Pengaruh Perbedaan Jumlah *Hidden Layer* dan *Node* pada *Hidden Layer* terhadap Performa Model Klasifikasi Diabetes

I N A Suprana[*1]

[1] Universitas Atma Jaya Yogyakarta

E-mail: supranaarya@gmail.com[1]

Abstrak. Diabetes merupakan salah satu penyakit kronis yang serius. Diabetes umumnya ditandai dengan tubuh tidak membuat cukup insulin atau tidak dapat menggunakan insulin yang dibuat seefektif yang dibutuhkan. Diabetes dapat dikelompokkan menjadi empat tipe. Semua tipe diabetes dapat menyebabkan komplikasi pada berbagai bagian tubuh dan dapat meningkatkan risiko kematian dini. Pada artikel ini, dilakukan penelitian dalam membangun model menggunakan algoritma *multilayer perceptron* di WEKA untuk mengklasifikasikan seseorang menderita diabetes. *Dataset* yang digunakan adalah "*Diabetes Health Indicators Dataset*" yang bersumber dari Kaggle. Terdapat tujuh percobaan yang dilakukan yang mana hasil terbaik ditunjukkan oleh percobaan 7 berdasarkan parameter performa *F-measure*.

Kata kunci: Diabetes; *F-measure; multilayer perceptron*; WEKA

Abstract. *Diabetes is a serious chronic disease. Diabetes is generally characterized by the body not making enough insulin or not being able to use the insulin it makes as effectively as needed. Diabetes can be categorized into four types. All types of diabetes can cause complications in various parts of the body and can increase the risk of premature death. This research is building a model using the multilayer perceptron algorithm at WEKA to classify a person as suffering from diabetes. The dataset used is the "Diabetes Health Indicators Dataset" sourced from Kaggle. There were seven experiments conducted in which the best results were shown by experiment 7 based on the F-measure performance parameter.*

Keywords: Diabetes; *F-measure; multilayer perceptron*; WEKA

1. Pendahuluan

Diabetes merupakan salah satu penyakit kronis yang serius. Individu yang mempunyai penyakit ini kehilangan kemampuan untuk secara efektif mengatur kadar glukosa dalam darah dan dapat menyebabkan penurunan kualitas serta harapan hidup. Diabetes umumnya ditandai dengan tubuh tidak membuat cukup insulin atau tidak dapat menggunakan insulin yang dibuat seefektif yang dibutuhkan.

Berdasarkan laporan dari organisasi kesehatan dunia atau WHO yang diterbitkan pada tahun 2016 [1], diperkirakan terdapat 422 juta orang dewasa yang mengidap penyakit diabetes. Angka ini naik dibandingkan pada tahun 1980 di mana jumlah orang dewasa pengidap diabetes berjumlah 108 juta orang. Pada tahun 2012, diabetes menyebabkan sekitar 1,5 juta kematian. Peningkatan risiko penyakit

kardiovaskular dan penyakit lainnya diiringi dengan kadar glukosa darah yang lebih tinggi dari optimal menyebabkan tambahan 2,2 juta kematian. 43% dari 3,7 juta kematian ini terjadi sebelum usia 70 tahun. Negara-negara berpenghasilan rendah memiliki persentase lebih tinggi dibandingkan dengan negara-negara berpenghasilan tinggi dalam hal kasus kematian akibat glukosa darah tinggi atau diabetes yang terjadi sebelum usia 70 tahun.

Diabetes dapat dikelompokkan menjadi empat tipe [2]. Tipe 1, yaitu diabetes yang disebabkan oleh kerusakan autoimun pada sel β yang biasanya mengarah pada defisiensi insulin absolut. Tipe 2, yaitu diabetes yang disebabkan karena hilangnya progresif sekresi insulin sel β yang memadai, sehingga tubuh tidak dapat menggunakan insulin yang dihasilkannya dengan benar. Tipe selanjutnya adalah diabetes yang disebabkan oleh penyebab lain, seperti diabetes yang dipicu oleh obat atau bahan kimia dan diabetes yang disebabkan oleh eksokrin pankreas. Tipe terakhir adalah diabetes mellitus gestasional, yaitu diabetes yang didiagnosis pada trimester kedua atau ketiga pada orang hamil.

