ANTIVIRUS: Jurnal Ilmiah Teknik Informatika (p – ISSN: 1978 – 5232; e – ISSN: 2527 – 337X)
Vol. 17 No. 2 Mei 2023, pp. 278 – 285

EVALUASI KOMPARATIF METODE *MACHINE LEARNING* UNTUK MEMPREDIKSI PERUBAHAN HARGA SAHAM

Diterima Redaksi: 18 Juni 2023; Revisi Akhir:4 Mei 2024; Diterbitkan Online:14 Mei 2024

Galih Adhi Putratama[1)], Satya Maulana Fahreza[2)], Yudhistira Rakha Ramandhani[3)]
[1, 2, 3)] Program Studi Informatika, Fakultas Teknologi Informasi dan Sains Data, Universitas Sebelas Maret
[1, 2, 3)] Jl. Ir Sutami No.36, Kentingan, Kec. Jebres, Kota Surakarta, Jawa Tengah, kode pos : 57126
e-mail: galihadhip@student.uns.ac.id[1)], fahryreza13@student.uns.ac.id[2)], rakharamandhani@student.uns.ac.id[3)]

***Abstrak*:** *Memprediksi tren harga di pasar saham adalah tantangan yang kompleks karena banyak faktor ketidakpastian dan variabel yang mempengaruhi nilai pasar. Studi ini melakukan evaluasi komparatif terhadap tiga metode machine learning populer, yaitu Random Forest, K-Nearest Neighbors (KNN), dan XGBoost, untuk memprediksi perubahan harga saham. Hasil penelitian menunjukkan bahwa Random Forest memiliki nilai ROC yang paling tinggi, yaitu 0.98, sedangkan XGBoost memiliki performa yang lebih baik dalam hal akurasi, recall, dan presisi berturut-turut yaitu 0.93, 0.69, 0.77. Metode Windowing juga diterapkan pada dataset untuk mengatasi permasalahan overfitting. Penelitian ini memberikan wawasan penting bagi praktisi dan peneliti di bidang prediksi harga saham untuk memilih model terbaik berdasarkan matriks evaluasi yang lebih diutamakan.*

Kata Kunci— KNN, Prediksi, Random Forest, Windowing, XGBoost

***Abstract*:** *Forecasting price patterns in the stock market poses a complicated and intricate task due to numerous uncertain factors and variables that influence market value. This study conducts a comparative evaluation of three popular computational learning approaches, namely Random Forest, K-Nearest Neighbors (KNN), and XGBoost, for predicting stock price changes. The research findings indicate that Random Forest achieves highest ROC scores, which is 0.98, while XGBoost exhibits superior performance in relation to accuracy, recall, and precision of 0.93, 0.69, 0.77, respectively. The Windowing method is also applied to the dataset to address overfitting issues. This study offers valuable knowledge for professionals and researchers in the domain of stock price prediction, enabling them to choose the optimal model based on preferred evaluation metrics.*

Keywords— KNN, Random Forest, Stock price, XGBoost, Windowing

I. PENDAHULUAN

MEMPREDIKSI tren harga di pasar saham adalah tantangan yang kompleks karena banyak faktor ketidakpastian dan variabel yang mempengaruhi nilai pasar setiap harinya, seperti kondisi ekonomi, sentimen investor terhadap perusahaan tertentu, peristiwa politik, dan lain sebagainya [1]. Pasar saham rentan terhadap perubahan yang cepat, yang menyebabkan fluktuasi acak dalam harga saham [2]. Secara umum, pasar saham bersifat dinamis, non-parametrik, dan tidak teratur [3]. Oleh karena itu, pergerakan harga saham sering dianggap sebagai proses acak dengan fluktuasi yang lebih jelas dalam jangka pendek. Namun, beberapa saham cenderung mengikuti trend linier dalam jangka waktu yang lebih panjang. Investasi di pasar saham memiliki risiko tinggi karena sifat kacau dan tidak stabilnya perilaku saham [4]. Untuk mengurangi risiko tersebut, pengetahuan mendalam tentang pergerakan harga saham di masa depan sangat penting. Para *trader* cenderung membeli saham yang diprediksi akan meningkat nilainya di masa depan, sementara mereka cenderung menghindari saham yang diprediksi akan turun nilainya [1]. Oleh karena itu, prediksi tren harga pasar saham yang akurat menjadi krusial untuk memaksimalkan keuntungan dan meminimalkan kerugian. Terdapat beberapa metodologi yang digunakan untuk memprediksi perilaku harga saham, termasuk analisis teknis, peramalan deret waktu,

