ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

Выпускная квалификационная работа
по курсу «Data Science»

Тема:
Прогнозирование конечных свойств новых материалов (композиционных материалов)

Васягин Игорь Евгеньевич

ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

Этапы работы

1. Разведочный анализ данных, визуализация
2. Предобработка данных: удаление выбросов, стандартизация, разделение на тренинг и тест
3. Разработка и обучение моделей «Модуль упругости при растяжении, ГПа», «Прочность при растяжении, МПа» и нейросети для рекомендации «Соотношение матрица-наполнитель»
4. Тестирование моделей, анализ ошибок
5. Выводы

ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

Анализ данных

```
Int64Index: 1023 entries, 0 to 1022
Data columns (total 13 columns):
 #   Column                                Non-Null Count  Dtype
---  ------                                --------------  -----
 0   Соотношение матрица-наполнитель       1023 non-null   float64
 1   Плотность, кг/м3                      1023 non-null   float64
 2   Модуль упругости, ГПа                 1023 non-null   float64
 3   Количество отвердителя, м.%           1023 non-null   float64
 4   Содержание эпоксидных групп,%_2       1023 non-null   float64
 5   Температура вспышки, С_2              1023 non-null   float64
 6   Поверхностная плотность, г/м2         1023 non-null   float64
 7   Модуль упругости при растяжении, ГПа  1023 non-null   float64
 8   Прочность при растяжении, МПа         1023 non-null   float64
 9   Потребление смолы, г/м2               1023 non-null   float64
 10  Угол нашивки, град                    1023 non-null   float64
 11  Шаг нашивки                           1023 non-null   float64
 12  Плотность нашивки                     1023 non-null   float64
dtypes: float64(13)
```

```
Соотношение матрица-наполнитель          1014
Плотность, кг/м3                         1013
Модуль упругости, ГПа                    1020
Количество отвердителя, м.%              1005
Содержание эпоксидных групп,%_2          1004
Температура вспышки, С_2                 1003
Поверхностная плотность, г/м2            1004
Модуль упругости при растяжении, ГПа     1004
Прочность при растяжении, МПа            1004
Потребление смолы, г/м2                  1003
Угол нашивки, град                          2
Шаг нашивки                               989
Плотность нашивки                         988
dtype: int64
```

	count	mean	std	min	25%	50%	75%	max
Соотношение матрица-наполнитель	1023.0	2.930366	0.913222	0.389403	2.317887	2.906878	3.552660	5.591742
Плотность, кг/м3	1023.0	1975.734888	73.729231	1731.764635	1924.155467	1977.621657	2021.374375	2207.773481
Модуль упругости, ГПа	1023.0	739.923233	330.231581	2.436909	500.047452	739.664328	961.812526	1911.536477
Количество отвердителя, м.%	1023.0	110.570769	28.295911	17.740275	92.443497	110.564840	129.730366	198.953207
Содержание эпоксидных групп,%_2	1023.0	22.244390	2.406301	14.254985	20.608034	22.230744	23.961934	33.000000
Температура вспышки, С_2	1023.0	285.882151	40.943260	100.000000	259.066528	285.896812	313.002106	413.273418
Поверхностная плотность, г/м2	1023.0	482.731833	281.314690	0.603740	266.816645	451.864365	693.225017	1399.542362
Модуль упругости при растяжении, ГПа	1023.0	73.328571	3.118983	64.054061	71.245018	73.268805	75.356612	82.682051
Прочность при растяжении, МПа	1023.0	2466.922843	485.628006	1036.856605	2135.850448	2459.524526	2767.193119	3848.436732
Потребление смолы, г/м2	1023.0	218.423144	59.735931	33.803026	179.627520	219.198882	257.481724	414.590628
Угол нашивки, град	1023.0	44.252199	45.015793	0.000000	0.000000	0.000000	90.000000	90.000000
Шаг нашивки	1023.0	6.899222	2.563467	0.000000	5.080033	6.916144	8.586293	14.440522
Плотность нашивки	1023.0	57.153929	12.350969	0.000000	49.799212	57.341920	64.944961	103.988901

