МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное автономное образовательное учреждение
высшего образования
«Южно-Уральский государственный университет
(национальный исследовательский университет)»
Высшая школа электроники и компьютерных наук
Кафедра системного программирования

ДОПУСТИТЬ К ЗАЩИТЕ

Заведующий кафедрой, д.ф.-м.н.,
профессор

_____ Л.Б. Соколинский

«___»_____ 2024 г.

Разработка системы для предсказания успеваемости студентов на основе данных портала «Электронный ЮУрГУ»

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА
ЮУрГУ – 09.03.04.2024.308-569.ВКР

Научный руководитель,
профессор кафедры СП, д.ф.-м.н.,
доцент
_____ М.Л. Цымблер

Автор работы,
студент группы КЭ-403
_____ Д.В. Старостенок

Ученый секретарь
(нормоконтролер)
_____ И.Д. Володченко
«___»_____ 2024 г.

Челябинск, 2024 г.

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное автономное образовательное учреждение
высшего образования
«Южно-Уральский государственный университет
(национальный исследовательский университет)»
Высшая школа электроники и компьютерных наук
Кафедра системного программирования

УТВЕРЖДАЮ
Зав. кафедрой СП
_____ Л.Б. Соколинский
29.01.2024 г.

ЗАДАНИЕ
на выполнение выпускной квалификационной работы бакалавра

студенту группы КЭ-403
Старостенку Дмитрию Владимировичу,
обучающемуся по направлению
09.03.04 «Программная инженерия»

1. **Тема работы** (утверждена приказом ректора от 22.04.2024 г. № 764-13/12) Разработка системы для предсказания успеваемости студентов на основе данных портала «Электронный ЮУрГУ».
2. **Срок сдачи студентом законченной работы:** 03.06.2024 г.
3. **Исходные данные к работе**

3.1. Документация API «Электронного ЮУрГУ» DigitalTrace. [Электронный ресурс] URL: https://drive.google.com/drive/folders/1r3ZlV1grjMEfvKkqCSgp249zLo9mhuvD (дата обращения: 10.02.2024 г.).
3.2. VanderPlas J. Python Data Science Handbook: Essential Tools for Working with Data. // O'Reilly Media 2nd edition, 2022. – P. 591.

4. **Перечень подлежащих разработке вопросов**

4.1. Провести анализ и выполнить спецификацию предметной области.
4.2. Разработать модели для предсказания успеваемости.
4.3. Спроектировать интерфейс пользователя и модульную структуру приложения.
4.4. Выполнить реализацию, отладку и тестирование приложения. Провести вычислительные эксперименты по исследованию точности предсказания.

5. **Дата выдачи задания:** 29.01.2024 г.

Научный руководитель
профессор кафедры СП, д.ф.-м.н., доцент М.Л. Цымблер

Задание принял к исполнению Д.В. Старостенок

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

Актуальность

Образование стало незаменимой составляющей жизни современного общества. В связи с этим, повышение эффективности образовательного процесса и улучшение качества обучения являются одними из важнейших задач в данной области. Данные «Электронного ЮУрГУ», которые содержат информацию об успеваемости студентов, являются ценным ресурсом для анализа и предсказания успеха студентов.

Приложение позволит не только предсказывать успех студентов на основе их текущих оценок и статистики по их учебной деятельности, но и предоставлять студентам и преподавателям инструменты для повышения эффективности обучения.

Такое приложение также может быть полезно для администрации университета, так как оно позволит им отслеживать прогресс студентов и принимать меры для улучшения качества образования. Кроме того, оно может быть использовано для определения факторов, которые влияют на успеваемость студентов, и для разработки соответствующих программ для улучшения образовательной среды.

Постановка задачи

Целью данной работы является разработка системы для предсказания успеваемости студентов на основе данных «Электронного ЮУрГУ». В качестве исходных данных используется информация об абитуриентах и их успеваемости в виде журналов.

Для достижения поставленной цели необходимо решить следующие задачи.

1. Провести анализ предметной области и на его основе разработать модели для предсказания.
2. Разработать алгоритм получения данных с применением существующих методов предобработки.

3. Разработать приложение, выполняющее предсказание на основе моделей с помощью различных методов анализа данных.

4. Провести эксперименты, исследующие точность разработанных моделей данных.

Результатом выполнения данной работы будет создание приложения, которое позволит преподавателям и студентам получать предсказания успеваемости на основе имеющихся данных, что поможет улучшить качество обучения.

Структура и содержание работы

Дипломная работа состоит из введения, пяти разделов, заключения и списка литературы. Объем работы составляет 53 страницы, объем библиографического списка составляет 17 наименований.

В первом разделе, «Обзор работ по тематике исследования», содержится обзор на работы по тематике исследования.

Во втором разделе, «Анализ предметной области», описывается постановка задачи и описание данных, которые будут использоваться для анализа.

В третьем разделе, «Проектирование», определены требования к системе, описаны модели данных и структура приложения.

В четвертом разделе, «Реализация», приведены диаграммы системы и описана реализация компонентов системы, а также описано функциональное тестирование работы интерфейса.

В пятом разделе, «Вычислительные эксперименты», описаны эксперименты с разработанными моделями данных.

В приложении А представлен макет страницы обучения моделей предсказания.

В приложении Б представлен макет страницы итоговой реализации интерфейса обучения моделей предсказания.

В приложении В представлены изображения результатов предсказательных экспериментов.

1. ОБЗОР РАБОТ ПО ТЕМАТИКЕ ИССЛЕДОВАНИЯ

С ростом количества студентов в вузах, администрации университета становится все сложнее контролировать успеваемость каждого студента индивидуально. В связи с этим возникают потребности в создании системы, которая позволит предсказывать успеваемость студентов. Для этого необходимо произвести анализ данных о студентах, их предыдущих успехах и неудачах, а также данных об их прогрессе в течение текущего учебного года.

В работе [1] авторы используют искусственные нейронные сети (ИНС) для предсказания успеваемости студентов. Используется алгоритм обратного распространения ошибки для обучения ИНС и применяются методы статистической оценки для оценки точности модели. Для предобработки данных и извлечения признаков используются стандартные методы машинного обучения, такие как нормализация данных и кодирование категориальных признаков.

Анализ данных реализуется с использованием языка программирования Python и фреймворка TensorFlow. Данные предварительно обрабатываются и извлекаются признаки, после чего модель обучается и оценивается с помощью стандартных методов машинного обучения и статистических методов оценки.

Авторы работы [5] применяют алгоритм градиентного бустинга XGBoost для предсказания успеваемости. Для анализа использовались демографические данные, социально-экономические данные и данные об успеваемости, к которым применяется алгоритм XGBoost для построения регрессионной модели и предсказания успеваемости.

В работе [6] описано исследование, которое было проведено на базе нескольких смешанных курсов в медицинском университете. Для анализа использовались данные из учебной системы Moodle, которые были обработаны и преобразованы для анализа.

Анализ проводился с использованием смешанных линейных моделей, он выявил, что активное участие студентов в онлайн-обучении, такое, как

регулярные обсуждения на форуме и доступ к учебным материалам, оказывает значительное влияние на их академический успех.

Работа [12] показывает использование данных о взаимодействии студентов с системой управления обучением (LMS) в университете для создания предиктивных моделей, предсказывающих успеваемость студентов.

Основные результаты показали, что можно предсказывать успеваемость студентов с хорошей точностью, особенно в начальных моментах учебного процесса. Анализ также выявил шесть паттернов взаимодействия студентов с LMS, четыре из которых связаны с успеваемостью.

В работе [13] проведено исследование с целью изучения связи между онлайн-участием студентов и их успехами в учебе в течение четырехлетней программы обучения, а также выяснения, как эта связь развивается с течением времени и для каких студентов она меняется.

Эта статья также предлагает использование методов анализа последовательностей и скрытых марковских моделей для более детального изучения динамики успехов студентов и показывает, что студенты, которые поддерживают активное участие в учебном процессе в течение долгого времени достигают более высоких успехов. В то время как студенты, которые остаются неактивными в течение продолжительного периода, склонны к получению более низких оценок.

Автор работы [10] исследуют высокие достижения в математике и науке среди студентов в Ирландии. Исследование использует данные из программы международной оценки достижений учащихся. Методы анализа включают в себя двухуровневую бинарную логистическую регрессию для оценки влияния различных факторов, связанных со студентами, их семьями, классами и школами, на высокие достижения в математике и науке.

2. АНАЛИЗ ПРЕДМЕТНОЙ ОБЛАСТИ

В разделе описывается постановка задачи и описание данных, которые используются для предсказания.

Постановка задачи

Важной частью разрабатываемого приложения является анализ успеваемости студентов, который осуществляется на основе данных из журналов, которые хранятся в «Электронном ЮУрГУ». Журналы содержат информацию о студентах и их успеваемости. Важным этапом является выбор оптимальных методов анализа данных, которые учитывают особенности имеющихся данных, а также определение наиболее значимых факторов, влияющих на успеваемость студентов. Это позволит создать модели, которые точно предсказывают успеваемость студентов и предоставляют полезную информацию для принятия решений.

Для разработки приложения будут использоваться различные методы классификации, которые позволят выявить закономерности, которые могут быть полезны для предсказания будущей успеваемости студентов. Важным шагом в данном проекте будет выбор оптимальных методов анализа данных, учитывая особенности имеющихся данных, а также определение наиболее значимых факторов, влияющих на успеваемость студентов, и создание моделей, которые позволят предсказывать успеваемость на их основе.

Описание данных

Далее описаны основные атрибуты запросов к «Электронному ЮУрГУ», используемые для получения информации об учебном процессе в университете. Каждый запрос имеет свой набор атрибутов, соответствующих конкретной информации. Подробное описание каждого атрибута позволяет лучше понимать, какую информацию можно получить и как ее использовать для решения различных задач.

Диаграмма связей таблиц, представленных на рисунке 1, иллюстрирует структуру и взаимодействие данных, используемых в системе «Электронного ЮУрГУ». Она представляет собой схему, которая помогает понять, как различные элементы базы данных связаны между собой и каким образом они могут быть использованы для выполнения запросов.

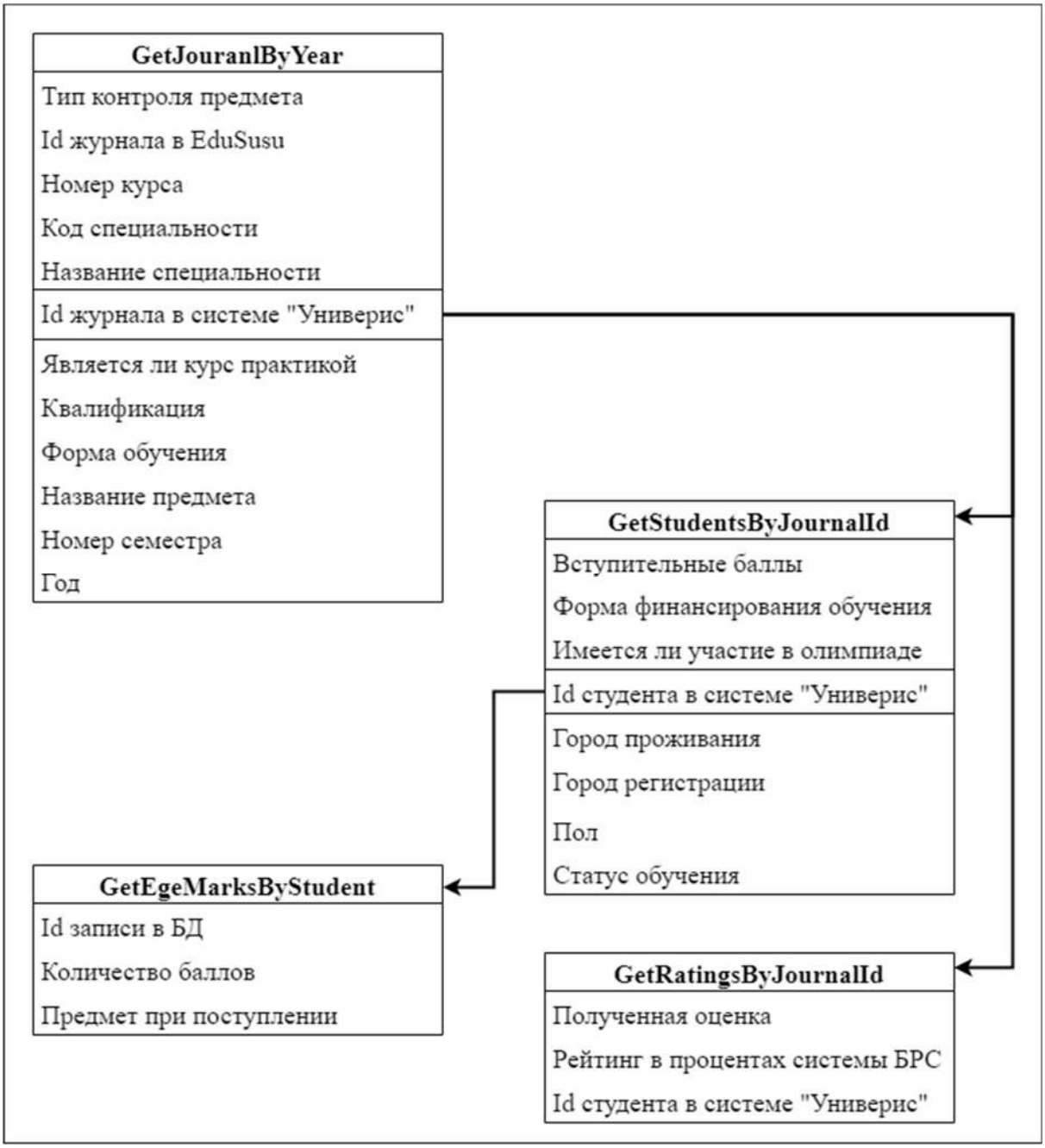

Рисунок 1 – Доступная для анализа структура
базы данных «Электронного ЮУрГУ»

Таблица 1 показывает структуру результата запроса «GetJournalByYear», который позволяет получить информацию о журналах по заданному году. В ней атрибут Id используется далее для получения информации о студентах и их успеваемости.

