Prediksi Kuat Tekan Semen Untuk Produk *Portland Composite Cement* (PCC) di PT. Semen Indonesia (Persero) Tbk. Menggunakan *Support Vector Regression* (SVR) Dengan *Feature Selection*

Rizki Febriasto, Ni Luh Putu Satyaning P.P dan Wibawati.
Departemen Statistika, Fakultas Matematika, Komputasi, dan Sains Data,
Institut Teknologi Sepuluh Nopember (ITS)
e-mail: pradnya@statistika.its.ac.id

***Abstrak*—Indonesia sebagai negara berkembang terus-menerus melakukan pembangunan dalam segala bidang kehidupan. Salah satu sektor yang selalu berkembang adalah sektor infrastruktur. Dengan menghadapi fenomena pertumbuhan infrastruktur ini, khususnya di sektor pembangunan perlu adanya hal yang menunjang dalam setiap pembangunan yang terjadi, salah satunya adalah material semen. Salah satu perusahaan yang bergerak di bidang produksi semen adalah PT. Semen Indonesia (Persero) Tbk. Terdapat beberapa produk yang dihasilkan oleh PT. Semen Indonesia (Persero) Tbk, salah satunya adalah Portland Composite Cement (PCC). Dalam penelitian ini akan dilakukan prediksi terhadap kuat tekan semen PCC di hari ke-28. Data yang digunakan yaitu bulan Juli 2018 hingga Maret 2019. Pada hasil analisis, hasil regresi linier telah terindikasi mengalami multikolinearitas, sehingga ditangani menggunakan PCR. Feature selection (RF-RFE) digunakan dan telah menghilangkan lima variabel. Dengan teknik machine learning yaitu SVR didapatkan model terbaik yaitu SVR dengan fungsi kernel Radial Basis Function (RBF) dengan complete feature. Feature selection tidak menghasilkan hasil yang lebih baik dibanding complete feature. Model terbaik digunakan untuk memprediksi kuat tekan pada bulan April 2019 yang menghasilkan kriteria terbaik RMSE sebesar 8,78.**

Kata Kunci—Kuat Tekan, Multikolinearitas, PCC, PCR, RBF, RFE, RMSE, SVR.

I. PENDAHULUAN

INDONESIA sebagai negara berkembang terus-menerus melakukan pembangunan dalam segala bidang kehidupan. Salah satu sektor yang selalu berkembang adalah sektor infrastruktur. Perkembangan bidang pembangunan juga selalu berkembang sejalan dengan semakin banyaknya penduduk yang berada di Indonesia. Berdasarkan proyeksi badan perencanaan pembangunan nasional (Bappenas) jumlah penduduk Indonesia pada tahun 2018 mencapai 265 juta jiwa. Dengan menghadapi fenomena pertumbuhan penduduk ini, khususnya di sektor pembangunan perlu adanya hal yang menunjang dalam setiap pembangunan yang terjadi, salah satunya adalah material-material pembangunan seperti beton, mortar dan semen. Kebutuhan semen selalu meningkat, hal ini ditunjukkan dengan penjualan semen yang mengalami kenaikan setiap tahunnya. Selama 14 tahun penjualan semen di Indonesia mulai tahun 2002 hingga 2016 meningkat sebesar 125% dari titik sebelumnya.

Salah satu perusahaan yang bergerak di bidang produksi semen adalah PT. Semen Indonesia (Persero) Tbk. Dengan bertambah pesatnya bisnis industri semen membuat semakin ketat pula persaingan antar industri dalam memperebutkan *customer* serta mempertahankan pasar yang ada. Salah satu penilaian inti dari semen adalah kuat tekan semen tersebut, jika kuat tekan semakin bagus, maka pelanggan akan semakin puas. Selama ini pengujian kuat tekan hanya dilakukan dengan mesin. Semen Indonesia harus melakukan *improvement* disetiap sistem yang berjalan. Maka dari itu, Semen Indonesia harus melakukan beberapa strategi agar *customer* tidak beralih ke produk semen lainnya. *Portland Composite Cement* (PCC) merupakan bahan pengikat hidraulis hasil penggilingan bersama-sama terak semen *portland* dan gips dengan satu atau lebih bahan anorganik, bahan anorganik tersebut antara lain terak tanur tinggi, pozolan, senyawa silikat, batu kapur dengan kadar total bahan anorganik sekitar 6%-35% dari massa *Portland Composite Cement* (PCC) [1].

Dalam penelitian ini menggunakan produk PCC (*Portland Composite Cement)*. Variabel yang digunakan adalah variabel kuat tekan dengan beberapa komposisi senyawa dan beberapa proses, kemudian akan dilakukan estimasi terhadap kuat tekan semen hari ke-28, pada hari sebelumnya semen masih beresiko mengalami perubahan kualitas dan tidak menjamin apakah dapat menghasilkan semen yang baik pada hari ke-28. Metode yang digunakan untuk mengestimasi kuat tekan semen PCC adalah metode *Support Vector Regression* (SVR) yang merupakan pengembangan metode dari *Support Vector Machine* (SVM). SVR memiliki tujuan memetakan vektor input ke dalam dimensi yang lebih tinggi. SVR juga digunakan karena beberapa proses produksi pada semen memiliki indikasi yang berhubungan di variabel prediktor yang satu dengan yang lain, sedangkan pada regresi linier sederhana tidak diperbolehkan adanya hubungan antara variabel prediktor. Variabel yang menunjang untuk memestimasi ini adalah senyawa kimia yang terkandung didalam semen, kehalusan semen, ekspansi dan beberapa variabel lainnya. Sebelum menggunakan SVR, akan dilakukan pemilihan variabel atau *Feature Selection* dengan metode *Recursive Feature Elimination* berbasis *Random Forest.*

Penelitian sebelumnya adalah Abdul (2018) yang menggunakan judul sistem pendukung keputusan pemberian bonus tetap memanfaatkan SVR bahwa pemilihan metode antara SVR dengan *neural network* didasari dengan nilai MSE, didapatkan bahwa nilai MSE pada SVR lebih rendah

dibandingkan *neural network*. Hasbi (2014) yang memprediksi kurs rupiah terhadap dollar amerika menggunakan SVR disimpulkan bahwa pada pengujian data *testing* kernel linier dan *ploynomial* menghasilkan dan eror yang kecil. Hendra (2012) yang menggunakan metode *feature selection* RFE menyimpulkan bahwa dengan RFE memiliki hasil RMSE yang lebih kecil dibandingkan *complete feature.*

