

**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»**

**Цель работы - прогнозирование конечных свойств
новых материалов (композиционных материалов)**

- Композиционные материалы — это искусственно созданные материалы, состоящие из нескольких других с четкой границей между ними. Композиты обладают теми свойствами, которые не наблюдаются у компонентов по отдельности. При этом композиты являются монолитным материалом, т. е. компоненты материала неотделимы друг от друга без разрушения конструкции в целом. Современные композиты изготавливаются из других материалов: полимеры, керамика, стеклянные и углеродные волокна, но данный принцип сохраняется. У такого подхода есть и недостаток: даже если мы знаем характеристики исходных компонентов, определить характеристики композита, состоящего из этих компонентов, достаточно проблематично.



Постановка задачи:

- ✓ **Цель решения задачи:** прогноз характеристик композиционного материала на основе имеющихся данных.
- ✓ **Входные данные:**
 - ✓ - общее описание свойств композиционного материала
 - ✓ - два датасета, которые содержат данные о количественных характеристиках различных свойств и составляющих композитного материала. Всего 13 характеристик.
 - ✓ - постановка задач для решения с помощью методов машинного обучения
 - решение задачи регрессии для прогнозирования двух из 13 представленных характеристик
 - разработка рекомендательной системы (задача регрессии) для прогнозирования показателя «Соотношение матрица-наполнитель»

Описательная статистика исходных данных:

	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

Из данных таблицы мы можем посмотреть следующие значения для каждого признака:

