ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

Образовательный центр МГТУ им. Н.Э. Баумана

Выпускная квалификационная работапо курсу "Data Science Pro"

Прогнозирование конечных свойств новых материалов (композиционных материалов)

Бобрович Татьяна Александровна

14 апреля 2024 года

ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

Постановка задачи

1. Изучить предметную область провести разведочный анализ данных
2. Разделить данные на тренировочную и тестовую выборки выполнить препроцессинг (предобаботку)
3. Подготовить модели для подбора
   сравнить модели с гиперпараметрами по умолчанию
4. Подобрать гиперпараметры с помощью с помощью поиска по сетке с перекрестной проверкой
5. Сравнить модели после подбора гиперпараметров и выбрать лучшую, разработать приложение

ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

Разведочный анализ данных

Предложенные для исследования наборы данных (DataSet)

X_bp.xlsx

1. Имеет индекс и 10 признаков
2. 1023 строки

X_nlp .xlsx

1. Имеет индекс и 3 признака
2. 1040 строк

Полученный набор данных в результате объединения по индексу

1. **Имеется 13 признаков**
2. **1023 строки**

ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

Разведочный анализ данных

Название	Файл	Тип данных	Непустых значений	Уникальных значений
Соотношение матрица-наполнитель	X_bp	float64	1023	1014
Плотность, кг/м3	X_bp	float64	1023	1013
модуль упругости, ГПа	X_bp	float64	1023	1020
Количество отвердителя, м.%	X_bp	float64	1023	1005
Содержание эпоксидных групп,%_2	X_bp	float64	1023	1004
Температура вспышки, С_2	X_bp	float64	1023	1003
Поверхностная плотность, г/м2	X_bp	float64	1023	1004
Модуль упругости при растяжении, ГПа	X_bp	float64	1023	1004
Прочность при растяжении, МПа	X_bp	float64	1023	1004
Потребление смолы, г/м2	X_bp	float64	1023	1003
Угол нашивки, град	X_nup	int64	1023	2
Шаг нашивки	X_nup	float64	1023	989
Плотность нашивки	X_nup	float64	1023	988

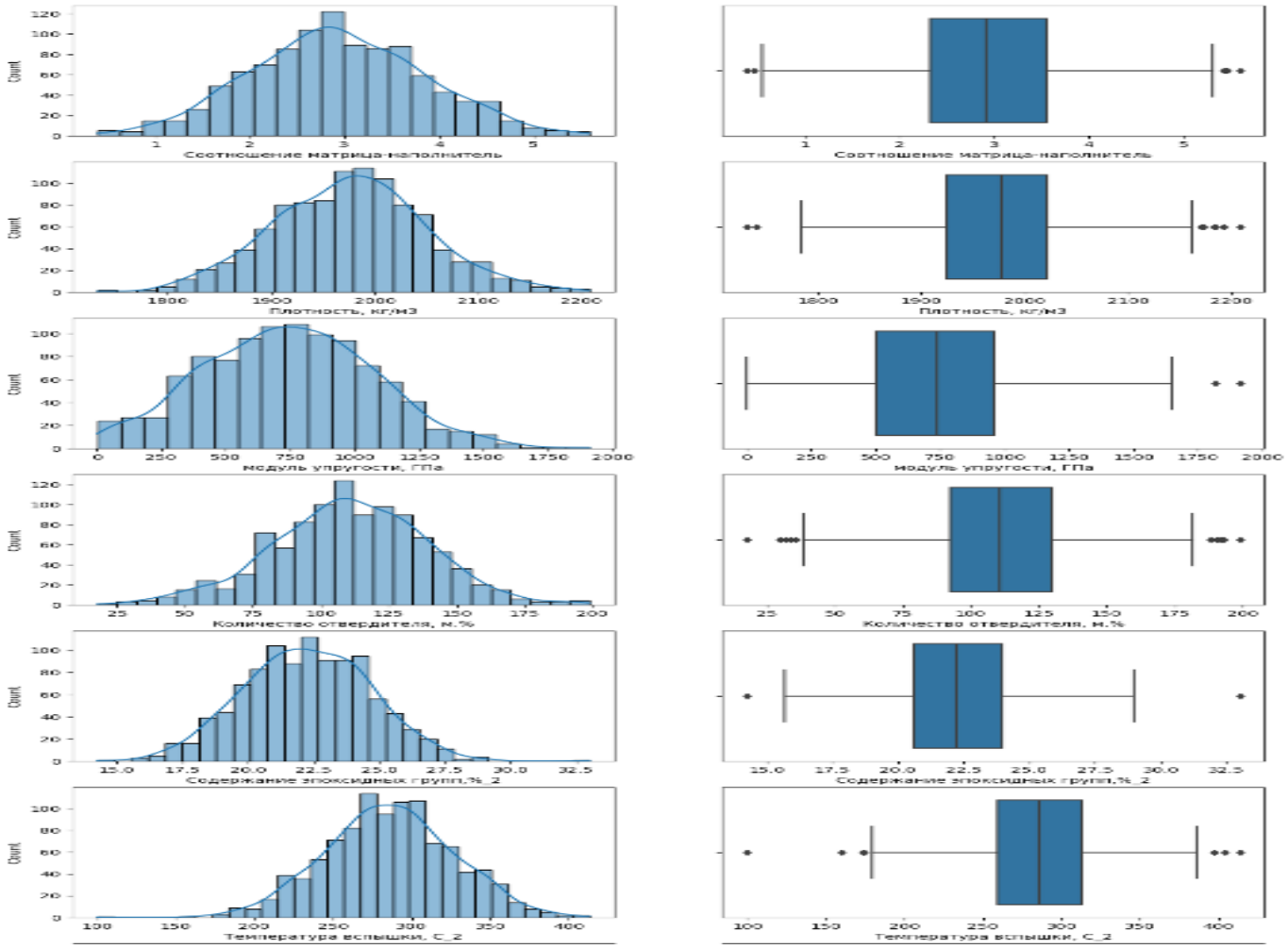
	Среднее	Стандартное отклонение	Минимум	Максимум	Медиана
Соотношение матрица-наполнитель	2.9304	0.9132	0.3894	5.5917	2.9069
Плотность, кг/м3	1975.7349	73.7292	1731.7646	2207.7735	1977.6217
модуль упругости, ГПа	739.9232	330.2316	2.4369	1911.5365	739.6643
Количество отвердителя, м.%	110.5708	28.2959	17.7403	198.9532	110.5648
Содержание эпоксидных групп, %_2	22.2444	2.4063	14.2550	33.0000	22.2307
Температура вспышки, С_2	285.8822	40.9433	100.0000	413.2734	285.8968
Поверхностная плотность, г/м2	482.7318	281.3147	0.6037	1399.5424	451.8644
Модуль упругости при растяжении, ГПа	73.3286	3.1190	64.0541	82.6821	73.2688
Прочность при растяжении, МПа	2466.9228	485.6280	1036.8566	3848.4367	2459.5245
Потребление смолы, г/м2	218.4231	59.7359	33.8030	414.5906	219.1989
Угол нашивки, град	44.2522	45.0158	0.0000	90.0000	0.0000
Шаг нашивки	6.8992	2.5635	0.0000	14.4405	6.9161
Плотность нашивки	57.1539	12.3510	0.0000	103.9889	57.3419

ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

Гистограммы распределения и диаграммы

Количественные, вещественные, положительные, нормально распределенные

Угол нашивки — категориальный, бинарный

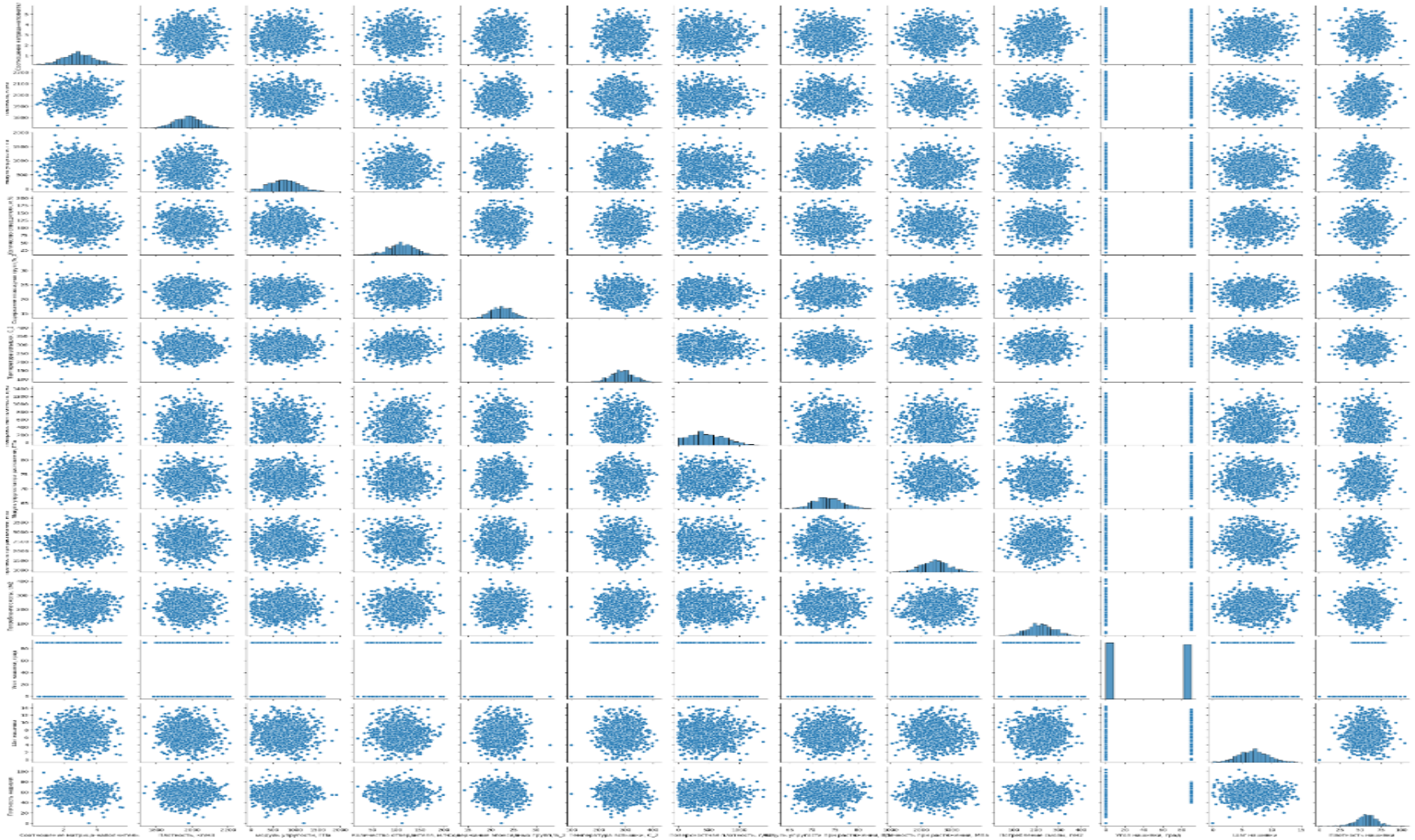


ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

