PREDICTIVE ANALYTICS USING STATISTICS

(UCS654)

ASSIGNMENT - SAMPLING

NAME – Ravinshu Kushwaha

ROLL NO.- 102003156

GROUP – 3CO7

	Simple Random	Systematic	Stratified	Cluster
Naive Bayes	0.7957	0.6403	0.9578	0.9868
SVM	0.5402	0.4774	0.9406	0.9462
KNN	0.9185	0.9467	0.9969	0.9949
Logistic Regression	0.9071	0.9184	0.9938	0.9944
Dummy Classifier	0.5267	0.5130	0.5000	0.6586

- Number of 1's class in original dataset: 9;
  Number of 0's class in original dataset: 763.
  Balancing of dataset done using SMOTE technique.
- After balancing the dataset contained 763 observations of both the classes.

SMOTE: Popular technique used to balance imbalanced datasets in machine learning. When a dataset is imbalanced, meaning that one class is under-represented compared to the other(s), the model trained on such a dataset may not perform well on the under-represented class.

SMOTE works by creating synthetic examples for the minority class, which helps to balance the dataset. It does this by randomly selecting an example from the minority class, and then selecting one or more of its nearest neighbours. It then creates new examples by interpolating between the original example and its selected neighbours. This results in new examples that are similar to the original ones, but are slightly different and help to augment the minority class.

SMOTE has been shown to be effective in improving the performance of machine learning models, particularly in cases where the minority class is significantly under-represented. However, it is important to note that SMOTE may not always work well for all types of datasets, and it is important to carefully evaluate the performance of the model before and after applying SMOTE.

- After calculating the of samples for various sampling methods, applied PyCaret library to all the samples to find the accuracies.
  The table above shows the accuracies of some models on samples obtained using different sampling methods.

- From the above table we can see that CLUSTER SAMPLING is most suitable for the provided dataset.