**MACHINE LEARNING FOR DATA SCIENCE**

**INSTRUCTONS**

The task is to predict the mortality rate (y) from the remaining attributes. Use linear regression and then SVM to show how well you can do this prediction (using RMSE). Then tune the SVM and find the best model you can generate. Calculate the error to show if you were able to improve on regular SVM.

**SOLUTION:**

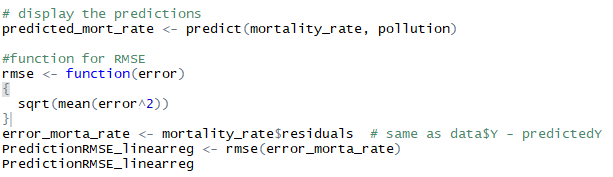
**Steps followed:**

1. Load the dataset into R studio

2. For the given pollution dataset, we can create the multilinear regression model to find the best fit line to classify the data and perform prediction with the same dataset as testing as we have less number of observations.

3. To find the accuracy of the model, we can calculate the Root Mean Squared Error, which is one of the most common accuracy measures for the evaluating the difference between the values of predicted models and the actual observations.

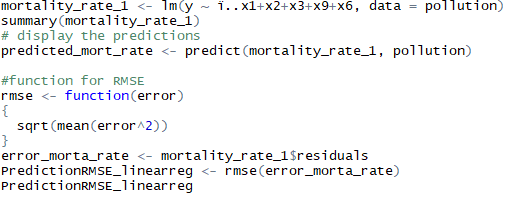




4. The RMSE of the above with all the variables is turned out to be 29.91123

5. From the summary of the model, we can notice that all the variables are not significant, so removing the

insignificant variables we can create one more regression model and calculate its RMSE to measure the accuracy



6. Now the RMSE value is captured as 34.89845

7. The lower RMSE is always the better, considering that there is no data cleaning, data preprocessing done. This might not be a right RMSE value.

8. Now let us proceed with the SVM (Support Vector Machine). The goal of SVMs is to find the optimal

hyperplane because it not only classifies the existing dataset but also helps predict the class of the unseen data. And the optimal hyperplane is the one which has the biggest margin.

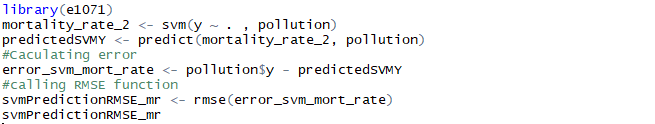
9. The easiest way to separate two classes of data is a line in case of 2D data and a plane in case of 3D data (like N dimensional space in our case). But it is not always possible to use lines or planes and one requires a nonlinear region to separate these classes.

10. Support Vector Machines handle such situations by using a kernel function

which maps the data to a different space where a linear hyperplane can be used to separate classes. This is

known as the **kernel trick** where the kernel function transforms the data into the higher dimensional feature

space so that a linear separation is possible.



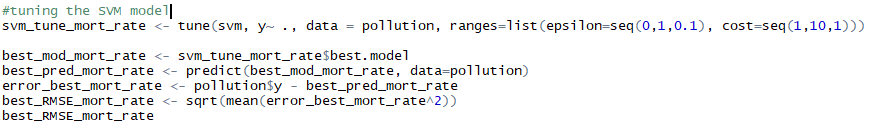


This SVM model’s RMSE value looks better than our regression model.

11. Now let us relax the requirements for the margins by tuning the parameter using tune() function in the SVM

model which will allow us to add little slack values/errors within the limited budget for every data. This is called

as Soft Margin SVM.





The RMSE value of the parameter tuned best SVM model is 21.47147 which seemed to be lowest error in all the

models. This clearly shows the improvement in the regular SVM model. Always the lower RMSE, the better the

accuracy is. Although soft margin SVM classifier gives little less accuracy than the hard margin MMC

(Maximal Margin Classifier) which is very sensitive and might over fit the data, this soft margin SVM classifier will be more robust with the new observations. So, we can conclude that the model created with SVM gives us the

better model to predict the mortality rate based on the given attributes.