II. DASAR TEORI

A. Pre-Processing

Pre-processing data merupakan sebuah langkah penting dalam data *minning* untuk membuat data lebih berkualitas. Data perlu dilakukan *pre-processing* karena dalam data mentah terdapat data yang tidak lengkap, *noise* dan tidak konsisten. Terdapat beberapa cara untuk melakukan *pre-processing* salah satunya adalah *data cleaning* yang merupakan proses untuk membersihkan data salah satunya adalah membuang data yang bersifat *outlier*. Yang kedua adalah *data integration*, yaitu merupakan integrase dari data-data yang digunakan seperti korelasi. Selanjutnya adalah *data transformation* yang membuat perubahan data ketika terdapat perbedaan satuan antar variabel dan pola dapat dipahami. Yang terakhir adalah *data reduction* yang bertujuan untuk mereduksi data ataupun variabel agar lebih mudah diolah tetapi tidak menghilangkan karakteristik data tersebut [2].

B. Recursive Feature Elimination

Feature selection adalah upaya untuk memilih fitur subset dari fitur asli yang paling berguna. *Feature extraction* adalah upaya untuk memetakan semua fitur ke dalam fitur baru yang lebih sedikit. Kelebihan *feature selection* dibandingkan *feature extraction* adalah akusis data yang lebih cepat. Oleh karena itu, pengurangan fitur pada data *hyperspectral* yang berupa *band* akan lebih baik menggunakan *feature selection* [3]. Salah satu dari banyak *feature selection* yang tersedia adalah RFE (*Recursive Feature Elimination*).

Salah satu metode yang dapat digunakan dengan RFE ini adalah *Random Forest*. RF menunjukkan kelebihannya antara lain dapat menghasilkan eror yang lebih rendah, memberikan hasil yang bagus dalam klasifikasi, dapat mengatasi data *training* dalam jumlah yang sangat besar secara efisien dan metode yang efektif untuk mengestimasi *missing value* [4].

Random Forest juga merupakan sebuah teknik *machine learning* yang dapat digunakan dengan dimensi data yang tinggi dan memperbolehkan hubungan non-linier yang berada di variabel prediktor, khususnya pada beberapa variabel prediktor yang memiliki hubungan sangat kuat. Tidak semua metode dapat digabungkan dengan RFE, terdapat beberapa metode yang memiliki keuntungan atau kerugian lebih, karena RFE membutuhkan model awal menggunakan set prediktor penuh, maka beberapa model tidak dapat digunakan ketika jumlah prediktor melebihi jumlah sampel. *Random Forest* merupakan salah satu model yang dapat digabungkan dengan RFE (RF-RFE).

C. K-*Fold Cross Validation*

Sebelum dilakukan tahap estimasi, dilakukan pembagian data menjadi *training* dan *testing*. *Cross validation* merupakan salah satu metode pembagian data. Metode ini mempartisi data ke dalam dua subset data yang berukuran sama, salah satu sebagai *training* dan salah satu sebagai data *testing*, kemudian dilakukan pertukaran fungsi dari subset sedemikian sehingga subset sebelumnya sebagai *training set* dan *test set*. Metode *k-fold cross validation* menggeneralisasi pendekatan ini dengan mensegmentasi data ke dalam *k* partisi berukuran sama. Selama proses, prosedur pembagian data *training* dan *testing* diulang sebanyak *k* kali, sehingga setiap subset akan menjadi data uji dari model. Proses yang telah dilakukan sebanyak *k* kali akan mendapatkan *k* buah nilai dari proses pembelajaran. Semua nilai performa ini akan dicari rata-ratanya dan nilai dengan rata-rata tertinggi akan dipilih sebagai model. *k-fold cross validation* memiliki kelebihan dapat mengklasifikasi dataset lebih efisien, namun metode ini memiliki kelemahan dalam proses komputasi yang digunakan akan lebih besar karena akan melakukan proses sebanyak *k* kali. *Cross validation* adalah bentuk sederhana dari statistik, jumlah *fold* standar untuk memprediksi tingkat eror dari data adalah dengan menggunakan 10-fold *cross validation* [5]. Selain untuk pembagian data *training* dan *testing*, teknik ini untuk melakukan validasi pada dataset untuk menemukan akurasi yang baik.

D. Regresi Linier

Analisis regresi merupakan sebuah alat statistik yang berguna untuk mendapatkan hubungan fungsional antara dua variabel atau lebih yaitu variabel respon dan prediktor. Tujuan dari analisis regresi agar mendapatkan pengaruh antara variabel prediktor terhadap responnya [5]. Bentuk persamaan umum regresi liner berganda adalah sebagai berikut :

$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_i x_{ki} + \varepsilon_i \qquad (1)$$

keterangan :

y_i : variabel respon ke-i dari model.

$\beta_0, \beta_1, \ldots, \beta_k$: parameter dari model.

$x_1, x_2, \ldots, x_k$: variabel prediktor ke-i dari model.

ε_i : galat model ke-i.

Dari model yang terbentuk, terdapat beberapa asumsi pada residual yang harus dipenuhi, yaitu asumsi identik, independen dan berdistribusi normal.Asumsi identik merupakan salah satu asumsi residual yang penting dari model regresi. Varians residual harus bersifat homoskedastisitas atau varians residual bersifat identic [5]. Suatu data dikatakan identik apabila plot residualnya menyebar secara acak dan tidak membentuk suatu pola tertentu [6]. Asumsi saling bebas *(independent)* atau autokorelasi residual, yang dilakukan untuk mengetahui apakah ada korelasi antar residual. Suatu data dikatakan independen apabila plot residualnya menyebar secara acak dan tidak membentuk suatu pola tertentu [6]. Pengujian Asumsi Residual berdistribusi normal dilakukan untuk melihat apakah residual memenuhi asumsi berdistribusi normal atau tidak. Kenormalan suatu data dapat dilihat dari plotnya. Apabila plot sudah mendekati garis linier, dapat dikatakan bahwa data tersebut memenuhi asumsi yaitu berdistribusi normal [6].