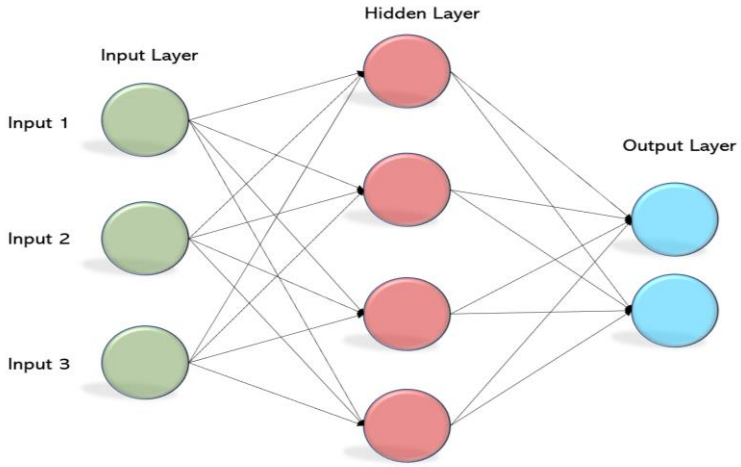
Semua tipe diabetes dapat menyebabkan komplikasi pada berbagai bagian tubuh dan dapat meningkatkan risiko kematian dini. Komplikasi seperti penyakit jantung, kehilangan penglihatan, amputasi kaki atau tungkai bawah, dan penyakit ginjal berhubungan dengan tingginya kadar gula yang tersisa dalam aliran darah bagi penderita diabetes. Meskipun tidak ada obat untuk diabetes, strategi seperti menurunkan berat badan, makan sehat, aktif berolahraga, dan menerima perawatan medis dapat mengurangi bahaya penyakit ini pada banyak pasien. Dengan adanya diagnosis dini terhadap penyakit diabetes dapat menyebabkan perubahan gaya hidup dan pengobatan yang lebih efektif.

Penelitian ini akan mencoba untuk membangun sebuah model yang memiliki ketepatan dalam melakukan klasifikasi terhadap adanya diabetes pada seseorang berdasarkan *dataset* yang digunakan. Model yang dibangun menggunakan algortima *multilayer perceptron* di aplikasi WEKA. Parameter utama pada *multilayer perceptron* ini adalah jumlah *hidden layer* dan jumlah *node* pada masing-masing *hidden layer*. Dengan dilakukannya penelitian ini diharapkan dapat ditemukan model yang baik dalam melakukan klasifikasi penderita diabetes.

2. Metode

2.1. Multilayer

Multilayer perceptron adalah jenis *neural network* yang paling dikenal dan paling sering digunakan. *Multilayer perceptron* terdiri dari tiga *layer*, yaitu *input layer*, *hidden layer*, dan *output layer* dengan elemen komputasi nonlinear. Dari satu *layer* dengan *layer* yang berdekatan, semua neuron atau *node* terhubung penuh ke sesama neuron atau *node* lainnya. Dalam proses komputasi, koneksi ini direpresentasikan sebagai bobot (intensitas koneksi). Bobot mempunyai peran penting dalam propagasi sinyal dalam jaringan. Bobot berisi pengetahuan tentang hubungan masalah dengan solusinya. Jumlah neuron pada *input layer* tergantung pada jumlah variabel independen dalam model, sedangkan jumlah neuron pada *output layer* sama dengan jumlah variabel dependen. Jumlah *output* neuron dapat tunggal atau lebih dari satu [3]. Gambar 1 menunjukkan alur informasi dari *input layer* menuju *output layer* melewati *hidden layer*.



