# USING DATA MINING TECHNIQUES IN CATALYTIC CRACKING PROCESS MODELING – A COMPARATIVE STUDY

**MARINOIU CRISTIAN**
*ASSOC. PROF.PH.D., PETROLEUM-GAS UNIVERSITY OF PLOIEŞTI*
*e-mail:marinoiu_c@yahoo.com*

***Abstract***

*Despite notable successes achieved in trying to obtain viable alternatives to the use of oil as a source of energy and raw materials, oil industry remains one of the main pillars of sustainable development of modern society . Continued modernization of the industry by using the latest technologies in all its important links - extraction, transport, refining - led to substantial reduction of its negative impact on the environment. Particularly, refineries benefited from a special technological contribution, which resulted in a significant plus in efficiency and reliability. In order to always maintain these high standards, the control of oil refining chemical processes must be based on simple but at the same time performant mathematical models. One way to achieve this objective is the use of modern data mining methods. In this paper we propose to compare the performance of three data mining methods in order to be used in catalytic cracking process modeling in a refinery.*

***Key words:*** *data mining , catalytic cracking, regularization, multicollinearity, overfitting,support vector machine*
***JEL Classification*** *: L71, C63*

## 1. Introduction

Since it was officially launched in 1972 at the UN Conference on Human Development in Stockholm [13], the concept of sustainable development is always in the attention of world's most important meetings on development. The 2030 Agenda for Sustenable Development [14] , adopted by United Nations General Assembly on September 25, 2015 sets the targets to be achieved for the realization of this concept. Industry and particularly the oil and gas industry had and still have an important role in shaping the world in all its essential economic and social aspects. In this way, according to [2, p.15], even if the growth rate of renewable energies and of that provided by nuclear and hydroelectric power is very high " oil, gas and coal remain the dominant source of energy powering the world for more than three-quarters of total energy supply in 2035". Petroleum products are vital not only in terms of their potential energy, but also for the petrochemical industry. These products are a result of the refining process that has three major steps: separation, conversion and treating. An important aspect of the conversion process is the cracking process, which can be done in two ways:

- cracking that takes place at low temperature and pressure, but in the presence of a catalyst, known as catalytic cracking;
- cracking that does not use catalyst but which is carried out at high temperatures and pressures, called pyrolisis or thermal decomposition.

In this paper we consider modeling catalytic cracking process because the efficient control of this chemical process, with potential financial consequences and environmental major problems, depends largely on its proper modeling.

## 2. The problem of modeling the catalytic cracking process

Over time, various designs have been proposed [1] [6], which attempt to accurately reflect the complexity of the catalytic cracking process, which is characterized by [5]:

- the multitude of chemical agents from various classes of hydrocarbons;
- the reaction kinetics, which includes primary and secondary reactions;

- the catalytic mechanism of reaction in which the activity and selectivity of the catalyst change rapidly and continuously.

Thus, kinetic models are among the most commonly used mathematical models proposed in the literature for catalytic cracking process modeling, from which we mention [7]: Weekman's model, the Gianneto model and the Mobile model. Among these we remark the Weekman's model, which is one of the first and most commonly used kinetic model [6 p.4], being considered both a „simple and robust" model [7]. Other classes of mathematical models, such as the model based on catalyst deactivation, riser model, regenerator model, fluid cracking reactor-regenerator model, take into account a structural, component based approach, of the process, which improves the accuracy of process modeling.However, the main criticism of these mathematical models, especially of kinetic models refers to the difficulty to closely reflect the complexity of the phenomenon in industrial conditions. This happens due to the presence of a large number of reactants and due to the significant computing effort, necessary both for identifying parameters and for simulations [6]. If properly used, data mining techniques can be a viable alternative to the models mentioned above due to their ability to detect patterns of complex phenomena from a set of available observations. Moreover, their simplicity represents a further advantage.

Therefore, in [5] and [4] simplified mathematical models have been proposed, based on multiple regression, respectively Ridge regression. In this paper we use three modern data mining techniques for modeling - LASSO (Least Absolute Shrinkage and Selection), SVR (Support Vector Regression) and PCR (Principal Components Regression) and we compare their performances. The study is based on data from the paper [1] and organized in a convenient form in Table 1.

