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

|  |  |
| --- | --- |
| ФАКУЛЬТЕТ | РАДИОТЕХНИЧЕСКИЙ |
| КАФЕДРА | СИСТЕМЫ ОБРАБОТКИ ИНФОРМАЦИИ И УПРАВЛЕНИЯ |

**РАСЧЕТНО-ПОЯСНИТЕЛЬНАЯ ЗАПИСКА**

***К НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ***

***НА ТЕМУ:***

|  |
| --- |
|  |
| ***Разработка и оценка моделей*** |
| ***машинного обучения*** |
|  |
|  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Студент | РТ5-61Б |  |  |  | Г. Г. Слкуни |
|  | (группа) |  | (подпись, дата) |  | (И.О. Фамилия) |
| Руководитель НИР |  |  |  |  | Ю.Е. Гапанюк |
|  |  |  | (подпись, дата) |  | (И.О. Фамилия) |
|  |  |  |  |  |  |

*2025 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) | |
| (подпись) |  | |
|  | (дата) | |

**ЗАДАНИЕ**

**на выполнение научно-исследовательской работы**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| по теме | | Разработка и оценка моделей методов машинного обучения | | | | | | | | | | | | | |
|  | | | |  | | | | | | | | | | | |
|  | | | |  | | | | | | | | | | | |
| Студент группы | | | | РТ5-61Б | | | |  | | | | | | | |
|  | | | | Слкуни Герман Грантович | | | | | | | | | | | |
| Направленность НИР (учебная, исследовательская, практическая, производственная, др.) | | | | | | | | | | | | | | | |
| ИССЛЕДОВАТЕЛЬСКАЯ | | | | | | | | | | | | | | | |
|  | | | | | | | | | | | | | | | |
| Источник тематики (кафедра, предприятие, НИР) | | | | | | | | | | КАФЕДРА | | | | | |
| График выполнения НИР: | | | | | |  | | | | | | | | | |
| 25% к |  | | нед., 50% к | | |  | нед., 75% к | |  | | | нед., 75% к |  | нед |
| ***Техническое задание:*** | | | | |  | | | | | | | | | | |
|  | | | | | | | | | | | | | | | |
|  | | | | | | | | | | | | | | | |
|  | | | | | | | | | | | | | | | |
|  | | | | | | | | | | | | | | | |
| ***Оформление научно-исследовательской работы:*** | | | | | | | | | | |  | | | | |
| Расчетно-пояснительная записка на \_\_\_\_\_ листах формата А4. | | | | | | | | | | | | | | | |
| Перечень графического (иллюстративного) материала (чертежи, плакаты, слайды и т.п.) | | | | | | | | | | | | | | | |
|  | | | | | | | | | | | | | | | |
|  | | | | | | | | | | | | | | | |
|  | | | | | | | | | | | | | | | |
|  | | | | | | | | | | | | | | | |
|  | | | | | | | | | | | | | | | |

Дата выдачи задания « 07 » \_февраля\_\_\_ 2025 г.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Руководитель НИР** |  |  |  | Ю.Е. Гапанюк |
|  |  | (подпись, дата) |  | (И.О. Фамилия) |
| **Студент** |  |  |  | Г. Г. Слкуни |
|  |  | (подпись, дата) |  | (И.О. Фамилия) |

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

СОДЕРЖАНИЕ

[ВВЕДЕНИЕ 4](#_Toc199437512)

[1. ПОСТАНОВКА ЗАДАЧИ 5](#_Toc199437513)

[2. ПОДГОТОВКА ДАННЫХ 6](#_Toc199437514)

[3. ИССЛЕДОВАТЕЛЬСКИЙ АНАЛИЗ ДАННЫХ 7](#_Toc199437515)

[4. ОБРАБОТКА И ПРЕОБРАЗОВАНИЕ ПРИЗНАКОВ 10](#_Toc199437516)

[5. ПОСТРОЕНИЕ И СРАВНЕНИЕ МОДЕЛЕЙ 11](#_Toc199437517)

