Penerapan *Linear Sampling* dan *Information Gain* pada Algoritma *Decision Tree* untuk Diagnosis Penyakit Diabetes

Gunawan[1], Ami Rahmawati[2], Satia Suhada[3], Taufik Hidayatulloh[4], Dede Wintana[5]

[1,4,5]Universitas Bina Sarana Informatika

[2,3]Sekolah Tinggi Manajemen Informatika dan Komputer Nusa Mandiri

[1]gunawan.gnz@bsi.ac.id, [2]ami.amv@nusamandiri.ac.id, [3]satia.shq@nusamandiri.ac.id, [4]taufik.tho@bsi.ac.id, [5]dede.dwe@bsi.ac.id

Diterima: 23 Juli 2021. Disetujui: 18 Desember 2021. Dipublikasikan: 25 Januari 2022.

Abstract - Diabetes which is assigned to be in the top 10 list of diseases that cause death in the last 10 years has increased. What was observed was that this increase occurred in developing countries with middle to lower social status. In Indonesia, diabetes is included in the top 10 diseases with a large number of sufferers. And more than that, diabetes becomes a comorbid that causes complications in Covid 19 patients. Then to detect diabetes more quickly and accurately, it is necessary to make research that can produce a better level of accuracy in order to detect diabetes. By using public dataset taken from the UCI repository consisting of 520 records, obtained from Diabetes Sylhet Hospital, Bangladesh. In this research, classification will be carried out using the Decision Tree Algorithm with optimization of Linear Sampling and Information Gain. After calculating using these methods and calculating the accuracy, the results obtained are 99.04% accuracy with a comparison with previous research which only used a Random Forest of 97.04%.

Keywords: Diabetes, Data mining, Linear Sampling, Information Gain

Abstrak-- Dalam rentang waktu 10 tahun terakhir diabetes ditetapkan telah masuk kedalam 10 daftar penyakit yang menyebabkan kematian yang mengalami kenaikan. Yang diamati kenaikan ini justru terjadi dinegara berkembang dengan status sosial menengah ke bawah. Di Indonesia diabetes masuk dalam 10 penyakit dengan jumlah penderita yang cukup banyak. Dan lebih daripada itu, diabetes menjadi komorbid yang menimbulkan komplikasi pada pasien covid 19. Maka untuk mendeteksi penyakit diabetes lebih cepat dan akurat, perlu dibuat penelitian yang dapat menghasilkan tingkat akurasi yang lebih baik agar dapat mendeteksi diabetes. Dengan menggunakan dataset publik yang diambil dari UCI repository yang terdiri dari 520 record, yang didapat dari RS Diabetes Sylhet, Bangladesh. Pada penelitian kali ini akan dilakukan klasifikasi menggunakan Algoritma *Decision Tree* dengan penerapan *linier sampling* dan *Information Gain*, dan setelah dilakukan perhitungan mengunakan metode tersebut dan dilakukan perhitungan akurasi maka diperoleh hasil akurasi sebesar 99,04% dengan perbandingan penelitian sebelumnya yang hanya menggunakan *Random Forest* sebesar 97,04%.

Kata kunci: Diabetes, Data mining, Linear Sampling, Information Gain

I. PENDAHULUAN

Selama kurun waktu dua dekade terakhir, diabetes masuk ke dalam 10 penyebab kematian teratas dan mengalami peningkatan sebanyak 70% sejak tahun 2000 [1]. Angka peningkatan tersebut justru terjadi pada negara berkembang dengan pendapatan menengah ke bawah [2]. Pada tahun 2019, *International Diabet Federation* (IDF) menuturkan jika pengidap diabet di dunia diperkirakan mencapai sedikitnya 463 juta orang pada rentang usia 20 tahun - 79 tahun. Jumlah penderita diabet ini diprediksi akan terus meningkat hingga 578 juta jiwa pada tahun 2030 dan 700 juta jiwa pada tahun 2045 [3]. Pada tahun 2019 Indonesia masuk dalam peringkat 10 besar dunia, sebagai negara dengan jumlah penderita diabet terbanyak dengan total sebanyak 10,7 juta orang. Pada penelitian [4], dan [8], hasil perbandingan algoritma diketahui bahwa Algoritma *Decision Tree* menggunakan teknik *cross-validation* menghasilkan akurasi sebesar 95,6%. Sementara itu, dataset publik yang sama juga diuji menggunakan algoritma *Decision Tree* C4.5 dan menghasilkan nilai akurasi sebesar 97,12% serta nilai AUC sebesar 0,994 [7]. Selain itu pada penelitian sebelumnya, Algoritma *Decision Tree* juga digunakan untuk menguji dataset *private* pada RSUP Dr. Sardjito [6].

