МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ
УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«РОССИЙСКИЙ ГОСУДАРСТВЕННЫЙ
ПЕДАГОГИЧЕСКИЙ УНИВЕРСИТЕТ им. А. И. ГЕРЦЕНА»

институт информационных технологий и технологического образования
кафедра информационных технологий и электронного обучения

Основная профессиональная образовательная программа
Направление подготовки 09.03.01 Информатика и вычислительная техника
Направленность (профиль) «Технологии разработки программного обеспечения»
форма обучения – очная

Курсовая работа

по дисциплине «Пакеты прикладных программ для статистической обработки и анализа данных»

Статистические методы и алгоритмы машинного обучения в анализе академической успеваемости: Исследование влияния подготовки на экзаменационные результаты

Обучающегося 3 курса
Чирцова Тимофея Александровича

_____

Руководитель:
д.п.н, профессор
_____ Власова Е. З.

«_____» _____ 2024г.

Санкт-Петербург
2023

Оглавление

ВВЕДЕНИЕ 3

Глава 1. ТЕОРЕТИЧЕСКАЯ ЧАСТЬ 5

1.1 Общие сведения о природе данных 5

1.2 Использованные инструменты анализа 7

Глава 2. ПРАКТИЧЕСКАЯ ЧАСТЬ 9

2.1 Обзор данных для анализа 9

2.2 Этап очистки данных 9

2.3 Общая статистика для режима подготовки и экзаменационного режима 10

2.4 Распределение вопросов по уровню сложности на основе медианных оценок студентов по этим вопросам. 12

2.5 Распределение пользователей по числу прохождений вопросов и среднему баллу. 13

2.6 Распределение пользователей по числу прохождений вопросов и числу уникальных вопросов 21

2.7 Распределение пользователей по отношению числа прохождений в режиме подготовки к числу прохождений в экзаменационном режиме 25

2.8 Использование машинного обучения 28

2.9 Поиск паттернов подготовки в случае частного экзамена. 31

2.10 Выбор новой целевой метрики 34

2.11 Вывод по Главе 2. 37

ЗАКЛЮЧЕНИЕ 38

ВВЕДЕНИЕ

Актуальность данной работы обусловлена стремительным развитием образовательных технологий и постоянно меняющимися подходами к обучению и оценке знаний студентов. В современном образовательном пространстве важность точной и объективной оценки академической успеваемости студентов не может быть недооценена. Это обусловлено необходимостью разработки эффективных учебных программ и методик, которые способны удовлетворить потребности как учебных заведений, так и самих студентов.

С учетом увеличения объема доступных образовательных данных и возможностей их анализа, возникает необходимость более глубокого понимания того, как различные факторы влияют на успеваемость студентов. В частности, актуальной становится задача исследования связи между уровнем подготовки студентов и их результатами на экзаменах. Понимание этих взаимосвязей позволит разработать более целенаправленные и эффективные образовательные стратегии, а также поможет студентам лучше готовиться к экзаменам.

Таким образом, данная работа направлена на анализ данных о подготовке студентов и их результатах на экзаменах с целью выявления закономерностей и возможных прогнозов успеваемости. Использование методов анализа данных и машинного обучения в этом контексте открывает новые горизонты для понимания и оптимизации процесса обучения.

Цель исследования заключается в анализе взаимосвязи между уровнем подготовки студентов и их успеваемостью на экзаменах, с последующей оценкой возможности прогнозирования результатов экзаменов на основе данных о подготовке.

Для достижения поставленной цели были определены следующие задачи:

1. Изучение существующих методик и подходов к оценке успеваемости студентов и анализу их подготовки;
2. Сбор и предобработка данных о подготовке студентов и их результатах на экзаменах;
3. Применение статистических методов и методов машинного обучения для анализа взаимосвязи между подготовкой студентов и их успеваемостью на экзаменах;
4. Оценка стабильности и постоянства результатов студентов с использованием метрик, таких как среднеквадратичное отклонение и перцентили;
5. Разработка и проверка гипотез о возможности прогнозирования результатов экзаменов на основе анализа данных о подготовке;
6. Анализ результатов исследования и формулирование выводов о влиянии уровня подготовки на успеваемость студентов на экзаменах.

Обьектом исследования является электронное тестирование студентов

Предметом исследования является методы выявления закономерностей между оценками студентов в режиме подготовки и экзаминационном режиме

Материалом исследования является статистика на платформе Study Ways для студентов ЛЭТИ по физике за вторую половину 2023 года.

Глава 1. ТЕОРЕТИЧЕСКАЯ ЧАСТЬ

1.1 Общие сведения о природе данных

В университете ЛЭТИ на протяжении последних двух лет активно используется образовательная платформа Study Ways, разработанная на кафедре данного вуза, для оценки подготовленности студентов по дисциплине физика. Эта платформа включает в себя специально разработанную систему тестирования, обладающую функциями защиты от списывания.

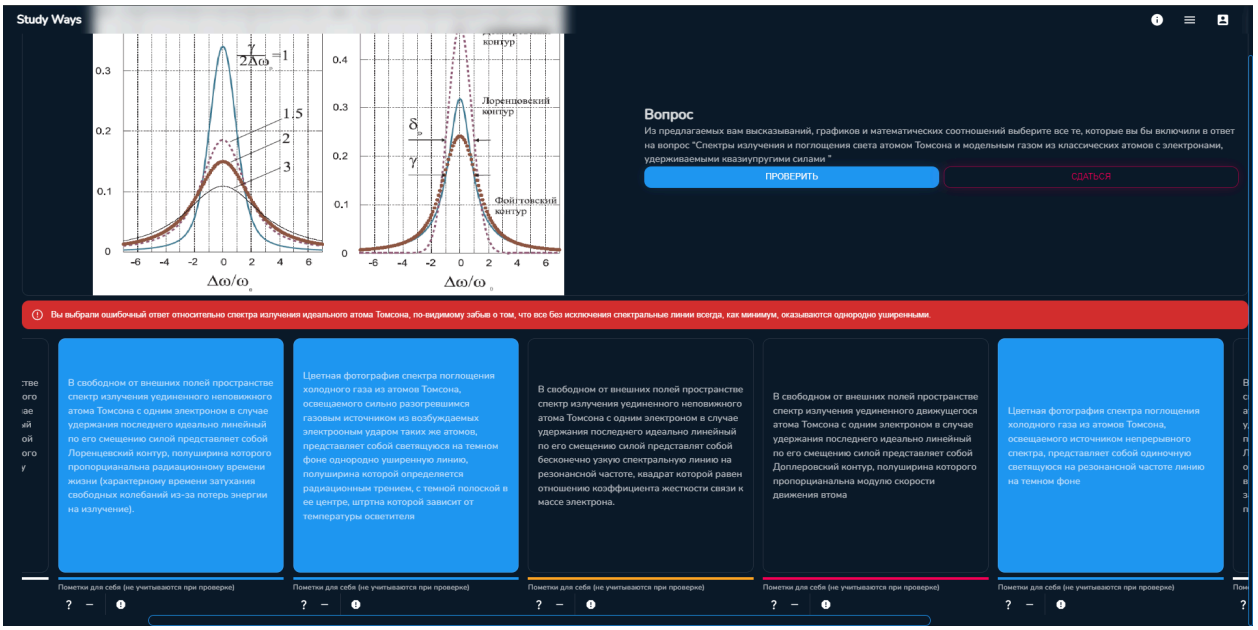


Рисунок 1. Интерфейс системы тестирования в платформе Study Ways

В рамках режима подготовки система предоставляет пользователям возможность выбора вопроса (или перехода к нему непосредственно из соответствующей лекции). Для каждого вопроса система генерирует до 10 вариантов ответов, расположенных в случайном порядке. Пользователю необходимо отметить все правильные ответы. После совершения выбора система осуществляет проверку, и в случае ошибки пользователя предоставляется подсказка по наиболее значительной ошибке, после чего пользователю снова предлагается выбрать правильные ответы. Пользователь

также имеет возможность прекратить попытку ответа на вопрос. Прохождение одного вопроса определяется либо выбором всех правильных ответов, либо отказом от дальнейшего ответа. Важно отметить, что после того как студент выбирает все правильные ответы и подтверждает свой выбор, тестирование завершается, предотвращая возможность создания скриншотов или сохранения ответов иными способами. Это является одной из мер, направленных на предотвращение списывания.

Студенты оцениваются по 100-балльной системе за прохождение каждого вопроса. Механизм начисления баллов устроен следующим образом: на каждой попытке прохождения вопроса студент выбирает ответы, для которых преподавателями заранее установлен уровень сложности - сложный, средний или легкий. Уровень сложности влияет на количество начисляемых баллов: чем выше сложность, тем больше баллов студент получает за правильный ответ и тем меньший штраф применяется за ошибку.

В процессе разработки системы штрафов за множественные попытки было принято решение отказаться от традиционной модели, предполагающей вычитание фиксированного количества баллов за каждую попытку. Вместо этого был выбран подход, при котором влияние каждой последующей попытки на общий балл уменьшается в 0.8 раза по сравнению с предыдущей. Такая система позволяет избежать недооценки усилий студента в ситуациях, когда он неоднократно пытается ответить на вопрос, правильно выбирая большинство ответов, но из-за большого количества попыток получает сниженный балл. Это способствует более справедливой и объективной оценке знаний студентов.

Экзаменационный режим отличается от режима подготовки тем, что в нем происходит добавление специальных вариантов ответов, помеченных преподавателями как "только для экзамена". Это существенно уменьшает вероятность ситуаций, когда студенты, используя информацию, полученную

во время подготовки, могут просто заучивать правильные ответы, вместо того чтобы развивать глубокое понимание материала.

Кроме того, в экзаменационном режиме преподаватели имеют возможность устанавливать ограничения на количество попыток ответа на каждый вопрос. Это позволяет более точно контролировать процесс тестирования и увеличивает его соответствие реальным экзаменационным условиям.

В контексте режима подготовки система Study Ways создает условия, аналогичные диалогу студента с максимально лояльным и понимающим преподавателем. В этом режиме система не только указывает на ошибки студента, но и предоставляет подсказки и направления для нахождения правильного решения, способствуя более эффективному и глубокому обучению.

1.2 Использованные инструменты анализа

В рамках данного исследования для оценки взаимосвязи между подготовкой студента и его результатами в экзаменационном режиме были применены следующие методы статистического анализа:

1. Кластерный анализ: Метод, используемый для группировки объектов в так называемые кластеры. Объекты внутри одного кластера обладают схожими свойствами, в то время как объекты разных кластеров отличаются между собой.

2. Метод K-means: Популярный метод кластеризации, целью которого является разделение набора данных на предопределенное количество

кластеров. Основная идея метода заключается в минимизации суммы расстояний от точек до центров их кластеров.

3. Случайный лес: Алгоритм машинного обучения, основанный на концепции ансамбля деревьев решений. Этот метод эффективен для задач классификации и регрессии и отличается высокой точностью и устойчивостью к переобучению.

