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»

Слушатель Литенец Александр Юрьевич

Москва, 2023

Содержание

Введение

Многокомпонентные материалы, состоящие из пластичной основы (матрицы), армированной наполнителями, обладающими высокой прочностью, жесткостью и т.д. Такие материалы называют композитными материалами или композитами.

Свойства нового материала количественно и качественно отличаются от свойств каждого из его составляющих. Многие композиты превосходят традиционные материалы и сплавы по своим механическим свойствам и в то же время они легче. Использование композитов обычно позволяет уменьшить массу конструкции при сохранении или улучшении ее механических характеристик.

Композитные конструкции позволяют значительно сократить расходы на строительство, содержание и ремонт и, одновременно, увеличить срок службы и межремонтные сроки.

Для ускорения создания и внедрения новых композитных материалов в промышленное применение, а также для удешевления себестоимости конечной продукции и как следствие повышение конкурентоспособности товаров крайне важно быстро и точно прогнозировать возможные результаты испытаний, а также подбирать состав для новых образцов. С данной задачей может помочь справиться анализ свойств и характеристик исследуемых материалов, а также их последующее прогнозирование.

Методология, используемая в данном курсе, позволяет решить данные задачи и предоставить промышленности инструменты для решения прикладных задач.

В процессе исследовательской работы были разработаны несколько моделей, способные с высокой вероятностью прогнозировать модули упругости при растяжении и прочности при растяжении, а также были созданы 2 нейронных сети, которые предлагают соотношение «матрицы - наполнитель».

1. Аналитическая часть

1.1 Постановка задачи

Целью данного работы является разработать модели для прогноза модуля упругости при растяжении, прочности при растяжении, а также соотношения «матрица-наполнитель».

Для проведения Аналитического исследования были предоставлены две excel-таблицы, содержащие информацию об используемом композитном материале. X_bp.xlsx с данными о параметрах базальтопластика, состоящий из 1024 строки и 11 столбцов (первый из которых содержит индексы),а также X_nup.xlsx, содержащий данные о нашивках углепластика, состоящий из 1041 строки и 4 столбцов(первый столбец также содержит индексы).

После объединения данных таблиц в единый датафрейм (объединение проводилось по индексу тип объединения INNER) была создана таблица, содержащая 1023 строки и 13 столбцов. Часть информации (17 строк таблицы способов компоновки композитов) не имеют соответствующих строк в таблице соотношений и свойств используемых компонентов композитов, поэтому были удалены.

Начальные свойства компонентов композиционных материалов разделены по следующим критериям:

- Соотношение матрица-наполнитель;
- Плотность;
- Модуль упругости;
- Количество отвердителя;
- Содержание эпоксидных групп;
- Температура вспышки;

– Поверхностная плотность;
– Модуль упругости при растяжении;
– Прочность при растяжении;
– Потребление смолы;
– Угол нашивки;
– Шаг нашивки;
– Плотность нашивки.

Общее количество параметров для анализа – 13.

При разведочном анализе данных в данном датасете не было выявлено пропусков. Также учитывая специфику данных, было принято решение считать, что в случае если в данных из первой таблицы будут встречены значения равные нулю, то их также следует считать пропуском. Для каждой колонки были получены среднее, медианное значение и другие параметры описательной статистики с помощью использования функции df.describe().T

	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

Рис.1 Пример вывода описательной статистики describe

При дальнейшем анализе предоставленных материалов были составлены диаграммы «ящик с усами», так как они позволяют визуально оценить возможные выбросы в каждом из столбцов, объединённого датасета.

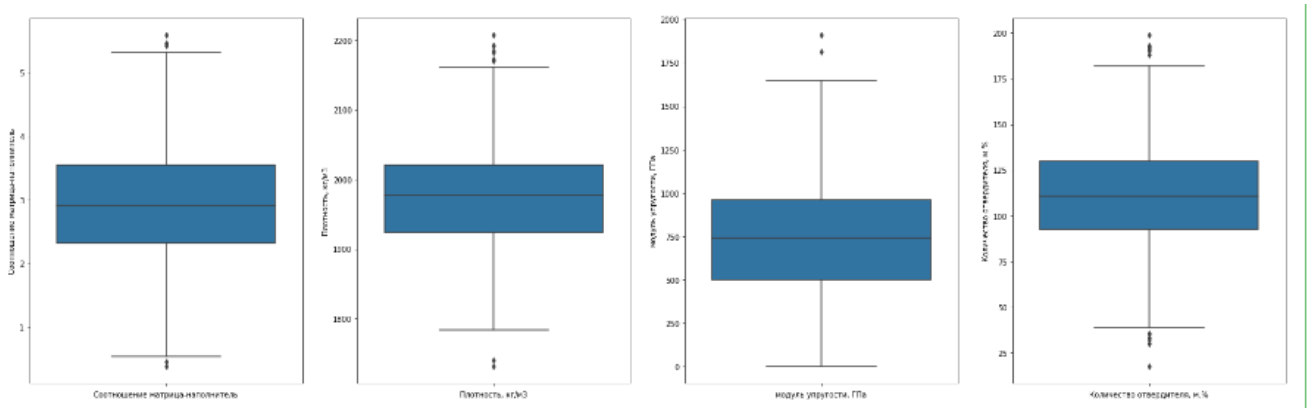


Рис 2. Диаграммы «Ящик с усами»

Для устранения выбросов в данных был использован метод «Трёх сигм», после применения которого в датасете стало насчитываться 996 строк при первом запуске метода и 996 при последующих. При этом 27 наблюдения (2,3% от общего количества наблюдений), которые содержали эти выбросы, были исключены из датасета.

После очистки данных от выбросов количество наблюдений составило 996 строк. Таким образом, можно сделать вывод, что исключение выбросов не оказало существенного влияния на размер выборки.

Для решения поставленной задачи необходимо обучить несколько моделей для прогноза модуля упругости при растяжении и прочности при растяжении. Написать нейронную сеть, которая будет рекомендовать соотношение матрица-наполнитель. Разработать приложение с графическим интерфейсом, которое будет выдавать прогноз соотношения «матрица-наполнитель».

1.2 Описание используемых методов

В ходе данного исследования решается задача о предсказания какой-ли численной характеристики объекта предметной области по определенному набору его параметров (атрибутов). Данная задача в рамках классификации категорий машинного обучения относится к классу задач регрессии. Цель любого алгоритма обучения с учителем — определить функцию потерь и минимизировать её. В решения поставленной задачи были были применены следующие методы:

- случайный лес;
- градиентный бустинг;
- многослойный перцептрон;
- Лассо;

Рассмотрим подробнее каждый из них.

В случайных лесах (RandomForestRegressor класс) каждое дерево в ансамбле строится из выборки, взятой с заменой (то есть выборкой начальной загрузки) из обучающего набора.

Кроме того, при разбиении каждого узла во время построения дерева наилучшее разбиение находится либо по всем входным характеристикам, либо по случайному подмножеству размера max_features.

