International Journal of Smart Energy Technology and Environmental Engineering
Volume 2, Issue 2, September 2023
http://globalpublisher.org/journals-1007/
DOI: 10.52547/setee.2.2.154

Behavior Analysis of High-Performance Concrete Using Data Mining Techniques

Behrouz Alibeyk [1*], Alireza Saraei [2]

1- Independent Researcher, Imperial College Business School
Toronto, Canada

2- Assistant Professor, Department of Mechanical Engineering, Islamic Azad University
Tehran, IRAN

*Corresponding Author: behrouz.alibeyk18@alumni.imperial.ac.uk

ABSTRACT

This study tried to predict mechanical behavior of high-performance concrete (HPC), specially the compressive strength of HPC using different data mining methods. HPC is a highly complex composite material and modelling of its dynamics is a real challenge. Moreover, compressive strength of HPC is nonlinear function of its ingredients. The results of several studies have represented that compressive strength of HPC depends on not only water/cement ratio but also some other additive ingredients. It is actually a function of cement, blast furnace slag, fly ash, water, superplasticizer, coarse aggregate, fine aggregate and age. The quantitative analysis in this study were conducted by using Principal components analysis (PCA) and Multiple Regression (MR) methods. For this purpose, some effective statistical analysis and modeling software such as MATLAB, MINITAB and R has been used. Analytical results suggested that Multiple Regression is effective for predicting behavior of HPC based on its compressive strength with respect to different ages.

Keywords: High-performance concrete, Data Mining, Concrete Strength, predictive techniques.

1. INTRODUCTION

Data mining is a process of extracting implicit, previously unknown, but potentially useful information and knowledge from a large quantity of incomplete, noisy, ambiguous and random data in the practical aspplication.

Main objective in this study is to implement different data mining techniques based on Knowledge Discovery in Databases (KDD).

Modeling the dynamics of HPC, which is a highly complex composite material, is extremely challenging. Concrete compressive strength is also a highly nonlinear function of ingredients and that's why optimizing the prediction accuracy of mechanical properties of these composites are really interesting and challenging.

2. RELATED WORK

Several studies have proposed approaches for modeling concrete compressive strength.
Oh et al. (1999) applied ANNs to optimize the proportion of four concrete ingredients (water, cement, fine aggregate, and coarse aggregate). They used a tool for minimizing uncertainty and errors in proportioning concrete mixes, which is a complicated, time-consuming, and uncertain task (Oh et al. 1999). [1]
Mostofi and Samaee (1995) applied multilayer perceptron ANNs to estimate HPC compressive strength (Mostofi and Samaee 1995).[2]
Yeh (1998) modeled HPC strength by using artificial neural networks as a function of cement, fly ash, blast-furnace slag, water, superplasticizer, coarse aggregate, fine aggregate, and age of testing and obtained promising results (Yeh 1998). [5] Kasperkiewicz and Dubrawski (1995) applied fuzzy-ARTMAP ANNs to predict the 28-day compressive strength of HPC mixes.[9]
Fazel Zarandi et al. (2008) developed a fuzzy polynomial neural network (FPNN) that combined fuzzy neural networks (FNNs) and polynomial neural networks (PNNs). Six different FPNN architectures were constructed. Each architecture had six input parameters (concrete ingredients) and one output parameter (28-day compressive strength of the mix-design). [10]
Trtnik et al. (2009) used an ultrasonic pulse technique, which is among the most common nondestructive techniques for assessing concrete properties. This testing method employs a portable ultrasonic nondestructive digital indicating tester to generate an ultrasonic pulse, which is transmitted through the concrete and received at the opposite surface. Upon receiving the pulse, the instrument amplifies it and measures the time it required to travel through the concrete.[11]
Gupta et al. (2006) presented a neural-fuzzy inference system for predicting the compressive strength of HPC.
Generally, previous studies applied similar ANN techniques with only minor modifications. [12]
The aim of this study was to exploit common data-mining techniques to model HPC compressive strength as a function of its primary ingredients and to improve strength prediction performance.

3. PROPOSED SOLUTION

Data description and preparation
The experimental data set used in this study was obtained from following data source:
Original Owner and Donor Prof. I-Cheng Yeh
Department of Information Management Chung-Hua University,
Hsin Chu, Taiwan 30067, R.O.C. E-mail: icyeh@ chu.edu.tw
TEL: 886-3-5186511

Date Donated: August 3, 2007

Table 1 represents a summary of information about all attributes in this data set. As shown, there are 9 attributes and 1030 instances in this data set. Attribute breakdown is 8 quantitative input variables and 1 quantitative output variable, and there is no missing attribute values. All tests were performed on 15 cm cylindrical specimens of concrete prepared by using standard procedures.

TABLE 1. DATA SET CHARACTERISTICS

Attribute	Unit	Minimum	Maximum	Average	Standard deviation
Cement	kg/m^3	102.0	540.0	281.168	104.506
Blast-furnace slag	kg/m^3	11.0	359.4	107.277	61.884
Fly ash	kg/m^3	24.5	200.1	83.862	39.989
Water	kg/m^3	121.8	247.0	181.567	24.354
Superplasticizer	kg/m^3	1.7	32.2	8.486	4.037
Coarse aggregate	kg/m^3	801.0	1,145.0	972.919	77.754
Fine aggregate	kg/m^3	594.0	992.6	773.580	80.176
Age of testing	Day	1.0	365.0	45.662	63.170
Concrete compressive strength	MPa	2.3	82.6	35.818	16.706

Data sets described in the literature often contain unexpected inaccuracies. For example, the class of fly ash may not be indicated. Another problem is that the superplasticizers may be produced by different manufacturers and have different chemical compositions (Kasperkiewicz and Dubrawski 1995). Concrete compressive strength not only is determined by the water to concrete ratio, but also by the other materials used in the mix. Concrete contains five ingredients other than cement and water. The multiple ingredients, in addition to the nonlinearity of concrete structures, complicate the computation of compressive strength. The following predictive techniques are proposed for applying these complex inputs when modeling the compressive strength of HPC.

Pre-processing Step

In this study, data mining techniques are used through KDD process to determine main and interaction effects of different components in a High-performance concrete (HPC) on its compression strength, and also to find a statically reliable model to be able to predict and infer the strength behavior of HPC based on introduced data set. For this purpose, different statistical and modelling software such as MATLAB, MINITAB and R were used in different steps.

The results of data exploration step showed that there is no missing data in this data set, and all attributes are quantitative. Moreover, compressive strength variable is considered as output attribute or response variable. So, it means in this study focused should be on supervised approaches, and fortunately it is possible to check the quality of modeling

In pre-processing step, different analysis have performed on data to find noise data or outliers, correlation between attributes, clusters among them and also Principal Component Analysis

(PCA) has performed to find out the possibility of dimension reduction on the data using MATLAB.

Fig. 1 shows boxplot of all attributes. As it is clear, there are not many outliers among the data. The most significant outliers are related to "Age" attribute. But in response variable "strength" data, there is no outliers. Based on this result, it was concluded to keep all data because it seemed all data are important and relevant.

