Ensemble methods

Tasks (Lab 9):
(**all tasks are scored**)

1. Implementation of bagging and boosting (AdaBoost)

   (a) Implement your own version of AdaBoost algorithm. You can use available implementations of decision trees (or other base learners). Below is a description of AdaBoost algorithm:

   Training:

   i. Define initial weights: $w_i = \frac{1}{n}$, $i = 1, \ldots, n$.
   ii. For $k = 1, \ldots, B$:
      A. Build classifier f_k with weights: w_i.
      B. Compute weighted classification error for k-th classifier:

$$\epsilon_k := \sum_{i=1}^{n} I[f_k(x) \neq y_i] w_i.$$

      C. Compute scaling factor: $\beta_k := \frac{\epsilon_k}{1-\epsilon_k}$.
      D. Add a pair (f_k, β_k) to the ensemble.
      E. For $i = 1, \ldots, n$:
      If $f_k(x_i) = y_i$ (example is correctly classified) then $w_i := w_i \beta_k$ (decrease weights).
      F. Normalize weights: $w_i = w_i / \sum(w_i)$.

   Prediction:

   - Prediction for new observation x:

$$\hat{y}(x) = \arg\max_y \sum_{k=1}^{B} I[f_k(x) = y] \log(\frac{1}{\beta_k})$$

   - Weighted voting.
   - Classifier with small errors (small value of β_k) have larger weights in majority voting.

2. (a) Compare the following ensemble methods:
   - Bagging
   - Boosting (AdaBoost), your implementation.
   - Random Forest
   - Single tree (without pruning)
   - Decision stump (tree of depth 1)
   - Gradient Boosting
   - XGboost

   (b) Consider one real dataset (you can choose it from e.g. UCI repository) and one artificial dataset generated as follows:
   - Generate $X_1, X_2, \ldots, X_{10} \sim N(0, 1)$.
   - Denote by $\chi^2_{10}(0.5)$ median of chi squared distribution with 10 degrees of freedom.
   - Set $Y = 1$ if $\sum_{j=1}^{10} X_j^2 > \chi^2_{10}(0.5)$ and $Y = -1$ otherwise.
   - Generate training data of size 2000 and testing data of size 10000.

   (c) Make a plot showing how the error changes with the number of iterations (for boosting and bagging).