungsi evaluasi dalam klasifikasi akurasinya [15]. Metode filter yang paling unggul adalah information gain, yaitu dengan mungukur banyaknya kehadiran dan ketidakhadiran kata untuk membantu mengklasifikasi keputusan yang tepat pada *class* apapun. Information gain adalah metode filter yang paling baik untuk mengklasifikasi [17].

2. PENELITIAN TERKAIT

Penelitian yang dilakukan oleh Kamadi V.S.R.P.Varma[a]Allam AppaRao[b]T.Sita Maha Lakshmi[a]P.V.Nageswara Rao[a], pada tahun 2014, yang menggunakan *fuzzy* pada algoritma decision tree. Dalam penelitian tersebut menggunakan fungsi *fuzzy* untuk meminimalkan *index gini* dan hanya mencapai akurasi sebesar 75.8 %.

Penelitian yang dilakukan diantaranya Jia Zhu, Qing Xie, & Kai Zheng, dimana meningkatkan akurasi diagnosis dini untuk mengurangi tingkat *error* data dengan mengkomparasi beberapa metode *Data Mining*. Penelitian ini berpendapat bahwa penyebab

spesifik dari penyakit rumit seperti type-2 diabetes mellitus (DMT2) belum diidentifikasi. Namun demikian, banyak peneliti ilmu kedokteran percaya bahwa penyakit komplikasi disebabkan oleh kombinasi genetik, faktor lingkungan, dan gaya hidup. Metode yang diusulkan tidak hanya akurasi lokal dan global tetapi juga keragaman antara pengklasifikasi dan lokal kesalahan generalisasi setiap classifier. Untuk mengevaluasi metode pada dua nyata DMT2 set data dan data set medis lainnya. Hasil menguntungkan menunjukkan bahwa metode menghasilkan akurasi signifikan melebihi pengklasifikasi individu dan metode-motode klasifikasi yang lainnya (Zhu, Xie, and Zheng 2015).

Penelitian oleh Kamer Kayaer and T. Yildirim dikembangkan struktur General Regression Neural Networks (GRNN) untuk mendiagnosis Pima diabetes India , diselidiki . Database Pima Indian Diabetes telah diperiksa dengan struktur jaringan saraf yang lebih kompleks. Hasil yang dicapai oleh penelitian sebelumnya dan hasil dari struktur GRNN dibandingkan dalam makalah ini . Kinerja Perceptron Standar Multilayer (MLP) dan Radial Basis Function (RBF) juga diuji untuk perbandingan karena mereka adalah struktur jaringan saraf yang paling umum dan sering digunakan .

Penelitian Dr.Kumar pada tahun 2012 menggunakan beberapa algorima, yaitu Fuzzy, *Neural network* dan *Case Based* untuk memprediksi penyakit diabetes. Dr.Kumar menyajikan suatu pendekatan diagnosis status diabetes yang menggunakan dua tahap yaitu, Tahap prediksi awal dan tahap prediksi akhir. Tahap awal mengadopsi dua kecerdasan dan pengetahuan teknik teknik komputasi seperti *fuzzy*, *Neural Network* dan *Case Based* sebagai pendekatan awal dan akan diolah kembali pada tahap akhir.

T.Jayalakshmi pada tahun 2010 [23]. Penelitian ini dilatar belakangi masalah pada penyakit diabetes yang terbagi menjadi 2 diabetes militus tipe 1 dan tipe 2, dan ini bertujuan untuk mempermudah pendeteksian dengan menggunakan *machine learning* yang dianggap sangat membantu nantinya. Untuk mengatasi masalah tersebut, jurnal penelitian ini peneliti menggunakan Neural Network dan menyajikan dengan menggunakan duapendekatanyang berbeda.

3. METODE YANG DIUSULKAN

Dengan membandingkan antara penelitian terkait sebelumnya yang dijadikan rujukan penelitian ini decision tree memiliki akurasi lemah dalm memprediksi penyakit diabetes. Dan kekurangan decession tree tersebut dapat diatasi dengan kelebihan naïve bayes. Jadi dapat disimpulkan bahwa dengan meningkatkan metode klasifikasi Naïve Bayes diperkirakan mampu menghasilkan akurasi yang lebih tinggi dibanding dengan fuzzy decision tree.