Gambar 1. Alur Informasi

Pada WEKA, inti dari algoritma *backpropagation* adalah bentuk turunan parsial $\frac{\partial C}{\partial w}$ dari *cost function* C terhadap setiap bobot w (atau bias b) dalam *neural network*. Bentuk turunan ini menunjukkan seberapa cepat nilai *cost* (C) berubah ketika nilai bobot dan bias diubah. Bentuk kuadrat dari *cost function* C adalah sebagai berikut:

$$C = \frac{1}{2n}\sum \|y(x) - a^L(x)\|^2 \quad (1)$$

Dari bentuk tersebut, n adalah jumlah total contoh pelatihan; x adalah penjumlahan dari contoh pelatihan individu; y = y(x) adalah keluaran yang diinginkan; L menunjukkan jumlah *layer* dalam jaringan; dan $a^L = a^L(x)$ adalah vektor keluaran aktivasi dari jaringan ketika x dimasukkan.

Setelah iterasi pertama, *error* pada *output layer* perlu dihitung dengan persamaan berikut:

$$\delta_j^L = \frac{\partial C}{\partial a_j^L}\sigma'(z_j^L) \quad (2)$$

Error pada *hidden layer* kemudian dihitung setelah *error* pada *output layer* dihitung. Persamaan *error* pada *layer* δ^l dalam hal *error* pada *layer* selanjutnya δ^{l+1} adalah sebagai berikut:

$$\delta^l = ((w^{l+1})^T\delta^{l+1}).\sigma'(z') \quad (3)$$

Ketika semua *error* telah ditemukan, *gradient descent* dihitung guna meminimalkan *cost function* dan menemukan solusi optimal [4].

2.2. Cross-validation

Cross-validation adalah sebuah metode statistik yang dapat digunakan untuk mengevaluasi dan membandingkan performa model atau algoritma [5]. Pada metode ini, *dataset* dipisahkan menjadi dua segmen, yaitu data pelatihan atau pembelajaran dan data pengujian atau evaluasi. Model atau algoritma dilatih oleh segmen pelatihan dan divalidasi oleh segmen pengujian. Bentuk dasar dari *cross-validation* adalah *k-fold cross-validation*.

Dalam *k-fold cross-validation*, data pertama-tama dibagi menjadi sejumlah k segmen atau lipatan berukuran sama (atau hampir sama). Selanjutnya dilakukan proses iterasi k segmen pelatihan dan pengujian sedemikian rupa. Salah satu *subset* pada k segmen akan digunakan sebagai data pengujian dan *subset* data k segmen lainnya berfungsi sebagai data pelatihan [5]. Dalam *data mining* dan *machine learning*, *k-fold cross-validation* yang umum digunakan adalah *k* = 10 *cross-validation*.

2.3. WEKA

Waikato Environment for Knowledge Analysis (WEKA) adalah *software machine learning* dan *data mining* yang dikembangkan dalam bahasa pemrograman Java oleh Universitas Waikato di Selandia Baru [6]. WEKA menyediakan tampilan *interface* yang memungkinkan penggunanya untuk menerapkan berbagai metode pada *task data mining* standar secara langsung pada *dataset*, seperti data *pre-processing*, *classification*, *clustering*, *association rule*, *evaluation*, *visualization*, dan *feature selection*. Beberapa metode yang terdapat pada WEKA di antaranya adalah algoritma C4.5 (C5), ID3, *K-means*, dan *Apriori*. WEKA juga mendukung berbagai format *file* untuk *data mining* termasuk .arff dan .csv [7].

