Лабораторная работа 2

Базовые алгоритмы классификации с использованием библиотеки sklearn

Провести обучение и классификацию данных. Выполнить следующие процедуры:

1) Загрузить данные с сайта, считать и вывести на экран названия колонок и размер датасета
2) Обработать пропуски (по возможности заполнить их или удалить)
3) Визуализировать данные: построить график (heatmap) отображающий корреляции признаков между собой и с целевой переменной (разметкой); построить гистограммы распределения признаков и ящичковые (boxplot) диаграммы признаков относительно целевой переменной (если признаков слишком много ограничиться несколькими).
4) Масштабировать данные
5) Провести обучение следующих классификаторов:
    - kNN(sklearn.neighbors.KNeighborsClassifier)
    - обучить дерево принятия решений, визуализировать его (используя sklearn.tree.export_graphviz и pydot)
    - SVM(sklearn.svm.SVC)

6) Провести обучение ансамблевых классификаторов (Random Forest, AdaBoost, Gradient Boost)

Подобрать оптимальные параметры для каждой модели:

- Число ближайших соседей для kNN
- Для SVM рассмотреть линейное ядро и rbf, с помощью решетчатого поиска (sklearn.grid_search.GridSearchCV) подобрать оптимальные «C» и «gamma»
- для ансамблевых методов найти оптимальные значения параметров с помощью решетчатого поиска

Среди выбранных оптимальных моделей каждого класса выбрать наилучшую (считая основным критерием метрику `f1-score`)

Отобразить sklearn.metrics.classification_report **и** sklearn.metrics.confusion_matrix

В качестве отчета – рабочий код в Jupyter notebook

Датасеты:

https://www.kaggle.com/data

https://archive.ics.uci.edu/ml/index.