МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
федеральное государственное бюджетное образовательное учреждение высшего образования
«Тольяттинский государственный университет»

Кафедра «Прикладная математика и информатика»
(наименование)

02.03.03 Математическое обеспечение и администрирование информационных систем
(код и наименование направления подготовки / специальности)

Мобильные и сетевые технологии
(направленность (профиль) / специализация)

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА (БАКАЛАВРСКАЯ РАБОТА)

на тему Исследование методов кластеризации и классификации данных для анализа поведения пользователей социальных сетей

Обучающийся В.Ю. Тульчинский
(Инициалы Фамилия) (личная подпись)

Руководитель к.п.н., доцент, О.В. Оськина
(ученая степень (при наличии), ученое звание (при наличии), Инициалы Фамилия)

Консультант И.Ю. Усатова
(ученая степень (при наличии), ученое звание (при наличии), Инициалы Фамилия)

Тольятти 2025

Аннотация

Тема выпускной квалификационной работы – «Исследование методов кластеризации и классификации данных для анализа поведения пользователей социальных сетей».

Ключевые слова: исследование методов, кластеризации данных, классификации данных, анализа поведения пользователей, социальные сети.

Объектом исследования бакалаврской работы являются социальные сети.

Предметом исследования бакалаврской работы являются методы кластеризации и классификации данных социальных сетей.

Цель бакалаврской работы – исследование методов кластеризации и классификации данных социальных сетей для анализа поведения их пользователей.

Методы исследования – методы интеллектуального анализа данных, методы модели анализа социальных сетей, методы и технологии разработки программного обеспечения.

Выпускная квалификационная работа состоит из введения, трех глав, заключения, списка используемой литературы и используемых источников.

Выпускная квалификационная работа включает 43 страницы текста, 16 рисунков, 4 таблицы и 26 источников.

Abstract

The topic of the final qualification work is «Research of methods of clustering and classification of data for the analysis of behavior of users of social networks».

Keywords: research of methods, data clustering, data classification, user behavior analysis, social networks.

The object of the bachelor's work research is social networks.

The subject of the bachelor's thesis is methods of clustering and classification of social network data.

The purpose of the bachelor's thesis is to study methods of clustering and classification of social network data to analyze the behavior of their users.

Research methods – methods of data mining, methods of social network analysis model, methods and technologies of software development.

The final qualifying work consists of an introduction, three chapters, a conclusion, and a list of references.

The final qualifying work consists of 43 pages of text, 16 figures, 4 tables and a list of used literature (26 sources).

Оглавление

Введение

Сегодня жизнь человека в основном наполнена цифровыми устройствами, такими как мобильные телефоны, ноутбуки, цифровое телевидение и т. д. В этом жизненно важную роль играет Интернет.

Социальные сети, такие как ВКонтакте, WhatsApp, Telegram и другие становятся независимыми. В этом нагруженном информацией мире, используя цифровые устройства, мы можем подключаться в любое время в любой точке мира. Технологический прогресс сделал возможным сбор данных с платформ социальных сетей с беспрецедентной скоростью и объемом.

Интеллектуальный анализ данных (Data mining) с помощью таких методов как кластеризация и классификация позволяет разработать гибкую, но интерпретируемую структуру моделирования для извлечения релевантных признаков поведения пользователя при публикации в социальных сетях с меньшими размерами на основе его активности при публикации с течением времени.

Извлеченные признаки затем можно использовать для анализа пользователей на основе их моделей вовлеченности и поведения.

Таким образом, исследование методов кластеризации и классификации данных для анализа поведения пользователей социальных сетей представляет научный и практический интерес.

Объектом исследования бакалаврской работы являются социальные сети.

Предметом исследования бакалаврской работы являются методы кластеризации и классификации данных социальных сетей.

Цель бакалаврской работы – исследование методов кластеризации и классификации данных социальных сетей для анализа поведения их пользователей.

Для достижения данной цели необходимо выполнить следующие задачи:

- выполнить постановку задачи исследования методов кластеризации и классификации данных для анализа поведения пользователей социальных сетей;
- проанализировать методы и алгоритмы кластеризации и классификации данных для анализа поведения пользователей социальных сетей;
- разработать и протестировать программу, реализующую алгоритмы кластеризации и классификации данных для анализа поведения пользователей социальных сетей.

«Методы исследования – методы интеллектуального анализа данных (Data mining), методы модели анализа социальных сетей, методы и технологии разработки программного обеспечения» [7].

Практическая значимость бакалаврской работы заключается в разработке программы, позволяющий производить анализ поведения пользователей социальных сетей.

Выпускная квалификационная работа состоит из введения, трех глав, заключения, списка используемой литературы и используемых источников.

Выпускная квалификационная работа включает 43 страницы текста, 16 рисунков, 4 таблицы и 26 источников.

Глава 1 Постановка задачи исследования методов кластеризации и классификации данных социальных сетей для анализа поведения их пользователей

1.1 Методы анализа социальных сетей

«Анализ социальных сетей (Social network analysis, SNA) – это подход к изучению отношений (связей) между отдельными лицами, группами, организациями или другими субъектами в рамках определенной сети с использованием сетей и теории графов» [8]. Анализ социальных сетей (АСС) помогает понять явления, возникающие в результате взаимодействия отдельных лиц или учреждений, и полезен для понимания результатов, зависящих от «социального капитала» или, в более общем плане, от формы и качества сотрудничества между субъектами. АСС по сути предоставляет набор репрезентативных методов для анализа социальных связей и предполагает, что социальные связи имеют значение, поскольку они влияют на поведение или передают информацию и ресурсы. По сравнению с другими методами анализа социальных явлений, начиная с микро- или мезоединиц, АСС подчеркивает важность структуры отношений, а не только их атрибутов и установок: общество рассматривается не как совокупность отдельных лиц, а как структура межличностных связей [20].

Субъектами, представленными сетью, могут быть отдельные лица, группы или организации. Отношения могут принимать множество значений в зависимости от темы и типа оцениваемой сети (например, коммуникационные связи, деловое сотрудничество, неформальные отношения, членские связи).

С помощью АСС сеть может быть представлена и проанализирована с использованием трех типов взаимосвязанных инструментов, которые могут использоваться как по отдельности, так и в сочетании [21]:

- сбор данных о взаимоотношениях (матрица);
- визуализация отношений (карты);

– оценка структуры сети (меры).

«Методы анализа социальных сетей (Social Network Analysis, SNA) представляют собой набор подходов и инструментов для изучения взаимодействий и отношений между участниками в социальных сетях. Эти методы позволяют выявлять структуру, динамику и особенности социальных связей» [3].

Основные методы анализа социальных сетей:

– моделирование сети**:** социальные сети представляются в виде графов, где узлы (вершины) представляют участников (людей, организации), а ребра (связи) — их взаимодействия (дружба, общение и т.д.);
– метрики центральности (степень центральности, посредническая центральность, гармоническая центральность и другие);
– анализ путей, например, исследование кратчайших путей между узлами для понимания, как информация или влияние распространяются в сети;
– динамика сети: анализ изменений в сети с течением времени, включая добавление или удаление узлов и связей;
– «методы машинного обучения: применение алгоритмов машинного обучения для предсказания связей, определения влияния узлов и анализа больших объемов данных.
– анализ больших данных: использование технологий обработки больших данных для анализа социальных сетей с учетом большого объема информации» [11].

Эти методы позволяют исследователям, маркетологам и аналитикам лучше понять социальные взаимодействия, выявить ключевых участников и оптимизировать стратегии взаимодействия в социальных сетях.

1.2 Методы анализа поведения пользователей социальных сетей

АСС является одним из самых перспективных направлений в аналитике и используется для отслеживания и анализа человеческого поведения, эмоций, социальных отношений, тенденций общения в Интернете. Социальные сети стали выдающейся новой технологией мониторинга человеческого поведения и эмоций в социальных сетях. Социальные сети являются неотъемлемой частью нашей повседневной жизни. Онлайновые социальные сети являются новой дисциплиной в изучении человеческого поведения и анализа эмоций. Анализ социальных сетей является мощным инструментом, который использовался для самых разных целей, чтобы получить ценную информацию о бизнесе и обществе [8].

Рассмотрим методы анализа поведения пользователей в социальных сетях.

Как показывает практика в последнее время для решения данной задачи широкое распространения получили методы интеллектуального анализа данных (Data mining) [7].

Это обусловлено появлением на ИТ-рынке гибких инструментов анализа, адаптированных для хорошей работы со сложными наборами данных, созданными в социальных сетях.

Рассмотрим метод анализа поведения человека, основанного на использовании методов извлечения ассоциативных правил [16].

Целью добычи ассоциативных правил является нахождение сильных ассоциативных правил из собранных массовых данных о поведении акторов и получение соответствующих знаний, содержащихся в данных, путем дальнейшего анализа.

Например, ассоциативный анализ данных о поведении студентов основан на архитектуре распределенного хранения и параллельных вычислений.