E. Principle Component Regression

Principle Component Regression (PCR) merupakan suatu teknik analisis yang mengkombinasikan antara analisis regresi dengan *Principal Component Analysis* (PCA). Analisis Regresi digunakan untuk mengetahui ada tidaknya hubungan antara variabel *dependent* dan *independent*,

sedangkan PCA pada dasarnya bertujuan untuk menyederhanakan variabel yang diamati dengan cara menyusutkan (mereduksi) dimensinya. Hal ini dilakukan dengan jalan menghilangkan korelasi di antara variabel independen melalui transformasi variabel asal ke variabel baru (merupakan kombinasi linear dari variabel-variabel asal) yang tidak saling berkorelasi. Dalam hal ini, PCR bisa menanggulangi kasus multikolinearitas. Dari p buah variabel asal dapat dibentuk p buah komponen utama, dipilih k buah komponen utama saja (k<p) yang telah mampu menerangkan keragaman data cukup tinggi (antara 80% sampai dengan 90%) [8]. Komponen utama yang dipilih tersebut (k buah) dapat mengganti p buah variabel asal tanpa mengurangi informasi.

Analisis regresi komponen utama (PCR) merupakan analisis regresi variabel *dependent* terhadap komponen-komponen utama yang tidak saling berkorelasi, regresi komponen utama dapat dinyatakan sebagai berikut :

$$Y = w_o + w_1K_1 + w_2K_2 + ... + w_mK_m + \varepsilon \quad (2)$$

$K_{1,2},K_3,\ldots,K_m$ menunjukan komponen utama yang dilibatkan dalam analisis regresi komponen utama, dimana besaran m lebih kecil daripada banyaknya variabel *independent* yaitu sejumlah p, serta Y sebagai variabel *dependent*. Komponen utama merupakan kombinasi linear dari variabel baku Z, sehingga:

$$\begin{aligned} K_1 &= a_{11}Z_1 + a_{21}Z_2 + ... + a_{p1}Z_p \\ K_2 &= a_{12}Z_1 + a_{22}Z_2 + ... + a_{p2}Z_p \\ &\vdots \\ K_m &= a_{1m}Z_1 + a_{2m}Z_2 + ... + a_{pm}Z_p \end{aligned} \quad (3)$$

Apabila $K1,K2,\ldots,Km$ dalam persamaan (3) didistribusikan kembali ke dalam persamaan regresi komponen utama, yaitu persamaan (2) maka diperoleh:

$$\begin{aligned} Y = w_o &+ w_1(a_{11}Z_1 + a_{21}Z_2 + ... + a_{p1}Z_p) + \\ &w_2(a_{12}Z_1 + a_{22}Z_2 + ... + a_{p2}Z_p) + ... + \\ &w_m(a_{1m}Z_1 + a_{2m}Z_2 + ... + a_{pm}Z_p) + \varepsilon \end{aligned} \quad (4)$$

Persamaan regresi linear dugaan komponen utama sebagai berikut:

$$Y = b_0 + b_1Z_1 + b_2Z_2 + ... + b_pZ_p \quad (5)$$

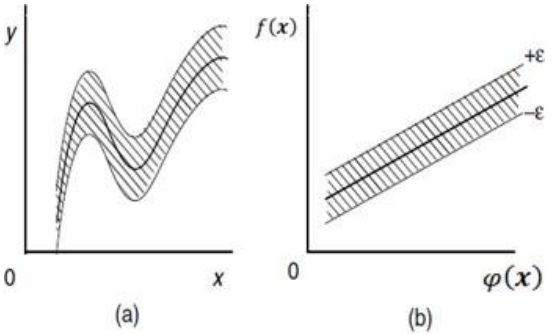
F. Support Vector Regression

Support Vector Regression (SVR) merupakan suatu metode *Machine Learning* yang dikembangkan dari metode *Support Vektor Machine* (SVM). Yang berbeda dari kedua metode ini adalah jika SVM satu kumpulan teknik klasifikasi, sedangkan SVR digunakan pada kasus regresiTujuan dari SVR adalah untuk menemukan sebuah fungsi sebagai suatu *hyperplane* (garis pemisah) berupa fungsi regresi yang mana sesuai dengan semua input data dengan sebuah error dan membuat error sekecil mungkin [9]. Tujuan lain dari SVR ini adalah untuk memetakan *vector* input ke dalam dimensi yang lebih tinggi. Misalkan sebuah fungsi berikut adalah garis regresi sebagai *optimal hyperplane* :

$$f(x) = w^T\varphi(x) + b \quad (6)$$

Pada regresi terdapat residual misalkan residual didefinisikan dengan mengurangkan *output* scalar y terhadap estimasi $f(x)$ yaitu $r = y - f(x)$ dengan:

$$E(r) = \begin{cases} 0, \text{untuk } |r| \le \varepsilon \\ |r| - \varepsilon, \text{untuk yang lain} \end{cases} \quad (7)$$



Gambar 1. *Insensitive zone* (a) *original input space*, dan (b) *feature space*

$D(x,y)$ adalah jarak terjauh *support vector* dari *hyperplane*, kemudian disebut margin. Memaksimalkan margin akan meningkatkan probabilitas data ke dalam radius $\pm\varepsilon$. Jarak dari *hyperplane* $D(x,y) = 0$ ke data adalah (x,y) adalah $|D(x,y)| / \|W^*\|$, dengan:

$$W^* = (1 - W^T)^T \quad (8)$$

Diasumsikan bahwa jarak maksimum data terhadap *hyperlane* adalah δ, maka estimasi yang ideal akan terpenuhi dengan :

$$\begin{aligned} &\frac{|D(x,y)|}{\|W^*\|} \le \delta \\ &|D(x,y)| \le \delta \|W^*\| \\ &\delta \|W^*\| = \varepsilon \end{aligned} \quad (9)$$

Oleh karena itu untuk memaksimalkan margin φ, diperlukan $\|W^*\|$ yang minimum, Optimasi penyelesaian masalah dengan bentuk *Quadratic Programming* adalah sebagai berikut :

$$\min \frac{1}{2}\|W^*\|^2 \quad (10)$$

dengan syarat :

$$y_i - W^T\varphi(x_i) - b \le \varepsilon \text{ untuk } i = 1,...,l \quad (11)$$

$$W^T\varphi(x_i) - y_i + b \le \varepsilon \text{ untuk } i = 1,...,l \quad (12)$$

Faktor $\|W^*\|$ dinamakan regulas. Meminimalkan $\|W^*\|$ akan membuat suatu fungsi setipis (*flat*) mungkin, sehingga bias mengontrol kapasitas fungsi (*function capacity*) [10].