count - количество значений

std - стандартное отклонение

25% - верхнее значение первого квартиля

75% - верхнее значение третьего квартиля

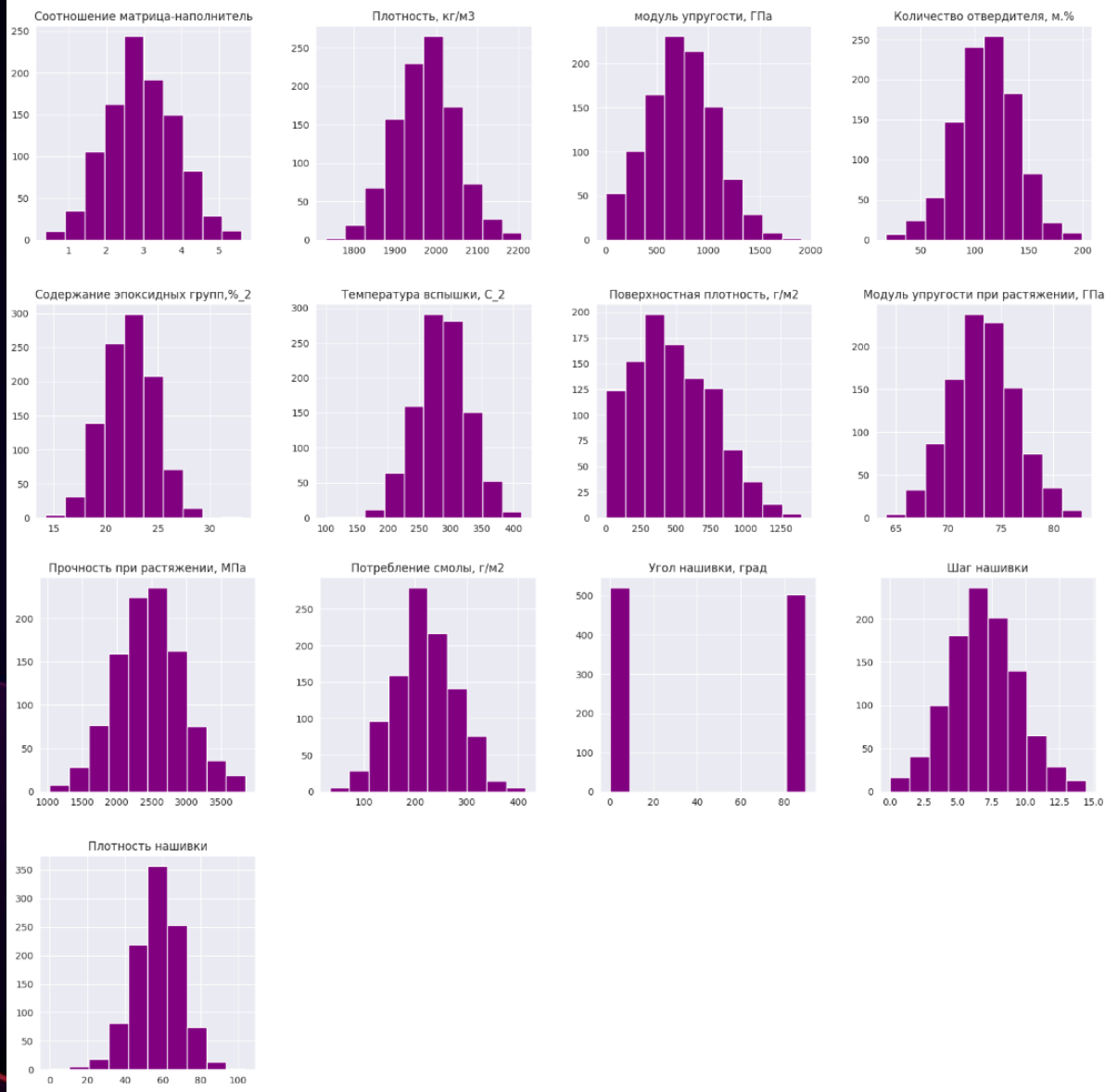
mean - среднее значение

min - минимум

50% - медиана

max - максимум

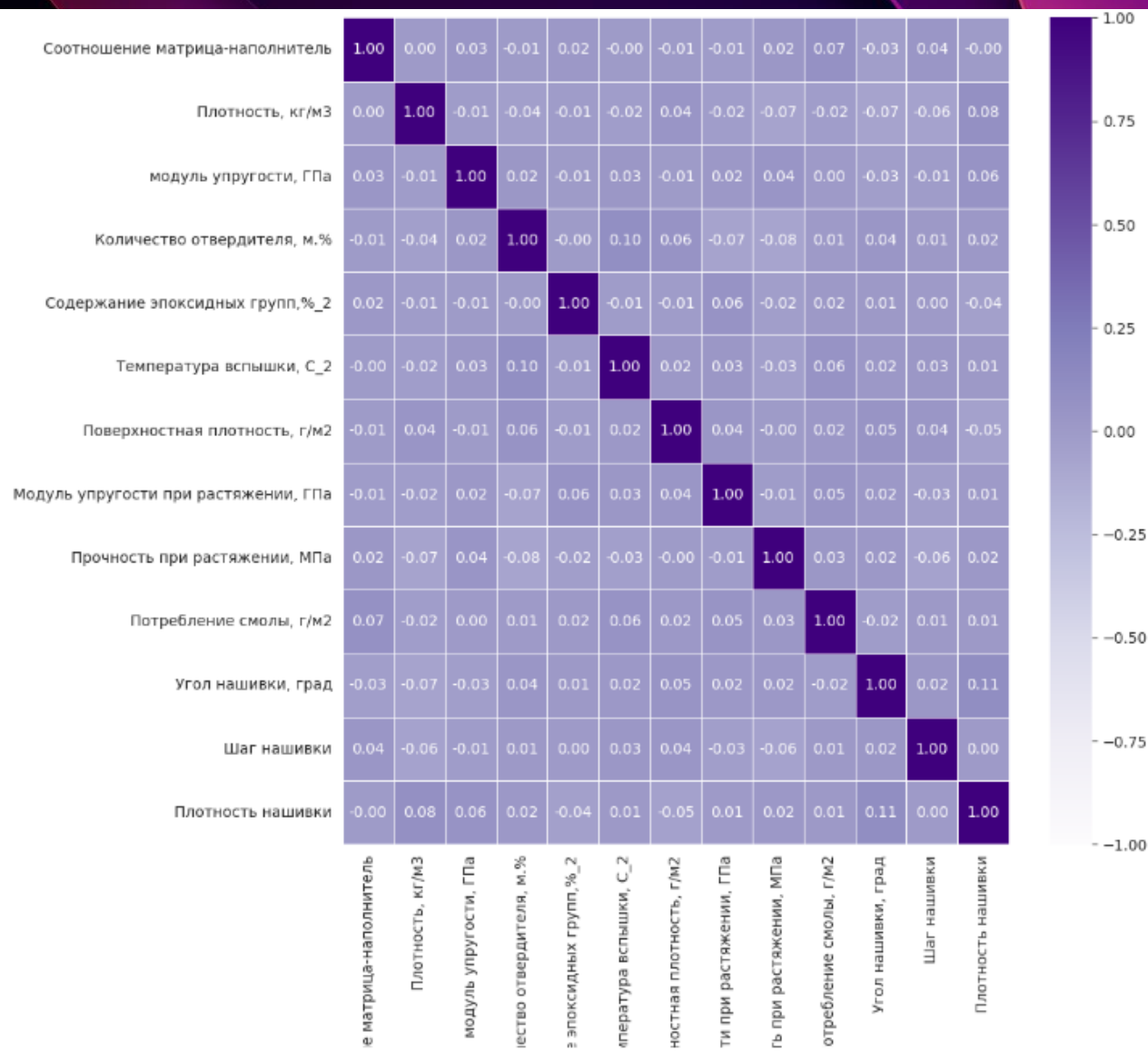
Гистограммы распределения:



Вывод:

1. Угол нашивки является дискретной величиной.
2. Значения Поверхностной плотности имеют Пуассоновское распределение.
3. Все остальные распределения близки к нормальному.

Тепловая карта коэффициентов корреляции:

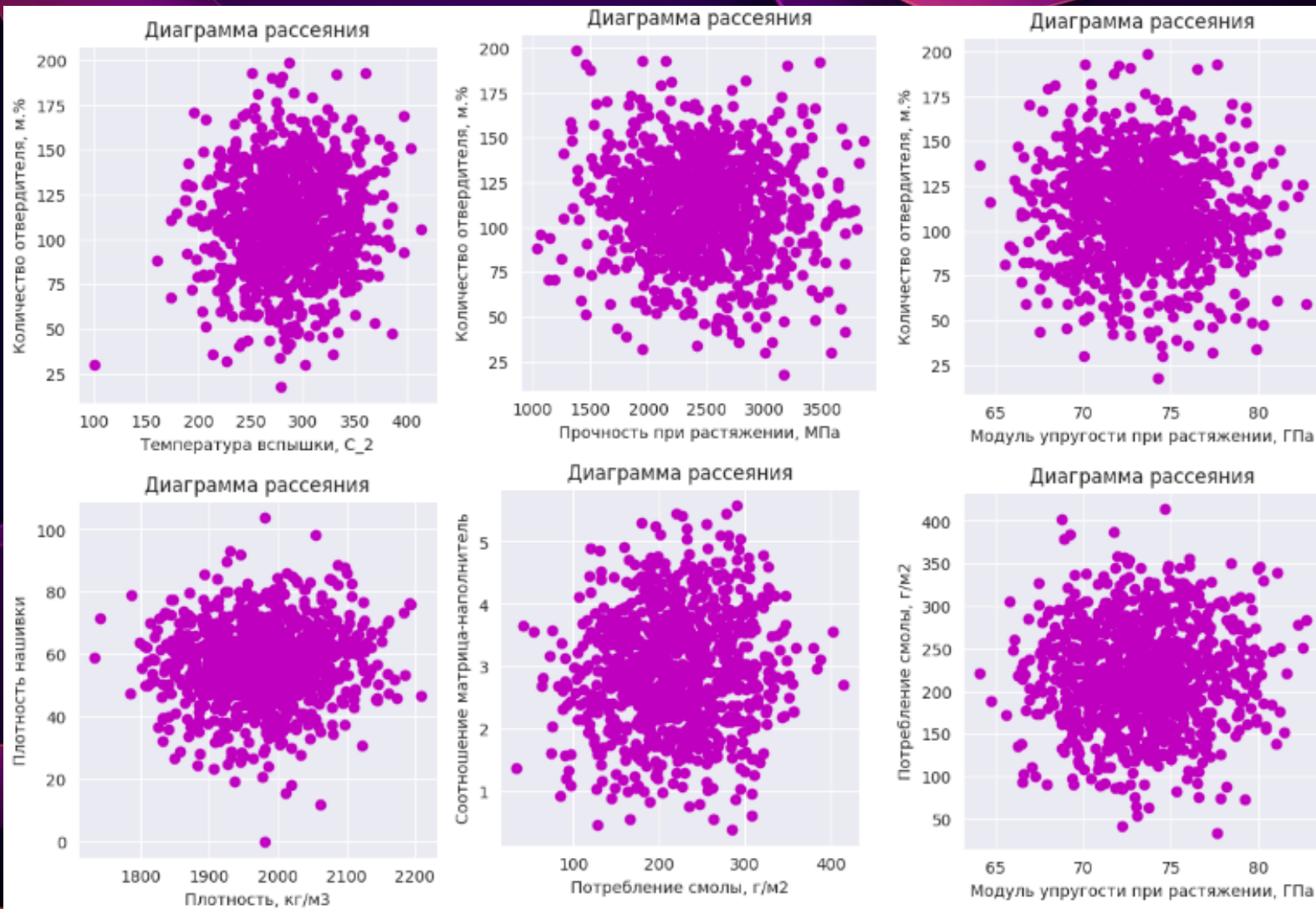


Лучше всего коррелируют между собой (по убыванию):

