Top-k Feature Selection untuk Deteksi Penyakit Hepatitis Menggunakan Algoritme Naïve Bayes

Riska Wibowo[1], Henny Indriyawati[2]
[1]Program Studi Teknik Informatika, [2]Program Studi Sistem Informasi,
Fakultas Teknologi Informasi dan Komunikasi, Universitas Semarang, Indonesia
Jl. Soekarno Hatta Tlogosari, Semarang 50196, Jawa Tengah, Indonesia
Email: [1]bowo@usm.ac.id, [2]henny@usm.ac.id

Masuk: 09 Mei 2019; Direvisi: 24 Januari 2020; Diterima: 01 April 2020

Abstract. *Becoming one of the society health problems in the world, hepatitis is an inflammation liver disease caused by a virus, bacterial infection, chemical substances including drugs and alcohol. In this research, for the dataset of hepatitis having high dimensionality, its value for each attribute was calculated using weight information gain method. Then, the attributes were selected by using top-k methods and were classified by using Naïve Bayes Algorithm respectively. This research showed that 9 out of 20 attributes had chosen to be the highest top-9 with an accuracy rate of 85.57%. Later on, this research can be useful for a consideration in a decision making process for various subjects related to feature selection and Naïve Bayes Algorithm method and also for predicting hepatitis.*
Keywords: *data mining, weight information gain, Naïve Bayes algorithm*

Abstrak. *Penyakit hepatitis merupakan masalah kesehatan masyarakat di dunia. Penyakit hepatitis merupakan penyakit peradangan hati yang disebabkan oleh virus, infeksi bakteri, zat-zat kimia termasuk obat-obatan dan alkohol. Pada penelitian ini, dataset hepatitis yang memiliki data berdimensi tinggi akan dihitung nilai bobot dari masing-masing atribut menggunakan metode weight information gain. Setelah dihitung nilai bobot dilakukan pemilihan atribut, atribut yang dipilih menggunakan metode top-k. Kemudian dilakukan klasifikasi menggunakan algoritme Naïve Bayes. Hasil penelitian menunjukkan dari 20 atribut, terpilih top-9 tertinggi dengan nilai akurasi 85.57%. Dengan adanya penelitian ini dapat digunakan sebagai bahan pertimbangan dan pengambilan keputusan pada berbagai bidang yang berkaitan dengan metode feature selection, algoritme Naïve Bayes, dan di dalam memprediksi penyakit hepatitis.*
Kata Kunci*: data mining, weight information gain, algoritma Naïve Bayes*

1. Pendahuluan

Penyakit hepatitis merupakan masalah kesehatan masyarakat di dunia. Penyakit hepatitis menyebabkan 1,34 juta kematian pada tahun 2015, jumlah itu sebanding dengan kematian yang di sebabkan oleh penyakit tuberkulosis dan lebih tinggi dari penyakit HIV. Penyakit hepatitis merupakan penyakit peradangan hati yang disebabkan oleh virus, infeksi bakteri, zat-zat kimia termasuk obat-obatan dan alkohol [1].

Melihat banyaknya kasus kematian yang disebabkan oleh penyakit hepatitis, maka diperlukan satu langkah dini sebagai upaya pencegahan penyakit hepatitis. Seiring dengan perkembangan dalam dunia IT *(Information and Technology),* kehadiran cabang ilmu di bidang data mining telah menarik banyak perhatian. Seperti melakukan prediksi untuk penentuan prediksi penyakit hepatitis.

Dataset hepatitis yang digunakan pada penelitian ini mempunyai atribut (*feature*) bertipe data numerik dan nominal. Dan mempunyai kelas bertipe data nominal, sehingga penerapan metode data mining masuk ke dalam metode klasifikasi. Pada penelitian[2], melakukan komparasi algoritme Naïve Bayes dan k-NN, hasil penelitian menunjukan akurasi algoritme

Naïve Bayes lebih baik di banding dengan algoritme k-NN. Menurut Wu [3], algoritme Naïve Bayes masuk ke dalam 10 besar algoritme terbaik, khususnya algoritme klasifikasi.

Algoritme Naïve Bayes merupakan teknik *machine learning* yang populer untuk klasifikasi, karena memiliki peforma yang baik, efisien dan sangat sederhana. Sebagai *classifier*, Naïve Bayes sangat efisien dan sederhana serta sangat sensitif terhadap seleksi fitur (*feature selection*) [4]. Dalam hal ini, Naïve Bayes membuktikan tingkat akurasi yang bagus saat klasifikasi dianggap seimbang.