2.4. Dataset

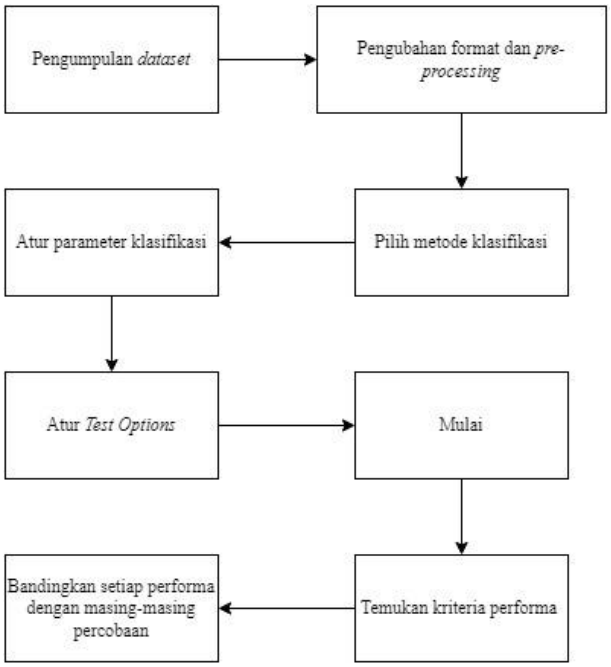
Pada percobaan klasifikasi ini, digunakan "*Diabetes Health Indicators Dataset*" yang bersumber dari situs Kaggle (https://www.kaggle.com/alexteboul/diabetes-health-indicators-dataset). "*Diabetes Health Indicators Dataset*" adalah *dataset* bersih yang terdiri dari 253.680 baris respon terhadap survei "*The Behavioral Risk Factor Surveillance System*" (BRFSS) tahun 2015 yang diinisiasi oleh Centers for Disease Control (CDC) yang diselenggarakan setiap tahunnya. *Dataset* ini memiliki 21 variabel fitur dan satu variabel target yang terbagi menjadi tiga kelas. Kelas 0 untuk tanpa diabetes atau diabetes hanya selama kehamilan, Kelas 1 untuk pradiabetes, dan Kelas 2 untuk diabetes. Penjelasan mengenai masing-masing target dan fitur pada *dataset* ini adalah sebagai berikut:

Diabetes	:	0 = Tidak diabetes, 1 = Pradiabetes, 2 = Diabetes.
HighBP	:	0 = Tekanan darah rendah, 1 = Tekanan darah tinggi.
HighChol	:	0 = Kolesterol rendah, 1 = Kolesterol tinggi.
CholCheck	:	0 = Tidak melakukan pemeriksaan kolesterol selama lima tahun terakhir, 1 = Melakukan pemeriksaan kolesterol selama lima tahun terakhir.
BMI	:	*Body mass index.*
Smoker	:	0 = Bukan perokok, 1 = Perokok.
Stroke	:	0 = Tidak menderita stroke, 1 = Menderita stroke.
HeartDiseaseorAttack	:	0 = Tidak menderita *coronary heart disease* (CHD), 1 = Menderita *coronary heart disease* (CHD).
PhysActivity	:	0 = Tidak aktif, 1 = Aktif.
Fruits	:	0 = Jarang makan buah, 1 = Sering makan buah.
Veggies	:	0 = Jarang makan sayur, 1 = Sering makan sayur.
HvyAlcoholConsump	:	0 = Jarang minum alkohol, 1 = Sering minum alkohol.
AnyHealthcare	:	AnyHealthcare : 0 = Tidak memiliki asuransi kesehatan, 1 = Memiliki asuransi kesehatan.
NoDocbcCost	:	0 = Tidak mampu ke dokter karena masalah biaya, 1 = Mampu ke dokter.
GenHlth	:	1 = Kondisi kesehatan luar biasa baik, 2 = Kondisi kesehatan sangat baik, 3 = Kondisi kesehatan baik, 4 =

		Kondisi kesehatan cukup baik, 5 = Kondisi kesehatan tidak baik.
MentHlth	:	Permasalahan kesehatan mental dalam 30 hari.
PhysHlth	:	Permasalahan kesehatan fisik dalam 30 hari.
DiffWalk	:	0 = Tidak mempunyai masalah dalam berjalan, 1 = Mempunyai masalah dalam berjalan.
Sex	:	0 = Perempuan, 1 = Laki-laki.
Age	:	13 lever kategori umur.
Education	:	Level edukasi, terdiri dari skala 1-6.
Income	:	Level pendapatan, terdiri dari skala 1-8.