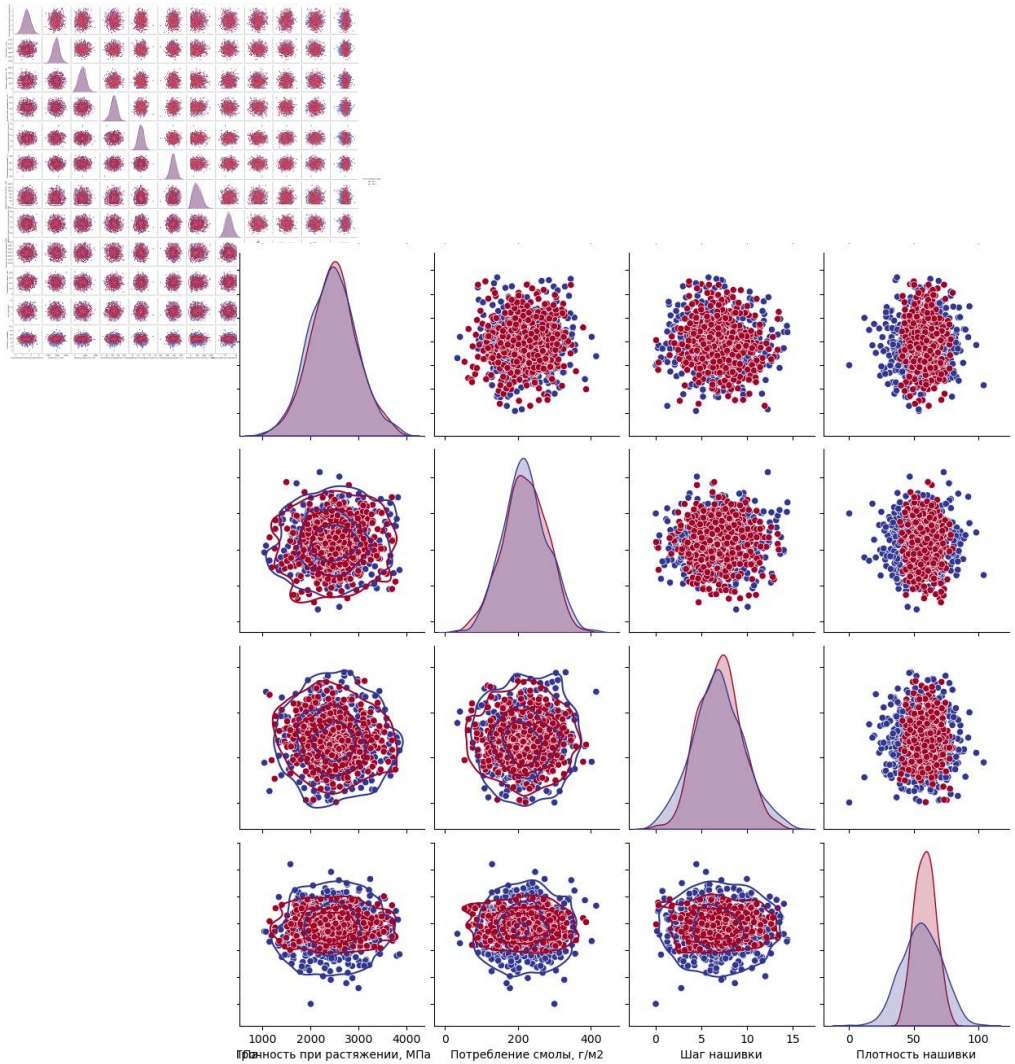
13 признаков – 1023 строки

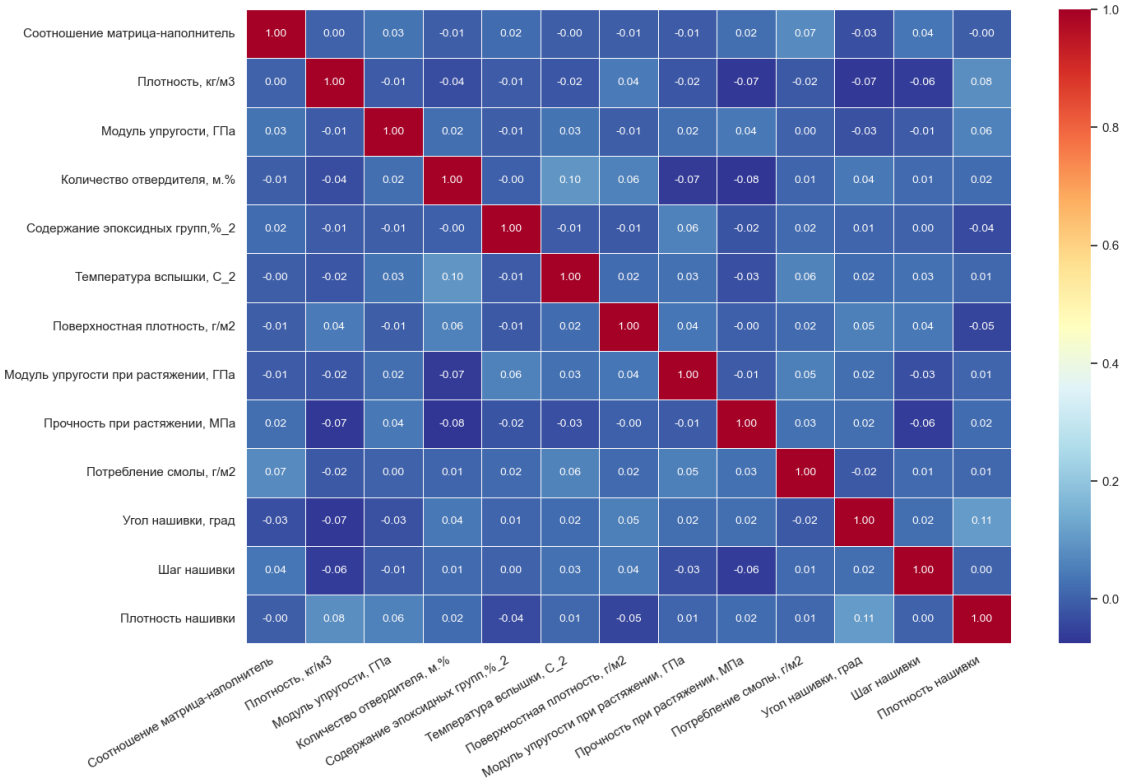
Отсутствие нулевых значений

Уникальность строк

ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

Визуализация



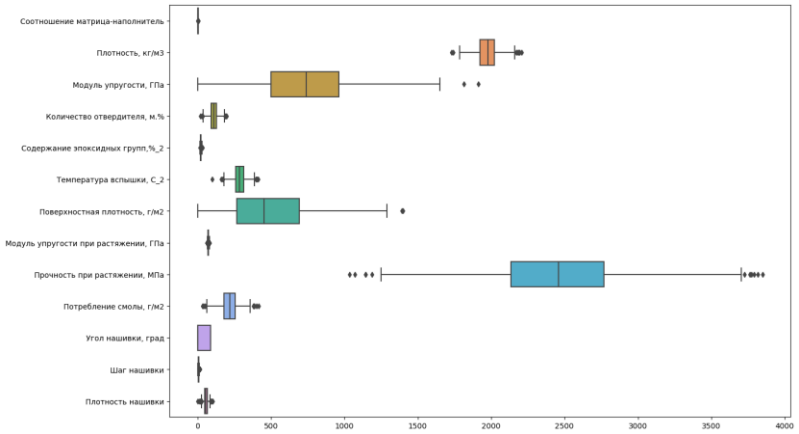


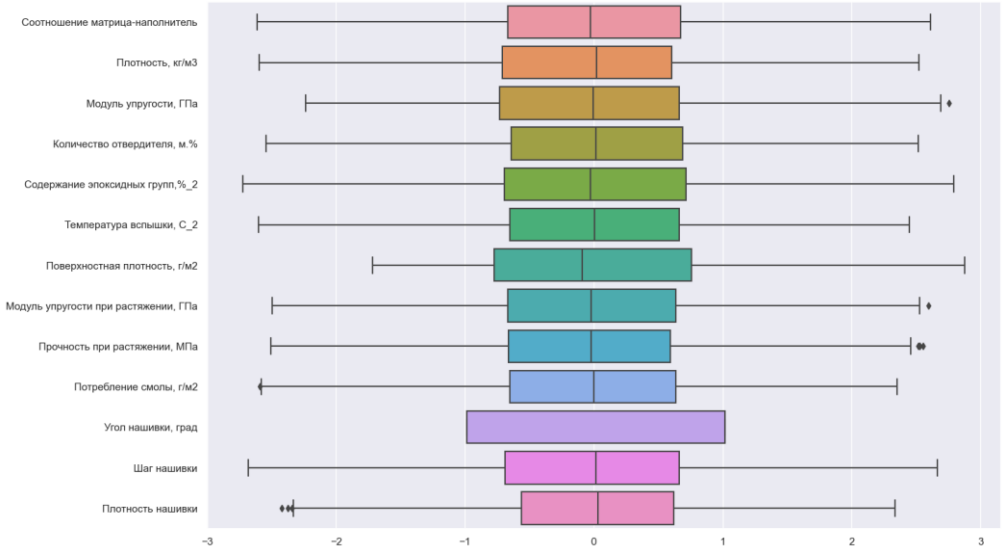
Значения Тепловой карты близки к 0

Попарные графики не демонстрируют корреляцию между признаками

ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

Обработка данных





	count	mean	std	min	25%	50%	75%	max
Соотношение матрица-наполнитель	936.0	-0.002047	0.278720	-0.736970	-0.202569	-0.008203	0.199412	0.652068
