BAB I
PENDAHULUAN

1.1 Latar Belakang

Sejak krisis tahun 1970-an, fluktuasi harga minyak mentah menjadi perhatian utama bagi para pemangku kebijakan, para analisis ekonomi, dan juga pelaku pasar. Lonjakan harga minyak dunia pada periode itu menyebabkan resesi global yang besar-besaran. Sejak itu, pasar minyak dunia terus mengalami fluktuasi harga yang dipengaruhi oleh faktor-faktor seperti permintaan global, kemampuan produksi kilang minyak, kondisi geopolitik antar negara khususnya negara produsen minyak, dan masih banyak lagi faktor lainnya (Fitri, 2023).

Minyak mentah (*crude oil*) menjadi salah satu komoditas memainkan peran yang vital dalam perekonomian global (Iftikhar dkk., 2023). Pergerakan harga minyak dunia seringkali dianggap sebagai cerminan kondisi ekonomi dunia dan perubahannya harga minyak mentah turut mempengaruhi harga-harga komoditi lainnya (Seru et al., 2023).

Salah satu standar acuan bagi harga minyak global adalah Brent Oil (Hesniati dkk., 2022). Hal ini karena sekitar dua pertiga minyak dunia dihargakan berdasarkan harga Brent Crude Oil (Charles Schwab, 2023). Brent Oil juga merupakan standar acuan harga minyak mentah bagi Indonesia. Dimana jika harga minyak mentah seperti Brent Oil mengalami kenaikan, maka membuat harga produk seperti bahan bakar minyak, aftur , dan lainnya akan ikut naik. Hal ini akan mengakibatkan inflasi naik, terutama di sektor transportasi (Rachman, 2023).

Indonesia sangat bergantung pada ketersediaan energi seperti minyak mentah yang menjadi bahan baku pembuatan bahan bakar minyak. Setiap tahun, kebutuhan energi di Indonesia terus mengalami peningkatan seiring dengan peningkatan jumlah penduduk. Peningkatan kebutuhan energi ini tidak berbanding lurus dengan jumlah produksi energi dalam negeri, khususnya minyak mentah. Hal ini mendorong pemerintah untuk melakukan impor untuk memenuhi kebutuhan energi dalam negeri yang membuat posisi Indonesia menjadi rentan terhadap kondisi pasar energi global (khususnya minyak mentah) (Sa'adah dkk., 2017).

Penting bagi kita untuk memprediksi pergerakan harga minyak global untuk mengurangi dampak yang ditimbulkan oleh fluktuasi harga minyak global dan juga membantu para pelaku pasar untuk membuat keputusan yang berkaitan dengan pasar energi. Prediksi harga minyak global memiliki juga tantangan yang besar, karena pergerakannya sangat dipengaruhi oleh beberapa faktor, seperti faktor fundamental, faktor non-fundamental, dan pengaruh dari kebijakan pasokan OPEC (Fauzannissa R. A. dkk., 2015).

Seiring berkembangnya teknologi informasi dan komunikasi, salah satu metode yang uk memprediksi pergerakan harga saham maupun harga komoditi adalah *machine learning* (Eka Patriya, 2020). *Machine learning* ebagai proses pemprogramman komputer yang memanfaatkan data nbangun model dan memungkinkan sistem untuk melakukan n tujuan mencapai kinerja yang optimal dalam mengekstraksi umpulan data (Triya dkk., 2024).

Random Forest merupakan salah satu metode *machine learning* berbasis *Decision Tree* yang menggabungkan beberapa pohon keputusan untuk membentuk model yang lebih akurat. *Random Forest* mampu untuk meningkatkan akurasi meskipun terdapat *missing value*, tahan terhadap *outlier*, dan efisien dalam penyimpanan data (Amna dkk., 2023).

Support Vector Regression adalah pengembangan dari *Support Vector Machine* untuk menyelesaikan masalah regresi. Perbedaan *Support Vector Machine* (SVM) dan *Support Vector Regression* (SVR) yaitu SVM digunakan untuk memperoleh *hyperplane* yang terbaik untuk 2 (dua) objek yang tidak terbatas jumlahnya dengan memaksimalkan margin (jarak) antara dua objek yang berbeda, sedangkan SVR digunakan untuk menemukan fungsi regresi yang memiliki deviasi tidak lebih dari margin ε dari nilai target agar fungsi tersebut tetap sederhana (Cahyono dkk., 2019).

Algoritma Random Forest dan Support Vector Regression lebih banyak digunakan dalam penelitian yang memprediksi menggunakan harga saham atau kripto. Seperti penelitian yang dilakukan oleh Saadah dan Salsabila (2021) yang berjudul "Prediksi Harga Bitcoin menggunakan Metode *Random Forest*" menyimpulkan bahwa memprediksi data bitcoin yang fluktuasinya tinggi dengan *Random Forest* dikatakan berhasil karena memberikan fitting yang sesuai dengan data sesungguhnya, serta menghasilkan nilai MAPE sebesar 1,5% atau dengan akurasi sekitar 98%. Sementara itu, penelitian yang dilakukan oleh Isra Miraltamirus dkk. (2023) yang menggunakan algoritma SVR berjudul "Prediksi Harga Saham PT Bank Syariah Indonesia Tbk menggunakan *Support Vector Regression*" yang menyimpulkan bahwa parameter terbaik adalah dengan kernel RBF, $C = 100$, $\varepsilon = 0{,}01$, dan $\gamma = 0{,}001$ dengan MAPE sebesar 0,0087 atau 0,87% atau akurasi sekitar 99,13%. Artinya kesalahan prediksi yang dihasilkan oleh model tersebut sangat kecil dan membuktikan bahwa *Support Vector Regression* menghasilkan hasil prediksi yang sangat baik.