Naïve Bayes adalah metode klasifikasi yang efektif, paling mudah dan banyak diuji untuk induksi *probabilistik* yang dikenal dengan nama lain *bayesian classifier*. Meskipun paling banyak digunakan tapi Naïve Bayes masih memiliki kelemahan yaitu hasil probabilitas kurang berjalan secara optimal dan sering salah pada attribut. Untuk mengatasi kelemahan Naïve Bayes yaitu dengan metode pembobotan data agar akurasi dari Naïve Bayes meningkat. Masalah tersebut dapat diatasi dengan AdaBoost untuk meningkatkan akurasi Naïve Bayes dengan melakukan pembobotan data, dan akan mengatasi masalah pada data yang salah klasifikasi.

Untuk mengidentifikasi fitur dan mengurangi jumlah data yang dianalisis membutuhkan model pemilihan fitur yaitu filter. Metode filter yang paling unggul adalah information gain, yaitu dengan mungukur banyaknya kehadiran dan ketidakhadiran kata untuk membantu mengklasifikasi keputusan yang benar pada *class* apapun.

Penelitian ini adalah penelitian mengenai diagnosa penyakit diabetes militus dengan menggunakan *Bayesian Classification*. Dengan menggunakan dataset Pima Indian Diabetes (PID) yang diperoleh dari *http://archive.ics.uci.edu/ml*. Dalam penelitian ini klasifikasi Naïve Bayes akan dibantu dengan metode pemilihan fitur information gain, dan dibantu teknik *boosting* yang digunakan yaitu AdaBoost. Hasil penelitian ini akan dibandingkan dengan fazzy decession tree.

3.1. *naïve bayes*

Naïve Bayes atau *Bayesian Classification*, dengan menggunakan Teorema Bayes. Teorema Bayes memiliki kemampuan klasifikasi yang setara dengan kemampuan yang dimiliki klasifikasi *decision tree* dan klasifikasi *neural network*. Berdasarkan beberapa penelitian *Naïve Bayes* memiliki akurasi dan kecepatan yang lebih baik ketika diaplikasikan ke dalam *database* yang memiliki data yang relative lebih banyak atau besar.

Bentuk umum Teorema Bayes:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad \ldots(1)$$

Dalam notasi ini *P*(*A*|*B*) berarti peluang kejadian *A* bila *B* terjadi dan *P*(*B*|*A*) peluang kejadian *B* bila *A* terjadi.

Klasifikasi ini digunakan sebagai pendukung pengambilan keputusan. Pada metode Naïve Bayes dalam mengambil keputusan semua atribute memberikan kontribusinya, dengan memberikan bobot yang sama, pada setiap atributnya dan setiap atribut tidak saling terikat dengan atribut yang lainya. Ketika atribut tidak saling berhubungan, maka nilai probabilitas didapatka dengan rumus berikut.

$$P(x_1, \ldots . x_k|C) = P(x_1|C)x_{\ldots\ldots}x \; P(x_k|C) \quad .(2)$$

Apabila atribut ke-i bukan data diskret tapi bersifat kontinu, maka P(x|C) diestimasi dengan fungsi densitas Gauss.

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{-(x-\mu)2}{2\sigma}} \quad .(3)$$

Pertama, dengan menghitung mean dan standar deviasi pada setiap variable, dengan menggunakan rumus.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_1 \quad .(4)$$

Dan untuk menghitung standar deviasi sendiri menggunakan rumus:

$$s = \sqrt{\frac{n\sum_{i=1}^{n} x_i^2 (\sum_{i=1}^{n} x_1)^2}{n(n-1)}} \quad .(5)$$

s^2= Varian
s = Standar Deviasi
Xi= x ke-i
$\bar{x}$ = (*mean*)
n = sampel

3.2 Seleksi Fitur

Setelah proses *removing* kemudian dilanjutkan dengan seleksi fitur, seleksi fitur yang digunakan adalah filter. Metode filter yang paling unggul adalah information gain, yaitu dengan mungukur banyaknya kehadiran dan ketidakhadiran kata untuk membantu mengklasifikasi keputusan yang tepat pada *class* apapun. Information gain adalah metode filter yang paling baik untuk mengklasifikasi [17].

Tahapan yang dilakukan Information Gain dalam pemilihan fitur, yaitu:

1. Setiap atribut dihitung nilai information gainya dalam dataset original.
2. Menentukan batas (*treshold*). Yaitu dengan mempertahankan atribut yang memiliki bobot yang sama dengan batas atau lebih besar dan akan membuang atribut yang barada dibawah batas
3. Mengurangi atribut untuk mempebaiki dataset.

