INTRODUCTION TO MACHINE LEARNING

Project (**by groups of 2 students at the most**)

Objectives: apply your skills in Machine Learning and practice Python programming

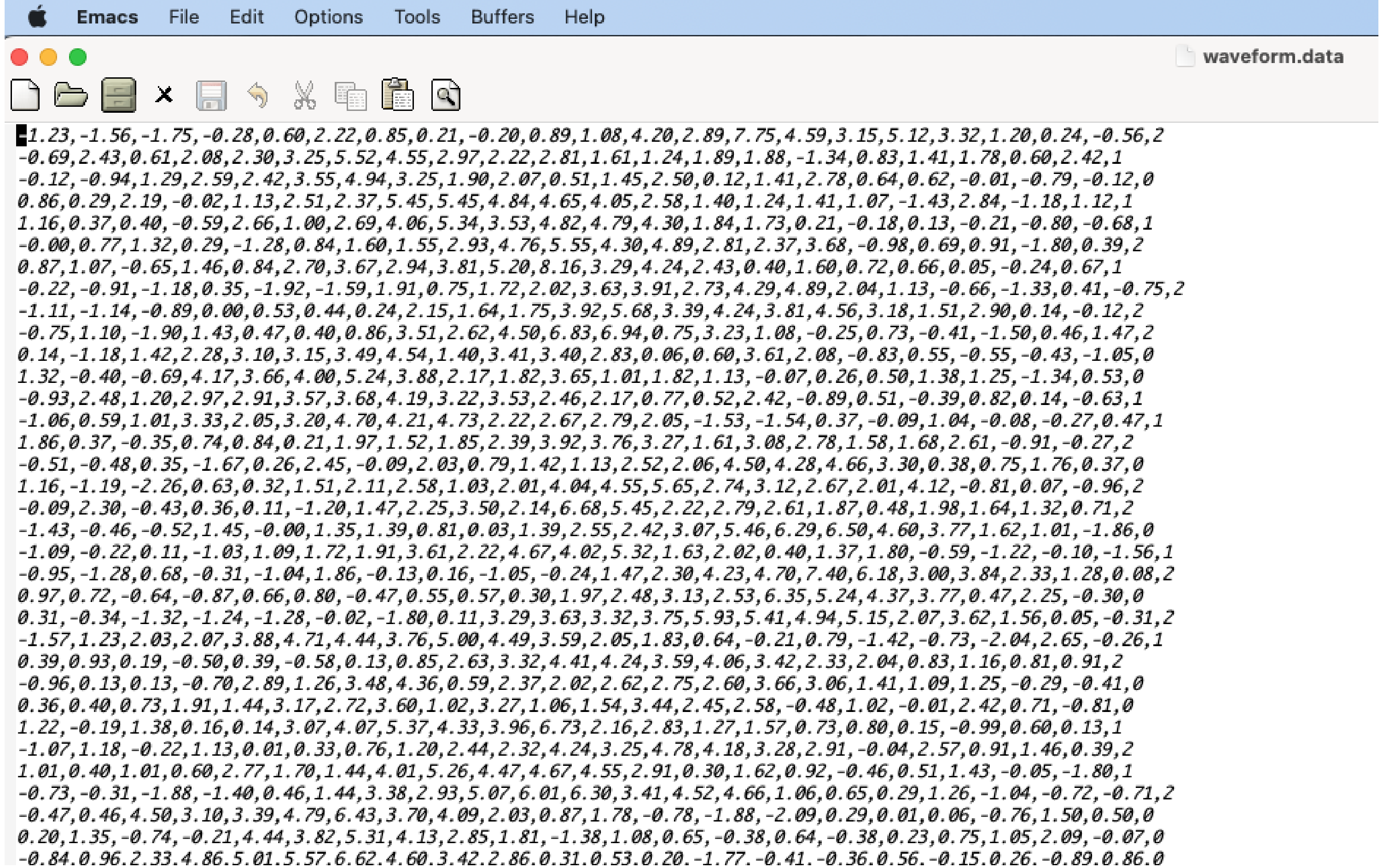
By October, 11th, send me an email with the composition of the team.
By December, 20th, upload on MOODLE your report (3 pages) along with your ipynb jupyter notebook file. Use the Latex style https://media.icml.cc/Conferences/ICML2024/Styles/icml2024.zip

Your report has to be self-content and must contain: title, authors, abstract, introduction, contributions, experimental set-up, analysis of the results, conclusion, references

