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»

Прогнозирование конечных свойств

новых материалов (композиционных материалов)

Слушатель Бадеева Елена Евгеньевна

Москва, 2022

# Содержание

Содержание 2

Введение 3

1. Аналитическая часть 4

1.1. Постановка задачи 4

1.2. Описание используемых методов 5

1.3. Разведочный анализ данных 5

2. Практическая часть 11

2.1. Предобработка данных 11

2.2. Разработка и обучение модели 12

2.3. Тестирование модели 14

2.4. Нейронная сеть для прогноза соотношения матрица - наполнитель 14

2.5. Разработка приложения 15

2.6. Репозиторий и результаты 16

Заключение 17

Библиографический список 18

Приложения 19

# Введение

Пояснительная записка подготовлена в рамках выпускной квалификационной работы по курсу «Data Science» на тему «Прогнозирование конечных свойств новых материалов (композиционных материалов)».

Решаемая задача основана на реальных производственных задачах Центра НТИ «Цифровое материаловедение: новые материалы и вещества» (структурное подразделение МГТУ им. Н. Э. Баумана).

Предполагается, что целью данной работы является проведение исследования с использованием методов, изученных на курсе «Data Science».

В результате работы формируется комплект документов, в частности - код в Jupiter notebook, графическая визуализация, презентация для защиты, данная пояснительная записка и другие материалы.

# 1. Аналитическая часть

## 1.1. Постановка задачи

Постановка задачи сформирована на основании сведений и требований к выпускной квалификационной работе по курсу «Data Science».

Dataset представлен в двух файлах формата excel:

* Файл X\_nup – Накладки углепластика;
* Файл X\_bp – Базальтопластик.

В файлах набор данных, состоящий из характеристик компонентов композитных материалов и характеристик производственного процесса. Общий объем dataset - 1040 измерений (строк) для каждой из 13 переменных:

* Угол нашивки, град;
* Шаг нашивки;
* Плотность нашивки;
* Соотношение матрица-наполнитель;
* Плотность, кг/м3;
* Модуль упругости, Гпа;
* Количество отвердителя, м.%;
* Содержание эпоксидных групп,%\_2;
* Температура вспышки, С\_2;
* Поверхностная плотность, г/м2;
* Модуль упругости при растяжении, ГПа;
* Прочность при растяжении, МПа;
* Потребление смолы, г/м2.

Необходимо разработать несколько моделей для прогнозирования следующих переменных:

* Модуль упругости при растяжении, ГПа;
* Прочность при растяжении, МПа.

Для переменной «Соотношение матрица-наполнитель» необходимо разработать нейронную сеть, которая будет рекомендовать соотношение матрица-наполнитель с использованием приложения.

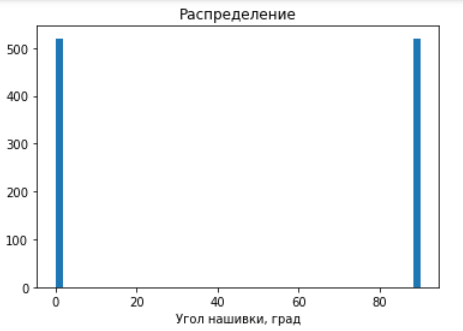
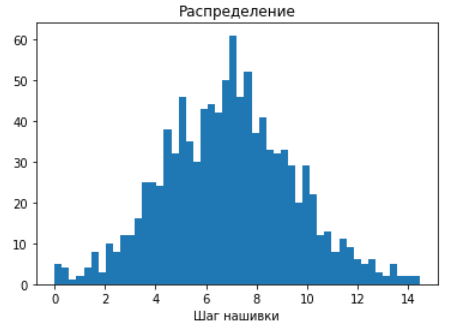
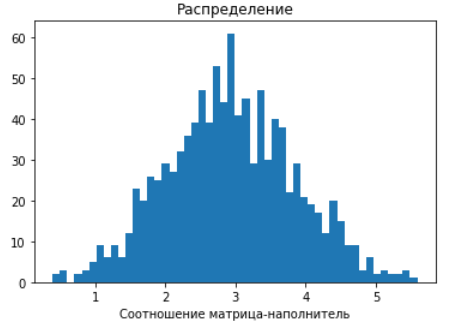
## 1.2. Описание используемых методов

Для реализации прогноза конечных свойств композитных материалов необходимо решить задачу регрессии.

