ПРАВИТЕЛЬСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ
ФГАОУ ВО НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

Факультет компьютерных наук

Образовательная программа «Прикладная математика и информатика»

УДК _____

Отчет об исследовательском проекте

на тему «Топологический анализ социальных данных»

(промежуточный, этап 2)

Выполнил:

студент группы БПМИ 198 _____ С.В. Пилипенко
Подпись И.О. Фамилия

29.04.2021
Дата

Принял:

руководитель проекта Айзенберг Антон Андреевич
Имя, Отчество, Фамилия

кандидат физико-математических наук, доцент
Должность, ученое звание

международная лаборатория алгебраической топологии и ее приложений
Место работы (Компания или подразделение НИУ ВШЭ)

Дата проверки _____2021 _____ _____
Оценка (по 10-тибалльной шкале) Подпись

Москва 2021 г.

Содержание

1 Основые термины и определения

Симплициальный комплекс *Симплициальным комплексом* на конечном множестве вершин M называется совокупность $K \subset 2^M$ подмножеств множества M, удовлетворяющая следующим двум условиям:

1. если $I \in K$ и $J \subset I$, то $J \in K$;
2. $\varnothing \in K$.

Симплекс *Симплексом* называются элементы симплециального комплекса K.

Порядковая переменная Переменная, которая принимает значения из конечного упорядоченного множества.

Номинальная переменная Переменная, которая принимает значения из конечного неупорядоченного множества.

2 Введение

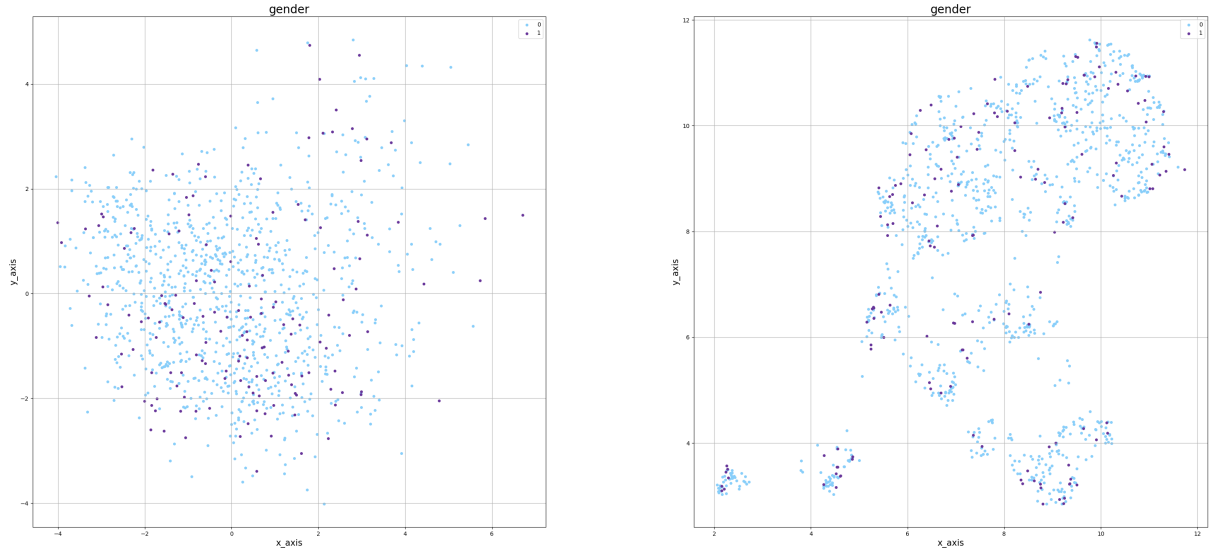
Следующим этапом моей работы над проектом были непосредственные манипуляции с данными с целью выявить в них какие-то закономерности. Для этого, пока что, были рассмотрены алгоритмы UMAP [2] и PCA [1]. Для достижения этой цели, нужно было написать код, который берет обработанный массив данных, представленный матрицей $A \in \mathbb{R}^{m \times n}$ со стандартизированными столбцами (из каждого вычли его среднее, и поделили на дисперсию). Чтобы упростить читаемость отчета и работу с исходными данными, я закодировал все «содержательные» вопросы — такие вопросы, по которым очень сложно делать хорошие предсказания (например, ввиду их большого количества), или те вопросы, по которым мы не будем делать предсказания и красить точки в пространстве. Также я выделил несколько гипотез относительно данных, которые хотел проверить:

1. Распределение ответов на опрос не зависит от гендера участников.
2. Распределение ответов учеников на опрос не зависит от класса, в котором ученик обучается.
3. Распределение ответов членов школьной администрации зависит от должности.

3 Проверка гипотез

Для того, чтобы проверить выдвинутые гипотезы, я построил и проанализировал 12 графиков: на каждом наборе данных (результаты опросов школьников и администрации) я применил алгоритмы UMAP и PCA, а потом сравнил их результаты. Интересно, что распределения у студентов, полученные с помощью PCA и UMAP, практически не отличаются друг от друга, в то время как распределение ответов администрации может меняться в зависимости от того, какую целевую переменную мы предсказываем. Этот эффект объясняется тем, что я использовал две различные техники кластеризации в процессе работы с данными. В случае администрации, я выделял целевую переменную, оставшиеся кодировал, нормализировал, а потом понижал размерность. В случае же ответов учеников, я заранее определил набор целевых переменных, по которым буду различать результаты эксперимента, и понижал размерность оставшихся данных.

3.1 Распределение ответов на опрос не зависит от гендера участников

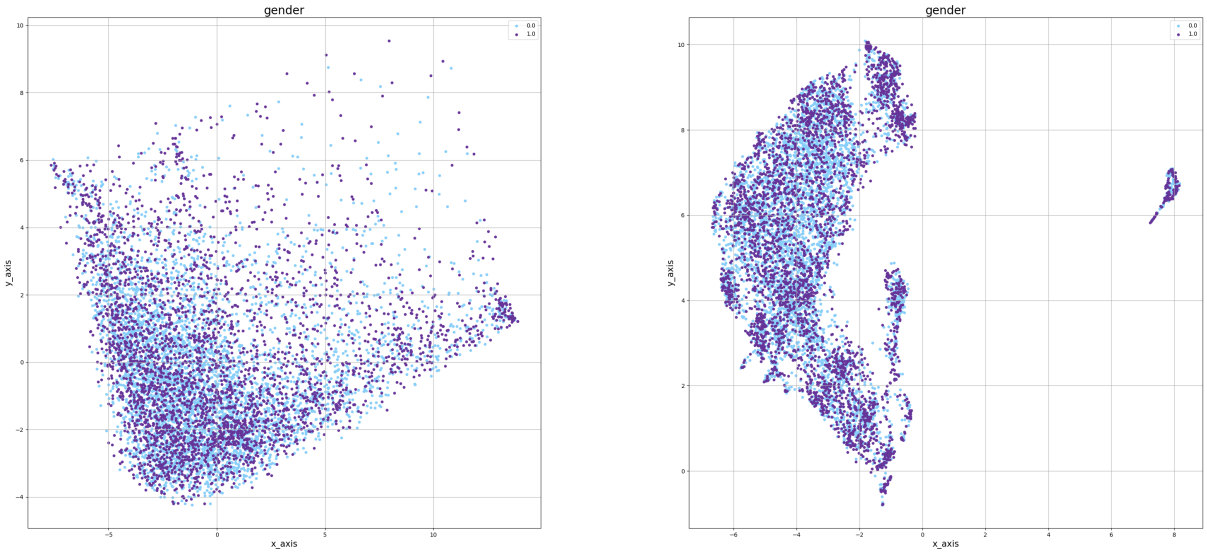


(a) PCA (b) UMAP

Рис. 1: Распределение ответов администрации в зависимости от пола

На построенных распределениях для администрации (рис. 1a и 1b) мы можем видеть, что вне зависимости от гендера респондентов, распределяются по кластерам они примерно

одинаково, с той лишь разницей, что женского персонала в составе администрации школ сильно больше, чем мужского. Аналогичное распределение мы можем наблюдать и на ответах школьников (рис. 2a и 2b), только в этом случае у нас примерно одинаковое количество респондентов разного пола. Таким образом, гипотеза подтверждается, ответы участников действительно не зависят от пола.

