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)»

# ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

**по курсу**

«Data Science»

по теме:

ПРОГНОЗИРОВАНИЕ КОНЕЧНЫХ СВОЙСТВ НОВЫХ МАТЕРИАЛОВ (КОМПОЗИЦИОННЫХ МАТЕРИАЛОВ)

Слушатель Перебатова Надежда Петровна

Москва, 2022

СОДЕРЖАНИЕ

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 1 | Аналитическая часть ……………………………………………………… | | | 3 |
|  | 1.1 | Постановка задачи …………………………………………………… | | 3 |
|  | 1.2 | Описание используемых методов …………………………………... | | 4 |
|  | 1.3 | Разведочный анализ данных ………………………………………… | | 5 |
| 2 | Практическая часть ……………………………………………………….. | | | 12 |
|  | 2.1 | Предобработка данных ……………………………………………… | | 12 |
|  | 2.2 | Построение и обучение моделей …………………………………… | | 15 |
|  |  | 2.2.1 | Линейная регрессия (Linear Regression) …………………….. | 16 |
|  |  | 2.2.2 | Случайный лес (Random Forest Regressor) ………………….. | 17 |
|  |  | 2.2.3 | Градиентный бустинг (Gradient Boosting Regressor) ………. | 19 |
|  |  | 2.2.4 | Сводные результаты по моделям ……………………………. | 21 |
|  |  | 2.2.5 | Нейронная сеть ……………………………………………….. | 22 |
| ЗАКЛЮЧЕНИЕ ……………………………………………………………….. | | | | 24 |
| СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ ………………………… | | | | 25 |
| Репозиторий в GitHub ………………………………………………………… | | | | 26 |

# Аналитическая часть

* 1. **Постановка задачи**

Тема: Прогнозирование конечных свойств новых материалов (композиционных материалов).

Описание: Композиционные материалы – это искусственно созданные материалы, состоящие из нескольких других с четкой границей между ними. Композиты обладают теми свойствами, которые не наблюдаются у компонентов по отдельности. При этом композиты являются монолитным материалом, т. е. компоненты материала неотделимы друг от друга без разрушения конструкции в целом. Яркий пример композита – железобетон. Бетон прекрасно сопротивляется сжатию, но плохо растяжению. Стальная арматура внутри бетона компенсирует его неспособность сопротивляться сжатию, формируя тем самым новые, уникальные свойства. Современные композиты изготавливаются из других материалов: полимеры, керамика, стеклянные и углеродные волокна, но данный принцип сохраняется. У такого подхода есть и недостаток: даже если мы знаем характеристики исходных компонентов, определить характеристики композита, состоящего из этих компонентов, достаточно проблематично. Для решения этой проблемы есть два пути: физические испытания образцов материалов, или прогнозирование характеристик. Суть прогнозирования заключается в симуляции представительного элемента объема композита, на основе данных о характеристиках входящих компонентов (связующего и армирующего компонента).

На входе имеются данные о начальных свойствах компонентов композиционных материалов (количество связующего, наполнителя, температурный режим отверждения и т. д.). На выходе необходимо спрогнозировать ряд конечных свойств получаемых композиционных материалов. Кейс основан на реальных производственных задачах Центра НТИ

«Цифровое материаловедение: новые материалы и вещества» (структурное подразделение МГТУ им. Н.Э. Баумана).

Актуальность: Созданные прогнозные модели помогут сократить количество проводимых испытаний, а также пополнить базу данных материалов возможными новыми характеристиками материалов, и цифровыми двойниками новых композитов.

Датасет со свойствами композитов. Объединение делать по индексу тип объединения INNER.

# Описание используемых методов

Поставленная задача относится к машинному обучению с учителем, задача регрессии.

Анализ, предобработка данных, построение моделей выполнены посредством языка программирования Python.

Обучение моделей для прогноза модуля упругости при растяжении и прочности при растяжении были построены следующие модели:

* Линейная регрессия (Linear Regression),
* Случайный лес (Random Forest Regressor),
* Градиентный бустинг (Gradient Boosting Regressor).

В качестве нейронной сети, которая будет рекомендовать соотношение матрица-наполнитель, построена простая модель глубокого обучения с помощью библиотеки Keras и функции активации ReLu.

Для оценки эффективности используемых методов машинного обучения и нейронных сетей использованы следующие показатели:

* MAE – средняя абсолютная ошибка,
* MSE – среднеквадратическая ошибка,
* RMSE – корень из среднеквадратической ошибки.

# Разведочный анализ данных

Разведочный анализ данных (Exploratory Data Analysis, EDA) представляет из себя предварительное исследование датасета с целью определения его основных характеристик, взаимосвязей между признаками, а также сужения набора методов, используемых для создания модели машинного обучения.

В основу настоящего исследования легли два датасета со свойствами композитов – X\_bp.xlsx (содержащий 1023 строки и 11 колонок) и X\_nup.xlsx (содержащий 1040 строк и 4 колонки).