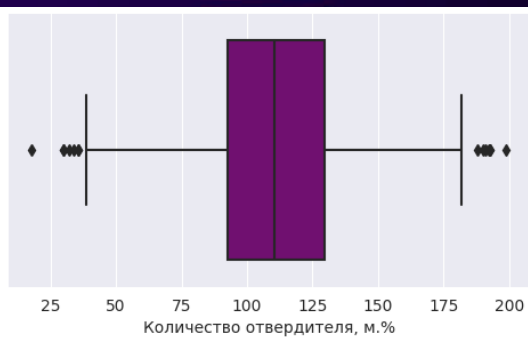
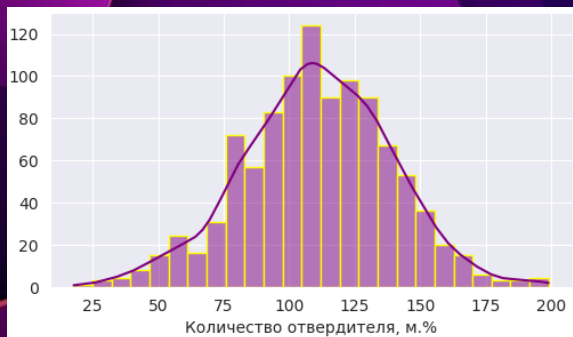
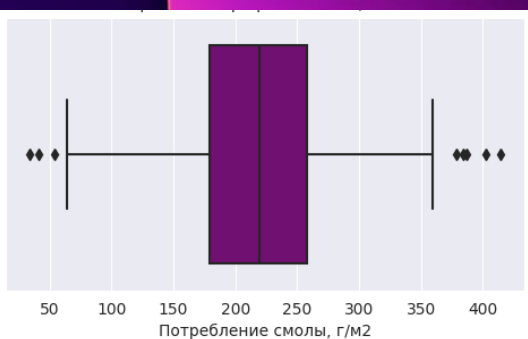
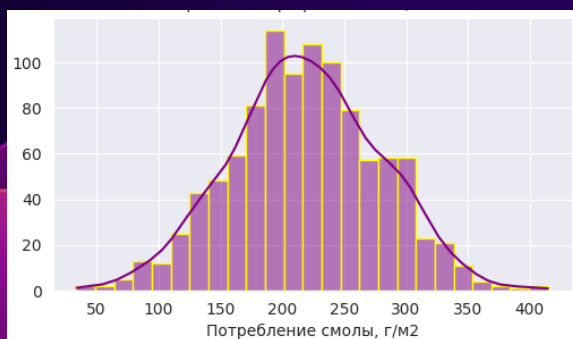
- 1) Угол нашивки и Плотность нашивки; 0,11
- 2) Температура вспышки и Количество отвердителя; 0,10
- 3) Плотность и Плотность нашивки; 0,08
- 4) Прочность при растяжении и Количество отвердителя – обратная корреляция; - 0,08
- 5) Потребление смолы и Соотношение матрица-наполнитель; 0,07
- 6) Модуль упругости при растяжении и Количество отвердителя – обратная корреляция; -0,07

Однако, стоит отметить, что все параметры коррелируют между собой очень слабо. Это также доказывают построенные диаграммы рассеивания.

