PEMODELAN RAMALAN PASARAN SAHAM MENGGUNAKAN ALGORITMA PEMBELAJARAN MESIN

AWANGKU MUHAMMAD AFFIF OMAR BIN AWANG RADUAN

UNIVERSITI KEBANGSAAN MALAYSIA

PEMODELAN RAMALAN PASARAN SAHAM MENGGUNAKAN ALGORITMA PEMBELAJARAN MESIN

AWANGKU MUHAMMAD AFFIF OMAR BIN AWANG RADUAN

PROJEK YANG DIKEMUKAKAN UNTUK MEMENUHI SEBAHAGIAN
DARIPADA SYARAT MEMPEROLEH
IJAZAH SARJANA SAINS DATA

FAKULTI TEKNOLOGI DAN SAINS MAKLUMAT
UNIVERSITI KEBANGSAAN MALAYSIA
BANGI

2024

PENGAKUAN

Saya akui karya ini adalah hasil kerja saya sendiri kecuali nukilan dan ringkasan yang tiap-tiap satunya telah saya jelaskan sumbernya.

16 Februari 2024

AWANGKU MUHAMMAD AFFIF OMAR BIN AWANG RADUAN
P117990

PENGHARGAAN

Bersyukur kehadrat Allah s.w.t kerana dengan limpah kurnianya projek Pemodelan Algoritma Ramalan Pasaran Saham Menggunakan Pembelajaran Mesin dapat disempurnakan. Setinggi-tinggi penghargaan dan ribuan terima kasih diucapkan kepada Dr. Afzan Binti Adam selaku penyelia yang telah memberi nasihat dan tunjuk ajar dengan penuh kesabaran. Ribuan terima kasih juga diucapkan kepada Fakulti Teknologi Sains Dan Maklumat kerana telah memberi kemudahan dan bantuan teknikal bagi menyiapkan projek ini.

Ucapan terima kasih tidak terhingga juga kepada keluarga tercinta terutama Ibu dan Bapa saya dengan berkat doa dan kata-kata sokongan yang sering diberikan sepanjang tempoh saya menyiapkan projek ini. Terima kasih atas kepercayaan yang diberikan kepada saya bagi menyiapkan projek ini. Akhir sekali, terima kasih juga kepada sahabat seperjuangan yang sering memberi kata semangat sepanjang tempoh saya menyiapkan projek ini.

ABSTRAK

Kaedah tradisional menganalisis ciri-ciri asas dan teknikal telah lama digunakan untuk meramalkan nilai pasaran saham. Dengan adanya model pembelajaran mesin, ramalan mengenai pasaran saham lebih tepat dan mudah. Pasaran saham menggunakan model algoritma pembelajaran mesin termasuk rangkaian neural buatan, mesin vektor sokongan, dan regresi linear telah dibangunkan. Bagi meramalkan perubahan harga saham dalam kesusasteraan akademik, tetapi tidak banyak kajian yang menggunakan set data terkini. Objektif utama kajian ini adalah untuk meramal pergerakan pasaran saham menggunakan teknik Pembelajaran Mesin (ML). Selain itu, analisis perbandingan prestasi model ramalan harga saham akan dibuat bagi mencari model yang tepat untuk kajian ini. Kaedah ML yang telah dihasilkan dan dicadangkan dalam kajian ini untuk membuat analisis dan ramalan pasaran saham secara mudah ialah Rangkaian Neural Buatan (ANN), Mesin Vektor Sokongan (SVM) dan Hutan Rawak (RF). Data pasaran saham daripada Apple Inc dengan label AAPL diperoleh daripada laman web Kaggle dan digunakan sebagai set data siri masa dalam kajian ini. Tarikh, Buka, Tutup, Tinggi, Rendah, Volum, Dividen dan Pembahagian Saham ialah lapan atribut yang secara kolektif membentuk set data. Dengan saiz sampel 1440, kajian ini memberi tumpuan dari tahun 2018 hingga 2023. Bagi meramalkan nilai harga tutup dalam kajian ini, analisis univariat telah digunakan. Kajian ini dijalankan dalam tiga langkah, bermula dengan pembersihan dan penyediaan data pasaran saham. Pada peringkat kedua, tiga teknik pembelajaran mesin iaitu ANN, SVM, dan RF telah digunakan dalam kaedah pengesahan silang dalam RapidMiner. Langkah ketiga ialah peringkat kecekapan ramalan dan penilaian ketepatan untuk tiga model yang dicadangkan. Bagi tujuan ini, RMSE, MAE, MSE, dan ralat relatif adalah empat metrik penilaian prestasi yang digunakan untuk menilai model. Keupayaan model untuk menjangkakan harga tutup saham ditunjukkan oleh nilai rendah keempat-empat petunjuk ini. Menurut keputusan percubaan, model SVM mempunyai kadar ralat terendah apabila ia berkaitan dengan ramalan pasaran saham dengan RMSE sebanyak 0.024+/-0.00, MAE 0.018+/-0.016, MSE 0.01+/-0.001 dan ralat relatif 1.93% +/-1.86%. Kajian ini telah berjaya mencapai objektif yang dinyatakan iaitu menggunakan model algoritma ANN, SVM dan RF untuk ramalan pasaran saham jangka pendek dan membandingkan tiga prestasi algoritma. Kesimpulannya, kajian itu menunjukkan bahawa model SVM yang dicadangkan mempunyai ketepatan yang lebih tinggi daripada ANN dan RF.

STOCK MARKET PREDICTIONS MODELING USING MACHINE LEARNING ALGORITHMS

ABSTRACT

The stock market has long been predicted using conventional techniques that analyse fundamental and technical features. In addition, the stock market has been predicted in the past using machine learning models including support vector machines, neural networks, and linear regression. In the academic literature, there have not been many studies that use this latest dataset to make predictions. The main objective for this study is to anticipate stock market movements using Machine Learning techniques. Additionally, to make analysis and compare the stock price prediction performances of the chosen algorithms and find the best model with the results presented in this study. The ML methods that have been selected in this study for stock market analysis and prediction are the Artificial Neural Network (ANN), Support Vector Machine (SVM), and Random Forest (RF) algorithms. The stock market data from Apple Inc with the label AAPL was obtained from the Kaggle website and utilised as the time series dataset for this study. Open, High, Low, Close prices, Volume, Date, Dividend, and Stock Split were the eight attributes which collectively made up the dataset. With a sample size of 1440, this study focused from year 2018 to 2023. In order to forecast the close prices values in this study, univariate analysis was used. This study was conducted in three steps, beginning with the cleaning and preparation of the stock market data. In the second stage, three machine learning techniques ANN, SVM, and RF were used in a cross-validation method. In the third step would be the prediction efficiency and accuracy evaluation stage for the three suggested models. The RMSE, MAE, MSE, and relative error are the four performance evaluation metrics that are used to assess the models. The three model ability to anticipate the closing price of stocks is demonstrated by the low values of these four indicators. According to the experiment results, SVM models had the lowest error rates when it came to stock market prediction with RMSE of 0.024+/-0.00, MAE 0.018+/-0.016, MSE 0.01+/-0.001 and relative error 1.93% +/-1.86%. The study also succeeded in achieving its stated objectives which were to employ ANN, SVM, and RF algorithms model for short-term stock market prediction and compare the three algorithms' performances. In conclusion, the study highlights that suggested algorithm SVM method has higher accuracy than ANN and RF.

KANDUNGAN

Halaman

SENARAI JADUAL

SENARAI ILUSTRASI

SENARAI SINGKATAN

ANN	Artificial Neural Network
SVM	Support Vector Machine
RF	Random Forest
RMSE	Root Mean Square Error
MSE	Mean Square Error
MAE	Mean Absolute Error
AAPL	Apple Inc.
ML	Machine Learning

BAB I

PENGENALAN

1.1 PENDAHULUAN

Turun naik pasaran saham merupakan corak yang sangat tidak dapat diramalkan menyebabkan kebanyakan peminat atau kaki pasaran saham ingin merebut peluang yang akan membolehkan mereka mendapat lebih banyak manfaat dengan menjangkakan pergerakan pasaran saham yang tepat. Secara umum, pasaran saham adalah dengan definisi pasaran di mana pembeli dan penjual berminat untuk membeli saham syarikat tertentu, dan harga saham ini berubah dengan ketara dari semasa ke semasa (Bansal et al. 2022). Perubahan harga ini adalah berdasarkan keadaan ekonomi semasa, sentimen pelabur terhadap syarikat tertentu, peristiwa politik dalam kerajaan sesebuah negara dan faktor lain hanyalah beberapa perkara yang tidak diketahui dan faktor yang mungkin mempengaruhi nilai pasaran saham pada hari tertentu. Ini akan menjadikannya sangat sukar untuk meramalkan pergerakan dalam kos pasaran saham (Bansal et al. 2022).

(Subasi et al. 2021a) melaporkan, terdapat banyak faktor dan ketidakpastian yang boleh menjejaskan nilai pasaran saham pada hari tertentu. Oleh kerana itu, terdapat kemungkinan bahawa perubahan mendadak dalam pasaran saham boleh mewujudkan perubahan yang tidak dapat diramalkan dalam kadar saham. (Patel et al. 2015) menyatakan bahawa sebelum mula melabur dalam saham, terdapat dua jenis analisis yang harus dilakukan. Pertama, analisis fundamental perlu dilakukan seperti pelabur perlu melihat ke dalam nilai intrinsik saham, prestasi industri, keadaan ekonomi semasa, dan lain-lain. Kedua, untuk analisis teknikal, pelabur perlu mengkaji nilai saham dan statistik yang diperoleh daripada aktiviti pasaran, seperti harga dan bilangan saham yang didagangkan pada hari sebelumnya.

Di dalam era globalisasi kini, terdapat banyak pekerjaan sebelum ini telah berubah dan diautomasikan termasuk buruh pasaran saham. Pasaran saham kini boleh diramalkan dengan mudah menggunakan beberapa kaedah yang tidak melibatkan kerja manual disebabkan oleh kemajuan teknologi semasa. Algoritma pembelajaran mesin (ML) adalah salah satu kaedah yang paling terkenal dan digunakan secara meluas untuk mencipta model ramalan. (Bansal et al. 2022) menyatakan bahawa ML adalah konsep di mana komputer belajar atau meramalkan perkara dengan bantuan pengetahuan dan latihan masa lalu, tanpa sebarang program luaran yang terlibat. Banyak algoritma telah digunakan untuk meramalkan pasaran saham, pada mulanya linear regresi klasik. Walau bagaimanapun, rangkaian neural buatan (ANN) dan mesin vektor sokongan (SVM) adalah dua algoritma yang kemudiannya digunakan secara meluas (Mintarya et al. 2023a). Setiap algoritma mempunyai cara corak pembelajaran dan meramalkan corak pasaran saham secara tersendiri (Patel et al. 2015).Selain pasaran saham, permintaan bekalan air, industri penjagaan kesihatan, dan disiplin lain semuanya boleh diramalkan menggunakan pelbagai kaedah ML.

Walau bagaimanapun, memilih algoritma yang betul berdasarkan jenis set data dan penggunaan yang dimaksudkan adalah sangat penting. Jenis set data mungkin bergantung pada waktu. Sebagai contoh, ketinggian dan berat kanak-kanak kecil, yang akan berubah seiring berjalannya waktu. Contoh data yang tidak bergantung pada waktu adalah nama seseorang yang akan tetap sama walaupun setelah satu dekad lamanya. Pasaran saham adalah sektor yang sangat bergantung pada masa di mana harga saham biasanya akan berubah setiap minit sebagai hasilnya. Oleh itu, analisis siri masa adalah metodologi yang banyak digunakan secara meluas.

Terdapat empat pendekatan utama yang terkenal yang digunakan untuk menganggarkan nilai pasaran saham iaitu analisis teknikal, ramalan siri masa, ML dan perlombongan data, pemodelan dan ramalan turun naik saham (Khaidem et al. 2016). Berikut adalah sedikit penjelasan secara umum mengenai empat pendekatan utama yang terkenal yang digunakan untuk mengganggarkan nilai pasaran saham. Dengan mengkaji corak statistik yang diperoleh daripada aktiviti perdagangan, seperti pergerakan harga dan jumlah, analisis teknikal adalah disiplin perdagangan yang digunakan untuk menilai pelaburan dan mencari peluang perdagangan. Ramalan siri masa adalah amalan

menganalisis data siri masa dengan statistik dan pemodelan untuk membuat ramalan dan membantu dalam membuat keputusan strategik. ML adalah proses mengajar komputer bagaimana untuk belajar dan memahami parameter yang diberikan, sedangkan perlombongan data direka untuk mengeluarkan peraturan dari sejumlah besar data. Dalam erti kata yang lebih mudah, perlombongan data hanyalah satu proses menjalankan penyelidikan untuk mencapai kesimpulan tertentu berdasarkan jumlah data yang diperoleh. Sebaliknya, ML belajar untuk menjalankan tugas yang mencabar dan menggunakan maklumat dan pengalaman yang dikumpulkan untuk menjadikan sistem lebih pintar.

(Engle & Patton 2001) pula menyatakan bahawa model turun naik sepatutnya dapat meramalkan turun naik kewangan. Hampir semua aplikasi kewangan model turun naik melibatkan meramalkan beberapa ciri pulangan pada masa depan. Model turun naik biasanya digunakan untuk menjangkakan magnitud pulangan mutlak, tetapi ia juga boleh digunakan untuk meramalkan kuantil atau termasuk juga dengan keseluruhan ketumpatan. Metodologi yang akan dibincangkan di dalam kertas kerja ini memberi tumpuan lebih kepada aplikasi ML dan Perlombongan Data di pasaran saham.

Terdapat beberapa penyelidikan baru-baru ini menunjukkan bagaimana ML dapat meningkatkan ramalan pasaran saham. Antara teknik ML yang dapat menambah baik ramalan pasaran saham adalah SVM dan hutan rawak (RF) (Liaw & Wiener 2002). Selain daripada SVM dan RF, rangkaian neural buatan (ANN), rangkaian saraf perlingkaran (CNN), rangkaian saraf berulang (RNN), dan rangkaian saraf dalam seperti memori jangka pendek panjang (LSTM) adalah beberapa teknik berasaskan rangkaian saraf yang telah menunjukkan hasil positif meramal pasaran saham (Oyeyemi et al. 2007 ; Li et al. 2017).

1.2 PENYATAAN MASALAH

Pemodelan dan ramalan pasaran kewangan telah menarik perhatian daripada kalangan ahli akademik dan penyelidik dari pelbagai bidang pengajian. Pasaran kewangan adalah idea khayalan di mana pertukaran antara pembeli dan penjual berlaku untuk aset kewangan seperti saham, kekukuhan, dan logam berharga. Corak ramalan atau harga

saham menggunakan teknik ML dan rangkaian saraf tiruan adalah masalah yang paling menarik untuk diteliti dalam persekitaran pasaran kewangan semasa terutamanya di pasaran saham. Tahap gangguan hingar yang tinggi, saiz sampel yang terhad, data yang tidak bergerak, dan data tidak linear membuat ramalan kewangan sesuatu tugas pemprosesan isyarat yang sungguh mencabar. Jurang maklumat yang tidak lengkap antara jumlah perdagangan saham pada masa lalu dan harga masa depan dirujuk sebagai ciri yang hingar.

