Moroccan Journal of Chemistry
ISSN : 2351-812X

Mor. J. Chem., 2023, Volume 11, Issue 1, Page 221-229
https://revues.imist.ma/index.php/morjchem

Ridge and Lasso Regression for Feature Selection of Overlapping Ibuprofen and Paracetamol UV Spectra

Suprapto S.[1] *, Ni'mah Y. L. [1], Baroroh V. [1], Harmami H. [1], Ulfin I. [1], Harmami H. [1], Purbaningtyas T. E.[2]

[1]*Department of Chemisty, Faculty of Science and Data Analytics, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia*

[2]*Analytical Chemistry Study Program, Universitas Islam Indonesia, Jogjakarta, Indonesia*

**Corresponding author, Email address: suprapto@chem.its.ac.id*

***Received** xx Jan 2023, **Revised** xx xxx 2023, **Accepted** xx xxx 2023*

***Citation**: Suprapto S., Ni'mah Y. L., Baroroh V., Harmami H., Ulfin I., Harmami H., Purbaningtyas T. E. (2023) Ridge and Lasso Regression for Feature Selection of Overlapping Ibuprofen and Paracetamol UV Spectra, Mor. J. Chem., 11(1), 221-229.*

DOI: https://doi.org/10.48317/IMIST.PRSM/morjchem-v11i1.31466

Abstract: Feature selection is a process of identifying and selecting a subset of the input variables that are most relevant to the target variables. Regularized regressions such as Ridge and Lasso apply penalties to reduce the weight of features in the linear model input. Ridge regression causes the regression weight to be very close to zero, but not zero. On the other hand, the Lasso makes the weight equal to zero. Lasso regression is recommended for models with a lot of features, but only a few are important. In this study, a quantitative analysis of a mixture of ibuprofen and paracetamol in pain relief tablets was performed. A training set consisting of 25 UV spectra was optimized by Linear, Ridge, and Lasso regression. Two sets of test solutions with known concentrations of ibuprofen and paracetamol were used for validation. The Lasso regression only uses some absorbance values as estimators but has good accuracy at predicting ibuprofen and paracetamol in the test and pain relief tablet samples that were comparable to linear and ridge regression. The ridge regression showed better recovery and RMSE compared to the linear and lasso regression.

Keywords: Regularized regression; Feature selection; Ridge regression; Lasso regression; Paracetamol; Ibuprofen

1. Introduction

The methods for the determination of ibuprofen and paracetamol in mixtures are very limited (Issa et al., 2011). Analytical techniques such as titrimetric, chromatographic, spectroscopic, electrophoretic, and electrochemistry have been studied in previous research (Erk et al., 2001). The spectrophotometric method is a simple and fast method for the determination of ibuprofen and paracetamol in a binary mixture in a tablet without prior separation. This method has been shown to have good accuracy and precision (El-Didamony et al., 2022; Sugiarso et al., 2020). Several spectrophotometric techniques use absorbance of UV-Vis spectra in a relatively wide range of wavelengths (Darwish et al., 2016). Using a relatively large number of estimators poses a risk of overfitting. It is unlikely that all of the absorbances in the wavelength range used as an estimator have a significant contribution to the regression model (Ali et al., 2017). The selection of estimator

features that are relevant to the response variable can be done using the stepwise regression, forward selection, or backward elimination method. The backward elimination process is carried out by including all estimators at the beginning of the analysis, reducing or even canceling the features that have no significant effect. This method requires many steps to optimize. The Ridge and Lasso regression models are known as regularization methods that provide information for backward elimination type feature selection (Rao et al., 2019). These models are simpler to implement for feature selection. Optimization can be done by tuning the hyperparameter in the models. Therefore, in this study, the absorbances at a wavelength from 200 to 270 nm were analyzed using the Ridge and Lasso regression to determine the reliability of the contribution of each estimator to the composition of paracetamol and ibuprofen solution.