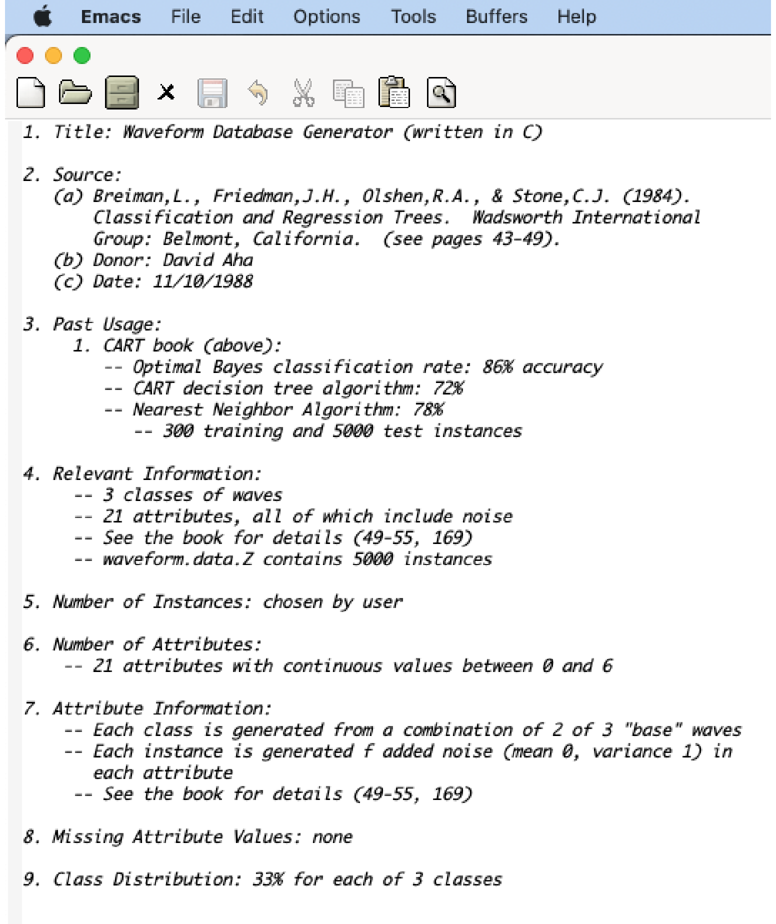
Input : the dataset **waveform.data (available on MOODLE)**

- 5000 data
- 3 classes of waves
- 21 attributes all of which include noise
- Optimal Bayes classification rate = 86% accuracy

