Eksplorasi Data Analisis dan Perbandingan Model Prediktif

Prediksi Harga Penutupan Saham Produk *Fast Retailing* 'Uniqlie'

Thomas Januardy
Program Studi Sistem Informasi, Fakultas Teknik dan Informatika, Universitas Multimedia Nusantara, Tangerang, Indonesia
thomas.januardy@student.umn.ac.id

***Abstract*—Retail is an activity that markets the end consumer to products or services for their own personal or household use. Retail business is a business that sells products or services to individual consumers or end consumers for their own use or not for resale (end users). With big data, as a retail entrepreneur, we will know the behavior and interest of consumers in an area that is of interest to the consumer market out there. The datas certainly has a great influence on an organization's development, such as in the industrial business sector, one of which is the retail business. However, the current movement of a company is not only valuing its products and merchandise, but also the capital market in the form of shares of ownership rights of a company. There are many factors that can determine a company has a value that can affect the value (shares) of the company. One of them is external factors based on the behavior patterns of capital market users who carry out stock transaction activities or can be referred to as technical analysis. The data will be analyzed using predictive models, such as Linear Regression, Random Forest, and Gradient Boosting, which have a basis for predicting a value. The purpose of this study is to predict stock prices and look for patterns created by user behavior patterns in the stock capital market based on previous years. The share price itself is the closing price which will be closed after the capital market is also closed and continued on the next day. This closing price becomes a benchmark for predictive analysis for the future by taking advantage of previous times that already have facts (data).**

Index Terms*—Retail business, Big Data, Data Mining, Prediction, Data Modelling, Exploratory Data Analysis, Linear Regression, Random Forest, Gradient Boosting, ASE

I. PENDAHULUAN

A. *Konsep Big Data dan Hubungannya dengan Bisnis Ritel*

Perkembangan teknologi telah beradaptasi dan terus mengikuti perkembangan zaman, yang semakin pesat tersebar luas dan mempengaruhi seluruh aspek kehidupan manusia sehari-hari yang tidak luput dari istilah bernama 'data' [1]. Data telah menjadi salah satu aspek penting, terutama di dunia digital, di mana semua informasi dapat memiliki catatan riwayat esensial yang bersumber dari sebuah data yang semakin melimpah, membentuk satu terminologi *'big data'* [1]. Setiap aktivitas manusia yang memanfaatkan teknologi, seperti internet, hampir semuanya menggunakan data sebagai infrastruktur berbasis agar dapat berjalan dalam sebuah sistem. Data adalah sekumpulan fakta "mentah" dari suatu proses pengamatan atau penelitian, sebelum menjadi data hasil yang sudah dikelola, yaitu informasi. Data dapat dibedakan menjadi data kuantitatif dan data kualitatif. Setelah data diubah menjadi informasi, fakta dapat digunakan untuk banyak hal, salah satunya sebagai bahan analisis strategis untuk menentukan suatu keputusan, baik data kualitatif maupun kuantitatif. Maka dari itu, keberadaan data sangat diperlukan oleh industri, terutama bisnis dan organisasi yang sudah melangkah masuk ke dalam dunia digital. Dengan data, sebuah organisasi dapat dengan mudah mendapatkan informasi untuk perancangan suatu strategi yang nantinya dapat memiliki manfaat serta keuntungan, baik untuk perkembangan organisasinya (dalam hal material/informasi). Data yang telah diolah menjadi informasi dapat membantu suatu keputusan yang terbaik berdasarkan fakta di lapangan yang ada, sehingga mempermudah analisis dan langkah strategi selanjutnya yang dapat dilakukan dalam sebuah perencanaan.

Big data adalah istilah sebuah konsep yang menggambarkan terdapatnya volume data dalam jumlah yang besar, baik terstruktur maupun tidak terstruktur yang memiliki tujuan dan manfaat yang beragam dalam penggunaannya. Data dengan tingkat yang besar dapat dianalisis untuk wawasan yang mengarah pada keputusan dan langkah bisnis strategis yang lebih baik [3]. Selain itu, *big data,* yang memegang peranan data sebagai sebuah fakta, juga menjadi aspek yang sering dibutuhkan oleh perusahaan besar saat ini sebagai akumulasi data yang sangat kompleks dan berguna bagi perusahaan.

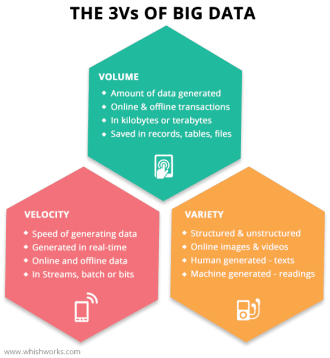
Dengan konektivitas jaringan teknologi dan internet yang semakin luas dan cepat, pengguna akan dengan sangat mudah untuk mendapatkan banyak koleksi informasi yang berkaitan dengan hampir setiap aspek kehidupan kita. Tetapi dibalik mudahnya

pencarian suatu data, proses pengelolaan analisis dengan jumlah yang sangat besar, memiliki perjalanan yang tidak dapat dilakukan dengan cara instan, melainkan melewati suatu prosedur tahapan yang profesional agar data tersebut dapat dikelolakan. Istilah *big data* sendiri mengacu pada data yang sangat besar, cepat atau kompleks sehingga sulit atau tidak mungkin untuk diproses menggunakan metode tradisional. Tindakan mengakses dan menyimpan sejumlah besar informasi untuk analitik telah ada sejak lama. Tetapi konsep big data mendapatkan momentum di awal tahun 2000-an ketika seorang analis industri Doug Laney mengartikulasikan definisi *big data* yang sekarang ke dalam berbagai karakteristik utama sebagai 'Tiga V':

- *Volume*: Organisasi mengumpulkan data dari berbagai sumber, termasuk transaksi bisnis, perangkat pintar (IoT), peralatan industri, video, media sosial, dan banyak lagi. Di masa lalu yang masih menggunakan metode tradisional, menyimpannya tentu akan menjadi masalah dan memakan banyak waktu, tetapi penyimpanan di era sekarang yang lebih mudah sudah terdapat pada *platform* seperti SAS dan Hadoop, yang telah meringankan beban.

