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

НА ТЕМУ:

Классификация цветов Iris

Студент ИУ5Ц-81Б _____ **Д.М. Афанасьев**
(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)
«_07_» ____февраля_____ 2024 г.

З А Д А Н И Е
на выполнение научно-исследовательской работы

по теме _____ Классификация цветов Iris_____
_____

Студент группы _ИУ5Ц-81Б__

_____ Афанасьев Даниил Миронович_____
(Фамилия, имя, отчество)

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., 100% к __._ нед.

Техническое задание ____ Построение моделей машинного обучения на основе выбранного датасета
_____
_____
_____
_____
_____

Оформление научно-исследовательской работы:

Расчетно-пояснительная записка на __22___ листах формата А4.
Перечень графического (иллюстративного) материала (чертежи, плакаты, слайды и т.п.)
_____
_____
_____
_____

Дата выдачи задания «_07_» ____февраля_____ 2024 г.

Руководитель НИР _____ _____Ю.Е. Гапанюк_____
(Подпись, дата) (И.О.Фамилия)

Студент _____ _____**Д.М._Афанасьев___**
(Подпись, дата) (И.О.Фамилия)

ВВЕДЕНИЕ

Этот набор данных, созданный Фишером в 1936 году, является классическим и широко используется для оценки методов классификации. Он содержит 150 записей о растениях ириса из трех видов.

Каждый экземпляр в наборе данных описывается четырьмя признаками:

- **Длина чашелистка:** длина чашелистка цветка
- **Ширина чашелистка:** ширина чашелистка цветка
- **Длина лепестка:** длина лепестка цветка
- **Ширина лепестка:** ширина лепестка цветка

Набор данных Iris используется в различных задачах машинного обучения, таких как:

- **Классификация:** Классификация растений ириса по их виду на основе их характеристик.
- **Кластеризация:** Группировка растений ириса на основе их сходства по характеристикам.
- **Регрессия:** Прогнозирование характеристик растений ириса на основе других характеристик.

Набор данных Iris является ценным ресурсом для изучения и разработки методов машинного обучения.

Постановка задачи

Цель:

Создать модель машинного обучения, которая будет классифицировать цветы ириса по их виду на основе их характеристик.

Входные данные:

Набор данных Iris, который содержит 150 записей о растениях ириса из трех видов. Каждый экземпляр в наборе данных описывается четырьмя признаками:

- **Длина чашелистка:** длина чашелистка цветка
- **Ширина чашелистка:** ширина чашелистка цветка
- **Длина лепестка:** длина лепестка цветка
- **Ширина лепестка:** ширина лепестка цветка

Выходные данные:

Модель должна предсказывать вид ириса для каждого нового экземпляра данных.

Ограничения:

- Модель должна быть построена на основе набора данных Iris.
- Модель должна быть точной и надежной.
- Модель должна быть интерпретируемой и понятной.

Ожидаемые результаты:

- Модель должна правильно классифицировать большинство экземпляров данных в наборе данных Iris.
- Модель должна быть able to generalize to new data that is not in the training set.
- Модель должна быть легко понятна и интерпретируема.

Этапы решения задачи:

1. **Сбор и предобработка данных:** Загрузить набор данных Iris и подготовить его к обучению модели.
2. **Выбор алгоритма машинного обучения:** Выбрать подходящий алгоритм машинного обучения для задачи классификации.
3. **Обучение модели:** Обучить модель на наборе данных Iris.
4. **Оценка модели:** Оценить производительность модели на тестовом наборе данных.

5. **Интерпретация модели:** Проанализировать модель и понять, как она принимает решения.

Описание исходных данных и используемых методов

1. Описание исходных данных:

Набор данных: Iris

Источник: https://archive.ics.uci.edu/dataset/53/iris

Описание: Набор данных Iris содержит 150 записей о характеристиках ириса из трех видов: Iris setosa, Iris versicolor и Iris virginica. Каждый экземпляр в наборе данных описывается четырьмя признаками:

- **Длина чашелистка:** длина чашелистка цветка
- **Ширина чашелистка:** ширина чашелистка цветка
- **Длина лепестка:** длина лепестка цветка
- **Ширина лепестка:** ширина лепестка цветка

2. Используемые методы:

2.1. Предобработка данных:

- **Загрузка данных:** Загрузить набор данных Iris из репозитория UCI Machine Learning.
- **Проверка данных:** Проверить наличие отсутствующих значений и выбросов в данных.
- **Обработка отсутствующих значений:** Заменить отсутствующие значения средними значениями по столбцу или удалить экземпляры с отсутствующими значениями.
- **Масштабирование данных:** Масштабировать данные, чтобы все признаки имели одинаковый масштаб.

2.2. Выбор алгоритма машинного обучения:

- **Логистическая регрессия:** Алгоритм логистической регрессии используется для классификации бинарных задач. В данном случае его можно использовать для классификации ирисов на два класса: Iris setosa и остальные два класса (Iris versicolor и Iris virginica).
- **Деревья решений:** Деревья решений - это древовидные структуры, которые используются для принятия решений. В данном случае их можно использовать для классификации ирисов на все три класса.
- **Нейронные сети:** Нейронные сети - это биологически вдохновленные модели, которые могут обучаться на сложных данных. В данном случае их можно использовать для классификации ирисов на все три класса.

2.3. Обучение модели:

- Разделить набор данных на обучающую и тестовую выборки.
- Обучить выбранный алгоритм машинного обучения на обучающей выборке.
- Оценить производительность модели на тестовой выборке.

