Perbandingan Hasil Prediksi Diagnosis pada *Indian Liver Patient Dataset* (ILPD) dengan Teknik *Supervised Learning* Menggunakan *Software* Orange

Ira Rosianal Hikmah[#1], Ray Novita Yasa[*2]

[#]*Program Studi Rekayasa Keamanan Siber, Politeknik Siber dan Sandi Negara*

[*]*Program Studi Rekayasa Kriptografi, Politeknik Siber dan Sandi Negara*

Jln. Raya Haji Usa, Putat Nutug, Bogor, Indonesia

[1]ira.rosianal@poltekssn.ac.id

[2]ray.novita@poltekssn.ac.id

Abstract*— The development of the volume of data every day has resulted in the need for data mining to obtain valuable and meaningful data. There are many data mining software that has been developed, both free and paid. One of the free data mining software is Orange. This software provides modeling, both supervised and unsupervised learning. Orange also provides model evaluation features, such as accuracy, precision, the time required for training and testing, specificity, and other evaluation measures. Therefore, Orange makes it easy for users to perform data mining. One of the users who need Orange is a user with a non-IT background, such as a health user who can make predictions for the diagnosis of a disease. Users do not need to focus on syntax to perform data mining. With Orange, healthcare users can easily and faster predict the diagnosis of the disease. This study uses Indian Liver Patient (ILPD) data from the UCI-Machine Learning Repository. The objective of the diagnosis is to determine whether the patient has a liver disorder or not. The methods that are used in this study are Decision Tree, Random Forest, SVM, Neural Network, Naïve Bayes, k-NN, and Logistic Regression. This study evaluates using a confusion matrix, accuracy level, precision level, training time, and testing time. The results show that the time required for training and testing is relatively short. With the data used, this study has proved that the four best methods based on accuracy are Logistic Regression, Neural Network, Random Forest, and Naïve Bayes..

***Keywords*— data mining, Orange, Supervised Learning, ILPD**

Abstrak*— Perkembangan volume data setiap hari mengakibatkan perlunya *data mining* untuk mendapatkan data berharga dan berguna. Terdapat banyak *data mining software* yang telah dikembangkan, baik gratis maupun berbayar. Salah satu *data mining software* yang gratis adalah Orange. *Sofware* ini menyediakan pemodelan, baik *supervised* maupun *unsupervised learning*. Orange juga menyediakan fitur evaluasi model, seperti akurasi, presisi, waktu yang dibutuhkan untuk *training* dan *testing*, spesifisitas, dan ukuran evaluasi lainnya. Oleh karena itu, dapat dikatakan bahwa Orange memudahkan pengguna untuk melakukan *data mining*. Salah satu pengguna yang membutuhkan Orange adalah pengguna dengan latar belakang non-IT, seperti pengguna bidang kesehatan yang dapat melakukan prediksi untuk diagnosis suatu penyakit. Pengguna tidak perlu berfokus pada sintaks untuk melakukan *data mining*. Dengan Orange, pengguna bidang kesehatan dapat memprediksi diagnosis suatu penyakit dengan lebih mudah dan lebih cepat. Penelitian ini menggunakan data *Indian Liver Patient* (ILPD) dari *UCI-Machine Learning Repository*. Targetnya adalah menentukan diagnosis pasien apakah memiliki ganguan hati atau tidak. Metode yang digunakan adalah *Decision Tree, Random Forest, SVM, Neural Network, Naïve Bayes, k-NN*, dan Regresi Logistik. Penelitian ini melakukan evaluasi dengan menggunakan *confusion matrix*, tingkat akurasi, tingkat presisi, waktu *training*, dan waktu *testing*. Hasil penelitian menunjukkan bahwa waktu yang dibutuhkan untuk *training* dan *testing* terbilang singkat. Dengan data yang digunakan, dalam penelitian ini diperoleh hasil pula empat metode terbaik berdasarkan tingkat akurasi adalah Regresi Logistik, *Neural Network, Random Forest*, dan *Naïve Bayes*.

Kata Kunci— *Data mining*, Orange, *Supervised Learning*, ILPD

I. PENDAHULUAN

Data meningkat dari hari ke hari dan sangat sulit bagi seseorang untuk menganalisis volume data yang besar untuk pengambilan keputusan yang sempurna. Oleh karena itu, diperlukan *data mining* untuk mengekstrak data yang berharga dan berguna dari data yang tersedia [1]. *Data mining* (penambangan data) merupakan proses penggalian informasi yang tersembunyi dari data yang ada untuk menemukan pola tertentu [2]. *Data mining* sering digunakan di berbagai bidang ilmu, seperti teknologi basis data, pembelajaran mesin (*machine learning*), statistika, pengambilan informasi, jaringan saraf tiruan, kecerdasan buatan, dan sebagainya [3].

Terdapat empat teknik yang digunakan untuk *machine learning,* yaitu *supervised learning, reinforcement learning, active learning*, atau *unsupervised learning*. Dua teknik yang umum digunakan adalah *supervised* dan *unsupervised learning* [4]. Terdapat tiga teknik *data mining*, yaitu klasifikasi, regresi, dan *clustering*. Klasifikasi merupakan analisis data prediktif di mana data terbagi menjadi beberapa kelas yang sudah diketahui dan diberikan label. Regresi merupakan proses melihat hubungan dan pengaruh pada suatu *dataset* yang terbagi menjadi satu variabel dependen dan satu atau lebih variabel independen. *Clustering* adalah analisis data di mana data dibagi menjadi beberapa *cluster* berdasarkan karakteristik tertentu dan belum ada label dari *cluster* tersebut

di awal analisis [5]. Klasifikasi dan regresi pada umumnya merupakan tipe pemodelan dari *supervised learning*, sedangkan *clustering* biasanya merupakan tipe pemodelan *unsupervised learning*.