2. Methodology

2.1 Chemicals and instruments

The materials used in this study were: ibuprofen (99.8 - 101.0 %, Sigma Aldrich), paracetamol (87.5 - 92.5%, Sigma Aldrich), methanol p.a. (Merck), and commercial pain relief tablets containing 350 mg paracetamol and 200 mg ibuprofen. Absorption spectra were recorded using a Genesys 10S spectrophotometer under the following conditions: quartz cell 10 mm; wavelength range 200–270 nm. Ridge and Lasso's regression was carried out using the sklearn linear model (Buitinck et al., 2013).

2.2 Experiments

Paracetamol 1000 mg/L stock solutions were prepared by dissolving paracetamol 0.1143 gram in 100 ml methanol. Ibuprofen 1000 mg/L stock solution was prepared by dissolving ibuprofen 0.1002 gram in 100 ml methanol. Standard solutions with 25 ibuprofen (6, 8, 10, 12, and 14 ppm) and paracetamol (6, 8, 10, 12, and 14 ppm) concentration variations were prepared. The 25 calibration solutions were measured with 5 replications so that 75 spectra were obtained. Twenty commercial pain relief tablets were weighed and ground to powder with a mortar. An amount of tablet powder corresponding to 200 mg ibuprofen (IB) and 350 mg paracetamol (PA) was weighed using an analytical balance. The medicinal powder was placed in a beaker and dissolved with a little methanol. The dissolved medicinal powder was transferred to a 100 ml volumetric flask using a funnel. The residue on the beaker was rinsed three times using 10 ml of methanol, then methanol was added up to the limit mark on the volumetric flask (Khoshayand et al., 2008). The solution was put in an ultrasound water bath for 5 minutes to speed up the dissolution process. After the sonication process, the mixture was introduced into the falcon tube and centrifugated for 10 minutes at 340 rpm. The supernatant of 0.05 ml was dissolved in 25 ml methanol in a volumetric flask. Furthermore, the absorbances of the solutions were measured using a UV-Vis spectrophotometer at a wavelength of 200-270 nm. This treatment was repeated three times to obtain triplicate data.

2.3 Chemometric elaboration

Ridge regression is a linear least-squares regression with l2 regularization. l2 regularization eliminates some common least-squares problems by penalizing the coefficients. The regularization minimizes the number of ridge coefficients. The l2 penalty applied is as shown in **Eqn 1**:

$$\min_w ||Xw - y||_2^2 + \alpha||w||_2^2 \qquad \textbf{Eqn. 1}$$

The complexity parameter α > 0 controls the shrinkage level of the regression coefficient: the larger the value, the higher the depreciation amount and thus the collinearity coefficient. Lasso regression is a linear least squares model with additional l1 regularization (**Eqn. 2**). The l1 penalizes insignificant coefficients to zero:

$$\min_w (1/2n\text{samples})\|Xw - y\|22 + \alpha\|w\|1 \qquad \textbf{Eqn. 2}$$

Lasso estimates the regression model by applying the least squares penalty (α||w||1) where α is constant while ||w||1 is the l1 regularization coefficient vector (Buitinck et al., 2013).

3. Results and Discussion

3.1 Spectrophotometric analysis and construction of the models

The structural similarity of ibuprofen and paracetamol based on Tanimoto similarity calculation in RDKit (Landrum, 2016) was 0.2051. The chemical structures of ibuprofen and paracetamol were shown in **Figure 1**. The substructure similarity of ibuprofen and paracetamol were shown in **Figure 2**. The absorption spectra of the training set were recorded in the range of 220–270 nm. **Figure 3 (a, b, and c)** shows the UV spectra of the training set, validation/test set, and pharmaceutical formulation sample.

