|  |  |
| --- | --- |
| 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\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

КАФЕДРА \_\_\_\_СИСТЕМЫ ОБРАБОТКИ ИНФОРМАЦИИ И УПРАВЛЕНИЯ\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

РАСЧЕТНО-ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

К НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ

по дисциплине \_\_\_\_\_\_\_\_\_\_\_\_\_ Технологии машинного обучения\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

по теме \_\_\_\_\_\_\_ Решение задачи машинного обучения \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Студент ИУ5Ц-84Б **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**  Е. С. Клеша

(Группа) (Подпись, дата) (И.О.Фамилия)

Руководитель **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**  Ю. Е. Гапанюк

(Подпись, дата) (И.О.Фамилия)

2024 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\_\_\_

(Индекс)

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

(И.О.Фамилия)

« \_\_\_ » \_\_\_\_\_\_\_ 2024 г.

**ЗАДАНИЕ**

**на выполнение научно-исследовательской работы**

по теме \_\_\_\_\_\_\_\_\_\_Решение задачи машинного обучения\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Студент группы \_\_\_ИУ5Ц-84Б\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Клеша Егор Сергеевич \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(Фамилия, имя, отчество)

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

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_исследовательская\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Источник тематики (кафедра, предприятие, НИР) \_кафедра \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

График выполнения НИР: 25% к \_3\_ нед., 50% к \_9\_ нед., 75% к \_12\_ нед., 100% к \_15 нед.

***Техническое задание***

Решение задачи машинного обучения. Результатом проекта является отчет, содержащий описания моделей, тексты программ и результаты экспериментов.\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

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

Расчетно-пояснительная записка на \_19\_\_ листах формата А4.

Перечень графического (иллюстративного) материала (чертежи, плакаты, слайды и т.п.)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Дата выдачи задания «\_\_» \_\_\_\_\_\_\_\_\_\_ 2024 г.

**Руководитель НИР**  \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Ю. Е. Гапанюк

(Подпись, дата) (И.О.Фамилия)

**Студент** \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Е. С. Клеша

(Подпись, дата) (И.О.Фамилия)

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

**СОДЕРЖАНИЕ**

[Введение 4](#_Toc164104565)

[Основная часть 6](#_Toc164104566)

[1. Импорт библиотек. 6](#_Toc164104567)

[2. Загрузка данных. 6](#_Toc164104568)

[3. Изучение данных с целью обнаружения закономерностей. Создание графических представлений для более глубокого понимания информации. Определение и заполнение пропущенных значений в наборе данных. 7](#_Toc164104569)

[4. Обработка данных 9](#_Toc164104570)

[5. Анализ корреляции данных с последующим извлечением выводов относительно возможности построения моделей машинного обучения. 11](#_Toc164104571)

[6. Определение критериев оценки качества моделей для последующего использования. 12](#_Toc164104572)

[7. Определение наилучших моделей для решения задачи классификации или регрессии. 13](#_Toc164104573)

[8. Создание обучающего и тестового наборов данных из исходного датасета. 13](#_Toc164104574)

[9. Создание начального решения (baseline) для выбранных моделей без настройки гиперпараметров. Модели обучаются на обучающем наборе данных, а затем оцениваются по качеству на тестовом наборе. 14](#_Toc164104575)

[10. Настройка гиперпараметров выбранных моделей с применением кросс-валидации. 15](#_Toc164104576)

[11. Применение оптимальных значений гиперпараметров для выбранных моделей и оценка их качества сравнительно с базовыми моделями. 16](#_Toc164104577)

[12. Вывод 17](#_Toc164104578)

[Заключение 18](#_Toc164104579)

[Список использованных источников информации 19](#_Toc164104580)

## Введение

Этот проект представляет собой самый всесторонний и полный источник информации о преступлениях, особенно убийствах, среди доступных на текущий момент в Соединенных Штатах. Он включает данные из Дополнительного отчета ФБР об убийствах с 1976 года по настоящее время, а также информацию, полученную в результате запросов по Закону о свободе информации о более чем 22 000 убийствах, которые не были включены в отчеты Министерства юстиции. Этот набор данных содержит информацию о возрасте, расе, поле, этнической принадлежности как жертв, так и преступников, а также об отношениях между ними и используемом оружии.

В качестве набора данных мы будем использовать набор данных, содержащий историю преступлений в США - <https://www.kaggle.com/datasets/murderaccountability/homicide-reports>.

