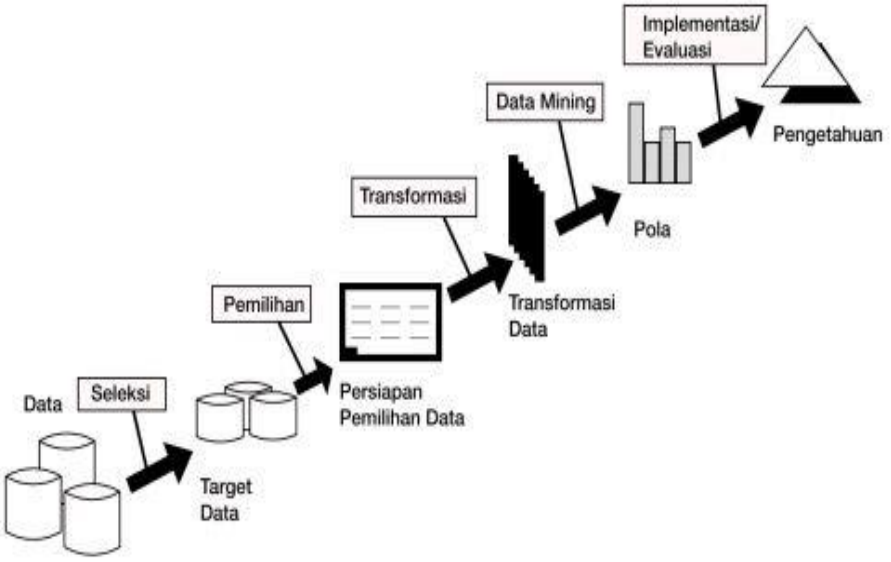
BAB II

LANDASAN TEORI

2.1 Knowledge Discovery in Database

Knowledge Discovery in Database (KDD) adalah proses mengungkap informasi tersembunyi yang sebelumnya tidak dikenal dalam database. Metode ini digunakan untuk menghasilkan pengetahuan yang dapat dimanfaatkan sebagai basis pengambilan keputusan. KDD melibatkan serangkaian langkah kompleks untuk menemukan pola asli dalam data yang dapat dipahami dan berguna bagi pengguna.[1]



Gambar 2.1 Tahapan Dalam Knowledge Discovery in Database

Sumber : Sigarmas.com

Salah satu tahapan dalam keseluruhan proses KDD adalah data mining. Seperti yang ditunjukan pada proses berikut :

1. *Selection* (Seleksi Data)

   Pada tahap ini, seleksi data dilakukan dengan memilih data yang relevan dari kumpulan data operasional. Data yang terpilih kemudian akan digunakan dalam proses berikutnya dalam *Knowledge Discovery in Database* (KDD).

2. *Preprocessing/Cleaning* (Pembersihan Data)

   Pada tahap ini, data yang duplikat atau inkonsisten dibuang, dan kesalahan seperti kesalahan ketik atau tipografi diperbaiki. Tujuannya adalah untuk memastikan bahwa data yang digunakan dalam proses mining bersih dan siap untuk dianalisis.

3. *Transformation* (Transformasi)

   Pada tahap ini, data yang masih belum terstruktur atau belum valid diubah menjadi format yang lebih jelas dan valid. Data yang sudah tertransformasi ini siap untuk diproses lebih lanjut dalam tahap mining.

4. Data Mining

   Pada tahap ini, algoritma atau metode pencarian pengetahuan diterapkan untuk menggali pola atau informasi tersembunyi dari data yang telah disiapkan sebelumnya. Teknik ini menggunakan berbagai metode statistik dan matematik.

5. *Interpretation/Evaluation* (Interpretasi/Evaluasi)

   Tahap terakhir ini melibatkan proses interpretasi hasil data mining, di mana pola yang ditemukan diubah menjadi informasi yang mudah dipahami dan dapat digunakan untuk pengambilan keputusan.