Для повышения качества визуализации проводимого анализа переменным были присвоены условные обозначения, как это представлено в таблице 1.

Таблица 1 – Условные обозначения переменных

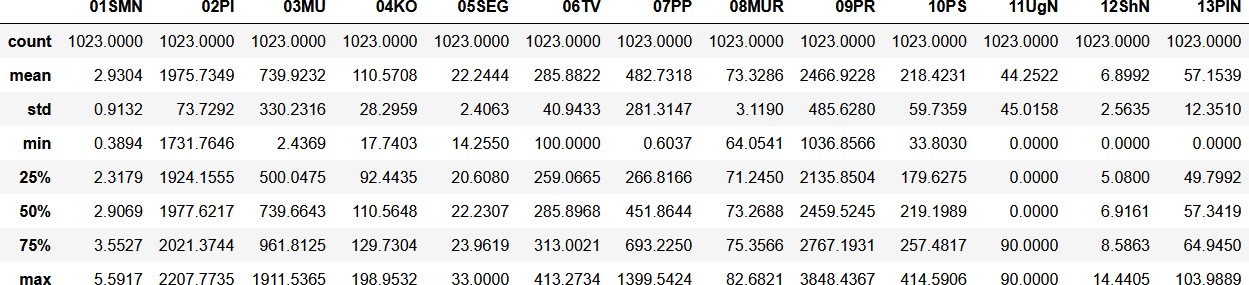
|  |  |  |
| --- | --- | --- |
| Условное  обозначение | Переменная | Датасет |
| 01SMN | Соотношение матрица-наполнитель | X\_bp.xlsx |
| 02Pl | Плотность, кг/м3 |
| 03MU | Модуль упругости, ГПа |
| 04KO | Количество отвердителя, м.% |
| 05SEG | Содержание эпоксидных групп, %\_2 |
| 06TV | Температура вспышки, С\_2 |
| 07PP | Поверхностная плотность, г/м2 |
| 08MUR | Модуль упругости при растяжении, Гпа |
| 09PR | Прочность при растяжении, Мпа |
| 10PS | Потребление смолы, г/м2 |
| 11UgN | Угол нашивки, град | X\_nup.xlsx |
| 12ShN | Шаг нашивки |
| 13PlN | Плотность нашивки |

Представленные датасеты в соответствии с заданными условиями были объединены по индексу тип объединения inner в один (unite\_df), содержащий 1023 строки и 13 колонок. Таким образом, 17 строк исходной информации датасета X\_nup.xlsx были исключены из дальнейшего анализа в связи с тем, что в датасете X\_bp.xlsx соответствующие им строки не имелись.

Проверка объединенного датасета на наличие пропусков и типа данных показала, что все данные являются количественными (float и int), пропусков нет, очистка не требуется.

Для объединенного датасета с помощью метода describe получены основные статистические характеристики по каждой переменной (таблица 2), построены для каждой из переменных гистограммы распределения (рисунок 1), диаграммы «ящик с усами» (рисунок 2), попарные графики рассеяния точек (матрица диаграмм рассеяния) (рисунок 3), а также корреляционная матрица в виде тепловой карты (таблица 3).

Таблица 2 – Основные статистические характеристики по каждой переменной



Примечание:

count – количество строк mean – среднее значение

std – стандартное отклонение min – минимальное значение 25% – первый квартиль

50% – второй квартиль, медиана 75% – третий квартиль

max – максимальное значение

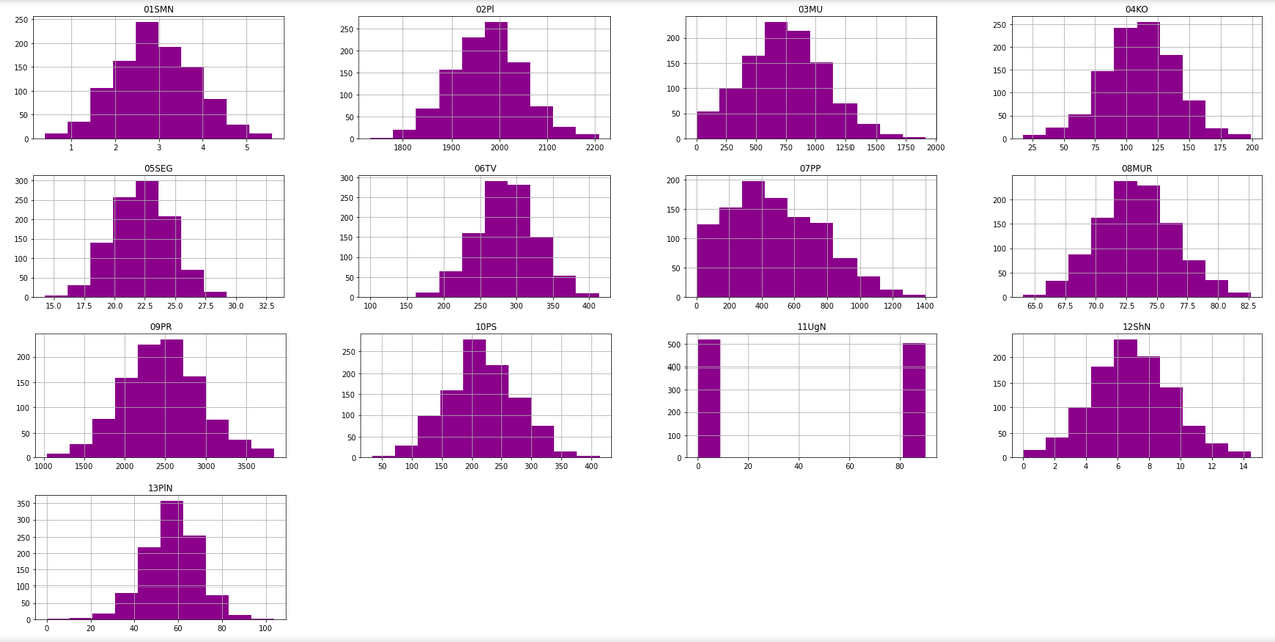


Рисунок 1 – Гистограммы распределения каждой из переменных

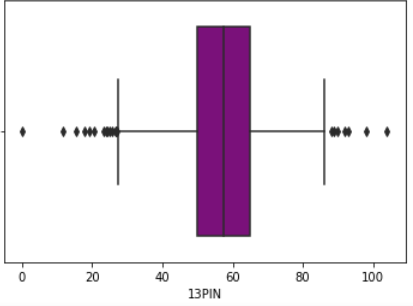
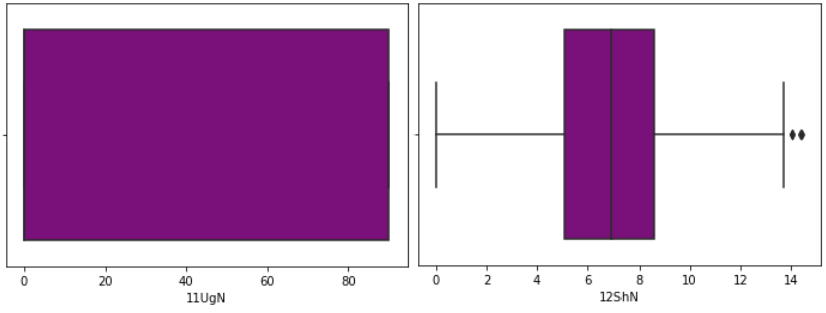
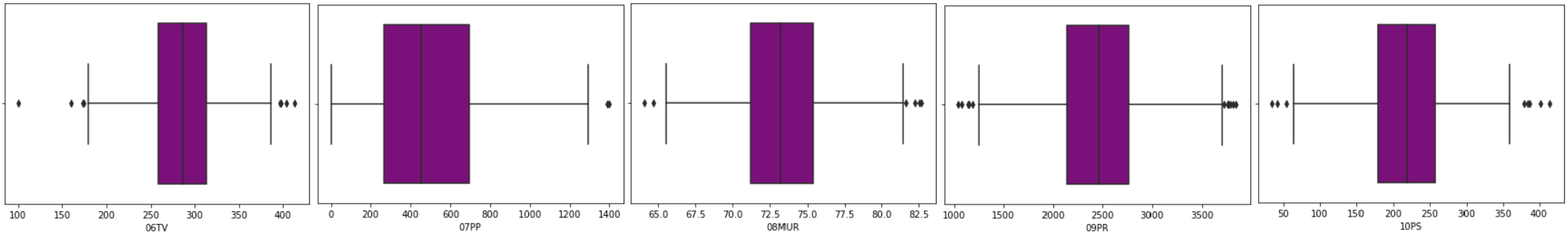
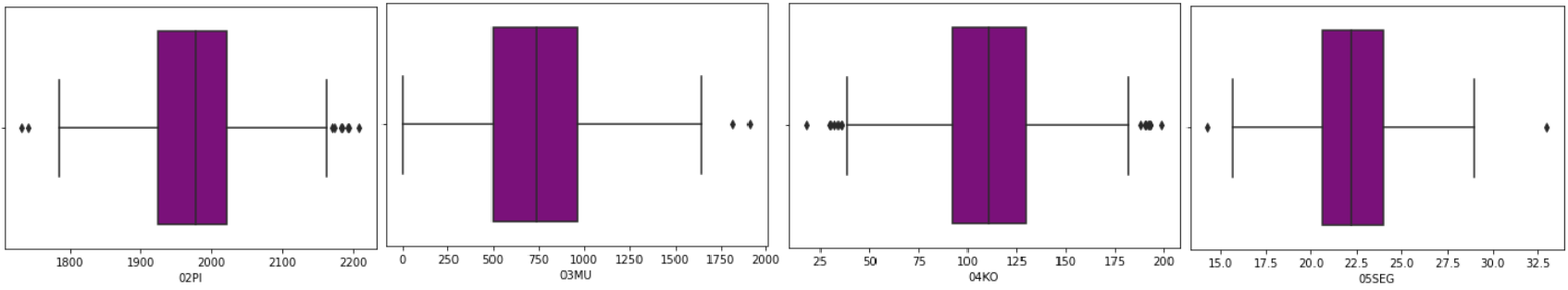
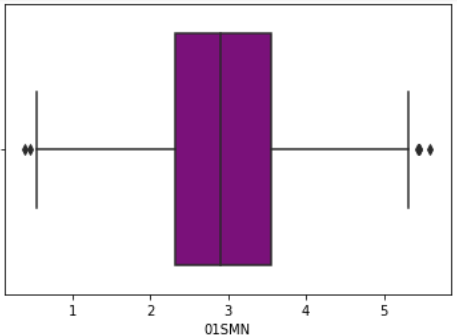