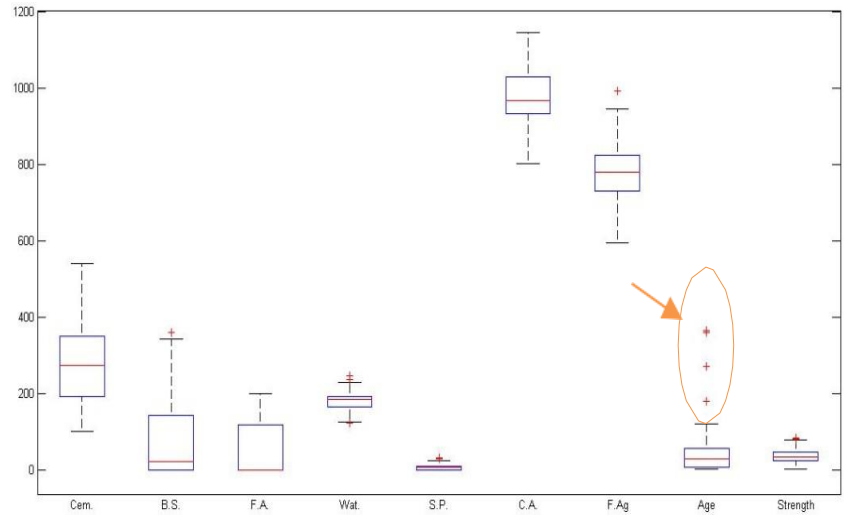

Fig. 1. Boxplot diagram of all attributes

Figures 2 and 3 represent correlation between different input variables and response variable (strength). The results of this analysis give good idea about impact of different attributes on each other. It is obvious that the amount of water and superplasticizer are strongly negative correlated. That means if amount of water in concrete increases, superplasticizer content should decrease to reach the same compressive strength. On the other hand, water and age are positively correlated. It is clear that increasing water content of a concrete results in increasing the age of concrete and it needs more time to reach the same strength result. Despite of what expected, cement content has no strong correlation with water content and age, but the amount of fine aggregation has a considerable negative impact on cement content. All other correlation between these attributes can be extracted from these figures.

	Cem.	B.S.	F.A.	Wat.	S.P.	C.A.	F.Ag	Age
Cem.	1.000	-0.275	-0.397	-0.082	0.093	-0.109	-0.223	0.082
B.S.	-0.275	1.000	-0.324	0.107	0.043	-0.284	-0.282	-0.044
F.A.	-0.397	-0.324	1.000	-0.257	0.377	-0.010	0.079	-0.154
Wat.	-0.082	0.107	-0.257	1.000	-0.657	-0.182	-0.451	0.278
S.P.	0.093	0.043	0.377	-0.657	1.000	-0.266	0.223	-0.193
C.A.	-0.109	-0.284	-0.010	-0.182	-0.266	1.000	-0.179	-0.003
F.Ag	-0.223	-0.282	0.079	-0.451	0.223	-0.179	1.000	-0.156
Age	0.082	-0.044	-0.154	0.278	-0.193	-0.003	-0.156	1.000

Fig. 2. Correlation plot between attributes

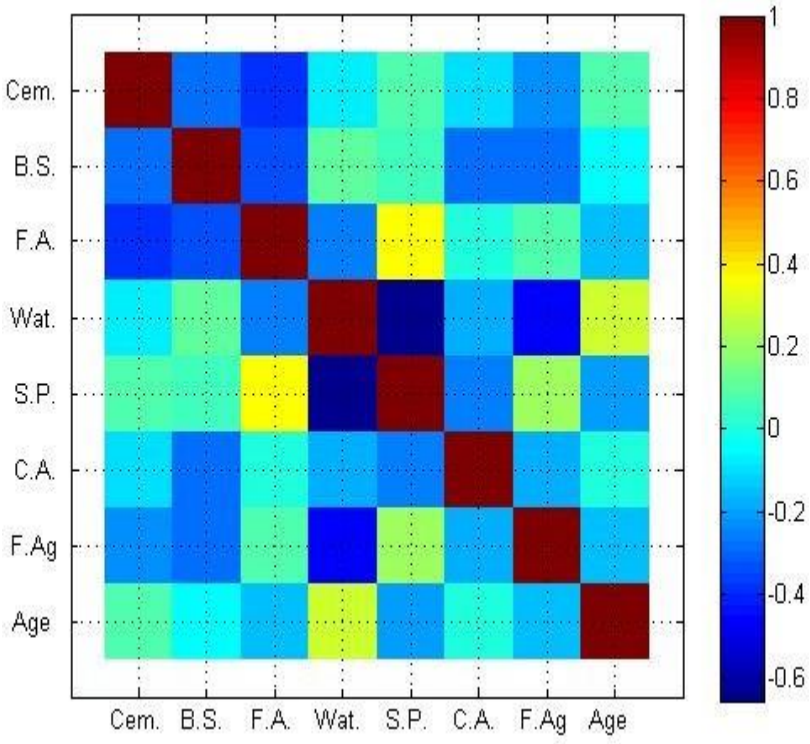

Fig. 3. Color coded correlation plot between attributes

TABLE 2. Main Effect Analysis of Variables on Response (Strength)

```
Call:
lm(formula = strength ~ cement + blast + flyash + water + superplast +
    coarseagg + fineagg + age, data = x)

Residuals:
    Min      1Q  Median      3Q     Max
-28.654  -6.302   0.703   6.569  34.450

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -23.331214  26.585504  -0.878 0.380372
cement        0.119804   0.008489  14.113  < 2e-16 ***
blast         0.103866   0.010136  10.247  < 2e-16 ***
flyash        0.087934   0.012583   6.988 5.02e-12 ***
water        -0.149918   0.040177  -3.731 0.000201 ***
superplast    0.292225   0.093424   3.128 0.001810 **
coarseagg     0.018086   0.009392   1.926 0.054425 .
fineagg       0.020190   0.010702   1.887 0.059491 .
age           0.114222   0.005427  21.046  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.4 on 1021 degrees of freedom
Multiple R-squared:  0.6155,    Adjusted R-squared:  0.6125
F-statistic: 204.3 on 8 and 1021 DF,  p-value: < 2.2e-16
```

Also, following scatterplot (Fig.4) clearly shows relationship between attributes and response variable as well as relationship among input variable themselves. The most important result of this figure is the effect of cement content on compressive strength of HPC. As a result, increase of cement causes increase of strength. But in most cases, to find the exact relationship between the attributes, it is necessary to implement other analysis methods. For this purpose, main effects and interaction effects of all attributes have been analyzed using MINITAB and R software.

As we have response variable (strength) in the data set, main effect of input variables address the impact of each input variable on the response variable. Interaction effects, on the other hand, refer to mutual effects of variables on the response.