2.5. Rancangan Eksperimen

Percobaan klasifikasi ini bertujuan untuk melatih model *multilayer perceptron* guna menemukan parameter jumlah *hidden layer* dan jumlah *node* pada *hidden layer* yang tepat, sehingga dapat menghasilkan model dengan performa terbaik. Langkah-langkah dalam melakukan percobaan tersebut tersaji pada Gambar 2



Gambar 2. Tahapan Penelitian

a. Pengumpulan *dataset*
"*Diabetes Health Indicators Dataset*" diperoleh dari situs yang banyak digunakan oleh komunitas *data scientist*, yaitu Kaggle (https://www.kaggle.com/alexteboul/diabetes-health-indicators-dataset).
b. Pengubahan format dan *pre-processing*
Dataset yang diperoleh dari Kaggle masih berformat .csv. Oleh sebab itu, dilakukan perubahan format *dataset* menjadi format .arff agar *dataset* dapat diolah di WEKA. Setelah pengubahan format *file*, bentuk fitur-fitur pada *dataset* kemudian disesuaikan. Fitur BMI, MentHlth, dan PhysHlth diubah menjadi bentuk numerik dan fitur sisanya diubah menjadi bentuk nominal.

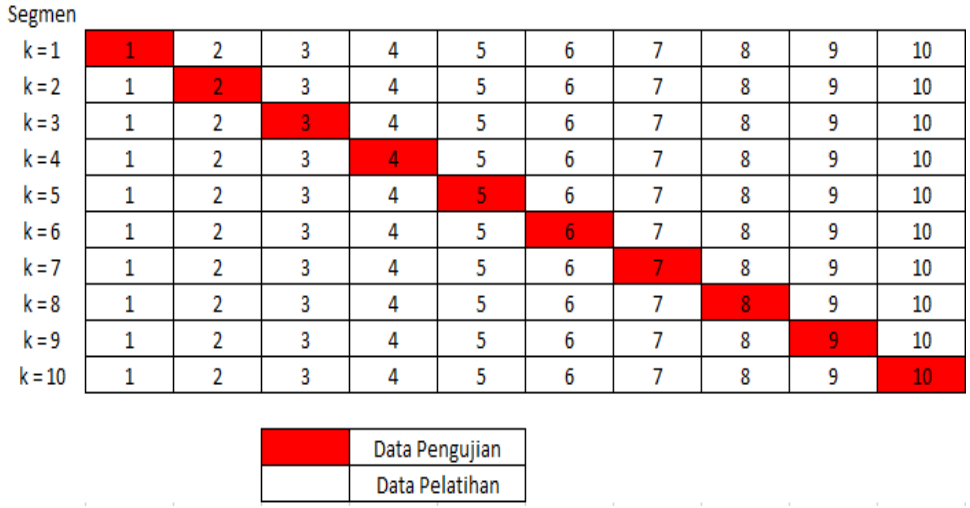
c. Pilih metode klasifikasi
Pada proses klasifikasi ini, algoritma yang digunakan adalah algoritma *multilayer perceptron*. Algoritma ini dipilih karena *dataset* yang digunakan tidak dapat dipisahkan secara linear (nonlinear) [3].

d. Atur parameter klasifikasi
Pada percobaan klasifikasi ini, pilih *classifier* MultilayerPerceptron pada WEKA. Atur parameter dengan ketentuan di antaranya adalah batchSize = 32, learningRate = 0.3, momentum = 0.2, dan trainingTime = 100. Pada parameter hiddenLayers, dilakukan berbagai percobaan. Percobaan tersebut sebagaimana ditunjukkan oleh Tabel 1.

