InfoTekJar Jurnal Nasional Informatika dan Teknologi Jaringan
ISSN 2540-7600 (Online)
ISSN 2540-7597 (Print)

KOMPARASI AKURASI METODE *CORRELATED NAIVE BAYES CLASSIFIER* DAN *NAIVE BAYES CLASSIFIER* UNTUK DIAGNOSIS PENYAKIT DIABETES

Hairani, Gibran Satya Nugraha, Mokhammad Nurkholis Abdillah, Muhammad Innuddin

STMIK Bumigora Mataram
Jl. Ismail Marzuki, Cilinaya, Cakranegara, Kota Mataram, Nusa Tenggara Bar. 83127, Indonesia
hairani@stmikbumigora.ac.id, gibransn@stmikbumigora.ac.id, nurkholisabdillah@stmikbumigora.ac.id, inn@stmikbumigora.ac.id

***Abstrak*— Penyakit diabetes merupakan salah satu penyakit paling banyak diderita oleh manusia seluruh dunia. Setiap tahun terjadi peningkatan kematian yang disebabkan oleh penyakit diabetes. Penyakit diabetes terjadi disebabkan oleh tubuh tidak menghasilkan insulin dalam jumlah yang cukup. Salah satu cara yang digunakan untuk mengurangi jumlah kematian yang disebabkan oleh penyakit diabetes adalah melakukan diagnosis secara dini. Salah satu teknik yang bisa digunakan adalah memanfaatkan teknik data mining. Untuk melakukan diagnosis penyakit diabetes dibutuhkan suatu metode yang memiliki akurasi terbaik. Pada penelitian ini melakukan komparasi metode Correlated-Naive Bayes Classifier dan Naive Bayes Classifier untuk mendapatkan akurasi terbaik sehingga dapat digunakan untuk diagnosis penyakit diabetes. Berdasarkan pengujian yang telah dilakukan menunjukkan bahwa metode *Correlated Naive Bayes Classifier* (CNBC) memperoleh akurasi terbaik dibandingkan dengan metode *Naive Bayes Classifier* (NBC) untuk *Dataset Pima indian Diabetes*. Tingkat akurasi metode *Correlated Naive Bayes Classifier* (CNBC) sebesar 67,15%, sedangkan metode *Naive Bayes Classifier* (NBC) sebesar 64,33%. Metode *Correlated Naive Bayes Classifier* (C-NBC) memiliki akurasi lebih tinggi dibandingkan metode *Naïve Bayes Classifier* (NBC) karena pada metode Correlated Naïve Bayes Classifier memperhitungkan nilai korelasi dari masing-masing atribut dataset terhadap Kelasnya. Dengan demikian penggunaan metode *Correlated Naïve Bayes Classifier* (C-NBC) dapat digunakan untuk melakukan diagnosis penyakit diabetes karena memiliki tingkat akurasi yang bagus dibandigkan metode *Naive Bayes Classifier*.**

***Keywords*—data mining; correlated naive bayes classifier; naïve bayes classifier; diabetes.**

I. PENDAHULUAN

Penyakit diabetes merupakan salah satu penyakit paling banyak diderita oleh manusia seluruh dunia. Menurut WHO (*Word Health Organization*) melaporkan bahwa penderita penyakit diabetes didunia mendekati jumlah 350 juta orang. Pada tahun 2012 dilaporkan sekitar 1,5 juta kematian disebabkan oleh penyakit diabetes, lebih dari 80% dari jumlah kematian tersebut terjadi di negara-negara berkembang [1]. WHO memprediksikan bahwa tahun 2030 penyakit diabetes menjadi salah satu dari 7 faktor penyebab utama terjadinya kematian didunia.

Selain itu bahaya yang ditimbulkan penyakit diabetes adalah kebutaan, amputasi, dan gagal ginjal diakibatkan kurangnya kesadaran masyarakat dunia tentang bahaya penyakit diabetes [2]. Penyakit diabetes disebabkan oleh pankreas tidak bisa memproduksi insulin yang cukup sehingga menyebabkan peningkatan produksi glukosa dalam darah (hiperglikemia). Insulin merupakan sebuah hormon berfungsi mengatur gula darah dan bertanggung jawan untuk menghasilkan energi yang dibutuhkan oleh manusia. Penyakit diabetes sendiri dibagi menjadi 2 jenis yaitu diabetes tipe 1 dan tipe 2 [3].

