**Assignment**

**Part No 1**

**(a)**

Class A = 100 -----> 20 correct and 80 wrong

Class B = 900 -----> 800 correct and 100 wrong

Total = 1000

**Confusion Matrix**

|  |  |  |
| --- | --- | --- |
|  | Predicted A | Predicted B |
| Actual A | 20 | 80 |
| Actual B | 100 | 800 |

**(b)**

**Accuracy:**

Accuracy is defined as the calculation of all the correct predictions by model divided by the total number of the data.

Accuracy = (True Positive + True Negative) / (Positive + Negative)

Accuracy = (20 + 800) / (100 + 900) = **0.82**

**Error Rate:**

Error Rate is defined as the calculation of a number of all wrong predictions by model divided by the total number of the data.

Error Rate = (False Positive + False Negative) / (Positive + Negative)

Error Rate = (80 + 100) / (100 + 900) = **0.18**

**(c)**

**Precision:**

Precision is defined as the calculation of a number of all correct positive predictions by model divided by the total number of correct positive and wrong positive predictions.

Precision = (True Positive) / (True Positive + False Positive)

Precision = 20 / (20 + 80) = **0.20**

**Recall:**

Recall is defined as the calculation of a number of correct positive predictions by model divided by the total number of correct positive predictions and wrong positive predictions.

Recall = (True Positive) / (True Positive + False Negative)

Recall = 20 / (20 + 100) = **0.16**

**F1-measure:**

F1 is a measure that takes into account both recall and precision.

F1-measure = 2 \* ((Precision \* Recall) / (Precision + Recall))

F1-measure = 2 \* ((0.20 \* 0.16) / (0.20 + 0.16))

F1-measure = 2 \* (0.032 / 0.36)

F1-measure = **0.18**

**(d)**

Accuracy is not a good metric to measure the performance of model predictions under class imbalance. Due to this issue majority classes are predicted by the model accurately but minority classes are not predicted accurately. High accuracy may be misleading in this case. SMOTE can be applied to the data for balancing the classes to overcome this problem.

**Part No 2**

**(a)**

The following are the stages involved in knowledge discovery and data mining project:

* **Goal-Setting and Application Understanding:**

In this stage, we will define our goal about how to start the process and through which levels we have to go through to achieve the goal of the insurance company that wants to leverage its vast warehouse of customer data to perform market segmentation to know what the market they operate in looks like.

* **Data Selection and Integration:**

After setting the objectives and goals, the relevant data related to product portfolio, customer demographics, and social media data have to be gained to perform analysis. We can use the Extract Transform Load technique to get the desired data.

* **Data Cleaning and Prepossessing:**

In this process, we will examine the data to get an idea about its quality. Outliers noisy and redundant data will be removed, missing values will be imputed and all the values will be converted to the same scale for modeling purposes. Unnecessary features will also be removed to reduce the complexity of the data.

* **Data Mining:**

In this stage, we analyze the data to identify the distribution of variables and the relationship between them by using statistical analysis and data visualization.

* **Modeling:**

In this step, we apply different models to get predictions. We measure the performance of models based on metrics. In classification, we use accuracy, recall, precision, and f1-score. In regression, we use root mean squared error, mean squared error and mean absolute error to identify model performance. We usually select the best-performing model in our future predictions work.

* **Pattern Evaluation and Interpretation:**

In this step, we identify the most important findings relevant to our goal and understand all the mechanisms through the insights gained from the data.

* **Knowledge Discovery and Use:**

The information generated from the analysis, visualizations, and modeling are now presented in a report or tables for actionable insights and to prepare a strategy for the future.

**(b)**

Both descriptive and predictive data mining approaches can help to get desirable results according to the goal. K-means clustering can be applied to perform customer segmentation. It has the advantage that we can know the number of clusters with the help of the elbow method. It may have a disadvantage that other unsupervised machine learning models may perform better.

**(c)**

The following are the steps to perform K-means clustering for customer segmentation:

1. Apply scaling on the data
2. Apply the elbow method to the data.
3. Plot the values to know about the best number of clusters.
4. Use the best number of clusters in the K-means clustering model.
5. Analyze the output values from the K-means clustering model.

**Part 3**

**(a)**

**Resub:**

Resub refers to the re-substitution method. In this method, we train the model with whole data and then do predictions using the same data values.

**Hold-out-10%:**

Hold-out-10% refers to the hold-out method with 90% training and 10% testing. In this method, we split our data into a 90% training set and a 10% testing set. We provide a training set to the model for training and then give a testing set for predictions.

**xVal-10f:**

xVal-10f refers to 10 fold cross-validation method. In this method, we create 10 equal pieces of data out of which 9 are used for training and 1 used for testing. This process takes process continuously so that each piece becomes a testing set out of all sets of data.

**LOOCV:**

LOOCV refers to Leave-One-Out Cross-Validation. In this method, we use each observation as testing data and all other samples as training data. We create n-modes for n samples in the data.

**(b)**

1. **Accuracy:**

* Resub has an average accuracy of 96.2%. It is the highest among all.
* Hold-out-10% has an average accuracy of 91.7%. It is the smallest among all.
* xVal-10f has an average accuracy of 94.1. It is the 3rd highest among all.
* LOOCV has an average accuracy of 94.15. It is the 2nd highest among all.

1. **Standard Deviation:**

* Resub has the smallest standard deviation among all.
* Hold-out-10% has the highest standard deviation among all.
* xVal-10f has the 2nd smallest standard deviation among all.
* LOOCV has 3rd smallest standard deviation among all.

1. **Computational Cost:**

* Resub has low computational cost as we just have to prepare only one model and pass the data just one time.
* Hold-out-10% has a low computational cost as compared to the resub method as we have to make validation data from the training set. So the amount of data is decreased in training and the model learns in small time.
* xVal-10f has a high computational cost as compared to resub and Hold-out-10%. xVal-10f makes predictions using ten models which becomes expensive.
* LOOCV has the highest computational cost among all of the above as we have to create n models for n observations in the data. So it becomes the most expensive of all of the above methods.

**(c)**

**Resub:**

We should use the re-substitution method when we have synthetically generated data or the data out of which we would not find any other sample.

**Hold-out-10%:**

When we have very large data and we are on short time, we can apply the Hold-out-10% method to initially build the model.

**xVal-10f:**

We use 10-fold cross-validation when having small data. We split our data into the training set, validation set, and testing set. In the training set, we made one split for validation and the other 9 splits for training. We repeat this process until each piece becomes a validation set in the training set.

**LOOCV:**

LOOCV is used when we have a small limited amount of data and we want as much training as possible when fitting to the algorithm. We should use LOOCV when we have samples less than 75 approximately.