Algoritma *Decision Tree* pada penelitian tersebut menghasilkan tingkat akurasi klasifikasi sebesar 88,42%. *Decision Tree* memilki kelemahan jika jumlah kelas dan atribut berjumlah banyak sehingga dapat menimbulkan overplap dan salah satu solusinya ada menggunakan *Random Forest*, namun *Random Forest* juga bisa kurang optimal jika menggunakan dataset yang tidak seimbang [4]. Merujuk hasil dari penelitian-penelitian tersebut, maka pada penelitian kali ini dataset publik tersebut akan kembali diuji menggunakan algoritma *Decision Tree* karena algoritma tersebut memiliki tingkat akurasi yang cukup tinggi, ujicoba ini akan dilakukan dengan penerapan linear sampling dan information gain.

II. METODE PENELITIAN

A. Tahapan Penelitian

Dalam penelitian ini digunakan model CRISP-DM *(Cross Industry Standard Process for Data Mining)* sebagai metode penelitian yang terdiri dari enam tahap, yaitu *business understanding, data understanding, data preparation, modeling, evaluation,* dan *deployment*.

1. Tahap *Business Understanding*

Pada tahapan ini difokuskan untuk penentuan tujuan dari penelitian yaitu menerapkan algoritma *Decision Tree* dan *dicision tree* menggunakan *linier sampling* dan *information gain* untuk mendiagnosis penyakit diabetes.

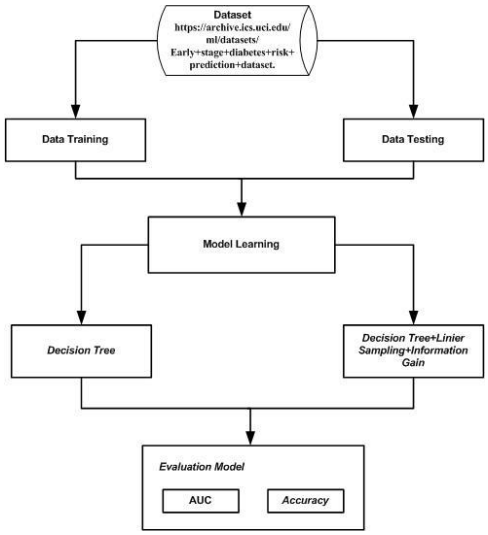
2. Tahap *Data Understanding*

Pada tahapan ini dilakukan proses pengumpulan data secara sekunder melalui data *repository UCI*, kemudian data diperiksa untuk mengetahui struktur data, semua data normal dan dapat digunakan untuk melanjutkan penelitian sehingga analisis data tidak banyak dilakukan dan evaluasi kualitas data pada dataset sangat baik. Data tersebut memiliki 520 *record* dengan 16 atribut prediktor dan 1 atribut hasil.

3. Tahap *Data Preparation*

Pada tahapan ini dilakukan pengolahan data penyakit diabetes, jumlah data yang diperoleh pada penelitian ini sebanyak 520 *record* yang terdiri dari pasien positif dan negatif diabetes. Dalam data ini tidak dilakukan *preprocessing* dikarenakan data sudah normal.

4. Tahap *Modelling*



Gambar 1. Model yang Diusulkan

Pada gambar 1 dijelaskan tahapan *modelling* yang dilakukan uji data sesuai dengan model yang akan diusulkan dengan menggunakan *tools RapidMiner* 9.6 dengan menggunakan Algoritma *Decision Tree* serta *linier sampling* dan *information gain* pada *data training* dan *data testing*.

5. Tahap *Evaluation*

Pada tahapan ini dilakukan evaluasi terhadap model yang terbentuk dengan menggunakan *confussion matrix* untuk mengetahui nilai akurasi dan AUC.

6. Tahap *Deployment*

Berdasarkan model yang telah terbentuk maka perlu diuji dengan menggunakan data baru dan dilakukan kembali evaluasi untuk keakuratan data.

B. Landasan Teori

1. Penyakit Diabetes

Diabetes merupakan penyakit non infeksi yang paling popular tidak hanya di Indonesia tetapi juga diseluruh dunia. Diabetes ini adalah suatu penyakit *metabolic* akibat gangguan *metabolism* karbohidrat yang ditandai dengan kadar gula di dalam darah yang tinggi *(hiperglikemia)*, sehingga disebut juga dengan sakit gula [9]. Pada penderita diabetes, ada gangguan keseimbangan antara transportasi gula ke dalam sel, gula yang disimpan di hati, dan gula yang dikeluarkan dari hati. Akibatnya, kadar gula dalam darah meningkat. Kelebihan ini keluar melalui *urine*. Oleh karena itu, *urine* menjadi banyak dan mengandung gula. Penyebab keadaan ini hanya dua.

Pertama, *pankreas* tidak mampu lagi memproduksi insulin. Kedua, sel tidak memberikan respons pada kerja insulin selaku kunci buat membuka pintu sel sehingga gula tidak bisa masuk ke dalam sel [10].

