PENERAPAN *DATA MINING* UNTUK KLASIFIKASI PENYAKIT *HEPATOCELLULAR CARCINOMA* MENGGUNAKAN ALGORITMA *NAÏVE BAYES*

Bambang Tri Rahmat Doni[1], Sari Susanti[2], Ade Mubarok[3]

[1]Universitas Adhirajasa Reswara Sanjaya
Email: geniuses17@gmail.com

[2]Universitas Adhirajasa Reswara Sanjaya
Email: sarisusanti@ars.ac.id

[3]Universitas Adhirajasa Reswara Sanjaya
Email: adem@ars.ac.id

Abstrak

Hepatocellular Carcinoma merupakan tumor ganas hati primer yang berasal dari hepatosit. Dalam dasawarsa terakhir terjadi perkembangan yang cukup berarti menyangkut penyakit Hepatocellular Carcinoma. Penelitian ini bertujuan untuk mengklasifikasikan tingkat kemungkinan hidup pasien yang telah di diagnosis menderita penyakit Hepatocellular Carcinoma dengan menggunakan penerapan metode data mining serta melakukan pengukuran terhadap performa algoritma Naïve Bayes dengan mengacu kepada Confusion Matrix dan Kurva ROC. Data sekunder yang digunakan merupakan data publik yang bernama HCC Survival Data Set yang dirilis pada tahun 2017 dan diperoleh dari website UCI Machine Learning Repository. Algortima Naïve Bayes merupakan salah satu algoritma yang terdapat dalam metode data mining yang menerapkan Teori Keputusan Bayes pada teknik klasifikasi dengan menggunakan cara pendekatan statistik yang bersifat fundamental dalam pengenalan pola. Teknik validasi yang digunakan menggunakan teknik 10-Fold Cross-Validation dengan rasio pembagian data sebesar 90:10. Perangkat lunak yang digunakan adalah RapidMiner Studio v9.5. Hasil penelitian menunjukkan bahwa hasil performa algoritma Naïve Bayes yang diukur menggunakan Confussion Matrix dengan nilai yang dihasilkan berupa nilai Akurasi sebesar 70,30%, Presisi sebesar 73,53% dan Recall sebesar 77,32% serta hasil performa yang diukur menggunakan Kurva ROC (Receiver Operating Characteristic) dengan nilai yang dihasilkan berupa nilai AUC sebesar 0.783 yang termasuk dalam kategori Fair Classification atau kategori Klasifikasi Cukup.

Kata Kunci: Penyakit *Hepatocelullar Carcinoma*, *HCC Survival Data Set*, *Naïve Bayes*, Klasifikasi.

Abstract

Hepatocellular Carcinoma is a primary liver malignant tumor derived from hepatocytes. In the last decade there is considerable development means concerning the disease of Hepatocellular Carcinoma's. This research aims to classify the possible life levels of patients who have been diagnosed with Hepatocellular Carcinoma by using the application of data mining methods and performing measurements of the Naïve Bayes algorithm performance's with reference to the Confusion Matrix and the ROC Curve. The secondary data used is a public data called HCC Survival Data Set which was released in 2017 and obtained from the UCI Machine Learning Repository website. Naïve Bayes algorithm is one of the algorithms contained in data mining methods that apply Bayes decision theory to the classification technique by using a statistical approach that is fundamental in the introduction of patterns. The validation technique is used using the 10-Fold Cross-Validation technique with a data sharing ratio of 90:10. The software used is RapidMiner Studio v 9.5. The results showed that Naïve Bayes algorithm performance's was measured using Confussion Matrix with the resulting value of accuracy value of 70.30%,

Naskah diterima 23 Januari 2021; direvisi 17 Februari 2021; diterbitkan 28 Februari 2021

precision at 73.53% and Recall of 77.32% and the performance results Measured using the Receiver Operating Characteristic with the resulting value in the form of a AUC of 0783 which belongs to the category of Fair Classification or sufficient classification category.

