PROSIDING SEMINAR NASIONAL PEMANFAATAN SAINS DAN TEKNOLOGI INFORMASI 2023
Vol. 1 No. 1 (2023)297-306

Penerapan Data Mining untuk Prediksi Pergerakan Harga Saham Menggunakan Algoritma *K-Nearest Neighbor*

Ihzan Sayid Muallif[1], Herdi Budiman[2], Natalis Ransi*[3]
[1,2,3]Program Studi Ilmu Komputer, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Halu Oleo, Kendari
e-mail: [1]**ihzansayidmuallif@gmail.com**, [2]herdi.budiman@uho.ac.id, *[3]natalis.ransi@uho.ac.id

Abstrak

Pergerakan harga saham merupakan fenomena yang kompleks dan dipengaruhi oleh berbagai faktor, sehingga menjadi tantangan dalam melakukan prediksi yang akurat. Algoritma *K-Nearest Neighbor* adalah salah satu metode klasifikasi yang digunakan dalam data mining untuk mengklasifikasikan data ke dalam kelas tertentu berdasarkan kedekatannya dengan tetangga terdekat dalam hal atribut atau variabel. Penelitian ini bertujuan untuk mengetahui implementasi metode *K-Nearest Neighbor* dalam memprediksi pergerakan harga saham. Penelitian ini menggunakan data historis harga saham sebagai data latih untuk melatih model *K-Nearest Neighbor*. Proses pelatihan melibatkan identifikasi dan perangkingan saham-saham yang memiliki karakteristik serupa berdasarkan data historis. Penelitian ini menambahkan beberapa variabel-variabel yang memiliki pengaruh terhadap pergerakan harga saham yang membedakan dari penelitian terdahulu. Pada penelitian ini, ditemukan model prediksi dengan tingkat *accuracy* tertinggi sebesar 62,54%, *precision* sebesar 64,14%, dan *recall* sebesar 92,08%. Hasil ini diperoleh dengan menggunakan nilai k=53 pada random data ke-6 dalam data set ke-2.

Kata kunci—K-Nearest Neighbor, Data Mining, Prediksi, Saham

Abstract

Stock price movements are a complex phenomenon and are influenced by various factors, making it a challenge to make accurate predictions. The K-Nearest Neighbor algorithm is one of the classification methods used in data mining to classify data into certain classes based on their proximity to the nearest neighbors in terms of attributes or variables. This study aims to determine the implementation of the K-Nearest Neighbor method in predicting stock price movements. This study uses historical stock price data as training data to train the KNearest Neighbor model. The training process involves identifying and ranking stocks with similar characteristics based on historical data. This research adds several variables that have an influence on stock price movements that differentiate it from previous research. In this study, it was found that a prediction model with the highest level of accuracy was 62.54%, precision was 64.41%, and recall was 92.08%. This result is obtained by using the value of k = 53 in the 6th random data in the 2nd data set.

Keywords—*K-Nearest Neighbor, Data mining, Prediction, Stock*

1. PENDAHULUAN

Menurut Otoritas Jaksa Keuangan (OJK) pada laman website-nya, saham dapat diartikan sebagai tanda penyertaan modal seseorang atau pihak (badan usaha) pada suatu perusahaan atau perseroan terbatas. Dengan menyertakan modal tersebut, maka pihak tersebut memiliki klaim (hak) atas pendapatan perusahaan, aset perusahaan, dan berhak hadir dalam rapat umum pemegang saham (RUPS). Saham merupakan salah satu produk pasar modal yang menjadi salah satu instrumen investasi untuk jangka panjang. Saham ialah bukti penyertaan modal di sebuah perusahaan. dengan membeli saham perusahaan, berarti anda menginvestasikan modal/dana yang nantinya akan dipergunakan oleh pihak manajemen untuk membiayai aktivitas operasional perusahaan[1]. Salah satu Perusahaan yang

menjual sahamnya kepada masyarakat adalah bank BCA dengan kode BBCA.JK.

Bank Central Asia (BCA) didirikan pada 10 agustus tahun 1955 dengan nama NV Perseroan Dagang Dan Industrie Semarang Knitting Faktory. Namanya kemudian berganti menjadi Central Bank Asia pada tahun 1957 dan mulai beroprasi pada 21 februari 1957 yang berkantor pusat di Jakarta. Nama Bank telah diubah beberapa kali hingga pada tahun 1975 nama Bank diubah menjadi PT Bank Central Asia (BCA). BCA melakukan IPO pertama kali pada 11 mei 2000 (bca.co.id). IPO (*Initial Public Offering*) adalah penawaran pertama di mana emiten baru saja melantai atau hadir di pasar saham. Emiten merupakan sebutan dari perusahaan yang menjual sahamnya di Bursa Efek Indonesia. Saham BCA merupakan salah satu saham yang sering membagikan *dividen* kepada para pemegang sahamnya, pembagiannya dapat dilihat pada *website* bank BCA.

Harga-harga saham selalu mengalami fluktuasi baik berupa kenaikan maupun penurunan. Pembentukan harga saham terjadi karena adanya permintaan dan penawaran atas saham tersebut. Permintaan dan penawaran atas suatu dipengaruhi banyak faktor, baik yang sifatnya spesifik berhubungan dengan saham tersebut (kinerja perusahaan dan industri dimana perusahaan tersebut berada) maupun faktor-faktor non ekonomi seperti kondisi sosial dan politik. Faktor ekonomi makro antara lain adalah nilai tukar, cadangan devisa, suku bunga, inflasi dan bank Indonesia *rate*. Sedangkan faktor ekonomi mikro antara lain *non performing loans* (NPL), *capital adequency ratio* (CAR), *loan to deposit ratio* (LDR), *return on asset* (ROA), *return on equity* (ROE), *net profit margin* (NPM), *net interest margin* (NIM),*earning per share* (EPS), dewan direksi, rapat umum pemegang saham (RUPS), beban operasional pendapatan operasional (BOPO) dan komite audit[2].