Ada beberapa strategi untuk mengetahui seberapa baik model yang terbentuk. Salah satunya adalah validasi silang. Metode validasi silang yang paling sederhana adalah metode *holdout* di mana data dibagi menjadi dua bagian, yaitu data *training* dan data *testing*. Model dilatih dan dibentuk dengan data *training* serta dievaluasi pada data *testing* [6]. Pengukuran kebaikan model dapat dilakukan dengan mengukur persentase prediksi dengan benar atau akurasinya [7]. *Confusion matrix* melakukan pengujian untuk memperkirakan objek yang benar dan salah [8]. Menurut Han dan Kamber, *confusion matrix* merupakan suatu alat yang memiliki fungsi untuk melakukan analisis *classifier* yang sudah baik dalam mengenali *tuple* dari kelas yang berbeda. Nilai dari *True Positive* (TP) dan *True Negative* (TN) memberikan informasi ketika *classifier* melakukan klasifikasi data bernilai benar. *False Positive* (FP) dan *False Negative* (FN) memberikan informasi ketika *classifier* salah dalam melakukan klasifikasi data [9].

Beberapa tahun yang lalu terdapat banyak perangkat lunak *data mining* dikembangkan. Beberapa dari perangkat lunak tersebut tersedia secara bebas dan gratis [1]. Perangkat lunak *open source* merupakan perangkat lunak komputer di mana *source code* tersedia untuk umum. Pengguna dapat menggunakan, memeriksa, memperbaharui, dan mendistribusikan kepada siapa saja untuk digunakan. Salah satu perangkat lunak *data mining* yang *open source* adalah Orange yang dapat digunakan untuk visualisasi data dan analisis data [1]. Orange menyediakan pemodelan, baik *supervised* maupun *unsupervised learning*. Bahkan dengan *widget test and score*, pengguna dapat menjalankan beberapa model sekaligus pada perangkat lunak tersebut. Orange juga menyediakan fitur evaluasi model seperti akurasi, presisi serta waktu yang dibutuhkan untuk *training* dan *testing*, spesifisitas, dan sebagainya [10]–[13]. Oleh karena itu, dapat dikatakan bahwa *software* Orange memudahkan bagi pengguna untuk melakukan *data mining* karena hanya dengan menggunakan fitur yang disediakan, hasil klasifikasi, regresi, atau *clustering* sudah dapat diperoleh.

Salah satu pengguna yang membutuhkan *software* Orange adalah pengguna di bidang kesehatan yang perlu melakukan prediksi terhadap diagnosis suatu penyakit. Pengguna bidang kesehatan tidak perlu berfokus pada sintaks untuk melakukan *data mining*. Dengan Orange, pengguna bidang kesehatan dapat memprediksi diagnosa suatu penyakit dengan lebih mudah dan lebih cepat. Penelitian ini menggunakan data *Indian Liver Patient* dari UCI-*Machine Learning Repository* dengan targetnya adalah menentukan diagnosa pasien apakah memiliki ganguan hati (liver) atau tidak (nonliver). Dilakukan klasifikasi terhadap data tersebut dengan menggunakan beberapa model pada metode *supervised learning* dan menggunakan perangkat lunak *open source* Orange. Penelitian ini juga melakukan evaluasi pada model yang terbentuk dengan menggunakan *confusion matrix*, tingkat akurasi, tingkat presisi, waktu *training*, dan waktu *testing*.

Penelitian ini diharapkan dapat menunjukkan bahwa penggunaan perangkat lunak *open source* Orange untuk *data mining* dapat mempercepat waktu. Pengguna hanya membuat *workflow* dan melakukan konfigurasi dengan klik tanpa membuat sintaks Python, walaupun *software* Orange memiliki *widget* Python Script. Pengguna dengan latar belakang non-IT, khususnya pada data penelitian ini, dapat melakukan klasifikasi di bidang kesehatan dengan *software* ini. Selain itu, dapat dipilih beberapa metode pada teknik *supervised learning* kemudian dibandingkan berdasarkan tingkat akurasi dan presisi tertinggi serta waktu *training* dan *testing* yang singkat sehingga diperoleh metode yang terbaik dalam memprediksi diagnosis penyakit.

II. METODOLOGI

A. Data Penelitian

Penelitian ini menggunakan *Indian Liver Patient Dataset* (ILPD) dari repositori UCI – Machine Learning Repository [14]. Data tersebut terdiri atas 583 kasus yang terdiri atas 416 pasien liver dan 167 pasien nonliver. Data tersebut dikumpulkan dari India. ILPD berisi 441 catatan pasien pria dan 142 pasien wanita. Setiap pasien yang usianya lebih dari 89 tahun, terdaftar sebagai usia 90 tahun. Data ini terdiri atas 11 atribut (keterangan), yaitu:

1. Usia Pasien
2. Jenis Kelamin
3. TB (*Total Bilirubin)*
4. DB (*Direct Bilirubin*)
5. Alkfos (*Alkaline Phosphatase*)
6. SGPT (*Alamine Aminotransferase*)
7. SGOT (*Aspartate Aminotransferase*)
8. TP (*Total Proteins*)
9. ALB (*Albumin*)
10. Rasio A/G (*Albumin & Globulin Ratio*)
11. Kelas (1 menunjukkan pasien liver dan 2 menunjukkan pasien nonliver)