Ramalan harga saham adalah usaha yang sukar dilakukan kerana ia bergantung kepada pelbagai pembolehubah, seperti iklim politik, ekonomi global, laporan kewangan korporat, prestasi, dan lain-lain lagi. Oleh itu, kaedah untuk menjangkakan nilai saham dengan berdasarkan melihat corak sejak beberapa tahun sebelumnya telah terbukti cukup membantu untuk membuat pergerakan pasaran saham, menjanakan keuntungan dan mengurangkan kerugian.

1.3 SOALAN PENYELIDIKAN

Berikut adalah soalan-soalan yang akan dikaji dan disiasat dalam projek ini:

1. Apakah pendekatan perlombongan data terbaik untuk meramalkan pasaran saham?

2. Apakah corak harga pasaran saham jangka pendek pada masa hadapan menggunakan model ANN, SVM dan RF akan memberikan ketepatan tertinggi?

1.4 OBJEKTIF KAJIAN

Berdasarkan tujuan kajian ini, ia akan memberi tumpuan kepada pelaksanaan menggunakan algoritma SVM, ANN dan RF menggunakan set data pasaran saham yang terdapat di laman web Kaggle.com dan membuat perbandingan dengan setiap algoritma untuk menunjukkan kesimpulan yang boleh dipercayai dan mudah difahami untuk para sarjana termasuk di bidang lain untuk melakukannya. Objektif kajian bagi kajian ini adalah seperti berikut :

1. Membangun model ramalan pasaran saham menggunakan algoritma ANN, SVM dan RF.

2. Menganalisis dan mencadangkan antara ketiga-tiga model untuk mencari yang terbaik.

1.5 SKOP KAJIAN

Meramalkan nilai masa depan saham kewangan syarikat adalah matlamat yang utama bagi ramalan pasaran saham. Penggunaan ML dalam teknologi ramalan pasaran saham adalah perkembangan yang baru dan terkini. Teknologi ini menghasilkan ramalan berdasarkan nilai indeks pasaran saham semasa dengan latihan pada nilai data masa lalu mereka. Set data yang digunakan di dalam penyelidikan ini adalah set data yang terdahulu yang telah tersedia ada di Yahoo Finance. Data tersebut telah dikumpulkan dan boleh diambil oleh sesiapa daripada laman web Kaggle untuk digunakan dan di analisa. Di dalam laman web Kaggle terdapat banyak set data daripada pelbagai syarikat dan set data yang akan digunakan bagi penyelidikan ini adalah set data berlabel AAPL. Set data AAPL ini mempunyai data daripada tahun 1980 sehingga 2023.

Set data ini mempunyai data harga saham dalam seminggu daripada hari Isnin sehingga hari Jumaat. Set data ini tidak mempunyai harga saham pada hari Sabtu dan Ahad kerana pada hari tersebut adalah hari cuti. Di dalam penyelidikan ini menggunakan kaedah data analisis univariat. Set data ini mempunyai lapan atribut dan atribut tersebut adalah Tarikh, Buka, Tinggi, Rendah, Tutup, Volum, Dividen, dan Pembahagian Saham. Bagi menganggarkan harga saham menggunakan set data yang dipilih untuk penyelidikan ini, terdapat tiga teknik yang dipilih dan digunakan di dalam kajian ini.

Antara teknik tersebut adalah, ANN, SVM, dan RF. Satu-satunya tujuan kertas ini adalah untuk meramalkan corak jangka pendek dalam penemuan harga saham. Di samping itu, ia akan mengkaji dan membandingkan ANN, SVM, dan RF untuk menentukan algoritma mana yang paling tepat dan berguna untuk meramalkan harga saham masa depan.

1.6 KEPENTINGAN PROJEK

Keupayaan untuk meramalkan arah pasaran saham sama ada naik atau turun boleh digunakan sebagai sistem cadangan awal untuk pelaburan jangka pendek dan ia juga berpotensi dapat digunakan sebagai sistem amaran awal pemegang saham jangka panjang mengenai masalah kewangan yang mungkin berlaku. Salah satu perkara utama dalam kertas kerja ini adalah untuk meramalkan harga pasaran saham dalam jangka pendek dan menunjukkan corak saham. Pertimbangan yang paling penting apabila memilih pendekatan ramalan adalah dengan meramalkan ketepatan. Kajian ini akan mencari antara ANN, SVM dan RF, yang antara ketiga-tiga model tersebut mempunyai ketepatan tertinggi untuk meramalkan harga pasaran saham.

Profesional yang melabur haruslah mengetahui dengan lebih lanjut mengenai ekonomi semasa, pasaran saham, atau sekuriti dengan melakukan analisis saham. Bagi memilih saham yang sesuai untuk perdagangan, sebuah sistem haruslah dibangunkan selepas profesional tersebut telah menilai data pasaran yang lama dan yang semasa. Bagi mengenalpasti titik masuk dan keluar dari segi pelaburan juga merupakan satu lagi komponen analisis saham.

Dengan menggunakan teknik ramalan pasaran saham yang sesuai dapat membolehkan seseorang individu itu menganggarkan titik masuk dan keluar mereka dengan lebih tepat. Oleh kerana peniaga sering masuk atau keluar dari pasaran pada saat-saat yang salah atau tidak tepat, mereka tidak dapat memanfaatkan semua peluang mereka yang berpotensi menjana keuntungan.

1.7 ORGANISASI PROJEK

Di dalam Bab 1 terdapat penjelasan secara umum mengenai kertas kajian ini. Gambaran keseluruhan idea yang dikeluarkan untuk kertas kajian ini turut disediakan di dalam bab ini. Bab ini juga memberikan penerangan umum mengenai isu yang sering dihadapi oleh penyelidik terdahulu bagi meramal pasaran saham. Matlamat kajian yang ingin diketengahkan dan juga kepentingan bagi melaksanakan projek ini juga turut dibincangkan di dalam bab ini.

Dalam Bab 2, kertas kerja ini mengkaji beberapa kesusasteraan yang dikaji semula yang berkaitan dengan Ramalan Pasaran Saham menggunakan ML atau bahkan menggunakan kaedah lain. Pengenalan ringkas mengenai ANN, SVM dan RF dan set data yang telah digunakan oleh penyelidik lain. Kajian dan ringkaskan hasil yang diperoleh daripada penyelidikan terdahulu yang menggunakan pendekatan ML dan bandingkan prestasi pendekatan yang berbeza.

Dalam Bab 3, kajian ini akan membentangkan tentang metodologi penyelidikan. Bab ini akan menerangkan langkah demi langkah pendekatan yang digunakan untuk penyelidikan ini dan menyediakan set data yang digunakan dalam eksperimen ini. Keperluan perkakasan dan perisian penyelidikan ini akan disertakan sekali di dalam bab ini. Di sini juga akan menerangkan pengukuran prestasi penyelidikan dan akan menunjukkan perbandingan antara ANN, SVM dan RF berserta dengan penerangan carta alir proses metodologi.

Dalam Bab 4, akan membincangkan hasil percubaan dan membincangkan ketepatan terbaik antara tiga model. Selepas itu, bab ini membentangkan perincian eksperimen ramalan pasaran saham menggunakan ML dan membuat perbandingan tiga model yang dipilih untuk eksperimen ini iaitu ANN, SVM dan RF. Bab ini merangkumi perincian eksperimen untuk membangunkan ramalan pasaran saham yang dicadangkan menggunakan ANN, SVM dan RF. Akhir sekali, Bab 5 akan memaparkan kesimpulan umum berdasarkan kepada kajian yang telah dilakukan, sumbangan penyelidikan, kekuatan dan batasan kajian, dan membincangkan beberapa pendekatan cadangan yang sesuai untuk sebarang kerja kajian pada masa hadapan.

BAB II

KAJIAN KESUSASTERAAN

2.1 PENGENALAN

Bab dua ini akan menerangkan lebih lanjut mengenai kajian kesusasteraan masa lalu yang dilakukan oleh orang lain yang berkaitan dengan kajian mengenai ramalan pasaran saham menggunakan ML. Ini adalah untuk membentangkan latar belakang dan gambaran keseluruhan kesusasteraan yang diperlukan mengenai ramalan pasaran saham yang berkaitan dengan matlamat kajian. Terdapat beberapa bahagian yang dibahagikan dalam bab dua iaitu untuk bahagian 2.2 adalah mengenai teknik yang digunakan dalam ML. Bagi bahagian 2.2 akan mempunyai dua lagi bahagian yang dibahagikan dengan lebih kecil iaitu 2.2.1 adalah mengenai pembelajaran yang diselia dan 2.2.2 akan menerangkan tentang pembelajaran yang tidak diselia. Bahagian 2.3 akan menerangkan tentang model yang akan digunakan untuk kajian ini iaitu ANN, SVM dan RF dan akan dibahagikan kepada tiga bahagian yang lebih kecil. Selepas itu, dalam bahagian 2.4 akan menunjukkan secara terperinci tentang kerja yang berkaitan dengan kajian ini. Seterusnya, 2.5 menerangkan penetapan kajian bagi kajian ini. Selepas itu, 2.6 menerangkan tentang kajian yang pernah dilakukan oleh penyelidik terdahulu dan yang terakhir ialah bahagian 2.7 bagi menyimpulkan pelajaran yang diperoleh dari kajian – kajian terdahulu dan bagaimana kajian ini akan diteruskan.

2.2 TEKNIK PEMBELAJARAN MESIN

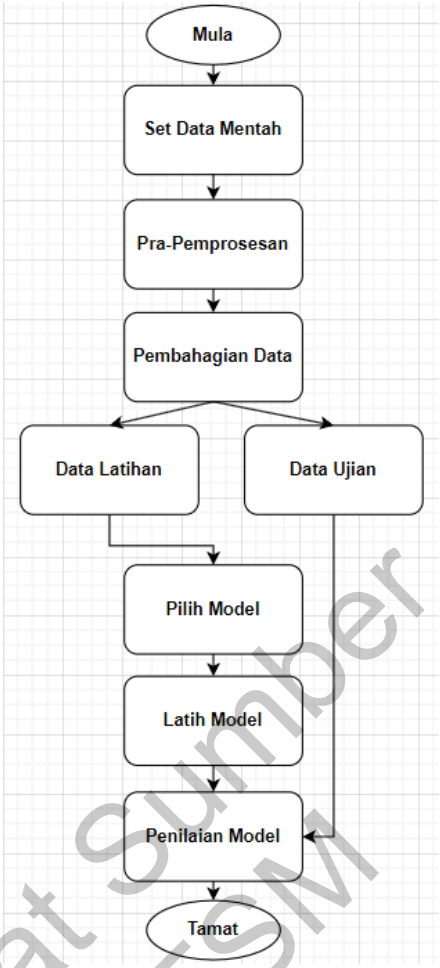
ML merupakan sebuah subset teknik kecerdasan buatan, yang membolehkan mesin menjadi lebih bagus dalam melakukan tugas dari masa ke masa. Ia juga boleh digunakan untuk mengautomasikan sesuatu pekerjaan dengan mengarahkan komputer untuk membina program sendiri dan membolehkan data melakukan kerja. Pembelajaran

terselia, pembelajaran tidak terselia, dan pembelajaran pengukuhan adalah tiga jenis kategori di mana ML boleh dipisahkan.

2.2.1 Pembelajaran Terselia

Pembelajaran yang terselia adalah subset kecil daripada ML di mana hasil data akan diramalkan oleh mesin menggunakan data latihan berlabel yang baik telah dilatih. Istilah "data berlabel" merujuk kepada maklumat yang telah dikaitkan dengan hasil yang diinginkan. Dalam pembelajaran yang terselia, data latihan yang diberikan kepada mesin berfungsi sebagai penyelia yang melatih mesin untuk meramalkan output dengan tepat. Ia mengikuti idea yang sama seperti bahawa seorang murid akan belajar di bawah bimbingan guru. Kaedah pembelajaran terselia melibatkan pemberian model ML data input yang betul beserta data output yang betul. Matlamat algoritma pembelajaran terselia adalah untuk mengenal pasti fungsi pemetaan yang akan menyambungkan pemboleh ubah data masuk dengan pemboleh ubah hasil keluaran (Anon. 2021a).

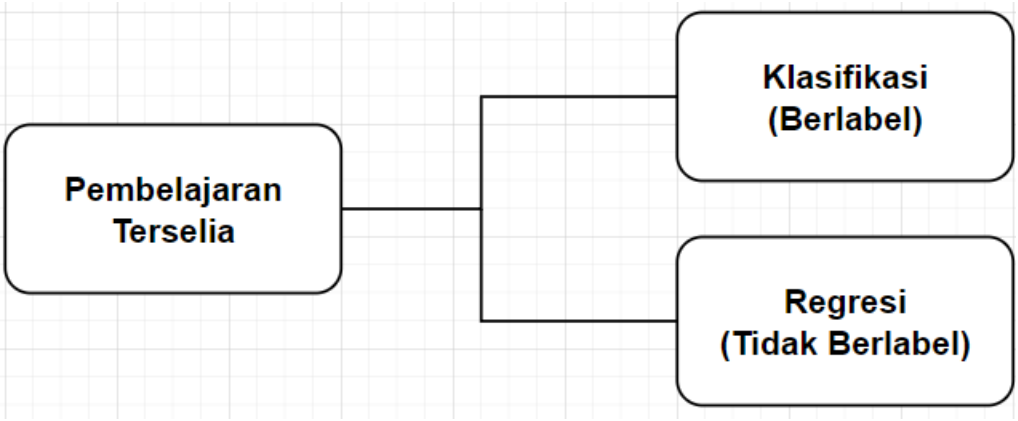
Berdasarka daripada Rajah 2.1 di bawah menunjukkan permulaan sehingga akhir langkah pembelajaran terselia. Bagi memulakan pembelajaran terselia pada mulanya haruslah memilih jenis set data apakah yang ingin digunakan. Set data tersebut haruslah mempunyai atribut yang telah dilabel atau dikelaskan. Seterusnya data tersebut haruslah melalui proses pra-pemprosesan di mana data tersebut akan dibersihkan sehingga ia boleh digunakan oleh model untuk dianalisa. Langkah seterusnya adalah data tersebut akan dibahagiakn kepada dua bahagian iaitu kepada set data latihan dan set data ujian. Selepas itu, pemilihan model yang sesuai haruslah dilakukan dan model tersebut haruslah model yang sesuai dengan set data. Terdapat banyak jenis model yang boleh digunakan bagi pembelajaran terselia seperti mesin vektor sokongan, pokok keputusan, regrasi linear dan banyak lagi.



Rajah 2.1 Rajah Pembelajaran Terselia

Adaptasi daripada: (Kumbure et al. 2022)

Seterusnya, model algoritma yang dipilih akan dilatih menggunakan set data latihan yang telah dibahagiakan. Pada setiap model algoritma akan memberi bacaan atau hasil yang berbeza - beza. Selepas model tersebut dilatih, ia akan digunakan pada set data ujian untuk membuat ramalan. Setiap model algoritma akan memberi bacaan ramalan yang berbeza. Akhir sekali, dengan bacaan hasil ramalan bagi setiap model algoritma yang dipilih akan melalui proses penilaian di mana setiap bacaan ramalan akan dibandingkan dan bacaan ramalan yang mempunyai ketepatan ramalan yang paling tepat.



