INFORM : Jurnal Ilmiah Bidang Teknologi Informasi dan Komunikasi
Vol. 4 No. 1, Januari 2019, P-ISSN : 2502-3470, E-ISSN : 2581-0367

Perbandingan Hasil Analisis Teknik Data Mining "Metode Decision Tree, Naive Bayes, Smo Dan Part" Untuk Mendiagnosa Penyakit Diabetes Mellitus

Nurahman[1], Prihandoko[2]
[1]Magister Teknik Informatika, Universitas AMIKOM Yogyakarta
[2]Fakultas Ilmu Komputer dan Teknologi Informasi, Universitas Gunadarma Jakarta
[1]nurrahman.ikhtiar@gmail.com, [2]pri@staff.gunadarma.ac.id

Abstract— Diabetes Mellitus (DM) is one of the diseases that can bring a person to death. The main cause of this disease is irregular or excessive lifestyle and food. Someone who develops diabetes will be marked by increased sugar levels. This happens because of a disruption in insulin secretion and insulin action or even in both. In many countries more and more patients with diabetes, if not stopped soon, it is estimated that people with diabetes will reach 642 by 2040 [1]. This study aims to choose the best data mining clasifiers in diagnosing *Diabetes Mellitus (DM)*. Diagnosis is based on computer systems using the feature selection method and classification of the *Pima Indians Diabetes dataset*. The feature selection method used is *Correlation based Featured Selection (CFS)*. The data mining classification results in this study indicate that SMO has the highest value of accuracy compared to other *Classifiers*.

Keywords - Data mining, Classifiers, Naive Bayes, SMO, Decision Tree and PART

Abstrak— Diabetes Mellitus (DM) merupakan salah satu penyakit yang dapat membawa seseorang berujung pada suatu kematian. Peyebab utama dari penyakit ini adalah pola hidup dan makanan yang tidak teratur atau berlebihan. Seseorang yang terserang penyakit *diabetes* akan ditandai dengan meningkat kadar gula. Hal ini terjadi karena adanya gangguan pada *sekresi insulin* dan kerja *insulin* atau bahkan pada keduanya. Di diberbagai negara semakin banyak pasien penyakit diabetes, jika tidak segera dihentikan maka diperkirakan penderita penyakit *diabetes* akan mencapai 642 jiwa pada tahun 2040 [1]. Penelitian ini bertujuan untuk memilih *Clasifiers* data mining yang terbaik dalam melakukan diagnosis terhadap penyakit *Diabetes Mellitus (DM)*. *Diagnosis* yang dilakukan berbasis sistem komputer dengan menggunakan metode seleksi fitur dan klasifikasi terhadap *Dataset Pima Indians Diabetes*. Metode seleksi fitur yang digunakan adalah *Correlation based Featured Selection (CFS)*. Hasil klasifikasi data mining pada penelitian ini menunjukkan bahwa SMO memiliki nilai akurasi tertinggi dibanding *Classifiers* yang lainnya.

Kata Kunci — Data mining, Classifiers, Naive Bayes, SMO, Decision Tree dan PART

I. PENDAHULUAN

Diabetes mellitus (DM) merupakan penyakit metabolik yang ditandai dengan peningkatan kadar gula darah akibat gangguan pada sekresi insulin, kerja insulin atau keduanya. Secara umum, penyakit ini dibagi atas dua tipe, yaitu tipe 1 dengan kerusakan sel beta pankreas akibat faktor autoimun, genetik atau idiopatik dan tipe 2 yang umumnya timbul akibat resistensi insulin terkait perubahan gaya hidup [2]. Seseorang yang terkena penyakit Diabetes mellitus dapat berakibat terserang beberapa penyakit lainnya, karena biasanya penderita Diabetes mellitus akan disertai pula terkena penyakit hipertensi, jantung, strok, retinopati, kanker, ginjal dan beberapa penyakit lainnya. Hal ini tentunya sangat berbaya dan penting untuk selalu dijaga kesehatan agar tidak terkena penyakit Diabetes mellitus.

Penyakit Diabetes Mellitus menyerang di berbagai negara, terutama di negara-negara berkembang. Hal ini terjadi dikarenakan adanya perubahan gaya hidup masyarakat. Pada negara berkembang hampir seluruh penyandang penyakit Diabetes Mellitus tipe-2 sebanyak 40% penderita ditemukan berasal dari golongan yang merubah gaya hidupnya dari tradisional menuju gaya hidup modern [3]. Data pada ditahun 2011 lalu diperkirakan penderita diabetes millitus hingga mencapai 366 juta jiwa dengan perbedaan kondisi antara populasi dan wilayah [4]. Data yang diperoleh dari Internasional Diabetes Federation (IDF) juga menunjukkan bahwa tingkat prevelensi global penderita Diabetes Mellitus pada tahun 2012 sebesar 8,3% dari populasi penduduk dunia dan mengalami peningkatan 382 kasus pada tahun 2013. Pada tahun 2035 IDF memperkirakan jumlah penderita penyakit Diabetes Mellitus akan mengalami peningkatan menjadi 55 % (592 juta) jiwa, usia penderita Diabetes Mellitus berkisar antara 40-59 tahun [5].