DOI: https://doi.org/10.35457/antivirus.v17i2.2871

pembelajaran mesin, penambangan data, serta pemodelan dan prediksi volatilitas saham menggunakan persamaan diferensial.

Pada penelitian [5], dilakukan percobaan untuk memprediksi pergerakan harga saham menggunakan metode Random Forest. Akan tetapi, terdapat sebuah masalah pada penelitian tersebut yaitu terjadi *data leakage* dan juga kesalahan dalam membagi *dataset* untuk *training* dan *testing* yang menyebabkan hasil penelitian yang dilakukan menjadi rancu karena mendapatkan hasil yang tidak seharusnya. Oleh karena itu, penulis mencoba menggunakan metode *machine learning* lainnya, yaitu metode K-Nearest Neighbour (KNN) dan metode Extreme Gradient Booster (XGBoost) untuk diterapkan pada objek penelitian yang sama dengan tujuan agar mengetahui nilai *accuracy*, *recall*, *precision*, dan juga *specificity* sehingga bisa menyimpulkan metode yang terbaik untuk digunakan sebagai dasar untuk prediksi pergerakan harga dari pasar saham. Selain itu, untuk mengatasi permasalahan yang dihadapi oleh penelitian sebelumnya [5] penelitian ini akan menggunakan metode Windowing untuk mengatasi permasalahan *data leakage*, *overfitting*, dan kesalahan pembagian *dataset* untuk *training* dan *testing*.

II. TINJAUAN PUSTAKA

A. *Random Forest*

Algoritma Random Forest adalah metode *machine learning* yang efektif dan populer yang menggabungkan prediksi dari beberapa pohon keputusan acak (*random decision tree*) untuk menghasilkan prediksi akhir. Setiap pohon keputusan dalam kumpulan (*forest*) dibangun dengan menggunakan subset acak dari data pelatihan dan fitur-fitur yang dipilih secara acak. Keputusan akhir diambil berdasarkan mayoritas suara (klasifikasi) atau rata-rata (regresi) dari prediksi pohon keputusan dalam kumpulan [6]. Random Forest efektif dalam mengatasi *overfitting* dan mampu mengatasi dataset yang besar dengan banyak fitur [7].

$$\hat{y}i = \sum_{j=1}^{m} wj \cdot hj(xi) \quad (1)$$

Tabel 1. Keterangan Rumus Random Forest

Simbol	Keterangan
$\hat{y}i$	prediksi untuk sampel ke-i
m	jumlah pohon keputusan
wj	Bobot pohon keputusan ke-j
hj(xi)	Prediksi pohon keputusan ke-j untuk sampel ke-i

B. *Extreme Gradient Boosting (XGBoost)*

Extreme Gradient Boosting atau bisa juga disebut dengan XGBoost adalah algoritma *intelligent retrieval* yang merupakan implementasi yang ditingkatkan dari algoritma Gradient Boosting. XGBoost adalah algoritma yang kuat dan efektif untuk masalah klasifikasi, regresi, dan perangkingan. XGB menggunakan pendekatan *ensemble* dan membangun model prediksi dengan menggabungkan sejumlah besar pohon keputusan kecil yang lemah secara berurutan [8]. Rumus umum XGBoost adalah sebagai berikut.