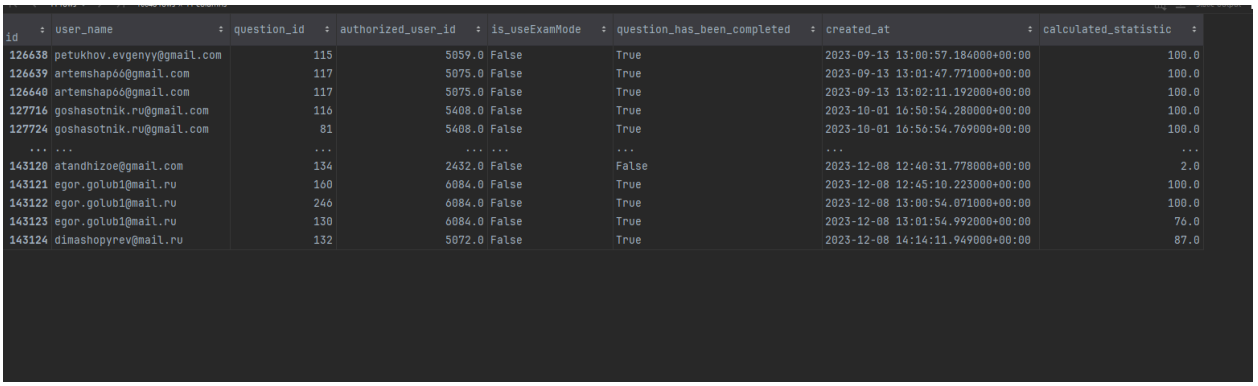
4. Метрика RSME: Среднеквадратичная ошибка, мера расхождения между предсказанными и фактическими значениями. Является стандартной метрикой для оценки точности моделей регрессии.

5. Перцентильное распределение: Статистическая мера, показывающая, какой процент наблюдений падает ниже определенного значения. В контексте данного исследования перцентили используются для анализа распределения оценок студентов.

Глава 2. ПРАКТИЧЕСКАЯ ЧАСТЬ

2.1 Обзор данных для анализа

В этой работе будут использованы данные полученные на платформе StudyWays за второе полугодье по предмету физика. Общий вид данных изображен на рисунке.



id	user_name	question_id	authorized_user_id	is_useExamMode	question_has_been_completed	created_at	calculated_statistic
126638	petukhov.evgenyy@gmail.com	115	5059.0	False	True	2023-09-13 13:00:57.184000+00:00	100.0
126639	artemshapóó@gmail.com	117	5075.0	False	True	2023-09-13 13:01:47.771000+00:00	100.0
126640	artemshapóó@gmail.com	117	5075.0	False	True	2023-09-13 13:02:11.192000+00:00	100.0
127716	goshasotnik.ru@gmail.com	116	5408.0	False	True	2023-10-01 16:50:54.280000+00:00	100.0
127724	goshasotnik.ru@gmail.com	81	5408.0	False	True	2023-10-01 16:56:54.769000+00:00	100.0
...	...	...	...	...	...	...	...
143120	atandhizoe@gmail.com	134	2432.0	False	False	2023-12-08 12:40:31.778000+00:00	2.0
143121	egor.golub1@mail.ru	160	6084.0	False	True	2023-12-08 12:45:10.223000+00:00	100.0
143122	egor.golub1@mail.ru	246	6084.0	False	True	2023-12-08 13:00:54.071000+00:00	100.0
143123	egor.golub1@mail.ru	130	6084.0	False	True	2023-12-08 13:01:54.992000+00:00	76.0
143124	dimashopyrev@mail.ru	132	5072.0	False	True	2023-12-08 14:14:11.949000+00:00	87.0

Рисунок 2.1 Фрагмент датасета

Как можно увидеть на представленном рисунке, набор данных включает следующую информацию: идентификатор попытки, идентификатор вопроса, идентификатор пользователя, индикатор, определяющий, являлся ли вопрос тренировочным или экзаменационным, индикатор, указывающий, был ли вопрос успешно пройден или студент не справился с заданием, дату и время завершения попытки, а также окончательный балл.

2.2 Этап очистки данных

Перед началом анализа данных необходимо провести тщательную очистку используемого датасета. Этот процесс включает удаление строк, которые не содержат информации об авторизованном пользователе, не имеют рассчитанного количества баллов или не указывают дату создания. В случае пользователей, набравших менее 0 баллов (возможная ситуация в рамках заданной системы оценки), баллы корректируются до 0. Кроме того,

осуществляется отбор данных, ограничивающийся периодом после середины лета 2023 года.

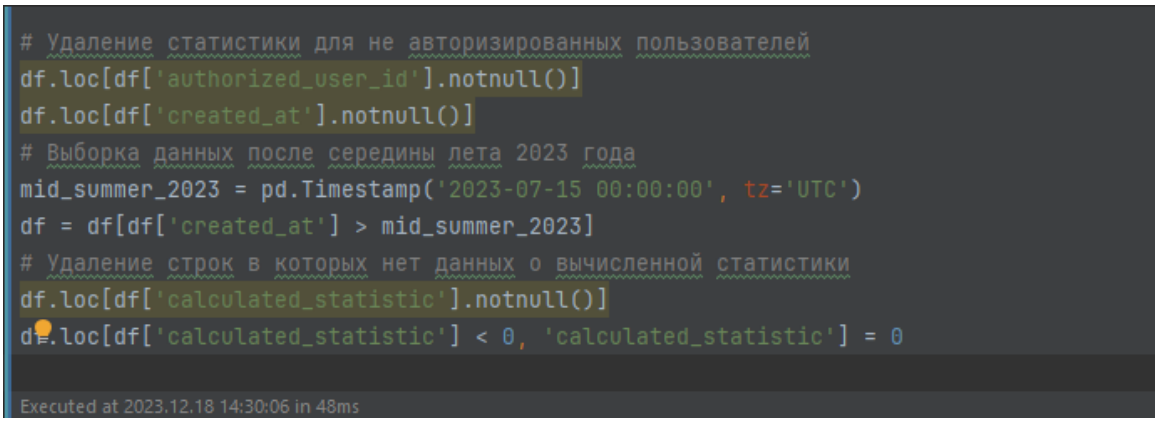


Рисунок 2.2 Код для очистки датасета

2.3 Общая статистика для режима подготовки и экзаменационного режима

В рамках первого этапа анализа предполагается построение графика, который демонстрирует распределение баллов в соответствии с обоими режимами сдачи тестов. Это позволит наглядно оценить и сравнить характеристики результатов, полученных в тренировочном и экзаменационном режимах.

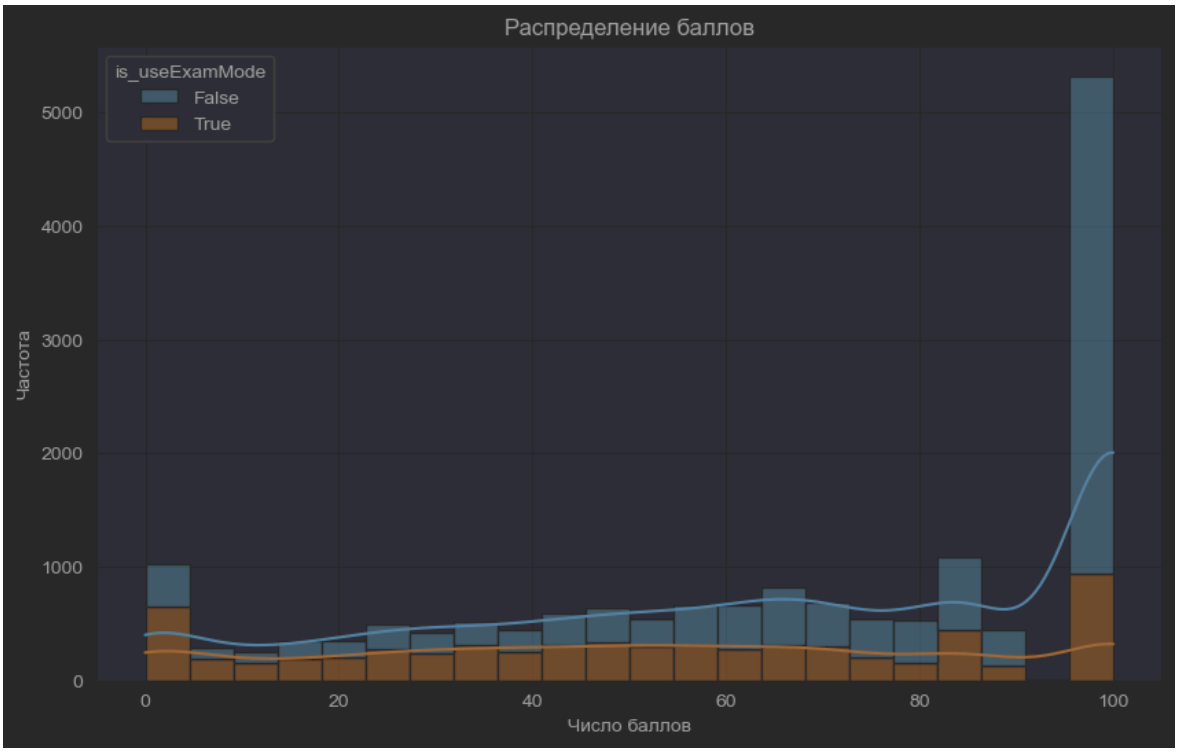


Рисунок 2.3 Столбчатый график распределения баллов

Из анализа представленного графика видно, что в режиме подготовки (обозначенном голубым цветом) студенты часто достигают максимального результата в 100 баллов, в то время как в экзаменационном режиме данное явление наблюдается менее выраженно. Для более детального понимания распределения результатов будут рассчитаны ключевые статистические показатели: среднее значение, медиана, а также 25-й и 75-й процентили.

is_useExamMode	count	mean	50%	25%	75%
False	10146.0	74.016558	84.0	55.0	100.0
True	6502.0	52.003999	52.0	27.0	78.0

Рисунок 2.4 Рассчитанные значения для всех прохождений в тренировочном и экзаменационном режиме

Из анализа представленного рисунка следует, что количество попыток в экзаменационном режиме примерно вдвое меньше по сравнению с

тренировочным режимом. Кроме того, наблюдается существенная разница в показателях успеваемости: средняя оценка в экзаменационном режиме оказывается на 20 пунктов ниже по сравнению с режимом подготовки, тогда как медианная оценка - на 30 пунктов ниже.

2.4 Распределение вопросов по уровню сложности на основе медианных оценок студентов по этим вопросам.

Следующий этап анализа предоставленных данных заключается в оценке равенства сложности вопросов. В рамках этого исследования были построены графики для обоих режимов тестирования. На этих графиках для каждого вопроса была вычислена медианная оценка, после чего результаты были упорядочены по убыванию для наглядного сравнения и выявления возможных различий в сложности вопросов между двумя режимами.

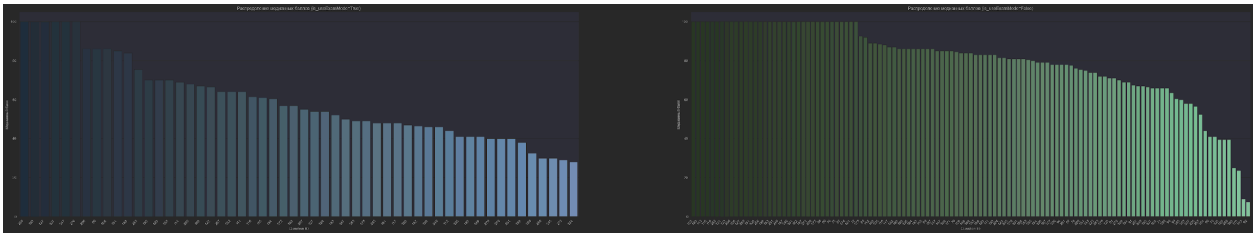


Рисунок 2.5 Столбчатый график распределения минимальной оценки по вопросам

На основе данных, представленных на графике, можно сделать вывод, что в режиме подготовки (изображенном справа) наблюдается значительное количество вопросов, по которым студенты часто набирают максимальные 100 баллов. Это указывает на то, что около 70% вопросов в данном режиме, по-видимому, не представляют существенной сложности для студентов. В контрасте с этим, в экзаменационном режиме количество вопросов, по которым студенты достигают максимального балла, значительно меньше. График для этого режима предполагает, что сложность вопросов более равномерно распределена по всем уровням, что свидетельствует о более строгой оценке.

2.5 Распределение пользователей по числу прохождений вопросов и среднему баллу.

