TOMSK POLYTECHNIC UNIVERSITY ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

Министерство науки и высшего образования Российской Федерации
федеральное государственное автономное
образовательное учреждение высшего образования
«Национальный исследовательский Томский политехнический университет» (ТПУ)

---

Школа Информационных технологий и робототехники
Направление подготовки 09.04.04 Программная инженерия
Отделение школы (НОЦ) Информационных технологий

## МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

| Тема работы |
|---|
| **Прогнозная модель для оценки успеваемости студентов университета по итогам текущего обучения.** |

УДК 004.81:005.853:378.4

Студент

| Группа | ФИО | Подпись | Дата |
|---|---|---|---|
| 8ПМ8И | Зяблецев Павел Андреевич | | 10.06.2020 |

Руководитель ВКР

| Должность | ФИО | Ученая степень, звание | Подпись | Дата |
|---|---|---|---|---|
| Доцент ОИТ ИШИТР | Губин Евгений Иванович | к.ф.-м.н. | | 10.06.2020 |

## КОНСУЛЬТАНТЫ ПО РАЗДЕЛАМ:

По разделу «Финансовый менеджмент, ресурсоэффективность и ресурсосбережение»

| Должность | ФИО | Ученая степень, звание | Подпись | Дата |
|---|---|---|---|---|
| Доцент ОСГН ШБИП | Меньшикова Екатерина Валентиновна | к.ф.н. | | 02.06.2020 |

По разделу «Социальная ответственность»

| Должность | ФИО | Ученая степень, звание | Подпись | Дата |
|---|---|---|---|---|
| Доцент ООД ШБИП | Горбенко Михаил Владимирович | к.т.н | | 10.06.2020 |

## ДОПУСТИТЬ К ЗАЩИТЕ:

| Руководитель ООП | ФИО | Ученая степень, звание | Подпись | Дата |
|---|---|---|---|---|
| Доцент ОИТ ИШИТР | Губин Евгений Иванович | к.ф.-м.н. | | 10.06.2020 |

Томск – 2020 г.

TOMSK POLYTECHNIC UNIVERSITY ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

Министерство науки и высшего образования Российской Федерации
федеральное государственное автономное
образовательное учреждение высшего образования
«Национальный исследовательский Томский политехнический университет» (ТПУ)

Школа Информационных технологий и робототехники
Направление подготовки (специальность) 09.04.04. Программная инженерия
Отделение школы (НОЦ) Информационных технологий

УТВЕРЖДАЮ:
Руководитель ООП
_____ 10.05.2020 Губин Е.И.
(Подпись) (Дата) (Ф.И.О.)

**ЗАДАНИЕ**
**на выполнение выпускной квалификационной работы**

В форме:

| Магистерской диссертации |
|---|

(бакалаврской работы, дипломного проекта/работы, магистерской диссертации)

Студенту:

| Группа | ФИО |
|---|---|
| 8ПМ8И | Зяблецев Павел Андреевич |

Тема работы:

| Прогнозная модель для оценки успеваемости студентов университета по итогам текущего обучения. | |
|---|---|
| Утверждена приказом директора (дата, номер) | 59-62/С |

| Срок сдачи студентом выполненной работы: | 13.06.2020 |
|---|---|

**ТЕХНИЧЕСКОЕ ЗАДАНИЕ:**

| | |
|---|---|
| **Исходные данные к работе** | Набор данных по успеваемости 8600 студентов |
| **Перечень подлежащих исследованию, проектированию и разработке вопросов** | 1. Анализ предметной области<br>2. Предварительная обработка данных<br>3. Модель машинного обучения<br>4. Графический интерфейс пользователя прогнозной модели<br>5. Финансовый менеджмент |

| | |
|---|---|
| | ресурсоэффективность и ресурсосбережение<br>6. Социальная ответственность |
| **Перечень графического материала** | 1. Описательные диаграммы предметной области<br>2. Матрица корреляций<br>3. Примеры графического интерфейса пользователя прогнозной модели |

**Консультанты по разделам выпускной квалификационной работы**
*(с указанием разделов)*

| Раздел | Консультант |
|---|---|
| Финансовый менеджмент, ресурсоэффективность и ресурсосбережение | Меньшикова Екатерина Валентиновна |
| Социальная ответственность | Горбенко Михаил Владимирович |
| Раздел на английском языке | Пичугова Инна Леонидовна |

**Названия разделов, которые должны быть написаны на русском и иностранном языках:**

1. Анализ предметной области

| | |
|---|---|
| **Дата выдачи задания на выполнение выпускной квалификационной работы по линейному графику** | 01.03.2020 |

**Задание выдал руководитель:**

| Должность | ФИО | Ученая степень, звание | Подпись | Дата |
|---|---|---|---|---|
| Доцент ОИТ ИШИТР | Губин Евгений Иванович | к.ф.-м.н. | | 10.06.2020 |

**Задание принял к исполнению студент:**

| Группа | ФИО | Подпись | Дата |
|---|---|---|---|
| 8ПМ8И | Зяблецев Павел Андреевич | | 10.06.2020 |

TOMSK POLYTECHNIC UNIVERSITY ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

Министерство науки и высшего образования Российской Федерации
федеральное государственное автономное
образовательное учреждение высшего образования
«Национальный исследовательский Томский политехнический университет» (ТПУ)

Школа Информационных технологий и робототехники
Направление подготовки (специальность) 09.04.04 Программная инженерия
Уровень образования Магистратура
Отделение школы (НОЦ) Информационных технологий
Период выполнения Весенний семестр 2019 /2020 учебного года

Форма представления работы:

| Магистерская диссертация |
|---|

(бакалаврская работа, дипломный проект/работа, магистерская диссертация)

**КАЛЕНДАРНЫЙ РЕЙТИНГ-ПЛАН**
**выполнения выпускной квалификационной работы**

| | |
|---|---|
| Срок сдачи студентом выполненной работы: | 10.06.2020 |

| Дата контроля | Название раздела (модуля) / вид работы (исследования) | Максимальный балл раздела (модуля) |
|---|---|---|
| 10.03.2020 | Раздел 1. Анализ предметной области | 15 |
| 10.04.2020 | Раздел 2. Предварительная обработка данных | 20 |
| 10.05.2020 | Раздел 3. Модель машинного обучения | 30 |
| 20.05.2020 | Раздел 4. Графический интерфейс пользователя прогнозной модели | 10 |
| 30.05.2020 | Раздел 5. Финансовый менеджмент, ресурсоэффективность и ресурсосбережение | 15 |
| 30.05.2020 | Раздел 6. Социальная ответственность | 10 |

**СОСТАВИЛ:**
**Руководитель ВКР**

| Должность | ФИО | Ученая степень, звание | Подпись | Дата |
|---|---|---|---|---|
| Доцент ОИТ ИШИТР | Губин Евгений Иванович | к.ф.-м.н. | | 10.06.2020 |

**СОГЛАСОВАНО:**
**Руководитель ООП**

| Должность | ФИО | Ученая степень, звание | Подпись | Дата |
|---|---|---|---|---|
| Доцент ОИТ ИШИТР | Губин Евгений Иванович | к.ф.-м.н. | | 10.06.2020 |

**ЗАДАНИЕ ДЛЯ РАЗДЕЛА**
**«ФИНАНСОВЫЙ МЕНЕДЖМЕНТ, РЕСУРСОЭФФЕКТИВНОСТЬ И РЕСУРСОСБЕРЕЖЕНИЕ»**

Студенту:

| Группа | ФИО |
|---|---|
| 8ПМ8И | Зяблецев Павел Андреевич |

| Школа | ИШИТР | Отделение школы (НОЦ) | ОИТ |
|---|---|---|---|
| Уровень образования | Магистратура | Направление/специальность | 09.04.04 Программная инженерия |

| **Исходные данные к разделу «Финансовый менеджмент, ресурсоэффективность и ресурсосбережение»:** | |
|---|---|
| *1. Стоимость ресурсов научного исследования (НИ): материально-технических, энергетических, финансовых, информационных и человеческих* | Стоимость материальных ресурсов определялась согласно прейскурантам компаний<br>Оклад руководителя – 33664 р.<br>Оклад инженера – 21760 р. |
| *2. Нормы и нормативы расходования ресурсов* | Накладные расходы 16 %;<br>Районный коэффициент 30%;<br>Норма амортизации ПЭВМ 33,33%;<br>Норма амортизации ПО 20% |
| *3. Используемая система налогообложения, ставки налогов, отчислений, дисконтирования и кредитования* | Коэффициент отчислений на уплату во внебюджетные фонды 30% |
| **Перечень вопросов, подлежащих исследованию, проектированию и разработке:** | |
| *1. Оценка коммерческого и инновационного потенциала НТИ* | Анализ потенциальных потребителей результатов исследования, оценка качества и перспективности проекта по технологии QuaD, SWOT-анализ |
| *2. Разработка устава научно-технического проекта* | Инициация проекта: определение заинтересованных сторон проекта, целей и результатов проекта |
| *3. Планирование процесса управления НТИ: структура и график проведения, бюджет, риски и организация закупок* | План проекта, определение трудоемкости выполнения работ, разработка графика проведения научного исследования, расчет бюджета разработки |
| *4. Определение ресурсной, финансовой, экономической эффективности* | Описание потенциального эффекта |
| **Перечень графического материала** *(с точным указанием обязательных чертежей):* | |
| *1. Матрица SWOT разработки*<br>*2. Диаграмма Ганта*<br>*3. График проведения и бюджет НТИ*<br>*4. Потенциальные риски* | |

| **Дата выдачи задания для раздела по линейному графику** | |
|---|---|

**Задание выдал консультант:**

| Должность | ФИО | Ученая степень, звание | Подпись | Дата |
|---|---|---|---|---|
| Доцент ОСГН ШБИП | Меньшикова Екатерина Валентиновна | к.ф.н | | |

**Задание принял к исполнению студент:**

| Группа | ФИО | Подпись | Дата |
|---|---|---|---|
| 8ПМ8И | Зяблецев Павел Андреевич | | |

# ЗАДАНИЕ ДЛЯ РАЗДЕЛА «СОЦИАЛЬНАЯ ОТВЕТСТВЕННОСТЬ»

Студенту:

| Группа | ФИО |
|---|---|
| 8ПМ8И | Зяблецев Павел Андреевич |

| Школа | ИШИТР | Отделение (НОЦ) | ОИТ |
|---|---|---|---|
| Уровень образования | Магистратура | Направление/специальность | 09.04.04 Программная инженерия |

Тема ВКР:

| Прогнозная модель для оценки успеваемости студентов университета по итогам текущего обучения. | |
|---|---|
| **Исходные данные к разделу «Социальная ответственность»:** | |
| 1. Характеристика объекта исследования (вещество, материал, прибор, алгоритм, методика, рабочая зона) и области его применения | Объект исследования – рабочее место программиста, разрабатывающего прогнозную модель для оценки успеваемости студентов университета по итогам текущего обучения, разработанная на персональном компьютере.<br><br>**Рабочая зона** – аудитория, оборудованная системой отопления, кондиционирования воздуха, с естественным и искусственным освещением. Рабочее место – стационарное, оборудованное персональным компьютером и оргтехникой. |
| Перечень вопросов, подлежащих исследованию, проектированию и разработке: | |
| **1. Правовые и организационные вопросы обеспечения безопасности:**<br>– Специальные (характерные при эксплуатации объекта исследования, проектируемой рабочей зоны) правовые нормы трудового законодательства;<br>– Организационные мероприятия при компоновке рабочей зоны. | – Трудовой кодекс Российской Федерации от 30.12.2001 ФЗ-197<br>– Федеральный закон от 27.07.2006 N 152-ФЗ (ред. От 25.07.2011) «О персональных данных»<br>– Рабочее место при выполнении работ сидя регулируется ГОСТом 12.2.032 – 78<br>– Организация рабочих мест с электронно-вычислительными машинами регулируется СанПиНом 2.2.2/2.4.1340 – 03 |
| **2. Производственная безопасность**<br>2.1. Анализ выявленных вредных и опасных факторов:<br>2.2. обоснование мероприятий по снижению воздействия: | Вредные факторы:<br>– Повышенный уровень электромагнитных излучений<br>– Отклонение показателей микроклимата<br>– Недостаточная освещенность рабочей зоны<br>– Повышенный уровень шума на рабочем месте |

| | |
|---|---|
| | – Умственное перенапряжение<br><br>Опасные факторы:<br>– Опасность поражения электрическим током<br>– статическое электричество короткое замыкание |
| **3. Экологическая безопасность:** | Воздействия на окружающую природную среду: утилизация люминесцентных ламп, компьютеров и другой оргтехники |
| **4. Безопасность в чрезвычайных ситуациях:** | Возможной чрезвычайной ситуацией техногенного характера для данной сферы деятельности является пожар в результате возгорания электропроводки, перегрева рабочих частей персонального компьютера. Создание общих правил поведения и рекомендаций во время пожара, разработка плана эвакуации, ознакомление с использованием огнетушителей типа ОУ-5. |

| | |
|---|---|
| **Дата выдачи задания для раздела по линейному графику** | **01.03.2020** |

**Задание выдал консультант:**

| Должность | ФИО | Ученая степень, звание | Подпись | Дата |
|---|---|---|---|---|
| Доцент ООД ШБИП | Горбенко Михаил Владимирович | к.т.н | | |

**Задание принял к исполнению студент:**

| Группа | ФИО | Подпись | Дата |
|---|---|---|---|
| 8ПМ8И | Зяблецев Павел Андреевич | | |

**Планируемые результаты обучения**

| Код | Результат обучения |
|---|---|
| Общие по направлению 09.04.04 «Программная инженерия» | |
| Р1 | Проводить научные исследования, связанные с объектами профессиональной деятельности |
| Р2 | Разрабатывать новые и улучшать существующие методы и алгоритмы обработки данных в информационно-вычислительных системах |
| Р3 | Составлять отчеты о проведенной научно-исследовательской работе и публиковать научные результаты |
| Р4 | Проектировать системы с параллельной обработкой данных и высокопроизводительные системы |
| Р5 | Осуществлять программную реализацию информационно-вычислительных систем, в том числе распределенных |
| Р6 | Осуществлять программную реализацию систем с параллельной обработкой данных и высокопроизводительных систем |
| Р7 | Организовывать промышленное тестирование создаваемого программного обеспечения |
| Профиль «Технологии больших данных»/ «Big data solutions» | |
| Р8 | Исследовать и анализировать большие данные, создавать их модели и интерпретировать структуры данных в таких моделях |
| Р9 | Понимать принципы создания, хранения, управления, передачи и анализа больших данных с использованием новейших технологий, инструментов и систем обработки данных в высокопроизводительных сетях |
| Р10 | Применять теорию распределенной системы управления базами данных к традиционным распределенным системам реляционных баз данных, облачным базам данных, крупномасштабным системам машинного обучения и хранилищам данных |

# Реферат

Выпускная квалификационная работа 108 с., 31 рис., 22 табл., 31 источник, 2 прил.

Ключевые слова: прогнозирование успеваемости студентов, модель машинного обучения, подготовка данных, набор данных, классификация.

Объектом исследования является процесс разработки прогнозной модели для оценки успеваемости студентов университета по итогам текущего обучения.

Цель работы – построение прогнозной модели итогов сессии студентов в зависимости от оценочных параметров текущей успеваемости.

В процессе исследования был проведен анализ основных проблем в рассматриваемой области, поставлены цели для их непосредственного выполнения. Также был проведена предварительная обработка данных для построения модели машинного обучения. После этого были построены различные модели машинного обучения, а также оценены качественные показатели каждой модели. После выбора оптимальной модели был создан графический интерфейс пользователя прогнозной модели.

В результате исследования была создана прогнозная модель успеваемости студентов вуза, а также графический интерфейс для её использования. Определенны наиболее значимые факторы при прогнозировании успеваемости у студентов.

Степень внедрения: на текущем этапе идет согласование о внедрении.

Область применения: образование.

Экономическая эффективность/значимость работы: данная разработка позволяет выявлять студентов, которые с большой вероятностью будут иметь академические задолженности, и применять управляющее воздействие раньше, чем появиться реальная проблема с успеваемостью. Это приводит к более эффективной учебно-воспитательной работе единого деканата ТПУ.

В будущем планируется внедрение данной прогнозной модели в информационную систему университета.

Оглавление

**Введение**

Уровень успеваемости студента в ВУЗе является своеобразной формой диагностики и прогнозирования степени отдачи будущего специалиста. В свою очередь успехи студентов – это показатель деятельности ВУЗа в решении учебно-воспитательных задач. Для того, чтобы решать данные задачи максимально эффективно требуется постоянная объективная оценка, корректировка и управление. Однако, без прогнозирования управление невозможно. Поэтому возникает необходимость прогнозирования успеваемости студентов на всех этапах обучения.

Имея сведения о тех студентах, которые вероятнее всего к концу семестра будут иметь академические задолженности, если не изменят текущую тенденцию, мы можем повлиять на студентов, тем самым повысить уровень их успеваемости.

Целью данной работы является создание прогнозной модели успеваемости студентов ТПУ. Наличие такой модели позволит уделять более пристальное внимание студентам, которые попадают в группу риска большого количества долгов по учебным дисциплинам, а как следствие, будут претендентами на отчисления. Определение таких студентов на ранних этапах позволит более детально и персонально работать с ними для того, чтобы они более успешно справлялись с учебной нагрузкой.

Исходя из цели, данная работа включает следующие задачи: обзор литературы посвященной данному вопросу, изучение используемых методов и алгоритмов, очистка и подготовка исходных данных, разработка прогнозной модели, апробирование результатов, создание графического интерфейса пользователя для модуля прогнозирования успеваемости студентов.

## 1. Анализ предметной области

Образование играет одну из важнейших ролей в любом государстве. От качества образования, существующего в конкретном обществе, во многом зависят темпы его экономического и политического развития, его нравственное состояние. Стремительное развитие информационных технологий позволяет автоматизировать многие сферы деятельности людей повышая их эффективность, и образование не является исключением. В данной работе речь пойдет о создании прогнозной модели успеваемости студентов по текущим оценкам с помощью технологий анализа данных.

Мерой измерения качества получаемого образования для конкретного студента служат его оценки по пройденным предметам. Если говорить об учебном заведении, то одной из мер качества предоставляемого им образования служит совокупность оценок его учащихся. Своевременные меры по оказанию помощи студентам, которые не справляются с учебной нагрузкой, являются одной из основных частей по учебно-воспитательной работе в ВУЗе, которые влияют на показатели качества образования в учебном заведении.

В последнее время, было сделано много различных изменений для улучшения качества образования в ВУЗах. Например, переход от традиционной системы оценок к бальным исключает возможность того, что студент, который не ходил на занятия весь семестр просто придет и сдаст экзамен. Для допуска к экзамену ему нужно набрать определенное количество балов, а для этого в свою очередь необходимо посещать занятия и выполнять текущие задания. Такой подход эффективно влияет на понимание учебного материала. Согласно многим исследованиям, информация, которая была изучена в течение длительного промежутка времени, сохранится на долгое время, в то время как, «зубрежка» в ночь перед экзаменом может дать лишь хороший результат на экзамене, который в итоге будет сдан, но остаточных знаний от предмета у студента вовсе не останется. Помимо этого, существует рубежный контроль, назначенная дата в середине семестра, когда определенная часть материала должна быть сдана. Однако, зная специфику студенческой жизни, многие

студенты все равно оставляют все на последний момент. Некоторым студентам удается сдать предмет, другие же остаются с долгом на другой семестр.

Проблема заключается в том, что управляющее воздействие со стороны единого деканата Томского Политехнического Университета начинается только после того факта, что у студента уже появилась задолженность. Таким образом, данные меры нельзя назвать превентивными. Главной задачей прогнозной модели является выделение таких студентов и проведение с ними определенных бесед до того, когда появится проблема в виде академической задолженности. В настоящее время данная мера реализована отчасти тем, что у каждой группы первокурсников есть куратор, и этот куратор сопровождает каждую группу вплоть до 2го курса, в расписании отводятся такое мероприятие как «час куратора», где он анализирует успеваемость студентов. Это является отличной практикой, и отказываться от нее нельзя, однако, тут играет роль человеческий фактор. Не всегда куратор может донести до студентов важность посещения занятий. Внедрение новой системы прогнозирования количества задолженностей в конце семестра по текущей успеваемости является важным этапом автоматизации процесса учебно-воспитательной работы. Таким образом, единый деканат сможет оказывать управляющее воздействие на студентов, попавших в группу риска, гораздо раньше возникновения реальной проблемы. Тем самым, появится возможность помочь студентам заканчивать семестр успешно и формировать профессиональные компетенции, а тех, кто не заинтересован в обучении, отчислять раньше и освобождать места для тех, кто действительно хочет и готов получать знания.

Основное внимание уделено именно системе прогнозирования, так как речь идет не просто о контроле посещаемости занятий студентом. Анализ данных показывает, что на конечный результат семестра влияет не только один фактор посещаемости, а целый комплекс различных данных. Например, существует целая категория студентов, которые уже работают по специальности, в основном это магистранты и студенты последних курсов. У таких студентов не всегда есть возможность посещения занятий, и тем не менее

они показывают отличные результаты на конец сессии, в силу того, что понимают многие аспекты профессии на более высоком уровне чем их одногруппники.

В действительности на успеваемость студента влияет огромное количество факторов, и одними из самых важных являются мотивация на учебу, моральное состояние, отношение с одногруппниками и так далее, но благодаря тому, что система будет выдавать проблемных студентов и с каждым будет возможность работать детально, выявлять какие проблемы существуют именно у этого студента, который рискует окончить семестр с большим количеством долгов. Выявление таких параметров как мотивация, целеустремленность, психологические данные возможно, но это требует большого числа тестов, которые должны проводится на регулярной основе. Данные методы не имеют высокой эффективности за счет сложности их проведения, а также проверки и интерпретации результатов.

После выяснения того, что данная проблема действительно является актуальной, стоит рассмотреть существующие методы ее решений.

На данный момент в открытом доступе нет материалов, посвященных внедрению и использованию системы прогнозирования успеваемости студента в ВУЗе. Хотя данная идея не является инновационной, примерно с 2010 года выходят статьи о прогнозировании успеваемости абитуриентов, прогнозировании успеваемости учащихся по определенному курсу, где за исходные данные как правило берутся следующие параметры: уровень текущих знаний по предмету, количество пропусков, промежуточный контроль, оценки по обеспечивающим курсам к дисциплине.

Далее рассмотрим алгоритмы машинного обучения, которые применяются для решения похожих задач в сфере прогнозирования успеваемости студентов. Будут приведены наиболее часто встречающиеся методы и алгоритмы, а также описание существующих работ с конкретными задачами, для которых эти методы применяются.

Некоторые источники применяют метод кластеризации, на наш взгляд это не совсем верно, так как у нас уже есть метки классов, а именно количество задолженностей, так что в данном случае следует применять методы классификации. Кластеризация относится к методам машинного обучения без учителя, а классификация и регрессия в свою очередь к обучению с учителем, так как имеют в качестве исходных данных маркеры для каждого класса. Методы кластерного анализа для оценки итоговой оценки по предмету были применены в данной статье. [1]

### 1.1 Алгоритм K-ближайших соседей

Один из наиболее часто встречающихся методов для решения аналогичной задачи – это Алгоритм K-ближайших соседей:

Алгоритм K-ближайших соседей (KNN, k nearest neighbors) – это тип управляемого алгоритма машинного обучения, который может использоваться как для классификации, так и для задач прогнозирования регрессии. Данный алгоритм является одним из самых простых для понимания, но тем не менее он доказал свою эффективность в ряде задач и используется не только в учебных целях. Алгоритм ближайших соседей также называют «ленивым» классификатором, потому что в процессе обучения он не строит какую-либо модель, а просто хранит данные. Все вычисления начинаются только тогда, когда необходимо классифицировать новые данные.

Суть данного алгоритма заключается в том, что прогнозирование значений новых данных основано на их близости к уже маркированным данным в обучающем наборе. Другими словами, если новая точка данных имеет среди своих ближайших соседей 4 точки класса *A* и 1 точку класса *B*, то эта новая точка будет определенна как класс *A*. Таким образом алгоритм *k*-ближайших соседей имеет 2 важнейших параметра, а именно метрика расстояния (Евклидово, Манхэттенское или Хэммингговское) и количество соседей, которое мы будем рассматривать. Визуальное представление работы алгоритма представлено на рисунке 1.

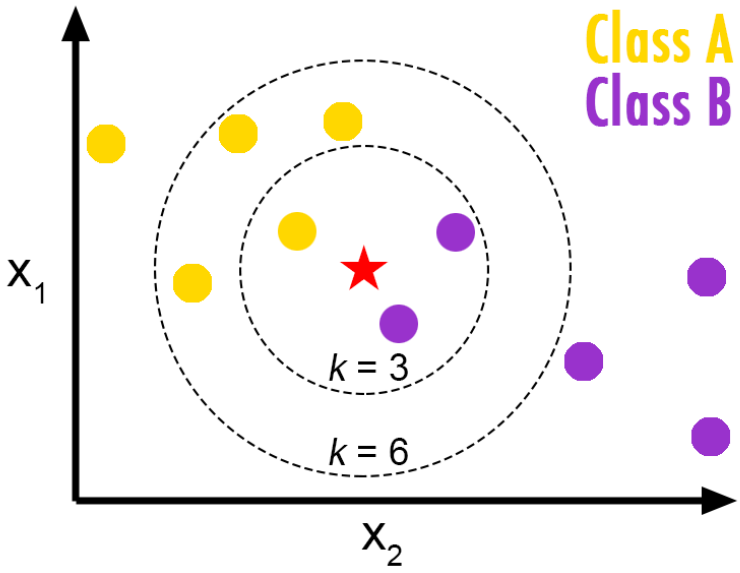



Рисунок 1. Работа алгоритма KNN

На данном рисунке изображен массив исходных точек, которые маркированы цветом в зависимости от принадлежности к классу. Звездочкой помечена точка, которую необходимо классифицировать. Все точки представлены в двумерном виде, по осям $X_1$ и $X_2$. Если установленное число ближайших соседей равно 3, то немаркированный объект будет относиться к классу *B*, если 6, то уже к классу *A*.

Данный алгоритм имеет много нюансов, например, мы можем добавить веса данным при голосовании в зависимости от близости к нашему немаркированному объекту. Именно эти нюансы делают алгоритм KNN актуальным для решения целого спектра задач.

Преимущества данного алгоритма:

- Простота понимания и интерпретации.
- Хорошо работает на нелинейных данных.
- Универсальный алгоритм, подходит как для задач классификации, так и регрессии.
- Имеет относительно высокую точность.

Недостатки алгоритма:

- Большой расход памяти на хранение всех данных в отличии от алгоритмов, которые используют построение моделей.

- Чувствительность к масштабу данных.
- Медленное прогнозирование в случае большого количества данных.
- Высокая чувствительность к “шуму” в данных.

В статье «Прогнозирование персональной успеваемости студентов в вузе» Будаевой А.А. [2] этот алгоритм используется для классификации оценки отдельного студента по каждому предмету на основании его прошлых оценок и оценок студентов прошлых курсов, с максимально аналогичными параметрами по этим предметам. В данной статье приведена достаточно высокая точность алгоритма для данной задачи, максимальная ошибка прогноза оценки составляла 0,55 балла. Однако учитывалось лишь 307 студентов одной специальности и ни о каком внедрении данной методики в систему университета речи не шло.

Также данный алгоритм используется в статье «Student performance prediction using support vector machine and k-nearest neighbor» [3], где прогнозируется оценка студента за экзамен по определенному предмету, на основании его оценок по предшествующим предметам, его посещаемости и оценок промежуточного контроля.

### 1.2 Метод опорных векторов

Метод опорных векторов (SVM – support vector machines) – это набор схожих алгоритмов вида «обучение с учителем», использующихся для решения задач классификации и регрессии. Данный метод является одним из наиболее популярных методов обучения и принадлежит к семейству линейных классификаторов. Особым свойством метода опорных векторов является непрерывное уменьшение эмпирической ошибки классификации и увеличение зазора. Поэтому этот метод также известен как метод классификатора с максимальным зазором.