- *Velocity*: Dengan pertumbuhan *'Internet of Things'*, aliran data ke bisnis dengan kecepatan yang belum pernah terjadi sebelumnya/sangat cepat dan harus ditangani secara tepat waktu. RFID, sensor, dan pengukur pintar mendorong kebutuhan untuk menangani aliran data ini secara hampir *real-time*.

- *Variety*: Data tersedia dalam semua jenis format – mulai dari data terstruktur dan numerik dalam basis data tradisional hingga dokumen teks tidak terstruktur, *email*, video, audio, dan transaksi keuangan [3].



Gambar 1. Konsep 'Tiga V'

Big data mengumpulkan baik sumber data yang terstruktur maupun tidak terstruktur, disamping mereka adalah data kuantitatif atau kualitatif. Data tidak terstruktur dapat berasal dari media sosial (Facebook, Instagram, postingan Twitter, dsb) [6]. Sedangkan sumber data terstruktur dapat berasal dari *database* internal, seperti dari suatu organisasi atau perusahaan [4]. Banyak yang percaya bahwa, bagi perusahaan yang melakukan pengelolaan data dengan benar, *big data* akan mampu mengeluarkan kemampuan dan nilai organisasi baru [1]. Penciptaan dan pengumpulan data di masyarakat saat ini tidak hanya terbatas pada fungsi yang berkaitan dengan keuangan dan konsumen. Singkatnya, perusahaan-perusahaan ini dapat memiliki gambaran yang jauh lebih lengkap tentang pelanggan dan operasi mereka dengan menggabungkan data tidak terstruktur dan terstruktur [2].

Data tersebut tentunya memiliki pengaruh yang besar terhadap suatu perkembangan organisasi, seperti di bidang bisnis industri, salah satunya bisnis ritel. Bisnis ritel memiliki suatu sistem transaksi yang digerakkan oleh dua pihak, baik dari segi penjual dan pembeli. Pihak penjualan memiliki basis data yang bersifat internal sehingga dapat diakses langsung oleh perusahaan. Sedangkan, dari pihak pembeli, kebutuhan untuk mengetahui kegiatan dan preferensi untuk melakukan transaksi tersebut tentunya dibutuhkan juga oleh perusahaan untuk strategi dan perencanaan baru, namun tidak semudah mengambil data internal. Disinilah kebutuhan akan pengelolaan data diperlukan, dengan fasilitas yang menunjang dan mendukung untuk keperluan *big data*. Mulai dari perilaku konsumen, seperti preferensi pemilihan produk, harga produk, hingga warna produk dapat diketahui apabila pengelolaan yang dilakukan sesuai dan tercapai tujuannya. Dengan *big data*, selaku pengusaha ritel akan mengetahui perilaku dan minat konsumen akan suatu bidang yang diminati pasar konsumen di luar sana. Apabila pengelolaan *big data* dapat dilakukan dengan baik, perusahaan ritel tentunya akan mendapatkan nilai lebih, seperti peningkatan penjualan, timbal balik pelayanan yang baik oleh konsumen, dan lain sebagainya.

Perusahaan-perusahaan yang bergerak pada sektor bisnis umumnya memiliki orientasi utama terhadap keuntungan dan pendapatan margin laba setinggi-tingginya. Berbagai informasi penting dapat dihasilkan dari *big data* yang dapat mendukung proses pengambilan keputusan bagi pimpinan perusahaan sebagai berikut: [5]

a) Mengetahui respons masyarakat terhadap produk-produk yang dikeluarkan melalui analisis sentimen di media sosial.

b) Membantu perusahaan mengambil keputusan secara lebih tepat dan akurat berdasarkan data.

c) Membantu meningkatkan citra perusahaan di mata pelanggan.

d) Perencanaan usaha, dengan mengetahui perilaku pelanggan seperti pada perusahaan telekomunikasi dan perbankan.

e) Mengetahui tren pasar dan keinginan konsumen.

Perusahaan-perusahaan yang bergerak pada sektor bisnis dapat memanfaatkan informasi-informasi berharga yang dihasilkan *big data* untuk mengoptimalkan proses pengambilan keputusan, agar target memaksimalkan raihan profit dapat tercapai [5]. Oleh karena itu, di era yang serba digital ini, penggunaan *big data* memiliki pengaruh serta manfaat yang besar bagi keberlangsungan perusahaan. *Big data* dipandang sebagai sebuah aset berupa informasi yang berharga, di mana berpeluang dalam menambah wawasan dan pengetahuan yang baik, seperti dari analisis pasar dan pola perilaku konsumen. Dengan *big data*, sebuah perusahaan pada akhirnya dapat membantu perusahaan menentukan perencanaan ke depan terhadap langkah apa yang selanjutnya akan dihadapi dalam menyusun strategi.

