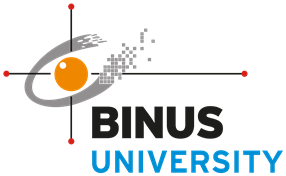
****

**LAPORAN AKHIR PROYEK MACHINE LEARNING**

**STOCKPEDIA : WEB APP PREDIKSI SAHAM MARKET CAP INDONESIA**

**Disusun Oleh:   
1. Bryan Sereno Sutanto - 2702245141 - angkatan B27**

**2. Peter Abraham Besinga - 2702253761 - angkatan B27**

**3. Natanael Timotius Thung - 2702242000 - angkatan B27**

**4. Hanzel Octavian Lesmana-2702241162-angkatan B27**

**UNIVERSITAS BINA NUSANTARA 2025**

**DAFTAR ISI**

| **DAFTAR ISI** | | 2 |
| --- | --- | --- |
| **BAB 1. PENDAHULUAN** | | 3 |
| * 1.1 Latar Belakang | | 3 |
| * 1.2 Rumusan Masalah | | 4 |
| * 1.3 Tujuan Program | | 4 |
| * 1.4 Manfaat Program | | 4 |
| **BAB 2. TINJAUAN PUSTAKA** | | 5 |
| * 2.1 Investment | | 5 |
| * 2.2 Supervised Machine Learning | | 6 |
| * 2.3 GridSearchCV Algorithm | | 9 |
| * 2.4 Ensemble Learning | | 9 |
| * 2.5 Flask App | | 11 |
| * 2.6 Metrics in Regression Algorithm | | 11 |
| * 2.7 Kesimpulan | | 12 |
| **BAB 3. TAHAP PELAKSANAAN** | | 13 |
| * 3.1 Data Gathering | | 13 |
| * 3.2 Data Preprocessing | | 13 |
| * 3.3 Model Training | | 14 |
| * 3.4 Model Testing | | 15 |
| * 3.5 Metrics & Evaluation | | 16 |
| * 3.6 Model Deployment | | 16 |
| **BAB 4. EVALUASI DAN HASIL** | | 17 |
| **LAMPIRAN** | | 20 |
| **DAFTAR PUSTAKA** | | 23 |

**BAB 1. PENDAHULUAN**

* 1. **Latar Belakang**

Saham merupakan salah satu instrumen investasi yang merepresentasikan kepemilikan seseorang atau suatu entitas dalam sebuah perusahaan. Dengan memiliki saham, investor atau pemegang saham berhak atas keuntungan yang diperoleh perusahaan, sesuai dengan proporsi jumlah saham yang dimiliki *(Bodie, Kane, & Marcus, 2018)*. Pasar saham sendiri menjadi salah satu pilihan investasi yang menjanjikan karena menawarkan potensi keuntungan yang tinggi dengan tingkat risiko yang relatif dapat dikelola.

Terdapat dua cara utama untuk memperoleh keuntungan dari investasi saham. Pertama adalah capital gain, yaitu keuntungan yang diperoleh dari selisih harga jual dan harga beli saham ketika harga saham mengalami kenaikan. Kedua adalah dividen, yaitu pembagian keuntungan perusahaan kepada pemegang saham, sebagai bentuk imbal hasil atas dana yang telah diinvestasikan ke dalam perusahaan.

Meskipun menawarkan potensi keuntungan yang besar, pasar saham juga dikenal dengan tingkat volatilitasnya yang tinggi. Harga saham dapat berubah dengan sangat cepat dan dipengaruhi oleh berbagai faktor, seperti kondisi ekonomi, laporan keuangan perusahaan, kebijakan pemerintah, serta sentimen pasar yang dinamis dan mudah berubah *(Fama, 1970)*.

Berikut adalah beberapa risiko utama dalam berinvestasi saham:

1. Risiko pasar – Terjadi akibat fluktuasi harga saham secara umum karena ketidakstabilan pasar.
2. Risiko likuiditas – Berkaitan dengan seberapa mudah saham dapat diperjualbelikan. Jika tidak ada pembeli, maka saham menjadi sulit untuk dijual.
3. Risiko fundamental perusahaan – Muncul akibat kondisi internal perusahaan, seperti hutang berlebih, manajemen buruk, atau kinerja keuangan yang menurun.
4. Risiko suku bunga – Ketika suku bunga naik, investor cenderung beralih ke instrumen lain yang dianggap lebih aman, sehingga harga saham bisa turun.
5. Risiko inflasi – Inflasi yang tinggi dapat menggerus daya beli dan mempengaruhi pendapatan perusahaan, yang berujung pada penurunan harga saham.
6. Risiko psikologis – Banyak investor terjebak oleh emosi, seperti rasa takut tertinggal (*fear of missing out* / FOMO), yang bisa menyebabkan keputusan investasi yang impulsif dan tidak rasional.

Oleh karena itu, memahami risiko-risiko tersebut sangat penting sebelum terjun ke dunia saham. Untuk membantu meminimalisasi risiko dan meningkatkan peluang keuntungan, kami menawarkan sebuah algoritma yang mampu memprediksi harga saham. Algoritma ini diharapkan dapat menjadi alat bantu yang berguna, terutama bagi investor pemula, dalam mengambil keputusan investasi yang lebih bijak dan berbasis data.

* 1. **Rumusan Masalah**

Berikut adalah rumusan masalah dari program kami :

1. Algoritma apa yang efektif dalam memprediksi saham ?
2. Apa yang harus diperhatikan dalam pembuatan algoritma agar hasil prediksi mendekati sempurna ?
3. Supervised Machine Learning apa yang lebih efektif dalam prediksi saham, regresi polinomial atau linear ?
4. Apakah algoritma machine learning masih efektif digunakan dalam melakukan prediksi saham ?  
   1. **Tujuan Program**

Berikut adalah tujuan dari program kami :

1. Melakukan pengembangan algoritma prediksi saham dengan machine learning
2. Menyediakan alat bantu untuk investor yang ingin belajar
3. Mengurangi resiko dalam mengalami kerugian akibat pasar saham
   1. **Manfaat Program**

Dengan adanya stockpedia ini tentunya kami mengharapkan agar masyarakat secara luas dapat belajar saham dengan aman dan nyaman.membantu masyarakat yang ingin belajar saham serta mebantu dalam pembelajaran dalam teknikal.membantu pemain saham dalam pengambilan keputusan.agar tidak terjadi kerugian dan dapat menghasilkan keuntungan yang maksimal.