$$\hat{y}i = \sum_{k=1}^{k} fk(xi) \quad (2)$$

Tabel 2. Keterangan Rumus XGBoost

Simbol	Keterangan
$\hat{y}i$	prediksi untuk sampel ke-i
k	jumlah pohon keputusan
$fk(xi)$	prediksi dari pohon ke-k untuk sampel ke-i

DOI: https://doi.org/10.35457/antivirus.v17i2.2871

C. *K-Nearest Neighbor (KKN)*

K-Nearest Neighbors atau yang biasa disebut dengan KNN adalah prosedur *machine learning* dan biasa diterapkan untuk pengelompokan dan regresi. Konsep dasar dari KNN adalah bahwa suatu data akan diklasifikasikan atau diprediksi berdasarkan mayoritas label dari tetangga terdekatnya. Dalam KNN, tetangga terdekat ditentukan berdasarkan jarak Euclidean atau metrik lainnya [9]. Menghitung jarak Euclidean antara dua titik dalam KNN bisa menggunakan rumus sebagai berikut.

$$Dxy = \sqrt{\sum_{i=1}^{n}(Xi - Yi)^2} \quad (3)$$

Tabel 3. Keterangan Rumus K-NN

Simbol	Keterangan
D	jarak *similarity*
x	data *training*
y	data *testing*
n	jumlah atribut pada individu antara 1 s.d. n
f	fungsi *similarity* atribut i antara kasus X dengan kasus Y
i	atribut individu dari 1 sampai dengan n

D. *Sliding Window Algorithm (SWA)*

Sliding Window atau biasa disebut *Windowing* merupakan pembuatan struktur dari sebuah data Transformasi dari yang sebelumnya data *time series* menjadi data *cross-sectional* dengan memperhatikan strukturnya. *Windowing* berfungsi untuk membagi data dengan ukuran yang cukup besar menjadi jendela (*window*) data yang lebih kecil [10]. Tujuan utama windowing adalah untuk mengatasi masalah sekuensialitas dan memfasilitasi pemrosesan data dalam bentuk jendela waktu yang lebih kecil. Ukuran dari *window* akan memengaruhi akurasi dan hasil prediksi. Oleh karena itu, ukuran *window* dapat diubah sesuai dengan kebutuhan dan data untuk memperoleh error terkecil.

III. METODE PENELITIAN

Penelitian ini akan dilakukan dengan melakukan beberapa tahapan yaitu menambahkan metode Windowing pada *dataset*, evaluasi model menggunakan metode Random Forest, KNN, XGBoost, kemudian menampilkan hasil grafik ROC. Alur penelitian ini dapat dilihat pada gambar 1.

Gambar 1. Tahapan Metode Penelitian

A. *Dataset*

Data yang diaplikasikan dalam penelitian berasal dari *repository* GitHub https://github.com/jmartinezheras/reproduce-stock-market-direction-random-forests.git. Dataset yang digunakan adalah AAPL.CSV dengan jangka waktu tahun 2010 sampai dengan 2014.

Tabel 4. Dataset

Date	Open	High	Low	Close	Adj Close	Volume
2010-03-31	33.641	33.801	33.494	33.571	22.726	107664900
2010-04-01	33.915	34.104	33.250	33.709	22.820	150786300
2010-04-05	33.568	34.072	33.538	34.070	23.063	171126900
2010-04-06	34.028	34.320	33.857	34.220	23.165	111754300

DOI: https://doi.org/10.35457/antivirus.v17i2.2871

B. Penambahan metode Windowing

Dalam prediksi harga saham, metode windowing digunakan untuk mengubah data historis harga saham menjadi serangkaian jendela waktu yang lebih kecil. Misalnya, jika data harga saham harian selama beberapa tahun, metode windowing dapat digunakan untuk menghasilkan jendela waktu harian atau mingguan. Setiap jendela waktu terdiri dari sejumlah titik data yang terurut secara sekuensial. Pendekatan ini memungkinkan analisis yang lebih terperinci pada setiap jendela waktu, sehingga memungkinkan model untuk mempelajari pola dan tren yang spesifik pada setiap periode waktu tersebut.

