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 Мещанова Анна Андреевна

Москва, 2023

Содержание

Введение

Композиционный материал – это искусственно созданный неоднородный сплошной материал, состоящий из двух или более нерастворимых друг в друге компонентов с чёткой границей раздела между ними. Композиционные материалы обладают комплексом свойств, которыми обладают компоненты, и свойствами, которыми отдельные компоненты не обладают. При этом композиты являются монолитным материалом, т. е. компоненты материала неотделимы друг от друга без разрушения конструкции в целом.

В большинстве композитов компоненты можно разделить на матрицу (отвержденное связующее) и включённые в неё армирующие элементы (волокна, нити, ткани и др.), равномерно распределенные в матрице и имеющие заданную пространственную ориентацию. В композитах конструкционного назначения наполнители обычно обеспечивают необходимые механические характеристики материала (прочность, жёсткость и т. д.), а матрица обеспечивает совместную работу наполнителей и защиту их от механических повреждений и агрессивной химической среды.

Полимерные композиционные материалы получают путем создания комбинации наполнителя и матрицы при заранее заданной пропорции обеих фаз с помощью определенных технологических приемов. В результате наполнения получают материалы, основные физические и механические свойства которых отличаются от свойств матрицы.

Актуальность темы обусловлена широким применением композиционных материалов в различных областях техники. Созданные прогнозные модели помогут сократить количество проводимых испытаний, а также пополнить базу данных материалов возможными новыми характеристиками материалов, и цифровыми двойниками новых композитов.

1 Аналитическая часть

1.1 Постановка задачи

В условиях расширения разнообразия материалов, используемых при проектировании новых композитов возрастает потребность в определении характеристики нового композита с наименьшими финансовыми затратами. Разработка композитных материалов является долгосрочным и дорогостоящим процессом, так как из свойств отдельных компонентов невозможно рассчитать конечные свойства композита, а для достижения определенных характеристик требуется большое количество различных комбинированных тестов. Сократить время и затраты на создание определенного материала могла бы помочь система поддержки производственных решений, построенная на принципах машинного обучения.

На входе задачи имеются данные о начальных свойствах компонентов композиционных материалов, на выходе необходимо спрогнозировать ряд конечных свойств получаемых композиционных материалов. Полученный от структурного подразделения МГТУ им. Н.Э. Баумана набор данных состоит из двух частей в формате MS Excel: X_bp (характеристики полученной в ходе экспериментов матрицы из базальтопластика) и X_nup (характеристики нашивки углепластика). Первый файл содержит 1 023 значения по следующим десяти параметрам:

1) Соотношение матрица-наполнитель;
2) Плотность, кг/м3;
3) Модуль упругости, Гпа;
4) Количество отвердителя, м.%;
5) Содержание эпоксидных групп,%_2;
6) Температура вспышки, С_2;
7) Поверхностная плотность, г/м2;
8) Модуль упругости при растяжении, Гпа;
9) Прочность при растяжении, Мпа;
10) Потребление смолы, г/м2.

Второй файл содержит 1 040 значений по трем параметрам:

1) Угол нашивки, град;
2) Шаг нашивки;
3) Плотность нашивки.

