|  |  |
| --- | --- |
| Gerb-BMSTU_01 | **Министерство науки и высшего образования Российской Федерации**  **Федеральное государственное бюджетное образовательное учреждение**  **высшего образования**  **«Московский государственный технический университет**  **имени Н.Э. Баумана**  **(национальный исследовательский университет)»**  **(МГТУ им. Н.Э. Баумана)** |

ФАКУЛЬТЕТ ГОЛОВНОЙ УЧЕБНО-ИССЛЕДОВАТЕЛЬСКИЙ И МЕТОДИЧЕСКИЙ ЦЕНТР ПРОФЕССИОНАЛЬНОЙ РЕАБИЛИТАЦИИ ЛИЦ С ОГРАНИЧЕННЫМИ ВОЗМОЖНОСТЯМИ ЗДОРОВЬЯ (ИНВАЛИДОВ)

КАФЕДРА СИСТЕМЫ ОБРАБОТКИ ИНФОРМАЦИИ И УПРАВЛЕНИЯ

**РАСЧЕТНО-ПОЯСНИТЕЛЬНАЯ ЗАПИСКА**

***К НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ***

***НА ТЕМУ:***

**Аналитическая система на основе машинного обучения**

Студент группы ИУ5Ц-92Б **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** А.Н. Свинцов

(код группы) (подпись, дата) (инициалы и фамилия)

Научный руководитель **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** В.Ю. Строганов

(подпись, дата) (инициалы и фамилия)

Оценка \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Руководитель от кафедры **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** В.И. Терехов

(подпись, дата) (инициалы и фамилия)

Оценка \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Москва, 2023 г.

**Министерство науки и высшего образования Российской Федерации**

**Федеральное государственное бюджетное образовательное учреждение**

**высшего образования**

**«Московский государственный технический университет имени Н.Э. Баумана**

**(национальный исследовательский университет)»**

**(МГТУ им. Н.Э. Баумана)**

УТВЕРЖДАЮ

Заведующий кафедрой ИУ5

\_\_\_\_\_\_\_\_\_\_\_\_  В.И. Терехов

« \_\_\_\_ » \_\_\_\_\_\_\_\_\_\_ 2023 г.

**ЗАДАНИЕ**

**на выполнение научно-исследовательской работы**

по теме: Аналитическая система на основе машинного обучения

**Студент группы** ИУ5Ц-92Б Свинцов Артемий Николаевич   
 (Фамилия имя отчество)

**Направленность НИР** (учебная, исследовательская, практическая, производственная, др.) исследовательская

**Источник тематики** (кафедра, предприятие, НИР) учебная тематика

**График выполнения НИР**: 25% к 5 нед., 50% к 9 нед., 75% к 13 нед., 100% к 16 нед.

***Техническое задание*:**Исследование существующих методов и технологий аналитики данных; анализ средств аналитики с помощью машинного обучения; анализ библиотек языка Python для обработки и анализа данных.

***Оформление научно-исследовательской работы:***

Расчетно-пояснительная записка, минимальный объем 12 листов формата А4.

Приложения: графический (иллюстративный) материал (чертежи, схемы, диаграмма и т.п.)

Дата выдачи задания «15» октября 2023 г.

Научный руководитель **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** В.Ю. Строганов

(подпись, дата) (инициалы и фамилия)

Студент группы ИУ5Ц-92Б **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** А.Н. Свинцов

(код группы) (подпись, дата) (инициалы и фамилия)

Примечание: Задание оформляется в двух экземплярах: один выдается студенту, второй хранится на кафедре.

