Improved nu-Support Vector Regression Model Based on Variable Selection and Brain Storm Optimization for Stock Price Forecasting

August 30,2019

Stock Price Prediction

- The act of trying to determine the future value of a company stock or other financial instrument traded on an exchange.
- The successful prediction of a stock's future price could yield significant profit.
- The efficient-market hypothesis suggests that stock prices reflect all currently available information and any price changes that are not based on newly revealed information thus are inherently unpredictable. Others disagree and those with this viewpoint possess myriad methods and technologies which purportedly allow them to gain future price information.

Overview

- Big data mining, analysis and forecasting always play a vital role in modern economic and industrial fields, and selecting an optimization model to improve time series' forecasting accuracy is challenging.
- A hybrid nu-Support Vector Regression(nu-SVR) model is developed by combining with principal component analysis (PCA) and brain storm optimization (BSO) for stock price index forecasting.
- Correlation analysis and PCA are conducted initially to select the input variables of the nu-SVR from 17 technical indicators, while the advanced BSO algorithm is used to search for optimal parameters of nu-SVR.

Understanding the problem

Input variable selection process and extracts the most representative information from an original high-dimensional dataset. forecasting model developed in this research.

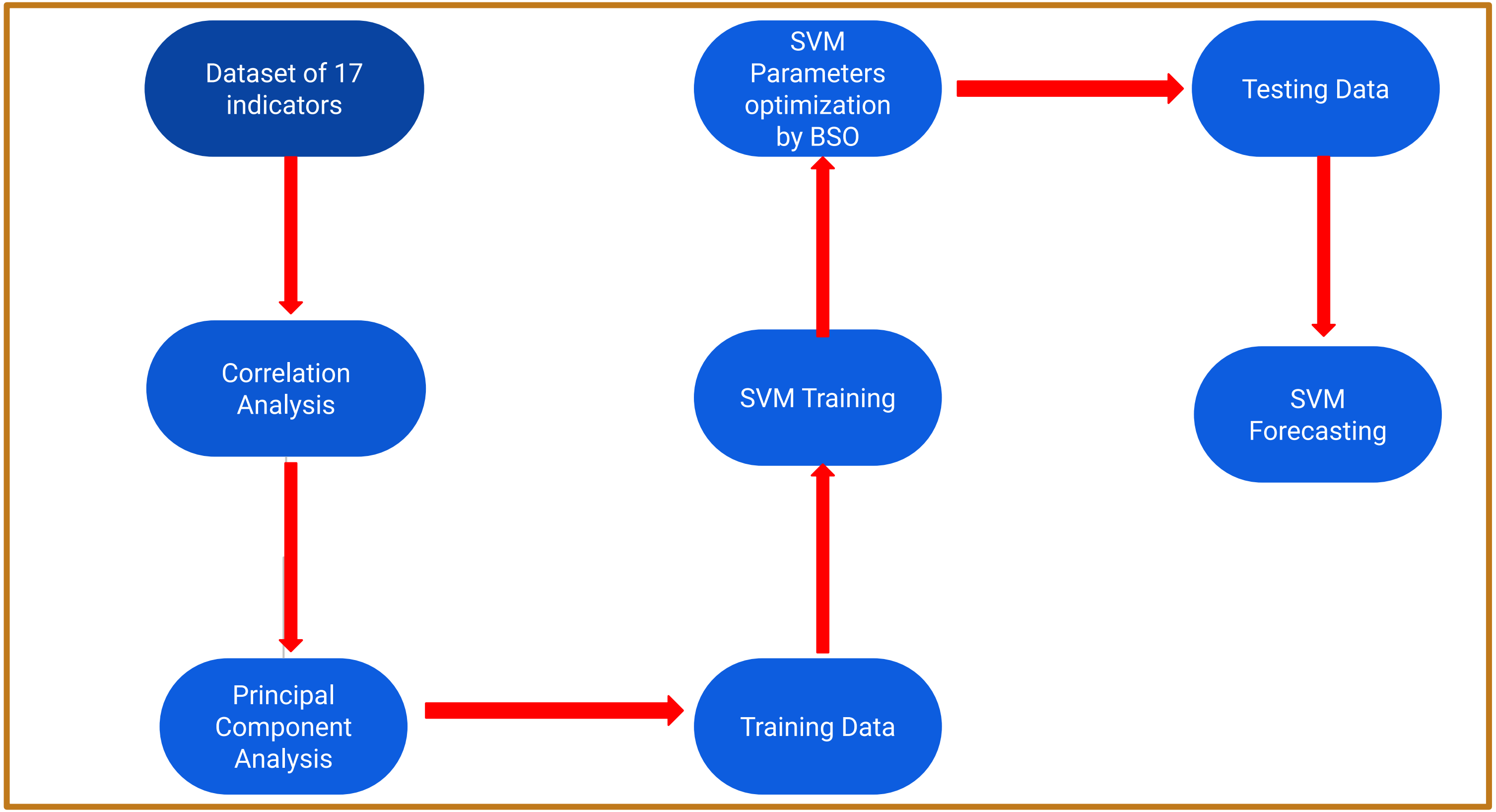
The proposed hybrid forecast strategy is compared with other existing nu-SVR related approaches in the literature, such as nu-SVR with default parameters , nu-SVR with Grid Search.

Lastly,to test the model's performance, a case studies concerning the Chinese stock market are conducted, and the simulation results will be discussed later on.

Why SVR?

- Support vector regression is a useful and powerful machine learning technique to recognize pattern of time series dataset.
- Conventional modeling techniques, inadequate for stock market price forecasting.
- Tackle the stock market fluctuations and yielding satisfactory forecasting precision.
- SVR has a global optimum and exhibits better prediction accuracy. It considers both the training error and the capacity of the regression model to avoid underfitting and overfitting problems in the training process.