Графики рассеяния точек

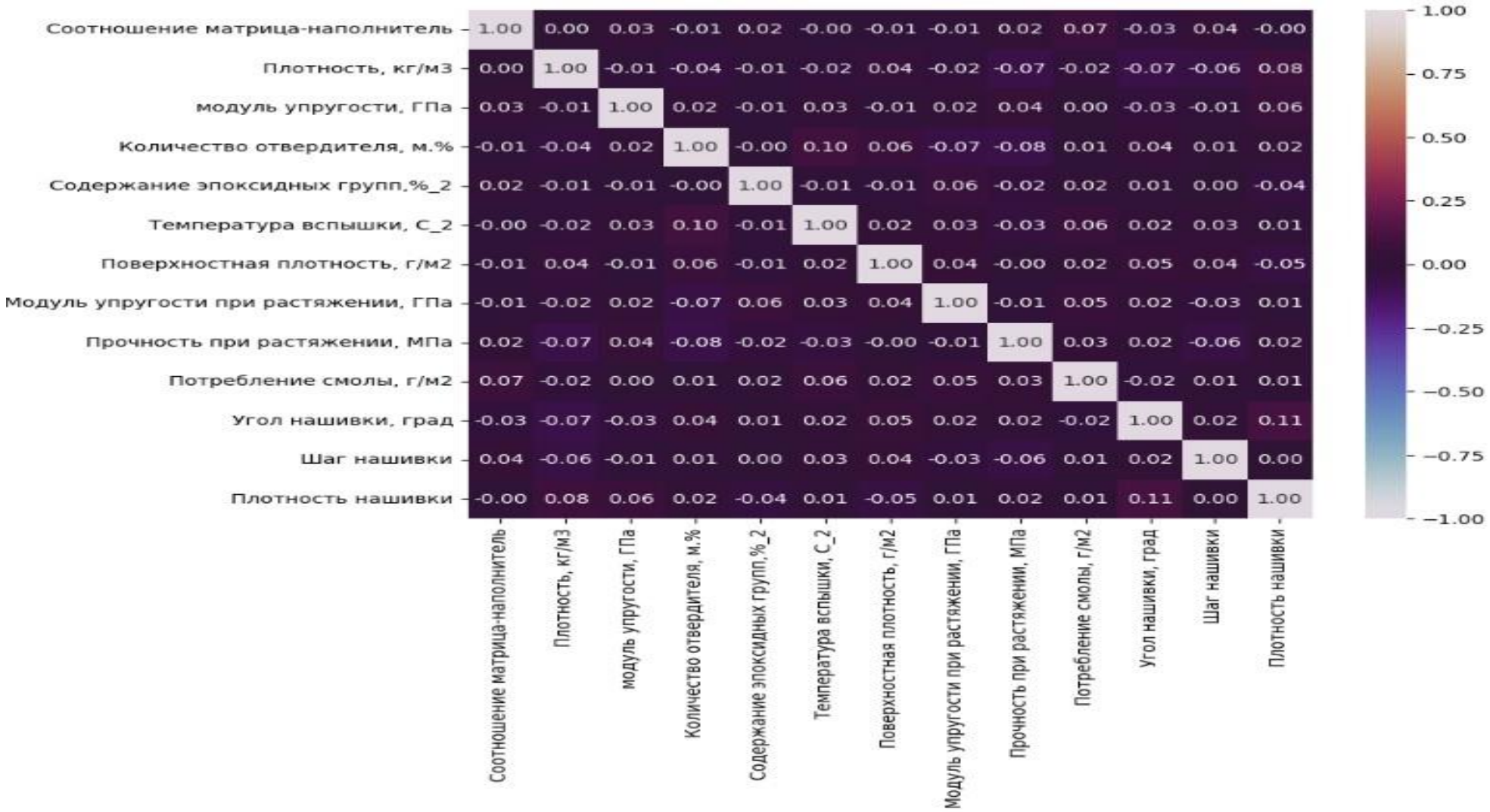
Выбросы наблюдаются

Зависимостей нет



ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

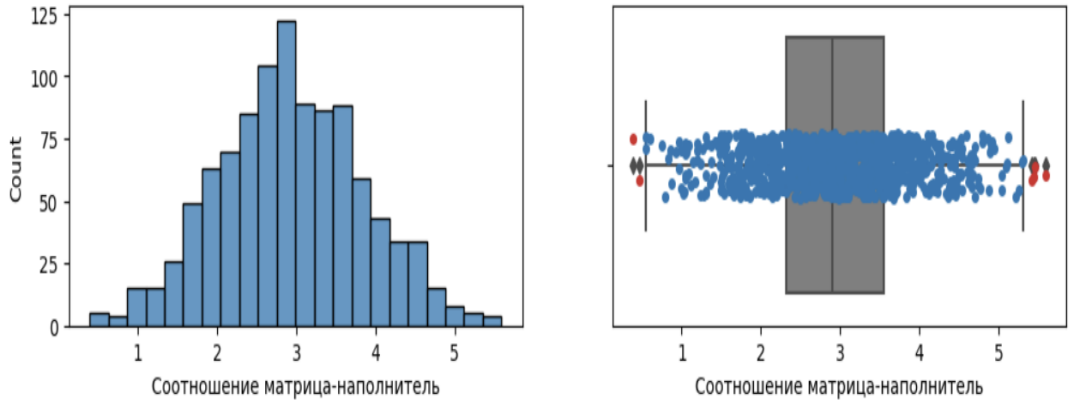
Матрица корреляции



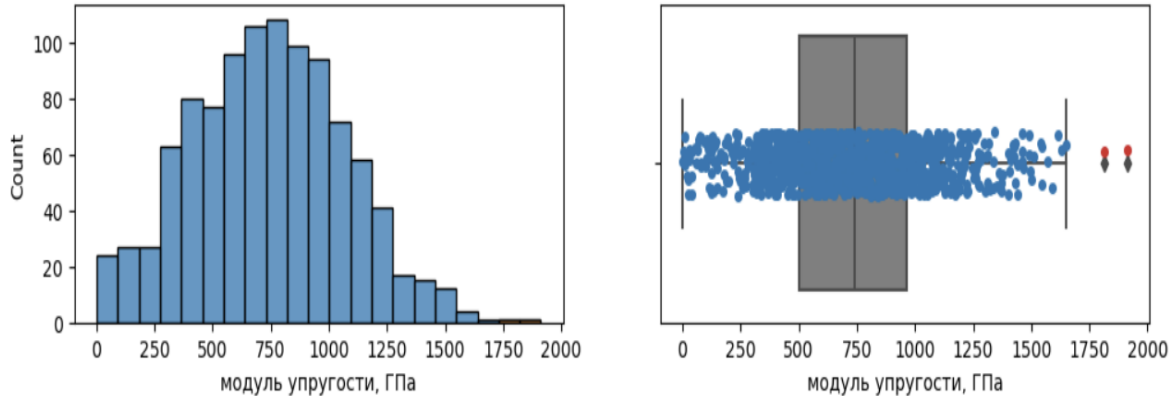
ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

