A comparative study on machine learning regression algorithms aplied to modeling gas centrifuge
/
Um estudo comparativo sobre algoritmos de regressão de aprendizagem de máquinas aplicado à modelagem de centrífugas a gás

DOI:10.34117/bjdv8n7-265

Recebimento dos originais: 23/05/2022
Aceitação para publicação: 30/06/2022

Delvonei Alves de Andrade
Doutor em Tecnologia Nuclear
Institution: Universidade de São Paulo (USP)
Address: Avenida Lineu Prestes, 2242, Cidade Universitária, São Paulo – SP, CEP: 05508-000
E-mail: delvonei@ipen.br

Roberto Navarro de Mesquita
Doutor em Engenharia Mecânica
Institution: Universidade de São Paulo (USP)
Address: Avenida Lineu Prestes, 2242, Cidade Universitária, São Paulo – SP, CEP: 05508-000

Natan Patussi Nascimento
Engenharia Química
Institution: Universidade de São Paulo (USP)
Address: Avenida Lineu Prestes, 2242, Cidade Universitária, São Paulo – SP, CEP: 05508-000
E-mail: patussi@marinha.mil.br

ABSTRACT
The gas Centrifuge is a very hard equipment to model, because it involves a gas dynamic with many complications, such as hypersonic waves and rarefied regions combined with continuous flow areas. Therefore, data analysis regressions remain currently a very important technique to understand and describe the problem in a practical way. This paper intends to apply and compare several regression techniques using machine learning, to obtain a hydraulic and a separative power model of gas centrifuge used in enrichment plants. For this purpose, a set of normalized data composed of 134 experimental lines was used, observing the variables of interest, the separation power (dU), and the waste pressure (Pw), through the following explanatory variables: feed flow (F), cut (q), and product pressure (Pp). The comparisons were presented between the results obtained for the models generated by the following: algorithms, multivariate regression, multivariate adaptive regression splines – MARS, bootstrap aggregating multivariate adaptive regression splines – Bagging MARS, artificial neural network – ANN, extreme gradient boosting – XGBoost, support vector regression– Poly SVR, radial basis Function support vector regression – RBF SVR, K-nearest neighbors – KNN and Stacked Ensemble. That way, to avoid overfitting and provide insights about generalization of the models in unseen data, during the training phase, the k-fold cross validation approach was used. Subsequently, the residuals were analyzed, and the models were compared by the

following metrics: Root mean square error – RMSE; Mean squared error – MSE; Mean absolute error – MAE; and Coefficient of determination – R2.

Keywords: gas centrifuge, Uranium enrichment, machine learning, multivariate regression, xgboost, artificial neural network, support vector machine, spline, k- nearest neighbors, multivariate adaptive regression splines

RESUMO

A Centrífuga a gás é um equipamento muito difícil de modelar, pois envolve uma dinâmica de gás com muitas complicações, tais como ondas hipersônicas e regiões raras combinadas com áreas de fluxo contínuo. Portanto, as regressões da análise de dados continuam sendo atualmente uma técnica muito importante para compreender e descrever o problema de forma prática. Este trabalho pretende aplicar e comparar várias técnicas de regressão usando o aprendizado de máquinas, para obter um modelo hidráulico e um modelo de potência separadora da centrífuga a gás usada em plantas de enriquecimento. Para este fim, foi utilizado um conjunto de dados normalizados composto de 134 linhas experimentais, observando as variáveis de interesse, a potência de separação (dU), e a pressão do resíduo (Pw), através das seguintes variáveis explicativas: fluxo de alimentação (F), corte (q), e pressão do produto (Pp). As comparações foram apresentadas entre os resultados obtidos para os modelos gerados pelos seguintes: algoritmos, regressão multivariada, splines de regressão adaptativa multivariada - MARS, bootstrap agregando splines de regressão adaptativa multivariada - Bagging MARS, rede neural artificial - ANN, reforço de gradiente extremo - XGBoost, regressão vetorial de suporte - Poly SVR, base radial Função de suporte de regressão vetorial - RBF SVR, K-nearest vizinhos - KNN e Stacked Ensemble. Desse modo, para evitar o ajuste excessivo e fornecer informações sobre a generalização dos modelos em dados não vistos, durante a fase de treinamento, foi utilizada a abordagem de validação cruzada k-fold. Posteriormente, os resíduos foram analisados, e os modelos foram comparados pelas seguintes métricas: Erro quadrático médio - RMSE; Erro quadrático médio - MSE; Erro absoluto médio - MAE; e Coeficiente de determinação - R2.

