Deteksi Penyakit *Dengue Hemorrhagic Fever* dengan Pendekatan *One Class Classification*

Zida Ziyan Azkiya[1)], Fatma Indriani[2)], Heru Kartika Chandra[3)]

[12)]*Program Studi Ilmu Komputer, Fakultas MIPA, Universitas Lambung Mangkurat*
Jl A Yani KM 36, Banjarbaru 70714
[1)]ziziyanazkiya@gmail.com
[2)]f.indriani@unlam.ac.id

[3)]*Program Studi Komputerisasi Akuntansi, Politeknik Negeri Banjarmasin*
Jl Brigjen H. Hasan Basry, Banjarmasin 70123
[3)]hkc131170@gmail.com

Abstrak— **Pada kasus deteksi penderita penyakit demam berdarah (*Dengue Hemorrhagic Fever*- DHF), data training yang tersedia umumnya hanya data pasien penderita positif. Sedangkan data orang normal (data negatif) tidak tersedia secara khusus. Pada makalah ini dipaparkan pembangunan model klasifikasi untuk deteksi DHF dengan pendekatan *One Class Classification* (OCC). Data yang digunakan pada penelitian ini adalah hasil uji darah dari laboratorium dari pasien penderita penyakit demam berdarah. Metode yang diteliti adalah *One-class Support Vector Machine* dan *K-Means*. Hasil yang diperoleh pada penelitian ini adalah untuk metode SVM memiliki nilai *precision* = 1,0, *recall* = 0,993, *f-1 score* = 0,997, dan tingkat akurasi sebesar 99,7% sedangkan dengan metode *K-Means* diperoleh nilai *precision* = 0,901, *recall* = 0,973, *f-1 score* = 0,936, dan tingkat akurasi sebesar 93,3%. Hal ini menunjukkan bahwa metode SVM sedikit lebih unggul dibandingkan dengan *K-Means* untuk kasus ini.**

Kata Kunci— **demam berdarah, Dengue Hemorrhagic Fever, K-Means, One Class Classification, OSVM**

Abstract— **Two class classification problem maps input into two target classes. In certain cases, training data is available only in the form of a single class, as in the case of Dengue Hemorrhagic Fever (DHF) patients, where only data of positive patients is available. In this paper, we report our experiment in building a classification model for detecting DHF infection using One Class Classification (OCC) approach. Data from this study is sourced from laboratory tests of patients with dengue fever. The OCC methods compared are One-Class Support Vector Machine and One-Class K-Means. The result shows SVM method obtained precision value = 1.0, recall = 0.993, f-1 score = 0.997, and accuracy of 99.7% while the K-Means method obtained precision value = 0.901, recall = 0.973, f- 1 score = 0.936, and accuracy of 93.3%. This indicates that the SVM method is slightly superior to K-Means for One-Class Classification of DHF patients.**

Keywords— **Dengue Hemorrhagic Fever, K-Means, One Class Classification, OSVM**

Article history:
Received 25 August 2017; Received in revised form 10 October 2017; 17 October 2017; Available online 28 October 2017

I. PENDAHULUAN

Klasifikasi dua kelas memetakan input ke dua macam kelas target. Namun pada kasus klasifikasi tertentu, data training yang tersedia hanya berupa satu kelas (kelas positif). Pada kasus penyakit demam berdarah (*Dengue Hemorrhagic Fever*), hanya data pasien penderita positif yang umum tersedia, sedangkan data bukan penderita (data negatif) sulit dicari atau tidak tersedia secara khusus. Oleh karena itu, makalah ini bertujuan memaparkan pembangunan model klasifikasi untuk identifikasi penyakit *Dengue Hemorrhagic Fever* (DHF) hanya dengan menggunakan data positif. Pendekatan seperti ini disebut dengan One Class Classification (Tax, 2001).