B. Klasifikasi dan Regresi

Klasifikasi *data mining* adalah penempatan objek-objek ke salah satu dari beberapa kelas yang telah ditentukan sebelumnya. Klasifikasi banyak digunakan untuk memprediksi kelas pada suatu label tertentu dengan membangun model berdasarkan data *training* lalu menggunakan model tersebut untuk mengklasifikasikan data *testing* [15]. Pada penelitian ini terdapat 2 kelas, yaitu kelas 1 menunjukkan objek pasien liver dan kelas 2 menunjukkan objek pasien nonliver. Regresi merupakan teknik *data mining* yang termasuk ke dalam *supervised learning*. Regresi digunakan untuk memprediksi target numerik, seperti regresi linier sederhana [16]. Jika variabel dependennya merupakan variabel kategori, maka Regresi Logistik dapat digunakan. Jika variabel dependennya hanya dua kategori, maka disebut Regresi Logistik binomial. Namun, jika lebih dari dua kategori, maka disebut dengan Regresi Logistik multinomial [17].

Penelitian ini menggunakan tujuh metode yang umum dipilih dan digunakan serta tersedia pada *software* Orange, yaitu model *Decision Tree*, *Random Forest*, *Support Vector Machine (SVM)*, *Neural Network*, *Naïve Bayes*, *k-NN*, dan Regresi Logistik. Berikut ini merupakan penjelasan singkat untuk ketujuh metode yang digunakan pada penelitian ini:

1) Decision Tree

Metode ini menyediakan sarana untuk mendapatkan metode peramalan dalam bentuk aturan yang mudah diterapkan. Aturan-aturan ini memiliki bentuk `If-Then` yang mudah diterapkan oleh pengguna. Pendekatan *data mining* ini menerapkan pendekatan sistem berbasis aturan [18]. Sebuah pohon terdiri dari *node* keputusan yang dihubungkan dengan cabang-cabang dari simpul akar sampai *node* daun (akhir). Setiap cabang akan diarahkan ke *node* lain atau ke *node* akhir untuk menghasilkan suatu keputusan [19].

2) Random Forest

Pada dasarnya *Random Forest* merupakan *Decision Tree* berganda (*multiple Decision Tree*) [18], tetapi menambahkan beberapa keacakan saat membuat pohon. Hal ini akan menghasilkan metode yang lebih baik dan lebih stabil [1].

3) Support Vector Machine (SVM)

Metode ini pertama kali dikenalkan oleh Vladimir Vapnik. Metode ini mendapatkan popularitas karena banyak fitur menarik dan kinerja yang menjanjikan. Karena kinerja penyederhanaannya tinggi, metode ini dianggap sebagai *classifier* yang baik karena tidak membutuhkan pengetahuan sebelumnya [20]. SVM menghasilkan fungsi pemetaan *input-output* dari data *traning* yang telah diberi label. Fungsi tersebut dapat berupa fungsi klasifikasi atau regresi. Selain landasan matematika yang kuat dalam teori pembelajaran statistik, SVM menunjukkan kinerja yang sangat kompetitif di berbagai aplikasi seperti diagnosis medis, bioinformatika, pengenalan wajah, pemrosesan gambar, dan *text mining* [18].

4) Neural Network

Neural Network cenderung bekerja lebih baik ketika ada hubungan yang rumit dalam data, seperti tingkat nonlinier yang tinggi. Dengan demikian metode ini cenderung menjadi model yang layak saat adanya ketidakpastian yang tinggi. Semakin baik kecocokan yang ditentukan, semakin lama waktu yang dibutuhkan *Neural Network* untuk berlatih, meskipun sebenarnya tidak ada cara untuk memprediksi secara akurat berapa lama waktu yang dibutuhkan model tertentu untuk belajar [18].

Dalam bentuknya yang paling sederhana, *Neural Network* adalah fungsi linier. Dibutuhkan *input* berupa atribut dan koefisien optimal (atau bobot) atribut dicari untuk menghasilkan *output* yang sedekat mungkin dengan data eksperimen. *Neural Network* bisa jauh lebih rumit dan fleksibel, bahkan bisa juga berbentuk nonlinier. Secara arsitektur, *Neural Network* memiliki banyak lapisan. Lapisan pertama adalah lapisan masukan (*input layer*) dan lapisan terakhir adalah lapisan keluaran (*output layer*). Di antara dua lapisan ini, bisa ada satu atau lebih lapisan tersembunyi (*hidden layer*). Informasi mengalir dari lapisan *input* ke lapisan tersembunyi dan ke lapisan *output* [17]. Ilustrasi bentuk *Neural Network* sederhana dapat dilihat pada Gambar 1.

5) Naïve Bayes

Naïve Bayes merupakan sekelompok algoritme probabilistik sederhana yang didasarkan pada Teorema Bayes (peluang bersyarat) [1]. Metode ini merupakan salah satu algoritme klasifikasi yang paling banyak digunakan [17].

