PERBANDINGAN ALGORITMA KLASIFIKASI UNTUK PREDIKSI PENYAKIT LIVER

Muhammad RizkiFahdia[1]

Program Studi Sistem Informasi STMIK Nusa Mandiri
www.nusamandiri.ac.id
rizki.muz@nusamandiri.ac.id

Abstract—*The liver is the largest organ in the human body, weighing around 3 lbs. The liver is located on the right side of the abdominal cavity below the diaphragm. Where the liver itself acts as a blood filter in the body. Almost all of the blood passes through the liver. liver damage causes complications for other diseases. Several studies were conducted to diagnose patients affected by liver disease. From the process above, it is possible to do research by finding correlations between attributes that influence liver disease. Data mining processing is one of the solutions in research to find linkages between data attributes to gain knowledge and patterns. In addition, it must be explained how far the level of cursation obtained from the model. By comparing the five classification algorithms specified, including Decision Tree (C4.5), Naive Beyes (NB), K-Nearest Neighbor (kNN), Log-R, Deep learning. one of the best algorithms is used for decision making in determining which patients are affected by liver disease. In this study we used a dataset obtained from the UCI repository consisting of 583 patients with 11 attributes. comparisons obtained from the five algorithms using algorithms with Rapidminer software to find out which is the most accurate in predicting liver disease. To determine the accuracy indicator, use the Area Under Curve (AUC) method and a different test (t-Test). The comparison of the five Decision Tree (C4.5) algorithms is the algorithm with the best results, with accuracy (72.56%) and AUC (0.594), to get better performance with Feature Extraction and Feature Selection, there is an increase to 73.24% (accuracy) and 0.602 (AUC).*

Keywords - Liver, AlgoritmaKlasifikasi

Abstrak - Hati merupakan oragn terbesar dalam tubuh manusia, beratnya sekitar 3 lbs. Hati letaknya berada di sisi kanan rongga perut di bawah diafragma. Dimana hati itu sendiri bertindak sebagai filter darah dalam tubuh. Hampir semua darah melewati liver. kerusakan liver menimbulkan komplikasi terhadap penyakit yang lain. Beberapa penelitian dilakukan untuk mendiagnosis pasien yang terkena penyakit liver. Dari permaslahan diatas hal yang memungkinkan adalah melakukan penelitian dengan cara menemukan korelasi antar atribut yang berpengaruh terhadap timbulnya penyakit liver. Pengolahan data mining menjadi salah satu solusi dalam penelitian untuk menemukan keterkaitan antar atribut data untuk mendapatkan pengetahuan dan pola. Selain itu harus dikatehui sejauh mana tingkat kaurasi yang didapat dari model. Dengan cara membandingkan lima algoritma klasifikasi yang ditentukkan, diantaranya Decision Tree (C4.5), Naive Beyes(NB), K-Nearest Neighbor(kNN), Log-R, Deep learning. salah satu algorima terbaik dijadikan untuk pengambilan keputusan dalam menentukan pasien yang terkena atau tidaknya terhadap penyakit liver. Dalam penelitian ini kami menggunakan dataset yang didapat dari repositori UCI yang terdiri dari 583 pasien catatan dengan 11 atribut. perbandingan yang didapat dari ke lima algoritma menggunakan algoritma dengan perangkat lunak Rapidminer untuk mengetahui mana yang paling akurat dalam memprediksi penyakit liver. Untuk menentukan indikator akurasi menggunakan metode Area Under Curve(AUC) dan uji beda (t-Test). Dari perbandingan kelima algoritma Decision Tree (C4.5) merupakan algoritma dengan hasil paling baik, dengan tingkat accuracy (72.56%) dan AUC (0.594), untuk mendapatkan performa yang lebih baik dilakukan Feature Extraction dan Feature Selection, ada kenaikan menjadi 73.24% (accuracy) dan 0.602(AUC).

Kata kunci - Liver, AlgoritmaKlasifikasi

PENDAHULUAN

Hatimerupakanoragnterbesardalamtubuhmanusia , beratnyasekitar 3 lbs. Hatiletaknyaberada di sisikananronggaperut di bawahdiafragma(Kumar & Katyal, 2018). Setiapgangguanhatidapatmenyebabkanperadanganakutatauperadangankronis, kelemahanhati, dan bahkandapatmerusak organ lain di dalamtubuh. Pada

berbagaigangguanhatiperlumembutuhkanperawatankhusus oleh praktisimedisatauprofesionaldalamperawatankesehatan(Kumar & Katyal, 2018)

Penelitisebelumnyasudahbanyak yang melakukanpenelitiantentangpenyakit liver denganmenggunakanberbagaialgoritma, diantaranyaadalahpenelitiantentangmemprediksikematianjangkapendek yang disebabkan salah satunya oleh penakit liver pada pasien yang kerumahsakituntukdirawatdengansirosismenggunakanalgoritmaregresilogistik dan jaringansarafmemorijangkapendek(Harrison et al., 2018).