Terjadinya kenaikan dan penurunan harga saham ini lah yang membuat investor saham melakukan Analisa maupun prediksi terhadap harga saham kedepannya. Salah satu cara yang dapat dilakukan untuk memprediksi harga saham adalah dengan menambang dan mengolah data saham yang terdapat pada pasar modal. Ada beberapa metode yang dapat digunakan untuk mengolah data, khususnya dalam kasus prediksi yaitu data mining.

Data mining adalah kegiatan mengekstraksi atau menambang pengetahuan dari data yang berukuran/berjumlah besar, informasi inilah yang nantinya sangat berguna untuk pengembangan. Definisi sederhana dari data mining adalah ekstraksi informasi atau pola yang penting atau menarik dari data yang ada di database yang besar[3]. Beberapa teknik dalam data mining adalah estimasi, *forecasting*, klasifikasi, asosiasi dan klastering.

Klasifikasi adalah bentuk analisis data yang mengekstraksi model yang menggambarkan kelas data penting. Model seperti itu, disebut pengklasifikasi, memprediksi label kelas kategoris (diskrit,tidak berurutan). klasifikasi data merupakan proses dua langkah yang terdiri dari langkah pembelajaran (dimana model klasifikasi dibangun) dan langkah klasifikasi (dimana model ini dipergunakan untuk memprediksi label kelas untuk data yang diberikan)[4]. Adapun algoritma yang dapat digunakan untuk klasifikasi dan prediksi salah satunya adalah K-Nearest Neighbor.

Algoritma K-Nearest Neighbor (KNN) adalah sebuah metode untuk melakukan klasifikasi terhadap objek berdasarkan data pembelajaran yang jaraknya paling dekat dengan objek tersebut. Data pembelajaran diproyeksikan ke ruang berdimensi banyak, dimana masing-masing dimensi merepresentasikan fitur dari data. Ruang ini dibagi menjadi bagian-bagian berdasarkan klasifikasi data pembelajaran[5].

Rumusan masalah pada penelitian ini adalah bagaimana implementasi metode *K-Nearest Neighbor* dalam memprediksi pergerakan harga saham. Adapun tujuan dari Penelitian ini adalah untuk mengetahui implementasi metode *K-Nearest Neighbor* dalam memprediksi pergerakan harga saham.

Data saham yang digunakan pada penelitian ini, didalamnya terdapat 9 variabel yaitu *Date*, *Open*, *High*, *Low*, *Close*, ROE (*Return On Equity*), ROA (*Return On Asset*), EPS (*Earnings Per Share*), dan *class* yang dimana menjadi salah satu pembeda dari penelitian yang telah ada sebelumnya. *Date* adalah variabel yang menunjukan waktu dalam setiap data saham yang ada. *Open* adalah harga perdagangan pertama suatu sebuah saham pada hari itu. Harga ini merupakan nilai pertama yang diajukan dan dieksekusi ketika pasar dibuka. *High* merupakan nilai harga paling

tinggi yang dicapai oleh suatu saham selama periode perdagangan pada hari itu. Hal ini menunjukkan titik harga puncak yang berhasil dicapai oleh saham dalam satu sesi perdagangan. *Low* memrupakan harga terendah yang dicapai oleh suatu saham saham selama periode perdagangan pada hari itu. Hal ini menujukkan titik harga paling rendah yang dapat dicapai oleh saham dalam satu sesi perdagangan. *Close* merupakan nilai perdagangan terakhir sebuah saham pada akhir sesi perdagangan dibursa saham pada hari itu. Harga ini merupakan nilai terakhir yang diajukan dan di eksekusi sebelum pasar ditutup.

ROE (*Return on Equity*) adalah suatu rasio yang digunakan untuk menunjukkan hubungan antara laba bersih dan ekuitas suatu perusahaan. Semakin tinggi rasio ROE, semakin besar pula laba bersih yang dihasilkan oleh dana yang diinvestasikan dalam ekuitas atau modal perusahaan. ROA (*Return on Asset*) adalah suatu rasio yang digunakan untuk mengevaluasi kemampuan perusahaan dalam menghasilkan laba bersih dari aset yang dimilikinya. EPS (*Earnings per Share*) adalah suatu rasio yang digunakan oleh pemegang saham untuk mengevaluasi kinerja manajemen atau perusahaan dalam mencapai laba. Semakin tinggi rasio EPS, semakin meningkat kesejahteraan investor. Sedangkan *Class* adalah kategori atau label yang ingin diprediksi oleh model dari data yang diberikan.

Uji korelasi merupakan teknik analisis yang termasuk dalam salah satu teknik pengukuran asosiasi atau hubungan (*measures of association*)[6]. Terdapat beberapa teknik statistik untuk analisis korelasi, salah satunya adalah uji korelasi pearson. Korelasi pearson berguna untuk menentukan hubungan antara dua variabel yang berskala interval (skala yang menggunakan angka sebenarnya), oleh karena itu korelasi termasuk dalam kategori uji statistik parametrik. Besarnya korelasi adalah 0 s/d 1. Hasil korelasi yang mendekati angka 1 memiliki hubungan yang sangat tinggi, sedangkan hasil korelasi yang mendekati angka 0 memiliki hubungan yang dapat dianggak tidak ada[7]. Adapun rumus korelasi pearson ditunjukkan pada Persamaan 1.

$$r_{xy} = \frac{n\Sigma xy - \Sigma x \Sigma y}{\sqrt{\{ n\Sigma x^2 - (\Sigma x)^2 \}} \sqrt{\{ n\Sigma y^2 - (\Sigma y)^2 \}}} \quad (1)$$

Keterangan :
r_{xy} : Koefisien korelasi r pearson
n : jumlah sampel
x : variabel bebas
y : variabel terikat

Metode normalisasi min-max merupakan metode normalisasi dengan melakukan transformasi linear terhadap data asli [8]. Tujuan utama normalisasi adalah untuk menangani perbedaan skala antara variabel dalam *dataset*, terutama ketika menggunakan algoritma analisis data atau pemodelan statistik yang sensitif terhadap perbedaan skala. Dengan melakukan normalisasi, setiap variabel akan memiliki dampak yang seimbang dalam analisis, sehingga menghasilkan hasil yang lebih akurat dan konsisten. Adapun perhitungan normalisasi ditunjukkan pada Persamaan 2.