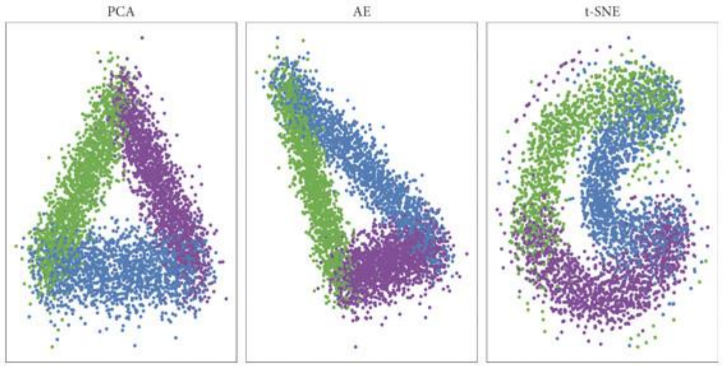
waveform.data



```
-1.23,-1.56,-1.75,-0.28,0.60,2.22,0.85,0.21,-0.20,0.89,1.08,4.20,2.89,7.75,4.59,3.15,5.12,3.32,1.20,0.24,-0.56,2
-0.69,2.43,0.61,2.08,2.30,3.25,5.52,4.55,2.97,2.22,2.81,1.61,1.24,1.89,1.88,-1.34,0.83,1.41,1.78,0.60,2.42,1
-0.12,-0.94,1.29,2.59,2.42,3.55,4.94,3.25,1.90,2.07,0.51,1.45,2.50,0.12,1.41,2.78,0.64,0.62,-0.01,-0.79,-0.12,0
0.86,0.29,2.19,-0.02,1.13,2.51,2.37,5.45,5.45,4.84,4.65,4.05,2.58,1.40,1.24,1.41,1.07,-1.43,2.84,-1.18,1.12,1
1.16,0.37,0.40,-0.59,2.66,1.00,2.69,4.06,5.34,3.53,4.82,4.79,4.30,1.84,1.73,0.21,-0.18,0.13,-0.21,-0.80,-0.68,1
-0.00,0.77,1.32,0.29,-1.28,0.84,1.60,1.55,2.93,4.76,5.55,4.30,4.89,2.81,2.37,3.68,-0.98,0.69,0.91,-1.80,0.39,2
0.87,1.07,-0.65,1.46,0.84,2.70,3.67,2.94,3.81,5.20,8.16,3.29,4.24,2.43,0.40,1.60,0.72,0.66,0.05,-0.24,0.67,1
-0.22,-0.91,-1.18,0.35,-1.92,-1.59,1.91,0.75,1.72,2.02,3.63,3.91,2.73,4.29,4.89,2.04,1.13,-0.66,-1.33,0.41,-0.75,2
-1.11,-1.14,-0.89,0.00,0.53,0.44,0.24,2.15,1.64,1.75,3.92,5.68,3.39,4.24,3.81,4.56,3.18,1.51,2.90,0.14,-0.12,2
-0.75,1.10,-1.90,1.43,0.47,0.40,0.86,3.51,2.62,4.50,6.83,6.94,0.75,3.23,1.08,-0.25,0.73,-0.41,-1.50,0.46,1.47,2
0.14,-1.18,1.42,2.28,3.10,3.15,3.49,4.54,1.40,3.41,3.40,2.83,0.06,0.60,3.61,2.08,-0.83,0.55,-0.55,-0.43,-1.05,0
1.32,-0.40,-0.69,4.17,3.66,4.00,5.24,3.88,2.17,1.82,3.65,1.01,1.82,1.13,-0.07,0.26,0.50,1.38,1.25,-1.34,0.53,0
-0.93,2.48,1.20,2.97,2.91,3.57,3.68,4.19,3.22,3.53,2.46,2.17,0.77,0.52,2.42,-0.89,0.51,-0.39,0.82,0.14,-0.63,1
-1.06,0.59,1.01,3.33,2.05,3.20,4.70,4.21,4.73,2.22,2.67,2.79,2.05,-1.53,-1.54,0.37,-0.09,1.04,-0.08,-0.27,0.47,1
1.86,0.37,-0.35,0.74,0.84,0.21,1.97,1.52,1.85,2.39,3.92,3.76,3.27,1.61,3.08,2.78,1.58,1.68,2.61,-0.91,-0.27,2
-0.51,-0.48,0.35,-1.67,0.26,2.45,-0.09,2.03,0.79,1.42,1.13,2.52,2.06,4.50,4.28,4.66,3.30,0.38,0.75,1.76,0.37,0
1.16,-1.19,-2.26,0.63,0.32,1.51,2.11,2.58,1.03,2.01,4.04,4.55,5.65,2.74,3.12,2.67,2.01,4.12,-0.81,0.07,-0.96,2
-0.09,2.30,-0.43,0.36,0.11,-1.20,1.47,2.25,3.50,2.14,6.68,5.45,2.22,2.79,2.61,1.87,0.48,1.98,1.64,1.32,0.71,2
-1.43,-0.46,-0.52,1.45,-0.00,1.35,1.39,0.81,0.03,1.39,2.55,2.42,3.07,5.46,6.29,6.50,4.60,3.77,1.62,1.01,-1.86,0
-1.09,-0.22,0.11,-1.03,1.09,1.72,1.91,3.61,2.22,4.67,4.02,5.32,1.63,2.02,0.40,1.37,1.80,-0.59,-1.22,-0.10,-1.56,1
-0.95,-1.28,0.68,-0.31,-1.04,1.86,-0.13,0.16,-1.05,-0.24,1.47,2.30,4.23,4.70,7.40,6.18,3.00,3.84,2.33,1.28,0.08,2
0.97,0.72,-0.64,-0.87,0.66,0.80,-0.47,0.55,0.57,0.30,1.97,2.48,3.13,2.53,6.35,5.24,4.37,3.77,0.47,2.25,-0.30,0
0.31,-0.34,-1.32,-1.24,-1.28,-0.02,-1.80,0.11,3.29,3.63,3.32,3.75,5.93,5.41,4.94,5.15,2.07,3.62,1.56,0.05,-0.31,2
-1.57,1.23,2.03,2.07,3.88,4.71,4.44,3.76,5.00,4.49,3.59,2.05,1.83,0.64,-0.21,0.79,-1.42,-0.73,-2.04,2.65,-0.26,1
0.39,0.93,0.19,-0.50,0.39,-0.58,0.13,0.85,2.63,3.32,4.41,4.24,3.59,4.06,3.42,2.33,2.04,0.83,1.16,0.81,0.91,2
-0.96,0.13,0.13,-0.70,2.89,1.26,3.48,4.36,0.59,2.37,2.02,2.62,2.75,2.60,3.66,3.06,1.41,1.09,1.25,-0.29,-0.41,0
0.36,0.40,0.73,1.91,1.44,3.17,2.72,3.60,1.02,3.27,1.06,1.54,3.44,2.45,2.58,-0.48,1.02,-0.01,2.42,0.71,-0.81,0