Диаграммы рассеивания на примере параметров с наибольшей корреляцией

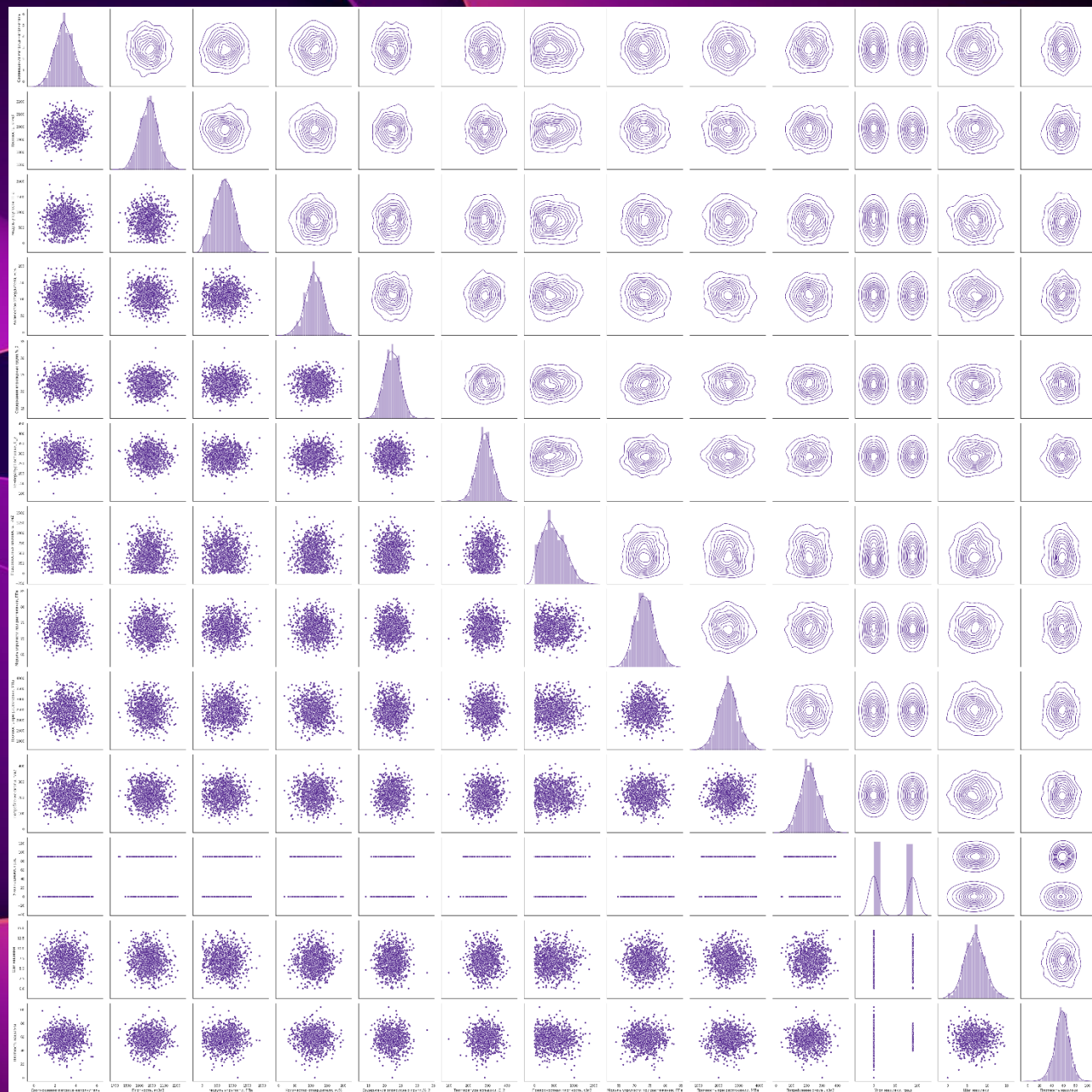


Диаграммы размаха («ящики с усами») с гистограммами распределения:



Попарные диаграммы рассеивания точек:

Чтобы понять взаимосвязь между всеми возможными парами числовых переменных с помощью использования библиотеки `seaborn` был построен попарный график, в верхнем-правом углу которого расположены графики плотности ядра, в нижнем-левом — графики рассеивания, между ними — графики распределения.



Предобработка данных:

В целях очищения датасета от выбросов (аномалий) предпринимались попытки выявления выбросов данными методами: Изолирующий лес, Метод трех сигм и Метод IQR (межквартильный размах)

Остановиться я решил на методе трёх сигм, т.к. он удаляет меньше выбросов, а именно - 24 строки (2,3% от общего количества наблюдений). Это решение связано с тем, чтобы сохранить большее количество данных для обработки, так как их распределение не говорит о присутствии явных аномалий. После очистки данных от выбросов количество наблюдений составило 999 строк. Таким образом, можно сделать вывод, что исключение выбросов не оказало существенного влияния на размер выборки.