2.2 Data Mining

Data mining adalah Proses sistematis yang dikenal sebagai data mining bertujuan untuk menemukan pola, hubungan, atau informasi tersembunyi yang berharga dari kumpulan data yang sangat besar dan kompleks dengan menggunakan berbagai teknik seperti algoritma, statistik, dan kecerdasan buatan. Tujuan dari proses ini adalah untuk menghasilkan nilai tambah dari wawasan baru yang sebelumnya tidak dapat diidentifikasi secara manual. Nilai tambah ini dapat digunakan untuk mendorong pengembangan ilmu pengetahuan dan inovasi di berbagai bidang, seperti bisnis, kesehatan, pendidikan, teknologi, dan keuangan.

Data mining tidak hanya mencari pola-pola yang menarik dalam data, tetapi juga memberikan pemahaman baru yang dapat digunakan untuk mengatasi berbagai permasalahan, meningkatkan efisiensi operasional, dan menciptakan peluang baru dalam berbagai sektor industri. Hal ini menjadikan data mining sebagai salah satu teknologi kunci dalam era data besar (*big data*) yang semakin penting dalam mendukung transformasi digital dan pengambilan keputusan berbasis data.[2]

Ada beberapa definisi data mining yang dikenal dari berbagai sumber, diantaranya adalah:

1. Data mining yaitu kegiatan meliputi pengumpulan, pemakaian data historis dalam menemukan keteraturan, pola, atau hubungan dalam dataset berukuran besar. Output dari data mining ini bisa dipakai untuk memperbaiki pengambilan keputusan di masa depan.[3]

2. Data mining adalah kegiatan analisa data untuk mencari suatu pola tertentu, dengan jumlah data yang besar dan bertujuan untuk menghasilkan informasi yang dapat digunakan dan dikembangkan lebih lanjut.[4]

3. Data mining adalah proses menemukan informasi yang berguna dari gudang basis data yang besar. Data mining juga dapat diartikan sebagai pengekstrakan informasi dari sekumpulan data besar untuk membantu dalam pengambilan keputusan.[5]

Berdasarkan beberapa definisi tersebut, dapat disimpulkan bahwa Data mining adalah proses atau kegiatan yang melibatkan pengumpulan dan analisis data historis dalam jumlah besar untuk menemukan pola, keteraturan, atau hubungan yang tersembunyi. Tujuan utama dari data mining adalah untuk menghasilkan informasi yang berguna yang dapat digunakan untuk mendukung pengambilan keputusan yang lebih baik di masa depan, serta untuk mengembangkan pengetahuan lebih lanjut dari data tersebut.

2.3 Defenisi Keuntungan

Keuntungan adalah hasil dari selisih antara total pendapatan dan total biaya dalam suatu periode tertentu. usaha Amanda Brownies, keuntungan dapat dihitung dengan mengurangi total biaya operasional, yang meliputi biaya tetap (seperti sewa dan gaji karyawan) dan biaya variabel (biaya Transportasi), Keuntungan yang optimal sangat bergantung pada faktor-faktor yang memengaruhi penjualan, seperti harga, volume penjualan, dan biaya operasional.

Meningkatkan keuntungan memerlukan pemahaman yang mendalam mengenai variabel-variabel yang berpengaruh dan bagaimana faktor-faktor tersebut berinteraksi dalam proses bisnis. Keuntungan merupakan selisih antara penerimaan dan pengeluaran dalam suatu bisnis. Keuntungan dapat dihitung dengan mengurangi total biaya (termasuk biaya tetap dan variabel) dari total pendapatan yang diperoleh dari penjualan produk. Secara umum, keuntungan terbagi menjadi dua jenis keuntungan jangka pendek dan keuntungan jangka panjang.[6]

Berikut adalah definisi konsep keuntungan berdasarkan berbagai sumber :

1. Keuntungan adalah manfaat finansial yang direalisasikan ketika pendapatan bisnis melebihi pengeluaran, biaya lainnya, dan pajak yang terlibat dalam aktivitas bisnis tersebut.
2. Secara operasional, keuntungan adalah perbedaan antara pendapatan yang direalisasikan dari transaksi selama satu periode dengan biaya yang dikeluarkan untuk memperoleh pendapatan tersebut.
3. Keuntungan atau laba merupakan selisih lebih pendapatan atas beban sehubungan dengan kegiatan usaha. Apabila beban lebih besar dari pendapatan, selisihnya disebut rugi.
4. Keuntungan adalah hasil finansial yang diperoleh ketika pendapatan dari bisnis melebihi total pengeluaran, biaya lainnya

Berdasarkan beberapa definisi tersebut, dapat disimpulkan bahwa Keuntungan adalah selisih positif antara pendapatan yang diperoleh dari aktivitas bisnis dengan total biaya, pengeluaran, dan pajak yang dikeluarkan untuk

menghasilkan pendapatan tersebut. Keuntungan menjadi hasil finansial yang tercapai ketika pendapatan melebihi biaya yang dikeluarkan, dan jika biaya lebih besar, maka akan disebut rugi.