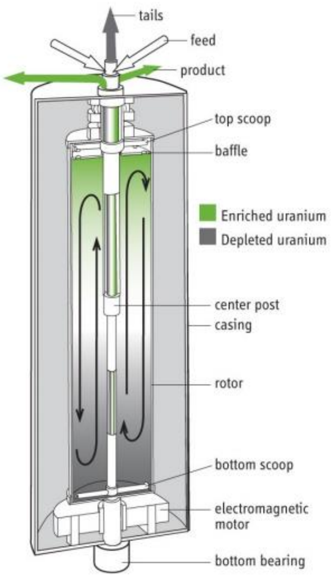
Palavras-chave: centrífuga a gás, enriquecimento de Urânio, aprendizagem da máquina, regressão multivariada, xgboost, rede neural artificial, máquina vetorial de apoio, estribo, k - vizinhos mais próximos, estribo de regressão adaptativa multivariada.

1 INTRODUCTION

Natural uranium compounds are mostly made up of U^{238} isotopes, however, the U^{235} isotope, which represents only 0.71% of its composition, is the isotope of greatest commercial interest. It happens because it is fissile when bombarded by slow neutrons, releasing an enormous amount of energy, which is converted into electricity in nuclear plants. That said, one of the most important stages of the nuclear fuel cycle is the enrichment process, which increases the proportion of U^{235} to usual concentrations between 2% and 5% for application in commercial reactors, like Pressurized Water Reactor - PWR and Boiling Water Reactor – BWR.

Currently, there are several technologies for enriching uranium; however, the gas centrifuge process remains the main way. The isotopic separation by gas centrifuge occurs with uranium hexafluoride – UF_6. In the process, the feed current is introduced in the cylindrical equipment, which rotates at high speed, and can generate a density gradient, where components with higher molecular weights accumulate in regions close to the cylinder wall. On the other hand, the lighter components accumulate in regions more distant from the wall, making it possible to extract two outputs currents, one more enriched (product) and the other impoverished (waste) in the isotope of interest ($^{235}UF_6$)[2]. In addition to the radial effect, it induces a vertical countercurrent flow, obtaining a multiplication of the elementary effect of radial separation. So, the difference between the top and bottom composition of the centrifuge rotor becomes greater than the difference in the radial direction for a given axial position. A countercurrent is generated by two basic mechanisms: a mechanical and a thermal one. The first occurs due to the positioning of a stationary collector at one end of the rotor and by a rotating plate at the other end; the second occurs due to a temperature gradient along the rotor wall, heating one end and cooling the other[3]. Figure 1 below shows the scheme of a gas centrifuge.

Figure 1: Scheme of a Gas Centrifuge[4].



A fluid dynamic analysis of the internal flow in a gas centrifuge is an extremely wide nonlinear problem, which may cover high vacuum regions, continuous regions, as well as transition regions. For this reason, this work will focus on empirical models of gas centrifuge by data analysis using algorithms of supervised machine learning regression.

In this sense, for the separative problem, the separative power – dU, was chosen as dependent variable. That represents the minimum energy required to obtain the separation of the inlet flux with a known mixed concentration, into two fluxes with different composition[5], so that the separative power could be defined as follows[6]:

$$\delta U = PG(c_P) + WG(c_W) - FG(c_F) \tag{1}$$

F, P and W are the fluxes of feed, product, and waste respectively and $G(c) = (2c - 1)\ln[c/(1 - c)]$ is the separative potential introduced by Fuchs and Peierls7. Therefore, it could be calculated by measuring the fluxes of P, W and F in each composition.

Another important dependent variable must be chosen to evaluate the hydraulic answer of the equipment when operating in cascade. For this reason, the wasting pressure – Pw also was chosen as a dependent variable.

The explanatory variables selected to this work were feed flow – F, cut – q, defined as a ratio between product and feed flux – P/F and product pressure – P_P. Other variables like temperature, scoop distances, feed position, and baffle size, were maintained unchanged during the experiments. Consequently, separative power and pressure of waste flux can be written as follows:

$$\delta U = f(P_P, \theta, F) \tag{2}$$

$$P_W = f(P_P, \theta, F) \tag{3}$$

2 RELATED WORKS

The first work using machine learning to model and optimize gas centrifuge was made by MIGLIAVACCA (1999)[8], who used artificial neural networks – ANN for this purpose. Years after, ANDRADE (2004)[9] utilized ANN for the detection of gross errors in process data.

