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: Певцев Станислав Владимирович

Москва, 2023

Содержание

Введение 3

1. Аналитическая часть 4
   1. Постановка задачи 4
   2. Описание используемых методов. 6
   3. Разведочный анализ данных 16
2. Практическая часть 23
   1. Предобработка данных 23
   2. Разработка и обучение модели. 25
   3. Написать нейронную сеть, которая будет рекомендовать соотношение «матрица-наполнитель» 28
   4. Разработка приложения 31
   5. Создание удалённого репозитория и загрузка. 33
   6. Заключение 34
   7. Список используемой литературы и веб ресурсы 35

Введение

Композиционные материалы представляют собой искусственно созданные материалы, которые состоят из нескольких компонентов с четкой границей между ними. Они обладают уникальными свойствами, которые не характерны для отдельных компонентов, и представляют собой монолитный материал, в котором компоненты неотделимы друг от друга без разрушения всей конструкции.

Структура композитных материалов состоит из матрицы - главного компонента, а также укрепляющих элементов, которые обычно называются наполнителями. Граница или поверхность раздела отделяет матрицу от наполнителя. Наполнитель равномерно распределен в матрице и имеет определенную пространственную ориентацию. Композитные материалы обладают комплексом свойств, которые не свойственны каждому компоненту по отдельности. Эти материалы могут быть модифицированы путем изменения объемной доли армирующих элементов в матричном материале, а также путем варьирования их размеров, формы, ориентации и связи по границе "матрица-наполнитель". Таким образом, возможно широкое регулирование свойств композитных материалов.

Данный подход имеет некоторые ограничения: хотя свойства исходных компонентов могут быть известны, определение свойств композита, состоящего из этих компонентов, представляет значительные трудности. Для преодоления этой проблемы могут быть использованы два пути: физические испытания образцов материалов или прогнозирование свойств. Суть прогнозирования заключается в моделировании представительного элемента объема композита на основе информации о свойствах составляющих компонентов (связующего и армирующего компонента). На входе предоставляются данные о начальных свойствах компонентов композитных материалов, таких как количество связующего, наполнителя, температурный режим отверждения и другие. На выходе требуется прогнозировать ряд конечных свойств получаемых композиционных материалов.

Разработка композитных материалов традиционно является длительным процессом, поскольку невозможно рассчитать конечные свойства композита из свойств отдельных компонентов. Для достижения необходимых характеристик необходимо проводить множество комбинированных тестов, что затрудняет задачу прогнозирования успешного решения и увеличивает затраты на разработку новых материалов. Одним из способов уменьшения временных и финансовых затрат на создание материалов может стать использование системы поддержки производственных решений на основе машинного обучения. Такая система может помочь в определении наиболее эффективных комбинаций компонентов для достижения необходимых свойств композита, что позволит снизить стоимость производства композитных материалов.

* + 1. **Аналитическая часть**
       1. **Постановка задачи**

В данной работе исследуется композит с матрицей из базальтопластика и нашивками из углепластика. Были предоставлены 2 файла: X\_bp.xlsx (с данными о параметрах, состоящий из 1023 строк и 10 столбцов данных) и X\_nup.xlsx (данными нашивок, состоящий из 1040 строк и 3 столбцов данных).

Для разработки моделей, позволяющих прогнозировать модуль упругости при растяжении, прочность при растяжении и соотношение "матрица-наполнитель", необходимо объединить два файла с данными. Объединение должно быть выполнено по типу INNER, что означает, что некоторая информация (17 строк таблицы X\_nup.xlsx) не имеет соответствующих строк в таблице X\_bp.xlsx и будет удалена. После объединения по индексу, общее количество строк в объединенном датасете составило 1023, содержащих 13 признаков. Дальнейшие исследования будут проводиться с использованием полученного объединенного датасета.

Требуется провести разведочный анализ данных, который включает построение гистограмм распределения каждой переменной, диаграмм ящика с усами и попарных графиков рассеяния точек. Кроме того, для каждого столбца необходимо определить среднее и медианное значения, проанализировать и исключить выбросы, а также проверить наличие пропущенных значений.

Перед обучением моделей для прогнозирования модуля упругости и прочности при растяжении необходимо провести предобработку данных, включающую удаление шумов и выбросов, а также нормализацию и стандартизацию.

Также требуется разработать нейронную сеть для рекомендации соотношения матрица-наполнитель и создать приложение с графическим интерфейсом, которое будет выдавать прогноз этого соотношения.

Кроме того, необходимо оценить точность модели на тренировочном и тестовом датасете, создать репозиторий в GitHub и разместить там код исследования, а также оформить файл README.

* + - 1. **Описание используемых методов**

Данная задача относится к категории машинного обучения, которая осуществляется при наличии учителя. Традиционно решается с помощью алгоритмов регрессии. Основной целью обучения с учителем является определение функции потерь и её минимизация. С целью получения наилучшего решения, были исследованы несколько методов, некоторые из которых были применены.:

* линейная регрессия (Linear regression);
* лассо регрессия (Lasso);
* гребневая регрессия (Ridge);
* эластичная регрессия (ElasticNet) ;
* градиентный бустинг(GradientBoostingRegressor);
* К-ближайших соседей (KNeighborsRegressor);
* дерево решений (DecisionTreeRegressor);
* случайный лес (RandomForest);
* градиентный бустинг (AdaBoostRegressor);
* стохастический градиентный спуск (SGDRegressor);
* метод опорных векторов (Support Vector Regression);
* многослойный перцептрон.

**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. 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.

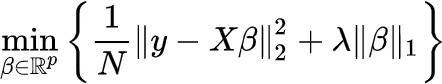
*Недостатки метода:* моделирует только прямые линейные зависимости; требует прямую связь между зависимыми и независимыми переменными; выбросы оказывают огромное влияние, а границы линейны.

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.

***Лассо регрессия (Lasso) –*** 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.

***Гребневая регрессия (Ridge) –*** 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.

https://machinelearningmastery.ru/img/0-789051-512943.png

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.

***Эластичная сеть*** (***ElasticNet) –*** 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.

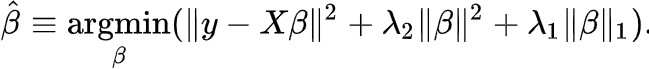
Компромисс смещения дисперсии - это компромисс между сложной и простой моделью, в которой промежуточная сложность, вероятно, является наилучшей.

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.



***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 иприменения; 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.

***Метод ближайших соседей - К-ближайших соседей (kNN - k Nearest Neighbours)*** 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.

***Дерево решений (DecisionTreeRegressor)*** – 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.

***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 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.

***Градиентный бустинг (AdaBoost)* –** это алгоритм, который работает по принципу перевзвешивания результатов. Есть деревья решений, а ансамбль из них это градиентный бустинг, задача решается с помощью градиентсного спуска. 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.

Он итеративно исправляет ошибки слабого классификатора и повышает точность путем объединения слабых учащихся.

Можно использовать многие базовые классификаторы с AdaBoost. AdaBoost не склонен к переоснащению.

*Недостатки метода:*

AdaBoost чувствителен к шумным данным.

AdaBoost обучается дольше линейной регрессии, классификация дольше чем при использовании логистической регрессии.

На AdaBoost сильно влияют отклонения, так как он пытается идеально подогнать каждую точку.

AdaBoost работает медленнее и чуть хуже, чем XGBoost. Но легче в понимании.

***Стохастический градиентный спуск (SGDRegressor)*** — это простой, но очень эффективный подход к подгонке линейных классификаторов и регрессоров под выпуклые функции потерь. Этот подход подразумевает корректировку весов нейронной сети, используя аппроксимацию градиента функционала, вычисленную только на одном случайном обучающем примере из выборки.

*Достоинства метода:* эффективен; прост в реализации; имеет множество возможностей для настройки кода; способен обучаться на избыточно больших выборках.