$$x_n = \frac{x_0 - x_{min}}{x_{max} - x_{min}} \quad (2)$$

Keterangan :
X_n : nilai baru untuk variabel x
x_0 : nilai lama untuk variabel x
x_{min} : nilai minimun dari suatu fitur
x_{max} : nilai maksimum dari suatu fitur

Algoritma *K-Nearest Neighbor* merupakan metode untuk mengelompokkan objek berdasarkan kemiripan dengan contoh-contoh latihan dalam ruang fitur. KNN menjadi teknik klasifikasi yang paling mendasar dan sederhana, terutama ketika pengetahuan awal tentang distribusi data minim atau bahkan tidak ada. Prinsip KNN melibatkan pengklasifikasian setiap sampel yang mirip dengan sampel-sampel di sekitarnya. Dalam hal ini, jika kelas suatu sampel belum diketahui, prediksi kelasnya dapat dibuat berdasarkan mayoritas kelas sampel-sampel tetangga terdekat [9]. Untuk memprediksi pergerakan harga saham menggunakan K-Nearest Neighbor, rumus *euclidean distance* ditunjukkan pada Persamaan 3.

$$d(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad (3)$$

Keterangan :
d(x,y) : Jarak
x_i : Sampel data *training*
y_i : Sampel data *testing*

n : Dimensi data
i : Variabel data

Confusion matrix adalah akurasi hasil klasifikasi yang diuji dengan cara membuat matrix kontingensi [10]. Untuk menghitung nilai akurasi dapat menggunakan *confusion matrix*. Adapun cara untuk menghitung nilai akurasi yang ditunjukkan oleh Tabel 1.

Tabel 1 *Confuion Matrix*

Predict Label	*Actual Label*	
	Positive	*Negative*
Positive	*True Positive* (TP)	*False positive* (FP)
negative	*False Negative* (FN)	*True Negative* (TN)

Keterangan :
TP : Merupakan data positif dan diprediksi benar
TN : Merupakan data negatif dan diprediksi benar
FP : Merupakan data negatif, namun diprediksi sebagai data positif
FN : Merupakan data positif, namun diprediksi sebagai data negatif

Accuracy adalah jumlah prediksi yang benar dibagi dengan keseluruhan sampel data. Nilai *acuraccy* ditunjukkan pada Persamaan 4.

$$accuracy = \frac{TP + TN}{(TN + TP + FP + FN)} \quad (4)$$

Recall merupakan seberapa banyak prediksi yang benar dari semua kelas positif. Nilai *recall* ditunjukkan pada Persamaan 5.

$$recall = \frac{TP}{(TP + FN)} \quad (5)$$

Precision merupakan rasio dari berapa banyak prediksi yang benar-benar positif dari semua kelas positif yang diprediksi dengan benar. Nilai *precision* ditunjukkan pada Persamaan 6.

$$precision = \frac{TP}{(TP + FP)} \quad (6)$$

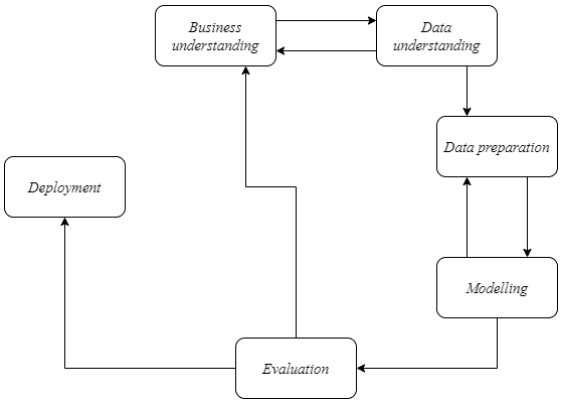
Cross Industry Standard Process For Data Mining (CRISP-DM) merupakan model dari proses yang menyediakan kerangka kerja untuk melaksanakan proyek data mining yang *independent* dari sektor industri dan teknologi yang digunakan. CRISP-DM bertujuan untuk membuat projek data mining yang besar, lebih cepat dan mudah untuk dikelolah [11]. Berikut adalah tahapan-tahapan dalam CRISP-DM :

a. *Business understanding*
Tahapan awal pada CRISP-DM adalah memahami masalah dan tujuan penelitian yang ingin dilakukan.
b. *Data understanding*
Pada tahap ini, data yang akan digunakan untuk penelitian dikumpulkan.
c. *Data preparation*
Setelah data dikumpulkan, pada tahap selanjutnya adalah mempersiapkan data untuk digunakan dalam penelitian.
d. *Modeling*
Pada tahap keempat ini, dilakukan pemilihan model, pengolahan data dan pelatihan model.
e. *Evaluation*
Selanjutnya, model yang telah dipilih dan dilatih akan dievaluasi untuk menilai kinerjanya.
f. *Deployment*
Tahap terakhir adalah penggunaan model untuk tujuan bisnis. Model dapat diimplementasikan dalam produksi dan digunakan untuk membuat keputusan bisnis atau dapat digunakan untuk melakukan tindakan yang sesuai dengan hasil analisis.

Rapidminer adalah alat perangkat lunak *open source* gratis untuk *data/text mining*. Rapidminer tersedia sebagai aplikasi yang berdiri sendiri untuk analisis data/teks dan sebagai mesin *data/text mining*. Rapidminer juga merupakan sistem yang mendukung desain dan dokumentasi proses data mining secara keseluruhan. Rapidminer menawarkan tidak hanya seperangkat operator yang hampir komprehensif, tetapi juga struktur yang mengekspresikan aliran kontrol proses [12].

2. METODE PENELITIAN

Tahapan prosedur yang digunakan pada penelitian ini adalah CRISP-DM. Tahapan prosedur ditunjukkan pada Gambar 1.