Таблица 1 – Описание атрибутов запроса «GetJournalByYear»

Атрибут	Семантика	Пример
CheckType	Тип контроля предмета	Зачет, экзамен, аттестация, практика
CourseEduId	Id журнала в системе «EduSusu»	152872
CourseNumber	Номер курса	С 1 по 6
DirectionCode	Код специальности	15.03.03
DirectionName	Название специальности	Прикладная механика
Id	Id журнала в системе Универис	5ef36ac9-67de-430c-a1dc-000d081c752c
GroupId	Id студенческой группы	97fe3c32-8b97-4ff7-837c-92b5f2edbfa5
IsPractice	Является ли курс практикой	true, false
Speciality	Квалификация	бакалавр, специалист
StudyForm	Форма обучения	очная, заочная
SubjectName	Название предмета	Электротехника и электроника
Term	Семестр, в котором проходит предмет	2
Year	Год	2021

Таблица 2 показывает структуру результата запроса «GetStudentsByJournalId», который показывает информацию о студентах. В ней атрибут Id используется для получения результатов ЕГЭ.

Таблица 2 – Описание атрибутов запроса «GetStudentsByJournalId»

Атрибут	Семантика	Пример
EnrollScore	Вступительные баллы	252
FinancialForm	Форма финансирования обучения	бюджет
HasOlymp Participation	Имеется ли участие в олимпиаде	false
Id	Id студента в системе Универис	ad69ba03-36fb-4c4e-8c75-14b26da8ef38
LiveCity	Город проживания	г.Челябинск
RegisterCity	Город регистрации	г.Челябинск
Sex	Пол	Мужской, женский
Status	Статус обучения	учится, отчислен

Таблица 3 показывает структуру результата запроса «GetRatingsByJournalId», который предоставляет информацию о рейтинге студентов по журналу.

Таблица 3 – Описание атрибутов запроса «GetRatingsByJournalId»

Атрибут	Семантика	Пример
Mark	Полученная оценка	3
Rating	Рейтинг в процентах по системе БРС	60.77
StudentId	Id студента в системе Универис	9b6db037-5e67-4331-bdb5-35c5d75d5235

Таблица 4 показывает структуру результата запроса «GetEgeMarksByStudentId», который предоставляет информацию о количестве баллов по ЕГЭ и предметах, с которыми поступал абитуриент.

Таблица 4 – Описание атрибутов запроса «GetEgeMarksByStudentId»

Атрибут	Семантика	Пример
Id	Id записи предмета	34306cea-7c32-4ad5-b923-eb3d123886c0
Mark	Количество баллов	64
Subject	Предмет, с которым поступал абитуриент	Математика

В целом, атрибуты запросов к «Электронному ЮУрГУ» направлены на получение различных характеристик студентов. Запросы позволяют извлекать данные о журналах, студентах и их академических достижениях, а также о вступительных баллах и участии в олимпиадах.

Эти данные служат основой для оценки успеваемости студентов и проведения сравнительного анализа по различным направлениям и специальностям. Такой подход помогает лучше понять динамику образовательных процессов и принять обоснованные решения для улучшения качества образования в университете.

3. ПРОЕКТИРОВАНИЕ

В разделе 3.1 представлены функциональные и нефункциональные требования к системе. В разделе 3.2 описаны варианты использования системы. В разделе 3.3 описывается предобработка данных. В разделе 3.4 представлено описание моделей предсказания. В разделе 3.5 представлен описан модуль предсказания, а в разделе 3.6 описан планируемый графический интерфейс.

3.1. Требования к системе

Функциональные требования.

В ходе анализа работы были определены следующие функциональные требования к разрабатываемой системе.

1. Система должна обеспечивать возможность интеграции с системой «Электронный ЮУрГУ» для извлечения учебных данных студентов, включая их сохранение для дальнейшего использования.
2. Система должна предоставлять предсказательные модели для анализа успеваемости студентов.
3. Система должна включать в себя механизмы для выбора и конфигурации алгоритмов классификации, а также для инициации процесса обучения аналитических моделей на основе полученных данных.
4. Система должна предусматривать возможности для сохранения и повторного использования обученных аналитических моделей, а именно для выполнения предсказания и анализа результатов с использованием этих моделей через автоматизированный запрос данных из «Электронного ЮУрГУ».

Нефункциональные требования.

Также были составлены нефункциональные требования, которые представлены далее.

1. Реализация интерфейса, обработки и анализа данных должна быть выполнена на языке программирования Python.

2. Данные для анализа должны быть получены из системы «Электронный ЮУрГУ».

3.2. Варианты использования системы

Для описания способов взаимодействия с системой была разработана диаграмма вариантов использования, изображенная на рисунке 2.

Рисунок 2 – Диаграмма вариантов использования

Первым актером, взаимодействующим с системой, является администратор, который осуществляет обучение моделей. Для него определены следующие варианты использования.

1. Извлечь данные. Позволяет администратору осуществлять извлечение данных системы «Электронного ЮУрГУ», путем запросов к API, либо загрузив файлы с сохраненными данными, запрошенными ранее.

2. Настроить параметры извлечения. Позволяет администратору задать параметры для извлечения данных из «Электронного ЮУрГУ», а именно: диапазон учебных годов, диапазон учебных курсов и высшую школу.

3. Сохранить данные. Позволяет администратору сохранить на свое устройство извлеченные из «Электронного ЮУрГУ» данные, чтобы иметь возможность воспользоваться ими в будущем.

4. Обучить модель. Позволяет администратору использовать извлеченные данные, которые он может использовать для обучения выбранной им модели, которую может использовать в анализе.

5. Настроить параметры обучения. Позволяет администратору задать параметры для дальнейшего обучения моделей, такие как: модель данных, аналитическую модель и гиперпараметры выбранной аналитической модели.

6. Сохранить модель. Позволяет администратору сохранить обученную на данных модель для последующего использования ее методистом.

Вторым актером, взаимодействующим с системой, является методист, который осуществляет использование обученных моделей. Для него определены следующие варианты использования.

1. Загрузить модель. Позволяет методисту загрузить обученную исследователем модель для дальнейшего использования.

2. Использовать модель. Позволяет методисту использовать загруженную ранее обученную исследователем модель и получить результат исследования.

3. Запросить данные. Позволяет методисту сделать запрос к «Электронному ЮУрГУ», чтобы получить данные для использования обученной модели.

4. Сохранить результат. Позволяет методисту сохранить результат анализа в виде документа с результатом предсказания.

3.3. Графический интерфейс

На рисунке 3 представлен макет окна извлечения данных из «Электронного ЮУрГУ». На макете изображены основные элементы интерфейса.

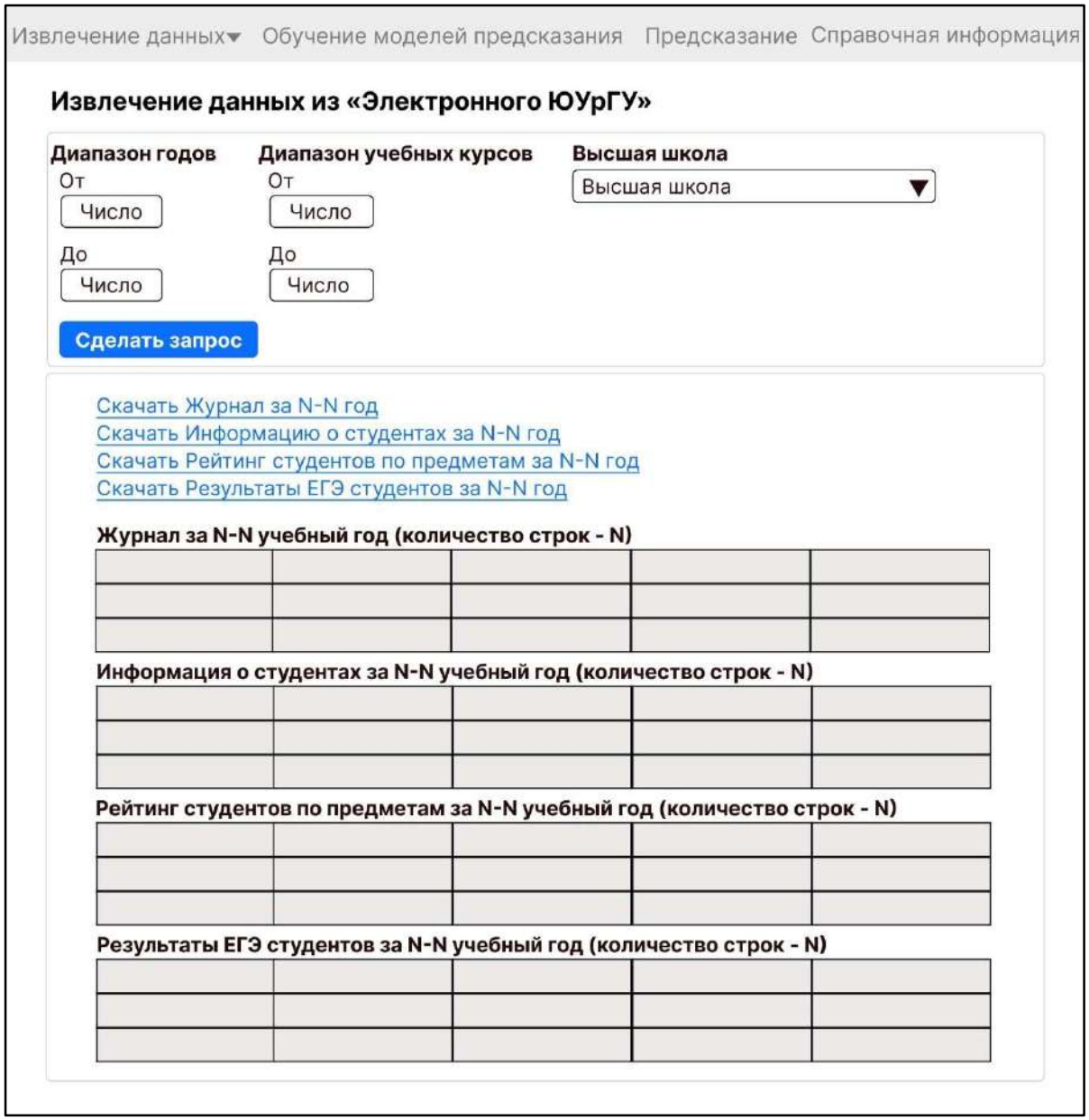

Рисунок 3 – Макет страницы извлечения данных
из «Электронного ЮУрГУ»

В верхней части находится меню переключения между страницами, далее расположены поля для настройки данных, получаемых из «Электронного ЮУрГУ», рядом с ними расположены кнопки для запроса по указанным настройкам. Ниже расположены ссылки для скачивания запрошенных файлов и таблицы, в которых отображаются данные, полученные после запроса.

На рисунке 4 представлен макет окна извлечения данных из файлов.

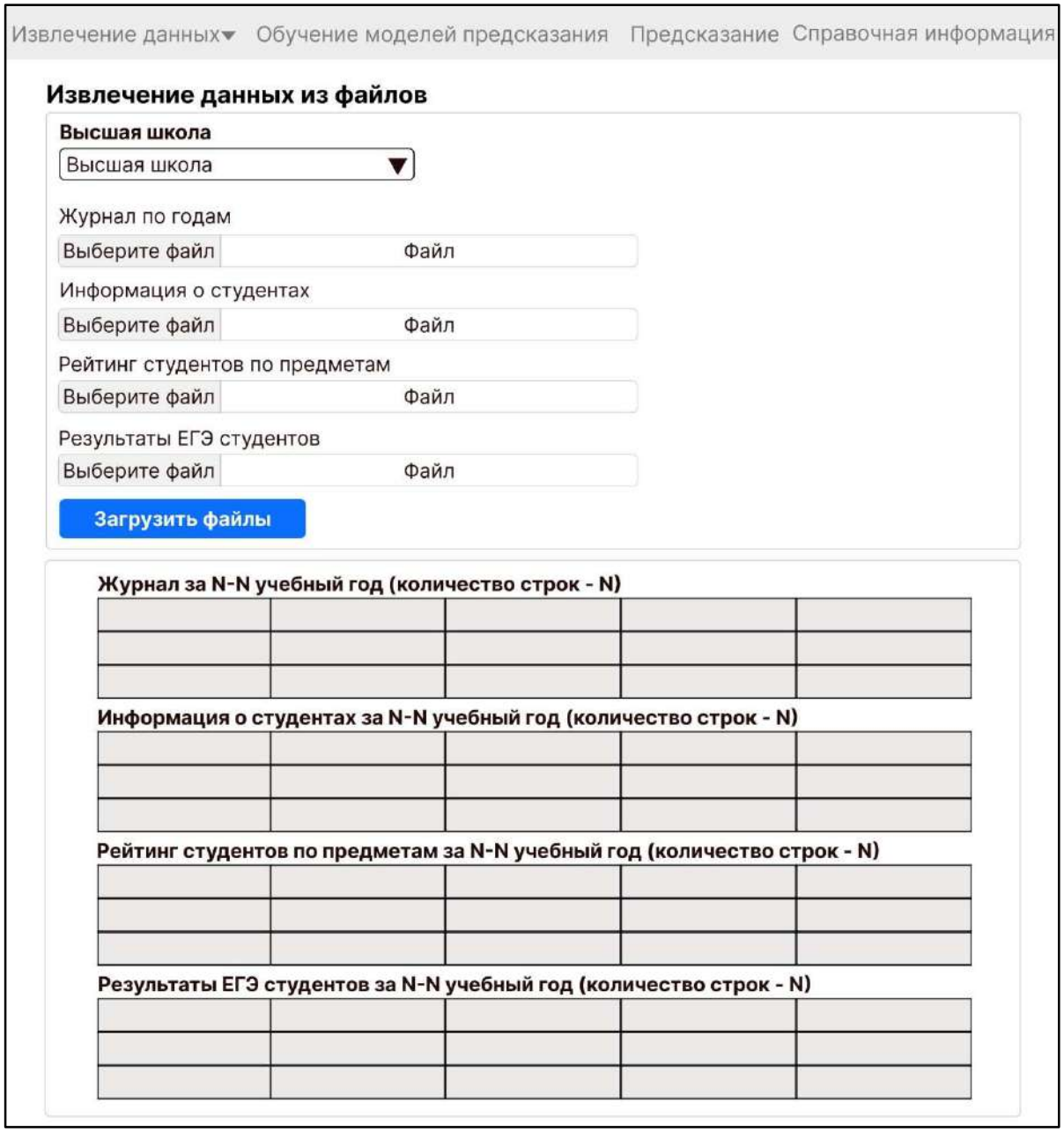

Рисунок 4 – Макет страницы загрузки ранее сохраненных данных

На странице также находится возможность переключения между окнами. Далее находится форма для загрузки файлов и выбора высшей школы загружаемых файлов. Ниже расположены таблицы загруженных файлов.

