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: Прусакова Александра Михайловна

Москва, 2023

Содержание

Содержание. 2

Введение 3

1. Аналитическая часть 4
   1. Постановка задачи 4
   2. Описание используемых методов. 5
   3. Разведочный анализ данных 14
2. Практическая часть 20
   1. Предобработка данных 20
   2. Разработка и обучение модели. 20
   3. Написать нейронную сеть, которая будет рекомендовать соотношение

«матрица-наполнитель» 23

* 1. Заключение 28
  2. Список используемой литературы и веб ресурсы 29

Введение

Композиционные материалы — это искусственно созданные материалы, состоящие из нескольких других с четкой границей между ними. Композиты обладают теми свойствами, которые не наблюдаются у компонентов по отдельности. При этом композиты являются монолитным материалом, т. е. компоненты материала неотделимы друг от друга без разрушения конструкции в целом. Яркий пример композита - железобетон. Бетон прекрасно сопротивляется сжатию, но плохо растяжению. Стальная арматура внутри бетона компенсирует его неспособность сопротивляться сжатию, формируя тем самым новые, уникальные свойства. Современные композиты изготавливаются из других материалов: полимеры, керамика, стеклянные и углеродные волокна, но данный принцип сохраняется. У такого подхода есть и недостаток: даже если мы знаем характеристики исходных компонентов, определить характеристики композита, состоящего из этих компонентов, достаточно проблематично. Для решения этой проблемы есть два пути: физические испытания образцов материалов, или прогнозирование характеристик. Суть прогнозирования заключается в симуляции представительного элемента объема композита, на основе данных о характеристиках входящих компонентов (связующего и армирующего компонента). На входе имеются данные о начальных свойствах компонентов композиционных материалов (количество связующего, наполнителя, температурный режим отверждения и т.д.). На выходе необходимо спрогнозировать ряд конечных свойств получаемых композиционных материалов. Кейс основан на реальных производственных задачах Центра НТИ

«Цифровое материаловедение: новые материалы и вещества» (структурное подразделение МГТУ им. Н.Э. Баумана). Актуальность: Созданные прогнозные модели помогут сократить количество проводимых испытаний, а также пополнить базу данных материалов возможными новыми характеристиками материалов, и цифровыми двойниками новых композитов.

Традиционно разработка композитных материалов является долгосрочным процессом, так как из свойств отдельных компонентов невозможно рассчитать конечные свойства композита. Для достижения

определенных характеристик требуется большое количество различных комбинированных тестов, что делает насущной задачу прогнозирования успешного решения, снижающего затраты на разработку новых материалов.

* + 1. Аналитическая часть
       1. Постановка задачи

Для работы доступны два файла: X\_bp.xlsx (данные параметров, 1023 строки и 10 столбцов) и X\_nup.xlsx (данные нашивок, 1040 строк и 3 столбца).

Эти два файла необходимо объединить, чтобы разработать модель для прогнозирования модуля упругости при растяжении, прочности при растяжении и соотношения наполнителей матрицы. Поскольку объединение имеет тип INNER, часть информации (17 строк в таблице X\_nup.xlsx) будет удалена, так как в таблице X\_bp.xlsx нет соответствующих строк.

Кроме того, необходимо провести разведочный анализ данных для построения гистограмм, графиков (box-and-whisker diagrams) и точечных парных диаграмм рассеяния распределения каждой переменной.

Для каждого столбца следует получить и проанализировать среднее и медиану, исключить и проверить выбросы. Предварительная обработка: удаление шума и выбросов, нормализация, стандартизация.

Обучить несколько моделей для прогнозирования модуля упругости при растяжении и прочности при растяжении. Создать нейронную сеть для рекомендации соотношения матрицы и наполнителя. Разработать приложение с графическим интерфейсом для прогнозирования соотношения матрицы и наполнителя. Оценить точность моделей на обучающих и тестовых наборах.

* + - 1. Описание используемых методов

Целью алгоритмов контролируемого обучения является определение и минимизация функции потерь, и для получения оптимального решения были рассмотрены (и частично применены) следующие методы:

− 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, который исследует связь между одной входной переменной и выходной переменной. Это один из самых простых и эффективных инструментов статистического моделирования. Для определения зависимости переменных используется прямая линия с наилучшим соответствием. Регрессионные модели дают несколько показателей: R2 (коэффициент детерминации) измеряет, насколько хорошо модель объясняет дисперсию в данных; если R2 равен 1, это означает, что модель объясняет все данные; если R2 равен 0,5, модель объясняет только 50% дисперсии в данных. R2 равен 0,5, это означает, что модель объясняет только 50% дисперсии данных. Остальная дисперсия не может быть объяснена; 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, который уменьшает сложность модели и предотвращает перебор, возникающий при простой линейной регрессии. Этот метод вводит дополнительный член регуляризации при оптимизации модели. Это обеспечивает более надежное решение; регрессия Лассо добавляет условие смещения к функции оптимизации, чтобы уменьшить коллинеарность и, следовательно, дисперсию модели. Однако вместо квадратичного смещения используется смещение по абсолютной величине; регрессия Лассо подходит для прогнозирования моделей временных рядов на основе регрессии, таких как авторегрессия.

*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 имеет проблемы с прогностическим качеством; дает ложноположительные результаты; случайно выбирает одну из коллинеарных переменных; не оценивает правильную форму связи между независимыми и зависимыми 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.

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)*** основан на поиске близлежащих объектов с известными значениями целевой переменной и сохранении данных в памяти для сравнения с новыми объектами. Алгоритм находит расстояние между запросом и всеми объектами и выбирает количество (k) примеров в данных, наиболее близких к запросу, путем голосования за наиболее часто встречающиеся метки (для задач классификации) или путем усреднения меток (для задач 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, вобщей 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

Данные необходимо обрабатывать и очищать, прежде чем их можно будет использовать в моделях машинного обучения. Необработанные данные могут содержать искажения или пропущенные значения, что может привести к неверным результатам моделирования. Однако удалять что-либо без веской причины также является ошибкой. Поэтому набор данных сначала должен быть проверен.