Назначение этих двух источников случайности — уменьшить дисперсию оценки леса. В самом деле, отдельные деревья решений обычно демонстрируют высокую дисперсию и имеют тенденцию переоснащаться. Внедренная случайность в лесах дает деревья решений с несколько несвязанными ошибками прогнозирования. Если взять среднее значение этих прогнозов, некоторые ошибки могут быть устранены. Случайные леса уменьшают дисперсию за счет комбинирования разных деревьев, иногда за счет небольшого увеличения

смещения. На практике уменьшение дисперсии часто бывает значительным, что дает в целом лучшую модель.

В отличие от исходной публикации от других вариантов , реализация scikit-learn, используемая в данной работе, объединяет классификаторы путем усреднения их вероятностного прогноза вместо того, чтобы позволить каждому классификатору голосовать за один класс.

Регрессия нейронной сети. Несмотря на то, что нейронные сети широко используются для углубленного обучения и моделирования сложных задач, таких как распознавание изображений, они легко адаптируются к задачам регрессии. Любой класс статистических моделей можно назвать нейронной сетью, если эти модели используют адаптивные весовые коэффициенты и могут использоваться для аппроксимации нелинейных функций входных данных. Таким образом, регрессия нейронной сети подходит для задач, которые нельзя решить с помощью более традиционных моделей.

Нейронная сеть выдаст прогнозируемое значение переменной, зависимое от множества входных параметров.

Перед тем, как производить прогноз, алгоритм обучается на тренировочном наборе данных — обучающей выборке. Каждая строка такой выборки содержит:

- в полях, обозначенных как входные — множество входных параметров;
- в поле, обозначенном как выходное — соответствующее входным параметрам значение зависимой переменной.

Технически обучение заключается в нахождении весов — коэффициентов связей между нейронами. В процессе обучения нейронная сеть способна выявлять сложные зависимости между входными параметрами и выходными, а также выполнять обобщение. Это значит, что в случае успешного обучения

нейронная сеть способна выдать верный результат на основании данных, которые отсутствовали в обучающей выборке, а также на неполных данных.

Лассо — это линейная модель, которая оценивает разреженные коэффициенты. Это полезно в некоторых контекстах из-за своей тенденции отдавать предпочтение решениям с меньшим количеством ненулевых коэффициентов, эффективно уменьшая количество функций, от которых зависит данное решение. По этой причине лассо и его варианты являются фундаментальными для области сжатого зондирования. При определенных условиях он может восстановить точный набор ненулевых коэффициентов «зашумленных», частично искажённых данных.

Математически лассо состоит из линейной модели с добавленным членом регуляризации.

Lasso использует координатный спуск в качестве алгоритма подбора коэффициентов. В качестве регрессии LASSO позволяет выявлять редкие модели.

KNeighborsRegressor предоставляет функциональные возможности для неконтролируемых и контролируемых методов обучения на основе соседей. Неконтролируемые ближайшие соседи — это основа многих других методов обучения, в частности множественного обучения и спектральной кластеризации. Обучение на основе контролируемых соседей бывает двух видов: классификация данных с дискретными метками и регрессия для данных с непрерывными метками.

Принцип, лежащий в основе методов ближайшего соседа, состоит в том, чтобы найти предопределенное количество обучающих выборок, ближайших по расстоянию к новой точке, и предсказать метку по ним. Количество выборок может быть заданной пользователем константой (обучение k-ближайшего соседа) или изменяться в зависимости от локальной плотности точек (обучение соседей на основе радиуса). Расстояние, как правило, может быть любой

метрической мерой: стандартное евклидово расстояние является наиболее распространенным выбором. Соседи на основе методов известны как не-обобщающего машины методы обучения, так как они просто «вспомнить» все его подготовки данных (возможно , превращается в быструю индексной структуры

Несмотря на свою простоту, функция «Ближайшие соседи» успешно справляется с большим количеством задач классификации и регрессии, включая рукописные цифры и сцены спутниковых изображений. Будучи непараметрическим методом, он часто бывает успешным в ситуациях классификации, когда граница решения очень нерегулярна.

1.3 Разведочный анализ данных

Для анализа взаимосвязей между столбцами таблицы была составлена тепловая карта:

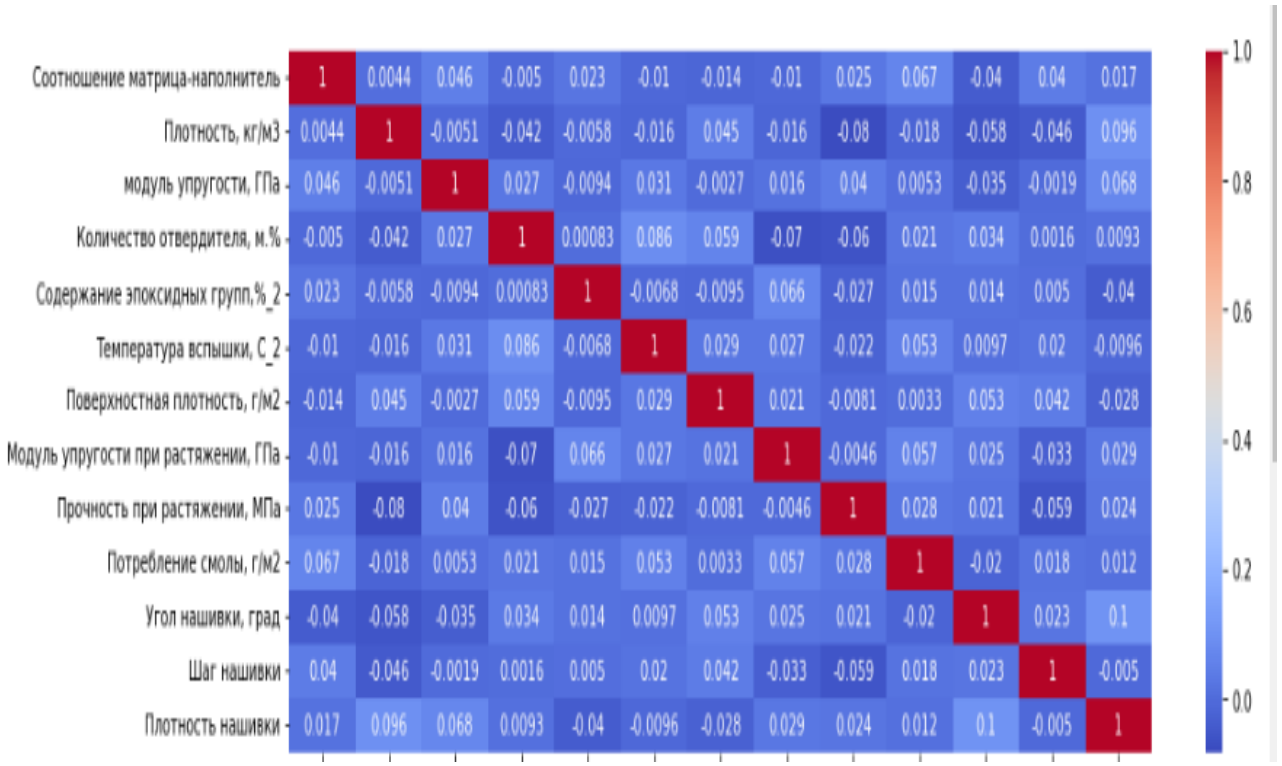


Рис. 3 Тепловая карта корреляции

Таким образом, лучше всего коррелируют между собой (по убыванию):

1) угол нашивки и плотность нашивки (коэффициент корреляции равен 0,11);

2) температура вспышки и количество отвердителя (коэффициент корреляции равен 0,10);

3) плотность и плотность нашивки (коэффициент корреляции равен 0,08);

4) прочность при растяжении и количество отвердителя (коэффициент корреляции равен -0,08);

5) потребление смолы и соотношение матрица-наполнитель (коэффициент корреляции равен 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.