После объединения указанных датафреймов по типу объединения INNER получен датасет объемом выборки в 1 023 значения, не содержащий пропусков и включающий только числовые значения (рис. 1).

```
<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 1 – Сведения о количестве значений и формате данных

Целевыми факторами определены модуль упругости при растяжении, прочность при растяжении, а также соотношение матрица-наполнитель.

1.2 Описание используемых методов

Данная задача является задачей регрессии и относится к машинному обучению с учителем. Цель любого алгоритма обучения с учителем —

определить функцию потерь и минимизировать её, поэтому для наилучшего решения были исследованы и применены следующие методы:

- линейная регрессия (Linear regression);
- гребневая регрессия (Ridge);
- лассо регрессия (Lasso);
- эластичная регрессия (ElasticNet);
- градиентный бустинг (Gradient boosting);
- метод К-ближайших соседей (K-Nearest neighbor);
- дерево решений (Decision tree);
- случайный лес (Random forest);
- адаптивный бустинг (AdaBoostRegressor);
- многослойный перцептрон (Multilayered perceptron).

LinearRegression

Линейная регрессия — 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. R2, или коэффициент детерминации, позволяет измерить, насколько модель может объяснить дисперсию данных. Если R-квадрат равен 1, это значит, что модель описывает все данные. Если же R-квадрат равен 0,5, модель объясняет лишь 50 процентов дисперсии данных. Оставшиеся отклонения не имеют объяснения. Чем ближе R2 к единице, тем лучше.

Достоинства метода: быстр и прост 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.

Чтобы улучшить Линейную модель путем обмена некоторой этой дисперсии с предвзятостью, чтобы уменьшить нашу общую ошибку. Это происходит при помощи регуляризации, в которой модифицируется функция стоимости, чтобы ограничить значения коэффициентов. Это позволяет изменить чрезмерную дисперсию на некоторое смещение, потенциально уменьшая общую ошибку.

Ridge

Гребневая регрессия – это регрессия, которая добавляет дополнительный штраф к функции стоимости, но вместо этого суммирует квадраты значений коэффициентов (норма L-2) и умножает их на некоторую постоянную лямбду. По сравнению с Лассо этот штраф регуляризации уменьшит значения коэффициентов, но не сможет принудительно установить коэффициент равным 0. Это ограничивает использование регрессии гребня в отношении выбора признаков. Однако, когда p> n, он способен выбрать более n релевантных предикторов, если необходимо, в отличие от Лассо. Он также выберет группы коллинеарных элементов, которые его изобретатели назвали «эффектом группировки».

Как и в случае с Лассо, мы можем варьировать лямбду, чтобы получить модели с различными уровнями регуляризации, где лямбда = 0 соответствует OLS, а лямбда приближается к бесконечности, что соответствует постоянной функции.

Анализ регрессии Лассо, так и Риджа показазывает, что ни один метод не всегда лучше, чем другой; нужно попробовать оба метода, чтобы определить, какой использовать.

Ридж-регрессию лучше применять, когда предсказательная способность набора данных распределена между различными характеристиками. Ридж-регрессия не обнуляет характеристики, которые могут быть полезны при составлении прогнозов, а просто уменьшает вес большинства переменных в модели.

Lasso

Лассо регрессия – это линейная модель, которая оценивает разреженные коэффициенты. Это простой метод, позволяющий уменьшить сложность модели и предотвратить переопределение, которое может возникнуть в результате простой линейной регрессии. Данный метод вводит дополнительное слагаемое регуляризации в оптимизацию модели. Это даёт более устойчивое решение. В регрессии лассо добавляется условие смещения в функцию оптимизации для того, чтобы уменьшить коллинеарность и, следовательно, дисперсию модели. Но вместо квадратичного смещения, используется смещение абсолютного значения. Лассо регрессия хорошо прогнозирует модели временных рядов на основе регрессии, таким как авторегрессии.

Достоинства метода: легко полностью избавляется от шумов в данных; быстро работает; не очень энергоёмко; способно полностью убрать признак из датасета; доступно обнуляет значения коэффициентов.

Недостатки метода: часто страдает качество прогнозирования; выдаёт ложное срабатывание результата; случайным образом выбирает одну из коллинеарных переменных; не оценивает правильность формы взаимосвязи между независимой и зависимой переменными; не всегда лучше, чем пошаговая регрессия.