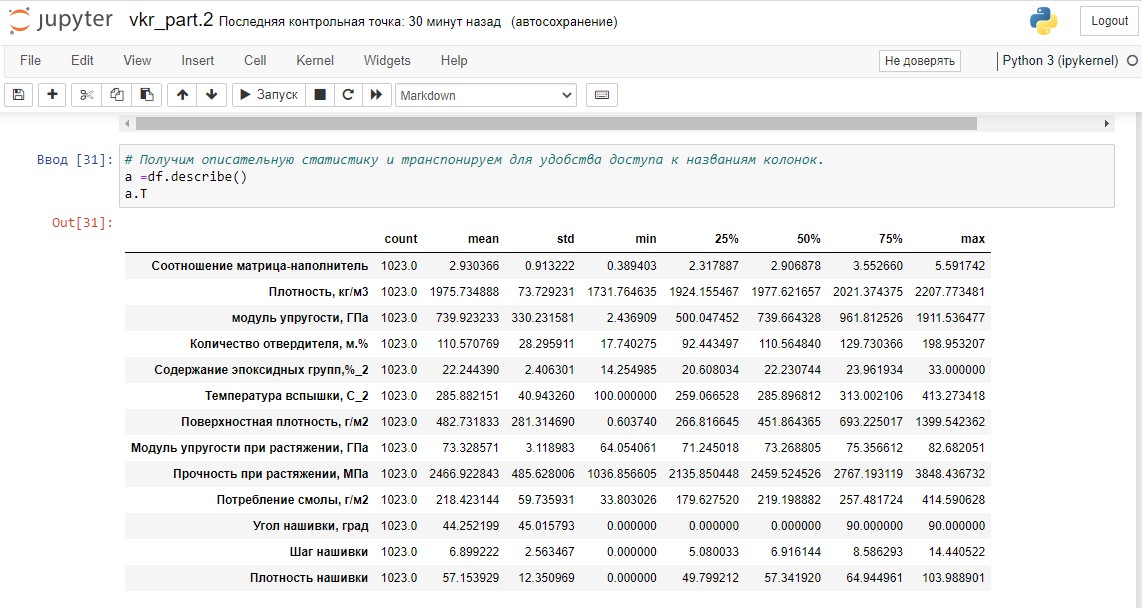
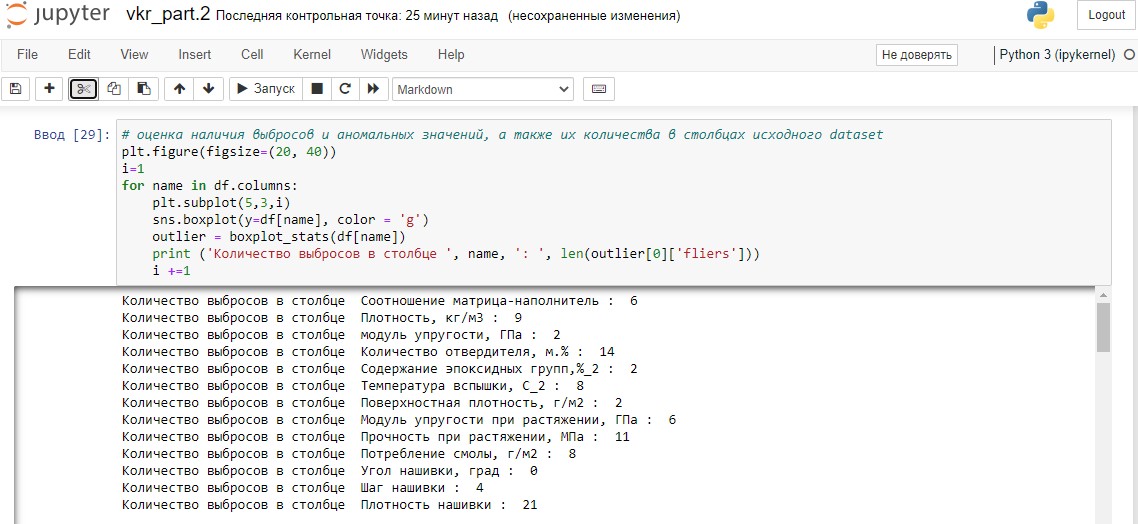


Рисунок 1 - Описательная статистика датасета

Цель разведочного анализа - получить первое представление о характере распределения переменных в исходном наборе данных, сформировать оценку качества исходных данных (пробелы, выбросы) и определить характер взаимосвязей между переменными, чтобы сформулировать гипотезы о модели машинного обучения, наиболее подходящей для решения поставленной задачи.

Рисунок 2 - Проверка датасета на наличие дубликатов

В ходе разведочного анализа использовались следующие инструменты: оценка статистических характеристик набора данных, гистограммы распределения каждой переменной, графики boxplot, диаграммы рассеяния по парам точек, тепловые карты, описательная статистика для каждой переменной, анализ выбросов и полного исключения, проверка на выбросы и перекрытия, корреляция Кендалла и корреляция Пирсона.