C. Evaluasi model KNN

Metode KNN dapat diterapkan dalam prediksi harga saham dengan menggunakan data historis dalam setiap jendela waktu yang dihasilkan melalui metode Windowing. Dalam konteks ini, algoritma KNN dapat digunakan untuk memprediksi pergerakan harga saham berdasarkan kumpulan tetangga terdekat pada data harga saham sebelumnya. Dengan menggunakan jendela waktu harian, Harga saham akan naik atau turun dapat diprediksi berdasarkan harga saham pada periode sebelumnya dengan KNN. KNN akan mencari K titik data harga saham terdekat dan mengambil mayoritas labelnya untuk memberikan prediksi.

D. Evaluasi model XGBoost

Dalam penelitian ini, penambahan metode XGBoost digunakan untuk memodelkan dan memprediksi pergerakan harga saham dalam setiap jendela waktu yang dihasilkan melalui metode Windowing. XGBoost adalah teknik *ensemble learning* yang menggabungkan beberapa model kecil (*weak model*) untuk membentuk model yang lebih kuat. Dalam prediksi harga saham, XGBoost dapat memperhitungkan berbagai fitur dan tren dalam data historis untuk menghasilkan prediksi yang lebih akurat.

E. Evaluasi model Random Forest

Random Forest juga merupakan algoritma *ensemble learning* yang bekerja dengan menggabungkan beberapa pohon keputusan (*decision tree*). Setiap pohon dalam Random Forest dilatih dengan menggunakan subset data dan fitur yang acak. Dengan menggabungkan hasil prediksi dari pohon-pohon tersebut, Random Forest memberikan prediksi yang lebih stabil dan akurat.

F. Receiver Operating Characteristic (ROC)

Receiver Operating Characteristic atau ROC adalah metode statistik yang digunakan untuk mengevaluasi dan memvisualisasikan performa suatu model klasifikasi biner [11]. ROC didasarkan pada evaluasi kinerja model yang menggunakan *trade-off* antara tingkat *True Positive Rate* (TPR) dengan tingkat *False Positive Rate* (FPR). TPR merupakan proporsi positif yang benar-benar teridentifikasi oleh model dengan benar, sedangkan FPR adalah proporsi negatif yang salah diidentifikasi positif. Dalam konteks ROC, biasanya digunakan variabel sensitivitas (TPR) dan spesifisitas (1 - FPR) sebagai sumbu X dan Y, sehingga menghasilkan kurva ROC.

IV. HASIL DAN PEMBAHASAN

Pada bagian ini, akan dijelaskan hasil dan analisis dari penelitian yang bertujuan untuk memprediksi harga saham dengan menggunakan beberapa metode *machine learning* dan metode *Windowing*.

A. Windowing

Pada penelitian ini, digunakan metode Windowing untuk mengubah data *time series* yang besar menjadi lebih kecil untuk memudahkan perhitungan sehingga memiliki hasil yang lebih akurat. Penelitian ini menggunakan ukuran *window* atau *window size* dengan nilai 5. Metode Windowing diterapkan pada saat pembagian *dataset* menjadi 80 persen untuk *training*, 20 persen untuk *testing*. Pelatihan dengan menggunakan 80% data bertujuan agar model memperoleh pemahaman yang kuat terhadap pola-pola dalam dataset. Selanjutnya, dengan menyisihkan 20% data untuk pengujian, evaluasi model menjadi lebih konsisten dan akurat karena menggunakan data yang tidak terlibat dalam proses pelatihan.

DOI: https://doi.org/10.35457/antivirus.v17i2.2871

B. Evaluasi Performa Model

Evaluasi dilakukan menggunakan hasil performa dari ketiga model *machine learning* yang telah dilakukan sebelumnya. Performa yang diukur adalah *accuracy*, *recall*, *precision*, dan *specificity*. Perhitungan dilakukan menggunakan *dataset training* dan *testing* untuk membandingkan performa model pada pembagian *dataset*. Hasilnya bisa dilihat pada tabel berikut.

