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»

Тема: «Прогнозирование конечных свойств новых материалов (композиционных материалов)»

Слушатель Кузьмина Е.В.

Москва, 2023

Содержание

Введение

Предметом исследования в данной работе выступают композитные материалы. Это искусственно созданные материалы, применяемые во многих сферах деятельности: строительстве, авиационной промышленности, медицине, автомобилестроении и многих других.

Композитные материалы – это многокомпонентные материалы, изготовленные из двух или более компонентов с существенно различными физическими или химическими свойствами, которые приводят к появлению нового материала с характеристиками, отличными от характеристик отдельных компонентов.

Сложность получения новых композитов заключается в прогнозировании свойств будущих материалов, поэтому для упрощения и удешевления исследований можно применять машинное обучение. Прогнозные модели помогут уменьшить количество проводимых испытаний.

В этой работе воспроизведено исследование с анализом данных, а также созданы модели и нейросеть с набором различных параметров для прогнозирования конечных свойств новых композитных материалов.

1. Аналитическая часть

1.1 Постановка задачи

Цель прогнозирования заключается в симуляции представительного элемента объема композита на основе данных о характеристиках входящих компонентов (связующего и армирующего компонента). Необходимо создать и обучить модели или нейронные сети.

Входные данные состоят из двух датасетов. В файле с физическими характеристиками базальтопластика (X_bp.xlsx) содержится 1023 строки и 10 признаков:

- Соотношение матрица-наполнитель;
- Плотность;
- Модуль упругости;
- Количество отвердителя;
- Содержание эпоксидных групп;
- Температура вспышки;
- Поверхностная плотность;
- Модуль упругости при растяжении;
- Прочность при растяжении;
- Потребление смолы.

	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
0	1.857143	2030.0	738.736842	30.00	22.267857	100.000000	210.0	70.0	3000.0	220.0
1	1.857143	2030.0	738.736842	50.00	23.750000	284.615385	210.0	70.0	3000.0	220.0
2	1.857143	2030.0	738.736842	49.90	33.000000	284.615385	210.0	70.0	3000.0	220.0
3	1.857143	2030.0	738.736842	129.00	21.250000	300.000000	210.0	70.0	3000.0	220.0
4	2.771331	2030.0	753.000000	111.86	22.267857	284.615385	210.0	70.0	3000.0	220.0

Рисунок 1 – X_bp, характеристики базальтопластика

Файл с геометрическими характеристиками нашивки углепластика (X_nup.xlsx) содержит 1040 строк и 3 параметра:

- Угол нашивки;

- Шаг нашивки;
- Плотность нашивки.

	Угол нашивки, град	Шаг нашивки	Плотность нашивки
0	0	4.0	57.0
1	0	4.0	60.0
2	0	4.0	70.0
3	0	5.0	47.0
4	0	5.0	57.0

Рисунок 2 – X_nup, характеристики углепластика

По заданию, таблицы были объединены в один датасет по индексу типом объединения INNER. Пример полученного датасета на рисунке 3.

	0	1	2	3	4
Соотношение матрица-наполнитель	1.857143	1.857143	1.857143	1.857143	2.771331
Плотность, кг/м3	2030.000000	2030.000000	2030.000000	2030.000000	2030.000000
модуль упругости, ГПа	738.736842	738.736842	738.736842	738.736842	753.000000
Количество отвердителя, м.%	30.000000	50.000000	49.900000	129.000000	111.860000
Содержание эпоксидных групп,%_2	22.267857	23.750000	33.000000	21.250000	22.267857
Температура вспышки, С_2	100.000000	284.615385	284.615385	300.000000	284.615385
Поверхностная плотность, г/м2	210.000000	210.000000	210.000000	210.000000	210.000000
Модуль упругости при растяжении, ГПа	70.000000	70.000000	70.000000	70.000000	70.000000
Прочность при растяжении, МПа	3000.000000	3000.000000	3000.000000	3000.000000	3000.000000
Потребление смолы, г/м2	220.000000	220.000000	220.000000	220.000000	220.000000
Угол нашивки, град	0.000000	0.000000	0.000000	0.000000	0.000000
Шаг нашивки	4.000000	4.000000	4.000000	5.000000	5.000000
Плотность нашивки	57.000000	60.000000	70.000000	47.000000	57.000000

Рисунок 3 – Объединённый датасет

Выходными переменными по объединённому датасету являются:

- Соотношение матрица-наполнитель;

- Модуль упругости при растяжении, ГПа;
- Прочность при растяжении, Мпа.

Для каждого параметра необходимо вычислить среднее и медианное значения, произвести анализ и исключение выбросов, проверить наличие пропусков и провести предварительную обработку данных, включая удаление выбросов. Также следует применить нормализацию и стандартизацию, и приступить к обучению моделей для выходных переменных. После, написать нейронную сеть, которая будет рекомендовать соотношение «матрица/наполнитель» и разработать по ней приложение с графическим интерфейсом. Затем нужно оценить точность модели на тренировочном и тестовом датасете, а также создать репозиторий в GitHub и разместить код исследования.

1.2 Описание используемых методов

В рамках классификации категорий машинного обучения, задача данной работы относится к машинному обучению с учителем и, чаще всего, это задача регрессии.

Задача регрессии в машинном обучении – это предсказание параметра Y по известному параметру X, то есть, модель прогнозирует значение метки по набору связанных компонентов. Метка здесь может принимать любое значение, а не просто выбирается из конечного набора значений, как в задачах классификации.

Алгоритмы регрессии моделируют зависимость меток от связанных компонентов, чтобы определить закономерности изменения меток при разных значениях компонентов. На вход алгоритма регрессии подается набор примеров с метками известных значений. Результатом работы алгоритма регрессии является функция, которая умеет прогнозировать значения метки для любого нового набора входных компонентов.