Peneliti yang memprediksi virus hepatitis C terkaitpenyakit liver denganmenggunakanjaringansaraf polynomial (Lara et al., 2015). Prediksikomorbiditasterkaitpenyakit liver menggunakanJaringan Bayesian yang Resampling dan DinamisdenganVariabelLaten(Yousefi et al., 2017). Peneliti lain melakukanpenelitiantentang diagnosis non-invasifpenyakithatiberlemak non-alkohol (NAFLD) menggunakanechogenisitasgambar ultrasound (Benjamin et al., 2017). Penelitiberikutnyamelakukanpenelitiantentangoptimalisasiaturanklasifikasi C5.0 Boosted MenggunakanAlgoritmaGenetikauntukPrediksipenyakithati(Hassoon et al., 2017).

Penelitiantentangperbandinganpendekatanpembelajaranmesinuntukprediksi fibrosis hatilanjut pada Pasien Hepatitis C Kronik(Hashem et al., 2018). Dari keseluruhanpenelitiantersebutsemuanyamenggunakan dataset yang beragam dan tentusajahasildaripenelitianmenyelesaikanpermasalahandari dataset masing-masing.

Pada penelitianinimenggunakandataset liver yang didapatdarirepositori UCI (Universal Child Immunization) yang terdiridari 583 pasiencatatandengan 11 atribut.

Target penelitianini juga mengacu pada atribut yang digunakanuntukmenentukanapakahpasienterkenapenyakit liver atautidak. Ketika doktermemutuskanuntukmemeriksapasienmakahal yang pertamadilakukanadalahmengajukantes bilirubin bersamadenganteslainnya (*alkaline phospatase, aspartate aminotransferase, alanine aminotransferase*). Biasanya liver di cekdenganmelakukanbeberapates, yang meliputites*bilirubin, kemudiantes alanine transaminase* (ALT), tes*asparate transaminase* (AST), dan juga tes*alkaline phosphatase* (ALP), *albumin* total protein dan lain-lain (Lika Aprilia Samadi, 2018). Informasi yang didapatdariberbagailliteraturmenunjukanadanyahubungan yang terkaitdengansejumlahatribut dataset yang digunakanuntukmemprediksiadanyapenyakit liver pada pasien.

Salah satucarauntukmenyelesaikanmasalah di atasadalahdenganmengolah data sehinggainformasi yang terkandungdidalamnyadapatdiambiluntukmendapatkanpola/pengetahuan yang diperlukan.

Dalamhaliniperan Data Mining digunakanuntukmengatasimasalahtentangpenyakit liver pada pasien, analisa yang akuratakanmemberikansolusiterbaikbagipenelitidalammengambilkeputusan, inisiatif yang baikuntukmemilihsuatumetode yang dapatdigunakanuntukmembantudalammemprediksipenyakit liver denganmenggunakanteknik CRISP-DM. Ada beberapahal yang dilakukandalampenelitianiniyaitumelakukananalisisfaktorgunamendapatkankorelasiantaratributhalinidijadikansebagaiacuandasardalammenemukanpola yang tepat, Melakukanperbandingan lima algoritmaklasifikasiuntukmenghasilkanalgoritmaterbaiksebagaitindaklanjutpenelitian yang lebihrelevan, juga melakukanperbandingandengan*feature extaction* dan *feature selection*. Data yang digunakanuntukpenelitianinibersumberdarialamat web: http://archive.ics.uci.edu/ ml/. Data tersebutmerupakanhasilpemeriksaandari 583 orang yang tersebardiwilayah Andhra Pradesh, India jumlahatributsebanyak 11, Model dibuatdenganmenggunakanRapidminer 8.