**Table 1**. The observed values of the input and output variables of a catalytic cracking process

| No. | $y$ | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ | $X_7$ |
|---|---|---|---|---|---|---|---|---|
| 1 | 52.3 | 0.9007 | 442.0 | 0.38 | 183.4 | 316 | 732.0 | 4.6 |
| 2 | 52.8 | 0.9029 | 441.5 | 0.25 | 183.1 | 311 | 730.0 | 4.5 |
| 3 | 52.8 | 0.9028 | 434.4 | 0.25 | 184.3 | 310 | 732.0 | 4.7 |
| 4 | 51.4 | 0.9043 | 448.6 | 0.29 | 189.7 | 310 | 725.0 | 4.6 |
| 5 | 52.4 | 0.9009 | 442.5 | 0.38 | 183.8 | 320 | 731.0 | 4.7 |
| 6 | 52.1 | 0.9039 | 440.0 | 0.25 | 182.4 | 310 | 734.0 | 4.5 |
| 7 | 52.8 | 0.9042 | 445.8 | 0.38 | 182.6 | 312 | 728.5 | 4.6 |
| 8 | 52.2 | 0.9050 | 445.0 | 0.32 | 183.7 | 319 | 733.0 | 4.5 |
| 9 | 52.8 | 0.9007 | 436.8 | 0.39 | 182.8 | 315 | 732.0 | 4.6 |
| 10 | 51.8 | 0.9014 | 440.2 | 0.28 | 182.7 | 316 | 733.0 | 4.5 |
| 11 | 52.3 | 0.9004 | 443.2 | 0.49 | 187.9 | 316 | 726.0 | 4.5 |
| 12 | 52.0 | 0.9020 | 436.0 | 0.23 | 191.1 | 324 | 734.0 | 4.4 |
| 13 | 53.0 | 0.9030 | 441.5 | 0.25 | 184.6 | 311 | 733.0 | 4.9 |
| 14 | 51.3 | 0.9068 | 449.6 | 0.43 | 182.2 | 314 | 727.0 | 4.6 |
| 15 | 52.7 | 0.9033 | 424.4 | 0.36 | 182.7 | 312 | 732.0 | 4.5 |
| 16 | 43.7 | 0.9217 | 438.2 | 2.14 | 173.6 | 314 | 727.5 | 4.8 |
| 17 | 45.4 | 0.9247 | 438.4 | 2.19 | 188.7 | 319 | 727.0 | 5.0 |

Source:[1]

The data from table 1 represent a selection volume of 17 from a series of observations made over a period of 90 days on the following characteristic variables for the catalytic cracking process: $y$ (gas productivity), $X_1$(density), $X_2$(volumetric temperature), $X_3$(sulphur content), $X_4$(feedstock flow) $X_5$(output heater feedstock temperature), $X_6$(catalyst temperature in regenerator system), $X_7$(catalyst /feedstock ratio). The variables $X_1$, $X_2$, $X_3$ reflect the characteristics of the raw materials that is processed, while the variables $X_4$, $X_5$, $X_6$, $X_7$ represent control variables and the variable $y$ is an output variable. Basically, the problem is to find a function $f$ so that the output variable $y$ could be better approximated depending on the control variables $X_4$, $X_5$, $X_6$, $X_7$ and on the variables $X_1$, $X_2$, $X_3$ which reflect the characteristics of raw material. An alternative is to try to

represent this approximation by a linear function $f$, determined by using the multiple linear regression model

$$y = \beta_0 + \beta_1 X_1 + \cdots + \beta_7 X_7 + \varepsilon \,, \quad (1)$$

where $\beta_0, \beta_1, \cdots, \beta_7$ are the parameters of the model, and $\varepsilon$ is the additive error of the model.

Because some of the variables $X_1, X_2, \cdots, X_7$ are strongly correlated, in order to solve the multiple regression model, the matrix of the regressor variables $X_1, X_2, \cdots, X_7$ is affected by the phenomenon of multicollinearity. The multicollinearity strongly distorts the stability of the estimations of the regression coefficients $\beta_i$ by affecting the credibility of the obtained model [15]. For this reason, in [4], the Ridge regression [12] was used, which is a regularization method that avoids the negative consequences of multicollinearity. In the following paragraphs we take into account the use of another regularization method, the LASSO method, as well as two other data mining techniques SVR [9] and PCR [3], and we compare them from the point of view of their performances with the help of *RMSE* (Root Mean Squared Error ) indicator, defined as :

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

where $y_i$is the observed value, and $\hat{y}_i$ is the correspondent value predicted by the model.