2. Data Mining

Data mining didefinisikan sebagai sebuah teknik untuk menggali pola (*pattern*) dari sebuah data yang berjumlah besar. Pola yang telah ditemukan tersebut akan berguna dalam mendukung pengambilan keputusan yang lebih cepat dan akurat [11]. Selain itu, pola tersebut juga memungkinkan kita untuk membuat sebuah prediksi data yang baru. Berikut ini merupakan peran utama dalam *data mining* [12]:

a. Deskripsi

Data mining dapat mendeskripsikan/ menggambarkan sebuah pola ataupun tren yang ada dalam sebuah data. Analisis dan eksplorasi sejumlah data dapat menghasilkan sebuah deskripsi atau gambaran yang baik, terutama jika dapat digambarkan menggunakan metode grafis.

b. Estimasi

Metode estimasi mirip dengan klasifikasi, hanya saja untuk estimasi label atau classnya berbentuk numerik dan bukan berupa kategorikal. Contohnya seperti memperkirakan jumlah poin dalam suatu pertandingan atau memperkirakan nilai rata-rata (IPK) pasca sarjana berdasarkan nilai rata-rata pendidikan sebelumnya.

c. Prediksi

Prediksi mirip dengan klasifikasi dan estimasi, akan tetapi untuk prediksi nilainya terletak di masa depan, kemudian datanya bersifat rentet waktu serta *class*-nya bersifat numerikal. Contohnya prediksi untuk harga saham beberapa bulan ke depan atau memprediksi persentase kenaikan kematian lalu lintas tahun depan jika terjadi peningkatan batas kecepatan.

d. Klasifikasi

Ciri khas dari sebuah klasifikasi, atribut datanya biasanya dapat berupa nilai numerik atau nominal, sedangkan classnya bersifat nominal. Contohnya menentukan kecurangan dalam suatu transaksi kartu kredit, menilai seorang calon nasabah yang memiliki resiko kredit lancar atau macet, atau juga mendiagnosis suatu penyakit tertentu.

e. Klastering

Untuk klastering mengacu pada sebuah kelompok *record*, hasil pengamatan atau sebuah kasus ke dalam kelas objek yang serupa. Klastering tidak mencoba untuk mengklasifikasikan, memperkirakan ataupun memprediksi. Contohnya klaster bunga iris, klastering jenis pelanggan, serta klastering sentimen warga.

f. Asosiasi

Tugas utama dari asosiasi adalah menemukan atribut mana yang paling sering muncul secara bersamaan. Dalam dunia bisnis paling sering dijumpai pada analisis keranjang belanja. Asosiasi berupaya mengungkap aturan tentang hubungan antara dua atribut atau lebih.

3. Klasifikasi

Klasifikasi dapat didefinisikan selaku pekerjaan yang melaksanakan pelatihan/ pembelajaran terhadap fungsi target yang memetakan tiap set fitur (*atribut*e) ke satu jumlah label kelas yang ada. Klasifikasi ialah suatu buat memperhitungkan objek informasi buat memasukannya kedalam kelas tertentu dari jumlah kelas yang ada. Klasifikasi menerapkan pembangunan model bersumber pada informasi latih yang terdapat, setelah itu memakai model tersebut untuk dibuatkan klasifikasi pada informasi yang baru [13]. Klasifikasi memiliki 4 komponen dasar yaitu: 1) *class*, merupakan variabel yang menjadi label atau hasil suatu objek; 2) *predictor*, merupakan variabel yang menjadi atribut dari data yang akan digunakan pada klasifikasi; 3) *training dataset*, merupakan data yang tekah memiliki label sebelumnya; 4) *testing dataset*, merupakan data baru yang akan dilakukan proses klasifikasi [14].

4. Algoritma *Decision Tree*

Algoritma *Decision Tree* adalah model prediktif yang dapat digunakan untuk mewakili model pengklasifikasi dan regresi. Dalam penelitian, operasi *Decision Tree* mengacu pada model hirarki keputusan. Pohon klasifikasi sering digunakan dalam bidang terapan seperti keuangan, pemasaran, teknik dan kedokteran [15]. Selain itu *Decision Tree* merupakan *supervised learning algorithm* yang mampu menganalisis data numerik dan kategorikal.

Selama proses klasifikasi atau regresi, model struktur pohon akan dibangun berdasarkan kumpulan data yang dipilih. Struktur pohon akan menjadi besar ketika data dipecah menjadi bagian-bagian yang lebih kecil. Dalam *Decison Tree* terdapat tiga simpul yang terlibat yaitu: simpul keputusan, simpul daun dan simpul akar. Simpul daun mewakili klasifikasi atau keputusan, sedangkan Simpul keputusan terdiri dari dua atau lebih cabang dan Simpul akar adalah simpul keputusan yang ditempatkan paling atas dari pohon. Simpul akar mewakili sebagai prediktor terbaik di

antara simpul keputusan lainnya. Kemampuan ini menjadikan pohon keputusan sebagai salah satu teknik yang dapat mengklasifikasikan data multi-kategori [16].

