Лабораторна робота 2

Базові алгоритми класифікації з використанням бібліотеки sklearn

Провести навчання і класифікацію даних. Виконати наступні процедури:
1) Завантажити дані, вивести на екран назви колонок і розмір датасета
2) Опрацювати пропуски (по можливості заповнити їх або видалити)
3) Візуалізувати дані: побудувати графік (heatmap), що відображає кореляції ознак між собою і з цільовою змінною (розміткою); побудувати гістограми розподілу ознак і boxplot-и ознак відносно цільової змінної (якщо ознак занадто багато обмежитися декількома)
4) нормалізувати дані
5) провести навчання наступних класифікаторів

- kNN
- дерево прийняття рішень і візуалізувати його
- SVM
- Random Forest
- AdaBoost

Підібрати оптимальні параметри для

- kNN
- для SVM за допомогою GridSearch підібрати оптимальні «C» і «gamma»
- за допомогою GridSearch підібрати оптимальні параметри для Random Forest і AdaBoost

Серед обраних оптимальних моделей кожного класу вибрати найкращу. Відобразити sklearn.metrics.classification_report і sklearn.metrics.confusion_matrix

Як звіт – робочий код в Jupyter notebook заливаєте на свій репозиторій на https://github.com/. Лінк відправляєте на пошту natsakh-ipt@lll.kpi.ua

Максимальний бал – 10, 6 за роботу + 4 захист.
Deadline 25.10.21, після цього терміну максимальний бал зменшується на 1 кожні 2 тижні

Дані брати тут:
https://www.kaggle.com/data
https://archive.ics.uci.edu/ml/index

Обов'язкова вимога – унікальність даних (датасети не повинні повторюватись), обраний датасет заносите в табличку, попередньо перевіривши, що його ще ніхто не обрав https://docs.google.com/document/d/1DGOxYoDMmC2Kpk3viUhAWXvARBB0S1k-q2alpimm_Ic/edit?usp=sharing

також дані, що розглядалися на практичних брати не слід (Іриси, Титанік, Wine). Також не слід обирати MNIST (рукописні цифри)