Гистограмма распределения и диаграмме «ящик с усами»

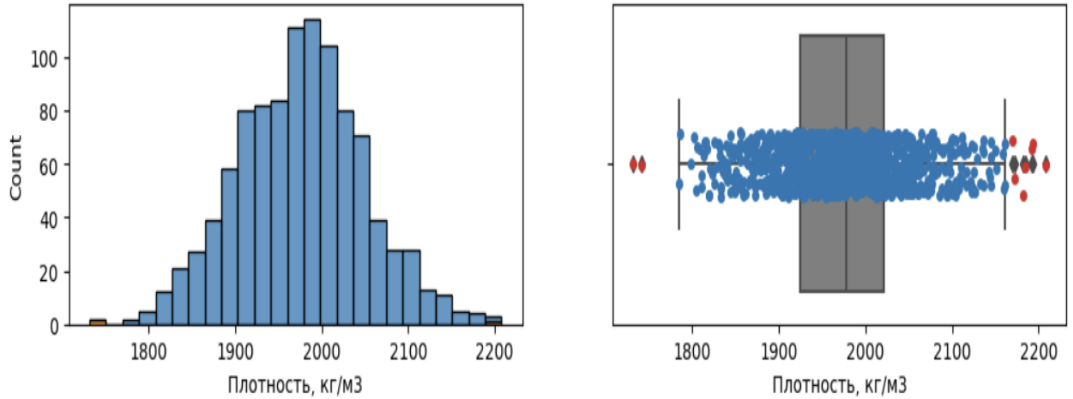
Соотношение матрица-наполнитель: 3s=0 iq=6



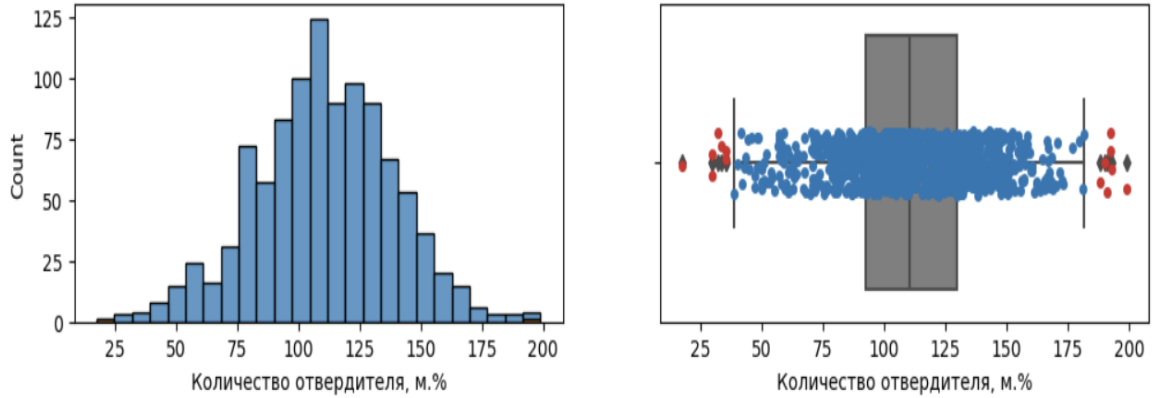
модуль упругости, ГПа: 3s=2 iq=2



Плотность, кг/м3: 3s=3 iq=9



Количество отвердителя, м.%: 3s=2 iq=14



ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

Предобработка данных

Предобработка данных:
✓ Исключение выбросов:
• Посчитаем количество значений методом 3 сигм и методом межквартильных расстояний;
• Исключим выбросы методом 3 сигм ;
• Проверим результат;
• Проверим чистоту датасета от выбросов ;
• Построим все возможные графики.

✓ Нормализация данных:
• Нормализуем данные MinMaxScaler()

```
df_norm.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 11 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   Соотношение матрица-наполнитель  1000 non-null   float64
 1   Плотность, кг/м3                1000 non-null   float64
 2   модуль упругости, ГПа           1000 non-null   float64
 3   Количество отвердителя, м.%     1000 non-null   float64
 4   Содержание эпоксидных групп,%_2  1000 non-null   float64
 5   Температура вспышки, С_2        1000 non-null   float64
 6   Поверхностная плотность, г/м2   1000 non-null   float64
 7   Потребление смолы, г/м2         1000 non-null   float64
 8   Угол нашивки, град              1000 non-null   float64
 9   Шаг нашивки                     1000 non-null   float64
 10  Плотность нашивки               1000 non-null   float64
dtypes: float64(11)
memory usage: 86.1 KB
```

```
# Категориальные данные при нормализации преобразовались в 0 и 1 и можно не применять OneHotEncoding
df_norm['Угол нашивки, град'].unique ()
```

```
array([0., 1.])
```

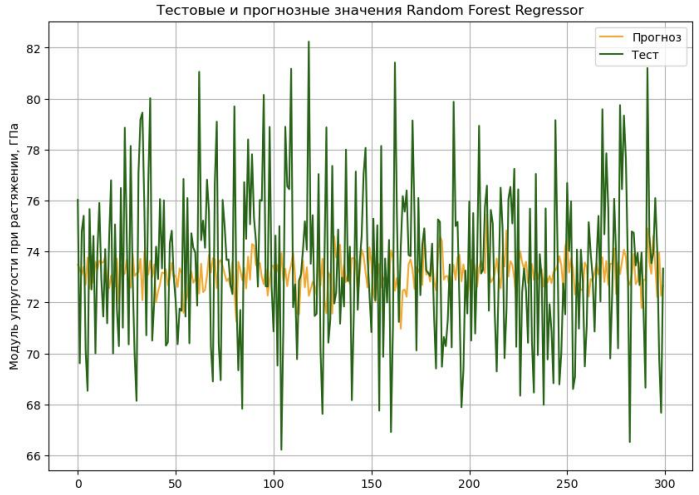
ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