Pada tahun 2015 IDF Diabetes Atlas Committee memperkirakan penderita penyakit diabetes tipe 1 juga menyerang anak dibawah usia 14 tahun. Bahkan jika penyakit diabetes ini tidak dihentikan maka diperkirakan penderita penyakit diabetes akan mencapai 642 jiwa pada tahun 2040 [1]. Semakin banyaknya penderita penyakit diabetes hingga Negara Indonesia pernah menduduki peringkat kelima di Dunia sebagai negara dengan jumlah penderita diabetes melitus terbanyak setelah Banglades, Bhutan, Cina dan India [6]. Kemudian ditahun 2015 pemeringktan Indonesia menjadi ke-7 di Dunia dengan estimasi penderita diabetes sebesar 10 juta jiwa [1].

DOI: 10.25139/inform.v3i2.1010

Diagnosis penyakit diabetes tidak hanya dilakuan atas dasar adanya glukosuria saja. Penentuan diagnosis *Diabetes Mellitus* perlu dilakukan pemeriksaan terhadap glukosa darah, pemeriksaan yang dianjurkan adalah pemeriksaan glukosa secara enzimatik dengan bahan darah plasma vena [7]. Dalam pemeriksaan juga ada tahapan-tahapan tertentu yang harus dilakukan. Bahkan dalam pemeriksaan juga biasa akan memperhatikan *Pregnan, Plasma, Pressure, Skin, Insulin, Mass, Predigree,* dan *Age.*

Diagnosis terhadap penyakit *Diabetes Mellitus* secara medis masih mengalami kesulitan dan bahkan mengalami reduksi data. Data medis yang memiliki sejumlah fitur yang tidak relevan, dan *redundant* dapat memberikan pengaruhi terhadap kualitas dari diagnosis penyakit [8]. Untuk mendukung mengenai *diagnosis* perlu menggunakan teknik data mining berbasis komputer agar dapat menggali informasi yang berharga dari kumpulan informasi *Diabetes Mellitus* [9]. Data mining merupakan sebuah proses terpadu dari analisis data yang terdiri dari serangkaian kegiatan yang berjalan berdasarkan pada pendefinisian tujuan dari apa yang akan dianalisis sampai pada interpretasi dan evaluasi hasil [10].

Pada penelitian terdahulu, banyak metode data mining yang telah digunakan untuk mendiagnosis penyakit Diabetes Mellitus. Penggunaan Metode C4.5 [11], [12], [13]. Metode Naive Bayes dilakukan oleh dkk [11], [14], [12], dan [13]. Metode SVM dilakukan oleh Zheng dkk [11], [12], [15] dan [16].

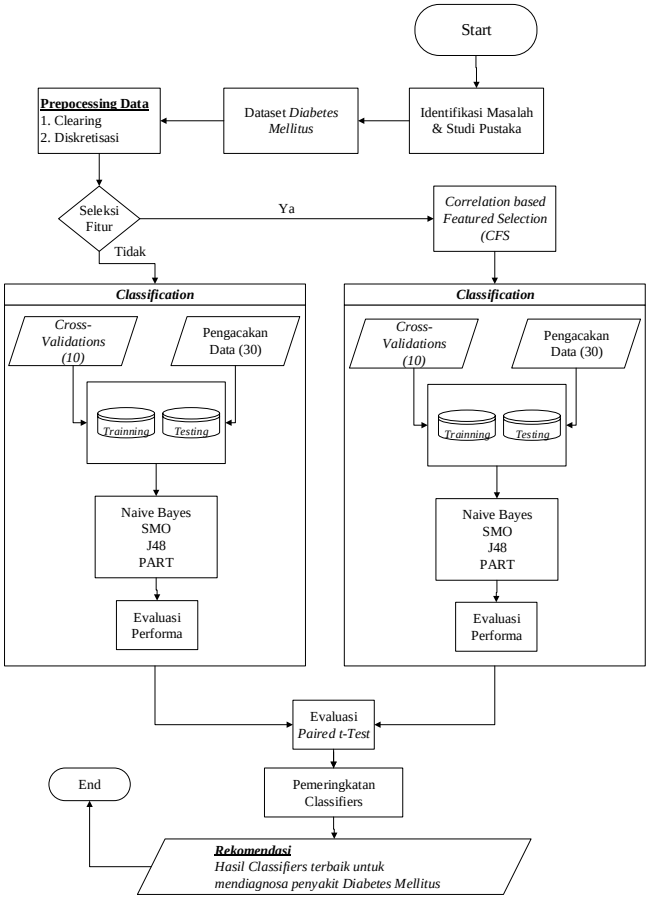
Berdasarkan penelitian-penelitain terdahulu ditemukan bahwa peneliti sebelumnya tidak melakukan reduksi data. Sedangkan dapat diketahui bahwa dataset *Diabetes Melitus* yang dimiliki para bidang kesehatan masih terdapat *redundant* data. Oleh karena itu, untuk menghasilkan diagnosis mengenai penyakit Diabetes Mellitus yang berkualitas perlu dilakukan seleksi pada fitur-fitur pada dataset. Pada penelitian ini akan mengusulkan sebuah proses seleksi fitur-fitur untuk mengatasi kekurangan pada penelitian sebelumnya. Dengan hasil seleksi fitur yang dilakukan peneliti diharapkan dapat meningkatkan performa *diagnosis Diabetes Mellitus*. Sehingga hasil dari penelitian ini dapat memberikan masukan kepada para ahli kesehatan.