**BAB 2. TINJAUAN PUSTAKA**

**2.1. Investment** Investasi merupakan salah satu aktivitas keuangan yang bertujuan untuk mengalokasikan dana pada aset tertentu dengan harapan memperoleh keuntungan di masa depan. Dalam dunia keuangan, terdapat berbagai instrumen investasi, seperti properti, obligasi, reksa dana, dan saham. Pemilihan instrumen investasi bergantung pada tingkat risiko, tujuan finansial, serta jangka waktu investasi yang diinginkan oleh investor.

Salah satu bentuk investasi yang paling populer adalah investasi di pasar saham. Saham menjadi pilihan banyak investor karena potensi imbal hasil yang tinggi dibandingkan dengan instrumen investasi lainnya. Namun, investasi saham juga memiliki risiko yang harus dipahami dengan baik agar investor dapat mengambil keputusan yang tepat.

Pada bagian berikutnya, akan dibahas lebih lanjut mengenai pengertian saham, pasar saham sebagai tempat transaksi investasi saham, serta tujuan utama dari aktivitas di dalamnya :

**2.1.1. Pengertian Saham**

Saham merupakan bukti kepemilikan modal seseorang atau badan usaha dalam sebuah perusahaan atau Perseroan Terbatas. Pemegang saham berhak atas pendapatan dan aset perusahaan serta memiliki hak untuk menghadiri Rapat Umum Pemegang Saham (RUPS). Di pasar sekunder, harga saham mengalami fluktuasi berdasarkan permintaan dan penawaran, yang dipengaruhi oleh berbagai faktor seperti kinerja perusahaan, kondisi industri, serta aspek makro ekonomi seperti suku bunga, inflasi, dan stabilitas politik.

**2.1.2 Pengertian Pasar Saham**

Pasar saham adalah tempat terjadinya transaksi jual beli saham yang melibatkan berbagai pihak seperti investor, spekulan, dan pemerintah. Pasar saham juga menjadi sarana perdagangan instrumen keuangan jangka panjang seperti saham, obligasi, dan reksa dana.

Selain itu, pasar saham berfungsi sebagai sumber pendanaan bagi perusahaan dan pemerintah, serta sebagai instrumen investasi bagi para pemilik modal. Untuk mengoptimalkan investasi, investor perlu melakukan analisis fundamental dan teknikal guna mengidentifikasi peluang serta memitigasi risiko.

**2.1.3 Tujuan Saham dan Pasar Saham**

Secara umum, ada dua tujuan utama dalam aktivitas pasar saham:

1. Mendapatkan Keuntungan – Investor dan spekulan berusaha memperoleh keuntungan maksimal melalui strategi investasi yang baik.
2. Pengambilan Keputusan Berkelanjutan – Investor yang cerdas mempertimbangkan faktor risiko dan jangka panjang dalam membuat keputusan investasi.

Pihak-pihak yang berperan dalam pasar saham antara lain investor, spekulan, emiten, penjamin emisi, serta pialang saham. Pemahaman yang baik mengenai pasar saham sangat penting bagi para investor untuk mengelola portofolio secara optimal dan mengurangi potensi risiko.

**2.2. Supervised Machine Learning**

Supervised learning adalah kategori [machine learning](https://cloud.google.com/learn/what-is-machine-learning) yang menggunakan set data berlabel untuk melatih algoritma guna memprediksi hasil dan mengenali pola. Tidak seperti [unsupervised learning](https://cloud.google.com/discover/what-is-unsupervised-learning), algoritma supervised learning diberi set data pelatihan berlabel untuk mempelajari hubungan antara input dan output.

Algoritma supervised machine learning memudahkan organisasi membuat model kompleks yang dapat membuat prediksi akurat. Oleh karena itu, model ini banyak digunakan di berbagai industri dan bidang, termasuk layanan kesehatan, pemasaran, jasa keuangan, dan lainnya.

**2.2.1. Cara Kerja Supervised Machine Learning**

Dalam supervised learning, data yang digunakan telah diberi label, yang berarti setiap data terdiri dari pasangan input (fitur) dan output yang benar (label). Algoritma menganalisis sejumlah besar data yang mengandung kedua komponen ini untuk mempelajari pola yang dapat digunakan dalam membuat inferensi terhadap nilai output yang diinginkan. Dengan demikian, ketika model diterapkan pada data baru, ia dapat melakukan prediksi berdasarkan pola yang telah dipelajarinya.

Sebagai contoh, apabila tujuan pembelajaran adalah mengenali gambar pohon, maka dataset yang digunakan harus berisi berbagai contoh gambar pohon beserta nama spesiesnya sebagai label. Algoritma kemudian akan menganalisis karakteristik khas yang dimiliki setiap jenis pohon berdasarkan data yang telah dilabeli. Setelah model dilatih, model tersebut dapat diuji dengan menampilkan gambar pohon yang belum pernah dilihat sebelumnya dan meminta model untuk mengidentifikasi spesiesnya.

Apabila model memberikan prediksi yang tidak akurat, proses pelatihan dapat terus dilakukan dengan menyesuaikan parameter dan menambahkan lebih banyak contoh data guna meningkatkan akurasi serta meminimalkan tingkat kesalahan (error). Setelah model melewati tahap pelatihan dan pengujian, model tersebut dapat digunakan untuk melakukan prediksi terhadap data baru berdasarkan pola yang telah dipelajari sebelumnya.

**2.2.2. Jenis-jenis Algoritma dalam Supervised Machine Learning**

Supervised machine learning memiliki berbagai jenis algoritma yang digunakan untuk tugas prediksi dan klasifikasi. Setiap algoritma memiliki pendekatan yang berbeda dalam mempelajari pola dari data latih serta menghasilkan output yang akurat. Berikut adalah beberapa algoritma yang umum digunakan dalam supervised learning beserta cara kerjanya:

1. **Linear Regression**

Regresi linear merupakan teknik yang digunakan untuk memodelkan hubungan antara variabel input dan output dengan menggunakan garis lurus. Algoritma ini sering diterapkan dalam kasus prediksi nilai kontinu, seperti perkiraan harga rumah atau suhu udara.