Identifikasi pasien penderita DHF dengan klasifikasi dua ataupun multi kelas sudah banyak diteliti. Beberapa penelitian terdahulu menggunakan metode pohon keputusan seperti (Astuti, Suciati, Mujiati, Ayu, Ristianah, & Lestari, 2016), (Haryanto, 2013), dan (Munah, 2015) dengan akurasi terbaik 88,71%. Sedangkan penelitian (Lesmana, Hikmah, & Karimah, 2014) menggunakan jaringan syaraf tiruan Backpropagation dengan akurasi 88,9%. (Suwardi, 2012) meneliti tiga metode sekaligus yaitu Backpropagation, Nearest-Neighbor, dan C4.5, dengan akurasi terbaik adalah Backpropagation 92,9%. Semua penelitian tersebut merupakan klasifikasi *2-class*, yaitu menggunakan data training dari kasus positif maupun kasus negatif, yang bersumber dari pemeriksaan awal pasien dan/atau catatan rekam medis.

Menurut (Khan & Madden,2010), pendekatan One Class Classification (OCC) banyak diterapkan pada berbagai bidang, diantaranya klasifikasi teks, pengenalan angka tulisan tangan, pengenalan wajah, analisis medis, bioinformatika, deteksi spam, deteksi anomali, serta deteksi

e-ISSN 2443-2555
doi: http://dx.doi.org/10.20473/jisebi.3.2.122-128

kerusakan mesin. Selain itu, OCC pernah diteliti untuk deteksi kebocoran listrik oleh (Nagi, Yap, Tiong, Ahmed, & Mohamad, 2010).

Pada bidang penyakit, OCC digunakan untuk memprediksi wabah flu burung berdasar data lingkungan dan tempat terjadinya wabah (Zhang, Lu, & Zhang, 2011). Di penelitian lain, (Cabral & de Oliveira, 2014) menerapkan OCC untuk diagnosis penyakit jantung.

Pada makalah ini dipaparkan deteksi *Dengue Hemorrhagic Fever* dengan pendekatan One Class Classification (OCC). Menurut (Khan & Madden,2014), metode OCC digolongkan menjadi dua macam yaitu SVM dan non-SVM. Oleh karena itu, metode OCC yang kami pilih adalah One Class SVM serta K-Means untuk mewakili dua kelompok tersebut. Makalah ini memberikan kontribusi terkait pengunaan metode OCC untuk kasus deteksi deman DHF.

II. METODE PENELITIAN

Berikut ini adalah langkah-langkah yang dilakukan.

A. Pengumpulan Data

Pada tahap pengumpulan data, data yang diperoleh berasal dari dua sumber. Sumber pertama yaitu data hasil uji cek lab pasien positif penderita penyakit *Dengue Hemorrhagic Fever* di kota Banjarbaru pada tahun 2013-2015. Data ini diperoleh dari Dinas Kesehatan Kota Banjarbaru, Kalimantan Selatan dan akan digunakan pada proses training. Sumber kedua yaitu hasil uji lab pasien normal yang diperoleh dari Laboratorium Klinik Panasea Kota Banjarbaru, Kalimantan Selatan. Data ini akan digunakan sebagai data pasien negatif. Proses training *one class classification* hanya membutuhkan data positif, namun untuk tahap evaluasi (*testing*) dibutuhkan juga data pasien negatif. Data negatif berjumlah jauh lebih sedikit karena hanya akan digunakan untuk *testing*.

Penelitian-penelitian sebelumnya untuk deteksi DHF menggunakan data rekam medis dan/atau catatan awal pemeriksaan. Pada penelitian ini, data tes laboratorium dipilih karena mudah tersedia sebagai data training. Selain itu untuk penggunaan deteksi setelah dilakukan training, data tes laboratorium mudah didapat sebagai data masukan oleh pengguna, tidak ambigu, serta eksak dan terstruktur.

B. Prapemrosesan Data

Tahap ini dilakukan agar data mentah siap diolah dengan metode One-Class SVM maupun One-Class K-means.

1) *Data Cleaning :* Pemrosesan *cleaning* yang dilakukan adalah penghapusan *instance* jika ada beberapa bagian atributnya yang kosong.