1. Fungsi Kernel

   Menurut Santosa (2007) [11]. banyak teknik *Machine Learning* yang dikembangkan dengan asumsi kelinieran, sehingga algoritma yang dihasilkan terbatas untuk kasus-kasus yang linier. Dengan metode kernel suatu data x di *input space* dipetakan ke *feature space* dengan dimensi yang lebih tinggi melalui φ. Fungsi Kernel yang digunakan adalah :

   1. Kerner Linier

      $$\varphi(x) = K(x,x') = x^Tx \quad (13)$$

   2. Kernel *Polynomial*

      $$\varphi(x) = K(x,x') = (\gamma(x^Tx) + 1)^d \quad (14)$$

   3. *Radial Basis Function* (RBF)

$$\varphi(x) = K(x, x') = \exp\left(-\gamma \|x - x_i\|^2\right) \quad (15)$$

Nilai $K(x_i, x_j)$ merupakan funsi kernel yang menunjukkan pemetaan linier pada *feature space*. Perlu dijelaskan bahwa nilai $K(x_i, x_j)$ tidak selalu bisa diekspresikan secara eksplisit sebaga kombinasi antara α, y dan $\varphi(x)$, karena dalam banyak kasus $\varphi(x)$ tidak diketahui dan sulit dihitung. Sedangkan x dan x' adalah pasangan dua data *training*. Parameter d > 0 merupakan konstanta. Fungsi kernel mana yang harus digunakan untuk subtitusi *dot product* di *feature space* sangat tergantung pada data karena fungsi kernel ini akan menentukan fitur baru dimana fungsi pemisah akan dicari.

G. Pemilihan Model Terbaik

Pemilihan model terbaik dapat dilakukan dengan berbagai metode. Hal ini dilakukan karena terciptanya beberapa model yang layak pakai. Metode yang digunakan dalam penelitian ini *Root Mean Square Error* (RMSE).

1. *Root Mean Square Error*

   RMSE (*Root Mean Square Error*) merupakan kriteria pemilihan model terbaik berdasarkan pada data yang telah dibagi menjadi data *training* dan data *testing*. Model terbaik dipilih yang memiliki nilai kriteria error terkecil [12].

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2} \quad (16)$$

H. Kuat Tekan Semen

Kuat tekan merupakan kekuatan tekan maksimum yang dapat dipikul benda persatuan luas. Semen yang dicampur dengan air dan menjadi semen hidraulis akan diuji kuat tekannya. Biasanya, kuat tekan diukur pada hari ke-3, hari ke-7, hari ke-14, hari ke-21 dan hari ke-28. Penentuan kuat semen mengacu kepada ASTM C 109/109M-02 (*Standard Test Method for Compressive Strength of Hydraulic Cement Mortar*) [13]. Metode uji ini melingkupi penentuan kuat tekan mortar semen hidraulis dengan menggunakan cetakan kubus berukuran sisi 50 mm.

III. METODOLOGI PENELITIAN

A. Sumber Data

Data yang digunakan dalam penelitian ini adalah data sekunder yang diperoleh pada produksi bulan Juli 2018 hingga Januari 2019. Data diperoleh dari PT. Semen Indonesia (Persero) Tbk. pabrik Tuban dengan produk yang diamati adalah produk *Portland Composite Cement* (PCC).

B. Struktur Data

Berikut merupakan variabel penelitian yang digunakan yang merupakan karakteristik kualitas kuat tekan dari produk *Portland Composite Cement* (PCC) pada hari ke-28, yang sesuai dengan standar perusahaan.

C. Langkah Analisis

Berikut ini adalah langkah analisis yang digunakan dalam estimasi kuat tekan produk *Portland Composite Cement* menggunakan data produksi.

1. Mengumpulkan data sekunder hasil produksi pada produk PCC di PT. Semen Indonesia (Persero) Tbk.
2. Melakukan *pre-processing* pada data produksi PCC.
3. Melakukan eksplorasi data produksi PCC.
4. Pembagian data yaitu data *training* dan data *testing* menggunakan k-*fold cross validation*. Label *fold* pada data dengan variabel prediktor lengkap maupun yang terpilih adalah sama.

Tabel 1.
Variabel Penelitian

Variabel	Keterangan	Satuan	Skala Data
Y	Kuat Tekan Hari ke-28	Kg/cm^2	Rasio
X_1	Oksida Silkon (SiO_2)	%	Rasio
X_2	Oksida Alumunium (Al_2O_3)	%	Rasio
X_3	Oksida Besi (Fe_2O_3)	%	Rasio
X_4	Oksida Kalsium (CaO)	%	Rasio
X_5	Magensium Oksida (MgO)	%	Rasio
X_6	Sulfur Trioksida (SO_3)	%	Rasio
X_7	*Freelime*	%	Rasio
X_8	*Insoluble Residu*	%	Rasio
X_9	*Loss of Ignition*	%	Rasio
X_{10}	*Blaine*	Luasan/massa	Rasio
X_{11}	*Residu*	%	Rasio

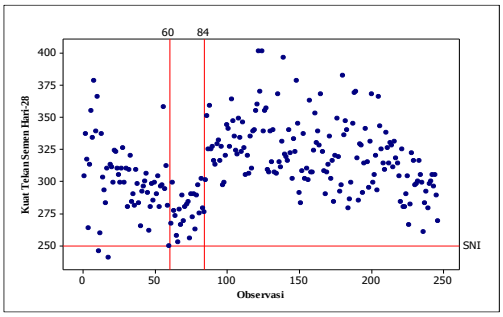
5. Melakukan pemilihan variabel dengan RFE berbasis *Random Forest* untuk menentukan variabel yang akan dianalisis selanjutnya.
6. Melakukan pemodelan dengan metode regresi linier.
   a. Pemodelan dengan regresi linier didahului dengan pemeriksaan hubungan antara semua variabel (prediktor dan respon). Variabel yang digunakan yaitu semua variabel.
   b. Kemudian, melakukan pemodelan regresi linier berganda.
   c. Melakukan pemeriksaan asumsi identik, independen dan distribusi normal. Jika terdapat pelanggaran asumsi, maka dilakukan penaggulangan.
   d. Menghitung nilai kesalahan RMSE sebagai pemilihan model terbaik.
7. Melakukan pemodelan dengan metode SVR.
   a. Melakukan pemodelan dengan SVR menggunakan semua variabel dan beberapa variabel yang terpilih dari langkah 4.
   b. Melakukan *tuning* parameter agar mendapatkan *parameter*.
   c. Mendapatkan model dan menghitung RMSE.
8. Pemilihan model terbaik dengan menggunakan kriteria yang telah dihitung yaitu RMSE.
9. Menarik kesimpulan dan saran.