Плотность, кг/м3	936.0	-0.005779	0.272050	-0.702335	-0.204961	0.006913	0.189967	0.725274
Модуль упругости, ГПа	936.0	-0.002458	0.281938	-0.649405	-0.217683	-0.001201	0.198212	0.713835
Количество отвердителя, м.%	936.0	0.007026	0.270797	-0.722259	-0.187585	0.005521	0.205621	0.734216
Содержание эпоксидных групп,%_2	936.0	-0.005639	0.281418	-0.696394	-0.207710	-0.007297	0.210553	0.667124
Температура вспышки, С_2	936.0	0.001436	0.273659	-0.793328	-0.195383	0.001091	0.194493	0.750782
Поверхностная плотность, г/м2	936.0	-0.002983	0.283842	-0.630006	-0.238900	-0.029166	0.213642	0.760946
Модуль упругости при растяжении, ГПа	936.0	-0.002247	0.275719	-0.672087	-0.203252	-0.006188	0.196761	0.709015
Прочность при растяжении, МПа	936.0	0.001053	0.269270	-0.652666	-0.203522	-0.005199	0.189479	0.752336
Потребление смолы, г/м2	936.0	-0.004237	0.273329	-0.691941	-0.195633	-0.000239	0.192672	0.747426
Угол нашивки, град	936.0	0.017126	0.307042	-0.600179	-0.281021	0.218917	0.302881	0.545564
Шаг нашивки	936.0	-0.000481	0.278545	-0.716617	-0.204390	0.005955	0.203298	0.704588
Плотность нашивки	936.0	0.007557	0.257474	-0.665016	-0.165191	0.010254	0.187391	0.656576