Датасет состоит из одного файла homicideReports.csv

Файл содержит следующие колонки:

* Record ID – Идентификатор записи;
* Agency Code – Код агентства;
* Agency Type – Тип агентства;
* City – Город;
* State – Государство;
* Year – Год;
* Month – Месяц;
* Incident – Инцидент;
* Crime Type – Тип преступления;
* Crime Solved – Раскрытое преступление;
* Victim Sex – Пол жертвы;
* Victim Age – Возраст жертвы;
* Victim Race – Гонка жертв;
* Victim Ethnicity – Этническая принадлежность жертвы;
* Perpetrator Sex – Пол преступника;
* Perpetrator Age – Возраст преступника;
* Perpetrator Race – Раса преступников;
* Perpetrator Ethnicity – Этническая принадлежность преступника;
* Relationship – Взаимоотношения;
* Weapon – Оружие;
* Victim Count – Количество жертв;
* Perperator Count – Количество преступников;
* Record Source – Источник записи.

В данной работе будем решать задачи регрессии.

## Основная часть

Ссылка на датасет: <https://www.kaggle.com/datasets/yassine9/dataset-spine>.

## Импорт библиотек.

****

## Загрузка данных.

Загрузим файлы датасета с помощью библиотеки Pandas.

Этот файл содержит данные в формате CSV, где информация разделена определенным символом, таким как запятая ",", точка с запятой ";" или табуляция. Перед чтением файла методом read\_csv рекомендуется определить этот символ-разделитель, чтобы правильно указать его с помощью параметра sep. Это можно сделать, открыв файл в любом текстовом редакторе и просмотрев его структуру.



Так как объем данных в наборе велик, мы решили ограничиться только первыми 15000 строками для упрощения анализа и обработки.

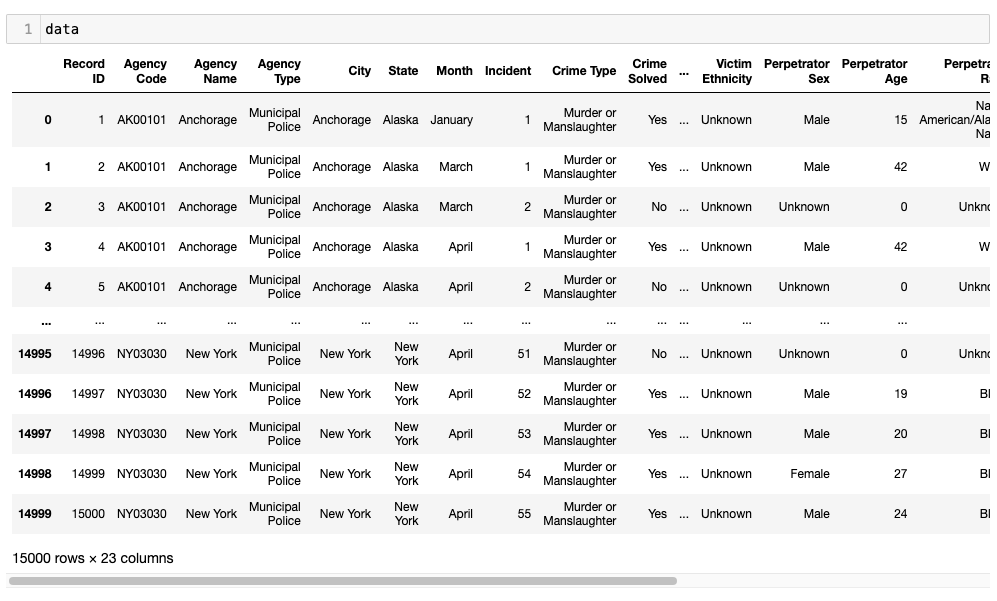


Из-за уменьшения количества строк, столбец Year перестает быть релевантным. В этом наборе данных будут использоваться только данные за один временной период.