For this purpose, two-thirds of the data (11 rows randomly selected from Table 1) were used as training data to build models, and the remaining third was used as necessary test data in order to calculate the *RMSE*. For facilitating the presentation we will make the following notations:

- *Idt* the set of the 11 indicators randomly sampled from the set of indices $\{1,2,\ldots,17\}$ which represent the selected lines from table 1 as training data. For the case analysed in this paper, the set of randomly selected indices is: $Idt = \{1,2,3,5,6,7,9,10,11,12,13,14\}$
- $\boldsymbol{X}(11x8)$ is the matrix of training data: a line from matrix $\boldsymbol{X}$ is a vector of $1x8$ dimension under the form $(1, \boldsymbol{x}_i^T), i \epsilon Idt$, where $\boldsymbol{x}_i$ is the column vector which contains the values of the variables $X_1$, $X_2$, $X_3$, $X_4$, $X_5$, $X_6$, $X_7$ stored in the $i$ line of table 1, and $\boldsymbol{x}_i^T$ is the transpose of the $\boldsymbol{x}_i$ vector;
- $\boldsymbol{y}(11x1)$ is the column vector of training data of components $\boldsymbol{y_i}$, $i \epsilon Idt$ ; each component $\boldsymbol{y_i}$ contains the value of the variable $y$, which is stored in line i of table 1;
- $\boldsymbol{\beta}(7x1)$ is the vector of the parameters of the model.

## 3. Using LASSO method

Basically, in Ridge regression the effect of the multicolinearity is removed by penalizing the multiple regression linear model (1). The penalty function is the Euclidean norm or $l_2$ of the vector $\boldsymbol{\beta}$, that is $\|\boldsymbol{\beta}\|_{l_2} = \sqrt{\beta_1^2 + \beta_2^2 + \cdots + \beta_p^2}$. The result of using this method is shrinking to zero the values of the regression estimated coefficients $\hat{\beta}_1, \hat{\beta}_2, \cdots \hat{\beta}_7$, without any of coefficients $\hat{\beta}_i$ being reduced to zero value. The advantage of using LASSO method developed in [10] lies in the fact that it also performs shrinking to zero of the coefficients of regression, but, unlike the Ridge regression method, a part of the coefficients are zero valued. In this case a selection of the involved regression variables is obtained and finally a more easily interpretable model.

The method consists in estimating the coefficients $\beta_i$ of the model by solving the following optimization problem,

$$\min_{\boldsymbol{\beta}}(\|\boldsymbol{y} - \boldsymbol{X\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|_{l_1}),\ \lambda>0, \quad (3)$$

where the penalty function is the norm $l_1$ or the Manhattan norm of the vector $\boldsymbol{\beta}$, meaning $\|\boldsymbol{\beta}\|_{l_1}=|\beta_1|+|\beta_2|+\cdots+|\beta_p|$, and $\lambda$ is the tuning parameter which controls the degree of penalty. The larger $\lambda$ is, the higher the number of coefficients set to zero and the smaller the absolute value of the nonzero coefficients. At the limit , when $\lambda=0$ the problem (3) is reduced to the problem of solving the multiple linear regression model, and when $\lambda=\infty$, the solution is $\widehat{\boldsymbol{\beta}}_{\boldsymbol{LASSO}}=\mathbf{0}$. The value of the parameter $\lambda$ is chosen in order to satisfy a certain optimality criterion. A frequently used criterion is to choose as an optimal value for $\lambda$ that value $\lambda min$ for which the mean cross-validated error is minimum.

In order to solve the problem (3) we use functions from the software package *glmnet* from *R* [8]. In figure 1 the graph of the cross-validated curve is presented, surrounded by curves generated by standard deviations for each value $\lambda$. The dotted vertical bars indicate an optimal value $\lambda$ for which the cross-validated error is minimal ( $\lambda min= 0.093$, $log\ (\lambda min)=-2.375$)), and , on the superior horizontal axis we find the number of nonzero coefficients $\hat{\beta}_i$, namely 4. Those are: $\hat{\beta}_0 = 101.85224688, \hat{\beta}_1 = -43.17604951,\quad \hat{\beta}_2 = -0.01817084\ ,\quad \hat{\beta}_5 = -0.02225328,\ \hat{\beta}_7 = 0.97913214$ .