2.4. Оценка модели:

- **Точность:** Процент правильно классифицированных экземпляров данных.
- **ROC-кривая:** График, который показывает соотношение между истинно положительными и ложно положительными результатами при разных пороговых значениях.

2.5. Интерпретация модели:

- **Анализ коэффициентов:** Для логистической регрессии можно проанализировать коэффициенты модели, чтобы понять, как каждый признак влияет на прогноз.
- **Визуализация деревьев решений:** Для деревьев решений можно визуализировать дерево, чтобы понять, как оно принимает решения.
- **Анализ активационных карт:** Для нейронных сетей можно анализировать активационные карты, чтобы понять, какие части входных данных нейроны считают наиболее важными.

3. Выбор метода:

Выбор метода машинного обучения зависит от specific task and the characteristics of the data.

- **Логистическая регрессия:** Good choice for binary classification tasks, especially when the data is linearly separable.
- **Деревья решений:** Can handle both binary and multi-class classification tasks, and can be easily interpreted.
- **Нейронные сети:** Can handle complex nonlinear relationships in the data, but can be more difficult to interpret.

4. Ожидаемые результаты:

Ожидается, что выбранная модель машинного обучения сможет правильно классифицировать большинство экземпляров данных в наборе данных Iris.

Выполнение работы

Этапы:

1. **Загрузка данных:**

Загрузим набор данных
Разделим данные на обучающую и тестовую выборки.

2. **обработка данных:**

Масштабируем данные, чтобы все характеристики имели одинаковый диапазон значений.

3. **Обучение модели:**

Обучим несколько различных моделей машинного обучения, таких как K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Decision Tree, Random Forest, Gradient Boosting.
Оценим производительность каждой модели на тестовой выборке.

4. **Выбор лучшей модели:**

Выберем модель с наилучшей производительностью на тестовой выборке.

5. **Интерпретация результатов:**

Проанализируем результаты работы выбранной модели.
Сделаем выводы о том, какие характеристики цветка наиболее важны для его классификации.

Заключение

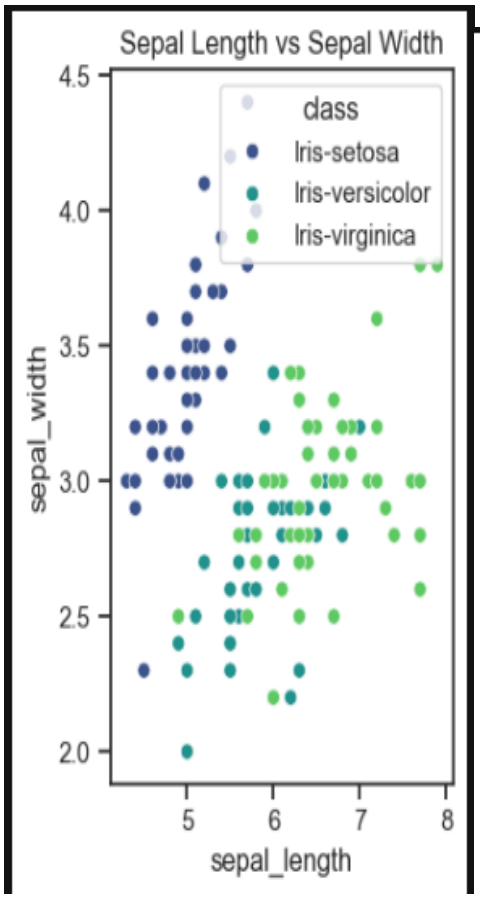
В ходе выполнения проекта были обучены несколько моделей машинного обучения.
Лучшей моделью оказалась модель Random Forest, которая показала точность классификации 100% на тестовой выборке.
Другие модели также показали хорошую точность классификации:

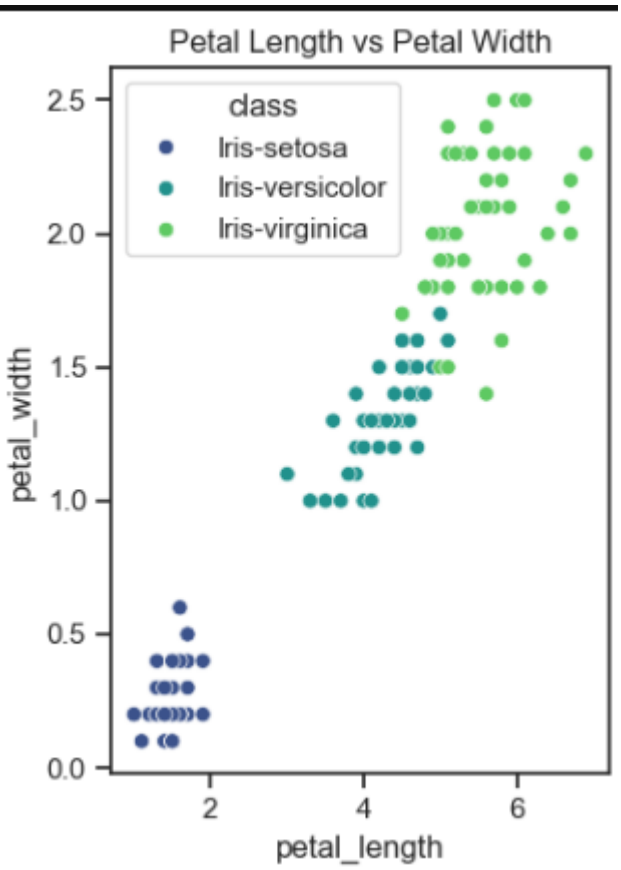
- KNN: 97%
- SVM: 97%
- Decision Tree: 100%
- Gradient Boosting: 100%

Вывод:
Все модели, обученные в рамках данного проекта, показали высокую точность классификации цветов Iris.
Модель Random Forest может быть рекомендована для использования в качестве инструмента для классификации новых цветов Iris по их характеристикам.