Будучи разными по физическому смыслу, данные сильно различаются между собой по абсолютным величинам. Работа аналитических моделей машинного обучения с такими показателями окажется некорректной: дисбаланс между значениями признаков может вызвать неустойчивость работы модели, ухудшить результаты обучения и замедлить процесс моделирования.

После нормализации все числовые значения входных признаков будут приведены к одинаковой области их изменения – некоторому узкому диапазону. Это позволит свести их вместе в одной модели и обеспечит корректную работу вычислительных алгоритмов. Для дальнейшей работы с данными, сравнения их между собой и составления модели машинного обучения была проведена нормализация данных методом «MinMaxScaler()». Этот метод нормализации включает масштабирование набора данных до диапазона [0, 1].

Описательная статистика после нормализации данных методом «MinMaxScaler()»:

	count	mean	std	min	25%	50%	75%	max
Соотношение матрица-наполнитель	999.0	0.489727	0.174701	0.0	0.371306	0.484288	0.608487	1.0
Плотность, кг/м3	999.0	0.467798	0.178722	0.0	0.341020	0.472391	0.579760	1.0
модуль упругости, ГПа	999.0	0.446887	0.198929	0.0	0.302135	0.448458	0.581067	1.0
Количество отвердителя, м.%	999.0	0.496747	0.170875	0.0	0.384427	0.495616	0.613450	1.0
Содержание эпоксидных групп,%_2	999.0	0.493097	0.179869	0.0	0.368588	0.492051	0.624540	1.0
Температура вспышки, С_2	999.0	0.488685	0.174877	0.0	0.371822	0.488391	0.606296	1.0
Поверхностная плотность, г/м2	999.0	0.371058	0.215125	0.0	0.206249	0.348503	0.534748	1.0
Модуль упругости при растяжении, ГПа	999.0	0.501023	0.167891	0.0	0.389296	0.496176	0.610020	1.0
Прочность при растяжении, МПа	999.0	0.508273	0.172193	0.0	0.390683	0.504890	0.613078	1.0
Потребление смолы, г/м2	999.0	0.512182	0.170414	0.0	0.401086	0.512933	0.625356	1.0
Угол нашивки, град	999.0	0.496496	0.500238	0.0	0.000000	0.000000	1.000000	1.0
Шаг нашивки	999.0	0.477203	0.177675	0.0	0.351355	0.478419	0.593879	1.0
Плотность нашивки	999.0	0.507132	0.163683	0.0	0.405778	0.510118	0.612960	1.0

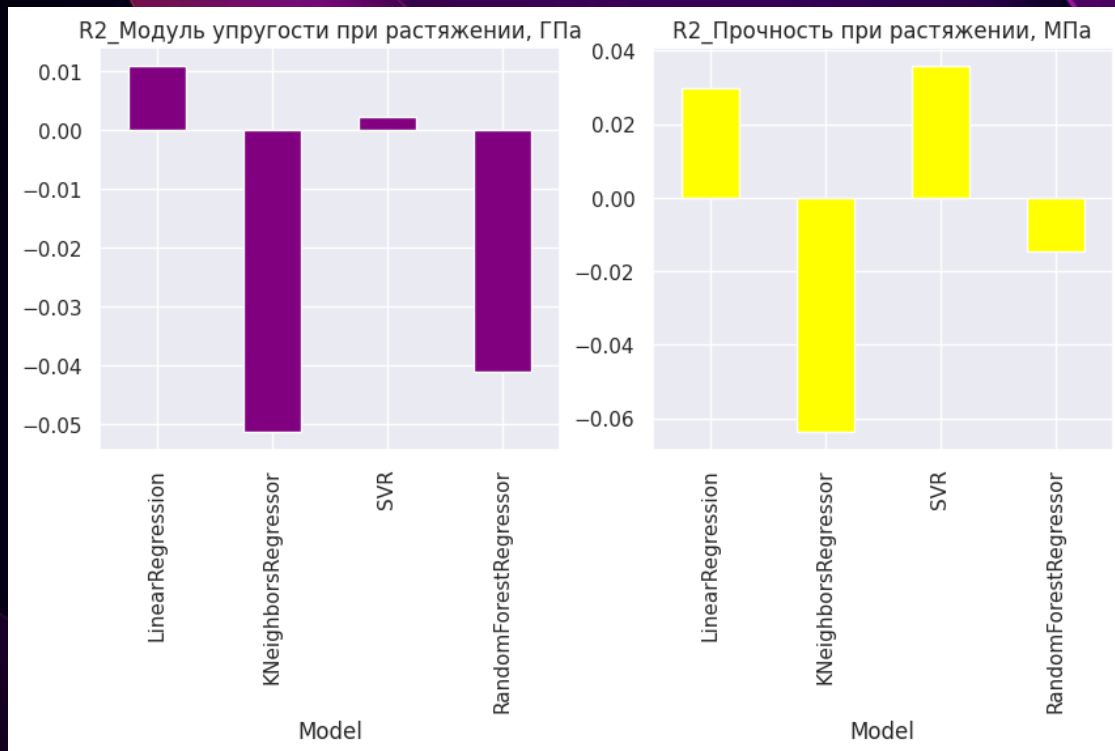
Таким образом, после предобработки данных получаем нормализованный датасет размером 999 строк, данные в котором приближены к нормальному распределению, за исключением угла нашивки (дискретная величина) и поверхностной плотности (Пуассоновское распределение).