Tabel 1. Hasil Percobaan

Percobaan	Jumlah *node* pada *hidden layer* 1	Jumlah *node* pada *hidden layer* 2
1	5	0
2	10	0
3	15	0
4	5	1
5	10	5
6	15	10
7	20	15

e. Atur *Test Option*
Percobaan ini menggunakan 10 *cross validation*. Nilai k = 10 membuat prediksi menggunakan 90% data, sehingga lebih mungkin untuk digeneralisasikan ke data yang lengkap [5]. Dalam 10 *cross-validation*, data dibagi menjadi 10 segmen berukuran kira-kira sama, sehingga didapatkan 10 segmen data untuk mengevaluasi kinerja model atau algoritma. Untuk masing-masing dari 10 segmen data tersebut, digunakan 9 *subset* untuk pelatihan dan 1 *subset* untuk pengujian seperti diilustrasikan pada Gambar 3.



Gambar 3. Pengujian Subset

f. Mulai
Klik tombol '*Start*" untuk melatih model dengan parameter yang telah ditentukan.

g. Temukan kriteria performa
   Setelah proses pelatihan selesai dilakukan, perhatikan nilai *F-measure* yang didapatkan untuk setiap kelas. Nilai *F-measure* akan digunakan sebagai kriteria performa model.
h. Bandingkan setiap performa dengan masing-masing percobaan
   Lakukan proses pelatihan sebanyak percobaan yang telah ditentukan, yaitu tujuh kali percobaan. Setelah itu, bandingkan nilai *F-measure* dari masing-masing percobaan untuk setiap kelas. Model dengan nilai *F-measure* tertinggi akan menjadi model terbaik.

3. Hasil dan Pembahasan

3.1. F-Measure

Hasil Parameter evaluasi yang digunakan untuk mengetahui seberapa baik model yang dibangun pada penelitian ini adalah *precision*, *recall*, dan *F-Measure*. Hasil dari *precision*, *recall*, dan *F-Measure* menghasilkan nilai antara 0 sampai 1. Model yang baik adalah model yang menghasilkan nilai mendekati 1 [8].

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

$$F\text{-}Measure = \frac{(2\times Precision\times Recall)}{(Precision+Recall)} \quad (6)$$

3.2. Eksperimen

Tujuh percobaan telah dilakukan pada penelitian ini menggunakan algoritma klasifikasi *multilayer perceptron* di WEKA untuk menemukan model terbaik berdasarkan parameter evaluasi *F-measure*. Model terbaik pada penelitian ini mengacu pada nilai tertinggi dan terbaik yang diperoleh berdasarkan parameter evaluasi *F-measure* untuk masing-masing kelas (kelas 0, 1, dan 2) pada masing-masing percobaan (percobaan 1 sampai 7). Masing-masing percobaan memiliki parameter konfigurasi yang hampir sama. Namun, yang membedakannya adalah jumlah *hidden layer* dan *node* yang terdapat pada masing-masing *hidden layer*. Tabel 2 menunjukkan hasil *F-measure* pada masing-masing kelas, yaitu kelas 0, kelas 1, dan kelas 2 untuk masing-masing percobaan.