**Содержание**

[**Введение** 4](#_Toc154327326)

[**Анализ методов и технологий аналитического исследования** 5](#_Toc154327327)

[Методы аналитики данных 5](#_Toc154327328)

[Описательная аналитика 6](#_Toc154327329)

[Диагностическая аналитика 6](#_Toc154327330)

[Предикативная аналитика 6](#_Toc154327331)

[Предписывающая аналитика 7](#_Toc154327332)

[**Технологии аналитических исследований** 7](#_Toc154327333)

[Краудсорсинг 8](#_Toc154327334)

[Смешение и интеграция данных 8](#_Toc154327335)

[Имитационное моделирование 9](#_Toc154327336)

[Машинное обучение и нейронные сети 10](#_Toc154327337)

[**Задачи и инструменты машинного обучения** 11](#_Toc154327338)

[Классификаторы 11](#_Toc154327339)

[Регрессоры 11](#_Toc154327340)

[Кластеризация 12](#_Toc154327341)

[Библиотеки для машинного обучения 13](#_Toc154327342)

[**Вывод** 16](#_Toc154327343)

[**Список используемых источников** 17](#_Toc154327344)

# **Введение**

В современном мире проведение аналитической работы в организациях — вне зависимости от их формы собственности — является неотъемлемым компонентом для повышения эффективности и своевременности принимаемых управленческих решений, конкурентоспособности и нормального функционирования самих организаций.

Такая аналитическая работа активно осуществляется в финансовой сфере, и это вполне понятно: банки, фонды и венчурные компании стремятся автоматизировать управление активами и минимизировать риски. В корпоративном секторе востребованы системы бюджетирования, комплексы управления закупками и складскими запасами. Кроме того, предприятия начинают проявлять интерес к системам управления инвестициями и займами, инструментам моделирования и прогнозирования.

Сейчас существует масса методов проведения аналитической работы. Решение о выборе методологии основывается на исходной задачи исследования, технических требованиях постановки задачи, входных данных, которые могут быть использованы для построения эксперимента.

В научно-исследовательской работы будет рассмотрено применение методов машинного обучения в задачах аналитики. Также рассмотрим другие методы проведения исследования, их отличия между собой по целям и методам, область их практического применения.

# **Анализ методов и технологий аналитического исследования**

## Методы аналитики данных

Аналитическое исследование – это детальный анализ каких-либо данных, на основе которых потом применяется решение. Самый банальный пример: проанализировали продажи в интернет-магазине и пришли к выводу, что синих дрелей покупают больше, чем красных. Логично же, что вывод сам собой напрашивается, какого товара следует закупить больше для дальнейшей перепродажи.

На самом деле возможности аналитики куда шире, чем в примере выше, особенно в нынешних реалиях с огромным количеством больших данных. BigData вносит свои коррективы в устоявшиеся методы и технологии, и они не становятся хуже, наоборот, модернизируются и компьютеризируются. Сейчас уже редко кто проводит полевые исследования, достаточно сделать пару кликов мышкой, чтобы открыть сводную таблицу в специальной программе.

Принято выделять 4 вида аналитики больших данных, отличающихся уровнем сложности работы с информацией и степенью человеческого участия (рис. 1).



Рисунок 1 - Виды аналитики больших данных

### Описательная аналитика

При проведении описательного аналитического исследования данные собираются, систематизируются, а затем заносятся в таблицу, чтобы после можно было выделить основные параметры. Описательная аналитика необходима, чтобы получать достоверную информацию об основных характеристиках данных. Этот метод приносит определенную пользу, однако с его помощью не получится определить причину возникновения той или иной ситуации, также нет возможности спрогнозировать вектор развития текущих бизнес-процессов.

### Диагностическая аналитика

Диагностическая аналитика — это второй уровень аналитики данных. Диагностическая аналитика — это процесс изучения данных для выявления причин конкретных событий. Это влечет за собой использование более сложных статистических методов, таких как регрессионный анализ и проверка гипотез. Диагностическая аналитика часто используется для точного определения источника проблемы или понимания основных причин конкретного результата. На этом уровне ИИ и машинное обучение можно использовать для проведения сложных исследований и выявления закономерностей, которые людям может быть сложно заметить.

### Предикативная аналитика

Придется подготовить такое аналитическое исследование, если необходимо не просто описать данные и зависимости между переменными в виде показателей, имеющих ряд значений, но и спрогнозировать динамику показателей в будущем, используя информацию, полученную за прошлые периоды.

Несмотря на то, что связи между переменными используют, чтобы спрогнозировать ситуацию, не во всех случаях получится найти причинно-следственную связь. На самом деле, это не всегда требуется, чтобы составить точный прогноз.