Для всех видов гетерогенных данных о поведении студентов из

нескольких источников алгоритм интеллектуального анализа правил ассоциации используется для извлечения элементов данных, которые соответствуют строгим правилам ассоциации, а значимые связи обнаруживаются посредством анализа.

Процесс анализа данных состоит из трех этапов: предварительная обработка данных, нахождение ассоциативных правил и получение соответствующих знаний (таблица 1) [25].

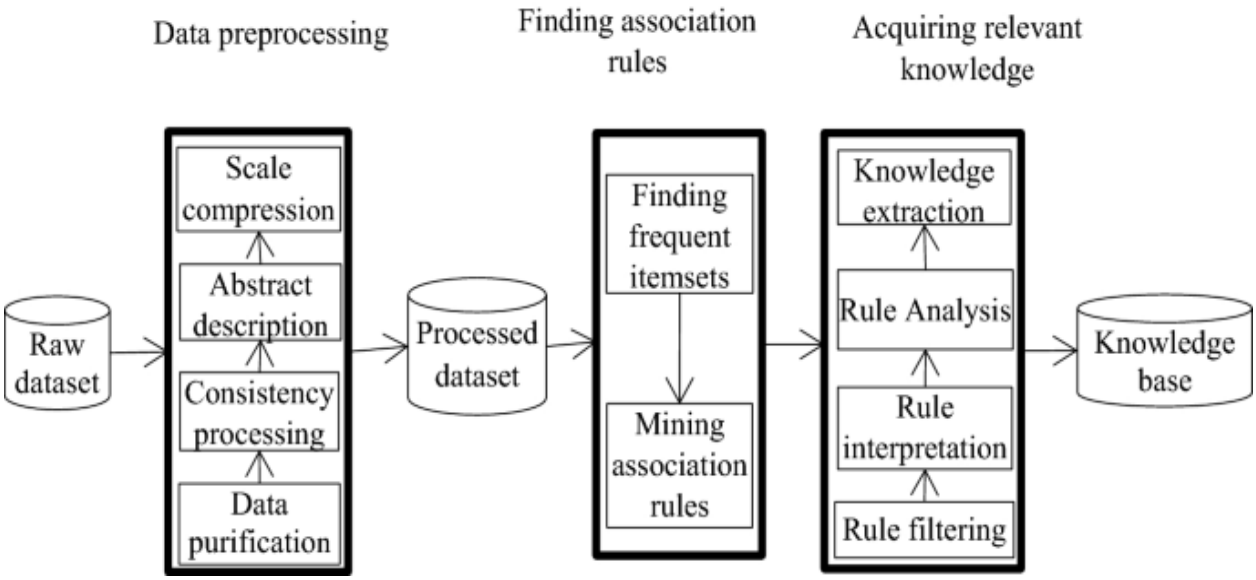

Рисунок 1 – Процесс извлечения ассоциативных правил из данных о поведении студентов

Для извлечения ассоциативных правил из данных о поведении студентов использован алгоритм Apriori [15].

Рассмотрим методы анализы поведения пользователей социальных сетей, основанные на применении метода кластеризации данных.

Определение характеристик акторов является важным первым шагом к пониманию процесса инноваций и экономического роста. В более общем плане исследователям в области социальных наук часто необходимо классифицировать данные об индивидуальном поведении в значимые группы, чтобы мы могли лучше описывать различия и сходства между людьми.

Для этого используется метод кластеризации данных социальных сетей [18].

Рассмотрим математическое описание основных методов кластеризации, используемых в анализе поведения в социальных сетях.

Начальной точкой многих кластерных исследований является $n \times p$ многомерная матрица X с n наблюдениями, каждое из которых описывается p различными характеристиками.

Для поведенческих наборов данных это можно интерпретировать как матрицу из n индивидов, где каждый индивид имеет p описательных переменных, таких как пол, возраст, выбор и т. д.

Для количественного измерения расстояния были предложены различные метрики расстояния между объектами из набора категориальных или непрерывных наблюдений. Категориальные данные обычно измеряются с точки зрения сходства, в то время как непрерывные данные обычно измеряется в несходстве (или расстоянии). Эти два типа мер в основном взаимозаменяемы, поскольку несут одинаковую информацию о расстоянии.

Когда характеристики каждого индивидуума измеряются как непрерывная переменная, расстояние между двумя индивидуумами i и j обычно количественно определяется индексом различия d_{ij}.

Предлагается множество мер различия, среди которых наиболее часто используемым является евклидово расстояние (1):

$$d_{ij} = \left[\sum_{k=1}^{p} \left(x_{ik} - x_{jk}\right)^2\right]^{\frac{1}{2}}, \quad (1)$$

где x_{ik} и x_{jk} являются, соответственно, k-м значением переменной p-мерных наблюдений для отдельных i и j.

В качестве альтернативы, расстояние городского квартала (расстояние Манхэттена) измеряет различие людей на прямолинейной конфигурации (2):

$$d_{ij} = \sum_{k=1}^{p} \left| x_{ik} - x_{jk} \right|. \quad (2)$$

Следует отметить, что, хотя метрики расстояния, упомянутые выше для категориальных данных, измеряют расстояние по сходству, а для непрерывных данных – по различию, в большинстве случаев эти две меры являются взаимозаменяемыми, используя следующую формулу (3):

$$d_{ij} = \sqrt{1 - s_{ij}}, \quad (3)$$

где s_{ij} – коэффициент Жаккара, определяемый по формуле (4) [4]:

$$s_{ij} = \frac{1}{p} \sum_{k=1}^{p} s_{ijk}. \quad (4)$$

Один из наиболее часто используемых критериев кластеризации объединяет $c_1(n,g)$ с индексом различия $h2(m)$ для представления общей суммы внутригруппового различия. Критерий также может быть показан эквивалентным внутригрупповому критерию суммы квадратов, полученному непосредственно из $n \times p$ многомерной матрицы X (5):

$$c_1^*(n,g) = \sum_{m=1}^{g} h_2(m) = \sum_{m=1}^{g} \sum_{l=1}^{n_m} (x_l^m - x^{\bar{m}})' (x_l^m - x^m). \quad (5)$$

где $h(m)$ – индекс различия, m – номер группы.

Таким образом, кластеризация в анализе поведения в социальных сетях – это процесс группировки узлов (участников) на основе их характеристик и взаимодействий. Основная цель кластеризации – выявление сообществ или

групп, которые имеют плотные внутренние связи и разреженные связи с другими группами.

Рассмотрим методы анализы поведения пользователей социальных сетей, основанные на применении метода классификации данных.

Классификация данных в анализе поведения в социальных сетях – это процесс, в ходе которого алгоритмы машинного обучения используются для предсказания меток (классов) на основе характеристик (признаков) участников или их взаимодействий. Ниже приведено математическое описание основных методов классификации, применяемых в этой области.

Данные о поведении пользователей в социальных сетях могут быть представлены в виде матрицы X (6):

$$X = \begin{bmatrix} x_1^T \\ x_2^T \\ \dots \\ x_n^T \end{bmatrix}, \quad (6)$$

где $x_i \in \mathrm{R}$ – вектор признаков для i-го пользователя, а n – количество пользователей. Признаки могут включать такие данные, как количество постов, лайков, комментариев, временные метки взаимодействий и т.д.

Метки классов могут быть представлены в виде вектора y (7):

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{bmatrix}, \quad (7)$$

где y_i – метка класса (например, позитивный или негативный пользователь, активный или пассивный и т.д.).

Рассмотрим методы классификации, используемые для анализа поведения пользователей.

Логистическая регрессия используется для бинарной классификации.

Модель описывается следующим образом (8):

$$P(y_i = 1|x_i) = \sigma(w^T x_i + b), \quad (8)$$

где $\sigma(z) = \frac{1}{1+e^{-z}}$ – логистическая функция, w – вектор весов, b – смещение.

Метод опорных векторов (SVM) ищет гиперплоскость, которая максимизирует разделяющую границу между классами.

Математически это можно выразить как (9):

$$maximize \frac{2}{\|w\|} subject\ to\ y_i(w^T x_i + b) \geq 1\ \forall_i, \quad (9)$$

где w – вектор весов, b – смещение, y_i – метка класса.

Деревья решений строятся путем последовательного разбиения данных на подмножества. Каждый узел дерева представляет собой условие на признак, а листья – метки классов. Алгоритм выбирает признак, который максимизирует уменьшение неопределенности (например, с использованием критерия Джини или энтропии) (10):

$$Gini(D) = 1 - \sum_{k=1}^{K} p_k^2 , \quad (10)$$

где p_k – доля экземпляров класса k во множестве D.

Случайный лес – это ансамблевый метод, который строит множество деревьев решений и объединяет их предсказания. Каждый узел дерева принимается на основе случайного подмножества признаков. Предсказание осуществляется по следующему принципу (11):

$$\hat{y} = mode(y_1, y_2, \dots, y_T). \quad (11)$$

Таким образом, рассмотренные методы и метрики позволяют

классифицировать участников социальных сетей на основе их поведения, выявляя паттерны и тенденции, что может быть полезно для маркетинга, анализа мнений и других областей.