В приложении А представлены макеты окна обучения моделей предсказания. На форме расположена форма выбора модели и ее параметров, по которым будет обучаться модель. После нее таблица с составленной моде-

лью данных. Далее расположены результаты тестирования обученной модели с таблицей предсказанных результатов, после которой расположены метрики, кнопка для скачивания обученной модели и графики.

На рисунке 5 представлен макет окна предсказания.

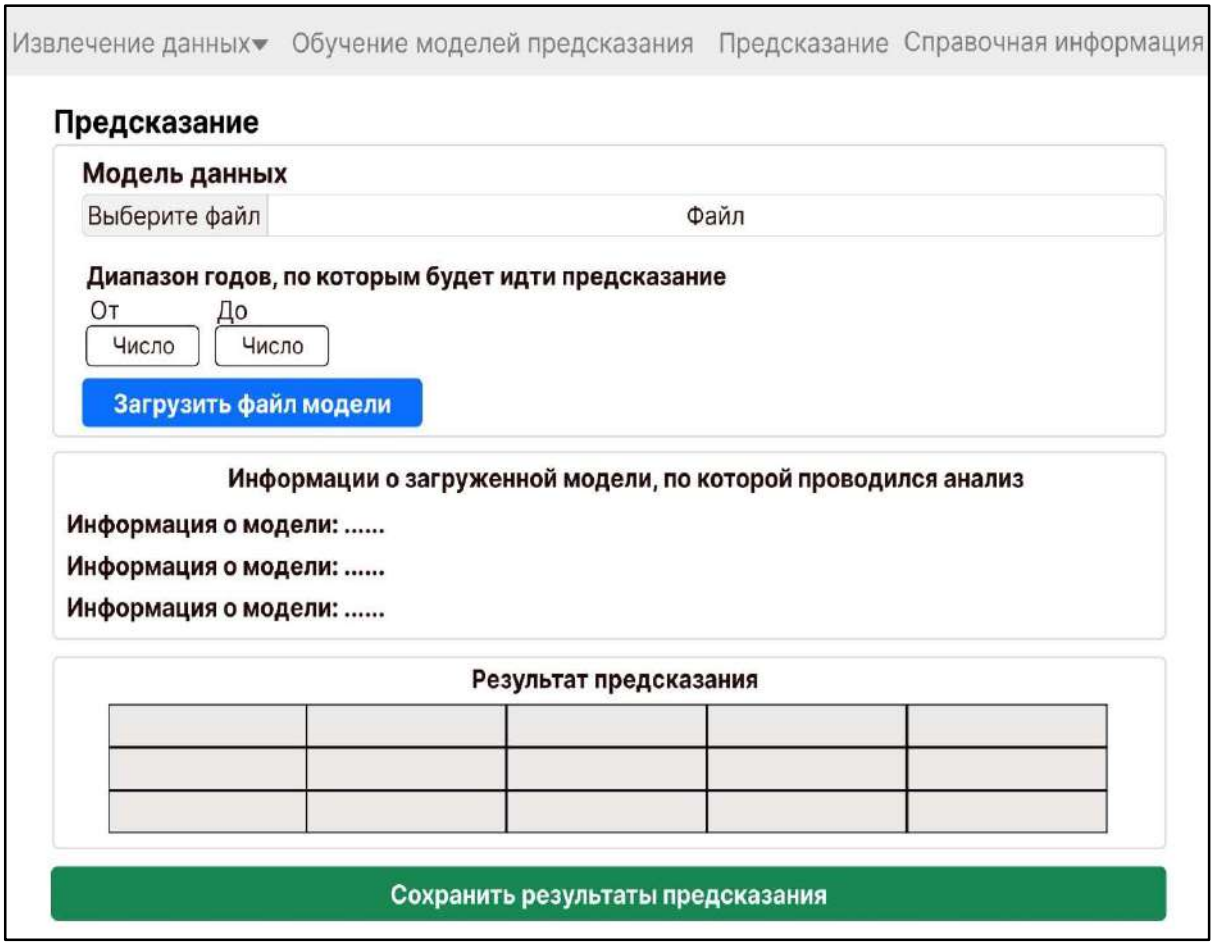

Рисунок 5 – Макет страницы предсказания

На форме расположены элементы для выбора файла обученной модели и выбора диапазона годов, по которым будет сделан запрос для предсказания. Ниже расположена информация о загруженной модели и таблица с данными о студентах и результатами предсказания загруженной модели и выполненного по годам запроса. В нижней части находится кнопка, которая позволяет сохранить результаты предсказания.

На рисунке 6 представлен макет страницы справочной информации.

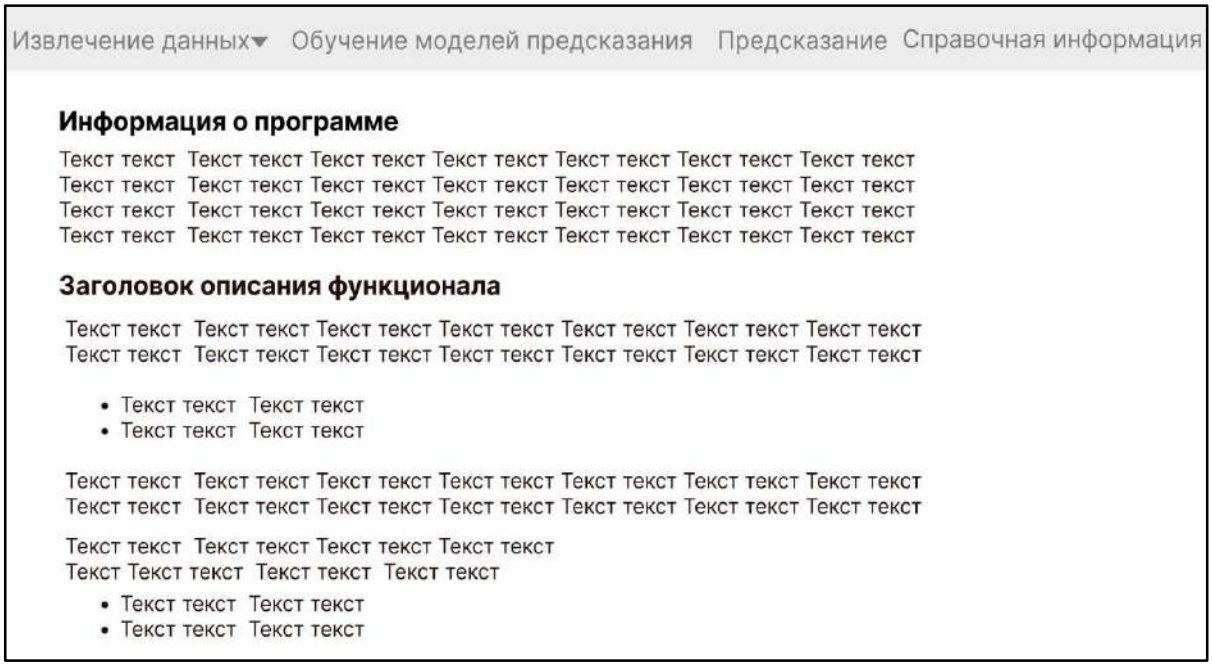

Рисунок 6 – Макет информации о программе

Также имеется меню с переключения между окнами, далее располагается информация о программе, ее функционале и том, что пользователю нужно делать по шагам, для анализа данных в программе.

На рисунке 7 представлена иерархия диалоговых форм.

Рисунок 7 – Иерархия диалоговых форм

Иерархия предполагает окно извлечения данных, в котором администратор производит извлечение данных, которые далее используются для анализа данных, после, методист получает возможность выполнить предсказание по загружаемым данным.

В целом, макеты окон являются интуитивно понятными и предоставляют пользователю интерфейс для работы с данными из «Электронного ЮУрГУ», обучения моделей и их использования.

3.4. Модели предсказания

Модели предсказания являются неотъемлемой частью анализа данных. В условиях большого объема данных, необходимых для анализа, применение моделей предсказания позволяет автоматизировать и ускорить процесс принятия решений на основе данных.

В задаче дипломной работы, модели предназначены для предсказания будущих результатов учебной деятельности студентов на основе данных, хранимых в системе «Электронного ЮУрГУ». Для того чтобы предсказания были точными и достоверными, необходимо правильно сформировать модели для анализа. Правильный выбор атрибутов, использующихся в моделях, играет важную роль в точности предсказания.

Далее описаны модели, разработанных для предсказания учебной деятельности студентов. Эти модели учитывают различные атрибуты, полученные из «Электронного ЮУрГУ», чтобы точно предсказать их будущие результаты и выявить ключевые факторы, влияющие на их успех.

Для обозначения основных данных о студентах, которые повторяются в различных моделях, принято условное обозначение группы данных «нативные данные студента» (таблица 5).

Таблица 5 – Нативные данные студента

№	Название поля	Тип данных	Семантика
1	Sex	bool	Пол
2	EgeMark1	int	Дециль ЕГЭ1 абитуриента
3	EgeMark2	int	Дециль ЕГЭ2 абитуриента
4	EgeMark3	int	Дециль ЕГЭ3 абитуриента
5	RegisterCity	bool	Является ли студент иногородним (местный/иногородний)
6	FinancialForm	bool	Форма финансирования обучения студента (бюджет/контракт)

Далее описаны сформированные модели.

1. Модель предсказания оценки дисциплины за указанный семестр (таблица 6). Предназначена для предсказания оценки, которую студент получит по определенному предмету в семестре обучения.

Таблица 6 – Модель предсказания оценки дисциплины за указанный семестр

№	Название поля	Тип данных	Семантика
Входные данные			
1	Нативные данные студента		Таблица 5
Выходное предсказываемое поле			
2	Mark	int	Оценка по предмету за N семестр

2. Модель предсказания оценки дисциплины, проходящей в указанном диапазоне семестров (таблица 7). Предназначена для прогнозирования оценки, которую студент получит по указанному предмету в последнем указанном семестре обучения.

Таблица 7 – Модель предсказания оценки дисциплины, проходящей в указанном диапазоне семестров

№	Название поля	Тип данных	Семантика
Входные данные			
1	Нативные данные студента		Таблица 5
2	MarkN	int	Оценка по предмету за N семестр
3	MarkN+1	int	Оценка по предмету за N+1 семестр
Выходное предсказываемое поле			
4	MarkN+2	int	Оценка по предмету за N+2 семестр

3. Модель предсказания оценки по практикам (таблица 8).

Таблица 8 – Модель предсказание успеваемости прохождения практик

№	Название поля	Тип данных	Семантика
Входные данные			
1	Нативные данные студента		Таблица 5
2	DecileSum RatingStudent{n}Sem	int	Дециль суммы рейтингов студента по всем дисциплинам N семестра
3	DecileSum RatingStudent{n+1}Sem	int	Дециль суммы рейтингов студента по всем дисциплинам N+1 семестра
Выходное предсказываемое поле			
4	InternshipSuccess	int	Успешность прохождения практики

4. Модель предсказания отчисления студентов по экзаменам/зачетам (таблица 9). Предоставляет возможность предсказать, будет ли студент отчислен или продолжит обучение после указанных семестров.

Таблица 9 – Модель предсказания вероятности отчисления

№	Название поля	Тип данных	Семантика
Входные данные			
1	Нативные данные студента		Таблица 5
2	DecileSum RatingStudent{n}Sem	int	Дециль суммы рейтингов студента по всем дисциплинам N семестра
3	DecileSum RatingStudent{n+1}Sem	int	Дециль суммы рейтингов студента по всем дисциплинам N+1 семестра
4	DecileMedian RatingCredits{n}Sem	int	Дециль медианного рейтинга студента по дисциплинам экзаменов/зачетов N семестра
5	DecileMedian RatingCredits{n+1}Sem	int	Дециль медианного рейтинга студента по дисциплинам экзаменов/зачетов N+1 семестра
Выходное предсказываемое поле			
6	Status	bool	Отчислен/Учится

5. Модель предсказания успеваемости на основе рейтинга по экзаменам/зачетам (таблица 10). Предоставляет возможность предсказать оценку, которую студент получит по экзаменам или зачетам.

