****

**INT 247 CA2**

**E Coli classification**

**Name - Mrityunjay Kumar Mishra**

**Reg. no. - 11804112**

**Roll no. 47**

**Mo no.- 9340913502**

**Email - mmrityunjay727@gmail.com**



|  |  |
| --- | --- |
| **Abstract**: This data contains protein localization sites |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Data Set Characteristics:** | Multivariate | **Number of Instances:** | 336 | **Area:** | Life |
| **Attribute Characteristics:** | Real | **Number of Attributes:** | 8 | **Date Donated** | 1996-09-01 |
| **Associated Tasks:** | Classification | **Missing Values?** | No | **Number of Web Hits:** | 253334 |

**Data Set Information:**

The references below describe a predecessor to this dataset and its development. They also give results (not cross-validated) for classification by a rule-based expert system with that version of the dataset.

Reference: "Expert System for Predicting Protein Localization Sites in Gram-Negative Bacteria", Kenta Nakai & Minoru Kanehisa, PROTEINS: Structure, Function, and Genetics 11:95-110, 1991.

Reference: "A Knowledge Base for Predicting Protein Localization Sites in Eukaryotic Cells", Kenta Nakai & Minoru Kanehisa, Genomics 14:897-911, 1992.

**Attribute Information:**

1. Sequence Name: Accession number for the SWISS-PROT database

2. mcg: McGeoch's method for signal sequence recognition.

3. gvh: von Heijne's method for signal sequence recognition.

4. lip: von Heijne's Signal Peptidase II consensus sequence score. Binary attribute.

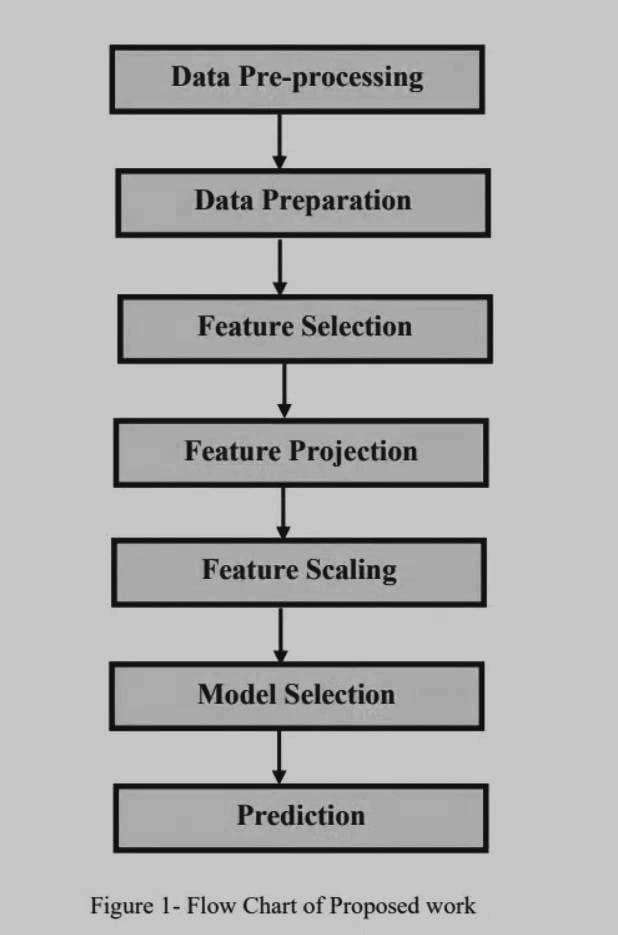
5. chg: Presence of charge on N-terminus of predicted lipoproteins. Binary attribute.

6. aac: score of discriminant analysis of the amino acid content of outer membrane and periplasmic proteins.

7. alm1: score of the ALOM membrane spanning region prediction program.

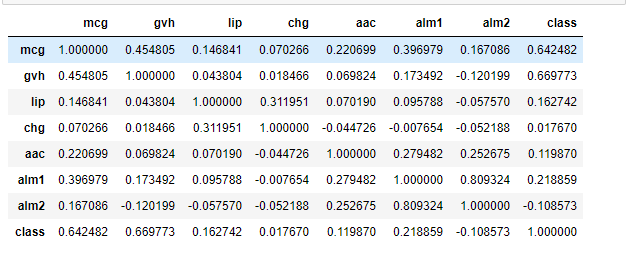
8. alm2: score of ALOM program after excluding putative cleavable signal regions from the sequence.

**METHEDOLOGY**

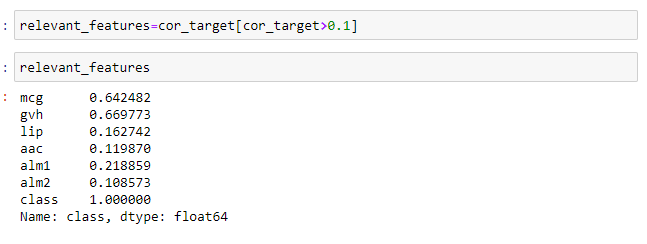
****

**● Gathering Data :-** I have used the Ecoli dataset from openml because it has a classification column so that we can train and test our model. The testing on this dataset will also work for the original raw dataset from UCI Machine Learning ECOLI dataset.

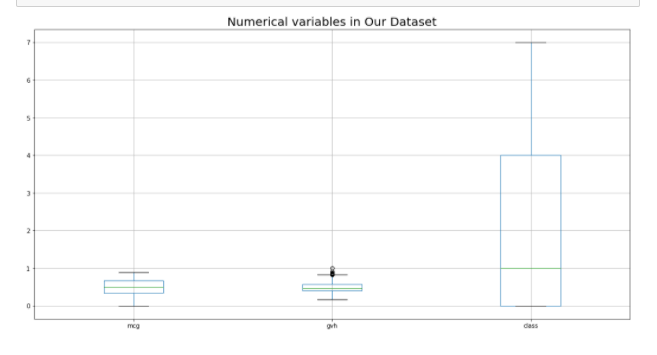
**● Data Preprocessing :-** First, Created a correlation matrix to see the correlation between all the features and target feature ie Class



**Extracted the highly correlated feature :-**

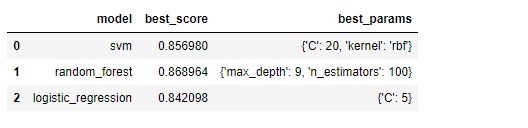
****

Using those features and label as the target Class. Then normalize the data and seeing correlation between the features. Then applying our classifiers to see the result. Detailed data preprocessing is mentioned in the code.

Box plot after preprocessing :- 

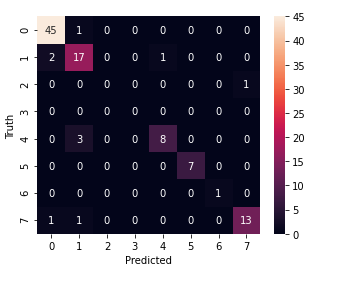
**RESULT AND DISCUSSION**

It is a classification problem so I have predicted the test samples using random forest ,svm, and Logistic regression.



As it is clearly visible that Random Forest is best classifier for our prediction

**CONFUSION MATRIX FOR THE SAME:**

****

**CONCLUSION**

Random Forest Classifier is the best fitted classifier for our prediction.It has cross\_val\_score greater than svm and logistic regression.

On training and testing on our dataset, Our model gave an accuracy of greater than 85%.

#### **Key learning :-**

Learned about confusion matrix,classifiers,feature extraction techniques,how to process raw data,feature scaling,how to make models with better accuracy using data preprocessing and pandas profiling.