Для сравнения методов Data mining для анализа поведения пользователей социальных сетей составлена таблица 1.

Таблица 1 – Характеристики методов анализа поведения пользователей социальных сетей

Метод	Преимущества	Недостатки
Анализ правил ассоциации	Выявление скрытых паттернов: может обнаруживать неожиданные связи между различными действиями пользователей, что может помочь в понимании их поведения. Простота интерпретации: правила ассоциации обычно представлены в понятной форме что облегчает интерпретацию результатов. Гибкость: может применяться к различным типам данных и могут быть адаптированы для анализа различных аспектов поведения пользователей. Не требуют предварительных предположений. Поддержка больших объемов данных.	Сложность в интерпретации больших наборов правил. Проблема избыточности: многие правила могут быть избыточными или неинформативными. Не учитывают временные аспекты. Необходимость в предварительной обработке данных. Ограниченная способность к предсказанию. Проблема с редкими событиями. Правила ассоциации могут не выявлять редкие, но важные паттерны, так как такие события могут не встречаться достаточно часто в данных.

Продолжение таблицы 1

Метод	Преимущества	Недостатки
Кластеризация	Помогает выявить группы пользователей с похожими интересами и поведением, что может быть полезно для таргетирования маркетинговых кампаний. Позволяет адаптировать контент под интересы различных сегментов пользователей, что повышает вовлеченность и удовлетворенность. Позволяет выявить эмоциональные реакции пользователей, что помогает в управлении репутацией и реагировании на негативные отзывы. Помогает находить пользователей, которые могут оказать значительное влияние на другие группы, что полезно для сотрудничества с ними. Позволяет исследовать сообщества и их динамику, что может быть полезно для понимания взаимодействий между пользователями.	Конфиденциальность данных. Предвзятость алгоритмов. Динамичность данных. Сложность интерпретации. Качество кластеризации напрямую зависит от качества и полноты исходных данных.

Продолжение таблицы 1

Метод	Преимущества	Недостатки
Классификация	Позволяет предсказывать поведение пользователей на основе их предыдущих действий, что может быть полезно для персонализации контента и рекомендаций. Помогает структурировать данные, разбивая их на категории, что упрощает анализ и интерпретацию результатов. Алгоритмы классификации могут автоматизировать процессы, такие как фильтрация спама, определение токсичных комментариев и сегментация пользователей. Поддержка принятия решений. Адаптивность. Многообразие методов.	Необходимость размеченных данных. Переобучение. Сложность интерпретации. Чувствительность к шуму. Необходимость в регулярном обновлении. Проблемы с балансом классов.

Как следует из таблицы, рассмотренные методы данных являются мощными инструментом для анализа поведения пользователей в социальных сетях, но их эффективность зависит от качества данных, выбора алгоритма и правильной настройки модели.

Выводы по главе 1

В соответствии с заданием выбираем в качестве методов для анализа поведения пользователей в социальных сетях методы кластеризации и классификации данных.

Глава 2 Обзор и анализ алгоритмов кластеризации и классификации данных для анализа поведения пользователей социальных сетей

Проанализируем алгоритмы кластеризации и классификации данных для анализа поведения пользователей социальных сетей.

2.1 Обзор и анализ алгоритмов кластеризации данных для анализа поведения пользователей социальных сетей

Рассмотрим характеристики алгоритма k-means.

Как указано в названии, алгоритмы k-means подчеркивают среднее значение кластеров.

Блок-схема классического алгоритма k-means показана на рисунке 2 [13].

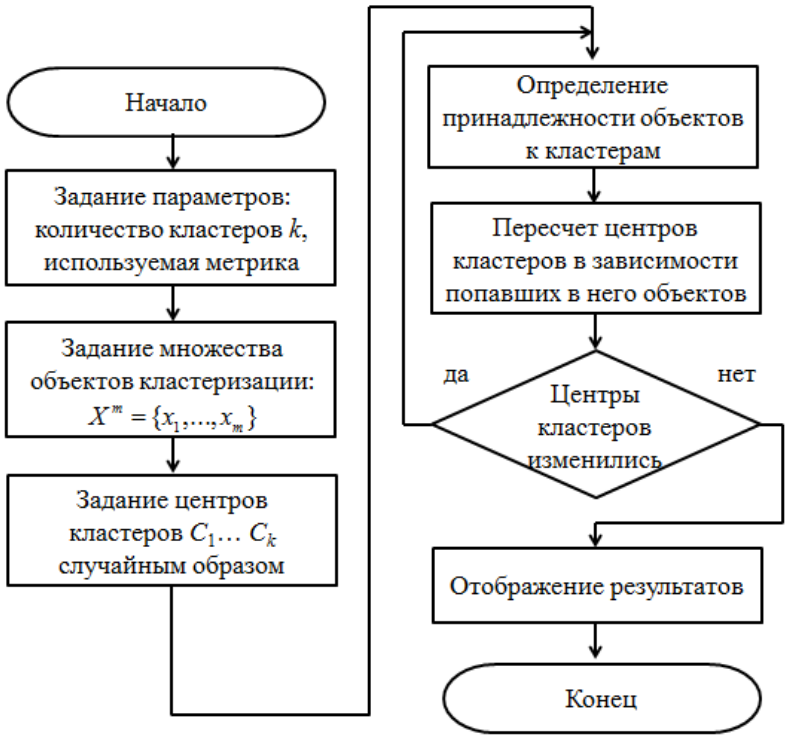

Рисунок 2 – Блок-схема алгоритма k-means

В общем, все алгоритмы k-means включают итеративные обновления кластеров путем одновременного перемещения объектов в кластер, среднее значение которого ближе всего, а затем пересчета среднего значения кластера.

В частности, все алгоритмы k-means содержат следующие шаги:

Шаг 1. g начальных скоплений определяются для каждого кластера p-мерным вектором, $\tilde{x}^m = (\tilde{x}_1^m, \tilde{x}_2^m, \dots, \tilde{x}_p^m$, где $\tilde{x}_k^m$ обозначает $\underline{k}$-ю характеристику начального элемента кластера m. Квадрат евклидова расстояния между i-м объектом и начальным скоплением кластера m просто вычисляется как (12):

$$d_{i\tilde{x}^m}^2 = \sum_{k=1}^{p} (x_{ik} - \tilde{x}^m)^2. \qquad (12)$$

Сравнивая результат уравнения (X) для объекта с каждым начальным объектом (их g), мы относим объект к кластеру, где результат минимизируется.

Шаг 2. После того, как все объекты были распределены по тому или иному кластеру, среднее значение кластера получается путем взятия среднего значения по всем объектам, которые попадают в каждый кластер. Это делается для каждого измерения p-характеристик (13):

$$\tilde{x}^m = \left(\tilde{x}_1^m, \tilde{x}_2^m, \dots, \tilde{x}_p^m\right). \qquad (13)$$

Вышеуказанное среднее значение кластеров $\tilde{x}^m$ затем может заменить начальные скопления $\tilde{x}^m$ и использоваться для вычисления квадрата расстояния между каждым объектом и каждым центроидом кластера, как в уравнении (X).

Объекты снова перемещаются в кластер, который дает наименьшую меру квадрата расстояния.

Шаг 3. Шаг 2 повторяется. Для каждого повторения старое среднее кластера заменяется вычисленным на основе последнего членства. Процесс повторяется до тех пор, пока ни один объект не изменит членство. Хотя все алгоритмы k-средних пытаются минимизировать внутригрупповую сумму квадратов отклонений от (группового) среднего, они могут отличаться друг от друга в деталях. В зависимости от конкретного используемого набора данных эти различия могут существенно влиять на результаты кластеризации. Здесь мы прослеживаем несколько важных различий этих самых популярных алгоритмов.

Рассмотрим характеристики алгоритма k-meadian [11].

В последние годы алгоритм k-meadian привлек все большее внимание.

Этот алгоритм перемещает объект в группу, медиана которой ближе всего к нему в соответствии с определенной мерой расстояния.

Численно конкретная процедура кластеризации выполняется так же, как и k-средние, за исключением того, что критерий кластеризации имеет вид (14):

$$c_2^*(n,g) = \sum_{m=1}^{g} \sum_{l=1}^{n_m} |(x_l^m - \tilde{x}^m)|\,, \qquad (14)$$

где $\tilde{x}^m$ относится к медианному вектору m-го кластера.

Первоначальная идея использования медианы вместо среднего значения заключается в том, чтобы уменьшить влияние выбросов.

Существуют также вариации алгоритма k-meadian с точки зрения того, как выбираются начальные скопления и как объекты меняются местами между кластерами.

Для выбора алгоритма кластеризации для решения задачи анализа поведения пользователей социальной сети разработана таблица 2.

