**Classification – Assignment**

1.) Identify your problem statement

* Stage 1 – Domain Selection – Numerical Values in the Dataset

Machine Learning

* Stage 2 – Learning Selection – Input and Output are well defined

Supervised Learning

* Stage 3 – Regression or Classification – Classifying the output based on the input parameter

Classification

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

The dataset contains total 399 rows × 25 columns. The first 24 columns are considered as input data and the last column is considered as output or target value.

3.) Mention the pre-processing method if you’re doing any (like converting

string to number – nominal data)

In the dataset, there were total 12 columns represent the categorical value. The categorical value has been changed into numnercial value using pd.get\_dummies. StandardScaler preprocessing has been done to bring the dataset in a standard range.

4.) 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.

Many models have been created with the evaluation metrics such as accuracy table, scoring and roc\_auc.

5.) All the research values of each algorithm should be documented. (You

can make tabulation or screenshot of the results.)

**Logistic Regression:**



**K Nearest Neighbor:**



**GaussianNB:**



**BernoulliNB:**



**Support Vector Machine:**



**Decision Tree:**



**Random Forest:**



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

**Final Model**

**Logistic Regression** and **Support Vector Machine** models are the best models. For this dataset these two models work well and the prediction is close to 100%. The accuracy is 99% and the roc\_auc value is 1 for both models.