Rajah 2.2 Jenis Pembelajaran Terselia

Sumber: (Anon. 2021a)

Terdapat dua jenis pembelajaran terselia dalam ML:

1. Regresi

   Sekiranya terdapat hubung kait antara pembolehubah data masuk dan data keluar bermakna prosedur regresi telah digunakan. Ia digunakan untuk meramalkan pembolehubah berterusan seperti cuaca, corak pasaran, dan lain-lain lagi. Antara algoritma regresi yang terkenal dan sering digunakan adalah regresi linear, pokok regresi, regresi bukan linear, regresi linear bayesian, dan regresi polinomial.

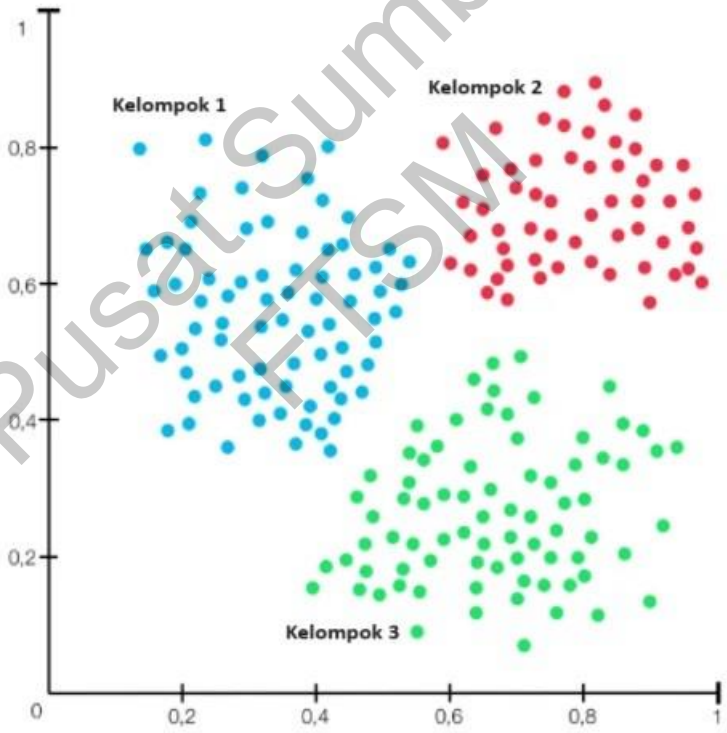
2. Pengelasan

   Apabila pemboleh ubah data hasil keluar adalah berkategori dan terdapat hanya dua kelas pengelasan digunakan. Contohnya seperti Ya-Tidak, Lelaki-Perempuan, Benar-Palsu, dan lain-lain. Kaedah pengelasan sering digunakan untuk menapis dan membezakan e-mail Spam. Algoritma pengelasan yang terkenal dan sering digunakan adalah hutan rawak, pokok keputusan, regresi logistik, mesin vektor sokongan.

2.2.2 Pembelajaran Tidak Terselia

Pembelajaran tanpa penyeliaan, kadang-kadang dirujuk sebagai penemuan pengetahuan, menggunakan data latihan yang tidak dilabelkan, tidak dikelaskan atau dikategorikan. Objektif utama pembelajaran tanpa penyeliaan adalah untuk mengenal

pasti corak tersembunyi dan menarik dalam data yang tidak dilabel. Teknik pembelajaran tanpa penyeliaan adalah berbeza dengan pembelajaran terselia. Ia tidak boleh digunakan untuk menyelesaikan masalah regresi atau pengelasan secara langsung kerana teknik ini tidak diketahui nilai hasil pengeluarannya. Pendekatan pembelajaran tanpa penyeliaan yang paling popular yang digunakan untuk menjalankan analisis data dan menemui corak atau kumpulan tersembunyi dalam data adalah pendekatan berkelompok. Analisis jujukan gen, penyelidikan pasaran, dan pengiktirafan objek adalah beberapa contoh aplikasi analisis berkelompok. Dalam pembelajaran tanpa penyeliaan, pengelompokan, pengesanan anomali, rangkaian saraf, dan kaedah untuk membangunkan model pembolehubah yang tersembunyi adalah antara algoritma yang sering digunakan (El Bouchefry & de Souza 2020).



Rajah 2.3 Rajah Pendekatan Berkelompok

Sumber: (Marzell 2021)

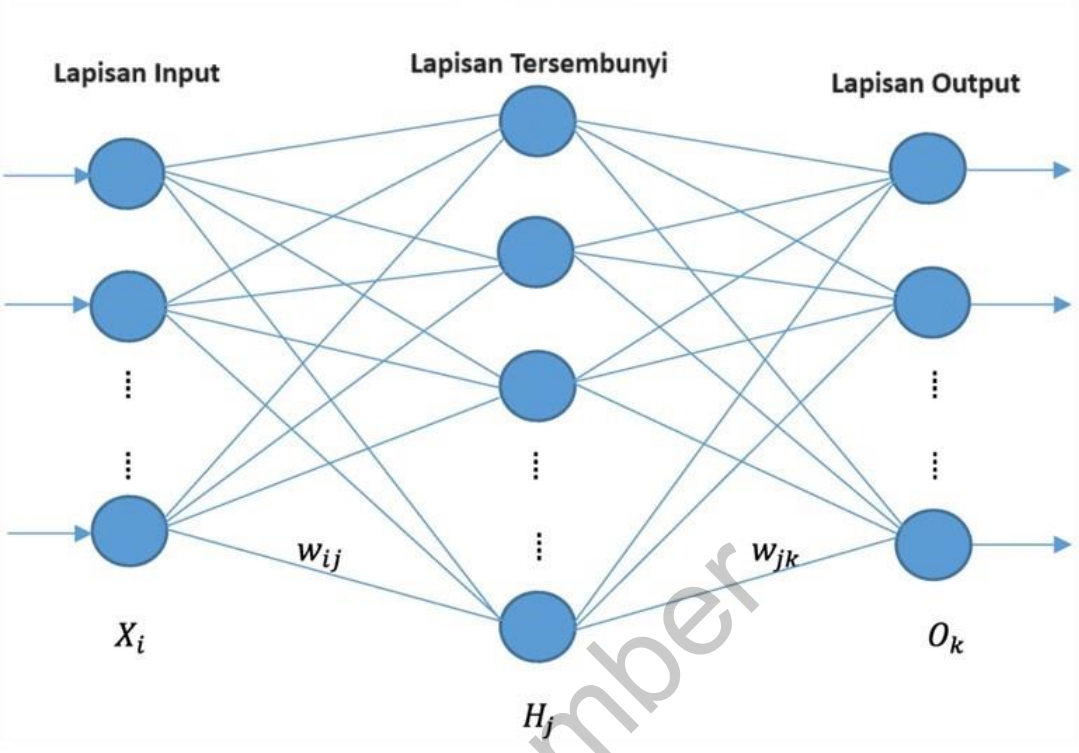
2.3 KAEDAH PENGELASAN

Terdapat banyak jenis kaedah pengelasan yang wujud dan telah digunakan untuk meramalkan turun naik pasaran saham. Tetapi di dalam kajian ini hanya tiga jenis kaedah pengelasan yang terkenal dan akan digunakan. Kaedah pengelasan tersebut adalah ANN, SVM dan RF.

2.3.1 Rangkaian Neural Buatan (ANN)

Rangkaian neural buatan (ANN) atau rangkaian saraf adalah algoritma pengiraan. Ia direka untuk meniru bagaimana "neuron" dalam sistem biologi manusia atau haiwan mungkin berkelakuan. Model komputer yang dipanggil ANN dimodelkan selepas sistem saraf pusat haiwan. Selain pengecaman corak, ia juga mampu melakukan ML. Ini diwakili sebagai rangkaian "neuron" bersambung yang boleh mengira nilai daripada input. Graf berorientasi adalah apa itu rangkaian saraf. Ia terdiri daripada nod yang disambungkan oleh lengkok yang menggunakan analogi biologi mewakili neuron. Ia menyerupai dendrit dan sinaps. Manakala pada setiap nod, setiap lengkok mempunyai berat yang berkaitan dengannya.

Mengunakan nilai yang diperoleh nod sebagai input dan tentukan fungsi pengaktifan di sepanjang lengkok masuk dengan melaraskan untuk berat lengkok. Kaedah pemprosesan maklumat adalah ANN. Ia berfungsi sama seperti bagaimana otak manusia mengendalikan maklumat. Sebilangan besar unit pemprosesan yang saling berkaitan membentuk ANN, yang bekerjasama untuk memproses data. Menggunakan kaedah ini juga mendapat hasil yang berguna daripadanya. Perlombongan data menggunakan rangkaian saraf dengan baik. Forensik, kewangan, dan pengiktirafan corak, antara bidang lain adalah beberapa contoh. Selepas latihan yang sesuai, ia juga boleh digunakan untuk pengelasan data dalam jumlah data yang besar (Sharma 2017).



Rajah 2.4 Perwakilan Menyeluruh ANN

Sumber: (Zhong & Enke 2019)

Rajah 2.4 menunjukkan contoh reka bentuk ANN. ANN mempunyai beberapa lapisan:

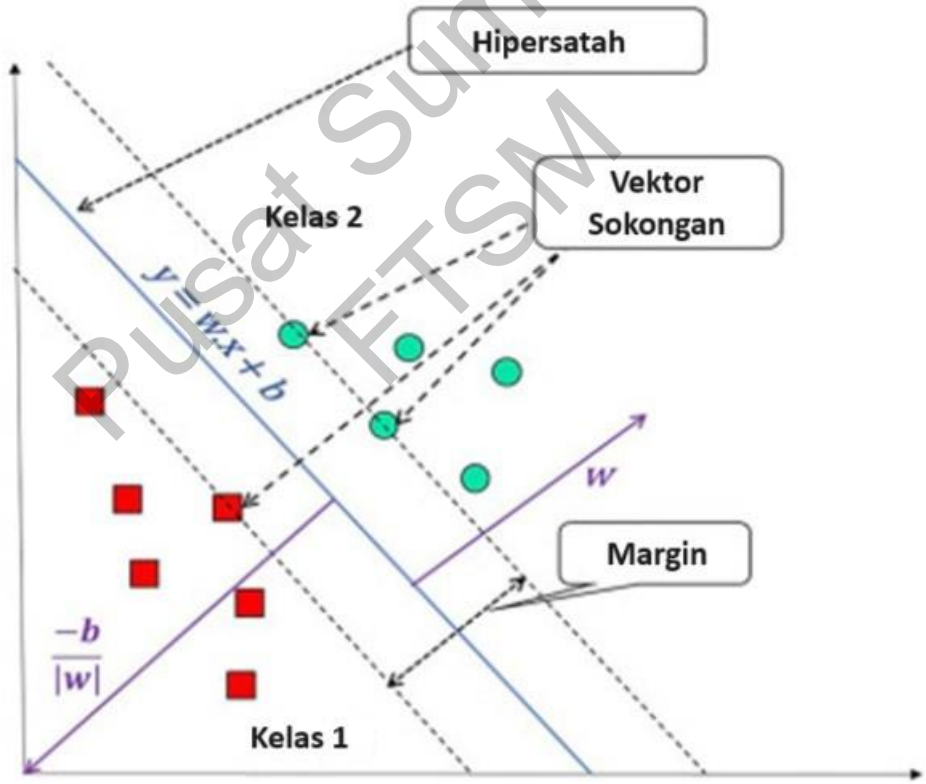
1. Lapisan input mempunyai aktiviti unit input mewakili maklumat mentah yang boleh masuk ke dalam rangkaian.

2. Lapisan tersembunyi adalah untuk menentukan aktiviti setiap unit tersembunyi. Aktiviti unit input dan berat pada sambungan antara input dan unit tersembunyi. Di lapisan ini mungkin terdapat satu atau lebih lapisan tersembunyi.

3. Lapisan output mempunyai tingkah laku unit output bergantung kepada aktiviti unit tersembunyi, berat antara unit tersembunyi dan output.

2.3.2 Mesin Vektor Sokongan (SVM)

Salah satu algoritma yang terkenal untuk pembelajaran terselia ialah Mesin Vektor Sokongan (SVM), dan ia digunakan untuk menyelesaikan masalah pengelasan dan regresi. Kaedah SVM bertujuan untuk membina garis terbaik atau sempadan keputusan yang boleh membahagikan ruang n-dimensi ke dalam beberapa kelas supaya membolehkannya mengelaskan titik data baru dengan cepat dan tepat pada masa akan datang. Istilah "hyperplane" merujuk kepada sempadan keputusan optimum ini. Bagi mencipta hyperplane, SVM memilih titik melampau dan vektor. Vektor sokongan yang merujuk kepada contoh-contoh yang melampau ini adalah nama algoritma yang dikenali sebagai SVM (Anon. 2021b). Rajah 2.5 di bawah menunjukkan contoh reka bentuk SVM.

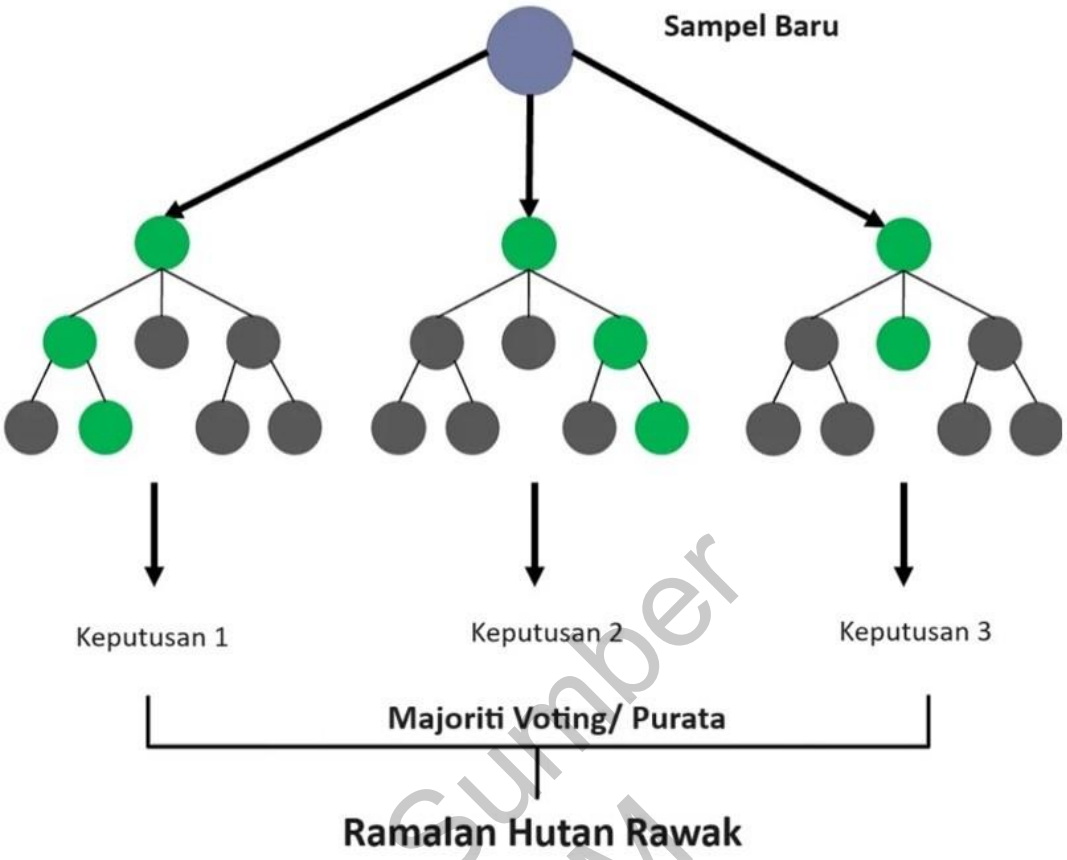


Rajah 2.5 Perwakilan Menyeluruh SVM

Sumber: (Rani et al. 2022)

2.3.3 Hutan Rawak (RF)

Salah satu algoritma yang paling kerap dan sering digunakan di kalangan saintis data ialah hutan rawak (RF). Algoritma ML yang diselia seperti RF sering digunakan dalam isu pengelasan dan regresi. Pada pelbagai sampel, ia menghasilkan pokok keputusan dan menggunakan purata mereka untuk pengelasan dan undi majoriti untuk regresi. Leo Breiman dan Adele Cutler adalah pencipta algoritma ML fleksibel yang dikenali sebagai RF ini. Bagi menjana ramalan atau pengelasan, ia menggunakan kumpulan beberapa pokok keputusan. Algoritma RF menghasilkan hasil yang disatukan dan lebih tepat dengan mengintegrasikan output pokok-pokok ini. Penerimaannya yang luas terhasil daripada faktor kebolehsuaian dan mesra pengguna yang membolehkannya menangani isu pengelasan dan regresi. Kaedah ini amat bagus dalam mengendalikan set data yang rumit dan mengurangkan masalah overfitting dan menjadikannya alat yang berguna untuk pelbagai aplikasi ramalan ML. Keupayaan Algoritma RF untuk mengendalikan set data dengan kedua-dua pembolehubah berterusan seperti dalam hal regresi dan pembolehubah kategori seperti dalam hal pengelasan, adalah salah satu ciri yang paling penting. Bagi tugas yang melibatkan pengelasan dan regresi, FR berfungsi dengan amat baik (Sruthi 2023). Rajah 2.6 di bawah menunjukkan contoh reka bentuk RF.



