Assignment

In this assignment, we will compare Naive Bayesian and Decision Tree Classification for identifying normal vs. non-normal fetus status based on fetal cardiograms.

For the dataset, we use "fetal cardiotocography data set" at UCI:
`https://archive.ics.uci.edu/ml/datasets/Cardiotocography`

Dataset Description: From the website: "2126 fetal cardiotocograms (CTGs) were automatically processed and the respective diagnostic features measured. The CTGs were also classified by three expert obstetricians and a consensus classification label assigned to each of them. Classification was both with respect to a morphologic pattern (A, B, C. ...) and to a fetal state (N, S, P). Therefore the dataset can be used either for 10-class or 3-class experiments."

We will focus on the "fetal state". We will combine labels "S" (suspect) and "P" (pathological) into one class "A" (abnormal). We will focus on predicting "N" (normal) vs. "A" ("Abnormal"). For a detailed description of features, please visit the above website.

The data is an Excel (not csv) file. For ways to process excel files in Python, see `https://www.python-excel.org/`

You will use the following subset of 12 numeric features:

1. LB - FHR baseline (beats per minute)
2. ASTV - percentage of time with abnormal short term variability
3. MSTV - mean value of short term variability
4. ALTV - percentage of time with abnormal long term variability
5. MLTV - mean value of long term variability
6. Width - width of FHR histogram
7. Min - minimum of FHR histogram
8. Max - Maximum of FHR histogram
9. Mode - histogram mode
10. Mean - histogram mean
11. Median - histogram median
12. Variance - histogram variance

You will consider the following set of 4 features depending on your facilitator group.

- Group 1: LB, ALTV, Min, Mean
- Group 2: ASTV, MLTV, Max, Median
- Group 3: MSTV, Width, Mode, Variance
- Group 4: LB, MLTV, Width, Variance

For each of the questions below, these would be your features.

Question 1:

1. load the Excel ("raw data" worksheet) data into Pandas dataframe
2. combine NSP labels into two groups: N (normal - these labels are assigned) and Abnormal (everything else) We will use existing class 1 for normal and define class 0 for abnormal.

Question 2: Use Naive Bayesian NB classifier to answer these questions:

1. split your set 50/50, train NB on X_{train} and predict class labels in X_{test}
2. what is the accuracy?
3. compute the confusion matrix

Question 3: Use Decision Tree to answer these questions:

1. split your set 50/50, train NB on X_{train} and predict class labels in X_{test}
2. what is the accuracy?
3. compute the confusion matrix

Question 4: Recall that there are two hyper-parameters in the random forest classifier: N - number of (sub)trees to use and d - max depth of each subtree

Use Random Forest classifier to answer these questions:

1. take $N = 1, \ldots, 10$ and $d = 1, 2, \ldots, 5$. For each value of N and d, split your data into X_{train} and X_{test}, construct a random tree classifier (use "entropy" as splitting criteria - this is the default) Train you r classifier on X_{train} and compute the error rate for X_{test}
2. plot your error rates and find the best combination of N and d.
3. what is the accuracy for the best combination of N and k?
4. compute the confusion matrix using the best combination of N and d

Question 5: Summarize your results for Naive Bayesian, decision tree and random forest in a table below and discuss your findings.

Model	TP	FP	TN	FN	accuracy	TPR	TNR
naive bayesian							
decision tree							
random forest							