Домашнее задание по машинному обучению №3

Дедлайн на полный балл – **16 марта**

Дедлайн на половинный балл – **23 марта**

Датасеты – spam.csv и cancer.csv

1. (2) Посчитайте ROC AUC для пороговых условий по всем признакам для обоих датасетов. Найдите 3 лучших признака для каждого датасета и постройте по ним ROC кривые.

2. (1) Разделите датасеты на train и validate выборки (80/20%), предварительно перемешав точки внутри датасет (можно и воспользоваться готовым методом). Убедитесь, что пропорции классов в подвыборках сохранились.

3. (5) Реализуйте алгоритм построения дерева с тремя вариантами функции impurity (Misclassification error, Entropy, Gini Impurity). Для каждого варианта найдите лучшую глубину дерева (по accuracy на validate) в интервале от 1 до 10.

4. (1) Для трех лучших вариантов из предыдущей задачи постройте ROC кривые и посчитайте ROC AUC.

Подсказка: по сути, нам нужно выдавать пропорцию (вероятность) класса в листе как новый признак, и по нему строить пороговое условие.