Оглавление

[Цели проекта 2](#_Toc41482985)

[Информация о данных 2](#_Toc41482986)

[Этапы работы над проектом 3](#_Toc41482987)

[Первичный анализ данных 3](#_Toc41482988)

[Анализ бинарных признаков 3](#_Toc41482989)

[Анализ номинативных категориальных признаков 3](#_Toc41482990)

[Анализ вещественных (числовых) признаков 3](#_Toc41482991)

[Анализ упорядоченных категориальных признаков 3](#_Toc41482992)

[Анализ столбца studytime, granular 4](#_Toc41482993)

[Анализ корреляции признаков 4](#_Toc41482994)

[Тест Стьюдента на значимость признаков 4](#_Toc41482995)

[Формирование выводов 4](#_Toc41482996)

[Дополнительно: построение модели линейной регрессии 4](#_Toc41482997)

[Ответы на вопросы саморефлексии 4](#_Toc41482998)

[Какова была ваша роль в команде? 4](#_Toc41482999)

[Какой частью своей работы вы остались особенно довольны? 4](#_Toc41483000)

[Что не получилось сделать так, как хотелось? Над чем ещё стоит поработать? 4](#_Toc41483001)

[Что интересного и полезного вы узнали в этом модуле? 4](#_Toc41483002)

[Что является вашим главным результатом при прохождении этого проекта? 5](#_Toc41483003)

[Какие навыки вы уже можете применить в текущей деятельности? 5](#_Toc41483004)

[Планируете ли вы дополнительно изучать материалы по теме проекта? 5](#_Toc41483005)

# Цели проекта

Предлагается поучаствовать в одном из проектов UNICEF — международного подразделения ООН, чья миссия состоит в повышении уровня благополучия детей по всему миру.

**Цель проекта** — 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 проекта**:

* Провести анализ данных
* Выявить наиболее значимые признаки для построения модели
* Дополнительно: построить модель для предсказания результатов госэкзамена по математике для каждого ученика школы
* Дополнительно: провести оценку полученной модели

# Информация о данных

В датасете представлены следующие данные:

* school — аббревиатура школы, в которой учится ученик
* sex — пол ученика ('F' - женский, 'M' - мужской)
* age — возраст ученика (от 15 до 22)
* address — тип адреса ученика ('U' - городской, 'R' - за городом)
* famsize — размер семьи('LE3' <= 3, 'GT3' >3)
* Pstatus — статус совместного жилья родителей ('T' - живут вместе 'A' - раздельно)
* Medu — образование матери (0 - нет, 1 - 4 класса, 2 - 5-9 классы, 3 - среднее специальное или 11 классов, 4 - высшее)
* Fedu — образование отца (0 - нет, 1 - 4 класса, 2 - 5-9 классы, 3 - среднее специальное или 11 классов, 4 - высшее)
* Mjob — работа матери ('teacher' - учитель, 'health' - сфера здравоохранения, 'services' - гос служба, 'at\_home' - не работает, 'other' - другое)
* Fjob — работа отца ('teacher' - учитель, 'health' - сфера здравоохранения, 'services' - гос служба, 'at\_home' - не работает, 'other' - другое)
* reason — причина выбора школы ('home' - близость к дому, 'reputation' - репутация школы, 'course' - образовательная программа, 'other' - другое)
* guardian — опекун ('mother' - мать, 'father' - отец, 'other' - другое)
* traveltime — время в пути до школы (1 - <15 мин., 2 - 15-30 мин., 3 - 30-60 мин., 4 - >60 мин.)
* studytime — время на учёбу помимо школы в неделю (1 - <2 часов, 2 - 2-5 часов, 3 - 5-10 часов, 4 - >10 часов)
* failures — количество внеучебных неудач (n, если 1<=n<3, иначе 0)
* schoolsup — дополнительная образовательная поддержка (yes или no)
* famsup — семейная образовательная поддержка (yes или no)
* paid — дополнительные платные занятия по математике (yes или no)
* activities — дополнительные внеучебные занятия (yes или no)
* nursery — посещал детский сад (yes или no)
* higher — хочет получить высшее образование (yes или no)
* internet — наличие интернета дома (yes или no)
* romantic — в романтических отношениях (yes или no)
* famrel — семейные отношения (от 1 - очень плохо до 5 - очень хорошо)
* freetime — свободное время после школы (от 1 - очень мало до 5 - очень мого)
* goout — проведение времени с друзьями (от 1 - очень мало до 5 - очень много)
* health — текущее состояние здоровья (от 1 - очень плохо до 5 - очень хорошо)
* absences — количество пропущенных занятий
* score — баллы по госэкзамену по математике
* studytime, granular - загадочный 30й столбец, из соображений логики наверняка как-то связанный со studytime

У нас есть 30 колонок данных, score - целевая переменная. Среди признаков, судя по описанию датасета, у нас есть следующие виды признаков:

* бинарные **(school, sex, address, famsize, Pstatus, schoolsup, famsup, paid, activities, nursery, higher, internet, romantic)**
* категориальные (номинативные): **Mjob,Fjob,reason,guardian**
* категориальные (порядковые): **Medu, Fedu, studytime, failures, traveltime, famrel, freetime, goout, health**
* вещественные (числовые): **age, absences**

Также есть признак **studytime,granular**, смысл которого нужно было понять в процессе анализа.