Gambar 1 Prosedur Penelitian CRISP-DM

a. *Business Understanding*

Pada tahap ini yaitu memahami masalah dan tujuan penelitian. Dalam prediksi pergerakan harga saham, tujuan bisa berupa memprediksi pergerakan harga saham suatu perusahaan.

b. *Data Understanding*

Pada tahap ini, data yang akan digunakan untuk melakukan prediksi harga saham dikumpulkan. Data yang digunakan dapat berupa data historis harga saham.

c. *Data Preparation*

Setelah data dikumpulkan, tahap selanjutnya adalah mempersiapkan data untuk digunakan dalam model prediksi. Pada tahap ini, dilakukan seleksi variabel, uji korelasi, penghilangan *missing value*, dan *preprocessing data* lainnya seperti pembagian data *testing* dan data *training* serta normalisasi data untuk menghasilkan data yang berkualitas agar dapat digunakan dalam model prediksi.

d. *Modelling*

Pada tahap ini, model prediksi yang digunakan adalah *k-nearest neighbor*. Model ini akan mempelajari pola dari data historis harga saham dan dapat digunakan untuk memprediksi pergerakan harga saham.

e. *Evaluation*

pada tahap ini dilakukan uji performa model menggunakan data testing. Performa model dapat diukur menggunakan *confusion matrix* meliputi *accuracy*, *precision* dan *recall*.

f. *Deployment*

pada tahap ini dilakukan perbandingan dari tiga data untuk melihat data dengan jenis label terbaik.

Pada penelitian ini, menggunakan data sekunder sebagai sumber data. Data sekunder adalah data yang didapatkan dari database yang sudah ada sebelumnya. Data sekunder didapatkan dari platform Yahoo Finance dengan nama *dataset* BBCA.JK. Sedangkan data faktor diperoleh dari laporan tahunan diwebsite resmi PT Bank Central Asia Tbk.

Instrumen yang digunakan pada penelitian ini terbagi menjadi dua, yaitu perangkat keras (hardware) dan perangkat lunak (software). Instrumen penelitian ditunjukkan pada Tabel 2.

Tabel 2 Instrumen Penelitian

Perangkat Keras	Perangkat Lunak
Laptop Lenovo *processor* AMD Ryzen 7. RAM 8 GB SSD 512 GB	Sistem operasi Windows 11
	Microsoft Excel
	Anaconda
	Rapidminer Studio versi 10.1.001

3. HASIL DAN PEMBAHASAN

3.1 Data Preparation

Pada fase ini dilakukan pengecekan missing value pada *dataset*, uji korelasi, pembagian data training dan data testing serta penentuan class kepada 3 *dataset* yang akan diuji. Data saham yang akan digunakan dalam penelitian kali ini berjumlah 3225 data dengan 9 variabel.

a. Penentuan *Class*

Untuk memaksimalkan hasil prediksi, dilakukan penentuan class kepada 3 *dataset* yang sama untuk dilakukan pengujian. Penentuan *class* ditunjukkan pada Tabel 3.

Tabel 3 Jumlah Pembagian *Class* pada 3 *Dataset*

Data	Class	Jumlah	Persentase
1	Naik	1465	45,43%

	Tetap	415	12,87%
	Turun	1345	41,7%
2	Naik	1880	58.3%
	Turun	1345	41,7%
3	Naik	1465	45,43%
	Turun	1760	54,57%

Pada Tabel 2, terdapat 3 *dataset* dengan pembagian *class*. *Dataset* 1 terdirir atas 3 *class* yaitu naik dengan persentase 45,43%, tetap 12,87% dan turun 41,7%. *Dataset* 2 terdiri atas 2 *class* yaitu naik 58,3% dan turun 41,7% yang mana pada data ini *class* tetap dianggap menjadi *class* naik. Pada *dataset* 3 terdiri atas 3 *class* yaitu naik 45,43% dan turun 54,57% yang mana pada data ini *class* tetap dianggap sebagai *class* turun. Ketiga data saham tersebut memiliki jumlah variabel dan jumlah data yang sama. Sehingga pembagian *class* yang menjadikan ketiga data set tersebut berbeda.

b. *Missing Value*

Pengecekan *missing value* penting dilakukan kepada data saham yang akan digunakan karena nilai yang hilang memiliki potensi untuk mempengaruhi hasil analisis dan pemodelan data. Pengecekan *missing value* pada data saham ditunjukkan pada Gambar 2.

```
In [63]: import pandas as pd

         # Baca dataset dari CSV
         dataset = pd.read_csv('BBCAJK.csv')

         # Periksa apakah ada nilai yang hilang (missing value) dalam dataset
         if dataset.isnull().values.any():
             print("Terdapat nilai yang hilang (missing value) dalam dataset.")
             # Jumlah nilai yang hilang pada setiap kolom
             print("\nJumlah nilai yang hilang pada setiap kolom:")
             print(dataset.isnull().sum())
         else:
             print("Tidak ada nilai yang hilang (missing value) dalam dataset.")

         Tidak ada nilai yang hilang (missing value) dalam dataset.
```

Gambar 2 Pengecekan *Missing Value*

Pada Gambar 2, menunjukkan bahwa tidak terdapat missing value pada data saham yang akan digunakan.