IV. ANALISIS DAN PEMBAHASAN

Pada penelitian ini, analisis dan pembahasan mencakup beberapa tahap yaitu eksplorasi data, *feature selection* dengan RFE, pembagian data *training* dan *testing* dengan k-*fold cross validation*, analisis regresi linear serta identifikasi asumsinya, kemudian penggunaan metode *support vector regression* serta pemilihan model terbaik. Tahapan prediksi akan dilakukan setelah mendapatkan model terbaiknya.

A. *Eksporasi Data*

Eksplorasi data bertujuan unutk mengetahui karakteristik data secara visual dan secara inferensia. Secara visual dapat ditunjukkan dengan beberapa *tools*. Berikut adalah pola data dari kuat tekan semen pada hari ke 28 dari bulan Juli 2018 hingga Maret 2019.



Gambar 2. Pola Data Kuat Tekan Semen Hari ke-28.

Gambar 2 memberikan visualisasi tentang pola data kuat tekan semen *PCC* pada hari ke-28 di *Mill* 8 pabrik Tuban, bahwa kuat tekan semen memiliki pola yang fluktuatif seiring berjalannya produksi. Terlihat pula bahwa pada observasi ke-60 atau tepatnya tanggal 5 September 2018 hingga observasi ke-84 yaitu 1 Oktober 2018 mengalami penurunan kuat tekan semen, kuat tekan semen PCC mengalami penurunan dibandingkan kuat tekan semen pada bulan lainnya. Berdasarkan acuan SNI 15-2049-2004 tentang semen portland komposit (PCC) bahwa kuat tekan semen PCC pabrik semen Tuban, hanya 3 observasi atau sekitar 1% dari seluruh produksi dari bulan Juli 2018 hingga Maret 2019. Berikut adalah eksplorasi data secara inferensia.

Tabel 2.
Statistika Deskriptif

Variabel	N	SNI	Rata-rata	Min.	Maks.	Median	Varians
Kuat Tekan (Y)	246	250	312,28	241	401	310	852,34

Tabel 2 memberikan informasi bahwa terdapat 246 data kuat tekan semen PCC yang digunakan. Rata-rata kekuatan tekan semen adalah 312,28 kg/cm^2 , dengan nilai varians yang besar dan berbanding lurus dengan Gambar 2 yang mengindikasikan bahwa data kuat tekan semen PCC ini sangat beragam dengan nilai minimum tekan semen adalah 241 kg/cm^2 dan nilai maksimum sebesar 401 kg/cm^2.

B. *K-Fold Cross Validation*

Cross Validation merupakan salah satu teknik pembagian data menjadi *training* dan *testing*. Penelitian kali ini akan menggunakan 10 fold. Dengan jumlah data kuat tekan semen sebanyak 246 observasi, maka setiap *subset* akan memiliki 24 dan 25 observasi. Pengambilan data sebanyak 10 *fold* bersifat *random* dan tidak ada pengulangan dalam metode ini.

C. *Analisis Regresi Linier*

Analisis regresi linear pada penelitian kali ini menggunakan variabel dependen kuat tekan semen PCC pada hari ke-28 dengan variabel independennya sebanyak 11 variabel. Sebelum melakukan pemodelan, perlu diketahui hubungan antara variabel dependen dengan masing-masing variabel independen.

Hasil hubungan antar variabel menggunakan *pearson correlation* yang terletak pada Lampiran 3. memperlihatkan bahwa hasil hubungan antara beberapa variabel independen dengan variabel dependen. Terlihat bahwa hanya 6 variabel independen yang berhubungan dengan variabel dependen, sedangkan variabel lainnya tidak berhubungan. Selain itu, banyak variabel independen yang memiliki korelasi yang kuat dengan variabel independen lain yang menyebabkan tidak terpenuhinya asumsi *non*-multikolinearitas pada model regresi linier. Untuk melihat apakah benar-benar terdapat multikolinearitas didalam penelitian ini, maka akan dilihat nilai *Variance Inflation Factor* (VIF).

Tabel 3.
Variance Inflation Factor

Predictor	VIF
x_1	14,656
x_2	16,098
x_3	2,397
x_4	2,077
x_5	1,448
x_6	2,151
x_7	1,101
x_8	2,573
x_9	4,701
x_{10}	3,057
x_{11}	1,084

Tabel 3 memperlihatkan nilai VIF pada variabel SiO (x_1) dan Al (x_2) melebihi angka 10, yang menunjukkan bahwa terdapat multikolinearitas pada penelitian ini dan yang menyebabkan terjadinya adalah variabel SiO dan Al. Dikarenakan fenomena ini, sehingga penelitian tidak dapat dilanjutkan ke analisis regresi linier.

Terdapat beberapa cara untuk menanggulangi fenomena multikolinearitas, salah satu cara yang termudah adalah dengan mengeluarkan variabel independen yang mempunyai korelasi yang kuat dengan variabel independen lain dan tidak berkorelasi dengan variabel dependen, namun solusi ini mempunyai resiko terbuangnya informasi data. Alternatif lainnya dengan menggunakan *Principal Component Regression* (PCR), dengan tujuan mereduksi variabel.

D. *Principle Component Regression*

PCR merupakan teknik analisis regresi yang dikombinasikan dengan teknik analisis komponen utama (PCA). Model PCR akan dibentuk menggunakan keseluruhan data dengan *complete feature*. Lalu, model PCR akan divalidasi menggunakan 10 *fold cross validation* untuk mendapatkan nilai RMSE.

Sebelum melakukan PCR, akan dicari nilai *eigenvalue* yang dalam analisis faktor terdapat beberapa komponen yang merupakan variabel. Setiap faktor mewakili variabel yang dianalisis. Kemampuan setiap faktor mewakili variabel yang dianalisis ditunjukkan oleh besarnya varians yang dijelaskan, yang disebut *eigenvalue*. Kriteria pemilihan faktor dengan melihat nilai *eigenvalue* yang bernilai lebih besar sama dengan satu. Dalam analisis ini akan dilakukan pemodelan menggunakan data kuat tekan semen PCC Juli 2018 hingga Maret 2019. Berikut adalah hasil analisisnya.

