**Московский государственный технический университет им. Н.Э.Баумана**

|  |  |  |
| --- | --- | --- |
|  |  |  |

**Отчет по лабораторной работе №6 по курсу**

**Технологии машинного обучения**

«**Ансамбли моделей машинного обучения**»

4

(количество листов)

|  |  |
| --- | --- |
| ИСПОЛНИТЕЛЬ: |  |
| студент группы **ИУ5-62** | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |
|  | (подпись) |
| **Писарчук Н.** | "25" мая 2019 г. |

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |

Москва, МГТУ - 2019

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Задание:**

1. Выберите набор данных (датасет) для решения задачи классификации или регресии.
2. В случае необходимости проведите удаление или заполнение пропусков и кодирование категориальных признаков.
3. С использованием метода train\_test\_split разделите выборку на обучающую и тестовую.
4. Обучите две ансамблевые модели. Оцените качество моделей с помощью одной из подходящих для задачи метрик. Сравните качество полученных моделей.
5. Произведите для каждой модели подбор значений одного гиперпараметра. В зависимости от используемой библиотеки можно применять функцию GridSearchCV, использовать перебор параметров в цикле, или использовать другие методы.
6. Повторите пункт 4 для найденных оптимальных значений гиперпараметров. Сравните качество полученных моделей с качеством моделей, полученных в пункте 4.

**Выполнение**:

Я выбрала датасет Red Wine Quality, содержащий оценки красных вин по 10-ти бальной шкале. Выполненное задание представлено в виде ноутбука Lab5 – red wine.ipynb. Пропуски данных отсутствуют.

Описание датасета:

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

**target:** quality - output variable (based on sensory data, score between 0 and 10)

Ансамблевые модели:

1. Случайный лес – Random Forest Classifier
2. Градиентный бустинг – Gradient Boosting Classifier

Метрики оценки модели:

1. Accuracy
2. Accuracy for classes
3. Precision
4. Recall

Сравнение accuracy до и после grid search (train, test)

Random forest

Default parameters

0.9865951742627346, 0.6604166666666667

Метка Accuracy

3 0.0

4 0.0

5 0.8164251207729468

6 0.6307692307692307

7 0.46153846153846156

8 0.3333333333333333

Grid search

0.8588025022341377 0.6645833333333333

Метка Accuracy

3 0.0

4 0.0

5 0.7922705314009661

6 0.7025641025641025

7 0.34615384615384615

8 0.0

Gradient boosting

Default parameters

0.8918677390527256 0.6541666666666667

Метка Accuracy

3 0.0

4 0.047619047619047616

5 0.7777777777777778

6 0.6461538461538462

7 0.4807692307692308

8 0.3333333333333333

Grid search

0.9526362823949955 0.68125

Метка Accuracy

3 0.0

4 0.047619047619047616

5 0.782608695652174

6 0.7128205128205128

7 0.46153846153846156

8 0.3333333333333333

Gradient boosting лучше распознает разные классы, чем Random forest