Graphical Abstract



Dataset of 17 indicators

Correlation Analysis

Principal Component Analysis

Training Data

SVM Training

SVM Parameters optimization by BSO

Testing Data

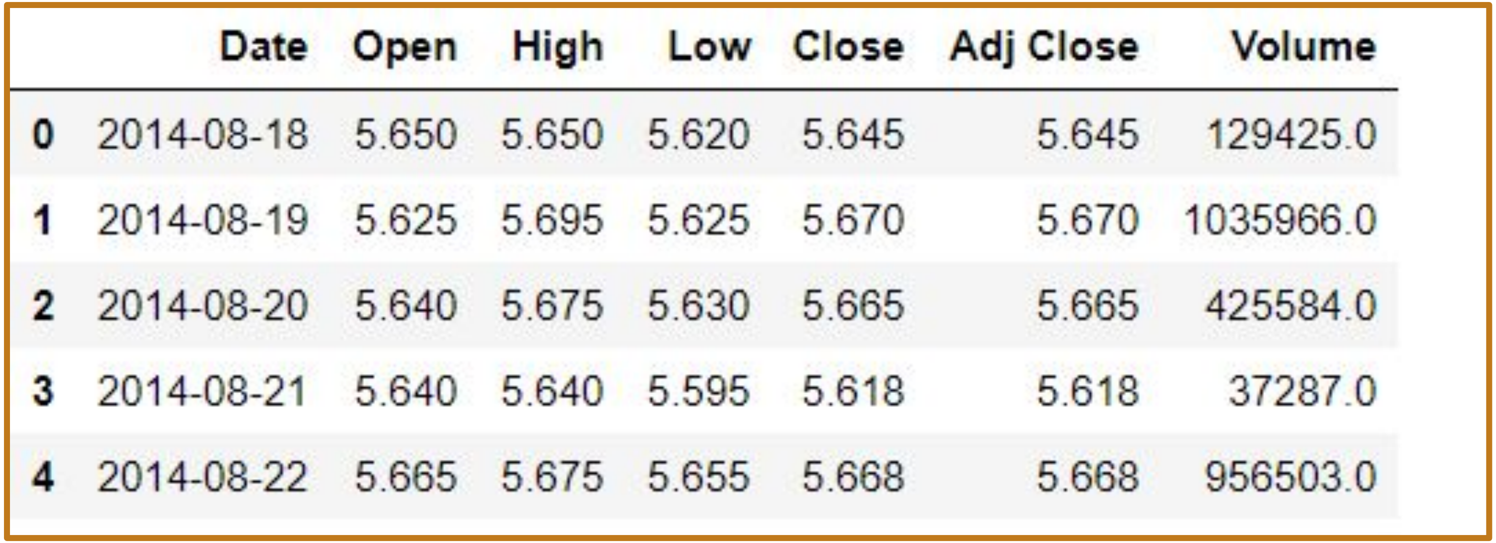
SVM Forecasting

Technical Indicators

- Technical indicators are heuristic or mathematical calculations based on the price, volume, or open interest of a security or contract used by traders who follow technical analysis.
- Technical analysts or chartists look for technical indicators in historical asset price data in order to judge entry and exit points for trades.
- Example : Moving Averages, Relative Strength Index(RSI) , Moving average convergence-divergence (MACD).

Original DataSet

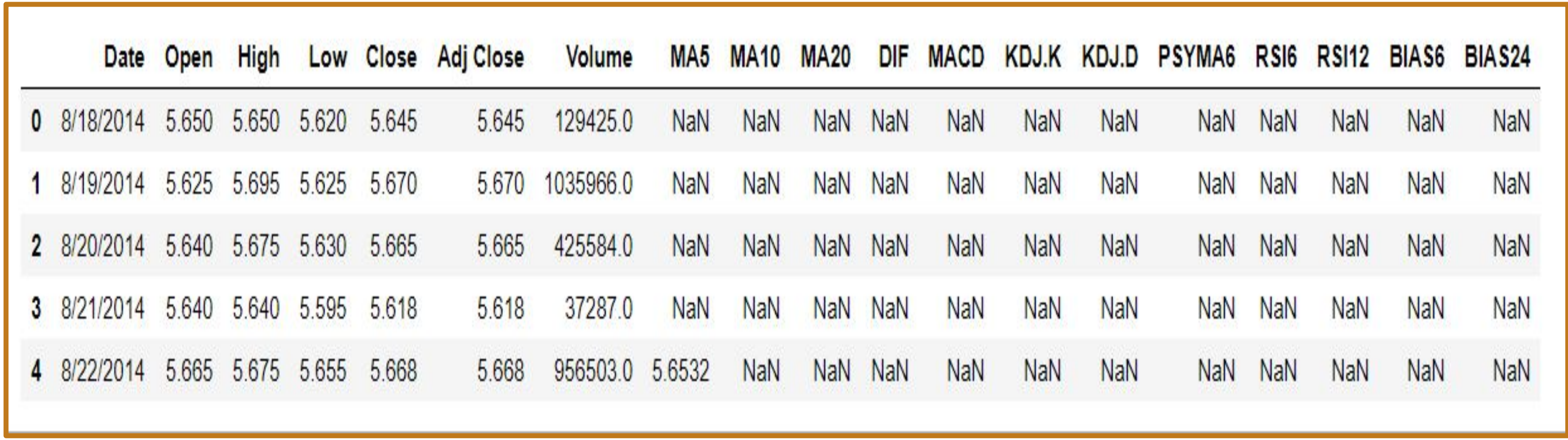
CSI300



	Date	Open	High	Low	Close	Adj Close	Volume
0	2014-08-18	5.650	5.650	5.620	5.645	5.645	129425.0
1	2014-08-19	5.625	5.695	5.625	5.670	5.670	1035966.0
2	2014-08-20	5.640	5.675	5.630	5.665	5.665	425584.0
3	2014-08-21	5.640	5.640	5.595	5.618	5.618	37287.0
4	2014-08-22	5.665	5.675	5.655	5.668	5.668	956503.0

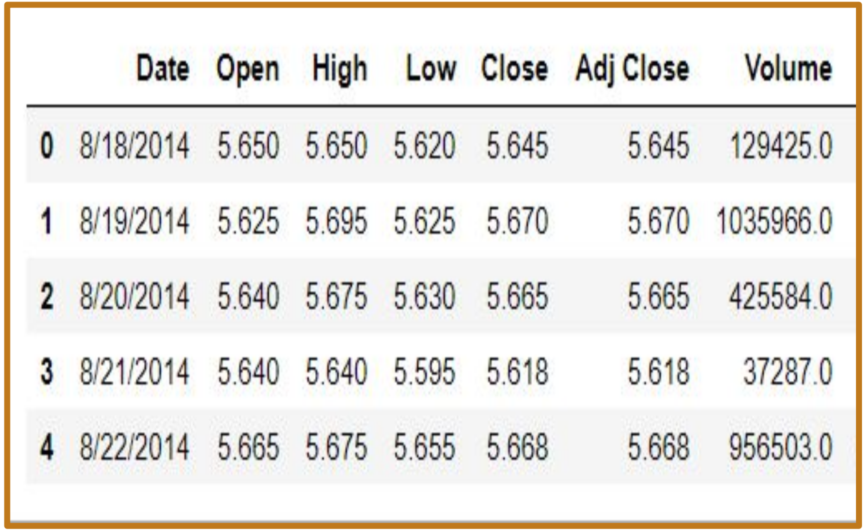
Prepared Dataset

CSI300



	Date	Open	High	Low	Close	Adj Close	Volume	MA5	MA10	MA20	DIF	MACD	KDJ.K	KDJ.D	PSYMA6	RSI6	RSI12	BIAS6	BIAS24
0	8/18/2014	5.650	5.650	5.620	5.645	5.645	129425.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	8/19/2014	5.625	5.695	5.625	5.670	5.670	1035966.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	8/20/2014	5.640	5.675	5.630	5.665	5.665	425584.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3	8/21/2014	5.640	5.640	5.595	5.618	5.618	37287.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
4	8/22/2014	5.665	5.675	5.655	5.668	5.668	956503.0	5.6532	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Indicators