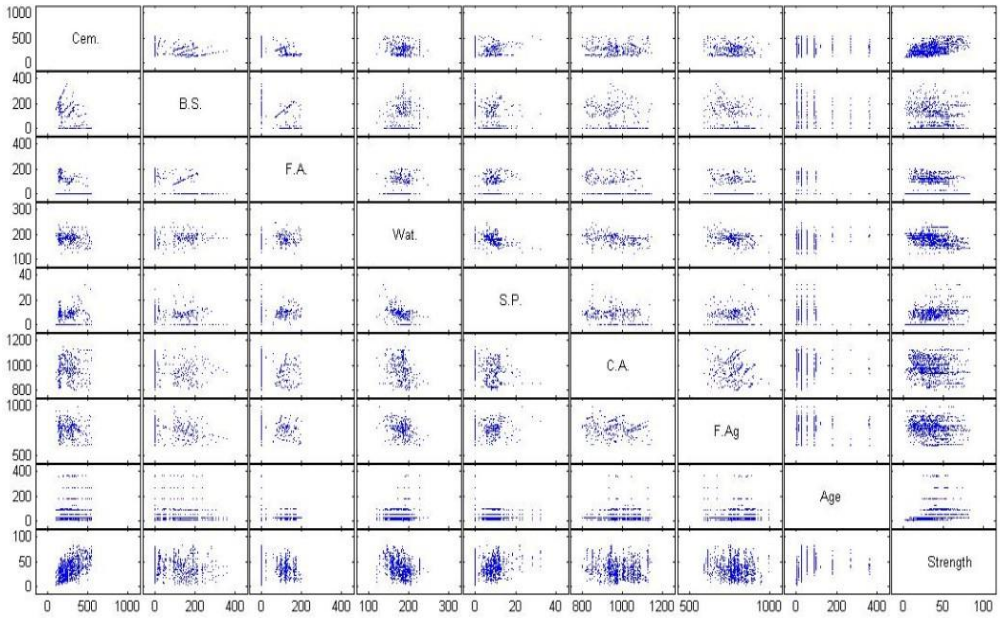

Fig. 4. Scatterplot of attributes

Table 2 represents main effects analysis results by R. The result shows that find 5 variables (Cement, Blast, Fly ash, Water and Age) have significant main effect on our response variable. Superplasticizer has less effect and 2 other variables (Coarse aggregate and fine aggregate) has no significant main effect on strength. Among all attributes, age has the most significant effect on strength. That's why in this study even outliers of this attribute have been kept. Other important variables are cement, blast, fly ash and water respectively.

Practically, interaction effects of variables are more important than main effects, because in real situation all ingredients are together to form a HPC. So, these interaction effects have been evaluated by R and MINITAB software. To do so, all variables have been evaluated two by two and their interaction effect on response variable have been investigated.

The results of these analysis show that not all attributes have
significant interaction effects, but most of them have. It should also be noted that in presence of interaction effects, the impact of main effects of individual attributes may change. That's why, interaction effects are more important to explore behavior of different components.

As a sample of these interaction effects analysis, Table 3 shows results of interaction effect analysis of water and superplasticizer variables on strength. As it can be seen, not only these two variables have significant interaction effects on strength (based on their t-value in the table), but also main effects of them as well as other attributes have changed.

TABLE 3. interaction Effect Analysis of Variables on Response (water: superplasticizer interaction effects)

```
Call:
lm(formula = strength ~ cement + blast + flyash + water + superplast +
    coarseagg + fineagg + age + water * superplast, data = x)

Residuals:
    Min      1Q  Median      3Q     Max
-30.273  -6.100   0.287   6.971  32.865

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      0.828724  26.603012   0.031   0.9752
cement           0.116889   0.008391  13.930  < 2e-16 ***
blast            0.096902   0.010081   9.612  < 2e-16 ***
flyash           0.065267   0.013101   4.982 7.40e-07 ***
water           -0.245582   0.043401  -5.658 1.98e-08 ***
superplast      -2.261218   0.481073  -4.700 2.95e-06 ***
coarseagg        0.015449   0.009278   1.665   0.0962 .
fineagg          0.017001   0.010573   1.608   0.1082
age              0.121464   0.005519  22.010  < 2e-16 ***
water:superplast 0.015648   0.002894   5.408 7.94e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.26 on 1020 degrees of freedom
Multiple R-squared:  0.6262,    Adjusted R-squared:  0.6229
F-statistic: 189.9 on 9 and 1020 DF,  p-value: < 2.2e-16
```

Fig. 5 shows the plot related to interaction effect analysis of water – superplasticizer on strength.

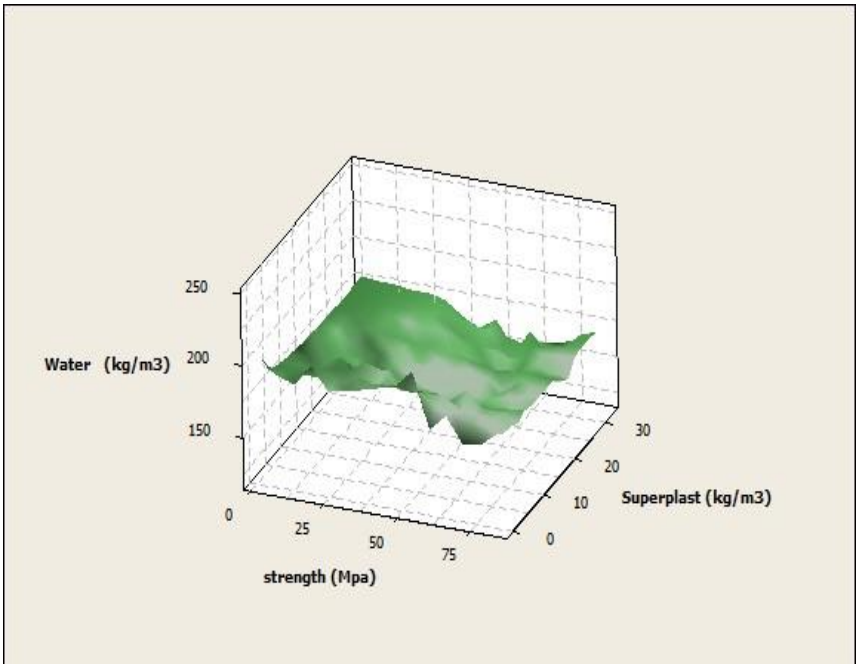

Fig. 5. 3D Interaction Effects Plot of Water and Superplasticizer

Results of all interaction effect analysis show that following attributes have significant interaction effect on strength of HPC:
Water – Superplasticizer
Water – Coarse aggregate
Water – Age
Water – Fine aggregate
Coarse aggregate – Superplasticizer
Age – Superplasticizer
Fly ash – Age
Cement – Superplasticizer
Cement – Age

Principle Component Analysis (PCA)
Principal component analysis (PCA) is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. The number of principal components is less than or equal to the number of original variables. This transformation is defined in such a way that the first principal component has the largest possible variance (that is, accounts for as much of the variability in the data as possible), and each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to the preceding components. The resulting vectors are an uncorrelated orthogonal basis set. The principal components are orthogonal because they are the eigenvectors of the covariance matrix, which is symmetric. PCA is sensitive to the relative scaling of the original variables. After pre-processing step, a Principal Component Analysis (PCA) has been performed to find out the possibility of dimension reduction of attributes. Figures 6 and 7 show the results of achieved principal components based on the experimental data set.