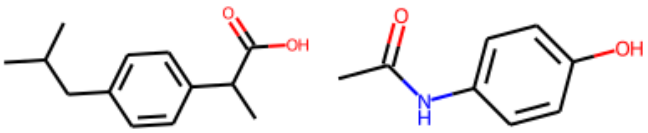

Figure 1. The chemical structure of ibuprofen and paracetamol

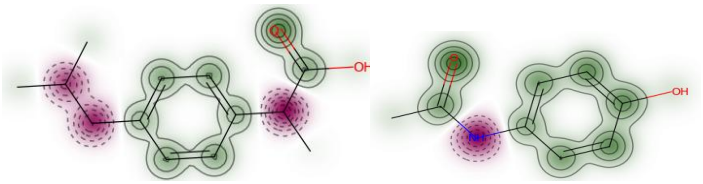

Figure 2. The chemical structure similarity of ibuprofen and paracetamol (green color indicated similar substructures, purple color indicated dissimilarity)

Figure 3 shows peaks at about 225 nm and 250 nm. The peaks at 225 nm were the absorption of ibuprofen and the peaks at 250 nm were the peaks of paracetamol (Hoang et al., 2014). The concentration of the validation solution was set to avoid 'leaking' in the calibration or training solution. So, the absorbance was not similar to the training solution. In general, the spectra of train, test, and sample solutions have similar patterns.

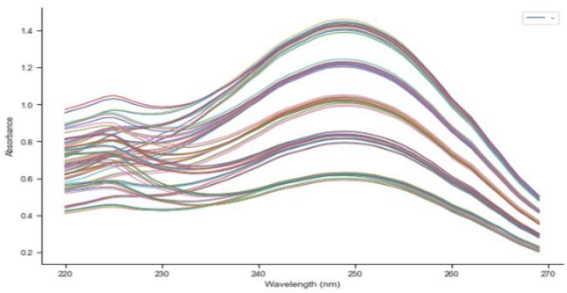

Figure 3a. Temperature profile for biomass conversion

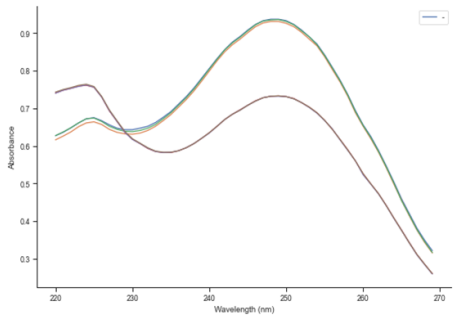

Figure 3b. Temperature profile for hybrid co-conversion

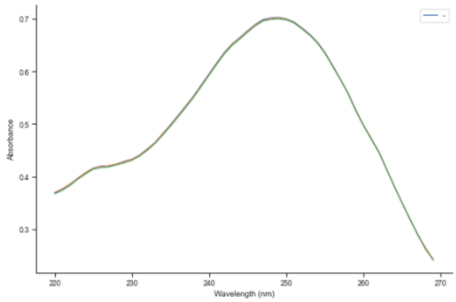

Figure 3c. Temperature profile for hybrid co-conversion

Feature selection through regularized regression was proposed to overcome the overfitting problem in linear regression models. The regularization factor that affected the number of predictors involved in the model was the alpha value. The RMSE and recovery of train, test, and sample were optimized at a certain range of alpha to accommodate the best recovery with proportional RMSE for train, test, and sample data. The ibuprofen RMSE and recovery of train, test, and sample as a function of alpha using Ridge regression were shown in **Figure 4** and **Figure 5**, respectively.