Rajah 2.6 Perwakilan Menyeluruh RF

Sumber: (Yehoshua 2023)

2.4 PENILAIAN

Jadual 2.1 Matriks Kekeliruan

		Nilai Diramal	
		Ramalan Betul	Ramalan Salah
Nilai Sebenar	Ramalan Betul	Positif Sebenar(TP)	Positif Palsu(FP)
Nilai Sebenar	Ramalan Salah	Negatif Palsu(FN)	Negatif Sebenar(TN)

Matriks kekeliruan adalah teknik untuk menilai prestasi model pengelasan dalam ML yang terselia terutamanya dalam tugas pengelasan binari. Dengan membandingkan ramalan model dengan label kelas yang sebenar, ia memberikan gambaran menyeluruh mengenai ramalan tersebut. Ketepatan, kepersisan, kepekaan, dan skor F1 adalah empat matriks prestasi yang akan digunakan dalam kajian ini untuk menilai ramalan bagi

setiap algoritma. Dengan menggunakan matriks kekeliruan, ia dapat mengira keempat - empat langkah prestasi untuk kajian ini. Jadual 2.1 menunjukkan contoh matriks kekeliruan bagi kajian ini.

Matriks kekeliruan adalah dua kali dua matriks yang terdiri daripada empat bahagian yang berbeza:

Positif Sebenar (TP): bilangan situasi yang diklasifikasikan dengan tepat sebagai tergolong dalam kelas positif.

Negatif Sebenar (TN): bilangan situasi yang dikelaskan dengan tepat sebagai kelas negatif.

Positif Palsu (FP): Bilangan situasi yang salah diklasifikasikan sebagai positif apabila fakta tersebut tergolong dalam kategori negatif.

Negatif Palsu (FN): Bilangan situasi yang salah diklasifikasikan sebagai negatif apabila fakta tersebut tergolong dalam kategori positive.

2.4.1 Ralat Punca Min Kuasa Dua (RMSE)

Berdasarkan daripada (Christie & Neill 2021) menyatakan bahawa ralat punca min kuasa dua (RMSE) boleh dijelaskan sebagai punca kuasa dua min semua kesilapan. RMSE digunakan secara meluas sebagai metrik ralat yang biasa digunakan untuk ramalan berangka. Walaupun ia bergantung pada skala, RMSE adalah metrik yang berguna untuk perbandingan meramalkan ralat antara model atau konfigurasi model untuk pembolehubah tertentu tetapi tidak antara pembolehubah. Formula di bawah ialah persamaan untuk RMSE:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad \ldots(2.1)$$

$y_i - Nilai\ Sebenar$
$\hat{y}_i - Nilai\ Jangkaan$
$n - Bilangan\ Titik\ Data$

2.4.2 Ralat Min Kuasa Dua (MSE)

Berdasarkan daripada (Tyagi et al. 2022) menyatakan bahawa ralat min kuasa dua (MSE) merupakan yang paling kerap digunakan di dalam model regresi dengan menggunakan nilai sasaran berterusan sebagai pemboleh ubah bebas. Ia dinyatakan sebagai perbezaan min kuasa dua antara output yang dihasilkan dan apa yang diramalkan. Jika nilai MSE model itu rendah, ia menunjukkan bahawa model tersebut mempunyai nilai ralat yang rendah. Jika ralat model itu besar , nilai MSE juga akan turut meningkat. Berikut adalah formula bagi MSE:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad \ldots(2.2)$$

$y_i - Nilai\ Sebenar$
$\hat{y}_i - Nilai\ Jangkaan$
$n - Bilangan\ Titik\ Data$

2.4.3 Ralat Mutlak (MAE)

Berdasarkan daripada (Schneider & Xhafa 2022) menyatakan bahawa ralat mutlak (MAE) merupkan salah satu cara pengukuran ralat yang paling kerap digunakan untuk isu regresi. Di dalam penggunaan MAE, nilai meningkat secara linear apabila bilangan ralat meningkat. Purata nombor ralat mutlak digunakan untuk mengira skor MAE. Fungsi matematik yang dipanggil “mutlak” menukarkan nilai kepada positif. Oleh itu, apabila mengira MAE, perbezaan antara nilai yang dijangkakan dan nilai yang sebenar akan sentiasa positif, tidak kira sama ada ia positif atau negatif. Berikut adalah formula ralat mutlak:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad \ldots(2.3)$$

$y_i - Nilai\ Sebenar$
$\hat{y}_i - Nilai\ Jangkaan$
$n - Bilangan\ Titik\ Data$

2.4.4 Ralat Relatif

Berdasarkan daripada (Taylor 1997) dengan menggunakan ralat relatif dapat menunjukkan bagaimana berhampiran nilai yang diukur berbanding kepada nilai sebenar. Peratusan nombor sebenar digunakan untuk menggambarkan ralat relatif. Ketepatan pengukuran ditentukan oleh ralat relatif. Ia boleh dinyatakan sebagai nilai sebenar dibahagikan dengan nilai mutlak perbezaan antara nilai yang diukur dan benar. Kesalahan relatif tidak pernah negatif. Berikut ialah formula ralat relatif:

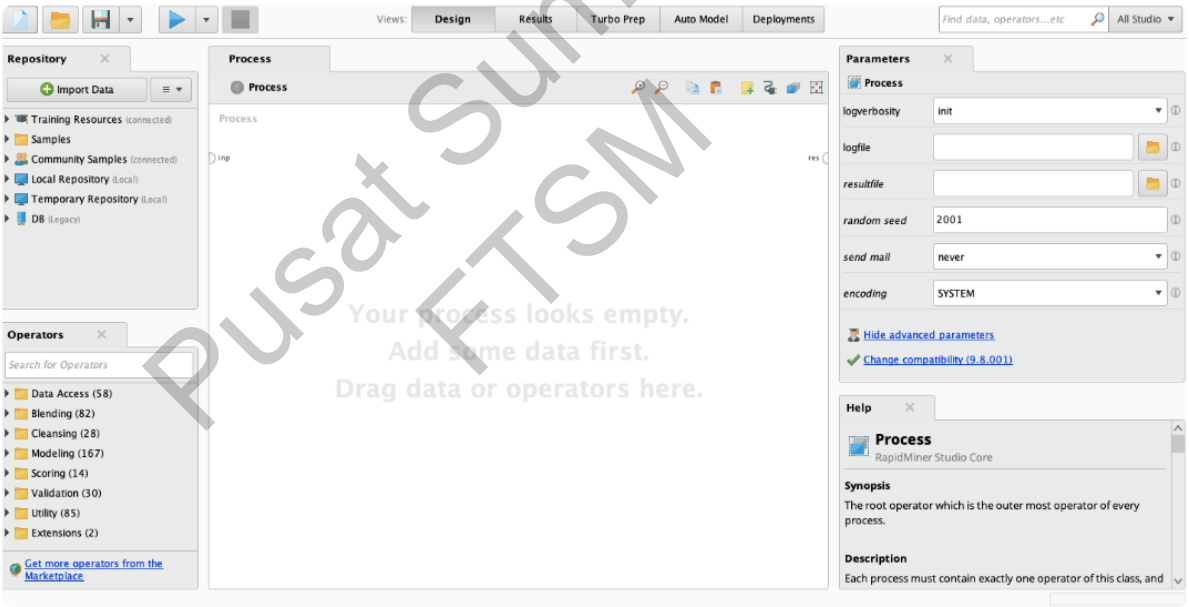
$$Ralat\ Relatif\ = \frac{|y_i - \hat{y}_i|}{|y_i|} \times 100\% \quad \ldots(2.4)$$

y_i – Nilai Sebenar
$\hat{y}_i$ – Nilai Jangkaan

2.5 PENETAPAN EKSPERIMEN

Piawaian persekitaran eksperimen memainkan peranan penting dalam penerapan teknik ML. Oleh itu, adalah lebih baik untuk menerangkan keperluan persekitaran yang digunakan untuk melaksanakan peringkat model meramal sebelum melihat hasil ramalan. Program perisian yang digunakan untuk eksperimen bagi mendapatkan hasilnya dipanggil sebagai RapidMiner dan ia akan dibentangkan dan dijelaskan dalam bab ini. Berdasarkan pada (Troy & Mich 2022) RapidMiner merupakan persekitaran untuk sains data yang mengkaji kesan keseluruhan data sesebuah organisasi. Prosedur untuk perlombongan data dan ML seperti pemuatan dan transformasi data, penyediaan data dan visualisasi, pemodelan statistik dan analisis ramalan, penilaian dan penggunaan juga ditawarkan oleh RapidMiner.

Java adalah bahasa pengaturcaraan yang digunakan untuk membina RapidMiner. Antara muka unit grafik (GUI) ditawarkan oleh RapidMiner bagi mencipta dan menguruskan aliran kerja analisis. Di dalam RapidMiner, aliran kerja tersebut dirujuk sebagai "Proses" dan ia melibatkan banyak "Operator". Dalam proses ini, setiap pengendali menyelesaikan satu tugas dan output dari satu pengendali menjadi input untuk pengendali berikutnya. Secara alternatif adalah dengan menggunakan enjin sebagai API atau menggunakannya dari program lain. Barisan arahan membolehkan panggilan fungsi tertentu. Skrip R dan Python boleh digunakan di dalam skim pembelajaran, model dan algoritma RapidMiner (Norris 2013). Plugin yang boleh diakses melalui pasaran RapidMiner juga boleh digunakan dalam perisian RapidMiner. Pembangun aplikasi boleh berkongsi algoritma analisis data mereka kepada komuniti melalui platform pasaran RapidMiner (Ohri 2011 ; Jungermann 2011).



Rajah 2.7 GUI RapidMiner

Sumber : (Arnaldo 2021)

Rajah 2.7 menunjukkan GUI utama bagi sistem RapidMiner. Bahagian utama RapidMiner adalah di bahagian tengah iaitu bahagian “Proses” dan di bahagian bawah kiri menunjukkan bahagian “Operator”.

2.6 PEMBELAJARAN MESIN DALAM PASARAN SAHAM

Pasaran saham adalah pasaran di mana saham syarikat dengan penyenaraian saham boleh didagangkan. Ia diiktiraf sebagai jenis pasaran sekunder. Sebuah firma mesti mewujudkan kehadiran di salah satu bursa saham yang diiktiraf sebelum ia dapat menyenaraikan sahamnya untuk dijual kepada pelabur di pasaran terbuka. Sebaik sahaja proses ini telah berjaya dilakukan, promoter mesti menjual sebahagian besar saham syarikat kepada pelabur runcit di khalayak ramai, selepas itu perdagangan tambahan boleh berlaku di pasaran sekunder atau bursa saham. Di pasaran saham, mungkin terdapat persaingan antara dua atau lebih perniagaan bursa saham dengan fungsi dan amalan perdagangan yang sama. Kebanyakan dagangan pasaran saham berlaku melalui akaun dagangan dan dematerialisasi. Orang awam boleh mendapat manfaat daripada keupayaan bursa saham untuk menyalurkan dan mengumpulkan simpanan mereka sementara perniagaan mendapat manfaat daripada kemasukan modal ke dalam usaha mereka.

Berdasarkan pada kajian daripada (Vijayarani et al. 2020), ia adalah untuk menerangkan dan mengkaji semula kaedah yang lebih praktikal untuk meramalkan pembangunan saham dengan lebih tepat. Kajian ini menggunakan set data perbelanjaan pasaran saham dari tahun sebelumnya. SVM dan algoritma RF telah digunakan dalam set data kajian ini. Kajian yang dicadangkan juga menunjukkan bagaimana rangka kerja ramalan digunakan dalam konteks praktikal dan bagaimana kesukaran dengan ketepatan atribut umum bekait. Dengan membandingkan keberkesanan pelbagai algoritma didapati bahawa algoritma RF adalah algoritma yang paling munasabah untuk meramalkan harga pasaran saham berdasarkan banyak titik data dari data sejarah.

Kajian yang dilakukan oleh (Mittal et al. 2018) memberi tumpuan kepada ramalan harga pasaran pelbagai cryptocurrency berdasarkan corak sejarah mereka. Dengan mengkaji faktor-faktor yang berkaitan dengan harga mata wang kripto, kajian ini cuba untuk mengetahui dan mengesan corak harian di pasaran untuk mata wang kripto. Bagi menjalankan kajian ini, set data dengan lebih daripada sembilan ciri yang berkaitan dengan data harga cryptocurrency yang dikumpulkan setiap hari dalam tempoh enam bulan telah digunakan. Perubahan harian dalam harga mata wang kripto diramalkan menggunakan model regresi linear. Selain itu, menurut penyelidikan yang

dibentangkan oleh (Ghani et al. 2019), ia ditentukan bahawa alat statistik terbaik untuk graf dan gambaran berjadual hasil ramalan ialah Microsoft Excel dengan tumpuan khusus pada Regresi Linear (LR), Purata Pergerakan Tiga Bulan (3MMA), Melicinkan Eksponen (ES), dan Ramalan Siri Masa. Data untuk saham Google (GOOG), Apple (AAPL), dan Amazon (AMZN) diperolehi daripada Yahoo Finance digunakan untuk kajian ini. Selepas melaksanakan kajian ini, didapati bahawa hasil daripada penggunaan algoritma ES telah dapat menjangkakan hala tuju pasaran saham dengan tepat untuk bulan yang akan datang dengan baik berbanding LR dan 3MMA.