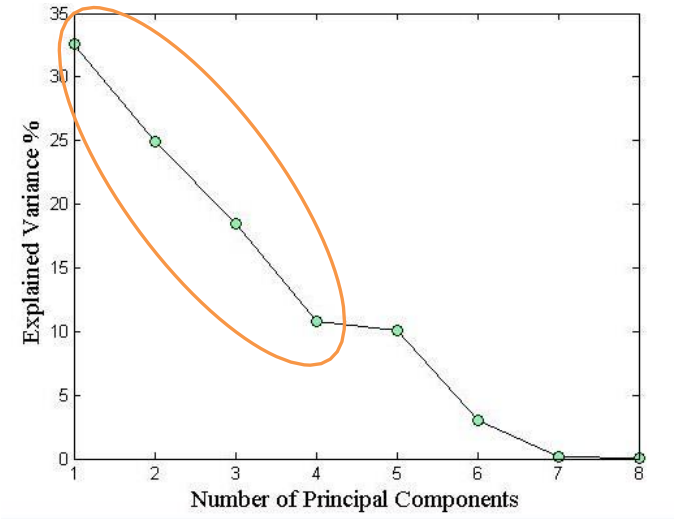

Fig. 6. Number of Principal Component's diagram

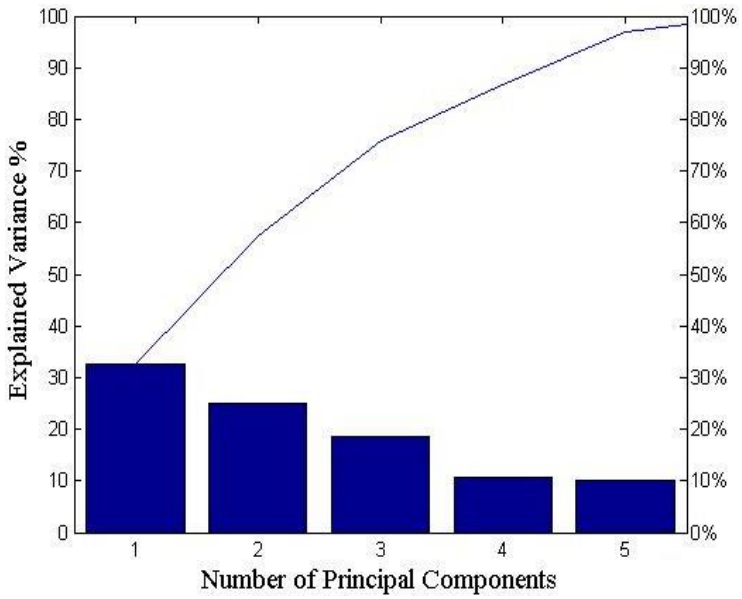

Fig. 7. Pareto Plot of Principal Component's diagram

This result represents that the dimension of attributes can be reduced to 4 principle components. That's a very interesting result and it could be very useful especially when the data set contains many different variables and it is difficult to analyze them from different point of views.
Fig. 8 shows interaction effect of component 1 and component 2 as a result of PCA analysis.

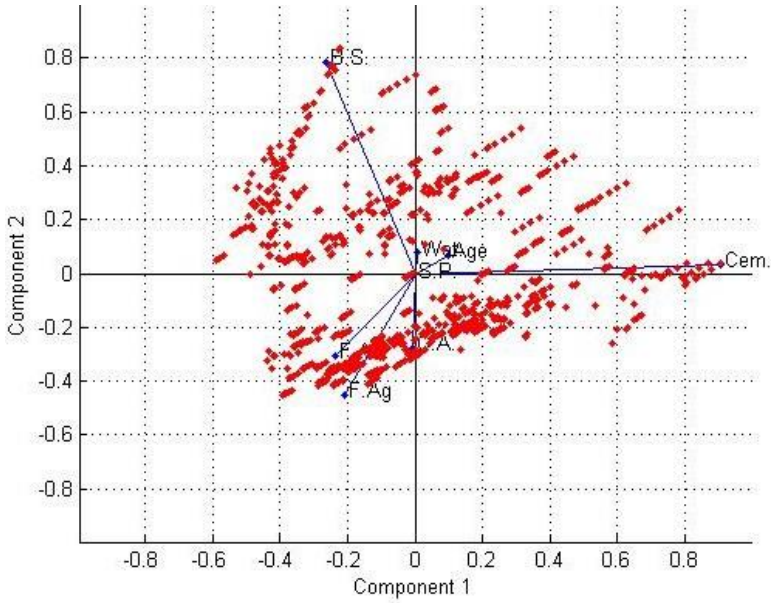

Fig. 8. Interaction effect between component 1 and component 2

Similarly, Fig. 9 visualize interaction effects of 3 first components in 3D diagram.

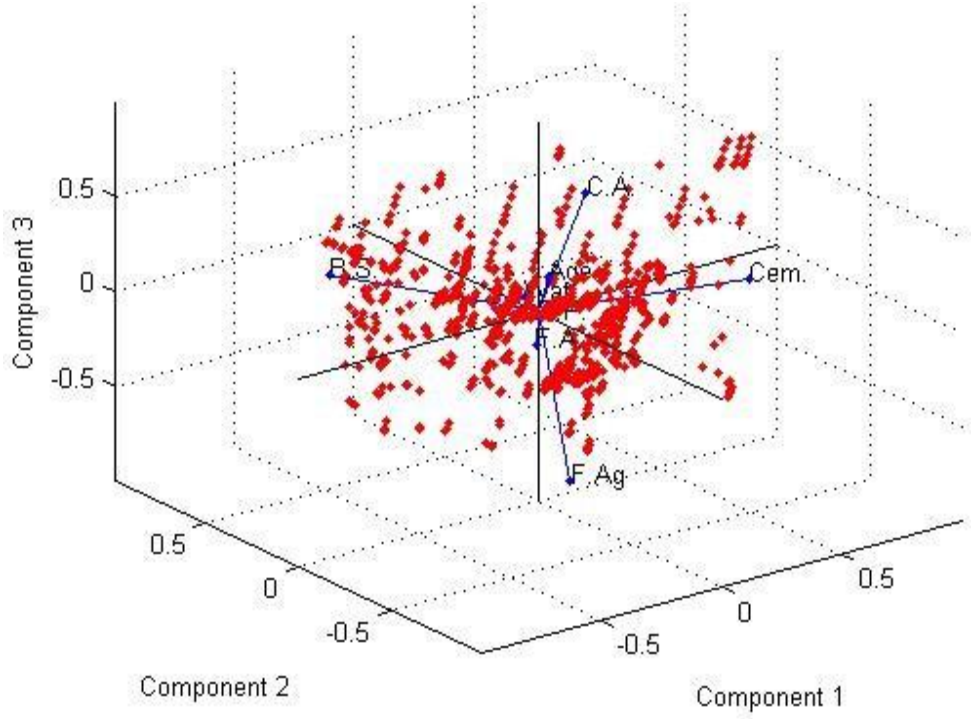

Fig. 9. 3D diagram Interaction effect between component 1 and component 2 and component 3

Although the result are considerable, but in this study these results have been neglected and it was preferred to use principal data set because in this case, it is not a big data and it is not necessary to use PCA results to reach a model for predicting HPC behavior. Main reason to do PCA was to determine the possibility of dimension reduction of such a data in case of very big data set.

LINEAR MULTIPLE REGRESSION

Multiple regression (MR) models depict the relationship among two or more variables. The computational problem addressed by multiple regression is fitting a plane to an n- dimensional space where n = number of independent variables to a number of points.

For a system with n inputs (independent variables) and one output (dependent variable), Y, the general least square (or linear regression) problem is to determine unknown parameters, bi, of the linear model, as shown in Eq. (1):

$$Y = C + b_1 * X_1 + b_2 * X_2 + b_3 * X_3 + \ldots + b_{n-1} * X_{n-1} + b_n * X_n \quad (1)$$

In the proposed regression model, Y represents concrete compressive strength, and b1; b2;…; b8 are regression coefficients.

The Xi values represent cement, blast-furnace slag, fly ash, water, superplasticizer, coarse aggregate, fine aggregate, and age; and C is the estimated constant. Regression analysis estimates the unbiased values of the regression coefficients b1;

b2;…; b8 against the training data set.
As all attributes in this data set are continuous attributes and this is a supervised data mining project, one of the best methods to model the behavior of HPC is MR.
In this study Multiple Regression has performed using R software. For this purpose, at start multiple linear regression method has been examined. Table 4 shows the result of this step. It illustrates the regression equation and all coefficients b1 to b8. Moreover, Analysis of Variance information is given in this table.