*Недостатки метода:* требует ряд гиперпараметров; чувствителен к масштабированию функций; может не сходиться или сходиться слишком медленно; функционал многоэкстремален; процесс может "застрять" в одном из локальных минимумов; возможно переобучение.

***Метод опорных векторов (Support Vector Regression)*** – этот бинарный линейный классификатор был выбран, потому что он хорошо работает на небольших датасетах. Данный алгоритм – это алгоритм обучения с учителем, использующихся для задач классификации и регрессионного анализа, это контролируемое обучение моделей с использованием схожих алгоритмов для анализа данных и распознавания шаблонов. Учитывая обучающую выборку, где алгоритм помечает каждый объект, как принадлежащий к одной из двух категорий, строит модель, которая определяет новые наблюдения в одну из Категорий. Модель метода опорных векторов – отображение данных точками в пространстве, так что между наблюдениями отдельных категорий имеется разрыв.

Каждый объект данных представляется как вектор (точка) в p-мерном пространстве. Он создаёт линию или гиперплоскость, которая разделяет данные на классы.

*Достоинства метода:* для классификации достаточно небольшого набора данных. При правильной работе модели, построенной на тестовом множестве, вполне возможно применение данного метода на реальных данных. Эффективен при большом количестве гиперпараметров. Способен обрабатывать случаи, когда гиперпараметров больше, чем количество наблюдений. Существует возможность гибко настраивать разделяющую функцию. Алгоритм максимизирует разделяющую полосу, которая, как подушка безопасности, позволяет уменьшить количество ошибок классификации.

*Недостатки метода:* неустойчивость к шуму, поэтому в работе была проведена тщательнейшая работа с выбросами, иначе в обучающих данных шумы становятся опорными объектами-нарушителями и напрямую влияют на построение разделяющей гиперплоскости; для больших наборов данных требуется долгое время обучения; достаточно сложно подбирать полезные преобразования данных; параметры модели сложно интерпретировать, поэтому были рассмотрены и другие методы.

***Многослойный персептрон (MLPRegressor)*** — это алгоритм обучения с

учителем, который изучает функцию f(⋅):Rm→Ro обучением на наборе данных, где m — количество измерений для ввода и o- количество размеров для вывода.

Это искусственная нейронная сеть, имеющая 3 или более слоёв персептронов. Эти слои - один входной слой, 1 или более скрытых слоёв и один выходной слой персептронов.

*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.

***Используемые метрики качества моделей:***

***R2 (коэффициент детерминации)*** 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.

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

* + - 1. **Разведочный анализ данных**

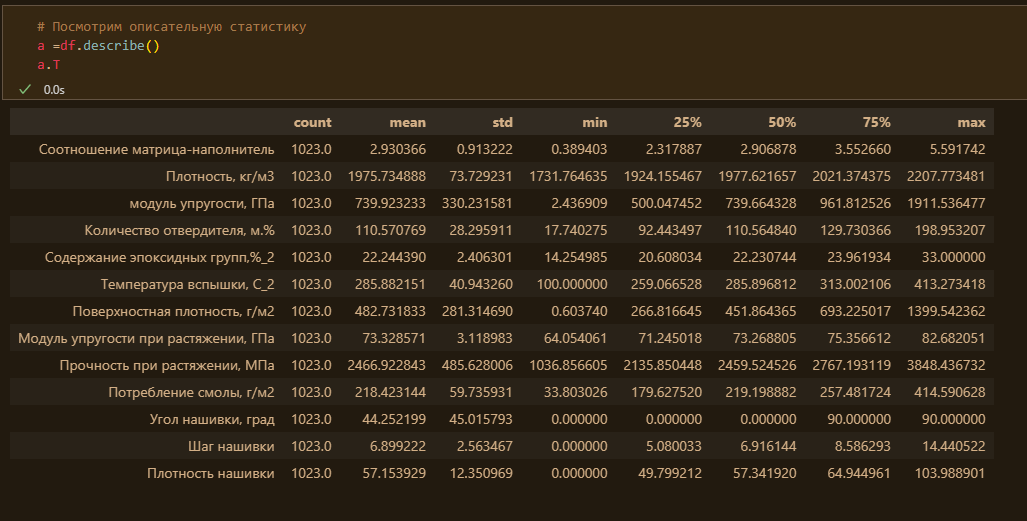
Перед использованием в моделях машинного обучения, данные должны быть обработаны и очищены. Исходные данные могут содержать ошибки и пропущенные значения, что может привести к неверным результатам в процессе моделирования. Однако, произвольное удаление данных также не является правильным подходом. Поэтому перед началом использования набора данных необходимо проанализировать его содержание.

Рисунок 1. Описательная статистика датасета

Цель проведения разведочного анализа заключается в получении первоначальных представлений о характеристиках распределений переменных исходного набора данных, а также в формировании оценки качества исходных данных путем выявления пропусков и выбросов. Кроме того, разведочный анализ позволяет обнаружить характер взаимосвязей между переменными, что является необходимым для последующего выдвижения гипотез о наиболее подходящих моделях машинного обучения для решения задачи.

Для выполнения разведочного анализа данных используются различные инструменты, включая оценку статистических характеристик датасета, гистограммы для отображения распределения каждой из переменных, диаграммы boxplot, попарные графики рассеяния точек, тепловые карты, описательную статистику для каждой переменной, анализ и полное исключение выбросов, проверку наличия пропусков и дубликатов, а также оценку корреляции, используя методы, такие как корреляция Кендалла Пирсона.

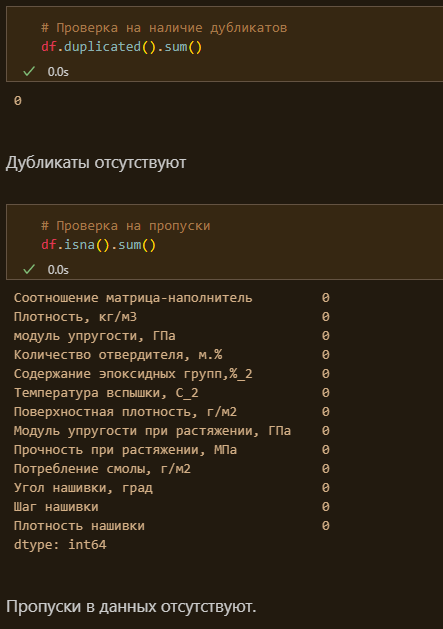


Рисунок 2. Проверка датасета на наличие дубликатов и пропусков

Оценка количества выбросов

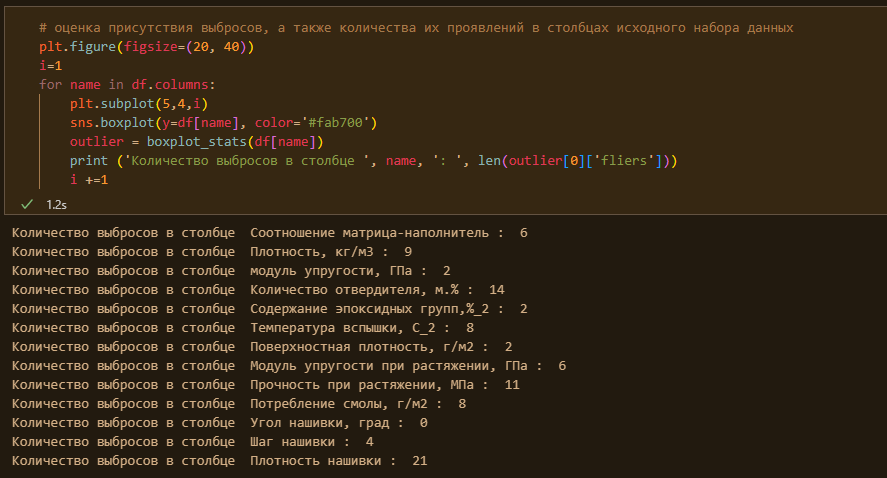
Для определения выбросов используются методы трех сигм и межквартильного расстояния. В данном случае удалим способом межквартильного расстояния.