B. Pemahaman Bisnis dan Permasalahannya

Perusahaan *retail* memiliki peranan penting dalam membantu perekonomian suatu negara karena tingkat permintaan penyediaan produk dalam bentuk barang dan jasa oleh konsumen sangat tinggi, sehingga terjadi perputaran keuangan antara penyedia produk dengan konsumen [8]. Perusahaan ritel membuat dan menyalurkan produk serta jasa dalam skala yang besar, dan membagi-bagi pasokannya terhadap pelaku usaha beserta pasarnya dalam skala yang relatif kecil [7]. Tetapi seiring perkembangannya zaman dan konsep teknologi, turut membuka peluang bagi usaha ritel untuk melakukan inovasi baru dalam memasarkan produknya. Terutama dengan hadirnya teknologi internet yang dapat mencakup kawasan secara global dan tidak terbatas, menjadikan internet sebagai media yang dapat mewadahi jajaran dan pusat pemasaran yang berbasis digital. Seperti yang diketahui juga bahwa pengguna internet dari tahun ke tahun mengalami peningkatan jumlah pengguna seiring *platform* yang disediakan dalam jaringan internet. Perusahaan *retail* harus menyediakan sistem pelayanan produk yang berbasis *online* (*E-commerce*) karena daya beli atau tingkat konsumsi masyarakat ke perusahaan retail menurun disebabkan penyediaan produk di bisnis berbasis *online* tersebut tidak kalah dari perusahaan *retail* [8]. Tapi di era globalisasi saat ini, setiap perusahaan harus menyiapkan strategi baru, salah satunya perusahaan *retail*, dilihat dari perkembangan teknologi saat ini sangat cepat. Perusahaan ritel harus mengikuti arus perkembangan teknologi apabila perusahaan ritel tetap ingin menjaga keeksistensian perusahaan dan menjaga konsumennya dengan cara perusahaan *retail* harus menyediakan bisnis yang berbasis online juga, agar daya beli atau tingkat konsumsi konsumen tidak menurun ke perusahaan tersebut. Apabila hal itu tidak dilakukan, kemungkinan yang terjadi adalah adanya ancaman bagi perusahaanr ritel, karena seperti yang terjadi saat ini sudah banyak bermunculan bisnis yang berbasis *online* yang menyediakan produk dalam bentuk barang dan jasa, dan hal tersebut akan memberikan banyak dampak ke perusahaan *retail*, salah satunya adalah akan berdampak ke harga saham perusahaan *retail* [9].

Pasar modal (saham) merupakan salah satu instrumen ekonomi yang memiliki peran penting bagi perekonomian suatu negara karena pasar modal menjalankan dua fungsi sekaligus yaitu fungsi ekonomi dan fungsi keuangan. Pasar modal memberikan berbagai alternatif untuk para investor selain berbagai investasi lainnya, seperti menabung di bank, membeli tanah, asuransi, emas dan sebagainya. Pasar modal merupakan penghubung antara investor (pihak yang memiliki dana) dengan perusahaan (pihak yang memerlukan dana jangka panjang) ataupun institusi pemerintah melalui perdagangan instrumen melalui jangka panjang, seperti surat berharga yang meliputi surat pengakuan utang, surat berharga komersial (*commercial paper*), saham, obligasi, tanda bukti hutang, waran (*warrant*), dan *right issue.* Jenis pasar modal menurut Sunariyah (2000) jenis-jenis pasar modal dibagi menjadi 4 (empat), antara lain: Pasar Perdana (*Primary Market*), Pasar Sekunder (*Secondary Market*), Pasar Ketiga (*Third Market*) dan Pasar Keempat (*Fourth Market*). Sedangkan instrumen pasar modal terdiri dari Saham (*Stock*), Obligasi (*Bond*), *Right*, *Warrant* dan Opsi [9].

Setiap pihak, termasuk seorang individu dapat memanfaatkan fasilitas pasar modal untuk melakukan investasi, salah satunya adalah dengan investasi di perusahaan yang terdapat di pasar modal, salah satunya perusahaan ritel. Terdapat banyak faktor yang dapat menentukan suatu perusahaan memiliki nilai yang dapat mempengaruhi nilai (saham) perusahaan tersebut. Mulai dari perencanaan dan hasil projek yang menguntungkan, inovasi baru, hingga hasil penjualan tiap tahunnya yang berbentuk dividen. Macam-macam instrumen yang terdapat pada saham, antara lain harga pembukaan *(open price)*, harga penutupan (*closing price)*, harga tertinggi *(high)*, harga terendah *(low)*, volume, *stock trading*, dan lain sebagainya yang dipatok berdasarkan interval waktu tertentu. Terdapat banyak faktor yang dapat menentukan pergerakkan harga saham dari waktu ke waktu. Faktor tersebut biasa terjadi karena adanya faktor internal, seperti *launching* produk baru, lakunya suatu produk, dan lain sebagainya, tetapi tidak menutup kemungkinan dari faktor eksternal, yaitu pola perilaku di pasar modal saham terhadap suatu perusahaan tersebut. Sebagai pengguna pasar modal yang memiliki modal dan saham, tentunya pengguna akan memiliki aset dan bebas melakukan kegiatan jual-beli menurut preferensinya. Maka dari itu, terciptanya faktor eksternal berdasarkan pola perilaku pengguna pasar modal yang melakukan kegiatan transaksi saham atau dapat disebut sebagai analisis secara teknikal.

Tujuan dari penelitian ini adalah untuk memprediksi harga saham dan mencari pola yang tercipta akibat pola perilaku pengguna di pasar modal saham berdasarkan tahun-tahun sebelumnya. Harga saham sendiri merupakan harga penutupan *(closing price)* yang akan ditutup setelah pasar modal juga ditutup dan dilanjutkan pada hari berikutnya. Harga penutupan ini menjadi patokan untuk analisis prediksi untuk waktu ke depan dengan memanfaatkan waktu sebelum-sebelumnya yang sudah memiliki fakta. Metode yang digunakan menggunakan model *linear regression* untuk memprediksi nilai, *random forest*, dan *gradient boosting* untuk membandingkan keakuratan nilai yang dihasilkan dari data yang sudah ada, dengan variable *'close'* sebagai target variabel yang akan diprediksi. Hasil analisis diharapkan dapat menjadi patokan bahkan keakuratan harga dalam mempersiapkan strategi maupun perencanaan untuk melakukan kegiatan jual beli saham di pasar modal perusahaan ritel 'Uniqlie' dari tahun 2012-2017 dan seterusnya.