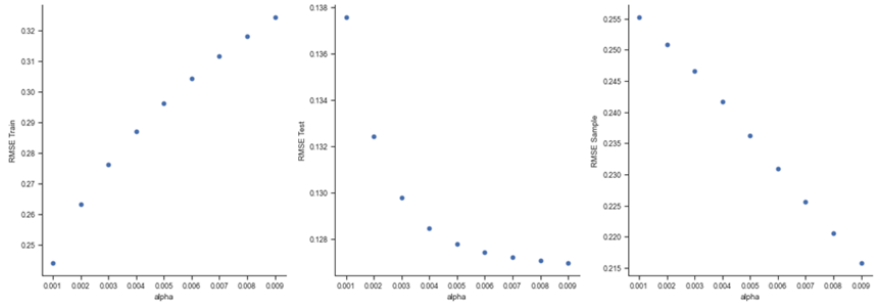

Figure 4. RMSE of ibuprofen train, test, and sample sets as a function of alpha

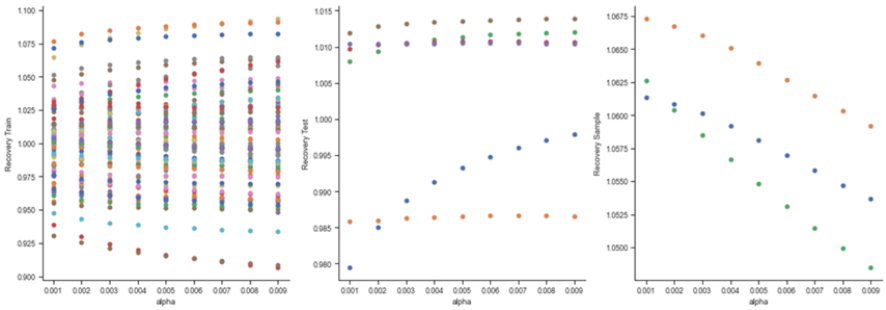

Figure 5. Recovery of ibuprofen train, test, and sample sets as a function of alpha

The RMSE and recovery of train, test, and sample show different trends when the alpha was increased, as shown in **Figures 4** and **Figure 5**. The recoveries of ibuprofen either in train, test, or sample were lies in the range of 0.90 to 1.10.

Alpha optimization for RMSE and recovery of paracetamol using Ridge regression were shown in **Figure 6** and **Figure 7**, respectively.

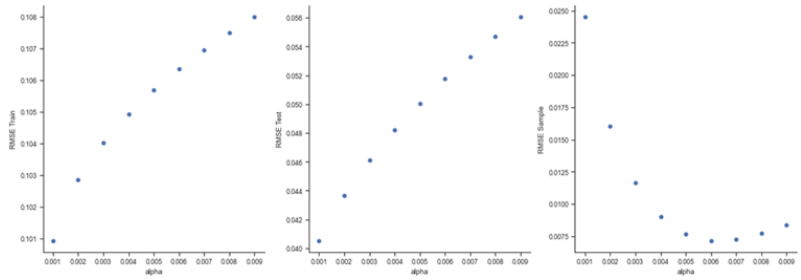

Figure 6. The influence of alpha on paracetamol train, test, and sample RMSE

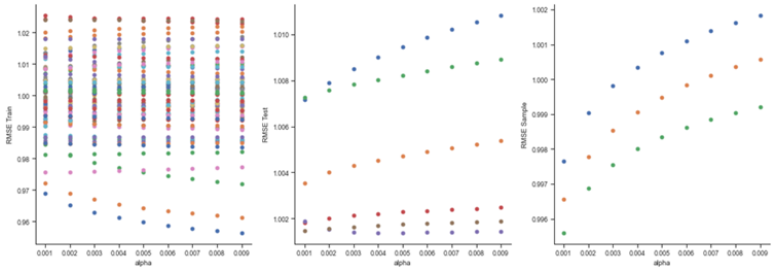

Figure 7. The influence of alpha on paracetamol train, test, and sample recovery

Paracetamol recoveries were lies at about 1 either for train, test, or sample in the alpha ranges that optimized using Ridge regression. To accommodate both RMSE and recovery of ibuprofen and paracetamol, an alpha of 0.005 was applied for ibuprofen and paracetamol prediction. The optimization of the Lasso regression alpha value was carried out in a similar way to the Ridge regression. **Figure 8** and **Figure 9** show the RMSE and recovery of ibuprofen using Lasso regression.