Struktur dalam paper ini adalah sebagai berikut. Bab 1 Pendahuluan. Bab 2 menjelaskan tentang langkah penelitian menggunakan metode yang digunakan. Bab 3 menampilkan hasil dan diskusi tentang hasil penelitian. Dan yang terakhir adalah pada Bab 4 kami membahas kesimpulan.

II. METODE PENELITIAN

Pada penelitian ini menggunakan *Dataset Pima Indians Diabetes* yang terdiri dari 8 fitur. Penelitian dilakukan dengan manggambarkan fenomena secara *numberik* yang memandang setiap *realitas* / gejala / *fenomena* itu dapat diklasifikasikan, *relative* tetap, *konkrit*, teramati, terukur, dan memiliki hubungan gejala bersifat sebab akibat. Sehingga penelitian ini dapat dikategorikan menjadi penelitian yang bersifat *kuantitatif*. Nilai data *numerik* pada penelitian ini, diperoleh dari dataset yang telah tersedia dari sumbernya. Data *numerik* akan diolah menjadi suatu informasi yang dapat menyelesaikan permasalahan dalam penelitian.

Menemukan masalah hingga dapat menyelesaikannya menjadi suatu informasi yang relevan diperlukan ketelitian dalam penelitian. Gambar 1 menunjukkan Bagan alur penelitian yang akan dilakukan. Bagan alur penelitian ini dapat membantu dalam melaksanakasn penelitian agar tetap konsisten terhadap tujuan penelitian. Penelitian ini dilakukan mulai dari identifikasi masalah dengan memahami berbagai sumber-sumber referensi terkait dalam penelitian. Kemudian dari dataset yang digunakan akan dipertimbangkan mengenai fitur-fitur yang dimiliki dataset tersebut. Dilanjutkan berbagai tahapan klasifikasi dengan menggunakan dataset penderita penyakit *Diabetes Mellitus*. Berikut ini adalah gambar alur penelitian yang dilakukan.



Gambar 1. Bagan Alur Penelitian

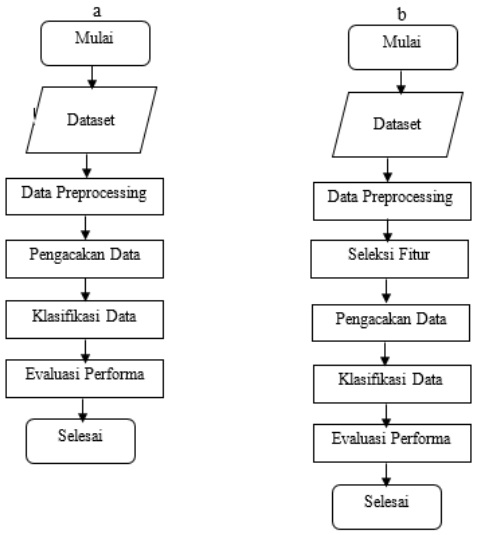
DOI: 10.25139/inform.v3i2.1010

Pendekatan yang digunakan dalam penelitian ini adalah dengan memahami cara kerja algoritma-algoritma *data mining*. Setiap algoritma akan menghasilkan masing-masing nilai performa akurasi yang berbeda-beda. Perbedaan dari setiap nilai performa akurasi akan dilakukan pemeringkatan hingga memperoleh algoritma yang terbaik dalam mendiagnosis penyakit *Diabetes Mellitus*.