Model regresi linear bekerja dengan mencari garis terbaik yang dapat mewakili hubungan antara variabel input dan output berdasarkan data latih. Semakin baik garis yang terbentuk, semakin akurat hasil prediksi yang diperoleh.

1. **Logistic Regression**

Regresi logistik digunakan dalam masalah klasifikasi, yaitu ketika output yang diprediksi berbentuk kategori, seperti "ya" atau "tidak". Berbeda dengan regresi linear yang memprediksi nilai kontinu, regresi logistik menggunakan fungsi sigmoid untuk mengubah hasil prediksi menjadi probabilitas.

Teknik ini banyak diterapkan dalam analisis risiko kredit, diagnosis medis, dan deteksi spam. Dengan menggunakan ambang batas tertentu, model dapat menentukan kategori mana yang paling sesuai berdasarkan probabilitas yang dihasilkan.

1. **K-Nearest Neighbors (KNN)**

Algoritma K-Nearest Neighbors (KNN) bekerja dengan cara membandingkan data baru dengan data yang telah tersedia dalam dataset. Model ini menentukan kelas suatu data berdasarkan mayoritas tetangga terdekatnya.

Sebagai contoh, jika mayoritas dari lima tetangga terdekat suatu titik data tergolong dalam kategori A, maka data tersebut akan diklasifikasikan sebagai kategori A. Teknik ini banyak digunakan dalam sistem rekomendasi, pengenalan pola, serta deteksi anomali.

1. **Support Vector Machine (SVM)**

Support Vector Machine (SVM) merupakan algoritma yang digunakan untuk klasifikasi dan regresi dengan mencari hyperplane terbaik yang dapat memisahkan kelas dalam ruang multidimensi. Algoritma ini bekerja dengan menemukan batas optimal yang memisahkan kategori data dengan margin terbesar.

SVM sering diterapkan dalam pengenalan wajah, klasifikasi teks, serta bioinformatika. Dengan kemampuannya dalam menangani data berdimensi tinggi, SVM dapat memberikan hasil yang akurat meskipun jumlah data yang digunakan tidak terlalu besar.

1. **Polynomial Regression**

Secara matematis, regresi polinomial adalah model regresi yang melihat hubungan antara variabel independen (x) dan variabel dependen (y) dengan menggunakan fungsi polinomial sehingga dihasilkan suatu grafik yang tidak berbentuk lurus seperti grafik regresi linear. Dalam penerapannya dalam algoritma machine learning, regresi polinomial sendiri tergolong dalam supervised machine learning yang berarti hasil dari prediksi model dipandu oleh data berlabel, yaitu pasangan input (𝑥x) dan output (𝑦y). Dengan kata lain, model dilatih menggunakan dataset yang sudah memiliki label, sehingga model dapat belajar pola hubungan antara variabel independen dan dependen. Setelah proses pelatihan selesai, model dapat digunakan untuk memprediksi nilai baru berdasarkan pola yang telah dipelajari.

Dalam implementasinya, regresi polinomial sering digunakan untuk memodelkan hubungan non-linear dalam data yang tidak dapat dijelaskan dengan regresi linear sederhana. Dengan menambahkan derajat polinomial yang lebih tinggi, model dapat menangkap pola yang lebih kompleks. Namun, perlu diperhatikan bahwa jika derajat polinomial terlalu tinggi, model bisa mengalami overfitting, di mana model terlalu menyesuaikan diri dengan data pelatihan dan kurang mampu melakukan generalisasi terhadap data baru. Oleh karena itu, penting untuk melakukan validasi model dan menerapkan teknik regulasi seperti Ridge Regression atau Lasso Regression guna menghindari kompleksitas yang berlebihan.

**2.3 GridSearchCV Algorithm**

GridSearchCV *(Grid Search Cross Validation)* adalah teknik optimasi hyperparameter yang digunakan untuk menemukan kombinasi parameter terbaik dalam model machine learning. Teknik ini bekerja dengan cara melakukan pencarian secara sistematis melalui seluruh kombinasi parameter yang telah ditentukan sebelumnya, kemudian mengevaluasi performa model untuk setiap kombinasi menggunakan metode cross-validation. GridSearchCV bekerja dengan membuat kisi-kisi atau *grid* dari semua kemungkinan parameter yang user ingin uji. Cross-validation membagi dataset menjadi beberapa subset (fold), di mana setiap fold akan digunakan sebagai data validasi sementara fold lainnya digunakan sebagai data pelatihan.

**2.3.1. Keuntungan GridSearchCV Algorithm**

Berikut adalah keuntungan penggunaan GridSearchCV Algorithm untuk menentukan hyperparameter :

* Pencarian yang komprehensif dan sistematis terhadap semua kombinasi parameter
* Menggunakan cross-validation untuk memberikan estimasi performa yang lebih robust
* Mudah diimplementasikan dan dapat digunakan dengan berbagai algoritma machine learning
* Memberikan hasil yang dapat direproduksi

**2.4 Ensemble Learning**

Ensemble learning adalah teknik dalam machine learning yang menggabungkan prediksi dari beberapa model untuk menghasilkan prediksi yang lebih akurat dan robust dibandingkan dengan menggunakan model tunggal. Konsep dasar dari ensemble learning adalah bahwa dengan menggabungkan beberapa model yang berbeda, kelemahan dari satu model dapat dikompensasi oleh kekuatan model lainnya.

### **2.4.1. Prinsip Dasar Ensemble Learning**

Ensemble learning bekerja berdasarkan prinsip "wisdom of crowds" atau kebijaksanaan kelompok, di mana kombinasi dari beberapa pendapat atau prediksi cenderung lebih akurat daripada pendapat individual. Dalam konteks machine learning, ini berarti bahwa rata-rata dari prediksi beberapa model seringkali lebih baik daripada prediksi dari model terbaik secara individual.

Ensemble learning efektif karena berbagai model dapat memiliki bias dan varians yang berbeda. Ketika model-model ini digabungkan, bias dan varians keseluruhan dapat berkurang, sehingga menghasilkan prediksi yang lebih stabil dan akurat.

