**ЧАСТИ РАБОТЫ**

[**АВТОРСКИЙ ВКЛАД ВЫБОРА ФИЧ** 1](#_Toc163560644)

[ИСПОЛЬЗУЕМЫЕ ТЕРМИНЫ 1](#_Toc163560645)

[МОТИВАЦИЯ АВТОРА 1](#_Toc163560646)

[ВАРИАНТЫ ИСПОЛЬЗОВАНИЯ ПРОГРАММНОГО МОДУЛЯ 1](#_Toc163560647)

[ВОЗМОЖНОСТЬ МОДИФИКАЦИИ 1](#_Toc163560648)

[СБОР ДАННЫХ 1](#_Toc163560649)

[**ВЫБОР ФИЧ** 1](#_Toc163560650)

[2.2 СБОР ДАННЫХ 1](#_Toc163560651)

[2.3 MACHINE LEARNING 1](#_Toc163560652)

[СБОР ДАННЫХ 1](#_Toc163560653)

### **АВТОРСКИЙ ВКЛАД ВЫБОРА ФИЧ**

Наибольшего успеха в поиске проблем неудовлетворительной успеваемости студентов могут добиться лица, наиболее приближенные к самим студентам в нашем случае автор - студент. Так что автор попробует внести свою лепту в данную работу.

объединение разных групп форм обучения и тд и тп является плохим признаком как так данные имеют крайне разные характеры и хаарактеристик.

Автором был проведён экспертный анализ по выбору фич, так автор напрямую узнал о мотивации студентов. Были выявлены следующие фичи:

- \*скажите о вашей мотивации не о плюшках а именно о мотивации самой учёбы\*

Также включу эмпирические данные, что не есть хорошо, но исключать данные факторы - упущение большой части важной информации:

\* отсрочка для представителей мужского пола, так как исследуемые студенты из РФ регионов, то нужно учитывать особенности территориального аспекта

## ИСПОЛЬЗУЕМЫЕ ТЕРМИНЫ

Для грамотного рассмотрения дальнейшего текста, а также для понимания хода повествования, существует раздел “Глоссарий” [1], в котором представлены определения основных понятий используемых в ВКР, чтобы читатель полноценно понимал о чём идёт речь.

Отметим, что часть определений трактуется неоднозначно, указана лишь одна из возможных и подходящих для данной работы трактовка.

Если какое-то определение не дано, значит автор считает, что большинство читателей будет иметь представление о термине.

## МОТИВАЦИЯ АВТОРА

Выбор темы данной ВКР аргументирован. Автор, на момент написании труда, два с половиной года изучал машинное обучение и предметные области науки Data Science. Первый имеет опыт построения моделей на известной платформе Kaggle. Что касается математического аппарата, то его исследователь развивает по сей день при помощи академической учёбы (школа, университет) и изучения материалов в свободное время в интернете. По поводу владения современными инструментами проблем не имеется, автор обучается на IT-направлении в высшем учебном заведении, что подразумевает под собой программирование и изучение работы электронно-вычислительных устройств, также создатель ВКР дополнительно проходит образовательную программу в школе программирования **XXX** (соответствующие проекты в приложении). Создатель труда решил не пренебрегать своими интересами в данной ВКР, поэтому выбрал тему для данной работы наиболее приближенную ему.

Данный труд будет использоваться создателем в дальнейшем не только в рамках высшего образования, но и в роли пет-проекта в профессиональной деятельности. Так что качество работы обеспечивается как личностными интересами, так академические.

## ВАРИАНТЫ ИСПОЛЬЗОВАНИЯ ПРОГРАММНОГО МОДУЛЯ

Данный программный модуль можно будет использовать как открытый веб-сервис для пользователей, которые хотят узнать информацию о себе как о студентах и об своей успеваемости. Для этого пользователю нужно будет ввести необходимую информацию и проследовать инструкции. Так в модульном журнале есть оценка для прогноза балла абитуриента, так,возможно, моё решение будет лучшей альтернативой для оценки успевемости студента. А также система возможно позволит найти тех индивидов, которые не хотят учиться по тем или иныч причинам, в следствии чего, не стоит будет с ними дольше возиться, а также государству продолжать спонсировать их пребывание.