Sebelumnya, penelitian terkait penggunaan algoritma Random Forest dan Support Vector Regression dalam memprediksi harga lebih banyak difokuskan pada instrumen keuangan seperti saham atau mata uang kripto. Sementara itu, dalam memprediksi harga komoditi energi seperti minyak mentah, pendekatan yang lebih umum digunakan lebih sering berasal dari metode deep learning seperti LSTM, ANN, dan metode statistik klasik seperti ARIMA. Penelitian yang secara khusus membandingkan analisis algoritma Random Forest dan Support Vector Regression dalam konteks prediksi harga minyak mentah masih terbatas. Selain itu, studi yang ada lebih banyak menilai performa model dari sisi kuantitatifnya menggunakan MAPE atau RMSE, tanpa memberikan gambar bagaimana model tersebut merespons fluktuasi dari harga aktual. Maka dari itu, penelitian ini bertujuan untuk memberikan perbandingan performa model, tidak hanya secara kuantitatif, namun memberikan juga analisis terhadap pola prediksi yang dihasil


kan dengan harga aktual, sehingga memberikan pemahaman yang
nan dan kekurangan dari model prediksi.

masalahan yang telah dijabarkan serta hasil penelitian terdahulu,
untuk meneliti tentang peramalan atau prediksi harga minyak
judul **"Analisis Algoritma Random Forest dan Support Vector**
alam Prediksi Harga Minyak Brent Oil (BZ=F)".

1.2 Rumusan Masalah

Penelitian ini memiliki rumusan masalah, yaitu bagaimana perbandingan analisis algoritma Random Forest dan Support Vector Regression pada prediksi harga minyak mentah Brent Crude Oil periode 03 Februari 2020 sampai 03 Februari 2023 ?

1.3 Tujuan Penelitian

Penelitian ini bertujuan untuk membandingkan analisis algoritma Random Forest dan Support Vector Regression pada prediksi harga minyak mentah Brent Crude Oil periode 03 Februari 2020 sampai 03 Februari 2023.

1.4 Landasan Teori

1.4.1 Prediksi

Menurut Widiarni dan Mustakim (2023), prediksi adalah metode peramalan yang dilakukan secara sistematis dan bertahap untuk memperhitungkan kejadian dimasa yang akan datang berdasarkan data historis dan data terkini. Tujuan utama dari prediksi bukanlah menghasilkan jawaban yang pasti, melainkan untuk mendapatkan kemungkinan yang bisa saja terjadi di masa mendatang. Prediksi juga memiliki peran sebagai input dalam proses perencanaan dan pengambilan keputusan, karena mampu memberikan gambaran mengenai potensi kejadian dalam situasi tertentu.

1.4.2 Minyak Mentah

Minyak mentah merupakan salah satu sumber energi utama yang berkontribusi lebih dari sepertiga konsumsi energi global. Guncangan harga minyak memiliki pengaruh besar terhadap aktivitas makroekonomi melalui berbagai mekanisme. Perubahan harga yang tiba-tiba dapat menciptakan ketidakpastian, yang berakibat pada menurunnya investasi bisnis serta berkurangnya output ekonomi secara keseluruhan. Mempelajari pola pergerakan minyak mentah menjadi faktor penting dalam merumuskan kebijakan ekonomi serta proses pengambilan keputusan (Băra dkk., 2024).

1.4.1.1 Minyak Brent Crude Oil

Minyak Brent Crude Oil merupakan salah satu tolak ukur utama dalam pasar minyak global, dengan sekitar dua pertiga dari seluruh kontrak minyak mentah mengacu padanya. Minyak mentah ini mencakup minyak yang berasal dari empat ladang di Laut Utara, yaitu Brent, Forties, Osberg, dan Ekofisk. Minyak mentah ini memiliki karakteristik yang ringan dan manis sehingga sangat diminati untuk diolah menjadi bahan bakar
roduk lainnya yang memiliki permintaan yang tinggi.
dakstabilan harga minyak Brent dapat memberikan dampak besar
investasi dalam sektor energi. Volatilitas harga yang tinggi dapat
nvestor untuk enggan berinvestasi dalam proyek energi terbarukan
dikarenakan adanya risiko tinggi dari ketidakpastian harga energi.
a energi yang stabil, maka akan menciptakan kondisi yang lebih

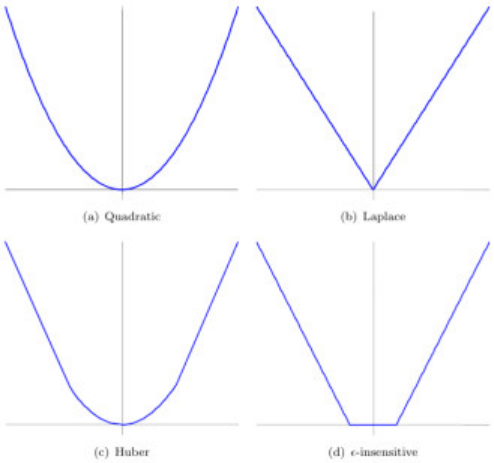
mendukung bagi investasi dalam proyek energi berkelanjutan yang sifatnya jangka panjang (Băra dkk., 2024).

1.4.3 Machine Learning

Menurut Bintoro dkk. (2024), *Machine learning* atau pembelajaran mesin merupakan suatu bidang ilmu yang memberikan kelebihan kepada sistem komputer untuk belajar dan meningkatkan kemampuannya berdasarkan data tanpa harus diprogram secara eksplisit untuk masalah/*problem* tertentu. *Machine learning* tidak hanya membantu sistem dalam membuat keputusan, tetapi juga memungkinkan prediksi tren di masa depan dengan lebih akurat. Hal ini memberikan keuntungan bagi para pemangku kepentingan untuk menggunakannya.