2.4 Regresi Linier

Regresi linier adalah salah satu metode statistik yang digunakan untuk menganalisis hubungan antara satu variabel independen (bebas) dengan satu variabel dependen (tergantung). Dalam analisis bisnis, regresi linier sering digunakan untuk memprediksi nilai suatu variabel. [7] Metode ini menggunakan persamaan linear yang menghubungkan antara variabel independen dan variabel dependen.

2.4.1 Persamaan Regresi Linier

Persamaan Regresi linear dapat dinyatakan sebagai berikut:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

Di mana:

y : Variabel dependen (yang diprdiksi)

x : Variabel independen (faktor yang mempengaruhi y)

β_0 : *Intersep,* yaitu nilai y dan x bernilai nol.

β_1 : Koefisien reqresi (mengukur pengaruh x terhadap y)

ε : *Eror* (selisih antara nilai aktual dan prediksi).

Regresi linier sangat berguna dalam menganalisis hubungan yang sederhana antara variabel-variabel yang dapat diukur secara kuantitatif. Namun, regresi linier memiliki keterbatasan dalam menangani hubungan yang lebih

kompleks, terutama ketika hubungan antara variabel tidak linier.[8] Namun, regresi linier sederhana memiliki keterbatasan, terutama jika hubungan yang dianalisis melibatkan lebih dari satu faktor. Dalam kasus tersebut, Regresi Linier Berganda menjadi solusi yang lebih tepat.

2.4.2 Regresi Linier Berganda

Regresi linier berganda adalah metode statistik yang digunakan untuk digunakan untuk menganalisis hubungan antara satu variabel dependen (y) dengan dua atau lebih variabel indenpenden (X). Metode ini lebih kompleks dibandingkan regresi linier sederhana, tetapi memberikan hasil yang komprehensif karena mempertimbangkan berbagai faktor secara simultan. [9] Tujuan dari analisis regresi linier berganda adalah untuk memprediksi nilai variabel tak bebas atau response (Y) jika nilai variabel-variabel bebas atau *predictor* ($x_1, x_{2,.....,} x_n$). diketahui. Disamping itu juga untuk mengetahui arah hubungan antara variabel tak bebas dengan variabel-variabel bebas.[10]

2.4.3 Persamaan Regresi Linier Berganda

Berikut adalah model persamaan Regresi Linier Berganda :

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ...+\beta_n x_n + \varepsilon$$

Di mana:

y : Variabel dependen (yang diprdiksi)

$x_1, x_{2,....,} x_n$: Variabel independen (faktor-faktor yang mempengaruhi y)

β_0 : *Intersep,* yaitu nilai y ketika semua variabel x bernilai nol.

β_1, $\beta_{2,....,}$ β_n : Koefisien reqresi masing-masing variabel *x* .koefisien ini menunjuukan sejauh mana perubahan pada variabel *x* mempengaruhi variabel *y*.

ε : *Eror* (selisih antara nilai aktual dan prediksi).

Regresi linier berganda berguna untuk menganalisis dan memahami hubungan antara satu variabel dependen dengan beberapa variabel independen, mengukur pengaruh masing-masing variabel independen terhadap variabel dependen, serta membuat prediksi berdasarkan hubungan linear yang diidentifikasi.[11]

2.5 Support Vector Machine (SVM)

Support Vector Machine (SVM) metode pada machine learning yang dapat digunakan untuk menganalisis data dan mengurutkannya ke dalam salah satu dari dua kategori. SVM bekerja dengan cara memetakan data ke ruang fitur berdimensi tinggi dan kemudian mencari hyperplane yang optimal untuk memisahkan dua kelas pola. Fungsi kernel digunakan untuk mengklasifikasikan data non-linier dan mengubah data *nonlinear* menjadi data linier.[12]

Selain digunakan untuk klasifikasi, SVM juga dapat diterapkan pada masalah regresi, yang dikenal sebagai *Support Vector Regression* (SVR). Dalam kasus regresi, SVM berusaha menemukan fungsi yang memiliki margin tertentu di sekitar data. Ini membuat SVM sangat fleksibel untuk berbagai jenis masalah, baik klasifikasi maupun prediksi.SVM dapat menjadi alat yang sangat kuat untuk menghasilkan model prediksi yang akurat dan andal.