Рисунок 3. Начальное количество выбросов

Рисунок 4. Начальный boxplot

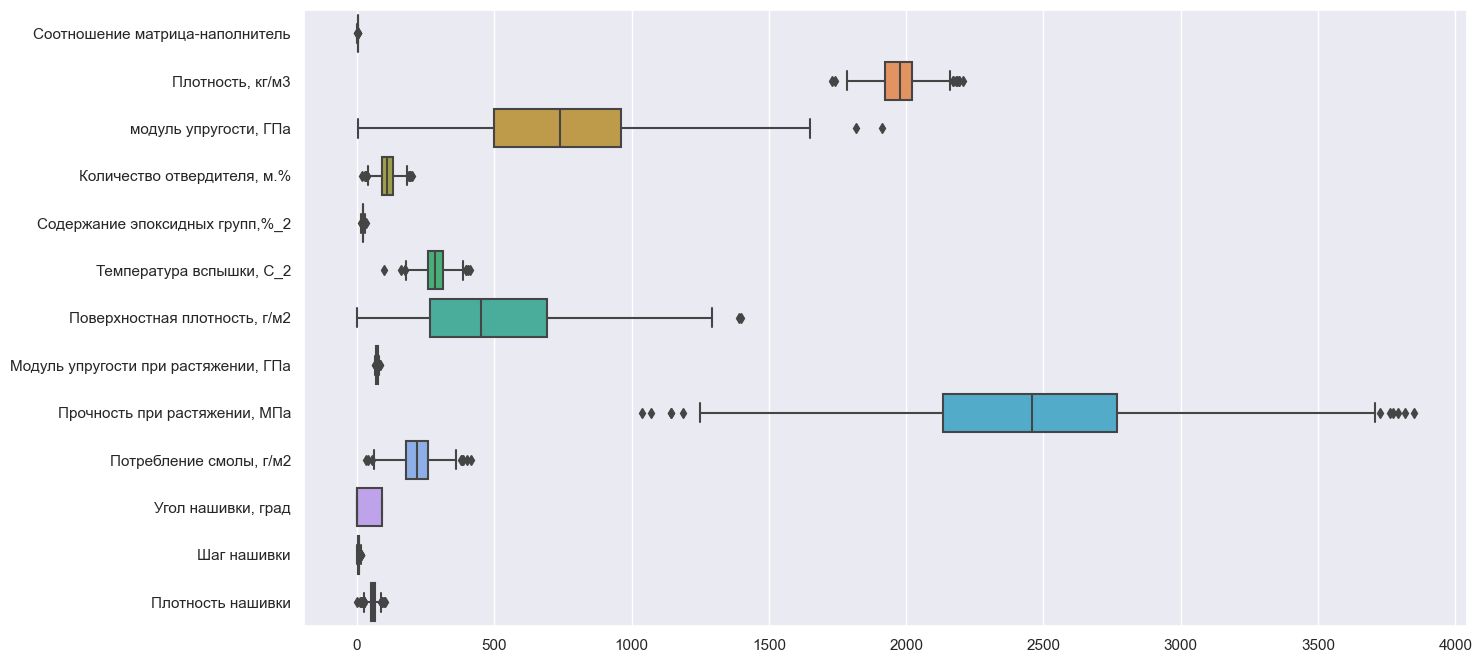
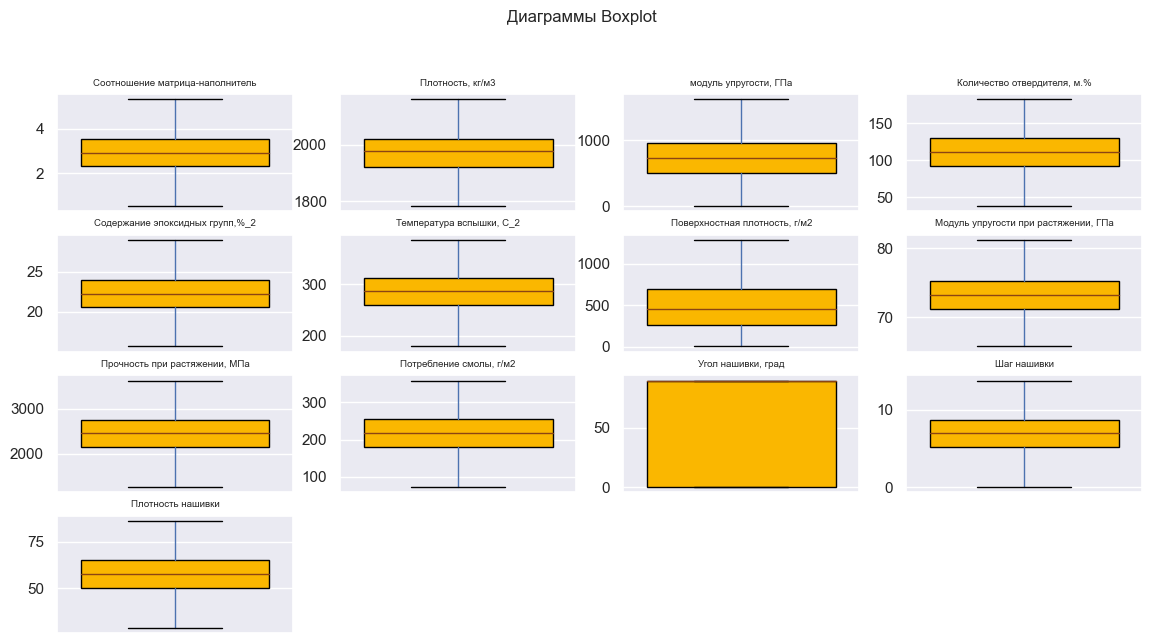