Для прогноза модуля упругости при растяжении и прочности при растяжении были использованы следующие методы решения задачи множественной регрессии с помощью Python:

- 1) *Линейная регрессия* – метод «LinearRegression»;
- 2) *Метод k-ближайших соседей* – метод «KneighborsRegressor»;
- 3) *Метод опорных векторов с линейным ядром* – метод «SVR»;
- 4) *Лес принятия решений (случайный лес)* – метод «RandomForestRegressor».

На тестирование модели оставили 30 процентов данных, а на остальных происходило обучение моделей. Далее для каждого из параметров (модуль упругости при растяжении и прочность при растяжении) были построены модели регрессии.

Результаты оценки качества моделей для решения задач множественной регрессии



Оценка качества моделей проведена с помощью расчёта коэффициентов детерминации, которые показывает долю вариации результативного признака под влиянием факторного признака. При отсутствии связи эмпирический коэффициент детерминации равен нулю, а при функциональной связи — единице.

MSE: 1.47; MAE: 1.18 - для модели на основе линейной регрессии

Из графиков видно, что ни одна из моделей не справилась с задачей. При этом хуже всего показывает себя метод ближайших соседей. Метод линейной регрессии и метод опорных векторов дают прогнозы, приближённые к простому усреднению.

Для прогнозирования соотношения матрица-наполнитель написаны модели с использованием однослойного и многослойного персептрона

Все модели обучались на 100 эпохах.

Размер тестовой выборки составил 30 процентов.

При построении линейной модели зависимость соотношения матрица-наполнитель была установлена от потребления смолы.

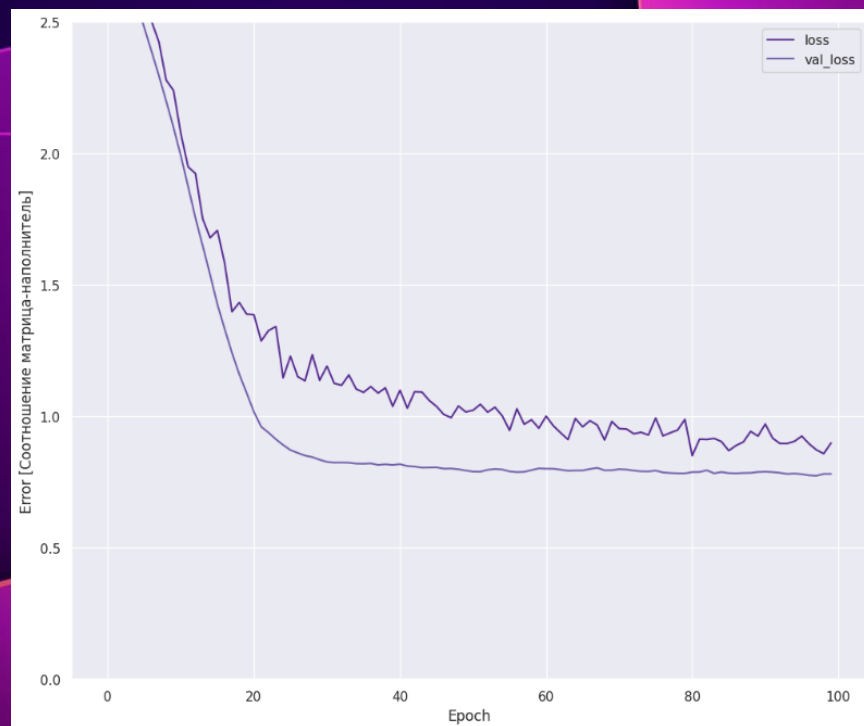
График предсказанных значений соотношения матрица-наполнитель в модели (dnn_potrsmoly_model):



При построении многослойного персептрона после слоя нормализации добавлены слои:

- слой с 32 нейронами и активационной функцией «tanh»,
- слой с пакетной нормализацией (BatchNormalization),
- слой с методом «прореживания» (Dropout),
- выходной слой.