Berdasarkan kajian dilakukan oleh (Pranav & Kumar 2023), kajian ini adalah untuk mencipta teknik ramalan pasaran saham inovatif yang membandingkan algoritma regresi linear dengan algoritma regresi tambahan. Regresi Tambahan dan Regresi Linear adalah dua teknik perlombongan data yang menjadikannya mudah untuk menganalisis dan meramalkan harga saham digunakan dalam kajian ini. Dengan menggunakan semua kaedah moden, kajian ini menunjukkan bahawa algoritma yang dicadangkan iaitu Algoritma Regresi Linear mempunyai ketepatan yang lebih tinggi daripada Algoritma Regresi Tambahan. Selepas itu, kajian yang dilakukan oleh (Madeeh & Abdullah 2021) menunjukkan bahawa untuk menggunakan teknik ML yang berkesan bagi mencipta model yang boleh dipercayai untuk ramalan pasaran saham. K-Nearest Neighbour (K-NN) dan Random Forest (RF) merupakan dua pendekatan ML yang terselia digunakan dalam kajian ini untuk menjangkakan pergerakan pasaran saham. Set data yang digunakan bagi kajian ini merupakan set data Bursa Saham New York (NYSE) daripada Kaggle. Set data yang digunakan terdiri daripada enam belas syarikat dan antaranya daripada syarikat yang besar iaitu AAPL, GOOGL, AMZN, MSFT dan lain-lain lagi. Hasil kajian menunjukkan bahawa kedua-dua model yang dicadangkan adalah sangat tepat dengan model RF mempunyai ketepatan ramalan terbaik jika dibandingkan dengan model K-NN untuk ramalan pasaran saham.

Terdapat banyak teknik ML dan algoritma yang telah digunakan dan diperhalusi dari masa ke masa untuk meramalkan beribu-ribu pasaran saham di seluruh dunia. Berdasarkan Jadual 2.2 di bawah akan meringkaskan jumlah rujukan mengenai ramalan pasaran saham yang telah dilakukan termasuk teknik dan kaedah pemilihan ciri yang digunakan.

Jadual 2.2 Jadual Ringkasan Kajian Kesusasteraan

No	Tajuk	Penerangan	Kaedah yang digunakan	Kaedah Pemodelan	Set Data
1	Machine learning for liquidity prediction on Vietnamese stock market (Khang et al. 2021)	Ia menumpukan perhatian kepada mewujudkan model ML khusus untuk meramalkan kecairan. Ia boleh dikatakan bahawa Model LSTM membolehkan ramalan dengan nilai MSE terendah.	Kaedah ML bagi meramal Regrasi	Persepsi Pelbagai Lapisan (MLP), Memori Jangka Pendek Panjang (LSTM) dan model regresi linear.	Bursa Saham Vietnam daripada tahun 2011 – 2019. Bursa Saham Ho Chi Minh (HOSE) dan Bursa Saham Hanoi (HNX).
2	Heterogeneous graph knowledge enhanced stock market prediction (Xiong et al. 2021)	Dengan menggunakan hubungan antara perkataan, peristiwa, dan maklumat kontekstual dalam bahasa kewangan untuk meramalkan pergerakan pasaran saham. Oleh itu, teks kewangan adalah halus, sederhana, dan kasar maklumat boleh diwakili sepenuhnya dalam graf heterogen. Sistem HGM-GIF mempunyai empat bahagian: fungsi graf heterogen berbilang bijirin untuk mencipta graf heterogen, pengekod nod berbilang bijirin untuk pengekodan nod ke dalam pembenaman, agregator maklumat heterogen berurutan untuk pemodelan dan pengumpulan data mengenai corak hubungan maklumat berbilang bijirin, dan peramal pergerakan pasaran saham untuk meramalkan arah pergerakan harga saham berakhir.	Graf Heterogeneos	Pembina graf heterogen pelbagai bijirin. Pengekod nod berbilang bijirin. Pengagregat maklumat heterogen berurutan. Peramal pergerakan pasaran saham.	Artikel berita kewangan yang tersedia secara umum dari Reuters sejak Oktober 2006 sehingga Disember 2015.
3	Machine learning approaches in stock market prediction: A systematic literature review	Kajian ini mengkaji artikel yang relevan mengenai strategi ML untuk ramalan pasaran saham. Kajian ini menjalankan kajian literatur sistematik (SLR) untuk mencapai matlamat. Kajian ini pada mulanya	Kajian Literatur Sistematik (SLR)	NN, SVM, LSTM, RF, KNN	Kajian ini mengkaji tiga puluh kajian yang berkaitan dengan

bersambung…

…sambungan

	(Mintarya et al. 2023)	membangunkan soalan penyelidikan, dan ia mengumpulkan maklumat yang berkaitan dari artikel jurnal dan ulasan kesusasteraan. Kajian ini menganalisa 30 penyelidikan mengenai strategi ML atau model untuk ramalan pasaran saham. Ia menggunakan versi diubahsuai senarai semak Bahan Laporan Pilihan untuk Ulasan Sistematik dan Meta-Analisis (PRISMA) untuk menilai kertas kerja mereka. Kajian ini menggunakan rangkaian saraf dan menyokong mesin vektor dalam strategi mereka.			model ML dan teknik untuk ramalan pasaran saham. Data yang digunakan adalah daripada artikel jurnal dan ulasan kesusasteraan dari tahun 2012 – 2020.
4	Stock market prediction using machine learning (Subasi et al. 2021)	Kajian ini membuat perbandingan ramalan menggunakan input daripada 7 pengelas yang berbeza. Kajian ini menggunakan aplikasi perlombongan data dan ML di pasaran saham sebagai metodologi mereka dalam kajian ini. Di samping itu, penemuan perbandingan dinilai untuk ketepatan. Ia membuat kesimpulan bahawa Bagging dan Hutan Rawak kedua-duanya dilakukan di atas jangkaan semasa menggunakan set data yang bocor.	Ramalan menggunakan ML	KNN, SVM, Pokok Keputusan, RF Bagging, AdaBoost	Sistem Sebut Harga Automatik. Persatuan Peniaga Sekuriti Nasional (NASDAQ), Bursa Saham New York (NYSE), Nikkei, dan Bursa Saham Financial Times (FTSE).
5	The applications of artificial neural networks, support vector machines, and long–short term memory for stock market prediction (Chhajer et al. 2022)	Ini adalah gambaran keseluruhan ML dan kecerdasan buatan sebagai teknik analisis ramalan pasaran saham. Ia merangkumi kelebihan dan kekurangan menggunakan ML untuk meramalkan pasaran saham serta beberapa peluang dan risiko yang berkaitan dengan berbuat demikian. Kajian ini juga mengkaji penggunaan tiga teknologi ML seperti rangkaian neural buatan, mesin vektor sokongan, dan memori jangka pendek dalam ramalan pergerakan pasaran saham.	Ramalan menggunakan ML dan kecerdasan buatan.	ANN, SVM, LSTM	Bursa Saham New York (NYSE)

bersambung…

…sambungan

6	Stock market prediction with high accuracy using machine learning techniques (Bansal et al. 2022)	Kajian ini mengkaji lima algoritma iaitu Jiran K-Terdekat, Regresi Linear, Menyokong Regresi Vektor, dan Memori Jangka Pendek Panjang untuk meramalkan nilai saham 12 perniagaan pasaran saham India yang terkenal. Laporan ini juga menyoroti beberapa kaedah yang lebih berkesan dan boleh dipercayai untuk meramalkan perubahan pasaran saham. Algoritma Pembelajaran Mendalam (DL) melepasi semua algoritma lain untuk ramalan harga saham atau siri masa dan memberikan penemuan dengan ketepatan yang tinggi, menurut kesimpulan kertas yang berdasarkan hasil analisis menyeluruh data yang telah dibentangkan.	Ramalan menggunakan ML.	KNN, LR, SVM, LSTM	Bursa Saham Bombay
7	Stock closing price prediction using Machine Learning Techniques (Vijh et al. 2020)	Rangkaian neural buatan dan teknik Hutan Rawak telah digunakan untuk meramalkan harga tutup saham bagi syarikat dalam pelbagai industri pada hari berikutnya. Data kewangan digunakan untuk menjana pembolehubah tambahan yang digunakan sebagai input model. Maklumat itu dikumpulkan dari yahoo finance dan terdiri daripada lima syarikat yang berbeza. Model dianalisis menggunakan RMSE dan MAPE yang merupakan dua metrik strategi biasa. Kedua-dua penunjuk ini mempunyai nilai rendah yang menunjukkan bahawa model berkesan dalam meramalkan harga tutup saham. Berdasarkan kajian nilai RMSE, MAPE dan MBE, ANN meramalkan harga saham lebih tepat daripada RF.	Ralat peratusan mutlak min (MAPE), Ralat kuasa dua root (RMSE), Ralat Bias Purata (MBE)	ANN, RF	Yahoo Finance
8	A new prediction NN framework design for individual stock based on the industry environment (Zhu et al. 2022)	Bagi penguraian mod variasi hibrid dan model unit berulang berpagar bertindan (VMD-StackedGRU), kajian ini membina modul ramalan dan modul persekitaran. Maklumat saham individu dimasukkan ke dalam modul ramalan, dan maklumat industri dimasukkan ke dalam modul persekitaran. Berdasarkan modul ramalan dan	(VMD-StackedGRU)	Model VMD StackedGRU (GRUI-GRUE)	Bank domestik Cina

bersambung…

…sambungan

		modul persekitaran, model VMD StackedGRU (GRUI-GRUE) yang dicadangkan dalam kertas kerja ini didapati mempunyai prestasi ramalan yang sangat tepat, dengan modul alam sekitar terbukti memainkan peranan penting dalam mengawal dan menyekat modul ramalan. Kajian ini menangani jurang dengan menawarkan cara untuk menjamin unjuran harga saham kewangan yang lebih baik berdasarkan keadaan industri.			
9	Machine learning techniques and data for stock market forecasting: A literature review (Kumbure et al. 2022)	Perhatian kajian ini adalah pada pasaran saham yang dikaji dalam kesusasteraan serta pelbagai faktor yang digunakan sebagai input dalam pendekatan ML yang digunakan untuk meramal pasaran ini. Kajian ini melihat kepada 138 penerbitan jurnal dari tahun 2000 hingga 2019. Ia memberi tumpuan kepada penyelidikan yang dilakukan antara tahun 2000 dan 2019 dan hanya menulis mengenai kajian yang menggunakan ML untuk meramalkan harga saham. Kajian ini mencapai kesimpulan dari penyiasatan mereka bahawa teknik LSTM telah mengatasi teknik pembelajaran mendalam yang lain dari segi keberkesanan dan ketepatan ramalan.	Kajian Literasi	ANN, SVM, Teori Kabur, DL, Pemilihan Ciri	Jurnal Ramalan Pasaran Saham dari tahun 2000 – 2019.
10	An improved deep learning model for predicting stock market price time series (Liu & Long 2020)	Pendekatan unik untuk meramalkan harga penutupan saham dikemukakan dalam kajian ini. Kedua-dua prapemprosesan dan prepemprosesan data menggunakan model mesin pembelajaran ekstrem yang lebih bagus (ORELM), yang berdasarkan transformasi gelombang empirikal (EWT). Komponen asas bingkai campuran, peramal rangkaian pembelajaran mendalam yang dibina di atas rangkaian memori jangka pendek panjang (LSTM), dioptimumkan bersama oleh kaedah keciciran dan algoritma pengoptimuman kawanan zarah (PSO). Setiap algoritma dalam seni bina	Ramalan menggunakan pembelajaran mendalam	Transformasi Gelombang Empirikal (EWT), Mesin Pembelajaran Ekstrem Yang Lebih Mantap (ORELM), Memori Jangka Pendek Panjang (LSTM), Pengoptimuman Kawanan Zarah (PSO)	Indeks Standard and Poor's 500 (S&P 500) dari 2010 hingga 2013. China Min Sheng Bank (CMSB) dari 2013 hingga 2016. Purata Perindustrian Dow Jones (DJI)

bersambung…

…sambungan

		hibrid boleh menggunakan semua keupayaannya untuk meningkatkan ketepatan ramalan. Keputusan ujian mereka menunjukkan bahawa seni bina hibrid yang disediakan dalam kajian ini boleh digunakan untuk memantau pasaran saham dan mempunyai ketepatan ramalan terbaik.			dari 2014 hingga 2017.
11	Stock market prediction based on statistical data using machine learning algorithms (Akhtar et al. 2022)	Kajian ini memberi tumpuan kepada mencari versi terbaru untuk meramalkan kos permintaan inventori. Ia memperkenalkan dan menilai alat yang lebih berguna untuk meramalkan pergerakan saham yang kurang sensitif. Malah menyediakan data mentah set data adalah sesuatu yang boleh ditumpukan oleh komposisi kajian ini. Reka bentuk kajian ini mungkin menumpukan perhatian kepada penyediaan data mentah set data. Ia dapat memantau bagaimana kayu sewenang-wenangnya digunakan, membantu mesin vektor pada set data, dan melihat output yang dihasilkannya. Komposisi ini juga menawarkan versi literasi sistem untuk meramalkan kehidupan saham dalam permintaan yang kompetitif.	Ramalan menggunakan ML	Pengelasan, RF, SVM	Kaggle
12	Social Media and Stock Market Prediction: A Big Data Approach (Awan et al. 2021)	Kajian ini menggunakan aplikasi PySpark, alat yang boleh diskalakan, cepat, mudah diintegrasikan dengan alat lain, dan dapat melakukan tugas lebih baik daripada model konvensional bagi mencipta beberapa model ML dengan Spark MLlib. Sepuluh firma saham terkemuka dikaji menggunakan model MLlib untuk regresi linear, regresi linear umum, hutan rawak, dan pokok keputusan. Sepuluh simbol firma saham yang digunakan sebagai data adalah AAPL, Yahoo, AMZN, Gold, FB, IBM, DELL, GOOG dan NFLX. Data kajian termasuk harga saham sejarah. Kedua-dua model pengelasan regresi logistik dan model Naive Bayes digunakan	Ramalan menggunakan ML	Regrasi Linear(LR), Pokok Keputusan(DT), Regresi linear umum(GLR), RF, Naive Bayes(NB), Regresi logistik	Twitter, Yahoo Finance, Kaggle

bersambung…

…sambungan

		dalam penyiasatan ini.			
13	Predicting Stock Prices Using Machine Learning Techniques (Karthikeyan et al. 2021)	Kajian ini meramalkan saham menggunakan model Purata Bergerak, dan hasil yang diperolehnya mempunyai sedikit kesilapan. Ini menunjukkan bahawa anggaran kajian ini telah dijalankan adalah munasabah dalam menganalisis dan meramalkan maklumat tersebut. Kajian ini menggunakan Model Purata Bergerak untuk menjalankan eksperimen, dan ia memberikan hasil terbaik sambil mempunyai Nilai RMSE terendah. Nilai RMSE rendah model menunjukkan betapa tepatnya bacaan. Ia menunjukkan bahawa Purata Pergerakan berprestasi lebih baik jika dibandingkan dengan teknik Regresi. Kajian ini menunjukkan bahawa teknik ramalan siri masa, seperti Purata Bergerak, melaksanakan dengan tepat dan memberikan lebih ketepatan jika dibandingkan dengan Regresi Linear untuk ramalan harga saham seperti yang ditunjukkan oleh nombor RMSE (Ralat Punca Min Square) dan contoh graf.	Ramalan menggunakan ML	LR, Model Purata Bergerak, KNN	Kaggle
14	Machine Learning for Predicting Stock Market Movement using News Headlines (Liu et al. 2020)	Kajian ini akan membandingkan teknik Pembelajaran Mendalam (DL) dan ML (ML) untuk meramalkan pergerakan Indeks Dow berdasarkan tajuk berita. Dalam kajian ini, dua jenis teknik yang berbeza akan digunakan berbanding dengan semua pendekatan lain. Bagi menentukan sejauh mana model ini menangani situasi yang tidak pernah berlaku sebelum ini yang disebabkan oleh pandemik. Kajian ini akan mengambil kira dua set data iaitu satu yang mengandungi tarikh sebelum pandemik COVID-19 dan satu lagi yang merangkumi tarikh sepanjang pandemik sehingga tarikh 17 Jun 2020.	Ramalan menggunakan ML dan pembelajaran mendalam.	TF-IDF, GloVe, Convolutional Neural Rangkaian (CNN), LR, SVM, RF	Dow Jones Industrial Average (or Dow Index), Reddit

2.7 KESIMPULAN

Dalam bab dua ini telah menerangkan lebih lanjut mengenai kesusasteraan kajian ini dan juga lebih lanjut mengenai kajian yang telah dilakukan oleh penyelidik terdahulu yang berkaitan dengan kajian mengenai ramalan pasaran saham menggunakan ML. Bab ini juga telah menerangkan mengenai pembelajaran yang terselia dan pembelajaran yang tidak terselia. Dalam bab ini juga memberi penjelasan mengenai kaedah pengelasan yang akan digunakan dalam kajian ini. Selepas itu, bab ini juga menunjukkan jenis penilaian yang akan digunakan bagi mengukur hasil ramalan bagi setiap model.