Таблица 2 – Сравнение характеристик алгоритмов кластеризации

Алгоритм	Преимущества	Недостатки
k-means	«Простота: метод K-средних прост в понимании и реализации, что делает его доступным для новичков. Эффективность: он может быстро обрабатывать большие наборы данных, поскольку он эффективен с точки зрения вычислений по сравнению с методами иерархической кластеризации» [13]. Динамическая кластеризация: экземпляры могут менять кластеры при повторном вычислении центроидов, что обеспечивает гибкость в назначении кластеров. Более плотные кластеры: метод K-средних имеет тенденцию формировать более плотные кластеры по сравнению с иерархической кластеризацией, что может привести к более определенным группировкам.	Определение оптимального количества кластеров (K) может быть сложной задачей и часто требует знания предметной области или таких методов, как метод локтя. Чувствительность к инициализации: выходные данные алгоритма сильно зависят от начального размещения центроидов, что может привести к разным результатам при разных запусках. Чувствительность к порядку: Порядок точек данных может повлиять на конечный результат кластеризации, что может привести к непоследовательным результатам. Чувствительность к масштабированию: K-средние чувствительны к масштабу данных; нормализация или стандартизация могут значительно изменить результаты. Геометрические ограничения: k-means испытывает трудности с кластерами, имеющими сложную форму или разную плотность, поскольку предполагает сферические кластеры.

Продолжение таблицы 2

Алгоритм	Преимущества	Недостатки
k-median	Устойчивость к выбросам: менее чувствительна к выбросам. Интерпретируемость: может быть более интерпретируемым в определенных контекстах, особенно в асимметричных распределениях. Простота: относительно прост для понимания и реализации Динамическая кластеризация: позволяет динамически переназначать точки данных кластерам на основе медианы, что может привести к более точной кластеризации в определенных сценариях. Применимость к различным метрикам: можно адаптировать для работы с различными метриками расстояния, что делает ее универсальной для различных типов данных.	Сложность вычислений: может быть более вычислительно интенсивным, особенно для больших наборов данных. Как и в случае с K-means, определение оптимального количества кластеров может быть сложной задачей и часто требует дополнительных методов или эвристик. Проблемы сходимости: может сходиться к локальным оптимумам Чувствительность к инициализации: начальный выбор медиан может повлиять на конечный результат кластеризации, что приведет к разным результатам при разных запусках. Ограничен числовыми данными.

По результатам сравнения для кластеризации выбран алгоритм k-means.

2.2 Обзор и анализ алгоритмов классификации данных для анализа поведения пользователей социальных сетей

Рассмотрим характеристики алгоритма логистической регрессии.

«Логистическая регрессия – это контролируемый алгоритм машинного обучения, который выполняет задачи бинарной классификации, предсказывая вероятность результата, события или наблюдения. Модель выдает бинарный или дихотомический результат, ограниченный двумя возможными результатами: да/нет, 0/1 или истина/ложь.

Логическая регрессия анализирует связь между одной или несколькими независимыми переменными и классифицирует данные в дискретные классы. Она широко используется в предиктивном моделировании, где модель оценивает математическую вероятность того, принадлежит ли экземпляр определенной категории или нет» [26].

Логистическая функция представлена следующими формулами (15), (16):

$$Logit(pi) = \frac{1}{1 + \exp(-pi)}\ , \tag{15}$$

$$\ln\left(\frac{pi}{1 - pi}\right) = Beta_0 + Beta_1 * X_1 + \cdots Beta_k * K_k. \tag{16}$$

В этом уравнении логистической регрессии logit(pi) является зависимой или ответной переменной, а X – независимой переменной. Параметр Beta, или коэффициент, в этой модели обычно оценивается с помощью оценки максимального правдоподобия. Этот метод проверяет различные значения бета с помощью нескольких итераций для оптимизации наилучшего соответствия логарифмических шансов.

Рассмотрим характеристики алгоритма дерева решений.

«Дерево решений – это древовидная структура, похожая на блок-схему, где внутренний узел представляет признак (или атрибут), ветвь представляет правило принятия решения, а каждый конечный узел представляет результат.

Самый верхний узел в дереве решений известен как корневой узел. Он учится разделять на основе значения атрибута. Он разделяет дерево рекурсивным способом, называемым рекурсивным разделением. Эта структура, похожая на блок-схему, помогает вам в принятии решений. Это визуализация, похожая на блок-схему, которая легко имитирует мышление на

уровне человека. Поэтому деревья решений легко понять и интерпретировать (рисунок 3)» [22].

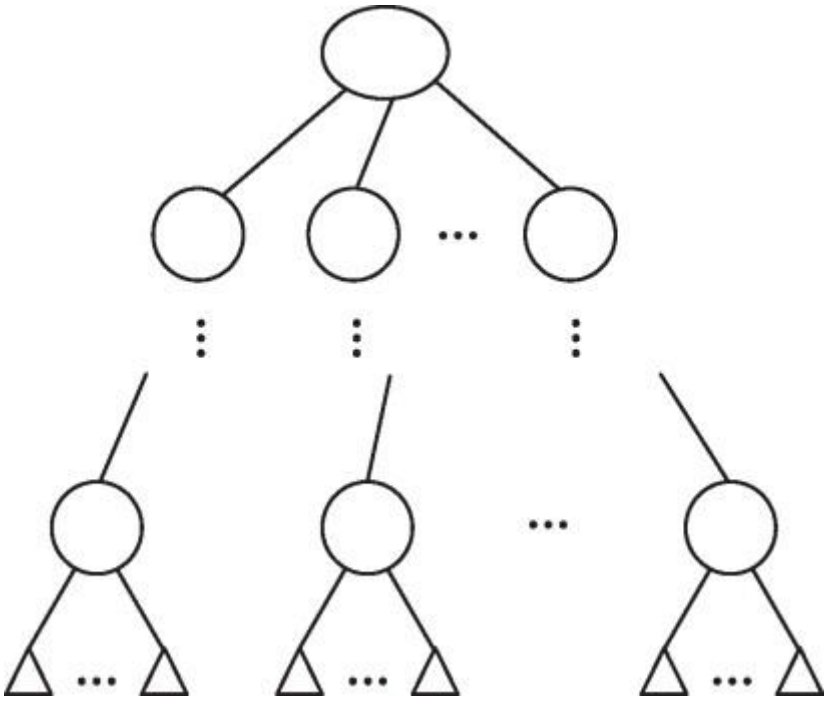

Рисунок 3 – Пример дерева решений

Для выбора алгоритма классификации для решения задачи анализа поведения пользователей социальной сети разработана таблица 3.

Таблица 3 – Сравнение характеристик алгоритмов классификации

Алгоритм	Преимущества	Недостатки
Логистическая регрессия	Простота и интерпретируемость. Быстрота обучения. Малое количество параметров. Модель предоставляет вероятностные оценки. Отсутствие предположений о распределении. Подходит для линейных зависимостей.	Линейная зависимость. Чувствительность к выбросам. Требование к большому объему данных. Проблемы с мультиколлинеарностью. Невозможность моделирования сложных зависимостей. Требование к независимости наблюдений.

Продолжение таблицы 3

Алгоритм	Преимущества	Недостатки
Дерево решений	Простота и интерпретируемость. Небольшие требования к предварительной обработке данных. Способность обрабатывать как числовые, так и категориальные данные. Выявление взаимодействий между переменными. Работа с большими объемами данных. Гибкость	Проблема переобучения. Нестабильность. Склонность к смещению. Проблемы с предсказанием. Ограниченная способность к экстраполяции. Неоптимальное разделение.

Принимая во внимание успешную практику использования представленных алгоритмов для решения задача классификации данных социальных сетей, выбираем алгоритм логистической регрессии.

Выводы по главе 2

Принимая во внимание успешную практику использования алгоритмов k-means и логистической регрессии, выбираем указанные алгоритмы для решения задач анализа поведения пользователей социальных сетей.

Глава 3 Реализация и тестирование программы анализа поведения пользователей социальной сети

3.1 Выбор средства разработки программы

Для реализации алгоритма поиска ассоциативных правил в большом наборе данных используем язык Python [2], [6].

Для выбора среды разработки сравним характеристики двух сред разработки: Jupyter Notebook и Replit.com.

Рассмотрим возможности среды Jupyter Notebook.

«Jupyter Notebook – это веб-приложение с открытым исходным кодом, которое позволяет пользователям создавать и обмениваться документами, содержащими живой код, уравнения, визуализации и повествовательный текст. Оно широко используется в науке о данных, машинном обучении, научных вычислениях и академических исследованиях благодаря своей способности сочетать выполнение кода с расширенным форматированием текста (рисунок 4)» [12].