Рисунок 2 – Диаграммы «ящик с усами» для каждой из переменных

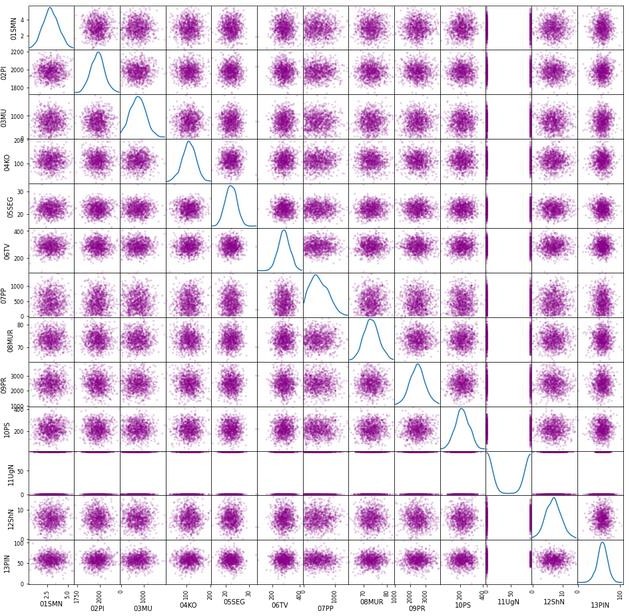
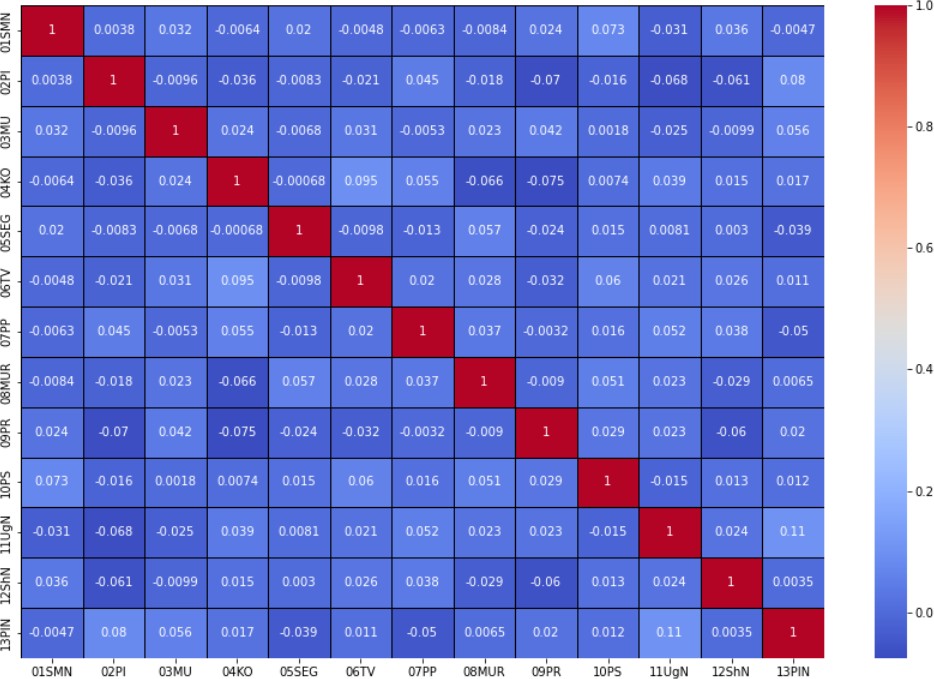


Рисунок 3 – Попарные графики рассеяния точек (матрица диаграмм рассеяния)

Как видно из данных рисунка 3, попарное сравнение признаков зависимости не выявило.

Таблица 3 – Корреляционная матрица в виде тепловой карты



Корреляционная матрица зависимости не выявила.