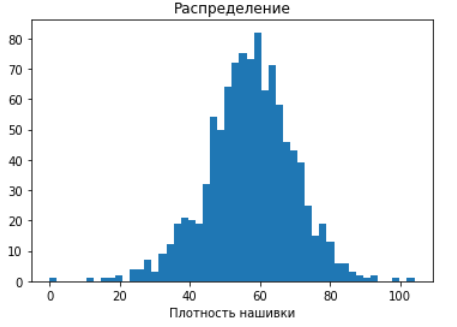
На первом этапе за базовую модель принимается линейная регрессия - tf.keras.Sequential.

На следующем этапе используется многослойный персептрон с последующими экспериментами для выявления оптимальных настроек - keras.Sequential.

## 1.3. Разведочный анализ данных

Для разведочного анализа данных были использованы методы, указанные ниже.

Для построения гистограмм распределения для каждой переменной использовался метод matplotlib.pyplot.



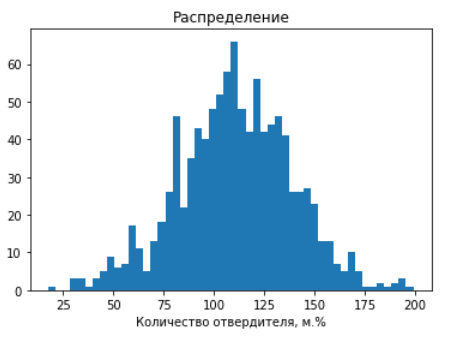
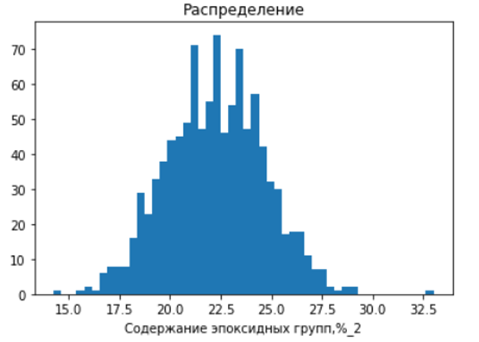
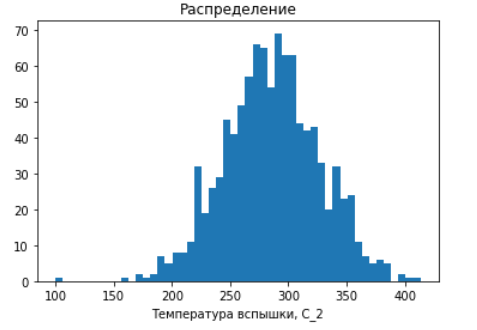
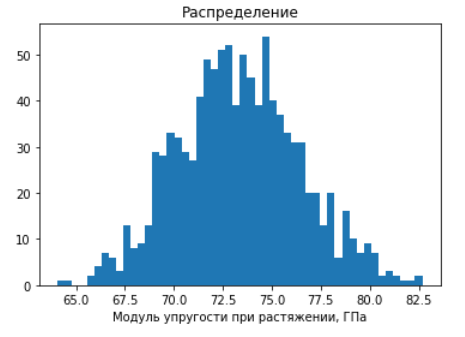