[6. НАСТРОЙКА ГИПЕРПАРАМЕТРОВ 12](#_Toc199437518)

[7. ВЕБ-ПРИЛОЖЕНИЕ 14](#_Toc199437519)

[ЗАКЛЮЧЕНИЕ 15](#_Toc199437520)

[СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ 16](#_Toc199437521)

ВВЕДЕНИЕ

Машинное обучение на сегодняшний день является одним из наиболее активно развивающихся направлений в области анализа данных. Оно широко применяется для решения различных задач, включая классификацию, регрессию, кластеризацию и многое другое. В данной работе рассматривается задача классификации: необходимо из данных о шахматной партии определить, выиграют белые или черные.

Целью настоящего исследования является построение, обучение и сравнение нескольких моделей машинного обучения, включая ансамблевые методы, а также проведение полного цикла подготовки и анализа данных. В процессе работы производится оценка качества моделей по различным метрикам, подбор гиперпараметров, формирование выводов о качестве решений, а также демонстрация полученного результата в виде веб-приложения.

1. ПОСТАНОВКА ЗАДАЧИ

Имеется открытый набор данных о шахматных партиях с Lichess. На основе этих данных необходимо спрогнозировать, кто победит – игрок, играющий за белые фигуры или игрок, играющий за черные фигуры. Но также необходимо учитывать вариант ничьи. В зависимости от этого задача формализуется либо как задача бинарной классификации, либо как многоклассовая задача. Целевая переменная — winner (black, white, draw).

Для решения задачи требуется:

1. Выполнить разведочный анализ данных;
2. Обработать пропуски и закодировать категориальные переменные;
3. Провести масштабирование признаков;
4. Сформировать обучающую и тестовую выборки;
5. Построить не менее пяти моделей (включая две ансамблевые);
6. Оценить их качество по нескольким метрикам (точность, полнота, F1-мера);
7. Настроить гиперпараметры моделей;
8. Сравнить результаты и обосновать выбор финальной модели.

# ПОДГОТОВКА ДАННЫХ

Выбранный набор включает такие признаки, как:

1. rated — рейтинговая ли игра,
2. white\_rating — рейтинг белого игрока,
3. black\_rating — рейтинг черного игрока,
4. turns — количество ходов,
5. opening\_ply— количество ходов в дебюте,

Из датасета была убрана информации об уникальном идентификаторе партии, дате партии, названии турнира и т.д., т.к. эти данные не несут полезной нагрузки для нашего анализа.

1. ИССЛЕДОВАТЕЛЬСКИЙ АНАЛИЗ ДАННЫХ

Разведочный анализ данных позволил выявить особенности распределения признаков и их связь с целевой переменной. Это важный этап, позволяющий сформировать гипотезы и принять решения по обработке и отбору признаков.

Построим гистограмму, демонстрирующую распределение победителей.

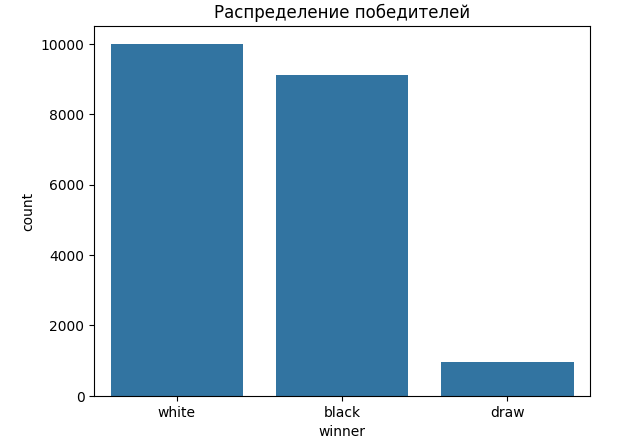


Рисунок 1 – Распределение победителей

Нетрудно заметить, что число случаев, когда произошла ничья гораздо меньше, чем других исходов. Это говорит о том, что данный исход лучше исключить, т.к. в ином случае классы будут несбалансированные, что повлечет низкую точность классификации.