II. TINJAUAN TEORITIS

A. Data Mining

Proses pencarian dan penggalian informasi dari tumpukan data *(big data)* dengan jumlah yang besar merupakan proses utama dari *data mining*, tujuan utama dari pengolahan data adalah untuk menghasilkan informasi baru [9]. *Data mining* adalah proses menemukan pengetahuan yang menarik dari sejumlah besar data yang disimpan dalam basis data, gudang data, atau tempat penyimpanan informasi lainnya. Sejumlah teknik data mining telah dilakukan pada data mining pendidikan untuk meningkatkan kinerja siswa seperti regresi, algoritma genetika, klasifikasi teluk, pengelompokan *k-means*, aturan asosiasi, prediksi, dan lain sebagainya. Teknik *data mining* dapat digunakan di segala bidang untuk meningkatkan pemahaman kita tentang proses pembelajaran untuk fokus pada mengidentifikasi, mengekstraksi dan mengevaluasi variabel yang terkait dengan proses belajar. *Data mining* juga dikenal dengan Knowledge Discovery in Database (KDD), yaitu suatu proses yang secara otomatis mencari data dalam ruang memori yang sangat besar dari data untuk menemukan pola dengan menggunakan teknik seperti klasifikasi asosiasi atau klastering [10].

B. Linear Regression

Regresi adalah teknik membangun model yang bertujuan mencari nilai yang digunakan untuk memprediksi data masukan diberikan. Regresi adalah ukuran statistik yang digunakan untuk menentukan kekuatan hubungan antar variabel dependen (tidak mandiri) dan independen (mandiri). Cara kerja utama untuk membuat prediksi adalah dengan membangun model regresi bekerja dengan mencari hubungan antara satu atau lebih variabel bebas atau prediktor dengan variabel (X) dependen atau respon (Y).Model regresi linier memodelkan hubungan antara variabel skalar dan satu atau lebih variabel jelas.

Secara umum, model *linear regression* dibagi menjadi dua jenis, yaitu *simple linear regression* dan *multiple linear regression*. *Simple linear regression* merupakan hubungan antara satu variabel dependen dengan satu variabel independen, sedangkan *multiple linear regression* merupakan hubungan antara satu variabel dependen dengan dua atau lebih variabel independen [11].

Metode *linear regression* dapat digunakan untuk mencari nilai *closing price* dari seluruh variabel prediktor yang ada, seperti *opening price,* harga tertinggi, harga terendah, volume, dan *stock trading* yang tersedia dalam suatu data. Maka, variabel dependen (Y) dapat diartikan sebagai *closing price*, dan variabel independen (X) terdiri dari *opening price*, harga tertinggi, harga terendah, volume, dan *stock trading*.

A. Random Forest

Random forest merupakan sebuah metode *ensemble* dan *supervised*. Metode *ensemble* merupakan cara untuk meningkatkan akurasi metode klasifikasi dengan cara mengkombinasikan metode klasifikasi. Metode *supervised* dapat digunakan untuk memprediksi keluaran dari data masukan [12]. Pada umumnya, *random forest* terdiri dari beberapa *decision tree* dari proses *data mining*. *Random forest* merupakan proses klasifikasi dari seluruh masing-masing struktur 'pohon', di mana masing-masing pohon melemparkan unit suara untuk kelas paling populer di *input* X [13]. Maka dengan itu, sebuah *random forest* akan mengumpulkan dari sekian banyak model *decision tree* ke dalam suatu jenis klasifikasi untuk suatu kelas.

Metode dan cara kerja dari model *random forest* diawali dengan pembentukan 'pohon' pertama dan semakin banyak untuk memutuskan suatu keputusan. Di lain sisi, keputusan akhir akan ditentukan oleh hasil keputusan terbanyak dari 'pohon' yang telah dibangun. Konsep tersebut dinamakan *majority voting*, yang mengambil suara terbanyak dari anggota yang ada [14].

B. Gradient Boosting

Salah satu metode *machine learning* adalah *boosting*. *Boosting* adalah strategi *ensemble* yang membagi data *training* menjadi beberapa bagian dan masing-masing bagian diterapkan ke dalam model yang berbeda atau satu model dengan preferensi yang berbeda, kemudian hasilnya dikombinasikan berdasarkan 'suara' terbanyak [15]. *Gradient boosting* adalah teknik *supervised learning* berbasis *decision tree*. Algoritma dimulai dari menghasilkan pohon klasifikasi awal dan terus menyesuaikan pohon baru melalui minimalisasi fungsi kerugian [16].

Algoritma *gradient boosting* bekerja secara sekuensial, yaitu secara berurutan menambah setiap

prediktor sebelumnya yang kurang cocok dengan prediksi ke *ensemble*, dan memastikan kesalahan yang dibuat sebelumnya diperbaiki. Penggambaran sederhana konsep *ensemble* adalah keputusan-keputusan dari berbagai mesin pembelajaran digabungkan, kemudian untuk kelas yang menerima mayoritas 'suara' adalah kelas yang akan diprediksi oleh keseluruhan *ensemble* [17].

III. METODOLOGI

Metode penelitian adalah langkah yang dilakukan dalam rangka untuk mengumpulkan informasi atau data, serta melakukan evaluasi pada data yang telah didapatkan tersebut. Metode penelitian memberikan suatu gambaran rancangan penelitian yang meliputi prosedur penelitian dan langkah-langkah yang harus dikerjakan, waktu penelitian, sumber data, dan dengan langkah apa data-data tersebut diperoleh dan tahap selanjutnya diolah dan dianalisis untuk mendapatkan suatu kesimpulan.

A. *Objek Penelitian*

Objek penelitian yang digunakan sebagai bahan penelitian adalah untuk memprediksi harga penutupan *(closing price)* dengan variable 'Close' menggunakan pola dari tahun-tahun sebelumnya (2012-2016). *Data set* telah memenuhi kriteria minimum, ditentukan untuk tujuan penelitian dengan 7 variabel dan sekitar 1,233 data mengenai instrumen saham perusahaan Uniqlie dari tahun 2012-2017. Variabel terdiri dari 'Date' (*category*), 'Open' (*measure*), 'High' (*measure*), 'Low' (*measure*), 'Close' (*measure*), 'Volume' (*measure*), dan 'Stock Trading' (*measure*). Karena 'Date' merupakan variabel kategorikal, penggunaan *date* hanya diterapkan sebagai patokan dalam melihat suatu harga berdasarkan tanggal dan selebihnya tidak digunakan dalam model prediktif. Dengan variabel-variabel instrumen harga saham beserta datanya tersebut, cukup untuk diterapkan dalam model prediktif sehingga keakuratan dan kecocokan harga dapat menghasilkan hasil prediksi yang maksimal.