Для наилучшего решения в процессе исследования были применены следующие методы:

- Случайный лес;
- Линейная регрессия;
- Градиентный бустинг;
- Гребневая регрессия (ридж-регрессия);
- Лассо-регрессия.

Случайный лес (RandomForest) — это множество решающих деревьев. Универсальный алгоритм машинного обучения с учителем, представитель ансамблевых методов. Если точность дерева решений оказалось недостаточной, мы можем множество моделей собрать в коллектив.

Достоинства метода: не переобучается; не требует предобработки входных данных; эффективно обрабатывает пропущенные данные, данные с большим числом классов и признаков; имеет высокую точность предсказания и внутреннюю оценку обобщающей способности модели, а также высокую параллелизуемость и масштабируемость.

Недостатки метода: построение занимает много времени; сложно интерпретируемый; не обладает возможностью экстраполяции; может недообучаться; трудоёмко прогнозируемый; иногда работает хуже, чем линейные методы.

Линейная регрессия (Linear regression) — это алгоритм машинного обучения, основанный на контролируемом обучении, рассматривающий зависимость между одной входной и выходными переменными. Это один из самых простых и эффективных инструментов статистического моделирования. Она определяет зависимость переменных с помощью линии наилучшего соответствия. Модель регрессии создаёт несколько метрик. R^2, или коэффициент детерминации, позволяет измерить, насколько модель может объяснить дисперсию данных. Если R^2 равен 1, это значит, что модель описывает все данные. Если же R^2 равен 0,5, модель объясняет лишь 50 процентов дисперсии

данных. Оставшиеся отклонения не имеют объяснения. Чем ближе R^2 к единице, тем лучше.

Достоинства метода: быстрый и простой в реализации; легко интерпретируем; имеет меньшую сложность по сравнению с другими алгоритмами.

Недостатки метода: моделирует только прямые линейные зависимости; требует прямую связь между зависимыми и независимыми переменными; выбросы оказывают огромное влияние, а границы линейны.

Градиентный бустинг (Gradient Boosting) — это ансамбль деревьев решений, обученный с использованием градиентного бустинга. В основе данного алгоритма лежит итеративное обучение деревьев решений с целью минимизировать функцию потерь. Основная идея градиентного бустинга: строятся последовательно несколько базовых классификаторов, каждый из которых как можно лучше компенсирует недостатки предыдущих. Финальный классификатор является линейной композицией этих базовых классификаторов.

Достоинства метода: новые алгоритмы учатся на ошибках предыдущих; требуется меньше итераций, чтобы приблизиться к фактическим прогнозам; наблюдения выбираются на основе ошибки; прост в настройке темпа обучения и применения; легко интерпретируем.

Недостатки метода: необходимо тщательно выбирать критерии остановки, иначе это может привести к переобучению; наблюдения с наибольшей ошибкой появляются чаще; слабее и менее гибок, чем нейронные сети.

Ридж-гегрессия (Ridge Regression) – это метод регрессионного анализа, который используется для уменьшения переобучения модели и улучшения ее обобщающей способности. Он основан на добавлении штрафа (регуляризации) к сумме квадратов ошибок (RSS) в функции потерь, которая минимизируется при обучении модели. Этот штраф добавляет некоторую величину к диагональным элементам матрицы признаков, что приводит к уменьшению весов признаков и предотвращает переобучение. Метод может быть использован для решения

проблемы мультиколлинеарности, когда признаки в модели сильно коррелируют между собой.

Достоинства метода: помогает снизить переобучение модели, что может произойти, когда количество признаков превышает количество наблюдений; уменьшает дисперсию оценок коэффициентов регрессии, что может произойти, когда в данных присутствует мультиколлинеарность; может улучшить качество модели при наличии мультиколлинеарности.

Недостатки метода: усложняет интерпретацию коэффициентов регрессии, так как они могут быть сильно сжаты; не удаляет признаки из модели, а только уменьшает их вес, что может привести к тому, что некоторые признаки останутся в модели, несмотря на то, что они не являются значимыми; может быть неэффективным для больших данных, так как он требует вычисления обратной матрицы, что может быть вычислительно затратным.

Метод лассо-регрессии (Lasso Regression) – это метод регрессионного анализа, который используется для выбора наиболее значимых переменных из набора предикторов. Он является модификацией метода гребневой регрессии (Ridge Regression) и использует L1-регуляризацию для уменьшения весов модели.

Достоинства метода: позволяет выбрать наиболее значимые переменные из набора предикторов, что упрощает модель и уменьшает риск переобучения; может работать с большим количеством предикторов, что делает его полезным для анализа данных с большим числом переменных; может использоваться для уменьшения размерности данных, что упрощает анализ и уменьшает риск переобучения.

Недостатки метода: выбор наиболее значимых переменных может быть субъективным и зависеть от выбранного значения параметра регуляризации; может быть неустойчивым при наличии коррелированных предикторов, что может привести к выбору неправильных переменных; требует выбора параметра регуляризации, что может быть сложным и требовать определенных знаний и опыта.

Все вышеперечисленные задачи в данной работе решены на языке Python с использованием библиотек Pandas, NumPy, Matplotlib, Seaborn и Tensorflow, Scikit-Learn.

Python — высокоуровневый язык программирования общего назначения с динамической строгой типизацией и автоматическим управлением памятью, ориентированный на повышение производительности разработчика, читаемости кода и его качества, а также на обеспечение переносимости написанных на нём программ.

Pandas — это библиотека машинного обучения, представляющая структуры данных высокого уровня и большой диапазон инструментов для анализа. Отличительной чертой Pandas считается возможность переводить сложнейшие операции с информацией, используя всего одну либо две команды. Данная библиотека содержит массу способов для объединения данных, их группировки и фильтрации.

К особенностям Pandas относятся:

- возможность упростить манипуляции данными;

- поддержка сортировки, визуализации и прочих опций.

Pandas обеспечивает широкую гибкость, функциональность, если эксплуатировать ее с иными библиотеками.