Pada penelitian-penelitian terdahulu, sudah dilakukan penelitian klasifikasi di bidang kesehatan dengan menggunakan teknik atau algoritme *data mining*. Penelitian sebelumnya yang masing-masing dilakukan oleh referensi [4][5] sudah mencoba menggunakan beberapa metode atau algoritme *data mining* diantaranya adalah *Naive Bayes*, *K-NN*, *NBTree*, dan *Decision Tree*. Salah satu penelitian yang dilakukan oleh referensi [5] menggunakan algoritme *Decision Tree*, *Naive Bayes*, dan *NBTree* untuk klasifikasi penyakit liver. Berdasarkan hasil penelitian yang dilakukan menunjukan bahwa metode *NBTree* memiliki akurasi yang paling tinggi bila dibandingkan dengan *Decision Tree dan Naive Bayes yaitu* sebesar 67,01%.

Penelitian yang dilakukan oleh referensi [4] mengembangkan metode *Correlated-Naive Bayes Classifier* atau C-NBC. C-NBC merupakan sebuah pengembangan dari metode *Naive Bayes Classfier* (NBC) dengan menambahkan paramater korelasi antar atribut terhadap kelas. Dengan memperhitungkan nilai korelasi dari masing-masing atribut vektor X terhadap kelas Y dapat meningkatkan akurasi.

Berdasarkan hasil pengujian yang dilakukan menggunakan 4 dataset yaitu yaitu Dataset *Iris*, Dataset *Balance-Scale*, Dataset *Haberman*, dan Dataset *Servo*, menunjukkan bahwa metode C-NBC mengalami kenaikan akurasi rata-rata sebesar 13,3% dibandingkan metode NBC.

Shah & Jivani [6] menggunakan algoritme *Random Forest*, *Naive Bayes*, dan *K-NN* untuk klasifikasi penyakit kanker payudara. Berdasarkan hasil penelitian yang dilakukan menunjukan bahwa metode *Naive Bayes* memiliki akurasi yang paling tinggi bila dibandingkan dengan *Random Forest dan K-NN* yaitu sebesar 95,99%. Kelemahan penelitian ini adalah tidak menjelaskan tahapan data *pre-processing* sebelum ketahapan proses klasifikasi dengan teknik *data mining*.

Berdasarkan uraian masalah diatas untuk mengurangi jumlah kematian yang disebabkan oleh penyakit diabetes adalah melakukan diagnosis secara dini. Salah satu teknik yang bisa digunakan adalah memanfaatkan teknik data mining. Untuk melakukan diagnosis penyakit diabetes dibutuhkan suatu metode yang memiliki akurasi terbaik, sehingga pada penelitian ini melakukan komparasi beberapa metode klasifikasi data mining yaitu metode *Correlated-Naive Bayes Classifier* dan *Naive Bayes Classifier* untuk mendapatkan akurasi terbaik sehingga dapat digunakan untuk diagnosis penyakit diabetes secara efektif.

II. TINJAUAN PUSTAKA

A. Data Mining

Data Mining merupakan suatu proses untuk menggali pengetahuan yang dibutuhkan dari sejumlah data besar. Data mining dan *Knowledge Discovery in Database* (KDD) secara bergantian menjelaskan proses penggalian informasi tersembunyi dalam kumpulan data yang besar. Sebenarnya kedua istilah tersebut memiliki konsep berbeda, tetapi berkaitan satu sama lain dan salah satu tahap dalam proses KDD adalah *Data Mining*. *Knowledge Disccovery in Database* (KDD) merupakan proses yang bertujuan untuk menggali, menganalisis, dan mengekstrak sejumlah data yang besar menjadi sebuah informasi atau pengetahuan yang berguna.

Adapun langkah penting dalam proses *Knowledge Discovery in Database* (KDD) seperti berikut [7] :

1.*Data Cleaning*

Data *cleaning* merupakan proses untuk membuang duplikasi data, memeriksa data yang tidak konsisten, dan memperbaiki kesalahan pada data, seperti kesalahan penulisan, data yang hilang.