***Keywords**: Hepatocelullar Carcinoma Disease, HCC Survival Data Set, Naïve Bayes, Classification.*

1. Pendahuluan

Kanker merupakan suatu penyakit yang disebabkan oleh pertumbuhan sel-sel jaringan tubuh yang tidak normal. Sel-sel kanker akan berkembang dengan cepat, tidak terkendali, dan akan terus membelah diri, selanjutnya menyusup ke jaringan di sekitarnya *(invasive)* dan terus menyebar melalui jaringan ikat, darah, dan menyerang organ-organ penting serta saraf tulang belakang. Dalam keadaan normal, sel hanya akan membelah diri jika ada penggantian sel-sel yang telah mati dan rusak. Sebaliknya, sel kanker akan membelah terus meskipun tubuh tidak memerlukannya, sehingga akan terjadi penumpukan sel baru. Penumpukan sel tersebut mendesak dan merusak jaringan normal, sehingga mengganggu organ yang ditempatinya (Sugiarti, 2015).

Salah satu jenis kanker yang patut diwaspadai adalah kanker hati atau dalam bahasa medis biasa disebut juga dengan *Hepatocellular Carcinoma (HCC)*. Kanker hati merupakan salah satu jenis kanker yang paling sering ditemui oleh masyarakat (Yulianto, Kuzairi, & Hasanah, 2016), hal ini terbukti dari data yang didapat dari data Globocan yang menyebutkan bahwa di tahun 2018 terdapat 18,1 juta kasus baru dengan angka kematian sebesar 9,6 juta kematian, dimana 1 dari 5 laki-laki dan 1 dari 6 perempuan di dunia mengalami kejadian kanker. Data tersebut juga menyatakan 1 dari 8 laki-laki dan 1 dari 11 perempuan, meninggal karena kanker (Kemenkes, 2019).

Penderita dinyatakan mengidap kanker *Hepatocellular Carcinoma (HCC)* setelah didiagnosa memiliki kanker tahap lanjut, sehingga diperlukan diagnosis lebih awal untuk mendeteksi kanker *Hepatocellular Carcinoma (HCC)* agar pasien dapat memiliki kesempatan untuk selamat dari penyakit tersebut (Nugraha, Shidiq, & Rahayu, 2019).

Data mining merupakan kegiatan mengekstrak informasi atau pengetahuan *(knowledge)* penting dari suatu set data berukuran besar yang meliputi pengumpulan, pemakain data historis untuk menentukan pola keteraturan, pola hubungan dengan menggunakan teknik tertentu (Santosa & Umam, 2018). Dalam dunia kesehatan penggunaan metode data mining telah banyak membantu dunia kesehatan dalam membuat prediksi mengenai masalah kesehatan yang dihadapi (Amalia, 2018). Terdapat beberapa teknik pada data mining yang dapat digunakan untuk mengelola hasil diagnosis pada dunia medis, salah satunya adalah dengan menggunakan teknik klasifikasi. Teknik klasifikasi pada data mining juga pernah digunakan pada penelitian (Susanti, 2019), (Ramdhani et al., 2018) & (Ramdhani, 2015).

Salah satu metode yang sering digunakan untuk menerapkan teknik klasifikasi adalah metode *Naïve Bayes*. *Naïve Bayes* merupakan sebuah metode untuk teknik pengklasifikasian dengan konsep probabilistik sederhana yang menghitung sekumpulan probabilitas dengan menjumlahkan frekuensi dan kombinasi nilai dari dataset yang diberikan. Adapun keuntungan dalam penggunaan metode *Naïve Bayes* sebagai metode untuk menerapkan teknik klasifikasi yaitu metode ini hanya membutuhkan jumlah data pelatihan *(Training Data)* yang kecil untuk menentukan estimasi paremeter yang diperlukan dalam proses pengklasifikasian (Saleh, 2015).

2. Metode Penelitian

Pada penelitian ini akan menggunakan metodologi penelitian secara keseluruhan yang didasarkan pada konsep *Knowledge Discovery in Database (KDD)*. Berikut adalah tahapan yang dilakukan dalam penelitian ini :

Gambar 1. Desain Penelitian
Sumber: Penelitian (2019)

Data Selection

Pada tahapan awal ini, penulis mengambil data sekunder yang telah disediakan oleh website *UCI Machine Learning Repository* dengan judul *HCC Survival Data Set* yang dipublikasikan pada tahun 2017 dengan berisikan data pasien sebanyak 165 orang dan setiap data tersebut memiliki atribut atau fitur data sebanyak 49 atribut/fitur dengan jumlah pasien yang terdiagnosis 0 atau *Dies* (Meninggal) sebanyak 63 pasien dan jumlah pasien yang terdiagnosis 1 atau *Lives* (Selamat) sebanyak 102 pasien.

