Муниципальное бюджетное образовательное учреждение

«Гимназия №1» города Липецка

Амвросов Георгий Дионисович

«Математика в Data Science»

Руководитель проекта

Токарева Инна Александровна

учитель математики

2023 год

Оглавление

|  |  |  |
| --- | --- | --- |
| 1. | Пояснительная записка | 2 стр. |
| 2. | Что такое Data Science | 4 стр. |
| 3. | Области применения Data Science | 5 стр. |
| 4. | Математика в Data Science | 6-14 стр. |
| 5. | Проект | 15-17 стр. |
| 6. | Выводы | 18 стр. |
| 7. | Список литературы | 19 стр. |

* + - 1. **Пояснительная записка**

Ещё с давних времен человечество пыталось упростить и автоматизировать все, что только возможно. И вот, спустя много столетий человечество изобрело компьютеры и Интернет, вследствие чего у нас началась эра информационных технологий, одно из достижений которого является Data Science.

В своей работе я постараюсь объяснить, почему Data Science на столько востребовано и какова роль математики во всем этом.

**Цель моей работы:** объяснить, почему Data Science сейчас очень востребован.

**Проектным продуктом** программа, угадывающая заболевания в грудной клетке по снимку.

**Ход выполнения работы**

**Работа над проектом велась поэтапно**

1. Изучение Data Science (сентябрь - октябрь)

2. Выяснение причин востребованности Data Science (ноябрь)

3. Изучение основ математики для Data Science (декабрь)

4. Прочтение литературы, связанной с Data Science (январь)

5. Изучение принципов работы нейронных сетей и машинного обучения (февраль)

6. Написание программы (март)

7. Предзащита проекта (март)

8. Защита проекта (апрель)

**Методы исследования:**

* изучение теоретического материала о Data Science,
* консультация с людьми, связанными с изучением Data Science.

**2. Что такое Data Science**

Каждый день человечество генерирует примерно 2,5 квинтиллиона байт различных данных. Они создаются буквально при каждом клике и пролистывании страницы, не говоря уже о просмотре видео и фотографий в онлайн-сервисах и соцсетях.

Термин Data Science вошел в обиход в середине 1970-х с подачи датского ученого-информатика Петера Наура. Data Science — междисциплинарная область на стыке статистики, математики, системного анализа и машинного обучения, которая охватывает все этапы работы с данными. Она предполагает исследование и анализ сверхбольших массивов информации и ориентирована в первую очередь на получение практических результатов.

При этом, человек, который занимается Data Science, называется Data Scientist. Data Scientist — это эксперт по аналитическим данным, который обладает техническими навыками для решения сложных задач, а также любопытством, которое помогает эти задачи ставить. Они частично математики, частично компьютерные ученые и частично трендспоттеры.

Таким образом еще в 2012 году позицию дата-сайентиста журналисты назвали самой привлекательной работой XXI века.

**3. Области применения Data Science**

По данным компании Kaggle, которая представляет собой профессиональную социальную сеть для специалистов описываемой нами области, сегодня data science аналитика используется бизнесом любого масштаба. IDC и Hitachi отмечают, что 78% предприятий подтверждают серьезный рост количества анализируемой и используемой информации за последнее время.

Перечислим отрасли, в которых Data Science активно используется для решения текущих задач:

* онлайн-торговля и развлекательные сервисы: рекомендательные системы для пользователей;
* здравоохранение: прогнозирование заболеваний и рекомендации по сохранению здоровья;
* логистика: планирование и оптимизация маршрутов доставки;
* финансы: скоринг, обнаружение и предотвращение мошенничества;
* промышленность: предиктивная аналитика для планирования ремонтов и производства;
* недвижимость: поиск и предложение наиболее подходящих покупателю объектов;
* спорт: отбор перспективных игроков и разработка стратегий игры.

Исходя из приведенных мной аргументов и доводов, мы можем сделать вывод, что область применения data science действительно очень обширна, а также мы понимаем, что data science значительно облегчает нашу жизнь и за этим направлением будущее.

**4. Математика в Data Science**

Математика — это неотъемлемый элемент в Data Science. Ведь на самом деле без них невозможно по-настоящему глубоко анализировать и систематизировать огромные массивы данных.  
  
