

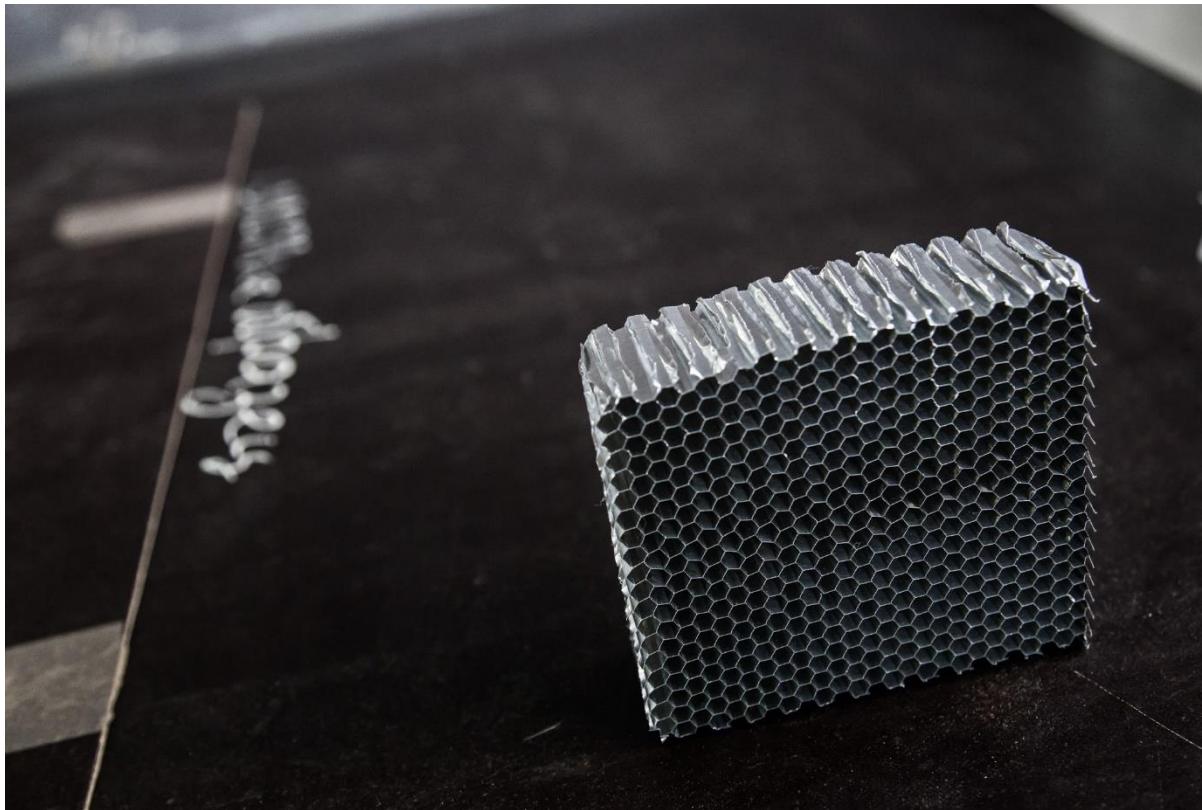
**ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ  
РАБОТА**  
**по курсу**  
**«Data Science»**

Слушатель

Курцева Лариса Юрьевна

Москва, 2023

Задача данной работы заключается в прогнозировании характеристик компонентов композиционных материалов на основе данных о составе композитов с использованием подхода, ориентированного на данные.



В настоящее время композиты являются неотъемлемой частью нашей жизни и широко используются в различных направлениях. Значимость композитов и необходимость в них постоянно растут. Современные композиты также демонстрируют преимущества перед традиционными материалами, в том числе в долговечности, прочности и легковесности. Необходимо также и снижение стоимости производства композитов, отсюда мы делаем вывод, что данная работа актуальна и значима.

Описание датасета

Столбец	Тип данных	Пропуски	Уникальные значения
Соотношение матрица-наполнитель	float64	нет	1014
Плотность, кг/м3	float64	нет	1013
модуль упругости, ГПа	float64	нет	1020
Количество отвердителя, м.%	float64	нет	1005
Содержание эпоксидных групп,%_2	float64	нет	1004
Температура вспышки, С_2	float64	нет	1003
Поверхностная плотность, г/м2	float64	нет	1004
Модуль упругости при растяжении, ГПа	float64	нет	1004
Прочность при растяжении, МПа	float64	нет	1004
Потребление смолы, г/м2	float64	нет	1003
Угол нашивки, град	float64	нет	2
Шаг нашивки	float64	нет	989
Плотность нашивки	float64	нет	988

Признак «Угол нашивки, град» считается категориальным и представлен только двумя значениями 0° и 90°, который при дальнейшем препроцессинге и с помощью кодировщика LabelEncoder будет преобразован в числовой.