При анализе результатов студентов с целью прогнозирования их успеваемости на экзаменах можно предположить, что студент, демонстрирующий высокие результаты в учебе, будет стремиться к тщательной подготовке к экзамену, зачету или проверочной работе. Таким образом, построение графика для тренировочного режима, на котором по оси абсцисс (X) отложено общее количество попыток, а по оси ординат (Y) — средняя оценка, позволит визуализировать эту закономерность. Согласно данному подходу, наиболее успешные студенты будут располагаться в правой верхней части графика, поскольку они совершают множество попыток и достигают высоких оценок. Хорошие студенты окажутся в левой верхней части, демонстрируя хорошие результаты при меньшем количестве попыток. Наименее успешные студенты будут сосредоточены в левой нижней части графика, указывая на низкие оценки и небольшое количество попыток. Правая нижняя часть графика, вероятно, будет пустой, так как в тренировочном режиме студенты имеют возможность неоднократно проходить тесты, что делает многократное получение низких оценок маловероятным.

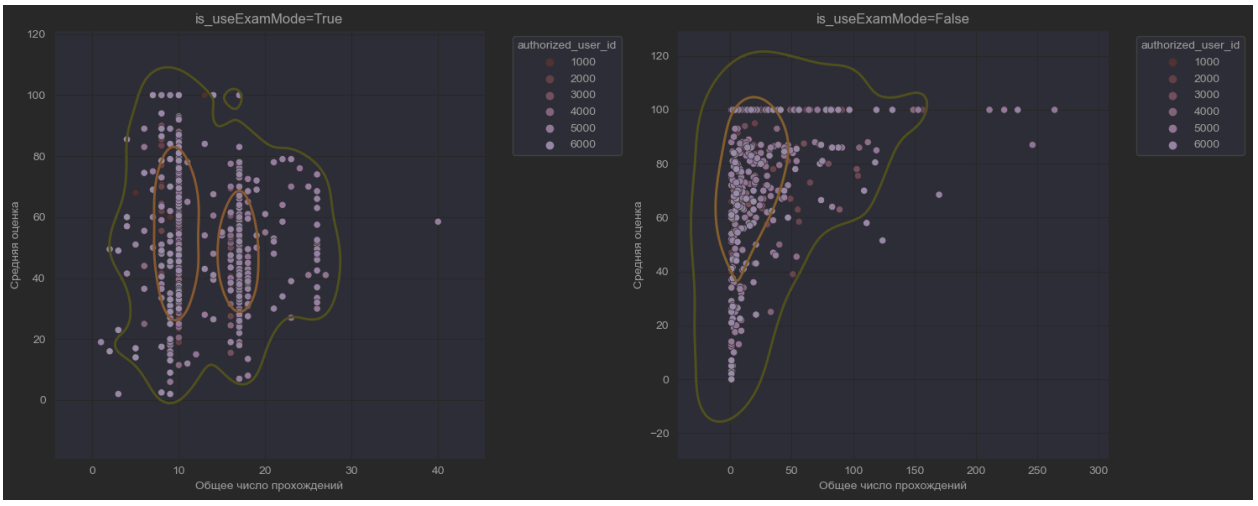


Рисунок 2.6 Точечные графики распределения пользователей по числу прохождений и количеству баллов для тренировочного и экзаменационного режима

На основании анализа графиков, отображающих результаты студентов в тренировочном и экзаменационном режимах, можно заметить интересные закономерности. На графике для тренировочного режима (справа) большинство студентов выполняет до 50 попыток, при этом их результаты варьируются в пределах 60-80 баллов. На графике для экзаменационного режима (слева) видны две группы пользователей: первая делает около 10 попыток, вторая — около 18.

Учитывая сложность прямого вывода из этих данных, целесообразно рассмотреть вариант кластеризации пользователей в тренировочном режиме и последующего применения полученных меток кластеризации к данным экзаменационного режима. При этом можно использовать различные методы кластеризации: агломеративная кластеризация, K-средних, DBSCAN.

Выполним кластеризацию всеми тремя приведенными способами с целью выявления наиболее удачного для наших целей.

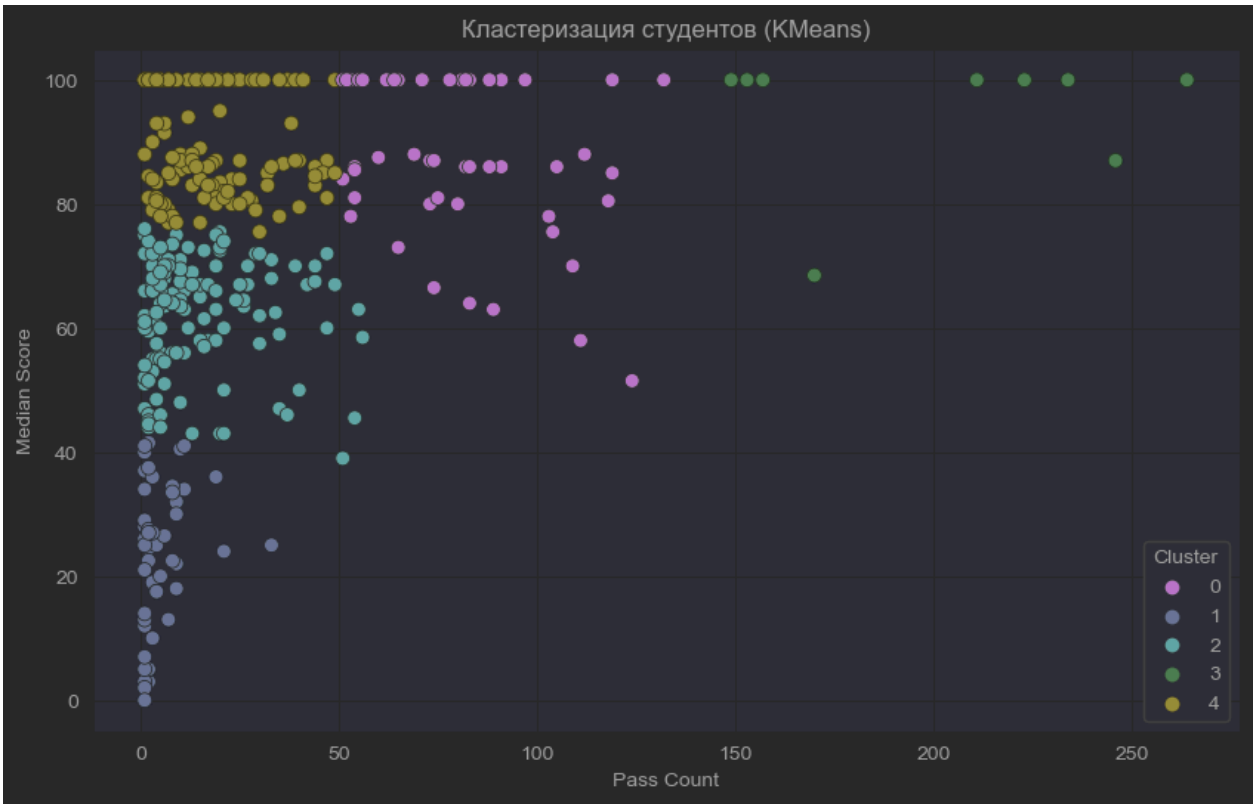


Рисунок 2.7 Точечные графики распределения пользователей по числу прохождений и количеству баллов для тренировочного режима с наложением меток кластеризации методом KMeans

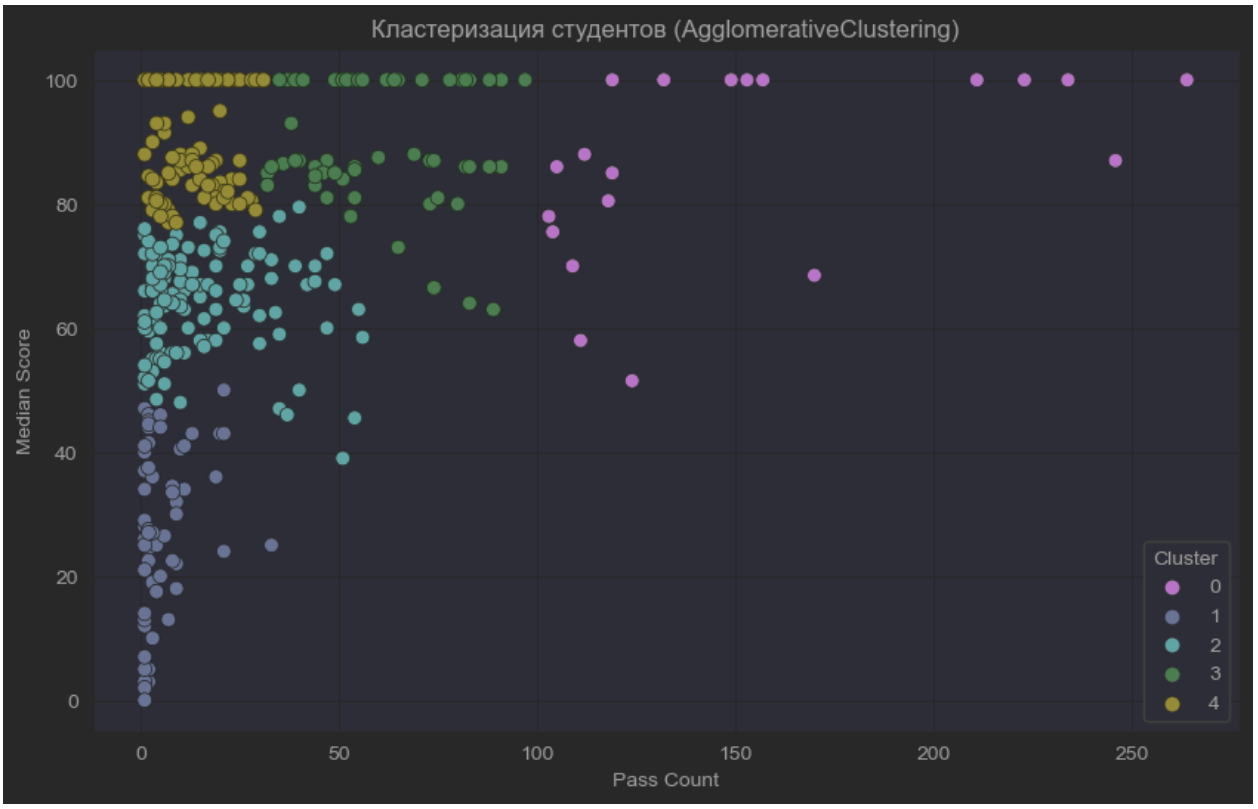


Рисунок 2.7 Точечные графики распределения пользователей по числу прохождений и количеству баллов для тренировочного режима с наложением меток кластеризации методом Agglomerative Clustering