TABLE 4. Multiple linear regression results

```
The regression equation is
Strength       (Mpa) = - 23.2 + 0.120 Cement (kg/m3) + 0.104 Blast (kg/m3)
                       + 0.0879 FlyAsh (kg/m3) - 0.150 Water    (kg/m3)
                       + 0.291 Superplas (kg/m3) + 0.0180 CoarseAgg  (kg/m3)
                       + 0.0202 FineAgg    (kg/m3) + 0.114 Age       (day)


Predictor               Coef   SE Coef      T      P
Constant               -23.16     26.59  -0.87  0.384
Cement (kg/m3)      0.119785  0.008489  14.11  0.000
Blast (kg/m3)        0.10385   0.01014  10.25  0.000
FlyAsh (kg/m3)       0.08794   0.01259   6.99  0.000
Water    (kg/m3)    -0.15030   0.04018  -3.74  0.000
Superplast (kg/m3)   0.29069   0.09346   3.11  0.002
CoarseAgg  (kg/m3)  0.018030  0.009394   1.92  0.055
FineAgg    (kg/m3)   0.02015   0.01070   1.88  0.060
Age       (day)     0.114226  0.005427  21.05  0.000


S = 10.3998   R-Sq = 61.5%   R-Sq(adj) = 61.2%


Analysis of Variance

Source            DF      SS     MS       F      P
Regression         8  176745  22093  204.27  0.000
Residual Error  1021  110428    108
Total           1029  287173


Source              DF  Seq SS
Cement (kg/m3)       1   71172
Blast (kg/m3)        1   22957
FlyAsh (kg/m3)       1   21636
Water    (kg/m3)     1   11459
Superplast (kg/m3)   1    1360
CoarseAgg  (kg/m3)   1     253
FineAgg    (kg/m3)   1       1
Age       (day)      1   47905
```

Fig. 10 shows the residuals result from this equation. In next step, this multiple linear regression equation should be tested to be sure that it validate all assumptions and this model is creditable.

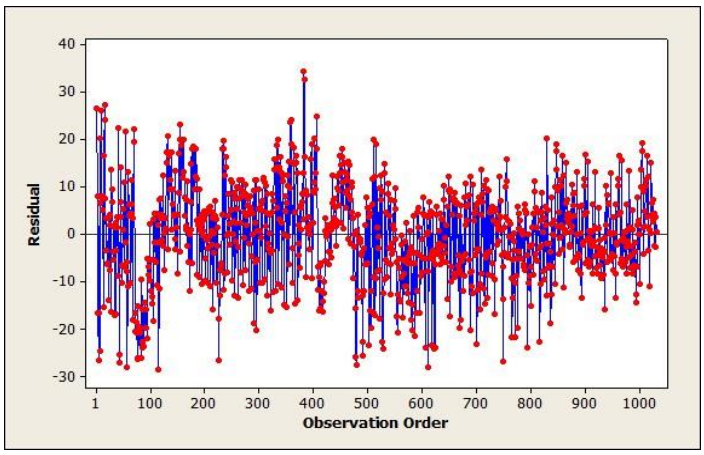

Fig. 10. Residuals of Multiple Linear Regression Equation.

First assumption is Normal Probability Distribution of the residuals. To verify this assumption Figures 11 and 12 have been considered. These figures compensate that distribution of residuals of this equation is normal and this assumption is validated.

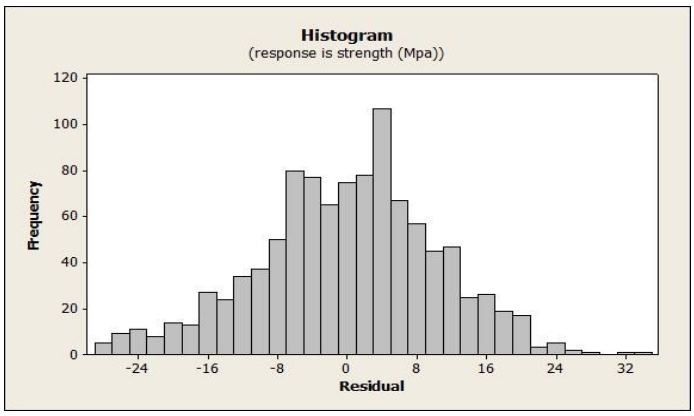

Fig. 11. Histogram of the Residuals

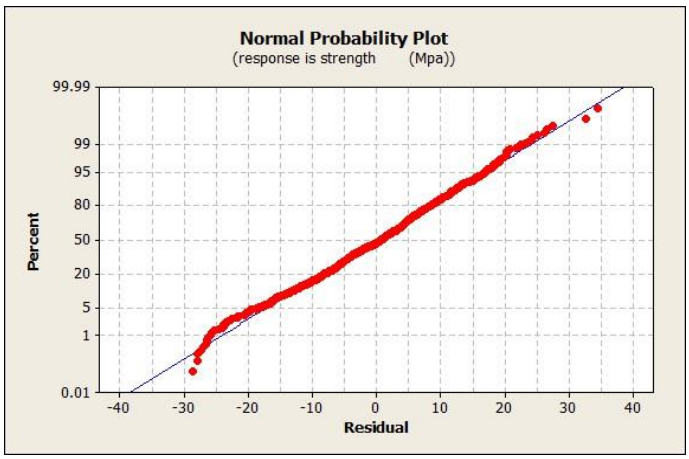

Fig. 12. Norm plot of the Residuals

Fig. 13 shows the result related to verify second assumption of this regression model. As it is clear, there is a pattern in this plot and this pattern demonstrate a relationship between the residuals and fitted values. So, second assumption is violated and it shows that this multiple linear regression is not a suitable model to predict behavior of concrete, and it results that relationship between these attributes is not linear.

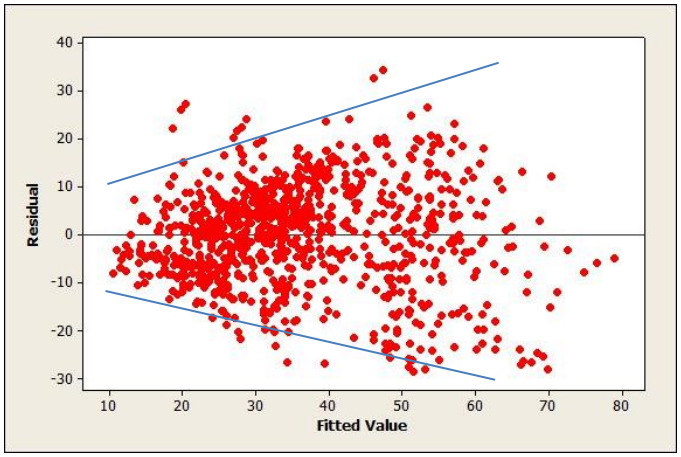

Fig. 13. Residuals vs fitted values

Logarithmic Multiple Regression

Previous outcomes demonstrated that relationship of concrete components as well as its compressive strength is not linear. As a result, it is necessary to look for a Non-linear Multiple Regression equation. Regarding to this result, Logarithmic Multiple Regression equation has been tested. Table 5 shows the equation and its estimated coefficients. Also, analysis of variance results have been represented.