Для специалиста Data Science важны следующие направления математики:

* статистика;
* теория вероятностей;
* математический анализ;
* линейная алгебра

**4.1. Теория вероятностей и математическая статистика**

В теорию вероятности и математическую статистику data science входят:

* Формула Байса
* Распределение Пуассона и биноминальный закон распространения
* Описательная статистика
* Функция распределения и плотность распределения вероятностей. Равномерное и нормальное распределение. Центральная предельная теорема
* Доверительные интервалы. А/В – тестирование
* Линейная регрессия
* Дисперсионный анализ. Логистическая регрессия

|  |  |  |
| --- | --- | --- |
| Размещение | Перестановка | Сочетание |
| Число упорядоченных комбинаций подмножества k объектов из множества n объектов. | Сколькими комбинациями мы можем отсортировать наши n объектов | Размещение при игнорировании порядка. Для этой величины наборы (A, B) и (B, A) идентичны |
|  |  |  |

|  |
| --- |
| Условная вероятность и формула Байеса |
| Смысл условной вероятности – оценивать вероятность некоторого события при условии, что осуществилось иное событие, предшествующее финальному.  Важные аспекты:  ● A&B обозначает составное событие, при котором выполнились и А, и В. Причём, поскольку В наступает раньше А, вероятность составного события определяется как произведение условной вероятности «А при В» на вероятность В.  ● Формула Байеса позволяет оценивать условную вероятность через срабатывание «В при А» |
|  |

**4.2. Математический анализ**

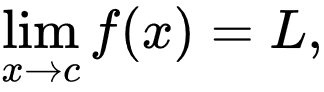
В математический анализ data science входят:

* Дифференциальное исчисление
* Интегральное исчисление
* Теория функций действительного переменного
* Теория функций комплексного переменного
* Приближение функций
* Теория дифференциальных уравнений
* Теория интегральных уравнений
* Вариационное исчисление
* Функциональный анализ и некоторые другие математические дисциплины

Прежде чем начать рассматривать некоторые разделы мат. анализа, давайте ознакомимся с пределом функции:

В [математике](https://en.wikipedia.org/wiki/Mathematics) предел - это значение, к которому приближается [функция](https://en.wikipedia.org/wiki/Function_(mathematics)), когда входные данные приближаются к некоторому [значению](https://en.wikipedia.org/wiki/Value_(mathematics)). Пределы необходимы для [исчисления](https://en.wikipedia.org/wiki/Calculus) и [математического анализа](https://en.wikipedia.org/wiki/Mathematical_analysis) и используются для определения [непрерывности](https://en.wikipedia.org/wiki/Continuous_function), [производных](https://en.wikipedia.org/wiki/Derivative) и [интегралов](https://en.wikipedia.org/wiki/Integral).

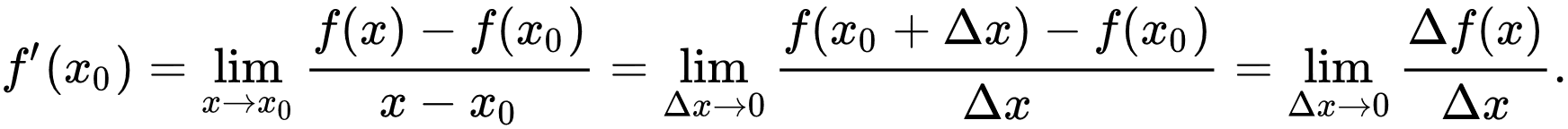
В формулах предел функции обычно записывается как:



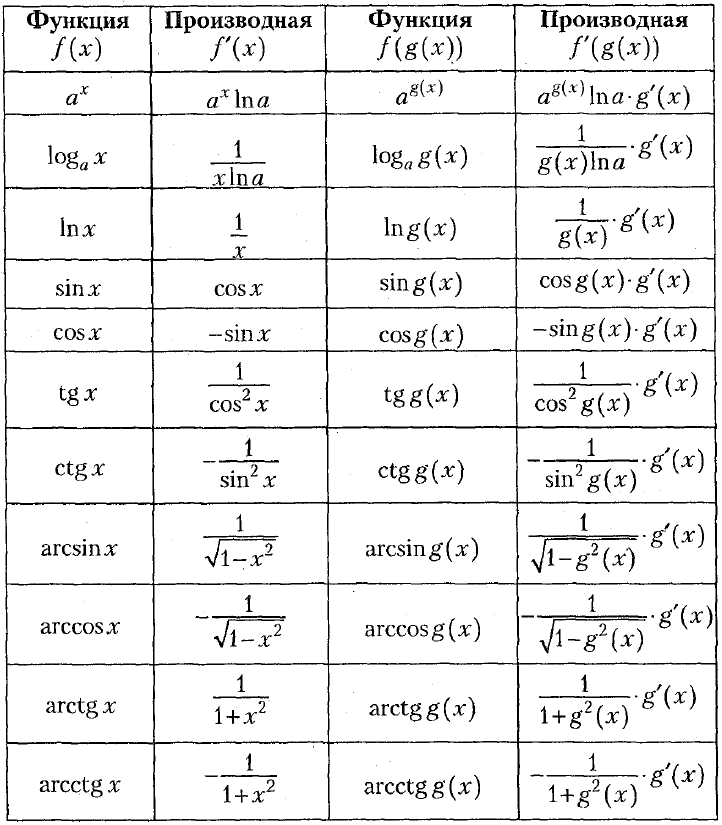
Дифференциальное исчисление

Дифференцирование — это действие вычисления производной, в свою очередь производная – это понятие, которое характеризует скорость изменения функции в данной точке.