Лассо-регрессию следует использовать, когда есть несколько характеристик с высокой предсказательной способностью, а остальные бесполезны. Она обнуляет бесполезные характеристики и оставляет только подмножество переменных.

ElasticNet

Эластичная сеть – это регрессия, которая включает в себя термины регуляризации как L-1, так и L-2. Это дает преимущества регрессии Лассо и Риджа. Было установлено, что он обладает предсказательной способностью лучше, чем у Лассо, хотя все еще выполняет выбор функций. Поэтому получается лучшее из обоих методов, выполняя выбор функции Лассо с выбором группы объектов Ridge.

Elastic Net поставляется с дополнительными издержками на определение двух лямбда-значений для оптимальных решений.

Компромисс смещения дисперсии – это компромисс между сложной и простой моделью, в которой промежуточная сложность, вероятно, является наилучшей.

Лассо, Ридж-регрессия и Эластичная сеть – это модификации обычной линейной регрессии наименьших квадратов, которые используют дополнительные штрафные члены в функции стоимости, чтобы сохранить значения коэффициента небольшими и упростить модель.

Лассо полезно для выбора функций, когда наш набор данных имеет функции с плохой предсказательной силой.

Регрессия гребня полезна для группового эффекта, при котором коллинеарные элементы могут быть выбраны вместе.

Elastic Net сочетает в себе регрессию Лассо и Риджа, что потенциально приводит к модели, которая является простой и прогнозирующей.

GradientBoostingRegressor

Градиентный бустинг — 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, или это может привести к переобучению, наблюдения с наибольшей ошибкой появляются чаще; слабее и менее гибок чем нейронные сети.

KneighborsRegressor

Метод К-ближайших соседей ищет ближайшие объекты с известными значения целевой переменной и основывается на хранении данных в памяти для сравнения с новыми элементами. Алгоритм находит расстояния между запросом и всеми примерами в данных, выбирая определенное количество примеров (k), наиболее близких к запросу, затем голосует за наиболее часто встречающуюся метку (в случае задачи классификации) или усредняет метки (в случае задачи регрессии).

Достоинства метода: прост в реализации и понимании полученных результатов; имеет низкую чувствительность к выбросам; не требует построения модели; допускает настройку нескольких параметров; позволяет делать дополнительные допущения; универсален; находит лучшее решение из возможных; решает задачи небольшой размерности.

Недостатки метода: замедляется с ростом объёма данных; не создаёт правил; не обобщает предыдущий опыт; основывается на всем массиве доступных исторических данных; невозможно сказать, на каком основании строятся ответы; сложно выбрать близость метрики; имеет высокую зависимость результатов классификации от выбранной метрики; полностью перебирает всю обучающую выборку при распознавании; имеет вычислительную трудоемкость.

DecisionTreeRegressor

Дерево решений – метод автоматического анализа больших массивов данных. Это инструмент принятия решений, в котором используется древовидная структура, подобная блок-схеме, или модель решений и всех их возможных результатов, включая результаты, затраты и полезность. Дерево принятия решений - эффективный инструмент интеллектуального анализа данных и предсказательной аналитики. Алгоритм дерева решений подпадает под категорию контролируемых алгоритмов обучения. Он работает как для непрерывных, так и для категориальных выходных переменных. Правила генерируются за счёт обобщения множества отдельных наблюдений

(обучающих примеров), описывающих предметную область. Регрессия дерева решений отслеживает особенности объекта и обучает модель в структуре дерева прогнозированию данных в будущем для получения значимого непрерывного вывода. Дерево решений один из вариантов решения регрессионной задачи, в случае если зависимость в данных не имеет очевидной корреляции.

Достоинства метода: помогают визуализировать процесс принятия решения и сделать правильный выбор в ситуациях, когда результаты одного решения влияют на результаты следующих решений, создаются по понятным правилам; просты в применении и интерпретации; заполняют пропуски в данных наиболее вероятным решением; работают с разными переменными; выделяют наиболее важные поля для прогнозирования.

Недостатки метода: ошибаются при классификации с большим количеством классов и небольшой обучающей выборкой; имеют нестабильный процесс (изменение в одном узле может привести к построению совсем другого дерева); имеет затратные вычисления; необходимо обращать внимание на размер; ограниченное число вариантов решения проблемы.

RandomForestRegressor

Случайный лес — это множество решающих деревьев.