Fig. 14 shows the residuals result from this equation.

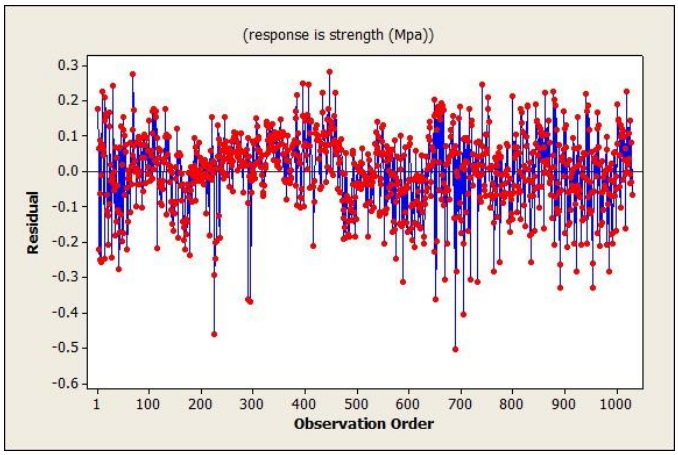

Fig. 14. Residuals of Logarithmic Regression Equation.

TABLE 5. Logarithmic regression results

```
Logarithmic Regression Analysis: strength

The regression equation is
strength       (Mpa) = 2.70 + 0.786 Cement (kg/m3) + 0.0659 Blast (kg/m3)
                       + 0.0293 FlyAsh (kg/m3) - 1.09 Water   (kg/m3)
                       + 0.0608 Superplast (kg/m3) - 0.015 CoarseAgg  (kg/m3)
                       - 0.394 FineAgg   (kg/m3) + 0.291 Age      (day)


Predictor               Coef   SE Coef      T      P
Constant               2.695     1.032   2.61  0.009
Cement (kg/m3)       0.78586   0.03301  23.81  0.000
Blast (kg/m3)       0.065885  0.005398  12.21  0.000
FlyAsh (kg/m3)      0.029300  0.005635   5.20  0.000
Water   (kg/m3)      -1.0866    0.1276  -8.52  0.000
Superplast (kg/m3)   0.06077   0.01338   4.54  0.000
CoarseAgg  (kg/m3)   -0.0150    0.1548  -0.10  0.923
FineAgg   (kg/m3)    -0.3943    0.1242  -3.17  0.002
Age      (day)      0.290916  0.006671  43.61  0.000


S = 0.108729   R-Sq = 79.5%   R-Sq(adj) = 79.4%


Analysis of Variance

Source            DF       SS      MS       F      P
Regression         8  46.9137  5.8642  496.04  0.000
Residual Error  1021  12.0702  0.0118
Total           1029  58.9840


Source              DF   Seq SS
Cement (kg/m3)       1  13.1442
Blast (kg/m3)        1   4.7293
FlyAsh (kg/m3)       1   4.0748
Water   (kg/m3)      1   1.1397
Superplast (kg/m3)   1   0.7060
CoarseAgg  (kg/m3)   1   0.1347
FineAgg   (kg/m3)    1   0.5010
Age      (day)       1  22.4841
```

In next step, logarithmic regression equation should be tested to be sure that it validate all assumptions and this model is significantly creditable to model this data.
First assumption is Normal Probability Distribution of the residuals. To verify this assumption Figures 15 and 16 have been considered. Fig. 15 shows that distribution of these residuals is not completely normal.

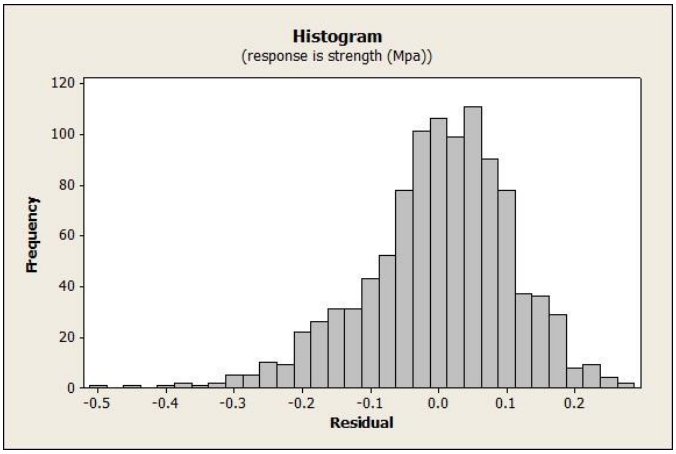

Fig. 15. Histogram of the Residuals

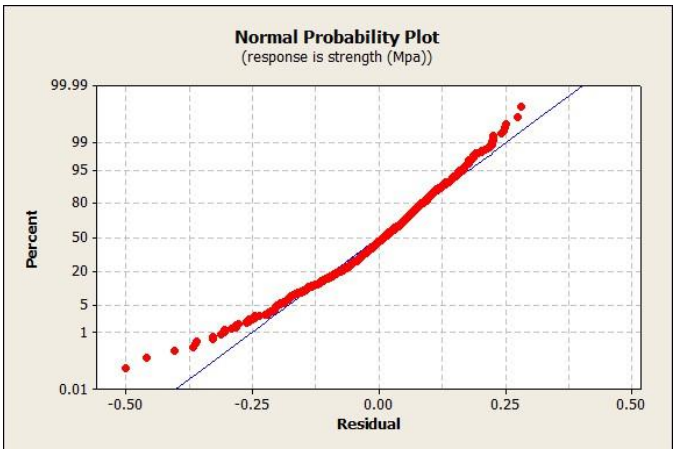

Fig. 16. Norm plot of the Residuals

Normal probability plot of the residuals in Fig.16 gives a clearer view for this result. It shows that the residuals don't make a straight line in their norm plot completely. So, the first assumption is violated and it shows Logarithmic Multiple Regression is not a suitable model for this data.

Fig. 17 shows another prove for this fact. It represents the result related to verify second assumption of this regression model. As it is clear, there is again a pattern in this plot and this pattern demonstrate a relationship between the residuals and fitted values. So, second assumption is also violated and it demonstrates that this logarithmic regression is not a suitable model to predict behavior of concrete.

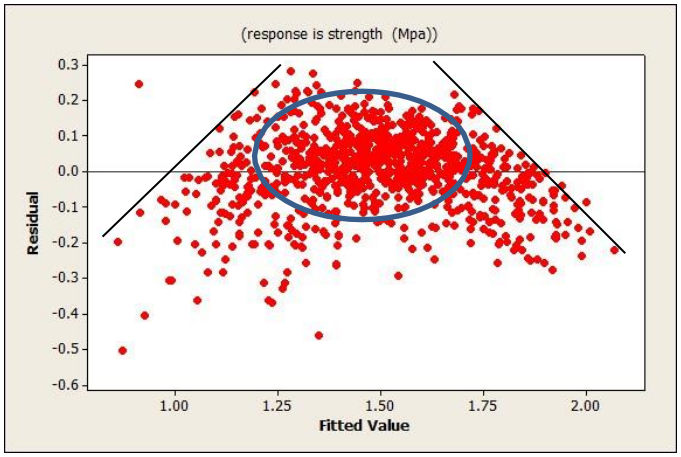

Fig. 17. Residuals vs fitted values

Root Square Regression

Regarding previous outcomes, Root Square Regression model has been considered, and Table 6 shows the results.