3.3 *Ada Boost*

AdaBoost (*Adaptive Boosting)* ilmu algoritma yang ditemukan oleh Y Freund, RE Schapire (Wu et al. 2007). Ini adalah suatu algoritma dan dapat digunakan bersama dengan algoritma lain untuk meningkatkan kinerja algoritma tersebut. Ada beberapa Algoritma yang memiliki kemampuan klasifikasi lemah seperti *Decision Trees*, *Bayesian Network*, *Random Forests*, dan lain-lain (Korada, Kumar, and Deekshitulu 2012). *Boosting* adalah metode yang digunakan pengklasifikasi untuk meningkatkan akurasinya. Metode pengklasiikasi digunakan sebuah sub rutin untuk membangun sebuah pengklasifikasi yang akurat. Salah satu ide utama algoritma AdaBoost adalah menjaga distribusi atau set bobot (Wang 2012).

Pada data yang memilik permasalahan pada klasifikasi, dimana input vector *x1*, *x2*, sampai dengan, *xN* dan mempunyai target *t1*,*t2*, sampai dengan *tN* dimana *tn* $\in$ *{-1,1}*. Setiap data akan diberika parameter bobot *Wn* (nilai awal *Wn* adalah *1/N* untuk semua data). Pada proses proses pengklasifikasi awal menggunakan bobot data untuk fungsi *y(x)* $\in$ *{-1,1}*. Pada setiap tahapan algoritma, AdaBoost memproses klasifikasi dengan koefesien untuk mengatur bobot, dengan memberikan bobot yang lebih tinggi pada data yang masih salah dalam klasifikasi. Selanjutnya dilakukan proses dengan mengkombinasikan ke dalam bentuk *committee* dengan menggunakan koefisien untuk memberikan bobot yang berbeda pada data dengan klasifikasi yang berbeda pula.

Metode AdaBoost:

Inisialisasi bobot data {*Wn*} dengan *Wn (m)* = 1/N untuk n = 1, 2, …, N.

1. *Form m = 1, ..., M:*
   a. *Training Ym(x)* dengan membuang fungsi yang salah (*error function*) sebagai berikut:

$$J_m = \sum_{n=1}^{N} W_n^{(m)} I(Y_m\ (X_n) \neq t_n)\ .(9)$$

b. Evaluasi kesalahan

$$\varepsilon_m = \frac{\sum_{n=1}^{N} w_n^{(m)} I(y_m(x_n) \neq t_n}{\sum_{n=1}^{N} w_n^{(m)}} \quad \ldots(10)$$

Dan kemudian digunakan evaluasi

$$\alpha_m = In\left\{\frac{1-\varepsilon_m}{\varepsilon_m}\right\} \quad .(11)$$

c. Memperbaiki (*update*) bobot data

$$w_n^{(m+1)} = w_n^{(m)} exp\{\alpha_m\ I\ (y_m(x_n) \neq t_n)\} \quad .(12)$$

2. Membuat prediksi menggunakan model terakhir sebagai berikut

$$Y_m(X) = sign\left(\sum_{m=1}^{M} \alpha_m\, y_m(x)\right) \quad .(13)$$

4. HASIL PENELITIAN

Penelitian ini menggunakan Rapidminer dan dalam penelitianya melakukan beberapa langkah, sebelum dataset diproses ke langkah selanjutnya dataset diproses terlebih dahulu agar bisa diklasifikasi, berikut adalah tahapan prosesnya:

4.1. Pengumpulan Data

Data yang digunakan adalah dataset Pima Indian Diabetes (PID) yang diperoleh dari *http://archive.ics.uci.edu/ml*, berisi data sekitar 768 *data record.* Data awal berupa file Notepad dan kemudian dirubah dalam bentuk Excel. Yang data tersebut terdiri dari 9 atribut diantanya Usia dalam kandungan, kandungan gula dalm kulit selama 2 jam (mg / dl), tekanan darah diastolik- (mm Hg), ketebalan lipatan kulit - (mm), insulin Serum 2 jam- (mu U / ml), berat badan- (kg / m2), Riwaya keturunan Diabetes, Umur, dan hasil uji dengan label 1 Positif (terkena diabetes) dan label 0 Negatif: (tidak menderita diabetes).