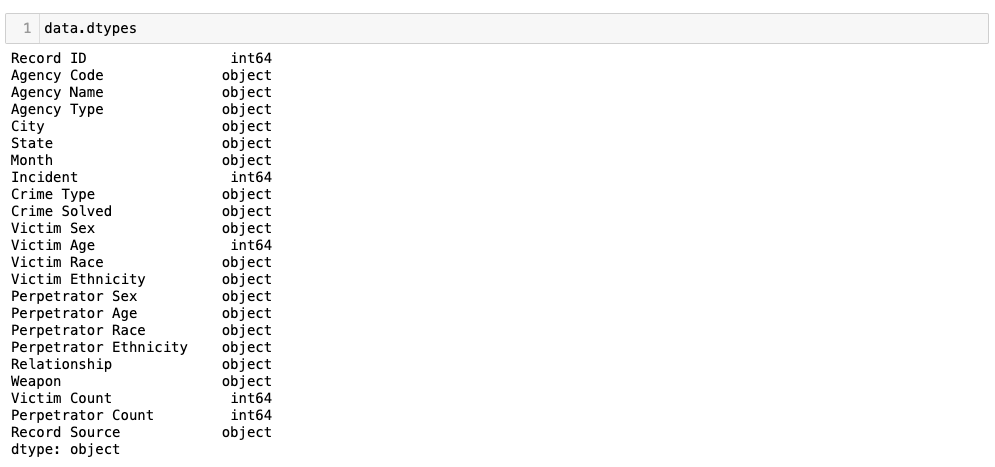


## Изучение данных с целью обнаружения закономерностей. Создание графических представлений для более глубокого понимания информации. Определение и заполнение пропущенных значений в наборе данных.

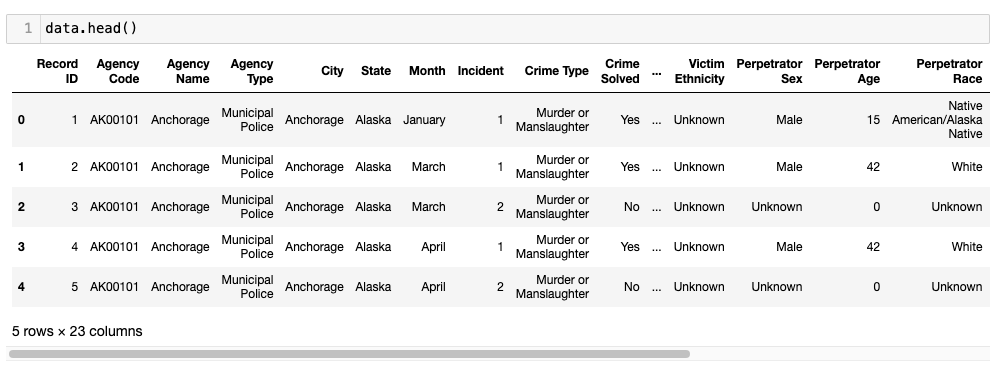
Основные характеристики набора данных.

Первые и последние 5 строк датасета. 

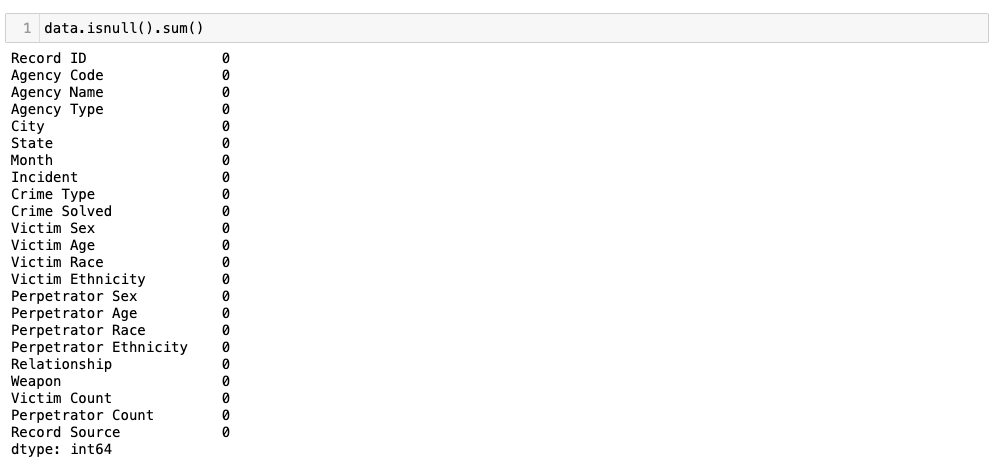
Список колонок с типами данных.



5 первых строк датасета.