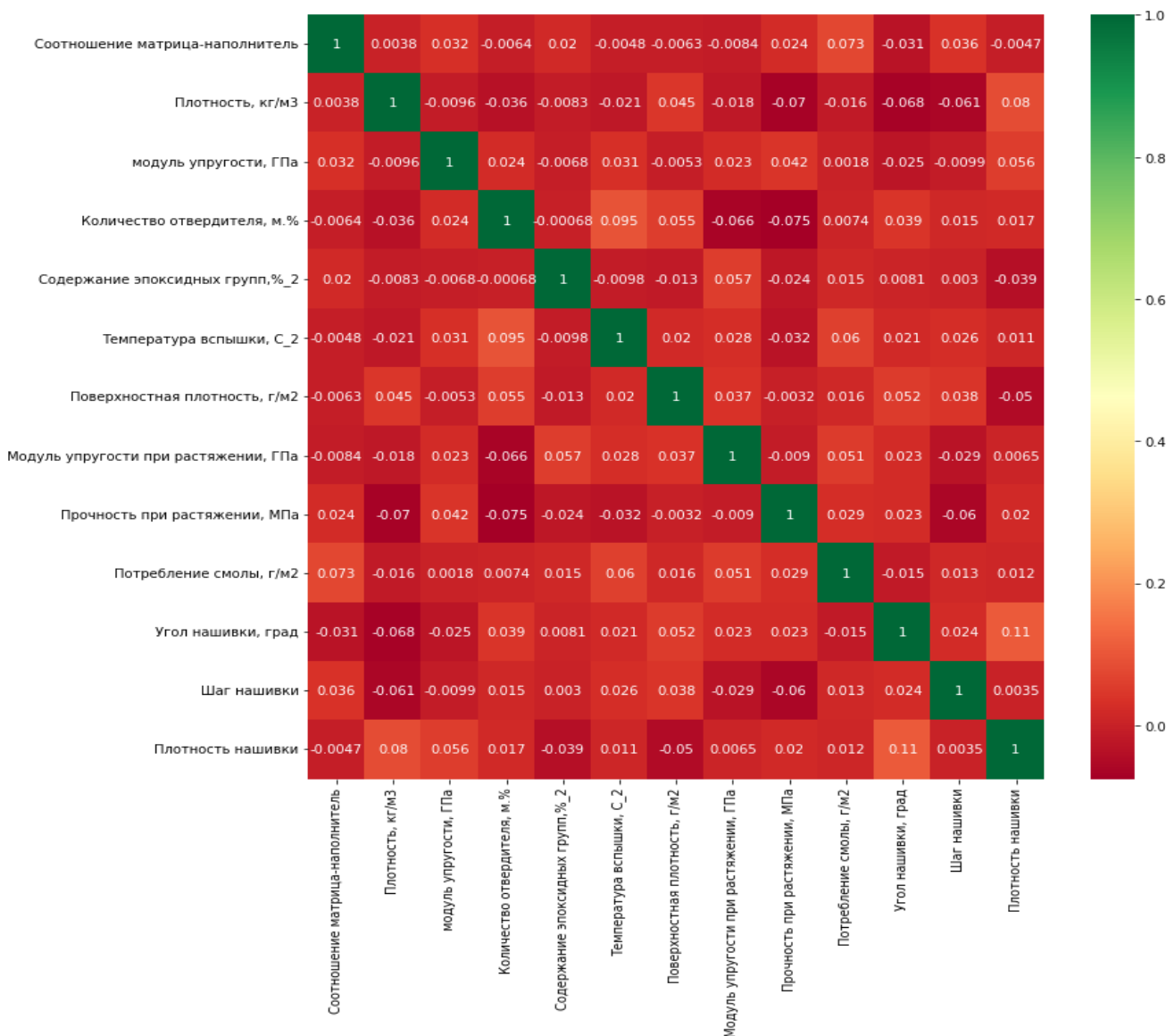
- пропусков значений в датасете нет;
- дубликатов нет;
- тип данных float64;

# Статистический анализ датасета

Столбец	Среднее значение	Медианное значение	Минимальное значение	Максимальное значение
Соотношение матрица-наполнитель	2.9304	2.9069	0.3894	5.5917
Плотность, кг/м3	1975.7349	1977.6217	1731.7646	2207.7735
модуль упругости, ГПа	739.9232	739.6643	2.4369	1911.5365
Количество отвердителя, м.%	110.5708	110.5648	17.7403	198.9532
Содержание эпоксидных групп,%_2	22.2444	22.2307	14.2550	33.0000
Температура вспышки, С_2	285.8822	285.8968	100.0000	413.2734
Поверхностная плотность, г/м2	482.7318	451.8644	0.6037	1399.5424
Модуль упругости при растяжении, ГПа	73.3286	73.2688	64.0541	82.6821
Прочность при растяжении, МПа	2466.9228	2459.5245	1036.8566	3848.4367
Потребление смолы, г/м2	218.4231	219.1989	33.8030	414.5906
Угол нашивки, град	44.2522	0.0000	0.0000	90.0000
Шаг нашивки	6.8992	6.9161	0.0000	14.4405
Плотность нашивки	57.1539	57.3419	0.0000	103.9889