	Date	Open	High	Low	Close	Adj Close	Volume
0	8/18/2014	5.650	5.650	5.620	5.645	5.645	129425.0
1	8/19/2014	5.625	5.695	5.625	5.670	5.670	1035966.0
2	8/20/2014	5.640	5.675	5.630	5.665	5.665	425584.0
3	8/21/2014	5.640	5.640	5.595	5.618	5.618	37287.0
4	8/22/2014	5.665	5.675	5.655	5.668	5.668	956503.0

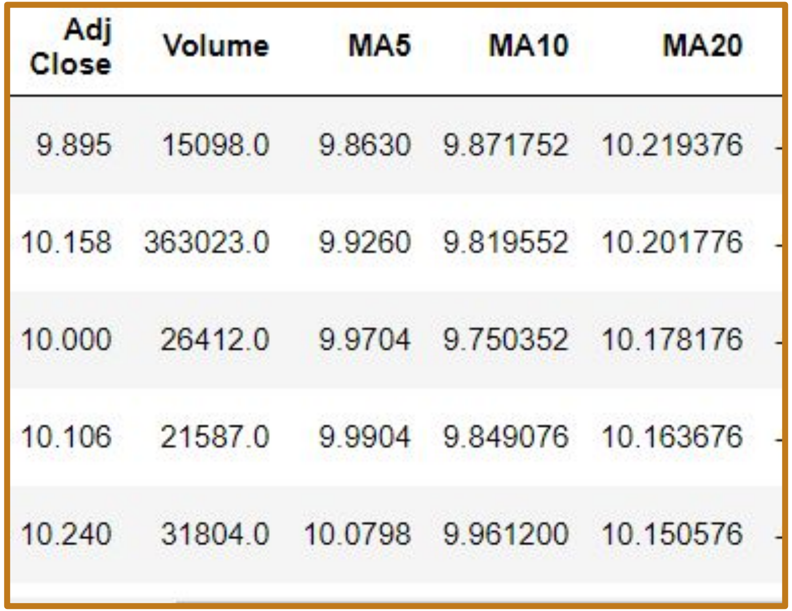
- **Open** : Opening price of the current day
- **High :** Highest price of the current day
- **Low :** Lowest price of the current day
- **Close :** Closing Price of the current day
- **Volume :** Measures the number of a stock's shares that are traded on a stock exchange in a day.

Contd.

Adj Close (Target Variable): The adjusted closing price factors in anything that might affect the stock price after the market closes . It is considered to be the true price of that stock and is often used when examining historical returns or performing a detailed analysis of historical returns.

MA (MA5,MA10,MA20): The simple moving average calculates the arithmetic mean of a security over a number (n) of time periods, A.

$$A = (A_1+A_2+...+A_n)/n$$



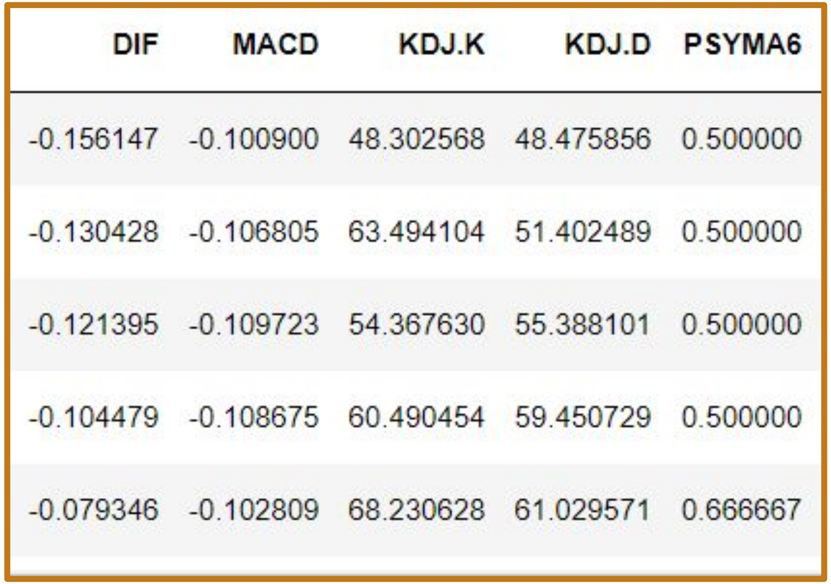
Adj Close	Volume	MA5	MA10	MA20
9.895	15098.0	9.8630	9.871752	10.219376
10.158	363023.0	9.9260	9.819552	10.201776
10.000	26412.0	9.9704	9.750352	10.178176
10.106	21587.0	9.9904	9.849076	10.163676
10.240	31804.0	10.0798	9.961200	10.150576

Contd.

DIF : EMA$_{12}$ - EMA$_{26}$

where ,
EMA is Exponential Moving Average (weighted moving average that gives importance more to the recent data)

MACD : Moving Average Convergence Divergence. It is a trend following Momentum indicator that shows the relationship between two moving averages of a security’s price. It is calculated as 9- period moving Average of DIF.