B. *Metode Pengumpulan Data*

Metode pengumpulan yang digunakan adalah dengan mengambil *data set* dari salah satu situs pencarian *data set* terkenal bernama "Kaggle". Situs *web* tersebut memiliki banyak kumpulan data *online*, terutama dari komunitas sains dan praktisi *machine learning*. Kaggle juga memungkinkan untuk menemukan dan mempublikasikan kumpulan data, menjelajahi dan membangun model di lingkungan ilmu data berbasis *web*, bekerja dengan ilmuwan data dan insinyur pembelajaran mesin lainnya, dan mengikuti kompetisi untuk memecahkan tantangan ilmu data. Alasan pengambilan *data set* di aplikasi Kaggle karena kelengkapan isi data dari *website* sangat lengkap dan bebas diakses oleh umum. Setiap orang dapat mengakses data dan dapat memprosesnya. Kaggle memiliki banyak fitur bermanfaat seperti buku catatan sebagai sarana bagi pengguna untuk menampilkan data yang dibutuhkan.

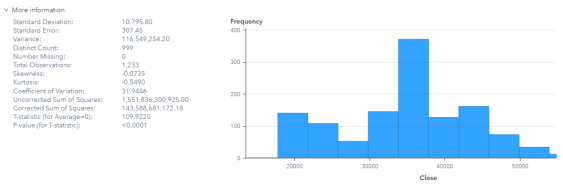
IV. HASIL DAN ANALISIS DATA

A. *Validasi Data*

Dari keseluruhan *data set,* dibuat sebuah partisi dengan skala 70:30 untuk melakukan *simple random sampling* terhadap data yang akan digunakan. Keseluruhan 7 variabel yang tersedia pada *data set* dilakukan pemeriksaan dan pengecekan untuk segala informasi harga yang berkaitan dengan suatu instrumen harga. Dari 7 variabel yang ada, tidak ditemukan data yang hilang, sehingga dapat disimpulkan bahwa semua variabel aman digunakan dan siap untuk berproses ke tahap selanjutnya.

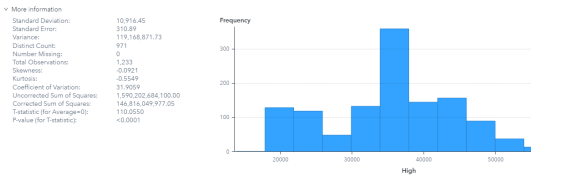
Berikut adalah gambar dan penjelasan informasi secara singkat mengenai masing-masing variabel *measure*:

a) Close, dengan nilai minimal 13,720, maksimal 61,930, rata-rata 33,795.43, total 41,669,765.



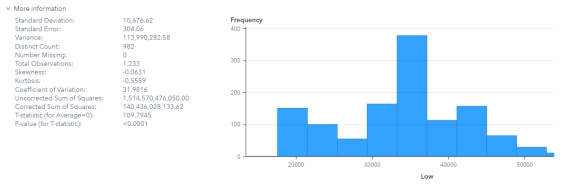
Gambar 2. Variable Close

b) High, dengan nilai minimal 13,840, maksimal 61,970, rata-rata 34,214.47, total 42,186,440.



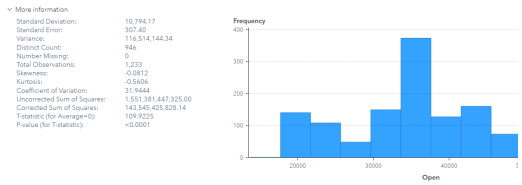
Gambar 3. Variable High

c) Low, dengan nilai minimal 13,600, maksimal 60,740, rata-rata 33,383.59, total 41,161,970.



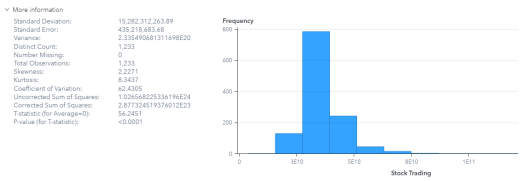
Gambar 4. Variable Low

d) Open, dengan nilai minimal 13,720, maksimal 61,550, rata-rata 33,790.49, total 41,663,675.



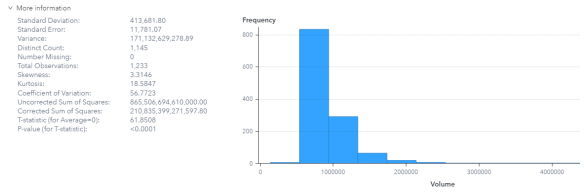
Gambar 5. Variable Open

e) Stock Trading, dengan nilai minimal 3,966,140, maksimal 146,045,000,000 , rata-rata 24,478,929,378.88, total 30,182,519,918,000.



Gambar 6. Variable Stock Trading

f) Volume, dengan nilai minimal 139,100, maksimal 4,937,300 , rata-rata 728,668, total 898,448,500.

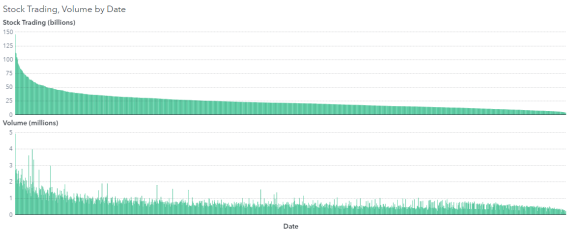


Gambar 7. Variable Volume

B. *Eksplorasi Data*

Eksplorasi data dilakukan dengan pembuatan figur dan grafik menggunakan fitur dengan variabel-variabel yang ada. Karena variabel Date merupakan variabel kategorikal, hampir keseluruhan grafik dalam eksplorasi data ditinjau dan dipatok berdasarkan Date. Eksplorasi data dibagi menjadi 3 jenis berdasarkan penempatan variabel harga yang saling berhubungan. Berikut adalah gambar dan penjelasan informasi secara singkat mengenai masing-masing jenis eksplorasi data:

- Stock Trading, Volume by Date

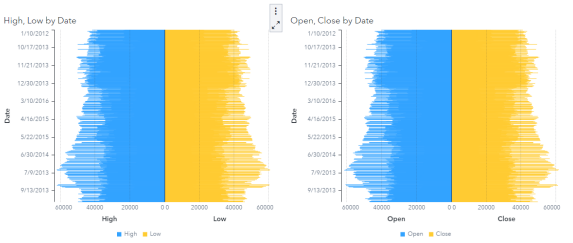


Gambar 8. Stock Trading, Volume by Date

Eksplorasi data dalam hubungan antar variabel Stock Trading dan Volume berdasarkan Date. Terlihat bahwa pada gambar, kedua pola Stock Trading dan Volume memiliki bentuk grafik yang hampir sama, tetapi dengan *detail*/rinci yang berbeda. Stock Trading dipatok berdasarkan harga per miliar, sedangkan Volume berdasarkan harga per juta. Walaupun berbeda, Volume yang terdapat di pasar modal memiliki pola yang berbentuk sama dengan keberadaan Stock yang diadakan. Mulai dari yang paling tertinggi (kiri), ke yang paling terendah (kanan), dengan

Stock Trading dan Volume tertinggi pada tanggal 3 Agustus 2018 sebesar 146,045,000,000 dan 4,937,300, serta terendah pada 13 Agustus 2012 sebesar 3,966,140,000 dan 229,500.

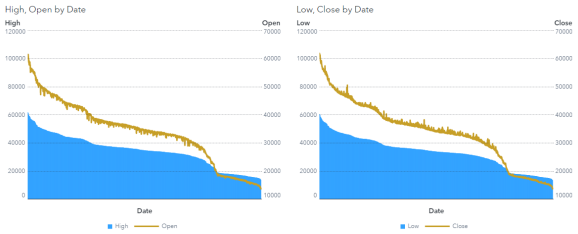
- High, Low and Open, Close by Date



Gambar 9. High, Low and Open, Close by Date

Eksplorasi data dalam grafik berdasarkan jenis instrumen yang sama (harga buka/tutup, tertinggi/terendah). Grafik dari setiap jenis memiliki pola pergerakkan rata-rata yang hampir sama, dengan perbedaan minor di antara jenis maupun tanggal. Dapat disimpulkan bahwa setiap harga tertinggi maupun pembukaan, memiliki sifat pola pengulangan terhadap harga terendah dan penutupnya. Tetapi tidak menutup kemungkinan terjadinya perbedaan, baik itu mayor maupun minor. Namun, jika ditarik secara rata-rata menyimpulkan hasil yang mirip.

- High, Open and Low, Close by Date

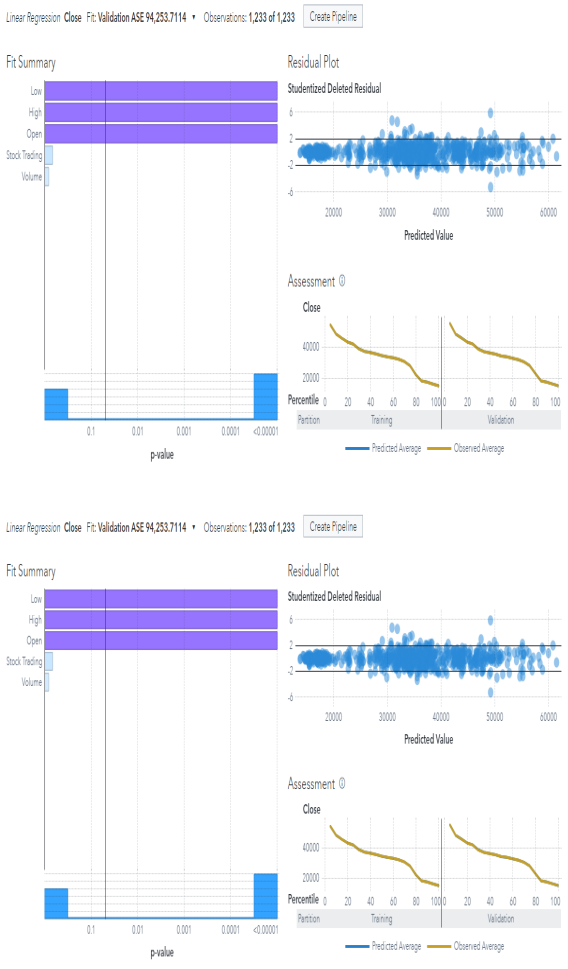


Gambar 10. High, Open and Low, Close by Date

Eksplorasi data dalam grafik berdasarkan jenis instrumen yang bersilangan dengan bentuk eksplorasi data sebelumnya (harga buka/tertinggi, harga tutup/terendah). Grafik menunjukkan nilai dengan pengulangan yang sama, seperti pada eksplorasi data sebelumnya. Menunjukkan pola yang relatif mirip, namun terdapat perbedaan minor pada setiap harga maupun tanggalnya.

C. *Modeling*

1) *Linear Regression*

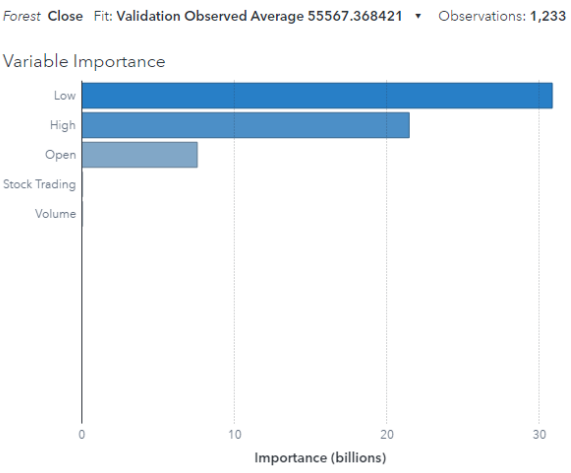


Gambar 11. Linear Regression

Metode *linear regression* di atas digunakan untuk mencari nilai *closing price* dari seluruh variabel prediktor yang ada, seperti *opening price,* harga tertinggi, harga terendah, volume, dan *stock trading* yang tersedia dalam suatu data. Variabel dependen (Y) dapat diartikan sebagai *closing price* (Close sebagai respon), dan variabel independen (X) terdiri dari *opening price*, harga tertinggi, harga terendah, volume, dan *stock trading* (Open, High, Low, Volume, Stock Trading sebagai prediktor dan pendukung). Terlihat pada *summary*, bahwa untuk pengecekan prediksi untuk variabel Close, yang hanya berkualifikasi hanyalah nilai berformat harga, yaitu Low, High, Open dengan alpha 0.05/-log (alpha) 1.30103. Sedangkan Stock Trading dan Volume bersifat sebagai satuan unit.