Таблица 10 – Модель предсказания успеваемости на основе рейтинга по экзаменам/зачетам

№	Название поля	Тип данных	Семантика
Входные данные			
1	Нативные данные студента		Таблица 5
2	DecileSum RatingStudent{n}Sem	int	Дециль суммы рейтингов студента по всем дисциплинам N семестра
3	DecileSum RatingStudent{n+1}Sem	int	Дециль суммы рейтингов студента по всем дисциплинам N+1 семестра
4	DecileMedian RatingCredits{n}Sem	int	Дециль медианного рейтинга студента по дисциплинам экзаменов/зачетов N семестра
5	DecileMedian RatingCredits{n+1}Sem	int	Дециль медианного рейтинга студента по дисциплинам экзаменов/зачетов N+1 семестра
Выходное предсказываемое поле			
6	DecileMedian RatingCredits{n+2}Sem	int	Оценка студента по дисциплинам экзаменов/зачетов N+2 семестра

6. Модель предсказания успеваемости при завершении обучения (таблица 11). Предоставляет возможность предсказать успешность завершения обучения студента, по децилям суммы всех дисциплин указанного диапазона семестров.

Таблица 11 – Модель предсказания успеваемости при завершении обучения

№	Название поля	Тип данных	Семантика
Входные данные			
1	Нативные данные студента		Таблица 5
2	DecileMedian RatingCredits{N}Sem	int	Дециль медианного рейтинга студента по дисциплинам экзаменов/зачетов N семестра
3	DecileMedian RatingCredits{N+1}Sem	int	Дециль медианного рейтинга студента по дисциплинам экзаменов/зачетов N+1 семестра
Выходное предсказываемое поле			
4	AvgScoreFinalRating	int	Средний балл итогового рейтинга

Эти модели представляют собой инструменты прогнозирования, которые не только облегчают анализ данных, но и служат фундаментом для планирования образовательных процессов.

Прогностические модели, разработанные для оценки будущих достижений студентов, выполняют важную роль в адаптации учебных программ и индивидуальных планов обучения. Они позволяют более точно выявлять сильные и слабые стороны студентов, что дает возможность преподавателям своевременно корректировать образовательные стратегии. Это, в свою очередь, способствует более персонализированному подходу к обучению, повышению мотивации и академических успехов студентов.

Таким образом, использование данных моделей не только облегчает принятие решений на основе объективных данных, но и стимулирует постоянное совершенствование образовательных процессов. В долгосрочной перспективе это ведет к формированию более компетентных и успешных специалистов, готовых к вызовам современного мира.

4. РЕАЛИЗАЦИЯ

Для визуализации высокоуровневого представления, организации компонентов системы и зависимостей между ними, компоненты на диаграмме представляют собой модули разработанных классов, составлена диаграмма компонентов (рисунок 8).

Структура диаграммы представлена в виде модулей, которые в данном контексте представляют собой условное обозначение направления функционала классов.

Модуль извлечения данных позволяет получить данные из системы «Электронного ЮУрГУ» для анализа, которые с помощью модуля предобработки возможно очистить и сохранить необходимые данные. Модуль моделей предсказания формирует наборы данных для анализа, а модуль анализа данных позволяет обучить по сформированным предсказательным моделям аналитические модели и оценить их точность. Функционал модулей объединен в интерфейс взаимодействия, который позволяет взаимодействия с остальными модулями. Полный набор исходных текстов проекта в репозитории [17].

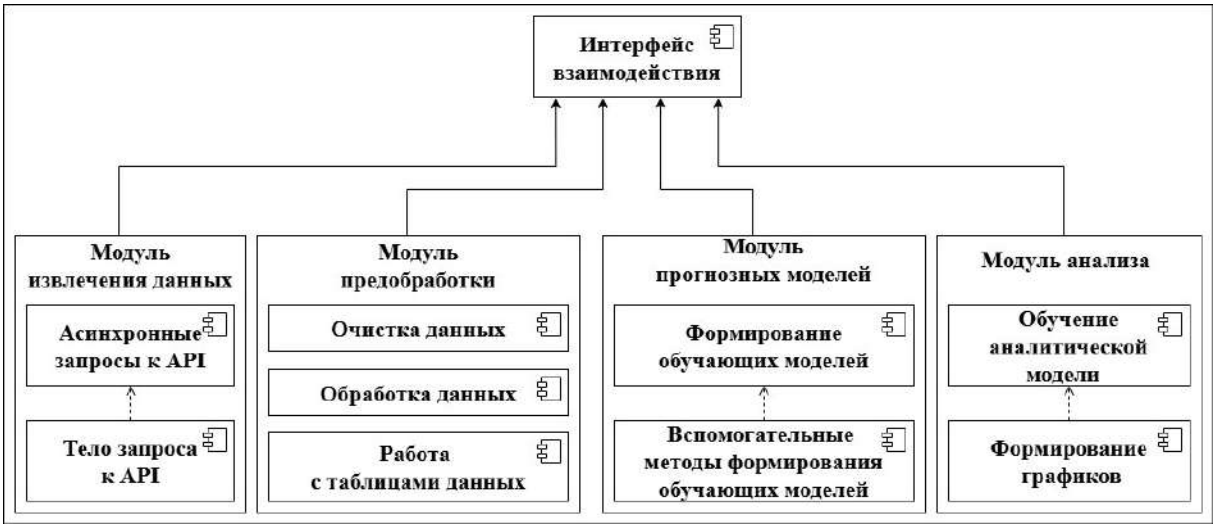

Рисунок 8 – Диаграмма компонентов

4.1. Программные средства реализации

Разработка приложения для предсказания успеваемости студентов была выполнена с использованием языка программирования Python 3 [11] со следующим основным набором библиотек.

1. Pandas – это библиотека Python для анализа и обработки данных. Она предоставляет удобные и мощные инструменты для работы с таблицами, временными рядами и другими форматами данных [9].

2. Numpy – это библиотека Python для научных вычислений. Она предоставляет многомерные массивы и матрицы, а также функции для работы с ними [8].

3. Asyncio – это библиотека Python для асинхронного программирования. Она позволяет создавать асинхронные функции и сопрограммы, которые могут выполняться параллельно без использования потоков [3].

4. Aiohttp – это библиотека Python для создания асинхронных веб-приложений. Она предоставляет клиентские и серверные инструменты для работы с протоколами HTTP и WebSocket [2].

5. Scikit-Learn – это библиотека Python для машинного обучения. Она предоставляет инструменты для обучения моделей машинного обучения, включая классификацию, регрессию, кластеризацию, обработку текстов и другие задачи [14].

6. Matplotlib – это библиотека Python для визуализации данных. Она предоставляет инструменты для создания графиков, диаграмм, карт и других видов визуализации [7].

7. Tenacity – это библиотека Python для обработки ошибок и повторной попытки выполнения операций. Она предоставляет механизмы для обработки ошибок сетевых запросов, баз данных и других операций, которые могут временно недоступны или нестабильны [15].

8. Django – это фреймворк Python для создания веб-приложений. Она предоставляет инструменты для создания и управления базами данных, обработки запросов, авторизации пользователей, взаимодействия с другими веб-сервисами и многое другое [4].

4.2. Модуль извлечения данных

Модуль извлечения состоит из классов `DigitalTrace` и `AsyncRequests`. Класс `DigitalTrace` формирует основу для запросов к API системы «Электронного ЮУрГУ», содержит поле BASE_URL, которое задает основной путь для GET запросов к системе «Электронного ЮУрГУ». Класс `DigitalTrace` имеет следующие методы.

1. Метод `get_journal_by_year` – позволяет получить журналы по диапазону введенных лет. Входные данные: диапазон учебных лет, по которым необходимо сделать запрос. Выходные данные: JSON с журналами по введенным годам.

2. Метод `get_students_by_journal_id` – позволяет получить информацию о студентах. Входные данные: id журнала в системе «Электронного ЮУрГУ. Выходные данные: JSON с информацией о студентах по журналу.

3. Метод `get_rating_by_journal_id` – позволяет получить информацию о рейтинге студентов. Входные данные: id журнала в системе «Электронного ЮУрГУ. Выходные данные: JSON с информацией о рейтинге по журналу.

4. Метод `get_ege_marks_by_student_id` – позволяет получить информацию о результатах ЕГЭ абитуриентов. Входные данные: id студента в системе «Электронный ЮУрГУ». Выходные данные: JSON с информацией о результатах ЕГЭ абитуриентов.

Методы класса `AsyncRequests` предназначены для осуществления асинхронных запросов, созданных в `DigitalTrace`.

Все методы, используют механизм семафоров для ограничения количества одновременно выполняемых запросов к API, чтобы избежать перегрузки удаленного сервера. Ограничение на количество задач позволяет эффективно использовать ресурсы и уменьшить нагрузку на сервер.

Также используется декоратор retry, который позволяет повторять выполнение асинхронной функции в случае возникновения ошибки при запросе к API. Это повышает устойчивость кода к сбоям, так как вместо сбоя при первой неудачной попытке выполнения задачи, код будет повторно пытаться выполнить задачу, что увеличит вероятность успешного выполнения операции в конечном итоге.

4.3. Модуль предобработки данных

Необработанные данные из источников, таких как API, часто содержат множество ошибок и неточностей, которые могут привести к неправильным выводам и оценкам. Поэтому необходимо проводить предобработку данных, чтобы выделить необходимые данные, а также гарантировать точность и качество анализа. Этот процесс может включать в себя удаление дубликатов, заполнение пропущенных значений, преобразование категориальных переменных в числовые значения и масштабирование данных.

Модуль предобработки включает в себя классы `DataClear`, `TransformationsOverDataframe` и `PreprocessingAnalysis`.

Класс `DataClear` содержит в себе методы, необходимые для того, чтобы убрать лишние значения из выборки, имеет атрибут INSTITUTES, который представляет из себя словарь с ключами в виде названий высших школ ЮУрГУ и их значений в виде списка с кодом специальностей, который будут использованы для фильтрации по определенной высшей школе. Методы класса `DataClear`.

1. Метод `institutional_exclusion` – позволяет получить список кодов высших школ, для исключения их из запроса к API. Выходные данные: названия высших школ, который необходимо будет убрать из набора

данных. Выходные данные: список значений ключей указанных высших школ.

2. Метод `drop_rows_in_journal` – позволяет удалить указанные строки в dataframe из указываемого столбца в формате ключ-значение. Входные данные: строки для удаления. Выходные данные: dataframe с удаленными строками.

3. Метод `keep_rows_in_journal` – позволяет сохранить только указанные строки в dataframe из указываемого столбца в формате ключ-значение. Входные данные: строки для удаления. Выходные данные: dataframe с указанными сохраненными строками.

В классе `TransformationsOverDataframe` содержится метод `ege_marks_transpose`, который позволяет транспонировать строки результатов ЕГЭ студентов в dataframe, формируя столбцы вступительных предметов. Входные данные: результат, возвращаемый методом получения результатов ЕГЭ. Выходные данные: dataframe с транспонированными строками результатов ЕГЭ студентов.

Класс `PreprocessingAnalysis` используемый для обработки моделей данных перед передачей их в аналитическую модель, содержит следующие методы.

Метод `data_synthesis` – позволяет выполнить синтез данных с помощью алгоритма ADASYN. Входные данные: dataframe модели данных. Выходные данные: dataframe с дополнительными данными.

Метод `emission_removal` – Позволяет найти межквартильный размах, определить границы выбросов и удалить их из передаваемой модели данных. Входные данные: dataframe модели данных и прогнозируемый столбец. Выходные данные: dataframe с удаленными выбросами.

Метод `data_remove_noise_with_dbscan` – Позволяет удалить шумы из модели данных используя алгоритм кластеризации DBSCAN. Входные данные: dataframe модели данных. Выходные данные: dataframe с удаленными шумами.

4.4. Модуль моделей предсказания

Для решения задачи предсказания успеваемости студентов на основе полученных атрибутов и составленных моделей были выбраны два алгоритма машинного обучения: случайный лес и градиентный бустинг.

Алгоритм случайного леса использует ансамбль решающих деревьев для предсказания. В случайных лесах деревья в ансамбле строятся из выборки, взятой с заменой (то есть выборкой начальной загрузки) из обучающего набора. Каждое дерево обучается на подмножестве данных и делает предсказание. Затем предсказания всех деревьев усредняются, чтобы получить окончательное предсказание. Этот метод хорошо работает для больших наборов данных и может обрабатывать большое количество признаков.

Градиентный бустинг также использует ансамбль решающих деревьев. Однако в отличие от случайного леса, градиентный бустинг строит деревья последовательно, каждое следующее дерево исправляет ошибки предыдущего. Этот метод хорошо работает для малых и средних наборов данных и может достигать высокой точности предсказания [10].

Выбор между алгоритмами случайного леса и градиентного бустинга зависит от специфики задачи и доступных данных. В обозначенной задаче оба алгоритма могут быть использованы для предсказания успеваемости студентов. Они помогут выявить взаимосвязи между различными факторами и их влиянием на успеваемость.

В модуле моделей предсказания содержатся классы `PredictiveModels` и `ModelsSimplification`.

Класс `PredictiveModels` позволяет формировать модели предсказания данных по полученной и предобработанной выборке из «Электронного ЮУрГУ», содержит следующие методы.

1. Метод `model_subjectName` – используется для формирования моделей предсказания оценки дисциплины за указанный семестр, предсказания оценки дисциплины, проходящей в указанном диапазоне семестров и

предсказания успеваемости прохождения практик. Входные данные: название предмета, начало диапазона семестров, конец диапазона семестров, параметр добавления дополнительного столбца. Выходные данные: dataframe с сформированной моделью.

