ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА
по курсу «Data Science»

Тема: Прогнозирование конечных свойств новых материалов (композиционных материалов)

Трещева Елена

ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

Описание задачи

- Основной задачей является поиск новых композитных материалов, которые могут иметь лучшие свойства чем существующие материалы.
- С помощью анализа данных и разработки моделей машинного обучения необходимо разработать систему предсказаний свойств новых композитных материалов.
- Для решения этой задачи могут быть использованы различные методы машинного обучения, такие как различные подвиды линейной регрессия, случайный лес, нейронные сети. Каждый из этих методов имеет свои преимущества и недостатки, поэтому в процессе выбора необходимо выбрать модели машинного обучения, которая больше подойдет для предсказания характеристик композитных материалов наиболее точно на основании предоставленного набора данных для исследования (данные о характеристиках базальтопластика и о углепластика).

ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

Качество данных

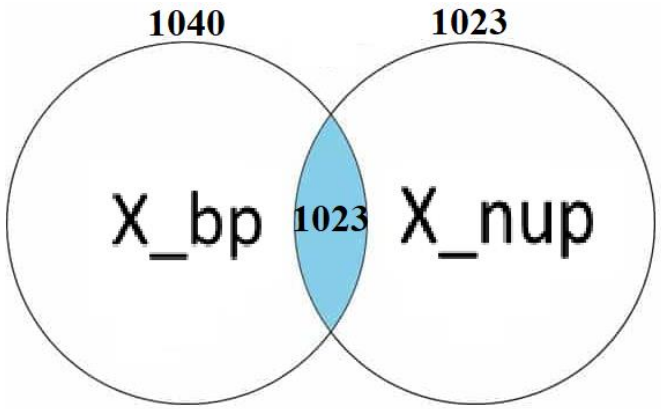
- Для проведения исследования получены два набора данных

X_bp.xlsx базальтопластика в файле 1023 строки и 10 полей с данными;

X_nup.xlsx, файл, углепластика в файле 1040 сток и 3 поля с данными;

Объединить поля можно по полю-индексу, который находится в каждом файле

- Количество записей после объединения наборов данных X_bp ∩ X_nup составил 1023 записи
- Данные были загружены в DataFrame при помощи билиотеки Pandas Python и сохранены в csv Формате после преобразования.



ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

Качество данных

- Около 5% записей имело целочисленный тип данных, при том что остальная часть набора данных Является вещественными числами с плавающей запятой с точностью до 12 знаков после запятой.
- В результате выявлено, при помощи пользовательской функции, что поле поле «Плотность нашивки» - имеет наибольшее число целочисленных данных.

```
def find_astype_int_records(df, persent = 5):
    max_variation_col_nm = ""
    max_variation_col_cnt = 0
    for cols in df.columns:
        total_row_cnt = df[cols].count()
        cnt_astype_int = len([el for el in (df[cols].astype(int) - df[cols]) if el == 0])
        print("Всего записей {} из них целыми являются {}, что составляет: {}%".format(total_row_cnt
                                                                                        , cnt_astype_int
                                                                                        , round(cnt_astype_int/total_row_cnt,2)*100))
        if (round(cnt_astype_int/total_row_cnt, 2)*100) < persent:
            if cnt_astype_int > max_variation_col_cnt:
                max_variation_col_nm = cols
    if max_variation_col_nm == "":
        print("Столбцов - отклонений не найдено")
    else:
        print("Столбец {} имеет наибольшее число отклонений".format(max_variation_col_nm))
    return max_variation_col_nm
max_int_values_found = find_astype_int_records(inner_df)
print(max_int_values_found)
```

ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

Качество данных

- После проведения анализа о распределении целочисленных данных – строки, содержащие такие данные были удалены, с использованием следующего условия:

число не имеет знаков после запятой

- После очистки данных проверка

была проведена повторно

, как видно на изображении целые

числа присутствуют только в одном

Поле «Угол нашивки», последнее

Имеет только два значения [0, 90]

```
1  max_int_values_found = find_astype_int_records(inner_df_res)
2  print(max_int_values_found)
3  inner_df = inner_df_res
```

```
Всего записей 983 из них целыми являются 0, что составляет: 0.0%
Всего записей 983 из них целыми являются 0, что составляет: 0.0%
Всего записей 983 из них целыми являются 0, что составляет: 0.0%
Всего записей 983 из них целыми являются 0, что составляет: 0.0%
Всего записей 983 из них целыми являются 0, что составляет: 0.0%
Всего записей 983 из них целыми являются 0, что составляет: 0.0%
Всего записей 983 из них целыми являются 0, что составляет: 0.0%
Всего записей 983 из них целыми являются 0, что составляет: 0.0%
Всего записей 983 из них целыми являются 0, что составляет: 0.0%
Всего записей 983 из них целыми являются 0, что составляет: 0.0%
Всего записей 983 из них целыми являются 983, что составляет: 100.0%
Всего записей 983 из них целыми являются 0, что составляет: 0.0%
Всего записей 983 из них целыми являются 0, что составляет: 0.0%
Столбцов - отклонений не найдено
```

ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

Описание используемых методов

- Регрессия:
  - Линейная регрессия - анализа и моделирование зависимостей между целевой переменной и одним или несколькими входными признаками
  - Регрессия методом k-ближайших соседей - предсказании значения целевой переменной для нового наблюдения
  - Случайный лес - создании множества решающих деревьев с различными характеристиками, и использовании их для прогнозирования целевой переменной
  - Лассо регрессия - минимизация функции потерь, то есть расстояния между прогнозированными значениями целевой переменной и фактическими значениями
- Нейронные сети – сводится к настройке весов между нейронами и определении оптимального количество нейронов в каждом слое сети.

ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

Разведочный анализ данных

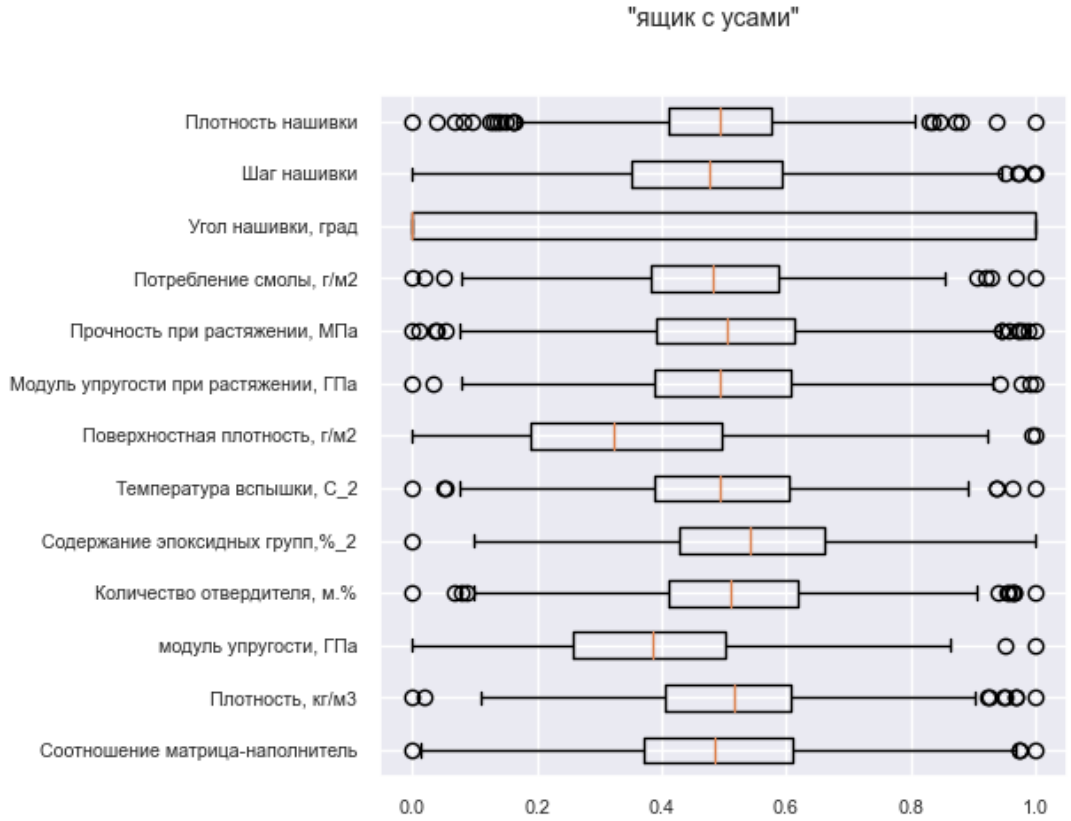
- изучении основных характеристик и закономерностей произвольных данных с помощью графиков, статистического анализа и визуализации, выявления выбросов и аномалий, и определения связей между переменными ('mean', '50%’)

	count	mean	std	min	25%	50%	75%	max	median
Соотношение матрица-наполнитель	983.00	2.93	0.92	0.39	2.32	2.91	3.56	5.59	2.91
Плотность, кг/м3	983.00	1975.98	73.76	1731.76	1925.40	1977.66	2021.28	2207.77	1977.66
модуль упругости, ГПа	983.00	737.74	329.63	2.44	497.87	738.96	962.21	1911.54	738.96
Количество отвердителя, м.%	983.00	110.49	28.30	17.74	92.11	110.10	129.88	198.95	110.10
Содержание эпоксидных групп,%_2	983.00	22.23	2.41	14.25	20.55	22.21	23.98	28.96	22.21
Температура вспышки, С_2	983.00	285.66	40.96	160.26	258.40	285.41	313.06	413.27	285.41
Поверхностная плотность, г/м2	983.00	482.19	281.69	0.60	267.14	451.31	694.34	1399.54	451.31
Модуль упругости при растяжении, ГПа	983.00	73.32	3.12	64.05	71.29	73.24	75.36	82.68	73.24
Прочность при растяжении, МПа	983.00	2464.78	484.96	1036.86	2136.60	2456.40	2759.08	3848.44	2456.40
Потребление смолы, г/м2	983.00	218.68	59.94	33.80	179.63	217.48	257.63	414.59	217.48
Угол нашивки, град	983.00	44.22	45.02	0.00	0.00	0.00	90.00	90.00	0.00
Шаг нашивки	983.00	6.90	2.57	0.04	5.12	6.89	8.56	14.44	6.89
Плотность нашивки	983.00	57.15	12.36	11.74	49.80	57.34	64.94	103.99	57.34