Dalam *Decision Tree* , node dapat memiliki lebih dari dua anak atau cabang tergantung pada kondisi pengujian atribut dan atribut yang dipilih. Untuk memisahkan node, langkah-langkah pemilihan atribut dengan berbagai implementasi diterapkan. Ukuran pemilihan atribut dalam node yang sama juga dapat bervariasi untuk cabang biner atau cabang *multiway* [17]. Beberapa ukuran pemilihan atribut yang umum diantaranya:

a. *Entropy* : Diterapkan secara hierarki di setiap tingkat pohon keputusan untuk menjumlahkan keragaman yang ditunjukkan oleh nilai atribut target di setiap kategori, dan disimpulkan menggunakan satu hingga banyak atribut pendukung di tingkat hierarki yang lebih tinggi [18]. Rumus *Entropy* terlihat pada rumus 1:

$$Entropy\ (S) = \sum_{i=1}^{c} - P_i log_2 P_i \qquad (1)$$

dengan *S* adalah himpunan kasus. *C* adalah jumlah partisi *S*. *Pi* adalah proporsi dari *Si* terhadap *S*.

b. *Gain* : Untuk mengukur efektivitas masing-masing atribut kriteria dalam mengklasifikasikan data. Pada algoritma *Decision Tree* nilai *Gain* digunakan sebagai dasar pembentukan node atau akar dan cabang pohon keputusan [19]. Rumus Gain terlihat pada rumus 2:

$$Gain\ (S, A) = Entropy\ (S) - \sum_{v \in Values\ (A)} \frac{|Sv|}{|S|} Entropy(Sv)_i \qquad (2)$$

dengan *A* adalah variabel. *V* adalah nilai yang mungkin untuk variabel *A*. *Sv* adalah jumlah sampel untuk nilai v. *S* adalah jumlah sampel untuk seluruh sampel data. $Entropy(Sv)$ adalah entropy untuk sampel yang memiliki nilai v.

c. *Information Gain* : *Information Gain* didasarkan pada *Entropy*. Perbedaan antara *Entropy* kelas dan *Entropy* bersyarat kelas dan fitur yang dipilih. Ini mengukur kegunaan fitur f dalam klasifikasi yaitu, perbedaan *Entropy* dari sebelum ke setelah pemisahan himpunan L pada fitur f. Dengan kata lain, *Information Gain* mengukur pengurangan ketidakpastian setelah memisahkan himpunan pada fitur. Jika nilai *Information Gain* meningkat, berarti fitur f lebih berguna untuk klasifikasi [20]. Rumus *Information Gain* seperti yang terlihat pada rumus 3:

$$IG\ (L, F) = Entropy\ (L) - \sum_{v=1}^{V} \frac{|Lv|}{|L|} (\text{Entropy (Lv)}) \qquad (3)$$

dengan asumsi bahwa *v* adalah nilai berbeda untuk fitur *f*. |*Lv*| adalah mewakili subset *L* dengan *f=v*.

5. Kurva ROC – AUC (*Area Under Curve*)

Kurva ROC (*Receiver Operating Characteristics*) merupakan perbandingan visual untuk model klasifikasi. Selain itu kurva ini dapat mengukur serta menggambarkan kinerja sebuah klasifikasi [21]. Untuk mengukur nilai kurva ROC digunakanlah teknik AUC (*Area Under Curve*) sebagai berikut:

a. 0.90-1.00 = *Execellent Classification*

b. 0.80-0.90 = *Good Classification*

c. 0.70-0.80 = *Fair Classification*

d. 0.60-0.70 = *Poor Classification*

e. 0.50-0.60 = *Failure*

III. HASIL DAN PEMBAHASAN

A. Dataset

Dataset yang digunakan dalam penelitian ini ialah dataset publik yang didapatkan dari website https://ics.uci.edu dataset dapat di unduh melalui link: https://archive.ics.uci.edu/ml/datasets/Early+stage+diabetes+risk+prediction+dataset.. Dataset ini dikumpulkan di rumah sakit Diabetes Bangladesh melalui kuesioner yang disebar dan di isi oleh 520 partisipan yang mengalami penyakit dan gejala diabetes [8]. Berikut ini merupakan deskripsi dataset diabetes seperti terlihat pada Tabel 1 dan Tabel 2.

