Техническая документация.

Модель анализа качества вина по его химическим свойствам.

Пререквизиты для запуска

- Компьютер(Windows 7+, Linux, Mac OS).
- Установленный Python 3.
- Установленные для Python библиотеки numpy, sckit, pandas, matplotlib последних стабильных версий.

Входные данные

На вход программе подаются 2 файла с данными: обучающий датасет (можно использовать приложенный к проекту, можно создать свой) и датасет для вина, качество которого мы желем определить. Обучающий и запускаемый датасеты должны отличаться только тем, что в запускаемом датасете должен отсутствовать признак качества вина, все остальные признаки должны совпадать и по названию, и по расположению.

Формат датасета.

Стандартный формат включает в себя 12 числовых полей: 11 признаков и 1 зависимую переменную(„quality”), представляющую из себя оценку от 1 до 10. Можно добавлять свои признаки: они должны быть числовыми и не пустыми. Признак зависимой переменной с именем „quality” обязательно должен присутствовать в обучающем датасете.

Стандарные признаки

- **Fixed acidity**: most acids involved with wine or fixed or nonvolatile (do not evaporate readily)
- **Volatile acidity**: the amount of acetic acid in wine, which at too high of levels can lead to an unpleasant, vinegar taste
- **Citric acid**: found in small quantities, citric acid can add 'freshness' and flavor to wines
- **Residual sugar**: the amount of sugar remaining after fermentation stops, it's rare to find wines with less than 1 gram/liter and wines with greater than 45 grams/liter are considered sweet
- **Chlorides**: the amount of salt in the wine
- **Free sulfur dioxide**: the free form of SO2 exists in equilibrium between molecular SO2 (as a dissolved gas) and bisulfite ion; it prevents microbial growth and the oxidation of wine

- **Total sulfur dioxide**: amount of free and bound forms of S02; in low concentrations, SO2 is mostly undetectable in wine, but at free SO2 concentrations over 50 ppm, SO2 becomes evident in the nose and taste of wine
- **Density**: the density of water is close to that of water depending on the percent alcohol and sugar content
- **pH**: describes how acidic or basic a wine is on a scale from 0 (very acidic) to 14 (very basic); most wines are between 3-4 on the pH scale
- **Sulphates**: a wine additive which can contribute to sulfur dioxide gas (S02) levels, wich acts as an antimicrobial and antioxidant
- **Alcohol**: the percent alcohol content of the wine

Запуск скрипта

Для запускающего скрипта (script.py) определены 4 параметры(по порядку следования):

1) Путь к запускаемому датасету (обязательный признак)
2) Путь к обучающему датасету (опциональный признак, по-умолчанию используется тренировочный датасет, поставляемый с моделью)
3) Бинарный признак необходимости перенастроить модель(опциональный признак, должен быть установлен в True, если вы добавляете данные в или используете свой обучающий датасет)
4) Путь к файлу с результами(опциональный признак, по умолчанию результат сохраняется в файл result.csv в директории скрипта)

Запускается скрипт командой python script.py path_to_launched_dataset [optional_args].

Выходные данные

Столбец из оценок вина, где в i-й строчке стоит оценка для вина из i-й строчки запускаемого датасета и confusion matrix для обучившийся модели (файл «confusion matrix for random forest classifier.png»).

На консоль выводятся confusion matrix для обучающей и тестовой подвыборок обучающего датасета и среднее арифметическое среднего взвешенного f1 score для получившейся модели.