In 2005, ANDRADE and MIGLIAVACCA (2005)[10,11] modeled the separative power of gas centrifuges by using multivariate regression with a covariance matrix. That technique has some advantages, because it is easily interpretable and transportable to other program languages, leading to an empirical equation.

In the same year, CRUS (2005) [12] presented a work of modeling the separative parameters via hybrid neural networks, which established a correlation between the variables applied in practice with the theoretical parameters of the model that describe the internal flow of the gas centrifuge.

Thereby, previous works that used data analysis to model gas centrifuge was restricted to the use of artificial neural networks and multivariate regressions, so this work has innovations in the application of a greater variability of machine learning techniques and a comparison between the results.

3 METODOLOGY

The data set used in this work was extracted from MIGLIAVACCA (1999)[8], which is composed of 134 normalized lines obtained in the isotopic separation process, so that the integrity of information and details about the process are preserved, as shown in Table 1.

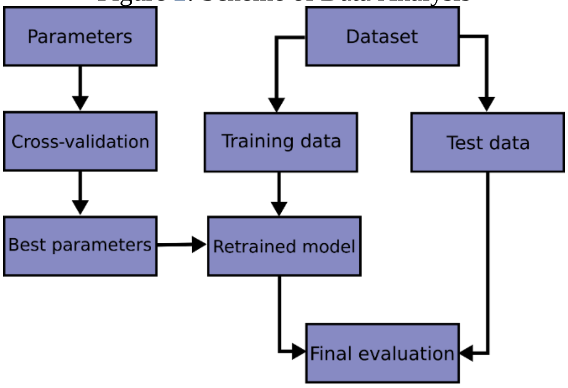
Table 1: Data Set

F	P_P	q	P_W	dU
0,1000	0,2330	0,5317	0,3857	0,6446
0,2338	0,2343	0,5221	0,4205	0,6715
0,6331	0,2333	0,5355	0,4950	0,8830
0,2338	0,3335	0,4993	0,4604	0,7240
0,3669	0,3321	0,5345	0,4805	0,7714
...	...	...	...	...

The program used in the data analysis was the free software Rstudio®[13], developed to create an environment for statistical computing and graphics. Besides, there are specific functions and packages that facilitate the practice of modeling via machine learning, such as Tidymodels[14] and Tidyverse[15], which promote practicalities in the resampling, cross-validation and hyperparameter optimization of the algorithms.

Initially, the database was splitted into two parts, 80% for training and 20% for testing. Then, to avoid overfitting, in order to get insight about how the models will generalize in unseen data and reach a better relationship between bias and variance, the k-fold cross validation approach to choose the best hyperparameters was used, by selecting the hyperparameters which achieved better metric of root mean square error – RMSE. The scheme in Figure 2 describes the training and evaluation process for each model developed.

Figure 2: Scheme of Data Analysis



4 REGRESSION TECHNIQUES

In this paper, nine regression techniques were used, according to the following algorithms: a) Multivariate polynomial regression; b) Multivariate adaptive regression splines – MARS [16]; c) Multivariate ad

aptive regression splines coupled with the bagging technique - Bagging MARS [17]; d) K- Nearest Neighbors – KNN [18]; e) Artificial Neural Networks – ANN [19]; f) Polynomial Kernel Function Support Vector Regression – Poly SVR [20]; g) Radial Basis Function Support Vector Regression – RBF SVR [21]; h) Extreme Gradient Boost Machine – XGBoost [22]; i) Stacked Ensemble [19], in which all the previous algorithms were used as weak learners.

5 RESULTS AND DISCUSSION

To better analyze the variables studied and make comparisons between the suggested models, a multivariate polynomial regression was performed, as below.