TABLE 6. Square Root regression results

```
Root square Regression Analysis
The regression equation is
strength       (Mpa) = - 1.62 + 0.312 Cement  (kg/m3) + 0.108 Blast  (kg/m3)
                       + 0.0615 FlyAsh  (kg/m3) - 0.318 Water    (kg/m3)
                       + 0.154 Superplast (kg/m3) + 0.0946 CoarseAgg  (kg/m3)
                       + 0.0375 FineAgg   (kg/m3) + 0.229 Age       (day)


Predictor               Coef   SE Coef      T      P
Constant              -1.622     2.856  -0.57  0.570
Cement  (kg/m3)      0.31232   0.01576  19.82  0.000
Blast  (kg/m3)      0.107918  0.008176  13.20  0.000
FlyAsh  (kg/m3)     0.061496  0.008749   7.03  0.000
Water    (kg/m3)    -0.31829   0.06347  -5.01  0.000
Superplast (kg/m3)   0.15405   0.02766   5.57  0.000
CoarseAgg  (kg/m3)   0.09461   0.03452   2.74  0.006
FineAgg    (kg/m3)   0.03747   0.03281   1.14  0.254
Age       (day)      0.229249  0.006757  33.93  0.000


S = 0.736927   R-Sq = 74.3%   R-Sq(adj) = 74.1%


Analysis of Variance

Source            DF       SS      MS       F      P
Regression         8  1604.90  200.61  369.41  0.000
Residual Error  1021   554.47    0.54
Total           1029  2159.37


Source              DF  Seq SS
Cement  (kg/m3)      1  518.52
Blast  (kg/m3)       1  198.79
FlyAsh  (kg/m3)      1  178.76
Water    (kg/m3)     1   57.96
Superplast (kg/m3)   1   17.38
CoarseAgg  (kg/m3)   1    7.76
FineAgg    (kg/m3)   1    0.70
Age       (day)      1  625.04
```

Table 6 illustrates root square regression equation as well as related coefficients and analysis of variance information which are all the outcome reached by R software.
Fig. 18 shows the residuals result from the last equation.

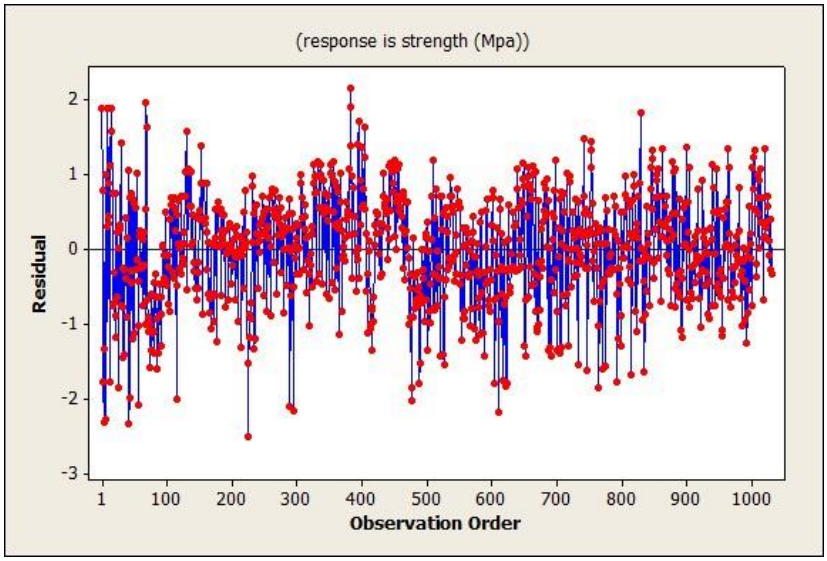

Fig. 18. Residuals of Multiple Linear Regression Equation.

As before, for first step, normality assumption of residuals' probability distribution should be considered and tested. Figures 19 and 20 show these results. As it is clear, this probability distribution is visually normal and it validate first assumption in this analysis.

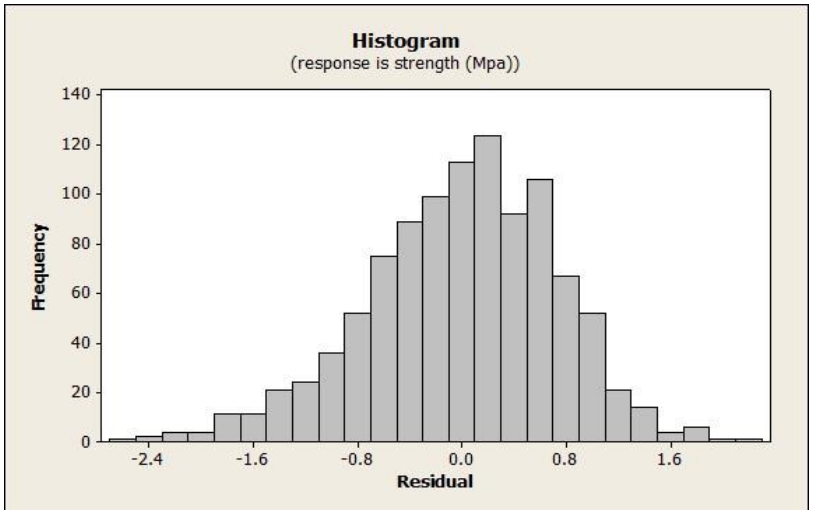

Fig. 19. Histogram of the Residuals

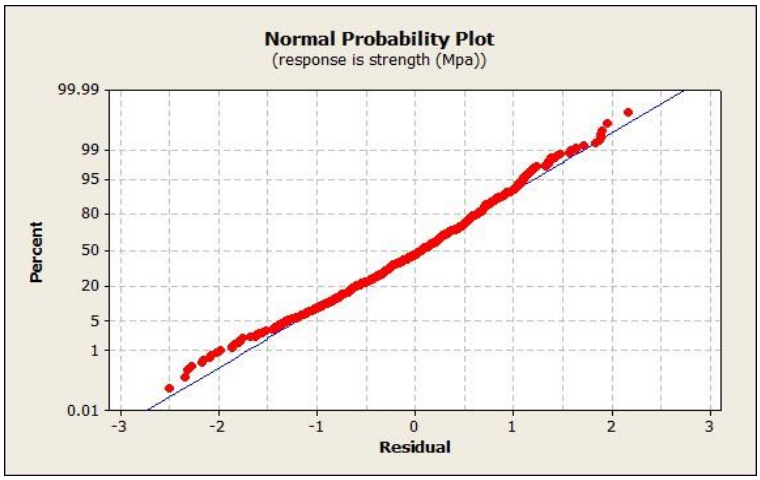

Fig. 20. Norm plot of the Residuals

Fig. 21 shows the result related to verify second assumption of this regression model. As it is clear, there is no pattern in this plot. So, second assumption is verified and it shows that this root square regression is a suitable model to predict behavior of concrete, and it results that relationship between these attributes is not linear.