METODE PENELITIAN

Metode Correlation Matrix digunakanuntukmengetahuihubunganantarfaktordalammendeteksipasien yang terkenapenyakit liver. Kemudianpenelitiandibandingkandengan 5 algoritmaumum, yaitu Decision Tree (C4.5), Naive Bayes (NB), K-Nearest Neighbor (kNN), Logistic Regression (LogR) dan Deep Learning, yang kemudianakandidapatkanalgoritmaterbaikuntukmenentukanpenyakit liver. Selanjutnyadigunakanperbandingan 4 metode Feature Extraction (PCA) Feature Selection (FS) antara lain Forward Selection, Information Gain, dan Backward Elimination yang kemudiandiperolehmetode Feature Selection yang paling baikdalammeningkatkanperformaalgoritma yang akandigunakantersebut. Penelitiandilakukandenganmetode CRISP-DM yang terdiridari 6 fase.

Eksperimendilakukan pada laptop berbasis Core i3 2.53 GHz CPU, 8 GB RAM dan sistemoperasi Windows 10 Enterprise 32-bit. Aplikasi yang digunakanuntukpenelitianyaitu RapidMiner 8.0.

Metode CRISP-DM meliputi

a. Business Understanding
b. Data Understanding
c. Data Preparation
d. Modeling
e. Evaluasi

HASIL DAN PEMBAHASAN

A. Business Understanding

1. Motivasi:
   Industrikesehatanmemerlukananalisa yang akurat dan efektifuntukmenghasilkanprediksipenyakit pada pasien, agar masalah dan kebutuhanpasienteratasidenganbaik.
   Inginmengetahuibagaimana data yang adadiolahdapatmenghasilkanpengetahuanatau model yang akurat.

2. Objektif:
   dalamhaliniakandilakukanpencarianrelasiantarfaktor yang salingberpengaruh

B. Data Understanding

Dataset yang digunakandalampenelitianinidiambidar UCI Machine Learning Repository denganalamat:https://archive.ics.uci.edu/ml/datasets/ILPD+(Indian+Liver+Patient+Dataset).

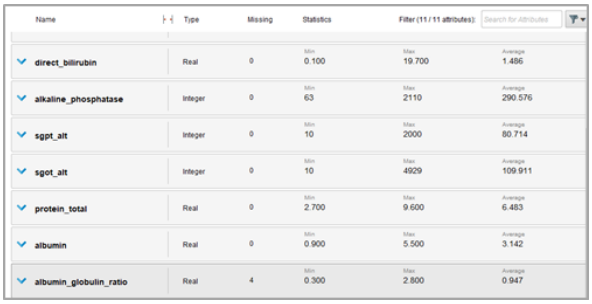
Kumpulan data iniberisi 416 catatanpasien liver dan 167 catatanpasien non liver. Kumpulan data diambildariutaratimur Andhra Paradesh India. selector adalah label kelasyangdigunakanuntukmembagikedalamkelompokpasien liver atautidak. Kumpulan data iniberisi 441 catatanpasienlaki-laki dan 142 catatanpasienperempuan, Setiappasien yang umurnya 89 tahuntercatatberusia 90 tahun Dataset denganatributsebagaiberikut:

1. Usia (umurpenderita)
2. Gender daripasien
3. TB (Total Bilirubin)
4. DB (Bilirubin Langsung)
5. Alkphos (alkaline phospotase)
6. SGPT alamine aminotransferase
7. Sulfataspartat aminotransferase
8. TP (total Protein)
9. ALB (albumin)
10. Rasio Albumin dan Globulin Rasio A/G
11. Pasienterkena liver atautidak (1:ya, 2:no)

C. Data Preparation

Atributakandisesuaikandengankorelasiatributnya, bisaberupanumerikmaupun nominal dan menggunakan label. dalamhaliniketerangannyaadalah 1=(terkena liver) atau 2=(tidakterkena).

Kemudiansetelahdievaluasikualitas data ternyatadatasetnya missing (kosong) pada salah satuatributyaituRasio Albumin dan Globulin Rasio A/G sudahkonsistensehinggadapatdimodelkan. Selainitu juga pada beberapaatributmenunjukannilai yang abnormal (tidakwajar)



Name	Type	Missing	Statistics	Filter (11 / 11 attributes):	
direct_bilirubin	Real	0	Min 0.100	Max 19.700	Average 1.486
alkaline_phosphatase	Integer	0	Min 63	Max 2110	Average 290.576
sgpt_alt	Integer	0	Min 10	Max 2000	Average 80.714
sgot_alt	Integer	0	Min 10	Max 4929	Average 109.911
protein_total	Real	0	Min 2.700	Max 9.600	Average 6.483
albumin	Real	0	Min 0.900	Max 5.500	Average 3.142
albumin_globulin_ratio	Real	4	Min 0.300	Max 2.800	Average 0.947

Gambar 1. Statistik Dataset Pasien Liver masihterdapat missing dan noisy

Kemudianmelakukan Preparation data dengan operator *Replace Missing value* (menggantinilai yang hilangdengannilai rata-rata, dan Operator *Filter Example* (menghapusnilai yang tidakwajar).