Универсальный алгоритм машинного обучения с учителем, представитель ансамблевых методов. Если точность дерева решений оказалось недостаточной, мы можем множество моделей собрать вместе.

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.

AdaBoostRegressor

Адаптивный бустинг – это алгоритм, который работает по принципу перевзвешивания результатов и используется для повышения производительности алгоритмов машинного обучения. Он лучше всего работает со слабыми обучающими алгоритмами, поэтому такие модели могут достигнуть точности гораздо выше случайной при решении задачи классификации. Наиболее распространенными алгоритмами, используемыми с AdaBoost, являются одноуровневые деревья решений. Алгоритм AdaBoost учится на ошибках, больше концентрируясь на сложных участках, с которыми от столкнулся в процессе предыдущей итерации обучения. На каждой итерации дается вес алгоритмам. Каждый новый алгоритм корректирует ошибки предыдущих до получения хорошего результата. Все прогнозы объединяются с помощью голосования для получения окончательного прогноза.

Достоинства метода: AdaBoost легко реализовать, достаточно класса моделей и их количества; итеративно исправляет ошибки слабого классификатора и повышает точность путем объединения слабых учащихся; не склонен к переоснащению; можно использовать многие базовые классификаторы с AdaBoost.

Недостатки метода: чувствителен к шумным данным; обучается дольше линейной регрессии, классификация дольше чем при использовании логистической регрессии; на AdaBoost сильно влияют отклонения, так как он пытается идеально подогнать каждую точку.

Perceptron

Многослойный персептрон — это искусственная нейронная сеть, имеющая три или более слоёв персептронов: один входной слой, один или более скрытых слоёв и один выходной слой персептронов (рис. 2).

Достоинства метода: построение сложных разделяющих поверхностей; возможность осуществления любого отображения входных векторов в

выходные; легко обобщает входные данные; не требует распределения входных векторов; изучает нелинейные модели.

Недостатки метода: имеет невыпуклую функцию потерь; разные инициализации случайных весов могут привести к разной точности проверки; требует настройки ряда гиперпараметров; чувствителен к масштабированию функций.

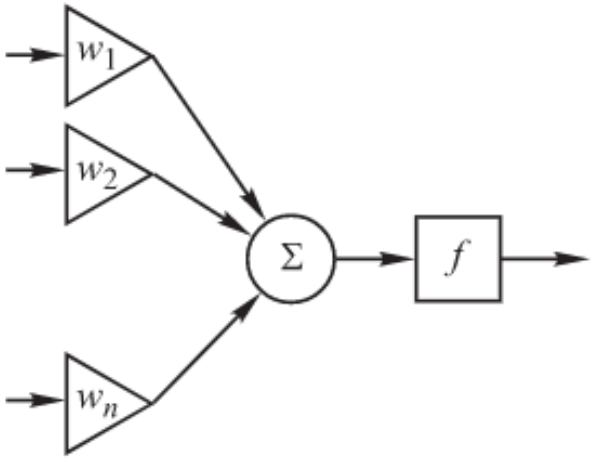

Рисунок 2 – Структура персептрона

Используемые метрики качества моделей

Под ошибкой обучения понимают разность между желаемым (целевым) и фактическим выходом модели на примерах обучающего множества. Особенно большую роль играет ошибка при обучении нейронной сети, поскольку производная ошибки на обучающем множестве используется для расчета коррекции весов нейронов. Если нейронная сеть имеет несколько выходных нейронов, то ошибка обучения определяется как средний квадрат ошибок на каждом выходе:

Ошибка обучения является показателем точности настройки модели на обучающем множестве и может использоваться в качестве условия остановки обучения. Однако она не позволяет оценить точность работы модели с новыми данными, не участвовавшими в процессе обучения.

Среднеквадратическая ошибке MSE (Mean Squared Error) принимает значениях в тех же единицах, что и целевая переменная. Чем ближе к нулю MSE, тем лучше работают предсказательные качества модели.

Средняя абсолютная ошибка MAE (Mean Absolute Error) рассчитывается как среднее абсолютных разностей между целевыми значением и значением, предсказанным моделью на данном обучающем примере в процессе обучения. В отличие от среднеквадратических ошибок, где используется квадрат разности, MAE является линейной оценкой, поэтому вес разностей одинаков независимо от диапазона.

Коэффициент детерминации R2 измеряет долю дисперсии, объяснённую моделью, в общей дисперсии целевой переменной. Если он близок к единице, то модель хорошо объясняет данные, если же он близок к нулю, то качество прогноза идентично средней величине целевой переменной (т. е. очень низкое). Отрицательные значение коэффициента детерминации означают плохую объясняющую способность модели.

Точная настройка модели

Определившись со списком используемых моделей, следует учесть необходимость проведения их точной настройки. Во избежание ручного перебора гиперпараметров используется класс GridSearchCV из библиотеки Scikit-Learn, выполняющий решетчатый поиск. Для этого нужно лишь сообщить ему, с какими параметрами нужно поэкспериментировать и какие значения опробовать, а GridSearchCV оценит все возможные комбинации значений гиперпараметров, применяя перекрестную проверку.