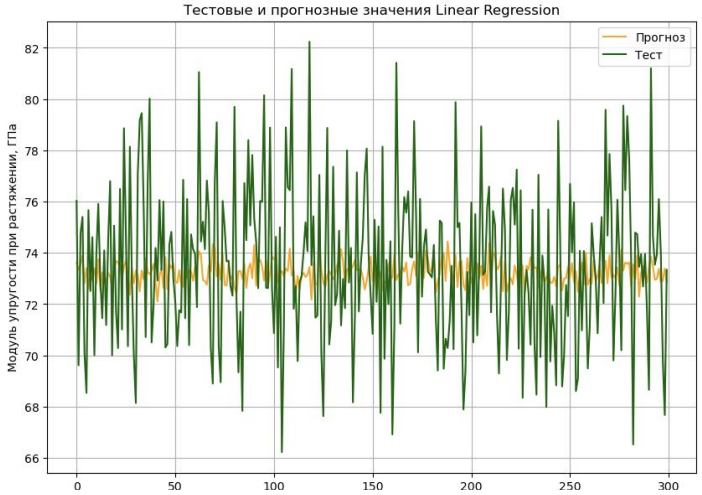
Рассмотренные модели

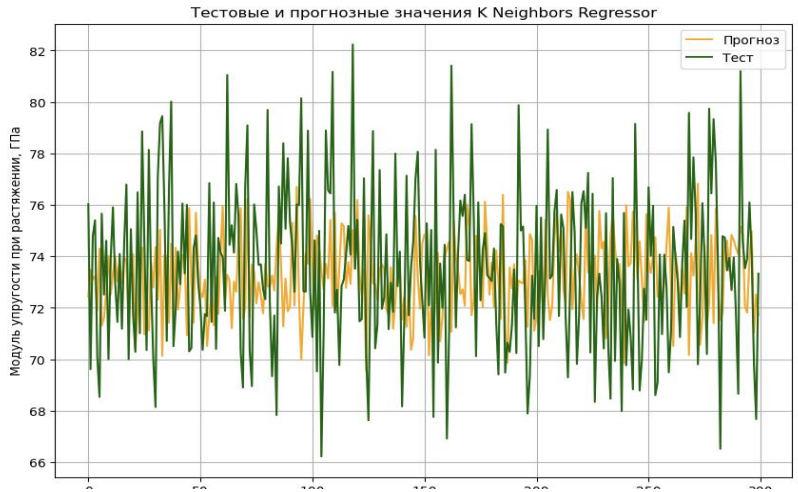
1. Линейная регрессия
2. Метод опорных векторов для регрессии
3. Метод k-ближайших соседей
4. Деревья решений
5. Градиентный бустинг
6. Случайный лес
7. Нейронная сеть

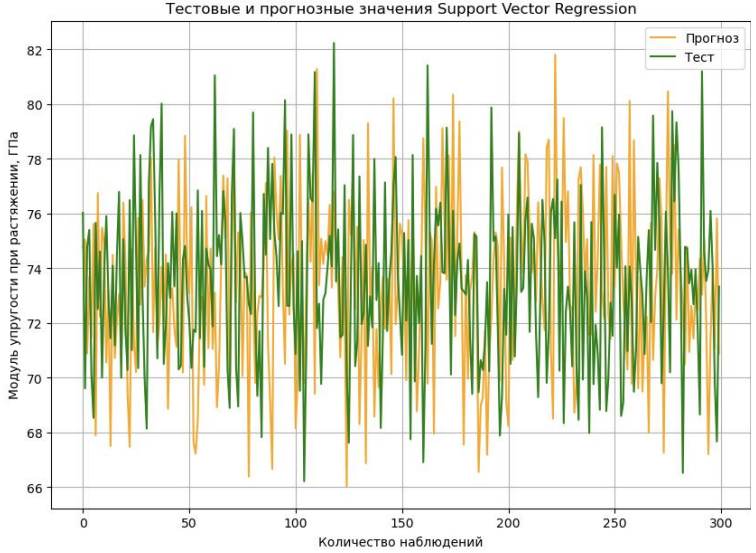
ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

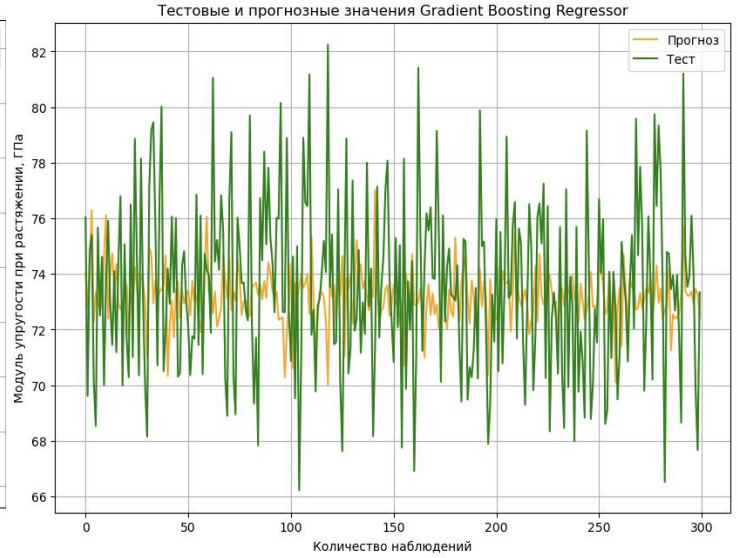
Модели прогноза «Модуль упругости при растяжении, Гпа»