Рисунок 4 – Главная страница среды Jupyter Notebook

Вот некоторые ключевые особенности и аспекты Jupyter Notebook:

- интерактивное кодирование: пользователи могут писать и выполнять код в реальном времени, что особенно полезно для тестирования и отладки;
- «поддержка нескольких языков: хотя Jupyter изначально поддерживал Python, теперь он поддерживает множество языков программирования с помощью «ядер». К ним относятся R, Julia, Scala и другие;
- поддержка форматированного текста: пользователи могут включать форматированный текст, изображения, ссылки, уравнения LaTeX и многое другое, что позволяет создавать исчерпывающую документацию вместе с кодом;
- визуализация данных: Jupyter Notebook может отображать визуализации непосредственно в документе с помощью таких библиотек, как Matplotlib, Seaborn и Plotly» [12];
- совместное использование блокнотов: блокнотами можно легко делиться с другими в различных форматах, включая HTML, PDF и Markdown, или их можно делиться через такие платформы, как GitHub или JupyterHub;
- расширения и настройка: пользователи могут улучшить функциональность Jupyter с помощью различных расширений и плагинов, что позволяет создавать настраиваемые рабочие процессы и функции;
- интеграция с библиотеками науки о данных: Jupyter обычно используется с популярными библиотеками науки о данных, такими как NumPy, Pandas и TensorFlow, что делает его основным продуктом в сообществе науки о данных.

Рассмотрим возможности среды Replit.com (рисунок 5) [17].

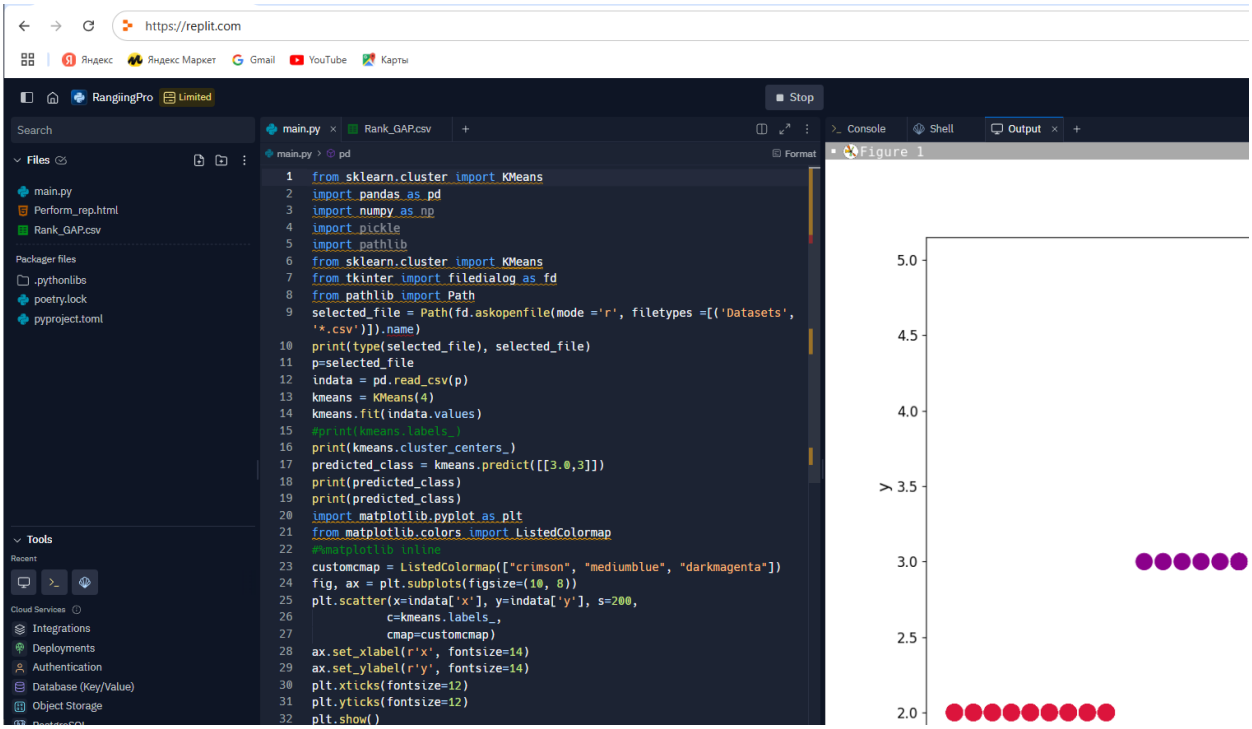

Рисунок 5 – Главная страница среды Replit.com

Replit.com – это интегрированная среда разработки (IDE), которая позволяет пользователям писать, запускать и совместно работать над кодом на более чем 50 языках программирования. Она предоставляет удобную облачную платформу для создания проектов, что делает ее доступной как для новичков, так и для опытных разработчиков.

Replit поддерживает совместную работу в реальном времени, что позволяет нескольким пользователям работать над одним проектом одновременно.

Основные особенности Replit:

– поддержка нескольких языков: Replit поддерживает широкий спектр языков программирования, включая Python, JavaScript, Java, HTML и CSS и многие другие. Эта универсальность делает его подходящим для различных проектов по кодированию;
– совместная работа в реальном времени: пользователи могут приглашать других людей для редактирования и совместной работы

над своими проектами в реальном времени, что улучшает возможности командной работы и обучения;

- «интегрированная среда разработки: платформа включает в себя мощную IDE с такими функциями, как подсветка синтаксиса, инструменты отладки и контроль версий, что позволяет пользователям эффективно писать и управлять своим кодом;
- мгновенное выполнение кода: пользователи могут мгновенно запускать свой код и видеть результаты в окне предварительного просмотра, что облегчает быструю разработку и тестирование» [17].

Replit имеет активное сообщество, где пользователи могут делиться своими проектами, искать помощь и учиться друг у друга. Он также предлагает шаблоны и руководства, которые помогут новичкам начать работу.

Для сравнения рассмотренных сред программирования используем таблицу 4.

Таблица 4 – Сравнение характеристик сред Jupyter Notebook и Replit.com

Характеристика	Jupyter Notebook	Replit.com
Интерфейс	Ориентирован на научные исследования, с возможностью добавления текста в формате Markdown, визуализаций и интерактивных графиков. Удобен для анализа данных и создания отчетов	«Предоставляет более традиционную среду разработки с редактором кода, терминалом и возможностью запуска приложений. Более ориентирован на разработку программного обеспечения и совместную работу» [17]
Установка и доступ	Обычно устанавливается локально на компьютере с использованием Anaconda или pip. Также доступен в облачных вариантах, таких как Google Colab.	Полностью облачная платформа, не требует установки. Доступен через веб-браузер, что позволяет легко начинать работу без необходимости настройки окружения

Продолжение таблицы 4

Характеристика	Jupyter Notebook	Replit.com
Совместная работа	Поддерживает совместную работу, но в основном через экспорт и обмен файлами (например, .ipynb). Для реального времени требуется использование сторонних решений, таких как JupyterHub или Google Colab	Предоставляет встроенные функции совместной работы, позволяя нескольким пользователям редактировать код одновременно в реальном времени. Это делает его идеальным для учебных целей и командных проектов
Цели использования	Идеален для анализа данных, машинного обучения, визуализации данных и научных исследований. Часто используется в образовательных целях для демонстрации концепций программирования и анализа	Более универсален и подходит для создания веб-приложений, игр, скриптов и других проектов. Хорошо подходит для обучения программированию и совместной разработки
Визуализация и графика	Поддерживает сложные визуализации с использованием библиотек, таких как Matplotlib, Seaborn и Plotly. Позволяет интегрировать графики непосредственно в документ	Основной фокус на коде и приложениях, хотя можно использовать библиотеки для графики, но возможности визуализации не так развиты, как в Jupyter

По результатам сравнения целей использования и с учетом предпочтений разработчика выбираем среду разработки Jupyter Notebook.

3.2 Реализация и тестирование алгоритма

Для построения диаграмм UML программы кластеризации и классификации данных для анализа поведения пользователей социальных сетей использован онлайн-сервис Visual Paradigm [24].

Для представления функционального аспекта программы построена диаграмма вариантов использования, показанная на рисунке 6 [23].