1.4.4 *Support Vector Regression*

Support Vector Regression (SVR) merupakan penerapan dari *Support Vector Machine* (SVM) untuk masalah regresi. Dalam regresi, *output* yang dihasilkan berupa bilangan nyata atau kontinu (Saputra dkk., 2019). Dalam permasalahan Regresi, *Suppor Vector Machine* (SVM) memperkenalkan alternatif *loss function*. *Loss function* ini harus dimodifikasi ke dalam agar mencakup ukuran dan jarak. Gambar 1 mengilustrasikan 4 (empat) kemungkinan *loss function* yang dapat digunakan.



Gambar 1 *Loss Function*

Loss function pada Gambar 1(a). setara dengan kriteria *least square error* konvensional. Sementara itu, *loss function* pada Gambar 1(b). merupakan *Laplacian loss function* yang lebih tahan terhadap outlier dibandingkan dengan *quadratic loss* gusulkan *loss function* pada Gambar 1(c). sebagai *robust loss* t optimal ketika distribusi data tidak diketahui. Namun, ketiga *loss* nghasilkan *support vector* yang tidak menyebar. Untuk mengatasi angkan *loss function* pada Gambar 1(d) sebagai pendekatan *function*, sehingga memungkinkan terbentuknya himpunan *support* ebar (Gunn, 1997).

SVR bertujuan untuk mencari *hyperplane* (garis pemisah) terbaik berupa fungsi kernel dengan membuat error sekecil mungkin dan memaksimalkan margin. Nilai error pada SVR dianggap 0 jika titik-titik masih diantara batas deviasi ($+\varepsilon$ sampai $-\varepsilon$). Suatu fungsi SVR dianggap sangat baik jika batas deviasinya mendekati 0. Pengertian support vector pada SVR adalah data training yang terletak pada batas ε dan diluar batas ε. Jika nilai ε naik, maka jumlah dari support vector akan berkurang. Fungsi regresi $f(x)$ dengan batas deviasinya sama dengan 0 merupakan fungsi regresi terbaik dan dituliskan sebagai berikut.

$$f(x) = w^T \varphi(x) + b \tag{1}$$

dengan,
$f(x)$ = fungsi prediksi
w = vektor pembobot
$\varphi(x)$ = fungsi yang memetakan x dalam suatu dimensi yang lebih tinggi
b = bias

Simbol $\varphi(x)$ menunjukkan suatu titik dalam ruang fitur F yang merupakan hasil pemetaan x didalam ruang input X. Koefisien w dan b berperan dalam meminimalkan fungsi risiko. Dengan meminimalkan fungsi risiko, suatu fungsi dapat dibuat menjadi sesederhana mungkin, sehingga kapasitasnya tetap terkontrol (Maghfiroh, 2018). Koefisien w dan b diestimasi dengan cara meminimalkan fungsi risiko (*risk function*), yang didefinisikan ke dalam persamaan berikut :

$$R(f(x)) = \frac{C}{n}\sum_{t=1}^{n} L_\varepsilon\big(y_t, f(x_t)\big) + \frac{1}{2}\|w\|^2 \tag{2}$$

dimana,

$$L_\varepsilon\big(y_t, f(x_t)\big) = \begin{cases} 0, & |y_t - f(x_t)| \le \varepsilon \\ |y_t - f(x_t)| - \varepsilon, & \text{lainnya} \end{cases} \tag{3}$$

Dengan $L_\varepsilon\big(y_t, f(x_t)\big)$ merupakan ε-*insensitive loss function* atau fungsi yang mengukur risiko empiris yang dikenai penalti jika error $|y_t - f(x_t)|$ lebih besar dari ε. Faktor $\|w\|^2$ disebut dengan regularisasi, jika meminimalkan $\|w\|^2$ maka akan diperoleh fungsi setipis mungkin , sehingga *function capacity* dapat dikontrol. Faktor lain yang dapat menunjang ketelitian fungsi dengan memperhitungkan *empirical error* dengan menggunakan ε-*insensitive loss function*. ε adalah nilai toleransi terhadap *error* (Suhartono, 2024). Parameter C dan ε merupakan *hyper*-parameter yang sudah ditentukan sebelumnya

itik berada dalam rentang $\pm\varepsilon$ atau berada dalam batas toleransi ebut dengan *feasible*. Namun jika diluar rentang tersebut maka ebut *infeasible*. Problem optimasi untuk kondisi *feasible* dapat samaan.

$$F = min\frac{1}{2}\|w\|^2 \quad (4)$$

yang memenuhi batasan $\begin{cases} y_t - f(x_t) \le \varepsilon, & t = 1,2,\dots,n \\ f(x_t) - y_t \le \varepsilon, & t = 1,2,\dots,n \end{cases}$

dimana,
y_t = nilai aktual periode ke-t
$f(x_t)$ = nilai dugaan periode ke-t

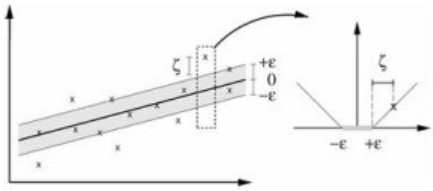
Variabel *slack* ξ dan ξ^* bisa digunakan untuk menangani masalah dimana nilai pengamatan berada diluar rentang batas ε (Smola dan Schölkopf, 2004). Pada kondisi *infeasible* dapat ditambahkan variabel *slack* (ξ_t, ξ_t^*) untuk menyelesaikan masalah *infeasible* sehingga dapat dirumuskan sebagai berikut :

$$IF = min\frac{1}{2}\|w\|^2 + C\sum_{t=1}^{n}(\xi_t + \xi_t^*) \quad (5)$$

yang memenuhi $\begin{cases} y_t - w.\varphi(x_t) - b \le \varepsilon + \xi_t, & t = 1,2,\dots,n \\ w.\varphi(x_t) + b - y_t \le \varepsilon + \xi_t^*, & t = 1,2,\dots,n \\ \xi_t, \xi_t^* \ge 0, & t = 1,2,\dots,n \end{cases}$