При составлении тепловой карты корреляции между столбцами датасета было выяснено, что линейные зависимости между столбцами крайне низки (максимальное значение составило 0,11), что однако не исключает возможность наличия нелинейных зависимостей.

Для более подробного изучения данных были составлены диаграммы распределения, на основании которых было сформулировано предположение о том, что в датасете присутствуют смешанные данные двух различных

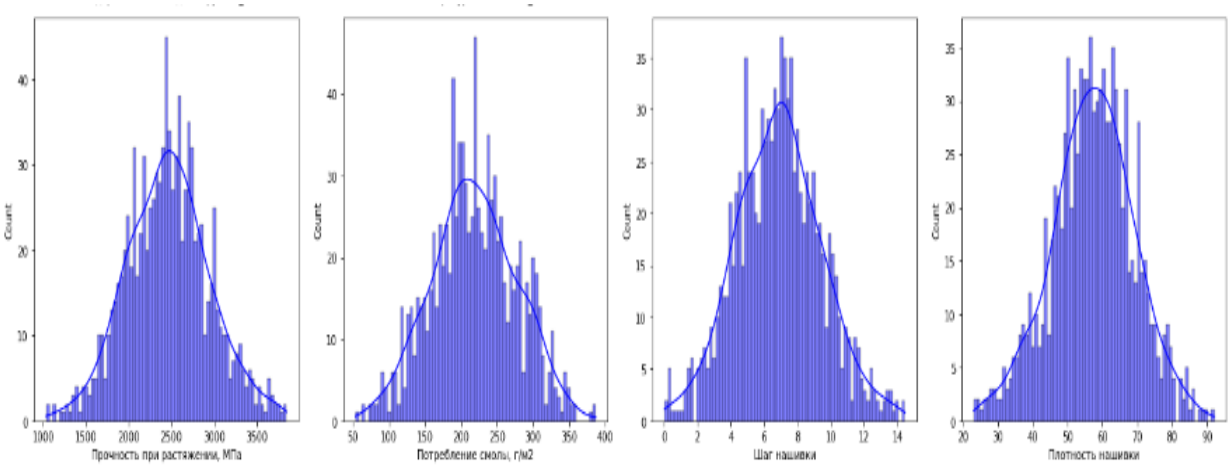


Рис. 5 диаграммы распределения

композитных материалов. (на некоторых графиках можно заметить бимодальность распределения).

Дополнительно был построен график плотности распределения значений

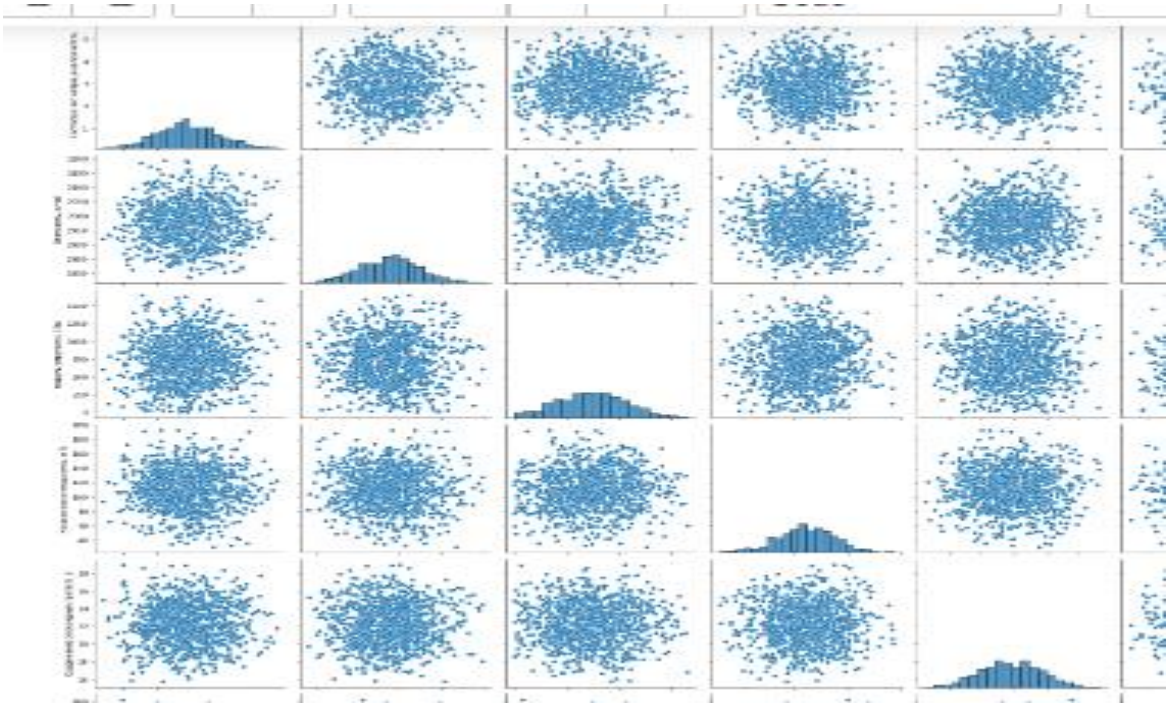


Рис.6 График плотности распределения

Для подтверждения или опровержения гипотезы бимодальности был построен график «локтя» на основании метода кластеризации k means. Данный график подтвердил предположение о бимодальной природе этого набора данных и при дальнейшей работе было принято решение провести отдельное изучения как датафрейма целиком так и двух его отдельных кластеров. Алгоритм k-средних в ходе своей работы делит набор образцы переданных ему данных на К непересекающиеся кластер, каждый из которых описывается средним образцов в кластере. Число кластеров необходимо задать предварительно и требует отдельного метода подбора.

