e-ISSN 2549-7472
Edumatic: Jurnal Pendidikan Informatika
Vol. 5 No. 1, Juni, 2021, **Hal. 89-98**
DOI: 10.29408/edumatic.v5i1.3336

# Klasifikasi Penyakit Diabetes Menggunakan Metode *CFS* dan *ROS* dengan Algoritma J48 Berbasis Adaboost

**Dian Pramadhana**
Program Studi Teknik Informatika, Politeknik Baja Tegal
email: dianpramadhana@gmail.com

(Received: 20 April 2021/ Accepted: 29 Mei 2021 / Published Online: 20 Juni 2021)

**Abstrak**

Diabetes adalah penyakit yang terjadi akibat kadar glukosa di dalam darah tinggi. Para peneliti berusaha untuk mencegah berkembangnya komplikasi dengan menggunakan teknik data mining. Salah satu teknik yang digunakan dalam data mining adalah klasifikasi. Tujuan dari penelitian ini untuk meningkatkan hasil akurasi pengklasifikasian penyakit diabetes agar lebih baik dan optimal. Metode yang digunakan pada penelitian ini adalah *Correlation Feature Selection (CFS)* sebagai seleksi atribut, *Random Over Sampling* untuk menangani data yang tidak seimbang dan AdaBoost untuk meningkatkan kinerja algoritma J48 agar hasil yang didapat lebih optimal. Berdasarkan hasil penelitian yang telah dilakukan, menunjukan bahwa *Correlation Feature Selection* untuk seleksi atribut dan *Random Over Sampling* untuk menangani ketidakseimbangan kelas dengan algoritma J48 berbasis Adaboost terbukti dapat meningkatkan hasil klasifikasi penyakit diabetes dengan akurasi sebesar 92,3%. Disarankan untuk penelitian selanjutnya dapat menerapkan metode lain agar hasil akurasi yang diperoleh lebih optimal untuk dijadikan perbandingan.

**Kata Kunci**: Adaboost, *Correlation Feature Selection*, Diabetes Melitus, J48, *Random Over Sampling*

***Abstract***

*Diabetes was a disease that occurs due to high blood-glucose levels. Researchers tried to prevent complications from developing by using data mining techniques. One of the techniques used in data mining was classification. The purpose of this study improves the accuracy of the classification of diabetes for a better and optimal result. The method in this study is Correlation Feature Selection (CFS) as attribute selection, Random Over Sampling to handle unbalanced data and AdaBoost to improve the performance of the J48 algorithm so the result obtained best. Based on the result of this study, showed that Correlation Feature Selection for attribute selection and Random Over Sampling to handle imbalance's class with the Adaboost-based J48 algorithm proved can increase the results of the diabetes classification with an accuracy of 92.3%. For the further research recommended to apply other methods so that the accuracy results obtained are more optimal for comparison.*

***Keywords**: Adaboost, Correlation Feature Selection, Diabetes Mellitus, J48, Random Over Sampling*

## PENDAHULUAN

Diabetes Mellitus atau yang sering di kenal Diabetes adalah penyakit yang terjadi karena kadar gula pada darah tinggi yang disebabkan oleh tubuh yang tidak bisa melepaskan insulin dengan normal (Kemenkes., 2018). Para peneliti dan praktisi memusatkan perhatiannya untuk mendeteksi kondisi DM dan mencegah atau menghambat berkembangnya komplikasi (Khairani, 2019). Untuk mendukung hal ini dapat digunakan teknik data mining untuk menggali informasi yang berharga dari kumpulan informasi atau histori data diabetes (Mardi, 2017).

Data mining adalah proses pencarian pola-pola yang tersembunyi berupa pengetahuan yang tidak diketahui sebelumnya dari sekumpulan data (Takdirillah, 2020). Kegiatan yang meliputi pengumpulan, pemakaian data historis untuk menemukan keteraturan, pola dan hubungan dalam set data berukuran besar. Teknik yang digunakan untuk pengekstrakan pengetahuan dalam data mining salah satunya yaitu klasifikasi. Klasifikasi merupakan bagian penting dari data mining, Klasifikasi merupakan sebuah proses untuk menemukan sekumpulan model yang membedakan kelas data, yang bertujuan untuk dapat memperkirakan kelas dari suatu objek yang kelasnya tidak diketahui (Permana & Dewi, 2021). Ada beberapa algoritma yang bisa untuk melakukan klasifikasi data diantaranya J48, nearest neighbor, dan Support Vector Machine (Kaunang, 2019).