Konstanta C merupakan parameter yang menentukan besar penalti terhadap kesalahan prediksi yang melebihi toleransi ε. Semakin besar nilai dari C, maka model tidak mentoleransi kesalahan prediksi yang melebihi ε dan berusaha keras meminimalkan margin tersebut (Suhartono, 2024). Parameter ε digunakan untuk mengontrol lebar zona regresi. Semakin besar nilai ε maka akan semakin besar toleransi model terhadap kesalahan kecil, sehingga model tidak terlalu mencoba mengikuti data aktual terlalu dekat (Cahyono dkk., 2019).

Dalam algoritma SVR, ε ekuivalen dengan tingkat akurasi data training. Nilai ε yang kecil berhubungan dengan nilai yang tinggi pada variabel slack dengan akurasi tinggi. Sebaliknya nilai ε yang besar berhubungan dengan nilai variabel slack yang kecil dan akurasi yang rendah. Gambar 2 menunjukkan penambahan variabel *slack* ξ dan ξ^* (Suhartono, 2024). Variabel *slack* yang berada dalam rentang ε bernilai 0 (nol), tetapi semakin jauh dari batas ε, nilainya akan semakin besar. *Support vector* merupakan titik-titik dalam data latih yang terletak tepat digaris batas atau diluar batas ε. Semakin kecil nilai ε, maka akurasi model meningkat, jumlah support vector bertambah, dan nilai variabel *slack* juga semakin besar (Saputra dkk., 2019).



Gambar 2 Ilustrasi Penambahan Variabel *Slack*

Sumber : (Smola dan Schölkopf, 2004)

Problem Optimasi dari SVR adalah adalah masalah optimasi terbatas (*constrained* optimization). Pendekatan dasar yang dapat digunakan untuk menyelesaikan problem optimasi dari SVR adalah fungsi Lagrange dari fungsi objektif (yang selanjutnya disebut sebagai fungsi objektif primal) beserta kendala-kendala yang menyertainya dengan memperkenalkan serangkaian variabel dual. Solusi untuk permasalahan optimasi persamaan (5) dapat diselesaikan dengan memakai *primal Lagrangian* yang dituliskan sebagai berikut :

$$\begin{aligned} L(w,b,\xi,\xi^*,\alpha_t,\alpha_t^*,\beta_t,\beta_t^*) = \frac{1}{2}\|w\|^2 + C\left[\sum_{t=1}^{n}(\xi_t+\xi_t^*)\right] \\ -\sum_{t=1}^{n}\beta_t\left[w.\varphi(x_t)+b-y_t+\varepsilon+\xi_t\right] \\ -\sum_{t=1}^{n}\beta_t^*\left[y_t-w.\varphi(x_t)-b+\varepsilon+\xi_t^*\right] \\ -\sum_{t=1}^{n}\left[\alpha_t\xi_t+\alpha_t^*\xi_t^*\right] \end{aligned} \tag{6}$$

Persamaan (6) atau L merupakan fungsi *Lagrangian* dan $\alpha_t, \alpha_t^*, \beta_t, \beta_t^*$ adalah pengali Lagrange non-negatif (*lagrangian multiplier nonnegative*) (Smola dan Schölkopf, 2004). Persamaan (6) diminimalkan pada variabel primal w, b, ξ, ξ^* dan dimaksimalkan ke dalam bentuk *lagrangian multiplier nonnegative* $\alpha_t, \alpha_t^*, \beta_t, \beta_t^*$ (Hong, 2009). Oleh karena itu variabel dual pada persamaan (6) harus memenuhi kendala $\beta_t, \beta_t^* \geq 0$. Pencarian parameter yang optimal ditentukan melalui turunan parsial fungsi lagrange terhadap w, b, ξ, ξ^* yang dituliskan pada persamaan-persamaan berikut :

$$\frac{\partial L}{\partial w} = w - \sum_{t=1}^{n}(\beta_t-\beta_t^*)\varphi(x_t) = 0 \tag{7}$$

$$\frac{\partial L}{\partial b} = \sum_{t=1}^{n}(\beta_t^*-\beta_t) = 0 \tag{8}$$

$$\frac{\partial L}{\partial \xi} = C-\beta_t-\alpha_t = 0 \tag{9}$$

$$\frac{\partial L}{\partial \xi^*} = C - \beta_t^* - \alpha_t^* = 0 \tag{10}$$

Maka, kondisi Karush-Kuhn-Tucker teraplikasikan untuk model regresi dan dengan menggunakan persamaan (6) didapatkanlah *dual lagrangian* pada persamaan (1) dengan mensubstitusikan persamaan (7), (8), (9), dan 10 ke dalam persamaan (6). Dapat dituliskan kembali sebagai berikut.

$$\begin{aligned} L(\beta_t, \beta_t^*) = & \sum_{t=1}^{n} y_t(\beta_t - \beta_t^*) - \varepsilon \sum_{t=1}^{n} (\beta_t + \beta_t^*) \\ & -\frac{1}{2} \sum_{t=1}^{n} \sum_{u=1}^{n} (\beta_t - \beta_t^*)(\beta_u - \beta_u^*)\, K(x_t, x_u) \end{aligned} \tag{11}$$

dengan batasan,

(i) $\sum_{t=1}^{n} (\beta_t - \beta_t^*) = 0$
(ii) $0 \le \beta_t \le C, \quad t = 1,2,\dots,n\,; \quad 0 \le \beta_t^* \le C, \quad t = 1,2,\dots,n$