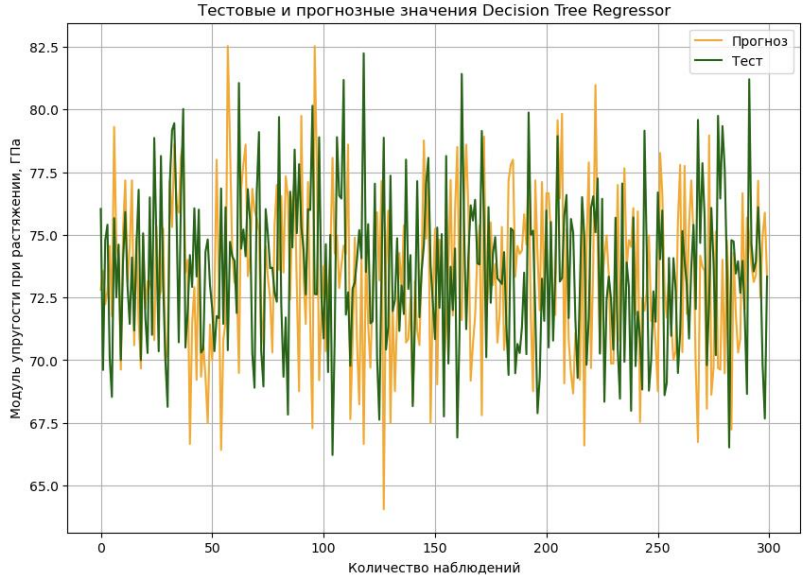












ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

Модели прогноза «Модуль упругости при растяжении, Гпа»

	Регрессор	MAE
0	RandomForest	2.578553
1	Linear Regression	2.510989
2	KNeighbors	2.803555
3	Support Vector	3.505371
4	GradientBoosting	2.586746
5	DecisionTree	3.506381
6	RandomForest1_GridSearchCV	2.537475
7	KNeighbors1_GridSearchCV	2.768574
8	DecisionTree1_GridSearchCV	2.527006

По умолчанию

```
Random Forest Regressor Results Train:
Test score: 0.40
Random Forest Regressor Results:
RF_MAE:  3
RF_MAPE: 0.03
RF_MSE: 10.19
RF_RMSE: 3.19
Test score: -0.07
```

```
Linear Regression Results Train:
Test score: 0.02
Linear Regression Results:
lr_MAE:  3
lr_MAPE: 0.03
lr_MSE: 9.63
lr_RMSE: 3.10
Test score: -0.01
```

```
K Neighbors Regressor  Results Train:
Test score: 0.24
K Neighbors Regressor  Results:
KNN_MAE:  3
KNN_MAPE: 0.04
KNN_MSE: 12.45
KNN_RMSE: 3.53
Test score: -0.30
```

После подбора гиперпараметров

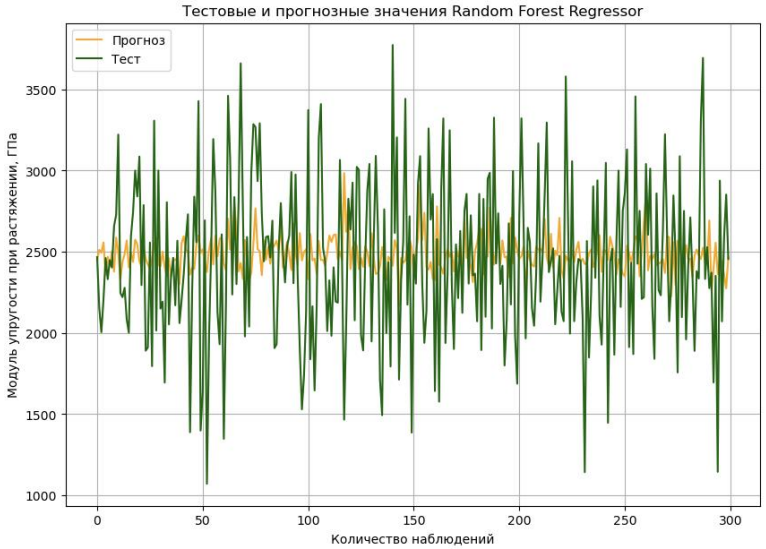
```
Support Vector Regression Results Train:
Test score: 0.91
Support Vector Regression Results:
SVR_MAE: 4
SVR_MAPE: 0.05
SVR_MSE: 19.65
SVR_RMSE: 4.43
Test score: -1.05
```

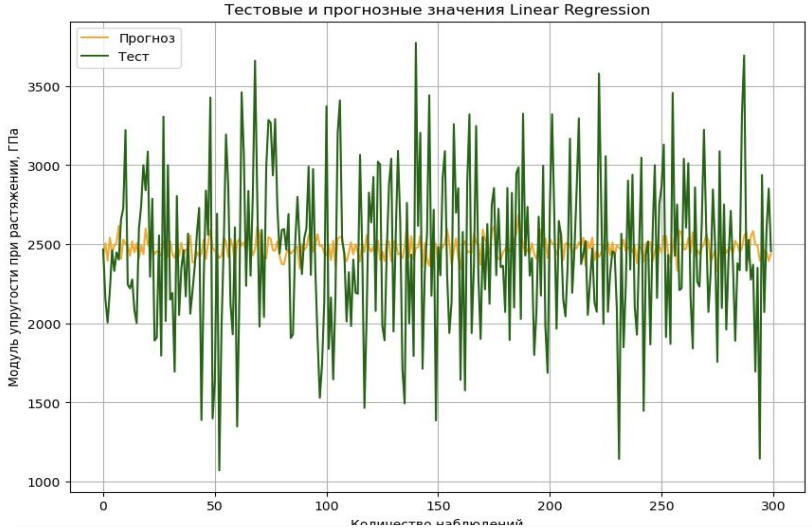
```
Gradient Boosting Regressor Results Train:
Test score: 0.49
Gradient Boosting Regressor Results:
GBR_MAE:  3
GBR_MAPE: 0.04
GBR_MSE: 10.25
GBR_RMSE: 3.20
Test score: -0.07
```