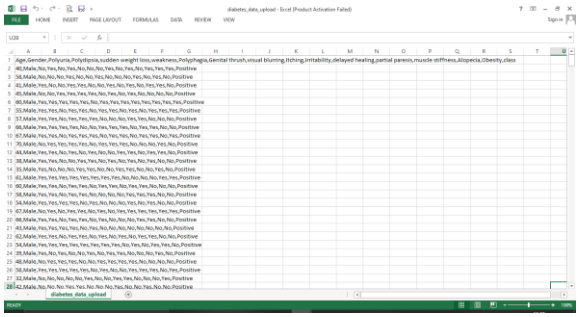
TABEL I. DESKRIPSI DATASET DIABETES

Dataset Diabetes	Jumlah Atributes	Jumlah Instance
	16	520

TABEL II. DESKRIPSI ATRIBUT DATASET

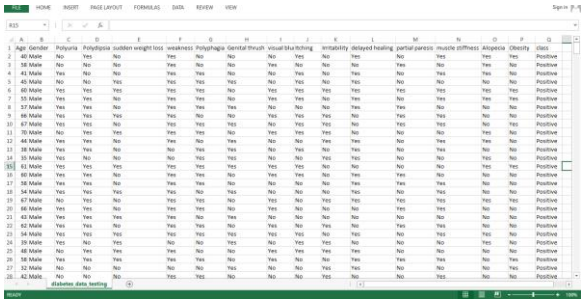
Atributes	Value
Age	1.16-90
Sex	1.Male, 2.Femlae
Polyuria	1.Yes, 2.No
Polydipsia	1.Yes, 2.No
Sudden weight loss	1.Yes, 2.No
Weakness	1.Yes, 2.No
Polyghagia	1.Yes, 2.No
Genital thrush	1.Yes, 2.No
Visual blurring	1.Yes, 2.No
Itching	1.Yes, 2.No
Irritability	1.Yes, 2.No
Delayed healing	1.Yes, 2.No
Partial paresis	1.Yes, 2.No
Muscle stiffness	1.Yes, 2.No
Alopecia	1.Yes, 2.No
Obesity	1.Yes, 2.No
Class	1.Positive, 2.Negative.

Setelah dataset berhasil diunduh dari website https://ics.uci.edu dataset berupa file coma delimiter csv seperti yang terlihat pada gambar 2 berikut.



Gambar 2. Dataset Diabetes format CSV

Dataset kemudian dirubah format filenya dari csv ke dalam format excel 97-2003 seperti pada gambar 3.

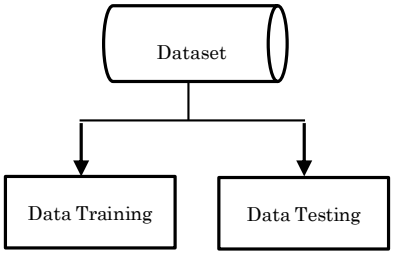


Gambar 3. Dataset Diabetes format Excel

B. *Implementasi*

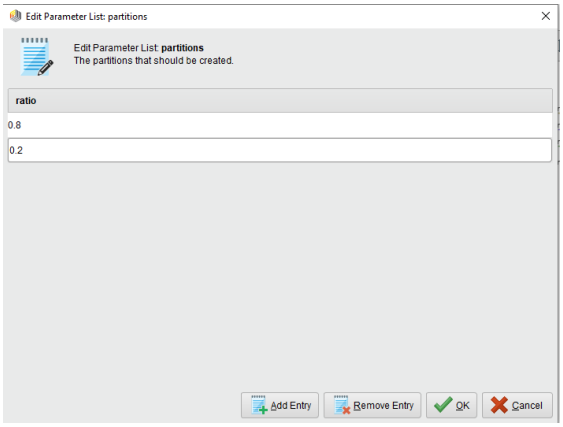
Uji coba dilakukan setelah dataset dirubah kedalam format excel. Ujicoba dengan melakukan *training* sebanyak dua kali terhadap dataset diabetes menggunakan aplikasi *Rapid Miner* 9.6 untuk mengetahui performa terbaik dari Algoritma *Decision Tree* dengan melihat perbandingan hasil ujicoba pertama dan kedua.

Uji coba pertama dilakukan *training* menggunakan Algoritma *Decision Tree* secara *default* Rapid Miner 9.6 tanpa melakukan perubahan properties apapun pada Algoritma *Decision Tree* dan uji coba kedua *training* pada dataset diabetes dengan *Decision Tree* menggunakan *linier sampling* yaitu membagi sampel menjadi dua partisi tanpa merubah urutan dari sampelnya.

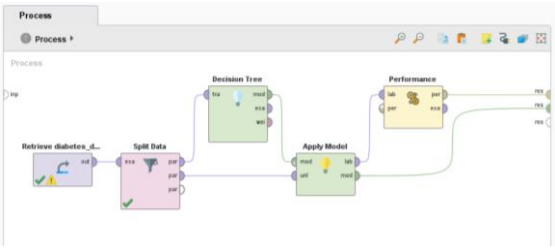


Gambar 4. Metode *Linier Sampling*

Dataset pada tahap *training* dibagi menjadi dua partisi yaitu 80% untuk data *training* dan 20% untuk data *testing*, seperti yang terlihat pada gambar 4 dan visualisasi pembagian data terlihat pada gambar 5. Sehingga pembagian dataset menjadi 416 data *training* dan 104 data *testing*.



Gambar 5. Split Data *Training* dan *Testing*



Gambar 6. Proses Model Rapid Miner *Decision Tree* Diabetes

Gambar 6 menjelaskan model ujicoba pertama dataset yang di bagi menjadi data *training* dan data *testing* dengan pembagian 80% data *training* dan 20% data *testing*. Data *training* yang 80% kemudian di-*training* menggunakan Algoritma *Decision Tree* secara *default* kemudian model yang telah dibangun dengan ujicoba data *training* diterapkan terhadap data *testing* dan diukur performanya.