Основную идею метода опорных векторов можно проиллюстрировать на примере: на плоскости имеются точки, маркированные на 2 класса, которые линейно разделимы. В таком случае результирующей функцией будет плоскость, которая разделяет эти классы. Однако, возможно провести множество гиперплоскостей, которые разделят данные классы. Для того чтобы найти оптимальную гиперплоскость, необходимо найти максимальную сумму векторов нормали от класса *A* и класса *B*. Визуальное отображение данного метода можно увидеть на рисунке 2. На данном рисунке опорными векторами будут являться перпендикулярные нормалям, которые изображены на рисунке 2.

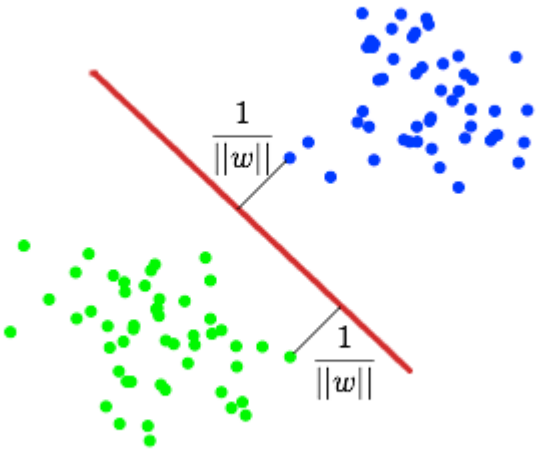



Рисунок 2. Метод опорных векторов.

Формальное описание данного метода следующее: пусть мы имеем обучающую выборку $\{(X_1, C_1), (X_2, C_2), \ldots, (X_i, C_i)\}$, где:

$X_i$ – это $p$-мерный вещественный вектор

$C_i$ – значение 1 или -1, которое принимает класс

Метод опорных векторов строит классифицирующую функцию в виде:

$$F(x) = sign([w, x] + b), \text{ (1) где,}$$

[ , ] – скалярное произведение

$w$ – нормальный вектор к разделяющей гиперплоскости

$b$ – вспомогательный параметр

Таким образом, можно записать это все в виде задачи оптимизации, которая имеет решение и при этом единственное.

$$\begin{cases} \|\mathbf{w}\|^2 \to \min \\ c_i(\mathbf{w} \cdot \mathbf{x_i} - b) \geq 1, \quad 1 \leq i \leq n. \end{cases} \quad (2)$$

Данная задача решается методом квадратичного программирования и с помощью множителей Лагранжа.

Был рассмотрен случай, когда существуют 2 разделимых класса. На практике почти всегда классы линейно не разделимы, и стоит задача классификации более чем 2х классов. Для решения задачи с линейно неразделимыми классами мы позволяем классификатору допускать ошибки на обучающей выборке. Запишем уравнения для такого допущения.

$$\begin{cases} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n} \xi_i \to \min_{w,b,\xi_i} \\ c_i(\mathbf{w} \cdot \mathbf{x_i} - b) \geq 1 - \xi_i, \quad 1 \leq i \leq n \\ \xi_i \geq 0, \quad 1 \leq i \leq n \end{cases} \quad (3)$$

Где C – параметр настройки метода,

$\mathcal{E}_i$ – величина допустимой ошибки.

Для решения мультиклассовых задач используется обобщенный метод опорных векторов за счет того, что переход к классификации на множество классов осуществляется разбиением на 2 класса, таких как подходящий класс и не подходящий. Данная стратегия также называется «один против всех» и используется для применения бинарных классификаторов для мультиклассовых задач.

Преимущества алгоритма:

- Задача хорошо изучена и имеет единственное решение.
- Принцип оптимальной разделяющей гиперплоскости приводит к уверенной классификации.

- Эквивалентен двухслойной нейронной сети, где число нейронов на скрытом слое определяется автоматически как число опорных векторов.

Недостатки:

- Неустойчивость к шуму, выбросы в исходных данных напрямую влияют на построение разделяющей гиперплоскости.
- Нет отбора признаков.
- Необходимо подбирать методы построения ядер и спрямляющих пространств отдельно для каждой задачи.

Данный алгоритм используется в статье «Student performance prediction using support vector machine and k-nearest neighbor» [3], где прогнозируется оценка студента за экзамен по определенному предмету, на основании его оценок по предшествующим предметам, его посещаемости и оценок промежуточного контроля.

### 1.3 Нейронные сети

Помимо вышеизложенных методов для прогнозирования успеваемости студентов используют еще и нейронные сети. Встречается достаточно много упоминаний про возможность использования нейронных сетей для решения данной задачи, однако информации о реальном внедрении таких прогнозных моделей не встречается.

Нейронные сети – это математические модели, построенные по принципу организации и функционирования биологических нейронных сетей. Нейронные сети не программируются в привычном смысле этого слова, они обучаются. В процессе обучения нейронная сеть способна выявлять сложные зависимости между входными данными и выходными, а также выполнять обобщение. После обучения сеть способна предсказать будущее значение некой последовательности на основе нескольких предыдущих значений.

Иллюстрация принципа работы нейронной сети представлена на рисунке 3. Нейронная сеть состоит из нейронов, слоев и синапсов. Нейроны изображены в виде узлов разного цвета. Все узлы одного цвета относятся к одному слою нейронной сети. Синапсы – это линии, которые связывают нейроны одного слоя с нейронами другого слоя. Синапсы имеют всего один параметр, и это вес.

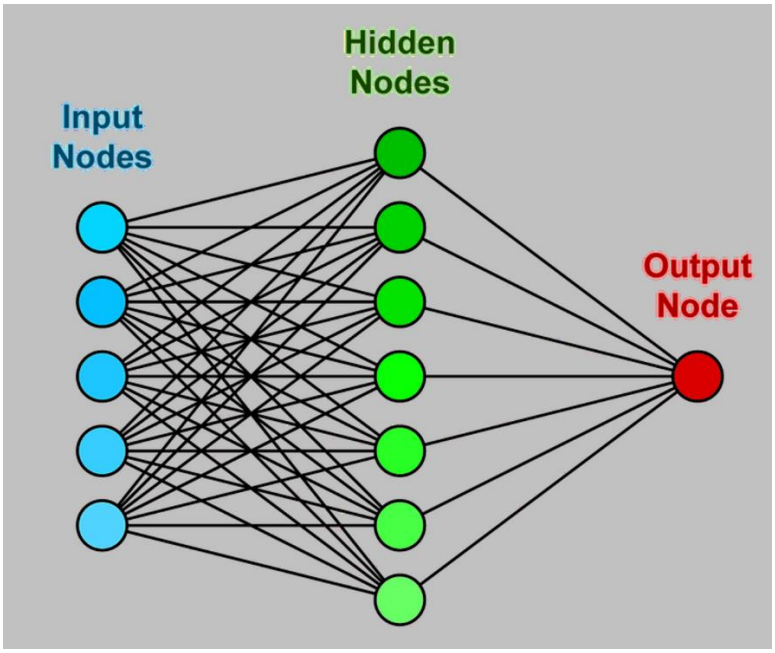



Рисунок 3. Нейронная сеть.

Каждый нейрон выполняет определенную математическую функцию, таким образом на вход ему попадает множество значение, а на выходе одно. Таким образом, на выходе получается определенное значение, которое выдала уже обученная нейронная сеть.

Преимущества:

- Устойчивость к шумам входных данных.
- Самообучение и «креативность». Возможность решения таких задач, которые не решаются другими алгоритмами.
- Адаптация к изменениям, переобучение.

Недостатки:

- Для больших сетей невозможность заранее даже приблизительно оценить время обучения сети.

- Сложность интерпретации результата.
- Приблизительность полученного ответа.

Рассмотрим более подробно применения нейронной сети для решения задачи прогнозирования успеваемости студентов. В статье «Опыт прогнозирования успеваемости студентов при помощи нейросетевой технологии» Ясинского И.Ф. [4] нейросетевая модель обучена, чтобы предсказывать будет ли студент перспективным. Решается задача бинарной классификации абитуриентов по таким входным данным как номер школы, оценки по физике и математике и профессии родителей.

В статье «Анализ и прогнозирование успеваемости студентов на основе радиальной базисной нейронной сети» [5] нейронная сеть применяется для прогнозирования оценки по курсу информатики.

В статье «Прогнозирование успеваемости студентов первого курса по результатам сдачи единого государственного экзамена» Харламовой И.Ю. [6] рассмотрена задача классификации студентов по результатам ЕГЭ.

### 1.4 Вывод по разделу

В данном разделе были рассмотрены алгоритмы, которые чаще всего используются для решения проблемы прогнозирования успеваемости учащихся на основе различных исходных данных и решают разные задачи, будь то успеваемость по конкретному предмету или общая картина успеваемости во всех дисциплинах. Рассмотренные примеры прогнозирования не носят системного характера, а являются лишь попытками приблизиться к решению данной проблемы. Очевидно, что для успешного решения поставленной задачи необходимо применить несколько методов и сравнить их результаты.

## 2. Предварительная обработка данных

Качество моделей машинного обучения очень сильно зависит от качества исходных данных. Однако, реальные данные очень часто плохо структурированы, имеют пропущенные значения, шумы, а также ошибочные значения. При условии, что данные не были качественно подготовлены никакие настройки алгоритмов машинного обучения не смогут обеспечить высокую прогнозную точность этих моделей. Подготовка данных перед их анализом занимает около 80 процентов рабочего времени специалиста по машинному обучению, но эта работа является необходимой.

В данной работе для подготовки данных будет использована методика, описанная в статье Губина Е.И. «Методика подготовки больших данных для прогнозного анализа» [7].

Для ознакомления с данными и их подготовкой для дальнейшего анализа будет использован язык программирования Python и его библиотеки. Выбор языка программирования Python обусловлен высокой производительностью при обработке данных, простотой и большим количеством библиотек для машинного обучения. Python является одним из лучших языков программирования для работы с данными. В данной работе будут использоваться следующие библиотеки Python:

- NumPy. Данная библиотека добавляет поддержку больших многомерных массивов и матриц, а также высокоуровневые команды математических функций с очень высокой производительностью над этими массивами [11].
- Pandas. Программная библиотека для обработки и анализа данных.
- Seaborn. Библиотека для визуализации данных, основанная на другой библиотеке программного языка Python Matplotlib.
- Scikit-learn. Библиотека машинного обучения на с открытым исходным кодом.

- PyQt5 – набор расширений графического фреймворка Qt для языка программирования Python, выполненный в виде расширения Python. Данная библиотека практически полностью реализует возможности Qt и позволяет создавать графический интерфейс для программ написанных на Python.

Первым делом, для того чтобы работать с данными с помощью Python необходимо их преобразовать в формат, с которым работает данный язык и его библиотеки. В данном случае таким форматом является DataFrame из библиотеки Pandas.

После преобразования, имеется возможность ознакомиться с основными характеристиками исходного датасета. (Рисунок 4)

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8551 entries, 0 to 8550
Data columns (total 24 columns):
Форма обучения                                                  8551 non-null object
Квалификация                                                    8551 non-null object
Курс                                                            8551 non-null int64
Специальность                                                   8551 non-null object
Профиль                                                         6676 non-null object
Выпуск. отдел.                                                  8551 non-null object
Выпуск. школа                                                   7988 non-null object
Группа                                                          8551 non-null object
Обуч. подразд.                                                  8551 non-null object
ID                                                              8551 non-null int64
Форма финансирования                                            8551 non-null object
Страна                                                          8533 non-null object
Гражданство                                                     8551 non-null object
Пол                                                             8551 non-null object
Дата рождения                                                   8551 non-null object
Академ отпуск (действующий) - да / нет                          8551 non-null object
Всего                                                           8551 non-null int64
Положительных                                                   8551 non-null int64
Неудовлетворительных                                            8551 non-null int64
Дисциплины по которым получены неудовлетворительные оценки      6370 non-null object
Пропусков по дисциплинам по которым получены неудовлетворительные оценки    8551 non-null int64
Всего часов по дисциплинам по которым получены неудовлетворительные оценки  8539 non-null float64
Всего часов пропусков в семестре                                8551 non-null int64
Всего часов аудиторных занятий в семестре                       8470 non-null float64
dtypes: float64(2), int64(7), object(15)
```

Рисунок 4. Наличие нулевых значений у параметров

Можно наблюдать, что в кортежи некоторых атрибутов имеют нулевые значения, для большей наглядности следует посмотреть в каких именно атрибутах они имеются. (Рисунок 5)

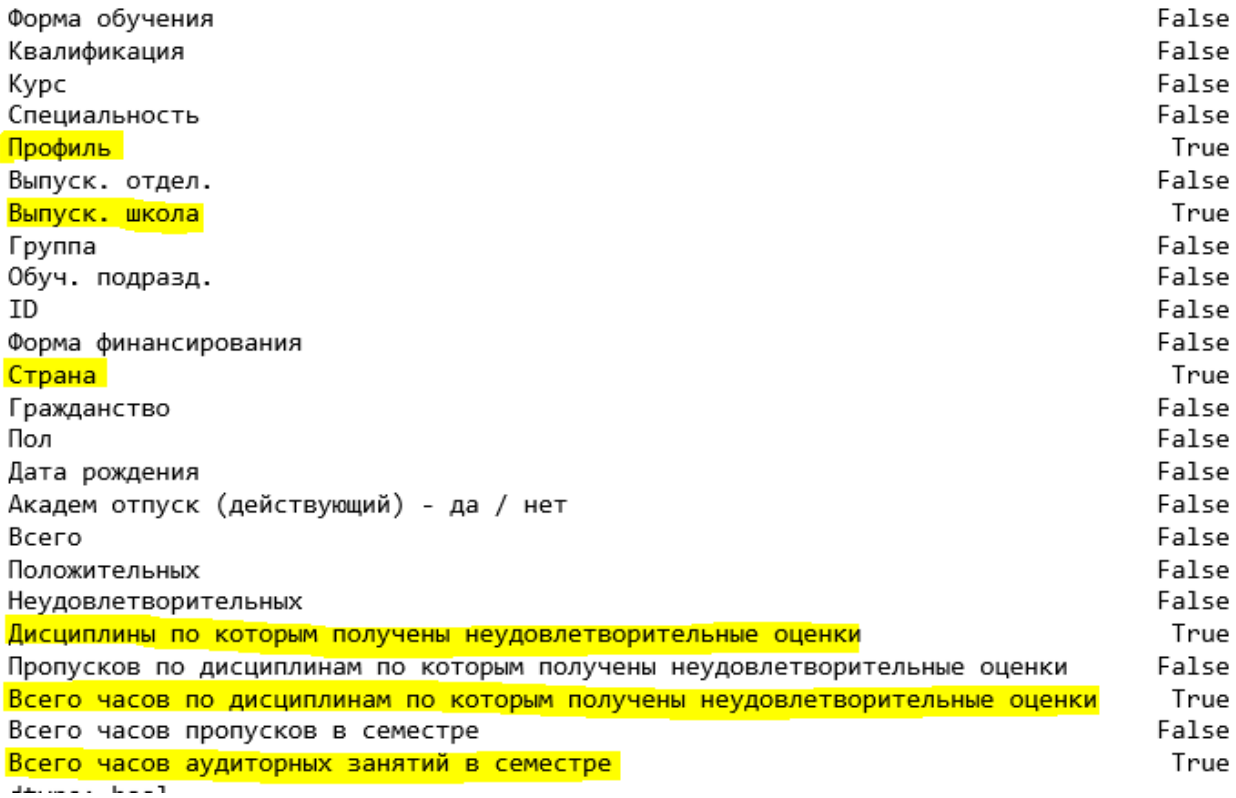
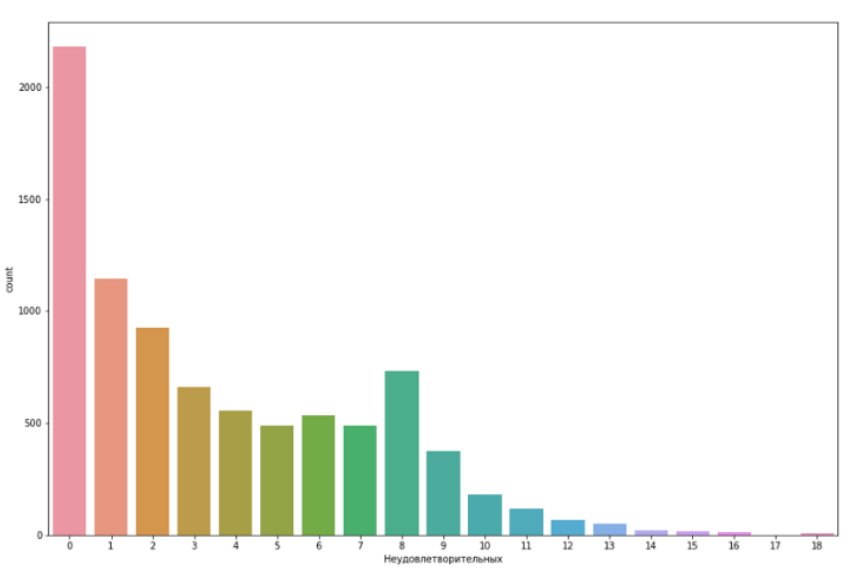



Рисунок 5. Наличие нулевых значений у параметров

Можно видеть, что атрибут «Дисциплины по которым получены неудовлетворительные оценки» сам по себе предполагает наличие «миссингов», в то время как «профиль» и «страна» скорее всего пропущенное значение в результате заполнения базы данных.

Далее будут рассмотрены данные в самых простых разрезах, для того чтобы понять начальный вектор подготовки данных.

В первую очередь рассмотрим количество (а вместе и с ним качество, которое оценим по количеству долгов) должников (Рисунок 6).




| | Неудовлетворительных |
|---|---|
| 0 | 2181 |
| 1 | 1146 |
| 2 | 926 |
| 3 | 660 |
| 4 | 557 |
| 5 | 489 |
| 6 | 535 |
| 7 | 487 |
| 8 | 733 |
| 9 | 372 |
| 10 | 179 |
| 11 | 117 |
| 12 | 66 |
| 13 | 50 |
| 14 | 20 |
| 15 | 14 |
| 16 | 11 |
| 17 | 1 |
| 18 | 7 |

Name: ID, dtype: int64

Рисунок 6. Количество неудовлетворительных оценок

Чтобы учитывать данный атрибут (количество неудовлетворительных оценок) в дальнейшем необходимо разделить данный домен значений атрибута на группы, так как разница между студентом без долгов и с одним долгом больше, чем между студентами с 18 и 17 долгами, где по сути это уже не играет принципиального значения.

Было решено распределить студентов на 3 группы в зависимости от количества долгов:

- 1 группа “успешно” – 0 долгов
- 2 группа “долги” – от 1 до 6 долгов
- 3 группа “много долгов” – от 7 и больше (максимально 18 долгов)

Для того чтобы было удобнее анализировать этот критерий, был создан дополнительный столбец в датасете, где указано к какой группе принадлежит студент.

Рассмотрим общее количество студентов в этих группах:

| группа долгов | |
|---|---|
| долги | 4313 |
| много долгов | 2057 |
| успешно | 2181 |

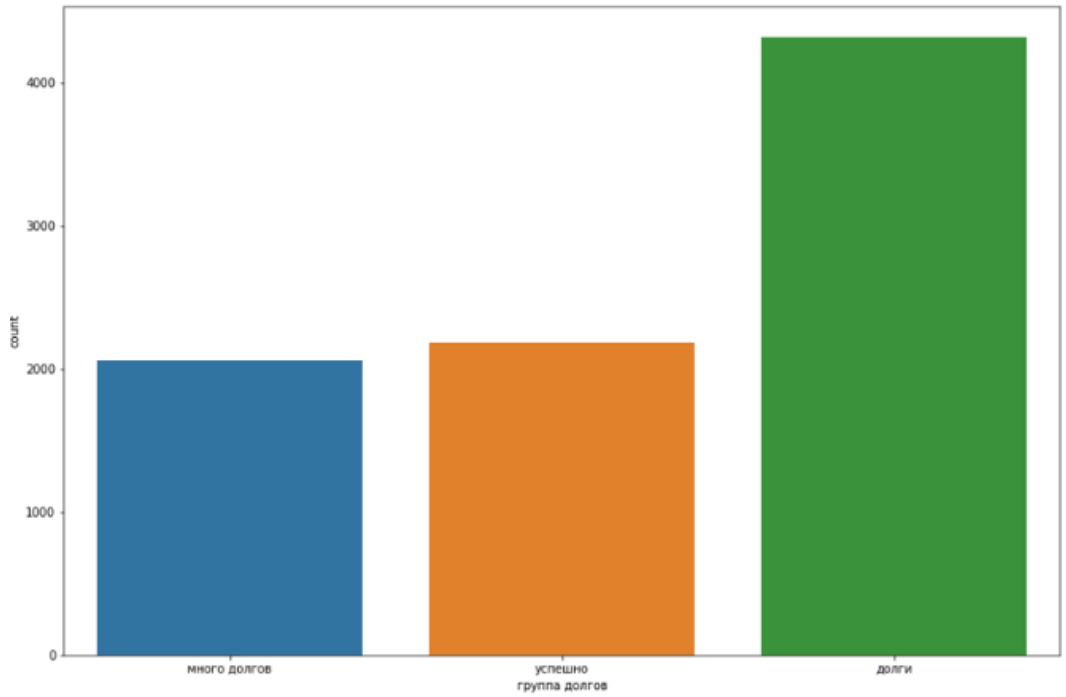



Рисунок 7. Распределение студентов по «группам долгов»

Рассмотри состав студентов по курсам (и квалификации)

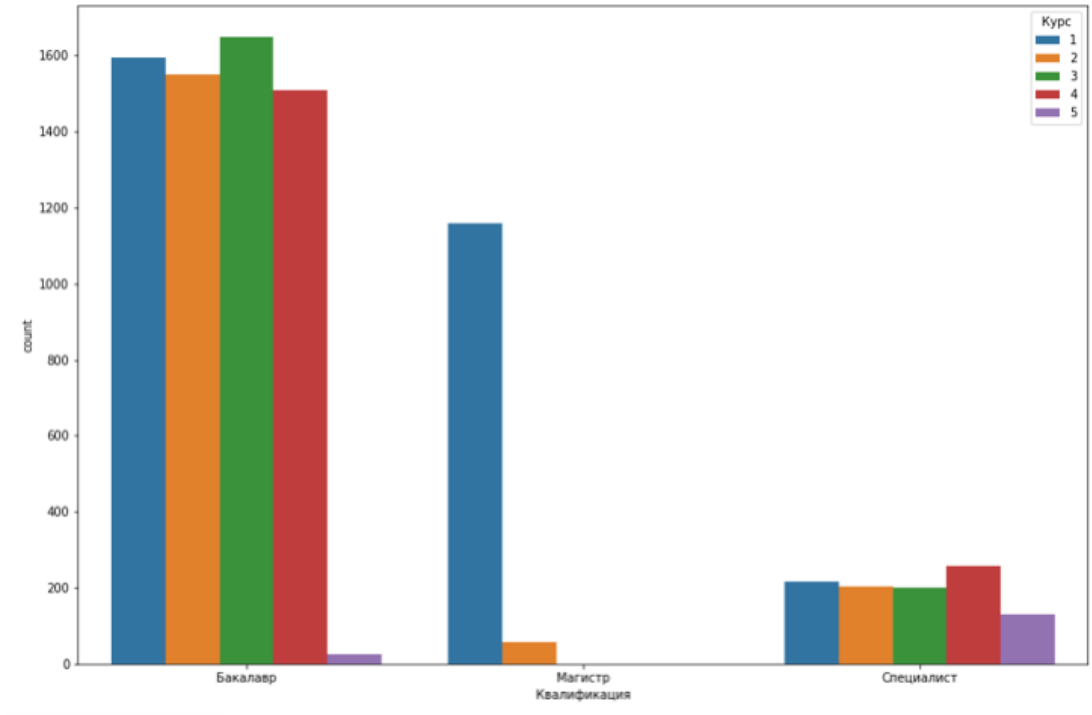



Рисунок 8. Распределение по квалификации и курсу обучения

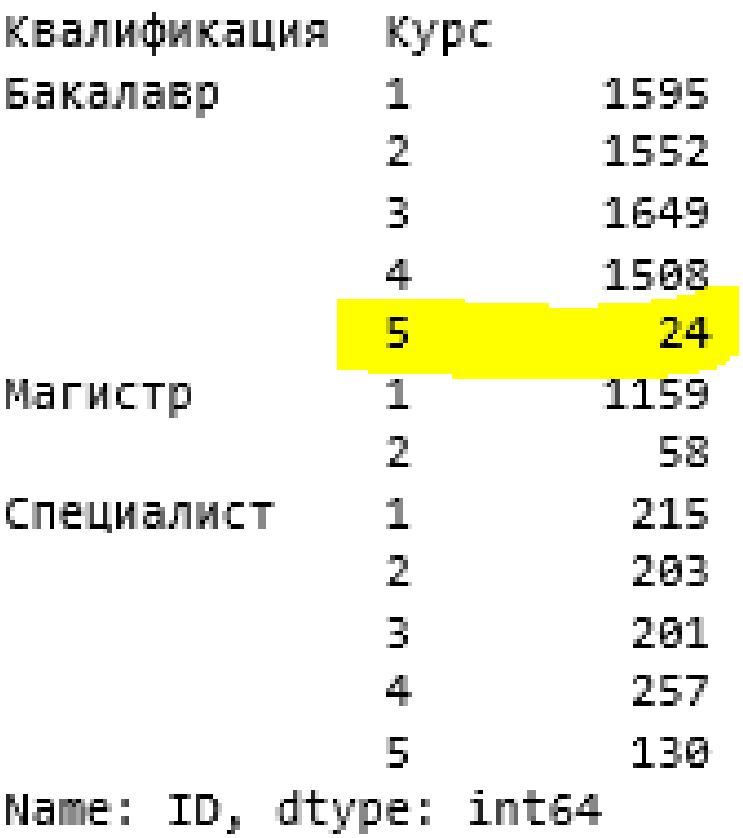



Рисунок 9. Распределение по квалификации и курсу обучения

Состав студентов примерно равномерен среди бакалавриата, однако можно видеть такую группу студентов как бакалавры 5го курса. Тут скорее всего была допущена ошибка при занесении в базу данных, и эту группу студентов следует удалить, так как она не является значимой, а относится к «выбросам».

Рассмотрим группы в зависимости от Формы обучения и квалификации.