Tabel 4.
Eigenvalue

Faktor	*Eigenvalue*	Proporsi	Komulatif
1	3,4639	0,315	0,315
2	1,8505	0,168	0,483
3	1,5367	0,14	0,623
4	1,0768	0,098	0,721
5	0,9046	0,082	0,803
6	0,8895	0,081	0,884

Faktor	*Eigenvalue*	Proporsi	Komulatif
7	0,5644	0,051	0,935
8	0,2933	0,027	0,962
9	0,2446	0,022	0,984
10	0,1428	0,013	0,997
11	0,0329	0,003	1

Pada Tabel 4 terlihat bahwa nilai *eigenvalue* yang memiliki nilai lebih dari satu sampai di faktor empat, sehingga jumlah faktor yang terbentuk adalah empat dengan faktor yang dengan keempat faktor ini dapat menjelaskan data sebesar 72,1%. Langkah selanjutnya adalah dengan mencari nilai skor komponen yang terbentuk untuk membentuk persamaan regresi komponen utama, berikut adalah hasilnya.

Tabel 5.
Koefisien PCR

Variabel	Faktor 1	Faktor 2	Faktor 3	Faktor 4
Z_1	-0,50068	0,02702	0,04558	0,03216
Z_2	-0,50752	0,09249	0,08563	-0,02254
Z_3	-0,38385	-0,09894	0,11251	0,07979
Z_4	0,08497	-0,50530	0,07712	0,57451
Z_5	0,04033	-0,09476	0,53391	-0,45016
Z_6	0,21251	0,24892	-0,58534	-0,17937
Z_7	-0,00965	0,27268	-0,04212	0,51350
Z_8	-0,29523	0,52017	-0,02928	-0,04613
Z_9	0,41803	0,26134	0,22594	-0,00228
Z_{10}	0,16842	0,38505	0,52348	0,20694
Z_{11}	0,00505	-0,30523	0,01775	-0,34435

Berdasarkan Tabel 5 yang memperlihatkan keempat faktor yang terbentuk dapat dibentuk persamaan regresi sebagai berikut.

$$K_1 = -0{,}50068\,Z_1 - 0{,}50752\,Z_2 - 0{,}38385\,Z_3 + 0{,}08497\,Z_4 + \ldots + 0{,}00505\,Z_{11}$$
$$K_2 = 0{,}02702\,Z_1 + 0{,}09249\,Z_2 - 0{,}09894\,Z_3 - 0{,}50530\,Z_4 + \ldots - 0{,}30523\,Z_{11}$$
$$K_3 = 0{,}04558\,Z_1 + 0{,}08563\,Z_2 + 0{,}11251\,Z_3 + 0{,}07712\,Z_4 + \ldots + 0{,}01775\,Z_{11}$$
$$K_4 = 0{,}03216\,Z_1 - 0{,}02254\,Z_2 + 0{,}07979\,Z_3 + 0{,}57451\,Z_4 + \ldots - 0{,}34435\,Z_{11}$$

Setelah skor komponen terbentuk, langkah selanjutnya adalah menentukan variabel-variabel baru untuk menggantikan variabel independen sebelumnya. Variabel pengganti juga berjumlah empat variabel karena mengikuti faktor yang terbentuk, berikut adalah variabel yang baru.

Tabel 6.
Skor Komponen Utama

i	K_{1i}	K_{2i}	K_{3i}	K_{4i}
1	-4,69357	0,77447	3,72359	-0,95912
2	-5,24951	-3,44305	0,73046	-0,99116
3	-2,03896	-1,56704	0,46048	-0,31848
4	-0,53945	-0,23737	2,40233	2,44069
5	-2,75136	-0,72927	2,37485	0,80409
6	-5,14789	-2,06380	-0,12851	-0,82029
...	...	...	...	...
...	...	...	...	...
244	2,11031	-0,15445	1,22751	-1,19461
245	1,43106	0,90606	1,80560	-1,70655
246	0,38554	0,93918	1,65935	-1,64986

Pada tabel 6 memberikan informasi bahwa terdapat empat variabel baru yaitu nilai skor komponen utama (K) yang didapatkan dari perkalian matrix nilai baku (Z) yang telah di standardisasi dengan matrix *eigenvectors*, yang kemudian di regresikan dengan variabel respon yaitu kuat tekan semen pada hari ke-28. Berikut adalah hasil regresinya.

Tabel 7.
Uji Parsial

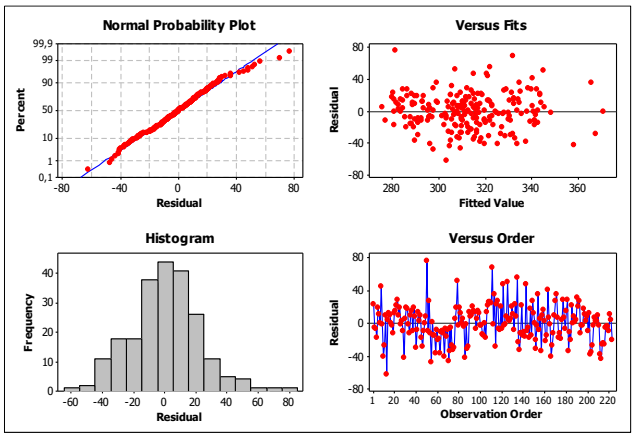
Prediktor	Koefisien	SE Koef	T	P	VIF
Konstan	312,276	1,445	216,18	0,000	
K_1	-3,1877	0,777	-4,10	0,000	1,000
K_2	-4,060	1,064	-3,82	0,000	1,000
K_3	-13,295	1,168	-11,39	0,000	1,000
K_4	-3,062	1,395	-2,20	0,029	1,000

Berdasarkan *output* yang disajikan di Tabel 7 terlihat bahwa nilai VIF disetiap variabel baru memiliki nilai yang kurang dari 10, sehingga fenomena multikolinearitas telah berhasil diatasi.

$$\hat{Y} = 312.28 - 3.18K_1 - 4.06K_2 - 13.29K_3 - 3.06K_4$$

$$\hat{Y} = 312.28 + 0{,}782\,x_1 + 0{,}173\,x_2 - 0{,}115\,x_3 - 1{,}004\,x_4 - 5{,}464\,x_5 + 6{,}643\,x_6 - 2{,}088\,x_7 - 0.640\,x_8 - 5{,}850\,x_9 - 9{,}694\,x_{10} + 2{,}042\,x_{11}$$

Hasil *output* juga memberikan keputusan bahwa setiap variabel independen berpengaruh signifikan terhadap model Kemudian setelah didapatkan model dan hasil regresinya, akan dilakukan pendeteksian asumsi regresi linear, berikut adalah hasilnya.



