THE POWER OF DATA SCIENCE IN ACCURATE MOST COMMON DISEASE DIAGNOSIS IN HEALTHCARE

Desi Noviyanti
Program Sarjana Matematika, Institut Pertanian Bogor, Bogor
Email: noviyantidesi01@gmail.com

1. PENDAHULUAN

Perkembangan kemajuan teknologi informasi dan komunikasi semakin pesat membutuhkan informasi yang cepat dan akurat bagi para pengambil keputusan, termasuk di dalamya manajemen rumah sakit. Rumah sakit yang menyimpan beragam jenis data membutuhkan pengolahan data yang tepat dan akurat sehingga dapat disajikan dalam bentuk laporan yang mudah dipahami (Handiwidjojo, 2009). Salah satu data yang sering dipakai rumah sakit adalah rekam medis pasien, yang merupakan catatan, berkas pasien, serta dokumen mengenai identitas, laporan pemeriksaan, informasi pengobatan, dan tindakan medis pasien pada sarana pelayanan utuk rawat jalan dan rawap inap baik yang dikelola pemerintah maupun dikelola swasta. Namun, informasi yang terdapat dalam rekam medis harus dijaga kerahasiannya dan tidak diperkenankan untuk disebarluaskan kepada pihak-pihak yang tidak berwenang (Amin, Cholil, Herdiansyah, & Negara, 2021).

Untuk mengatasi masalah tersebut, penggunaan *Electronic Healthy Record* (EHR) sangat direkomendasikan. Pada dasarnya, *Electronic Healthy Record* merupakan penggunaan perangkat teknologi informasi untuk pengumpulan, penyimpanan, pengolahan, serta peng-akses-an data yang tersimpan pada rekam medis mencakup berbagai informasi klinis pasien seperti riwayat medis, tanda-tanda vital, hasil tes laboratorium, dan catatan klinis. Tantangan pengaplikasian EHR dalam dunia kesehatan, antara lain banyak pihak yang mencurigai bahwa rekam medis elektronik tidak memiliki landasan hukum yang kuat terhadap unsur privasi maupun keamanan informasi, serta tantangan finansial karena rumah sakit perlu menyiapkan infrastuktur teknologi informasi. Meskipun demikian, EHR mempunyai dampak yang baik, yaitu dapat meningkatkan profesionalisme dan kinerja manajemen rumah sakit, karena para *stakeholder* akan menikmati kemudahan, kecepatan, dan kenyamanan pelayanan kesehatan (Handiwidjojo, 2009).

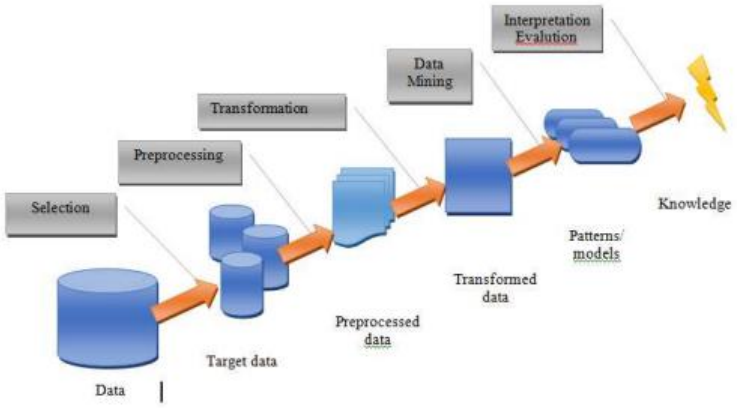
Rekam medis mengandung informasi mengenai diagnosis akhir yang kemudian digunakan untuk proses pengkodean. Standar klafisikasi penyakit ICD-10 (*International Statistical Classification of Diseases and Related Health Problems Tenth Revision*) digunakan untuk melakukan pengkodean tersebut. Penggunaan ICD-10 bertujuan untuk menyelaraskan nama dan klasifikasi penyakit serta faktor-faktor yang memengaruhi kesehatan (Mardi, 2018). ICD merupakan sebuah klasifikasi standar Internasional untuk mendiagnosa penyakit dalam semua tujuan epidemiologis manajemen kesehatan dan umum. Dahulu, data medis tersedia dalam jumlah yang besar dan sering dianggap sebagai sampah dan diabaikan. Namun, dengan perkembangan ilmu pengetahuan dan teknologi, konsep *data mining* telah berkembang. Saat ini, data tersebut dapat memberikan manfaat yang begitu banyak, baik untuk ilmu pengetahuan, bisnis maupun pengambilan keputusan (Amin, Cholil, Herdiansyah, & Negara, 2021).