Lagrange Multipliers pada persamaan (11) memenuhi persamaan $\beta_t \cdot \beta_t^* = 0$. Selanjutnya hitung nilai β_t dan β_t^* untuk menentukan vektor pembobot optimal yang diharapkan dari *hyperplane* regresi yang dirumuskan sebagai berikut,

$$w = \sum_{t=1}^{n} (\beta_t - \beta_t^*)\, \varphi(x) \tag{12}$$

Maka diperoleh fungsi regresi,

$$f(x) = \sum_{t=1}^{n} (\beta_t - \beta_t^*)\, K(x_t, x_u) + b \tag{13}$$

Nilai C didefinisikan oleh peneliti, sementara $K(x_t, x_u)$ adalah *dot product* dari fungsi kernel yang didefinisikan $K(x_t, x_u) \le \varphi(x_t), \varphi(x_u)$. Nilai kernel sama dengan hasil kali dalam dua vektor x_t dan x di ruang fitur $\varphi(x)$, yaitu $K(x_t, x) = \varphi(x_t)^T \varphi(x_u)$. Secara umum, ada 3 (tiga) contoh fungsi kernel yaitu :

1. Kernel Linear : $K(x_t, x_u) = x_t^T x_u$,
2. Kernel Radial Basis Function : $K(x_t, x_u) = exp\,(-\gamma \|x_t - x_u\|^2)$,
3. Kernel Polinomial : $K(x_t, x_u) = (\gamma\ x_t^T x_u + 1)^d$, d adalah derajat polinomial.

fitur dari satu sampel dalam dataset sementara x_u adalah vektor dari dataset yang sama. Pada kernel polinomial, r merupakan d merupakan derajat polinomial. Sedangkan γ pada kernel RBF untuk *dot product* (x_t, x_u) yang memengaruhi seberapa signifikan -fitur dalam vektor fitur. Fitur adalah karakteristik yang digunakan an data, jika dalam konteks *dataset*, fitur merupakan kolom-kolom

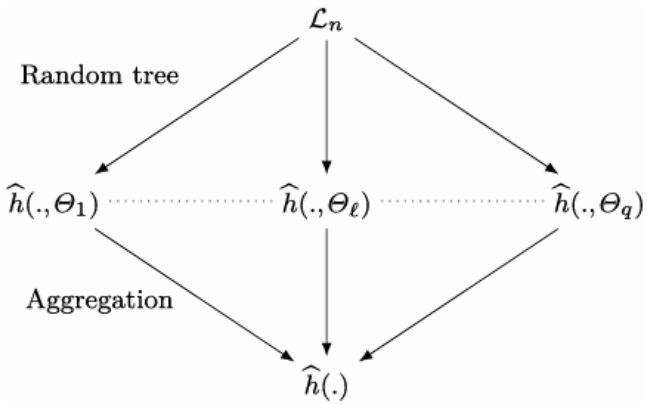
dalam tabel data. Sementara vektor fitur adalah kumpulan nilai dari fitur-fitur untuk satu sampel data yang disusun dalam bentuk vektor (Suhartono, 2024). Sampai sekarang, sulit untuk menentukan jenis kernel yang sesuai untuk pola data tertentu (Hong, 2009).

1.4.5 *Random Forest*

Perancangan model *Random Forest* telah dikembangkan sejak 1990-an. *Random Forest* mempunya kelebihan dalam hal klasifikasi ataupun regresi, serta mampu menangani variabel yang kategori dan kontinu, juga mampu menangani data yang hilang (*missing value*). *Random Forest* dikenal mampu untuk menangani data yang sifatnya kompleks, non-linear, dan mempunyai banyak fitur. Maka dari itu, *Random Forest* dianggap cocok untuk beragam jenis masalah prediksi dan pemodelan diberbagai jenis sektor (Penalun dkk., 2023).

Prinsip umum dari *Random Forest* adalah menggabungkan sejumlah pohon keputusan. Tujuannya adalah menemukan satu prediktor yang optimal dari sekumpulan gabungan prediktor (yang tidak harus optimal). Karena pohon di hutan ini diacak, maka *Random Forest* memungkinkan eksplorasi yang lebih luas dari semua kemungkinan pohon prediktor , yang pada praktiknya menghasilkan kinerja yang prediktif yang lebih baik (Annisa dkk., 2025).

Dikemukakan oleh Genuer dan Poggi (2020), misalkan $\left(h(\cdot,\theta_1),\ldots,h(\cdot,\theta_q)\right)$ adalah sekumpulan pohon prediktor dengan $\theta_1,\ldots,\theta_q$ adalah variabel acak yang identik dan independen dari $\mathcal{L}_n$. Prediktor *Random Forest* diperoleh dengan menggabungkan kumpulan pohon acak. Penggabungan (*aggregation*) digambar seperti berikut :



Gambar 3 Skema Umum *Random Forest*

Sumber : Genuer dan Poggi (2020),

ritanthia Christina Tanuwijaya dkk. (2020), prosedur dalam *Random* ıi berikut :

gambilan sampel acak berukuran n dengan pengembalian. Tahap apan *bootstrap.*

2. Dengan menggunakan sampel *bootstrap*, pohon dibangun hingga mencapai ukuran maksimum. Pembangunan pohon dilakukan dengan menerapkan *random feature selection* pada setiap proses pemilihan, yaitu q variabel penjelas yang dipilih secara acak.
3. Ulangi langkah 1 dan 2, sehingga terbentuk sebuah hutan yang terdiri atas beberapa pohon.