2. *Data Integration*

Pada tahapan ini melakukan proses menambah data yang sudah ada dengan data atau informasi lain yang relevan atau bisa disebut juga merupakan penggabungan data dari berbagai database kedalam satu database baru yang dibutuhkan oleh KDD.

3. *Data Selection*

Pada tahapan ini melakukan pemilihan data yang relevan dan dapat dilakukan analisis dari data operasional. Data hasil pemilihan disimpan dalam *database* yang terpisah.

4. *Data Transformation*

Pada tahapan ini melakukan proses transformasi data kedalam bentuk format tertentu sehingga data tersebut sesuai untuk proses data mining. Penelitian ini menggunakan tahapan transformasi data untuk mengubah tipe data non-numerik menjadi data numerik. Hal ini dilakukan untuk menghitung nilai korelasi (*R-Square*) pada algoritme *Correlated Naive Bayes Classifier*.

5. *Data Mining*

Pada tahapan ini proses untuk mencari pola atau informasi menarik dengan menggunakan teknik, metode atau algoritme tertentu. Penelitian ini menggunakan algoritme klasifikasi N*aive Bayes Classifier* dan *Correlated Naive Bayes Classifier*

6. *Pattern Evaluation*

Pada tahapan ini melakukan proses untuk mengidentifikasi pola-pola yang benar-benar menarik dari hasil data mining. Dalam tahap ini hasil dari teknik data mining berupa pola-pola yang khas maupun model prediksi yang dievaluasi untuk menilai apakah hipotesa yang ada memang tercapai atau tidak.

7. *Knowledge Presentation*

Pada tahap ini proses untuk menampilkan pola informasi yang dihasilkan dari proses data mining, visualisasi ini membantu mengkomunikasikan hasil data mining dalam bentuk yang mudah dimengerti.

B. Metode Correlated Naive Bayes Classifier

Metode *Correlated Naive Bayes Classifier* merupakan sebuah pengembangan dari metode *Naive Bayes*. Pada metode *Correlated Naive Bayes Classifier* memperhitungkan nilai korelasi (*R-Square*) antara variabel bebas (X) terhadap variabel terikat (Y). Penambahan paramater korelasi digunakan untuk mengukur tinggi rendahnya derajat hubungan antara variabel bebas (X) terhadap variabel terikat (Y).

Formula metode *Correlated Naive Bayes Classifier* untuk klasifikasi ditunjukkan pada persamaan 1.

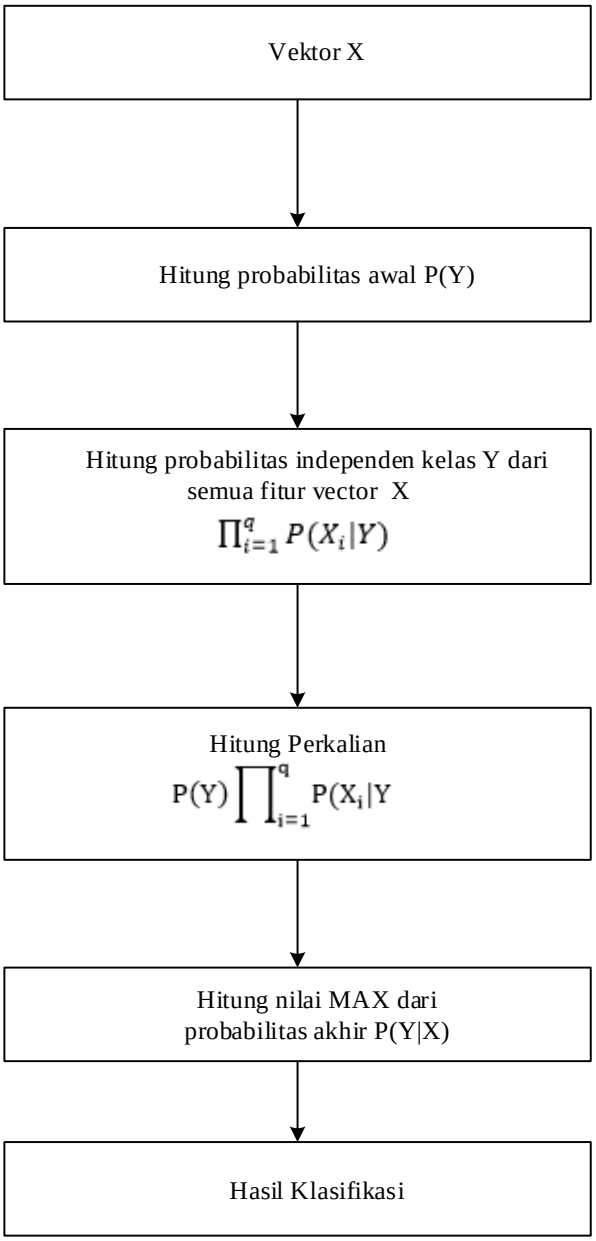
$$P(Y|X) = \frac{P(Y)\prod_{i=1}^{q} P(X_i|Y)^{\tau}.R(X_i|Y)}{P(X)} \qquad 1$$