Penderita diabetes memiliki berbagai gejala-gejala yang dialaminya. Beragamnya gejala penderita diabetes merupakan tantangan dalam penelitian ini. Penyelesaian dalam kasus ini diperlukan langkah-langkah dan teknik terbaik agar dapat memehami dan penyelesaian masalah dalam penelitian. Memehami suatu masalah dalam penelitian diperlukan informasi dan data-data pendukung. Untuk mendapatkan informasi dan data-data pendukung, maka metode pengumpulan data yang diterapkan adalah sebagai berikut:

1. Studi literatur, yaitu pengumpulan data penelitian dengan mempelajari buku, file atau dokumen yang diperlukan dalam penelitian ini. Hal-hal dibutuhkan yaitu mengenai literatur penyakit Diabetes Mellitus, penggunaan metode data mining seperti *C4.5, Naive Bayes, SMO dan PART*.
2. *Observasi*, yaitu data yang digunakan dalam penelitian ini diambil dari Dataset Pima Indians Diabetes. Sehingga dalam penelitian ini akan melakukan Eksperimen terhadapat dataset gejala-gejala penyakit Diabetes Mellitus yang bersumber dari GitHubGis https://gist.github.com/ktisha/c21e73a1bd1700294ef790c56c8aec1f.

Data mining memiliki berbagai jenis kategori untuk melakukan penelitian. Dalam penelitian ini memilih Klasifikasi data dengan menggunakan metode *Decision Tree (C4.5), Naive Bayes, SMO, dan PART*. Berikut adalah alur yang dilakukan dalam menganalisis penelitian ini :



Gambar 2. (a) klasifikasi *CFS*, (b) klasifikasi dengan *CFS*

Langkah pertama yang dilakukan adalah data preprocessing, kemudian pengacakan data. Berdasarkan Gambar 2(a) dan Gambar 2(b), proses diagnosis atau klasifikasi penyakit *Diabetes Mellitus* dilakukan dengan dua cara yaitu menggunakan seleksi fitur dan tanpa seleksi fitur. Hasil performa diagnosis yang diperoleh selanjutnya dievaluasi untuk memberikan perbandingan antara diagnosis dengan menggunakan seleksi fitur dan tanpa seleksi fitur.

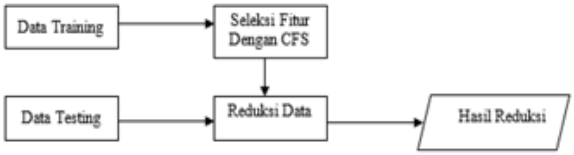
a. Data *Preprocessing*

Langkah awal yang dilakukan dalam pengecekan terhadap Dataset Pima Indians Diabetes yaitu tahapan Preprocessing. tahapan ini dilakukan menjadi 2 langkah yaitua:

a) *Cleaning*
b) *Diskretisasi data numerik*

b. Seleksi Fitur

Proses seleksi fitur ini dilakukan untuk mereduksi data Dataset Pima Indians Diabetes dan memilih fitur-fitur yang relevan terhadap diagnosis penyakit diabetes. Penelitian ini melakukan seleksi fitur menggunakan teknik *FCS*. Langkah-langkah dalam seleksi fitur dapat dilihat pada gambar 3 berikut.



Gambar 3. Seleksi fitur

Hasil dari *Preprocessing*, maka data akan dibagi menjadi dua bagian yaitu data training dan data testing. Proporsi data traning sebesar 2/3 sedangkan data testing sebesar 1/3 bagian. Proses train-test split dilakukan dengan disertai stratifikasi sehinggga proposi kelas positif dan kelas negatif pada data tranning dan testing akan sama. Selanjutnya proses seleksi fitur akan dilakukan pada data training. Kemudian data testing direduksi sesuai dengan hasil seleksi fitur pada data training.

c. Pengacakan Data

Proses pengacakan data pada Dataset Pima Indians Diabetes dapat mempengaruhi performa dari diagnosis penyakit diabetes. Hal tersebut karena adanya model klasifikasi yang dilakukan selama pelatihan terhadap data. Penerapan pngacakan terhadap data akan dilakukan setelah hasil dari preprocessing dan hasil reduksi.

d. Klasifikasi Data

Pada tahapan ini melakukan klasifikasi terhadapt data Dataset Pima Indians Diabetes dengan menggunakan Algoritma *Decision tree, naive bayes, SMO* dan *PART*. Performa klasifikasi yang disajikan dengan melakukan 10-fold cross validation pada dataset hasil pengacakan. Penerapan 10-fold

DOI: 10.25139/inform.v3i2.1010

cross validation dilakukan berdasarkan default yang diperoleh dari software weka yang biasanya digunakan secara umum dalam berbagai penelitian. Sembilan fold pertama data akan dijadikan sebagai training sedang folt terakhir akan dijadikan sebagai testing. Kemudian sembilan folt kedua akan dijadikan testing sedangkan sisanya kan dijadikan testing. Proses ini dilakukan terus menerus samapai semua folt pernah satu kali menjadi data testing. Pada tahapan ini software weka digunakan untuk mengklasifikasi Dataset Pima Indians Diabetes dengan menerapkan Algoritma *Decision tree, naive bayes, SMO* dan *PART*.