Data Pre-Processing/Cleaning

Pada tahap ini, penulis melakukan pre-processing terhadap data sekunder yang telah didapatkan sebelumnya dengan tujuan agar dapat menyesuaikan dan memperbaiki isi dari data sekunder tersebut sesuai dengan deskripsi yang diberikan oleh donatur data sekunder tersebut. Data pre-processing dalam penelitian ini dilakukan dengan cara memperbaiki data dan mengisi missing value yang ada dengan cara menggantinya menjadi nilai rata-rata dari nilai atribut/fitur keseluruhan dari masing-masing jenis atribut/fitur yang dilakukan untuk setiap masing-masing jenis atribut/fitur yang masih memiliki missing value. Pengisian missing value juga dilakukan untuk jenis atribut/fitur yang memiliki data bersifat nominal maupun ordinal dengan cara membulatkan nilai dari rata-rata untuk seluruh data pada tiap masing-masing jenis atribut/fitur dengan memberi nilai 1 jika nilai rata-rata dari jumlah data pada atribut/fitur tersebut bernilai > 0.5 dan memberikan nilai 0 jika nilai rata-rata dari jumlah data pada atribut/fitur tersebut bernilai <= 0.5. Untuk jenis atribut/fitur yang memiliki data bersifat numerik maupun rasio dilakukan pengisian missing value dengan cara mengisi missing value tersebut dengan nilai rata-rata dari jumlah data pada atribut/fitur tersebut dan hanya nilai dibelakang koma yang akan dibulatkan menjadi nilai terdekat dari koma.

Data Transformation

Pada tahap ini, penulis melakukan proses transformasi data dengan mengubah tipe data yang mengandung nilai continuous menjadi tipe data yang mengandung nilai kategorikal. Hal ini dilakukan agar kinerja dari algoritma *Naïve Bayes* yang digunakan bisa lebih optimal dalam mengklasifikasikan suatu kasus. Pada *HCC Survival Data Set* yang digunakan pada penelitian ini, terdapat jenis nilai atribut/fitur campuran antara jenis nilai atribut/fitur yang bersifat nominal maupun ordinal dengan jenis nilai atribut/fitur yang bersifat numerik maupun rasio. Proses transformasi data yang digunakan pada penelitian ini adalah metode Discretization, dimana dalam tahap ini nilai-nilai baku dari atribut/fitur dengan jenis numerik (misalnya; *Age at diagnosis*) akan diganti dan dikelompokkan dengan suatu rentang nilai yang telah ditentukan (misalnya; 20-25 tahun digolongkan menjadi Remaja Akhir). Proses ini dilakukan untuk mengubah semua jenis nilai atribut/fitur yang bersifat nominal maupun ordinal dan numerik maupun rasio menjadi jenis nilai atribut/fitur dengan jenis nilai interval maupun kategorikal. Berikut ini adalah transformasi untuk 49 jenis atribut/fitur yang dilakukan proses *Discretization*.

Tabel 1. Hasil Transformasi Atribut/Fitur Menggunakan *Discretization*

No.	Atribut/Fitur	Nilai
1.	Gender	1 = Male 0 = Female
2.	Symptoms	1 = Yes 0 = No
3.	Alcohol	1 = Yes 0 = No
4.	Hepatitis B Surface Antigen	1 = Yes 0 = No
5.	Hepatitis B e Antigen	1 = Yes 0 = No
6.	Hepatitis B Core Antibody	1 = Yes 0 = No