## ВОЗМОЖНОСТЬ МОДИФИКАЦИИ

Если сервис будет востребован у пользователей, то будет осуществлена его модернизация и дальнейшая модификация для использования его как полноценного продукта на рынке. Проект имеет практически безграничный потенциал к развитию, так как современные тенденции связанные с машинным обучением развиваются и будут развиваться. Что касается масштабируемости, то имеются прилагающиеся возможности. Например, реализовать портативную версию для мобильных устройств.

## СБОР ДАННЫХ

Автор понимает, что при анкетировании опрашиваемый может находится в том или ином эмоциональном состоянии на показывающие объективность показателей исследуемых признаков, но пытаться получить объективные данные касаемо настроения и мотивации студентов, крайне тяжёлая работа (это проведелние аналитических опросов прибегая к определённой аппаратуре), что не является возможным на данный момент.

От качества опроса будет зависить качество данных, фактически своими формулировками и вопросами я манипулирую областью ответов студентов.

Графическое представление в n мерном пространстве значений каких-то признаков. По графикам можн понять понять о внешним виде распределения признака, Выборку разделяют на обучающую, контрольную, тестовую.

Также можно нарисовать распределние получившихся результатов, дабы проанализировать общие данные по поводу людей.

### **ВЫБОР ФИЧ**

Сразу скажем, что реальные сильно влияющие факторы не могут быть определены сразу, наша задача определить их гарницы. То есть выбрать те фичи, которые с большой вероятностью будут иметь связь с успеваемость. Проверить же это мы сможем только по ходу аналитики данных и фич.

Очевидно, что чем больше факторов мы будем учитывать, тем точнее будут модели, однако избыточные данные нам не нужны, так как они нагрузят наши вычисления, а также могут сильно повлиять не резульаты исследования и онулировать наши старания. Так что наша задача отобрать достаточное количество и необходимое количество признаков.

Основные группы факторов опроса студентов:

* потенциал, уровень знаний, достигнутый к поступлению в вуз;
* мотивация студента к получению высшего образования и причины поступления;
* характер и мотивация студента;
* наличие занятности во внеучебное время;
* влияние среды, родителей;
* преподаватели;
* организация учебного процесса: расписание, работа деканата, библиотеки.

*Таблица 4. Описание переменных полученной анкеты*

|  |  |  |  |
| --- | --- | --- | --- |
| Название признака | Тип данных | Смысл | Пример |
| Ф. И. О. | String | Нужен для идентификации студентов | Иванов Иван Иванович |
| Возраст | Integer | Используется в исследовании, особенно важен при больших отклонениях от моды. | 25 лет |
| Время в часах проводимое за университетской учёбой в неделю | Double | Используется в исследовании, показывать степень затраченных усилий студента | 5.25 часов |
| Время в часах проводимое за развлечениями в неделю | Double | Используется в исследовании, показывает время отдыха студента. | 7.5 часов |
| Время в часах проводимое за хобби в неделю | Double | Используется в исследовании, показывает время для удовлетворения интересов студента. | 1.5 часа |
| Количество секций и кружков, студий | Integer | Показывать активность студента | 5 |
| Укажите количество членов в вашей семье | Integer | Нужен для исследования | 4 |

Пути решения проблем сбора данных:

**1) Параллельные данные**. Нередко оказывается целесообразно разработать две достаточно равноправные методики для измерения заданной социальной характеристики что позволяет установить обоснованность. Когда имеется несколько методови один испольнитель то это так назывемая случай совокупного измерения. Основная задача здесь – установить зависимость двух комплексных оценок с помощью методов регрессионого анализа;

В процессе отработки инструментов измерения в отношении надёжности целесообразна следующая последовательность основных этапов работы:

* предварительный контроль обоснованности методов измерения на стадии проб методики на небольших выборках;
* пилотаж методики и тщательная проверка устойчивости исходных данных и вторичных показателей
* усовершенствование методики и доработка всех её деталей и в итоге получение окончательного варианта методики для исследования

2.2 Моделирование

2.2.1 IDEF

2.2.2 UML

# 2.2 СБОР ДАННЫХ

В связи с развитием интернета и информационных технологий имеется возможность онлайн анкетирования, которое позволит провести опрос в разных вузах и регионах, что в свою очередь увеличит объём выборки, а также повысит её репрезентативность. Стоит и не забывать, что человек сможет пройти опрос в любое удобное ему время, что исключает эффект стресса по сравнению с очным анкетированием. Однако стоит быть готовым к тому, что качество анкетных данных может упасть в связи с возможной аннонимностью и “неограниченным” количеством попыток прохождения анкетирования. Если социологические данные, то нужно понимать, построить универсальную модель невозможно, так как имеют разные распределения. Лучше строить две модели, или быть уверенным, что у них общее распределение.