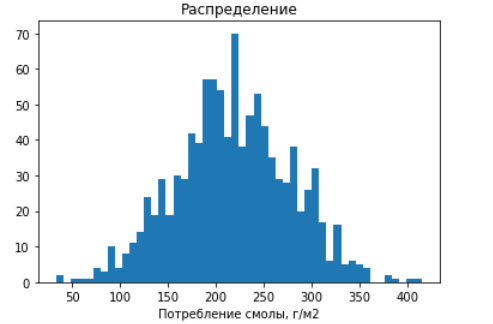
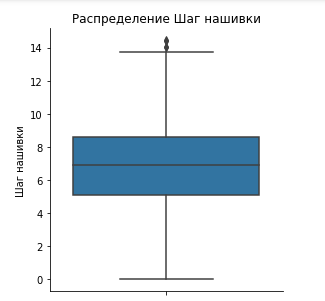
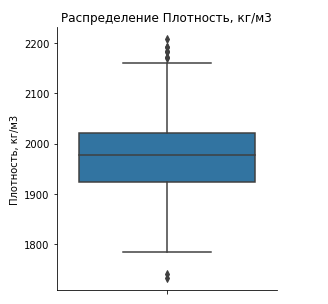
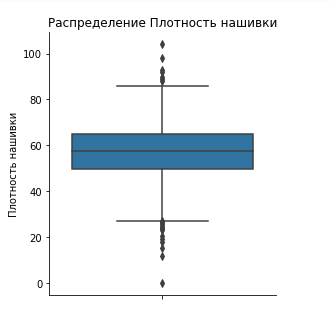
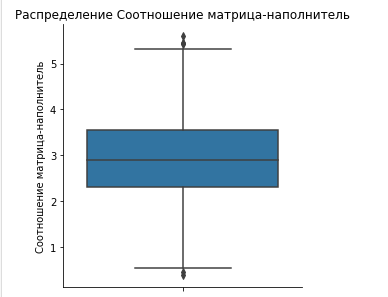
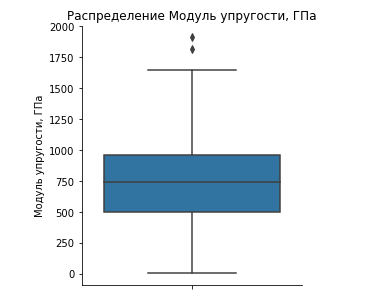
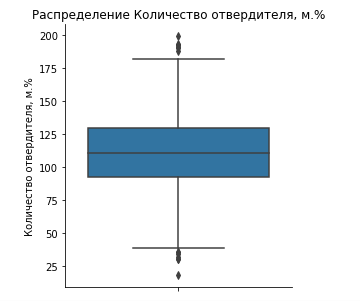
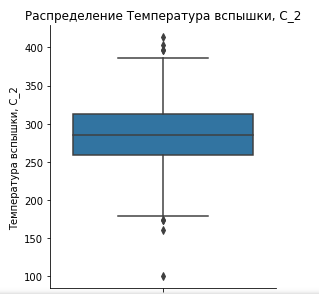
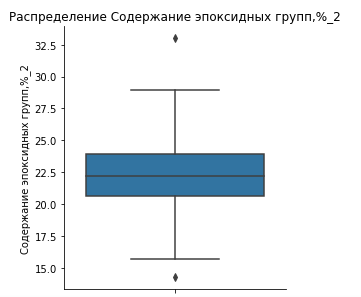


Рисунок 1 - Графики распределения по каждой переменной

Диаграммы «ящик с усами» для каждой переменной - sns.catplot:



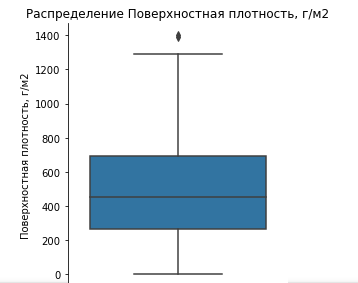
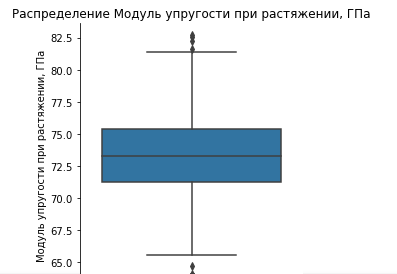
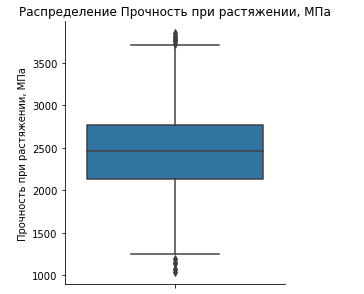
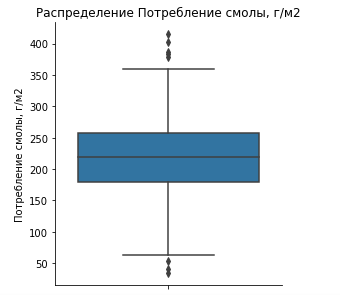