2. Метод `model_students_rating_status` – используется для формирования моделей предсказания вероятности отчисления по экзаменам/зачетам и предсказания успеваемости на основе рейтинга по экзаменам/зачетам. Входные данные: начало диапазона семестров, конец диапазона семестров, тип контроля предмета, параметр добавления дополнительного столбца. Выходные данные: dataframe с сформированной моделью.

3. Метод `model_avg_final_rating` – используется для формирования модели предсказания успеваемости при завершении обучения. Входные данные: начало диапазона семестров, конец диапазона семестров. Выходные данные: dataframe с сформированной моделью.

Класс `ModelsSimplification` содержит вспомогательные методы для формирования моделей в `PredictiveModels`, имеющийся в нем метод `decile_sum_rating_students_by_semestr` позволяет вычислить дециль суммы рейтингов студента по всем дисциплинам семестра. А метод `decile_median_rating_disciplin_check_type_by_semestr`, вычисляет дециль медианного рейтинга дисциплины зачета или экзамена переданного номера семестра.

4.5. Модуль анализа данных

В модуле анализа данных содержатся классы `DataAnalysis` и `GraphsBuilder`.

Класс `DataAnalysis` позволяет обучать определенные аналитические модели по сформированным предсказательным моделям данных. Содержит метод `train_by_model`, входные данные которого: dataframe со сформированной моделью данных, столбцом, значения которого будут предсказываться, типом модели для обучения, размером выборки, которая

будет использоваться для тестирования обученной модели и набором гиперпараметров. Выходные данные: обученная модель, оценки ее точности, данные для построения графиков и таблицу с предсказанными результатами.

Класс `GraphsBuilder` предназначен для построения графиков обученной модели, содержит методы.

1. Метод `feature_importance_graph` – позволяет формировать график важности признаков, отображая гистограмму входных данных, оказывающих наибольшее влияние на результат обучения. Входные данные: обученная аналитическая модель, столбцы входных. Выходные данные: график важности признаков.

2. Метод `plot_predictions` – позволяет формировать график, показывающий совпадение истинных и предсказанных значений обученной моделью на гистограмме. Входные данные: данные предсказываемого столбца, предсказанные значения по обученной модели. Выходные данные: график совпадения истинных и предсказанных значений.

3. Метод `plot_confusion_matrix` – позволяет формировать матрицу ошибок, отображающую предсказанные и истинные значения принадлежности данных к классам. Входные данные: данные предсказываемого столбца, предсказанные значения по обученной модели. Выходные данные: график матрицы ошибок предсказанных и истинных классов.

4.6. Реализация интерфейса

Интерфейс представляет собой веб-приложение, реализованное на основе фреймворка Django. Дополнительно для формирования визуальной составляющей используется HTML, CSS, JavaScript и Bootstrap.

Одной из ключевых страниц интерфейса является страница извлечения данных из «Электронного ЮУрГУ». На данной странице пользователю предоставляется возможность получить данные путем запроса к API системы «Электронного ЮУрГУ» по заданным параметрам.

С использованием разработанных модулей, данные получаются, затем осуществляется их очистка и обработка, далее пользователь имеет возможность просмотреть запрошенные данные и скачать их в формате CSV файлов. Этот функционал обеспечивает пользователю удобство при получении и сохранении данных.

На рисунке 9 показан выполненный интерфейс извлечения данных из API «Электронного ЮУрГУ».

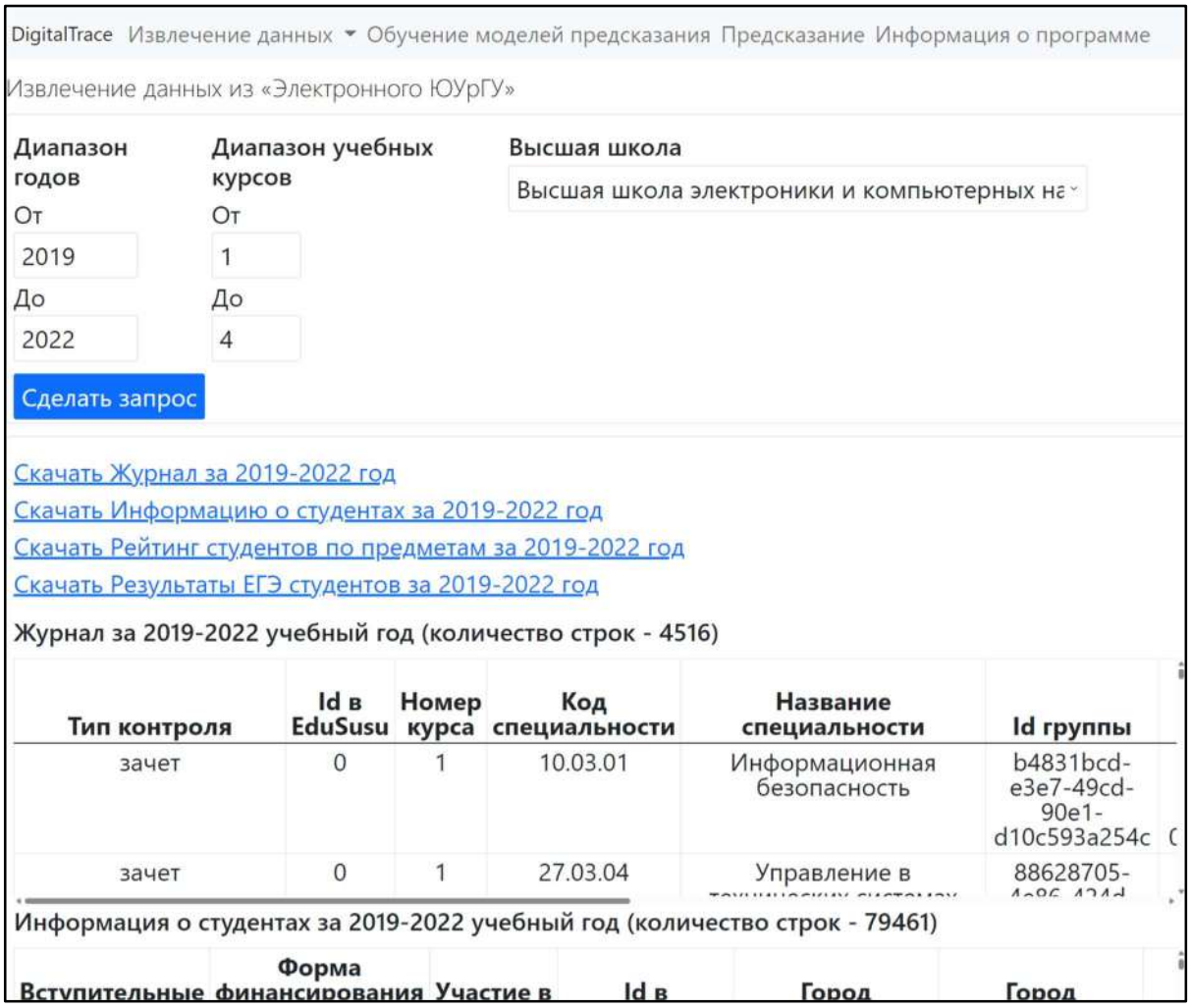

Рисунок 9 – Интерфейс извлечения данных из API

Кроме того, была разработана страница извлечения данных из CSV файлов, которая предоставляет пользователю возможность загрузить предварительно сохраненные файлы в приложении. На рисунке 10 показан разработанный интерфейс извлечения из CSV файлов.

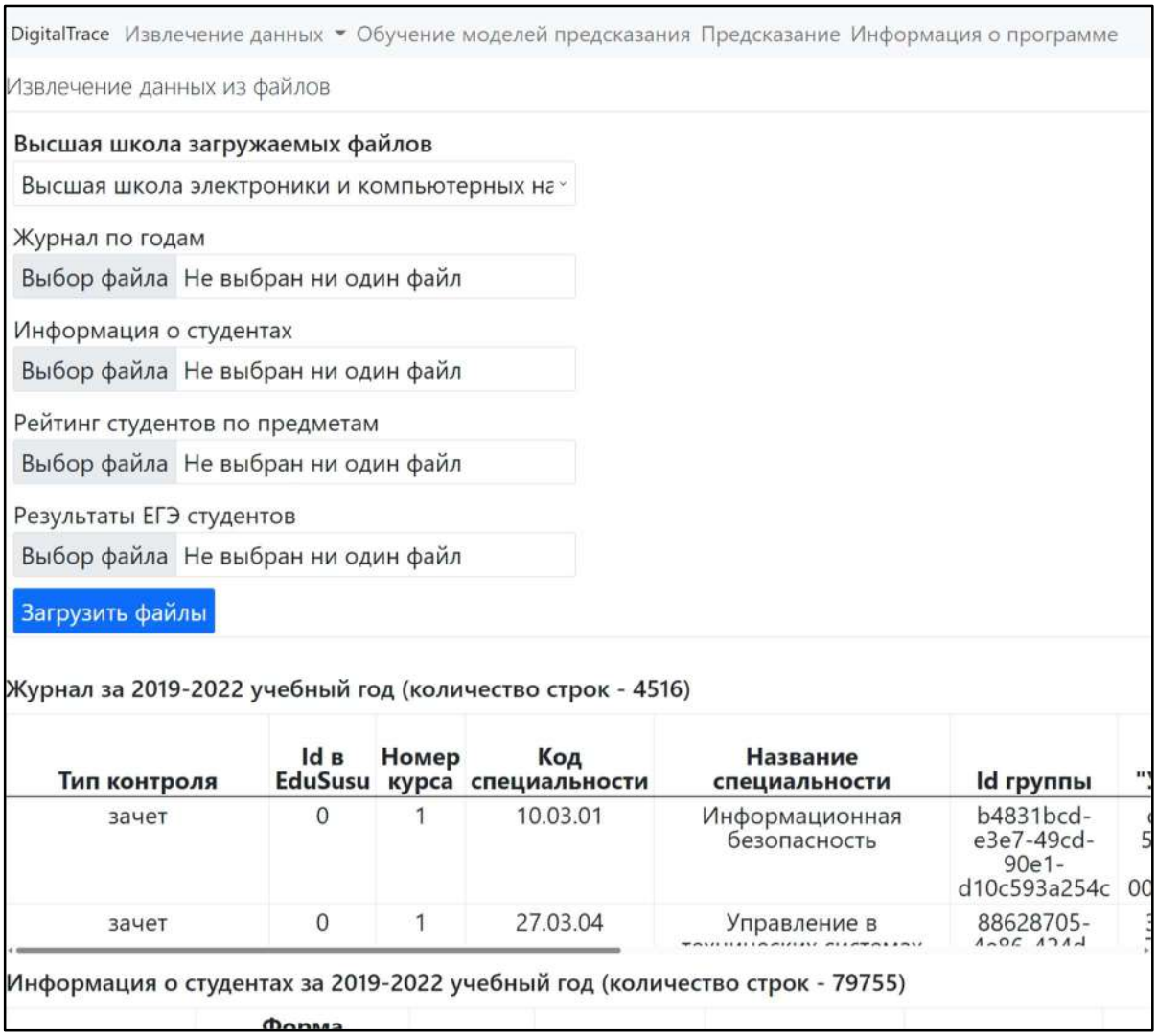

Рисунок 10 – Интерфейс извлечения данных из CSV файлов

Разработанные модули позволяют осуществить извлечение и обработку данных из загруженных файлов, что позволяет пользователям использовать повторно ранее запрошенные данные.

Также в интерфейсе была реализована страница обучения моделей предсказания. Пользователям предоставляется возможность задавать модель данных, а также ее настройки, по которой будет сформирована выборка и аналитическую модель, которая будет производить анализ по задаваемым далее гиперпараметрам.

После обучения предсказательной модели, у пользователя есть возможность увидеть таблицу результата обучения модели, а именно таблицу предсказанных результатов, кнопку для скачивания файла в формате joblib,

метрики точности обученной модели и диаграммы. Описанная страница находится в приложении Б.

Для предсказания, реализована страница, на которой имеется возможность загрузить файл обученной модели, а также указать диапазон годов, по которым будет сделан запрос к API для получения данных и выполнено предсказание с использованием аналитической модели из загруженного файла. Описанная страница представлена на рисунке 11.

