МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ «МОСКОВСКИЙ АВИАЦИОННЫЙ ИНСТИТУТ (национальный исследовательский университет)»

Институт (Филиал) № 8 «Компьютерные науки и прикладная математика» **Кафедра** 806

Группа М8О-408Б-20 **Направление подготовки** 01.03.02 «Прикладная математика и информатика»

Профиль Информатика

Квалификация: бакалавр

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА БАКАЛАВРА

на тему: Кластеризация учащихся с использованием машинного обучения на основе анкетных ответов в свободной форме

Автор ВКРБ:	Морозов Артем Борисович	(_____)
Руководитель:	Левинская Мария Александровна	(_____)
Консультант:	—	(_____)
Консультант:	—	(_____)
Рецензент:	—	(_____)

К защите допустить

Заведующий кафедрой № 806 Крылов Сергей Сергеевич (_____)

____ мая 2024 года

Москва 2024

РЕФЕРАТ

Выпускная квалификационная работа бакалавра состоит из 59 страниц, 50 рисунков, 10 использованных источников, 1 приложения.

УЧЕНИКИ, ПСИХОЛОГИЧЕСКИЕ ПОРТРЕТЫ, PYTHON, PANDAS, KMEANS, КЛАСТЕРИЗАЦИЯ, JUPYTER, АНАЛИЗ АНКЕТНЫХ ОТВЕТОВ

Объектом исследования в предоставленной дипломной работе являются ученики образовательных организаций, принимающие участие в национальных исследованиях качества образования.

Цель работы – кластеризовать учащихся на основе анкетных ответов в свободной форме при помощи машинного обучения, получить психологические портреты учащихся.

Дипломная работа состоит из следующих частей: очистки исходных данных, анализа и разбора свободных ответов учащихся, выделения набора параметров, по которым можно разделить учащихся, а также непосредственной кластеризации и формирования конкретных групп среди общей выборки.

Результатом проделанной работы являются кластеры учащихся, их психологические портреты – группы, внутри которых ученики максимально похожи между собой.

Результаты проведенного анализа предназначены для проведения оценки динамики развития школ и работы учителей в однородных группах по контексту.

Практическая значимость исследования заключается в нахождении зависимости успеваемости учащихся от контекстных факторов, полученные результаты позволят принимать более правильные управленческие решения в сфере образования.

СОДЕРЖАНИЕ

ТЕРМИНЫ И ОПРЕДЕЛЕНИЯ 4
ПЕРЕЧЕНЬ СОКРАЩЕНИЙ И ОБОЗНАЧЕНИЙ 5
ВВЕДЕНИЕ 6
1 АНАЛИЗ ПРЕДМЕТНОЙ ОБЛАСТИ 9
1.1 Необходимость оценки качества образования 9
1.2 Что такое НИКО 10
1.3 Цели и задачи исследования НИКО 10
1.4 Цели и задачи диагностики НИКО 12
1.5 Анкетирование и репрезентативная база данных НИКО 13
2 РАЗБОР И АНАЛИЗ СВОБОДНЫХ ОТВЕТОВ 17
2.1 Технологии для разбора свободных ответов 17
2.2 Очистка и преобразование изначальной базы данных 22
2.3 Разбор свободных ответов 25
2.4 Формирование индексов 30
2.5 Анализ полученной статистики 32
3 КЛАСТЕРИЗАЦИЯ 39
3.1 Технологии для кластеризации 39
3.2 Кластеризация при помощи полученных индексов 40
3.3 Кластеризация при помощи машинного обучения 45
3.4 Итоговый результат 50
ЗАКЛЮЧЕНИЕ 57
СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ 58
ПРИЛОЖЕНИЕ А Исходный код 59

ТЕРМИНЫ И ОПРЕДЕЛЕНИЯ

В настоящей выпускной квалификационной работе бакалавра применяют следующие термины с соответствующими определениями:

Репрезентативная база данных – база данных, в которой информация отражает многообразие и особенности исследуемых объектов

Датасет – структурированный набор данных

Датафрейм – двумерный массив с метками столбцов и строк, похожий на таблицу в реляционной базе данных

Кластер – группа объектов, которые обладают схожими характеристиками или свойствами

Кластеризация – одна из задач машинного обучения, которая подразумевает разделение исходной выборки на кластеры, в которых объекты больше похожи друг на друга, чем на объекты в других кластерах

Классификация – одна из задач машинного обучения, которая подразумевает присвоение объектам меток классов на основании их схожести между собой

Рособрнадзор – федеральный орган исполнительной власти, осуществляющий функции по контролю и надзору в сфере образования и науки

ПЕРЕЧЕНЬ СОКРАЩЕНИЙ И ОБОЗНАЧЕНИЙ

В настоящей выпускной квалификационной работе бакалавра применяют следующие сокращения и обозначения:

ОО – общеобразовательная организация

ПК – первичный ключ

ПБ – первичный балл

ФО – федеральный округ

ВКРБ – выпускная квалификационная работа бакалавра

РФ – Российская Федерация

НИКО – национальные исследования качества образования

ФИОКО – Федеральный институт оценки качества образования

ФГОС – Федеральный государственный образовательный стандарт

PIRLS – международное исследование качества чтения и понимания текста

TIMSS – международное мониторинговое исследование качества школьного математического и естественнонаучного образования

PISA – международная программа по оценке образовательных достижений учащихся

ВВЕДЕНИЕ

В различных стратегических документах, ориентированных на развитие образования в РФ, едва ли не самой главной целью является повышение качества образования.

Вопрос качества образования является ключевым аспектом в таких документах, как, например, государственная программа «Развитие образования на 2018–2025 годы», а также национальный проект «Образование» на период с 2018 по 2024 год.

Приоритетами в этих документах являются сохранение и укрепление лидирующих позиций России в международных рейтингах качества образования, таких как PIRLS, TIMSS и PISA. Вместе с этим стремление к улучшению качества образования прослеживается не только глобально, но и в чуть более узконаправленных инициативах, таких как, например, федеральный проект «Современная школа».

В рамках подпрограммы «Совершенствование управления системой образования» акцент делается, в основном, на обеспечении участия России в международных исследованиях качества образования, а также на совершенствовании процедур оценки уровня освоения образовательных программ.

В этом контексте особую важность приобретает разработанная методология и критерии оценки качества образования, которые, в свою очередь, утверждены компетентными органами РФ и направлены на соответствие требованиям федеральных образовательных стандартов.

Все эти усилия направлены на создание сбалансированной, качественной и надежной системы оценки качества образования, которая позволит стабильно получать надежную и достоверную информацию о состоянии образовательных систем на различных уровнях – от региональных и до муниципальных.

Такая система по своей задумке должна также обеспечивать возможность оценки состояния отдельных компонентов системы образования, затрагивая каждый, даже самый труднодоступный уголок РФ.

Результатом введенных новшеств и разработанных проектов и систем должно являться не только повышение качества образования в РФ, но и улучшение конкурентоспособности российской системы образования на мировой арене в целом.

Таким образом, эффективное решение данной задачи заключается в регулярном проведении исследований по качеству образования на всех уровнях. Эти исследования основаны на сборе и анализе обширного объема данных о состоянии систем образования. И программа НИКО является одним из самых ярких примеров такого подхода.

Данная ВКРБ служит довольно крепким фундаментом для дальнейших, более точных исследований НИКО по части качества образования и его объективной оценки.

Главной целью дипломной работы является получение кластеров учащихся 6-8 классов, принимавших участие в исследовании НИКО, чтобы оценивание качества образования проводилось не на уровне общей школьной или учебной группы, а внутри независимых групп учащихся, в которых они имеют схожие начальные характеристики и контекстные данные.

Дипломная работа актуальна и с теоретической, и с практической точек зрения, так как нередки ситуации, когда заслуги конкретно взятых преподавателей или ОО и, наоборот, их недостатки «нивелируются» необъективным подходом к оценке качества проделанной ими работы. Так происходит в силу того, что педагогам приходится иметь дело с абсолютно разными по набору характеристик и человеческих качеств учениками. При этом в одной и той же школе, в одном и том же классе, зачастую, учатся так же совершенно разные дети.

В процессе выполнения работы были решены следующие задачи:

– была проведена очистка данных в деперсонифицированной репрезентативной базе НИКО, позволившая сформировать более подходящий датасет для исследования;

– были разобраны анкетные ответы учащихся при помощи машинного обучения и преобразованы в категориальные переменные;

– была проведена кластеризация при помощи машинного обучения учащихся исходя из их контекста.

Основным результатом работы являются кластеры школьников – их психологические портреты, полученные из общей выборки учащихся, участвовавших в анкетировании НИКО 2022.

Данная работа является частью большого проекта под названием «Кластеризация общеобразовательных организаций с учетом психологического портрета учащихся». Результаты классификации анкетных ответов учащихся, полученные на одном из этапов этой работы, были переданы второму участнику проекта для исследования тенденций в ОО и их последующей кластеризации.

Результаты ВКРБ направлены на оценку функционирования школ и работы учителей в группах с сопоставимым контекстом.

Эти результаты могут быть применены для уточнения образовательных программ ОО и более гибкого определения успехов и динамики обучения в группах, выявленных в процессе работы, а также для дальнейших исследований НИКО в будущем.

1 АНАЛИЗ ПРЕДМЕТНОЙ ОБЛАСТИ

1.1 Необходимость оценки качества образования

Проект ФИОКО, занимающийся обеспечением информационно-аналитического и методического сопровождения исследований качества образования всех уровней и созданный в октябре 2015 года, в своих отчетах ежегодно подчеркивает необходимость оценки качества образования и имеющихся на текущий момент обучающих курсов.

Более того, если подойти к данному вопросу глобальнее, то закон «Об образовании в Российской Федерации» и вовсе определяет воспитание и образование учащихся как многогранные процессы, охватывающие не только передачу знаний, но и формирование ценностных ориентиров, развитие нравственных и социокультурных убеждений, уважение к общечеловеческим ценностям, нравственным нормам и установленным порядкам.

В контексте современной ОО образование и воспитание представляют собой активную деятельность, направленную на создание условий для полноценного развития личности каждого ученика, его самореализации и успешной интеграции в общество, а в дальнейшем во взрослую жизнь.

Такой подход к ОО открывает множество новых горизонтов и возможностей для систематической оценки эффективности образовательного процесса.

При этом важно учитывать не только уровень усвоения знаний, но и степень внутреннего развития конкретно взятого человека, его способность к самостоятельному и критическому мышлению, анализу ситуации и принятию осознанных решений. Это подразумевает оценку образовательных достижений учеников не только в контексте формальных учебных программ, но и с учетом их развития как личностей.

Кроме того, современное образование должно поддерживать активный процесс самообразования, поэтому оценка образования должна также включать анализ готовности учеников к самостоятельному поиску и

использованию информации, умению быстро адаптироваться к меняющимся условиям современного мира. Ученики не должны бояться новых технологий, а, наоборот, с энтузиазмом познавать ранее неизведанные аспекты той или иной отрасли жизни.

Таким образом, оценка образования становится неотъемлемой частью образовательного процесса, направленного на формирование всесторонне развитой личности.

1.2 Что такое НИКО

Согласно стратегии научно-технологического развития РФ, основная цель научно-технологического прогресса в стране заключается в создании эффективной системы, способствующей независимости и конкурентоспособности России. Чтобы достичь этой цели, необходимо полностью использовать интеллектуальный потенциал страны и обеспечить условия для проявления талантливой молодежи в сфере технологий и инноваций. Одним из инструментов для достижения этой задачи, как уже было оговорено ранее, являются НИКО [2].

Эта программа, начавшаяся в 2014 году по инициативе Рособрнадзора, осуществляет регулярные исследования по различным учебным предметам на различных уровнях общего образования.

В рамках НИКО проводятся диагностические работы и анкетирование учащихся. Для выбора участников исследования используется репрезентативная выборка с частичным квотированием по различным параметрам, включая федеральные округа, типы и виды ОО.

1.3 Цели и задачи исследования НИКО

Основными целями исследования, рассматриваемого НИКО, являются:

– оценка достижения установленных ФГОС результатов обучения в основной школе;

– идентификация текущих проблем в воспитательной работе с учащимися и формулирование рекомендаций для их решения на уровнях федеральных, региональных, муниципальных и школьных программ;

– анализ результатов национальных исследований качества образования в 6 и 8 классах по достижению личностных результатов, включая готовность к саморазвитию и личностному самоопределению, мотивацию к обучению, социальные компетенции и правосознание;

– оценка сформированности метапредметных результатов, включая умения и способности в области регулятивных, познавательных и коммуникативных действий, и их применение в учебной, познавательной и социальной практике;

– анализ эффективности методик и подходов, используемых в образовательном процессе, на достижение поставленных целей;

– исследование долгосрочных перспектив в области образования и разработка стратегий для адаптации образовательной системы к изменяющимся потребностям общества и рынка труда.

Основными задачами исследования НИКО являются:

– оценка уровня подготовки учащихся 6 и 8 классов в образовательных учреждениях, реализующих программы общего образования, через комплексную диагностику;

– сбор, анализ и обработка информации, описывающей образовательные процессы в учебных заведениях;

– подготовка аналитических отчетов на основе результатов исследования;

– разработка рекомендаций по использованию полученных данных;

– организация общественно-профессионального обсуждения результатов исследования.

1.4 Цели и задачи диагностики НИКО

Цели диагностики НИКО определяются в соответствии с основными принципами образовательной программы и направлены на оценку уровня усвоения учащимися основных знаний, умений и навыков, предусмотренных учебным планом.

Основными целями диагностики НИКО являются:

– содействие развитию уважительного отношения к культурному, религиозному, социальному и мировоззренческому многообразию;

– содействие активному участию в общественной жизни и школьном самоуправлении;

– способствование формированию морального сознания через осознанное принятие нравственных ценностей;

– стимулирование участия в различных образовательных и общественных мероприятиях, способствующих формированию лидерских навыков.

Диагностические задания НИКО, в свою очередь, формируются в соответствии с требованиями Федерального государственного образовательного стандарта для образования в РФ.

Основными задачами диагностики НИКО являются:

– патриотическое воспитание: уважение к Родине и ее наследию, освоение ценностей гуманизма и демократии, зарождения чувства ответственности перед Отчизной;

– формирование ответственного отношения к обучению: готовность к саморазвитию и самообразованию, а также осознанный выбор профессионального пути;

– формирование уважения и чувства долга перед семьей, учение о семье как о важной ценности в жизни человека;

– развитие эстетического восприятия: изучение художественного наследия народов России.

1.5 Анкетирование и репрезентативная база данных НИКО

Как было сказано выше, НИКО ежегодно проводит одинаковое анкетирование для учащихся 6-8 классов, собирая деперсонифицированную репрезентативную базу данных. В данной работе используется база НИКО 2021-2022.

Так, после прохождения учеником анкетирования, можно многое узнать о его личности. Например, на рисунке 1 представлены вопросы семейного характера, на которые ученик отвечает в процессе опроса:

Расскажите немного о своей семье.

Кто в неё входит?

_____

Кем из членов семьи Вы гордитесь? Объясните, почему.

_____

Какие традиции существуют в Вашей семье?

_____

Что Вы обычно делаете вместе с другими членами семьи?

_____

Рисунок 1 – Вопросы про семью в анкетировании НИКО

Помимо семейных вопросов, учащиеся делятся информацией как о своем населенном пункте, так и о стране в целом, как на рисунках 2 и 3.

Расскажите немного о своём населённом пункте.

Как называется Ваш населённый пункт?

_____

Какие достопримечательности есть в Вашем населённом пункте?

_____

Что, по Вашему мнению, необходимо сделать в первую очередь, чтобы улучшить жизнь в Вашем населённом пункте?

_____

Рисунок 2 – Вопросы про населенный пункт в анкетировании НИКО

Расскажите немного о нашей стране.

Какими военными победами нашей страны Вы гордитесь?

_____

Какие российские деятели искусства прошлого и настоящего известны всему миру?

_____

Рисунок 3 – Вопросы про нашу страну в анкетировании НИКО

В анкетировании есть и такие вопросы, ответы на которые не вошли в итоговую репрезентативную базу. Пример таких вопросов мы можем увидеть на рисунке 4 и на рисунке 5.

29. Согласны ли Вы со следующими утверждениями?

Дайте ответ в каждой строке.

	НЕТ	Скорее НЕТ, чем ДА	Скорее ДА, чем НЕТ	ДА
Я часто чувствую себя одиноким				
Я чувствую поддержку со стороны окружающих в трудных ситуациях				
Я часто чувствую непонимание со стороны окружающих				
В школе у меня есть друзья				

30. Согласны ли Вы со следующими утверждениями?

Дайте ответ в каждой строке.

	НЕТ	Скорее НЕТ, чем ДА	Скорее ДА, чем НЕТ	ДА
В случае трудностей я всегда могу обратиться к своим родным				
Я часто ссорюсь с родителями и (или) другими членами моей семьи				
У меня хорошие отношения с родителями (или с теми, кто их заменяет)				
Дома я чувствую себя очень спокойно				

Рисунок 4 – Некоторые вопросы из анкетирования НИКО

3. Согласны ли Вы со следующими утверждениями?

Дайте ответ в каждой строке.

	НЕТ	Скорее НЕТ, чем ДА	Скорее ДА, чем НЕТ	ДА
Я интересуюсь информацией о здоровье и здоровом образе жизни				
Вредные привычки человека могут вредить не только его здоровью, но и здоровью окружающих				
Следование правилам здорового образа жизни поможет мне добиться успехов в будущем				
Здоровье мало зависит от образа жизни: оно либо дано от природы, либо нет				
Способы лечения многих болезней можно найти в Интернете, поэтому к врачу идти необязательно				
Я знаю, как оказать первую помощь при травме				
Я знаю, как вызвать скорую медицинскую помощь				
Если у меня будет выбор, как провести появившееся свободное время – выйти на прогулку или посидеть за компьютером (со смартфоном), – я, пожалуй, выберу последнее				
Бывает, что я употребляю слабоалкогольные напитки (пиво, коктейли и др.)				
Я ограничиваю себя в употреблении сладких газированных напитков, мучного и сладкого				
В моей семье регулярны занятия физкультурой и спортом: зарядка, фитнес, бег, лыжи, пешие прогулки, подвижные игры и т. д.				