# Практическая часть

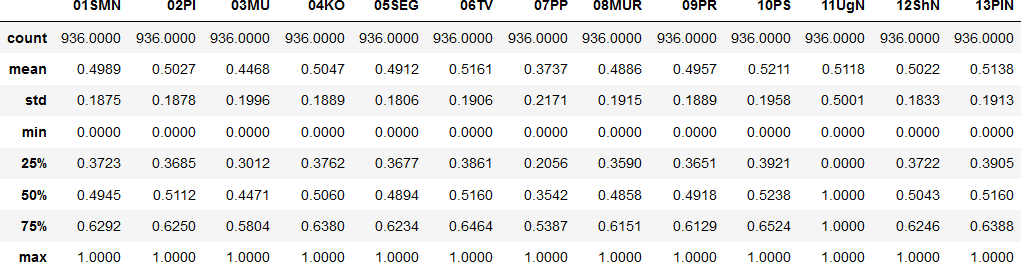
* 1. **Предобработка данных**

С помощью метода межквартильного расстояния было найдено 64 выброса. После удаления выбросов в датасете осталось 936 строк и 13 колонок.

Далее данные были нормализованы (norm\_df).

Для нормализованного датасета с помощью метода describe получены основные статистические характеристики по каждой переменной (таблица 4), построены для каждой из переменных диаграммы «ящик с усами» (рисунок 4), а также корреляционная матрица в виде тепловой карты (таблица 5).

Таблица 4 – Основные статистические характеристики по каждой переменной после удаления выбросов и нормализации



Примечание:

count – количество строк mean – среднее значение

std – стандартное отклонение min – минимальное значение 25% – первый квартиль

50% – второй квартиль, медиана 75% – третий квартиль

max – максимальное значение

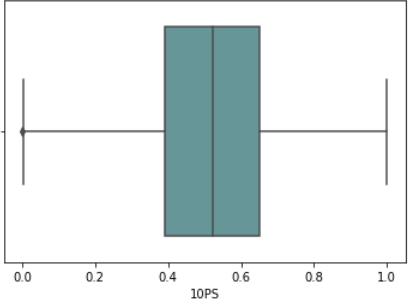
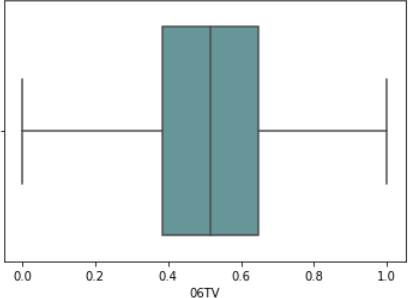
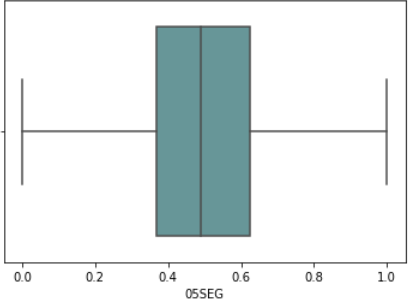
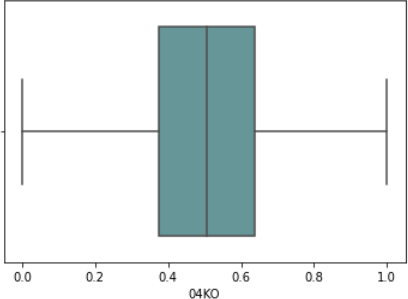
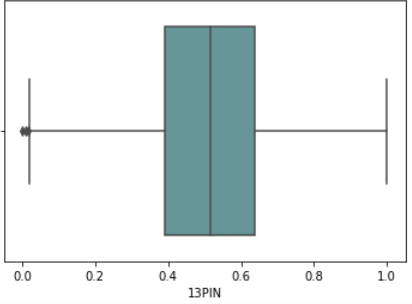
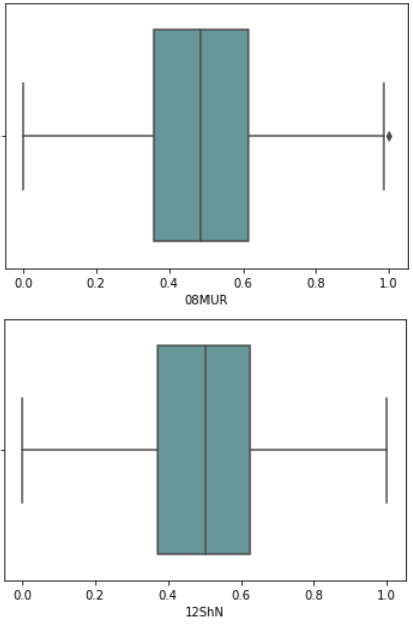
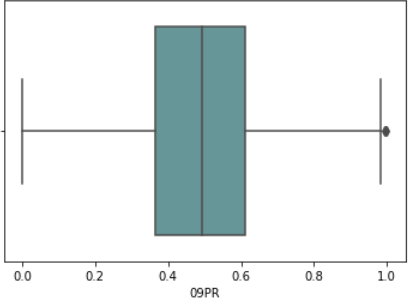
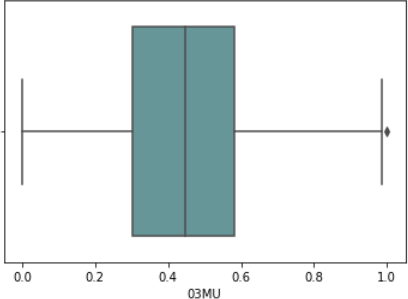
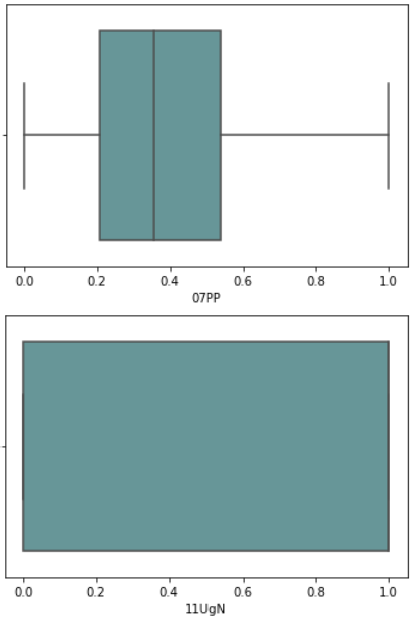
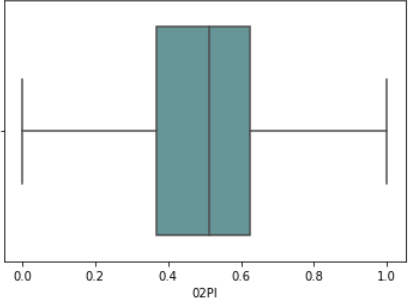
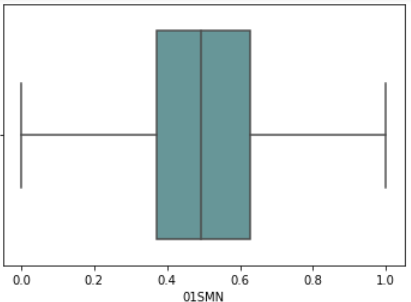
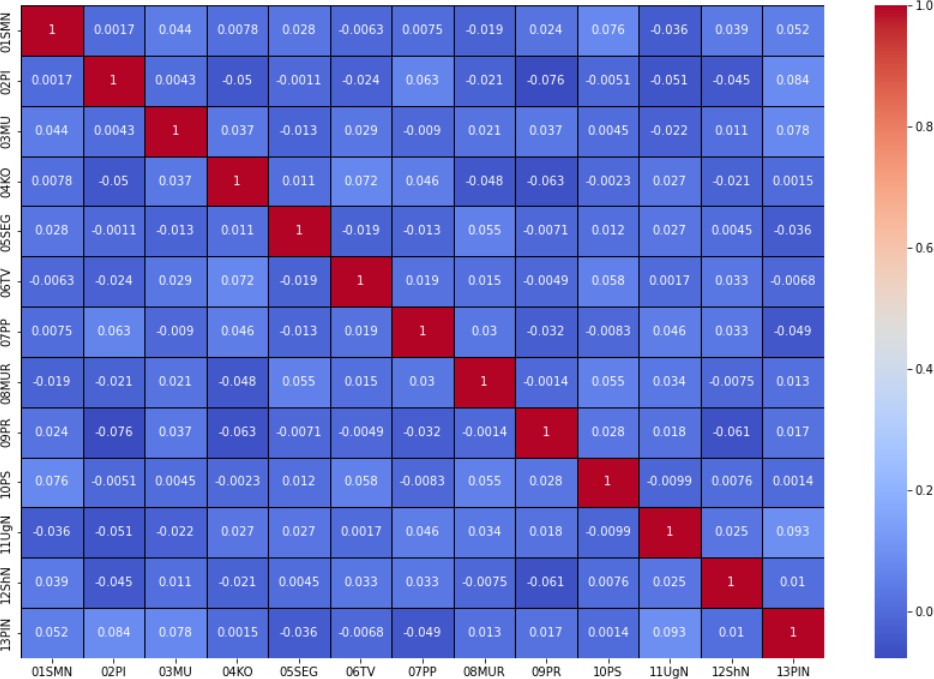


Рисунок 4 – Диаграммы «ящик с усами» для каждой из переменных после удаления выбросов и нормализации



Таблица 5 – Корреляционная матрица в виде тепловой карты после удаления выбросов и нормализации

В результате удаления выбросов корреляция между признаками несколько изменилась, но по-прежнему практически отсутствует.

# Построение и обучение моделей

Для прогноза двух выходных параметров – «Модуль упругости при растяжении» (08MUR) и «Прочность при растяжении» (09PR) было проведено построение, обучение и тестирование трех моделей:

* Линейная регрессия (Linear Regression),
* Случайный лес (Random Forest Regressor),
* Градиентный бустинг (Gradient Boosting Regressor).

Для построения моделей в качестве датасета использована нормализованная выборка (norm\_df), которая была разделена на тестовую (30%) и тренировочную (70%).

# Линейная регрессия (Linear Regression)

learn.