Рисунок 5. Boxplot после удаления выбросов



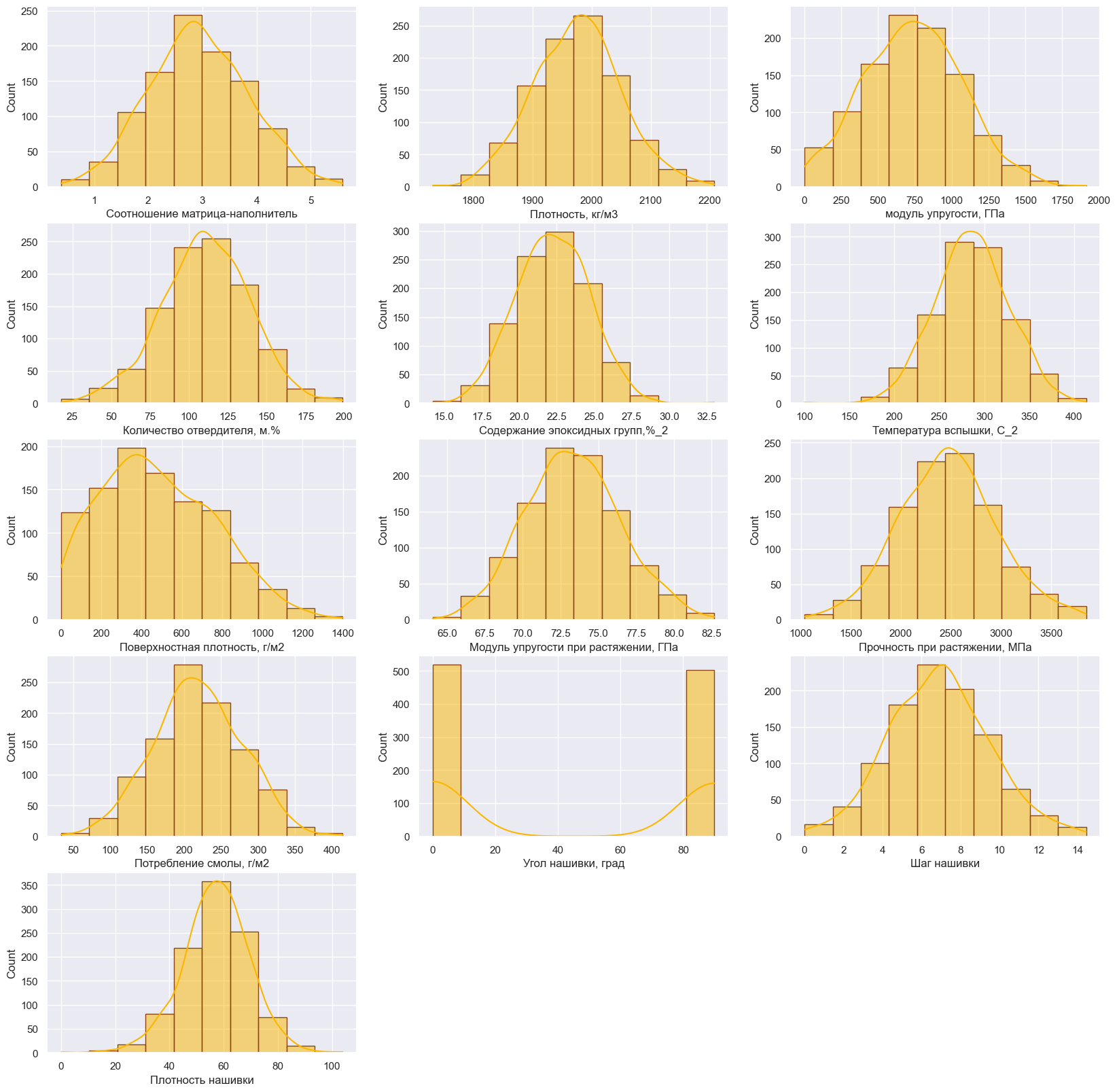


Рисунок 6. Гистограммы распределения

Гистограммы показывают нормальное распределение, за исключением признака «Угол нашивки», который имеет всего два значения 0 и 90 градусов

На попарных графиках распределения не видно корреляции между признаками.

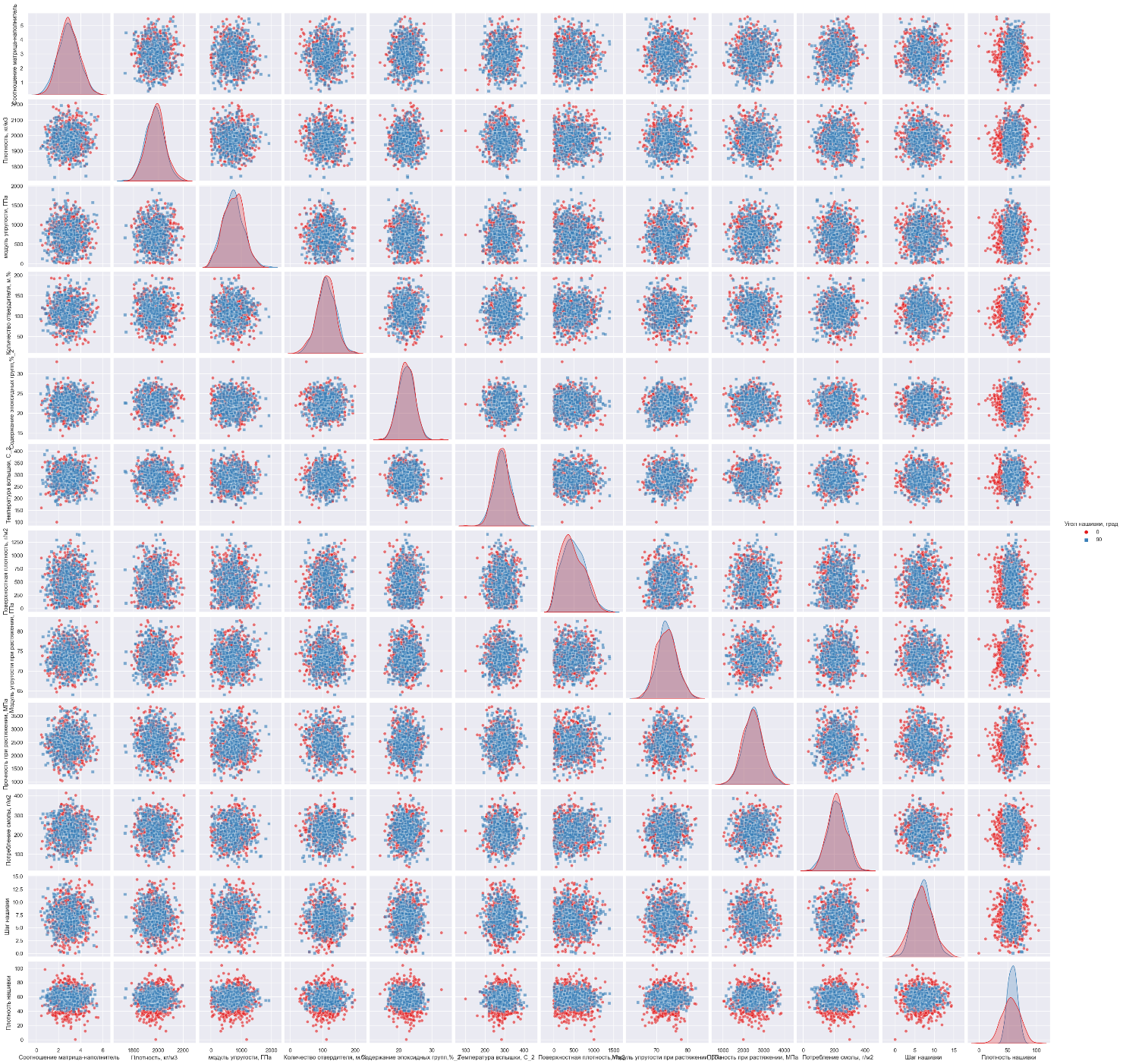


Рисунок 7. Попарные графики рассеяния точек с выделением значений Угол нашивки

Единственная зависимость, которую можно отметить – это меньшая дисперсия значений Плотности нашивки при 90 градусов Угла нашивки, по сравнению со значением 0 градусов.

Тепловая карта показывает практически отсутствие корреляции между признаками и целевыми переменными