Tabel 5. Hasil performa pada dataset training

	Accuracy	Recall	Precision	Specificity
RandomForest	0.96	0.94	0.88	0.96
XGBoost	0.95	0.94	0.85	0.95
KNN	0.75	0.35	0.41	0.86

Tabel 6. Hasil performa pada dataset testing

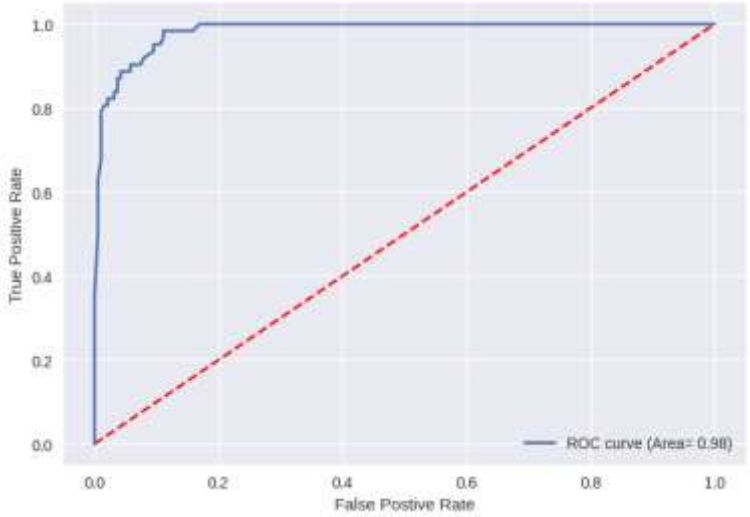
	Accuracy	Recall	Precision	Specificity
RandomForest	0.92	0.66	0.73	0.96
XGBoost	0.93	0.69	0.77	0.96
KNN	0.85	0.00	0.00	1.00

Berdasarkan tabel 5 dan tabel 6, Random Forest memeroleh hasil akurasi yang tertinggi pada saat *training* dengan akurasi 0,96. Akan tetapi, XGBoost (Extreme Gradient Boosting) memeroleh hasil akurasi yang tertinggi pada saat *testing* dengan akurasi 0,93. Dengan mempertimbangkan seluruh matriks evaluasi, XGBoost merupakan model terbaik dalam kasus ini. Dengan akurasi sebesar 0,93, recall sebesar 0,69, presisi sebesar 0,77, dan spesifisitas sebesar 0,96, XGBoost menunjukkan kinerja yang lebih baik secara keseluruhan apabila dibandingkan dengan model Random Forest dan KNN.

C. Receiver Operating Characteristic (ROC)

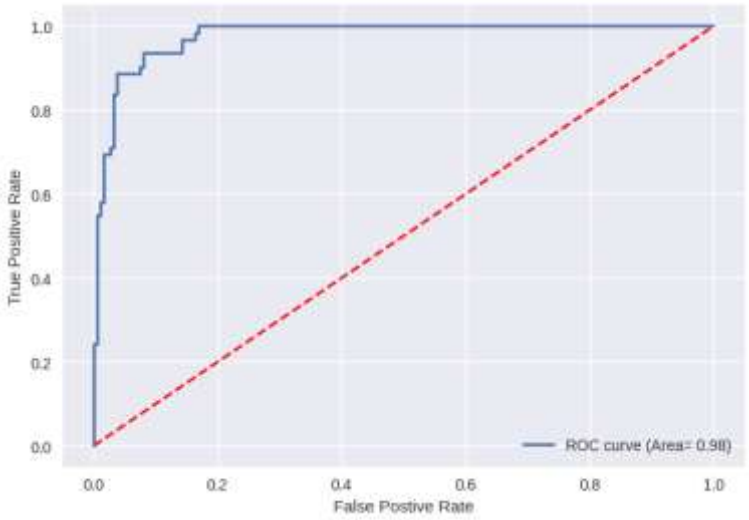
Untuk mengevaluasi kinerja model prediksi harga saham dapat digunakan dengan menampilkan ROC (*Receiver Operating Characteristic*). ROC memberikan gambaran tentang seberapa baik model dapat membedakan antara pergerakan harga saham yang benar (naik atau turun). Area di bawah kurva ROC (AUC-ROC) digunakan sebagai metrik evaluasi untuk menentukan seberapa baik model dapat membedakan pergerakan harga saham. Semakin tinggi AUC-ROC, semakin baik kinerja model