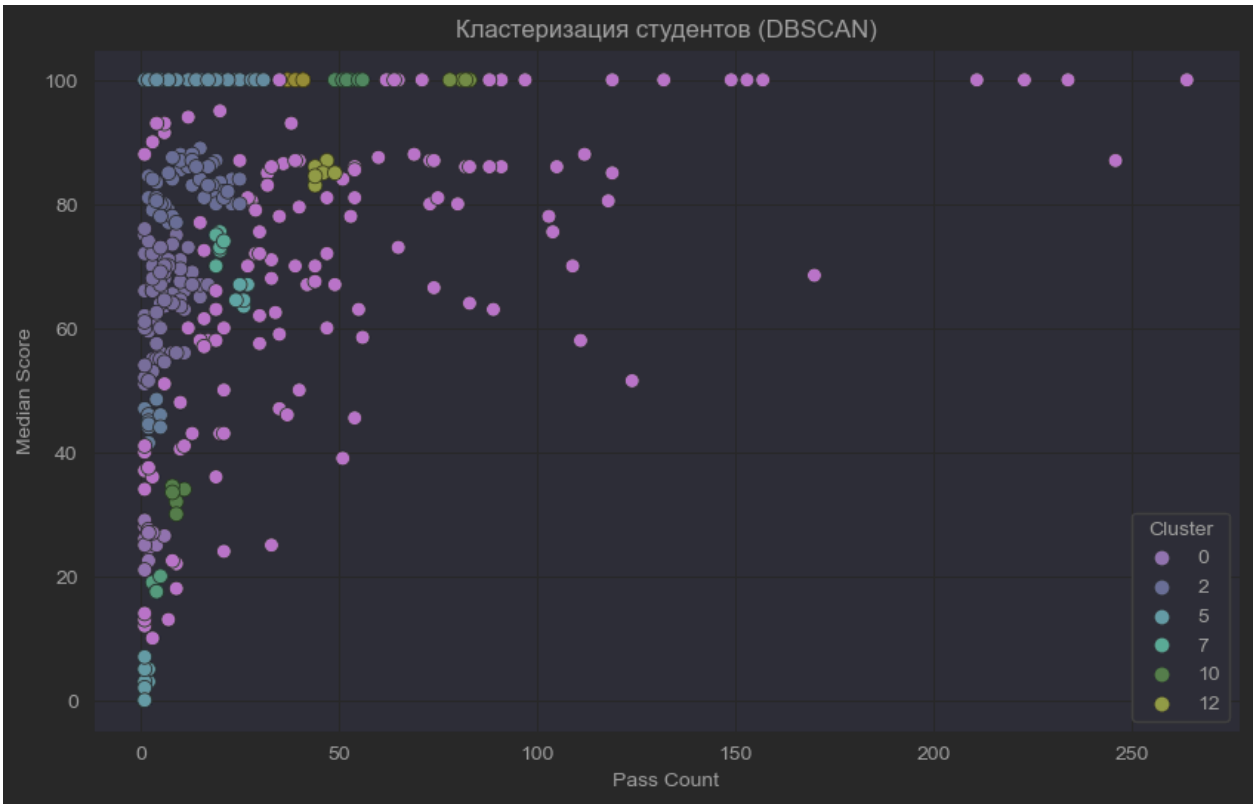


Рисунок 2.8 Точечные графики распределения пользователей по числу прохождений и количеству баллов для тренировочного режима с наложением меток кластеризации методом DBSDCAN

В результате проведенного анализа было установлено, что кластеризация с использованием метода KMeans демонстрирует наиболее значимые результаты. Она позволяет классифицировать студентов на основе их подготовки и достигнутых результатов на следующие группы: группа 1 - студенты, которые уделяют подготовке минимальное количество времени и получают низкие оценки; группа 2 - студенты с умеренными результатами; группа 4 - студенты, показывающие высокие результаты; группа 0 - студенты, активно занимающиеся подготовкой и получающие высокие баллы; группа 3 - студенты, наиболее интенсивно готовящиеся и достигающие высоких результатов.

Наложим метки кластеризации на график в экзаменационном режиме.

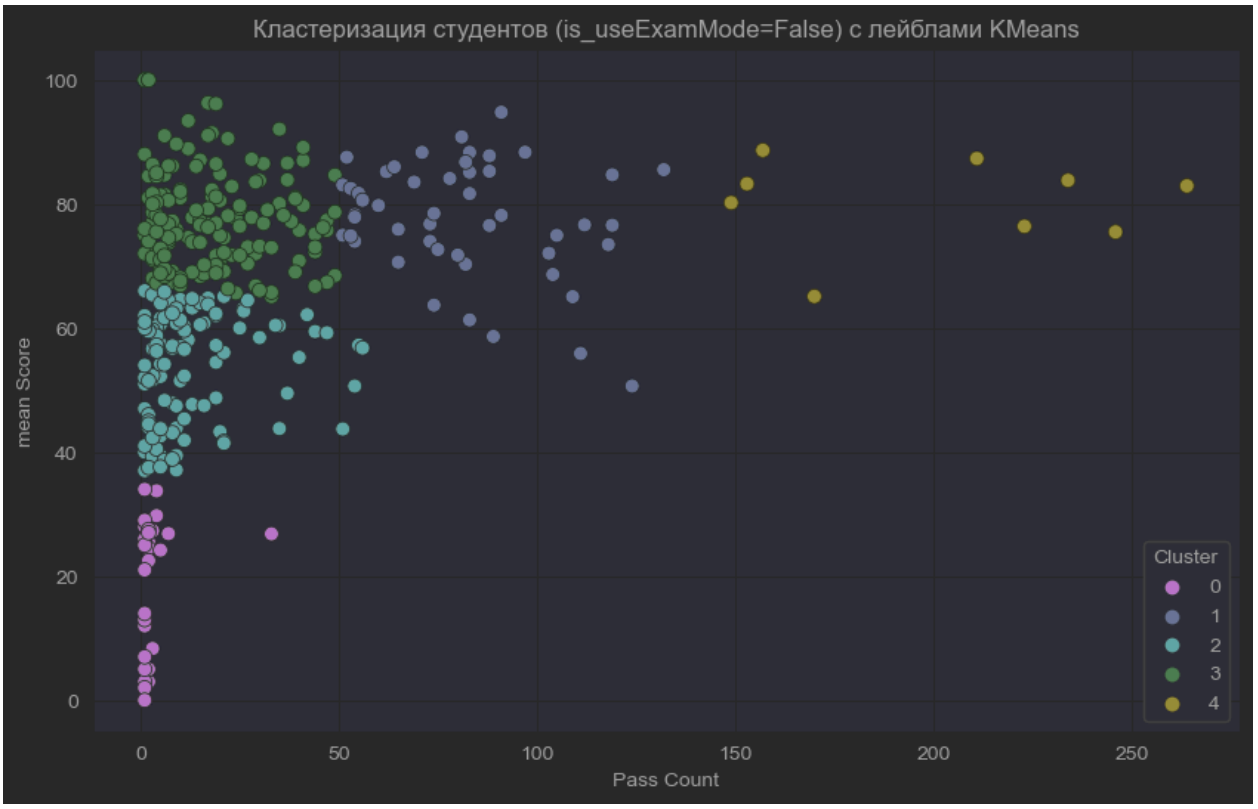


Рисунок 2.9 Точечные графики распределения пользователей по числу прохождений и количеству баллов для тренировочного режима с наложением меток кластеризации методом KMeans

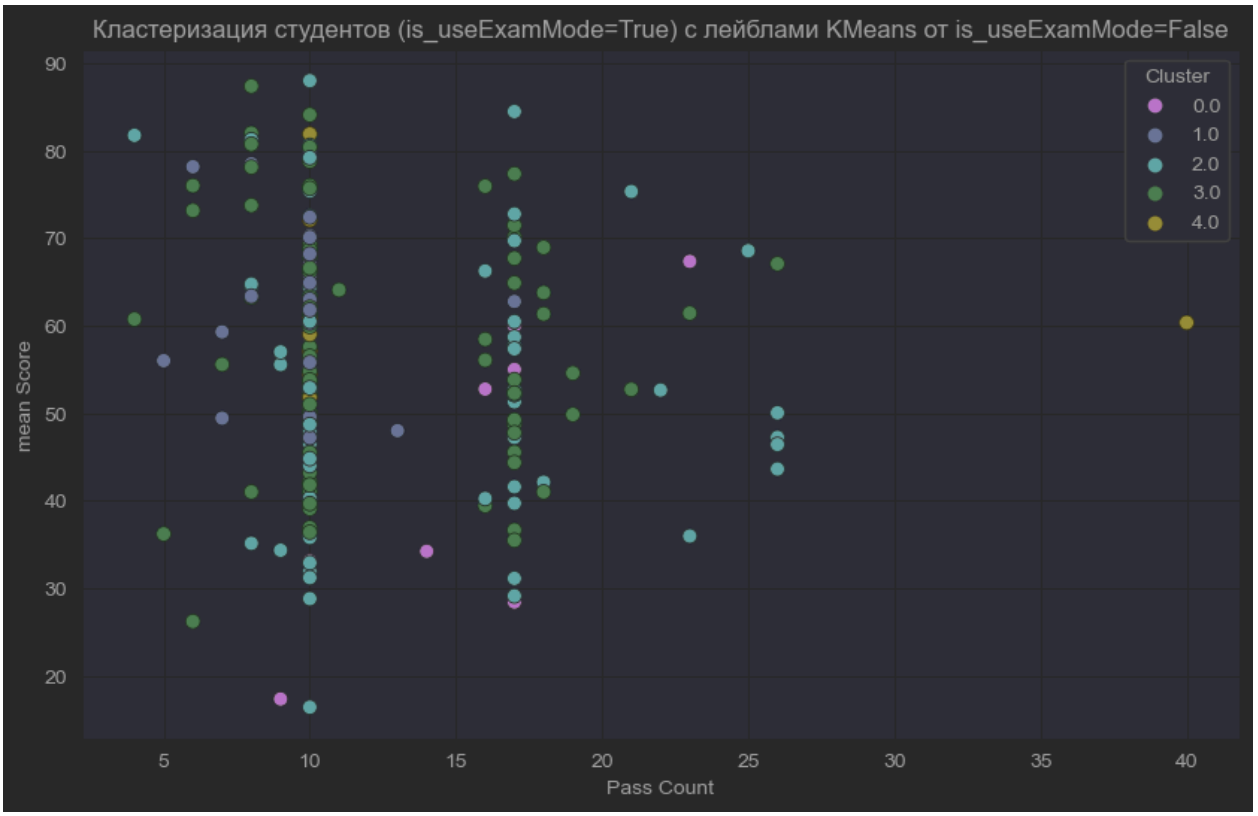


Рисунок 2.9 Точечные графики распределения пользователей по числу прохождений и количеству баллов для экзаменационного режима с наложением меток кластеризации методом KMeans, полученных из предыдущего графика

На основе анализа текущих графиков не удается четко определить степень взаимосвязи между принадлежностью к определенному кластеру и итоговыми оценками студентов. Для более детального изучения этой связи целесообразно построить столбчатые диаграммы, отражающие распределение баллов в экзаменационном режиме для каждого кластера. Это позволит более наглядно оценить различия в успеваемости студентов в зависимости от их кластерной принадлежности.

Рисунок 2.10 Столбчатые графики распределения баллов в экзаменационном режиме для каждого из кластеров, полученного на рисунке 2.9

Из анализа представленных графиков можно заметить, что кластеры 4 и 1 демонстрируют более высокие результаты по сравнению с кластерами 0, 2 и 3, причем кластер 0 выделяется особенно низкими результатами. Рассматривая график кластеризации для тренировочного режима, можно сделать вывод, что кластеры 0, 2 и 3 объединяют студентов, которые проводили меньше времени на подготовку, особенно это касается кластера 0, где студенты не только готовились меньше, но и получали низкие оценки в процессе подготовки. Вычислим средние оценки для всех кластеров.

KMeans	count	mean
0.0	64.0	53.476996
1.0	38.0	62.196674
2.0	10.0	44.652192
3.0	9.0	62.861111
4.0	102.0	58.235868

Рисунок 2.11 Вычисленные значения количества и средней оценки для каждого из кластеров, полученного на рисунке 2.9

Анализ средних оценок по кластерам показывает, что, несмотря на кажущуюся схожесть результатов, реальное влияние кластерной принадлежности на итоговые оценки становится заметным при учете структуры экзамена, состоящего из 5-6 вопросов. В таких условиях студенты из кластеров 3 и 1 могли бы достичь наивысшей оценки, равной 5, в то время как студенты из кластера 4 скорее всего получили бы оценку 4. Студенты из кластеров 2 и 0, согласно этим данным, получили бы оценки на уровне 3.