NumPy — основная библиотека Python, которая упрощает работу с векторами и матрицами. Содержит готовые методы для разных математических операций: от создания, изменения формы, умножения и расчета детерминант матриц, до решения линейных уравнений и сингулярного разложения. Это значительно повышает производительность и, соответственно, ускоряет время выполнения работы.

Matplotlib — низкоуровневая библиотека для создания двумерных диаграмм и графиков. С ее помощью можно отображать широкий спектр визуализаций: линейные и точечные диаграммы, диаграммы с областями, гистограммы, круговые диаграммы, диаграммы «стебель-листья», контурные графики, поля векторов и спектрограммы.

Seaborn — библиотека более высокого уровня, чем matplotlib. С ее помощью проще создавать специфическую визуализацию: тепловые карты, временные ряды и скрипичные диаграммы.

TensorFlow — это библиотека AI, которая помогает разработчикам создавать крупномасштабные нейронные сети со многими слоями, используя графики потоков данных. TensorFlow также облегчает построение моделей глубокого обучения, продвигает современную технологию ML / AI и позволяет легко развертывать приложения на базе ML. TensorFlow достаточно эффективен, когда дело доходит до классификации, восприятия, понимания, обнаружения, прогнозирования и создания данных.

Scikit-learn – библиотека, которая основана на NumPy и SciPy. В ней есть алгоритмы для машинного обучения и интеллектуального анализа данных: кластеризации, регрессии и классификации.

Особенностями Scikit-Learn являются: возможность извлечения элементов из текстов и картинок; перекрестная проверка – множество различных методов проверки точности контролируемой модели на невидимой информации; большое количество алгоритмов машинного обучения; возможность осуществления дорогостоящих задач.

1.3 Разведочный анализ данных

Прежде чем передать данные в работу моделей машинного обучения, необходимо обработать и очистить их. Необработанные данные могут содержать искажения и пропущенные значения, что может негативно отразиться на работе моделей.

Цель разведочного анализа – получение первоначальных представлений о характерах распределений переменных исходного набора данных, формирование оценки качества исходных данных (наличие пропусков, выбросов), выявление характера взаимосвязи между переменными с целью

последующего выдвижения гипотез о наиболее подходящих для решения задачи моделях машинного обучения.

В качестве инструментов разведочного анализа используется: оценка статистических характеристик датасета; гистограммы распределения каждой из переменной; диаграммы ящика с усами; попарные графики рассеяния точек; тепловая карта; описательная статистика для каждой переменной; анализ и полное исключение выбросов; проверка наличия пропусков и дубликатов; ранговая корреляция Пирсона.

Команда df.info() выводит общую информацию о датасете: количество строк и столбцов, количество значений, название переменных, тип данных. Результат команды представлен на рисунке 4.

```
<class 'pandas.core.frame.DataFrame'>
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   int64  
 11  Шаг нашивки                           1023 non-null   float64
 12  Плотность нашивки                     1023 non-null   float64
dtypes: float64(12), int64(1)
memory usage: 111.9 KB
```

Рисунок 4 – Общая информация о датасете

Для того, чтобы оценить взаимосвязи между признаками была составлена корреляционная матрица, которая представлена на рисунке 5.

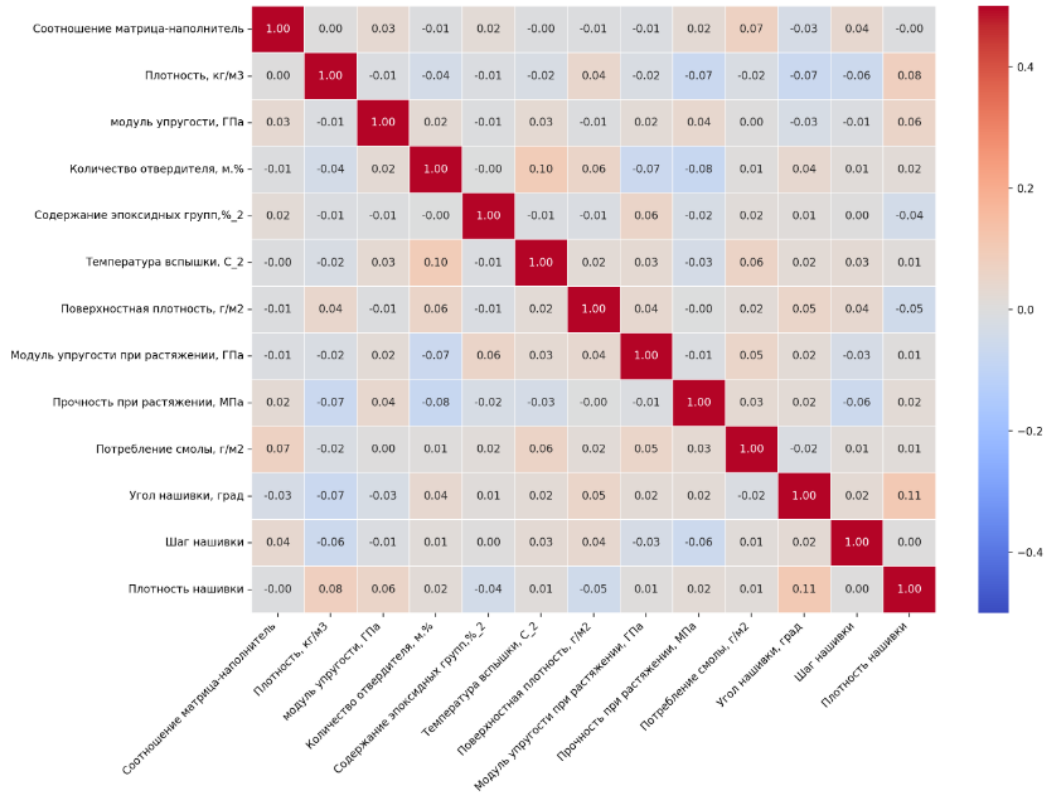

Рисунок 5 – Корреляционная матрица признаков

Проанализировав матрицу, можно сделать вывод, что признаки между собой не имеют линейной зависимости, так как между ними низкий коэффициент корреляции.

