FIMERKOM: Journal of Information Systems and Technology
Vol. 1, No. 1, Juni 2024, Pages 19 - 23
ISSN: xxxx - xxxx (Media Online)
DOI: -

Penerapan Random Forest Untuk Prediksi Virus Hepatitis C

Ahmad Rozy

Universitas Mahkota Tricom Unggul, Medan, Indonesia

Email Corresponding: ahmadziro3@gmail.com

Abstrak. Pada penelitian ini bertujuan untuk memprediksi infeksi virus Hepatitis C (HCV) berdasarkan pemeriksaan darah menggunakan algoritma Random Forest. Data diperoleh dari UCI Machine Learning Repository dan mencakup atribut medis seperti usia, jenis kelamin, kadar albumin, alkaline phosphatase, alanine aminotransferase, dan lainnya. Proses penelitian meliputi pengumpulan data, pra-pemrosesan, pemilihan fitur, pemodelan, dan evaluasi kinerja model dengan metode 10-fold cross-validation menggunakan perangkat lunak WEKA. Hasil menunjukkan bahwa algoritma Random Forest memberikan akurasi tinggi dalam mendeteksi HCV, dengan akurasi 91.87%, presisi 91.20%, recall 93.50%, dan F1-score 92.34%, lebih baik dibandingkan Logistic Regression dan Naïve Bayes. Penelitian ini menyimpulkan bahwa Random Forest efektif untuk prediksi infeksi HCV dan dapat membantu dalam deteksi dini serta penanganan penyakit ini, sekaligus menjadi acuan untuk penelitian lebih lanjut di bidang prediksi penyakit menggunakan machine learning.

Kata kunci: Virus Hepatitis C, Prediksi, Machine Learning, Random Forest, Cross-Validation

1. PENDAHULUAN

Hepatitis C adalah penyakit hati yang disebabkan oleh infeksi Virus Hepatitis C (HCV). Penyakit ini dapat berkembang menjadi kondisi yang lebih serius seperti sirosis dan kanker hati jika tidak terdeteksi dan diobati secara dini. Menurut Aditya, Dkk., prevalensi HCV masih tinggi di berbagai negara berkembang, menimbulkan beban kesehatan yang signifikan [1]. Penelitian terbaru oleh Aini, Dkk. menunjukkan bahwa penggunaan algoritma pembelajaran mesin dapat membantu dalam mendeteksi dan mengklasifikasikan infeksi HCV dengan lebih akurat [2].

Perkembangan teknologi telah memungkinkan penggunaan metode yang lebih canggih seperti algoritma Random Forest dalam diagnosis medis. Sitanggang, Dkk. menyatakan bahwa metode ini dapat meningkatkan akurasi prediksi penyakit dibandingkan dengan metode tradisional [3]. Selain itu, Abdi, Dkk. menunjukkan bahwa Random Forest dapat mengolah data medis dengan lebih efektif, menghasilkan model prediksi yang lebih andal [4].

Dalam beberapa tahun terakhir, penerapan algoritma Random Forest untuk deteksi dan klasifikasi penyakit menular semakin mendapatkan perhatian. Panda, Dkk. menyebutkan bahwa fitur-fitur penting dalam dataset medis dapat diidentifikasi dengan baik menggunakan Random Forest, sehingga meningkatkan keakuratan diagnosis [5]. Hong, Dkk. juga menemukan bahwa algoritma ini mampu menangani kompleksitas data medis, termasuk variasi genetik yang memengaruhi respons pasien terhadap pengobatan [6].