Algoritme Naïve Bayes memiliki kekurangan yaitu sangat sensitif pada fitur yang terlalu banyak, sehingga membuat akurasi menjadi rendah. Upaya untuk meningkatkan kinerja algoritme Naive Bayes pada data berdimensi tinggi dengan menggunakan beberapa pendekatan, salah satu metode yang digunakan adalah seleksi fitur (*feature selection*) [5]. Terdapat tiga teknik yang digunakan dalam metode seleksi fitur (*feature selection*) yaitu teknik yaitu *filter, wrapper*, dan *hybrid* [6]. Teknik *filter* merupakan hubungan/relevansi antar atribut berdasarkan sifat intrinsik dari data. Teknik *wrapper* memilih attribut berdasarkan dari evalusai kinerja *classifier.* Sedangkan teknik *hybrid* adalah menggabungkan antara teknik *filter* dan teknik *wrapper.*

Pada penelitian ini, menggunakan, metode *weight information gain.* Metode *weight information gain* masuk ke dalam teknik *filter.* Teknik *filter* lebih cepat digunakan dibandingkan teknik *wrapper* dan teknik *hybrid*, selain itu teknik *filter* lebih baik dan mudah diterapkan dari pada teknik *wrapper* dan teknik *hybrid* [7].

Pada penelitian ini dataset hepatitis yang memiliki data berdimensi tinggi akan dihitung nilai bobot dari masing-masing atribut menggunakan metode *weight information gain*. Setelah dihitung nilai bobot dilakukan pemilihan atribut, atribut yang dipilih menggunakan metode *top*-k. Algoritme klasifikasi yang digunakan adalah Naïve Bayes. Dengan adanya penelitian ini dapat digunakan sebagai bahan pertimbangan dan pengambilan keputusan pada berbagai bidang yang berkaitan dengan metode *feature selection,* algoritme Naïve Bayes, dan di dalam memprediksi penyakit hepatitis.

Dari latar belakang diatas, identifikasi masalah pada penelitian ini bahwa data yang mempunyai atribut banyak dan masing-masing atribut yang tidak relevan menyebabkan algoritme klasifikasi Naïve Bayes menjadi rendah. Berdasarkan latar belakang dan identifikasi masalah diatas, maka rumusan masalah pada penelitian ini adalah bagaimana penerapan metode *feature selection* dan algoritme Naïve Bayes pada data berdimensi tinggi untuk deteksi penyakit hepatitis. Tujuan dari penelitian ini adalah menerapkan metode pembobotan data dengan metode *weight information gain* dan pemilihan atribut dengan metode *top-k.* Kemudian klasifikasi menggunakan algoritme Naive Bayes pada data berdimensi tinggi pada dataset penyakit hepatitis.

2. Tinjauan Pustaka

2.1. Hepatitis

Hepatitis merupakan penyakit peradangan hati yang disebabkan oleh virus, infeksi bakteri, zat-zat kimia termasuk obat-obatan dan alkohol. Penyakit hepatitis merupakan masalah kesehatan masyarakat di dunia. Penyakit hepatitis menyebabkan 1,34 juta kematian pada tahun 2015, jumlah itu sebanding dengan kematian yang di sebabkan oleh penyakit tuberkulosis dan lebih tinggi dari penyakit HIV. Gejala hepatitis terbagi dalam 4 tahap yaitu fase *inkubasi*, fase *prodromal*, fase *icterus*, dan fase *konvalesen* [1].

2.2. Data Mining

Data mining adalah proses menemukan pola-pola yang menarik dan pengetahuan dari data yang besar. Sumber data dapat mencakup *database, warehouses, web, repository* , atau informasi lainnya [8]. Data mining menjadi salah satu cabang bidang ilmu keilmuan dan populer dalam dunia komputer. Sehingga beberapa perusahaan-perusahaan besar terus mengembangkan dan menyempurnakan metodologi dalam data mining. Berdasarkan peran utama data mining dibagi menjadi beberapa kelompok, yaitu: estimasi, prediksi, klasifikasi, klastering, dan asosiasi.