Saat ini, Algoritma J48 merupakan algoritma pengklasifikasian yang sering digunakan karena mempunyai struktur yang sederhana dan mudah untuk diinterpretasikan, namun Algoritma J48 memiliki beberapa kelemahan dalam proses klasifikasi seperti mengkonsumsi terlalu banyak waktu, J48 tidak memiliki kemampuan belajar tambahan yang baik, beberapa atribut yang tidak relevan menyebabkan efek buruk pada pembangunan pohon keputusan, dan kurangnya kemampuan belajar dari dataset tidak seimbang (Mardi, 2017).

*Feature selection* merupakan bagian penting untuk mengoptimalkan kinerja dari *classifier* (Harianto et al., 2020). Penggunaan *feature selection* yang tepat pada proses klasifikasi dapat meningkatkan hasil akurasi. Seleksi fitur dapat dilakukan dengan menghilangkan fitur yang tidak relevan dan redundan sehingga proses klasifikasi lebih optimal (Ferone, 2018). Ada beberapa metode yang dapat digunakan dalam seleksi fitur salah satunya yaitu *Correlation feature selection* (CFS). Fitur yang tidak relevan akan menurunkan performa *machine learning*, sedangkan fitur yang redundan akan membuat *machine learning* bekerja lebih lama (Heyong & Ming, 2019). Pemilihan subset fitur yang relevan dari sekumpulan fitur yang besar sangat penting untuk meningkatkan performa *machine learning* (Castellano, 2017).

Dalam proses pembelajaran algoritma *machine learning*, jika rasio kelas minoritas dan kelas mayoritas berbeda secara signifikan, *machine learning* cenderung didominasi oleh kelas mayoritas dan kelas minoritas sedikit dikenali. Sebagai hasilnya, klasifikasi akurasi kelas minoritas mungkin rendah bila dibandingkan keakurasian klasifikasi kelas mayoritas. Dengan menggunakan metode *Random Over Sampling* (ROS) dapat meningkatkan kelas minoritas sehingga kelas minoritas akan dapat dikenali, hal ini dapat meningkatkan kemampuan algoritma *machine learning* bekerja lebih baik, karena dapat mengenali sampel kelas minoritas dari sampel mayoritas (Li et al., 2018). Dalam menyelesaikan permasalahan ketidakseimbangan kelas Resample (ROS) akan memodifikasi data training dengan melakukan sampel ulang dataset asli, baik pada kelas minoritas ataupun kelas mayoritas agar seimbang (Syukron & Subekti, 2018).

*Boosting* adalah metode umum untuk meningkatkan kinerja algoritma belajar apapun. Salah satu algoritma boosting yang populer adalah Adaboost. AdaBoost adalah kependekan dari *Adaptive Boosting*. AdaBoost dapat digunakan untuk mengurangi kesalahan algoritma belajar "lemah" secara signifikan. Pada umumnya, metode boosting bisa meningkatkan ketelitian pada proses klasifikasi dan prediksi. Kerja boosting dalam meningkatkan ketelititian pada klasifikasi dan prediksi dengan cara membangkitkan kombinasi dari suatu model, tetapi hasil klasifikasi atau prediksi yang dipilih yaitu model yang memiliki bobot paling besar. sehingga, pada tiap model yang dibangkitkan memiliki atribut berupa nilai bobot (Listiana & Muslim, 2017).

Beberapa penelitian dalam menangani fitur yang tidak relevan dan ketidakseimbangan kelas pada data Diabetes Melitus telah dilakukan oleh para peneliti diantaranya adalah penelitian yang dilakukan oleh Hayashi & Yukita, hasil menunjukkan bahwa pengambilan sampel Re-RX dengan J48 memberikan aturan ekstraksi yang lebih akurat, ringkas, dan dapat

diinterpretasikan namun hasil akurasi yang di peroleh hanya mencapai 83,83% (Hayashi & Yukita, 2016). Selanjutnya Penelitian yang dilakukan oleh Bunkhumpornpat & Sinapiromsaran menggunakan DBSMOTE dengan J48, dalam penelitian ini dataset kelas minoritas dan kelas mayoritas diseimbangkan, namun nilai AUC yang dihasilkan sebesar 0.789 (Bunkhumpornpat & Sinapiromsaran, 2017) dan yang terakhir adalah penelitian dari Srinivas, pada penelitian ini menggunakan Algoritma J48 dengan OSID3 dan USRF hasil akurasi yang diperoleh dengan OSID3 sebesar 49,35% dan USRF 81,24% (Srinivas, 2017). Dalam penelitian-penelitian yang sudah dilakukan sebelumnya masih perlu adanya peningkatan akurasi dalam pengklasifikasian penyakit diabetes agar lebih optimal.