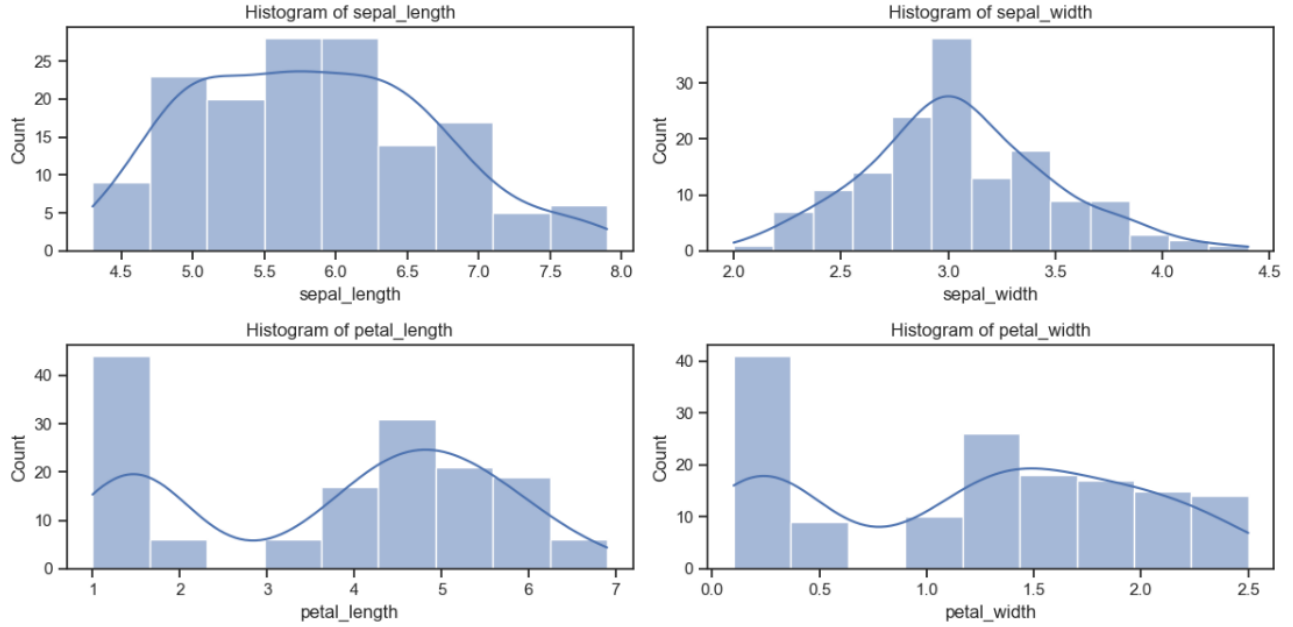
Были построены такие важные графики:

График распределение ширины и длины чашелистника и лепестков.