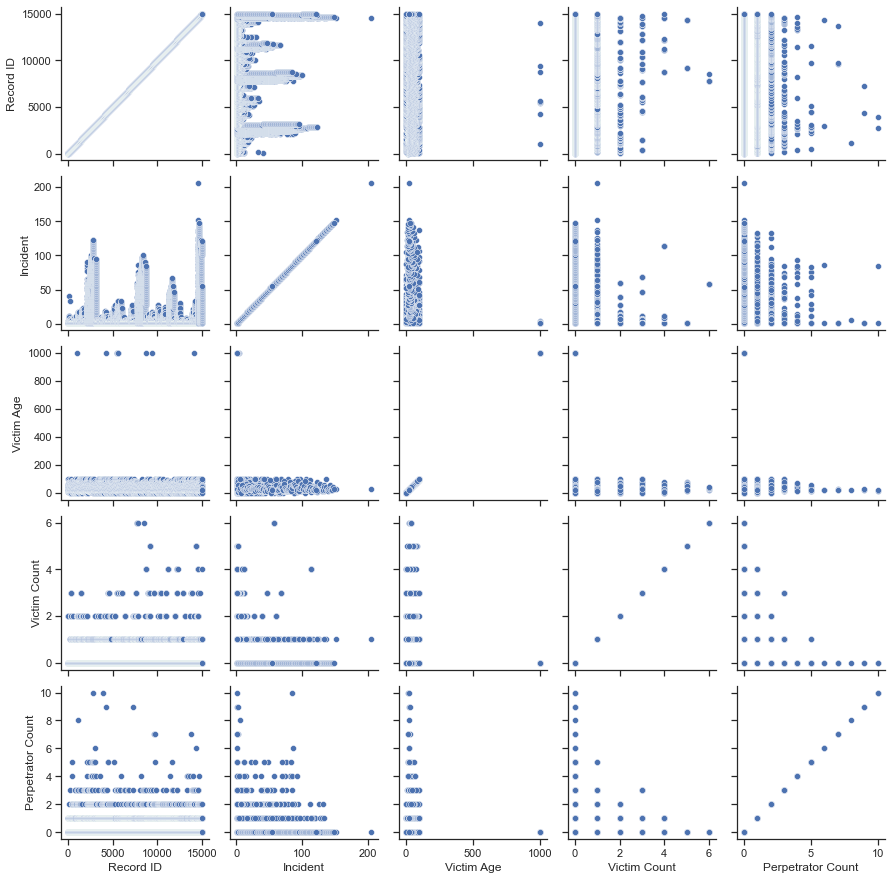


Проверка на наличие пропущеных значений



Создание графических представлений для более глубокого понимания информации.





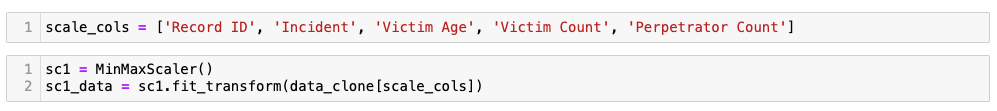
## Обработка данных

Кодирование категориальных признаков.

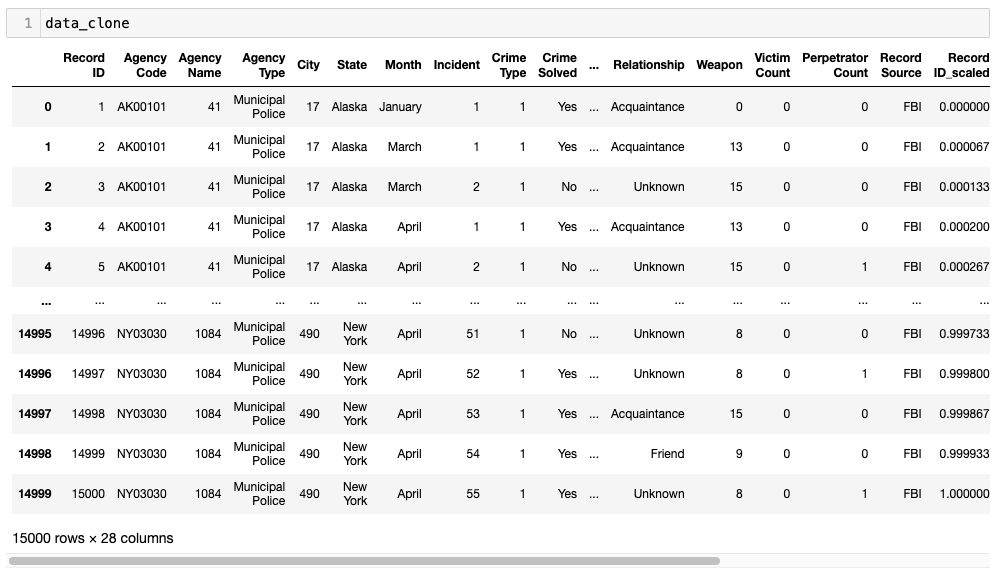


Масштабирование данных.