Построение линейной регрессии произведено с помощью библиотеки scikit-

Прогноз модуля упругости при растяжении. В таблице 6 представлены

результаты обучения линейной регрессии, округленные до 6 знаков после запятой.

Таблица 6 – Результаты обучения линейной регрессии (08MUR)

|  |  |  |  |
| --- | --- | --- | --- |
| Модель | MAE | MSE | RMSE |
| Линейная регрессия (Linear Regression) | 0,160464 | 0,037892 | 0,194658 |

График совпадения предсказанных и актуальных значений и гистограмма распределения ошибки продемонстрированы на рисунках 5 и 6 соответственно.

|  |  |
| --- | --- |
| Рисунок 5 – График совпадения предсказанных и актуальных значений  (08MUR) | Рисунок 6 – Гистограмма распределения ошибки (08MUR) |

Прогноз прочности при растяжении. В таблице 7 представлены результаты обучения линейной регрессии, округленные до 6 знаков после запятой.

Таблица 7 – Результаты обучения линейной регрессии (09PR)

|  |  |  |  |
| --- | --- | --- | --- |
| Модель | MAE | MSE | RMSE |
| Линейная регрессия (Linear Regression) | 0,150596 | 0,034967 | 0,186994 |

График совпадения предсказанных и актуальных значений и гистограмма распределения ошибки продемонстрированы на рисунках 7 и 8 соответственно.

|  |  |
| --- | --- |
| Рисунок 7 – График совпадения предсказанных и актуальных значений  (09PR) | Рисунок 8 – Гистограмма распределения ошибки (09PR) |

# Случайный лес (Random Forest Regressor)

Построение модели Случайный лес произведено с помощью библиотеки

scikit-learn.

Для данной модели проведен поиск гиперпараметров с помощью поиска по сетке с перекрестной проверкой методом GridSearchCV, количество блоков равно 10.

Прогноз модуля упругости при растяжении. В таблице 8 представлены результаты обучения модели Случайный лес, округленные до 6 знаков после запятой.

Таблица 8 – Результаты обучения модели Случайный лес (08MUR)

|  |  |  |  |
| --- | --- | --- | --- |
| Модель | MAE | MSE | RMSE |
| Случайный лес  (Random Forest Regressor) | 0,162886 | 0,039740 | 0,199348 |

График совпадения предсказанных и актуальных значений и гистограмма распределения ошибки продемонстрированы на рисунках 9 и 10 соответственно.

|  |  |
| --- | --- |
| Рисунок 9 – График совпадения предсказанных и актуальных значений  (08MUR) | Рисунок 10 – Гистограмма распределения ошибки (08MUR) |

Прогноз прочности при растяжении. В таблице 9 представлены результаты обучения модели Случайный лес, округленные до 6 знаков после запятой.

Таблица 9 – Результаты обучения модели Случайный лес (09PR)

|  |  |  |  |
| --- | --- | --- | --- |
| Модель | MAE | MSE | RMSE |
| Случайный лес  (Random Forest Regressor) | 0,152456 | 0,035131 | 0,187433 |

График совпадения предсказанных и актуальных значений и гистограмма распределения ошибки продемонстрированы на рисунках 11 и 12 соответственно.

|  |  |
| --- | --- |
| Рисунок 11 – График совпадения предсказанных и актуальных значений  (09PR) | Рисунок 12 – Гистограмма распределения ошибки (09PR) |

# Градиентный бустинг (Gradient Boosting Regressor)

Построение модели Градиентный бустинг произведено с помощью библиотеки scikit-learn.

Для данной модели проведен поиск гиперпараметров с помощью поиска по сетке с перекрестной проверкой методом GridSearchCV, количество блоков равно 10.

Прогноз модуля упругости при растяжении. В таблице 10 представлены результаты обучения модели Градиентный бустинг, округленные до 6 знаков после запятой.

Таблица 10 – Результаты обучения модели Градиентный бустинг (08MUR)

|  |  |  |  |
| --- | --- | --- | --- |
| Модель | MAE | MSE | RMSE |
| Градиентный бустинг (Gradient Boosting Regressor) | 0,173699 | 0,045519 | 0,213352 |

График совпадения предсказанных и актуальных значений и гистограмма распределения ошибки продемонстрированы на рисунках 13 и 14 соответственно.

|  |  |
| --- | --- |
| Рисунок 13 – График совпадения предсказанных и актуальных значений  (08MUR) | Рисунок 14 – Гистограмма распределения ошибки (08MUR) |

Прогноз прочности при растяжении. В таблице 11 представлены результаты обучения модели Градиентный бустинг, округленные до 6 знаков после запятой.