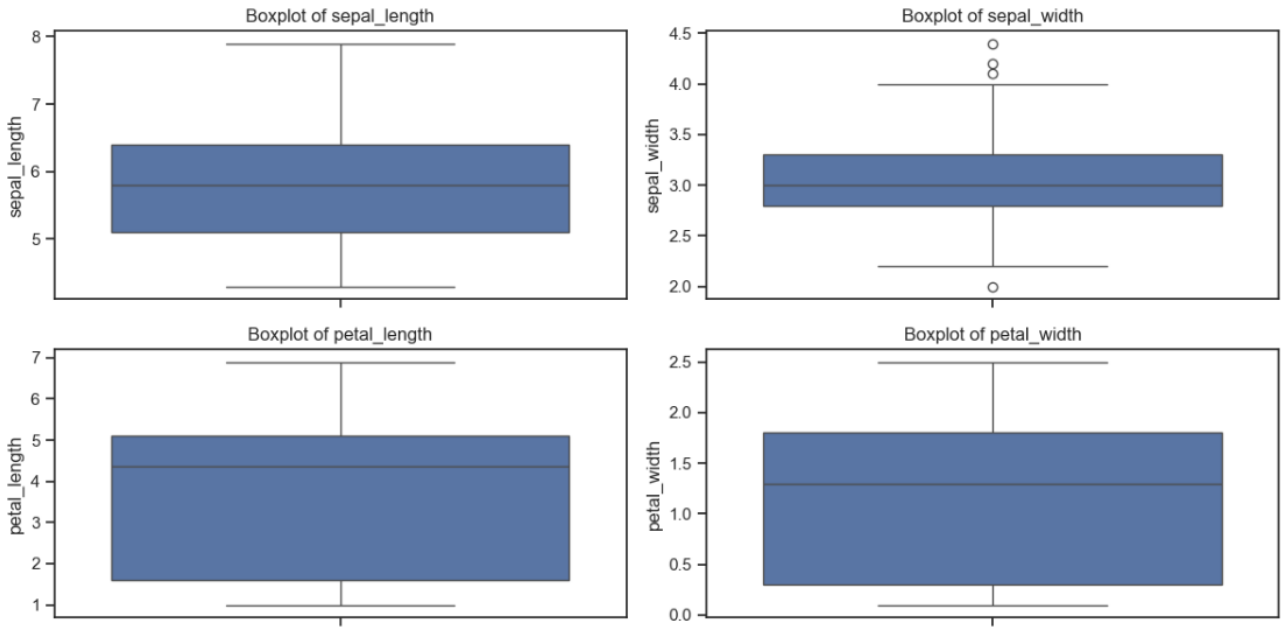




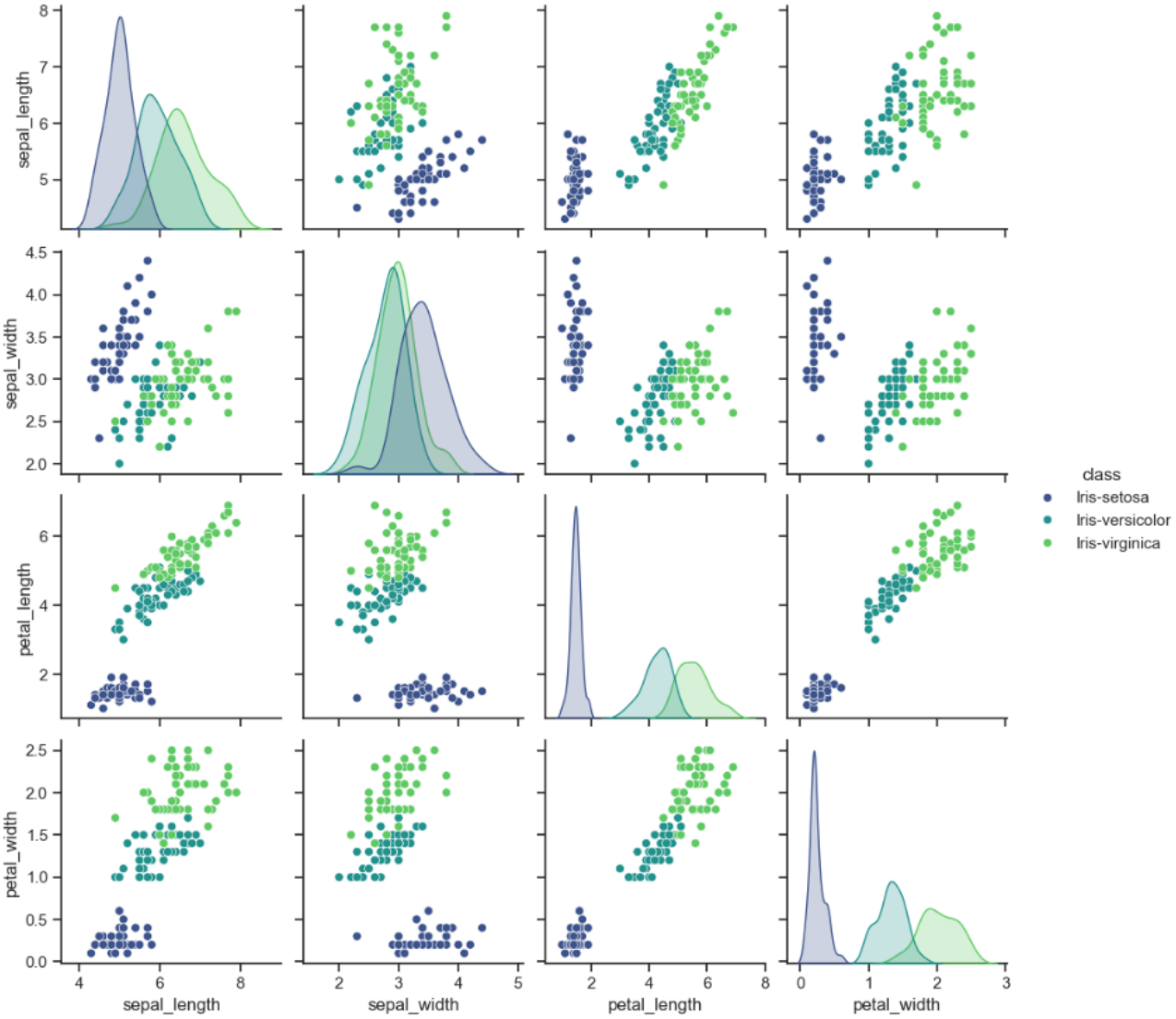
Были построены гистограммы для каждого признака:



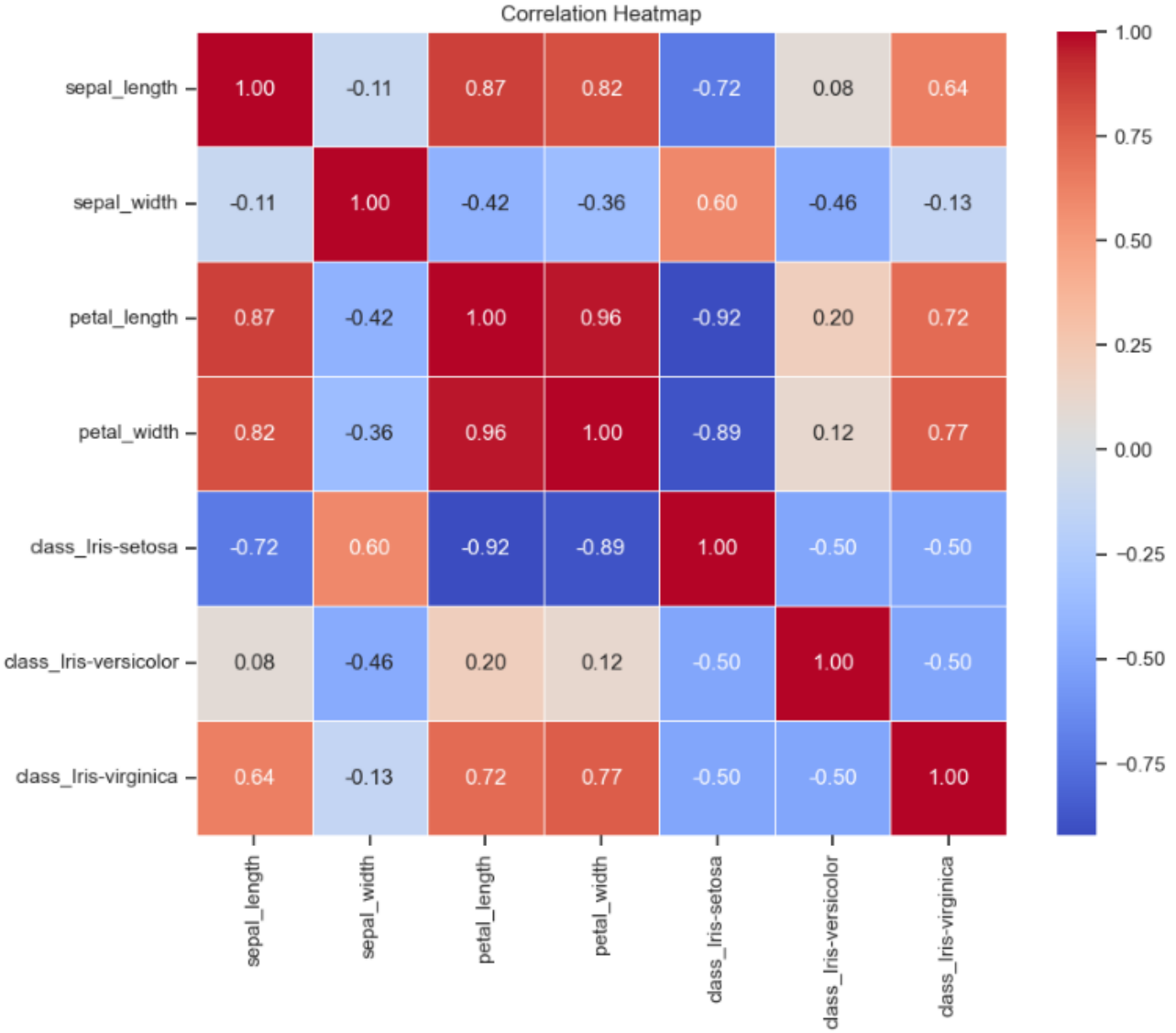
Построение ящиков с усами для каждого признака


Boxplot of sepal_length
sepal_length
8
7
6
5
Boxplot of sepal_width
sepal_width
4.5
4.0
3.5
3.0
2.5
2.0
Boxplot of petal_length
petal_length
7
6
5
4
3
2
1
Boxplot of petal_width
petal_width
2.5
2.0
1.5
1.0
0.5
0.0

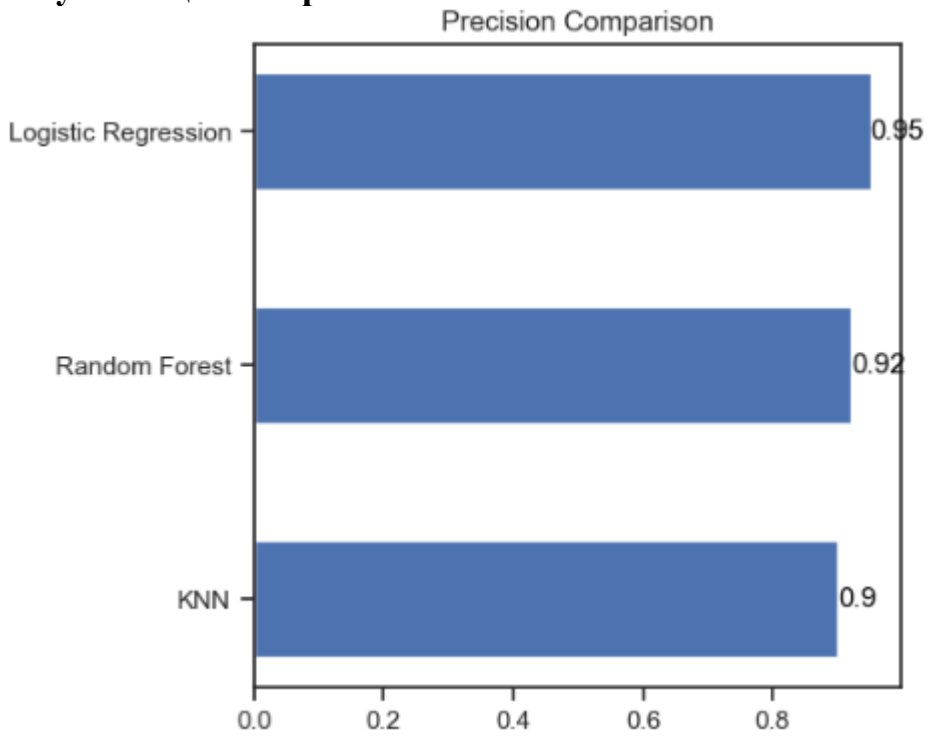
Построение матрицы рассеяния для пар признаков