6) k-Nearest Neighbour (k-NN)

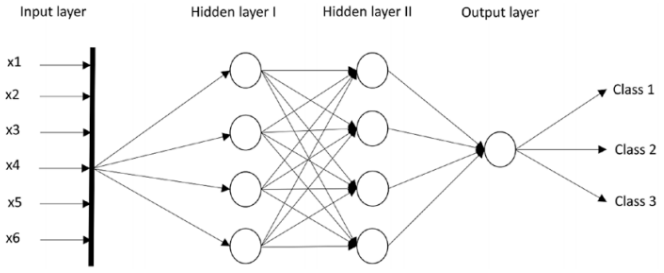
Metode ini merupakan pengklasifikasi (*classifier*) sederhana yang menyimpan semua kasus yang tersedia kemudian menghasilkan kasus baru berdasarkan pengukuran serupa, misalnya, fungsi jarak [1]. Di sini, *k* menunjukkan bahwa keputusan didasarkan pada *k* tetangga. Misalnya, di antara 10.000 sampel yang diketahui, ketika kasus baru A terjadi dan perlu memprediksi hasil A, maka akan menemukan 11 (biasanya ganjil) tetangga terdekat A dari sampel. Berdasarkan hasil mayoritas dari 11 tetangga (k = 11 dalam kasus ini), hasil dari A diprediksi [17].

Ketika berbicara tentang *machine learning*, ada dua jenis pelajar (metode): pelajar yang bersemangat (*eager learner*) dan pelajar yang malas (*lazy learner*). Metode *k-NN* termasuk ke dalam *lazy learner* karena tidak membuat model terlebih dahulu dan tidak menghasilkan serangkaian parameter atau aturan berdasarkan data *training*. Sebagai gantinya, ketika data penilaian masuk, *lazy learner* menggunakan seluruh rangkaian data pelatihan untuk mengklasifikasikan data penilaian secara dinamis. Dalam kasus *k-NN*, tetangga terdekat dari titik data penilaian dihitung secara dinamis menggunakan seluruh data *training*. Karena tidak ada set parameter yang dihitung sebelumnya, *k-NN* juga merupakan metode *data mining* nonparametric [17].

Pekerjaan utama dalam *k-NN* adalah menentukan tetangga yang terdekat. Salah satu cara untuk mengukur kedekatan adalah dengan menghitung jarak antara dua titik data. Jarak Euclidean, Manhattan, dan Chebyshev adalah yang paling banyak digunakan untuk mengukur kedekatan dan titik data numerik. Untuk atribut diskrit, jarak Hamming dapat digunakan [17].

7) Regresi Logistik

Regresi Logistik dapat dianggap sebagai kasus khusus dari regresi linier ketika yang diprediksi bersifat kategorik. Regresi



Gambar 1 Ilustrasi *Neural Network* sederhana [17]

Logistik adalah model statistik yang mampu memperkirakan probabilitas terjadinya suatu peristiwa. Beberapa data yang menarik dalam studi regresi mungkin berskala ordinal atau nominal. Karena analisis regresi memerlukan data numerik, maka dilakukan pengkodean pada variabel [18].

Misalkan 1 mewakili kasus ketika peristiwa terjadi dan 0 mewakili ketika peristiwa tidak terjadi. Oleh karena itu, $P(1)$ adalah peluang kejadian dan $P(0) = 1 - P(1)$. Misalkan terjadinya peristiwa tergantung pada variabel independen X_1, X_2, …, X_n. Dalam Regresi Logistik, *log* peluang adalah fungsi linier dari X_i, dengan $i = 1, \ldots, n$ yang digambarkan pada persamaan berikut [17]:

$$\ln\left(\frac{P(1)}{1-P(1)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \qquad (1)$$

Jika persamaan (1) diselesaikan, maka diperoleh:

$$P(1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n)}} \qquad (2)$$

dengan β_i dengan $i = 0, 1, \ldots, n$ merupakan koefisien dari Regresi Logistik [18].

C. Evaluasi

Setelah model terbentuk, dilakukan evaluasi. Ukuran yang digunakan dalam penelitian ini adalah tingkat akurasi dan presisi serta waktu *training* dan *testing*. Formula untuk tingkat akurasi dapat dihitung berdasarkan hasil *confusion matrix* sebagai berikut [21]:

$$\text{Akurasi} = \frac{TP + TN}{TP + TN + FP + FN} \qquad (3)$$

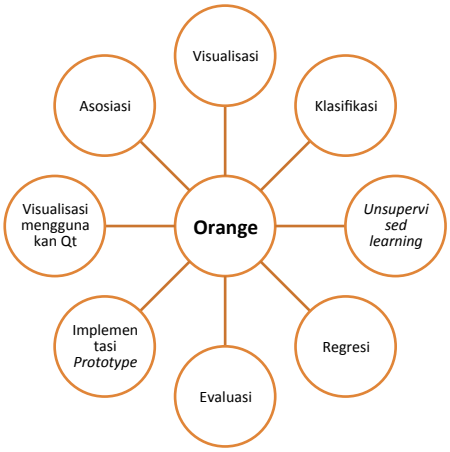
Selanjutnya, presisi merupakan nilai prediksi positif, atau dapat dihitung dengan formula berikut [1]:

$$\text{Presisi} = \frac{TP}{TP + FP} \qquad (4)$$

True Positive (TP) dan *True Negative* (TN) memberikan informasi ketika *classifier* melakukan klasifikasi data bernilai benar, sedangkan *False Positive* (FP) dan *False Negative* (FN) memberikan informasi ketika *classifier* salah dalam melakukan klasifikasi data [9]. Selain itu, penelitian ini juga menggunakan waktu *training* dan *testing* dari setiap model yang dibentuk.