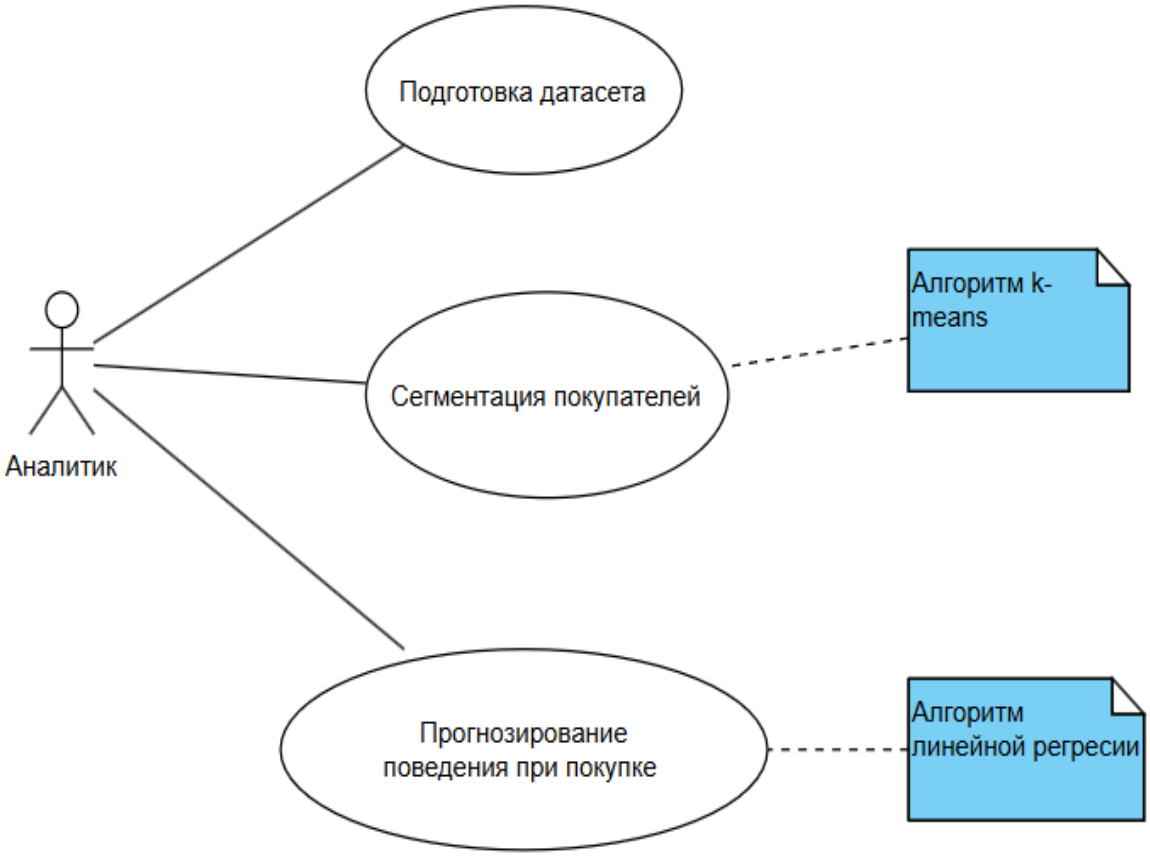

Рисунок 6 – Диаграмма вариантов использования программы

Рассмотрим использование метода кластеризации для сегментации пользователей социальной сети торгового центра [10].

Сегментация клиентов требует от компании сбора определенной информации/данных о клиентах, ее анализа и выявления закономерностей, которые можно использовать для создания сегментов.

Сегментация позволяет маркетологам лучше адаптировать свои маркетинговые усилия к различным подгруппам аудитории. Эти усилия могут относиться как к коммуникациям, так и к разработке продукта.

Сегментация клиентов производится с помощью алгоритма k-means.

В качестве входного датасета использован свободно распространяемый датасет Mall_Customers.csv [14].

На рисунке 7 показан код и результат загрузки и подготовки датасета.

```
import pandas as pd
customer_data = pd.read_csv("Mall_Customers.csv", index_col="CustomerID")
customer_data.info()
customer_data.head()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 200 entries, 1 to 200
Data columns (total 4 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Gender                  200 non-null    object
 1   Age                     200 non-null    int64
 2   Annual Income (k$)      200 non-null    int64
 3   Spending Score (1-100)  200 non-null    int64
dtypes: int64(3), object(1)
memory usage: 7.8+ KB
```

CustomerID	Gender	Age	Annual Income (k$)	Spending Score (1-100)
1	Male	19	15	39
2	Male	21	15	81
3	Female	20	16	6
4	Female	23	16	77
5	Female	31	17	40

Рисунок 7 – Код и результат загрузки и подготовки датасета

Для поддержки методов и алгоритмов машинного обучения используем библиотеку sickit-lean языка Python (рисунок 8).

```
from sklearn.cluster import KMeans
inertia_values = [
    get_clustering_inertia(k, customer_data.iloc[:, 1:]) for k in range(1, 11)
]
```

Рисунок 8 – Код импорта алгоритма k-means

В качестве метрики кластеризации использован коэффициент силуэта.

«Коэффициент силуэта является мерой того, как хорошо образцы кластеризируются с образцами, которые подобны себе. Кластеризирующиеся модели с высоким коэффициентом контура, являются плотными, где образцы

в том же кластере подобны друг другу и хорошо разделенные, где образцы в различных кластерах не очень похожи друг на друга» [1].

Функция silhouette_samples sklearn вычисляет коэффициенты силуэта для каждого образца данных и назначенного ему кластерного центроида. silhouette_score возвращает средний коэффициент силуэта для кластеризации с учетом данных и назначенных ему кластерных центроидов (рисунок 9).

```
from sklearn.metrics import silhouette_samples
from sklearn.metrics import silhouette_score
silhouette_avgs = []

for k in range(2, 11):
    kmeans = KMeans(
        n_clusters=k, max_iter=100, algorithm="lloyd", n_init=100, random_state=42
    ).fit(customer_data)
    cluster_labels = kmeans.predict(customer_data)

    _, ax = plt.subplots()
    si = plot_silhouettes(kmeans, customer_data, labels=cluster_labels, ax=ax)
    silhouette_avgs.append(si)

silhouette_avgs = np.array(silhouette_avgs)
```

Рисунок 9 – Код вычисления коэффициента силуэта

Проверив как среднее значение коэффициента силуэта, так и форму областей кластеров, можно убедиться, что модель для k=6 является наилучшей кластеризацией.

Этот вывод дополнительно подтверждается построением графика средних коэффициентов для каждой модели (рисунки 10 и 11).

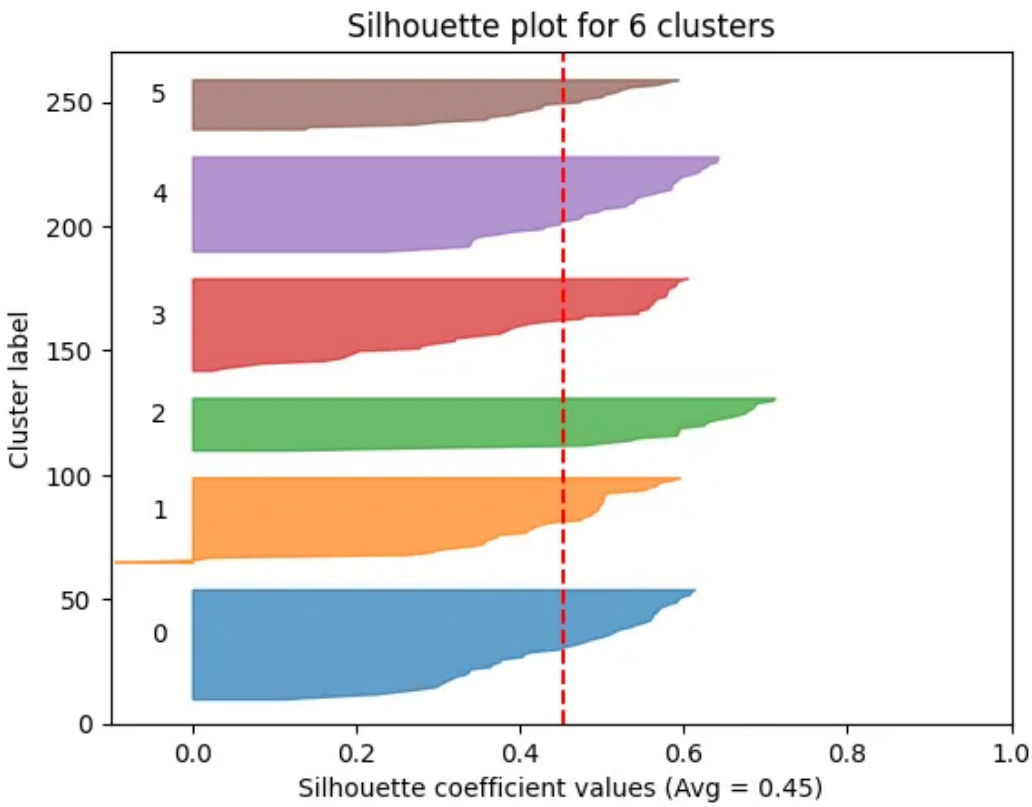

Рисунок 10 – График силуэта для 6 кластеров

```
_, ax = plt.subplots()
ax.plot(range(2, 11), silhouette_avgs, "o-")
ax.axvline(x=silhouette_avgs.argmax() + 2, color="lightgreen", linestyle="--")
ax.set_title("Оптимальное количество кластеров")
ax.set_xlabel("Количество кластеров ($ k $)")
ax.set_ylabel("Средняя ширина контура")
plt.show()
```

Рисунок 11 – Нахождение оптимального количества кластеров

На рисунке 12 показан график сегментация клиентов торгового центра с использованием кластеризации.

