PREDICTIVE ANALYTICS USING STATISTICS

(UCS654)

ASSIGNMENT - SAMPLING

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Simple Random | Systematic | Stratified | Cluster |
| Random Forest | 0.9971 | 0.9962 | 0.9984 | 0.9944 |
| Ridge Classifier | 0.9791 | 0.9849 | 0.9812 | 0.9906 |
| Decision Tree | 0.9941 | 0.9925 | 0.9969 | 0.9906 |
| Logistic Regression | 0.9941 | 0.9923 | 0.9938 | 0.9944 |
| SVM | 0.9615 | 0.9439 | 0.9406 | 0.9462 |

* The original dataset had 9 observations belonging to class 1 and 763 observations belonging to class 0. To balance the dataset, the SMOTE technique was used to create synthetic observations for the minority class. After applying SMOTE, the dataset had 763 observations for both classes.
* SMOTE creates synthetic observations by selecting examples close in feature space, drawing a line between them, and creating a new observation at a random point along that line. The process involves selecting a random observation from the minority class, finding k nearest neighbors (typically 5), selecting a random neighbor, and creating a synthetic observation between the two in feature space. This process can be repeated to generate as many synthetic observations as required. To balance the class distribution, random under-sampling was applied to the majority class, followed by SMOTE oversampling of the minority class.
* The PyCaret library was used to evaluate the accuracy of various models on the different samples obtained using various sampling methods. The table shows the accuracy results for some models on the different samples.