D. Software Orange

Orange merupakan perangkat lunak *open source* untuk *machine learning* dan *data mining* yang ditulis dengan bahasa Python. Orange dikembangkan oleh Laboratorium Bioinformatika, Fakultas Ilmu Komputer dan Informasi, Universitas Ljubljana [22]. Fitur yang disediakan pada *software* Orange dapat dilihat pada Gambar 2.



Gambar 2 Fitur *software* Orange [1]

Berikut ini beberapa *widget* pada *software* Orange yang digunakan dalam penelitian ini: [10]

1) File

Widget ini diklik dua kali untuk membuka dan memilih file data yang akan digunakan. Output dari widget ini adalah data.

2) Select Columns

Widget ini diklik dua kali untuk menentukan atribut target pada data. Pada widget ini yang dimasukkan adalah data.

3) Data Table

Widget ini diklik dua kali untuk melihat data dalam bentuk spreadsheet. Pada widget ini yang dimasukkan adalah data.

4) Test and Score

Widget ini diklik dua kali untuk melakukan pemodelan, prediksi, dan evaluasi model. Pada widget ini yang dimasukkan adalah data dan learner (model) dengan keluarannya adalah prediksi dan hasil evaluasi.

5) Data Sampler

Widget ini diklik dua kali untuk membagi data menjadi data sampel dan *out-of-sample*, atau yang disebut dengan *remaining data*. Akan muncul tampilan dan pilih metode pembagiannya *fixed proporsion*, *fixed sample size*, dsb. Pada *widget* ini yang dimasukkan adalah data. Jika pada *widget test and score* diterapkan pengulangan *random sampling*, maka *widget* ini tidak dibutuhkan.

6) Tree, Random Forest, Neural Network, Naïve Bayes dan k-NN

Widget ini dapat digunakan dengan dua bentuk, yaitu untuk menghasilkan model dan sebagai *learner*. Jika keluarannya adalah suatu model, maka yang dimasukkan adalah data. Namun, jika keluaran sebagai *learner*, maka tidak perlu ada masukan pada *widget* ini.

7) SVM

Widget ini memiliki tiga *output* yaitu model, sebagai *learner*, dan *support vector*.

8) Logistic Regression

Widget ini memiliki tiga *output* yaitu model, sebagai *learner*, dan *coefficients*.

9) Confusion Matrix

Widget ini diklik dua kali untuk menampilkan tabulasi silang, atau dalam hal ini disebut dengan *confusion matrix*, yang berisi jumlah data yang *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negative* (FN) untuk setiap model/*learner* yang dijalankan.

E. Langkah-Langkah Pengolahan Data

Secara umum proses klasifikasi data mencakup dua langkah. Langkah pertama adalah membangun model, atau aturan klasifikasi, dengan mempelajari data *training* dengan label kelas terkait. Langkah kedua adalah menggunakan model tersebut untuk melakukan klasifikasi terhadap data *testing* untuk mendapatkan keakuratan dari model atau aturan klasifikasi tersebut [22].

Langkah-langkah secara rinci dalam pengolahan data penelitian ini adalah sebagai berikut:

1) Preprocessing

Pada tahap ini dilakukan pembersihan data untuk menghindari hal-hal, seperti *incomplete*, *noisy,* dan *inconsistent*.

2) Proses integrasi data

Pada tahap ini, dilakukan pengubahan atau konversi data ke jenis data yang diinginkan. Penelitian ini mengubah variabel target, yaitu angka 1 dengan "Liver" dan angka 2 dengan "Non".

3) Menentukan target data

4) Proses data mining

Tahap ini merupakan tahap pembentukan model *supervised learning*. Proses ini menggunakan *software* Orange dengan menggunakan fitur *test & score* [12]. Pada tahap ini dilakukan pengambilan sampel (*sampling*) secara acak untuk membaginya menjadi data *training* dan data *testing* dengan proporsi 80%:20%. Penelitian ini melakukan pengulangan *sampling* sebanyak 100 kali (maksimum pengulangan yang dapat dilakukan oleh *software* Orange). Pengulangan perlu dilakukan supaya hasil yang didapatkan lebih akurat, mengingat bahwa proses validasi silang dilakukan dengan *sampling*. Selain itu, pengulangan juga dapat meningkakan ketelitian. Jika jumlah ulangan semakin banyak, maka penelitian semakin meningkat dan tidak salah dalam pengambilan keputusan. Hal ini karena pengulangan dapat menambah cakupan penarikan kesimpulan. Pada tahap ini juga dilakukan perhitungan evaluasi model. Ukuran evaluasi yang digunakan dalam penelitian ini adalah akurasi, presisi, serta waktu *training* dan *testing*.

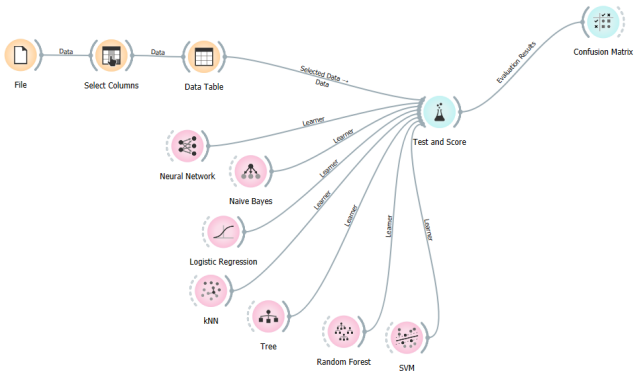
5) Pengetahuan

Tahap ini merupakan tahap terakhir dalam penelitian. Tahap ini diharapkan dapat memberikan pengetahuan, baik dalam proses *data mining* maupun dalam penggunaan *software* Orange untuk *data mining* agar dapat dimanfaatkan untuk kepentingan pemerintah dan masyarakat, khususnya di bidang non-IT.