Рисунок 5 – Вопрос из анкетирования НИКО

Таким образом был получен набор данных НИКО 2021-2022 в 45 тысяч записей, на основании которого строились исследования в представленной дипломной работе.

На рисунке 6 и рисунке 7 мы можем наблюдать «сырой» вид данных:

№	Класс	Пол	РУ Теку	МА Тек	Расскажите немного о своей семье. Кто в неё входит?	Кем из членов семьи Вы гордитесь? Объясните, почему.	Какие традиции существуют в Вашей семье?	Что Вы обычно делаете вместе с другими членами семьи?	Как называется Ваш населённый пункт?
1964193	6	Мальчики	3	4	Мама, папа,2 брата ,и д	Мамой. Она помога	В новый год ждат	Общаюсь ,гуляю с н	Санкт-Петербург
1964588	6	Мальчики	4	4	папа мама брат я и две	мамой потому что с	их нет	готовим	санкт питербург
1964265	6	Девочки	3	4	отчим,мама,я,сестричка	отчем- он работает	их нет	гуляем,ужинаем.	Санкт- Петербург
1964812	6	Девочки	4	4	мама брат тетя двоюро,	никем	нету	могу лишь побыть с	красносельский район
1965005	6	Девочки	4	3	мама папа и я	мама и папа	мы путишествуем	гуляем веселимся и	Санкт Петербург
1964478	6	Девочки	4	5	мама, папа, сестра,	Я горжусь всей свое	у нас много тради	мы с моей семьей к	Санкт-Петербург, Крас
1964973	6	Девочки	4	3	мама папа брат сестра	папой потому что о	Пасха, Новый год,	гуляем по городу, с	Санкт-Петербург
1965931	6	Девочки	4	3	мама папа я брат бабуш	я горжусь своей сем	никаких	гуляем	Санкт Петербург
1964983	6	Мальчики	3	4	папа мама дедушка баб	сынам мамы братот	стричать новый го	играем едим и гул	санкт питербур
1965954	6	Девочки	5	4	мама,папа,бабушка	Бабушка потому что	никакие	идем в магазины	Санкт-Петербург
1966120	6	Девочки	5	5	Все + старший брат	всеми, потому что я	Очень популярные	Проводим все обыч	Санкт-Петербург.
1968491	6	Мальчики	2	3	мама ,папа, брат, бабуш	дедушкой потому чт	-	-	-
1966117	6	Мальчики	4	4	родственники	отцом, так как он д	-	занимаемся своими	Санкт-Петербург
1966125	6	Девочки	5	4	Я, мама, папа.	Я горжусь каждым с	В нашей семье мн	Гуляем, веселимся.	Санкт-Петербург.
1967149	6	Девочки	3	3	я (младшая дочь) Ира (м	я горжусь всеми, по	отмечать всем вме	с сестрой и собакой	Санкт Петербург
1966123	6	Девочки	3	3	мама,бабушка,прабабуш	всеми, потому что о	помогать кому ни	розговариваем	город Санкт-петербург
1967344	6	Мальчики	5	5	Папа,мама,бабушка,дед	Папой.Потому что о	Поздравлять всех	Гуляем,общаемся.	Санкт-Петербург
1963858	6	Мальчики	4	3	папа мама сестра 2 баб	папой по тому что о	отмечать новый го	играю в игры разгов	Санкт-Петербург
1967728	6	Мальчики	3	4	мама,папа,сестра и я	мамой и папой они	день рождения,ме	гуляем , проводим в	Всеволожск Микрорай
1966078	6	Девочки	4	4	мама, папа, я, 2 бабушк	папой и мамой они	на пасху бабушка	гуляем, ходим по ма	Санкт-Петербург
1967752	6	Мальчики	4	4	мама папа сестра я	папа зарабатывает	нету	играем	санкт петербург
1967187	6	Мальчики	3	3	мама сестра бабушка п	НЕ буду говорить	Нету	НЕ буду говорить	кудрово
1967372	6	Девочки	3	3	мама, брат, папа	брат, мама. брат ум	праздновать праз	общаюсь, гуляю.	Санкт-Петербург
1967718	6	Девочки	3	3	мама папа бабушка бра	мамой. она много р	на новый год мы г	говорим как дела на	санкт петербург

Рисунок 6 – Исходный набор данных

Регион_т	Какие достопримечательности есть в Вашем населённом пункте?	Что, по Вашему мнению, необходимо сделать в первую очередь, чтобы улучшить жизнь в Вашем населённом пункте?	Какими военными победами нашей страны Вы гордитесь?	Какие российские деятели искусства прошлого и настоящего известны всему миру?
г. Санкт-Петербург	фонтаны эрмитаж петр	больше дорог, больше освящен	победа в Великой Отече	Анна Павлова, Ломоносов
г. Санкт-Петербург	эрмитаж,александрискийй	столб ,аничков дворед,синий	победа в 1945	
г. Санкт-Петербург	Спас на Крови, Исаакиев	Не мусорить или подключать бс	Великая Отечественная	Шишкин, Пушкин
г. Санкт-Петербург	кунск камера ,зимний дв	убирать за собой мусор	война 1941-1945	Александр Сергеевич Пушкин
г. Санкт-Петербург	Разбитое кольцо,статуя	Улучшать космическую дейтель	Победой над Наполеон	Незнаю
г. Санкт-Петербург	Исаакиевский собор, Ка	Позаботиться об экологии.	Отечественная война, В	Шишкин, Чайковский, Глинка,
г. Санкт-Петербург	Казанский собор,собор	построить экологические мусор	Великая Отечественая в	скульптуры
г. Санкт-Петербург	Разорванное кольцо. Ра	повысить зарплату и понизить н	победой во второй мир	Шишкин Лермонтов Пушкин
г. Санкт-Петербург	Александрийская колон	Убрать весь мусор по улицам	С 1941-1945 года Велик	Михаил Юрьевич Лермонтов
г. Санкт-Петербург		уехать		
г. Санкт-Петербург	Зимний дворец, Эрмита	добавить уличных туалетов	Я горжусь РОССИЕЙ. Мь	Я не помню.
г. Санкт-Петербург	их очень много. наприм	убираться на улице , сократить	победой в 2 мировой в	шишкин, пушкин
г. Санкт-Петербург	на улице Бульвар Менде	неплохо для начала узнать, что	победа в ВОв 1812 года	я сомневаюсь, что он известен за
г. Санкт-Петербург	Эрмитаж, Медный всадн	отчистить леса, озёра и моря от	Велика Отечественная в	писание стихов и рассказов
г. Санкт-Петербург	медный всадник, мосты	улучшить инфроструктуру	победа во 2 мировой в	я не знаю
г. Санкт-Петербург	много красивых цветов	построит какие-нибудь строите	победа над вашистами	-
г. Санкт-Петербург	Эрмитаж, Медные всадн	Помыть старые красивые здани	победой над золотой о	А.С. Пушкин, Рахманинов
г. Санкт-Петербург	парки, дворцы, музеи	построить торговый центр	Великая Отечественная	Пушкин, Менделеев
г. Санкт-Петербург	Исакиевский собор, Мед	Убирать за собой мусор. Не губ	Великая Отечественная	Пётр Первый.
г. Санкт-Петербург	В моем городе их очень	Надо следить за экологией	вторая мировая война	Станиславский, Репин и еще
г. Санкт-Петербург	Петропавловская крепос	Помочь людям нуждающимся	Те кто победил немцов	Александр Сергеевич Пушкин
г. Санкт-Петербург	памятник петру первому	скейт парки для самокатеров и	победой в великую оте	александер сергеевич пушкин
г. Санкт-Петербург	их много	ни чего всё идеально	2 мировой войной,возм	А.С.Пушкин-великий писатель.
г. Санкт-Петербург	не знаю	ничего, я считаю тут и так хорош	Вторая Мировая война	Александр Сергеевич Пушкин

Рисунок 7 – Исходный набор данных

Как мы можем видеть, помимо полученных ответов от учеников, в репрезентативной базе НИКО содержатся поля, содержащие класс, пол ученика, его ПК в базе данных, школьные оценки по русскому языку и математике, а также проставленный работниками НИКО на основании ответа учащегося его регион.

Также в базе НИКО, в файле «Диагностическая работа», содержится поле ПБ для каждого учащегося в 20-балльной системе. Если оценки по русскому языку и по математике – это результат их учебы исключительно в пределах своих школ, оцениваемый по 5-балльной системе, то ПБ – это результат диагностики НИКО, объективная метрика успеваемости каждого конкретно взятого ученика.

В дальнейшем значение ПБ для нас будет очень важно, так как это одна из основополагающих характеристик, которая будет наглядно показывать отличия учеников в одном кластере от учеников в других кластерах.

2 РАЗБОР И АНАЛИЗ СВОБОДНЫХ ОТВЕТОВ

2.1 Технологии для разбора свободных ответов

В данной ВКРБ для разбора свободных ответов были использованы регулярные выражения и алгоритмы машинного обучения для задач классификации.

Регулярные выражения – это очень мощный инструмент для работы с текстовыми данными и поиском в тексте какого-то заранее известного паттерна. Если говорить в общем, то регулярные выражения можно описать как последовательности символов, используемые для поиска и манипуляций с текстом на основе заданных шаблонов.

Регулярные выражения в зависимости от их предназначения и функционала могут быть условно разделены на следующие:

– обычные символы, которые при поиске шаблона в тексте сопоставляются сами с собой;

– специальные символы, такие как точка, которая используется для обозначения любого символа;

– символы-множители, которые могут обозначать количество повторений предшествующего выражения – например, символ «+» соответствует одному и более повторениям;

– символы классов – например, [a-z] соответствует любой строчной букве от a до z;

– альтернативные регулярные выражения – например, a|b может соответствовать как строчной букве a, так и строчной букве b;

– якоря, которые помогают в том случае, когда, предположим, нужно четко обозначить начало и конец строки;

– другие типы, использующиеся редко и в более узких целях.

На рисунке 8 представлены шаблоны некоторых регулярных выражений:

Образцы шаблонов	
([A-Za-z0-9-]+)	Буквы, числа и знаки переноса
(\d{1,2}\/\d{1,2}\/\d{4})	Дата (напр., 21/3/2006)
([^\s]+(?=\.(jpg|gif|png))\.\2)	Имя файла jpg, gif или png
(^[1-9]{1}$\|^[1-4]{1}[0-9]{1}$|^50$)	Любое число от 1 до 50 включительно
(#?([A-Fa-f0-9]){3}(([A-Fa-f0-9]){3})?)	Шестнадцатиричный код цвета
((?=.*\d)(?=.*[a-z])(?=.*[A-Z]).{8,15})	От 8 до 15 символов с минимум одной цифрой, одной заглавной и одной строчной буквой (полезно для паролей).
(\w+@[a-zA-Z_]+?\.[a-zA-Z]{2,6})	Адрес email
(\<(/?[^\>]+)\>)	HTML теги

Рисунок 8 – Шаблоны некоторых регулярных выражений

. При всех своих очевидных достоинствах, таких как гибкость и универсальность, регулярные выражения обладают рядом недостатков, которые иногда могут очень сильно затруднить работу с ними. Вот некоторые из главных недостатков регулярных выражений как инструмента для работы с текстом:

– в силу своего многообразия регулярные выражения могут быть очень трудны для понимания и для чтения, так как требуют должной теоретической подкованности и опыта работы с ними;

– сложный синтаксис, из-за которого при написании сложного регулярного выражения могут возникнуть неочевидные проблемы;

– при всей своей универсальности регулярные выражения, по-хорошему, не должны являться первым выбором для задач разбора вложенных текстов или иерархических структур.

Иными словами, регулярные выражения хороши для поиска какого-то паттерна в тексте, сравнительно быстры в написании и не требуют дополнительных затрат, однако реальный мир гораздо сложнее, и зачастую

нужно не просто найти какой-то шаблон в корпусе текстов, а классифицировать эти корпуса по содержанию, то есть присвоить каждому тексту одной или нескольких заранее определенных категорий или меток на основе его содержания. К ярким примерам таких задач можно отнести, например, определение, является ли электронное письмо спамом, определение тона сообщения как негативного, нейтрального или позитивного, а также разбор новостей по тематикам – спорт, политика, искусство и т.д.

Задача разбора свободных ответов может быть сведена к задаче классификации текста, так как нам так же нужно сделать какие-то выводы об ответе учащегося на тот или иной анкетный вопрос в виде метки, категориальной переменной, которая отражала бы богатство его ответа. Очевидно, что для таких задач регулярные выражения довольно ограничены в своем функционале, так как не учитывают тональность текста и его непосредственную суть, а просто выполняют наивный поиск без учета контекста. В этот момент на помощь приходит машинное обучение, как на рисунке 9:

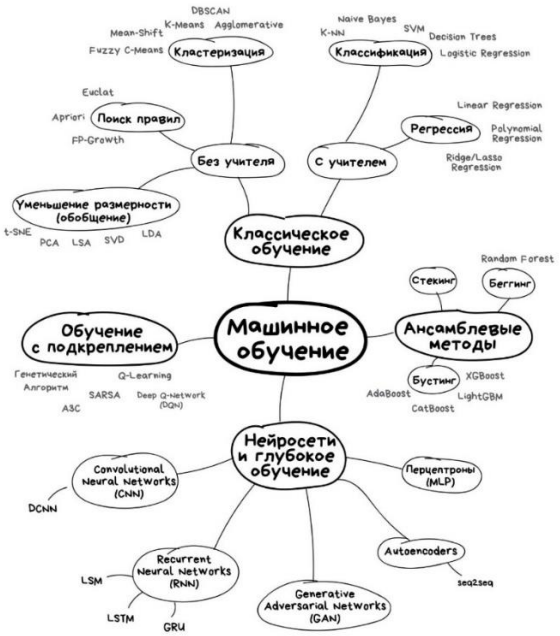

Рисунок 9 – Мир машинного обучения

Машинное обучение – это раздел искусственного интеллекта, который занимается разработкой алгоритмов и моделей, которые позволяют компьютерам обучаться на основе некоторой выборки и принимать решения без явного программирования. Вместо того, чтобы явно задавать компьютеру инструкции, как решать определенную задачу, в машинном обучении мы предоставляем алгоритму большой объем данных, на основании которых он самостоятельно настраивает параметры модели, чтобы лучше выполнять поставленную задачу.

Машинное обучение широко применяется в разных областях, таких как обработка естественного языка, медицина, финансы, прогнозирование и т.д.

Основные три типа задач машинного обучения – это задача классификации, задача регрессии и задача кластеризации. Задачи классификации и регрессии относятся к методам машинного обучения с учителем – это значит, что для выполнения этих задач требуется набор данных, который содержит как входные признаки, так и соответствующие им выходные метки или целевые значения. Методы машинного обучения с учителем используют этот набор данных для обучения модели, которая может классифицировать новые данные на основе уже известных примеров.

Задача же кластеризации относится к методам машинного обучения без учителя, так как она не требует наличия заранее определенных меток или ответов для данных. Вместо этого алгоритмы кластеризации группируют данные на основе их сходства без учитывания целевых значений.

Этот тип задач является примером обучения без учителя, поскольку модель не обучается на основе предоставленных меток, а, напротив, находит внутренние закономерности, необходимые для кластеризации.

Возвращаясь к задаче классификации, важно отметить, что в классической ее постановке имеется набор объектов, каждый из которых описывается набором признаков. Эти признаки могут быть числовыми,

категориальными или даже текстовыми. Для каждого объекта известна его метка класса, то есть категория, к которой он относится.

Целью алгоритма классификации является построение модели, способной классифицировать новые или неизвестные объекты на основе их признаков. Эта модель обучается на обучающем наборе данных, где для каждого объекта известна его метка класса. Затем модель проверяется на тестовом наборе данных, который не использовался при обучении, с целью оценить производительность и качество полученной модели.

В настоящее время существует очень и очень много алгоритмов машинного обучения для подобных задач, каждый из которых имеет свои особенности и применения.

Несмотря на многообразие алгоритмов машинного обучения для задач классификации, можно выделить следующие методы:

– метод логистической регрессии, суть которого заключается в моделировании вероятности принадлежности объекта к определенному классу с использованием логистической функции;

– метод опорных векторов, суть которого заключается в поиске оптимальной разделяющей гиперплоскости между классами;

– метод наивного байесовского классификатора, основанный на теореме Байеса и предполагающий независимость между признаками;

– метод K-ближайших соседей, основанный на принципе близости, где объекты классифицируются на основе классов их ближайших соседей в пространстве признаков;

– нейросетевой метод, основанный на применении какой-либо из нейросетей для задачи классификации;

– метод случайного леса, основанный на построении множества деревьев решений во время обучения и выдающий итоговую классификацию путем голосования или усреднения результатов всех деревьев.

Задача же классификации текстов является подмножеством задач классификации и может быть успешно решена одним из вышеописанных методов машинного обучения.

2.2 Очистка и преобразование изначальной базы данных

Как было сказано выше, в изначальной выборке НИКО 2021-2022 содержится около сорока пяти тысяч записей, однако перед началом работы с ней нужно провести очистку данных, а также для удобства работы с ней в интерактивной среде программирования Jupyter Notebook выполнить некоторые преобразования [7].

Проблема изначальной выборки заключается в том, что в репрезентативной базе содержится немалое количество записей с множеством пустых полей, а также полей с прочерками, несвязными символами и т.д. Это связано с тем, что далеко не все учащиеся 6-х и 8-х классов по-честному проходили анкетирование и делились информацией о себе. Примеры таких записей мы можем наблюдать на рисунке 10.