2.3. Weight Information Gain

Weight Information gain adalah metode pembobotan tiap variabel yang paling umum dari atribut evaluasi [7]. Dalam menghitung *information gain* harus memahami suatu aturan lain terlebih dahulu yang disebut *entropy*. Di dalam bidang *Information Theory*, kita sering menggunakan *entropy* sebagai suatu parameter untuk mengukur heterogenitas (keberagaman) dari suatu kumpulan sampel data. Jika kumpulan sampel data semakin heterogen, maka nilai *entropy* semakin besar [9]. Metode *weight information gain* berbasis pada penghitungan nilai e*ntropy* (persamaan 1) dan penghitungan $information\ gain$ (persamaan 2).

$$Entropy = \sum_{i}^{C} - p_i \log_2 p_i \tag{1}$$

$$Gain(S, A) = Entropy(S) - \sum_{v \epsilon Value(A)} \frac{|S_v|}{|S|} Entropy(S_v) \tag{2}$$

2.4. Algoritme Naïve Bayes

Algoritme Naïve Bayes merupakan teknik *machine learning* yang populer untuk klasifikasi teks, karena memiliki peforma yang baik, efisien dan sangat sederhana. Sebagai *classifier*, Naïve Bayes sangat efisien dan sederhana serta sangat sensitif terhadap seleksi fitur [4]. Algoritme Naïve Bayes merupakan algoritme klasifikasi yang menggunakan metode probabilitas dan statistik. Algoritme ini diusulkan oleh ilmuwan Inggris *Revered Thomas Bayes*. Algoritme ini memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya, oleh karena itu dikenal sebagai Teorema *Bayes*. Teorema tersebut dikombinasikan dengan Naïve[10]. Teorema Bayesian menghitung nilai *posterior probability* $P(H|X)$ menggunakan probabilitas $P(H)$, $P(X)$, dan $P(X|H)$ (persamaan 3).

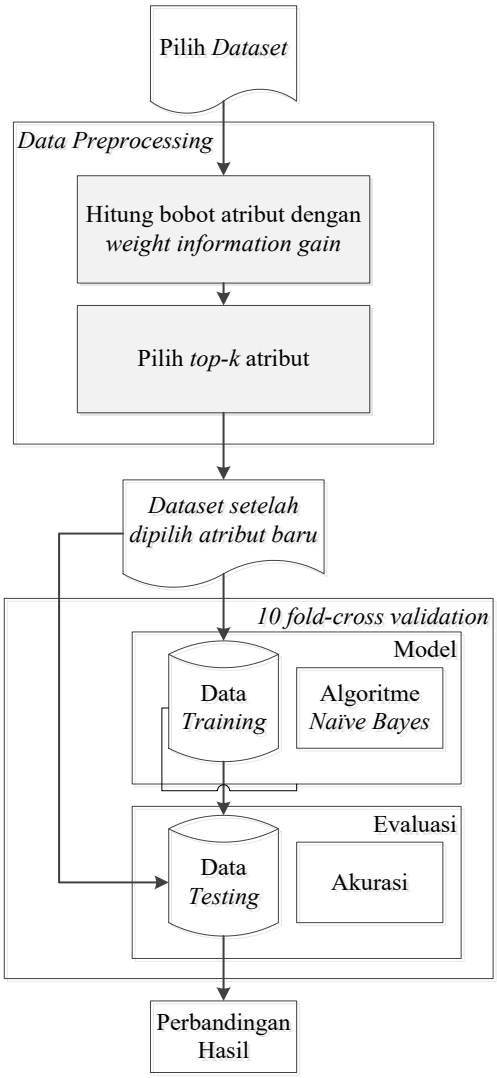
$$P(H|X) = \frac{P(X|H).P(H)}{P(X)} \tag{3}$$

2.5. Metodologi Penelitian Eksperimen

Metodologi penelitian yang digunakan pada penelitian ini menggunakan metode eksperimen. Metode penelitian eksperimen adalah ujicoba yang dikontrol oleh peneliti sendiri untuk melakukan investigasi hubungan kausal (hubungan sebab-akibat) [11]. Langkah-langkah penelitian meliputi: (1) Analisa permasalahan dan tinjauan pustaka. Penelitian ini diawali dengan mengumpulkan *technical paper* dan *survey paper* dengan topik data berdimensi tinggi pada algoritme Naïve Bayes. (2) Pengumpulan *dataset*, *dataset* yang digunakan pada penelitian ini adalah *dataset public* UCI *machine learning repository,* yaitu dataset hepatitis. (3) Pengolahan data, pada penelitian ini *dataset* yang mengandung *missing value* diganti dengan nilai rata-rata dari masing-masing atribut. (4) Metode yang diusulkan, pada tahapan ini *dataset* hepatitis terlebih dahulu dilakukan seleksi terhadap atribut-atribut berdimensi tinggi. Pada tahapan awal untuk penentuan atribut yang nantinya akan digunakan dalam perhitungan algoritme klasifikasi yaitu metode *information gain.* Atribut yang dipilih menggunakan metode *top-k,* dan metode klasifikasi yang digunakan adalah Naïve Bayes. (5) Tahapan eksperimen pada penelitian ini menggunakan metode CRISP-DM *(Cross Industry standard Process for Data Mining)*, merupakan salah satu metodologi dalam data mining. (6) Evaluasi hasil, setelah dilakukan eksperimen terhadap semua *dataset* dengan metode yang diusulkan maka menghasilkan nilai-nilai akurasi yang kemudian hasil tersebut dievaluasi dan divalidasi. Dari hasil tersebut dapat ditarik kesimpulan dari penelitian eksperimen ini.