Apabila kurva semakin dekat ke batas kiri dan batas atas ruang ROC, ini mengindikasikan bahwa tes tersebut akurat. Semakin dekat kurva ke batas atas dan kiri, semakin akurat tes tersebut. Jika kurva mendekati diagonal 45 derajat dari ruang ROC, itu berarti tes tersebut tidak akurat. Kurva ROC dapat digunakan untuk memilih model yang optimal dan membuang yang tidak optimal. Berikut adalah perbandingan hasil ROC dari setiap model pada saat *training*.

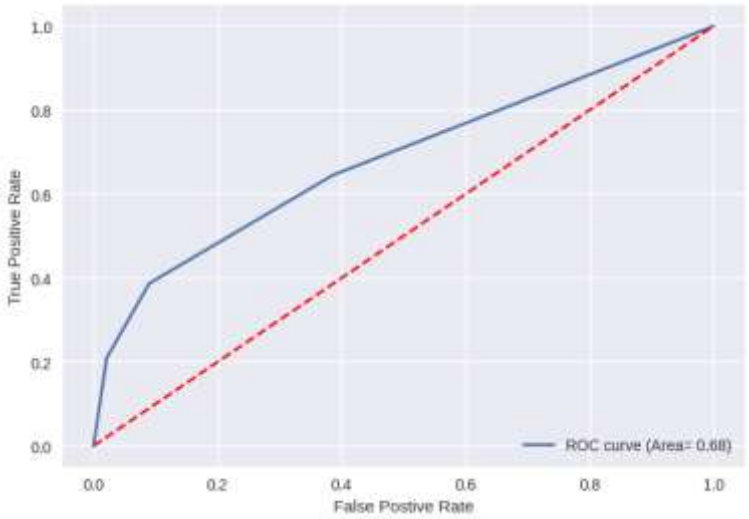


Gambar 2. ROC model Random Forest pada saat training

DOI: https://doi.org/10.35457/antivirus.v17i2.2871

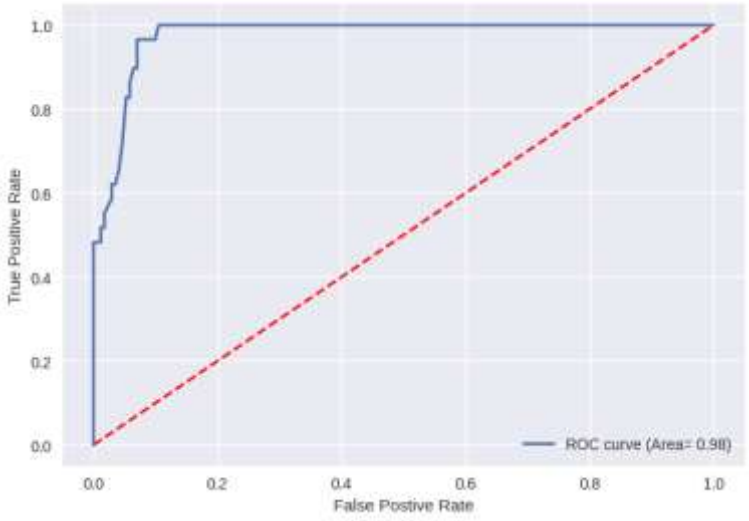


Gambar 3. ROC model XGBoost pada saat training



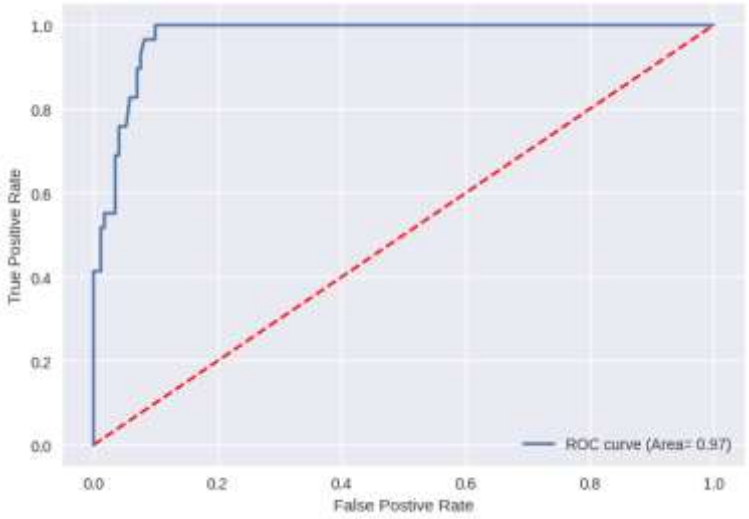
Gambar 4. ROC model KNN pada saat training

Setelah model dilatih dengan menggunakan *dataset training*, kemudian dilakukan tahap pengujian menggunakan *dataset testing*. Tujuan penggunaan *dataset testing* adalah untuk mengevaluasi sejauh mana model dapat mengeneralisasi informasi yang diperoleh dari data pelatihan ke data yang belum pernah dilihat sebelumnya.