4.2. Pengolahan Awal

Dataset yang didapatkan kemudian dilakukan tahap awal dengan melakukan preprocessing untuk mendapatkan data yang lebih akurat, cara yang digunakan adalah *data cleaning*, yaitu dengan proses (*removing)* atau membuang data yang memiliki atribut kosong/*missing value*. Dari dataset penyakit diabetes tersebut terdapat sebanyak 768 *record*, setelah dialakukan proses proses awal didapatkan data yang memiliki atribut tidak lengkap berjumlah 237 *record* dan data yang memiliki atribut lengkap sejumlah 531 *record*. Kemudian data tersebut dipisahkan dan digunakan dalam proses penelitian yang terdiri dari 268 *record* berisi data yang memiliki label positif terkena diabetes dan 263 *record* berisi data yang memiliki label negatif.

4.3. Information Gain

Dalam perhitungan *information gain* untuk dataset PID dibutuhkan beberapa perhitungan. Dalam data ini atribut label adalah registrasi memiliki nilai m adalah 2 yaitu label 1 *(diabetic)* berarti positif terkena diabetes dan label 0 *(non diabetic)* yang tidak terkena diabetes. Jumlah *record* dalam data ini adalah sebanyak 531 sehingga nilai s=531. Dengan rincian 268 record berlabel positif dan 263 record berlabel negatif. Dengan demikian nilai s1=268 salah satu data sampel dari S didalam kelas C1 serta s2=263 salah data sampel dari S di kelas C2. Berdasarkan data tersebut informasi gain dari kelas tersebut adalah:

$$I(s_1, s_2) = -\frac{268}{531} log2 \frac{268}{531} - \frac{263}{531} log2 \frac{263}{531} = 0{,}999936$$

Misalkan atribut yang akan dihitung adalah *Number of times pregnant* dengan hanya memiliki 17 nilai berbeda yaitu 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, dan 17. Proses selanjutnya kemudian atribut tersebut digunakan untuk mengelompokan S kedalam 18 subset. Sehingga S=S11+S12+……+S181+S182 = 531 sampel. Jika Sj merupakan jumlah sampel pada masing masing subset *Number of times pregnant* maka informasi harapan dari subset S1 Sakit dan S1 Sehat adalah sebagai berikut:

$$I(s_{11}, s_{12}) = -\frac{38}{75} log2 \frac{38}{75} - \frac{37}{75} log2 \frac{37}{75} = 0{,}999872$$

Dan informasi harapan dihitung sampai dengan S18. Dengan rincian seperti tabel 43 sebagai berikut:

Tabel 3 Informasi Harapan

S	Informasi Harapan
S1	0.999872
S2	0.857559
S3	0.838008
…	…..

S16	0.000000
S17	0.000000
S18	0.000000

Kemudian nilai *entropy* dari atribut *Number of times pregnant* dapat dihitung dari informasi harapan berdasarkan pemisahan kedalam subset berikut:

$$E(\mathrm{A}) = \frac{38+37}{531}0.999872 + \frac{29+74}{531}0.857559 + \ldots\ldots + \frac{1+0}{531}0$$

$$E(\mathrm{A}) = 0.877296$$

Dengan demikian nilai *information gain* dari atribut *Number of times pregnant* (A1) adalah:

Gain (A1) = |I (S1,S2) – *E* (A1)|
= |0, 0,999936 - 0.877296|
= **0,12264**

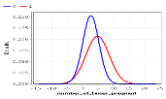
Cara yang sama dilakukan untuk menghitung nilai *information gain* untuk semua atribut yang ada. Kemudian dari nilai *information gain* yang telah diketahui maka atribut diurutkan dari nilai *information gain* tertinggi sampai dengan atribut dengan nilai *information gain* terendah.

Tebel 4 Nilai *information gain*

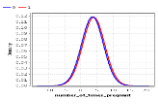
Atribute	Weight
2_hours_serum_insulin	1
Triceps_skin_fold_thickness	0,557
Age	0,419
2_hour_plasma_glucose_concentration_in_oral_glucose	0,415
number_of_times_pregnant	0,192
Body_mass_index	0,166
diastolic_blood_pressure	0,056
Diabetes_pedigree function	0

4.4. Bosting

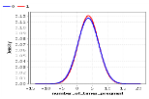
AdaBoost ini bekerja untuk menghasilkan nilai mean dan standar deviasi, yang bersumber dari nilai probabilitas yang dihasilkan. Dari hasil penghitungan pembobotan yang kemudian dihitung tingkat error sebuah data dan kemudian akan diperbaiki sampai mendapatkan nilai mean dan standar deviasi yang optimal dan dapat dilihat pada klasifikasi.