Рисунок 2 - Ящики с усами для каждой переменной

Для определения наличия выбросов в данных использовался метод df.quantile.

Изображение выглядит как текст

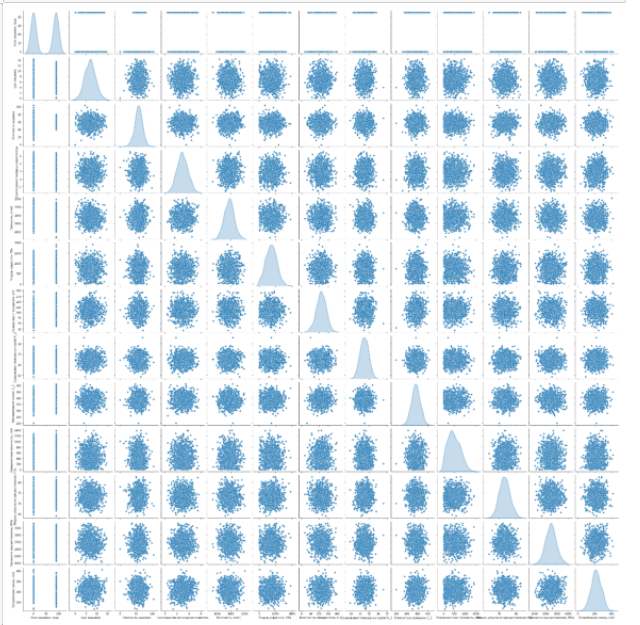
Автоматически созданное описаниеНаличие пропусков было выявлено с использованием df.isna():

Рисунок 3 - Наличие пропусков в каждой переменной

Изображение выглядит как стол

Автоматически созданное описаниеОпределение среднего и медианного значений переменных - df.describe()

Рисунок 4 - Среднее и медианное значение всех переменных

Попарные графики рассеяния точек построены с помощью sns.pairplot.

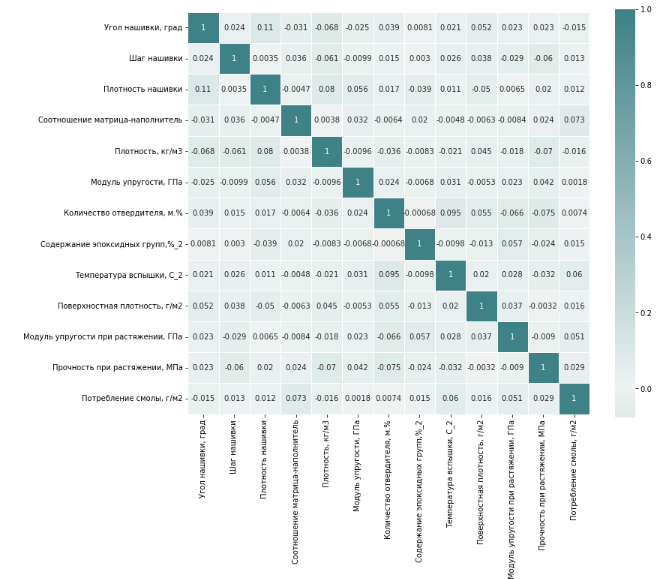
Корреляция признаков - df.corr() и визуализация в виде тепловой карты - sns.heatmap, а также инструмент в deductor.