Thus, cross validation was applied to choose the degree of the polynomial and after the complete polynomial regression; the terms' parameters were selected by backward method, with the purpose to obtain greater robustness and significance of the estimated parameters. At end, the regressions showed below in Table 2 were obtained, where βi is the estimated parameter:

$$P_W = \beta_0 + \beta_1 P_P + \beta_2 P_P^2 + \beta_3 F + \beta_4 P_P\theta + \beta_5 \theta F + \beta_6 P_P^2\theta + \beta_7 P_P\theta^2 + \beta_8 P_P^2 F + \beta_9\theta^2 F + \beta_{10} P_P F\theta \quad (4)$$

$$\delta U = \beta_0 + \beta_1 P_P + \beta_2\theta + \beta_3\theta^2 + \beta_4\theta^3 + \beta_5 F^2 + \beta_6 F + \beta_7\theta^2 F + \beta_8 P_P F\theta \quad (5)$$

Table 2: Estimated Parameters for Separative Power – δU and Waste Pressure – P_W Regressions

Parameters	**P_W Model Estimatitive**	**P_W Model Standart Error**	**P_W Model *t* - value**	**P_W Model *p* – value**	**δU Model Estimatitive**	**δU Model Standart Error**	**δU Model *t* - value**	**δU Model *p* – value**
β_0	0.12912	0.03166	4.078	9.32e-05	-0.27759	0.06284	-4.418	2.55e-05
β_1	2.00847	0.26699	7.523	2.73e-11	-0.34728	0.03898	-8.909	2.62e-14
β_2	-1.31786	0.32487	-4.057	0.000101	4.81532	0.33197	14.505	< 2e-16
β_3	-1.41960	0.12453	-11.400	< 2e-16	-7.82183	0.73275	-10.675	< 2e-16
β_4	-2.57090	0.49715	-5.171	1.25e-06	3.14943	0.48929	6.437	4.39e-09
β_5	4.80774	0.43201	11.129	< 2e-16	-1.17191	0.10435	-11.230	< 2e-16
β_6	1.97845	0.48194	4.105	8.43e-05	1.40187	0.10287	13.628	< 2e-16
β_7	0.90806	0.29334	3.096	0.002568	-0.76637	0.19347	-3.961	0.000141
β_8	0.81418	0.20645	3.944	0.000152	0.82022	0.14024	5.849	6.40e-08
β_9	-1.66908	0.31994	-5.217	1.03e-06	-	-	-	-
B_{10}	-2.62012	0.40396	-6.486	3.71e-09	-	-	-	-

Table 3 shows the metrics of the predictive results of all algorithms used. It is possible to verify the metrics RMSE, MSE, MAE and R2 for the models of separative power and waste pressure.

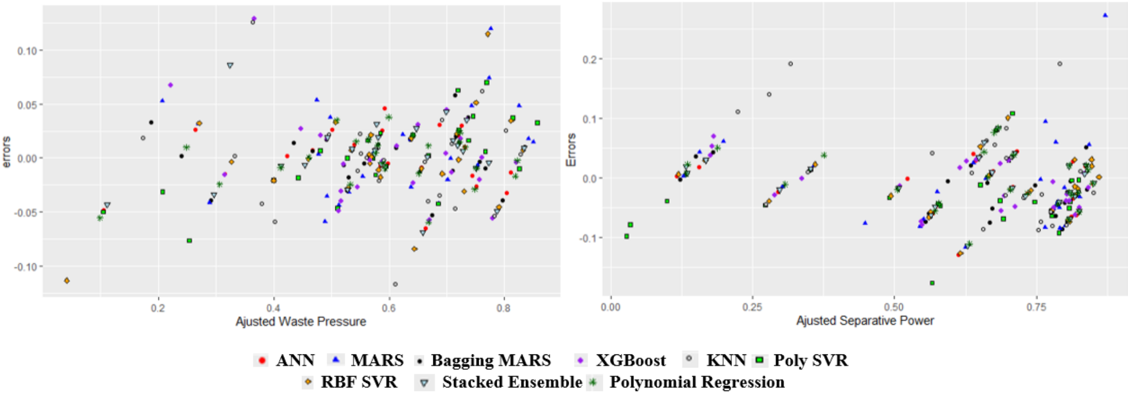
Table 3 : Comparison of Machine Learning Regression Techniques for Test Set Data

Algorithm	**MSE dU**	**MAE dU**	**RMSE dU**	**R2 dU**	**MSE Pw**	**MAE Pw**	**RMSE Pw**	**R2 Pw**
ANN	0,001910	0,0334048	0,043707	0,9656786	0,0006767	0,0212199	0,026013	0,97529
MARS	0,006133	0,0582456	0,0783154	0,889806	0,001786	0,0338141	0,042261	0,93477
Bagging MARS	0,002265	0,0383384	0,0475936	0,959303	0,000585	0,0186536	0,0241868	0,97864
Poly SVR	0,004334	0,0549332	0,0658354	0,9221277	0,001127	0,0260079	0,0335706	0,95884
RBF SVR	0,002340	0,0382443	0,0483755	0,9579549	0,0017451	0,0281615	0,0417748	0,93627
XGBoost	0,001732	0,0362098	0,0416216	0,9688756	0,0015615	0,0288764	0,0395164	0,94297
KNN	0,006396	0,0628507	0,0799757	0,8850842	0,001962	0,0311137	0,0442943	0,92835
Stacked Ensemble	0,002112	0,0386568	0,0459572	0,9620536	0,0007805	0,0228197	0,0279382	0,97149
Polynomial Regression	0,002035	0,0370322	0,0451151	0,9634315	0,0005717	0,0190563	0,0239112	0,97912