Выбросы удаляем методом IQR

Стандартизируем датасет

Разделяем на тренинг и тест

ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

Разработка, обучение моделей

```python
def build_model(hp):

    model = keras.Sequential()

#    model.add(layers.Flatten())
    model.add(LayerNormalization())

    for i in range(hp.Int('num_layers', 1, 5)):
        model.add(layers.Dense(units=hp.Int(f'units_{i}', min_value=4,
                                            max_value=12, step=2),
                               activation=hp.Choice('activation', ['relu', 'tanh']),
                              )
                )

    if hp.Boolean('dropout'):
        model.add(layers.Dropout(rate=0.25))

    model.add(layers.Dense(1,))

    optimizer = hp.Choice('optimizer', ['sgd','rmsprop','adam'])

    learning_rate = hp.Choice('learning_rate', values = [.1, .01, .001, .0001])

    model.compile(optimizer=optimizer,
                  loss='mean_squared_error',
                  metrics=['mean_squared_error'],
                 )

    return model
```

```python
param_decision = {'criterion': ['squared_error', 'friedman_mse', 'absolute_error'],
                  'splitter': ['best', 'random'],
                  'max_depth': range (5, 15, 1),
                  'min_samples_leaf': range (2, 4, 1),
                  'min_samples_split': range (2, 4, 1),
                  'max_features': range(1, 11, 1)
                 }

param_random = {'criterion': ['squared_error', 'friedman_mse', 'absolute_error'],
                'max_depth': range (1, 3, 1),
                'min_samples_leaf': range (2, 6, 1),
                'min_samples_split': range (2, 10, 1),
                'n_estimators': range (6, 12, 1)
               }

param_gradient = {'loss': ['squared_error', 'absolute_error', 'huber', 'quantile'],
                  'criterion': ['friedman_mse', 'squared_error'],
                  'max_features': ['sqrt', 'log2'],
                  'max_depth': range (1, 5, 1)
```

```python
Bayes = kt.BayesianOptimization(hypermodel=build_model, objective='val_loss',
                                max_trials=50, executions_per_trial=3, overwrite=True)

Random = kt.RandomSearch(hypermodel=build_model, objective='val_loss',
                         max_trials=50, executions_per_trial=3, overwrite=True)

Hyperband = kt.Hyperband(hypermodel=build_model, objective='val_loss',
                         max_epochs=50, executions_per_trial=3, overwrite=True)
```

Модели для поиска лучших параметров

GridSearchCV & Tuner Keras

ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

Тестирование моделей

Тестовые и прогнозные значения: GRADIENT BOOSTING

Тестовые и прогнозные значения: LASSO

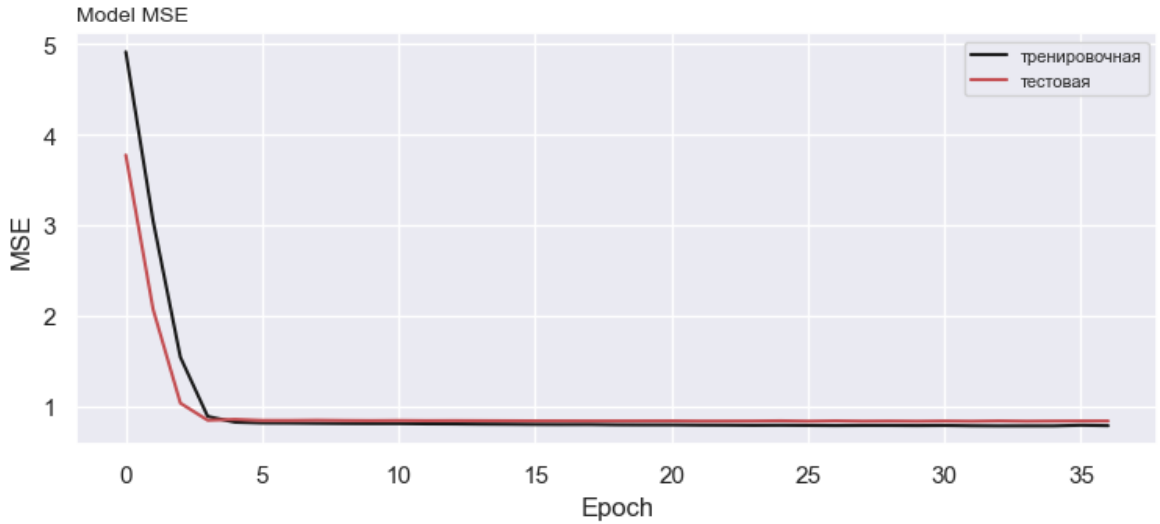
[Dummy, Elastic, LASSO, SVR]
около нулевой результат