2.5.1 Konsep Dasar SVM untuk Regresi

Konsep SVM menitik beratkan pada *risk minimization*, yaitu untuk mengestimasi suatu fungsi dengan cara meminimalkan batas atas dari *generalization error*, sehingga SVM mampu mengatasi *overfitting*. Fungsi regresi dari metode SVM adalah sebagai berikut:

$$F(x) = w^T x + b$$

dimana w merupakan sebuah vector pembobot, (T) merupakan sebuah fungsi yang memetakan x ke dalam suatu dimensi dan b merupakan faktor bias.

2.5.2 Fungsi *Kernel*

Fungsi *kernel* pada SVM digunakan untuk memetakan data *non-linear* ke dimensi yang lebih tinggi agar hubungan kompleks lebih mudah ditemukan, dengan jenis yang sering digunakan seperti *Linear Kernel* untuk hubungan *linier*, *Polynomial Kernel* untuk hubungan *non-linear* dengan derajat tertentu, Radial *Basis Function* (RBF) *Kernel* untuk fleksibilitas tinggi, dan *Sigmoid Kernel* yang menyerupai fungsi aktivasi pada jaringan saraf. Beberapa jenis kernel yang sering digunakan adalah:

1. *Linear Kernel*

   Kernel ini digunakan ketika data dapat dipisahkan secara linier. Rumus matematikanya adalah:

   $$K(x, x') = x \cdot x'$$

   Di sini, x dan x′ adalah vektor fitur dari dua titik data.

2. *Polynomial Kernel*

   Kernel ini digunakan untuk data yang memiliki hubungan polinomial,

sehingga dapat memodelkan hubungan yang lebih kompleks daripada linear. Rumusnya adalah :

$$K(x, x') = (x \cdot x' + c)^d$$

Di mana *c* adalah konstanta dan *d* adalah derajat polinomial.

3. *Radial Basis Function* (RBF)

   Kernel ini sangat efektif untuk menangani data yang lebih kompleks dengan pola yang tidak linier. Fungsi kernel RBF memiliki rumus :

$$K(x, x') = \exp\left(-\frac{||x - x'||^2}{2\sigma^2}\right)$$

   Di mana σ adalah parameter yang mengontrol seberapa besar pengaruh data titik tersebut.

Kernel-kernel ini digunakan untuk mengubah ruang fitur data sehingga masalah klasifikasi dapat diselesaikan dengan lebih efektif, terutama dalam kasus data yang tidak linier.[13]

2.6 Perbandingan Regresi Linier dan *Support Vector Machine* (SVM)

Perbandingan antara regresi linier dan Support Vector Machine (SVM) telah banyak diteliti dalam berbagai konteks. Berikut adalah beberapa temuan dari penelitian terkait:

1. Analisis Kinerja CPU Sebuah penelitian membandingkan regresi linier dan SVM untuk estimasi kinerja CPU. Meskipun detail spesifik tidak tersedia

dalam cuplikan, studi ini menunjukkan bahwa kedua metode memiliki kelebihan dan kekurangan masing-masing dalam konteks tersebut. [14]

2. Machine Learning untuk Memprediksi Jumlah Penjualan, Stok, dan Jumlah Tanam pada Hidroponik Tilung Farm Penelitian ini membandingkan algoritma Regresi Linier dan SVM dalam memprediksi data stok, transaksi, dan jumlah tanam. Hasilnya menunjukkan bahwa algoritma Regresi Linier menghasilkan nilai MSE, MAE, dan MAPE yang lebih kecil dibandingkan dengan SVM, sehingga dianggap lebih baik dalam memprediksi data tersebut.[15]