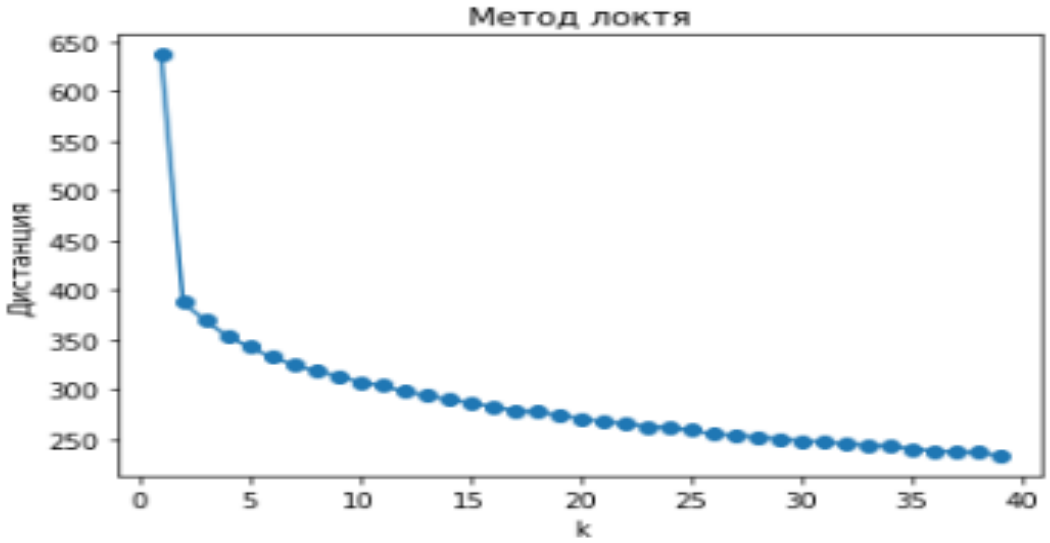


Рис.7 График локтя

В связи с тем, что разброс минимальных и максимальных значений между столбцами достаточно велик для улучшения работы прогнозных моделей была проведена нормализация данных в диапазоне от 0 до 1. Для этого использовался метод MinMaxScaler(). Данная функция масштабирует данные таким образом,

чтобы они находились между заданным минимальным и максимальным значением, часто между нулем и единицей.

Рис. 8 Описательная статистика после нормализации

2.Практическая часть

	count	mean	std	min	25%	50%	75%	max
Соотношение матрица-наполнитель	996.0	0.489425	0.174654	0.0	0.370964	0.484284	0.608022	1.0
Плотность, кг/м3	996.0	0.468036	0.178910	0.0	0.341039	0.472967	0.579921	1.0
модуль упругости, ГПа	996.0	0.447416	0.198975	0.0	0.303019	0.448658	0.582408	1.0
Количество отвердителя, м.%	996.0	0.496990	0.171017	0.0	0.384551	0.496653	0.614168	1.0
Содержание эпоксидных групп,%_2	996.0	0.492885	0.180083	0.0	0.368458	0.491095	0.624852	1.0
Температура вспышки, С_2	996.0	0.488576	0.175082	0.0	0.371321	0.488205	0.606379	1.0
Поверхностная плотность, г/м2	996.0	0.370791	0.215318	0.0	0.205775	0.348438	0.534571	1.0
Модуль упругости при растяжении, ГПа	996.0	0.500751	0.167970	0.0	0.388664	0.495558	0.609604	1.0
Прочность при растяжении, МПа	996.0	0.508025	0.172210	0.0	0.390414	0.504890	0.612932	1.0
Потребление смолы, г/м2	996.0	0.494378	0.176148	0.0	0.379054	0.495415	0.611101	1.0
Угол нашивки, град	996.0	0.497992	0.500247	0.0	0.000000	0.000000	1.000000	1.0
Шаг нашивки	996.0	0.476411	0.177319	0.0	0.350266	0.477391	0.593526	1.0
Плотность нашивки	996.0	0.495708	0.170531	0.0	0.390120	0.498893	0.606506	1.0
Кластер	996.0	0.519076	0.499887	0.0	0.000000	1.000000	1.000000	1.0

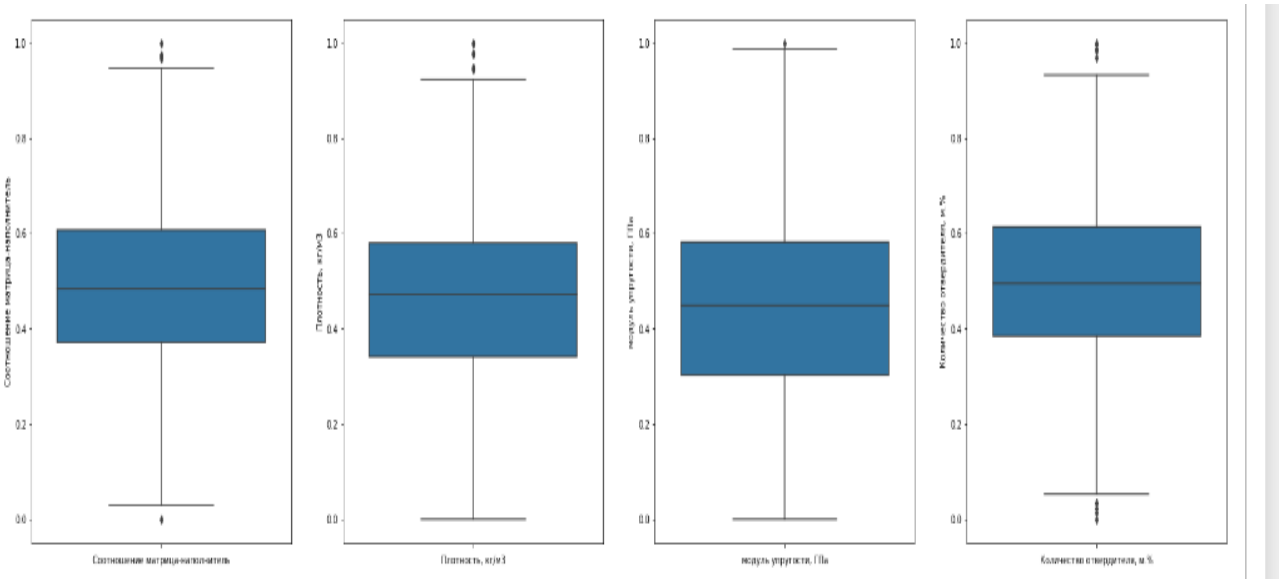
2.1 Предобработка данных

Предоставленные для решения задачи данные были предобработаны.

В целях очищения датасета от выбросов (аномалий) был выбран метод исключения выбросов с помощью правила трёх сигм. Анализ выбросов в данных методом позволяет определить аномальные значения в нестационарных рядах с распределением близким к нормальному. Основу данного метода анализа составляет расчет среднего значения ряда и среднеквадратичного отклонения.

После удаления выбросов , а также нормализации повторно выведем график «ящик с усами» дабы оценить изменения произошедшие с датасетом.

Рис 9. Ящики с усами после нормализации



Для изучения гипотезы бимодально посредством использования метода k-means в нормализованном датасете были выявлены два кластера 0-й и 1-й. Предположительно в них содержатся разные виды композитов. В нулевом кластере содержится 479 строк.