№	Расскажите немного о своей семье. Кто в неё входит?	Кем из членов семьи Вы гордитесь? Объясните, почему.	Какие традиции существуют в Вашей семье?	Что Вы обычно делаете вместе с другими членами семьи?	Какие достопримечательности есть в Вашем населённом пункте?	Что, по Вашему мнению, необходимо сделать в первую очередь, чтобы улучшить жизнь в Вашем населённом пункте?	Какими военными победами нашей страны Вы гордитесь?	Какие российские деятели искусства прошлого и настоящего известны всему миру?
1965914	мать отцец			живем		уехать		
1965931	мама папа я брат бабуш	я горжусь своей сем	никаких	гуляем				
1968491	мама ,папа, брат, бабуш	дедушкой потому чт	-	-	-	-	-	-
1966117	родственники	отцом, так как он дл	-	занимаемся своими	например Исаакиевский	привести его в порядок (имеетс	победа во 2-ой мирово	Шишкин к примеру
2071085								
2070769								
2070840	Мама,я,старший брат.	Мама	незнаю					
2073990	Мама							
2074371	Мама, бабушка, дедуш	Некем	Никакие				Великая отечественная	война
2080641	мама, дедушка, младша	я горжусь своей ма	каждое лето ездит	в свободное время мы смотрим фотоальбом			Великая Отечественная	.
2085292	мама, папа(частично) , бабушка		никакие	Обсуждаю наши с мамой проблемы				
2083408	Папа,мама,брат, бабушка, дедушка						Великая отечественная	война
2083416	Я,брат,мама,папа	Собой					Всеми	Нет

Рисунок 10 – Примеры записей с множественными пропусками

Именно поэтому перед началом работы с данной базой данных необходимо сделать правильную очистку, чтобы подобные записи не влияли на итоговый результат работы.

Для начала при помощи библиотеки Pandas загрузим данную выборку в датафрейм, каждому населенному пункту сопоставим округ, в котором он

находится, а также выполним преобразование названий колонок для удобства работы так, как это показано на рисунке 11:

Number	Class	Sex	Rus	Math	Family	Pride	Traditions	Activities	Region	Sights
1964895	6	Девочки	3	3	бабушка мама папа брат	мама потому что оба мне заботится	помогать друг другу	собираемся в месте и гуляем по вечерам	Северо-Западный федеральный округ	фонтаны эрмитаж петропавловская крепость
1964302	6	Девочки	5	5	я, папа, мама, брат, брат.	мамой, потому что она очень умная, а папой, по...	на день рождения любого члена семьи всегда веш...	гуляем, ходим по магазинам, играем в настольны...	Северо-Западный федеральный округ	Спас на Крови, Исаакиевский собор, Медный всад...
1965852	6	Мальчики	4	4	Мама,Папа,собака,я	Я горжусь всеми членами моей семьи	покупать что нибудь вкусное в пятницу	разговариваем, гуляем ,играем	Северо-Западный федеральный округ	кунск камера ,зимний дворец,петрапавловская кр...
1953331	6	Мальчики	3	3	Папа,мама, две сестры, бабушка,я	Папа он очень много зарабатывает и дает мне де...	В моей семье традиций нет.	Кушаю,общаюсь,играю	Северо-Западный федеральный округ	Разбитое кольцо,статуя Петра первого
1964793	6	Девочки	5	4	мама, папа, бабушка, дедушка, брат.	Я горжусь своей мамой. Она очень часто может у...	Отмечать праздники вместе, ездить летом на дачу.	Общаемся, живем вместе, смотрим фильмы, гуляем...	Северо-Западный федеральный округ	Исаакиевский собор, Казанский собор, Цветок жи...
...	...	...	...	...	...	...	...	...	...	...
2076893	8	Мальчики	4	4	Из близких родственников: Мама, Папа, Брат, Дв...	Братом, потому что он закончил гимназию на зол...	Праздновать все традиционные праздники	Общаемся по душам.	Сибирский федеральный округ	Много (Одни из примеров):Памятники колбасе, ру...

Рисунок 11 – Пример внесенных изменений удобства работы

Далее выполним очистку данных. Ее идея состоит в следующем:

– для начала мы временно удалим знаки пунктуации во всех полях датафрейма, и если после этого останется пустое поле, ставим туда специальный символ, обозначающий пустоту - NaN;

– после чего мы удаляем те записи, где больше двух полей со значением NaN;

– затем мы удаляем те записи, где хотя бы в одной колонке нет русских букв (это означает, что данный ученик подошел к анкетированию безответственно и писал информацию, не соответствующую тематике вопроса);

– следующим шагом мы удалим записи с пустыми полями, которые заполнялись работниками НИКО – поля с полом, регионом, классом, уникальным номером и оценками по русскому языку и математике.

При помощи подобной очистки данных мы убрали около 6 тысяч так называемых шумовых данных, оставив только те, которые будут полезны в нашем исследовании.

Однако на этом базовые преобразования не заканчиваются. В дальнейшем вместе с научным руководителем было принято два немаловажных решения, а именно:

– из общей выборки были удалены колонки, связанные с ответами учащихся про улучшение жизни в населенном пункте и про гордость семьи, так как они по сравнению с другими вопросами характеризуют учащегося как личность в меньшей степени и не являются основополагающими для дальнейших исследований;

– Северо-Кавказский федеральный округ был объединен с Южным федеральным округом по географическому принципу в силу своей маленькой выборки, как можно увидеть на рисунке 12:

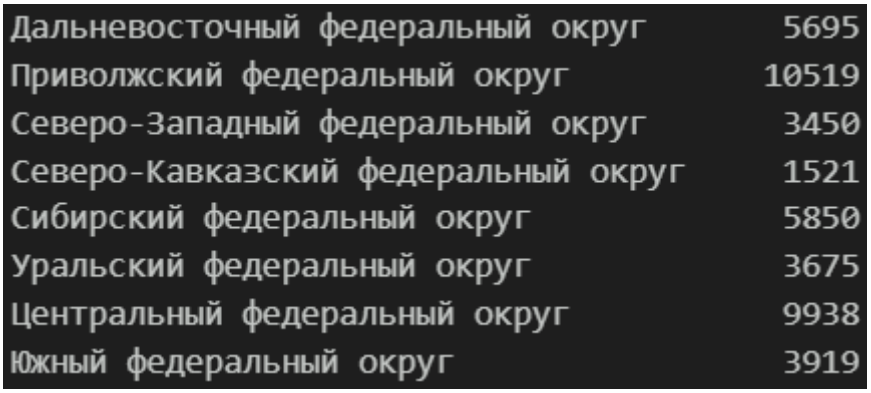

Рисунок 12 – Количество записей в базе данных по каждому из округов

2.3 Разбор свободных ответов

Разбор свободных ответов начался с разбора поля состава семьи учеников при помощи регулярных выражений. На рисунке 13 представлены изначальные ответы учащихся на соответствующий вопрос:

Family
мать отец кот рыбки
В мою семью входит 2 брата и родители
мама,папа,два брата,я
Мама, папа, брат, я.
семья
В мою семью входит Мама,папа и младшая сестра.
все
В мою семью входят мои родители и моя кошка.
ПАПА МАМА БАБУШКА БРАТЬЯ СЕСТРЫ ДЕДУШКА
родители,3 брата,4 сестры и я.
МАМА ПАПА СИСТРА БРАТ
Я и бабушка
я не смогу перечислить всех членов семьи
родители
я,брат,брат,брат,папа,мама
мои родители, братья,сестра,бабушка и дедушка
мама отец брат старший и я
мать папа сестра
Мама, сестрёнка.
В мою семью входят: мои родители, бабушка с дедушкой и мой младший брат.
Родители, братья, бабушки и дедушки, тёти, дяди.
Я, мои братья Артём, Саша и Фёдор, а также сестра Виолетта и мама с папой.
не скажу
Родители Бабушка дедушка
отчим мама я
папа и я
мачеха,папа,я

Рисунок 13 – Ответы учеников на вопрос про их состав семьи

В процессе работы был придуман следующий алгоритм разбора данного поля:

– первым делом в ответах учащихся удалялась вся пунктуация, а сами ответы приводились в нижний регистр;

– далее слово, обозначающее количество, было заменено числом – например, слово «два» было заменено цифрой 2;

– затем была применена следующая эвристика: слово, идущее после цифры n в ответе учащегося, повторялось n раз, затирая саму цифру;

– после чего для получения итогового результата применялись регулярные выражения по следующему принципу: если учащийся не привел никакой релевантной информации, то в колонку типа семьи ставилось «Нет четкого ответа», если учащийся привел информацию о биологических родителях и о ком-то из старших родственников, то в колонку типа семьи ставилось «Расширенная», если учащийся привел информацию о биологических родителях и минимум еще о двух детях в семье, то в колонку типа семьи ставилось «Многодетная», если учащийся привел информацию только о биологических родителях, то в колонку типа семьи ставилось

«Полная», в отличном от вышеописанных случаев в колонку типа семьи ставилось «Неполная».

Более наглядную работу алгоритма мы можем увидеть на рисунке 14:

	Family		Family
2	мать отец кот рыбки	2	Полная
3	В мою семью входит 2 брата и родители	3	Многодетная
4	мама,папа,два брата,я	4	Многодетная
5	Мама, папа, брат, я.	5	Полная
6	семья	6	Нет четкого ответа
7	В мою семью входит Мама,папа и младшая сестра.	7	Полная
8	все	8	Нет четкого ответа
9	В мою семью входят мои родители и моя кошка.	9	Полная
10	ПАПА МАМА БАБУШКА БРАТЬЯ СЕСТРЫ ДЕДУШКА	10	Расширенная
11	родители,3 брата,4 сестры и я.	11	Многодетная
12	МАМА ПАПА СИСТРА БРАТ	12	Многодетная
13	Я и бабушка	13	Неполная
14	я не смогу перечислить всех членов семьи	14	Нет четкого ответа
15	родители	15	Полная
16	я,брат,брат,брат,папа,мама	16	Многодетная
17	мои родители, братья,сестра,бабушка и дедушка	17	Расширенная
18	мама отец брат старший и я	18	Полная
19	мать папа сестра	19	Полная
20	Мама, сестрёнка.	20	Неполная
21	В мою семью входят: мои родители, бабушка с дедушкой и мой младший брат.	21	Расширенная
22	Родители, братья, бабушки и дедушки, тёти, дяди.	22	Расширенная
23	Я, мои братья Артём, Саша и Фёдор, а также сестра Виолетта и мама с папой.	23	Многодетная
24	не скажу	24	Нет четкого ответа
25	Родители Бабушка дедушка	25	Расширенная
26	отчим мама я	26	Неполная
27	папа и я	27	Неполная
28	мачеха,папа,я	28	Неполная

Рисунок 14 – Итоговый вид поля состава семьи учащихся

Остальные поля были разобраны при помощи алгоритма машинного обучения для классификации RandomForest, или же случайный лес [5]. Идея этого алгоритма заключается в создании большого количества решающих деревьев в процессе обучения и использовании их для получения более точных и стабильных прогнозов. Из обучающего набора данных создается несколько случайных выборок с возвращением. Это означает, что в каждой выборке могут содержаться повторяющиеся записи, а некоторые записи могут быть пропущены, после чего для каждой выборки с заменой строится решающее дерево. При построении каждого дерева на каждом узле случайным образом выбирается подмножество признаков. Это делается для уменьшения корреляции между деревьями. Финальным этапом для заданного объекта выполняется классификация путем проведения «голосования» результатов

всех деревьев. На рисунке 15 изображен общий пример работы данного алгоритма:

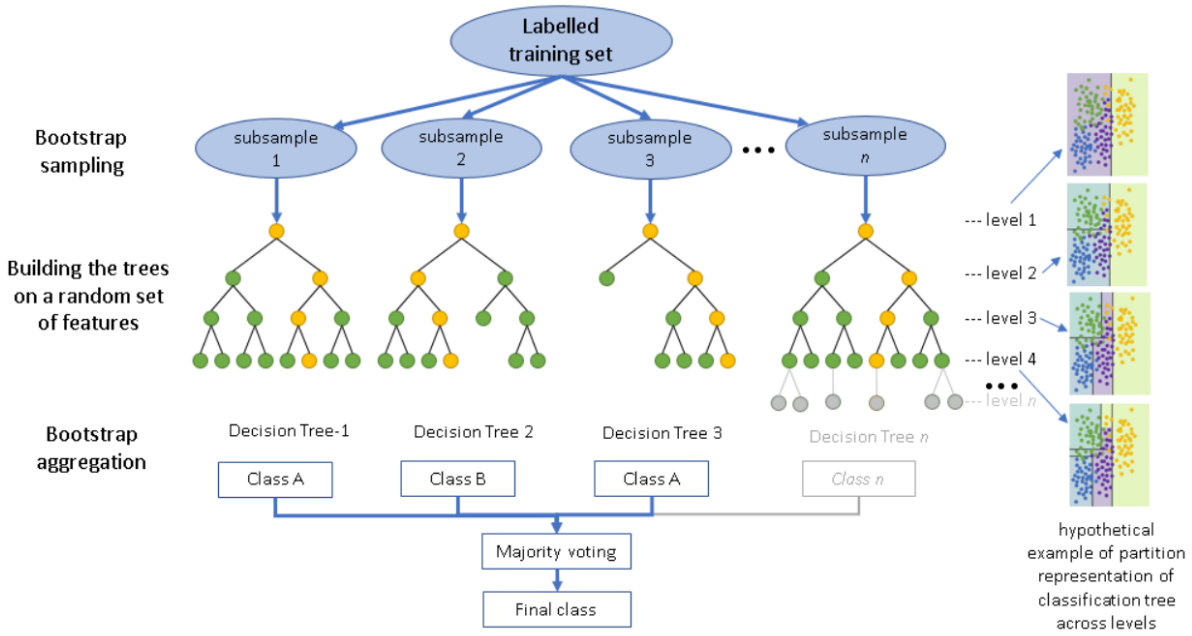

Рисунок 15 – Принцип работы случайного леса

Данный алгоритм стоит особняком среди остальных методов классического машинного обучения для задач классификации из-за того, что обладает множеством преимуществ, таких как устойчивость к переобучению, хорошая обобщающая способность, а также богатый функционал для обработки большого количества признаков. Помимо него, другой алгоритм классификации в данной работе не применялся, так как полученные случайным лесом метрики получились очень хорошие и полностью устраивали.

Перед применением данного алгоритма важно было сделать 2 шага: разделить ответы учащихся на 3 группы и предварительно качественно их обработать. Для обработки ответов была применена токенизация, удаление ненужных и стоп-слов, лемматизация и другие методы работы с текстом. Помогли в этом две python-библиотеки, которые хорошо работают с русским языком и вычленяют из текста непосредственно его содержимое: ntlk и pymorphy2. Библиотека ntlk хороша в части синтаксического анализа фразы, а

pymorphy2 в части морфологического. Вместе они являются довольно сильным инструментом для обработки текста перед применением к нему модели машинного обучения.

Так как изначально ответы учащихся представлялись в сыром виде и не были отмечены, для каждого вопроса были вручную размечены 6 тысяч ответов учащихся на 3 категории по следующему принципу: 1 – учащийся не привел релевантной информации по вопросу, 2 – учащийся привел минимум релевантной информации (например, указал только один памятник культуры или известного деятеля), 3 – учащийся привел много релевантной информации (например, указал 2 и более достопримечательности своего населенного пункта). На 80% этих размеченных данных модель обучалась, а на 20% (что составляет 1200 строк для каждого вопроса) она предсказывала итоговую метку. Результаты превзошли все ожидания – на рисунке 16 мы можем наблюдать посчитанные метрики для процесса обучения модели на ответах учащихся про традиции в их семье:

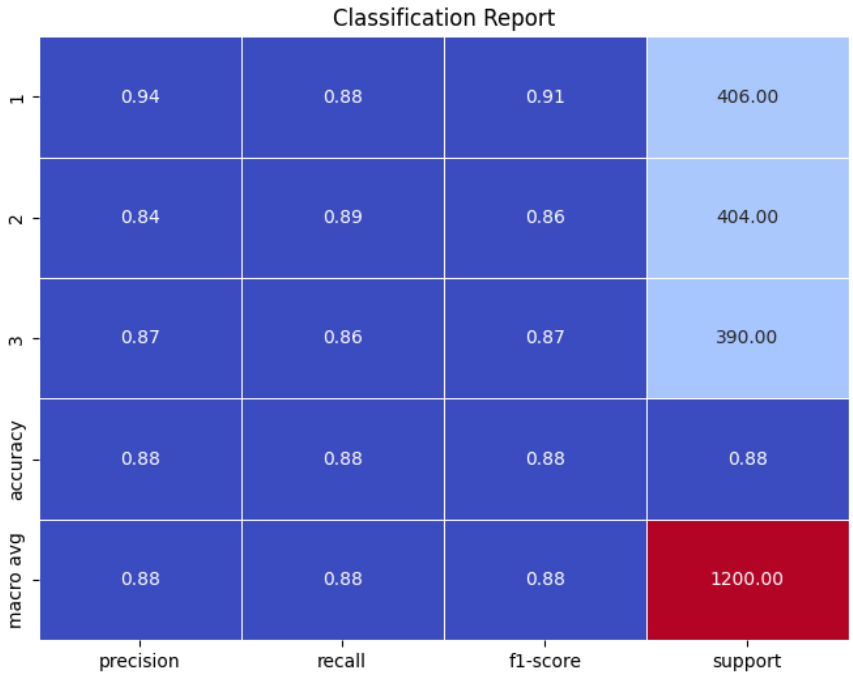

Рисунок 16 – Результат работы модели случайного леса на наших данных

На представленном рисунке цифры 1, 2 и 3 означают метку для ответа учащегося, число 1200 в правом нижнем углу означает количество ответов, на которых модель работала. Точность 88% указывает на то, что созданная модель классификации может быть успешно применена на других, неотмеченных данных. На рисунке 17 представлен один из результатов работы нашей модели:

Traditions	TraditionsAnswers
помогать друг другу	1
на день рождения любого члена семьи всегда вешаем одно поздравление	2
покупать что нибудь вкусное в пятницу	2
В моей семье традиций нет.	1
Отмечать праздники вместе, ездить летом на дачу.	3
на каждый новый год я пою	2
Каждую неделю мы смотрим фильмы в домашнем кинотеатре. У нас он свой.	2
рождество	2
каждый год ездим на пляж	2
На новый год мы все вместе собираемся у кого-то дома. И нам не мешает , что мы живем в разных странах.	2
ми лубим сабиратса на празники	1
Мы вместе празднуем день рождения моей младшей сестры	2
В новый год ждать пока не будет ровно 12 посла празнования смотрим фильм	2
их нет	1
их нет	1
нету	1

Рисунок 17 – Результат классификации моделью машинного обучения ответов учащихся

2.4 Формирование индексов

Итак, на прошлом этапе мы успешно разобрали анкетные ответы и получили следующую информацию:

– состав семей учащихся;

– переменную, обозначающую «богатство» традиций в семьях учащихся;

– переменную, обозначающую «богатство» активностей в семьях учащихся;

– переменную, обозначающую осведомленность учащегося в культурном аспекте своего населенного пункта;

– переменную, обозначающую интеллектуальную подкованность учащегося в сфере искусства;

– переменную, обозначающую историческую оснащенность учащегося.