Проанализируем, какой рейтинг у игроков в датасете:

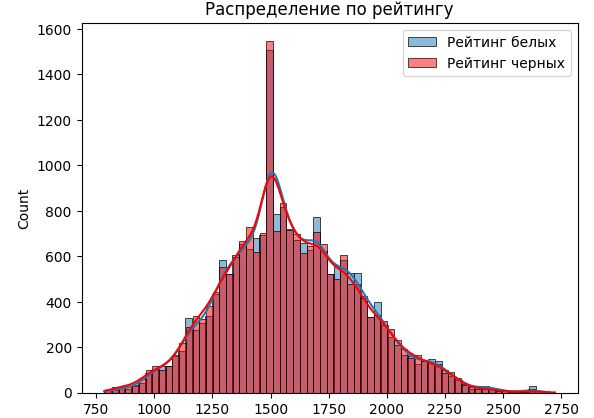


Рисунок 2 – Распределение по рейтингу

Из графика можно сделать вывод, что в среднем рейтинг белых и черных игроков одинаков.

Построим график разницы в рейтинге с победителем партии:

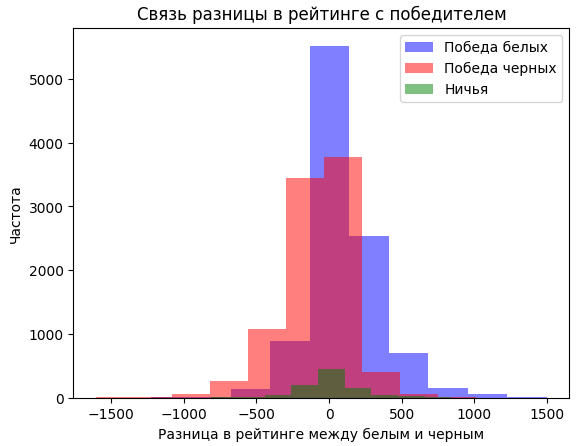


Рисунок 3 – Разница в рейтинге с побидетелем

Из графика видно, что если у игроков одинаковый рейтинг, то шансы на победу у них равны. А в случае отклонения, то шанс победить у игрока с большим рейтингом выше.

Для комплексного понимания взаимосвязей была построена тепловая карта корреляций числовых переменных с целевой переменной.

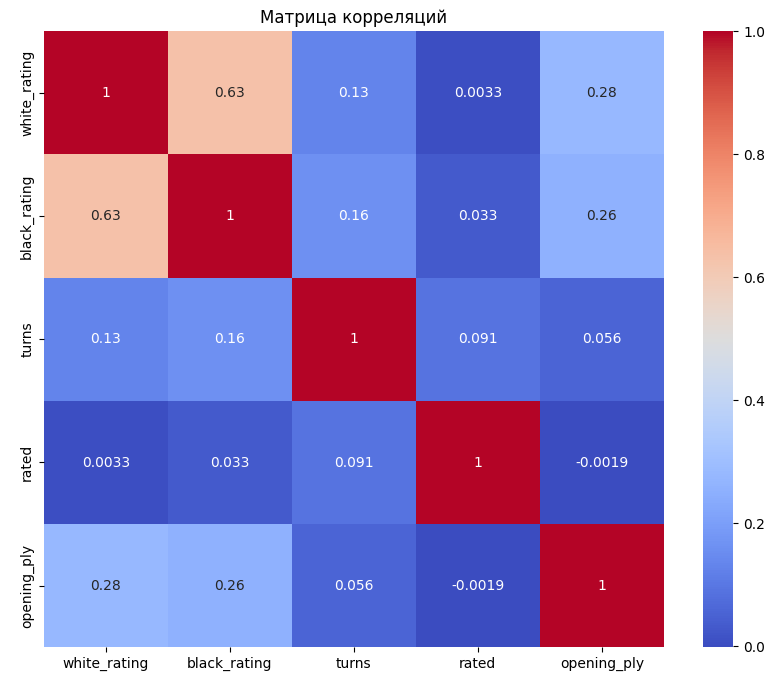


Рисунок 4 – Корреляционная матрица признаков

Матрица корреляций показывает взаимосвязи между признаками в датасете.