Pada Penelitian ini digunakan penerapan *Correlation Feature Selection* (CFS) sebagai seleksi atribut dan teknik *Resample* (*Random Over Sampling*) untuk menangani ketidakseimbangan kelas. Pada proses klasifikasi algoritma yang digunakan yaitu J48. Untuk meningkatkan kinerja J48 dalam pengklasifikasian maka di tambahkan Adaboost. AdaBoost dipilih untuk membantu meningkatkan nilai akurasi dengan cara memberi bobot pada tiap atribut, agar hasil kinerja algoritma J48 lebih optimal.

## METODE

Penelitian ini adalah penelitian eksperimen dengan menggunakan dataset pima indian diabetes. Penelitian ini bertujuan untuk menguji model klasifikasi yang terbaik untuk klasifikasi penyakit diabetes. Tahapan dari penelitian ini dimulai dari pengumpulan data, kemudian menentukan metode yang akan digunakan, eksperimen dan pengujian sampai hasil evaluasi dan validasi.

Pada tahap pengumpulan data, *Dataset* yang digunakan dalam penelitian ini berasal dari repositori publik *online* dari laman *www.archive.ics.uci.edu, Pima Indian population in Arizona, USA. National Institute of Diabetes and Digestive and Kidney Diseases.* Data pima diabetes melitus dengan jumlah data 768, jumlah atribut 8 dan terdiri dari 2 kelas yaitu kelas positif dan kelas negatif.

Metode yang digunakan pada penelitian ini menggunakan *Corelattion feature selection* (CFS) sebagai seleksi atribut pada data diabetes, CFS sebagai seleksi atribut akan menghilangkan atribut yang tidak relavan dan redundan, dimana atribut yang memiliki korelasi yang tinggi satu sama lain atribut tersebut menandakan redundan. Sedangkan atribut yang berkorelasi rendah pada kelas, atribut tersebut tidak relevan. Sehingga atribut yang tidak relevan dan redundan akan di hilangkan atau dihapus. Kemudian digunakan teknik *resample* (ROS) untuk menyeimbangkan kelas. Pada tahap *Resample* (*Random Over Sampling*) dalam menangani ketidakseimbangan kelas yaitu data kelas minoritas dipilih satu per satu secara acak, selanjutnya ditambahkan ke dalam data latih. Proses pemilihan dan penambahan ini dilakukan perulangan sampai jumlah data kelas minoritas sama dengan jumlah kelas mayoritas (seimbang). Selanjutnya dilakukan tahap klasifikasi menggunakan Algoritma J48 berbasis addabost, addabost digunakan untuk mengurangi kesalahan pada algoritma J48 dengan meningkatkan ketelitian pada proses klasifikasi.

Dalam melakukan eksperimen dan pengujian data menggunakan bantuan *tools Wekato Environment for knowletge Analysis* (WEKA). Proses awal dilakukan seleksi atribut pada data menggunakan CFS untuk menghilangkan atribut yang kurang mendukung. Kemudian dilakukan *Resample* untuk menghasilkan sample acak dari kumpulan data sehingga kelas minoritas seimbang dengan jumlah kelas mayoritas. Selanjutkan untuk proses pengklasifikasian menggunakan algoritma *decision tree* J48 dengan penambahan Adaboost untuk mengurangi kesalahan dari algoritma J48 dengan secara konsisten menghasilkan kinerja pengklasifikasi yang lebih baik. Selanjutnya dilakukan teknik pengujian menggunakan *10-fold cross validation.* Untuk mengetahui tingkat akurasi dari metode

klasifikasi decision tree J48 pada penelitian ini akan menggunakan evaluasi *conffusion matrix* untuk penilaian *accuracy*, *recall*, *precision* dan *AUC.*

## HASIL DAN PEMBAHASAN

### Hasil

Data uji yang digunakan dalam penelitian ini berupa dataset yang diambil dari repositori publik *online* dari laman *www.archive.ics.uci.edu*. yaitu data pima diabetes melitus dengan jumlah data sebanyak 768.

### Penerapan CFS *(Correlation Feature Selection)*

Pada penelitian ini dilakukan pengolahan awal yaitu dengan menggunakan CFS. Sebelum dilakukan perhitungan CFS, terlebih dahulu masing-masing atribut dihitung korelasinya antar artibut dengan kelas dan atribut dengan atribut. CFS akan memilih fitur yang relevan dan tidak redudan berdasarkan nilai merit yang tertinggi. Dalam penelitian ini untuk memilih subset atau sekumpulan atribut yang mempunyai nilai merit tertinggi menggunakan algoritma *forward selection search* (fss). *forward selection search* mencari subset dengan mula-mula atribut 0 kemudian ditambahkan satu atribut. Kemudian menghitung meritnya, dan seterusnya sampai semua atribut diujicoba. Kumpulan atribut dengan nilai merit tertinggi selanjutnya dipilih, dimana subset tersebut adalah hasil dari seleksi atribut berbasis korelasi (CFS). Berikut hasil rangkuman nilai merit tertinggi dari perhitungan berdasarkan rumus yang disajikan pada tabel 1.