D. Modeling

Merupakan proses memilihteknik data mining denganmenentukanalgoritma yang akandigunakan. Tool yang dipakaiyaitu RapidMiner versi 8. Untukmendapatkanhubunganantaratributdigunakan Korelasi Matrix yang mampumendeskripsikanbentuk dan kekuatanhubunganantaratributtersebut.

UntukmetodeKlasifikasi, denganmembandingkan 5 Algoritmayaitu: Decision Tree, Naive Bayes, K-NN, Logistic Regression dan deep learning. untukmengetahuinilai accuracy dan AUC.

Digunakan Uji beda (T-Test) Untukdilakukanperbandingankinerja (performa) darikelimaalgoritmatersebut. Dengan uji beda (T-test) dapatdiketahuiAkurasiuntukKlasifikasi dan perbedaansignifikandarikelimaalgoritmadarimasing-masingmetodeKlasifikasitersebutuntukdianalisasehinggadapatdiketahui yang terbaik.

Untukmemperbaikikinerja (performa) darimasing-masingmetodedapatdigunakanmetodeFeatur Extraction, Feature Selection yang terdiridaribeberapajenis, sedangkan yang digunakandalampenelitianiniadalah :

1. Feature Exctrationmenggunakan Principal Component analysis
2. Feature Selection

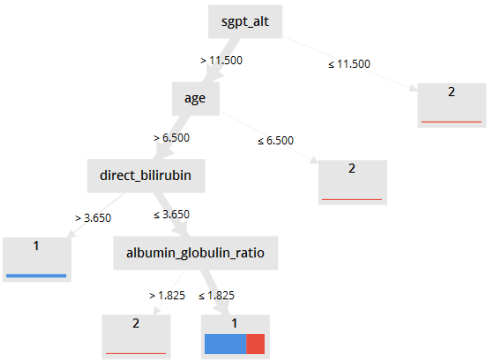
Filter (information gain), Wrapper (Forward Selection dan Backward Elimination).

Feature Selection dapatmemperbaikiperformakarenadapatmeningkatkanAkurasi. Untukmembandingkanmetode Feature Selection tersebut juga digunakan uji beda (t-Test). Hasil dari uji bedatersebutdigunakanuntukmengetahuimetode feature Selection yang terbaik.

E. Evaluasi

Dari hasilpemodelantersebutdapatdiketahuibahwa yang paling berpengaruhdalammenentukanPasien yang terkenapenyakit liver adalahsgpt_alt (tesdarah yang dilakukandokterterhadap alanine aminotransferase untukmemeriksakerusakan liver)

Jika sgpt_altdengankadarkurangdari 11.500 pasientidakterkenapenyakithati, dan jikakadarsgpt_altlebihbesardari 11.500 makaselanjutnyamemeriksaumurpasien, direct_bilirubin dan albumin_globulin_ratio. Pohonkeputusan (decision tree) yang digunakanuntukmengetahuihubunganantaratribut :



Gambar 2. polaklasifikasipasienterkena liver

Pola/pengetahuan yang didapatkanadalahsebagaiberikut

1. sgpt _alt <= 11.500 tidakterkenapenyakit liver
2. sgpt_alt>=11.5, umur<= 6.5 tahuntidakterkena liver
3. sgpt_alt>=11.5, umur>=6.5 tahun dan direct bilirubin > 3.65 terkena liver
4. sgptalamine>=11.5, umur>=6.5 tahun dan direct bilirubi<= 3.65 dan ratio albumin globulin > 1.825 tidakterkena Liver
5. sgpt_alt>=11.5, umur>=6.5 tahun dan direct bilirubin <= 3.65 dan ratio albumin globulin <=1.825 sebagianbesarterkena liver

Grafikdari Hasil Corelation Matrix dapatdiketahui pada masing-masingatribut yang memilikikorelasiterhadappasien yang terkenapenyakit liver ataupuntidak.

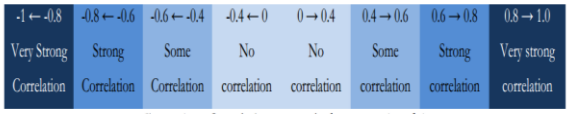
Tabel yang didapatkandarihasilpemodelandapat juga diketahuihubunganantarfatordintaranyaadalahhubungan positif (berbandinglurus) Sepertihubunganantaraliver_patient dan albumin. Jika nilai correlation negatifmakahubunganyanegatif (berbandingterbalik)

Lalu untukhubungannegatif (berbandingterbalik) yaituantaraLiver_patient dan direct-bilirubin dimanasemakinrendahtingkatdirect_bilirubinmakakemungkinantidakterkenapenyakit liver.

