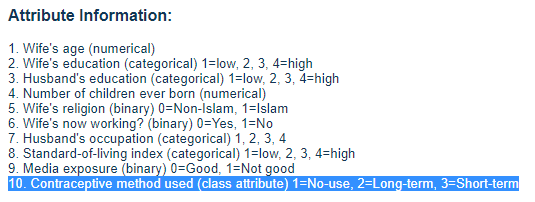
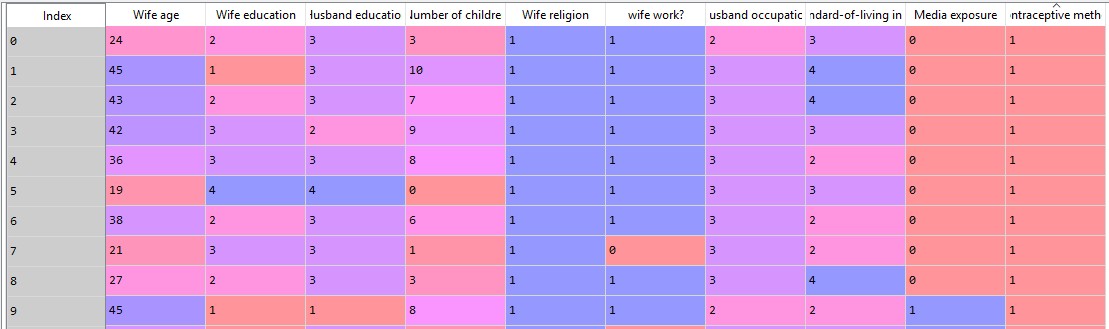
**Classification**

This dataset (1474 rows and 10 columns) is a subset of the 1987 National Indonesia Contraceptive Prevalence Survey. The samples are married women who were either not pregnant or do not know if they were at the time of interview. The problem is to predict the current contraceptive method choice (no use, long-term methods, or short-term methods) of a woman based on her demographic and socio-economic characteristics. The full information about the dataset is the following:

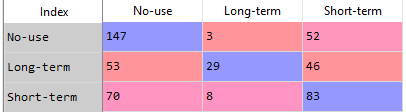
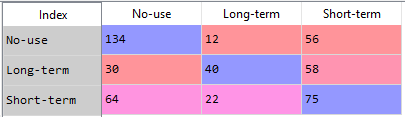


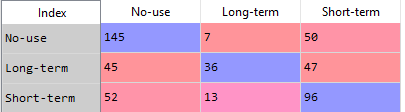
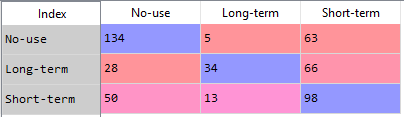
**(****Contraceptive method is Y-variable)**

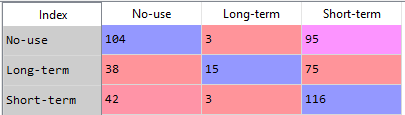


I converted all the categorical variables to dummy variables and deleted the first dummy for each variable. Also, some non-significant columns are deleted using back elimination method. The full report for the 5 algorithms (logistic regression, KNN, ***SVM,*** random forest, logistic regression (one vs one)) I used for this classification model are the followings: (clearly, we can see that SVM is the best algorithm for the prediction of this model. However, it still gives an accuracy of 0.56, which is very low. The simple explanation for this is that these X (explanatory) variables are not good predictors to predict the preference of Contraceptive method used by married woman)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | logistic regression | KNN | ***SVM*** | random forest | logistic regression (one vs one) |
| accuracy | 0.5275 | 0.5071 | ***0.5642*** | 0.5418 | 0.4786 |
| CV (full df) | 0.5078 | 0.5275 | ***0.5499*** | 0.5261 | 0.4752 |
| macro\_precision | 0.5760 | 0.5084 | ***0.5798*** | 0.5725 | 0.5617 |
| CV (full df) | 0.5003 | 0.5118 | ***0.5323*** | 0.5210 | 0.4945 |
| macro\_recall | 0.4899 | 0.4806 | ***0.5318*** | 0.5126 | 0.4508 |
| CV (full df) | 0.4755 | 0.5029 | ***0.5142*** | 0.5016 | 0.4780 |

logistic regression  KNN

SVM random forest

logistic regression (one vs one)