Tabel 1. Rangkuman Nilai Merit Tertinggi

| No | Atribut | Nilai Merit |
|---|---|---|
| **1** | A6 | 0,272 |
| **2** | A6.A8 | 0,49 |
| **3** | A6.A8.A2 | 0,620 |
| **4** | A6.A8.A2.A7 | 0,764 |
| **5** | A6.A8.A2.A7.A3 | 0,736 |
| **6** | A6.A8.A2.A7.A3.A5 | 0,748 |
| **7** | A6.A8.A2.A7.A3.A5.A1 | 0,597 |
| **8** | A6.A8.A2.A7.A3.A5.A1.A4 | 0,608 |

Hasil tabel 1 menunjukan nilai merit tertinggi di peroleh pada nomer 4 yaitu atribut 6, atribut 8, atribut 2, dan atribut 7. Hal ini menunjukan bahwa atribut tersebut relevan terhadap kelas dan tidak redudan terhadap atribut lainnya, sehingga selain atribut 6, 7, 8 dan 2 akan dihilangkan karena atribut tersebut tidak relevan dan redundan.

### Penerapan *Resample (Random Over Sampling)*

Setelah melakukan proses CFS tahap selanjutnya dalam penelitian ini akan dilakukan teknik resample untuk menyeimbangkan kelas (kelas minoritas dengan kelas mayoritas) pada dataset diabetes. Tahap Resample (*Random Over Sampling*) dalam menangani ketidakseimbangan kelas pada data diabetes yaitu, langkah pertama pemilihan dataset kemudian hitung kelas mayoritas dan kelas minoritas, selanjutnya menghitung nilai selisih antara kelas mayoritas dengan kelas minoritas, diperoleh nilai selisih antara kelas mayoritas dan kelas minoritas yaitu sebanyak 232 record, kelompokan record kelas minoritas. Selanjutnya tambahankan record kelas minoritas satu per satu secara acak ke dalam data latih kelas minoritas dengan cara menduplikasi record data kelas minoritas. Penambahan record ini diulang sesuai dengan jumlah dari selisih, sehingga jumlah data kelas minoritas akan sama dengan jumlah data kelas mayoritas. Berikut hasil data setelah dilakukan teknik resample menggunakan tools WEKA dengan jumlah data baru sebanyak 1000 data (kelas mayoritas 500 dan kelas minoritas 500).

**Metode J48**

Pada penelitian ini metode klasifikasi yang digunakan adalah J48. Dataset yang digunakan yaitu data diabetes setelah dilakukan CFS dan Resample. Hasil pengukuran kinerja CFS dan Resample dengan metode klasifikasi J48 pada data diabetes, dengan hasil *accuracy* diperoleh sebesar 86.1 %, nilai *recall* sebesar 0,916, dan nilai *precision* sebesar 0,825. Nilai AUC yang didapat yaitu sebesar 0,913 yang ditunjukan pada tabel 2.

Tabel 2. Hasil Pengkuran *Accuracy*

| Pengukuran | Nilai |
|---|---|
| ***Accuracy*** | 86.1 % |
| ***Recall*** | 0,916 |
| ***Precision*** | 0,825 |
| ***AUC*** | 0,913 |

**J48 Berbasis Adaboost**

Hasil eksperimen dan pengujian dengan data diabetes yang sudah dilakukan proses CFS dan Resample menggunakan algoritma J48 berbasis adaboost menunjukan nilai accuracy sebesar 92,3%, nilai recall 0,942, nilai precision 0,908 dan AUC 0,959 yang ditunjukan pada tabel 3.