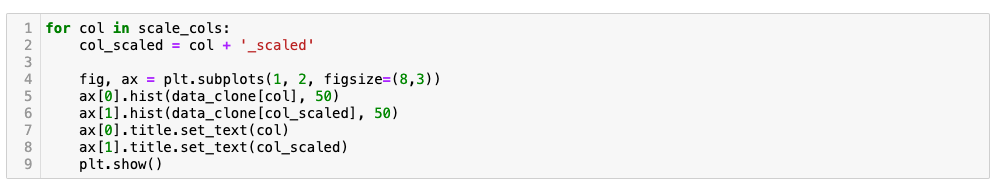
Для повышения эффективности алгоритмов машинного обучения выполним нормализацию данных.

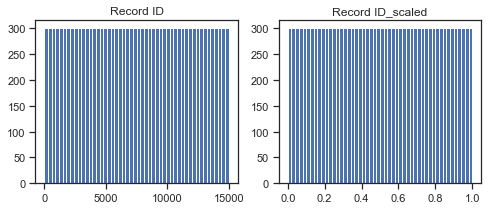


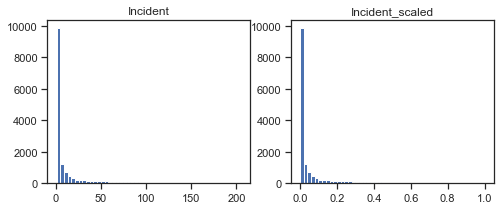
Включим в набор данных отмасштабированные значения. 

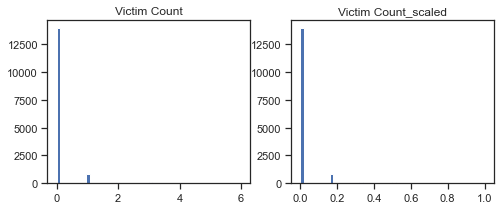


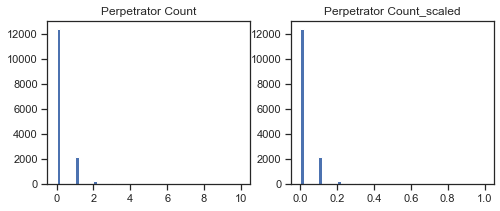
Проверяем сохранение формы распределения данных после масштабирования.



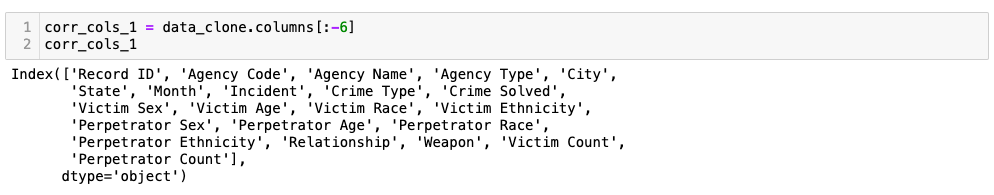


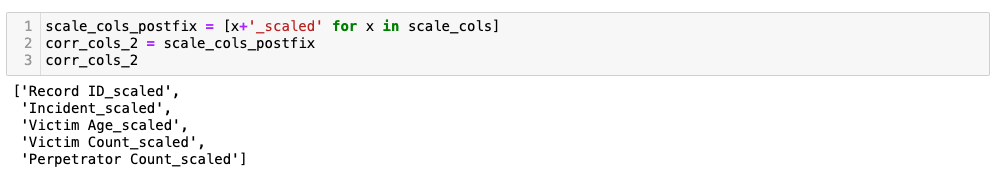


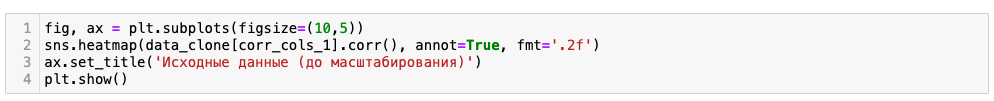




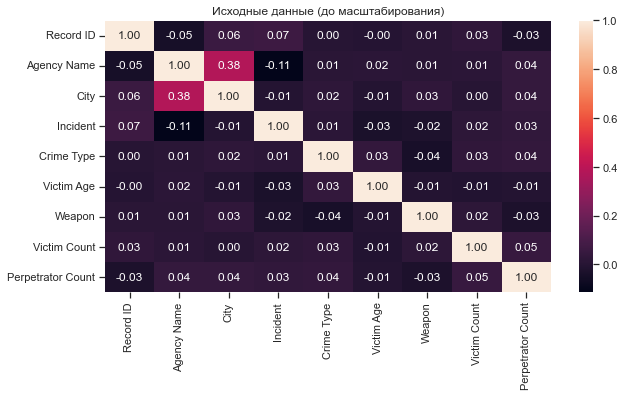
## Анализ корреляции данных с последующим извлечением выводов относительно возможности построения моделей машинного обучения.