### **2.4.2. Jenis-jenis Ensemble Learning**

1. **Bagging *(Bootstrap Aggregating)***

Bagging adalah teknik ensemble yang melatih beberapa model dengan menggunakan subset yang berbeda dari dataset pelatihan. Setiap subset dibuat dengan cara bootstrap sampling, yaitu pengambilan sampel dengan pengembalian dari dataset asli. Model yang paling terkenal menggunakan teknik bagging adalah Random Forest, yang menggabungkan banyak decision tree.

1. **Boosting**

Boosting adalah teknik ensemble yang melatih model secara berurutan, di mana setiap model baru fokus pada kesalahan yang dibuat oleh model sebelumnya. Contoh algoritma boosting yang populer adalah AdaBoost, Gradient Boosting, dan XGBoost. Teknik ini sangat efektif untuk mengurangi bias dan meningkatkan akurasi prediksi.

1. **Stacking**

Stacking adalah teknik ensemble yang menggunakan meta-learner untuk mempelajari cara terbaik menggabungkan prediksi dari beberapa base model. Base model dilatih pada dataset pelatihan, kemudian prediksi mereka digunakan sebagai input untuk meta-learner yang akan menghasilkan prediksi final.

1. **Voting**

Voting adalah teknik ensemble yang menggabungkan prediksi dari beberapa model dengan cara "voting" atau pemungutan suara untuk menghasilkan prediksi final. Terdapat dua jenis voting utama:

* Hard Voting *(Majority Voting)*: Setiap model memberikan satu suara untuk kelas prediksinya, dan kelas dengan suara terbanyak menjadi prediksi final.
* Soft Voting *(Weighted Voting)*: Menggunakan probabilitas atau confidence score dari setiap model dengan cara menghitung rata-rata probabilitas dari semua model. Kelas dengan probabilitas tertinggi menjadi prediksi final. Teknik ini umumnya memberikan hasil yang lebih baik karena mempertimbangkan tingkat keyakinan setiap model.

**2.5 Flask App**

Flask adalah framework web mikro yang ditulis dalam bahasa Python. Flask dirancang untuk membuat aplikasi web dengan cepat dan mudah, dengan kemampuan untuk dikembangkan menjadi aplikasi yang lebih kompleks. Dalam konteks machine learning, Flask sering digunakan untuk membuat web application yang dapat melayani model machine learning sebagai API atau interface pengguna.

Flask memiliki beberapa karakteristik utama yang membuatnya populer di kalangan developer. Pertama, Flask bersifat lightweight dan minimalis, menyediakan komponen dasar yang diperlukan untuk membuat aplikasi web tanpa banyak overhead. Kedua, Flask sangat fleksibel dan dapat dikustomisasi sesuai kebutuhan, memungkinkan developer untuk memilih komponen dan library tambahan yang diperlukan.

Flask juga memiliki learning curve yang relatif rendah, membuatnya mudah dipelajari oleh developer yang baru memulai dengan web development. Selain itu, Flask memiliki dokumentasi yang lengkap dan community yang aktif, sehingga mudah untuk menemukan solusi ketika menghadapi masalah.

**2.6 Metrics in Regression Algorithm**

Dalam machine learning, algoritma regresi memiliki beberapa metrik evaluasi yang dapat digunakan untuk mengukur seberapa baik model dapat memprediksi nilai kontinu berdasarkan data input. Pemilihan metrik yang tepat sangat penting karena akan mempengaruhi interpretasi kinerja model serta proses perbaikannya. Beberapa metrik umum yang sering digunakan dalam regresi antara lain:

1. **Mean Absolute Error (MAE)**  
    MAE mengukur rata-rata dari selisih absolut antara nilai prediksi dan nilai aktual. Metrik ini menunjukkan seberapa besar kesalahan secara rata-rata tanpa mempertimbangkan arah kesalahan (positif atau negatif). Semakin kecil nilai MAE, semakin baik performa model.
2. **Mean Squared Error (MSE)**  
    MSE menghitung rata-rata kuadrat dari selisih antara nilai prediksi dan nilai aktual. Metrik ini memberikan penalti yang lebih besar untuk kesalahan yang besar karena perhitungan kuadrat. MSE sering digunakan ketika kita ingin sangat memperhatikan kesalahan besar.
3. **Root Mean Squared Error (RMSE)**  
    RMSE adalah akar kuadrat dari MSE. Metrik ini memberikan interpretasi kesalahan dalam satuan yang sama dengan target aslinya, sehingga lebih mudah dipahami dalam konteks domain data.
4. **R-squared (R²)**  
    R² mengukur seberapa besar proporsi variabilitas dalam data target yang dapat dijelaskan oleh model. Nilai R² berkisar antara 0 hingga 1, di mana nilai mendekati 1 menunjukkan bahwa model mampu menjelaskan sebagian besar variasi dalam data target.

Pemahaman dan penggunaan metrik-metrik ini sangat penting untuk mengevaluasi dan membandingkan performa berbagai model regresi secara objektif sehingga menghasilkan evaluasi akurat suatu model.

**2.7 Kesimpulan**

Berdasarkan tinjauan diatas, kami memutuskan untuk menggunakan ensemble model dengan gabungan linear, polynomial, dan random forest regressor. Untuk model ensemble-nya sendiri, kami menggunakan voting regressor dengan default param, yaitu *hard-voting*. Flask app kami gunakan sebagai metode untuk menunjukan grafik kami dengan bantuan HTML, CSS, dan Javascript. Kami juga menggunakan framework seperti *AoS* untuk mempercantik website deployment kami.

**BAB 3. TAHAP PELAKSANAAN**

**3.1 Data Gathering**

Tahap pengumpulan data merupakan fondasi awal dari proses pengembangan model prediksi harga saham. Pada tahap ini, data historis harga saham dikumpulkan dari sumber terpercaya, yaitu Yahoo Finance, kami mengambil menggunakan pustaka bernama yfinance. Data yang kami ambil yaitu berupa: Open,High, Low, Close, Volume, dan Adj close. selain itu, data dilengkapi dengan waktu dan kode perusahaan untuk membedakan data dari perusahaan yang berbeda. Data ini nantinya akan diubah ke bentuk DataFrame agar dapat dilihat apakah data relevan dan bisa digunakan.