2) *Data Integration :* Data yang diperoleh merupakan data penderita penyakit DHF pada tahun 2013–2015. Data yang akan digunakan terdiri dari beberapa tabel. Untuk data tahun 2013 terdiri dari 12 tabel yang dipisah namun masih dalam file yang sama dan dibagi sesuai dengan bulan yaitu dari Januari-Desember. Sedangkan pada tahun 2014 dan 2015 tabel tidak terpisah berdasarkan bulan. Data yang diperoleh merupakan data penderita penyakit DHF pada tahun 2013–2015. Setelah penggabungan didapat 1059 data pasien positif dan 150 data negatif. Atribut yang digunakan ditampilkan di Tabel 1.

3) *Data Transformation :* Pada tahap ini, akan dilakukan proses normalisasi MinMax [0,1] untuk menormalisasi tiap atribut pada dataset agar nilai rentangnya seragam antara 0 dan 1.

4) *Pembentukan dataset training dan testing :* Untuk proses training klasifikasi, data yang digunakan adalah sebanyak 909 data pasien DHF (kelas positif). Sedangkan untuk testing, data yang digunakan terdiri dari 150 pasien DHF (kelas positif) dan 150 bukan penderita (kelas negatif). Jumlah data untuk hasil akhir pembagian dataset *training* dan *testing* dapat dilihat di Tabel 2.

TABEL 1. ATRIBUT DATA PASIEN

No	Nama atribut
1	Jenis kelamin
2	Hemoglobin
3	Trombosit
4	Hematokrit
5	Kelas (positif/negatif)

TABEL 2. ALOKASI DATA UNTUK TRAINING DAN TESTING

Data	Training	Testing
Positif	909	150
Negatif	0	150

Selanjutnya dilakukan dua buah eksperimen klasifikasi dengan dua metode OCC yang berbeda, yaitu One-Class SVM dan One-Class K-means.

C. One-Class SVM

Formulasi SVM reguler adalah memaksimalkan margin garis pemisah antara dua kelas data. Untuk SVM dengan pendekatan *one-class classification,* (Schölkopf, Williamson, Smola, Shawe-Taylor, & Platt) memodifikasi SVM menjadi pencarian hiperplane dengan jarak pemisah maksimal antara titik-titik data yang diberikan (data positif) dengan titik origin. Hasil training berupa fungsi biner yang memetakan apakah suatu titik uji memiliki probabilitas tinggi berada di daerah data training (positif) atau di luar daerah (negatif). Pada formulasi One-class SVM ini ada tambahan parameter yaitu v (nu) yang bernilai antara 0 dan 1. Nilai v ini menentukan batas maksimal fraksi *outlier* pada dataset training, sekaligus batas

minimal fraksi jumlah support vector terhadap jumlah data training.

Tingkat keakuratan metode One-Class SVM dalam memprediksi bergantung pada optimasi nilai parameter yang digunakan dalam suatu penelitian, parameter tersebut berupa gamma dan nu. Parameter gamma bernilai harus lebih besar dari 0, sedangkan parameter nu bernilai di bawah 1. Untuk mendapatkan nilai akurasi yang terbaik, optimasi parameter diperlukakn. Langkah-langkah yang dilakukan yaitu Training One-Class SVM dan Testing One-Class SVM. Training One-Class SVM menggunakan dataset training dengan berbagai nilai gamma (γ) dan nu (v). Pelaksanaan tahap ini menggunakan library libsvm pada Java (https://www.csie.ntu.edu.tw/~cjlin/libsvm/). Nilai gamma yang diujicoba yaitu γ=[0,001; 0,01; 0,1; 1; 10; 100; 1000) dan nilai nu yaitu v=[0,001; 0,01; 0,1]. Kernel yang digunakan adalah kernel RBF. Testing One-Class SVM menggunakan dataset testing

D. One-Class K-Means

Penerapan metode K-Means untuk *one class classification* dilakukan dengan langkah-langkah di bawah ini. Langkah 1-6 adalah metode K-Means reguler (*unsupervised*) yang diterapkan pada data training. Kluster-kluster yang terbentuk mendeskripsikan kelas positif. Pada langkah 7-8, tiap butir data testing akan diuji apakah masuk *boundary* salah satu kluster (positif) atau tidak masuk kluster manapun (negatif).

1) Menentukan nilai K: Nilai k yang dicoba yaitu k = [1..15]. Clustering data dibantu oleh tool Weka.

2) *Menentukan K sentroid:* menentukan titik pusat kluster awal secara random.