Прежде всего нужно выявить связи между переменными. После этого можно оценить вероятность того, что произойдет какое-либо событие. К примеру, что реакция потенциальных клиентов на рекламную кампанию будет положительной и они захотят покупать продукцию.

### Предписывающая аналитика

Предписывающая аналитика, также известная как прескриптивная аналитика, является формой аналитического подхода, который использует данные и статистику полученную на предыдущих этапах аналитики для определения оптимальных действий или рекомендаций в сложных бизнес-ситуациях. Поэтому это более продвинутая стадия аналитического процесса по сравнению с дескриптивной (описательной) и прогностической (прогнозирующей) аналитикой.

Вместо простого анализа прошлых данных или прогнозирования будущих событий, предписывающая аналитика использует алгоритмы и модели, чтобы определить наилучшие решения и принять действия для достижения определенных целей. Она может использовать методы машинного обучения, искусственного интеллекта и оптимизации для определения оптимальных стратегий, которые могут помочь бизнесу принимать более осознанные решения и повышать свою эффективность.

# **Технологии аналитических исследований**

Для проведения аналитического исследования мало лишь собрать большие данные, необходимо правильно использовать их, к примеру, чтобы составлять прогнозы развития бизнеса либо для проверки маркетинговых гипотез. Для работы данные должны быть структурированы, после этого можно приступать к анализу. Рассмотрим различные методы и технологии BigData, их применение для различных задач аналитики.

### Краудсорсинг

Краудсорсинг – это практика использования знаний группы ради общей цели. Больше всего она полезна при решении сложных проблем инновационным образом или для упрощения сложных процессов.

Этот термин впервые был применён Джеффом Хоувом в статье 2006 года о подобной практике. Хотя краудсорсинг в той или иной форме существовал уже многие века, популярность его начала увеличиваться примерно с развитием электронной коммерции, социальных сетей и культуры смартфонов. Больше всего повлияла на рост интереса к этой практике увеличившаяся связность между людьми всего мира.

Для успешного краудсорсинга компании первым делом нужно разбить крупный проект на отдельные микрозадачи. Работники объединяются для решения этих микрозадач мелкими кусочками, что, по сути, ускоряет процесс.

Способ, которым компания собирает работников для выполнения этих задач, часто коррелирует с типом задач, которые нужно выполнить. Компания может использовать цифровое пространство, иногда называемое платформой краудсорсинга или сайтом микрозадач, для объединения этих работников в одном месте и поручения им микрозадач.

### Смешение и интеграция данных

При работе с полученной из нескольких источников разнородной информацией стоит собрать ее воедино для комплексного анализа. Автоматически загружать данные в базу нельзя, поскольку в источниках они могут быть записаны в различных форматах и с отличающимися параметрами. В таком случае выполняют интеграцию и смешение, в процессе которых разнородные данные сводят к единому формату.

Для интеграции можно использовать стандартный метод, основанный на ETL-процессе (извлечение данных из источника, преобразование к единому формату и очищение от избыточной информации, загрузка в хранилище). Экосистема Hadoop и БД NoSQL имеют свои инструменты для извлечения, преобразования и загрузки информации. После процедуры интеграции, происходит анализ и дальнейшая обработка материалов.

### Имитационное моделирование

В ряде случаев бывает необходимо узнать, как будут вести себя одни показатели, если изменить другие. К примеру, какие изменения продаж произойдут при повышении стоимости товара. Выполнять подобное тестирование в реальности не всегда возможно, поскольку это мероприятие достаточно затратное и может привести к финансовым потерям. В таких ситуациях аналитическое исследование рынка осуществляется с помощью имитационной модели.

Предположим, нужно узнать, какое влияние оказывают различные факторы на продажи торговой точки. Потребуются следующие данные: продажи, цены, количество покупателей. Используя эту информацию, получится создать модель магазина.