Selanjutnya dalam menghitung prediksi dalam *Random Forest* untuk kasus regresi adalah dengan mencari rata-rata prediksi dari setiap pohon regresi yang dituliskan kedalam persamaan berikut.

$$f(x) = \frac{1}{q}\sum_{j=1}^{q} h(x, \theta_j) \tag{14}$$

1.4.6 Normalisasi Data

Normalisasi dilakukan dengan menggunakan metode *Min-Max* untuk mengonversi data ke dalam rentang 0 hingga 1. Tujuannya adalah meningkatkan ketepatan perhitungan nurmerik dan meningkatkan akurasi dalam proses peramalan (YULI ARTINI dkk., 2024). Normalisasi umumnya diperlukan jika terdapat atribut-atribut dari dataset yang memiliki skala atau rentang yang berbeda. Sebagai contoh, terjadi ketimpangan dimana adanya data yang terlampau tinggi dan ada data yang terlampau rendah. Apabila tidak dinormalisasi maka konsekuensinya akan terjadi dilusi atau sebuah data akan dianggap tidak penting karena skalanya nilainya kecil, meski sebenarnya sama-sama penting dengan data yang memiliki nilai skala yang besar. Dengan kata lain, jika terdapat sebuah kumpulan data yang memiliki nilai-nilai yang berbeda-beda, semua derajat kepentingannya dianggap sama. Ketika terdapat banyak karakteristik pada dataset, namun nilainya bervariasi atau memiliki skala atau rentang yang berbeda, model yang dibangun dapat menghasilkan prediksi yang tidak akurat (Trivusi, 2022)

$$x'_t = \frac{(x_t - x_{min})}{x_{max} - x_{min}}, \qquad t = 1,2,\dots,n \tag{15}$$

Dimana,

x'_t = Hasil normalisasi data
x_t = Nilai data yang akan dinormalisasi
x_{min} = Nilai *minimum* (terkecil) dari dataset
x_{max} = Nilai *maximum* (terbesar) dari dataset

i Data

i bertujuan untuk mengembalikan hasil peramalan ke skala aslinya. si, hasil prediksi dapat dibandingkan secara langsung dengan data a lebih mudah diinterpretasikan dan dianalisis dalam konteks yang RTINI dkk., 2024)

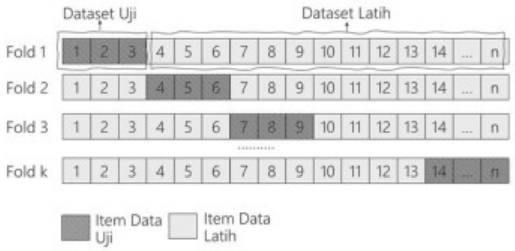
$$x_t = x_t'(x_{max} - x_{min}) + x_{min}, \qquad t = 1,2,\dots,n \tag{16}$$

Dimana,

x_t' = Hasil normalisasi data-t
x_t = Nilai data yang akan dinormalisasi
x_{min} Nilai *minimum* (terkecil) dari dataset
x_{max} = Nilai *maximum* (terbesar) dari dataset

1.4.8 *K-Fold Cross Validation*

K-fold adalah metode yang membagi dataset menjadi dua bagian, yaitu data latih dan data uji, dalam k kelompok dengan proporsi yang sama pada setiap kelompok. Proses ini dilakukan secara berulang sebanyak k kali untuk melatih dan mengevaluasi model. Nilai k merupakan bilangan bulat, dimana jika k bernilai 1, maka metode ini sama dengan Holdout, dimana data hanya dibagi satu kali menjadi data latih dan data uji tanpa adanya proses validasi silang.

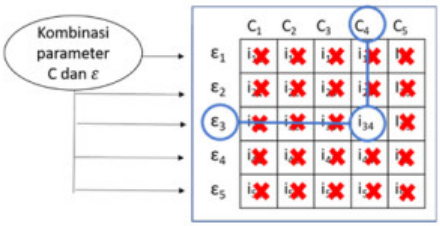


Gambar 4 Ilustrasi *K-Fold Cross Validation*

10-Fold Cross Validation merupakan salah satu metode *K-Fold* yang sering direkomendasikan untuk pemilihan model terbaik karena menghasilkan estimasi akurasi yang lebih stabil dan kurang bias dibandingkan *Cross-Validation* biasa. Dalam metode ini, data dibagi menjadi 10 bagian (*fold*) dengan ukuran yang hampir sama. Setiap iterasi, 9 *fold* digunakan untuk melatih model, sedang 1 *fold* digunakan untuk pengujian. Proses ini diulang sebanyak 10 kali sehingga setiap *fold* berkesempatan menjadi data uji, guna memastikan evaluasi model yang lebih menyeluruh (Id, 2021).

1.4.9 *Grid Search*

Algoritma Grid Search bekerja dengan mencoba berbagai kombinasi parameter secara sistematis untuk menemukan kombinasi paramater terbaik. Setiap pasangan parameter
: *grid*, lalu diuji satu per satu dengan membandingkan nilai galatnya.
· yang menghasilkan galat terkecil akan dipilih sebagai paramater
el yang akan digunakan (Saputra dkk., 2019). Gambar 5
ode *Grid Search*.



Gambar 5 Ilustrasi Metode *Grid Search*

1.4.10 Evaluasi Model

Secara umum evaluasi model digunakan untuk mengukur tingkat akurasi atau jumlah *error* pada model yang dibangun. Tujuannya yaitu menilai seberapa optimal model dalam menyelesaikan suatu permasalahan serta menguji keandalan dataset yang digunakan. Evaluasi ini juga membantu dalam menentukan apakah dataset tersebut relevan dan cukup kuat untuk diekstraksi informasinya. Selain itu, evaluasi model juga digunakan untuk membandingkan berbagai algoritma yang digunakan. Dengan melakukan perbandingan ini, dapat diketahui algoritma mana yang menghasilkan model paling bagus dan sesuai dengan kebutuhan analisis (Muttaqin, Wahyu Wijaya Widiyanto dkk., 2023). Model evaluasi yang umum digunakan diantaranya *Mean Absolute Percentage Error* (MAPE) dan *Root-Mean Square Error* (RMSE).