Berdasarkan daripada bacaan kajian terdahulu, kajian daripada (Ghani et al. 2019) ini merupakan kajian yang paling hampir dengan kajian yang dicadangkan tetapi terdapat beberapa perbezaan. Antara perbezaan kajian (Ghani et al. 2019) dengan kajian yang dicadangkan adalah ia hanya menggunakan data tujuh bulan sahaja iaitu daripada Januari hingga Julai 2019 dan data yang digunakan adalah data tidak terkini berbanding kajian yang dicadangkan. Selain itu, penyelidikan daripada (Ghani et al. 2019) juga membuat ramalan menggunakan kaedah pembelajaran mesin menggunakan model khusus iaitu LR, 3MMA, ES dan Ramalan Siri Masa. Berdasarkan daripada (Ghani et al. 2019) melaporkan didapati bahawa hasil daripada penggunaan algoritma ES telah dapat menjangkakan hala tuju pasaran saham dengan tepat untuk bulan yang akan datang dengan baik berbanding LR dan 3MMA. Seterusnya, berlandaskan kajian yang dicadangkan ini akan memberi tumpuan kepada menggunakan tiga jenis model yang berbeza iaitu ANN, SVM, dan RF. Tiga jenis model ini dipilih bagi kajian ini adalah kerana berdasarkan bacaan kajian terdahulu didapati tiga jenis model ini kerap digunakan untuk meramal pasaran saham dan memberi hasil ramalan yang bagus. Ketiga-tiga model yang digunakan dalam kajian dicadangkan ini kemudiannya akan dianalisis dan dibandingkan antara satu sama lain berdasarkan beberapa penilaian untuk mengenal pasti model yang memberi hasil yang terbaik.

Kaedah yang digunakan bagi kajian ini adalah kaedah pembelajaran mesin berbanding menggunakan kaedah siri masa kerana terdapat lebih banyak model pembelajaran mesin yang dapat digunakan untuk meramal pasaran saham. Selain itu, hasil daripada pembacaan kajian terdahulu juga menunjukkan penggunaan kaedah

pembelajaran mesin mempunyai banyak cara penggunaan model untuk membuat ramalan pasaran saham itu lebih tepat. Seterusnya, tiga algoritma peramal akan dilarikan dan dibandingkan untuk mencari model ramalan yang terbaik. Di dalam kajian ini memilih menggunakan perisian RapidMiner berbanding daripada perisian lain seperti python dan Weka kerana perisian ini mempunyai antara muka unit grafik yang lebih mudah untuk difahami. Selain itu ia juga menawarkan penggunaan model algoritma secara automasi.

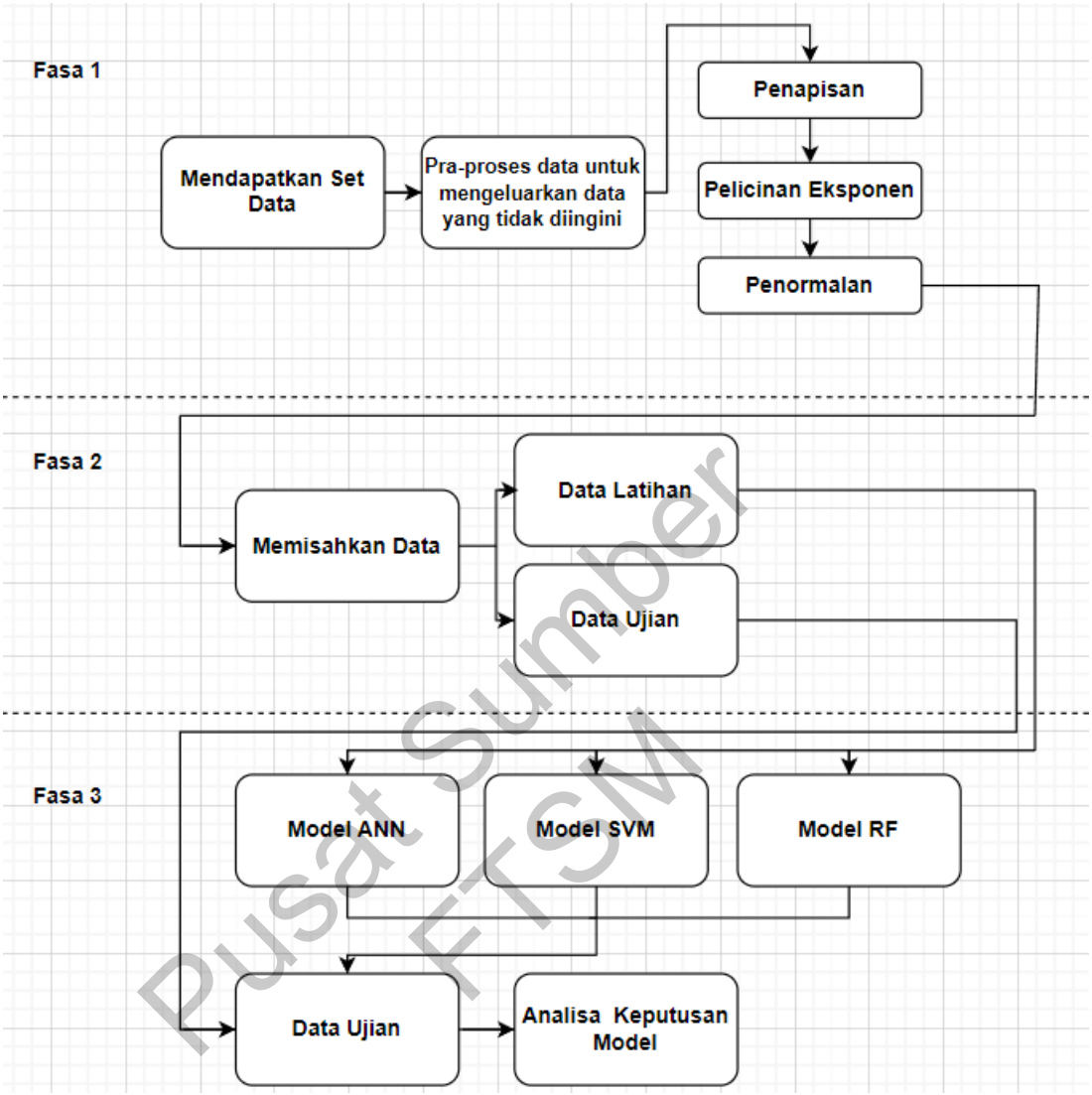
BAB III

KAEDAH KAJIAN

3.1 PENGENALAN

Bab tiga ini mempunyai metodologi kajian yang akan dibincangkan secara lebih terperinci. Di dalam bab ini juga akan menerangkan secara terperinci mengenai pelan reka bentuk kajian yang akan dijalankan. Prosedur penyediaan akan membentangkan beberapa langkah bagaimana objektif kajian yang dilaksanakan akan dibincangkan di dalam pendekatan kajian ini. Selepas itu, bab ini akan menerangkan tentang ilustrasi set data dan butirannya. Berikutan itu, akan ada memperkenalkan mengenai proses cadangan yang akan digunakan untuk kajian ini. Terdapat tiga fasa utama yang akan diterangkan dan akan dipecahkan lagi kepada beberapa bahagian kecil bagi menjelaskan secara khusus untuk setiap fasa. Akhir sekali akan memberi ringkasan atas metodologi kajian yang digunakan bagi kajian ini.

3.2 TEKNIK YANG DICADANGKAN



Rajah 3.1 Rajah Sistem Jalan Kerja ML

Adaptasi daripada: (Rauschert et al. 2020)

ML adalah berdasarkan idea bahawa pengetahuan boleh diekstrak dari data (Kubat 2017). Aliran kerja umum bagi strategi berasaskan pembelajaran yang diselia yang digunakan untuk ramalan pasaran saham ditunjukkan dalam Rajah 3.1. Pemilihan data siri masa adalah langkah pertama dalam proses. Berdasarkan kajian ini, pada permulaannya data harga saham dan maklumat yang berkaitan dengan harga saham seperti berita kewangan yang dikaitkan dengan firma pasaran saham di dalam data pasaran saham dari tempoh masa tertentu akan dipilih. Jika tugas melibatkan cabaran pengelasan, kelas sasaran sama ada telah diketahui atau akan ditentukan. Data yang

berkaitan mesti terlebih dahulu melalui pra-pemprosesan, yang bermula dengan membersihkan set data mentah dan mengasingkan sebarang data yang tidak lengkap atau jelas dan tidak perlu. Data siri masa asas, seperti data harga tutup, kemudiannya boleh digunakan untuk mencipta penunjuk teknikal.

Bagi menggunakan pengurangan penskalaan dan dimensi, data akan diproses terlebih dahulu untuk mendapatkan pemboleh ubah penting dan untuk membersihkan data yang tidak perlu. Data pra-proses sering digunakan untuk membina model ramalan yang tepat (Chen et al. 2019). Berdasarkan Rajah. 3.1, terdapat data pra-pemprosesan pada rajah tersebut, tetapi ia perlu diperhatikan bahawa langkah ini adalah pilihan kerana ia biasanya bergantung kepada domain yang telah dipilih. Data input biasanya dipisahkan ke dalam data latihan, pengesahan dan set ujian untuk tujuan ini. Akhir sekali, ramalan dilakukan sama ada dalam pengelasan menggunakan model pengelasan latihan atau regresi menggunakan model regresi yang telah dilatih.

3.3 SET DATA YANG DIGUNAKAN

Komuniti yang didedikasikan untuk pemodelan ramalan dan analisis maklumat web dipanggil Kaggle. Ia juga mengandungi set data dari bidang lain yang terlalu banyak oleh pelombong rekod. Cara yang paling terkenal untuk ramalan dan menggambarkan statistik menurut saintis disediakan oleh fakta berwarna-warni. Ia membolehkan pengguna saintis data menggunakan set data mereka untuk mencipta model dan bekerjasama dengan pakar pengetahuan rekod kreatif untuk menyelesaikan masalah pengetahuan statistik sebenar yang berwarna-warni (Grigoryan 2016). Set data siri masa yang digunakan untuk kajian yang dicadangkan ini dimuat turun daripada laman web Kaggle daripada (Bajaj 2023) dan data yang dipilih berlabel AAPL. AAPL adalah simbol saham bagi sebuah syarikat bernama Apple Inc. Set data siri masa ini merupakan set data pasaran saham yang dikumpulkan daripada Yahoo Finance. Kerana saiz dan keselamatannya yang besar, yang membolehkan perkongsian awam atau swasta bergantung kepada keadaan, kajian ini menggunakan set data Kaggle. Walau bagaimanapun, set data ini boleh didapati dalam apa yang sering di rujuk sebagai set data mentah. Pengumpulan maklumat permintaan saham mengenai firma tertentu membentuk set data.

Dokumen set data yang akan digunakan ini dinamakan sempena setiap simbol saham dan mengandungi data saham sejarah untuk semua saham dalam Yahoo Finance API. Set data saham sejarah telah dipilih daripada bank data terkenal Kaggle untuk tujuan kajian ini. Set data ini bersaiz 139 MB, terdiri daripada data harga saham sejarah untuk pelbagai saham, termasuk indeks besar seperti S&P 500 dan NASDAQ serta syarikat yang lebih kecil dan kurang terkenal. Atribut yang termasuk dalam set data adalah tarikh, terbuka, tinggi, rendah, tutup, volum, dividen dan pembahagian saham. Set data siri masa yang digunakan di dalam kajian ini adalah data siri masa di mana data tersebut telah direkodkan sebanyak lima kali sehari sepanjang tahun daripada hari isnin sehingga hari jumaat iaitu pada hari bekerja dan tidak pada hari hujung minggu seperti pada hari sabtu dan hari ahad. Set data mentah ini mempunyai rekod lebih daripada 5 tahun dan set data ini direkod sehingga September 2023. Analisis ini hanya akan menggunakan saham sebuah syarikat. Jadual 3.1 di bawah menunjukkan contoh set data sejarah dengan ciri data berstruktur.

Jadual 3.1 Contoh wakil set data sejarah bagi syarikat AAPL

Tarikh	Buka	Tinggi	Rendah	Tutup	Volum	Dividen	Pembahagian Saham
Dec 12, 1980	0.099583767 3544883	0.100016716 227917	0.099583767 3544883	0.099583767 3544883	469033 600	0	0
Dec 15, 1980	0.094821396 6962838	0.094821396 6962838	0.094388447 701931	0.094388447 701931	175884 800	0	0
Dec 16, 1980	0.087774979 40939089	0.087774979 40939089	0.087342619 89593506	0.087342619 89593506	105728 000	0	0
Dec 17, 1980	0.089504398 40555191	0.089936763 54420131	0.089504398 40555191	0.089504398 40555191	864416 00	0	0
Dec 18, 1980	0.092099338 76991272	0.092531698 0768195	0.092099338 76991272	0.092099338 76991272	734496 00	0	0
Dec 19, 1980	0.097720034 42049026	0.098152399 48182175	0.097720034 42049026	0.097720034 42049026	486304 00	0	0
Dec 22, 1980	0.102476067 84105301	0.102909206 66057306	0.102476067 84105301	0.102476067 84105301	373632 00	0	0
Dec 23, 1980	0.106800436 97357178	0.107232802 11752925	0.106800436 97357178	0.106800436 97357178	469504 00	0	0

Jadual 3.2 di bawah menunjukkan perihalan yang berkaitan dengan atribut dalam set data.