```
Форма обучения  Квалификация
Заочная         Бакалавр        2090
                Специалист       270
Очная           Бакалавр        4238
                Магистр         1145
                Специалист       736
Очно-заочная    Магистр           72
Name: ID, dtype: int64
```

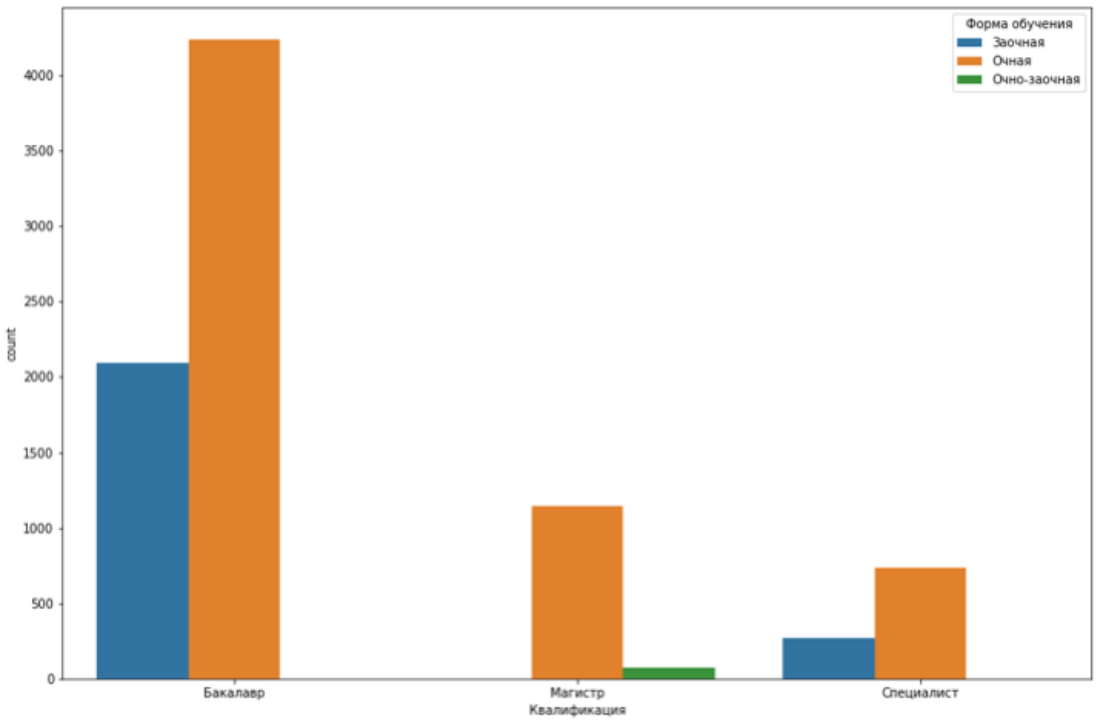



Рисунок 10. Распределение студентов по «группам долгов»

Как видно из графика, сравнить формы обучения на предмет корреляции с количеством неудовлетворительных оценок мы может только у бакалавриата и понять есть ли разницу между очной и заочной формой обучения в процентах.

Определим процентное соотношение студентов среди этих 6 групп по параметрам успешности.

| Форма обучения | Квалификация | группа долгов | |
|---|---|---|---|
| Заочная | Бакалавр | долги | 853 |
| | | много долгов | 1004 |
| | | успешно | 233 |
| | Специалист | долги | 170 |
| | | много долгов | 38 |
| | | успешно | 62 |
| Очная | Бакалавр | долги | 2117 |
| | | много долгов | 642 |
| | | успешно | 1479 |
| | Магистр | долги | 691 |
| | | много долгов | 228 |
| | | успешно | 226 |
| | Специалист | долги | 434 |
| | | много долгов | 121 |
| | | успешно | 181 |
| Очно-заочная | Магистр | долги | 48 |
| | | много долгов | 24 |

```
Диаграмма процентного соотношения студентов с долгами и без для группы: Бакалавр, Очная форма обучения
группа долгов
долги           2117
много долгов     642
успешно         1479
Name: ID, dtype: int64
```

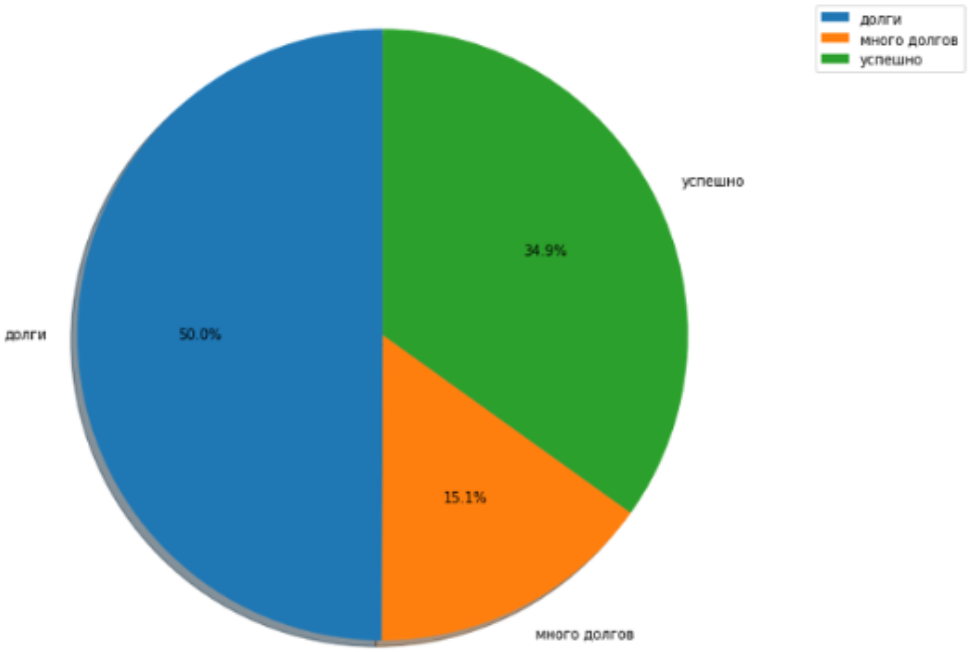



Рисунок 11. Распределение студентов по «группам долгов»

```
Диаграмма процентного соотношения студентов с долгами и без для группы: Бакалавр, Заочная форма обучения
группа долгов
долги            853
много долгов    1004
успешно          233
Name: ID, dtype: int64
```

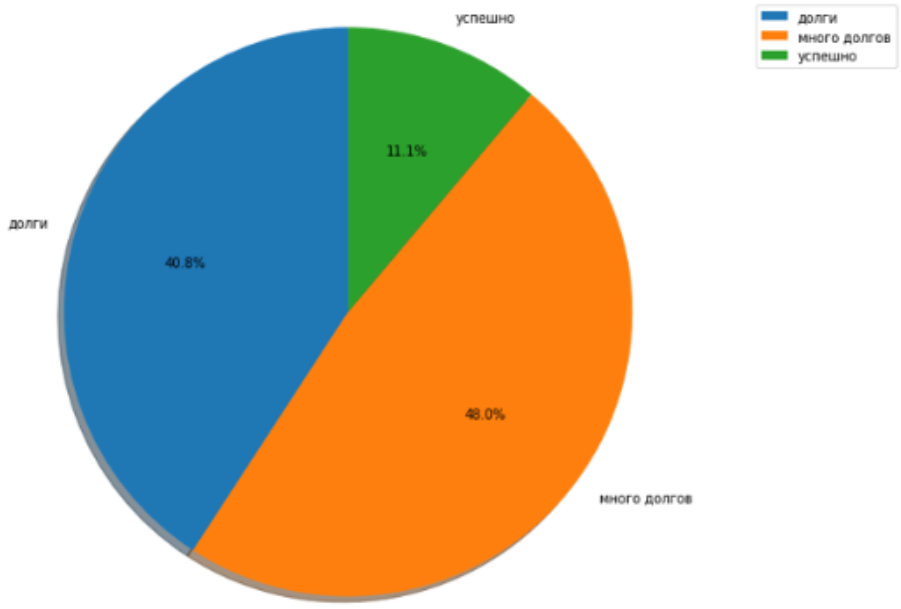



Рисунок 12. Распределение студентов по «группам долгов»

```
Диаграмма процентного соотношения студентов с долгами и без для группы: Специалист, Очная форма обучения
группа долгов
долги           434
много долгов    121
успешно         181
Name: ID, dtype: int64
```

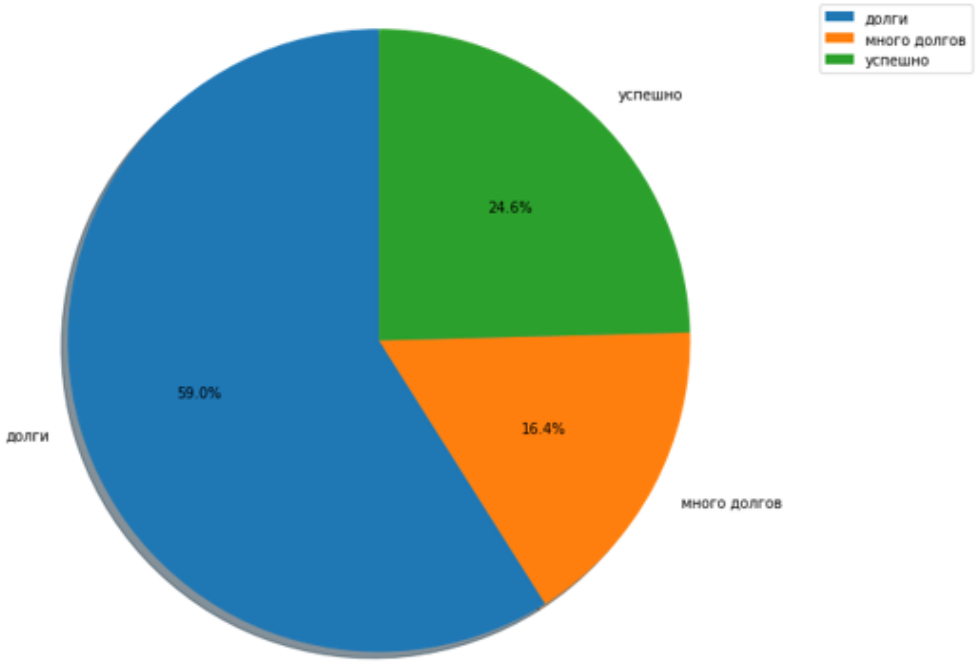



Рисунок 13. Распределение студентов по «группам долгов»

```
Диаграмма процентного соотношения студентов с долгами и без для группы: Специалист, Заочная форма обучения
группа долгов
долги           170
много долгов     38
успешно          62
Name: ID, dtype: int64
```

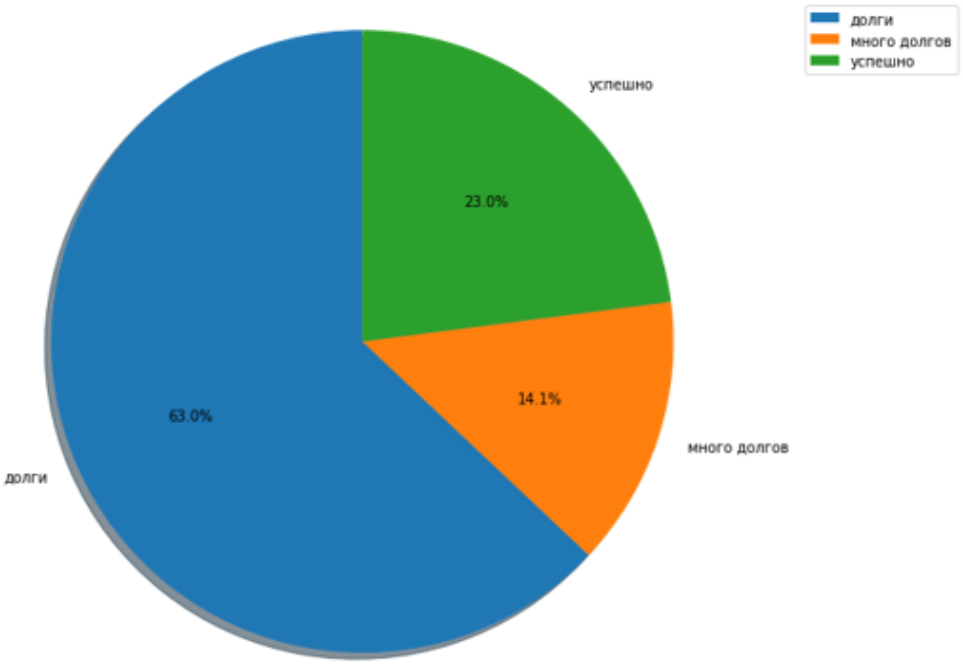



Рисунок 14. Распределение студентов по «группам долгов»

```
Диаграмма процентного соотношения студентов с долгами и без для группы: Магистр, Очная форма обучения
группа долгов
долги           691
много долгов    228
успешно         226
Name: ID, dtype: int64
```

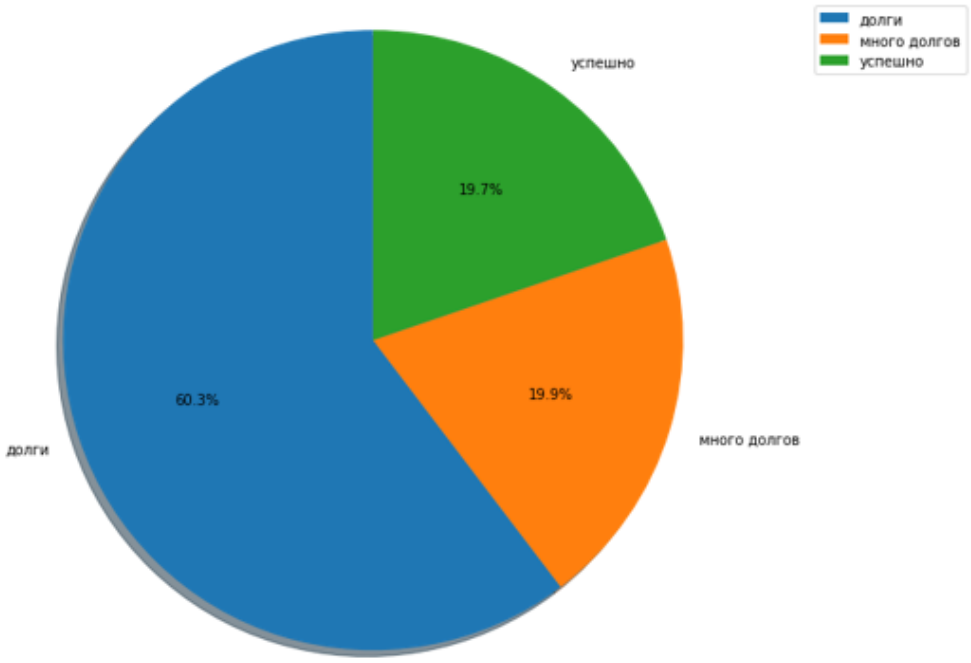



Рисунок 15. Распределение студентов по «группам долгов»

```
Диаграмма процентного соотношения студентов с долгами и без для группы: Магистр, Очно-заочная форма обучения
группа долгов
долги           48
много долгов    24
Name: ID, dtype: int64
```

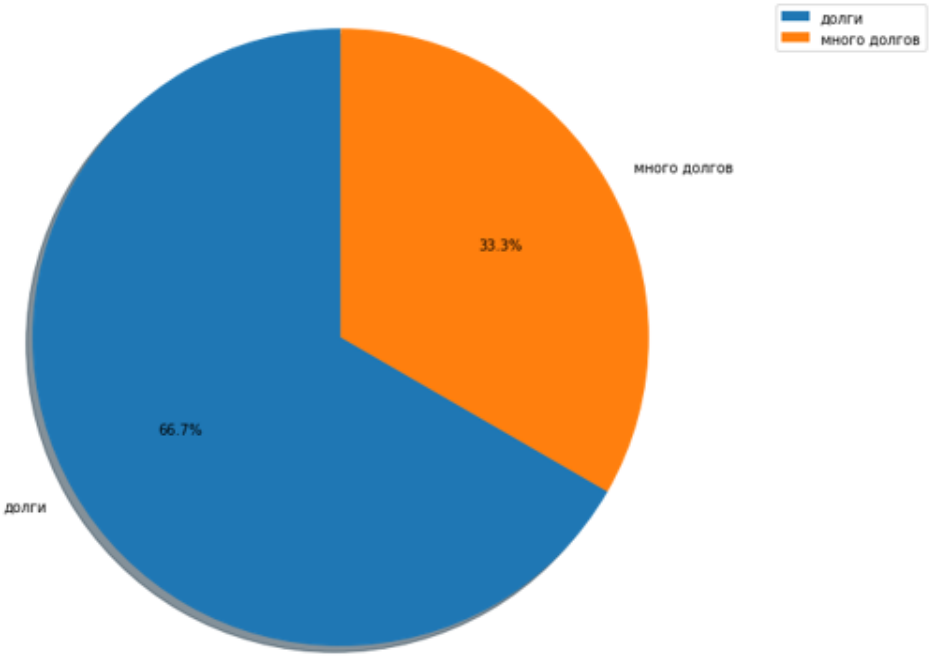



Рисунок 16. Распределение студентов по «группам долгов»

```
Пол
Женский    2501
Мужской    6050
Name: ID, dtype: int64
```

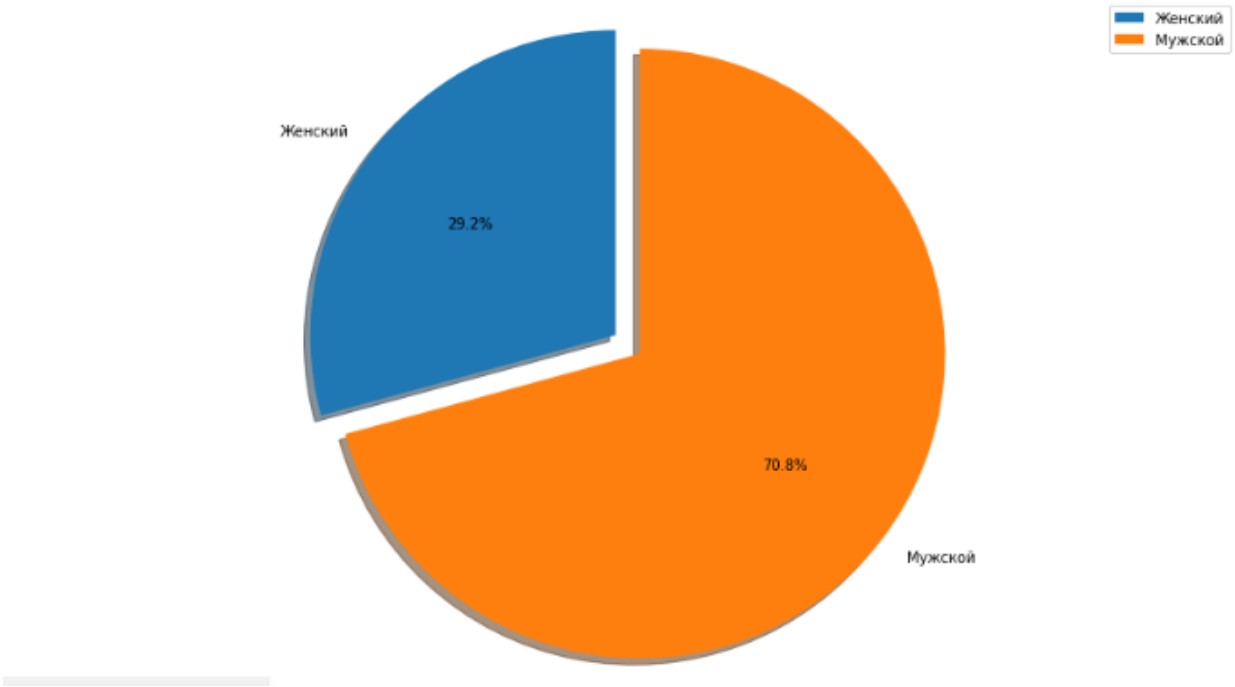



Рисунок 17. Распределение студентов по половому признаку

| Выпуск. школа | группа долгов | |
|---|---|---|
| Инженерная школа информационных технологий и робототехники | долги | 543 |
| | много долгов | 309 |
| | успешно | 266 |
| Инженерная школа неразрушающего контроля и безопасности | долги | 421 |
| | много долгов | 181 |
| | успешно | 197 |
| Инженерная школа новых производственных технологий | долги | 451 |
| | много долгов | 201 |
| | успешно | 254 |
| Инженерная школа природных ресурсов | долги | 1170 |
| | много долгов | 463 |
| | успешно | 717 |
| Инженерная школа энергетики | долги | 1029 |
| | много долгов | 408 |
| | успешно | 499 |
| Инженерная школа ядерных технологий | долги | 507 |
| | много долгов | 160 |
| | успешно | 210 |
| Школа базовой инженерной подготовки | долги | 1 |
| | много долгов | 1 |

| Курс | группа долгов | |
|---|---|---|
| 1 | долги | 1645 |
| | много долгов | 739 |
| | успешно | 585 |
| 2 | долги | 1166 |
| | много долгов | 590 |
| | успешно | 57 |
| 3 | долги | 978 |
| | много долгов | 513 |
| | успешно | 359 |
| 4 | долги | 464 |
| | много долгов | 195 |
| | успешно | 1106 |
| 5 | долги | 60 |
| | много долгов | 20 |
| | успешно | 74 |

Среднее количество неудовлетворительных оценок в выпускающей школе

| Обуч. подразд. | |
|---|---|
| Инженерная школа информационных технологий и робототехники | 3.923214 |
| Инженерная школа неразрушающего контроля и безопасности | 3.197747 |
| Инженерная школа новых производственных технологий | 3.512141 |
| Инженерная школа природных ресурсов | 3.407234 |
| Инженерная школа энергетики | 3.264463 |
| Инженерная школа ядерных технологий | 3.319270 |
| Исследовательская школа химических и биомедицинских технологий | 3.476190 |
| Школа инженерного предпринимательства | 6.214022 |

Name: Неудовлетворительных, dtype: float64

Медиана неудовлетворительных оценок в выпускающей школе

| Обуч. подразд. | |
|---|---|
| Инженерная школа информационных технологий и робототехники | 3 |
| Инженерная школа неразрушающего контроля и безопасности | 2 |
| Инженерная школа новых производственных технологий | 2 |
| Инженерная школа природных ресурсов | 3 |
| Инженерная школа энергетики | 2 |
| Инженерная школа ядерных технологий | 2 |
| Исследовательская школа химических и биомедицинских технологий | 4 |
| Школа инженерного предпринимательства | 8 |

Name: Неудовлетворительных, dtype: int64

Топ по среднему количеству неуд. Оценок на студента по разным специальностям.

| Специальность | |
|---|---|
| 38.03.01 Экономика | 7.509009 |
| 38.03.02 Менеджмент | 7.111888 |
| 38.04.02 Менеджмент | 6.222222 |
| 54.03.01 Дизайн | 5.529412 |
| 13.04.01 Теплоэнергетика и теплотехника | 5.323077 |
| 15.04.06 Мехатроника и робототехника | 5.000000 |
| 27.04.04 Управление в технических системах | 4.812500 |
| 09.04.02 Информационные системы и технологии | 4.769231 |
| 15.03.04 Автоматизация технологических процессов и производств | 4.755352 |
| 15.03.01 Машиностроение | 4.649770 |
| 12.04.02 Оптотехника | 4.588235 |
| 09.04.04 Программная инженерия | 4.583333 |
| 01.04.02 Прикладная математика и информатика | 4.541667 |
| 09.04.01 Информатика и вычислительная техника | 4.377778 |
| 54.04.01 Дизайн | 4.333333 |

Топ специальностей по медиане неуд оценок.

| Специальность | |
|---|---|
| 38.03.01 Экономика | 8.0 |
| 38.03.02 Менеджмент | 8.0 |
| 38.04.02 Менеджмент | 6.5 |
| 09.04.04 Программная инженерия | 6.0 |
| 15.04.06 Мехатроника и робототехника | 5.5 |
| 15.03.04 Автоматизация технологических процессов и производств | 5.0 |
| 09.04.02 Информационные системы и технологии | 4.0 |
| 13.03.01 Теплоэнергетика и теплотехника | 4.0 |
| 15.03.01 Машиностроение | 4.0 |
| 21.03.01 Нефтегазовое дело | 4.0 |
| 27.04.05 Инноватика | 4.0 |
| 54.03.01 Дизайн | 4.0 |
| 01.04.02 Прикладная математика и информатика | 3.5 |
| 12.04.04 Биотехнические системы и технологии | 3.5 |
| 27.04.04 Управление в технических системах | 3.5 |

Также из текущего файла были получены данные о количестве неудовлетворительных оценок по каждому предмету по всему университету.

| Название предмета | Количество должников |
|---|---|
| Учебно-исследовательская работа студентов(Зач.) | 1860 |
| Прикладная физическая культура(Зач.) | 953 |
| Иностранный язык для программ академической мобильности (английский). A2.2(Зач.) | 718 |
| Иностранный язык (английский)(Экз.) | 685 |
| Профессиональная подготовка на английском языке(Зач.) | 668 |
| Иностранный язык (английский)(Зач.) | 643 |
| Творческий проект(Зач.) | 591 |
| Второй иностранный язык (немецкий). A1.1(Зач.) | 560 |
| Профессиональный иностранный язык (английский)(Зач.) | 524 |
| Физика 1(ДЗ) | 515 |
| Математика 2(ДЗ) | 488 |
| Метрология | 436 |
| Научно-исследовательская работа в семестре(Зач.) | 423 |
| Управление проектами(Зач.) | 395 |
| стандартизация и сертификация 1.1(Зач.) | 388 |

Как видно из таблицы, предметы не очень показательны, так как, например, предмет УИРС встречается во всех школах и специальностях, поэтому логично что по нему больше всего долгов.

Количество должников по предметам по специальности «38.03.01 Экономика»:

| Название предмета | Количество должников |
|---|---|
| Учебно-исследовательская работа студентов(Зач.) | 211 |
| Экономика и нормирование труда(Зач.) | 62 |
| Прикладная физическая культура(Зач.) | 58 |
| Теория бухгалтерского учета(Экз.) | 57 |
| Теория бухгалтерского учета(КР) | 57 |
| Математика 2.4(Экз.) | 57 |
| Этика деловых отношений(Зач.) | 56 |
| Эконометрика(Зач.) | 56 |
| Иностранный язык (английский)(Экз.) | 56 |
| Корпоративные финансы(Экз.) | 51 |
| Корпоративные финансы(КР) | 51 |
| Профессиональный иностранный язык (английский)(Зач.) | 49 |
| Международные стандарты учета и финансовой отчетности (Экз.) | 49 |
| банки(Экз.) | 49 |
| кредит | 49 |

## 3. Модель машинного обучения

Таким образом, после анализа всех параметров, были выявлены следующие, которые могут в большей степени влиять на успеваемость студента, а именно:

1. Форма обучения
2. Квалификация
3. Курс
4. Специальность
5. Академ отпуск (действующий) - да / нет
6. Всего часов пропусков в семестре
7. Всего часов аудиторных занятий в семестре

Данные параметры будут использованы в машинном обучении с учителем с целью классификации студента на предмет количества задолженностей.

Рассмотрим корреляции между выбранными параметрами.

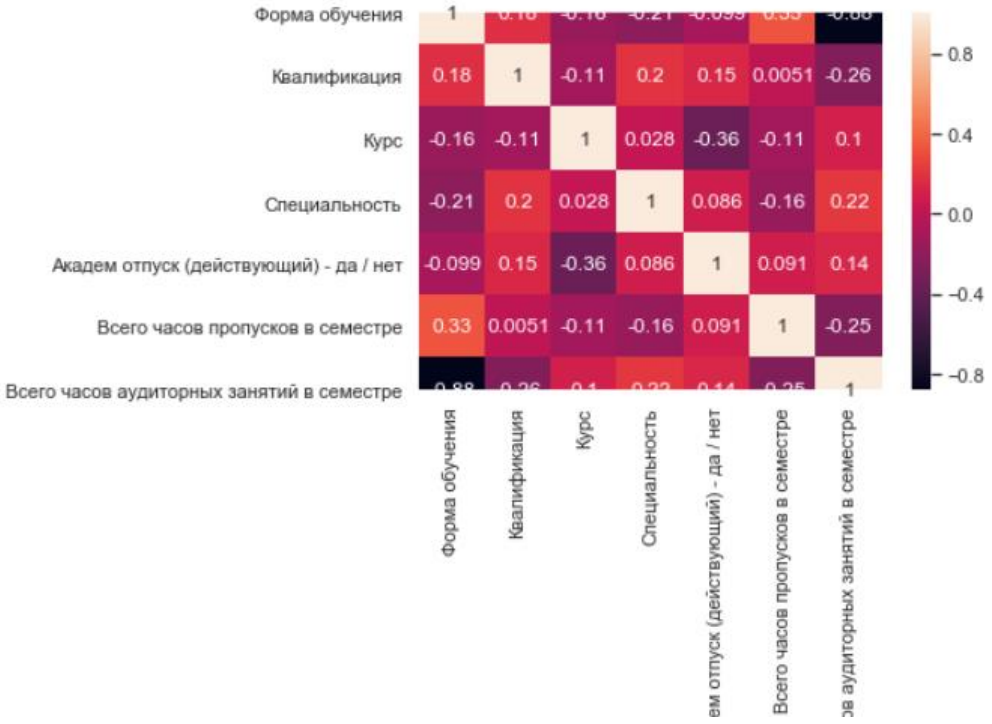



Рисунок 18. Корреляция признаков

Для того чтобы приступить к обучению модели, нам необходимо преобразовать все параметры в числовые представления. Для этого будет использована уже имеющаяся в Python функция LabelEncoder из библиотеки sklearn.preprocessing.