**3.2 Data Preprocessing**

Setelah melakukan gathering dan mengetahui NULL value pada setiap individu data, kami melakukan proses pembersihan data agar data yang digunakan menghasilkan prediksi yang maksimal. Proses preprocessing dimulai dengan melakukan remove data yang lebih dari 50% isinya NULL dan melakukan proses *imputing* untuk data yang memiliki data NULL lebih sedikit dari 50%. Dalam kasus kami, stock 'GOTO.JK' dan ‘PRAY.JK’ merupakan data yang kami buang sedangkan ‘BHAT.JK’ merupakan data yang perlu dilakukan imputing. Metode yang kami pilih adalah *backward fill (bfill) dan forward fill (ffill)*. Data yang sudah bersih kemudian di-ekspor sebagai file csv dengan nama “stock\_data\_reshaped.csv”. Proses preprocessing kemudian dipecah menjadi dua untuk mengakomodasi dua model yang hendak kami pakai :

1. Pada dataset pertama, kami melakukan proses pembersihan terhadap data yang akan digunakan untuk training model ensemble. Setelah loading data csv, kami melakukan encoding kolom kategorikal ke kolom seperti *Ticker*. Kami melakukan ini menggunakan *LabelEncoder* agar dapat diolah secara maksimal dengan algoritma yang hendak kami pakai. Untuk mengecek nilai apa yang dienkode oleh fungsi *LabelEncoder* ke *Ticker* kami, kami melakukan pengecekan dengan *mapping\_ticker.* Dataset pertama disimpan dalam variabel *“training\_data”.* Kolom-kolom yang kami prioritaskan adalah *“Price” ,”Days” ,”Open” , “High” , “Low” ,”Volume” ,“Ticker” ,“Close” ,*dan *“Ticker\_encoded”*. Setelah preprocessing ini, data dibagi menjadi data pelatihan dan dana pengujian, kami melakukan pembagian dengan perbandingan 80 train : 20 test menggunakan fungsi *train\_test\_split*. Data-data ini disimpan dalam variabel *“X\_train” , “y\_train” ,”X\_test" ,* dan *“y\_test”*.
2. Pada dataset kedua, kami melakukan hal yang sama tetapi ada sedikit perbedaan pada bagian encoding. Setelah melakukan remove NULL dan imputing, kami melakukan feature engineering terhadap dataset yang sudah di load. Feature engineering yang kami lakukan berupa perhitungan closing value setiap 7 hari dan disimpan di dalam variabel bernama lag. Hal ini dilakukan untuk merekam pergerakan historis harga penutupan, yang berguna sebagai input prediktif. Selain lag\_1 hingga lag\_7, kami juga melakukan engineering column sehingga column akhir dari dataset kedua ini adalah *'Open', 'High', 'Low', 'Volume', ‘lag\_1’* hingga *‘lag\_7’,* dan *‘Ticker’.* Data-data ini disimpan dalam variabel *‘X\_all\_df’* dan *‘Y\_all\_df’*. Preprocessing untuk dataset kedua ini dilakukan pada bagian blok code *“Polynomial Regressor model (Future Prediction)”* dan bukan pada block “*2. Data Preprocessing”*.

Alasan kami membuat dua dataset spesifik untuk program ini untuk memastikan data yang hendak kami prediksi memiliki bias sekecil mungkin. Model dengan dataset 1 dapat memberikan prediksi yang cukup baik untuk closing value dalam range 1 tahun pertama tetapi hal tersebut tidak dapat memastikan hal yang lebih akurat untuk data lebih dari 1 tahun karena dataset 1 hanya memberikan informasi berupa closing value. Dataset 2 digunakan untuk membantu prediksi lebih dari 1 tahun karena features dalam dataset memungkinkan prediksi yang dilakukan lebih akurat.

**3.3 Data Training**

Pada tahapan ini kami melatih beberapa model machine learning menggunakan data yang telah diproses sebelumnya. Untuk dataset pertama, kami menggunakan tiga model utama yaitu: Linear Regression, Polynomial Regression, dan Random Forest Regressor. Masing-masing model dibungkus dalam pipeline dan dilatih dengan data pelatihan (X\_train,y\_train). Setelah itu, dilakukan pencarian parameter terbaik dengan GridSearchCVuntuk mencari model yang baik. Parameter yang ditemukan oleh algoritma grid kemudian dimasukan kedalam pipeline baru sebelum digabungkan kedalam sebuah model ensemble menggunakan *Voting Regressor* dan disimpan dalam pipeline ensemble. Model pertama dapat ditemukan pada code block “*Ensemble Regressor (Close Prediction)”.*

Untuk dataset kedua, kami hanya menggunakan satu model yaitu Polynomial Regression. Alasan kami memilih algoritma ini adalah polynomial regression mendukung terjadinya lengkungan dalam grafik, sesuatu yang linear regression tidak dapat lakukan. Polynomial regression juga lebih baik dalam memprediksi terjadinya naik dan turun pada suatu grafik karena tidak 100% terpaku pada tren grafik yang dihasilkan oleh training dataset. Model kedua dapat ditemukan pada bagian blok code *“Polynomial Regressor model (Future Prediction)”*.