Figure 3 shows the errors around the predicted values for each algorithm. It is possible to verify the homoscedasticity of the experimental data since there is a random distribution around the predicted values. Furthermore, the model of separative power generated by XGBoost algorithm reached better metrics than the others. On the other hand, the best model for waste pressure was developed by Bagging MARS algorithm.

However, it is important to highlight that models which use ANN, stacked ensemble, or polynomial regression techniques reached very good results for both variables. According to this reasoning, multivariate polynomial regression becomes more attractive than the other techniques because it is a special case of linear regression and presents an equation as a result, which makes easier to interpret the outputs coefficients and to transfer the model to other programming languages. Furthermore, it is a faster and simpler technique than the others. Although the case in study was a problem of optimization, with more variables, like temperature, scoops distances, feed position, baffle size, this results and performance of multivariate polynomial regression could not be the same and other techniques may be more attractive.

Figure 3: Error Distribution Around Predict Values for Test Set Data



Figures 4 to 12 show the contour plots at constant product pressure of all models.

Figure 4: View of contour curves at constant product pressure (Pp = 0.5) according to polynomial multivariate regression model. a) Waste pressure - PW; and b) Separative Power - dU.

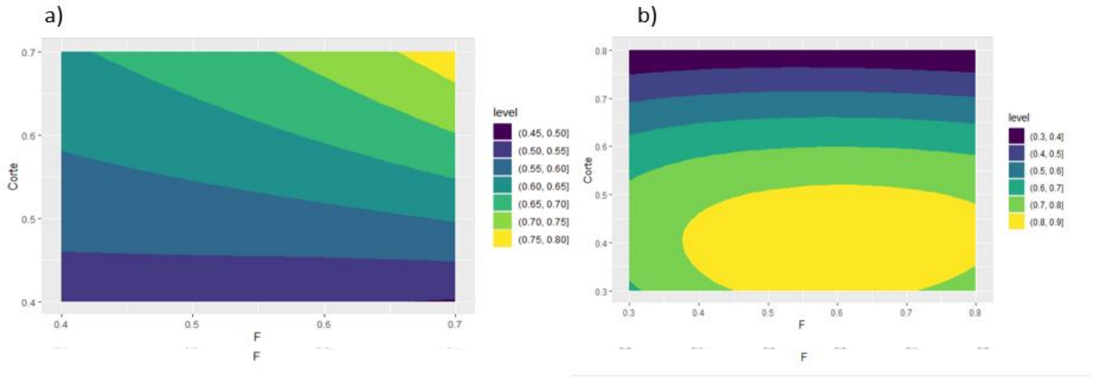


Figure 4: View of contour curves at constant product pressure ($P_P = 0.5$) according to XGBoost model. a) Waste pressure - PW; and b) Separative Power - dU.

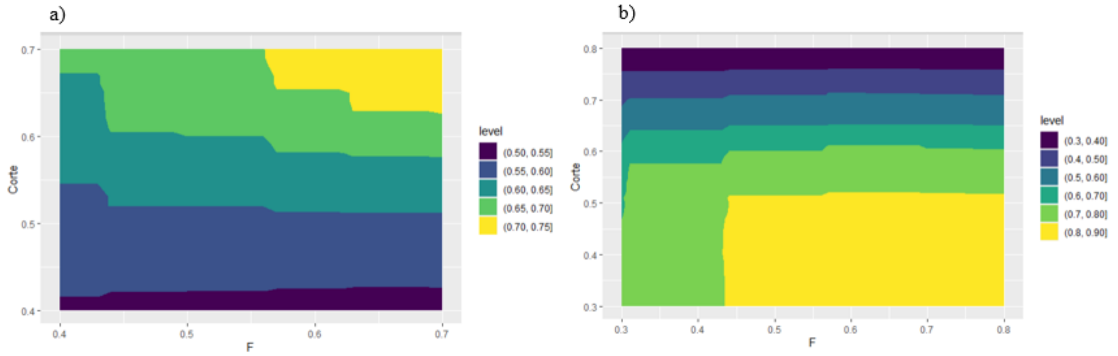


Figure 5 View of contour curves at constant product pressure ($P_P = 0.5$) according to XGBoost model. a) Waste pressure - PW; and b) Separative Power - dU.