После этого можно внести необходимые изменения, например поднять либо снизить стоимость товара, изменить число консультантов, увеличить количество клиентов. От этих факторов зависят другие показатели. Проведя такой эксперимент, получится выбрать наиболее эффективные решения, чтобы затем реализовать их на практике.

Имитационное моделирование чем-то напоминает предиктивную аналитику. Однако в данном случае предугадывание будущих событий происходит по гипотетическим, а не реальным данным.

Имитационную модель получится создать без больших данных. Однако чем больше информации, тем более точной окажется модель, поскольку в ней будет учтено максимальное количество факторов.

### Машинное обучение и нейронные сети

Стандартные компьютеры могут считать, однако не всегда способны решать задачи, с которыми легко справляется человеческий мозг. К примеру, машина не способна понять, что «Дрель Bosch 10 мАч», «Дрель Бош 10», «Дрель Бошь акк 10» ничем не отличаются друг от друга.

Для того чтобы машина «думала», как человеческий мозг, необходимо создать в ней структуру, напоминающую мозг человека. Нейронные сети – именно такая структура. Они сделаны из огромного количества искусственный нейронов, при обучении которые создают связи, чтобы стало возможным анализировать данные.

Принцип работы нейронных сетей следующий: после получения информации они пропускают данные через сеть нейронов, чтобы выйти на результат. К примеру, отнести входные данные к какой-либо группе.

Для нормального функционирования нейросети необходимо обучить ее, данная процедура носит название «машинное обучение».

Для наглядности рассмотрим пример. Предположим, необходимо обучить нейросеть, чтобы она могла по изображению определять пол человека. Сделать это можно следующим образом:

- Строим нейросеть — программируем искусственные нейроны, чтобы они воспринимали входные данные, а затем создавали связи.

- Передаем нейросети очищенную выборку — базу лиц, которые отмечены как мужские или как женские. В результате нейросеть поймет, чем отличаются люди разного пола, и сможет самостоятельно определять по фотографии, кто на ней изображен: мужчина или женщина.

- Тестируем нейросеть — передаем ей новую очищенную выборку, однако не указываем, на каких изображения женские лица, а где мужские. После этого можно будет судить о том, насколько часто нейросеть допускает ошибки и приемлемо ли это для вас.

Нейросети подходят для сортировки информации, классификации данных на основе входной информации, после чего принимается определенное решение.

# **Задачи и инструменты машинного обучения**

### Классификаторы

Классификатор — это процесс, позволяющий сказать, к какой группе будет относиться тот или иной объект. Например, у кошек есть разные характеристики: длина хвоста, цвет шерсти, масса тела и другие параметры. По ним мы можем определить, к какой породе относится кошка. Если мы решаем эту задачу с помощью алгоритма, то этот алгоритм будет называться классификатором.

Алгоритм, часто применяемый для классификации — дерево принятия решений. Если мы хотим построить дерево условий для распределения котов по породам, на моменте обучения алгоритм строит дерево условий, задавая первый вопрос.

Этот алгоритм удобен с точки зрения бизнес-интерпретации результатов, так как мы не всегда можем сразу определить, по каким свойствам у нас разделились группы.