Jadual 3.2 Penerangan Mengenai Atribut Set Data

Atribut	Keterangan
Tarikh	Tarikh hari perdagangan..
Buka	Harga pembukaan saham untuk hari itu.
Tutup	Harga penutupan saham untuk hari itu.
Tinggi	Harga tertinggi yang dicapai oleh saham pada hari itu.
Rendah	Harga terendah yang dicapai oleh saham pada hari itu.
Volum	Jumlah saham yang didagangkan pada hari itu.
Dividen	Sebarang dividen dibayar pada hari itu.
Pembahagian Saham	Sebarang pembahagian saham yang berlaku pada hari itu.

3.4 FASA 1

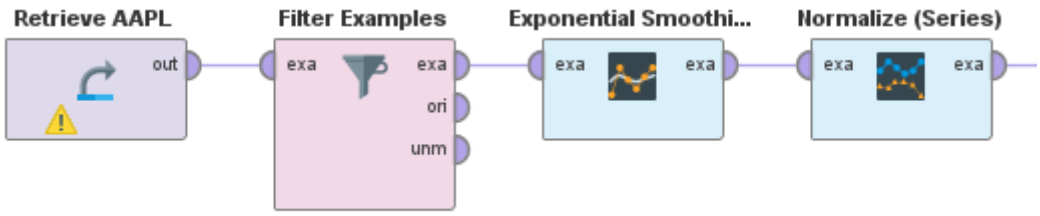
3.4.1 Pra-Pemprosesan

Pra-pemprosesan data terutamanya untuk projek yang melibatkan sejumlah data yang besar. Ini adalah langkah penting yang melibatkan transformasi data rawak dan mentah dengan cara yang meningkatkan kualitinya dengan mengeluarkan atau membersihkan titik yang tidak diingini, menyeragamkannya untuk kegunaan biasa dan membolehkannya menghasilkan data berharga. Kualiti data mempunyai kesan yang lebih besar daripada kuantiti data ketika menghasilkan hasil yang sangat baik. Pra-pemprosesan data juga melibatkan pembersihan data, pengasingan data atau organisasi, penskalaan data, penormalan data dan penyeragaman, serta pengekodan kategori data. Kaedah yang digunakan untuk set data yang mempunyai nilai yang hilang adalah dengan strategi min atribut untuk mengisi nilai yang hilang. Ia berfungsi dengan menggantikan nilai yang hilang untuk atribut tertentu dengan nilai purata atribut tersebut. Penormalan data digunakan untuk fasa pra-pemprosesan kerana ia melibatkan perubahan nilai data ke dalam julat tertentu seperti antara 0 dan 1 atau -l dan 1 supaya proses berfungsi. Bagi teknik perlombongan, strategi ini amat berguna. Ciri-ciri data diskalakan menggunakan penormalan yang juga boleh digunakan untuk mempercepatkan proses pembelajaran (Suad & Wesam 2017).

Terdapat sejumlah yang besar harga saham dari masa terdahulu mengikut rekod yang diperoleh daripada Kaggle. Pembinaan model ramalan bermula dengan pra-pemprosesan. Berdasarkan banyak data yang di muat turun daripada Kaggle. Data yang dipilih untuk kajian ini ialah AAPL. Data daripada AAPL terdiri daripada tahun 1980 sehingga tarikh terkini iaitu pada tahun 2023. Seperti yang ditunjukkan dalam Jadual 3.2, set data terdiri daripada 10759 sampel dan 8 atribut. Rajah 3.2 menunjukkan statistik set data sebelum pra-pemprosesan. Berdasarkan statistik tersebut data ini tidak mempunyai sebarang sampel data yang hilang daripada setiap atribut.

Name	Type	Missing	Statistics			Filter (8 / 8 attributes):
Date	Date-time	0	Earliest date Dec 12, 1980	Latest date Sep 21, 2023	Duration 15622 days	
Open	Real	0	Min 0.038	Max 195.975	Average 18.284	
High	Real	0	Min 0.038	Max 197.963	Average 18.487	
Low	Real	0	Min 0.038	Max 195.017	Average 18.087	
Close	Real	0	Min 0.038	Max 196.185	Average 18.295	
Volume	Integer	0	Min 0	Max 7421640800	Average 323196109.245	
Dividends	Real	0	Min 0	Max 0.240	Average 0.001	
Stock Splits	Real	0	Min 0	Max 7	Average 0.002	

Rajah 3.2 Statistik Set Data Sebelum Pra-pemprosesan



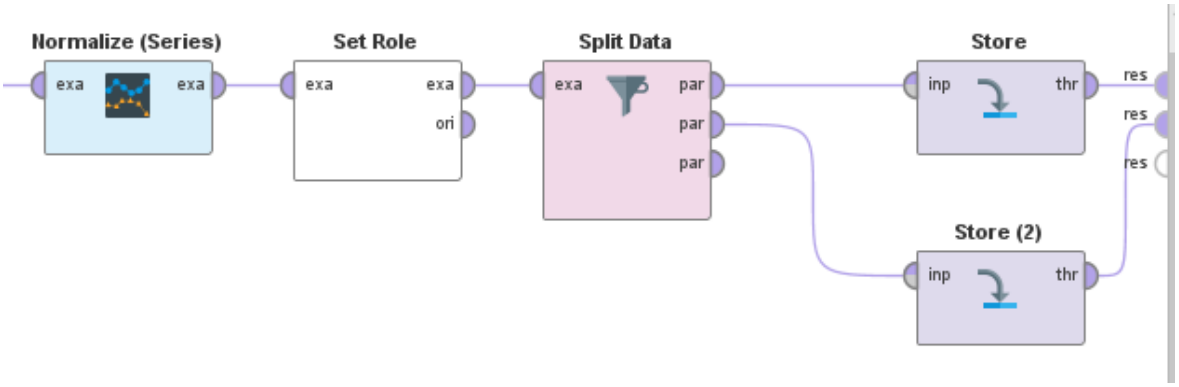
Rajah 3.3 Pra-pemprosesan

Pertama sekali, set data mentah akan dibawa ke tetingkap "Proses" dalam RapidMiner selepas diimport ke repositori. Data mentah AAPL dipaparkan sebagai

"Operator" dalam Rajah 3.3. Pengendali data mentah akan dihubungkan dengan pengendali penapis, ini membolehkan data mentah ditapis bermula pada 1 Januari 2018 dan berterusan hingga 21 September 2023 iaitu data yang paling baharu. Khususnya, matlamat eksperimen ini adalah untuk menjangkakan pasaran saham menggunakan tiga algoritma yang dipilih, membandingkan output model dan menentukan mana yang paling tepat. Inilah sebab mengapa data dari 2018 telah ditapis. Berikutan itu, operator pelicinan eksponen untuk set data siri masa akan dilampirkan pada operator pengendali penapis. Berdasarkan daripada (Anon. 2023), salah satu teknik ramalan yang amat berguna untuk ramalan jangka pendek ialah dengan pelicinan eksponen.

Apabila data lebih jauh, kaedah ini akan memberikan berat yang lebih kecil secara eksponen dan memberikan berat yang lebih besar kepada data yang lebih baharu. Model ini membuat andaian bahawa data masa lalu dan masa hadapan ini akan sedikit menyerupai satu sama lain. Pelicinan eksponen hanya mengiktiraf jumlah permintaan atau nilai purata yang mewakili turun naik data dari masa ke semasa sebagai corak dari data permintaan masa lalu. Terdapat sebahagian daripada kelancaran eksponen yang dikenali sebagai Alpha. Pilihan ini menetapkan berapa banyak pengaruh nilai sebelumnya atau seberapa kuat kesan melicinkan. Jika nilai alpha adalah tinggi, ia akan menyebabkan nilai data tidak dilicinkan. Sebagai contoh, jika nilai alpha adalah 1, data tidak akan dilicinkan. Dalam percubaan ini, nilai alfa akan diletakkan sebagai 0.5, kerana ia memberi sedikit kesan melicinkan sambil mengekalkan corak yang dapat dilihat.

Berikutan itu, operator menormalkan untuk data siri masa akan dilampirkan kepada operator pelicinan eksponen. Ketepatan dan keberkesanan model ramalan boleh dipengaruhi oleh varians dalam skala, julat, dan taburan yang terdapat dalam data siri masa. Penormalan data melalui penggunaan operator penormalan dapat membantu dalam membuang data luaran daripada siri data dan menyusun semula nilai siri ke dalam julat yang sama sambil mengekalkan kestabilan ciri data. Teknik penormalan data misalnya dapat mengurangkan bilangan nilai yang sama dalam set data. Oleh itu, penormalan juga dapat mengosongkan ruang simpanan ingatan tambahan untuk data yang akan datang dengan memadam data saiz besar yang tidak perlu.



Rajah 3.4 Pecahan Data Ujian Dan Latihan

Rajah 3.4 menunjukkan bahawa menormalkan operator disambungkan dengan operator peranan set. Mengubah suai peranan satu atau lebih ciri boleh dilakukan dengan operator peranan set. Cara di mana atribut ini dikendalikan oleh operator lain dijelaskan dalam peranannya. Peranan boleh dikategorikan sebagai istimewa atau biasa, tetapi pada sebelum pra pemprosesan, data hanya mempunyai atribut peranan biasa sahaja. Atribut dengan nama Tutup akan mempunyai fungsi set label yang ditetapkan untuk eksperimen ini dan atribut tarikh akan mempunyai peranan set ID. Maksud peranan label adalah peranan unik yang akan mencipta atribut untuk berfungsi sebagai atribut sasaran untuk operasi pembelajaran dan peranan id adalah peranan khas yang bertindak sebagai pengenal pasti untuk set data. Selepas itu, operator peranan set disambungkan kepada operator data berpecah dan kemudiannya berpecah kepada dua operator simpanan. Operator data berpecah berfungsi sepagai pembahagi data dan data ini akan dibahagi kepada dua jenis bahagian data. Data bahagian pertama adalah data latihan yang mempunyai 70% data dan data bahagian kedua mempunyai 30% data yang dinamakan data ujian. Operator simpan hanya berfungsi sebagai penyimpan data ke dalam fail di dalam komputer.

Row No.	Date	Close	Open	High	Low	Volume	Dividends	Stock Splits
1	Dec 7, 2021	1.255	167.390	169.865	166.658	120405400	0	0
2	Dec 8, 2021	1.330	170.410	174.201	168.994	116998900	0	0
3	Dec 9, 2021	1.362	173.162	174.984	172.182	108923700	0	0
4	Dec 10, 2021	1.428	173.459	177.835	172.944	115402700	0	0
5	Dec 13, 2021	1.423	179.310	180.310	173.776	153237000	0	0
6	Dec 14, 2021	1.407	173.499	175.964	170.489	139380400	0	0
7	Dec 15, 2021	1.448	173.360	177.706	170.588	131063300	0	0
8	Dec 16, 2021	1.398	177.488	179.330	169.044	150185800	0	0
9	Dec 17, 2021	1.362	168.232	171.736	167.994	195432700	0	0
10	Dec 20, 2021	1.330	166.598	168.875	165.786	107499100	0	0
11	Dec 21, 2021	1.347	169.845	171.469	167.430	91185900	0	0
12	Dec 22, 2021	1.382	171.311	174.102	170.430	92135300	0	0
13	Dec 23, 2021	1.405	174.093	175.083	173.518	68356600	0	0
14	Dec 27, 2021	1.458	175.320	178.617	175.300	74919600	0	0
15	Dec 28, 2021	1.474	178.359	179.518	176.746	79144300	0	0
16	Dec 29, 2021	1.483	177.538	178.825	176.360	62348900	0	0
17	Dec 30, 2021	1.475	177.676	178.765	176.310	59773000	0	0
18	Dec 31, 2021	1.465	176.310	177.439	175.488	64062300	0	0
19	Jan 3, 2022	1.505	176.053	181.052	175.934	104487900	0	0
20	Jan 4, 2022	1.501	180.805	181.112	177.330	99310400	0	0
21	Jan 5, 2022	1.452	177.815	178.369	172.895	94537600	0	0
22	Jan 6, 2022	1.397	170.974	173.548	169.925	96904000	0	0

Rajah 3.5 Contoh Data Selepas Pra-Pemprosesan

Rajah 3.5 menunjukkan contoh data mentah yang sudah di pra-pemprosesan. Ia menunjukkan bahawa data ini telah ditapis bermula daripada Januari 1, 2018 hingga ke tarikh paling baharu iaitu pada September 21, 2023. Data ini juga menunjukkan pada atribut tarikh dan tutup telah menjadi atribut istimewa. Data atribut tutup juga telah melalui proses kelancaran dan proses penormalan.

3.5 FASA 2

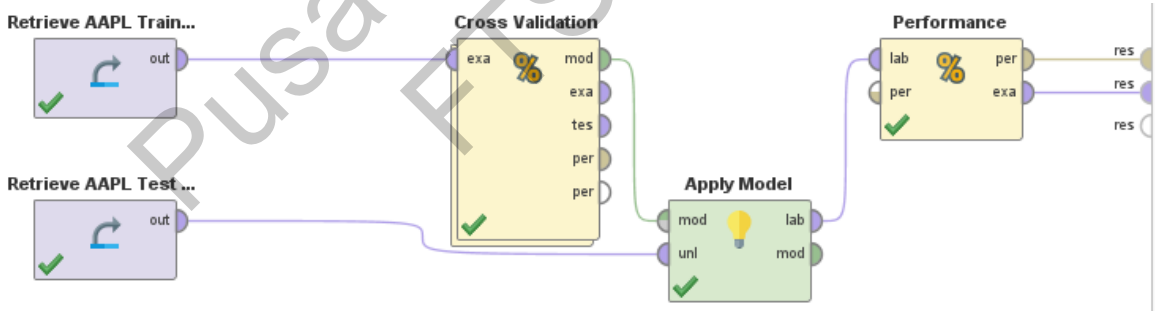
3.5.1 Pemisahan Data Ke Dalam Set Data Latihan Dan Set Data Ujian

Ia adalah perlu untuk membahagikan set data kepada set data latihan dan set data ujian sebagai langkah seterusnya berikutan langkah-langkah pra-pemprosesan. Set data akan dibahagikan kepada set latihan dan ujian di dalam kajian ini pada nisbah 70:30. Nisbah ini dipilih kerana ia akan berfungsi untuk keperluan kajian analisis ramalan yang melibatkan banyak data berubah-ubah. Bagi menguji model, hanya 30% daripada nilai set data atau hari dagangan digunakan dengan baki 70% digunakan untuk latihan. Oleh

kerana siri masa meramalkan sifat penyelidikan dan kerana saling bergantung antara data atau atribut, ia diperlukan untuk melatih model kebanyakannya daripada data sejarah. Proses pengesahan silang memerlukan sebahagian daripada data juga. Selain itu, baris rawak telah diberikan kepada set latihan dan set ujian untuk memastikan persampelan rawak yang boleh meningkatkan prestasi model semasa ujian. Operator Data Berpecah dalam aplikasi RapidMiner mengawal pengendalian bahagian set data ke dalam data latihan dan ujian. Oleh kerana ia mempengaruhi seberapa baik fungsi model, pembahagian data latihan dan ujian ke dalam perkadaran masing-masing adalah pembolehubah penting. Overfitting model boleh disebabkan oleh data latihan dan ujian yang sama, sedangkan kurang sesuai kerana boleh disebabkan oleh nilai set data yang sangat berbeza. Oleh itu, adalah penting untuk mempunyai nisbah set data latihan dan ujian yang sesuai untuk memperoleh gambaran sebenar prestasi model yang berbeza dan seterusnya akan menghasilkan penemuan yang boleh dipercayai.