Рисунок 3 - Начальное количество выбросов

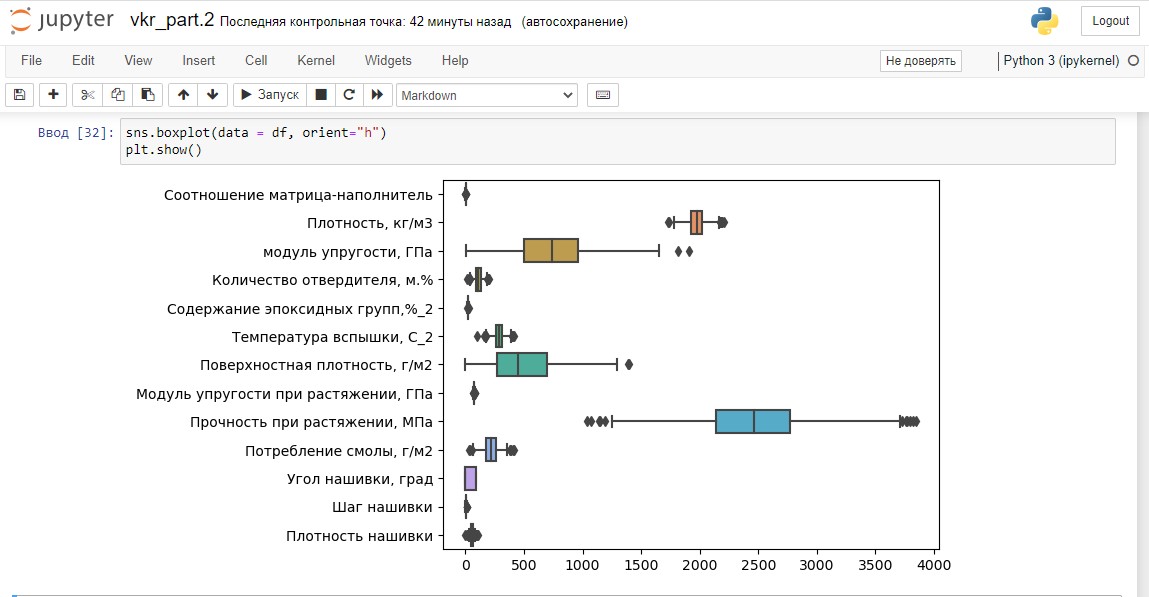
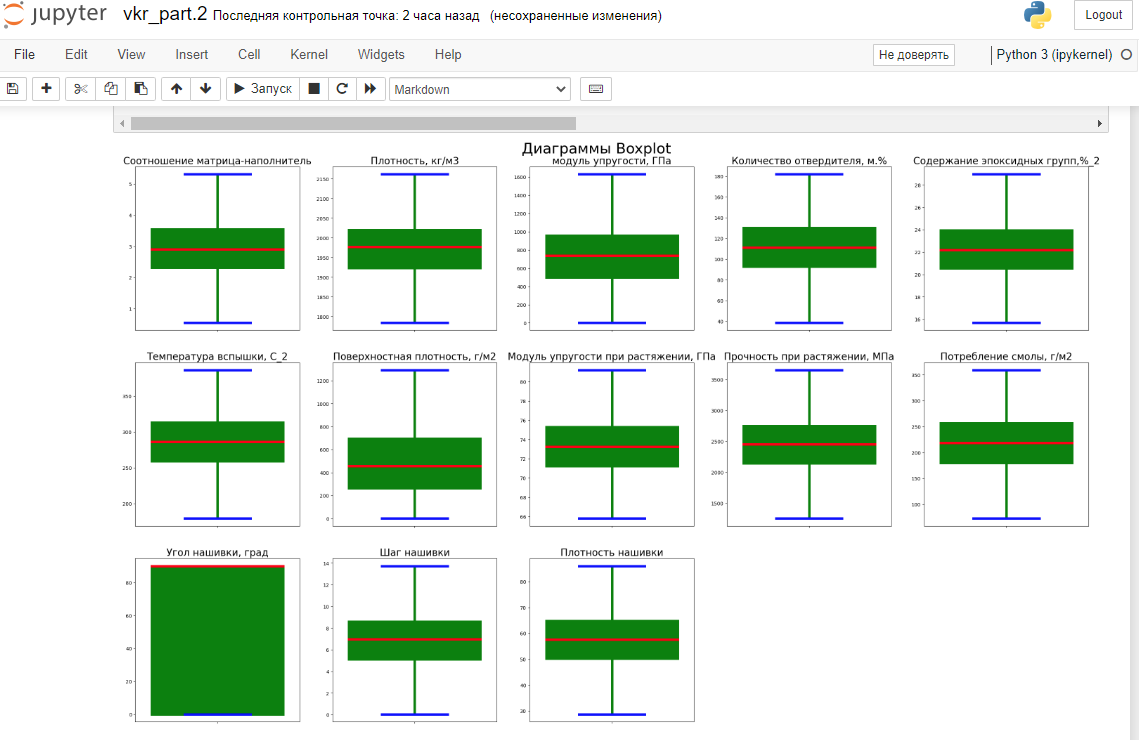


Image 4 - Начальный boxplot

Для удаления выбросов используются методы трех сигм и межквартильного

расстояния. В данном случае удалим способом межквартильного расстояния для максимальной чистоты, так как используем методы чувствительные к выбросам.

Рисунок 5 - Boxplot после удаления выбросов

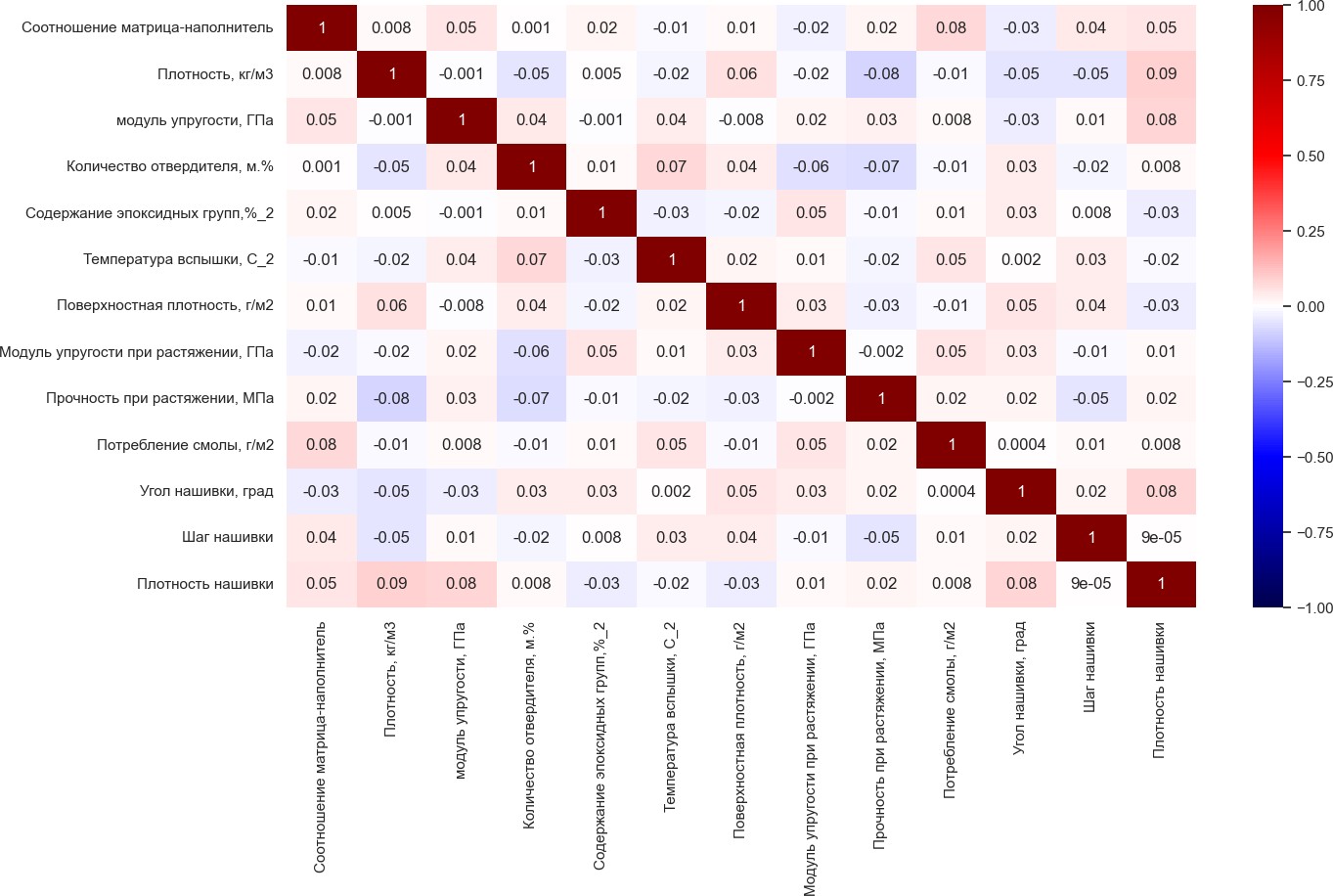


Рисунок 6 - Тепловая карта с корреляцией данных

Тепловая карта показывает практически отсутствие корреляции между признаками и целевыми переменными.