DIF	MACD	KDJ.K	KDJ.D	PSYMA6
-0.156147	-0.100900	48.302568	48.475856	0.500000
-0.130428	-0.106805	63.494104	51.402489	0.500000
-0.121395	-0.109723	54.367630	55.388101	0.500000
-0.104479	-0.108675	60.490454	59.450729	0.500000
-0.079346	-0.102809	68.230628	61.029571	0.666667

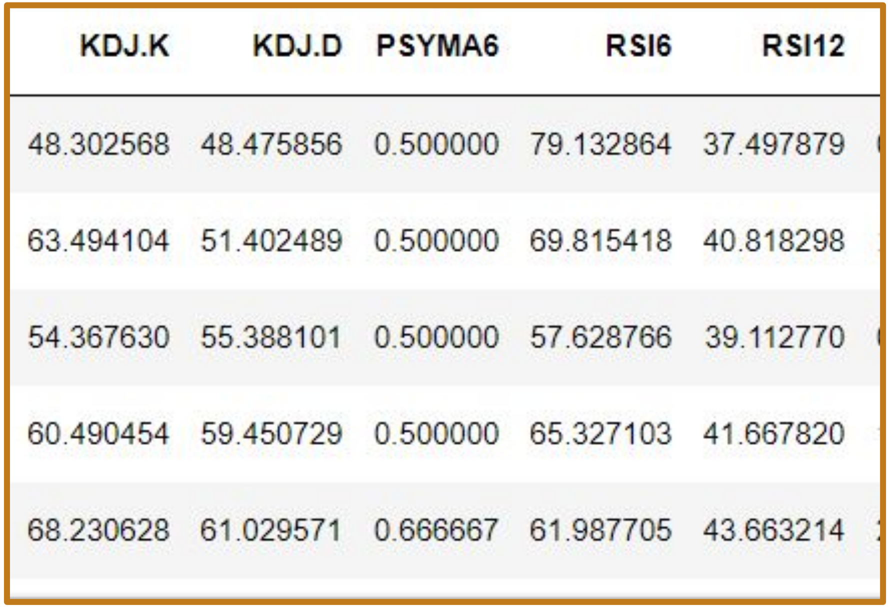
Contd.

KDJ: A stochastic oscillator is a momentum indicator comparing a particular closing price of a security to a range of its prices over a certain period of time.

KDJ.K : $(C-L_{14})/(H_{14}-L_{14})*100$

KDJ.D : 3-period moving average of KDJ.K

PSYMA6 : Psychological line is defined as ratio of number of rising periods to the total number of periods. Here number of periods are 6.



KDJ.K	KDJ.D	PSYMA6	RSI6	RSI12
48.302568	48.475856	0.500000	79.132864	37.497879
63.494104	51.402489	0.500000	69.815418	40.818298
54.367630	55.388101	0.500000	57.628766	39.112770
60.490454	59.450729	0.500000	65.327103	41.667820
68.230628	61.029571	0.666667	61.987705	43.663214

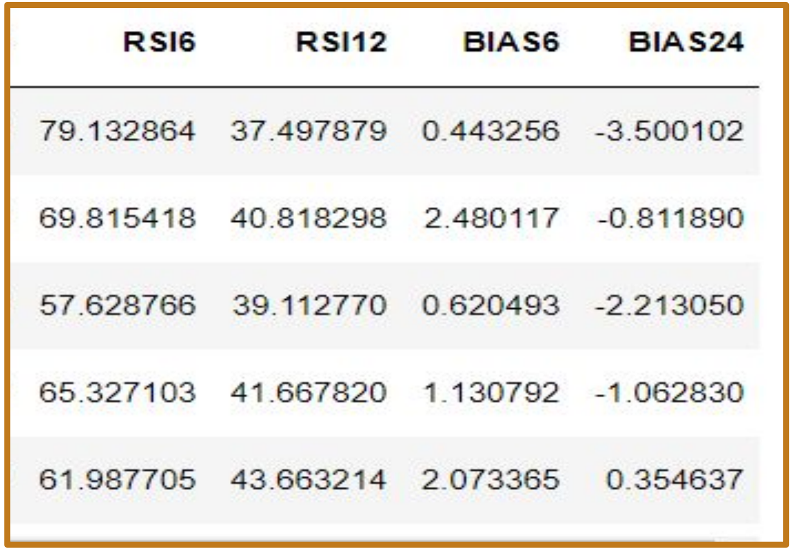
Contd.

RSI(RSI6 , RSI12) : Relative Strength Index.It is a momentum indicator that measures the magnitude of recent price changes to evaluate overbought or oversold conditions in the price of a stock or other asset.It compares recent gains to the recent losses.

RSI = 100 – [100 / (1 + (Average of Upward Price Change / Average of Downward Price Change))]

BIAS(BIAS6, BIAS24) : It is a difference between closing price and moving average line.It indicates the nature of market returning back to average line.

$$BIAS_n = (Close - MA_n)/MA_n * 100$$

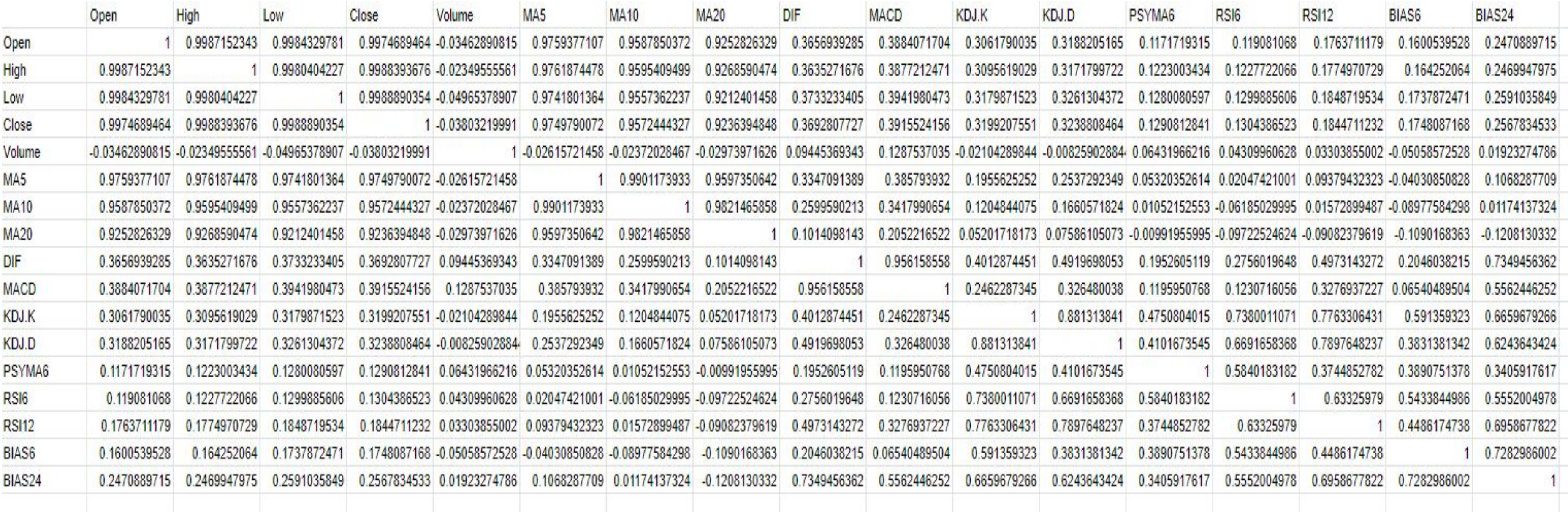


RSI6	RSI12	BIAS6	BIAS24
79.132864	37.497879	0.443256	-3.500102
69.815418	40.818298	2.480117	-0.811890
57.628766	39.112770	0.620493	-2.213050
65.327103	41.667820	1.130792	-1.062830
61.987705	43.663214	2.073365	0.354637