	Соотношение матрица-наполнитель	Плотность, кг/м3	модуль упругости, ГПа	Количество отвердителя, м.%	Содержание эпоксидных групп,%_2	Температура вспышки, С_2	Поверхностная плотность, г/м2	модуль упругости при растяжении, ГПа	Прочность при растяжении, МПа	Потребление смолы, г/м2	Угол нашивки, град	н
4	0.419084	0.307448	0.488508	0.502800	0.495653	0.482823	0.162230	0.321894	0.698235	0.499322	0.0	(
5	0.417519	0.282954	0.323358	0.502800	0.495653	0.482823	0.293938	0.592578	0.271429	0.199341	0.0	(
6	0.608883	0.356437	0.538297	0.608021	0.418887	0.549664	0.293938	0.592578	0.271429	0.199341	0.0	(
9	0.478238	0.503403	0.986997	0.608021	0.418887	0.549664	0.782031	0.754989	0.342563	0.739306	0.0	(
10	0.232351	0.405426	0.500652	0.608021	0.418887	0.549664	0.363665	0.502350	0.504591	0.499322	0.0	(
...	...	...	...	...	...	...	...	...	...	...	...	
988	0.464630	0.216487	0.603275	0.713605	0.442213	0.610057	0.619050	0.474319	0.500297	0.761100	1.0	(
989	0.645183	0.318788	0.411813	0.497392	0.771294	0.592226	0.486359	0.648369	0.508635	0.295889	1.0	(
991	0.361750	0.410540	0.552781	0.350139	0.333908	0.657301	0.161609	0.489229	0.480312	0.214363	1.0	(
993	0.555750	0.460227	0.251612	0.494656	0.623085	0.325580	0.572959	0.578197	0.578340	0.549139	1.0	(
995	0.657131	0.259472	0.251903	0.609147	0.888354	0.553803	0.587373	0.555208	0.647135	0.423589	1.0	(

479 rows × 13 columns

Рис. 10 Нулевой кластер

В первом кластере содержится 517 строк.

	Соотношение матрица-наполнитель	Плотность, кг/м3	модуль упругости, ГПа	Количество отвердителя, м.%	Содержание эпоксидных групп,%_2	Температура вспышки, С_2	Поверхностная плотность, г/м2	Модуль упругости при растяжении, ГПа	Прочность при растяжении, МПа	Потребление смолы, г/м2	Угол нашивки, град	н
0	0.282131	0.601381	0.447061	0.123047	0.607435	0.482823	0.162230	0.321894	0.698235	0.499322	0.0	(
1	0.282131	0.601381	0.447061	0.608021	0.418887	0.549664	0.162230	0.321894	0.698235	0.499322	0.0	(
2	0.457857	0.601381	0.455721	0.502800	0.495653	0.482823	0.162230	0.321894	0.698235	0.499322	0.0	(
3	0.457201	0.527898	0.452685	0.502800	0.495653	0.482823	0.162230	0.321894	0.698235	0.499322	0.0	(
7	0.604139	0.772842	0.861312	0.608021	0.418887	0.549664	0.782031	0.754989	0.342563	0.739306	0.0	(
...	...	...	...	...	...	...	...	...	...	...	...	
983	0.272501	0.683367	0.616274	0.869147	0.585397	0.451202	0.016390	0.577357	0.317909	0.649241	1.0	(
987	0.241590	0.564082	0.509192	0.448286	0.405565	0.425802	0.476135	0.817488	0.526985	0.536604	1.0	(
990	0.455435	0.529138	0.565962	0.694099	0.428655	0.436063	0.049973	0.193793	0.734705	0.528661	1.0	(
992	0.587163	0.650588	0.268550	0.712271	0.294428	0.350746	0.271207	0.480019	0.470745	0.192531	1.0	(
994	0.637396	0.691520	0.448724	0.684130	0.267818	0.444436	0.496511	0.540754	0.368070	0.430704	1.0	(

517 rows × 13 columns

Рис. 11. Первый кластер

Для того чтобы оценить важность того или иного столбца для прогноза интересующих значений было решено оценить вклад каждого отдельного параметра в прогноз методом Случайного леса. Данный эксперимент показал что наиболее значимыми с точки зрения данного метода являются:

- Плотность, кг/м3
- Количество отвердителя, м.%
- Содержание эпоксидных групп,%_2
- Температура вспышки, С_2
- Потребление смолы, г/м2
- Плотность нашивки

	4	6	3	8	10	2	7	0	9	5	1
feature	Количество отвердителя, м.%	Температура вспышки, С_2	Плотность, кг/м3	Потребление смолы, г/м2	Плотность нашивки	модуль упругости, ГПа	Поверхностная плотность, г/м2	Соотношение матрица-наполнитель	Шаг нашивки	Содержание эпоксидных групп,%_2	Угол нашивки, град
importance	0.114191	0.113564	0.111418	0.106192	0.101097	0.090164	0.08808	0.087701	0.087648	0.086944	0.013002

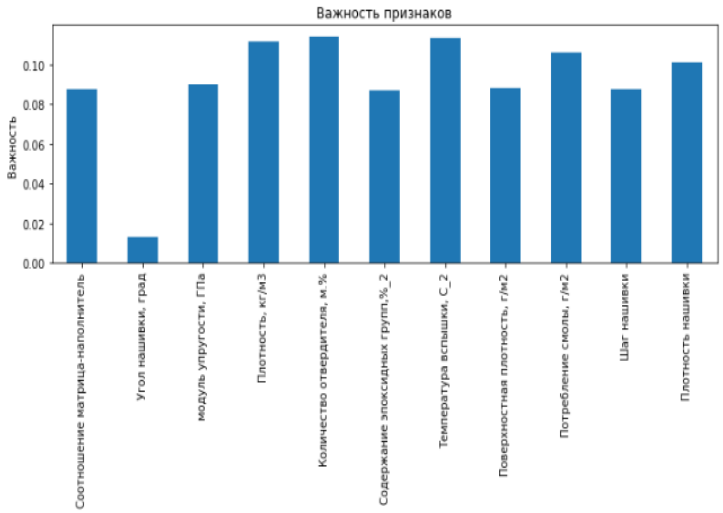


Рис. 12 Значимость параметров

Для проверки гипотезы о возможности использовании только 6 наиболее значимых параметров также отдельно просчитаем данные из всего датафрейма, нулевого и первого кластеров без учета незначимых параметров.

2.2 Разработка и обучение модели

При прогнозе модуля упругости при растяжении и прочности при растяжении были использованы следующие следующие датасеты были разбиты на тестовую и тренировочную выборку (70% на train и 30% test) :

- Объединённый нормализованный датасет(x_norm_train, x_norm_test, y_norm_train, y_norm_test)
- 0-й нормализованный кластер(x_0_train, x_0_test, y_0_train, y_0_test)
- 1-й нормализованный кластер(x_1_train, x_1_test, y_1_train, y_1_test)
- Шесть «главных» колонок из объединённого нормализованного датасета(x_6_train, x_6_test, y_6_train, y_6_test)
- Шесть «главных» колонок из 0-го кластера(x_06_train, x_06_test, y_06_train, y_06_test)
- Шесть «главных» колонок из 1-го кластера(x_16_train, x_16_test, y_16_train, y_16_test)

Для удобства проведения расчетов были созданы словарь, хранящий тестовые и тренировочные значения выборок xy_list,а также таблица в которую будут записываться результаты испытаний для каждого метода results. Для каждого из приведенных далее методов будет написан отдельная функция, упрощающая работу с данными.