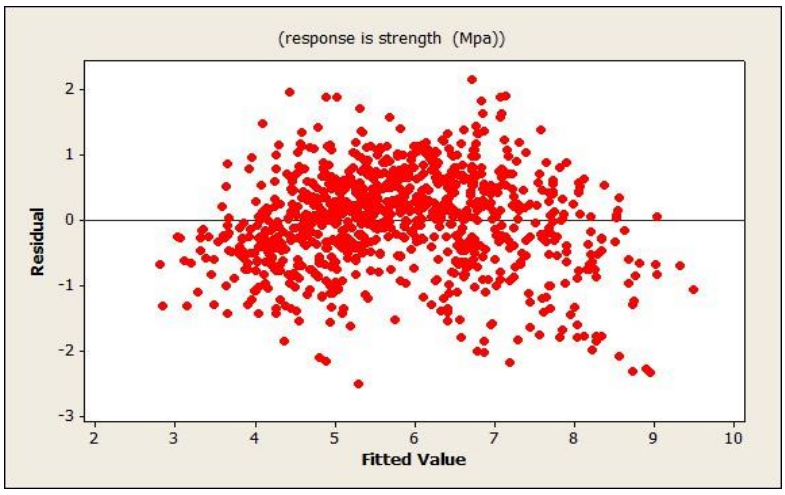

Fig. 21. Residuals vs fitted values of Root Square Regression

CONCLUSION

This study developed a data-mining approach and performance measures to predict compressive strength and assess the prediction reliability for HPC. In this project we first tried to perform some data exploration techniques. Fortunately there was no missing values in our data set. The nature of the chosen data set was supervised since we have the output value, due to this we had to use MR (Multiple Regression) analysis as out main data mining technique. After exploring the data we try to detect the outliers using box plots by using Minitab and R software. The result of the outlier detection process showed us that “water”, as one of the main

components, has the biggest number of outliers amongst the other components. But due to the accuracy of the final results, we decided not to eliminate those outliers.

Furthermore, we defined the correlation between different variables using scatter plots, By Minitab. We also defined main effects and interaction effects and we draw the 3D diagrams accordingly using Minitab and R.

As the next step, we decided to perform PCA analysis on the data set, to decrease the dimensionality of the data set. According to the results of PCA analysis, we are able to decrease the number of our main components to four instead of six, but again for the sake of getting more accurate results, we decided to keep our data set as is.

Moreover, we applied MR data mining analysis. First we applied simple linear regression analysis. Based on our results we could observe that a pattern exists scatter lot of residuals vs fitted values. So we changes the regression analysis type and performed logarithmic regression analysis. Similarly a pattern was observed in the graph of residuals vs fitted values. Therefore, we tired another type of regression analysis, namely square root regression analysis and the result of this type of the regression analysis was promising.

According to the results of the main effects and interaction effects, 5 of the components has the most effect on the compressive strength of HPC.

Acknowledgments

The writers would like to thank UC Irvine Machine Learning Repository (http://archive.ics.uci.edu/ml/) and Professor I- Cheng
Yeh for sharing the experimental data set.

REFERENCES

1. Oh, J., Lee, I., Kim, J., and Lee, G. (1999). “Applications of neural networks for proportioning of concrete mixes.” ACI Mater. J., 96(1), 51–59.
2. Mostofi, D., and Samaee, A. (1995). “HPC strength prediction using artificial neural network.” J. Comput. Civ. Eng., 9(4), 44–49.
3. Naresh Kumar Nagwani and Shirish V. Deo, ‘’ Estimating the Concrete Compressive Strength Using Hard Clustering and Fuzzy Clustering Based Regression Techniques,’’ National Institute of Technology Raipur, Raipur, Chhattisgarh 492010, India, Hindawi Publishing Corporation, The Scientific World Journal, Volum2014, Article ID 381549.
4. Jui-Sheng Chou, “Concrete compressive strength analysis using a combined classification and regression technique,” Department of Construction Engineering, National Taiwan University of Science and Technology, 43 Sec.4, Keelung Road, Taipei, Taiwan, Automation in Construction 24 (2012) 52–60
5. I.-C. Yeh, “MODELING OF STRENGTH OF HIGH-PERFORMANCE CONCRETE USING ARTIFICIAL NEURAL NETWORKS,” Department of Civil Engineering, Chung-Hua University, 30 Tung Shiang, Hsin Chu, 30067, Taiwan, Republic of China, Received July 3, 1998; in final form September 14, 1998.
6. M. F. M. Zain, Suhad M. Abd, K. Sopian, M. Jamil1 , Che-Ani A.I, “Mathematical Regression Model for the Prediction of Concrete Strength,” Faculty of Engineering and

Built Environment, Universiti Kebangsaan Malaysia, 43600 UKM Bangi, Selangor Darul Ehsan, Malaysia

7. M.F.M Zain and S.M.Abd, "Multiple Regression Model for Compressive Strength Prediction of High Performance Concrete," Department of Civil Engineering, Diyala University, Iraq, Journal of Applied Science 9(1):155-160, 2009.
8. U. Atici, "Prediction of the strength of mineral admixture concrete using multivariable regression analysis and an artificial neural network," Engineering Faculty, Nigde University, Nigde 51245, Turkey, Expert Systems with Applications 38 (2011) 9609–9618.
9. Kasperkiewicz, J., and Dubrawski, A. (1995). "HPC strength prediction using artificial neural network." J. Comput. Civ. Eng., 9(4), 279–284.
10. Kasperkiewicz, J., and Dubrawski, A. (1995). "HPC strength prediction using artificial neural network." J. Comput. Civ. Eng., 9(4), 279–284.
11. A. Fazel Zarandi, M. H., Türksen, I. B., Sobhani, J., and Ramezanianpour, (2008). "Fuzzy polynomial neural networks for approximation of the compressive strength of concrete." Appl. Soft Comput., 8(1), 488–498.
12. Trtnik, G., Kavcic, F., and Turk, G. (2009). "Prediction of concrete strength using ultrasonic pulse velocity and artificial neural networks." Ultrasonics, 49(1), 53–60.
13. Gupta, R., Kewalramani, M. A., and Goel, A. (2006). "Prediction of concrete strength using neural-expert system." J. Mater. Civ. Eng.,18(3), 462–466.
14. Amir Hossein Alavi Amir Hossein Gandomi,"A robust data mining approach for formulation of geotechnical engineering systems", Engineering Computations, Vol. 28 Iss 3 pp. 242 – 274, 2011
15. Mehrnoosh Ebrahimi, and Ali Akbar Niknafs, "Increasing Cement Strength Using Data Mining Techniques," International Conference Data Mining, Civil and Mechanical Engineering (ICDMCME'2015) Feb. 1- 2,Bali, Indonesia,2015
16. Jui-Sheng Chou, Chih-Fong Tsai, "Concrete compressive strength analysis using a combined classification and Regression technique," Department of Construction Engineering, National Taiwan University of Science and Technology, 43 Sec.4, Keelung Road, Taipei, Taiwan, Automation in Construction 24 (2012)
17. Ramin Hosseini Kupaei, U. Johnson Alengaram, and Mohd Zamin Jumaat, "The Effect of Different Parameters on the Development of Compressive Strength of Oil Palm Shell Geopolymer Concrete," Department of Civil Engineering, Faculty of Engineering, University of Malaya, 50603 Kuala Lumpur, Malaysia, Received 1 July 2014; Accepted 27 August 2014; Published 28 October 2014
18. Jui-Sheng Chou, Chien-Kuo Chiu, Mahmoud Farfoura, Ismail Al- Taharwa, "Optimizing the Prediction Accuracy of Concrete Compressive Strength Based on a Comparison of Data-Mining Techniques," JOURNAL OF COMPUTING IN CIVIL ENGINEERING © ASCE / MAY/JUNE 2011 / 253
19. Rakesh A. More* and Prof. S.K. Dubey, "Effect of Different Types of Water on Compressive Strength of Concrete," International Journal on Emerging Technologies 5(2): 40-50(2014)