Mean Absolute Percentage Error (MAPE) adalah ukuran statistik yang digunakan untuk menilai akurasi prediksi dalam metode peramalan. MAPE banyak digunakan karena mudah dipahami dan diterapkan dalam berbagai bidang untuk mengukur tingkat akurasi suatu prediksi. Berikut kriteria penilaian MAPE :

Tabel 1 Kriteria Nilai MAPE

Nilai MAPE	Kriteria
$< 10\%$	Sangat Baik
$10\% - 20\%$	Baik
$20\% - 50\%$	Cukup Baik
$> 50\%$	Buruk

Dapat dipahami bahwa jika nilai dari MAPE-nya kurang dari 50%, maka masih dapat digunakan. Sebaliknya, model tidak dapat digunakan lagi jika nilai MAPE sudah diatas 50%. Adapun rumus perhitungannya adalah (Muttaqin, Wahyu Wijaya Widiyanto dkk., 2023) :

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{y_t - f(x_t)}{y_t}\right| \times 100\% \qquad (17)$$

ıyaknya data
ıktual
Peramalan

Root Mean Square Error (RMSE) merupakan metode yang digunakan untuk mengukur perbedaan antara nilai prediksi suatu model dengan nilai observasi atau label sebenarnya. RMSE diperoleh dengan mengambil akarkuadrat dari *Mean Square Error* (MSE). Rentang nilai dari RMSE dimulai dari 0 hingga ∞, dimana semakin kecil nilai RMSE, semakin akurat model dalam melakukan prediksi. Nilai RMSE yang rendah menunjukkan bahwa hasil prediksi model mendekati nilai observasi. Oleh karena itu, model dengan RMSE yang lebih kecil dianggap lebih akurat dibandingkan Model dengan RMSE yang besar. Adapun rumus perhitungannya adalah (Id, 2021) :

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_t - f(x_t))^2}{n}} \tag{18}$$

Dimana,

y_t = Nilai hasil observasi
$f(x_t)$ = Nilai data observasi
n = Jumlah/banyaknya data
t = Urutan data


PDF
Optimized using
trial version
www.balesio.com

BAB II
METODOLOGI PENELITIAN

2.1 Pendekatan dan Jenis Penelitian

Pendekatan yang digunakan dalam penelitian ini adalah pendekatan kuantitatif deskriptif dengan metode penelitian eksperimental yang bersifat prediktif. Jenis penelitian yang digunakan dalam penelitian ini adalah kuantitatif deskriptif, karena fokus utamanya pada data numerik yaitu harga penutupan harian minyak Brent Crude Oil.

2.2 Waktu dan Tempat Penelitian

Waktu dilaksanakannya penelitian ini mulai dari November 2024. Tempat penelitian yaitu di Laboratorium Big Data Prodi Ilmu Aktuaria Departemen Matematika Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Hasanuddin.

2.3 Objek Penelitian

Objek dari penelitian ini adalah harga penutupan harian minyak Brent Crude Oil.

2.4 Jenis dan Sumber Data

Jenis data yang digunakan dalam penelitian ini adalah data kuantitatif, dengan data yang diambil berupa data sekunder yaitu data harga penutupan harian minyak Brent Crude Oil selama 3 tahun, periode 03 Februari 2020 sampai dengan 03 Februari 2023. Sumber data dalam penelitian ini diperoleh dari salah satu situs keuangan yaitu *Yahoo Finance*. Alat bantu penelitian yang digunakan adalah *Google Colaboratory*.

2.5 Metode Pengumpulan Data

Metode pengumpulan data dalam penelitian ini yaitu berupa studi literatur sejenis yang artinya mengumpulkan data dengan mencari referensi dari buku, skripsi, jurnal, serta artikel sejenis dari penelitian yang telah dilakukan sebelumnya dan berkaitan dengan model algoritma *Random Forest* dan *Support Vector Regressor*.

2.6 Metode Analisis Data

Dalam penelitian ini, analisis data dilakukan menggunakan *Support Vector Regression* (SVR) dan *Random Forest*. Langkah-langkahnya adalah sebagai berikut :

a. *Support Vector Regression* (SVR)

Tahapan analisis data menggunakan *Support Vector Regression* (SVR) adalah sebagai berikut :

data harga harian minyak Brent Crude Oil (BZ=F) dari situs *Yahoo*
03 Februari 2020 sampai 03 Februari 2023.
iabel penelitian yang akan diteliti yaitu harga penutupan (*close*)
ude Oil.
gecekan *missing value* terhadap data harga penutupan (*close*)
ude Oil.

4. Membuat fitur X (previous *close*) dan Y (*close*)
5. Melakukan normalisasi data dengan *MinMaxScaler*.
6. Melakukan split data dengan membagi data menjadi data latih (train) dan data uji (test).
7. Menerapkan metode *Grid Search* pada metode *Support Vector Regression* untuk menemukan kombinasi parameter terbaik.
8. Melakukan training dan testing model dengan *Support Vector Regression* pada data latih (*train*) dan data uji (*test*) dengan menggunakan kombinasi parameter terbaik yang diperoleh dari metode *Grid Search* untuk mendapatkan nilai prediksinya.
9. Melakukan denormalisasi data untuk mengembalikan nilai data ke ukuran atau skala awal.
10. Menggabungkan data latih (train) dan data uji (test) untuk mengevaluasi data secara keseluruhan.
11. Mengevaluasi kinerja model menggunakan *Mean Absolute Percentage Error* (MAPE) dan *Root-Mean Squared Error* (RMSE).
12. Memvisualisasikan hasil prediksi yang telah dilakukan menggunakan *Support Vector Regression* (SVR) dengan line-plot.
13. Melakukan Analisis pada harga minyak Brent Oil menggunakan Model *Support Vector Regression* (SVR).