2.7 Evaluasi Model

Evaluasi algoritma digunakan untuk mengukur performa model yang diterapkan, dalam hal ini regresi linier dan SVM. Tiga metrik utama yang digunakan untuk mengevaluasi model ini adalah *Mean Absolute Error (MAE), Root Mean Squared Error (RMSE),* Metrik-metrik ini membantu dalam menilai seberapa baik model memprediksi data yang ada.[16]

2.7.1 Mean Absolute Error (MAE)

MAE mengukur rata-rata selisih absolut antara nilai prediksi dan nilai aktual. Metrik ini memberikan gambaran yang jelas mengenai seberapa besar rata-rata kesalahan prediksi yang terjadi, tanpa memperhitungkan arah kesalahan (positif atau negatif). Semakin kecil nilai MAE, semakin akurat model prediksi. MEA dihitung dengan rumus:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Di mana y_i adalah nilai aktual dan $\hat{y}_i$ adalah nilai prediksi.[17]

2.7.2 Root Mean Squared Error (RMSE)

RMSE mengukur akar kuadrat dari rata-rata kuadrat selisih antara nilai prediksi dan nilai aktual. RMSE sangat sensitif terhadap outlier karena memberi penalti lebih besar pada kesalahan besar. Rumus:

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

Di mana y_i adalah nilai aktual, $\hat{y}_i$ adalah nilai prediksi dan n adalah jumlah data RMSE akan memberikan nilai dalam satuan yang sama dengan data asli, yang memudahkan interpretasi kesalahan dalam konteks masalah yang sedang dianalisis.[18]

2.7.3 R-squared (R^2)

Mengukur proporsi variansi dalam data yang dapat dijelaskan oleh model. Nilai R^2 berkisar antara 0 dan 1, di mana nilai 1 berarti model menjelaskan seluruh variansi data, sedangkan nilai 0 berarti model tidak dapat menjelaskan variansi sama sekali.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}$$

Di mana y_i adalah nilai aktual, $\hat{y}_i$ adalah nilai prediksi dan $\bar{y}$ adalah nilai rata-rata nilai aktual.

2.8 Alat Bantu Pemrograman untuk Implementasi

Dalam penelitian ini, berbagai perangkat lunak digunakan untuk menganalisis dan membandingkan metode *Regresi linier* dan *Support Vector Machine* (SVM). Berikut adalah beberapa perangkat lunak yang digunakan.

1. *Microsoft Excel*

   Microsoft Excel adalah perangkat lunak yang mudah digunakan dan sangat populer dalam dunia bisnis dan analisis data sederhana. Fitur analisis data bawaan seperti *Analysis ToolPak* memungkinkan pengguna untuk melakukan analisis regresi linier tanpa perlu keahlian teknis mendalam.

Gambar 2.2 Logo Excel

Sumber : Wikipedia.org

   Excel sangat populer karena kemampuannya dalam mengelola data dalam bentuk tabel, melakukan perhitungan otomatis menggunakan rumus, serta menyajikan data dalam bentuk grafik dan diagram yang mudah dipahami.

2. RapidMiner

   RapidMiner adalah *platform* perangkat lunak open-source untuk analisis data dan pembelajaran mesin. Dengan antarmuka grafis yang user-friendly analisis data berbasis *drag-and-drop* yang mendukung berbagai algoritma *machine learning*, termasuk Regresi Linier, Regresi Linier Berganda, dan

Support Vector Machine (SVM), serta memungkinkan evaluasi dan perbandingan model secara otomatis. [19]

Gambar 2.3 Logo RapidMiner

Sumber : Reveneer.io

RapidMiner menyediakan berbagai alat untuk ekstraksi data, pemrosesan data, pembuatan model statistik dan prediksi, dan evaluasi hasil model.

2.9 Metode Penelitian

Penelitian ini menggunakan metode kuantitatif dengan fokus pada analisis penjualan produk Amanda Brownies.Tujuan utamanya adalah untuk meningkatkan keuntungan dengan membandingkan efektivitas model Regresi Linier dan *Support Vector Machine* (SVM) dalam memprediksi penjualan. Sampel penelitian adalah data transaksi penjualan harian September hingga November 2024. Metode ini dilakukan untuk menentukan algoritma yang lebih baik berdasarkan akurasi prediksi, kestabilan model, dan relevansinya dalam pengambilan keputusan bisnis.