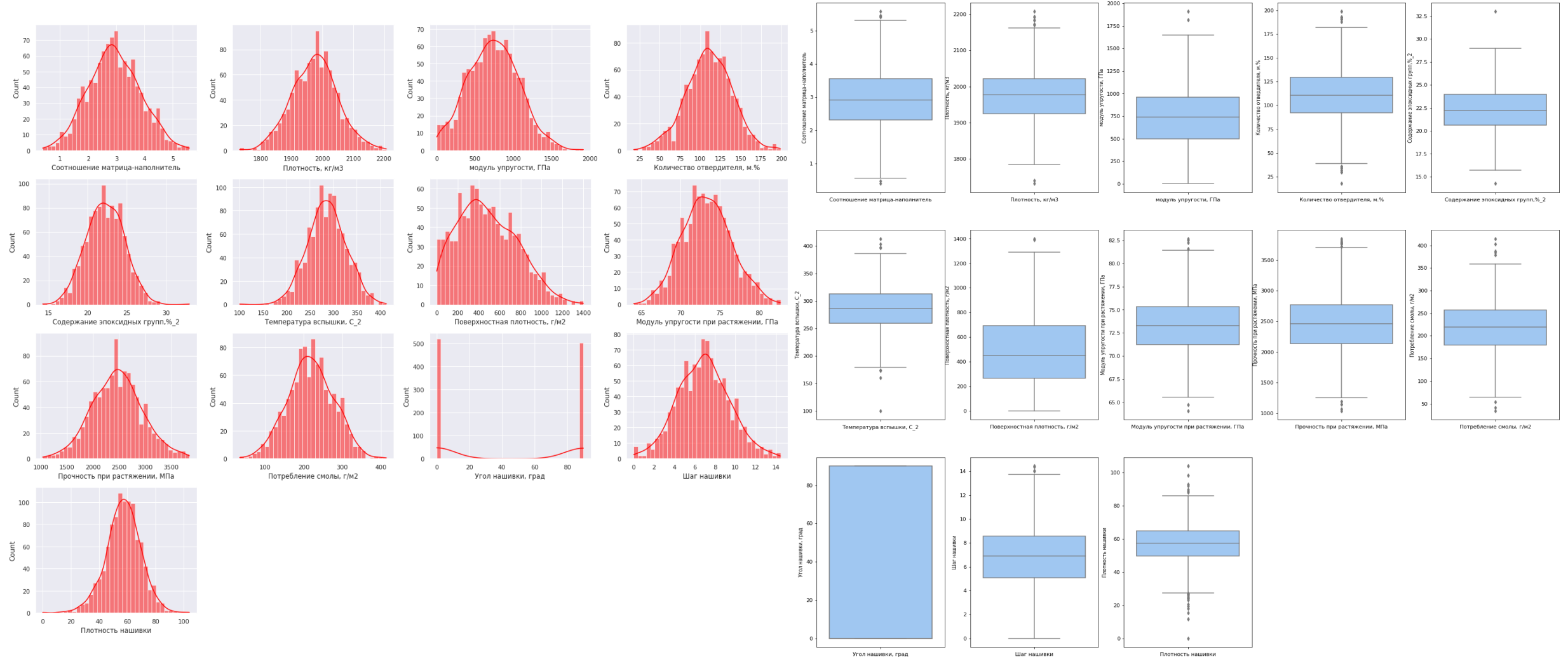
# Корреляция

Корреляция – взаимозависимость двух или нескольких случайных величин.

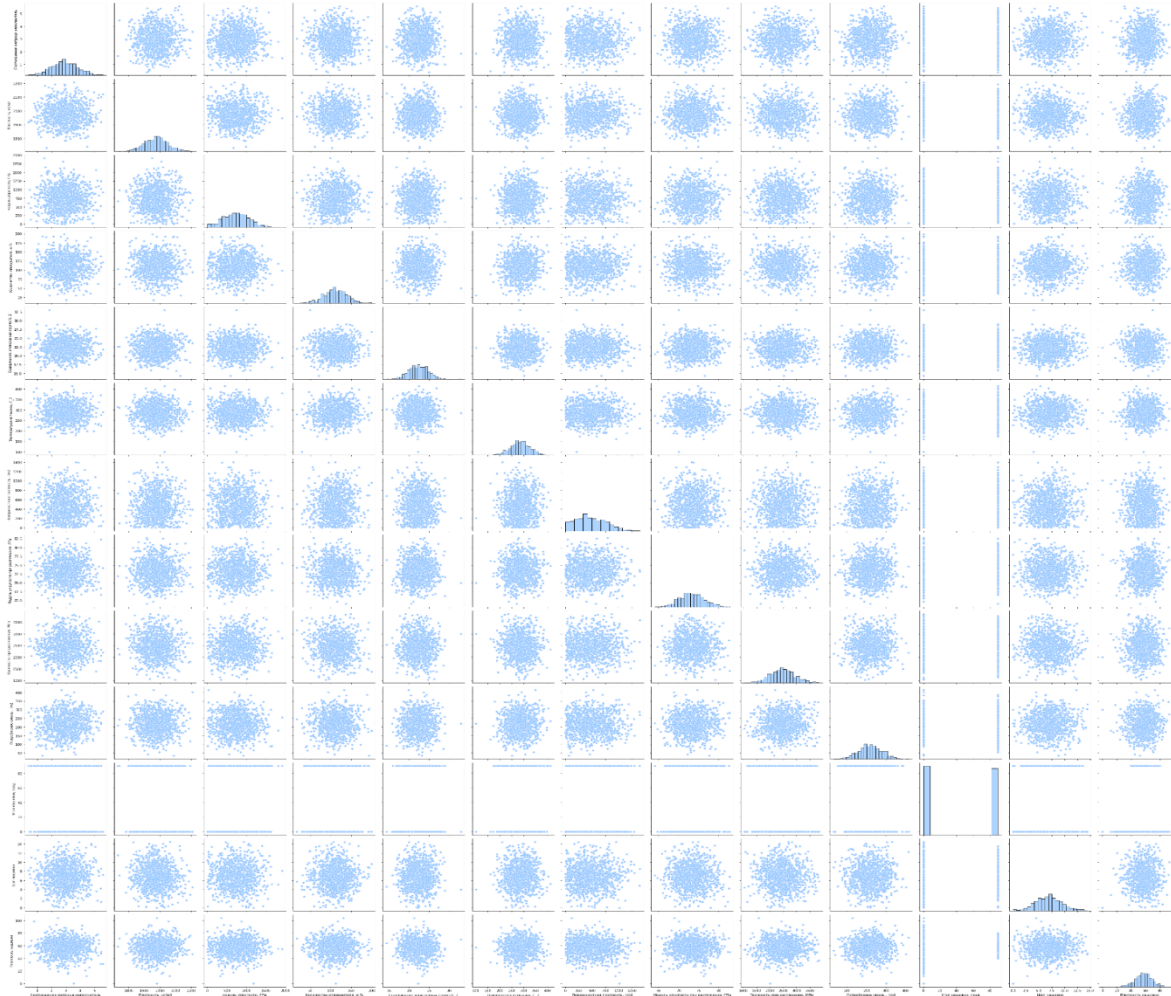


Тепловая карта, представленная на рисунке, показывает, что взаимосвязь между двумя или несколькими переменными нашего датасета очень слабая. Но надо понимать, что корреляционная зависимость отражает только взаимосвязь между переменными, но не говорит о причинах и следствиях.