```
X = students_df[['Форма обучения', 'Квалификация', 'Курс', 'Специальность',
        'Академ отпуск (действующий) - да / нет', 'Всего часов пропусков в семестре',
       'Всего часов аудиторных занятий в семестре']]
Y = students_df['группа долгов']
```

```
from sklearn.preprocessing import LabelEncoder
le_f = LabelEncoder()
X['Форма обучения'] = le_f.fit_transform(X['Форма обучения'].values)
le_k = LabelEncoder()
X['Квалификация'] = le_k.fit_transform(X['Квалификация'].values)
le_s = LabelEncoder()
X['Специальность'] = le_s.fit_transform(X['Специальность'].values)
le_a = LabelEncoder()
X['Академ отпуск (действующий) - да / нет'] = le_a.fit_transform(X['Академ отпуск (действующий) - да / нет'].values)
```

После преобразований мы получаем следующее: X – это датафрейм всех параметров, уже преобразованный в числовое значение, Y- это метки.

```
X.head()
```

| | Форма обучения | Квалификация | Курс | Специальность | Академ отпуск (действующий) - да / нет | Всего часов пропусков в семестре | Всего часов аудиторных занятий в семестре |
|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 4 | 64 | 1 | 0 | 1400.0 |
| 1 | 1 | 0 | 4 | 16 | 0 | 16 | 440.0 |
| 2 | 1 | 0 | 1 | 23 | 1 | 364 | 464.0 |
| 3 | 1 | 0 | 1 | 47 | 1 | 2 | 464.0 |
| 4 | 1 | 1 | 1 | 27 | 0 | 0 | 384.0 |

Для того чтобы приступить к созданию модели поделим наши данные на обучающую и тестовую выборки.

```
from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test = train_test_split(X, Y, test_size=0.3, random_state=0)
```

Первый классификатор, который будет использован это логистическая регрессия.

```
from sklearn.linear_model import LogisticRegression
lr = LogisticRegression(C=1000.0, random_state=0 )
lr.fit(x_train,y_train)
```

```
LogisticRegression(C=1000.0, class_weight=None, dual=False, fit_intercept=True,
                   intercept_scaling=1, l1_ratio=None, max_iter=100,
                   multi_class='warn', n_jobs=None, penalty='l2',
                   random_state=0, solver='warn', tol=0.0001, verbose=0,
                   warm_start=False)
```

```
print(lr.score(x_train, y_train))
print(lr.score(x_test, y_test))
```

```
0.6668915500084331
0.6336088154269972
```

```
pred_y = lr.predict(x_test)
from sklearn import metrics
# Model Accuracy, how often is the classifier correct?
print(metrics.classification_report(pred_y, y_test))
```

```
              precision    recall  f1-score   support

       долги       0.84      0.60      0.70      1732
много долгов       0.48      0.74      0.58       433
     успешно       0.40      0.69      0.51       376

    accuracy                           0.63      2541
   macro avg       0.57      0.67      0.59      2541
weighted avg       0.71      0.63      0.65      2541
```

Данные метод показал, что определенная зависимость есть и параметры выбраны верно. Далее мы попробуем другие виды классификации и сравним результаты.

Применим метод опорных векторов.

```
from sklearn import svm
clf = svm.SVC()
clf.fit(x_train,y_train)
```

```
SVC(C=1.0, cache_size=200, class_weight=None, coef0=0.0,
    decision_function_shape='ovr', degree=3, gamma='auto_deprecated',
    kernel='rbf', max_iter=-1, probability=False, random_state=None,
    shrinking=True, tol=0.001, verbose=False)
```

```
print(clf.score(x_train, y_train))
print(clf.score(x_test, y_test))
```

```
0.8794063079777366
0.7197953561589925
```

```
pred_y = clf.predict(x_test)
from sklearn import metrics
# Model Accuracy, how often is the classifier correct?
print(metrics.classification_report(pred_y, y_test))
```

```
              precision    recall  f1-score   support

       долги       0.86      0.68      0.76      1564
много долгов       0.54      0.85      0.66       421
     успешно       0.64      0.74      0.69       556

    accuracy                           0.72      2541
   macro avg       0.68      0.76      0.70      2541
weighted avg       0.76      0.72      0.73      2541
```

Здесь мы можем видеть, что метод опорных векторов предрасположен к переобучению, так как достаточно большой разброс в 10% между тренировочным набором и тестовым в определении меток. Тем не менее результат на тестовой выборке почти на 10 процентов выше, чем у классификатора на основе логистической регрессии.

Применим 3ий классификатор “случайный лес”

```
from sklearn.ensemble import RandomForestClassifier
clas = RandomForestClassifier(max_depth=15, random_state=0)
clas.fit(x_train,y_train)
```

```
RandomForestClassifier(bootstrap=True, class_weight=None, criterion='gini',
                       max_depth=15, max_features='auto', max_leaf_nodes=None,
                       min_impurity_decrease=0.0, min_impurity_split=None,
                       min_samples_leaf=1, min_samples_split=2,
                       min_weight_fraction_leaf=0.0, n_estimators=10,
                       n_jobs=None, oob_score=False, random_state=0, verbose=0,
                       warm_start=False)
```

```
print(clas.feature_importances_)
```

```
[0.01858244 0.03035815 0.12210839 0.2179611  0.07945399 0.31382953
 0.21770641]
```

```
print(clas.score(x_train, y_train))
print(clas.score(x_test, y_test))
```

```
0.8775510204081632
0.7898465171192444
```

```
pred_y = clas.predict(x_test)
from sklearn import metrics
# Model Accuracy, how often is the classifier correct?
print(metrics.classification_report(pred_y, y_test))
```

```
              precision    recall  f1-score   support

       долги       0.82      0.78      0.80      1311
много долгов       0.74      0.85      0.79       579
     успешно       0.77      0.76      0.77       651