Важным этапом разведочного анализа является выявление выбросов в данных и их устранение. Один из наиболее эффективных методов знакомства с выбросами является диаграмма «ящик с усами» или Boxplot. Такой вид диаграммы в удобной форме показывает медиану, нижний и верхний квартили, минимальное и максимальное значение выборки и выбросы.

Нижний и верхний концы ящика соответствуют первому и третьему квартилям (25% и 75% квантилям соответственно), а горизонтальная линия внутри ящика – медиане. Верхний «ус» продолжается вверх вплоть до максимального значения, но не выше полуторного межквартильного расстояния от верхней кромки ящика. Аналогично нижний «ус» продолжается вниз до

минимального значения, но не ниже полуторного межквартильного расстояния от нижней кромки ящика. Концы «усов» обозначаются небольшими горизонтальными линиями. А за пределами «усов» значения изображаются в виде отдельных точек – эти значения можно считать выбросами. Диаграммы Boxplot представлены на рисунке.

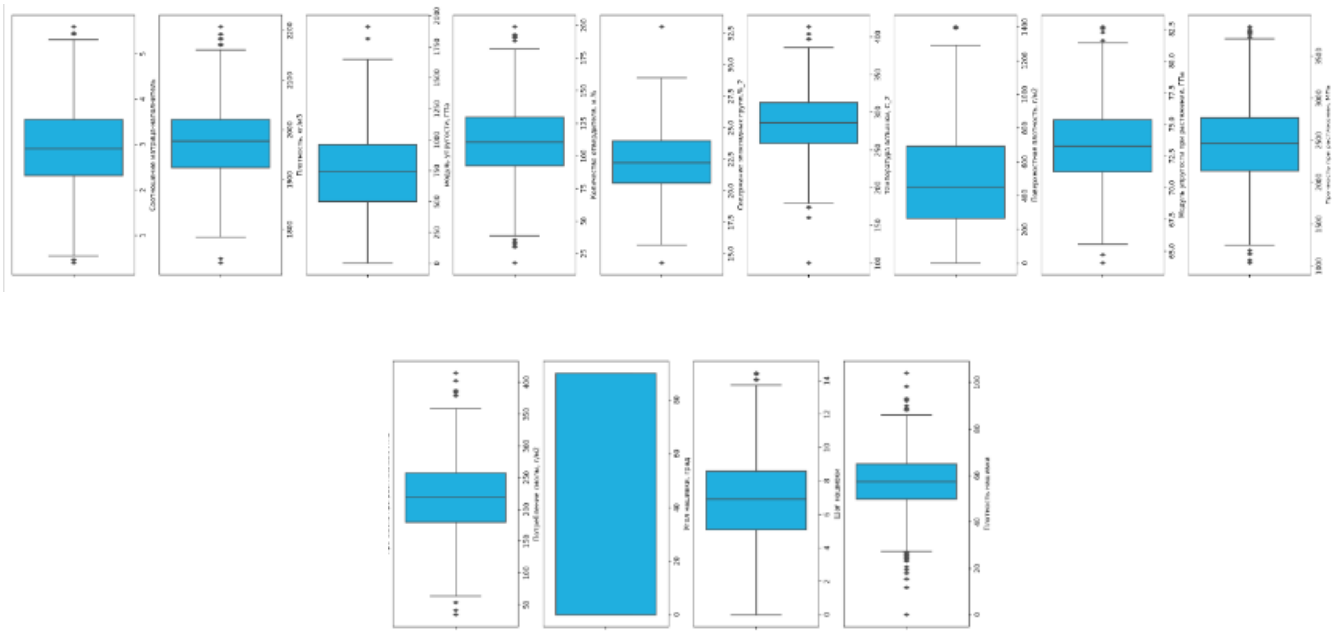

Рисунок 6 – Диаграммы Boxplot или «ящик с усами»

На диаграмме видно, что выбросы есть по всем характеристикам, кроме «Угол нашивки, град».

Следующим шагом в работе с данными было построение попарных графиков рассеяния точек. Присвоив каждой оси переменную, можно определить, существуют ли отношения или корреляция между этими двумя переменным. Отображаемые на диаграммах рассеяния паттерны позволяют увидеть разные типы корреляции. Среди них могут быть: положительная (оба значения увеличиваются), отрицательная (одно значение увеличивается, в то время как второе уменьшается), нулевая (отсутствие корреляции), линейная, экспоненциальная и подковообразная. Сила корреляции определяется по тому, насколько близко расположены друг от друга точки на графике. Данный график показал очень слабую зависимость между переменными датасета. Также имеем

возможность еще раз увидеть наличие некоторого количества выбросов – точки на графике, которые значительно удалены от общего кластера.