Keterangan :

X = data dengan kelas yang belum diketahui.
Y = hipotesis data X merupakan suatu kelas spesifik.
P(X|Y) = probabilitas hipotesis Y berdasarkan kondisi Y.
P(Y) = probabilitas awal hipotesis Y (prior probability).
$\prod_{i=1}^{q} P(X_i|Y)$ = probabilitas setiap atribut dari data X berdasarkan kondisi hipotesis Y.
$R(X_i|Y)$= R Square setiap atribut dari data X berdasarkan

kondisi hipotesis Y.
τ = bilangan laplacian.
P(X) = probabilitas dari X.

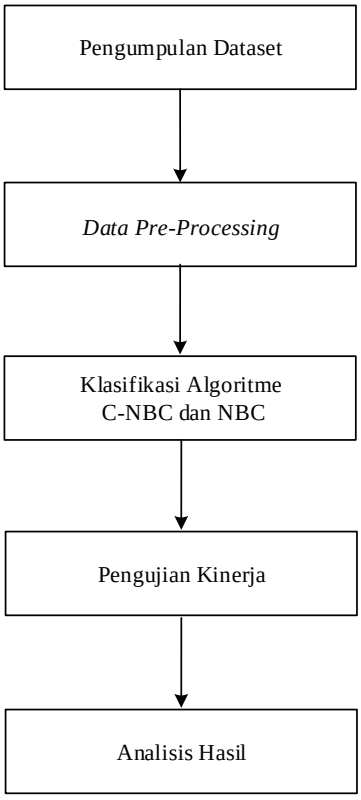
Untuk proses klasifikasi metode *Correlated Naïve Bayes Classifier* ditunjukkan pada Gambar 1.



Gbr. 1 Proses Skema *Klasifikasi Naïve Bayes Classifier*

III. METODE PENELITIAN

Penelitian ini dilakukan beberapa tahapan yang ditunjukkan pada Gambar 2 [8].



Gbr. 2 Tahapan-Tahapan Penelitian

A. Pengumpulan Dataset

Tahapan pertama yang dilakukan adalah mengumpulkan dataset *Pima Indian Diabetes* yang diperoleh dari *UCI Repository*. Dataset Pima Indian Diabetes memiliki jumlah 768 data, 9 atribut, dan 2 kelas. Adapun detail atribut *Dataset Pima Indian Diabetes* seperti ditunjukkan pada Tabel 1.

TABEL I
ATRIBUT DATASET PIMA INDIAN DIABETES

No.	Atribut	Label
1.	Number of times pregnant	Preg
2.	Plasma glucose concentation	Plas
3.	Diastoloc blood pressure (mm/Hg)	Pres
4.	Triceps skin fold thickness (mm)	Skin
5.	2-Hour serum insulin	Insu
6.	Body mass index (kg/m2)	Mass
7.	Diabetes pedigree function	Pedi
8.	Age (years)	Age
9.	Class (Tested Negative and Tested Posotive)	Class

B. Data Pre-Processing

Tahapan data *pre-processing* merupakan proses *data mining* yang pertama kali dilakukan untuk mendapatkan kualitas data sebelum dilakukan proses klasifikasi seperti transformasi data. Tahap transformasi data merupakan salah