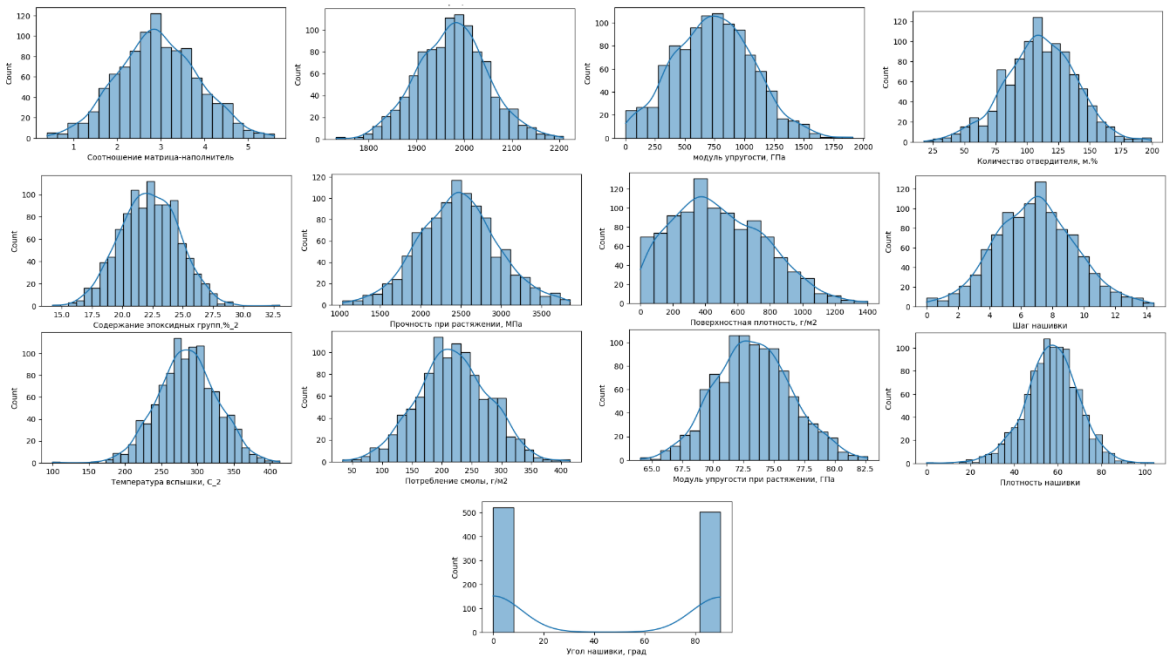
- Так же была выполнена проверка на отсутсвие пропущенных значений (тепловая карта)

ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

Разведочный анализ данных



Диаграммы «Ящик с усам»



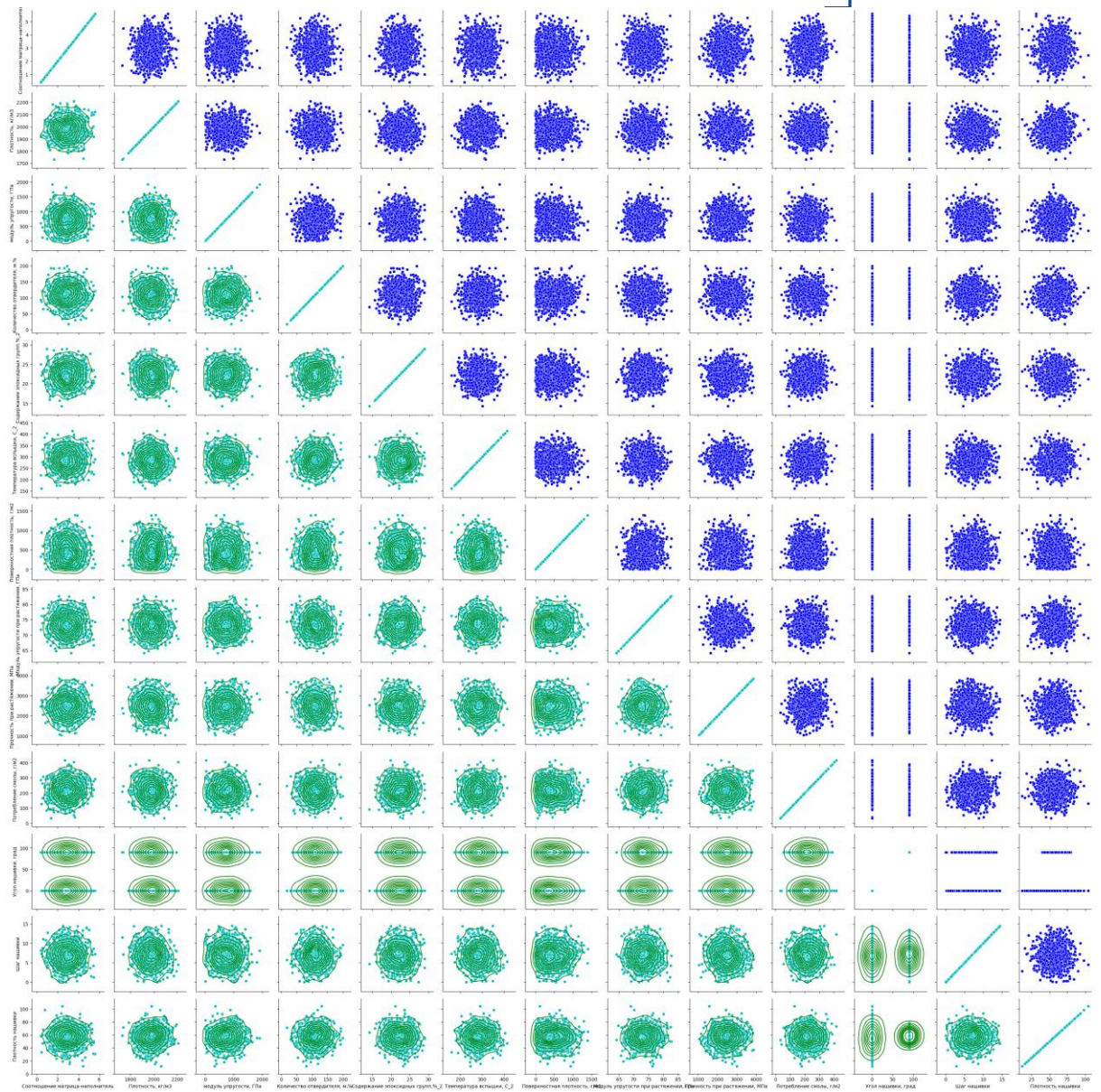
Диаграммы «Распределения данных по каждой переменной»
На нижнем график – поле «Угол нашивки»

ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

Разведочный анализ данных

- Попарные графики рассеивания атрибутов (признаков) данных

На данном наборе данных не выявлены явные закономерности



ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

Разведочный анализ данных. Выбросы

- Было получено распределение выбросов по атрибутам данных методами: 3-х сигм и межквартильного интервала
- По результатам из набора

Данных удалены строки, которые содержали выбросы установленные при помощи метода 3-х сигм

```
метод 3-х сигм: 0 метод irq: 4 - атрибут Соотношение матрица-наполнитель
метод 3-х сигм: 3 метод irq: 9 - атрибут Плотность, кг/м3
метод 3-х сигм: 2 метод irq: 2 - атрибут модуль упругости, ГПа
метод 3-х сигм: 2 метод irq: 11 - атрибут Количество отвердителя, м.%
метод 3-х сигм: 1 метод irq: 1 - атрибут Содержание эпоксидных групп,%_2
метод 3-х сигм: 2 метод irq: 7 - атрибут Температура вспышки, С_2
метод 3-х сигм: 2 метод irq: 2 - атрибут Поверхностная плотность, г/м2
метод 3-х сигм: 1 метод irq: 6 - атрибут Модуль упругости при растяжении,
метод 3-х сигм: 0 метод irq: 13 - атрибут Прочность при растяжении, МПа
метод 3-х сигм: 3 метод irq: 8 - атрибут Потребление смолы, г/м2
метод 3-х сигм: 0 метод irq: 0 - атрибут Угол нашивки, град
метод 3-х сигм: 0 метод irq: 5 - атрибут Шаг нашивки
метод 3-х сигм: 6 метод irq: 20 - атрибут Плотность нашивки
```

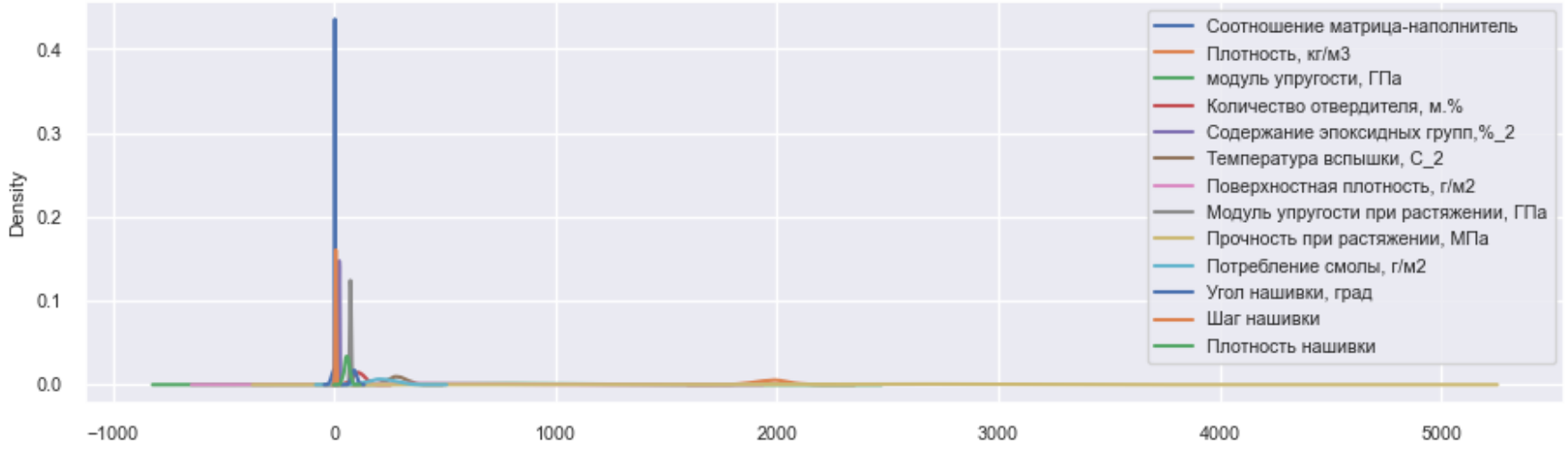
```
Соотношение матрица-наполнитель          920
Плотность, кг/м3                         920
модуль упругости, ГПа                    920
Количество отвердителя, м.%              920
Содержание эпоксидных групп,%_2          920
Температура вспышки, С_2                 920
Поверхностная плотность, г/м2            920
Модуль упругости при растяжении, ГПа     920
Прочность при растяжении, МПа            920
Потребление смолы, г/м2                  920
Угол нашивки, град                       920
Шаг нашивки                              920
Плотность нашивки                        920
dtype: int64
```

