**Classification Assignment** 

**Problem Statement or Requirement:**

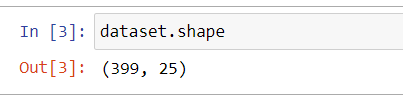
A requirement from the Hospital, Management asked us to create a predictive model which will predict the Chronic Kidney Disease (CKD) based on the several parameters. The Client has provided the dataset of the same.

1. Identify your problem statement

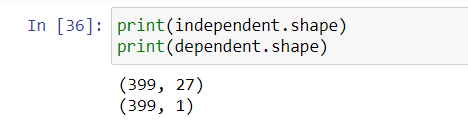
* The dataset is “**Chronic Kidney Disease**”. By using this data need to predict the Kidney disease based on the record.
* Step 1 – Dataset contains **Numeric data** with ordinal values. So, the domain is **Machine Learning**.
* Step 2 – **Learning** – Here input and output are clearly given so it comes under “**Supervised Learning**”.
* Step 3 – It is Supervised Learning by using the possibility of data in output column (i.e.) classification of disease yes/no so, it is a **classification problem**.

1. Tell basic info about the dataset (Total number of rows, columns)

* Total number of rows, columns: **399 rows, 28 columns.**

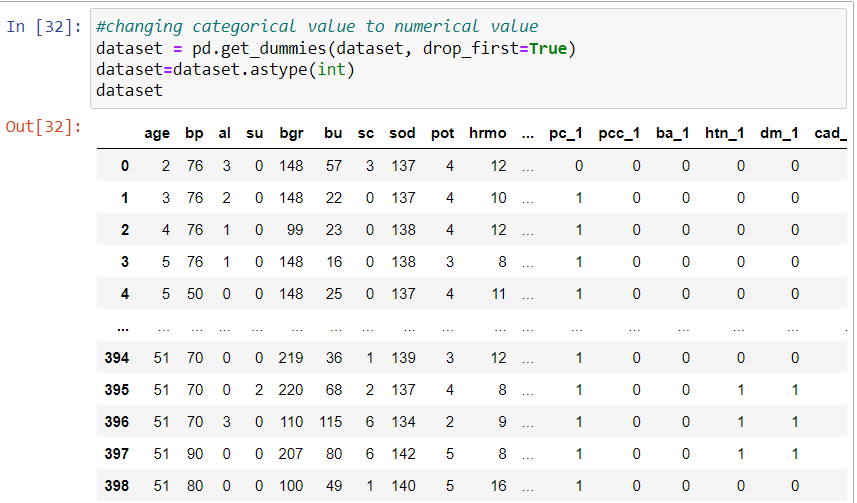


* For independent (i.e.) input columns: **399 rows, 27 columns.**
* For dependent (i.e.) input columns: **399 rows, 1 column.**



1. Mention the pre-processing method if you’re doing any (like converting string to number – nominal data)

* In dataset some columns contain **categorical value** so in pre-processing step did **one-hot encoding** using **get\_dummies.**



1. Develop a good model with good evaluation metric. You can use any machine learning algorithm; you can create many models. Finally, you have to come up with final model.

Developed a classification model: -

* Random Forest classifier
* Decision Tree Classifier
* KNN
* SVM
* Naïve Bayes classifier
* Logistic regression

Random Forest and Logistic Regression is giving Higher Accuracy of 99%.

1. All the research values of each algorithm should be documented. (You can make tabulation or screenshot of the results.)

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| S. No | Algorithm | Recall Yes-1 | Recall  No-0 | Precision  Yes-1 | Precision  No-0 | F1 Score Yes-1 | F1 Score  No-0 | ROC\_AUC | Accuracy |
| 1 | Random Forest classifier | 0.99 | 1.00 | 1.00 | 0.98 | 0.99 | 0.99 | 0.999 | 0.99 |
| 2 | Decision Tree Classifier | 0.93 | 0.98 | 0.99 | 0.89 | 0.96 | 0.93 | 0.953 | 0.95 |
| 3 | SVM | 0.98 | 1.00 | 1.00 | 0.96 | 0.99 | 0.98 | 0.987 | 0.98 |
| 4 | Logistic regression | 0.99 | 1.00 | 1.00 | 0.98 | 0.99 | 0.99 | 0.993 | 0.99 |
| 5 | KNN | 0.68 | 0.84 | 0.88 | 0.62 | 0.77 | 0.72 | 0.763 | 0.74 |
| 6 | Gaussian NB | 0.96 | 1.00 | 1.00 | 0.94 | 0.98 | 0.97 | 0.945 | 0.98 |
| 7 | Complement NB | 0.73 | 0.98 | 0.98 | 0.69 | 0.84 | 0.81 | 0.856 | 0.83 |
| 8 | Multinomial  NB | 0.73 | 0.98 | 0.98 | 0.69 | 0.84 | 0.81 | 0.856 | 0.83 |

1. Mention your final model, justify why u have chosen the same.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| S. No | Algorithm | Recall Yes-1 | Recall  No-0 | Precision  Yes-1 | Precision  No-0 | F1 Score Yes-1 | F1 Score  No-0 | ROC\_AUC | Accuracy |
| 1 | Random Forest classifier | 0.99 | 1.00 | 1.00 | 0.98 | 0.99 | 0.99 | 0.999 | 0.99 |
| 4 | Logistic regression | 0.99 | 1.00 | 1.00 | 0.98 | 0.99 | 0.99 | 0.993 | 0.99 |

* Here Logistic regression and Random Forest Giving better accuracy than other model.
* While comparing random forest and Logistic regression above ROC\_AUC is higher in RF than Logistic regression.