PERBANDINGAN KINERJA ALGORITMA RANDOM FLOREST CLASSIFIER DAN LIGHTGBM CLASSIFIER UNTUK PREDIKSI PENYAKIT JANTUNG

Filbert Duran[1)*], Frederico Wijaya[2)], Yakin Rianto Hulu[3)], Mawaddah Harahap[4)], Agung Prabowo[5)]
filbertduran140803@gmail.com[1)*], fredericowijaya45@gmail.com[2)], yakin22hulu@gmail.com[3)], mawaddah@unprimdn.ac.id[4)], agung.prabowo2610@gmail.com[5)]

[1,2,3,4,5)]Universitas Prima Indonesia

Received: 10 Maret 2024 Accepted: 30 April 2024 Published: 5 Mei 2024

*filbertduran140803@gmail.com

Kata Kunci: algoritma random florest classifier, lightgbm classifier, pneyakit jantung, prediksi

Abstrak : Penyakit jantung merupakan masalah kesehatan serius yang dapat dicegah dan diobati. Dengan menjaga gaya hidup sehat, melakukan pemeriksaan kesehatan secara rutin, dan mengikuti anjuran dokter[1], risiko penyakit jantung dapat dikurangi. Random Forest Classifier (RFC) bagaikan hutan pohon keputusan yang bekerja sama untuk menghasilkan prediksi yang lebih jitu. Algoritma ini tergolong handal dan fleksibel, mampu menangani berbagai tugas klasifikasi dan regresi. Kelebihannya, RFC menawarkan akurasi tinggi, tahan terhadap overfitting, dan mudah diinterpretasikan[2]. RFC adalah algoritma machine learning yang kuat dengan banyak keunggulan, namun perlu dipertimbangkan pula keterbatasannya dalam hal komputasi dan fleksibilitas[3]. LightGBM merupakan algoritma machine learning yang kuat dan efisien untuk klasifikasi dan regresi. Kecepatan, akurasi, dan kemudahan penggunaannya menjadikannya pilihan yang menarik untuk berbagai aplikasi[4]. Dari hasil yang didapat dari penelitian ini adalah metode RFC dan LightGBM dapat disimpulkan bahwa metode RFC merupakan metode yang tergolong efektif dalam analisis penyakit jantung dengan akurasi prediksi dari model adalah 95,37%., dapat dikatakan bahwa metode Random Florest Classifier cocok untuk melakukan analisis penyakit jantung bedasarkan dataset yang ada.

PENDAHULUAN

Random Forest Classifier (RFC) bagaikan hutan pohon keputusan yang bekerja sama untuk menghasilkan prediksi yang lebih jitu. Algoritma ini tergolong handal dan fleksibel, mampu menangani berbagai tugas klasifikasi dan regresi. Kelebihannya, RFC menawarkan akurasi tinggi, tahan terhadap overfitting, dan mudah diinterpretasikan. Namun, bagaikan pohon yang membutuhkan waktu untuk tumbuh, membangun hutan RFC bisa memakan waktu lama, terutama pada dataset besar. Selain itu, fleksibilitasnya terbatas, sehingga kurang optimal untuk tugas-tugas tertentu seperti klasifikasi multi-kelas[5]. LightGBM Classifier (LightGBM) bagaikan jagoan klasifikasi yang tak tertandingi. Algoritma ini terkenal dengan kecepatannya yang luar biasa dan hasil prediksi yang jitu, menjadikannya pilihan utama untuk berbagai tugas klasifikasi. Rahasia di balik kekuatan LightGBM terletak pada algoritma Gradient Boosting Decision Tree (GBDT) yang dioptimalkan untuk meningkatkan efisiensi dan skalabilitas. Mirip dengan Random Forest Classifier, LightGBM membangun hutan pohon keputusan untuk menghasilkan prediksi yang lebih akurat dan stabil. LightGBM merupakan algoritma machine learning yang kuat dan efisien untuk klasifikasi dan regresi. Kecepatan, akurasi, dan kemudahan penggunaannya menjadikannya pilihan yang menarik untuk berbagai aplikasi[6].

