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

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

**Отчет по лабораторной работе №2 по курсу**

**Технологии машинного обучения**

«Изучение библиотек обработки данных»

2

(количество листов)

|  |  |
| --- | --- |
| ИСПОЛНИТЕЛЬ: |  |
| студент группы **ИУ5-62** | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |
|  | (подпись) |
| **Писарчук Н.** | "27" февраля 2019 г. |

Москва, МГТУ - 2019

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

**Задание:**

#### **Часть 1.**

Выполните первое демонстрационное задание "demo assignment" под названием "Exploratory data analysis with Pandas" со страницы курса <https://mlcourse.ai/assignments>

Условие задания - <https://nbviewer.jupyter.org/github/Yorko/mlcourse_open/blob/master/jupyter_english/assignments_demo/assignment01_pandas_uci_adult.ipynb?flush_cache=true>

Набор данных можно скачать здесь - <https://archive.ics.uci.edu/ml/datasets/Adult>

Пример решения задания - <https://www.kaggle.com/kashnitsky/a1-demo-pandas-and-uci-adult-dataset-solution>

#### **Часть 2.**

Выполните следующие запросы с использованием двух различных библиотек - [Pandas](https://pandas.pydata.org/) и [PandaSQL](https://github.com/yhat/pandasql):

* один произвольный запрос на соединение двух наборов данных
* один произвольный запрос на группировку набора данных с использованием функций агрегирования

Сравните время выполнения каждого запроса в Pandas и PandaSQL.

* Сформировать отчет и разместить его в своем репозитории на github.

**Выполнение**:

Первая часть задания выполнена в ноутбуке assignment01\_pandas\_uci\_adult1.ipynb.

Вторая часть задания выполнена в ноутбуке pandas\_vs\_pandasql.ipynb. Для исследования я выбрала датасет об использовании телефонов, состоящий из двух файлов user\_usage.csv и user\_device.csv (признаки: название устройства, платформа и ее версия, кол-во исходящих звонков и смс…)

Вывод: время обработки запроса pandasql в разы больше pandas.