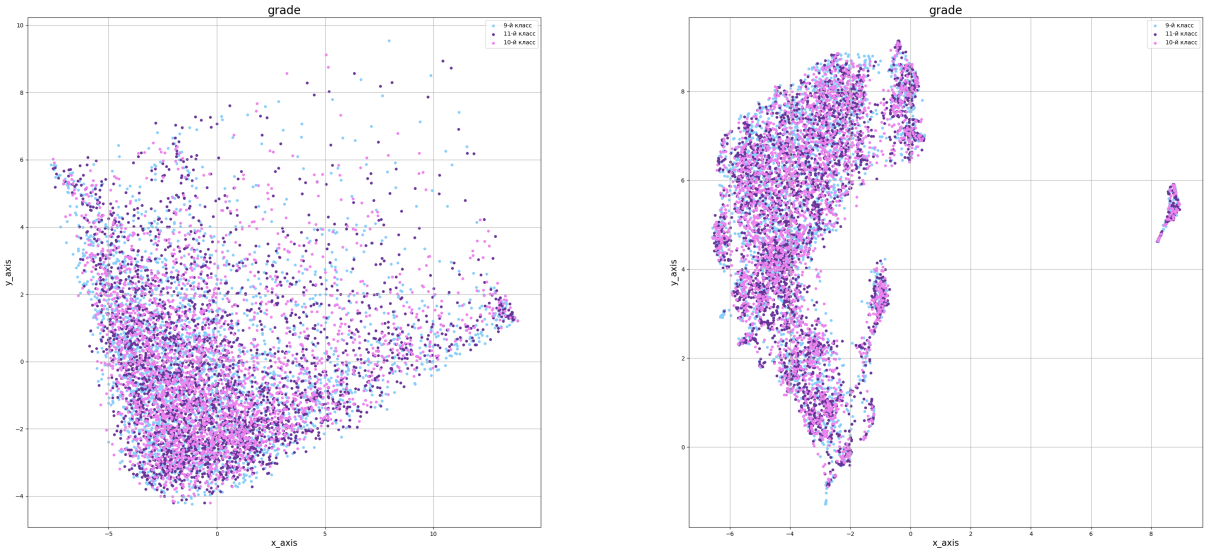


(a) PCA

(b) UMAP

Рис. 2: Распределение ответов школьников в зависимости от пола

3.2 Распределение ответов учеников на опрос не зависит от класса, в котором ученик обучается



(a) PCA

(b) UMAP

Рис. 3: Распределение ответов студентов в зависимости от класса

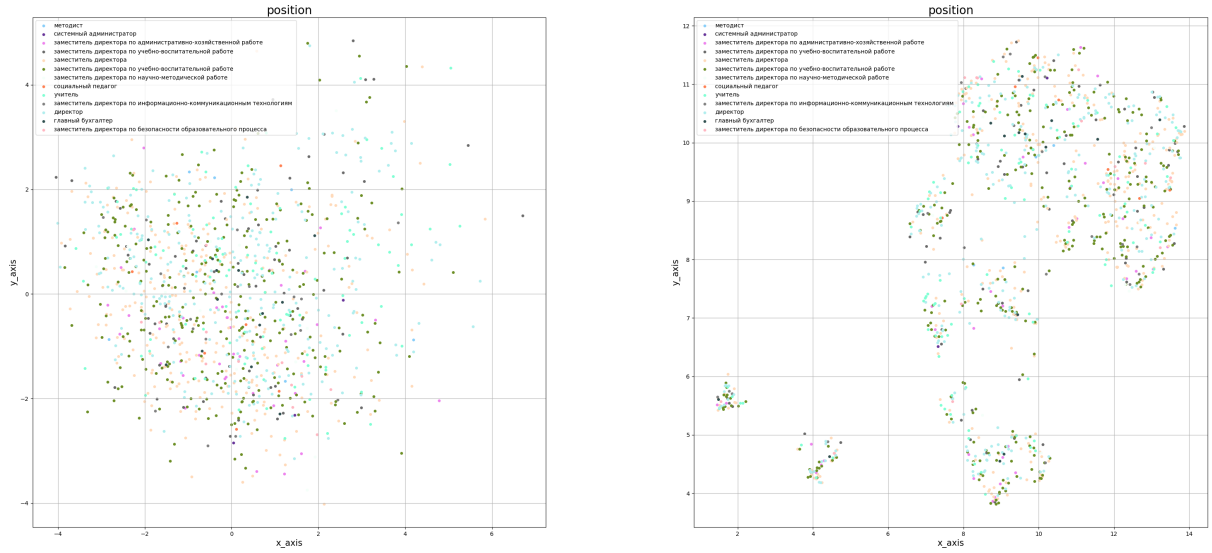
На рис. 3a и 3b мы можем заметить две вещи:

1. Распределение абсолютно не отличается от того, которое было в секции 3.1, что не удивительно, учитывая тот факт, что я просто подменил целевую переменную, по которой раскрашиваю точки в пространстве.

2. Распределение по классам весьма равномерное в двух кластерах в случае рис. 3b.

Таким образом, класс учеников не влияет на их ответы в опросе про техническую оснащенность школы. Это, на самом деле, весьма логично, ведь если школа использует IT технологии в образовательном процессе отдельной параллели, то она будет эти же технологии распространять и на все остальные параллели.

3.3 Распределение ответов членов школьной администрации зависит от должности



(a) PCA (b) UMAP

Рис. 4: Распределение ответов администрации школы в зависимости от должности

Как мы можем наблюдать, моя гипотеза не подтвердилась, поскольку почти все должности присутствуют в каждом из полученных кластеров на рис. 4a и 4b. Это говорит, как я предполагаю, о широкой осведомленности почти каждого сотрудника школьной администрации о технологической оснащенности школы.

4 План дальнейшей работы

Следующей задачей является задача классификации респондентов в зависимости от их ответов. Я попробую предсказывать школу человека по его ответам в опросе. Также, в процессе предварительной обработки данных, я выбросил некоторые неструктурированные данные. В частности, я проигнорировал все ответы на вопрос про использование конкретных электронных образовательных систем, хотя мне кажется, что в этих данных могут быть какие-то интересные закономерности. Эти данные сложно поддаются обработке, однако я все равно попробую их почистить и по ним покластеризовать ответы участников опроса.

Также, планируется перебрать абсолютно все категориальные признаки, чтобы выявить, какой именно среди них влияет на образование кластеров среди ответов сотрудников школьной администрации (рис. 1b, 4b).

Список литературы

1. *Jolliffe I.* Principal Component Analysis // International Encyclopedia of Statistical Science / под ред. M. Lovric. — Berlin, Heidelberg : Springer Berlin Heidelberg, 2011. — С. 1094—1096. — ISBN 978-3-642-04898-2. — DOI: 10.1007/978-3-642-04898-2_455. — URL: https://doi.org/10.1007/978-3-642-04898-2_455.
2. *McInnes L.*, *Healy J.*, *Melville J.* UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. — 2018. — URL: https://arxiv.org/pdf/1802.03426.pdf.

5 Приложение

Таблица 1: Календарный план работ по проекту

Дата	Название задачи
12.03.2021	Ознакомление с методичкой по симплициальным комплексам и гомологиям
22.03.2021	Изучение работы классических алгоритмов анализа данных на соц. опросах
30.03.2021	Изучение работы топологических алгоритмов анализа данных на соц. опросах
14.04.2021	Сравнительный анализ результатов и их визуализация
26.04.2021	Отчет по КТ2
01.05.2021	Изучение алгоритмов машинного обучения с учителем
07.05.2021	Разработка модели (нейронной сети), которая будет предсказывать по ответам ученика его школу
15.05.2021	Итоговый отчет по проекту