После выбора лучшей модели с оптимальными параметрами идет ее внедрение в продакшн путем разработки приложения, способного предсказать выходное значение после введения пользователем входных данных.

Таким образом, пайплайн машинного обучения можно изобразить следующим образом (рис. 3):

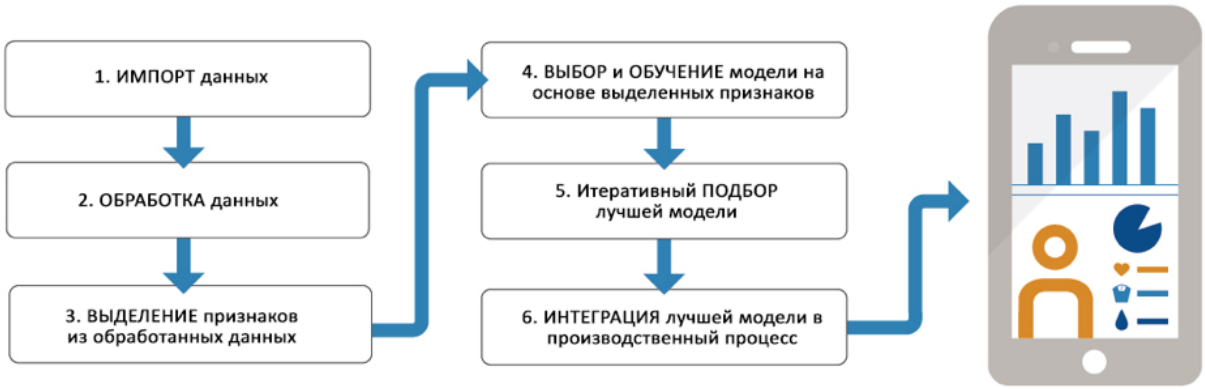

Рисунок 3 – Пайплайн машинного обучения

1.3 Разведочный анализ данных

Разведочный анализ — это предварительный анализ данных с целью выявления наиболее общих зависимостей, закономерностей и тенденций, характера и свойств анализируемых данных, законов распределения анализируемых величин, обнаружения отклонений и аномалий. Он применяется для нахождения связей между переменными в ситуациях, когда априорные представления о природе этих связей недостаточны или отсутствуют. Результаты разведочного анализа необходимы в разработке наилучшей стратегии углубленного анализа, для дальнейшего выдвижения гипотез, уточнения особенностей применения тех или иных математических методов и моделей.

К основным методам разведочного анализа относится процедура анализа распределений переменных, корреляционный анализ с целью поиска коэффициентов, превосходящих по величине определенные пороговые значения, факторный анализ, дискриминантный анализ, многомерное шкалирование, визуальный анализ гистограмм и т. д.

Исходная описательная статистика признаков датасета по композиционным материалам приведена в таблице 1.

Таблица 1 – Описательная статистика признаков

	count	mean	std	min	max
Соотношение матрица-наполнитель	1023.0	2.930366	0.913222	0.389403	5.591742
Плотность, кг/м3	1023.0	1975.734888	73.729231	1731.764635	2207.773481
модуль упругости, ГПа	1023.0	739.923233	330.231581	2.436909	1911.536477
Количество отвердителя, м.%	1023.0	110.570769	28.295911	17.740275	198.953207
Содержание эпоксидных групп,%_2	1023.0	22.244390	2.406301	14.254985	33.000000
Температура вспышки, С_2	1023.0	285.882151	40.943260	100.000000	413.273418
Поверхностная плотность, г/м2	1023.0	482.731833	281.314690	0.603740	1399.542362
Модуль упругости при растяжении, ГПа	1023.0	73.328571	3.118983	64.054061	82.682051
Прочность при растяжении, МПа	1023.0	2466.922843	485.628006	1036.856605	3848.436732
Потребление смолы, г/м2	1023.0	218.423144	59.735931	33.803026	414.590628
Угол нашивки, град	1023.0	44.252199	45.015793	0.000000	90.000000
Шаг нашивки	1023.0	6.899222	2.563467	0.000000	14.440522
Плотность нашивки	1023.0	57.153929	12.350969	0.000000	103.988901

Разведочный анализ следует начинать с обработки пропусков в данных и удаления дублирующих записей, так как они не только искажают статистические показатели датасета, но и снижают качество обучения модели. В нашем примере дубликаты и пропуски отсутствуют.

Следующим шагом рассмотрим данные наглядно с помощью различных диаграмм (рис. 4).

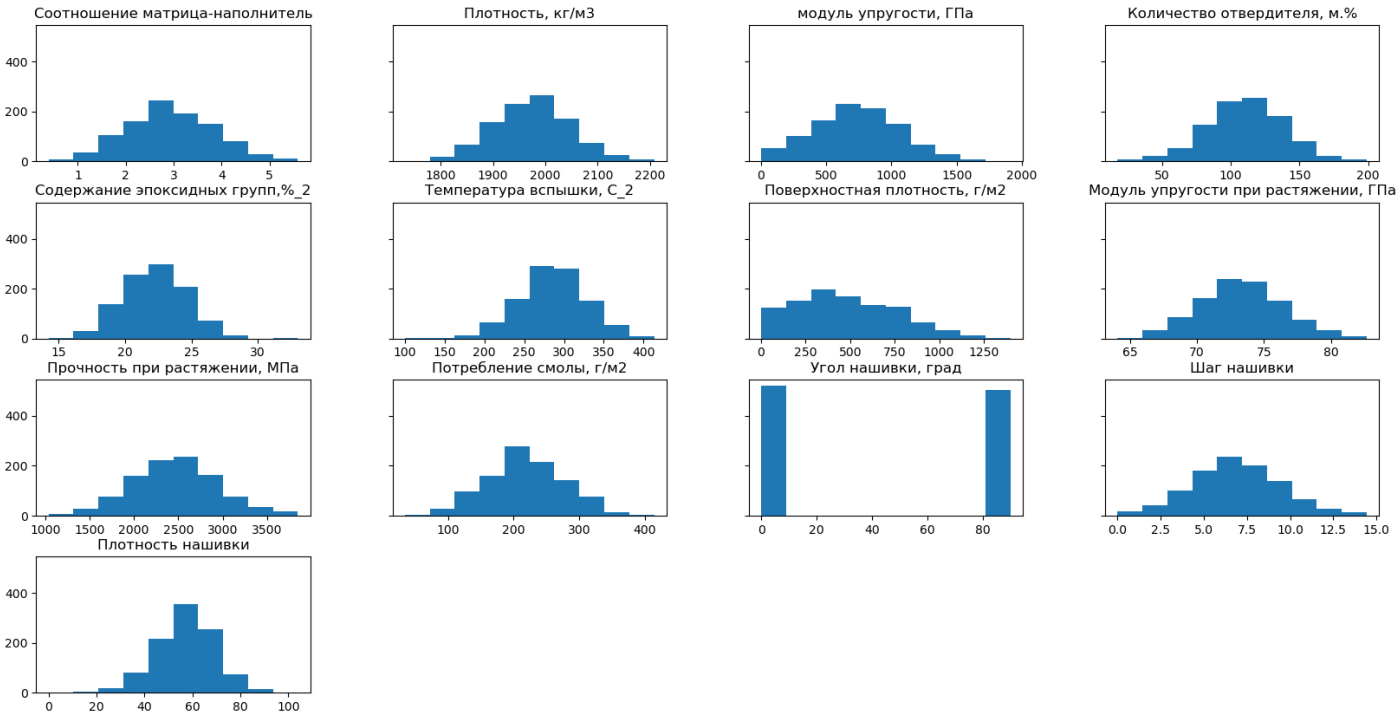

Рисунок 4 – Гистограммы распределения значений

Визуально распределение каждого показателя близко к нормальному (поверхностная плотность чуть смещена), кроме угла нашивки, где содержится всего два значения (рис. 5).

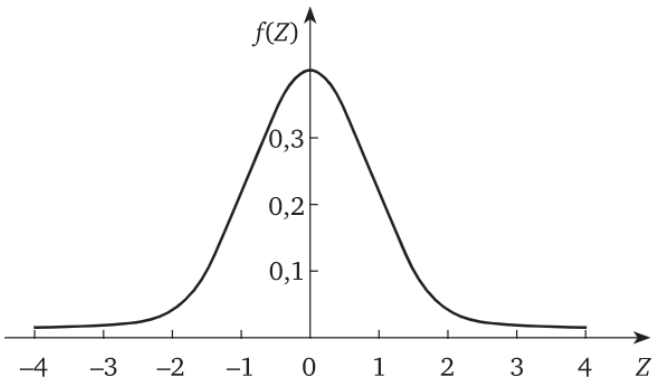

Рисунок 5 – Нормальная функция распределения

Для корректной работы большинства моделей желательна сильная зависимость выходных переменных от входных и отсутствие зависимости между входными переменными.

Построим парные графики рассеяния, позволяющие понять попарные отношения между различными переменными в наборе данных (рис. 6).

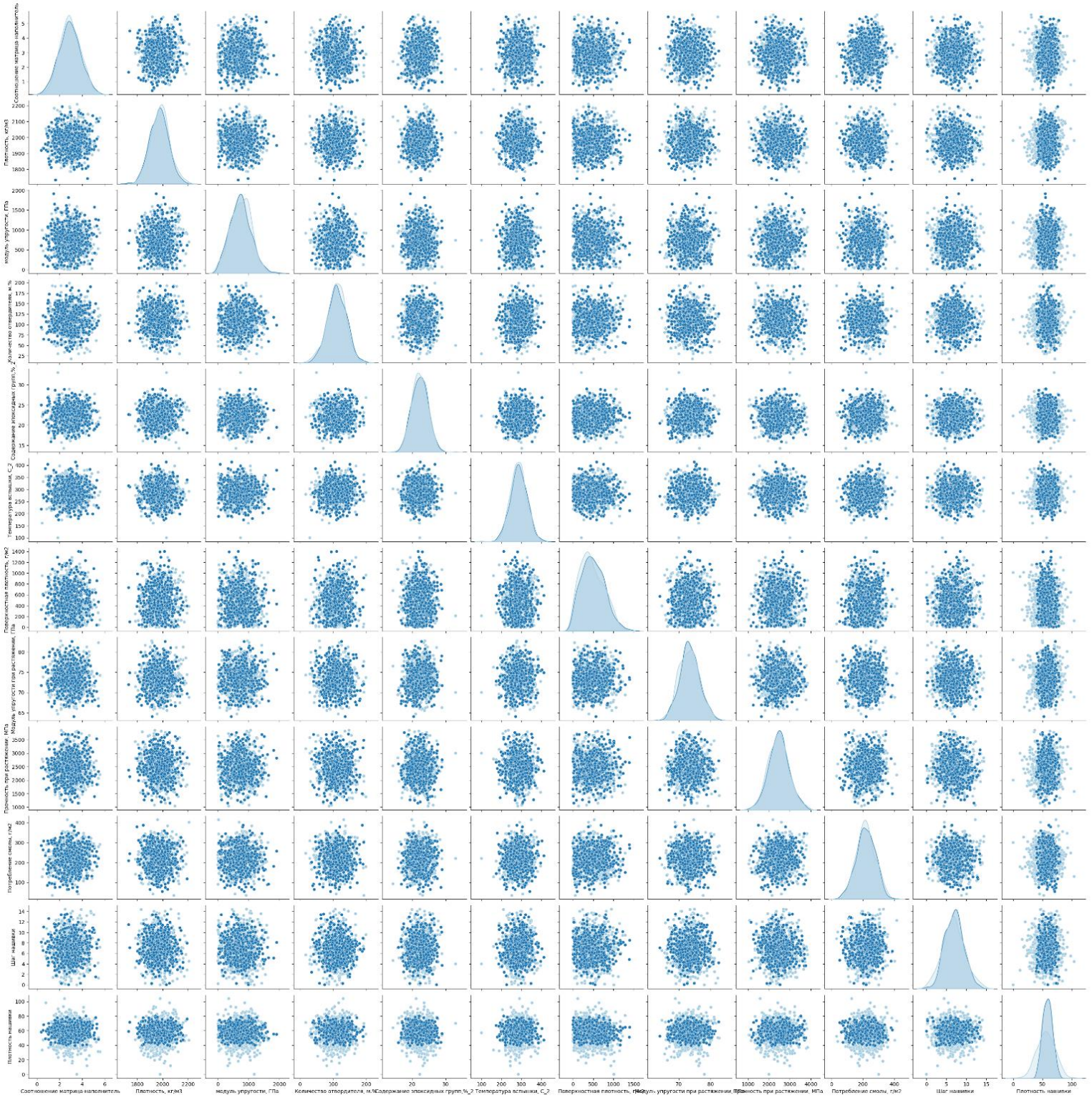

Рисунок 6 – Диаграмма попарного рассеяния

Наличие всего двух вариантов значений параметра «Угол нашивки» условно делает данную переменную категориальной, что можно учесть при построении парных графиков рассеяния. Однако видим, что корреляции между признаками по форме «облаков точек» в основном не наблюдается, за

исключением чуть меньшей дисперсии значений Плотности нашивки при 90 градусов Угла нашивки по сравнению со значением 0 градусов.

Отсутствие связи видно также и на тепловой карте (рис. 7). По матрице корреляции мы очевидно, что все коэффициенты близки к нулю, что означает отсутствие линейной зависимости между признаками.

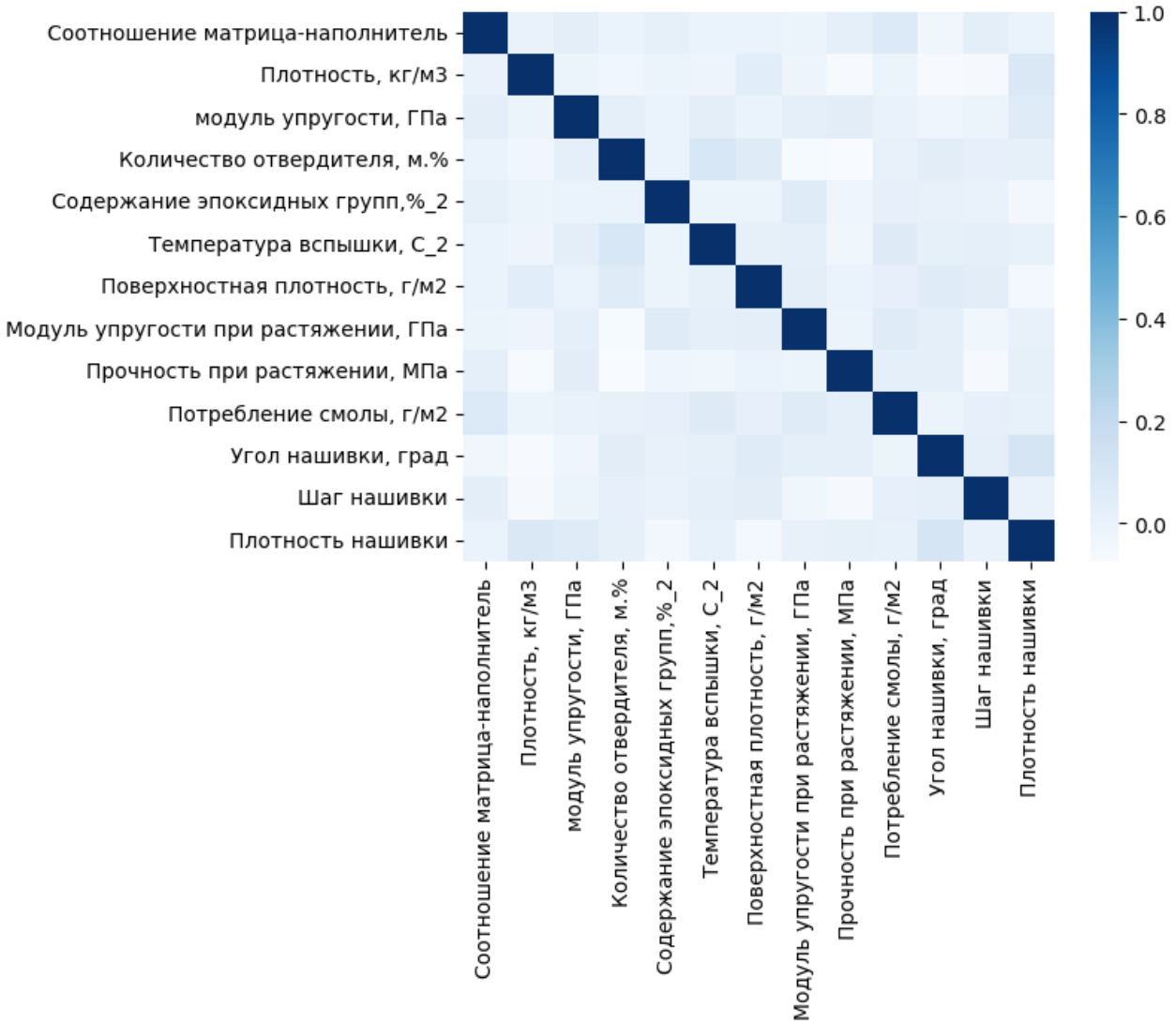

Рисунок 7 – Тепловая карта корреляции

Диаграммы размаха Boxplot являются удобным способом визуального представления групп числовых данных через квартили. Прямые линии, исходящие из ящика, называются «усами» и используются для обозначения степени разброса (дисперсии) за пределами верхнего и нижнего квартилей.

Выбросы иногда отображаются в виде отдельных точек, находящихся на одной линии с усами.