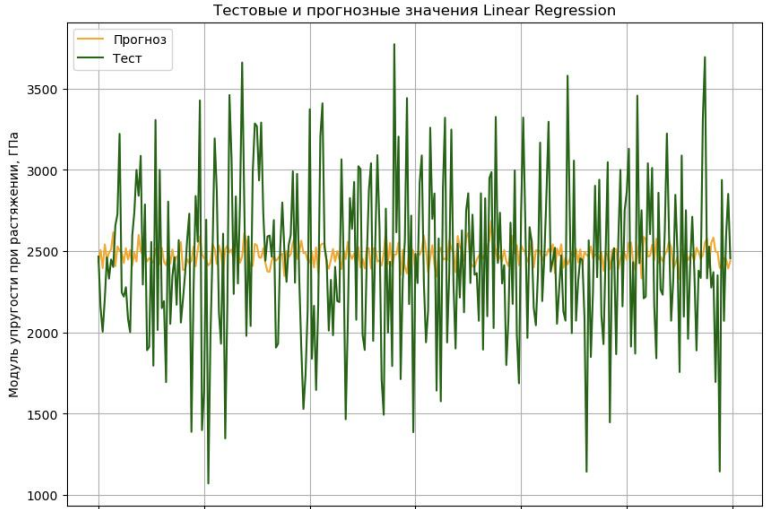
```
Decision Tree Regressor Results Train:
Test score: 1.00
Decision Tree Regressor Results:
DTR_MAE:  4
DTR_MSE: 20.67
DTR_RMSE: 4.55
DTR_MAPE: 0.05
Test score: -1.16
```

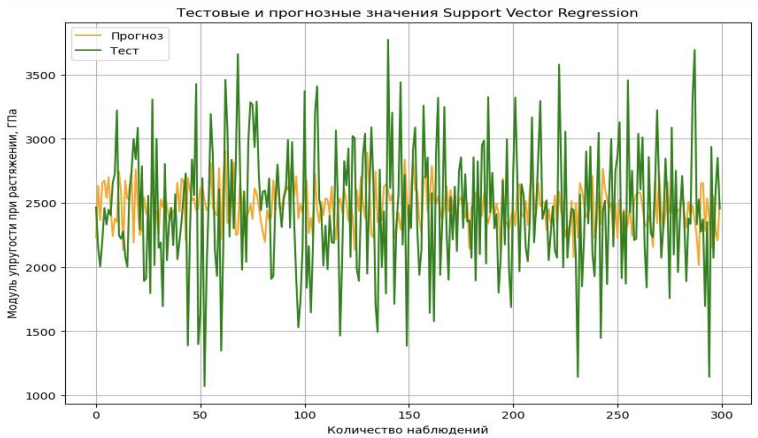
ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

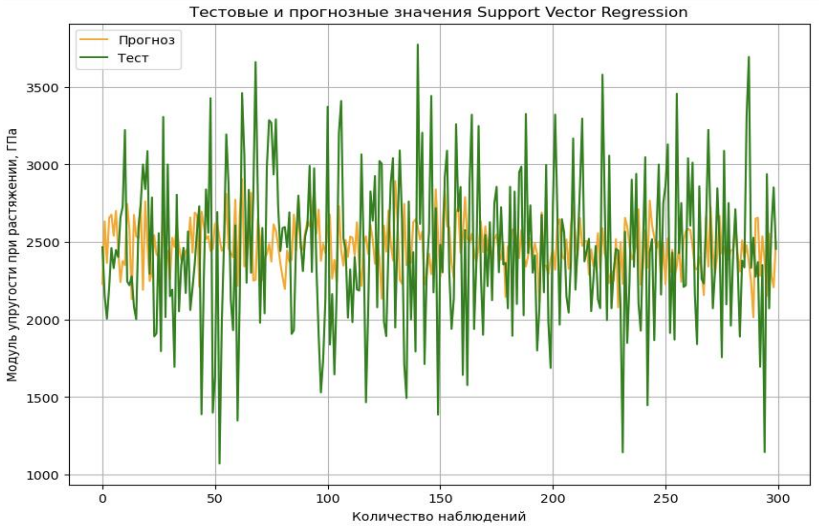
Модели прогноза «Прочность при растяжении, МПа, Гпа»