c. Uji Korelasi

Uji korelasi dilakukan untuk menguji hubungan antara 2 variabel. Pada data saham yang digunakan pengujian menggunakan uji korelasi pearson. Pengujian ditunjukkan pada Gambar 3.

```
In [65]: import pandas as pd
         from scipy.stats import pearsonr

         # Baca dataset dari CSV
         dataset = pd.read_csv('BBCAJK.csv')

         # Menghapus nilai NaN jika ada di seluruh dataset
         dataset = dataset.dropna()

         # Mengambil hanya kolom dengan tipe data numerik untuk perhitungan korelasi
         numeric_columns = dataset.select_dtypes(include=[float, int])

         # Menghitung korelasi Pearson untuk setiap pasangan kolom numerik
         correlation_matrix = numeric_columns.corr()

         # Menampilkan koefisien korelasi antara setiap pasangan kolom
         print("Koefisien Korelasi Pearson:")
         print(correlation_matrix)

         Koefisien Korelasi Pearson:
                  Open      High       Low     Close       ROE       ROA       EPS
         Open     1.000000  0.999737  0.999809  0.999631  0.952472  0.943883  0.934472
         High     0.999737  1.000000  0.999665  0.999777  0.953111  0.943658  0.934945
         Low      0.999809  0.999665  1.000000  0.999799  0.952575  0.944722  0.934503
         Close    0.999631  0.999777  0.999799  1.000000  0.952826  0.944199  0.934769
         ROE      0.952472  0.953111  0.952575  0.952826  1.000000  0.953640  0.969069
         ROA      0.943883  0.943658  0.944722  0.944199  0.953640  1.000000  0.957943
         EPS      0.934472  0.934945  0.934503  0.934769  0.969069  0.957943  1.000000
```

Gambar 3 Uji Korelasi

Pada Gambar 3, menunjukkan bahwa Ketika variabel-variabel yang ada dalam data saham diuji menggunakan uji korelasi pearson, Menurut [7] jika hasil yang didapatkan memiliki rentan nilai 0,90-1,00 yang berarti memiliki korelasi atau hubungan sangat tinggi antar variabel.

d. Data *Training* dan Data *Testing*

Data *training* digunakan untuk mengembangkan model menggunakan algoritma KNN. Model ini kemudian diuji pada data *testing* untuk mengukur tingkat akurasi dalam memprediksi pergerakan saham. Pembagian serta penentuan data *training* dan data *testing* dapat dilakukan menggunakan pemrograman *python*, ditunjukkan pada Gambar 4.

```
In [11]: from sklearn.model_selection import train_test_split

         # X adalah matriks fitur
         # y adalah vektor target
         X = dataset[['Date', 'Open', 'High', 'Low', 'Close', 'ROE', 'ROA', 'EPS']]
         y = dataset['class']
         # Membagikan data menjadi data pelatihan dan data pengujian secara acak dengan proporsi 70:30
         X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=0)

         # random_state adalah nilai yang digunakan untuk menghasilkan urutan acak yang sama setiap kali kode dijalankan, sehingga memasti
```

Gambar 4 Pembagian Data *Training* dan *Tesiting*

Pada Gambar 4 telah dilakukan pembagian data *training* dan data *testing* menggunakan *python* dengan perbandingan jumlah data 70:30, di mana x merupakan variabel bebas dan y merupakan variabel terikat. Setelah pembagian tersebut, terdapat 2275 data pada data *training* dan 968 data pada data *testing*. Informasi terkait data *training* dan data *testing* ditunjukkan pada Gambar 5 dan Gambar 6.

```
In [12]: X_train
Out[12]:
```

	Date	Open	High	Low	Close	ROE	ROA	EPS
891	19/08/2013	2070	2070	1900	1900	306211	317123	139585
2245	18/01/2019	5340	5425	5315	5425	470911	552244	195315
670	19/09/2012	1570	1600	1570	1570	260708	263530	97561
1570	27/05/2016	2600	2605	2580	2600	381826	447312	169919
1140	29/08/2014	2240	2370	2240	2240	324344	362354	137090
...	...	...	...	...	...	...	...	...
763	07/02/2013	2010	2010	1980	2010	306211	317123	118648
835	23/05/2013	2200	2210	2160	2180	306211	317123	107861
1653	30/09/2016	3195	3200	3100	3140	381826	447312	161827
2607	25/06/2020	5700	5745	5640	5725	520422	495874	216450
2732	04/01/2021	6800	6855	6720	6835	608457	563747	258097

2257 rows × 8 columns

Gambar 5 Data *Training*

In [14]: X_test

Out[14]:

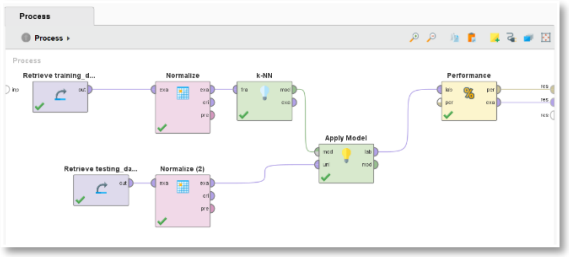
	Date	Open	High	Low	Close	ROE	ROA	EPS
1460	14/12/2015	2600	2640	2590	2640	326995	376297	149184
2822	18/05/2021	6400	6455	6365	6390	608457	563747	303644
599	06/06/2012	1440	1450	1410	1440	260708	263530	92915
2604	22/06/2020	5595	5660	5530	5540	520422	495874	216450
2791	30/03/2021	6370	6455	6370	6395	608457	563747	234634
...	...	...	...	...	...	...	...	...
1627	24/08/2016	3060	3070	3055	3060	381826	447312	154472
402	19/08/2011	1600	1600	1550	1600	230770	237875	94609
1231	09/01/2015	2605	2615	2585	2585	326995	376297	142080
2428	02/10/2019	6010	6070	6010	6050	470911	552244	195315
2366	08/07/2019	5935	5960	5865	5880	470911	552244	195315

968 rows × 8 columns

Gambar 6 Data *Testing*

3.2 *Modelling*

Sebelum dilakukan pemodelan algoritma KNN, terlebih dahulu dilakukan penentuan nilai k. Penentuan nilai k yang akan digunakan tidak memiliki aturan yang baku. Pada penelitian ini dilakukan beberapa kali percobaan nilai k untuk menentukkan nilai k terbaik sehingga didapatkan nilai k = 53 memiliki tingkat akurasi yang lebih tinggi dibandingkan nilai k lain. Implementasi algoritma KNN dilakukan menggunakan *software* rapidminer. Pengguanaan *software* rapidminer untuk *modelling* dapat dilihat pada Gambar 7.