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

- случайный лес;
- KNeighborsRegressor;
- многослойный перцептрон;
- Лассо;

Рассмотрим применение RandomForestRegressor для разных наборов данных. Для данной модели были использованы следующие настройки

random_state=42 – зерно для фиксации результатов

n_estimators=50 – число деревьев в лесу. Входе экспериментов было выяснено что дальнейшее увеличение числа деревьев либо приносит крайне незначительные результаты или вызывает переобучение леса в целом.

criterion='squared_error' - Функция для измерения качества разделения. Также была проверена результативность: “absolute_error”, “friedman_mse”, “poisson”, но эффективность уступала 'squared_error'.

max_depth=4 - максимальная глубина дерева. Данное значение оказалось оптимальным в ходе ряда тестов, так как дальнейшее увеличение привело к переобучению модели, а уменьшение к снижению достигаемых результатов.

Рассмотрим работу Lasso с следующими гиперпараметрами:

alpha=1.0 - Константа, которая контролируя силу регулярицции

max_iter - 1000 Максимальное количество итераций.

tol=0.0001 - Допуск для оптимизации

random_state=42 - Начальное значение генератора псевдослучайных чисел.

При применении KNeighborsRegressor использовались следующие гиперпараметры:

n_neighbors=25 - Количество соседей, которые будут использоваться по умолчанию

algorithm='auto' - Алгоритм, используемый для вычисления ближайших соседей

p=2 - Параметр мощности для метрики Минковского

metric='minkowski' - Метрика, используемая для вычисления расстояния.

2.3 Тестирование модели

На каждой из данных таблиц были проведены сходные вычисления для поиска оптимального способа разбиения датасета и вычеслены следующие метрики качества:

- r2 train коэффициент детерминации на обучающей выборке
- r2 test коэффициент детерминации на тестовой выборке
- mse среднеквадратичная ошибка
- mae средняя абсолютная ошибка
- mape Средняя абсолютная ошибка в процентах

Для результативности работы RandomForestRegressor разбивка исходного два датасета привела к увеличению производительности. Для первого кластера

коэффициент детерминации на тестовой выборке составил 0.06 и 0.02 для нулевого , в то время как для исходного датасета он составил -0,016. Вычленение

	train_r2	test_r2	mse	mae	mape
Полный датасет	0.135939	-0.015642	0.028650	0.134170	3.783847e+12
0-й кластер	0.239273	0.020304	0.028387	0.131582	7.879145e+12
1-й кластер	0.221050	0.064647	0.023527	0.122880	3.494982e-01
Полный датасет 6 столбцов	0.121267	0.001264	0.028174	0.132902	3.718998e+12
0-й кластер 6 столбцов	0.222402	-0.009753	0.029255	0.134236	7.869201e+12
1-й клстер 6 столбцов	0.213179	0.060093	0.023636	0.123519	3.462081e-01

и дальнейшая обработка отдельно шести «главных» признаков при работе данным методом значимых результатов не принесла.

Рис. 13 результаты работы Случайного леса

Эксперименты с подбором гиперпараметров для модели Lasso не принесли статистически значимого результата и мало отличались от стандартных настроек данной модели. Также не было выявлено значительной разницы между результатами полных датасетов и датасетов с 6 «главными» столбцами. Разбивка на кластеры дает незначительное улучшение по сравнению в «цельным» датафреймом.

	train_r2	test_r2	mse	mae	mape
Полный датасет	2.220446e-16	-0.009483	0.028476	0.132860	3.810883e+12
0-й кластер	-5.551115e-16	-0.003623	0.029093	0.134513	8.189343e+12
1-й кластер	3.885781e-16	-0.003551	0.025251	0.125529	3.714942e-01
Полный датасет 6 столбцов	2.220446e-16	-0.009483	0.028476	0.132860	3.810883e+12
0-й кластер 6 столбцов	-5.551115e-16	-0.003623	0.029093	0.134513	8.189343e+12
1-й клстер 6 столбцов	3.885781e-16	-0.003551	0.025251	0.125529	3.714942e-01

Рис. 13 результаты работы Lasso

Результаты работы модели KNeighborsRegressor оказались хуже чем работа «средней» модели. При это потребовалось увеличить число соседей до 25, так как при меньшем количестве происходило переобучение модели. Различие в результатах на разных выборках статистически не значимо.

	train_r2	test_r2	mse	mae	mape
Полный датасет	0.068137	-0.036886	0.029237	0.135432	3.534803e+12
0-й кластер	0.054912	-0.016316	0.029444	0.135089	7.204661e+12
1-й кластер	0.040216	-0.027797	0.025863	0.127571	3.773910e-01
Полный датасет 6 столбцов	0.048213	-0.040667	0.029343	0.135491	3.582180e+12
0-й кластер 6 столбцов	0.054864	-0.040367	0.030154	0.137815	7.844449e+12
1-й клстер 6 столбцов	0.065094	-0.025240	0.025799	0.127145	3.725394e-01

Рис. 14 результаты работы KNeighborsRegressor

В результатом работы данных трех моделей решить поставленную задачу о создании инструмента для предсказаний параметров композитного материала не удалось. Все три модели показали крайне низкие результаты. При этом была сформулирована и не отвергнута гипотеза о наличии в исходной выборке двух композитных материалов.

2.4 Написание нейронной сети, которая рекомендует соотношение матрица-наполнитель

Для прогнозирования соотношения матрица-наполнитель написаны модель с использованием многослойного персептрона. В этих целях была задействована библиотека keras.

Функция активации определяет выходное значение нейрона в зависимости от результата взвешенной суммы входов и порогового значения. В качестве такой функции был выбран гиперболический тангенс («relu»).

Для создания данной нейросети был использован входной слой с 20 -ю нейронами и активационной функцией «relu».

Следующим слоем использовался слой с методом «прореживания» Dropout(0.2), который позволяет бороться с переобучением модели за счет выключения из работы части нейронов. Dropout «выключает» нейроны с вероятностью p и, как следствие, оставляет их включенными с вероятностью $q=1-p$.

Далее в персептроне используется скрытый слой с 10 нейронами и активационной функций «relu»

Выходной слой состоит из одного нейрона и активационной функцией «linear».