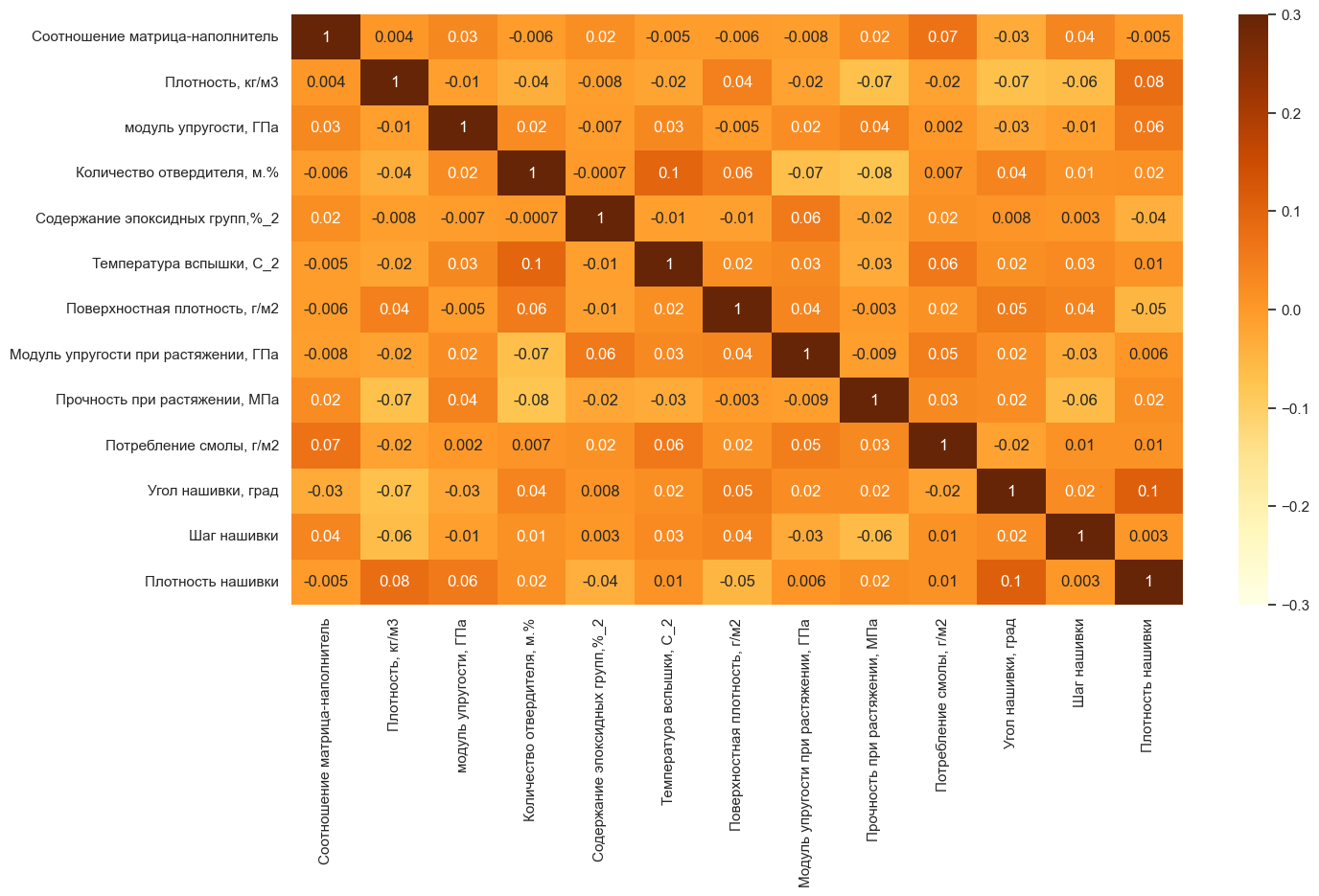


Рисунок 8. Тепловая карта с корреляцией данных

Максимальная корреляция между плотностью нашивки и углом нашивки 0.11, значит нет зависимости между этими данными. Корреляция между всеми параметрами очень близка к 0, корреляционные связи между переменными не наблюдаются.

* + 1. **Практическая часть**
       1. **Предобработка данных**

Стандартизируем датасет с очищенными данными.