На основании этих переменных было принято решение создать три так называемых «индекса» учащегося: индекс семейной жизни, индекс культуры и индекс знания истории.

Эти индексы, по сути своей, являются категориальными переменными, подчеркивающими всесторонний контекст учащегося. Они говорят нам о том, насколько развита та или иная область его жизни.

Индекс семейной жизни учащегося формировался как среднее значение (округленное в большую сторону) между переменной, обозначающей «богатство» традиций в его семье, и переменной, обозначающей «богатство» активностей в его семье – например, если у какого-либо учащегося первая переменная равна единице (не приведена ни одна традиция), а вторая переменная равна двойке (указана одна семейная активность), то его индекс семейной жизни будет равен двойке, что в конечном итоге означает довольно крепкий, среднестатистический семейный контекст.

Индекс культуры учащегося формировался как среднее значение (округленное в большую сторону) между переменной, обозначающей его осведомленность в культурном аспекте своего населенного пункта, и переменной, обозначающей его интеллектуальную подкованность в сфере искусства – например, если у какого-либо учащегося первая переменная равна двойке (указана одна достопримечательность населенного пункта), а вторая переменная равна тройке (указано два и более деятелей искусства), то его индекс культуры будет равен тройке, что в конечном итоге означает отличный культурный контекст.

Индекс знания истории учащегося брался как переменная, обозначающая историческую оснащенность учащегося.

Таким образом из необработанных данных анкетирования НИКО мы получили выборку, которую можно кластеризовать, на которой можно считать разные метрики и собирать важную статистику, о которой будет сказано в пункте 2.5 дипломной работы.

На рисунке 18 представлен итоговый результат разбора свободных ответов:

Number	Class	Sex	Rus	Math	Family	Region	History Knowledge Index	Family Life Index	Culture Index
1964895	6	Девочки	3	3	Расширен	Северо-За	2	1	3
1964302	6	Девочки	5	5	Многодет	Северо-За	2	2	3
1965852	6	Мальчики	4	4	Полная	Северо-За	2	3	3
1953331	6	Мальчики	3	3	Расширен	Северо-За	2	2	2
1964793	6	Девочки	5	4	Расширен	Северо-За	2	3	3
1963425	6	Девочки	4	4	Неполная	Северо-За	2	3	2
1963939	6	Мальчики	4	4	Неполная	Северо-За	2	2	3
1964625	6	Мальчики	4	4	Расширен	Северо-За	2	2	2
1964230	6	Мальчики	4	4	Расширен	Северо-За	2	2	2
1963753	6	Девочки	4	4	Полная	Северо-За	2	3	2
1962660	6	Девочки	5	5	Расширен	Северо-За	2	2	2
1951227	6	Мальчики	4	5	Расширен	Северо-За	2	2	2
1964158	6	Мальчики	4	4	Расширен	Северо-За	2	2	2
1964224	6	Мальчики	3	3	Расширен	Северо-За	1	2	1
1964476	6	Девочки	5	4	Расширен	Северо-За	2	2	3
1965066	6	Мальчики	3	3	Полная	Северо-За	2	1	2
1963800	6	Девочки	4	4	Многодет	Северо-За	2	2	2
1963973	6	Девочки	5	5	Полная	Северо-За	2	3	2
1964193	6	Мальчики	3	4	Многодет	Северо-За	2	3	2
1964588	6	Мальчики	4	4	Расширен	Северо-За	2	1	2
1964265	6	Девочки	3	4	Неполная	Северо-За	2	2	2
1964812	6	Девочки	4	4	Неполная	Северо-За	2	1	2
1965005	6	Девочки	4	3	Полная	Северо-За	1	2	2
1964478	6	Девочки	4	5	Полная	Северо-За	2	3	3
1964973	6	Девочки	4	3	Многодет	Северо-За	2	2	2
1964983	6	Мальчики	3	4	Расширен	Северо-За	2	2	2
1965954	6	Девочки	5	4	Расширен	Северо-За	2	1	2
1966120	6	Девочки	5	5	Неполная	Северо-За	2	1	2

Рисунок 18 – Результат разбора анкетных ответов учащихся

Разобранные ответы учащихся (индексные переменные, регион и состав семьи) были переданы второму участнику проекта для дальнейшего исследования тенденций внутри общеобразовательных организаций.

2.5 Анализ полученной статистики

В процессе анализа полученной статистики было выяснено, что поле типа семьи должно быть видоизменено. Связано это с тем, что записей, где в колонке типа семьи написано «Нет четкого ответа», значительно меньше

относительно записей с другими типами семьи. Мы можем видеть данную тенденцию на рисунке 19:

Многодетная	6131
Неполная	6933
Нет четкого ответа	470
Полная	13729
Расширенная	12499

Рисунок 19 – Количество записей для каждого типа семьи

Также при подсчете средних оценок по русскому языку и по математике для каждого типа семей по округам было замечено, что значения для полных и расширенных во многих округах едва ли не равны. Примеры такого феномена мы можем наблюдать на рисунках 20, 21 и 22:

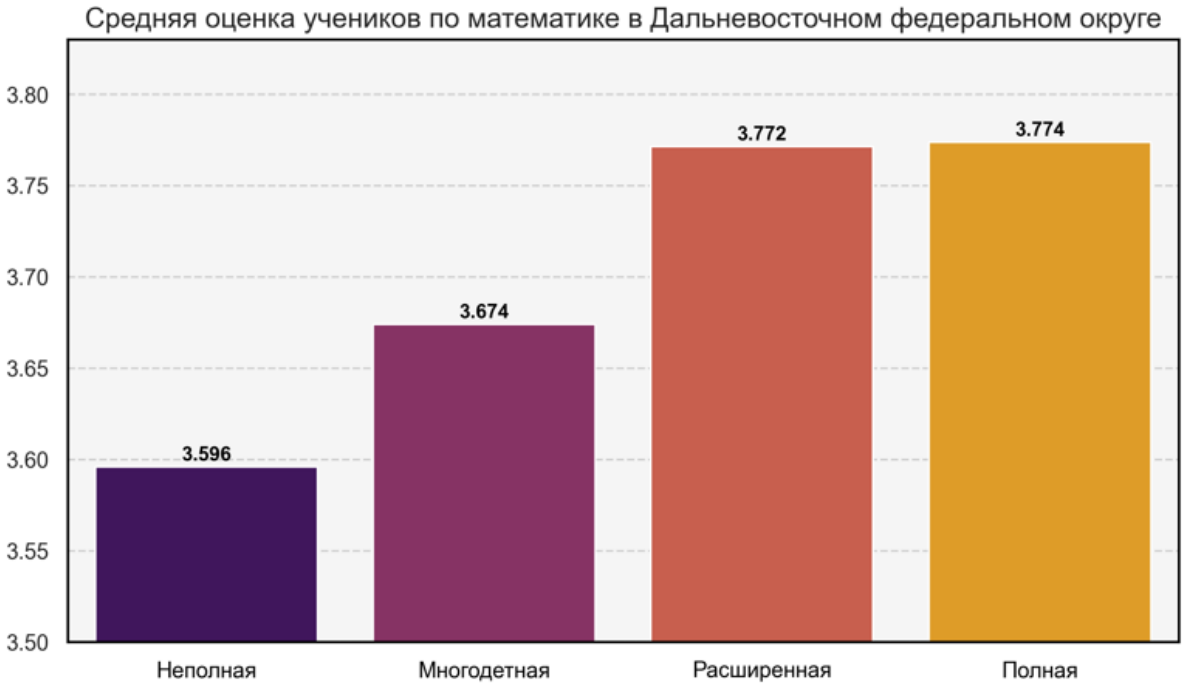

Рисунок 20 – Средняя оценка учеников по математике на Дальнем Востоке

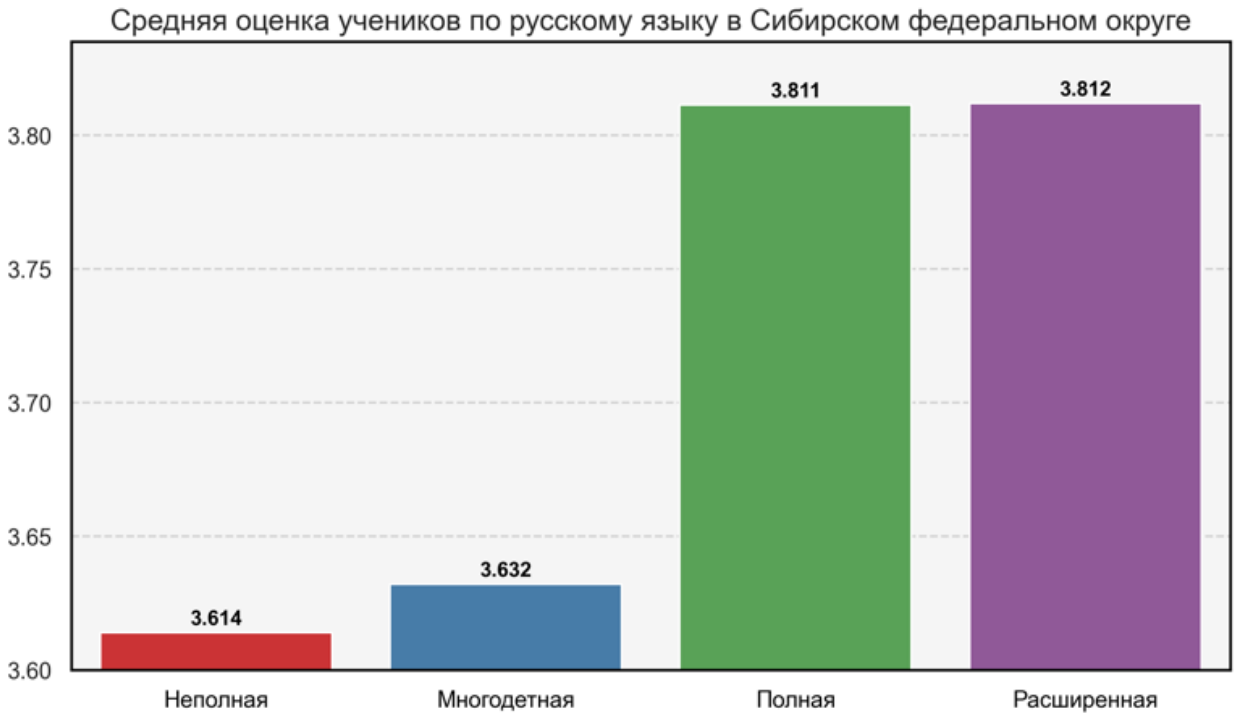

Рисунок 21 – Средняя оценка учеников по русскому языку в Сибири

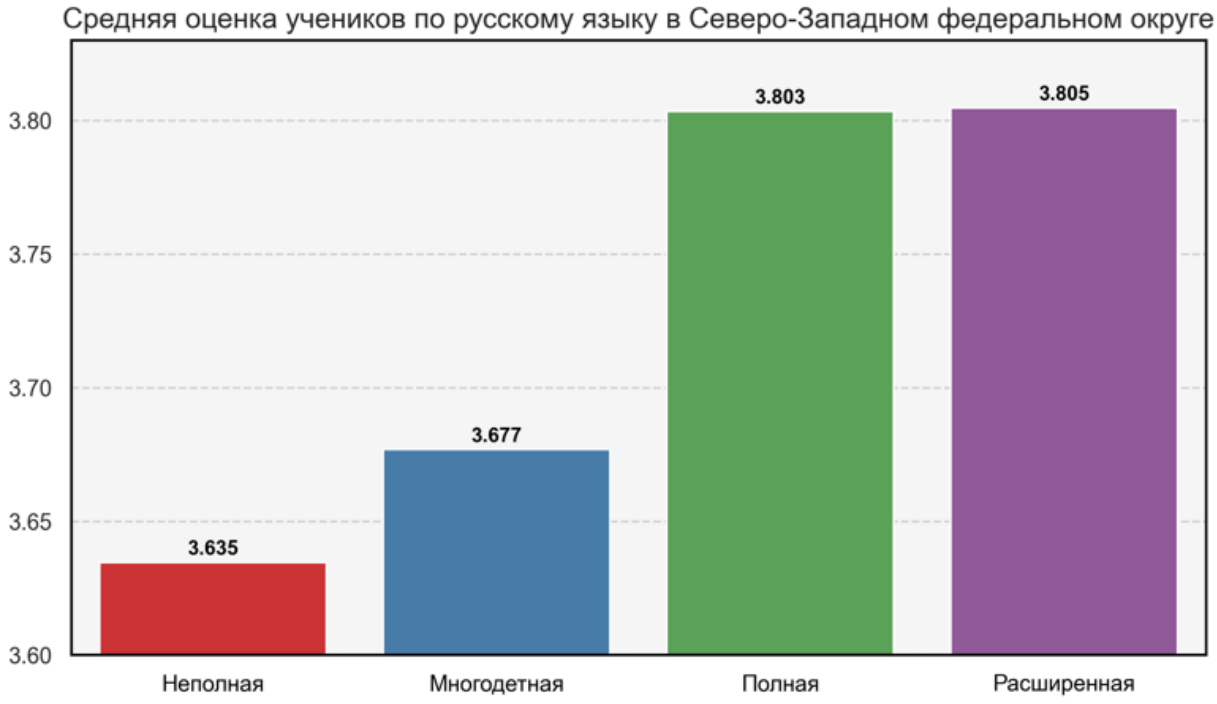

Рисунок 22 – Средняя оценка учеников по русскому языку на Северо-Западе

В связи с чем было принято решение остановиться на трех типах семьи: неполной, многодетной-полной (ранее многодетная), полной-немногодетной (объединение полной и расширенной).

Итоговое распределение средних оценок по русскому языку внутри округов для каждого типа семьи изображено на рисунке 23:

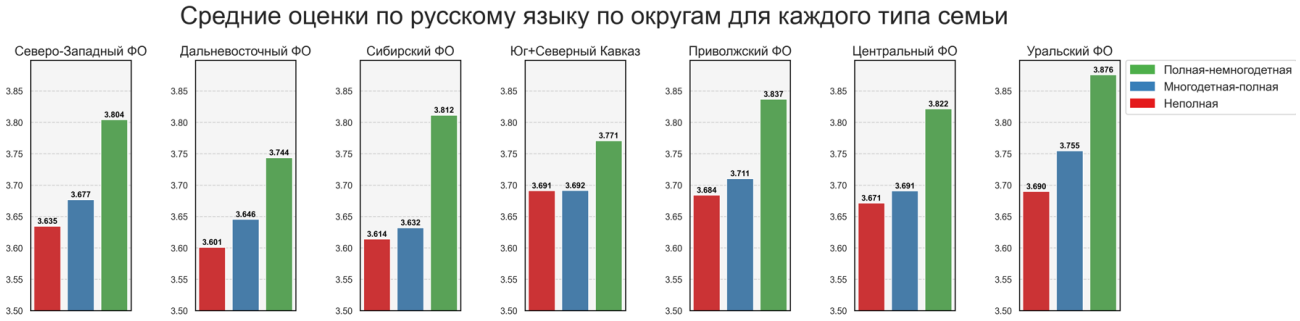

Рисунок 23 – Средние оценки по русскому языку внутри округов для каждого типа семьи

Несмотря на то, что цифры на графике видны недостаточно хорошо, тенденция сразу бросается в глаза: на каждом из графиков у нас слева-направо идут неполная, многодетная-полная и полная-немногодетная семьи.

Итоговое распределение средних оценок по математике внутри округов для каждого типа семьи изображено на рисунке 24:

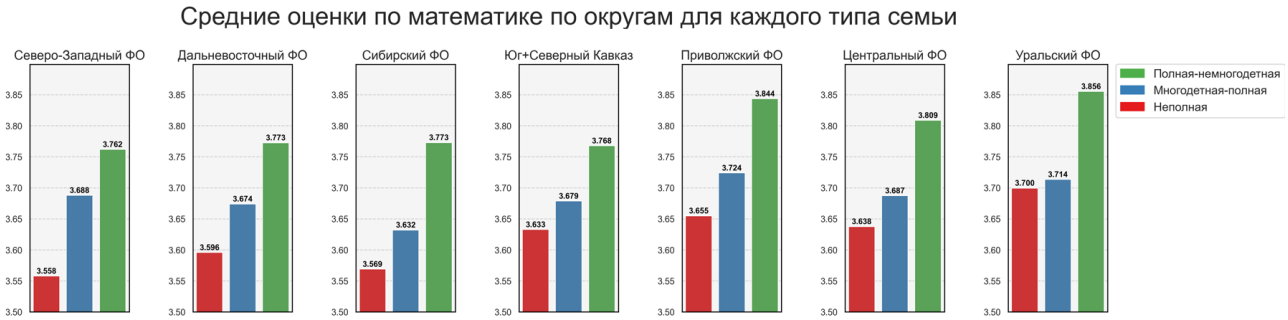

Рисунок 24 – Средние оценки по математике внутри округов для каждого типа семьи

Здесь мы можем заметить, что тенденция точно такая же: слева-направо по оценкам идут неполная, многодетная-полная и полная-немногодетная семьи.