Это позволяет подтвердить гипотезу о том, что анализ сочетания общего числа попыток и средних баллов в тренировочном режиме может быть эффективным инструментом для предсказания оценок студентов на экзамене. Таким образом, данные графики и кластеризация предоставляют ценную информацию для прогнозирования успеваемости студентов.

2.6 Распределение пользователей по числу прохождений вопросов и числу уникальных вопросов

Принимая во внимание, что в режиме подготовки студенты имеют возможность проходить одни и те же вопросы многократно, что может привести к заучиванию ответов и, как следствие, к искажению реальных знаний, целесообразно провести кластеризацию с учетом не только общего

количества попыток, но и разнообразия вопросов, с которыми студенты сталкивались в процессе подготовки. Такой подход позволит получить более объективное представление об их уровне подготовки.

Для этого можно рассчитать два ключевых показателя для каждого студента: количество уникальных вопросов, с которыми он столкнулся в ходе подготовки, и общее количество прохождений этих вопросов. Эти данные затем могут быть использованы для кластеризации студентов, что позволит выделить различные группы на основе их подхода к учебному процессу. Это может быть особенно полезно для выявления тех студентов, которые могут стремиться к механическому запоминанию ответов, в отличие от тех, кто ищет более глубокое понимание материала.

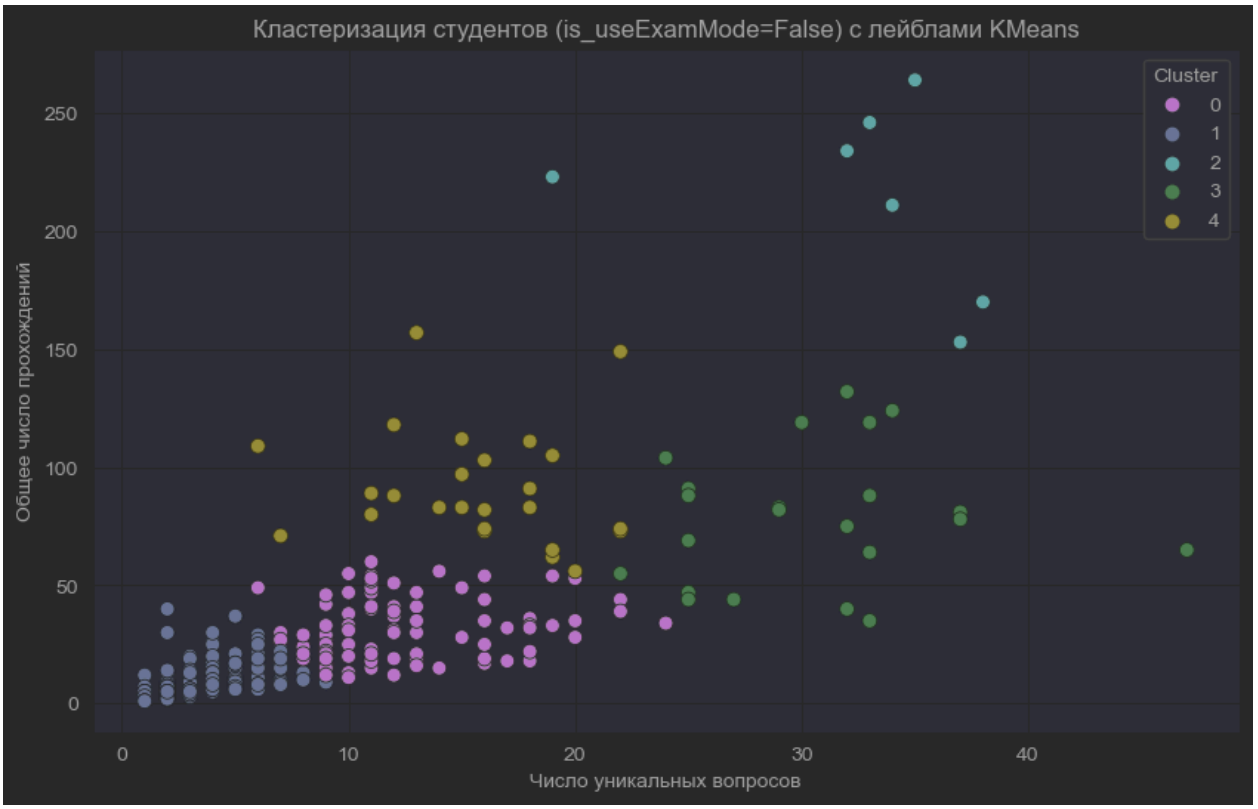


Рисунок 2.12 Точечный график распределения пользователей по критериям числа уникальных пройденных ответов и общему числу прохождений

Исходя из анализа графика, можно заметить, что кластер 2 выделяется существенно высоким количеством уникальных вопросов и общим числом попыток, что свидетельствует о широком охвате материала и активной подготовке. В то же время, кластер 4, имея сопоставимое количество попыток с кластером 3, демонстрирует меньшее число уникальных вопросов, что может указывать на более узкую специализацию или сосредоточение на определенных темах. Кластеры 0 и 1, с другой стороны, характеризуются низкими показателями как по количеству попыток, так и по количеству уникальных вопросов, что может отражать менее интенсивную подготовку.

Для дальнейшего анализа и лучшего понимания того, как различные стратегии подготовки влияют на успехи на экзаменах, целесообразно построить графики распределения баллов в экзаменационном режиме для каждого из кластеров. Это позволит увидеть, есть ли корреляция между количеством и разнообразием попыток в тренировочном режиме и финальными оценками на экзамене.

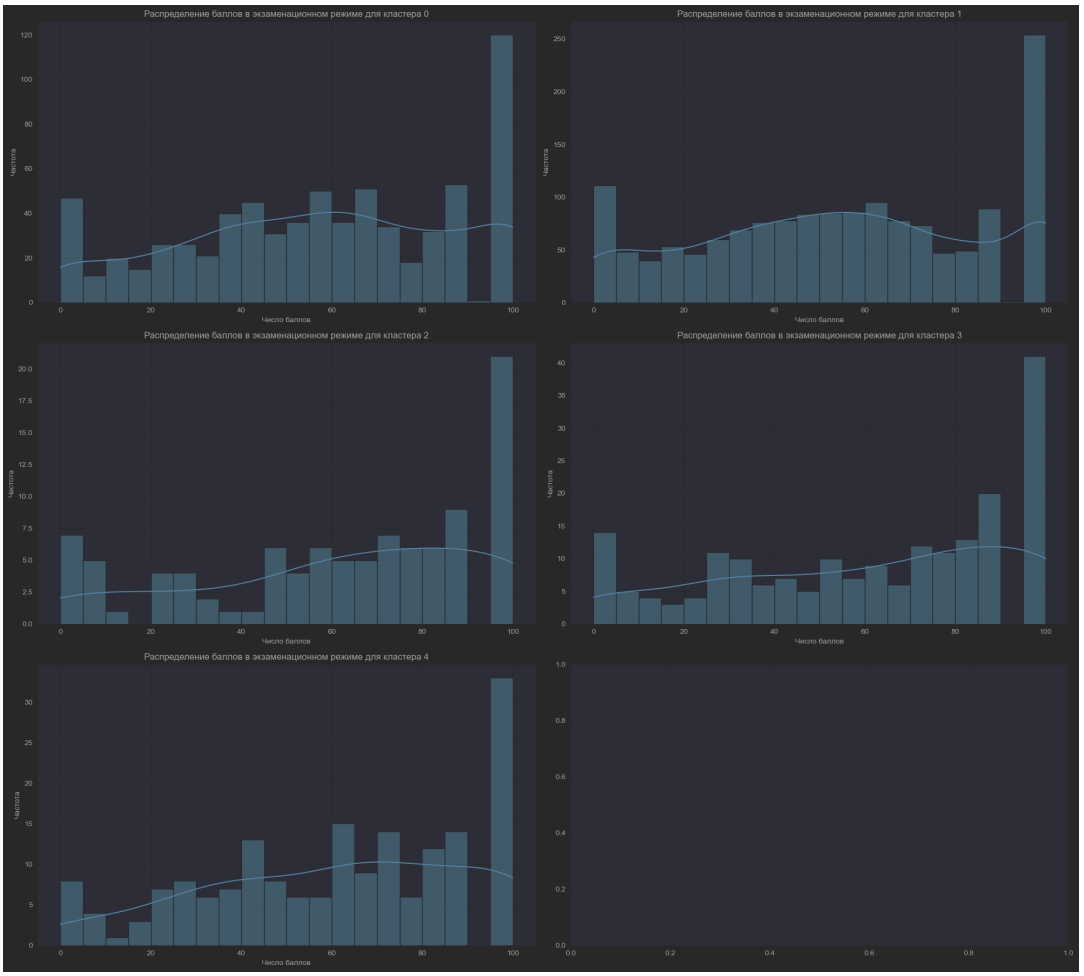


Рисунок 2.13 Столбчатый график распределения баллов в экзаменационном режиме для каждого кластера

Анализируя представленные графики, можно отметить, что кластеры 2 и 3 демонстрируют наиболее высокие результаты в экзаменационном режиме, в то время как кластеры 0 и 4 показывают чуть более скромные достижения, а кластер 1 - наименьшие. Это наблюдение соответствует первоначальным ожиданиям, основанным на анализе кластеризации для режима подготовки. Тем не менее, различия между кластерами в экзаменационном режиме являются относительно незначительными. Вычислим среднее число баллов для экзаменационного режима для всех кластеров.

	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
count	714.000000	1522.000000	100.000000	198.000000	180.000000
mean	57.231092	54.633377	62.140000	60.823232	61.066667
std	30.336238	31.087895	31.688197	32.100478	29.365468
min	0.000000	0.000000	0.000000	0.000000	0.000000
25%	36.000000	31.000000	44.750000	33.000000	40.000000
50%	58.000000	55.000000	67.000000	67.000000	64.000000
75%	84.000000	81.000000	86.000000	86.000000	85.000000
max	100.000000	100.000000	100.000000	100.000000	100.000000

Рисунок 2.14 Вычисленные значения среднего, стандартного отклонения и перцентелей баллов в экзаменационном режиме для каждого кластера

Из анализа полученных данных видно, что кластер 1 показывает заметно худшие результаты по сравнению с остальными группами, при этом результаты кластера 0 немного лучше. Кластеры 2, 3 и 4 демонстрируют примерно одинаковые показатели, причем различия между ними находятся в пределах статистической погрешности. Таким образом, можно заключить, что гипотеза о возможности предсказания результатов экзамена на основе анализа общего числа попыток и количества уникальных вопросов в тренировочном режиме не нашла своего подтверждения.

2.7 Распределение пользователей по отношению числа прохождений в режиме подготовки к числу прохождений в экзаменационном режиме

В контексте данного исследования следует подчеркнуть, что различия в количестве самостоятельных и контрольных работ в разных учебных группах приводят к ситуации, при которой общее число попыток ответов на вопросы в режиме подготовки у студентов различных групп коррелирует с числом проведенных самостоятельных работ. В таких условиях использование абсолютного количества попыток в тренировочном режиме как метрики оценки подготовленности может быть недостаточно информативным. Более релевантным подходом является применение отношения числа попыток в тренировочном режиме к числу попыток в экзаменационном режиме, что позволит более точно оценить уровень подготовки студентов. Построим

график, на оси X которого отложим отношение, а на оси Y число людей с таким отношением.