    accuracy                           0.79      2541
   macro avg       0.78      0.80      0.79      2541
weighted avg       0.79      0.79      0.79      2541
```

Случайный лес показал наилучший результат, хотя разброс между обучающей и тестовой выборкой достаточно велик (10%), точность прогноза тестовой выборки составляет 79 процентов. Это достаточно высокий показатель точности. Так же были опробован случайный лес с различной максимальной глубиной и в результате опыта было выявлено, что глубина больше 15, не дает существенного результата влияющего на точность определения тестовой выборки.

Таким образом можно сделать вывод, что в первую очередь задача по определению успеваемости студента на основании таких параметров как:

- Форма обучения

- Квалификация
- Курс
- Специальность
- Академ отпуск (действующий) - да / нет
- Всего часов пропусков в семестре
- Всего часов аудиторных занятий в семестре

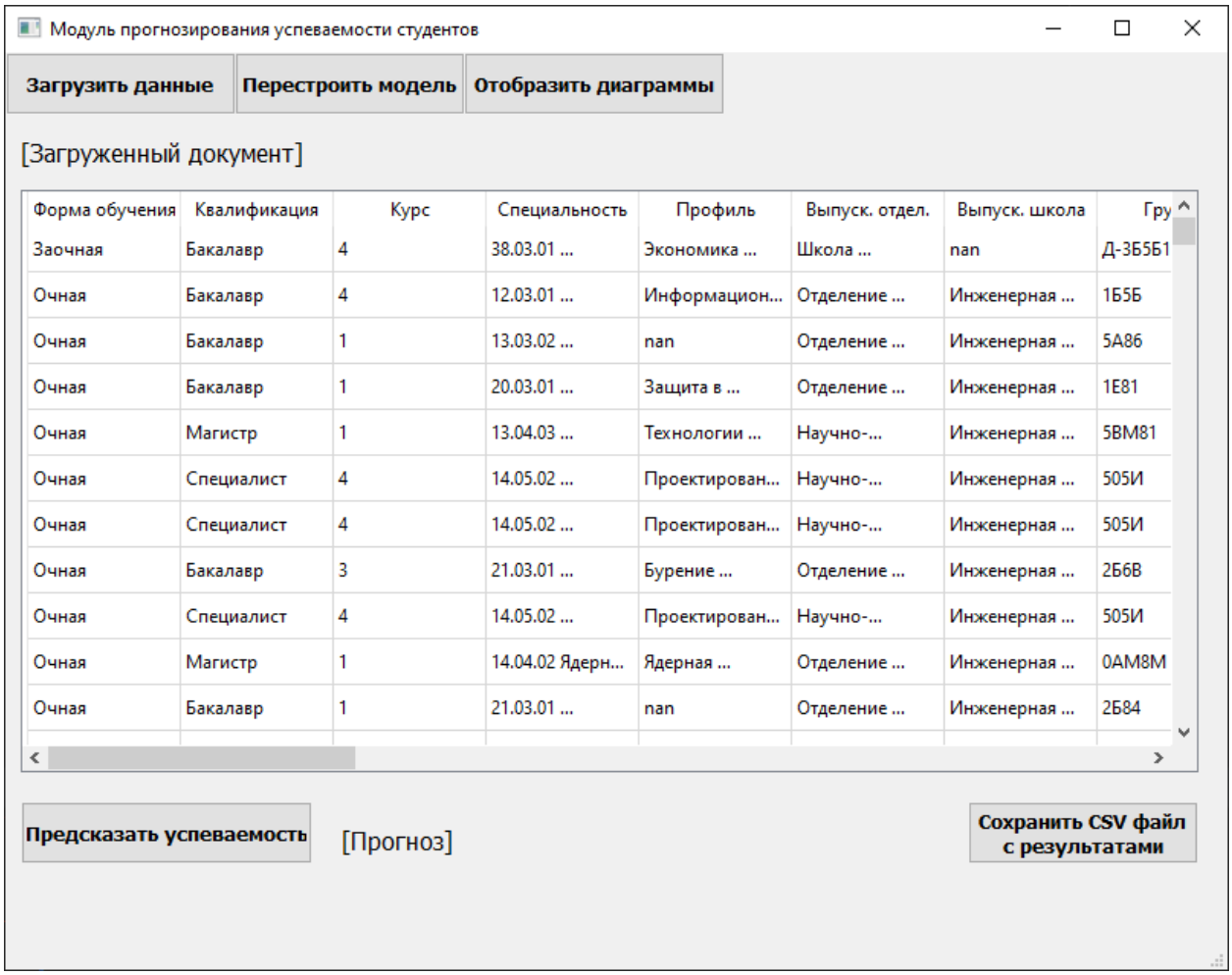
Также в результате анализа было выявлено, что наиболее значимый вклад в модель вносят 3 параметра, а именно:

- Всего часов пропусков в семестре
- Всего часов аудиторных занятий в семестре
- Курс

В дальнейшем точность модели также могут повысить дополнительные параметры, например, хобби студента, участие в социальной жизни университета и тд.

**4. Графический интерфейс пользователя прогнозной модели.**

В ходе преддипломной практики был создан графический интерфейс для будущего модуля прогнозирования в системе, который позволяет загружать информация о студентах в формате файла excel и отображать его на экране. Встроенная модель прогнозирование позволяет сделать оценку количества академических задолженностей на конец семестра для конкретного студента или для всех сразу и вывести ее на экран. Также модуль позволяет добавить данные к уже существующей модели, для того чтобы перестроить ее и повысить ее прогнозную точность. На рисунке 19 изображен главное меню модуля.




| Форма обучения | Квалификация | Курс | Специальность | Профиль | Выпуск. отдел. | Выпуск. школа | Гру |
|---|---|---|---|---|---|---|---|
| Заочная | Бакалавр | 4 | 38.03.01 ... | Экономика ... | Школа ... | nan | Д-3Б5Б1 |
| Очная | Бакалавр | 4 | 12.03.01 ... | Информацион... | Отделение ... | Инженерная ... | 1Б5Б |
| Очная | Бакалавр | 1 | 13.03.02 ... | nan | Отделение ... | Инженерная ... | 5А86 |
| Очная | Бакалавр | 1 | 20.03.01 ... | Защита в ... | Отделение ... | Инженерная ... | 1Е81 |
| Очная | Магистр | 1 | 13.04.03 ... | Технологии ... | Научно-... | Инженерная ... | 5ВМ81 |
| Очная | Специалист | 4 | 14.05.02 ... | Проектирован... | Научно-... | Инженерная ... | 505И |
| Очная | Специалист | 4 | 14.05.02 ... | Проектирован... | Научно-... | Инженерная ... | 505И |
| Очная | Бакалавр | 3 | 21.03.01 ... | Бурение ... | Отделение ... | Инженерная ... | 2Б6В |
| Очная | Специалист | 4 | 14.05.02 ... | Проектирован... | Научно-... | Инженерная ... | 505И |
| Очная | Магистр | 1 | 14.04.02 Ядерн... | Ядерная ... | Отделение ... | Инженерная ... | 0АМ8М |
| Очная | Бакалавр | 1 | 21.03.01 ... | nan | Отделение ... | Инженерная ... | 2Б84 |

Рисунок 19. Главное меню модуля

После нажатия кнопки загрузить файл, открывается проводник и через него следует выбрать файл, который необходимо добавить (рисунок 20)

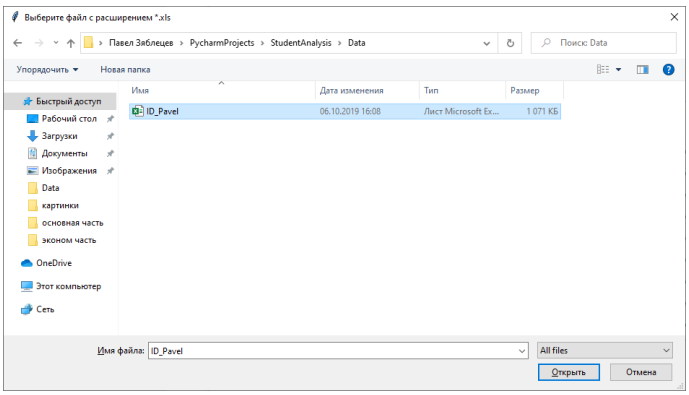
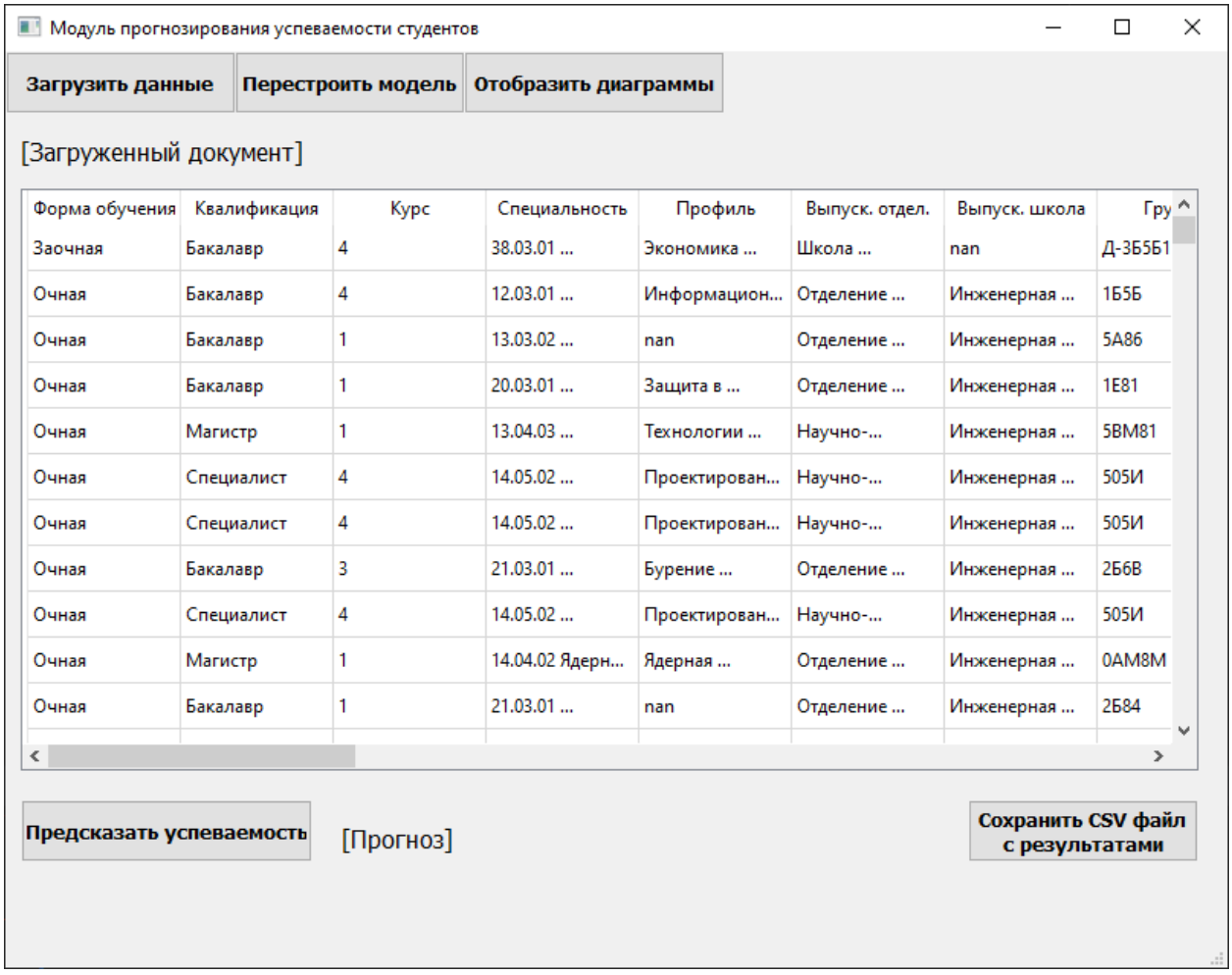



Рисунок 20. Окно выбора файла для загрузки

После этого расположение и название файла отобразится наверху (рисунок 21)




| Форма обучения | Квалификация | Курс | Специальность | Профиль | Выпуск. отдел. | Выпуск. школа | Гру |
|---|---|---|---|---|---|---|---|
| Заочная | Бакалавр | 4 | 38.03.01 ... | Экономика ... | Школа ... | nan | Д-3Б5Б1 |
| Очная | Бакалавр | 4 | 12.03.01 ... | Информацион... | Отделение ... | Инженерная ... | 1Б5Б |
| Очная | Бакалавр | 1 | 13.03.02 ... | nan | Отделение ... | Инженерная ... | 5А86 |
| Очная | Бакалавр | 1 | 20.03.01 ... | Защита в ... | Отделение ... | Инженерная ... | 1Е81 |
| Очная | Магистр | 1 | 13.04.03 ... | Технологии ... | Научно-... | Инженерная ... | 5ВМ81 |
| Очная | Специалист | 4 | 14.05.02 ... | Проектирован... | Научно-... | Инженерная ... | 505И |
| Очная | Специалист | 4 | 14.05.02 ... | Проектирован... | Научно-... | Инженерная ... | 505И |
| Очная | Бакалавр | 3 | 21.03.01 ... | Бурение ... | Отделение ... | Инженерная ... | 2Б6В |
| Очная | Специалист | 4 | 14.05.02 ... | Проектирован... | Научно-... | Инженерная ... | 505И |
| Очная | Магистр | 1 | 14.04.02 Ядерн... | Ядерная ... | Отделение ... | Инженерная ... | 0АМ8М |
| Очная | Бакалавр | 1 | 21.03.01 ... | nan | Отделение ... | Инженерная ... | 2Б84 |

Рисунок 21. Главное меню модуля

После нажатия кнопки построить модель, модель машинного обучения перестраивается в соответствие с добавленными данными (рисунок 22).

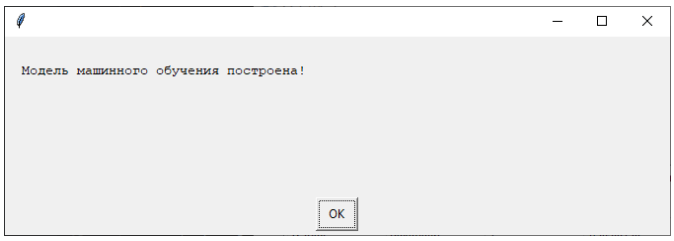



Рисунок 22. Сообщение об успешном построении модели

Также в окне основного меню можно видеть кнопку «Работа с графиками». Нажав на нее, откроется дополнительное окно. В этом окне существует возможность выбрать тип графика и данные, по которым мы хотим его построить (рисунок 23).

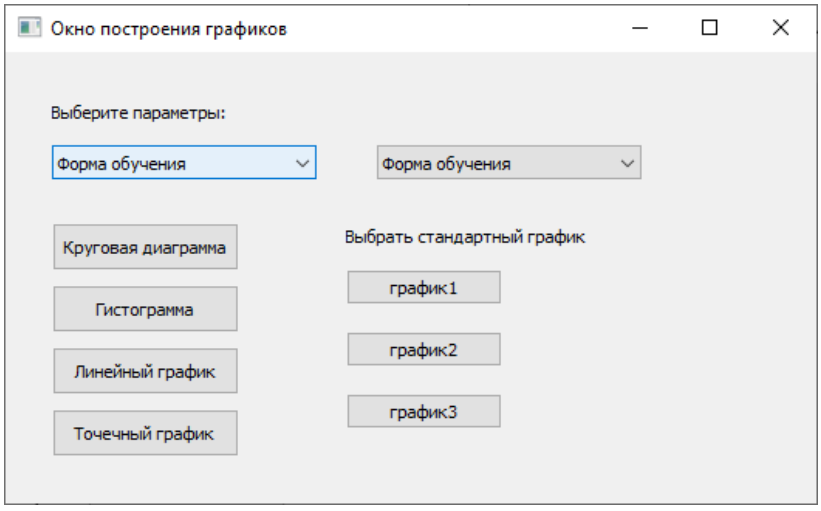



Рисунок 23. Окно построения графиков

Параметры для построения графиков можно выбрать из выпадающего списка (рисунок 24).

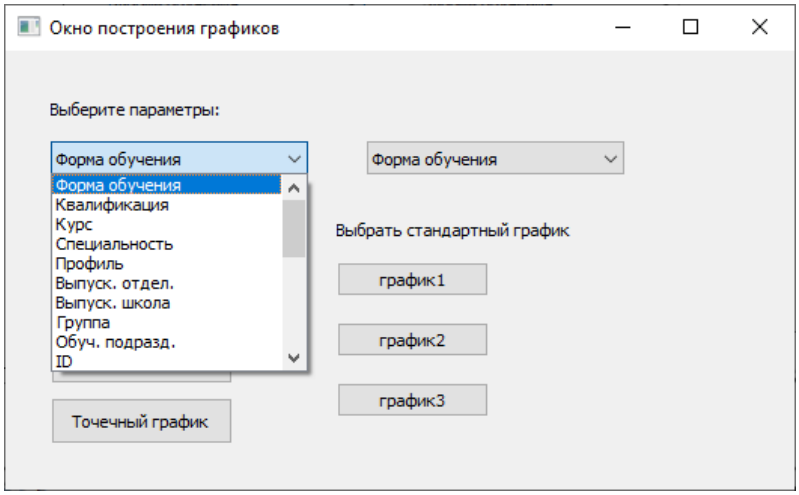



Рисунок 24. Окно построения графиков

Например, в первом случае, была выбрана круговая диаграмма с данными о форме обучения. Во втором случае был выбрана гистограмма и качестве данных это курс. (рисунок 25)

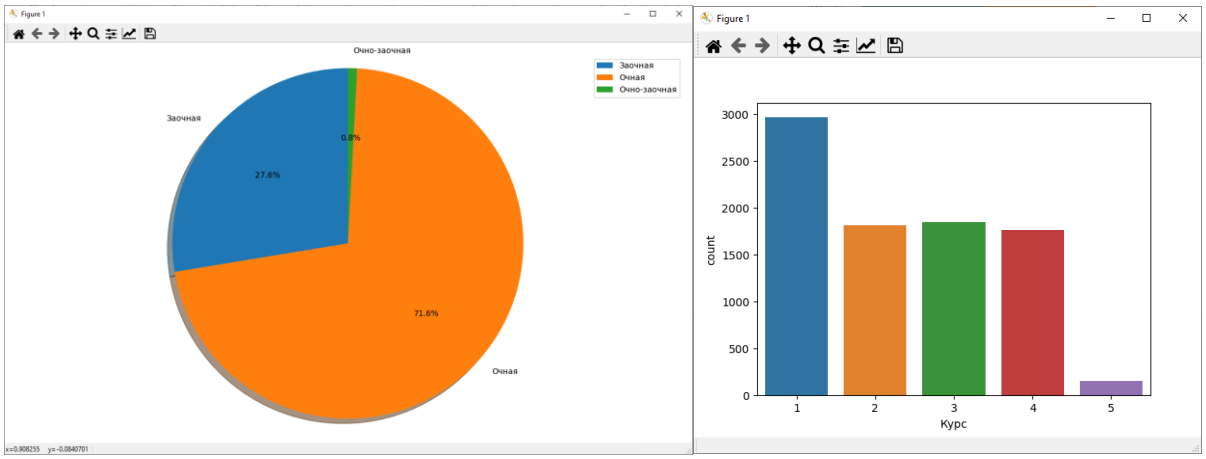



Рисунок 25. Примеры графиков

Так же существует опция построения более сложных графиков, которые есть по умолчанию. Например, первый график – это гистограмма распределения студентов по квалификации и курсу. (рисунок 26)

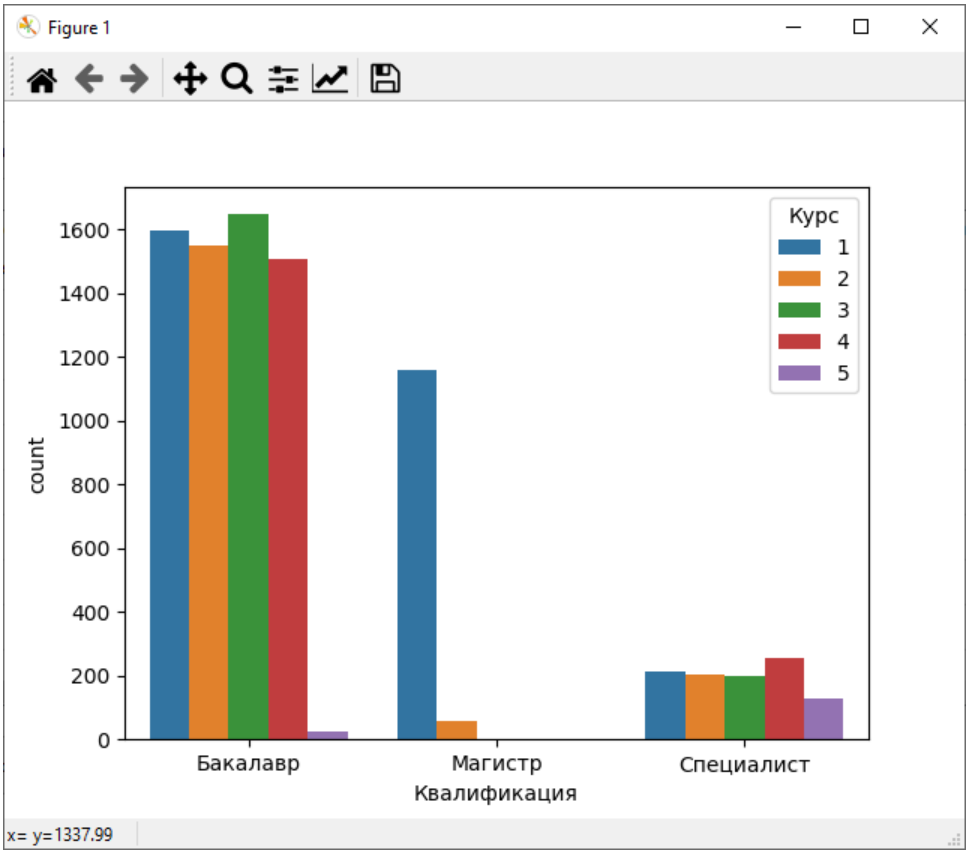
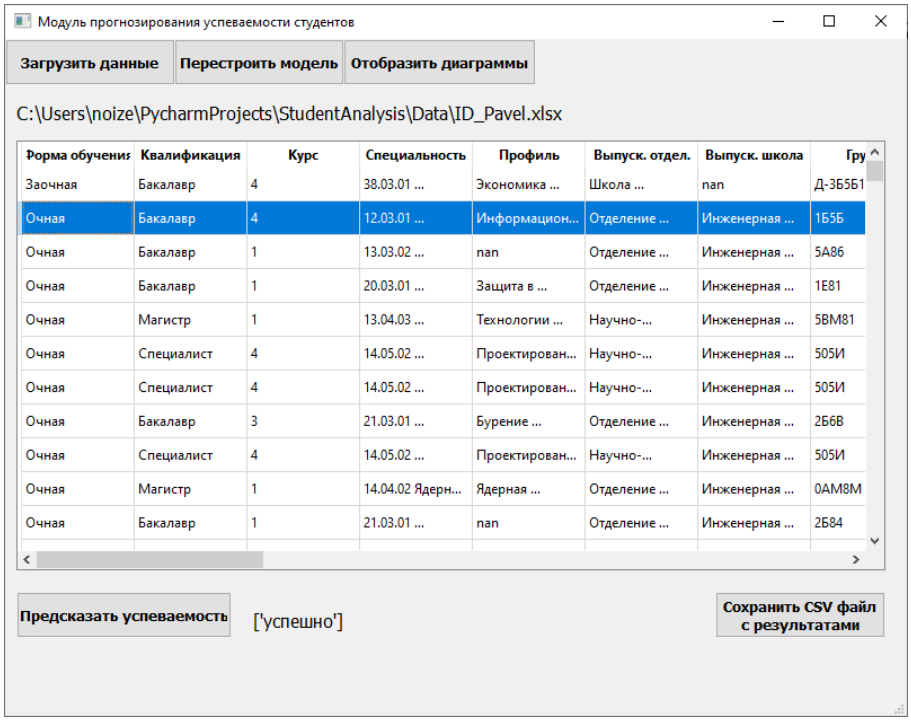



Рисунок 26. Примеры графиков

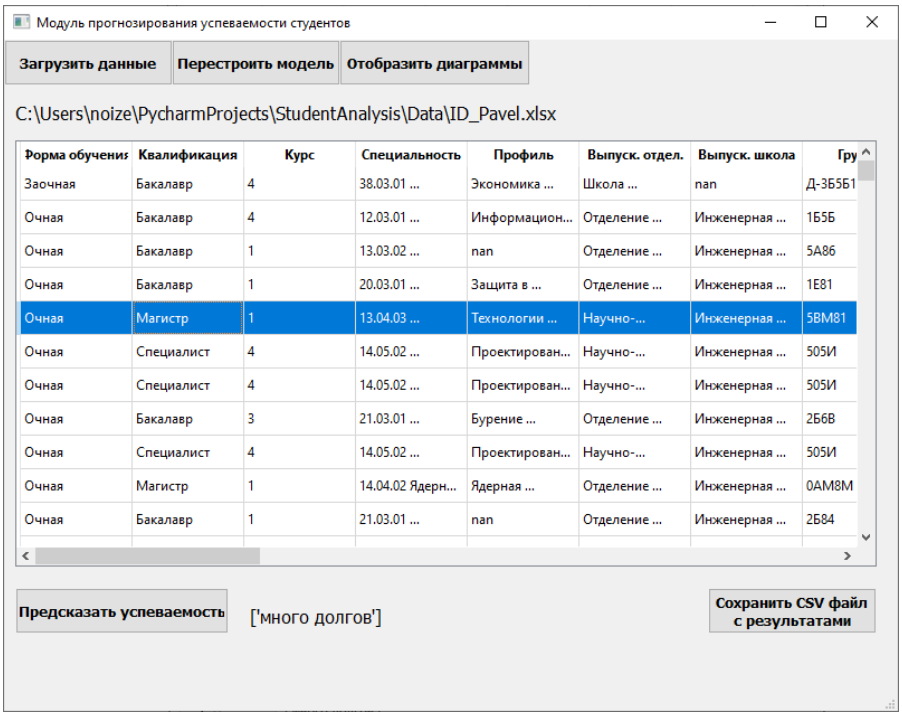
Данная возможность построения графиков позволяет произвести первичную аналитику данных и изучить некоторые тренды.

Так как основная функция данной систему это прогноз, то мы можем выбрать конкретного студента и посмотреть предсказание для него.




| Форма обучения | Квалификация | Курс | Специальность | Профиль | Выпуск. отдел. | Выпуск. школа | Гру |
|---|---|---|---|---|---|---|---|
| Заочная | Бакалавр | 4 | 38.03.01 ... | Экономика ... | Школа ... | nan | Д-3Б5Б1 |
| Очная | Бакалавр | 4 | 12.03.01 ... | Информацион... | Отделение ... | Инженерная ... | 1Б5Б |
| Очная | Бакалавр | 1 | 13.03.02 ... | nan | Отделение ... | Инженерная ... | 5А86 |
| Очная | Бакалавр | 1 | 20.03.01 ... | Защита в ... | Отделение ... | Инженерная ... | 1Е81 |
| Очная | Магистр | 1 | 13.04.03 ... | Технологии ... | Научно-... | Инженерная ... | 5ВМ81 |
| Очная | Специалист | 4 | 14.05.02 ... | Проектирован... | Научно-... | Инженерная ... | 505И |
| Очная | Специалист | 4 | 14.05.02 ... | Проектирован... | Научно-... | Инженерная ... | 505И |
| Очная | Бакалавр | 3 | 21.03.01 ... | Бурение ... | Отделение ... | Инженерная ... | 2Б6В |
| Очная | Специалист | 4 | 14.05.02 ... | Проектирован... | Научно-... | Инженерная ... | 505И |
| Очная | Магистр | 1 | 14.04.02 Ядерн... | Ядерная ... | Отделение ... | Инженерная ... | 0АМ8М |
| Очная | Бакалавр | 1 | 21.03.01 ... | nan | Отделение ... | Инженерная ... | 2Б84 |

Рисунок 27. Главное меню модуля




| Форма обучения | Квалификация | Курс | Специальность | Профиль | Выпуск. отдел. | Выпуск. школа | Гру |
|---|---|---|---|---|---|---|---|
| Заочная | Бакалавр | 4 | 38.03.01 ... | Экономика ... | Школа ... | nan | Д-3Б5Б1 |
| Очная | Бакалавр | 4 | 12.03.01 ... | Информацион... | Отделение ... | Инженерная ... | 1Б5Б |
| Очная | Бакалавр | 1 | 13.03.02 ... | nan | Отделение ... | Инженерная ... | 5А86 |
| Очная | Бакалавр | 1 | 20.03.01 ... | Защита в ... | Отделение ... | Инженерная ... | 1Е81 |
| Очная | Магистр | 1 | 13.04.03 ... | Технологии ... | Научно-... | Инженерная ... | 5ВМ81 |
| Очная | Специалист | 4 | 14.05.02 ... | Проектирован... | Научно-... | Инженерная ... | 505И |
| Очная | Специалист | 4 | 14.05.02 ... | Проектирован... | Научно-... | Инженерная ... | 505И |
| Очная | Бакалавр | 3 | 21.03.01 ... | Бурение ... | Отделение ... | Инженерная ... | 2Б6В |
| Очная | Специалист | 4 | 14.05.02 ... | Проектирован... | Научно-... | Инженерная ... | 505И |
| Очная | Магистр | 1 | 14.04.02 Ядерн... | Ядерная ... | Отделение ... | Инженерная ... | 0АМ8М |
| Очная | Бакалавр | 1 | 21.03.01 ... | nan | Отделение ... | Инженерная ... | 2Б84 |

Рисунок 28. Главное меню модуля

# 5. Финансовый менеджмент, ресурсоэффективность и ресурсосбережение

## 5.1 Предпроектный анализ

### 5.1.1 Потенциальные потребители разработки

Диссертация посвящена разработке прогнозной модели для оценки успеваемости студентов университета по итогам текущего обучения. Возможность превентивной оценки успеваемости студента на конец семестра по его текущей успеваемости позволит применять корректирующие меры до того, как у студента появились серьезные проблемы в виде академических задолженностей. На сегодняшний день единый деканат ТПУ начинает работать с неуспевающими студентами, только после того как семестр закончился, и выяснилось что студент имеет задолженности по каким-то предметам. Прогнозная модель позволит добавить новый модуль в электронную систему университета, который будет давать рекомендации о работе с потенциальными студентами-должниками, что позволит принимать превентивные меры по устранению академических задолженностей.

Потенциальным потребителем, в первую очередь, является Национальный исследовательский Томский политехнический университет.

Актуальность данного раздела заключается в важности понимания коммерческой составляющей научно-технических проектов и оценки коммерческой ценности разработки.

Целью данного раздела является анализ совокупности факторов, которые определяют коммерческую привлекательность разработки, ее перспективность и успешность.

Основными задачами является оценка перспективности разработки, ее готовности к коммерциализации, выявление потенциальных угроз, а также расчет стоимости и составление графика проведения работ.

### 5.1.2 Технология QuaD

Технология QuaD позволяет оценить перспективность разработки на рынке и целесообразность вложения средств в научно-исследовательский

проект. Результаты оценки, проведенной в табличной форме, представлены в таблице 1.

Таблица 1 – QuaD-анализ разработки

| Критерии оценки | Вес критерия | Средний балл | Макси-мальный балл | Относительное значение (3/4) | Средневзвешенное значение (5х2) |
|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 |
| Производительность | 0,1 | 90 | 100 | 0,9 | 0,09 |
| Отказоустойчивость | 0,15 | 90 | 100 | 0,9 | 0,135 |
| Унифицированность | 0,05 | 80 | 100 | 0,8 | 0,035 |
| Безопасность | 0,05 | 70 | 100 | 0,7 | 0,035 |
| Потребность в ресурсах памяти | 0,1 | 95 | 100 | 0,95 | 0,095 |
| Функциональная мощность | 0,1 | 60 | 100 | 0,6 | 0,06 |
| Простота эксплуатации | 0,15 | 90 | 100 | 0,9 | 0,135 |
| Масштабируемость | 0,05 | 60 | 100 | 0,6 | 0,03 |
| Конкурентоспособность продукта | 0,05 | 50 | 100 | 0,5 | 0,025 |
| Перспективность рынка | 0,05 | 60 | 100 | 0,6 | 0,03 |
| Цена | 0,1 | 60 | 100 | 0,6 | 0,06 |
| Финансовая эффективность научной разработки | 0,05 | 50 | 100 | 0,5 | 0,025 |
| Итого | 1 | | | | **0,76** |

По результатам оценки качества и перспективности можно утверждать, что перспективность текущей разработки выше среднего. Улучшить данную разработку можно путем повышения конкурентоспособности продукта и снижением цены. Финансовую эффективность научной разработки повысить не представляется возможным на данном этапе, так как это долгосрочное вложение в качество образования и ждать быстрой окупаемости или финансовой выгоды не стоит.

### 5.1.3 SWOT-анализ

SWOT-анализ разработанного протокола представляет собой двухэтапный комплексный анализ разработки. Результаты анализа представлены в таблице 2.

Таблица 2 – Матрица SWOT разработки

| | **Сильные стороны научно-исследовательского проекта:**<br>С1. Инновационность разработки (нет упоминаний о функционировании таких систем)<br>С2. Модель протестирована на большом количестве реальных данных<br>С3. Простота в использовании для конечного пользователя.<br>С4. Наличие бюджетного финансирования.<br>С5. Гибкость разработки | **Слабые стороны научно-исследовательского проекта:**<br>Сл1. Срок внедрения в университете, в связи с доказательной базой прогнозной модели<br>Сл2. Сложность внедрения модуля в общую электронную систему университета<br>Сл3. Отсутствие аналогов, на которые можно было бы опираться при внедрении<br>Сл4. Слабая доказательная база самой модели. |
|---|---|---|
| **Возможности:**<br>В1. ТПУ станет одним их первых вузов кто внедрит систему автоматического мониторинга успеваемости.<br>В2. Привлечение специалистов ТПУ для работы над проектом<br>В3. Появление дополнительного спроса от других вузов<br>В4. Публикация о проекте в тематических журналах<br>В5. Повышение стоимости конкурентных разработок | Данная разработка является инновационной и позволит вузу быть лидером в данной сфере, а значит это возможность для сотрудничества с другими вузами, заинтересованными в данной тематике. Данная разработка является востребованной и масштабируемым средством автоматизации оценки успеваемости учащихся. | При внедрении может возникнуть ряд проблем, связанный с доказательной базой точности этой модели, что также является возможностью для исследований данного вопроса и выступления на различных конференциях. |
| **Угрозы:**<br>У1. Отсутствие желания использования данной системы у сотрудников единого деканата.<br>У2. Изменение функциональных требований к готовому модулю<br>У3. Отказ от технической поддержки проекта после внедрения<br>У4. Отказ от внедрения из-за доказательной базы модели машинного обучения. | Наступление сценариев большинства угроз маловероятен, но тем не менее имеет место быть. Данная разработка за счет простоты использования и всей сложности в алгоритме построения прогнозной модели будет пользоваться успехом у конечного пользователя. | Отсутствие аналогов и слабость доказательной базы может повлечь скептическое отношение к использованию системы у конечного пользователя. Требуется объяснение того, что система автоматизирует процесс работы со студентами, а не усложняет его. |

Данная разработка обладает рядом возможностей в условиях низкой вероятности возникновения угроз. Разработка спроектирована таких образом, что сильные стороны предусматривают трудности с внедрением данного модуля в систему, а также возникновению задач по доказательной базе к модели машинного обучения.

### 5.1.4 Оценка готовности разработки к коммерциализации

Одной из важных задач в ходе выполнения данного раздела является оценка готовности разработки к коммерциализации. Оцениваемыми параметрами являются как научная, так и коммерческая составляющая. Таблица 3 представляет собой бланк оценки степени готовности разработки к коммерциализации.

Таблица 3 – Бланк оценки степени готовности разработки к коммерциализации

| № п/п | Наименование | Степень проработанности разработки | Уровень имеющихся знаний у разработчика |
|---|---|---|---|
| 1. | Определен имеющийся научно-технический задел | 4 | 4 |
| 2. | Определены перспективные направления коммерциализации научно-технического задела | 3 | 4 |
| 3. | Определены отрасли и технологии (товары, услуги) для предложения на рынке | 2 | 2 |
| 4. | Определена товарная форма научно-технического задела для представления на рынок | 2 | 2 |
| 5. | Определены авторы и осуществлена охрана их прав | 2 | 2 |
| 6. | Проведена оценка стоимости интеллектуальной собственности | 1 | 1 |
| 7. | Проведены маркетинговые исследования рынков сбыта | 1 | 1 |
| 8. | Разработан бизнес-план коммерциализации научной разработки | 1 | 1 |
| 9. | Определены пути продвижения научной разработки на рынок | 3 | 4 |

| | | | |
|---|---|---|---|
| 10. | Разработана стратегия (форма) реализации научной разработки | 5 | 5 |
| 11. | Проработаны вопросы международного сотрудничества и выхода на зарубежный рынок | 1 | 1 |
| 12. | Проработаны вопросы использования услуг инфраструктуры поддержки, получения льгот | 2 | 2 |
| 13. | Проработаны вопросы финансирования коммерциализации научной разработки | 2 | 3 |
| 14. | Имеется команда для коммерциализации научной разработки | 3 | 3 |
| 15. | Проработан механизм реализации разработки | 5 | 5 |
| | **ИТОГО БАЛЛОВ:** | 37/75 | 40/75 |

Поскольку данная разработка является индивидуальным проектом для уникального научного проекта для Томского Политехнического Университета, не предполагающем дальнейший выход на рынок, коммерциализация данного продукта не является целесообразной. Более того, организация, являющаяся потенциальным потребителем, является некоммерческой. В связи с этим провести полноценную оценку перспективы коммерциализации не представляется возможным. По результатам оценки можно утверждать, что данный проект еще не готов к коммерциализации, главным образом, с точки зрения сбыта разработки и финансирования коммерциализации.

Данную разработку возможно коммерциализировать способом передачи ноу-хау. При условии, что данная разработка покажет хорошие результаты после внедрения в ТПУ, то есть вероятность что ей заинтересуются и другие ВУЗы.

### 5.2 Инициация разработки

В рамках инициации разработки формулируются цели и ожидаемые результаты работы. Также определяются заинтересованные стороны разработки возможные ограничения. Заинтересованные в данной разработке стороны представлены в таблице 4.

Таблица 4 – Заинтересованные стороны разработки

| **Заинтересованные стороны** | **Ожидания заинтересованных сторон** |
|---|---|
| Высшие учебные заведения | Повышение качества мониторинга успеваемости студентов |
| Разработчики в области обработки больших данных | Получения опыта и создание новых и эффективных методологий для обработки больших объемов данных |

Цели и результат проекта отображены в таблице 5.

Таблица 5 – Цели и результат разработки

| | |
|---|---|
| **Цели разработки:** | Прогнозная модель для оценки успеваемости студентов университета по итогам текущего обучения |
| **Ожидаемые результаты разработки:** | 1) Формализованное описание методологии<br>2) Программная реализация алгоритмов |
| **Критерии приемки результата разработки:** | 1) Точность работы алгоритмов<br>2) Эффективность работы алгоритмов |
| **Требования к результату разработки:** | **Требования:** |
| | Формализованное описание работы методологии |
| | Точность работы алгоритмов |

В таблице 6 представлена рабочая группа разработки, определена роль и основные функции каждого участника в разработке.

Таблица 6 – Рабочая группа разработки

| **№ п/п** | **ФИО, основное место работы, должность** | **Роль в разработке** | **Функции** | **Трудо-затраты, час.** |
|---|---|---|---|---|
| 1 | Губин Евгений Иванович, ТПУ ОИТ ИШИТР, доцент | Научный руководитель | Утверждение основных разделов, выдача заданий к исполнению, координирование деятельности исполнителя | 30 |
| 2 | Зяблецев Павел Андреевич, ТПУ ОИТ ИШИТР, магистрант гр. 8ПМ8И | Исполнитель | Исполнение поставленных задач | 180 |
| ИТОГО | | | | 210 |

Данный раздел отражает тот факт, что выполняемая работа имеет довольно большой объем. Заинтересованные стороны проекта ожидают достаточно высококачественные результаты, которые необходимо достичь исполнителю.

## 5.3 Планирование управления разработкой

### 5.3.1 Иерархическая структура работ

Иерархическая структура работ для данной разработки представляет собой детализацию укрупненной структуры работ, продемонстрированной на рисунке 29.

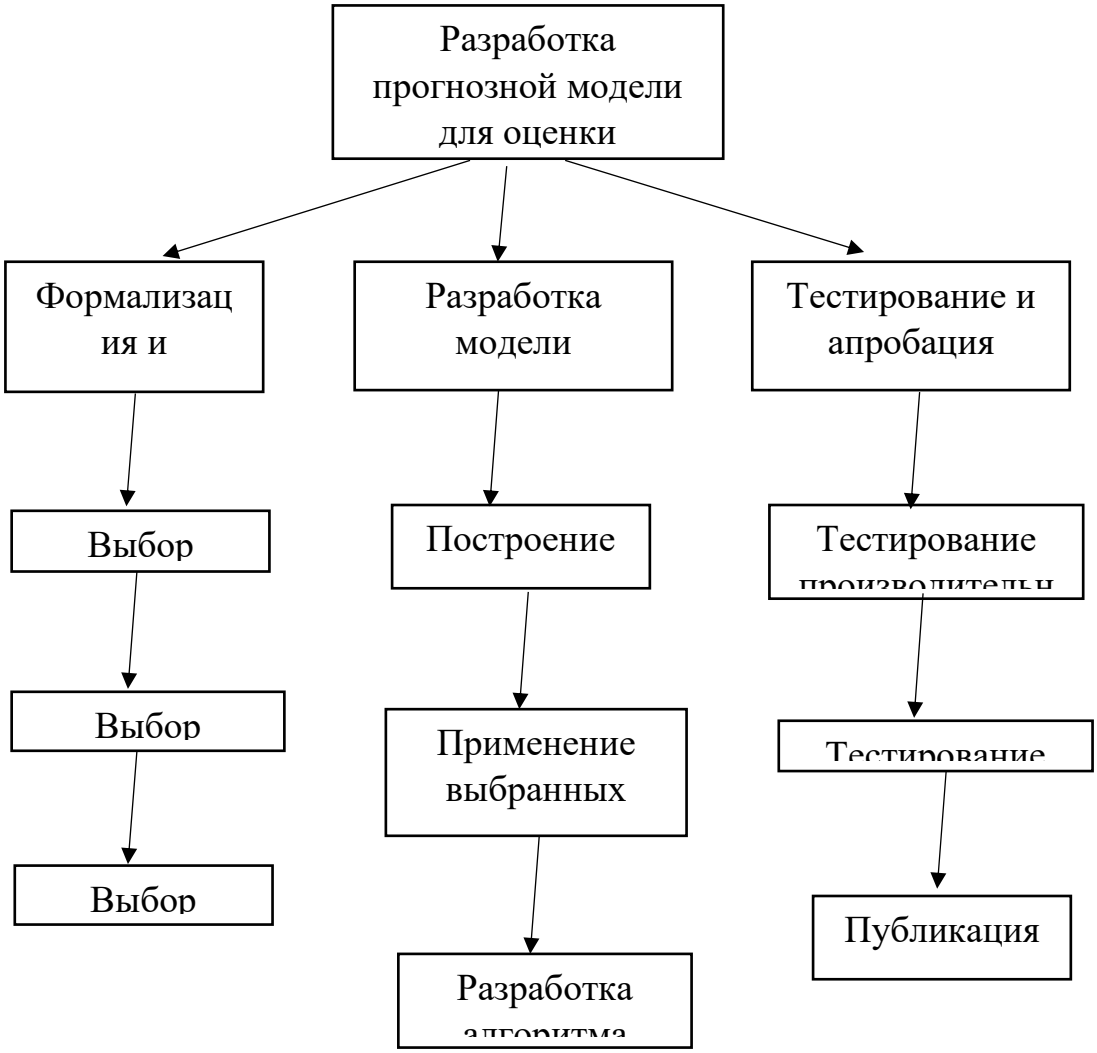



Рисунок 29. Иерархическая структура работ по проведению разработки

Задачи по созданию данной разработки разделены на три основных блока: формализация и разработка, реализация, а также тестирование и апробация.

### 5.3.2 План разработки

Чтобы отразить ключевые события по ведению разработки, необходимо составить календарный план. План представлен в таблице 7

Таблица 7 – Календарный план разработки

| Код работы | Название | Длительность, дни | Дата начала работ | Дата окончания работ | Состав участников |
|---|---|---|---|---|---|
| 1 | Выбор научного руководителя магистерской работы | 1 | 01.09.19 | 01.09.19 | Зяблецев Павел Андреевич |
| 2 | Составление и утверждение темы магистерской работы | 2 | 03.09.19 | 04.09.19 | Губин Евгений Иванович |
| 3 | Составление календарного плана-графика выполнения магистерской работы | 2 | 05.09.19 | 06.09.19 | Зяблецев Павел Андреевич, Губин Евгений Иванович |
| 4 | Выявление требований к разработке | 7 | 07.09.19 | 14.09.19 | Зяблецев Павел Андреевич, Губин Евгений Иванович |
| 5 | Подбор и изучение литературы по теме магистерской работы | 25 | 15.09.19 | 13.10.19 | Зяблецев Павел Андреевич |
| 6 | Анализ предметной области | 15 | 15.10.19 | 31.10.19 | Зяблецев Павел Андреевич |
| 7 | Предварительная обработка данных | 30 | 01.11.19 | 05.12.19 | Зяблецев Павел Андреевич |
| 8 | Разработка прогнозной модели | 80 | 06.12.19 | 08.03.20 | Зяблецев Павел Андреевич |
| 9 | Тестирование | 20 | 09.03.20 | 01.04.20 | Зяблецев Павел Андреевич |
| 10 | Анализ полученных результатов, сравнительная оценка точности прогноза | 4 | 02.04.20 | 05.04.20 | Зяблецев Павел Андреевич, Губин Евгений Иванович |

| | | | | | |
|---|---|---|---|---|---|
| 11 | Согласование выполненной работы с научным руководителем | 4 | 06.04.20 | 10.04.20 | Зяблецев Павел Андреевич, Губин Евгений Иванович |
| 12 | Выполнение других частей работы (финансовый менеджмент, социальная ответственность) | 30 | 11.04.20 | 11.05.20 | Зяблецев Павел Андреевич |
| 13 | Подведение итогов, оформление работы | 10 | 20.05.20 | 1.06.20 | Зяблецев Павел Андреевич, Губин Евгений Иванович |

### 5.3.2.1 Определение трудоемкости выполнения работ

Трудоемкость выполнения научного исследования оценивается экспертным путем в человеко-днях и носит вероятностный характер, завися от множества трудно учитываемых факторов. Для определения ожидаемого значения трудоемкости $t_{ожі}$ используется формула:

$$t_{ожі} = \frac{3tmini+2tmaxi}{5}, \qquad (1)$$

где $t_{ожі}$ – ожидаемая трудоемкость і-й работы чел.-дн;

$t_{maxi}$ – минимально возможная трудоемкость выполнения заданной работы, чел.-дн.;

$t_{mini}$ – минимально возможная трудоемкость выполнения заданной работы, чел.-дн.

Промежуточные расчеты представлены в таблице 9.

Таблица 9 – Временные показатели проведения разработки

| Наименование работы | Исполнители работы | Трудоемкость работ, чел-дни | | | Длительность работ, дни | |
|---|---|---|---|---|---|---|
| | | tmin | tmax | tож | Тр | Тк |
| Выбор научного руководителя магистерской работы | Зяблецев П.А. | 1 | 2 | 1,4 | 1 | 1 |

| | | | | | | |
|---|---|---|---|---|---|---|
| Составление и утверждение темы магистерской работы | Губин Е.И. | 1 | 3 | 1,8 | 2 | 2 |
| Составление календарного плана- графика выполнения магистерской работы | Зяблецев П.А | 1 | 3 | 1,8 | 2 | 2 |
| | Губин Е.И. | 1 | 3 | 1,8 | 2 | 2 |
| Выявление требований к разработке | Зяблецев П.А | 5 | 10 | 7 | 7 | 8 |
| | Губин Е.И. | 5 | 10 | 7 | 7 | 8 |
| Подбор и изучение литературы по теме магистерской работы | Зяблецев П.А | 22 | 30 | 25,2 | 25 | 29 |
| Анализ предметной области | Зяблецев П.А | 10 | 22 | 14,8 | 15 | 17 |
| Предварительная обработка данных | Зяблецев П.А | 20 | 45 | 30 | 30 | 35 |
| Разработка прогнозной модели | Зяблецев П.А | 60 | 110 | 80 | 80 | 89 |
| Тестирование | Зяблецев П.А | 10 | 35 | 20 | 20 | 24 |
| Анализ полученных результатов, сравнительная оценка производительности протокола | Зяблецев П.А | 2 | 7 | 4 | 4 | 4 |
| | Губин Е.И. | 2 | 7 | 4 | 4 | 4 |
| Согласование выполненной работы с научным руководителем | Зяблецев П.А | 2 | 7 | 4 | 4 | 5 |
| | Губин Е.И. | 2 | 7 | 4 | 4 | 5 |
| Выполнение других частей работы (финансовый менеджмент, социальная ответственность) | Зяблецев П.А | 20 | 45 | 30 | 30 | 35 |
| Подведение итогов, оформление работы | Зяблецев П.А | 7 | 14 | 9,8 | 10 | 12 |
| | Губин Е.И. | 7 | 14 | 9,8 | 10 | 12 |

### 5.3.2.2 Разработка графика проведения разработки

На рисунке 30 представлена диаграмма Ганта с планом выполнения работ, где М – магистрант (Зяблецев Павел Андреевич), НР – научный руководитель (Губин Евгений Иванович).

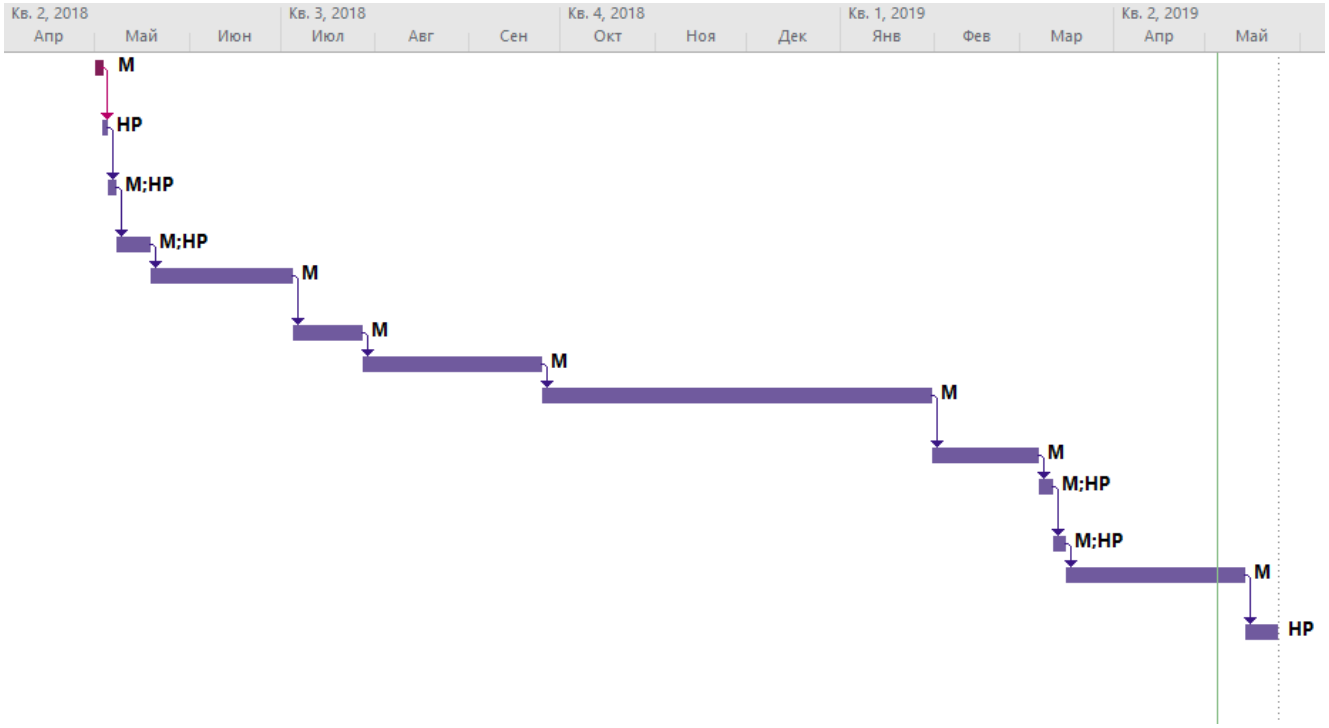



Рисунок 30. График проведения разработки

### 5.3.3 Бюджет разработки

Для данной разработки бюджет состоит из следующих пунктов:

1) Материальные затраты;
2) Амортизационные отчисления;
3) Основная заработная плата исполнителей темы;
4) Дополнительная заработная плата исполнителей темы;
5) Страховые отчисления;
6) Накладные расходы.

### 5.3.3.1 Расчет материальных затрат разработки

Для проведения исследования какие-либо специальные материалы и комплектующие не приобретались. Единая сумма на канцелярские принадлежности составляет 2000 рублей.

### 5.3.3.2. Расчет амортизационных отчислений

Поскольку для проведения исследований специальное дорогостоящее оборудование не приобреталось, при расчете затрат учитывается только амортизация. Первоначальная стоимость ПК магистранта, используемого для проведения исследований, составляет 65000 рублей. Срок полезного использования данной машины – 3 года, из которых 9 месяцев машина использовалась для написания ВКР.

Норма амортизации:

$$А_н = \frac{1}{n} \times 100\% = \frac{1}{3} \times 100\% = 33.33\ \% \qquad (2)$$

Годовые амортизационные отчисления:

$$А_г = 65000 \times 0{,}33 = 21450 \text{ рублей} \qquad (3)$$

Ежемесячные амортизационные отчисления:

$$А_м = \frac{21450}{12} = 1787{,}5 \text{ рублей} \qquad (4)$$

Итоговая сумма амортизации основных средств:

$$А = 1787{,}5 \times 9 = 16087{,}5 \text{ рублей} \qquad (5)$$

Таким образом затраты на амортизацию ПК составляют 16087,5 рублей.

Годовая лицензия программного среды разработки для Python составляет 199$ (или 13000 рублей в рублевом эквиваленте), т.е. срок полезного использования данного ПО составляет 1 год.

Ежемесячные амортизационные отчисления:

$$А_м = 13000 / 12 = 1083 \text{ рублей} \qquad (6)$$

Итоговая сумма амортизации основных средств:

$$A=1083 \times 9=9747 \text{ рублей} \qquad (7)$$

Суммарная стоимость остального программного обеспечения с перманентной лицензией составляет 42000 рублей.

Норма амортизации:

$$A_н = \frac{1}{n} \times 100\ \% = \frac{1}{5} \times 100\ \% = 20\ \% \qquad (8)$$

Годовые амортизационные отчисления:

$$A_г = 42000 \times 0{,}2 = 8400 \text{ рублей} \qquad (9)$$

Ежемесячные амортизационные отчисления:

$$A_м = 8400 / 12 = 700 \text{ рублей} \qquad (10)$$

Итоговая сумма амортизации основных средств:

$$A = 700 \times 9 = 6300 \text{ рублей} \qquad (11)$$

Амортизация остального программного обеспечения составляет 6300. Итого амортизация ПО составляет 16047 рублей.

Таблица 8. Расчет затрат на амортизацию

| Наименование | Затраты, руб. |
|---|---|
| Амортизация ПК | 16087,5 |
| Амортизация ПО | 16047 |
| **Итого** | **32134,5** |

### 5.3.3.3. Основная заработная плата исполнителей темы

Исполнителями темы выступают научный руководитель и инженер. Оклад руководителя в ТПУ без районного коэффициента составляет 35120 рубля, оклад инженера – 21760 рублей. Баланс рабочего времени для 6-дневной

недели, по которой учитывается рабочее время преподавателей и студентов, представлен в таблице.

Таблица 9. Баланс рабочего времени (для 6-дневной недели)

| Показатели рабочего времени | Дни |
|---|---|
| Календарные дни | 365 |
| Нерабочие дни (праздники/выходные) | 66 |
| Потери рабочего времени (отпуск/невыходы по болезни) | 56 |
| Действительный годовой фонд рабочего времени | 243 |

Таблица 10. Расчет основной заработной платы

| Исполнители | Здн, руб. | Кпр | Кд | Кр | Тр | Зосн |
|---|---|---|---|---|---|---|
| Инженер | 1860,02 | 0,3 | 0,3 | 1,3 | 180 | 334803,6 |
| Научный руководитель | 3126,40 | 0,3 | 0,3 | 1,3 | 30 | 93792 |
| | | | | | **Итого:** | **428595,6** |

$$З_{дн} = \frac{Зм \times М}{Fд} = \frac{21760 \times (1+Кпр+Кд) \times Кр \times 10,4}{243} = 1860,02 \text{ руб} \quad (12)$$

$$З_{дн} = \frac{Зм \times М}{Fд} = \frac{35120 \times (1+0,3+0,3) \times 1,3 \times 10,4}{243} = 3126,40 \quad (13)$$

Расчет основной заработной платы (для инженера и руководителя соответственно):

$$З_{осн} = З_{дн} \times Т_р \quad (14)$$

$$Зосн_{инж} = 1860,02 \times 180 = 334\,803,6 \quad (15)$$

$$Зосн_{рук} = 3126,40 \times 30 = 93\,792 \quad (16)$$

### 5.3.3.4. Дополнительная заработная плата исполнителей темы

Пусть дополнительная заработная плата составляет 10% от основной заработной платы. Тогда зарплаты инженера и руководителя соответственно будут высчитываться по формуле:

$$З_{доп} = З_{осн} \times 0,1 \qquad (16)$$

$$Здоп_{инж} = 334\,803,6 \times 0,1 = 33\,480,36 \qquad (17)$$

$$Здоп_{рук} = 93\,792 \times 0,1 = 9379,2 \qquad (18)$$

Дополнительная заработная плата исполнителей равна 33480,36 и 9379,2 соответственно.

### 5.3.3.5. Отчисления во внебюджетные фонды (страховые отчисления)

Составляют 30% от заработной платы (основная + дополнительная). Таким образом страховые взносы составляют 141 436,53 рублей.

$$Отч = (З_{доп} + З_{осн}) \times 0,3 \qquad (19)$$

$$Отч = (428\,595,6 + 42\,859,5) \times 0,3 = 141\,436,53 \qquad (20)$$

### 5.3.3.6. Накладные расходы

Накладные расходы составляют 16% от суммы материальных затрат, затрат на специальное оборудование, затрат на основную заработную плату, затрат на дополнительную заработную плату и страховых взносов.

Таким образом, сумма накладных расходов равна:

N = (2000 +32134,5+ 428595,6 + 42859,56 + 141436,53) × 0,16 = 103524,19

### 5.3.3.7. Формирование бюджета затрат научно-исследовательского разработки

Таблица 11. Бюджет затрат

| Наименование | Сумма, руб. | Удельный вес, % |
|---|---|---|
| Материальные затраты | 2000 | 0,27 |
| Затраты на специальное оборудование | 32134,5 | 4,28 |

| | | |
|---|---|---|
| Затраты на основную заработную плату | 428595,6 | 57,1 |
| Затраты на дополнительную заработную плату | 42859,56 | 5,71 |
| Страховые взносы | 141436,53 | 18,84 |
| Накладные расходы | 103524,19 | 13,79 |
| Общий бюджет | 750550 | 100 |

Общий бюджет разработки составляет 750550 рубля.

### 5.3.4 Риски разработки

Проведение любого научно-исследовательского проекта сопряжено с возникновением различных рисков. Предварительное определение рисков помогает своевременному принятию мер по предотвращению возникновения угроз или минимизации их последствий.