7.	Hepatitis C Virus Antibody	1 = Yes 0 = No
8.	Cirrhosis	1 = Yes 0 = No
9.	Endemic Countries	1 = Yes 0 = No
10.	Smoking	1 = Yes 0 = No
11.	Diabetes	1 = Yes 0 = No
12.	Obesity	1 = Yes 0 = No
13.	Hemochromatosis	1 = Yes 0 = No
14.	Arterial Hypertension	1 = Yes 0 = No
15.	Chronic Renal Insufficiency	1 = Yes 0 = No
16.	Human Immunodeficiency Virus	1 = Yes 0 = No
17.	Nonalcoholic Steatohepatitis	1 = Yes 0 = No
18.	Esophageal Varices	1 = Yes 0 = No
19.	Splenomegaly	1 = Yes 0 = No
20.	Portal Hypertension	1 = Yes 0 = No
21.	Portal Vein Thrombosis	1 = Yes 0 = No
22.	Liver Metastasis	1 = Yes 0 = No
23.	Radiological Hallmark	1 = Yes 0 = No
24.	Age at diagnosis	20-25 = Remaja Akhir 26-35 = Dewasa Awal 36-45 = Dewasa Akhir 46-55 = Lansia Awal 56-65 = Lansia Akhir 66-95 = Manula
25.	Grams of Alcohol per day	0-166,7 = Low 166,8-333,4 = Medium 333,5-500 = High
26.	Packs of cigarets per year	0-170 = Low 171-340 = Medium 341-510 = High
27.	Performance Status	0 = Active 1 = Restricted 2 = Ambulatory 3 = Selfcare 4 = Disabled 5 = Dead
28.	Encephalopathy degree	1 = None 2 = Grade I/II 3 = Grade III/IV
29.	Ascites degree	1 = None 2 = Mild 3 = Moderate to Severe
30.	International Normalised Ratio	0-1,67 = Low 1,68-3,34 = Medium 3,35-5 = High
31.	Alpha-Fetoprotein (ng/mL)	0-603448,67 = Low 603448,68-1206897,34 = Medium 1206897,35-1810346 = High
32.	Haemoglobin (g/dL)	0-6,34 = Low 6,35-12,68 = Medium 12,69-19 = High
33.	Mean Corpuscular Volume (f1)	0-40 = Low 41-80 = Medium 81-120 = High
34.	Leukocytes (G/L)	0-4333,34 = Low 4333,35-8666,68 = Medium 8666,69-13000 = High
35.	Platelets (G/L)	0-153000 = Low 153001-306000 = Medium 306001-459000 = High
36.	Albumin (mg/dL)	0-1,67 = Low 1,68-3,34 = Medium 3,35-5 = High
37.	Total Bilirubin (mg/dL)	0-13,67 = Low 13,68-27,34 = Medium 27,35-41 = High
38.	Alanine transaminase (U/L)	0-140 = Low 141-280 = Medium 281-420 = High
39.	Aspartate transaminase (U/L)	0-184,34 = Low 184,35-368,68 = Medium 368,69-553 = High
40.	Gamma glutamyl transferase	0-525 = Low 526-1050 = Medium 1051-1575 = High
41.	Alkaline phosphatase (U/L)	0-326,67 = Low 326,68-653,34 = Medium 653,35-980 = High
42.	Total Proteins (g/dL)	0-34 = Low 35-68 = Medium 69-102 = High
43.	Creatinine (mg/dL)	0-2,67 = Low 2,68-5,34 = Medium 5,35-8 = High
44.	Number of Nodules	0 = Very Low 1 = Low 2 = Medium 3 = High 4 = Very High 5 = Extremely High
45.	Major dimension of nodule	0-7,34 = Low 7,35-14,68 = Medium 14,69-22 = High
46.	Direct Bilirubin (mg/dL)	0-10 = Low 11-20 = Medium 21-30 = High
47.	Iron (mcg/dL)	0-81,34 = Low 81,35-162,68 = Medium 162,69-244 = High
48.	Oxygen Saturation (%)	0-42 = Low 43-84 = Medium 85-126 = High
49.	Ferritin (ng/mL)	0-743,34 = Low 743,35-1486,68 = Medium 1486,69-2230 = High

Sumber: Penelitian (2019)