2. LANDASAN TEORI

Istilah *Data Mining* dan *Knowledge Discovery in Database* (KDD) adalah dua istilah yang sering digunakan untuk menggambarkan proses menemukan informasi tersembunyi dalam basis data yang besar. Meskipun kedua istilah ini berbeda dalam konsep, tetapi berkaitan satu sama lain. *Data mining* adalah langkah penting dalam proses KDD secara keseluruhan (Mardi, 2018). *Data mining* sudah digunakan dalam beberapa penelitian, seperti prediksi cuaca dan prediksi kriteria nasabah.

Menurut Taranu (2015), *Knowledge Dicovery in Database* (KDD) merupakan proses pemindaian otomatis dari data yang berukuran besar sehingga mendapatkan pola yang beguna yang dapat dianggap sebagai pengetahuan tentang data. Proses KDD secara garis besar, sebagai berikut:

a. Data Selection.
b. *Pre-processing* atau *Cleaning*.
c. *Transformation.*
d. *Data Mining*.
e. *Interpretation* atau *Evaluation*.

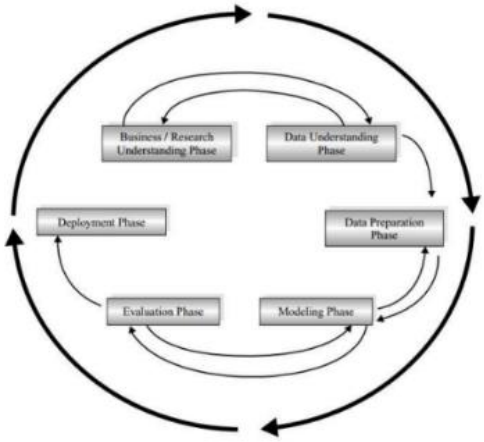


Gambar 1. *Knowledge Dicovery Process* (Taranu, 2015)

Menurut dicoding, *Data Mining* merupakan proses pengumpulan dan pengolahan data yang bertujuan untuk mengekstrak informasi penting pada data. Proses tersebut dilakukan menggunakan perangkat lunak dengan bantuan perhitungan statistika, matematika, ataupun teknologi *Artifical Intelligence* (AI). Menurut Purwanto, Primajaya, & Voutama (2020), pada tahun 1996, Komisi eropa membentuk sebuah konsorsium perusahaan yang dinekal dengan nama CRISP DM (*The Cross-Industry Standard Process for Data Mining*) yang telah diakui sebagai proses standar dalam *data mining* yang dapat diterapkan di berbagai sektor industry. CRISP-DM terbagi menjadi enam tahapan, antara lain:

a. *Business Understanding Phase*
b. *Data Understanding Phase*
c. *Data Preparation Phase*
d. *Modeling Phase*
e. *Evaluation Phase*

f. *Deployment Phase*



Gambar 2. Proses *Data Mining* menurut CRISP-DM ((Mardi, 2018)

Data mining dapat dibagi menjadi deskriptif dan prediktif. Deskriptif bertujuan untuk interpretasi data dan memastikan suatu hipotesis, sedangkan prediktif bertujuan untuk memprediksi hasil yang diinginkan. Metode yang digunakan untuk melakukan data mining prediktif dan deskriptif, antara lain *classification* dan *regression*, *association rule*, *clustering, text mining,* serta *link analysis* (Taranu, 2015).

