

**МИНОБРНАУКИ РОССИИ**

**федеральное государственное бюджетное образовательное учреждение**

**высшего образования**

**«Московский государственный технологический университет «СТАНКИН»**

**(ФГБОУ ВО «МГТУ «СТАНКИН»)**

|  |  |
| --- | --- |
| **Институт**  **информационных**  **технологий** | **Кафедра информационных**  **технологий и**  **вычислительных систем** |

|  |  |  |
| --- | --- | --- |
| **Ердоган Дениз Ердалович** | | |
| **Выпускная квалификационная работа по направлению подготовки 09.03.01 «Информатика и вычислительная техника»  профиль «Программное обеспечение средств вычислительной техники и автоматизированных систем»** | | |
| **на тему:**  **«Исследование инструментов Data Science и разработка программного модуля для анализа успеваемости студента»** | | |
|  | | |
|  | Регистрационный № \_\_\_\_\_\_\_\_\_\_ | |
| **Заведующий кафедрой ИТиВС**  **к. т. н. , доцент** |  | **Новоселова О. В.** |
|  | подпись |  |
| **Руководитель**  **науч. преп, ст. преп.** |  | **Бекмурзаев В. А.** |
|  | подпись |  |
| **Студент группы ИДБ-20-02** |  | **Ердоган Д. Е.** |
|  | подпись |  |

Москва 2023 г.



**МИНОБРНАУКИ РОССИИ**

**федеральное государственное бюджетное образовательное учреждение**

**высшего образования**

**«Московский государственный технологический университет «СТАНКИН»**

**(ФГБОУ ВО «МГТУ «СТАНКИН»)**

|  |  |
| --- | --- |
| **Институт**  **информационных**  **технологий** | **Кафедра информационных технологий и вычислительных систем** |

|  |  |
| --- | --- |
|  | **«УТВЕРЖДАЮ»**  Заведующий кафедрой ИТиВС  к. т. н. , доц. Новоселова О. В.  \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  «\_\_\_»\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_2023 г. |

**Задание**

на выполнение выпускной квалификационной работы   
по направлению подготовки 09.03.01 «Информатика и вычислительная техника»  
профиль «Программное обеспечение средств вычислительной техники и автоматизированных систем»

|  |  |
| --- | --- |
| Студент группы ИДБ-20-02 Ердоган Д. Е. | Научный руководитель  Бекмурзаев В. А. |

**Тема: «Исследование инструментов Data Science и разработка программного модуля для анализа успеваемости студента»**

1. **Описание задания на выполнение ВКР**
   1. **Тип ВКР –** исследовательскаяработа.
   2. **Цель исследования –** повышение прогнозирования успеваемости студента.
   3. **Объект исследования –** методы Data Science.
   4. **Предмет исследования –** статистические данные студентов вузов.
   5. **Методы исследования –** статистический анализ, дедукция, машинное обучение, извлечение и обработка данных, визуализация данных, системный анализ.
   6. **Задачи исследования:**
      1. Проанализировать и изучить исследования и работы связанные с выявлением корреляции между успеваемостью студента и его данных.
      2. Определить размер выборки, ключевые параметры, собрать и подготовить данные для анализа.
      3. Произвести обучение моделей на подготовленных данных и вывести зависимости.
      4. Интерпретировать полученные результаты, их визуалиция, разработать программное обеспечения для получения аналитической сборки по успеваемости студента.
2. **Требования к выполнению ВКР**
   1. **Соблюдение требований законодательной базы и стандартов**
      1. Образовательная программа высшего образования в бакалавриате ФГБОУ ВО «МГТУ «СТАНКИН» по направлению подготовки 09.03.01 «Информатика и вычислительная техника» для направленности (профиля) подготовки «Программное обеспечение средств вычислительной техники и автоматизированных систем» (утв. 07.04.2016 (изменена и дополнена)).
      2. Внутренний нормативный документ. П 01-04/264/2017. Положение о государственной итоговой аттестации по образовательным программам высшего образования – программам бакалавриата, программам специалитета и программам магистратуры: [утверждено приказом ректора от 31.08.2017 г. №431/1, одобрено решением ученого совета Университета от 31.08.2017 г., протокол № 10/17] – ФГБОУ ВО «МГТУ «СТАНКИН».
      3. Внутренний нормативный документ. П 01-04/438/2021. Положение о выпускной квалификационной работе обучающихся по образовательным программам высшего образования – программам бакалавриата, программам специалитета, программам магистратуры: [утверждено приказом врио ректора от 30.03.2021 г. №177/1, одобрено решением ученого совета Университета от 25.12.2020 г., протокол № 11/20] – ФГБОУ ВО «МГТУ «СТАНКИН».
   2. **Дополнительные требования**
      1. Оформление раздела «Список литературы» по национальному стандарту ГОСТ Р 7.0.100-2018 “Библиографическая запись. Библиографическое описание. Общие требования и правила составления”.
      2. Результаты исследования должны быть опубликованы в виде научных статей и тезисов докладов (не менее 1)
   3. **Срок сдачи оформленной квалификационной работы на кафедру – вторая декада мая 2024 г.**