Таблица 14 – Определение рисков

| № п/п | Наименование риска | Описание риска |
|---|---|---|
| 1 | Политические | Риск отказа от продолжения сотрудничества потенциального заказчика в связи с обострением политической обстановки. |
| 2 | Технологические | Безвозвратная утеря большого процента исходных данных, на которые опирается разработка, в ходе работы. |
| 3 | Финансовые | Прекращение финансирования проекта. |
| 4 | Технические | Сбой или поломка оборудования, связанного с хранилищами данных, на которые опирается разработка. |

Таблица 15 – Оценка вероятности рисков

| № п/п | Наименование риска | Оценка вероятности риска (низкая, средняя, высокая) |
|---|---|---|
| 1 | Политические | Низкая |
| 2 | Технологические | Низкая |
| 3 | Финансовые | Низкая |
| 4 | Технические | Низкая |

Таблица 16 – Оценка уровня потерь

| № п/п | Наименование риска | Оценка уровня потерь (низкий, средний, высокий) |
|---|---|---|
| 1 | Политические | Высокий |
| 2 | Технологические | Средний |
| 3 | Финансовые | Высокий |
| 4 | Технические | Низкий |

Таблица 17 – Оценка уровня потерь

| **Матрица вероятности рисков/потерь** | | | | |
|---|---|---|---|---|
| | | Уровень потерь | | |
| | | Высокий | Средний | Низкий |
| Вероятность | Высокая | Высокий | Высокий | Умеренный |
| | Средняя | Существенный | Существенный | Незначительный |
| | Низкая | Умеренный | Умеренный | Незначительный |

Таблица 18 – основные мероприятия по снижению рисков

| № п/п | Наименование риска | Мероприятия по снижению риска |
|---|---|---|
| 1 | Политические | Заключение контракта о сотрудничестве на четко обозначенный период. |
| 2 | Технологические | Разграничение уровня доступа пользователей. Создание бэкапов данных. |
| 3 | Финансовые | Своевременное принятие мер по подготовке отчетности по текущим работам и подача заявки на новые. |
| 4 | Технические | Соблюдение протокола безопасности |

Основными рисками при выполнении разработки можно назвать технологические и технические риски, связанные с хранилищами данных о научном эксперименте в области обработки больших объемов данных, при этом наибольшую угрозу представляют собой технологические риски. Среди прочих рисков стоит отметить, что финансовые риски связаны, в основном, с финансированием. Предотвращения данного риска возможно в случае своевременного предоставление релевантной документации в научные фонды. Политические риски наименее вероятны, однако стоит отметить, что в случае

наступления предполагаемого сценария, подобные риски имеют серьезные последствия для такого уникального научного проекта.

## 5.2 Определение потенциального эффекта разработки

Потенциальным пользователям разработки является Томский Политехнический Университет, поэтому метод оценки абсолютной эффективности исследования не подходит для данного исследования. Поскольку в ходе выполнения магистерской диссертации разрабатывался только один вариант разработки, провести оценку сравнительной эффективности исследования не представляется возможным.

Данная разработка ориентирована на конкретного потребителя, которыми является Томский Политехнический Университет. Стоит отметить, что, поскольку потенциальный рынок сбыта мал, коммерциализация данной разработки остается открытым вопросом. Бланк оценки степени готовности к коммерциализации указывает на то, что готовность в коммерциализации находится на нижнем пороге уровня выше среднего.

Общая продолжительность исследования составляет 9 месяцев.

Потенциальная стоимость исследования составляет около 750 тыс. рублей. При этом специальное дорогостоящее оборудование не закупалось. Данная цена является конкурентоспособной, поскольку стоимость решений для подобных вычислительных структур «из коробки», не ориентированных на конкретного потребителя, находятся в том же ценовом диапазоне.

Данный проект сопряжен с малыми рисками, однако стоит обратить внимание на технологические риски, последствия которых могут нанести существенный урон подобным исследованиям, а также при продолжении работы над данным проектом. Проект не имеет прямых аналогов в обозначенных условиях обработки больших объемов данных в высших учебных заведениях.

**Выводы по разделу**

Проведено комплексное описание и анализ финансово-экономических аспектов выполненной работы.

Составлен перечень проводимых работ, их исполнителей и продолжительность выполнения этапов работ, составлен линейный график.

Рассчитана смета затрат на выполнение проекта, проведен расчет себестоимости и прибыли проекта.

Определены показатели эффективности проекта и проведена оценка его эффективности.

## Глава 6. Социальная ответственность

Объектом исследования выступает рабочее место программиста, разрабатывающего прогнозную модель для оценки успеваемости студентов университета по итогам текущего обучения. Использование данной модели позволит моделировать количество академических задолженностей для конкретного студента в конце семестра по итогам текущего обучения. Преждевременная оценка академических успехов студентов позволит более детально работать с целевой группой учащихся, которые находятся в группе должников и таким образом система позволит улучшить качество учебно-воспитательной работы в университете.

Рабочей зоной при разработке данной информационной системы является учебная аудитория №207 Кибернетического центра Томского Политехнического Университета, оборудованная системой отопления, кондиционирования воздуха, с естественным и искусственным освещением. Рабочее место – стационарное, оборудованное персональным компьютером и оргтехникой.

Характеристики учебной аудитории:

- Длина рабочего помещения – 5 метров, ширина – 4 метра, высота – 2,5 метра;
- Площадь помещения – 20 $м^2$;
- Объем помещения – 50 $м^3$;
- Присутствует естественная вентиляция: окно, дверь и вытяжное вентиляционное отверстие;
- В помещение имеется естественное освещение, а также установлено искусственное освещение.

В данной работе освещен комплекс мер организационного, правового, технического и режимного характера, которые минимизируют негативные последствия проектирования информационной системы, а также рассматриваются вопросы техники безопасности, охраны окружающей среды и

пожарной профилактики, даются рекомендации по созданию оптимальных условий труда.

### 6.1 Правовые и организационные вопросы обеспечения безопасности

Трудовые отношения между работодателем и работником регулируются таким правовым актом как Трудовой кодекс Российской федерации от 30.12.2001 N 197-ФЗ. В ТК РФ [13], а также в соответствии с Конституцией РФ, признаются свобода труда, выбор и согласие на него, а также выбор профессии и деятельности. Запрещаются принудительный труд, дискриминация по какому-либо признаку. Гарантируются справедливые и достойные условия труда.

ТК РФ регламентирует порядок разрешения индивидуальных и коллективных трудовых споров, особенности труда женщин, детей и людей пенсионного возраста, права и обязанности работодателей и работников, нормы рабочего времени, порядок оплаты труда и виды компенсаций во вредных условиях труда, а также особенности социального страхования.

В соответствии со ст. 111 ТК РФ, рабочая неделя (в т.ч. шестидневная) не должна превышать 40 часов в неделю. Воскресенье является выходным днем.

В соответствии со ст. 212 ТК РФ, работодатель обязан обеспечить безопасные условия труда, а также обязательное социальное страхование работников от несчастных случаев на производстве и профессиональных заболеваний.

В соответствии со ст. 142 ТК РФ, в случае задержки выплаты заработной платы на срок более 15 дней работник имеет право, известив работодателя в письменной форме, приостановить работу на весь период до выплаты задержанной суммы, кроме ряда перечисленных случаев.

Данные о работнике, предоставляемые работодателю, обрабатываются только с согласия самого работника и охраняются Федеральным Законом от 27.07.2006 N 152-ФЗ (ред. от 25.07.2011) «О Персональных Данных» [14].

Обязательно предусмотрен предварительный медосмотр при приеме на работу и периодические медосмотры. Каждый сотрудник обязан пройти инструктаж по технике безопасности перед приемом на работу и в дальнейшем, должен быть пройден инструктаж по электробезопасности и охране труда. Предприятие обеспечивает рабочий персонал всеми необходимыми средствами индивидуальной защиты [28].

Рабочие места должны соответствовать требованиям ГОСТ 12.2.032-78 «ССБТ. Рабочее место при выполнении работ сидя. Общие эргономические требования» [29] и ГОСТ 12.2.061-81 «ССБТ. Оборудование производственное. Общие требования безопасности к рабочим местам» [27].

Работа с применением персональных компьютеров сопряжена со значительными зрительными и нервно-психологическими нагрузками, что повышает требования к организации труда пользователей ПК. Конструкция рабочей мебели должна обеспечивать возможность индивидуальной регулировки, соответственно росту работающего, и создавать удобную позу. Часто используемые предметы труда и органы управления должны находиться в оптимальной рабочей зоне. Конструкция рабочего стола должна обеспечивать оптимальное размещение на рабочей поверхности используемого оборудования с учетом его количественных и конструктивных особенностей, а также характера выполняемой работы. Высота рабочей поверхности стола должна регулироваться в пределах 680-800 мм, при отсутствии такой возможности его высота должна быть не менее 725 мм.

На поверхности рабочего стола для документов необходимо предусматривать размещение специальной подставки, расстояние которой от глаз должно быть аналогичным расстоянию от глаз до клавиатуры.

Модульными размерами рабочей поверхности стола, на основании которых должны рассчитываться конструктивные размеры, следует считать: ширину 800, 1000, 1200 и 1400 мм, глубину 800 и 1000 мм при нерегулируемой его высоте, равной 725 мм. Под столешницей рабочего стола должно быть

свободное пространство для ног с размерами по высоте не менее 600 мм, по ширине 500 мм, по глубине 650 мм.

Конструкция рабочего стула должна обеспечивать поддержание рациональной рабочей позы при работе, что позволит изменять позу для снижения статического напряжения мышц шейно-плечевой области и спины для предупреждения развития утомления.

## 6.2 Производственная безопасность

Опасные и вредные факторы при разработке и эксплуатации проектируемого решения. Факторы по ГОСТ 12.0.003-74 «Опасные и вредные производственные факторы. Классификация» [16] представлена в таблице 19.

Таблица 19. Основные элементы производственного процесса, формирующие опасные и вредные факторы

| Наименование видов работ и параметров производственного процесса | Факторы (ГОСТ 12.0.003-74 ССБТ ) [16] | | Нормативные документы |
|---|---|---|---|
| | Вредные | Опасные | |
| Работа с ПК, устройствами ввода и вывода информации | Повышенный уровень электромагнитного излучения | Опасность поражения электрическим током | СанПиН 2.2.2/2.4.1340-03 [17]<br>ГОСТ Р 50571. 17-2000 [18]<br>ГОСТ 12.1.045–84 ССБТ [24] |
| | Несоответствие параметрам микроклимата | Статическое электричество | ГОСТ 12.1.005-88 [15]<br>ГОСТ 12.4.124-83 [30] |
| | Недостаточная освещенность рабочего места | Короткое замыкание | СанПиН 2.2.1/2.1.1.1278-03 [19]<br>ГОСТ 26522-85 [29] |
| | Повышенный уровень шума | | СанПиН 2.2.4.3359-16 [20] |
| | Умственное перенапряжение | | ТОИ Р-45-084-01 [21] |

### 6.2.1 Повышенный уровень электромагнитных излучений

Когда все устройства персонального компьютера включены, в районе рабочего места программиста, формируется сложное по структуре электромагнитное поле. Реальную угрозу для пользователя компьютера

представляют электромагнитные поля. Известно, что монитор персонального компьютера является источником [26]:

- электростатического поля;
- слабых электромагнитных излучений в низкочастотном и высокочастотном в диапазонах (2 Гц – 400 кГц);
- ультрафиолетового излучения;
- инфракрасного излучения;
- излучения видимого диапазона.

В организме человека под влиянием электромагнитного излучения монитора происходят значительные изменения гормонального состояния, специфические изменения биотоков головного мозга, изменение обмена веществ. Пыль, притягиваемая электростатическим полем монитора, иногда становится причиной дерматитов лица, обострения астматических симптомов, раздражения слизистых оболочек [17, 26].

Для снижения воздействия электромагнитного излучения следует применять мониторы с пониженным уровнем излучения, также устанавливать защитные экраны, придерживаться регламентированного режима труда и отдыха, а также проводить регулярную гигиеническую уборку помещения.

Нормы электромагнитных полей, создаваемых ПЭВМ приведены в таблице 20, в соответствии с СанПиН 2.2.2/2.4.1340-03.

Таблица 20 - Временные допустимые уровни ЭМП, создаваемых ПЭВМ

| **Наименование параметров** | | **ВДУ ЭМП** |
|---|---|---|
| Напряженность электрического поля | В диапазоне частот 5 Гц – 2кГц | 25 В/м |
| | В диапазоне частот 2 кГц –400 кГц | 2,5 В/м |
| Плотность магнитного потока | В диапазоне частот 5 Гц – 2кГц | 250 нТл |
| | В диапазоне частот 2 кГц –400 кГц | 25 нТл |
| Электростатический потенциал экрана видеомонитора | | 500 В |

Для оценки соблюдения уровней необходим производственный контроль (измерения). В случае превышения уровней необходимы организационно-технические мероприятия (защита временем, расстоянием, экранирование источника, либо рабочей зоны, замена оборудования, использование СИЗ).

На рабочем месте уровень электромагнитного излучения не превышает допустимых норм, регламентированных СанПиН 2.2.2/2.4.1340-03. Для минимизации вредного влияния электромагнитного излучения на организм время работы за компьютером сокращено и чередуется с временем отдыха.

### 6.2.2 Несоответствие параметрам микроклимата

Оптимальное состояние воздушной среды должно обеспечивать ощущение теплового комфорта в течение 8-часового рабочего дня, не вызывать отклонений в состоянии здоровья. Энергетические затраты организма измеряются в ккал/ч (Вт) и по затраченной энергии работы разделяются на категории. Так работа программиста относится к категории Iа – интенсивность энергозатрат до 120 ккал/ч (до 139 Вт) [15]. Работы производятся в основном сидя и сопровождаются незначительным физическим напряжением. Допустимые параметры микроклимата на рабочем месте для категории Iа приведены в таблице 21.

Таблица 21. Допустимые величины показателей воздушной среды на рабочих местах производственных помещений по СанПиН 2.2.2/2.4.1340-03

| Сезон года | Категория тяжести выполняемых работ | Температура, °С | | Относительная влажность, % | | Скорость движения воздуха, м/сек | |
|---|---|---|---|---|---|---|---|
| | | Факт. | Допуст. | Факт. | Допуст. | Факт. | Допуст. |
| Холодный | Iа | 22-24 | 20-25 | 60 | 15-75 | 0,1 | 0,1 |
| Теплый | Iа | 23-25 | 21-28 | 60 | 15-75 | 0,1 | 0,2 |

Температура воздуха в рабочем помещении в холодное время года поддерживается в диапазоне (согласно измерениям термометром) от 21 до 23°С,

в теплое – от 23 до 25°С. Влажность в соответствии с нормами (согласно измерениям гигрометром) колеблется около 60%. Для поддержания соответствующих микроклиматических параметров используются системы отопления и вентиляции, а также проводится кондиционирование воздуха в помещении.

### 6.2.3 Недостаточная освещенность рабочего места

Для обеспечения нормативных условий работы необходимо провести оценку освещенности рабочей зоны в соответствие с СанПиН 2.2.1/2.1.1.1278-03. Правильное освещение рабочих мест и помещений является важным условием для создания безопасных и благоприятных условий труда. Все поле зрения должно быть освещено равномерно – это является основным гигиеническим требованием. Другими словами, уровень естественного освещения рабочего места и яркость дисплея компьютера должны быть приблизительно одинаковыми, т.к. яркий свет в зоне периферийного зрения заметно увеличивает глазное напряжение, что приводит к их быстрой утомляемости. Для снижения отраженной блескости наряду с перечисленными выше рекомендуются следующие мероприятия:

Для внутренней отделки интерьера помещений с компьютерами должны использоваться диффузно отражающие материалы с коэффициентом отражения для потолка 0,7 – 0,8, для стен 0,5 – 0,6, для пола – 0,3 - 0,5.

Дизайн ПЭВМ должен предусматривать окраску корпуса в спокойные мягкие тона с диффузным рассеиванием света. Корпус ПЭВМ, клавиатура и другие блоки и устройства ПЭВМ должны иметь матовую поверхность с коэффициентом отражения 0,4 – 0,6 и не иметь блестящих деталей, способных создавать блики.

Для освещения помещений с ПЭВМ рекомендуется применять светильники с зеркальными параболическими решетками. Применение светильников без рассеивателей или экранирующих решеток нежелательно.

Персоналу, эксплуатирующему компьютеры с черными экранами, не рекомендуется использование светлой или блестящей одежды.

Для кабинета информатики и учебных кабинетов норматив по СанПиН 2.2.1/2.1.1.1278-03 (искусственное общее освещение) минимальные значения освещённости - 300 люкс.

**Расчет освещения учебной аудитории №207 КЦ ТПУ**

Размеры помещения: длина А = 5 м, ширина В = 4 м, высота Н = 2,5 м. Высота рабочей поверхности $h_{рп}$ = 0,8 м. Требуется создать освещенность Е = 300 лк.

Коэффициент отражения оклеенных светлыми обоями стен $R_c$ = 30 % [31], потолка по типу Армстронг со светлыми потолочными плитами $R_п$ = 50 % (в таблице [31] нет значения для данного типа потолка, однако самым близким по параметрам будет значение для светло окрашенного деревянного потолка). Коэффициент запаса для люминесцентных ламп k = 1.5 [31], коэффициент неравномерности для люминесцентных ламп Z = 1.1 [31].

Рассчитываем систему общего люминесцентного освещения. Выбираем светильники типа ОД, интегральный критерий расположения светильников данного типа λ = 1,4 [31].

Н – высота помещения;

$h_c$ – расстояние светильников от перекрытия (свес);

$h_п = H – h_c$ – высота светильника над полом, высота подвеса;

$h_{рп}$ – высота рабочей поверхности над полом;

$h = h_п – h_{рп}$ – расчётная высота, высота светильника над рабочей поверхностью.

Приняв $h_c$ = 0,2 м, получаем h = 2,5 - 0,2 - 0,8 = 1,5 м;

Для расчета расстояний между ближайшими светильниками используем формулу 1 из учебного практикума [19].

$$L = \lambda \times h, \quad (1)$$

$L$ – расстояние между соседними светильниками или рядами (если по длине (А) и ширине (В) помещения расстояния различны, то они обозначаются $L_A$ и $L_B$).

Оптимальное расстояние $l$ от крайнего ряда светильников до стены рекомендуется принимать равным $L/3$.

$$L = 1,4 \times 1,5 = 2,1 \text{ м.};$$

$$L/3 = 0,7 \text{ м.}$$

Находим индекс помещения по формуле 2 из учебного практикума [31]:

$$i = S/(h \times (A+B)), \quad (2)$$

$$i = 20 / (1,5 * (4+5)) = 1,48$$

По таблице из справочника определяем коэффициент использования светового потока [31]:

$$\eta = 0.52$$

Размещаем светильники в 2 ряда. В каждом ряду можно установить 2 светильника типа ОД мощностью 40 Вт (с длиной 1,23 м), при этом разрывы между светильниками в ряду составят 60 см. Изображаем в масштабе план помещения и размещения на нем светильников. Учитывая, что в каждом светильнике установлено две лампы, общее число ламп в помещении N = 8.

Рассчитаем световой поток лампы по формуле 3 из учебного практикума [31]:

$$Ф = (E_н \times S \times K_З \times Z) / N \times \eta, \quad (3)$$

Где Ф – световой поток лампы;

Ен – нормируемая минимальная освещённость по СНиП 23-05- 95, лк; S – площадь освещаемого помещения, м2;

Кз – коэффициент запаса, 1,5 – для люминесцентных ламп [31];

Z – коэффициент неравномерности освещения, отношение Ecp/Emin. Для люминесцентных ламп при расчётах берётся равным 1,1 [31];

N – число ламп в помещении; η – коэффициент использования светового потока.

$$\Phi = \frac{300 \times 20 \times 1{,}5 \times 1{,}1}{8 \times 0{,}52} = 2380 \text{ Лм.}$$

Определяем потребный световой поток ламп в каждом из рядов:

По таблице из справочника выбираем ближайшую стандартную лампу – ЛД 40 Вт с потоком 2300 лм. Делаем проверку выполнения условия по формуле 4 из учебного практикума [31]:

$$-10\% \le \frac{\Phi_{л.станд} - \Phi_{л.\,расч}}{\Phi_{л.станд}} 100\% \le +20\%, \quad (4)$$

Получаем:

$$-10\% \le -3{,}5\ \% \le +20\%$$

Используемый тип и план размещения светильников в аудитории с люминесцентными лампами не выбивается из диапазона, поэтому можно предположить, что план и размещение светильников оказалось верным.

Определяем электрическую мощность осветительной установки

$$P = 8 \times 40 = 320 \text{ Вт.} \quad (5)$$

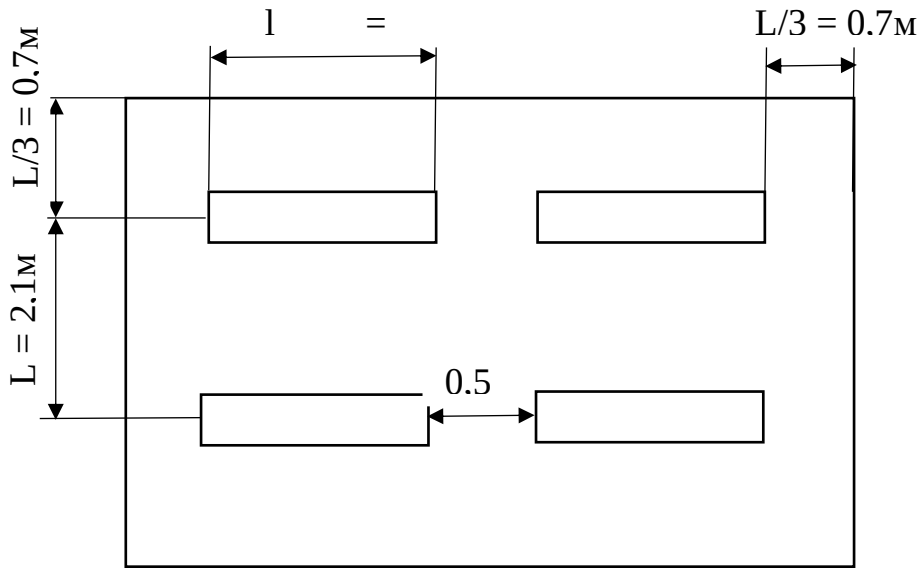



Рисунок 31. План помещения и размещения светильников с люминесцентными лампами

### 6.2.4 Повышенный уровень шума

Шум ухудшают условия труда, оказывают вредное воздействие на организм человека. Действие шума различно: он затрудняет разборчивость речи, вызывает снижение работоспособности, повышает утомляемость, вызывает необратимые изменения в органах слуха человека, ослабляет внимание, ухудшает память, снижает реакция, увеличивает число ошибок при работе.

Источником шума в учебных помещениях, эксплуатирующих вычислительную технику, являются сами вычислительные машины (встроенные в стойки ЭВМ вентиляторы, принтеры и т.д.), центральная система вентиляции и кондиционирования воздуха и другое оборудование.

На рабочем месте программиста исследователя предельно допустимый уровень звука приведен таблице 22.

Таблица 22. Предельно допустимый уровень звука по СанПиН 2.2.4.3359-16

| Рабочие места | Уровень звука, дБА |
|---|---|
| Учебные кабинеты, аудитории | 80 |
| Лаборатории при учебных кабинетах | 80 |

При выполнении основной работы уровень звука не должен превышать 80 дБА. По субъективным ощущениям шумовая обстановка на рабочем месте исследователя соответствует норме.

Снизить уровень шума в помещениях можно использованием звукопоглощающих материалов с максимальными коэффициентами звукопоглощения в области частот 63-8000 Гц для отделки стен и потолка помещений. Дополнительный звукопоглощающий эффект создают однотонные занавески из плотной ткани, повешенные в складку на расстоянии 15-20 см от ограждения. Ширина занавески должна быть в 2 раза больше ширины окна.

Уровень шума на рабочем месте не превышает предельно допустимый, регламентированный СанПиН 2.2.4.3359-16, и не препятствует комфортной

работе за ЭВМ.

### 6.2.5 Умственное перенапряжение

Для видов трудовой деятельности устанавливается 3 категории тяжести и напряженности работы с компьютером, которые определяются согласно таблице 23.

Таблица 23. Категория работы по тяжести и напряженности по ТОИ Р 45-084-01

| Категория работы по тяжести и напряженности | Уровень нагрузки за рабочую смену при видах работы на ПК | | |
|---|---|---|---|
| | Группа А<br>Количество знаков | Группа Б<br>Количество знаков | Группа В<br>Время работы, ч |
| III | До 60000 | До 40000 | До 6,0 |

При 8-часовой рабочей смене и работе на ПК регламентированные перерывы следует устанавливать:

Для третьей категории работ – через 1,5- 2,0 часа от начала рабочей смены и через 1,5-2,0 часа после обеденного перерыва продолжительностью 20 минут каждый или продолжительностью 15 минут через каждый час работы.

### 6.2.6. Нарушение правил электробезопасности

Электробезопасность – система организационных и технических мероприятий и средств, обеспечивающих защиту людей от вредного и опасного для жизни воздействия электрического тока, электрической дуги, электромагнитного поля и статического электричества.

Согласно СанПиН 2.2.2/2.4.1340-03. «Гигиенические требования к персональным электронно-вычислительным машинам и организации работы», рабочее место должно находиться в безопасной зоне, которое не характеризуется наличием таких условий, как повышенная влажность (относительная влажность воздуха длительно превышает 75%), высокая температура (более 35°С), токопроводящая пыль, токопроводящие полы,

возможность одновременного соприкосновения к имеющим соединения с землей металлическим элементам и металлическим корпусам электрооборудования.

Электрические установки, к которым относится ПК, представляют для человека большую потенциальную опасность, так как в процессе эксплуатации или проведения профилактических работ человек может коснуться комплектующих компьютера, находящихся под напряжением.

Специфическая опасность – корпуса ПК и прочего оборудования, оказавшегося под напряжением в результате повреждения или пробоя изоляции, не подают каких-либо сигналов, которые предупреждают человека об опасности. Причинами электропоражений являются: провода с поврежденной изоляцией, розетки сети без предохранительных кожухов.

Для защиты от поражения электрическим током все токоведущие части должны быть защищены от случайных прикосновений кожухами, корпус устройства должен быть заземлен. Заземление выполняется изолированным медным проводом сечением 1.5 мм, который присоединяется к общей шине заземления с общим сечением 48 м при помощи сварки. Общая шина присоединяется к заземлению, сопротивление которого не должно превышать 4 Ом.

Согласно ГОСТ Р 50571.17-2000. «Требования по обеспечению безопасности электроустановок в зданиях. Выбор мер защиты в зависимости от внешних условий. Защита от пожара», питание устройства в помещении, в котором выполнялась работа, осуществляется от силового щита через автоматический предохранитель, который срабатывает при коротком замыкании нагрузки. Для снижения величин возникающих разрядов применяются покрытия из антистатического материала.

Программист исследователь работает с электроприборами: компьютером (монитор, системный блок, компьютерная мышь и клавиатура).

В данном случае существует опасность электропоражения:

- при непосредственном прикосновении к токоведущим частям во время ремонта ПК;
- при прикосновении к нетоковедущим частям, оказавшимся под напряжением (в случае нарушения изоляции токоведущих частей ПК);
- при соприкосновении с полом, стенами, оказавшимися под напряжением;
- имеется опасность короткого замыкания в высоковольтных блоках: блоке питания и блоке дисплейной развёртки.

Токи статического электричества, наведенные в процессе работы компьютера на корпусах монитора, системного блока и клавиатуры, могут приводить к разрядам при прикосновении к этим элементам. Такие разряды опасности для человека не представляют, но могут привести к выходу из строя компьютера. Для снижения величин токов статического электричества используются нейтрализаторы, местное и общее увлажнение воздуха, полы с антистатической пропиткой.

Рабочие место программиста исследователя оборудовано таким образом, чтобы исключить взаимное соприкосновение кабелей и шнуров питания соседних компьютеров.

К организационно-техническим мероприятиям относится первичный инструктаж по технике безопасности. Первичный инструктаж по технике безопасности является обязательным условием для допуска к работе в данном помещении.

### 6.3 Экологическая безопасность

Разрабатываемый проект не имеет влияния на окружающую среду, так как само решение разрабатывается и используется внутри персональных компьютеров, которые могут стать источниками различных загрязнений.

Защита литосферы. Согласно ГОСТ Р 56397-2015 «Техническая экспертиза работоспособности радиоэлектронной аппаратуры, оборудования

информационных технологий, электрических машин и приборов. Общие требования» пункт 5.8.1, после проведения технической экспертизы если оборудование не ремонтопригодно, то оно признается неработоспособным и рекомендуется к списанию (замене); в случае деградационного отказа оборудования и нецелесообразности его ремонта и модернизации даются рекомендации о необходимости его списания и утилизации. [22]

Согласно «Методики проведения работ по комплексной утилизации вторичных драгоценных металлов из отработанных средств вычислительной техники», утвержденной Государственным Комитетом РФ по телекоммуникациям от 19 октября 1999г. В п.3.1.3. «Технология разборки универсальных ЭВМ» расписаны 4 этапа разборки и подготовки к утилизации внутренних частей ПК.

В результате выполнения этапов формируется партия сырья, включающая сортировку электронного лома по типу, проведение расчета количества ячеек, соединителей, серебросодержащих кабельных изделий, ячеек и типовых элементов замены, содержащие драгоценные металлы, а также партии черных и цветных металлов и сплавов (медь, сталь, никель, латунь, бронза, алюминий, дюралюминий, свинцово - оловянные припои) направляются на переработку на заводы ВДМ, полупроводниковые приборы (диоды, транзисторы), микросхемы в металлических и металлокерамических корпусах, а также конденсаторы в металлических корпусах демонтируются с плат и сортируются по типу, интегральные микросхемы в пластмассовых корпусах (серии 155, 551 и пр.) демонтируются и собираются отдельно, керамические конденсаторы типа КМ и резисторы после демонтажа также собираются отдельно.

На рабочем месте программиста используются 24 люминесцентных ламп ЛБ40, Согласно ГОСТ 12.3.031-83 «Работы со ртутью. Требования безопасности» п.2.1. все ртутьсодержащие отходы и вышедшие из строя приборы, содержащих ртуть, подлежат сбору и возврату для последующей регенерации ртути в специализированных организациях. В п.2.2. К работе по

замене и сбору отработанных ртутьсодержащих ламп допускаются только электромонтеры. Главным условием при замене и сборе отработанных ртутьсодержащих ламп является сохранение герметичности. В п.2.13. Факт сдачи ртутьсодержащих отходов подтверждается возвращением паспорта на вывоз отходов с отметкой о приеме представителя специализированного предприятия [23].

### 6.4 Безопасность в чрезвычайных ситуация

К наиболее вероятным ЧС можно отнести следующие: пожар (взрыв) в здании, авария на коммунальных системах жизнеобеспечения, землетрясение. Источниками возгорание может стать электропроводка, внутренние работающие устройства ПК, взрывоопасные предметы в помещении исследователя согласно ГОСТ 12.1.044-89 «Система стандартов безопасности труда. Пожаровзрывоопасность веществ и материалов. Номенклатура показателей и методы их определения» [24].

Превентивными мерами по предупреждению ЧС могут служить системы звукового и визуального оповещения персонала лаборатории и кабинетов об опасности, обучение персонала, методам работы с компьютером, наличие средств пожаротушения и информационных досок с планами эвакуации.

В случае угрозы возникновения ЧС необходимо отключить электропитание, вызвать по телефону пожарную команду, эвакуировать людей из помещения согласно плану эвакуации. При наличии небольшого очага пламени можно воспользоваться подручными средствами с целью прекращения доступа воздуха к объекту возгорания. В качестве подручных средств можно использовать углекислотные огнетушители ОУ-5 высокого давления с зарядом жидкой двуокиси углерода (по ГОСТ 8050-85).