Tabel 3. Pengukuran Nilai *Accuracy, Recall, Precision,* dan *AUC*

| Pengukuran | Nilai |
|---|---|
| ***Accuracy*** | 92.3 % |
| ***Recall*** | 0,942 |
| ***Precision*** | 0,908 |
| ***AUC*** | 0,959 |

**Evaluasi dan Validasi Hasil**

Tabel 4. *Confussion Matrix*

| | **POSITIF** | **NEGATIF** |
|---|---|---|
| **POSITIF** | 471 (True Positive) | 29 (False Negative) |
| **NEGATIF** | 48 (False Positive) | 452 (True Negative) |

Berikut adalah perhitungan confussion matrix :

1. Nilai akurasi (*acc*) adalah proporsi jumlah prediksi yang benar dapat dihitung dengan rumus :

$$Akurasi = \frac{TP + TN}{TP + TN + FP + FN}$$
$$= \frac{471 + 452}{471 + 452 + 48 + 29}$$
$$= \frac{923}{1000} = 0{,}923\ (92{,}3\%)$$

2. *Sensitivity* (*recall*) atau *True Positive Rate* (TP *rate)* yaitu mengakui sebuah kasus yang diamati positif, dapat dihitung dengan rumus :

$$Recall = \frac{TP}{TP + FN}$$
$$= \frac{471}{471 + 29} = \frac{471}{500}$$

$$= 0{,}942\ (94{,}2\%)$$

3. *Precision* adalah tingkat ketepatan antara informasi yang diminta oleh pengguna dengan jawaban yang diberikan oleh sistem dapat dihitung dengan rumus :

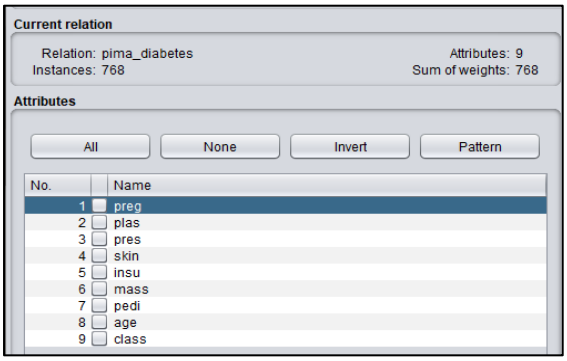
$$Precision\ Positive = \frac{TP}{TP + FP}$$
$$= \frac{471}{471 + 48}$$
$$= \frac{471}{519}$$
$$= 0{,}908$$

4. *AUC* adalah *Area Under the ROC (Receiver Operating Characteristic) Curve* (ROC atau AUC) adalah ukuran numerik untuk membedakan kinerja model dan menunjukkan seberapa sukses dan benar peringkat model dengan memisahkan pengamatan positif dan negatif**,** dapat dihitung dengan rumus:

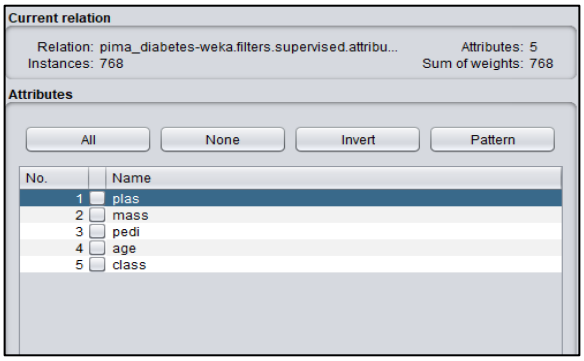
$$AUC = \frac{1 + TP_{rate} - FP_{rate}}{2}$$
$$= \frac{1 + 0.942 - 0.096}{2}$$
$$= 0{,}959$$

**Pembahasan**

Berdasarkan Pengujian model yang telah dilakukan dengan penerapan metode CFS (*Correlation feature selection*) dan Resample (ROS) pada data diabetes yang berjumlah 768 dengan menggunakan algoritma J48 berbasis Adaboost dengan tahap pengujian awal dilakukan proses CFS (*Correlation feature selection*) untuk menyeleksi atribut yang relevan dan tidak redundan pada dataset pima diabetes. Hasil seleksi atribut menggunakan CFS dapat dilihat pada gambar 1 dan gambar 2.




Gambar 1. Proses sebelum dilakukan CFS




Gambar 2. Proses setelah dilakukan CFS

Dari data yang sebelumnya berjumlah 8 atribut yaitu *Pregnant*, *Plasma-Glucose*, *Diastolic Blood-Pressure*, *Tricepts Skin Fold Thickness*, *Insulin*, *Body Mass Index*, *Diabetes Pedigree Function* dan *Age* diperoleh 4 atribut terbaik dari 8 atribut tersebut. 4 atribut yang terpilih berdasarkan pengujian menunjukan bahwa atribut tersebut relevan terhadap kelas dan tidak redundan terhadap atribut lain, 4 atribut yang terpilih berdasarkan pengujian yaitu *Plasma-Glucose*, *BMI (Body Mass Index)*, *Diabetes Pedigree Function* dan *Age* sedangkan atribut *Pregnant*, *Diastolic Blood-Pressure*, *Tricepts Skin Fold Thickness* dan *Insulin* tidak dipakai karena tidak relevan. Setelah terpilih atribut yang sudah relevan kemudian dilakukan

