Minor Exercise 1 : Data Discovery

**Due April 13th**

Using a chosen dataset from the below links (except Iris and Boston) compare and contrast with each of the following methods

-       Regression

-       Clustering

-       SVM

-       Ensemble

-       Neural Network

Submission should be in the form of a document outlining data assumptions made before applying models to dataset (e.g. increase in rooms, increase in price), the strengths and weakness to each model, the accuracy of each model using cross validation and a conclusion which will outline the best model for the chosen dataset and why.

Dataset Links:

<https://archive.ics.uci.edu/ml/index.php>

<https://www.kaggle.com/datasets>

Minor Exercise 2: Text Classiification

**Due April 13th**

Using the Newsgroups dataset and the following machine learning methods

-       SVM

-       Naive Bayes

-       Neural Network

You will need to use the techniques of **stopping**(removing small insignificant words eg I, the, you etc)**, stemming**(removing the endings of words eg -ed -ing) and **use of TF/IDF** (Term Frequency over Item Document Frequency) to aid in the classification of the type of news report

This is a task which will require you to do some feature engineering to get decent accuracy

The submission will be the source code which will output a confusion matrix and overall accuracy of each classifier

 Dataset:

sklearn.datasets import fetch\_20newsgroups

<https://archive.ics.uci.edu/ml/index.php>

<https://www.kaggle.com/datasets>

Visualising Twitter Data (Updated)

**Project Phases:**

**Phase 1**: Use tweepy to gain a critical mass of tweets regarding a chosen topic and annotate the sentiment of the tweets; positive, negative, neutral.

    -     Clean and organise the data, remove duplicates and use techniques from text classification exercise such as stopping and stemming to  prepare the dataset.

**Phase 2**: Create a model which can accurately assign a  polarity to a group of new tweets within the chosen topic. Test your model against a different domain eg politics model on sports or tech.

**---------------------------------------------------------------------**

**Deliverables**

1. A dataset with annotated training data for sentiment polarity

2. A model which predicts polarity of given tweet

3. A report which includes a comparison with VADER(an off the shelf sentiment analysis tool)