Data Mining

Pada tahap ini, penulis melakukan proses penggalian data untuk menemukan suatu pengetahuan dari sekumpulan data yang ada dengan menggunakan metode atau teknik

tertentu yang biasa dikenal dengan istilah *Data Mining*. Pada penelitian ini, penulis menggunakan teknik klasifikasi untuk menggali pengetahuan yang dapat dihasilkan dari data sekunder *HCC Survival Data Set* dengan menerapkan algoritma *Naïve Bayes* sebagai algoritma untuk mengatasi masalah klasifikasi data dan menggunakan metode *k-Fold Cross-Validation* sebagai metode untuk memvalidasi data serta membagi data secara acak menjadi *k* subhimpunan (biasanya disebut *fold*) yang saling bebas sehingga masing-masing fold berisi 1/k bagian data yang kemudian masing-masing himpunan data berisi *(k – 1)* fold untuk data training dan *1 fold* untuk data testing dengan rasio pembagian data sebesar *90:10* yaitu 90% digunakan untuk *data training* dan 10% sisanya digunakan untuk *data testing*. Jumlah *fold* atau subhimpunan untuk memvalidasi data yang digunakan dalam penelitian ini adalah *10-Fold Validation*.

Proses data mining dilakukan dengan menggunakan software *RapidMiner Studio v9.5* dengan mengusung update terbaru dari segi performa maupun tampilan.

Evaluation

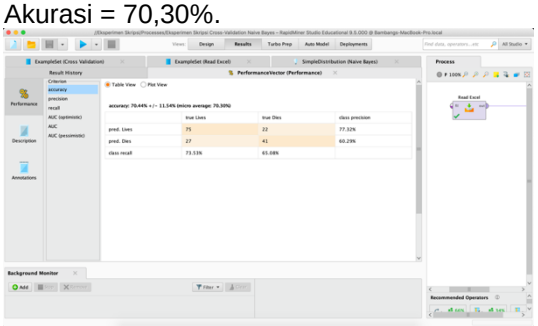
Pada tahap ini akan ditampilkan evaluasi terhadap kualitas dan efektifitas dari model yang telah dibangun. Proses evaluasi akan menggunakan metode Confusion Matrix serta Kurva ROC *(Receiver Operating Characteristic)* untuk mengetahui nilai akurasi yang dihasilkan dari model yang telah dibangun sebelumnya dengan menggunakan fitur dari software *RapidMiner Studio v9.5*. Tahap evaluasi dengan menggunakan software *RapidMiner Studio v9.5* dilakukan secara otomatis dibelakang layar ketika mengeksekusi proses model yang telah dibangun sebelumnya. Hasil evaluasi dapat dilihat pada bagian tab Result setelah mengeksekusi model yang telah dibangun.

3. Hasil dan Pembahasan

Penerapan teknik klasifikasi dan penggunaan algoritma *Naïve Bayes* pada *HCC Survival Data Set* dimaksudkan untuk mengetahui dan untuk mendapatkan hasil performa yang baik pada masalah pengklasifikasian tingkat kemungkinan hidup pasien yang telah didiagnosis menderita penyakit *Hepatocellular Carcinoma* berdasarkan probabilitas atau kemungkinan yang dihasilkan dari beberapa gejala penyakit tersebut. Hasil performa yang optimal dapat diraih atau tidak, akan terlihat pada proses akhir eksperimen yang dilakukan. Eksperimen pada *HCC Survival Data Set* dilakukan dengan menggunakan teknik klasifikasi yang ditujukan untuk dapat memprediksi kemungkinan pasien dengan data gejala penyakit tertentu dapat dinyatakan 0 atau *Dies* (Meninggal) dan 1 atau *Lives* (Selamat) dengan menerapkan algortima *Naïve Bayes* untuk digunakan sebagai algoritma yang akan digunakan dalam teknik klasifikasi serta menggunakan metode *k-Fold Cross-Validation* untuk memvalidasi data set yang telah dipilih.

3.1. Hasil Eksperimen Berupa Nilai *Accuracy*

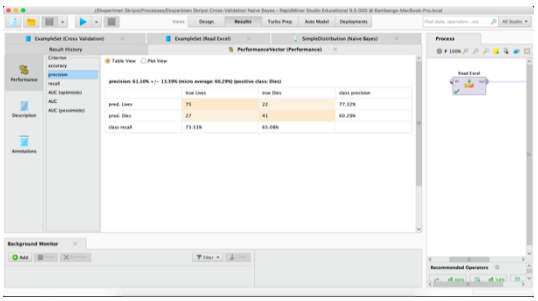
Dari confusion matrix yang telah dihasilkan tersebut, dapat dihasilkan nilai akurasi dari model yang sudah dibangun sebelumnya yang dapat dirumuskan dengan cara seperti dibawah ini:
TP = 75; TN = 41; P = 97; N = 68
Akurasi = (TP+TN)/(P+N)
Akurasi = (75+41)/(97+68)
Akurasi = 70,30%.