Gambar 3. Asumsi IIDN.

Dapat dilihat secara visual dari Gambar 3 bahwa pada pendeteksian asumsi kenormalan dapat dilihat dari gambar *normal probability plot* yang terlihat bahwa *plot* mengikuti garis kenormalan (berwarna biru) sehingga secara visual telah mengikuti asumsi distribusi normal, hal ini didukung oleh gambar *histogram* yang membentuk distribusi normal. Gambar *versus fits* menjelaskan bahwa data tidak membentuk suatu pola dan tersebar merata sehingga secara visual dapat diputuskan bahwa data bersifat identik, sedangkan pada gambar *versus order* yang memperlihatkan secara visual bahwa data tersebar merata diatas maupun dibawah angka nol, sehingga secara visual telah memenuhi asumsi independen. Kemudian dilanjutkan dengan validasi model PCR.

Validasi dalam kali ini menggunakan metode 10 *fold cross validation*. Sebanyak 10 *fold* yang terbagi menjadi data *training* dan data *testing*. Berikut adalah hasilnya.

Tabel 8.
Validasi Model PCR

Fold	RMSE
Fold 1	22,58
Fold 2	22,17
Fold 3	20,06
Fold 4	22,09
Fold 5	24,38
Fold 6	16,62
Fold 7	22,71
Fold 8	23,16
Fold 9	26,78
Fold 10	23,39
Rata-Rata	22,39

Tabel 8 menunjukkan hasil perhitungan dari kriteria model terbaik yaitu RMSE menggunakan metode PCR. Indikator rata-rata dari kesepuluh *fold* yang memiliki nilai RMSE untuk model PCR sebesar 22,39, yang kemudian nilai dibandingkan dengan metode lainnya.

E. Support Vector Regression

Model SVR akan dibentuk menggunakan keseluruhan data dengan *complete feature* dan *feature selection* dengan RFE, lalu validasi menggunakan 10 *fold cros validation* untuk mendapatkan nlai kriteria model terbaik yaitu RMSE masing-masing model. Metode yang digunakan dalam SVR kali ini adalah *Kernel-Linear*, *Kernel-Polynomial* dan *Kernel-Radial Basis Function*. Sebelum melakukan analisis SVR, perlu diketahui bahwa dengan RFE berbasis *Random Forest*, variabel dengan *ranking* lima teratas merupakan variabel yang perlu dikeluarkan menurut metode RF-RFE yaitu Blaine, SO_3, LOI, MgO dan S_1O_2.

1. *Kernel-Linear*

Kernel-Linear merupakan sebuah metode yang ketika digunakan, data akan terpisah oleh sebuah garis linier yang disebut *hyperlane*. Berikut adalah hasilnya.

Tabel 9.
Tuning Parameter Kernel-Linear

Feature	C	RMSE
Complete	1	23,510
RF-RFE	1	29,881

Tabel 9 diatas menunjukkan hasil *tuning* parameter menggunakan *kernel-linear complete feature* dan RFE yang hasilnya lebih baik menggunakan *complete* feature yaitu mendapatkan nilai *parameter cost* (C) yaitu senilai 1, dengan nilai kriteria RMSE yaitu sebesar 23,51. Kemudian dilanjutkan dengan fungsi *kernel* selanjutnya yaitu *kernel polynomial*.

2. *Kernel-Polynomial*

Kernel-polynomial merupakan *kernel* yang bersifat *non-linear*. *Kernel-polynomial* memiliki fungsi khusus untuk memetakan ke *feature space* yang biasanya berbentuk kurva parabola. Berikut adalah hasil *tuning parameter* untuk *kernel-polynomial*.

Tabel 10.
Tuning Parameter Kernel-Polynomial

Feature	*Degree*	*Gamma*	*C*	RMSE
Complete	3	0,01	1	22,91489
RF-RFE	1	0,001	0,25	29,7996

Tabel 10 merupakan *tuning parameter* untuk *kernel polynomial complete feature* dan RFE, didapatkan hasil bahwa *complete feature* lebih baik dibandingkan RFE. Terdapat dua *parameter* yang digunakan untuk *complete feature* yaitu *gamma* dan *cost* (C). Dengan kriteria terbaik yaitu RMSE sebesar 22,92. Didapatkan *parameter* yang paling optimal adalah 0,01 (*gamma*) dan 1 (*cost*) dengan nilai *degree* yaitu 3.

3. *Kernel*-RBF

Kernel-RBF (*Radial Basis Function*) merupakan salah satu dari beberapa *kernel* untuk kasus *non-linear*. *Kernel* ini memetakan suatu data ke dimensi yang lebih tinggi dan membentuk kurva yang fleksibel sehingga dapat mengikuti pola data yang digunakan. Berikut ini hasil dari *tuning parameter* dengan metode *kernel*-RBF untuk data kuat tekan semen PCC.

Tabel 11.
Tuning Parameter Kernel-RBF

Feature	*Gamma*	C	RMSE
Complete	0,030096	1	22,03268
RF-RFE	0,423572	0,5	27,26652

Tabel 11 menunjukkan hasil *tuning parameter* untuk SVR dengan *kernel* RBF *complete feature* dan RFE. Hasil yang terbaik adalah RBF dengan *complete* feature. Terdapat dua *parameter* untuk *kernel*-RBF yaitu *gamma* dan *cost* (C). Dengan kriteria RMSE sebesar 22,03, maka *parameter* yang terpilih dan paling optimal untuk data kuat tekan PCC ini adalah *gamma* bernilai 0,030096 dan nilai *cost* sebesar 1.

F. Pemilihan Model Terbaik

Pemilihan model terbaik dalam penelitian ini menggunakan kriteria pemilihan adalah RMSE. Model-model yang telah terbentuk akan dibandingkan dengan kriteria tersebut. Berikut adalah hasil dari ketiga kriteria pemilihan model terbaik.