3) Menghitung jarak: tiap data (1…n) dihitung jarak terhadap tiap kluster.

4) Menentukan posisi kluster: yang ditentukan oleh jarak yang terdekat.

5) *Menentukan pusat kluster baru :* pusat kluster baru ditentukan.

6) *Mengulangi langkah 3* : menghitung jarak tiap data terhadap tiap cluster hingga tidak ada perubahan cluster.

7) mencari jarak maksimum (maximum distance) : Setelah didapat cluster, maka mencari jarak maksimum (maximum distance) tiap kluster. Jarak maksimum masing-masing kluster adalah sebagai boundary dari kluster tersebut.

8) *Perhitungan data testing :* testing K-Means menggunakan dataset testing. Suatu data uji akan diberi nilai sebagai kelas positif jika berada dalam boundary salah suatu kluster hasil training.

III. HASIL DAN PEMBAHASAN

Metrik yang digunakan untuk evaluasi adalah *precision, recall, f-1 score,* dan akurasi. Pada kasus di penelitian ini, nilai *recall* lebih penting dari *precision*. Hal tersebut dikarenakan untuk penderita DHF, deteksi pasien yang sebenarnya berpenyakit tetapi dianggap tidak berpenyakit (*false negative)* lebih penting daripada pasien yang sebenarnya tidak berpenyakit tetapi dianggap berpenyakit (*false positive*). Jika pasien yang dianggap berpenyakit tetapi sebenarnya pasien tersebut tidak berpenyakit, maka akan dilakukan penanganan selanjutnya secara medis dan dapat teridentifikasi bahwa pasien memang tidak berpenyakit. Sedangkan jika pasien yang dianggap tidak berpenyakit tetapi sebenarnya pasien tersebut berpenyakit, maka tidak ada tahapan penanganan medis selanjutnya dan hal ini bisa berakibat fatal pada pasien.

Hasil testing untuk metode one class *support vector machine* dengan berbagai nilai gamma dan nu ditunjukkan pada Tabel 3 (*precision*), Tabel 4 (*recall*), Tabel 5 (*f-1 score*), dan Tabel 6 (akurasi). Precision secara umum baik jika γ >= 0,1 atau v >= 0,01. Sedangkan *recall* baik jika γ di antara 0,1-10. Untuk hasil keseluruhan berdasar *f-1 score* dan akurasi, maka parameter terbaik didapat jika γ=1 dan v=0,01.

Hasil testing untuk metode K-Means ditampilkan pada Tabel 7. Dapat dilihat bahwa nilai *precision* yang dihasilkan bernilai di atas 0,5, namun nilai *precision* sempurna diperoleh saat menggunakan jumlah kluster 1, 2, 3, 9, 10, 11, 13, 14, dan 15. Untuk *recall*, nilai yang dihasilkan cukup beragam dengan nilai *recall* terburuk pada saat menggunakan 14 cluster yaitu bernilai 0,230, sedangkan nilai *recall* terbaik diperoleh saat menggunakan 4, 5, 6, 7, dan 8 cluster, yaitu 0,993.

TABEL 3. HASIL *PRECISION* PADA ONE-CLASS SVM

γ / v	0,001	0,01	0,1	1	10	100	1000
0,001	0,645	0,75	0,887	0,993	1,0	1,0	1,0
0,01	0,702	0,993	0,993	1,0	1,0	1,0	1,0
0,1	1,0	1,0	1,0	1,0	1,0	1,0	1,0

TABEL 4. HASIL *RECALL* PADA ONE-CLASS SVM

γ / *v*	0,001	0,01	0,1	1	10	100	1000
0,001	0,607	1	0,993	0,993	0,96	0,847	0,153
0,01	0,787	0,993	0,993	0,993	0,96	0,847	0,180
0,1	0,893	0,9	0,887	0,88	0,887	0,87	0,213

TABEL 5. HASIL *F-1 SCORE* PADA ONE-CLASS SVM

γ / *v*	0,001	0,01	0,1	1	10	100	1000
0,001	0,625	0,857	0,937	0,993	0,979	0,917	0,266
0,01	0,742	0,993	0,993	0,997	0,979	0,917	0,305
0,1	0,944	0,947	0,94	0,936	0,94	0,925	0,352