После удаления выбрросов набор данных имеет 13 полей и 920 записей

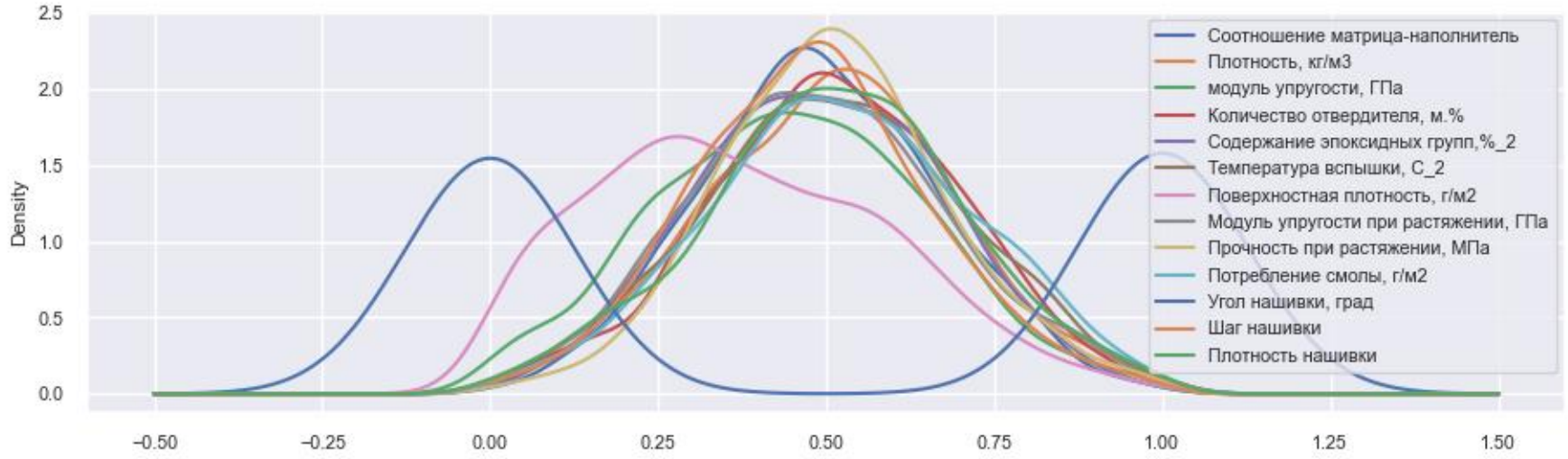
ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

Описание задачи

- Оценка плотности ядра до нормализации:



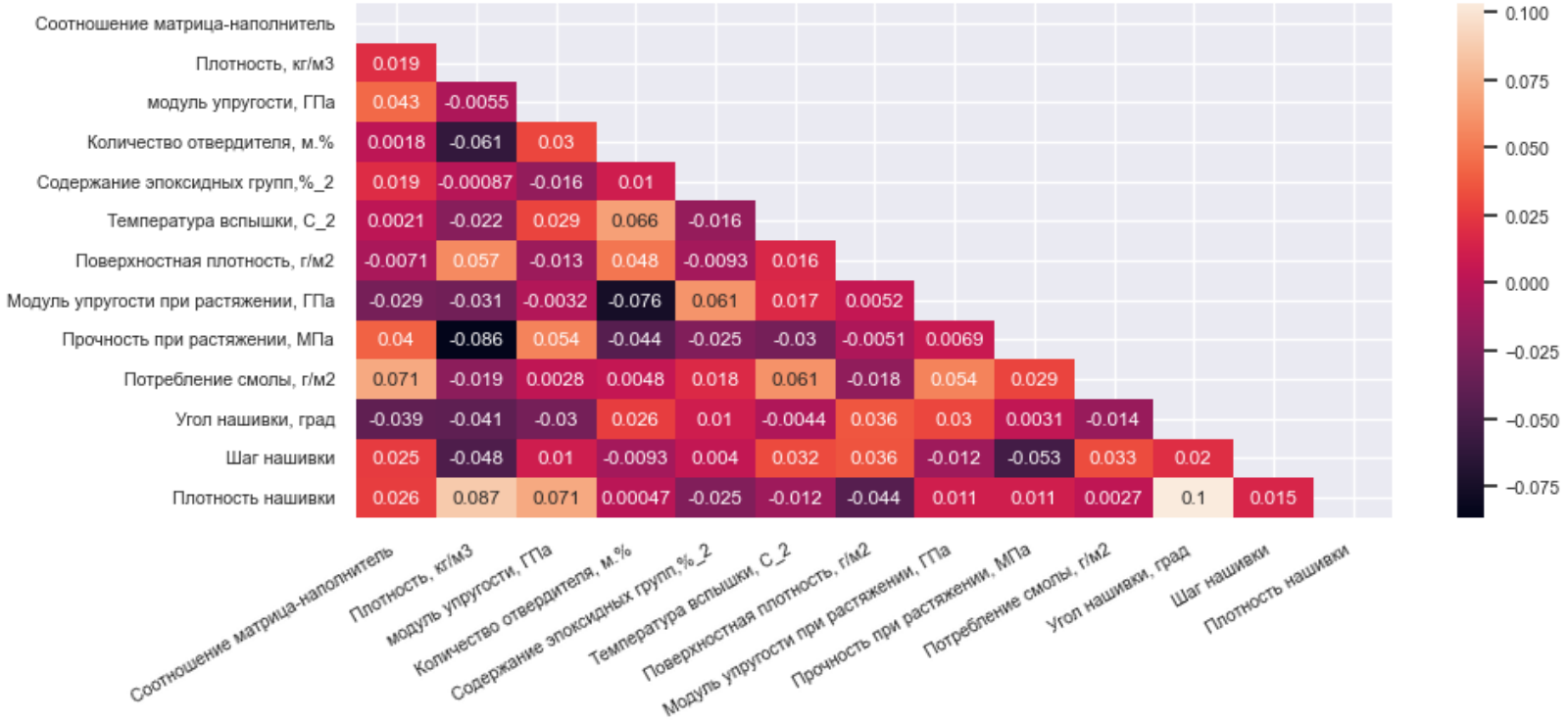
- Оценка плотности ядра после Нормализации при помощи MinMaxScaler:



ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

Разведочный анализ данных. Корреляции

- Составлена матрица корреляции после нормализации данныъ



- Сильных или явных признаков корреляции не выявлено

ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

Разработка и обучение модели

- Применяемые модели:
  - Линейная регрессия; sklearn.linear_model.LinearRegression
  - Лассо регрессия; sklearn.linear_model.LassoCV
  - Случайный лес; sklearn.ensemble. GradientBoostingRegressor
  - Метод К ближайших соседей; sklearn.model_selection.GridSearchCV
  - Нейронная сеть; tensorflow.keras.models. Sequential
- Оценка точности модели при помощи инструментов библиотеки sklearn.metrics :
  - средняя квадратичная ошибка - mean_squared_error;
  - средняя абсолютная ошибка - mean_absolute_error;
  - коэффициент детерминации - r2_score;

Все дата сет делится на тренировочный и основной при помощи метода train_test_split sklearn.model_selection

ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

Разработка и обучение модели

- Архитектура нейронной сети для прогнозирования модуля упругости при растяжении

Layer (type) Output Shape Param #

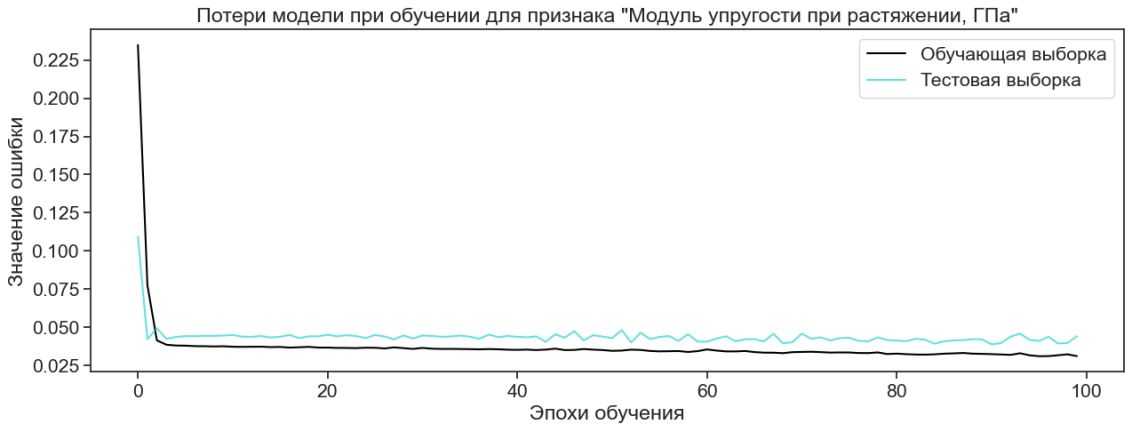
dense (Dense) (None, 128) 1664

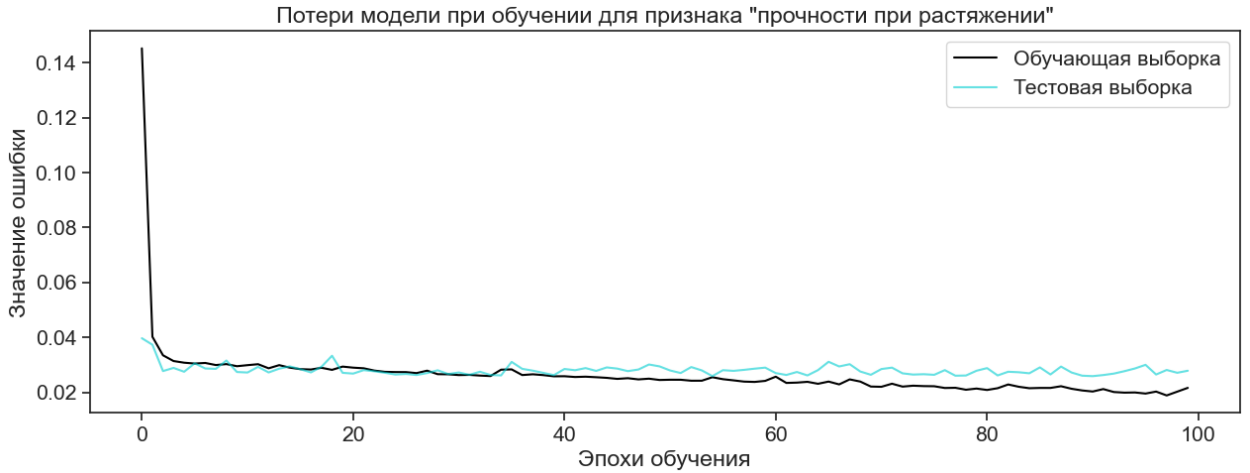
dense_1 (Dense) (None, 8) 1032