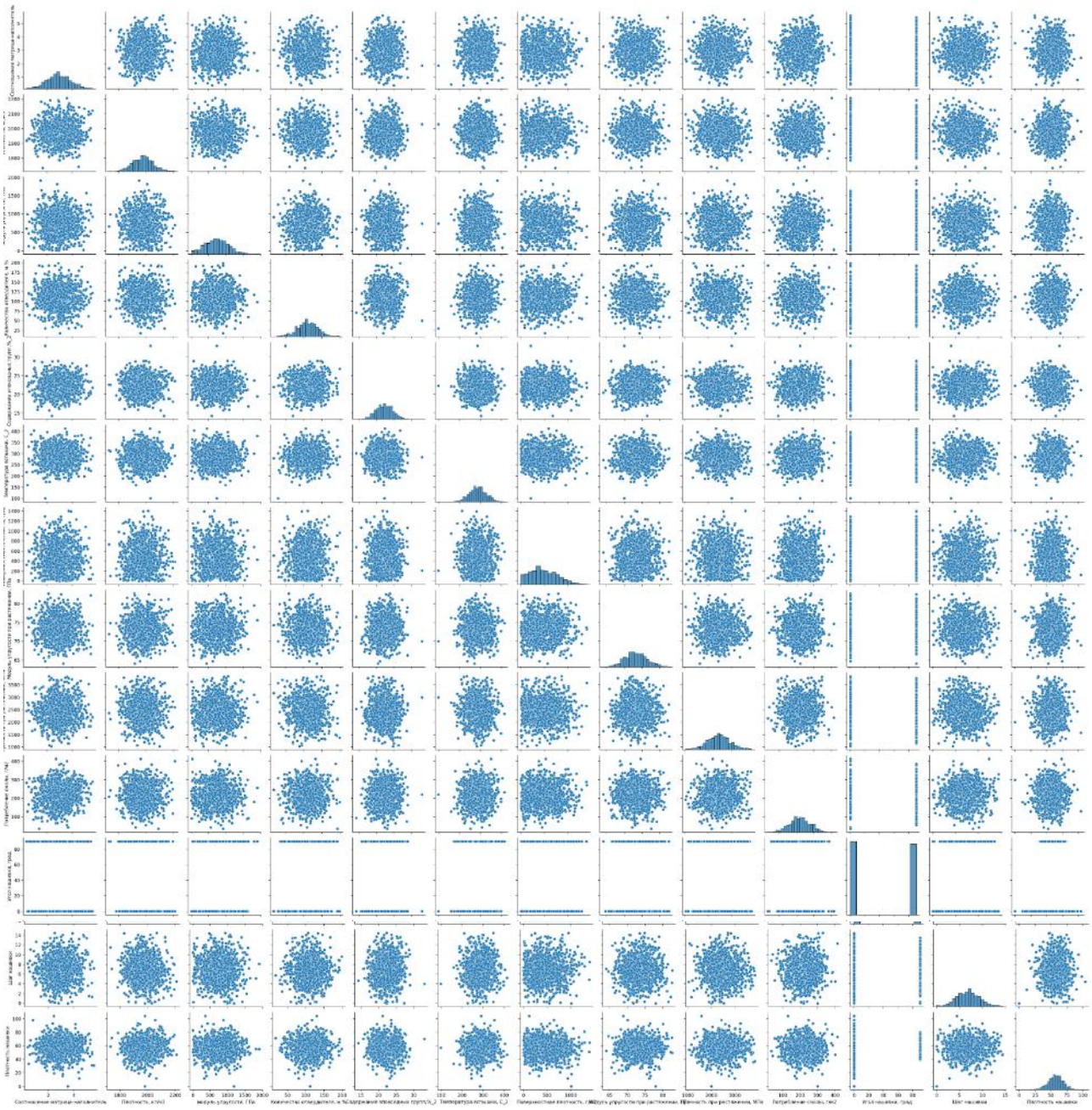

Рисунок 7 – Попарные графики рассеяния точек

2 Практическая часть

2.1 Предобработка данных

К задачам предварительной обработки данных относятся:

- Очистка данных;
- Редактирование данных;
- Заполнение пропусков.

При очистке данных удаляют устаревшие данные, дубликаты, аномалии, пропуски и ошибки. В данном датасете мы проведем очистку от выбросов методом межквартильного интервала.

Метод межквартильного интервала – это метод статистической обработки данных, который используется для определения выбросов в наборе данных. Он основан на интерквартильном расстоянии (IQR), которое является разницей между 75-м и 25-м процентилями данных. Для определения выбросов в наборе данных используется формула:

нижняя граница = Q1 - 1,5 * IQR

верхняя граница = Q3 + 1,5 * IQR

Если значение данных находится за пределами этих границ, то оно считается выбросом.

После удаления выбросов на рисунке 8 можно увидеть, что размерность датасета уменьшилась.

	1	3	4	5	6	7	8	9	10	11	...	1013	1014	1015	1016	1017	1018	1019	1020	1021	1022
Соотношение матрица-наполнитель	1.857143	1.857143	2.771331	2.767918	2.569620	2.561475	3.557018	3.532338	2.919678	2.877358	...	2.310394	1.646235	2.806563	3.745862	2.758727	2.271346	3.444022	3.280604	3.705351	3.808020
Плотность, кг/м3	2030.000000	2030.000000	2030.000000	2000.000000	1910.000000	1900.000000	1930.000000	2100.000000	2160.000000	1990.000000	...	1931.146887	2014.772547	1872.864660	1914.629424	2000.506141	1952.087902	2050.089171	1972.372865	2066.799773	1890.413468
модуль упругости, ГПа	738.736842	738.736842	753.000000	748.000000	807.000000	535.000000	889.000000	1421.000000	933.000000	1628.000000	...	554.010341	841.064806	996.018683	680.683701	934.564388	912.855545	444.732634	416.836524	741.475517	417.316232
Количество отвердителя, м.%	50.000000	129.000000	111.860000	111.860000	111.860000	111.860000	129.000000	129.000000	129.000000	129.000000	...	96.749782	102.979906	146.199194	110.979100	143.021859	86.992183	145.981978	110.533477	141.397963	129.183416
Содержание эпоксидных групп,% 2	23.750000	21.250000	22.267857	22.267857	22.267857	22.267857	21.250000	21.250000	21.250000	21.250000	...	22.146487	21.073367	21.559290	25.922635	21.379518	20.123249	19.599769	23.957502	19.246945	27.474763
Температура вспышки, С 2	284.615385	300.000000	284.615385	284.615385	284.615385	284.615385	300.000000	300.000000	300.000000	300.000000	...	214.827727	271.490843	313.900486	309.796388	273.852679	324.774576	254.215401	248.423047	275.779840	300.952708
Поверхностная плотность, г/м2	210.000000	210.000000	210.000000	210.000000	210.000000	380.000000	380.000000	1010.000000	1010.000000	1010.000000	...	56.242761	615.168127	799.634090	628.364550	65.105965	209.198700	350.660830	740.142791	641.468152	758.747882
Модуль упругости при растяжении, ГПа	70.000000	70.000000	70.000000	70.000000	70.000000	75.000000	75.000000	78.000000	78.000000	78.000000	...	78.143609	79.154469	72.815552	76.030555	67.633752	73.090961	72.920827	74.734344	74.042708	74.309704
Прочность при растяжении, МПа	3000.000000	3000.000000	3000.000000	3000.000000	3000.000000	1800.000000	1800.000000	2000.000000	2000.000000	2000.000000	...	1939.307550	2518.516089	2443.482888	2466.925422	3102.539548	2387.292495	2360.392784	2662.906040	2071.715856	2856.328932
Потребление смолы, г/м2	220.000000	220.000000	220.000000	220.000000	220.000000	120.000000	120.000000	300.000000	300.000000	300.000000	...	87.270139	232.428214	307.265172	152.184720	229.780372	125.007669	117.730099	236.606764	197.126067	194.754342
Угол нашивки, град	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	90.000000	90.000000	90.000000	90.000000	90.000000	90.000000	90.000000	90.000000	90.000000	90.000000
Шаг нашивки	4.000000	5.000000	5.000000	5.000000	5.000000	7.000000	7.000000	7.000000	7.000000	9.000000	...	7.683346	5.048503	5.240448	8.057020	8.736592	9.076380	10.565614	4.161154	6.313201	6.078902
Плотность нашивки	60.000000	47.000000	57.000000	60.000000	70.000000	47.000000	57.000000	60.000000	70.000000	47.000000	...	62.785021	59.837798	52.044507	47.067229	60.277805	47.019770	53.750790	67.629684	58.261074	77.434468

