Factors affecting classification performance/accuracy.

1. **Data scaling/Normalization**: As we had earlier seen in the Data pre-processing unit normalization plays a very important role in data mining algorithms especially in classification. If the data is not normalized then the classification algorithm may be biased towards a particular set of attributes hence resulting in poor accuracy values. The normal methods of Decimal Scaling, Min-Max normalization or Zero mean normalization can be applied here in this case.
2. **Presence of outliers:** Outliers are data points which are inconsistent with the majority of the data. Outliers are also points which are tough to classify because we can say that they do not possess the similar feature vector properties as the majority of the data. So classifying outliers is a risky task and can affect the accuracy adversely. A good method which can be employed here in advance is to remove the outliers. There are basically few good methods to do this which are Clustering and Curve fitting. Please find more information on Curve fitting in the References section.
3. **Presence of noise:** Noise can typically be defined as the random error or variance in a measured variable of which the two typical types are inconsistent values for features or classes. Noise is typically a minority in the dataset. Noise can be removed by using the following methods which are Smoothing and Clustering.
4. **Redundant attributes:** Generally data sometimes may have few sets of attributes which do not provide much information to the classifier. It means that these attributes do not form important entries in the feature vector. Removal of such attributes can speed up the performance of the classifier. This can be done by using information such as Information gain , Mutual information of attributes and Chi-square ratio. Dimensionality reduction is another broad area in this aspect which provides us with novel algorithms like PCA(Principal component analysis), LDA(Linear Discriminant Analysis) (Details of which are not necessary right now). These methods help in reducing the feature space effectively.
5. **Overfitting and Underfitting:** This has been explained in the previous resource document. Overfitting generally degrades the classifier accuracy. The methods mostly used to avoid overfitting will be looked in the Decision trees module which will come up later in the unit.
6. **Lack of diverse training data:** As already explained that over algorithm uses the training data to build the classification model so it is generally expected that the training data consists of instances which are different from each other. It would also be ideal that none of the instances provide same kind of information. This would help the classifier in learning more kinds of patterns. At the same time the quantity

of training data is also very important because low quantities of training data could lead to overfitting. There is a general field called Active Learning in machine learning which deals with this specific problem of choosing most informative training samples for classification.

7. **Wrong estimation methods:** Classification accuracy ideally should not be measured in a single experiment. Cross validation is a very good way of measuring accuracy which uses a leave one out kind of procedure and averages the accuracy for all the iterations. But still accuracy is a subjective measure and it cannot provide us with full information on the performance of the classifier. There are still other better metrics such as precision recall etc which can also be applied. Details of all these will be provided in the last module of this unit which deals with Evaluation of Classifiers.