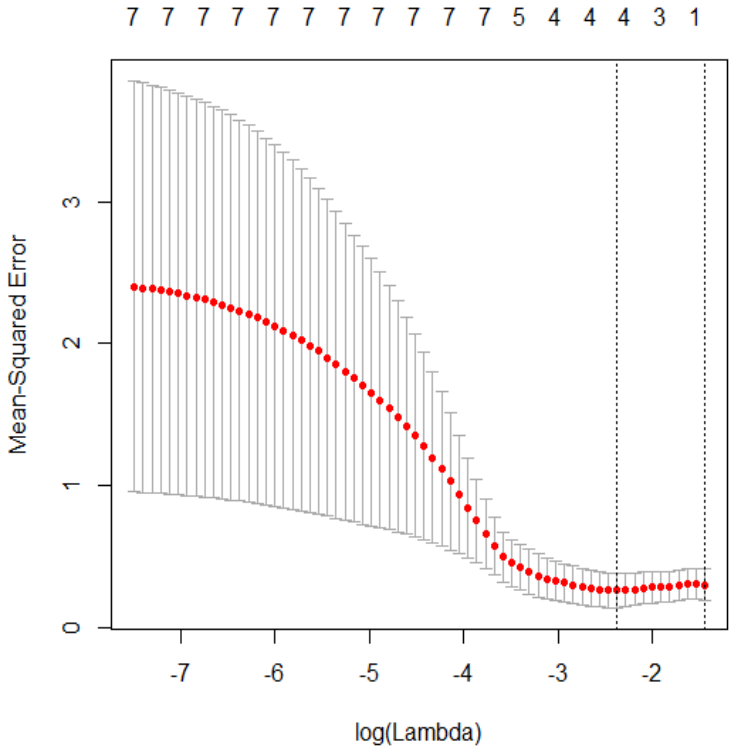


**Figure 1.** Graph of the cross-validated curve. Source: made by the author using functions from R

The predicted values for the test data are: $52.60269, 52.26220,\ 51.80672, 51.75812, 51.99920,$ and $RMSE = 4.62467$.

## 4. Using SVR method

The SVR method is the regression version (developed by Vapnik in the work [11]) of the SVM (Support Vector Machines) method. The variant called $\varepsilon-$Support Vector Regression allowes the determination of a function $f$ of form $f(x)=\beta_0+\boldsymbol{\beta}^T\boldsymbol{x}$ provided that the permissible deviation in

the point $\boldsymbol{x} = \boldsymbol{x_i}$ to $y_i$ is not more than $\varepsilon$. This approach leads to solving a convex optimization problem [9] :

$$\min_{\boldsymbol{\beta}} \frac{1}{2}\|\boldsymbol{\beta}\|^2 \quad (4)$$

with the restrictions:

$$|y_i - \beta_0 - \boldsymbol{\beta^T x_i}| \leq \varepsilon$$

One way to avoid situations in which problem (4) is not feasible is to accept that one part of training data has the error greater than $\varepsilon$, as reflected mathematically by introducing in the model the variables $z_i$ and $z_i^*$ called slack variables. In this way the problem (4) is

$$\min_{\boldsymbol{\beta}}\left(\frac{1}{2}\|\boldsymbol{\beta}\|^2 + C\sum_{i=1}^{l}(z_i + z_i^*)\right) \quad (5)$$

with the restrictions:

$$y_i - \beta_0 - \boldsymbol{\beta^T x_i} \leq \varepsilon + z_i,\ z_i \geq 0$$
$$\beta_0 + \boldsymbol{\beta^T x_i} - y_i \leq \varepsilon + z_i^*,\ z_i^* \geq 0$$
$$C > 0,\ z_i \geq 0,\ z_i^* \geq 0,$$

where the value of $C$ parameter is chosen so that it reflects the compromise between the requirement that the function $f$ is a smooth function and the requirement to not greatly exceed the stated value for the error $\varepsilon$. On the other hand, minimizing the norm of $\beta$ parameter has the effect of reducing the complexity of the model and and thus avoiding the phenomenon of overfitting.