13 rows × 922 columns

Рисунок 8 – Датасет после удаления выбросов методом межквартильного интервала

Для того, чтобы удостовериться в «чистоте» данных, были произведены 2 повторных проверки тем же методом. Обнаруженные выбросы были также удалены.

Следующей задачей предварительной обработки данных является редактирование. Данные могут быть записаны с ошибками или в разных форматах, поэтому их нужно корректировать.

Датасет с характеристиками композитных материалов содержит числовые типы данных. Их минимальные и максимальные значения можно оценить на рисунке.

	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

Рисунок 9 – Распределение значений до нормализации данных

Что касается числовых данных, то, чтобы привести их к единому формату, можно преобразовать значения в диапазон от 0 до 1. Такое преобразование возможно сделать при помощи методов нормализации данных.

Одним из методов нормализации данных является MinMaxScaler. Этот метод используется для масштабирования признаков в диапазоне от 0 до 1. Он работает путем пересчета значений признаков на основе их минимального и максимального значений в наборе данных.

Формула для MinMax-нормализации следующая:

x_scaled = (x - x_min) / (x_max - x_min)

где:

x_scaled - отмасштабированное значение признака;

x - оригинальное значение признака;

x_min - минимальное значение признака в наборе данных;

x_max - максимальное значение признака в наборе данных.

	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	Шаг нашивки	Плотность нашивки
count	922.000000	922.000000	922.000000	922.000000	922.000000	922.000000	922.000000	922.000000	922.000000	922.000000	922.000000	922.000000	922.000000
mean	0.499412	0.502904	0.451341	0.506200	0.490578	0.516739	0.373295	0.487343	0.503776	0.507876	0.510846	0.503426	0.503938
std	0.187858	0.188395	0.201534	0.186876	0.180548	0.190721	0.217269	0.196366	0.188668	0.199418	0.500154	0.183587	0.193933
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.371909	0.368184	0.305188	0.378514	0.366571	0.386228	0.204335	0.353512	0.373447	0.374647	0.000000	0.372844	0.376869
50%	0.495189	0.511396	0.451377	0.506382	0.488852	0.516931	0.354161	0.483718	0.501481	0.510143	1.000000	0.506414	0.504310
75%	0.629774	0.624719	0.587193	0.638735	0.623046	0.646553	0.538397	0.617568	0.624299	0.642511	1.000000	0.626112	0.630842
max	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000

Рисунок 10– Распределение значений после нормализации данных

Третьей задачей предобработки датасета является заполнение пропусков, но в данной работе нет такой необходимости.

2.2 Разработка и обучение модели

Для решения поставленной задачи данные были разделены на обучающую и тестовую выборки в соотношении 70/30. Разработка и обучение моделей машинного обучения осуществлялась для двух выходных параметров: «Прочность при растяжении» и «Модуль упругости при растяжении» отдельно.

С помощью поиска по сетке с перекрестной проверкой были найдены лучшие гиперпараметры для каждой модели.

Для решения задачи предсказания модуля упругости при растяжении и прочности при растяжении были использованы следующие методы:

- Линейная регрессия;
- Случайный лес;

- Градиентный бустинг;
- Ридж-регрессия;
- Лассо-регрессия.

2.3 Тестирование модели

Оценка качества работы каждой из моделей выполнялась с помощью вычисления среднеквадратичной ошибки (MSE), средней абсолютной ошибки (MAE), средней абсолютной ошибки в процентах (MAPE), коэффициента детерминации (R2).

Показатели точности работы моделей для параметра «Модуль упругости при растяжении» представлены на рисунке.