С помощью Boxplot можно увидеть, каковы ключевые значения, (средний показатель, медиана 25-го перцентиля и т. д.), существуют ли выбросы и каковы их значения, симметричны ли данные и насколько плотно они сгруппированы, смещены ли данные и в каком направлении (рис. 8).

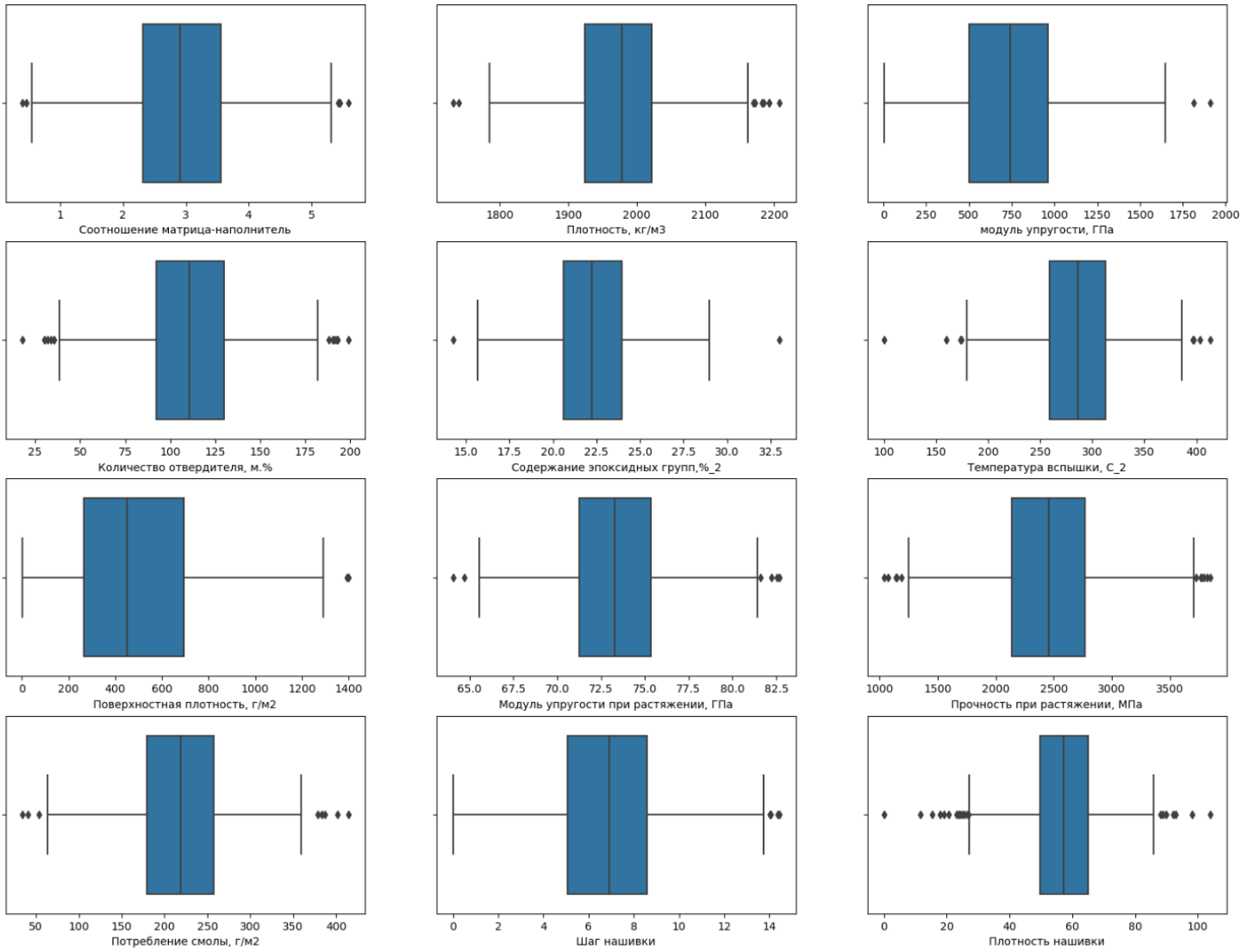

Рисунок 8 – Боксплоты для каждого признака

В нашем примере мы снова видим некоторое смещение данных по показателю Поверхностная плотность, а также большое количество выбросов практически у каждого параметра

Наличие выбросов может исказить предсказательную способность моделей, поэтому их следует обработать. Удаление или сохранение выбросов, в основном, зависит от трех факторов:

1) Область анализа и вопрос исследования. В некоторых областях обычно удаляют посторонние значения, поскольку они часто возникают из-за сбоев в процессе. В других областях отклонения сохраняются, потому что они содержат ценную информацию.

2) Устойчивость тестов. Например, наклон простой линейной регрессии может значительно варьироваться даже с одним выбросом, тогда как непараметрические тесты обычно устойчивы к ним.

3) Дальность выбросов от других наблюдений. Некоторые наблюдения, рассматриваемые как выбросы, на самом деле не являются экстремальными значениями по сравнению со всеми другими наблюдениями, в то время как другие потенциальные выбросы могут быть действительно отстающими от остальных наблюдений.

На рис. 9 видим, что значения признаков лежат в разных диапазонах.

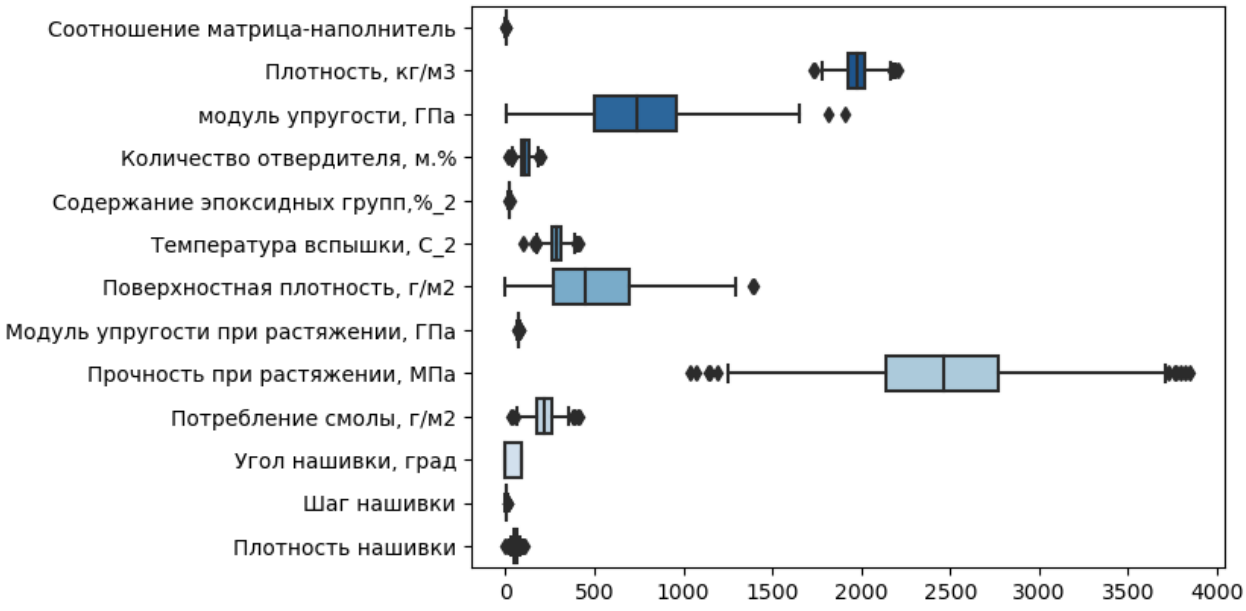

Рисунок 9 – Боксплоты до предобработки

2 Практическая часть

2.1 Предобработка данных

Предварительная обработка данных является важным шагом в процессе интеллектуального анализа данных. Очистка данных используется для обнаружения, исправления или удаления ошибочных записей в наборе данных. Нормализация данных используется для стандартизации диапазона значений независимых переменных или признаков данных.

Для оценки и очистки выбросов используются методы трех сигм и межквартильных интервалов. Удалим выбросы способом межквартильного расстояния, проведя три итерации для полного избавления. В итоге в датасете остаются 922 значения (табл. 2).

Таблица 2 – Описательная статистика признаков после очистки

	count	mean	std	min	max
Соотношение матрица-наполнитель	922.0	2.927964	0.895472	0.547391	5.314144
Плотность, кг/м3	922.0	1974.118744	71.040648	1784.482245	2161.565216
модуль упругости, ГПа	922.0	736.119982	327.607008	2.436909	1628.000000
Количество отвердителя, м.%	922.0	111.136066	26.753228	38.668500	181.828448
Содержание эпоксидных групп,%_2	922.0	22.200570	2.393926	15.695894	28.955094
Температура вспышки, С_2	922.0	286.181128	39.420764	179.374391	386.067992
Поверхностная плотность, г/м2	922.0	482.429070	280.437329	0.603740	1291.340115
Модуль упругости при растяжении, ГПа	922.0	73.303464	3.025864	65.793845	81.203147
Прочность при растяжении, МПа	922.0	2461.491315	453.564734	1250.392802	3654.434359
Потребление смолы, г/м2	922.0	218.048059	57.137475	72.530873	359.052220
Угол нашивки, град	922.0	45.976139	45.013829	0.000000	90.000000
Шаг нашивки	922.0	6.931939	2.514184	0.037639	13.732404
Плотность нашивки	922.0	57.562887	11.122204	28.661632	86.012427

Диаграммы Boxplot демонстрируют отсутствие каких-либо выбросов (рис. 10).

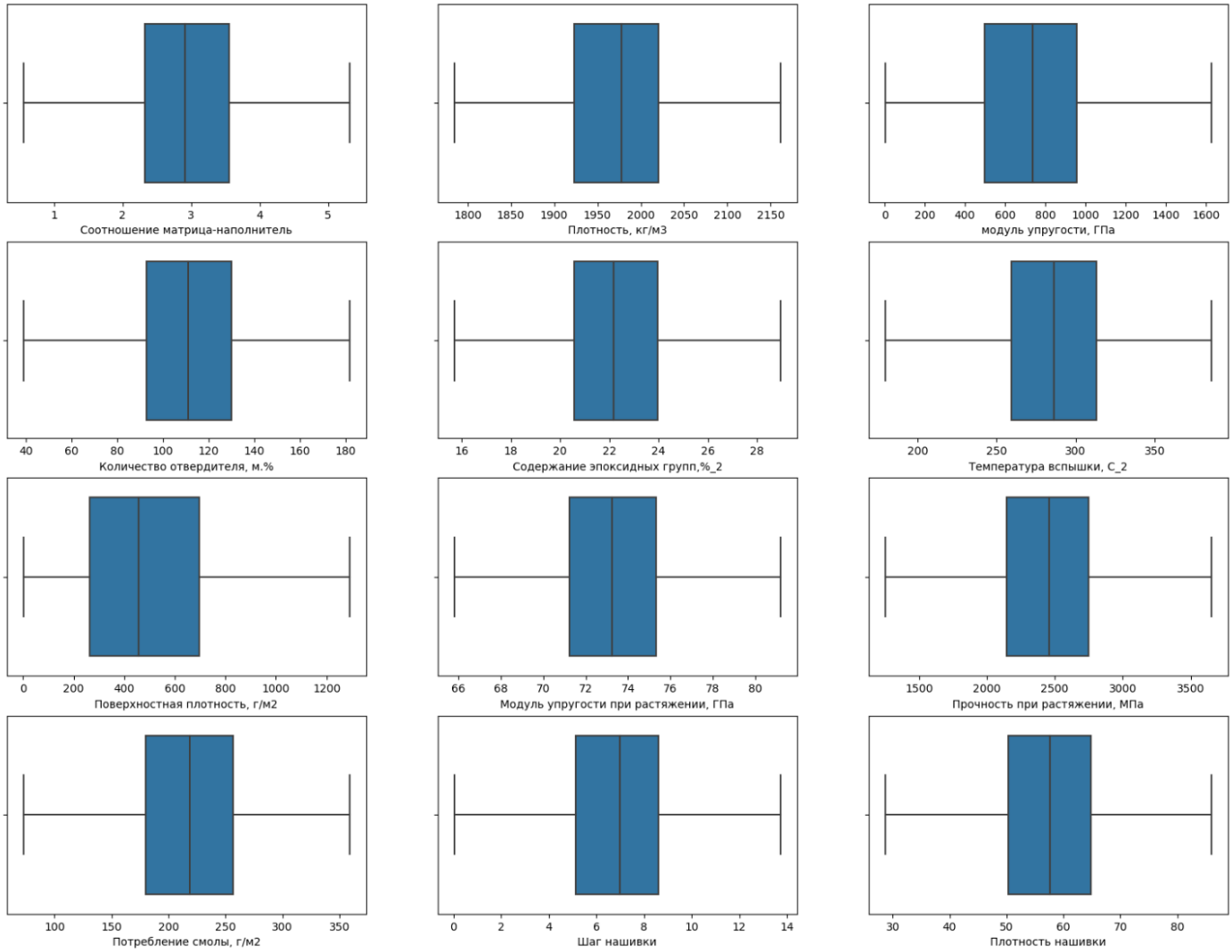

Рисунок 10 – Боксплоты для каждого признака после очистки

Важной частью препроцессинга является приведение различных данных в самых разных единицах измерения и диапазонах значений к единому виду, который позволит сравнивать их между собой. Кроме того, условно категориальный показатель «Угол нашивки» путем нормализации примет значения 0 и 1 без проведения дополнительной обработки. Для нормализации датасета был применён метод MinMaxScaler.

Как итог, получаем обработанный датасет со значениями в диапазоне от 0 до 1 (табл. 3).

Таблица 3 – Описательная статистика признаков после нормализации

	count	mean	std	min	max
Соотношение матрица-наполнитель	922.0	0.499412	0.187858	0.0	1.0
Плотность, кг/м3	922.0	0.502904	0.188395	0.0	1.0
модуль упругости, ГПа	922.0	0.451341	0.201534	0.0	1.0
Количество отвердителя, м.%	922.0	0.506200	0.186876	0.0	1.0
Содержание эпоксидных групп,%_2	922.0	0.490578	0.180548	0.0	1.0
Температура вспышки, С_2	922.0	0.516739	0.190721	0.0	1.0
Поверхностная плотность, г/м2	922.0	0.373295	0.217269	0.0	1.0
Модуль упругости при растяжении, ГПа	922.0	0.487343	0.196366	0.0	1.0
Прочность при растяжении, МПа	922.0	0.503776	0.188668	0.0	1.0
Потребление смолы, г/м2	922.0	0.507876	0.199418	0.0	1.0
Угол нашивки, град	922.0	0.510846	0.500154	0.0	1.0
Шаг нашивки	922.0	0.503426	0.183587	0.0	1.0
Плотность нашивки	922.0	0.503938	0.193933	0.0	1.0