Как мы помним, помимо школьных оценок, в нашем распоряжении есть также ПБ для каждого учащегося – значение от 0 до 20, которую он получил в рамках проведенной НИКО диагностической работы.

На рисунке 25 представлены средние значения первичного бала (НИКО 2022) по округам для каждого типа семьи:

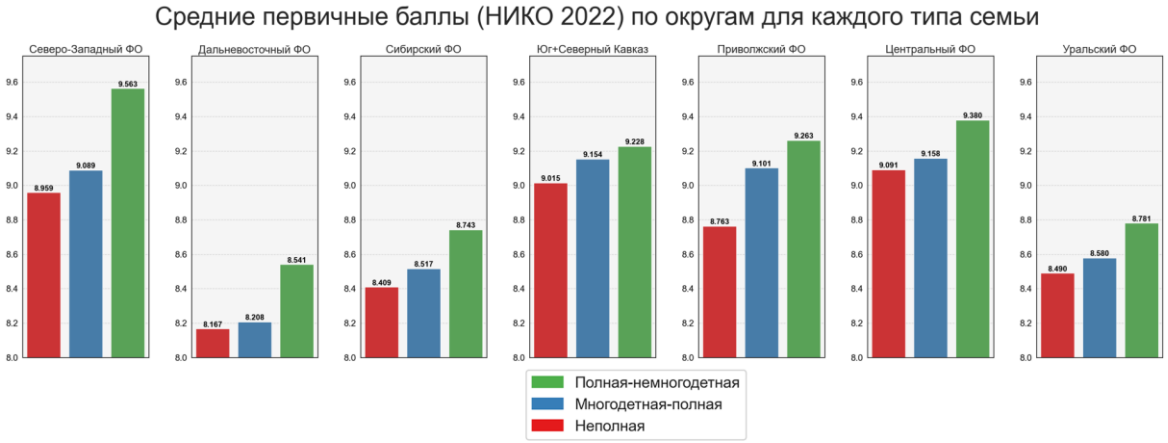

Рисунок 25 – Средние первичные баллы (НИКО 2022) по округам для каждого типа семьи

Видим ту же самую тенденцию, что и в случае со школьными оценками. Помимо распределения переменных, отвечающих за успеваемость учащихся, интересно также внутри все тех же округов посмотреть на средние значения наших полученных индексов.

На рисунках 26, 27 и 28 в разрезе округов представлены средние значения индекса знания истории, индекса культуры и индекса семейной жизни учащихся соответственно:

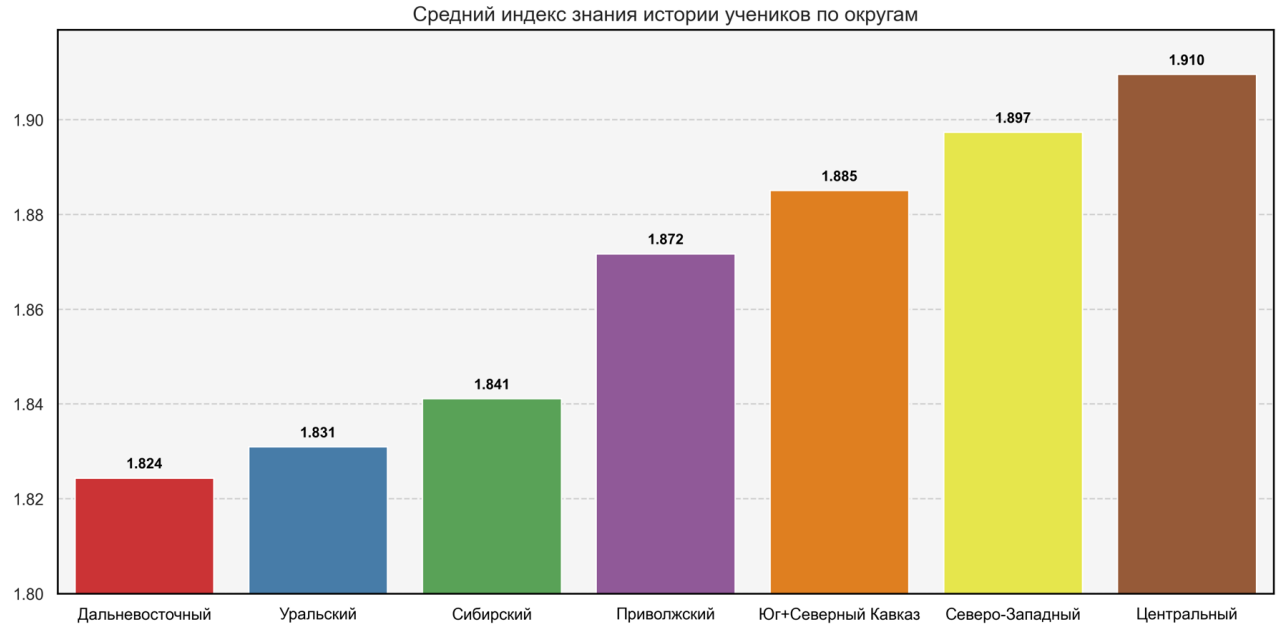

Рисунок 26 – Средний индекс знания истории учеников по округам

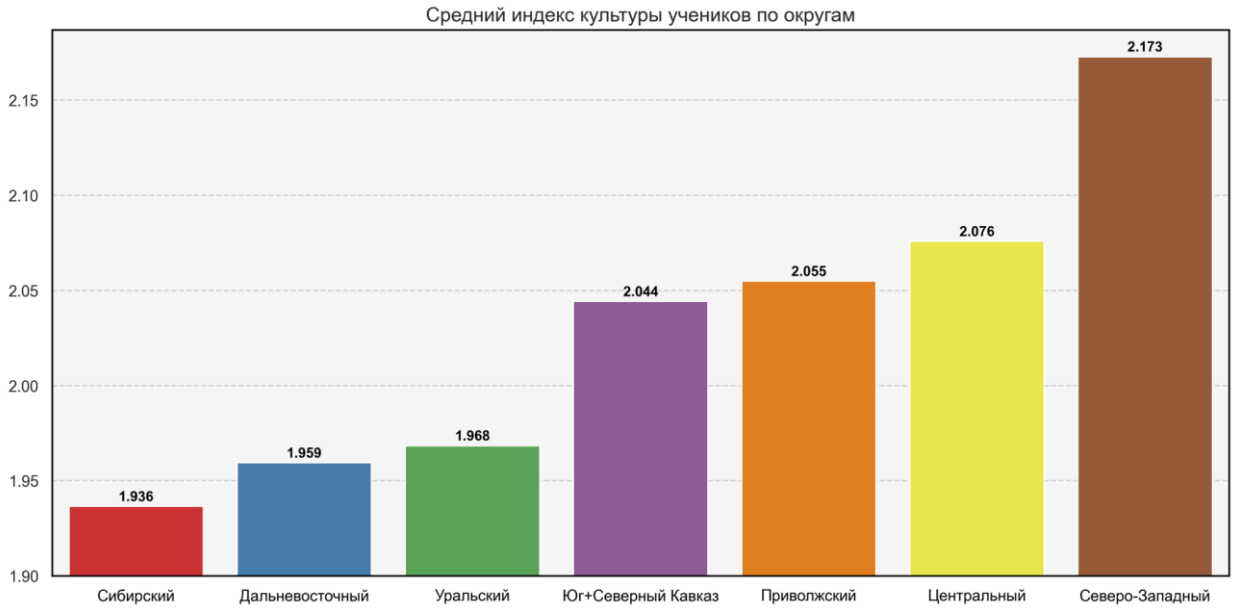

Рисунок 27 – Средний индекс культуры учеников по округам

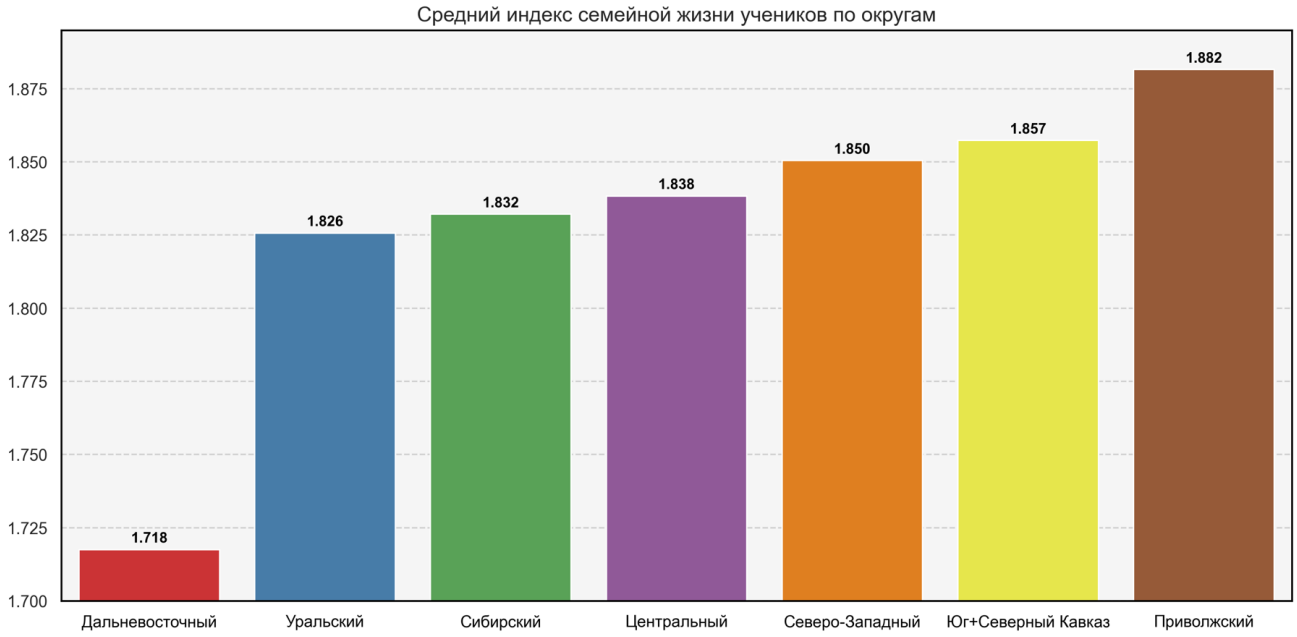

Рисунок 28 – Средний индекс семейной жизни учеников по округам

Говоря о среднем значении исторического индекса, нельзя не упомянуть о лидерстве на изображенной диаграмме Центрального и Северо-Западного федеральных округов. Действительно, ни для кого не секрет, что именно в этих регионах чтят историческое наследие нашей страны – в них очень много учащихся, ежегодно побеждающих в олимпиадах по истории на разных уровнях.

Что же касается культурной составляющей, то тут, конечно же, безоговорочным фаворитом среди всех округов является Северо-Западный, ведь недаром говорят, что Санкт-Петербург – это культурная столица нашей страны. Как и в случае с историческим индексом, Центральный федеральный округ здесь также занимает одну из ведущих позиций.

Статистика, изображенная на рисунке 28, также вряд ли будет для кого-то удивлением, так как Приволжский федеральный округ славится своим чтением семейных традиций – в частности, можно выделить такие республики, как Марий Эл, Татарстан и Удмуртию. Среди Южного и Северо-Кавказского федеральных округов, которые в приведенной выше статистике идут сразу

после Приволжья, нужно отметить республику Калмыкию и Дагестанскую республику.

Полученное распределение играет очень важную роль в понимании образовательных и личностных тенденций учащихся внутри каждой административно-территориальной единицы РФ, а также побуждает к множеству всесторонних размышлений и рассуждений.

На основании собранной статистики четко прослеживаются преимущества и недостатки в разрезе каждого округа, и при правильном использовании полученная информация может быть крепким фундаментом для дальнейших улучшений и нововведений в вопросе подхода к образованию и воспитанию учащихся во всех отдельно взятых регионах.

3 КЛАСТЕРИЗАЦИЯ

3.1 Технологии для кластеризации

Как уже говорилось выше, задача кластеризации является одной из основополагающих задач в машинном обучении.

Концептуально, кластеризация представляет собой процесс выявления внутренних структур в данных без заранее заданных меток классов. Она позволяет автоматически обнаруживать скрытые паттерны и группировать данные на основе их сходства, что может быть особенно полезно в случаях, когда отсутствует явное разделение на классы или когда необходимо предварительно исследовать данные перед применением более сложных моделей.

Существует множество алгоритмов для задач кластеризации, однако в большинстве своем выделяют следующие алгоритмы:

– алгоритм K-means, суть которого заключается в разбиении данных на заранее заданное количество кластеров, минимизируя сумму квадратов расстояний от каждой точки до центра своего кластера – данный алгоритм, пожалуй, является самым известным;

– алгоритм иерархической кластеризации, суть которого заключается в построении иерархии кластеров, объединяя или разделяя их на основе определенных критериев, таких как расстояние между кластерами;

– алгоритм DBSCAN, который основывается на плотности данных и идентифицирует кластеры как области с высокой плотностью точек, разделенные областями с низкой плотностью;

– алгоритм Mean Shift, ищущий локальные плотные области в пространстве данных и перемещает центры кластеров в направлении увеличения плотности;

– другие, менее популярные алгоритмы кластеризации, использующиеся для узкого круга задач.

3.2 Кластеризация при помощи полученных индексов

На этапе 2.4 мы успешно получили индексы учащихся – результат преобразования их свободных ответов в категориальные переменные. Теперь необходимо приступать к непосредственной кластеризации учащихся для получения особенно выделяющихся подгрупп среди основной выборки.

На рисунке 29 представлены полученные индексные группы, а также количество учащихся в каждой из этих групп:

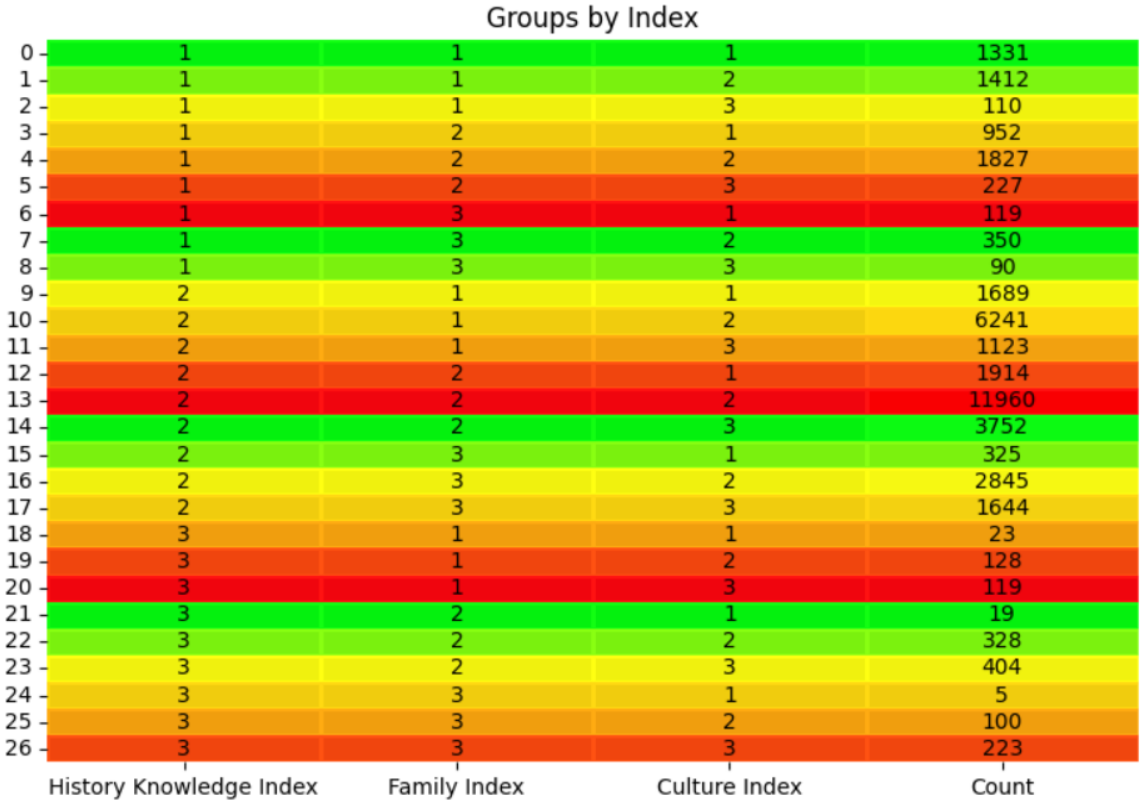

Рисунок 29 – Полученные индексные группы учащихся

Изначальная идея состояла в том, чтобы провести кластеризацию по четырем полям: по трем индексам и типу семьи учащегося, однако в реальности же выходило так, что группы формировались, как правило, на основании полученных индексных групп: например, получалось так, что учащиеся, которые по каждому из критериев имеют значение 2, составляли полностью отдельный кластер, а в другой тем временем мог попасть совершенно разный контингент учащихся. В связи с чем изначальным этапом

кластеризации было принято сделать более расширенные индексные группы, а слишком маленькие группы удалить из общей выборки, так как их можно трактовать как выбросы. Например, группа «3-3-1» (сокращение от History Knowledge Index: 3, Family Life Index: 3, Culture Index: 1) представляют только 5 человек, поэтому в дальнейшем она не рассматривается, так как требует более детального исследования и выбивается из общей тенденции.