Для демонстрации распределения значений, выбросов и потенциальных взаимосвязей между переменными, построены следующие диаграммы, гистограммы и графики:



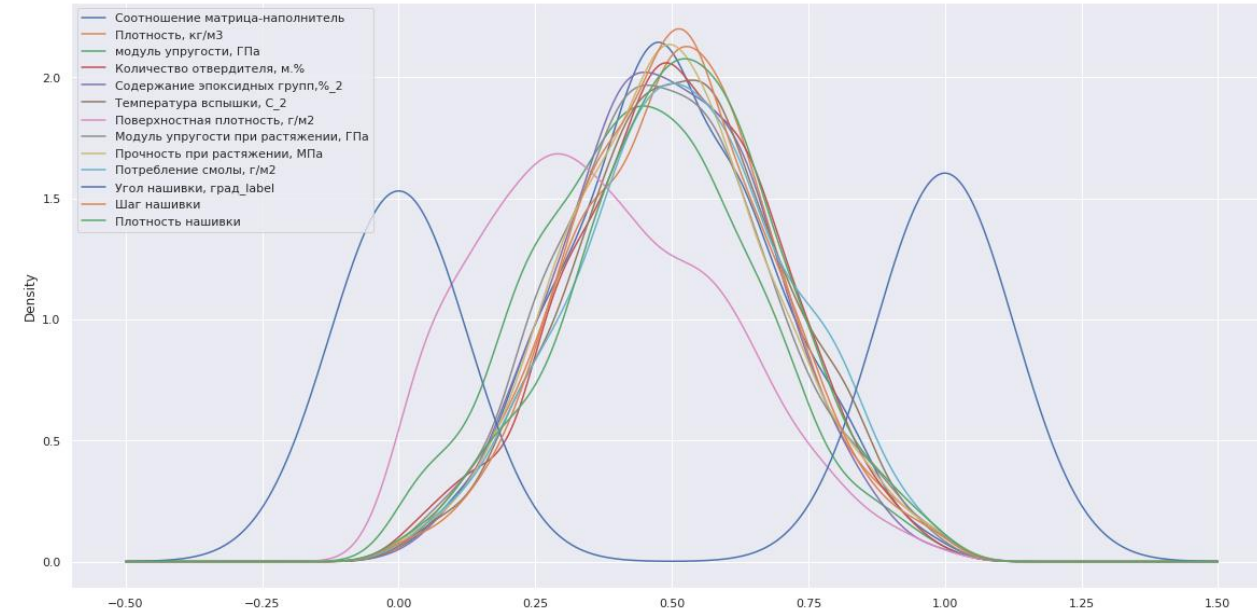
## Попарные диаграммы рассеяния точек



До нормализации

До нормализации плотность ядер на диаграммах рассеяния точек показывает равномерное распределение признаков. График оценки плотности ядер показывает также нормальное распределение после нормализации.

## Оценка плотности ядер



После нормализации



## Разбивка данных

По заданию известно, что необходимо построить модели регрессии и нейронную сеть для прогнозирования значений:

- «Прочности при растяжении, Мпа»;
- «Модуля упругости при растяжении, Гпа»;
- «Соотношения матрица-наполнитель».

Нормализованный датасет разделяется на 2 подвыборки: обучающую и тестовую, 0,7 и 0,3 от датасета соответственно, и определяется целевая переменная  $y$ . Над всеми тремя признаками, проводится идентичная работа, только меняется переменная, в зависимости от нашей задачи.

Ниже пример для одного признака.

```
y = df_norm_df['Соотношение матрица-наполнитель']  
x = df_norm_df.drop(('Соотношение матрица-наполнитель'), axis = 1)  
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.3, random_state=42)
```

```
print(x_train.shape)  
print(x_test.shape)  
print(y_train.shape)  
print(y_test.shape)
```

```
(655, 12)  
(281, 12)  
(655,)  
(281,)
```



## Методы, использованные для решения задачи

Предсказание значения целевой переменной — это задача регрессии. Регрессия — это зависимость среднего значения какой-либо величины от некоторой другой величины или нескольких величин. Таким образом, задача регрессии заключается в получении вещественного числа.

Существует несколько методов регрессии. При работе с датасетом использовались следующие модели, для определения, какая из них более работоспособна, дает меньшие ошибки и т.д.:

- линейная регрессия;
- ридж (гребневая) регрессия;
- регрессия по методу «лассо»
- регрессор случайного леса;
- градиентный бустинг;
- нейронная сеть.

# Тестирование моделей

Тестирование моделей показало следующие ошибки, приведенные в таблицах ниже.

