**Описание проекта «Разведывательный анализ данных (EDA)»**

**→ основные цели и задачи проекта:**

**Цель**: отследить влияние условий жизни учащихся в возрасте от 15 до 22 лет на их успеваемость по математике, чтобы на ранней стадии выявлять студентов, находящихся в группе риска

**Задачи**:

1. Первичная обработка данных
2. Устранить выбросы в числовых переменных
3. Оценить количество уникальных значений для номинативных переменных
4. Преобразовать данные при необходимости
5. Провести корреляционный анализ количественных переменных
6. Отобрать не коррелирующие переменные
7. Проанализировать номинативные переменные и устранить те, которые не влияют на предсказываемую величину
8. Сформулировать выводы относительно качества даннызи тех переменных, которые будут использоваться в дальнейшем

**→ краткая информация о данных:**

В датасете содержатся данные об успеваемости учащихся из 2х школ, а также характеристики этих учащихся (30 колонок). Всего 395 записей.

**→ этапы работы над проектом:**

1. Посмотреть на данные.
2. Проверить данные на пустые значения.
3. Проверить данные на дублированные/полностью скоррелированные значения.
4. Проверить данные на наличие выбросов.
5. Отобрать данные, пригодные для дальнейшего построения модели.
6. Снова и снова возвращаться к предыдущим пунктам, пока модель не заработает как надо.

**→ ответы на вопросы саморефлексии:**

1. Какова была ваша роль в команде?

К сожалению командной работы, как таковой, не получилось, поэтому был исполнителем и руководителем проекта ☺ Не совсем понятно в чем тут могло быть разделение на роли, ведь каждый должен был сделать всю часть работы, чтобы усвоить пройденный материал.

1. Какой частью своей работы вы остались особенно довольны?

Особенно доволен тем, что удается находить нужную информацию для выполнения проекта, так как теории иногда не хватает, и приходится доучиваться по ходу выполнения проекта.

1. Что не получилось сделать так, как хотелось? Над чем ещё стоит поработать?

Еще стоит поработать над теоретическим материалам, азы матстатистика необходимо хорошо знать и понимать, еще есть пробелы над котрыми предстоит поработать.

1. Что интересного и полезного вы узнали в этом модуле?

Хоть и был знаком с основами матстатистики, но они уже давно забылись, и все равно слушая видеолекции постоянно открывал для себя то-то новое или понимал, то что раньше было не понятно.

1. Что является вашим главным результатом при прохождении этого проекта?

Понимание основ работы с данным, считаю, что для себя я сделал большой шаг вперед. Стало более осмысленным, чем раньше.

1. Какие навыки вы уже можете применить в текущей деятельности?

- анализ данных на пустые значения, выбросы, дублирование

- построение графиков, таблиц корреляции, анализ номинативных переменных

7. Планируете ли вы дополнительно изучать материалы по теме проекта?

Да, обязательно. Как минимум скачал несколько книг по матстатистики, считаю, что на этом этапе это самое важное. Как написать код – это полбеды, сейчас много информации на эту тему и готовых кейсов, а вот понимать *какой именно код* писать – это основная задача.