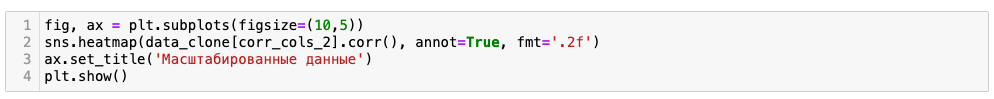




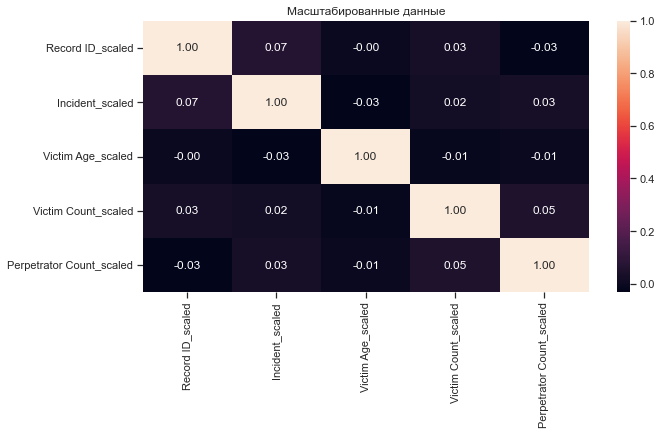


Исходные данные.





Масштабированные данные.



## Определение критериев оценки качества моделей для последующего использования.

Для оценки качества классификационной задачи выберем следующие метрики:

Сохранение и визуализация метрик.

Создадим класс для записи и отображения метрик эффективности построенных моделей.



## Определение наилучших моделей для решения задачи классификации или регрессии.

Для решения задачи регрессии мы выберем следующие модели:

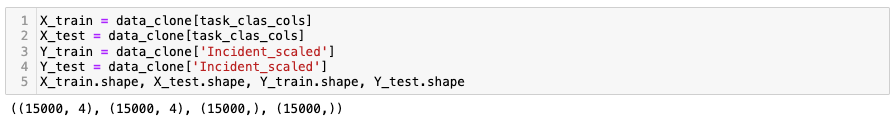
* Линейная регрессия
* Машина опорных векторов
* Решающее дерево
* Метод ближайших соседей
* Градиентный бустинг
* Случайный лес

## Создание обучающего и тестового наборов данных из исходного датасета.

Используя отмасштабированные данные, разделим выборку на обучающий и тестовый наборы с помощью фильтрации.

Выборки для задачи классификации.





## Создание начального решения (baseline) для выбранных моделей без настройки гиперпараметров. Модели обучаются на обучающем наборе данных, а затем оцениваются по качеству на тестовом наборе.

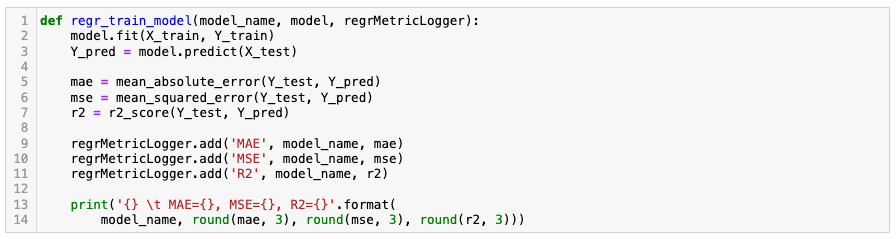
Решение задачи регрессии.

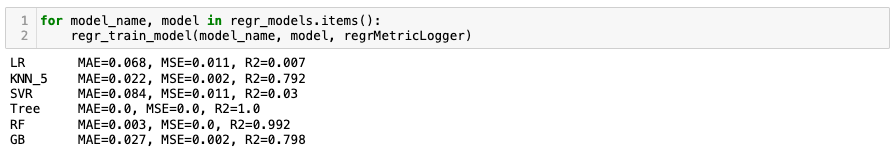
Модели.