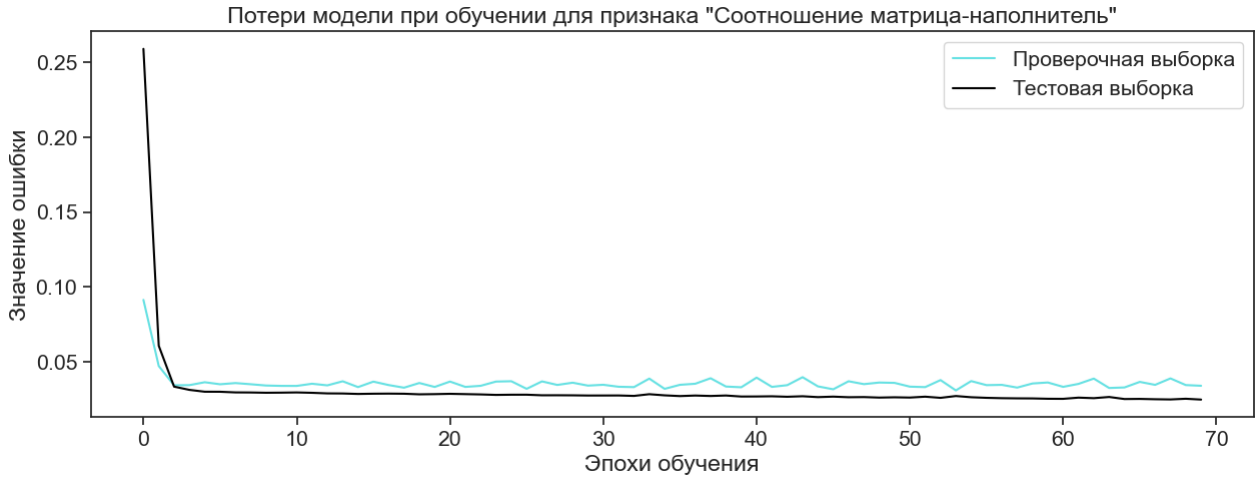
dense_2 (Dense) (None, 8) 72

dense_3 (Dense) (None, 1) 9 - выходной слой с одним нейроном для задачи рег

Total params: 2,777 Trainable params: 2,777 Non-trainable params: 0







ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

Разработка и обучение модели

- **Оценка точности работы модели**

Модуль упругости при растяжении

	Регрессор	MAE	MSE	R2
1	ЛАССО-регрессия	0.156849	0.156849	-0.000028
0	Линейная регрессия	0.155982	0.037360	-0.004109
2	Случайный лес	0.159042	0.159042	-0.039755
4	Нейросеть	0.160830	0.039832	-0.070556
3	Метод К-ближайших соседей	0.216157	0.216157	-0.881344

Прочность при растяжении

	Регрессор	MAE	MSE	R2
2	Случайный лес	0.141529	0.141529	0.004428
0	Линейная регрессия	0.145424	0.033081	-0.004109
1	ЛАССО-регрессия	0.143141	0.143141	-0.004375
4	Нейросеть	0.153287	0.036876	-0.127655
3	Метод К-ближайших соседей	0.205078	0.205078	-1.105589

Соотношение матрица - наполнитель

	Регрессор	MAE	MSE	R2
1	ЛАССО-регрессия	0.148999	0.148999	-0.001983
0	Линейная регрессия	0.151212	0.034224	-0.004109
4	Нейросеть	0.152025	0.033793	-0.005034
2	Случайный лес	0.151666	0.151666	-0.024878
3	Метод К-ближайших соседей	0.198215	0.198215	-0.789185

- В результате выбрана модель линейной Регрессии. Все выбранные показали себя одинаково неэффективно.