1. ОБРАБОТКА И ПРЕОБРАЗОВАНИЕ ПРИЗНАКОВ

Перед построением моделей обработаем датасет:

1. Преобразование переменной rated к integer типу;
2. Удаление ненужных признаков (идентификатор и т.д.);
3. Удаление вариантов, где исход – ничья;
4. Кодирование категориальных признаков;
5. Масштабирование численных признаков.

5. ПОСТРОЕНИЕ И СРАВНЕНИЕ МОДЕЛЕЙ

Разделив датасет на обучающую и тестовую выборку, построим и обучим следующие модели:

1. Логистическая регрессия (LogisticRegression)
2. Дерево решений (DecisionTreeClassifier)
3. Случайный лес (RandomForestClassifier)
4. Градиентный бустинг (GradientBoostingClassifier)
5. Метод опорных векторов (SVM)

Все модели были обучены на одной и той же выборке (80% обучающая, 20% тестовая), и поначалу использовались базовые параметры без настройки.

По результатам базового обучения была получена следующая статистика:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Accuracy | Precision | F1-score |
| Логистическая регрессия | 0,65 | 0,65 | 0,65 |
| Дерево решений | 0,71 | 0,71 | 0,71 |
| Случайный лес | 0,68 | 0,68 | 0,68 |
| Градиентный бустинг | 0,73 | 0,73 | 0,73 |
| Метод опорных векторов | 0,65 | 0,65 | 0,65 |

1. НАСТРОЙКА ГИПЕРПАРАМЕТРОВ

После проведения GridSearchCV и повторной оценки моделей наблюдается улучшение производительности у всех алгоритмов:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Accuracy | Precision | F1-score |
| Логистическая регрессия | 0,65 | 0,65 | 0,69 |
| Дерево решений | 0,75 | 0,75 | 0,75 |
| Случайный лес | 0,69 | 0,69 | 0,71 |
| Градиентный бустинг | 0,89 | 0,89 | 0,89 |
| Метод опорных векторов | 0,59 | 0,66 | 0,7 |

Построим сравнительные гистограммы:

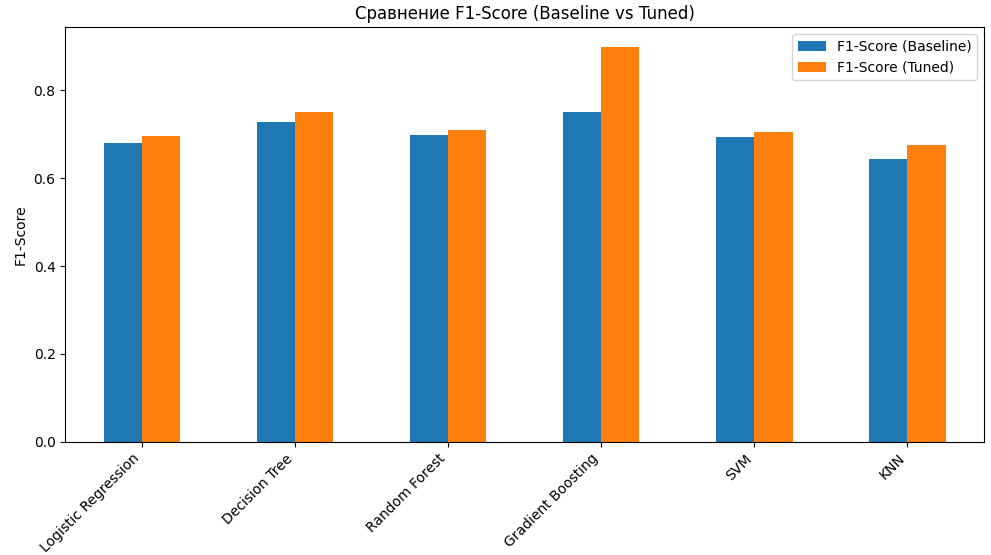


Рисунок 5 – Сравнение моделей по F1-Score

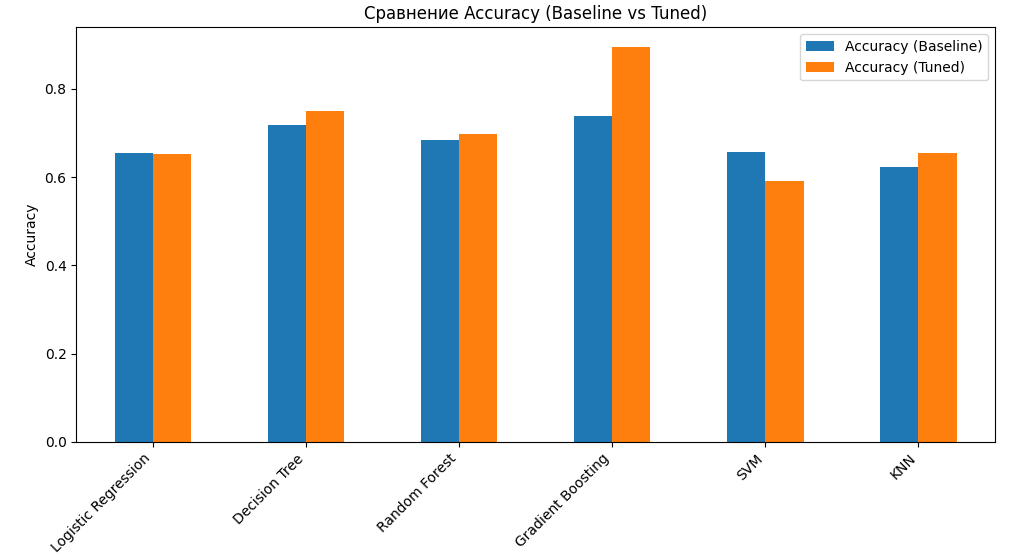


Рисунок 6 – Сравнение моделей по Accuracy

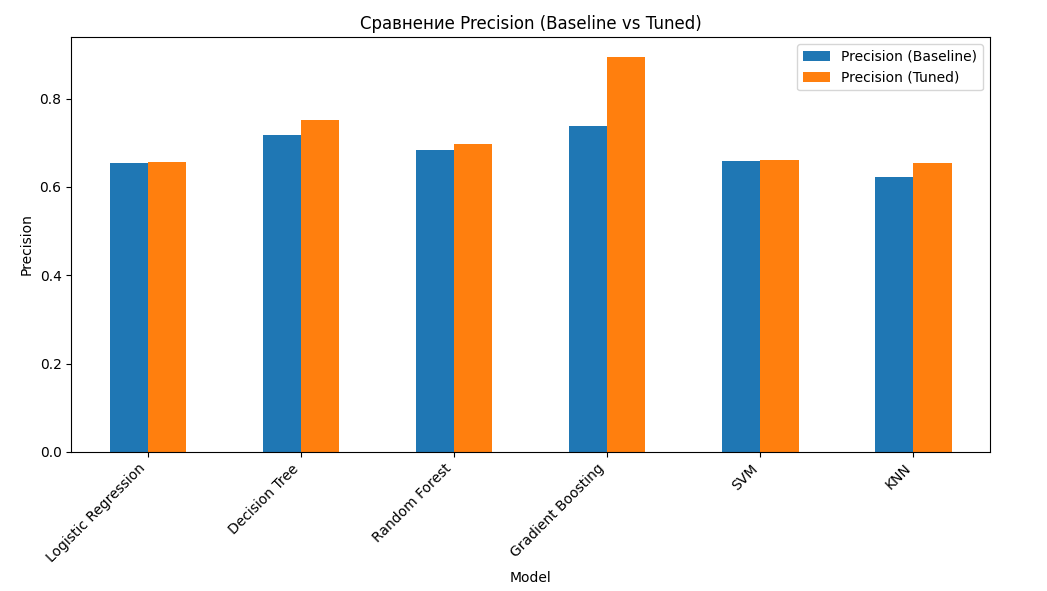
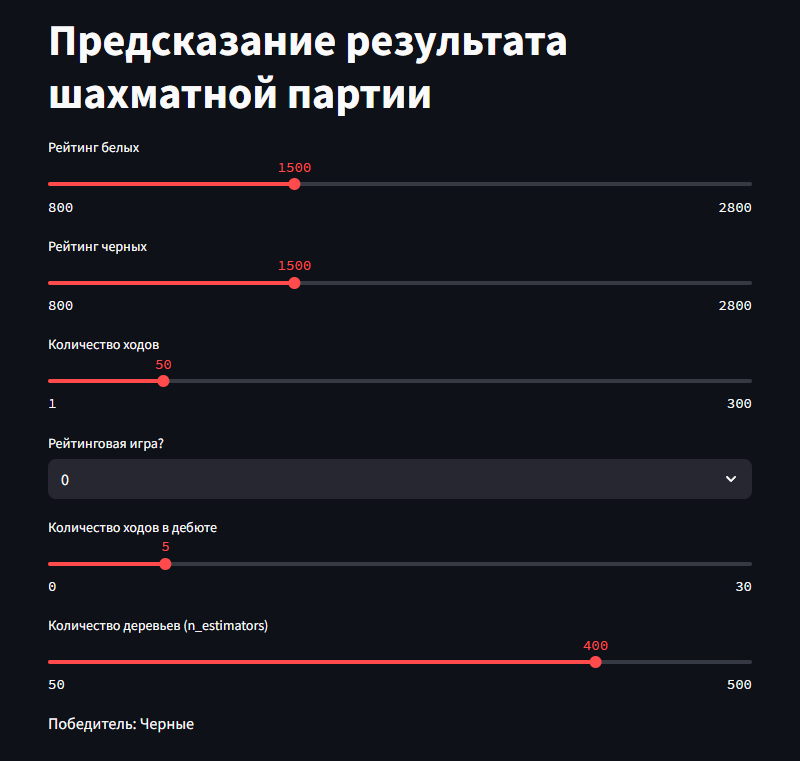