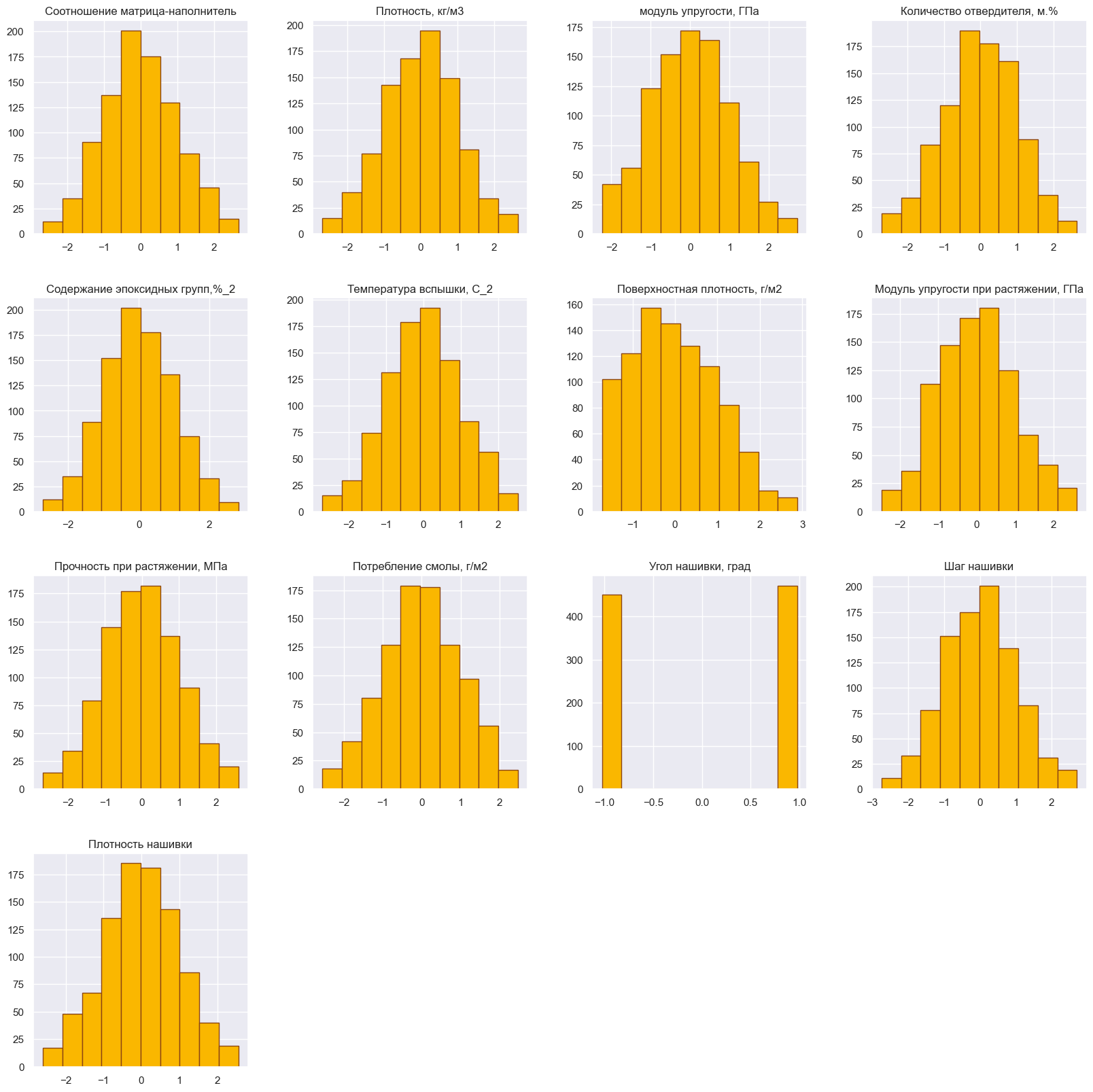


Рисунок 9. Гистограммы после стандартизации

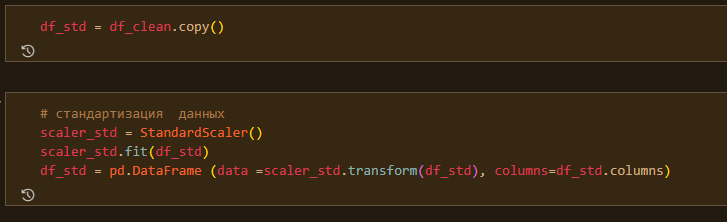


Рисунок 10. Стандартизация датасета

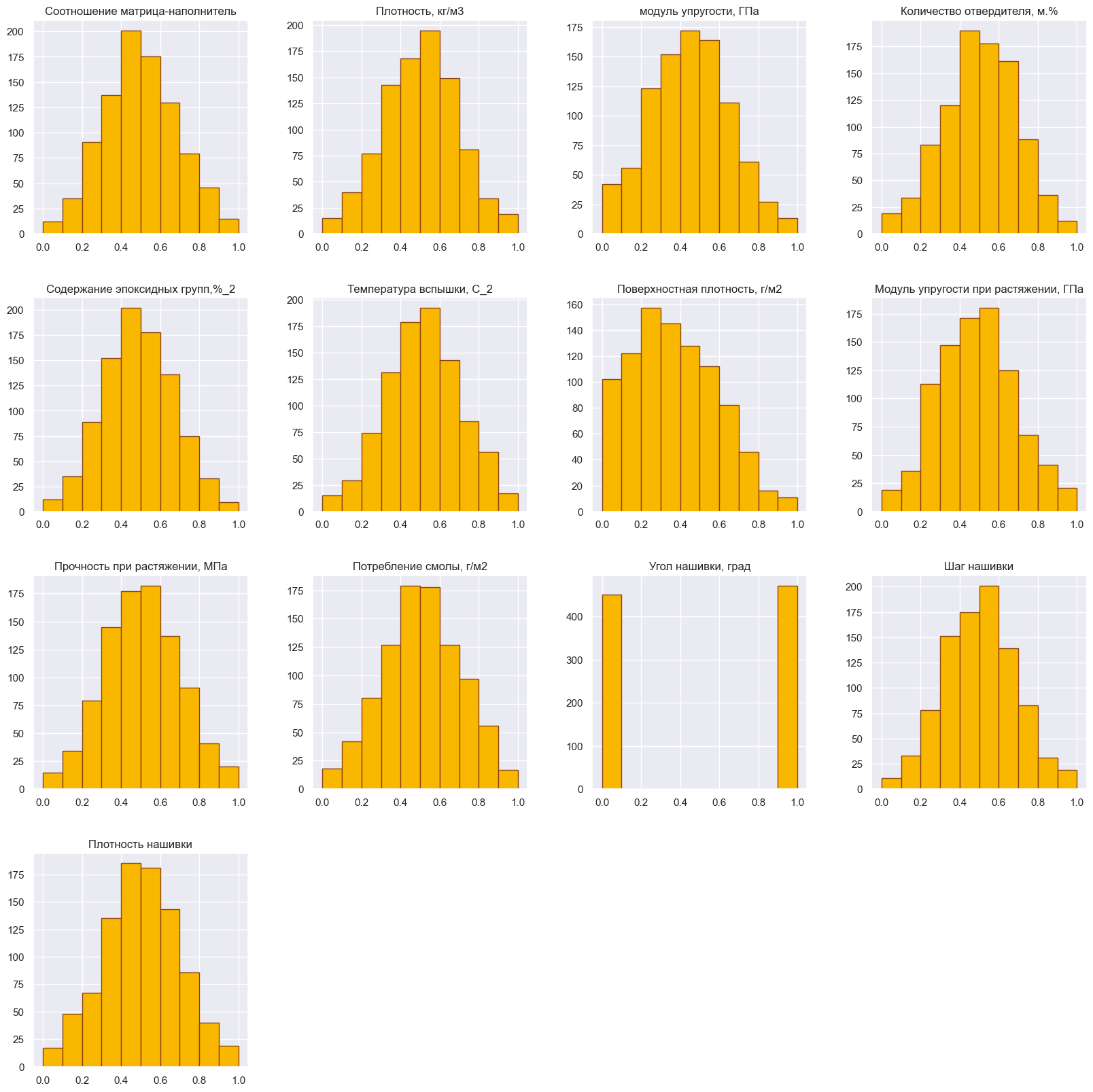
Теперь нормализуем значения, для этого применив MinMaxScaler.

Рисунок 11. Гистограммы после нормализации

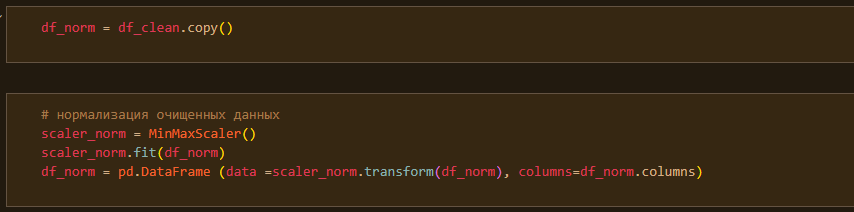


Рисунок 12. Нормализация датасета

* + - 1. **Разработка и обучение модели**

Разработка и обучение моделей машинного обучения осуществлялась для двух выходных параметров: «Прочность при растяжении» и «Модуль упругости при растяжении» отдельно. Для решения применим все методы, описанные выше.

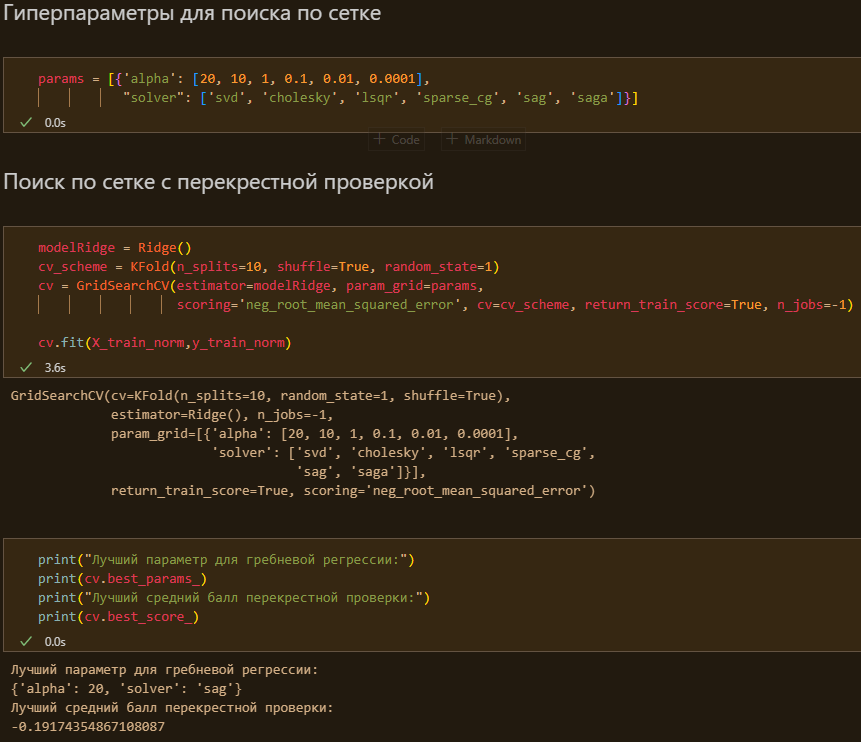


Рисунок 13. Поиск гиперпараметров по сетке

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

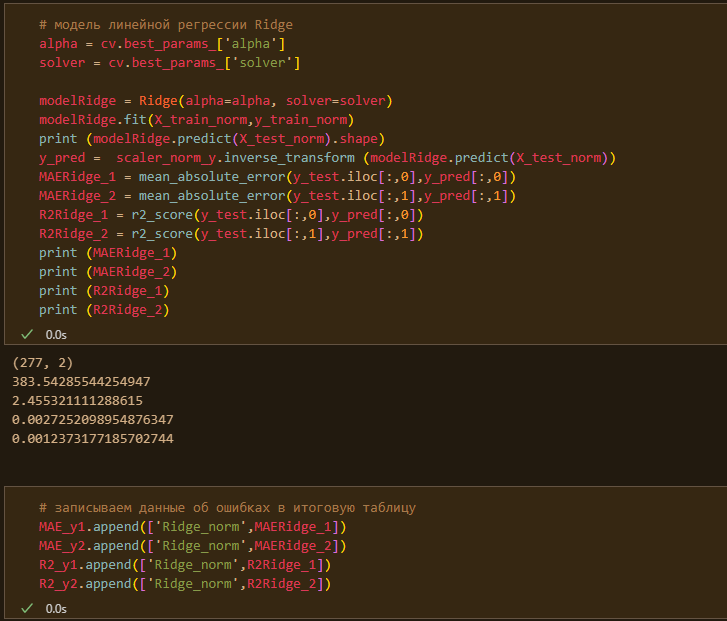
* разделение нормализованных данных на обучающую и тестовую выборки (в соотношении 70 на 30%);
* обучение моделей на нормализованных значениях;
* сравнение моделей по метрике МАЕ;
* поиск гиперпараметров, по которым будет происходить оптимизация модели, с помощью выбора по сетке и перекрёстной проверки;
* Оценка полученных результатов работы моделей. В качестве параметра оценки выбран также коэффициент детерминации (R2).