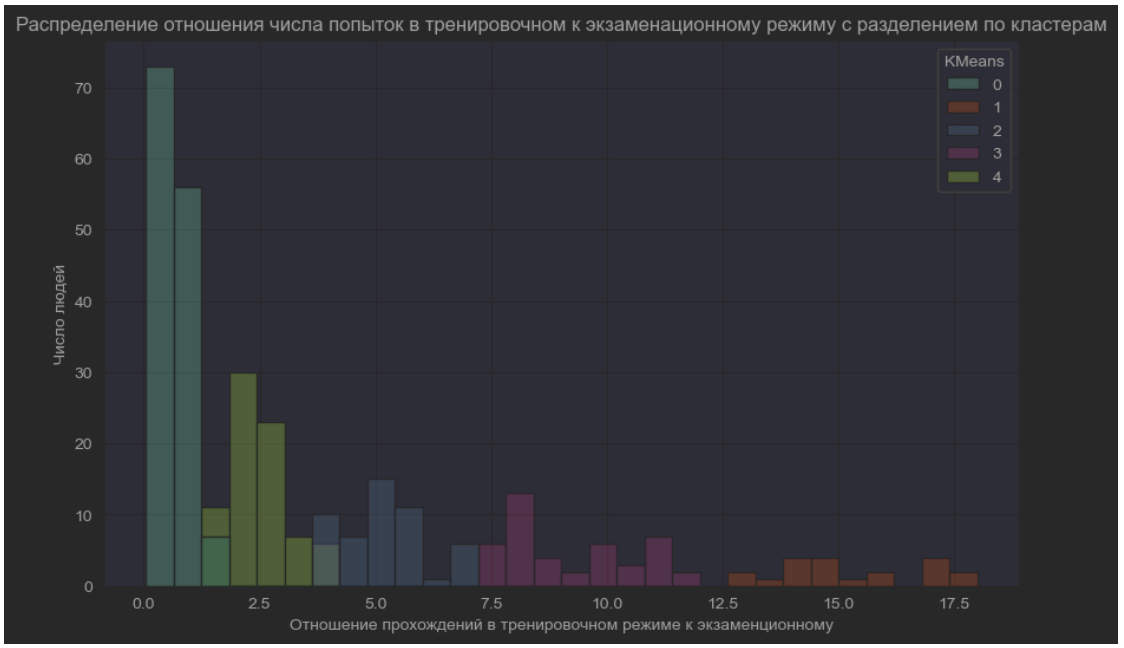


Рисунок 2.15 Столбчатый график распределения пользователей по коэффициенту отношения числа прохождений в тренировочном режиме к экзаменационному с разделением на кластеры методом KMeans

Основываясь на анализе данных, представленных на рисунке, можно выдвинуть гипотезу о том, что студенты, относящиеся к кластерам 4 и 0, скорее всего, покажут наименее удовлетворительные результаты на экзамене. В то же время предполагается, что студенты из кластеров 2 и 3 демонстрируют более высокий уровень подготовки и, соответственно, получат более хорошие оценки. Студенты из кластера 1, согласно этой гипотезе, могут ожидать наилучших результатов. Построим графики распределения баллов в экзаменационном режиме для каждого кластера.

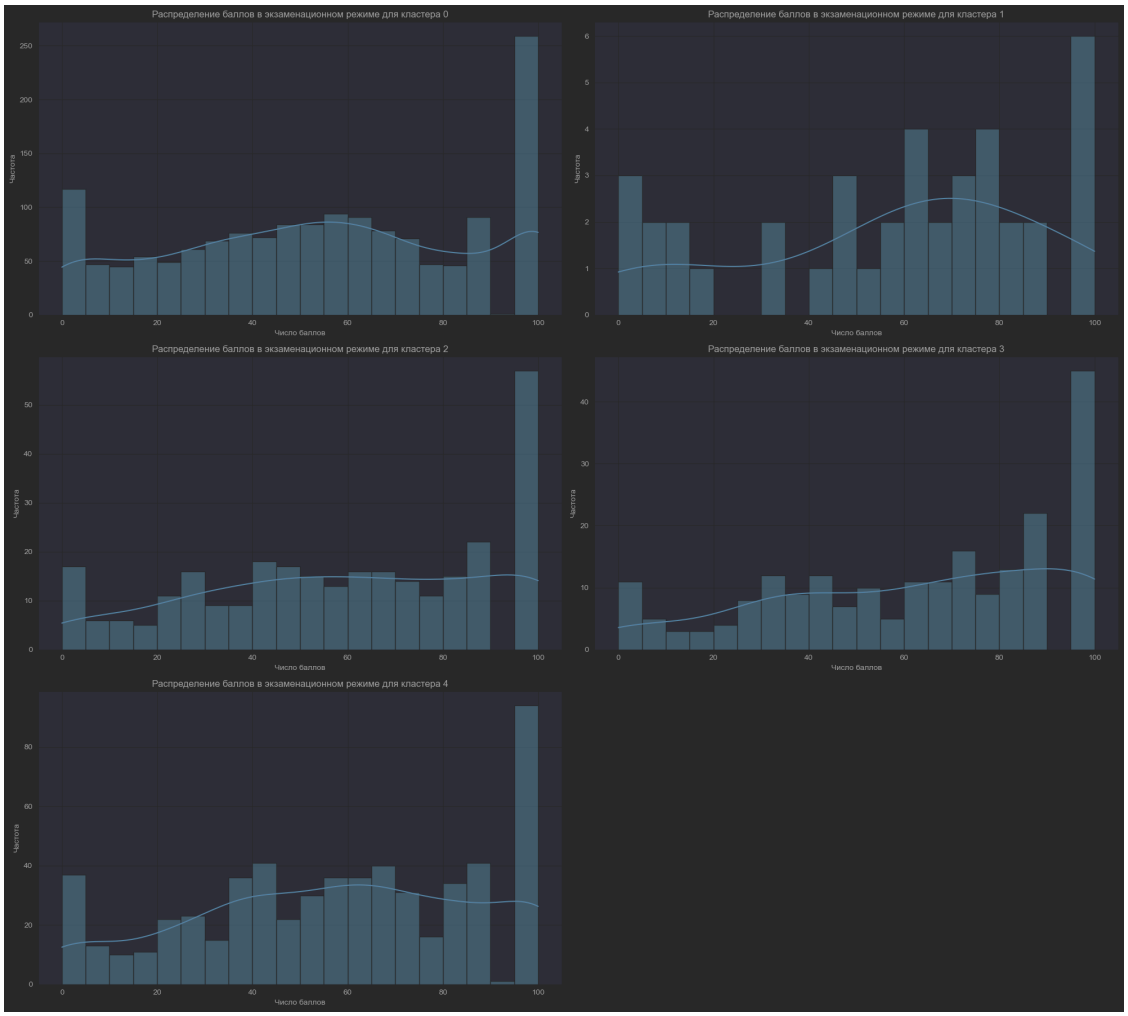


Рисунок 2.16 Столбчатый график распределения баллов в экзаменационном режиме для каждого кластера

Изучая представленные на рисунке данные, становится ясно, что определить наличие существенных различий в результатах между различными кластерами достаточно сложно. Для более глубокого понимания и анализа результатов, полезным шагом будет расчет среднего количества баллов, полученных студентами каждого кластера в экзаменационном режиме. Вычислим среднее число баллов для экзаменационного режима для всех кластеров.

	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
count	1536.000000	40.000000	293.000000	216.000000	589.000000
mean	54.376953	57.925000	59.167235	62.467593	57.483871
std	31.328633	31.303631	30.658459	30.519210	29.918628
min	0.000000	0.000000	0.000000	0.000000	0.000000
25%	30.000000	40.000000	36.000000	38.750000	36.000000
50%	55.000000	63.500000	61.000000	67.000000	59.000000
75%	81.000000	79.250000	85.000000	86.250000	84.000000
max	100.000000	100.000000	100.000000	100.000000	100.000000

Рисунок 2.17 Вычисленные значения баллов в экзаменационном режиме для каждого кластера

Анализ средних значений баллов по кластерам предоставляет более ясное представление о различиях в результатах между группами студентов. Согласно полученным данным, кластеры 3 и 2 продемонстрировали более высокие результаты по сравнению с другими группами, тогда как кластер 0 показал наименьшие показатели успеваемости. Однако результаты, полученные студентами из кластера 1, противоречат этой общей тенденции, нарушая ожидаемую закономерность.

В свете этих наблюдений можно сделать вывод, что гипотеза о том, что отношение количества попыток в режиме подготовки к количеству попыток в экзаменационном режиме является надежным предиктором результатов студентов на экзамене, не находит подтверждения.

2.8 Использование машинного обучения

Статистический анализ, проведенный в рамках данного исследования, не позволил достичь однозначных выводов о способности предсказать результаты экзаменов студентов. Это может быть обусловлено рядом факторов, включая отсутствие ключевых данных, таких как идентификатор экзамена или самостоятельной работы, идентификатор группы студентов, а также недостаточная дифференциация между заранее известными и неожиданными контрольными работами. В связи с этим, на основе имеющихся данных, предсказание оценок студентов на экзамене может иметь

высокую степень неопределенности, приближаясь к уровню случайных предположений.

Тем не менее, существует вероятность, что в данных могут присутствовать скрытые закономерности, неочевидные для прямого наблюдения, но доступные для выявления с помощью методов машинного обучения. В рамках данного исследования была применена модель регрессии на основе алгоритма случайного леса. Для анализа использовались данные о среднем значении, сумме, количестве и стандартном отклонении баллов в режиме подготовки, идентификаторе пользователя и количестве уникальных вопросов, с которыми студенты сталкивались. Целевой переменной для предсказания был выбран средний балл студентов за тесты в экзаменационном режиме.

```python
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import mean_squared_error
import numpy as np

# Агрегация тренировочных данных
training_data = df[df['is_useExamMode'] == False].groupby('authorized_user_id').agg({
    'calculated_statistic': ['mean', 'sum', 'count', 'std'],
    'question_id': ['count'],
}).reset_index()

# Преобразование MultiIndex в обычные столбцы
training_data.columns = ['authorized_user_id'] + ['train_' + '_'.join(col).strip() for col in training_data.columns.values[1:]]

# Агрегация экзаменационных данных
exam_data = df[df['is_useExamMode'] == True].groupby('authorized_user_id')['calculated_statistic'].mean().reset_index()
exam_data.columns = ['authorized_user_id', 'exam_average']

# Объединение данных
combined_data = pd.merge(training_data, exam_data, on='authorized_user_id')

# Предобработка данных
combined_data.fillna(0, inplace=True)

# Разделение данных
X = combined_data.iloc[:, 0:-1]  # Все столбцы, кроме средней оценки за экзамен
y = combined_data.iloc[:, -1]  # Средняя оценка за экзамен

# Разделение на обучающую и тестовую выборки
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=100)

# Создание и обучение модели
model = RandomForestRegressor()
model.fit(X_train, y_train)
```

Рисунок 2.18 Код для обучения модели искусственного интеллекта часть 1

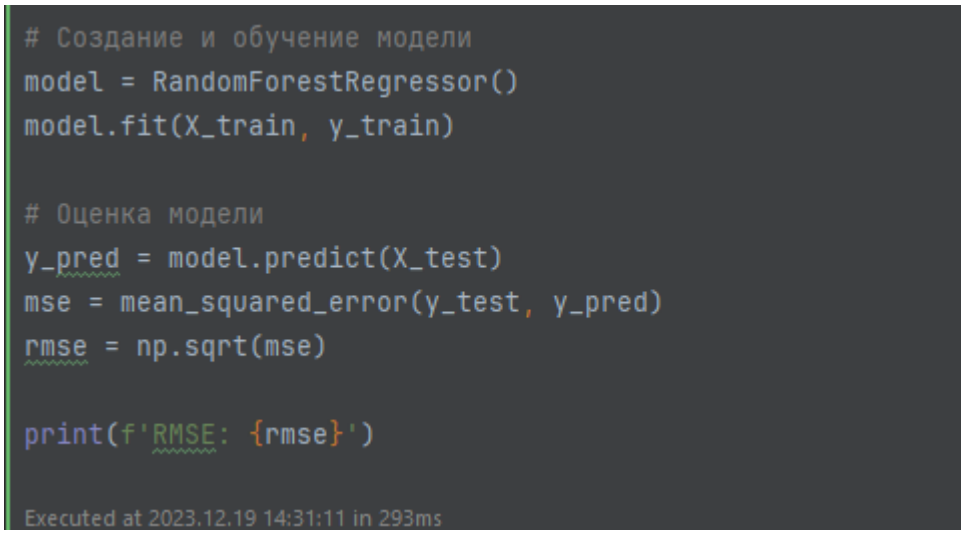


Рисунок 2.19 Код для обучения модели искусственного интеллекта часть 2

Полученные результаты, где среднеквадратичная ошибка (RSME) составила 15, на первый взгляд могут казаться достаточно обнадеживающими. Однако, если сравнить их с результатами, достигнутыми с помощью модели наивного прогноза, которая демонстрирует ошибку в 14 баллов, становится ясно, что применение машинного обучения в данном случае не привело к значительному улучшению предсказательной способности модели.