Определение производной через предел:



В математике существует таблица производных:

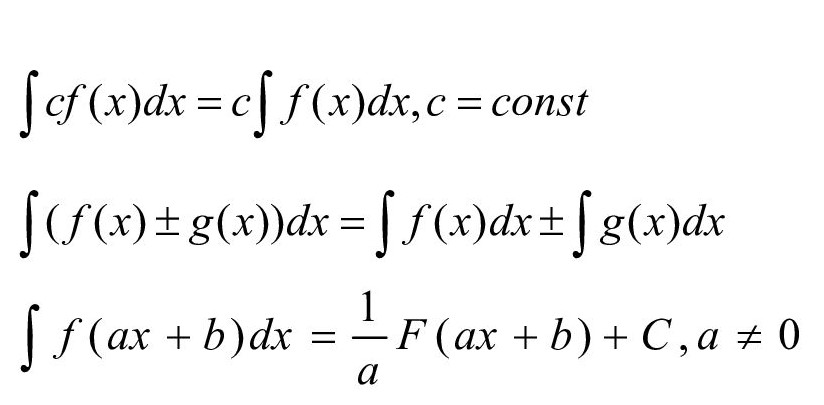


Основные правила дифференцирования: Изображение выглядит как текст

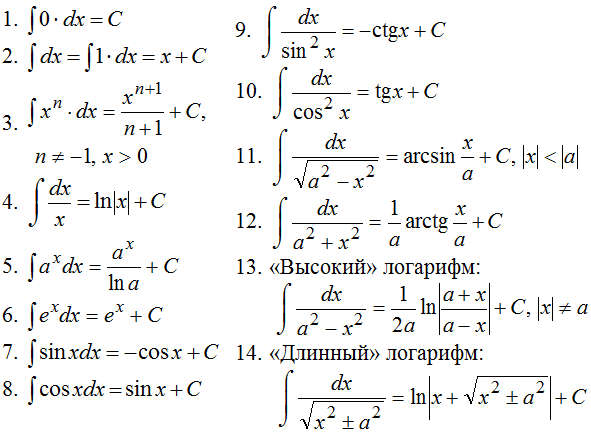
Автоматически созданное описание

Интегрирование

По сути, интегрирование является обратным процессом дифференцирования.



И так же, как и у дифференцирования существует таблица интегралов:



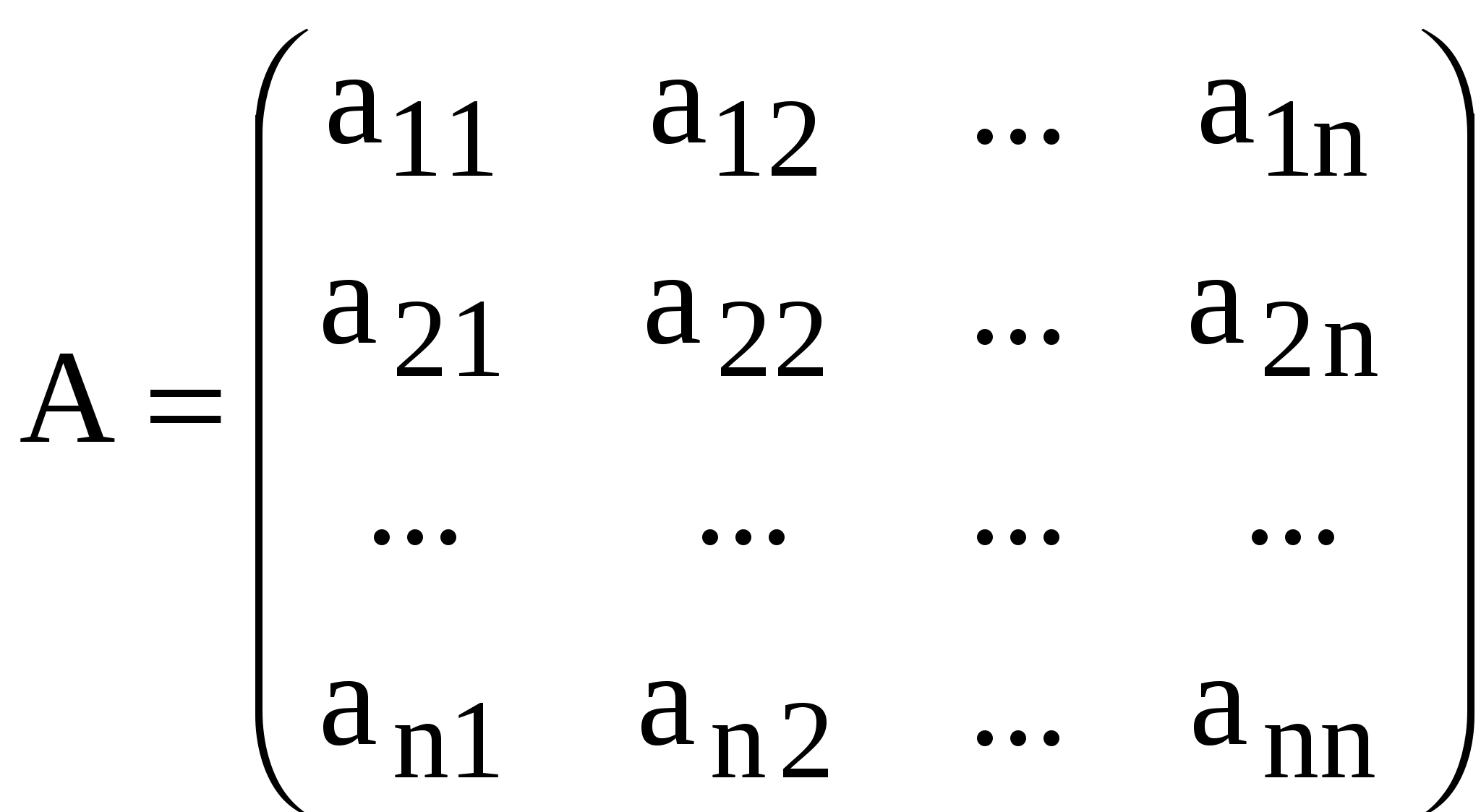
**4.2. Линейная алгебра**

В линейную алгебру data science входят:

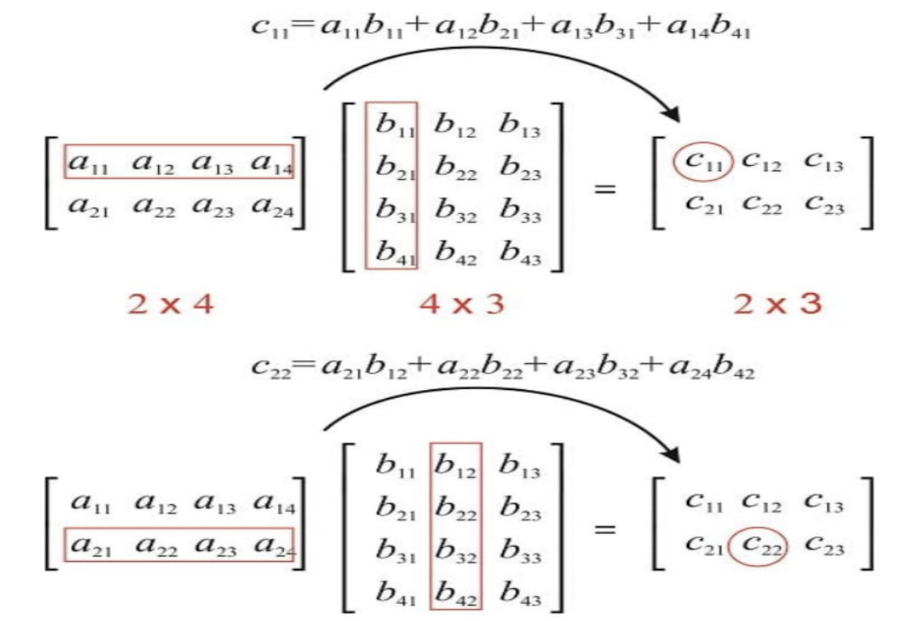
* Линейное пространство
* Матрицы и матричные операции
* Линейные преобразования
* Системы линейных уравнений
* Сингулярное разложение матриц

Для ознакомления с линейной алгеброй в data science предлагаю ознакомиться с базовой работой над матрицами, поскольку именно матрицы являются неотъемлемой частью data science.