Рисунок 5 - Корреляция признаков



Рисунок 6 – Корреляция с выходными полями из deductor

# 2. Практическая часть

## 2.1. Предобработка данных

Данные прошли предобработку согласно заданию с применением методов изученных на курсе.

Задача объединения dataset была решена с использованием функция ВПР в Excel.

Изображение выглядит как стол

Автоматически созданное описаниеПропуски удалены df = df.dropna().

Рисунок 7 - Предобработка данных для удаления пропусков

Изображение выглядит как стол

Автоматически созданное описаниеНормализация dataset произведена с использованием - min\_max\_scaler.fit\_transform(df2).

Рисунок 8 - Произведена нормализация dataset

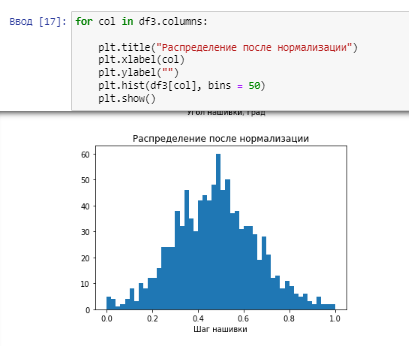
Графики распределения для каждого признака после нормализации –

Рисунок 9 - Пример графика распределения после нормализации

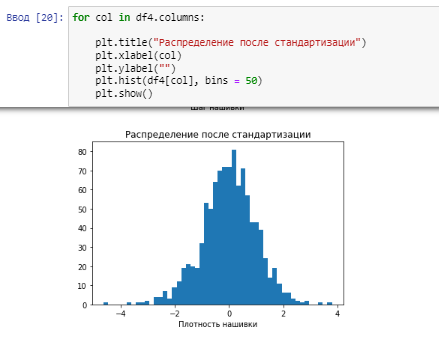
Была произведена стандартизация dataset.

Рисунок 10 - Пример графика распределения после стандартизации

## 2.2. Разработка и обучение модели

Для прогноза модуля упругости при растяжении и прочности при растяжении данные были разделены в соотношении:

- 70% обучение модели;

- 30% тестирование модели.

Были использованы следующие модели:

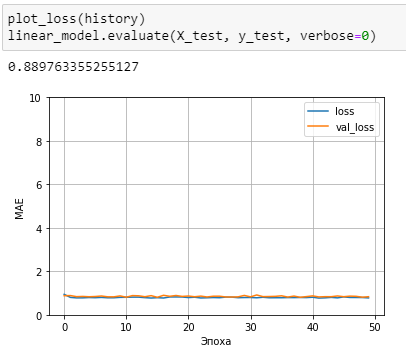
* Модель 1 - линейная регрессия leaner\_model;

Рисунок 11 - Ошибка модели

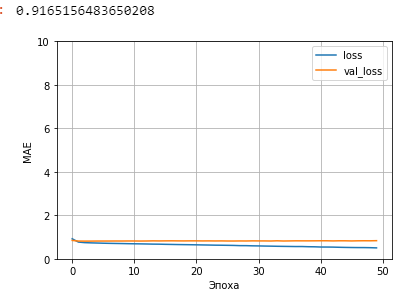
* Модель - многослойный персептрон 2 dnn\_model;

Рисунок 12 - Ошибка модели

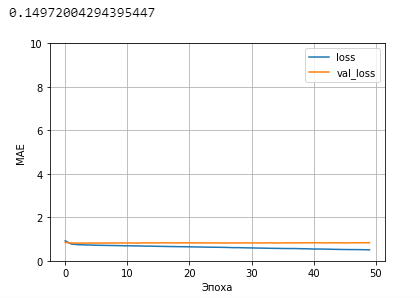
* Модель 3 – проведены эксперименты с многослойным персептроном dnn\_model2.

Рисунок 13 - Ошибка модели

## 2.3. Тестирование модели

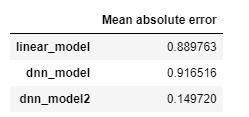
Для каждой модели были определены средние абсолютные ошибки на тренировочной и тестирующей части выборке. Тестовая часть одинаково преобразована для всех моделей (удалены пропуски и проведена нормализация).