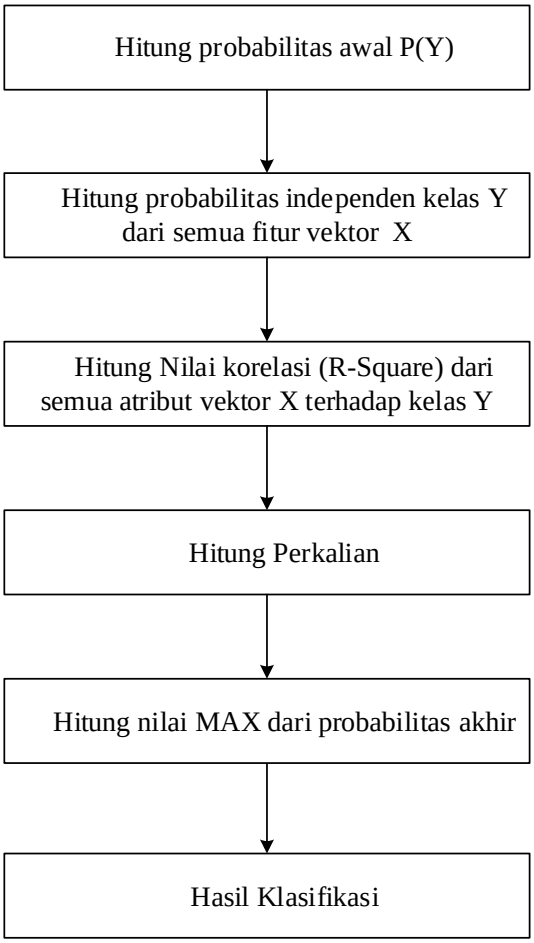
satu bagian terpenting dari proses data *pre-processing*. Transformasi data dilakukan dalam penelitian ini untuk mengubah tipe data atribut dataset pima indian diabetes yang memiliki tipe data non-numerik menjadi tipe data numerik. Hal ini dilakukan untuk memudahkan perhitungan nilai korelasi (*R-Square*) antar atribut terhadap kelas pada metode *Correlated Naive bayes Classifier* pada tahap klasifikasi. Hasil dari transformasi tipe data non-numerik menjadi numerik ditunjukkan pada Tabel 2.

TABEL 2
TRANSFORMASI DATA KELAS KE TIPE NUMERIK DATASET PIMA INDIAN DIABETES

No.	Data Non Numerik	Data Numerik
1.	Tested Negative	1
2.	Tested Positive	2

C. Klasifikasi

Setelah melalui tahapan *pre-processing*, dilanjutkan ketahap klasifikasi. Klasifikasi merupakan salah satu tugas dari data mining. Klasifikasi adalah proses untuk menentukan suatu item dari dataset kedalam label kelas. Penelitian ini menggunakan metode klasifikasi *Naive Bayes Classifier* (NBC) dan *Correlated-Naive Bayes Classifier* (C-NBC). Proses klasifikasi menggunakan kedua metode tersebut terdiri dari beberapa proses. Alur proses klasifikasi metode *Correlated Naive Bayes Classifier* ditunjukkan pada Gambar 3.



Gbr. 3 Proses *Klasifikasi Naïve Bayes Classifier*

Pengujian kedua metode tersebut menggunakan *k-fold cross validation* yang bertujuan untuk membagi data kedalam *k-fold* yang diberikan.

IV. HASIL DAN PEMBAHASAN

Pada tahapan ini berisi hasil dari tahapan-tahapan penelitian berdasarkan metode penelitian ditunjukkan pada Gambar 2. Tahapan-tahapan penelitian tersebut terdiri dari pengumpulan *Dataset Pima Indian Diabetes* dari *UCI Repository*. Tahapan selanjutnya adalah data *pre-processing*, kemudian proses klasifikasi menggunakan metode *Naive Bayes Classifier* (NBC) dan *Correlated Naive Bayes Classifier* (C-NBC). Tahapan terakhir yaitu pengukuran kinerja berdasarkan akurasi dari metode yang digunakan.