|  |  |  |
| --- | --- | --- |
| Исполнитель |  | Ердоган Д. Е. |
| Научный руководитель,  Старший преподаватель |  | Бекмурзаев В. А. |

**ГРАФИК**

выполнения выпускной квалификационной работы (ВКР) бакалавра на тему

«**Исследование инструментов Data Science и разработка программного модуля для анализа успеваемости студента**»

студента гр. ИДБ-20-02 Ердогана Дениза Ердаловича

|  |  |  |
| --- | --- | --- |
| **Мероприятия** | **Сроки выполнения** | **Отметка руководителя о выполнении** |
| 1. Сбор и анализ существующих методов сбора социологической информации. | до 27 ноября 2023 г. |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

График составлен 10 октября 2023 г.

|  |  |  |
| --- | --- | --- |
| С графиком ознакомлен |  | Ердоган Д. Е. |
|  | (подпись) |  |
|  |  |  |
| Руководитель |  | Бекмурзаев В.А. |
|  | (подпись) |  |

Оглавление

[Глава 1 - Введение 7](#_Toc153130349)

[1.1 Основные понятия 7](#_Toc153130350)

[1.2 Актуальность работы 8](#_Toc153130351)

[1.3 Определение основных этапов Data Science проекта 8](#_Toc153130352)

[1.4 Этап сбора данных 12](#_Toc153130353)

[Список используемой литературы 16](#_Toc153130354)

# Глава 1 - Введение

### 1.1 Основные понятия

Цель работы заключается в рассмотрении Data Science (DS) инструментов и разбора различных задач связанных с успеваемостью студентов.

Для грамотного рассмотрения дальнейшего текста, а также для точного осознавания повествования, стоит дать определения основных понятий с примерами, чтобы слушатель полноценно представлял о чём идёт речь.

Нужно отметить, что часть из определений неоднозначно трактуется, автор указал лишь одну из возможных трактовок уместных и подходящих для данной работы.

*Таблица 1. Основные понятия по дипломной работе*

|  |  |  |
| --- | --- | --- |
| Название основного понятия | Определение основного понятия | Пример использования связанный с дипломной работой |
| **Данные** | совокупность сведений, зафиксированных на определённом носителе в форме, пригодной для постоянного хранения, передачи и обработки | Год рождения, пол, имя, фамилия, … студента |
| **Информация** | данные имеющие ценность для субъекта | Баллы по ЕГЭ, средний балл в школе,… важные для исследования и исследователя |
| **Генеральная совокупность** | совокупность всех возможных социальных объектов, которые подлежат изучению в пределах объекта исследования | Все существующие студенты мира |
| **Выборочная совокупность (выборка)** | множество случаев с помощью определённой процедуры выбранных из генеральной совокупности для участия в исследовании | Студенты учавствующие в исследовании |
| **Big Data (Большие данные)** | структурированные или неструктурированные массивы данных большого объема | Таблица значений признаков миллиона студентов |
| **Machine Learning (Машинное обучение)** | Наука создания алгоритмов (моделей), которые используют данные, как опыт | Построение ML моделей для прогнозирования успеваемости студента |

Если какое-то определение не было внесёно в данную таблицу, значит автор считает, что большинство читателей будет иметь понимание о термине.

### 1.2 Актуальность работы

Повышению успеваемости студентов высших учебных заведений всегда уделяется большое внимание. Хорошая успеваемость — свидетельство надлежащего усвоения содержания образовательной программы и залог получения высококлассного специалиста, востребованного обществом.

Нахождения зависимостей между признаками студентов и их успеваемостью может помочь в будущем студентам и преподавателям осознавать причинно-следственную связь между их характеристиками и их успеваемостью, а также снять внимание с тех факторов, которые на самом деле не влияют на успеваемость.

### 1.3 Определение основных этапов Data Science проекта

Так в Data Science-проекте имеется три основных раздела, в каждом из которых 3 этапа.