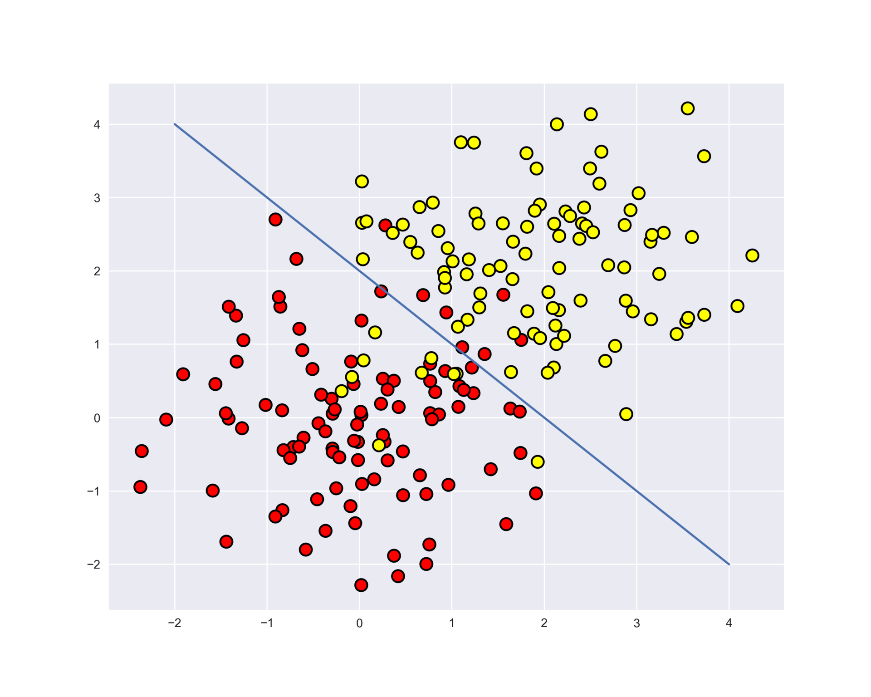


Рисунок 2 – Классификация

### Регрессоры

Регрессор – это алгоритм, который не относит предмет исследования к определенному классу, а присваивает ему определенное число. Пример — алгоритм кредитного скоринга: у нас есть возраст заемщика, трудовой стаж, зарплата — и требуется рассчитать, через какое время клиент сможет выплатить кредит.

Самый простой такой алгоритм — линейная регрессия. Представим себе, что наши объекты - это точки на плоскости (рис. 2). Наша задача — сделать так, чтобы прямая, которая будет проходить на плоскости, лежала как можно ближе ко всем точкам. Тем самым мы зададим линейные коэффициенты между входными данными и выходным значением. Подобный алгоритм прост и не требует особых затрат. Им удобно пользоваться, если у нас много признаков и мало объектов.

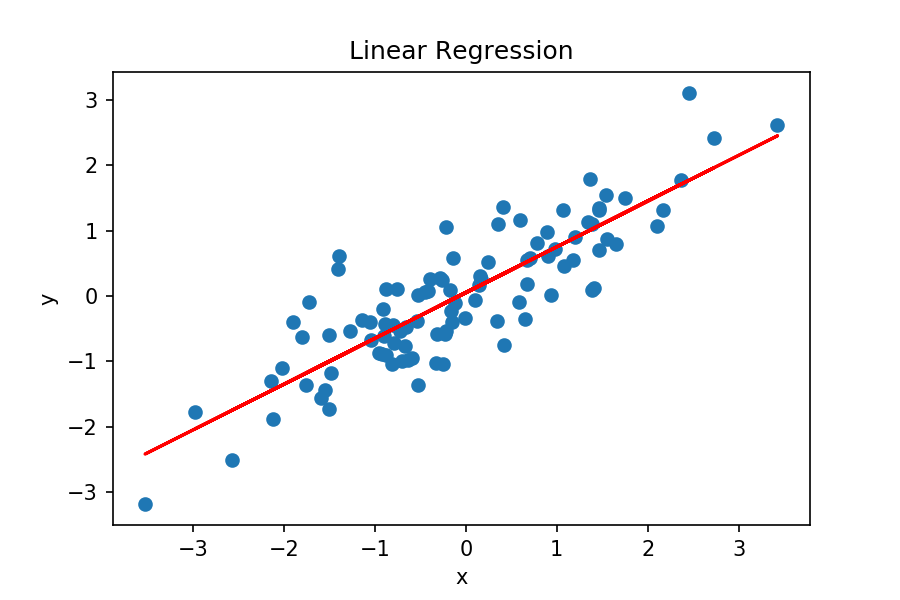


Рисунок 3 – Линейная регрессия

### Кластеризация

Кластеризация отвечает на вопросы о том, как разбить исследуемые объекты на группы и чем объекты внутри одной группы похожи.

Самый популярный алгоритм кластеризации – метод ближайших соседей. Допустим, необходимо разделить наш набор на 3 группы (рис. 3). Наши объекты – снова точки на плоскости. Мы выбираем случайным образом центры наших групп, затем смотрим расстояние от центра группы до точек, ближайших к этому центру группы. После мы смещаем центры таким образом, чтобы расстояние до точек своей группы оказывалось меньше, чем до точек другой группы. Через нескольких итераций у нас получатся хорошо разделенные группы.