Gambar 2. Hasil Akurasi Evaluasi Model *k-Fold Cross-Validation*
Sumber: Penelitian (2019)

3.2. Hasil Eksperimen Berupa Nilai *Precision*

Dari confusion matrix yang telah dihasilkan tersebut, dapat dihasilkan nilai presisi dari model yang sudah dibangun sebelumnya yang dapat dirumuskan dengan cara seperti dibawah ini:
TP = 75; FP = 27
Presisi = TP/(TP+FP)
Presisi = 75/(75+27)
Presisi = 73,53%.



Gambar 3. Hasil Presisi Evaluasi Model *k-Fold Cross-Validation*

3.3. Hasil Eksperimen Berupa Nilai *Recall*

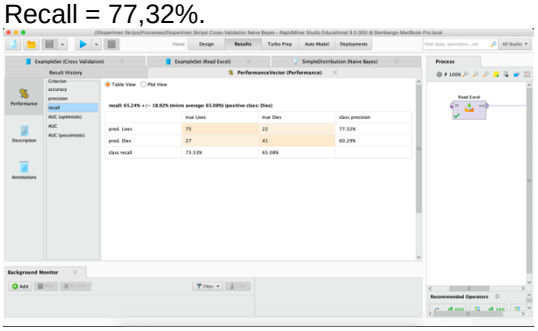
Dari confusion matrix yang telah dihasilkan tersebut, dapat dihasilkan nilai recall dari model yang sudah dibangun sebelumnya yang dapat dirumuskan dengan cara seperti dibawah ini:

TP = 75; P = 97

Recall = TP/P

Recall = 75/97

Recall = 77,32%.

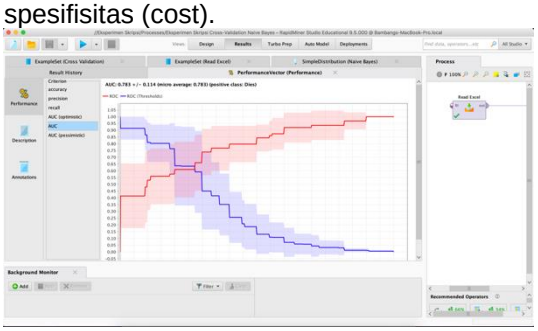


Gambar 4. Hasil Recall Evaluasi Model *k-Fold Cross-Validation*

Sumber: Penelitian (2019)

3.4. Hasil Eksperimen Berupa Kurva ROC

Dari visualisasi grafik Kurva ROC yang telah dihasilkan tersebut, Kurva ROC menunjukkan akurasi dan membandingkan klasifikasi secara visual dengan false positive rate (specificity) sebagai garis horizontal dan true positive rate (sensitivity) sebagai garis vertikal dengan menggambarkan tawar menawar antara sensitivitas (benefit) dan spesifisitas (cost).



Gambar 5. Visualisasi Grafik Kurva ROC

Sumber: Penelitian (2019)

Dari grafik Kurva ROC tersebut dapat dihasilkan nilai AUC (Area Under Receiver Operating Characteristics Curve) dari model yang sudah dibangun sebelumnya dengan nilai AUC sebesar 0.783 yang termasuk didalam kategori Fair Classification atau bisa disebut juga model dengan kategori Klasifikasi Cukup.

3.5. Perhitungan Manual Algoritma Naïve Bayes

Untuk mencari nilai prediksi pada data testing dalam tahap penerapan algoritma Naïve Bayes untuk masalah klasifikasi ini, dilakukan dengan cara menghitung probabilitas prior, probabilitas bersyarat, probabilitas setiap kelas dan probabilitas posterior untuk mencari hasil dari probabilitas atau kemungkinan dari setiap masing-masing data.

Berikut adalah contoh singkat perhitungan manual proses klasifikasi dengan menggunakan algoritma *Naïve Bayes* pada salah satu data pasien yang diambil secara acak dari dalam data set yang digunakan.