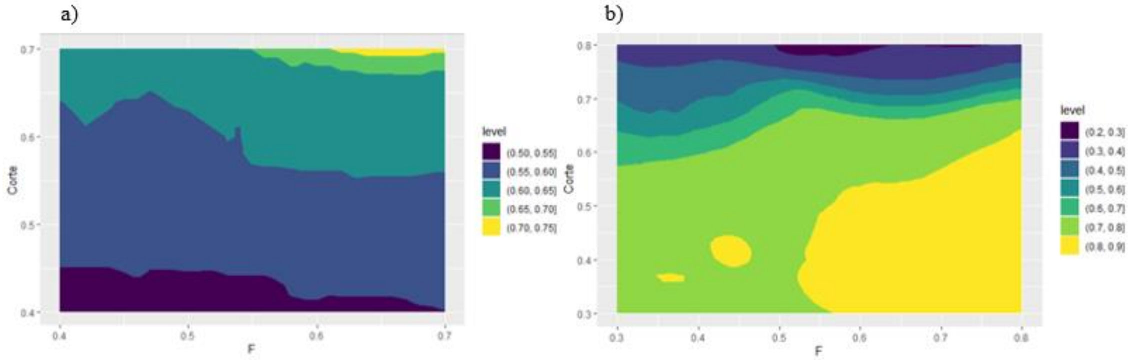


Figure 6: View of contour curves at constant product pressure ($P_P = 0.5$) according to Poly SVR model. a) Waste pressure - PW; and b) Separative Power - dU.

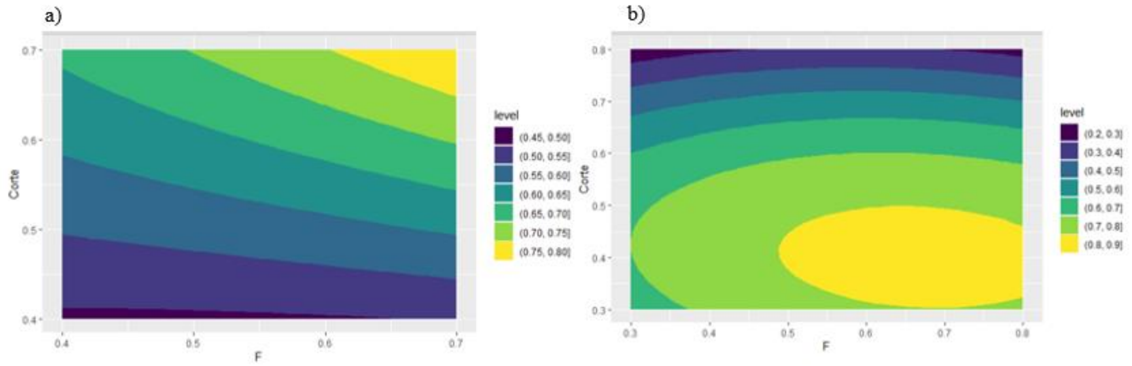


Figure 7: View of contour curves at constant product pressure (P_P = 0.5) according to bagging MARS model. a) Waste pressure - PW; and b) Separative Power - dU.