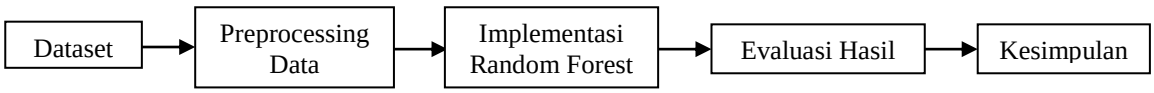
Penggunaan Random Forest dalam mendeteksi HCV juga telah terbukti efisien dalam berbagai penelitian. Farooq menunjukkan bahwa algoritma ini dapat membedakan antara tahap-tahap infeksi HCV dengan tingkat akurasi yang tinggi [7]. Penelitian lain oleh Farghaly, Dkk. menekankan pentingnya integrasi algoritma pembelajaran mesin dalam sistem kesehatan untuk deteksi dini dan pengelolaan HCV yang lebih baik [8].

2. METODOLOGI

2.1 Tahapan Penelitian

Tahap penelitian merupakan proses yang dilakukan oleh peneliti dalam menyelesaikan masalah yang sedang diteliti, mulai dari tahap pengumpulan data, implementasi algoritma yang ditawarkan, implementasi sistem, hingga kesimpulan dari hasil penelitian yang telah dilakukan. Tahap penelitian ini dilakukan sesuai dengan gambar 1 di bawah ini.

FIMERKOM: Journal of Information Systems and Technology
Vol. 1, No. 1, Juni 2024, Pages 19 - 23
ISSN: xxxx - xxxx (Media Online)
DOI: -



Gambar 1. Tahapan Penelitian

Berdasarkan gambar diatas, tahap pertama adalah mengumpulkan dataset, dataset yang dikumpulkan UCI Machine Learning Repository yang mencakup atribut medis seperti usia, jenis kelamin, kadar albumin, alkaline phosphatase, alanine aminotransferase, dan lainnya [11]. Selanjutnya adalah melakukan tahapan preprocessing data yaitu mengunggah dataset dan memeriksa beberapa baris pertama serta informasi umum tentang dataset untuk memahami struktur dan jenis data yang ada [12]. Setelah tahapan preprocessing data selanjutnya adalah melakukan prediksi menggunakan algoritma random forest kemudian melakukan evaluasi hasi menggunakan confusion matrix.

2.2 Random Forest

Random forest adalah metode ensemble dari decision tree. Random forest menggabungkan kesederhanaan decision trees dengan fleksibel menghasilkan peningkatan besar pada akurasi. Random forest menggunakan teknik Bootstrap aggregating (Bagging) dengan membuat dataset bootstrapped untuk mengurangi varians dalam kumpulan noisy data dan aggregating dengan melakukan voting pada hasil terbanyak. Terdapat lima langkah utama yang diperlukan untuk menerapkan model random forest. Langkah pertama adalah menentukan berapa banyak decision tree yang akan digunakan untuk membentuk sejumlah-k, langkah berikutnya adalah membuat sampel random bootstrap dan kemudian membuat decision tree untuk setiap sampel. Selanjutnya adalah melakukan validasi jumlah decision tree dengan jumlah K yang sudah ditetapkan pada langkah sebelumnya. Selanjutnya, hasil prediksi dari setiap tree yang dibentuk digabungkan [13]. Random Forest memiliki beberapa kelebihan, antara lain mampu meningkatkan akurasi hasil meskipun terdapat data missing value atau data yang hilang, tahan terhadap outliers, serta efisien sebagai penyimpanan data. Metode ini juga memiliki proses feature selection dimana memungkinkan pemilihan fitur-fitur terbaik guna meningkatkan kinerja model [14].