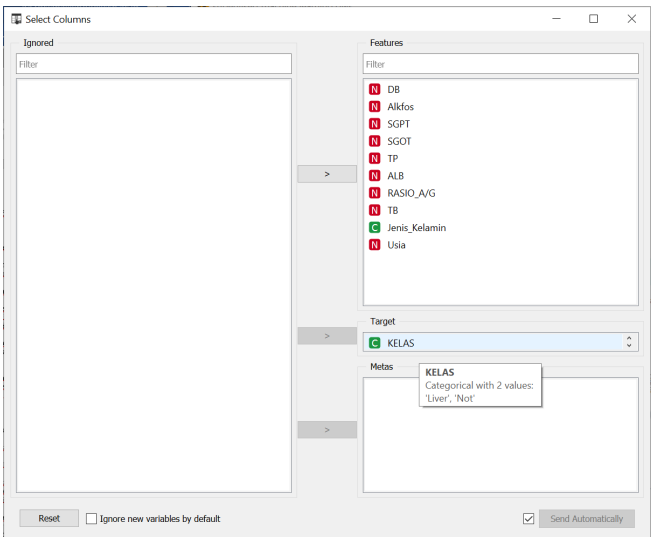
III. HASIL DAN PEMBAHASAN

Gambar 3 merupakan *workflow* penelitian dengan *software* Orange. Pada Gambar 3 terlihat bahwa terdapat tiga bagian, yang berwarna kuning, merah muda, dan hijau. Bagian yang berwarna kuning merupakan tahap persiapan data, yaitu *import* data, menentukan atribut bebas dan atribut target, dan membuat tampilan secara *spreadsheet*. Bagian yang berwarna merah menunjukkan metode *supervised learning* yang akan diterapkan. Penelitian ini menggunakan tujuh metode untuk menghasilkan *learner*. Pada bagian yang berwarna hijau akan diperoleh prediksi dan hasil evaluasi.

Salah satu hal terpenting dalam *data mining* klasifikasi adalah menentukan atribut target. Ketika data diimpor, maka semua variabel atau atribut akan dianggap sebagai atribut bebas. Oleh karena itu, *widget selected columns* perlu dilibatkan untuk memilih atribut target yang akan menjadi fokus penelitian. Proses penggunaan *widget selected columns* dapat dilihat pada Gambar 4.



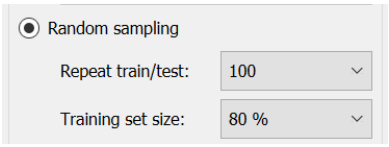
Gambar 3 *Workflow* penelitian dengan *software* Orange



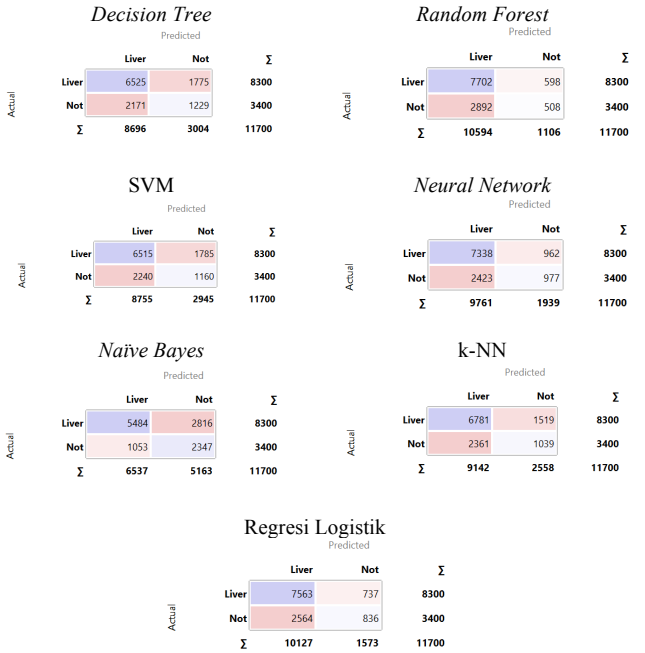
Gambar 4 Melakukan pemetaan atribut bebas dan atribut target

Karena penelitian ini menggunakan *widget test and score* dan memilih pembagian data *training* dan data *testing* dengan proporsi 80%:20% secara *random sampling* dengan pengulangan sebanyak 100 kali, maka penelitian ini tidak menggunakan *widget Data Sampler* yang telah dijelaskan pada bagian sebelumnya. Tampilan pemilihan *random sampling* pada *widget test and score* dapat dilihat pada Gambar 5.