3. Metodologi Penelitian

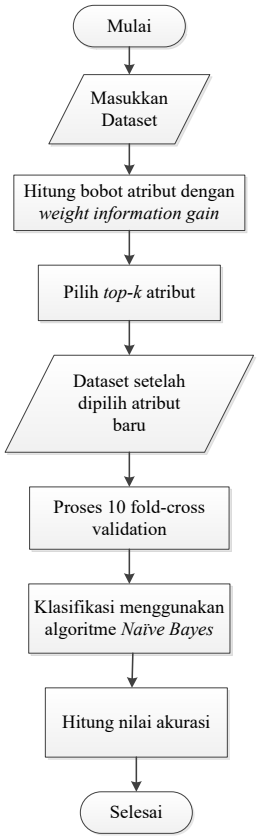
Pada tahapan ini dijelaskan metode yang diusulkan pada penelitian. Dataset hepatitis terlebih dahulu akan dilakukan seleksi terhadap atribut-atribut berdimensi tinggi yang ada pada dataset tersebut. Pada tahap awal untuk penentuan atribut yang nantinya akan dibuang atau dipakai dalam perhitungan algoritme klasifikasi yaitu dengan metode *information gain.* Setelah dihitung nilai bobot dilakukan pemilihan atribut, atribut yang dipilih menggunakan metode *top-k*. Metode klasifikasi yang digunakan adalah Naïve Bayes, sehingga diperoleh nilai akurasi terbaik. Setelah mendapatkan nilai akurasi terbaik kemudian tahapan *deployment* yaitu pembuatan aplikasi *User Interface* dari hasil data mining untuk tampilan kepada *user* berupa hasil model yang dibuat. Tahapan metode yang diusulkan dapat dilihat pada Gambar 1.



Gambar 1. Metode yang Diusulkan

Flowchart menggambarkan langkah-langkah urutan metode yang diusulkan dalam penelitian ini, dapat disajikan dengan sistematis menggunakan simbol. *Flowchart* metode yang digunakan dapat dilihat pada Gambar 2. Berikut adalah langkah-langkah metode yang diusulkan dalam penelitian ini: (1) Siapkan *dataset*. (2) Hitung bobot atribut dengan menggunakan metode *weight information gain.* (3) Pemilihan atribut dengan menggunakan metode *top-k.* (4) *Dataset* setelah dipilih atribut baru. (5) *Dataset* yang telah dipilih atribut baru tesebut, dibagi dua yaitu menjadi data training dan data testing menggunakan *10 fold-cross validation.* (6) Klasifikasi

menggunakan algoritme Naïve Bayes. (7) Hitung nilai akurasi berdasarkan table *confusion matrix.*

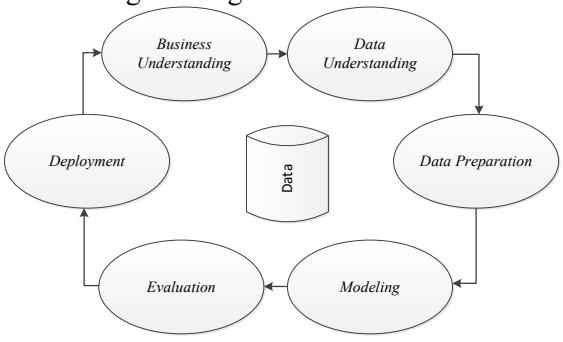


Gambar 2. ***Flowchart Metode yang Diusulkan***

4. Hasil dan Diskusi

4.1. Eksperimen Menggunakan CRISP-DM

Eksperimen ini menggunakan program bantu Rapidminer versi 8.1 dan Microsoft Excel 365. Langkah-langkah penelitian menggunakan metode *Cross Industry Standard Process Data Mining* (CRISP-DM). Data mining menjadi salah satu cabang bidang keilmuan baru dan populer dalam dunia komputer. Sehingga beberapa perusahaan-perusahaan besar terus mengembangkan dan menyempurnakan metodologi dalam data mining, salah satu metodologi yang berhasil dirumuskan adalah CRISP-DM *(Cross Industry Standard Process for Data Mining).* CRISP-DM diperkenalkan pada tahun 1999 oleh empat perusahaan besar, yaitu perusahaan pembuat mobil Daimler-Benz, produsen perangkat keras dan perangkat lunak NRC Corp, penyedia asuransi OHRA, dan perusahaan pembuat perangkat statistik SPSS, Inc. Gambar 3 menunjukkan 6 (enam) tahapan dalam metodologi Berikut langkah-langkah metode CRISP-DM:



Gambar 3. Metode CRISP-DM ***(Cross Industry Standard Process for Data Mining)***

4.2. Business Understanding

Pada penelitian ini penulis tertarik dalam bidang penelitian data mining untuk kesehatan. Topik penelitian yang diangkat adalah hepatitis. Penulis ingin mengidentifikasi apakah seseorang tersebut terserang penyakit hepatitis atau tidak terserang penyakit hepatitis dengan menerapkan metode data mining yaitu Naïve Bayes.

4.3. Data Understanding

Dataset yang digunakan pada penelitian ini berasal *dataset public* yang sering digunakan yaitu UCI *machine learning repository. Dataset* hepatitis dapat diunduh melalui situs *https://archive.ics.uci.edu/ml/datasets.html*. *Dataset* hepatitis yang terdiri dari 20 atribut dan 155 *row* data. Tabel 1 menunjukkan deskripsi dataset hepatitis.

Tabel 1. Deskripsi Dataset Hepatitis

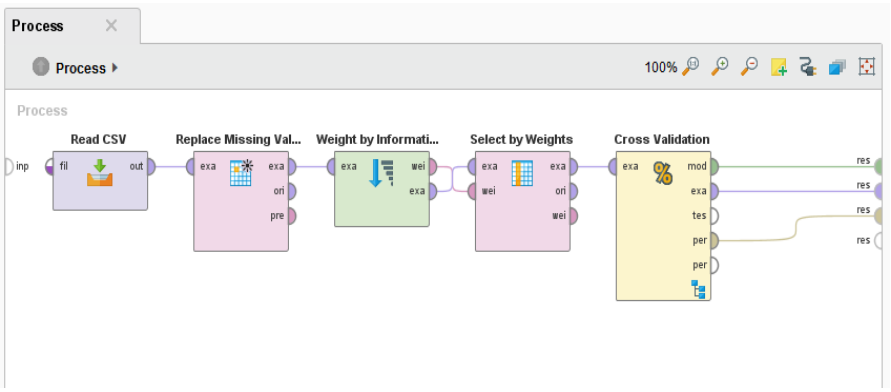
No	*Attribut Name*	*Description*	*Type Data*	*Value*
1	*age*	usia dalam tahun	integer	7 – 78 tahun
2	*sex*	jenis kelamin	*binominal*	*male, female*
3	*steroid*	obat anti inflamasi	*binominal*	*no, yes*
4	*antivirals*	kualitas hidup	*binominal*	*no, yes*
5	*fatigue*	kelelahan	*binominal*	*no, yes*
6	*malaise*	rasa tidak nyaman, lemas	*binominal*	*no, yes*
7	*anorexia*	kehabisan oksigen	*binominal*	*no, yes*
8	*liver big*	liver membesar	*binominal*	*no, yes*
9	*liver firm*	liver keras	*binominal*	*no, yes*
10	*spleen palpable*	organ limfatik	*binominal*	*no, yes*
11	*spiders*	pembuluh darah abnormal	*binominal*	*no, yes*
12	*ascites*	penumpukan cairan	*binominal*	*no, yes*
13	*varices*	aliran darah menuju hati tersumbat	*binominal*	*no, yes*
14	*bilirubin*	pigmen berwarna kuning kecokelatan di dalam empedu	numerik	0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1, 1.1, 1.2, 1.3, 1.4, 1.5, 1.6, 1.7, 1.8, 1.9, 2, 2.2, 2.3, 2.4, 2.5, 2.8, 2.9, 3, 3.2, 3.5, 3.9, 4.1, 4.1, 4.6, 4.8, 7.6, 8
15	*alk phosphate*	enzim *hydrolase* yang di temukan di dalam hati	numerik	26, 30, 34, 40, 44, 45, 46, 48, 50, 52, 53, 54, 55, 56, 57, 58, 59, 60, 62, 63, 65, 67, 68, 70, 71, 72, 74, 75, 76, 78, 80, 81, 82, 84, 85, 86, 89, 90, 92, 95, 96, 100, 102, 103, 107, 109, 110, 114, 115, 118, 119, 120, 1223, 125, 126, 127, 130, 133, 135, 138, 139, 141, 147, 150, 155, 160, 164, 165, 166, 167, 168, 175, 179, 180, 181, 191, 215, 230, 243, 256, 280, 295
16	*sgot*	enzim-enzim pada hati yang akan meningkat jumlahnya di dalam tubuh jika hati mengalami kerusakan baik kerusakan fungsi hati secara akut maupun kronis	numerik	14, 15, 16, 18, 19, 20, 23, 24, 25, 28, 29, 30, 31, 32, 33, 34, 38, 39, 40, 42, 43, 44, 45, 46, 48, 49, 52, 53, 54, 55, 58, 59, 60, 63, 64, 65, 68, 69, 70, 75, 78, 80, 81, 86, 89, 90, 92, 98, 100, 101, 110, 114, 117, 118, 120, 123, 128, 136, 140, 142, 144, 145, 150, 152, 153, 156, 157, 166, 173, 181, 182, 200, 224, 225, 227, 231, 242, 249, 269, 271, 278, 420, 528, 648
17	*albumin*	protein yang paling banyak terdapat dalam aliran darah dan diproduksi oleh hati	numerik	2.1, 2.2, 2.4, 2.6, 2.7, 2.8, 3, 3.1, 3.3, 3.4, 3.5, 3.6, 3.7, 3.8, 3.9, 4, 4.1, 4.2, 4.3, 4.4, 4.5, 4.6, 4.7, 4.8, 4.9, 5, 5.3, 6.4
18	*protime*	pembekuan hati	numerik	0, 21, 23, 29, 30, 31, 32, 35, 36, 38, 39, 40, 41, 42, 43, 46, 47, 48, 50, 51, 52, 54, 56, 57, 58, 60, 62, 63, 64, 66, 67, 70, 72, 73, 74, 75, 76, 77, 78, 80, 84, 85, 90, 100
19	*histology*	kondisi dan fungsi jaringan dalam hubungannya dengan penyakit	*binominal*	no, yes
20	*class*	atribut yang berisi dua pernyataan	*binominal*	*die, live*

