

# Отчет по домашней работе от 03.11.20

Михаил Михайлов, Ельцов Даниил

10 ноября 2020 г.

## Содержание

|          |  |          |
|----------|--|----------|
| <b>1</b> | <b>Постановка задачи</b>                 | <b>2</b> |
| <b>2</b> | <b>Используемый датасет</b>              | <b>2</b> |
| <b>3</b> | <b>Описание решения</b>                  | <b>2</b> |
| 3.1      | Разметка предметов по профилям . . . . . | 2        |
| 3.2      | Обработка данных . . . . .               | 2        |
| 3.3      | Построение классификатора . . . . .      | 2        |
| <b>4</b> | <b>Результаты</b>                        | <b>3</b> |

## Резюме

Был построен классификатор, который по количеству дипломов от школы по каждому предмету определяет ее профиль. Источник данных — датасет победителей и призеров олимпиад московских школ (**источник**).

На вход программы подается название школы, после чего строится вектор данных, содержащий информацию о количестве дипломов (диплом призера и победителя считаются равноценными), который прогоняется через классификатор.

## Постановка задачи

Построить классификатор который определит к какому из профилей принадлежит школа:

- физико-математический,
- естественно-научный,
- обществоведческий,
- языковой,
- гуманитарный,
- многопрофильный/без профиля/не определено.

## Используемый датасет

Датасет взят с [kaggle.com](https://www.kaggle.com) — [источник](#).

Состоит из одной csv таблицы, содержащей около 70 тысяч строк, формата:

Полное название школы, Краткое название школы, Тип олимпиады (ММО/Всеросс), Этап, Класс, Предмет, Год, ID

## Описание решения

### Разметка предметов по профилям

- *Категория физико-математических профилей*: астрономия, физика, информатика, математика, информатика и информационно-коммуникационные технологии (ИКТ).
- *Категория естественно-научных профилей*: биология, география, химия, экология.
- *Категория обществоведческих профилей*: изобразительное искусство, история, мировая художественная культура (МХК), обществознание, право, экономика, искусство (МХК), бюджетная грамотность.
- *Категория языковых профилей*: французский язык, итальянский язык, китайский язык, английский язык, испанский язык, немецкий язык, латынь.
- *Категория гуманитарных профилей*: филология, русский язык, литература, лингвистика.
- *Категория других предметов*: технология, основы безопасности жизнедеятельности, физическая культура, робототехника, информационные технологии в профессиональной деятельности.

## Обработка данных

Строится таблица:

Полное имя школы -- вектор: (Количество дипломов категории 1, ..., Количество дипломов категории 6)

## Построение классификатора

Для классификатора зафиксированы константы:

- **THRESHOLD** — порог числа дипломов, необходимого для того чтобы можно было определить профиль.
- **MAIN\_SUBJECT\_THRESHOLD** — минимальный процент дипломов, необходимый для выделения главного профиля.
- **MIDDLE\_SUBJECT\_THRESHOLD** — минимальный процент дипломов, необходимый для установки многопрофильности школы.
- **EPS** — порог разницы процентов между дипломами, необходимый для установки многопрофильности школы.

Если значение какой-то координаты в пересчете на проценты превосходит **MAIN\_SUBJECT\_THRESHOLD**, то профиль школы соответствует профилю этой координаты.

Если значение каких-то двух координат в пересчете на проценты превосходит **MIDDLE\_SUBJECT\_THRESHOLD**, то школа имеет несколько профилей.

## Результаты

Обработка данных — Михаил Михайлов

Построение классификатора — Даниил Ельцов

Отчет — Михаил Михайлов

Код