TINJAUAN LITERATUR

Penelitian terkait dengan prediksi penyakit jantung telah dilakukan pada penelitian sebelumnya, pada penelitian sebelumnya mengunakan berbagai macam jenis Algoritma Machine Learning, peneltian yang mengunakan jenis Algoritma klasifikasi pada machine learning seperti Random Forest, Support Vector Machines, Gradient Boosting Machines, XGBOOST, Light GBM telah dilakukan oleh Wahyu Nugraha. Pada penelitian yang telah dilakukan mendapatkan hasil sebuah aplikasi website untuk memprediksi penyakit Jantung Kardiovaskular. Berdasarkan penelitiannya yang dilakukan menggunakan dataset Kaggle dataset repository (Heart Failure Prediction) dapat disimpulkan bahwa model klasifikasi menggunakan XGBOOST rata-rata memperoleh nilai tertinggi baik menggunakan pengukuran accuracy, F1 score maupun AUC. Model klasifikasi SVM memperoleh nilai rata-rata hampir mirip dengan model Gradient Boosting. Sedangkan untuk model pengukuran dengan nilai terendah diperoleh dengan model klasifikasi LightGBM.[7]

METODE PENELITIAN

1. Analisis dan Pembahasan Riset Penelitian
    a. Input data

    Sumber data set yang dipakai diambil dari kaggle.com, yang berisi data umur, gender, max heart rate, dll. Data-data tersebut kita perlukan untuk melakukan penelitian ini.

Out[3]:

	age	sex	chest pain type	resting bp s	cholesterol	fasting blood sugar	resting ecg	max heart rate	exercise angina	oldpeak	ST slope	target
0	40	1	2	140	289	0	0	172	0	0.0	1	0
1	49	0	3	160	180	0	0	156	0	1.0	2	1
2	37	1	2	130	283	0	1	98	0	0.0	1	0
3	48	0	4	138	214	0	0	108	1	1.5	2	1
4	54	1	3	150	195	0	0	122	0	0.0	1	0
...	...	...	...	...	...	...	...	...	...	...	...	...
1185	45	1	1	110	264	0	0	132	0	1.2	2	1
1186	68	1	4	144	193	1	0	141	0	3.4	2	1
1187	57	1	4	130	131	0	0	115	1	1.2	2	1
1188	57	0	2	130	236	0	2	174	0	0.0	2	1
1189	38	1	3	138	175	0	0	173	0	0.0	1	0

1190 rows × 12 columns

Gambar 1 Melakukan Input data

    b. Membersihkan data

    Proses ini berguna untuk mengidentifikasi dan memperbaiki kesalahan yang timbul saat mengolah data dari berbagai sumber. Dengan menghapus ketidakkonsistenan dan ketidakakuratan, data cleaning dapat memastikan bahwa suatu kumpulan data dapat dipercaya. Jadi data yang tidak penting atau kosong dihapus pada tahap ini.

Out[5]:

	count	mean	std	min	25%	50%	75%	max
age	1190.0	53.720168	9.358203	28.0	47.0	54.0	60.00	77.0
sex	1190.0	0.763866	0.424884	0.0	1.0	1.0	1.00	1.0
chest pain type	1190.0	3.232773	0.935480	1.0	3.0	4.0	4.00	4.0
resting bp s	1190.0	132.153782	18.368823	0.0	120.0	130.0	140.00	200.0
cholesterol	1190.0	210.363866	101.420489	0.0	188.0	229.0	269.75	603.0
fasting blood sugar	1190.0	0.213445	0.409912	0.0	0.0	0.0	0.00	1.0
resting ecg	1190.0	0.698319	0.870359	0.0	0.0	0.0	2.00	2.0
max heart rate	1190.0	139.732773	25.517636	60.0	121.0	140.5	160.00	202.0
exercise angina	1190.0	0.387395	0.487360	0.0	0.0	0.0	1.00	1.0
oldpeak	1190.0	0.922773	1.086337	-2.6	0.0	0.6	1.60	6.2
ST slope	1190.0	1.624370	0.610459	0.0	1.0	2.0	2.00	3.0
target	1190.0	0.528571	0.499393	0.0	0.0	1.0	1.00	1.0

Gambar 2 cleaning data(1)

    c. Comparing model

    Untuk memilih model pembelajaran mesin terbaik untuk kumpulan data tertentu, penting untuk mempertimbangkan fitur atau parameter model . Parameter dan tujuan model membantu mengukur fleksibilitas model, asumsi, dan gaya belajar.

