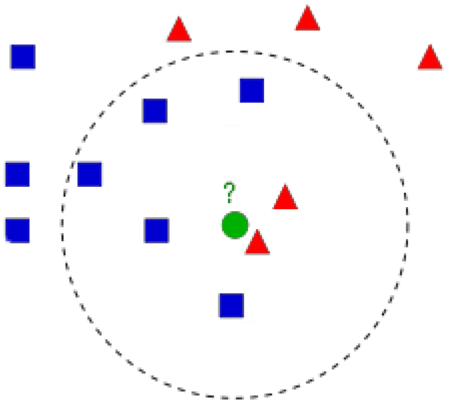
1 - Classification Challenge

K-Nearest Neighbors (kNN) is one of the simplest of classification algorithms available for supervised learning. The idea is to search for closest matches of the test data in order to classifying it; you can find more details on kNN and how it works online (YouTube, blogs,..)
A practical application of kNN can be summarised as following: given the below image where we have Blue Squares representing boys and Red Triangles girls, we want to classify the gender green circle (test data) based on closest neighbors.
In this example we set k=7 hence we explore the 7 nearest neighbors and by voting, we classify the green circle as a boy since we have 5 boys and 2 girls within the neighborhood (dotted line)



Exercise

We have a dataset containing details of primary school students.
The Training file contains students information including id (UserID), gender and other features such as:
Ageyears,Handed,Height,Foot_Length,Arm_Span,Languages_spoken,Travel_to_School,Travel_time_to_School,Reaction_time,Score_in_memory_game.
As per above paragraph, we want to estimate the gender of the 20 students in the Test file.
From here you can download the two files:

Training CSV file:
https://drive.google.com/file/d/0B5mKxO6QIkSjcXN0U20zTFNMZDg/view?usp=sharing
Test CSV file:
https://drive.google.com/file/d/0B5mKxO6QIkSjNUxpclc1cFQ5dG8/view?usp=sharing

For this exercise, the classification output should be a csv having 2 columns as per following example:

```
UserID,Gender
1,F
2,F
3,F
4,M
5,F
6,F
7,M
8,M
9,F
10,F
11,F
12,M
13,M
14,M
15,M
16,F
17,F
18,M
19,M
20,M
```

Questions:

1 - Implement an algorithm that classifies the gender of the Test students by using kNN. The code can be in your favorite language (C/C++, Python, Java, Matlab, R, ...).
Make sure you attach the source code in your answer.

2 - Save your classification output in CSV format and attach it to your answer

3 - Using a cross-validation method please comment on the classification performances you observe.
3.1 - which is the optimal value for K?

How could you improve the classification (Attach additional code if needed)
4.1 - try to apply feature selection in order to improve classification accuracy. Does it work?
4.4 - Instead of kNN try a different classifier, have you improved the classification accuracy?
4.5 - finally, suggest other ideas that can lead to classification accuracy enhancement.

Ideas on how to start implementing your answer:

A\ Save on your local PC Train.csv and Test.csv
B\ With your favorite language read both CSV files
C\ Import a library that implements kNN or write your own version; here following some links anyhow you could employ different ones.

java http://java-ml.sourceforge.net/content/classification-basics
http://docs.opencv.org/2.4/modules/ml/doc/k_nearest_neighbors.html
C/C++ http://docs.opencv.org/2.4/modules/ml/doc/k_nearest_neighbors.html
python http://scikit-learn.org/stable/modules/neighbors.html
R https://stat.ethz.ch/R-manual/R-devel/library/class/html/knn.html
Matlab https://it.mathworks.com/help/stats/classificationknn-class.html

D\ Prepare your data in order to be compatible with the algorithm you decided to use. Please note that each student is identified by a UserID which needs to be removed when running the classification.

E\ Classify the 20 students in Test.csv and save the classification output in a CSV file.
F\ Carry on working and answering as many questions as you can.