Сложность этого алгоритма заключается в том, что объекты не всегда хорошо делятся на группы — в связи с этим трудно оценить корректность результата даже с помощью специальной оценки.

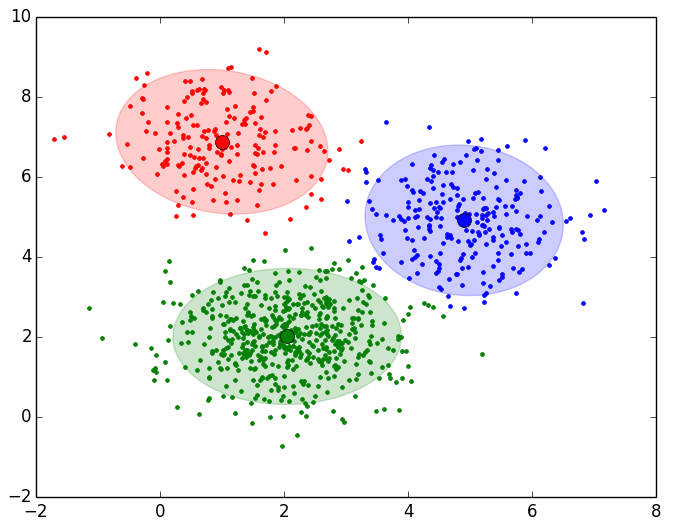


Рисунок 4 – Кластеризация

Помимо рассмотренных моделей, существует также алгоритмы нейронных сетей и обучения с подкреплением. Они являются более продвинутыми средствами машинного обучения и приближают модель к уровню искусственного интеллекта.

## Библиотеки для машинного обучения

Pandas – это библиотека, предоставляющая высокоуровневые структуры данных и инструменты для анализа данных на языке Python. Она предоставляет гибкие и эффективные структуры данных, такие как DataFrame, для манипуляции и анализа табличных данных. Pandas обладает мощными функциями для фильтрации, сортировки, группировки, агрегирования и визуализации данных, делая ее идеальным инструментом для работы с данными в области машинного обучения и анализа данных.

NumPy – это библиотека численных вычислений на языке Python. Она предоставляет многомерные массивы (ndarrays), которые являются эффективной структурой данных для хранения и манипуляции числовыми данными. NumPy обладает мощными функциями для выполнения операций линейной алгебры, математических операций, случайных генераций чисел и других вычислительных задач. Она является основной зависимостью многих других библиотек для научных вычислений и анализа данных на языке Python.

Pandas и NumPy часто используются вместе для обработки, анализа и предварительной подготовки данных перед применением моделей машинного обучения. Одна из преимуществ Pandas и NumPy заключается в их интеграции с другими библиотеками машинного обучения, такими как TensorFlow и Scikit-learn, что обеспечивает удобный и эффективный рабочий процесс в области анализа данных и машинного обучения.

TensorFlow – это открытая платформа машинного обучения, разработанная компанией Google. Она предоставляет инструменты для создания и обучения нейронных сетей, а также для выполнения других задач машинного обучения. TensorFlow обладает широкой функциональностью и поддерживает различные языки программирования, включая Python.

Keras – это высокоуровневая нейронная сетевая библиотека, написанная на языке Python. Она построена поверх TensorFlow и предоставляет простой и интуитивно понятный интерфейс для создания и обучения нейронных сетей. Keras позволяет быстро прототипировать и экспериментировать с различными моделями глубокого обучения.

Scikit-learn – это библиотека машинного обучения на языке Python, которая предоставляет множество алгоритмов и инструментов для решения задач классификации, регрессии, кластеризации и прогнозирования. Scikit-learn является одной из самых популярных библиотек для начинающих и профессионалов в области машинного обучения.