The problem (5) can be easily solved in its dual variant, which allows the extension to the non-linear variant by introducing different models of kernels $K$, in the model, including the following:

- linear kernel, $K(x_1, x_2) = x_1^T x$
- gaussian (radial basis function –rbs), $K(x_1, x_2) = \exp(-\gamma\|\mathrm{x}_1 - \mathrm{x}_2\|^2)$
- polinomial, $K(x_1, x_2) = (1 + x_1^T x)^k, k\epsilon\{2,3\cdots\}$
- sigmoid, $K(x_1, x_2) = \tanh(\gamma x_1^T x + C)$

For our problem we chose the kernel *rbs*, and the optimal values for γ and $C$ were selected with the help of the function *svm_tune* from the package *e1071* of *R*. This function allows the automatic selection of the best performant model from a multitude of SVR models generated by the variation of $C$ and γ parameters, through a method of exhaustive grid search. For each model we calculate (by using the cross-validated method) the mean square error. The model which has the lowest mean square error is selected as being the best model. For the $\varepsilon$ parameter the default value is used from the function *svm*_tune, meaning $\varepsilon = 1/data\ size = 1/11 = 0.1$ Therefore, the model having the lowest mean square error and the parameters $\gamma = 7$, $C = 0.1$, $\varepsilon = 0.1$ is selected. With this model, the predicted values on test data are $52.41891\ 52.39006\ 52.08078\ 52.08078\ 51.80972$, and $RMSE = 4.818338$. Figure 2 represents the graph of the values of the mean square errors calculated on the grid generated by the variation of γ and $C$ parameters. According to the legend, the darkest areas represent the lowest values of the mean square error.

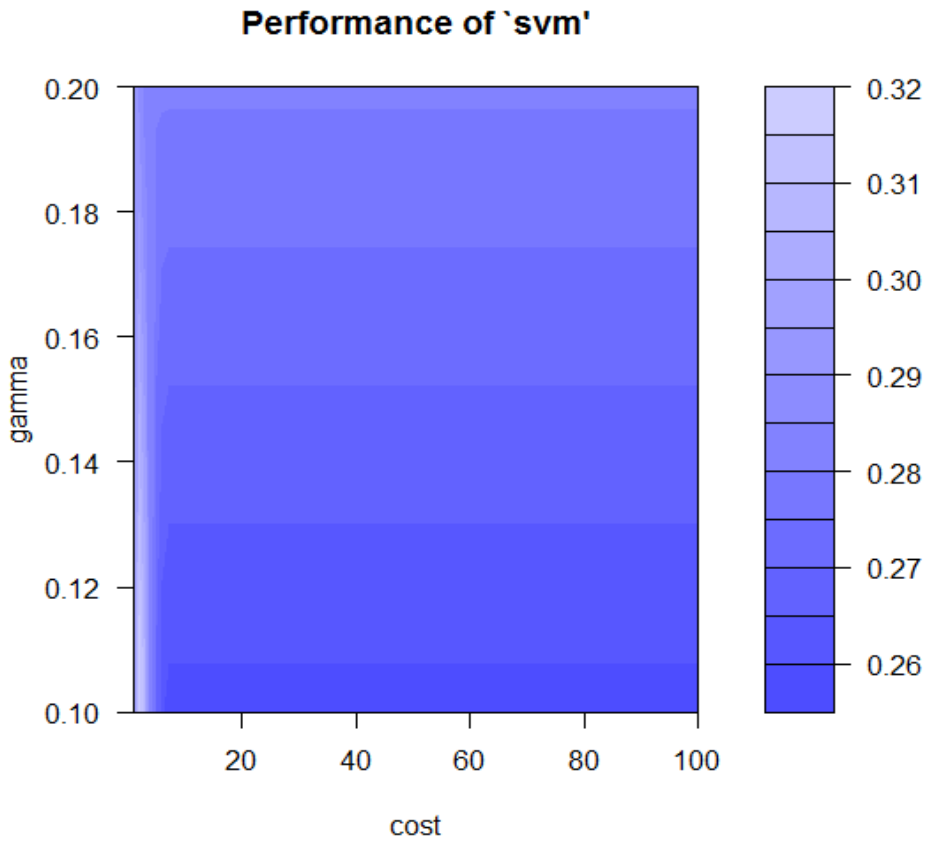


**Figure 2.** Performance of SVR models for different values of $C$ and γ parameters.
Source: made by the author using functions from R

## 5. Using PCR method

Let we denote by $Z_1, Z_2, \cdots, Z_m$ the first $m$ main components of matrix $\boldsymbol{X}$, $m < 7$. If a large part from data variation is explained by these components and if they have a high degree of correlation with the dependent variable $y$, then the linear regression model [3],

$$y = \theta_0 + \theta_1 Z_1 + \cdots + \theta_m Z_m + \varepsilon, \quad (6)$$

also called principal components regression, may be a better choice than (1), at least for the following reasons:

- the principal components $Z_1, Z_2, \cdots, Z_m$ are uncorrelated and the danger of multicolinearity is removed;
- the obtained model has fewer variables and consequently the risk of overfitting is diminished.