Setelah proses training tersebut, kami melakukan export model menggunakan pickle. Terdapat 3 model utama yang kami export, yaitu *“close\_predict.pkl”, “future\_predict.pkl”,* dan *“mapping\_ticker.pkl”.* Model close dan future fungsinya sudah kami jelaskan pada bagian preprocessing sedangkan *mapping\_ticker.pkl* digunakan untuk melihat mapping ticker pada dataset pertama *(mengacu pada encoding ticker di proses preprocessing dataset 1).*

**3.4 Model Testing**

Untuk melakukan testing, diperlukan 2 fungsi utama agar program dapat berjalan, yaitu *forecast\_to\_date* dan *plot\_forecasted\_close.* 2 Fungsi ini digunakan secara berurut, berikut adalah step-by-step cara kerja fungsinya yang sudah kami rincikan agar lebih mudah dimengerti :

1. Konversi tanggal target:  
   Ubah target\_date\_str (dalam format string) menjadi objek datetime.
2. Ambil data terakhir:  
   Salin data historis untuk *ticker\_symbol* dari data. Pastikan indeks berupa datetime dan simpan tanggal terakhir (current\_date).
3. Ambil nilai encoding ticker:Gunakan *mapping\_ticker* untuk mendapatkan label numerik dari ticker dari fungsi yang sudah kita export tadi.
4. Forecast Loop :  
   Ulangi hingga current\_date mencapai target\_date:

a. Membuat ulang fitur lag yang sudah dibuat pada proses pembuatan dataset 2. Fungsi fitur lag adalah untuk menghitung average prediksi close setiap 7 hari. Hal ini memungkinkan loop berjalan terus hingga waktu yang diinginkan. Contoh, jika seseorang ingin melakukan prediksi saham 14 hari dari sekarang, maka model akan melakukan prediksi dari hari 1 - 7, 2 - 8, 3 - 9, dan seterusnya hingga hari akhir berada pada hari ke 14.

b. Prediksi Open, High, Low, dan Volume dengan model *future\_predict\_model.pkl* menggunakan fitur lag dan label ticker.

c. Hitung jumlah hari sejak awal data: Digunakan sebagai fitur tambahan untuk memprediksi harga Close.

d. Prediksi Close dengan menggunakan model *close\_predict\_model.pkl* untuk memprediksi harga penutupan berdasarkan fitur-fitur:

* hari ke-n
* Open
* High,
* Low
* Volume
* Ticker.

e. Buat baris data baru:Gabungkan hasil prediksi ke dalam satu baris dengan index tanggal berikutnya.

f. Tambahkan ke data historis dengan menggabungkan *new\_row\_df* dengan *last\_known*.  
  
g. Perbarui tanggal saat ini (current\_date) ke hari berikutnya.

1. Return hasil:  
   Kembalikan last\_known, yaitu data historis yang sudah ditambah dengan prediksi sampai tanggal target.
2. Dengan hasil dari fungsi *forecast\_to\_date*, fungsi *plot\_forecasted\_close* memanggil fungsi *forecast\_to\_date* agar dapat diubah menjadi line graph dengan periode 1 bulan dari tanggal yang diminta.

**3.5 Metrics & Evaluation**

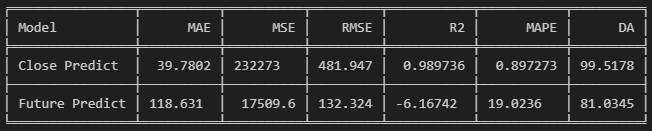
Evaluasi performa model dilakukan dengan berbagai metrik seperti, Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), R-Squared (R²), Mean Absolute Percentage Error (MAPE), dan Directional Accuracy (DA). Metrik-metrik ini memberikan gambaran yang komprehensif terkait ketepatan prediksi model dan kemampuan model dalam mengikuti arah perubahan nilai. Semua hasil evaluasi disusun dalam tabel agar mudah dibandingkan antar model. Selain evaluasi numerik, dilakukan juga visualisasi hasil prediksi seperti scatter plot antara nilai aktual dan prediksi, residual plot untuk melihat kesalahan model, serta plot deret waktu yang menunjukkan perbandingan antara nilai aktual dan nilai prediksi. Visualisasi ini membantu dalam mengidentifikasi pola atau anomali yang mungkin tidak terlihat dari angka saja.

**3.6 Model deployment**

Deploy model dilakukan dengan bantuan flask app, HTML CSS JS, dan framework AoS serta aset 3D spline untuk mempercantik website yang hendak kami buat.

**BAB 4. EVALUASI DAN HASIL**

Berdasarkan metrik yang kami gunakan, berikut adalah tabel hasil training model kami :