TABEL 6. HASIL *ACCURACY* PADA ONE-CLASS SVM

γ / *v*	0,001	0,01	0,1	1	10	100	1000
0,001	0,637	0,833	0,993	0,993	0,98	0,923	0,577
0,01	0,727	0,993	0,993	0,997	0,98	0,923	0,590
0,1	0,947	0,95	0,943	0,94	0,943	0,925	0,607

Untuk *f-1 score* pada K-Means, nilai yang dihasilkan juga cukup beragam dengan nilai *f-1 score* paling rendah atau terburuk pada saat menggunakan 14 cluster, sedangkan nilai *f-1 score* tertinggi atau terbaik diperoleh saat menggunakan 12 cluster yaitu 0,936. Nilai akurasi terbaik adalah 93% dengan k=12 dan k=15.

TABEL 7.HASIL *PRECISION, RECALL, F-1 SCORE* DAN *ACCURACY K-MEANS*

k	*Precision*	*Recall*	*F-1 Score*	*Accuracy*
1	1,0	0,460	0,630	73%
2	1,0	0,673	0,804	83.6%
3	1,0	0,530	0,696	76.7%
4	0,619	0,993	0,762	69%
5	0,634	0,993	0,774	71%
6	0,627	0,993	0,768	70%
7	0,639	0,993	0,778	71.67%
8	0,634	0,993	0,774	71%
9	1,0	0,660	0,795	83%
10	1,0	0,493	0,661	74.67%
11	1,0	0,440	0,611	72%
12	0,901	0,973	0,936	93%
13	1,0	0,640	0,780	82%
14	1,0	0,230	0,378	61.67%
15	1,0	0,867	0,928	93%

Walaupun *f-1 score* terbaik didapat pada k=12, namun jika hanya menekankan nilai *recall* seperti yang dijelaskan di awal Bagian III, maka kita bisa memilih model yang lebih sederhana, yaitu jika jumlah kluster k=4. Nilai *recall* pada k=4 bahkan lebih baik dari pada k=12 (0,993 dibanding 0,973). Jika hanya menekankan pada recall, maka nilai *precision* k=4 yang hanya 0,619 bisa diabaikan.

Tabel 8 menunjukkan model terbaik pada One-Class SVM dibandingkan dengan model terbaik pada K-means. Untuk metode One-Class SVM diperoleh nilai *precision* = 1,0, *recall* = 0,993, *f-1 score* = 0,997, dan *accuracy* sebesar 0,997 atau 99,7% dengan memasukkan parameter gamma = 1 dan nu = 0,01. Sedangkan untuk metode K-Means diperoleh nilai *precision* = 0,901, *recall* = 0,973, *f-1 score* = 0,936, dan *accuracy* sebesar 93,3 % dengan jumlah kluster k=12. Diperoleh kesimpulan bahwa metode One Class SVM memiliki nilai *precision*, *recall*, *f-1 score*, dan *accuracy* yang bernilai lebih tinggi dibandingkan dengan metode K-Means.

TABEL 8.HASIL PRECISION, RECALL, F-1 SCORE, DAN ACCURACY PADA METODE OSVM DENGAN PARAMETER TERBAIK DAN K-MEANS DENGAN CLUSTER TERBAIK

No	**Pengujian** Parameter	**OSVM** γ=1, *v=0,01*	**K-Means** k=12
1	*Precision*	1,0	0,901
2	*Recall*	0,993	0,73
3	*F1-Score*	0,997	0,936
4	*Accuracy*	99,7%	93,3%

Seperti sudah dipaparkan di bagian Pendahuluan, keunggulan pendekatan *One Class Classification* adalah bisa membuat model klasifikasi dengan training dari data kelas positif saja. Hal ini cocok untuk kasus deteksi demam DHF dimana hanya data penderita yang tersedia (data positif). Dari segi kinerja klasifikasi, pendekatan OCC bahkan lebih unggul daripada penelitian deteksi DHF yang diteliti sebelumnya. Metode *One Class* SVM dan K-Means untuk

OCC memiliki akurasi masing-masing sebesar 99,7% dan 93,3%.