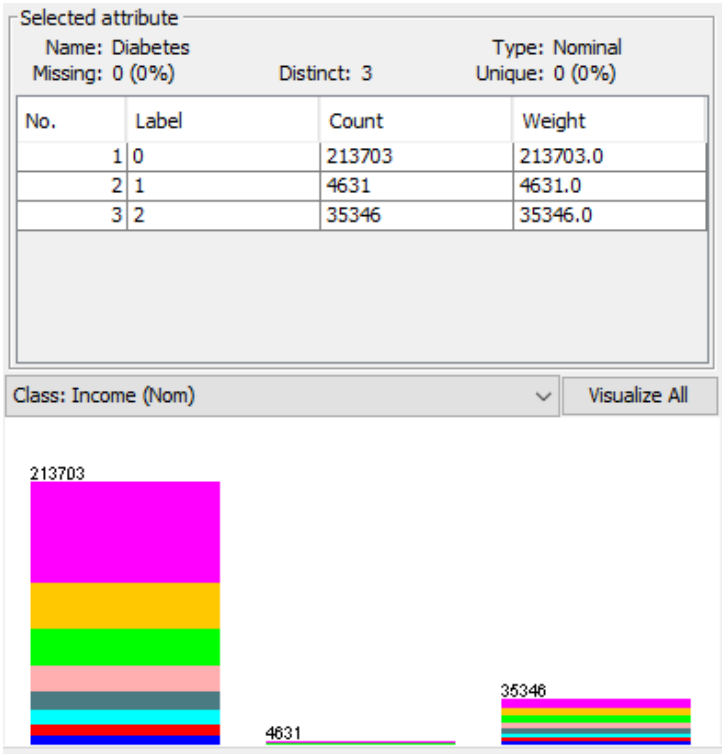
Tabel 2. Hasil *F-measure*

Percobaan	Jumlah *node* pada *hidden layer* 1	Jumlah *node* pada *hidden layer* 2	*F-measure* pada masing-masing kelas		
			0	1	2
1	5	0	0,915	?	0,223
2	10	0	0,915	?	0,205
3	15	0	0,915	?	0,189
4	5	1	0,914	?	0,125
5	10	5	0,915	?	0,185
6	15	10	0,916	?	0,147
7	20	15	0,916	?	0,228

Berdasarkan Tabel 2, diperoleh hasil yang beragam untuk setiap percobaan. Pada model percobaan 1, 2, dan 3 yang menggunakan satu *hidden layer* didapatkan nilai *F-measure* untuk kelas 0 bernilai sama, yaitu 0,915. Sedangkan untuk kelas 1, nilai *F-measure* yang diperoleh adalah berbentuk tanda tanya ("?"). Untuk kelas 2, nilai *F-measure* yang diperoleh untuk masing-masing percobaan berbeda-beda. Percobaan 1 mendapatkan nilai 0,223. Percobaan 2 mendapatkan 0,205. Percobaan 3 mendapatkan nilai paling kecil, yaitu 0.189.

Pada percobaan dengan menggunakan dua *hidden layer*, yaitu pada percobaan 4, 5, 6, dan 7 diperoleh hasil yang beragam juga. Pada kelas 0, percobaan 6 dan 7 menghasilkan nilai *F-measure* yang sama, yaitu 0,916 disusul oleh percobaan 5 dengan nilai 0,915 dan percobaan 4 dengan nilai 0,914. Untuk kelas 1, semua percobaan dengan dua *hidden layer* menghasilkan nilai *F-measure* yang sama dengan percobaan yang menggunakan satu *hidden layer*, yaitu berbentuk tanda tanya ("?"). Sebaliknya, untuk kelas 2 hasil yang beragam diperoleh. Percobaan 7 memperoleh nilai tertinggi dengan 0,228 dilanjutkan oleh percobaan 5 dengan 0,185. Percobaan 6 menghasilkan nilai 0,147 dan percobaan 4 menghasilkan nilai paling kecil dengan nilai 0,125. Berdasarkan hasil percobaan dari percobaan 1 hingga percobaan 7 yang telah dilakukan, terlihat bahwa percobaan 7 menunjukkan model terbaik berdasarkan parameter *F-measure*, dengan kelas 0 mendapat nilai sebesar 0.916, kelas 1 dengan nilai berbentuk tanda tanya ('?"), dan kelas 2 dengan nilai 0.228.

Munculnya tanda tanya ("?") untuk nilai *F-measure* pada kelas 1 di WEKA memperlihatkan bahwa terdapat pembagian 0 yang terjadi pada kelas 1 ketika menghitung nilai *F-measure*. Dengan adanya pembagian 0 menunjukkan bahwa model gagal menebak anggota dari kelas 1. Hal tersebut diperkuat oleh pengelompokkan data yang memperlihatkan bahwa *dataset* yang digunakan merupakan *imbalance dataset* atau *dataset* yang pengklasifikasiannya tidak seimbang. Ketidakseimbangan *dataset* itu ditunjukkan oleh Gambar 4.