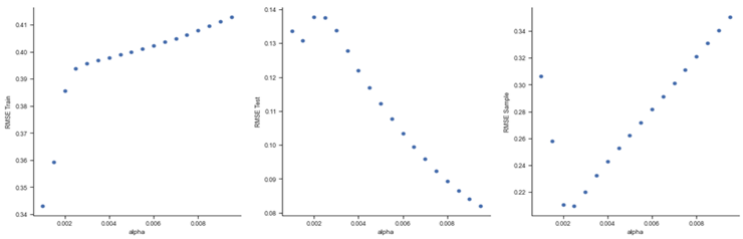

Figure 8. The RMSE of ibuprofen as a function of Lasso regression alpha

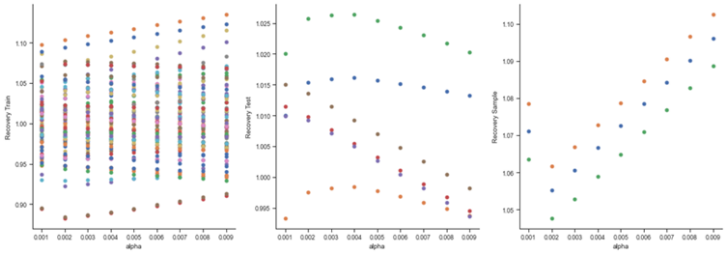

Figure 9. The recovery of ibuprofen as a function of Lasso regression alpha

The RMSE of Lasso regression tends to increase as alpha increases for train and sample data but decreases in test data. The recovery of ibuprofen using Lasso regression lies at about 1, so, to simplify the models, an alpha of 0.005 was also applied in Lasso regression for ibuprofen prediction.

Figures 10 and Figure 11 show the RMSE and recovery of paracetamol using Lasso regression. In general, the paracetamol RMSE increases as the Lasso regression alpha increases.

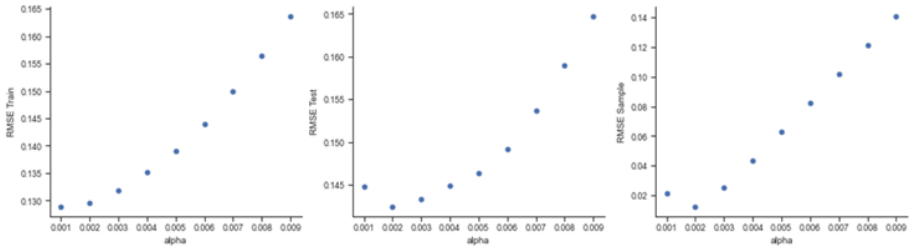

Figure 10. The paracetamol RMSE as Lasso regression alpha function

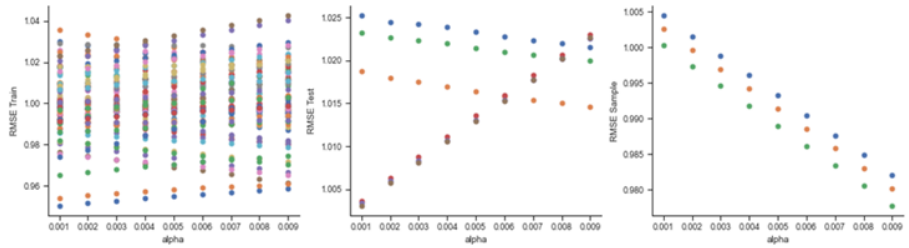

Figure 11. The paracetamol recovery as Lasso regression alpha function

The paracetamol recovery for pain relief drug samples decreases as alpha increases, but, all paracetamol training, test, and sample recoveries lie at about 1 for an alpha value of 0.005. Linear regression models involve absorbance at all wavelengths. Ridge models shrink the coefficient or slopes of absorbances that do not have a significant influence on regression models. Lasso regression models shrink the regression coefficients that have no significant effect to zero, thus Lasso models only involve absorbances at certain wavelengths, as shown in **Figure 12** for ibuprofen and **Figure 13** for paracetamol.