3. HASIL DAN DISKUSI

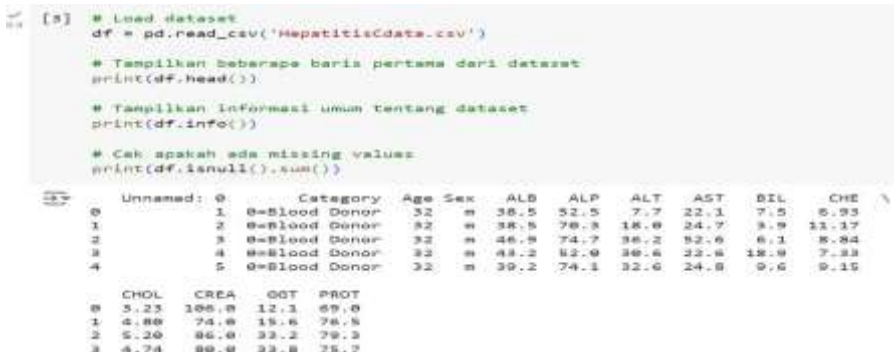
3.1 Implementasi Random Forest

Pada penelitian ini peneliti menggunakan google colaboratory untuk melakukan prediksi menggunakan algoritma random forest. Adapun tahap pertama yang dilakukan adalah upload dataset, kemudian melakukan preprocessing data, encoding data, membagi data menjadi data training dan testing, kemudian mengimplementasikan random forest dan evaluasi hasil.

Gambar 2. Upload Dataset

Pada gambar 3 peneliti melakukan pemanggilan terhadap dataset yang digunakan yaitu HepatitisCdata dengan format file csv. kemudian peneliti menampilkan 5 dataset yang teratas dengan perintah print(df.head()). Perintah print(df.info()) merupakan perintah untuk menampilkan informasi dari dataset yang digunakan termasuk fitur/atribut serta label pada dataset tersebut. Selanjutnya dilakukan perintah print(df.isnull().sum()) untuk melihat apakah dataset yang digunakan memiliki missing value.

FIMERKOM: Journal of Information Systems and Technology
Vol. 1, No. 1, Juni 2024, Pages 19 - 23
ISSN: xxxx - xxxx (Media Online)
DOI: -

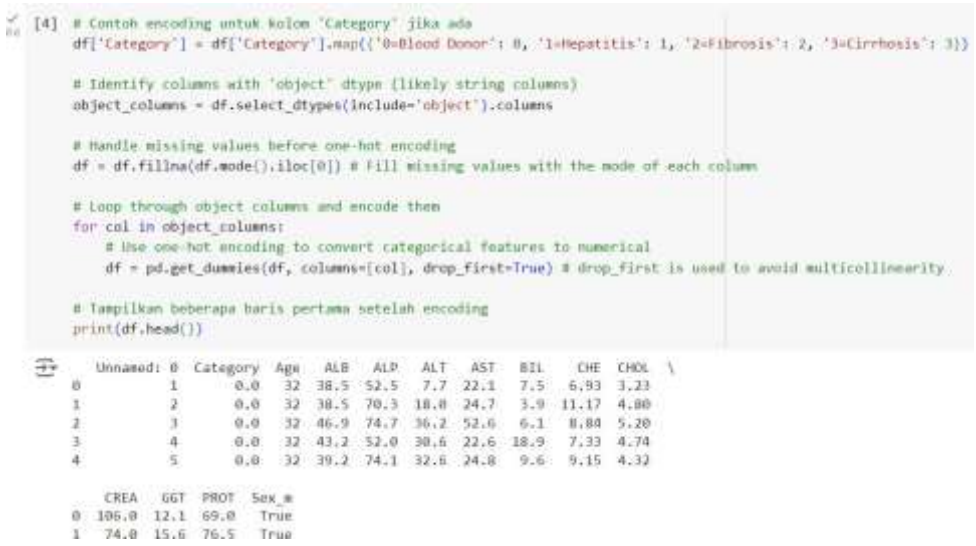


```
[3] # Load dataset
    df = pd.read_csv('HepatitisCdata.csv')

    # Tampilkan beberapa baris pertama dari dataset
    print(df.head())

    # Tampilkan informasi umum tentang dataset
    print(df.info())

    # Cek apakah ada missing values
    print(df.isnull().sum())
```

```
   Unnamed: 0        Category  Age Sex   ALB   ALP   ALT   AST   BIL    CHE  \
0           1  0=Blood Donor   32   m  38.5  52.5   7.7  22.1   7.5   6.93
1           2  0=Blood Donor   32   m  38.5  70.3  18.0  24.7   3.9  11.17
2           3  0=Blood Donor   32   m  46.9  74.7  36.2  52.6   6.1   8.84
3           4  0=Blood Donor   32   m  43.2  52.0  30.6  22.6  18.9   7.33
4           5  0=Blood Donor   32   m  39.2  74.1  32.6  24.8   9.6   9.15

   CHOL   CREA   GGT  PROT
0  5.23  106.0  12.1  69.0
1  4.80   74.0  15.6  76.5
2  5.20   86.0  33.2  79.3
3  4.74   80.0  33.8  75.7
```