**Вывод по разделу**

В результате выполнения работы по разделу «Социальная ответственность» был освещен комплекс мер организационного, правового, технического и режимного характера, которые минимизируют негативные последствия вредных и опасных факторов проектирования информационной системы. Были рассмотрены вопросы техники безопасности, охраны окружающей среды и пожарной профилактики, даны рекомендации по созданию оптимальных условий труда. В результате анализа было установлено, что аудитория, в которой выполняется работа, удовлетворяет всем требованиям нормативных документов в области охраны труда и окружающей среды, а также пожарной безопасности.

## Заключение

В результате выполнения магистерской диссертации были выполнены следующие задачи:

1. Проанализированы методы, которые используются для решения аналогичных задач в сфере образования.
2. Проведена предварительная обработка данных
3. Построена модель машинного обучения, способная прогнозировать успеваемость студентов на конец семестра.
4. Выявлены наиболее значимые признаки в определении успеваемости студентов.
5. Создан графический интерфейс пользователя прогнозной модели успеваемости студентов

Так как при предварительной обработке данных были использованы метки для обозначения «успешности студента», то тип машинного обучения был с учителем. Были апробированы 3 модели машинного обучения с учителем, а именно: логистическая регрессия, метод опорных векторов и случайный лес. Наилучший результат на тестовой выборке показал классификатор «случайный лес». Данный алгоритм является оптимальным так, как обладает следующими достоинствами:

Помимо этого, был создан графический интерфейс для модуля по прогнозированию успеваемости студентов на конец семестра по текущим оценкам.

**Список публикаций и научных достижений**

*Участие в конференциях:*

1. XVl Международная научно-практическая конференция студентов, аспирантов и молодых ученых «Молодёжь и современные и информационные технологии» с докладами:

- Анализ социальных данных с помощью технологий Big Data
- Использование OLAP технологии для выявления группы риска острого инфаркта миокарда
- Интеллектуальный анализ данных для определения группы риска сердечно сосудистых заболеваний

*Премии, звания, стипендии:*

1. Повышенная государственная стипендия по научно-исследовательской деятельности, весна 2018/2019 учебного года, весна 2019/2020 учебного года.
2. Стипендия Правительства РФ студентам ТПУ, обучающимся по специальностям или направлениям подготовки, соответствующим приоритетным направлениям модернизации и технологического развития российской экономики, весна 2018/2019 учебного года, весна 2019/2020 учебного года.
3. Стипендия Эразмус+ - некоммерческая программа Европейского союза по обмену студентами и преподавателями между университетами стран-партнёров. (Erasmus+ programme (KA107))

*Научные стажировки:*

1. Академический обмен в рамках программы Erasmus+ (Испания, г. Мадрид, Мадридский Политехнический Университет)

*Публикации:*

1. Зяблецев П. А. Использование OLAP технологии для выявления группы риска острого инфаркта миокарда // Молодежь и современные информационные технологии: сборник трудов XVI Международной научно-

практической конференции студентов, аспирантов и молодых ученых, Томск, 3-7 Декабря 2018. - Томск: ТПУ, 2019 - С. 154-155

2. Зяблецев П. А. Интеллектуальный анализ данных для определения группы риска сердечно сосудистых заболеваний // Молодежь и современные информационные технологии: сборник трудов XVI Международной научно-практической конференции студентов, аспирантов и молодых ученых, Томск, 3-7 Декабря 2018. - Томск: ТПУ, 2019 - С. 152-153

3. Журбич Н. И., Зяблецев П. А. Анализ данных с помощью технологий Big Data // Информационные технологии в науке, управлении, социальной сфере и медицине: сборник научных трудов V Международной конференции: в 2 т., Томск, 17-21 Декабря 2018. - Томск: ТПУ, 2018 - Т. 1 - С. 255-257.

4. Зяблецев П. А. Анализ многомерных данных для определения группы риска сердечно-сосудистых заболеваний // Информационные технологии в науке, управлении, социальной сфере и медицине: сборник научных трудов V Международной конференции: в 2 т., Томск, 17-21 Декабря 2018. - Томск: ТПУ, 2018 - Т. 1 - С. 263-266

5. Журбич Н. И., Зяблецев П. А. Анализ социальных данных с помощью технологий Big Data // Молодежь и современные информационные технологии: сборник трудов XVI Международной научно- практической конференции студентов, аспирантов и молодых ученых, Томск, 3-7 Декабря 2018. - Томск: ТПУ, 2019 - С. 148-149.

6. Журбич Н. И., Зяблецев П. А. Разработка виртуального полигона в Unity 3D // Информационные технологии в науке, управлении, социальной сфере и медицине: сборник научных трудов V Международной конференции: в 2 т., Томск, 17-21 Декабря 2018. - Томск: ТПУ, 2018 - Т. 1 - С. 252-255

7. Зяблецев П. А., Журбич Н. И. Выявление факторов риска острого инфаркта миокарда с помощью OLAP технологии // Информационные технологии в науке, управлении, социальной сфере и медицине: сборник научных трудов V Международной конференции: в 2 т., Томск, 17-21 Декабря 2018. - Томск: ТПУ, 2018 - Т. 1 - С. 267-270.

## Список литературы


1) В.А. Шевченко Прогнозирование успеваемости студентов на основе методов кластерного анализа // Вестник ХНАДУ. 2015. №68. URL: https://cyberleninka.ru/article/n/prognozirovanie-uspevaemosti-studentov-na-osnove-metodov-klasternogo-analiza (дата обращения: 18.02.2020).
2) ПРОГНОЗИРОВАНИЕ ПЕРСОНАЛЬНОЙ УСПЕВАЕМОСТИ СТУДЕНТОВ В ВУЗЕ. Будаева А.А. В сборнике: IT-Технологии: развитие и приложения XV Ежегодная Международная научно-техническая конференция: Сборник докладов. 2018. С. 9-16.
3) Al-Shehri H. et al. Student performance prediction using support vector machine and k-nearest neighbor //2017 IEEE 30th Canadian Conference on Electrical and Computer Engineering (CCECE). – IEEE, 2017. – С. 1-4.
4) Ясинский И.Ф. Опыт прогнозирования успеваемости студентов при помощи нейросетевой технологии // Вестник ИГЭУ. – 2007. – № 4. – С. 1 – 4.
5) Прошкина Е. Н., Балашова И. Ю. Анализ и прогнозирование успеваемости студентов на основе радиальной базисной нейронной сети [Текст] // Технические науки: традиции и инновации: материалы III Междунар. науч. конф. (г. Казань, март 2018 г.). — Казань: Молодой ученый, 2018. — С. 24-28. — URL https://moluch.ru/conf/tech/archive/287/13683/ (дата обращения: 22.02.2020).
6) Харламова И.Ю. Прогнозирование успеваемости студентов первого курса по результатам сдачи единого государственного экзамена // Базис. – 2017. – № 1(1). – С. 57 – 59.
7) Губин Е.И. Методика подготовки больших данных для прогнозного анализа //Наука и бизнес: Пути развития, № 3(105) 2020, с. 27-31
8) Моисеев Василий Борисович, Зубков Александр Фёдорович, Деркаченко Валентин Николаевич Прогнозирование успеваемости студентов по общепрофессиональным и специальным дисциплинам на основе регрессионных моделей // Научно-технические ведомости Санкт-Петербургского государственного политехнического университета. Информатика, телекоммуникации и управление. 2010. №6 (113). URL: https://cyberleninka.ru/article/n/prognozirovanie-uspevaemosti-studentov-po-obscheprofessionalnym-i-spetsialnym-distsiplinam-na-osnove-regressionnyh-modeley (дата обращения: 20.02.2020).
9) Апатова Н.В., Гапонов А.И., Майорова А.Н. ПРОГНОЗИРОВАНИЕ УСПЕВАЕМОСТИ СТУДЕНТОВ НА ОСНОВЕ НЕЧЕТКОЙ ЛОГИКИ // Современные наукоемкие технологии. – 2017. – № 4. – С. 7-11; URL: http://top-technologies.ru/ru/article/view?id=36630 (дата обращения: 20.02.2020).
10) РУСАКОВ, Сергей Владимирович; РУСАКОВА, Ольга Леонидовна; ПОСОХИНА, Кристина Андреевна. НЕЙРОСЕТЕВАЯ МОДЕЛЬ ПРОГНОЗИРОВАНИЯ ГРУППЫ РИСКА ПО УСПЕВАЕМОСТИ СТУДЕНТОВ ПЕРВОГО КУРСА. Международный научный журнал «Современные информационные технологии и ИТ-образование», [S.l.], v. 14, n. 4, p. 815-822, dec. 2018. ISSN 2411-1473. Доступно на: <http://sitito.cs.msu.ru/index.php/SITITO/article/view/479>. (дата обращения: 20.02.2020).
11) Breazley, D. Python Cookbook, Third Edition / D. Breasley, B. K. Jones. – USA: O'Reilly Media, 2013. – 688 p.
12) McKinney, W. Python for Data Analysis. – USA: O'Reilly Media, 2013. – 453 p.
13) Трудовой кодекс Российской Федерации от 30.12.2001 N 197-ФЗ (ред. от 01.04.2019).
14) Федеральный Закон от 27.07.2006 N 152-ФЗ (ред. от 25.07.2011) «О Персональных Данных».
15) ГОСТ 12.1.005-88 Система стандартов безопасности труда (ССБТ). Общие санитарно-гигиенические требования к воздуху рабочей зоны (с Изменением N 1).

16) ГОСТ 12.0.003-74 Система стандартов безопасности труда (ССБТ). Опасные и вредные производственные факторы. Классификация (с Изменением N 1).
17) СанПиН 2.2.2/2.4.1340-03. Гигиенические требования к персональным электронно-вычислительным машинам и организации работы.
18) ГОСТ Р 50571.17-2000. Электроустановки зданий. Часть 4. Требования по обеспечению безопасности. Глава 48. Выбор мер защиты в зависимости от внешних условий. Раздел 482. Защита от пожара.
19) СанПиН 2.2.1/2.1.1.1278-03. Гигиенические требования к естественному, искусственному и совмещенному освещению жилых и общественных зданий.
20) СанПиН 2.2.4.3359-16. Санитарно-эпидемиологические требования к физическим факторам на рабочих местах.
21) ТОИ Р-45-084-01. Типовая инструкция по охране труда при работе на персональном компьютере.
22) ГОСТ Р 56397-2015. Техническая экспертиза работоспособности радиоэлектронной аппаратуры, оборудования информационных технологий, электрических машин и приборов. Общие требования.
23) ГОСТ 12.3.031-83. Система стандартов безопасности труда. Работы со ртутью. Требования безопасности.
24) ГОСТ12.1.044-89. Система стандартов безопасности труда. Пожаровзрывоопасность веществ и материалов. Номенклатура показателей и методы их определения.
25) ГОСТ 12.2.032-78. Система стандартов безопасности труда. Рабочее место при выполнении работ сидя. Общие эргономические требования.
26) ГОСТ 12.1.045–84 ССБТ. Электростатические поля. Допустимые уровни на рабочих местах и требования к проведению контроля.
27) ГОСТ12.2.061-81. Система стандартов безопасности труда. Оборудование производственное. Общие требования безопасности к рабочим местам.
28) Трудовой кодекс Российской Федерации от 30.12.2001 N 197-ФЗ (ред. от 05.02.2018)
29) ГОСТ 26522-85 Короткие замыкания в электроустановках. Термины и определения.
30) ГОСТ 12.4.124-83 Система стандартов безопасности труда (ССБТ). Средства защиты от статического электричества. Общие технические требования
31) Безопасность жизнедеятельности: практикум / Ю.В. Бородин, М.В. Василевский, А.Г. Дашковский, О.Б. Назаренко, Ю.Ф. Свиридов, Н.А. Чулков, Ю.М. Федорчук. — Томск: Изд-во Томского политехнического университета, 2009. — 101 с.

**Приложение А**
(справочное)

# Overview of Student Performance Forecasting Methods

Студент

| Группа | ФИО | Подпись | Дата |
|---|---|---|---|
| 8ПМ8И | Зяблецев Павел Андреевич | | |

Руководитель ВКР

| Должность | ФИО | Ученая степень, звание | Подпись | Дата |
|---|---|---|---|---|
| доцент ОИТ ИШИТР | Губин Е.И. | к.ф.-м.н. | | |

Консультант – лингвист отделения иностранных языков ШБИП

| Должность | ФИО | Ученая степень, звание | Подпись | Дата |
|---|---|---|---|---|
| Старший преподаватель ОИЯ ШБИП | Пичугова И.Л. | – | | |

## Introduction

The grades of student performance at the university is a kind of indicator, which shows the future specialist’s level of professionalism. Alternately, the success of students is an indicator of the university's activity in solving educational problems. In this case, it is needed constant objective assessment, adjustment and management to solve these problems as efficiently as possible. However, control is impossible without forecasting. Therefore, there is a need to predict student performance at all stages of their studying.

Having information about those students who are most likely to have academic debts by the end of the semester if they do not change the current trend, we can influence students, thereby improving their academic performance.

The aim of this master's thesis is to create a forecasting model of student performance at TPU. The presence of such a model will make it possible to pay closer attention to students who fall into the risk group of a large number of debts in academic disciplines, and as a result, they will be applicants for expulsion. The identification of such students in the early stages will allow to do more detailed and personal work with them so that they can overcome this study load more successfully.

Based on the goal, this work includes the following tasks: reviewing the literature on this issue, studying the methods and algorithms used, cleaning and preparing the source data, developing a predictive model, testing the results, creating a graphical user interface for the student performance forecasting module.

## 1. Domain Analysis

Education plays one of the most important roles in any nation. The quality of education, existing in a particular society, has highly influence on the pace of its economic and political development, its moral condition. The rapid development of information technology allows to automatize many areas of people's activities increasing their effectiveness, and education is no exception. In this paper, we will focus on creating a forecasting model of student performance according to current estimates using data analysis technologies.

For a particular student the measure of the quality of education is his assessment of the completed subjects. If we talk about the educational institution, then one of the measures of the quality of the education provided by him is a set of assessments of his students. Taking measures to assist students, who cannot cope with the workload, in time is one of the main parts of the educational work at the university, which affects the quality of education.

Recently, many different changes have been made to improve the quality of education in universities. For example, the transition from the traditional grading system to point system eliminates the possibility that a student who has not attended classes for the whole semester will simply come and pass an exam. For admission to the exam, he needs to score a certain number of points, and for this, in turn, he needs to attend classes and perform current tasks. This approach effectively influences on the understanding of educational material. According to many studies, information, that has been studied over a long period of time, will remain for a long time, while cramming the night before the exam can only give a good result in the exam, which will eventually be passed, but the student will not have any residual knowledge of subject at all. In addition, there is a midterm control, a fixed date in the middle of the semester, when a certain part of the material must be delivered. However, knowing the specifics of student life, many students still leave everything till the last moment. Some students manage to pass the subject, while others remain in debt for another semester.

The problem is that the controlling influence of the unified dean's office of Tomsk Polytechnic University begins only after the fact that the student has already had a debt. Thus, these measures cannot be called preventive. The main task of the forecast model is to highlight such students and conduct certain discussions with them until the problem arises in the form of academic debt. Currently, this measure has been implemented partly by the fact that each group of freshmen has a curator and this curator accompanies each group up to the 2nd year, the schedule includes such event as "curator's hour", where he analyzes students' academic performance. This is an excellent practice and it is impossible to abandon it, however, the human factor plays a role here. The curator may not always convey to students the importance of attending classes. The introduction of a new system for predicting the amount of failed academic assignments at the end of the semester by current performance is an important step in the automation of the educational process. Thus, a unified dean's office will be able to exert a controlling influence on students who are at risk, much earlier than the real problem will begin. Thus, it will be possible to help students finish the semester successfully and to form professional competencies, and those who are not interested in studying, to expel earlier and create spots for those who really want and are ready to receive knowledge.

The main attention is paid to the forecasting system, since it is not just about controlling student attendance. Data analysis shows that the end result of a semester is influenced not only by one performance factor, but by a whole complex of different data. For example, there is a whole category of students who are already working in their specialty, mainly undergraduates and final year students. Such students do not always have the opportunity to attend a lesson and nevertheless they show excellent results at the end of the session, due to the fact that they understand many aspects of the profession at a higher level than their classmates.

In fact, a huge number of factors affect student performance, and one of the most important is the motivation to study, morale, relations with classmates, and so on. However, due to the fact that the system will give out problematic students and it will be possible for each to work in detail, to identify which problems exist precisely

with this student, who runs the risk of ending a semester with a lot of debts. Identification of such parameters as motivation, determination, psychological data is possible, but this requires a large number of tests, which should be carried out on a regular basis. These methods do not have high efficiency due to the complexity of their implementation, as well as verification and interpretation of the results.

After finding out that this problem is really relevant, it is worth considering the existing methods for solving it.

At the moment, there are no open access materials on the introduction and use of the student performance forecasting system at a university. Although this idea is not innovative, from 2010s articles have been published on forecasting student performance and forecasting student performance at a particular course, where the following parameters are usually taken as initial data: level of current knowledge in the subject, number of passes, intermediate control, assessments of courses for discipline.

Next, we consider machine learning algorithms that are used to solve similar problems in the field of forecasting student performance. The most common methods and algorithms will be given, as well as a description of existing work with specific tasks for which these methods are applied.

Some sources use the clustering method, in our opinion this is not entirely true, since we already have class labels, namely the amount of debts, so in this case, classification methods should be applied. Clustering refers to methods of machine learning without a teacher, and classification and regression, in turn, to learning with a teacher, since they have markers for each class as source data. Cluster analysis methods for assessing the final grade for a subject were applied in this paper. [1]

### 1.1 K-Nearest Neighbor Algorithm

One of the most common methods for solving a similar problem is the K-Nearest Neighbor Algorithm:

The k-nearest neighbors algorithm (KNN, k-nearest neighbors) is a type of managed machine learning algorithm that can be used both for classification and for regression of prediction tasks. This algorithm is one of the easiest to understand, but nevertheless it has proved its effectiveness in a number of tasks and is used not only for educational purposes. The algorithm of the closest neighbors is also called the "lazy" classifier, because in the process of learning it does not build any model, but simply stores data. All calculations begin only when it is necessary to classify new data.

The essence of this algorithm is that forecasting the value of new data is based on their proximity to already marked data in the training set. In other words, if a new data point has 4 class A points and 1 class B point among its nearest neighbors, then this new point will be defined as class A. Thus, the k-nearest neighbors algorithm has 2 most important parameters, namely the distance metric (Euclidean , Manhattan or Hamming) and the number of neighbors that we will consider. A visual representation of the algorithm is presented in Figure 1.

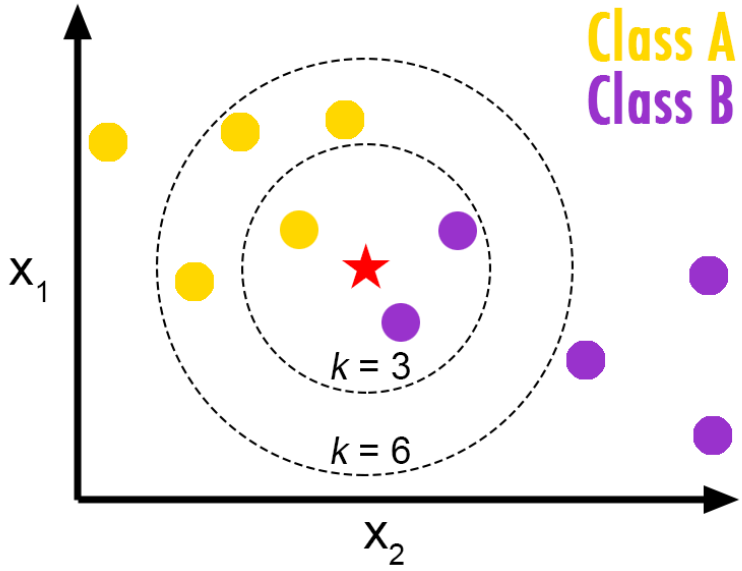



Figure 1. Operation of the KNN algorithm

This figure shows an array of starting points, which are color-coded depending on class membership. An asterisk indicates the point to be classified. All points are presented in two-dimensional form, along the axes $X_1$ and $X_2$. If the set number of

nearest neighbors is 3, then the unmarked object will belong to class B, if 6, then to class A.

This algorithm has many nuances, for example, we can add weight to the data when voting, depending on the proximity to our unmarked object. Exactly these nuances make the KNN algorithm relevant for solving a whole range of problems.

Advantages of this algorithm are as follows:

- Ease of understanding and interpretation.
- Works well on non-linear data.
- Universal algorithm, suitable for both classification and regression problems.
- Has a relatively high accuracy.

Disadvantages of the algorithm are as follows:

- Large memory consumption for storing all data, unlike algorithms that use model building.
- Sensitivity to data scale.
- Slow prediction in case of a large amount of data.
- High sensitivity to “noise” in the data.

In the article "Prediction of personal performance of students in high school" by A. Budaeva [2] this algorithm is used to classify the individual student's grades in each subject based on his past grades and grades of past students, with the most similar parameters in these subjects. This article provides a sufficiently high accuracy of the algorithm for this task, the maximum error of the forecast was 0.55 points. However, only 307 students of the same specialty were taken into account and there was no talk of applying this methodology to the university system.

Also, this algorithm is used in the article “Student performance prediction using support vector machine and k-nearest neighbor” [3], where the student is predicted to score for an exam in a particular subject, based on his grades in previous subjects, his attendance and estimates of intermediate control.

### 1.2 Support vector machines method

The support vector machines method (SVM) is a set of similar algorithms of the “learning with a teacher” type, which are used to solve classification and regression problems. This method is one of the most popular teaching methods and belongs to the family of linear classifiers. A special property of the support vector method is a continuous decrease in the empirical classification error and an increase in the gap. Therefore, this method is also known as a method of the classifier with the maximum gap.

The basic idea of the support vector method can be illustrated by an example: on the plane there are points marked in 2 classes that are linearly separable. In this case, the resulting function will be the plane that separates these classes. However, it is possible to draw many hyperplanes that separate these classes. In order to find the optimal hyperplane, it is necessary to find the maximum sum of normal vectors from class A and class B. The visual display of this method can be seen in Figure 2. In this figure, the support vectors are perpendicular to the normals, which are shown in Figure 2.

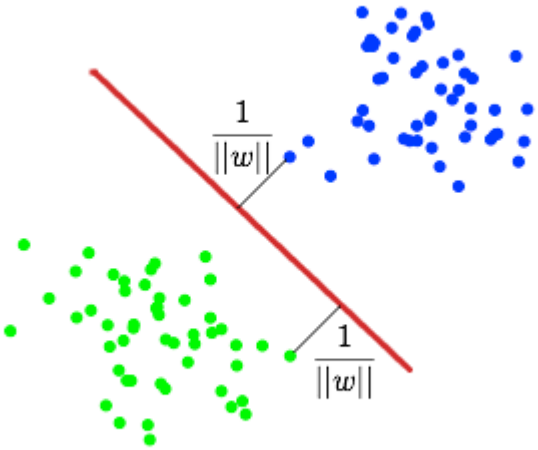



Figure 2. The method of support vector machines

The formal description of this method is as follows: suppose we have a training sample $\{(X_1, C_1), (X_2, C_2), \ldots, (X_i, C_i)\}$, where:

$X_i$ – a p-dimensional real vector

$C_i$ – a value of 1 or -1 that the class takes

The support vector method constructs a classification function in the form:

$$F(x) = sign([w, x] + b), \text{ (1) где,}$$

[ , ] – scalar product

$w$ – normal vector to the dividing hyperplane

$b$ – auxiliary parameter

Thus, all of this can be written in the form of an optimization problem that has a solution and the only one.

$$\begin{cases} \|\mathbf{w}\|^2 \to \min \\ c_i(\mathbf{w} \cdot \mathbf{x_i} - b) \geq 1, \quad 1 \leq i \leq n. \end{cases} \quad (2)$$

This problem is solved by the method of quadratic programming and with the help of Lagrange multipliers.

The case where there are 2 separable classes was considered. In practice, almost always classes are not linearly separable and the task is to classify more than 2 classes. To solve the problem with linearly inseparable classes, we allow the algorithm to make mistakes on the training set. Let us write down the equations for this assumption.

$$\begin{cases} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n} \xi_i \to \min_{w,b,\xi_i} \\ c_i(\mathbf{w} \cdot \mathbf{x_i} - b) \geq 1 - \xi_i, \quad 1 \leq i \leq n \\ \xi_i \geq 0, \quad 1 \leq i \leq n \end{cases} \quad (3)$$

where C – method setting parameter,

$\mathcal{E}_i$ – margin of error.

To solve multiclass problems, a generalized method of support vectors is used due to the fact that the transition to classification into many classes is carried out by splitting into 2 classes, such as a suitable class and not a suitable one. This strategy is also called “one against all” and is used to apply binary classifiers for multiclass problems.

Advantages of the algorithm are as follows:

- The problem is well studied and has a unique solution.
- The principle of optimal dividing hyperplane leads to a confident classification.
- Equivalent to a two-layer neural network, where the number of neurons in the hidden layer is automatically determined as the number of reference vectors.

There are some disadvantages:

- Noise instability, outliers in the source data directly affect the construction of the separating hyperplane.
- No feature selection.
- It is necessary to select methods for constructing kernels and rectifying spaces separately for each task.

This algorithm is used in the article “Student performance prediction using support vector machine and k-nearest neighbor” [3], where the student is predicted to score for an exam in a particular subject based on his grades in previous subjects, his attendance, and intermediate control scores.

### 1.3 Neural networks

In addition, neural networks are used to predict student performance. There are a lot of references on the possibility of using neural networks to solve this problem, however, information on the actual implementation of such predictive models is not found.

Neural networks are mathematical models built on the principle of organization and functioning of biological neural networks. Neural networks are not programmed in the usual sense of the word, they are trained. In the learning process, the neural network is able to identify complex relationships between input and output, as well

as perform generalization. After training, the network is able to predict the future value of a certain sequence based on several previous values.

An illustration of the principle of operation of a neural network is shown in Figure 3.

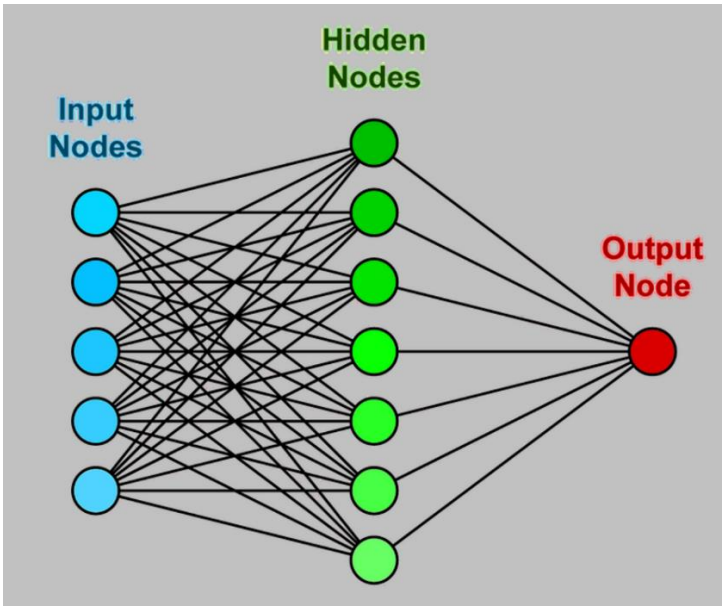



Figure 3. Neural network.

A neural network consists of neurons, layers, and synapses. Neurons are depicted as nodes of different colors. All nodes of the same color belong to one layer of the neural network. Synapses are lines that connect the neurons of one layer to the neurons of another layer. Synapse has only one parameter and this is its weight.

Each neuron performs a certain mathematical function, so it receives a lot of values at the input, and one at the output. Thus, at the output, a certain value is obtained that the already trained neural network produced.

There are some advantages:

- Immunity to input noise.
- Self-training and “creativity”. The ability to solve problems that cannot be solved by other algorithms.
- Adaptation to change, retraining.

There are some disadvantages:

- For large networks, the inability to even approximately estimate the network training time in advance.

- Difficulty in interpreting the result.
- The approximate response received.

The use of a neural network to solve the problem of predicting student performance is required to be considered in more detail. In the article "Experience in predicting student performance using neural network technology" by Yasinsky I.F. [4] the neural network model is trained to predict whether a student will be promising. The problem of binary classification of applicants by such input data as the school number, grades in physics and mathematics, and the profession of parents is being solved.

In the article “Analysis and forecasting of student performance based on the radial basis neural network” [5], the neural network is used to predict grades at the computer science course.

In the article "Predicting the performance of first-year students based on the results of the unified state exam" by I. Kharlamova [6] the task of classifying students according to the results of the exam is considered.

**Conclusion**

We have examined the algorithms that are most often used to solve the problem of forecasting student performance based on various input data and even in different subjects, whether it is performance in a particular subject or the overall picture of performance in all disciplines. The considered forecasting examples are not systemic in nature, but are only attempts to get closer to solving this problem. Obviously, it is necessary to apply several methods and compare the results in order to solve the problem successfully, since we do not know in advance the relationships between the data. In addition, it has been revealed that the linear dependencies are best handled by such an algorithm as the support vector method, and the implicit dependencies by neural networks or the random forest algorithm.

## Приложение Б. Листинг кода по предварительной подготовке данных.