Рисунок 14. Модель Ridge регрессия

Модели после настройки гиперпараметров показали результат немного лучше. В результате все модели показали примерно одинаковый результат: ошибка MAE примерно равна стандартному отклонению, значения R2 находятся около нуля, то есть все модели предсказывают результат сопоставимый со средним значением. Можно считать, что все примененные модели не справились с задачей, результат неудовлетворительный.

Для улучшения работы алгоритмов можно уменьшить количество признаков. Посмотрим коэффициенты вклада признаков в результат в моделях Lasso и Ridge.

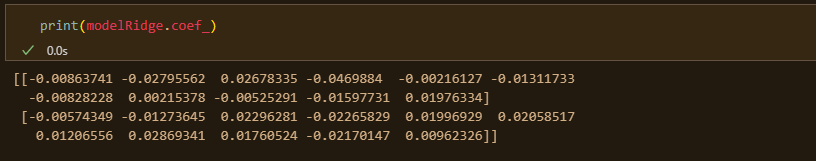


Рисунок 15. Коэффициенты признаков Ridge регрессии

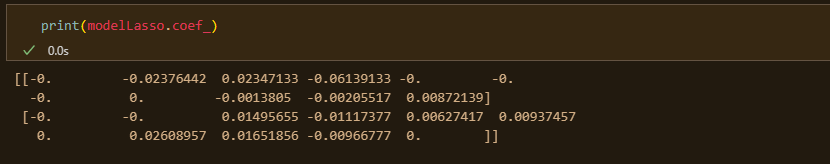


Рисунок 16. Коэффициенты признаков Lasso регрессии

Как видно, коэффициенты вклада в результат близки для всех признаков к нулю. Поэтому нет смысла в уменьшении размерности.

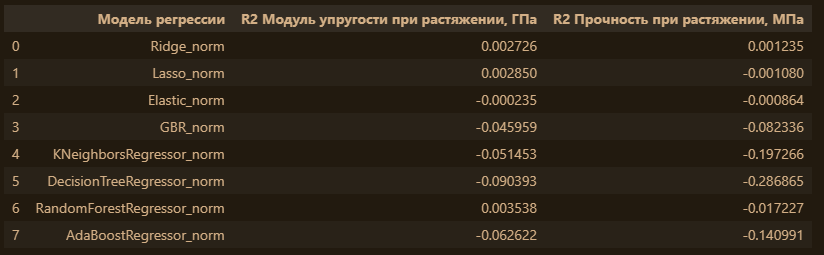


Рисунок 17. Оценка R2 результатов работы моделей



Рисунок 18. Оценка MAE результатов работы моделей

* 1. **Нейронная сеть для рекомендации «Соотношение матрица-наполнитель».**

Загружаем очищенный датасет, для X удаляем целевой столбец

«Соотношение матрица-наполнитель» и сохраняем его в у. Далее разбиваем на обучающую и тестовую выборку (для вычисления MAE). Затем делаем нормализацию, и снова разбиение данных уже для подачи на вход нейросети.

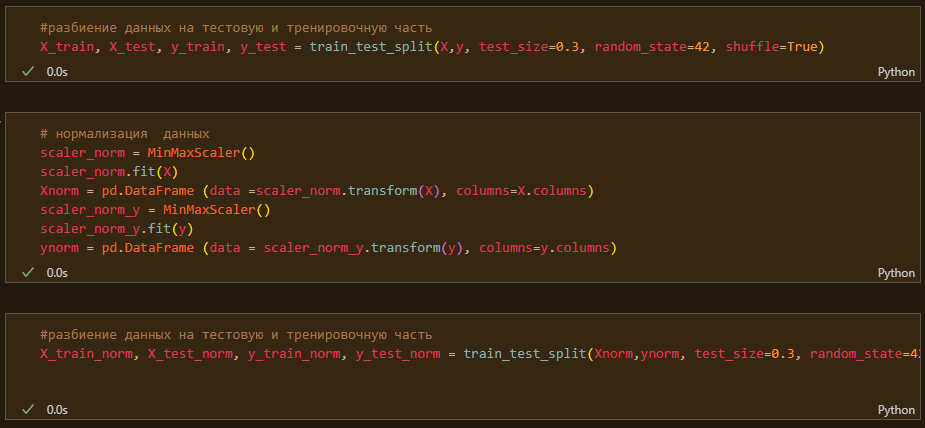


Рисунок 19. Разбиение и нормализация данных для нейросети

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



Рисунок 20. Ошибка MAE по результатам работы нейросетей с различной архитектурой



Рисунок 21. Код нейросети с наименьшей ошибкой



Рисунок 22. Визуализация работы нейросети с наименьшей ошибкой

Все нейросети показали схожий результат с ошибкой MAE чуть меньшей, чем среднее отклонение.

**Разработка приложения**

Создание приложения для расчета параметра «Соотношение матрица- наполнитель». Данное приложение — это основной файл Flask, папка templates, с Шаблоном html - страницы, папка mn\_model\_nn c сохранённой моделью.

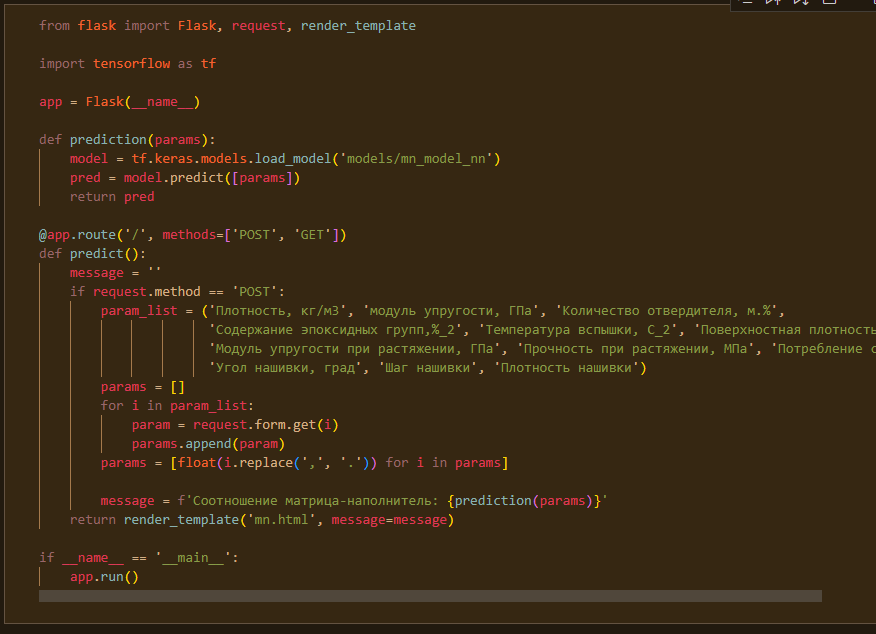


Рисунок 23. Код приложения

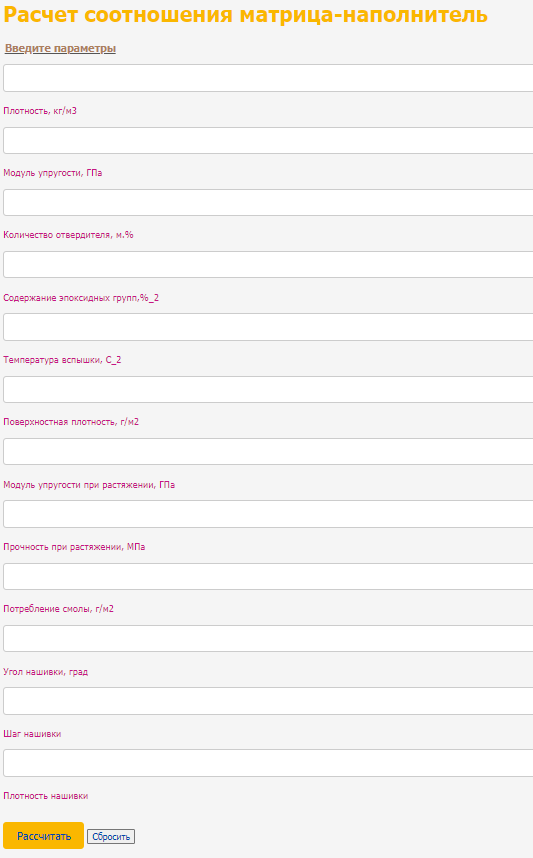


Рисунок 24. Форма пользовательского приложения для ввода параметров

На выходе пользователь получает результат прогноза для значения параметра «Соотношение «матрица – наполнитель»»