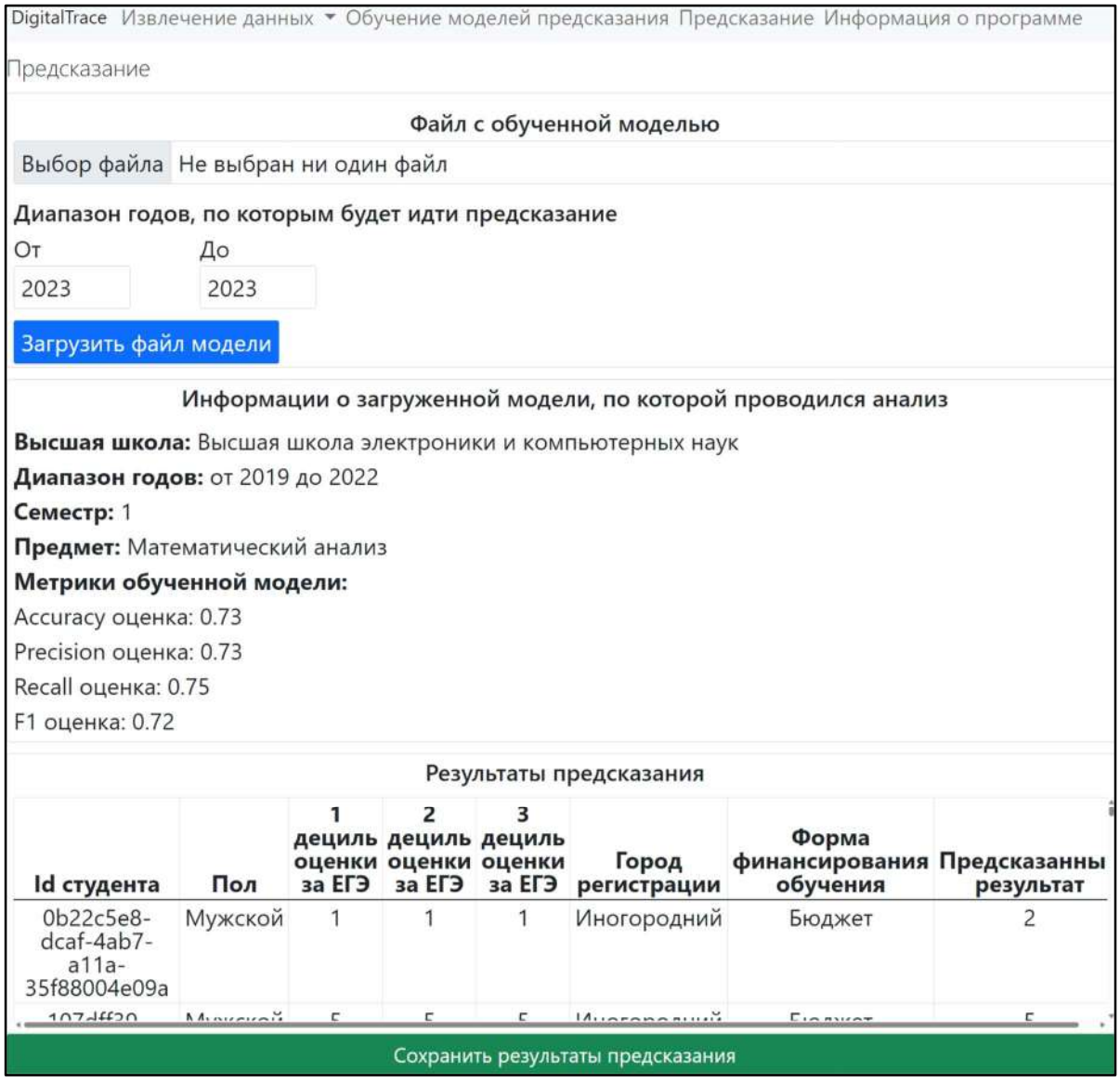

Рисунок 11 – Интерфейс страницы предсказания

Дополнительно, в интерфейсе была включена страница с информацией о программе, где предоставляется пользователю подробное описание функциональности приложения и инструкцию по использованию.

4.7. Тестирование интерфейса

Проверка работы интерфейса заключается в проведение функционального тестирования (таблица 12).

Таблица 12 – Тестирование интерфейса

№	Название теста	Шаги	Ожидаемый результат	Тест пройден?
1	Получение из API данных Высшей школы электроники и компьютерных наук за 2019–2021 год	1. На странице извлечения данных из API указать диапазон годов от 2019 до 2021. 2. Диапазон учебных курсов указать от 1 до 1. 3. Высшую школу указать «Высшая школа электроники и компьютерных наук». 4. Нажать кнопку «Сделать запрос».	После завершения всех запрос таблицы, полученных данных, корректно отображаются, имеется возможность скачать их в формате csv	Да
2	Получение из API данных Политехнического института за 2019–2019 год	1. На странице извлечения данных из API указать диапазон годов от 2019 до 2019. 2. Диапазон учебных курсов указать от 1 до 1. 3. Высшую школу указать «Политехнический институт». 4. Нажать кнопку «Сделать запрос».	После завершения всех запрос таблицы, полученных данных, корректно отображаются, имеется возможность скачать их в формате csv	Да
3	Проверка валидации получения данных из API	1. На странице извлечения данных из API указать диапазон годов от 2021 до 2019. 2. Нажать кнопку «Сделать запрос».	Система дает предупреждение о том, что год «От» должен быть меньше или равен году «До»	Да
4	Проверка валидации получения данных из файлов	1. На странице извлечения данных из фалов, добавить файлы по пунктам. – Журнал по годам; – Информация о студентах. 2. Нажать кнопку «Загрузить файлы».	Система дает предупреждение, что необходимо загрузить остальные файлы	Да
5	Проверка проведения анализа данных	1. Сделать извлечение данных из API или файлов 2. Перейти на страницу «Анализ данных» 3. Корректно указать параметры для анализа 4. Нажать кнопку «Обучить модель»	Отображается результат обучения модели с возможностью сохранения	Да

Продолжение таблицы 12

№	Название теста	Шаги	Ожидаемый результат	Тест пройден?
6	Скачивание полученных из API данных Политехнического института за 2019–2019 год	1. На странице извлечения данных из API указать диапазон годов от 2019 до 2019. 2. Диапазон учебных курсов указать от 1 до 1. 3. Указать высшую школу «Политехнический институт». 4. Нажать кнопку «Сделать запрос». 5. Нажать на гиперссылки. − Скачать Журнал за 2019-2019 год; − Скачать Информацию о студентах за 2019-2019 год; − Скачать Рейтинг студентов по предметам за 2019-2019 год; − Скачать Результаты ЕГЭ студентов за 2019-2019 год.	Таблицы скачиваются в формате csv	Да
7	Проверка получения данных из файлов	1. На странице извлечения данных из файлов, добавить файлы по пунктам: − Журнал по годам; − Информация о студентах; − Рейтинг студентов по предметам; − Результаты ЕГЭ студентов. 2. Нажать кнопку «Загрузить CSV файлы»	Добавленные csv файлы загружаются в программу, данные возможно использовать	Да
8	Проверка сохранения обученной модели	1. Сделать извлечение данных из API или файлов 2. Перейти на страницу «Анализ данных» 3. Корректно указать параметры для анализа 4. Нажать кнопку «Обучить модель» 5. Нажать кнопку «Сохранить обученную модель»	Результат обучения скачивается в формате joblib	Да

Окончание таблицы 12

№	Название теста	Шаги	Ожидаемый результат	Тест пройден?
9	Проверка проведения предсказания	1. Перейти на страницу «Анализ данных» 2. Загрузить файл с обученной моделью 3. Корректно указать параметры для предсказания 4. Нажать кнопку «Загрузить файл модели»	Отображается результат предсказания с возможностью сохранения	Да
10	Проверка сохранения результата предсказания	1. Перейти на страницу «Анализ данных» 2. Загрузить файл с обученной моделью 3. Корректно указать параметры для предсказания 4. Нажать кнопку «Загрузить файл модели» 5. Нажать кнопку «Сохранить результаты предсказания»	Результат обучения скачивается в формате csv	Да
11	Проверка валидации ввода данных на странице анализа данных	1. Сделать извлечение данных из API или файлов 2. Перейти на страницу «Анализ данных» 3. Указать в любом параметре некорректное значение 4. Нажать кнопку «Обучить модель»	Отображается ошибка о некорректном значении параметра для обучения	Да

В ходе проведения функционального тестирования были проверены различные аспекты работы системы. Тестирование показало, что система корректно обрабатывает запросы.

5. ВЫЧИСЛИТЕЛЬНЫЕ ЭКСПЕРИМЕНТЫ

Эксперименты формирования моделей предсказания

Неотъемлемой частью анализа данных, является проведение экспериментов, которые позволят увидеть качество предсказания разработанных моделей при изменении различных параметров.

Для достоверной и объективной оценки качества анализа классификационных моделей необходимо применять специализированные метрики.

Одной из распространенных метрик является точность (accuracy), которая измеряет долю правильно классифицированных наблюдений по отношению к общему числу наблюдений. Однако, точность может быть недостаточной для полного исследования модели, так как она не учитывает возможные неравенства важности различных классов и может оказаться искаженной в случае несбалансированных данных.

Для более полной и надежной оценки качества классификации используются также следующие метрики.

1. Полнота (recall), которая измеряет способность модели обнаруживать положительные классы из общего числа истинных положительных классов. Полнота является особенно важной метрикой в задачах, где ложноотрицательные предсказания имеют серьезные последствия.

2. Точность (precision), которая измеряет способность модели предсказывать правильно положительные классы из общего числа положительных предсказания. Точность позволяет оценить, насколько надежными являются положительные предсказания модели.

3. F-мера (F1-score), которая является гармоническим средним между точностью и полнотой. F-мера позволяет совместно учитывать точность и полноту модели и является полезной метрикой в случаях, когда важны и точность, и полнота предсказания.

Далее описано проведение экспериментов, заключающееся в предсказании на реальных данных, полученных путем запроса к API системы «Электронного ЮУрГУ».

Таблица 13 представляет параметры и итоговые метрики модели предсказания оценки дисциплины за указанный семестр, на рисунке 12 отображена полученная матрица ошибок после обучения этой модели, по которой видно, что имеется сильный разброс в точности истинных и предсказываемых значений.

Таблица 13 – Эксперимент модели предсказания оценки дисциплины за первый семестр

Высшая школа	**Диапазон годов**	**Семестр**	**Предмет**	**Аналитическая модель**
ВШЭКН	2019–2022	1	Математический анализ	Случайный лес

Проведено увеличение выборки	Проведено удаление выбросов	Проведено удаление шумов
Нет	Нет	Нет

Гиперпараметры			
Размер тестовой выборки	0,2	Минимальное количество объектов в узле	2
Количество деревьев решений	1000	Минимальное количество объектов в листьях деревьев решений	1
Максимальная глубина деревьев решений	10	Количество признаков при построении деревьев решений	3

Метрики			
Accuracy	Precision	Recall	F1
0,38	0,34	0,36	0,35

Рисунок 12 – Матрица ошибок модели в таблице 13

Таблица 14 также представляет параметры и итоговые метрики модели предсказания оценки дисциплины за указанный семестр, но можно увидеть, как изменение параметров обучения влияет на результат обучения модели. После увеличения выборки и удаления шумов, а также изменения гиперпараметров аналитической модели, итоговая точность повысилась в 2 раза, что также можно увидеть в матрице ошибок на рисунке 13.

Таблица 14 – Эксперимент модели предсказания оценки дисциплины за первый семестр, с подобранными параметрами

Высшая школа	**Диапазон годов**	**Семестр**	**Предмет**	**Аналитическая модель**
ВШЭКН	2019–2022	1	Математический анализ	Случайный лес

Проведено увеличение выборки	Проведено удаление выбросов	Проведено удаление шумов
Да	Нет	Да

Гиперпараметры			
Размер тестовой выборки	0,2	Минимальное количество объектов в узле	2
Количество деревьев решений	2000	Минимальное количество объектов в листьях деревьев решений	1
Максимальная глубина деревьев решений	20	Количество признаков при построении деревьев решений	1

Метрики			
Accuracy	Precision	Recall	F1
0,73	0,73	0,75	0,72

Рисунок 13 – Матрица ошибок модели в таблице 14

Таблица 15 представляет параметры и итоговые метрики модели предсказания оценки дисциплины, проходящей в указанном диапазоне семестров.

Таблица 15 – Эксперимент модели предсказания оценки дисциплины, проходящей в указанном диапазоне семестров

Высшая школа	**Диапазон годов**	**Диапазон семестров**	**Предмет**	**Аналитическая модель**
ПИ	2019–2021	1–3	Физическая культура и спорт	Случайный лес

Проведено увеличение выборки	Проведено удаление выбросов	Проведено удаление шумов
Нет	Нет	Да

Гиперпараметры			
Размер тестовой выборки	0,2	Минимальное количество объектов в узле	4
Количество деревьев решений	2000	Минимальное количество объектов в листьях деревьев решений	1
Максимальная глубина деревьев решений	30	Количество признаков при построении деревьев решений	3

Метрики			
Accuracy	Precision	Recall	F1
0,82	0,82	0,82	0,82

Таблица 16 представляет параметры и итоговые метрики модели предсказания оценки по практикам.

Таблица 16 – Эксперимент модели предсказания оценки по практикам

Высшая школа	**Диапазон годов**	**Семестр**	**Предмет**	**Аналитическая модель**
ВШЭКН	2019–2022	2	Учебная практика	Случайный лес

Проведено увеличение выборки	Проведено удаление выбросов	Проведено удаление шумов
Да	Нет	Нет

Гиперпараметры			
Размер тестовой выборки	0,2	Минимальное количество объектов в узле	2
Количество деревьев решений	1500	Минимальное количество объектов в листьях деревьев решений	2
Максимальная глубина деревьев решений	20	Количество признаков при построении деревьев решений	1

Метрики			
Accuracy	Precision	Recall	F1
0,76	0,67	0,65	0,64

Таблица 17 представляет параметры и итоговые метрики модели предсказания отчисления студентов по экзаменам/зачетам.

Таблица 17 – Эксперимент модели предсказания отчисления студентов по экзаменам/зачетам

Высшая школа	**Диапазон годов**	**Диапазон семестров**	**Тип контроля предмета**	**Аналитическая модель**
ПИ	2019–2022	1–4	Экзамен	Градиентный бустинг

Проведено увеличение выборки	Проведено удаление выбросов	Проведено удаление шумов
Да	Да	Нет

Гиперпараметры			
Размер тестовой выборки	0,2	Минимальное количество объектов в узле	4
Количество деревьев решений	1000	Минимальное количество объектов в листьях деревьев решений	1
Максимальная глубина деревьев решений	20	Количество признаков при построении деревьев решений	2

Метрики			
Accuracy	Precision	Recall	F1
0,91	0,91	0,91	0,91

Таблица 18 представляет параметры и итоговые метрики модели предсказания успеваемости на основе рейтинга по экзаменам/зачетам.