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 dense (Dense)               (None, 20)                260

 dropout (Dropout)           (None, 20)                0

 dense_1 (Dense)             (None, 10)                210

 dense_2 (Dense)             (None, 1)                 11

=================================================================
Total params: 481
Trainable params: 481
Non-trainable params: 0
```

Рис. 15 Параметры многослойного персептрона на общей выборке

Все модели обучались на 100 эпохах, что является приемлемым для данного датасета. Размер тестовой выборки составил 30 процентов.

Данная нейросеть была применена к нескольким наборам данных:

К полному нормализованному датасету

Рис 16. Динамика обучения персетрона на полном датасете

Также персептрон был применен на урезанных данных из датасетов, разбитых на нулевом кластере:

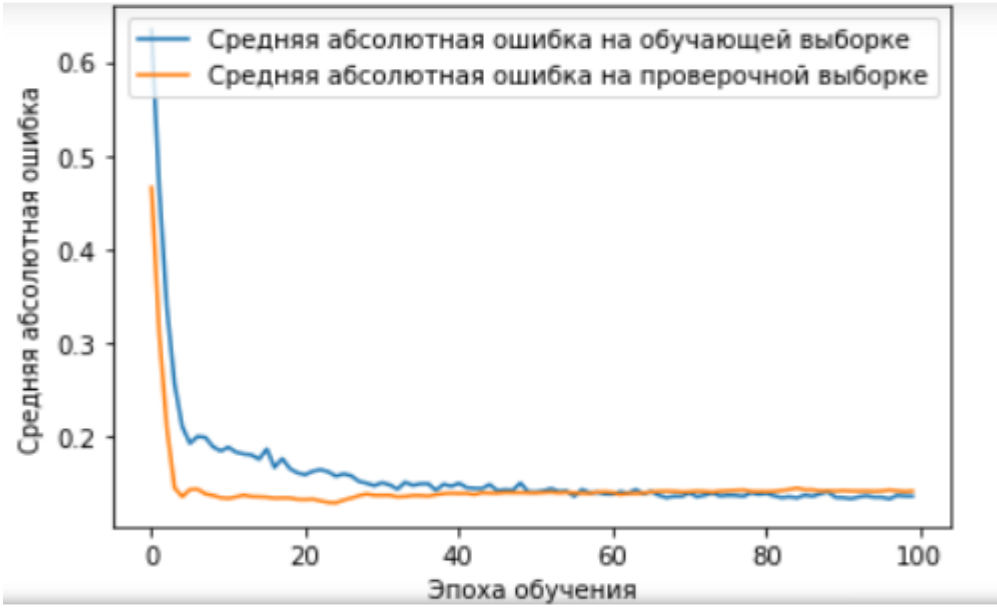


Рис 16. Динамика обучения персетрона на нулевом кластере

```
Model: "sequential_1"
_________________________________________________________________
 Layer (type)                 Output Shape              Param #
=================================================================
 dense_3 (Dense)              (None, 20)                260

 dropout_1 (Dropout)          (None, 20)                0

 dense_4 (Dense)              (None, 10)                210

 dense_5 (Dense)              (None, 1)                 11

=================================================================
Total params: 481
Trainable params: 481
Non-trainable params: 0
_________________________________________________________________
None
```

Рис. 17 Параметры многослойного персептрона на нулевом кластере.

Далее сеть была применена для расчета на первом кластере:

```
_________________________________________________________________
 Layer (type)                 Output Shape              Param #
=================================================================
 dense_6 (Dense)              (None, 20)                260