Таблица 1 Сравнение моделей

|  |  |  |  |
| --- | --- | --- | --- |
| Модель | dnn\_model2 | dnn\_model | linear\_model |
|  | keras.Sequential | keras.Sequential | tf.keras.Sequential |
| loss | MAE | MAE | MAE |
| activation | sigmoid | relu |  |
| layers.Dense | 200/20/10/2 | 40/40/2 | 2 |
| optimizer | RMSprop | Adam |  |
| layers.Dropout | 0.6 |  |  |

Обоснование выбора модели: для модели dnn\_model2 выявлена наименьшая абсолютная средняя ошибка.

## 2.4. Нейронная сеть для прогноза соотношения матрица - наполнитель

Для прогнозирования соотношения матрица – наполнитель была выбрана модель dnn\_model2. Активационная функция sigmoid, оптимизатор RMSprop, функция потерь MAE, 3 скрытых слоя определены экспериментальным путем, на выходе 1 нейрон.

## 2.5. Разработка приложения

Для логической части приложения была выбрана модель прогнозирования соотношения матрица – наполнитель.

Для отображения результата прогнозирования пользователю необходимо ввести известные характеристики (тип float) и нажать кнопку «Отправить», ответ будет выведен в строке «Результат».

Изображение выглядит как стол

Автоматически созданное описаниеРабота приложения показана на рисунках ниже.

Рисунок 14 - Внесение данных для расчета прогноза пользователем

Изображение выглядит как стол

Автоматически созданное описание

Рисунок 15 - Получение результата прогнозирования

## 2.6. Репозиторий и результаты

Создан репозиторий в GitHub, где размещен код исследования, оформлен файл README.

Страница слушателя: https://github.com/ElenaBadeeva/

Созданный репозиторий: 2022\_Final\_qualifying\_work

Коммиты в репозитории:

# Заключение

В заключение следует отметить, что в результате исследования было построено несколько моделей для прогнозирования конечных свойств композитных материалов. Описанные выше подходы возможно использовать только как обучающие, для промышленного использования не подходят.

Цель работы, а именно проведение исследования с использованием методов, изученных на курсе «Data Science», была достигнута. Результаты проекта сформированы.

# Библиографический список

1. Бизли Д. Python. Подробный справочник: учебное пособие. – Пер. с англ. – СПб.: Символ-Плюс, 2010. – 864 с., ил.

2. Силен Дэви, Мейсман Арно, Али Мохамед. Основы Data Science и Big Data. Python и наука о данных. – СПб.: Питер, 2017. – 336 с.: ил.

3. Траск Эндрю. Грокаем глубокое обучение. – СПб.: Питер, 2019. – 352 с.: ил.

4. Машинное обучение. Конструирование признаков: принципы и техники для аналитиков/ Элис Чжен, Аманда Казари; [перевод с английского М.А. Райтман]. – Москва: Эксмо, 2022. – 240 с.

5. Быстрый старт. Минимальное приложение: – Режим доступа: https://flask-russian-docs.readthedocs.io/ru/latest/quickstart.html

6. Обучение и оценка модели с Keras: - Режим доступа: https://habr.com/ru/post/485890/

7. Советы по оптимизации гиперпараметров: – Режим доступа: https://www.machinelearningmastery.ru/grid-search-hyperparameters-deep-learning-models-python-keras/

8. Метрики и оценки: количественная оценка качества прогнозов: - Режим доступа: https://scikit-learn.ru/3-3-metrics-and-scoring-quantifying-the-quality-of-predictions/

# Приложения

1. Файл «12\_Прогнозирование конечных свойств новых материалов»
2. Файл «matrix\_filler2»
3. Файл «02\_Flask\_matriix\_filler»
4. Презентация «Выпускная квалификационная работа\_Бадеева Елена Евгеньевна»