Output dari *widget confusion matrix* dan *test and score* untuk hasil evaluasi dapat dilihat pada Gambar 6, 7, 8, dan 9. Gambar 6 menunjukkan *confusion matrix* untuk ketujuh metode yang digunakan dalam penelitian ini. Hasil prediksi benar adalah kondisi saat *True Positive* (aktual liver diprediksi liver) dan *True Negative* (aktual nonliver diprediksi nonliver). Jumlah total *True Positive* dan *True Negative* untuk metode *Decision Tree, Random Forest, Support Vector Machine (SVM), Neural Network, Naïve Bayes, k-NN,* dan *Regresi Logistik*, masing-masing adalah 7754, 8210, 7675, 8315, 7831, 7820, dan 8399. Berdasarkan hasil klasifikasi yang benar, didapatkan metode terbaik adalah Regresi Logistik yang diikuti dengan *Neural Network, Random Forest,* dan *Naïve Bayes*. Berdasarkan *confusion matrix* dapat dihitung tingkat akurasi dan presisi dari setiap metode yang dapat dilihat pada Gambar 7 dan 8.



Gambar 5 Pemilihan *random sampling* dengan pengulangan 100 kali dan proporsi data *training* dan data *testing* adalah 80%:20%



Gambar 6 *Confusion matrix* setiap model penelitian

Berdasarkan Gambar 7, dapat dilihat bahwa model dengan tingkat akurasi tertinggi adalah model Regresi Logistik yang diikuti dengan *Neural Network, Random Forest, Naïve Bayes, k-NN, Decision Tree,* dan SVM. Berbeda dengan tingkat akurasi yang mengukur kemampuan model memprediksi dengan benar, tingkat presisi disebut dengan nilai prediksi positif yaitu proporsi dengan hasil tes positif. Berdasarkan Gambar 8, terlihat bahwa model dengan tingkat presisi tertinggi adalah model *Naïve Bayes* yang diikuti dengan *Neural Network, Decision Tree,* Regresi Logistik, SVM, k-NN, dan *Random Forest*.

Gambar 9 menunjukkan evaluasi untuk ketujuh model berdasarkan waktu yang dibutuhkan untuk *training* dan *testing*.

Evaluation Results

Model	CA
Logistic Regression	0.718
Neural Network	0.711
Random Forest	0.702
Naive Bayes	0.669
kNN	0.668
Tree	0.663
SVM	0.656

Gambar 7 Hasil evaluasi berdasarkan tingkat akurasi model

Evaluation Results

Model	Precision
Naive Bayes	0.839
Neural Network	0.752
Tree	0.750
Logistic Regression	0.747
SVM	0.744
kNN	0.742
Random Forest	0.727

Gambar 8 Hasil evaluasi berdasarkan tingkat presisi model

Evaluation Results

Model	Train time [s]	Test time [s]
Tree	9.388	0.016
Naive Bayes	1.039	0.190
Logistic Regression	30.212	0.216
Neural Network	101.701	0.500
Random Forest	2.893	0.577
SVM	4.665	0.605
kNN	0.796	1.015

Gambar 9 Hasil evaluasi berdasarkan waktu *training* dan *testing*

Terlihat bahwa model *k-NN* membutuhkan waktu *training* paling cepat, tetapi saat *testing* membutuhkan waktu yang paling lama. Waktu *training* tercepat kedua adalah model Naïve Bayes dengan waktu *testing*-nya juga peringkat kedua dengan selisih 0,184 detik dibandingkan model *Decision Tree* yang waktu *testing*-nya adalah 0,016 detik. Model Regresi Logistik membutuhkan waktu yang cukup lama saat *training*, yaitu sekitar 30,212 detik, diikuti dengan model *Neural Network* yang membutuhkan waktu paling lama sekitar 101,701 detik. Walaupun *training* membutuhkan waktu yang lama, Regresi Logistik dan *Neural Network* membutuhkan waktu yang cukup cepat saat *testing*, masuk ke peringkat ketiga dan keempat. Model *Random Forest* dapat dikatakan membutuhkan waktu yang cukup cepat, baik *training* maupun *testing*, dengan masing-masing waktunya adalah 2,893 detik (peringkat ketiga) dan 0,577 detik (peringkat kelima). Perlu dipahami bahwa hasil evaluasi berdasarkan waktu yang dibutuhkan untuk *training* dan *testing* model tergantung pada kemampuan dan spesifikasi *device.* Dalam penelitian ini, spesifikasi *device* dapat dilihat pada Tabel I.

TABEL I
SPESIFIKASI *DEVICE* YANG DIGUNAKAN

Processor	Intel Core i7-10510U CPU @ 1.80GHz 2.30 GHz
RAM	16 GB
OS	Windows 10
Tipe Sistem	64-bit OS

IV. SIMPULAN

Berdasarkan hasil penelitian terlihat bahwa penggunaan *software* Orange untuk *data mining* dapat mempercepat waktu karena pengguna hanya tinggal membuat *workflow* dan melakukan konfigurasi dengan klik tanpa membuat sintaks *Python*, walaupun *software* Orange memiliki *widget Python Script*. Pengguna dengan latar belakang non-IT, khususnya pada data penelitian di bidang kesehatan atau bidang lainnya, seperti pertanian, dsb., dapat melakukan *data mining* klasifikasi dengan *software* ini.

Selanjutnya terlihat bahwa waktu yang dibutuhkan untuk *training* dan *testing* model terbilang singkat. Penelitian ini menggunakan data ILDP yang bertujuan memprediksi diagnosis apakah seseorang menderita liver atau tidak. Dari tujuh metode yang dipilih, empat metode terbaik berdasarkan tingkat akurasi adalah Regresi Logistik, *Neural Network*, *Random Forest*, dan *Naïve Bayes*. Metode *Neural Network*, walaupun menghasilkan tingkat akurasi terbesar kedua, membutuhkan waktu *training* paling lama. Metode *Random Forest* menghasilkan tingkat akurasi terbesar ketiga dan membutuhkan waktu *training* lebih cepat, tetapi waktu *testing* lebih lama dibandingkan Regresi Logistik.