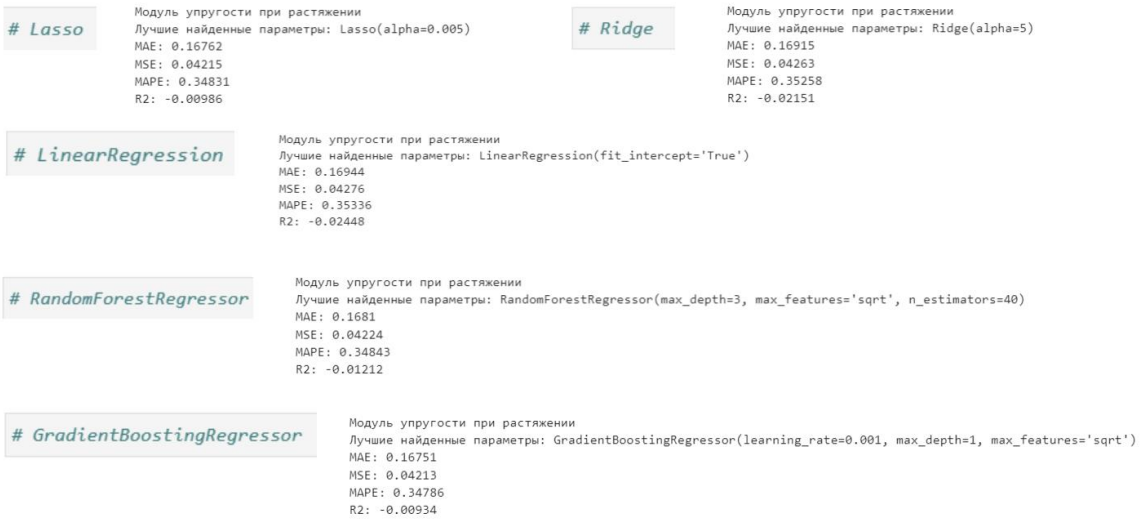

Рисунок 11 – Ошибки модели предсказания для параметра «Модуль упругости при растяжении»

Показатели точности работы моделей для параметра «Прочность при растяжении» представлены на рисунке 12.

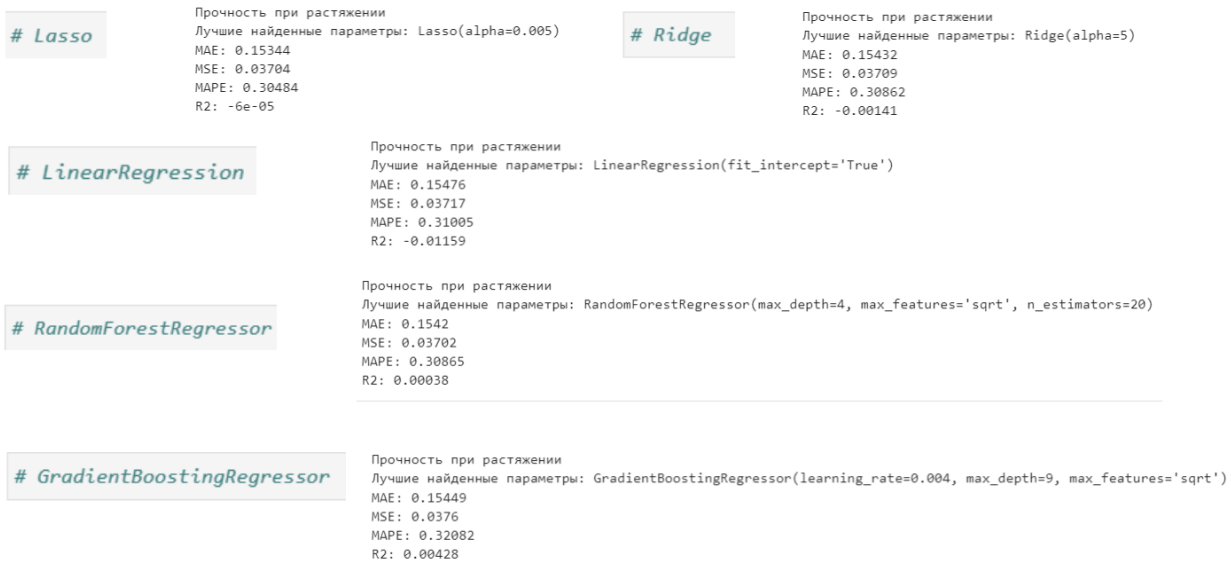

Рисунок 12 – Ошибки модели предсказания для параметра «Прочность при растяжении»

Результаты построения и обучения моделей, к сожалению, не дали значительного положительного результата. Наименьшая ошибка в предсказании для признака «Модуль упругости при растяжении» получилась у модели градиентного бустинга, в предсказании для признака «Прочность при растяжении» - у модели «случайный лес». Но в целом разброс показателей ошибок незначителен.

2.4 Нейронная сеть

Для построения рекомендательной системы признака «Соотношение матрица-наполнитель» использовали многослойный персептрон.

Первым шагом необходимо разделить очищенный от выбросов датасет на выходные данные в виде колонки «Соотношение матрица-наполнитель» и входные данные, которые включают все остальные колонки. Разделить входные и выходные данные на тренировочную и тестовую части в соотношении 70 и 30% с помощью train_test_split, после нормализовать данные используя TensorFlow.layers.Normalization.

После этого можно создавать нейронную сеть с помощью Sequential – это модель в библиотеке Keras, позволяющая создать нейронную сеть прямого распространения путем последовательного добавления слоев. Результат на рисунке.

```
nn_model = Sequential([
                    normalizer,
                    Dense(8, activation = 'relu'),
                    Dense(8, activation = 'relu'),
                    Dense(1)
                    ])

nn_model.compile(loss = 'mean_squared_error', optimizer = tf.keras.optimizers.Adam(0.0005))
```

```
nn_model.summary()
```

```
Model: "sequential_6"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 normalization (Normalizatio  (None, 12)              3
 n)

 dense_23 (Dense)            (None, 8)                 104

 dense_24 (Dense)            (None, 8)                 72

 dense_25 (Dense)            (None, 1)                 9

=================================================================
Total params: 188
Trainable params: 185
Non-trainable params: 3
```

Рисунок 13 – Информация о модели нейронной сети

Модель нейронной сети имеет следующие настраиваемые гиперпараметры:

- входной слой нормализации признаков;
- скрытые слои -;
- активационная функция скрытых слоев; relu - выполняет простое нелинейное преобразование поданных на вход данных (x). Возвращает x, если x > 0 и 0 в противном случае. Отличается высокой скоростью вычисления;
- нейронов в каждом скрытом слое: по 8;

- выходной слой с 1 нейроном (т.е. для одного признака), так как на выходе выводится одно значение для введенных данных;
- метод Adam (adaptive moment estimation) – оптимизационный алгоритм, используемый для обучения сети, основная функция которого – изменение весов для уменьшения ошибки сети в процессе обучения. Для каждого нейрона алгоритм изменяет веса индивидуально;
- оценка качества модели при помощи loss-функция: MeanSquaredError (MSE).