Gambar 4. Ketidakseimbangan Dataset

Pada Gambar 4 jumlah *instance* yang termasuk kelas 0 berjumlah 213.703. Selanjutnya kelas 2 dengan jumlah *instance* 35.346. Jumlah *instance* terkecil diperoleh oleh kelas 1, yaitu sebanyak 4.631.

4. Kesimpulan

Tujuh percobaan telah dilakukan untuk melakukan klasifikasi terhadap tiga kelas, yaitu kelas 0, 1, dan 2 pada "*Diabetes Health Indicators Dataset*" yang diperoleh dari situs Kaggle. Tujuh percobaan tersebut dibagi menjadi dua bagian, yaitu percobaan dengan satu *hidden layer* dan dua *hidden layer*. Masing-masing percobaan menggunakan *node* yang berbeda-beda sesuai pada Tabel 2. Dari ketujuh percobaan tersebut, diperoleh hasil yang beragam. Berdasarkan kriteria performa *F-measure*, percobaan 7 memperoleh nilai terbaik yang mana untuk kelas 0 memperoleh nilai 0,916, kelas 1 memperoleh nilai tanda tanya ("?"), dan kelas 2 mendapatkan nilai 0,228. Meskipun percobaan 7 menghasilkan performa terbaik dari semua percobaan, tetapi hasil ini bukanlah hasil paling maksimal mengingat munculnya tanda tanya ("?") untuk kelas 1 akibat *imbalance dataset*. Selain itu, penelitian ini hanya terbatas pada tujuh percobaan. Untuk memperoleh hasil lainnya yang lebih optimal, perlu dilakukan percobaan lanjutan dengan menggunakan pengaturan parameter, khususnya pada parameter jumlah *hidden layer* dan jumlah *node* pada *hidden layer*.

Referensi

[1] WHO Global Report, "Global Report on Diabetes," *Isbn*, vol. 978, p. 11, 2016, [Online]. Available: http://www.who.int/about/licensing/copyright_form/index.html%0Ahttp://www.who.int/about/licensing/copyright_form/index.html%0Ahttp://www.who.int/about/licensing/copyright_form/index.html%0Ahttps://apps.who.int/iris/handle/10665/204871%0Ahttp://www.who.int/a.

[2] D. Care and S. S. Suppl, "2. Classification and diagnosis of diabetes: Standards of medical care in diabetes-2021," *Diabetes Care*, vol. 44, no. January, pp. S15–S33, 2021, doi: 10.2337/dc21-S002.

[3] Y. S. Park and S. Lek, *Artificial Neural Networks: Multilayer Perceptron for Ecological Modeling*, vol. 28. Elsevier, 2016.

[4] A. Music and S. Gagula-Palalic, "Classification of Leaf Type Using Multilayer Perceptron, Naive Bayes and Support Vector Machine Classifiers," *Southeast Eur. J. Soft Comput.*, vol. 5, no. 2, 2016, doi: 10.21533/scjournal.v5i2.119.

[5] P. Refaeilzadeh, L. Tang, H. Liu, L. Angeles, and C. D. Scientist, "Cross-Validation," *Encycl. Database Syst.*, 2020, doi: 10.1007/978-1-4899-7993-3.

[6] E. G. Kulkarni and R. B. Kulkarni, "WEKA Powerful Tool in Data Mining General Terms," *Int. J. Comput. Appl.*, vol. 5, no. Rtdm, pp. 975–8887, 2016.

[7] K. Manchandia, N. Khare, and M. Agrawa, "WEKA AS A DATA MINING TOOL TO ANALYZE STUDENTS' ACADEMIC PERFORMANCES USING NAÏVE BAYES CLASSIFIER- A SURVEY," *Int. J. Eng. Sci. Res. Technol. WEKA*, vol. 6, no. 3, pp. 431–434, 2017.

[8] E. Haddi, X. Liu, and Y. Shi, "The role of text pre-processing in sentiment analysis," *Procedia Comput. Sci.*, vol. 17, no. December, pp. 26–32, 2013, doi: 10.1016/j.procs.2013.05.005.