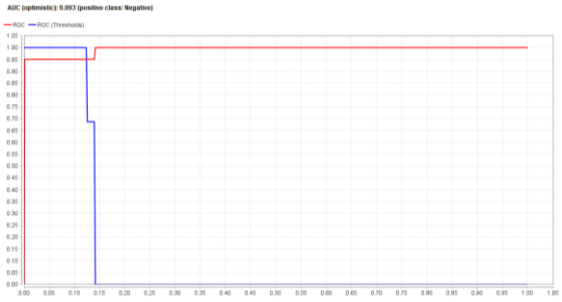
Hasil model dari uji coba pertama menunjukan performa klasifikasi *Decision Tree* tanpa menggunakan optimasi apapun, hasil akurasi yang didapatkan dari uji coba pertama ini hanya menghasilkan klasifikasi dengan akurasi sebesar 90,38% dengan sebaran kelas negatif dan positif sebagai berikut:

TABEL III. CONFUSION MATRIX *DECISION TREE* DATA SET DIABETES UJI COBA PERTAMA

Klasifikasi yang benar	**Diklasifikasikan sebagai**	
	Positif	Negatif
Positif	56	2
Negatif	8	38

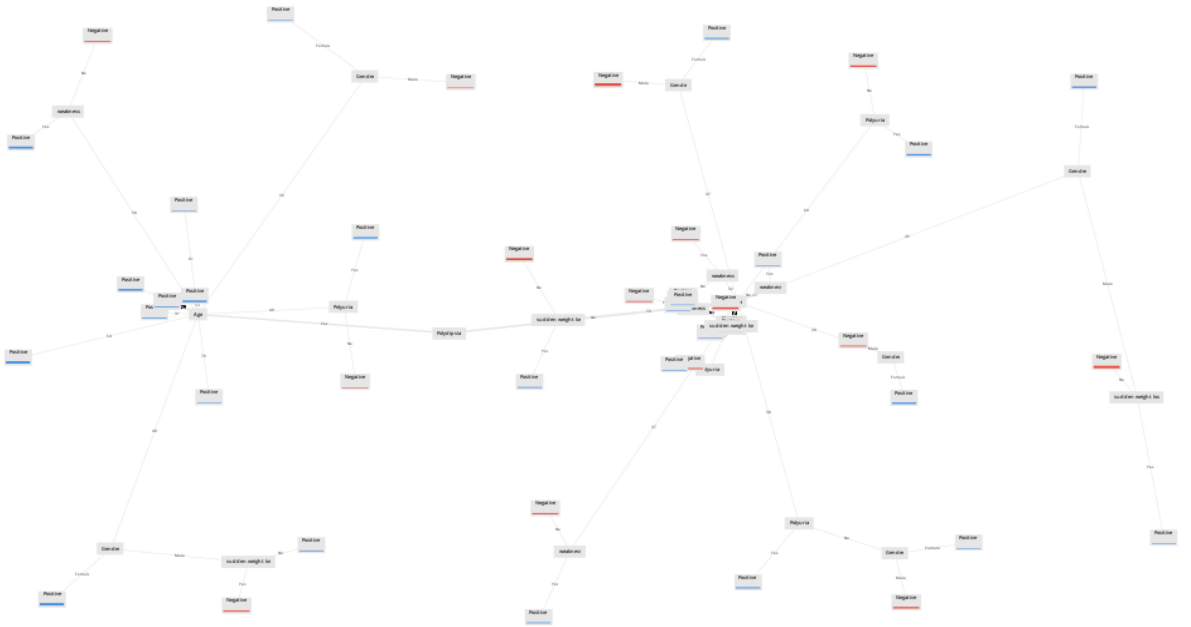
Berdasarkan Tabel 3 hasil ujicoba pertama *Decision Tree* secara default terhadap dataset diabetes dari 104 data *testing* dapat diketahui bahwa sebanyak 56 data positif diklasifikasikan sebagai positif dan sebanyak 38 data negatif diklasifikasikan sebagai negatif, dari hasil uji coba pertama masih ditemukan klasifikasi yang salah yaitu 2 data positif diklasifikasikan sebagai negatif dan 8 data negatif diklasifikasikan sebagai positif.

Selain menggunakan *confusion matrix* untuk mengukur performa hasil ujicoba pertama, dilakukan juga pengukuran performa menggunakan *area under curve* atau AUC dengan hasil seperti yang terlihat pada gambar 7.



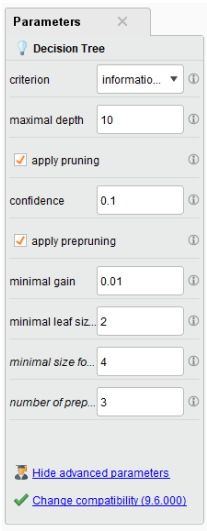
Gambar 7. Nilai AUC *Decision Tree* Uji Coba Pertama

Nilai AUC yang didapatkan dari uji coba pertama algoritma *Decision Tree* secara *default* terhadap dataset diabetes adalah 0,9 dengan nilai diagnosa *excellent classification*.