Данное решение обосновано тем, что во многих группах учеников получилось очень мало относительно других групп, а сами по себе эти учащиеся практически не отличаются. Поэтому некоторые из маленьких групп были объединены с ближайшими по контексту группами. Для более наглядного понимания происходящего можно взглянуть на рисунок 30, на котором на трехмерной диаграмме изображена каждая из 27 изначальных индексных групп в виде круга, причем размер круга соответствует размеру самой группы:

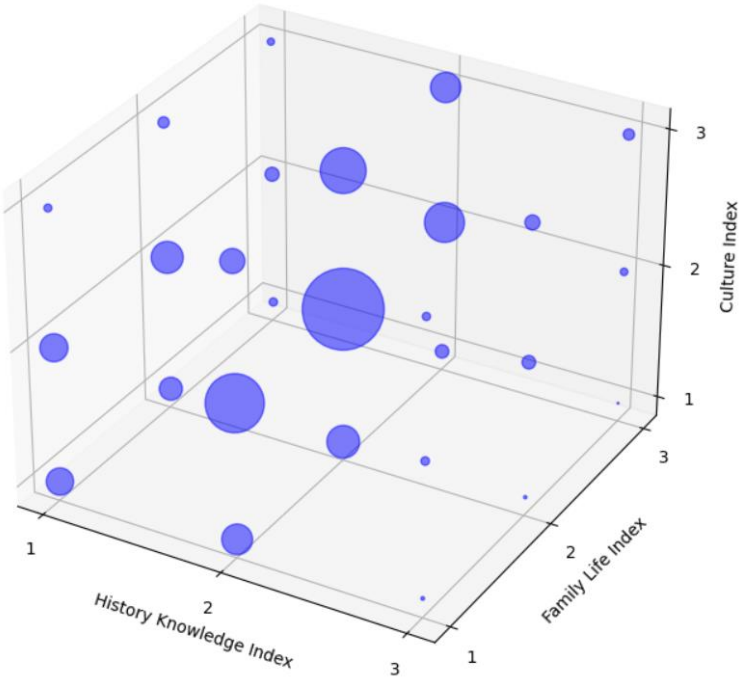

Рисунок 30 – Трехмерное представление изначальных индексных групп

После проделанных манипуляций было выделено 8 больших групп, как на рисунке 31:

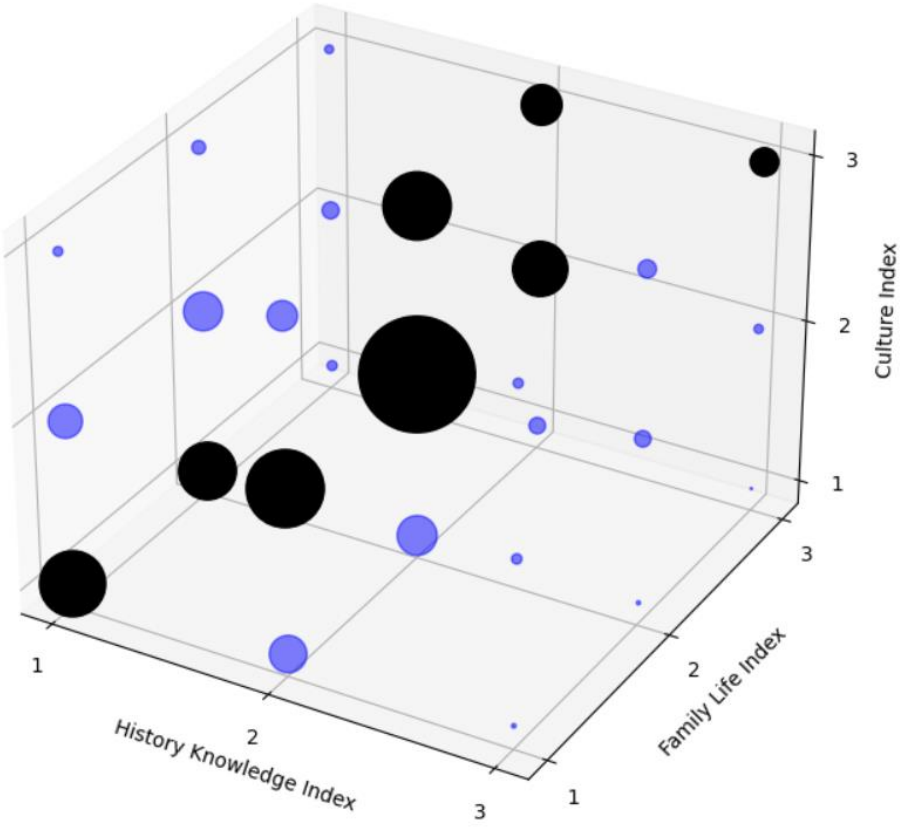

Рисунок 31 – Трехмерное представление объединенных индексных групп

Так, например, группы «2-3-1», в составе которой было 325 человек, перешла в состав группы «2-3-2», в составе которой было 2845 человек, а группа «3-1-2», в составе которой было 128 человек, перешла в группу «2-1-2», в составе которой было 6241 человека.

Полный состав объединенных групп мы можем видеть на рисунке 32:

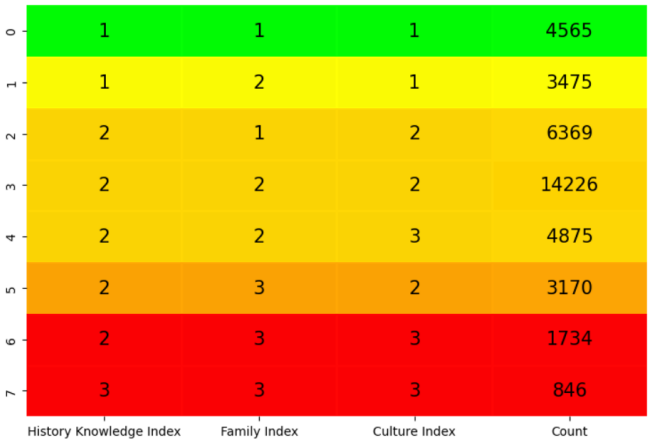

Рисунок 32 – Полученные объединенные индексные группы учащихся

При детальном рассмотрении этих групп было выяснено, что группы 6 и 7 помимо того, что похожи по контексту, так еще и похожи по успеваемости и, как следствие, по своему составу, поэтому данные две группы уже могут быть нашим первым полученным кластером и будут содержать самых сильных учеников в выборке. На рисунке 33 представлены средние оценки и среднее значение первичного балла в группах 6 и 7:

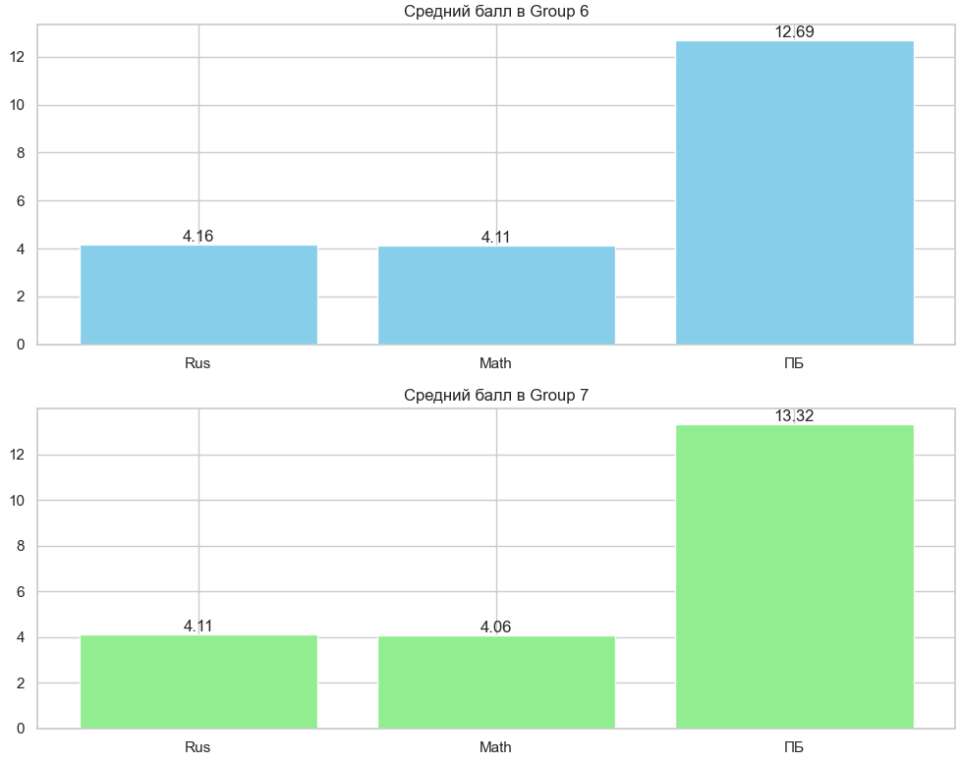

Рисунок 33 – Успеваемость в 6 и 7 индексных группах

В то время как в самых слабых по контексту группах 0 и 1 ученики различаются больше, поэтому каждая из них будет представлять свой отдельный кластер, о чем свидетельствует рисунок 34:

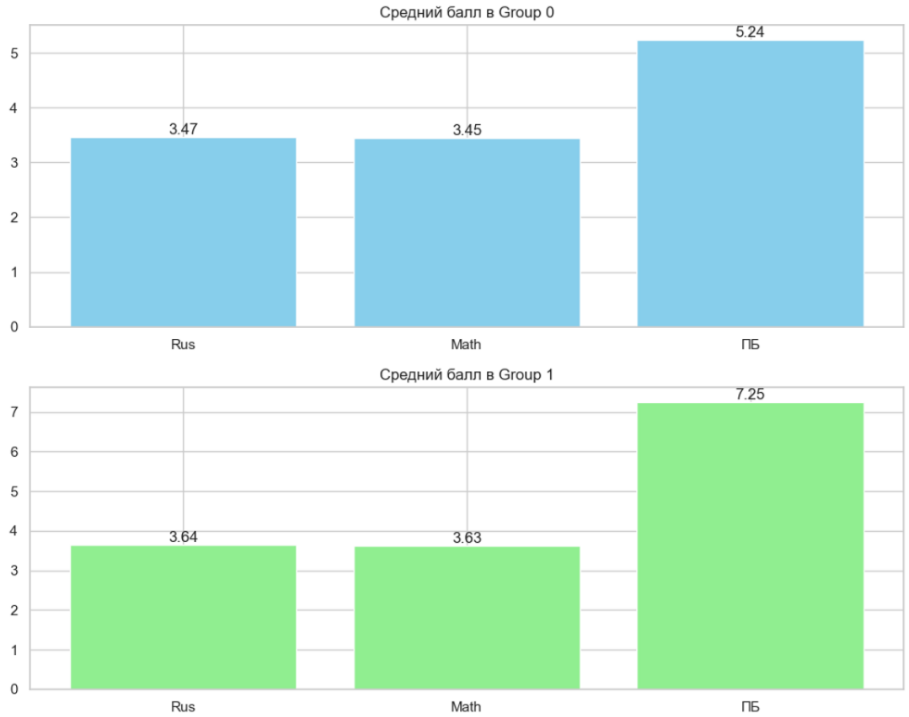

Рисунок 34 – Успеваемость в 0 и 1 индексных группах

Что же касается остальных групп, то в ходе анализа выяснилось, что группа 5 также является уникальной в своем роде, не похожей на группы сильных и средних учеников, и поэтому выделяется в отдельный кластер, а вот группы 2-4, исходя из своего состава, более разрозненны и требуют дополнительного изучения, а именно – кластеризации при помощи машинного обучения, о чем будет сказано в пункте 3.3 работы.

Таким образом, в этой главе мы из сырых разобранных анкетных ответов приблизились к более осмысленной и реальной картине того, с какими именно учащимися имеем дело, а также сделали уверенный шаг к тому, чтобы получить принципиально разные группы учащихся, каждая из которых имеет свою уникальность и заслуживает быть выделенной среди остальных.

3.3 Кластеризация при помощи машинного обучения

В качестве алгоритмов кластеризации были рассмотрены 3 алгоритма: DBSCAN, иерархическая кластеризация и кластеризация K-Means. Однако и иерархическая кластеризация, и алгоритм DBSCAN требуют слишком много

оперативной памяти. Из-за этого иерархическая кластеризация на репрезентативной базе НИКО так и не отработала, завершившись с ошибкой превышения памяти, а результат работы DBSCAN не оправдал ожиданий, так как минимальным количество кластеров выдал 214, что, конечно же, много для нашей задачи. Вероятно, это связано с тем, что данный алгоритм зачастую используется для более сложных задач с более сложными элементами, для которых количество кластеров не имеет значения и может достигать очень большого количества.

Именно поэтому было принято решение остановиться на алгоритме K-Means – самом популярном и простом, но при всем при этом очень эффективном.

Его основная идея состоит в разделении набора данных на заранее определенное количество кластеров (k) так, чтобы объекты внутри одного кластера были максимально похожи между собой, а объекты из разных кластеров были максимально различны. Алгоритм его работы можно описать следующим образом:

– выбираются центры кластеров, называемые центроидами;

– для каждой точки данных вычисляется расстояние до каждого центроида, в результате чего каждая точка присваивается к тому кластеру, чей центроид находится ближе всего;

– после того как все точки были присвоены к кластерам, центр каждого кластера пересчитывается как среднее значение всех точек данных, принадлежащих этому кластеру, в результате чего получается новое положение центроида;

– шаги 2 и 3 повторяются до тех пор, пока центроиды не стабилизируются или пока не будет выполнен критерий останова, которым в большинстве случаев является максимальное количество итераций;

– после завершения работы алгоритма каждая точка данных принадлежит одному из кластеров, а центроиды представляют собой центры этих кластеров.

На рисунке 35 представлен наглядный пример работы алгоритма K-means:

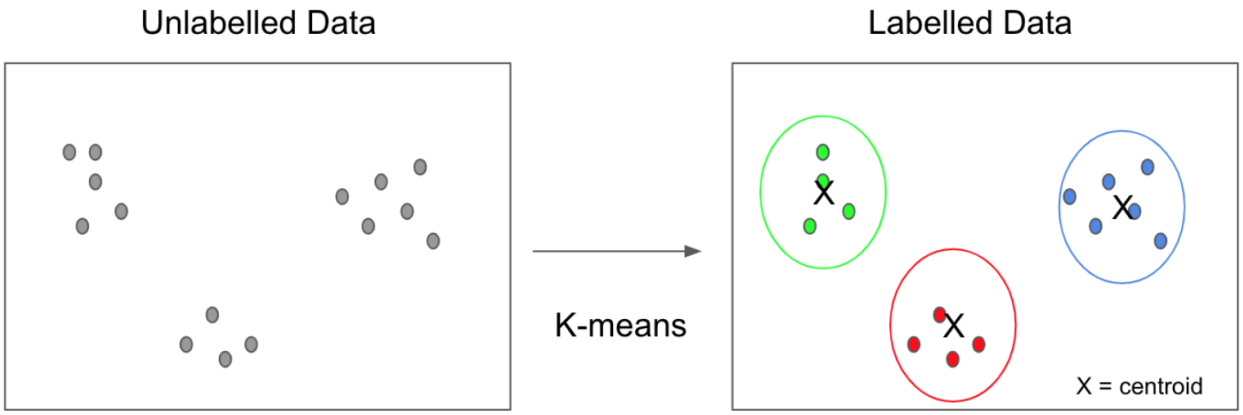

Рисунок 35 – Пример работы алгоритма кластеризации K-means

Так как алгоритм K-Means перед кластеризацией требует количество кластеров, на которые исходная выборка должна разбиться, то было необходимо подобрать это число. Сделать это получилось при помощи распространенного метода локтя – графического метода выбора оптимального числа кластеров в алгоритмах кластеризации, который помогает определить оптимальное количество кластеров на основе изменения внутригрупповой дисперсии или суммы квадратов внутрикластерных расстояний с увеличением числа кластеров.

Основная идея метода заключается в том, чтобы выбрать точку на кривой, после которой добавление дополнительного кластера не приводит к значительному уменьшению внутригрупповой дисперсии.

Мы, по сути, запускаем n раз наш алгоритм кластеризации, где n – это разность крайнего числа кластеров и начального, где на каждой итерации считается наша инерция. Далее строится график внутригрупповой дисперсии в зависимости от количества кластеров, на котором ищется так называемый

«локоть» – точка на этой кривой, где изменение становится менее значительным.

Данная точка на графике указывает на оптимальное количество кластеров. Это тот момент, когда добавление еще одного кластера перестает значительно уменьшать внутригрупповую дисперсию относительно других ее значений. Поэтому количество кластеров, соответствующее этой точке, выбирается как оптимальное число кластеров для набора исходных данных.