Gambar 7 Model klasifikasi KNN menggunakan rapidminer

Pada Gambar 4.3 menunjukkan tampilan proses pemodelan KNN. Data *training* dihubungkan ke operator *normalize* untuk menormalisasikan data agar rentan nilai dari setiap variabelnya sama yaitu skala 0 sampai dengan 1. Kemudian dari dari operator *normalize* dihubungkan ke operator KNN untuk dilakukan perhitungan jarak serta penentuan nilai k. Selanjutnya operator KNN dihubungkan ke operator *apply model*. Sedangkan data *testing* dihubungkan ke operator *normalize* yang selanjutnya dihubungkan ke operator *apply model* untuk dilakukan prediksi antara data *training* dan data *testing*. Kemudian, operator *apply model* dihubungkan ke operator *performance classification* untuk menguji performa model dan dari operator *performance classification* dihubungkan ke *result* agar dapat melihat hasil.

3.3 *Evaluation*

Pada tahap ini, 3 data yang telah disiapkan akan dilakukan masing-masing 10 kali pengacakan data untuk setiap pembagian data *testing* dan data *training* (90:10, 80:20, 70:30, 60:40, 50:50) dengan tujuan mencari kombinasi data *training* dan data *testing* yang memberikan nilai akurasi tertinggi. *Confusion matrix* digunakan sebagai alat untuk menguji kinerja model. Terdapat tiga komponen yaitu *accuracy*, *precision* dan *recall*. *Accuracy* adalah metrik yang umum digunakan untuk mengukur performa model yaitu rasio antara data yang diprediksi dengan benar oleh model terhadap total data. Semakin tinggi *acuraccy*, semakin baik kemampuan model dalam memprediksi data dengan benar secara keseluruhan. *Precision* adalah metrik yang mengukur seberapa baik model memprediksi data positif dengan tingkat spesifikitas yang tinggi yaitu mengukur rasio antara data positif yang diprediksi dengan benar oleh model terhadap total data yang diprediksi sebagai positif. Semakin tinggi *precision*, semakin rendah kemungkinan model memprediksi data negatif sebagai positif. Sedangkan *recall* adalah metrik yang menilai sejauh mana kemampuan model dalam mengenali semua data positif yang benar-benar ada dengan mengukur proporsi data positif yang diprediksi dengan benar oleh model terhadap keseluruhan data positif yang ada. Semakin tinggi *recall*, kemungkinan besar

model akan menghindari kesalahan melewatkan data positif menjadi lebih rendah. Hasil yang didapatkan dari *dataset* 1 ditunjukkan pada Tabel 4.

Tabel 4 Hasil *Random Data* dari *Dataset* 1

Random data	Komponen	50:50	60:40	70:30	80:20	90:10
1	*Accuracy*	43,71 %	41,32 %	43,18 %	41,09 %	44,27 %
	Precision	45,45 %	41,80 %	45,39 %	43,29 %	42,06 %
	Recall	64,23 %	64,58 %	54,92 %	57,19 %	65,69 %
2	*Accuracy*	42,47 %	42,48 %	42,36 %	41,24 %	37,77 %
	Precision	44,42 %	43,88 %	44,31 %	43,47 %	40,30 %
	Recall	60,14 %	57,12 %	56,73 %	58,07 %	55,86 %
3	*Accuracy*	44,64 %	42,48 %	43,18 %	44,81 %	45,51 %
	Precision	46,73 %	46,27 %	44,25 %	45,07 %	50,00 %
	Recall	54,82 %	49,42 %	62,33 %	63,10 %	59,88 %
4	*Accuracy*	42,90 %	44,11 %	43,90 %	47,29 %	44,89 %
	Precision	46,15 %	46,32 %	43,57 %	49,74 %	43,90 %
	Recall	54,12 %	57,19 %	63,47 %	63,19 %	63,83 %
5	*Accuracy*	42,59 %	41,55 %	42,46 %	40,47 %	47,06 %
	Precision	43,11 %	42,05 %	43,18 %	41,65 %	48,21 %
	Recall	64,52 %	69,49 %	65,82 %	57,00 %	62,25 %
6	*Accuracy*	44,27 %	42,79 %	44,32 %	42,95 %	42,11 %
	Precision	45,32 %	44,91 %	44,85 %	43,01 %	43,88 %
	Recall	64,47 %	59,43 %	66,28 %	57,44 %	59,72 %
7	*Accuracy*	43,21 %	41,16 %	42,36 %	43,72 %	39,94 %
	Precision	46,13 %	44,67 %	44,80 %	45,60 %	44,95 %
	Recall	61,36 %	58,07 %	56,07 %	58,86 %	58,94 %
8	*Accuracy*	43,71 %	41,40 %	42,77 %	44,65 %	42,72 %
	Precision	45,50 %	42,24 %	43,40 %	45,30 %	46,34 %
	Recall	56,59 %	63,51 %	69,05 %	64,83 %	62,09 %
9	*Accuracy*	42,53 %	40,93 %	39,05 %	42,33 %	44,27 %
	Precision	42,38 %	41,56 %	40,09 %	45,17 %	50,52 %
	Recall	73,55 %	60,96 %	61,56 %	57,67 %	58,08 %
10	*Accuracy*	43,15 %	41,78 %	42,98 %	42,79 %	42,72 %
	Precision	45,00 %	43,79 %	42,60 %	48,87 %	48,26 %
	Recall	56,18 %	58,56 %	60,79 %	62,37 %	62,58 %

Pada Tabel 4 menunjukkan hasil *random data* dari *dataset* 1, terbukti bahwa akurasi tertinggi terjadi pada *random data* 5 dengan pembagian data *training* dan data *testing* 90:10 yang memperoleh nilai *accuracy* sebesar 47,06%. Selanjutnya untuk hasil yang didapatkan dari data 2 ditunjukkan pada Tabel 5.