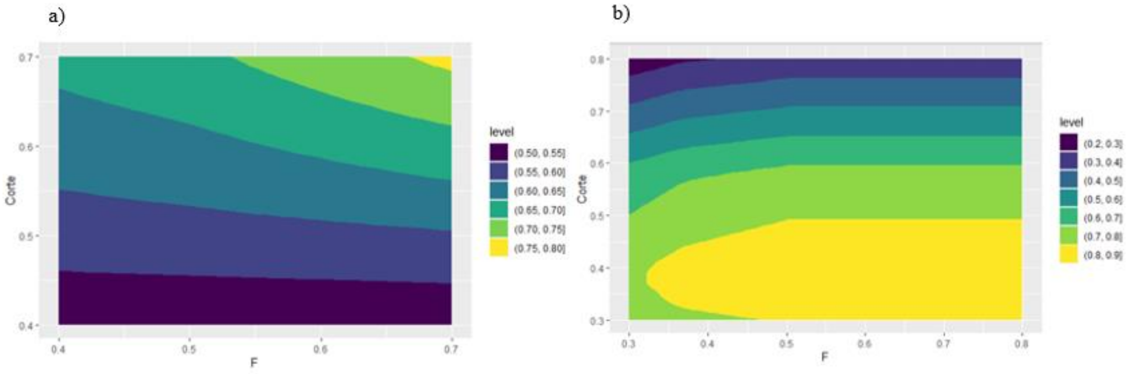


Figure 8: View of contour curves at constant product pressure (P_P = 0.5) according to MARS model. a) Waste pressure - PW; and b) Separative Power - dU.

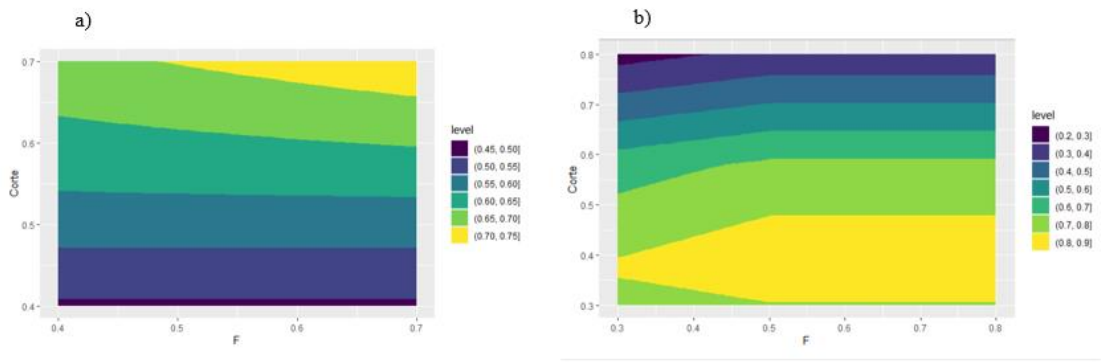
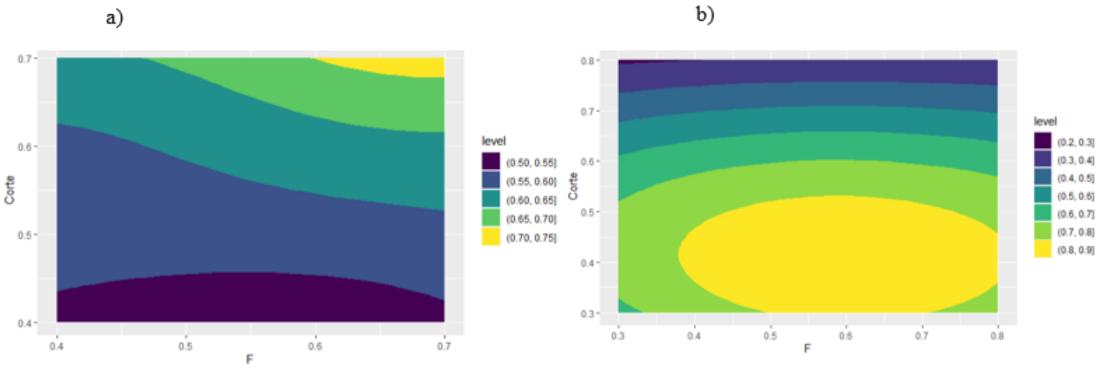


Figure 9: View of contour curves at constant product pressure (P_P = 0.5) according to RBF SVR model. a) Waste pressure - PW; and b) Separative Power - dU.



6 CONCLUSION

This paper presented a comparison between some machine learning regression applied to modeling the separative power and waste pressure gas centrifuge, resulting in nine regressions for each variable.

It is concluded that the best machine learning technique to assess the separative power was XBoost, while the best model for prediction of waste pressure was obtained by the Bagging MARS algorithm, because these methods presented the best metrics. However, when the problem is restricted to evaluate only three explicative variables, such as feed flux, product pressure, and cut, it is recommended to use multivariate polynomial regression, because it is simpler, interpretable, and the results reached metrics similar to the best algorithms.

REFERENCES

WHITAKER, J.M. Uranium Enrichment Plant Characteristics: A Training Manual for IAEA. Tennessee, Oak Ridge National Laboratory, 2005.