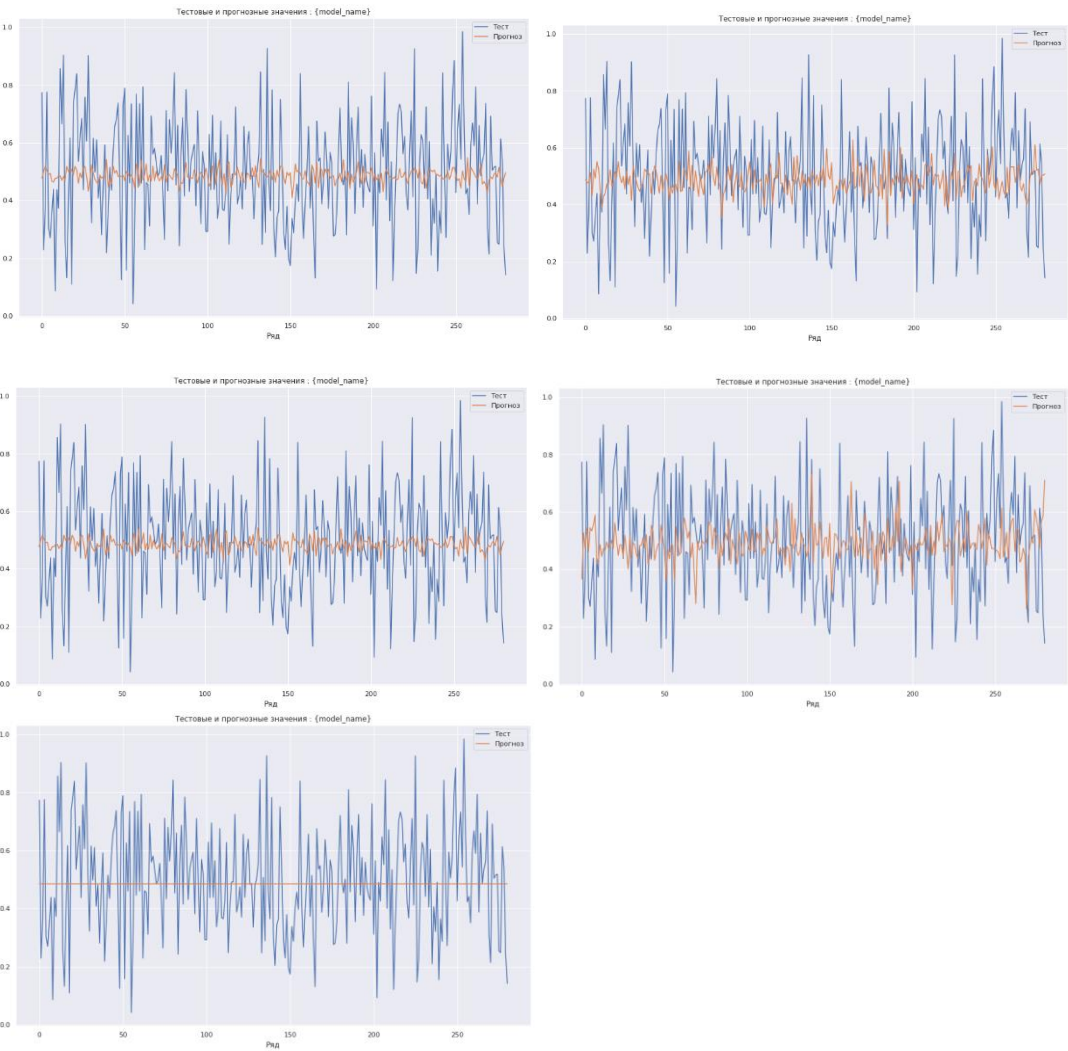
Для прогнозирования  
«Модуля упругости при растяжении, Гпа»

Модель	Коэф. детерм. на обуч. выборке	Коэф. детерм. на тест. выборке	MSE	MAE	MAPE	Max Error	Cross val score (mean)	Стандартное отклонение
Линейная регрессия	0.015	-0.005	0.035	0.152	0.453	0.483	-0.061	0.057
Ридж регрессия	0.015	-0.005	0.035	0.152	0.453	0.484	-0.059	0.056
«Лассо» регрессия	0.0	-0.006	0.035	0.151	0.450	0.500	-0.031	0.040
Регрессор случайного леса	0.848	-0.019	0.035	0.152	0.452	0.504	-0.114	0.078
Градиентный бустинг	0.456	-0.099	0.038	0.157	0.474	0.569	-0.133	0.125

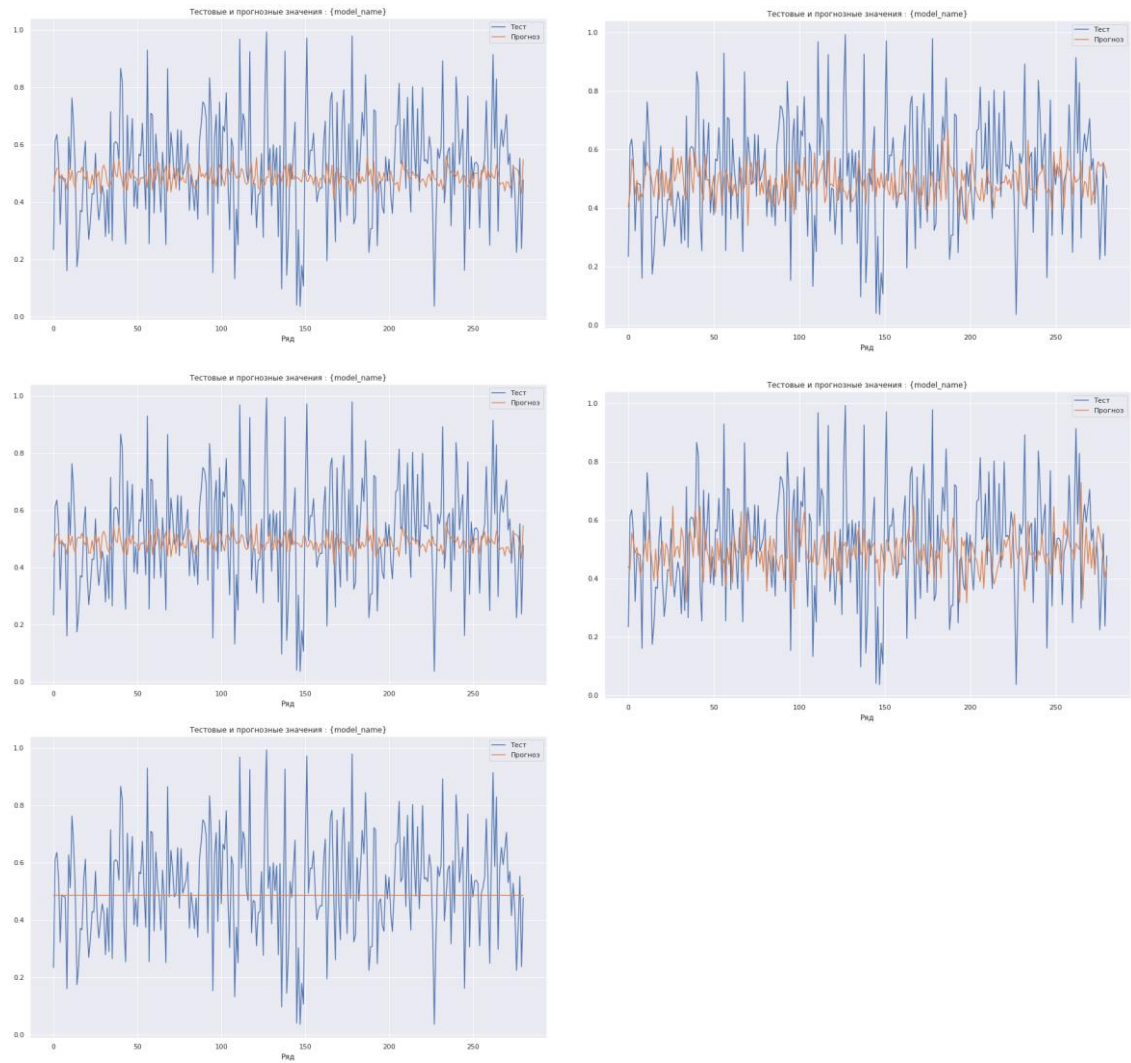
Для прогнозирования  
«Прочности при растяжении, Мпа»