Таблица 18 – Эксперимент модели предсказания успеваемости на основе рейтинга по экзаменам/зачетам

Высшая школа	**Диапазон годов**	**Диапазон семестров**	**Тип контроля предмета**	**Аналитическая модель**
ВШЭКН	2019–2022	1–4	Зачет	Градиентный бустинг

Проведено увеличение выборки	Проведено удаление выбросов	Проведено удаление шумов
Да	Нет	Нет

Гиперпараметры			
Размер тестовой выборки	0,2	Минимальное количество объектов в узле	2
Количество деревьев решений	1000	Минимальное количество объектов в листьях деревьев решений	1
Максимальная глубина деревьев решений	30	Количество признаков при построении деревьев решений	3

Метрики			
Accuracy	Precision	Recall	F1
0,75	0,72	0,70	0,69

Таблица 19 представляет параметры и итоговые метрики модели предсказания успеваемости при завершении обучения.

Таблица 19 – Эксперимент модели предсказания успеваемости при завершении обучения

Высшая школа	**Диапазон годов**	**Диапазон семестров**	**Аналитическая модель**
ПИ	2019–2022	2–3	Случайный лес

Проведено увеличение выборки	Проведено удаление выбросов	Проведено удаление шумов
Да	Да	Нет

Гиперпараметры			
Размер тестовой выборки	0,2	Минимальное количество объектов в узле	4
Количество деревьев решений	1500	Минимальное количество объектов в листьях деревьев решений	1
Максимальная глубина деревьев решений	30	Количество признаков при построении деревьев решений	2

Метрики			
Accuracy	Precision	Recall	F1
0,81	0,71	0,72	0,71

Проведенные эксперименты показали, что используемые модели дают оптимальные итоговые точности метрик, а также могут быть улучшены при подборе входных параметров.

Эксперименты анализа данных

Далее, были проведены эксперименты использования обученных моделей, которые позволят увидеть результат предсказания на неизвестных для обученной модели данных.

На рисунке 5 приложения В представлен эксперимент, проведенный на данных ВШЭКН за 2023 год, показывающий предсказание оценки дисциплины за 1 семестр по математическому анализу, данных 2019–2022 годов со следующими метриками:

1) accuracy – 0,73;
2) precision – 0,73;
3) recall – 0,75;
4) f1 – 0,72.

На рисунке 6 приложения В представлен эксперимент, проведенный на данных ВШЭКН за 2023 год, показывающий предсказание отчисления

студентов за 1–2 семестр по зачетам, данных 2019–2022 годов со следующими метриками:

1) accuracy – 0,84;
2) precision – 0,84;
3) recall – 0,84;
4) f1 – 0,84.

На рисунке 7 приложения В представлен эксперимент, проведенный на данных ПИ за 2022–2023 года, показывающий предсказание успеваемости на основе рейтинга за 1–4 семестр по зачетам, данных 2019–2021 годов со следующими метриками:

1) accuracy – 0,73;
2) precision – 0,69;
3) recall – 0,65;
4) f1 – 0,66.

Проведение экспериментов по анализу данных имеет ключевое значение для оценки и улучшения качества предсказательных моделей. Такие эксперименты позволяют не только проверить, но и оптимизировать различные параметры моделей, что в свою очередь приводит к более точным и надежным предсказаниям.

В результате проведенных тестов видно, что правильная настройка гиперпараметров и предварительная обработка данных существенно повышают эффективность моделей. Эти эксперименты помогли выявить сильные и слабые стороны текущих методов, дав ценные результаты для дальнейших улучшений.

ЗАКЛЮЧЕНИЕ

В рамках данной работы было разработано приложение для предсказания успеваемости студентов на основе данных портала «Электронного ЮУрГУ», оно позволяет предсказывать успех студентов на основе их текущих оценок и статистики учебной деятельности и предоставляет инструмент для повышения эффективности обучения студентов и преподавателей. При этом решены следующие задачи.

1. Проведен анализ предметной области и на его основе разработаны модели предсказания.
2. Разработан алгоритм получения данных с применением существующих методов предобработки.
3. Разработано приложение, выполняющее предсказание на основе моделей с помощью различных методов анализа данных.
4. Проведены эксперименты, исследующие точность разработанных моделей данных.

Разработанное приложение имеет практическое применение и может быть полезно как студентам, так и преподавателям, позволяя им получать предсказания успеваемости и использовать их для улучшения учебного процесса. Кроме того, администрация университета может воспользоваться приложением для отслеживания прогресса студентов и разработки программ для улучшения образовательной среды.

ЛИТЕРАТУРА

1. Ahajjam T., Moutaib M., Aissa H., Azrour M., Farhaoui Y., Fattah M. Predicting Students' Final Performance Using Artificial Neural Networks. // Big Data Mining and Analytics, 2022. – Vol. 5, no 4. – 294–301 pp.

2. Aiohttp. [Электронный ресурс] URL: https://docs.aiohttp.org/en/stable/ (дата обращения: 10.02.2024 г.).

3. Asyncio. [Электронный ресурс] URL: https://docs.python.org/dev/library/asyncio.html (дата обращения: 10.02.2024 г.).

4. Django. [Электронный ресурс] URL: https://www.djangoproject.com/ (дата обращения: 10.02.2024 г.).

5. Jeganathan S., Lakshminarayanan A. R., Ramachandran N., Tunze G. B. Predicting Academic Performance of Immigrant Students Using XGBoost. // International Journal of Information Technology and Web Engineering (IJITWE), 2022. – Vol. 17, no 1. – 1–19 pp.

6. Jovanović J., Saqr MJoksimović S., Gašević D. Students matter the most in learning analytics: The effects of internal and instructional conditions in predicting academic success. // Computers & Education, 2021. – Vol. 172, no 1. – 1–13 pp.

7. Matplotlib. [Электронный ресурс] URL: https://matplotlib.org/ (дата обращения: 10.02.2024 г.).

8. Numpy. [Электронный ресурс] URL: https://numpy.org/ (дата обращения: 10.02.2024 г.).

9. Pandas. [Электронный ресурс] URL: https://pandas.pydata.org/ (дата обращения: 10.02.2024 г.).

10. Pitsia V. Examining high achievement in mathematics and science among post-primary students in Ireland: a multilevel binary logistic regression analysis of PISA data. // Large-scale Assessments in Education, 2022. – Vol.10, no 1. – 1–30 pp.

11. Python. [Электронный ресурс] URL: https://www.python.org (дата обращения: 10.02.2024 г.).

12. Riestra-González M., Paule-Ruíz Maria del Puerto, Ortin F. Massive LMS log data analysis for the early prediction of course-agnostic student performance. // Computers & Education, 2021. – Vol. 163, no 1. – 1–38 pp.

13. Saqr M., López-Pernas S., Helske S., Hrastinski S. The longitudinal association between engagement and achievement varies by time, students' profiles, and achievement state: A full program study. // Computers & Education, 2023. – Vol.199, no. 7. – 1–21 pp.

14. Sklearn. [Электронный ресурс] URL: https://scikit-learn.org/stable/ (дата обращения: 10.02.2024 г.).

15. Tenacity. [Электронный ресурс] URL: https://tenacity.readthedocs.io/en/latest/ (дата обращения: 10.02.2024 г.).

16. Ансамблевые методы. [Электронный ресурс] URL: https://scikit-learn.ru/1-11-ensemble-methods/# (дата обращения: 10.02.2024 г.).

17. Исходный проект системы. [Электронный ресурс] URL: https://github.com/Hlebnica/PredictingStudentPerformance (дата обращения: 24.05.2024 г.).