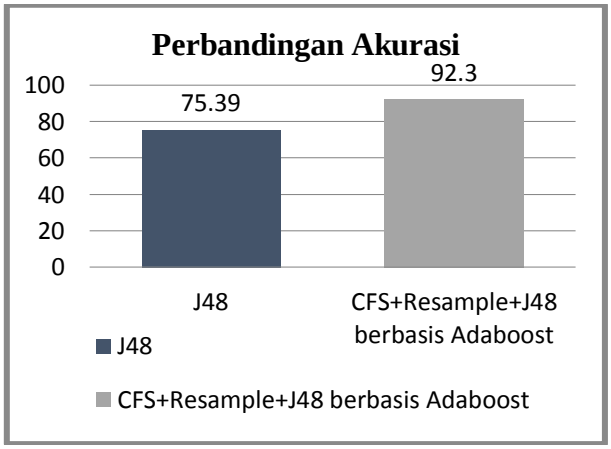
Resample (ROS). Resample (ROS) diterapakan untuk menangani ketidakseimbangan kelas pada data diperoleh data baru yaitu sejumlah 1000 data (500 kelas positif dan 500 kelas negatif) dari sebelumnya hanya berjumlah 768 data (500 negatif dan 268 positif) dapat dilihat pada Tabel 5.

Tabel 5. Dataset diabetes setelah dilakukan Resample (ROS)

| No | A2 (PG) | A6 (BMI) | A7 (DPF) | A8 (AGE) | KELAS |
|---|---|---|---|---|---|
| **1** | 102 | 30.8 | 0.400 | 26 | Negatif |
| **2** | 128 | 21.1 | 0.268 | 55 | Negatif |
| **3** | 57 | 21.7 | 0.735 | 67 | Negatif |
| **4** | 80 | 32.0 | 0.174 | 22 | Negatif |
| **5** | 143 | 42.4 | 1.076 | 22 | Negatif |
| **:** | : | : | : | : | : |
| **:** | : | : | : | : | : |
| **996** | 162 | 49.6 | 0.364 | 26 | Positif |
| **997** | 188 | 32.0 | 0.682 | 22 | Positif |
| **998** | 171 | 43.6 | 0.479 | 26 | Positif |
| **999** | 166 | 25.8 | 0.587 | 51 | Positif |
| **1000** | 143 | 36.6 | 0.254 | 51 | Positif |

Pengujian klasifikasi dilakukan menggunakan algoritma J48 berbasis Adaboost, hasil pengujian adaboost memberikan perbaikan bobot sehingga nilai bobot setelah dilakukan adaboost mengalami kenaikan dari nilai akurasi sebelum penerapan adaboost didapatkan nilai akurasi sebesar 86,6% setelah penerapan adaboost nilai akurasi yang dihasilkan menjadi 92,4 %. Dengan perbaikan bobot yang dilakukan oleh adaboost akan meminimalkan kesalahan dari pengklasifikasian yang dilakukan oleh algoritma J48 sehingga dapat meningkatkan nilai akurasi.

Perbandingan grafik hasil nilai akurasi yang diperoleh sebelum dan sesudah penerapan metode yang digunakan dalam penelitian ini pada data pima indian diabetes dapat dilihat pada gambar 3.