Tabel 12.
Pemilihan Model Terbaik

Metode		RMSE
PCR		22,39
SVR -	Linear	22,90
Complete	Polynomial	22,48
Feature	RBF	22,19
SVR -	Linear	29,88
Feature	Polynomial	29,79
Selection	RBF	27,27

Tabel 12 memberikan informasi bahwa, dari beberapa model yang terbentuk, kriteria model terbaik adalah pada metode *Support Vector Regression Kernel-Radial Basis Function* dengan *Complete Feature* yang memiliki nilai kriteria RMSE adalah 22,19. Model yang terbentuk adalah sebagai berikut.

$$\hat{f}(x) = \hat{w}_i^T \exp(-0{,}030096 \left\| x - x_i \right\|^2) + \hat{b}$$

Tabel 13.
Koefisien Parameter Model SVR

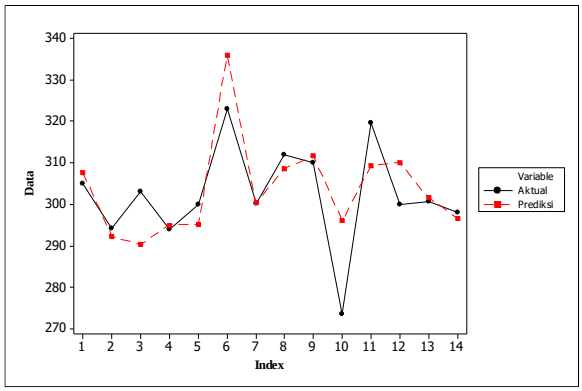
Variabel	$\hat{w}_i^T$
x_1	3,427
x_2	3,266
x_3	-2,751
x_4	0,173
x_5	-6,875
x_6	7,776
x_7	-5,090
x_8	-3,249
x_9	-8,365
x_{10}	-9,747
x_{11}	1,073

Tabel 13 merupakan sebuah nilai pada setiap parameter $\hat{w}_i^T$ yang terdapat pada model, dengan parameter $\hat{b}$ yang menjadi *intercept*. Setelah mendapatkan model makan dilakukan prediksi pada bulan April 2019, berikut adlaah hasilnya.

G. Prediksi Kuat Tekan Semen

Dengan model *kernel-Radial Basis Function* yang terpilih sebagai metode terbaik dalam penelitian ini disertai *parameter*, dilakukan prediksi kuat tekan semen PCC pada hari ke-28 pada bulan April 2019 di *Mill* 8, berikut adalah hasilnya.

Gambar 4 memberikan informasi bahwa, terdapat 14 data kuat tekan semen PCC di bulan April 2019 di *Mill* 8. Secara visual nilai prediksi dengan model SVR *kernel*-RBF telah mengikuti data aktual, namun pada observasi keenam dan kesepuluh yaitu tanggal 6 April 2019 dan 12 April 2019 cukup memiliki perbedaan yang jauh pada besaran kuat tekan semen PCC, dengan prediksi ini didapatkan nilai kesalahan RMSE adalah 8,78.



Gambar 4. Prediksi Kuat Tekan Semen April 2019.

V. KESIMPULAN DAN SARAN

A. *Kesimpulan*

Berdasarkan analisis dan pembahsan yang telah dilakukan, diperoleh kesimpulan sebagai berikut:

1. Pada hasil analisis regresi linier terindikasikan mengalami multikolinearitas, sehingga perlu ditangani menggunakan metode PCR yang mereduksi variabel menjadi empat faktor. Kemudian dilakukan pemodelan dengan metode SVR dengan fungsi *kernel* menghasilkan bahwa, *complete feature* lebih memiliki hasil yang baik dibandingkan *feature selection* (RF-RFE). *Feature selection* tidak cocok untuk digunakan penelitian ini. SVR *complete feature* juga menghasilkan model dan *parameter* yang lebih baik dibandingkan PCR. Metode yang terpilih adalah *Kernel – Radial Basis Function* dengan *complete feature*.
2. Hasil prediksi menunjukkan bahwa dengan model *Kernel*-RBF telah memprediksi kuat tekan semen PCC di bulan April 2019 dengan baik, terlihat secara visual bahwa secara menyeluruh telah memprediksi dengan baik.

B. *Saran*

Berdasarkan kesimpulan, maka saran yang dapat diberikan untuk penelitian selanjutnya adalah peneliti dapat mengambil informasi lebih banyak dalam pengaruh kuat tekan semen, karena variabel-variabel yang digunakan dalam penelitian ini tidak sepenuhnya menggambarkan kuat tekan semen secara keseluruhan, diduga variabel-variabel seperti manusia, *setting* mesin, cuaca dan sebagainya juga turut mempengaruhi kuat tekan semen dan diharapkan pengambilan data dengan periode yang lebih panjang, agar pemodelan dan prediksi lebih akurat.

DAFTAR PUSTAKA

[1] SNI 15-2049-2004 Semen Portland

[2] Alfarisi. *Data Preprocessing - Konsep Pembelajaran Data Mining*. *Steetmit.com*. [Dikutip: 9 Desember 2018.] https://steetmit.com/education/@alfarisi/data-preprocessing-konsep-pembelajaran-datamining?sort=new#comments, 2017.

[3] Casasent D, Nakariyakul S. 2004. *Hyperspectral Ratio Feature Selection: Agricultural Product Inspection Example.* Nondestructive Sensing for Food Safety, Quality, and Natural Resources 5587, Proceedings of SPIE; Philadelphia, 26 Oktober 2004. hlm 133-143.

[4] Breiman, Leo. 2001. *Random Forests*. *Machine Learning* 45:5-32.

[5] Drapper, N., H. Smith. 1992. *Analisis Regresi Terapan Edisi Kedua*. Terjemahan oleh Bambang Sumantri. Gramedia Pustaka Utama, Jakarta.

[6] Sudjana. 2005. *Metode Statistika*. Bandung : Tarsito.

[7] Daniel, W.W. 1989. *Statistika Non Parametrik Terapan*. Jakarta: PT. Gramedia.

[8] Johnson R. Dan D. Wichern. 2010. *Applied Multivariate Statistical Analysis*. Sixth Edition, Prentince Hall. New Jersey.

[9] Scholkopf B., and Smola A. 2002. *Learning With Kernel*. MIT Press.

[10] Abe, S. 2005. *Support Vector Machine for Pattern Classification*. Springer - Verlag. London Limited.

[11] Santosa, B. 2007. *Data Mining Teknik Pemanfaatan Data untuk Keperluan Bisnis*. Graha Ilmu: Yogyakarta.

[12] Gooijer, Jan G. De dan Hyndman, Rob J. (2006). *25 Years Of Time Series Forecasting*. International Journal of Forcasting vol. 22, no. 443-473.

[13] ASTM Standards, 2002, ASTM C 109/C 109M – 02. *Standard Test Method for Compressive Strength of Hydraulic Cement Mortars (Using2-in. or 50- mm Cube Specimens)*. ASTM International, West Conshohocken, PA.