Dalam kasus mengetahui penyakit terbanyak berdasarkan rekam medis pasien, digunakan *Decision Tree*, dimana metode ini menyaring sesuatu melalui pohon keputusan, apakah suatu data lolos atau tidak terhadap saringan yang telah dibuat dengan proses yang cukup cepat. Metode *decision tree* merupakan salah satu metode Teknik klasifikasi dalam data mining. Pohon keputusan berguna untuk mengekplorasi data, menemukan hubungan yang tersirat antara sejumlah calon variable input dengan sebuah variable output atau target. Dalam pembentukan pohon keputusan digunakan algoritma C4.5, karena memiliki kelebihan yaitu mudah dimengerti, fleksibel, dan menarik karena dapat divisualisasikan dalam bentuk gambar pohon keputusan.

3. METODOLOGI

a. Nilai Entrophy

$$Entrophy\ (S) = \sum_{i=1}^{n} - pi.log2pi$$

Keterangan:
S : himpunan kasus
n : jumlah partisi *S*
pi : proporsi S_i terhadap *S*

b. Nilai Gain

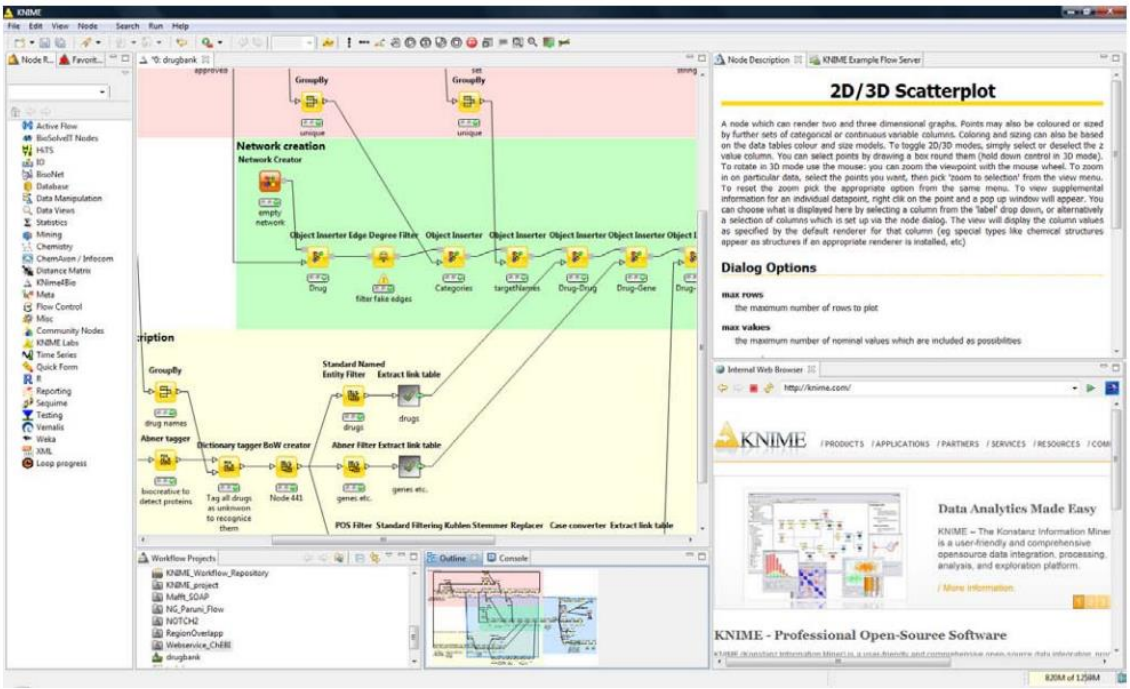
$$Gain\ (S, A) = Entrophy\ (S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} * Entrophy\ (S_i)$$