Correlation Analysis

- Correlation refers to how close two variables are to having a linear relationship with each other.
- Features with high correlation are more linearly dependent and hence have almost the same effect on the independent variable. So, when two features have high correlation, we can drop one of the two features.

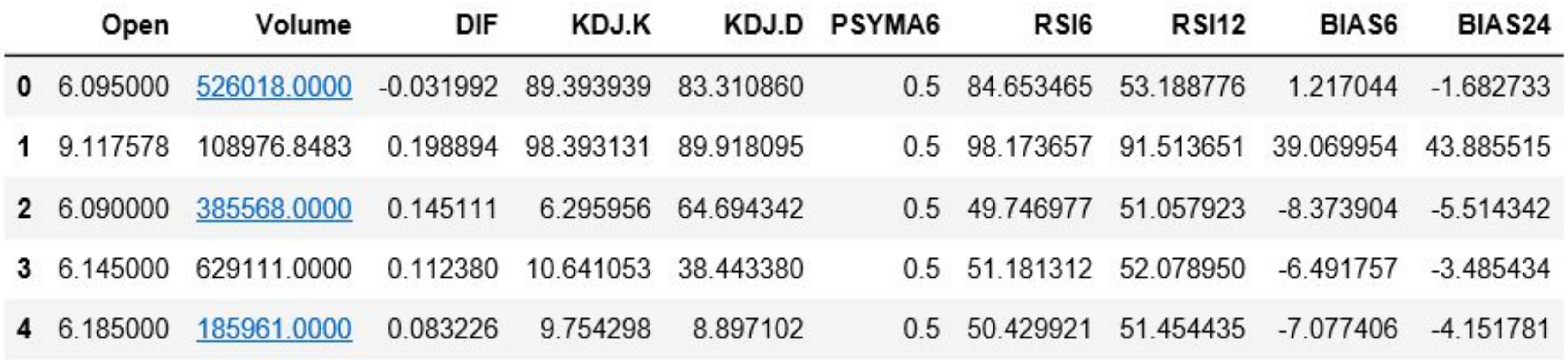
Correlation Matrix



	Open	High	Low	Close	Volume	MA5	MA10	MA20	DIF	MACD	KDJ.K	KDJ.D	PSYMA6	RSI6	RSI12	BIAS6	BIAS24
Open	1	0.9987152343	0.9984329781	0.9974689464	-0.03462890815	0.9759377107	0.9587850372	0.9252826329	0.3656939285	0.3884071704	0.3061790035	0.3188205165	0.1171719315	0.119081068	0.1763711179	0.1600539528	0.2470889715
High	0.9987152343	1	0.9980404227	0.9988393676	-0.02349555561	0.9761874478	0.9595409499	0.9268590474	0.3635271676	0.3877212471	0.3095619029	0.3171799722	0.1223003434	0.1227722066	0.1774970729	0.164252064	0.2469947975
Low	0.9984329781	0.9980404227	1	0.9988890354	-0.04965378907	0.9741801364	0.9557362237	0.9212401458	0.3733233405	0.3941980473	0.3179871523	0.3261304372	0.1280080597	0.1299885606	0.1848719534	0.1737872471	0.2591035849
Close	0.9974689464	0.9988393676	0.9988890354	1	-0.03803219991	0.9749790072	0.9572444327	0.9236394848	0.3692807727	0.3915524156	0.3199207551	0.3238808464	0.1290812841	0.1304386523	0.1844711232	0.1748087168	0.2567834533
Volume	-0.03462890815	-0.02349555561	-0.04965378907	-0.03803219991	1	-0.02615721458	-0.02372028467	-0.02973971626	0.09445369343	0.1287537035	-0.02104289844	-0.00825902884	0.06431966216	0.04309960628	0.03303855002	-0.05058572528	0.01923274786
MA5	0.9759377107	0.9761874478	0.9741801364	0.9749790072	-0.02615721458	1	0.9901173933	0.9597350642	0.3347091389	0.385793932	0.1955625252	0.2537292349	0.05320352614	0.02047421001	0.09379432323	-0.04030850828	0.1068287709
MA10	0.9587850372	0.9595409499	0.9557362237	0.9572444327	-0.02372028467	0.9901173933	1	0.9821465858	0.2599590213	0.3417990654	0.1204844075	0.1660571824	0.01052152553	-0.06185029995	0.01572899487	-0.08977584298	0.01174137324
MA20	0.9252826329	0.9268590474	0.9212401458	0.9236394848	-0.02973971626	0.9597350642	0.9821465858	1	0.1014098143	0.2052216522	0.05201718173	0.07586105073	-0.00991955995	-0.09722524624	-0.09082379619	-0.1090168363	-0.1208130332
DIF	0.3656939285	0.3635271676	0.3733233405	0.3692807727	0.09445369343	0.3347091389	0.2599590213	0.1014098143	1	0.956158558	0.4012874451	0.4919698053	0.1952605119	0.2756019648	0.4973143272	0.2046038215	0.7349456362
MACD	0.3884071704	0.3877212471	0.3941980473	0.3915524156	0.1287537035	0.385793932	0.3417990654	0.2052216522	0.956158558	1	0.2462287345	0.326480038	0.1195950768	0.1230716056	0.3276937227	0.06540489504	0.5562446252
KDJ.K	0.3061790035	0.3095619029	0.3179871523	0.3199207551	-0.02104289844	0.1955625252	0.1204844075	0.05201718173	0.4012874451	0.2462287345	1	0.881313841	0.4750804015	0.7380011071	0.7763306431	0.591359323	0.6659679266
KDJ.D	0.3188205165	0.3171799722	0.3261304372	0.3238808464	-0.00825902884	0.2537292349	0.1660571824	0.07586105073	0.4919698053	0.326480038	0.881313841	1	0.4101673545	0.6691658368	0.7897648237	0.3831381342	0.6243643424
PSYMA6	0.1171719315	0.1223003434	0.1280080597	0.1290812841	0.06431966216	0.05320352614	0.01052152553	-0.00991955995	0.1952605119	0.1195950768	0.4750804015	0.4101673545	1	0.5840183182	0.3744852782	0.3890751378	0.3405917617
RSI6	0.119081068	0.1227722066	0.1299885606	0.1304386523	0.04309960628	0.02047421001	-0.06185029995	-0.09722524624	0.2756019648	0.1230716056	0.7380011071	0.6691658368	0.5840183182	1	0.63325979	0.5433844986	0.5552004978
RSI12	0.1763711179	0.1774970729	0.1848719534	0.1844711232	0.03303855002	0.09379432323	0.01572899487	-0.09082379619	0.4973143272	0.3276937227	0.7763306431	0.7897648237	0.3744852782	0.63325979	1	0.4486174738	0.6958677822
BIAS6	0.1600539528	0.164252064	0.1737872471	0.1748087168	-0.05058572528	-0.04030850828	-0.08977584298	-0.1090168363	0.2046038215	0.06540489504	0.591359323	0.3831381342	0.3890751378	0.5433844986	0.4486174738	1	0.7282986002
BIAS24	0.2470889715	0.2469947975	0.2591035849	0.2567834533	0.01923274786	0.1068287709	0.01174137324	-0.1208130332	0.7349456362	0.5562446252	0.6659679266	0.6243643424	0.3405917617	0.5552004978	0.6958677822	0.7282986002	1