Рисунок 7 – Сравнение моделей по Precision

Наилучшие результаты по F1-мере, точности и полноте показала модель градиентного бустинга после настройки гиперпараметров. Будем использовать его для следующего этапа.

1. ВЕБ-ПРИЛОЖЕНИЕ

Реализуем веб-приложение для демонстрации влияния гиперпараметров на точность модели Градиентный бустинг. Используем фреймворк Streamlit.



**Рисунок 8 – Веб-приложение**

Изменяя положение параметров, будем получать разные предсказания шахматной партии.

ЗАКЛЮЧЕНИЕ

Результаты настоящей работы показали, что при системном подходе к решению задачи классификации можно существенно повысить точность моделей за счёт следующих факторов:

1. Тщательная обработка и расширение признаков;
2. Корректное кодирование и масштабирование данных;
3. Использование продвинутых моделей и подбор гиперпараметров;
4. Анализ метрик, позволяющих делать взвешенные выводы.
5. Наилучшей моделью была признана градиентного бустинга, которая после настройки показала F1-score выше 0.85, а также хорошую сбалансированность между точностью и полнотой. Это делает ее наиболее подходящей для решения поставленной задачи.

Полученные модели и подготовленный код могут быть легко адаптированы для других задач классификации, что демонстрирует универсальность применённого подхода. Кроме того, была реализована сохранённая модель и масштабировщик, которые могут использоваться в продуктивной среде, а также возможно их внедрение в веб-интерфейс с использованием streamlit.

**СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ**

1. Kaggle: Chess Game Dataset (https://www.kaggle.com/datasets/datasnaek/chess)
2. Géron, Aurélien. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow — O'Reilly, 2019.
3. Документация Scikit-learn — <https://scikit-learn.org/>
4. Материалы курса "Машинное обучение", OpenAI, Stepik, Coursera
5. Python Software Foundation — <https://www.python.org/>
6. Визуализация и EDA: <https://seaborn.pydata.org/>, <https://matplotlib.org/>