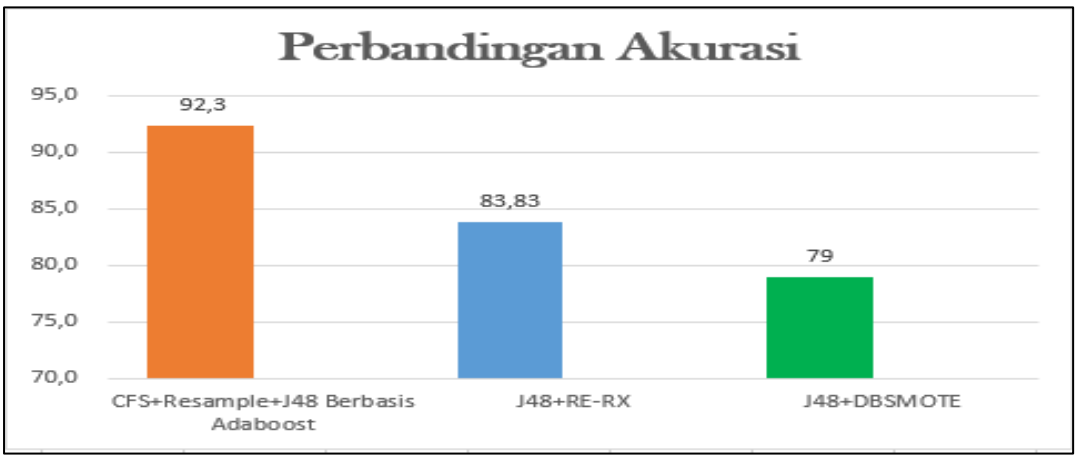



Gambar 3. Perbandingan Akurasi

Hasil nilai *accuracy, recall, precision* dan nilai AUC dengan penerapan metode CFS dan Resample pada algoritma J48 berbasis Adaboost dapat dilihat pada gambar 4. Selanjutnya, gambar 5 menunjukan grafik perbandingan tingkat akurasi hasil penerapan metode CFS+Resample dengan algoritma J48 berbasis adaboost dan hasil akurasi pada penelitian sebelumnya.

```
Time taken to build model: 0.36 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         923               92.3    %
Incorrectly Classified Instances        77                7.7    %
Kappa statistic                          0.846
Mean absolute error                      0.0764
Root mean squared error                  0.2684
Relative absolute error                 15.2749 %
Root relative squared error             53.6792 %
Total Number of Instances             1000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,904    0,058    0,940      0,904    0,922      0,847    0,959     0,958     0
                 0,942    0,096    0,908      0,942    0,924      0,847    0,959     0,947     1
Weighted Avg.    0,923    0,077    0,924      0,923    0,923      0,847    0,959     0,952

=== Confusion Matrix ===

   a   b   <-- classified as
 452  48 |   a = 0
  29 471 |   b = 1
```

Gambar 4. Hasil *Accuracy, Recall, Precision* dan *AUC* Pada Weka




Gambar 5. Perbandingan Nilai Akurasi dengan Penelitian Sebelumnya

Berdasarkan hasil pada gambar 5 menunjukan perbedaan hasil penelitian ini dengan penelitian sebelumnya yaitu terletak pada metode yang digunakan dan tingkat akurasi yang diperoleh. Pada beberapa penelitian sebelumnya yang menggunakan algoritma J48 didapatkan nilai akurasi yang masih belum optimal. Seperti penelitian yang dilakukan oleh (Hayashi & Yukita, 2016) dengan menggunakan *Reorder-Reaction (Re-RX)* dengan J48 menunjukan hasil klasifikasi yang cukup akurat. Namun, karena sifat rekursifnya *Reorder-Reaction (Re-RX)* cenderung menghasilkan lebih banyak aturan sehingga dalam proses pengklasifikasian cukup memakan waktu dan nilai akurasi yang didapat masih kurang optimal yaitu sebesar 83,83%. Penelitian lain yang dilakukan oleh (Bunkhumpornpat & Sinapiromsaran, 2017) dengan menggunakan DBSMOTE dengan J48, pada penelitian ini dataset kelas minoritas dan kelas mayoritas diseimbangkan, namun nilai akurasi yang dihasilkan masih sebesar 79%.

Berdasarkan dari beberapa penelitian sebelumnya yang sudah dilakukan bahwa nilai akurasi yang didapatkan masih kurang optimal. Sedangkan pada penelitian ini untuk mengoptimalkan algoritma J48 dalam pengklasifikasian penyakit diabetes digunakan metode seleksi atribut CFS untuk memilih atribut-atribut yang relevan sehingga atribut yang tidak relevan akan dipangkas dan tidak dipakai, sehingga dengan berkurangnya atribut yang

digunakan akan mempercepat proses pengklasifikasian. Setelah melakukan seleksi atribut pada dataset diabetes dilakukan penyeimbangan kelas menggunakan *Resample* (*Random Over Sampling*) dimana metode ini akan menjadikan kelas minoritas dan kelas mayoritas menjadi seimbang, sehingga kelas minoritas dapat dikenali. Setelah proses CFS dan Resample kemudian dilakukan perbaikan bobot menggunakan adaboost sehingga nilai bobot mengalami kenaikan.

Penggunaan metode CFS dan Resample pada J48 berbasis Adaboost untuk klasifikasi penyakit diabetes menunjukan hasil akurasi yang lebih optimal yaitu sebesar 92,3 %, hal ini menunjukan bahwa metode yang digunakan dalam penelitian ini dapat memperbaiki nilai akurasi dibandingkan dengan penelitian sebelumnya.