Density number of times pregnant pembobotan ke -1



Density number of times pregnant pembobotan ke -2



Density number of times pregnant pembobotan ke -3

4.5. Klasifikasi

Apabila atribut ke-i bukan data diskret tapi bersifat kontinu, maka P(x|C) diestimasi dengan fungsi densitas Gauss. Berdasarkan dataset PID, kemudian dibangun sebuah system untuk menentukan label 1 *(diabetic)* yang artinya positif terkena diabetes dan label 0 *(non diabetic)* yang berarti negatif. Apabila diterpakan pada data yang baru, untuk mendapatkan label dari data tersebut dengan cara berikut

Pertama, penghitungan mean dan standar deviasi pada variable yang dihasilkan, seperti yang sudah dijelaskan di sebelumnya dan karena data ini bersifat diskret. Maka mengunakan diestimasi dengan fungsi densitas Gauss. Data ini digunakan untuk menghitung probabiltas dan prediksi pada setiap *record* yang ada. *Kedua,* hitung probabilitas setiap kategori apabila kita memprediksi salah satu *record* yang ada didalam dataset. Misalkan mengambil data dari *record* positif seperti tabel 5 sebagai berikut:

Tabel 5 Sampel dataset penghitungan probabilitas Naïve Bayes

R	A1	A2	A3	A4	A5	A6	A7	A8	L
1	1	89	66	23	94	28.1	0.167	21	0

2	1	103	30	38	83	43.3	0.183	33	0
3	3	126	88	41	235	39.3	0.704	27	0
429	1	119	86	39	220	45.6	0.808	29	1
430	0	107	62	30	74	36.6	0.757	25	1
431	2	128	78	37	182	43.3	1.224	31	1

Menghitung probabilitas menggunakan persamaan (5) untuk keperluan tersebut. Untuk menghitung dokumen 431 dengan label positif bagaiamana mengetahui prediksi dari Naïve Bayes dengan seleksi fitur information gain dan AdaBoost. Dengan atribut A1= 2, A2= 128, A3=78, A4= 37, A5=182, A6=43.3, A7=1.224 dan A8=31 dengan label 1(Diabitic). Maka berdasarkan persamaan (3).

f (A1=2|Label=0)

$$= \frac{1}{\sqrt{2\,\pi(3.412)}} e^{\frac{-(2-3.687)2}{2(3.412)}} = 1.90$$

f (A1=2|Label=1)

$$= \frac{1}{\sqrt{2\,\pi(3.228)}} e^{\frac{-(2-3.739)2}{2(3.288)}} = 2.01$$

f (A2=128|Label=0)

$$= \frac{1}{\sqrt{2\,\pi(30.347)}} e^{\frac{-(128-120.630)2}{2(30.347)}} = 1.48x10^{1}$$

f (A2=128|Label=1)

$$= \frac{1}{\sqrt{2\,\pi(32.006)}} e^{\frac{-(128-129.035)2}{2(32.006)}} = 1.05$$

f (A3=78|Label=0)

$$= \frac{1}{\sqrt{2\,\pi(14.543)}} e^{\frac{-(78-70.919)2}{2(14.543)}} = 4.901x10^{1}$$

f (A3=78|Label=1)

$$= \frac{1}{\sqrt{2\,\pi(15.916)}} e^{\frac{-(78-70.801)2}{2(15.916)}} = 4.249x10^{1}$$

f (A4=37|Label=0)

$$= \frac{1}{\sqrt{2\,\pi(11.204)}} e^{\frac{-(37-29.267)2}{2(11.204)}} = 2.9206x10^{2}$$

f (A4=37|Label=1)

$$= \frac{1}{\sqrt{2\,\pi(14.393)}} e^{\frac{-(37-26.795)2}{2(14.393)}} = 3.460x10^{2}$$

f (A5=182|Label=0)

$$= \frac{1}{\sqrt{2\,\pi(126.819)}} e^{\frac{-(182-150.956)2}{2(126.819)}} = 3.253x10^{5}$$

f (A5=182|Label=1)