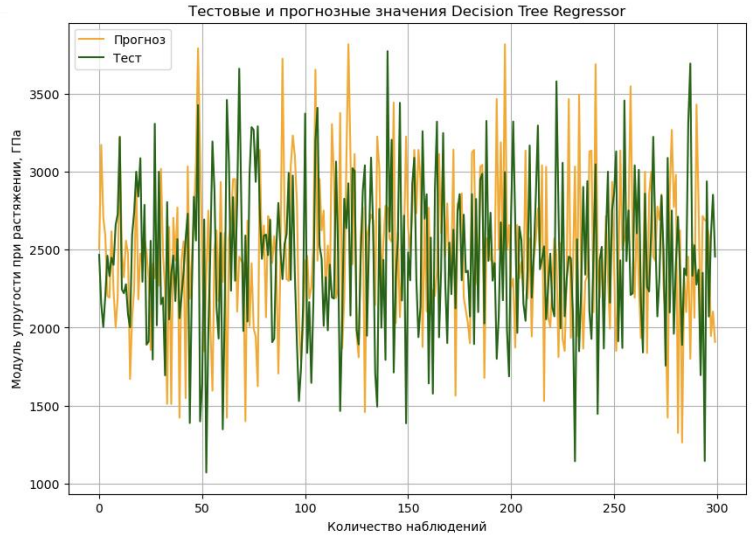








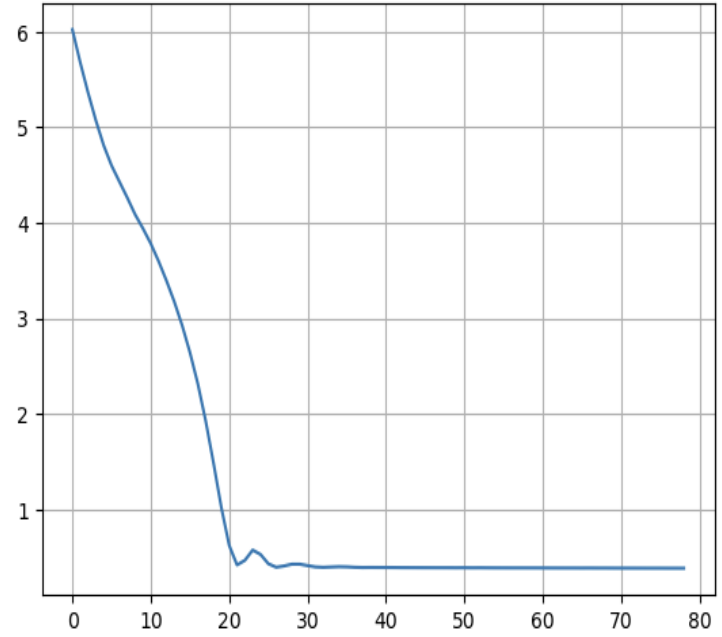


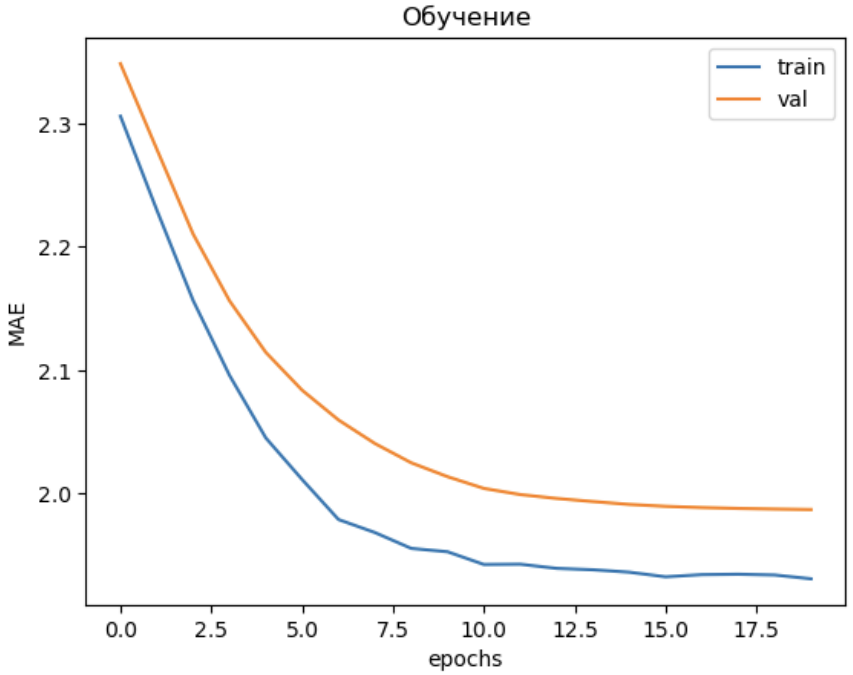


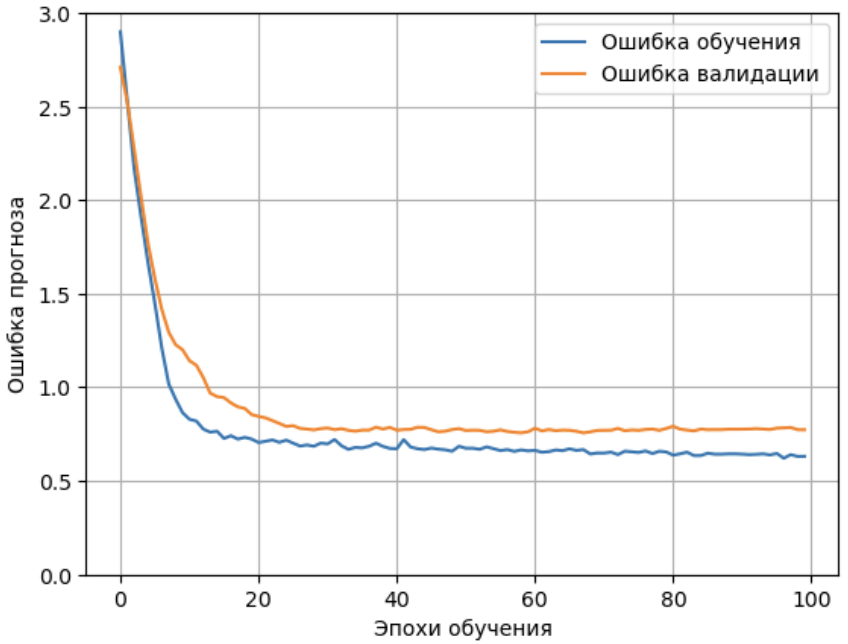
ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

Модель соотношения матрица-наполнитель

ИНС MLPRegressor из библиотеки sklearn








```
MAE3: 0.7694991587628883
MSE3: 0.9045963312234566
RMSE3: 0.9511026922595985
```

ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

Результаты работы

В результате исследования предложенных заказчиком датасетов различными методами, зависимостей, позволяющих обучить эффективную модель не получено

1. Недостаточное количество данных: если датасет слишком маленький, модель может не иметь достаточно информации для обучения.

2. Низкое качество данных: если данные содержат ошибки, пропуски или неточности, модель может быть затруднено обучиться.

3. Неподходящий выбор модели: некоторые модели могут быть неэффективными для конкретного типа данных или задачи.

4. Недостаточное исследование признаков: возможно, не все признаки были правильно обработаны или использованы при построении модели.

6. Недостаточное понимание задачи: иногда недостаточное понимание целей и требований заказчика может привести к неправильному подходу к построению модели.

5. Недостаточное исследование гиперпараметров: выбор оптимальных гиперпараметров модели также может оказать значительное влияние на ее эффективность.

ЦЕНТР
ДОПОЛНИТЕЛЬНОГО
ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

do.bmstu.ru

Спасибо за внимание!