Tabel 2. Perhitungan Manual Naïve Bayes

No.	Nilai	Probabilitas Bersyarat	Probabilitas Atribut
1.	Male	P(Gender = Male | Class Attribute = Lives) = 81/102 P(Gender = Male | Class Attribute = Dies) = 52/63	0,794 0,825
2.	No	P(Symptoms = No | Class Attribute = Lives) = 41/102 P(Symptoms = No | Class Attribute = Dies) = 12/63	0,402 0,190
3.	Yes	P(Alcohol = Yes | Class Attribute = Lives) = 74/102 P(Alcohol = Yes | Class Attribute = Dies) = 48/63	0,725 0,762
4.	No	P(Hepatitis B Surface Antigen = No | Class Attribute = Lives) = 91/102 P(Hepatitis B Surface Antigen = No | Class Attribute = Dies) = 58/63	0,892 0,921

5.	No	P(Hepatitis B e Antigen = No | Class Attribute = Lives) = 102/102 P(Hepatitis B e Antigen = No | Class Attribute = Dies) = 62/63	1 0,984

Untuk dapat menghitung probabilitas *posterior* dari contoh salah satu data pasien tersebut, diperlukan nilai probabilitas *prior* dari data set tersebut seperti perhitungan yang dilakukan dibawah ini:

Probabilitas *Prior Lives* = P(*Class Attribute* = *Lives*) = $\frac{102}{165}$ = 0,618
Probabilitas *Prior Dies* = P(*Class Attribute* = *Dies*) = $\frac{63}{165}$ = 0,382

Setelah itu, diperlukan perhitungan probabilitas untuk setiap jenis atribut yang dilakukan sebagai berikut:

$$P(X|Class\ Attribute = Lives)$$
$$= P(\text{Gender} = \text{Male}|Class\ Attribute = Lives)$$
$$\times P(\text{Symptoms} = \text{No}|Class\ Attribute = Lives)$$
$$\times P(\text{Alcohol} = \text{Yes}|Class\ Attribute = Lives)$$
$$\times P\left(\text{Hepatitis B Surface Antigen} = \text{No}\middle|\begin{matrix}Class\ Attribute\\ = Lives\end{matrix}\right)$$
$$\times P(\text{Hepatitis B e Antigen} = \text{No}|Class\ Attribute = Lives)$$
$$== 0{,}794 \times 0{,}402 \times 0{,}725 \times 0{,}892 \times 1$$
$$= \mathbf{1{,}66049E - 10}$$
$$P(X|Class\ Attribute = Dies)$$
$$= P(Gender = Male|Class\ Attribute = Dies)$$
$$\times P(Symptoms = No|Class\ Attribute = Dies)$$
$$\times P(Alcohol = Yes|Class\ Attribute = Dies)$$
$$\times P\left(Hepatitis\ B\ Surface\ Antigen = No\middle|\begin{matrix}Class\ Attribute\\ = Dies\end{matrix}\right)$$
$$\times P(Hepatitis\ B\ e\ Antigen = No|Class\ Attribute = Dies)$$
$$== 0{,}825 \times 0{,}190 \times 0{,}762 \times 0{,}921 \times 0{,}984$$
$$= \mathbf{1{,}32742E - 11}$$

Adapun perhitungan yang dilakukan untuk mengetahui nilai probabilitas *posterior* dari contoh salah satu data pasien tersebut adalah sebagai berikut:

$$P(X \mid Class\ Attribute = Lives) \times P(Class\ Attribute = Lives)$$
$$= 1{,}66049\text{E} - 10 \times 0{,}618 = \mathbf{1{,}02648E - 10}$$
$$P(X \mid Class\ Attribute = Dies) \times P(Class\ Attribute = Dies)$$
$$= 1{,}32742\text{E} - 11 \times 0{,}382 = \mathbf{5{,}06832E - 12}$$

Berdasarkan hasil perhitungan manual tersebut, dapat diketahui bahwa probabilitas *posterior P*(*Class Attribute* = *Lives*) lebih besar daripada probabilitas *posterior P*(*Class Attribute* = *Dies*). Sehingga hasil klasifikasi yang diperoleh untuk contoh salah satu data pasien tersebut bernilai ***Lives*** atau **Selamat**.