Proses klasifikasi dilakukan setelah melalui tahapan data *pre-processing*. Klasifikasi merupakan salah satu tugas dari data mining. Klasifikasi adalah proses untuk menentukan suatu item dari dataset kedalam label kelas. Penelitian ini menggunakan metode klasiikasi *Naive Bayes Classifier* (NBC) dan *Correlated Naive Bayes Classifier* (C-NBC).

Proses klasifikasi dengan algoritme *Naïve Bayes Classifier* (NBC) dan *Correlated Naïve Bayes Classifier* (C-NBC) dilakukan dengan aplikasi Java. Proses klasifikasi dilakukan

dengan teknik pengujian dengan algoritme *Naïve Bayes Classifier* (NBC) dan *Correlated Naive Bayes Classifier* (C-NBC) terhadap dataset dengan metode *10-Fold Cross Validation* dari hasil pengacakan dataset sebanyak 30 kali secara *random*.

Dataset yang digunakan pada penelitian ini yaitu *Dataset Pima Indian Diabetes*. Setelah melakukan proses klasifikasi menggunakan algoritme *Naive Bayes Classifier* (NBC) dan *Correlated Naive Bayes Classifier* (C-NBC) dari hasil pengacakan dataset sebanyak 30 kali secara *random* dengan metode *10-Fold Cross Validation* diperoleh rata-rata akurasi *Dataset Pima Indian Diabetes* yang ditunjukkan pada tabel 3 dan Tabel 4.

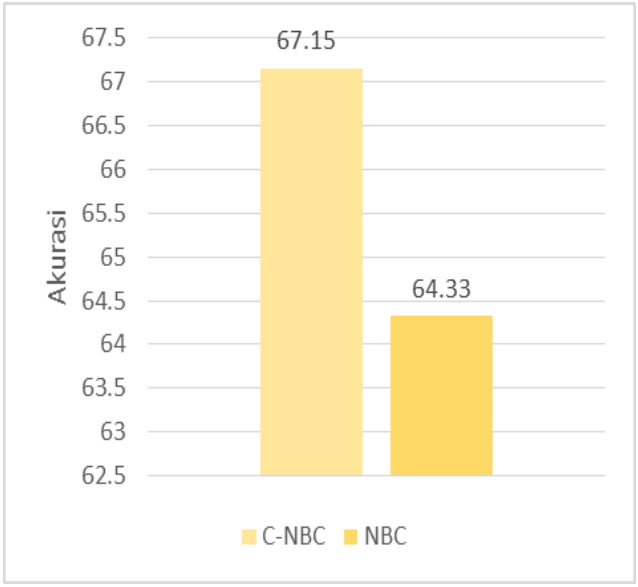
TABEL 3
RATA-RATA AKURASI DATASET PIMA INDIAN DIABETES

No	C-NBC	NBC
1.	67,89	63,95
2.	67,5	63,95
3.	67,37	64,34
4.	66,71	63,82
5.	67,37	65,53
6.	66,97	64,47
7.	67,37	64,87
8.	67,24	63,29
9.	67,37	63,68
10.	67,24	64,74
11.	67,11	65,53
12.	66,84	64,34
13.	66,18	61,97

TABEL 4
RATA-RATA AKURASI DATASET PIMA INDIAN DIABETES
(LANJUTAN)

No	C-NBC	NBC
14.	67,11	64,08
15.	67,5	64,74
16.	67,37	63,29
17.	66,97	63,29
18.	66,58	65
19.	67,63	64,47
20.	67,11	65,66
21.	67,5	64,61
22.	66,71	63,55
23.	67,24	63,68
24.	66,84	64,34
25.	67,11	66,97
26.	67,37	64,61
27.	66,97	63,68
28.	67,24	63,95
29.	67,24	65,13
30.	66,71	64,34

Untuk mempermudah melihat perbandingan akurasi algoritme *Correlated Naive Bayes Classifier* dan *Naive Bayes Classifier* terhadap dataset *Pima Indian Diabetes* berdasarkan hasil pengujian yang dilakukan ditunjukkan pada Gambar 4.