2.2.1 РЕАЛИЗАЦИЯ СПОСОБА СБОРА ДАННЫХ

2.2.2 СБОР ДАННЫХ

2.2.3 АНАЛИЗ ДАННЫХ

2.2.4 ВЫВОДЫ ПО СОБРАННЫМ ДАННЫМ

# 2.3 MACHINE LEARNING

Улучшение модели стоит лишь до того момента, пока есть приемленные изменения, добавлять всё больше и больше слоёв нет смысла, так как вычисления будут больше и дольше. В зависимости от количества имеющихся данных надо аккуратно выбирать что модель что метрики что функции потерь.

**2.3.1 КЛАССИФИКАЦИЯ**

**2.3.2 КЛАСТЕРИЗАЦИЯ**

**2.3.3 РЕГРЕССИЯ**

**2.3.4 КЛАССИФИКАЦИЯ**

## СБОР ДАННЫХ

Нужно отметить, что для объективности испытания нам необходим математический аппарат, так как в ином случае мы не сможем доказать правоту наших утверждений.

При работе с исследованием важен все этапы, так что помимо самих алгоритмов стоит уделить большое внимание сбору информации.

Изучаемы признаки бывают: внутренние и внешние.

**Внутреннии признаки** – определяются личностью студента.

**Внешнии признаки** – определеяются установленной средой, в которой студент живёт.

Так как студенты из разных групп, разных направлений и разных курсов, мы постараемся выделить те признаки, которые универсальны для всех, дабы не строить индивидуальную модель для каждого.

Разработка электронной анкеты. Сбор данных осуществляется в ходе письменного опроса по анкете, включающей N вопросов. Структура анкеты отражает факторы, обозначенные в теоретических вопросах исследования.

Для студентов разных ВУЗ-ов мы построим следующую таблицу:

*Таблица 3. Пример таблицу студентов по ВУЗ-у*

|  |  |  |  |
| --- | --- | --- | --- |
| Направление обучения | Генеральная совокупность | Университет X | Отклонение в % |
| 1. ИВТ | 4,56 | 4,64 | 1,8 |
| 1. Психология | 4,00 | 4,15 | 3,8 |
| 1. Инноватика | 4,41 | 4,59 | 4,1 |

ДИАГРАММА КОМПОНЕНТОВ.

Диаграмма компонентов - это структурная диаграмма языка унифицированного моделирования, она описывает особенности физического представления системы. Диаграмма компонентов позволяет определить архитектуру разрабатываемой системы, установив зависимости между программными компонентами. Диаграмма компонентов предоставляет общую картину архитектуры системы, помогает разработчикам и архитекторам лучше понять ее структуру и взаимосвязи, а также является полезным инструментом для коммуникации и документирования архитектурных решений.

" отписать: мол данная работа может использоваться как пример полноценного проекта специалиста DS, а также надлежащим для устройства на работу"

4) ПОСТАНОВКА РЕШАЕМЫХ ЗАДАЧ (можно в аннотации, подумать)

5) СБОР ДАННЫХ:

- анализ литературы (график сравнения встречаемости фич, возможно, зарубежной)

- добавка своих фич (с будущей приоритезацией или удаления)

- аргументирование выбранного способа сбора данных

- проведение анкетирования (практический опыт)

- реализация оформления анкетирования

- макет анкеты (желательно фото)

- Обзор полученных данных (на момент опыт 400... записей)

- Вывод

6) АНАЛИЗ ДАННЫХ (подготовка, визуализация результатов):

- первичный анализ

- анализ при помощи мат. аппарата

- анализ при помощи "новых" инструментов

7) МОДЕЛИРОВАНИЕ (к каждому в конце вывод, желательно визуализировать все этапы, как раньше решались подобные задачи):

- мат. модели без ПО

- Classic ML модели

- DL модели

- Общие выводы

8) СРАВНЕНИЕ МОДЕЛЕЙ

9) СОСТАВЛЕНИЕ ПРОГРАММНОГО МОДУЛЯ (ДЕСКТОП или ВЕБ СЕРВИС)

- возможные варианты создания модуля

- аргументация выбранного способа

10) РЕЗЮМИРОВАНИЕ

- СЛАБЫЕ МЕСТЕ (где неоднозначно интерпретация и возможные влияющие факторы)

- ВОЗМОЖНОСТЬ К МАСШТАБИРУЕМОСТИ

11 - 9) ИСПОЛЬЗУЕМЫЕ СРЕДСТВА (либы, языки, мат аппарат, утилиты)

- аргументирование выбранных средств: "Первостепенно стояла задача использовать весь инструментарий полученый за время обучения..."