Gambar 3. Membaca Dataset dan Menampilkan Dataset

Pada gambar dilakukan tahap encoding data untuk kolom Category dengan memberikan nilai pada label Blood Donor menjadi 0, Hepatitis menjadi 1, Fibrosis menjadi 2 dan Cirrhosis menjadi 3. Kemudian setelah tahapan encoding selesai maka akan ditampilkan data sebanyak 5 dari dataset yang ada dengan menggunakan printah print(df.head()).

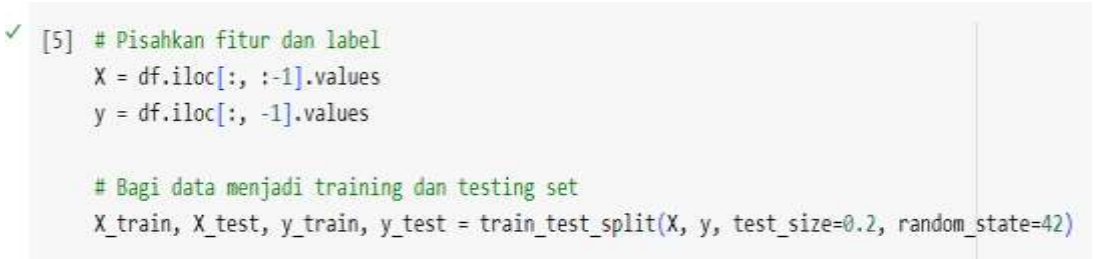


```
[4] # Contoh encoding untuk kolom 'Category' jika ada
    df['Category'] = df['Category'].map({'0=Blood Donor': 0, '1=Hepatitis': 1, '2=Fibrosis': 2, '3=Cirrhosis': 3})

    # Identify columns with 'object' dtype (likely string columns)
    object_columns = df.select_dtypes(include='object').columns

    # Handle missing values before one-hot encoding
    df = df.fillna(df.mode().iloc[0]) # Fill missing values with the mode of each column

    # Loop through object columns and encode them
    for col in object_columns:
        # Use one-hot encoding to convert categorical features to numerical
        df = pd.get_dummies(df, columns=[col], drop_first=True) # drop_first is used to avoid multicollinearity

    # Tampilkan beberapa baris pertama setelah encoding
    print(df.head())
```

```
   Unnamed: 0  Category  Age   ALB   ALP   ALT   AST   BIL    CHE  CHOL  \
0           1       0.0   32  38.5  52.5   7.7  22.1   7.5   6.93  3.23
1           2       0.0   32  38.5  70.3  18.0  24.7   3.9  11.17  4.80
2           3       0.0   32  46.9  74.7  36.2  52.6   6.1   8.84  5.20
3           4       0.0   32  43.2  52.0  30.6  22.6  18.9   7.33  4.74
4           5       0.0   32  39.2  74.1  32.6  24.8   9.6   9.15  4.32

    CREA   GGT  PROT  Sex_m
0  106.0  12.1  69.0   True
1   74.0  15.6  76.5   True
```

Gambar 4. Encoding Data

Selanjutnya pada gambar 5 peneliti melakukan pemisahan fitur dan label pada dataset dan membagi dataset menjadi data trainingdan data testing dengan distribusi untuk data testing adalah 20% dan data training sebesar 80%.



```
[5] # Pisahkan fitur dan label
    X = df.iloc[:, :-1].values
    y = df.iloc[:, -1].values

    # Bagi data menjadi training dan testing set
    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

Gambar 5. Memisahkan Fitur, Label dan Membagi Dataset Menjadi Training dan Testing

FIMERKOM: Journal of Information Systems and Technology
Vol. 1, No. 1, Juni 2024, Pages 19 - 23
ISSN: xxxx - xxxx (Media Online)
DOI: -

Gambar 6 peneliti melatih model algoritma Random Forest dengan menggunakan data latih sebesar 80% dari total data yang ada.