Attribut...	liver_pa...	age	gender	tot_bilir...	direct_...	alkaline...	sgpt_alt	spot_alt	protein...	albumin	albumin...
liver_pati...	1	-0.136	-0.089	-0.231	-0.240	-0.212	-0.202	-0.219	0.031	0.158	0.160
age	-0.136	1	0.058	0.014	0.018	0.031	-0.076	-0.029	-0.189	-0.263	-0.209
gender	-0.089	0.058	1	0.116	0.126	0.045	0.100	0.115	-0.083	-0.098	-0.014
tot_biliru...	-0.231	0.014	0.116	1	0.979	0.227	0.200	0.308	0.004	-0.230	-0.200
direct_bil...	-0.240	0.018	0.126	0.979	1	0.235	0.200	0.294	0.006	-0.224	-0.189
alkaline_...	-0.212	0.031	0.045	0.227	0.235	1	0.160	0.128	0.005	-0.141	-0.224
sgpt_alt	-0.202	-0.076	0.100	0.200	0.200	0.160	1	0.802	-0.008	-0.036	-0.064
spot_alt	-0.219	-0.029	0.115	0.308	0.294	0.128	0.802	1	-0.025	-0.128	-0.130
protein_t...	0.031	-0.189	-0.083	0.004	0.006	0.005	-0.008	-0.025	1	0.785	0.233
albumin	0.158	-0.263	-0.098	-0.230	-0.224	-0.141	-0.036	-0.128	0.785	1	0.682
albumin...	0.160	-0.209	-0.014	-0.200	-0.189	-0.224	-0.064	-0.130	0.233	0.682	1

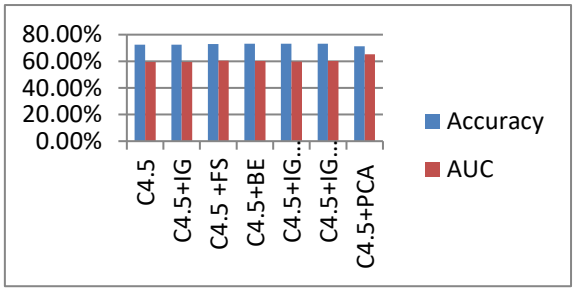
Gambar 3. Tabelhasil Correlation matrix hubunganantarfaktoratribut

Selainitudapatdiketahui juga kekuatanhubungannyasemakinbesarnilaicorrelationnyamakasemakinkuat/banyakhubungangannyabegitu juga sebaliknyasemakinkecilnilaicorrelationnyamakasemakinlemah/sedikithubungannyacontohnyaadalahhubunganantarasgot_altdengansgpt_altkedua serum inimemilikikorelasi yang kuatsehingganilaicorrelationnyabesar. Pada kenyataanyajikadoktermemeriksakadarsgottinggimakauntukmenentukanpenyakit liver makasgpt juga harusdiketahuikadarnya. Sedangkanuntuk yang hubungannyalemah/sedikitcontohnyaadalahhubungannantaraPasien yang terkena liver dengantidakselaludisebabkan oleh salah satufaktorpenyebabnya. dengantekananudarakarenamemilikinilai correlation yang kecil. Sedangkan yang nilaicorrelationnyasangatkecilataubisadikatakantidak berhubungankarenanilaicorrelationnyakurangdari 0.4 seperti yang ditunjukkan pada gambarhasil correlation matrix berikutini :



Gambar 4. Kekuatanhubunganantarfaktor

Dari kelimaalgoritma yang dipilihmenunjukanhasildari uji beda (t-Test) diketahuibahwa pada Klasifikasi model yang terbaikadalah decision Tree(C4.5) karenamemilikitingkatAkurasi yang lebihtinggi dan memilikiperbandingan yang signifikanterhadapalgoritma yang lain.



Gambar 5. Accuracy dan AUC pada 5 Algoritma

Dari hasiltersebutdapatdiketahuibahwaurutan model Algoritmaterbaik : 1. C4.5 2. Log R 3. kNN 4. DP 5. NB.