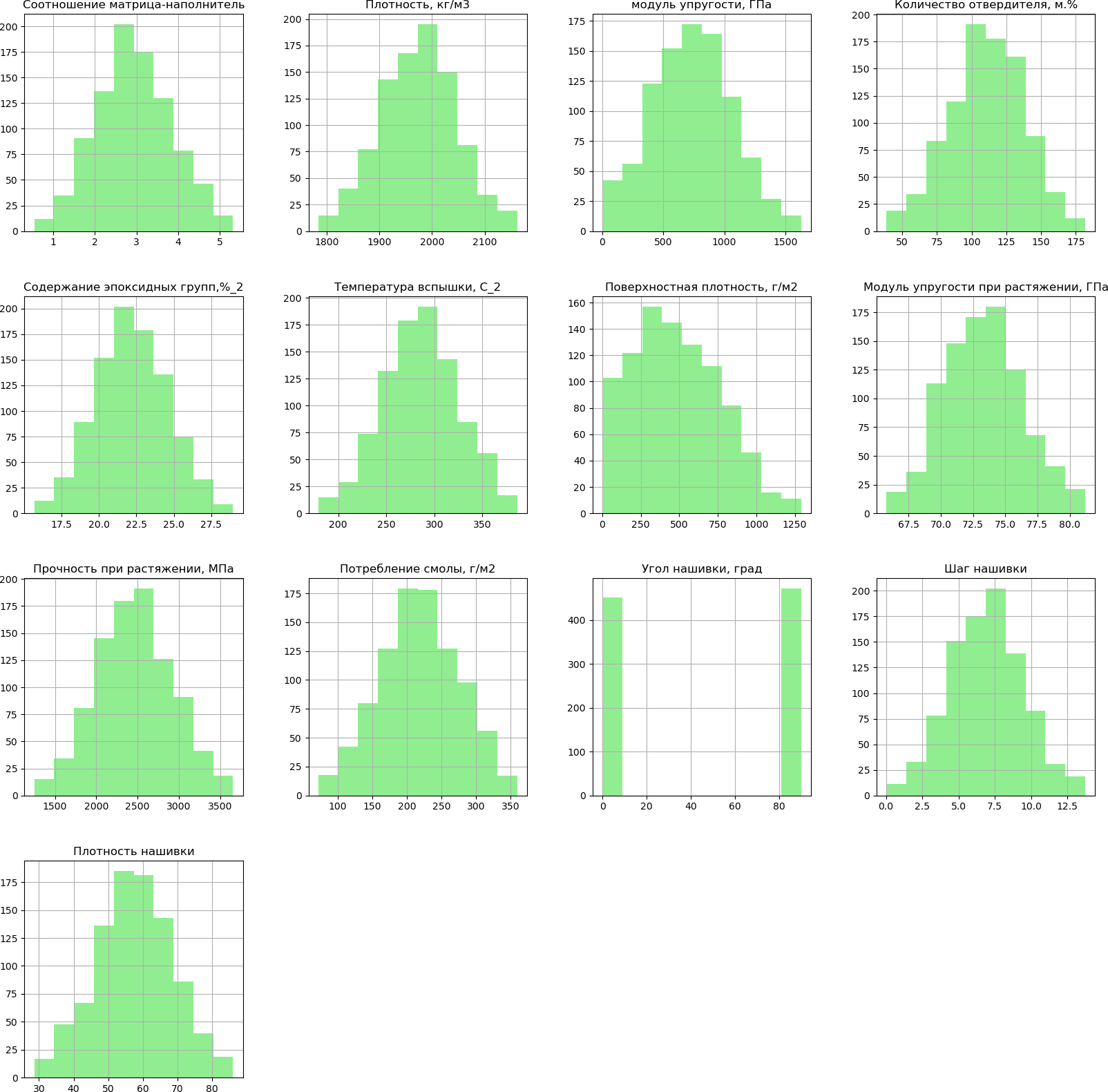


Рисунок 7 - Гистограммы распределения

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

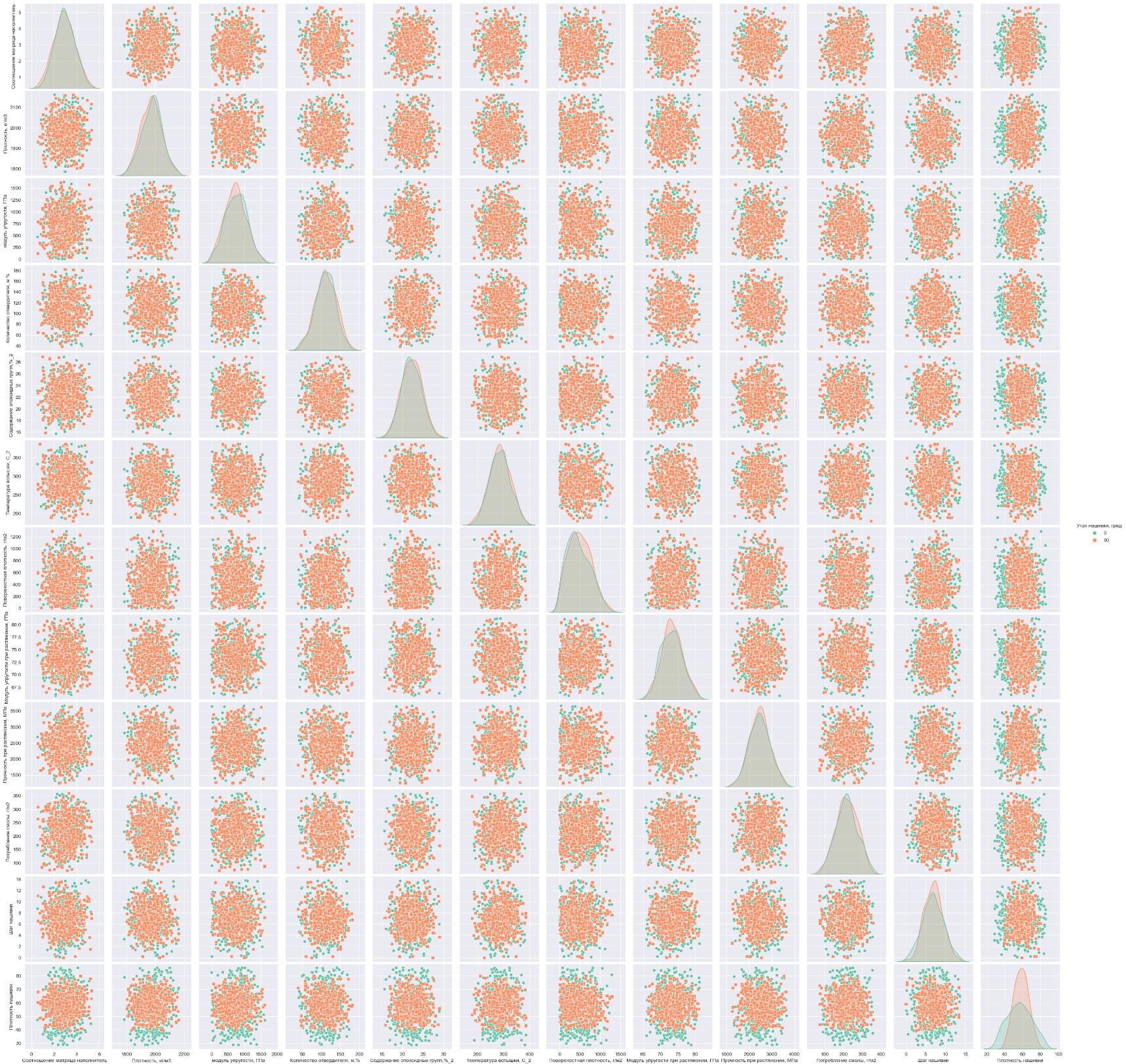


Рисунок 8 - Попарные графики рассеяния точек с выделением значений Угол нашивки

На попарных графиках распределения не видно корреляции между признаками.

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