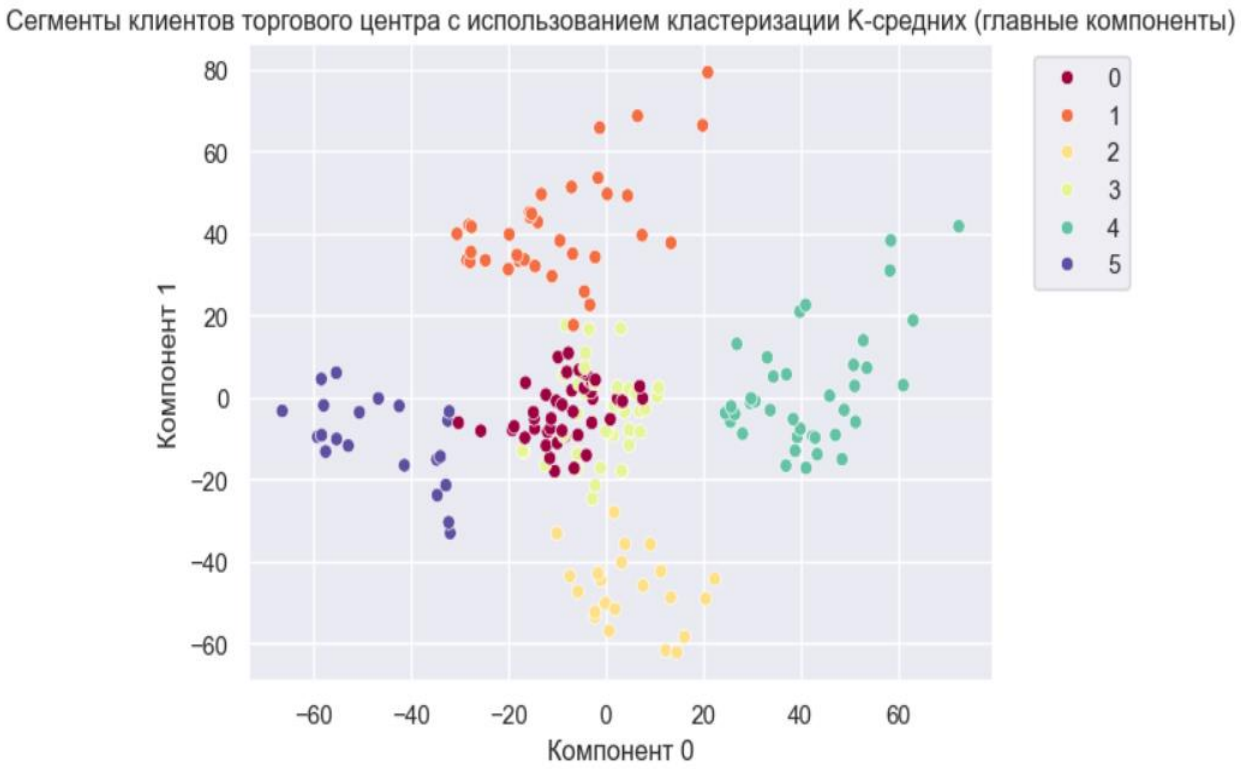

Рисунок 12 – График сегментация клиентов торгового центра

Для анализа поведения клиентов методом классификации данных использован датасет Social_Network_Ads.csv, который содержит информацию о людях и их реакции на определенную рекламную кампанию в социальных сетях [19].

Набор данных включает такие признаки, как «Возраст» (Age), «Предполагаемая зарплата» (EstimatedSalary) и «Куплено» (Purchased) (двоичная переменная, указывающая, совершил ли человек покупку после просмотра рекламы) (рисунок 13).

```
# Чтение данных
ads = pd.read_csv('social_network_ads.csv')
ads.head()
```

	Age	EstimatedSalary	Purchased
0	19	19000	0
1	35	20000	0
2	26	43000	0
3	27	57000	0
4	19	76000	0

Рисунок 13 – Структура датасета

Набор данных подходит для задач бинарной классификации, где целью может быть предсказание того, совершит ли человек покупку на основе возраста и предполагаемой зарплаты. Методы разведочного анализа данных (EDA) можно применять для понимания закономерностей и корреляций в наборе данных перед построением прогностических моделей [5].

Построены графики распределения покупок по группам (рисунки 14 и 15).

```
sns.FacetGrid(ads, hue='Purchased', height=3) \
    .map(sns.distplot, 'Age', bins=20, kde=True) \
    .add_legend()
plt.title("Распределение по возрасту")
sns.FacetGrid(ads, hue='Purchased', height=3) \
    .map(sns.distplot, 'EstimatedSalary', bins=20, kde=True) \
    .add_legend()
plt.title("Распределение по предполагаемой заработной платы")
plt.tight_layout()
plt.show()

```

Рисунок 14 – Код построения графиков распределения покупок

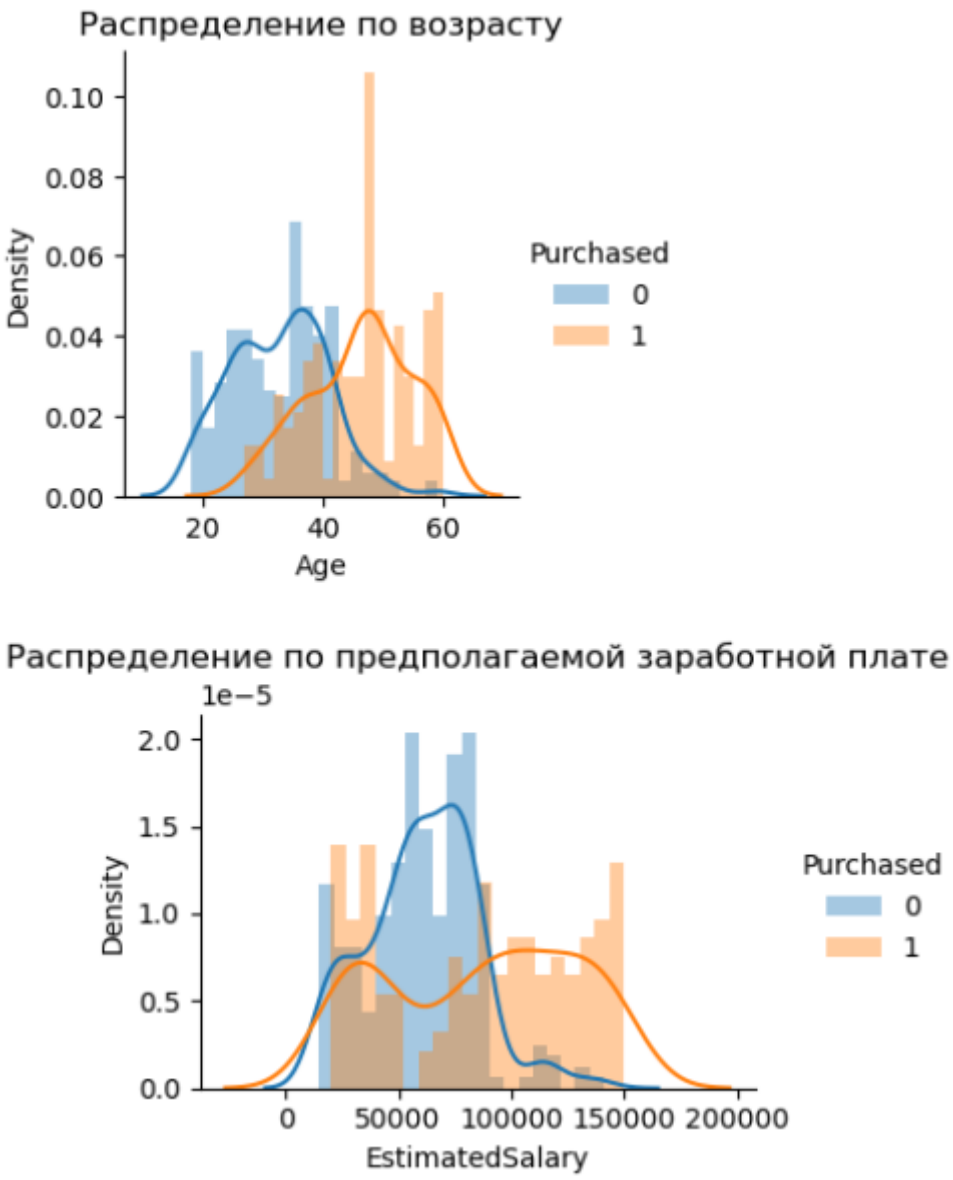

Рисунок 15 – Графики распределения покупок по группам

Как показано на графиках, возрастное распределение показывает довольно нормальный разброс с широким диапазоном возрастов, представленных в наборе данных. Пожилые люди демонстрируют более высокую вероятность совершения покупки. Люди с более высокими предполагаемыми зарплатами могут иметь более высокую склонность к совершению покупок, но эта связь требует дальнейшего анализа.