b. *Random Forest*

Tahapan analisis data menggunakan *Random Forest* adalah sebagai berikut :

1. Mengumpulkan data harga harian minyak Brent Crude Oil (BZ=F) dari situs Yahoo Finance periode 03 Februari 2020 sampai 03 Februari 2023.
2. Menentukan variabel penelitian yang akan diteliti yaitu harga penutupan (*close*) minyak Brent Crude Oil.
3. Melakukan pengecekan missing value terhadap data harga penutupan (*close*) minyak Brent Crude Oil.
4. Membuat fitur X (Previous *close*) dan Y (*close*)
5. Melakukan normalisasi data dengan *MinMaxScaler*.
6. Melakukan split data dengan membagi data menjadi data latih (train) dan data uji (test).
7. Menerapkan metode *Grid Search* untuk menemukan kombinasi parameter terbaik.
8. Melakukan training dan testing model dengan *Random Forest* pada data latih (train) dan data uji (test) dengan menggunakan kombinasi parameter terbaik yang diperoleh dari metode *Grid Search* untuk mendapatkan nilai prediksinya.
9. Melakukan denormalisasi data untuk mengembalikan nilai data ke ukuran atau skala awal.

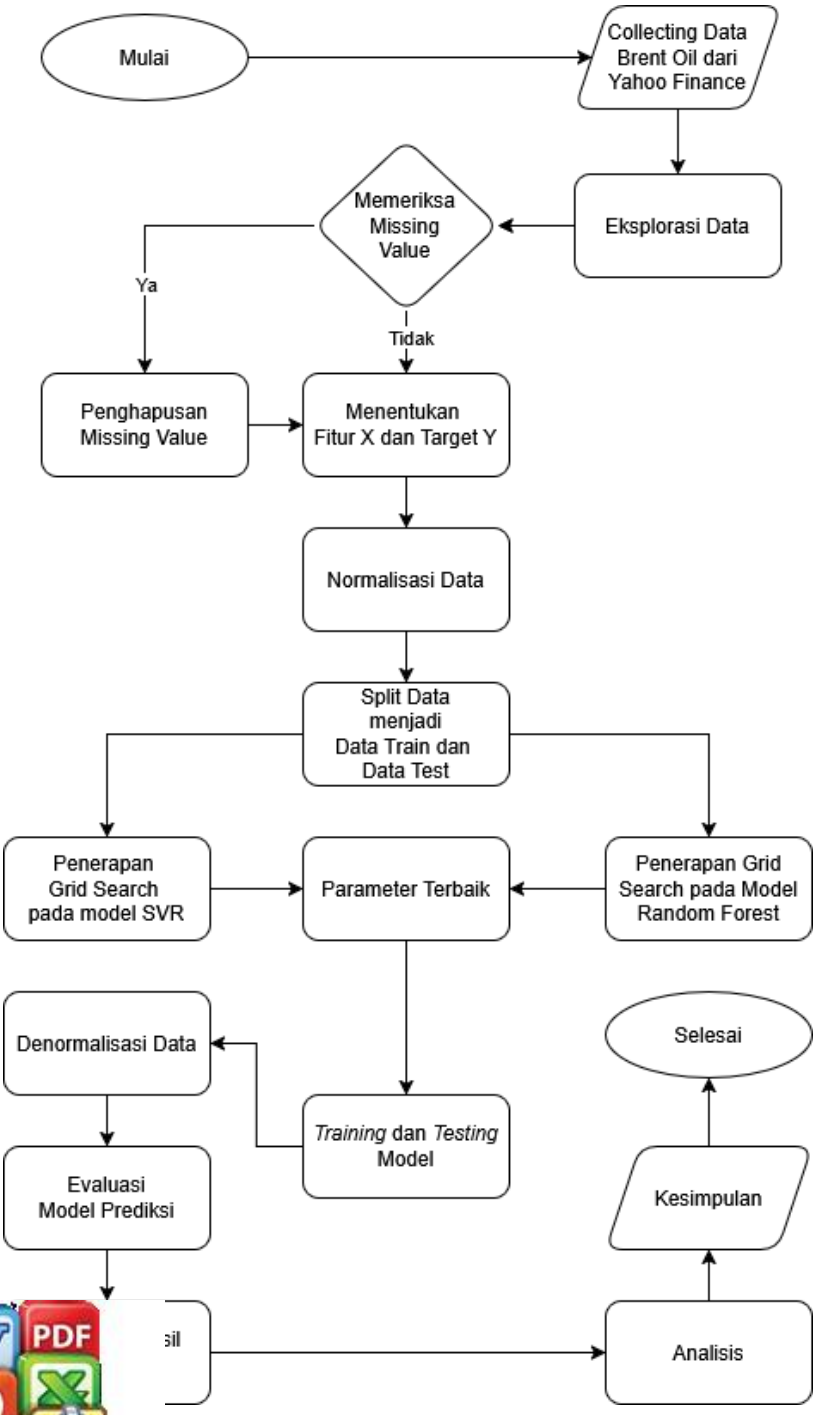
ı data latih (train) dan data uji (test) untuk mengevaluasi data secara

inerja model menggunakan *Mean Absolute Percentage Error* ɔt-Mean Squared Error (RMSE).

an hasil prediksi yang telah dilakukan menggunakan *Random* ne-plot.

13. Melakukan Analisis pada harga minyak Brent Oil menggunakan Model *Random Forest.*

2.7 Alur Kerja



Gambar 6 Alur Kerja