Максимальная корреляция между плотностью пятен и углом наклона пятен составляет 0,11, и корреляция между этими данными отсутствует. Корреляции между всеми параметрами очень близки к нулю, и корреляции между переменными не обнаружено.

* + 1. Практическая часть
       1. Предобработка данных

По условиям задания нормализуем значения. Для этого применим MinMaxScaler.

* + - 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.

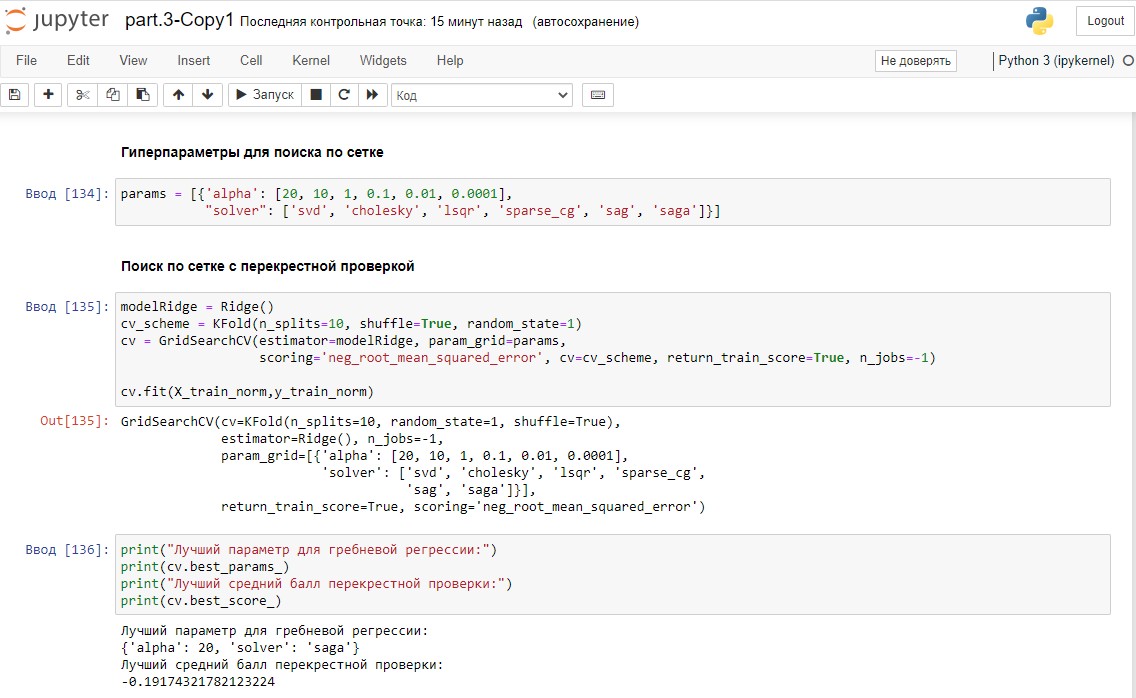
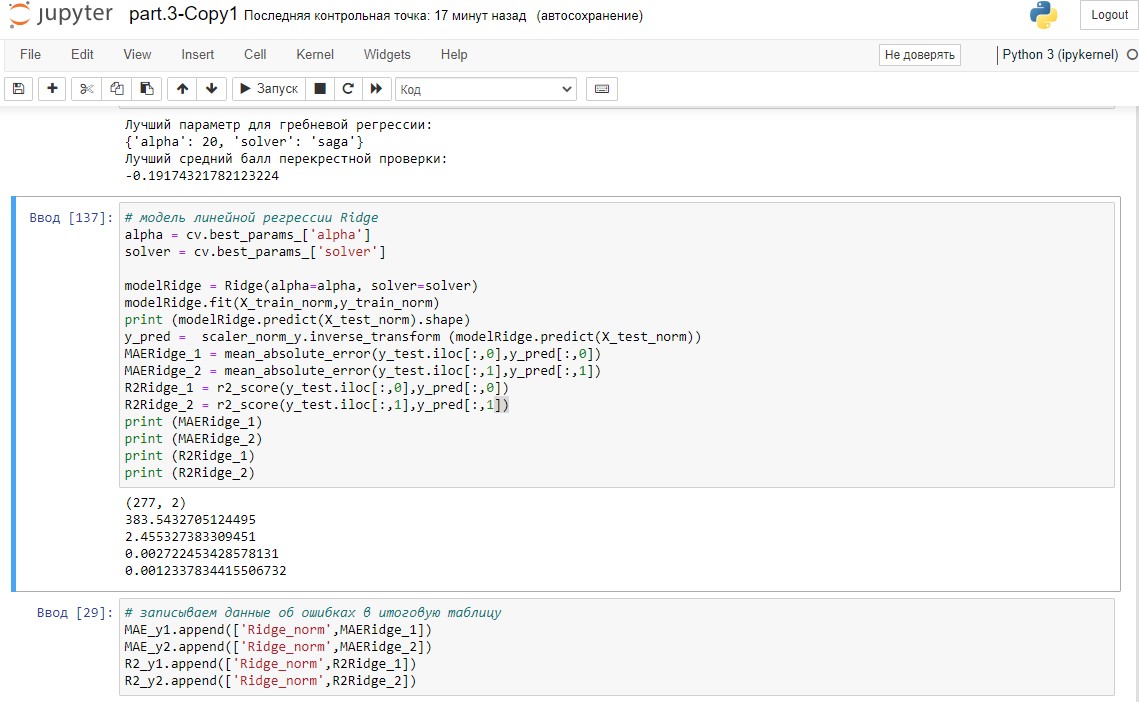