Keterangan:

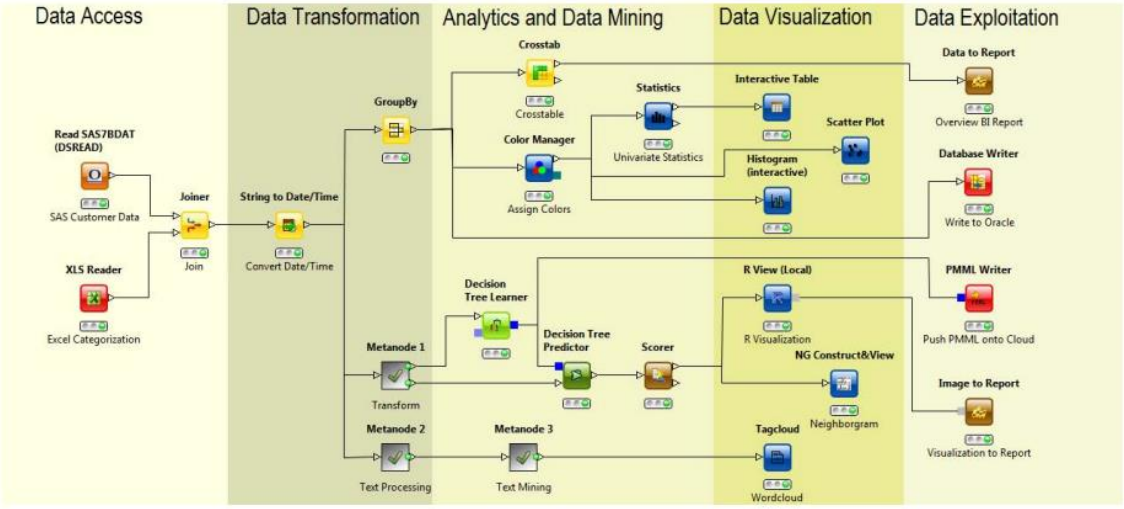
S : himpunan kasus
A : fitur
n : jumlah partisi S
pi : proporsi S_i terhadap S
$|S_i|$: proporsi S_i terhadap S
$|S|$: jumlah kasus dalam S

c. *Software* KNIME

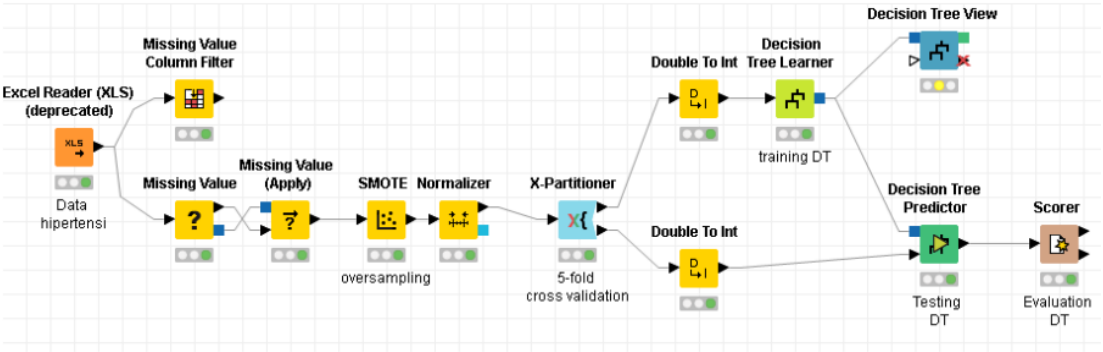
KNIME adalah platform modular *data mining* open sourrce yang bertujuan untuk memecahkan masalah tersebut dengan menyediakan platform yang dapat diperluas dengan mudah dengan integrasi alat baru, memiliki sistem data yang kuat, dan menciptakan alur kerja untuk mendokumentasikan langkah-langkah yang dilakukan oleh alur kerja secara detail (Fillbrunn, et al., 2017). *Software* KNIME ditulis menggunakan Bahasa Java dan berbasis pada lingkungan Eclipse sehingga dapat diperluas melalui plug-in yang memberikan lebih banyak fungsi. Oleh karena itu, KNIME memungkinkan digunakan di banyak bidang yang berbeda (Mazanetz, Marmon, Reisser, & Morao, 2012).