Table 2 shows the estimated values of prediction error using RMSEP (Root Mean Squared Error of Prediction) made by the cross-validated method and the cumulative percentage of variance explained according to the first $m$, $m = 1,2 \cdots 7$ main components of the data for the catalytic craking model.

**Table 2**. The *RMSE* values and the cumulative percentage of variance explained according to the number of main components

| **Components number** | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| **RMSEP values** | 0.5567 | 0.7045 | 0.6177 | 0.7045 | 0.5614 | 0.6428 | 1.539 |
| **% Variance explained** | 38.82 | 67.58 | 81.43 | 90.47 | 97.00 | 99.02 | 100.00 |

Source: made by the author using results from R

Figure 3 shows the graph of the values RMSEP estimated by the cross-validation (CV) and by bias cross-validation method (adjCV). Given that, for the number of components $m = 5$, we have a RMSEP value close to its minimum value and, at the same time, the first five main components explain 97% of data variation, we consider that the value $m = 5$ is a good choice for our model.

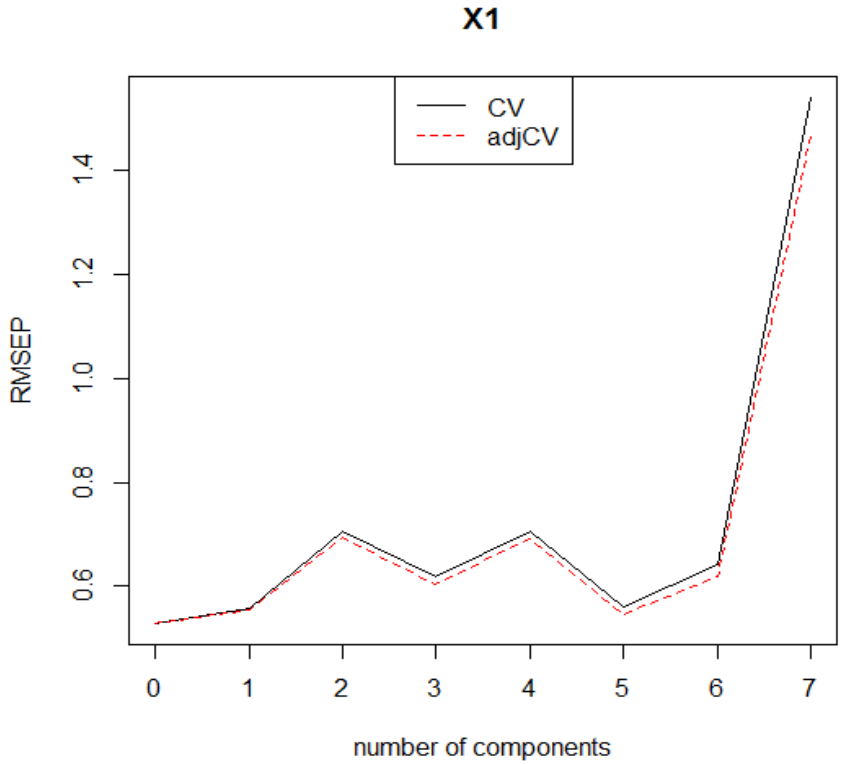


**Figure 3**. The graph of cross-validation estimate curve (continuous line-CV) and of bias cross-validation estimate curve (dashed line-adjCV). Source: made by the author using functions from R

With the choice made ($m = 5$), the predictions of the model PCR, obtained on test data, are: $53.07053,\ 52.48596, 51.18058, 51.42133, 51.44710$, and $RMSE = 4.338177$.

## 6 . Conclusions and future work

Given the importance of effective and safe control of the catalytic cracking process in a refinery, the mathematical model of the process used for this purpose should be simple and should accurately reflect its basic characteristics. In this paper we proposed to use three modern data mining techniques to build such models, starting from a set of observations made on the input and output variables of a catalytic cracking process.