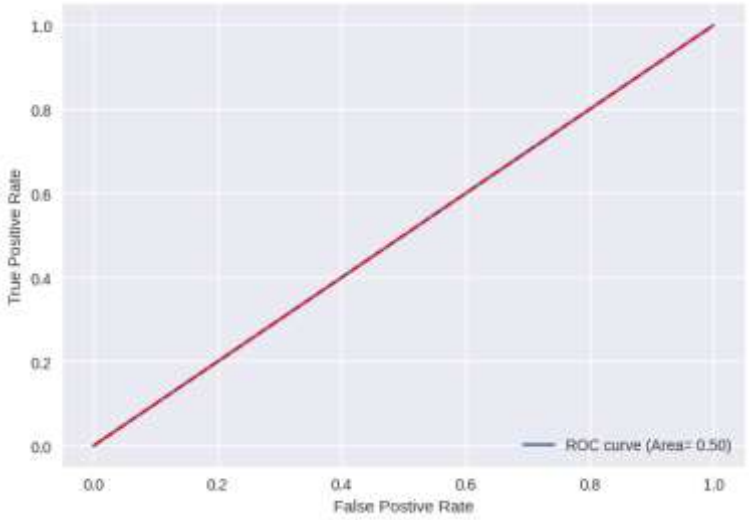


Gambar 5. ROC model Random Forest pada saat testing

DOI: https://doi.org/10.35457/antivirus.v17i2.2871



Gambar 6. ROC model XGBoost pada saat testing



Gambar 7. ROC model KNN pada saat testing

Seperti yang dapat dilihat dari kurva ROC di atas, model Random Forest terbukti menjadi model yang paling optimal. Area di bawah kurva ROC adalah parameter yang sangat penting untuk mengevaluasi kinerja model pengklasifikasi biner. Akurasi diukur dari area di bawah kurva ROC. Area 1 menunjukkan pengklasifikasi yang sempurna, area 0,9 - 0,6 menunjukkan pengklasifikasi yang cukup baik, dan area 0,5 menunjukkan pengklasifikasi yang kurang baik serta menghasilkan *output* acak. Area di bawah kurva ROC berada di atas 0,9 untuk kedua model. Ini berarti dua model pengklasifikasi tersebut sangat baik sedangkan untuk model KNN memiliki nilai terendah dalam setiap ROC.