IV. KESIMPULAN

Dengan metode One Class SVM diperoleh nilai precision = 1,0, recall = 0,993, f-1 score = 0,997, dan accuracy sebesar 0,997 atau 99,7% dengan menggunakan parameter gamma = 1, epsilon = 0,01, dan nu = 0,01. Sedangkan dengan metode K-Means diperoleh nilai precision = 0,901, recall = 0,973, f-1 score = 0,936, dan accuracy sebesar 93,3 % dengan menggunakan k=12 cluster.

Metode One Class SVM memiliki nilai precision, recall, f-1 score, dan accuracy yang bernilai sedikit lebih tinggi dibandingkan dengan metode K-Means. Hal itu berarti di dalam penelitian ini metode One Class SVM sedikit lebih unggul dibandingkan dengan metode K-Means. Namun, baik metode One Class SVM ataupun metode K-Means dapat digunakan untuk penelitian ini, hanya saja untuk tingkat keakuratannya tergantung nilai optimasi parameter untuk OSVM dan jumlah cluster untuk K-Means.

Hasil F-Score tersebut menunjukkan bahwa metode *one-class classification* berpotensi digunakan untuk mengidentifikasi penderita DHF berdasarkan uji tes darah di laboratorium. Kinerja model ini juga jauh lebih tinggi dibanding metode klasifikasi *2-class/multiclass* yang pernah diteliti oleh peneliti lain. Namun untuk lebih memastikannya, diperlukan validasi dengan data yang lebih banyak, disamping juga mengujinya dengan metode klasifikasi biasa.

DAFTAR PUSTAKA

Astuti, T., Suciati, Mujiati, I., Ayu, D., Ristianah, V., & Lestari, W. (2016). Penerapan Algoritme J48 Untuk Prediksi Penyakit Demam Berdarah. *Telematika , 9* (2), 1-10.

Cabral, G., & de Oliveira, A. (2014). One-class Classification for heart disease diagnosis. *2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (hal. 2551-2556). IEEE.

Haryanto, T. (2013). *Prediksi Penyakit Demam Berdarah Dan Typhus Dengan Algoritma C5.0.* Skripsi, Telkom University.

Khan, S., & Madden, M. (2010). A Survey of Recent Trends in One Class Classification. *Artificial Intelligence and Cognitive Science*, (hal. 188-197).

Khan, S., & Madden, M. (2014). One-class classification: taxonomy of study and review of techniques. *The Knowledge Engineering Review , 29* (3), 345-374.

Lesmana, I., Hikmah, F., & Karimah, R. (2014). Model Prediktif Identifikasi Tersangka Tuberkulosis Dan Demam Berdarah Menggunakan Data Mining. *Seminar Nasional Teknologi Informasi dan Multimedia* (hal. 2.02.1-2.02.6). STMIK AMIKOM Yogyakarta.

Munah, M. (2015). *Implementasi Algoritma C4.5 Untuk Mengklasifikasi Penyakit Tipes Dan DBD.* Universitas Dian Nuswantoro.

Nagi, J., Yap, K., Tiong, S., Ahmed, S., & Mohamad, M. (2010). Nontechnical Loss Detection for Metered Customers in Power Utility Using Support Vector Machines. *IEEE Transactions on Power Delivery* (hal. 1162-1171). IEEE.

Schölkopf, B., Williamson, R., Smola, A., Shawe-Taylor, J., & Platt, J. Support vector method for novelty detection. *Proceedings of the 12th International Conference on Neural Information Processing Systems (NIPS'99).*

Suwardi, U. (2012). Komparasi Algoritma Backpropagation, Nearest Neighbor, Dan Decision Tree Untuk Mendeteksi Penyakit Demam Berdarah Pada Pasien Opname. *Jurnal Teknologi Informasi , 8* (1), 57-67.

Tax, D. (2001). *One-class classification.* Desertasi, TU Delft.

Zhang, J., Lu, J., & Zhang, G. (2011). Combining one class classification models for avian influenza outbreaks. *IEEE Symposium on Computational Intelligence in Multicriteria Decision-Making (MDCM).* IEEE.