Pada *predicted value,* terlihat data dengan pola yang sama menimbulkan paling banyak kepadatan pada jarak harga antara 30,000 sampai 40,000 berdasarkan *studentized deleted residuals. Studentized Deleted Residuals* sendiri merupakan sebuah *error* yang timbul dan dapat digunakan untuk mendeteksi suatu *outlier* pada nilai residual. Dengan adanya prediksi sampai ke 60,000, terdapat kemungkinan bahwa harga akan terus meningkat secara pelan, dengan pertimbangan terdapat kepadatan juga pada harga 0 sampai 20,000.

2) *Random Forest*

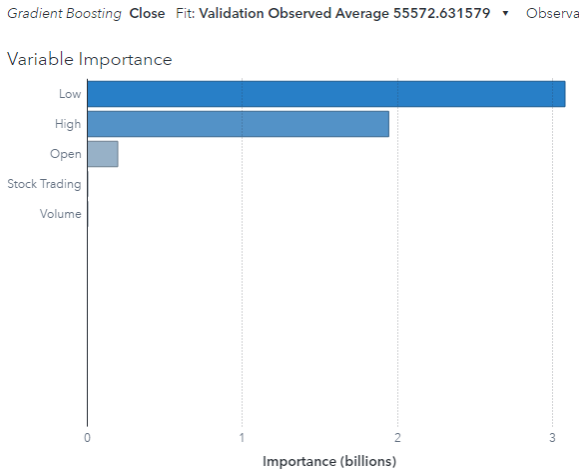


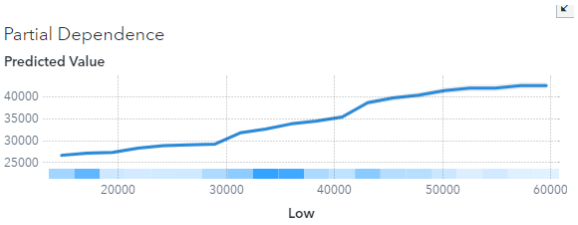
Gambar 12. Random Forest

Metode *random forest* memiliki faktor pendukung dalam *Variable Importance* yang sama dalam memprediksi respon Close, yakni Low, High, dan Open, terkait dengan instrumen harga.

Karena terdapatnya partisi, *data set* dapat menggunakan partisi sebagai validasi keakuratan data, yang menimbulkan kesamaan pola antar plot *out-of-bag*, *validation,* beserta *training* yang menyatu dengan figur *out-of-bag* berdasarkan *Average Square Error* (ASE). Di puncak tertinggi, *validation* sekitar 870,000 dengan jumlah *tree* 1, diikuti *out-of-bag* dan *training* di sekitar 609,000 dengan jumlah *tree* 1. Di terendah, terdapat 2 *validation* dengan jumlah *tree* 99 dan 100, sekitar 380,000, diikuti oleh masing-masing 2 *out-of-bag* dan *training* dengan jumlah *tree* 99 dan 100 di sekitar 270,000.

3) *Gradient Boosting*





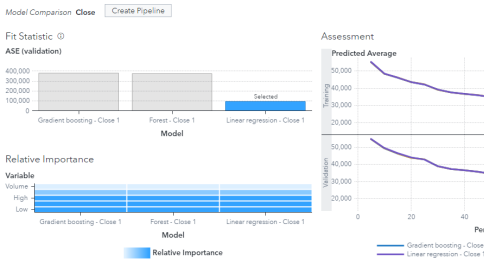
Gambar 13. Gradient Boosting

Gradient boosting memiliki sistematis yang relatif mirip dengan *random forest. Variable Imporatance* masih sama diikuti oleh kedua model lainnya yang merupakan instrumen harga, yaitu Low, High, dan Open. Perbedaannya terhadap *random forest*, yaitu *error plot* menjadi *iteration plot* di model ini.

Berdasarkan *iteration plot,* ASE dari kedua *training* dan *validation* memiliki tingkat dan pola yang sama, membentuk penyatuan dua figur menjadi satu. Di puncak tertinggi, dengan jumlah *tree* 1, di *training* sekitar 94,000,000 dan di *validation* sekitar 95,000,000. Di terendah, dengan jumlah *tree* 50, di *training* sekitar 207,000 dan *validation* sekitar 361,000.

Partial Dependence juga cenderung menunjukkan kepadatan di jarak harga 30,000 sampai 40,000, dan kepadatan ringan di jarak 0 sampai 20,000, serta sedikit kemungkinan *outlier* pada nilai 60,000 ke atas.