Tabel 5 Hasil *Random Data* dari *Dataset* 2

Random data	Komponen	50:50	60:40	70:30	80:20	90:10
1	*Accuracy*	57,78 %	56,98 %	56,30 %	56,90 %	52,94 %
	Precision	60,19 %	58,32 %	57,84 %	59,12 %	53,76 %
	Recall	85,83 %	90,09 %	89,25 %	88,22 %	86,71 %
2	*Accuracy*	57,04 %	55,81 %	55,89 %	60,00 %	51,39 %
	Precision	57,66 %	56,73 %	57,71 %	62,43 %	53,18 %
	Recall	94,38 %	92,33 %	86,87 %	87,91 %	81,61 %
3	*Accuracy*	56,79 %	56,28 %	57,23 %	57,98 %	54,18 %
	Precision	59,07 %	58,09 %	58,06 %	59,65 %	53,63 %
	Recall	87,51 %	87,25 %	90,97 %	89,24 %	91,72 %
4	*Accuracy*	55,80 %	57,29 %	56,30 %	53,95 %	56,66 %
	Precision	57,55 %	58,83 %	58,14 %	55,73 %	59,01 %
	Recall	90,89 %	88,86 %	87,88 %	88,43 %	87,43 %
5	*Accuracy*	56,79 %	55,89 %	54,86 %	56,43 %	56,35 %
	Precision	59,36 %	57,24 %	55,54 %	57,67 %	59,22 %
	Recall	85,43 %	91,35 %	93,30 %	89,70 %	86,53 %
6	*Accuracy*	55,86 %	58,06 %	58,78 %	57,05 %	62,85 %
	Precision	56,66 %	58,86 %	60,91 %	60,44 %	64,14 %
	Recall	93,31 %	90,59 %	87,44 %	84,95 %	92,08 %
7	*Accuracy*	57,16 %	56,59 %	56,71 %	54,42 %	59,13 %
	Precision	58,16 %	58,64 %	58,35 %	56,06 %	59,93 %
	Recall	92,39 %	86,99 %	88,41 %	89,01 %	91,01 %
8	*Accuracy*	55,61 %	57,29 %	58,26 %	55,66 %	56,97 %
	Precision	58,07 %	58,49 %	60,86 %	57,90 %	60,00 %
	Recall	85,41 %	90,08 %	85,37 %	86,93 %	83,94 %
9	*Accuracy*	58,15 %	57,75 %	55,06 %	57,98 %	57,89 %
	Precision	58,90 %	59,24 %	56,29 %	60,14 %	59,14 %
	Recall	93,30 %	90,11 %	91,03 %	86,68 %	88,24 %
10	*Accuracy*	57,84 %	58,53 %	54,44 %	55,19 %	57,28 %

Random data	Komponen	50:50	60:40	70:30	80:20	90:10
	Precision	59,66 %	59,57 %	55,18 %	56,60 %	58,93 %
	Recall	88,68 %	91,01 %	90,24 %	89,32 %	87,77 %

Pada Tabel 5 menunjukkan hasil *random data* dari *dataset* 2, terbukti bahwa akurasi tertinggi terjadi pada *random data* 6 dengan pembagian data *training* dan data *testing* 90:10 yang memperoleh nilai *accuracy* sebesar 62,85%. Kemudian untuk hasil yang didapatkan dari data 3 ditunjukkan pada Tabel 6.

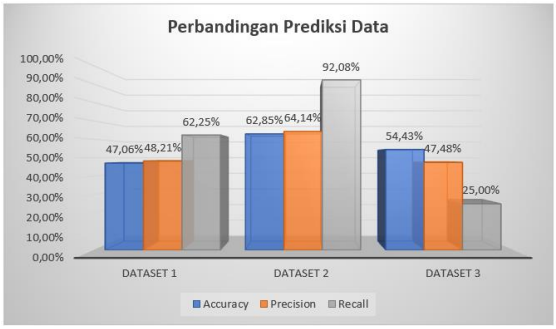
Tabel 6 Hasil *Random Data* dari *Dataset* 3

Random data	Komponen	50:50	60:40	70:30	80:20	90:10
1	*Accuracy*	51,77 %	52,48 %	50,72 %	53,02 %	47,68 %
	Precision	55,21 %	47,41 %	42,79 %	47,06 %	44,44 %
	Recall	68,20 %	18,30 %	19,91 %	24,49 %	17,28 %
2	*Accuracy*	51,58 %	50,23 %	51,14 %	51,78 %	52,63 %
	Precision	45,71 %	41,40 %	42,26 %	42,54 %	45,57 %
	Recall	23,53 %	26,64 %	25,93 %	19,59 %	24,66 %
3	*Accuracy*	52,82 %	53,57 %	53,10 %	50,85 %	49,23 %
	Precision	44,90 %	45,39 %	46,56 %	52,00 %	43,24 %
	Recall	18,06 %	22,34 %	27,98 %	16,20 %	20,78 %
4	*Accuracy*	51,33 %	52,87 %	52,17 %	50,54 %	52,63 %
	Precision	43,63 %	45,83 %	42,53 %	42,47 %	46,15 %
	Recall	25,31 %	26,55 %	26,18 %	20,88 %	20,27 %
5	*Accuracy*	52,32 %	53,72 %	50,00 %	52,71 %	51,70 %
	Precision	44,74 %	46,63 %	41,96 %	47,01 %	51,43 %
	Recall	30,22 %	26,43 %	20,98 %	21,21 %	22,78 %
6	*Accuracy*	50,46 %	51,32 %	51,24 %	50,54 %	51,08 %
	Precision	43,29 %	41,55 %	43,40 %	43,90 %	41,67 %
	Recall	33,89 %	20,34 %	23,13 %	24,08 %	16,89 %
7	*Accuracy*	51,58 %	53,10 %	53,00 %	55,19 %	51,39 %
	Precision	45,08 %	46,69 %	42,35 %	41,07 %	40,86 %
	Recall	27,99 %	21,64 %	31,86 %	26,64 %	27,14 %
8	*Accuracy*	51,46 %	51,09 %	50,10 %	52,56 %	53,25 %
	Precision	43,31 %	42,58 %	41,34 %	46,01 %	39,29 %
	Recall	28,79 %	30,70 %	16,37 %	25,60 %	24,81 %
9	*Accuracy*	54,43 %	49,53 %	54,03 %	49,61 %	48,61 %
	Precision	47,48 %	40,80 %	45,21 %	45,16 %	47,37 %
	Recall	25,00 %	20,47 %	28,10 %	22,58 %	22,22 %
10	*Accuracy*	51,70 %	52,95 %	50,00 %	51,47 %	51,39 %
	Precision	42,66 %	45,22 %	42,42 %	40,76 %	44,78 %
	Recall	26,47 %	28,10 %	21,83 %	31,39 %	20,00 %

Pada Tabel 6 menunjukkan hasil *random data* dari *dataset* 3, terbukti bahwa akurasi tertinggi terjadi pada *random data* 9 dengan pembagian data *training* dan data *testing* 50:50 yang memperoleh nilai *accuracy* sebesar 54,43%.