Визуально также заметно уравнивание масштабов разброса значений (рис. 11)

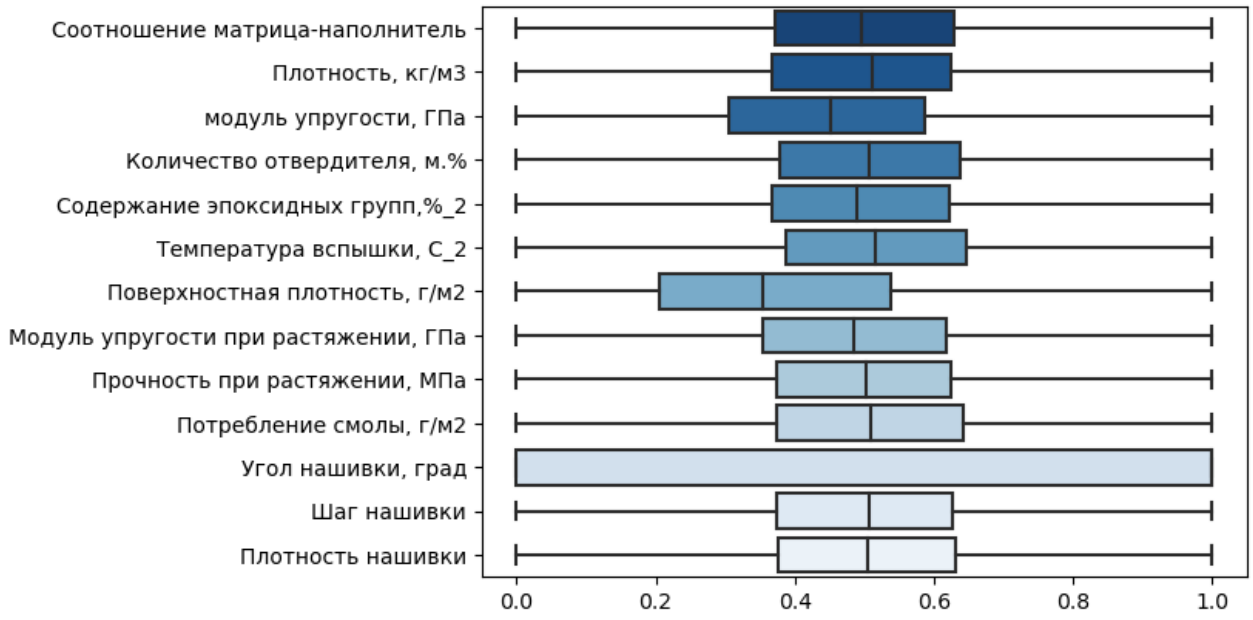

Рисунок 11 – Боксплоты после предобработки

2.2 Разработка и обучение модели

Для построения прогноза модуля упругости при растяжении и прочности при растяжении применены описанные выше методы с использованием библиотеки Scicit-learn.

Порядок разработки модели для каждого параметра и для каждого выбранного метода можно разделить на следующие этапы:

1) разделение нормализованных данных на обучающую и тестовую выборки (в соотношении 70 на 30%);

2) обучение моделей на нормализованных значениях;

3) сравнение моделей по метрикам;

4) поиск гиперпараметров, по которым будет происходить оптимизация модели, с помощью выбора по сетке и перекрёстной проверки;

5) оценка полученных результатов работы моделей.

2.3 Тестирование модели

В качестве метрик для оценки моделей взяты средняя абсолютная ошибка (MAE), среднеквадратическая ошибка (MSE) и коэффициент детерминации (R2). Ниже приведены сводные таблицы результатов тестирования моделей (табл. 4 и 5).

Таблица 4 – Сравнение ошибок моделей модуля упругости при растяжении

	Модель регрессии	MAE	MSE	R2
0	Linear	2.6109355505901184	10.152935631185066	-0.024476898005094228
1	Ridge	2.5998136023821914	10.084913812910749	-0.017613190412159918
2	Lasso	2.5948108078737024	10.09572202309215	-0.018703787461314736
3	ElasticNet	2.5801929025662833	9.998351572705044	-0.00887867080603466
4	GradientBoosting	2.6998404476642666	11.112730470692213	-0.12132451682377177
5	KNeighbors	2.6225621597024773	10.325154738425082	-0.04185458098976058
6	DecisionTree	2.6342990118377716	10.509499850482133	-0.06045583243299113
7	RandomForest	2.576793337919035	10.024106937432755	-0.011477503017891744
8	AdaBoost	2.6353405208350638	10.432782936591103	-0.052714750560453716
9	NeuralNetwork	2.580495886956309	9.966996628786234	-0.005714815853147925

Таблица 5 – Сравнение ошибок моделей прочности при растяжении

	Модель регрессии	MAE	MSE	R2
0	Linear	372.05238450611114	214801.90979576955	-0.003529768950333745
1	Ridge	369.8832593534413	213955.87645116603	0.0004228013393881014
2	Lasso	368.9355592328838	213322.56630912074	0.0033815533408245724
3	ElasticNet	368.8669394882762	214058.58303888745	-5.703200261408803e-05
4	GradientBoosting	382.86326924208987	225302.159467887	-0.05258572537734851
5	KNeighbors	400.581653879789	250632.24504259886	-0.1709249661618053
6	DecisionTree	400.581653879789	250632.24504259886	-0.1709249661618053
7	RandomForest	368.94893692224065	213550.7865727877	0.0023153345689093108
8	AdaBoost	380.4897225346174	224358.12034276538	-0.04817528337516097
9	NeuralNetwork	1213.5428445429625	1686732.611086863	-6.880220381161292

Как видим, все модели показали неудовлетворительный результат предсказания каждого из целевых показателей. Если выбирать лучших из худших, то имеет смысл обратить внимание на регрессию ElasticNet и искусственную нейронную сеть для прогноза модуля упругости при растяжении, а для прогноза прочности при растяжении – случайный лес и регрессии Ridge и Lasso.

2.4 Написание нейронной сети

Для рекомендации соотношения матрица-накопитель построены несколько вариантов искусственной нейронной сети с использованием библиотеки TensorFlow.

Создаем архитектуру нейронной сети и запускаем обучение. Оценивая результаты, меняем параметры нейросети: количество нейронов, функции активации, количество слоев, добавление слоя Dropout.

Так же, как и с прогнозами модуля упругости при растяжении и прочности при растяжении, ни одна из моделей нейросети не показала достаточного результата (табл. 6).

Таблица 6 – Сравнение ошибок моделей ИНС

	Версия нейросети	MAE	MSE	R2
0	Нейросеть 1	2.386060661491529	6.45774411548042	-7.4474735694538445
1	Нейросеть 2	0.7137162029574031	0.7679923222381272	-0.004622469957904052
2	Нейросеть 3	0.7104158222844613	0.7644645654075289	-7.757443281297682e-06
3	Нейросеть 4	0.7143640262384645	0.775958358616658	-0.015042963640772733
4	Нейросеть 5	0.7116793262948142	0.7653706815770234	-0.001193061824467545
5	Нейросеть 6	0.7131466493755869	0.7667935002491898	-0.003054272630024002
6	Нейросеть 7	1.19195047989152	1.9860870722335884	-1.598030790521507
7	Нейросеть 8	0.7101724567930656	0.7637227176227177	0.0009626649585182667
8	Нейросеть 9	0.711454842236497	0.7661455688434341	-0.0022067036760575753
9	Нейросеть 10	0.7133339033046622	0.7730611124592092	-0.011253031754097309

Наименьшие средняя абсолютная и среднеквадратическая ошибки, а также положительный коэффициент детерминации оказались у восьмой модели нейросети со следующей архитектурой (рис. 12):