Penelitian ini tidak mengatur *seed* terlebih dahulu sehingga saat melakukan *running* dimungkinkan mendapat hasil yang berbeda. Oleh karena itu, penelitian ini dilakukan pengulangan hingga 100 kali. Untuk penelitian selanjutnya, dapat dilakukan pengaturan *seed* sehingga bisa diperoleh hasil yang seragam. Selain itu, dapat dilakukan penelitian lanjutan dengan menggunakan data terbaru dan menggunakan metode lainnya sehingga memungkinkan ditemukannya metode lain yang menghasilkan tingkat akurasi yang lebih baik.

DAFTAR REFERENSI

[1] R. Ratra dan P. Gulia, "Experimental evaluation of open source data mining tools (WEKA and Orange)," *Int. J. Eng. Trends Technol.*, vol. 68, no. 8, hlm. 30–35, 2020.

[2] M. PhridviRaj dan C. GuruRao, "Data mining - past, present and future - A Typical Survey on Data Stream," *INTER-ENG ProcediaTechnology*, vol. 12, hlm. 255–263, 2013.

[3] J. Han, M. Kamber, dan J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. USA: Morgan Kaufman Publisher, 2012.

[4] S. Shalev-Shwartz dan S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*. New York: Cambridge University Press, 2014.

[5] M. Roos, "A data analysis demonstrator for managing customer experience in a partnering ventur," Tesis, Faculty of Engineering, Stellenbosch University, 2019.

[6] T. Wendler dan S. Grottrup, *Data Mining with SPSS Modeler: TheORY, Exercises and Solution*, 2nd ed. Switzerland: Springer Nature Switzerland, 2021.

[7] A. Jose, M. Philip, L. T. Prasanna, dan M. Manjula, "Comparison of Probit and Logistic Regression models in the analysis of dichotomous outcomes," *Curr. Res. Biostat.*, vol. 10, no. 1, hlm. 1–19, 2020, doi: 10.3844/amjbsp.2020.1.19.

[8] F. Gorunescu, *Data Mining Concepts, Model and Techniques, Vol. 12*. Berlin: Springer, 2011.

[9] J. Han dan M. Kamber, *Data Mining: Concepts and Techniques Tutorial*. San Francisco: Morgan Kaufman Publisher, 2001.

[10] B. Zupan dan Dems, "Introduction to data mining," 2011. https://file.biolab.si/notes/2018-05-intro-to-datamining-notes.pdf. [10 Agustus 2021].

[11] Orange, "Orange Data Mining: Fruitful and Fun." [Daring]. Tersedia: https://orangedatamining.com/.

[12] J. Demsar dan B. Zupan, "Orange: data mining fruitful and fun," *Informatica*, vol. 37, hlm. 55–60, 2013.

[13] Orange Data Mining, "Orange Data Mining Library Documentation Release 3."

[14] UCI, "ILPD (Indian Liber Patient Dataset)." https://archive.ics.uci.edu/ml/datasets/ILPD+(Indian+Liver+Patient+Dataset) [20 Agustus 2021].

[15] P. Meilina, "Penerapan data mining dengan metode klasifikasi menggunakan *Decision Tree* dan Regresi," *J. Teknol. Univ. Muhammadiyah Jakarta*, vol. 7, no. 1, hlm. 11–20, 2015.

[16] S. Sansgiry, M. Bhosle, dan K. Sail, "Factors that affect academic performance among pharmacy students," *Am. J. Pharm. Educ.*, vol. 70, no. 5, artikel 104, 2006.

[17] H. Zhou, *Learn Data Mining Through Excel: A Step-by-Step Approach for Understanding Machine Learning Methods*. USA: Apress, 2020.

[18] D. L. Olson dan D. Wu, *Predictive Data Mining Models*, 2nd ed. Singapore: Springer, 2020.

[19] Larose dan T. Daniel, *Discovering Knowledge in Data: An Introduction to Data Mining*. USA: John Wiley & Sons, 2005.

[20] S. Dash, S. K. Pani, S. Balamurugan, dan A. Abraham, *Biomedical Data Mining for Information Retrieval: Methodologies, Techniques and Applications*. USA: Scrivener Publishing, 2021.

[21] D. M. W. Powers, "Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness and Correlation," Adelaide, 2007. [Daring]. Tersedia: http://arxiv.org/abs/2010.16061.

[22] A. Naik dan L. Samant, "Correlation review of classification algorithm using data mining tool: WEKA, Rapidminer, Tanagra, Orange, and Knime," *Procedia Comput. Sci.*, vol. 85, no. 2016, hlm. 662–668, 2016, doi: 10.1016/j.procs.2016.05.251.

Ira Rosianal Hikmah, kelahiran Jakarta. Menempuh pendidikan program sarjana di Program Studi Matematika FMIPA Universitas Indonesia dan melanjutkan program magister di Program Studi Statistika Institut Pertanian Bogor. Minat riset terkait bidang matematika, statistika, aktuaria, manajemen risiko, dan *machine learning*.

Ray Novita Yasa, kelahiran Wates. Menempuh pendidikan program sarjana di Program Studi Pendidikan Matematika STKIP MPL dan melanjutkan program magister di Program Studi Matematika Institut Teknologi Bandung. Minat riset terkait bidang matematika terapan, teori graf, *machine learning*, dan *deep learning*.