Pohon Keputusan (*Decision Tree*) merupakanmetodeklasifikasi dan prediksi yang sangatkuat dan terkenal. Metodepohonkeputusanmengubahfakta yang sangatbesarmenjadipohonkeputusan yang merepresentasikanaturan. Aturandapatdenganmudahdipahamidenganbahasaalami. Aturanini juga dapatdiekspresikandalambentukbahasa basis data seperti SQL untukmencari *record* pada kategoritertentu. Pohonkeputusan juga bergunauntukmengeksplorasi data, menemukanhubungantersembunyiantarasejumlahcalonvariabel input dengansebuahvariabel target. Karena pohonkeputusanmemadukanantaraeksplorasi data dan pemodelan, pohonkeputusaninisangatbagussebagailangkahawaldalam proses pemodelanbahkanketikadijadikansebagai model akhirdaribeberapateknik lain(J R Quinlan, 1993)
Theoremauntukmenghitungnilai pada Decision Tree adalahsebagaiberikut :
MenghitungnilaiEntrophy :
Netrophy(S) = $\sum_{j=1}^{k} -\text{pj}\log 2 \;\; pj$

S : himpunankasus
k : jumlahpartisi S
pj : probabilitas yang didapatdarijumlah (ya/tidak) dibagi total kasus.

Untukmenghitung gain digunakanrumusberikut :
Gain (S,A) = Entrophy (s) – $\sum_{i=1}^{n} |\frac{si}{s}| * entrhopy(si)$

S : himpunankasus
A : atribut
n : jumlahpartisiatribut A
|si| : jumlahkasus pada partisikei
|S| : jumlahkasusdalam S

Untukmengetahuiakurasidari Decision Tree inidapatdilihatdengan Confusion Matrix berikutini :

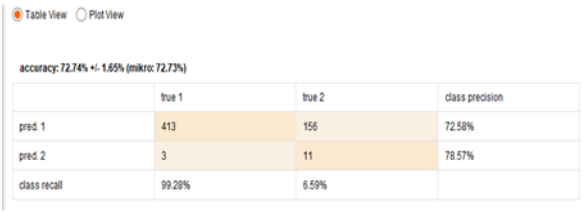
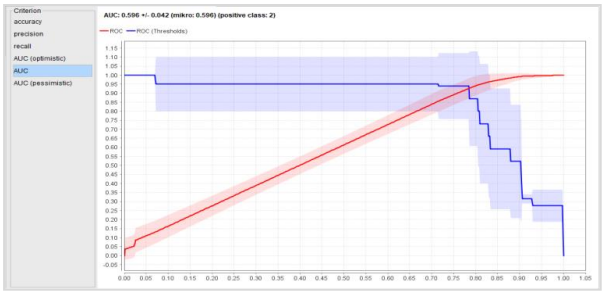


Table View Plot View

accuracy: 72.74% +/- 1.65% (mikro: 72.73%)

	true 1	true 2	class precision
pred. 1	413	156	72.58%
pred. 2	3	11	78.57%
class recall	99.28%	6.59%	

Gambar 6. Confusion Matrix Decision tree

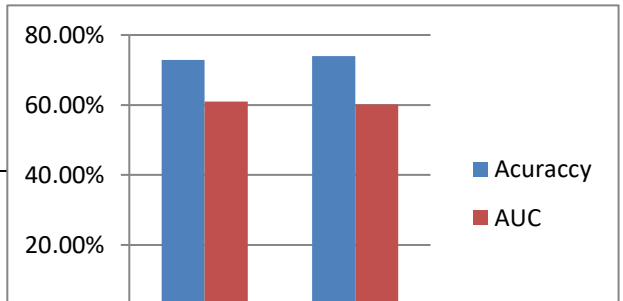


Gambar 7. Kurva ROC-AUC untuk Decision tree

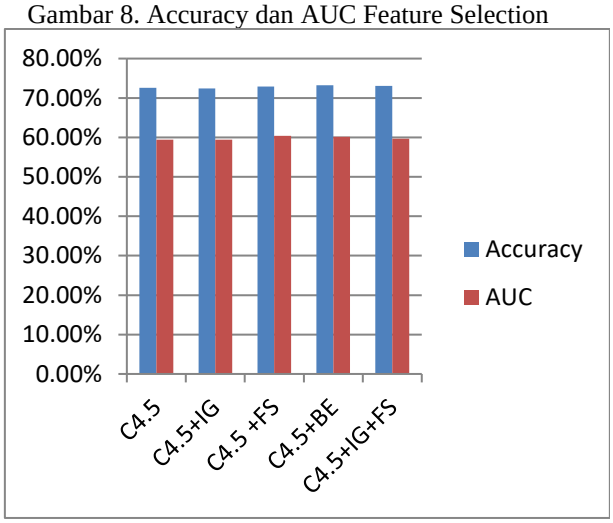
KategoriKlasifikasi AUC:
1. 0.90 - 1.00 = excellent classification
2. 0.80 - 0.90 = good classification
3. 0.70 - 0.80 = fair classification
4. 0.60 - 0.70 = poor classification
5. 0.50 - 0.60 = failure

Dari kurva ROC-AUC model Decision tree memiliki AUC sebesar 0.596 iniberartitermasukdalamkategoriklasifikasirendah .