Рисунок 9- Поиск гиперпараметров по сетке

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 10 - Модель Ridge регрессия

После настройки гиперпараметров производительность моделей немного улучшилась. В результате все модели показали практически одинаковые результаты: ошибка MAE была почти равна стандартному отклонению, а значение R2 было около нуля, что означает, что все модели предсказывали результаты, сравнимые 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

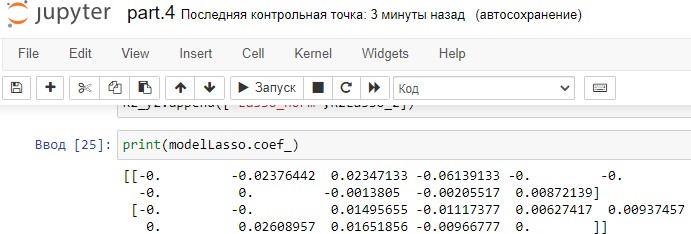
Lasso и Ridge.

Рисунок 11 – Коэффициенты признаков Lasso регрессии

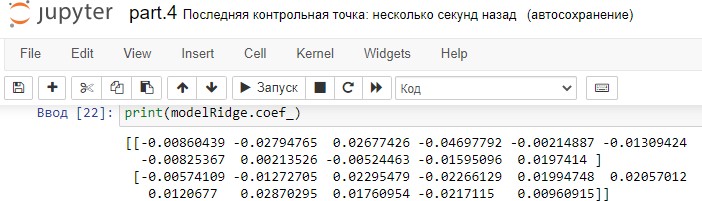
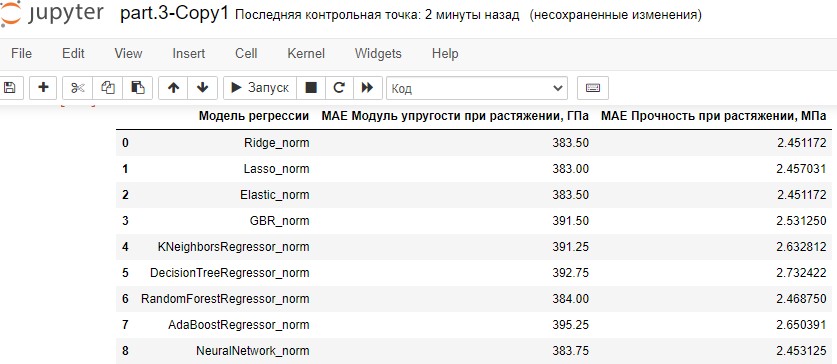


Рисунок 12 – Коэффициенты признаков Ridge регрессии

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



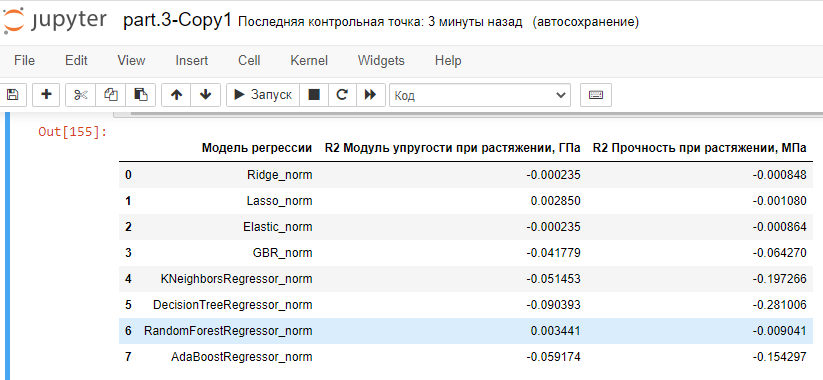
Рисунок 13 - Оценка MAE результатов работы модели

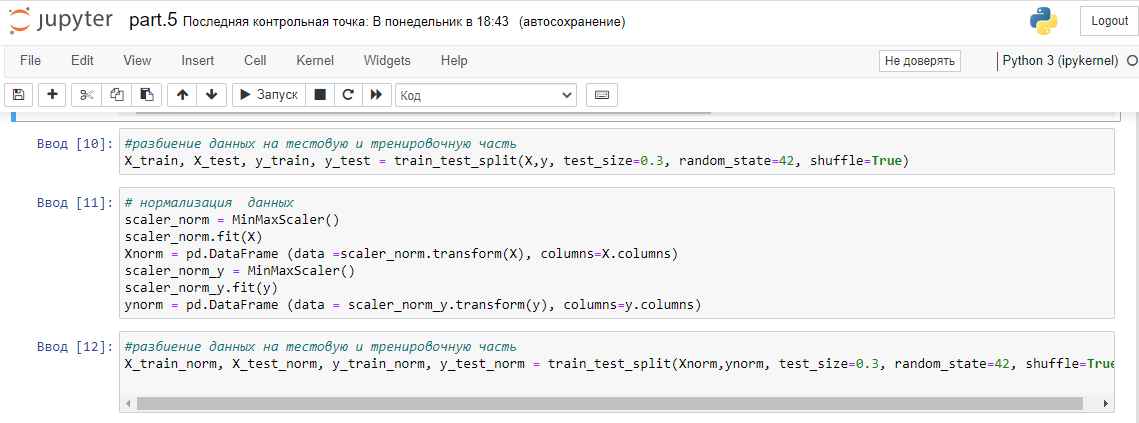
Рисунок 14 - Оценка R2 результатов работы моделей

* 1. Нейронная сеть для рекомендации «Соотношение

матрица-наполнитель».

Загружаем очищенный датасет, для X удаляем целевой столбец

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

Рисунок 15 – 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.