14) ИСПОЛЬЗУЕМЫЕ ТЕМЫ (теоремы, формулы и т. д.)

15) Благодарность и контакты для связи

... доработка

ГЛАВА 1. ПЕРВЫЙ ЭТАП РАБОТЫ 11

1.1 ИСПОЛЬЗУЕМЫЕ ТЕРМИНЫ 11

1.1.1 ОБЩАЯ ЧАСТЬ 11

1.1.2 МАТЕМАТИЧЕСКАЯ ЧАСТЬ 12

1.1.3 MACHINE LEARNING ЧАСТЬ 13

1.1.4 DEEP LEARNING ЧАСТЬ 13

1.2 МОТИВАЦИЯ АВТОРА 14

1.2 АРГУМЕНТИРОВАНИЕ НАДОБНОСТИ И ВЫБОРА 15

1.3 МЕСТО ДЛЯ ИСПОЛЬЗОВАНИЯ 15

ROAD MAP ДЕЙСТВИЙ 15

Определение основных этапов Data Science проект 16

АНАЛИЗ ИМЕЮЩИЙСЯ ЛИТЕРАТУРЫ 21

ГЛАВА 2. ПРАКТИЧЕСКАЯ ЧАСТЬ 44

СБОР ДАННЫХ 44

ВЫБОР ФИЧ 45

АВТОРСКИЙ ВКЛАД ВЫБОРА ФИЧ 47

СПЕЦИАЛИСТ ПО РАБОТЕ С ДАННЫМИ 48

Успеваемость важный фактор для мониторинга знаний социума государства. Примечательно то, что способы для анализа успеваемости существовали издавна. Фактически, с появлением оценивающих моделей и рейтинговых систем специалисты стали анализировать и рассматривать успеваемость социальных групп. Например: следующими образами:

• оценка знаний и навыков учащихся через тесты, контрольные работы, экзамены;

• оценивание работы и участия учащихся в классе, активность на занятиях, выполнение домашних заданий;

• присуждение баллов, оценок или степеней на основе результатов оценивания и сравнение их с установленными стандартами или нормами.

Делать начали это для того, чтобы регулировать успеваемость социальных групп. Так выгоняют из учебных заведений тех, кто плохо учится, чтобы не тратить на них ресурсы. Или поощряют тех, кто хорошо учится.

Сами по себе данные об успеваемости ничего не дают, свою сущность они раскрывают в работе аналитиков и специалистов предметной области.

Также успеваемость стала качественной меткой обучающегося. В большинстве случаев, к отличникам относятся с уважением, так как зачастую для получения хорошей оценки, человеку нужно приложить должное количество усилий. И наоборот, к неуспевающим по учебё относятся с некой предвзятостью (если не брать во внимание физиологический особенности обучающегося), так как последние не имеют дисциплину и желание развиваться в академической деятельности.

то добавим пару своих признаков для исследования. Так участие в активностях.

В ходе были изучены работы разных типов и форматов связанные с ВКР: курсы (на различных языках), книги, видеоматериалы.

Успеваемость:

\* со стороны студента: мониторинг качества осваиваемых дисциплин, показатель качества себя как специалиста в будущем.

\* со стороны университета: диагностика подготовки специалистов для государственных нужд. То есть успеваемость студентов в какой-то степени является показателем качества проводимого образования университетом.

По мнению авторов, анализ результатов ЕГЭ может быть также использован для выявления среди первокурсников «групп риска» по успеваемости.

Если рассматривать тему данного исследования во всемирной паутине, то можно несколько десятков международных работ раскрывающих и решающих проблему по своему. Но, возможно, данная работа способствует модернизации уровня текущего образования.

Очевидно, что хотелось бы иметь контролируемую систему образования, в которой у нас были: мониторинговые системы (чтобы посмотреть, что с успеваемостью студентов на данный момент), прогностические модели и расчёты (чтобы понимать в каком состоянии рано или поздно окажется система), изменяемость (чтобы проанализировав аналитическую модель системы, мы смогли повлиять на её состояние в будущем).