# Coursework Description

## Ensemble Classifier Evaluation

You have been retained as a data scientist and suppose you have collected a dataset from UCI, excluding IRIS, of already-classified instances and you have to build an ensemble type classifier.

Ensembles can give you a boost in accuracy on your dataset. You can create ensembles of machine learning algorithms in R. There are three main techniques (Boosting, Bagging and Stacking) that you can create an ensemble of machine learning algorithms in R.

The three most popular methods for combining the predictions from different models are:

- **Bagging**: Building multiple models (typically of the same type) from different subsamples of the training dataset.
- **Boosting**: Building multiple models (typically of the same type) each of which learns to fix the prediction errors of a prior model in the chain.
- **Stacking**. Building multiple models (typically of differing types) and supervisor model that learns how to best combine the predictions of the primary models.

You can combine the predictions of multiple *caret* models using the ***caretEnsemble*** package. Given a list of caret models, the ***caretStack()*** function can be used to specify a higher-order model to learn how to best combine the predictions of sub-models together

This assignment is focused on **Bagging** and **Stacking** and on how you can continue to ratchet up the accuracy of the models on your own datasets.

### Bagging Algorithms

The base type bagging machine learning algorithms that will be examined in this assignment are:

- Bagged CART,
- Random Forest

### Stacking Algorithms

The base type stacking machine learning algorithms that will be examined in this assignment are

- Classification and Regression Trees (CART),
- K-Nearest Neighbors (KNN),
- Naïve Bayes (NB)

***Main-Question*: How will you know how good your ensemble classifier is? Under which conditions ensemble learning is useful?**

## 1st Task: Data Set Selection and Visualisation

You need to select a data set of your own choice (i.e. you may use a dataset already used before in the lab, or from the literature review) for the purposes of building training and validating the above type of classifiers (Bagging, Stacking). With the aid of R package visualise and justify the properties of the selected data set.

**[15 Marks]**

## 2nd Task: Formation of Training and Test Sets

Assuming we have collected one large dataset of already-classified instances, you need to look into methods of forming training and test sets from this single dataset in R as described below.

**Repeated k-fold Cross Validation**

The process of splitting the data into k-folds can be repeated a number of times; this is called Repeated k-fold Cross Validation (repeatedcv). The final model accuracy is taken as the mean from the number of repeats.

**[10 Marks]**

## 3rd Task: Build Train and Test a Bagging type Classifier

You need to construct, train and test a Bagging type classifier in R, based on Bagged CART and Random Forest base classifiers. Train and test the Bagging classifier using the training and test sets generated based on the method tried as part of the **2nd Task**.

**[20 Marks]**

## 4th Task: Build Train and Test a Stacking type Classifier

You need to construct, train and test a Stacking type classifier in R, based on (CART, KNN, NB). Train and test your Stacking classifier using the training and test sets generated based on the method tried as part of the **2nd Task**.

**[25 Marks]**

## 5th Task: Measure Performance

For each type of ensemble type classifier calculate and display the following performance related metrics in R. Critically comment on the importance of each metric for each type of ensemble type classifier. Use the library library(ROCR)

1. Confusion matrix
2. Precision vs. Recall
3. Accuracy
4. ROC(receiver operating characteristic curve)
5. RAUC (receiver under the curve area)
6. Training time
7. Testing time
8. Based on the above Metrics briefly discuss, how we can increase the reliability and consistency of the data classification task at hand.

**[30 Marks]**

# Coursework Marking scheme

The Coursework will be marked based on the following marking criteria:

## 1st Task: Data Set Selection and Visualisation

| | |
|---|---|
| • Data Set summary of main properties | 5 |
| • Visualisation in R of main data set properties | 5 |
| • Feature Selection | 5 |

## 2nd Task: Formation of Training and Test Sets

Formation of training and test sets from in R using the methods below.

| | |
|---|---|
| • Repeated CV for Bagging type classifier | 5 |
| • Repeated CV for Stacking type classifier | 5 |

## 3rd Task: Build Train and Test a Bagging type Classifier type Classifier

| | |
|---|---|
| • Building of Random Forest type classifier in R | 5 |
| • Building of Bagged CART type classifier in R | 5 |
| • Testing of Bagging type classifier | 10 |

## 4th Task: Build Train and Test a Stacking type Classifier

| | |
|---|---|
| • Building of Stacking CART classifier in R | 5 |
| • Building of Naïve Bayes type classifier in R | 5 |
| • Building of K-NN type classifier in R | 5 |
| • Testing of Stacking type classifier in R | 10 |

## 5th Task: Measure Performance

| | |
|---|---|
| • Confusion matrix estimation | 4 |
| • Precision vs. Recall estimation | 4 |
| • Accuracy estimation | 4 |
| • ROC(receiver operating characteristic curve) plot | 3 |
| • RAUC (receiver under the curve area) plot | 3 |
| • Training time | 3 |
| • Testing time | 3 |
| • Based on the above Metrics briefly discuss, how we can increase the reliability and consistency of the data classification task at hand. | 6 |