Матрицы имеют вид:



Базовыми операциями являются умножение и сложение матриц:



Изображение выглядит как текст, устройство, панель управления, ночное небо

Автоматически созданное описание

**5. Проект**

Буквально несколько столетий назад человечество не знало большинство болезней и как с ними бороться. А даже если и знали, что за болезнь у человека, врачи, да и сами люди, это поздно замечали и совсем не оставалось времени на лечение недуга, впоследствии чего человек погибал. Сейчас же медицина находится совершенно на другом уровне и может лечить и выявлять большинство заболеваний на ранних стадиях. Но это все равно не всегда спасает человека, поскольку со временем появляется все больше коварных болезней, многие из которых очень опасны и даже могут привести к летальному исходу. Одной из самых больших за последние время по числу заболеваний и числу летальных исходов, является COVID-19. Этот вирус можно обнаружить по снимку грудной клетки человека, но бывает, что человеческий глаз попросту не может отличить снимок больного человека от снимка здорового человека, из-за мелких деталей, незаметных невооруженному глазу, а также неопытному специалисту.

Одним из решением всех этих проблем может являться моя программа, которая по снимку грудной клетки способна отличить здорового человека, от человека с заболеванием. А также может различать COVID-19, пневмонию и туберкулез.

При написании кода программы было перепробовано несколько вариантов обучения с предобработкой снимков грудной клетки и без, а также с увеличением числа слоев нейронной сети и увеличением числа эпох. По итогу лучшим результатом является точность ~ 81%. Программу можно продолжать совершенствовать посредством смены модели нейронной сети на более сложную, увеличением числа материала для обучения сети, а также более тщательной предобработкой снимков и увеличением числа эпох. В таком случаем возможно будет повысить точность >90%, но, к сожалению, чем лучше работает модель, тем больше и совершенней требуются вычислительные мощности, коими я не обладаю. Поэтому я считаю, что результат ~81% вполне приемлем для домашней станции проверки снимков на какое-либо заболевание.

(0.5, 1.0, 'Предсказание COVID19')



Мы можем видеть, что программа смогла отличить COVID-19 от остальных заболеваний.

В конце я хочу немного рассказать, как в моей программе используются различные разделы из математики, без которых невозможно работать в Data Science. Моя модель преобразует снимок в матрицу, с которой может работать. И в этом ей помогает Линейная алгебра. Так же находить зависимость одной переменной от другой помогает линейная регрессия, которая в свою очередь является разделом теории вероятности и математической статистики. Ну и на вершине стоит математический анализ, посредством которого проходят различные дифференциальные вычисления и прочие процессы работы со значениями переменных.

Подводя итог, я хочу сказать, что моя программа является одним из способов различения заболеваний по снимку, но всегда нужно учитывать, что модель не всегда правильно предсказывает и имеет ~81% правильных ответов. Поэтому по поводу снимков пока что лучше обращаться к квалифицированным специалистам с большим опытом работы в сфере различных инфекционных заболеваний.

Если кто-то захочет посмотреть исходный код, то ниже я прилагаю ссылку на Google Colab: <https://colab.research.google.com/drive/1W9TLW7cNl4X6oYMOl_IgNb3Q-blDxHxZ?usp=sharing>

**6. Вывод**

На основе проведенного мной исследования можно сделать вывод, что математика всегда являлась, является и будет являться неотъемлемой частью всех сфер деятельности человека. Как раз одной из таких сфер является Data Science. Это направление является достаточно молодым, но уже стало важной частью бизнеса, медицины и прочих сфер. В свою очередь Data Science не было бы возможным без математики, которая является фундаментом в данной науке. В ходе работы над проектом мы выяснили:

1. Что такое Data Science
2. Какие разделы математики входят в Data Science
3. Базовые элементы математического анализа, математической статистики, теории вероятности и линейной алгебры
4. Написали небольшую программу и выяснили ее преимущества и недостатки
5. Пользу науки Data Science для современного общества

Источники:

<https://trends.rbc.ru/trends/industry/cmrm/60740ae09a7947fd64bbed56>

<https://youtu.be/kwLz1JI_SFU>

<https://gb.ru/blog/data-science/>

<https://habr.com/ru/company/netologyru/blog/329068/>

<https://habr.com/ru/company/citymobil/blog/557416/>

<http://dspace.bsu.edu.ru/bitstream/123456789/20926/1/Tarasova_Osnovy_matemat_analiz_18.pdf#:~:text=Математический%20анализ%20охватывает%20достаточно%20большую,и%20некоторые%20другие%20математические%20дисциплины>

<https://en.wikipedia.org/wiki/Limit_(mathematics)>

<https://ru.wikipedia.org/wiki/Производная_функции>

<https://habr.com/ru/company/plarium/blog/442772/>