2.9.1 Variabel Penelitian

Penelitian ini menganalisis data penjualan Amanda Brownies dengan fokus pada variabel jumlah barang terjual dan total pedapataan yang memper-ngaruhi pola penjualan. Algoritma regresi linier dan SVM digunakan untuk

memprediksi keuntungan berdasarkan hubungan antar variabel. Berikut adalah variabel studi.

1. Jumlah Produk Terjual

   Data jumlah unit Amanda Brownies yang terjual setiap hari selama tiga bulan terakhir.

2. Jumlah Pendapatan

   Data pendapatan yang diperoleh setiap hari dari penjualan Amanda Brownies.

2.9.2 Metode Pengumpulan Data

Data yang digunakan dalam penelitian ini diperoleh melalui beberapa teknik pengumpulan data sebagai berikut:

1. Wawancara

   Wawancara adalah teknik pengumpulan data yang dilakukan dengan cara bertanya langsung kepada narasumber untuk memperoleh informasi mendalam .

2. Dokumentasi

   Dokumentasi adalah teknik pengumpulan data dengan cara mengumpulkan dan menganalisis dokumen atau arsip yang sudah tersedia, seperti laporan penjualan, laporan keuangan.

3. Observasi

   Observasi adalah teknik pengumpulan data yang dilakukan dengan mengamati langsung kegiatan yang terjadi di lapangan,

2.9.3 Kerangka kerja penelitian

Kerangka kerja penelitian yang akan dilakukan disajikan pada tabel dibawah ini yaitu sebagai berikut :

Tabel 2.1 Kerangka kerja penelitian

No	Kegiatan	Nov	Des	Jan	Feb	Mar
1	Penetuan topik penelitian	✓				
2	Pendefinisian Masalah	✓				
3	Menganalisa Masalahan	✓				
4	Menentukan Tujuan	✓				
5	Pengumpulan Data	✓	✓			
6	Cleaning Data	✓	✓			
7	Transformasi Data	✓	✓	✓		
8	Merancang Algoritma	✓	✓	✓		
9	Pengujian Algoritma			✓	✓	
10	Evaluasi Akhir				✓	✓
11	Pengajuan Seminar Proposal				✓	
12	Seminar Proposal				✓	

Catatan :

* Durasi setiap kegiatan dapat disesuaikan dengan kebutuhan penelitian.

* Waktu pelaksanaan bergantung pada kalender akademik dan kesiapan.

2.10 Peneltian Terdahulu

Judul	Nama Peneliti	Tahun	Hasil
Analisis Perbandingan Model Regresi Linier dan *Support Vector Machine* untuk Estimasi Kinerja CPU	Aji Suatmaji Prasetiyo, Anita Adelia Syahfitri, Maria Susye Marni Berek, Naratama Rizky Rawi Saputra	2024	Algoritma Regresi Linier menghasilkan nilai RMSE 61,223, yang lebih rendah dibandingkan dengan *Support Vector Machine* (SVM) dalam estimasi kinerja CPU. Regresi Linier digunakan untuk estimasi kinerja CPU.
Implementation of GridSearch to Improve the Performance of the Support Vector Regression (SVR) *Model for Predicting Product Sales* at Rohman Jaya[20]	Ahmad Baidowi, Eko Fitra Firmanda, Ahmad Hudawi AS, Abu Tholib,	2024	Penggunaan optimasi *GridSearch* pad a SVR meningkatkan akurasi prediksi penjualan dengan nilai MAPE dari 40,39% (tanpa optimasi) menjadi 0,45% (dengan optimasi).
Perbandingan Metode *Algoritma Support Vector Regression* dan *Multiple Linear Regression* Untuk Memprediksi Stok Obat[21]	Lit Malem Ginting, Marojahan MT.Sigiro, Esra Delima Manurung, Juan Jasa Putra Sinurat	2021	*Multiple Linear Regression* (MLR) menunjukkan MSE sebesar 1.41348%, lebih baik dari *Support Vector Regression* (SVR) yang memiliki MSE 0.00135%. MLR lebih akurat dalam memprediksi stok obat dibandingkan SVR.