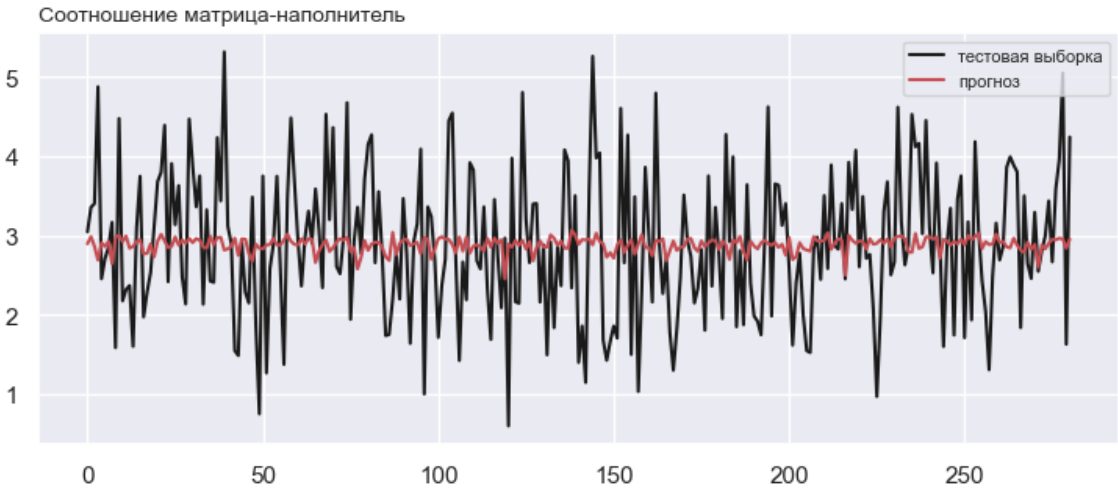
[Decision Tree, Gradient Boosting, Random Forest, K-Neighbors, Linear, Ridge]
визуально лучший график

ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

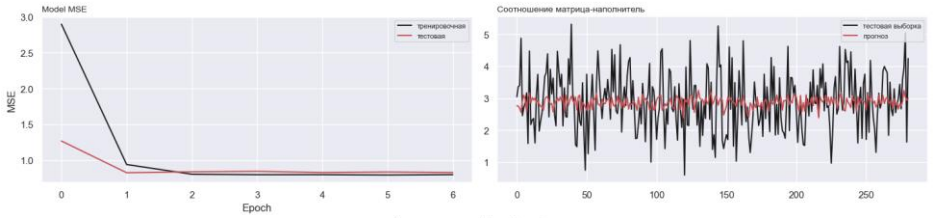
Тестирование моделей

Класс тюнера: Random Search

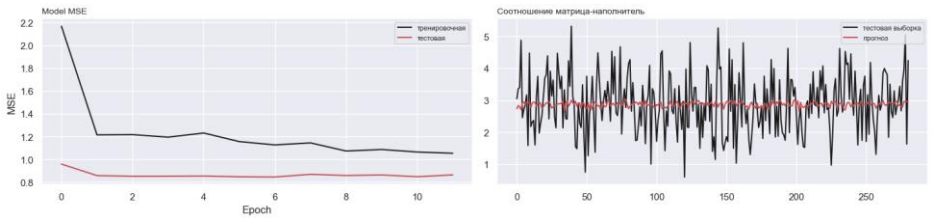




Класс тюнера: Bayesian Optimization



Класс тюнера: Hyperband



[BayesianOptimization,
RandomSearch,
Hyperband]