Correlation Analysis of our data

- We compare the correlation between features and remove one of two features that have a correlation higher than 0.9.
- After removing features, we obtain :

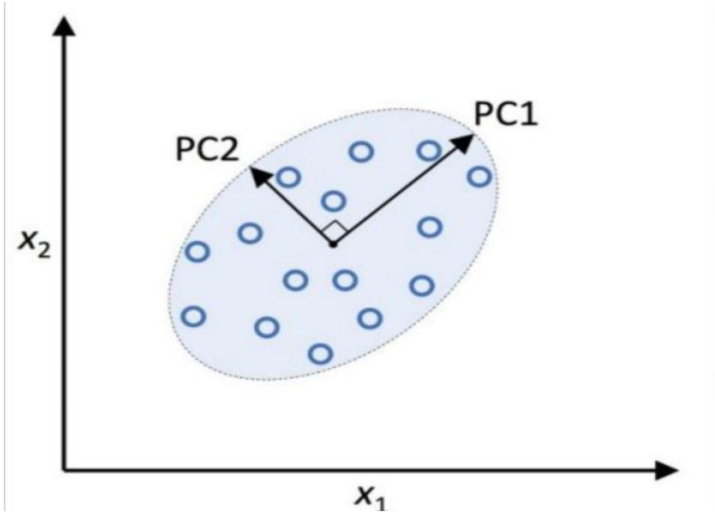


	Open	Volume	DIF	KDJ.K	KDJ.D	PSYMA6	RSI6	RSI12	BIAS6	BIAS24
0	6.095000	526018.0000	-0.031992	89.393939	83.310860	0.5	84.653465	53.188776	1.217044	-1.682733
1	9.117578	108976.8483	0.198894	98.393131	89.918095	0.5	98.173657	91.513651	39.069954	43.885515
2	6.090000	385568.0000	0.145111	6.295956	64.694342	0.5	49.746977	51.057923	-8.373904	-5.514342
3	6.145000	629111.0000	0.112380	10.641053	38.443380	0.5	51.181312	52.078950	-6.491757	-3.485434
4	6.185000	185961.0000	0.083226	9.754298	8.897102	0.5	50.429921	51.454435	-7.077406	-4.151781

Removed Features : High , Low , Close , MA5 , MA10 , MA20 , MACD

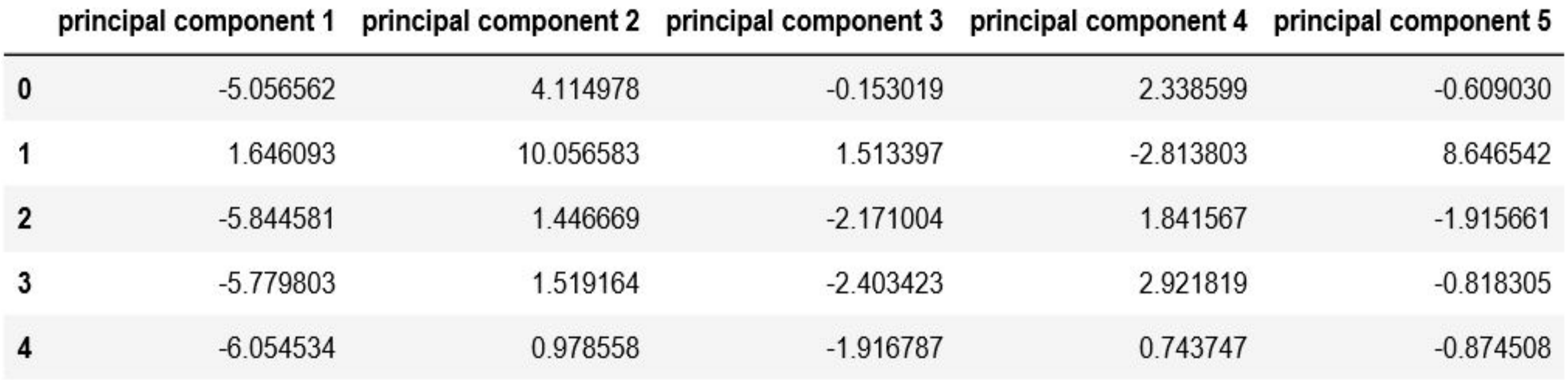
PCA

- PCA stands for Principal Component Analysis.
- It is a statistical procedure that converts a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called **principal components**.
- It helps us to identify patterns in data based on the correlation between features i.e.it aims to find the directions of maximum variance in high-dimensional data and projects it onto a new subspace with equal or fewer dimensions than the original one.
- The Principal Components are orthogonal to each other.