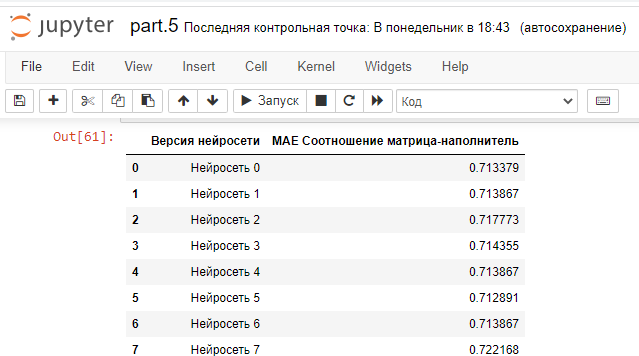


Рисунок 16 – Ошибка MAE по результатам работы нейросетей с различной архитектурой

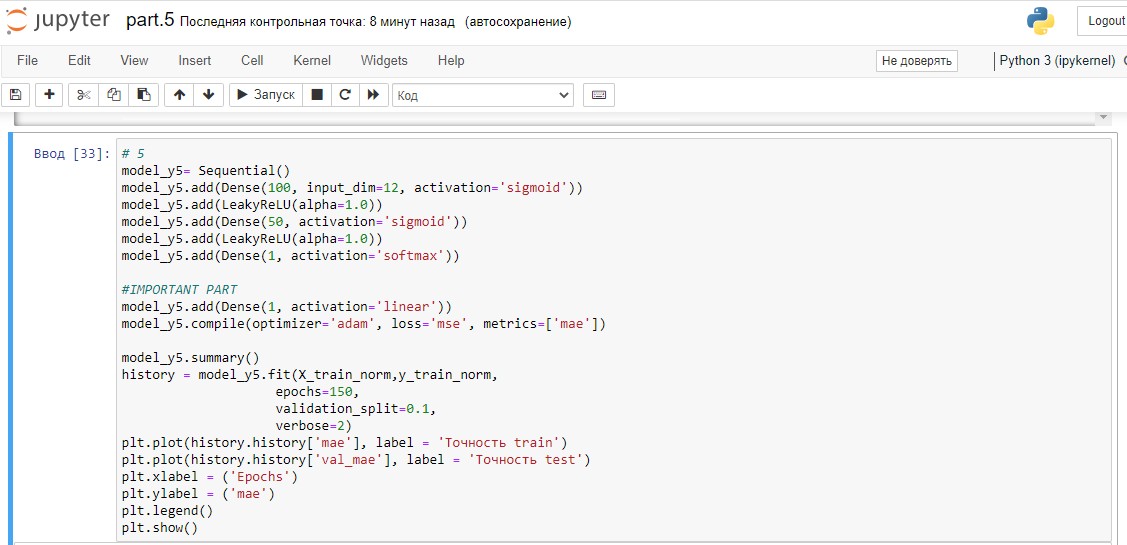


Рисунок 17 — Код нейросети с наименьшей ошибкой

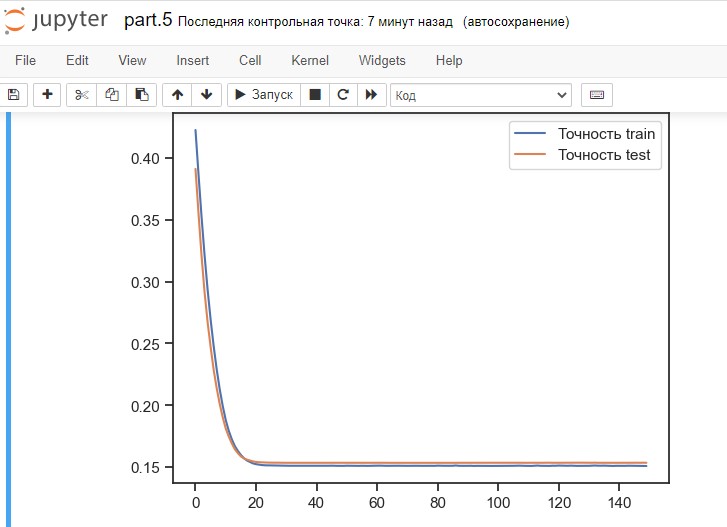


Рисунок 18 — Визуализация работы нейросети с наименьшей ошибкой

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.