Далее было проведено обучение модели на тренировочных данных при помощи метода fit со следующими параметрами:

- аргумент validation_split позволяет автоматически зарезервировать часть тренировочных данных для валидации. Это необходимо для того, чтобы иметь возможность обучить модель и оценить результаты работу с данными параметрами, не затрагивая тестовую выборку. Значением аргумента является доля данных, которые должны быть зарезервированы, в нашем случае это 30% тренировочных данных;
- verbose – режим вывода информации о процессе обучения нейронной сети;
- epoch – количество повторений циклов обучения для всей выборки данных. В данном случае их 100.
- Итог обучения модели представлен на рисунке.

```
Epoch 100/100
15/15 [==============================] - 0s 7ms/step - loss: 0.0351 - val_loss: 0.0457
Wall time: 10.4 s
```

Рисунок 14 – Обучение модели нейронной сети

По результатам обучения необходимо построить график, на котором две кривых: отображение среднеквадратической ошибки модели на тестовых (голубая линия) и валидационных данных (оранжевая линия) относительно числа итераций. На рисунке можно увидеть, что линии идут рядом, ошибка

постепенно снижается и выходит на плато, где остается приблизительно на одном уровне до конца обучения.

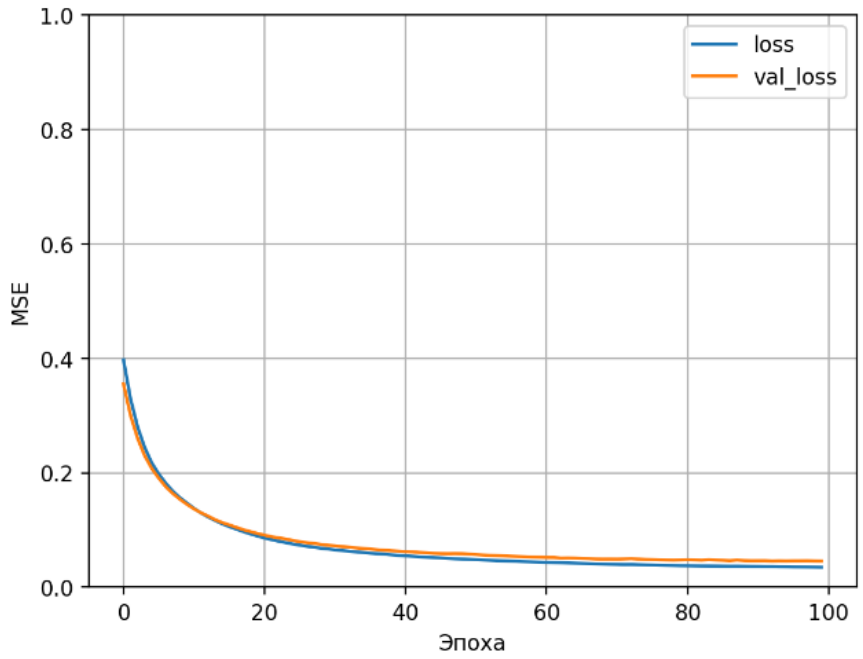

Рисунок 15 – Визуализация ошибки модели нейронной сети

2.5. Создание удаленного репозитория и загрузка результатов работы на него.

Ссылка на репозиторий в Github: **https://github.com/11airplanes/VKR**

Заключение

В рамках проведенного исследования были изучены теоретические основы методов машинного обучения и основные библиотеки языка программирования Python, которые являются важными инструментами для анализа данных. В практической части работы были применены изученные методы машинного обучения и построения моделей на реальных данных.

Несмотря на то, что в данной работе не удалось разработать эффективную модель для прогнозирования конечных свойств новых композиционных материалов на основе данных об их составе и структуре, точность предсказанных значений была ниже среднего, был получен значительный практический опыт в области анализа, визуализации и предобработки данных, создания нейронных сетей и моделей машинного обучения.

Список использованной литературы

1) Композиционные материалы: Справочник /Под. ред. В.В. Васильева, Ю.М.Тарнопольского. –М.: Машиностроение, 1990. –512 с.

2) 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.

3) Силен Дэви, Мейсман Арно, Али Мохамед. Основы Data Science и Big Data. Python и наука о данных. – СПб.: Питер, 2017. – 336 с.: ил.

4) Платформа scikit-learn [Электронный ресурс]: – Режим доступа: https://scikit-learn.org/stable/ (дата обращения: 21.03.2023).

5) Библиотека Seaborn- Режим доступа: https://seaborn.pydata.org/. (дата обращения 22.03.2023)

6) Язык программирования Python- Режим доступа: https://www.python.org/. (дата обращения 12.03.2023)

7) Библиотека Pandas – Режим доступа: https://pandas.pydata.org/ (дата обращения 17.03.2023)

8) Библиотека Sklearn – Режим доступа: https://scikit-learn.org/stable/ (дата обращения 23.03.2023)

9) Библиотека Pandas- Режим доступа: https://pandas.pydata.org/. (дата обращения 22.03.2023)

10) Библиотека Matplotlib- Режим доступа: https://matplotlib.org/. (дата обращения 18.03.2023)

11) Библиотека Tensorflow: Режим доступа: https://www.tensorflow.org/. (дата обращения 24.03.2023)