Gbr. 4 Perbandingan Akurasi Metode NBC dan C-NBC pada Dtaaset Pima Indian Diabetes

Berdasarkan Gambar 4 diatas, ditunjukkan bahwa nilai akurasi tertinggi pada *Dataset Pima Indian Diabetes* di peroleh dengan metode *Correlated Naive Bayes Classifier* (C-NBC) yaitu sebesar 67,15% dibandingkan dengan metode *Naive Bayes Classifier* (NBC) sebesar 64,33%. Hasil tertinggi untuk akurasi pada dataset tersebut diperoleh oleh metode C-NBC dibandingkan metode NBC adalah karena pada metode *Correlated Naive Bayes Classifier* (C-NBC) memperhitungkan nilai korelasi (*R-Square*) dari masing-masing atribut dataset terhadap kelasnya. Hal ini selaras dengan pendapat dar referensi [9] mengatakan bahwa penambahan parameter perhitungan korelasi pada metode *Naive Bayes Classifier* dapat meningkatkan akurasi secara *significant*.

V. KESIMPULAN

Berdasarkan pengujian yang telah dilakukan menunjukkan bahwa metode *Correlated Naive Bayes Classifier* (CNBC) memperoleh akurasi terbaik dibandingkan dengan metode

Naive Bayes Classifier (NBC) untuk *Dataset Pima indian Diabetes*. Tingkat akurasi metode *Correlated Naive Bayes Classifier* (CNBC) sebesar 67,15%, sedangkan metode *Naive Bayes Classifier* (NBC) sebesar 64,33%. Metode *Correlated Naive Bayes Classifier* (C-NBC) memiliki akurasi lebih tinggi dibandingkan metode *Naïve Bayes Classifier* (NBC) karena pada metode *Correlated Naïve Bayes Classifier* memperhitungkan nilai korelasi (*R-Square*) dari masing-masing atribut *dataset* terhadap kelasnya. Dengan demikian penggunaan metode *Correlated Naïve Bayes Classifier* (C-NBC) dapat digunakan untuk melakukan diagnosis penyakit diabetes karena memiliki tingkat akurasi yang bagus.

UCAPAN TERIMA KASIH

Terima kasih saya ucapkan kepada LPPM STMIK Bumigora Mataram atas dukungannya dalam terlaksananya penelitian ini.

DAFTAR PUSTAKA

[1] WHO, "World Diabetes Day 2015," 2005. [Online]. Available: http://www.who.int/diabetes/wdd_2015/en/. [Accessed: 24-Apr-2018].

[2] Who, "10 Facts on Diabetes," 2016. [Online]. Available: http://www.who.int/features/factfiles/diabetes/en/. [Accessed: 24-Apr-2018].

[3] Who, "Diabetes," 2015. [Online]. Available: http://www.who.int/diabetes/en/. [Accessed: 24-Apr-2018].

[4] B. A. Muktamar, N. A. Setiawan, and T. B. Adji, "Pembobotan Korelasi pada Naive Bayes Classifier," *Semin. Nas. Teknol. Inf. dan Multimed. 2015 STMIK AMIKOM Yogyakarta, 6-8 Februari 2015*, no. 1, pp. 43–47, 2015.

[5] S. N. N. Alfisahrin and T. Mantoro, "Data Mining Techniques for Optimization of Liver Disease Classification," *2013 Int. Conf. Adv. Comput. Sci. Appl. Technol.*, pp. 379–384, 2013.

[6] C. Shah and A. G. Jivani, "Comparison of Data Mining Classification Algorithms for Breast Cancer Prediction," *Comput. Commun. Netw. Technol. (ICCCNT), 2013 Fourth Int. Conf.*, vol. 4, pp. 4–7, 2013.

[7] J. Han and M. Kamber, *Data Mining Concepts and Techniques*, Second. Francisco: Morgan Kaufmann, 2006.

[8] B. A. Muktamar, N. A. Setiawan, and T. B. Adji, "Correlated Naive Bayes Classifier," Universitas Gadjah Mada, 2015.

[9] B. A. Muktamar, N. A. Setiawan, and T. B. Adji, "Analisis Perbandingan Tingkat AKurasi Algoritma Naive Bayes Classifier dengan Correlated-Naive Bayes Classifier," *Semin. Nas. Teknol. Inf. dan Multimed. 2015 STMIK AMIKOM Yogyakarta, 6-8 Februari 2015*, pp. 49–54, 2015.