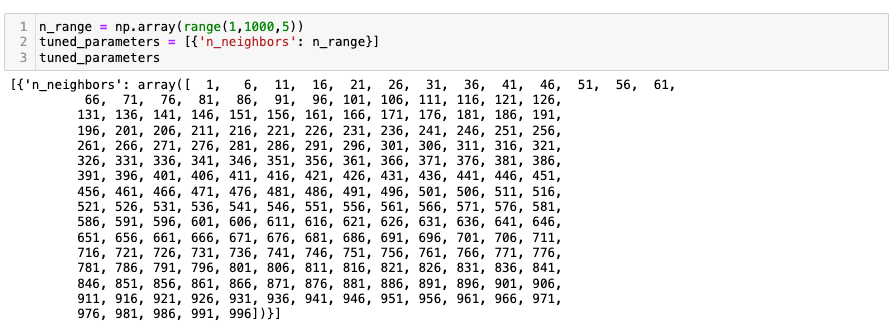
Сохранение метрик.

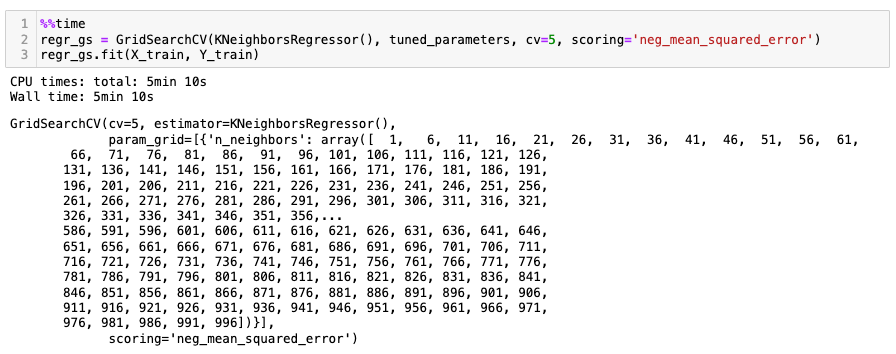




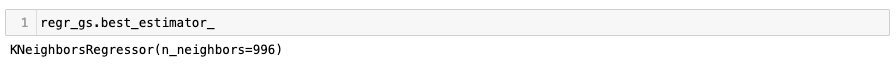


## Настройка гиперпараметров выбранных моделей с применением кросс-валидации.

****



Лучшая модель.

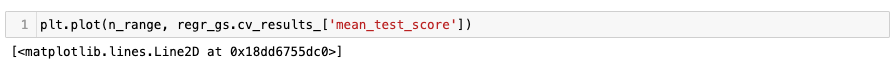


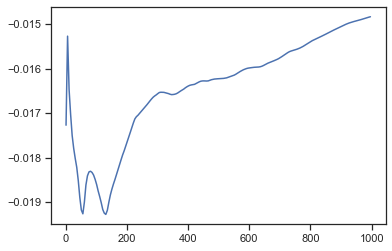
Лучшее значение параметров.



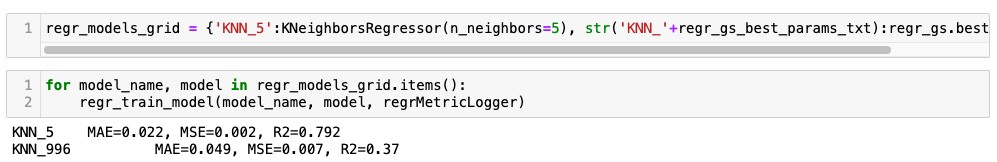


Изменение качества на тестовой выборке в зависимости от К-соседей.





## Применение оптимальных значений гиперпараметров для выбранных моделей и оценка их качества сравнительно с базовыми моделями.

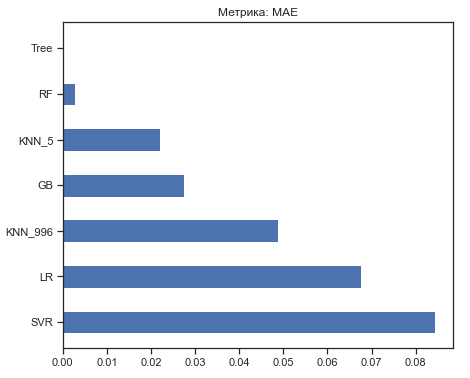


Сделаем выводы о производительности построенных моделей, основываясь на выбранных метриках.