3.6 FASA 3

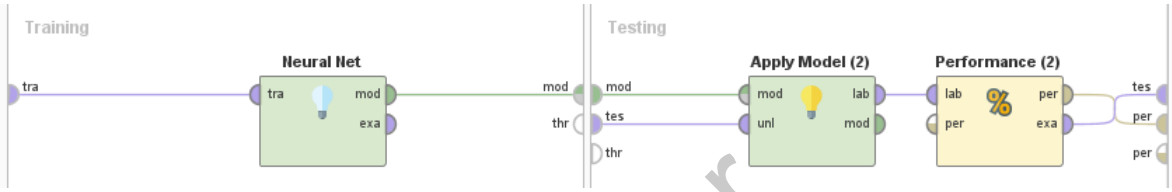
3.6.1 Proses Latihan ANN



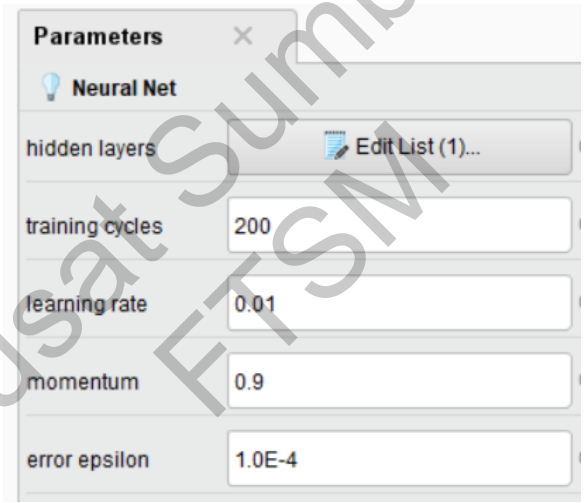
Rajah 3.6 Proses Latihan dan Ujian ANN

Berdasarkan Rajah 3.6 di atas menunjukkan proses latihan dan ujian bagi model ANN. Set data latihan yang telah dibahagi ketika proses pemisahan data di fasa kedua mengandungi 70% data dan data ini mengandungi sebanyak 1008 data. Rajah 3.6 menggambarkan bagaimana pengendali set data latihan akan dipautkan kepada moderator yang dikenali sebagai moderator pengesahan silang. Teknik statistik yang dipanggil pengesahan silang digunakan untuk menilai sejauh mana prestasi model ML terutamanya terhadap ketepatannya. Apabila model ramalan digunakan ia berfungsi sebagai perlindungan terhadap terlebih pemadanan terutamanya apabila jumlah data

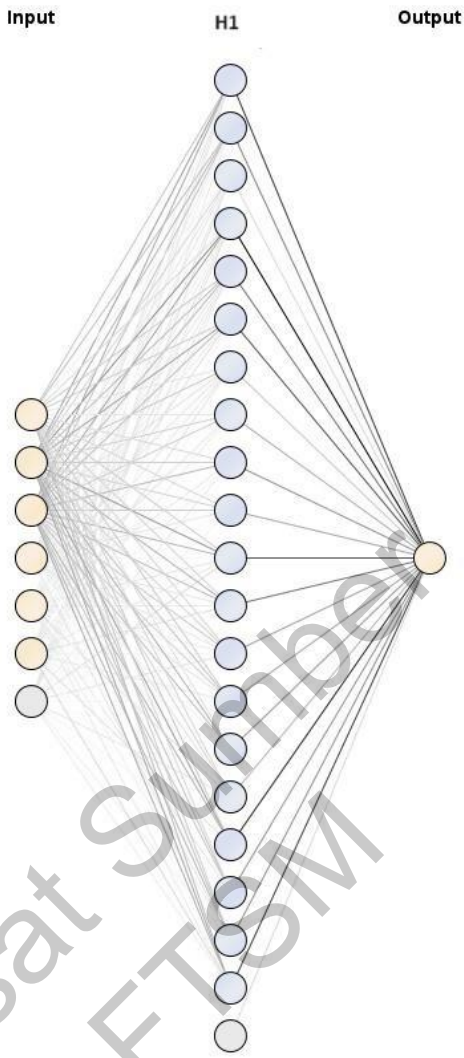
mungkin terhad (Anon. 2022). Operator yang melakukan pengesahan silang merupakan sebuah operator bersarang. Ia terdiri daripada dua subproses iaitu satu bahagian untuk ujian dan satu bahagian lagi untuk latihan. Model latihan dilakukan menggunakan subproses latihan. Subproses ujian kemudiannya akan menggunakan model yang telah dipelajari ketika di subproses latihan. Semasa fasa ujian, prestasi model tersebut akan dinilai.



Rajah 3.7 Sub-Proses Pengesahan Silang ANN



Rajah 3.8 Parameter ANN

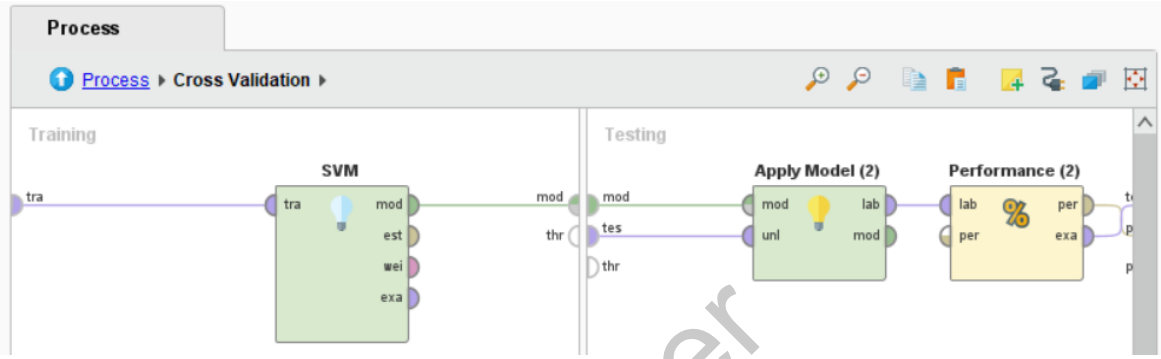


Rajah 3.9 Bentuk Rajah ANN

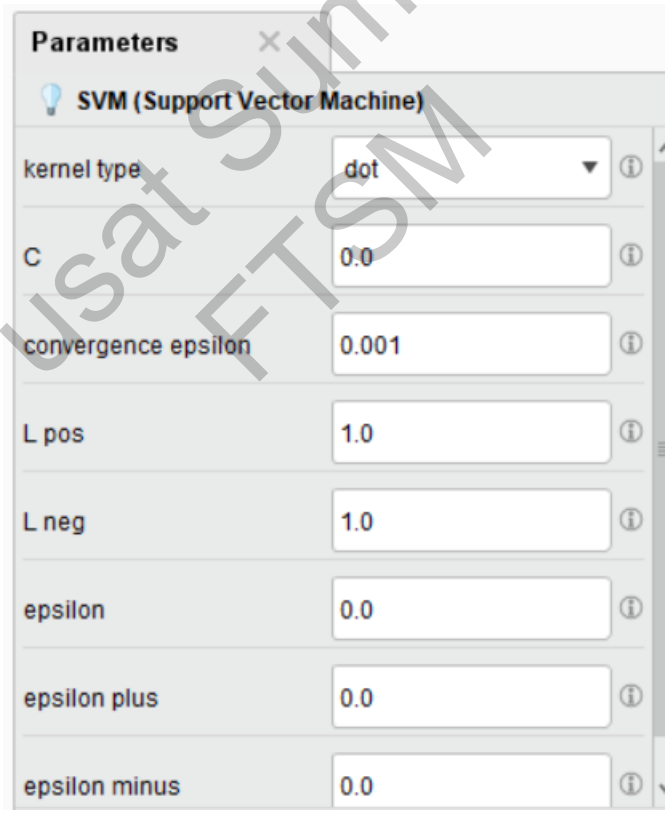
Subproses ujian dan latihan terkandung dalam operator pengesahan silang bersarang. Rajah 3.7 menggambarkan bahawa model ANN hadir dalam subproses latihan. Di dalam kajian ini, operator ANN menggunakan parameter sedia ada yang disediakan oleh RapidMiner seperti di Rajah 3.8. Bagi operator ANN ini akan mempunyai bahagian lapisan input, satu bahagian lapisan tersembunyi yang diberi nama H1 dan lapisan output seperti di Rajah 3.9. Berdasarkan Rajah 3.9 ini juga menunjukkan bentuk rajah ANN di dalam RapidMiner. Selepas itu, data terlatih akan disambungkan kepada operator memohon model dan operator prestasi untuk mempelajari set data ujian. Seterusnya, model latihan akan digunakan untuk set data ujian menggunakan operator memohon model sebaik sahaja data latihan telah dilatih melalui operator pengesahan silang. Akhir sekali, penilaian prestasi akan dipaparkan

dengan sambungan operator memohon model kepada operator prestasi. Hasil model ANN akan dipaparkan oleh operator prestasi.

3.6.2 Proses Latihan SVM



Rajah 3.10 Sub-Proses Pengesahan Silang SVM

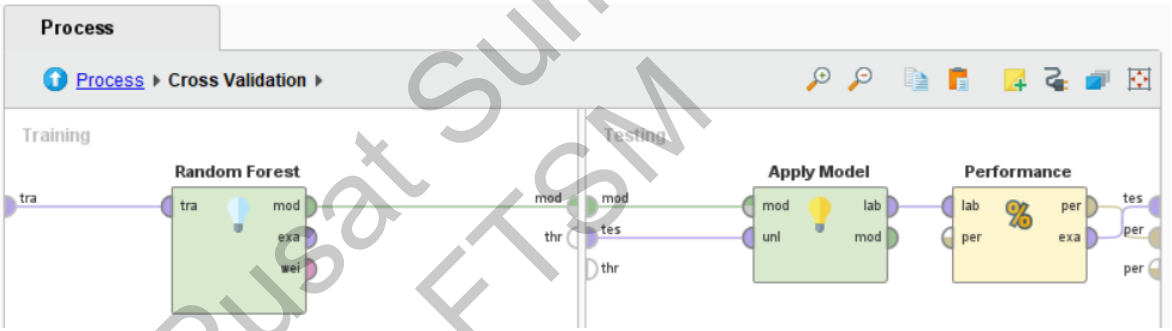


Rajah 3.11 Parameter SVM

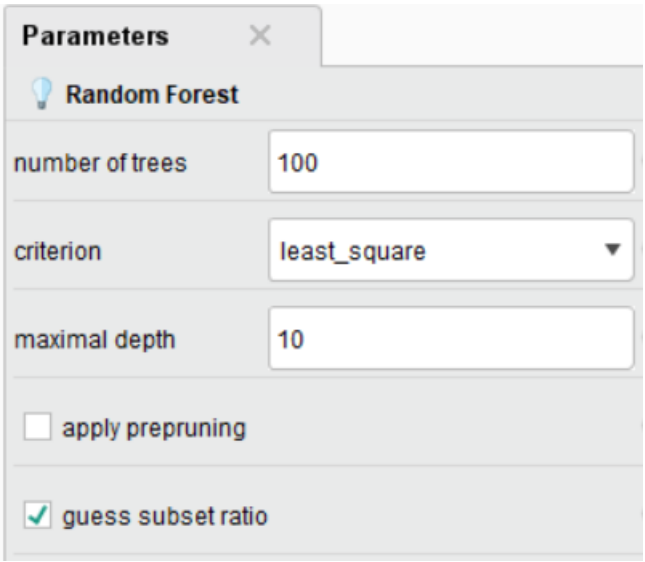
Berdasarkan Rajah 3.6 di atas juga menunjukkan proses latihan dan ujian bagi model SVM. Berdasarkan Rajah 3.10 menunjukkan bagaimana pengendali set data latihan akan dipautkan kepada moderator yang dikenali sebagai moderator pengesahan silang.

Teknik statistik yang dipanggil pengesahan silang digunakan untuk menilai sejauh mana prestasi model ML terutamanya terhadap ketepatannya. Operator SVM ini juga menggunakan parameter yang sedia ada diberikan oleh Rapidminer seperti di Rajah 3.11. Subproses ujian dan latihan terkandung dalam operator pengesahan silang bersarang. Rajah 3.10 menggambarkan bahawa model SVM hadir dalam subproses latihan. Selepas itu, data terlatih akan disambungkan kepada operator memohon model dan operator prestasi untuk mempelajari set data ujian. Seterusnya, model latihan akan digunakan untuk set data ujian menggunakan operator memohon model sebaik sahaja data latihan telah dilatih melalui operator pengesahan silang. Akhir sekali, penilaian prestasi akan dipaparkan dengan sambungan operator memohon model kepada operator prestasi. Hasil model SVM akan dipaparkan oleh operator prestasi.

3.6.3 Proses Latihan RF



Rajah 3.12 Sub-Proses Pengesahan Silang RF



Rajah 3.13 Parameter RF

Berdasarkan Rajah 3.6 di atas juga menunjukkan proses latihan dan ujian bagi model RF. Berdasarkan Rajah 3.12 menunjukkan bagaimana pengendali set data latihan akan dipautkan kepada moderator yang dikenali sebagai moderator pengesahan silang. Teknik statistik yang dipanggil pengesahan silang digunakan untuk menilai sejauh mana prestasi model ML terutamanya terhadap ketepatannya. Operator RF ini juga menggunakan parameter yang sedia ada diberikan oleh Rapidminer seperti di Rajah 3.13. Subproses ujian dan latihan terkandung dalam operator pengesahan silang bersarang. Rajah 3.12 menggambarkan bahawa model RF juga ada dalam subproses latihan. Selepas itu, data terlatih akan disambungkan kepada operator memohon model dan operator prestasi untuk mempelajari set data ujian. Seterusnya, model latihan akan digunakan untuk set data ujian menggunakan operator memohon model sebaik sahaja data latihan telah dilatih melalui operator pengesahan silang. Akhir sekali, penilaian prestasi akan dipaparkan dengan sambungan operator memohon model kepada operator prestasi. Hasil model RF akan dipaparkan oleh operator prestasi.

3.7 KEPERLUAN PENYELIDIKAN

Eksperimen yang dilakukan di dalam kajian ini menggunakan aplikasi RapidMiner dan menggunakan sistem operasi seperti yang di terangkan di bahagian 3.7.1 dan 3.7.2.

3.7.1 Aplikasi RapidMiner

Aplikasi RapidMiner merupakan sebuah aplikasi sumber terbuka, percuma dan alat perlombongan data. RapidMiner menyokong sistem pengendalian Macintosh, Linux dan Unix sebagai tambahan kepada sistem pengendalian Windows. Sebagai analisis data, ia ditawarkan sebagai program kendiri dan sebagai enjin perlombongan data untuk integrasi ke dalam produk pengguna sendiri. Pengguna boleh mencari dan memuat turun aplikasi perisian RapidMiner, bersama-sama dengan sambungan dan dokumentasinya. Pengguna boleh menggunakan aplikasi untuk beberapa tugas perlombongan data dan teks selepas memuat turun dan memasang versi yang betul. Setelah difahami, ia mempunyai antara muka pengguna yang logik dan bermaklumat dengan reka bentuk berasaskan grafik yang mudah digunakan.