Модель	Коэф. детерм. на обуч. выборке	Коэф. детерм. на тест. выборке	MSE	MAE	MAPE	Max Error	Cross val score (mean)	Стандартное отклонение
Линейная регрессия	0.025	-0.047	0.035	0.148	0.473	0.507	-0.065	0.083
Ридж регрессия	0.025	-0.045	0.035	0.148	0.473	0.507	-0.063	0.079
«Лассо» регрессия	0.0	-0.03	0.035	0.147	0.473	0.507	-0.031	0.042
Регрессор случайного леса	0.854	-0.097	0.037	0.154	0.484	0.539	-0.087	0.085
Градиентный бустинг	0.539	-0.123	0.038	0.156	0.473	0.536	-0.154	0.126

# Визуализация работы моделей для «Модуля упругости при растяжении, ГПа»



# Визуализация работы моделей для «Прочности при растяжении, МПа»



## Обучение моделей нейронной сети для признака «Соотношение матрица-наполнитель»

Параметры первой нейросети:

- входной слой 12 признаков;
- выходной слой 1 признак;
- 3 скрытых слоя;
- первый скрытый слой содержит 8 нейронов;
- второй скрытый слой содержит 8 нейронов;
- третий скрытый слой содержит 8 нейронов;
- функция активации –relu;
- на последнем скрытом слое функция активации – linear;
- оптимизатор – adam;
- функция потерь - mean\_absolute\_error;
- количество эпох – 20.

Параметры второй нейросети:

- входной слой 12 признаков;
- выходной слой 1 признак;
- 3 скрытых слоя;
- первый скрытый слой содержит 32 нейрона;
- второй скрытый слой содержит 16 нейронов;
- третий скрытый слой содержит 8 нейронов;
- функция активации –tanh;
- на последнем скрытом слое функция активации – linear;
- оптимизатор – adam;
- функция потерь - mean\_absolute\_error;
- количество эпох – 20.

# Архитектура первой нейронной сети

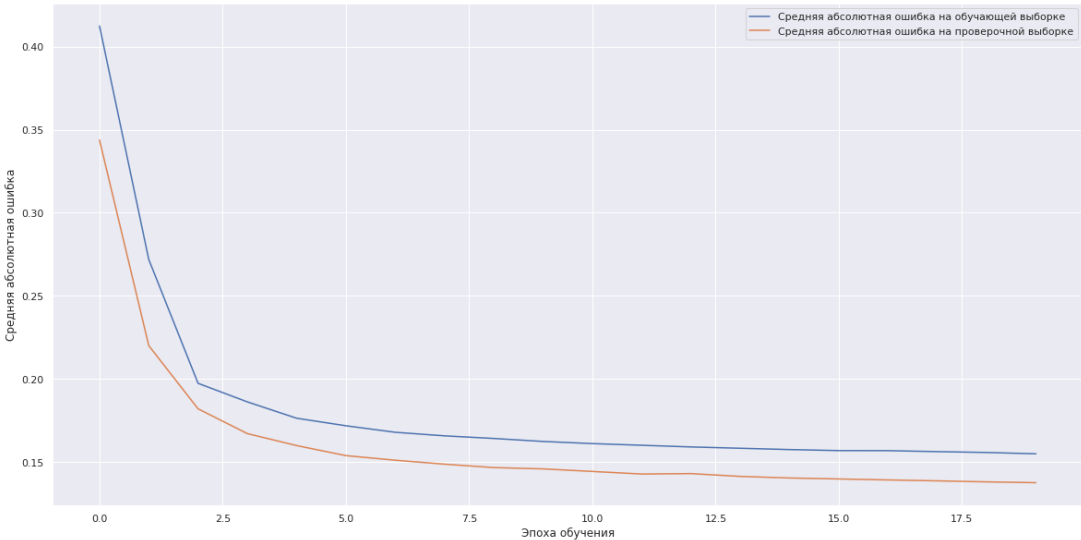
Model: "sequential"

Layer (type)	Output Shape	Param #
dense (Dense)	(None, 8)	104
dense_1 (Dense)	(None, 8)	72
dense_2 (Dense)	(None, 8)	72
dense_3 (Dense)	(None, 1)	9
Total params: 257		
Trainable params: 257		
Non-trainable params: 0		
None		

# Результаты работы нейронной сети

	loss	mae	val_loss	val_mae	epoch
15	0.156793	0.156793	0.139715	0.139715	15
16	0.156753	0.156753	0.139119	0.139119	16
17	0.156166	0.156166	0.138575	0.138575	17
18	0.155659	0.155659	0.137963	0.137963	18
19	0.154861	0.154861	0.137469	0.137469	19

# Качество обучения нейронной сети



# Метрики нейронной сети

21/21 [=====] - 0s 2ms/step  
Коэффициент детерминации на обучающей выборке: -0.04765695615654186  
9/9 [=====] - 0s 3ms/step  
Коэффициент детерминации на тестовой выборке (R2): -0.02897547710113435  
MSE для нейронной сети: 0.034  
MAE для нейронной сети: 0.149  
MAPE для нейронной сети: 0.453  
Max error для нейронной сети: 0.488

## Архитектура второй нейронной сети

Model: "sequential\_1"