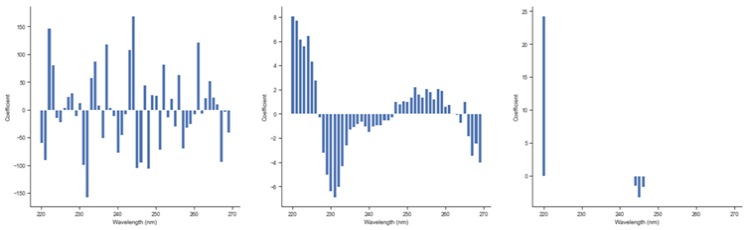

Figure 12. The regression coefficient for ibuprofen prediction

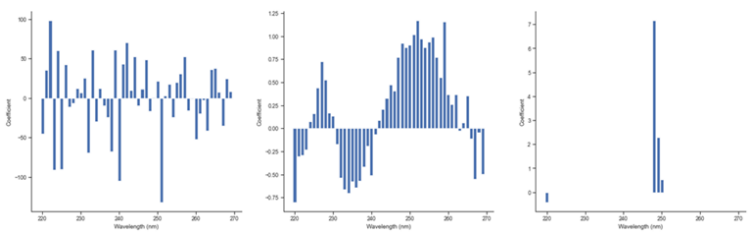

Figure 13. The regression coefficient for paracetamol prediction

Ridge regression coefficients for ibuprofen have positive values at wavelengths 220 to 227 nm and small positive values at 248 to 263 nm. Coefficients for paracetamol prediction, on the other hand, have high positive values from 243 to 266 nm and small positive values at 225 to 231 nm. Negative coefficients were observed for absorbances at other wavelengths. Most extreme shrinkage was observed for Lasso regression models. Ibuprofen Lasso regression coefficient only involves absorbance at 221 nm, with a value of 24.2710, and 245-247 nm with values of - 1.5096, -3.2499, and -1.7087. Paracetamol Lasso regression coefficients, on the other hand, have positive values at 249-250 nm with coefficient values of 7.1539 and 2.2885. The regression curves for ibuprofen and paracetamol training datasets were shown in **Figure 14** and **Figure 15**, respectively.

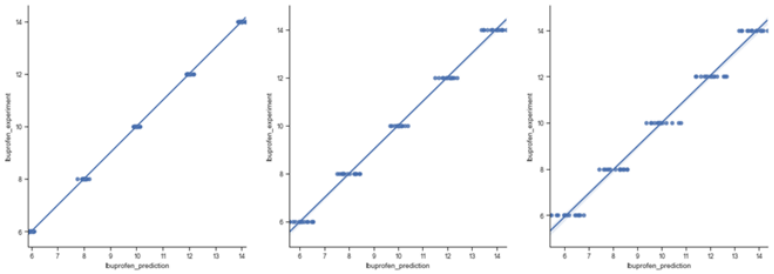

Figure 14. Ibuprofen regression curve of training/calibration data

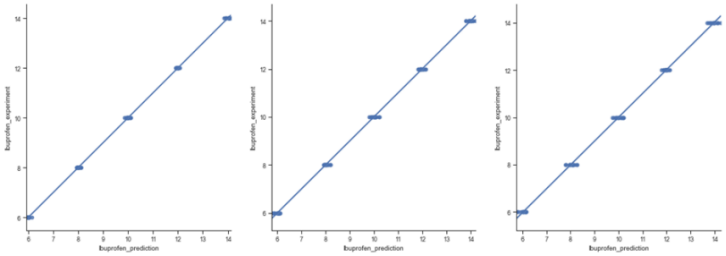

Figure 15. Paracetamol regression curve of training/calibration data

3.2. Spectrophotometric analysis of Validating and Sample Solutions

Ibuprofen test solution prediction using Linear, Ridge, and Lasso, **Figure 16**, shows that all of the models perform well in predicting test solutions. The ibuprofen prediction in sample or pain relief tablets shows that Ridge and Lasso perform better than Linear regression, as shown in **Figure 17**.