Gambar 8. *Decision Tree* Ujicoba Pertama

Gambar 8 menunjukkan sebaran *root* dari algoritma *Decision Tree* untuk ujicoba pertama. Setelah uji coba pertama dilakukan selanjutnya ujicoba kedua dengan menerapkan algoritma *Decision Tree* dengan *linier sampling* dan *Decision Tree*. Pembagian dataset untuk data *training* dan data *testing* dilakukan menggunakan *linier sampling* sehingga didapatkan pembagian dataset 416 untuk data *training* dan 104 untuk data *testing*. Pemilihan kriteria dari dataset diabetes pada *Decision Tree* menggunakan *information gain* untuk mencari fitur terbaik untuk diklasifikasikan serta menggunakan pruning dengan *confidence 0,1*. Seperti yang terlihat pada gambar 9.



Gambar 9. Properties *Decision Tree*

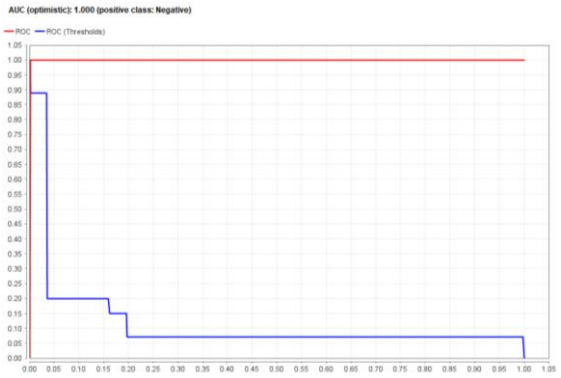
Hasil dari model yang dibangun kemudian diukur performanya dengan akurasi prediksi dan *Area Under Curve* (AUC).

TABEL IV. *CONFUSION MATRIX DECISION TREE* DATASET DIABETES UJICOBA KEDUA

Klasifikasi yang benar	**Diklasifikasikan sebagai**	
	Positif	Negatif
Positif	55	0
Negatif	1	48

Berdasarkan Tabel 4 dari 104 *data testing* dapat diketahui bahwa sebanyak 55 data positif diklasifikasikan sebagai positif dan sebanyak 48 data negatif diklasifikasikan sebagai negatif, 0 data positif diklasifikasikan sebagai negatif dan 1 data negatif diklasifikasikan sebagai positif. Sehingga didapatkan nilai akurasi untuk model yang dibangun adalah 99,04% hasil dari perhitungan:

$$Accuracy = \frac{55+48}{55+0+1+48} \text{ x } 100\% = 0{,}990$$



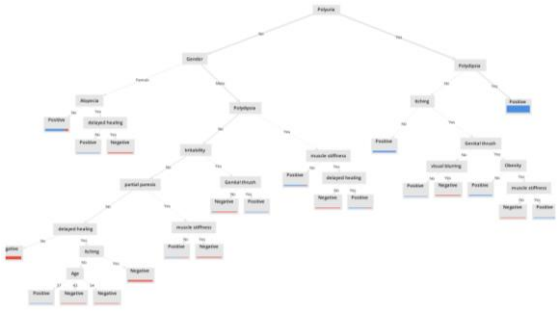
Gambar 10. Nilai AUC *Decision Tree* Ujicoba Kedua

Nilai AUC yang didapatkan dari model yang diusulkan adalah 1 dengan nilai diagnosa *excellent classification*. Dapat dilihat pada gambar 10. Kemudian untuk sebaran *root* pada percobaan kedua dapat dilihat pada gambar 11.

TABEL V. PERBANDINGAN PERFORMA MODEL DENGAN PENELITIAN SEBELUMNYA

Penelitian	**Metode**	**Akurasi**
Islam (2019) [8]	*Random Forest*	97,04%
Model *Decision Tree*	*Decision Tree*	90,38%
Model yang diusulkan	*Decision Tree +Linear Sampling+Information Gain*	99,04%

Perbandingan performa dari model yang diujicobakan dengan penelitian sebelumnya dapat dilihat pada Tabel 5 dengan akurasi yang didapatkan 99,04%.



Gambar 11. *Decision Tree* Diabetes

IV. KESIMPULAN DAN SARAN

Algoritma *Decision Tree* dapat menghasilkan sebuah klasifikasi yang baik karena dapat menemukan hubungan yang saling berkaitan antara

variabel dan atribut serta digambarkan dalam bentuk pohon keputusan. Pada penelitian sebelumnya yang hanya menggunakan algoritma *Decision Tree* saja akurasi yang didapat sebesar 90,38%. Sedangkan penelitian yang dilakukan kali ini dengan menghitung optimalisasi menggunakan metode *Linear Sampling* dan *Information Gain*, algoritma *Decision Tree* ternyata mendapatkan peningkatan akurasi yang lebih baik dari hasil sebelumnya dengan hasil 99,04%.