4.4. Data Preparation

Dalam pengolahan dan visualisasi data pada penelitian ini menggunakan aplikasi bantu Rapidminer versi 8.1. Berdasarkan data *understanding* yang sudah dijelaskan pada sub bab sebelumnya, dapat dilihat bahwa pada *dataset* hepatitis terdiri dari 20 (duapuluh) atribut. Terdapat atribut yang mengandung data *missing value* yaitu *steroid, fatigue, malaise, anorexia, liver big, liver firm, spleen palpable, spiders, ascites, varices, bilirubin, alk phosphate, sgot, albumin, dan protime*. *Dataset* yang mengandung *missing value* diganti dengan nilai rata-rata dari masing-masing atribut [8].

4.5. Modeling

Pada tahapan *business understanding* dijelaskan bahwa tujuan dari studi kasus ini adalah *top-k feature selection* untuk deteksi penyakit hepatitis. Berdasarkan tipe data pada tahapan *data understanding*, diketahui bahwa tipe data pada semua atribut adalah numerik, sedangkan tipe data pada kelas adalah *binominal*. Oleh sebab itu, peranan data mining klasifikasi akan diterapkan. Algoritme yang digunakan pada peranan data mining klasifikasi adalah Naïve Bayes. Supaya mendapatkan hasil evaluasi, perlu dimodelkan masing-masing algoritme menggunakan Rapidminer. Gambar 4 menunjukan tahapan *modeling* pada penelitian ini.