Данный процесс был запущен на наших данных: в процессе кластеризации алгоритмом K-means по школьным оценкам, ПБ, составу семьи и полу метод локтя проходился по каждому количеству кластеров от 1 до 7, считал дисперсию и визуализировал ее. Полученный результат изображен на рисунке 36:

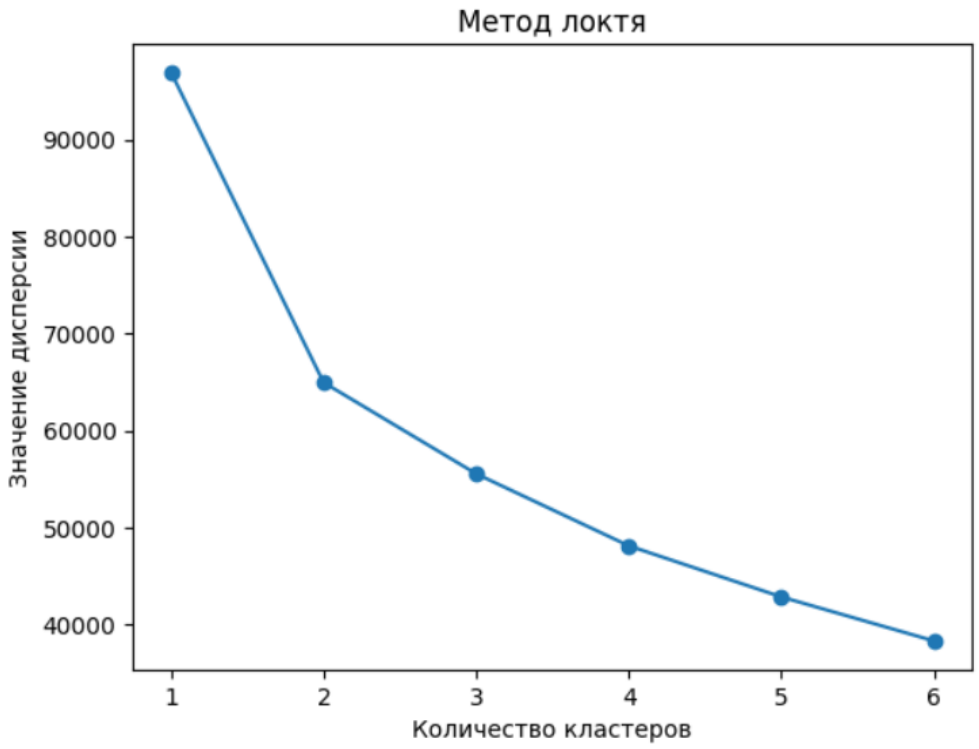

Рисунок 36 – Метод локтя

Так как весь процесс кластеризация наша инерция продолжала уменьшаться, то нужно посмотреть на другие метрики, чтобы четко определить, какое количество кластеров оптимально.

Для этого была посчитана разность между последовательными значениями дисперсии, посчитанная по формуле $inertia_{k+1} - inertia_k$, где k – количество кластеров. Результат данной разности изображен на рисунке 37:

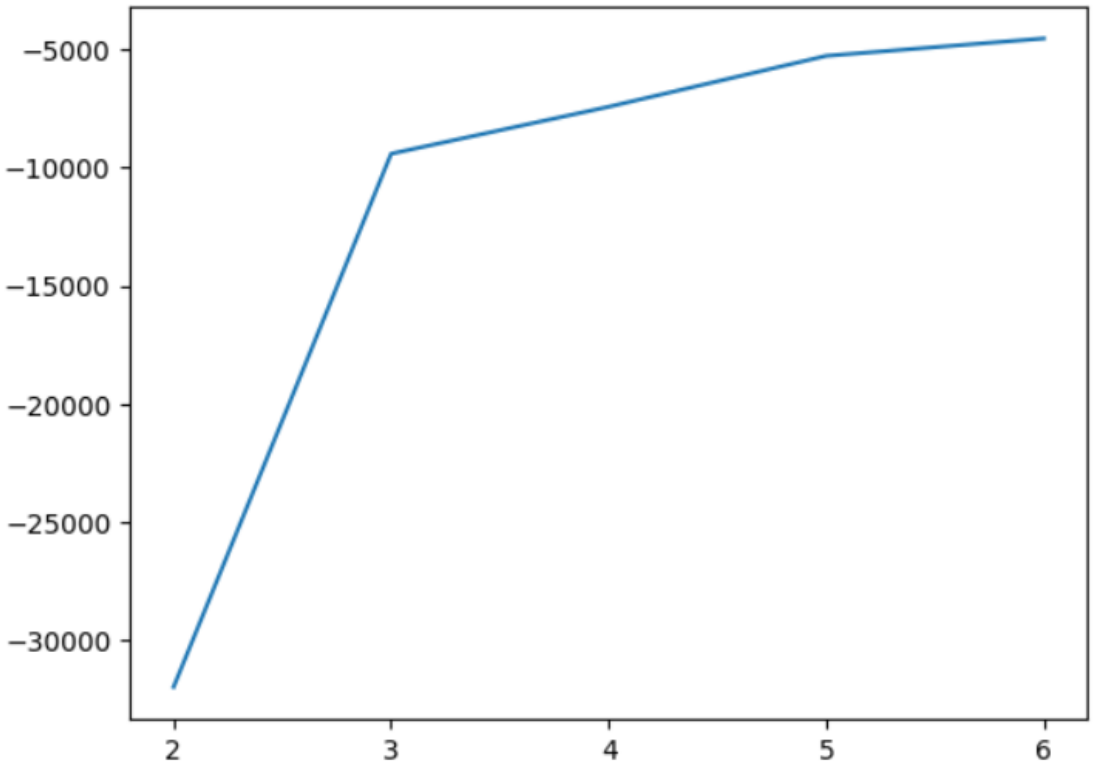

Рисунок 37 – Разность между последовательными значениями инерции

Данная формула показывает нам то, как произошло изменение инерции относительно следующего разбиения. Иными словами, было ли следующее разбиение оптимальнее предыдущего. Видим, что при разбиении на 2 кластера разница в дисперсии составила более 30000, а при разбиении на 4 и более кластера дисперсия менялась крайне незначительно.

На основании полученных результатов было вычислено отношение последовательных разностей инерций по формуле $k_{opt} = argmin((inertia_{k+1} - inertia_k) / (inertia_k - inertia_{k-1}))$.

В данной формуле разница между инерциями на текущей итерации делится на разницу между инерциями предыдущей итерации. Она показывает,

насколько наши улучшения масштабны по сравнению с предыдущим изменением. Результат вычисления формулы представлен на рисунке 38:

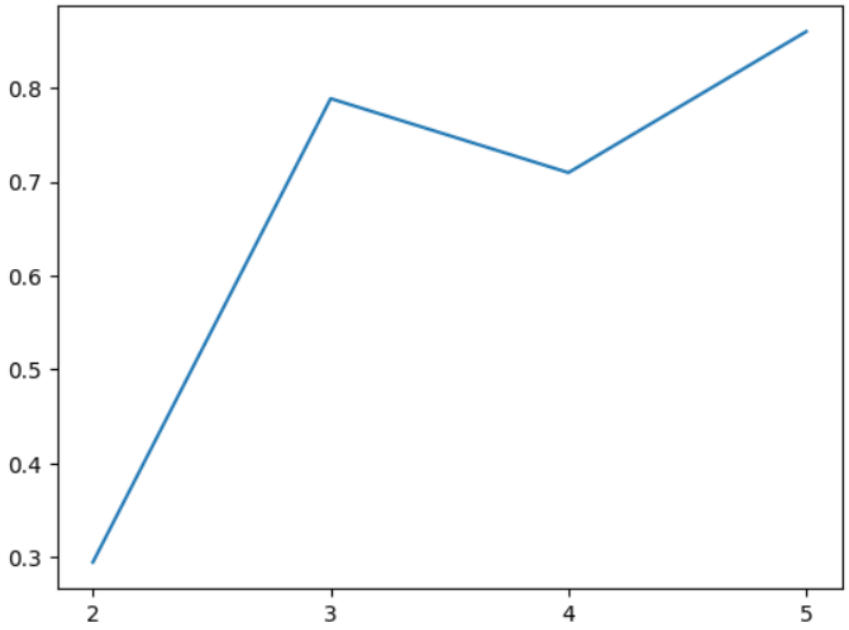

Рисунок 38 – Оптимальное количество кластеров

Заметим, что оптимальным количеством кластеров будет минимальное значение графика, представленного на рисунке 39, так как это значение является точкой, где происходит наибольшее сокращение в отношении значений разностей между последовательными элементами. То есть, переход от значения 1 к значению 2 как к количеству кластеров для нас оптимальнее перехода от 3 к 4 и так далее – именно на значении 2 и была проведена кластеризация при помощи K-Means [9].

Успеваемости в полученных машинным обучением кластерах изображены на рисунке 39:

	Rus	Math	ПБ
Cluster			
0	3.209842	3.166538	9.036628
1	4.248023	4.273848	9.711135

Рисунок 39 – Успеваемость в кластерах, полученных при помощи K-Means

Таким образом, объединенные в предыдущей главе дипломной работы группы учеников со «средним» контекстом, были разбиты на 2 группы в ходе кластеризации.

3.4 Итоговый результат

Итак, перед процессом кластеризации мы имели 27 групп учеников, а после проделанных изменений, манипуляций и проведенной кластеризации их стало 6. Стоит отметить, что, исходя из того, что учащихся из полных семей больше всего, а мальчиков и девочек в изначальной выборке примерно поровну, и несмотря на то что данные поля участвовали в кластеризации, в кластерах данная тенденция также сохраняется. Основным отличием, как мы уже поняли, является успеваемость и контекст.

Исходя из увиденного, группу учеников с индексами «1-1-1» можно назвать «Троечниками с худшим контекстом», учеников с индексами «1-2-1» «Троечниками с плохим контекстом», учеников с индексами «2-3-2» «Примерными семьянинами», учеников с индексами «2-3-3» и «3-3-3» «Всесторонне развитыми отличниками», а полученные на прошлом шаге 2 большие группы можно назвать «Среднестатистическими учениками» и «Хорошистами со слабым контекстом» соответственно.

Важно сделать оговорку, что в последнем случае «слабый контекст» означает то, что у этой группы контекстные переменные хуже, чем у других хорошистов и отличников.

Также очень важным и ценным наблюдением для нашей работы является то, что главным фактором, который может служить для разделения учеников на принципиально разные группы, является семейный контекст. Мы видим, что ученики в группе «1-2-1» кардинально отличаются от более слабых учеников, а контингент группы «2-3-2» отличается от контингента в другой – более сильной группе учеников.

Распределение учащихся в разрезе каждого кластера представлено на рисунке 40:

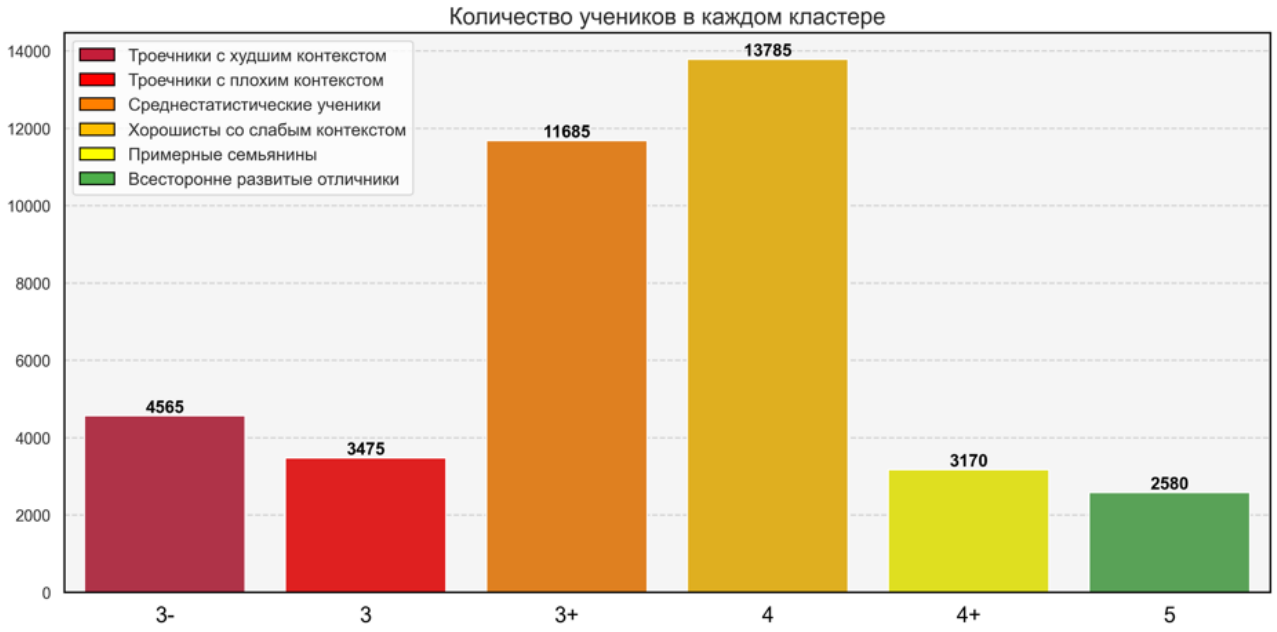

Рисунок 40 – Количество учащихся в каждом из полученных кластеров

Доказательством качественного разделения наших учеников на группы может служить, например, график по среднему ПБ диагностирования НИКО для каждого кластера:

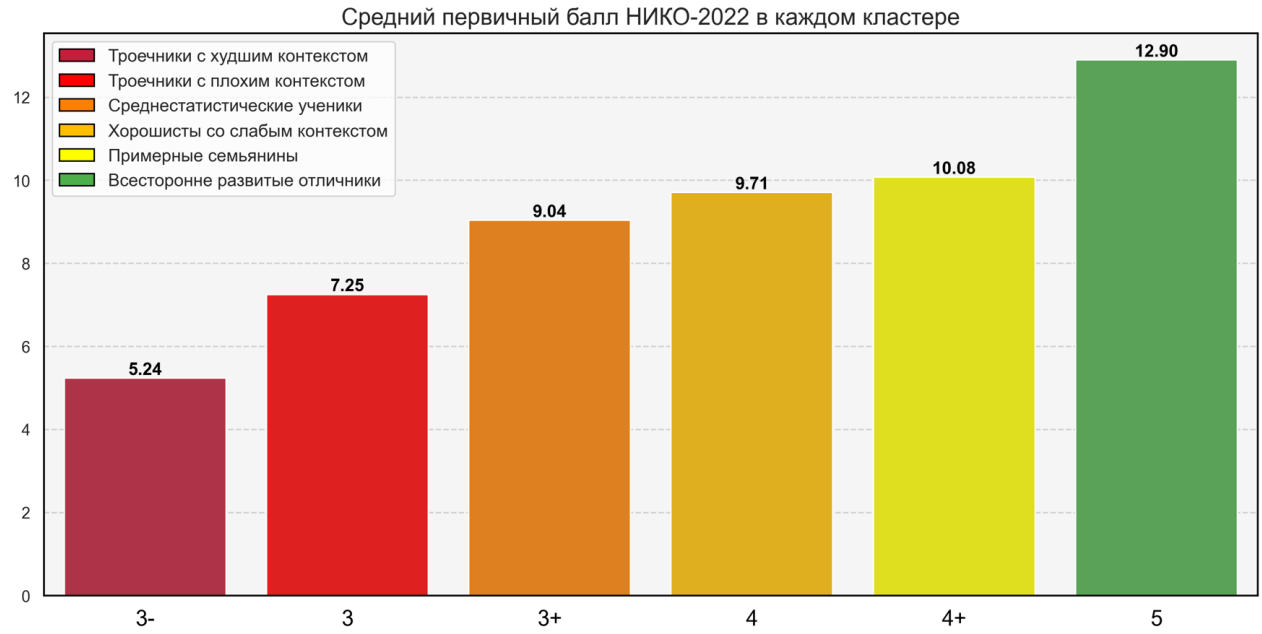

Рисунок 41 – Средний первичный балл НИКО-2022 в каждом кластере

А также графики корреляций между оценками на исходной выборке и в некоторых из кластеров, как на рисунках 42-46:

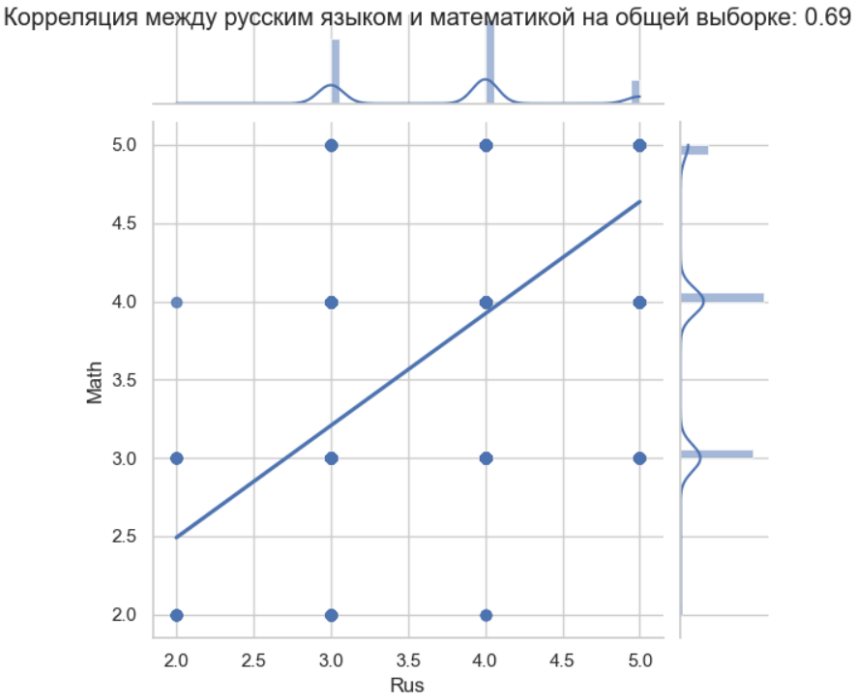

Рисунок 42 – Корреляция между русским языком и математикой на общей выборке

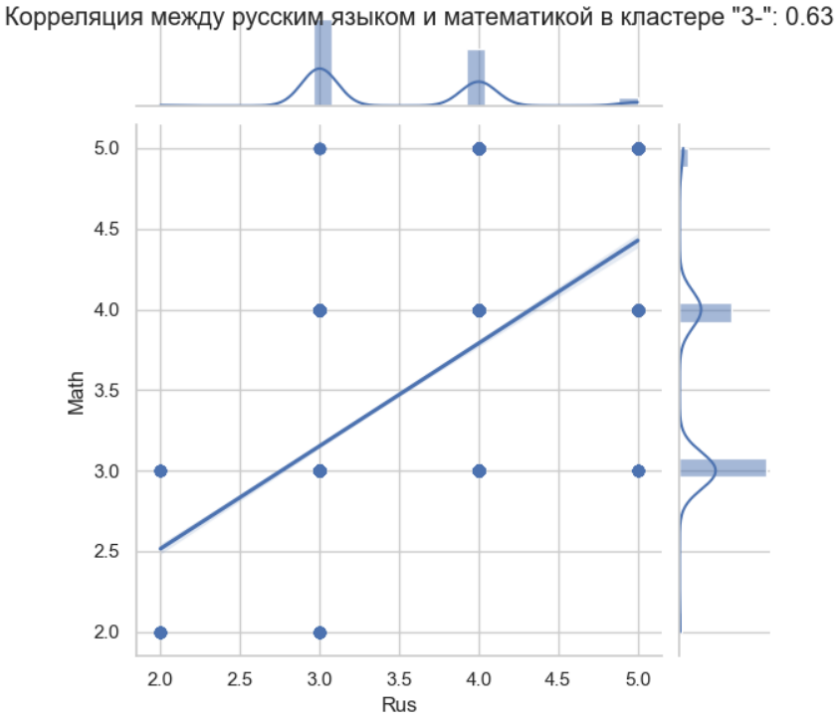

Рисунок 43 – Корреляция между русским языком и математикой в кластере «3-»