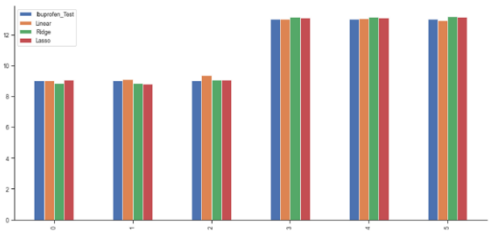

Figure 16. Ibuprofen test solution prediction using Linear, Ridge, and Lasso regression

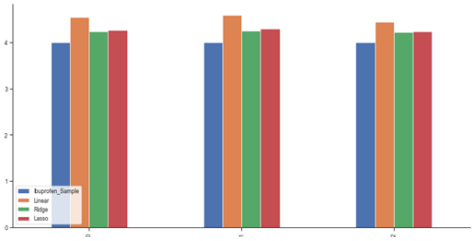

Figure 17. Ibuprofen in pain relief tablets prediction using Linear, Ridge, and Lasso regression

Paracetamol test solution prediction using Linear, Ridge, and Lasso regression was shown in **Figure 18**. Paracetamol prediction using Linear, Ridge, and Lasso regression for the test solution did not show any significant differences. But, similar to ibuprofen, Ridge, and Lasso regression paracetamol prediction in sample or pain relief tablets has better performance than the Linear model, as shown in **Figure 19**. This remark was confirmed by the recovery values of each model as shown in **Figure 20**.

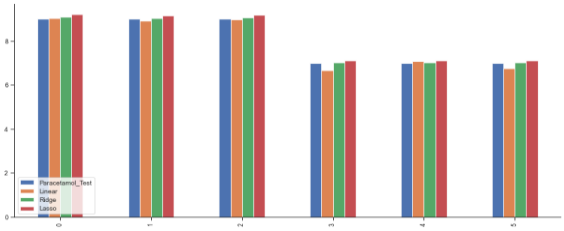

Figure 18. Paracetamol test solution prediction using Linear, Ridge, and Lasso regression

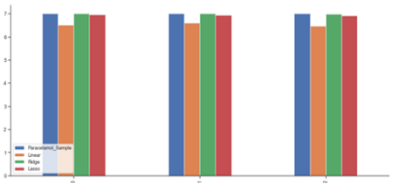

Figure 19. Paracetamol sample solution prediction using Linear, Ridge, and Lasso regression

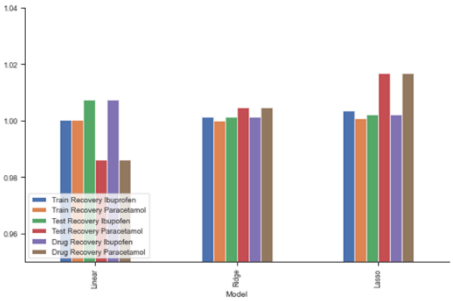

Figure 20. Recovery of ibuprofen and paracetamol in test and sample solution using Linear, Ridge, and Lasso regression

The RMSE of ibuprofen and paracetamol in test and sample solution using Linear, Ridge, and Lasso regression (**Figure 21**) confirm that the Linear regression model show some indication of overfitting as it is good in predicting training/calibrating data but has higher RMSE for test and sample solution. Ridge and Lasso RMSE for ibuprofen in calibrating data were higher than those of paracetamol as well as test and sample data.

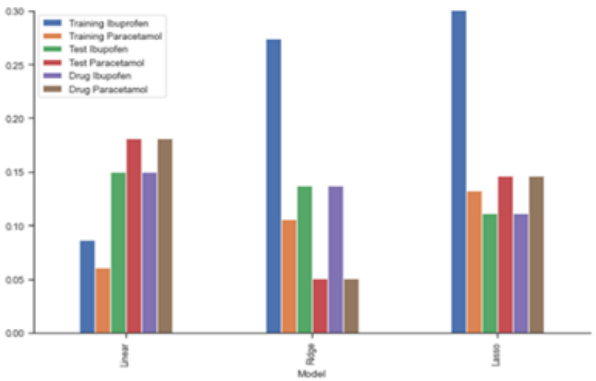

Figure 21. RMSE of ibuprofen and paracetamol in test and sample solution using Linear, Ridge, and Lasso regression

Conclusion

The recovery and RMSE of Linear, Ridge, and Lasso regression show that, although Lasso regression uses only a few absorbance values at a certain wavelength as estimators, it can perform very well in predicting ibuprofen and paracetamol in test and pain relief tablet samples that are comparable to Linear and Ridge regression. Ridge regression shows better recovery as well as RMSE compared to Linear and Lasso regression. In conclusion, Ridge and Lasso's regression shows good performance by putting a penalty on regression coefficients, by shrinkage their values in Ridge regression, or by putting the coefficient to zero for Lasso regression. Ridge and Lasso's regression using a few estimators, makes the regression model simple and avoids overfitting.