ПРИЛОЖЕНИЯ

Приложение А. Макет страницы обучения моделей предсказания

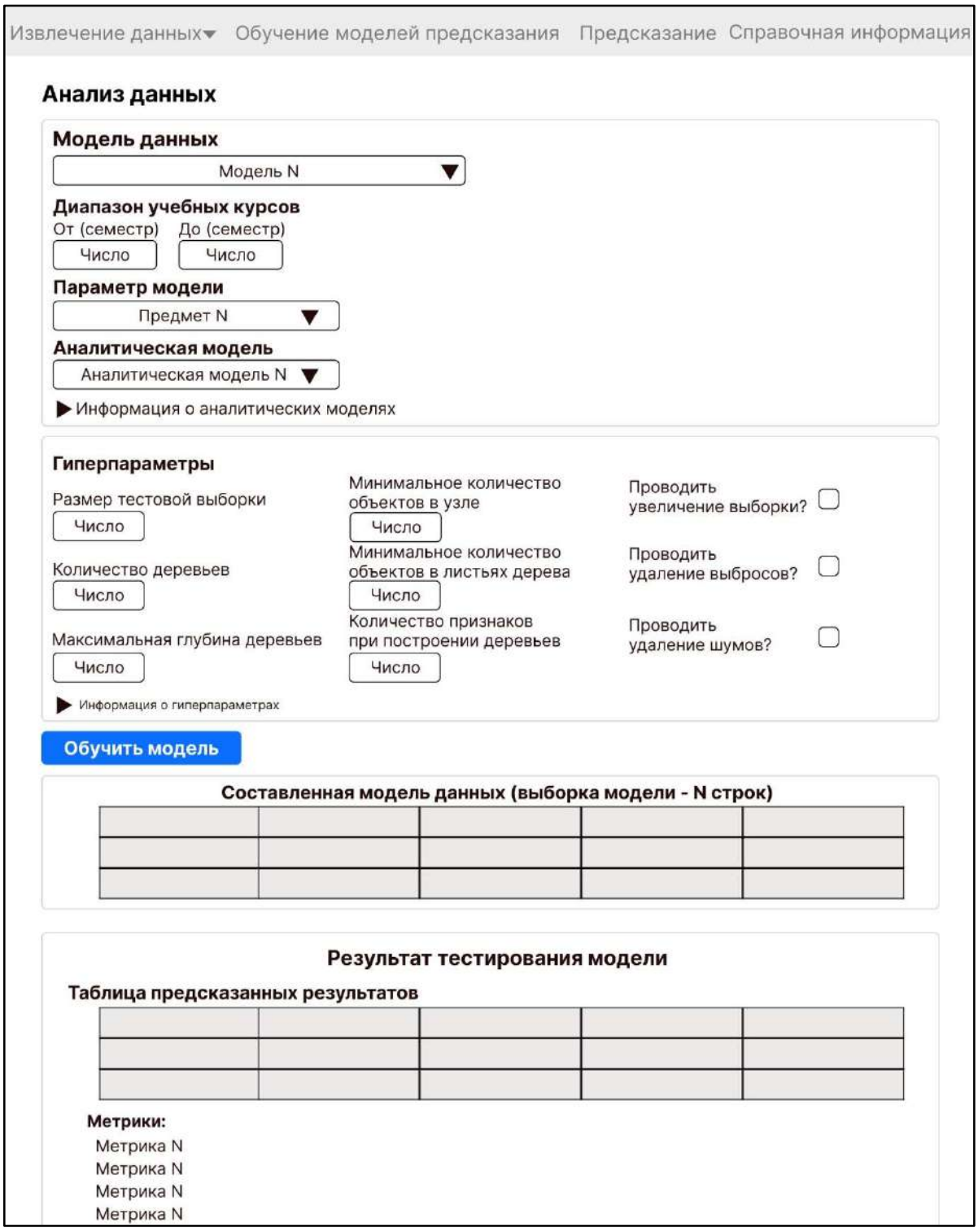

Рисунок 1 – Верхняя часть макета страницы анализа данных

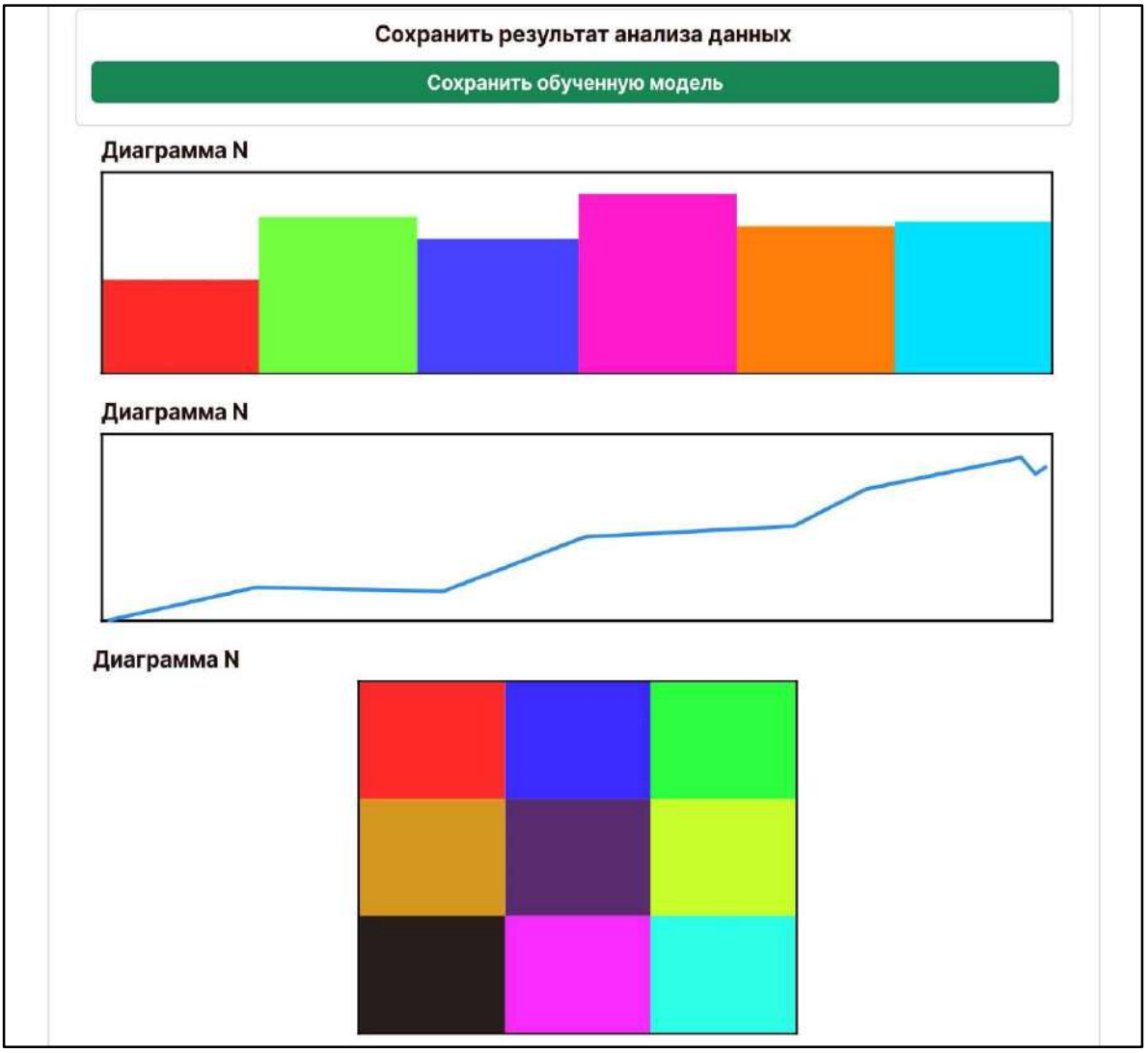

Рисунок 2 – Нижняя часть макета страницы анализа данных

Приложение Б. Итоговая реализация интерфейса обучения моделей предсказания

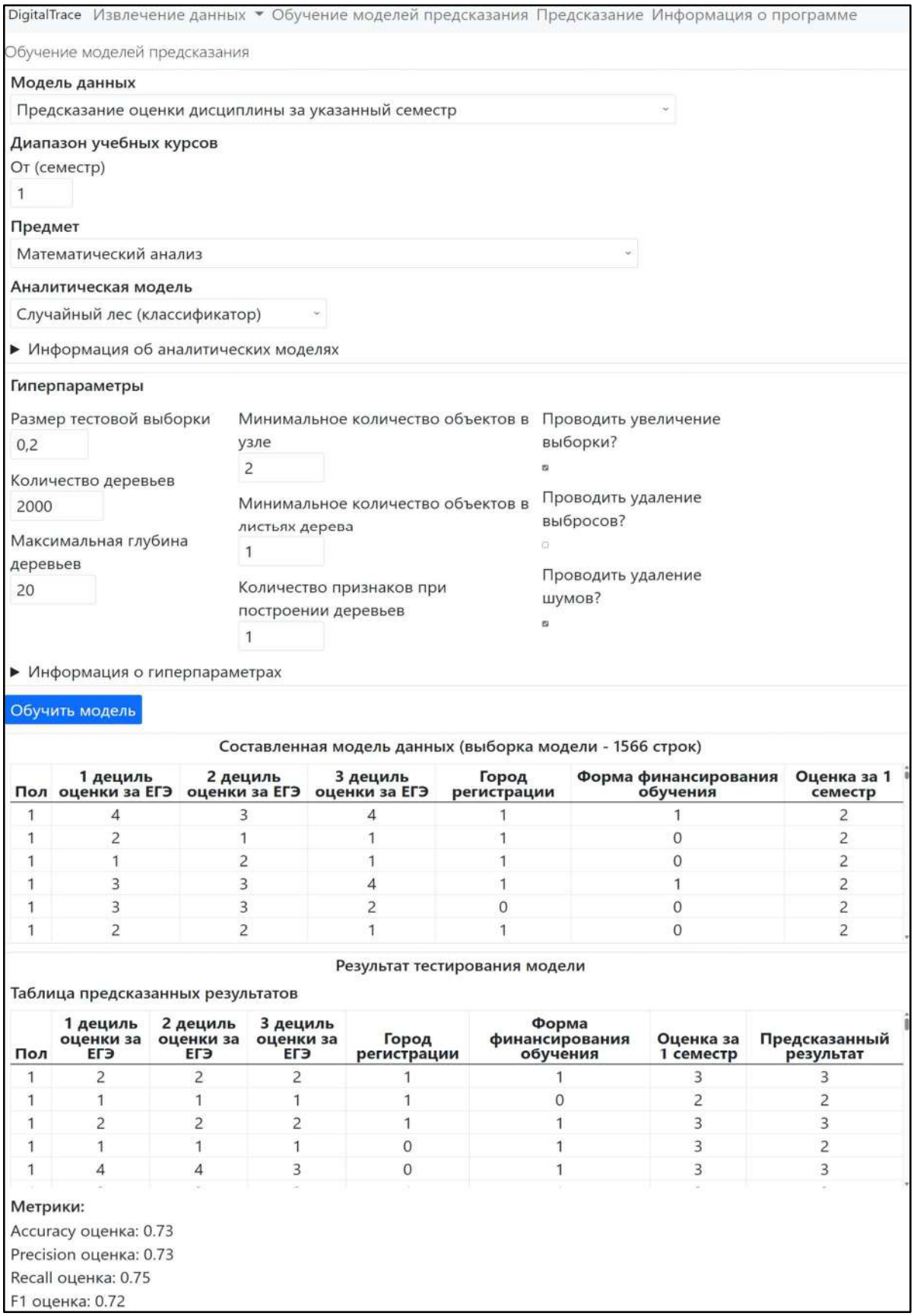

Рисунок 3 – Интерфейс обучения моделей предсказания

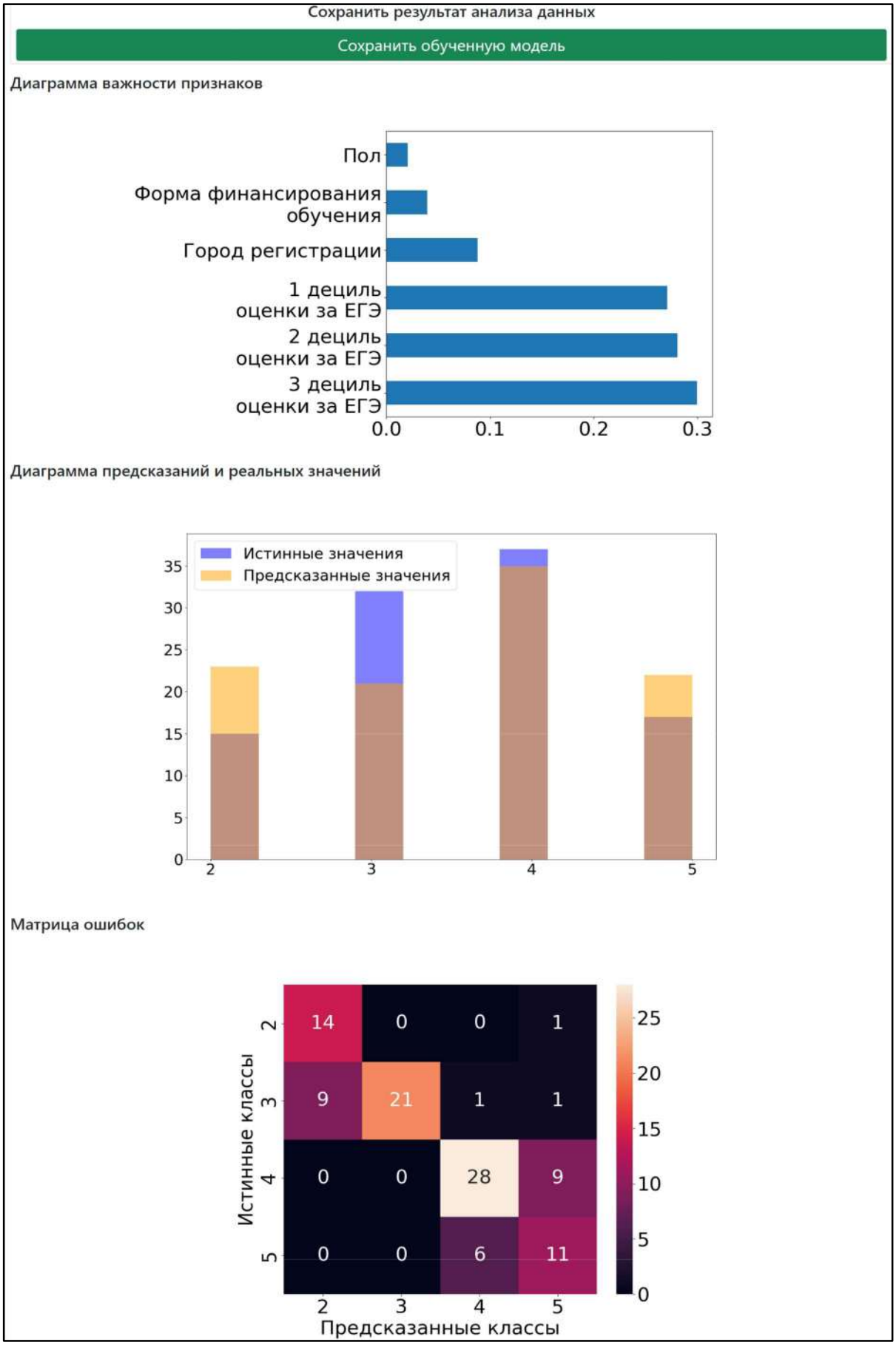

Рисунок 4 – Графики в интерфейсе обучения моделей предсказания

Приложение В. Результаты предсказательных экспериментов

Id студента	Пол	1 дециль оценки за ЕГЭ	2 дециль оценки за ЕГЭ	3 дециль оценки за ЕГЭ	Город регистрации	Форма финансирования обучения	Предсказанный результат
0b22c5e8-dcaf-4ab7-a11a-35f88004e09a	Мужской	1	1	1	Иногородний	Бюджет	2
107dff39-cc02-45cc-baaa-798da48ac31b	Мужской	5	5	5	Иногородний	Бюджет	5
1e85052e-d031-45b5-bb89-bea778599759	Мужской	2	2	1	Иногородний	Бюджет	3
2d32a7ab-8793-4e2a-bd34-774b693a24dc	Женский	2	2	1	Иногородний	Бюджет	3
4831659c-56f2-49fd-9d24-e151f3cb88db	Мужской	3	4	2	Иногородний	Бюджет	3
915c0f4c-3740-4547-a984-b66836a2439a	Женский	2	2	5	Иногородний	Бюджет	3
98478dfd-31c8-458f-8c47-0cbc74d06bf4	Мужской	2	4	5	Иногородний	Бюджет	3
abf5c2d5-b52e-4f75-9327-b4a66d95081a	Мужской	4	3	4	Иногородний	Бюджет	5
25b2f8c2-fd30-489b-ab40-cb4226f12c19	Мужской	1	1	1	Иногородний	Бюджет	2
3837d3ad-4842-4cd6-8e9c-e6fca609bd26	Мужской	4	3	5	Иногородний	Бюджет	3
5048d615-377f-4152-a9b4-fdfe3a71d4b1	Мужской	3	3	2	Иногородний	Бюджет	3

Рисунок 5 – Эксперимент предсказания оценки

1 дециль оценки за ЕГЭ	2 дециль оценки за ЕГЭ	3 дециль оценки за ЕГЭ	Город регистрации	Форма финансирования обучения	Дециль суммарного рейтинга за 1 семестр	Дециль медианного рейтинга по зачетам за 1 семестр	Дециль суммарного рейтинга за 2 семестр	Дециль медианного рейтинга по зачетам за 2 семестр	Предсказанный результат
1	1	1	Иногородний	Бюджет	3	4	3	3	Будет отчислен
5	5	5	Иногородний	Бюджет	3	5	3	3	Не будет отчислен
2	2	1	Иногородний	Бюджет	3	4	3	3	Не будет отчислен
2	2	1	Иногородний	Бюджет	2	3	3	2	Будет отчислен
3	4	2	Иногородний	Бюджет	3	5	3	3	Не будет отчислен
2	2	5	Иногородний	Бюджет	3	5	3	3	Не будет отчислен
2	4	5	Иногородний	Бюджет	3	5	3	3	Не будет отчислен
4	3	4	Иногородний	Бюджет	2	3	3	2	Не будет отчислен

Рисунок 6 – Эксперимент предсказания отчисления

2 дециль оценки за ЕГЭ	3 дециль оценки за ЕГЭ	Город регистрации	Форма финансирования обучения	Дециль суммарного рейтинга за 1 семестр	Дециль медианного рейтинга по зачетам за 1 семестр	Дециль суммарного рейтинга за 2 семестр	Дециль медианного рейтинга по зачетам за 2 семестр	Дециль суммарного рейтинга за 3 семестр	Дециль медианного рейтинга по зачетам за 3 семестр	Дециль суммарного рейтинга за 4 семестр	Предсказанный результат
1	1	Иногородний	Бюджет	2	3	2	2	2	2	2	2
1	3	Иногородний	Бюджет	3	2	5	4	2	2	4	5
2	1	Иногородний	Бюджет	3	2	4	3	2	3	2	2
1	1	Челябинск	Бюджет	4	4	5	4	2	2	2	2
4	2	Иногородний	Бюджет	2	2	5	4	2	2	4	4
2	2	Иногородний	Бюджет	3	3	5	4	2	2	2	2
3	2	Иногородний	Бюджет	4	3	5	4	2	2	2	2
1	1	Челябинск	Бюджет	2	2	5	4	2	2	2	2

Рисунок 7 – Эксперимент предсказания успеваемости по рейтингу