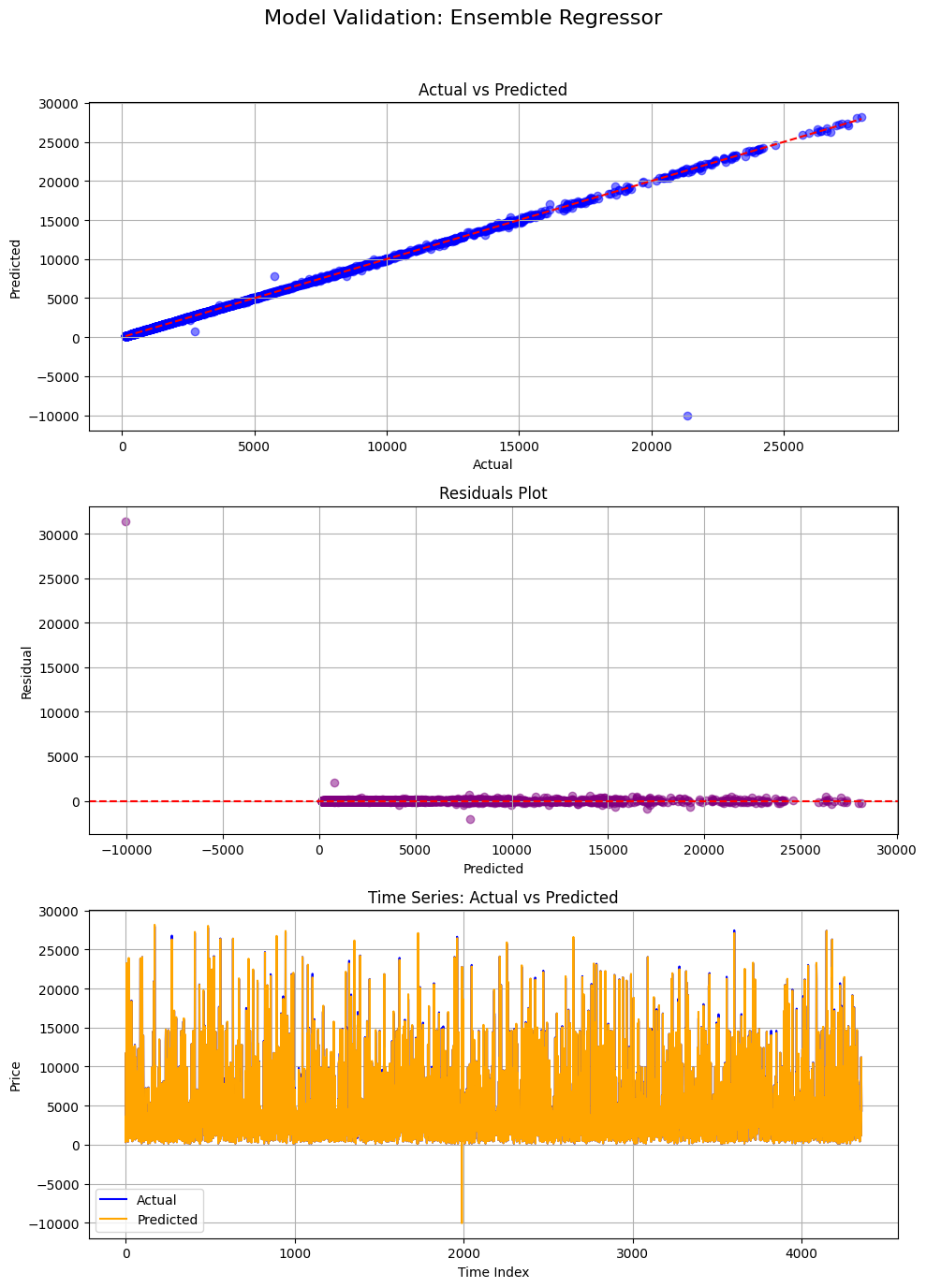


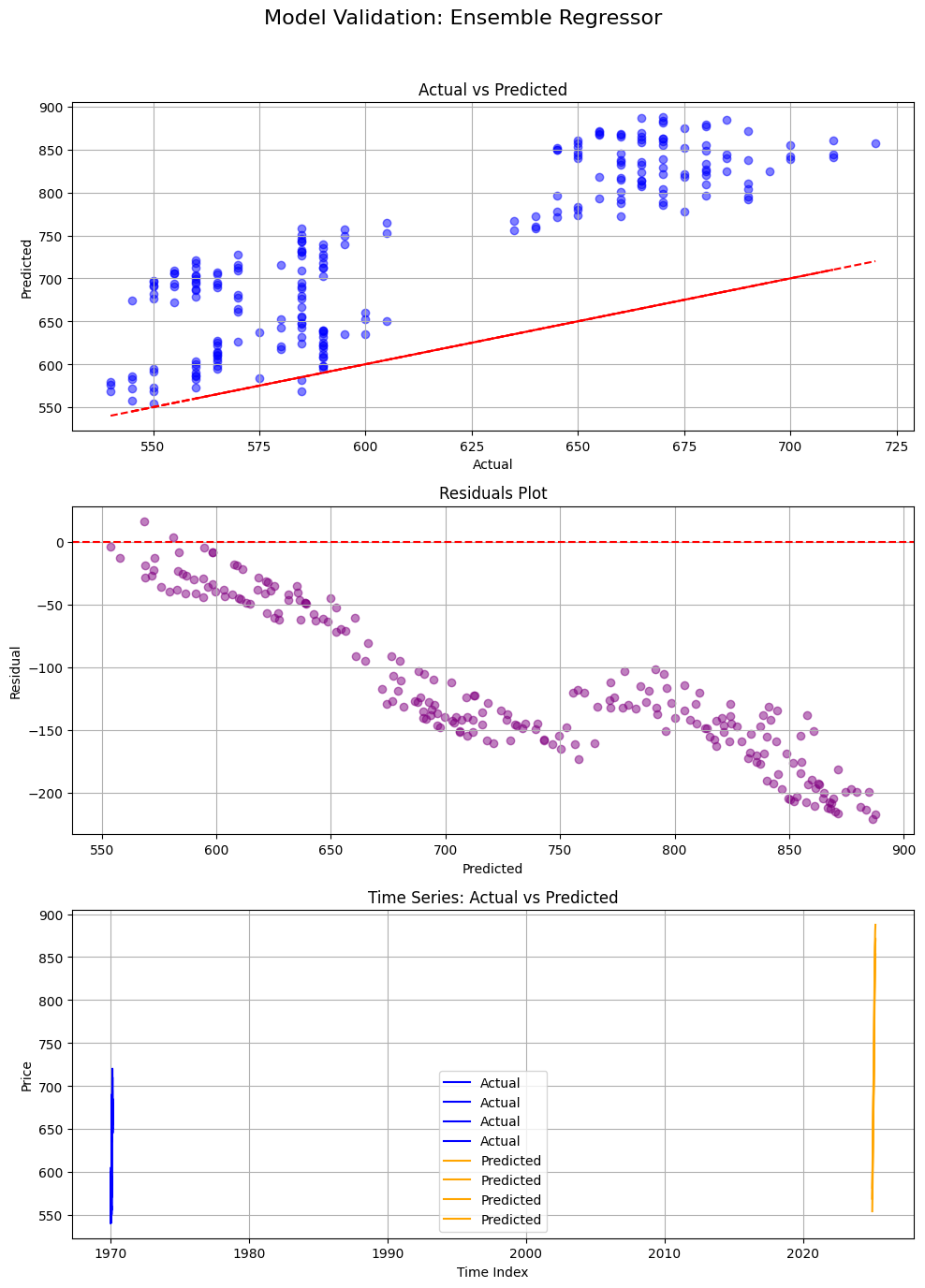
Model Close Predict berhasil mencapai skor R² yang memuaskan sebesar 98%, menunjukkan bahwa model ini mampu menjelaskan sebagian besar variasi harga penutupan saham. Sementara itu, model Future Predict justru menunjukkan skor negatif. Nilai ini mencerminkan adanya bias dalam model kami — yang memang disengaja.

Kami menyadari bahwa prediksi saham menggunakan machine learning tidak akan pernah sepenuhnya akurat, mengingat pergerakan harga saham dipengaruhi oleh banyak faktor eksternal seperti kondisi ekonomi, berita global, keputusan perusahaan, serta sentimen pasar — yang semuanya tidak dapat sepenuhnya dimasukkan ke dalam dataset kami. Selain itu, pergerakan harga saham bersifat real-time dan dinamis, sehingga pendekatan prediktif berbasis machine learning menghadapi keterbatasan.