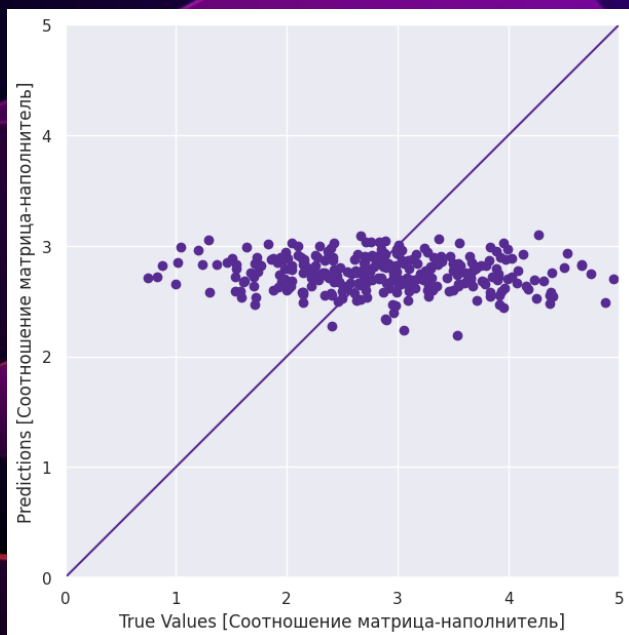
График изменения ошибки тестовой и валидационной выборки в процессе обучения модели, которая учитывает все имеющиеся признаки (dnn_model):



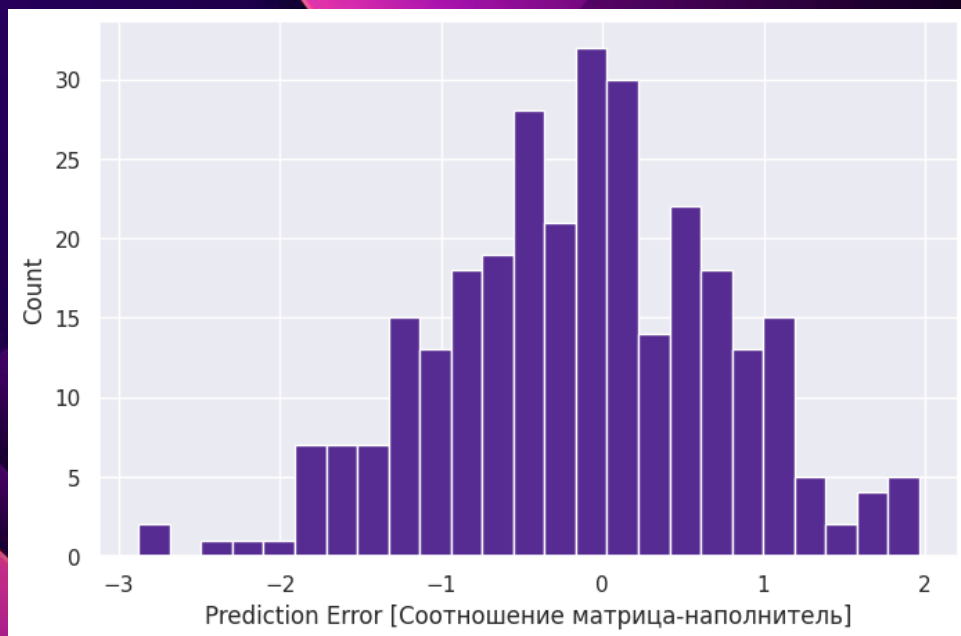
Результаты оценки качества моделей для прогноза соотношения матрица-наполнитель с помощью метрики MAE:


Mean absolute error [Соотношение матрица-наполнитель]	
potrsmoly_model	0.700175
linear_model	0.691898
dnn_potrsmoly_model	0.678404
dnn_model	0.710185

График истинных и предсказанных значений соотношения матрица-наполнитель:



Гистограмма распределения ошибки:





Благодарю за
уделенное время!