Out[6]:

	Model	Accuracy
0	Logistic Regression	0.861345
1	Decision Tree Classifier	0.915966
2	Random Forest Classifier	0.941176
3	Support Vector Classifier	0.890756
4	K-Nearest Neighbors Classifier	0.886555
5	XGBoost Classifier	0.928571
6	LightGBM Classifier	0.945378

Gambar 3 Comparing model

d. Melakukan train dan validation data

Training dataset adalah himpunan data yang digunakan untuk melatih atau membangun model. Kemudian, *validation dataset* adalah himpunan data yang digunakan untuk mengoptimasi saat melatih model. Model dilatih menggunakan *training dataset*, kemudian kinerja saat latihan tersebut diuji menggunakan *validation dataset*. Hal ini bertujuan untuk melihat kemampuan model pada saat *training* apakah dapat mengenal pola secara umum. *Validation dataset* juga dapat digunakan untuk melihat akurasi dari model yang dibuat

Data Scaling and Grid Search

In [7]:
```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)

rf = RandomForestClassifier()

param_grid = {
    'n_estimators': [200, 300, 400],
    'max_depth': [40, 45, 50],
    'min_samples_split': [2, 5, 10]
}


grid_search = GridSearchCV(estimator=rf, param_grid=param_grid, cv=5, scoring='accuracy')
grid_search.fit(X_train_scaled, y_train)


print("Best parameters:", grid_search.best_params_)
print("Best score:", grid_search.best_score_)
```

```
Best parameters: {'max_depth': 50, 'min_samples_split': 2, 'n_estimators': 200}
Best score: 0.9107026729126482
```

Gambar 4 Scaling data

Modeling

In [8]:
```
best_model = grid_search.best_estimator_
best_model.fit(X_train_scaled, y_train)
```

Out[8]:
```
RandomForestClassifier(max_depth=50, n_estimators=200)
```
In a Jupyter environment, please rerun this cell to show the HTML representation or trust the notebook.
On GitHub, the HTML representation is unable to render, please try loading this page with nbviewer.org.

Gambar 5 Visualisasi data

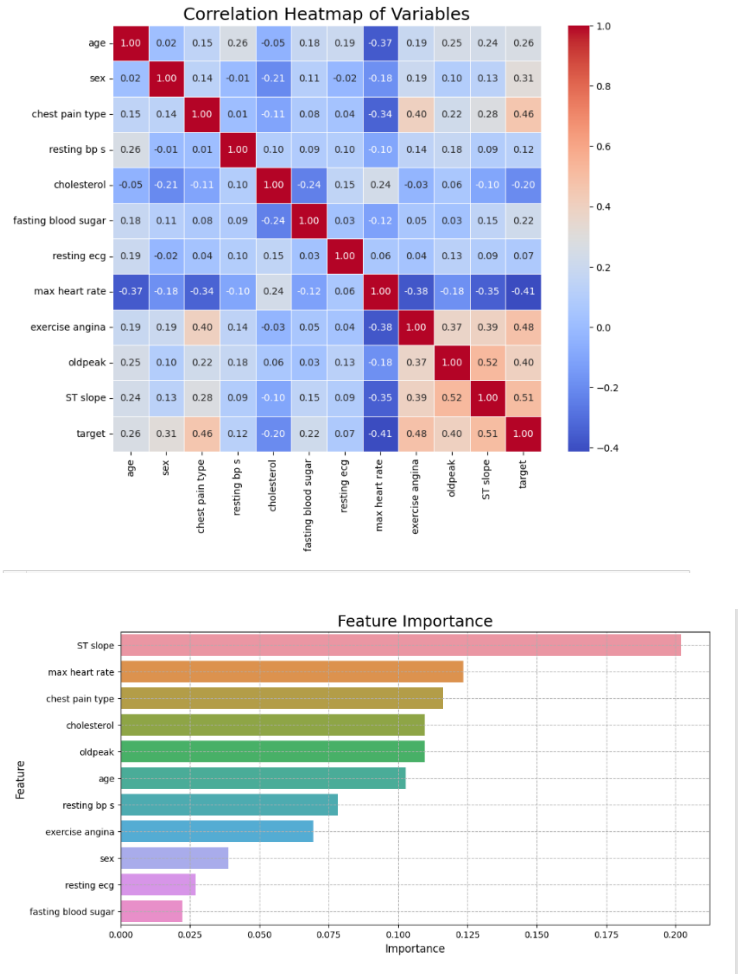
Memilih model RFC sebagai model utama yang akan digunakan bedasarkan hasil akurasi terbaik yang didapatkan