```
import pandas as pd
import seaborn as sns

students_df = pd.read_excel('ID_Pavel.xlsx')

students_df.head(10)

# функция отнесения студента в группу в зависимости от колличества долгов
def label_student (n_fails):
    if n_fails == 0:
        return("успешно")
    if n_fails > 0 and n_fails <= 6:
        return("долги")
    if n_fails > 6:
        return("много долгов")

# создание дополнительного столбца в дата сете, где указывается группа студента в
зависимости от количества долгов
students_df['группа долгов'] = students_df.apply(lambda row:
label_student(row['Неудовлетворительных']), axis=1)

students_df.info()

students_df.isna().any()

import matplotlib.pyplot as plt
def draw_pie_diagram(name):
    print(name)
    name.plot(kind='pie')
    plt.axis('equal')
    plt.show()

def draw_pie_diagram_advanced(name, lst=[], exp=None):
    print(name)
    if exp != None:
           exp = exp
    fig1, ax1 = plt.subplots(figsize=(12,7))
    ax1.pie(name, explode=exp, labels=lst, autopct='%1.1f%%', shadow=True,
startangle=90)
    # Equal aspect ratio ensures that pie is drawn as a circle
    ax1.axis('equal')
    plt.tight_layout()
    plt.legend()
    plt.show()


print(students_df.groupby(['Неудовлетворительных'])['ID'].count())
print(students_df.groupby(['Неудовлетворительных'])['ID'].count()/len(students_df.in
dex)*100)
f,ax = plt.subplots(figsize=(15,10))
sns.countplot(students_df['Неудовлетворительных'])

stud_sex = students_df.groupby(['Пол'])['ID'].count()
draw_pie_diagram_advanced(stud_sex, ['Женский', 'Мужской'], [0.1, 0])

# sns.countplot(students_df['Форма обучения'])
print(students_df.groupby(['Форма обучения'])['ID'].count())
```

```
print(students_df.groupby([ 'Квалификация', 'Курс'])['ID'].count())
print(students_df.groupby([ 'Квалификация',
'Курс'])['ID'].count()/len(students_df.index)*100)
f,ax = plt.subplots(figsize=(15,10))
sns.countplot(students_df['Квалификация'],hue=students_df['Курс'])

print(students_df.groupby([ 'Форма обучения', 'Квалификация'])['ID'].count())
print(students_df.groupby([ 'Форма обучения',
'Квалификация'])['ID'].count()/len(students_df.index))
f,ax = plt.subplots(figsize=(15,10))
sns.countplot(students_df['Квалификация'],hue=students_df['Форма обучения'])

print(students_df.groupby([ 'Форма обучения', 'Квалификация', 'группа
долгов'])['ID'].count())
print(students_df.groupby([ 'Форма обучения', 'Квалификация', 'группа
долгов'])['ID'].count()/len(students_df.index))

print(students_df.groupby([ 'группа долгов'])['ID'].count())
print(students_df.groupby([ 'группа долгов'])['ID'].count()/len(students_df.index))
f,ax = plt.subplots(figsize=(15,10))
sns.countplot(students_df[ 'группа долгов'])

def draw_diagramm(kval, form):
    print("Диаграмма процентного соотношения студентов с долгами и без для группы:
"+ str(kval)+ ', ' + str(form) +' форма обучения')
    stud_debts = students_df[(students_df['Форма обучения']==form) &
(students_df['Квалификация']==kval)].groupby(['группа долгов'])['ID'].count()
    draw_pie_diagram_advanced(stud_debts, ['долги', 'много долгов', 'успешно'])

# kval_list = ["Магистр"]
# form_list = ['Очная', "Очно-заочная"]
# for kval in kval_list:
#     for form in form_list:
#         draw_diagramm(kval, form)


# print("Диаграмма процентного соотношения студентов с долгами и без для группы:
Магистр, Очно-заочная форма обучения")
# stud_debts = students_df[(students_df['Форма обучения']=='Очно-заочная') &
(students_df['Квалификация']=='Магистр')].groupby(['группа долгов'])['ID'].count()
# draw_pie_diagram_advanced(stud_debts, ['долги', 'много долгов'])

stud_debts = students_df[students_df.Квалификация=='Бакалавр'].groupby(['группа
долгов'])['ID'].count()
draw_pie_diagram_advanced(stud_debts, ['долги', 'много долгов', 'успешно'])

pr1 = students_df.groupby([ 'Специальность'])['Неудовлетворительных'].median()
# print(pr1)
# type(pr1)
pr1.nlargest(15)
# pr1.nsmallest(10)

f,ax = plt.subplots(figsize=(15,10))
sns.countplot(students_df['Выпуск. школа'],hue=students_df['группа долгов'])



students_df = students_df[students_df['Всего часов аудиторных занятий в
семестре'].notnull()]
```

```
Y = students_df['Неудовлетворительных']


X = students_df[['Форма обучения', 'Квалификация', 'Курс', 'Специальность',
        'Академ отпуск (действующий) - да / нет', 'Всего часов пропусков в
семестре',
       'Всего часов аудиторных занятий в семестре']]
Y = students_df['группа долгов']

from sklearn.preprocessing import LabelEncoder
le_f = LabelEncoder()
X['Форма обучения'] = le_f.fit_transform(X['Форма обучения'].values)
le_k = LabelEncoder()
X['Квалификация'] = le_k.fit_transform(X['Квалификация'].values)
le_s = LabelEncoder()
X['Специальность'] = le_s.fit_transform(X['Специальность'].values)
le_a = LabelEncoder()
X['Академ отпуск (действующий) - да / нет'] = le_a.fit_transform(X['Академ отпуск
(действующий) - да / нет'].values)

X.info()

sns.set(font_scale= 1)
hm = sns.heatmap(X.corr(), cbar=True, annot=True)
plt.show()

X.head()

from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test = train_test_split(X, Y, test_size=0.3,
random_state=0)

from sklearn.linear_model import LogisticRegression
lr = LogisticRegression(C=1000.0, random_state=0 )
lr.fit(x_train,y_train)


print("Тренеровочная: {0}".format(lr.score(x_train, y_train)))
print("Тестовая: {0}".format(lr.score(x_test, y_test)))


pred_y = lr.predict(x_test)
from sklearn import metrics
# Кросс- валидация модели
print(metrics.classification_report(pred_y, y_test))


from sklearn import svm
clf = svm.SVC()
clf.fit(x_train,y_train)

print("Тренеровочная: {0}".format(clf.score(x_train, y_train)))
print("Тестовая: {0}".format(clf.score(x_test, y_test)))


pred_y = clf.predict(x_test)
from sklearn import metrics
# Model Accuracy, how often is the classifier correct?
print(metrics.classification_report(pred_y, y_test))

from sklearn.ensemble import RandomForestClassifier
```

```
clas = RandomForestClassifier(max_depth=15, random_state=0)
clas.fit(x_train,y_train)



print(clas.feature_importances_)

print(clas.score(x_train, y_train))
print(clas.score(x_test, y_test))


pred_y = clas.predict(x_test)
from sklearn import metrics
# Model Accuracy, how often is the classifier correct?
print(metrics.classification_report(pred_y, y_test))


from sklearn.model_selection import train_test_split
X = students_df[['Курс']].values.astype(int)
y = students_df['Неудовлетворительных'].values.astype(int)

import numpy as np
from sklearn.linear_model import LinearRegression
model = LinearRegression()
model.fit(X, y)
y_pred = model.predict(X)
print('Slope: {:.2f}'.format(model.coef_[0]))
print('Intercept: {:.2f}'.format(model.intercept_))


plt.scatter(X, y)
plt.plot(X, model.predict(X), color='red', linewidth=2);

model = LinearRegression()
model.fit(x_train, y_train)
y_pred = model.predict(x_test)
print('Slope: {:.2f}'.format(model.coef_[0]))
print('Intercept: {:.2f}'.format(model.intercept_))

print(model.score(x_train, y_train))
print(model.score(x_test, y_test))


test_pred_y = model.predict(x_test)
train_pred_y = model.predict(x_train)
from sklearn import metrics
# Model Accuracy, how often is the classifier correct?
# print(metrics.classification_report(pred_y, y_test))

from sklearn.metrics import mean_absolute_error, mean_squared_error,
median_absolute_error, r2_score

print('MSE train: {:.3f}, test: {:.3f}'.format(
        mean_squared_error(y_train, train_pred_y),
        mean_squared_error(y_test, test_pred_y)))
print('R^2 train: {:.3f}, test: {:.3f}'.format(
        r2_score(y_train, train_pred_y),
        r2_score(y_test, test_pred_y)))

X.info()
```

```
xx = X[['Курс', 'Специальность', 'Всего часов пропусков в семестре', 'Всего часов
аудиторных занятий в семестре']]

from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test = train_test_split(xx, Y, test_size=0.3,
random_state=0)


from sklearn.ensemble import RandomForestClassifier
clas = RandomForestClassifier(max_depth=15, random_state=0)
clas.fit(x_train,y_train)

print(clas.feature_importances_)

print(clas.score(x_train, y_train))
print(clas.score(x_test, y_test))

pred_y = clas.predict(x_test)
from sklearn import metrics
# Model Accuracy, how often is the classifier correct?
print(metrics.classification_report(pred_y, y_test))

super_df = students_df
from sklearn.preprocessing import LabelEncoder
le_f = LabelEncoder()
super_df['Форма обучения'] = le_f.fit_transform(super_df['Форма обучения'].values)
le_k = LabelEncoder()
super_df['Квалификация'] = le_k.fit_transform(super_df['Квалификация'].values)
le_s = LabelEncoder()
super_df['Специальность'] = le_s.fit_transform(super_df['Специальность'].values)
le_a = LabelEncoder()


for i in range(1, 6):
    print("Для {0} курса ".format(i))
    u_df = super_df[students_df['Курс']==i]
    u_df['Академ отпуск (действующий) - да / нет'] = le_a.fit_transform(u_df['Академ
отпуск (действующий) - да / нет'].values)
    x1 = u_df[['Форма обучения', 'Квалификация', 'Курс', 'Специальность',
           'Академ отпуск (действующий) - да / нет', 'Всего часов пропусков в
семестре',
           'Всего часов аудиторных занятий в семестре']]
    Y = u_df['группа долгов']
    x_train, x_test, y_train, y_test = train_test_split(x1, Y, test_size=0.3,
random_state=0)
    clas = RandomForestClassifier(max_depth=15, random_state=0)
    clas.fit(x_train,y_train)
    print(clas.score(x_train, y_train))
    print(clas.score(x_test, y_test))
    pred_y = clas.predict(x_test)
    from sklearn import metrics
    # Model Accuracy, how often is the classifier correct?
    print(metrics.classification_report(pred_y, y_test))


for i in le_k.transform(list(le_k.classes_)):
    print("Для {0}".format(le_k.inverse_transform(i)))
    u_df = super_df[students_df['Квалификация']==i]
    u_df['Академ отпуск (действующий) - да / нет'] = le_a.fit_transform(u_df['Академ
отпуск (действующий) - да / нет'].values)
    x1 = u_df[['Форма обучения', 'Квалификация', 'Курс', 'Специальность',
```

```
            'Академ отпуск (действующий) - да / нет', 'Всего часов пропусков в
семестре',
           'Всего часов аудиторных занятий в семестре']]
    Y = u_df['группа долгов']
    x_train, x_test, y_train, y_test = train_test_split(x1, Y, test_size=0.3,
random_state=0)
    clas = RandomForestClassifier(max_depth=15, random_state=0)
    clas.fit(x_train,y_train)
    print(clas.score(x_train, y_train))
    print(clas.score(x_test, y_test))
    pred_y = clas.predict(x_test)
    from sklearn import metrics
    # Model Accuracy, how often is the classifier correct?
    print(metrics.classification_report(pred_y, y_test))


x_train, x_test, y_train, y_test = train_test_split(x1, Y, test_size=0.3,
random_state=0)
clas = RandomForestClassifier(max_depth=15, random_state=0)
clas.fit(x_train,y_train)
print(clas.score(x_train, y_train))
print(clas.score(x_test, y_test))
pred_y = clas.predict(x_test)
from sklearn import metrics
# Model Accuracy, how often is the classifier correct?
print(metrics.classification_report(pred_y, y_test))


students_df.head(10)

subj_df = students_df[students_df['Дисциплины по которым получены
неудовлетворительные оценки'].notnull()]
subj_df = subj_df[subj_df['Специальность']=='38.03.01 Экономика']

subj_df.head()

subject_dict = {}
for i in subj_df['Дисциплины по которым получены неудовлетворительные оценки']:
    clear_data = i.replace(', ', ',').split(',')
    for j in clear_data:
        if j in subject_dict:
            subject_dict[j] = (subject_dict[j]+1)
        else:
            subject_dict[j] = 1
print(subject_dict)


from collections import OrderedDict
d_sorted_by_value = OrderedDict(sorted(subject_dict.items(), key=lambda x: x[1]))


from beautifultable import BeautifulTable
table = BeautifulTable()
table.column_headers = ['Название предмета', 'Количество должников']

count = 0
for key in reversed(d_sorted_by_value):
    table.append_row([key, d_sorted_by_value[key]])
    count += 1
    if count == 15:
        break
```

```
print(table)


# функция отнесения студента в группу в зависимости от колличества долгов
def label_student1(missings, lessons):
    p = missings/lessons
    return p


students_df['пропуски'] = students_df.apply(lambda row: label_student1(row['Всего
часов пропусков в семестре'],
                                                                       row['Всего
часов аудиторных занятий в семестре']), axis=1)


j_df = students_df[students_df['пропуски'] < 0.1]


j_df.head(15)


# plt.scatter(students_df['пропуски'], students_df['Неудовлетворительных'])
plt.plot(students_df['Неудовлетворительных'], students_df['пропуски'], color='red',
linewidth=2);
```