Hasil metode Feature Extraction, Feature selection pada klasifikasi yang terbaikadalah Backward Elimination (BE) sehinggga model terbaik yang digunakanadalah C4.5+BE. Dengan Feature Selection iniperformanyalebihbaikdarisebelumnyakarenaakurasimeningkatdari 72.74% menjadi 73.24% dan AUC meningkatdari 0.596 menjadi 0.602.



Gambar 8. Accuracy dan AUC Feature Selection



Gambar 9. Grafik Accuracy dan AUC pada C4.5 sebelum dan sesudahmenggunakan BE

A	B	C	D
	0.720 +/- 0.026	0.718 +/- 0.030	0.724 +/- 0.020
0.720 +/- 0.026		0.889	0.743
0.718 +/- 0.030			0.646
0.724 +/- 0.020			

Gambar 10. Hasil Uji beda (T-Test) Feature Selection untuk C4.5

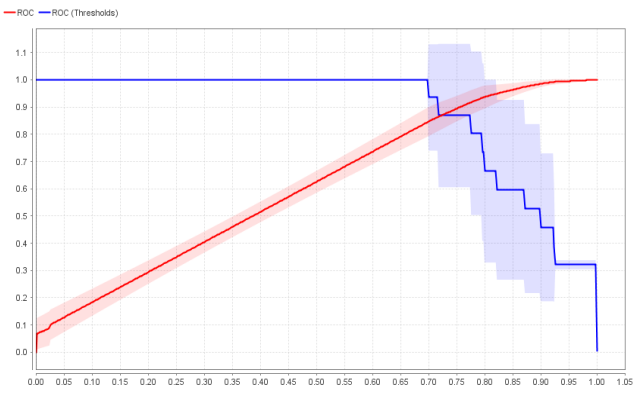
Dari hasiltersebutdapatdiketahuibahwaurutan model terbaiknyaadalah: 1. C4.5+BE 2. C4.5+FS 3. C4.5+FILTER

Table View Plot View

accuracy: 72.37% +/- 1.98% (mikro: 72.37%)

	true 1	true 2	class precision
pred. 1	393	153	71.98%
pred. 2	2	13	86.67%
class recall	99.49%	7.83%	

Gambar 11. Confusion Matrix C4.5+BE



Gambar 12. Kurva ROC-AUC untuk C4.5+BE

Dari kurva ROC-AUC model NB+BE memiliki AUC sebesar 0.606 iniberartitermasukdalamkategori Poor Classification.

F. Deployment

Pola sebagaipengetahuan yang dihasilkanuntukdijadikaninformasibarudalam proses data mining. Pola pengetahuantersebutdidapatdarimetodeKorelasi, Klasifikasi, dan Feature Selection untukmenentukanPenyakit Liver berdasarkan dataset ILPD(India Liver Patient Dataset). Atribut yang ada pada dataset tersebutkorelasinyasangatkecil. Dalammenentukanklasifikasipenyakit liver yang paling menentukanadalahsgpt_alt (serum yang terkandungdalamdarah). KeakuratanKlasifikasidapatditingkatkandenganmenggunakan Backward Elimination sehinggadapatmenghasilkankeputusanklasifikasi yang lebihakuratPengetahuan yang diperolehdapatdigunakansebagaidasardalammengambilkeputusanuntukmenentukanpasien yang terkenapenyakit liver.