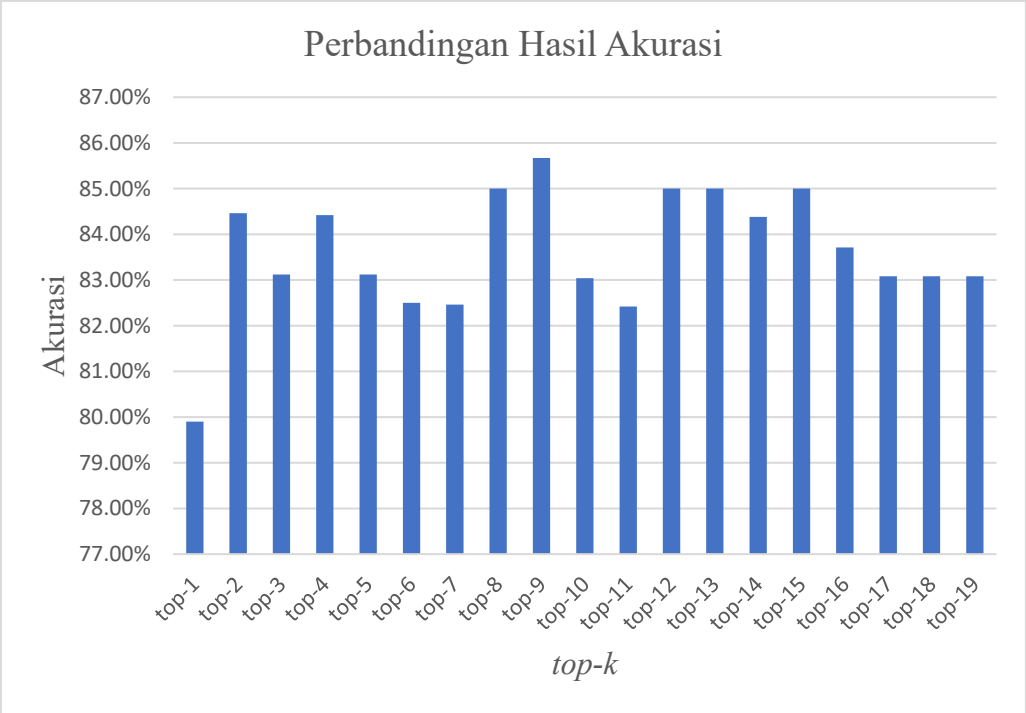
Gambar 4. Tahapan *Modeling*

4.6. Evaluation

Peranan data mining yang digunakan untuk mengekstraksi data menjadi informasi, pola, dan pengetahuan pada tahapan modeling adalah peranan data mining klasifikasi dengan algoritme Naïve Bayes. *Top-k* atribut akan dibandingkan hasil evaluasinya dan hasil evaluasi terbaik akan digunakan untuk tahapan deployment. Pengukuran evaluasi pada peranan data mining klasifikasi adalah mengukur akurasi, penghitungan akurasi berdasarkan pada *confusion matrix* yang terbentuk. Dimana nilai TP (*true positif*) adalah jumlah data yang diprediksi benar dan kenyataannya benar, nilai TN (*true negative*) adalah jumlah data yang diprediksi salah dan kenyataannya salah, FP (*false positif*) adalah jumlah data yang diprediksi benar, tapi kenyataannya salah, sedangkan FN (*false negatif*) adalah jumlah data yang diprediksi salah, tapi pada kenyataanya benar. Formula untuk menghitung akurasi menggunakan persamaan.

$$Akurasi = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

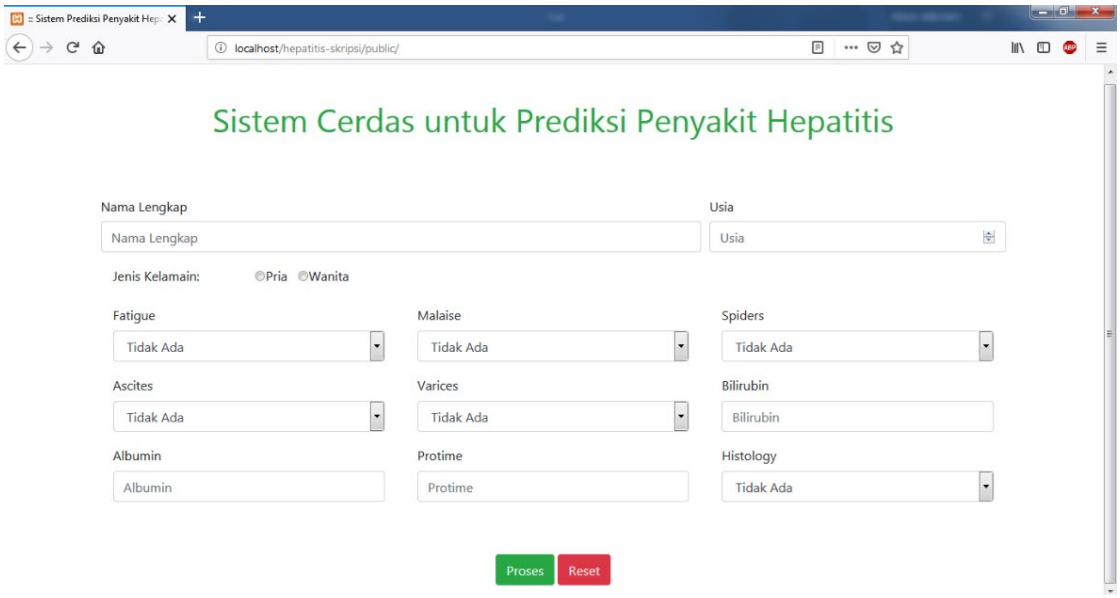
Dari *confusion matrix* tersebut dihitung nilai akurasi masing-masing *top-k*, sehingga didapatkan hasil akurasi. Perbandingan hasil akurasi masing-masing *top-k* dapat dilihat pada Gambar 5.