e. Evaluasi Performa

Tahapan Evaluasi performa dilakukan dengan melihat nilai akurasi dari setiap *Classifiers*. Rumus akurasi yang digunakan adal sebagai berikut [17]:

$$Akurasi = \frac{tp + tn}{tp + tn + fp + fn}$$

Namun untuk mengetahui tingkat signifikansi perbedaan antara performa yang dihasilkan oleh masing-masing algoritme klasifikasi, perlu dilakukan uji Paired t-Test. Berikut ini langkah-langkah dalam melakukan uji *Paired t_Test* [18]*:*

Langkah 1 : Hitung $m_d = m_x - m_y$, (m_d adalah perbedaan kedua *mean*)
Langkah 2 : Hitung $\sigma_d{}^2 = \frac{\sigma_x{}^2}{k} - \frac{\sigma_y{}^2}{k}$, ($\sigma_d{}^2$ adalah perbedaan kedua *variance*)
Langkah 3 : Hitung $t = \frac{m_d}{\sqrt{\sigma_d{}^2 / k}}$, (Bentuk standar dari m_d disebut nilai *t* statistik)
Langkah 4 : Nilai *confidence limit z* dapat diperoleh dari tabel distribusi *student* dengan derajat kebebasan *k-1*
Langkah 5 : Karena uji ini merupakan *two-tailed test*, maka tingkat signifikansi (α) yang digunakan harus dibagi menjadi dua ($\alpha/2$), kemudian cari nilai *z* yang bersesuaian dengan tabel distribusi *student*.
Langkah 6 : Jika $t \leq -z$ atau $t \geq z$ maka perbedaan antara kedua skema pembelajaran adalah signifikan

III. HASIL DAN DISKUSI

Dalam bab ini, kami membahas mengenai tahapan analisis terhadap dataset yang digunakan. Kemudian melakukan seleksi terhadap fitur-fitur dataset. Hasil seleksi fitur akan dianalisis menggunakan metode data mining klasifkasi. *Classifiers* yang digunakan yaitu *Naive Bayes, SMO, J48,* dan *PART*. Hasil dari nilai performa akurasi akan dilakukan uji *Pairend t-Test* untuk menentukan signifikansi dari setiap nilai akurasi *Classifiers.*

A. Dataset

Dataset pada penelitian ini memiliki jumlah *instance* sebanyak 768, dengan data numerik. Dataset memiliki atribut-atribut atau fitu yang mendeskripsikan setiap *instance.* Berikut ini adalah daftar atribut pada penelitian ini :

TABEL I
ATRIBUT PADA DATASET

No	Indikator	Keterangan	Fitur
1	*Number of times pregnant*	Berapa kali hamil	*Pregnan*
2	*Plasma glucose concentration a 2 hours in an oral glucose tolerance test*	Konsentrasi glukosa plasma 2 jam dalam tes toleransi glukosa oral	*Plasma*
3	*Diastolic blood pressure (mm Hg)*	Tekanan darah diastolik (mm Hg)	*Pressure*
4	*Triceps skin fold thickness (mm)*	Ketebalan lipatan kulit trisep (mm)	*Skin*
5	*2-Hour serum insulin (mu U/ml)*	Insulin serum 2-Jam (mu U / ml)	*Insulin*
6	*Body mass index (weight in kg/(height in m)^2)*	Indeks massa tubuh (berat tubuh dalam kg / (tinggi dalam m) ^ 2)	*Mass*
7	*Diabetes pedigree function*	Riwayat diabetes dalam keluarga	*Pedigree*
8	*Age (years)*	Umur (tahun)	*Age*
9	*Class*	Kelas	*Class*

Setiap fitur pada tabel 1 memiliki nilai *instance* yang beragam. Nilai *instance* dataset dapat disajikan kedalam bentuk numerik *range.* Pada tabel 2 disajikan *range* nilai dataset yang digunakan pada penelitin ini.

TABEL II
RANGE NILAI DATASET

Atribut	Nilai
Pregnan	0 – 17
Plasma	0 – 199
Pressure	0 – 122
Skin	0 – 99
Insulin	0 – 846
Mass	0 – 67.1
Predigree	0.078 – 2.420
Age	21 – 81
Class	*Tested_Positif, Tested_Negative*

Pada tabel 2 menunjukkan bahwa setiap atribut memiliki *range* nilai masing-masing. *Range* nilai pada dataset tersebut adalah numerik. Nilai numerik pada dataset ini akan dilakukan analisis sesuai tahapan pada metode penelitian.

B. Preprocessing Data

Pada tahapan ini dataset yang berjumlah 768 *instance* akan dilakukan tahapan *Preprocessing Data.* Tahapan *Preprocessing* dilakukan menjadi dua tahapan yaitu tahapan *Clearing* dan tahapan *diskretisasi.*

DOI: 10.25139/inform.v3i2.1010

1) Clearing

Preprocessing pertama dilakuakan adalah tahapan cleaning data atau tahapan penghapusan data terhadap missing value yang terdapat pada Dataset Pima Indians Diabetes. Pada dataset ini akan dilakukan penghapusan terhadap data duplikasi, pemeriksaan data yang inkonsisten, dan memperbaiki kesalahan data (seperti kesalahan cetak/tipografi). Dataset *diabetes mellitus* peneitian ini adalah sebagai berikut.