PCA of our data

After performing correlation analysis on our dataset, we obtained a new dataset with less related features. To this new dataset we applied PCA to obtain the top 5 principal components.



	principal component 1	principal component 2	principal component 3	principal component 4	principal component 5
0	-5.056562	4.114978	-0.153019	2.338599	-0.609030
1	1.646093	10.056583	1.513397	-2.813803	8.646542
2	-5.844581	1.446669	-2.171004	1.841567	-1.915661
3	-5.779803	1.519164	-2.403423	2.921819	-0.818305
4	-6.054534	0.978558	-1.916787	0.743747	-0.874508

Support Vector Machine

- SVM (Support-Vector Machines) are supervised learning models that analyze data used for classification and regression analysis.
- A SVM constructs a hyperplane or set of hyperplanes in a high- or infinite-dimensional space, which can be used for classification,regression, or other tasks like outliers detection.
- When SVM is used for regression it is called Support Vector Regression (SVR).
- The model produced by SVR depends only on a subset of the training data, because the cost function for building the model ignores any training data close to the model prediction. Here, we are using SVR to predict the Adjusted Closing Price.

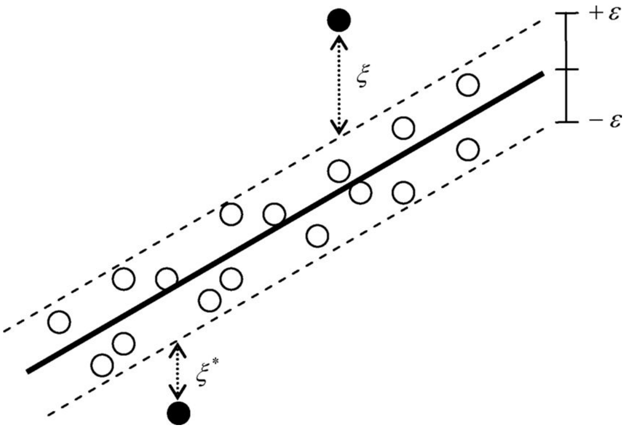
ε-Support Vector Regression`

- ε - SVR is used for working with continuous values.
- In SVR we try to fit the error within a certain threshold (unlike simple regression where we try to minimize error).
- In other words , our basic objective is to basically consider the points that are within the boundary line. And the best fit line is the line hyperplane that has maximum number of points.
- ε - loss function is insensitive with respect to the error as long as the error is less than ε .

- Suppose our Boundary Line is at a distance of ε i.e. the lines are at '+ε' and -'ε' from the hyperplane.
- Suppose the hyperplane is a straight line going through the Y axis.
  - Equation of hyperplane : Wx + b=0
- So the equation of boundary lines are :

  Wx + b= +ε and Wx + b= -ε respectively
- So for any linear hyperplane the equation that satisfy SVR is:

  ε ≤ y - Wx - b ≤ + ε stating y=Wx + b => y - Wx + b = 0



υ(nu)-SVR

υ(nu)-SVR automatically minimizes the ε- insensitive loss function, and causes that the SVR formulation changes by using a new υ parameter whose value is between the [0,1] interval.

The υ(nu) parameter is used to control the number of support vectors, it implies that the υ(nu)-SVR allows data compression and generalizes the prediction error bounds.

υ(nu) gives an upper bound on the fraction of training errors and a lower bound of the fraction of support vectors.

Let the training data be $\{(x_1,y_1),...,(x_l,y_l)$ such that $x_i \epsilon R^n$,i=1,2,...,l and $y_i \epsilon R$. For best estimation we minimize $|| \omega^2 ||$under the precision of ε .

$$\text{Min} \; : \; ½ \; || \, \omega^{2} ||$$

Subject to constraints :

$$Y_i - \omega \, x_i - b <= \varepsilon$$

$$\omega \, x_i + b - Y_i <= \varepsilon$$

The primal problem of v-SVR is:

$$min\ \frac{1}{2}\omega^{T}\cdot\omega + C\left(v\cdot\varepsilon + \frac{1}{l}\sum_{i=1}^{l}(\xi_{i}+\xi_{i}^{*})\right),$$

$$s.t.\begin{cases}(\omega\cdot\varphi(x_{i})+b)-y_{i}\leq\varepsilon+\xi_{i}, \\ y_{i}-(\omega\cdot\varphi(x_{i})+b)\leq\varepsilon+\xi_{i}^{*}, \\ \varepsilon\geq 0, \xi_{i}\geq 0, \xi_{i}^{*}\geq 0.\end{cases}$$

Where $\phi(x_i)$ maps x_i into a higher-dimensional space.

C is regularisation parameter.

v is an upper bound on the fraction of training errors and a lower bound of the fraction of support vectors

ξ_i ξ_i^* are slack variables

Brain Storm Optimization : The Heart of Hybrid Model

While trying to understand brain storm optimization, a storm may be created in your brain.

Brain Storm Optimization : The Heart of Hybrid Model

Brain storm optimization algorithm is a new and promising swarm intelligence algorithm, which simulates the human brainstorming process.

Here we will use this to optimize the parameters **C** & ***v(nu)*** of SVR Model.

Algorithm

```
Initialization: Randomly generate N potential solutions (individuals).

while not reached the predetermined maximum number of iterations
do
    Clustering: Cluster N individuals into M clusters.
    New individual generation: randomly select one or two cluster(s) to generate new individual.
    Selection: The newly generated individual is compared with the existing individual, the better one is kept.
```

Clustering

```
**Clustering:** Cluster $N$ individuals into $M$ clusters by k-means clustering algorithm.
Rank individuals in each cluster and record the best individual as its cluster center in each cluster.
Randomly generate a value $R^{clustering}$ in the range [0, 1).
**if** the value $R^{clustering}$ is smaller than a predetermined probability $P^{clustering}$ **then:**
    Randomly select a cluster center.
    Randomly generate an individual to replace the selected cluster center.
```

Combine two individual

$$X = R * X_1 + (1 - R) * X_2$$

R - random value within (0,1)

New individual generation

$$x_{\text{new}}^{i} = x_{\text{old}}^{i} + \xi(t) \times \text{rand}()$$

$$\xi(t) = \text{logsig}\left(\frac{0.5 \times T - t}{k}\right) \times \text{rand}()$$

x_{new}^{i} x_{old}^{i} - are the i_{th} dimension of x_{new} and x_{old}

$\text{rand}()$ - generate uniformly distributed random numbers in the range [0, 1)

T - max no. of iterations

t - current iteration

k - coefficient to change logsig() function’s slope

New individual generation

New individual generation: randomly select one or two cluster(s) to generate new individual.

```
Randomly generate a value $R^{generation}$ in the range [0, 1).
if the value $R^{generation}$ is less than a probability $P^{generation}$ then:
    Randomly select a cluster, and generate a random value $R^{onecluster}$ in the range [0, 1).
    if the value $R^{onecluster}$is smaller than a predetermined probability $P^{onecluster}$ then:
        Select the cluster center and add random values to it to generate new individual.
```

New individual generation

```
    else:
        Randomly select a individual from this cluster and add
        random value to the individual to generate new individual.
else:
    Randomly select two clusters to generate new individual.
    Generate a random value $R^{twocluster}$ in the range [0, 1).
    if the value $R^{twocluster}$ is less than a predetermined probability
    $P^{twocluster}$ then:
        The two cluster centers are combined and then added with
        random values to generate new individual.
```

New individual generation

else:

Two individuals from each selected cluster are randomly selected to be combined and added with random values to generate new individual.