1.22,-0.19,1.38,0.16,0.14,3.07,4.07,5.37,4.33,3.96,6.73,2.16,2.83,1.27,1.57,0.73,0.80,0.15,-0.99,0.60,0.13,1
-1.07,1.18,-0.22,1.13,0.01,0.33,0.76,1.20,2.44,2.32,4.24,3.25,4.78,4.18,3.28,2.91,-0.04,2.57,0.91,1.46,0.39,2
1.01,0.40,1.01,0.60,2.77,1.70,1.44,4.01,5.26,4.47,4.67,4.55,2.91,0.30,1.62,0.92,-0.46,0.51,1.43,-0.05,-1.80,1
-0.73,-0.31,-1.88,-1.40,0.46,1.44,3.38,2.93,5.07,6.01,6.30,3.41,4.52,4.66,1.06,0.65,0.29,1.26,-1.04,-0.72,-0.71,2
-0.47,0.46,4.50,3.10,3.39,4.79,6.43,3.70,4.09,2.03,0.87,1.78,-0.78,-1.88,-2.09,0.29,0.01,0.06,-0.76,1.50,0.50,0
0.20,1.35,-0.74,-0.21,4.44,3.82,5.31,4.13,2.85,1.81,-1.38,1.08,0.65,-0.38,0.64,-0.38,0.23,0.75,1.05,2.09,-0.07,0
-0.84.0.96.2.33.4.86.5.01.5.57.6.62.4.60.3.42.2.86.0.31.0.53.0.20.-1.77.-0.41.-0.36.0.56.-0.15.0.26.-0.89.0.86.0
```



```
1. Title: Waveform Database Generator (written in C)

2. Source:
   (a) Breiman,L., Friedman,J.H., Olshen,R.A., & Stone,C.J. (1984).
       Classification and Regression Trees.  Wadsworth International
       Group: Belmont, California.  (see pages 43-49).
   (b) Donor: David Aha
   (c) Date: 11/10/1988

3. Past Usage:
     1. CART book (above):
        -- Optimal Bayes classification rate: 86% accuracy
        -- CART decision tree algorithm: 72%
        -- Nearest Neighbor Algorithm: 78%
           -- 300 training and 5000 test instances

4. Relevant Information:
     -- 3 classes of waves
     -- 21 attributes, all of which include noise
     -- See the book for details (49-55, 169)
     -- waveform.data.Z contains 5000 instances

5. Number of Instances: chosen by user

6. Number of Attributes:
    -- 21 attributes with continuous values between 0 and 6

7. Attribute Information:
    -- Each class is generated from a combination of 2 of 3 "base" waves
    -- Each instance is generated f added noise (mean 0, variance 1) in
       each attribute
     -- See the book for details (49-55, 169)

8. Missing Attribute Values: none

9. Class Distribution: 33% for each of 3 classes
```



List of possible experiments to perform (non exhaustive)

1. **Tune the best *k* of a *k*NN classifier** by cross-validation (plot the accuracies over the validation subset w.r.t. *k*) from 4000 randomly drawn **training** examples (you will keep apart 1000 waves for the **test** set).

2. **Reduce the complexity** by implementing and running **your own** Data Reduction algorithms, as studied in class on the training data. Compare the accuracy (with a 1NN) on the 1000 test waves before and after reduction of the training set.

3. Using the original dataset, compare (in terms of time) one of the two methods studied in class for **speeding-up the calculation** of the 1NN with a brute force 1NN algorithm.

4. **Split the 4000 data into 3000 training and 1000 validation data. Generate artificially imbalancy** in the training data (e.g. one class is more and more minority) and analyze the impact on the accuracy on the 1000 test waves (using the best k obtained in 1/). Then, tune **k** w.r.t. the F-measure using the validation data and see if this allows to control the accuracy decrease on the test data.