 dropout_2 (Dropout)          (None, 20)                0

 dense_7 (Dense)              (None, 10)                210

 dense_8 (Dense)              (None, 1)                 11

=================================================================
Total params: 481
Trainable params: 481
Non-trainable params: 0
```

Рис. 18 Параметры многослойного персептрона на первом кластере

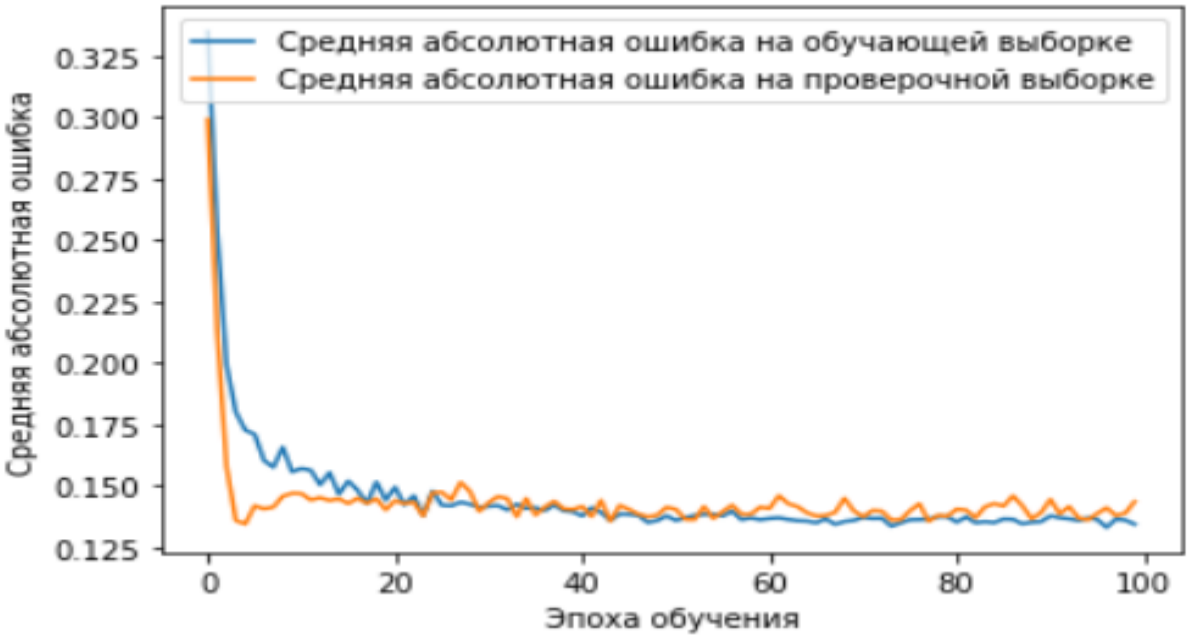


Рис 19. Динамика обучения персептрона на первом кластере

Для данной нейронной сети были построены график, отражающий истинные и предсказанные моделью значения соотношения матрица-наполнитель , что наглядно позволяет оценить точность предсказаний.

Анализируя результаты достигнуты при работе нейросети можно также заявить что заявленная задача о предсказании параметров композитного материала не достигнута. При этом результаты для усеченных на кластеры выборок показали худшие результаты по сравнению с полным датасетом, но отличие незначительно так как достигнутые результаты находится на невысоком уровне. Предположительно для анализа и решения данной задачи необходимо выявить скрытые закономерности в имеющихся данных. При работе нейросети набор данных в кластерах мог быть недостаточным для полноценного обучения нейросетей.

2.5 Разработка приложения

Разработано веб-приложение для рекомендательной системы «Прогноз показателя «Соотношение матрица-наполнитель». Для разработки был использован интерпретатор Python для запуска веб-приложения Flask.

На рисунке 16 изображена веб-страница, на которой запустив приложение, пользователь должен ввести в появившуюся форму 12 входных параметров. После ввода исходных данных в соответствующие окна, пользователь должен нажать кнопку «Отправить». После этого приложение выводит расчётное значение соотношения матрица-наполнитель.

2.6 Создание удаленного репозитория и загрузка результатов работы

Создан репозиторий в GitHub, где размещён код исследования. Оформлен файл README.

Страница слушателя: https://github.com/Ckomap

Созданный репозиторий: https://github.com/Ckomap/ВКР.

Заключение

В ходе решения задачи прогнозирования конечных свойств новых композиционных материалов были изучены основные теоретические основы методов машинного обучения. Проведён анализ предоставленных данных, а также получены прогнозы ряда конечных свойств получаемых композиционных материалов.

При этом в практической части задачи были применены методы машинного обучения на базе библиотек Python. Проведён анализ результатов, полученных с помощью созданных моделей, для выбора наиболее точной из них.

Полученные в ходе решения задачи модели не приносят положительного результата, так как имеют высокий уровень ошибки в предсказаниях. Однако, при проведения данного исследования получен опыт выбора наиболее подходящих методов для решения задач регрессии, опыт настройки моделей, выбора гиперпараметров, а также оценки качества моделей по разным метрикам.

Таким образом, можно сделать вывод о том, что данная работа позволила выработать навыки решения задач с помощью методов машинного обучения, определить направление для дальнейшего постижения науки о данных (Data science).

Библиографический список

1 ГОСТ 32794-2014 Композиты полимерные. - Введ. 2015-09-01. - М.: Стандартинформ, 2015

2 ГОСТ Р 57970-2017 Композиты углеродные. Углеродные композиты, армированные углеродным волокном. – Введ. 2018-06-01. - М.: Стандартинформ, 2018

3 СП 28.13330.2012. «Защита строительных конструкций от коррозии». Актуализированная редакция СНиП 2.03.11-85. Дата введения 2013.01.01.

4 Библиотека Keras - инструмент глубокого обучения. Реализация нейронных сетей с помощью библиотек Theano и TensorFlow / пер. с англ. Слинкин А. А. - М.: ДМК Пресс, 2018. - 294 с.

5 Аллен Б. Дауни – Основы Python. Научитесь думать как программист / Аллен Б. Дауни ; пер. с англ. С. Черникова ; [науч. ред. А. Родионов]. — Москва: Манн, Иванов и Фербер, 2021. — 304 с.

6 Avdeeva A., Shlykova I., Perez M., Antonova M., Belyaeva S.Chemical properties of reinforcing fiberglass in aggressive media. MATEC Web of Conferences. 2016. Vol. 53. 01004.

7 Астапов Р.Л., Мухамадеева Р.М. Автоматизация подбора параметров машинного обучения и обучение модели машинного обучения // Актуальные научные исследования в современном мире. 2021. № 5-2 (73). С. 34-37.

8 Barabanshchikov Y., Belyaeva S., Avdeeva A.and Perez M. Fiberglass Reinforcement for Concrete (2015) Applied Mechanchanics and Materials, Pp. 475-481

9 Вичугова A.Data Preparation: полет нормальный – что такое нормализация данных и зачем она нужна [Электронный ресурс]: – Режим доступа: https://www.bigdataschool.ru/blog/нормализация-feature-transformation-data-preparation.html (дата обращения: 13.06.2022).

10 Билл Любанович. Простой Python. Современный стиль программирования. — СПб.: Питер, 2016. — 480 с.: ил. — (Серия «Бестселлеры O’Reilly»).

11 Бринк Х. Машинное обучение. / Х. Бринк, Дж. Ричардс, М. Феверолф. – СПб.: Питер,2017. 336 с.

12 Вандер Плас Дж. Python для сложных задач: наука о данных и машинное обучение. - СПб.: Питер, 2018. - 576 с.

13 Гиздатуллин А.Р., Хозин В.Г., Куклин А.Н., Хуснутдинов А.М. «Особенности испытаний и характер разрушения полимеркомпозитной арматуры».

14 Горбунов П.М., Мацкевич Ю.А., Чубарь А.В. Машинное обучение. Автоматизация подбора модели машинного обучения // Робототехника и искусственный интеллект. 2021. С. 155-160.

15 Джалилов Ш.А. Метод расчета параметров множественной линейной регрессии // Достижения науки и образования. 2020. № 3 (57). С. 24-28.

16 Джулли, Пал: Библиотека Keras - инструмент глубокого обучения / пер. с англ. А. А. Слинкин.- ДМК Пресс, 2017. – 249 с.

17 Жерон, Орельен. Прикладное машинное обучение с помощью Scikit-Learn и TensorFlow: концепции, инструменты и техники для создания интеллектуальных систем.Пер. с англ. - СпБ.: ООО Альфа-книга: 2018. - 688 с.

18 Flach P. Machine learning. The art and science of building algorithms. pp. 118-142.

19 Кузнецов И.Н. Пример решения задачи множественной регрессии с помощью Python [Электронный ресурс]: – Режим доступа: https://habr.com/ru/post/206306/ (дата обращения: 13.06.2022).

20 Makusheva N.Yu., Kolosova N.B. Comparative analysis of metal reinforcement and fibre-reinforced plastic rebar Construction of Unique Buildings and Structures, 2014, No10 (25) Pp. 60-72

21 Плас Дж. Вандер. Python для сложных задач: наука о данных и машинное обучение. / Дж. Вандер Плас. – СПб.: Питер, 2018. 576 с.

22 Сохина С.А., Немченко С.А. Машинное обучение. Методы машинного обучения // Современная наука в условиях модернизационных процессов: проблемы, реалии, перспективы. 2021. С. 165-168.

23 Субботина С.А., Шлыкова И.Д., Авдеева А.А., Одинокова Г.В., Соколова Н.В. Виды композитных материалов: стеклопластик, углепластик, базальтопластик // Синергия наук. 2017. № 18. − С. 641-645.

24 Таршхоева Ж.Т. Zзык программирования Python. Библиотеки Python // Молодой ученый. 2021. № 5 (347). С. 20-21.

25 Токарев В. Сравнение арматурных прутьев из базальтопластика и углепластика [Электронный ресурс]: – Режим доступа: http://cemgid.ru/sravnenie-armaturnyx-prutev-iz-bazaltoplastika-i-ugleplastika.html (дата обращения: 12.06.2022).

26 Щелконогов А.Н. Разработка простейших нейросетей в keras // Математическое и программное обеспечение вычислительных систем. 2019. С. 51-53.

27 Выбор алгоритмов машинного обучения Azure [Электронный ресурс]: – Режим доступа: https://docs.microsoft.com/ru-ru/azure/machine-learning/how-to-select-algorithms (дата обращения: 13.06.2022).

28 Нейросеть (регрессия) [Электронный ресурс]: – Режим доступа: https://help.loginom.ru/userguide/processors/datamining/neural-network-regression.html (дата обращения: 13.06.2022).

29 Машинное обучение [Электронный ресурс]: – Режим доступа: https://ru.wikipedia.org/wiki//Машинное_обучение (дата обращения: 13.06.2022).

30 Комплексная платформа машинного обучения с открытым исходным кодом [Электронный ресурс]: – Режим доступа: https://www.tensorflow.org/ (дата обращения: 12.06.2022).

31 Платформа scikit-learn [Электронный ресурс]: – Режим доступа: https://scikit-learn.org/stable/ (дата обращения: 12.06.2022).