Gambar 3. Workbench KNIME. Editor alur kerja (panel tengah), repositori node (panel kiri), deskripsi node (panel kanan atas(, perambanan web (panel kanan bawah), dan daftar alur kerja dan gambaran alur kerja saat ini (panel bawah). (Mazanetz, Marmon, Reisser, & Morao, 2012)



Gambar 4. Alur kerja KNIME yang menggambarkan berbagai langkah dalam proses data mining (Mazanetz, Marmon, Reisser, & Morao, 2012)



Gambar 5. Alur kerja KNIME untuk *decision tree* (Santoni, Chamidah, & Matondang, 2020)

4. HASIL DAN PEMBAHASAN

a. *Business Understanding*

Business Understanding merupakan tahapan pemahaman bisnis yang berfokus pada latar belakang masalah dan tujuan penelitian. Pemahaman bisnis tersebut, selanjutnya diubah menjadi sebuah rencana awal data mining untuk menjadi tujuan yang didasarkan pada fungsi analisis algoritma C4.5. Tujuannya yaitu supaya dapat mengetahui data penyakit terbanyak berdasarkan data rekam medis sehingga dapat dimanfaatkan sebaik mungkin. Data yang digunakan yaitu data rekam medis rumah sakit di salah satu wilayah kota di Indonesia, yang diperoleh dari Dinas Kesehatan kota tersebut. Tujuan data mining dalam penelitian ini adalah mengetahui model terbaik untuk memprediksi penyakit terbanyak yang dialami oleh masyarakat di Kota tersebut.

b. *Data Understanding*

Data Understanding merupakan tahapan pemahaman data yang berkaitan dengan rekapitulasi rekam medis, data tersebut diperoleh dari Dinas Kesehatan di suatu kota di Indonesia, khususnya penyakit yang pernah terjadi di wilayah tersebut. Data yang digunakan yaitu data yang dikumpulkan dari Dinas Kesehatan di suatu kota di Indonesia terkait dengan riwayat penyakit yang dialami oleh masyarakat di daerah tersebut. Parameter data yang digunakan antara lain, jenis kelamin, usia pasien, pengelompokkan wilayah, serta pengelompokkan kode ICD-10.

c. *Data Preparation*

Data preparation mencakup semua kegiatan untuk membangun dataset yang akan dimasukkan ke dalam tools pemodelan dari data mentah. Pemodelan dataset tersebut diolah menggunakan algoritma C4.5. Parameter Dataset yang diambil dalam kasus ini yaitu, jenis kelamin, usia pasien, pengelompokkan wilayah, dan pengelompokkan kode ICD-10. Dari data yang didapatkan, kemudian varibel-variabel tertentu dikelompokkan sebagai berikut:

i. Usia

Menurut WHO, pengelompokkan usia kesehatan mendasar dikelompokkan menjadi tiga kelompok jenjang usia. Usia <15 tahun klasifikasinya bayi dan anak-anak, usia >15 tahun dan <50 tahun klasifikasinya muda dan dewasa, dan usia >50 tahun klasifikasinya adalah tua.

ii. Pengelompokkan Wilayah

Dalam rekam medis keterangan alamat ditulis secara lengkap. Namun, pada kasus ini, Parameter yang diambil hanya kecamatan saja, karena alamat kecamatan dapat mewakili suatu Kota tersebut secara keseluruhan, sedangkan Parameter lainnya seperti jalan, kabupaten/kota, kode pos, dan seterusnya tidak diperlukan.