e. Menampilkan dan menguji akurasi dari analisis yang dilakukan

Out[10]:

	precision	recall	f1-score	support
0	0.961538	0.934579	0.947867	107.000000
1	0.947761	0.969466	0.958491	131.000000
accuracy	0.953782	0.953782	0.953782	0.953782
macro avg	0.954650	0.952023	0.953179	238.000000
weighted avg	0.953955	0.953782	0.953715	238.000000

Gambar 6 Menguji hasil olahan data



Gambar 7 Mengvisualisasikan hasil uji olahan data

HASIL PENELITIAN

Dari hasil yang didapat dari penelitian ini adalah metode RFC dan LightGBM dapat disimpulkan bahwa metode RFC merupakan metode yang tergolong efektif dalam analisis penyakit jantung dengan akurasi prediksi dari model adalah 95,37%. Riset ini juga memberi pengalaman yang cukup banyak kepada Saya dalam membuat riset menggunakan machine learing, mengajarkan Saya tahapan-tahapan dalam pengolahan data, dan meningkatkan kemampuan analisis. Kegiatan riset ini juga melatih Saya dalam menggunakan kedua metode tersebut untuk melakukan analisis mendalam dalam suatu dataset.

PEMBAHASAN

Penelitian ini membangun sebuah model untuk melakukan prediksi produksi tanaman padi. Pada proses sebelumnya telah dilakukan proses rangkaian untuk pembuatan model RFC (Random Forest Classifier) dan LightGBM[8]. Random Forest Classifier adalah algoritma machine learning yang kuat dengan banyak keunggulan, namun perlu dipertimbangkan pula keterbatasannya dalam hal komputasi dan fleksibilitas. Dan LightGBM adalah algoritma machine learning yang kuat dan efisien untuk klasifikasi dan regresi. Kecepatan, akurasi, dan kemudahan penggunaannya menjadikannya pilihan yang menarik untuk berbagai aplikasi[9]. Berdasarkan hasil kerja dan perbandingan antara kedua metode tersebut bisa disimpulkan hasil penelitian ini dimenangkan oleh Random Florest Classifier Bedasarkan pengujian yang telah dilakukan pada dataset tersebut, penelitian ini mendapatkan hasil akurasi 95,37% yang dapat dikatakan bahwa metode Random Florest Classifier cocok untuk melakukan analisis penyakit jantung bedasarkan dataset yang ada.