JORDAN, I. Separação dos isótopos de urânio pelo processo da centrifugação em fase gasosa, Informação IPEN 3 -IPEN-Inf-3, 1980.

ANDRADE, D.A.; ANDRADE, D. A.; BASTOS, J. L. F. . Thermal hydrodynamical analysis of a countercurrent gas centrifuge. ANNALS OF NUCLEAR ENERGY, England, v. 25, n.11, p. 859-888.

MIGLIORINI, P.J. Modeling and Simulation of Gas Centrifuge Cascades for Enhancing the Efficiency of IAEA Safeguards. 2013. 142 f. Tese (PhD em Engenharia). University of Virginia, Charlottesville, VA, 2013.

VILLANI, S. (ed.), Uranium Enrichment, Énergoatomizdat, Moscow (1983).

BOGOVALOV, S. BORMAN, V. Separative Power of an Optimised Concurrent Gas Centrifuge. Nuclear Engineering and Technology. Volume 48, Issue 3, June 2016, Pages 719-726.

FUCHS, K. PEIERLS, R. Separation of isotopes, Selected Scientific Papers of Sir Rudolf Peierls: (With Commentary) R.H. Dalitz, Sir Rudolf Peierls (Eds.), World scientific (1997), pp. 303-320 (DTA rept. MS 12A, 1941).

MIGLIAVACCA, S.C.P. Modelagem do comportamento separativo de ultracentrígugas via rede neural. Tese de Doutorado. Universidade de São Paulo, São Paulo, 1999.

ANDRADE, M,C,V. Aplicação de Redes Neurais para Detecção de Erros Grosseiros em Dados de Processo de Separação de Isótopos de Urânio por Ultracentrifugação. Dissertação de Mestrado. Universidade de São Paulo, São Paulo, 2004.

MIGLIAVACCA, E. ; ANDRADE, D. A. ; ANDRADE, D.A. . Multivariate analysis with covariance matrix applied to separative power modeling of a gas centrifuge. Annals of Nuclear Energy, v. 35, p. 534-538, 2008.

ANDRADE, D.A.; ANDRADE, D. A. ; MIGLIAVACCA, E. . Ultracentrifuge Separative Power Modeling with Multivariate Regression using Covariance Matrix. PASSAGES DE PARIS (APEB-FR), França, v. 2, p. 91-102, 2005.

CRUZ, M.U.L. Modelagem dos Parâmetros Separativos de Ultracentrífugas para Enriquecimento de Urânio Através de Modelos de Redes Neurais Híbridas. Dissertação de Mestrado. Universidade de São Paulo, São Paulo, 2005.

RStudio Team. RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/. 2020

KUHN et al. Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles. https://www.tidymodels.org. 2020.

WICKHAM et al. Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686. 2019

Stephen MILBORROW. Derived from mda:mars by Trevor Hastie and Rob Tibshirani. Uses Alan Miller's Fortran utilities with Thomas Lumley's leaps wrapper. earth: Multivariate Adaptive Regression Splines. R package version 5.3.0. https://CRAN.R-project.org/package=earth. 2020.

Max KUHN. baguette: Efficient Model Functions for Bagging. R package version 0.1.0. https://CRAN.R-project.org/package=baguette. 2020.

Klaus SCHLIEP and Klaus HECHENBICHLER. kknn: Weighted k-Nearest Neighbors. R package version 1.3.1. https://CRAN.R-project.org/package=kknn. 2016.

VENABLES, W. N & RIPLEY, B. D. Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0. 2002.

Alexandros KARATZOGLOU, Alex SMOLA, Kurt HORNIK, Achim ZEILEIS. kernlab - An S4 Package for Kernel Methods in R. Journal of Statistical Software 11(9), 1-20. URL http://www.jstatsoft.org/v11/i09/. 2014.

Tianqi Chen, Tong He, Michael Benesty, Vadim Khotilovich, Yuan Tang, Hyunsu Cho, Kailong Chen, Rory Mitchell, Ignacio Cano, Tianyi Zhou, Mu Li, Junyuan Xie, Min Lin, Yifeng Geng and Yutian Li. xgboost: Extreme Gradient Boosting. R package version 1.3.2.1. https://CRAN.R-project.org/package=xgboost. 2021.

Simon Couch and Max Kuhn. stacks: Tidy Model Stacking. R package version 0.1.0. https://CRAN.R-project.org/package=stacks. 2020.