$$= \frac{1}{\sqrt{2\,\pi(127.197)}} e^{\frac{-(182-117.039)2}{2(127.197)}} = 1.192x10^{24}$$

f (A6=43.3|Label=0)

$$= \frac{1}{\sqrt{2\,\pi(7.33)}} e^{\frac{-(43.3-33.349)2}{2(7.33)}} = 4.149x10^{5}$$

f (A6=43.3|Label=1)

$$= \frac{1}{\sqrt{2\,\pi(6.422)}} e^{\frac{-(43.3-34.507)2}{2(6.422)}} = 6.822x10^{4}$$

f (A7=1.224|Label=0)

$$= \frac{1}{\sqrt{2\,\pi(0.342)}} e^{\frac{-(1.224-0.515)2}{2(0.342)}} = 1.32$$

f (A7=1.224|Label=1)

$$= \frac{1}{\sqrt{2\,\pi(0.341)}} e^{\frac{-(1.224-0.542)2}{2(0.341)}} = 1.30$$

f (A8=31|Label=0)

$$= \frac{1}{\sqrt{2\,\pi(11.499)}} e^{\frac{-(31-32.496)2}{2(11.499)}} = 1.23$$

f (A8=31|Label=1)

$$= \frac{1}{\sqrt{2\,\pi(9.776)}} e^{\frac{-(31-33.083)2}{2(9.776)}} = 1.58$$

Dengan lebih detail nilai frekuensi relative yang dihasilkan dari record 431 dengan 8 atribut dan 2 label bisa dilihat pada tabel 6 berikut ini.

Tabel 6 hasil dari *record* 431.

(f)	Label 0	Label 1
A1	1.90	2.01
A2	$1.48x10^{1}$	1.05
A3	$4.901x10^{1}$	$4.249x10^{1}$
A4	$2.9206x10^{2}$	$3.460x10^{2}$
A5	$3.253x10^{5}$	$1.192x10^{24}$
A6	$4.149x10^{5}$	$6.822x10^{4}$
A7	1.32	1.30
A8	1.23	1.58

Sehingga:

Likehoold 0 (Non Diabetic) $= 6.263\ x\ 10^{16}$

Likehoold 1 (Diabetic) $= 5.136\ x\ 10^{34}$

Dengan melakukan normalisasi terhadap likehood dapat dihitung nilai probabilitas sehingga jumlah yang diperoleh =1.

Probabilitas 0 (Non Diabetic) =

$$\frac{6.263\ x\ 10^{16}}{6.263\ x\ 10^{16}+5.136\ x\ 10^{34}} = 0.00$$

Probabilitas 1 (Diabetic) =

$$\frac{5.136\ x\ 10^{34}}{6.263\ x\ 10^{16}+5.136\ x\ 10^{34}} = 1.00$$

Dari nilai probabilitas tertinggi ada pada label 1 (Diabetic) dan dapat disimpulkan record 431 didiagnosis terkena penyakit diabetes. Dan hasilnya sesuai dengan data yang sebenarnya.

5. HASIL PENELITIAN

5.1. Confunsion Matrix.

Hasil pengukuran dengan menggunakan *confusion matrix* menampilkan perbandingan dari hasil akurasi yang hanya menggunakan klasifikasi model Naïve Bayes saja dan dibandingkan dengan model Naïve Bayes yang sudah ditambahkan Information Gain dan metode AdaBoost. Untuk hasilnya bias dilihat pada Tabel 7 dan tabel 8.

Tabel 7 Confution Matrix naïve bayes.

Accuracy: 74.01 % +/- 5.29 % (mikro: 74.01 %)			
	True 0	True 1	CP
Pred. 0	287	65	81,53 %
Pred. 1	74	113	60,42 %
Cr	79,50 %	63,48 %	

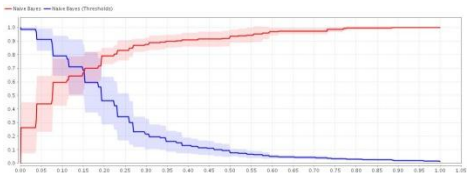
Tabel 8 Confution Matrix naïve bayes, information gain dan AdaBoost.