KESIMPULAN

Penelitiandapatmenggunakanbeberapaperan data mining yaitu : Korelasi, Klasifikasi, dan Feature Extraction, Feature Selection (Filter dan Wapper)

1. HubunganAntarFaktorPasien yang terkenaPenyakit liver. Faktor yang paling mempengaruhipasien yang terkenapenyakit liver adalahsgpt_altdengankadar>= 11.500. Jika kadarsgpt_alt<= 11.500 tidakterkenapenyakit liver. Apabilasgpt_alt>= 11.500, umur<=6.5 tahuntidakterkena liver. sgpt _alt>=11.5, umur>=6.5 tahun dan kadar direct bilirubin > 3.65 terkena liver. Sgpt_alt>=11.5, umur>=6.5 tahun dan direct bilirubin <= 3.65 dan ratio albumin globulin > 1.825 tidakterkena liver. sgpt_alt>=11.5, umur>=6.5 tahun dan direct bilirubin <= 3.65 dan ratio albumin globulin <=1.825 sebagianbesarterkena liver
2. Perbandingan 5 AlgoritmaKlasifikasiPasien yang terkenapenyakit liver. Algoritmaterbaiknyaadalah C4.5 denganAkurasi 72.56 %, AUC 0.594, dan tidakmemilikiperbedaansignifikanterhadapalogitma yang lain. Sehinggaalgoritma C4.5 dapatdigunakanuntukklasifikasidalammenentukanpasien yang terkenapenyakit liver.
3. Peningkatanperformadengan Feature Selection Feature Selection dapatmengurangifaktor/atribut

yang tidakterlaluberpengaruhsehinggadapatmeningkatkanAkurasi dan AUC. Sedangkanmetode Feature Selection yang terbaikuntuk C4.5 pada penelitianiniadalah Backward Elimination (BE) modelnyamenjadi C4.5+BE. Metodedengan model C4.5+BE memilikitingkatakurasi yang baiksehinggahasilnyacukupakuratdengandemikian metodeinidapatdigunakansebagairekomendasidallammembantumengambilkeputusan yang tepatuntukmenentukanpasein yang terkena liver.

Untukpenelitianmendatangdapatmenggunakanalgoritma yang lain agar akurasinyalebihbaiklagiatau dataset yang adaatributnyabisaditambahkan agar kinerjadarialgoritma yang adamenunjukanperforma yang lebihbaiklagi.

SelanjutnyauntuklebihmeningkatkanperformadariAlgoritmatersebutdapatdilakukandenganmenggunakan metode Feature Selection atau Feature Extraction lain yang dapatmeningkatkanperformaalgoritmamenjadilebihcepat dan lebihakurat.

REFERENSI

Benjamin, A., Zubajlo, R., Thomenius, K., Dhyani, M., Kaliannan, K., Samir, A. E., & Anthony, B. W. (2017). Non-invasive diagnosis of non-alcoholic fatty liver disease (NAFLD) using ultrasound image echogenicity. *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, 2920–2923. https://doi.org/10.1109/EMBC.2017.8037468

Harrison, E., Chang, M., Hao, Y., & Flower, A. (2018). Using machine learning to predict near-term mortality in cirrhosis patients hospitalized at the University of Virginia health system. *2018 Systems and Information Engineering Design Symposium, SIEDS 2018*, 112–117. https://doi.org/10.1109/SIEDS.2018.8374719

Hashem, S., Esmat, G., Elakel, W., Habashy, S., Raouf, S. A., ElHefnawi, M., Eladawy, M., & ElHefnawi, M. (2018). Comparison of Machine Learning Approaches for Prediction of Advanced Liver Fibrosis in Chronic Hepatitis C Patients. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *15*(3), 861–868. https://doi.org/10.1109/TCBB.2017.2690848

Hassoon, M., Kouhi, M. S., Zomorodi-Moghadam, M., & Abdar, M. (2017). Rule Optimization of Boosted C5.0 Classification Using Genetic Algorithm for Liver disease Prediction. *2017 International Conference on Computer and Applications, ICCA 2017*, 299–305. https://doi.org/10.1109/COMAPP.2017.8079783

Kumar, S., & Katyal, S. (2018). Effective Analysis and Diagnosis of Liver Disorder by Data Mining. *Proceedings of the International Conference on Inventive Research in Computing Applications, ICIRCA 2018*, 1047–1051. https://doi.org/10.1109/ICIRCA.2018.8596817

Lara, J., Khudyakov, Y., Rossi, L., & Vaughan, G. (2015). Highlights: Predicting the cross-immunoreactivity of hepatitis C virus hyper-variable region 1 peptides using polynomial neural networks. *IEEE*, 1–1. https://doi.org/10.1109/iccabs.2015.7344732

Yousefi, L., Saachi, L., Bellazzi, R., Chiovato, L., & Tucker, A. (2017). Predicting Comorbidities Using Resampling and Dynamic Bayesian Networks with Latent Variables. *Proceedings - IEEE Symposium on Computer-Based Medical Systems*, *2017-June*, 205–206. https://doi.org/10.1109/CBMS.2017.32