Построена тепловая карта корреляции



Визуализация метрик



Построение ROC-curve

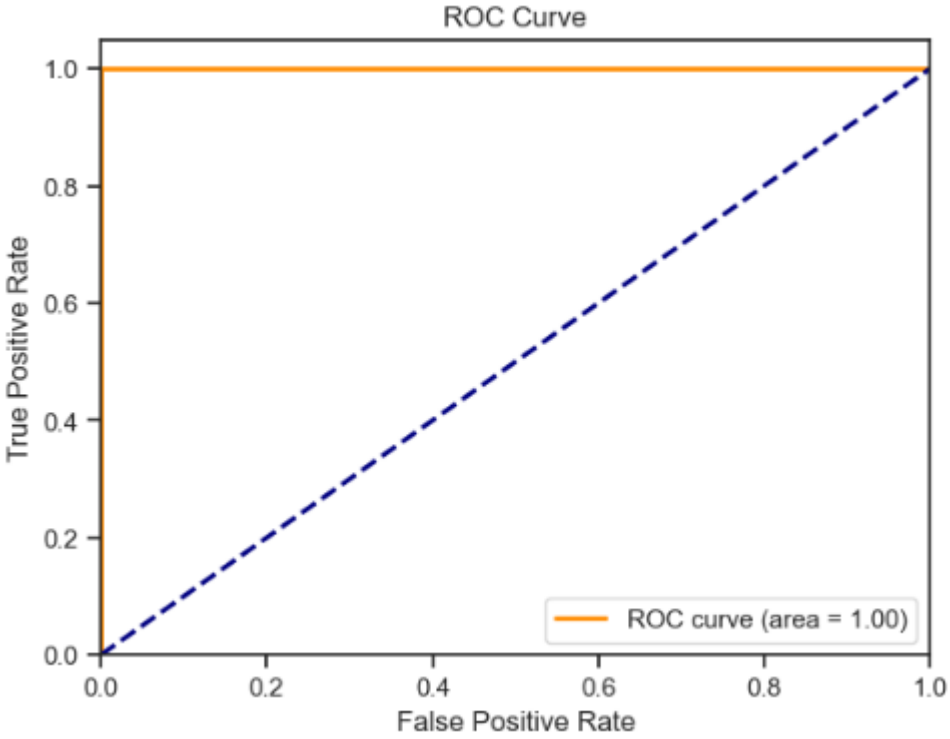


График классификаций

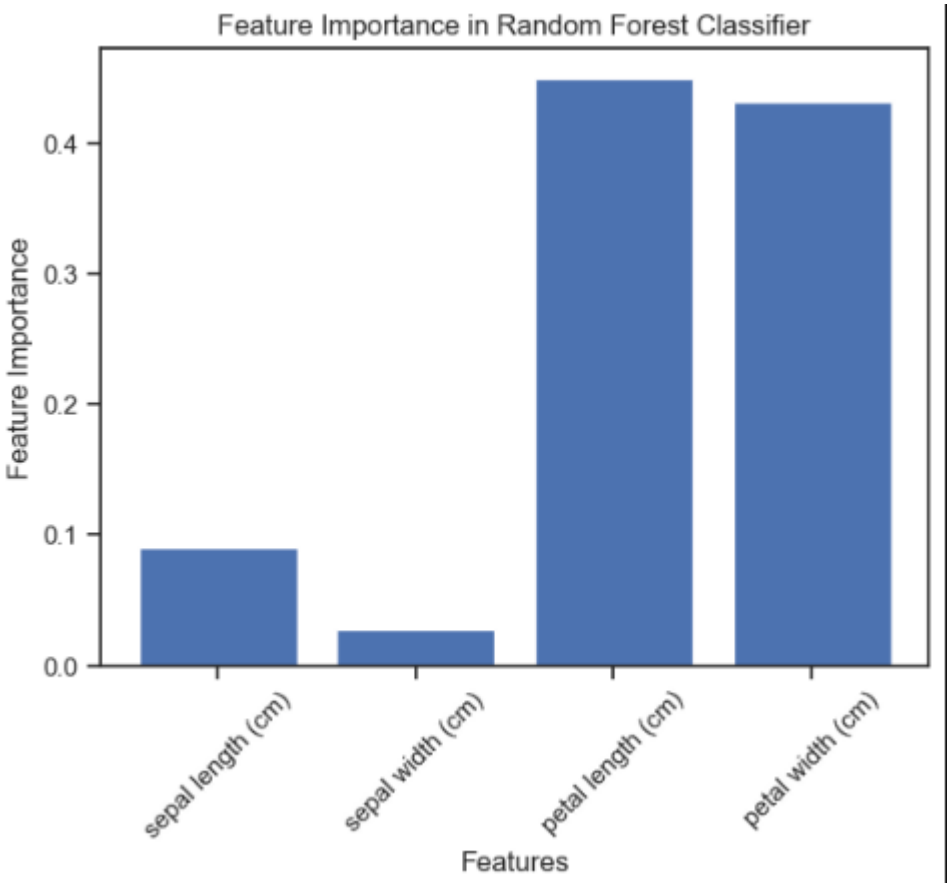


График метрики MAE

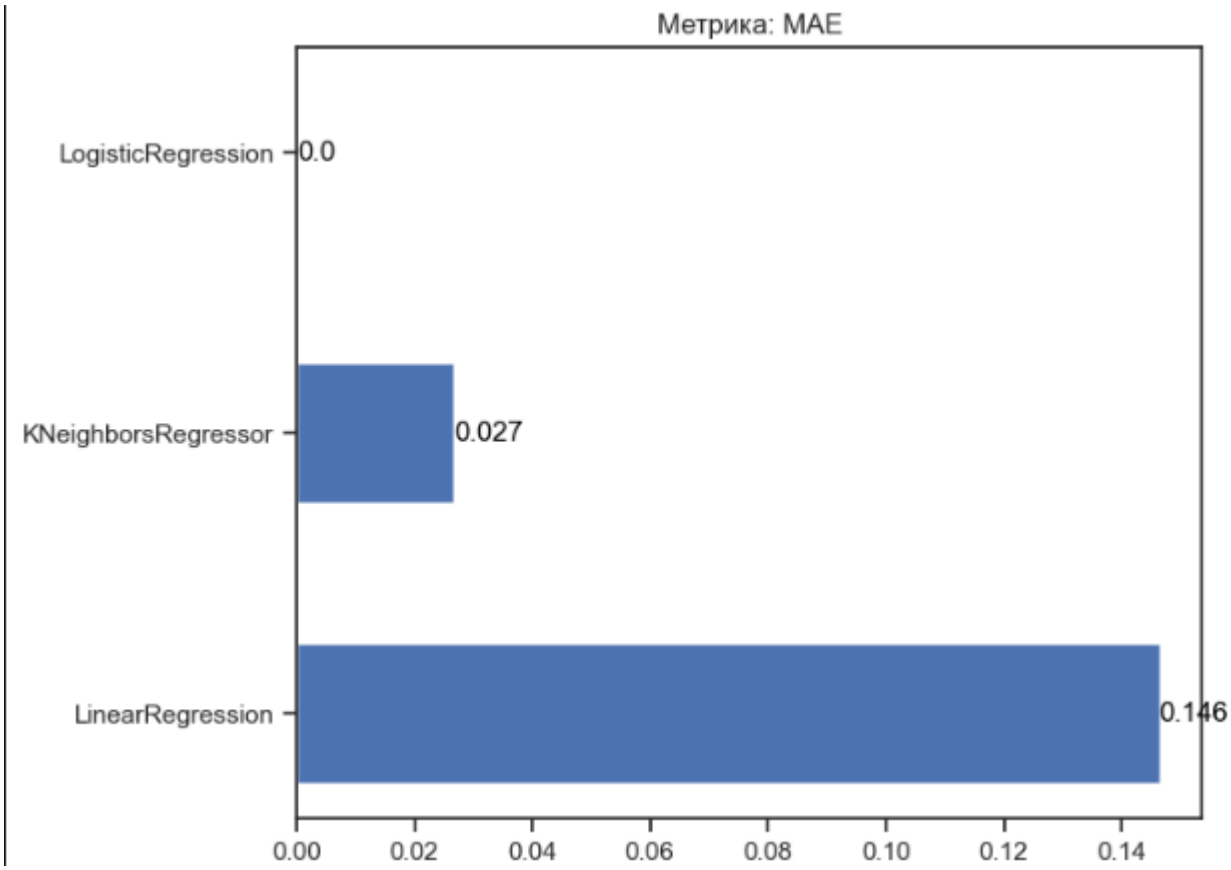