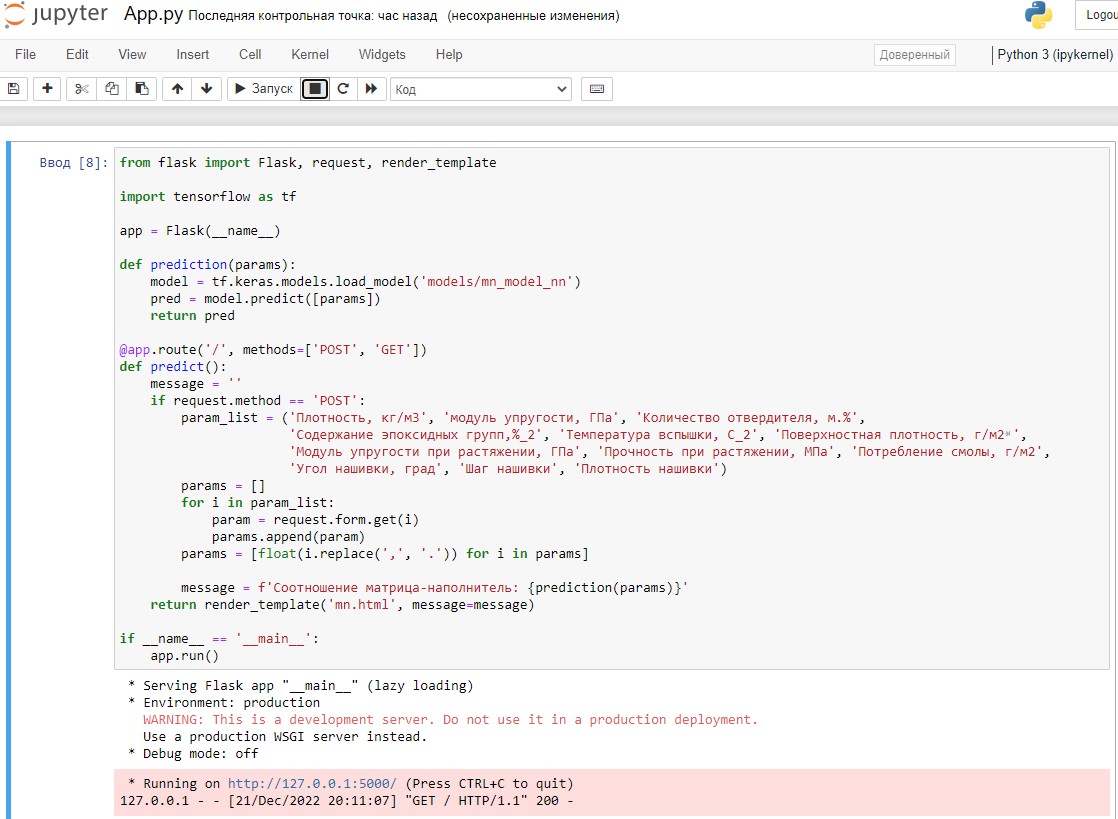


Рисунок 19 - Код приложения

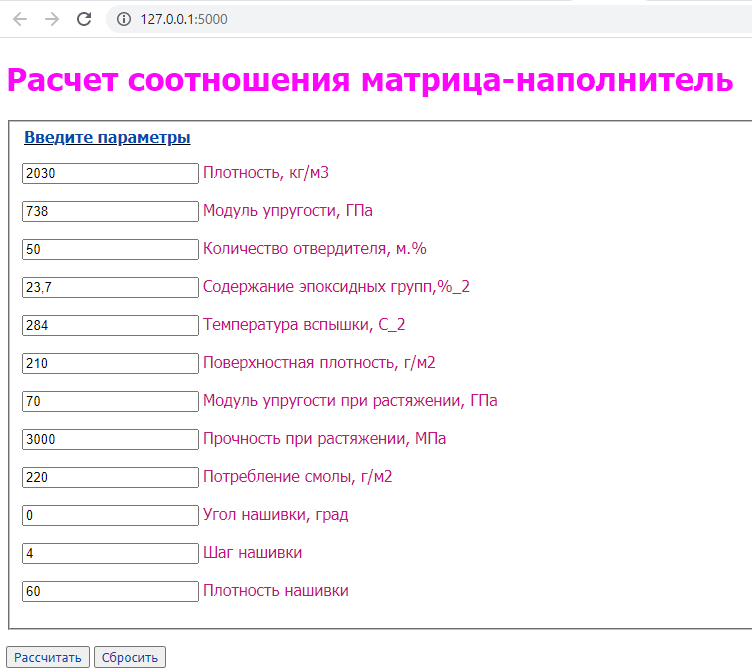
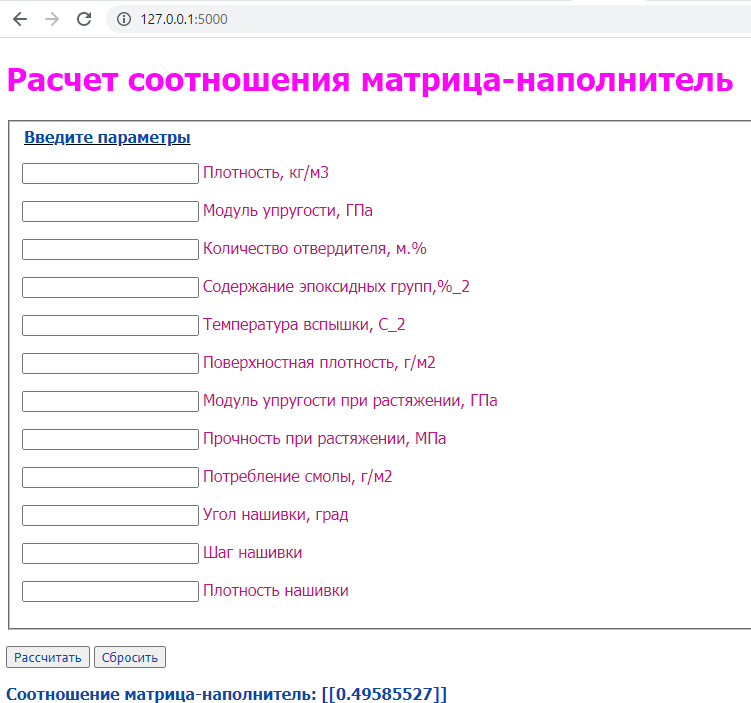


Рисунок 20 - Форма пользовательского приложения для ввода параметров

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



* 1. Рисунок 21 – Результат расчета

Создание удалённого репозитория и загрузка Репозиторий был создан на github.com по адресу :

* 1. Заключение

Данная исследовательская работа позволяет сделать некоторые основные выводы по данному вопросу. Распределение данных, полученных из объединенного набора данных, близко к нормальному распределению, а корреляция между парами признаков близка к нулю. Подход, использованный для разработки модели, не дал надежных прогнозов. Использованная регрессионная модель не показала эффективности в прогнозировании свойств композитного материала.

Можно сделать вывод, что по свойствам материала невозможно определить взаимосвязь между матрицей и наполнителем. Этот факт не говорит о невозможности прогнозирования свойств композитных материалов на основе предоставленного набора данных, но он может указывать на недостатки базы данных, подхода, использованного для прогнозирования, и необходимость пересмотра инструментов прогнозирования.

Необходимы дополнительные исходные данные, получение новых результирующих признаков в результате математических преобразований, консультации с профильными экспертами и новые исследования. В целом, не изучив материаловедение, нельзя шагнуть в проблему экспериментального анализа свойств композитных материалов и предсказать конечные свойства и характеристики композитных материалов без получения удовлетворительных результатов. Учитывая отсутствие корреляции между признаками, можно сделать вывод, что существующий набор алгоритмов не решает поставленную задачу, возможно, с трудом или не решает вовсе.

* 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.
11. Документация по библиотеке numpy: https://numpy.org/doc/1.22/user/index.html#user.
12. Документация по библиотеке pandas: https://pandas.pydata.org/docs/user\_guide/index.html#user-guide.
13. Документация по библиотеке scikit-learn: https://scikit-learn.org/stable/user\_guide.html.
14. Документация по библиотеке seaborn: https://seaborn.pydata.org/tutorial.html.
15. Документация по библиотеке Tensorflow: https://[www.tensorflow.org/overview](http://www.tensorflow.org/overview)
16. Документация по языку программирования python: https://docs.python.org/3.8/index.html.
17. Иванов Д.А., Ситников А.И., Шляпин С.Д – Композиционные материалы: учебное пособие для вузов, 2019. 13 с.
18. Краткий обзор алгоритма машинного обучения Метод Опорных Векторов (SVM) : https://habr.com/ru/post/428503/
19. Ларин А. А., Способы оценки работоспособности изделий из композиционных материалов методом компьютерной томографии, Москва, 2013, 148 с.
20. [Макс](https://your-scorpion.ru/) Ветков. Градиентный бустинг(AdaBoost) https://your- scorpion.ru/gradient-boosting-adaboost
21. Материалы конференции: V Всероссийская научно-техническая конференция «Полимерные композиционные материалы и производственные технологии нового поколения», 19 ноября 2021 г.
22. Плас Дж. Вандер, Python для сложных задач: наука о данных и машинное обучение. Санкт-Петербург: Питер, 2018, 576 с
23. 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.
24. Роббинс, Дженнифер. HTML5: карманный справочник, 5-е издание.: Пер. с англ. - М.: ООО «И.Д. Вильямс»: 2015. - 192 с.: ил