Accuracy: 79.10 % +/- 4.24 % (mikro: 79.10 %)			
	True 0	True 1	CP
Pred. 0	218	66	76,76 %
Pred. 1	45	202	81,78 %
Cr	82,89 %	75,37 %	

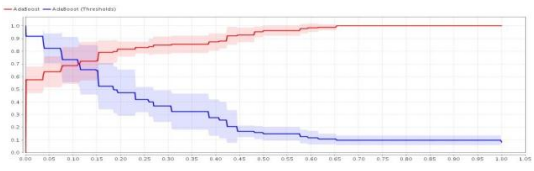
$$accuracy = \frac{202 + 218}{218 + 45 + 66 + 202} = 79{,}10\ \%$$

Dari tabel *Confusion matrix* di atas menunjukan algoritma Naïve Bayes hanya memeliki akurasi 74.01% sedangkan yang sudah ditambahkan Information Gain dan metode AdaBoost akurasinya menjadi 79.10%, akurasi naik 5,09 % dari sebelumnya.

5.2 Kurva ROC (Reciver Operating Characteristic)



Kurva Roc Untuk Model Naive Bayes



Kurva Roc Untuk Model Naive Bayes, Information Gain Dan AdaBoost

Kurva ROC (Reciver Operating Characteristic) diatas menunjukkan algoritma Naïve Bayes memiliki nilai AUC sebesar 0.66 yang artinya *Poor classification* (Rendah) jika algoritma Naïve Bayes yang menggunakan Information Gain dan metode AdaBoost yang meningkatkan nilai AUC yang memiliki nilai AUC 0.72 yang berarti *Fair classification.*

Dalam hasil penelitian, menunjukkan meningkatnya akurasi Naïve Bayes dan meningkatnya akurasi prediksi penyakit diabetes pada dataset PID. Dengan menggunakan algoritma Naïve Bayes dan Information Gain dan metode AdaBoost mempunyai akurasi 79.10%. Akurasi naik 5.09 % dari sebelumnya.

Tabel 9 perbandingan akurasi penelitian terkait

No	*Method*	*Author*	*Accuracy (%)*
1	K-NN	Ster.Dobnikar[18]	71,9
2	MLP	Ster.Dobnikar[18]	75,2
3	SGFDT	Hongze Qie, Haitang Zhang[19]	74,09
4	RBT	Kayaer. Yildirim[20]	68,23
5	Naïve Bayes	Friedman[21]	74,5
6	C 4.5	J.R. Quinlan[5]	65,06
7	Fuzzy Decision Tree	Kamadi V.S.R.P. Varma, [4]	75,8
8	Information Gain+AdaBoost & Naïve Bayes		79,1

Dari penelitian yang dilakukan dalam mengolah data dengan menggunakan AdaBoost dan pemilihan fitur yaitu Information Gain, dapat meningkatkan akurasi klasifikasi Naïve Bayes pada dataset PID.

6. KESIMPULAN

Untuk meningkatkan akurasi algoritma decession tree dalam diagnosa penyakit diabetes dengan dataset PID, salah satu klasifikasi yang digunakan yaitu klasifikasi Naive Bayes. Hal ini dikarenakan Naïve Bayes dapat mengatasi kelemahan yang dimiliki dcession tree.

Dari penelitian yang dilakukan dalam mengolah data dengan menggunakan AdaBoost dan Information Gain untuk pemilihan fitur, dapat meningkatkan akurasi dari klasifikasi Naïve Bayes pada dataset PID ke dalam bentuk positif dan negatif. Akurasi Decession Tree dengan

menambahkan *Fuzzy* mencapai 75.8 %. Sedangkan akurasi Naïve Bayes dengan menggunakan metode boosting AdaBoost dan Information Gain, meningkatkan akurasi mencapai 79.10%. Peningkatan sebesar 4.2%.

Maka dari itu, terbukti bahwa Naïve Bayes dengan menambahkan Information Gain dan AdaBoost, meningkatkan akurasi pada prediksi penyakit diabetes.

Daftar Pustaka

Bluma, Avrim L, and Pat Langley. 1997. "Artificial Intelligence Selection of Relevant Features and Examples in Machine." *Articial Intelligence* 97 (97): 245–71.

Chen, Jingnian, Houkuan Huang, Shengfeng Tian, and Youli Qu. 2009. "Feature Selection for Text Classification with Na??ve Bayes." *Expert Systems with Applications* 36 (3 PART 1). Elsevier Ltd: 5432–35. doi:10.1016/j.eswa.2008.06.054.