In order to obtain the three models the same training data and test data were used for the performance evaluation. The values of the performance indicator used- *RMSE* are very similar in magnitude: 4.62467 for LASSO method, 4.818338 for SVR method and 4.338177 for the PCR method. For this reason we consider that all three analyzed methods provide similar models regarding the performance, the best being, from this point of view, the model provided by PCR method, which has the lowest *RMSE*. On the other hand, if we want to also have an easily interpretable model, the LASSO method can be a good choice because it also provides a simple

model of the dependence between the output variable $y$ and the input variables $X_i$.That is : $y = 101.85 - 43.18X_1 - 0.02X_2 - 0.02X_5 + 0.98X_7$.

The study started in this paper can be extended by using other data mining techniques in modeling catalytic cracking process. In this regard, in the future we intend to test the efficacy of using neural networks and ElasticNet methods. These techniques look promising since they have the ability to capture the depth of the nonlinear character of this process. Also, the choice of ElasticNet method is motivated by the fact that it is an extension of already used Ridge and LASSO methods: the penalty function that occurs here is a linear combination of the penalty functions of these methods controlled by a parameter$\alpha\epsilon[0,1]$. Because for $\alpha = 0$ the method is reduced to Ridge regression, and for $\alpha = 1$ to the LASSO method, the study will focus in this case on determining the optimal parameter value $\alpha\epsilon(0,1)$, for the considered data.

## Bibliography

**[1] Belitzianis, M**., *Conducerea evoluată a sistemului reactor-regenerator din instalaţia de cracare catalitică*, Teză de doctorat, Universitatea Petrol-Gaze din Ploieşti, 1992

**[2] BP Energy Outlook,** 2017 Edition, *Energy overview - the base case,* available at http://www.bp.com/en/global/corporate/energy-economics/energy-outlook/energy-overview-the-base-case.html [accessed on 13.07.2017]

**[3] Hastie, T., Tibshirani, R, Friedman, J.,**2009, *The elements of statistical learning, Data mining, inference and prediction*, Second edition, Springer Series in Statistics

**[4] Marinoiu, C**., *A Ridge regression model of the cracking process,* Buletinul Universităţii Petrol-Gaze din Ploieşti, Seria Matematică, Informatică şi Fizică, no.1/2009, pp. 65-70

**[5] Pătrăscioiu, C., Marinoiu, C**., *Soluţii numerice pentru modelarea statistică a procesului de cracare catalitică,* Revista de Informatică economică, nr. 10/1999, pp. 53-61

**[6] Pinheiro, C.I.C., Fernandes J. L., Domingues, L., Chambel A. J. S., Graca, I., Oliveira N.M.C., Cerqueira, H. S., Ribeiro, F.R.,** *Fluid Catalytic Cracking (FCC) Process Modeling ,Simulation and Control***,** Industrial & Engineering Chemistry Research, 2012, 51, pp.1-29

**[7] Popa, C, Pătrăşcioiu, C.,** *The model predictive control system for the fluid catalytic cracking unit*, Advances in dynamical systems and control, 2011, pp. 95-100, available at https://www.researchgate.net/publication/228953174_The_model_predictive_control_system_for_the_fluid_catalytic_cracking_unit [accessed on 15.08..2017]

**[8] R Core Team** , 2017, *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, URL https://www.R-project.org

**[9]Smola, A. J., and Şcholkopf, B**., *A tutorial on support vector regression*,no.14, pp.199-222,2004, Kluwer Academic Publishers

**[10] Tibshirani, R**., 1996, *Regression Shrinkage and Selection via LASSO*, Journal of the Royal Statistical Society, Serie B (Methodological), Volume 58, ISSUE 1 , 267-288

**[11] Vapnik, V**. , *The nature of statistical learning theory*, Springer, 1995

**[12] Vinod, H. U**., *Recent advances in regression methods*, Marcel Dekker, New Zork, 1981

**[13]United Nations** -Sustainable Knowledge Platform, *United Nations Conference on the Human Environment (Stockholm Conference)***,** available at https://sustainabledevelopment.un.org/milestones/humanenvironment**,** accessed on 13.07.2017]

**[14] United Nations** -Sustainable Knowledge Platform, *Transforming our world: the 2030 Agenda for Sustainable Development* available at https://sustainabledevelopment.un.org/post2015/transformingourworld [accessed on 13.07.2017]

**[15**] *** *Multicollinearity*, available at http://www.statisticssolutions.com/multicollinearity/, [accessed on 13.07.2017]