показали схожие результаты

ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

Ошибки

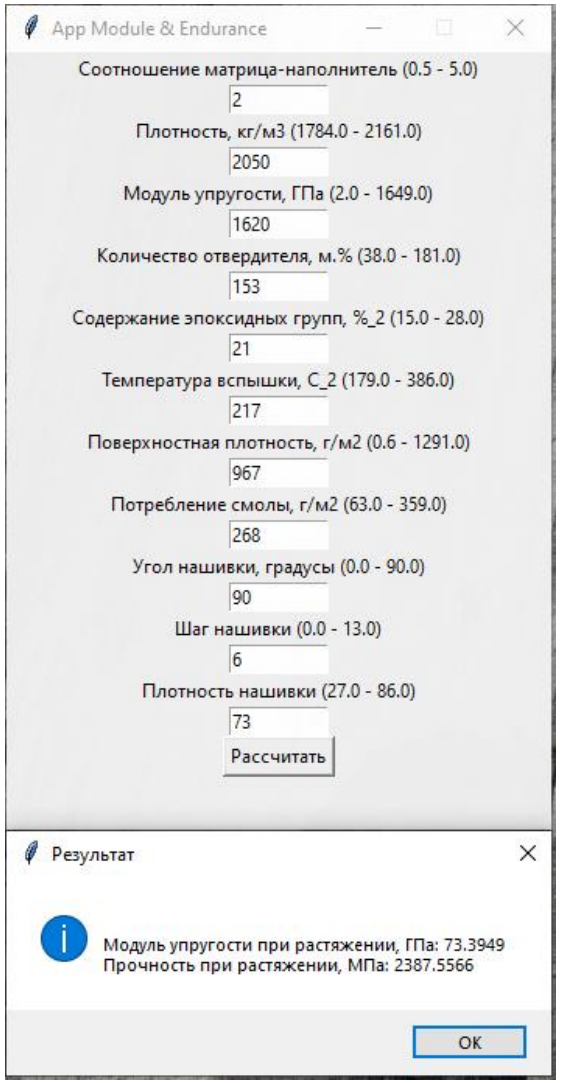
	name	feature	mse	rmse	mae	r2	mape
1	LINEAR	модуль	0.072562	0.269373	0.217889	0.019288	1.352428
2		прочность	0.071976	0.268284	0.217090	-0.031393	9.434725
3	DUMMY	модуль	0.074034	0.272091	0.220546	-0.000602	1.002576
4		прочность	0.069822	0.264239	0.213066	-0.000531	6.574755
5	LASSO	модуль	0.074034	0.272091	0.220546	-0.000602	1.002576
6		прочность	0.069822	0.264239	0.213066	-0.000531	6.574755
7	RIDGE	модуль	0.072577	0.269401	0.217926	0.019085	1.346937
8		прочность	0.071897	0.268136	0.216948	-0.030259	9.296989
9	ELASTIC	модуль	0.074034	0.272091	0.220546	-0.000602	1.002576
10		прочность	0.069822	0.264239	0.213066	-0.000531	6.574755
11	K-NEIGHBORS	модуль	0.074317	0.272611	0.221565	-0.004426	1.445657
12		прочность	0.072537	0.269328	0.215942	-0.039435	183.634952
13	DECISION TREE	модуль	0.074954	0.273778	0.221138	-0.013047	1.360824
14		прочность	0.070986	0.266432	0.215070	-0.017206	21.991708
15	RANDOM FOREST	модуль	0.073929	0.271900	0.220563	0.000807	1.122600
16		прочность	0.071029	0.266512	0.214664	-0.017816	68.978934
17	GRADIENT BOOSTING	модуль	0.076182	0.276012	0.223623	-0.029645	1.687428
18		прочность	0.071058	0.266568	0.216150	-0.018241	2.355036
19	SVR	модуль	0.073975	0.271984	0.220236	0.000186	1.096575
20		прочность	0.069821	0.264236	0.212943	-0.000504	18.598072

	name	mse	rmse	mae	r2	mape
1	Bayesian Optimization	0.824209	0.907859	0.735155	0.019836	0.314291
2	Random Search	0.835548	0.914083	0.741446	0.006351	0.318499
3	Hyperband	0.844739	0.919097	0.741930	-0.004579	0.317525

Регрессоры (по обоим признакам в отдельности)
и нейросеть дали высокий уровень ошибок

ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

Приложение



Приложение использует K-Neighbors

Модель использует лучшие параметры, полученные на стандартизированных и нормализованных данных, очищенных от выбросов

Обучение модели происходит на фактических данных, очищенных от выбросов

ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

Заключение

1. Распределение данных предложенного Датасета соответствует нормальному
2. Корреляция между парами признаков стремится к нулю
3. Использованные при разработке моделей методы не позволили получить достоверные прогнозы
4. Применённые модели регрессии и нейросеть не показали эффективности в прогнозировании свойств композитов
5. В ходе выполнения выпускной работы были выполнены все задачи

ЦЕНТР
ДОПОЛНИТЕЛЬНОГО
ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

do.bmstu.ru