**Question 1**

A major problem in bioinformatics analysis or medical science is in attaining the correct diagnosis of certain important information. For the ultimate diagnosis, normally, many tests generally involve the clustering or classification of large scale data. All of these test procedures are said to be necessary in order to reach the ultimate diagnosis. However, on the other hand, too many tests could complicate the main diagnosis process and lead to the difficulty in obtaining the end results, particularly in the case where many tests are performed. This kind of difficulty could be resolved with the aid of machine learning which could be used directly to obtain the end result with the aid of several artificial intelligent algorithms which perform the role as classifiers.

**Dataset Information**

The data used in this investigation is the breast cancer data. For training and testing, only 75% of the overall data is used for training and the rest is used for testing the accuracy of the classification of the selected classification methods.

The ARFF file is provided to load the data into WEKA software for the classification of the Cancer.

1. Indicate total number of instances, the number of attributes and number of samples under each class of power quality problems along with a bar graph.
2. The data load into WEKA is used to train the data mining algorithms: J48, SVM and Random Forest for classification purpose. After training, the algorithms are tested based on the given training set and as well as using stratified 10-fold cross-validation.
   1. Indicated the results obtained after testing the algorithms using training set.
   2. Indicated results obtained after testing the algorithms using stratified 10-fold cross validation
3. Compare results of classifiers based on precision, recall, f1 score, and which take less time.
4. Talk about any possible challenges with this problem that could arise throughout the classification process.

**Question 2**

There can be an imbalance dataset provided for classification. By imbalance dataset we mean that one of the two class has very less number of samples compared to number of samples in the other class,|C2| << |C1|.Then C2 is called the minority class, and C1 is called the majority class. The minority class is of our interest. The machine learning algorithm always performs well if it is given balanced dataset, but this is not always the case , as an example the dataset for fraud detection ,will have less number of fraud transactions than genuine transaction. Anomaly detection, medical diagnostic and fault monitoring are other examples. The prediction in case of unbalanced dataset is biased towards majority class.

The approach to solve this problem is sampling based approach. This is also known as data level approach. It works by artificially balancing the instances of class in the dataset. To artificially balance the class we apply resampling technique, such as random under sampling the majority class, random oversampling of minority class, and Synthetic Minority Over-Sampling Technique (SMOTE).

* In random undersampling of majority class approach we try to balance the class distribution in the dataset by randomly throwing away some data samples from majority class.
  + Weka filter path **weka.filters.supervised.instance.SpreadSubsample**
* In random oversampling of minority class approach we balance the class distribution by the random replication of minority class instances,to increase their number.
  + Weka filter path **weka.filters.supervised.instance.Resample**
* To reduce the problem of overfitting a method of creating synthetic instances of minority class is used. This technique is known as the synthetic minority over-sampling technique (SMOTE).
  + Weka filter path **weka.filters.supervised.instance.SMOTE**

1. Data preprocessed before evaluation

|  |  |  |
| --- | --- | --- |
|  | Instances of Majority Class | Instances of Minority Class |
| Without Handling Class imbalance problem(No Filter Applied) |  |  |
| Random Under Sampling of Majority Class (Filter Applied: SpreadSubsample) |  |  |
| Random Over Sampling of Minority Class (Filter Applied: Resample) |  |  |
| SMOTE (Filter Applied: SMOTE) |  |  |

1. The data load into WEKA is used to train the data mining algorithms: J48, Naïve Bayes , and SVM for classification purpose. For evaluation split the data set into two sets, training dataset and test dataset. The splitting is generally two-thirds training and one-third test. We use a training data set to build the classifier and then we use this classifier for prediction on test dataset.
   1. Indicated the results obtained after testing the algorithms using training set.
   2. Indicated results obtained after testing the algorithms using testing set.
2. Compare results of classifiers based on precision, recall, f1 score, and which take less time.
3. Talk about any possible challenges with this problem that could arise when we do undersampling of majority class, random oversampling of minority class and synthetic minority over-sampling technique (SMOTE).