График метрики MSE

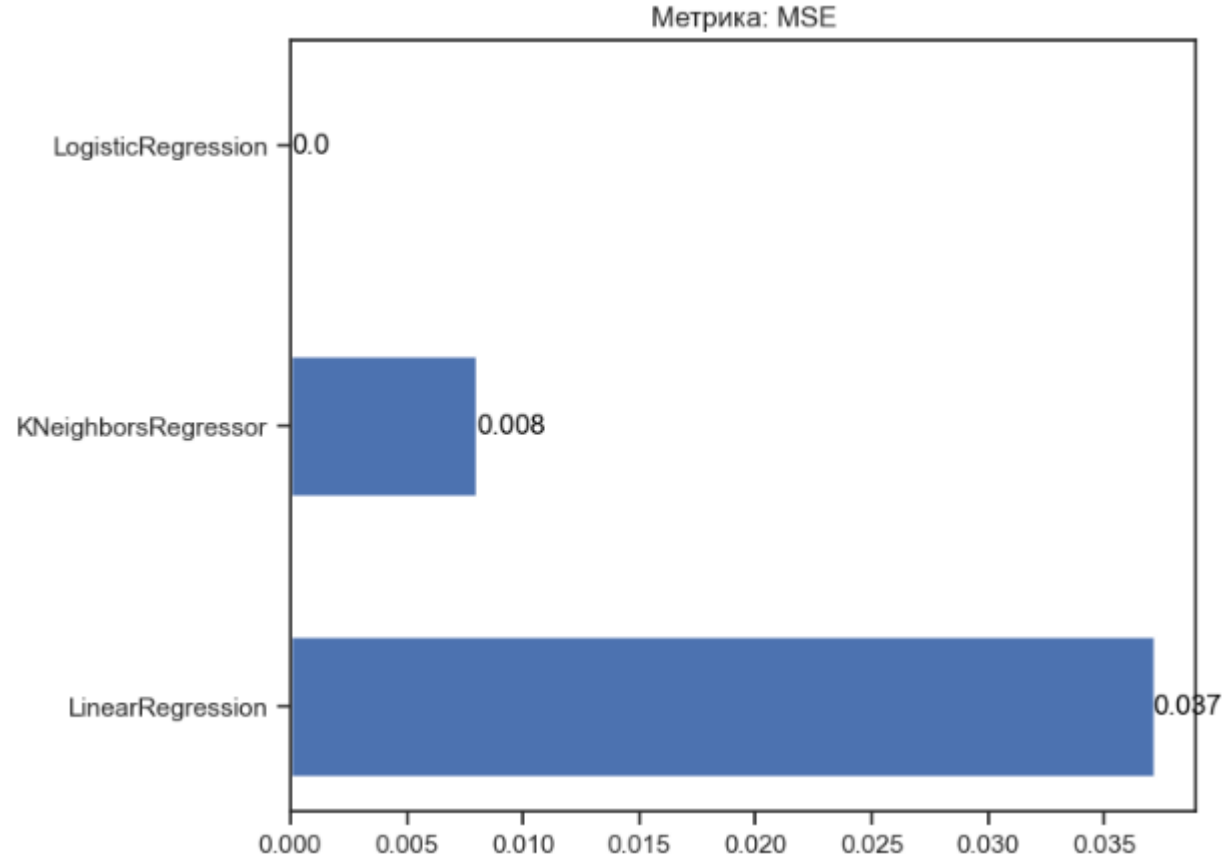
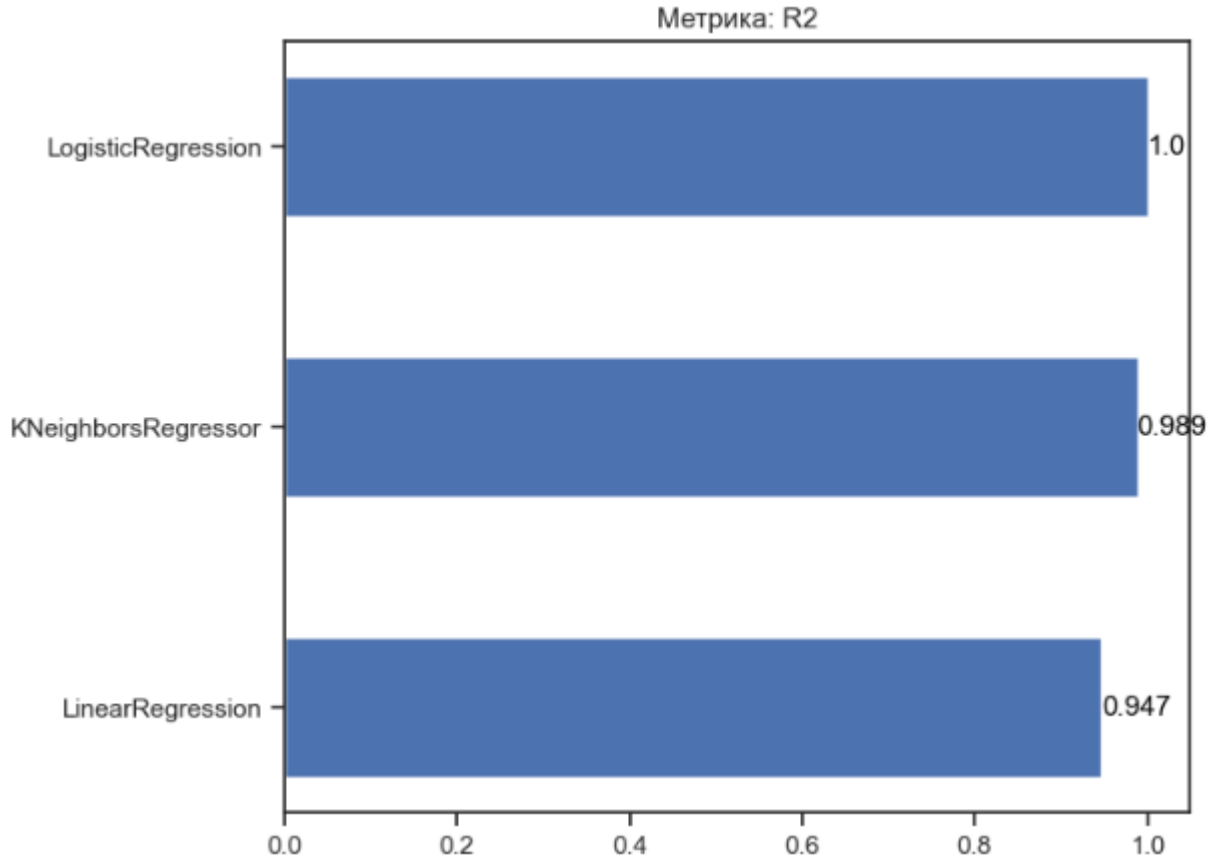


График метрики R2



Источники

- codingfamily.net/machine-learning/a-simple-machine-learning-process-example-supervised-just-to-start-with/
- www.cloudiqtech.com/machine-learning-an-introduction/
- github.com/AnityaGan9urde/would-you-survive-titanic-2.0
- www.yourdatateacher.com/2022/06/06/which-models-are-interpretable/
- github.com/AijajKhan/Digit-Recognizer-by-Support-Vector-Machines
- github.com/KrishGupta-rgb/headbrain