Это наблюдение подчеркивает важность выбора подходящего метода машинного обучения и оценки его эффективности в сравнении с более простыми базовыми моделями. Тот факт, что и другие примененные алгоритмы, включая XGBoost и различные виды регрессии, также не смогли дать удовлетворительных результатов, указывает на потенциальные сложности, связанные с качеством, структурой исходных данных или с самой природой задачи.

2.9 Поиск паттернов подготовки в случае частного экзамена.

Переориентация исследования на анализ подготовки студентов к конкретному экзамену, к которому они заранее готовились, представляется более целесообразной стратегией. Этот подход позволяет сравнивать результаты студентов в более однородных и контролируемых условиях, что способствует повышению объективности и достоверности выводов.

В рамках такого исследования подготовлены две выборки данных: одна содержит результаты экзаменационного тестирования по выбранному предмету, а другая – записи тренировочных попыток, выполненных студентами в течение месяца перед экзаменом. Такое разделение данных позволяет точнее анализировать и сопоставлять уровень подготовленности и успеваемость студентов.

Для выявления закономерностей и группировки студентов по уровню подготовленности применена кластеризация методом K-means. Построен график распределения студентов по среднему баллу на экзамене, что является релевантной метрикой, учитывая, что анализируется один экзамен. Наложение результатов кластеризации на данный график предоставляет возможность визуально оценить, как группы студентов, сформированные на основе их подготовленности, коррелируют с их итоговыми оценками на экзамене.

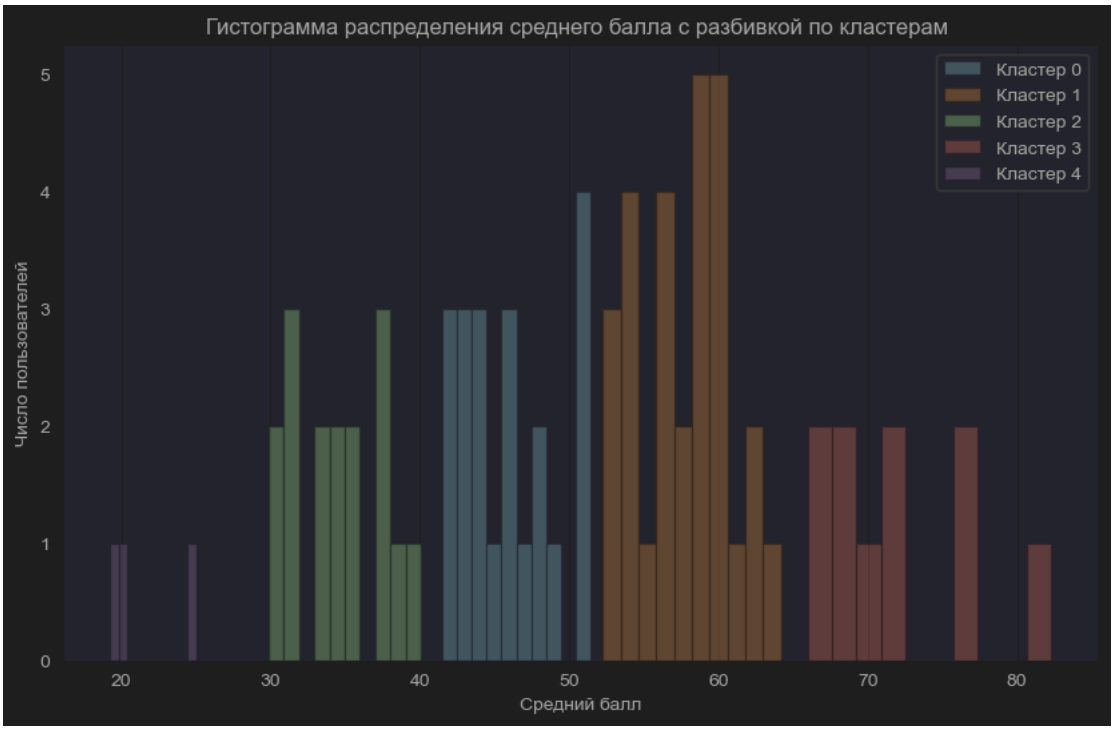


Рисунок 2.20 Столбчатый график распределения пользователей по среднему баллу в экзаменационном режиме в случае частного экзамена с наложением меток кластера методом KMeans

Анализируя полученные результаты с учетом проведенной кластеризации, можно сделать вывод о том, что наиболее высокие достижения демонстрируют студенты, отнесенные к 3-му кластеру, в то время как студенты 4-го кластера показали наименьшие результаты на экзамене.

Для дополнительной оценки уровня подготовки студентов различных кластеров был построен точечный график, отображающий взаимосвязь между средним баллом на экзамене и общим количеством попыток в тренировочном режиме с разделением по кластерам. Этот график позволяет оценить, насколько тщательно студенты каждого кластера подходили к процессу подготовки.

Такой подход дает возможность не только сравнить окончательные оценки студентов на экзамене, но и увидеть, как различные уровни подготовки влияют на итоговые результаты. Это позволяет получить более полное представление о стратегиях обучения и подготовки, применяемых студентами, и оценить их эффективность в контексте экзаменационных результатов.

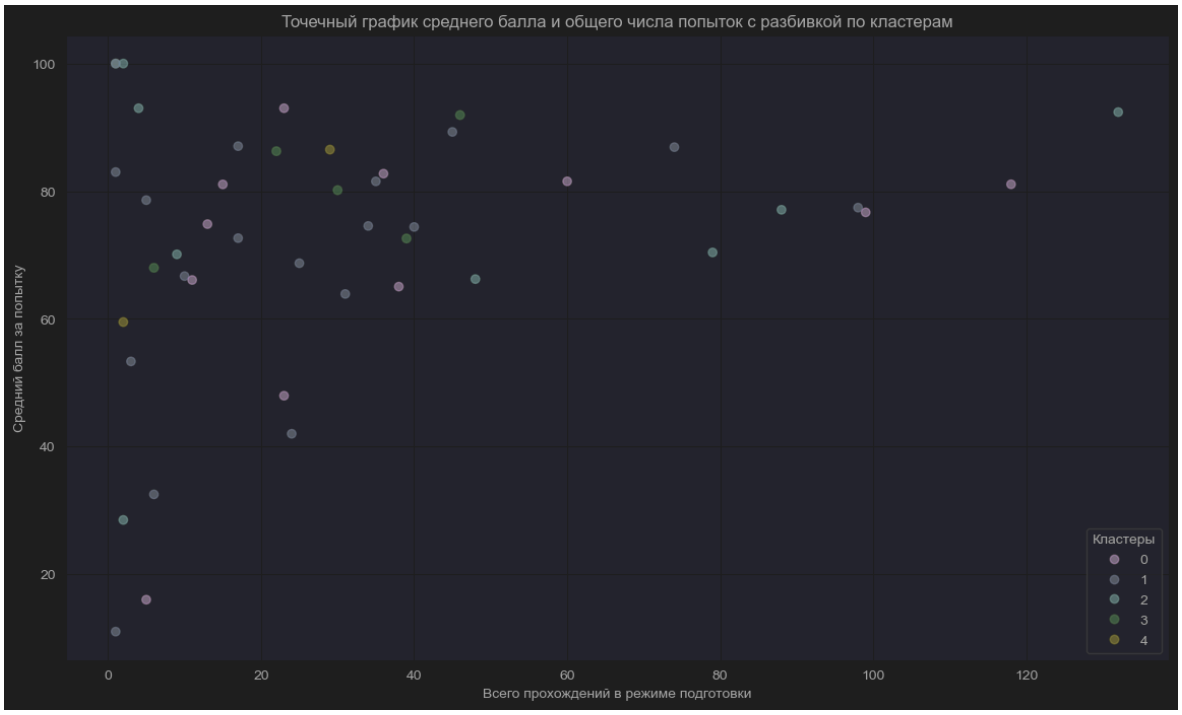


Рисунок 2.21 Точечный график распределения пользователей по критериям числа попыток в режиме подготовки и среднему баллу с наложением меток кластера из предыдущего шага

Из анализа представленного изображения следует, что студенты из кластеров 2 и 0 продемонстрировали наибольшую активность в режиме подготовки, выполняя больше всего попыток. В то же время, студенты, отнесенные к 3-му кластеру, которые показали наилучшие результаты на экзамене, в среднем затрачивали меньше времени на подготовку.

Для дальнейшего количественного анализа было выполнено определение медианного количества попыток прохождений вопросов для каждого кластера. Подсчет медианных значений позволяет получить более точное представление о характерной активности студентов в тренировочном режиме в каждой группе, минимизируя искажения, вызванные экстремальными значениями. Это может дать более глубокое понимание взаимосвязи между уровнем подготовленности и успехом на экзамене в различных кластерах.

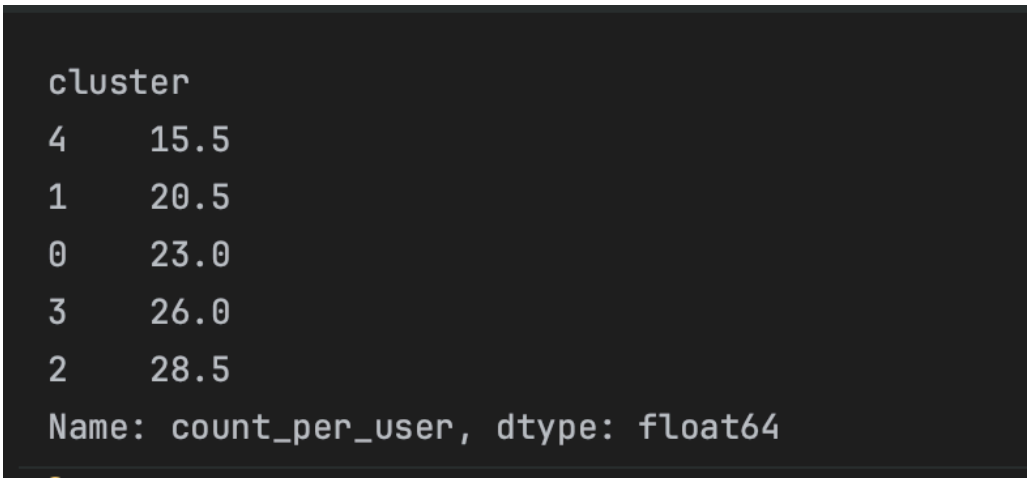


Рисунок 2.22 Медианное число попыток прохождения студентами вопросов в режиме подготовки с разделением по кластерам

На основе анализа изображения становится очевидным, что между оценкой, полученной студентами на экзамене, и медианным количеством попыток в режиме подготовки не наблюдается заметной корреляции. Это указывает на то, что количество попыток, предпринятых студентами при подготовке, не обязательно свидетельствует об их успеваемости на экзамене.