iii. Pengelompokkan Kode ICD-10

Pengkodean diagnose berdasarkan ICD-10 dapat dikelompokkan berdasarkan bab sesuai dengan ketentu ICD-10, sebagai berikut:

BAB	KODE	PENYAKIT
I	A00 – B99	Infeksi dan Parasit
II	C00 – C99	Neoplasma Ganas
III	D00 – D48	Neoplasma In Situ dan Jinak
IV	D50 – D89	Penyakit Darah dan Alat Pembuat Darah, Mekanisme Imun
V	E00 – E90	Penyakit Endokrin, Nutrisi dan Metabolik
VI	F00 – F99	Gangguan Jiwa dan Perilaku
VII	H00 – H59	Penyakit Susunan Syarat
VIII	H60 – H95	Penyakit Mata dan Adnexa
IX	I00 – I99	Penyakit Telinga dan Proses Mastoid
X	J00 – J99	Penyakit Pembuluh Darah
XI	K00 – K93	Penyakit Saluran Nafas
XII	L00 – L99	Penyakit Saluran Cerna
XIII	M00 – M99	Penyakit Kulit dan Jaringan Bawah Kulit
XIV	N00 – N99	Penyakit Otot dan Jaringan Ikat
XV	O00 – O99	Penyakit Sistem Kemih Kelamin
XVI	P00 – P96	Kehamilan, Persalinan, dan Nifas

XVII	Q00 – Q99	Kondisi tertentu berawal dari Masa Perinatal
XVIII	R00 – R99	Malformasi Bawaan, Deformasi, dan Abnormalitas Kromosom
XIX	S00 – T98	Cedera, Keracunan, dan Faktor Eksternal
XX	V01 – Y98	Penyakit dan Kematian akibat Faktor Eksternal
XXI	Z00 – Z99	Faktor yang berpengaruh pada Status Kessehatan dan Kontak dengan Fasilitas Pelayanan Kesehatan
XXII	U00 – U99	Kode untuk Penggunaan Khusus