## SIMPULAN

Berdasarkan hasil penelitian yang telah dilakukan, diperoleh hasil bahwa penggunaan seleksi fitur dengan metode *CFS (Correlation Feature Selection)* dan *Resample (Random Over Sampling)* untuk menangani ketidakseimbangan kelas serta penambahan adaboost cukup berpengaruh terhadap nilai akurasi yang didapatkan pada algoritma J48, dengan menghasilkan nilai akurasi yang lebih baik yaitu sebesar 92,3%. Dengan penambahan metode CFS dan *Resample (Random Over Sampling)* serta adaboost dalam pengklasifikasian penyakit diabetes dapat meningkatkan kinerja Algoritma J48 sehingga membuat semakin lebih optimal dibandingkan tanpa menggunakan metode CFS dan *Resample (Random Over Sampling)* serta adaboost. Hal ini menunjukan penelitian yang telah dilakukan menghasilkan nilai akurasi yang lebih baik dan optimal untuk pengklasifikasian penyakit diabetes dibandingkan dari penelitian sebelumnya yang hanya menghasilkan akurasi 83,83%.

## REFERENSI


Bunkhumpornpat, C., & Sinapiromsaran, K. (2017). DBMUTE : density-based majority under-sampling technique. *Knowledge and Information Systems*, *50*(3), 827–850. https://doi.org/10.1007/s10115-016-0957-5

Castellano, J. G. (2017). Adaptative CC4.5 : Credal C4.5 with a rough class noise estimator. *Expert Systems With Applications*. https://doi.org/10.1016/j.eswa.2017.09.057

Ferone, A. (2018). Feature selection based on composition of rough sets induced by feature granulation. *International Journal of Approximate Reasoning*, *101*, 276–292. https://doi.org/10.1016/j.ijar.2018.07.011

Harianto et al. (2020). Optimasi Algoritma Naïve Bayes Classifier untuk Mendeteksi Anomaly dengan Univariate Fitur Selection. *Edumatic: Jurnal Pendidikan Informatika*, *4*(2), 40–49. https://doi.org/10.29408/edumatic.v4i2.2433

Hayashi, Y., & Yukita, S. (2016). Informatics in Medicine Unlocked Rule extraction using Recursive-Rule extraction algorithm with J48graft combined with sampling selection techniques for the diagnosis of type 2 diabetes mellitus in the Pima Indian dataset. *Informatics in Medicine Unlocked*, *2*, 92–104. https://doi.org/10.1016/j.imu.2016.02.001

Heyong, W., & Ming, H. (2019). Supervised Hebb rule based feature selection for text classification. *Information Processing and Management*, *56*, 167–191. https://doi.org/10.1016/j.ipm.2018.09.004

Kaunang, F. J. (2019). Penerapan Algoritma J48 Decision Tree Untuk Analisis Tingkat Kemiskinan di Indonesia. *CogITo Smart Journal*, *4*(2), 348–357. https://doi.org/10.31154/cogito.v4i2.141.348-357

Kemenkes. (2018). Riset Kesehatan Dasar (RIKESDAS).

Khairani. (2019). *Hari Diabetes Sedunia Tahun 2018*. Jakarta: Kementerian Kesehatan RI.

Li, F., Zhang, X., Zhang, X., Du, C., Xu, Y., & Tian, Y. C. (2018). Cost-sensitive and hybrid-attribute measure multi-decision tree over imbalanced data sets. *Information Sciences*,

*422*, 242–256. https://doi.org/10.1016/j.ins.2017.09.013

Listiana, E., & Muslim, M. A. (2017). Penerapan Adaboost Untuk Klasifikasi Support Vector Machine Guna Meningkatkan Akurasi Pada Diagnosa Chronic Kidney Disease. *Prosiding SNATIF*, 875–881.

Mardi, Y. (2017). Data Mining : Klasifikasi Menggunakan Algoritma C4.5. *Jurnal Edik Informatika*, *2*(2), 213–219.

Permana, B. A. C., & Dewi, I. K. (2021). Komparasi Metode Klasifikasi Data Mining Decision Tree dan Naïve Bayes Untuk Prediksi Penyakit Diabetes. *Infotek : Jurnal Informatika Dan Teknologi*, *4*(1), 63–69.

Srinivas, K. (2017). A Case Study On Class Imbalance Diabetes Data Using. *International Journal of Advances in Electronics and Computer Science*, (8), 13–16.

Syukron, A., & Subekti, A. (2018). Penerapan Metode Random Over-Under Sampling dan Random Forest Untuk Klasifikasi Penilaian Kredit. *Jurnal Informatika*, *5*(2), 175–185. https://doi.org/10.31311/ji.v5i2.4158

Takdirillah, R. (2020). Penerapan Data Mining Menggunakan Algoritma Apriori Terhadap Data Transaksi Sebagai Pendukung Informasi Strategi Penjualan. *Edumatic : Jurnal Pendidikan Informatika*, *4*(1), 37–46. https://doi.org/10.29408/edumatic.v4i1.2081