Таблица 11 – Результаты обучения модели Градиентный бустинг (09PR)

|  |  |  |  |
| --- | --- | --- | --- |
| Модель | MAE | MSE | RMSE |
| Градиентный бустинг (Gradient Boosting Regressor) | 0,160442 | 0,038558 | 0,196361 |

График совпадения предсказанных и актуальных значений и гистограмма распределения ошибки продемонстрированы на рисунках 15 и 16 соответственно.

|  |  |
| --- | --- |
| Рисунок 15 – График совпадения предсказанных и актуальных значений  (09PR) | Рисунок 16 – Гистограмма распределения ошибки (09PR) |

# Сводные результаты по моделям

Сводные результаты обучения моделей по прогнозу модуля упругости при растяжении представлены в таблице 12.

Таблица 12 – Сводные результаты обучения моделей (08MUR)

|  |  |  |  |
| --- | --- | --- | --- |
| Модель | MAE | MSE | RMSE |
| Линейная регрессия (Linear Regression) | 0,160464 | 0,037892 | 0,194658 |
| Случайный лес  (Random Forest Regressor) | 0,162886 | 0,039740 | 0,199348 |
| Градиентный бустинг (Gradient Boosting Regressor) | 0,173699 | 0,045519 | 0,213352 |

Сводные результаты обучения моделей по прогнозу прочности при растяжении

представлены в таблице 13.

Таблица 13 – Сводные результаты обучения моделей (09PR)

|  |  |  |  |
| --- | --- | --- | --- |
| Модель | MAE | MSE | RMSE |
| Линейная регрессия (Linear Regression) | 0,150596 | 0,034967 | 0,186994 |
| Случайный лес  (Random Forest Regressor) | 0,152456 | 0,035131 | 0,187433 |
| Градиентный бустинг (Gradient Boosting Regressor) | 0,160442 | 0,038558 | 0,196361 |

Все модели оценки не смогли выявить зависимости. Полученные результаты по всем построенным моделям можно оценить как неудовлетворительные.

Самую низкую ошибку по метрикам регрессии показала линейная регрессия, это один из параметров, по которому можно оценить эту модель как наиболее подходящую.

# Нейронная сеть

Для рекомендации параметра «соотношение матрица-наполнитель» построена простая модель глубокого обучения с помощью библиотеки Keras и функции активации ReLu.

Архитектура нейронной сети может быть описана следующим образом: входной слой – 64 нейрона; три скрытых уровня – по 64 нейрона каждый; выходной слой – 1 нейрон.

Количество эпох, показывающих сколько раз модель подвергается воздействию обучения, было определено равным 100.

В качестве датасета использована нормализованная выборка (norm\_df), которая была разделена на тестовую (30%) и тренировочную (70%).

В качестве оценки результатов обучения нейронной сети определены следующие метрики: средняя квадратическая ошибка (MSE) и средняя абсолютная ошибка (MAE), рассчитанные на тестовом наборе данных и округленные до 6 знаков после запятой (таблица 14).

Таблица 14 – Результаты обучения нейронной сети

|  |  |  |
| --- | --- | --- |
| Модель | MAE | MSE |
| Нейронная сеть | 0,212306 | 0,067789 |

Нейронная сеть также не смогла выявить зависимости, результат неудовлетворительный.

# ЗАКЛЮЧЕНИЕ

Ни одна из представленных в исследовании моделей не смогла выявить зависимости и показать удовлетворительный результат.

Это может быть связано с несколькими факторами:

* проблемы с набором данных,
* проблемы нормализация данных,
* проблемы реализации модели,
* проблемы обучения модели.

# СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Брюс П. Практическая статистика для специалистов Data Science: Пер. с англ. – СПб.: БХВ-Петербург, 2018. – 304 с.
2. Грас Д. Data Science. Наука о данных с нуля: Пер. с англ. – 2-е изд., перераб. И доп. – Спб.: БХВ-Петербург, 2021. – 416 с.
3. Гудфеллоу Я., Бенджио И., Курвилль А. Глубокое обучение / пер. с анг.

А.А. Слинкина. – 2-е изд., испр. – М.: ДМК Пресс, 2018. – 652 с.

1. Дауни А. Основы Python. Научитесь думать как программист / пер. с англ. С. Черникова ; [науч. ред. А. Родионов]. – Москва: Манн, Иванов и Фербер, 2021.

– 304 с.

1. Pandas [Электронный ресурс] : – Режим доступа: <http://www.mysql.ru/docs/man/Reference.html>(дата обращения: 20.04.2022).
2. Matplotlib [Электронный ресурс] : – Режим доступа: https://matplotlib.org/ (дата обращения: 20.04.2022).
3. Scikit-learn [Электронный ресурс] : – Режим доступа: https://scikit- learn.org/stable/

(дата обращения: 20.04.2022).

1. Data scientist [Электронный ресурс] : – Режим доступа: https://data- scientists.ru/ (дата обращения: 20.04.2022).
2. Машинное обучение, нейронные сети, искусственный интеллект [Электронный ресурс] : – Режим доступа: https://[www.machinelearningmastery.ru](http://www.machinelearningmastery.ru/) (дата обращения: 20.04.2022).

# Репозиторий в GitHub

Результаты исследования размещены в репозитории GitHub: https://github.com/Perebatoff/PerebatovaNadejdaProject