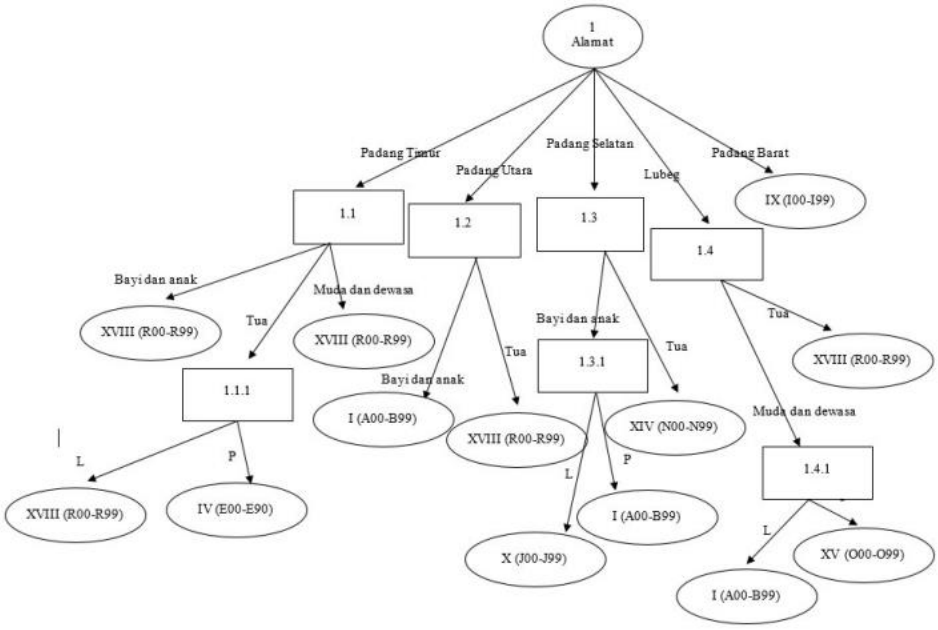
d. *Modelling*

Pemodelan adalah proses dimana dalam penelitiannya melibatkan *data mining*. Teknik pemodelan untuk kasus ini yaitu klasifikasi prediksi dengan menggunakan algoritma C4.5 *decision tree*. Klasifikasi algoritma C4.5 digunakan untuk memprediksi penyakit terbanyak yang dialami oleh masyarakat di suatu Kota di Indonesia. Dalam pembuatan *decision tree*, perlu diketahui terlebih dahulu nilai *entrophy* dan *gain*.

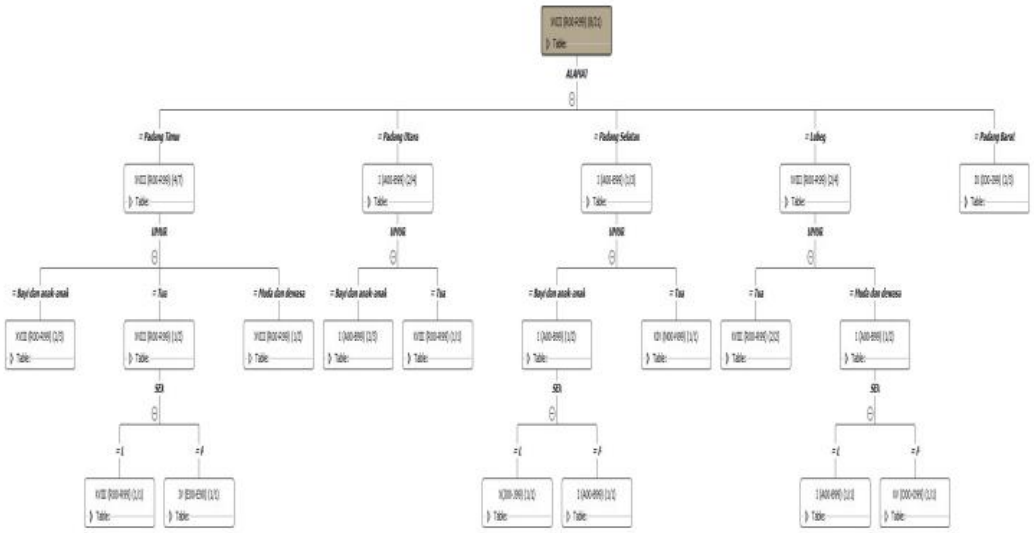
Langkah-langkah dalam pembuatan *decision tree* yaitu, pertama menyiapka *data training* atau *dataset*, ke-dua menentukan akar dari pohon. Akar yang diambil yaitu parameter usia, pengelompokkan wilayah, dan pengelompokkan kode ICD-10, lalu dihitung nilai *entrophy*-nya, setelah dihitung nilai *entrohpy*-nya lakukanlah perhitungan nilai *gain*. Nilai *gain* yang paling tinggi menjadi akar akar pertama dari *decision tree*. Ulangi langkah ke-dua hingga semua *record* terpartisi dan proses partisi *decision tree* akan berhenti pada saat semua *record* dalam simpul N mendapat kelas yang sama, tidak ada atribut di dalam *record* yang dipartisi lagi, dan tidak ada *record* di dalam cabang yang kosong (Mardi, 2018).

e. *Evaluation*

Proses evaluasi dilakukan dengan bantuan *software* KNIME (*Konstanz Information Miner*) untuk mengetahui keakuratan dari data yang telah diolah. Proses perhitungan dan pembuatan *decision tree* yang dilakukan secara manual, selanjutnya diuji dengan menggunakan salah satu *software data mining*, yaitu KNIME (Konstanz Information Miner). Dari proses tersebut, akan dihasilkan *decision tree* dan aturan yang sesuai. Berikut contoh *decision tree* dengan proses manual dan bantuan *software* KNIME:



Gambar 6. Hasil Node Decision Tree secara Manual (Mardi, 2018)



Gambar 7. Hasil *Node Decision Tree* menggunakan KNIME (Mardi, 2018)

f. *Deployment*

Dari permasalahan yang sudah dipaparkan dari tahap *business understanding* yaitu memprediksi penyakit yang banyak dialami, dari kasus ini dapat diketahui bahwa algoritma C4.5 merupakan algoritma yang terbaik. Berdasarkan kasus tersebut, dapat disimpulkan bahwa tujuan dari kasus ini tercapai dengan baik.

5. PENUTUP

a. Simpulan

Berdasarkan proses data mining merurut CRISP-DM, untuk mengetahui akurasi dalam diagnosis penyakit yang paling banyak dialami oleh masyarakat di suatu kota di Indonesia yaitu dengan menggunakan decision tree dalam pengklasifikasiannya. Pengambilan data dari kasus ini dapat diperoleh dari Dinas Kesehatan di suatu kota di Indonesia, Parameter yang digunakan antara lain, jenis kelamin, usia, pengelompokkan wilayah, dan pengelompokkan kode ICD-10. Lalu data tersebut dihitung mengenai nilai *entrophy* dan nilai *gain*. Setelah itu, Pembuatan *decision tree* menggunakan algoritma C4.5 karena mudah digunakan, dianalisa, dan dideskripsikan. Dalam mengetahui akurasi suatu penyakit yang paling banyak di alami yaitu dengan membandingkan *decision tree* manual dengan *decision tree* menggunakan *software* KNIME.

b. Saran

Perlunya menggunakan metodologi lain seperti naïve bayes untuk pengklasifikasian lebih lanjut, karena naïve bayes merupakan salah satu metode yang cocok untuk klasifikasi biner dan multiclass. Dengan menggunakan variatif metode lain dapat meningkatkan nilai akurasi algoritma. Selain itu, perlunya penjelasan lebih lanjut mengenai *deployment*, karena perlunya penambahan *deployment* tersebut maka perlu juga ditambahkan referensi jurnal mengenai hal tersebut.

6. REFERENSI

Amin, Z. A., Cholil, W., Herdiansyah, M. I., & Negara, E. S. (2021). Analisa Rekam Medis Elektronik untuk menentukan Diagnosa Medis dalam kategori BAB ICD 10 menggunakan Machine Learning. *Sistem dan Teknologi Informasi*.

Fillbrunn, A., Dietz, C., Pfeuffer, J., Rhan, R., Landrum, G. A., & Berthold, M. R. (2017). KNIME for Redroducible Cross-Domain Analysis of Life Science Data. *Biotechnology*, 149-156.

Handiwidjojo, W. (2009). Rekam Medis Elektronik. *EKSIS, II*, 36-41.

Mardi, Y. (2018). Data Mining Rekam Medis untuk menentukan penyakit terbanyak menggunakan Decision Tree C4.5. *Sains dan Informatika*, 40-53. doi:10.22216/jsi.v4i1.3077

Mazanetz, M. P., Marmon, R. J., Reisser, C. B., & Morao, I. (2012). Drug Discovery Applications for KNIME: An Open Source Data Mining Platform. *Medicinal Chemistry*, 1965-1979.

Purwanto, A., Primajaya, A., & Voutama, A. (2020). Penerapan Algoritma C4.5 dalam Prediksi Potensi Tingkat Kasus Pneumonia di Kabupaten Karawang. *Sistem dan Teknologi Informasi*, 390-396.

Santoni, M. M., Chamidah, N., & Matondang, N. (2020). Prediksi Hipertensi menggunakan Decision Tree, Naive Bayes dan Artificial Neural Network pada Software KNIME. *Techno.COM*, 353-363.

Taranu, I. (2015). Data Mining in Healthcare: Decision Making and Precision. *Database System Journal*, 33-40.