Bias negatif pada model Future Predict sengaja diterapkan sebagai bentuk proteksi bagi pengguna. Tujuannya adalah agar ketika terjadi kesalahan prediksi, dampaknya tidak merugikan pengguna secara signifikan. Dengan kata lain, lebih baik pengguna tidak membeli saham karena prediksi yang terlalu konservatif, daripada membeli dan mengalami kerugian besar. Strategi ini kami anggap sebagai pendekatan yang lebih bijak dan bertanggung jawab dalam konteks investasi berbasis AI.

Selain skor metrik diatas, kami juga melakukan model validation yang menghasilkan grafik sebagai berikut : *(close predict dan future predict secara berurutan)*





**LAMPIRAN**

Berikut adalah lampiran link google drive untuk akses kode dan aset yang kami gunakan dalam membangun STOCKPEDIA :

[Machine Learning Project](https://drive.google.com/drive/folders/1xEUBy0nk8W-Z6u4j_nrCbnpbuPU5AiYm)

Lampiran link google drive ini mengandung :

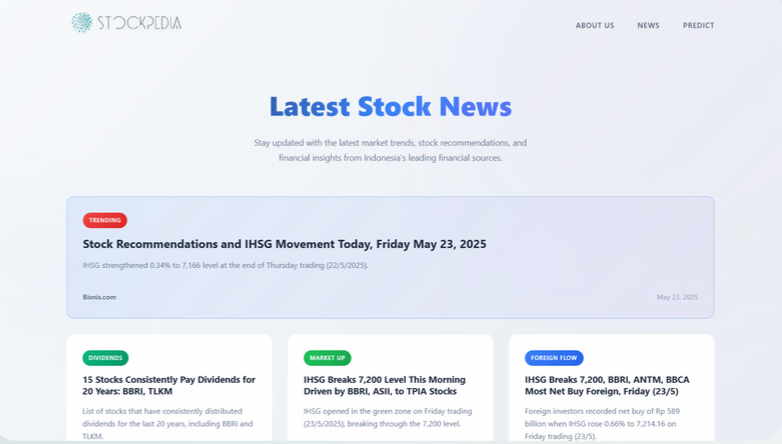
* Source code model
* Source code deployment
* Document proposal
* Video demonstrasi web app
* Contribution Statement
* Powerpoint presentasi STOCKPEDIA

Berikut adalah lampiran berupa tampilan gambar dari STOCKPEDIA :

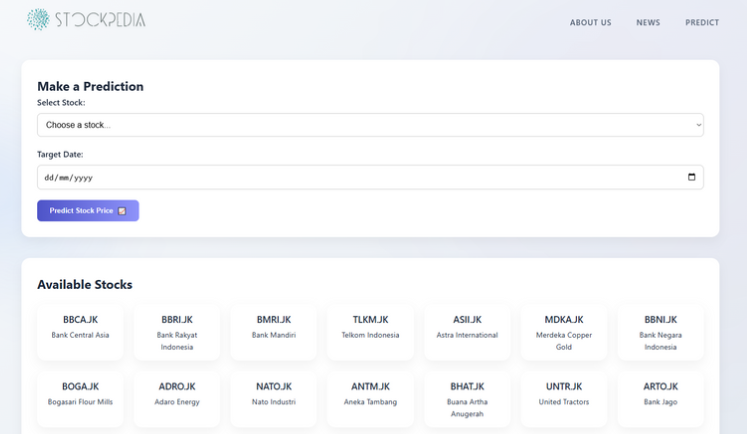
1. Home page



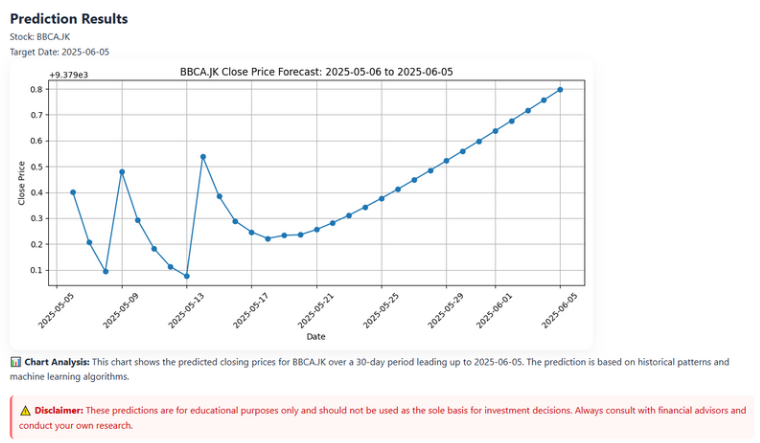
1. News page



1. Predict page - before pressing predict



1. Predict page - after pressing predict



**DAFTAR PUSTAKA :**

* **Bodie, Z., Kane, A., & Marcus, A. J. (2018). *Investments* (11th ed.). McGraw-Hill Education.**
* **Fama, E. F. (1970). “Efficient Capital Markets: A Review of Theory and Empirical Work.” *The Journal of Finance, 25*(2), 383-417.**
* **Bodie, Z., Kane, A., & Marcus, A. J. (2018). *Investments* (11th ed.). McGraw-Hill Education.**
* **Antara News. (2024, Agustus 8). *Pengertian saham dan pasar saham*.** [**https://www.antaranews.com/berita/4246663/pengertian-saham-dan-pasar-saham**](https://www.antaranews.com/berita/4246663/pengertian-saham-dan-pasar-saham)
* **DQLab. (2023, January 5). *Supervised Learning: Tipe Machine Learning Populer.*** [**https://dqlab.id/supervised-learning-tipe-machine-learning-populer**](https://dqlab.id/supervised-learning-tipe-machine-learning-populer)
* **Yankson, Matilda. 2025. *“Types of Regression Models and analysis.”* Analytics Vidhya.** [**https://www.analyticsvidhya.com/blog/2022/01/different-types-of-regression-models/**](https://www.analyticsvidhya.com/blog/2022/01/different-types-of-regression-models/)**.**