3.4 *Deployment*

Berdasarkan hasil *evaluation* sebelumnya, pada tahap ini akan dilakukan perbandingan antara hasil prediksi dari ketiga *dataset* tersebut. *Accuracy* prediksi tertinggi masing-masing dari ketiga data yaitu *dataset* 1, *dataset* 2 dan *dataset* 3 ditunjukkan pada Gambar 8.



Gambar 8 Grafik Perbandingan Prediksi *Dataset*

Berdasarkan Gambar 4.7 dapat dilihat hasil prediksi dari masing-masing *dataset*. Pada *dataset* 1, nilai *accuracy* tertinggi didapatkan sebesar 47,06%, *precision* 48,21%, dan *recall* 62,25%. Kemudian pada *dataset* 2, nilai *accuracy* tertinggi didapatkan sebesar 62,85%, *precision* 64,14%, dan *recall* 92,08%. Sedangkan pada *dataset* 3, nilai *accuracy* tertinggi didapatkan sebesar 47,06%, *precision* 48,21%, dan *recall* 62,25%.

4. KESIMPULAN

Berdasarkan hasil penelitian yang telah dilakukan, maka ditarik kesimpulan bahwa

implementasi algoritma *k-nearest neighbor* dalam memprediksi pergerakan harga saham mendapatkan hasil yang cukup baik dengan nilai akurasi tertinggi pada *dataset* 2 dan pada *random data* ke 6 dengan rasio data *training* dan data *testing* 90:10 serta nilai K=53 yaitu *accuracy* 62,54%, *precision* 64,14% dan *recall* 92,08%. Prediksi harga saham merupakan hal yang kompleks dan dipengaruhi oleh banyak faktor, sehingga perlu dilakukan analisis lebih lanjut dan pengembangan model yang lebih canggih untuk meningkatkan performa prediksi harga saham secara keseluruhan.

5. SARAN

Saran untuk penelitian selanjutnya, menggunakan model atau metode algoritma data mining lainnya serta menambahkan variabel yang memiliki korelasi ataupun pengaruh dalam pergerakan harga saham.

4. UCAPAN TERIMA KASIH

Terima kasih yang tulus kepada dosen pembimbing atas bimbingan, arahan, serta dukungan yang telah diberikan selama tahap penelitian ini.

DAFTAR PUSTAKA

[1] H. Z. Aikin, S. U. Sh, L. W. P. Suhartana, dan M. H. Sh, "Pengantar Hukum Perusahaan," Kencana, 2016

[2] S. A. Purnama and B. Rikumahu, "Analisis Faktor yang Mempengaruhi Harga Saham Menggunakan Metode Principal Component Analysis (Studi pada Sub Sektor Perbankan Saham LQ45 yang terdaftar di Bursa Efek Indonesia Periode 205-2019)," *e-Proceeding Manag.*, vol. 7, no. 2, pp. 5240–5247, 2020.

[3] H. Susanto and S. Sudiyatno, "Data mining untuk memprediksi prestasi siswa berdasarkan sosial ekonomi, motivasi, kedisiplinan dan prestasi masa lalu," *J. Pendidik. Vokasi*, vol. 4, no. 2, pp. 222–231, 2014, doi: 10.21831/jpv.v4i2.2547.

[4] J. Han, J. Pei, and H. Tong, "Data mining: concepts and techniques," Morgan Kaufmann, 2022.

[5] W. Yustanti, "Algoritma K-Nearest Neighbour untuk Memprediksi Harga Jual Tanah," *J. Mat. Stat. dan komputasi*, vol. 9, no. 1, pp. 57–68, 2012.

[6] B. Subandriyo, "Buku Ajar Analisis Kolerasi dan Regresi," *Diklat Stat. Tingkat Ahli BPS Angkatan XXI*, p. 31, 2020.

[7] J. Sarwono, *Metode Penelitian Kuantitatif & Kualitatif*, Pertama. Yogyakarta: graha Ilmu, 2006.

[8] M. S. Yuniarto and E. A. Sarwoko, "Implementasi Metode K-Nearest Neighbor untuk Diagnosis Kanker Kolorektal dengan Biomarker Micro-RNA," *J. Masy. Inform.*, vol. 11, no. 1, pp. 35–48, 2020.

[9] S. B. Imandoust and M. Bolandraftar, "Application of K-Nearest Neighbor (KNN) Approach for Predicting Economic Events : Theoretical Background," *Int. J. Eng. Res. Appl.*, vol. 3, no. 5, pp. 605–610, 2013.

[10] B. S. Nila Hapsari Nawangwulan, Ir Bambang Sudarsono, "Analisis Pengaruh Lahan Pertanian Terhadap Hasil Produk Tanaman Pangan Di Kabupaten Pati," *J. Geod. UUndip*, vol. 2, no. 2, pp. 127–140, 2013.

[11] R. Wirth and J. Hipp, "CRISP-DM: towards a standard process model for data mining. Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining, 29-39," *Proc. Fourth Int. Conf. Pract. Appl. Knowl. Discov. Data Min.*, no. 24959, pp. 29–39, 2000,

[12] G. Miner, J. Elder IV, A. Fast, T. Hill, R. Nisbet, and D. Delen, "Practical text mining and statistical analysis for non-structured text data applications," Academic Press, 2012.