```
# модель ИНС 8
ANN_8 = Sequential()
ANN_8.add(Dense(100, input_dim=12, activation='sigmoid'))
ANN_8.add(Dropout(0.5))
ANN_8.add(LeakyReLU(alpha=1.0))
ANN_8.add(Dense(50, activation='sigmoid'))
ANN_8.add(LeakyReLU(alpha=1.0))
ANN_8.add(Dense(25, activation='softmax'))
ANN_8.add(Dense(1, activation='linear'))
ANN_8.compile(optimizer='adam', loss='mse', metrics=['mae'])
ANN_8.summary()
history = ANN_8.fit(X_train_norm,y_train_norm, epochs=20, validation_split=0.1, verbose=2)
plot_ANN(history)
```

Рисунок 12 – Архитектура выбранной ИНС

2.5 Разработка приложения

Несмотря на то, что пригодных к внедрению моделей получить не удалось, пайплайн требует разработать функционал приложения.

Веб-приложение для предсказания соотношения матрица-наполнитель разработано с помощью фреймворка Flask, реализованы функции ввода и проверки входных параметров, получение и отображение прогноза выходных параметров.

Скриншот приложения приведен на рисунке 13.

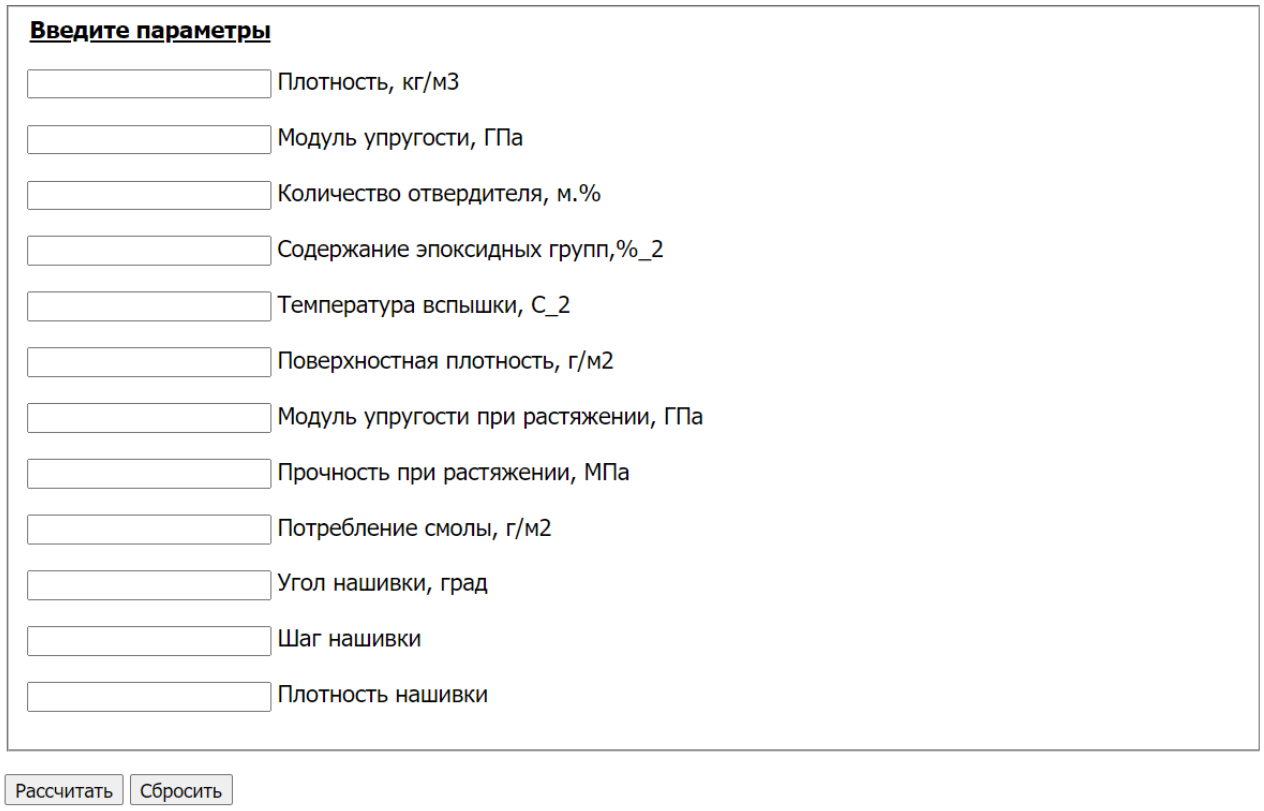

Рисунок 13 – Интерфейс приложения

2.6 Создание удаленного репозитория

Удаленный репозиторий создан на GitHub.com по следующему адресу: https://github.com/Ann-Dreamer/BMSTU_2023

Заключение

Композиционные материалы все более востребованы в различных сегментах рынка, поэтому высока актуальность решения задач прогнозирования характеристик композита на основе известных характеристик исходных компонентов.

Результаты проведения настоящей работы позволяют сделать некоторые основные выводы по теме. Распределение полученных данных в объединённом датасете близко к нормальному, корреляция между парами признаков близка к нулю. Использованные при разработке моделей подходы не позволили получить сколь-нибудь достоверных прогнозов.

Данный факт не указывает на то, что прогнозирование характеристик композитных материалов на основании предоставленного набора данных невозможно, но может указывать на недостатки базы данных, подходов, использованных при прогнозе, необходимости пересмотра инструментов для прогнозирования.

Необходимы дополнительные вводные данные, получение новых результирующих признаков в результате математических преобразований, консультации экспертов предметной области, новые исследования. В целом прогнозирование конечных свойств/характеристик композитных материалов без изучения материаловедения, погружения в вопрос экспериментального анализа характеристик композитных материалов не демонстрирует удовлетворительных результатов. Проработка моделей и построение прогнозов требует внедрения в процесс производных от имеющихся показателей для выявления иного уровня взаимосвязей.

Библиографический список

1. Берикашвили, В. Ш. Статистическая обработка данных, планирование эксперимента и случайные процессы: учебное пособие для вузов / Берикашвили В. Ш., Оськин С. П. — 2-е изд., испр. и доп. — М.: Юрайт, 2021. — 163 с.

2. Грас Д. Data Science. Наука о данных с нуля: Пер. с англ. — 2-е изд., перераб. и доп. — СПб.: БХВ-Петербург, 2021. — 416 с.: ил.9.

3. 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.: ил. — Парал. тит. англ.

4. Первушин, Ю.С., Жернаков, В.С. Проектирование и прогнозирование механических свойств однонаправленного слоя из композиционного материала: Учебное пособие / Ю. С. Первушин, В. С. Жернаков; Уфимск. гос. авиац. техн. ун-т. — Уфа, 2002. — 127 с.

5. Сидняев, Н. И. Нейросети и нейроматематика: учебное пособие / Н. И. Сидняев, П. В. Храпов; под ред. Н. И. Сидняева. — Москва: Издательство МГТУ им. Н. Э. Баумана, 2016. — 83, [3] с.: ил.

6. Документация по библиотеке keras [Электронный ресурс]: — Режим доступа: https://keras.io/api/. (дата обращения: 18.04.2023).

7. Документация по библиотеке matplotlib [Электронный ресурс]: — Режим доступа: https://matplotlib.org/stable/users/index/. (дата обращения: 18.04.2023).

8. Документация по библиотеке numpy [Электронный ресурс]: — Режим доступа: https://keras.io/api/. (дата обращения: 18.04.2023).

9. Документация по библиотеке pandas [Электронный ресурс]: — Режим доступа: https://pandas.pydata.org/docs/user_guide/index.html/. (дата обращения: 18.04.2023).

10. Документация по библиотеке scikit-learn [Электронный ресурс]: — Режим доступа: https://scikit-learn.org/stable/user_guide.html/. (дата обращения: 18.04.2023).

11. Документация по библиотеке seaborn [Электронный ресурс]: — Режим доступа: https://seaborn.pydata.org/tutorial.html/. (дата обращения: 18.04.2023).

12. Документация по библиотеке tensorflow [Электронный ресурс]: — Режим доступа: https://www.tensorflow.org/tutorials/. (дата обращения: 18.04.2023).

13. Документация по фреймворку flask [Электронный ресурс]: — Режим доступа: https://flask.palletsprojects.com/en/1.1.x/. (дата обращения: 18.04.2023).

14. Документация по языку программирования python [Электронный ресурс]: — Режим доступа: https://docs.python.org/3/. (дата обращения: 18.04.2023).

15. Машинное обучение [Электронный ресурс]: — Режим доступа: https://exponenta.ru/wfml/. (дата обращения: 18.04.2023).

16. Loginom Wiki [Электронный ресурс]: — Режим доступа: https://wiki.loginom.ru/. (дата обращения: 18.04.2023).