References

Ali, S.N., Sultana, N., Qayoom, A., Arayne, M.S., (2017). Synthesis and spectrophotometric study of omeprazole charge transfer complexes with bromothymol blue, methyl orange and picric acid. *Mor. J. Chem.* 5, 5–226. https://doi.org/10.48317/IMIST.PRSM/morjchem-v5i1.6551

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., Layton, R., Vanderplas, J., Joly, A., Holt, B., Varoquaux, G., (2013) API design for machine learning software: experiences from the scikit-learn project. *arXiv*:1309.0238 [cs].

Darwish, H.W., Elzanfaly, E.S., Saad, A.S., Abdelaleem, A.E.-B., (2016) Full spectrum and selected spectrum based multivariate calibration methods for simultaneous determination of betamethasone dipropionate, clotrimazole and benzyl alcohol: Development, validation and application on commercial dosage form. *Spectrochim. Acta A Mol. Biomol. Spectrosc.* 169, 50–57. https://doi.org/10.1016/j.saa.2016.06.001

El-Didamony, A.M., Ramadan, G.M., Moustafa, M.A., (2022) Spectrofluorimetric Determination of Citalopram and Desvenlafaxine Antidepressant Drugs in Micellar Medium. *Mor. J. Chem.* 10, 10–247. https://doi.org/10.48317/IMIST.PRSM/morjchem-v9i3.22269

Erk, N., Özkan, Y., Banoğlu, E., Özkan, S.A., Şentürk, Z., (2001) Simultaneous determination of paracetamol and methocarbamol in tablets by ratio spectra derivative spectrophotometry and LC. *J Pharm Biomed Anal* 24, 469–475. https://doi.org/10.1016/S0731-7085(00)00447-7

Hoang, V.D., Ly, D.T.H., Tho, N.H., Minh Thi Nguyen, H., (2014) UV Spectrophotometric Simultaneous Determination of Paracetamol and Ibuprofen in Combined Tablets by Derivative and Wavelet Transforms. *The Scientific World Journal* 2014, e313609. https://doi.org/10.1155/2014/313609

Issa, Y.M., Zayed, S.I.M., Habib, I.H.I., (2011) Simultaneous determination of ibuprofen and paracetamol using derivatives of the ratio spectra method. *Arab. J. Chem.* 4, 259–263. https://doi.org/10.1016/j.arabjc.2010.06.044

Khoshayand, M.R., Abdollahi, H., Shariatpanahi, M., Saadatfard, A., Mohammadi, A., (2008) Simultaneous spectrophotometric determination of paracetamol, ibuprofen and caffeine in pharmaceuticals by chemometric methods. *Spectrochim. Acta A Mol. Biomol. Spectrosc.* 70, 491–499. https://doi.org/10.1016/j.saa.2007.07.033

Landrum, G., (2016) RDKit: Open-Source Cheminformatics Software [WWW Document]. URL http://www.rdkit.org/ (accessed 3.15.22).

Rao, H., Shi, X., Rodrigue, A.K., Feng, J., Xia, Y., Elhoseny, M., Yuan, X., Gu, L., (2019) Feature selection based on artificial bee colony and gradient boosting decision tree. *Appl. Soft Comput.* 74, 634–642. https://doi.org/10.1016/j.asoc.2018.10.036

Sugiarso, R.D., Zahiroh, N., AnnisaRahmayanti, Putro, H.S., Ediati, R., (2020) Development of the Analytical Method of Iron through Optimization Reducing Agent Ability of Iron (III) to Iron(II). *Mor. J.Chem.* 8, 8–077. https://doi.org/10.48317/IMIST.PRSM/morjchem-v8i1.19127