Saran untuk penelitian selanjutnya, optimasi algoritma *Decision Tree* dapat diimplementasikan pada sebuah aplikasi, sehingga pohon keputusan yang telah dihasilkan dapat dirasakan langsung manfaatnya, terutama oleh siapapun yang akan melakukan deteksi dini penyakit diabetes.

REFERENSI

[1] WHO, "The Top 10 Causes of Death," World Health Organization, 2020.

[2] WHO, "Diabetes," World Health Organization, 2020.

[3] KEMENKES RI, "INFODATIN Pusat Data dan Informasi Kementerian Kesehatan RI," Kementerian Kesehatan Republik Indonesia, Jakarta Selatan, 2020.

[4] Pangastuti, S. S. (2018). *Perbandingan Metode Ensemble Random Forest Dengan Smote-Boosting Dan Smote-Bagging Pada Klasifikasi Data Mining Untuk Kelas Imbalance (Studi Kasus: Data Beasiswa Bidikmisi Tahun 2017 di Jawa Timur)-A Comparison Of The Ensemble Random Forest Methods With Smote-Boosting And Smote-Bagging On Data Mining Classification For Imbalance Class* (Doctoral dissertation, Institut Teknologi Sepuluh Nopember).

[5] N. Nurdiana and A. Algifari, "Studi Komparasi Algoritma ID3 dan Algoritma Naive Bayes Untuk Klasifikasi Penyakit Diabetes Mellitus," *INFOTECH Journal*, pp. 18-23, 2020.

[6] M. F. Salim and S. , "Analisis Rekam Medis Pasien Diabetes Mellitus Melalui Implementasi Teknik Data Mining di RSUP Dr. Sardjito Yogyakarta," *JKesV - Jurnal Kesehatan Vokasional*, pp. 167-174, 2017.

[7] F. M. Hana, "Klasifikasi Penderita Penyakit Diabetes Menggunakan Algoritma *Decision Tree* C4.5," *Jurnal Sistem Komputer dan Kecerdasan Buatan* , pp. 32-39, 2020.

[8] M. M. F. Islam, R. Ferdousi, S. Rahman and H. Y. Bushra, "Likelihood Prediction of Diabetes at Early Stage Using Data Mining Techniques," in *Computer Vision adn Machine Intelligence in Medical Image Analysis*, 2019.

[9] I. M. P. Dwipayana and I. M. S. Wirawan, Tanya Jawab Seputar Kencing Manis (Diabetes Mellitus) dan Sakit Maag (Gastritis), Ponorogo: Uwais Inspirasi Indonesia, 2018.

[10] H. Tandra, Segala Sesuatu yang harus Anda Ketahui Tentang Diabetes Panduan Lengkap Mengenal dan Mengatasi Diabetes dengan Cepat dan Mudah Edisi Kedua dan Paling Komplit, Jakarta: PT Gramedia Pustaka Utama, 2017.

[11] I. H. Witten, E. Frank, M. A. Hall and C. J. Pal, Data Mining - Practical Machine Learning Tools and Techniques - Fourth Edition, Chennai: Elsevier, 2017.

[12] D. T. Larose, Discovering Knowledge in Data, New Jersey: Wiley-Interscience, 2005.

[13] D. P. Utomo and M. , "Analisis Komparasi Metode Klasifikasi Data Mining dan Reduksi Atribut Pada Dataset Penyakit Jantung," *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 4, no. 2, pp. 437-444, 2020.

[14] A. P. Ayudhitama and U. Pujianto, "Analisa 4 Algoritma dalam Klasifikasi Penyakit Liver Menggunakan Rapid Miner," *JIP (Jurnal Informatika Polinema)*, vol. 6, no. 2, pp. 1-9, 2020.

[15] L. Rokach and O. Maimon, Data Mining With *Decision Tree* s Theory and Applications 2nd Edition, Singapore: World Scientific Publishing, 2015.

[16] H. Fujita and A. Selamat, Advancing Technology Industrialization Through Intelligent Software Methodologies, Tools and Techniques, Netherlands: IOS Press BV, 2019.

[17] B. Makhabel, Learning Data Mining with R, Birmingham, UK: Packt Publishing, 2015.

[18] X. Li and C. Claramunt, "A Spatial Entropy-Based *Decision Tree* for Classification of Geographical Information," *Transition in GIS*, vol. 10, no. 3, pp. 451-467, 2006.

[19] E. Buulolo, Data Mining Untuk Perguruan Tinggi, Yogyakarta: Deepublish, 2020.

[20] S. Tangirala, "Evaluating the Impact of GINI Index and Information Gain on Classification using *Decision Tree* Classifier Algorithm," *(IJACSA) International Journal of Advanced Computer Science and Applications*, vol. 11, no. 2, pp. 612-619, 2020.

[21] S. Bahri, A. Wibowo, R. Wajhillah and S. Suhada, Data Mining; Algoritma Klasifikasi dan Penerapannya Dalam Aplikasi, Yogyakarta: Graha Ilmu, 2019.