Gambar 5. Perbandingan Hasil Akurasi Masing-Masing *Top-k*

4.7. Deployment

Berdasarkan hasil pada tahapan evaluasi (Gambar 5), dapat disimpulkan bahwa hasil akurasi *top-9* lebih tinggi daripada *top-k* lainnya. Sehingga pada tahapan deployment, sistem cerdas *top-9 feature selection* akan dibangun untuk prediksi penyakit hepatitis menggunakan algoritme Naïve Bayes dengan bahasa pemrograman PHP. Gambar 6 menunjukkan aplikasi sistem cerdas untuk prediksi penyakit hepatitis.



Gambar 6. Sistem Cerdas untuk Prediksi Penyakit Hepatitis

5. Simpulan dan Saran

Algoritme Naive Bayes merupakan salah satu algoritme *machine learning* yang populer, karena sederhana sehingga memiliki efisiensi komputasi yang tinggi dan memiliki akurasi yang baik. Metode Naïve Bayes memiliki kekurangan yaitu sangat sensitif pada fitur yang terlalu banyak, sehingga membuat akurasi menjadi rendah. Sebuah metode diusulkan untuk meningkatkan kinerja dari algoritme Naïve Bayes, yaitu pembobotan atribut dengan menggunakan metode *weight information gain.* Setelah dihitung nilai bobot dilakukan pemilihan atribut, atribut yang dipilih menggunakan metode *top*-k. Hasil penelitian menunjukkan dari 20 atribut, terpilih *top*-9 atau 9 atribut tertinggi dengan nilai akurasi 85.57%.

Referensi

[1] World Health Organization. (2017). *Global Hepatitis Report*.

[2] L. Dey, S. Chakraborty, A. Biswas, B. Bose, and S. Tiwari, Sentiment analysis of review datasets using naïve bayes‘ and K-NN classifier, *Int. J. Inf. Eng. Electron. Bus.*, vol. 8(4), pp. 54–62, 2016, [Online] doi: 10.5815/ijieeb.2016.04.07.

[3] X. Wu and V. Kumar, The Top Ten Algorithm in Data Mining. Boca Raton: Taylor & Francis Group, 2009.

[4] J. Chen, H. Huang, S. Tian, and Y. Qu, Feature selection for text classification with naïve bayes, *Expert Syst. Appl.*, vol. 36(3) PART 1, pp. 5432–5435, 2009, [Online] doi: 10.1016/j.eswa.2008.06.054.

[5] R. S. Wahono and N. S. Herman, Genetic feature selection for software defect prediction, *Adv. Sci. Lett.*, vol. 20(1), pp. 239–244, 2014, [Online] doi: 10.1166/asl.2014.5283.

[6] G. Chen and J. Chen, A novel wrapper method for feature selection and its applications, *Neurocomputing*, vol. 159(1), pp. 219–226, 2015, [Online] doi: 10.1016/j.neucom.2015.01.070.

[7] V. Bolón-Canedo, I. Porto-Díaz, N. Sánchez-Maroño, and A. Alonso-Betanzos, A framework for cost-based feature selection, *Pattern Recognit.*, vol. 47(7), pp. 2481–2489, 2014, [Online] doi: 10.1016/j.patcog.2014.01.008.

[8] J. Han, M. Kamber, and J. Pei, Data Mining Concepts and Techniques, 3rd ed. Waltham: Elsevier Inc., 2012.

[9] J. Suntoro and C. N. Indah, Average weight information gain untuk menangani data berdimensi tinggi menggunakan algoritma c4.5, *Jurnal Buana Informatika,* vol. 8(3), pp. 131–140, 2017.

[10] Bustami, Penerapan algoritma naive bayes, *J. Inform.*, vol. 8(1), pp. 884–898, 2014, [Online] doi: 10.1364/OFC.2009.OWD2.

[11] C. W. Dawson, Projects in Computing and Information Systems, vol. 2. United States of America: Addison-Wesley, 2011.