4. Kesimpulan

Pada penelitian ini dilakukan eksperimen terhadap algoritma *Naïve Bayes* untuk mengklasifikasi tingkat kemungkinan hidup pasien yang telah didiagnosis menderita penyakit Hepatocellular Carcinoma. Untuk mencari hasil optimal dari klasifikasi tersebut, diterapkan metode validasi untuk mengetahui performa klasifikasi berdasarkan *Confusion Matrix* dan Kurva ROC dengan menggunakan metode validasi *k – Fold Cross Validation*. Kesimpulan yang didapat dari hasil penelitian ini adalah :

1. Telah diterapkan algoritma *Naïve Bayes* untuk mengklasifikasikan tingkat kemungkinan hidup pasien yang telah didiagnosis menderita penyakit *Hepatocellular Carcinoma* berdasarkan probabilitas atau kemungkinan yang dihasilkan dari beberapa gejala penyakit yang didapatkan dari catatan medis setiap individu pasien.
2. Telah diketahui hasil performa algoritma *Naïve Bayes* dalam mengklasifikasi penyakit berdasarkan *HCC Survival Data Set* yang diukur menggunakan *Confussion Matrix* dengan nilai yang dihasilkan berupa nilai Akurasi sebesar 70,30%, Presisi sebesar 73,53% dan Recall sebesar 77,32% dengan menggunakan metode validasi *10 – Fold Cross Validation.*
3. Telah diketahui hasil performa algoritma *Naïve Bayes* dalam mengklasifikasi penyakit berdasarkan *HCC Survival Data Set* yang diukur menggunakan kurva ROC *(Receiver Operating Characteristic)* dengan nilai yang dihasilkan berupa nilai AUC *(Area Under Receiver Operating Characteristics Curve)* sebesar 0.783 yang termasuk didalam kategori *Fair Classification* atau bisa disebut juga model dengan kategori Klasifikasi Cukup.

Referensi

Amalia, H. (2018). Perbandingan Metode Data Mining SVM dan NN Untuk Klasifikasi Penyakit Ginjal Kronis. Jurnal PILAR Nusa Mandiri, 1-6.

Kemenkes. (2019). *Hari Kanker Sedunia 2019*. https://www.kemkes.go.id/article/view/19020100003/hari-kanker-sedunia-

2019.html

Nugraha, F. S., Shidiq, M. J., & Rahayu, S. (2019). Analisis Algoritma Klasifikasi Neural Network Untuk Diagnosis Penyakit Kanker Payudara . Jurnal PILAR Nusa Mandiri, 149-156.

Ramdhani, Y. (2015). Komparasi Algoritma LDA Dan Naïve Bayes Dengan Optimasi Fitur Untuk Klasifikasi Citra Tunggal Pap Smear. *Jurnal Informatika*, *2*(2).

Ramdhani, Y., Susanti, S., Adiwisastra, M. F., & Topiq, S. (2018). Penerapan Algoritma Neural Network Untuk Klasifikasi Kardiotokografi. *Jurnal Informatika*, *5*(1), 43–49. https://doi.org/10.31311/ji.v5i1.2832

Saleh, A. (2015). Implementasi Metode Klasifikasi Naïve Bayes Dalam Memprediksi Besarnya Penggunaan Listrik Rumah Tangga. Citec Journal, 207-217.

Santosa, B., & Umam, A. (2018). Data Mining dan Big Data Analytics Teori dan Implementasi Menggunakan Python dan Apache Spark. Yogyakarta: Penebar Media Pustaka.

Sugiarti, M. (2015). Pengaruh Khemoterapi Terhadap Jumlah Trombosit Pasien Penderita Kanker di RS Abdul Moeloek Provinsi Lampung. Jurnal Analis Kesehatan, 450-455.

Susanti, S. (2019). Klasifikasi Kemampuan Perawatan Diri Anak dengan Disabilitas Menggunakan SMOTE Berbasis Neural Network. *Jurnal Informatika*, *6*(2), 175-184.

Yulianto, T., Kuzairi, & Hasanah, R. (2016). Mplementasi Metode Lagrange Untuk Optimasi Penyakit Kanker Hati. Unisda Journal of Mathematics and Computer Science, 62-68.