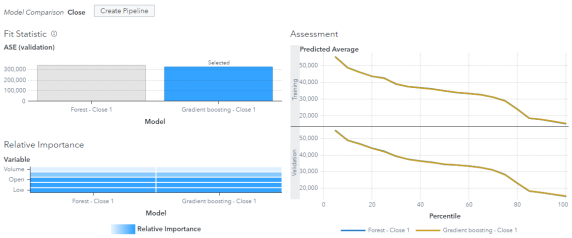
D. Model Comparison



Gambar 14. Model Comparison

Dari perbandingan antara ketiga model prediktif yang telah dipaparkan, *linear regression* dinyatakan sebagai model prediktif terbaik dengan tingkat ASE paling kecil di antara kedua lainnya. ASE sendiri merupakan sebuah tolak ukur penilaian suatu model berdasarkan data validasi dan diteruskan dengan menghitung rata-rata kuadrat residualnya. Jumlah dari rata-rata yang *error* akan dihitung sebagai jumlah dari keseluruhan ASE. ASE memiliki keunggulan dalam kemudahannya dalam mengukur validasi data dan dijelaskan, terutama untuk model yang memiliki respon yang berkelanjutan. Pemilihan ASE sebagai tolak ukur penilaian sesuai dalam menentukan model terbaik dalam memprediksi suatu data. ASE yang dimiliki oleh ketiga model tersebut, diungguli *linear regression* dengan nilai terkecil sekitar 93,954. Posisi lainnya, di *random forest* sekitar 373,661 dan *gradient boosting* sekitar 381,400.

Tetapi, hasil perhitungan akan begitu tertimpang jika *linear regression* merupakan model dengan tipe regresi. Maka, perbandingan model lainnya dapat dilakukan tanpa *linear regression* dengan membandingkan model berdasarkan basis *data mining* yang sama menggunakan *decision tree*, yaitu *random forest* dan *gradient boosting*.



Gambar 15. Model Comparison 2

Terlihat bahwa tanpa adanya *linear regression,* basis model yang menggunakan *decision tree* sebagai klasifikasi kelasnya, menunjukkan *gradient boosting* lebih baik dibanding *random forest* dengan nilai perbandingan ASE yang begitu tipis. *Gradient boosting* di sekitar 327,080 dan *random forest* di sekitar 341,681.

V. Conclusion (Kesimpulan)

Berdasarkan analisis data dan pembahasan yang telah dipaparkan, model yang baik digunakan untuk memprediksi nilai dari harga penutupan di *data set* tersebut merupakan model *linear regression*, dengan variabel dependen berupa Close, dan independen berupa Open, High, Low, Stock Trading, dan Volume.

Harga penutupan berjarak sekitar dari 14,000 hingga 62,000. Rata-rata harga penutupan adalah 34,000. Sebagian besar data lain (986 dari 1,233) memiliki harga penutupan antara 17,000 dan 47,000. Harga tinggi terbaik membedakan yang tertinggi (10% teratas) dan terendah (10% terbawah) harga penutupan. Namun, terdapat satu data yang mungkin merupakan *outlier*, dengan Close lebih besar dari atau sama dengan 62,000.

Variabel Close/harga penutupan mungkin memiliki hubungan positif yang kuat dengan variabel High/harga tertinggi. Kedua hubungan variabel tersebut membentuk sebuah hubungan linier. Untuk setiap 1 peningkatan di High, Close meningkat juga sebesar 0,988. Rata-rata High terdapat pada sekitar 34,000, atau berkisar dari 14,000 hingga 62,000.

Referensi

[1] Pujianto, A., Mulyati, A., & Novaria, R. (2018). Pemanfaatan Big Data dan Perlindungan Privasi Konsumen di Era Ekonomi Digital. *Majalah Ilmiah BIJAK, 15(2)*, 127-137.

[2] Davenport, T. H., Barth, P., & Bean, R. (2012). How 'Big Data' is Different. *MIT Sloan Management Review, FALL 2012, 54(1),* 22.

[3] Davenport, T. H., & Dyché, J. (2013). Big Data in Big Companies. *International Institute for Analytics*.

[4] SAS, "Big Data: What it is and why it matters", diakses pada 9 Juni 2022 dari https://www.sas.com/en_id/insights/big-data/what-is-big-data.html#:~:text=Big%20data%20is%20a%20term,with%20the%20data%20that%20matters.

[5] Maryanto, B. (2017). Big Data dan Pemanfaatannya dalam Berbagai Sektor. *Media Informatika*, 16(2), 14-19

[6] Watson, H. J. (2014). Tutorial: Big Data Analytics: Concept, technology and application. *Communication for the association for information systems, 34*(65), 1247-1268.

[7] Kotni, V. (2011). Impact of retail services on retail sales. *Journal of Business and Retail Management Research (JBRMR),* 6(1).

[8] Salwis, & Niu, F. A. L. (2018). Analisis Harga Saham Perusahaan *Retail* Di Bursa Efek Indonesia Yang Menerapkan *E-Commerce*. *Journal Economic and Business Of Islam*. *3(1).* 1-12.

[9] Muliono, R. (2017). Implementasi Algoritma Apriori Pada Data Benchmark Kosarak Dan Mushrooms. *JITE : JOURNAL OF INFORMATICS AND TELECOMMUNICATION ENGINEERING, 1*(1), 34-41.

[10] E, M., & S, K. (2015). Penerapan Metode K-Means Untuk Clustering Produk Online Shop Dalam Penentuan Stok Barang. *Jurnal Bianglala Informatika, 3*(1).

[11] Herwanto, H. W., Widiyaningtyas, T., & Indriana, P. (2019). Penerapan Algoritme Linear Regression untuk Prediksi Hasil Panen Tanaman Padi. *JNTETI, 8(4)*.

[12] Han, J. (2012). Data Mining Concepts and Techniques Third Edition. *USA:Elsevier*.

[13] Breiman, L. (2001). Machine Learning. *Berkeley: University of California*.

[14] Haristu, R. (2019). PENERAPAN METODE RANDOM FOREST UNTUK PREDIKSI *WIN RATIO* PEMAIN *PLAYER UNKNOWN BATTLEGROUND*. *UNIVERSITAS SANATA DHARMA*.

[15] Essam, A. D. (2019). Comparison between XGBoost, LightGBM and CatBoost Using a Home Credit Dataset. International Journal of Computer and Information Engineering, 13(1), 6–10.

[16] Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics*.

[17] Suryana, S. E., Warsito, B., & Suparti. (2021). PENERAPAN *GRADIENT BOOSTING* DENGAN *HYPEROPT* UNTUK MEMPREDIKSI KEBERHASILAN TELEMARKETING BANK. *Jurnal Gaussian, 10(4),* 617-623.