PyTorch – это фреймворк глубокого обучения, разработанный компанией Facebook. Он предоставляет инструменты для создания и обучения нейронных сетей. PyTorch отличается гибкостью и простотой использования, а также обладает хорошей поддержкой для операций на GPU. Его динамический граф вычислений делает его особенно подходящим для исследований и прототипирования новых моделей.

XGBoost – это библиотека градиентного бустинга, которая предоставляет реализации алгоритмов градиентного бустинга деревьев решений. Она отличается высокой производительностью и эффективностью, а также имеет множество возможностей для настройки моделей и обработки данных.

Seaborn – это библиотека визуализации данных на языке Python, основанная на matplotlib. Она предоставляет высокоуровневые интерфейсы и стилевые настройки, позволяющие легко создавать информативные и красивые графики. Seaborn позволяет исследовать и визуализировать данные, обнаруживать закономерности, отображать распределения, связи и шаблоны в данных.

Особенности Seaborn включают в себя широкий спектр графических объектов, таких как гистограммы, ящики с усами, точечные графики, тепловые карты, линейные графики, пространственные графики и многое другое. Библиотека также обладает мощными функциями для создания множественных графиков, комбинирования разных типов графиков и настройки цветовой палитры.

Seaborn предлагает простой и интуитивно понятный интерфейс, который упрощает создание визуализаций данных с минимальными усилиями. Библиотека также предоставляет стильные настройки по умолчанию, которые делают графики приятными визуально и удобными для понимания. Seaborn можно использовать в различных областях анализа данных, исследовательского анализа, построения моделей и презентации результатов.

Эти библиотеки можно считать основными для аналитической работы. Поскольку машинное обучения имеет широкий спектр задач, для каждой из них используются разные средства работы.

**Вывод**

Выбор метода исследования – важный этап в подготовке решения аналитических задач. В данной работе были рассмотрены различные методы аналитического исследования, а также отдельно изучены средства машинного обучения. Были выделены основные задачи, решаемые с помощью данного метода, а также рассмотрены базовые библиотеки языка Python для выполнения поставленных задач.

Работа позволила определить средства разработки, которые я применю в дальнейшей работе над этой темой.

# **Список используемых источников**

1. https://cyberleninka.ru/article/n/informatsionno-analiticheskaya-sistema-naznachenie-rol-svoystva/viewer - статья «Информационно-аналитическая система: назначение, роль, свойства».
2. <https://habr.com/ru/companies/pgk/articles/717328/> - статья «7 кругов аналитики».
3. <https://cyberleninka.ru/article/n/zadachi-funktsii-i-metodika-provedeniya-analiticheskih-issledovaniy/viewer> - статья «Задачи, функции и методика проведения аналитических исследований».
4. <https://surgay.ru/blog/analiticheskoe-issledovanie/> - статья «Аналитическое исследование: классические методы и современный подход»
5. <https://proglib.io/p/big-data-metody-i-tehniki-analiza-bolshih-dannyh-2021-08-31> - ProgLib / Путеводитель по BigData для начинающих: методы и техника анализа больших данных.
6. <https://github.com/ugapanyuk/courses_current/wiki/COURSE_TMO_SPRING_2023/> - учебный курс «Технологии машинного обучения» / МГТУ им. Н.Э. Баумана / 2023 г.
7. <https://cyberleninka.ru/article/n/tehnologiya-mnogoaspektnogo-analiticheskogo-issledovaniya-kak-metod-mashinnogo-obucheniya/viewer> - статья «Технологии многоаспектного аналитического исследования как метод машинного обучения».
8. <https://cyberleninka.ru/article/n/obzor-prikladnyh-osnov-ispolzovaniya-analitiki-dannyh-i-mashinnogo-obucheniya-v-prognozirovanii-sprosa/viewer> - статья «Обзор прикладных основ использования аналитики данных и машинного обучения в прогнозировании спроса»
9. Методические рекомендации по подготовке и защите выпускной квалификационной работы бакалавра. / Кротов Ю.Н. [Электронный ресурс] – URL: https://drive.google.com/file/d/1pEcfTr3xDdJ81Hxz2F6GcbtNV1n3dan6/view.