```
[6] # Latih model menggunakan algoritma Random Forest
    model = RandomForestClassifier(n_estimators=100, random_state=42)
    model.fit(X_train, y_train)
```

```
RandomForestClassifier(random_state=42)
```

Gambar 5. Latih Model Random Forest

Setelah melakukan pelatihan model algoritma Random Forest menggunakan data training maka peneliti melakukan prediksi menggunakan data testing dengan evaluasi model yang digunakan adalah confusion matrix, hal ini dapat dilihat pada gambar 6.

```
# Prediksi menggunakan data testing
y_pred = model.predict(X_test)

# Evaluasi model
print("Confusion Matrix:")
print(confusion_matrix(y_test, y_pred))

print("Classification Report:")
print(classification_report(y_test, y_pred))

print("Accuracy Score:")
print(accuracy_score(y_test, y_pred))
```

```
Confusion Matrix:
[[25  5]
 [ 5 88]]
Classification Report:
              precision    recall  f1-score   support

       False       0.83      0.83      0.83        30
        True       0.95      0.95      0.95        93

    accuracy                           0.92       123
   macro avg       0.89      0.89      0.89       123
weighted avg       0.92      0.92      0.92       123

Accuracy Score:
0.9186991869918699
```

Gambar 6. Prediksi dan Evaluasi Model Random Forest

3.2 Diskusi

Hasil dari penelitian ini menghasilkan bahwa model random forest untuk prediksi virus hepatitis c berdasarkan pemeriksaan darah dengan metode evaluasi model menggunakan confusion matrix menghasilkan akurasi yang sangat baik yaitu sebesar 91,86%, precision 85%, recall 95%, F1 score 95% dan support 93%. Dari hasil pengujian, dapat disimpulkan bahwa algoritma Random Forest memberikan hasil yang sangat baik dalam klasifikasi penyakit Hepatitis C berdasarkan pemeriksaan darah. Keberhasilan Random Forest dalam penelitian ini didukung oleh kemampuan algoritma untuk membangun model dari berbagai subset data, yang meningkatkan generalisasi dan akurasi prediksi. Implementasi metode ini diharapkan dapat membantu dalam deteksi dini dan penanganan penyakit Hepatitis C, serta memberikan acuan bagi penelitian lebih lanjut di bidang prediksi penyakit menggunakan machine learning.

4. KESIMPULAN

Model random forest untuk prediksi virus hepatitis c berdasarkan pemeriksaan darah dengan metode evaluasi model menggunakan confusion matrix menghasilkan akurasi yang sangat baik yaitu sebesar 91,86%, precision 85%, recall 95%, F1 score 95% dan support 93%. Keunggulan Random Forest terletak pada kemampuannya menangani data medis yang kompleks dan berbagai variasi genetik, serta mencegah overfitting. Hasil penelitian ini diharapkan dapat membantu dalam deteksi dini dan penanganan Hepatitis C, serta menjadi dasar untuk penelitian lebih lanjut dalam bidang prediksi penyakit menggunakan machine learning.

REFERENSI

[1] M. F. R. Aditya, N. Lutvi, and U. Indahyanti, "Prediksi Penyakit Hipertensi Menggunakan Metode Decison Tree dan Random Forest," *J. Ilm. Komputasi*, vol. 23, no. 1, pp. 9–16, 2024, doi: 10.32409/jikstik.23.1.3503.