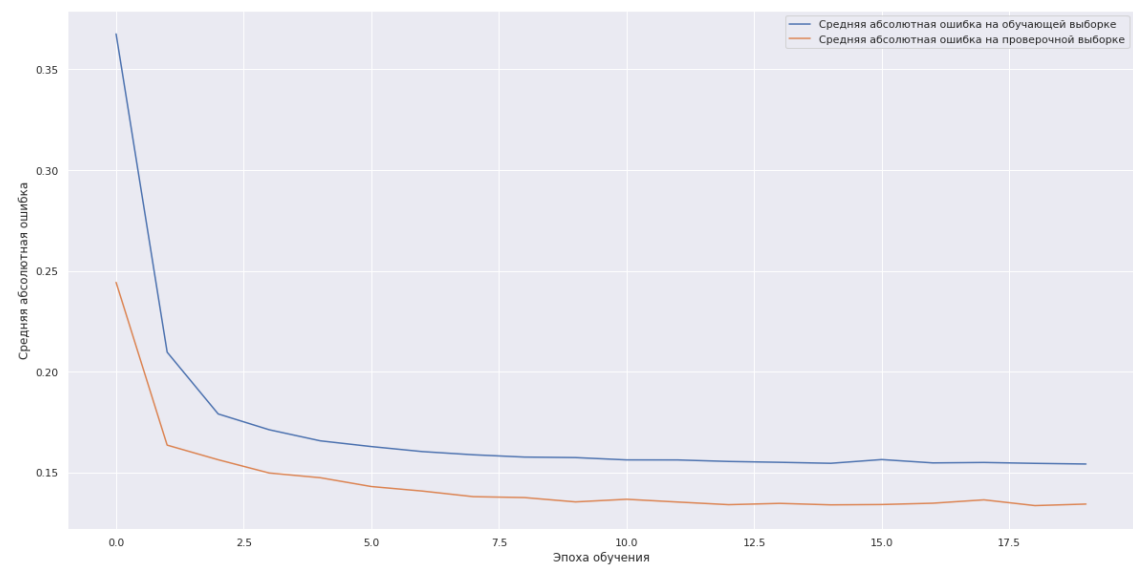
Layer (type)	Output Shape	Param #
dense_4 (Dense)	(None, 32)	416
dense_5 (Dense)	(None, 16)	528
dense_6 (Dense)	(None, 8)	136
dense_7 (Dense)	(None, 1)	9
Total params: 1,089		
Trainable params: 1,089		
Non-trainable params: 0		

None

## Результаты работы нейронной сети

	loss	mae	val_loss	val_mae	epoch
15	0.037117	0.156406	0.030753	0.134106	15
16	0.036764	0.154751	0.030406	0.134755	16
17	0.036540	0.154963	0.030695	0.136426	17
18	0.036384	0.154519	0.030528	0.133568	18
19	0.036397	0.154197	0.030371	0.134333	19

## Качество обучения нейронной сети



## Метрики нейронной сети

```
21/21 [=====] - 0s 1ms/step
Коэффициент детерминации на обучающей выборке: 0.008
9/9 [=====] - 0s 1ms/step
Коэффициент детерминации на тестовой выборке (R2): -0.09
MSE для нейронной сети: 0.037
MAE для нейронной сети: 0.155
MAPE для нейронной сети: 0.466
Max error для нейронной сети: 0.469
```

Сводная информация по метрикам модели нейронной сети для прогнозирования «Соотношения матрица-наполнитель»

Модель	Коэф. детерм. на обуч. выборке	Коэф. детерм. на тест. выборке	MSE	MAE	MAPE	Max Error
Нейронная сеть 1	-0.029	0.009	0.037	0.155	0.466	0.469
Нейронная сеть 2	0.008	-0.09	0.037	0.155	0.466	0.496

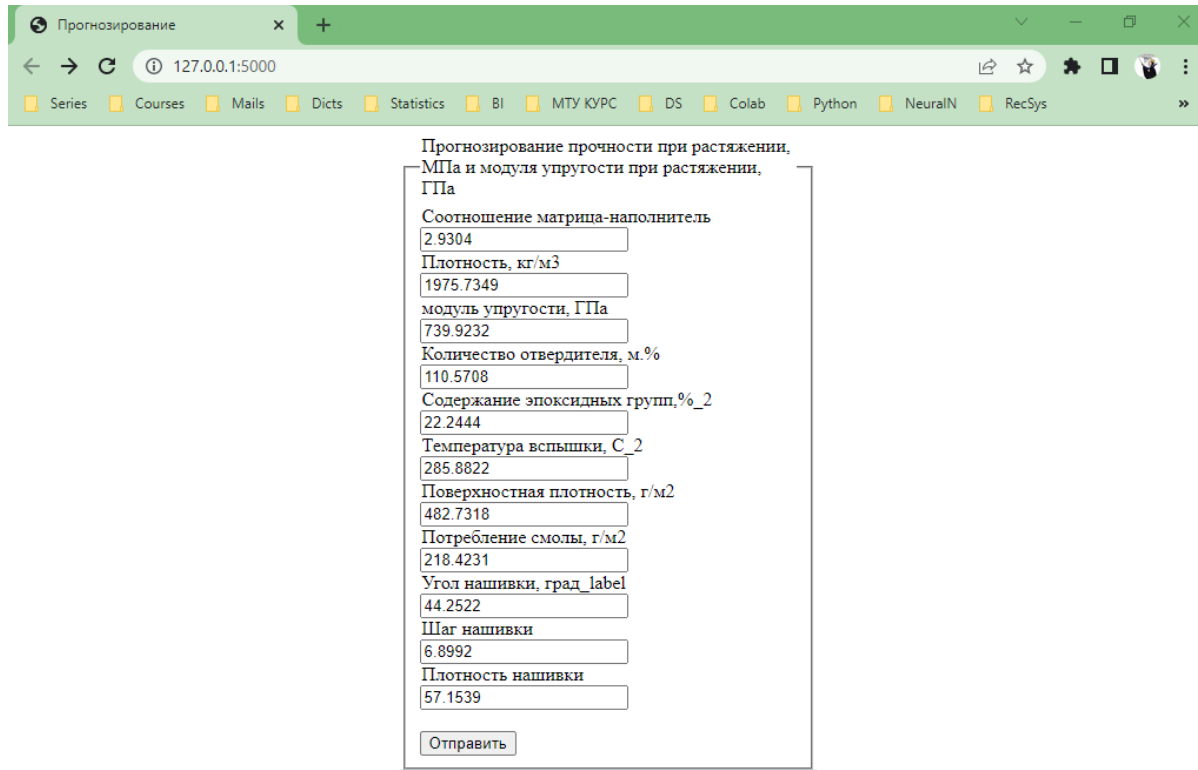
Основным показателем эффективности работы моделей регрессии является метрика R2 или коэффициент детерминации на тестовой выборке, т.е. то, как целевая переменная прогнозируется моделью. Его значения, как правило, варьируются от 0 до 1, чем выше, тем лучше, т.е. чем выше значение, тем лучше модель предсказывает значения. Одновременно с этим очень высокий коэффициент показывает переобученность модели. В нашем случае коэффициенты по всем моделям отрицательные, т.е. модели вообще не понимают признаков. Причина не в данных, а в обработке датасета. Перекрестная проверка и поиск гиперпараметров по сетке наши модели практически не улучшили, что еще раз говорит о вероятно неверном подходе к препроцессингу данных.