Отсутствие прямой взаимосвязи между этими двумя показателями опровергает первоначальную гипотезу о том, что более интенсивная подготовка (выраженная в количестве попыток) прямо коррелирует с лучшими результатами экзамена.

2.10 Выбор новой целевой метрики

Исходя из проведенного анализа, можно утверждать, что первоначальная методология оказалась некорректной по ряду причин. Во-первых, анализ основывался на сравнении баллов, полученных за разные вопросы, что предполагает одинаковую сложность всех вопросов, что, однако, не всегда соответствует действительности. Во-вторых, итоговая оценка на экзамене определяется не абсолютным значением суммы баллов, что могло бы быть аналогом среднего показателя, а позицией студента в общем распределении результатов по данному экзамену. Следовательно, более адекватной метрикой для оценки результатов студентов является их позиция в процентильном распределении по экзамену.

Для выполнения анализа в перцентелях от распределения по экзаменам был вычислен средний балл студента по каждому экзамену.

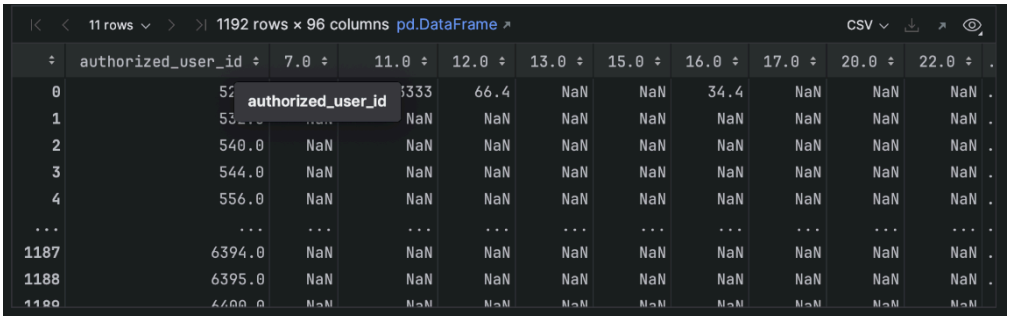


Рисунок 2.23 Средние баллы студентов по экзаменам

На представленном изображении заметно, что в ряде ячеек наблюдаются пропущенные значения, обусловленные тем фактом, что студенты сдавали различные экзамены. В дальнейшем, после обработки исходных данных, был проведен расчет перцентилей для каждого студента. Этот шаг позволил количественно оценить и сравнить уровень успеваемости студентов, учитывая их позиционирование в общем распределении результатов по всем экзаменам. Перцентильные значения представляют собой эффективный способ классификации результатов студентов, поскольку они отражают не только их абсолютные оценки, но и относительную успеваемость по сравнению с другими учащимися.

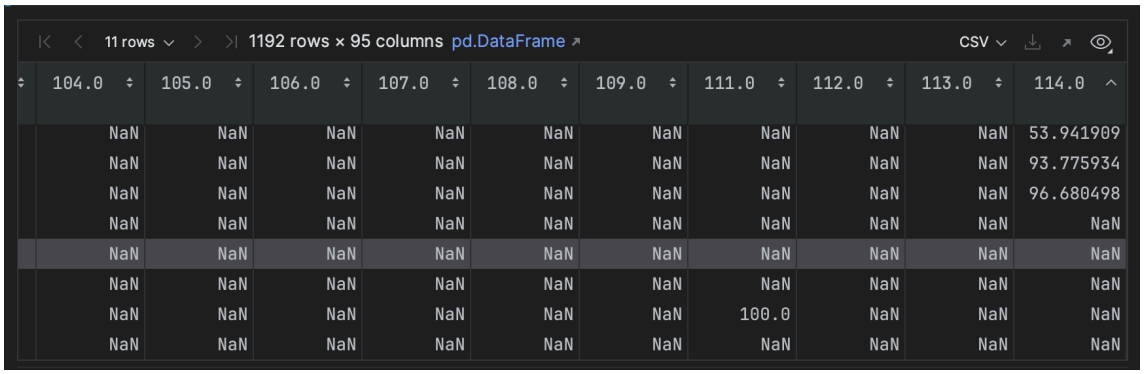


Рисунок 2.24 Перцентель в который попали студенты по каждому экзамену

В рамках данного исследования была сформулирована гипотеза о том, что перцентильное распределение оценок студентов является постоянным, то

есть каждый студент стабильно занимает схожие позиции в рейтинге своей группы. Это предположение, если оно подтвердится, позволит утверждать, что студенты демонстрируют стабильные результаты относительно других участников группы.

Для проверки этой гипотезы и понимания степени стабильности перцентильного распределения каждого студента использовалось среднеквадратичное отклонение. Важно отметить, что расчет данного показателя проводился индивидуально для каждого студента, что позволило более точно оценить их стабильность в получении оценок.

Кроме того, для дополнительного анализа была применена кластеризация, в результате которой студенты были разделены на 5 групп. Это позволило углубить понимание различий в уровне стабильности оценок среди разных категорий студентов и выявить потенциальные закономерности в их академической успеваемости.

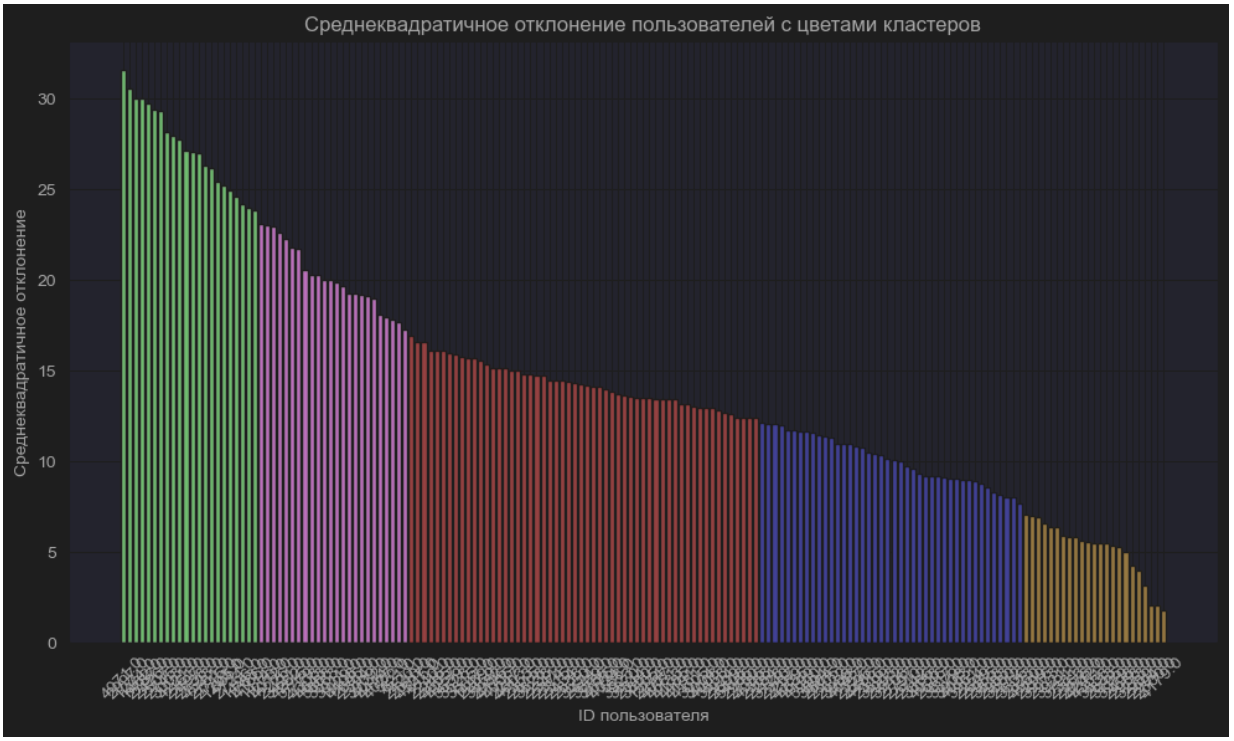


Рисунок 2.25 Столбчатый график распределения студентов по среднеквадратичному отклонению из перцентеля на экзамене

Из данных, представленных на графике, следует, что для большинства студентов среднеквадратичное отклонение их перцентильного ранга оказывается меньше 15. Это означает, что положение студентов в группе в среднем колеблется не более чем на одну шестую от их обычного места, указывая на постоянство их оценок в академическом контексте. В свете этого

можно утверждать, что наивный метод прогнозирования оценок, основанный на предположении о том, что студенты будут получать оценки, схожие со средними показателями их предыдущих результатов, оказывается эффективным.

2.11 Вывод по Главе 2.

В этой главе было проведено исследование взаимосвязи между учебной подготовкой студентов и их оценками на экзаменах. Анализ показал, что прямая корреляция между этими факторами не всегда очевидна, что указывает на сложность и многоаспектность процессов обучения. Особое внимание уделено анализу подготовки к конкретному экзамену и применению кластерного анализа, но предположение о возможности предсказания итогов экзаменов на основе данных о подготовке не подтвердилось.

ЗАКЛЮЧЕНИЕ

В данной курсовой работе проведен тщательный анализ взаимосвязи между академической подготовкой студентов и их успехами на экзаменах. Исследование охватило различные метрики, включая среднеквадратичное отклонение перцентильного ранга и средние баллы студентов, и показало, что значительная часть студентов демонстрирует стабильность в академических достижениях.

Тем не менее, важно отметить, что множество попыток установить прямую корреляцию между уровнем подготовки и итоговыми оценками на экзаменах не привели к однозначным результатам. Это подчеркивает многофакторность и сложность процессов обучения и оценки. Так, влияние таких аспектов, как количество и качество подготовки, не всегда может быть однозначно измерено и отделено от других факторов в рамках стандартных академических показателей.

Особое внимание в работе было уделено анализу подготовки студентов к конкретному экзамену, что позволило провести сравнение результатов в более однородных условиях. Были применены методы кластеризации и анализа различных стратегий подготовки, что пролило свет на разнообразие подходов студентов к учебному процессу. Однако, несмотря на эти усилия, гипотеза о возможности предсказания результатов экзамена на основе анализа подготовительной деятельности не нашла своего подтверждения.

Таким образом, данная работа демонстрирует сложность применения методов машинного обучения и анализа данных в образовательной сфере, подчеркивая необходимость более глубокого понимания многочисленных факторов, влияющих на академическую успеваемость студентов. Результаты данного исследования могут быть использованы для разработки более эффективных образовательных стратегий и подходов к оценке студенческих достижений, а также для дальнейших исследований в данной области.

СПИСОК ЛИТЕРАТУРЫ

1. Образовательная платформа Study Ways // Study Ways URL: https://sw-university.com/ (дата обращения: 10.01.2024).
2. Попов А. В. Тестирование как метод контроля качества знаний студентов // Труды Санкт-Петербургского государственного института культуры. - 2013. - №371. - С. 283.
3. Соломченко М. А. Проблемы оценки знаний студентов с помощью тестирования // Казанский педагогический журнал. - 2009. - №9. - С. 34.
4. Гуменникова Ю. В. Статистическая обработка результатов тестирования студентов // Вестник Самарского государственного технического университета. Серия: Психолого-педагогические науки. – 2015
5. Вялкова О. С., Ельцова Валентина, Юрьевна Ситникова, Светлана Юрьевна Проблемы тестирования студентов технического вуза // Высшее образование сегодня. – 2016
6. Трубилин А. И. Система оценки знаний и рейтингового тестирования студентов // Политематический сетевой электронный научный журнал Кубанского государственного аграрного университета. – 2016
7. Тусюк С.К., Белянская Е.С. Диагностическое Интернет - тестирование студентов - первокурсников в ТулГУ // Известия Тульского государственного университета. Технические науки. - 2012