* 1. **Создание удалённого репозитория и загрузка**

Репозиторий был создан на github.com по адресу: https://github.com/Prayonis/DS-2022-4.0

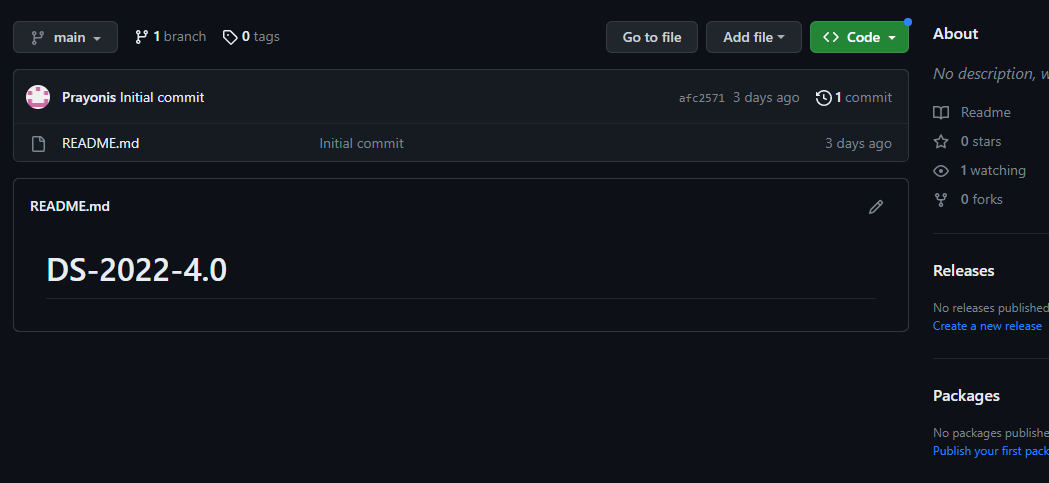


Рисунок 25. Загруженные файлы

* 1. **Заключение**

Представленное исследование дает возможность извлечь основные выводы в связи с темой исследования. Распределение полученных данных в объединенном датасете имеет близкое к нормальному распределение, в то время как корреляция между парами признаков приближается к нулю. Несмотря на использование различных методов в разработке моделей, не удалось получить достоверных прогнозов. Модели регрессии, примененные в исследовании, не показали эффективности в прогнозировании свойств композитов.

Исходя из имеющихся данных, можно сделать вывод о невозможности точного определения соотношения между матрицей и наполнителем композитных материалов исключительно на основе свойств материалов. Однако этот факт не обязательно свидетельствует о невозможности прогнозирования характеристик композитных материалов с помощью имеющихся данных. Вместо этого, он может указывать на недостатки базы данных, использованных подходов и инструментов для прогнозирования, требующих пересмотра и усовершенствования.

Для более точного прогнозирования свойств композитных материалов требуются дополнительные исходные данные, которые могут быть получены с помощью математических преобразований и консультаций экспертов в области материаловедения, а также через проведение новых исследований. Применение методов машинного обучения для прогнозирования свойств композитных материалов без учета материаловедения и экспериментального анализа характеристик композитов не демонстрирует достаточно точных результатов. В связи с отсутствием корреляции между признаками можно сделать вывод, что текущие алгоритмы не решают поставленную задачу, возможно, из-за ее сложности или неразрешимости.

* 1. **Список используемой литературы и веб ресурсы.**

1. Alex Maszański. Метод k-ближайших соседей (k-nearest neighbour) https://proglib.io/p/metod-k-blizhayshih-sosedey-k-nearest-neighbour2021-07-19.
2. Andre Ye. 5 алгоритмов регрессии в машинном обучении, о которых вам сле-дует знать: https://habr.com/ru/company/vk/blog/513842/
3. [Anthony Schams](https://medium.com/%40acschams?source=post_page-----8bf81991d0c5--------------------------------). Смещение, дисперсия и регуляризация в линейной регрессии: лассо, хребет и эластичная сеть - различия и использовании https://machinelearningmastery.ru/bias-variance-and-regularization-in-linear- regression-lasso-ridge-and-elastic-net-8bf81991d0c5/
4. Devpractice Team. Python. Визуализация данных. Matplotlib. Seaborn. Mayavi. - devpractice.ru. 2020. - 412 с.: ил.
5. Абросимов Н.А Методика построения разрешающей системы уравнений динамического деформирования композитных элементов конструкций (Учебно-методическое пособие), ННГУ, 2010
6. Абу-Хасан Махмуд, Масленникова Л. Л. Прогнозирование свойств композиционных материалов с учётом наноразмера частиц и акцепторных свойств катионов твёрдых фаз, статья 2006 год
7. Гафаров, Ф.М., Галимянов А.Ф. Искусственные нейронные сети и приложения: учеб. пособие /Ф.М. Гафаров, А.Ф. Галимянов. – Казань: Издательство Казанского университета, 2018. – 121 с.
8. Грас Д. Data Science. Наука о данных с нуля: Пер. с англ. - 2-е изд., перераб. и доп. - СПб.: БХВ-Петербурr, 2021. - 416 с.: ил.9.
9. Документация по библиотеке keras: <https://keras.io/api/>.
10. Документация по библиотеке matplotlib:

<https://matplotlib.org/stable/users/index.html>.

1. Документация по библиотеке numpy: <https://numpy.org/doc/1.22/user/index.html#user>.
2. Документация по библиотеке pandas: <https://pandas.pydata.org/docs/user_guide/index.html#user-guide>.
3. Документация по библиотеке scikit-learn: <https://scikit-learn.org/stable/user_guide.html>
4. Документация по библиотеке seaborn: <https://seaborn.pydata.org/tutorial.html>.
5. Документация по библиотеке Tensorflow: https://[www.tensorflow.org/overview](http://www.tensorflow.org/overview)
6. Документация по языку программирования python: <https://docs.python.org/3.8/index.html>.
7. Иванов Д.А., Ситников А.И., Шляпин С.Д – Композиционные материалы: учебное пособие для вузов, 2019. 13 с.
8. Краткий обзор алгоритма машинного обучения Метод Опорных Векторов (SVM) : <https://habr.com/ru/post/428503/>
9. Ларин А. А., Способы оценки работоспособности изделий из композиционных материалов методом компьютерной томографии, Москва, 2013, 148 с.
10. [Макс](https://your-scorpion.ru/) Ветков. Градиентный бустинг(AdaBoost) https://your- scorpion.ru/gradient-boosting-adaboost
11. Материалы конференции: V Всероссийская научно-техническая конференция «Полимерные композиционные материалы и производственные технологии нового поколения», 19 ноября 2021 г.
12. Плас Дж. Вандер, Python для сложных задач: наука о данных и машинное обучение. Санкт-Петербург: Питер, 2018, 576 с
13. Реутов Ю.А.: Прогнозирование свойств полимерных композиционных материалов и оценка надёжности изделий из них, Диссертация на соискание учёной степени кандидата физико-математических наук, Томск 2016.
14. Роббинс, Дженнифер. HTML5: карманный справочник, 5-е издание.: Пер. с англ. - М.: ООО «И.Д. Вильямс»: 2015. - 192 с.: ил
15. Руководство по быстрому старту в flask: – Режим доступа: https://flaskrussian-docs.readthedocs.io/ru/latest/quickstart.html. (дата обращения: 09.06.2022)