TABEL III
DATASET DIABETES MILLETUS

Pregnan	*Plasma*	*Pressure*	*Skin*	*Insulin*	*Mass*	*Pedigree*	*Age*	*Class*
6	148	72	35	0	33.6	0.627	50	*tested_positive*
1	85	66	29	0	26.6	0.351	31	*tested_negative*
8	183	64	0	0	23.3	0.672	32	*tested_positive*
1	89	66	23	94	28.1	0.167	21	*tested_negative*
0	137	40	35	168	43.1	2.288	33	*tested_positive*
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
1	93	70	31	0	30.4	0.315	23	*tested_negative*

Pada tahapan clearing yang telah dilakukan tidak menemukan mising value data. Sehingga data sebanyak 678 ini akan di lanjutkan ketahap Preprocessing berikutnya, agar data menjadi data yang berkualitas baik untuk penelitian ini.

2) *Diskretisasi*

Pada tahapan *diskretisasi* dataset akan dilakukan pengelompokan nilai pada suatu interval tertentu. Diskretisasi ditentukan dengan melihat frekuensi nilai yang hampir sama. Hasil dari *diskretisasi* menggunakan aplikasi weka 3.8 dapat dilihat pada tabel 4 berikut ini.

TABEL IV
HASIL DIKRETISASI

Fitur	**Hasil Diskret**				
Pregnan	(-inf-6.5]	(6.5-inf)			
Plasma	(-inf-99.5]	(99.5-127.5]	(127.5-154.5]	(154.5-inf)	
pressure	(-inf-69]	(69-inf)			
Skin	(-inf-7.5]	(7.5-23.5]	(23.5-31.5]	(31.5-inf)	
Insulin	(-inf-7]	(7-14.5]	(14.5-87.5]	(87.5-121]	(121-inf)
Mass	(-inf-27.85]	(27.85-inf)			
Predigree	(-inf-0.5275]	(0.5275-inf)			
Age	(-inf-24.5]	(24.5-28.5]	(28.5-inf)		
Class	0	1			

Pada tabel 4 diatas kemudian dilakukan pengkategorian nilai dengan menggunakan nilai 0 sampai nilai 4. Nilai kategori juga disesuaikan dengan kebutuhan setiap fitur masing-masin. Pemberian nilai kategori tersebut akan memberi kemudahan dalam penelitian.

C. Seleksi Fitur

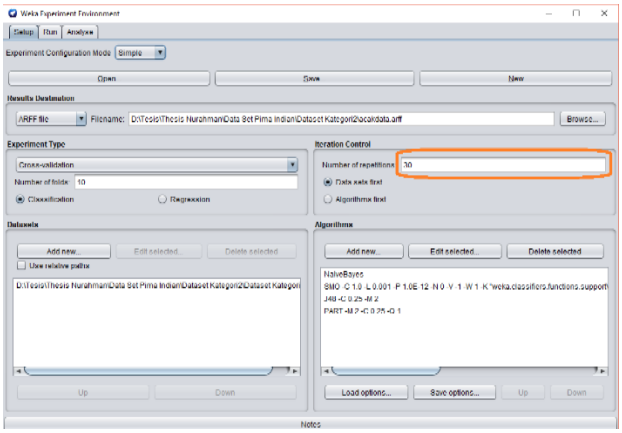
Proses Seleksi Fitur dilakukan dengan Teknik seleksi fitur *Correlation based Featured Selection (CFS)*. Seleksi fitur ini dilakukan untuk memilih fitur yang memiliki korelasi tinggi terhadap kelas. Sehingga hasil dari seleksi ini akan memilih hanya beberapa fitur yang akan digunakan dalam penelitian ini. Tabel 5 adalah hasil dari seleksi fitur menggunakan Teknik seleksi fitur *Correlation based Featured Selection (CFS)*.

TABEL V
HASIL SEBELUM DAN SESUDAH DILAKUKAN SELEKSI FITUR

Atribut Sebelum Seleksi Fitur	Atribut Sesudah Seleksi Fitur
Pregnan, Plasma,Pressure, Skin, Insulin, Mass, Pedigree, Age, Class	*Plasma, Insulin,Mass, Pedigree, Age, Class*

D. Pengacakan Data

Pengacakan Dataset dilakukan sebanyak 30 kali dengan menggunakan Aplikasi weka 3.8 pada setiap *Classifiers* yang digunakan. Pengacakan dilakukan terhadap dataset sebelum seleksi fitur dan dataset setelah seleksi fitur dengan metode CFS. Tahapan pengacakan dilakukan dengan memberikan nilai 30 pada kolom number of repetitions. Penentuan 30 kali acak data dapat dilihat pada gambar 4 yang telah dilakukan pada aplikasi weka 3.8.