KESIMPULAN

Penyakit jantung merupakan masalah kesehatan serius yang dapat dicegah dan diobati. Dengan menjaga gaya hidup sehat, melakukan pemeriksaan kesehatan secara rutin, dan mengikuti anjuran dokter, risiko penyakit jantung dapat dikurangi[7]. Random Forest Classifier (RFC) bagaikan hutan pohon keputusan yang bekerja sama untuk menghasilkan prediksi yang lebih jitu. Algoritma ini tergolong handal dan fleksibel, mampu menangani berbagai tugas klasifikasi dan regresi. Kelebihannya, RFC menawarkan akurasi tinggi, tahan terhadap overfitting, dan mudah diinterpretasikan[10]. RFC adalah algoritma machine learning yang kuat dengan banyak keunggulan, namun perlu dipertimbangkan pula keterbatasannya dalam hal komputasi dan fleksibilitas. LightGBM merupakan algoritma machine learning yang kuat dan efisien untuk klasifikasi dan regresi. Kecepatan, akurasi, dan kemudahan penggunaannya menjadikannya pilihan yang menarik untuk berbagai aplikasi[11]. Bedasarkan pengujian yang telah dilakukan pada dataset tersebut, penelitian ini mendapatkan hasil akurasi 95,37% yang dapat dikatakan bahwa metode Random Florest Classifier cocok untuk melakukan analisis penyakit jantung bedasarkan dataset yang ada[12]. Untuk penelitian selanjutnya dapatlebih di kembangkan lagi untuk data yang akan digunakan agar model yang dibuat lebih bagus.

REFERENCES

[1] A. A. Mahottama, "PREVALENSI HIPERTENSI PADA PENDERITA PENYAKIT JANTUNG KORONER (PJK) DI RSUP SANGLAH DENPASAR MARET – SEPTEMBER 2019 FAKULTAS KEDOKTERAN UNIVERSITAS UDAYANA." Accessed: Apr. 26, 2024. [Online]. Available: https://ojs.unud.ac.id/index.php/eum/article/view/72265

[2] X. Wang, E. R. Andrinopoulou, K. M. Veen, A. J. J. C. Bogers, and J. J. M. Takkenberg, "Statistical primer: An introduction to the application of linear mixed-effects models in cardiothoracic surgery outcomes research - a case study using homograft pulmonary valve replacement data," *European Journal of Cardio-thoracic Surgery*, vol. 62, no. 4, Oct. 2022, doi: 10.1093/ejcts/ezac429.

[3] A. Liaw and M. Wiener, "Classification and Regression by randomForest," 2002. [Online]. Available: http://www.stat.berkeley.edu/

[4] G. Ke *et al.*, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree." [Online]. Available: https://github.com/Microsoft/LightGBM.

[5] yanli Liu, yourong Wang, and jian Zhang, "New Machine Learning Algorithm: Random Forest." Accessed: Apr. 26, 2024. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-642-34062-8_32

[6] R. Kanth, "How Does Random Forest Work." Accessed: Apr. 26, 2024. [Online]. Available: https://www.analyticsvidhya.com/blog/2023/02/how-does-random-forest-work/

[7] W. Nugraha, "PREDIKSI PENYAKIT JANTUNG CARDIOVASCULAR MENGGUNAKAN MODEL ALGORITMA KLASIFIKASI." [Online]. Available: https://www.researchgate.net/publication/358561013

[8] J. Sanchez, *Random Forest Algorithm And Parameters by Jason Sanchez*, (2015). Accessed: Apr. 26, 2024. [Online Video]. Available: https://www.youtube.com/watch?v=GofidHdilas

[9] S. Raharjo, "Cara Mengatasi Soal Angket yang Tidak Valid." Accessed: Apr. 26, 2024. [Online]. Available: https://www.konsistensi.com/2014/03/mengatasi-angkettidak-valid.html

[10] G. G. Kurniawati, "Machine Learning Python: Kenali Dua Library Python Terbaik, PyTorch vs TensorFlow." Accessed: Apr. 26, 2024. [Online]. Available: https://dqlab.id/belajar-machine-learning-python-bersama-dqlab-550

[11] S. Karray, "LightGBM: an Effective Decision Tree Gradient Boosting Method to Predict Customer Loyalty in the Finance Industry." Accessed: Apr. 26, 2024. [Online]. Available: https://ieeexplore.ieee.org/document/8845529

[12] B. So, "Enhanced Gradient Boosting for Zero-Inflated Insurance Claims and Comparative Analysis of CatBoost, XGBoost, and LightGBM," Jul. 2023, [Online]. Available: http://arxiv.org/abs/2307.07771