Стратегии таргетирования для рекламы в социальных сетях могут учитывать персонализированные подходы, основанные на возрастных и

предполагаемых сегментах заработной платы.

Для бинарной классификации использован алгоритм логистической регрессии, которая направлена на прогнозирование поведения при покупке («Куплено») [9] (рисунок 16).

```
Ввод [16]:
from sklearn.linear_model import LogisticRegression
logit = LogisticRegression(max_iter=500)

Ввод [17]:
logit.fit(X_train,y_train)

Out[17]:
LogisticRegression
LogisticRegression(max_iter=500)

Ввод [18]:
y_pred_logit = logit.predict(X_test)

Ввод [19]:
from sklearn.metrics import classification_report,confusion_matrix,accuracy_score

Ввод [20]:
# Матрица путаницы
cm = confusion_matrix(y_pred_logit,y_test)
cm

Out[20]:
array([[24,  4],
       [ 2, 10]], dtype=int64)

Ввод [21]:
cr = classification_report(y_pred_logit,y_test)
print(cr)

              precision    recall  f1-score   support

           0       0.92      0.86      0.89        28
           1       0.71      0.83      0.77        12

    accuracy                           0.85        40
   macro avg       0.82      0.85      0.83        40
weighted avg       0.86      0.85      0.85        40
```

Рисунок 16 – Код и результаты прогнозирования поведения при покупке

Модель достигла точности 85%, что подчеркивает ее эффективность в прогнозировании того, совершит ли клиент покупку на основе возраста и предполагаемой зарплаты.

Таким образом, тестирование разработанной программы подтвердило ее работоспособность.

Выводы по главе 3

Сегментация позволяет маркетологам лучше адаптировать свои маркетинговые усилия к различным подгруппам аудитории. Эти усилия могут относиться как к коммуникациям, так и к разработке продукта.

Сегментация клиентов производится с помощью алгоритма k-means.

Модель для k=6 является наилучшей кластеризацией.

Стратегии таргетирования для рекламы в социальных сетях могут учитывать персонализированные подходы, основанные на возрастных и предполагаемых сегментах заработной платы.

Для бинарной классификации использован алгоритм логистической регрессии, которая направлена на прогнозирование поведения при покупке.

Модель логистической регрессии достигла точности 85%, что подчеркивает ее эффективность в прогнозировании того, совершит ли клиент покупку на основе возраста и предполагаемой зарплаты.

Таким образом, тестирование разработанной программы подтвердило ее работоспособность.

Заключение

Бакалаврская работа посвящена актуальной проблеме исследования методов кластеризации и классификации данных для анализа поведения пользователей социальных сетей.

Выполненные в рамках бакалаврской работы задачи представлены следующими основными результатами:

- произведена постановка задачи исследования методов кластеризации и классификации данных для анализа поведения пользователей социальных сетей;
- дан обзор и произведен анализ алгоритмов кластеризации данных для анализа поведения пользователей социальных сетей. Принимая во внимание успешную практику использования алгоритмов k-means и логистической регрессии, выбираем указанные алгоритмы для решения задач анализа поведения пользователей социальных сетей;
- выполнены реализация и тестирование программы анализа поведения пользователей социальной сети. Метод кластеризации использован для сегментации клиентов торгового центра. Для бинарной классификации использован алгоритм логистической регрессии, которая направлена на прогнозирование поведения при покупке. Модель логистической регрессии достигла точности 85%, что подчеркивает ее эффективность в прогнозировании того, совершит ли клиент покупку на основе возраста и предполагаемой зарплаты.

Таким образом, тестирование разработанной программы подтвердило эффективность методов кластеризации и классификации данных для анализа поведения пользователей социальных сетей. Результаты бакалаврской работы могут представлять интерес для разработчиков и специалистов, занимающимся исследованием методов кластеризации и классификации данных для анализа поведения пользователей социальных сетей.

Список используемой литературы и используемых источников

1. Библиотека scikit-learn [Электронный ресурс]. URL: https://scikit-learn.ru/ (дата обращения: 10.12.2024).
2. Дроботун Н. В., Рудков Е. О., Баев Н. А. Алгоритмизация и программирование. Язык Python : учебное пособие. СПб.: СПбГУПТД, 2020. 119 с. URL: https://www.iprbookshop.ru/102400.html (дата обращения: 02.12.2024).
3. Замолоцких В. С., Сидоренко В.Г. Применение теории графов для анализа социальных сетей : учебное пособие. М. : МИИТ, 2020. 73 с. URL: https://www.iprbookshop.ru/115876.html (дата обращения: 06.12.2024).
4. Индекс Жаккара [Электронный ресурс]. URL: https://planetcalc.ru/9651/ (дата обращения: 06.12.2024).
5. Разведочный анализ (EDA) [Электронный ресурс]. URL: https://habr.com/ru/companies/otus/articles/752434/ (дата обращения: 10.12.2024).
6. Сузи Р. А. Язык программирования Python : учебное пособие. М.: ИНТУИТ, Ай Пи Ар Медиа, 2020. 350 с. URL: https://www.iprbookshop.ru/97589.html (дата обращения: 02.12.2024).
7. Чубукова И. А. Data Mining : учебное пособие. М.: ИНТУИТ, Ай Пи Ар Медиа, 2024. 469 с. URL: https://www.iprbookshop.ru/133907.html (дата обращения: 02.12.2024).
8. CfP: Role of Social Network Analysis in Human Behaviour and Emotion Analysis for Business Intelligence [Электронный ресурс]. URL: https://link.springer.com/journal/13278/updates/25868284 (дата обращения: 02.12.2024).
9. Customer Behavior Analysis for Social Media Ads [Электронный ресурс]. URL: https://www.kaggle.com/code/sakshisatre/customer-behavior-analysis-for-social-media-ads/notebook (дата обращения: 06.12.2024).
10. Gonçalves C. Customer Segmentation using K-Means clustering

[Электронный ресурс]. URL: https://medium.com/@camilolgon/customer-segmentation-using-k-means-clustering-9e5e11a3165a (дата обращения: 10.12.2024).

11. Jinhua Li et al. Robust K-Median and K-Means Clustering Algorithms for Incomplete Data [Электронный ресурс]. URL: https://onlinelibrary.wiley.com/doi/full/10.1155/2016/4321928 (дата обращения: 10.12.2024).

12. Jupyter Notebook [Электронный ресурс]. URL: https://jupyter.org/ (дата обращения: 02.12.2024).

13. K-means Clustering – Introduction [Электронный ресурс]. URL: https://www.geeksforgeeks.org/k-means-clustering-introduction/ (дата обращения: 10.12.2024).

14. Mall Customers [Электронный ресурс]. URL: https://www.kaggle.com/datasets/shwetabh123/mall-customers (дата обращения: 10.12.2024).

15. Rai A. An Overview of Association Rule Mining & its Applications [Электронный ресурс]. URL: https://www.upgrad.com/blog/association-rule-mining-an-overview-and-its-applications/ (дата обращения: 02.12.2024).

16. Raihan M. et al. Human Behavior Analysis using Association Rule Mining Techniques, 11th ICCCNT 2020 July 1-3, 2020 - IIT – Kharagpur.

17. Replit.com [Электронный ресурс]. URL: https://replit.com/~ (дата обращения: 02.12.2024).

18. Rong R. and Houser D. Exploring Network Behavior Using Cluster Analysis, Discussion Paper, Interdisciplinary Center for Economic Science, George mason University, 2014.

19. Social Network Ads [Электронный ресурс]. URL: https://www.kaggle.com/datasets/rakeshrau/social-network-ads (дата обращения: 10.12.2024).

20. Social network analysis (SNA) [Электронный ресурс]. URL: https://erc.undp.org/methods-center/methods/methodological-fundamentals-for-

evaluations/social-network-analysis (дата обращения: 02.12.2024).

21. Social Network Analysis [Электронный ресурс]. URL: https://www.publichealth.columbia.edu/research/population-health-methods/social-network-analysis (дата обращения: 02.12.2024).

22. The Decision Tree Algorithm [Электронный ресурс]. URL: https://www.datacamp.com/tutorial/decision-tree-classification-python (дата обращения: 10.12.2024).

23. Unified Modeling Language (UML) Diagrams [Электронный ресурс]. https://www.geeksforgeeks.org/unified-modeling-language-uml-introduction/ (дата обращения: 12.11.2002.12.202424).

24. Visual Paradigm [Электронный ресурс]. URL: https://online.visual-paradigm.com/ (дата обращения: 12.11.2024).

25. Wang T. et al. Student Behavior Data Analysis Based on Association Rule Mining. Int J Comput Intell Syst 15, 32 (2022). https://doi.org/10.1007/s44196-022-00087-4.

26. What Is Logistic Regression? [Электронный ресурс]. URL: https://www.ibm.com (дата обращения: 10.12.2024).