V. KESIMPULAN DAN SARAN

Berdasarkan analisis dan pembahasan sebelumnya, dapat disimpulkan bahwa Random Forest menunjukkan nilai ROC yang lebih tinggi. Di sisi lain, XGBoost memberikan performa yang lebih baik dalam hal akurasi, recall, dan presisi. Oleh karena itu, pilihan model terbaik antara Random Forest dan XGBoost tergantung pada preferensi dan tujuan penggunaan, serta matriks evaluasi yang lebih diutamakan. Jadi, kedua model dapat dianggap sebagai opsi yang baik, tergantung pada kebutuhan spesifik pengguna.

DOI: https://doi.org/10.35457/antivirus.v17i2.2871

DAFTAR PUSTAKA

[1] Raharjo, S. (2006). Kiat Membangun Aset Kekayaan. Jakarta, Indonesia, 2006.
[2] Meidiawati, K., & Mildawati, T. (Februari, 2016). Pengaruh size, growth, profitabilitas, struktur modal, kebijakan dividen terhadap nilai perusahaan. *Jurnal Ilmu Dan Riset Akuntansi (JIRA)*. [Online]. *5(2).* Tersedia : http://jurnalmahasiswa.stiesia.ac.id/index.php/jira/article/view/1536
[3] Devitra, J. (Juni, 2013). Kinerja Keuangan dan Efisiensi Terhadap Return Saham Perbankan di Bursa Efek Indonesia Periode 2007-2011. *Jurnal Keuangan dan Perbankan*. [Online]. *15(1)*, hal. 38–53. Tersedia : https://journal.perbanas.id/index.php/jkp/article/view/181
[4] Mariana, S, "Pengaruh Faktor Fundamental, Faktor Teknikal Dan Risiko Sistematik Terhadap Harga Saham Pada Sektor Perbankan Indeks InfoBank15 Yang Terdaftar Di Bursa Efek Indonesia Periode 2015-2018," disertasi doctor, Program Studi Magister Akuntansi, Sekolah Pasca Sarjana Universitas Widyatama, Indonesia, 2020.
[5] Khaidem, L., Saha, S., & Dey, S. R. (April, 2016). Predicting the direction of stock market prices using random forest. [Online]. *arXiv preprint arXiv:1605.00003*. Tersedia : https://arxiv.org/abs/1605.00003
[6] Breiman, L. (Oktober, 2001). Random forests. *Machine learning*. [Online]. *45*, hal. 5–32. Tersedia : https://link.springer.com/article/10.1023/a:1010933404324
[7] Cutler, D. R., Edwards Jr, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., & Lawler, J. J. (November, 2007). Random forests for classification in ecology. *Ecology*. [Online]. *88(11)*, hal. 2783–2792. Tersedia : https://esajournals.onlinelibrary.wiley.com/doi/abs/10.1890/07-0539.1
[8] Chen, T., dan Guestrin, C, "Xgboost: A scalable tree boosting system". Dalam Proc. 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016, hal. 785-794.
[9] Cover, T., & Hart, P. (Januari, 1967). Nearest neighbour pattern classification. *IEEE transactions on information theory*. [Online]. *13(1)*, hal. 21–27. Tersedia : https://ieeexplore.ieee.org/document/1053964
[10] Wahyuni, R. E. (Juli, 2021). Optimasi Prediksi Inflasi dengan Neural Network Pada Tahap Windowing: Adakah Pengaruh Terhadap Window Size?. *Jurnal Ilmiah "Technologia"*. [Online]. *12(3)*. Hal. 176–181. Tersedia : https://ojs.uniska-bjm.ac.id/index.php/JIT/article/view/5181
[11] Fawcett, T. (Juni, 2006). An introduction to ROC analysis. *Pattern Recognition Letters*. [Online]. *27(8)*, hal. 861–874. Tersedia : https://www.sciencedirect.com/science/article/abs/pii/S016786550500303X

DOI: https://doi.org/10.35457/antivirus.v17i2.2871