Gambar 4. Penentuan Pengacakan *Dataset*

E. Klasifikasi Data

Klasifikasi data dilakukan dengan berbagai *Classifiers* diantaranya yaitu *Naive Bayes, SMO, J48, dan PART*. Pada gambar 1 menunjukan *Experiment Type* yang digunakan adalah *Cross-Validations* dengan ditentukan *Number Of Folds 10. Artinya Experiment Type Cross-Validations* ini dilakukan dengan membagi data sebanyak 768 menjadi 10 bagian. Data yang telah dibagi menjadi 10 bagian ini akan diacak sebanyak 30 kali dengan menggunakan *Naive Bayes, SMO, J48, dan PART*. Setiap proses klasifikasi dan pengacakan akan menghasilkan nilai *Num_true_positives* (TP), *Num_true_*

DOI: 10.25139/inform.v3i2.1010

negatives (TN), Num_false_positives (FP), dan *Num_false_ negatives (FN).*

Nilai TP, TN, FP, dan FN yang diperoleh ada dua yaitu nilai-nilai sebelum seleksi fitur dan nilai-nilai setelah seleksi fitur. Nilai tersebut akan dignakan untuk mencari nilai akurasi pada setiap *Classifiers.*

F. Evaluasi Performa

Performa klasifikasi dengan *10-Fold Cross-Validations* ini diimplementasikan dengan menggunakan empat *Classifiers* yaitu *Naive Bayes, SMO, J48, dan PART.* Ditentukannya 10 pembagian pada Dataset Pima Indians Diabetes yang diacak sebanyak 30 kali, maka menghasilkan nilai akurasi sebanyak 300 data. Hasil nilai akurasi sebelum dilakukan seleksi fitur dan sesudah seleksi fitur disajikan dalam bentuk persentase (%).

TABEL VI
NILAI PERFORMA AKURASI

Algoritma	RATA-RATA AKURASI	
	Full Feature%	***CFS%***
Naive Bayes	76.52272731	77.95876054+
SMO	77.58971296	77.96707678
PART	74.40510368	76.60554796+
J48	75.67128049	76.88288904+

Tanda (+) dan (-) pada tabel 6 menunjukan bahwa penerapan *Classifiers* terhadap dataset yang menggunakan metode seleksi fitur (*CFS*) mengalami peningkatan atau mengalami penurunan secara signifikan dibandingkan dengan dataset tanpa menggunakan metode seleksi fitur. Secara statistik tingkat akurasi dapat diukur menggunakan uji Paired t-Test dengan tingkat signifikansi sebesar 5%.

Langkah berikutnya yang dilakukan dalam penelitian ini yaitu uji *Paired t-Test* terhadap nilai performa akurasi. Hasil dari uji *Paired t-Test* dapat diketahui bahwa nilai akurasi sebelum seleksi fitur dibandingkan dengan nilai akurasi setelah seleksi fitur mengalami peningkatan secara signifikan pada setiap *Classifiers,* keculai pada klasifikasi SMO. Sehingga dapat dinyatakan bahwa teknik seleksi fitur dengan metode *Correlation based Featured Selection (CFS)* dapat meningkatkan nilai performa akurasi secara signifikan.

Langkah selanjutnya yaitu memilih *Classifiers* terbaik dalam mendiagnosis penyakit diabetes. Pemilihan ini dilakukan dengan membuat skema pemeringkatan antara nilai performa akurasi pada setiap *Classifiers.* Tabel 8 menunjukkan skema pemeringkatan nilai akurasi setiap *Classifiers.*

TABEL VII
SKEMA PEMERINGKATAN NILAI AKURASI CLASSIFIERS

Perbandingan	*Classifiers* Terpilih	Nilai Akurasi Terpilih
Naive Bayes vs SMO	SMO	77.96707678
Naive Bayes vs J48	*Naive Bayes*	77.95876054+
Naive Bayes vs PART	*Naive Bayes*	77.95876054+
SMO vs Naive Bayes	*SMO*	77.96707678
SMO vs J48	*SMO*	77.96707678
SMO vs PART	*SMO*	77.96707678
J48 vs Naive Bayes	*Naive Bayes*	77.95876054+
J48 vs SMO	*SMO*	77.96707678
J48 vs PART	*J48*	76.88288904+
PART vs Naive Bayes	*Naive bayes*	77.95876054+
PART vs SMO	*SMO*	77.96707678
PART vs J48	*J48*	76.88288904+

Berdasarkan skema pemeringkatan pada tabel 7, maka dapat diuraikan menjadi sebagai berikut :

1) Terpilih *SMO* memiliki nilai performa akurasi tertinggi diantara *Classifiers* yang lain dengan jumlah pemeringkatan 6
2) *Naive Bayes* mendapat peringkat setelah *SMO* dengan jumlah pemeringkatan 4
3) *J48* mendapat pemeringkatan setelah *Naive Bayes dengan jumlah pemeringkatan 2*
4) *PART* merupakan sebuah algoritma klasifikasi dengan hasil nilai akurasi terendah dibandingkan algoritma yang lainnya.