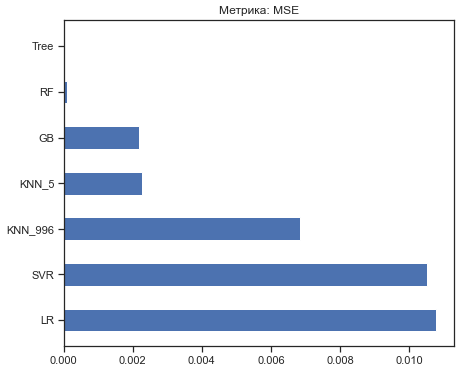
Метрики качества модели.



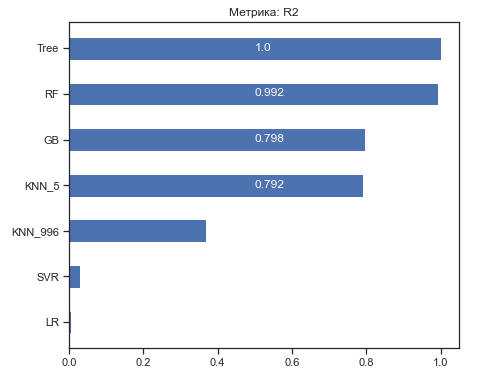












## Вывод

С учетом двух из трех использованных метрик, наилучшей моделью оказалась та, которая базируется на методе опорных векторов (SVR).

## Заключение

В рамках нашего научного исследования мы изучили набор данных "История преступлений США" с помощью различных инструментов анализа данных, таких как Pandas, Matplotlib, Seaborn, Numpy и Streamlit, используя язык программирования Python в Jupyter Notebook.

Мы начали с анализа структуры данных, заполнили пропущенные значения, определили признаки для построения моделей, провели кодирование категориальных признаков, выполнили масштабирование данных и создали дополнительные признаки для улучшения качества моделей.

Проведен корреляционный анализ данных, который позволил сделать промежуточные выводы о возможности построения моделей машинного обучения.

Для оценки качества моделей были выбраны метрики, включая точность, полноту, F1-меру и ROC AUC. Мы рассмотрели более пяти различных моделей, в том числе базовые и ансамблевые методы.

Сформированы обучающая и тестовая выборки, построены базовые решения для выбранных моделей, а также подобраны оптимальные значения гиперпараметров с использованием методов кросс-валидации.

В итоге проведена оценка качества построенных моделей и их сравнение с базовыми моделями, что позволило сделать выводы о наилучших подходах к решению задачи классификации или регрессии, основываясь на выбранных метриках.

## Список использованных источников информации

1. Абдрахманов, М. И. Devpractice Team. Pandas. Работа с данными. / М. И. Абдрахманов — 2-е изд. — devpractice.ru. 2020. - 170 с.: [Электронный ресурс]. // URL: <https://coderbooks.ru/books/python/pandas_rabota_s_dannymi_abdrahmanov_2020/> (дата обращения: 15.03.2024)
2. Абдрахманов, М. И. Devpractice Team. Python. Визуализация данных. Matplotlib. Seaborn. Mayavi. — devpractice.ru. 2020. - 412 с.: [Электронный ресурс]. // URL: <https://coderbooks.ru/books/python/python_vizualizaciya_dannyh_abdrahmanov_2020/> (дата обращения: 21.03.2024)
3. Методические указания по программному обеспечению «Pandas»: [Электронный ресурс]. // URL: <https://pandas.pydata.org/> (дата обращения: 20.02.2024)
4. Методические указания по программному обеспечению «Seaborn»: [Электронный ресурс]. // URL: <https://seaborn.pydata.org/> (дата обращения: 22.02.2024)
5. Метрики качества. Метрики классификации. Подсчет количества ошибок, доли правильных ответов, точности, полноты. [Электронный ресурс]. // URL: <https://data-scientists.ru/metriki-kachestva-metriki-klassifikacii-podschet-kolichestva-oshibok-doli-pravilnyh-otvetov> (дата обращения: 01.03.2024)
6. Регрессионные модели в Python: [Электронный ресурс]. // URL: <https://nagornyy.me/it/regressionnye-modeli-v-python/> (дата обращения: 11.03.2024)
7. Учебник по машинному обучению: [Электронный ресурс]. // URL: <https://education.yandex.ru/handbook/ml> (дата обращения: 01.03.2024)