Diponegoro, Universitas, and Ahmad Fatoni. 2014. "Implementasi Model Pohon Kepututusan Untuk Mengklasifikasi Masa Studi Mahasiswa Menggunakan Algoritma C4 . 5."

Hu, Wei, and Weiming Hu. 2005. "Network-Based Intrusion Detection Using Adaboost Algorithm." *IEEE/WIC/ACM International Conference on Web Intelligence*.

Jayalakshmi, T., and a. Santhakumaran. 2010. "A Novel Classification Method for Diagnosis of Diabetes Mellitus Using Artificial Neural Networks." *2010 International Conference on Data Storage and Data Engineering*, February. Ieee, 159–63. doi:10.1109/DSDE.2010.58.

Kim, Yu-hwan. 2000. "Text Filtering by Boosting Naive Bayes Classifiers." *Artificial Intelligence*, 168–75.

Korada, Naveen Kumar, N Sagar Pavan Kumar, and Y V N H Deekshitulu. 2012. "Implementation of Naive Bayesian Classifier and Ada-Boost Algorithm Using Maize Expert System." *International Journal of Information Sciences and Techniques (IJIST)* 2 (3): 63–75.

Langley flamingostan ford edu, P a T Langley, Stephanie Sage, S a G E Flamingo, and Stanford Edu. 1993. "Institute for the Study of Learning and Expertise 2451 High Street, Palo Alto, CA 94301," no. 1990: 399–406.

Moraes, Rodrigo, João Francisco Valiati, and Wilson P. Gavião Neto. 2013. "Document-Level Sentiment Classification: An Empirical Comparison between SVM and ANN." *Expert Systems with Applications* 40 (2). Elsevier Ltd: 621–33. doi:10.1016/j.eswa.2012.07.059.

Parthiban, G, and C Abdul Hakkeem College. 2011. "Diagnosis of Heart Disease for Diabetic Patients Using Naive Bayes Method." *International Journal of Computer Applications* 24 (3): 7–11.

Seminar, Prosiding, and Nasional Aplikasi. 2012. "Klasifikasi Teks Dengan Naïve Bayes Classifier (Nbc) Untuk Pengelompokan Teks Berita Dan Abstract Akademis," no. 2011: 269–77.

States, United. 2014. "National Diabetes Statistics Report , 2014 : Data Sources , Methods , and References for Estimates of Diabetes and Its Burden in the United States." *National Diabetes Statistic Report*.

Uysal, Alper Kursat, and Serkan Gunal. 2012. "A Novel Probabilistic Feature Selection Method for Text Classification." *Knowledge-Based Systems* 36 (DECEMBER): 226–35. doi:10.1016/j.knosys.2012.06.005.

Varma, Kamadi V.S.R.P., Allam Appa Rao, T. Sita Maha Lakshmi, and P.V. Nageswara Rao. 2014. "A Computational Intelligence Approach for a Better Diagnosis of Diabetic Patients." *Computers & Electrical Engineering* 40 (5). Elsevier Ltd: 1758–65. doi:10.1016/j.compeleceng.2013.07.003.

Wang, Ruihu. 2012. "AdaBoost for Feature Selection, Classification and Its Relation with SVM, A Review." *Physics Procedia* 25. Elsevier Srl: 800–807. doi:10.1016/j.phpro.2012.03.160.

Wu, Xindong, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, et al. 2007. *Top 10 Algorithms in Data Mining*. *Knowledge and Information Systems*. Vol. 14. doi:10.1007/s10115-007-0114-2.

Xindong Wu, Vipin Kumar. 2009. *The Top Ten Algorithm in Data Mining*.

Zaidi, Na, and J Cerquides. 2013. "Alleviating Naive Bayes Attribute Independence Assumption by Attribute Weighting." *Machine Learning Research* 14: 1947–88.

Zhang, Yiduo, D Ph, Timothy M Dall, Sarah E Mann, Yaozhu Chen, Jaana Martin, Victoria Moore, Alan Baldwin, Viviana A Reidel, and William W Quick. 2009. "The Economic Costs of Undiagnosed Diabetes." *The*

Economic Cost Of Undiagnosed Diabetes 12 (2).

Zhu, Jia, Qing Xie, and Kai Zheng. 2015. "An Improved Early Detection Method of Type-2 Diabetes Mellitus Using Multiple Classifier System." *Information Sciences* 292 (January). Elsevier Inc.: 1–14. doi:10.1016/j.ins.2014.08.056.