# Разработка приложения

Разработка приложения проводилась на фреймворке Flask на языке Python.

Скриншоты работы приложения.



Прогнозирование

127.0.0.1:5000

Series Courses Mails Dicts Statistics BI MTY KYPC DS Colab Python NeuralN RecSys

Прогнозирование прочности при растяжении, МПа и модуля упругости при растяжении, ГПа

Соотношение матрица-наполнитель  
2.9304

Плотность, кг/м<sup>3</sup>  
1975.7349

модуль упругости, ГПа  
739.9232

Количество отвердителя, м.%  
110.5708

Содержание эпоксидных групп, %<sub>2</sub>  
22.2444

Температура вспышки, C<sub>2</sub>  
285.8822

Поверхностная плотность, г/м<sup>2</sup>  
482.7318

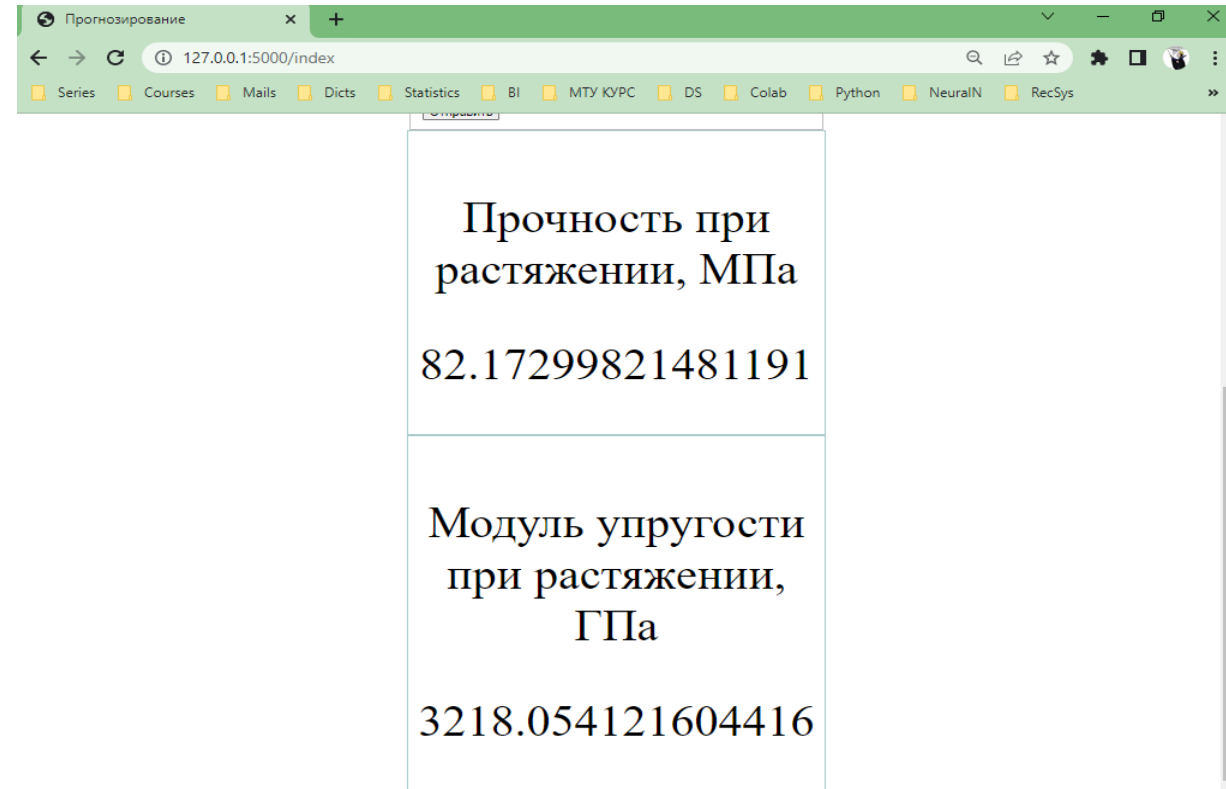
Потребление смолы, г/м<sup>2</sup>  
218.4231

Угол нашивки, град\_label  
44.2522

Шаг нашивки  
6.8992

Плотность нашивки  
57.1539

Отправить



Прогнозирование

127.0.0.1:5000/index

Series Courses Mails Dicts Statistics BI MTY KYPC DS Colab Python NeuralN RecSys

Прочность при растяжении, МПа

82.17299821481191

Модуль упругости при растяжении, ГПа

3218.054121604416