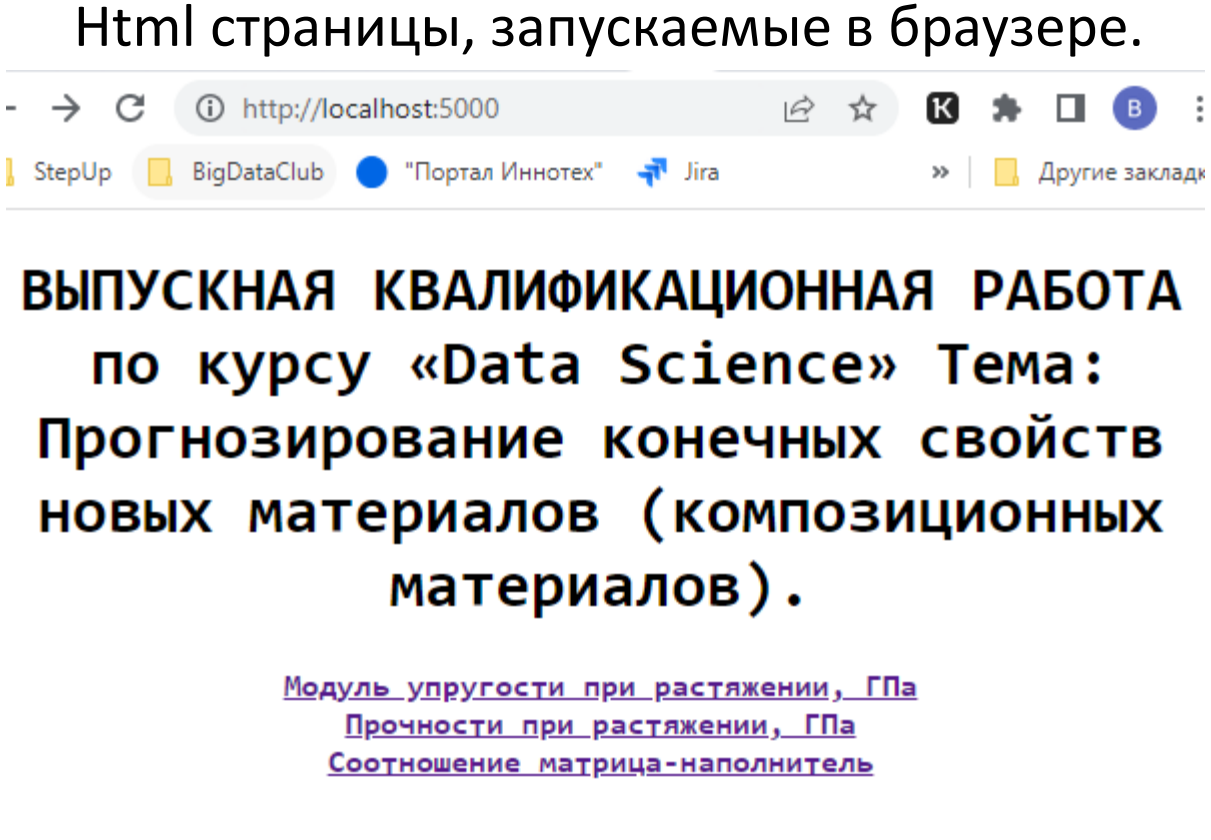
ЦЕНТР ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

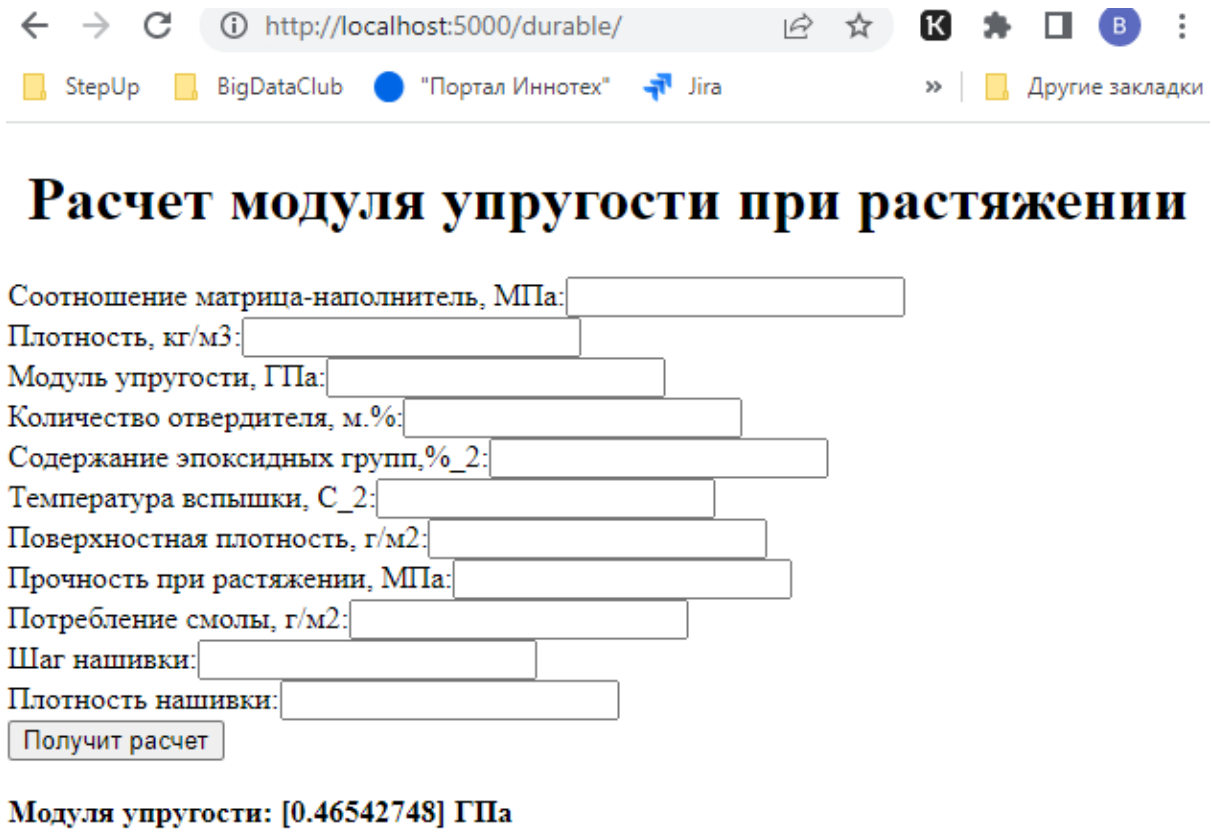
Разработка модели. Выгрузка

- Все подготовленные модели выгружены в формат pkl
- Разработан Web server к которому могут обращаться

Flask
web development, one drop at a time

Html страницы, запускаемые в браузере.





ЦЕНТР
ДОПОЛНИТЕЛЬНОГО
ОБРАЗОВАНИЯ
МГТУ им. Н.Э. Баумана

do.bmstu.ru