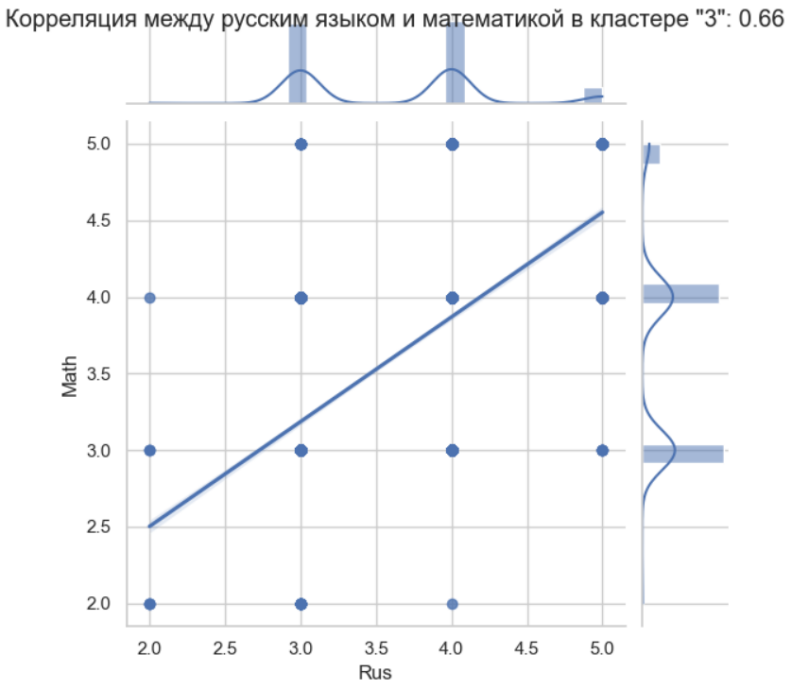

Рисунок 44 – Корреляция между русским языком и математикой в кластере «3»

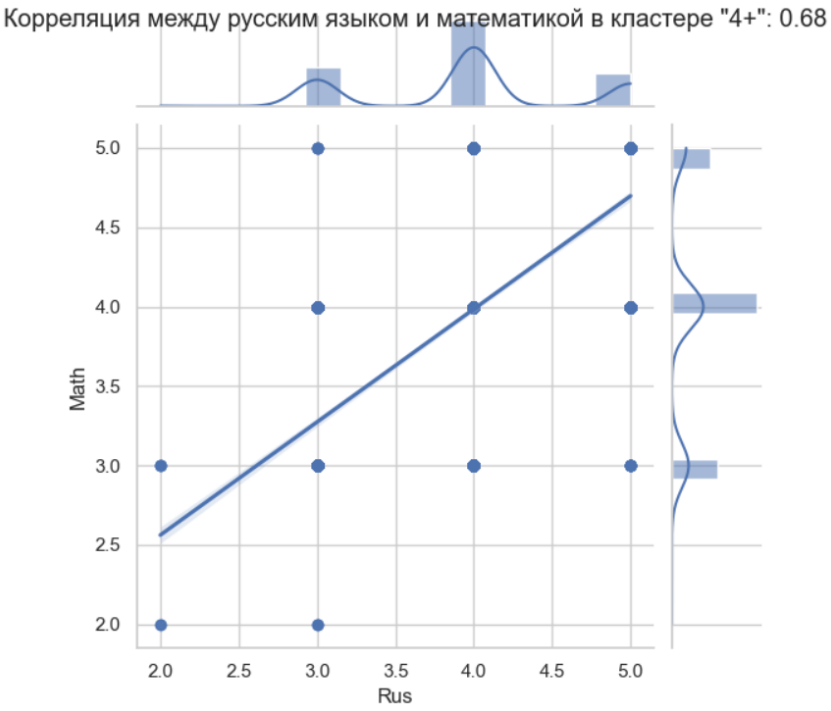

Рисунок 45 – Корреляция между русским языком и математикой в кластере «4+»

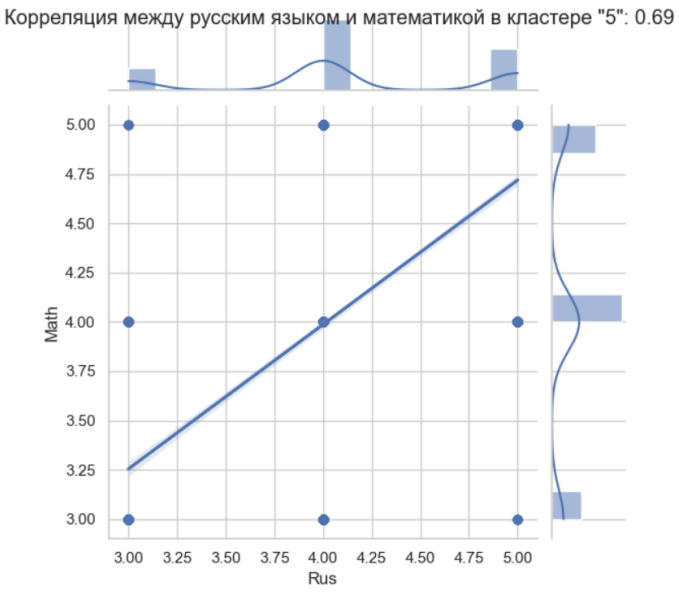

Рисунок 46 – Корреляция между русским языком и математикой в кластере «5»

Как мы можем заметить, уровень корреляции на общей выборке сохраняется в кластерах, что может говорить о качестве проведенного разбиения учащихся на группы.

На рисунках 47 и 48 представлены тепловые карты кластеров с самой слабой и с самой сильной успеваемостями соответственно:

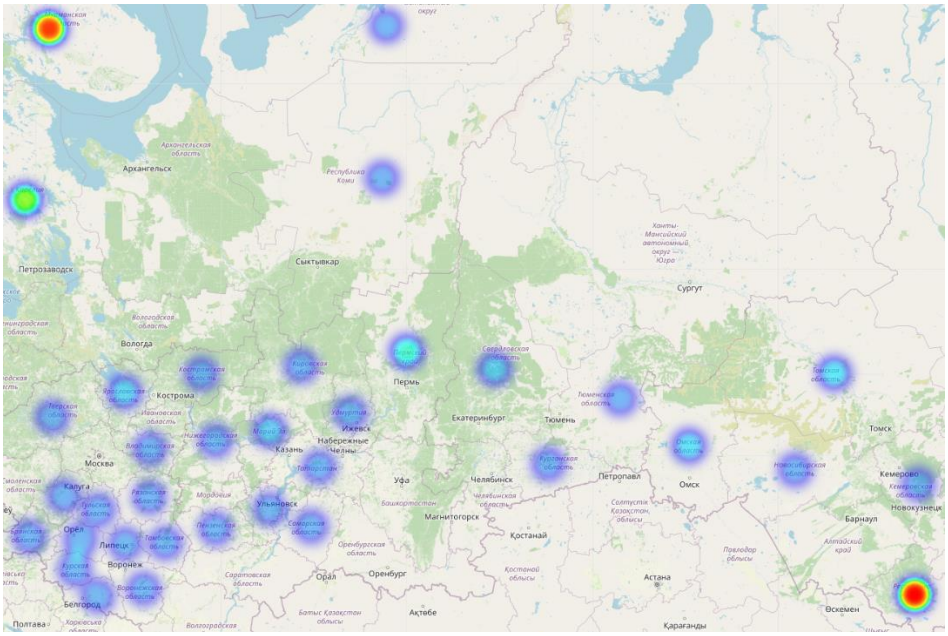

Рисунок 47 – Географическое расположение учащихся с самой слабой успеваемостью

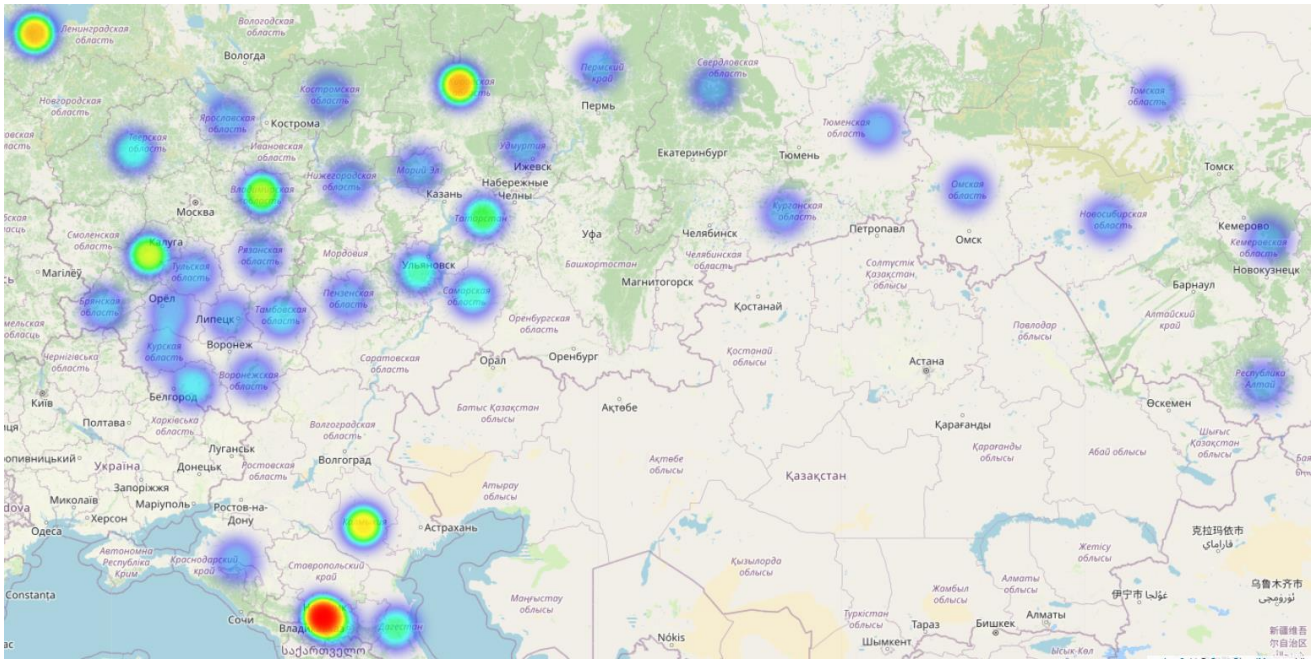

Рисунок 48 – Географическое расположение учащихся с самой сильной успеваемостью

Как мы можем заметить, учащиеся с самой слабой успеваемостью, в основном, располагаются в Мурманской области и в Алтае, а с самой сильной успеваемостью – в Кабардино-Балкарии.

Интересно также посмотреть на школьные оценки по русскому языку и математике в разрезе каждого кластера, как на рисунках 49 и 50:

Рисунок 49 – Средняя оценка по русскому языку в каждом кластере

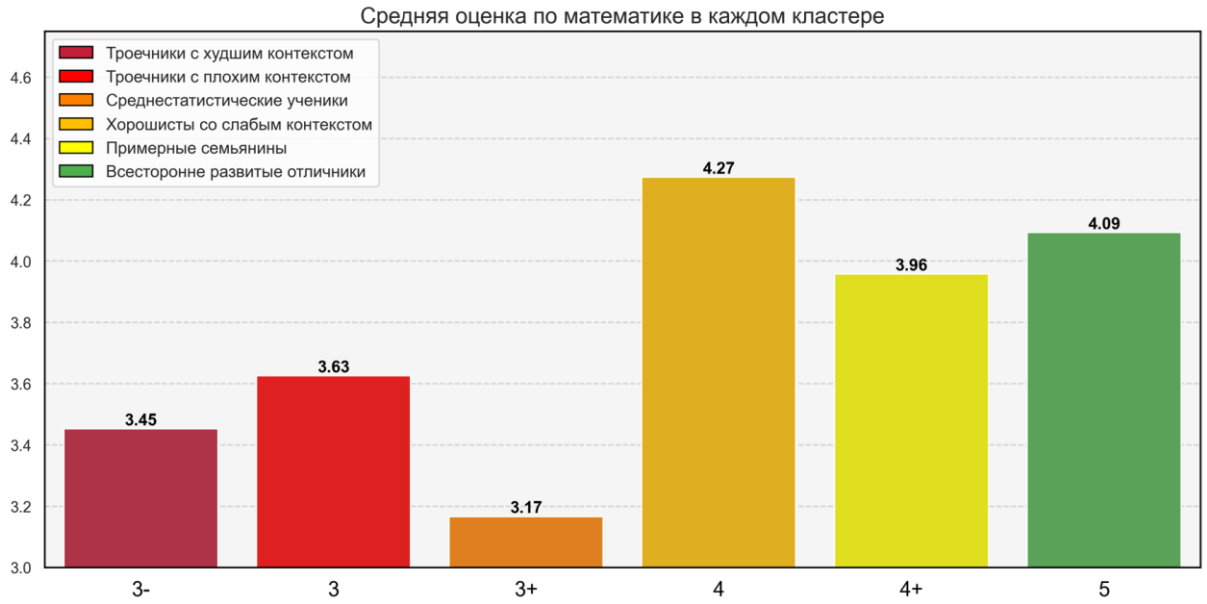

Рисунок 50 – Средняя оценка по математике в каждом кластере

На первый взгляд, все выглядит не так однозначно и радужно, как на прошлых графиках. Однако важно держать в голове тот факт, что проставление оценки в школе – это процесс, который, как правило, связан не только с истинными знаниями того или иного учащегося, но и с другими внешними факторами, такими как, например, отношение учителя к ученику.

Объективность оценки (ПБ), полученной каждым учащимся за диагностическую работу НИКО, наоборот, не подлежит сомнению. Именно поэтому такие проекты, как НИКО, в нынешнее время представляют собой особую ценность – они показывают реальную картину нашего образования.

ЗАКЛЮЧЕНИЕ

В результате выполнения выпускной квалификационной работы бакалавра были проведены исследования, направленные на потенциальное изменение подхода к оценке качества образования.

Основная цель работы заключалась в получении кластеров учащихся для более точных методов оценки образовательных тенденций.

Для достижения этой цели были выполнены ряд задач, включая очистку данных в репрезентативной базе НИКО, преобразование анкетных ответов учащихся в категориальные переменные и проведение кластеризации учащихся на основе их контекста.

Разобранные анкетные ответы были переданы второму участнику проекта для исследования общеобразовательных организаций и их кластеризации с учетом полученных нами контекстных переменных учащихся.

Основным результатом работы стали выделенные на странице 52 кластеры школьников, представляющие собой психологические портреты учащихся, полученные из общей выборки анкетирования НИКО 2022.

Полученные результаты позволяют оценить эффективность функционирования школ и работы учителей в группах с сопоставимым контекстом. Эти данные могут быть использованы для уточнения образовательных программ ОО, а также для более гибкого определения успехов и динамики обучения в различных группах. В дальнейшем предполагается использование этих результатов для проведения более глубоких исследований в области образования.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Национальный институт качества образования –URL: https://fioco.ru/ru/osoko/niko/ (дата обращения 16.02.2024).
2. Федеральный институт оценки качества образования –URL: https://www.fioco.ru/ (дата обращения 21.02.2024).
3. Введение в машинное обучение и его алгоритмы –URL: https://habr.com/ru/articles/448892/ (дата обращения 01.03.2024).
4. Обзор методов классификации в машинном обучении с помощью Scikit-learn –URL: https://tproger.ru/translations/scikit-learn-in-pythonl (дата обращения 01.03.2024).
5. Введение в pandas: анализ данных на Python –URL: https://khashtamov.com/ru/pandas-introduction/ (дата обращения 02.03.2024).
6. Использование Jupyter для разведочного анализа данных –URL: https://habr.com/ru/companies/wunderfund/articles/792970/ (дата обращения 10.03.2024).
7. Алгоритмы кластеризации в машинном обучении –URL: https://moluch.ru/archive/342/77003/ (дата обращения 07.03.2024).
8. Кластеризация: алгоритмы k-means и c-means –URL: https://habr.com/ru/articles/67078/ (дата обращения 01.04.2024).
9. Обзор алгоритмов кластеризации данных –URL: https://habr.com/ru/companies/piter/articles/752258/ (дата обращения 15.04.2024).
10. Себастьян Рашка. Python and machine learning. // Packt, 2017. – 418 с

ПРИЛОЖЕНИЕ А
Исходный код