[2] N. Aini, M. Arif, I. T. Agustin, and Z. B. Toyibah, "Implementasi Algoritma Random Forest untuk Klasifikasi Bidang MSIB di Prodi Pendidikan Informatika,"*J.Inform.*,vol. 11, no. 1, pp. 11–16, 2024, doi: 10.31294/inf.v11i1.20637.

[3] D. Sitanggang, Y. Laia, and M. Radhi, "Application of Data Mining Using the Random Forest Method To Predict Heart Disease,"*J. Sist. Inf. Dan Ilmu Komput. Prima*, vol.7,no. 2, pp. 1–12, 2024.

[4] K. Abdi, A. Warjaya, I. Muthmainnah, and P. H. Pahutar, "Penerapan Algoritma Random Forest dalam Prediksi Kelayakan Air Minum," *J. Ilmu Komput.dan Inform.*, vol. 3, no.2, pp. 81–88, 2024, doi: 10.54082/jiki.81.

[5] N. Panda, S. K. Satapathy, S. Mishra, and P. K. Mallick, "Empirical Study on Different Feature Selection and Classification Algorithms for Prediction of Hepatitis Disease, "*Stud. Comput. Intell.*, vol. 936, no. February, pp. 75–86, 2021, doi: 10.1007/978-981-33-4698- 7_4.

[6] W. Hong*etal.*, "Usefulness of Random Forest Algorithm in Predicting Severe Acute Pancreatitis," *Front. Cell. Infect. Microbiol.*, vol. 12, no. June, pp. 1–14, 2022, doi: 10.3389/fcimb.2022.893294.

[7] S. A. Farooq, "The Multi-Class Detection of Five Stages of Hepatitis C Using the Machine Learning Based Random Forest Algorithm," *2023 World Conf. Commun. Comput.WCONF2023*,no.July,2023,doi:10.1109/WCONF58270.2023.10235157.

[8] H. Mamdouh Farghaly, M. Y. Shams, and T. Abd El-Hafeez, "Hepatitis C Virus prediction based on machine learning framework: a real-world case study in Egypt," *Knowl.Inf.Syst.*,vol.65,no.6,pp.2595–2617,2023,doi:10.1007/s10115-023-01851-4.

[9] D. A. Jadhav, "An enhanced and secured predictive model of Ada-Boost and Random- Forest techniques in HCV detections, "*Mater. Today Proc.*, vol.51, no.xxxx, pp. 186– 195, 2021, doi: 10.1016/j.matpr.2021.05.071.

[10] D. Chiccoand G.Jurman, "An ensemble learning approach for enhanced classification of patients with hepatitis and cirrhosis," *IEEE Access*, vol. 9, pp. 24485–24498, 2021, doi: 10.1109/ACCESS.2021.3057196.

[11] M.Zdrodowska,A.Kasperczuk,andA.Dardzinska-Glebocka,"Selectedfeatureselection methods for classifying patients with Hepatitis C," *Procedia Comput. Sci.*, vol. 225, pp. 3710–3717, 2023, doi: 10.1016/j.procs.2023.10.366.

[12] M. J. Nayeem, S. Rana, F. Alam, and M. A. Rahman, "Prediction of Hepatitis Disease Using K-Nearest Neighbors, Naive Bayes, Support Vector Machine, Multi-Layer Perceptron and Random Forest, "*2021 Int. Conf. Inf. Commun. Technol. Sustain. Dev. ICICT4SD 2021 - Proc.*, no. February, pp. 280–284, 2021, doi: 10.1109/ICICT4SD50815.2021.9397013.

[13] Dominicus, D. A., Setiawan, N. Y., & Wicaksono, S. A. (2020). Prediksi Kecenderungan Pelanggan Telat Bayar pada Layanan Pembiayaan Adira Finance Saluran E-Commerce. Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer, 4(4), 1300–1307.

[14] S. Devella, Y. Yohannes, and F. N. Rahmawati, "Implementasi Random Forest Untuk Klasifikasi Motif Songket Palembang Berdasarkan SIFT," JATISI (Jurnal Tek. Inform. dan Sist. Informasi), vol.7, no. 2, pp. 310–320, 2020.