Рассмотрим основные разделы:

- Работа с бизнес-требованиями: data scientist ничего не знает о данных. Его задание - вникнуть в постановку задачи, понять, какой результат требуется получить от проекта, узнать про участников и заинтересованных лиц. Далее в соответствии с определенной задачей он должен решить, какой метод использовать для решения задачи.

Результатом этого шага будут требования к данным: что нам может понадобиться для успешного решения;

- Работа с данными: data scientist приступает к поиску данных для решения задачи: узнаёт, какие источники ему доступны, и формируем выборку, с которой в дальнейшем будет работать. После того как данные собраны, необходимо провести ряд исследований, чтобы лучше понимать, как устроена выборка:

* Исследовать: центральное положение, вариабельность;
* Выявить корреляции между признаками;
* Построить графики распределения.

После этого этапа можно приступать к подготовке данных. Как правило, этот этап самый трудоемкий процесс. В зависимости от того, насколько качественно он выполнен, зависит успех всего проекта;

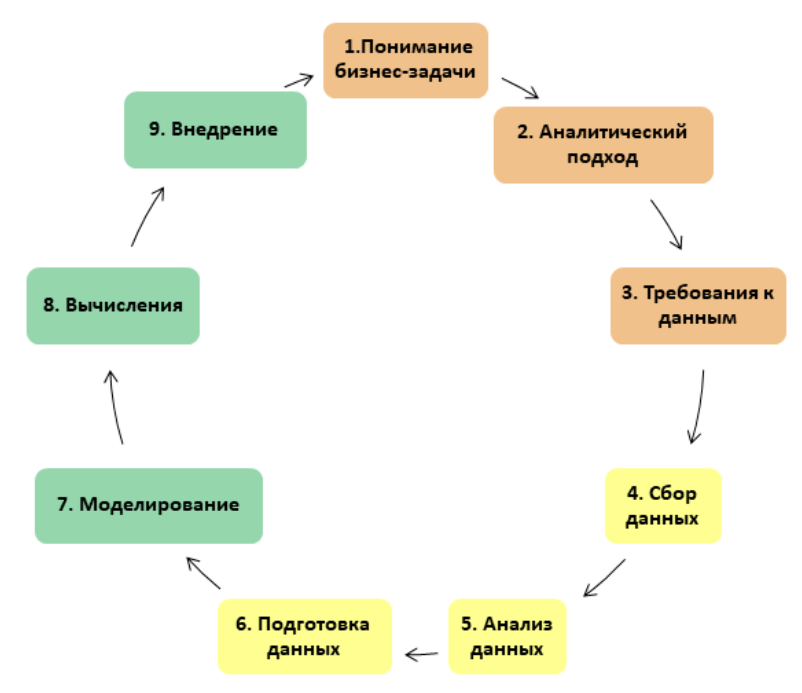
- Разработка и внедрение: после того, как данные готовы, можно приступать к разработке и внедрению. Программируем модель, прогоняем на обучающей выборке, проверяем на тестовой, если результат устраивает, то демонстрируем заказчику, внедряем, собираем фидбэк, если нет, то дорабатываем до удовлетворительного результата.

Теперь рассмотрим подробнее продемонстрированные выше разделы по этапам:

*Таблица 2. Описание основных этапов Data Science проекта*

|  |  |
| --- | --- |
| Название этапа | Описание этапа |
| Понимание бизнес-задачи | Необходимо определить цель исследования: что является проблемой? Почему проблема должна быть решена? Кого затрагивает проблема?  Главное: по каким метрикам будет оцениваться успешность проекта? Иными словами, необходимо выявить цель заказчика. |
| Аналитический подход | Нужно выбрать аналитический подход для решения бизнес-задачи. Выбор подхода зависит от того, какой тип ответа нужно получить в итоге:  - если ответ должен быть вида да/нет, то нужен классификатор;  - если ответ значение численного признака, то используются регрессионные модели;  - промежуточный вариант перечисленных деревья решений;  - если нужно определить вероятность, необходимо использовать предиктивную модель;  - если необходимо выявить связи, используется дескриптивный подход. |
| Требования к данным | Когда определена цель исследования и выбран подход, необходимо определиться с тем, какие данные позволят дать искомый ответ. Нужно подготовить требования к данным: контент, форматы, источники. |
| Сбор данных | На этом этапе выполняется сбор данных из имеющихся источников: убеждаемся, что источники доступны, надежны и могут быть использованы для получения искомых данных в требуемом качестве.  После необходимо понять, получили ли мы те данные, какие хотели. На этой стадии можно пересмотреть требования к данным и принять решения о необходимости дополнительных данных. Могут быть выявлены лакуны в данных и составлен план, как их закрыть или найти замену. |
| Анализ данных | Анализ данных включает в себя все работы по конструированию выборки. На этом этапе необходимо получить ответ на вопрос: репрезентативны ли собранные данные относительно поставленной задачи?  Здесь используется описательная статистика. Она применяется ко всем переменным, которые будут использоваться в выбранной модели:  - исследуется центральное положение;  - ищутся выбросы и выполняется оценка вариабельности;  - строятся гистограммы распределения переменных;  - визуализируются данные;  - выполняется попарное сравнение: вычисляются корреляции между переменными, чтобы определить, какие из них связаны и насколько. Если найдутся значительные корреляции между переменными, какие то из них могут быть отброшены, как избыточные. |
| Подготовка данных | На этом этапе мы перерабатываем данные в такую форму, чтобы с ними было удобно работать: удаляем дубликаты, обрабатываем отсутствующие или неверные данные, проверяем и при необходимости исправляем ошибки форматирования. Также на этом этапе мы конструируем набор факторов, с которым на следующих этапах будет работать машинное обучение: проводим извлечение и отбор признаков, которые потенциально помогут решить бизнес-задачу. Ошибки на этом этапе могут оказаться критическими для всего проекта, поэтому к нему стоит отнестись особенно внимательно: избыточное количество признаков может привести к тому, что модель будет переобучена, а недостаточное — к тому, что модель будет недообучена. |
| Построение модели | Выбор модели, как можно было заметить, осуществляется в самом начале работы и зависит от бизнес-задачи. Таким образом, когда тип модели определен и имеется обучающая выборка, аналитик разрабатывает модель и проверяет, как она работает на созданном на этапе 6 наборе признаков. |
| Применение модели | Применение модели идет в тесной связке с собственно построением модели: вычисления чередуются с настройкой модели. На этом этапе мы должны ответить на вопрос, отвечает ли построенная модель бизнес-задаче. Вычисление модели имеет две фазы: проводятся диагностические измерения, которые помогают понять, работает ли модель, так как задумано. Если используется предиктивная модель, может использоваться дерево решений, чтобы понять, что выдача модели соответствует изначальному плану. На второй фазе проводится проверка статистической значимости гипотезы. Она необходима, чтобы убедиться, что данные в модели правильно используются и интерпретируются и полученный результат выходит за пределы статистической погрешности. |
| Внедрение | Если модель дает нам удовлетворительный ответ на поставленный вопрос, этот ответ должен начать приносить пользу. Когда модель разработана, и аналитик уверен в результате своей работы, необходимо познакомить заказчика с разработанным инструментом. Имеет смысл привлечь не только владельца продукта, но и других заинтересованных лиц: маркетинг, разработчиков, системные администраторы: всех, кто хоть как то может оказать влияние на дальнейшее использование результатов проекта. Далее необходимо переходить к внедрению. Внедрение может происходить поэтапно, например, на ограниченную группу пользователей или в тестовом окружении. Также необходимо наладить систему фидбэка, чтобы отслеживать, насколько успешно разработанная модель справляется с поставленной задачей. Через некоторое время этот фидбэк будет полезен для того, чтобы усовершенствовать модель. |

Таким образом, нет предела совершенству: даже внедренная модель никогда не может считаться идеальной. Если визуализировать таблицу, то получится следующий рисунок:



*Рис. 1 Основные этапы Data Science проекта*

### 1.4 Этап сбора 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 |

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

# Список используемой литературы

1. Толстова Ю. Н. МАТЕМАТИЧЕСКАЯ СТАТИСТИКА ДЛЯ СОЦИОЛОГОВ : учеб. пособие / Толстова Ю. Н. — 2-е изд.. — Москва: Юрайт, 2021 — 258 c.;
2. Вьюгин В. В, МАТЕМАТИЧЕСКИЕ ОСНОВЫ МАШИННОГО ОБУЧЕНИЯ И ПРОГНОЗИРОВАНИЯ [Текст] / Вьюгин В. В, — 1-ое изд.. — Москва: , 2013-2020 — 484 c.;
3. Савельев В. Статистика и котики [Текст] / Савельев В. — 1-ое изд.. — Москва: АСТ, 2018 — 170 c.;
4. Академия наук СССР Математические методы в социологическом исследовании [Текст] / Академия наук СССР — 1-ое изд.. — Москва: Наука, 1981 — 33 c.;
5. Бурков А. МАШИННОЕ ОБУЧЕНИЕ без лишних слов [Текст] / Бурков А. — 2-ое изд.. — Санкт-Петербург: Питер, 2020 — 192 c.