The newly generated individual is compared with the existing individual, the better one is kept and recorded as the new individual.

Comparison (on data without PCA)

N, M	Parameters (C, v(nu))	MSE
10, 3	4.991334013932322, 0.0921352710860659	1.575584710367192
20, 5	9.61886154564763, 0.36053403613909435	1.586625157707029
100, 5	3.176187078668917, 0.10458726757454902	1.578113048092591
100, 8	4.523995631637805, 0.11309771359861912	1.575494627693644
100, 10	3.261952930497494, 0.10581950203594881	1.577586695244474
-	1, 0.5 (default parameters of NuSVR)	1.596384053712726

Comparison (on data with PCA)

N, M	Parameters (C, v(nu))	R2 Score	MSE
10, 3	7.530235833029731, 0.9088633624406102	0.93742277230554	0.104833535215573
20, 5	6.5229372321284895, 0.9123245773592755	0.93740346822209	0.104865874700510
100, 5	5.862091352846379, 0.8704365799098428	0.93778811578701	0.104221487508427
100, 8	5.82667970562841, 0.9062638174511335	0.93743215861750	0.104817810641285
100, 10	6.761216940704663, 0.8703308301588357	0.93777881780896	0.104237064099865
-	1, 0.5 (default parameters of NuSVR)	0.82214737650813	0.297950547746450

Summary

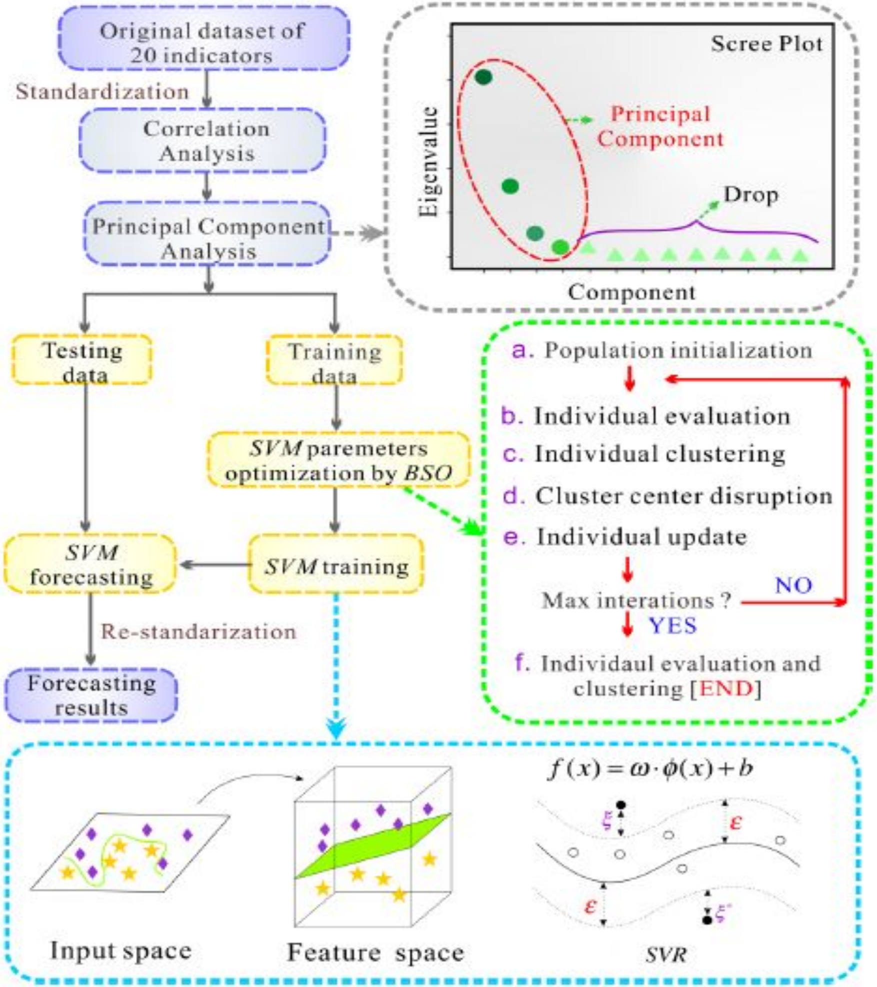


Fig. 1 The flowchart of the proposed hybrid model

Conclusion

- From the previous Tables , we can conclude that the results with without PCA are not that much good .
- Even with BSO, we can optimize upto a certain limit.
- Now with PCA results are better but the there us a certain threshold for N , M after which the reduction in error ceases.
- Data preprocessing techniques such as correlation analysis and PCA are employed to select appropriate variables for a primitive SVR model. Additionally, BSO is for the first time integrated with a SVR model for the purpose of parameter optimization. So the improved model is better than the original one.

Future Work

- Other clustering algorithms can be used to observe the impact on result.
- A new algorithm to optimize parameters in brain storm optimization can be created if you are daring enough. (just a thought)

References

- Jianzhou Wang, Ru Hou, Chen Wang, Lin Shen, Improved-Support Vector Regression Model Based on Variable Selection and Brain Storm Optimization for Stock Price Forecasting, Applied Soft Computing Journal (http://dx.doi.org/10.1016/j.asoc.2016.07.024)
- A Comprehensive Survey of Brain Storm Optimization Algorithms by Shi Cheng, Yifei Suny, Junfeng Chenz, Quande Qinx, Xianghua Chux, Xiujuan Lei, Yuhui Shi from School of Computer Science, Shaanxi Normal University, Xi'an, China. (https://doi.org/10.1109/CEC.2017.7969498)
- Analytics Vidhya
- https://www.investopedia.com/terms/t/technicalindicator.asp
- https://finance.yahoo.com/quote/399001.SZ/history?period1=1409250600&period2=1567017000&interval=1d&filter=history&frequency=1d

THANK YOU