# Этапы работы над проектом

## Первичный анализ данных

На данном этапе было установлено количество пропусков в каждом признаке, первично выбраны методы заполнения пропусков, а также удалены пропуски в целевой переменной score.

## Анализ бинарных признаков

На данном этапе номинативные признаки, в которых было всего 2 значения, были трансформированы в бинарные. Для каждого признака был построена гистограмма распределения значений, заполнены пропуски медианой.

## Анализ номинативных категориальных признаков

Для номинативных признаков также были устранены пропуски с помощью моды, построены боксплоты с распределением данных в каждом признаке, сделан первичный вывод о возможной значимости признаков.

## Анализ вещественных (числовых) признаков

Для числовых признаков были построены гистограммы распределений, данные были протестированы на выбросы по формуле интерквартильного расстояния. Пропуски были заполнены медианой. Выбросы отсеивались исходя из комбинации формулы и здравого смысла.

## Анализ упорядоченных категориальных признаков

На данном этапе были построены гистограммы распределений значений признаков, выявлены и устранены выбросы, связанные, вероятно, с ошибкой в данных при заполнении. Пропуски были заполнены медианой по признаку.

## Анализ столбца studytime, granular

Для данного признака также было построено распределение его значений, выявлено количество каждого из уникальных значений. Был выявлена полная корреляция со столбцом studytime. Вероятно, значения столбца studytime,granular были получены путем умножения значений из столбца studytime на -3. Было принято решение удалить данный признак из датасета.

## Анализ корреляции признаков

Была сформирована матрица корреляции признаков и выявлены признаки с корреляцией более 0.5. Из данных признаков был сформирован новый, использующий их данные, а сами признаки были удалены из датасета.

## Тест Стьюдента на значимость признаков

Все признаки были протестированы с помощью теста Стьюдента, выявляющего статистически значимые различия в признаках. Таким образом, важные признаки были отобраны для дальнейшего построения модели регрессии.

## Формирование выводов

На основе проведенного анализа данных были сформированы выводы по количеству пропусков, выбросов, чистоте данных, корреляции и важности признаков. Были даны рекомендации по наиболее важным признакам, которые предлагается далее использовать в модели.

## Дополнительно: построение модели линейной регрессии

На данном этапе на основе обработанного датасета была построена модель линейной регрессии с нормализацией данных с помощью RobustScaler, и была проведена оценка модели с помощью метрик MSE и RMSE.

# Ответы на вопросы саморефлексии

## Какова была ваша роль в команде?

В данном проекте я участвовала в роли аналитика данных с последующей разработкой рекомендации по использованию признаков для построения модели.

## Какой частью своей работы вы остались особенно довольны?

На мой взгляд, в итоге у меня получилось обобщить и систематизировать информацию по видам признаков, так как в первой версии моего решения каждый столбец был проанализирован отдельно, не объединяясь в подгруппы признаков.

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

В качестве пробной проверки правильности выделения признаков, я построила модель регрессии, которая показала не слишком хорошие, на мой взгляд, результаты на 100-балльной шкале оценок. Стоит еще поработать над добавлением\удалением признаков с последующим анализом качества модели, а также возможно стоит добавить нелинейность в виде комбинации каких-то признаков или их степеней.

## Что интересного и полезного вы узнали в этом модуле?

Этот модуль помог систематизировать знания о том, как работать с датасетом, на что следует обращать внимание. Окончательно закрепилось понимание, к каким категориям относить признаки, каким шагам следовать, чтобы очистить и систематизировать данные.

## Что является вашим главным результатом при прохождении этого проекта?

Главным результатом является первая попытка анализа EDA, систематизация информации и знаний и видах признаков, более глубокое понимание этапов анализа данных и инструментов для анализа.

## Какие навыки вы уже можете применить в текущей деятельности?

Появились навыки анализа данных с построением распределений данных, устранением выбросов, наблюдением корреляции признаков. Навыки, конечно, нужно еще расширять и углублять, но более прочное понимание базы как раз к этому и будет подталкивать.

## Планируете ли вы дополнительно изучать материалы по теме проекта?

Да, планирую, так как анализ данных – это одна из важнейших составляющих в процессе подготовки к построению моделей.