Hasil skema klasifikasi dengan mempertimbangkan nilai peforma akurasi dapat disimpulkan bahwa Algoritma *SMO* merupakan algoritma yang tepat untuk digunakan dalam mendiagnosis penyakit *diabetes mellitus.* Nilai performa akurasi yang diperoleh dari SMO adalah 77.96707678.

IV. KESIMPULAN

Pada penelitian ini dapat disimpulkan bahwa teknik selesi fitur dengan menggunakan metode *Correlation based Featured Selection (CFS)* dapat meningkatkan performa akurasi secara signifikan. Hasil skema pemilihan algoritma menunjukkan bahwa algoritma SMO memiliki nilai performa akurasi yang lebih tinggi dibandingkan dengan algoritma yang lain.

REFERENSI

[1] D. Cavan, J. R. Fernandes, S. Webber, K. Ogurtsova, dan L. Makaroff, Ed., *IDF Diabetes Atlas Seventh edition*, 7 ed. International Diabetes Federation, 2015.

[2] Trihon dan N. Mboi, "Riset Kesehatan Dasar," *Badan Penelit. Dan Pengemb. Kesehat. Kementeri. Kesehat. RI*, hlm. 304, 2013.

[3] E. Shakibazadeh, B. Larijani, D. Shojaeezadeh, A. Rashidian, M. Forouzanfar, dan L. Bartholomew, "Patients' Perspectives on Factors that Influence Diabetes Self-Care," *Iran. J. Public Health*, vol. 40, no. 4, hlm. 146–158, Des 2011.

DOI: 10.25139/inform.v3i2.1010

[4] J. E. Shaw, R. A. Sicree, dan P. Z. Zimmet, "Global estimates of the prevalence of diabetes for 2010 and 2030," *Diabetes Res. Clin. Pract.*, vol. 87, no. 1, hlm. 4–14, Jan 2010.
[5] L. Guariguata, T. Nolan, J. beagley, U. Linnenkamp, dan olivier Jacqmain, Ed., *IDF Diabetes Atlas Sixth edition*, 6 ed. International Diabetes Federation, 2013.
[6] Bustam, *Epidemologi Penyakit Tidak Menular*. Jakarta: PT. Rineka Cipta, 2009.
[7] Perkeni, *Konsensus Pengelolaan dan Pencegahan Diabetes Melitus Tipe 2*. Perkumpulan Endokrinologi Indonesia, 2011.
[8] N. Chu, L. Ma, J. Li, P. Liu, dan Y. Zhou, "Rough set based feature selection for improved differentiation of traditional Chinese medical data," dalam *2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery*, Yantai, China, 2010, hlm. 2667–2672.
[9] I. P. D. Lesmana, "Perbandingan Kinerja Decision Tree J48 dan ID3 Dalam Pengklasifikasian Diagnosis Penyakit Diabetes Mellitus," vol. 2, no. 2, hlm. 10, 2012.
[10] P. Giudici dan S. Figini, *Applied Data Mining for Business and Industry*. Chichester, UK: John Wiley & Sons, Ltd, 2009.
[11] T. Zheng *dkk.*, "A machine learning-based framework to identify type 2 diabetes through electronic health records," *Int. J. Med. Inf.*, vol. 97, hlm. 120–127, Jan 2017.
[12] D. Sisodia dan D. S. Sisodia, "Prediction of Diabetes using Classification Algorithms," *Procedia Comput. Sci.*, vol. 132, hlm. 1578–1585, 2018.
[13] E. S. Kundari, "Perbandingan Kinerja Metode Naive Bayes dan C4.5 Dalam Pengklasifikasian Penyakit Diabetes Melitus di Rumah Sakit Kumala Siwi Kudus," hlm. 8.
[14] M. Maniruzzaman *dkk.*, "Comparative approaches for classification of diabetes mellitus data: Machine learning paradigm," *Comput. Methods Programs Biomed.*, vol. 152, hlm. 23–34, Des 2017.
[15] J. A. Putra dan A. L. Akbar, "Klasifikasi Pengidap Diabetes Pada Perempuan Menggunakan Penggabungan Metode Support Vector Machine dan K-Nearest Neighbour," vol. 1, no. 2, hlm. 6, 2016.
[16] M. Alehegn, R. Joshi, dan D. P. Mulay, "Analysis and Prediction of Diabetes Mellitus using Machine Learning Algorithm," hlm. 8.
[17] F. Gorunescu, "Data Mining (Intelligent Systems Reference Library, 12)," vol. 12, hlm. 370, 2011.
[18] I. H. Witten dan E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*. Elsevier, 2011.

DOI: 10.25139/inform.v3i2.1010